HYBRID BIOMOLECULAR MODELING

EDITED BY: Slavica Jonic, Osamu Miyashita and Isabelle Callebaut

PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-699-4
DOI 10.3389/978-2-88945-699-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

HYBRID BIOMOLECULAR MODELING

Topic Editors:
Slavica Jonic, Sorbonne Université, UMR CNRS 7590, Muséum National d'Histoire Naturelle, IRD, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie - IMPMC, France
Osamu Miyashita, RIKEN Center for Computational Science, Japan
Isabelle Callebaut, Sorbonne Université, UMR CNRS 7590, Muséum National d'Histoire Naturelle, IRD, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie - IMPMC, France

Models of biomolecular structure and dynamics are often obtained by combining simulation or prediction approaches (e.g., comparative modeling, Molecular Dynamics (MD) simulations, Normal Mode Analysis (NMA), etc.) with experimental approaches (e.g., Nuclear Magnetic Resonance (NMR), X-ray crystallography, Small-Angle X-ray Scattering (SAXS), Electron Microscopy (EM), etc.). Such hybrid modeling extends the capabilities of experimental techniques, by enriching structural information and facilitating dynamics studies of biomolecules. This eBook contains articles on methodological developments, applications, and challenges of hybrid biomolecular modeling that have been collected in the framework of the Frontiers Research Topic entitled "Hybrid Biomolecular Modeling".

Citation: Jonic, S., Miyashita, O., Callebaut, I., eds. (2019). Hybrid Biomolecular Modeling. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-699-4

Table of Contents

EDITORIAL
published: 09 November 2018
doi: 10.3389/fmolb.2018.00098

Editorial: Hybrid Biomolecular Modeling

Slavica Jonic[1]*, *Osamu Miyashita*[2] *and Isabelle Callebaut*[1]

[1] *Sorbonne Université, UMR CNRS 7590, Muséum National d'Histoire Naturelle, IRD, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, Paris, France,* [2] *RIKEN Center for Computational Science, Kobe, Japan*

Keywords: structure, dynamics, Conformational variability, hybrid models, biomolecules

Editorial on the Research Topic

Hybrid Biomolecular Modeling

Models of biomolecular structure and dynamics are often obtained by combining simulation or prediction approaches [e.g., comparative modeling, Molecular Dynamics (MD) simulations, Normal Mode Analysis (NMA), etc.] with experimental approaches [e.g., Nuclear Magnetic Resonance (NMR), X-ray crystallography, Small-Angle X-ray Scattering (SAXS), Electron Microscopy (EM), etc.] (Sali et al., 2015) (**Figure 1**). Such hybrid modeling extends the capabilities of experimental techniques, by enriching structural information and facilitating dynamics studies of biomolecules. This e-Book contains articles on methodological developments, applications, and challenges of hybrid biomolecular modeling that have been collected in the framework of the Frontiers Research Topic entitled "Hybrid Biomolecular Modeling."

An example of hybrid modeling is fitting of structures of protein domains obtained by X-ray crystallography, NMR, or structure prediction into EM density maps of protein complexes (Kawabata, 2008; Birmanns et al., 2011; Tjioe et al., 2011; Yang et al., 2012). This allows obtaining high-resolution models of complexes when this cannot be achieved using a single experimental technique, as is often the case with large and flexible complexes (Cottevieille et al., 2008; Ciferri et al., 2012; Brown et al., 2014). This problem is addressed in the article by Habeck. The article is focused on a Bayesian inference approach to integrative biomolecular modeling by combining X-ray crystallography and cryo-EM data, but Habeck also discusses the computational challenges of this approach in a more general context of integrating other experimental data such as cross-linking/mass spectrometry and solid-state NMR data. The proposed approach is based on probabilistic models for cryo-EM maps and Markov chain Monte Carlo sampling of model structures from the posterior distribution.

Computational methods have been developed to predict the interactions between the protein subunits based on their shape complementary, electrostatic interactions, solvation energy, and statistical potential energy derived from the structural databases. This is known as molecular docking and one of its main challenges is the design of a reliable scoring function to assess the model quality. Inspired by the application of X-ray Free-Electron Lasers (XEFL) data in scoring models of conformational changes of complexes (Tokuhisa et al., 2016), Wang and Liu propose to use single-particle XFEL data for a more reliable scoring of models obtained by docking methods.

Computational approaches based on NMA or MD simulations have been developed to explore the conformational space of a model and identify the conformation (in this space) that best agrees with experimental data (Trabuco et al., 2008; Gorba and Tama, 2010; Jin et al., 2014). Devaurs et al. address this problem in the context of modeling based on experimental hydrogen/deuterium exchange (HDX) data. HDR data is often interpreted using an X-ray crystallography structure or a conformational ensemble obtained by MD simulations, though their correspondence with the HDR data is often not enough satisfactory. Devaurs et al. propose

OPEN ACCESS

Edited by:
Dieter W. Heermann,
Universität Heidelberg, Germany

Reviewed by:
Peter Virnau,
Johannes Gutenberg-Universität Mainz, Germany

Correspondence:
Slavica Jonic
slavica.jonic@upmc.fr

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *20 September 2018*
Accepted: *23 October 2018*
Published: *09 November 2018*

Citation:
Jonic S, Miyashita O and Callebaut I (2018) Editorial: Hybrid Biomolecular Modeling. Front. Mol. Biosci. 5:98. doi: 10.3389/fmolb.2018.00098

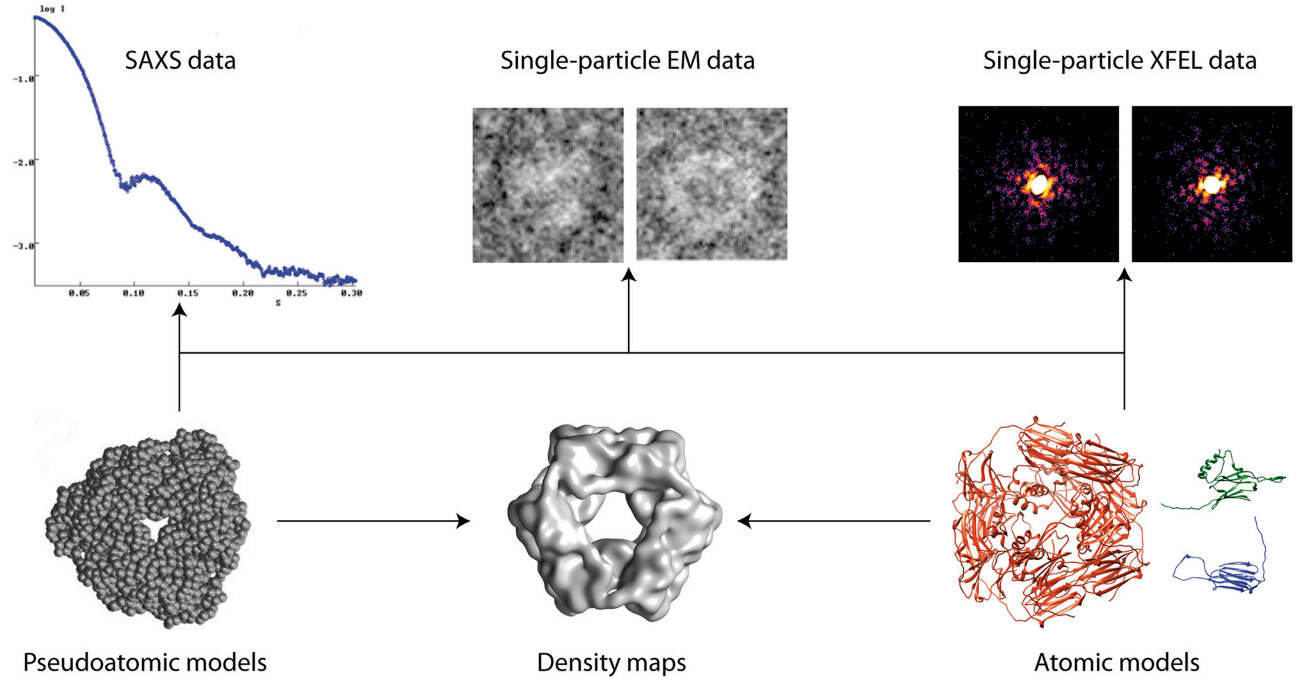


FIGURE 1 | Hybrid modeling of biomolecular structures by fitting experimental data (arrows). Other data can also be used (e.g., NMR, cross-linking, FRET, etc.).

to select a single conformation that best fits the HDX data, from the conformational ensemble obtained with an extensive coarse-grained conformational sampling (of the given X-ray crystallography structure) that is biased with the information on the protein regions that produce the largest discrepancies with the HDX data.

Prischi and Pastore review an integrative structural modeling methodology that they have developed to determine the structure of weakly interacting molecular complexes. It combines NMR, SAXS, site directed mutagenesis, molecular docking, and MD simulations, and has been used by the authors and other groups to gain structural information on several iron-sulfur cluster (ISC) biogenesis complexes. The authors review these applications and discuss the advantages and limitations of this methodology as well as the future directions to improve it.

Woods et al. show a new application of an approach combining MD simulations, evolutionary sequence analysis, and Terahertz spectroscopy that they have developed to probe dynamics and allostery in rhodopsin. They show how the binding of the chlorophyll derivative, chlorin-e6 (Ce6) allosterically excites evolutionarily conserved communication pathways in rhodopsin that connect the ligand-binding site and the rest of the receptor.

Hsieh et al. present a NMA approach to analyze the dynamics of Dengue and Zika virus capsids based on their high-resolution cryo-EM models. They relate the differences identified in the dynamics of the E proteins in the two capsids to the differences observed in the two high-resolution models. They discuss the work that should be done in the future in order to fully characterize the dynamics of the two viruses.

Intrinsically disordered peptides and proteins present a challenge to experimental characterization of their functional conformations. Olson explores simulation techniques that could be used to build a computational framework for capturing conformational ensembles of such peptides and proteins. He explores temperature-based replica exchange methods for conformational ensemble sampling with implicit solvent models, as well as, explicit/implicit solvent hybrid replica exchange methods to capture the conformational ensemble of an intrinsically disordered peptide derived from the Ebola virus protein VP35. The author points out that intrinsically disordered peptides and proteins can be used as benchmarks to develop accurate methods for modeling conformational transitions.

The permeability of a cell membrane can be increased under the influence of an electric field of sufficient magnitude, which is known as membrane electroporation. Wriggers et al. address the problem of experimental and theoretical investigation of membrane electroporation. They extend, to the context of lipid bilayers and solvents, a statistical approach that they have originally developed for detecting allosteric signatures in MD simulations of well-structured proteins. This method is based on transforming time-domain information from MD trajectories into spatial heat maps that can be visualized on 3D molecular structures or in the form of interaction networks. The method is multiscale in the time domain and uses a mutual information approach for statistical bridging between the fast (local variables recorded by MD) and slow (global rate of change) time series. The mutual information method used with proteins was adapted to lipids and solvents by developing a new approach to probability density function estimation of random variables, which was described in a separate article (Kovacs et al.) in this e-Book.

We hope that this e-Book will be useful to experimentalists and method developers and that it will stimulate further use and development of hybrid biomolecular modeling methods. We thank all authors, co-authors, and reviewers for their contribution to this Research Topic and acknowledge the support from Frontiers Team members.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Birmanns, S., Rusu, M., and Wriggers, W. (2011). Using Sculptor and Situs for simultaneous assembly of atomic components into low-resolution shapes. *J. Struct. Biol.* 173, 428–435. doi: 10.1016/j.jsb.2010.11.002

Brown, A., Amunts, A., Bai, X. C., Sugimoto, Y., Edwards, P. C., Murshudov, G., et al. (2014). Structure of the large ribosomal subunit from human mitochondria. *Science* 346, 718–722. doi: 10.1126/science.1258026

Ciferri, C., Lander, G. C., Maiolica, A., Herzog, F., Aebersold, R., and Nogales, E. (2012). Molecular architecture of human polycomb repressive complex 2. *Elife* 1:e00005. doi: 10.7554/eLife.00005

Cottevieille, M., Larquet, E., Jonic, S., Petoukhov, M. V., Caprini, G., Paravisi, S., et al. (2008). The subnanometer resolution structure of the glutamate synthase 1.2-MDa hexamer by cryoelectron microscopy and its oligomerization behavior in solution: functional implications. *J. Biol. Chem.* 283, 8237–8249. doi: 10.1074/jbc.M708529200

Gorba, C., and Tama, F. (2010). Normal mode flexible fitting of high-resolution structures of biological molecules toward SAXS data. *Bioinform. Biol. Insights* 4, 43–54. doi: 10.4137/BBI.S4551

Jin, Q., Sorzano, C. O., de la Rosa-Trevín, J. M., Bilbao-Castro, J. R., Núñez-Ramírez, R., Llorca, O., et al. (2014). Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure* 22, 496–506. doi: 10.1016/j.str.2014.01.004

Kawabata, T. (2008). Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophys. J.* 95, 4643–4658. doi: 10.1529/biophysj.108.137125

Sali, A., Berman, H. M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S. K., et al. (2015). Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* 23, 1156–1167. doi: 10.1016/j.str.2015.05.013

Tjioe, E., Lasker, K., Webb, B., Wolfson, H. J., and Sali, A. (2011). MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map. *Nucleic Acids Res.* 39, W167–W170. doi: 10.1093/nar/gkr490

Tokuhisa, A., Jonic, S., Tama, F., and Miyashita, O. (2016). Hybrid approach for structural modeling of biological systems from X-ray free electron laser diffraction patterns. *J. Struct. Biol.* 194, 325–336. doi: 10.1016/j.jsb.2016.03.009

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16, 673–683. doi: 10.1016/j.str.2008.03.005

Yang, Z., Lasker, K., Schneidman-Duhovny, D., Webb, B., Huang, C. C., Pettersen, E. F., et al. (2012). UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J. Struct. Biol.* 179, 269–278. doi: 10.1016/j.jsb.2011.09.006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

METHODS
published: 22 March 2017
doi: 10.3389/fmolb.2017.00015

Bayesian Modeling of Biomolecular Assemblies with Cryo-EM Maps

Michael Habeck[1,2]*

[1] *Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany,* [2] *Felix Bernstein Institute for Mathematical Statistics in the Biosciences, University of Göttingen, Göttingen, Germany*

A growing array of experimental techniques allows us to characterize the three-dimensional structure of large biological assemblies at increasingly higher resolution. In addition to X-ray crystallography and nuclear magnetic resonance in solution, new structure determination methods such cryo-electron microscopy (cryo-EM), crosslinking/mass spectrometry and solid-state NMR have emerged. Often it is not sufficient to use a single experimental method, but complementary data need to be collected by using multiple techniques. The integration of all datasets can only be achieved by computational means. This article describes Inferential structure determination, a Bayesian approach to integrative modeling of biomolecular complexes with hybrid structural data. I will introduce probabilistic models for cryo-EM maps and outline Markov chain Monte Carlo algorithms for sampling model structures from the posterior distribution. I will focus on rigid and flexible modeling with cryo-EM data and discuss some of the computational challenges of Bayesian inference in the context of biomolecular modeling.

Keywords: cryo-EM, modeling, Bayesian inference, Markov chain Monte Carlo, inferential structure determination

OPEN ACCESS

Edited by:
Slavica Jonic,
IMPMC, Sorbonne Universités - CNRS UMR 7590, UPMC Univ Paris 6, MNHN, IRD UMR 206, France

Reviewed by:
Gunnar F. Schroeder,
Forschungszentrum Jülich, Germany
Riccardo Pellarin,
Centre National de la Recherche Scientifique (CNRS), France

***Correspondence:**
Michael Habeck
mhabeck@gwdg.de

Specialty section:
This article was submitted to Biophysics,
a section of the journal
Frontiers in Molecular Biosciences

Received: *01 December 2017*
Accepted: *07 March 2017*
Published: *22 March 2017*

Citation:
Habeck M (2017) Bayesian Modeling of Biomolecular Assemblies with Cryo-EM Maps.
Front. Mol. Biosci. 4:15.
doi: 10.3389/fmolb.2017.00015

1. INTRODUCTION

Thanks to groundbreaking advances in experimental techniques it has become possible to study the structure of large biological assemblies at increasingly higher resolution. Traditionally, high-resolution biomolecular structure determination was only possible by X-ray crystallography or nuclear magnetic resonance (NMR) in solution (Berman et al., 2000). The application of NMR and X-ray crystallography to larger systems remained challenging due to the sheer size of the system and/or because it was difficult to find suitable crystallization conditions. More recently, emerging methods such as cryo-electron microscopy (cryo-EM) (Frank, 2002; Orlova and Saibil, 2004; Chiu et al., 2005), crosslinking/mass spectrometry (Gingras et al., 2007; Rappsilber, 2011) and solid-state NMR (Yan et al., 2013) have started to provide exciting insights into the structure of large macromolecular assemblies that was previously very difficult, if not impossible to obtain. In particular, cryo-EM has reached near-atomic and in some cases even atomic resolution over the last 5 years (Bai et al., 2015; Fischer et al., 2015; Khatter et al., 2015). The EM databank (EMDB) (Lawson et al., 2011) stores an increasing number of high-resolution EM reconstructions. Several biologically essential assemblies that resisted high-resolution studies have recently been characterized by cryo-EM including spliceosomal complexes (Yan et al., 2015; Agafonov et al., 2016; Galej et al., 2016; Rauhut et al., 2016; Wan et al., 2016), eukaryotic ribosomes (Anger et al., 2013; Khatter et al., 2015), and transcription initiation complexes (Plaschka et al., 2015).

Although several powerful experimental techniques are available that allow us to study the structure of large biomolecular systems, we need computational methods that assist us in integrative modeling with diverse structural data (Sali et al., 2003; Robinson et al., 2007; Ward et al., 2013). The reasons for developing new computational methods are both of a principled and practical nature.

Structural models built from hybrid data should be as objective as possible and ideally not be biased by a human modeler, therefore automated computational modeling tools are indispensable (Karaca and Bonvin, 2013; Villa and Lasker, 2014; Schröder, 2015). The models should be compatible with all of the available data, which might come from different experimental sources. The modeling software should also be able to integrate data-independent prior information about the system.

Most existing refinement and modeling software focuses on structural data of a particular type. For example, a number of software packages for X-ray structure refinement or modeling with NMR restraints exist. To use these packages for modeling with hybrid data is often difficult and involves some sort of tweaking. We therefore need a versatile software that can integrate diverse types of structural information (Russel et al., 2012).

Every software for integrative modeling with hybrid data has to address the following questions: How much weight should the various pieces of information be given? How to deal with datasets that (partially) contradict some of the other datasets? Obviously, the weights can have a strong impact on the final structure (Brünger, 1992; Habeck et al., 2006), and it would be desirable to choose the weights in a data-driven, self-adaptive fashion. Because the individual datasets themselves typically provide only ambiguous structural information, we have to fit the model against all data simultaneously to obtain the least ambiguous result. What is a good representation of the remaining uncertainty about the structure? We need to represent the ambiguity of the structural model adequately.

The software should also be able to integrate data of varying resolution. A common scenario is that high-resolution information about the subunits in isolation is available (Esquivel-Rodríguez and Kihara, 2013), such that modeling the complex appears to be simple: we just need to put the pieces together. However, even in this seemingly simple situation several issues need to be considered.

The formation of the complex is often accompanied by a conformational change in the subunits (Gerstein et al., 1994). How much should we deviate from the known structures of the free subunits in order to fit the data of the complex? If the data is sparse (e.g., crosslinking or NMR data) or of a medium resolution, there is the risk of overfitting the data.

Another practical problem is the enormous size of the systems that can comprise tens of thousands up to millions of atoms. Is there enough information to determine the position of all atoms? Or should we rather lower our goal and aim for a coarse-grained, intermediate resolution model?

At the source of many of these issues is the question of how to deal with uncertainty in the data and about our model. We need a mathematical framework to quantitatively represent any uncertainty in the process that takes us from the input data to the final model. The framework should allow us to follow the propagation of the uncertainty about a biomolecular structure as we combine data from diverse sources and to compute structural error bars that reflect the degree of uncertainty.

Bayesian probability theory is a unique and objective mathematical framework for quantitative inference from limited, diverse and uncertain information (Cox, 1946; Jaynes, 2003; MacKay, 2003). The essence of the Bayesian approach is that any probability should be interpreted as incomplete information about a quantity rather than a frequency of occurrence. Highly ambiguous and uncertain information results in multi-modal distributions that are spread out over many parameter values. Markov chain Monte Carlo (MCMC) methods (Liu, 2001) allow us to apply the Bayesian formalism in practice even to highly complex data and models.

More than a decade ago, Bayesian methods have been introduced for protein structure determination from solution NMR data (Rieping et al., 2005; Habeck, 2012). In this article, I will describe recent developments in Bayesian integrative modeling with hybrid data.

2. METHODS

2.1. Inferential Structure Determination

Inferential structure determination (ISD) is the first strictly statistical approach to biomolecular modeling (Habeck et al., 2005a; Rieping et al., 2005). Originally ISD was developed for solution NMR data on small protein domains (Rieping et al., 2008; Habeck, 2012). But the basic principle can be applied to large systems and diverse structural data (Bayrhuber et al., 2008; Shahid et al., 2012; Habenstein et al., 2015).

At the core of the ISD approach is a probabilistic formulation of the structure determination problem. We have to distinguish two principal types of information that guide us in the modeling of a biomolecular structure: the experimental data D and data-independent prior information I about biomolecular structures. All the information is encoded statistically through conditional probabilities. The probability:

$$\Pr(D|\theta, I)$$

quantifies how probable it is to observe data D if the actual configuration of the system is θ. $\Pr(D|\theta, I)$ is called the *likelihood* function. The prior probability:

$$\Pr(\theta|I)$$

expresses what we know about reasonable system configurations θ without observing any data.

Probability calculus allows us to combine both types of information and to derive a *posterior* distribution over all conformational degrees of freedom by invoking Bayes' theorem (Jaynes, 2003):

$$\Pr(\theta|D, I) = \frac{1}{\Pr(D|I)} \Pr(D|\theta, I) \Pr(\theta|I) .$$

The posterior $\Pr(\theta|D, I)$ expresses what we know about the unknown structure given the experimental data D and our prior knowledge I. The probability $\Pr(D|I)$ (the so-called model evidence) can be ignored if we are only interested in estimating θ, because $\Pr(D|I)$ does not depend on θ. However, if we aim to compare different prior or modeling assumptions, it will be important to calculate $\Pr(D|I)$ (Habeck, 2011; Mechelke and Habeck, 2012, 2014; Knuth et al., 2015).

Often, we need to introduce additional unknown parameters to express our prior information or to model the experimental data. Let's denote these parameters by ξ; in statistical parlance, ξ are *nuisance parameters*. It is straightforward to infer both θ and ξ from the experimental data. All we need to do is to introduce a prior probability for the model parameters ξ and to invoke Bayes' theorem on the joint parameter space:

$$\Pr(\theta, \xi | D, I) = \frac{1}{\Pr(D|I)} \Pr(D|\theta, \xi, I) \Pr(\theta|I) \Pr(\xi|I).$$

where we assumed that θ and ξ are independent *a priori*: $\Pr(\theta, \xi|I) = \Pr(\theta|I)\Pr(\xi|I)$. It is straightforward to relax this assumption if necessary.

The posterior probability $\Pr(\theta, \xi|D, I)$ encodes all available information about the unknown parameters. In biomolecular structure determination, the posterior is typically too complex to do any further analytical calculations. By drawing Monte Carlo samples from $\Pr(\theta, \xi|D, I)$ we generate a finite approximation of the posterior (Liu, 2001). These samples can be used to compute expectations and variances over the unknown parameters and thereby estimate the parameters and compute error bars.

2.2. Probabilistic Models for Hybrid Data

Before we can launch an ISD calculation, we need to choose a likelihood $\Pr(D|\theta, \xi, I)$ and the priors $\Pr(\theta|I)$ and $\Pr(\xi|I)$. The application of ISD to multiple datasets D_i is straightforward: $\Pr(D|\theta, \xi, I) = \prod_i \Pr(D_i|\theta, \xi)$. Each dataset is described independently with an appropriate probabilistic model; all datasets are integrated by simply multiplying all factors representing the various datasets. Because probabilities for different datasets are calibrated (they all normalize to one), there is no issue of weighing the different datasets relative to each other.

We use a Boltzmann distribution as a prior over the conformational degrees of freedom:

$$\Pr(\theta|I) = \frac{1}{Z} \exp\{-E(\theta)\} \tag{1}$$

where $E(\theta)$ is a force field. ISD currently supports two force fields: a quartic repulsion term that lacks any attractive interaction, and a linearly ramped Lennard-Jones potential (see Habeck, 2011; Mechelke and Habeck, 2012 for more details). The prior distribution $\Pr(\theta|I)$ allows us to restrict the conformational degrees of freedom such that reasonable model structures are preferred (for example, structures that are free of atom-atom clashes and have well-packed interfaces). The prior distribution over the model parameters $\Pr(\xi|I)$ is typically of a standard form and chosen such that sampling with MCMC is straightforward.

2.2.1. Probabilistic Model for EM Maps

The result of a cryo-EM study is a 3D reconstruction of the structure, which typically comes in the form of a regular cubic grid with equal grid spacing in all three spatial directions. To construct a probabilistic model for 3D reconstructions, we first need a mathematical relation that allows us to compute a theoretical density map from a given structure θ. ISD's current model for density maps is quite simple. The theoretical map is obtained from an atomic model by placing spherical Gaussians of the same size and weight at each atom. The theoretical density at 3D position x is:

$$\rho(x; \theta, \sigma) = \sum_k \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\left\{-\frac{1}{2\sigma^2}\|x - x_k(\theta)\|^2\right\} \tag{2}$$

where the index k runs over all atoms that contribute to the density and $x_k(\theta)$ is the 3D position of the k-th atom in the structure parameterized by the conformational degrees of freedom θ. The theoretical density map can be interpreted as a blurred version of an atomic map with infinite resolution:

$$\rho(x; \theta, \sigma) = g_\sigma * \rho(x; \theta, 0) \quad \text{with} \quad \rho(x; \theta, 0) = \sum_k \delta[x - x_k(\theta)]$$

where δ is the Dirac delta function, g_σ is a Gaussian blur kernel with bandwidth σ and $*$ denotes a 3D convolution. Model (2) is admittedly simplistic and valid only for modeling protein complexes at intermediate to low resolutions. For high-resolution maps and/or the modeling of protein/nucleic acid complexes the model should also incorporate atom-wise weights (proportional to atom mass) as well as scattering and temperature factors.

Let us assume that experimental values ρ_n are available at positions x_n ($n = 1, \ldots, N$) which are typically the centers of voxels that make up a cubic grid. The discrepancy between the experimental map ρ_n and the theoretical map $\rho(x_n; \theta, \sigma)$ can be assessed with a Gaussian distribution. Alternative error models for density maps have been proposed (Vasishtan and Topf, 2011), but the Gaussian model is still the most widely used model.

The likelihood function resulting from a Gaussian model is:

$$\begin{aligned}\Pr(\rho|\theta, \xi, I) &= \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}\left[\rho_n - \alpha\rho(x_n; \theta, \sigma)\right]^2\right\} \\ &= \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_n \left[\rho_n - \alpha\rho(x_n; \theta, \sigma)\right]^2\right\}\end{aligned} \tag{3}$$

where the calibration factor α was introduced. There are three nuisance parameters $\xi = (\sigma, \alpha, \lambda)$. Typically, the bandwidth of the blur kernel σ is set to a constant value which depends on the resolution of the map. For example, the default value in Chimera (Pettersen et al., 2004) is $\sigma = 0.225 \times$ resolution. For this fixed choice of the bandwidth, σ can be absorbed into the background information I. However, it is also possible to estimate σ along with the other nuisance parameters and the conformational degrees of freedom.

To estimate the scaling parameter, we have to look at the conditional posterior distribution:

$$\Pr(\alpha \mid \lambda, \theta, D, I) \propto \Pr(\alpha|I) \times \exp\left\{-\frac{\lambda\|\rho(\theta,\sigma)\|^2}{2}\left(\alpha - \frac{\sum_n \rho_n \rho(x_n;\theta,\sigma)}{\|\rho(\theta,\sigma)\|^2}\right)^2\right\}$$

where $\|\rho\| = \sqrt{\sum_n \rho_n^2}$. The second factor is a Gaussian centered about the estimator:

$$\hat{\alpha}(\theta,\sigma) = \frac{\sum_n \rho_n \rho(x_n;\theta,\sigma)}{\|\rho(\theta,\sigma)\|^2} \tag{4}$$

which is the slope of a straight line relating the calculated volume $\rho(x_n;\theta,\sigma)$ to the observed density ρ_n.

The Gaussian model is directly related to the cross-correlation coefficient, which is often used to compare EM maps. To see this, let's integrate out the unknown scaling factor α. If we ignore the fact that α should be positive and choose a uniform (improper) prior over α (i.e., $\Pr(\alpha|I) = \text{const}$), we can analytically integrate out α to obtain a new likelihood that no longer depends on α (this procedure is also called marginalization in Bayesian statistics, Habeck et al., 2005a):

$$\Pr(\rho|\theta,\lambda,I) = \int d\alpha\ \Pr(\rho|\theta,\alpha,\lambda,I)\ \Pr(\alpha|I) \propto \lambda^{(N-1)/2} \exp\left\{-\frac{\lambda\,\|\rho\|^2}{2}[1 - C^2(\theta)]\right\} \tag{5}$$

where

$$C(\theta) = \frac{\sum_n \rho_n\, \rho(x_n;\theta,\sigma)}{\|\rho\|\ \|\rho(\theta,\sigma)\|}$$

is the cross-correlation between the experimental and the theoretical map. The effective likelihood function (Equation 5) attains its maximum when the cross-correlation coefficient is one. Whenever we assess the goodness of fit between the model and the experimental map by means of the cross-correlation coefficient, we implicitly assume that the error of the EM map follows a Gaussian distribution.

The parameter λ is the inverse variance of the Gaussian likelihood (Equation 3) and called the *precision* of the model (Bernardo and Smith, 2009). It is also possible to estimate the precision λ of the fit between the experimental and the theoretical density map. The parameter λ assesses how well the experimental and theoretical map agree on average. For large λ, the experimental map is very reliable and imposes a strong force on the model to adapt itself such that the calculated map reproduces the observed map as closely as possible. Assuming Jeffreys's prior for the precision, i.e., $\Pr(\lambda|I) = 1/\lambda$, the conditional posterior of the precision is a Gamma distribution (Habeck et al., 2006):

$$\Pr(\lambda|\theta,\alpha,\rho,I) \propto \lambda^{N/2-1} \exp\{-\lambda E_{\text{map}}(\theta,\alpha)\} \tag{6}$$

where the least-squares residual

$$E_{\text{map}}(\theta,\alpha) = \frac{1}{2}\sum_n [\,\rho_n - \alpha\rho(x_n;\theta,\sigma)\,]^2$$

is the restraint energy resulting from the Gaussian model of the experimental EM map. The expected value of the precision given the experimental map ρ and all unknown parameters is the inverse mean-squared error:

$$\hat{\lambda}(\theta,\alpha) \approx \frac{N}{2\,E_{\text{map}}(\theta,\alpha)}\,. \tag{7}$$

Estimator (Equation 7) tells us that the precision of the map increases when the fit between the observed map and the calculated map improves. This seems reasonable, but there is a problem.

Typically, EM maps are surrounded by bordering layers of low density voxels ($\rho_n \approx 0$). If we classify all voxels into N_1 voxels that contain density of the biomolecular assembly and N_0 voxels that carry only noise or zero density, we have $N = N_0 + N_1$. By increasing N_0 (e.g., by zero padding) the goodness of fit E_{map} does not change or changes only very little, such that we can artificially increase the apparent precision of the density map simply by increasing N_0:

$$\hat{\lambda}(\theta,\alpha) \approx \frac{N_0 + N_1}{2\,E_{\text{map}}(\theta,\alpha)} \geq \frac{N_1}{2\,E_{\text{map}}(\theta,\alpha)}\,.$$

To obtain a realistic estimate of λ, we should only fit those voxels that carry real density.

In principle, the task of classifying voxels into noise and non-noise voxels is an inference problem in itself: we would have to introduce a mask that tells us whether a voxel carries true signal or not. For the sake of simplicity we do not introduce an adaptive mask that we estimate along with with the model parameters, but restrict the fitting to voxels that are likely to carry the true signal. These voxels are identified in a couple of preparatory steps, which I will outline in the next section.

If we look at the conditional posterior of the conformational degrees of freedom θ, we find that:

$$\Pr(\theta|\xi,\rho,I) \propto \exp\{-E(\theta) - \lambda E_{\text{map}}(\theta,\alpha)\}\,. \tag{8}$$

By taking the negative logarithm of the posterior probability, we obtain a hybrid energy function (Jack and Levitt, 1978; Brünger and Nilges, 1993; Habeck et al., 2005a):

$$E_{\text{hybrid}}(\theta) = E(\theta) + \lambda\, E_{\text{map}}(\theta,\alpha)\,. \tag{9}$$

The precision acts as a weighting factor for the EM map (Habeck et al., 2006). If λ is too large, the forces from the EM term can bias the final structure (overfitting). Therefore, it is important to obtain a realistic estimate of λ.

2.2.2. Preparation of EM Maps

ISD carries out several preparatory steps before modeling with EM maps starts: thresholding, cropping, decimation, and masking. These steps improve the speed of fitting and are necessary to obtain a meaningful estimate of the precision of the density map.

Typically the user provides a threshold $\rho_{\min}$ above which the density shows the particle. ISD clips the density at $\rho_{\min}$, i.e., all values greater than the threshold are set to the threshold. After clipping, the density is shifted by subtracting the threshold such that the smallest experimental density is zero:

$$\rho_n \leftarrow \begin{cases} \rho_n - \rho_{\min} & ; \ \rho_n \geq \rho_{\min} \\ 0 & ; \ \rho_n < \rho_{\min} \end{cases} \tag{10}$$

After thresholding all $\rho_n \geq 0$. To reduce the map to those voxels that carry the real signal, a cropping operation is applied to reduce the 3D grid to a minimum size. Cropping removes bordering layers which only contain zero-density voxels analogous to an auto crop in image processing programs.

To represent the assumption that the structure is entirely covered by the thresholded density map, ISD introduces a box prior, which confines the system to lie inside the interior of a cubic box that coincides with the boundary of the 3D map. The box is parameterized by its lower left and upper right corner where the lower left corner is located at the origin of the 3D grid on which the thresholded EM map is evaluated. The box has a soft boundary which is implemented as a logistic function with finite steepness γ:

$$s_\gamma(x) = \frac{1}{1 + e^{-\gamma x}} \tag{11}$$

where typically $\gamma = 1\text{Å}^{-1}$. The complete prior over the conformational degrees of freedom is:

$$\Pr(\theta|I) \propto \exp\{-E(\theta)\} \prod_k \prod_{d=1}^{3} s_\gamma(x_{kd}(\theta) - l_d)\, s_\gamma(u_d - x_{kd}(\theta)) \tag{12}$$

where l_d, u_d are the spatial coordinates of the lower left / upper right corner of the bounding box of the EM map and $x_{kd}(\theta)$ are the spatial coordinates of the k-th atom.

The Gaussian likelihood (Equation 3) is only valid for voxels that carry signal. Let us introduce a binary mask $m_n \in \{0, 1\}$ which indicates for each voxel, if it carries signal ($m_n = 1$) or noise ($m_n = 0$). The modified Gaussian likelihood is:

$$\Pr(\rho|\theta, \xi, I) = \left(\frac{\lambda}{2\pi}\right)^{\sum_n m_n/2} \exp\left\{-\frac{\lambda}{2} \sum_n m_n [\, \rho_n - \alpha\, \rho(x_n; \theta, \sigma)\,]^2\right\}. \tag{13}$$

As mentioned above, the mask m_n should in principle be also considered an unknown parameter and therefore be estimated along with the other unknown quantities. However, this is currently not implemented in ISD and therefore m is part of the background information I.

Another parameter that we have to consider is the spacing of the EM map. The Gaussian likelihood assumes that the discrepancy between the experimental and calculated map is independent from voxel to voxel and shows no spatial correlations. However, this assumption is violated when the size of the voxels becomes too small. By resampling the experimental map on a finer grid, we could artificially increase the number of data points, which would result in an increase of the estimated weight λ. Therefore, EM maps are typically downsampled in ISD such that the spacing is roughly $2 \times \sigma$. A more rigorous treatment that accounts for spatial correlations between neighboring voxels is currently under development.

2.2.3. Conformational Degrees of Freedom

ISD supports multiple parameterizations for biomolecular systems. ISD typically decouples internal degrees of freedom from rigid external degrees of freedom, although modeling based on Cartesian coordinates is also supported. In case we want to model the internal flexibility of the subunits of a biomolecular assembly, ISD uses dihedral angles to parameterize the atom positions. The external degrees of freedom are three translational and three rotational degrees of freedom. To parameterize the rotation matrices, ISD uses a Lie group representation (Gallego and Yezzi, 2015). It is also possible to model symmetric assemblies by using virtual copies of the symmetry mates. ISD supports cyclic, dihedral and helical symmetry. The parameters of a helical symmetry can be estimated along with the conformational degrees of freedom.

To sample the conformational degrees of freedom θ, ISD uses the gradient of the log posterior probability (i.e., the gradient of the hybrid energy). Typically it is straightforward to compute the gradient with respect to the Cartesian coordinates. The Cartesian gradient is mapped onto the conformational degrees of freedom by virtue of the chain rule. This requires us to evaluate the Jacobian of the parameterization. In case of dihedral angles, there is an efficient recursive algorithm that avoids building up the full Jacobian matrix by traversing the tree of covalent bonds.

2.3. Markov Chain Monte Carlo for Biomolecular Modeling

The posterior probability $\Pr(\theta, \xi|D, I)$ encodes everything that can be said about the conformational degrees of freedom θ and the nuisance parameters ξ in the light of the experimental data D and our modeling assumptions I. Because $\Pr(\theta, \xi|D, I)$ is a high-dimensional probability distribution that is not suited for analytical computations, we explore $\Pr(\theta, \xi|D, I)$ by drawing random samples from it. Sampling from $\Pr(\theta, \xi|D, I)$ is based on Markov chain Monte Carlo (MCMC) (Liu, 2001). An MCMC algorithm simulates a Markov chain over (θ, ξ) space whose stationary distribution is the posterior $\Pr(\theta, \xi|D, I)$. After convergence of the Markov chain, the generated θ, ξ are valid samples from $\Pr(\theta, \xi|D, I)$. The samples can be used to compute expected values, variances and other statistics that characterize the posterior distribution. If we were to construct a multi-dimensional histogram from the θ, ξ samples, it would approximate the posterior distribution. The longer we run the Markov chain, the closer we get to the posterior distribution.

2.3.1. Gibbs Sampling

Gibbs sampling (Geman and Geman, 1984) is an iterative MCMC algorithm that decomposes sampling from $\Pr(\theta, \xi | D, I)$ into two successive steps, which are repeated:

$$\begin{aligned} \theta^{(t+1)} &\sim \Pr(\theta \mid \xi^{(t)}, D, I) \\ \xi^{(t+1)} &\sim \Pr(\xi \mid \theta^{(t+1)}, D, I) \end{aligned} \tag{14}$$

where t is an iteration index (pseudo time) and the superindex (t) marks samples generated in the t-th iteration; the notation $\sim$ means "sampled from." It can be shown that the Gibbs sampler (Equation 14) generates valid samples from the joint distribution $\Pr(\theta, \xi | D, I)$.

To implement a Gibbs sampler, we need to compute the conditional posterior distributions $\Pr(\theta \mid \xi, D, I)$ and $\Pr(\xi \mid \theta, D, I)$. The conditional posterior over the conformational degrees of freedom involves the hybrid energy (Equation 9):

$$\Pr(\theta \mid \xi, D, I) \propto \exp\left\{-\lambda E_{\mathrm{map}}(\theta, \alpha) - E(\theta)\right\}. \tag{15}$$

Sampling of the nuisance parameters is most easily done by applying a Gibbs sampling strategy to $\Pr(\xi \mid \theta, D, I)$ itself. We break down the second step in scheme (14) into the generation of α and λ samples according to:

$$\begin{aligned} \alpha^{(t+1)} &\sim \Pr(\alpha \mid \lambda^{(t)}, \theta^{(t+1)}, D, I) \\ \lambda^{(t+1)} &\sim \Pr(\lambda \mid \alpha^{(t+1)}, \theta^{(t+1)}, D, I) \end{aligned} \tag{16}$$

The conditional posteriors for the individual nuisance parameters, e.g., $\Pr(\lambda \mid \alpha, \theta, D, I)$, have been discussed in the previous section. Often these distributions are of a standard form and can be sampled directly using random number generators. For example, the conditional posterior of the precision λ is a Gamma distribution (Equation 6). Efficient algorithms for generating variates from a Gamma distribution exist (Devroye, 1986).

2.3.2. Hamiltonian Monte Carlo

Sampling the conformational degrees of freedom θ from the conditional posterior (Equation 9) is the most challenging step in an ISD calculation. Typically, the conformational degrees of freedom are highly coupled, and $\Pr(\theta|\xi, D, I)$ exhibits multiple peaks. A powerful variant of Metropolis Monte Carlo (Metropolis et al., 1957) is the Hybrid Monte Carlo method, also known as Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2010). The improvement over the simple Metropolis sampler is achieved by using a more efficient proposal step. In the standard version of Metroplis Monte Carlo, new candidate structures are proposed by randomly perturbing a conformational degree of freedom. The perturbation is either accepted or rejected depending on whether it produced an acceptable change in the hybrid energy or not. This kind of proposal results in a random walk in conformational space, which explores the space very inefficiently, because typically we can only apply small perturbations to the structure without increasing the hybrid energy by an unacceptable amount.

HMC proposes the candidate structure by running a short molecular dynamics trajectory where the hybrid energy plays the role of a force field. This has the advantage that the moves in structure space are adapted to the shape of the posterior distribution and that the conformational degrees of freedom change conjointly rather than one by one. HMC is several orders of magnitude more efficient than random walk Metropolis Monte Carlo, but comes at an additional computational cost. To run the proposal trajectory, one needs to calculate the gradient of the hybrid energy with respect to the conformational degrees of freedom. Since ISD uses non-Cartesian parameterizations, the gradient can be quite involved. Thanks to the chain rule we can break the computation of the gradient into two steps: First, the Cartesian gradient is calculated. In a second step, the Cartesian gradient is projected into the space of the conformational degrees of freedom. ISD implements this projection for dihedral angles and the rotational degrees of freedom of a rigid-body transformation.

2.3.3. Replica-Exchange Simulation

The posterior distribution arising in an application of ISD, is quite complex and typically shows multiple modes. As we will see in Section 3.3, the posterior distribution encountered in integrative modeling with cryo-EM data is often sharply peaked and exhibits isolated peaks. It is highly challenging to draw conformational samples from such a posterior distribution. ISD uses replica-exchange simulations (also known as parallel tempering) (Swendsen and Wang, 1986; Geyer, 1991) to address the sampling problem.

There are two factors that contribute to the posterior, the prior and the likelihood, and both are difficult to simulate in their own right. Therefore, ISD controls the complexity of each factor independently by introducing two "temperatures" (Habeck et al., 2005b). The first parameter, the inverse temperature $\beta \in [0, 1]$, scales the likelihood:

$$\left[\Pr(D|\theta, \xi, I)\right]^{\beta};$$

for $\beta = 1$ we obviously recover the original likelihood, for $\beta = 0$ we completely switch off the data.

The second parameter controls the shape of the conformational prior. Because the non-bonded interactions $E(\theta)$ span many orders of magnitude, it is highly inefficient to work with the standard Boltzmann ensemble which scales down the non-bonded energy when the temperature is increased. Instead of the Boltzmann ensemble, ISD uses the Tsallis ensemble to smooth out non-bonded interaction (Habeck et al., 2005b) and simulates:

$$\left[1 + (q-1)(E(\theta) - E_{\mathrm{min}})\right]^{-q/(q-1)}$$

where $q \geq 1$ is the so-called Tsallis q and E_{min} has to be chosen such that $E(\theta) > E_{\mathrm{min}}$ for all structures. For $q = 1$, we recover the standard Boltzmann prior (Equation 1).

The choice of the tempering schedule (i.e., the sequence of β and q) is difficult and crucial. We have to trade-off efficiency vs.

ergodicity of sampling. With increasing number of temperatures, the overlap between the replicas increases which results in an elevated swapping rate. But with increasing number of replicas the time for round trips increases quadratically, because states diffuse across different temperatures (i.e., there is no directed exchange of states that would aim for rapid mixing of states across different temperatures) (Earl and Deem, 2005). Therefore, we would rather choose a minimal number of replicas such that the smallest swapping rate is maintained.

3. RESULTS

In this section, I will illustrate Bayesian integrative modeling with hybrid data focusing on EM maps.

3.1. Flexible Fitting with Hamiltonian Monte Carlo

ISD can fit known structures and structural models into EM maps. In flexible fitting, we are trying to change the internal structure of a biomolecule so as to better fit an experimental EM map. A couple of software packages for flexible fitting has been published. Normal mode and elastic network methods (Delarue and Dumas, 2004; Tama et al., 2004; Hinsen et al., 2005; Schröder et al., 2007; Jolley et al., 2008; Tan et al., 2008) boost transitions along the principal directions of structural change. Molecular dynamics (MD) based methods (Orzechowski and Tama, 2008; Trabuco et al., 2008) combine a density fitting score with a full-fledged force field. Real-space refinement in Cartesian and internal coordinates, originally developed for X-ray crystallographic data, has been adapted to cryo-EM maps (Fabiola and Chapman, 2005). Rigid-body modeling with Flex-EM (Topf et al., 2008) freezes secondary structure elements and keeps just the linker regions flexible. Fragment-based structure prediction methods such as Rosetta has been combined with density map refinement (DiMaio et al., 2009).

ISD uses dihedral angles to parameterize the structures of the subunits of a macromolecular complex. In addition to the dihedral angles, each subunit has six external degrees of freedom that describe a rigid transformation of the subunit (three translational and three rotational degrees of freedom). The complete list of dihedral angles as well as the translational and rotational degrees of freedom from all subunits makes up the conformational degrees of freedom θ.

To study flexible fitting with ISD, let us first look at a specific example. Adenylate kinase (AK) is a widely used test system to predict and simulate conformational changes in proteins (see e.g., Orzechowski and Tama, 2008; Beckstein et al., 2009; Whitford et al., 2009). AK adopts two conformational states: an open state in which no ligands are bound and a closed state. The overall difference between both states is an RMSD of $\sim$ 7 Å. The conformational change can be understood as a rigid-body movement of three domains relative to each other: CORE, LID, and NMP-bind. During the conformational change, these three domains maintain their internal structure (Müller et al., 1996; Whitford et al., 2009).

I ran local posterior sampling with HMC starting from the open state (PDB code 4ake) and fitted it into a simulated EM map of the closed state (PDB code 1ake) at 10 Å resolution. **Figure 1A** shows the evolution of the RMSD to the initial and target structures during flexible fitting. The simulation starts at an RMSD of about 7 Å and rapidly improves it by optimizing the agreement with the experimental and theoretical maps. This is reflected by the evolution of the cross-correlation coefficient (see **Figure 1B**), which increases as the RMSD to the target structure decreases. After less than 200 steps of HMC sampling the fitted structure has an RMSD $<$ 1 Å to the target structure and a cross-correlation of almost 100%. During flexible fitting, the structure of the three domains remains intact. This is reflected by the fact that the RMSD restricted to those Cα atoms that belong to the same domain changes only little compared to the change in the overall RMSD (see **Figure 1C**). Thus, the HMC sampler preserves the integrity of the input structure and introduces larger scale changes only in a few hinge regions.

3.2. Flexible Fitting Benchmark

To systematically validate local flexible fitting of EM maps with ISD, I applied HMC sampling of the posterior distribution to a benchmark proposed by Topf et al. (2008) to test their Flex-EM method. The Flex-EM benchmark comprises various medium sized proteins and simulated EM maps at different resolutions ranging from 4 to 14 Å. For each flexible fitting task of the single-domain subset, I launched an HMC sampler starting from the initial structure as provided by the benchmark. The initial structure was obtained by homology modeling based on a template structure that shows an alternative conformational state. The task is to deform the homology model such that it better agrees with a simulated EM map showing a different conformational state.

Figure 2 shows the results of a flexible fitting benchmark from Topf et al. (2008). In all cases, ISD improves the fit of the initial structure quite significantly and achieves cross-correlation coefficients above 95%. Moreover, the RMSDs of the final structures fitted with ISD are systematically better than the fits obtained with Flex-EM.

Although flexible fitting with HMC performs well in practice, there are still conceptual problems with this approach. Sampling with HMC does not explore the full posterior distribution, but stays in the vicinity of the initial structure. A truly Bayesian approach, however, aims to explore the entire posterior distribution by using, for example, a full-blown replica simulation. However, global sampling of the posterior will result in many alternative fits of the EM map that will show a large RMSD to the target structure, because the force fields implemented in ISD cannot distinguish between the target structure and other globular structures that fit the density map. A remedy is to not only use the known structure that is fitted against the EM map as the initial structure, but also to develop a probabilistic model that allows for deformations of the known structure. Such a model is currently under development.

3.3. Global Fitting of Symmetric Assemblies

Global sampling of the posterior distribution is currently only possible in ISD, if the internal structure of the subunits is kept

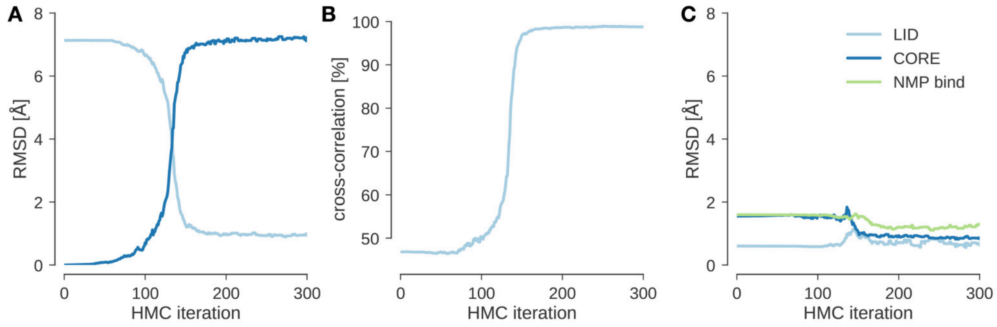


FIGURE 1 | Flexible fitting of adenylate kinase into a 10 Å map. (A) Evolution of the RMSD to the initial structure (4ake) shown in dark blue and the target structure (1ake) shown in light blue. **(B)** Evolution of the cross-correlation coefficient during flexible fitting. **(C)** RMSD reduced to Cα atoms that are part of the same rigid domain.

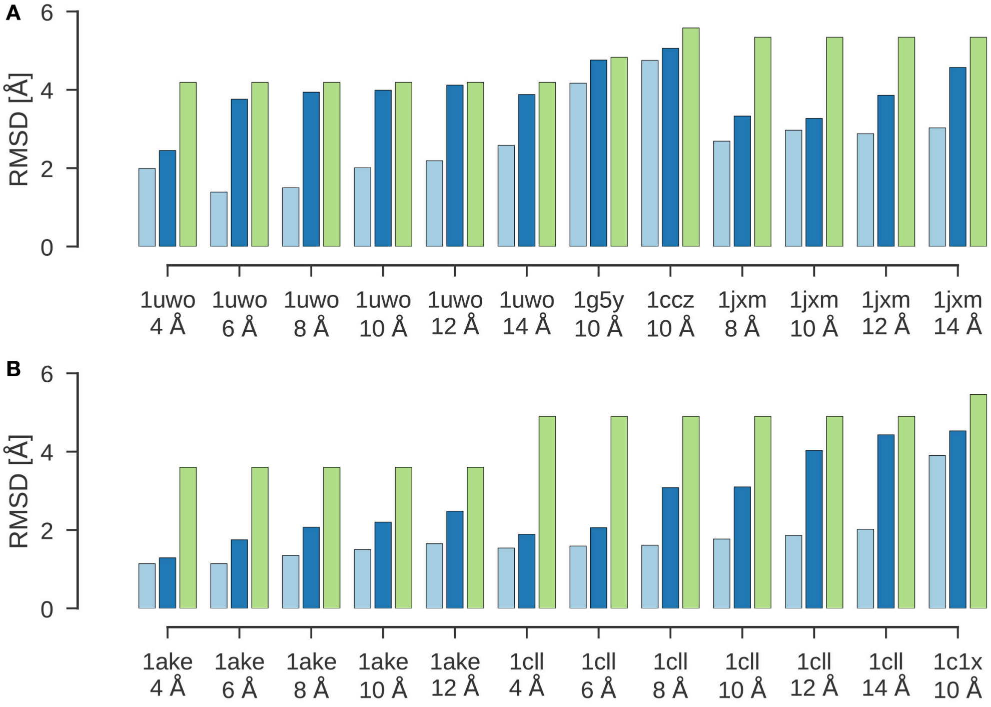


FIGURE 2 | Flexible fitting benchmark. Shown are the RMSD values for the final results of flexible fitting with ISD (light blue) and Flex-EM (dark blue) in comparison to the RMSD of the initial structure to the target structure (green). **(A)** Flexible fitting results for 1uwo, 1g5y, 1ccz, 1jxm. **(B)** Flexible fitting results for 1ake, 1cll, 1c1x.

fixed. The only degrees of freedom are the six external degrees of freedom parameterizing a global rotation and translation of each subunit. The sampling problem arising in global fitting of EM maps is quite severe. To see this, let us first study sampling from the prior (Equation 12), which is the Boltzmann ensemble confined by a soft box containing the experimental density map. Sampling from this prior is a sort of toy version of the density fitting problem. Instead of fitting the assembly against the density map, our aim is to generate non-clashing configurations that lie inside a box which contains the thresholded map. This is an instance of a 3D packing problem, which is NP-hard.

Let us look at a specific example: The symmetric chaperonin GroEL has been studied extensively by cryo-EM, X-ray crystallography and NMR. A 3D reconstruction of GroEL at a resolution of 4.1 Å is available (EMD-6422). The original map spans 240^3 voxels. The EMDB entry suggests a user-defined threshold of $\rho_{\min} = 3.5$ for visualizing the map. After thresholding (Equation 10) and cropping, the grid has $135 \times 133 \times 133$ voxels, i.e., only $\sim$ 17% of the original volume carries information that is useful for structural modeling. The 3D cropping operation results in a box that spans a volume of $144.5 \times 142.3 \times 142.3\,\text{Å}^3$. This example illustrates that thresholding and cropping can achieve a drastic reduction in the

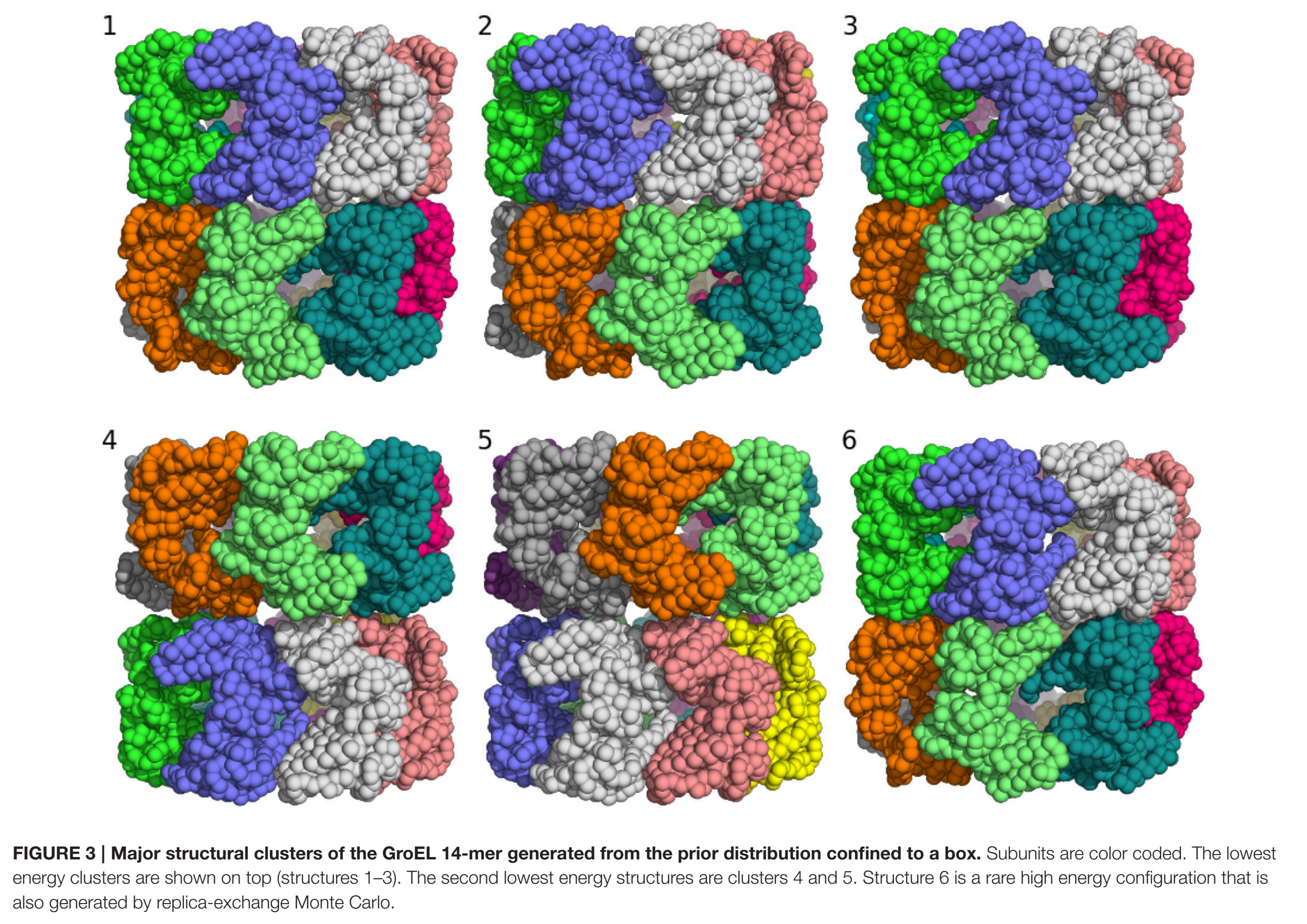


FIGURE 3 | Major structural clusters of the GroEL 14-mer generated from the prior distribution confined to a box. Subunits are color coded. The lowest energy clusters are shown on top (structures 1–3). The second lowest energy structures are clusters 4 and 5. Structure 6 is a rare high energy configuration that is also generated by replica-exchange Monte Carlo.

number of grid points that have to be evaluated during density fitting.

GroEL exhibits a seven-fold tetrahedral symmetry (D7). Therefore, our task is to sample configurations of the 14-mer that fit inside the box and minimize the overlap between atoms from different subunits. I used a Tsallis replica simulation to sample structures of the GroEL 14-mer. There are only six conformational degrees of freedom: three rotational and three translational degrees of freedom, which determine the position and orientation of a single GroEL subunit. The positions and orientations of the other 13 subunits are generated by the action of the D7 symmetry operator.

Although this is a low-dimensional sampling problem, it turns out to be surprisingly hard. I needed 59 replicas in the Tsallis ensemble to achieve an average swap rate of 38%. If the non-bonded interactions are fully switched on, there are only few arrangements that fit into the box without producing significant clashes between atoms from different subunits. As a consequence, the box prior exhibits a few isolated peaks. The shape of the prior distribution is reminiscent of a golf-course energy landscape and quite different from the funnel-shaped energy landscape imposed by distance restraints.

Clustering of the sampled rigid-body degrees of freedom yields six groups of symmetric assemblies that fit into the box (see **Figure 3** and **Table 1**). Each group is defined very precisely with an ensemble RMSD ranging between 0.13 and 0.23 Å over the entire 14-mer. The tightness of the clusters shows that there is only a discrete set of arrangements that fits into the box. The first three clusters achieve the lowest non-bonded energies $E(\theta)$. The energy of the next two clusters is elevated by 70 units. Replica-exchange Monte Carlo occasionally also samples a high-energy structure (cluster 6). The first five clusters show the same arrangement of the seven-membered ring formed by chains A–G. The RMSD of these chains to the arrangement in the crystal structure is below 0.8 Å; only the last cluster shows a higher RMSD of 4.7 Å. The major difference between the clusters is in how the rings are arranged relative to each other. In clusters 1, 2, 3, and 6, the two rings are oriented in the same fashion as in the crystal structure (with the termini facing each other), whereas clusters 4 and 5 show an inverted orientation.

Posteriors based on distance data such as those arising in NMR applications exhibit a continuum of high-probability structures. The Markov chain is guided to the most likely structures by a funnel-shaped probability landscape. The distributions arising in EM fitting problems show a very different landscape with multiple isolated peaks that carry similar probability mass and therefore all contribute significantly to the posterior. Rigid-body modeling with EM maps can be viewed as

TABLE 1 | Summary of a clustering analysis of the prior ensemble of GroEL.

Cluster	av. energy	Population [%]	Ensemble RMSD	RMSD (7-mer) [Å]	RMSD (14-mer) [Å]
1	228.8	22.8	0.2	0.8	7.8
2	234.0	23.1	0.2	0.7	9.0
3	234.1	23.1	0.1	0.7	13.4
4	301.7	19.3	0.2	0.8	71.5
5	301.7	11.5	0.2	0.8	80.2
6	995.5	0.2	0.1	4.7	8.6

Six major clusters have been identified. Listed are their average non-bonded energy, the RMSD to the average structure within each cluster (precision) and the RMSD (accuracy) to the crystal structure (PDB code 1oel) for a single ring (chains A–G) and the entire 14-mer (chains A–N).

a 3D packing problem. In case of GroEL, the packing constraint from the prior box and the D7 symmetry already determine the overall structure of the assembly to a large degree without any use of the density map. But the tests also show that even sampling from the prior alone can be quite challenging.

The minimum energy assembly sampled from the prior fits the density map only poorly with a cross-correlation of $\sim$ 10%. Refining the assembly in the presence of the map improves the cross-correlation to 55% and decreases the RMSD of the entire 14-mer to 1.1 Å.

3.4. Multi-Body Modeling of GroEL/ES

In general rigid-body modeling applications, we have to fit multiple rigid bodies into an EM map. I will use the GroEL/ES complex to illustrate multi-body fitting with ISD. GroEL/ES is formed by GroEL and the cochaperonin GroES. GroES interacts with one of the seven-membered rings formed by GroEL after a conformational change has been induced in the subunits. Therefore, the structures of the two GroEL 7-mers are no longer identical, and we have to fit three rigid bodies: one subunit of free GroEL (PDB code 1aon, chain A), one subunit of GroEL in complex with GroES (1aon, chain H), and one subunit of GroES (1aon, chain O). Each of the three subunits is duplicated by the action of a 7-fold cyclic symmetry. The symmetry mates are not represented explicitly, but generated from each of the three rigid bodies. Forces that act on the symmetry mates are backprojected onto the subunit. Therefore, we have a total of 18 conformational degrees of freedom.

I used ISD to fit GroEL/ES into a 23.5 Å map (Ranson et al., 2001) (EMD-1046). To shortcut the convergence of posterior sampling, I first ran a replica simulation with a Cα representation of the subunits and switched off the non-bonded interactions. With this strategy, the sampler rapidly generates models that achieve a cross-correlation of 96% (see **Figure 4D**). Inspection of the structures shows that there are two clusters which differ only in the structure of the GroES subunit. The structure of the two GroEL rings is already very close to the crystal structure (1aon) with an RMSD of 3.5 ± 0.5 Å over the 14-mer formed by the GroEL subunits (**Figure 4A**). The GroES 7-mer arranges in two versions of the ring: One is the correct structure with an RMSD of 2.1 ± 0.6 Å to the crystal structure. The second structure is incorrect with an RMSD of 20.0 ± 0.3 Å. Both structures are almost equally populated. The correct structure is adopted by 51.3% of the structures; the population of the incorrect assembly is 47.7% (see **Figure 4B**). There is a tiny fraction with a population of $\sim$1% that shows a third arrangement of the GroES subunit (RMSD 9.17 ± 0.51 Å). **Figure 4C** shows the distribution of the RMSD over the entire assembly.

In a refinement step, I used a full-atom representation of the subunits and switched on the non-bonded energy terms. The RMSD to the crystal structure drops to 1.4 Å without compromising the fit to the EM map: the cross-correlation coefficient of the full-atom structure is still 96%.

3.5. Estimation of the Precision of an EM Map

As outlined in Section 2.2.1, it is challenging to obtain a good estimate of the precision of an EM map, because an EM map typically contains many zero-density voxels in addition to the non-noise voxels, but only voxels carrying a real signal should contribute to the precision. To identify which voxels carry true signal, we would have to first solve the fitting problem. Therefore, both problems, the estimation of a well-fitting structure and the construction of a good mask, are highly related. Moreover, the errors (i.e., the discrepancy between the experimental and calculated maps) are spatially correlated, but the Gaussian model (3) treats them as completely independent observations, which also results in an artificial increase in the precision. The reason for the latter effect is the following: If errors are correlated, the effective number of data points is smaller than the number of voxels (Sivia, 2004). According to Equation (7) the precision of the map is proportional to the number of voxels for the simple Gaussian model, the precision will therefore be overestimated, if the errors between neighboring voxels are correlated.

Let us illustrate the various factors that influence the precision for a concrete example. **Figure 5** shows the distribution of the discrepancy between the experimental and the calculated density map for the GroEL/ES map analyzed in the previous section. The Gaussian likelihood assumes that this distribution has a bell-shaped curve whose width is determined by the precision λ. The distribution of the discrepancy $\epsilon_n = \rho_n - \rho(x_n; \theta, \sigma)$ is shown in (**Figures 5A–D**) for various stages of preprocessing. The original map contains many low-density voxels that lead to a very sharp, dominating peak at zero in the distribution of ϵ_n (**Figure 5A**). Cropping (**Figure 5B**) and subsequent decimation (**Figure 5C**) chops away many of the zero-density voxels and decreases the detrimental effect of the low-density voxels. However, the distribution of ϵ_n is only captured well by a Gaussian, if we mask out low-density voxels (see **Figure 5D**). The effect of the preprocessing steps on the estimated precision is shown in **Figure 5E**. Each of the preparation steps lowers the estimated precision by orders of magnitude.

4. CONCLUSION

This article discusses how ISD incorporates EM maps into a structure calculation and demonstrates some aspects of Bayesian integrative modeling with EM data. The Bayesian framework is

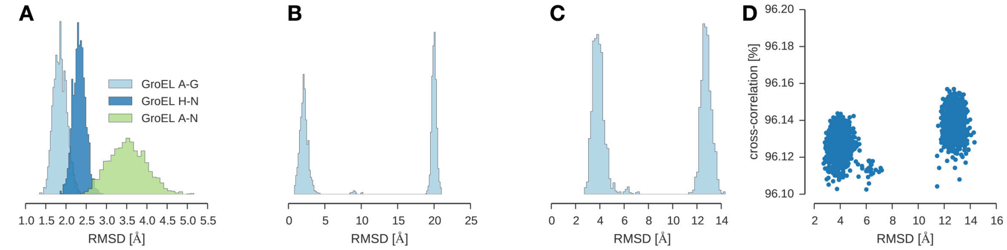


FIGURE 4 | Multi-body modeling of GroEL/ES. Shown is the RMSD between structural models obtained by posterior sampling with ISD and the crystal structure (PDB code 1aon). **(A)** RMSD for GroEL subunits for both 7-membered rings (chains A–G and chains H–N) and for the entire 14-mer (chains A–N). **(B)** RMSD for GroES (chains O–U) **(C)** RMSD for the entire 21-mer. **(D)** Correlation between the overall RMSD (21-mer) and the cross-correlation coefficient.

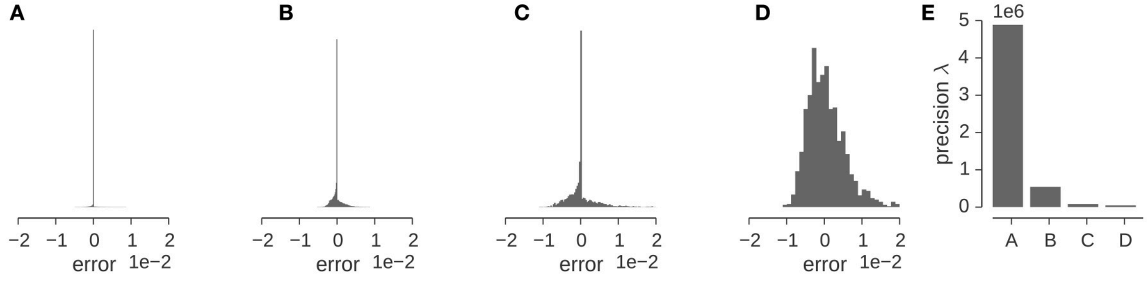


FIGURE 5 | Estimation of the precision λ of the GroEL/ES map. (A–D) Show the distribution of the "error" (or discrepancy) between the experimental and calculated maps $\rho_n - \rho(x_n; \theta, \sigma)$. Error distribution for the full map **(A)**, full map after cropping **(B)**, the downsampled and cropped map **(C)**, the downsampled, cropped and masked map **(D)**. **(E)** Estimated precision for the different input maps used in multi-body fitting.

highly suited to address issues in structural modeling with hybrid data such as how to weigh multiple datasets relative to each other. The major bottleneck of an inferential structure determination is conformational sampling. The posterior distribution arising in EM fitting poses a challenging sampling problem, which can be overcome with replica-exchange Monte Carlo.

The article does not cover crosslinking/mass spectrometry and solid-state NMR, which are complementary methods for characterizing the structure of large assemblies. ISD has also been used to model biomolecular assemblies from solid-state NMR data. For example, we have used ISD to compute the structure of the membrane domain of the trimeric autotransporter adhesin YadA (Shahid et al., 2012). We modeled a fully flexible subunit in the presence of a cyclic trimer symmetry. Although the data are highly ambiguous due to the imprecision of solid-state NMR restraints and the trimer symmetry, ISD was able to determine the correct structure of the YadA membrane anchor domain. Another example is our recent structure of a type 1 pilus FimA from *E. coli* (Habenstein et al., 2015). Here solid-state NMR and scanning electron microscopy data were combined with solution NMR data to estimate the internal structure of the subunit as well as the parameters of the helical symmetry of the FimA pilus. Also modeling with crosslinking data is possible with ISD, e.g., Carstens et al. (2016) discuss chromosome structure modeling. However, the use of crosslinking data for modeling macromolecular complexes still needs to be benchmarked thoroughly. A common scenario is to combine cryo-EM with crosslinking data, which also needs to be tested systematically with ISD. A Bayesian approach to modeling macromolecular assemblies with crosslinking data has been proposed recently by Ferber et al. (2016).

Future work will focus on various aspects of modeling with hybrid data. One goal is to develop a better model for EM maps that incorporates the various preprocessing steps discussed in Section 2.2.2. The model will incorporate a mask that will be estimated along with the other unknown parameters. Moreover, we will develop a likelihood function that accounts for spatial correlations between errors in the density map. Another goal is to support modeling with coarse-grained representations of biomolecular systems (Tozzini, 2005; Saunders and Voth, 2013). Especially, for very large systems it will be critical to work with a multiscale representation to enable exhaustive conformational sampling. We are already using highly coarse-grained models for modeling the 3D structure of chromosomes and genomes from chromosome conformation capture data (Carstens et al., 2016).

AUTHOR CONTRIBUTIONS

MH designed and performed research and wrote the manuscript.

FUNDING

The author acknowledges funding from the German Research Foundation (DFG) (SFB 860, Project B09).

REFERENCES

Agafonov, D. E., Kastner, B., Dybkov, O., Hofele, R. V., Liu, W. T., Urlaub, H., et al. (2016). Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science* 351, 1416–1420. doi: 10.1126/science.aad2085

Anger, A. M., Armache, J. P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D. N., et al. (2013). Structures of the human and Drosophila 80S ribosome. *Nature* 497, 80–85. doi: 10.1038/nature12104

Bai, X. C., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., et al. (2015). An atomic structure of human γ-secretase. *Nature* 525, 212–217. doi: 10.1038/nature14892

Bayrhuber, M., Meins, T., Habeck, M., Becker, S., Giller, K., Villinger, S., et al. (2008). Structure of the human voltage-dependent anion channel. *Proc. Natl. Acad. Sci. U.S.A.* 105, 15370–15375. doi: 10.1073/pnas.0808115105

Beckstein, O., Denning, E. J., Perilla, J. R., and Woolf, T. B. (2009). Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open closed transitions. *J. Mol. Biol.* 394, 160–176. doi: 10.1016/j.jmb.2009.09.009

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bernardo, J. M., and Smith, A. F. M. (2009). *Bayesian Theory*. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons.

Brünger, A. T. (1992). The free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355, 472–474.

Brünger, A. T., and Nilges, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. *Q. Rev. Biophys.* 26, 49–125. doi: 10.1017/S0033583500003966

Carstens, S., Nilges, M., and Habeck, M. (2016). Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS Comput. Biol.* 12:e1005292. doi: 10.1371/journal.pcbi.1005292

Chiu, W., Baker, M. L., Jiang, W., Dougherty, M., and Schmid, M. F. (2005). Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure* 13, 363–372. doi: 10.1016/j.str.2004.12.016

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *Am. J. Phys.* 14, 1–13. doi: 10.1119/1.1990764

Delarue, M., and Dumas, P. (2004). On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6957–6962. doi: 10.1073/pnas.0400301101

Devroye, L. (1986). *Non-uniform Random Variate Generation*. New York, NY: Springer Verlag.

DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., and Baker, D. (2009). Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol.* 392, 181–190. doi: 10.1016/j.jmb.2009.07.008

Duane, S., Kennedy, A. D., Pendleton, B., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* 195, 216–222. doi: 10.1016/0370-2693(87)91197-X

Earl, D. J., and Deem, M. W. (2005). Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7, 3910–3916. doi: 10.1039/b509983h

Esquivel-Rodríguez, J., and Kihara, D. (2013). Computational methods for constructing protein structure models from 3D electron microscopy maps. *J. Struct. Biol.* 184, 93–102. doi: 10.1016/j.jsb.2013.06.008

Fabiola, F., and Chapman, M. S. (2005). Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure* 13, 389–400. doi: 10.1016/j.str.2005.01.007

Ferber, M., Kosinski, J., Ori, A., Rashid, U. J., Moreno-Morcillo, M., Simon, B., et al. (2016). Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nat. Methods* 13, 515–520. doi: 10.1038/nmeth.3838

Fischer, N., Neumann, P., Konevega, A. L., Bock, L. V., Ficner, R., Rodnina, M. V., et al. (2015). Structure of the *E. coli* ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* 520, 567–570. doi: 10.1038/nature14275

Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* 31, 303–319. doi: 10.1146/annurev.biophys.31.082901.134202

Galej, W. P., Wilkinson, M. E., Fica, S. M., Oubridge, C., Newman, A. J., and Nagai, K. (2016). Cryo-EM structure of the spliceosome immediately after branching. *Nature* 537, 197–201. doi: 10.1038/nature19316

Gallego, G., and Yezzi, A. (2015). "A compact formula for the derivative of a 3-d rotation in exponential coordinates". *J. Math. Imaging Vis.* 51, 378–384. doi: 10.1007/s10851-014-0528-x

Geman, S., and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI* 6, 721–741. doi: 10.1109/TPAMI.1984.4767596

Gerstein, M., Lesk, A. M., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry* 33, 6739–6749. doi: 10.1021/bi00188a001

Geyer, C. J. (1991). "Markov chain Monte Carlo maximum likelihood," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (Fairfax Station, VA: Interface Foundation of North America), 156–163.

Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 8, 645–654. doi: 10.1038/nrm2208

Habeck, M. (2011). Statistical mechanics analysis of sparse data. *J. Struct. Biol.* 173, 541–548. doi: 10.1016/j.jsb.2010.09.016

Habeck, M. (2012). "Inferential structure determination from nmr data," in *Bayesian Methods in Structural Bioinformatics* eds T. Hamelryck, K. Mardia, and J. Ferkinghoff-Borg (Berlin; Heidelberg: Springer), 287–311. doi: 10.1007/978-3-642-27225-7_12

Habeck, M., Nilges, M., and Rieping, W. (2005a). Bayesian inference applied to macromolecular structure determination. *Phys. Rev. E* 72:031912. doi: 10.1103/PhysRevE.72.031912

Habeck, M., Nilges, M., and Rieping, W. (2005b). Replica-exchange Monte Carlo scheme for Bayesian data analysis. *Phys. Rev. Lett.* 94, 0181051–0181054. doi: 10.1103/PhysRevLett.94.018105

Habeck, M., Rieping, W., and Nilges, M. (2006). Weighting of experimental evidence in macromolecular structure determination. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1756–1761. doi: 10.1073/pnas.0506412103

Habenstein, B., Loquet, A., Hwang, S., Giller, K., Vasa, S. K., Becker, S., et al. (2015). Hybrid structure of the type 1 pilus of uropathogenic *Escherichia coli*. *Angew. Chem. Int. Ed. Engl.* 54, 11691–11695. doi: 10.1002/anie.201505065

Hinsen, K., Reuter, N., Navaza, J., Stokes, D. L., and Lacapère, J. J. (2005). Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys. J.* 88, 818–827. doi: 10.1529/biophysj.104.050716

Jack, A., and Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Cryst. Sect. A* 34, 931–935. doi: 10.1107/S0567739478001904

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge: Cambridge University Press.

Jolley, C. C., Wells, S. A., Fromme, P., and Thorpe, M. F. (2008). Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophys. J.* 94, 1613–1621. doi: 10.1529/biophysj.107.115949

Karaca, E., and Bonvin, A. M. (2013). Advances in integrative modeling of biomolecular complexes. *Methods* 59, 372–381. doi: 10.1016/j.ymeth.2012.12.004

Khatter, H., Myasnikov, A. G., Natchiar, S. K., and Klaholz, B. P. (2015). Structure of the human 80S ribosome. *Nature* 520, 640–645. doi: 10.1038/nature14427

Knuth, K. H., Habeck, M., Malakar, N. K., Mubeen, A. M., and Placek, B. (2015). Bayesian evidence and model selection. *Digit. Signal Process.* 47, 50–67. doi: 10.1016/j.dsp.2015.06.012

Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., et al. (2011). EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39, D456–D464. doi: 10.1093/nar/gkq880

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York, NY: Springer.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge: Cambridge University Press.

Mechelke, M., and Habeck, M. (2012). Calibration of Boltzmann distribution priors in Bayesian data analysis. *Phys. Rev. E* 86:066705. doi: 10.1103/PhysRevE.86.066705

Mechelke, M., and Habeck, M. (2014). Bayesian weighting of statistical potentials in NMR structure calculation. *PLoS ONE* 9:e100197. doi: 10.1371/journal.pone.0100197

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A., and Teller, E. (1957). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Müller, C. W., Schlauderer, G. J., Reinstein, J., and Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* 4, 147–156. doi: 10.1016/S0969-2126(96)00018-4

Neal, R. M. (2010) "MCMC using hamiltonian dynamics," in *The Handbook of Markov Chain Monte Carlo* eds S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Chapman & Hall/CRC Press), 113–162.

Orlova, E. V., and Saibil, H. R. (2004). Structure determination of macromolecular assemblies by single-particle analysis of cryo-electron micrographs. *Curr. Opin. Struct. Biol.* 14, 584–590. doi: 10.1016/j.sbi.2004.08.004

Orzechowski, M., and Tama, F. (2008). Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.* 95, 5692–5705. doi: 10.1529/biophysj.108.139451

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., et al. (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature* 518, 376–380. doi: 10.1038/nature14229

Ranson, N. A., Farr, G. W., Roseman, A. M., Gowen, B., Fenton, W. A., Horwich, A. L., et al. (2001). ATP-bound states of GroEL captured by cryo-electron microscopy. *Cell* 107, 869–879. doi: 10.1016/S0092-8674(01)00617-1

Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* 173, 530–540. doi: 10.1016/j.jsb.2010.10.014

Rauhut, R., Fabrizio, P., Dybkov, O., Hartmuth, K., Pena, V., Chari, A., et al. (2016). Molecular architecture of the *Saccharomyces cerevisiae* activated spliceosome. *Science* 353, 1399–1405. doi: 10.1126/science.aag1906

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science* 309, 303–306. doi: 10.1126/science.1110428

Rieping, W., Nilges, M., and Habeck, M. (2008). ISD: a software package for Bayesian NMR structure calculation. *Bioinformatics* 24, 1104–1105. doi: 10.1093/bioinformatics/btn062

Robinson, C. V., Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. *Nature* 450, 973–982. doi: 10.1038/nature06523

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., et al. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 10:e1001244. doi: 10.1371/journal.pbio.1001244

Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. *Nature* 422, 216–225. doi: 10.1038/nature01513

Saunders, M. G., and Voth, G. A. (2013). Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* 42, 73–93. doi: 10.1146/annurev-biophys-083012-130348

Schröder, G. F. (2015). Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr. Opin. Struct. Biol.* 31, 20–27. doi: 10.1016/j.sbi.2015.02.016

Schröder, G. F., Brunger, A. T., and Levitt, M. (2007). Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* 15, 1630–1641. doi: 10.1016/j.str.2007.09.021

Shahid, S. A., Bardiaux, B., Franks, W. T., Krabben, L., Habeck, M., van Rossum, B. J., et al. (2012). Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. *Nat. Methods* 9, 1212–1217. doi: 10.1038/nmeth.2248

Sivia, D. S. (2004). "Some thoughts on correlated noise," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23RD International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, vol. 707 (Melville, NY: AIP Publishing), 303–313. doi: 10.1063/1.1751374

Swendsen, R. H., and Wang, J.-S. (1986). Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* 57, 2607–2609. doi: 10.1103/PhysRevLett.57.2607

Tama, F., Miyashita, O., and Brooks, C. L. (2004). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.* 147, 315–326. doi: 10.1016/j.jsb.2004.03.002

Tan, R. K., Devkota, B., and Harvey, S. C. (2008). YUP.SCX: coaxing atomic models into medium resolution electron density maps. *J. Struct. Biol.* 163, 163–174. doi: 10.1016/j.jsb.2008.05.001

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure* 16, 295–307. doi: 10.1016/j.str.2007.11.016

Tozzini, V. (2005). Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* 15, 144–150. doi: 10.1016/j.sbi.2005.02.005

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16, 673–683. doi: 10.1016/j.str.2008.03.005

Vasishtan, D., and Topf, M. (2011). Scoring functions for cryoEM density fitting. *J. Struct. Biol.* 174, 333–343. doi: 10.1016/j.jsb.2011.01.012

Villa, E., and Lasker, K. (2014). Finding the right fit: chiseling structures out of cryo-electron microscopy maps. *Curr. Opin. Struct. Biol.* 25, 118–125. doi: 10.1016/j.sbi.2014.04.001

Wan, R., Yan, C., Bai, R., Huang, G., and Shi, Y. (2016). Structure of a yeast catalytic step I spliceosome at 3.4 resolution. *Science* 353, 895–904. doi: 10.1126/science.aag2235

Ward, A. B., Sali, A., and Wilson, I. A. (2013). Biochemistry. Integrative structural biology. *Science* 339, 913–915. doi: 10.1126/science.1228565

Whitford, P. C., Noel, J. K., Gosavi, S., Schug, A., Sanbonmatsu, K. Y., and Onuchic, J. N. (2009). An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins* 75, 430–441. doi: 10.1002/prot.22253

Yan, C., Hang, J., Wan, R., Huang, M., Wong, C. C., and Shi, Y. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* 349, 1182–1191. doi: 10.1126/science.aac7629

Yan, S., Suiter, C. L., Hou, G., Zhang, H., and Polenova, T. (2013). Probing structure and dynamics of protein assemblies by magic angle spinning NMR spectroscopy. *Acc. Chem. Res.* 46, 2047–2058. doi: 10.1126/science.aac7629

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 25 April 2017
doi: 10.3389/fmolb.2017.00023

Determining Complex Structures using Docking Method with Single Particle Scattering Data

Hongxiao Wang and Haiguang Liu *

Complex Systems Division, Beijing Computational Science Research Center, Beijing, China

Protein complexes are critical for many molecular functions. Due to intrinsic flexibility and dynamics of complexes, their structures are more difficult to determine using conventional experimental methods, in contrast to individual subunits. One of the major challenges is the crystallization of protein complexes. Using X-ray free electron lasers (XFELs), it is possible to collect scattering signals from non-crystalline protein complexes, but data interpretation is more difficult because of unknown orientations. Here, we propose a hybrid approach to determine protein complex structures by combining XFEL single particle scattering data with computational docking methods. Using simulations data, we demonstrate that a small set of single particle scattering data collected at random orientations can be used to distinguish the native complex structure from the decoys generated using docking algorithms. The results also indicate that a small set of single particle scattering data is superior to spherically averaged intensity profile in distinguishing complex structures. Given the fact that XFEL experimental data are difficult to acquire and at low abundance, this hybrid approach should find wide applications in data interpretations.

Keywords: hybrid method, single particle scattering, x-ray free electron laser, docking, molecular complex

OPEN ACCESS

Edited by:
Osamu Miyashita,
RIKEN Advanced Institute for Computational Science, Japan

Reviewed by:
Yasumasa Joti,
Japan Synchrotron Radiation Research Institute, Japan
Khanh Dao Duc,
University of Pennsylvania, USA
Robert Rambo,
Diamond Light Source, UK

****Correspondence:***
Haiguang Liu
hgliu@csrc.ac.cn

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *01 December 2016*
Accepted: *29 March 2017*
Published: *25 April 2017*

Citation:
Wang H and Liu H (2017) Determining Complex Structures using Docking Method with Single Particle Scattering Data. Front. Mol. Biosci. 4:23. doi: 10.3389/fmolb.2017.00023

INTRODUCTION

In crowded cellular environment, protein molecules often form complexes to fulfill their functions. Thus, the study of protein complex structures and dynamics is critical for the understanding of molecular mechanism (Eisenberg et al., 2000; Bader et al., 2003; Krissinel and Henrick, 2007). Because protein complexes are mostly stabilized by non-covalent interactions, their stability is under strong influence of solvent conditions, making it difficult to form molecular crystals that can yield strong diffraction signals. The nuclear magnetic resonance (NMR) spectroscopy has been widely applied to structure determination of relatively small molecular systems, but the degeneracy of NMR signals in large protein complexes challenges the model reconstructions (Bax and Grzesiek, 1993; Mainz et al., 2013; Göbl et al., 2014; Shen and Bax, 2015). Other experimental approaches that do not require crystallization include small angle X-ray scattering (SAXS) methods that obtain rotational averaged scattering intensity profile, from which structural information can be extracted to build low resolution 3D models (Konarev et al., 2006; Liu et al., 2012). Biochemistry techniques, such as cross-linking, mutagenesis, or single molecule fluorescence experiments can reveal critical interacting regions at complex interfaces, for example. The SAXS and biochemistry assay data bear a common problem: the information deficiency, compared to X-ray crystallography or NMR, does not allow a high resolution 3D structure determination. The data interpretation therefore heavily depends on computational modeling.

Recent advances in single particle imaging (SPI) methods using cryogenic electron microscopy (cryo-EM) or the emerging X-ray Free Electron Laser (XFEL) provide a new opportunity to study the molecular complex structure and dynamics (Emma et al., 2010; Chapman et al., 2011; Seibert et al., 2011; Cheng, 2015; Cheng et al., 2015; Schlichting, 2015). The cryo-EM single particle imaging technology has achieved significant breakthroughs, mostly thanks to the development of direct electron detecting device, model reconstruction algorithms, and sample handling, and automated data collection (Scheres, 2012; Cheng, 2015; Cheng et al., 2015). The resolution of 3D reconstruction models from cryo-EM data has been reported to atomic resolution, and the molecular size can be smaller than 100 kDa (Merk et al., 2016). The XFELs with their unprecedented peak brilliance realized a new experimental mode, "diffract before damage," to overcome the X-ray dosage limitations, making it possible to collect high resolution X-ray diffraction signals from non-crystal single molecule samples in principle (Neutze et al., 2000; Bogan et al., 2008; Seibert et al., 2011; Munke et al., 2016). Since the commissioning of the world's first hard XFEL facility, the Linac Coherent Light Source (LCLS), collective efforts have been made to push forward the application of XFEL in structure determination using single particle diffraction approach, and progress has been achieved toward high resolution structure determinations (Aquila et al., 2015; Munke et al., 2016). Nevertheless, both cryo-EM single particle imaging and XFEL single particle diffraction require tremendous amount of data measured at orientations that span SO(3) rotation space to assemble into a finely sampled 3D diffraction volume, from which 3D structures can be reconstructed. It is still a limiting step to obtain such experimental datasets, especially for XFEL single particle diffraction cases (Aquila et al., 2015). Experimental challenges include sample purification, injection, and alignment to the X-ray incidence beam etc., making the data collection very tedious and inefficient. Because of the low hit-rate (the chance for XFEL pulses hitting on individual clean sample particle) and the limited XFEL resources all over the world (only LCLS in SLAC national laboratory and the SACLA in RIKEN SPring-8 center are currently commissioned), collecting a full dataset which may include millions of single particle scattering patterns is still beyond present reach as routine experiments. Therefore, the data analysis methods in cryo-EM single particle imaging is not yet practical for XFEL single particle scattering data interpretation. The computational challenges for XFEL data analysis are summarized in a recent review (Liu and Spence, 2016).

Computational docking methods have been developed for protein complex structure prediction based on the structures of protein subunits. The Critical Assessment of PRedicted Interactions (CAPRI) contests have been organized and progress has been reported in the proceedings published after each evaluation (Janin, 2005; Lensink et al., 2016). One of the major challenges in protein complex structure prediction is to design reliable scoring functions for model quality assessment. The scoring functions for docking usually incorporate the following terms to rank the predicted models: the shape complementary between protein subunits, electrostatic interactions, solvation energy, and statistical potential energy derived from protein structure databases. Although encourage progress is obtained, a satisfactory scoring function is still needed (Gray et al., 2003; Vreven et al., 2015). The aforementioned XFEL single particle scattering data can be valuable in improving the ranking of protein complex structures generated using docking method, even for the cases that the dataset is not sufficient for high resolution structure determination. As a matter of fact, similar ideas have been implemented for SAXS data, which can be incorporated in model evaluation (Mattinen et al., 2002; Zheng and Doniach, 2002; Förster et al., 2008; Schneidman-Duhovny et al., 2012; Schindler et al., 2016). In this work, we extend this approach to XFEL single particle scattering data, inspired by the application of XFEL data in modeling of protein conformation changes (Tokuhisa et al., 2016). Using Zdock program(Chen et al., 2003), structure decoy sets are generated for several selected protein complexes, and the power of ranking using the original Zdock score, the SAXS score, and the single particle scattering score is studied. The simulation results suggest that the XFEL single particle data has the most information that best distinguish the correct models from the rest in the decoy sets. The problems in experimental data based model selection and the challenges in scoring function calculation are discussed.

METHODS

Single Particle Scattering Pattern Simulations

The scattering pattern simulation for a given protein structure is a forward problem, which is straightforward by using the Fourier transform of the electron density represented with atomic positions. In this work, the structural form factors $\boldsymbol{F}(\boldsymbol{q})$ is calculated using the direct summation of scattered wavefunctions, i.e.,

$$\boldsymbol{F}(\boldsymbol{q}) = \sum_j f_j(q) e^{i\boldsymbol{q} \cdot \boldsymbol{r}_j} \tag{1}$$

where the $\boldsymbol{q}$ is a vector in Fourier space, corresponding to the momentum transfer of the X-rays, defined as $\boldsymbol{q} = \boldsymbol{2\pi}(\boldsymbol{K_0} - \boldsymbol{K_i})$, $\boldsymbol{K_i}$ and $\boldsymbol{K_0}$ are the incidence and scattered wave vectors. $f_i(q)$ and $\boldsymbol{r}_j$ are the form factor and position of atom j. The atomic form factor depends on the magnitude of momentum transfer $\mathrm{q} = |\mathrm{q}|$; the values can be looked up in the International Table for Crystallogrophy. For a forward scattering experiment, the momentum transfer q can be calculated as

$$\mathbf{q} = \frac{4\,\boldsymbol{\pi}\,\mathbf{sin}\,\boldsymbol{\theta}}{\lambda} \tag{2}$$

and $\boldsymbol{2\theta}$ is the scattering angle that can be calculated based on the distance between sample and detector and the pixel location information, λ is the wavelength of X-rays. Based on the construction of Ewald sphere, for a given model at any specified orientation, the structure form factor $\boldsymbol{F}(\boldsymbol{q})$ at momentum transfer $\boldsymbol{q}$ that is mapped to the pixel position on 2D detector can be calculated using Equation (1). Then the squared modulus of the structure factors is taken for scattering intensity, i.e., $I(\mathbf{q}) = ||F(\mathbf{q})||^2$. For experimental data, Poisson noise was added to simulate the statistics error occurred during photon detection. On top of this, background noise was simulated by adding random photons following a Gaussian distribution at desired noise levels.

The key parameters for the pattern simulations can be found in Table S1 in the Supplementary Material. Experimental scattering intensity is proportional to the incidence beam intensity $\boldsymbol{I_0}$, which can be used to scale the intensity values recorded with detector. Therefore, $\boldsymbol{I_0}$ in this study has an immediate impact to the resolutions of scattering signals. In the simulations presented here, the incidence beam intensity was not explicitly considered. Instead, $\boldsymbol{I_0}$ was used as a scaling factor to set the highest measurable resolutions. In the simulations presented in this paper, we set the highest measurable resolution shell to be 4Å, where the average number of photons recorded at each pixel in this resolution is 1. This requires the photon flux is 1–2 order of magnitudes higher than the current XFELs, such as the LCLS, whose photon flux is about 10^{12} photons/pulse/μm^2.

The patterns for the native structures of the complexes are first simulated at random orientations in SO(3) rotation space (or a subspace) as the "experimental data"; then the patterns for the predicted models are generated with two orientation sampling approaches: (1) using the *same* orientations as the "experimental data" to study the ranking power of the scoring functions under ideal situations; and (2) using orientations specified by Euler angles spanning SO(3) rotation space. In the latter case, the orientations will be determined by computing the cross-correlation between "experimental patterns" and "model patterns," therefore the discretizing step size is important for finding the correctly matched orientations. All patterns are simulated to 4 Å resolution.

Protein Complex Generation Using Z-dock Program

The protein complex structures were generated using the Z-dock program developed by Weng's group in University of Massachusetts. Using Z-dock program, protein complexes were generated and 1,000 structures with high Z-dock scores were saved for single particle scattering pattern simulations. The root-mean-square-deviation (RMSD) values of these predicted models compared to the native (correct) complex structure are also recorded.

Scoring Function Based on X-ray Scattering Data

The scoring function for Z-dock program is based on molecular shape complementary, electrostatic interaction, and solvation energy etc. Higher scores indicate better chance to be the correct model. With simulated X-ray scattering data, the chi-score is used to measure the difference between datasets to reflect the structural differences. For single particle scattering data composed of N scattering patterns, having intensity values in M pixels, the SPI chi-score is defined as:

$$\chi^2_{\mathrm{spi}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M}\sum_{m=1}^{M}\left(\frac{I^{(n,m)}_{model} - I^{(n,m)}_{data}}{\sigma^{(n,m)}_{data}}\right)^2 \tag{3}$$

where $I^{(n,m)}$ is the intensity value in *n-th* pattern at pixel position m, and $\sigma^{(n,m)}$ is the associated standard deviation in the simulation data, $\sigma^{(n,m)} = (I^{(n,m)})^{1/2}$according to the Poisson noise distribution. The subscripts, *model* and *data,* refer to the values corresponding to the structures generated by Z-dock, and the values corresponding to the correct model (*data* means the simulated experimental data; while *model* means the theoretical value calculated from the predicted models). Note that the *n-th* model pattern must be in the same orientation as the *n-th* "experimental" pattern for Equation (3) to be valid. In reality, orientation is unknown during the chi-score calculation for real experimental data. Therefore, orientation matching must be carried out by minimizing the chi-score for each experimental pattern with respect to all possible orientations of the model. The Equation (3) becomes:

$$\chi^2_{\mathrm{spi}} = \frac{1}{N}\sum_{n=1}^{N}\min_{n'}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\frac{I^{(n',m)}_{model} - I^{(n,m)}_{data}}{\sigma^{(n,m)}_{data}}\right)^2\right) \tag{4}$$

where $\{n'\}$ is the set of patterns computed for any predicted model. For finer sampled orientation space using discretized euler angles, the number of model patterns grows rapidly, so the pair-wise orientation matching is very time consuming, and we offer a possible remedy in the following sub-section.

Instead of comparing single particle patterns at matched orientations, the SAXS profiles can be obtained from experiments, or from the virtual "SAXS" pattern by summing the single particle patterns. Specifically, SAXS profile is obtained by aggregating the single particle scattering data, then averaging over the angular direction, i.e.,

$$\begin{aligned} I_{\mathrm{SAXS}}(q) &= \frac{1}{N}\sum_{n=1}^{N}\frac{\int_{\phi=0}^{2\pi} I^n(q,\phi)d\phi}{\int_{\phi=0}^{2\pi} d\phi} \\ &= \frac{1}{2\pi N}\sum_{n=1}^{N}\int_{\phi=0}^{2\pi} I^n(q,\phi)d\phi \end{aligned} \tag{5}$$

$I^n(q,\phi)$ is the intensity value at polar coordinate (q,ϕ) specified by the radial component q and the azimuth angle ϕ for the *n-th* pattern. The chi-score can be calculated as:

$$\chi^2_{\mathrm{SAXS}} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{I_{\mathrm{SAXS,\,model}}(q_k) - I_{\mathrm{SAXS,\,data}}(q_k)}{\sigma_{\mathrm{SAXS,\,data}}(q_k)}\right)^2 \tag{6}$$

Orientation Matching

In order to find the orientation that best matches each "experimental" pattern, it is necessary to generate an orientation grid that spans SO(3) rotation space by discretizing three Euler angles. The step size for discretization is critical to the accuracy of orientation match. The step size can be estimated by matching the highest resolutions of 2D scattering patterns.

In order to find the best matched orientations, theoretical patterns must be simulated for all discretized orientations (after removing symmetric redundancies if there are any). Then each "experimental" pattern must be compared to all theoretical patterns for the theoretical model. The best matched pattern is identified by finding the lowest chi-scores compared to each experimental pattern. It is very computational expensive to evaluate chi-scores for all "experiment-model" pattern pairs at pixel levels. For example, if each rotational Euler angle is discretized to *n* values, to find orientations of *m* experimental patterns, there will be n^3m evaluations of 2D matrix comparison. This computational challenging problem can be sorted out in several approaches, and here we offer two solutions.

First, for the simulation case, as a proof-of-principle, we artificially confine our rotational degree of freedom within a subspace of SO(3) defined by the Euler angles ($-22.5° \leq \alpha,\beta,\gamma \leq 22.5°$). This does not solve the problem in actual applications to experimental data, which are certainly not confined to this subspace, yet this operation allows quick assessment of the effects of grid size.

The second solution is to reduce the "experimental" pattern to its angular auto-correlation, which does not depend on the in-plane rotation angle (Kam, 1977; Liu et al., 2013; Huang and Liu, 2016). The angular auto-correlation function (AC) is defined as:

$$AC(q, \Delta\phi) = \int_0^{2\pi} I(q, \phi)I(q, \phi + \Delta\phi)d\phi \qquad (7)$$

where $I(q,\phi)$ is the intensity at pixel specified using polar coordinate (q,ϕ). This requires a pre-processing of the "experimental" patterns and the theoretical patterns computed from predicted models. The AC transformation removes the in-plane rotation dependence of the scattering pattern, making the AC function depend on two Euler angles that specify a direction perpendicular to the scattering pattern. Then the AC functions are used for pairwise comparison for scoring (i.e., chi-scores of AC functions are calculated), rather than comparing each scattering pattern with every reference pattern. It can be shown that the extra overhead calculation has benefit in reducing the computational complexity from $O(n^3M)$ to $O(n^2M)$, where *n* is the number of grids for each Euler angle, and *M* is the number of experimental patterns. The computational complexity for overhead computing of AC function is $O((n^3+M)^*k)$, where n^3, M are the numbers of theoretical patterns and "experimental" patterns respectively, k is the number of discretization of in-plane rotation angle. The advantage is obvious if M>>k.

RESULTS

In this section, using simulation data with the docking decoys, we will answer four questions: (1) how many single particle scattering patterns are needed for the scoring function to converge; (2) how do the scoring functions compare to each other in terms of ranking the predicted models; (3) how does the orientation mismatching affect accuracy of the scoring functions; (4) how to speed up the orientation matching by using reduced representations of scattering patterns.

The molecular complex systems are selected from Benchmark 5.0 on Z-dock server (Vreven et al., 2015). The models are depictured in **Figure 1** and major features are summarized in **Table 1**. The native structures are available at http://liulab.csrc.ac.cn/download/zdock/.

The Convergence of Scoring Function

Both the SPI-score and SAXS-score (Equations 3, 6) need a good number of patterns to reach convergence. The first task is to determine the lower limit of this number using simulation data. Experimentally, the SAXS profile can be obtained without too much technical challenge, and even high throughput data collection is possible for standard SAXS experiments. We focus on the convergence of SPI-score in this section, because high quality single particle scattering patterns are still very difficult to obtain, even at X-ray free electron laser facilities. This is also one of the major motivations of this work, through which we hope to demonstrate that the hybrid approach for data analysis can improve the performance of both computational modeling and the XFEL data interpretation using a small set of data.

Regarding the convergence question, the SPI-score was computed with different numbers of single particle scattering patterns. The convergence can be monitored by plotting SPI-score as a function of pattern numbers. The purpose of the convergence test is to ensure that the scores are consistent and independent of number of measurement. **Figure 2A** shows the convergence of scoring function for 60 decoy structures of complex#1 (3AAD). Here, the goal is to find the minimum number of patterns required to yield a reliable scoring function. To rule out other factors, the orientation for each pattern was taken as known information, i.e., the exactly matched orientation was used for comparison. The actual cases where orientation assignment is required are considered in the following sections. As shown in **Figure 2A**, the SPI-scores have large fluctuations when the number of patterns is small, then converges quickly when the number approaches 1,000. Similar trends were observed for other complexes, and for this reason we use 1,000 scattering patterns in the SPI-score calculations through the study. It is worthwhile to note that the minimum number of scattering patterns required for a converged SPI-score varies for each system, depending on complex size, binding mode, and complex structure. The number 1,000 is a compromised choice between accuracy and speed. The SPI-scores for different predicted models are well separated when the SPI-score reach convergence, indicating that the converged SPI-scores can be used to assess the quality of the molecular complexes. In **Figure 2B**, for each decoy model, we compared the SPI-scores with 1,000 patterns and those with 2,000 patterns, the two sets of scores are perfectly lined up around y = x. Therefore, simulation results indicate that the convergence can be reached when number of patterns is above 1,000. In other words, the minimum number of patterns required to reliably scoring the predicted models is 1,000, which is feasible with current instruments at XFEL facilities.

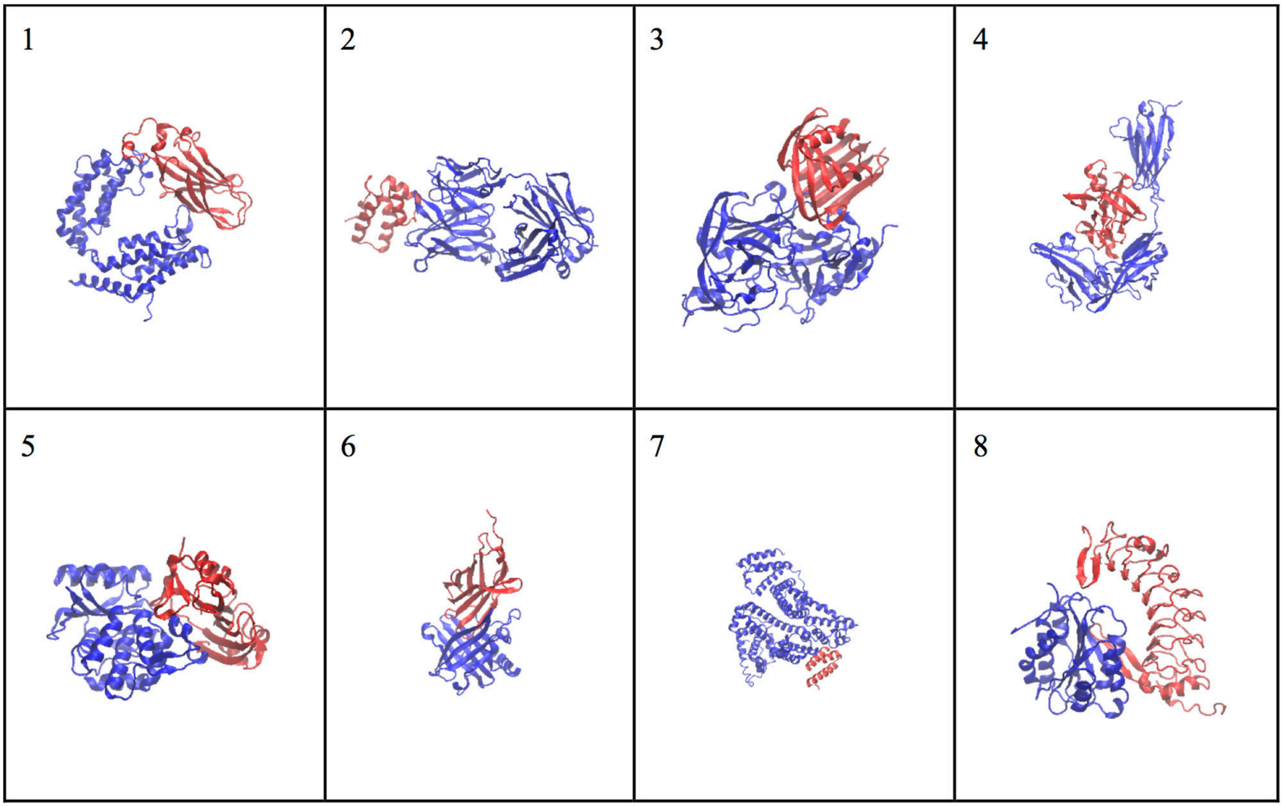


FIGURE 1 | The native structures for the molecular complexes used in this study. Two subunits are colored in blue and red in each complex, where the blue subunit is fixed and the red subunit moves around the blue subunit to generate complex structures.

TABLE 1 | The characteristics of the molecular complexes.

ID	Complex PDB code	Subunit 1 (S1)	Subunit2 (S2)	No. atom of S1	No. atom of S2	No. Residue of S1	No. Residue of S2	Difficulty in Zdock	No. atom of complex	No. Residue of complex
1	**3AAD_A:D**	1EQF_A	1TEY_A	2,164	1,231	243	144	Difficult	3,395	387
2	**2B42_B:A**	2DCY_A	1T6E_X	2,604	1,443	341	171	Easy	4,047	512
3	**1E6J_HL:P**	1E6O_HL	1A43_	3,275	577	397	69	Easy	3,852	466
4	**1IRA_Y:X**	1G0Y_R	1ILR_1	2,499	1,139	294	138	Difficult	3,638	432
5	**1JTG_B:A**	3GMU_B	1ZG4_A	2,021	1,234	242	155	Easy	3,255	397
6	**3BX7_A:C**	3BX8_A	3OSK_A	1,389	897	163	111	Middle	2,286	274
7	**2VDB_A:B**	3CX9_A	2J5Y_A	4,345	436	528	52	Easy	4,781	580
8	**1M10_A:B**	1AUQ_	1M0Z_B	1,601	2,087	184	254	Middle	3,688	438

*The complex structures are shown in **Figure 1**, labeled with the complex ID.*

The Comparison of Three Scoring Functions

The power of ranking for each scoring function can be evaluated by studying the correlation between the scores and model differences. The RMSD is one of the most commonly used measurements for model comparison. In **Figure 3**, the ranking power for SPI-score and SAXS-score are summarized for the complex#1, which belongs to "difficult" docking case. As shown in **Figure 3**, the scattering plots clearly show that both SPI-scores and SAXS-scores are positively correlated with the RMSD values in general. For the case of complex#1, the correlation coefficients between SPI-score and RMSD is 0.59, and the correlation coefficient between SAXS-score and RMSD is smaller, giving a value of 0.36. To better quantify the ranking power of the scoring functions, a probability distribution function of RMSD, $P^{(\mathrm{RMSD},n)}$, was computed for top n selected models. Specifically, the probability for a model differing from the native structure by a particular RMSD value was calculated for n models with lowest scores. The probability distribution functions are plotted in (**Figures 3B,E**), where the $P^{(\mathrm{RMSD},n)}$ for n = 25, 100, 1,000

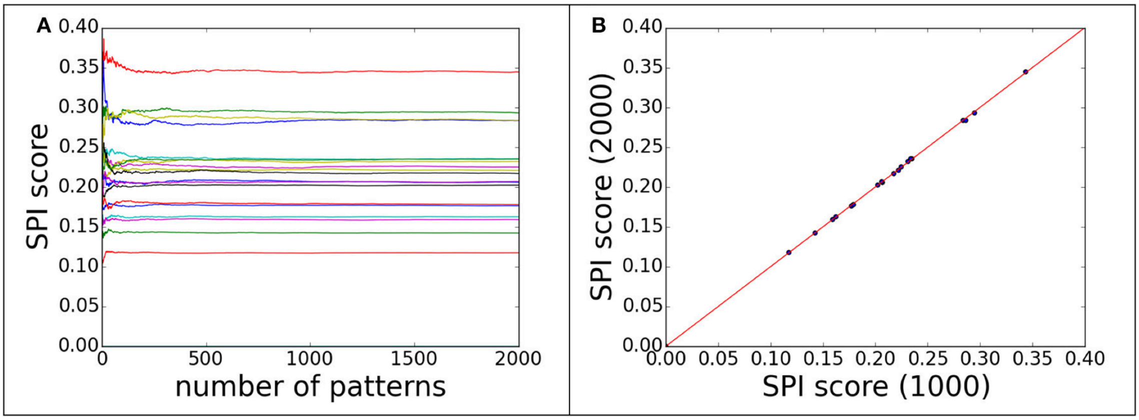


FIGURE 2 | The convergence of SPI-score for patterns with correct orientations. 60 decoys from complex#1 are used to demonstrate the convergence progress of SPI-score. **(A)** the SPI-score is plotted as a function of pattern quantity, each line represent the SPI-score of one predicted decoy model by comparing model patterns to "experimental" data. **(B)** The comparison of SPI-scores computed using 1,000 or 2,000 scattering patterns, whose orientations are random.

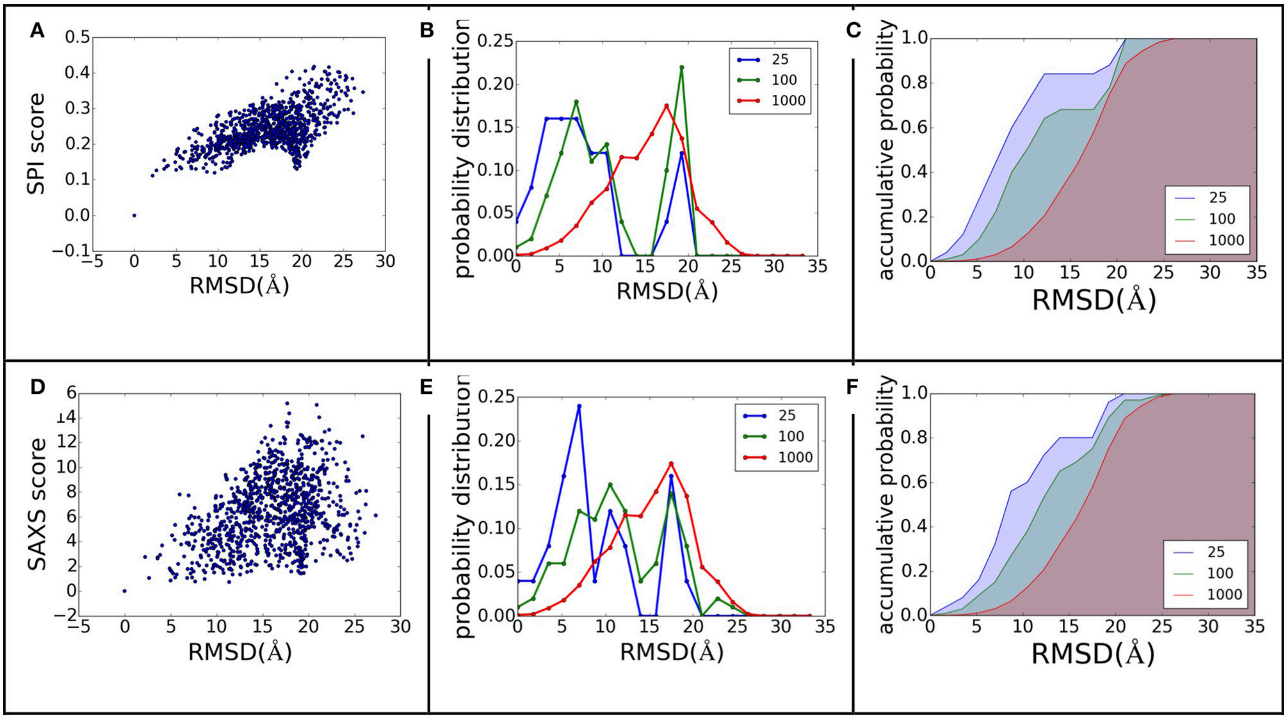


FIGURE 3 | The ranking power comparison between SPI-score and SAXS-score. (A,D) the scatter plot of scores as a function of RMSD. **(B,E)** the probability distribution function of RMSD for the selected models. The three curves correspond to the distribution function of top 25, top 100, and all models. **(C,F)** The accumulative probability functions corresponding to the three distributions in **(B,E)**. The green and blue shaded area indicates the gain of ranking power by selecting subsets of models.

(all) are calculated and compared. Based on the probability distribution and the correlation coefficients between the scoring function and the RMSD, it is clear that both SPI-scores and SAXS-scores are capable of selecting models that have lower RMSD values with respect to the native structure, while the SPI-scores have stronger selecting power. The probability of selecting models with lower RMSD values is increased after model ranking using either SPI-score or SAXS-score. This increasing trend is more pronounced for the ranking using SPI-scores. The probability function is converted to accumulative probability function by integration, as shown in (**Figures 3C,F**). On the other side, the scoring function from the Z-dock can select a few best matched models from predicted models, the overall ranking power is not as good as the SPI or SAXS scoring functions (data not shown). This makes the z-dock scoring function vulnerable to insufficient model generation. The SPI-score is a more powerful function not only because it can be used to select the lowest RMSD models, but also because the model ranking is consistent with structure differences.

The same analyses were carried out for eight complexes, as described in **Figure 1** and **Table 1** (for the other seven complexes, see Supplementary Material). To quantify the ranking power, we define a new parameter, the area under the accumulative probability curve (AUC, area under curve), similar to the measure of classification power. For each accumulative probability distribution curve, the area is calculated by integration. The x-axis, the range of RMSD, can be normalized to the fraction of the largest RMSD value in the decoy sets. Therefore, the AUC has a largest possible area of 1.0, as an extreme case when all models are ranked in the same order as the RMSD with respect to the native structure. Under this definition, larger AUC values correspond to more powerful ranking method. We calculated AUC at three levels of selection (top 25, top 100, and all models) for each method (SPI-scoring, SAXS-scoring, and Z-dock scoring), same as the demonstration example in **Figure 3**. In **Table 2**, the AUC statistics are summarized, suggesting that SPI-score has better performance in terms of ranking power, compared to SAXS-score. There is one exception in the case of complex#2, where the ranking power of SAXS-score is slightly better than that of the SPI-score.

The Effects of Orientation Mismatching

As mentioned in the previous section, the scoring functions can be reliably obtained from about 1,000 single particle scattering patterns, which are feasible to collect with the current XFEL experimental technologies. However, the results in the previous section are obtained based on a strong assumption that the orientations of the models are "exactly" matched to the orientation of native structure. It is known that orientation determination is challenging using computational methods, which utilize cross correlations between patterns by matching "experimental data" to the "model data" at discretized orientations.

During the orientation matching, the actual orientation can be deviated from the computational matched orientation. The mismatching can happen at two levels, as schematically illustrated in **Figure 4**: (1) the discretized orientations for the "model" patterns are not fine enough to match the "exact" orientation but rounding up to the nearest orientations of the "exact" orientation, and this finite discretization is unavoidable due to the limitation of computing power; (2) the orientations

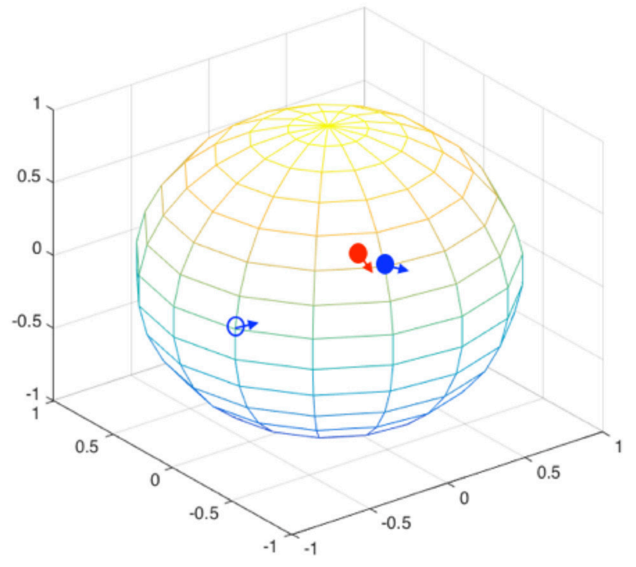


FIGURE 4 | Orientation mismatching scenarios. Two rotation angles can be mapped to the points on a sphere, the third angle is the in-plane rotation indicated using the arrow at each point. The red solid circle and the associate arrow indicate the orientation of one experimental pattern, the blue circles and arrows indicate possible orientations. The orientation deviation of solid blue circle from the correct values (red solid circle) is due to the discretization of SO(3) rotation space; and the orientation mismatching to the open blue circle is attributed to large conformational difference. For the models that are similar to the correct complex structure, the orientations are likely to be identified to the vicinity of correct orientations (see **Figure 5**).

TABLE 2 | The performance of scoring functions.

Complex ID	Z-dock			SAXS			SPI		
	Top 25	Top 100	All	Top 25	Top 100	All	Top 25	Top 100	All
1	0.53	0.55	0.55	0.71	0.64	0.54	**0.74**	0.65	0.54
2	0.77	0.78	0.78	**0.86**	0.83	0.78	0.84	0.83	0.78
3	0.68	0.65	0.57	0.78	0.65	0.56	**0.83**	0.71	0.56
4	0.54	0.46	0.37	0.76	0.69	0.36	**0.77**	0.70	0.36
5	0.68	0.57	0.52	0.75	0.63	0.53	**0.85**	0.65	0.51
6	0.78	0.75	0.62	0.72	0.68	0.51	**0.83**	0.75	0.51
7	0.69	0.63	0.60	0.82	0.74	0.58	**0.88**	0.85	0.58
8	0.59	0.56	0.48	0.73	0.64	0.49	**0.78**	0.67	0.48

The numbers are the AUC values for the selected models using the corresponding scoring functions. The numbers in bold font indicate the highest ranking power for top 25 models.

for the best "experimental-model" pattern pairs judged by the chi-score or correlation functions are not matched, meaning that the matching is messed up by conformational differences. In this section, we implicitly considered both factors by not providing orientation information during pattern matching process. The SPI-score is calculated using the modified formula (Equation 4).

Using complex#1 (3AAD) as an example again, the orientation mismatching effects are studied. The matching results are summarized in **Figure 5**, which shows the deviation of the Euler angles from the correct orientation. For models with smaller RMSD values, most of the recovered orientations are indeed close to the orientations of "experimental" data, suggesting that the major orientation mismatching is due to the discretization of SO(3) rotation space. For the models with larger RMSD values, the success rate of determining the pattern orientations are lower, which can be explained as the consequences of conformational changes that overwhelm orientation variation effects. The statistics of the orientation deviation are summarized in Table S2. It is interesting to observe that the second rotation angle, β, is more accurately recovered using the reference matching approach than the other two angles. Using simulation data, we mapped the landscape of SPI-score due to the orientation differences. The results reveal that the SPI-score landscape around the β rotation is smoother relatively, suggesting that mismatching due to finite discretization of β angle can be tolerated. In other words, the chance of recovering the orientation within the vicinity of correct β angle is higher.

Using the subset of SO(3) rotation space, we studied the case of discretized representations of the orientations using step size of 3 degrees. The results show that the orientation matching is reasonable, and the ranking power is similar to the ideal cases discussed in the previous section. The AUC for top 25 models is 0.72 vs. 0.74 for the ideal case for complex#1 (see **Table 3**). Nevertheless, as the discretization step size increases, the SPI-score becomes less accurate. As a result, the ranking power of the SPI-score is reduced. When the orientation sampling is fine enough (step size of 3 degrees is sufficient in this simulation), the SPI-score outperforms the SAXS-score, which does not depend on orientation matching. The optimal discretization of SO(3) rotation space has to be chosen under the considerations of (1) the computational cost and (2) the accuracy of orientation matching. For the latter concern, the discretization step size should match the resolution of the scattering signals. For low resolution data, larger discretization step sizes can be tolerated. This may provide an opportunity of implementing multilevel model selection method to speed up the overall computing: using low resolution data to rule out a set of very unlikely models, and using higher resolutions to narrow down the best matched models.

In order to quantify the effects of background noise to the ranking results, the signal-noise-ratio (SNR, defined as

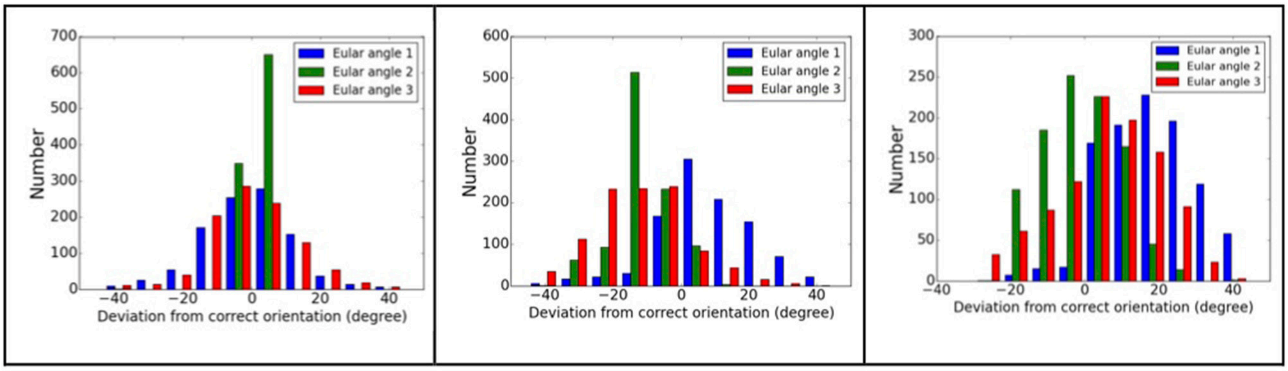


FIGURE 5 | Orientation matching results. The dependency of matching accuracy on the conformational differences, the deviations from correct orientations for three models: left to right, the RMSD values are 2.2, 10.5, and 15.1 Å. Larger RMSD values correspond to larger deviation from correct orientations.

TABLE 3 | Comparison of three methods for orientation matching.

	Scattering pattern			Radial profile			Correlation pattern		
Number of selected models	1,000	100	25	1,000	100	25	1,000	100	25
AUC (RMSD)*	0.54	0.66	0.72	0.54	0.63	0.73	0.54	0.62	0.72
AUC (s-score)#	0.37	0.69	0.80	0.37	0.63	0.74	0.37	0.59	0.73
Computing time (seconds)$		211.47			1.07			14.30	

[*] *The AUC (area under curve) using RMSD as the measure for model difference.*

[#] *The AUC (area under curve) using s-score as the measure for model difference.*

[$] *Computing time needed for orientation matching for one pattern: for raw patterns, comparing to 4,096 (16ˆ3) patterns; for radial profile, comparing to 256 (16ˆ2) lines; for correlation function, comparing to 256 (16ˆ2) auto-correlation patterns.*

The results are for complex#1, and the reference patterns from models are in subspace of SO(3), with discretized euler angles cover a range of [−22.5°, 22.5°] using step size of 3°.

the ratio between variances of signals and noise) was varied from 100 to 0.1 logarithm spaced. The results presented in the previous sections were essentially the same with small variations in the ranking, although the absolute values of scores are larger for low SNR (i.e., larger noises for same level of signals).

Speed Up the Matching of Orientations

The pairwise pattern comparison requires the exhaustive sampling SO(3) rotation space using three euler angles. The pairwise 2D pattern comparison is expensive computationally, limiting the applications of this approach to large dataset. It has been found that some preprocessing of the raw scattering data can reduce computational cost for downstream analysis. First, the in-plane rotation angle can be decoupled from the other two rotations, by using an angular auto-correlation function (Huang and Liu, 2016). In this case, the computational complexity can be reduced significantly by converting the raw scattering patterns to auto-correlation functions, which are used for comparison instead of the scattering patterns. We compared the performance of the new SPI-score based on the auto-correlation functions to original SPI-score in **Table 3**. The results show that the ranking power is maintained to be similar, and the computational time is reduced by a factor of 14.8. Furthermore, each pattern can be reduced to a radial profile (1D) by integrating over the azimuth angle, yielding a curve that is similar to SAXS curve. Because the scores computed using the radial profile representation are essentially an average of chi-scores between matched patterns (i.e., additional information are obtained by *minimizing* the differences between experimental data and reference model), it is different from SAXS curve that is the average of radial profiles (by assuming random orientation distributions). The results show that this radial profile, although with compressed information, can be used for pairwise pattern comparisons. The score computed from radial profiles after orientation matching has a ranking power comparable to the SPI-score, as shown in **Table 3**. This radial profile representation further reduces the computing time by another 13.4 folds (∼200 times faster than using raw pattern comparison). It is worthwhile to point out that both reduced representations do not need to sample the in-plane rotation, therefore, significantly reducing computing time of generating model patterns as well.

DISCUSSIONS

X-Rays Only See Electron Distributions, Not Sequential Information

X-ray scattering/diffraction is due to the interaction with electrons, so the subject under probing is the electron density map. In crystallography, the atomic models are built to the electron maps by incorporating information of amino acid sequences. Without considering the sequences, the information from X-ray scattering is not sufficient to describe full features of atomic models, especially when the resolution of X-ray scattering signal is worse than atomic resolution. We observed several cases that the low SPI-scores correspond to the predicted models with large RMSD values (see **Figures 3A,D**). A closer examination of the corresponding models reveals that the predicted docking site is correct, but the docking pose (i.e., the orientation of the docking subunit) is opposite to the correct model. The symmetry of protein molecules can also introduce confusions in the analysis of X-ray scattering data. For example, in **Figure 6**, the fixed subunit molecule has a 2-fold pseudo symmetry, making it hard to distinguish the native binding modes from its symmetric counterpart. This explains some observations where the SPI-score (or SAXS-score) positively correlates to the RMSD values for models that are similar to the native structure, but the trend becomes reversed for very large RMSD values (lower SPI-scores correspond to models with larger RMSD).

An alternative measurement for structural differences is to treat each model as a point cloud, which ignores the sequence and connections between these points. Then, the spatial correlations between two models can be computed by maximizing their overlaps. The correlation coefficients can be calculated as the following:

$$cc = \frac{\langle \rho_1(\mathbf{r})\rho_2(\mathbf{r})\rangle - \langle \rho_1(\mathbf{r})\rangle\langle \rho_2(\mathbf{r})\rangle}{\sigma_1\sigma_2} \qquad (8)$$

where $\rho_{1/2}(\mathbf{r})$is the electron density of model 1 or 2 at position $\mathbf{r}$, $\sigma_{1/2}^2$ is the variance of model 1 or 2. We applied the model alignment method described in SASTBX programs (Liu et al., 2012). Briefly, the models are shifted such that the centers of mass coincide with the real space origin, then the relative orientations of the models are optimized by finding the largest overlaps

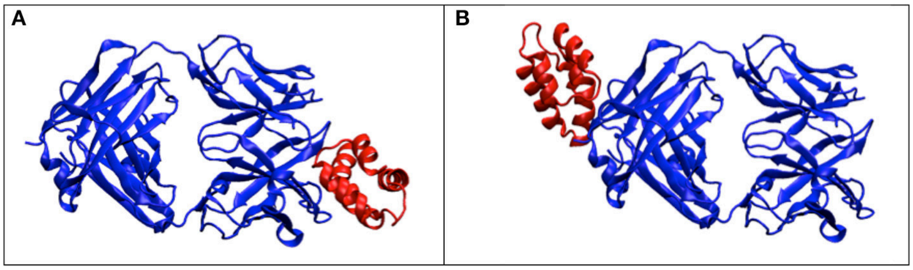


FIGURE 6 | The effects of symmetry. The dimer complex has a pseudo-symmetry (blue color), which may reduce the model ranking power of scattering data based scoring functions. **(A)** the correct structure for the complex; **(B)** the model that has similar electron density to **(A)** after rotation, but differs significantly from **(A)** in terms of RMSD.

between models. The computing is sped up by sampling three Euler angles with fast fourier transform (FFT) algorithm. In order to be consistent with RMSD that is a distance measure, we define a model difference parameter, *s-score* $s = 1.0 - cc$, to gauge the ranking power of SPI-score or SAXS-score. Using complex complex#1 (3AAD) as an example, the ranking power for scattering based scoring functions is summarized in **Figure 7**. The comparison to **Figure 3** suggests that the X-ray scattering data is more useful in describing electron density maps. In order to compare structures that have sequential and connection information, it is necessary to incorporate knowledge of physics and chemistry. When considering the docking problem, the biochemical properties at the interface are crucial, so the model evaluation should include physicochemical terms.

Joint Scoring Function Is Needed to Outperform Individual Functions

We examined the relation between SPI-scores and the SAXS-scores by computing the correlation coefficients (See Table S3 in Supplementary Material). The results suggest positive correlation between the two scoring functions, with varying correlation strength (0.12 to 0.81). This variation suggests that the two scoring functions contain different structural information. As shown in the Result section, the SPI-score is better in ranking the models, so it is natural to include the SPI-score in the joint scoring function.

The built-in scoring function of Z-dock is not sufficient in ranking the models, but it has its merit by design, which incorporates physicochemical terms and geometry complementary properties. The model ranking by each scoring approach is unlikely to outperform the combined scores. The optimized IRAD (integration of residue- and atom-based potentials for docking) function was reported to improve the model ranking by combining several scoring functions (Vreven et al., 2011). We re-ranked the models using z-rank program where IRAD functions are implemented (Pierce and Weng, 2007). However, the model ranking power is increased modestly in this case, mainly because the Z-dock program has a built-in scoring function that give comparable ranking power as IRAD scores.

In order to explore the potential of joint scoring functions, we experimented one method of combination using SPI-score and Z-dock score using a voting system: first, the Z-scores are calculated for each model with either SPI-score or Z-dock score, then the Z-scores are combined to give an overall ranking. The experiment for complex model (#1) dataset does not yield significant improvement. This suggests that it is not trivial to combine the scores from different evaluation methods, because hybrid does not mean simple linear combination. Designing better ways to combine different scoring functions are subjects of future studies.

Hybrid Approach Can Be Applied to Incomplete Dataset

Although the idea in this work is about applying experimental data in SPI or SAXS in the ranking of docking models, the impact of modeling to the data interpretation is equally significant. As mentioned, the XFEL facilities are scarce resources, although

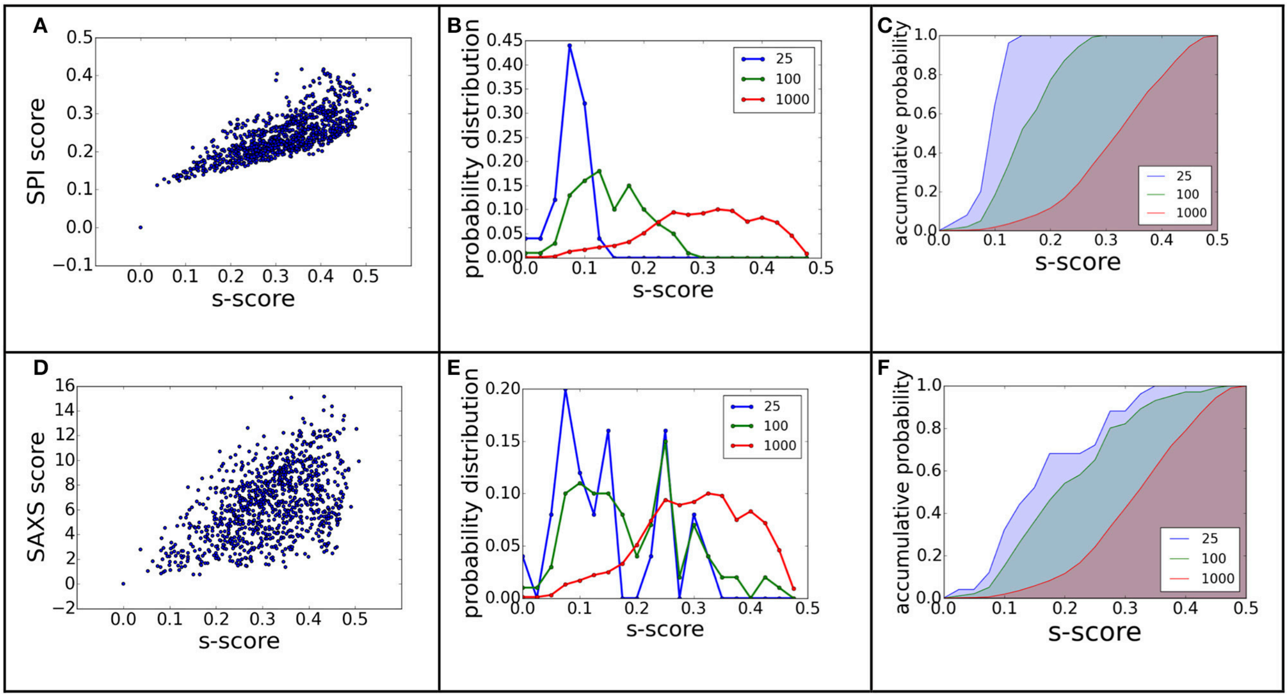


FIGURE 7 | The ranking power revisited using electron density map differences. The figure **(A–F)** caption is the same as **Figure 3**, except that the model difference is measured using *s-score*, instead of RMSD.

more XFEL facilities will be commissioned in the near future, there are still some technological challenges to carry out high throughput single particle scattering experiments. It is not practical to collect complete datasets for model reconstructions that are solely based on experimental data yet. If computational modeling, such as molecular docking or protein structure prediction, is integrated in the data interpretation, it is possible to determine structures from a much smaller dataset (∼1,000 patterns in the simulation cases). In other words, the hybrid approach turns a reverse modeling (from intensity to electron density map) problem to a ranking problem of the predicted models. Given the advances in high performance computing, sampling algorithms will be capable of generating diverse models, in which the correct structure is very likely to be included. Then the model ranking and selection criteria is the key to model determination.

In a related research field, the cryogenic electron microscopy (CryoEM), the projection images of molecules are detected. Several algorithms have been developed to reconstruct detailed 3D structures based on projection images. In general, such dataset must be composed of a large number of images (at the order of 10 thousands to 100 thousands), in order to obtain high resolution structures. For relative low resolution model reconstruction, it is feasible to obtain an *ab initio* density map with <1,000 patterns using the maximum likelihood method (Ekeberg et al., 2015). A global assignment of orientations is also reported for simulation data using common line algorithm for fewer than 1,000 patterns (Singer and Shkolnisky, 2011). The hybrid approach reported here can potentially be used to select the models at higher resolutions with similar amount of data, given the availability of high resolution structures of docking subunits.

CONCLUSION

The development of XFEL and its application in single particle imaging requires fast and reliable methods to interpret experimental data, especially when the dataset is not sufficient to convert scattering signals to a unique structural model. In this work, we demonstrated that single particle experimental data is valuable in ranking the predicted models, and this hybrid approach can be one solution for structure determination with limited XFEL data.

AUTHOR CONTRIBUTIONS

HL designed the research, HW carried the simulations, both authors analyzed the data and contributed to the manuscript writing.

ACKNOWLEDGMENTS

This project is supported by the natural science foundation of china (Award number: 11575021) and the 1,000 talent program to HL. The project was also partially supported by the NSAF grant (award number: U1430237, U1530401). This computational work is supported by a Tianhe-2JK computing time award at the Beijing Computational Research Center (CSRC).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmolb.2017.00023/full#supplementary-material

REFERENCES

Aquila, A., Barty, A., Bostedt, C., Boutet, S., Carini, G., dePonte, D., et al. (2015). The linac coherent light source single particle imaging road map. *Struct. Dyn.* 2, 41701. doi: 10.1063/1.4918726

Bader, G. D., Betel, D., and Hogue, C. W. V (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Res.* 31, 248–250. doi: 10.1093/nar/gkg056

Bax, A., and Grzesiek, S. (1993). "Methodological Advances in Protein NMR," in *NMR of Proteins* eds G. M. Clore and A. M. Gronenborn (London: Macmillan Education UK), 33–52.

Bogan, M. J., Benner, W. H., Boutet, S., Rohner, U., Frank, M., Barty, A., et al. (2008). Single particle X-ray diffractive imaging. *Nano Lett.* 8, 310–316. doi: 10.1021/nl072728k

Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., et al. (2011). Femtosecond X-ray protein nanocrystallography. *Nature* 470, U73–U81. doi: 10.1038/nature09750

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins Struct. Funct. Genet.* 52, 80–87. doi: 10.1002/prot.10389

Cheng, Y. (2015). Single-Particle Cryo-EM at crystallographic resolution. *Cell* 161, 450–457. doi: 10.1016/j.cell.2015.03.049

Cheng, Y., Grigorieff, N., Penczek, P. A., and Walz, T. (2015). A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* 161, 438–449. doi: 10.1016/j.cell.2015.03.050

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* 405, 823–826. doi: 10.1038/35015694

Ekeberg, T., Svenda, M., Abergel, C., Maia, F. R. N. C., Seltzer, V., Claverie, J.-M., et al. (2015). Three-dimensional reconstruction of the giant mimivirus particle with an X-Ray free-electron laser. *Phys. Rev. Lett.* 114:98102. doi: 10.1103/PhysRevLett.114.098102

Emma, P., Akre, R., Arthur, J., Bionta, R., Bostedt, C., Bozek, J., et al. (2010). First lasing and operation of an angstrom-wavelength free-electron laser. *Nat. Photonics* 4, 641–647. doi: 10.1038/nphoton.2010.176

Förster, F., Webb, B., Krukenberg, K. A., Tsuruta, H., Agard, D. A., and Sali, A. (2008). Integration of Small-Angle X-Ray scattering data into structural modeling of proteins and their assemblies. *J. Mol. Biol.* 382, 1089–1106. doi: 10.1016/j.jmb.2008.07.074

Göbl, C., Madl, T., Simon, B., and Sattler, M. (2014). NMR approaches for structural analysis of multidomain proteins and complexes in solution. *Prog. Nucl. Magn. Reson. Spectrosc.* 80, 26–63. doi: 10.1016/j.pnmrs.2014.05.003

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., et al. (2003). Protein–protein docking with simultaneous optimization of Rigid-body displacement and side-chain conformations. *J. Mol. Biol.* 331, 281–299. doi: 10.1016/S0022-2836(03)00670-3

Huang, L., and Liu, H. (2016). Fast algorithm for determining orientations using angular correlation functions and Bayesian statistics. *BioRxiv* 1–14. doi: 10.1101/074732

Janin, J. (2005). Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein Sci.* 14, 278–283. doi: 10.1110/ps.041081905

Kam, Z. (1977). Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules* 10, 927–934. doi: 10.1021/ma60059a009

Konarev, P. V., Petoukhov, M. V., Volkov, V. V., and Svergun, D. I. (2006). ATSAS 2.1, a program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* 39, 277–286. doi: 10.1107/S0021889806004699

Krissinel, E., and Henrick, K. (2007). Inference of Macromolecular assemblies from Crystalline State. *J. Mol. Biol.* 372, 774–797. doi: 10.1016/j.jmb.2007.05.022

Lensink, M. F., Velankar, S., and Wodak, S. J. (2016). Modeling protein-protein and protein-peptide complexes: CAPRI 6 th edition. *Proteins* 85, 359–377. doi: 10.1002/prot.25215

Liu, H., Hexemer, A., and Zwart, P. H. (2012). The Small Angle Scattering ToolBox (SASTBX): an open source software for biomolecular small angle scattering. *J. Appl. Crystallogr.* 45, 587–593. doi: 10.1107/s0021889812015786

Liu, H., Poon, B. K., Saldin, D. K., Spence, J. C. H., and Zwart, P. H. (2013). Three-dimensional single-particle imaging using angular correlations from X-ray laser data. *Acta Crystallogr. A* 69, 365–373. doi: 10.1107/S0108767313006016

Liu, H., and Spence, J. C. H. (2016). XFEL data analysis for structural biology. *Quant. Biol.* 4, 159–176. doi: 10.1007/s40484-016-0076-z

Mainz, A., Religa, T. L., Sprangers, R., Linser, R., Kay, L. E., and Reif, B. (2013). NMR spectroscopy of soluble protein complexes at one mega-dalton and beyond. *Angew. Chemie Int. Ed.* 52, 8746–8751. doi: 10.1002/anie.201301215

Mattinen, M.-L., Pääkkönen, K., Ikonen, T., Craven, J., Drakenberg, T., Serimaa, R., et al. (2002). Quaternary structure built from subunits combining, NMR and small-angle x-ray scattering data. *Biophys. J.* 83, 1177–1183. doi: 10.1016/S0006-3495(02)75241-7

Merk, A., Bartesaghi, A., Banerjee, S., Falconieri, V., Rao, P., Davis, M. I., et al. (2016). Breaking Cryo-EM resolution barriers to facilitate drug discovery. *Cell* 165, 1698–1707. doi: 10.1016/j.cell.2016.05.040

Munke, A., Andreasson, J., Aquila, A., Awel, S., Ayyer, K., Barty, A., et al. (2016). Coherent diffraction of single Rice Dwarf virus particles using hard X-rays at the linac coherent light source. *Sci. Data* 3:160064. doi: 10.1038/sdata.2016.64

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E., and Hajdu, J. (2000). Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* 406, 752–757. doi: 10.1038/35021099

Pierce, B., and Weng, Z. (2007). ZRANK: Reranking protein docking predictions with an optimized energy function. *Proteins Struct. Funct. Bioinforma.* 67, 1078–1086. doi: 10.1002/prot.21373

Scheres, S. H. W. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* 180, 519–530. doi: 10.1016/j.jsb.2012.09.006

Schindler, C. E. M., de Vries, S. J., Sasse, A., and Zacharias, M. (2016). SAXS data alone can generate high-quality models of protein-protein complexes. *Structure* 24, 1387–1397. doi: 10.1016/j.str.2016.06.007

Schlichting, I. (2015). Serial femtosecond crystallography: the first five years. *IUCrJ* 2, 246–255. doi: 10.1107/S205225251402702X

Schneidman-Duhovny, D., Kim, S., Sali, A., Hura, G., Menon, A., Hammel, M., et al. (2012). Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct. Biol.* 12:17. doi: 10.1186/1472-6807-12-17

Seibert, M. M., Ekeberg, T., Maia, F. R. N. C., Svenda, M., Andreasson, J., Jönsson, O., et al. (2011). Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature* 470, 78–81. doi: 10.1038/nature09748

Shen, Y., and Bax, A. (2015). Homology modeling of larger proteins guided by chemical shifts. *Nat. Methods* 12, 747–750. doi: 10.1038/nmeth.3437

Singer, A., and Shkolnisky, Y. (2011). Three-dimensional structure determination from common lines in Cryo-EM by eigenvectors and semidefinite programming. *SIAM J. Imaging Sci.* 4, 543–572. doi: 10.1137/090767777

Tokuhisa, A., Jonic, S., Tama, F., and Miyashita, O. (2016). Hybrid approach for structural modeling of biological systems from X-ray free electron laser diffraction patterns. *J. Struct. Biol.* 194, 325–336. doi: 10.1016/j.jsb.2016.03.009

Vreven, T., Hwang, H., and Weng, Z. (2011). Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci.* 20, 1576–1586. doi: 10.1002/pro.687

Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., et al. (2015). Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* 427, 3031–3041. doi: 10.1016/j.jmb.2015.07.016

Zheng, W., and Doniach, S. (2002). Protein structure prediction constrained by solution X-ray scattering data and structural homology identification. *J. Mol. Biol.* 316, 173–187. doi: 10.1006/jmbi.2001.5324

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 10 March 2017
doi: 10.3389/fmolb.2017.00013

Coarse-Grained Conformational Sampling of Protein Structure Improves the Fit to Experimental Hydrogen-Exchange Data

Didier Devaurs[1], *Dinler A. Antunes*[1], *Malvina Papanastasiou*[2,3], *Mark Moll*[1], *Daniel Ricklin*[2,4], *John D. Lambris*[2] and *Lydia E. Kavraki*[1]*

[1] Department of Computer Science, Rice University, Houston, TX, USA, [2] Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA, [3] Broad Institute of MIT & Harvard, Cambridge, MA, USA, [4] Department of Pharmaceutical Sciences, University of Basel, Basel, Switzerland

Monitoring hydrogen/deuterium exchange (HDX) undergone by a protein in solution produces experimental data that translates into valuable information about the protein's structure. Data produced by HDX experiments is often interpreted using a crystal structure of the protein, when available. However, it has been shown that the correspondence between experimental HDX data and crystal structures is often not satisfactory. This creates difficulties when trying to perform a structural analysis of the HDX data. In this paper, we evaluate several strategies to obtain a conformation providing a good fit to the experimental HDX data, which is a premise of an accurate structural analysis. We show that performing molecular dynamics simulations can be inadequate to obtain such conformations, and we propose a novel methodology involving a coarse-grained conformational sampling approach instead. By extensively exploring the intrinsic flexibility of a protein with this approach, we produce a conformational ensemble from which we extract a *single* conformation providing a good fit to the experimental HDX data. We successfully demonstrate the applicability of our method to four small and medium-sized proteins.

Keywords: protein conformational sampling, coarse-grained conformational sampling, molecular dynamics, experimental data fitting, hydrogen/deuterium exchange, mass spectrometry, nuclear magnetic resonance spectroscopy, X-ray crystallography

OPEN ACCESS

Edited by:
Slavica Jonic,
IMPMC, Sorbonne Universités - CNRS UMR 7590, UPMC Univ Paris 6, MNHN, IRD UMR 206, France

Reviewed by:
Paolo De Los Rios,
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Patrice Koehl,
University of California, Davis, USA

****Correspondence:***
Lydia E. Kavraki
kavraki@rice.edu

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *11 November 2016*
Accepted: *24 February 2017*
Published: *10 March 2017*

Citation:
Devaurs D, Antunes DA, Papanastasiou M, Moll M, Ricklin D, Lambris JD and Kavraki LE (2017) Coarse-Grained Conformational Sampling of Protein Structure Improves the Fit to Experimental Hydrogen-Exchange Data. Front. Mol. Biosci. 4:13. doi: 10.3389/fmolb.2017.00013

1. INTRODUCTION

Hydrogen/deuterium exchange (HDX) is a chemical phenomenon in which hydrogen atoms of molecules are exchanged with deuterium atoms of the solvent (Engen et al., 2011). Contrary to other structural biology techniques, such as nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography, HDX experiments cannot reveal the three-dimensional structure of a molecule, but they can provide valuable structural information (Huang and Chen, 2014). This has led to numerous applications for the analysis of protein structure and conformational changes, as well as protein folding and interactions (Pirrone et al., 2015). As they monitor HDX over time (see Section 2.1), HDX detected by mass spectrometry (HDX-MS) experiments also allow studying protein dynamics (Wei et al., 2013). HDX-MS has benefited from the development of various computational tools (Claesen and Burzykowski, 2016), and has proven useful in the study of challenging systems, such as molecular complexes or membrane proteins (Harrison and Engen, 2016). Additionally, HDX-MS is having a deep impact in drug discovery and drug development (Deng et al., 2016), where it has helped characterize various biopharmaceuticals (Pirrone et al., 2015) and innate immunity proteins (Schuster et al., 2007; Sfyroera et al., 2015; Papanastasiou et al., 2017), among others.

Despite the clear benefits of monitoring HDX for structural analysis, it is sometimes difficult to interpret experimental HDX data. This data may be reported as protection factors (Jaswal, 2013), often visualized on a protein *heat map* (Huang and Chen, 2014) built using a structural model reported in the Protein Data Bank (PDB, RRID:SCR_012820), if available. However, it has been suggested that the correspondence between these structural models and experimental HDX data can be inadequate, especially for models produced by X-ray crystallography (Radou et al., 2014). This is due to the difference in nature between HDX data and crystallographic data: only HDX data can reflect the inherent variability of a specific protein state. As a result, it has been argued that experimental HDX data should rather be interpreted using a conformational ensemble produced by a molecular dynamics (MD) simulation (Best and Vendruscolo, 2006; Radou et al., 2014). However, this method can also fail at expressing the variability of a protein state in the same way as experimental HDX data does. In a previous study, we have observed that a single conformation extracted from a conformational ensemble produced by an MD simulation could provide a better fit to experimental HDX data than the whole ensemble (Devaurs et al., 2016). Therefore, it is reasonable to try and fit experimental HDX data using a single protein conformation; this can also be computationally advantageous.

In this paper, we propose a novel methodology to obtain a single conformation providing a good fit to the experimental HDX data collected for a protein, after confirming that crystal structures and conformations produced by MD simulations might not be good choices. Our methodology involves a coarse-grained conformational sampling tool that allows exploring the flexibility of a protein by generating a conformational ensemble, starting from the crystal structure of this protein (see Section 3.3). We evaluate our methodology on four small and medium-sized proteins that correspond to two scenarios: for three proteins, both the HDX data and the crystal structure are known to describe their native state; for one protein, the HDX data and crystal structure are known to describe two different states (see Section 3.4). The evaluation results show that our methodology can successfully produce conformations that provide a good fit to the experimental HDX data, for these four proteins (see Section 4).

A critical element of any method aiming to analyze the correspondence between a protein's structure and its HDX data is the definition of an *HDX prediction model*. Indeed, in such a method, some HDX data has to be derived from the protein's structure; then, one can assess the goodness-of-fit between this structurally-derived HDX data and the experimentally-observed HDX data. By comparing different protein conformations, it is then possible to determine which conformation provides the best estimates for the experimental HDX data (see Section 3.3). The challenge here is that, although numerous HDX prediction models have been proposed, none of them has yet been widely recognized and adopted by researchers in this field (see Section 2.2). Furthermore, a recent evaluation study has shown the limitations of several existing models (Skinner et al., 2012b). To mitigate this issue, we have integrated in our methodology the model that performed best in that evaluation study (see Section 3.1). Our approach compensates for the current limitations and achieves a successful application of this HDX prediction model (see Section 5). This is accomplished by using coarse-grained conformational sampling as a way to extensively explore the intrinsic flexibility of a given protein.

2. BACKGROUND

2.1. Hydrogen/Deuterium Exchange (HDX) in Proteins

Hydrogen exchange is a chemical phenomenon in which hydrogen atoms of proteins are exchanged with hydrogens in the surrounding solvent (Engen et al., 2011). Intuitively, the extent to which different parts of a protein are subjected to this exchange is influenced by their solvent accessibility and by the protein's structure (Wei et al., 2013). Therefore, researchers have worked on quantifying hydrogen exchange, as a way to gain information on a protein's structure. This is made possible by the fact that this exchange takes place with any isotope of hydrogen, such as deuterium. If a protein, initially kept in a regular water solution (H_2O), is placed in a "heavy water" solution (D_2O), the hydrogen in the protein will exchange with the deuterium in the solvent. This phenomenon is referred to as hydrogen/deuterium exchange (HDX).

Using experimental techniques sensitive to differences between hydrogen isotopes, one can monitor HDX (Englander et al., 1997; Engen et al., 2011). In the 1970s, nuclear magnetic resonance (NMR) spectroscopy was the main approach to measure HDX, leveraging the differences in magnetic properties of hydrogen and deuterium (Huang and Chen, 2014). However, HDX-NMR experiments were hindered by practical weaknesses of NMR, such as the limit on the size of proteins that could be investigated. In the 1990s, advances in mass spectrometry (MS) made this technique an interesting alternative to measure HDX. HDX-MS experiments rely on that the mass of deuterium is about twice the mass of hydrogen: deuterium uptake (i.e., the amount of deuterium incorporated in the protein) thus corresponds to an increase in mass. Some advantages of HDX-MS over HDX-NMR are that it requires only small quantities of protein sample, and that there is no strong limitation on the size of proteins that can be studied (Jaswal, 2013).

In HDX experiments, only the exchange rates of amide hydrogens (i.e., hydrogens attached to backbone nitrogens, referred to as amide nitrogens) are monitored (Engen et al., 2011); at least this represents what is most often assumed, in a slightly simplified view of the hydrogen exchange phenomenon. As a result, HDX experiments can generate at most one measurement per amino acid residue, for all amino acids of the protein, except for proline residues and for the N-terminus of the polypeptide chain (i.e., the first amino acid in the chain) because they do not possess an amide N–H group. In HDX-NMR experiments, results are acquired at the residue level (i.e., at the level of amide groups themselves), but obtaining a good coverage of the protein is very challenging. As explained in what follows, in HDX-MS experiments, results are most often acquired at the peptide level (i.e., deuterium uptake is measured for various proteolytic peptides extracted from the protein), and usually yield a good coverage of the protein. Note that, although we do not provide details on this, obtaining HDX-MS data at the residue level is feasible (Rand et al., 2009; Kan et al., 2013).

The hydrogen-exchange rate of a given amino acid can vary up to several orders of magnitude, depending on various conditions, such as solution pH and temperature (Brier and Engen, 2008). Even though this differs among amino acids, exchange rates are generally the lowest when pH is around 2.5 and temperature is around 0°C. The exchange rate of a residue in an unstructured peptide is only affected by its adjacent amino acids; this “intrinsic” exchange rate, denoted by k^{int}, can be predicted (Bai et al., 1993; Connelly et al., 1993). On the other hand, the exchange rate of a residue in a protein is influenced by additional factors, such as its solvent accessibility and the protein's structure; therefore, this experimentally-observed exchange rate, denoted by k^{obs}, is slower than k^{int} (Wei et al., 2013). To quantify the extent to which amide hydrogens are protected from being exchanged in a protein, one can define the *protection factor* of every amino acid i by $P_i = k_i^{\text{int}} / k_i^{\text{obs}}$. In HDX-NMR experiments, results are often reported as a list of (logarithms of) protection factors.

On the other hand, HDX-MS experiments produce richer information. A typical experiment starts by equilibrating a protein in H_2O at room temperature under physiological conditions (pH 7–8). Then, the protein is diluted with excess D_2O for the HDX to occur. At various time points, a small quantity of solution is sampled. The HDX reaction is quenched in the sample by adding acid to lower pH to 2.5, and by cooling it to 0°C. Proteins in the sample are then digested using acidic proteases (such as pepsin) that are active under quenching conditions. This proteolytic digestion generates numerous peptides, which are portions of the protein typically 6–20 amino acids in length. The sample is then introduced into a chromatography system, to separate the peptides and automatically send them for MS analysis. This analysis allows identifying the peptides generated by the proteolytic digestion and quantifying their deuterium uptake. As the digestion and MS analysis are repeated at various time points, HDX-MS experimental results are usually reported as a set of deuterium-uptake kinetic curves for various peptides (Huang and Chen, 2014).

A crucial technical aspect of HDX-MS experiments is known as *back-exchange*. This is the process by which the deuterium atoms incorporated by the peptides exchange back to hydrogens. This happens when the sample is prepared for MS analysis because all the required steps (quenching, enzymatic digestion, desalting, chromatographic separation) are performed in H_2O solution. On the one hand, back-exchange is beneficial because it enables fast-exchanging side-chain positions to revert to hydrogens, which greatly facilitates the MS identification of peptides by limiting mass changes to amide groups (Wei et al., 2013). On the other hand, back-exchange can become detrimental if slower-exchanging amide groups start reverting to hydrogens, which means losing the information generated by the experiment (Mayne, 2016). To mitigate this problem, all experimental steps have to be performed rapidly, at low temperature.

Unfortunately, back-exchange of amide groups cannot be totally avoided, which affects several aspects of HDX-MS experiments. First, depending on the kind of performed analysis, the measurements produced by the mass spectrometer might have to be corrected for back-exchange (Engen et al., 2011). Second, because terminal positions of a polypeptide chain are more susceptible to back-exchange than other positions, the analysis of peptide-level deuterium-uptake curves has to account for it. More specifically, if the HDX experienced by a given peptide is considered as the average HDX undergone by its amino acids (as done in Section 3.1), the first two amino acids in the chain have to be ignored (Konermann et al., 2011; Huang and Chen, 2014). Indeed, after digestion, the first amino acid of the peptide becomes an amine-terminus, therefore losing its deuterium; as a result, the second amino acid usually undergoes back-exchange as well (Mayne, 2016).

2.2. Hydrogen Exchange Estimated from Protein Structure

Numerous theoretical models have been suggested to formalize a relationship between local and/or global structural properties of a protein and the level of hydrogen exchange it undergoes locally. However, none of these models has yet been largely accepted by the scientific community. Several of them have also shown limitations in a recent evaluation study (Skinner et al., 2012b). In this section, we mention the ideas that prevailed in the early days of the research on hydrogen exchange mechanisms, and introduce various models proposed during the past 10 years.

Early attempts to connect hydrogen-exchange mechanisms with protein structure, in the 1970s, were based on accessibility or penetration models. A common view was that solvent-accessible hydrogens located at the protein's surface would exchange rapidly, and that buried hydrogens would exchange more slowly. In other words, protection from exchange was thought to be positively correlated with atom burial or, equivalently, negatively correlated with solvent penetration in the protein matrix. However, it is now well recognized that atom burial is not the primary factor in characterizing hydrogen exchange (Konermann et al., 2011; Skinner et al., 2012b). Indeed, hydrogen-bonded amide groups at the surface can exchange as slowly as deeply-buried amide groups. A variant of this early model of hydrogen exchange based on solvent penetration became popular in the 1980s: hydrogen exchange was thought to be positively correlated with solvent accessibility surface area (SASA). Although this correlation is in general relatively weak (Skinner et al., 2012b; Radou et al., 2014), it has been observed for surface loops of non-globular proteins (Truhlar et al., 2006). This model has been used in qualitative studies of hydrogen exchange (Petruk et al., 2013), sometimes including rigidity properties for increased accuracy (Sljoka and Wilson, 2013).

To explain the fact that even solvent-exposed hydrogens can exchange very slowly, several protein properties have been investigated. For example, there have been some attempts to show that hydrogen exchange is modulated by electrostatic effects on the relative acidity of amides (Anderson et al., 2008; Avbelj and Baldwin, 2009; Hernández et al., 2009; LeMaster et al., 2009). Although this appears to be true in specific cases, in general, no correlation can be expected between protection from hydrogen exchange and changes in relative acidity of amides evaluated via electrostatic calculations (Skinner et al., 2012b). On the other hand, participation in hydrogen bonds is usually recognized as a strong determinant of protection from hydrogen exchange (Skinner et al., 2012b). However, approaches that consider only hydrogen bonding to explain protection from exchange, such as those described in Ma and Nussinov (2011) and Park et al. (2015), are not expected to generalize well. Therefore, some attempts have been made to combine several factors, such as N–H coupling constants and residue fluctuation (Brand et al., 2007).

The most successful approaches to date have been those that combine packing density with various properties related to protein dynamics. On the one hand, some approaches, such as the COREX family of tools, have attempted to link hydrogen exchange to large segmental unfolding reactions (Hilser et al., 2006; Wrabl et al., 2011; Liu et al., 2012). However, a drawback of COREX is that it heavily relies on SASA for doing so. On the other hand, other approaches have attempted to link hydrogen exchange to local interactions (Wu et al., 2009; Gogonea et al., 2010; Craig et al., 2011). Among them, the approach we have adopted in our work relies on the combined evaluation of hydrogen bonding and packing density (Vendruscolo et al., 2003; Best and Vendruscolo, 2006; Gsponer et al., 2006; Kieseritzky et al., 2006; Radou et al., 2014). It is based on a phenomenological equation approximating hydrogen-exchange protection, which is detailed in Section 3.1. Of note, there has been an attempt to predict the coefficients of this phenomenological equation from a protein's amino acid sequence (Tartaglia et al., 2007). Other methods have similarly focused on estimating structural parameters related to hydrogen exchange, directly from protein sequence (Dovidchenko et al., 2009; Lobanov et al., 2013).

3. MATERIALS AND METHODS

3.1. Phenomenological Approximation of Hydrogen Exchange

As mentioned in Section 2.1, the levels of hydrogen exchange observed in different parts of a protein are known to be partly influenced by its local structure. Several theoretical models have been proposed to formalize a relationship between a protein's conformation and the corresponding hydrogen exchange (see Section 2.2). However, none of them benefits from a consensus of the scientific community, and several of them have shown limitations (Skinner et al., 2012b). Among these models, we chose the one that seemed the most promising, based on its performance in a recent comparative study (Skinner et al., 2012b) and on the number of publications in which it features (Vendruscolo et al., 2003; Best and Vendruscolo, 2006; Gsponer et al., 2006; Kieseritzky et al., 2006; Tartaglia et al., 2007; Radou et al., 2014).

The model we use to estimate hydrogen exchange from a protein's conformation relies on the definition of a phenomenological expression to approximate the protection factors (cf. Section 2.1) of the protein's residues (Vendruscolo et al., 2003). In this theoretical model, it is assumed that protection from hydrogen exchange results from the presence of hydrogen bonds involving amide groups and from the packing density of atoms around these amide groups. More precisely, the protection factor of residue i in conformation C, $P_i(C)$, is derived from the phenomenological expression

$$\ln P_i(C) = \beta^{\mathrm{h}} N_i^{\mathrm{h}}(C) + \beta^{\mathrm{c}} N_i^{\mathrm{c}}(C) \ , \tag{1}$$

where $N_i^{\mathrm{h}}(C)$ is the number of hydrogen bonds formed by the amide hydrogen of residue i, and $N_i^{\mathrm{c}}(C)$ is the number of so-called "atom contacts" (which is used to quantify packing density) involving residue i. Parameters β^{h} and β^{c} were estimated by fitting experimental hydrogen-exchange data from seven proteins, which lead to: $\beta^{\mathrm{h}} = 2$ and $\beta^{\mathrm{c}} = 0.35$ (Best and Vendruscolo, 2006).

Instead of being estimated from a single conformation, hydrogen exchange can also be estimated from a conformational ensemble. In that case, protection factors are computed as ensemble averages. Given a set of conformations, S, the protection factor of residue i with respect to S is derived from

$$\ln P_i(S) = \frac{1}{|S|} \sum_{C \in S} \ln P_i(C) \ . \tag{2}$$

The way hydrogen bonds and atom contacts are accounted for has changed over the years, following the evolution of the theoretical model (Vendruscolo et al., 2003; Best and Vendruscolo, 2006). Additionally, not all the details of the methodology have been published. Building on this model, we define hydrogen bonds and atom contacts in the following way:

- We only consider the hydrogen bonds maintaining secondary structure elements because they are more important than other hydrogen bonds in protecting amide groups from exchange. More specifically, only main-chain oxygens are considered as potential acceptors, when an amide nitrogen is regarded as potential donor. We count only the acceptor oxygens that are within a *cutoff* distance of 2.4 Å from the amide hydrogen. Additionally, when estimating $N_i^{\mathrm{h}}(C)$, oxygens from residues $i-2, \ldots, i+2$ are not considered as potential acceptors. This is justified by the fact that α-helices, 3_{10}-helices and β-sheets are formed by N−H···O=C hydrogen bonds involving residues that are at least three positions apart in the protein's sequence.
- The number of contacts, $N_i^{\mathrm{c}}(C)$, is defined as the number of heavy atoms (i.e., non-hydrogen atoms) in any residue, apart from residues $i-2, \ldots, i+2$, within a *cutoff* distance of 6.5 Å from the amide hydrogen of residue i. Note that these contacts are not restricted to secondary structure elements.

The residues' protection factors derived from Equation (1) can be directly compared to protection factors obtained from an HDX-NMR experiment. On the other hand, HDX-MS experiments produce deuterium-uptake curves of peptides extracted from a protein. Therefore, a similar kind of data has to be derived from the protein's structure to allow for a comparison with HDX-MS data. For that, we consider that the deuterium uptake of a residue follows pseudo-first-order kinetics (Brier and Engen, 2008; Konermann et al., 2011; Huang and Chen, 2014). Knowing that $P_i = k_i^{\text{int}} / k_i^{\text{obs}}$, the fraction of deuterium incorporated by residue i at time t can be expressed as

$$d_i(t) = 1 - \exp(-k_i^{\text{obs}}\, t) = 1 - \exp(-(k_i^{\text{int}}/P_i)\, t) \ . \qquad (3)$$

As k_i^{int} is known (Bai et al., 1993; Connelly et al., 1993), $d_i(t)$ can be derived from the protein's conformation by calculating P_i. The deuterium uptake of a peptide can be considered as an average over the residues it contains. Therefore, the fraction of deuterium incorporated by peptide j at time t is

$$D_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} d_i(t) \ , \qquad (4)$$

where n_j is the number of residues containing an exchangeable amide hydrogen in peptide j (Radou et al., 2014). Note that, in addition to the N-terminal amino acid and to prolines, we systematically exclude from the average the second amino acid (even if it contains an amide group) because of back-exchange (see Section 2.1) (Konermann et al., 2011; Huang and Chen, 2014). Using Equation (4), one can obtain deuterium-uptake curves for various peptides, from any protein conformation.

3.2. Goodness-of-Fit between Structurally-Derived and Experimental HDX Data

Using the HDX prediction model presented in Section 3.1, one can derive HDX data from a protein's conformation and compare it to the experimental HDX data. Then, assessing the goodness-of-fit between structurally-derived and experimentally-observed HDX data can be done as follows:

- When dealing with HDX-NMR data (i.e., protection factors of residues), one can obtain a histogram of differences by computing, for every residue i, the error $|\ln P_i^{\text{der}} - \ln P_i^{\text{obs}}|$, where P_i^{der} is the structurally-derived protection factor and P_i^{obs} is the experimentally-observed protection factor. This histogram can be aggregated into an average over all residues (as done in Section 4.1): $\frac{1}{n}\sum_{i=1}^{n} |\ln P_i^{\text{der}} - \ln P_i^{\text{obs}}|$, where n is the number of protein residues for which measurements have been obtained in the HDX-NMR experiment. Alternatively, one can compute the R^2 correlation coefficient between the series $\{\ln P_i^{\text{der}}\}_{i=1}^{n}$ and $\{\ln P_i^{\text{obs}}\}_{i=1}^{n}$ (as done in Section 4.2).
- With HDX-MS data (i.e., deuterium-uptake curves of peptides), one can obtain a histogram of differences (as done in Section 4.3) by computing, for every peptide j, the error $\sum_{t \in T} |D_j^{\text{der}}(t) - D_j^{\text{obs}}(t)|$, where T is the list of experimental time points, $D_j^{\text{der}}(t)$ is the structurally-derived deuterium uptake at time t, and $D_j^{\text{obs}}(t)$ is the experimentally-observed deuterium uptake at time t. This histogram can also be aggregated into an average difference over all peptides (as done in Section 4.3).

3.3. Conformation Providing the Best Fit to Experimental HDX Data

The question that remains is: which conformation should the HDX data be derived from to obtain a good fit to the experimentally-observed HDX data? Several studies have shown that conformations reported in the PDB (and more specifically crystal structures) do not provide good estimates for experimental HDX data (Radou et al., 2014; Devaurs et al., 2016). This can be explained by the very nature of HDX data: as it reflects the inherent flexibility of a molecule, in theory, it cannot be accurately predicted from a single conformation. Therefore, it was suggested that hydrogen exchange should be estimated from an ensemble of conformations extracted from an MD simulation, to account for the variability of a protein's structure (Best and Vendruscolo, 2006). Our previous study shows that this methodology also has limitations: better estimates of the experimental HDX data can sometimes be obtained from a single conformation extracted from a conformational ensemble produced by an MD simulation than from the whole ensemble (Devaurs et al., 2016). This shows that, in the context of the structural analysis of experimental HDX data, it is relevant to try and fit this data using a single conformation.

In this work, using computational methods that can sample protein conformations, we aim to obtain a single conformation that can help analyze the experimental HDX data collected for a protein. As PDB conformations produced by X-ray crystallography do not generally provide good estimates for experimental HDX data, they are usually not the best choice for a structural analysis of this HDX data. In spite of this, in our experiments, we systematically evaluate the goodness-of-fit achieved when comparing experimentally-observed HDX data against HDX data derived from a PDB conformation. This provides a baseline against which other methods can be compared. The two methods we evaluate in this study are MD simulations and coarse-grained conformational sampling.

3.3.1. MD Simulations

In this study, all MD simulations were performed with the GROMACS v4.6.5 package (Pronk et al., 2013) using the GROMOS96 (53a6) force field. A cubic box was defined with at least 9 Å of liquid layer around the protein (the exact dimensions were different for each protein), using SPC water model and periodic boundary conditions. An appropriate number of sodium (Na^+) and chloride (Cl^-) counter-ions were added to neutralize the system, with final concentration of 0.15 mol/L. The algorithms *v-rescale* ($\tau_t = 0.1$ ps) and *parrinello-rhaman* ($\tau_p = 2$ ps) were used for temperature and pressure coupling, respectively. Cutoff values of 1.2 nm were used both for van der Waals and Coulomb interactions, with Fast Particle-Mesh Ewald (PME) electrostatics. For all MD simulations, the production stage was preceded by (i) three steps of Energy Minimization (alternating steepest-descent and conjugate gradient) and (ii) eight steps of Equilibration. The Equilibration stage started with position restraints for all heavy atoms (5,000 $\mathrm{kJ^{-1}mol^{-1}nm^{-1}}$) and a temperature of 310 K, for a period of 300 ps, to allow for the formation of solvation layers. The temperature was then reduced to 280 K and the position restraints were gradually reduced. This process was followed by a gradual increase in temperature (up to 300 K). Together, these Equilibration steps represent the first 500 ps of each simulation. During the production stage, the system was held at constant temperature (300 K) without restraint. The MD simulations were run on various high-performance computers, using between 32 and 144 threads, depending on the size of the protein; the production stage lasted between 150 and 300 ns (additional protein-specific information is provided in Section 3.4). Then, we estimated HDX data as an average over the ensemble of conformations produced by a simulation. We also derived HDX data from every single conformation extracted from such a conformational ensemble.

3.3.2. Structured Intuitive Move Selector (SIMS)

In this paper, we propose a new methodology to obtain a better fit to experimental HDX data, using conformations produced by a coarse-grained conformational sampling approach. For that, we use a computational framework, called Structured Intuitive Move Selector (SIMS), that was developed to explore a protein's conformational space (Gipson et al., 2013). This framework integrates methods known as sampling-based motion-planning algorithms, initially proposed in the field of robotics to randomly explore high-dimensional spaces (Hsu et al., 1999; Şucan and Kavraki, 2010). Using these methods, exploring a protein's conformational space consists of incrementally building a graph whose nodes are conformations and whose edges represent potential transitions between them (Al-Bluwi et al., 2012; Gipson et al., 2012). SIMS follows a "coarse-grained" approach, similarly to MD-like methods using coarse-grained force fields (Davtyan et al., 2012), Monte-Carlo-based simulations (Sim et al., 2012; Boomsma et al., 2013), methods using elastic network models (López-Blanco and Chacón, 2016), or other robotics-inspired conformational sampling methods (Devaurs et al., 2013, 2015).

In SIMS, the exploration starts from a known conformation of the protein (usually, a crystal structure available in the PDB) and aims at producing new conformations by perturbing existing ones. Conformational sampling involves perturbations of the protein's structure, referred to as *protein moves*. These moves are common perturbation strategies, such as loop sampling, rigid-body motion (i.e., fix one loop's end and move the other end), random perturbation of backbone dihedral angles, and overall energy minimization. To implement these moves and calculate molecular energy, SIMS relies on the Rosetta modeling software (Das and Baker, 2008; Kaufmann et al., 2010). Additionally, SIMS involves an energy threshold to filter the conformations it generates. Note that, by varying this threshold, SIMS can be made more permissive than a typical MD simulation, with respect to the energy of the conformations it generates.

SIMS involves an internal-coordinate representation of proteins in which bond lengths and bond angles are assumed to be constant. Additionally, taking into account the planarity of peptide bonds, the associated torsion angles are restricted to their trans conformation (i.e., $\omega = 180°$). In SIMS, a protein's conformation is represented by a vector of backbone (φ, ψ) dihedral angles. Side chains are not explicitly modeled in a conformation, but they are automatically optimized by Rosetta when a move is performed. As a result, a protein composed of $N + 1$ residues is modeled with $2N$ degrees of freedom. Such a coarse-grained model has long been shown to provide a good approximation of a protein's behavior (Levitt, 1976).

In SIMS, proteins are decomposed into *fragments* on which moves are applied. Fragments are specific sets of residues that can be defined automatically, based on secondary structure, or that can be chosen by a domain expert. A fragment can be a protein domain, a single secondary structure element (or several of them), a single residue, or several (non-necessarily contiguous) residues. Using these fragments, one can favor the sampling of specific regions of the protein during conformational exploration. Indeed, based on how flexible some regions are expected to be, fragments are assigned probabilities to be perturbed during conformational sampling (Gipson et al., 2013). These probabilities can reflect available expert knowledge of the protein, reported experimental data (such as B factors) or predicted data resulting from a computational analysis (Fox et al., 2011). In some of the experiments presented here, we use discrepancies between experimentally-observed and structurally-derived HDX data to define these probabilities and therefore guide conformational exploration.

Our experimental methodology can be summarized as follows: First, we use SIMS to perform a conformational exploration starting from the crystal structure of a protein, without using any sampling bias. From the ensemble of conformations generated by SIMS, we determine which conformation provides the best estimates of the HDX data, using Equations (1)–(4). If a good fit is obtained, no additional run of SIMS is performed. If the goodness-of-fit is too low, we run SIMS again, using the largest discrepancies between experimentally-observed and structurally-derived HDX data as a sampling bias: the protein regions where these discrepancies are the largest are assigned higher probabilities to be sampled. We repeat this process a given number of times, or stop when a conformation providing a good fit to the HDX data is obtained.

A single run of SIMS lasted 24 h and was performed on four threads of a 3.6 GHz Intel i7-4790 quad-core CPU. For small proteins, we ran SIMS only once, but for the largest one, we ran SIMS five consecutive times (see Section 4 for more details). For comparison, if the aforementioned MD simulations were run on the same computer, 24 h of computation would yield only 5–15 ns of simulation, therefore requiring days to weeks for a whole simulation, depending on protein size.

3.4. Studied Proteins and Experimental HDX Data

First, we use two small proteins (CI2 and Im7) to illustrate the concepts involved in our methodology. As they have been extensively studied, they represent useful benchmarks. Then, we analyze two medium-size proteins (SN and C3d) that represent more challenging targets for our methodology. Note that we consider two kinds of HDX data: HDX-NMR for CI2, Im7 and SN; and HDX-MS for C3d.

3.4.1. Chymotrypsin Inhibitor 2 (CI2)

We consider a truncated form of chymotrypsin inhibitor 2 (CI2) composed of 64 amino acids (PDB 1TM1), where residue 1 corresponds to residue 20 of the full protein. The main secondary structure elements of CI2 are the following: residues Val13 to Asp23 form an α-helix; residues Gln28 to Pro33 and residues Arg46 to Val51 form two β-strands. As a simple system for folding studies, CI2 has been the subject of several HDX-NMR experiments (Itzhaki et al., 1997; Neira et al., 1997). Protection factors for more than half of CI2's residues have been reported. However, as done in other studies (Best and Vendruscolo, 2006), we only use the protection factors associated with local hydrogen-exchange mechanisms characteristic of CI2's native state. Therefore, we only consider the following 14 residues (whose protection factors are given in parentheses): Leu8 (8.1), Val9 (9.9), Val13 (7.2), Ala16 (7.1), Lys17 (6.6), Lys18 (8.2), Gln22 (9.5), Ala27 (6.7), Gln28 (8.2), Asp52 (8.5), Asn56 (8.4), Ala58 (9), Gln59 (10.5), Val63 (7.4). Note that these protection factors are given as $\ln P$, based on published exchange rates (Itzhaki et al., 1997). Three trajectories of CI2 were obtained by running MD simulations with a 150 ns production stage.

3.4.2. Bacterial Immunity Protein Im7

The bacterial immunity protein Im7 is a single-domain α protein composed of 86 residues (PDB 1AYI). Im7's native state comprises four α-helices: residues Glu12 to Lys24 (I), residues Asp32 to Thr45 (II), residues Thr51 to Tyr56 (III), and residues Glu66 to Asn79 (IV). Helices I and II form an N-terminal helical hairpin, and helix IV is located along the open end of this hairpin. Im7 has been shown to fold through an on-pathway intermediate whose structure is significantly different from that of its native state (Gorski et al., 2004). In this non-native state, helices I, II and IV are conserved, but helix III is not formed. A computational analysis of this intermediate state has shown that helices I, II, and IV are not organized as they are in the native state (Gsponer et al., 2006). This analysis was based on protection factors of 26 residues, obtained via an HDX-NMR experiment aimed at characterizing Im7's folding intermediate (Gorski et al., 2004). Here, we consider the same residues (whose protection factors are given in parentheses): Asp9 (8), Tyr10 (9.2), Thr11 (10.2), Val16 (11.5), Gln17 (11.6), Leu18 (11.2), Glu21 (8.5), Glu23 (6), Leu37 (6.1), Leu38 (7.5), Phe41 (6), Val42 (10.3), Leu53 (3.5), Ile54 (3.6), Tyr55 (4.4), Tyr56 (5.6), Gly67 (8.1), Val69 (8.8), Ile72 (9), Lys73 (9.5), Glu74 (9), Trp75 (8.8), Arg76 (9.9), Ala77 (9.8), Ala78 (8), Lys85 (6.8). These protection factors are given as $\ln P$, based on published exchange rates (Gorski et al., 2004). Three trajectories were obtained by running MD simulations with a 200 ns production stage.

3.4.3. Staphylococcal Nuclease (SN)

Micrococcal nuclease, or Staphylococcal nuclease (SN), is a mixed α/β protein composed of 149 amino acids organized in two domains (PDB 1SNO). The first domain (residues 1–98) belongs to the oligonucleotide/oligosaccharide-binding-fold (or OB-fold) superfamily. It consists of a five-stranded β-barrel with Greek key topology, capped by an α-helix (residues Gly55 to Glu67) located between the third and fourth strands. The five β-strands are: residues Lys9 to Ala17, residues Thr22 to Tyr27, residues Gln30 to Leu36, residues Ile72 to Phe76, residues Gly88 to Ala94. The second domain (residues 99–149) contains two α-helices: residues Val99 to Arg105, and residues Glu122 to Lys134. SN also contains two minor β-strands. HDX-NMR experiments have been performed on a double mutant of SN with similar structure but increased stability, to characterize its native state (Skinner et al., 2012b). This allowed measuring hydrogen-exchange rates for most residues and deriving corresponding protection factors. Here, we use 100 of these protection factors: residues of the N and C terminals that are missing from the crystal structure (PDB 1SNO) are not considered. Note that protection factors were reported as $\log_{10} P$, instead of $\ln P$ (Skinner et al., 2012b).

3.4.4. Complement Protein C3d

C3d is a fragment of the complement component C3 (Nagar et al., 1998; Hammel et al., 2007). It is a single-domain α protein composed of 297 residues (PDB 2GOX), where residue 1 corresponds to residue 991 of the full C3 molecule. C3d contains twelve α-helices and five 3_{10}-helices that are organized into an α-α barrel where most consecutive helices alternate between the inside and the outside. Based on previous notations (Nagar et al., 1998), the core of the barrel consists of the following six parallel α-helices: α_1 (residues Glu22 to Thr41), α_3 (residues Thr86 to Leu102), α_5 (residues Lys149 to Ala164), α_8 (residues Ser196 to Met209), α_{10} (residues Gln236 to Leu253), and α_{12} (residues Ser278 to Asp295). It is surrounded by another set of six parallel helices (running anti-parallel to those of the core) comprising one 3_{10}-helix, T_1 (residues Ala7 to Leu13), and five α-helices: α_2 (residues Leu49 to Arg70), α_4 (residues Ser107 to Lys121), α_7 (residues Ser174 to Asn189), α_9 (residues Pro215 to Thr223), and α_{11} (residues Phe256 to Gln269). In previous work, we performed an HDX-MS experiment and several MD simulations on C3d, to characterize its native state (Devaurs et al., 2016). In this paper, as in our previous study, we use the deuterium-uptake data obtained for 81 peptides extracted from C3d.

4. RESULTS

We now report the results we have obtained for the four proteins introduced in Section 3.4. First, a comparative analysis of CI2 and Im7 sheds light on the issues encountered when trying to fit experimental HDX data with the different methods presented in Section 3.3. Then, we examine two medium-size proteins: SN and C3d.

4.1. Chymotrypsin Inhibitor 2 vs. Bacterial Immunity Protein Im7

Our first set of results aims at highlighting differences and similarities that exist between two possible scenarios: (i) the case where the HDX data and the crystal structure describe the same state of the protein, and (ii) the case where they describe two different states. As mentioned in Section 3.4.1, the HDX-NMR data (i.e., protection factors of residues) obtained for CI2 is characteristic of its native state, whose structure has been described (PDB 1TM1). On the contrary, the HDX-NMR data gathered for Im7 is known to characterize a non-native folding intermediate (see Section 3.4.2) that is structurally different from Im7's native state (PDB 1AYI).

The comparison between these two scenarios is illustrated in **Figure 1**. The native conformations of CI2 and Im7, as reported in the PDB, are depicted in green using the ribbon model. The bar charts show that the HDX data derived from the PDB conformations (i.e., the crystal structures reported in the PDB) using Equation (1) does not match well the experimentally-observed HDX data: the average difference between structurally-derived and experimentally-observed protection factors (see Section 3.2) is close to 3. Although this is not surprising in the case of Im7 (because the HDX data and the crystal structure describe different states), it is important to note that the HDX estimates are equally bad in the case of CI2 (even though the HDX data and the crystal structure describe the same state).

For both CI2 and Im7, we performed three MD simulations. We observe that, as suggested in Radou et al. (2014), deriving HDX data from the ensemble of conformations extracted from each MD simulation leads to a better fit to experimentally-observed HDX data than if the PDB conformation is used. However, it also appears that the best fit is usually obtained with a single conformation selected within these MD ensembles, which confirms our previous results on C3d (Devaurs et al., 2016). The bar charts in **Figure 1** show the differences between the experimentally-observed and structurally-derived protection factors, when deriving these protection factors from the MD conformation providing the best fit. It is clear that using this MD conformation yields a better fit to the experimental data than using the PDB conformation, but not drastically. In the case of Im7, the limited improvement was expected: our MD simulations were meant to sample the native state; they were not long enough to observe a transition to the folding intermediate. Even in the case of CI2, we will show that better results can be obtained.

We used SIMS to sample the conformational space of CI2 and Im7, starting the exploration from their PDB conformations, without any bias. From the sets of conformations generated by SIMS, we extracted the conformation yielding the best fit between structurally-derived and experimentally-observed HDX data. The bar charts in **Figure 1** show differences between experimentally-observed and structurally-derived protection factors, when deriving them from the SIMS conformation with the best fit. This conformation yields a significantly better fit to experimental HDX data than the PDB and MD conformations. The SIMS conformations for CI2 and Im7 are depicted in red using the ribbon model in **Figure 1**. In the case of CI2, the SIMS conformation is very similar to the PDB conformation: differences occur mostly in side-chain positions and not in backbone structure. This was expected because the HDX data and the crystal structure describe the same state. This also highlights the strong impact that even small structural variations can have when estimating protection factors with Equation (1). In the case of Im7, the SIMS conformation providing the best fit to the experimental HDX data is significantly different from the PDB conformation, which confirms that the HDX data and the crystal structure describe different states.

4.2. Staphylococcal Nuclease

A recent evaluation study of various models for deriving hydrogen exchange from a protein's structure involved HDX-NMR data gathered for SN (see Section 3.4.3) (Skinner et al., 2012b). The study concludes that, at least for SN, none of the evaluated models can produce HDX data that fits well the experimentally-observed HDX data. The best results are achieved by the model based on Equation (1), with a correlation coefficient $R^2 = 0.51$ between the structurally-derived and experimentally-observed protection factors. That study follows the methodology in Best and Vendruscolo (2006), estimating the protection factors of SN's residues using an ensemble of conformations extracted from an MD simulation. However, that study does not consider estimating HDX data from the PDB conformation alone, using Equation (1). Interestingly, using this PDB conformation, we obtained a correlation coefficient $R^2 = 0.69$ between the structurally-derived and experimentally-observed protection factors (see Section 3.2). Note that better results can be achieved with our novel methodology.

We used SIMS to explore the conformational space of SN, starting from its PDB conformation, without introducing any bias. From the ensemble of conformations generated by SIMS, we extracted the conformation providing estimates of protection factors that best fit the experimental HDX data. This yields a correlation coefficient $R^2 = 0.78$ between the structurally-derived and experimentally-observed protection factors, as shown in **Figure 2**. Importantly, the SIMS conformation is very similar to the PDB conformation: only small structural differences are observed at the backbone level (see **Figure 2**). This confirms that the HDX data and the crystal structure both describe SN's native state.

4.3. Complement Protein C3d

In previous work (Devaurs et al., 2016), we performed an HDX-MS experiment on C3d and obtained deuterium-uptake curves for 86 peptides. As in that previous study, we restrict the current analysis to the 81 peptides whose data is the most reliable. This HDX data is expected to describe the native state of C3d when present alone in solution. However, once again, deuterium-uptake curves derived from the PDB conformation of C3d (PDB 2GOX) using Equations (1)–(4) do not fit well the experimental data (see **Figure 3**). The average difference between the experimentally-observed and structurally-derived deuterium-uptake curves across all peptides is 1.23 (see Section 3.2). Discrepancies are especially significant in the region of C3d comprising residues Met191 to Ala242. As shown in Devaurs et al. (2016), this is not due to structural differences between the native state of C3d and the conformation observed during the HDX-MS experiment, but rather to the limitations of predicting HDX data using crystal structures.

We carried out three MD simulations to sample the variability of C3d's native state (Devaurs et al., 2016). Using the ensemble

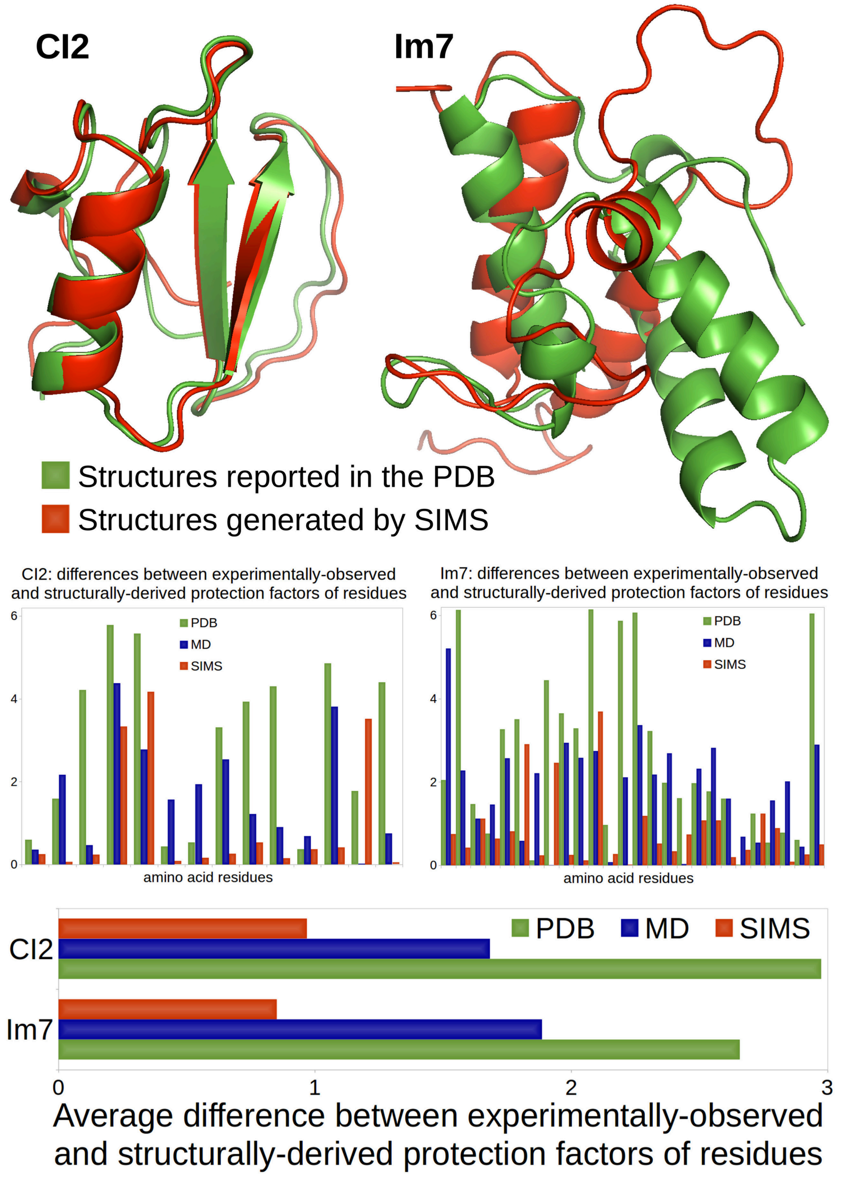


FIGURE 1 | Comparison of a case where HDX data and crystal structure describe the same state (CI2) and a case where they describe different states (Im7). CI2's HDX data is characteristic of its native state, described in the PDB. Im7's HDX data characterizes a non-native folding intermediate that structurally differs from the native state reported in the PDB. All conformations are represented using the ribbon model: conformations reported in the PDB are colored in green; conformations produced by SIMS that provide the best estimates of the experimental HDX data are colored in red. The bar chart at the bottom shows the average difference (across residues) between experimentally-observed and structurally-derived HDX data (i.e., protection factors), when deriving this data using conformations in the PDB (green), conformations produced by MD which best fit the HDX data (blue), or conformations generated by SIMS which best fit the HDX data (red). The bar charts in the middle show these differences for all residues separately.

of conformations extracted from each simulation allows deriving deuterium-uptake curves of peptides that fit the experimental data better than when using the PDB conformation. However, an important conclusion of our previous study is that: using a single conformation extracted from these MD ensembles produces even better results (Devaurs et al., 2016). The conformation providing the estimates of deuterium-uptake curves that best fit the experimental HDX data is referred to as the MD conformation. It yields a decrease in the average difference (0.89) between structurally-derived and experimentally-observed HDX data. Despite the improvement in goodness-of-fit, large discrepancies remain (see **Figure 3**), especially in the region [Met191-Ala242] of C3d (Devaurs et al., 2016).

To sample C3d's conformational space more extensively, we carried out the following iterative process with SIMS: using the PDB conformation as input, we ran SIMS once without

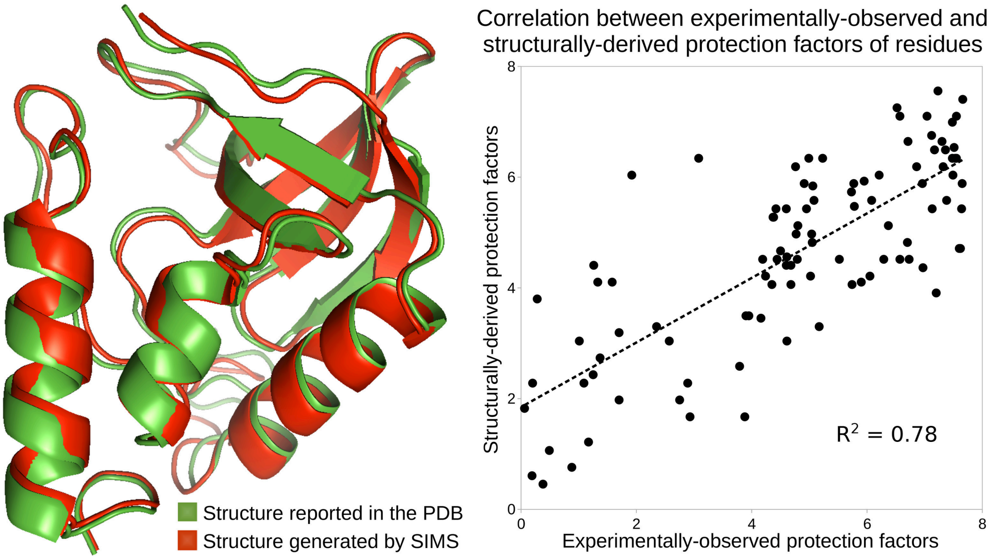


FIGURE 2 | Analysis of the native state of SN. Conformations of SN are depicted using the ribbon model: the conformation reported in the PDB, in green, and the conformation generated by SIMS which provides estimates of protection factors that best fit the experimental HDX data, in red. The plot shows the correlation between the experimentally-observed HDX data and the HDX data derived from the SIMS conformation. The correlation coefficient is 0.78.

introducing any bias; then, we ran SIMS four times, using the discrepancies between structurally-derived and experimentally-observed HDX data as a sampling bias. This bias is introduced in the following way: at the end of each run, we select the conformation generated by SIMS providing estimates of deuterium-uptake curves that best fit the experimental HDX data, and we determine the regions of C3d where discrepancies are the largest; then, in the following run, these regions are assigned higher probabilities to be sampled (cf. Section 3.3.2). This SIMS-based iterative process generated a conformation providing estimates of deuterium-uptake curves that fit well the experimental HDX data (see **Figure 3**). Using this SIMS conformation, the average difference between the experimentally-observed and structurally-derived deuterium-uptake curves across all peptides decreases to 0.6. Importantly, the SIMS conformation is very similar to the PDB conformation: all the helices forming the α-α barrel are conserved; only two short helices have unfolded. The α-α barrel of the SIMS conformation (radius of gyration: 19 Å) is only slightly wider than the α-α barrel of the PDB conformation (radius of gyration: 18 Å). This confirms that the HDX data and the crystal structure both describe C3d's native state. It also confirms that the native state of C3d is relatively stable, with little flexibility recorded by the HDX-MS experiment. Finally, note that using the SIMS conformation can provide an improved structural analysis of C3d, by refining the HDX data from the peptide level to the residue level (Devaurs et al., 2016).

5. DISCUSSION

We chose to include in our methodology the phenomenological approximation of hydrogen-exchange protection, as expressed by Equation (1), because it seemed to be the most promising HDX prediction model. Indeed, it performed best at predicting experimental HDX data, when compared to several other models (Skinner et al., 2012b). Even though the goodness-of-fit achieved with this model was not impressive (Skinner et al., 2012b), our work demonstrates how it can be used successfully.

Our results clearly show that using the conformation of a protein as reported in the PDB does not provide good estimates of experimental HDX data. This confirms what was observed in previous similar studies (Radou et al., 2014; Devaurs et al., 2016). This was also indirectly acknowledged when this HDX prediction model was first proposed (Vendruscolo et al., 2003). In an attempt to consider structural dynamics, it was suggested that HDX data should be derived as an average over an ensemble of conformations produced by an MD simulation (Best and Vendruscolo, 2006).

We have indeed observed that computing an average over an MD ensemble provides better estimates of experimental HDX data than using a single PDB conformation. However, as shown in our previous work (Devaurs et al., 2016), this study confirms that using a single conformation carefully extracted from the MD ensemble usually provides even better estimates. In other words, the MD conformation that provides the best estimates within the MD ensemble performs generally better than the whole ensemble. Note that we do not claim that this MD conformation constitutes a better representation of a protein's state than a PDB conformation or an MD ensemble. In theory, the best estimates for experimental HDX data would be obtained by computing an average over an ensemble of conformations best representing a protein's state and its inherent flexibility. However, in the same way as estimates derived from two similar conformations can significantly differ, estimates derived from two similar ensembles

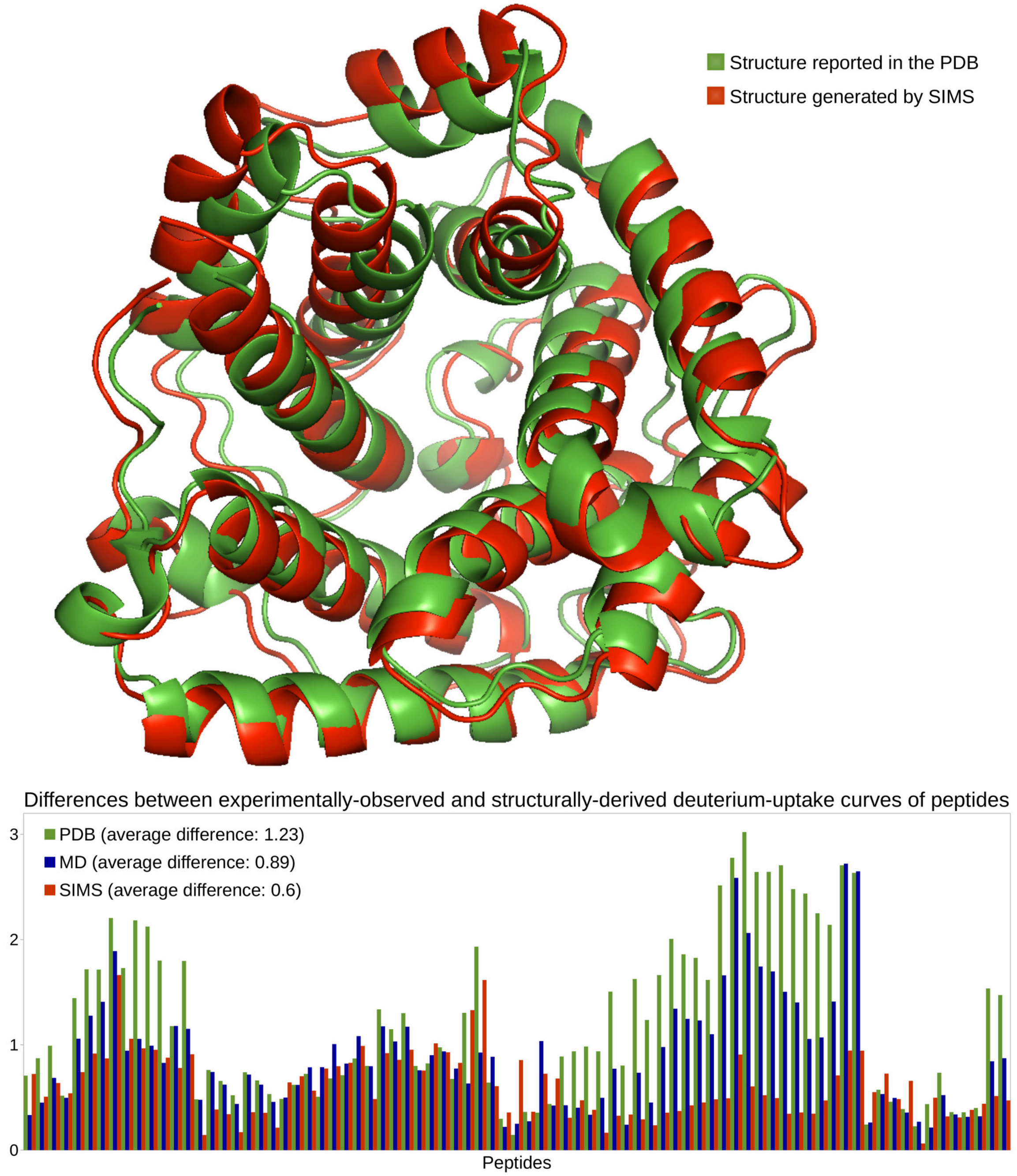


FIGURE 3 | Analysis of the native state of C3d. Conformations of C3d are depicted using the ribbon model: the conformation reported in the PDB, in green, and the conformation generated by SIMS which provides estimates of deuterium-uptake curves that best fit the experimental HDX data, in red. The plot shows differences between the experimentally-observed and structurally-derived deuterium-uptake curves, for all peptides, when deriving this data from the PDB conformation (green), the MD conformation (blue) or the SIMS conformation (red). The legend also includes the average differences across all peptides.

can be very different. In practice, it is thus more convenient to generate many conformations and select the one providing the best estimates than to find the best conformational ensemble.

The fact that numerous conformations have to be generated in order to obtain good estimates of experimental HDX data, and that a PDB conformation is not enough, is also linked to weaknesses of the HDX prediction model based on Equation (1). The first limitation of this model is its lack of robustness: it is very sensitive to small variations in the protein structure. As well illustrated by the case of CI2, two conformations that are very similar at the backbone level and present differences only in their side-chain conformations can produce very different

HDX estimates. The second limitation of this model is that it only partially reflects the mechanisms underlying hydrogen exchange. For example, it does not consider any dynamic aspect of proteins. Therefore, it could be interesting to develop a more accurate model by accounting for additional structural and dynamic properties of proteins (Skinner et al., 2012a). Since such a model has not been proposed yet, we believe it is best to compensate for the weaknesses of the current model by performing conformational sampling.

6. CONCLUSION

When performing a structural analysis of HDX data collected for a protein, a premise to an accurate analysis is to use a conformation that matches this data. Several studies, including ours, show that crystal structures reported in the PDB are not a good choice because they often provide bad estimates of experimental HDX data. Because HDX data reflects the inherent flexibility of a protein, a conformational ensemble should ideally provide better estimates than a single conformation. However, our work has shown that this is not always the case with a conformational ensemble produced by an MD simulation. Therefore, it is perfectly justified to try and fit experimental HDX data using a single conformation. In this paper, we have shown that this can be done using a coarse-grained conformational sampling tool to explore a protein's conformational space. The specific tool we used, called SIMS, yields a conformational ensemble from which one can extract a conformation providing a good fit to the experimental HDX data. Note that we do not claim that a conformation produced by SIMS is a better representation of a protein's state than its crystal structure. Besides the improved accuracy, another advantage of using SIMS is its efficiency: a conformation providing a good fit to experimental HDX data can be obtained at a fraction of the computational cost of running a traditional MD simulation. Finally, we believe that other conformational sampling methods could produce similar results, in terms of accuracy and efficiency. The achievement of our study mostly consists of revealing the technicalities that must be addressed for such methods to be successful.

Our methodology relies on the use of an HDX prediction model defining how to derive HDX data from a protein's structure. This model is based on a phenomenological approximation of the protection factors of a protein's residues. Despite its limitations, this model enables our methodology to successfully produce a conformation fitting the experimental HDX data. Another interesting benefit of this model is that, besides the validation of experimental HDX data, in the case of HDX-MS experiments, it offers the possibility to refine the HDX data from the peptide level to the residue level (Radou et al., 2014; Devaurs et al., 2016). This has the potential to enhance applications of the HDX-MS technique (Pirrone et al., 2015).

As part of our future work, we intend to apply our methodology to larger proteins, to evaluate its scalability. Since coarse-grained conformational sampling scales better than MD, we expect our methodology to be even more beneficial with large proteins. Additionally, we plan to investigate several useful applications of this work. First, as demonstrated with Im7, our method can be used to obtain a structural model of a non-native state of a protein when only its native state is described in the PDB and only HDX data is available for this non-native state. Second, although we applied our method only to cases where the experimental HDX data was expected to characterize a single protein conformation because a single conformer was assumed to be present in solution, it could be applied to more complex cases, where several conformers are involved. Indeed, if structurally-derived HDX data better fits experimentally-observed HDX data when deriving it from a small set of structurally-different conformations (i.e., two or three, or a handful of conformations) than when deriving it from a single conformation, we can suspect that several protein conformers are present together in solution.

AUTHOR CONTRIBUTIONS

DR, JL, LK, MM suggested the initial idea behind this work. DD conceived the study, including the methodological choices and the selection of proteins, and implemented the scripts required for the analysis. MP performed the HDX-MS experiment. DA performed the MD simulations. DD carried out the analysis of the results and wrote the paper. DA, DR, JL, LK, MM, MP reviewed and commented on the paper.

FUNDING

This work was supported in part by the National Science Foundation under Grant CCF 1423304, as well as the National Institutes of Health under Grants R21CA209941, AI068730, and AI030040. Computational simulations were run on equipment supported in part by the Data Analysis and Visualization Cyberinfrastructure funded by NSF under Grant OCI 0959097, on equipment supported by the Cyberinfrastructure for Computational Research funded by NSF under Grant CNS 0821727, as well as on equipment supported in part by the Big-Data Private-Cloud Research Cyberinfrastructure MRI-award funded by NSF under grant CNS 1338099 and by Rice University.

ACKNOWLEDGMENTS

The authors are thankful to Emanuele Paci and Gaël Radou for providing information on their implementation of the HDX prediction model based on the phenomenological approximation of protection factors. All protein structures are depicted by images produced with the PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

REFERENCES

Al-Bluwi, I., Siméon, T., and Cortés, J. (2012). Motion planning algorithms for molecular simulations: a survey. *Comput. Sci. Rev.* 6, 125–143. doi: 10.1016/j.cosrev.2012.07.002

Anderson, J. S., Hernández, G., and LeMaster, D. M. (2008). A billion-fold range in acidity for the solvent-exposed amides of *Pyrococcus furiosus* rubredoxin. *Biochemistry* 47, 6178–6188. doi: 10.1021/bi800284y

Avbelj, F., and Baldwin, R. L. (2009). Origin of the change in solvation enthalpy of the peptide group when neighboring peptide groups are added. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3137–3141. doi: 10.1073/pnas.0813018106

Bai, Y., Milne, J. S., Mayne, L., and Englander, S. W. (1993). Primary structure effects on peptide group hydrogen exchange. *Proteins* 17, 75–86. doi: 10.1002/prot.340170110

Best, R. B., and Vendruscolo, M. (2006). Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. *Structure* 14, 97–106. doi: 10.1016/j.str.2005.09.012

Boomsma, W., Frellsen, J., Harder, T., Bottaro, S., Johansson, K. E., Tian, P., et al. (2013). PHAISTOS: a framework for Markov chain Monte Carlo simulation and inference of protein structure. *J. Comput. Chem.* 34, 1697–1705. doi: 10.1002/jcc.23292

Brand, T., Cabrita, E. J., Morris, G. A., Günther, R., Hofmann, H.-J., and Berger, S. (2007). Residue-specific NH exchange rates studied by NMR diffusion experiments. *J. Mag. Res.* 187, 97–104. doi: 10.1016/j.jmr.2007.03.021

Brier, S., and Engen, J. R. (2008). "Hydrogen exchange mass spectrometry: principles and capabilities," in *Mass Spectrometry Analysis for Protein-Protein Interactions and Dynamics*, ed M. Chance (Hoboken, NJ: John Wiley & Sons, Inc.), 11–43.

Claesen, J., and Burzykowski, T. (2016). Computational methods and challenges in hydrogen/deuterium exchange mass spectrometry. *Mass Spectrom. Rev.* doi: 10.1002/mas.21519. [Epub ahead of print].

Connelly, G. P., Bai, Y., Jeng, M.-F., and Englander, S. W. (1993). Isotope effects in peptide group hydrogen exchange. *Proteins* 17, 87–92. doi: 10.1002/prot.340170111

Craig, P. O., Lätzer, J., Weinkam, P., Hoffman, R. M., Ferreiro, D. U., Komives, E. A., et al. (2011). Prediction of native-state hydrogen exchange from perfectly funneled energy landscapes. *J. Am. Chem. Soc.* 133, 17463–17472. doi: 10.1021/ja207506z

Das, R., and Baker, D. (2008). Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* 77, 363–382. doi: 10.1146/annurev.biochem.77.062906.171838

Davtyan, A., Schafer, N. P., Zheng, W., Clementi, C., Wolynes, P. G., and Papoian, G. A. (2012). AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* 116, 8494–8503. doi: 10.1021/jp212541y

Deng, B., Lento, C., and Wilson, D. J. (2016). Hydrogen deuterium exchange mass spectrometry in biopharmaceutical discovery and development – A review. *Anal. Chim. Acta* 940, 8–20. doi: 10.1016/j.aca.2016.08.006

Devaurs, D., Bouard, L., Vaisset, M., Zanon, C., Al-Bluwi, I., Iehl, R., et al. (2013). MoMA-LigPath: a web server to simulate protein-ligand unbinding. *Nucleic Acids Res.* 41, 297–302. doi: 10.1093/nar/gkt380

Devaurs, D., Molloy, K., Vaisset, M., Shehu, A., Siméon, T., and Cortés, J. (2015). Characterizing energy landscapes of peptides using a combination of stochastic algorithms. *IEEE Trans. Nanobiosci.* 14, 545–552. doi: 10.1109/TNB.2015.2424597

Devaurs, D., Papanastasiou, M., Antunes, D. A., Abella, J. R., Moll, M., Ricklin, D., et al. (2016). "Native state of complement protein C3d analyzed via hydrogen exchange and conformational sampling," in *Proceedings of International Conference on Intelligent Biology and Medicine (ICIBM)* (Houston, TX).

Dovidchenko, N. V., Lobanov, M. Y., Garbuzynskiy, S. O., and Galzitskaya, O. V. (2009). Prediction of amino acid residues protected from hydrogen-deuterium exchange in a protein chain. *Biochemistry (Moscow)* 74, 888–897. doi: 10.1134/S0006297909080100

Engen, J. R., Wales, T. E., and Shi, X. (2011). "Hydrogen exchange mass spectrometry for conformational analysis of proteins," in *Encyclopedia of Analytical Chemistry*, ed R. Meyers (Hoboken, NJ: John Wiley & Sons, Ltd.).

Englander, S. W., Mayne, L., Bai, Y., and Sosnick, T. R. (1997). Hydrogen exchange: the modern legacy of Linderstrøm-Lang. *Protein Sci.* 6, 1101–1109. doi: 10.1002/pro.5560060517

Fox, N., Jagodzinski, F., Li, Y., and Streinu, I. (2011). KINARI-Web: a server for protein rigidity analysis. *Nucl. Acids Res.* 39(Suppl. 2):W177–W183. doi: 10.1093/nar/gkr482

Gipson, B., Hsu, D., Kavraki, L. E., and Latombe, J.-C. (2012). Computational models of protein kinematics and dynamics: beyond simulation. *Annu. Rev. Anal. Chem.* 5, 273–291. doi: 10.1146/annurev-anchem-062011-143024

Gipson, B., Moll, M., and Kavraki, L. E. (2013). SIMS: a hybrid method for rapid conformational analysis. *PLoS ONE* 8:e68826. doi: 10.1371/journal.pone.0068826

Gogonea, V., Wu, Z., Lee, X., Pipich, V., Li, X., Ioffe, A. I., et al. (2010). Congruency between biophysical data from multiple platforms and molecular dynamics simulation of the double-super helix model of nascent high-density lipoprotein. *Biochemistry* 49, 7323–7343. doi: 10.1021/bi100588a

Gorski, S. A., Le Duff, C. S., Capaldi, A. P., Kalverda, A. P., Beddard, G. S., Moore, G. R., et al. (2004). Equilibrium hydrogen exchange reveals extensive hydrogen bonded secondary structure in the on-pathway intermediate of Im7. *J. Mol. Biol.* 337, 183–193. doi: 10.1016/j.jmb.2004.01.004

Gsponer, J., Hopearuoho, H., Whittaker, S. B.-M., Spence, G. R., Moore, G. R., Paci, E., et al. (2006). Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7. *Proc. Natl. Acad. Sci. U.S.A.* 103, 99–104. doi: 10.1073/pnas.0508667102

Hammel, M., Sfyroera, G., Ricklin, D., Magotti, P., Lambris, J. D., and Geisbrecht, B. V. (2007). A structural basis for complement inhibition by *Staphylococcus aureus*. *Nat. Immunol.* 8, 430–437. doi: 10.1038/ni1450

Harrison, R. A., and Engen, J. R. (2016). Conformational insight into multi-protein signaling assemblies by hydrogen–deuterium exchange mass spectrometry. *Curr. Opin. Struct. Biol.* 41, 187–193. doi: 10.1016/j.sbi.2016.08.003

Hernández, G., Anderson, J. S., and LeMaster, D. M. (2009). Polarization and polarizability assessed by protein amide acidity. *Biochemistry* 48, 6482–6494. doi: 10.1021/bi900526z

Hilser, V. J., García-Moreno, B., Oas, T. G., Kapp, G., and Whitten, S. T. (2006). A statistical thermodynamic model of the protein ensemble. *Chem. Rev.* 106, 1545–1558. doi: 10.1021/cr040423+

Hsu, D., Latombe, J.-C., and Motwani, R. (1999). Path planning in expansive configuration spaces. *Int. J. Comput. Geom. Appl.* 9, 495–512. doi: 10.1142/S0218195999000285

Huang, R. Y.-C., and Chen, G. (2014). Higher order structure characterization of protein therapeutics by hydrogen/deuterium exchange mass spectrometry. *Anal. Bioanal. Chem.* 406, 6541–6558. doi: 10.1007/s00216-014-7924-3

Itzhaki, L. S., Neira, J. L., and Fersht, A. R. (1997). Hydrogen exchange in chymotrypsin inhibitor 2 probed by denaturants and temperature. *J. Mol. Biol.* 270, 89–98. doi: 10.1006/jmbi.1997.1049

Jaswal, S. S. (2013). Biological insights from hydrogen exchange mass spectrometry. *Biochim. Biophys. Acta* 1834, 1188–1201. doi: 10.1016/j.bbapap.2012.10.011

Kan, Z.-Y., Walters, B. T., Mayne, L., and Englander, S. W. (2013). Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16438–16443. doi: 10.1073/pnas.1315532110

Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., and Meiler, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 49, 2987–2998. doi: 10.1021/bi902153g

Kieseritzky, G., Morra, G., and Knapp, E.-W. (2006). Stability and fluctuations of amide hydrogen bonds in a bacterial cytochrome *c*: a molecular dynamics study. *J. Biol. Inorg. Chem.* 11, 26–40. doi: 10.1007/s00775-005-0041-1

Konermann, L., Pan, J., and Liu, Y.-H. (2011). Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* 40, 1224–1234. doi: 10.1039/C0CS00113A

LeMaster, D. M., Anderson, J. S., and Hernández, G. (2009). Peptide conformer acidity analysis of protein flexibility monitored by hydrogen exchange. *Biochemistry* 48, 9256–9265. doi: 10.1021/bi901219x

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104, 59–107. doi: 10.1016/0022-2836(76)90004-8

Liu, T., Pantazatos, D., Li, S., Hamuro, Y., Hilser, V. J., and Woods, V. L., Jr. (2012). Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. *J. Am. Soc. Mass Spectrom.* 23, 43–56. doi: 10.1007/s13361-011-0267-9

Lobanov, M. Y., Suvorina, M. Y., Dovidchenko, N. V., Sokolovskiy, I. V., Surin, A. K., and Galzitskaya, O. V. (2013). A novel web server predicts amino acid residue protection against hydrogen-deuterium exchange. *Bioinformatics* 29, 1375–1381. doi: 10.1093/bioinformatics/btt168

López-Blanco, J. R., and Chacón, P. (2016). New generation of elastic network models. *Curr. Opin. Struct. Biol.* 37, 46–53. doi: 10.1016/j.sbi.2015.11.013

Ma, B., and Nussinov, R. (2011). Polymorphic triple β-sheet structures contribute to amide hydrogen/deuterium (H/D) exchange protection in the Alzheimer amyloid β42 peptide. *J. Biol. Chem.* 286, 34244–34253. doi: 10.1074/jbc.M111.241141

Mayne, L. (2016). "Chapter thirteen - hydrogen exchange mass spectrometry," in *Isotope Labeling of Biomolecules - Applications, Vol. 566 of Methods in Enzymology*, ed Z. Kelman (Cambridge, MA: Academic Press), 335–356.

Nagar, B., Jones, R. G., Diefenbach, R. J., Isenman, D. E., and Rini, J. M. (1998). X-ray crystal structure of C3d: A C3 fragment and ligand for complement receptor 2. *Science* 280, 1277–1281. doi: 10.1126/science.280.5367.1277

Neira, J. L., Itzhaki, L. S., Otzen, D. E., Davis, B., and Fersht, A. R. (1997). Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis. *J. Mol. Biol.* 270, 99–110. doi: 10.1006/jmbi.1997.1088

Papanastasiou, M., Koutsogiannaki, S., Sarigiannis, Y., Geisbrecht, B. V., Ricklin, D., and Lambris, J. D. (2017). Structural implications for the formation and function of the complement effector protein iC3b. *J. Immunol.* doi: 10.4049/jimmunol.1601864. [Epub ahead of print].

Park, I.-H., Venable, J. D., Steckler, C., Cellitti, S. E., Lesley, S. A., Spraggon, G., et al. (2015). Estimation of hydrogen-exchange protection factors from MD simulation based on amide hydrogen bonding analysis. *J. Chem. Inf. Model.* 55, 1914–1925. doi: 10.1021/acs.jcim.5b00185

Petruk, A. A., Defelipe, L. A., Rodríguez Limardo, R. G., Bucci, H., Marti, M. A., and Turjanski, A. G. (2013). Molecular dynamics simulations provide atomistic insight into hydrogen exchange mass spectrometry experiments. *J. Chem. Theory Comput.* 9, 658–669. doi: 10.1021/ct300519v

Pirrone, G. F., Iacob, R. E., and Engen, J. R. (2015). Applications of hydrogen/deuterium exchange MS from 2012 to 2014. *Anal. Chem.* 87, 99–118. doi: 10.1021/ac5040242

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29, 845–854. doi: 10.1093/bioinformatics/btt055

Radou, G., Dreyer, F. N., Tuma, R., and Paci, E. (2014). Functional dynamics of hexameric helicase probed by hydrogen exchange and simulation. *Biophys. J.* 107, 983–990. doi: 10.1016/j.bpj.2014.06.039

Rand, K. D., Zehl, M., Jensen, O. N., and Jørgensen, T. J. (2009). Protein hydrogen exchange measured at single-residue resolution by electron transfer dissociation mass spectrometry. *Anal. Chem.* 81, 5577–5584. doi: 10.1021/ac9008447

Schuster, M. C., Chen, H., and Lambris, J. D. (2007). "Hydrogen/deuterium exchange mass spectrometry: potential for investigating innate immunity proteins," in *Current Topics in Innate Immunity, Vol. 598 of Advances in Experimental Medicine and Biology*, ed J. D. Lambris (New York, NY: Springer), 407–417.

Sfyroera, G., Ricklin, D., Reis, E. S., Chen, H., Wu, E. L., Kaznessis, Y. N., et al. (2015). Rare loss-of-function mutation in complement component C3 provides insight into molecular and pathophysiological determinants of complement activity. *J. Immunol.* 194, 3305–3316. doi: 10.4049/jimmunol.1402781

Sim, A. Y., Levitt, M., and Minary, P. (2012). Modeling and design by hierarchical natural moves. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2890–2895. doi: 10.1073/pnas.1119918109

Skinner, J. J., Lim, W. K., Bédard, S., Black, B. E., and Englander, S. W. (2012a). Protein dynamics viewed by hydrogen exchange. *Protein Sci.* 21, 996–1005. doi: 10.1002/pro.2081

Skinner, J. J., Lim, W. K., Bédard, S., Black, B. E., and Englander, S. W. (2012b). Protein hydrogen exchange: testing current models. *Protein Sci.* 21, 987–995. doi: 10.1002/pro.2082

Sljoka, A., and Wilson, D. (2013). Probing protein ensemble rigidity and hydrogen-deuterium exchange. *Phys. Biol.* 10:056013. doi: 10.1088/1478-3975/10/5/056013

Şucan, I. A., and Kavraki, L. E. (2010). "Kinodynamic motion planning by interior-exterior cell exploration," in *Algorithmic Foundations of Robotics VIII*, eds G. S. Chirikjian, H. Choset, M. Morales, and T. Murphey (Berlin: Springer-Verlag), 449–464.

Tartaglia, G. G., Cavalli, A., and Vendruscolo, M. (2007). Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure* 15, 139–143. doi: 10.1016/j.str.2006.12.007

Truhlar, S. M., Croy, C. H., Torpey, J. W., Koeppe, J. R., and Komives, E. A. (2006). Solvent accessibility of protein surfaces by amide H/^{2}H exchange MALDI-TOF mass spectrometry. *J. Am. Soc. Mass Spectrom.* 17, 1490–1497. doi: 10.1016/j.jasms.2006.07.023

Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2003). Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. *J. Am. Chem. Soc.* 125, 15686–15687. doi: 10.1021/ja036523z

Wei, H., Tymiak, A. A., and Chen, G. (2013). "Hydrogen/deuterium exchange mass spectrometry for protein higher order structure characterization," in *Characterization of Protein Therapeutics Using Mass Spectrometry*, ed G. Chen (New York, NY: Springer), 305–341.

Wrabl, J. O., Gu, J., Liu, T., Schrank, T. P., Whitten, S. T., and Hilser, V. J. (2011). The role of protein conformational fluctuations in allostery, function, and evolution. *Biophys. Chem.* 159, 129–141. doi: 10.1016/j.bpc.2011.05.020

Wu, Z., Gogonea, V., Lee, X., Wagner, M. A., Li, X., Huang, Y., et al. (2009). Double superhelix model of high density lipoprotein. *J. Biol. Chem.* 284, 36605–36619. doi: 10.1074/jbc.M109.039537

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

REVIEW
published: 13 March 2017
doi: 10.3389/fmolb.2017.00012

Hybrid Methods in Iron-Sulfur Cluster Biogenesis

Filippo Prischi[1] *and Annalisa Pastore*[2, 3]*

[1] *School of Biological Sciences, University of Essex, Colchester, UK,* [2] *Maurice Wohl Institute, King's College London, London, UK,* [3] *Molecular Medicine Department, University of Pavia, Pavia, Italy*

Hybrid methods, which combine and integrate several biochemical and biophysical techniques, have rapidly caught up in the last twenty years to provide a way to obtain a fuller description of proteins and molecular complexes with sizes and complexity otherwise not easily affordable. Here, we review the use of a robust hybrid methodology based on a mixture of NMR, SAXS, site directed mutagenesis and molecular docking which we have developed to determine the structure of weakly interacting molecular complexes. We applied this technique to gain insights into the structure of complexes formed amongst proteins involved in the molecular machine, which produces the essential iron-sulfur cluster prosthetic groups. Our results were validated both by X-ray structures and by other groups who adopted the same approach. We discuss the advantages and the limitations of our methodology and propose new avenues, which could improve it.

Keywords: frataxin, NMR, molecular complexes, small angle X-ray scattering, structural biology, iron-sulfur cluster machinery, hybrid methods

OPEN ACCESS

Edited by:
Slavica Jonic,
UPMC Sorbonne Universités, France

Reviewed by:
Michael Sattler,
Technische Universität München, Germany
Ivaylo Ivanov,
Georgia State University, USA

***Correspondence:**
Annalisa Pastore
annalisa.pastore@crick.ac.uk

Specialty section:
This article was submitted to Biophysics,
a section of the journal
Frontiers in Molecular Biosciences

Received: *14 November 2016*
Accepted: *23 February 2017*
Published: *13 March 2017*

Citation:
Prischi F and Pastore A (2017) Hybrid Methods in Iron-Sulfur Cluster Biogenesis. Front. Mol. Biosci. 4:12.
doi: 10.3389/fmolb.2017.00012

INTRODUCTION

Biophysical approaches that make the combined and integrated use of different methodologies are named "hybrid techniques." Their use in Structural Biology has rapidly caught up in the last ca. 20 years (Sunnerhagen et al., 1996; Improta et al., 1998; Putnam et al., 2007; Tuukkanen and Svergun, 2014; Delaforge et al., 2015; Kachala et al., 2015; Milles et al., 2015; Sali et al., 2015; Prischi and Pastore, 2016; Venditti et al., 2016). A particularly useful application of hybrid techniques is the use of a combination of high and low resolution techniques which first target the local structure of a molecule (a domain or a complex component) and then reconstruct the full picture of the assembly (the so-called cut-and-paste approach) (Grishaev et al., 2005, 2008; Parsons et al., 2008; Deshmukh et al., 2013). Hybrid methods have, for instance, been successfully introduced to gain structural insights of complexes with different sizes which would be unaffordable if approached by only one technique (Wüthrich, 2001). One of the very first examples of hybrid methods was our study based on a combination of small angle scattering (SAXS) and nuclear magnetic resonance (NMR) to approach the arrangement of the domains of titin, a giant muscle modular protein containing more than 300 copies of two all-β sequence motifs, the fibronectin type 3 and the immunoglobulin-like modules (Improta et al., 1998). More recently, Michael Sattler and co-workers extended the approach to the study of RNA-protein interactions (Gabel et al., 2006; Madl et al., 2011; Huang et al., 2014). In 2010, we implemented a robust methodology which brings together NMR, SAXS, *in site* directed mutagenesis, ITC and other techniques to study weak complexes. This methodology has proven particularly effective for proteins of the iron-sulfur (FeS) clusters biogenesis machine, a highly conserved and essential metabolic pathway (Zheng et al., 1998). These proteins share important features which make particularly useful the application of hybrid methods to their study:

all the components of this cellular machine (i) tend to form transient interactions, making co-purification of the complexes difficult to impossible; (ii) have different binding affinities from each other; (iii) compete for the same binding sites; (iv) have different likelihood to crystallize, which often results in proteins forming crystals alone and not as part of the complex. In addition, many of the complexes are, although relatively large for NMR studies, too small for cryo-electron microscopy studies (Nogales and Scheres, 2015). As a result, high-resolution structures of most of these protein complexes are still not available. Here, we review our approach, discuss its successes and clarify the limitations. We also suggest ways to circumvent specific problems.

THE PARADIGMATIC EXAMPLE OF THE IRON-SULFUR CLUSTER BIOGENESIS COMPONENTS

Present ubiquitously in nearly all life forms, FeS clusters are protein inorganic prosthetic groups involved in a multitude of biological functions, such as electron transfer, gene expression regulation, thiolation, photosynthesis, nitrogen fixation, metal trafficking, substrate binding, DNA repair/replication and RNA modification (Johnson et al., 2005; Mettert and Kiley, 2015). FeS clusters are formed from iron ions and inorganic sulfide. Due to the toxic nature of these elements, formation of intracellular FeS clusters does not occur spontaneously, but all organisms have evolved protein machineries for the production of clusters. The FeS cluster assembly (ISC) system is a highly conserved factory found both in prokaryotes and eukaryotes and capable of providing FeS clusters to a wide range of apo-proteins. In particular, the eukaryote ISC machine is found in the matrix space of mitochondria and is distinct from the system that produces the clusters in the cytosol (Lill, 2009; Rouault, 2015). In *E. coli*, which is most studied as a model system because of its lower complexity, the ISC machine is composed of eight genes clustered in an operon, *iscRSUA-hscBA-fdx-iscX* (Takahashi and Nakamura, 1999; **Figure 1**). The operon is controlled by IscR, a transcriptional repressor (Schwartz et al., 2001) followed, in the order, by genes coding for a cysteine desulfurase (IscS) (Schwartz et al., 2000), a scaffold protein upon which clusters are built (IscU) (Agar et al., 2000), an A-type carrier with unclear function (IscA) (Krebs et al., 2001; Ollagnier-de-Choudens et al., 2001), a co-chaperone/chaperone system that is thought to facilitate cluster transfer from IscU to the final acceptor (hscA and hscB) (Chandramouli and Johnson, 2006), an electron donor ferredoxin (Fdx) (Yan et al., 2013b) and a protein with unknown function (IscX or YfhJ) (Pastore et al., 2006). This system constitutes the so called core assembly machine. The formation of FeS clusters by the core machine starts with the production of S^0 from L-cysteine by IscS, followed by reduction of S^0 to S^{2-} by Fdx (Yan et al., 2013b) and ends with the incorporation of Fe^{2+} or Fe^{3+} and formation of a [2Fe-2S] cluster on IscU (Agar et al., 2000). It is still unclear how the iron is delivered to the system.

Among these proteins, the crucial ones are IscS (NFS1 in human) and IscU (ISCU in human). IscS is a pyridoxal 5′-phosphate (PLP)-dependent desulfurase. PLP is not only necessary for the catalysis of L-Cys to L-Ala, but is also important for structural stabilization. CD spectra of recombinant IscS without PLP revealed in fact that the protein is completely unfolded, albeit proteolytically stable and not prone to aggregation (Prischi et al., 2010b). IscU, a 10 kDa protein, is predominantly in a monomeric state in solution and binds to IscS to accept the sulfur, which will form the clusters. The function of IscS and IscU are regulated by the protein frataxin (FNX). This is an essential protein highly conserved both in prokaryotes (where takes the name of CyaY) and eukaryotes where it is present in mitochondria. FXN was first identified for its connection to Friedreich's ataxia (Campuzano et al., 1996), a progressive neurodegenerative disease caused by an expansion of a GAA trinucleotide repeat within the first intron of the FXN gene, which results in reduced levels of FXN (Campuzano et al., 1996, 1997). Studies on the yeast frataxin homolog (YFH1) helped to understand that reduced levels of FXN causes loss of function of FeS cluster containing enzymes, increased amount of free radicals and iron deposits in mitochondria (Babcock et al., 1997; Foury and Cazzalini, 1997; Koutnikova et al., 1997; Rötig et al., 1997). Proteins from the FXN family bind weakly ferrous ions (Kd 4 μM) and ferric ions (Bou-Abdallah et al., 2004). These features are strongly conserved: human FXN is able to bind Fe^{2+} and Fe^{3+} in a similar way (Yoon and Cowan, 2003). Ability to weakly bind iron could be in agreement with the hypothesis that the protein functions as an iron chaperone, but the way FXR binds iron is unusual. The FXN fold, which is composed of two α-helices packed against an anti-parallel β-sheet (Cho et al., 2000), does not share any similarity with any other known iron binding proteins, like ferritins, ferredoxins or hemoglobins (Harrison and Arosio, 1996). It is also unusual that iron coordination occurs solely through carboxylate residues and no conserved histidine, cysteine, or tyrosine - residues usually found in iron binding motifs - are present in frataxins (Nair et al., 2004). Finally, cation binding is highly unspecific since, in addition to iron, frataxins bind to diamagnetic Ca^{2+}, Zn^{2+}, Lu^{3+}, and paramagnetic ions Mn^{2+}, Co^{2+}, Gd^{3+}, Eu^{3+}, and Yb^{3+} (Nair et al., 2004). Twenty years have passed since these initial studies which have made clear that FXN is connected to FeS cluster formation, but the exact function of FXN remains elusive. Different theories have been proposed: (i) FXN is the iron chaperone that delivers Fe^{2+} or Fe^{3+} to IscU (Yoon and Cowan, 2003); (ii) FXN acts as a scavenger that is able to sequester mitochondrial iron through formation of high-molecular-weight aggregates and to maintain it in a bioavailable form (Adamec et al., 2000). We have proposed a third hypothesis which is currently the most accredited: (iii) FXN acts as an iron sensor that regulates the amount of FeS cluster formed to match the concentration of the available acceptors (Adinolfi et al., 2009). Our model, which proposes a completely new function of FXN, is based on studies that rely on the demonstration that FXN binds to the IscS/IscU complex in an iron dependent manner (Prischi et al., 2010a). To gain information on this ternary complex, we adopted a hybrid approach, which relied on NMR, SAXS, site directed mutagenesis, molecular docking and molecular dynamics simulations.

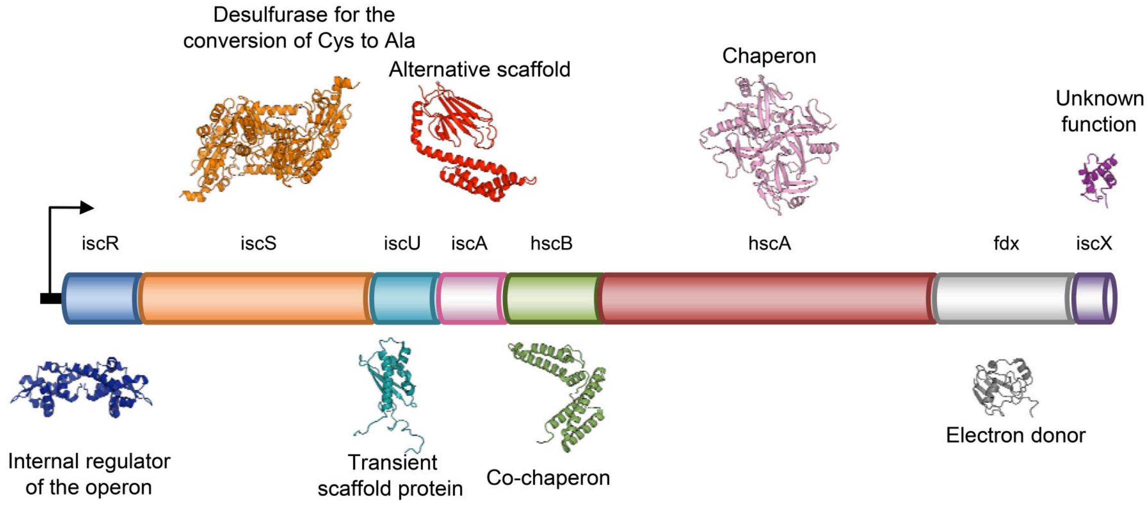

Desulfurase for the conversion of Cys to Ala
Alternative scaffold
Chaperon
Unknown function
iscR
iscS
iscU
iscA
hscB
hscA
fdx
iscX
Internal regulator of the operon
Transient scaffold protein
Co-chaperon
Electron donor

FIGURE 1 | Schematic representation of the genetic organization of the bacterial ISC system in operon.

THE DEVELOPMENT OF A HYBRID METHOD

The rationale of our hybrid method develops through the following logic steps (**Figure 2**):

Step 1: Identifying the ISC Interactome

The network of interactions between IscS, IscU, and CyaY was probed by NMR spectroscopy. We exploited the well-known concept that the spectrum of a molecule is very sensitive to the chemical environment. Protein-protein interaction cause changes in the chemical environment of the reporter nucleus. This means that titration of a protein with another molecule results in shifts in the position of some or all resonances in the spectrum or Chemical Shift Perturbation (CSP) (Roberts, 1993; Zuiderweg, 2002), which can then be used to map the regions involved in the interaction. Typically, we titrated a ^{15}N labeled component of the *isc* operon with another un-labeled protein. However, the resonance line widths depend inversely on the tumbling time (Bloembergen et al., 1948) and, thus, the larger the complex, the broader are the line widths up to spectral disappearance. Many of the ISC components have sizes well within the limits of NMR observation except for IscS, which is an obligate dimer of 90 kDa, and the chaperone HscA. This meant that we could alternatively titrate the low size proteins (i.e., adding unlabeled protein A into ^{15}N labeled protein B and, viceversa, unlabeled protein B into ^{15}N labeled protein A) and map the interacting site on both proteins. The case was quite different when adding the 90 kDa IscS to a smaller component. In this case the result would strongly depend on the regime of exchange of the complex.

The most common NMR experiment used to measure CSP is the two-dimensional ^{15}N heteronuclear single-quantum coherence NMR ([^{1}H,^{15}N]-HSQC NMR), a method that allows the detection of correlations between ^{15}N nucleus and ^{1}H nucleus which are covalently bound. Titration of IscS into ^{15}N labeled IscU caused complete disappearance of IscU signal from the [^{1}H,^{15}N]-HSQC NMR spectra at a 1:0.7 IscU:IscS molar ratio (Prischi et al., 2010b) without previous CSP. Absence of detectable CSP for these titrations and disappearance of the IscU signal indicates binding but also suggests that the process is under an intermediate-to-slow exchange regime in the NMR time range (**Figures 3A–C**). The exchange regime is the rate k_{ex} at which a nucleus switches from one conformation to another (in this case a "free state" to a "bound state"). The NMR linewidths depend on the populations of each state, the relative values of the exchange rate k_{ex} and the chemical shift difference $\Delta\nu$. In the slow exchange regime ($k_{ex} << |\Delta\nu|$), signals from both states are observed at their distinct chemical shifts, intensities and linewidths; if the regime is fast ($k_{ex} >> |\Delta\nu|$), a single peaks will be observed at the chemical shift between free and bound conformations weighed according to the populations; if it is in an intermediate regime ($k_{ex} \approx |\Delta\nu|$), a single peak is observed between the two states but due to the presence at the same time of the free state and the bound state, the resulting resonance is broadened (Kleckner and Foster, 2011). In our case IscU alone corresponds to the "free state", while the IscU-IscS complex is the "bound state". Since the spectra are completely unperturbed until we reach a 1:1 ratio IscU:IscS, we can deduce that this is not in a fast exchange regime in the NMR time range and we can deduce that the process is an intermediate-to-slow exchange regime. We would expect then to see peaks for both the free state and the bound one. However, the high molecular weight of the complex causes that the bound state is outside the limit of NMR observation and we do not observe it. This did not allow us to map the interaction surface of IscU on IscS, a problem often observed in the NMR studies of complexes. We also did not observe CSP when titrating directly IscU and CyaY in the absence of IscS but in this case the spectra of the two proteins, individually labeled in turn, where completely unaffected. This meant no

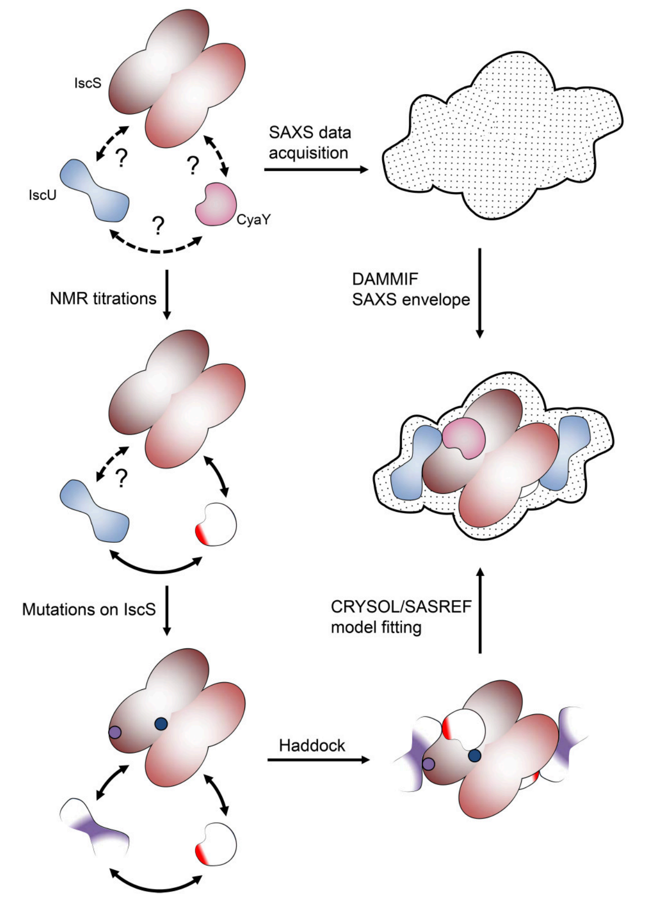


FIGURE 2 | Pipeline of the hybrid methodology described in this review. At the beginning we only know that proteins interact but do not know anything about their mutual orientation. NMR allowed us to map the surface of interaction of the lower molecular weight proteins. The surface involved in interaction on IscS was identified by a combination of mutations and SAXS experiments.

direct interaction in contrast with studies carried out on the human and yeast proteins, where a direct interaction between the scaffold protein and FXN was observed (Yoon and Cowan, 2003; Correia et al., 2009; Leidgens et al., 2010), suggesting a different behavior of the bacterial proteins. The difference could perhaps be ascribed to the lack of the N-terminal extension which, in eukaryotes, is part of the mitochondrial signal and absent in prokaryotes. Finally, when titrated with IscS, the spectrum of CyaY remains visible and shows clear CSP, which allowed us to map the interaction on a specific surface (**Figures 3D–F**). CyaY interacts with IscS using a negatively charged surface area localized on α1, β1 and α1β1 and β1β2 loops (Adinolfi et al., 2009). Interestingly, this negatively charged surface is the same involved in iron binding (Yoon and Cowan, 2003; Bou-Abdallah et al., 2004; Nair et al., 2004). We then tested for competition between CyaY and IscU binding on IscS by [^{1}H,^{15}N]-HSQC NMR spectra titrating ^{15}N labeled IscU with up to an equimolar amount of unlabeled IscS with unlabeled CyaY. Presence of competition should cause dissociation of ^{15}N labeled IscU from IscS, resulting in the reappearance or increase of the NMR signal, in a way proportional to the amount of competitor added. We did not observe competition with IscU (Adinolfi et al., 2009; Prischi et al., 2010a). We could thus conclude that both CyaY and IscU bind to IscS, but not each other and obtain the surface of interaction on CyaY from NMR only.

CSP data did rule out the presence of a direct interaction or competition between CyaY and IscU, but this did not automatically exclude binding between the two when in the presence of IscS. We titrated ^{2}H, ^{15}N double-labeled CyaY with unlabeled IscU and IscS up to a 1:1:1 molar ratio. ^{2}H labeling reduces spin-spin relaxation, a parameter inversely proportional to the linewidths of the resonance in the spectrum. This results in narrower linewidths and thus provide higher resolution (Gardner and Kay, 1998). We observed a new set of spectral perturbations, which we attributed to a direct contact of the residues involved with IscU. Once mapped onto CyaY structure, these residues clustered on the anti-parallel β-sheet surface of CyaY (**Figure 4A**). In particular, the conserved Trp61 in CyaY was found to be involved in the interaction with IscU. These data are in agreement with studies on human FXN, where it was shown that the exposed side chain of Trp155 (equivalent to CyaY Trp61) is indispensable for FXN-ISU binding (Correia et al., 2009; Leidgens et al., 2010).

To map the surface of interaction on IscS we used site directed mutagenesis. We designed mutations of IscS targeting solvent exposed residues. We aimed to abolish interaction with ISC components, while keeping IscS stable and functional. We titrated ^{15}N labeled IscU and CyaY with five different IscS mutants, i.e., IscS_R220E/R223E/R225E, IscS_I314E/M315E, IscS_K101E/K105E, IscS_E334S/R340S and IscS_R39E/W45E (Prischi et al., 2010a). IscS_R220E/R223E/R225E triple-mutant, in which a positively charged patch formed mainly by arginines close to the dimer interface was inverted in charge, does not bind CyaY (**Figure 4B**; Prischi et al., 2010a). This strongly supported the assumption that binding between these proteins is driven by electrostatic interactions and is in agreement with our competition studies. Differently, in IscS_I314E_M315E we inserted two charged residues into an uncharged/hydrophobic patch. This mutant has a reduced affinity for IscU, which, due to a change in the exchange regime, caused chemical shift perturbation of the ^{15}N-labeled IscU HSQC spectrum. This not only allowed us to identify IscS interacting surface, but also to identify the IscU residues involved in IscS binding (**Figure 4A**; Prischi et al., 2010a). All other IscS mutants behaved like the wild type in titration experiments and provided us with solid controls (Prischi et al., 2010a).

Step 2: Restrained Docking Simulation

To gain a visual impression and understand the relative orientation of proteins in the complex, we generated models of the central ISC machine using NMR restrained molecular docking simulations. We used the docking software HADDOCK (Dominguez et al., 2003). HADDOCK can incorporate NMR or other distance restraints and implement them as "ambiguous interaction restraints" (AIRs). The software forces the protein interfaces to come together without imposing a particular

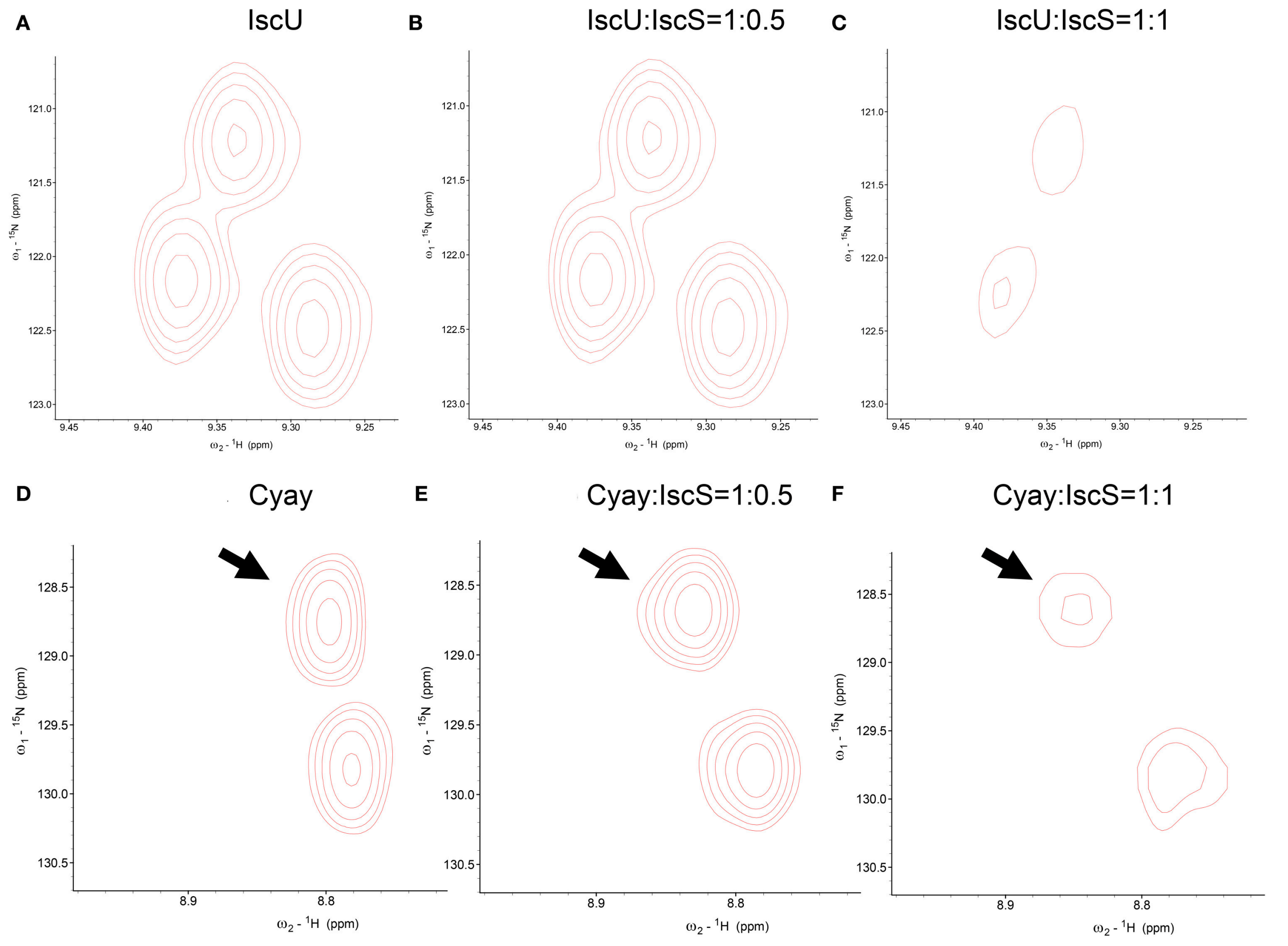


FIGURE 3 | Comparison among close-ups from HSQC spectra of 15N uniformly labeled IscU. (A) From left to right, IscU alone; **(B)** IscU in the presence of unlabeled IscS at a molar ratio of 1:0.5; **(C)** The same as in **(B)** but at a molar ration of 1:1. No detectable CSP of IscU peaks and concomitant disappearance of the IscU spectra when saturated with IscS suggest that the process is under an intermediate-to-slow exchange regime in the NMR time range. For comparison, The spectra were all recorded at 25°C and 600 MHz and visualized using the same signal to noise ratio. **(D)** CyaY alone; **(E)** CyaY in the presence of unlabeled IscS at a molar ration of 1:0.5; **(F)** The same as in **(E)** but at a 1:1 molar ratio. The spectra were all recorded at 25°C and 600 MHz and visualized using the same signal to noise ratio.

orientation. Using the AIRs that we determined experimentally during Step 1, we obtained different families of complexes, which differed by details but all reported IscU bound on the opposite tips of the IscS dimer, roughly close to the N-terminus. CyaY was instead consistently located near the cavity that contains the active site of IscS, spatially close but not overlapping with IscU. While these results could have been sufficient for having a first rough model of the IscS-IscU-CyaY complex, we felt that further validation was needed to confirm independently the relative positions of the three proteins.

Step 3: Validation through SAXS Data

The generated models were then experimentally verified and re-scored using SAXS data. SAXS is a solution technique that allows to study the shape, conformation and assembly state of proteins and, more in general, macromolecular complexes (Mertens and Svergun, 2010). Despite being a low-resolution technique, SAXS is well suited for the study of flexible systems and intrinsically disordered proteins (Wang et al., 2011), which are major limitation in X-ray crystallography and cryo-EM. It also allows the study of proteins in solution in nearly-native conditions. SAXS experiments are not time consuming. Recent hardware improvements allow high-throughput studies (Round et al., 2008). A monochromatic X-ray beam is scattered by the protein sample in solution. At low (below 0.1 Å^{-1}) and medium momentum transfer (s) (between 0.1 Å^{-1} and 0.25/0.3 Å^{-1}) scattering angle, we obtain different information about the system (**Figure 5A**). Above 0.3 Å^{-1} the noise masks the signal and above 0.5 Å^{-1} data are collected at wide angle. This technique is not called SAXS anymore, but WAXS (Wide Angle X-ray Scattering) (Graewert and Svergun, 2013). At low s it is possible to extrapolate the radius of gyration R_g, which provide information about the size of the protein (Grant et al., 2015; Kikhney and Svergun, 2015). In order to obtain a reliable R_g, it is important to

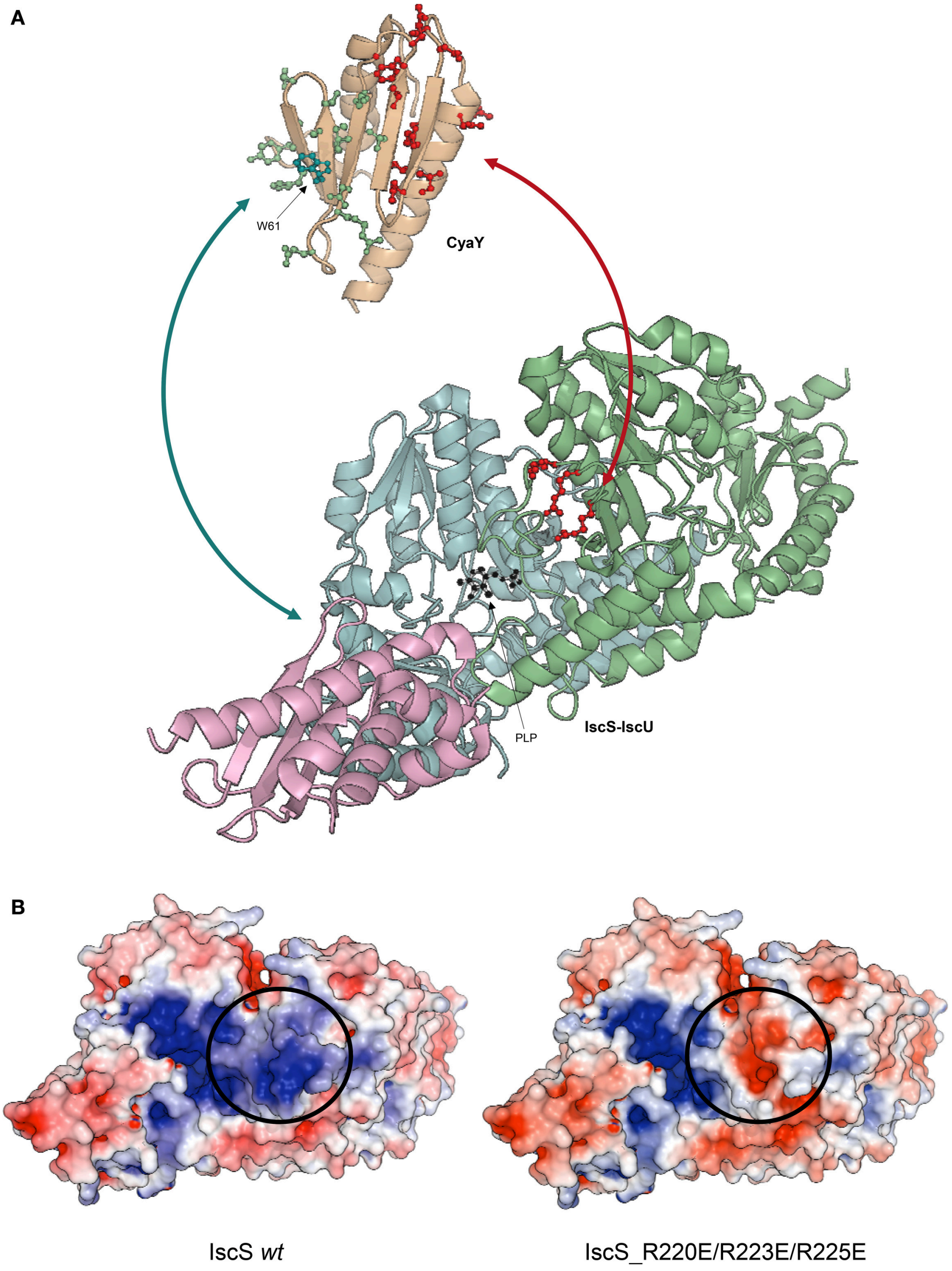
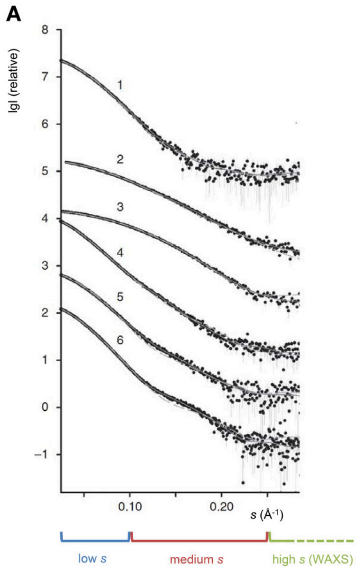


FIGURE 4 | Ribbon representation of CyaY and IscS-IscU interactions. (A) IscU is shown in pink, while IscS monomers are colored in pale cyan and light green with side chains of residues mutated indicated explicitly: R220E/R223E/R225E (red). PLP in IscS active site is shown in black. Side chains of CyaY residues exhibiting CSP are explicitly shown: residues interacting with IscS are in red (Trp14, Leu15, Glu19, Asp22, Asp23, Trp24, Asp25, Asp27, Ser28, Asp29, Ile30, Asp31, Cys32, Glu33, Ile34, Leu39, Thr42, Phe43, Glu44, and Gly46), while residues interacting with IscU are in green (Thr40, Ile41, Lys48, Ile50, Asp52, Arg53, Glu55, Trp61, Leu62, Ala63, Thr64, Gln66, Gly68, Tyr69, and His70). **(B)** Electrostatic surface of unbound Iscs. The circle indicates the position of the positively charged residues involved in binding.

have a precise measurement of the sample concentration before SAXS measurements. Medium *s* provides information about the shape of the protein. More precisely it is possible to obtain the D_{max}, which provides measurement of the maximum dimension of the protein (Svergun, 1992).

From these measurements it is possible to build a low-resolution envelop, which represents the shape of the protein studied. The shape is reliable only when the system under study is monodisperse. In fact, it is always possible to obtain envelops from SAXS data, but poly-disperse samples do not



B

	R_g (Å)	D_{max} (Å)	MM_{SAXS}(kDa)	MM_{exp} (kDa)	χ_{ab}	χ_{rb}
1 IscS	31.0±0.4	109±5	85±10	90.18	1.09	1.25
2 IscU	18.1±0.3	64±5	13±4	13.849	1.03	1.48
3 Cyay	15.2±0.3	50±3	12±4	12.231	1.01	1.2
4 IscS/IscU	35.0±0.5	121±5	112±10	117.878	1.21	1.5
5 IscS/CyaY	31.1±0.4	109±5	105±10	114.642	1.08	1.85
6 IscS/IscU/CyaY	34.1±0.5	119±5	122±10	142.34	1.18	1.78

FIGURE 5 | SAXS profiles of the complexes. (A) The X-ray scattering patterns from IscS (1), IscU (2), CyaY (3) binary complexes IscS/IscU (4) and IscS/CyaY (5) and ternary complex IscS/CyaY/IscU (6). Plots display the logarithm of the scattering intensity as a function of momentum transfer (s). At the bottom it is highlighted with a blue box the low *s* region of the SAXS curve, in red the medium and in green the high. The experimental data are displayed as dots with error bars, the scattering from typical ab initio models computed by DAMMMMIF as full lines and the calculated curves from the high-resolution (for proteins alone) and rigid body models (for complexes) computed by CRYSOSOL/SASREF as dashed lines. The successive curves are displayed down by one logarithmic unit for clarity (figure adapted from Prischi et al., 2010a). **(B)** Table summarizing SAXS data. R_g is the radius of gyration; D_{max} is the maximum size of the particle; MM_{SAXS} is the molecular mass calculated from SAXS data; MM_{exp} experimental molecular mass of the solute and χ_{ab} and χ_{rb} values for the fit curves from *ab initio* models and from high resolution models (for proteins alone) and rigid body modeling (for complexes) using CRYSOL/SASREF, respectively.

generate envelops that represent the real shape of the protein. For example, in the study of PERK N-terminal domain, the protein was in a dynamic equilibrium between a dimer and a tetramer (Carrara et al., 2015). In this case it is not possible to obtain protein shape information, but SAXS is still informative. It had to be assumed that the resulting shape was a weighted average of dimer and tetramer shapes. The factor of weight had to be obtained from independent techniques, such as analytical ultracentrifuge (AUC), from which the relative populations of PERK dimer and tetramer in solution were estimated. The SAXS curve was then back-calculated using PERK dimer and tetramer crystal structures, weighted according to their relative abundance in solution, and fitted on experimental data. Agreement between experimental and back-calculated SAXS curves provided a confirmation that the PERK oligomeric structures were not a crystallographic artifact, but representative of the oligomeric state of PERK in solution (Carrara et al., 2015).

Luckily, IscS, IscU, and CyaY are all mono-disperse in solution. It was thus possible to obtain reliable information about their shapes from SAXS only. We collected SAXS data for each of the individual components, as well as for the binary (IscS-IscU, IscS-CyaY) and tertiary (IscS-IscU-CyaY) complexes (Prischi et al., 2010a). As previously mentioned, due to the dynamic nature of the Fe-S cluster machinery, the binding affinities of IscU and CyaY for IscS are relatively low: $K_{dIscU\text{-}IscS}$ = 1.3 ± 0.2 μM and $K_{dCyaY\text{-}IscS}$ = 18.5 ± 2.4 μM (Prischi et al., 2010a). We were able to isolate IscU-IscS complex using Size-Exclusion Chromatography (SEC) (Prischi et al., 2010b), but not CyaY-IscS and IscS-IscU-CyaY. In these cases we directly mixed proteins in solution prior data collection. Knowledge of relative K_d allowed us to estimate the optimal proteins ratios in order to maximize formation of the (Prischi et al., 2010a). It is worth mentioning that, despite not being available when we collected our data, a new methodology, particularly useful when collecting SAXS data on protein complexes, is Size-Exclusion Chromatography in line with SAXS (SEC–SAXS) (Mathew et al., 2004). SEC-SAXS is useful for separating pure systems that are under monomer-oligomer equilibrium or to further purify the sample before SAXS data are collected (particularly indicated for low stability protein which tend to form soluble aggregates). SEC–SAXS is available as a continuous-flow sample delivery option at BioCAT (Advanced Photon Source, U.S.A.) (Mathew et al., 2004), SWING (Soleil, France) (David and Perez, 2009), the SAXS beam line at the Australian Synchrotron, BM29 (ESRF, France), BL23A1 (NSRRC, Taiwan), B21 (Diamond, U.K.) and P12 (DESY, Hamburg) (Blanchet et al., 2015). SEC-SAXS has however limitations and it shouldn't be used as a purification step (Jeffries et al., 2016).

Ab initio envelops were generated using the software DAMMIF (Franke and Svergun, 2009). The high-resolution PDB structures 1P3W (Cupp-Vickery et al., 2003) for IscS and 1SOY (Nair et al., 2004) for CyaY were fitted into the SAXS envelops by rigid body modeling. Two different structures were used for IscU: one solved by NMR (PDB ID 1Q48) (Ramelot et al., 2004) and one by X-ray crystallography (PDB ID 2Z7E) (Shimomura et al., 2008). The two structures have a similar overall secondary structure content, but while in the NMR structure the first 25 residues are in a random coil conformation, the crystal structure is more compact and the N-Terminus forms a α-helix which makes contacts with α3 and the α5α6 loop. The quality of fitting of a 3D structure on a SAXS envelop can be visually ascertained, and can be more accurately estimated using the χ^2 (Svergun, 1999). χ^2 tells us how well the back-calculated scattering intensity from a 3D structure fits the experimental SAXS data. Fitting of the isolated CyaY and IscS resulted excellent, with a χ^2 of respectively 1.01 and 1.09 and an estimated molar mass of respectively 12 kDa ± 4 kDa (expected 12.231 kDa) and 85 kDa ± 10 kDa (expected 90.180 kDa) (**Figures 5A,B**). Of the two structures, 1Q48 fitted better the SAXS data collected for the isolated IscU in agreement with the dynamic nature of isolated IscU in solution (Kim et al., 2009; Prischi et al., 2010b), with a χ^2 of 1.03 and an estimated molar mass of 13 kDa ± 4 kDa (expected 13.849 kDa) (**Figures 5A,B**). It is also strongly recommended to check the residuals of the difference between experimental and back calculated SAXS data (i.e., whether these are random and not systematic).

Step 4: Experimental Validation of the Models

The software DAMMIF (Franke and Svergun, 2009) was used to generate envelops from SAXS data. DAMMIF, a fast version of DAMMIN (Svergun, 1999), carries out an *ab initio* shape determination by simulated annealing using a single phase Dummy Atom Model (DAM). The DAM is represented by a tightly packed group of beads, which mimic, but do not resemble, real atoms. Each bead has a known scattering pattern and the software puts beads together so that the accumulated scattering resembles the experimental data. The software used to generates back-calculated curves and fit them on experimental data is the CRYSOL software (Svergun et al., 1995). CRYSOL requires a 3D structure/object as an input and then, taking into account the contribution for each atom, it evaluates the scattering intensity.

Despite having a 10–20 Å resolution ($2\pi/s_{max}$), the SAXS envelops of the binary and tertiary complexes resulted evidently different from those of the single components. We first tested whether SAXS data were sufficient to generate meaningful binary and tertiary complexes using the SASREF software (Petoukhov and Svergun, 2005). SASREF tries to build the quaternary structure of a complex using the structures of the subunits and the solution scattering data. It is particularly useful because it can work with multiple data set(s), which allows working with SAXS data from sub-complexes and creating contrast series. SASREF build the complex structure without steric clashes using a simulated annealing protocol, which minimize differences between the experimental scattering data and the back-calculated SAXS curve of the model being built.

We inputted in SASREF the SAXS data and the high-resolution structures of the single components but the complexes obtained with this approach did not generate reliable models, since they were not in agreement with our binding studies. Instead, we docked our HADDOCK structures into SAXS envelops. HADDOCK models were used to generate back-calculated SAXS curves, which were fitted on experimental data. Based on χ^2, we selected the best fitting model, which was an experimentally verified model of FeS machinery complexes. Selecting the "best fitting model" could be problematic if two similar HADDOCK models have small orientation differences, which are clearly not distinguishable at SAXS resolution. In this context, HADDOCK is particularly well suited, because it first generates a maximum of 1,000 structures and then groups them according to their relative RMSD. By aligning these generated models using the interface residues of the first molecule, the RMSD (more correctly called interface-ligand RMSD) is calculated for the interface residues (less than 10 Å distance from the first molecule) of the second molecule (Dominguez et al., 2003). In our case, all structures HADDOCK grouped within the same group had RMSD < 7.5 Å. For each group, we used the structures with overall lower energy (evaluated by HADDOCK). Analysis of the binary complexes confirmed that CyaY sits near the IscS dimer interface and the active site, while IscU is located on the periphery of the IscS dimer and is aligned with the long axis of IscS. Accordingly, the IscS-IscU (R_g = 35 Å, D_{max} = 121 Å) envelop is more elongated than the IscS alone (R_g = 31 Å, D_{max} = 109 Å), while the IscS-CyaY envelop is more globular (Prischi et al., 2010a; Yan et al., 2013b; **Figure 4B**).

Step 5: Comparison with X-Ray Crystal Protein Complexes

A limitation of this procedure is the absence of a tool to predict/model major structural changes upon formation of a complex. HADDOCK can simulate small conformational changes during the molecular dynamics refinement, but the final model strongly depends on the initial 3D structures provided: HADDOCK assumes a key-in-the-lock model. If a protein has significantly different structures in the free and bound states, HADDOCK (like any other protein docking software) will fail to generate a reliable model. For the IscS-IscU complex, we found that the model generated from HADDOCK did not fit the SAXS envelop as well as the single components did. However, a crystal structure of the IscS-IscU complex (PDB ID 3LVL) (Shi et al., 2010) became available while we were carrying out our studies. This structure is in perfect agreement with our NMR and mutant binding data and fits the SAXS envelop better than the HADDOCK model. This is due to IscU going through a structural rearrangement upon binding, with a formation of a α-helix in the N-terminus, similar to the one seen in 2Z7E (Shimomura et al., 2008). IscU has an optimal orientation for FeS cluster formation, with the surface containing three conserved cysteines pointing toward Cys328 in IscS loop (Shi et al., 2010). The distance between IscS active site and IscU is around 12 Å, suggesting the presence of major conformational changes happening during FeS cluster formation.

We then used 3LVL (Shi et al., 2010) for modeling the tertiary complex, IscS-IscU-CyaY. Interestingly, the model confirmed that, despite not being able to interact between each other directly, CyaY and IscU can interact once bound to IscS. This structure helped us to explain an inhibitory effect of CyaY on FeS cluster formation: enzymatic studies had showed that the tertiary complex is "less dynamic" than the binary ones with CyaY creating an additional anchoring point between IscS and IscU (Prischi et al., 2010a). This is in agreement with the observation that CyaY binding increases the affinity of IscU for IscS thus reducing the dissociation rates of the complex (the k_{off} for the disassembly of the IscS/IscU complex is 0.8 s^{-1} in the absence of CyaY, vs. 0.006 s^{-1} in the presence of CyaY).

A STEP FORWARDS: MOLECULAR DYNAMIC SIMULATIONS

From our studies it emerged that the dynamic nature of the ISC proteins is a key factor in their functions. To feature this dynamical behavior, we thus complemented our previous data with extensive (400 ns) molecular dynamic (MD) simulations (di Maio et al., 2017) of the IscS-IscU complex, both in the presence and in the absence of CyaY. We showed that the binary IscS-IscU complex is stably folded in line with our SAXS evidence (Prischi et al., 2010a), but IscU adopts a likely functionally relevant pivotal motion around the interface with IscS. This means that, despite being firmly attached to IscS, IscU maintains some degree of flexibility upon complex formation, which can be connected to their low binding affinity and the need of IscU to deliver FeS cluster to protein acceptors. At the same time, the pivotal motions observed in the MD simulations suggest that IscS-IscU interface is "fluid," with IscU side chains at the interface being trapped in several local minima. This was confirmed by NMR experiments (di Maio et al., 2017).

During the trajectory, the IscS catalytic loop containing Cys328 moves spontaneously and shifts from a mostly 3_{10}-helical structure to a β-turn/3_{10}-helix equilibrium, bringing Cys328 from 12 Å to 9 Å from the FeS cluster binding site on IscU (di Maio et al., 2017). This is in agreement with the IscS-IscU X-ray structure of the *A. fulgidus* (PDB ID 4EB5) (Marinoni et al., 2012), which brilliantly captured the delivery stage of FeS cluster from IscS to IscU. In 4EB5, the FeS cluster is bound to the Cys of the IscS catalytic loop and is about 6 Å away from the IscU FeS cluster binding site (Marinoni et al., 2012).

In agreement with our previous studies (Prischi et al., 2010a), the simulations showed that the tertiary complex IscS-IscU-CyaY is more stable and that CyaY reduces the structural fluctuations of the IscS-IscU complex (di Maio et al., 2017). The most striking feature of the complex is the absence of motions of the IscS catalytic loops (one for each protomer) over the same timescale, due to CyaY steric hindrance and a salt bridge between CyaY Arg53 and IscS Glu334. This model brings us back to the beginning of this review, where we described the possible roles of FXN. The model we generated recapitulated in an elegant way our enzymatic data and provides a mechanistic explanation of how CyaY slows down FeS cluster formation (Adinolfi et al., 2009; Prischi et al., 2010a).

EXTENSION OF THE METHODOLOGY TO OTHER ISC COMPLEXES

The approach described here has now been adopted by others (Kim et al., 2014, 2015) also to elucidate other ISC complexes,

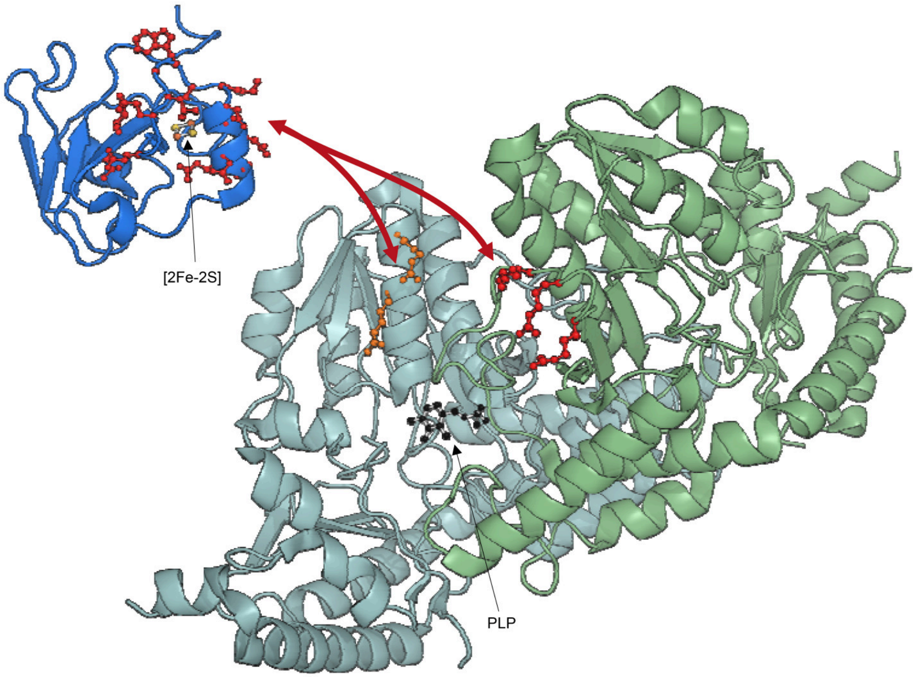


FIGURE 6 | Ribbon representation of Fdx, YfhJ, and IscS interaction. IscS monomers are colored in pale cyan and light green with side chains of residues mutated indicated explicitly: R112E/R116E (orange), R220E/R223E/R225E (red). PLP in IscS active site is shown in black. Side chains of holo-Fdx residues exhibiting CSP (Ile54, Val55, Gln68, Glu69, Asp70, Asp71, Met72, Leu73, Asp74, Lys75, Ala76, Trp77, Gly78, Leu79, Glu80, Glu82) are explicitly shown in red. Fdx is loaded with a [2Fe-2S] cluster. YfhJ is colored in light blue and side chains of residues exhibiting CSP (Leu3, Lys4, Glu10, Ile11, Glu13, Ala14, Asp17, Leu58, Trp61, Leu62, Asp63, Glu64) are explicitly shown in blue.

TABLE 1 | Hybrid method breakdown.

	PROS	CONS
NMR	• Proteins are in solution; • Detects presence of protein interaction; • Allows identification of residues involved in protein interaction; • Solve structure of small complexes (<30 kDa) at atomic resolution; • Allows to measure protein dynamics in solution	• According to the exchange regime in the NMR time range it is not always possible to identify residues involved in interactions → Site Directed Mutagenesis can be used to modify affinity and hence the exchange regime; • When the system crystallizes→ X-Ray Crystallography and SAXS can be a valid alternative approach.
X-Ray Crystallography	• Solves structure of a protein complex at atomic resolution.	• Not all proteins or protein complexes crystallize.
SAXS	• Proteins are in solution; • Detects presence of protein interaction; • Generate low resolution (10-20Å) models of proteins and protein complexes;	• Requires 3D structures solved by NMR or X-Ray Crystallography; • Unreliable protein complexes models built based only on SAXS data → Docking software (HADDOCK) can be used to generate models; • Provides shape of the protein or protein complexes in solution; • Generates reliable protein envelops only for monodisperse samples. • Detect conformational changes.
Site directed mutagenesis	• Allows to lower proteins affinities; • Allows to abolish protein interactions.	• May require the production of several different mutant clones in order to find residues involved in protein interaction; • Does not provide overall structural information.
ITC & Fluorescence Spectroscopy	• Detects presence of protein interaction; • Measure affinities of protein complexes; →Allows to predict proteins exchange regime in the NMR time range.	• Requires Site Directed Mutagenesis in order to identify residues involved in protein interactions; • Does not provide overall structural information.
Protein-Protein Docking simulation (HADDOCK)	• Generate 3D structures of protein complexes by forcing the protein interfaces to come together without imposing a particular orientation.	• Requires 3D structures solved by NMR or X-Ray Crystallography; • Reliable only in presence of experimental interaction restraints →NMR and Site Directed Mutagenesis can be used to identify residues involved in protein interaction.
Molecular Dynamics simulations	• Allows to measure and observe dynamical features of proteins and proteins complexes.	• Requires a 3D structure or an experimentally verified model.

Flowchart of the pros and cons of the different techniques part of the hybrid method adopted for the study of Iron-sulfur cluster machinery.

increasing the robustness of the methodology. A study of the complex between IscS and YfhJ was published (Kim et al., 2014). YfhJ behaves similarly to CyaY as it is able to bind both Fe (II) and Fe (III) using an electrostatic negative surface, which is the same area involved in IscS binding (**Figure 6**; Pastore et al., 2006). YfhJ also competes for the same site of CyaY on IscS in agreement with previous mutation studies (Shi et al., 2010).

We have ourselves recently applied this hybrid method to model the IscS complex with Fdx, a FeS cluster dependent protein which is known to provide electrons to cellular reactions. Fdx is not-functional and devoid of tertiary structure in the absence of the cluster (Yan et al., 2013a). As for IscU, the spectrum of labeled Fdx disappears completely upon addition of unlabeled IscS. To circumvent the problem, we titrated ^{2}H, ^{15}N double-labeled holo-Fdx with IscS using [^{2}H,^{15}N]-SOFAST HMQC NMR experiments (Yan et al., 2013b). This experiment requires a shorter acquisition time compared to HSQC and is thus more suitable for unstable samples. We could then identify the residues of Fdx involved in IscS binding cluster, which reside near a uniform acidic patch on the α2-α3 loop (**Figure 6**; Yan et al., 2013b).

NMR competition studies revealed that Fdx and CyaY compete for the same site of IscS (Yan et al., 2013b). A Fdx-IscS SAXS verified model showed that Fdx sits in a position similar to that of CyaY near the active site. This was utterly validated by creating a new IscS mutant (IscS_R112E/R116E), which interferes with Fdx binding (Yan et al., 2013b). Assuming that the two proteins exploit their functions in different times during the cluster biogenesis, competition could represent a fascinating regulation mechanism. Superimposition of the Fdx-IscS and IscS-IscU models reveals that the Fdx C-terminus (which contains two key residues for electron transfer reactions, Tyr101 and His105) points toward the interface between IscS and IscU. This nicely explains how, after production of S^0 from L-cysteine by IscS, Fdx could reduce S^0 to S^{2-} (Yan et al., 2013b, 2015).

To add surprise to surprise, we have more recently shown that also the co-chaperone HscB binds to IscS in the same binding pocket, a result further validated by cross-linking experiments (Puglisi et al., 2016). This implies a picture in which IscS acts as a central platform on which several of the other bacterial ISC proteins bind and typically form 1:1 complexes (Pastore et al., 2006; Adinolfi et al., 2009; Prischi et al., 2010a; Yan et al., 2013b). It remains for us to understand why and how several different components of the same pathway compete for the same site. We suggested that this is a regulatory significance, which could operate through allosteric responses and involve the binding sites on each of the protomers present in the IscS dimer.

CONCLUSIONS

In conclusions, we have in this review gone through a methodology (**Table 1**), which has allowed us to gain information on a 110 kDa complex with hybrid techniques. The method can in principle be applied also to larger complexes. The most successful cases are anyway those which involve an appreciable charge of shape of the complex, leading to a clear difference of the SAXS envelop between the isolated components and the complex. Limitations are currently dictated by the number of restraints available and by their distance tolerance: restraints which can allow a tolerance of more than 11 Å, as it is the case for cross-linking studies, are informative but only if several distances are available. It would also be useful to develop HADDOCK and other software to deal with the specific problems of hybrid methods. Some attempts along this line have already been made but more effort would be welcome in the future. It appears particularly useful, in a future perspective, to flank NMR and SAXS studies to other techniques, such as fluorescence, isothermal calorimetry, AUC and cross-linking to obtain more complete and complementary information. As a word of caution though, very good care should anyway be paid to the validation of the results. False positives can be easily obtained if assuming the presence of the wrong species in solution. It remains nevertheless clear that hybrid methods have open a new perspective to the size and complexity of the complexes which can be studied by Structural Biology and, more importantly to the possibility of tackle not only stable and rigid assemblies but also weakly interacting dynamical machines.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The work was supported by MRC (U117584256).

REFERENCES

Adamec, J., Rusnak, F., Owen, W. G., Naylor, S., Benson, L. M., Gacy, A. M., et al. (2000). Iron-dependent self-assembly of recombinant yeast frataxin: implications for Friedreich ataxia. *Am. J. Hum. Genet.* 67, 549–562. doi: 10.1086/303056

Adinolfi, S., Iannuzzi, C., Prischi, F., Pastore, C., Iametti, S., Martin, S. R., et al. (2009). Bacterial frataxin CyaY is the gatekeeper of iron-sulfur cluster formation catalyzed by IscS. *Nat. Struct. Mol. Biol.* 16, 390–396. doi: 10.1038/nsmb.1579

Agar, J. N., Krebs, C., Frazzon, J., Huynh, B. H., Dean, D. R., and Johnson, M. K. (2000). IscU as a scaffold for iron-sulfur cluster biosynthesis: sequential assembly of [2Fe-2S] and [4Fe-4S] clusters in IscU. *Biochemistry* 39, 7856–7862. doi: 10.1021/bi000931n

Babcock, M., de Silva, D., Oaks, R., Davis-Kaplan, S., Jiralerspong, S., Montermini, L., et al. (1997). Regulation of mitochondrial iron accumulation by Yfh1p, a putative homolog of frataxin. *Science* 276, 1709–1712. doi: 10.1126/science.276.5319.1709

Blanchet, C. E., Spilotros, A., Schwemmer, F., Graewert, M. A., Kikhney, A., Jeffries, C. M., et al. (2015). Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *J. Appl. Crystallogr.* 48, 431–443. doi: 10.1107/S160057671500254X

Bloembergen, N., Purcell, E. M., and Pound, R. V. (1948). Relaxation effects in nuclear magnetic resonance absorption. *Phys. Rev.* 73, 679–746. doi: 10.1103/PhysRev.73.679

Bou-Abdallah, F., Adinolfi, S., Pastore, A., Laue, T. M., and Dennis Chasteen, N. (2004). Iron binding and oxidation kinetics in frataxin CyaY of *Escherichia coli*. *J. Mol. Biol.* 341, 605–615. doi: 10.1016/j.jmb.2004.05.072

Campuzano, V., Montermini, L., Lutz, Y., Cova, L., Hindelang, C., Jiralerspong, S., et al. (1997). Frataxin is reduced in Friedreich ataxia patients and is associated with mitochondrial membranes. *Hum. Mol. Genet.* 6, 1771–1780. doi: 10.1093/hmg/6.11.1771

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., et al. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 271, 1423–1427. doi: 10.1126/science.271.5254.1423

Carrara, M., Prischi, F., Nowak, P. R., and Ali, M. M. (2015). Crystal structures reveal transient PERK luminal domain tetramerization in endoplasmic reticulum stress signaling. *EMBO J.* 34, 1589–1600. doi: 10.15252/embj.201489183

Chandramouli, K., and Johnson, M. K. (2006). HscA and HscB stimulate [2Fe-2S] cluster transfer from IscU to apoferredoxin in an ATP-dependent reaction. *Biochemistry* 45, 11087–11095. doi: 10.1021/bi061237w

Cho, S. J., Lee, M. G., Yang, J. K., Lee, J. Y., Song, H. K., and Suh, S. W. (2000). Crystal structure of *Escherichia coli* CyaY protein reveals a previously unidentified fold for the evolutionarily conserved frataxin family. *Proc. Natl. Acad. Sci. U.S.A.* 97, 8932–8937. doi: 10.1073/pnas.160270897

Correia, A. R., Ow, S. Y., Wright, P. C., and Gomes, C. M. (2009). The conserved Trp155 in human frataxin as a hotspot for oxidative stress related chemical modifications. *Biochem. Biophys. Res. Commun.* 390, 1007–1011. doi: 10.1016/j.bbrc.2009.10.095

Cupp-Vickery, J. R., Urbina, H., and Vickery, L. E. (2003). Crystal structure of IscS, a cysteine desulfurase from *Escherichia coli*. *J. Mol. Biol.* 330, 1049–1059. doi: 10.1016/S0022-2836(03)00690-9

David, G., and Perez, J. (2009). Combined sampler robot and high-performance liquid chromatography: a fully automated system for biological small-angle X-ray scattering experiments at the Synchrotron SOLEIL SWING beamline. *J. Appl. Crystallogr.* 42, 892–900. doi: 10.1107/S0021889809029288

Delaforge, E., Milles, S., Bouvignies, G., Bouvier, D., Boivin, S., Salvi, N., et al. (2015). Large-scale conformational dynamics control H5N1 influenza polymerase PB2 binding to importin α. *J. Am. Chem. Soc.* 137, 15122–15134. doi: 10.1021/jacs.5b07765

Deshmukh, L., Schwieters, C. D., Grishaev, A., Ghirlando, R., Baber, J. L., and Clore, G. M. (2013). Structure and dynamics of full-length HIV-1 capsid protein in solution. *J. Am. Chem. Soc.* 135, 16133–16147. doi: 10.1021/ja406246z

di Maio, D., Chandramouli, B., Yan, R., Brancato, G., and Pastore, A. (2017). Understanding the role of dynamics in the iron sulfur cluster molecular machine. *Biochim. Biophys. Acta.* 1861, 3154–3163. doi: 10.1016/j.bbagen.2016.07.020

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125, 1731–1737. doi: 10.1021/ja026939x

Foury, F., and Cazzalini, O. (1997). Deletion of the yeast homologue of the human gene associated with Friedreich's ataxia elicits iron accumulation in mitochondria. *FEBS Lett.* 411, 373–377. doi: 10.1016/S0014-5793(97)00734-5

Franke, D., and Svergun, D. I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* 42(Pt 2), 342–346. doi: 10.1107/S0021889809000338

Gabel, F., Simon, B., and Sattler, M. (2006). A target function for quaternary structural refinement from small angle scattering and NMR orientational restraints. *Eur. Biophys. J. Biophys. Lett.* 35, 313–327. doi: 10.1007/s00249-005-0037-3

Gardner, K. H., and Kay, L. E. (1998). The use of H-2, C-13, N-15 multidimensional NMR to study the structure and dynamics of proteins. *Annu. Rev. Biophys. Biomol. Struct.* 27, 357–406. doi: 10.1146/annurev.biophys.27.1.357

Graewert, M. A., and Svergun, D. I. (2013). Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). *Curr. Opin. Struct. Biol.* 23, 748–754. doi: 10.1016/j.sbi.2013.06.007

Grant, T. D., Luft, J. R., Carter, L. G., Matsui, T., Weiss, T. M., Martel, A., et al. (2015). The accurate assessment of small-angle X-ray scattering data. *Acta Crystallogr. D. Biol. Crystallogr.* 71(Pt 1), 45–56. doi: 10.1107/S1399004714010876

Grishaev, A., Tugarinov, V., Kay, L. E., Trewhella, J., and Bax, A. (2008). Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *J. Biomol. NMR* 40, 95–106. doi: 10.1007/s10858-007-9211-5

Grishaev, A., Wu, J., Trewhella, J., and Bax, A. (2005). Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *J. Am. Chem. Soc.* 127, 16621–16628. doi: 10.1021/ja054342m

Harrison, P. M., and Arosio, P. (1996). The ferritins: molecular properties, iron storage function and cellular regulation. *Biochim. Biophys. Acta* 1275, 161–203. doi: 10.1016/0005-2728(96)00022-9

Huang, J. R., Warner, L. R., Sanchez, C., Gabel, F., Madl, T., Mackereth, C. D., et al. (2014). Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. *J. Am. Chem. Soc.* 136, 7068–7076. doi: 10.1021/ja502030n

Improta, S., Krueger, J. K., Gautel, M., Atkinson, R. A., Lefevre, J. F., Moulton, S., et al. (1998). The assembly of immunoglobulin-like modules in titin: implications for muscle elasticity. *J. Mol. Biol.* 284, 761–777. doi: 10.1006/jmbi.1998.2028

Jeffries, C. M., Graewert, M. A., Blanchet, C. E., Langley, D. B., Whitten, A. E., and Svergun, D. I. (2016). Preparing monodisperse macromolecular samples for successful biological small-angle X-ray and neutron-scattering experiments. *Nat. Protoc.* 11, 2122–2153. doi: 10.1038/nprot.2016.113

Johnson, D. C., Dean, D. R., Smith, A. D., and Johnson, M. K. (2005). Structure, function, and formation of biological iron-sulfur clusters. *Annu. Rev. Biochem.* 74, 247–281. doi: 10.1146/annurev.biochem.74.082803.133518

Kachala, M., Valentini, E., and Svergun, D. I. (2015). Application of SAXS for the Structural Characterization of IDPs. *Adv. Exp. Med. Biol.* 870, 261–289. doi: 10.1007/978-3-319-20164-1_8

Kikhney, A. G., and Svergun, D. I. (2015). A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* 589(19 Pt A), 2570–2577. doi: 10.1016/j.febslet.2015.08.027

Kim, J. H., Bothe, J. R., Alderson, T. R., and Markley, J. L. (2015). Tangled web of interactions among proteins involved in iron-sulfur cluster assembly as unraveled by NMR, SAXS, chemical crosslinking, and functional studies. *Biochim. Biophys. Acta* 1853, 1416–1428. doi: 10.1016/j.bbamcr.2014.11.020

Kim, J. H., Bothe, J. R., Frederick, R. O., Holder, J. C., and Markley, J. L. (2014). Role of IscX in iron-sulfur cluster biogenesis in *Escherichia coli*. *J. Am. Chem. Soc.* 136, 7933–7942. doi: 10.1021/ja501260h

Kim, J. H., Fuzéry, A. K., Tonelli, M., Ta, D. T., Westler, W. M., Vickery, L. E., et al. (2009). Structure and dynamics of the iron-sulfur cluster assembly scaffold protein IscU and its interaction with the cochaperone HscB. *Biochemistry* 48, 6062–6071. doi: 10.1021/bi9002277

Kleckner, I. R., and Foster, M. P. (2011). An introduction to NMR-based approaches for measuring protein dynamics. *Biochim. Biophys. Acta* 1814, 942–968. doi: 10.1016/j.bbapap.2010.10.012

Koutnikova, H., Campuzano, V., Foury, F., Dollé, P., Cazzalini, O., and Koenig, M. (1997). Studies of human, mouse and yeast homologues indicate a mitochondrial function for frataxin. *Nat. Genet.* 16, 345–351. doi: 10.1038/ng0897-345

Krebs, C., Agar, J. N., Smith, A. D., Frazzon, J., Dean, D. R., Huynh, B. H., et al. (2001). IscA, an alternate scaffold for Fe-S cluster biosynthesis. *Biochemistry* 40, 14069–14080. doi: 10.1021/bi015656z

Leidgens, S., De Smet, S., and Foury, F. (2010). Frataxin interacts with Isu1 through a conserved tryptophan in its beta-sheet. *Hum. Mol. Genet.* 19, 276–286. doi: 10.1093/hmg/ddp495

Lill, R. (2009). Function and biogenesis of iron-sulphur proteins. *Nature* 460, 831–838. doi: 10.1038/nature08301

Madl, T., Gabel, F., and Sattler, M. (2011). NMR and small-angle scattering-based structural analysis of protein complexes in solution. *J. Struct. Biol.* 173, 472–482. doi: 10.1016/j.jsb.2010.11.004

Marinoni, E. N., de Oliveira, J. S., Nicolet, Y., Raulfs, E. C., Amara, P., Dean, D. R., et al. (2012). (IscS-IscU)2 Complex structures provide insights into Fe2S2 biogenesis and transfer. *Angew. Chem. Int. Ed.* 51, 5439–5442. doi: 10.1002/anie.201201708

Mathew, E., Mirza, A., and Menhart, N. (2004). Liquid-chromatography-coupled SAXS for accurate sizing of aggregating proteins. *J. Synchrotron Radiat.* 11(Pt 4), 314–318. doi: 10.1107/S0909049504014086

Mertens, H. D., and Svergun, D. I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.* 172, 128–141. doi: 10.1016/j.jsb.2010.06.012

Mettert, E. L., and Kiley, P. J. (2015). How is Fe-S cluster formation regulated? *Annu. Rev. Microbiol.* 69, 505–526. doi: 10.1146/annurev-micro-091014-104457

Milles, S., Mercadante, D., Aramburu, I. V., Jensen, M. R., Banterle, N., Koehler, C., et al. (2015). Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. *Cell* 163, 734–745. doi: 10.1016/j.cell.2015.09.047

Nair, M., Adinolfi, S., Pastore, C., Kelly, G., Temussi, P., and Pastore, A. (2004). Solution structure of the bacterial frataxin ortholog, CyaY: mapping the iron binding sites. *Structure* 12, 2037–2048. doi: 10.1016/j.str.2004.08.012

Nogales, E., and Scheres, S. H. (2015). Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol. Cell* 58, 677–689. doi: 10.1016/j.molcel.2015.02.019

Ollagnier-de-Choudens, S., Mattioli, T., Takahashi, Y., and Fontecave, M. (2001). Iron-sulfur cluster assembly: characterization of IscA and evidence for a specific and functional complex with ferredoxin. *J. Biol. Chem.* 276, 22604–22607. doi: 10.1074/jbc.M102902200

Parsons, L. M., Grishaev, A., and Bax, A. (2008). The periplasmic domain of TolR from Haemophilus influenzae forms a dimer with a large hydrophobic groove: NMR solution structure and comparison to SAXS data. *Biochemistry* 47, 3131–3142. doi: 10.1021/bi702283x

Pastore, C., Adinolfi, S., Huynen, M. A., Rybin, V., Martin, S., Mayer, M., et al. (2006). YfhJ, a molecular adaptor in iron-sulfur cluster formation or a frataxin-like protein? *Structure* 14, 857–867. doi: 10.1016/j.str.2006.02.010

Petoukhov, M. V., and Svergun, D. I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* 89, 1237–1250. doi: 10.1529/biophysj.105.064154

Prischi, F., Konarev, P. V., Iannuzzi, C., Pastore, C., Adinolfi, S., Martin, S. R., et al. (2010a). Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. *Nat. Commun.* 1:95. doi: 10.1038/ncomms1097

Prischi, F., and Pastore, A. (2016). Application of nuclear magnetic resonance and hybrid methods to structure determination of complex systems. *Adv. Exp. Med. Biol.* 896, 351–368. doi: 10.1007/978-3-319-27216-0_22

Prischi, F., Pastore, C., Carroni, M., Iannuzzi, C., Adinolfi, S., Temussi, P., et al. (2010b). Of the vulnerability of orphan complex proteins: the case study of the *E. coli* IscU and IscS proteins. *Protein Expr. Purif.* 73, 161–166. doi: 10.1016/j.pep.2010.05.003

Puglisi, R., Yan, R., Adinolfi, S., and Pastore, A. (2016). A New Tessera into the interactome of the isc operon: a novel interaction between HscB and IscS. *Front. Mol. Biosci.* 3:48. doi: 10.3389/fmolb.2016.00048

Putnam, C. D., Hammel, M., Hura, G. L., and Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* 40, 191–285. doi: 10.1017/S0033583507004635

Ramelot, T. A., Cort, J. R., Goldsmith-Fischman, S., Kornhaber, G. J., Xiao, R., Shastry, R., et al. (2004). Solution NMR structure of the iron-sulfur cluster assembly protein U (IscU) with zinc bound at the active site. *J. Mol. Biol.* 344, 567–583. doi: 10.1016/j.jmb.2004.08.038

Roberts, G. C. K. (1993). *NMR of macromolecules : a practical approach*. Oxford: Oxford University Press.

Rötig, A., de Lonlay, P., Chretien, D., Foury, F., Koenig, M., Sidi, D., et al. (1997). Aconitase and mitochondrial iron-sulphur protein deficiency in Friedreich ataxia. *Nat. Genet.* 17, 215–217. doi: 10.1038/ng1097-215

Rouault, T. A. (2015). Mammalian iron-sulphur proteins: novel insights into biogenesis and function. *Nat. Rev. Mol. Cell Biol.* 16, 45–55. doi: 10.1038/nrm3909

Round, A. R., Franke, D., Moritz, S., Huchler, R., Fritsche, M., Malthan, D., et al. (2008). Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. *J. Appl. Crystallogr.* 41(Pt 5), 913–917. doi: 10.1107/S0021889808021018

Sali, A., Berman, H. M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S. K., et al. (2015). Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* 23, 1156–1167. doi: 10.1016/j.str.2015.05.013

Schwartz, C. J., Djaman, O., Imlay, J. A., and Kiley, P. J. (2000). The cysteine desulfurase, IscS, has a major role in *in vivo* Fe-S cluster formation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 97, 9009–9014. doi: 10.1073/pnas.160261497

Schwartz, C. J., Giel, J. L., Patschkowski, T., Luther, C., Ruzicka, F. J., Beinert, H., et al. (2001). IscR, an Fe-S cluster-containing transcription factor, represses expression of *Escherichia coli* genes encoding Fe-S cluster assembly proteins. *Proc. Natl. Acad. Sci. U.S.A.* 98, 14895–14900. doi: 10.1073/pnas.251550898

Shi, R., Proteau, A., Villarroya, M., Moukadiri, I., Zhang, L., Trempe, J. F., et al. (2010). Structural basis for Fe-S cluster assembly and tRNA thiolation mediated by IscS protein-protein interactions. *PLoS Biol.* 8:e1000354. doi: 10.1371/journal.pbio.1000354

Shimomura, Y., Wada, K., Fukuyama, K., and Takahashi, Y. (2008). The asymmetric trimeric architecture of [2Fe-2S] IscU: implications for its scaffolding during iron-sulfur cluster biosynthesis. *J. Mol. Biol.* 383, 133–143. doi: 10.1016/j.jmb.2008.08.015

Sunnerhagen, M., Olah, G. A., Stenflo, J., Forsen, S., Drakenberg, T., and Trewhella, J. (1996). The relative orientation of Gla and EGF domains in coagulation factor X is altered by Ca2+ binding to the first EGF domain. A combined NMR-small angle X-ray scattering study. *Biochemistry* 35, 11547–11559. doi: 10.1021/bi960633j

Svergun, D., Barberato, C., and Koch, M. H. J. (1995). CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28, 768–773. doi: 10.1107/S0021889895007047

Svergun, D. I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* 25, 495–503. doi: 10.1107/S0021889892001663

Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing (vol 76, pg 2879, 1999). *Biophys. J.* 77, 2896–2896.

Takahashi, Y., and Nakamura, M. (1999). Functional assignment of the ORF2-iscS-iscU-iscA-hscB-hscA-fdx-ORF3 gene cluster involved in the assembly of Fe-S clusters in *Escherichia coli*. *J. Biochem.* 126, 917–926. doi: 10.1093/oxfordjournals.jbchem.a022535

Tuukkanen, A. T., and Svergun, D. I. (2014). Weak protein-ligand interactions studied by small-angle X-ray scattering. *FEBS J.* 281, 1974–1987. doi: 10.1111/febs.12772

Venditti, V., Egner, T. K., and Clore, G. M. (2016). Hybrid approaches to structural characterization of conformational ensembles of complex macromolecular systems combining NMR residual dipolar couplings and solution X-ray scattering. *Chem. Rev.* 116, 6305–6322. doi: 10.1021/acs.chemrev.5b00592

Wang, X., Watson, C., Sharp, J. S., Handel, T. M., and Prestegard, J. H. (2011). Oligomeric structure of the chemokine CCL5/RANTES from NMR, MS, and SAXS data. *Structure* 19, 1138–1148. doi: 10.1016/j.str.2011.06.001

Wüthrich, K. (2001). The way to NMR structures of proteins. *Nat. Struct. Biol.* 8, 923–925. doi: 10.1038/nsb1101-923

Yan, R., Adinolfi, S., Iannuzzi, C., Kelly, G., Oregioni, A., Martin, S., et al. (2013a). Cluster and fold stability of *E. coli* ISC-type ferredoxin. *PLoS ONE* 8:e78948. doi: 10.1371/journal.pone.0078948

Yan, R., Adinolfi, S., and Pastore, A. (2015). Ferredoxin, in conjunction with NADPH and ferredoxin-NADP reductase, transfers electrons to the IscS/IscU complex to promote iron-sulfur cluster assembly. *Biochim. Biophys. Acta Proteins Proteomics* 1854, 1113–1117. doi: 10.1016/j.bbapap.2015.02.002

Yan, R., Konarev, P. V., Iannuzzi, C., Adinolfi, S., Roche, B., Kelly, G., et al. (2013b). Ferredoxin competes with bacterial frataxin in binding to the desulfurase IscS. *J. Biol. Chem.* 288, 24777–24787. doi: 10.1074/jbc.M113.480327

Yoon, T., and Cowan, J. A. (2003). Iron-sulfur cluster biosynthesis. Characterization of frataxin as an iron donor for assembly of [2Fe-2S] clusters in ISU-type proteins. *J. Am. Chem. Soc.* 125, 6078–6084. doi: 10.1021/ja027967i

Zheng, L., Cash, V. L., Flint, D. H., and Dean, D. R. (1998). Assembly of iron-sulfur clusters. Identification of an iscSUA-hscBA-fdx gene cluster from *Azotobacter vinelandii*. *J. Biol. Chem.* 273, 13264–13272. doi: 10.1074/jbc.273.21.13264

Zuiderweg, E. R. (2002). Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* 41, 1–7. doi: 10.1021/bi011870b

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 12 December 2017
doi: 10.3389/fmolb.2017.00085

Chlorophyll-Derivative Modulation of Rhodopsin Signaling Properties through Evolutionarily Conserved Interaction Pathways

Kristina N. Woods[1]*, *Jürgen Pfeffer*[2] *and Judith Klein-Seetharaman*[3]

[1] Lehrstuhl für BioMolekulare Optik, Ludwig-Maximilians-Universität, München, Germany, [2] Bavarian School of Public Policy, Technical University of Munich, München, Germany, [3] Warwick Medical School, University of Warwick, Coventry, United Kingdom

Retinal is the light-absorbing chromophore that is responsible for the activation of visual pigments and light-driven ion pumps. Evolutionary changes in the intermolecular interactions of the retinal with specific amino acids allow for adaptation of the spectral characteristics, referred to as spectral tuning. However, it has been proposed that a specific species of dragon fish has bypassed the adaptive evolutionary process of spectral tuning and replaced it with a single evolutionary event: photosensitization of rhodopsin by chlorophyll derivatives. Here, by using a combination of experimental measurements and computational modeling to probe retinal-receptor interactions in rhodopsin, we show how the binding of the chlorophyll derivative, chlorin-e6 (Ce6) in the intracellular domain (ICD) of the receptor allosterically excites G-protein coupled receptor class A (GPCR-A) conserved long-range correlated fluctuations that connect distant parts of the receptor. These long-range correlated motions are associated with regulating the dynamics and intermolecular interactions of specific amino acids in the retinal ligand-binding pocket that have been associated with shifts in the absorbance peak maximum (λ_{max}) and hence, spectral sensitivity of the visual system. Moreover, the binding of Ce6 affects the overall global properties of the receptor. Specifically, we find that Ce6-induced dynamics alter the thermal stability of rhodopsin by adjusting hydrogen-bonding interactions near the receptor active-site that consequently also influences the intrinsic conformational equilibrium of the receptor. Due to the conservation of the ICD residues amongst different receptors in this class and the fact that all GPCR-A receptors share a common mechanism of activation, it is possible that the allosteric associations excited in rhodopsin with Ce6 binding are a common feature in all class A GPCRs.

Keywords: Terahertz spectroscopy, protein allostery, small-molecule allosteric agonist, Chlorine Compounds, protein dynamics

OPEN ACCESS

Edited by:
Isabelle Callebaut,
Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Simone Brogi,
University of Siena, Italy
Steffen Wolf,
Albert Ludwigs University of Freiburg, Germany

****Correspondence:***
Kristina N. Woods
kristina.woods@physik.lmu.de

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *20 February 2017*
Accepted: *28 November 2017*
Published: *12 December 2017*

Citation:
Woods KN, Pfeffer J and Klein-Seetharaman J (2017) Chlorophyll-Derivative Modulation of Rhodopsin Signaling Properties through Evolutionarily Conserved Interaction Pathways. Front. Mol. Biosci. 4:85. doi: 10.3389/fmolb.2017.00085

INTRODUCTION

The dimmest habitats on earth appear at night and in the depths of the ocean (Warrant, 2004). The greatest challenge for vision in these habitats is capture of photons, and the way these photons are post-processed. The primary photoreceptor in eyes, rhodopsin, is a major target for adaptation to different light conditions. Rhodopsin as the prototypical member of the GPCR-A family adopts an overall organization of seven transmembrane (TM) helices that form a bundle. Within the GPCR family there is a large sub-group of opsins, representing opsin sequences in different photoreceptor cell types and organisms. All opsins covalently bind retinal, a vitamin A derivative, at the interface between the transmembrane (TM) and extracellular domains (ECDs). Visual signal transduction is initiated by photon-induced isomerization of *11-cis* retinal to *all-trans* retinal. This event is sensed by the TM domain which undergoes a conformational change that results in the activated state, Metarhodopsin II (Meta-II) via Meta-I. Meta-II—unlike dark state rhodopsin—binds and activates the G protein, transducin G_t, ultimately leading to receptor hyperpolarization. Signal desensitization is initiated by phosphorylation of the C-terminus of rhodopsin by rhodopsin kinase, followed by binding of arrestin, preventing further binding of G_t to rhodopsin. *In vitro*, Meta-II decays to opsin and free retinal, the half-life of which depends on the lipid environment (Farrens and Khorana, 1995). In addition, a storage form of rhodopsin, Meta-III, emerges in parallel with Meta-II, from the Meta-I state. The Meta-III state of rhodopsin—unlike Meta-II—has a protonated retinal Schiff base and decays into opsin and free retinal on significantly longer time scales (Heck et al., 2003b,a; Vogel et al., 2003, 2004; Stehle et al., 2014).

One mechanism of adaptation to dim habitats is spectral tuning. Spectral tuning refers to adjustment of the absorbance maxima in the spectra of the photoreceptors. Spectral tuning can be hard-wired by genetic variation of the rhodopsin sequences changing the interactions between the retinal and the protein ("opsin shift") (Nathans, 1990). This allows adaptation to the wavelengths that are maximally transmitted under different environmental conditions, such as sun-light in the shade, sun-light penetrating water, moon-light or bioluminescence generated by deep-sea fish (Douglas et al., 1998; Fishkin et al., 2004). Genetic adaptation is also the mechanism by which cone and rod opsins absorb at different wavelengths (Douglas et al., 1999). However, spectral tuning can also be achieved by interaction of the rhodopsins with small molecules. In xanthorhodopsin, a bacteriorhodopsin-like proton pump in the halophilic eubacterium *Salinibacter ruber*, a carotenoid is bound in addition to retinal (Balashov et al., 2005). Light energy absorbed by the carotenoid is transferred to the retinal with a quantum efficiency of ∼40% (Balashov et al., 2005) and light is funneled to the retinal similar to the photosynthetic light harvesting complex. In the deep-sea fish *Malacosteus niger* (dragon fish), porphyrin/chlorophyll derivatives, including chlorin e6 (Ce6), may act as photosensitizers rendering its rhodopsin sensitive to wavelengths that other deep-sea fish respond to as a result of genetic adaptations of their retinal binding pockets or retinal replacement (Douglas et al., 1998, 1999; Isayama et al., 2006). An investigation of different chlorophyll derivatives in rod outer segments on bleaching rates of the chromophore by red light highlighted Ce6 as the molecule with the strongest photosensitizing effect on bovine rhodopsin (Washington et al., 2004). The hypothesis of a binding pocket for Ce6 and energy transfer to retinal resembling that of light harvesting in photosynthesis was proposed (Washington et al., 2004; Isayama et al., 2006). The photosensitizing effects of chlorophyll-derivatives observed in salamander and deep-sea fish can be reproduced qualitatively using bovine rhodopsin, *in vitro* (Washington et al., 2004). Furthermore, electroretinogram recordings in mice injected with Ce6 suggested that the sensitivity of rhodopsin can be broadened to blue and red light in the presence of Ce6 (Washington et al., 2007). Indeed, the photosensitizing effect of Ce6 and other porphyrin compounds in vision has been reported as a side-effect during photodynamic therapy (Kimura, 1987).

To understand by what Ce6 binding enhances bleaching rates of rhodopsin in red light, we used a novel approach which we recently developed to probe dynamics and allostery in rhodopsin (Woods et al., 2016): through a combination of molecular dynamics simulations, evolutionary sequence analysis and Terahertz (THz) spectroscopy we show that Ce6 binding excites evolutionarily conserved communication pathways in rhodopsin that establish connections between the ligand-binding site and the rest of the receptor.

RESULTS

Experimental Detection of the Conformational Ensemble Dynamics and Intermolecular Changes in Rhodopsin When Bound with the Allosteric Modulator Ce6

Global Fluctuations of Dark-State Rhodopsin Bound with Ce6

The <100 cm^{-1} region of the infrared spectrum is sensitive to global, internal fluctuations that describe the intrinsic dynamics of the receptor. These globally, correlated thermal fluctuations provide a mechanism for sampling the ensemble of conformations that comprise the free energy landscape of possible receptor conformations (Frauenfelder et al., 1991). Hence, the modes detected in the experimental THz spectrum in this region provide direct information about the sampling of conformational substates in rhodopsin. An inspection of the low frequency modes of dark-state rhodopsin in the <100 cm^{-1} spectral region in **Figure 1A** in the presence and absence of Ce6 reveals that the region of the spectrum above 50 cm^{-1} is dramatically altered when contrasting the two states of the receptors. For instance, in the Ce6-bound receptor in **Figure 1B** there is a prominent peak at 80 cm^{-1} and a general increase in the absorption peak intensity at a frequency above 50 cm^{-1} when contrasted with the dark-state receptor that is not bound with Ce6 (**Figure 1** and Woods et al., 2016). The differences in the spectra of the two receptors suggest that binding of Ce6 alters the conformational ensemble dynamics in rhodopsin.

In our previous work on the inactive-receptor in the unbound state (Woods et al., 2016), we have found there is an equilibrium of both inactive and active-state protein conformational fluctuations in the dark-state protein. This conformational heterogeneity in the inactive receptor indicates that rhodopsin samples a diverse set of functional structures even before any activation event has taken place. In particular, the heterogeneity of global structural fluctuations detected experimentally is

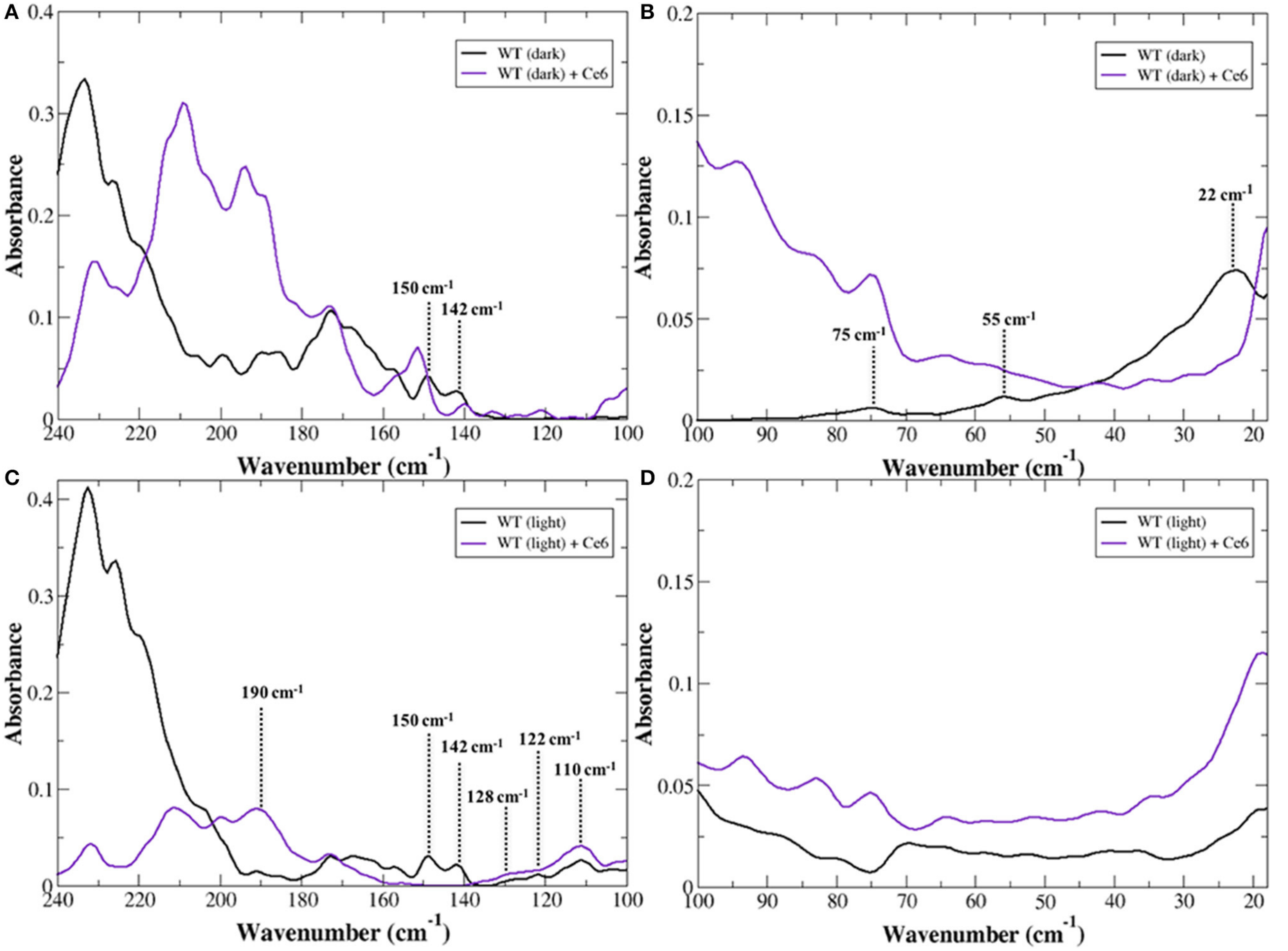


FIGURE 1 | Experimental THz spectrum of dark-state rhodopsin in the unbound state (black line) and when bound with Ce6 (purple line) in the **(A)** 240–100 cm^{-1} spectral region and in the **(B)** 100–20 cm^{-1}, spectral region. Experimental THz spectrum of Meta-II in the unbound state (black line) and when bound with Ce6 (purple line) in the **(C)** 240–100 cm^{-1} spectral region and in the **(D)** 100–20 cm^{-1} spectral region.

intimately tied with multiple *pre-existing* allosteric associations in the inactive receptor. Peaks at approximately 75 and 55 cm^{-1} in the experimental spectrum of the unbound receptor were identified as rhodopsin global fluctuations in an inactive-type conformation. Specifically, the peak at 55 cm^{-1} was found to be associated with a retinal (polyene chain) torsional fluctuation that is coupled with a protein, global backbone torsion, whereas the peak at 75 cm^{-1} is associated with a retinal torsional oscillation that is coupled with collective out-of-plane protein side-chain fluctuations (Woods et al., 2016). In addition to the inactive-state conformational fluctuations, we also uncovered a weaker band at approximately 40 cm^{-1} in the experimental spectrum that was later found to be associated with transient interactions of receptor amino acid side-chains that supported a more active-like rhodopsin conformation. These retinal-induced transient interactions of the more active-like receptor conformation were hypothesized to serve as a necessary precursor in the mechanism that eventually leads to the active-state receptor. Therefore, the enhancement of the ~75 cm^{-1} mode in the Ce6-bound receptor suggests that Ce6 stabilizes a specific conformational state of rhodopsin—the ground state.

Ce6-Induced Intra- and Inter-Protein Interaction Networks in Dark-State Rhodopsin

In the 100–250 cm^{-1} region of the experimental spectrum we detect motions in rhodopsin that reveal more localized intermolecular interactions (Woods, 2014b) such as interhelical contacts as well as helical interactions with the solvent. In general, peaks in the experimental spectrum in the 100–160 cm^{-1} are related to protein intra- and intermolecular interactions. For example in **Figure 1A**, the peaks at approximately 150 and 140 cm^{-1} in the dark-state spectrum of the unbound receptor were found to be associated with interhelical and solvent-induced H-bonding interactions in our earlier study (Woods et al., 2016), while peaks ≥170 cm^{-1} are predominately associated with solvent-solvent interactions of water molecules in the receptor hydration shell. A comparison of dark-state rhodopsin in the unbound and unbound state in **Figure 1A** suggests that binding of Ce6 alters the inter- and intra-protein interactions in rhodopsin but has a much stronger effect on an extensive network of solvent H-bonds that stabilize the ground state of the receptor (Woods, 2014a).

MD simulation of Ce6-Induced Long-Range Correlated Fluctuations and the Effect on Dark-State Rhodopsin Global Motions

Allostery in proteins enables the activity of one site in a protein to modulate function at another spatially distinct region (Hawkins and McLeish, 2006; Fenwick et al., 2011; Motlagh et al., 2014). Recent experimental and computation investigations on a number of proteins and enzymes have demonstrated that allosteric signal transmission is mediated by protein local structural fluctuations (Whitten et al., 2005; Daily and Gray, 2007; Pandini et al., 2013). To assess if Ce6 allosterically affects retinal-protein interactions in rhodopsin, we carried out MD simulations of dark-state bound rhodopsin in the presence and absence of Ce6. We find that Ce6 binds only weakly on the cytoplasmic surface (Supplementary text, *Ce6–ligand binding affinity* and Figures S2, S3) of the receptor. Experimental fluorescence measurements of Ce6 binding to rhodopsin also reveal micromolar binding affinity (unpublished results). Despite the weak binding, Ce6 has a strong effect on the long-range correlated fluctuations of the bound state when contrasted with the unbound receptor. This can be deduced from our analysis of the MD simulation induced localized structural fluctuations (LSFs) and the consequent collective dynamics in the receptor that arise from the addition of Ce6 to rhodopsin. For instance, **Figure 2** reveals the development of long-range correlated fluctuations in dark-state rhodopsin with Ce6 that significantly involve contributions from the intracellular loops close to where Ce6 binding takes place. Particularly, Ce6-induced long-range correlated fluctuations involving intracellular loop 2 (CL2) and extracellular loop 2 (EL2) are coupled with a structurally conserved collection of aromatic and polar residues

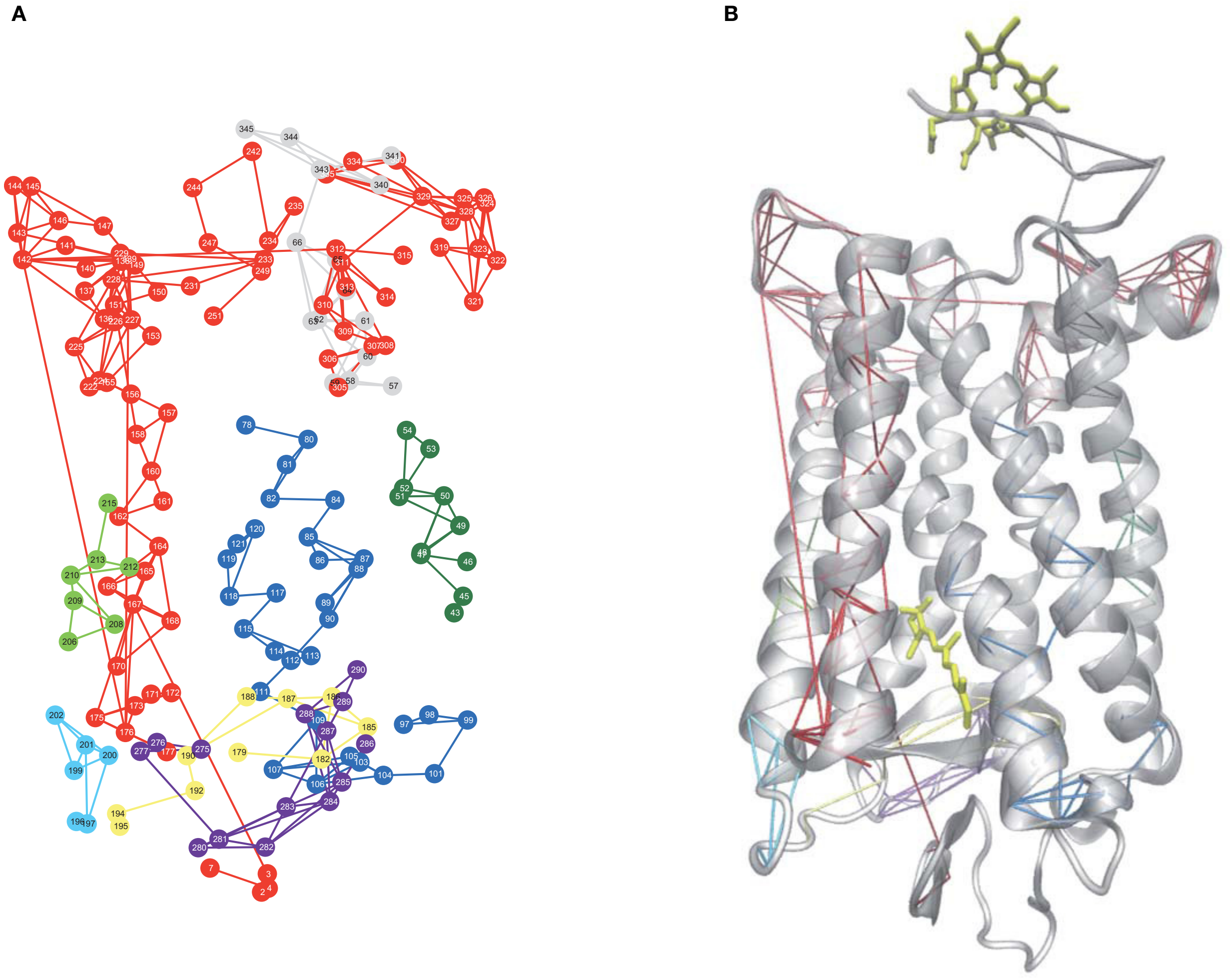


FIGURE 2 | (A) A 2-D network mapping of the LSFs from the MD simulation of dark-state rhodopsin when bound to Ce6 and **(B)** a cartoon representation of rhodopsin showing the mapping of the LSFs from **(A)** onto the protein 3-D structure.

at the extracellular end of H4 (Pro170, Pro171, Tyr175, Ser176, Arg177, and Tyr178) leading from the ligand-binding pocket to the EC domain. This group of GPCR-A wide conserved long-distance fluctuations modulates the H-bonding environment of residues and solvent molecules lining the retinal ligand-binding pocket (**Figure 2B**). Specifically, side-chain fluctuations of residues such as Glu122, Trp126, and Phe261 in Ce6 rhodopsin have increased (H-bonding) interactions with the retinal β-ionone ring due to the CL2–EL2 correlation when contrasted with unbound receptor (Figure S1b). The Ce6-induced correlated fluctuations reorient Glu122 so that the there is a rearrangement of the H-bonding network surrounding the ring such that it stabilizes the retinal in a defined conformation (**Figure 2** and Figure S4). In our previous work (Woods et al., 2016) we have conjectured that the ability of the retinal to assume multiple conformational states is intertwined with the formation of multiple, distinct signaling pathways in both the inactive and active receptor. In line with what was observed in the experimental detection of the global modes of the dark-state receptor in **Figure 1**, we find from the MD analysis of the Ce6-induced structural fluctuations in rhodopsin, that binding of the allosteric modulator assists in stabilizing the receptor in a specific conformation, namely the ground state because of the modified electrostatic interactions in the vicinity of the retinal ring that accompany Ce6-binding. This conclusion is further supported by a mapping of the low-frequency torsional dynamics of the retinal from the dark-state MD simulations of rhodopsin in both the unbound and bound states. The absence of the 40 cm^{-1} mode in the MD retinal torsional spectrum of Ce6-bound rhodopsin in **Figure 3A** clearly shows that Ce6 stabilizes only a subset of the receptor conformations. Both the 80 and 65 cm^{-1} represent the effect of retinal torsional dynamics from interactions with the inactive-like conformation of rhodopsin, whereas the 40 cm^{-1} mode reflects the influence on the retinal from protein interactions in a more active-like conformation (Woods et al., 2016).

Ce6-induced long-distance induced fluctuations involving CL1 and CL3 are instrumental in altering the interhelical packing within the receptor core. Correlated fluctuations involving these two loops modulate the dynamics of TM1and TM2 (**Figure 2**) and subsequently destabilize a conserved cluster of water molecules near Gly90 in helix 2 that maintain the shape of the ligand-binding pocket. In particular, the Ce6-induced distortion of the binding pocket weakens interactions in the ligand-binding region that stabilize residues forming the receptor hydrophobic core such as Gly114, Ala117, Thr118, and Gly120–Glu122 and concurrently, also disrupts a conserved network (Angel et al., 2009) of water molecules near Gly90 in helix 2 that stabilizes the Glu113 salt bridge of the protonated Schiff base (PSB). These Ce6-induced changes within the receptor ligand-binding pocket allow Glu113 to move further away from the PSB. This perturbation of the PSB H-bonding network increases the amplitude of the dynamic structural fluctuations of the counterion and affects the thermal stability of the entire receptor. The reduced electrostatic interactions from residues surrounding the PSB, resulting from the movement of Glu113 away from the PSB, decreases the energy gap between the ground and excited state of the receptor (Rajamani et al., 2011). The reduction of the energy difference between the two states also reduces the barrier for activation; hence, it reduces receptor thermal stability (Figure S4). Numerous previous studies on rhodopsin (Lin et al., 1998; Rajamani et al., 2011; Imamoto and Shichida, 2014) have also shown that the reduction in the energy difference between the ground and excited state reduces the thermal stability of the receptor and is also directly correlated with the red-shift of the λ_{max} of the receptor. Although, it is important to point out that previous thermal denaturing studies (Balem et al., 2009) on inactive-state rhodopsin found an increase in the helical content of the receptor when bound to Ce6. In other words, the helical regions of the receptor are stabilized in the dark-state receptor when bound to Ce6 (when contrasted with the unbound receptor). These results are in direct contrast to both the MD simulation studies and (THz) experimental studies carried out in this investigation. Although, the disparity in interpretations could be a consequence of what the distinctive experimental methods measure. The melting data is primarily sensitive to changes in the helical content of the receptor, whereas the THz

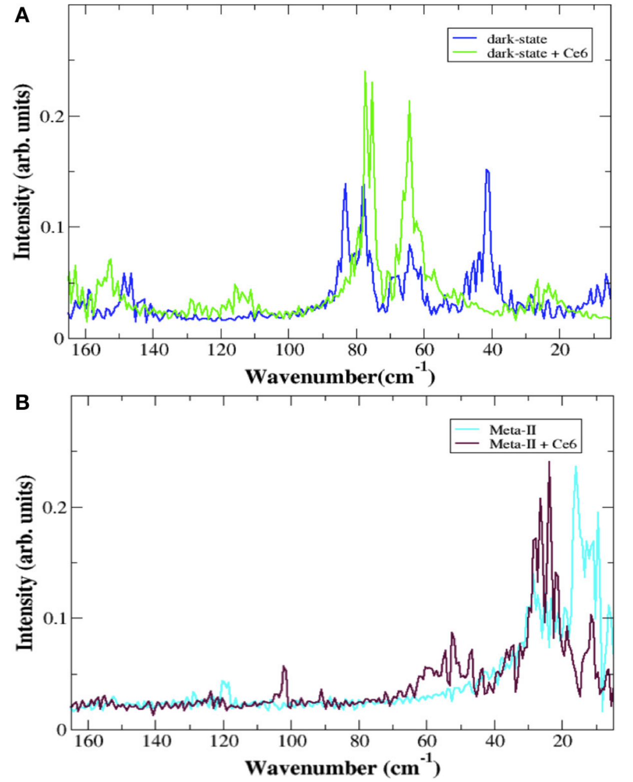


FIGURE 3 | The torsional spectrum of the retinal from rhodopsin MD simulations in **(A)** the dark-state (blue line) and in the dark-state when bound to Ce6 (green line) and **(B)** in Meta-II (cyan line) and Meta-II when bound to Ce6 (purple line). The peaks at 15 and 25 cm^{-1} in **(B)** are related to torsional fluctuations of the C-20 methyl group near the terminus of the polyene chain and a collective chain-twisting oscillation, respectively in Meta-II whereas the peak close to 60 cm^{-1} in **(B)** is associated with a chain torsion coupled with a retinal ring bending motion of a more inactive-type rhodopsin structure. In all cases, the torsion of the retinal is defined by the angle created by the C5-, C9-, and C13- methyl groups.

data ($\leq$100 cm^{-1}) is sensitive to the global changes in the receptor as a whole. The initial findings from this investigation suggest that the principal changes in thermal stability arise from modifications around the ligand-binding site in the Ce6-bound receptor and these alterations influence the global stability of the receptor.

The deformation of the retinal ligand-binding pocket due to the Ce6-induced correlated fluctuations of CL1 and CL3 is further stabilized by the rearrangement of conserved network of water molecules in EL2. An analysis of the CL1-CL3 Ce6-induced correlated dynamics in **Figure 2** also reveals alterations in solvent-protein interactions that include the water molecules shared between the side-chains of Glu181 and Ser186 that directly connect the dynamics of the EL2 with the retinal binding-pocket (Figure S5). The CL1-CL3 coordinated dynamics shift Glu181 further away from Ser186 by means of the introduction of an additional water molecule into the retinal binding pocket that also links the backbone atoms of Glu181 and Ser186 (while retaining the water coordinated side-chain linkage of the two residues). This results in the overall rearrangement of EL2 through the correlated dynamics of helix 3 via the conserved Cys110-Cys187 disulfide bond. The correlated movement shifts Ile189 (as well as the entire β_4 loop of EL2) in the direction of the receptor N-terminus. The end result of the altered Ce6-induced dynamics of dark-state rhodopsin is a more open ligand-binding pocket that is created by the larger separation between the retinal and extracellular region of the receptor. The increased distance between Ile189 and the retinal disrupts the H-bonding network of the PSB but in doing so, would also likely lead to a decrease in receptor thermal stability (associated with higher levels of dark-noise) as well as increase the rate of hydrolysis of the PSB. In fact, earlier experimental studies (Imamoto and Shichida, 2014; Yanagawa et al., 2015) focusing on the thermal activation rate of rhodopsin have revealed a direct correlation between the thermal stability of the dark-state of the receptor and that of the lifetime of the active-state intermediate Meta-II. In this regard, both Glu122 and Ile189 were identified as two residues that play a crucial role in suppressing thermal fluctuations in the retinal-binding pocket in rhodopsin, which ultimately imparts the low dark noise characteristic of the receptor.

It is also interesting to return to the experimental detection of the global modes of the dark-state receptor (**Figure 1A**) and to consider the observed changes in the conformational ensemble dynamics that were detected with Ce6-binding. Previous experimental studies probing the conformational stability of Ce6 in (dark-state) rhodopsin have indicated that Ce6 stabilizes the helices of the receptor. This earlier work is in line with the experimentally detected population shift in the conformational ensemble dynamics of the Ce6-bound receptor (when contrasted with the unbound receptor) that we have detected in this investigation on rhodopsin (**Figure 1A**). But interestingly, is in direct contrast to the identified changes in the ligand-binding pocket dynamics of the bound receptor that we have uncovered from MD simulation. In the latter case, the Ce6-induced dynamics appears to create a closer potential energy surface between the ground and excited state (Ala-Laurila et al., 2004; Hofmann and Palczewski, 2015) of the retinal in Ce6-rhodopsin. This reduction in the energy difference between the two states would, in principal, create a thermally unstable ground state that consequently would shift the equilibrium toward the excited-state of the receptor. The disparity in interpretations that arise when examining the local dynamics of the retinal-binding pocket with approaches that map the global characteristics of the receptor suggests that induced long-range structural fluctuations may also play an important role in the Ce6 mechanism of spectral tuning in rhodopsin.

Experimental Detection of the Induced Changes in Global Dynamics and the Allosterically Coupled Motions Associated with Activation in Meta-II Rhodopsin When Bound by Ce6

Global Dynamics and Thermal Stability in Meta-II-Ce6

The global motions of Meta-II rhodopsin in the presence of Ce6 are dramatically altered as compared to dark rhodopsin with and without Ce6 in **Figure 1D**. In contrast to the dark-state unbound spectrum in **Figure 1B**, there are no clearly resolved vibrational bands in the <100 cm^{-1} region of the experimental spectrum. Based on previous work on rhodopsin (Woods et al., 2016), this likely indicates that the major protein modes excited after isomerization are red-shifted to very low frequencies. Despite the lack of spectral features, one can still deduce information about the stability of the receptor states based on the general shape of the spectrum. For instance, it is apparent that the Ce6-bound Meta-II state in **Figure 1C** is far more oscillatory above 70 cm^{-1} when contrasted with the unbound Meta-II state. The oscillatory nature of the Meta-II-Ce6 spectrum indicates that there is instability in the internal modes of the receptor. Moreover, the increased instability detected in the global modes of the Ce6-bound receptor also implies a general decrease in the thermal stability of the Ce6-bound activated-state of the receptor.

Experimental Detection of Alterations in the Allosteric Interaction Network Associated with Activation in Meta-II-Ce6

An inspection of the higher frequency spectra of the Meta-II states of rhodopsin in **Figure 1C** strongly suggests that there are major modifications in the interhelical interactions in Meta-II-Ce6 when compared with the unbound active-state receptor. For example, the absence of the ~140 and 150 cm^{-1} modes in the Ce6 receptor bound spectrum suggests that both the interhelical packing and the solvent-protein interactions in the bound receptor are dramatically altered in Meta-II-Ce6. Both modes (at 140 and 150 cm^{-1}) are strongly influenced by retinal torsional dynamics associated with both the β-ionone ring and the polyene chain, indicating that the coupling between the retinal ligand and its immediate protein environment is considerably changed in the Ce6-activated state. In our earlier study (Woods et al., 2016) of Meta-II rhodopsin we determined that the vibrational modes at approximately 110, 120, and 130 cm^{-1} are spectral markers of the water-mediated pathway of activation in Meta-II that connects the dynamics in the ligand-binding pocket with the G-protein binding site in the ICD. The vibrational modes represent anharmonic, solvent-mediated fluctuations of protein backbone and side-chain atoms that form allosteric sites in Meta-II. The water-mediated allosteric sites create a coherent, defined pathway of signal propagation in Meta-II that connect the dynamics of the retinal taking place in the ligand-binding pocket with intermolecular interactions taking place in the receptor intracellular domain. The fact that both spectra in **Figure 1C** have prominent vibrational bands at about 110, 120, and 130 cm^{-1} implies that activation in the Ce6 bound receptor still takes place, but based on the other prominent differences in the spectrum of the distinct Meta-II states, it is likely that the activation mechanism in Meta-II-Ce6 is somehow altered. One clue of how the pathway of activation in Meta-II-Ce6 may differ from Meta-II can be discerned from the solvent—solvent interaction region of the experimental spectrum (>170 cm^{-1}) in **Figure 1C**. A comparison of the spectra makes it readily apparent that solvent interaction network in Meta-II is disrupted with Ce6 binding. The dramatic drop in intensity in Meta-II-Ce6 in the 200–240 cm^{-1} spectral region suggests that the addition of the allosteric modulator (Ce6) disrupts the solvent H-bonding network that both stabilizes and aids in the efficiency of the active Meta-II intermediate binding with the G-protein. Although interestingly, in the same solvent-interaction region, there is a new peak in the Meta-II-Ce6 spectrum centered at 190 cm^{-1} that possibly hints at a new set of solvent interactions that support an alternative pathway of activation in the Ce6 bound receptor. In previous THz studies on globular proteins (Woods, 2010, 2014a) we have identified an equivalent peak at approximately 190 cm^{-1} in the protein-solvent coupling region of the experimental spectrum that describes interaction dynamics directly tied with the formation of long-range communication channels in the protein. Particularly, in the previous instances we found that the anharmonic dynamics of the solvent molecules within the hydration shell coupled with protein motions to promote long-range coherence pathways in the protein three-dimensional structure.

MD simulation of Meta-II bound with Ce6

Localized Structural Fluctuations in Meta-II-Ce6 and the Disruption of the Active-State Activation Pathway

A network representation of the Meta-II LSFs from the MD simulations of Ce6-bound and -unbound rhodopsin structures are shown in **Figures 4A,B** and a mapping of the LSF components onto the two Meta-II structures is shown in **Figures 4C,D**. One of the distinguishing features that we detect in the Meta-II-Ce6 LSF when contrasted with unbound Meta-II is a disruption in the structural overlap of interactions that creates an activation pathway from the ligand-binding region to the cytoplasmic surface. Specifically, in the Meta-II-Ce6 LSF in **Figure 4B** we find that there is a disconnect between the network of interactions linking rearrangements taking place in the CWxP motif (in the retinal-binding pocket) with a conserved intracellular pathway of intermolecular associations consisting of: a conserved network of internal water molecules (Nygaard et al., 2010), conserved residues in TM1 and TM2, and residues comprising the NPxxY motif in helix 7. This conserved pathway dynamically connects residues from the protein interior (in contact with the retinal) with residues in the cytoplasmic surface that are crucial for G_T binding. In **Figure 4** and Figure S6, a comparison of the unbound and bound receptor structures from the MD simulation reveals that key water molecules involved with the activation pathway have been displaced in the Meta-II-Ce6 structure. The water-mediated H-bonding connection between Ala260 and Asn302 is disrupted in Meta-II-Ce6 and consequently, the connection between Trp265 in helix 6 and Asn302 in helix 7 is weakened. The modification in the conserved water-mediated network in Meta-II-Ce6 effects the rearrangement of H-bonding associations at the intracellular side of the receptor that form the G-protein binding site in the active state. Specifically, the rupture of the Ala260–Asn302 water-mediated H-bond linkage diminishes the correlation between the rotation "toggle" switch of the CWxP motif comprised of conserved residues Tyr268, Trp265, and Phe261 on helix 6 and the structural changes that take within the intracellular core of the receptor involving Met257, Tyr306, and Tyr223 that lead the "breaking" of the ionic-lock between Arg135 on helix 3 and Glu247 on helix 6. The ionic-lock connects the intracellular side of TM3 and TM6 in the active state and breaking of the lock is a crucial step in forming the G-protein binding site. In the MD simulations carried out in this study, we find evidence that the ionic-lock is not fully stabilized in the open form (broken) in the active-state of the Ce6-bound receptor (Figure S7).

Ce6-Induced Correlated Fluctuations and the Creation of an Altered Activation Pathway in Meta-II-Ce6

The reason for the disruption in the activation pathway becomes more apparent when analyzing the induced interactions and correlated motions that accompany Ce6 binding in Meta-II-Ce6. In **Figure 5**, we find that Ce6-induced correlated fluctuations introduce new LSFs in the vicinity of the receptor ligand-binding pocket. An illustration is provided in a comparison of the correlated dynamics in both receptors involving helix 4 near the ligand-binding pocket in **Figures 4C,D**. In the unbound, active-state receptor there is a strong correlation between residues in helix 4 with residues residing in the N-terminus region of the receptor. The helix 4–N-terminus correlated motion is associated with stabilizing the activation pathway that connects the dynamics taking place within the retinal pocket with the dynamics occurring in the cytoplasmic side of the receptor. For instance, in a previous computational study on squid rhodopsin it was revealed that Ala167 (in helix4) in addition to Ala304 (helix 7) and internal water molecules in the intracellular region of helix 7 are instrumental in creating a maximally connected H-bonding pathway that links the active-state retinal binding pocket with the cytoplasmic region (Bondar et al., 2011).

The absence of the (helix 4 - N-terminus) long-range connection in Meta-II-Ce6 suggests that the Ce6 bound receptor may support an alternate pathway for activation. For instance, in **Figure 4A** we observe that the binding of Ce6 excites new collective fluctuations in Meta-II that involve residues in CL3

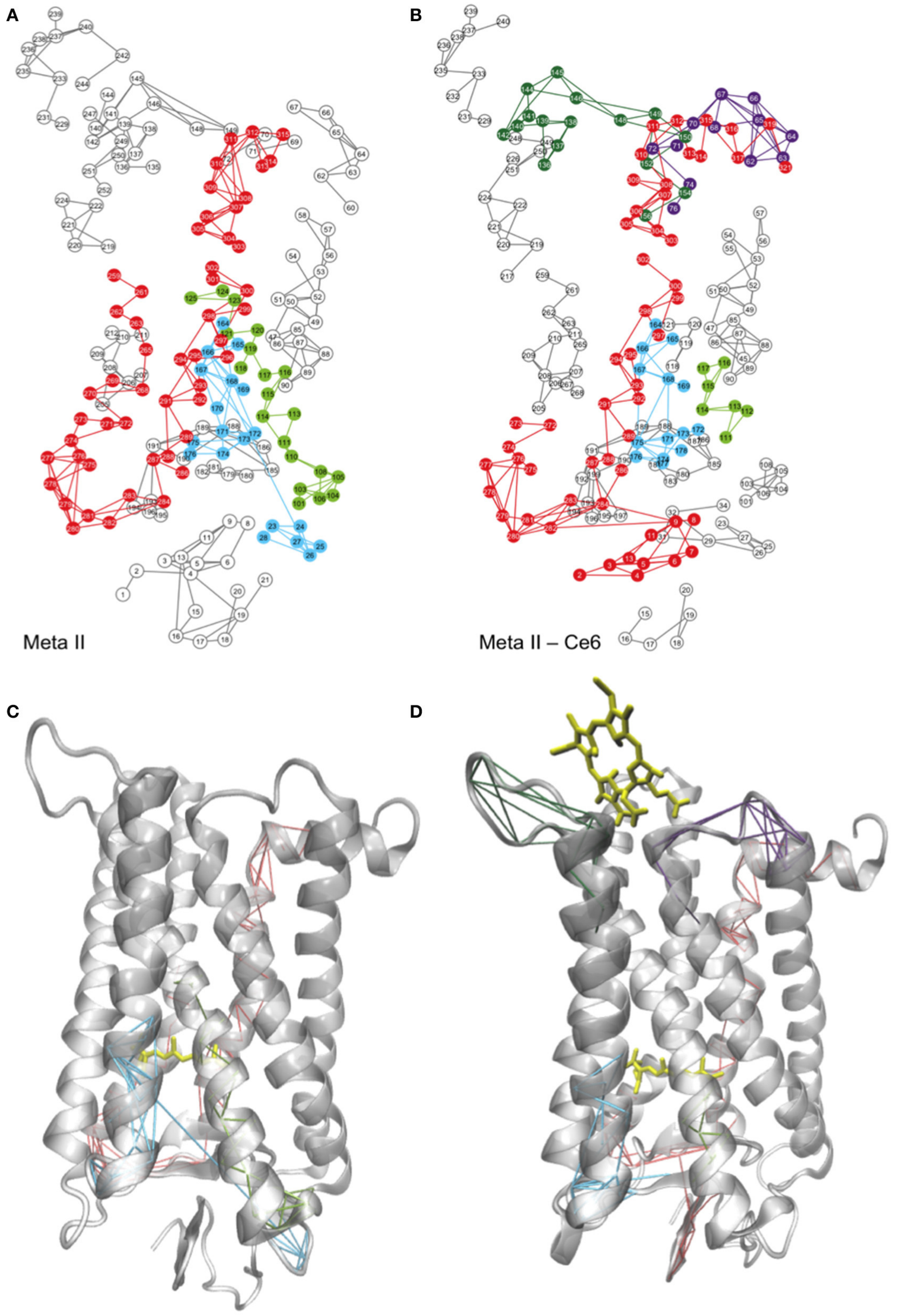


FIGURE 4 | A 2-D mapping of the LSFs of **(A)** Meta-II and **(B)** Meta-II-Ce6 from the MD simulations of rhodopsin and the corresponding mapping of the LSFs onto a cartoon representation of the 3-D structure of **(C)** Meta-II and **(D)** Meta-II-Ce6.

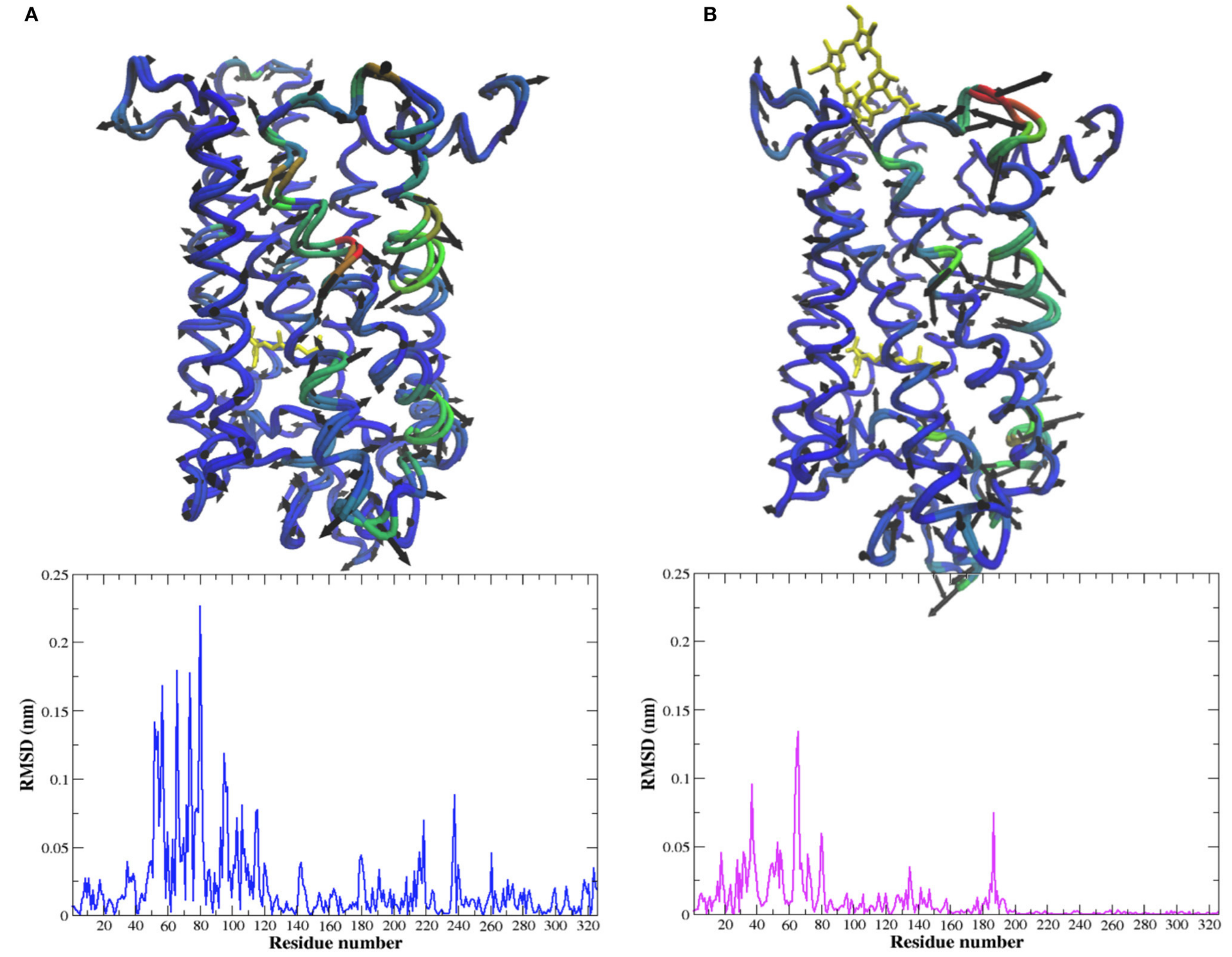


FIGURE 5 | C-α representation of the dominant PCA mode (PCA1) and the corresponding per residue RMSD from the MD simulation of **(A)** Meta-II and **(B)** Meta-II-Ce6 where the areas in blue show regions with less (translational) mobility and the areas in red show regions with more (translational) mobility.

near the Ce6-binding site that are coupled with collective fluctuations of residues in both EL2 and EL3. The Ce6-induced dynamical fluctuations have a prominent effect on the shape of the ligand-binding pocket and consequently, the allosteric interactions that determine the signal communication pathway from the ligand-binding site to the rest of the protein. Specifically, a comparison of the LSFs of the unbound and bound active-state receptor in **Figures 4A,B** reveals that Ce6-induced fluctuations disrupt long-range correlations between EL1 and residues that constitute the hydrophobic core of the receptor, particularly Ala117, Thr118, and Gly120–Glu122. In the unbound receptor, this long-range correlation is associated with maintaining the shape of ligand binding pocket (**Figures 4A,C**). The disruption EL1–receptor core correlated dynamics in Meta-II-Ce6 promotes a deformation of the ligand binding region that consequently allows residues surrounding β-ionone ring, such as Glu122 and Pro215 on helix 5, to shift closer to the retinal as well as to the extracellular side of helix 6. This shift simultaneously alters a stable network of polar interactions within the core of the receptor that connects the ligand-binding site (involving helices 3, 5, and 6) with the rest of the protein. Explicitly, the shift of helix 5 ligand-binding residues promotes a shift in the hydrophobic interactions involving Gly121 and Leu125 on helix 3 and Phe261 on helix 6. The adjustment of the hydrophobic packing interaction allows the intracellular side of the helix (helix 3) to move closer to helix 6. Further, the helix 3 shift is connected with a correlated fluctuation involving the motion of residues Phe261, Trp265, and Tyr268 on helix 6 with that of EL2, such that the movement of these residues promotes an overall tilt of the IC side of H6 and as a consequence also reduces the distance between TM3 and TM6 (**Figure 6**). The outcome is interhelical packing changes in the core of the molecule that shift TM2 and TM8 upward in the direction of the C-terminus of the receptor and tilts TM7 inward toward the hydrophobic core. Consequently, the altered ligand-binding cavity shape also supports a 7TM structure with a weakened but still associated

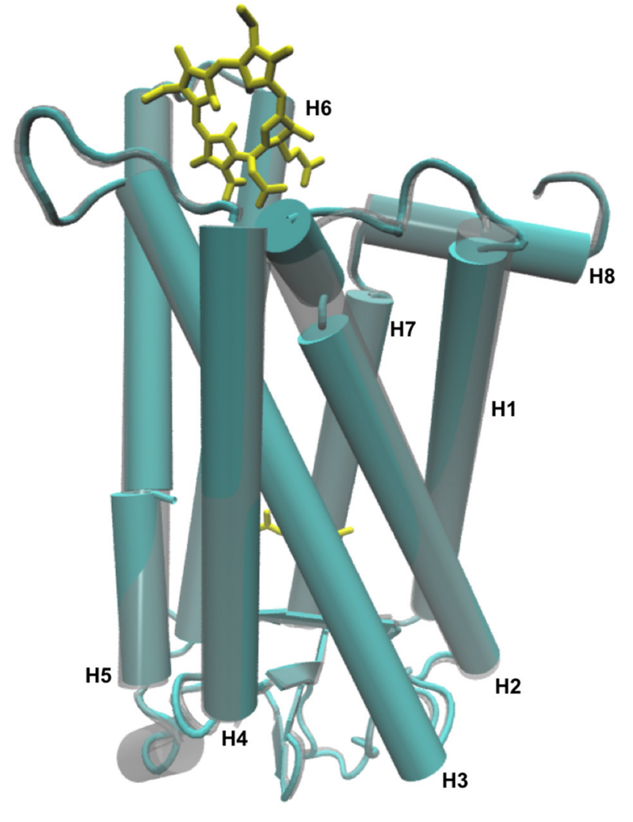


FIGURE 6 | Overlap of a cartoon representation of the 3-D structure of Meta-II (cyan) and Meta-II-Ce6 (gray) from the MD simulations of Meta-II.

Arg135–Glu247 "ionic-lock" (Figure S7) and a ligand-binding site with increased distance from EL2. In fact, the amplitude of the dynamical oscillations of the β_4 loop of EL2 is far greater in the Ce6-bound receptor when compared with unbound Meta-II. This suggests that the increased distance between the binding pocket and the ECD in Meta-II-Ce6 provides an environment for the ligand that possess far less protection from hydrolysis and therefore a reduction in overall receptor stability when contrasted with the unbound receptor. Hence, the effect of the loss of long-range interaction connecting core residues with EL1 in Meta-II-Ce6 is the deformation of the ligand-binding pocket such that it modifies the intra-protein interactions that determine the shape of the G-protein binding site. And it also alters the contact interactions between the ligand-binding site and EL2 that are instrumental in forming the activation pathway in Meta-II.

The Ce6-induced deformation of the Meta-II ligand-binding pocket also supports a new set of associations that creates an alternative pathway for signal propagation when compared with the unbound receptor (**Figure 4**). The induced long-range correlated fluctuation extending from the ECD to ICD connect the dynamics taking place in the N-terminus of the receptor with fluctuations in EL3 and residues in helix 7. From an analysis of the Ce6-induced collective dynamics in **Figure 4D**, we find that the long-range interactions create large-scale torsional fluctuations in a cluster of hydroxyl residues in EL3 that separate helices 6 and 7. The residues with the largest contribution to the induced EL3 dynamics include Thr277, Ser281, and Pro285. The substantial increase in the magnitude of EL3 fluctuations subsequently alters the amplitude of rotational fluctuations of residues in helices 6 and 7 that line the retinal-binding pocket. These particular residues have previously been identified as having an important role in signal propagation and activation in rhodopsin and include Trp265, Pro267 - Ala271, Pro291, and Ala295. The large-scale induced-torsional fluctuations of these activation residues are instrumental in linking the dynamics of the ligand-binding site with the conserved pathway of residues forming the NPxxY motif. Their amplified motion compensates for the disruption in the intricate network of intermolecular interactions that form the activation pathway in Meta-II and in its place creates an anharmonic connection of associations that translate changes taking place in retinal binding region with the structural changes taking place in the ICD.

Together, the changes in the localized interactions induced by Ce6-binding result in conformational structural changes in Meta-II that promote and altered NPxxY motif and an altered activation pathway. Specifically, the binding of the allosteric modulator Ce6 supports a series of long-range interactions that result in structural differences in the TM3/TM6 distance of the active receptor as well as packing interactions between TM1/TM2/TM7 helices (**Figure 6**) when compared with the unbound receptor. These modifications support a narrowed G_T binding site that is likely less efficient in binding the G-protein and adopts a more structurally dynamic ligand-binding cavity that would have a direct effect on the thermal stability of the active-state receptor.

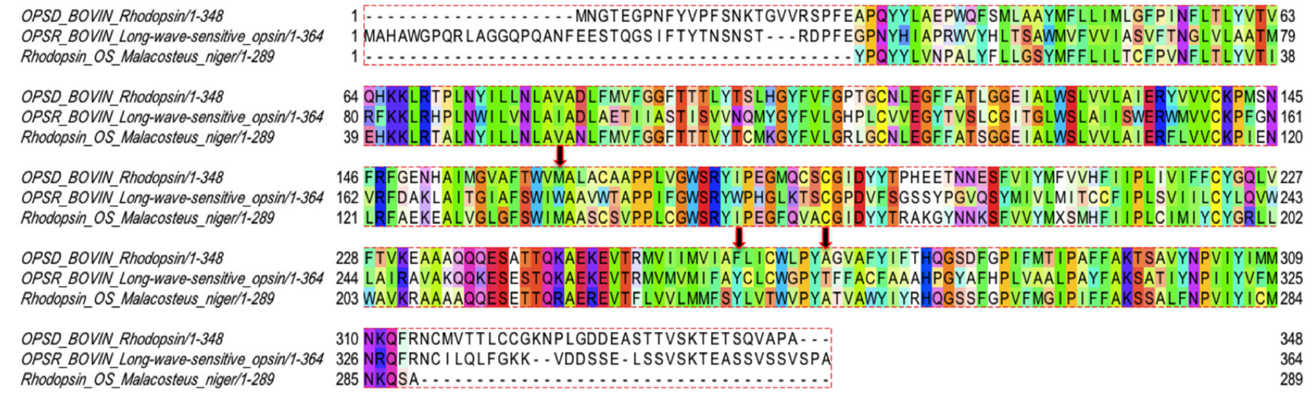
DISCUSSION

Implications of Ce6 Effects on Rhodopsin for Deep-Sea Ocean Vision and Spectral Tuning

It has been proposed that Ce6 plays a role in modifying the receptor-chromophore interactions in rhodopsin that adapts the *Malacosteus niger* (*M. niger*) dragon fish visual system (Douglas et al., 1998, 1999, 2016; Kenaley et al., 2014). Ce6-induced modification in intra- and inter-protein interactions permit the *M. niger* species to emit far-red light from suborbital photophores, in addition to the blue bioluminescence that is normally emitted by deep-sea dragon fish. The suspected mechanism of the enhancement of long-wavelength sensitivity in the dragon fish is via spectral tuning (Kenaley et al., 2014). Spectral tuning is a molecular mechanism that shifts the optical properties of the retinal that regulate the absorbance maximum of the absorption of light. This can happen through evolutionary processes in which case, it involves a combination of adaptation and positive selection of key residues that directly interact with the retinal. Or alternatively, the tuning mechanism can be induced by adding small molecules that externally modulate opsin-retinal interactions (Washington et al., 2004, 2007; Isayama et al., 2006; Balem et al., 2009). Thus, the interaction of *specific* amino acids of the rod-opsin protein with the chromophore ultimately determines the peak absorbance (λ_{max}) and therefore, also the spectral sensitivity of the pigment. It is well known that rhodopsin evolved from cone opsins (Imamoto and Shichida, 2014). Ancestral pigments were cone pigments and rod pigments evolved later in response to the necessity to function in dim light conditions. Rod opsins have a prolonged active-state. This is achieved by having high thermal stability, which promotes both slow thermal exchange and slow recovery. The advantage is that this allows for a longer signaling state, which both amplifies the response signal and also increases the overall sensitivity of the photoreceptor cell. Cone cells on the other hand, produce an active-state that is thermally unstable. In this respect, they also have a much faster signaling state and a faster rate of regeneration when contrasted with rod cells. These particular characteristics are best suited for operating in daylight conditions where the photoreceptor cells are under pressure to function under successive light stimuli.

We have previously proposed that Ce6 binding in rhodopsin takes place in the ICD of the receptor (Balem et al., 2009). The MD simulations and molecular docking predictions carried out in this investigation support this supposition. Thus, under this assumption, the Ce6-induced modulation of rhodopsin protein-ligand interactions would have to occur through an allosteric mechanism. However, the inferred mechanism of the Ce6-induced changes of rhodopsin functional dynamics from this study in many ways resonates more like an account of the *evolutionary* pattern of mutational changes that eventually transitioned ancestral cone cells into rod cells (Nathans, 1990; Lin et al., 1998; Rajamani et al., 2011; Imamoto and Shichida, 2014). In other words, the detected Ce6 modulation of rhodopsin functional properties mirrors the mechanism of spectral tuning brought about by mutational changes of specific amino acids in visual phototransduction evolution. For example, the observed "tuning" sites for Ce6-induced intermolecular and structural changes that we have detected in our experiments on bovine rhodopsin bound with Ce6 can be succinctly explained by contrasting them with the effects of well-known, critical mutational differences in the sequences of the two types of visual pigments. An illustration is given in **Figure 7** were we have assembled an alignment of three different receptor sequences that include bovine rhodopsin, long-wavelength bovine opsin (i.e., red cone opsin), and *M. niger* rhodopsin. In the aligned sequences, bovine rhodopsin represents a characteristic rod opsin, whereas the long-wavelength and *M. niger* sequences represent a cone opsin and a red-shifted rod opsin respectively. The aligned sequences clearly highlight some of the most distinguishable differences of rhodopsin when compared with the other two spectrally shifted opsins (**Figure 7**). The alignment also provides further insight into the nature of the observed induced dynamical and structural changes that lead to differences in the detected spectral properties of Ce6-bound and unbound rhodopsin. But more importantly, it offers a deeper understanding about the evolution of critical long-range interactions involved in the transmission of the excitation signal from the binding site of the chromophore to the cytoplasmic surface. These long-range interactions in visual pigments have been extensively tuned over time to improve the sensitivity and stability of the more evolved rod receptors and their pathway of adaptation is tied with essential global changes in the molecular structure of the receptor that have been tailored to stabilize them.

Two of the most recognized critical residue changes in rod cells vs. that of cone cells are found at position 122 and 189 (in rhodopsin numbering) in the receptor sequences. The substitution of Ile for Glu at position 122 is known to be associated with a dramatic increase (blue-shift) in the wavelength of the photosensitivity of various cone cells when contrasted with rod cells (Lin et al., 1998; Lewis et al., 2006; Imamoto and Shichida, 2014). Similarly the substitution of Pro or Met for Ile at position 189 is related with increasing both the decay rate of the signaling state (Figures S9c, S10e,f) as well as the reconstitution rate of the chromophore in cone opsins (Janz et al., 2003; Imamoto and Shichida, 2014; Yanagawa et al., 2015). Despite this general knowledge, there is still the remaining challenge of developing an overall sense of how these single point mutations impact major properties of the receptor. For instance, in particular cases local perturbations of proteins have been shown to produce global changes that have an inclusive effect



```
OPSD_BOVIN_Rhodopsin/1-348                      1 -------------------MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLYVTV 63
OPSR_BOVIN_Long-wave-sensitive_opsin/1-364      1 MAHAWGPQRLAGGQPQANFEESTQGSIFTYTNSNST---RDPFEGPNYHIAPRWVYHLTSAWMVFVVIASVFTNGLVLAATM 79
Rhodopsin_OS_Malacosteus_niger/1-289            1 --------------------------------------------YPQYYLVNPALYFLLGSYMFFLILTCFPVNFLTLYVTI 38

OPSD_BOVIN_Rhodopsin/1-348                     64 QHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSN 145
OPSR_BOVIN_Long-wave-sensitive_opsin/1-364     80 RFKKLRHPLNWILVNLAIADLAETIIASTISVVNQMYGYFVLGHPLCVVEGYTVSLCGITGLWSLAIISWERWMVVCKPFGN 161
Rhodopsin_OS_Malacosteus_niger/1-289           39 EHKKLRTALNYILLNLAVANLFMVFGGFTTTVYTCMKGYFVLGRLGCNLEGFFATSGGEIALWSLVVLAIERFLVVCKPIEN 120

OPSD_BOVIN_Rhodopsin/1-348                    146 FRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPEGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLV 227
OPSR_BOVIN_Long-wave-sensitive_opsin/1-364    162 VRFDAKLAITGIAFSWIWAAVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMITCCFIPLSVIILCYLQVW 243
Rhodopsin_OS_Malacosteus_niger/1-289          121 LRFAEKEALVGLGFSWIMAASCSVPPLCGWSRYIPEGFQVACGIDYYTRAKGYNNKSFVVYMXSMHFIIPLCIMIYCYGRLL 202

OPSD_BOVIN_Rhodopsin/1-348                    228 FTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAVYNPVIYIMM 309
OPSR_BOVIN_Long-wave-sensitive_opsin/1-364    244 LAIRAVAKQQKESESTQKAEKEVTRMVMVMIFAYCLCWGPYTFFACFAAAHPGYAFHPLVAALPAYFAKSATIYNPIIYVFM 325
Rhodopsin_OS_Malacosteus_niger/1-289          203 WAVKRAAAAQQESETTQRAEREVTFLVVLMMFSYLVTWVPYATVAWYIYRHQGSSFGPVFMGIPIFFAKSSALFNPVIYICM 284

OPSD_BOVIN_Rhodopsin/1-348                    310 NKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA---                                         348
OPSR_BOVIN_Long-wave-sensitive_opsin/1-364    326 NRQFRNCILQLFGKK--VDDSSE-LSSVSKTEASSVSSVSPA                                         364
Rhodopsin_OS_Malacosteus_niger/1-289          285 NKQSA-------------------------------------                                         289
```

FIGURE 7 | Sequence alignment of bovine rhodopsin (OPSD_BOVIN), long-wavelength bovine rhodopsin (OPSR_BOVIN), and *M. Niger* rhodopsin (Rhodopsin_OS_Malacosteus_niger) rendered with Taylor coloring.

on protein function and/or stability. In rhodopsin we identify three such spots, where substitution of select residues have a significant destabilizing effect on both the global functional properties of the receptor and the nature of the signaling pathway of activation. It is particularly through these sites that Ce6 modulates the functional dynamics of the visual pigments through long-range, evolutionary-conserved interactions that establish communication between the chromophore and the rest of the receptor (Figure S6).

Mechanisms of Allostery in Rhodopsin and the Generality to Other Visual Receptors

We have uncovered three specific allosteric sites in our investigation of Ce6-bound rhodopsin that are utilized to modulate the functional properties of the receptor. These sites are activated by Ce6 in the ICD of the receptor and are associated with moderating the coupling mechanism of the different structural components in the receptor that are necessary for signal activation and propagation in rhodopsin (Wolf and Grünewald, 2015). Particularly, through these allosteric sites, Ce6 modifies the signaling properties of rhodopsin by acting on distinct structural constraints that rearrange in response to activation and have a direct role in signal propagation.

Coupling between the Ligand-Binding Site and the ECD in Ce6-Bound Rhodopsin

One such allosteric site is Ala269 in rhodopsin. Ce6-induced interactions involving Ala269 have a direct role in altering the activation pathway in Meta-II rhodopsin (**Figure 7**) due to its role in stabilizing the ligand-binding site (Tsukamoto et al., 2010). For instance, Ce6 binding promotes reduced interactions between Ala269 and the β-ionone ring of the agonist. This change in residue-agonist binding affects the thermal stability of the entire receptor. The mechanism of the diminished receptor stability was deduced from an analysis of the Ce6-induced collective dynamics in Meta-II (**Figures 4C,D**). From the analysis we found that the Ce6-induced long-range interactions in rhodopsin create packing defects in the retinal-binding pocket that are allosterically transmitted to structurally conserved hydroxyl residues in EL3. The result is Ce6-induced, large-scale fluctuations in EL3 that directly modify the amplitude of the dynamics of residues in helices 6 and 7 that surround the retinal ligand. Thus, it is the increase in dynamics of these particular residues (in helix 6 and 7) that have a direct impact on the coupling mechanism that links signaling components (residues) in the ligand-binding site with those in the ECD. Their coupling is directly correlated with the basal activity of the receptor. The Ce6-induced dynamics in both dark-state and Meta-II rhodopsin foster a more structurally dynamic ligand-binding pocket that reduces the thermal stability of the receptor. We find evidence of the diminished stability in both our experimental (**Figure 1D**) and MD simulation analyses (**Figure 5B**) of Meta-II-Ce6.

Analogously, in the sequence alignment of the visual pigments in **Figure 7** we notice a substitution of Thr269Ala in the cone pigment when compared with both rhodopsin sequences. The replacement of threonine for alanine at position 269 in the receptor structure would account for less stability in the ligand pocket of the cone pigment due to the introduction of a bulky side-chain that would disrupt the packing interactions of the β-ionone ring with the ligand-binding residue (Supplementary text, *MD simulation of Meta-II mutations*, Figures S9b,f, S10c,d, S11, S12).

Previous studies on rhodopsin have noted a relationship between ligand-binding pocket structural flexibility, the thermal stability of the receptor, and the receptor active state lifetime (Janz et al., 2003; Ala-Laurila et al., 2004; Yanagawa et al., 2015). They conjectured that the higher thermal stability of rhodopsin when contrasted with cone cells represents an evolutionarily adapted trade-off of photoreceptor speed for a high detection threshold. The residue difference (at position 269) in the sequence alignment of the cone sequence vs. that of the rod cells parallels the Ce6-induced changes that we have observed in dark-state and Meta-II rhodopsin. In both instances, instability in the ligand-bonding pocket is directly correlated with the thermal stability of the receptor. Furthermore, in on our own previous investigation on rhodopsin we have clearly established a correlation between the flexibility of the agonist ring with receptor conformational stability. Therefore, Ce6-binding in rhodopsin alters the coupling mechanism between the ECD and ligand-binding site, and accordingly also strongly modifies the thermal stability of the receptor. Unfortunately, it is not possible to precisely measure the Meta-II decay rates in the presence of Ce6 because Ce6 quenches tryptophan fluorescence (Balem et al., 2009) but there is compelling evidence from the analysis of the residual fluorescence that Meta-II does decay faster in the presence of Ce6 (Balem et al., 2009).

Coupling between Ligand-Binding Site and the IC Region in Ce6-Bound Rhodopsin

Phe261 is another allosteric site that has been uncovered in our analyses of Ce6-bound rhodopsin. Phe261 (helix 6) is a highly conserved residue in the GPCR-A family. It, along with Gly121 (helix 3), forms a hydrophobic micro-domain in the interior of rhodopsin that rearranges during activation. The rearrangement of hydrophobic interactions between helix 3 and helix 6 are essential in translating retinal conformational changes that take place during activation into helical rearrangements in the intracellular region of the receptor that are propagated to the cytoplasmic surface. The mechanism of signal propagation from the ligand-binding site to the cytoplasmic region involves an intricate network of conserved H-bonding interactions that tightly couple the ligand-binding site with the IC region of the receptor (Fritze et al., 2003; Brown et al., 2009). This allosteric network of H-bonding interactions has been found in other class-A receptors, although the extent of coupling between the distinct components of the network varies amongst the different receptors. In our analyses on Ce6-bound rhodopsin, we find that the allosterically coupled motions leading from the ligand-binding site to the G-protein binding site is disrupted by the loss of conserved water molecules that comprise part of the NPxxY motif. Specifically, the Ce6-induced interactions in Meta-II lead to the loss of the conserved water molecule shared between Met257, Phe261, and Asn302 (Figure S6), which ultimately leads to hydrophobic packing defects in the protein interior and a less stable signaling pathway. The result is a weaker coupling between the allosteric components and a coarser translation of ligand-binding changes to the G-protein side of the receptor.

In the sequence alignment of the visual pigments in **Figure 7**, we observe a major change in the amino acid residue at position 261 in rhodopsin when compared with both the cone pigment and the *M. niger* rhodopsin sequence. Particularly, the substitution of Tyr261Phe in the cone pigment is known to lead to a much "looser" coupling between the ligand-binding site and the intracellular connections that lead to the G-protein binding region (Supplementary text, *MD simulation of Meta-II mutations*, Figures S9d,h, S10g). The replacement of tyrosine for phenylalanine at position 261 modifies the allosteric network of the cone pigment in a similar manner to what has been observed in the Ce6-induced disturbances of the intracellular signaling regions of rhodopsin (**Figure 4**, Figure S9g). It is well known that cone pigments maintain a more heterogeneous active-state after activation and less efficient G-protein activation when compared with rhodopsin (Imamoto and Shichida, 2014). The Tyr261Phe substitution has also been credited with a 10 nm blue-shift of the λ_{max} between cone and rod pigments (Chan et al., 1992), suggesting that the residue change is somehow connected with the modulation of the signaling network from the ligand-binding site. The stability and homogeneity of the active-state of rhodopsin is unique in the class-A receptors. The low basal activity and high photon detection efficiency of the receptor is attributed to the succinct coupling of the allosteric network of signaling components of the active-state.

Coupling between the G-Protein Binding Site and Retinal Ligand-Binding Site in Ce6-Bound Rhodopsin

We have also identified Met163 as an allosteric site in rhodopsin that has a central role in forming the G-protein binding site in the active-state protein. In our LSF analysis of the active-state of rhodopsin (**Figures 4B,D**) we found that long-range interactions between residues near Met163 and the N-terminus of the receptor are central in stabilizing the signaling network of interactions in Meta-II. Subsequently, in our analysis of Meta-II Ce6-induced long-range interactions (involving Met163) we determined that the Ce6-induced disrupted connection between helix 4 and the N-terminus allowed for the deformation of the ligand-binding pocket such that it modified the intra-protein interactions that determine the shape of the G-protein binding surface site (**Figure 6**). The disruption in the long-range interactions in Meta-II-Ce6 also accounted for the alteration in the contact interactions between the ligand-binding site and EL2 that are instrumental in forming the signaling pathway in the active-state in visual pigments.

Overall, we find that binding of Ce6 weakens conserved interactions that allosterically link the dynamics in the ligand-binding pocket with conformational fluctuations taking place at the receptor G-protein binding site. The decoupling of the distinct regions of the allosteric network result in a mixture of photo-intermediate conformations in the active-state of the receptor. For instance, interhelical distance fluctuations in the IC region of Meta-II-Ce6, due to instabilities in the conformational coupling of the structural elements of the receptor (Figure S7), alter the population of conformations in the active-state ensemble. An examination of the retinal torsional dynamics (**Figure 3B**) and the conformational ensemble dynamics (Figure S8) from the MD simulation of Meta-II reveals the interconversion between two dominant conformations in the active-state receptor when bound to Ce6. The dominant conformation (Meta-II$_{\mathrm{i}}$) has reduced distance between the IC region of helices 3 and 6 and an increased distance between the ligand-binding site and EL2, relative to the initial X-ray crystal structure of Meta-II used for the MD simulations. The secondary conformation (Meta-II$_{\mathrm{ii}}$) deviates only slightly from the initial Meta-II structure. The Meta-II$_{\mathrm{i}}$ conformation would presumably have a higher propensity for chromophore hydrolysis in the active-state and a conformation that is less efficient for G_T activation compared with Meta-II$_{\mathrm{ii}}$. Thus, the two photo-intermediate structures are associated with distinct intracellular signaling pathways that have ramifications on both the maximum G_T activity of the bound receptor as well as overall receptor thermal stability. In support of this conclusion, we have found in previous studies that the presence of Ce6 does indeed reduce G_T activation in a concentration dependent fashion (Balem et al., 2009).

Referring again to the sequence alignment of the visual pigments in **Figure 7** we observe that the cone pigment has a tryptophan in position 163, whereas both rhodopsin sequences possess a methionine. The Trp163Met substitution in the cone pigment sequence introduces an amino acid with a large aromatic side-chain in a key position that would significantly alter the close packing interactions of helices 3–5 that support the active-state receptor structure (Supplementary text, *MD simulation of Meta-II mutations*, Figures S9a,e, S10a,b). Furthermore, the residue exchange would also considerably weaken crucial long-range allosteric connections between the ligand-binding site and IC residues (Figures S9a,e) associated with G-protein binding (when contrasted with Meta-II of the rod cell sequences). The effect would be a substantially weaker coupling between the allosteric components of the signaling pathway and a shift in the population of photo-intermediate states toward a conformation that overwhelming supports a reduced affinity for both the G-protein and the chromophore (retinal). It is widely recognized that cone pigments have a higher rate of thermal activation and much faster decay of the photo-activated pigment compared with rod pigments. The decay of Meta-II is the rate-limiting step for the termination of the light response in visual pigments and hence, one of the key factors that distinguishes rod cells from cone cells. For this reason, it has been conjectured that the differences in the two types of visual pigments are attributed to a *few* strategic amino acid changes that were acquired during the evolution of rod cells for dim light sensitivity. The long-lived G_T activation state with high efficiency is typical of rhodopsins, implying that rhodopsin acquired these specific characteristics over time to explicitly refine its function for single-photon detection sensitivity.

Ce6-induced Modulation of Conserved Allosteric Sites in Rhodopsin-Like (GPCR-A) Receptors and New Strategies for Drug Design of GPCR Allosteric Modulators

Developing a clear understanding of how specific ligand modulators moderate topographically distinct allosteric sites in receptor families is one of the essential steps for the successful design of small-molecule allosteric drugs. The evolution of protein allosteric sites involves pairwise interactions that are responsible for the propagation of conformational changes from the ligand-binding site to a distal functional site. In visual receptors, these allosteric sites are linked with protein coevolved residue pairs (residues with evolutionary correlated mutational patterns) that have been selected for their regulatory properties and conformational ensemble dynamics that are directly involved with fine-tuning spectral sensitivity. In this work, we propose that the small-molecule allosteric modulator Ce6 binds specifically in the CP domain of rhodopsin. It acts by not only moderating the spectral sensitivity of rhodopsin, but on a higher level by modulating GPCR-A-wide conserved allosteric sites that facilitate coupling of receptor structural and functional domains (**Figure 8**, Figures S13, S14, and Supplementary text, *Ce6 modulation of GPCR-A conserved allosteric sites in rhodopsin*). Principally, we find that Ce6 binding in rhodopsin allosterically alters the receptor structure by mediating conserved long-range conformational fluctuations that modulate access to the retinal ligand-binding pocket. Why is this significant? Every currently known GPCR transmits a ligand-binding signal originating in the EC and/or TM domain to the CP domain via conformational fluctuations that take place in the TM domain. The CP domain is the site where the G-protein recognizes the active conformation of the receptors, and unlike other 7TM regions, the CP interface of the various GPCR-A receptors is relatively conserved. Therefore, these conserved residue positions at the CP interface form part of a signaling mechanism that allows the distinct receptors to bind and activate G-proteins by utilizing a common construction of coevolutionary residue pair interactions that form an allosteric regulatory network to the rest of the receptor. We conjecture that Ce6 modulates these same GPCR-A shared allosteric regulatory networks. For this reason, Ce6 may offer an initial stepping-stone for comprehending and creating small-molecule allosteric modulators that offer precise control of GPCR signaling pathways.

CONCLUSIONS

In this investigation, we identify long-range conserved interactions in rhodopsin that are excited by the binding of an allosteric modulator (Ce6) in the cytoplasmic domain of the receptor. The excited structural fluctuations modify fundamental signaling processes that control receptor long-range interactions and are common in all Class-A GPCRs. In this specific case, we find that Ce6 stabilizes specific receptor conformations that alter the coupling mechanism between the distinct domains of the receptor and hence modify the GPCR signaling components that define rhodopsin function. These results provide deeper insight into the evolutionary coupled interactions in Class-A GPCRs that modulate the mechanism for coupling the ligand-binding site with both the ECD and G-protein binding sites and offers a foothold for elucidating how GPCR signaling pathways may be manipulated for allosteric drug discovery and design. Ce6 provides a clear illustration of a how small-molecule ligand modulator, associated with moderating the allosteric interaction networks and signaling pathways linked with spectral sensitivity in visual receptors, may also provide valuable insight into the (molecular) mechanisms that enhance or inhibit selective receptor pathways connected with evolutionary principles of allosteric regulation.

MATERIALS AND METHODS

Sample Preparation

Rhodopsin samples were purified in detergent micelles composed of dodecyl maltoside (DM). The choice of DM as a detergent is justified because the conformational changes in rhodopsin in DM are virtually identical to those seen in liposomes (Kusnetzow et al., 2006). Rhodopsin samples were obtained through transient transfection or from stable cell lines. Transient transfection of COS-1 cells was carried out as described (Oprian et al., 1987), with the exception that the cells were harvested 72 h after transfection. Tetracycline inducible HEK293S stable cell lines were established as described previously (Reeves et al., 2002). Both types of cells were solubilized with 1% (w/v) DM for 1 h and the proteins were purified by 1D4 immuno-affinity chromatography in 0.05% DM as described (Hwa et al., 1999). Briefly, after solubilization of the cells, the suspension was centrifuged for 30 min at 35,000 rpm and 4°C. The supernatant was mixed with 1D4 Sepharose beads (approximate binding capacity of 1 μg rhodopsin/ μl of resin) for at least 6 h at 4°C. The resin was then washed with 50 bed volumes of 0.05% (w/v) DM in PBS followed by 10 bed volumes of 0.05% (w/v) DM in 2 mM Na_2HPO_4/NaH_2PO_4 (pH 6.0). WT and mutant proteins were eluted with 70 μM C-terminal nonapeptide (TETSQVAPA) in 0.05% (w/v) DM in 2 mM Na_2HPO_4/NaH_2PO_4 (pH 6.0). The initial concentration of the sample was determined by UV absorbance and subsequently diluted to a concentration of 200 μm in preparation for the THz spectroscopy experiments.

Ce6 was obtained from Frontier Scientific, Logan, UT. Ce6 was added to rhodopsin samples from a 100 mM DMSO stock solution or its dilutions, to a final concentration of 200 μm.

The rhodopsin samples used in the THz spectroscopy experiments were prepared by allotting 20 μL of the prepared sample onto a custom ordered diamond transmission window (Specac Co., United Kingdom). Excess water from the solution was initially removed by applying a low, steady flow of N_2 gas over the sample droplet for approximately 3 min. The resulting

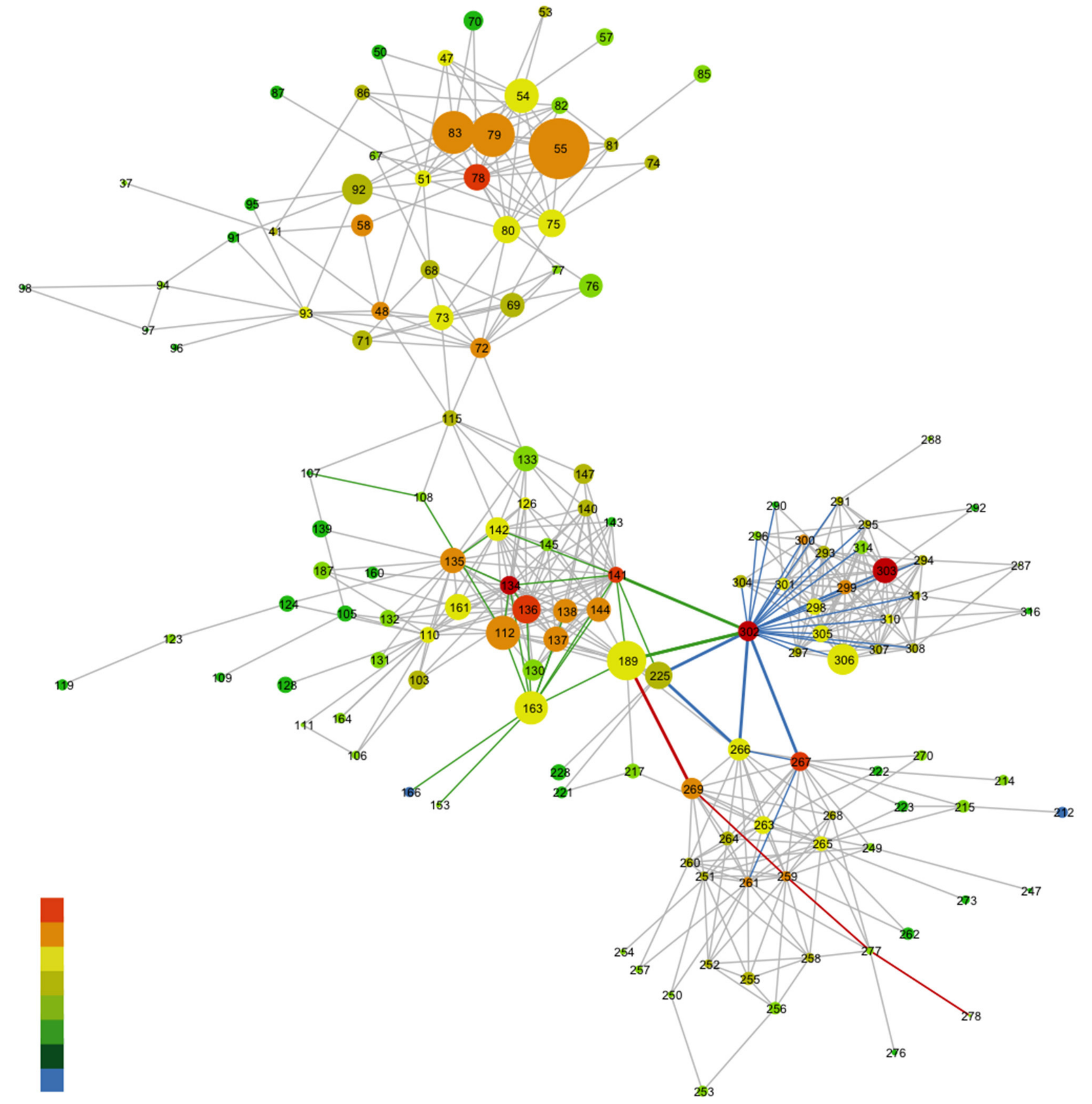


FIGURE 8 | Network representation of the MSA of the rhodopsin-like family of proteins. Nodes represent a specific amino acid and gray edges represent the lines between the nodes. The rainbow coloring of the nodes characterizes the degree to which a given amino acid takes part in the MI network (denoting coevolution propensity). Red nodes have a high MI value and blue nodes a low MI value. The size of the nodes in this case corresponds to the conservation of the amino acid from the MSA. The colored edges highlight the associations between specific nodes. For more details refer to Supplementary Text, *Ce6 modulation of GPCR-A conserved allosteric sites in rhodopsin*. The reference structure for the MSA is bovine rhodopsin (opsd_bovin) with pdb ID 1u19.

sample was subsequently rehydrated by equilibrating the dried-off sample in a vacuum sealed container with the vapor pressure of a saturated salt solution at 20°C for a minimum of 3 days. A relative humidity (RH) of 97% was obtained from the vapor pressure of a saturated K_2SO_4 solution (Wexler and Hasegawa, 1954). The prepared rhodopsin sample was subsequently placed in a sealed transmission cell consisting of two diamond window substrates and a saturated salt solution was placed at the bottom of the cell to ensure that hydration was maintained throughout the experiment.

Illumination of the samples in all experiments was carried out with a Fiber-Lite DC 950 regulated illuminator by Dolan-Jenner industries.

THz Spectroscopy Experiments

The dark-state rhodopsin experiments were performed under dim-red light conditions and photo-isomerization was triggered with visual light excitation. The THz spectroscopy experiments were carried out on a Jasco FTIR - 6000 series spectrometer. The protein sample spectra were collected with a liquid helium cooled bolometer in the 15–250 cm^{-1} spectral range. The 15–100 cm^{-1} THz spectra were collected with a 25-micron beam splitter while the data in the 100–250 cm^{-1} spectral region was collected with a 12-micron beam splitter. For each transmission measurement a 25 mm diameter region of the protein sample was illuminated with the THz beam to determine the absorbance. In the spectral measurements presented, each scan consists of 16 averaged scans and the infrared data was collected with a spectral resolution of 4 cm^{-1}.

Molecular Dynamic Simulations

MD Simulation of Dark-State Rhodopsin and Meta-II

Each MD simulation consisted of a starting x-ray crystal structure taken from the PDB database. PDB structure 1 μ19 was used for the inactive (dark) state of rhodopsin and 3pxo was used for Meta II. In all simulations, the receptor was embedded in a hydrated lipid bilayer with all atoms represented explicitly. Specifically, the dark-state receptor and any resolved water molecules from the crystal structure were embedded in an equilibrated palmitoyloleoyl-phosphatidylcholine (POPC) bilayer consisting of 110 lipid molecules, and additional 7400 water molecules, and 100 mM NaCl (to neutralize the net charge of the system). The membrane system was built with the use of the g_membed tool in Gromacs. All titratable groups in the receptor were considered to be charged (Fahmy et al., 1993). The exceptions were Asp83 and Glu122, which were both neutral in both the dark-state and Meta II MD simulations. Also for the dark-state MD simulation, the Schiff base was protonated whereas Glu113 was deprotonated. For the Meta II simulation both the Schiff base and Glu113 were set to neutral. The active state receptor combined with the structural waters from the crystal structure was prepared in a similar manner to that of the dark-state. MD simulations were performed at 300 K using the Gromacs package (www.gromacs.org) version 5.0. The GROMOS96 43a2 force field parameters were utilized for the protein and the Berger lipid parameters were used for the lipid component of the membrane protein (Berger et al., 1997). The SPC water model was used for hydration and the ground-state retinal parameters (Bondar et al., 2011) for both the 11-*cis* and *all-trans* retinal chromophore were obtained from the Bondar group.

In the rhodopsin simulations, energy minimizations were initially carried out to reduce the number of unfavorable contacts between added solvent molecules and the receptor using a steepest descent method to a convergence tolerance of 0.001 kJ mol^{-1}. The energy minimization was followed by a MD run with constraints for 200 ps in which an isotropic force constant of 100 kJ mol^{-1} nm^{-1} was used on the protein and lipid atoms. During the restrained dynamics simulation, the temperature and pressure of the system were kept constant by weak coupling to a modified velocity rescaled Berendsen temperature (Berendsen et al., 1984) and pressure baths and in all cases the protein, lipid, water, and ions were coupled to the temperature and pressure baths separately. The output conformation from the MD simulation with constraints was used as the starting conformation for an initial 200 ns equilibrium MD simulation.

Six subsequent simulations were conducted where randomized conformations from the last 10 ns of the equilibrium simulations were used as starting point conformations for each distinct simulation. These subsequent simulations were carried out for an additional 500 ns and were eventually used to assess the picosecond time scale fluctuations in the receptor systems. The final simulations were carried out with a 1 fs time step where the bonds between the hydrogen and the other heavier atoms were restrained to their equilibrium values with the linear constraints (LINCS) algorithm (Hess et al., 1997). Particle mesh Ewald (PME) method (Essmann et al., 1995) was used to calculate the long-range electrostatic interactions in the simulation and was used with a real-space cutoff of 1.0 nm, a fourth order B-spline interpolation and a minimum grid spacing of 0.14 nm.

MD simulations of Meta-II mutants (single and double mutations) were carried out by creating a residue mutation(s) with DUET (http://biosig.unimelb.edu.au/duet/), a web server for studying missense mutations in proteins. Minimization and production run MD simulations on the mutated receptors were carried out in a manner analogous to that described for WT rhodopsin.

Trajectory snapshots, each containing a record of the atom positions and velocities at a particular instant in time, were saved every 100 fs during the production simulations.

Modeling and Docking Studies

Ce6 docking studies on dark-state rhodopsin and Meta-II were performed using the AutoDock 3.0 program (Goodsell et al., 1996). Modeling and docking studies of the G_t C-terminal high affinity peptide were performed using the MODELER (Sali et al., 1995; Fiser and Sali, 2003) and ClusPro docking software (Comeau et al., 2004), respectively. SwissDock (http://swissdock.vital-it.ch/) web services (Grosdidier et al., 2011a,b) were used as a secondary verification method to determine the binding sites of Ce6 on rhodopsin. In this case, we assumed a blind docking estimate of the binding modes comprising the most favorable energies in the docking calculations. The highest populated cluster of Ce6-rhodopsin predicted a conformation with Ce6 bound to rhodopsin near the retinal-binding site, whereas the lowest energy cluster favored a ligand orientation in the cytoplasmic region with binding specifically favored close to residues in CL2. The output of the docking results were visualized with the UCSF Chimera (http://www.cgl.ucsf.edu/chimera/) molecular modeling system (Pettersen et al., 2004). Subsequent MD simulations for dark-state rhodopsin with Ce6 and Meta-II-Ce6 were carried out in an analogous manner to what has been described for the unbound receptors in the previous section. The simulations comprising Ce6 at the CP interface were the only conformations found to stably bind the allosteric ligand (Ce6) with a consistent binding site. Ce6 binds weakly in the CP region of inactive-state rhodopsin with an itinerant path limited to regions near the receptor C-terminus. In Meta-II, Ce6 has a clear, preferred binding site involving residues comprising CL2. The simulations containing Ce6 bound to regions near the receptor *retinal-binding site* found the ligand migrating away from the receptor within the first 10 ns in *all* MD simulations conducted and were not considered further for analyses conducted in this study. Using the RMSD based on the distances between structures, the g_cluster algorithim in gromacs was used to determine the range of accessible conformations in the Meta-II-Ce6 simulations (Palczewski et al., 2000; Menon et al., 2001; Ahuja et al., 2009).

Principal Component Analysis (PCA)

Principal component analysis or PCA is generally employed to detect correlations in large data sets. In MD simulations, the method can be utilized to reveal the most important motions in proteins. In this study, principal component analyses (PCAs) were carried out by diagonalzing the covariance matrix $C_{ij=\langle(x_i-\langle x_i\rangle)(x_j-\langle x_j\rangle)\rangle}$, where x denotes protein atomic positions in the $3N$-dimensional conformational space and the angular brackets represent the averages over the MD trajectory. Translational and rotational motions were removed by a least squares fitting to a reference structure. The eigenvectors of $\boldsymbol{C}$ were determined by diagonalization with an orthonormal transformation matrix. The resulting eigenvectors from the transformation were used to determine the PCA modes with eigenvalues (λ) equivalent to the variance in the direction of the corresponding eigenvector. The MD trajectory was projected onto the principal modes to determine the principal components. The eigenvalues λ_i of the principal components denote the mean square fluctuation of the principal component i and are arranged so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{3N}$. Using this arrangement, the trajectories were filtered along the first principal component to analyze the collective dynamics taking place within the protein. The cosine content of the PCA modes presented were found to be less than 0.001.

Determining Rhodopsin-Ce6 Ligand Binding Affinity from MD Simulation with an Alchemical Pathway Method

The ligand-binding affinity calculations were performed with Gromacs. The dissociation of the ligand from the receptor to determine the free energy of binding was determined by decoupling the van der Waal and Coulombic interactions of the ligand from the receptor by using an alchemical pathway (Boyce et al., 2009). Specifically, the alchemical pathway begins from the most accessible ligand-receptor conformation from the production run (unrestrained) MD simulation described previously. Subsequently, a harmonic restraint is added to the ligand to "pull" it away from the receptor ligand-binding site over a series of restrained conformational intermediates. The set of restraints used for pulling the ligand away from the receptor are described by one distance, two angles, and three dihedral harmonic potentials. Particularly, the distance restraint is defined by bonded terms between the ligand and the protein. In (dark-state) rhodopsin the distant restraint is described by the hydrogen-bond shared between the Pro347 backbone oxygen and hydrogen atom on one of the ligand pyrrole rings. For Meta-II the distance restraint is defined by a hydrogen-bonding interaction between the Arg147 side-chain and the nitrogen atom on the Ce6 pyrroline ring. The ligand restraints were comprised of 13 distributed λ values (or windows) of 0.0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.25, 0.3, 0.5, 0.75, and 1.0. Initially, all the intermediate states were equilibrated. Each window was energy minimized using the steepest descent algorithm. The receptor-ligand system was subsequently simulated for an additional 1.0 ns in the canonical ensemble with harmonic position restraints applied to the heavy atoms (of the receptor and ligand) with a force constant of 1000 kJ mol^{-1} nm^{-2}. Langevin dynamics (Goga et al., 2012) was used to set the reference temperature of the system to 300 K. Then, a 1 ns position restrained run in the isothermal–isobaric ensemble was conducted by using the Berendsen coupling algorithm for a target pressure of 1 atm. Finally, the production runs were carried out using Langevin dynamics for 30 ns. In all simulations, the particle mesh Ewald (PME) algorithm was used for electrostatic interactions with a real space cutoff of 1.2 nm, a spline order of 6, a relative tolerance of 10^{-6}, and a Fourier spacing of 0.10 nm. The Verlet cutoff scheme was used with a van der Waals cut-off of 1.2 nm and a tolerance of 0.005 kJ mol^{-1} ps^{-1}. A long-range dispersion correction for energy and pressure was employed and an *additional* long-range dispersion correction (EXP-LR) (Shirts and Chodera, 2008) was also utilized to compensate for using a van der Waals cutoff in a non-isotropic receptor-ligand system. Free energies were computed with the Bennett Acceptance Ratio (BAR) using the g_bar tool in Gromacs. The final free energy value was determined from the mean of all the free energy values from the separate simulations and the error computed from the standard deviation of the separate runs.

Localized Structural Fluctuations (LSFs)

Localized structural fluctuations (LSFs) are local relaxations that reflect specific intramolecular and intermolecular induced protein fluctuations. LSFs have also been hypothesized to form the basis of allosteric signal propagation in proteins (Daily and Gray, 2007). The localized structural fluctuations from the MD simulations carried out on rhodopsin were calculated with the method of Pandini et al. (2013) that utilizes a structural alphabet (SA) to define protein local structural fluctuations that are described by a set of 25 canonical states composed of four-residue protein fragments. The four-residue fragments define the most probable protein local, conformational fluctuations in the protein 3-D structure. Structural correlations between local conformational changes of two protein fragments were calculated as a positional mutual information (MI) matrix between two column positions in the SA alignment.

Multiple Sequence Alignment (MSA), Residue Conservation, and Coevolution Analysis

Multiple sequence alignment (MSA) is a powerful computational tool for uncovering the long-term evolutionary record of a protein family. The interdependence of evolutionary history or coevolution, which can be obtained from MSAs, is also used in this study to predict intermolecular communication between residue pairs and therefore, allosteric coupling that is related to protein functionality. The Class A Rhodopsin-like MSA data set was retrieved from the GPCR database (http://gpcrdb.org/). The reference sequence and structure were set as opsd_bovin with the PDB code 1 μ19. The conservation and co-evolutionary analyses on the rhodopsin-like family of sequences were carried out with the MISTIC server (http://mistic.leloir.org.ar/index.php). The mutual information score as implemented in MISTIC is calculated between pairs of columns in the MSA. The frequency for each amino acid pair is calculated using sequence weighting along with low count corrections and compared with the expected frequency. It is assumed that mutations between amino acids are uncorrelated. The MI score is calculated as a weighted sum of the log ratios between the observed and expected amino acid pair frequencies. The MI scores were translated into MI z-scores by comparing the MI values for each pair of positions with a distribution of prediction scores obtained from a large set of randomized MSAs (Buslje et al., 2009). The z-score is then calculated as the number of standard deviations that the observed MI value falls above the mean value obtained from the randomized MSAs. A z-score threshold of 6.5 describes a sensitivity of 0.4 and a specificity of 0.95. MISTIC lists every MI value between two residues with a value $\geq$6.5.

Visualization of Networks

Only the top 500 MI network links and nodes from the MSA were visualized. The position of residues in the two-dimensional MSA networks was computed with a combination of classical scaling and stress minimization (Brandes and Pich, 2009). And the network groupings were based on community detection resulting from modularity maximization (Newman, 2006). The network layout and grouping were calculated using the software tool Visone (http://visone.info/).

AUTHOR CONTRIBUTIONS

KW performed and analyzed the THz experimental measurements and conducted and analyzed the MD simulations and the MD simulation calculations. KW wrote the manuscript. JP carried out the network analyses on the MSA datasets and created the 2-D and 3-D LSF network visualizations from the MD simulations. JK-S supervised the rhodopsin sample preparation, discussed the manuscript with KW and edited the text of the manuscript.

FUNDING

JK-S was funded by Marie Curie Incoming International Fellowship (IIF) from the European Union.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2017.00085/full#supplementary-material

REFERENCES

Ahuja, S., Crocker, E., Eilers, M., Hornak, V., Hirshfeld, A., Ziliox, M., et al. (2009). Location of the retinal chromophore in the activated state of Rhodopsin*. *J. Biol. Chem.* 284, 10190–10201. doi: 10.1074/jbc.M805725200

Ala-Laurila, P., Donner, K., and Koskelainen, A. (2004). Thermal activation and photoactivation of visual pigments. *Biophys. J.* 86, 3653–3662. doi: 10.1529/biophysj.103.035626

Angel, T. E., Chance, M. R., and Palczewski, K. (2009). Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8555–8560. doi: 10.1073/pnas.0903545106

Balashov, S. P., Imasheva, E. S., Boichenko, V. A., Antón, J., Wang, J. M., and Lanyi, J. K. (2005). Xanthorhodopsin: a proton pump with a light-harvesting Carotenoid Antenna. *Science* 309, 2061–2064. doi: 10.1126/science.1118046

Balem, F., Yanamala, N., and Klein-Seetharaman, J. (2009). Additive effects of chlorin e6 and metal ion binding on the thermal stability of rhodopsin *in vitro*. *Photochem. Photobiol.* 85, 471–478. doi: 10.1111/j.1751-1097.2009.00539.x

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684. doi: 10.1063/1.448118

Berger, O., Edholm, O., and Jähnig, F. (1997). Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. *Biophys. J.* 72, 2002–2013.

Bondar, A.-N., Knapp-Mohammady, M., Suhai, S., Fischer, S., and Smith, J. (2011). Ground-State Properties of the Retinal Molecule: from Quantum Mechanical to Classical Mechanical Computations of Retinal Proteins. *Theor. Chem. Acc.* 130, 1169–1183. doi: 10.1007/s00214-011-1054-1

Boyce, S. E., Mobley, D. L., Rocklin, G. J., Graves, A. P., Dill, K. A., and Shoichet, B. K. (2009). Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.* 394, 747–763. doi: 10.1016/j.jmb.2009.09.049

Brandes, U., and Pich, C. (2009). "An experimental study on distance-based graph drawing," in *Graph Drawing Lecture Notes in Computer Science,* eds I. G. Tollis and M. Patrignani (Berlin; Heidelberg: Springer), 218–229.

Brown, M. F., Martínez-Mayorga, K., Nakanishi, K., Salgado, G. F. J., and Struts, A. V. (2009). Retinal conformation and dynamics in activation of rhodopsin illuminated by solid-state 2H NMR Spectroscopy. *Photochem. Photobiol.* 85, 442–453. doi: 10.1111/j.1751-1097.2008.00510.x

Buslje, C. M., Santos, J., Delfino, J. M., and Nielsen, M. (2009). Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* 25, 1125–1131. doi: 10.1093/bioinformatics/btp135

Chan, T., Lee, M., and Sakmar, T. P. (1992). Introduction of hydroxyl-bearing amino acids causes bathochromic spectral shifts in rhodopsin. Amino acid substitutions responsible for red-green color pigment spectral tuning. *J. Biol. Chem.* 267, 9478–9480.

Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinforma Oxf. Engl.* 20, 45–50. doi: 10.1093/bioinformatics/btg371

Daily, M. D., and Gray, J. J. (2007). Local motions in a benchmark of allosteric proteins. *Proteins* 67, 385–399. doi: 10.1002/prot.21300

Douglas, R. H., Genner, M. J., Hudson, A. G., Partridge, J. C., and Wagner, H.-J. (2016). Localisation and origin of the bacteriochlorophyll-derived photosensitizer in the retina of the deep-sea dragon fish *Malacosteus niger*. *Sci. Rep.* 6:39395. doi: 10.1038/srep39395

Douglas, R. H., Partridge, J. C., Dulai, K. S., Hunt, D. M., Mullineaux, C. W., and Hynninen, P. H. (1999). Enhanced retinal longwave sensitivity using a chlorophyll-derived photosensitiser in Malacosteus niger, a deep-sea dragon fish with far red bioluminescence. *Vision Res.* 39, 2817–2832.

Douglas, R. H., Partridge, J. C., Dulai, K., Hunt, D., Mullineaux, C. W., Tauber, A. Y., et al. (1998). Dragon fish see using chlorophyll. *Nature* 393, 423–424. doi: 10.1038/30871

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* 103, 8577–8593. doi: 10.1063/1.470117

Fahmy, K., Jäger, F., Beck, M., Zvyaga, T. A., Sakmar, T. P., and Siebert, F. (1993). Protonation states of membrane-embedded carboxylic acid groups in rhodopsin and metarhodopsin II: a Fourier-transform infrared spectroscopy study of site-directed mutants. *Proc. Natl. Acad. Sci. U.S.A.* 90, 10206–10210.

Farrens, D. L., and Khorana, H. G. (1995). Structure and function in rhodopsin. Measurement of the rate of metarhodopsin II decay by fluorescence spectroscopy. *J. Biol. Chem.* 270, 5073–5076.

Fenwick, R. B., Esteban-Martín, S., and Salvatella, X. (2011). Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur. Biophys. J.* 40, 1339–1355. doi: 10.1007/s00249-011-0754-8

Fiser, A., and Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. *Meth. Enzymol.* 374, 461–491. doi: 10.1016/S0076-6879(03)74020-8

Fishkin, N., Berova, N., and Nakanishi, K. (2004). Primary events in dim light vision: a chemical and spectroscopic approach toward understanding protein/chromophore interactions in rhodopsin. *Chem. Rec.* 4, 120–135. doi: 10.1002/tcr.20000

Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science* 254, 1598–1603. doi: 10.1126/science.1749933

Fritze, O., Filipek, S., Kuksa, V., Palczewski, K., Hofmann, K. P., and Ernst, O. P. (2003). Role of the conserved NPxxY(x)$_{5,6}$F motif in the rhodopsin ground state and during activation. *Proc. Natl. Acad. Sci. U.S.A.* 100, 2290–2295. doi: 10.1073/pnas.0435715100

Goga, N., Rzepiela, A. J., de Vries, A. H., Marrink, S. J., and Berendsen, H. J. C. (2012). Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* 8, 3637–3649. doi: 10.1021/ct3000876

Goodsell, D. S., Morris, G. M., and Olson, A. J. (1996). Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* 9, 1–5. doi: 10.1002/(SICI)1099-1352(199601)9:1<1::AID-JMR24>3.0.CO;2-6

Grosdidier, A., Zoete, V., and Michielin, O. (2011a). Fast docking using the CHARMM force field with EADock, D. *S. S. J. Comput. Chem.* 32, 2149–2159. doi: 10.1002/jcc.21797

Grosdidier, A., Zoete, V., and Michielin, O. (2011b). SwissDock, a protein-small molecule docking web service based on EADock, D. *S. S. Nucleic Acids Res.* 39, W270–W277. doi: 10.1093/nar/gkr366

Hawkins, R. J., and McLeish, T. C. B. (2006). Coupling of global and local vibrational modes in dynamic allostery of proteins. *Biophys. J.* 91, 2055–2062. doi: 10.1529/biophysj.106.082180

Heck, M., Schädel, S. A., Maretzki, D., and Hofmann, K. P. (2003b). Secondary binding sites of retinoids in opsin: characterization and role in regeneration. *Vision Res.* 43, 3003–3010. doi: 10.1016/j.visres.2003.08.011

Heck, M., Schädel, S. A., Maretzki, D., Bartl, F. J., Ritter, E., Palczewski, K., et al. (2003a). Signaling states of rhodopsin. Formation of the storage form, metarhodopsin III, from active metarhodopsin II. *J. Biol. Chem.* 278, 3162–3169. doi: 10.1074/jbc.M209675200

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472. doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Hofmann, L., and Palczewski, K. (2015). Advances in understanding the molecular basis of the first steps in color vision. *Prog. Retin. Eye Res.* 49, 46–66. doi: 10.1016/j.preteyeres.2015.07.004

Hwa, J., Reeves, P. J., Klein-Seetharaman, J., Davidson, F., and Khorana, H. G. (1999). Structure and function in rhodopsin: further elucidation of the role of the intradiscal cysteines, Cys-110,−185, and−187, in rhodopsin folding and function. *Proc. Natl. Acad. Sci. U.S.A.* 96, 1932–1935.

Imamoto, Y., and Shichida, Y. (2014). Cone visual pigments. *Biochim. Biophys. Acta* 1837, 664–673. doi: 10.1016/j.bbabio.2013.08.009

Isayama, T., Alexeev, D., Makino, C. L., Washington, I., Nakanishi, K., and Turro, N. J. (2006). An accessory chromophore in red vision. *Nature* 443, 649–649. doi: 10.1038/443649a

Janz, J. M., Fay, J. F., and Farrens, D. L. (2003). Stability of Dark State Rhodopsin Is Mediated by a Conserved Ion Pair in Intradiscal Loop E-2. *J. Biol. Chem.* 278, 16982–16991. doi: 10.1074/jbc.M210567200

Kenaley, C. P., Devaney, S. C., and Fjeran, T. T. (2014). The complex evolutionary history of seeing red: molecular phylogeny and the evolution of an adaptive visual system in deep-sea dragonfishes (Stomiiformes: Stomiidae). *Evol. Int. J. Org. Evol.* 68, 996–1013. doi: 10.1111/evo.12322

Kimura, S. (1987). Photobiology of pheophorbide. *Photomed. Photobiol.* 9, 35–47.

Kusnetzow, A. K., Altenbach, C., and Hubbell, W. L. (2006). Conformational states and dynamics of rhodopsin in micelles and bilayers. *Biochem. Mosc.* 45, 5538–5550. doi: 10.1021/bi060101v

Lewis, J. W., Szundi, I., Kazmi, M. A., Sakmar, T. P., and Kliger, D. S. (2006). Proton movement and photointermediate kinetics in rhodopsin mutants. *Biochem. Mosc.* 45, 5430–5439. doi: 10.1021/bi0525775

Lin, S. W., Kochendoerfer, G. G., Carroll, K. S., Wang, D., Mathies, R. A., and Sakmar, T. P. (1998). Mechanisms of spectral tuning in blue cone visual pigments visible and raman spectroscopy of blue-shifted rhodopsin mutants. *J. Biol. Chem.* 273, 24583–24591. doi: 10.1074/jbc.273.38.24583

Menon, S. T., Han, M., and Sakmar, T. P. (2001). Rhodopsin: structural basis of molecular physiology. *Physiol. Rev.* 81, 1659–1688.

Motlagh, H. N., Wrabl, J. O., Li, J., and Hilser, V. J. (2014). The ensemble nature of allostery. *Nature* 508, 331–339. doi: 10.1038/nature13001

Nathans, J. (1990). Determinants of visual pigment absorbance: identification of the retinylidene Schiff's base counterion in bovine rhodopsin. *Biochem. Mosc.* 29, 9746–9752.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Nygaard, R., Valentin-Hansen, L., Mokrosinski, J., Frimurer, T. M., and Schwartz, T. W. (2010). Conserved water-mediated hydrogen bond network between TM-I, -II, -VI, and -VII in 7TM receptor activation. *J. Biol. Chem.* 285, 19625–19636. doi: 10.1074/jbc.M110.106021

Oprian, D. D., Molday, R. S., Kaufman, R. J., and Khorana, H. G. (1987). Expression of a synthetic bovine rhodopsin gene in monkey kidney cells. *Proc. Natl. Acad. Sci. U.S.A.* 84, 8874–8878.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., et al. (2000). Crystal structure of rhodopsin,: AG protein-coupled receptor. *Science* 289, 739–745. doi: 10.1126/science.289.5480.739

Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. (2013). GSATools: analysis of allosteric communication and functional local motions using a structural alphabet. *Bioinformatics* 29, 2053–2055. doi: 10.1093/bioinformatics/btt326

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Rajamani, R., Lin, Y.-L., and Gao, J. (2011). The opsin shift and mechanism of spectral tuning in rhodopsin. *J. Comput. Chem.* 32, 854–865. doi: 10.1002/jcc.21663

Reeves, P. J., Kim, J.-M., and Khorana, H. G. (2002). Structure and function in rhodopsin: a tetracycline-inducible system in stable mammalian cell lines for high-level expression of opsin mutants. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13413–13418. doi: 10.1073/pnas.212519199

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. *Proteins* 23, 318–326. doi: 10.1002/prot.340230306

Shirts, M. R., and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* 129:124105. doi: 10.1063/1.2978177

Stehle, J., Silvers, R., Werner, K., Chatterjee, D., Gande, S., Scholz, F., et al. (2014). Characterization of the simultaneous decay kinetics of metarhodopsin states II and III in rhodopsin by solution-state NMR spectroscopy. *Angew. Chem. Int. Ed Engl.* 53, 2078–2084. doi: 10.1002/anie.201309581

Tsukamoto, H., Terakita, A., and Shichida, Y. (2010). A pivot between helices V and VI near the retinal-binding site is necessary for activation in Rhodopsins. *J. Biol. Chem.* 285, 7351–7357. doi: 10.1074/jbc.M109.078709

Vogel, R., Siebert, F., Mathias, G., Tavan, P., Fan, G., and Sheves, M. (2003). Deactivation of rhodopsin in the transition from the signaling state meta II to meta III involves a thermal isomerization of the retinal chromophore C=N Double Bond. *Biochem. Mosc.* 42, 9863–9874. doi: 10.1021/bi034684+

Vogel, R., Siebert, F., Zhang, X.-Y., Fan, G., and Sheves, M. (2004). Formation of Meta III during the Decay of Activated Rhodopsin Proceeds via Meta I and Not via Meta, I. I. *Biochemistry (Mosc)* 43, 9457–9466. doi: 10.1021/bi049337u

Warrant, E. (2004). Vision in the dimmest habitats on earth. *J. Comp. Physiol. A. Neuroethol. Sens. Neural Behav. Physiol.* 190, 765–789. doi: 10.1007/s00359-004-0546-z

Washington, I., Brooks, C., Turro, N. J., and Nakanishi, K. (2004). Porphyrins as photosensitizers to enhance night vision. *J. Am. Chem. Soc.* 126, 9892–9893. doi: 10.1021/ja0486317

Washington, I., Zhou, J., Jockusch, S., Turro, N. J., Nakanishi, K., and Sparrow, J. R. (2007). Chlorophyll derivatives as visual pigments for super vision in the red. *Photochem. Photobiol. Sci. Off. J. Eur. Photochem. Assoc. Eur. Soc. Photobiol.* 6, 775–779. doi: 10.1039/b618104j

Wexler, A., and Hasegawa, S. (1954). Relative humidity-temperature relationships of some saturated salt solutions in the temperature range 0 to 50 C. *J. Res. Natl. Bur. Stand.* 53, 19–26.

Whitten, S. T., García-Moreno E. B. and Hilser, V. J. (2005). Local conformational fluctuations can modulate the coupling between proton binding and global structural transitions in proteins. *Proc. Natl. Acad. Sci. U.S.A.* 102, 4282–4287. doi: 10.1073/pnas.0407499102

Wolf, S., and Grünewald, S. (2015). Sequence, structure and ligand binding evolution of rhodopsin-like G protein-coupled receptors: a crystal structure-based phylogenetic analysis. *PLoS ONE* 10:e0123533. doi: 10.1371/journal.pone.0123533

Woods, K. N. (2010). Solvent-induced backbone fluctuations and the collective librational dynamics of lysozyme studied by terahertz spectroscopy. *Phys. Rev.* E 81:031915. doi: 10.1103/PhysRevE.81.031915

Woods, K. N. (2014a). The glassy state of crambin and the THz time scale protein-solvent fluctuations possibly related to protein function. *BMC Biophys.* 7:8. doi: 10.1186/s13628-014-0008-0

Woods, K. N. (2014b). Using THz time-scale infrared spectroscopy to examine the role of collective, thermal fluctuations in the formation of myoglobin allosteric communication pathways and ligand specificity. *Soft Matter* 10, 4387–4402. doi: 10.1039/c3sm53229a

Woods, K. N., Pfeffer, J., Dutta, A., and Klein-Seetharaman, J. (2016). Vibrational resonance, allostery, and activation in rhodopsin-like G protein-coupled receptors. *Sci. Rep.* 6:37290. doi: 10.1038/srep37290

Yanagawa, M., Kojima, K., Yamashita, T., Imamoto, Y., Matsuyama, T., Nakanishi, K., et al. (2015). Origin of the low thermal isomerization rate of rhodopsin chromophore. *Sci. Rep.* 5:11081. doi: 10.1038/srep11081

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 27 December 2016
doi: 10.3389/fmolb.2016.00085

Comparative Normal Mode Analysis of the Dynamics of DENV and ZIKV Capsids

Yin-Chen Hsieh[1], Frédéric Poitevin[2,3], Marc Delarue[4] and Patrice Koehl[1]*

[1] Department of Computer Science and Genome Center, University of California, Davis, Davis, CA, USA, [2] Department of Structural Biology, Stanford University, Stanford, CA, USA, [3] SLAC National Accelerator Laboratory, Stanford PULSE Institute, Menlo Park, CA, USA, [4] Unit of Structural Dynamics of Macromolecules, UMR 3528 du Centre National de la Recherche Scientifique, Institut Pasteur, Paris, France

Key steps in the life cycle of a virus, such as the fusion event as the virus infects a host cell and its maturation process, relate to an intricate interplay between the structure and the dynamics of its constituent proteins, especially those that define its capsid, much akin to an envelope that protects its genomic material. We present a comprehensive, comparative analysis of such interplay for the capsids of two viruses from the *flaviviridae* family, Dengue (DENV) and Zika (ZIKV). We use for that purpose our own software suite, DD-NMA, which is based on normal mode analysis. We describe the elements of DD-NMA that are relevant to the analysis of large systems, such as virus capsids. In particular, we introduce our implementation of simplified elastic networks and justify their parametrization. Using DD-NMA, we illustrate the importance of packing interactions within the virus capsids on the dynamics of the E proteins of DENV and ZIKV. We identify differences between the computed atomic fluctuations of the E proteins in DENV and ZIKV and relate those differences to changes observed in their high resolution structures. We conclude with a discussion on additional analyses that are needed to fully characterize the dynamics of the two viruses.

Keywords: proteins, normal modes, elastic network models, viruses, Dengue, Zika

OPEN ACCESS

Edited by:
Slavica Jonic,
Université Pierre et Marie Curie,
France

Reviewed by:
Yves-henri Sanejouand,
Unité Fonctionnalité et Ingénierie des Protéines (CNRS), France
Xavier Siebert,
University of Mons, Belgium

****Correspondence:***
Patrice Koehl
koehl@cs.ucdavis.edu

Specialty section:
This article was submitted to Biophysics,
a section of the journal
Frontiers in Molecular Biosciences

Received: *25 October 2016*
Accepted: *12 December 2016*
Published: *27 December 2016*

Citation:
Hsieh Y-C, Poitevin F, Delarue M and Koehl P (2016) Comparative Normal Mode Analysis of the Dynamics of DENV and ZIKV Capsids.
Front. Mol. Biosci. 3:85.
doi: 10.3389/fmolb.2016.00085

1. INTRODUCTION

A major goal of molecular biology is to understand at the atomic level the functions of macromolecules and/or biological nano-machines, which are believed to be intimately related to the dynamics of their three-dimensional structures and especially their collective degrees of freedom (Koehl, 2014; Bahar et al., 2015). Our current understanding of the dynamics of macromolecules is, however, largely incomplete. This arises because only a few experimental techniques are capable of collecting time-resolved structural data, and those that can collect those data are usually limited to a narrow time window (Fromme, 2015). Similarly, state-of-the-art computational methods are limited in scope (usually in the microsecond time-scale), because of limitations in computing power (Fengand et al., 2015).

An alternate and promising approach to molecular dynamics is to infer dynamics from static structures corresponding to locally stable states (Mahajan and Sanejouand, 2015), together with reliable coarse-graining approaches to bridge the time-scale gap (Saunders and Voth, 2013; López-Blanco and Chacón, 2016). Cartesian Normal Modes, for example, represent a class of movements around a local energy minimum that are both straightforward to calculate and biologically relevant

(Noguti and Go, 1982; Brooks et al., 1983; Levitt et al., 1985). The low-frequency part of the spectrum of normal modes is often associated with functional transitions, for instance, between two known states of the same macromolecule such as its apo (ligand-free) or holo (bound) form. The Elastic Network Model (ENM), introduced by Tirion in 1996, offers a particularly simple and efficient way to calculate these modes, allowing fast access to the collective dynamics of large complexes with no minimization issues as it enforces that the crystal structure is already at the energy minimum (Tirion, 1996). This model was later expanded as the Gaussian Network Model (Bahar et al., 1997) and the Anisotropic Network Model (Hinsen, 1998; Tama et al., 2000; Atilgan et al., 2001), which were shown to describe conformational changes remarkably well (Tama and Sanejouand, 2001; Delarue and Sanejouand, 2002; Zheng and Doniach, 2003; Mahajan and Sanejouand, 2015).

During the past few years, several web-servers performing on-line Normal Mode Analysis (NMA) have been set up and described: ElNemo (Suhre and Sanejouand, 2004), ENCoM (Frappier et al., 2015), Webnm@ (Tiwari et al., 2014), ANM 2.0 (Eyal et al., 2015), AD-ENM (Zheng and Doniach, 2003), NMSim (Kruger et al., 2012). We have extended and updated our own server, NOMAD-REF (Lindahl et al., 2006), with a new and user-friendlier interface, including a better visual representation of the results while at the same time enlarging the performances of the core calculation of Normal Modes in the framework of the ENM representation. New features include (i) a wider array of coarse-graining levels prior to the actual building of the ENM, and (ii) variants of the ENM that are based on a cutoff-free Delaunay tessellation of the set of atoms of the molecule of interest. With these features we depart from the original Elastic Network Model (Tirion, 1996), but keep most of its salient features, as the construction of the original Tirion Elastic Model remains available. We found for example that the Elastic Network coming from a Delaunay tessellation correctly handles PDB models with isolated domains and/or dangling ends (Xia et al., 2014). In addition, the performance of the calculation of Normal Modes has been improved to a point where it can deal with 100,000 atoms routinely, making it possible, for instance, to deal with entire virus capsids without having to resort to a symmetry-specific implementation (Simonson and Perahia, 1992; van Vlijmen and Karplus, 2005; Peeters and Taormina, 2009).

In the present paper, we show an application of some of the tools implemented in DD-NMA, the updated version of NOMAD-REF, to study the dynamics of viruses of the *flaviviridae* family, namely of Dengue virus and Zika virus.

Dengue virus (DENV) is a positive-sense RNA virus responsible for dengue fever, a tropical infectious disease whose incidence has increased drastically over the last decades, for which no prophylactic treatments are known (with the exception of eliminating the vector, i.e., mosquitoes). Today, about 3.9 billion people, or 50% of the world's population, live in areas where there is a risk of dengue transmission (Brady et al., 2012). Dengue is endemic in at least 128 countries in Asia, the Pacific, the Americas, Africa, and the Caribbean (Brady et al., 2012). The World Health Organization (WHO) estimates that close to 390 million infections occur yearly, of which 96 million manifest clinically (Bhatt et al., 2013). DENV is recognized as a potential threat to public health in the USA (Morens and Fauci, 2008). Of similar concerns are the recent outbreaks of ZIKA virus (ZIKV), another *flaviviridae* virus similar to DENV. The current ZIKV epidemic in the Americas is linked to a sudden increase in the reported cases of congenital microcephaly and Guillain Barré syndrome. This led the World Health Organization (WHO) in February 2016 to declare a "public health emergency of international concern" (WHO, 2016). As no treatments currently exist for the consequences of infections with those two viruses, and as their incidence is only expected to increase, basic research on their infection mechanisms becomes highly significant.

Flaviviridae genomes encode for ten different proteins, three structural proteins that form the virus particle, and seven non-structural (NS) proteins that are involved in its replication (for recent review see Meng et al., 2015). Structures of all four serotypes of DENV (Perera and Kuhn, 2008 and references therein; Zhang et al., 2012; Kostyuchenko et al., 2013, 2014; Fibriansah et al., 2015) and recently two structures of the same ZIKV strain have been published (Kostyuchenko et al., 2016; Sirohi et al., 2016). Those structures show the same global architecture, with their capsids having icosahedral symmetry consisting of 60 units, with each unit containing three copies of the E protein and three copies of protein M. The E protein is known to play a central role in many parts of the virus life cycle (Perera and Kuhn, 2008). A perhaps surprising idea that has emerged from years of studies of viruses is that their biology is deeply encoded in the dynamics of these proteins. Significant structural dynamics has been shown to occur during infection cycles, both at the level of individual proteins and at the quaternary structure level of the viral particle. These dynamics can be blocked by antibody binding (Lok et al., 2008; Teoh et al., 2012; Fibriansah et al., 2015). In addition, while the overall geometry of the viral capsid is identical in all those viruses and only small differences are observed at a finer structural scale, significant differences in stability are observed between those viruses. For example, while infection with DENV is significantly affected by temperature, infection with ZIKV remains constant at even relatively high temperatures (Kostyuchenko et al., 2016). To better understand differences between those two viruses, we investigate the dynamics of their capsid E proteins. We study those proteins independently, as well as the impact of packing imposed by the icosahedral arrangement of the virus capsid. We explore whether the differences observed, if any, are consistent with the differences observed between the structures of the capsids of DENV and ZIKV and their biological activities.

The paper is organized as follows. In the next section, we describe normal mode analysis (NMA) in the context of the Elastic Network Model. We provide an overview of the theory and discuss the different options for choosing its parameters, namely the choice of coarse-graining level, the choice of the elastic force constants, and the cutoff for selecting the pairs of atoms that belong to the ENM. In the following section, we provide a description of the algorithms used to implement NMA within our new server DD-NMA, with a special focus on scalability to large molecular systems. In the Results section, we discuss the applications of DD-NMA to study the dynamics of DENV and ZIKV, focusing on the differences and similarities of the dynamics of their capsid E protein. We conclude the paper with a brief discussion on future developments of normal mode analysis applied to viral structures.

2. NORMAL MODE ANALYSIS

2.1. Normal Mode Analysis Based on the Tirion Elastic Network Model

The Elastic Network Model (ENM) was originally introduced by Tirion (1996). It is a model that captures the geometry of the molecule of interest in the form of a network of inter-atomic connections, linked together with elastic springs. Two categories of normal mode analyses based on ENMs are widely used today, namely, the Gaussian Network Model (GNM) (Bahar et al., 1997; Haliloglu et al., 1997) and the anisotropic network model (ANM) (Tirion, 1996; Atilgan et al., 2001). Here we follow the latter model, in which the energy of the molecule is equated to the harmonic energy associated with these springs. This defines a quadratic energy on the inter-atomic distances. Let M be a biomolecule containing N atoms, with atom i characterized by its position $X_i = (x_i, y_i, z_i)$. The whole molecule is then described by a $3N$ position vector X. For two atoms i and j of M, we set $r_{ij} = |X_i - X_j|$ and $r^0_{ij} = |X^0_i - X^0_j|$ to be their Euclidean distances in any conformation X and in the ground-state conformation X^0 (usually the X-ray structure), respectively. The total potential V_{ENM} of the biomolecule is then set to:

$$V_{ENM}(X) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j>i} k_{ij}(r_{ij} - r^0_{ij})^2\Theta(R_c - r^0_{ij}) \tag{1}$$

where R_c is a cutoff distance, k_{ij} is the force constant of the "spring" formed by the pair of atoms i and j, and $\Theta(x)$ is the Heaviside unit step function, i.e., $\Theta(x) = 0$ if $x < 0$ and $\Theta(x) = 1$ otherwise. Both R_c and k_{ij} are discussed in detail below.

In the normal mode framework, the potential V_{ENM} is then approximated with a second-order Taylor expansion in the neighborhood of the ground state X^0:

$$\begin{aligned} V_{ENM}(X) \approx\ & V_{ENM}(X^0) + \nabla V_{ENM}(X^0)^T(X - X^0) \\ & + \frac{1}{2}(X - X^0)^T H (X - X^0) \end{aligned} \tag{2}$$

where ∇V_{ENM} and H are the gradient and Hessian of V_{ENM}, respectively. Note that based on Equation 1, $V_{ENM}(X^0) = 0$ and $\nabla V_{ENM}(X^0)$ is the null vector (i.e., X^0 is a global minimum of V_{ENM} by definition). The ENM energy is then simply

$$V_{ENM}(X) \approx \frac{1}{2}(X - X^0)^T H (X - X^0) \tag{3}$$

The 3×3 submatrix Hij of the Hessian H corresponding to two atoms i and j that are in contact is given by:

$$\begin{aligned} Hij &= -\frac{k_{ij}}{(r^0_{ij})^2}(X_i - X_j)(X_i - X_j)^T \\ &= -\frac{k_{ij}}{(r^0_{ij})^2}\begin{bmatrix} (x_i - x_j)^2 & (x_i - x_j)(y_i - y_j) & (x_i - x_j)(z_i - z_j) \\ (y_i - y_j)(x_i - x_j) & (y_i - y_j)^2 & (y_i - y_j)(z_i - z_j) \\ (z_i - z_j)(x_i - x_j) & (z_i - z_j)(y_i - y_j) & (z_i - z_j)^2 \end{bmatrix} \end{aligned} \tag{4}$$

and the 3×3 submatrix Hii on the diagonal of H is then given by:

$$Hii = -\sum_{j=1,N} Hij \tag{5}$$

In Cartesian coordinates, the equations of motion defined by the potential V_{ENM} are derived from Newton's equation:

$$\frac{d^2X}{dt^2} = -H(X - X^0) \tag{6}$$

Writing the solution to this equation as a linear sum of intrinsic motions (the "normal modes" of the system),

$$X_j = \sum_{k=k_0}^{3N} A_{jk}\alpha_k cos(\omega_k t + \delta_k) \tag{7}$$

we get a standard eigenvalue problem,

$$HA = MA\Omega \tag{8}$$

The eigenfrequencies ω are given by the elements of the diagonal matrix Ω, namely $\omega_i^2 = \Omega(i,i)$. The eigenvectors are the columns of the matrix A, and the amplitudes and phases, α_k and δ_k, are determined by initial conditions. The matrix M is a diagonal matrix containing the masses of the atoms. We note that the first six eigenvalues in Ω are equal to 0, as they correspond to global translations and rotations of the biomolecule. To characterize the internal motions of the biomolecule, the sum in Equation 8 runs then from $k_0 = 7$ up to $3N$, the number of degrees of freedom of the system.

2.2. Parametrization: Choosing the Representation of the Molecule

The first requirement when building an ENM is to define the set of atoms on which it is based. Although all atoms could be used, it appears natural to lower the dimensionality of the system, namely "coarse-graining," when large biomolecules are considered, or in the context of a harmonic approximation to its energy as is the case in ENM (Tozzini, 2005). Coarse-grained models have long been used for studying protein folding and aggregation. They enable the exploration of large length scales and time scales that are usually inaccessible to all-atom models in explicit solvent (Saunders and Voth, 2013; Kmiecik et al., 2016). Combined with enhanced configuration search methods, these simplified models with various levels of granularity offer the possibility to determine equilibrium structures and to compare folding kinetics and thermodynamics quantities with the corresponding values obtained by experimental techniques. In their pioneer work from 1976, Levitt and Warshel (1976) developed the foundation of coarse-graining for protein folding. They were able to fold the 58-residue BPTI protein within 6.5 Å from its experimental structure using a two-bead representation for each residue in the protein. This representation included the Cα and the centroid of the side chain to define a residue. They used an effective implicit solvent force field such that the atoms of the solvent need not be considered explicitly, and successive minimizations and normal mode thermalization to fold BPTI. Since then, various levels of granularity have been developed, from lattice representations to multi-bead representations, and from single atom to multiple-atom residue-level representations (Kmiecik et al., 2016). The positions of those beads are either defined by known atoms (usually the Cα), or by fitting to capture the dynamics of the full molecular systems (Zhang et al., 2008, 2011; Li et al., 2016). For all the analyses of virus structures considered in this paper, we used the Cα-only representation.

2.3. Parametrization: Choosing the Spring Force Constants

In the original ENM introduced by Tirion, the elastic constants k_{ij} are set to be the same for all pairs of atoms. In other models, k_{ij} vary for different pairs of atoms. For example, Ming and Wall (2005) employed an enhanced ENM in which the interactions of neighboring Cα atoms on the backbone were strengthened to reproduce the correct bimodal distribution of density-of-states from an all-atom model. Kondrashov et al. (2006) used a strategy in which they classified residue interactions into several categories corresponding to different physical properties. The elastic constants can also be adjusted to have the fluctuations of the atoms of the molecule computed from the equations of motions given by Equation (7) to match the atomic fluctuations captured experimentally and usually reported as B-factors. Many methods have been developed for that purpose (see for example Xia et al., 2013, 2014 and references therein). Among those methods, the one proposed by Erman (2006) is worth discussing. Erman developed an iterative algorithm to update the Kirchhoff matrix of a Gaussian Network Model, in which the connections of neighboring Cα atoms on the backbone of the protein of interest are fixed, and the strengths of the interactions between pairs of residues are varied until a good fit between experimental B-factors and computed fluctuations is obtained. While this approach generates a really good fit between those two representations of fluctuations, a significant number of the optimized spring force constants are found to be negative. While such negative values are not forbidden, they do hint at the possibility of overfitting. This is in accordance with (Fuglebakk et al., 2013), who recently suggested that such a refinement procedure leads to overfitting, and not to a better dynamic model for the molecule. As such, in this study we assign the same value for all k_{ij}, following the initial ENM of Tirion (1996).

2.4. Parametrization: The Cutoff Parameter R_c

In standard implementations, the cutoff distance R_c and the force constant k are set constant for all pairs of residues. Their values, however, differ between the two models. For example, the cutoff distance R_c for GNM is usually set in the range of 7 to 8 Å (Kundu et al., 2002) while in ANM larger values are usually considered in the range from 13 to 15 Å (Eyal et al., 2006). There are, however, no guidelines as to which values are best and sometimes different implementations lead to contradicting optimal values. To circumvent these discrepancies, several authors have proposed to include all pairs of residues in a protein and to assign different force constants to their corresponding springs that relate to their lengths at rest (see for example Hinsen, 1998; Kovacs et al., 2004; Yang et al., 2009). In these methods, the use of a plain cutoff distance is avoided. The number of pairs of atoms considered, however, is large and scales as N^2, where N is either the total number of atoms in the biomolecule considered, or its number of residues. Such a quadratic behavior makes these methods unfit for studying large systems. To study the capsids of DENV and ZIKA, we have considered a traditional cutoff ENM, with the cutoff set to 14 Å, unless specifically noted.

3. MATERIALS AND METHODS

We have used our own software package, DD-NMA, to perform all the analyses discussed in the Results section. In the following, we highlight some of the elements of DD-NMA that are relevant to the analysis of large systems. We note that DD-NMA is available as a web-based service at http://lorentz.dynstr.pasteur.fr/suny/index.php?id0=delaunaynma#welcome.

3.1. An Efficient Algorithm to Diagonalize the Hessian of the Elastic Potential V_{ENM}

The Hessian matrix of V_{ENM} is a $3N \times 3N$ symmetric, real-valued matrix whose elements are described by Equation (4). The theory described above calls for diagonalizing this matrix, as its eigenvalues and eigenvectors provide the frequencies and directions, respectively, of the normal modes of the molecular systems under study. While many methods exist for solving such an eigenvalue problem, see Golub and van der Vorst (2000), many of those methods break down when N becomes large, both in terms of computing time and memory requirements. The Hessian matrix is highly sparse as only a subset of all atom pairs are usually considered (see previous section for a discussion of this point), but this is not enough to offset the computing requirements as the matrix A of eigenvectors is usually not sparse. However, in her original paper, Tirion (1996) had recognized that the lowest frequency normal modes can capture most of the dynamics of the protein of interest. This observation has since been supported by further evidence that the lowest-frequency normal modes generated from ENM conform with conformational changes observed by X-ray and NMR experiments (Kim et al., 2002; Maragakis and Karplus, 2005; Kurkcuoglu et al., 2009) as well as with the results of MD simulations (Rueda et al., 2007; Orellana et al., 2010; Leioatts et al., 2012). While it is unclear as to how many of those low frequency normal modes need to be considered (Petrone and Pande, 2006), it remains that only a small fraction of the eigenvalues and eigenvectors of the Hessian matrix need to be computed, which leads to the opportunity to use powerful iterative algorithms for computing those quantities. The most successful family of such algorithms is based on the efficient Krylov subspace method, as it allows for targeting only a subset of the eigenvalue spectrum of a matrix. We provide below the rationale behind this method to compute the eigenvalues with lowest magnitude of the Hessian matrix.

An intuitive method for finding the largest eigenvalue of a given $N \times N$ matrix A is the power iteration. Starting with an initial random vector x, this method calculates Ax, A^2x, A^3x,... iteratively, storing and normalizing the result into x at every iteration. The corresponding sequence of Rayleigh quotient R_i

$$R_i = \frac{x^T A^i x}{x^T x} \tag{9}$$

converges to the largest eigenvalue of A, while x itself converges to the corresponding eigenvector. However, much potentially useful computation is wasted by using only the final result. This suggests that, instead, the so-called Krylov matrix is to be formed:

$$\mathcal{K}_n = \left[x \; Ax \; A^2x \; \dots \; A^{n-1}x \right] \tag{10}$$

The columns of this matrix are not orthogonal, but an orthogonal basis can be constructed via a stabilized Gram–Schmidt orthogonalization. The resulting vectors are a basis of the Krylov subspace, $\mathcal{K}_n$. The vectors of this basis give good approximations of the eigenvectors corresponding to the n largest eigenvalues of A. In a similar manner, the smallest eigenvalues of A can be computed by applying this strategy to either A^{-1}, or by applying a spectral shift, i.e., by computing the largest eigenvalues of $A - \lambda_{max}I$, where λ_{max} is the largest eigenvalue of A.

We use the ARPACK implementation of a variant of this approach, the implicitly restarted Arnoldi iteration method (Lehoucq et al., 1998).

3.2. Atomic Fluctuations Computed From Normal Modes

From the normal modes of the ENM, it is possible to compute the mean square fluctuations of the positions of the atoms according to:

$$< \Delta \mathbf{X}_i^2 > = \frac{k_B T}{m_i} \sum_{k=7}^{m} \frac{A_{ik}^2}{\omega_k^2} \tag{11}$$

where $\Delta \mathbf{X}_i$ and m_i are the displacement vector and mass of vertex i, respectively, k_B is the Boltzmann's constant, T is the temperature considered, A_{ij} is the i-th component of the j eigenvector A_j of the Hessian, and ω_i is its associated eigenvalue. The summation should run over all the modes of the system (excluding the six modes for rigid body transformations); it is truncated here to the first $m = 100$ modes, as those low frequency modes are usually responsible of most of the atomic fluctuations (see above).

3.3. Correlated Motions Within a Biomolecule

The Boltzmann distribution for the approximate potential of the ENM (see Equation 3) is described by a multivariate Gaussian distribution with a covariance matrix proportional to the inverse of the Hessian H. Because of the six rigid motions captured by the six normal modes with 0 frequencies, the inverse of H is in fact not properly defined. We can, however, compute a pseudo-inverse by ignoring those zero energy modes; this pseudo-inverse can be regarded as a covariance matrix of internal deformation:

$$C = \sum_{k=7}^{M} \frac{1}{\omega_k^2} A_k A_k^T \tag{12}$$

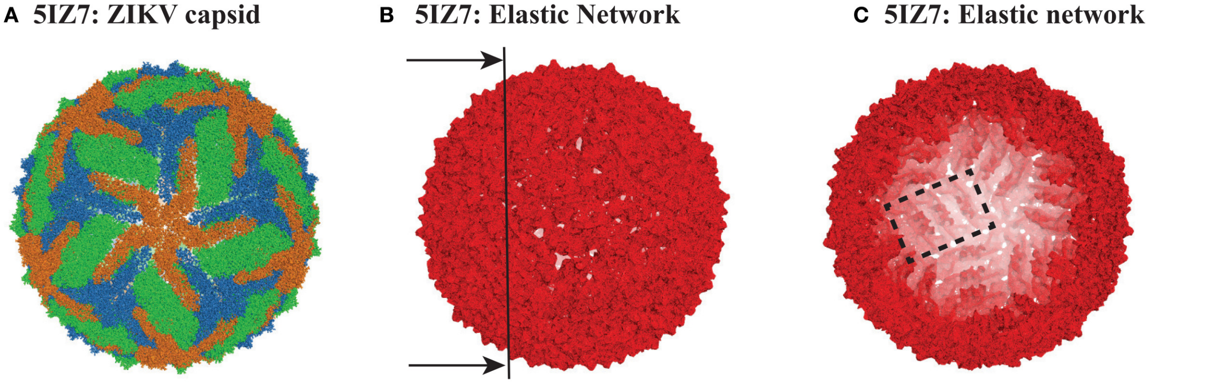


FIGURE 1 | The capsid of ZIKV. (A) Cartoon representation of the capsid of ZIKV (PDB file 5IZ7). The capsid includes 180 copies of protein E. The three E proteins from each asymmetric unit are colored green, orange, and blue. **(B)** The elastic network of the capsid of ZIKV, constructed from the Cα only, with a cutoff $R_C = 14$ Å. **(C)** Inside view of the elastic network, obtained by cutting the full elastic network along the plane shown as a line on **(B)**. Note that it is possible to identify rafts, as illustrated with one raft being contoured with a dashed rectangle (see text for details). All three panels were generated using Pymol (http://www.pymol.org).

where ω_k and A_k are the $k-th$ eigenvalues and eigenvectors, respectively. Note that C is a $3N \times 3N$ matrix. The summation extends from $k = 7$, the first non-zero mode, to M, the highest mode considered (up to $3N$). To obtain a scalar quantification of the correlation of the motions of two atoms i and j, a correlation matrix P is computed, following Ichiye and Karplus (1991):

$$P_{ij} = \frac{tr(C_{ij})}{\sqrt{tr(C_{ii})tr(C_{jj})}} \tag{13}$$

The values P_{ij} range from -1 to $+1$, with a negative correlation value indicating an anticorrelated motion, and a positive correlation value identifying a correlated pattern of dynamics between the two atoms considered. These values are stored into a cross-correlation matrices CCM that is used to visualize correlations of motion within the molecule under study.

4. RESULTS AND DISCUSSION

DENV and ZIKV are both members of the *flaviviridae* family. DENV serotype 1 and ZIKV (which are the focus of this study) share 53% sequence identity (Kostyuchenko et al., 2016). Their particles share a common fold, with their capsids having icosahedral symmetry. Those capsids are formed of 60 asymmetrical units, with each unit containing three copies of E protein (495 and 504 residues in DENV and ZIKV, respectively) and three copies of the membrane protein M (74 and 75 residues in DENV and ZIKV, respectively). The high resolution cryo-EM structures of all four serotypes of DENV, as well as the structure of one strain of ZIKV, are available in the Protein Data Bank (Zhang et al., 2012; Kostyuchenko et al., 2013, 2014; Fibriansah et al., 2015; Kostyuchenko et al., 2016; Sirohi et al., 2016). Here we focus on the structure of the mature form of serotype 1 of DENV, with PDB code 4CCT (Kostyuchenko et al., 2013), and of the equivalent mature form of ZIKV, as given by one of the recently published structures, with PDB code 5IZ7 (Kostyuchenko et al., 2016). Those two structures were shown to be very similar, with only small differences that will be discussed in light of their dynamics. A cartoon representation of ZIKV is given in **Figure 1A**. The DENV capsid shows the same architecture.

The PDB file for 4CCT only contains Cα atoms. For consistency, we used Cα only representations of 4CCT and 5IZ7. We isolated from those two files all the Cα atoms of the viral capsid. For both viruses, we considered E protein in four different environments: isolated, MONO, (corresponding to chain A in the asymmetric unit of 4CCT and chain B of the asymmetric unit of 5IZ7), within the asymmetric unit, UNIT, within a raft, RAFT, and within the whole capsid structure, FULL. The corresponding complexes MONO, UNIT, RAFT, and FULL contain 495, 1707, 3414, and 102420 residues for 4CCT, respectively, and 504, 1737, 3474, and 104220 residues for 5IZ7, respectively. We generated elastic networks for all those eight complexes using a cutoff procedure, with the cutoff set to 14 Å. **Figures 1B,C** illustrate the elastic network for the FULL complex for ZIKV (5IZ7). We note that this elastic network follows the surface of the capsid virus and does not include any edges that cross the interior of the capsid; this is a direct consequence of the cutoff that is used. The inside of the geometric structure formed by the elastic network reveals the presence of rafts (one such raft is shown inside a rectangle in **Figure 1C**), namely three dimers of E protein lying parallel to each other. Once the elastic networks were established, we computed the hundred lowest normal modes for each of them, using the procedure detailed in the Methods section.

We emphasize that the elastic networks for the full capsids were computed using the empty protein shells, following previous studies of viral particles using ENM and their normal modes (Tama and Brooks III, 2002, 2005; Kim et al., 2003; Chennubotla et al., 2005; Rader et al., 2005; Polles et al., 2013). This setting is expected to be satisfactory as the stability of the empty capsid is guaranteed by the geometric construction of the ENM, which makes up for the missing stabilizing interactions of the coat proteins and RNA. We note that the latter were not resolved in the cryo-EM structures we considered.

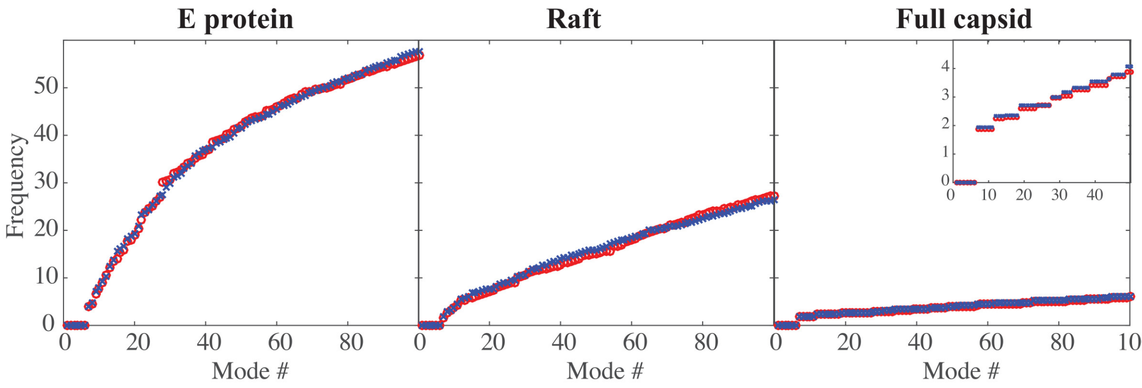


FIGURE 2 | Comparing the low frequencies of the normal modes of DENV and ZIKA. The frequencies of the first hundred normal modes of DENV (red circle, o) and ZIKV (blue cross, x) are plotted against the normal mode index (#), for the E protein by itself (left), for a raft (middle), and for the full capsid (right). The frequencies are in arbitrary units, as the force constants are also in arbitrary units. Note the decrease in the amplitude of those frequencies as the size of the complex increases. The insert in the right panel shows an enlargement for the first 50 normal modes; it highlights the degeneracy of the normal modes for a full capsid.

4.1. Characterizing the Low Frequency Normal modes of DENV and ZIKV

In **Figure 2** we compare the frequencies of the first hundred normal modes of the MONO, RAFT, and FULL complexes of DENV and ZIKV. As expected, the first six frequencies are found equal to zero, for all complexes considered, as those frequencies correspond to the rigid motions (three translations and three rotations). The larger the protein complex, the more the spectra of frequencies of the normal modes are shifted toward lower frequencies, indicating the presence of more collective motions in protein oligomers. The spectra of frequencies for the full capsids reveal the presence of degeneracy, namely repeating frequencies, that correspond to symmetries in the capsid. All the differences observed in the three complexes are conserved between DENV and ZIKV. We note also the nearly perfect match between the normal mode frequency spectra of the two viruses.

4.2. Correlated Dynamics of E Proteins in the Capsids of DENV and ZIKV

In **Figures 3**, **4** we assess the extent to which packing influences the dynamics of the E protein of DENV and ZIKV, respectively. For both viruses, the cross correlation matrices (CCM) for E protein vary significantly between the MONO, UNIT, and FULL complexes. The CCM for the E protein alone reveals significant positive correlations within each of the three domains I, II, and III. Domains II and III exhibit both positive and negative correlations in their atomic fluctuations, while the motions of domain I are only weakly correlated to the motions of domain II and III. When the dynamics of the E protein are studied in the context of the asymmetric unit, the same positive correlations are observed within each of the three domains. The interactions between the domains change significantly, however. In the UNIT complex, the dynamics of domain II are strongly anticorrelated to the dynamics of domain III, while domain

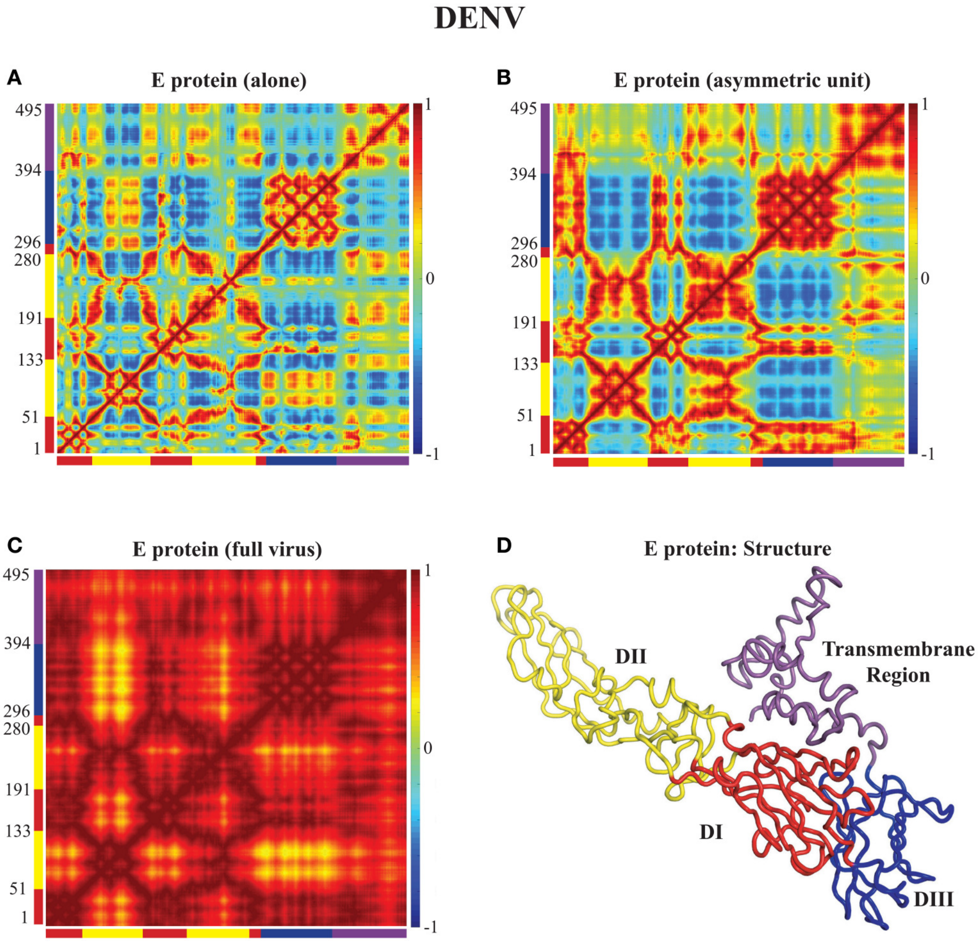


FIGURE 3 | Correlated motions in the DENV E protein. Cross Correlation Matrices (CCM) obtained from the 94 first non-zero modes for the E protein alone (MONO, **A**), the E protein in the asymmetric unit (UNIT, **B**), and the E protein in the whole capsid (FULL, **C**). Those plot show correlations between the motions of Cα atoms in each complex considered. Both axes of a matrix are the amino acid residue index. Each cell in a matrix shows the correlation between the motions of two residues (Cα atoms) in the protein on a range from −1 (anticorrelated, blue) to 1 (correlated, red), with 0 conferring no correlation. **(D)** The E protein is shown in cartoon mode. The color code for the structure in **(C)** as well as for the X and Y axes of the CCM plots in **(A)** to follows the standard designation of the E protein domains I (red), II (yellow), and III (blue). The transmembrane domain is shown in purple. Panel **(D)** was generated using Pymol.

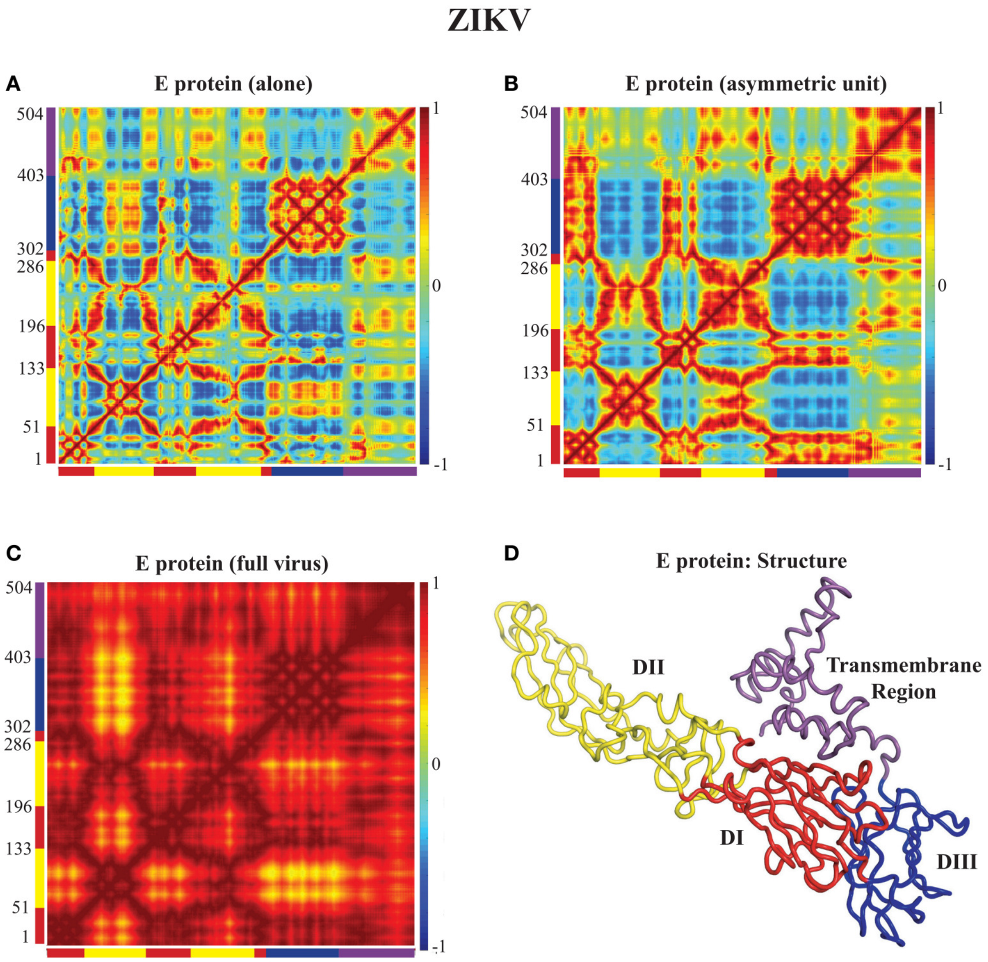


FIGURE 4 | Correlated motions in the ZIKV E protein. Cross Correlation Matrices (CCM) obtained from the 94 first non-zero modes for the E protein alone (MONO, **A**), the E protein in the asymmetric unit (UNIT, **B**), and the E protein in the whole capsid (FULL, **C**). **(D)** The E protein is shown in cartoon mode. Colors and layout follow the same schemes as in **Figure 3**.

I is correlated positively with domain III. In the full viral capsid, the internal dynamics of the E protein remain mostly as observed in the asymmetric unit. The only difference is the addition of a global positive correlation over the full protein that comes from concerted dynamics within the capsid. In all three oligomeric states, the transmembrane domain shows weak positive correlation with domain II.

All the differences in dynamics observed between isolated E proteins and E proteins in the whole capsid are conserved between DENV and ZIKV.

4.3. Correlated Dynamics of Rafts of E Proteins in the Capsids of DENV and ZIKV

Figures 3, **4** reveal the effects of packing in the viral capsid on the dynamics of one E protein. We performed a similar analysis on a larger structure of the capsid, namely a raft. A raft is formed from six E proteins forming 3 dimers arranged in a parallel manner, resulting from the combination of two asymmetrical units (see **Figure 5E**). The whole capsid contains 30 such rafts. In **Figures 5A–C**, we assess the extent to which packing influences the dynamics of such rafts for both DENV and ZIKV. In the CCM for the raft alone (**Figures 5A,B** for DENV and ZIKV, respectively) we clearly identify the six E proteins along the diagonal. Each of those E proteins exhibits dynamics correlation patterns equivalent to those observed in the E protein when it is in the asymmetrical unit. The interactions between the E proteins are consistent with the structure of the raft. The first E proteins of the two asymmetrical units, proteins E1A and E1B, show strong positively correlated dynamics. Those two proteins form a dimer in the raft. In contrast, proteins E2A and E3A in Unit A, and proteins E2B and E3B in Unit B have a pattern of interactions that include both positively correlated and negatively correlated motions, depending on their domains: for example, domains III have negative correlations between the two proteins, while domains II are positively correlated between the two proteins. The pair of proteins (E2A, E3A) shows weak correlated dynamics with the pair of proteins (E2B, E3B), with a chessboard pattern (i.e., positive correlations between E2A and E2B, and negative correlations between E3A and E3B).

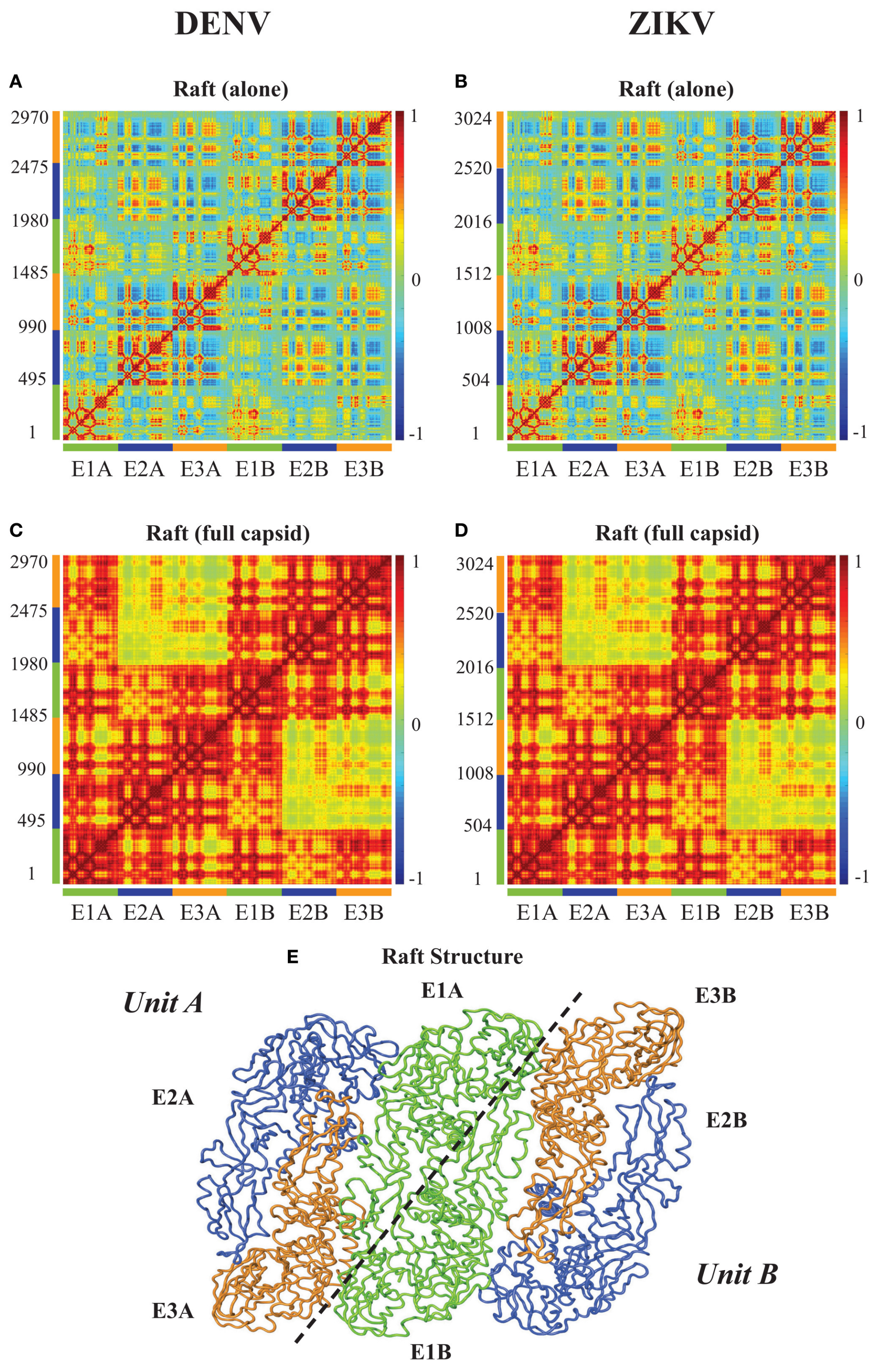


FIGURE 5 | Correlated motions in the a E protein raft. Cross Correlation Matrices (CCM) obtained from the 94 first non-zero modes for a E protein raft alone (UNIT), and a raft in the whole capsid (FULL) for DENV **(A,C)**, and for ZIKV **(B,D)**. X axes and Y axes are residue indices. The positions of the six E proteins are marked, with labels and color codes defined on the structure in **(E)**. **(E)** Cartoon model for the raft. Note that a raft includes two asymmetric units, labeled Unit A and Unit B. The first E protein of each unit, E1A and E1B form a dimer. Panel **(E)** was generated using Pymol.

The CCMs for a raft included in the whole capsid (**Figures 5C,D** for DENV and ZIKV, respectively) reveal different patterns than those described for the raft alone, highlighting again the impact of packing in the virus environment. There is a high level of positive correlation of motions within each of the units A and B. The proteins E1A and E1B that form a dimer at the center of the raft are mostly interacting with themselves in the raft alone, while they show strong levels of positive correlations with all three E proteins of the opposing unit in the raft when considered within the whole capsid. In contrast, the pairs of

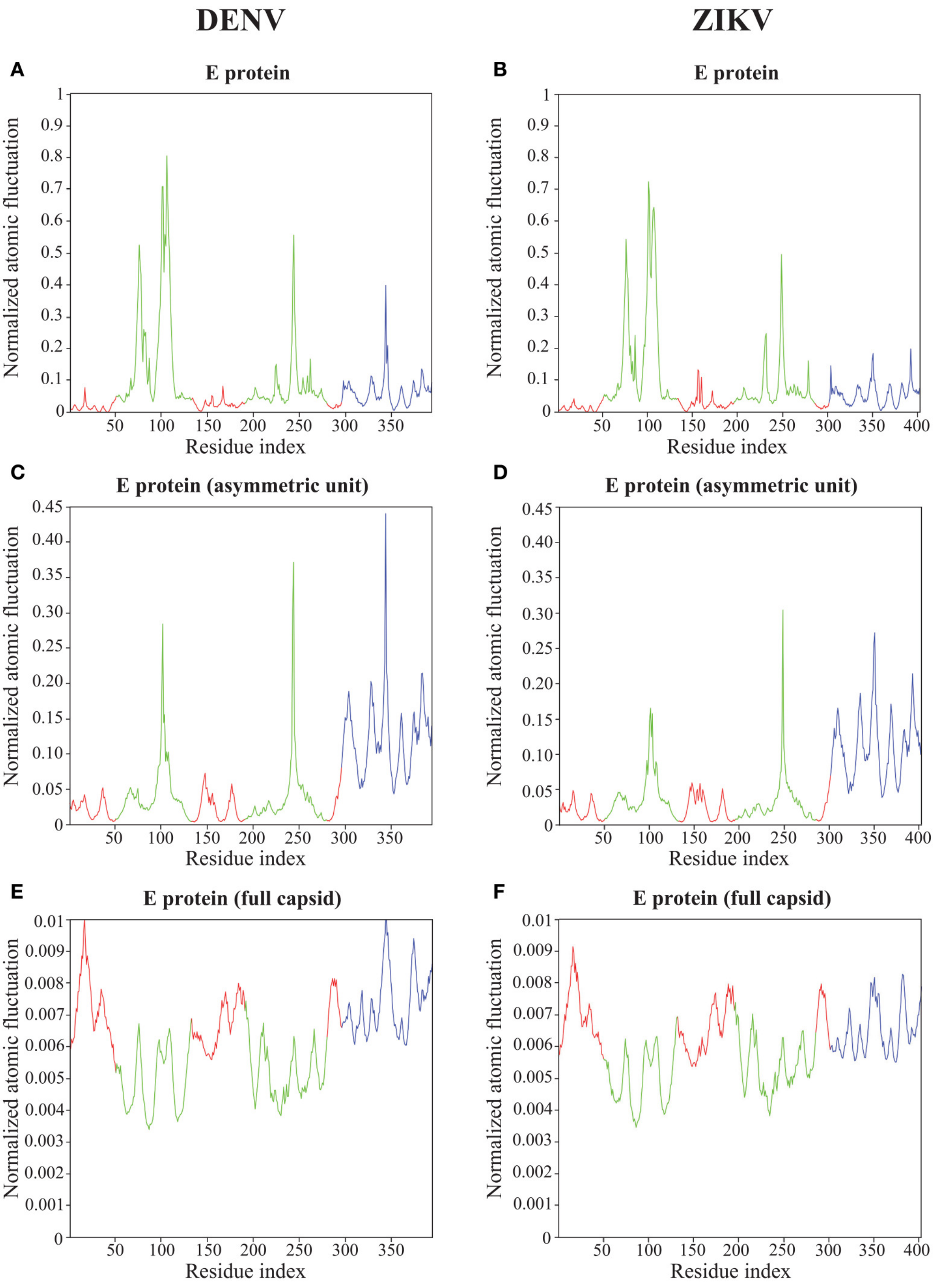


FIGURE 6 | Atomic fluctuations in the DENV and ZIKV E proteins. The atomic displacement fluctuations obtained from the 94 first non-zero modes for the E protein alone (MONO, **A,B**), the E protein in the asymmetric unit (UNIT, **C,D**), and the E protein in the whole capsid (FULL, **E,F**) are plotted as a function of the residue number for both DENV (PDB file 4CCT) and ZIKV (PDB file 5IZ7). The Y axis represents normalized displacements (see text for details). The color code follows the standard designation of the E protein domains for domains I (red) and III (blue), while domain II has been colored green to enhance visibility.

proteins (E2A, E3A) and (E2B, E3B) present significantly lower correlation when considered in the whole capsid compared to the raft alone. Such a behavior would favor concentration of concerted internal motions in a few E protein dimers at the center of the rafts in the whole viral capsid instead of a more uniform spread of concerted motions.

Similar to the findings for the dynamics of the E proteins, the differences in dynamics observed between isolated rafts and rafts in the whole capsid are conserved between DENV and ZIKV.

4.4. Atomic Fluctuations within the E Proteins of the Capsids of DENV and ZIKV

The normalized squared atomic fluctuations for each Cα atom in the E protein of DENV and ZIKV were calculated as the sum of their displacements along the first 94 non-zero modes, weighted by the reciprocal of the eigenvalues, as given by Equation (11). For both viruses, the calculation was performed in three states for the E protein, namely the MONO, UNIT, and FULL complexes described above. The absolute values of the amplitudes of the fluctuations computed using Equation (11) are somewhat arbitrary, as they depend on the parametrization of the elastic network, namely on the cutoff values R_c and the strength of the force constants k_{ij}. While it is possible to select those parameters such that a good fit is obtained between the computed fluctuations and experimental B-factors, we prefer not to, following the advice of Fuglebakk et al. (2013) that warned on possible overfitting problems. Instead, we just normalize the computed fluctuations for an atom i using:

$$< \Delta_N \mathbf{X}_i^2 > = \frac{< \Delta \mathbf{X}_i^2 > - \min(< \Delta \mathbf{X}^2 >)}{\max(< \Delta \mathbf{X}^2 >) - \min(< \Delta \mathbf{X}^2 >)} \tag{14}$$

where the min and max values are computed over all Cα atoms of the molecule considered. To enable comparison, we computed the min and max values from the fluctuations observed in the E protein alone, and applied those to normalize the fluctuations of all three states considered, i.e., MONO, UNIT, and FULL. Results for DENV and ZIKV are shown in **Figure 6**.

Not unexpectedly, the amplitude of the atomic fluctuations within the E protein decreases as the protein is more constrained, from a (normalized) range between 0 and 1 in the E protein alone (**Figures 6A,B**), to a range between 0 and 0.01 in the full capsid (**Figures 6E,F**). Of significance is the change in dynamics observed in the kl-loop between domains I and II (the DI-DII hinge, residues 280–290) between the stand alone E protein and the capsid. In the former, this loop region is predicted to be rigid, while in the latter it is found to be significantly more dynamic. This hinge is thought to be important to flip the domain DII to expose the fusion loop during the fusion event (Modis et al., 2003; Zhang et al., 2004; Kostyuchenko et al., 2016). In contrast, the HI-loop in the putative receptor binding domain DIII (residues 230–240) is found to be more dynamic in the E protein alone than in the whole capsid. DENV and ZIKV show the same dynamical behavior in both loops (the kl- and HI-loops).

The two plots showing the atomic fluctuations computed from normal modes in the E proteins are globally similar between DENV and ZIKV in all oligomeric states (**Figure 6**). There are, however, some localized differences that are worth discussing. There is a putative increase in dynamics in the region 150–160 in ZIKV compared to DENV that is most marked in the E protein monomer, but still present it its oligomeric states. This region corresponds to the Glycan loop, which contains a glycosylation site (Asn154 in ZIKV and Asn153 in DENV). It was found to be the region with the biggest structural differences (up to 6 Å) in the cryo-EM structures of ZIKV and DENV (Sirohi et al., 2016). Our calculations were performed in the absence of the sugar moities on the Asparagine. We believe however that our results highlight an intrinsic difference in the dynamics of the

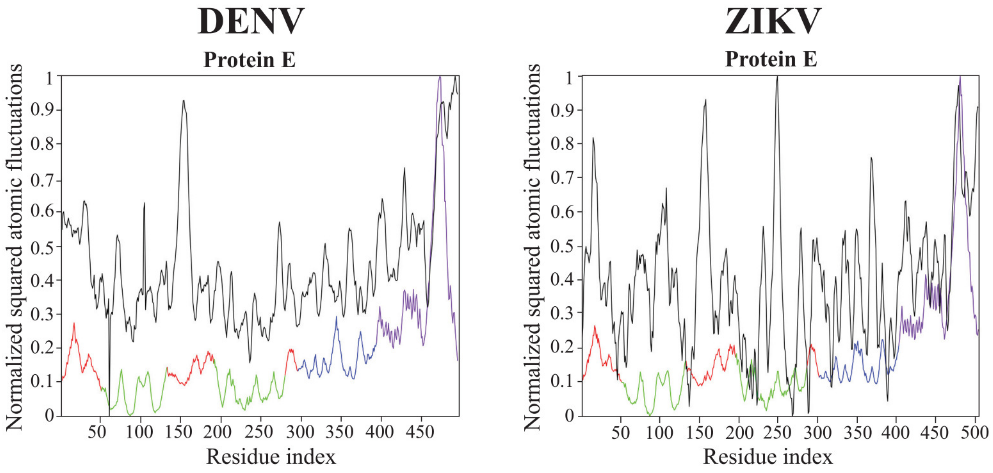


FIGURE 7 | Comparison of normalized experimental and computed atomic fluctuations in the DENV and ZIKV E proteins. The computed atomic displacement fluctuations were obtained from the 94 first non-zero modes of the whole capsid shell. The experimental fluctuations are taken from the cryo-EM structures of DENV (4CCT, Kostyuchenko et al., 2013) and ZIKV (5IZ7, Kostyuchenko et al., 2016) The color code for the computed atomic fluctuation is: E protein domain I, red, II, green, III, blue, and transmembrane domain, purple.

Glycan loops of DENV and ZIKV that is worth exploring further. In contrast to the Glycan loop, the region 340–350 is found to be less dynamic in ZIKV than in DENV in all oligomeric states of their E proteins. This region corresponds to the C strand and CD loop in domain III. Based on the differences in the structures of the DENV and ZIKV capsids, Kostyuchenko et al. (2016) hypothesized that the presence of an additional amino acid in the C strand in ZIKV was responsible for a rearrangement of the structure locally that is possibly responsible for the increased thermal stability of ZIKV. Our results indeed suggest a more rigid C strand in ZIKV compared to DENV. The exact relationship between this decrease in atomic fluctuations and thermal stability is unknown.

All results on dynamics presented above are based on atomic fluctuations and dynamic correlations computed from normal mode analyses. In **Figure 7** we compare those normalized computed atomic fluctuations for the Cα atoms of the E protein in the full capsid structure with the corresponding normalized experimental B-factors extracted from the PDB files 4CCT and 5IZ7 for DENV and ZIKV, respectively. Overall, the profiles show qualitative similarities over the full range of residues in E protein. The correlation coefficients between the experimental B-factors for DENV and ZIKV and the computed atomic fluctuations are 0.58 and 0.45, respectively. Those values are modest. We note that it would be possible to obtain significantly better correlations if the elastic constants k_{ij} assigned to the links of the networks were fitted to improve the match between B-factors and computed fluctuations. We also notice differences in relative amplitudes of the experimental and computed atomic fluctuations; these differences exist, however, between the experimental B factors for the two viruses and they could not be interpreted when analyzing the differences between the corresponding structures (Sirohi et al., 2016; Kostyuchenko et al., 2016).

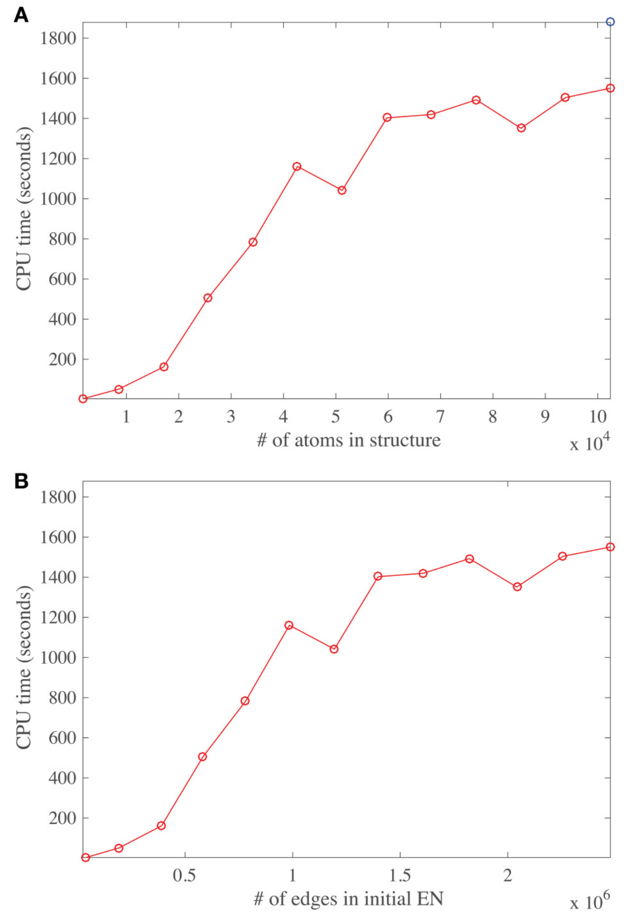


FIGURE 8 | Running time for DDNMA. The running time of the normal mode computation is plotted against the initial number of atoms **(A)**, and the initial number of edges in the corresponding elastic network, EN **(B)**. The timings are computed on a single Intel Core I7 processor running at 4.0 GHz with 8 GB of RAM.

4.5. Computing Time

The main task performed by DD-NMA when computing the normal modes of an elastic network is the diagonalization of the Hessian. For large systems, it is not feasible to perform the full diagonalization, both because of its time and memory complexities (both of order $O(N^3)$, where N is the number of atoms). Instead, only partial diagonalization is performed, with only the eigenvalues with the lowest amplitudes (usually 100) being computed. The method implemented is based on an iterative procedure. As discussed in the Material and Methods section, this procedure is efficient, of order $O(Mk + Nk^2 + k^3)$, where M is the number of non-zero elements in the sparse representation of the Hessian matrix, and k the number of eigenvalues that are computed. The first term corresponds to the matrix vector multiplications needed at each iteration, the second term relates to the Gram-Schmidt orthogonalization required to build the Krylov basis, and the last term is the cost of diagonalizing the matrix representing this basis. To test if we observe this expected behavior on real systems, we have experimented with systems of varying size. We have applied DD-NMA on parts of the capsids of DENV, with increasing number of asymmetrical units included, from one to sixty. For all systems, we extracted the Cα atoms, computed an all-atom elastic network with a cutoff of 14Å, and computed the 100 lowest frequency normal modes with DD-NMA. All those experiments were performed on a iMAC Apple computer with a 4.0 GHz Intel Core I7 processor, with 8 GB of memory. The computing times for DD-NMA are plotted against the initial numbers of atoms and edges in the all-atom elastic networks in **Figure 8**.

The number of non-zero elements in the Hessian matrix is directly proportional to the number of edges in the elastic network and implicitly proportional to the number of atoms in the protein, assuming constant density of atoms. Interestingly, the curves cpu time vs. number of atoms and vs. number of edges show three different regimes. For a relatively small number of atoms (below 20,000), and for a medium number of atoms (between 20,000 and 40,000), the cpu time is found to vary linearly, as expected, but with different slopes. The different slopes come from the relative weights of the two terms Mk and Nk^2 in the time complexity. For larger number of atoms, the behavior of the cpu time is found to be more erratic, with a slower rate of increase. We suspect that this behavior is due to cache issue. The time complexity of computing the product of the Hessian with a vector using the sparse representation of the Hessian is more complex than just being proportional to M, the number of non-zero elements of the Hessian H. Indeed, for very large matrices, it depends on their storage patterns. We have not tried to optimize this storage, which is most likely the reason for the erratic behavior. It does hint to possible improvement in the computation of the normal modes, by first re-ordering the Hessian using for example METIS (Karypis and Kumar, 1999).

We note that it takes approximately 30 min to compute the first hundred normal modes for a molecular system with hundred thousand atoms, on a single core, on a desktop computer. While this is not fast *per se*, it is still manageable. We do note that part of the codes for computing the eigenvalues of the Hessian can be parallelized; we are currently working on such an improvement.

5. SUMMARY AND CONCLUSIONS

Understanding the dynamics of viral capsids is of fundamental interest for modeling the key steps of viral life cycles. In this paper, we have described an implementation of normal mode analysis based on elastic network models that enables such analyses. This implementation is based on the known foundations in the domain (Tirion, 1996) and does not deviate significantly from other available implementations (Zheng and Doniach, 2003; Suhre and Sanejouand, 2004; Kruger et al., 2012; Tiwari et al., 2014; Eyal et al., 2015; Frappier et al., 2015). We discuss in details its parametrization, namely the choice of the coarse graining of the molecular system, the choice of the method for computing the elastic network, and the assignment of force constants to the resulting springs, and justify the choices we have implemented. We emphasize the need for efficient and robust algorithms for computing the normal modes of elastic networks, in particular when those networks include a very large number of nodes -in the hundred of thousands-, such as those derived for virus capsids. We have illustrated the application of our method to study the dynamics of the viral capsids of DENV serotype 1 and ZIKV. We have characterized the impact of the packing imposed by the capsids on their E proteins that play essential roles in receptor binding and fusion to the membrane of the host cells. We have identified differences in the atomic fluctuations of these proteins between DENV serotype 1 and ZIKV that are consistent with the structural differences observed using high resolution cryo-EM experimental structures (Kostyuchenko et al., 2016; Sirohi et al., 2016). In the future, we will consider two types of extensions of this first study that relate to the method itself as well as to its specific application to studying DENV and ZIKV.

First, we recognize that the need for a reasonable computational cost, when applying a method such as normal mode analysis to a large molecular system such as a virus capsid, implies that some sort of coarse graining is applied to such a system. Many options exist to reduce the dimensionality of the problem by selecting subsets of atoms, “beads,” to represent the system (Kmiecik et al., 2016). The positions of those beads are either defined by known atoms (usually the Cα), or by fitting to capture the dynamics of the full molecular system (Zhang et al., 2008, 2011; Li et al., 2016). The main difficulty in coarse-graining, however, is to design potential energy functions or force fields that retain the physics of the all-atom explicit solvent system in terms of structure, thermodynamics and dynamics (Riniker et al., 2012). While significant efforts have been made to guarantee that a coarse-grained model and its potential capture the complexity of the all-atom molecular system (Riniker et al., 2012; Saunders and Voth, 2013; Na et al., 2015; Zhang, 2015), we note that much less has been done to generate a true multi-scale representation of this system, i.e., to define a hierarchy of coarse-grained models with a coupling between those models. Our intention is to generate such a hierarchy; for this purpose, we will rely on the concept of renormalization group (RG) that is well known in physics (Wilson, 1975). We have implemented in DD-NMA a beta-version of such a method that performs iterative decimation of an elastic network. We will test this method on viral capsids once we have adapted the code to deal with hundreds of thousands of atoms.

Once the representation of the molecular system is chosen, the elastic network is defined as a set of links, with a link between two residues only if the distance between their Cα atoms is smaller than a given cutoff. As an alternative to this cutoff model, Xia et al. (2014) proposed to use all edges of the Delaunay triangulation of the selected atoms as an alternate elastic network. We believe that the use of Delaunay triangulation to define the ENM extends the range of applicability of NMA to the realm of less globular proteins. We will proceed in this direction and test this alternate definition of ENM to study the dynamics of viruses.

Our analyses of the dynamics of DENV and ZIKV capsids were based on naked, empty shells. There are many opportunities to extend this work. We are interested in generating plausible paths between different conformations of the virus capsids, such as the “breathing" induced by increase of temperature (Fibriansah et al., 2013), and the changes observed during the maturation of the virus. We will develop new methods to find such plausible paths in very large systems such as viral capsids, where “plausible" refers to a path with minimal frustration, also defined as the Minimum Action Path (MAP) (Olender and Elber, 1996; Eastman et al., 2001; Franklin et al., 2007; Vanden-Eijnden and Heymann, 2008; Zhou et al., 2008; Chandrasekaran et al., 2016). Finally, we plan to study the impact of glycolsylation of the E protein and/or antibody binding on the virus capsids onto their dynamical properties.

AUTHOR CONTRIBUTIONS

YH and FP performed research and analyzed data. MD and PK designed research and analyzed data. PK wrote the manuscript.

ACKNOWLEDGMENTS

PK acknowledges support from the Ministry of Education of Singapore through Grant Number: MOE2012-T3-1-008. FP acknowledges support from a Bioinfo:BipBip grant from the Agence Nationale de la Recherche (France) while he was at the Pasteur Institute.

REFERENCES

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80, 505–515. doi: 10.1016/S0006-3495(01)76033-X

Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Fold. Design* 2, 173–181. doi: 10.1016/S1359-0278(97)00024-2

Bahar, I., Cheng, M. H., Lee, J. Y., Kaya, C., and Zhang, S. (2015). Structure-encoded global motions and their role in mediating protein-substrate interactions. *Biophys. J.* 109, 1101–1109. doi: 10.1016/j.bpj.2015.06.004

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., et al. (2013). The global distribution and burden of Dengue. *Nature* 496, 504–507. doi: 10.1038/nature12060

Brady, O. J., Gething, P., Bhatt, S., Messina, J., Brownstein, J., Hoen, A., et al. (2012). Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl. Trop. Dis.* 6:e1760. doi: 10.1371/journal.pntd.0001760

Brooks, B., Bruccoleri, R., and Olafson, B. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* 4, 187–217. doi: 10.1002/jcc.540040211

Chandrasekaran, S. N., Dhas, J., Dokholyan, N. V., and Carter, C. W. Jr. (2016). A modified path algorithm rapidly generates transition states comparable to those found by other well established algorithms. *Struct. Dyn.* 3:012101. doi: 10.1063/1.4941599

Chennubotla, C., Rader, A. J., Yang, L. W., and Bahar, I. (2005). Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Phys. Biol.* 2, S173–S180. doi: 10.1088/1478-3975/2/4/s12

Delarue, M., and Sanejouand, Y. H. (2002). Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J. Mol. Biol.* 320, 1011–1024. doi: 10.1016/S0022-2836(02)00562-4

Eastman, P., Gronbech-Jensen, N., and Doniach, S. (2001). Simulation of protein folding by reaction path annealing. *J. Chem. Phys.* 114:3823. doi: 10.1063/1.1342162

Erman, B. (2006). The gaussian network model: precise prediction of residue fluctuations and application to binding problems. *Biophys. J.* 91, 3589–3599. doi: 10.1529/biophysj.106.090803

Eyal, E., Lum, G., and Bahar, I. (2015). The anisotropic network model web server at 2015 (anm 2.0). *Bioinformatics* 31, 1487–1489. doi: 10.1093/bioinformatics/btu847

Eyal, E., Yang, L. W., and Bahar, I. (2006). Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* 22, 2619–2627. doi: 10.1093/bioinformatics/btl448

Fengand, H., Costaouec, R., Darve, E., and Izaguirre, J. A. (2015). A comparison of weighted ensemble and markov state model methodologies. *J. Chem. Phys.* 142:214113. doi: 10.1063/1.4921890

Fibriansah, G., Ng, T. S., Kostyuchenko, V. A., Lee, J., Lee, S., Wang, J., et al. (2013). Structural changes in dengue virus when exposed to a temperature of 37°. *J. Virol.* 87, 7585–7592. doi: 10.1128/JVI.00757-13

Fibriansah, G., Tan, J. L., Smith, S. A., de Alwis, R., Ng, T. S., Kostyuchenko, V. A., et al. (2015). A highly potent human antibody neutralizes dengue virus serotype 3 by binding across three surface proteins. *Nat. Comm.* 6:6341. doi: 10.1038/ncomms7341

Franklin, J., Koehl, P., Doniach, S., and Delarue, M. (2007). Minactionpath: maximum likelihood trajectory for large-scale structural transitions in a coarse grained locally harmonic energy landscape. *Nucl. Acids. Res.* 35, W477–W482. doi: 10.1093/nar/gkm342

Frappier, V., Chartier, M., and Najmanovich, R. (2015). ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucl. Acids Res.* 43, W395–W400. doi: 10.1093/nar/gkv343

Fromme, P. (2015). XFELs open a new era in structural chemical biology. *Nat. Chem. Biol.* 11, 895–899. doi: 10.1038/nchembio.1968

Fuglebakk, E., Reuter, N., and Hinsen, K. (2013). Evaluation of protein elastic network models based on an analysis of collective motions. *J. Chem. Theory Comput.* 9, 5618–5628. doi: 10.1021/ct400399x

Golub, G., and van der Vorst, H. (2000). Eigenvalue computation in the 20th century. *J. Comput. Applied Math.* 123, 35–65. doi: 10.1016/S0377-0427(00)00413-1

Haliloglu, T., Bahar, I., and Erman, B. (1997). Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* 79, 3090–3093. doi: 10.1103/PhysRevLett.79.3090

Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins Struct. Func. Genet.* 33, 417–429.

Ichiye, T., and Karplus, M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins Struct. Func. Genet.* 11, 205–217. doi: 10.1002/prot.340110305

Karypis, G., and Kumar, V. (1999). A fast and highly quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.* 20, 359–392. doi: 10.1137/S1064827595287997

Kim, M. K., Jernigan, R. L., and Chirikjian, G. S. (2002). Efficient generation of feasible pathways for protein conformational transitions. *Biophys. J.* 83, 1620–1630. doi: 10.1016/S0006-3495(02)73931-3

Kim, M. K., Jernigan, R., and Chirikjian, G. (2003). An elastic network model of hk97 capsid maturation. *J. Struct. Biol.* 143, 107–117. doi: 10.1016/S1047-8477(03)00126-6

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chem. Rev.* 116, 7898–7936. doi: 10.1021/acs.chemrev.6b00163

Koehl, P. (2014). Mathematicss role in the grand challenge of deciphering the molecular basis of life. *Front. Mol. Biosci.* 1:2. doi: 10.3389/fmolb.2014.00002

Kondrashov, D. A., Cui, Q., and Phillips, G. N. Jr. (2006). Optimization and evaluation of a coarse-grained model of protein motion using X-ray crystal data. *Biophys. J.* 91, 2760–2767. doi: 10.1529/biophysj.106.085894

Kostyuchenko, V. A., Chew, P. L., Ng, T. S., and Lok, S. M. (2014). Near-atomic resolution cryo-electron microscopic structure of dengue serotype 4 virus. *J. Virol.* 88, 477–482. doi: 10.1128/JVI.02641-13

Kostyuchenko, V. A., Lim, E. X., , Zhang, S., Fibriansah, G., Ng, T. S., Ooi, J. S., et al. (2016). Structure of the thermally stable zika virus. *Nature* 533, 425–428. doi: 10.1038/nature17994

Kostyuchenko, V. A., Zhang, Q., Tan, J. L., Ng, T. S., and Lok, S. M. (2013). Immature and mature dengue serotype 1 virus structures provide insight into the maturation process. *J. Virol.*, 83:7700–7707. doi: 10.1128/JVI.00197-13

Kovacs, J., Chacon, P., and Abagyan, R. (2004). Predictions of protein flexibility: first-order measures. *Proteins* 54, 661–668. doi: 10.1002/prot.20151

Krüger, D. M., Ahmed, A., and Gohlke, H. (2012). NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. *Nucl. Acids Res.* 40, W310–W316. doi: 10.1093/nar/gks478

Kundu, S., Melton, J. S., Sorenson, D. C., and Phillips, G. N. Jr. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.* 83, 723–732. doi: 10.1016/S0006-3495(02)75203-X

Kurkcuoglu, O., Turgut, O. T., Cansu, S., Jernigan, R. L., and Doruker, P. (2009). Focused functional dynamics of supramolecules by use of a mixed-resolution elastic network model. *Biophys. J.* 97, 1178–1187. doi: 10.1016/j.bpj.2009.06.009

Lehoucq, R., Sorensen, D., and Yang, C. (1998). *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. Philadelphia, PA: SIAM. doi: 10.1137/1.9780898719628

Leioatts, N., Romo, T. D., and Grossfield, A. (2012). Elastic network models are robust to variations in formalism. *J. Chem. Theory Comput.* 8, 2424–2434. doi: 10.1021/ct3000316

Levitt, M., Sander, C., and Stern, P. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* 181, 423–447. doi: 10.1016/0022-2836(85)90230-X

Levitt, M., and Warshel, A. (1976). Computer simulation of protein folding. *Nature* 253, 694–698.

Li, M., Zhang, J. Z., and Xia, F. (2016). A new algorithm for construction of coarse-grained sites of large biomolecules. *J. Comp. Chem.* 37, 795–804. doi: 10.1038/253694a0

Lindahl, E., Azuara, C., Koehl, P., and Delarue, M. (2006). NORMAnDRef: visualization, deformation, and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucl. Acids. Res.* 34, W52–W56. doi: 10.1093/nar/gkl082

Lok, S. M., Kostyuchenko, V., Nybakken, G. E., Holdaway, H. A., Sukupolvi-Petty, A. B. S., Sedlak, D., et al. (2008). Binding of a neutralizing antibody to dengue virus alters the arrangement of surface glycoproteins. *Nat. Struct. Mol. Biol.* 15, 312–317. doi: 10.1038/nsmb.1382

López-Blanco, J. R., and Chacón, P. (2016). New generation of elastic network models. *Curr. Opin. Struct. Biol.* 37, 46–53. doi: 10.1016/j.sbi.2015.11.013

Mahajan, S., and Sanejouand, Y.-H. (2015). On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Arch. Biochem. Biophys.* 567, 59–65. doi: 10.1016/j.abb.2014.12.020

Maragakis, P., and Karplus, M. (2005). Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.* 352, 807–822. doi: 10.1016/j.jmb.2005.07.031

Meng, F., Badierah, R. A., Almehdar, H. A., Redwan, E. M., Kurgan, L., and Uversky, V. N. (2015). Unstructural biology of the dengue virus proteins. *FEBS J.* 282, 3368–3394. doi: 10.1111/febs.13349

Ming, D., and Wall, M. (2005). Allostery in a coarse-grained model of protein dynamics. *Phys. Rev. Lett.* 95:198103. doi: 10.1103/PhysRevLett.95.198103

Modis, Y., Ogata, S., Clements, D., and Harrison, S. (2003). A ligand-binding pocket in the dengue virus envelope glycoprotein. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6986–6991. doi: 10.1073/pnas.0832193100

Morens, D. M., and Fauci, A. S. (2008). Dengue and hemorrhagic fever. A potential threat to public health in the United States. *JAMA* 299, 214–216. doi: 10.1001/jama.2007.31-a

Na, H., Jernigan, R. L., and Song, G. (2015). Bridging between nma and elastic network models: preserving all-atom accuracy in coarse-grained models. *PLoS Comput. Biol.* 11:e1004542. doi: 10.1371/journal.pcbi.1004542

Noguti, T., and Go, N. (1982). Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature* 296, 776–778. doi: 10.1038/296776a0

Olender, R., and Elber, R. (1996). Calculation of classical trajectories with a very large time step: formalism and numerical examples. *J. Chem. Phys.* 105, 9299–9315. doi: 10.1063/1.472727

Orellana, L., Rueda, M., Ferrer-Costa, C., Lopez-Blanco, J. R., Chacon, P., and Orozco, M. (2010). Approaching elastic network models to molecular dynamics flexibility. *J. Chem. Theory Comput.* 6, 2910–2923. doi: 10.1021/ct100208e

Peeters, K., and Taormina, A. (2009). Group theory of icosahedral virus capsid vibrations: a top-down approach. *J. Theor. Biol.* 256, 607–624. doi: 10.1016/j.jtbi.2008.10.019

Perera, R., and Kuhn, R. (2008). Structural proteomics of dengue virus. *Curr. Opin. Microbiol.* 11, 369–377. doi: 10.1016/j.mib.2008.06.004

Petrone, P., and Pande, V. (2006). Can conformational change be described by only a few normal modes? *Biophys. J.* 90, 1583–1593. doi: 10.1529/biophysj.105.070045

Polles, G., Indelicato, G., Potestio, R., Cermelli, P., Twarock, R., and Micheletti, C. (2013). Mechanical and assembly units of viral capsids identified via quasi-rigid domain decomposition. *PLoS Comput. Biol.* 9:e1003331. doi: 10.1371/journal.pcbi.1003331

Rader, A. J., Vlad, D. H., and Bahar, I. (2005). Maturation dynamics of bacteriophage HK97 capsid. *Structure* 13, 413–421. doi: 10.1016/j.str.2004.12.015

Riniker, S., Allison, J. R., and van Gunsteren, W. F. (2012). On developing coarse-grained models for biomolecular simulation: a review. *Phys. Chem. Chem. Phys.* 14, 12423–12430. doi: 10.1039/c2cp40934h

Rueda, M., Chacon, P., and Orozco, M. (2007). Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure* 15, 565–575. doi: 10.1016/j.str.2007.03.013

Saunders, M. G., and Voth, G. A. (2013). Coarse-graining methods for computational biology. *Annu. Rev. Biophysics* 42, 73–93. doi: 10.1146/annurev-biophys-083012-130348

Simonson, T., and Perahia, D. (1992). Normal modes of symmetric protein assemblies. application to the tobacco mosaic virus protein disk. *Biophys. J.* 61, 410–427. doi: 10.1016/S0006-3495(92)81847-7

Sirohi, D., Chen, Z., Sun, L., Klose, T., Pierson, T., Rossmann, M., et al. (2016). The 3.8 å resolution cryo-em structure of zika virus. *Science* 352, 467–470. doi: 10.1126/science.aaf5316

Suhre, K., and Sanejouand, Y.-H. (2004). Elnémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucl. Acids Res.* 32, W610–W614. doi: 10.1093/nar/gkh368

Tama, F., and Brooks III, C. L. (2002). The mechanism and pathway of ph induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.* 318, 733–747. doi: 10.1016/S0022-2836(02)00135-3

Tama, F., and Brooks III, C. L. (2005). Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. *J. Mol. Biol.* 345, 299–314. doi: 10.1016/j.jmb.2004.10.054

Tama, F., Gadea, F. X., Marques, O., and Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* 41, 1–7. doi: 10.1002/1097-0134(20001001)41:1<1::AID-PROT10>3.0.CO;2-P

Tama, F., and Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng.* 14, 1–6. doi: 10.1093/protein/14.1.1

Teoh, E. P., Kukkaro, P., Teo, E. W., Lim, A. P., Tan, T. T., Yip, A., et al. (2012). The structural basis for serotype-specific neutralization of dengue virus by a human antibody. *Sci. Transl. Med.* 4:139ra83. doi: 10.1126/scitranslmed.3003888

Tirion, M. (1996). Large amplitude elastic motions in proteins from a single parameter, atomic analysis. *Phys. Rev. Lett.* 77, 1905–1908. doi: 10.1103/PhysRevLett.77.1905

Tiwari, S. P., Fuglebakk, E., Hollup, S. M., Skjaerven, L., Cragnolini, T., Grindhaug, S., et al. (2014). WEBnm@ v2.0: web server and services for comparing protein flexibility. *BMC Bioinformatics* 15:427. doi: 10.1186/s12859-014-0427-6

Tozzini, V. (2005). Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* 15, 144–150. doi: 10.1016/j.sbi.2005.02.005

van Vlijmen, H., and Karplus, M. (2005). Normal mode calculations of icosahedral viruses with full dihedral flexibility by use of molecular symmetry. *J. Mol. Biol.* 350, 528–542. doi: 10.1016/j.jmb.2005.03.028

Vanden-Eijnden, E., and Heymann, M. (2008). The geometric minimum action method for computing minimum energy paths. *J. Chem. Phys.* 128:061103. doi: 10.1063/1.2833040

WHO (2016). *Zika Strategic Response Framework and Joint Operations Plan (January-June)*. Geneva: WHO.

Wilson, K. (1975). The renormalization group: critical phenomena and the kondo problem. *Rev. Mod. Phys.* 47, 773–840. doi: 10.1103/RevModPhys.47.773

Xia, F., Tong, D., and Lu, L. (2013). Robust heterogeneous anisotropic elastic network model precisely reproduces the experimental b-factors of biomolecules. *J. Chem. Theory Comput.* 13, 3704–3714. doi: 10.1021/ct4002575

Xia, F., Tong, D., Yang, L., Wang, D., Hoi, S. C., Koehl, P., et al. (2014). Identifying essential pairwise interactions in elastic network model using the alpha shape theory. *J. Comp. Chem.* 35, 1111–1121. doi: 10.1002/jcc.23587

Yang, L., Song, G., and Jernigan, R. (2009). Protein elastic nmodels and the ranges of cooperativity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12347–12352. doi: 10.1073/pnas.0902159106

Zhang, X., Ge, P., Yu, X., Brannan, J., Bi, G., Zhang, Q., et al. (2012). Cryo-EM structure of the mature dengue virus at 3.5 å resolution. *Nat. Struct. Mol. Biol.* 20, 105–110. doi: 10.1038/nsmb.2463

Zhang, Y., Zhang, W., Ogata, S., Clements, D., Strauss, J. H., Baker, T. S., et al. (2004). Conformational changes of the flavivirus e glycoprotein. *Structure* 12, 1607–1618. doi: 10.1016/j.str.2004.06.019

Zhang, Z. (2015). Systematic methods for defining coarse-grained maps in large biomolecules. *Adv. Exp. Med. Biol.* 827, 33–48. doi: 10.1007/978-94-017-9245-5_4

Zhang, Z., Lu, L., Noid, W. G., Krishna, V., Pfaendtner, J., and Voth, G. A. (2008). A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys. J.* 95, 5073–5083. doi: 10.1529/biophysj.108.139626

Zhang, Z., Sanbonmatsu, K. Y., and Voth, G. A. (2011). Key intermolecular interactions in the e. coli 70s ribosome revealed by coarse-grained analysis. *J. Am. Chem. Soc.* 133, 16828–16838. doi: 10.1021/ja2028487

Zheng, W., and Doniach, S. (2003). A comparative study of motor-protein motions by using a simple elastic-network model. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13253–13258. doi: 10.1073/pnas.2235686100

Zhou, X., Ren W, E. W. (2008). Adaptive minimum action method for the study of rare events. *J. Chem. Phys.* 128:104111. doi: 10.1063/1.2830717

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 31 January 2017
doi: 10.3389/fmolb.2017.00003

On the Helix Propensity in Generalized Born Solvent Descriptions of Modeling the Dark Proteome

Mark A. Olson *

Department of Cell Biology and Biochemistry, Molecular and Translational Sciences Division, United States Army Medical Research Institute for Infectious Diseases, Fredrick, MD, USA

Intrinsically disordered proteins that populate the so-called "Dark Proteome" offer challenging benchmarks of atomistic simulation methods to accurately model conformational transitions on a multidimensional energy landscape. This work explores the application of parallel tempering with implicit solvent models as a computational framework to capture the conformational ensemble of an intrinsically disordered peptide derived from the Ebola virus protein VP35. A recent X-ray crystallographic study reported a protein-peptide interface where the VP35 peptide underwent a folding transition from a disordered form to a helix-β-turn-helix topological fold upon molecular association with the Ebola protein NP. An assessment is provided of the accuracy of two generalized Born solvent models (GBMV2 and GBSW2) using the CHARMM force field and applied with temperature-based replica exchange dynamics to calculate the disorder propensity of the peptide and its probability density of states in a continuum solvent. A further comparison is presented of applying an explicit/implicit solvent hybrid replica exchange simulation of the peptide to determine the effect of modeling water interactions at the all-atom resolution.

Keywords: molecular dynamics, free-energy landscape, intrinsically disordered proteins, explicit/implicit solvent model replica-exchange simulation

OPEN ACCESS

Edited by:
Isabelle Callebaut,
French National Centre for Scientific Research, France

Reviewed by:
Evgeni B. Starikov,
Independent Researcher all over the World, Germany
Kang K. L. Liu,
Brandeis University, USA

****Correspondence:***
Mark A. Olson
molson@compbiophys.org

Specialty section:
This article was submitted to Biophysics,
a section of the journal
Frontiers in Molecular Biosciences

Received: *29 November 2016*
Accepted: *12 January 2017*
Published: *31 January 2017*

Citation:
Olson MA (2017) On the Helix Propensity in Generalized Born Solvent Descriptions of Modeling the Dark Proteome.
Front. Mol. Biosci. 4:3.
doi: 10.3389/fmolb.2017.00003

INTRODUCTION

The large conformational heterogeneity and rapid dynamic transitions of intrinsically disordered peptides and proteins (IDPs) present a challenge to experimental boundaries in characterizing their functional form on rugged energy landscapes (Wright and Dyson, 1999, 2005). From a biological perspective, the broad interest in IDPs draws principally from their fundamental role in the regulation and function of cellular protein networks. Recent experimental studies have begun to unravel the interplay between "ordered chaos" of IDPs and their kinetic transition to a topological funnel of folded states (Arai et al., 2015). The contemporary view of this dynamic process is one that occurs by either an "induced-fit" of the IDP upon molecular association with a protein target or by target "fly casting" of a prefolded state in the disordered conformational ensemble of the IDP (see, e.g., Shoemaker et al., 2000; Arai et al., 2015).

Complementary to experimental studies are computer simulations which offer a powerful set of tools to understand IDPs at the all-atom level and their inherent plasticity to navigate a disordered network of microstates (see, e.g., Zhang and Chen, 2014; Chebaro et al., 2015; Bhowmick et al., 2016; Lee and Chen, 2016). Among the simulation methods, the generalized ensemble sampling technique of temperature-based replica exchange (T-ReX; Sugitaa and Okamoto, 1999; Ishikawa et al., 2001), also known as parallel tempering, has become an increasingly popular approach for exploring the energy landscape of proteins. Algorithms combined with T-ReX to generate protein configurations vary in their theoretical formulations and range from canonical molecular dynamics (MD) simulations to nontraditional methods that accelerate conformational sampling. Of the latter, examples includes coarse replica-exchange molecular dynamics (Peter et al., 2016), accelerated molecular dynamics (see, e.g., Miao et al., 2015), Hamiltonian switch Metropolis Monte Carlo (Mittal et al., 2014), all-atom multicanonical molecular dynamics (Higo et al., 2011) and self-guided Langevin dynamics (SGLD; Wu and Brooks, 2003), among others.

A computational strategy of reducing the complexity of all-atom simulations of proteins is the replacement of explicit water interactions with a continuum description of treating implicitly the bulk physical properties of solvation effects. The most common implicit solvent method for protein dynamics simulations is the generalized Born (GB) approximation. GB models are computationally faster than explicit solvent calculations and differ in their accuracy of reproducing Poisson-Boltzmann solvation energies for single protein conformations (see, e.g., Feig et al., 2004b). Application of GB solvent models to studies of IDPs has been reported by several laboratories (see, e.g., Ganguly and Chen, 2009; Click et al., 2010; Chebaro et al., 2015; Ganguly and Chen, 2015). To date the simulation results lack consensus on the accuracy of GB solvent models as a computational framework to capture the fold propensities of IDPs and their probability density of states on a conformational landscape. Particularly missing among the reported studies are comparative assessments of GB models of IDPs with those modeled by explicit all-atom solvent replica exchange simulations.

Given the current interests in characterizing the "Dark Proteome" which consists of "invisible" conformational states within the human, viral and microbial protein fold universe (Perdigão et al., 2015; Bhowmick et al., 2016), this work presents temperature-based replica exchange simulations of modeling an IDP derived from an Ebola virus protein. Ebola viruses are nonsegmented negative sense RNA viruses that cause severe hemorrhagic fever (Sanchez et al., 2006). An X-ray crystallographic structure was reported by Amarasinghe and coworkers (Leung et al., 2015) of the Ebola nucleoprotein NP in complex with a 28-residue peptide extracted from Ebola VP35 (peptide designated as NPBP). The NP-VP35 viral assembly is essential for virus replication and offers a protein target for therapeutic development. Experimental data reveals the NPBP peptide binds NP with high affinity and specificity, and acts by blocking NP oligomerization. The peptide undergoes a folding transition from a disordered form free in solution to a helix-β-turn-helix fold upon molecular association with NP (Leung et al., 2015).

Two different generalized ensemble sampling methods are applied based on combining T-ReX with the SGLD simulation method (Lee and Olson, 2010) and two different GB solvent models are examined to assess their accuracy in modeling the probability density of states of the NPBP peptide. One of the sampling methods is the conventional application of T-ReX with a static set of temperatures to explore the conformational landscape. The other technique is an adaptive T-ReX where the replica clients dynamically walk in temperature space in search of the optimal population density on a modeled energy function (Katzgraber et al., 2006; Trebst et al., 2006; Lee and Olson, 2011; Olson and Lee, 2014; Olson et al., 2016). The GB models analyzed are GBMV2 (generalized Born molecular volume; Lee et al., 2002, 2003) and the GBSW2 (generalized Born smoothing window; Im et al., 2003). The models differ in their dielectric-boundary descriptions with one of them constructed from an analytical formulation of the molecular volume (Lee et al., 2003).

The final simulation model applied to the NPBP peptide is an explicit/implicit solvent hybrid T-ReX/MD method (Chaudhury et al., 2012). The application of this simulation model is to investigate the effect of solvent resolution on the helix propensity and the search of conformational transitions. The idea behind the hybrid model is reducing the number of replica clients needed in explicit solvent simulations by replacing the contribution of explicit solvent energies in the Metropolis exchanges (Metropolis et al., 1953) with those of the GBMV2 solvent approximation. The hybrid model allows the same number of replica clients to be applied as in the GB solvent T-ReX/SGLD simulations of the NPBP peptide while retaining a higher resolution in conformational sampling on an explicit solvent landscape (Chaudhury et al., 2012; Olson and Lee, 2013).

COMPUTATIONAL METHODS

This section provides a brief outline of the computational methods applied in this work of modeling the NPBP peptide taken from the PDB 4YPI (**Figure 1**). Summarized are the sampling techniques and protocols as well as metrics to evaluate the simulation trajectories.

Replica Exchange Schemes

A general approach for conformational sampling is the application of T-ReX (see, e.g., Ishikawa et al., 2001). Unlike the well-established method of MD simulations at a single sampling temperature, T-ReX is a generalized ensemble method of applying multiple parallel simulations in which each replica is executed at a different temperature. In traditional applications of T-ReX, the temperatures T_1, T_2, ..., T_n, where n is the number of replica clients, are predetermined by a static (fixed) set of values that span a desired range. It is common to model the set of temperatures by a geometrically spaced sequence (Predescu et al., 2004) using $n-1$ intervals from the minimum temperature denoted as $T_1 = T_{\min}$ to the maximum $T_n = T_{\max}$

$$T_{i+1} = T_i(T_{\max}/T_{\min})^{\left[\frac{1}{n-1}\right]}, \tag{1}$$

where T_i is the temperature of the ith replica client illustrated in **Figure 1**.

An alternative to Equation (1) is an adaptive replica exchange method of allowing the clients to dynamically walk in

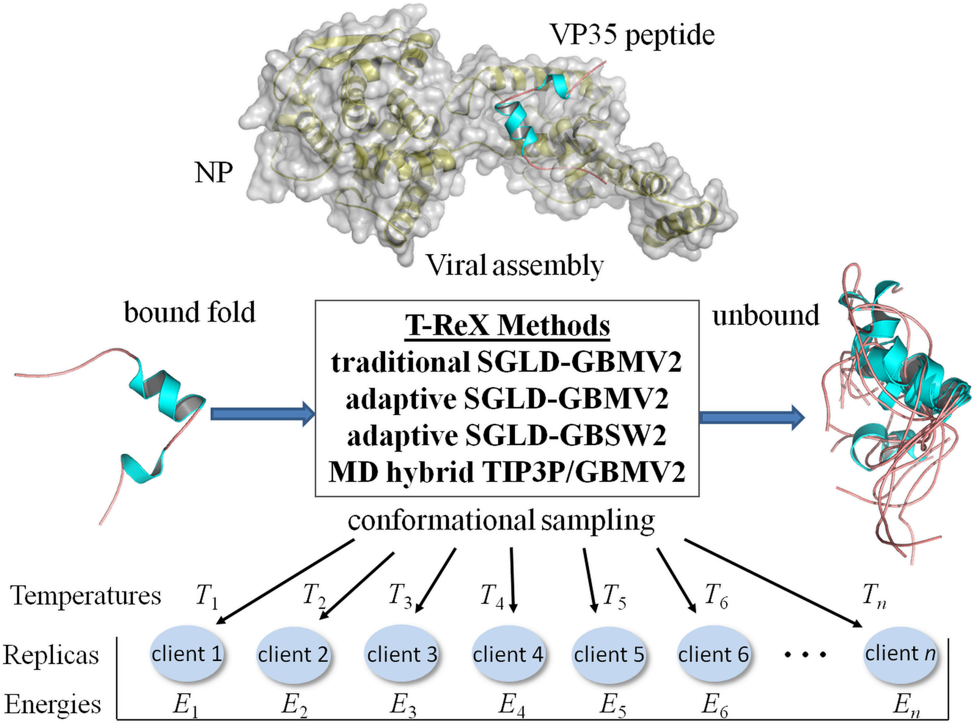

VP35 peptide
NP
Viral assembly
bound fold
unbound
T-ReX Methods
traditional SGLD-GBMV2
adaptive SGLD-GBMV2
adaptive SGLD-GBSW2
MD hybrid TIP3P/GBMV2
conformational sampling
Temperatures T_1 T_2 T_3 T_4 T_5 T_6 T_n
Replicas client 1 client 2 client 3 client 4 client 5 client 6 • • • client n
Energies E_1 E_2 E_3 E_4 E_5 E_6 E_n

FIGURE 1 | Computational strategies of modeling the Ebola virus VP35 peptide (PDB: 4YPI) in its unbound form using temperature-based replica exchange (T-ReX) simulation methods. The methods include: (1) GBMV2 solvent model applied with a traditional (static) set of temperatures spanning a range from a minimum temperature (T_1) to the upper extreme (T_n), where n is the number thermal windows (ensemble computing clients); (2) GBMV2 using an adaptive (dynamically walking) set of temperatures between T_1 and T_n; (3) GBSW2 solvent model applied by adaptive sampling; and (4) TIP3P/GBMV2 hybrid replica exchange method. Energies (E_i) used in the replica exchanges are described in the text. Molecular figures were drawn with PyMOL (www.pymol.org).

temperature space (Katzgraber et al., 2006; Trebst et al., 2006; Lee and Olson, 2011; Olson and Lee, 2014; Olson et al., 2016). In implementing the adaptive algorithm, each client is tagged as either "cold" or "hot" depending on the last temperature extreme it visited (Lee and Olson, 2011). Tracing of the clients is made by constructing histograms over temperature space, $n_{\text{cold}}(T)$ and $n_{\text{hot}}(T)$, where each bin accumulates the number of cold and hot clients visiting each temperature window. The fraction cold, $f_{\text{cold}}(T)$, of a client window at temperature T is the number of cold clients visiting that temperature divided by the total number of cold and hot client visits:

$$f_{\text{cold}}(T) = \frac{n_{\text{cold}}(T)}{n_{\text{cold}}(T) + n_{\text{hot}}(T)}. \tag{2}$$

Using the $f_{\text{cold}}(T)$ term, a thermal current is defined (Lee and Olson, 2011)

$$j = D(T)\,\eta(T)\,\frac{df_{\text{cold}}(T)}{dT}, \tag{3}$$

where $D(T)$ is the diffusivity and $\eta(T)$ is the probability that any client will reside at temperature T. The current j can be maximized by adjusting the temperatures such that $f_{\text{cold}}(T_i)$ increases linearly as a function of temperature index, i. Here in this work, a continuous function is constructed from the computed values of $f_{\text{cold}}(T_i)$ at the current set of temperatures, T_i, and new temperatures are searched for where $f_{\text{cold}}(T_i) = i/(N{-}1)$. To prevent all of the windows from clustering around the same temperature and depleting exchanges at the extremes, a constraint is applied where no neighboring temperatures can be more than two geometric spacing units apart,

$$\frac{T_{i+1}}{T} \leq \left(\frac{T_{\max}}{T_{\min}}\right)^{\left[\frac{2}{N-1}\right]} \tag{4}$$

with the lower and upper values of T_i set to $T_{\min}$ and $T_{\max}$, respectively.

The exchange of temperatures between neighboring replica clients, a and b, is determined by the Metropolis energy criteria (Metropolis et al., 1953)

$$p\left(a \leftrightarrow b\right) = \min\left[1, e^{(\beta_a - \beta_b)(E_b - E_a)}\right], \tag{5}$$

where $\beta_{\text{a}} = 1/k_B T_a$, k_B is Boltzmann's constant, T_a is the temperature of replica client a, and E_a is the potential energy of client a.

SGLD Simulation Models

For generating trajectories of the NPBP peptide, two methods were combined with T-ReX. The first is based on the SGLD simulation method developed by Wu and Brooks (2003). The SGLD equation of motion is given by

$$\dot{\mathbf{p}}_i = \mathbf{f}_i - \gamma_i \mathbf{p}_i + \mathbf{R}_i + \lambda \mathbf{g}_i, \tag{6}$$

where $\dot{\mathbf{p}}_i$ defines the rate of change of the momentum of particle i, $\mathbf{f}_i$ is the force acting on the particle, γ_i is the friction constant, $\mathbf{R}_i$ defines the random force and $\mathbf{g}_i$ is a memory function, which is scaled by an *ad hoc* guiding factor λ. The memory function $\mathbf{g}_i$ is defined by the moving average of momentum over an interval of time, L:

$$\mathbf{g}_i = \langle \mathbf{p}_i \rangle_L, \tag{7}$$

where $\langle \ldots \rangle_L$ denotes a local average. The time interval is further defined as $L = t_L/\delta t$, where t_L is the local averaging time and δt the time step along the simulation trajectory. It should be noted that because of the *ad hoc* force in Equation (6), the sampling algorithm deviates from a canonical ensemble (Lee and Olson, 2010; Wu and Brooks, 2011; Wu et al., 2012, 2016). For this work, the deviation is anticipated to be small for modeling a mini-protein (Lee and Olson, 2010), nevertheless the population distributions can be reweighted to remove the applied bias (Wu and Brooks, 2011).

In the SGLD simulations, solvent was represented by either the implicit solvent model GBMV2 (Lee et al., 2002, 2003) or GBSW2 (Im et al., 2003). The most noted difference between the two models is representation of the solvent excluded volume and the treatment of the dielectric interface. The GBMV2 parameters were selected to smooth the molecular volume by setting $\beta_s = -12$ and P3 = 0.65 (Yeh et al., 2008). The hydrophobic cavitation term was modeled by applying a phenomenological surface tension coefficient set to a value of 0.015 kcal/mol/Å^2. For applying GBSW2, the model was parameterized to fit the Lee-Richards molecular-surface Poisson results and required $w = 0.2$ Å, $a_0 = 1.2045$, and $a_1 = 0.1866$. The hydrophobic cavitation-energy tension term was set to 0.030 kcal/(molÅ^2).

The utilities and programming libraries of the Multiscale Modeling Tools for Structural Biology (MMTSB; Feig et al., 2004a) were used to carry out the T-ReX/SGLD simulations. The CHARMM simulation program (version c35b2) was applied as a modeling platform (Brooks et al., 2009). Simulations were carried out using 24 replica clients and the frequency of exchanges was set to every 1 ps of simulation. Temperatures were set at $T_{\min} =$ 300 K and $T_{\max} = 475$ K. Because the implicit solvent models GBMV2 and GBSW2 were originally developed for and have been extensively benchmarked with the CHARMM22 force field, this force field was applied with the CMAP backbone dihedral cross-term extension (Mackerell et al., 2004). An integration time step of 2 fs was used and parameters for SGLD consisted of the friction constant set to γ of 1 ps^{-1} for all heavy atoms, the guiding factor λ to a value of 1, and the averaging time t_L was set to 1 ps. These values were taken from previous studies of the SGLD model (Lee and Olson, 2010, 2011; Olson and Lee, 2014). Non-bonded interaction cutoff parameters for electrostatics and vdW terms were set at a radius of 22 Å with a 2-Å potential switching function. Covalent bonds between the heavy atoms and hydrogen atoms were constrained by the SHAKE algorithm (Ryckaert et al., 1977). The NPBP peptide was modeled for 200 ns of simulation time per thermal window, generating an ensemble of 4.8 μs.

Hybrid Simulation Model

The alternative method applied for generating trajectories of the NPBP peptide is an explicit/implicit solvent hybrid T-ReX/MD simulation (Chaudhury et al., 2012). In a typical explicit solvent T-ReX simulation the energies are given by

$$E_{\text{explicit}} = U^{\text{prot}}_{\text{all-atom}} + U^{\text{prot-solv}}_{\text{all-atom}} + U^{\text{solv-solv}}_{\text{all-atom}}, \tag{8}$$

where the first term describes the protein potential energy for a CHARMM-based molecular mechanics force field, the second term is the explicit protein-solvent interactions followed by the explicit solvent-solvent interactions. The all-atom solvent-solvent energy term requires significant number of replica-exchange clients to achieve adequate Metropolis updates (Chaudhury et al., 2012). In the hybrid T-ReX method, the dynamics of each replica moves on an explicit solvent landscape. During a Metropolis update, all waters are removed from a replica and the solvent energy term of the replica is calculated using the grid-based GBMV2 solvent model

$$E_{\text{implicit}} = U^{\text{prot}}_{\text{all-atom}} + \Delta G^{\text{prot-solv}}_{\text{GBMV2}}, \tag{9}$$

where $\Delta G^{\text{prot-solv}}_{\text{GBMV2}}$ is the free-energy term due to the implicit solvent contribution. After completion of the Metropolis exchanges, the explicit waters in each replica are replaced to their configurations prior to removal and the simulation continues according to Equation (8).

The NAMD code (Phillips et al., 2005) was applied for the 200-ns T-ReX/MD simulation with the CHARMM22+CMAP force field. The simulation cubic box size was set to 53.19 Å^3 and the number of waters was 4796. For modeling the waters the TIP3P potential was applied (Jorgensen et al., 1983). Nose'-Hoover thermostat was applied with a temperature coupling constant of 50 kcal/s^2. Given that the computational expense of the hybrid model relative to implicit solvent calculations is greater, the NAMD simulation parameters differ slightly from the T-ReX/SGLD simulations in that a smaller cutoff distance of 12 Å was applied with a switching distance of 8 Å. The integration time step remained identical to that used with the SGLD simulations and the SHAKE algorithm was similarly applied. Particle mesh Ewald was applied and combined with periodic boundary conditions.

Evaluation Metrics

To examine the trajectories generated by the simulations, the weighted histogram analysis method (WHAM; Ferrenberg and Swendsen, 1989; Kumar et al., 1992; Gallicchio et al., 2005) was applied to the data sets. The 2D density of states, Ω (q_1, q_2), for a molecular system, where q_1 and q_2 are a set of reaction coordinates of interest, is given by

$$\Omega\left(q_1,q_2\right)=\frac{\sum_{i=1}^{R}N_i\left(q_1,q_2\right)}{\sum_{j=1}^{R}n_i\exp\left(f_i-\beta_iE\right)},\tag{10}$$

where n_j is the number of data points in the jth simulation and β_j and T_j are Boltzmann's constant and temperature of the jth simulation, respectively. The function $N_i(q_1, q_2)$ is the histogram of (q_1, q_2) calculated from the ith simulation, and f_j is the scaled free energy obtained by solving the following equations self-consistently,

$$P_\beta\left(q_1,q_2\right)=\frac{\sum_{i=1}^{R}N_i\left(q_1,q_2\right)\exp\left(-\beta E\right)}{\sum_{j=1}^{R}n_i\exp\left(f_i-\beta_iE\right)}\tag{11}$$

and

$$\exp\left(-f_i\right)=\sum_{q_1,q_2}\Omega\left(q_1,q_2\right)\exp\left(-\beta E\right),\tag{12}$$

where $P_\beta(q_1, q_2)$ is the probability density at the inverse temperature β. From a density profile, a potential of mean force is determined from the relationship $W_T\left(q_1,q_2\right) = -\mathrm{R}T\log P_\beta\left(q_1,q_2\right)$, where R is the universal gas constant. For calculations presented here, q_1 = fractional helicity (f_H) of the peptide determined from DSSP (Kabsch and Sander, 1983) and q_2 = radius of gyration (R_g).

The trajectories were further analyzed by a Q score for the peptide. Q is the number of side-chain contacts in a generated conformation divided by the total number equivalent contacts in the X-ray crystal structure of NPBP. Values were computed for side-chain center-of-mass pairs (i,j), such that $j > i$ and whose distances are less than a cutoff of 4.2 Å. A sigmoidal function was applied (implemented in MMTSB) to effectively include residue pairs that are slightly further apart with a reduced weight. In addition to a Q score, pairwise Cα root-mean-square-deviation (RMSD) from the starting X-ray structure was computed for each peptide conformation in a generated ensemble of structures.

RESULTS AND DISCUSSION

Bound and Free NPBP

Figure 2 illustrates the X-ray crystallographic structure of the NPBP peptide extracted from the Ebola virus VP35 in association with the Ebola NP protein (Leung et al., 2015). The binding of NPBP occupies a functionally critical site on NP required for RNA synthesis. The peptide conformation is stabilized by a network of electrostatic interactions dominated by NP residues Arg240, Lys248, and Asp252. Using the DSSP secondary structure algorithm, NPBP (annotated as residues 20–47) shows segments Trp28 to Thr35 and Val40 to Asp42 as distinct helical conformations. The overall f_H is 0.4 and the bound form exhibits an R_g of 10.5 Å.

Experimental characterization of the secondary structure of the NPBP peptide free in solution by circular dichroism (CD) spectroscopy is reported to show the peptide as intrinsically disordered (Leung et al., 2015). When added to a solution of 50% trifluoroethanol (TFE), the NPBP peptide transitions from a coil to helical structures of ∼30–40% helicity, thus suggesting a strong underlying secondary-structure propensity. Predictions of secondary-structure without bias of the crystallographic structure estimate the NPBP peptide to encompass a consensus f_H ∼0.3 with probabilities >0.9 for helical formation in the sequence segment of Gly27 to Met34 (see, e.g., Kieslich et al., 2016).

Implicit Solvent T-ReX Simulations

To examine the accuracy of implicit solvent models to counterbalance the network of electrostatic interactions of the viral assembly interface that contribute to the stabilization of the NPBP helical fold and produce a conformational landscape with a predisposed helix propensity in bulk water, replica-exchange simulations were performed using different simulation strategies. The conformational sampling approach of SGLD was explored with two different GB solvent models and two different temperature-based replica-exchange methods. The first simulation model result shown in **Figure 2B** is the SGLD-GBMV2 with a static (fixed) set of temperatures in defining the replica-exchange protocol. The 2D profile $W_T\left(f_\mathrm{H},\mathrm{R_g}\right)$ computed at $T = 300\,\mathrm{K}$ using WHAM of the full ensemble shows a large manifold of conformational substates with a helix distribution of f_H ∼0–0.5. Several representative structures extracted from the basins are illustrated in **Figure 2E**. The conformational density takes place in R_g space of ∼8–11 Å and at the lower end of the population distribution non-structured states are observed to occupy a large range of R_g values and show the canonical feature of disorder.

Given the broad population distribution produced by a static set of temperatures in the T-ReX simulations, it is important to test whether the simulation model provided optimal sampling of the basins. To address this issue, an adaptive replica-exchange SGLD-GBMV2 simulation model was applied whereby allowing the clients to walk in temperature space to optimize the efficiency of exchanges between nearest-neighbor thermal windows at potential energy barriers separating conformational basins (Lee and Olson, 2011; Olson and Lee, 2014; Olson et al., 2016). The 2D profile from the adaptive T-ReX is illustrated in **Figure 2C** for $T = 300\,\mathrm{K}$ and the result is shown to retain the manifold of transient states similar to those sampled by the static T-ReX method, yet a population shift is observed toward an f_H ∼0.5 at the cost of reducing the density of unstructured conformations. The theoretical goal of the adaptive method is to enhance sampling of conformational transitions for a modeled potential energy surface. Early success of the method applied to a sharp phase transition of unfolding-folding of the protein SH3 showed better agreement with the experimental melting temperature than the traditional static approach (Lee and Olson, 2011). In addition, the adaptive method captured with greater accuracy the

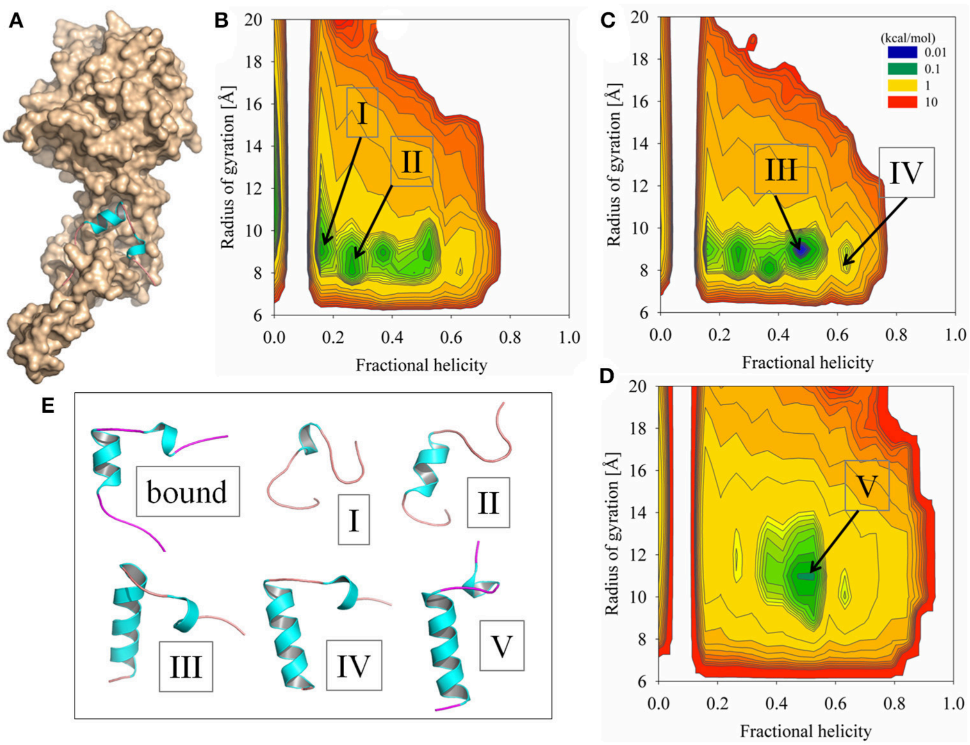


FIGURE 2 | Simulation results of sampling the Ebola virus VP35 NPBP peptide using GB-based solvent models combined with replica exchange methods. (A) X-ray crystallographic structure of the NPBP peptide bound to the Ebola NP (displayed as a molecular surface). **(B)** Probability density profile $W_T\,(f_\mathrm{H}, \mathrm{R_g})$ computed at $T = 300$ K and taken from the conformational ensemble modeled by the GBMV2 static T-ReX simulation method. The order parameters are fractional helicity and radius of gyration. **(C)** Probability density profile at $T = 300$ K from the adaptive T-ReX method with the GBMV2 solvent model. **(D)** Adaptive T-ReX with GBSW2 solvent model at the identical temperature. **(E)** Representative conformations extracted at $T = 300$ K from the simulations and are annotated at the indicated basins.

native state of SH3 extracted from the conformational ensemble. Given these earlier outcomes, and while the NPBP certainly lacks the folding cooperativity of SH3, the result suggests for the CHARMM22+CMAP/GBMV2 potential energy surface a NPBP "native" state of helix propensity near the value observed experimentally for the crystallographic bound conformation. Although the simulation shows a high rate of transitions among different basins, the overall population weight is inconsistent with the CD analysis in free solution. Because the potential energy surface is identical between the static and adaptive T-ReX methods, the less-efficient sampling approach will eventually converge to find a comparable $W_T\left(f_\mathrm{H}, \mathrm{R_g}\right)$.

To determine the bias of the GBMV2 solvent approximation on $W_T\left(f_\mathrm{H}, \mathrm{R_g}\right)$, adaptive T-ReX simulations were performed with a different implicit solvent model based on the GBSW2 approximation. Of the GB-based solvent models developed for protein dynamics, GBMV2 is one of the most accurate models in reproducing Poisson-Boltzmann theory with a Lee-Richards molecular surface (Feig et al., 2004b). The basis of GBMV2 is an analytical formulation of the molecular volume (Lee et al., 2003), while the less accurate but computationally much faster GBSW2 model is based on a smooth dielectric-boundary formulation constructed by applying a superposition of atomic-centered polynomials (Im et al., 2003). The dissimilarities between the two models in conformational sampling are clearly illustrated in **Figure 2D**. Application of GBSW2 significantly reduces the number of high-probability conformational excursions and leads to a folding funnel at f_H ~0.5. While the "optimized" f_H from the two different implicit solvent models is surprisingly similar, the limited disorder from the GBSW2 model in its current parameterization makes this solvent approximation less suitable for modeling IDPs (for an alternative parameterization of GBSW, see, e.g., Chen, 2010).

Figure 3 shows the probabilities of observing R_g as a function of three sampling temperatures taken from the ensemble. The GBMV2 model produced more compact states of NPBP than the crystallographic bound form, while GBSW2 yielded R_g values near the bound conformation. The observed difference between the solvent models can be partly attributed to the distinction in molecular surface representations, where different weights are applied to the surface-tension term that describes the hydrophobic free energy. In general, MD simulations of unfolded states are more compact and tend to favor helical structures than those found experimentally (Piana et al., 2014). By example, an experimental R_g for a unfolded 28 amino acids is estimated to be 13 Å (Kohn et al., 2004).

Also shown in **Figure 3** are the probability profiles of C_α-RMSD and the fraction of side-chain contacts similar to the

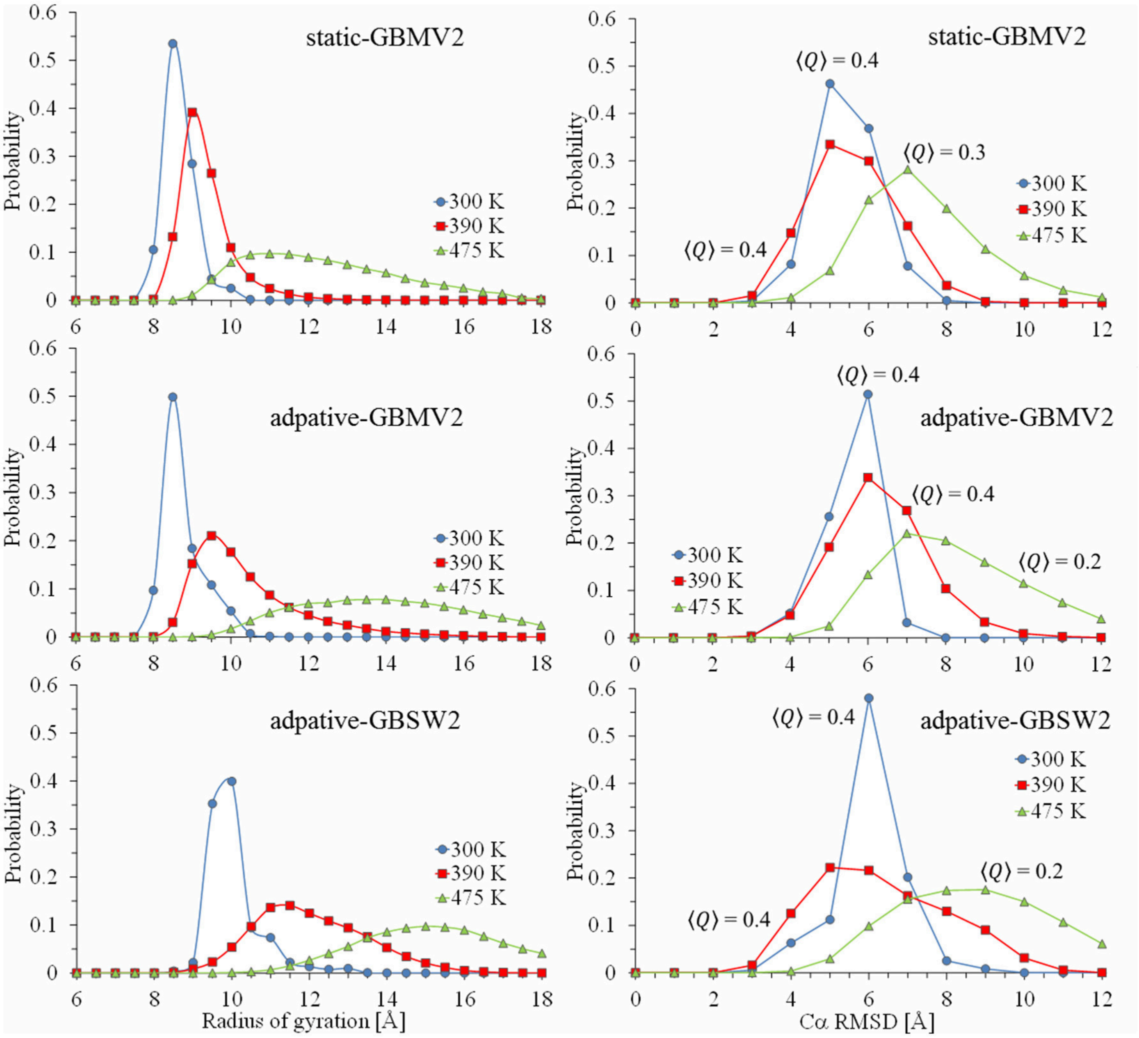


FIGURE 3 | Calculated probability profiles for sampling values of radius of gyration and Cα-RMSD from the starting bound conformation of the NPBP peptide. Plot lines colored blue represent quantities extracted at $T = 300$ K from the generated conformational ensembles, red represent values at 390 K and green at 475 K. From the top figure to bottom, simulation results are static T-ReX/GBMV2, adaptive T-ReX/GBMV2, and adaptive T-ReX/GBSW2.

starting conformation of NPBP. The ensemble average over contacts is denoted as <Q> and values <0.6 are considered unrelated to the starting structure. When combined with the analysis of the 2D profiles, the probabilities provide an interesting picture of the rare event of recognizing (via fly casting) a peptide conformation in the ensemble that is similar to the NPBP bound form. For the GBMV2 model and considering only the last 50 ns of simulation time, the lowest RMSD is 2.9 Å with $Q = 0.6$, and is clustered in the outer periphery of the highly-populated basin labeled as III in **Figure 2C**. This sparse cluster of low-RMSD states emerges with an f_{H} of 0.5 and R_{g} approaching 10 Å.

It is also important to understand the configurational stability of IDPs from the simulations and their fold propensities. The thermal unfolding profiles for NPBP are shown in **Figure 4A**. Consistent with the reduced number of transient states and their populations among the GB models, GBSW2 retains helicity over a greater thermal range. The aggregation of replica clients in the range of 360 K–425 K for the adaptive method (GBMV2 and GBSW2) is the effect of enhanced sampling of unfolding-folding transition points that stabilize helix formation. The statistical errors in the histograms for all model simulations are approximately f_{H} ±0.1 along the temperature profiles. Simulation convergence and the dominance of helix formation in NPBP can be further tested by conducting T-ReX simulations starting from a random coil state rather than the folded conformation. Although these additional simulations were executed only to 100 ns using the adaptive method, **Figure 4B** shows convergence to a folded state of helical conformations and establishes the strong helix propensity of applying CHARMM22+CMAP/GB descriptions.

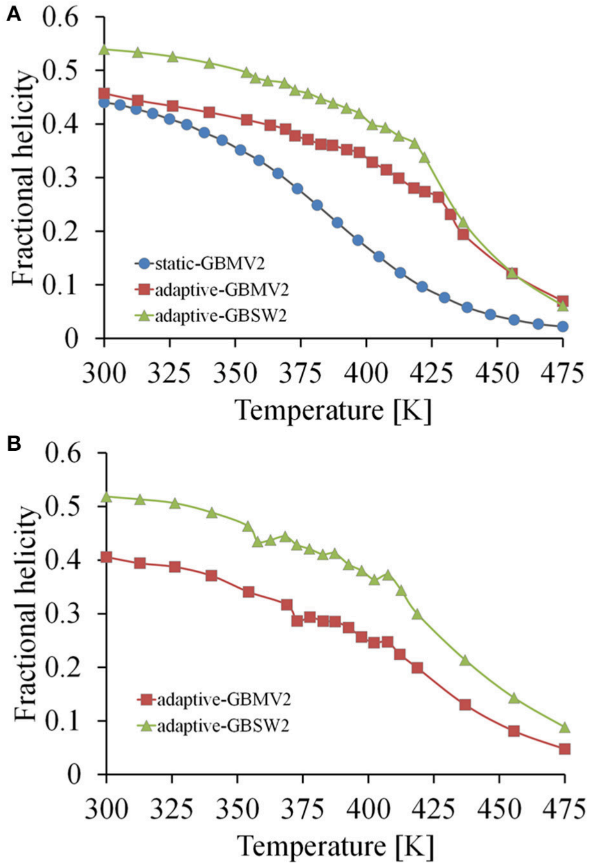


FIGURE 4 | Thermal unfolding profiles computed from the simulations of the Ebola VP35 NPBP peptide. (A) Profiles calculated from the starting folded conformation using the three simulation models of the static T-ReX/GBMV2 (blue colored line), adaptive T-ReX/GBMV2 (red colored line), and adaptive T-ReX/GBSW2 (green colored line). **(B)** Profiles calculated from the adaptive T-ReX simulations of starting from an unstructured (coil) peptide fold.

Explicit/Implicit Solvent Hybrid T-ReX/MD Simulation

The overweighting of secondary structure biases from the GBMV2 and GBSW2 solvent models is comparable to other studies of using different GB solvent models and parameterizations (Ganguly and Chen, 2009; Click et al., 2010; Chebaro et al., 2015). As a further test of the impact of the GBMV2 solvent model and its mean-field resolution of smearing out the details of the solvent on sampling conformational transitions of NPBP, the final simulation model tested is the explicit/implicit solvent hybrid T-ReX/MD method. This model generates peptide configurations on an explicit solvent (TIP3P) landscape while using the same number of replica clients as in the implicit solvent calculations. The latter is achieved by using the GBMV2 model in the Metropolis exchanges rather than explicit solvent. While the goal is to evaluate the simulation model in terms of a conformational landscape rather than unconstrained folding free energies to high accuracy, it is worth noting that replacement of energies in the Metropolis updates from an all-atom representation to a mean-field approximation can produce errors in the detailed balance required of a canonical ensemble (Chaudhury et al., 2012).

Figure 5 shows $W_T\left(f_{\mathrm{H}}, \mathrm{R_g}\right)$ at $T = 300\,\mathrm{K}$ from the WHAM calculation of the hybrid simulation model ensemble and the thermal unfolding profile. Several important observations can be made in comparison to the static GBMV2 model which best corresponds to the non-adaptive hybrid model. The most important distinction between the results is the striking difference in the favorable free energies and the network that shuttles conformations among the helical basins. While both sampling methods show sufficient plasticity among the states, the hybrid model shows a more quantifiable free-energy minimum at $f_{\mathrm{H}} = 0.26$ vs. 0.37 for the static GBMV2, and yields good agreement with secondary-structure predictions. The distinction in the potentials of mean force among the models is illustrated by considering a transition between an unstructured state and the free-energy minimum. For the static GBMV2, the transition ($f_{\mathrm{H}} = 0$; $R_g = 11$ Å) → ($f_{\mathrm{H}} = 0.37$; $R_g = 8$ Å) yields $\Delta G = -0.1$ kcal/mol, whereas for the adaptive model the transition from the same disordered state → ($f_{\mathrm{H}} = 0.47$; $R_g = 9$ Å) $\Delta G = -1.0$ kcal/mol, and for the hybrid model the transition → ($f_{\mathrm{H}} = 0.26$; $R_g = 9$ Å) yields $\Delta G = -1.7$ kcal/mol. Even though the static model exhibits a low-energy reversible transition to unstructured states and would appear to be in better agreement with the CD experiments (Leung et al., 2015), enhanced sampling of $P_\beta\left(f_{\mathrm{H}}, \mathrm{R_g}\right)$ by the adaptive method for this solvent description revealed a more costly transition to the densely populated $f_{\mathrm{H}} \sim 0.5$.

The lowest RMSD conformer for the hybrid model via the last 50 ns is 3.3 Å with $Q = 0.6$ and $R_g = 9.4$ Å. This conformer is illustrated in **Figure 5B** as the first structure depicted for the basin labeled III. The conformation is formed from a helical hairpin of residues Ser26-Met34 and Val40-Phe44. The top-rank conformer based on potential energies for the free-energy minimum at $f_{\mathrm{H}} = 0.26$ is illustrated as the first structure for basin I. This structure shows a 5-residue helix of Trp28-Met34. Among the highly populated basins, a distinction between the simulation models is the cluster at $f_{\mathrm{H}} = \sim 0.6$, where the hybrid model shows an improved free energy of population. Unlike the other basins, this basin lacks a direct low-energy pathway along the manifold of clusters.

A statistical average of the ensemble for the hybrid model computed from the multiple temperatures of the T-ReX simulation is illustrated in **Figure 5C** along with a comparison with the static GBMV2 model. Despite the differences in the potentials of mean force between the models, a simple statistical average without reweighting based on free energies shows remarkably similar f_{H} values at 300 K. Because of the lack of instantaneous relaxation of the explicit waters in contrast to GB approximations, the hybrid model shows a reduction in excursions of unfolded states at the upper R_g boundaries. Like many MD simulations of unfolded states with explicit solvent (Piana et al., 2014), a residual secondary-structure propensity is observed at 475 K.

The more compact favorable states observed in the explicit/implicit solvent hybrid model than that corresponding

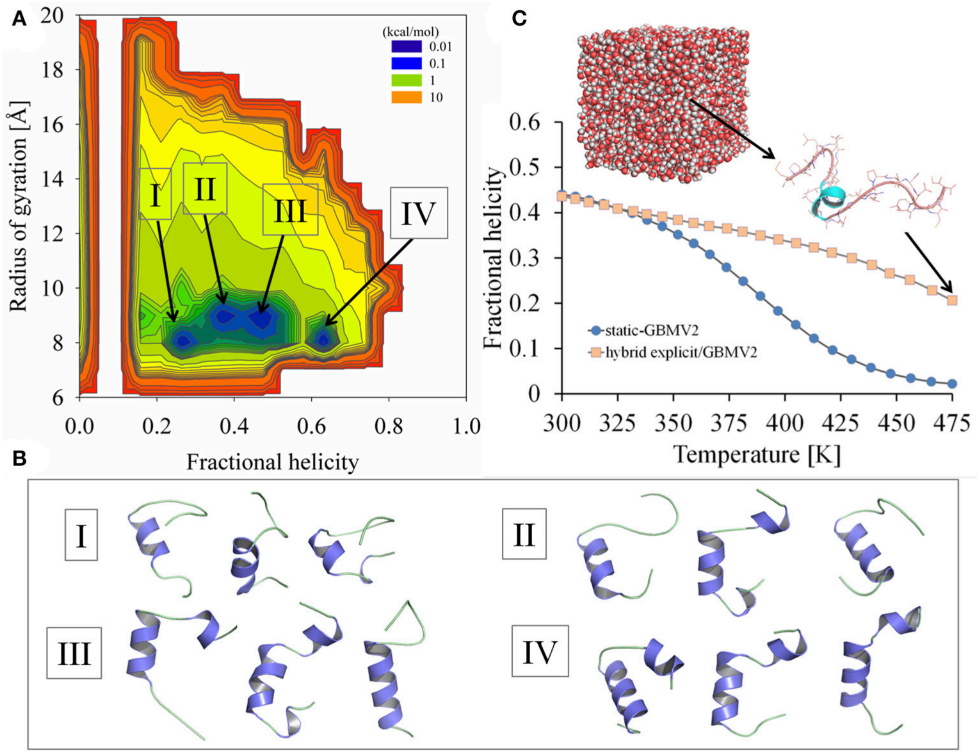


FIGURE 5 | Simulation results of sampling the Ebola virus VP35 NPBP peptide using the explicit/implicit solvent hybrid T-ReX/MD method. (A) Probability density profile W_T (f_H, R_g) computed at $T = 300$ K from sampling fractional helicity and radius of gyration. **(B)** Representative conformations extracted from the simulations are illustrated for selected basins. **(C)** Thermal unfolding profiles of the peptide computed using the explicit/implicit solvent hybrid T-ReX/MD method (light colored symbols) compared to the static T-ReX/SGLD method using GBMV2 (blue colored symbols). A representative structure is shown from the explicit solvent calculation.

to the bound NPBP conformation is unlikely due entirely to the GB model, but rather the additive force field (Piana et al., 2014). As noted above, the CHARMM22+CMAP force field was selected because of extensive benchmarks in reported studies of the GBMV2 and GBSW2 solvent descriptions to successfully model natively folded structures of proteins (see e.g., Yeh et al., 2008; Lee and Olson, 2010). While there are no reported studies of applying either GBMV2 or GBSW2 with the more refined CHARMM36m force field and its parameterization for TIP4P-based explicit solvent simulations (Huang et al., 2017), switching to this description may help reconcile the underestimated R_g values with those experimentally determined for unfolded states and reduce the overall weight and stabilization of secondary-structure propensities.

CONCLUSIONS

The current initiative to develop an atomistic understanding of "invisible" conformational states of the human/viral/bacterial proteomes requires an accurate computational framework for modeling conformational transitions within a disordered ensemble and their population density. The work presented here examined the application of temperature-based replica exchange simulations with different sampling methods and solvent descriptions of modeling an intrinsically disorder 28-residue peptide from the Ebola virus protein VP35. The X-ray crystallographic determination of the VP35 peptide bound to Ebola NP reports a helix-β-turn-helix fold of roughly 40% helical structure, whereas from CD experiments in free solution the peptide is unstructured. The simulations of the unbound peptide showed the selection of a GB solvent model combined with a replica-exchange sampling protocol can have a significant effect on the distribution of sampled populations. Overall, the tested GB models tend to favor a free-energy minimum of roughly 50% helical content for the peptide. The effect of an adaptive temperature-based replica exchange protocol compared to a traditional approach of a static set of temperatures was found to reduce the amount of unstructured states and shifted the ensemble to helical conformations with an extended peptide folding stabilization. A comparison with an explicit/implicit solvent hybrid MD-based replica exchange simulation showed that conformational sampling on an explicit solvent landscape leads to a free-energy minimum of ∼20% helicity, yet the overall conformational network underlying transient states resembles more of a helix-fold propensity in a solvent mixture of TFE-water rather than bulk water. The simulation results can be summarized as a benchmark for the testing of more refined CHARMM-based force fields and different GB model parameterizations.

The ultimate goal is to capture greater heterogeneity in conformational probabilities and reduce the over-stabilization of helix propensities in modeling intrinsically disordered peptides.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

Financial support for this work comes from US Department of Defense Threat Reduction Agency grant (DTRA 4.10011_07_RD_B). The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This article has been approved for public release with unlimited distribution.

ACKNOWLEDGMENTS

The author thanks Michael Lee for earlier collaborations on the adaptive temperature-based replica exchange method and assistance from Evan Olson on the graphics of 2D profiles.

REFERENCES

Arai, M., Sugase, K., Dyson, H. J., and Wright, P. E. (2015). Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Natl. Acad. Sci. U.S.A.* 112, 9614–9619. doi: 10.1073/pnas.1512799112

Bhowmick, A., Brookes, D. H., Yost, S. R., Dyson, H. J., Forman-Kay, J. D., Gunter, D., et al. (2016). Finding our way in the dark proteome. *J. Am. Chem. Soc.* 138, 9730–9742. doi: 10.1021/jacs.6b06543

Brooks, B. R., Brooks, C. L. III., Mackerell, A. D. Jr., Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30, 1545–1614. doi: 10.1002/jcc.21287

Chaudhury, S., Olson, M. A., Tawa, G., Wallqvist, A., and Lee, M. S. (2012). Efficient conformational sampling in explicit solvent using a hybrid replica exchange molecular dynamics method. *J. Chem. Theory Comput.* 8, 677–687. doi: 10.1021/ct200529b

Chebaro, Y., Ballard, A. J., Chakraborty, D., and Wales, D. J. (2015). Intrinsically disordered energy landscapes. *Sci. Rep.* 5:10386. doi: 10.1038/srep10386

Chen, J. (2010). Effective approximation of molecular volume using atom-centered dielectric functions in generalized Born models. *J. Chem. Theory Comput.* 6, 2790–2803. doi: 10.1021/ct100251y

Click, T. H., Ganguly, D., and Chen, J. (2010). Intrinsically disordered proteins in a physics-based world. *Int. J. Mol. Sci.* 11, 5292–5309. doi: 10.3390/ijms11125292

Feig, M., Karanicolas, J., and Brooks, C. L. III. (2004a). MMTSB Tool Set: Enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* 22, 377–395. doi: 10.1016/j.jmgm.2003.12.005

Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., and Brooks, C. L. III. (2004b). Performance comparison of generalized born and poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* 25, 265–284. doi: 10.1002/jcc.10378

Ferrenberg, A. M., and Swendsen, R. H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63, 1195–1198. doi: 10.1103/PhysRevLett.63.1195

Gallicchio, E., Andrec, M., Felts, A. K., and Levy, R. M. (2005). Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B.* 109, 6722–6731. doi: 10.1021/jp045294f

Ganguly, D., and Chen, J. (2009). Atomistic details of the disordered states of KID and pKID. Implications in coupled binding and folding. *J. Am. Chem. Soc.* 131, 5214–5223. doi: 10.1021/ja808999m

Ganguly, D., and Chen, J. (2015). Modulation of the disordered conformational ensembles of the p53 transactivation domain by cancer-associated mutations. *PLoS Comput. Biol.* 11:e1004247. doi: 10.1371/journal.pcbi.1004247

Higo, J., Nishimura, Y., and Nakamura, H. (2011). A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. *J. Am. Chem. Soc.* 133, 10448–10458. doi: 10.1021/ja110338e

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., et al. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14, 71–73. doi: 10.1038/nmeth.4067

Im, W., Lee, M. S., and Brooks, C. L. III. (2003). Generalized born model with a simple smoothing function. *J. Comput. Chem.* 24, 1691–1702. doi: 10.1002/jcc.10321

Ishikawa, Y., Sugita, Y., Nishikawa, T., and Okamoto, Y. (2001). Ab initio replica-exchange monte carlo method for cluster studies. *Chem. Phys. Lett.* 33, 199–206. doi: 10.1016/S0009-2614(00)01342-7

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi: 10.1063/1.445869

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Katzgraber, H. G., Trebst, S., Huse, D. A., and Troyer, M. (2006). Feedback-optimized parallel tempering Monte Carlo. *J. Stat. Mech. Theory Exp.* 2006:P03018. doi: 10.1088/1742-5468/2006/03/p03018

Kieslich, C. A., Smadbeck, J., Khoury, G. A., and Floudas, C. A. (2016). conSSert: consensus SVM model for accurate prediction of ordered secondary structure. *J. Chem. Inf. Model.* 56, 455–461. doi: 10.1021/acs.jcim.5b00566

Kohn, J. E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N., et al. (2004). Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12491–12496. doi: 10.1073/pnas.0403643101

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 13, 1011–1021. doi: 10.1002/jcc.540130812

Lee, K. H., and Chen, J. (2016). Multiscale enhanced sampling of intrinsically disordered protein conformations. *J. Comput. Chem.* 37, 550–557. doi: 10.1002/jcc.23957

Lee, M. S., Feig, M., Salsbury, F. R. Jr., and Brooks, C. L. III. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. *J. Comput. Chem.* 24, 1348–1356. doi: 10.1002/jcc.10272

Lee, M. S., and Olson, M. A. (2010). Protein folding simulations combining self-guided Langevin dynamics and temperature-based replica exchange. *J. Chem. Theory Comput.* 6, 2477–2487. doi: 10.1021/ct100062b

Lee, M. S., and Olson, M. A. (2011). Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. *J. Chem. Phys.* 134, 244111–224417. doi: 10.1063/1.3603964

Lee, M. S., Salsbury, F. R. Jr., and Brooks, C. L.III. (2002). Novel generalized born methods. *J. Chem. Phys.* 116, 10606–10614. doi: 10.1063/1.1480013

Leung, D. W., Borek, D., Luthra, P., Binning, J. M., Anantpadma, M., Liu, G., et al. (2015). An intrinsically disordered peptide from Ebola virus VP35 controls viral RNA synthesis by modulating nucleoprotein-RNA interactions. *Cell Rep.* 11, 376–389. doi: 10.1016/j.celrep.2015.03.034

Mackerell, A. D. Jr., Feig, M., and Brooks, C. L. III. (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* 25, 1400–1415. doi: 10.1002/jcc.20065

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Miao, Y., Feixas, F., Eun, C., and McCammon, J. A. (2015). Accelerated molecular dynamics simulations of protein folding. *J. Comput. Chem.* 36, 1536–1549. doi: 10.1002/jcc.23964

Mittal, A., Lyle, N., Harmon, T. S., and Pappu, R. V. (2014). Hamiltonian switch Metropolis Monte Carlo simulations for improved conformational sampling of intrinsically disordered regions tethered to ordered domains of proteins. *J. Chem. Theory Comput.* 10, 3550–3562. doi: 10.1021/ct5002297

Olson, M. A., and Lee, M. S. (2013). Structure refinement of protein model decoys requires accurate side-chain placement. *Proteins* 81, 469–478. doi: 10.1002/prot.24204

Olson, M. A., and Lee, M. S. (2014). Evaluation of unrestrained replica-exchange simulations using dynamic walkers in temperature space for protein structure refinement. *PLoS ONE* 9:e96638. doi: 10.1371/journal.pone.0096638

Olson, M. A., Legler, P. M., and Goldman, E. R. (2016). Comparison of replica exchange simulations of a kinetically trapped protein conformational state and its native form. *J. Phys. Chem. B.* 120, 2234–2240. doi: 10.1021/acs.jpcb.6b00233

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., et al. (2015). Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15898–15903. doi: 10.1073/pnas.1508380112

Peter, E. K., Shea, J. E., and Pivkin, I. V. (2016). Coarse kMC-based replica exchange algorithms for the accelerated simulation of protein folding in explicit solvent. *Phys. Chem. Chem. Phys.* 18, 13052–13065. doi: 10.1039/C5CP06867C

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802. doi: 10.1002/jcc.20289

Piana, S., Klepeis, J. L., and Shaw, D. E. (2014). Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* 24, 98–105. doi: 10.1016/j.sbi.2013.12.006

Predescu, C., Predescu, M., and Ciobanu, C. V. (2004). The incomplete beta function law for parallel tempering sampling of classical canonical systems. *Chem. Phys.* 120, 4119–4128. doi: 10.1063/1.1644093

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341. doi: 10.1016/0021-9991(77)90098-5

Sanchez, A., Geisbert, T. W., and Feldmann, H. (2006). "Filoviridae: Marburg and Ebola viruses," in *Fields Virology,* eds D. M. Knipe, P. M. Howley, R. A. Griffin, M. A. Martin, B. Roizman, and S. E. Straus (Philadelphia, PA: Lippincott Williams & Wilkins), 1409–1448.

Shoemaker, B. A., Portman, J. J., and Wolynes, P. G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U.S.A.* 97, 8868–8873. doi: 10.1073/pnas.160259697

Sugitaa, Y., and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314, 141–151. doi: 10.1016/S0009-2614(99)01123-9

Trebst, S., Troyer, M., and Hansmann, U. H. (2006). Optimized parallel tempering simulations of proteins. *J. Chem. Phys.* 124, 174903–174909. doi: 10.1063/1.2186639

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293, 321–331. doi: 10.1006/jmbi.1999.3110

Wright, P. E., and Dyson, H. J. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208. doi: 10.1038/nrm1589

Wu, X., and Brooks, B. R. (2003). Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* 381, 512–518. doi: 10.1016/j.cplett.2003.10.013

Wu, X., and Brooks, B. R. (2011). Toward canonical ensemble distribution from self-guided Langevin dynamics simulation. *J. Chem. Phys.* 134, 134108–134119. doi: 10.1063/1.3574397

Wu, X., Brooks, B. R., and Vanden-Eijnden, E. (2016). Self-guided Langevin dynamics via generalized Langevin equation. *J. Comput. Chem.* 37, 595–601. doi: 10.1002/jcc.24015

Wu, X., Damjanovic, A., and Brooks, B. R. (2012). Efficient and unbiased sampling of biomolecular systems in the canonical ensemble: a review of self-guided langevin dynamics. *Adv. Chem. Phys.* 150, 255–326. doi: 10.1002/9781118197714.ch6

Yeh, I. C., Lee, M. S., and Olson, M., A. (2008). Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. *J. Phys. Chem. B.* 112, 15064–15073. doi: 10.1021/jp802469g

Zhang, W., and Chen, J. (2014). Replica exchange with guided annealing for accelerated sampling of disordered protein conformations. *J. Comput. Chem.* 35, 1682–1689. doi: 10.1002/jcc.23675

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 25 April 2017
doi: 10.3389/fmolb.2017.00022

Computing Spatiotemporal Heat Maps of Lipid Electropore Formation: A Statistical Approach

Willy Wriggers [1, 2]*, Federica Castellani [3], Julio A. Kovacs [1, 2] and P. Thomas Vernier [3]

[1] Institute of Biomedical Engineering, Old Dominion University, Norfolk, VA, USA, [2] Department of Mechanical and Aerospace Engineering, Old Dominion University, Norfolk, VA, USA, [3] Frank Reidy Research Center for Bioelectrics, Old Dominion University, Norfolk, VA, USA

We extend the multiscale spatiotemporal heat map strategies originally developed for interpreting molecular dynamics simulations of well-structured proteins to liquids such as lipid bilayers and solvents. Our analysis informs the experimental and theoretical investigation of electroporation, that is, the externally imposed breaching of the cell membrane under the influence of an electric field of sufficient magnitude. To understand the nanoscale architecture of electroporation, we transform time domain data of the coarse-grained interaction networks of lipids and solvents into spatial heat maps of the most relevant constituent molecules. The application takes advantage of our earlier graph-based activity functions by accounting for the contact-forming and -breaking activity of the lipids in the bilayer. Our novel analysis of lipid interaction networks under periodic boundary conditions shows that the disruption of the bilayer, as measured by the breaking activity, is associated with the externally imposed pore formation. Moreover, the breaking activity can be used for statistically ranking the importance of individual lipids and solvent molecules through a bridging between fast and slow degrees of freedom. The heat map approach highlighted a small number of important lipids and solvent molecules, which allowed us to efficiently search the trajectories for any functionally relevant mechanisms. Our algorithms are freely disseminated with the open-source package *TimeScapes*.

Keywords: molecular dynamics, trajectory analysis, multiple time scales, mutual information, distance geometry, contact network

Edited by:
Slavica Jonic,
IMPMC, Sorbonne Universités - CNRS UMR 7590, UPMC Univ Paris 6, MNHN, IRD UMR 206, France

Reviewed by:
Rainer A. Böckmann,
University of Erlangen-Nuremberg, Germany
Francisco Monroy,
Complutense University of Madrid, Spain
Epameinondas Leontidis,
Massachussetts Institute of Technology, USA

Correspondence:
Willy Wriggers
wriggers@biomachina.org

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *04 January 2017*
Accepted: *28 March 2017*
Published: *25 April 2017*

Citation:
Wriggers W, Castellani F, Kovacs JA and Vernier PT (2017) Computing Spatiotemporal Heat Maps of Lipid Electropore Formation: A Statistical Approach. Front. Mol. Biosci. 4:22. doi: 10.3389/fmolb.2017.00022

1. INTRODUCTION

Membrane electroporation is a biomedical technique that artificially increases the permeability of cell membranes by applying short electric pulses (Neumann et al., 1982). Electroporation by an external electric field is attributed to the opening of discrete nanometer-sized pores in cell membranes: In some plasma and biomedical experiments, pulsed fields have high power (of the order of megavolts per meter) but short duration (of the order of nanoseconds) (Kohler et al., 2015), conditions that are easily accessible to atomistic molecular dynamics (MD) simulations. Other electroporative applications such as the electroinsertion of xenoproteins or electrofusion of cells are performed in experiments at much lower voltages and over longer time scales; in these cases, statistical theories may bridge between single-event poration times derived from MD simulations and slower experimental kinetics (Böckmann et al., 2008).

Direct experimental observations of electropore formation in biological membranes are not possible because of their small size and short duration. MD simulations of single-pore formation under an external electric field have consequently been of considerable interest for some time (Vernier and Ziegler, 2007; Böckmann et al., 2008; Ziegler and Vernier, 2008; Tokman et al., 2013; Kohler et al., 2015). Polar water molecules are known to play a key driving role in electroporation (**Figure 1**); however, no signature for pore initiation has yet been identified (Vernier and Ziegler, 2007; Ziegler and Vernier, 2008; Kohler et al., 2015), and Kohler et al. (2015) argued that a statistical framework would be needed for further development.

We have recently developed such a statistical approach for detecting allosteric signatures in protein MD simulations. *TimeScapes* is a Python-based program package that can be used to efficiently detect and characterize significant conformational changes in simulated biomolecular systems (Wriggers et al., 2009). We recently added a new functionality to TimeScapes that transforms time-domain information from MD trajectories into spatial heat maps (Kovacs and Wriggers, 2016) that can be visualized on 3D molecular structures or in the form of interaction networks. The method is multiscale in the time domain in that it uses statistical bridging between the fast, local variables recorded by MD and the slow, global rate of change of the simulated system that is characterized by a so-called activity function. In our work "activity" denotes a non-negative scalar function of time that quantifies the structural variability of the system (as introduced by Wriggers et al., 2009 and described in Kovacs and Wriggers, 2016). As simple example of an activity function is the RMS fluctuation in a sliding window. Low activity corresponds to quiescent periods of relative structural stability, whereas high activity corresponds to significant structural transitions between adjacent quiescent basins (Wriggers et al., 2009). Once the slow, global activity is quantified, the bridging between fast and slow time series can then be performed using either the Pearson cross-correlation or a nonlinear mutual information solver called Fast Information Matching (FIM).

In our recent work, we noted a potential weakness of FIM owing to the uniform Parzen window approach used in density estimation, which does not adapt well to activities that are zero-valued for some part of the simulation (Kovacs and Wriggers, 2016). In protein applications, we prefer the use of the sliding window RMS fluctuation activity that yields proper density histograms even for small systems and thereby avoids this issue. However, in the liquid (lipid or aqueous solvent) applications considered in this study, there is no stable structure that can be used as a reference for RMS fluctuation calculation. Instead, the distance geometry of intermolecular contacts is used; specifically, we use one of the two graph-based activities *TimeScapes* provides for contact networks. These graph-based activities (shown in **Figure 1** and further explained below) scale quadratically with the system size and rely on a spatial coarse-graining of the structure to reduce the computational complexity, resulting in potentially zero-valued activity functions unamenable to FIM analysis. The present generalization of our heat map analysis to lipids and solvents therefore required us to develop an adaptive bandwidth allocation for the mutual information solver, which was performed separately by Kovacs et al. (2017). The resulting Balanced Adaptive Density Estimation (BADE) code for mutual information calculations is more accurate and efficient and can replace the previously used FIM code (Kovacs and Wriggers, 2016) in future versions of our *TimeScapes* package.

The "Methods" section briefly describes the theory of heat map prediction with *TimeScapes* and the adaptations that are necessary to generalize the protein-based approach to lipid and solvent dynamics. We also describe MD protocols for the electroporation simulations conducted in this study. The "Results" section first establishes activity functions that are suitable for characterizing membrane pore formation before providing examples of lipid pore formation heat maps. We explore dependencies on critical parameters of the algorithm and show heat maps of the surrounding water-ion solutions. The "Conclusions" section presents the benefits and limitations of the current framework and discusses areas for future development.

2. METHODS

2.1. Transforming Distance Geometry Time Series into Spatial Heat Maps

In this paper, we study the time-dependent distance geometry between water, ion, or lipid pairs. Let $\{X_{i,j}(t)\}$ denote pairwise distances between such "residues" (a term commonly used in MD for covalently bonded molecules that are separated by a topology or force field), where i and j are suitably chosen indices. For residues that have more than one atom, such as water molecules or lipids, pairwise distances are defined by the position of characteristic atoms (e.g., water oxygens or lipid phosphorus atoms). The time-dependent distance geometry $X_{i,j}(t)$ comprises "fast" variables, that is, they exhibit fluctuations on time scales of the order of the frame length of the discrete MD trajectory. Furthermore, let $a(t)$ denote a scalar, non-negative "slow" activity function that describes the variability of the simulated system as a function of time, as described above. Finally, let $I(f, g)$ denote a statistical measure of dependence of two discrete random variables f and g (such as Pearson cross-correlation or mutual information). The coefficient

$$R_{X,a}(i,j) = I\left(\left|\frac{dX_{i,j}(t)}{dt}\right|,\ a(t)\right) \tag{1}$$

then provides an estimate of the spatial importance of local changes in the residue network for the global activity. In this work we are using absolute time-differentials of the fast variables for the statistical dependence analysis with the activity; this way both fast and slow timeseries correspond to a non-negative rate of change and are compatible. $R_{X,a}(i,j)$ values can then be used to rank all members of the family $\{X_{i,j}(t)\}$; this, after appropriate mapping to spatial features i (see below), yields a heat map of the importance of fast, local variables for slow, global activities. Our transformation of time series data to spatial images can be applied to various imaging modalities $X(t)$. However, in this study, we restrict our discussion to the abovementioned pairwise residue distances, because the distance geometry provides a suitable

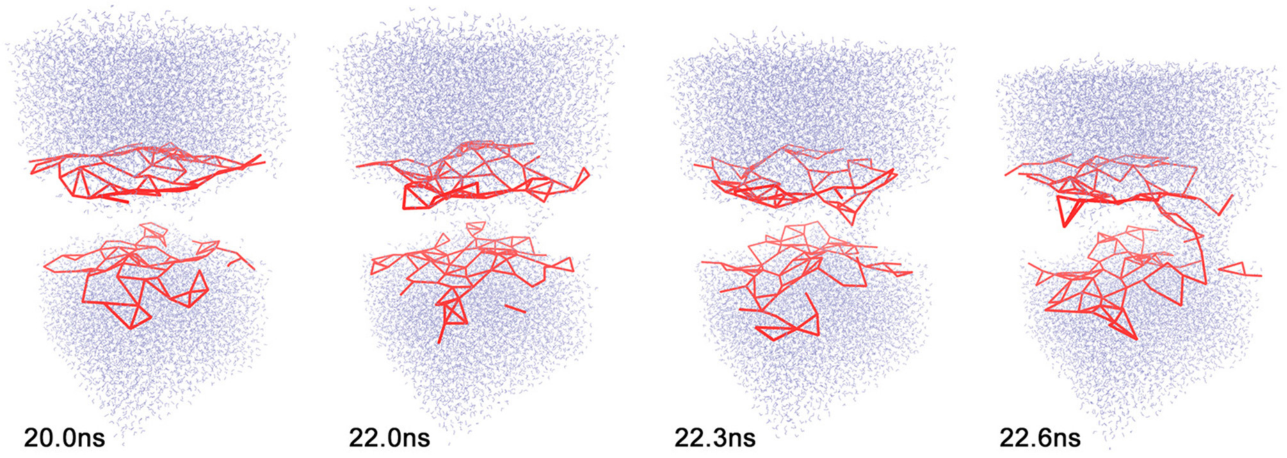
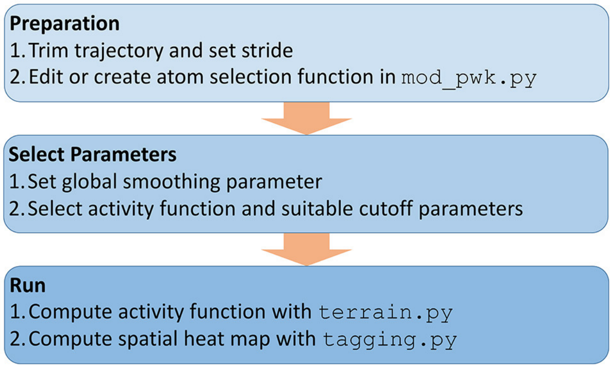


FIGURE 1 | Sequence of events during electropore formation. Only water (top and bottom) and the coarse-grained 11 Å cutoff graph (red) of lipid phosphorus atoms are shown for clarity. Frames of trajectory 1 (see text) are shown at simulation times t = 20.0, 22.0, 22.3, and 22.6 ns. All molecular graphics figures in the present paper were created using the VMD program (Humphrey et al., 1996).

characterization of interactions in the absence of a global frame of reference.



Preparation
1. Trim trajectory and set stride
2. Edit or create atom selection function in `mod_pwk.py`

Select Parameters
1. Set global smoothing parameter
2. Select activity function and suitable cutoff parameters

Run
1. Compute activity function with `terrain.py`
2. Compute spatial heat map with `tagging.py`

FIGURE 2 | Workflow of using *TimeScapes* for the lipid heat map analysis as described in the "Methods" section.

2.2. Lipid Heat Map Application Workflow in *TimeScapes*

Figure 2 shows an overview of the necessary analysis steps in our *TimeScapes* package (Wriggers et al., 2009). Before using *TimeScapes*, it is necessary to trim a trajectory to a time window of interest and to set the stride (trajectory time step). We selected time windows based on the timing of pore formation, which differed between the trajectories in this study. The end times were chosen by visual inspection when the pore size reached approximately 20% of the unit cell dimensions. The start times were chosen such that the window contained only the lead-up events immediately prior to pore formation, with full solvent perforation of the bilayer commencing at the 60% mark of the window. As a result, the poration process was normalized across the trajectory windows.

The heat maps are robust under variations in stride; however, the fastest variables captured in the analysis are limited by this choice. In our work, we selected strides of 1 and 10 ps that provided sufficient sampling of the time-dependent distance geometry.

Next, a user may have to modify the selection functions for representative atoms based on the atom and residue names defined by the force field (**Figure 2**). This step is necessary for the coarse-graining of the interaction networks (**Figure 1**). In this study, we added functions for selecting the lipid phosphorus atoms and water oxygens, which involved a straightforward edit of the available Python templates in the `mod_pwk.py` source file.

Few parameters must be set to run the required *TimeScapes* tools (**Figure 2**). An important choice is the temporal smoothing parameter that determines the temporal level of detail captured by the activity function $a(t)$. This parameter affects both the detection of events and the bandwidth of the activity function estimation, as described in Wriggers et al. (2009). As a rule of thumb, we recommend values of approximately 5% of the trajectory window length (actual numbers are provided in the figure captions below). The "Results" section shows an evaluation of the parameter space for ensuring that the resulting lipid heat maps are robust.

The global activity of the system $a(t)$, which is required for heat map analysis, can be computed from changes in a distance cutoff-based adjacency graph or from a so-called Generalized Masked Delaunay graph (Wriggers et al., 2009). In this work, we chose a cutoff graph (**Figure 1**) of lipid phosphorous atom distances because it decomposes structural changes into separate contact-forming and -breaking activity of adjacent lipids (**Figures 3**, **4**). The Generalized Masked Delaunay graph (not discussed here) is less affected by distances and is thus less capable of differentiating between forming and breaking events

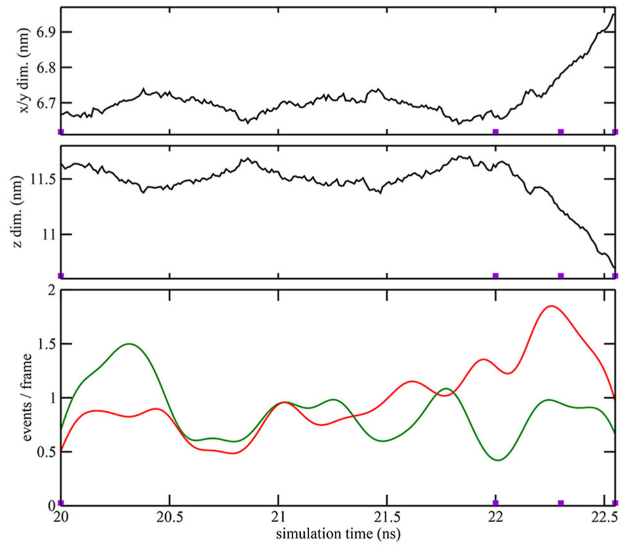


FIGURE 3 | Quantitative characterization of pore formation in ion-free trajectory 1. The periodic unit cell dimensions (top, center) and (bottom) cutoff graph forming (green) and breaking (red) activities are shown as a function of simulation time. The graph cutoff parameters were 11 and 13 Å, and the smoothing parameter was 200 ps. The violet markers indicate the times of the four snapshots shown in **Figure 1**.

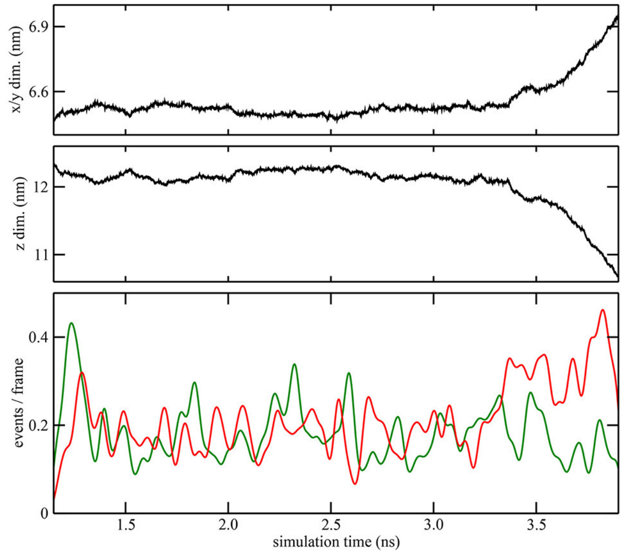


FIGURE 4 | Quantitative characterization of pore formation in ion-containing trajectory 2. The periodic unit cell dimensions (top, center) and (bottom) cutoff graph forming (green) and breaking (red) activities are shown as a function of simulation time. The graph cutoff parameters were 11 and 13 Å, and the smoothing parameter was 50 ps.

(Wriggers et al., 2009). *TimeScapes* also supports the calculation of RMS-fluctuation-based activity of Cartesian coordinates in a Gaussian-weighed sliding window. However, this approach requires a global frame of reference for least-squares fitting, such as a protein structure, which is not available in our liquid systems. Consequently, among the three activity functions available in *TimeScapes*, we only use the cutoff graph that can be computed using the `terrain.py` tool (with the user-provided phosphorus selection function). This graph requires the setting of two distance cutoff values for the event detection buffer. As a rule of thumb, the cutoff values should reflect the nearest-neighbor distances in the coarse grained model (i.e., lipid phosphorus atoms; **Figure 1**).

Finally, after computing the activity function using `terrain.py`, the program `tagging.py` uses it to compute the positive symmetric matrix $R_{X,a}(i,j)$ of the ranking coefficients between the time series $X_{i,j}(t)$ and $a(t)$ (**Figure 2**). This matrix quantifies the statistical dependence of every residue pair (i,j) with the activity function $a(t)$. As discussed in Kovacs and Wriggers (2016), the matrices $R_{X,a}(i,j)$ show a banded structure owing to the global nature of the statistical relationship between the activity and the concomitant change in distances from a particular residue to neighboring residues (Shaw et al., 2010). The banded structure of this matrix allows us to compress the columns of $R_{X,a}(i,j)$ to their average $R_{X,a}(i)$, so we can visualize the pairwise heat maps in three dimensions. We note that unlike in the earlier heat map projection to relatively stable protein structures (Kovacs and Wriggers, 2016), the heat maps in the present paper are projected to lipid or solvent molecules that undergo diffusive motion throughout the trajectory. Because we are essentially drawing an image on a moving canvas, it is valuable to visualize the results on a time-dependent trajectory instead of static structures.

The adaptation of *TimeScapes* to lipid systems in this study also required some updating of the source code, mainly to deal with the periodic boundaries and to read the unit cell box dimensions. The updated code will be released with version 1.5 at our web site, http://timescapes.biomachina.org.

2.3. Molecular Dynamics Simulations of Electroporation

Atomic-scale MD simulations of a symmetric phospholipid bilayer were performed using the GROMACS 4.6.6 software package (van der Spoel et al., 2005) on the Turing High Performance Computing cluster at Old Dominion University (Old Dominion University, 2017). A system containing only lipids and approximately 12,000 water molecules was created using the MemBuilder tool (Ghahremanpour et al., 2014). Four trajectories were generated for this paper. Trajectory 1 contained no ions. For trajectories 2–4 the built-in GROMACS function `genion` was used to replace bulk water molecules with Ca^{2+} and Cl^{-}. This generated an ionic solution comprising 20 calcium ions, 40 chloride ions, and approximately 12,000 water molecules. The CHARMM36 force field and TIP3P water model were used. The charge and size for both calcium and chloride ions were rescaled in accordance with Kohagen et al. (2014) to improve the ion-water interactions and to avoid unrealistic ion clustering. The simulation volume for both systems contained 128 (64 per leaflet) lipid molecules—1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylcholine (POPC)—with initial box dimensions of approximately 7 × 7 × 12 nm. The system was equilibrated for

1,500 ns to allow calcium-phospholipid binding (Vernier et al., 2009) and to stabilize the area per lipid. All simulations were performed under the NPT ensemble. A temperature of 310 K was maintained using a velocity rescaling algorithm (Bussi et al., 2007). A pressure of 1 bar was maintained using a Berendsen barostat (Berendsen et al., 1984) with relaxation time of 1 ps and compressibility of 4.5×10^{-5} bar^{-1} applied semi-isotropically in both normal and in-plane directions relative to the membrane. Bond lengths were constrained using the LINCS algorithm (Hess et al., 1997) for lipids and SETTLE (Miyamoto and Kollman, 1992) for water. Short-range electrostatic and Lennard-Jones interactions were cut off at 1.0 nm. Long-range electrostatics were calculated using the PME algorithm (Essmann et al., 1995), and boundary conditions were used to mitigate system size effects. The integration time step was 2 fs. An electric field of 400 MV/m was applied along the z-axis normal to the (x,y) bilayer plane. Under these conditions, pores form within approximately 3–20 ns. Trajectory 1 was run for 23.5 ns with a stride (trajectory saving time) of 10 ps. For systems containing calcium, three independent trials of lengths 5.3, 23.4, and 16.5 ns were run with a stride of 1 ps by assigning a randomized velocity to each atom after system equilibration. The trajectory windows selected to normalize the pore formation times (see above) were frames 2001–2256, 1150–3900, 19350–22100, and 13550–15300 for trajectories 1–4, respectively. The selection of these windows had the added benefit that any initial periodic box deformation of the system was already discounted (applying an electric field normal to the plane of a lipid bilayer under the NPT ensemble causes a reduction in the bilayer thickness and a corresponding change in the box dimensions, preceding and independent of pore formation).

3. RESULTS

3.1. Activity Functions Relevant for Pore Formation

As described in the "Methods" section, one of the prerequisites of the heat map analysis is the use of an activity function that characterizes the global change of the system. This work focuses on pore formation that introduces an anisotropic pressure in the system and by virtue of the NPT ensemble yields a compression of the periodic unit cell in the z-direction and an associated elongation in the x- and y-directions. The unit cell dimensions can therefore be used as a geometric marker for pore formation, as shown in the top and center plots of **Figures 3**, **4** for ion-free trajectory 1 and for one of the ion systems, trajectory 2, respectively. The data for trajectories 3 and 4 were similar to those of trajectory 2 and are omitted for brevity. The four simulation times in **Figure 3** can be compared visually to the corresponding snapshots in **Figure 1**.)

In addition to the geometric characterization of the poration process, we also computed the time-dependent cutoff graph (**Figure 1**) that decomposes structural changes in the lipid bilayer into separate contact-forming and -breaking events. The resulting lipid-forming and -breaking activities are plotted at the bottom of **Figures 3**, **4**. All four trajectories showed sustained breaking activity during pore formation that is not equally compensated for by the forming of lipid contacts. Consequently, we used the lipid-breaking activity (red graphs) as a measure of pore formation in our subsequent heat map analysis. The unit cell deformation was then used as an independent measure for validation.

3.2. Gallery of Lipid Pore Formation Heat Maps

The lipid heat maps shown in **Figure 5** visualize the importance of individual lipids for the contact-breaking activity associated with pore formation. This figure shows mutual information heat maps of trajectories 1–4 for both sides of the bilayer (as viewed in the +z and −z directions). In many, but not all, heat maps, we observe hot spots that are clearly associated with the emerging pore (most notably in views 1+, 2+, 4+, and 4−). Some cases also show outliers that are not associated with the pore (1−, 2−, 3+, and 3−). The heat maps are very valuable because they allow a user to focus on the relatively small number of statistically significant lipids. However, a detailed inspection of the trajectories does not reveal a consistent mechanism of action of these lipids. The outliers have their head groups exposed to the solvent, and they are important for a general destabilization of the bilayer that facilitates pore formation; however, there is no indication that the outliers participate in the actual poration event. The highlighted lipids that line the pore exhibit a disruption of their contact network; however, an inspection suggests that this appears to be a passive response to the tunneling of water molecules across the membrane.

3.3. Validation

We conducted a number of alternative heat map calculations to test the robustness of our approach and to validate the results shown in **Figure 5**. **Figure 6** shows the results when replacing mutual information **Figures 6a,c** with the Pearson cross-correlation **Figures 6b,d** and when replacing the graph-based lipid activity **Figures 6a,b** with the box dimensions **Figures 6c,d** in the trajectory 4 heat map. The Pearson cross-correlation is a simple, linear measure of statistical dependence. As implemented in `tagging.py`, negative correlations serve to measure the noise floor and are afterwards set to zero (only positive correlations between rates of change make physical sense, so negative correlations are deemed noise). However, the mutual information captures all non-linear dependencies (including negative linear correlations). Therefore, the resulting mutual information heat maps are smoother, whereas Pearson cross-correlation heat maps show higher dynamic range. Despite these differences, the two measures show comparable features.

Remarkably, heat maps are also largely unaffected by the type of activity function used (**Figure 6**). As shown in **Figures 3**. **4**, the red lipid-breaking activity graphs can be quite different from the box dimension graphs; however, the trajectory 4 heat map shows the same features in either case. This demonstrates that our contact-breaking activity is indeed a suitable measure for pore formation. Minor discrepancies in the heat maps in **Figure 6** are

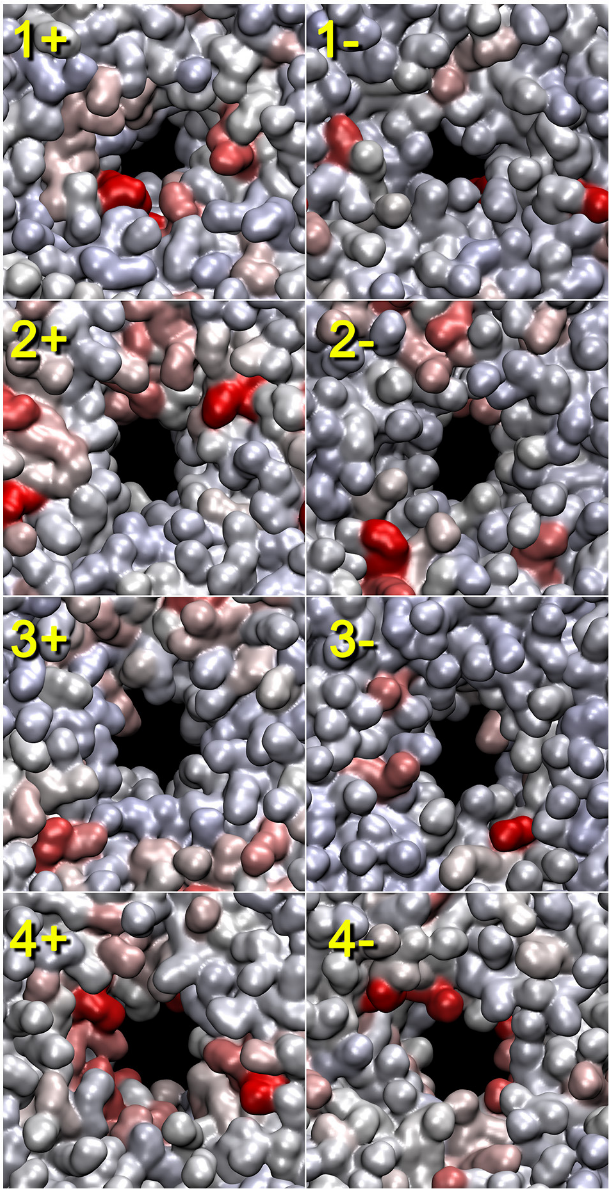


FIGURE 5 | Lipid heat maps indicating the statistical importance of individual lipids for the pore formation (as represented by the lipid contact breaking activity). The mutual information analysis for the lipid bilayer was conducted using `tagging.py` with the BADE solver described in Kovacs et al. (2017). The temporal smoothing parameter was 200 (trajectory 1) and 50 ps (trajectories 2–4). The front (+z direction) and rear (-z direction) views of the heat maps generated from the four trajectories are shown. Solvent molecules are omitted for clarity. Lipid heat maps in the present paper were rendered using QuickSurf mode in VMD (Humphrey et al., 1996) with a linear red-white-blue color scale (from high to low mutual information values). The pores and their symmetry mates were centered in the unit cell with image dimensions cropped to cell size. The 3D structures used for the rendering of the heat maps correspond to the last frame of the trajectory windows (see Section "2.3").

expected because the box dimensions probe for the size of the water tunnel, whereas the breaking activities probe for lipids that weaken the bilayer.

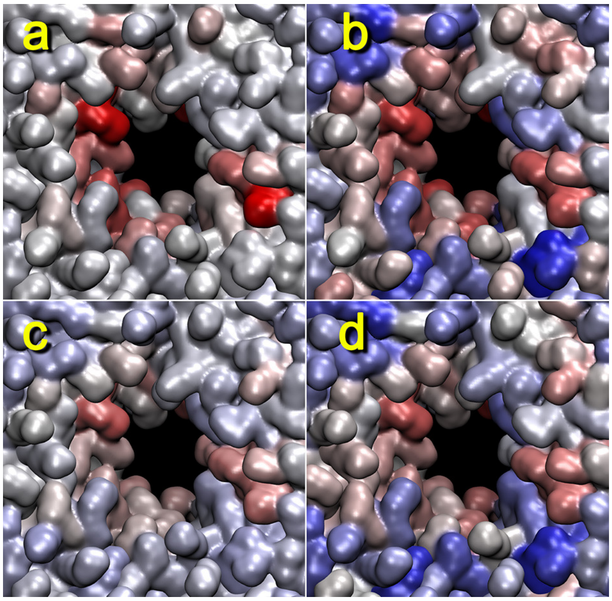


FIGURE 6 | Lipid pore formation heat map validation. The heat maps for trajectory 4 (+z direction view) generated with mutual information **(a,c)** or Pearson cross correlation **(b,d)** against activity data from the cutoff graph **(a,b)**, z dimension of the unit cell **(c)**, or x dimension of the unit cell **(d)** are shown. The parameters are otherwise the same as those in **Figure 5**. When considering the box dimensions, we note that the Pearson correlation expects (positively) co-correlated features (see text). Therefore, we have used the x dimension of the unit cell (identical to the y dimension); on the other hand, in the mutual information case, this distinction did not matter, and we used the z dimension for the analysis.

3.4. Evaluation of Parameter Space

As discussed earlier (**Figure 2**), the user must select several program parameters for the analysis, and it is worthwhile to investigate how sensitive the results are to such subjective choices.

We have used the box dimension as a benchmark for estimating the proper contact graph smoothing parameter in **Figure 7**. Although the lipid-breaking activity shows slightly different results, the overall appearance should be comparable. In **Figure 7**, we used three smoothing parameters: 20, 50, and 100 ps. All three cases had highlighted lipids at the pore; however, the smoothing parameter of 50 ps gave the closest match with the box dimension heat map, and that of 100 ps was a close second. This is in good agreement with our earlier rule of thumb, namely, to start the analysis with a smoothing of approximately 5% of the window width.

Figure 8 shows the dependence of the heat map on the distance cutoff values for the contact graph. As a rule of thumb, the cutoff values should reflect the nearest-neighbor distances in the coarse grained model (two values bracket a buffer zone for recrossing suppression, and the lower value is most important whereas the upper value is typically set 1–2 Å higher). For lipids, a visual inspection of the lipid phosphorus atoms suggests that a lower cutoff of 11 Å would be appropriate (**Figure 1**). **Figure 8** shows heat maps for cutoff values of 9–15 Å (with a 2 Å buffer).

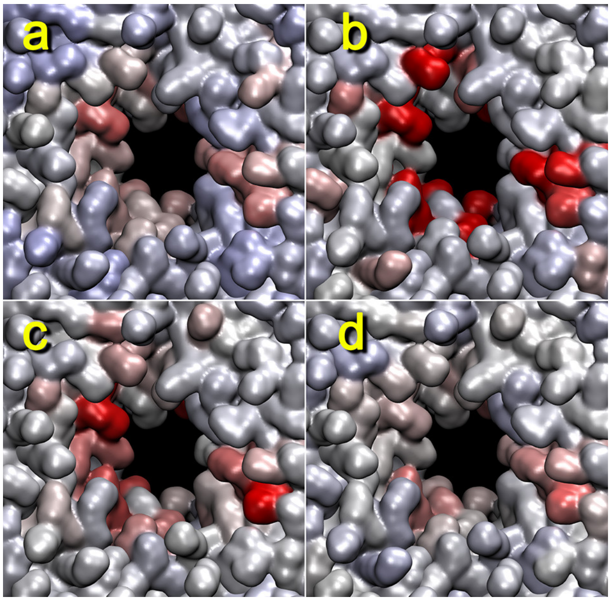


FIGURE 7 | Dependency of lipid heat maps on temporal smoothing parameter used in `terrain.py` and `tagging.py`. The heat maps for trajectory 4 (+z direction view) generated with mutual information against activity data from z dimension of the unit cell **(a)**, or the cutoff graph **(b–d)** with smoothing parameters 20 **(b)**, 50 **(c)**, and 100 ps **(d)** are shown. The parameters are otherwise the same as those in **Figure 5**.

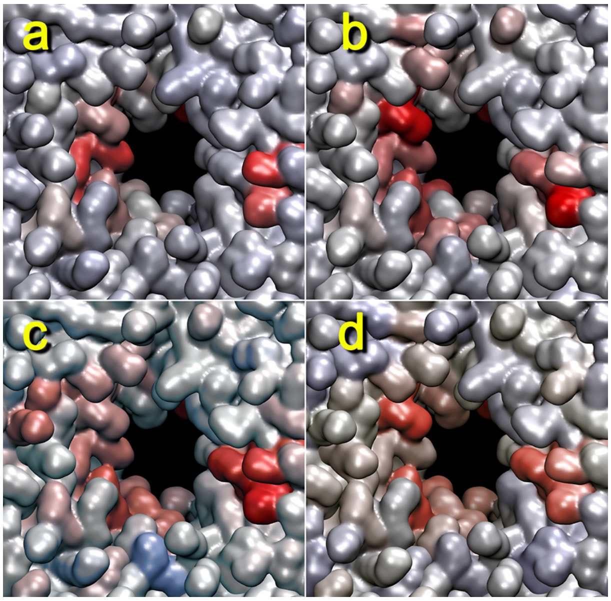


FIGURE 8 | Dependency of lipid heat maps on graph distance cutoff used in `terrain.py`. The heat maps for trajectory 4 (+z direction view) generated with mutual information against the cutoff graph with buffers of 9–11 **(a)**, 11–13 **(b)**, 13–15 **(c)**, and 15–17 Å **(d)** are shown. The parameters are otherwise the same as those in **Figure 2**.

The 9 Å value misses many of the lipid phosphorus contacts; however, results above 11 Å appear reasonably stable. Therefore, we used 11 Å for most of the analyses in this study.

3.5. Solvent Pore Formation Heat Maps

We also investigated whether we see any evidence that the global lipid dynamics, as described by the activity function, drives the solvent dynamics. Toward this end, we also generated a heat map for solvent molecules that were coarse-grained to one atom per molecule. Because we project the heat map on a moving canvas of rapidly diffusing solvent molecules, **Figure 9** shows the results for trajectory 1 as a function of time. **Figure 9** reveals both a temporal focusing of the solvent heat map on the pore formation time and a spatial focusing on the membrane-solvent interface (although there is no preferential association of the heat map with the emerging pore). The highlighted solvent molecules are clearly located at the lipid-solvent interface at 22.0 and 22.3 ns; however, before and after pore formation (20.0 and 22.6 ns) they are dispersed throughout the membrane by their diffusive motion. The spatio-temporal focusing shows the importance of the general weakening of interactions that precedes (or precipitates) the solvent perforation of the lipid bilayer (**Figures 3**, **4**). Inspecting the heat map further shows that the identity and origin of the tunneling water molecules in the pore is not determined by their position before the intrusion occurs; therefore, at 22.0 and 22.3 ns there is no accumulation of heat-map highlighted waters in the pore beyond the level that is generally observed at the interface. The dispersed heat map at 20.0 ns shows that water molecules in the pore can be from the interface or from the bulk. Sometimes, interfacial water can be seen climbing past the interface region, and then, ballistic water sails in from the bulk and replaces it.

Water and ions have one disadvantage, namely, that any coarse-graining on a per-molecule basis to compute the time-dependent distance geometry $X_{i,j}(t)$ is rather limited. For approximately 12,000 solvent molecules, we were able to achieve manageable analysis times (of the order of days) by using a 15 Å distance cutoff for significant interactions (as an argument passed to `tagging.py`) to reduce the number of pairwise water interactions.

4. CONCLUSIONS

Over the last several years, we have developed a statistical strategy for transforming MD simulation time series into spatial heat maps. The original purpose of this approach was to detect allosteric communication patterns in proteins, such as hinge bending and amino acid contact-forming and -breaking during folding and unfolding. Although this approach worked well for this purpose (Kovacs and Wriggers, 2016), one obvious limitation that impeded wider adoption was the exclusive focus on proteins. In this work, we have applied the algorithms for the first time to lipid and solvent interaction networks. This was motivated by our interest in the mechanism of the electroporation of cell membranes. This work has also prompted us to develop a faster and more robust mutual information solver that is described in the accompanying paper. Other generalizations of our heat map approach, such as to nucleic acids and mixed protein-membrane systems, are the subject of future work.

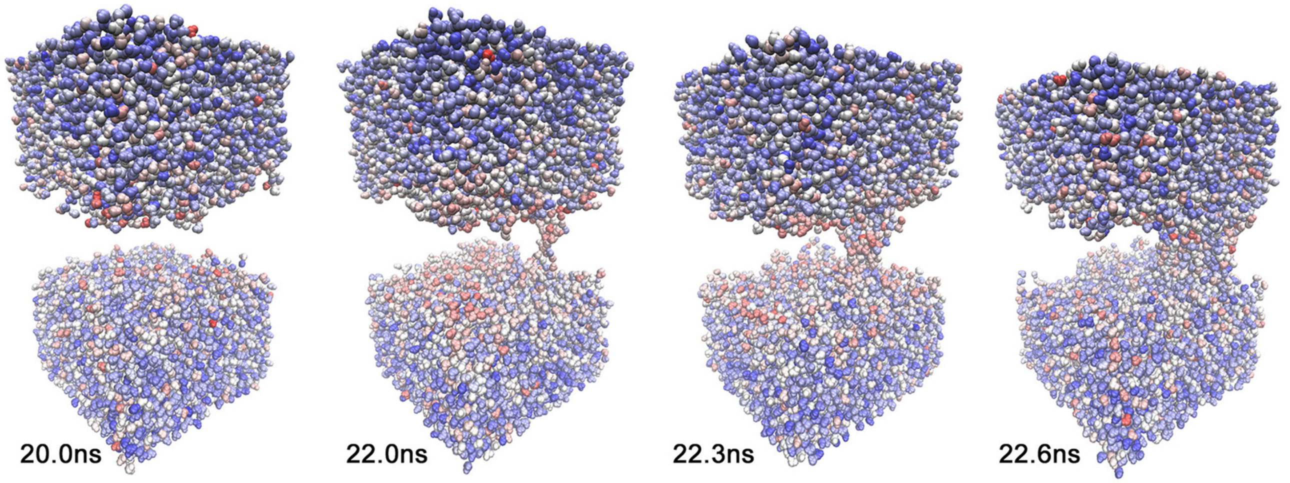


FIGURE 9 | Solvent heat map of trajectory 1 generated with mutual information against activity data from the lipid cutoff graph with smoothing parameter 200 ps. The lipids are omitted for clarity. The snapshots were taken at the same simulation times as in **Figure 1**, $t = 20.0$, 22.0, 22.3, and 22.6 ns. The solvent heat map was rendered using VDW mode in VMD (Humphrey et al., 1996) with a linear red-white-blue color scale (from high to low values).

The generalization of our numerical algorithm to aqueous solvents has revealed one limitation of our contact graph approach, namely, the quadratic scaling of the coarse-grained interaction network. This was not a problem for small proteins in the past work or for the 128 lipids in this study; however, the water molecules require at least one representative atom and cannot be coarse-grained further. (Force fields that group several water atoms together, such as the Martini force field, do exist, but the adequate modeling of intruding water fingers during electroporation requires full atomic detail). The mapping of pairwise interactions $R_{X,a}(i,j)$ before linear compression is the main performance bottleneck. We note that our choice of relative distance geometry $X_{i,j}(t)$ is rooted in the lack of a fixed reference in the liquid systems considered in this study. In the future, it would be desirable to find a reference-free but linearly scaling equivalent $X_i(t)$ that is suitable for statistical comparison with the activity function.

In the application to electroporation simulations, our heat map approach highlighted a small number of important lipids. This allowed us to efficiently search the trajectories for any mechanisms or patterns. While this is ongoing research and the causality remains unclear, the preliminary results obtained thus far suggest that pore lining lipids do not actively cause pore formation; instead, they rather passively follow the water perforation, which occurs first, as was proposed earlier by Tokman et al. (2013). This interpretation agrees with the result of our solvent heat map that showed only nonspecific interactions at the solvent-lipid interface but no sign of lipids driving the solvent at the pore location. Past efforts to identify a driving mechanism have always led to initial intruding water fingers—the phospholipids fall down their potential energy hill into the membrane interior *after* the water molecules (Vernier and Ziegler, 2007; Ziegler and Vernier, 2008; Kohler et al., 2015).

Even if the water molecules (but not lipids) play a driving role, a useful signature for pore initiation (nucleation) could exist among the lipid molecules. Perhaps promisingly, we found several lipids that were not associated with the growing pore but that indirectly contributed before (and during) pore formation through a weakening of the bilayer. Because the lipids are interchangeable and the results differ between trajectories, these "supporting events" seemingly occurred at random. However, our statistics could be limited by our choice of global activity functions, and more localized activity functions (that would reflect more specific degrees of freedom) could reveal a hidden nucleation mechanism of poration that has has eluded us thus far.

In summary, the proposed methodology provides new analyses for electroporation studies by transforming the temporal time series of simulations into spatial features. Additional future practical applications of this framework could include protein-lipid systems and studies of the effect of lipid- and water-soluble agents, in which allosteric mechanisms could be directly visualized on the embedded structures, as was the case in earlier applications to protein folding and hinge detection (Kovacs and Wriggers, 2016). All tools developed for this study will be documented and released in version 1.5 of the *TimeScapes* package that is freely available on our web site, http://timescapes.biomachina.org.

AUTHOR CONTRIBUTIONS

The *TimeScapes* analysis algorithms were designed, implemented, and applied by WW. The molecular dynamics trajectories were generated by FC. The mathematical theory of information matching (described in more detail in the accompanying paper) was created by JK. The experimental context and interpretation of this work were supervised by PV. The paper was written by WW, FC, and PV.

FUNDING

WW, FC, and JK were supported in part by the Frank Batten endowment to WW and by the National Institutes of Health grant R01GM62968. PV and FC were supported in part by the Air Force Office of Scientific Research (FA9550-15-1-0517 and FA9550-14-1-0123).

REFERENCES

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684–3690.

Böckmann, R. A., de Groot, B. L., Kakorin, S., Neumann, E., and Grubmüller, H. (2008). Kinetics, statistics, and energetics of lipid membrane electroporation studied by Molecular Dynamics simulations. *Biophys. J.* 95, 1837–1850. doi: 10.1529/biophysj.108.129437

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101. doi: 10.1063/1.2408420

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* 103, 8577–8593.

Ghahremanpour, M. M., Arab, S. S., Aghazadeh, S. B., Zhang, J., and van der Spoel, D. (2014). MemBuilder: a web-based graphical interface to build heterogeneously mixed membrane bilayers for the GROMACS biomolecular simulation program. *Bioinformatics* 30, 439–441. doi: 10.1093/bioinformatics/btt680

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graphics* 14, 33–38.

Kohagen, M., Mason, P. E., and Jungwirth, P. (2014). Accurate description of calcium solvation in concentrated aqueous solutions. *J. Phys. Chem. B* 118, 7902–7909. doi: 10.1021/jp5005693

Kohler, S., Levine, Z. A., García-Fernández, M. Á., Ho, M.-C., Vernier, P. T., Leveque, P., et al. (2015). Electrical analysis of cell membrane poration by an intense nanosecond pulsed electric field using an atomistic-to-continuum method. *IEEE Trans. Microwave Theory Techn.* 63, 2032–2040. doi: 10.1109/TMTT.2015.2418764

Kovacs, J. A., and Wriggers, W. (2016). Spatial heat maps from fast information matching of fast and slow degrees of freedom: application to molecular dynamics simulations. *J. Phys. Chem. B* 120, 8473–8484. doi: 10.1021/acs.jpcb.6b02136

Kovacs, J. A., Helmick, C., and Wriggers, W. (2017). A balanced approach to adaptive probability density estimation. *Front. Mol. Biosci.* 4:25. doi: 10.3389/fmolb.2017.00025

Miyamoto, S., and Kollman, P. A. (1992). SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* 13, 952–962.

Neumann, E., Schaefer-Ridder, M., Wang, Y., and Hofschneider, P. H. (1982). Gene transfer into mouse lyoma cells by electroporation and electroporative delivery of drugs and genes. *EMBO J.* 1, 841–845.

Old Dominion University (2017). *High Performance Computing*. Available online at: http://www.odu.edu/hpc

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., et al. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science* 330, 341–346. doi: 10.1126/science.1187409

Tokman, M., Lee, J. H., Levine, Z. A., Ho, M.-C., Colvin, M. E., and Vernier, P. T. (2013). Electric field-driven water dipoles: Nanoscale architecture of electroporation. *PLoS ONE* 8:e61111. doi: 10.1371/journal.pone.0061111

van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi: 10.1002/jcc.20291

Vernier, P. T., and Ziegler, M. J. (2007). Nanosecond field alignment of head group and water dipoles in electroporating phospholipid bilayers. *J. Phys. Chem. B* 111, 12993–12996. doi: 10.1021/jp077148q

Vernier, P. T., Ziegler, M. J., and Dimova, R. (2009). Calcium binding and head group dipole angle in phosphatidylserine-phosphatidylcholine bilayers. *Langmuir* 25, 1020–1027. doi: 10.1021/la8025057

Wriggers, W., Stafford, K. A., Shan, Y., Piana, S., Maragakis, P., Lindorff-Larsen, K., et al. (2009). Automated event detection and activity monitoring in long molecular dynamics simulations. *J. Chem. Theory Comput.* 5, 2595–2605. doi: 10.1021/ct900229u

Ziegler, M. J., and Vernier, P. T. (2008). Interface water dynamics and porating electric fields for phospholipid bilayers. *J. Phys. Chem. B* 112, 13588–13596. doi: 10.1021/jp8027726

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 25 April 2017
doi: 10.3389/fmolb.2017.00025

A Balanced Approach to Adaptive Probability Density Estimation

Julio A. Kovacs, Cailee Helmick and Willy Wriggers*

Department of Mechanical and Aerospace Engineering, Old Dominion University, Norfolk, VA, USA

Our development of a Fast (Mutual) Information Matching (FIM) of molecular dynamics time series data led us to the general problem of how to accurately estimate the probability density function of a random variable, especially in cases of very uneven samples. Here, we propose a novel Balanced Adaptive Density Estimation (BADE) method that effectively optimizes the amount of smoothing at each point. To do this, BADE relies on an efficient nearest-neighbor search which results in good scaling for large data sizes. Our tests on simulated data show that BADE exhibits equal or better accuracy than existing methods, and visual tests on univariate and bivariate experimental data show that the results are also aesthetically pleasing. This is due in part to the use of a visual criterion for setting the smoothing level of the density estimate. Our results suggest that BADE offers an attractive new take on the fundamental density estimation problem in statistics. We have applied it on molecular dynamics simulations of membrane pore formation. We also expect BADE to be generally useful for low-dimensional applications in other statistical application domains such as bioinformatics, signal processing and econometrics.

Keywords: adaptive density estimation, covariance ellipsoid, covariance smoothing, optimal number of nearest neighbors, R*-tree, visual criterion

OPEN ACCESS

Edited by:
Slavica Jonic,
IMPMC, Sorbonne Universités - CNRS UMR 7590, UPMC Univ Paris 6, MNHN, IRD UMR 206, France

Reviewed by:
Carlos Oscar Sanchez Sorzano,
National Center of Biotechnology (CSIC), Spain
Ion Grama,
University of Southern Brittany, France

****Correspondence:***
Julio A. Kovacs
jkovacs@odu.edu

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *30 November 2016*
Accepted: *07 April 2017*
Published: *25 April 2017*

Citation:
Kovacs JA, Helmick C and Wriggers W (2017) A Balanced Approach to Adaptive Probability Density Estimation. Front. Mol. Biosci. 4:25. doi: 10.3389/fmolb.2017.00025

1. INTRODUCTION

One of the most popular non-parametric density estimation methods is *kernel density estimation* (KDE), whereby the density is estimated by means of a sum of kernel functions centered at the sample points (Silverman, 1986; Wand and Jones, 1995):

$$\hat{f}(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M} K_H(\mathbf{x}-\mathbf{x}_j), \tag{1}$$

where $K_H(\mathbf{x}) = \det(H)^{-1/2}K(H^{-1/2}\cdot\mathbf{x})$, $K:\mathbb{R}^d \to \mathbb{R}$ being the d-variate kernel and M the data size. One of the most commonly used kernels is the Gaussian: $K(\mathbf{x}) = C_d\exp(-\|\mathbf{x}\|^2/2)$, with C_d a normalizing constant that depends on the dimension d. The $d \times d$ matrix H, called the *bandwidth matrix,* could either be fixed, or it could depend on the sample point $\mathbf{x}_j$ ("sample point estimator") or on the test point $\mathbf{x}$ ("balloon-type estimator").

Originally, we adopted a fixed-bandwidth KDE approach in our recent application to Fast (Mutual) Information Matching (FIM) of molecular dynamics time series data (Kovacs and Wriggers, 2016). The fixed-bandwidth approach is well mature and there exist a wide range of methods for bandwidth selection (see e.g., Jones et al., 1996 for a survey). Among these, the method by Sheather and Jones (1991) could be regarded as the *de facto* standard in the univariate case (Jones et al., 1996). Our application to molecular dynamics time series relies on non-negative activity functions (Kovacs and Wriggers, 2016). As discussed in more detail by Wriggers et al. (2017), the graph-based activity functions we typically use are zero during quiescent time periods of the simulation, leading to an uneven distribution of activity values with a strong peak at zero that is not amenable to fixed-bandwidth KDE approaches. In protein simulations we have therefore recommended to use a rms-fluctuation-based activity that gives a more even histogram (Kovacs and Wriggers, 2016). Unfortunately this is not an option for the membrane simulations in the accompanying paper (Wriggers et al., 2017), so we require a variable-bandwidth approach that can handle graph-based activity functions in that application.

The situation in regard to the variable-bandwidth KDE methods is less well developed. In fact, it has not been easy to make significant performance improvements by allowing the bandwidth to vary from point to point (Farmen and Marron, 1999). Several approaches have been proposed, with varying degrees of success across different types of data sets. One of the earliest approaches was that of Breiman, Meisel and Purcell, who used bandwidths proportional to the distance from each sample point to its *k*th nearest neighbor (Breiman et al., 1977). So, for dimensions $d > 1$ the j-dependent bandwidth matrices are scalar (i.e., multiples of the identity matrix). Later, Abramson (1982) proposed a square-root law, whereby the bandwidth at each point is taken to be inversely proportional to the square root of the density. Since the actual density is not known, a "pilot density" is needed, which is usually computed using a fixed-bandwidth method. Like the previous approach, in $d > 1$ it produces bandwidth matrices that are scalar.

One of the earliest alternative approaches to improve the performance of variable bandwidth estimators was proposed by Sain and Scott (1996): the binned kernel estimator, in which the support of the density is divided in m equal parts. Each of these subintervals yields a value of the bandwidth, which is then used for the kernels centered at points belonging to the corresponding subinterval. This method was extended to the multivariate setting by Sain (2002). Hazelton (2003) refined this approach (in the univariate case) by using cubic splines instead of piecewise-constant functions to model the bandwidths, showing improvements in the quality of the density estimates. However, these approaches are very slow, as they involve an optimization problem over many variables. Brewer (2000) showed improved results relative to Sain and Scott (1996) by using a Bayesian approach based on likelihood cross-validation, which works specially well for small sample sizes, and adds a local smoothing step to enhance the visual appeal of the density estimates. This method was extended by Zougab et al. (2014) to the multivariate case, in which the bandwidth matrices are not restricted to being diagonal. Like Brewer's approach, it works very well for small sample sizes, but the complexity scales quadratically with the sample size.

Attempts at alleviating the mentioned limitations include a class of methods that use convex combinations (i.e., linear combinations with non-negative coefficients adding up to 1) or mixtures of densities of certain types. Vapnik and Mukherjee (2000) used a mixture of Gaussian densities in which the coefficients are optimized by matching the sample's cumulative distribution function (CDF) with the CDF estimator. The Gaussian densities are isotropic (i.e., having scalar covariance matrices). Song et al. (2008) assume the density to be a convex combination of several prototype densities, and optimizes the coefficients by matching the mean estimators. The prototype densities are Gaussians with diagonal covariance matrices. Ganti and Gray (2011) proposed a density estimator in which the kernel functions are convex combinations of isotropic Gaussians of various widths. The expected outcome is that this would produce a richer set of function shapes which would compensate the limitation arising from using isotropic Gaussians. However, the quality of the resulting density estimates (judged by visual inspection) is questionable.

Several other interesting ideas have also been put forward. For instance, Katkovnik and Shmulevich (2002) described a univariate balloon-type estimator based on the "intersection of confidence intervals" (ICI) rule (i.e., shrinking sequences of intervals), for which, at each test point x, a fixed, arbitrary sequence of increasing bandwidth values is scanned until the ICI criterion is met, yielding the bandwidth for that point. Wu et al. (2007) used a cluster analysis of the set of nearest neighbors to derive the bandwidths at each sample point. The analysis is restricted to isotropic (scalar) bandwidth matrices. Shimazaki and Shinomoto (2010) used a "local MISE" criterion, which includes a window factor in the integral defining the mean squared integrated error (MISE) to derive local bandwidths in the univariate case. The "Rodeo" approach (Liu et al., 2007) is specially suited for high-dimensional data. The density is assumed to be the product of a non-parametric factor and a parametric one, which is known either completely or up to finitely many parameters. Bandwidth matrices are restricted to being diagonal, and a sparsity condition has to be imposed for the problem to be tractable.

Motivated by the various limitations of previous methods, here we propose a novel approach, which we call "BADE" (for Balanced Adaptive Density Estimation) that offers several desirable features: good scaling for large data sizes (sublinear complexity in M for $d = 1$ and 2); not restricted to diagonal bandwidth matrices; free of data-dependent parameters (the user does not need to make any choices). In fact, we are no longer dealing with bandwidth matrices *per se*, although there is a connection with kernel estimation through the "effective" number of neighbors (Section 2.1).

2. BALANCED ADAPTIVE DENSITY ESTIMATION

Let $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^d$ be a sample of size M drawn independently from an unknown d-dimensional distribution having probability density function $f : \mathbb{R}^d \to \mathbb{R}$. Let $\Sigma_\mathbf{P}$ be the covariance matrix of $\mathbf{P}$, which we use as an estimate of the covariance matrix of the true distribution. This matrix will be used later to give us an idea of the global size and shape of the whole sample.

Unlike most of the previous approaches, we do not use a kernel-based estimation approach. Instead, the basic idea is the following: for each probe point $\mathbf{x} \in \mathbb{R}^d$ where we want to estimate the density, we determine the set $N_k(\mathbf{x})$ of its k nearest neighbors among $\mathbf{P}$, and compute its covariance matrix:

$$\Sigma_k(\mathbf{x}) = \mathrm{Cov}(N_k(\mathbf{x})), \tag{2}$$

which gives us a basic description of the size and shape of the set of sample points near $\mathbf{x}$, by means of the "covariance ellipsoid" (or "inertia ellipsoid") defined by the eigenvectors and eigenvalues of this matrix. The volume (modulo a constant) of that ellipsoid is $V_k(\mathbf{x}) = \sqrt{\det \Sigma_k(\mathbf{x})}$. (Recall that the determinant is the product of the eigenvalues, which are the squares of the corresponding principal axes of the ellipsoid.) Then, our first version for the density estimate at $\mathbf{x}$ is:

$$\hat{f}(\mathbf{x}) = C \cdot \frac{k}{M\, V_k(\mathbf{x})}, \tag{3}$$

where C is a scaling constant. In practice, the density estimate is computed, omitting C, on a grid covering the sample $\mathbf{P}$, and then C is determined so that the integral of $\hat{f}$ is 1.

Of course, this expression is very reminiscent of the original proposal of Loftsgaarden and Quesenberry (1965), just with the volume of the ellipsoid in place of the volume of the sphere of radius equal to the distance from $\mathbf{x}$ to the kth nearest sample point. It is well known that Loftsgaarden and Quesenberry's method produce heavy tails and spiky density estimates (Silverman, 1986). The spikiness is due to the use of the simple kth nearest neighbor, which is highly variable. The use of $V_k(\mathbf{x})$ drastically decreases this variability, since this ellipsoid—being the covariance ellipsoid of a set of k neighbors— is much more stable than a domain (whether spherical or ellipsoidal) whose size is based simply on the distance to the kth neighbor. Note that this ellipsoid does not, in general, contain the set of neighbors on which it is based.

2.1. Fixing Heavy Tails

The basic idea described above still suffers from a number of drawbacks. First, as with the method of Loftsgaarden and Quesenberry (1965), this basic idea produces heavy tails—since the set of nearest neighbors remains virtually constant as the point $\mathbf{x}$ moves away (and exactly constant in the univariate case). This can be remedied by introducing a decay factor, giving an "effective" k:

$$k_e(\mathbf{x}) = k \cdot \exp\Big[-\frac{1}{2}(\mathbf{x} - \mu_k(\mathbf{x})) \cdot \Sigma_k(\mathbf{x})^{-1} \cdot (\mathbf{x} - \mu_k(\mathbf{x}))^T\Big], \tag{4}$$

where $\mu_k(\mathbf{x}) = \mathrm{Mean}(N_k(\mathbf{x}))$. This factor follows a decay rate corresponding to the distribution of the k nearest neighbors of $\mathbf{x}$, and is useful in the "interior" of the set $\mathbf{P}$ as well as the "exterior." Thus, our second version for the density estimate is

$$\hat{f}(\mathbf{x}) = C \cdot \frac{k_e(\mathbf{x})}{M\, V_k(\mathbf{x})}. \tag{5}$$

We note that due to the exponential decay of $k_e(\mathbf{x})$, this estimator is integrable. Incidentally, we can write Equation (5) as follows (using the kernel notation as in Equation 1):

$$\hat{f}(\mathbf{x}) = \frac{C}{M} \cdot k \cdot K_{\Sigma_k(\mathbf{x})}(\mathbf{x} - \mu_k(\mathbf{x})) \approx \frac{C}{M} \cdot \sum_{l=1}^{k} K_{\Sigma_k(\mathbf{x})}(\mathbf{x} - \mathbf{p}_{r_l}), \tag{6}$$

where $\{\mathbf{p}_{r_l} \mid 1 \le l \le k\} = N_k(\mathbf{x})$. Thus, the estimator given by Equation (5) is approximately like a balloon-type Gaussian kernel estimator, but based only on the k nearest neighbors of the probe point $\mathbf{x}$, instead of all the sample points.

2.2. Determination of *k*

A second drawback of our basic idea is: what should k be? Loftsgaarden and Quesenberry (1965) take it as independent of $\mathbf{x}$, depending only on the sample size M. We can improve on this by choosing k in such a way that the volume $V_k(\mathbf{x})$ of the covariance ellipsoid be a certain function of $f(\mathbf{x})$. Two common choices, in a sense antipodal to each other, are:

1. Volume = const. This would yield a k that is approximately proportional to the density.
2. Volume = const/f. This would yield a k that is approximately constant.

We found that neither of these extremes produces good density estimates: a constant volume is essentially like a histogram: it will not resolve sharp enough peaks, and will yield zero in regions where the sample points are widely spread; a constant k will tend to be too large in region of low density, and too small in regions of high density.

However, the geometric mean of both offers a good compromise: Volume = const/$\sqrt{f}$. (This is why we named our approach "balanced.") Hence, we have the equation

$$V_k(\mathbf{x}) = \mathrm{const}/\sqrt{f(\mathbf{x})}. \tag{7}$$

For $f(\mathbf{x})$ we can use, in this equation, the estimate $f(\mathbf{x}) \approx C_1 k / V_k(\mathbf{x})$. This allows us to solve the equation for $V_k(\mathbf{x})$:

$$V_k(\mathbf{x}) = \frac{C_2}{k}, \tag{8}$$

where C_2 (which depends on M but not on $\mathbf{x}$) subsumes the various constants. **Figure 1** depicts the situation graphically: when k is small, the left-hand side of the equation (i.e., $V_k(\mathbf{x})$) is small, while the right-hand side (C_2/k) is large, and vice versa. The point where the two curves cross gives the optimal k for this $\mathbf{x}$. Solving this equation is easy: keep increasing k by 1 until the inequality

$$V_k(\mathbf{x}) \cdot k < C_2 \tag{9}$$

no longer holds. This can be efficiently implemented in code by means of incremental nearest neighbor methods (Hjaltason and Samet, 1999), in which the cost of retrieving each additional neighbor is essentially $O(1)$ (see "Complexity," below).

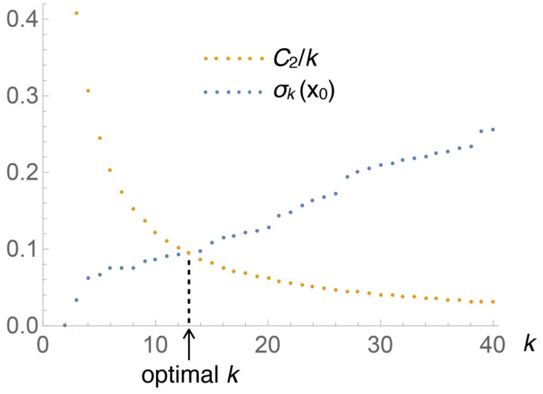


FIGURE 1 | Graphical illustration of the determination of the optimal number of nearest neighbors, k, to be used at each test point x (this particular plot corresponds to the univariate Old Faithful data set, at $x_0 = 3.59$). The intersection of both curves provides the solution to Equation (8).

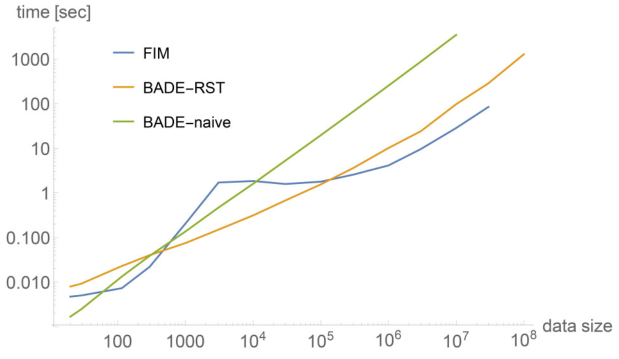


FIGURE 2 | Comparison between timings obtained by three methods applied to 2-dimensional data sets. The data sets were artificially generated to simulate a bimodal distribution, shown in **Figure 3A** for $M = 1,000$. See the section on Complexity for more details.

2.3. Determination of C_2

The constant C_2 in Equation (8) depends on M, d, and f. As pointed out in the introduction, one of our goals was to devise a method that does not depend on parameters that have to be adjusted for each particular data set. Our method satisfies this condition (as we'll describe in a moment) except for the dependence on the dimension d, which has to be worked out for each d. (We worked out the values for $d = 1$ and 2 since these are the ones that most interest us for our applications.)

In $d > 1$ the data needs to be rescaled so that each coordinate have unit variance. This is important not only for the derivation of the expression of the constants, but also for the correct functioning of the nearest neighbor search: if the rescaling were not done, then the neighbor search would be as if using ellipsoids instead of spheres for its distance queries. (We preferred this transformation rather than sphering—making the covariance matrix the identity—since the latter changes the correlations, being a skewed transformation.) In these conditions, we factor the volume of the covariance ellipsoid of the whole sample (square root expression in this equation) out of C_2:

$$C_2 = H_0 \sqrt{\det \Sigma_{\mathbf{P}}}. \tag{10}$$

It turns out that H_0 does not depend on f, but only on M and d. We verified this by means of a "visual criterion." The reason we chose this type of criterion, instead of a more objective one such as MISE, is that good MISE performance does not guarantee visually appealing density estimates (Farmen and Marron, 1999), which is one of our goals. One such visual criterion was proposed by Marron and Tsybakov (1995), in which the distance between the graphs of both functions is evaluated, instead of the vertical distance. Here we needed a different type of criterion to determine H_0: we ourselves examined by eye the density estimates resulting from an array of values of M and H_0, for various simulated densities. For each M and density, we looked for the minimum value of H_0 that yielded a density estimate that did not look undersmoothed. Even though this visual criterion might seem rather *ad hoc,* it actually yielded surprisingly good linear relationships in log/log scale on the H_0/M plane. The fitted lines, which were independent of the particular density, correspond to the following power laws:

$$H_0 = \begin{cases} 0.028\, M^{4/5} & \text{for } d = 1, \\ 0.162\, M^{2/5} & \text{for } d = 2. \end{cases} \tag{11}$$

It is interesting to note that the expression for $d = 2$ is nearly equal to the square root of that for $d = 1$: $\sqrt{0.028\, M^{4/5}} = 0.167\, M^{2/5}$. This is reassuring and adds confidence to our visual criterion, in addition to providing an obvious conjecture about the expression for H_0 for $d > 2$ (which we haven't tested).

A more theoretical justification of the expression for H_0 would probably be related to how the human visual system processes information. One possible approach could be the addition of a regularization term that would emulate visual perception. An intriguing link to the standard MISE theory in kernel density estimation is that the optimal bandwidth, in the 1-dimensional case, is proportional to $M^{-1/5}$, which is H_0/M (Wand and Jones, 1995).

We emphasize that this visual criterion was used only as a premise to determine the optimal dependence (on M) of the coefficient C_2. This optimal dependence is determined once and for all—the user does not need make any choices. However, the user could, with discretion, vary the coefficients in the formula for H_0 (Equation 11), to obtain density estimates with greater or lesser amount of smoothing than that provided by the values in Equation (11). As a rule of thumb, our visual tests (not shown) suggest to keep the variation within a factor 2 from the stated values.

2.4. Covariance Smoothing

To further improve the visual appeal of the density estimate given by Equation (5), we added an optional smoothing step to our method. The smoothing procedure was inspired by that of Brewer (2000), who averages the inverse variances of two neighboring sample points on either side of each sample point,

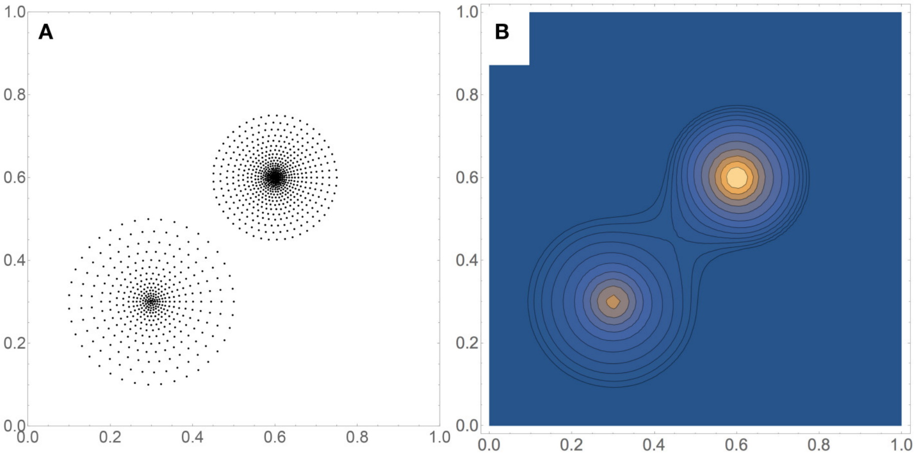


FIGURE 3 | (A) Example of a data set used for the timing comparison shown in **Figure 2**. Here $M = 1,000$. **(B)** Corresponding contour plot of the density estimate, which includes the covariance-smoothing step.

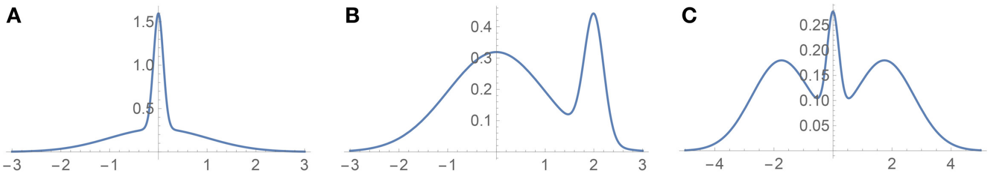


FIGURE 4 | Test densities (Gaussian mixtures) used in our univariate simulation study. (A) H3 (Kurtotic unimodal); **(B)** H4 (Skewed bimodal); **(C)** H5 (Trimodal).

producing estimates that are relatively free from unnecessary minor fluctuations. Since our approach is grid-based, we need a more sophisticated procedure, as averaging inverse variances of neighboring grid points would not be correct, since the spacing is arbitrary. We need to weight the contributions of all the grid points according to their respective covariance matrices and locations relative to the test point. Denoting the grid points by $\mathbf{x}_j$ ($j = 1, \ldots, G$), where G is the size of the grid, and the corresponding covariance matrices (Equation 2) by $\Sigma(\mathbf{x}_j)$ (omitting for clarity the subindex that indicated the number of nearest neighbors used), we define the smoothed precision matrices by:

$$\hat{\Sigma}_i^{-1} = \frac{\sum_{j=1}^{G} w_{i,j}\, \Sigma(\mathbf{x}_j)^{-1}}{\sum_{j=1}^{G} w_{i,j}}, \tag{12}$$

where the weights (influence of point j on point i) are given by

$$w_{i,j} = \frac{1}{\sqrt{\det \Sigma(\mathbf{x}_j)}} \cdot \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j) \cdot \Sigma(\mathbf{x}_j)^{-1} \cdot (\mathbf{x}_i - \mathbf{x}_j)^T\right]. \tag{13}$$

Thus, the contribution of each covariance matrix is in accordance with the value of the Gaussian function defined by it, at each of the grid points. This equation shows that the smoothing can be considered local, in the sense that points $\mathbf{x}_j$ where $\Sigma(\mathbf{x}_j)$ is large (where the density is low) or which are far from $\mathbf{x}_i$ contribute little, and only points that are close to $\mathbf{x}_i$ and with a small $\Sigma(\mathbf{x}_j)$ will contribute significantly to $\hat{\Sigma}_i$. (Note: "large" or "small" applied to a matrix mean that the volume of its ellipsoid—or equivalently, its determinant, or the product of its eigenvalues—is large or small.)

Since both the smoothing step just described and the main step (Equation 5) are local, we see that our method does not suffer from the non-locality issues that affect, for instance, one version of Abramson's square-root method (basically, extreme tail sample points affect the density estimate elsewhere too much; see Terrell and Scott, 1992; Hall et al., 1995 for details).

Finally, we also need the smoothed version of the "effective" k values (Equation 4). They are computed similarly to Equation (12):

$$\hat{k}_{e,i} = \frac{\sum_{j=1}^{G} w_{i,j}\, k_{e,j}}{\sum_{j=1}^{G} w_{i,j}}. \tag{14}$$

Then, the smoothed version of the density estimate is given by

$$\hat{f}(\mathbf{x}_i) = C \cdot \frac{\hat{k}_{e,i}}{\sqrt{\det \hat{\Sigma}_i}}, \tag{15}$$

where the constant C is determined by the condition of $\hat{f}$ integrating to 1.

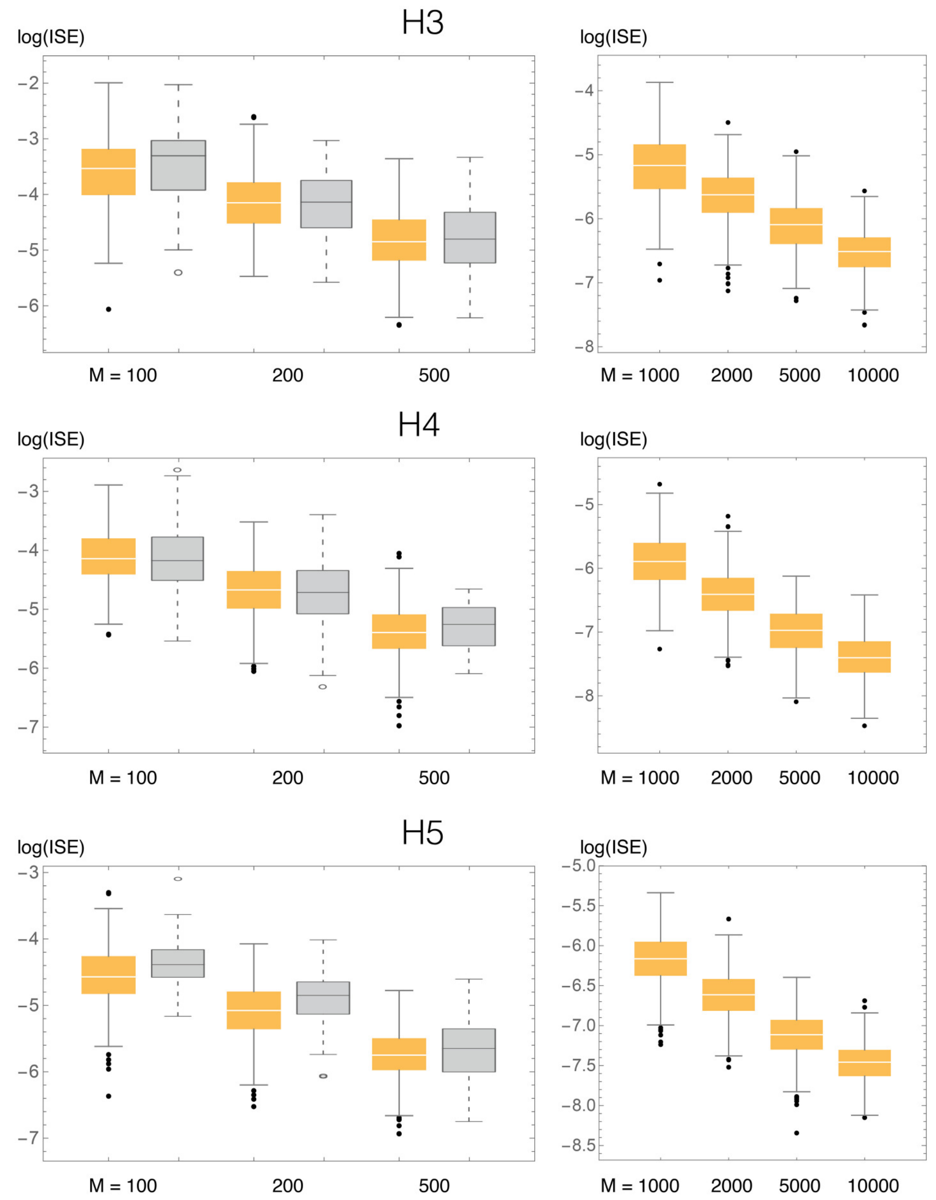


FIGURE 5 | ISE statistics for the three univariate test densities (H3, H4, H5), compaing our method (BADE, in orange) and VB3 (Hazelton, 2003, in gray). Also shown are results for larger values of the sample size *M*, not considered by Hazelton.

2.5. Complexity

We analyze separately the two steps of our method: the main estimator (Equation 5) and the (optional) covariance-smoothing step (Section 2.4).

The first, main step requires the incremental retrieval, for each test point **x**, of successive nearest neighbors. In two and higher dimensions, the R*-tree data structure (Beckmann et al., 1990) provides an effective means to implement such a retrieval procedure (Hjaltason and Samet, 1999). This procedure makes use of “priority queues” or heaps, one of its most efficient implementations being the *pairing heap* (Fredman et al., 1986), in which the cost of an insertion operation is $O(1)$. Using this implementation, the cost of finding k nearest neighbors among M data points in 2 dimensions turns out to be $O(k \log M)$ (Hjaltason and Samet, 1999). (The complexity analysis gets more complicated in higher dimensions; see Hjaltason and Samet (1999) for details. In dimension 1, determining the sequence of nearest neighbors is a simple logarithmic-time procedure which does not require the use of any special data structure.) The number k will vary from point to point, but always $k \leq M$, and so the per-point cost would be $\leq O(M \log M)$. However, this is not a typical situation, as the average k will usually be much less than M. If fact, more realistic estimates for the average k are of the order $O(M^{1/2})$ (Loftsgaarden and Quesenberry, 1965;

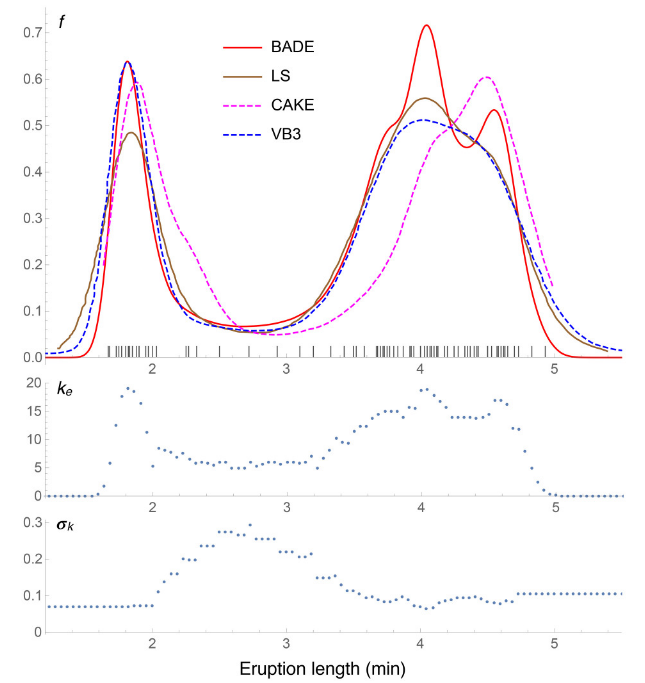


FIGURE 6 | Density estimates for the univariate Old Faithful data set, comparing our method (BADE) with LS (Brewer, 2000), CAKE (Ganti and Gray, 2011), and VB3 (Hazelton, 2003). The second and third rows show, respectively, the optimal effective number of nearest neighbors, and the standard deviation of the set of the k nearest neighbors, at each point of a regular grid of size 100. The marks on the x axis indicate the sample points.

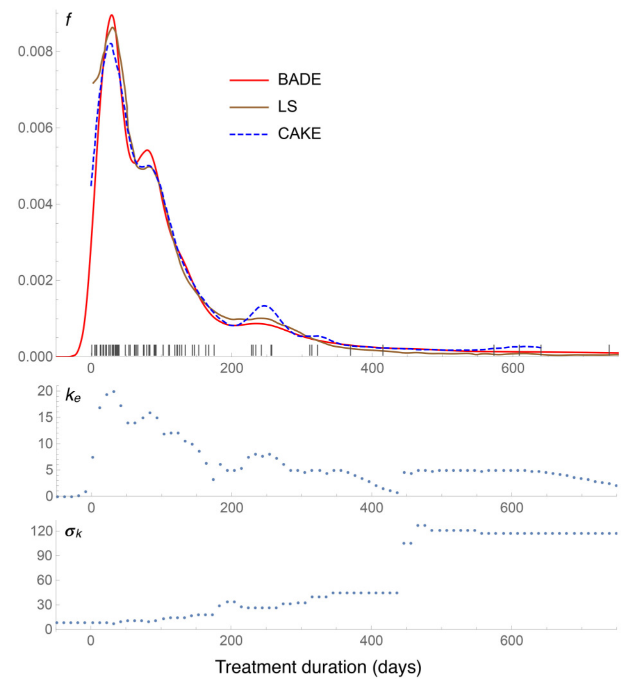


FIGURE 7 | Density estimates for the suicide data set, comparing our method (BADE) with LS (Brewer, 2000) and CAKE (Ganti and Gray, 2011). The second and third rows show, respectively, the optimal effective number of nearest neighbors, and the standard deviation of the set of the k nearest neighbors, at each point of a regular grid of size 100. The marks on the x axis indicate the sample points.

Silverman, 1986). Hence, the total cost of the main step can be approximated by

$$T_1 = \begin{cases} O(G \log M) & \text{for } d = 1, \\ O(GM^{1/2} \log M) & \text{for } d = 2. \end{cases} \tag{16}$$

where G is the size of the grid. We note that an incremental implementation of the nearest-neighbor search is essential to achieve this low complexity. Algorithms that are not incremental need to recompute the whole set of nearest neighbors each time one more is needed, with a significant deterioration in the efficiency.

Figure 2 shows a comparison between timings obtained by three methods applied to 2-dimensional data sets of a wide range of sizes, from $M = 20$ to 10^8. The data sets were artificially generated to simulate a bimodal distribution, shown in **Figure 3A** for $M = 1,000$, with the corresponding density estimate shown in **Figure 3B**. The three methods were: (a) FIM (using a fixed bandwidth) (Kovacs and Wriggers, 2016); (b) BADE-RST using the R*-tree to retrieve nearest neighbors; (c) BADE-naive using a naive way to retrieve nearest neighbors (i.e., by sorting the data points according to their distances to each probe point). We can see that FIM and BADE-RST have very similar asymptotics. In fact, FIM has a complexity of $O(M)$ (Kovacs and Wriggers, 2016), which is slightly worse than that of BADE-RST, although the constant is smaller for FIM. However, FIM only computes the mutual information, not the whole density function as BADE does, which introduces the factor G in Equation (16).

As for the second step (covariance smoothing), Equations (12) and (14) tell us that the cost will be

$$T_2 = O(G^2), \tag{17}$$

where the constant can be made quite small by summing each Gaussian function only over the ellipsoid where it has significant values (usually a small fraction of the total volume).

3. RESULTS

In order to evaluate the accuracy of BADE, we performed statistics of the ISE (Integrated Squared Error) for simulated samples taken from known distributions (**Figures 4**, **5** for the univariate case; **Figures 9**, **10** for the bivariate case). The ISE of an estimator $\hat{f}$ is defined as

$$\text{ISE} = \int \left(\hat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x} \approx \Delta x_1 \cdots \Delta x_d \cdot \sum_{j=1}^{G} \left(\hat{f}(\mathbf{x}_j) - f(\mathbf{x}_j)\right)^2. \tag{18}$$

Also, we considered some real data sets to compare the density estimates of BADE with those of previous methods (**Figures 6–8** for the univariate case; **Figures 11**, **12** for the bivariate case). The BADE results were computed using Equation (15), i.e., including the covariance-smoothing step.

3.1. Univariate Case

3.1.1. Simulated Examples

The three univariate simulated densities, all Gaussian mixtures, on which we tested our method are shown in **Figure 4**. They are densities 3, 4, and 5 used by Hazelton (2003), so we will refer to them in this paper as H3, H4, and H5:

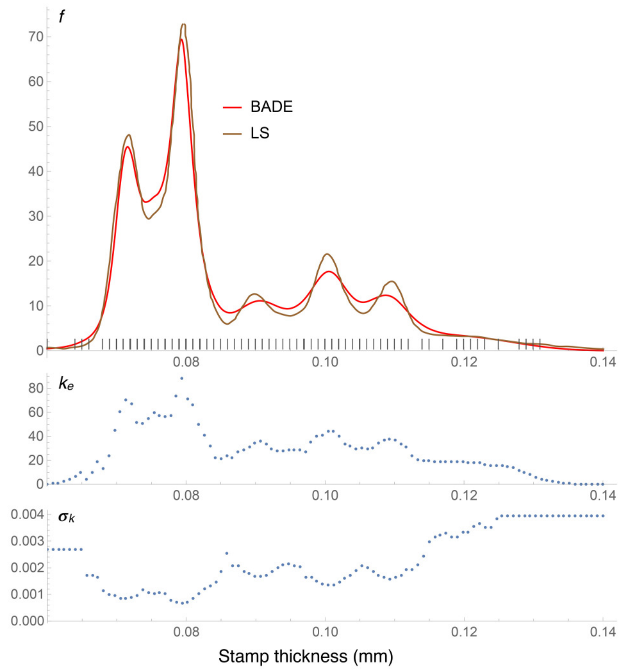


FIGURE 8 | Density estimates for the Hidalgo stamp data set, comparing our method (BADE) with LS (Brewer, 2000). The second and third rows show, respectively, the optimal effective number of nearest neighbors, and the standard deviation of the set of the *k* nearest neighbors, at each point of a regular grid of size 100. The marks on the *x* axis indicate the sample points; notice the equispacing due to rounding.

1. **H3** (Kurtotic unimodal, equal to density #4 in Marron and Wand, 1992): $\frac{2}{3}\mathrm{N}(0,1) + \frac{1}{3}\mathrm{N}(0,\frac{1}{10})$. (We denote the normal distribution with mean μ and standard deviation σ by $\mathrm{N}(\mu,\sigma)$.)
2. **H4** (Asymmetric bimodal, similar to density #8 in Marron and Wand, 1992): $\frac{4}{5}\mathrm{N}(0,1) + \frac{1}{5}\mathrm{N}(2,\frac{1}{5})$.
3. **H5** (Symmetric trimodal, similar to density #9 in Marron and Wand, 1992): $\frac{9}{20}\mathrm{N}(-\frac{7}{4},1)+\frac{9}{20}\mathrm{N}(\frac{7}{4},1)+\frac{1}{10}\mathrm{N}(0,\frac{1}{5})$. (Note the typo in Table 1 of Hazelton's paper in the equation for this density).

We compared the ISE statistics of our method, for each of the above three densities, with those of Hazelton (2003). They are displayed, in logarithmic scale, in **Figure 5**. We also considered larger sample sizes M, up to 10,000. For each density and sample size, 500 simulated samples were produced. We can observe that in most cases the ISE values of our method (BADE) are lower that those of Hazelton's method (VB3). The exceptions are H4 with $M = 100$ and 200, for which they are virtually the same. In some cases we note larger variability in BADE's ISE values than in VB3's. This is presumably due to a lesser degree of smoothing in BADE than in VB3.

Even though the differences in accuracy seem to be small in some cases, even a small consistent difference can be considered significant in this problem, as it has been difficult to make performance improvements in density estimation even when moving from fixed-bandwidth to variable-bandwidth methods (Terrell and Scott, 1992).

3.1.2. Real Examples

We tested our method on three univariate real data sets. Although not related to our intended application domain of molecular dynamics, the three data sets are widely used in the relevant statistics literature so that we can compare results easily with those from other methods:

1. **Univariate Old Faithful:** lengths, in minutes, of 107 eruptions of the Old Faithful geyser (Silverman, 1986).
2. **Suicide:** lengths, in days, of 86 treatment spells of control patients in a suicide study (Silverman, 1986).

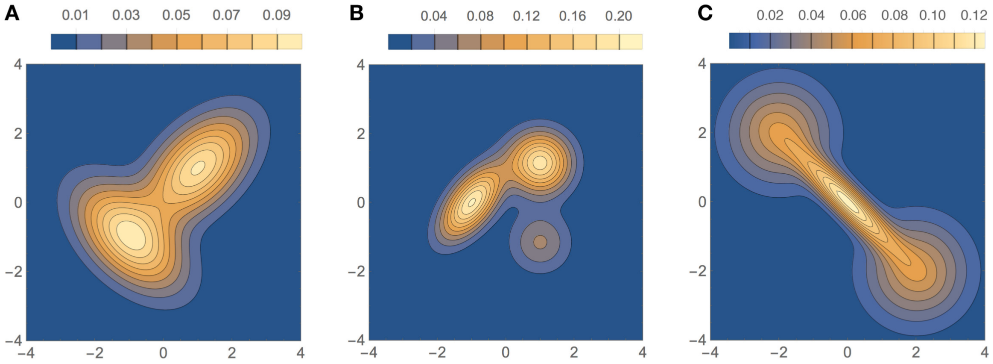


FIGURE 9 | Test densities (Gaussian mixtures) used in our bivariate simulation study. (A) F2 (bimodal); **(B)** F3 (trimodal); **(C)** F4 (dumbbell unimodal).

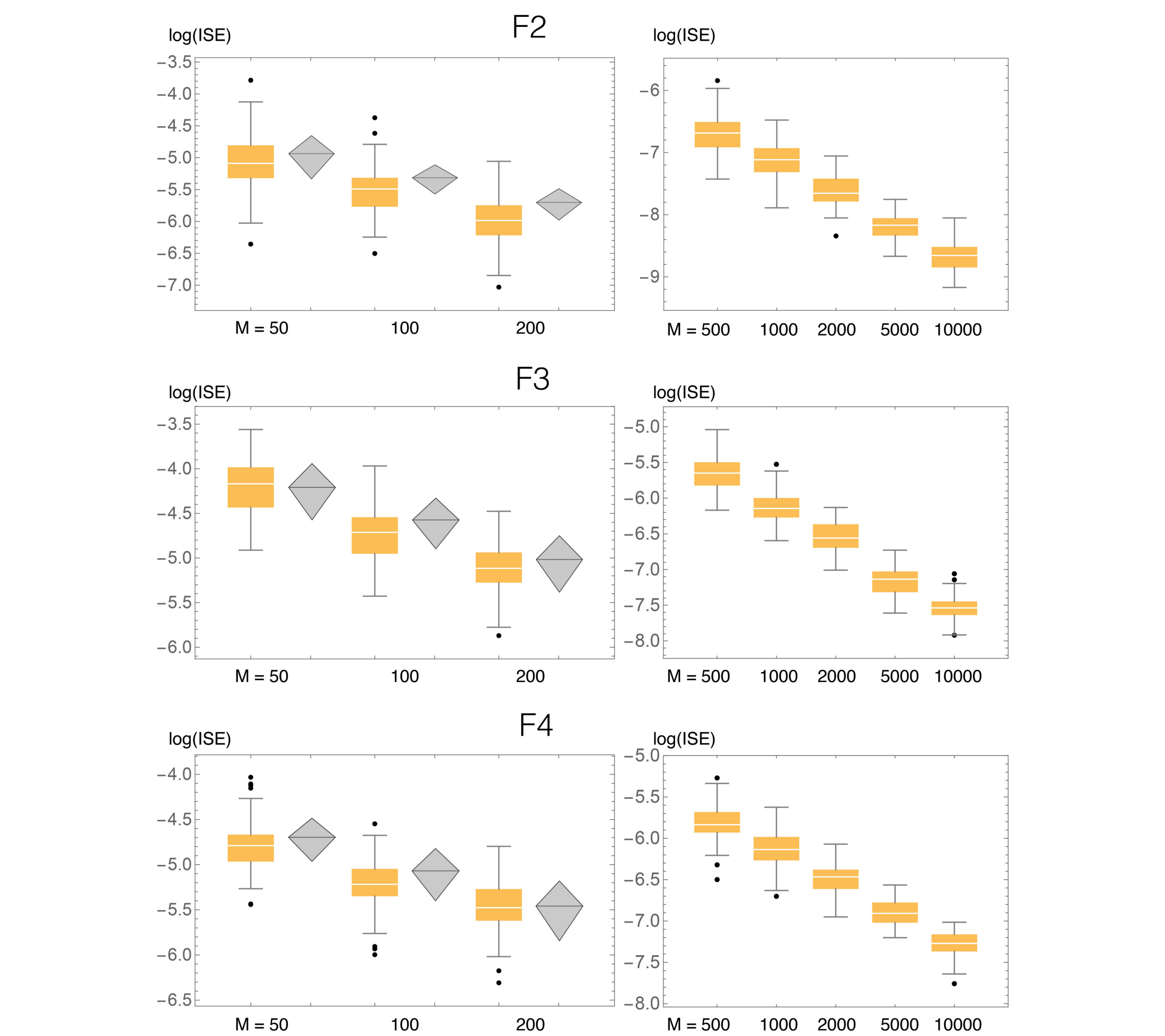


FIGURE 10 | ISE statistics for the three bivariate test densities (F2, F3, F4), comparing our method (BADE, in orange) and BABM (Zougab et al., 2014, in gray). Also shown are results for larger values of the sample size *M*, not considered by Zougab et al. Their results are shown by gray rhombi representing the mean plus and minus 1 standard deviation, which are the only data they reported.

3. **Hidalgo stamp:** paper thickness, in mm, of 485 stamps from the 1872 Hidalgo stamp issue (Izenman and Sommer, 1988). This data set is also available in the `locfit` package of the R software (Loader, 2013).

Results for the Old Faithful data set are shown in **Figure 6**, where the density estimate from our method, BADE, is compared to three others: LS (Brewer, 2000), CAKE (Ganti and Gray, 2011), and VB3 (Hazelton, 2003). We can see that the left peak matches quite well among the four methods, except that CAKE's estimate is somewhat shifted and has a wide shoulder, and LS's estimate has a lower value at this peak. As for the right peak, again CAKE's position is quite shifted to the right, and BADE's estimate shows a splitting in two submodes, which is visible just slightly in the other estimates. Finally, we observe that the other methods produce heavier tails than BADE (BADE will always produce "light," exponential tails due to the "effective" k (Equation 4). This k_e as a function of x is shown in the second row of the figure, before the covariance-smoothing step.) The third row of the figure shows the standard deviation $\sigma_k(x)$ (the one-dimensional analog of $V_k(\mathbf{x})$) of the set $N_k(x)$ of k nearest neighbors of x. Notice the balanced feature of the method: regions of higher k correspond to regions of lower σ_k, and vice versa.

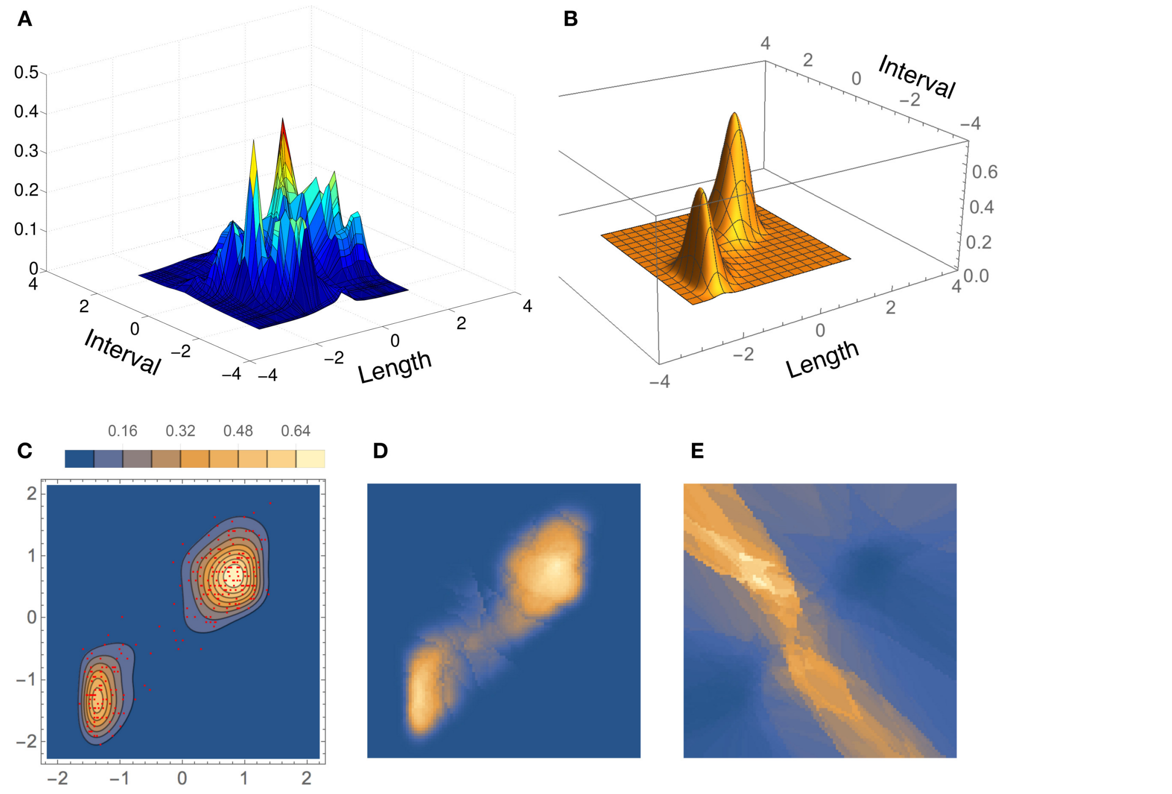


FIGURE 11 | Density estimates for the bivariate Old Faithful data set, comparing our method (BADE) with CAKE (Ganti and Gray, 2011). (A) CAKE estimate. **(B)** BADE estimate. **(C)** BADE estimate shown as a contour plot, along with the sample points. **(D)** Effective k values (i.e., k_e) on a grid of size 100 × 100. **(E)** V_k on the same grid. Both k_e and V_k are the ones before applying the covariance smoothing step.

The density estimates for the suicide data set are shown in **Figure 7**. We see a very good agreement among the three methods: BADE, LS (Brewer, 2000), and CAKE (Ganti and Gray, 2011). BADE shows a small satellite mode of the main peak, where LS and CAKE exhibit a small shoulder instead. On the other hand, CAKE is significantly more sensitive than BADE and LS to the sample points around $x = 250$, $x = 320$, and $x = 600$, while BADE and LS show only a small mode at around $x = 250$ and then they taper off. In this case we can see that that the exponential decay of k_e is slow as x grows, due to the large separation of the sample points at the right end, and hence the large σ_k values in that region.

The Hidalgo stamp comparison between BADE and LS is shown in **Figure 8**. In contrast with the Old Faithful example, here BADE's estimate looks more smoothed than LS's, but otherwise the position and number of modes is the same for both methods. This is interesting in connection with the results of the analysis carried out by Brewer (2000), whose LS method was the only one, among the ones considered in his comparison with previous methods, that revealed exactly five modes.

3.2. Bivariate case

3.2.1. Simulated Examples

The three bivariate simulated densities, all Gaussian mixtures, on which we tested our method are shown in **Figure 9**. They are densities F2, F3, and F4 used by Zougab et al. (2014), and we will refer to them by the same names:

1. **F2** (bimodal, similar to density H of Wand and Jones, 1993): $\frac{1}{2}\mathrm{N}[(1,1),\Sigma_1]+\frac{1}{2}\mathrm{N}[(-1,-1),\Sigma_2]$, where $\Sigma_1=\left(\begin{smallmatrix}1 & 1/2\\ 1/2 & 1\end{smallmatrix}\right)$ and $\Sigma_2=\left(\begin{smallmatrix}1 & -1/2\\ -1/2 & 1\end{smallmatrix}\right)$.
2. **F3** (trimodal, equal to density K of Wand and Jones, 1993): $\frac{3}{7}\mathrm{N}[(-1,0),\Sigma_1]+\frac{3}{7}\mathrm{N}[(1,2/\sqrt{3}),\Sigma_2]+\frac{1}{7}\mathrm{N}[(1,-2/\sqrt{3}),\Sigma_3]$, where $\Sigma_1=\left(\begin{smallmatrix}9/25 & 63/250\\ 63/250 & 49/100\end{smallmatrix}\right)$ and $\Sigma_2=\Sigma_3=\left(\begin{smallmatrix}9/25 & 0\\ 0 & 9/25\end{smallmatrix}\right)$.
3. **F4** ("dumbbell" unimodal): $\frac{4}{11}\mathrm{N}[(-2,2),\Sigma_1]+\frac{3}{11}\mathrm{N}[(0,0),\Sigma_2]+\frac{4}{11}\mathrm{N}[(2,-2),\Sigma_3]$, where $\Sigma_1=\Sigma_3=\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$ and $\Sigma_2=\left(\begin{smallmatrix}0.8 & -0.72\\ -0.72 & 0.8\end{smallmatrix}\right)$.

Results of ISE statistics comparing our method with that of Zougab et al. (2014) are displayed in **Figure 10**. Zougab et al.'s results were taken directly from their paper, but since they report just the mean and standard deviation of the ISE values, we show those two parameters as rhombi, whose horizontal line corresponds to the mean, and whose top and bottom vertices correspond to ±1 standard deviation from the mean. We also considered larger sample sizes, up to 10,000. For each density and sample size, 100 simulated samples were produced. We can observe that in most cases the ISE values of our method (BADE) are lower that those of Zougab et al.'s method (BABM). The exceptions are F3 with

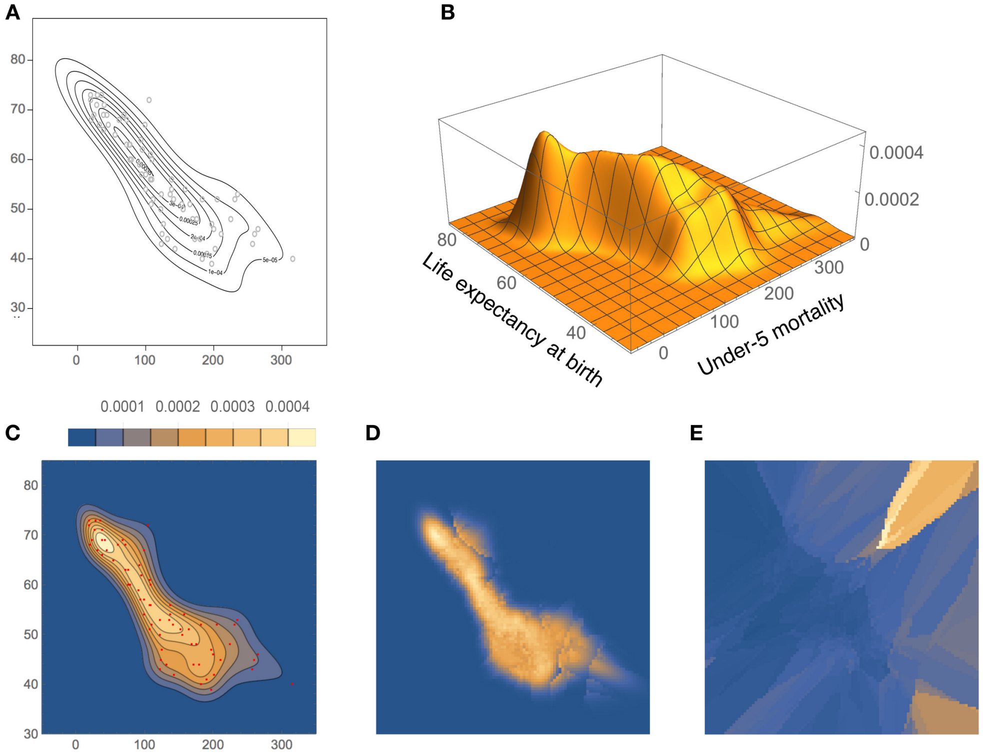


FIGURE 12 | Density estimates for the UNICEF data set, comparing our method (BADE) with BABM (Zougab et al., 2014). (A) BABM estimate. **(B)** BADE estimate. **(C)** BADE estimate shown as a contour plot, along with the sample points. **(D)** Effective k values (i.e., k_e) on a grid of size 100 × 100. **(E)** V_k on the same grid. Both k_e and V_k are the ones before applying the covariance smoothing step.

$M = 50$ and F4 with $M = 200$, for which they are very similar.

3.2.2. Real Examples

We tested our method on two bivariate real data sets, and compared the results with those from other methods. Again we chose data from outside our intended application in molecular dynamics to better compare with the available statistics literature:

1. **Bivariate Old Faithful:** length of eruptions vs. interval between consecutive eruptions, for 272 observations of the Old Faithful geyser (Härdle, 1991). These data are also available on the Internet, as extra material to Wasserman's book (Wasserman, 2004).
2. **UNICEF:** under-5 mortality (number of children who died under age 5 per 1,000 live births) vs. the average life expectancy (in years) at birth, for 73 countries with Gross National Income less than US$1,000 per annum per capita. These data are available from the UNICEF and also in the `ks` package of the R software (Duong, 2016).

Results for the bivariate Old Faithful data set are shown in **Figure 11**, where the density estimate from our method, BADE, is compared to that from CAKE (Ganti and Gray, 2011). (In this figure, we show the scaled coordinates in order to match CAKE's plot.) Both estimates show two main peaks; however, CAKE's estimate (**Figure 11A**) has, in addition, many other peaks that are not present in BADE's estimate, which is a clean bimodal density (**Figures 11B,C**). **Figures 11D,E** show, respectively, the effective number of nearest neighbors, k_e, and the area of the covariance ellipse, V_k (before the covariance-smoothing step). As in the univariate case, we see that regions of large k_e correspond to regions of small V_k, and vice versa.

The density estimates for the UNICEF data set, computed with our method and BABM (Zougab et al., 2014), are shown in **Figure 12**. Even though there is a good overall agreement between the two, BADE's estimate is apparently less smoothed than BABM's, resulting, in particular, in a shifted position of the mode toward the upper-left of the plot. In fact, BABM's estimate is virtually the same as that obtained using a fixed global bandwidth matrix (Zougab et al., 2014, Figure 3). **Figures 12D,E** show again the inverse relationship between k_e and V_k (before covariance smoothing).

4. CONCLUSION

We have implemented a novel adaptive density-estimation approach suitable for our statistical evaluation of membrane simulations in Wriggers et al. (2017).

Unlike most well known density estimation methods, ours is not based on kernels. Rather, it estimates the density at a given point directly, using the information about the sets of k nearest neighbors, finding the optimal k in an adaptive way, by balancing it with the size of the "covariance ellipsoid" of the set of nearest neighbors. Thus, the calculation does not involve solving costly optimization problems, and is free of data-dependent parameters. (However, in the optional smoothing step, one could vary the coefficients in Equation (11), to obtain density estimates with greater or lesser amount of smoothing).

We note that, specially in the context of fixed-bandwidth kernels, the covariance matrix could be considered a parameter which depends on the data. However, since in our approach it is not a fixed value, but rather a function of the point, we do not call it a parameter. Rather, the parameters are the coefficients in Equation (11), which are fixed (except for the optional smoothing variation) and do not depend on the data.

BADE is well suited for large data sizes. Methods that center a kernel function at each sample point become very expensive as the data size grows. Instead, BADE relies only on nearest-neighbor information, whose average required number $\bar{k}(M)$ is such that $\bar{k}(M)/M \to 0$ as $M \to \infty$, where M is the sample size (Loftsgaarden and Quesenberry, 1965). Thus, the main step scales very well with data size (sublinearly in one and two dimensions, Equation 16). On the contrary, methods such as that of Zougab et al. (2014) (with which we compared ours) scale quadratically with the data size and are thus restricted to smaller data sets.

Our method is free of restrictions on the bandwidth matrices, such as diagonal or scalar. In fact, we are no longer dealing with "bandwidth" matrices, but covariance matrices of sets of nearest neighbors.

BADE has been defined for data of any dimension; however, we have worked out the constants and made tests only for dimensions 1 and 2. It is most efficient in low dimensions, due to the need to compute nearest neighbors. For this, it takes advantage of the R*-tree data structure (Beckmann et al., 1990), which is, to the best of our knowledge, the most efficient one for nearest-neighbor search in low dimensions. In higher dimensions the R*-tree data structure becomes less efficient due to the increasing relative volume of the "corners" of the hyperrectangles, and so better adapted data structures would be preferable in this case (see Hjaltason and Samet, 1999 for details).

Our method was validated, both in the univariate and the bivariate settings, by ISE analyses on some simulated densities. These analyses consisted in generating a number of simulated samples (500 for the univariate case, 100 for the bivariate case) and measuring the integrated square error (ISE) between the density estimated from each sample and the actual density function. The ISE statistics were compared with similar results from previous approaches that were among the best available. In most cases we obtained lower errors, and in the remaining few cases the performance was virtually identical.

The apparent synergy between objective (low ISE) and subjective (visual appeal) criteria in our algorithm is a curious phenomenon that has also been observed by other researchers. Farmen and Marron (1999) pointed out that "visual error appears to be quite informative about performance," whereas Brewer (2000) stated that "subjective feeling about density estimates" produces "estimates relatively free from unnecessary minor fluctuations." Although the earlier work provides a rationale for including subjective criteria in our work, an open research question is whether aesthetics and objective error are covariant. Farmen and Marron have attempted to quantify visual appeal (Farmen and Marron, 1999) but they found that "good performance in MISE does not guarantee visually appealing curve estimates." In contrast, Hazelton (2003) states that "gains in ISE may understate the improvements in visual appeal," which seems to imply at least a weak dependence. A more systematic investigation of the objective value of subjective criteria could be the subject of future work.

The optional covariance-smoothing step in BADE yields very visually appealing density estimates, as our real-data examples show, but is not strictly necessary if all that's needed is a density estimate to perform further calculations. For instance, one of the applications for which we need bivariate density estimates is the computation of Mutual Information. In this case we don't need visually appealing functions, and thus we can save significant compute time.

At this time the algorithm is implemented as a C program. It will be freely disseminated as a part of release 1.5 of our software package *TimeScapes* (Kovacs and Wriggers, 2016). The web site for our software is http://timescapes.biomachina.org.

AUTHOR CONTRIBUTIONS

The mathematical theory of BADE was designed by JK. Experimental test data sets used in this work were prepared by CH. The larger project (including the accompanying paper) was supervised by WW. The paper was written by JK and WW.

FUNDING

This work was supported in part by the Frank Batten endowment and by National Institutes of Health grant R01GM62968 (to WW).

REFERENCES

Abramson, I. S. (1982). On bandwidth variation in kernel estimates—A square root law. *Ann. Statist.* 10, 1217–1223. doi: 10.1214/aos/1176345986

Beckmann, N., Kriegel, H.-P., Schneider, R., and Seeger, B. (1990). "The R*-tree: an efficient and robust access method for points and rectangles," in *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, SIGMOD '90* (New York, NY: ACM), 322–331. doi: 10.1145/93597.98741

Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* 19, 135–144. doi: 10.1080/00401706.1977.10489521

Brewer, M. J. (2000). A Bayesian model for local smoothing in kernel density estimation. *Stat. Comput.* 10, 299–309. doi: 10.1023/A:1008925425102

Duong, T. (2016). *ks: Kernel Smoothing*. R package version 1.10.3.

Farmen, M., and Marron, J. S. (1999). An assessment of finite sample performance of adaptive methods in density estimation. *Comput. Stat. Data Anal.* 30, 143–168. doi: 10.1016/S0167-9473(98)00070-X

Fredman, M. L., Sedgewick, R., Sleator, D. D., and Tarjan, R. E. (1986). The pairing heap: a new form of self-adjusting heap. *Algorithmica* 1, 111–129. doi: 10.1007/BF01840439

Ganti, R., and Gray, A. (2011). "CAKE: convex adaptive kernel density estimation," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, volume 15 of *JMLR Workshop and Conference Proceedings*, eds G. Gordon, D. Dunson, and M. Dudík (Fort Lauderdale, FL), 498–506.

Hall, P., Hu, T. C., and Marron, J. S. (1995). Improved variable window kernel estimates of probability densities. *Ann. Stat.* 23, 1–10. doi: 10.1214/aos/1176324451

Härdle, W. (1991). *Smoothing Techniques with Implementation in S, 1st Edn*. New York, NY: Springer-Verlag.

Hazelton, M. L. (2003). Variable kernel density estimation. *Aust. New Zealand J. Stat.* 45, 271–284. doi: 10.1111/1467-842X.00283

Hjaltason, G. R., and Samet, H. (1999). Distance browsing in spatial databases. *ACM Trans. Database Syst.* 24, 265–318. doi: 10.1145/320248.320255

Izenman, A. J., and Sommer, C. J. (1988). Philatelic mixtures and multimodal densities. *J. Am. Stat. Assoc.* 83, 941–953. doi: 10.1080/01621459.1988.10478683

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.* 91, 401–407. doi: 10.1080/01621459.1996.10476701

Katkovnik, V., and Shmulevich, I. (2002). Kernel density estimation with adaptive varying window size. *Patt. Recogn. Lett.* 23, 1641–1648. doi: 10.1016/S0167-8655(02)00127-7

Kovacs, J. A., and Wriggers, W. (2016). Spatial heat maps from fast information matching of fast and slow degrees of freedom: application to molecular dynamics simulations. *J. Phys. Chem. B.* 120, 8473–8484. doi: 10.1021/acs.jpcb.6b02136

Liu, H., Lafferty, J., and Wasserman, L. (2007). "Sparse nonparametric density estimation in high dimensions using the Rodeo," in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, volume 2 of JMLR Workshop and Conference Proceedings, eds M. Meila and X. Shen (San Juan), 283–290.

Loader, C. (2013). *locfit: Local Regression, Likelihood and Density Estimation*. R package version 1.5-9.1.

Loftsgaarden, D. O., and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.* 36, 1049–1051. doi: 10.1214/aoms/1177700079

Marron, J. S., and Tsybakov, A. B. (1995). Visual error criteria for qualitative smoothing. *J. Am. Stat. Assoc.* 90, 499–507. doi: 10.1080/01621459.1995.10476541

Marron, J. S., and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* 20, 712–736. doi: 10.1214/aos/1176348653

Sain, S. R. (2002). Multivariate locally adaptive density estimation. *Comput. Stat. Data Anal.* 39, 165–186. doi: 10.1016/S0167-9473(01)00053-6

Sain, S. R., and Scott, D. W. (1996). On locally adaptive density estimation. *J. Am. Stat. Assoc.* 91, 1525–1534. doi: 10.1080/01621459.1996.10476720

Sheather, S. J., and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B (Methodol.)* 53, 683–690.

Shimazaki, H., and Shinomoto, S. (2010). Kernel bandwidth optimization in spike rate estimation. *J. Comput. Neurosci.* 29, 171–182. doi: 10.1007/s10827-009-0180-4

Silverman, B. W. (1986). *Density Estimation, 1st Edn*. Boca Raton, FL: Chapman & Hall/CRC.

Song, L., Zhang, X., Smola, A., Gretton, A., and Schölkopf, B. (2008). "Tailoring density estimation via reproducing kernel moment matching," in *Proceedings of the 25th International Conference on Machine Learning, ICML '08* (New York, NY: ACM), 992–999.

Terrell, G. R., and Scott, D. W. (1992). Variable kernel density estimation. *Ann. Stat.* 20, 1236–1265. doi: 10.1214/aos/1176348768

Vapnik, V., and Mukherjee, S. (2000). "Support vector method for multivariate density estimation," in *Proceedings of the Conference on Neural Information Processing Systems (NIPS 1999),* volume 12 of *Advances in NIPS*, eds S. A. Solla, T. K. Leen, and K. Müller (Cambridge, MA: MIT Press), 659–665.

Wand, M. P., and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Am. Stat. Assoc.* 88, 520–528. doi: 10.1080/01621459.1993.10476303

Wand, M. P., and Jones, M. C. (1995). *Kernel Smoothing, 1st Edn*. London: Chapman and Hall.

Wasserman, L. (2004). *All of Statistics, 1st Edn.* New York, NY: Springer.

Wriggers, W., Castellani, F., Kovacs, J. A., and Vernier, P. T. (2017). Computing spatiotemporal heat maps of lipid electropore formation: a statistical approach. *Front. Mol. Biosci.* 4:22. doi: 10.3389/fmolb.2017.00022

Wu, T.-J., Chen, C.-F., and Chen, H.-Y. (2007). A variable bandwidth selector in multivariate kernel density estimation. *Stat. Prob. Lett.* 77, 462–467. doi: 10.1016/j.spl.2006.08.013

Zougab, N., Adjabi, S., and Kokonendji, C. C. (2014). Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation. *Comput. Stat. Data Anal.* 75, 28–38. doi: 10.1016/j.csda.2014.02.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership