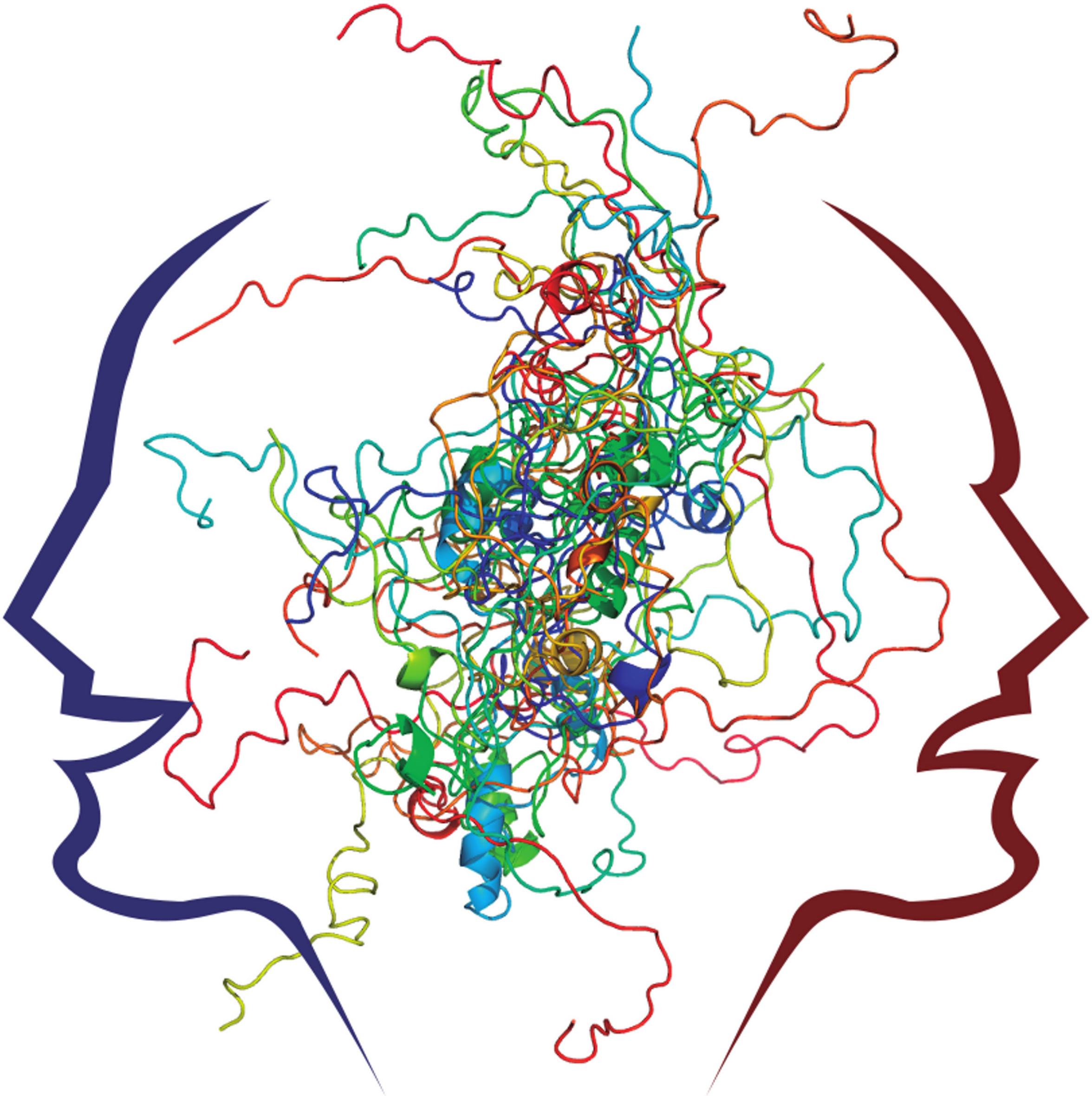
FUNCTION AND FLEXIBILITY: FRIEND OR FOE?

EDITED BY: Kris Pauwels and Peter Tompa

PUBLISHED IN: Frontiers in Molecular Biosciences



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88919-972-3
DOI 10.3389/978-2-88919-972-3

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.

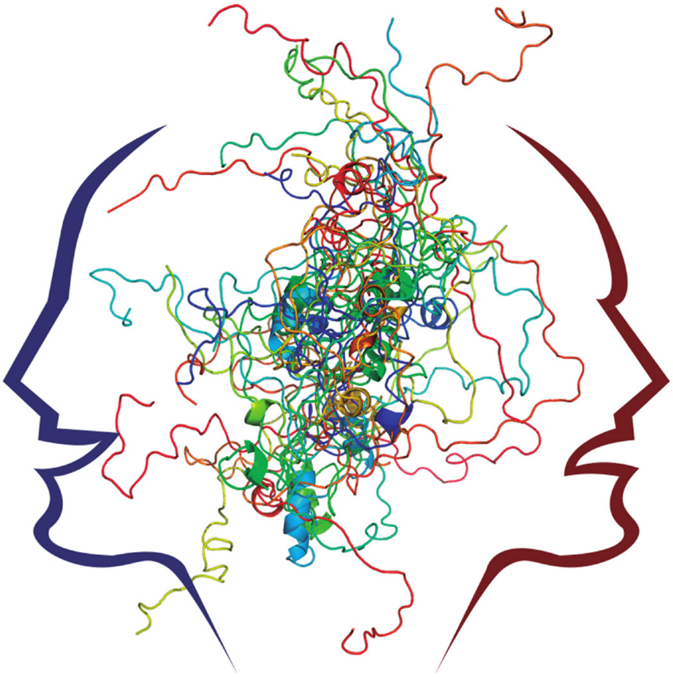
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

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**

FUNCTION AND FLEXIBILITY: FRIEND OR FOE?

Topic Editors:
Kris Pauwels, Vlaams Instituut voor Biotechnologie & Vrije Universiteit Brussel, Belgium
Peter Tompa, Vlaams Instituut voor Biotechnologie & Vrije Universiteit Brussel, Belgium;
Research Centre for Natural Sciences of the Hungarian Academy of Sciences, Hungary



Protein flexibility can be represented by a conformational ensemble. Sixteen different conformations of phosphorylated Sic1 are represented, which were generated based on NMR and SAXS data (entry 1AAA in the Protein Ensemble Database). Image by Luis Ferrer Campins, Sara Contreras Martos and Kris Pauwels.

Louis Sullivan (1856 - 1924) revolutionized architecture by designing the first skyscraper and he became famous by proclaiming that "form follows function". When x-ray crystallographers visualized the structures of proteins for the first time, the structural biology field embraced the view that "function follows form" as the 3D-architecture of proteins could unveil various aspects of their function. Despite the original "1 gene - 1 protein structure - 1 function" relationship, nowadays a far more complicated picture emerges where the flexibility and dynamics of a protein can play a central role in a multitude of functions. The ultimate form(s) that a protein adopt when interacting with (a) partner molecule(s) are the most biologically relevant and in this context Sullivan's quote is still appropriate: the conformation that the protein adopts follows from the function of that protein.

Despite the fact that many well-characterized proteins have a well-folded structure, there is a growing interest in the conformational flexibility within proteins. This flexibility is also a balanced phenomenon: excess of flexibility can be detrimental for protein behaviour, as well as the lack thereof. Notwithstanding its importance, studying intrinsically disordered protein regions or conformational rearrangements can be a very challenging. Therefore, flexibility can be perceived as a friend or a foe, depending on the context.

This e-book showcases the impact of the study of protein flexibility on the structural biology field and presents protein flexibility in the context of disease as well as its benign aspects. As detailed knowledge of the structural aspects of polypeptides remains essential to comprehend protein function, one of the future challenges for structural biology also lies with large macromolecular protein complexes. Also there the dynamics and flexibility are essential for proper functioning and molecular movement, which is an important aspect of living matter. This challenge stimulated the development of advanced techniques to study protein flexibility and the use of those techniques to address fundamental biological and biomedical problems. Those innovations should help us to unravel the intimate link between protein function and flexibility and explore new horizons.

Citation: Pauwels, K., Tompa, P., eds. (2016). Function and Flexibility: Friend or Foe? Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-972-3

Table of Contents

frontiers
in Molecular Biosciences

EDITORIAL
published: 07 July 2016
doi: 10.3389/fmolb.2016.00031

Editorial: Function and Flexibility: Friend or Foe?

Kris Pauwels [1, 2]* *and Peter Tompa* [1, 2, 3]

[1] *VIB Structural Biology Research Center, Vlaams Instituut voor Biotechnologie, Brussels, Belgium,* [2] *Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium,* [3] *Research Centre for Natural Sciences of the Hungarian Academy of Sciences, Institute of Enzymology, Budapest, Hungary*

Keywords: intrinsically disordered proteins, protein dynamics, protein flexibility, conformational ensemble, protein function

OPEN ACCESS

Edited and reviewed by:
Annalisa Pastore,
King's College London, UK

****Correspondence:***
Kris Pauwels
krpauwel@vub.ac.be

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

Received: *08 June 2016*
Accepted: *23 June 2016*
Published: *07 July 2016*

Citation:
Pauwels K and Tompa P (2016)
Editorial: Function and Flexibility:
Friend or Foe?
Front. Mol. Biosci. 3:31.
doi: 10.3389/fmolb.2016.00031

The Editorial on the Research Topic

Function and Flexibility: Friend or Foe?

Protein structural biology aims to link snapshots of three-dimensional macromolecular structures to their biological function. The high-resolution information that is obtained traditionally by x-ray crystallography or nuclear magnetic resonance (NMR) experiments is instrumental for understanding their functional properties, their biological roles, and their potential roles in diseases ("function follows form"). Yet, proteins are not rigid and/or static entities: their dynamics and flexibility are essential for proper functioning and molecular movement, which is an important aspect of living matter. Many proteins even completely lack a well-defined 3D-structure under physiological conditions, the so-called intrinsically disordered proteins (IDPs). Up to 35% of human proteins are predicted to possess intrinsically disordered regions (IDRs) of at least 30 consecutive disordered residues, that play important roles in cell signaling and regulation (Guharoy et al., 2015) ("flexibility facilitates function"). Therefore, this research topic covers the impact of the study of protein flexibility on the structural biology field.

The articles in this e-book feature plenty of examples where protein flexibility controls protein functionality. In their fascinating *Perspective*, Kern and colleagues provide an excellent overview of our actual mechanistic insights of how the anticancer drug Gleevec selectively inhibits the Abl kinase (Agafonov et al.). Their work showcases how rigorous kinetic and structural analysis yields definitive conclusions that selectivity is a function of a conformational change after binding (induced-fit) and the resulting slow dissociation rate of Gleevec from the Abl kinase, whereby the flexibility in the famous and highly conserved DFG-loop plays an important role (Agafonov et al., 2014; Wilson et al., 2015). By reconstructing the evolution of the energy landscape of kinases through the synergy of "old-fashioned" stopped-flow kinetics and "modern" ancestral sequence reconstruction, they advocate for the combined use of experimental studies and molecular dynamics approaches to find effective and selective kinase inhibitors.

The benign role of protein flexibility is also nicely illustrated in the *review* by Gontero and colleagues who demonstrate the central and multiple functionality of C- and N-terminal intrinsically disordered tails of globular proteins in photosynthetic organisms (Thieulin-Pardo et al.). They exemplify that protein flexibility at the N- and C-terminal extremities accounts for an increased number of binding partners and how new roles may emerge by the evolutionary addition of an intrinsically disordered extension. Indeed, often IDRs play a role in molecular recognition and binding events, whereby they can undergo a folding transition induced by the partner protein ("form follows function"). By a large scale thermodynamic assessment of mostly binary protein-protein interactions of ordered-ordered and ordered-disordered protein complexes, Kragelund and colleagues help shedding light on the debated role of kinetics and thermodynamics in the binding properties of IDPs (Teilum et al.). Through this capacity for interaction with other molecules, protein flexibility can also be linked to disease (Hubin et al., 2014; Uversky, 2014; Guharoy et al., 2015). Fraternalli and colleagues studied the localization of common and disease-related mutations within (dis)ordered protein regions (Lu et al.). They highlight that intra-domain ordered and intra-domain disordered regions show high propensity for disease-related mutations, while inter-domain disordered regions are enriched in common variants. Their analysis offers interesting perspectives for the further development of the field of protein flexibility and disorder. It also supports the fact that, in the field of IDPs, computational approaches play a major role. As such, Craveur et al. show that the concept of structural alphabets is suitable to analyze the dynamics and flexibility of proteins. In their comprehensive *review* they advocate that structural alphabets are required to begin to understand the complexity of protein flexibility by discriminating flexibility from mobility and deformability.

The IDP field is also one of the few areas in structural and molecular biology where the experiments provide support to computations to achieve an accurate understanding of the conformational properties of these complex proteins. Varadi et al. review the current characterizations of IDPs by combining computations and experiments. The *mini-review* identifies key developments in the field, including the employment of experimental data into structural refinement in search of the functional repertoire of IDPs. With regard to wet-lab experimental approaches, several emerging techniques allow to overcome some of the technical problems of studying IDPs and to obtain essential information on protein dynamics. In their *original research paper*, Barran and collaborators exemplify the potential of ion-mobility mass spectrometry to track conformational changes in unstructured proteins on a millisecond timescale (Dickinson et al.). They characterize the effect of two small molecule compounds RITA and nutlin-3 on their IDP targets with a multi-technique approach. The *minireview* by Belle and coworkers showcases the power of site-directed spin labeling with electron paramagnetic resonance to investigate flexible regions and fuzziness in proteins (Le Breton et al.). The information obtained by NMR can generate conformational ensembles that visualize the conformations that IDPs sample under functional conditions. Because protein disorder can be evaluated at the residue level with NMR, Nielsen and Mulder compiled a small database of disorder-containing proteins using experimental NMR chemical shift data in their *original research paper* that is felicitously entitled "There is Diversity in Disorder – 'In all Chaos there is a Cosmos, in all Disorder a Secret Order"'. They demonstrate that those proteins span the full spectrum of disorder, yet segregate into two classes: proteins mostly disordered but with small segments of order scattered along the sequence, or structured proteins with small segments of disorder inserted between the different structured regions. This study is also illustrative for the concept of "form and function follow (NMR) frequency."

Recently the $\underline{\text{D}}^3$-concept was introduced for IDPs by revealing the interconnections between protein intrinsic Disorder and Degenerative Diseases (Uversky, 2014). In analogy, it is opportune to introduce the $\underline{\text{F}}^3$-concept for flexible proteins, since "Function Follows Flexibility." Whereas in the past intrinsic disorder could cause frustration because IDRs were considered frivolous and flamboyant, their flirtatious behavior flaunted formidable features. We hope this e-book can stimulate the research community to finally stop fumbling for the fugacious forms of flexible proteins and bring their functional framing to fruition.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

PT is supported by the Odysseus grant G.0029.12 from Research Foundation Flanders (FWO) and KP is the recipient of a FWO postdoctoral fellowship (#1218713).

REFERENCES

Agafonov, R V., Christopher, W., Renee, O., Vanessa, B., and Dorothee, K. (2014). Energetic dissection of Gleevec's selectivity toward human tyrosine kinases. *Nat. Struct. Mol. Biol.* 21, 848–853. doi: 10.1038/nsmb.2891

Guharoy, M., Pauwels, K., and Tompa, P. (2015). SnapShot: intrinsic structural disorder. *Cell* 161, 1230–1230.e1. doi: 10.1016/j.cell.2015.05.024

Hubin, E., Van Nuland, N. A., Broersen, K., and Pauwels, K (2014). Transient dynamics of Aβ contribute to toxicity in Alzheimer's disease. *Cell. Mol. Life Sci.* 71, 3507–3521. doi: 10.1007/s00018-014-1634-z

Uversky, V. N. (2014). The triple power of D^3: protein intrinsic disorder in degenerative diseases. *Front. Biosci. (Landmark Ed)* 19, 181–258. doi: 10.2741/4204

Wilson, C., Agafonov, R. V., Hoemberger, M., Kutter, S., Zorba, A., Halpin, J., et al. (2015). Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* 347, 882–886. doi: 10.1126/science.aaa1823

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

PERSPECTIVE
published: 21 May 2015
doi: 10.3389/fmolb.2015.00027

Evolution and intelligent design in drug development

Roman V. Agafonov, Christopher Wilson and Dorothee Kern *

Howard Hughes Medical Institute and Department of Biochemistry, Brandeis University, Waltham, MA, USA

Sophisticated protein kinase networks, empowering complexity in higher organisms, are also drivers of devastating diseases such as cancer. Accordingly, these enzymes have become major drug targets of the twenty-first century. However, the holy grail of designing specific kinase inhibitors aimed at specific cancers has not been found. Can new approaches in cancer drug design help win the battle with this multi-faced and quickly evolving enemy? In this perspective we discuss new strategies and ideas that were born out of a recent breakthrough in understanding the molecular basis underlying the clinical success of the cancer drug Gleevec. An "old" method, stopped-flow kinetics, combined with old enzymes, the ancestors dating back up to about billion years, provides an unexpected outlook for future intelligent design of drugs.

Keywords: drug design, evolution, cancer drugs, protein kinases, conformational selection and induced fit, Gleevec

OPEN ACCESS

Edited by:
Peter Tompa,
Flanders Institute of Biotechnology (VIB), Belgium

Reviewed by:
Doriano Lamba,
Consiglio Nazionale delle Ricerche, Italy
Miquel Adrover,
Universitat de les Illes Balears, Spain

Correspondence:
Dorothee Kern,
Howard Hughes Medical Institute and Department of Biochemistry, Brandeis University, Waltham, MA 02452, USA
dkern@brandeis.edu

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

Received: *24 March 2015*
Accepted: *08 May 2015*
Published: *21 May 2015*

Citation:
Agafonov RV, Wilson C and Kern D (2015) Evolution and intelligent design in drug development. Front. Mol. Biosci. 2:27. doi: 10.3389/fmolb.2015.00027

The Beauty and Curse of Protein Kinases

Why are we more sophisticated than a yeast cell? One of the reasons is protein kinases, that exploded both in numbers (more than 500 in humans compared to 130 in yeast) and sophistication with the development of multicellularity (Richter and King, 2013). The evolution of specialized kinases enabled complex regulatory networks in higher organisms thereby providing a huge evolutionary edge. However, a crack in this machinery as little as a single point mutation in a kinase can cause cancer—an Achilles heel that has elevated protein kinases into the number one drug target of the twenty-first century (Cohen, 2002; Cohen and Alessi, 2013; Wang et al., 2014). The stringent requirements for catalyzing a chemical reaction that uncatalyzed would take about 7000 years (Stockbridge and Wolfenden, 2009; Kerns et al., 2015) resulted in a strong conservation of the active sites, which have thus been extensively targeted in cancer drug development. Unfortunately, inhibitors targeting the ATP binding site tend to be unselective due to this active site conservation, leading to unwanted side effects. The popularity of the field of protein kinase inhibition as well as alternative strategies such as inhibition of substrate binding and protein interaction sites is best reflected by a number of recent reviews (Wang et al., 2014 and a special issue in ACS Chemical Biology, 2015). In addition, new high-throughput assays are being constantly developed to facilitate screening of the compounds (Acker and Auld, 2014), however, the major goal of the pharmaceutical industry to develop specific kinase inhibitors remains a daunting challenge.

The Wonder Drug of the Century

Gleevec is an exception, as it has great specificity for the onco-protein BCR-Abl (Capdeville et al., 2002; Cohen et al., 2002). The BCR-Abl fusion protein results from reciprocal translocation between chromosome 9 and chromosome 22, widely known as the Philadelphia translocation, leading to a constitutively active kinase (Rowley, 1973; Daley and Baltimore, 1988). Gleevec was approved by the FDA for clinical use in 2001, and has proven to be remarkably successful in treating chronic myeloid leukemia (CML) and gastrointestinal stromal tumors. Its success generated tremendous enthusiasm in the scientific community and even general public, after the reports about "new ammunition in the war against cancer" and its outstanding effectiveness were picked up by the media (Lemonick and Park, 2001; Newsweek, 2001; Wade, 2001). Gleevec was viewed as a "proof of principle drug," which showed the possibility of rational design of an inhibitor that would specifically target a kinase of interest. Unfortunately, tireless efforts aimed at understanding the molecular mechanisms of Gleevec's selectivity over the last 20+ years were mostly unsuccessful, and the original expectations of a steady stream of new therapeutics emerging from basic research turned out to be overoptimistic. As reviewed recently (Cohen and Alessi, 2013; Wang et al., 2014) since Gleevec's triumph, approximately 20 new kinase inhibitors were developed and entered clinical use. This is a rather small number considering that there are more than 500 human kinases and multiple inhibitors are needed for each of them to combat the inevitable mutations that lead to drug resistance. A fundamental pitfall in drug development is a lack of understanding of the detailed biophysical mechanisms that make inhibitors successful.

Conformational Selection and the Famous "DFG-Loop"

In the search for the physical determinants of Gleevec selectivity, the DFG – loop (Asp-Phe-Gly), a 100% conserved element in the kinome (**Figure 1A**), stood out as a structural feature that differs between kinases that bind Gleevec tightly or weakly. In the x-ray structure of Abl, this loop adopts an "out" conformation in both the apo and Gleevec-bound protein, while in the closest homolog and weak binder Src kinase it occupies a binding-incompetent "in" conformation in the apo protein that would have to move into the "out" position to accommodate the drug (Xu et al., 1997; Schindler et al., 2000; Nagar et al., 2003; Seeliger et al., 2007). These structures, together with the fact that the active conformations look too similar to provide selectivity, shifted attention toward structural determinants of inactive conformations.

It was hypothesized that the preferential occupancy of the DFG-out state by Abl but not Src is the primary source of Gleevec selectivity. This model of an equilibrium between binding-incompetent (DFG-in) and competent state (DFG-out) (K_{CS}) being the source for differential drug affinities is a classical conformational selection mechanism (Cowan-Jacob et al., 2005; Dar et al., 2008; Shan et al., 2009; Aleksandrov and Simonson, 2010; Lovera et al., 2012; Lin and Roux, 2013; Lin et al., 2014) that has recently gained popularity in biology (see the special issue of Biophysical Chemistry and references within) (Biophysical Chemistry, 2014) (Scheme in **Figure 1E**). This hypothesis was further substantiated by the observation that less selective inhibitors such as Dasatinib do not differentiate between "in" and "out" conformations of the DFG-loop.

The elegance of this hypothesis, the direct observation of two different states of the DFG-loop in crystal structures and the excellent fit to the "expected model" of drug selectivity resulted in a wealth of literature focusing on this aspect of protein dynamics. A variety of approaches, both experimental (Vogtherr et al., 2006; Vajpai et al., 2008) and computational, were taken to quantify the free energy profile of the DFG-loop dynamics (Levinson et al., 2006; Aleksandrov and Simonson, 2010; Lovera et al., 2012; Lin and Roux, 2013; Lin et al., 2014; Meng et al., 2015). However, experimental studies of DFG-loop equilibrium in solution were complicated by high dynamics of this loop hampering quantification of this equilibrium. Some computational reports seemed to quite impressively quantitatively recapitulate the experimentally observed Gleevec affinities for the different kinases (Lin and Roux, 2013; Lin et al., 2014) despite the widely acknowledged current computational limitations for accurate energy calculations (Shaw et al., 2010; Piana et al., 2011; Lindorff-Larsen et al., 2012). Other computational studies were contradictory, and results varied depending on the methodology used. Despite the lack of direct experimental observation of the DFG-loop equilibrium, the DFG-loop hypothesis underlying selectivity became so popular that all active site kinase inhibitors were classified as class I (binding to both DFG-in and -out conformations) and class II (binding exclusively to the DFG-out state).

Although large screens hinted at a trend that class-II inhibitors may be more selective, many counterexamples of selective type I and promiscuous type II inhibitors were observed (Davis et al., 2011; Treiber and Shah, 2013). These data suggested that the DFG-loop may not be as essential for selectivity as initially thought. Paradoxically, despite its logical appeal, the DFG-loop conformational selection model did not lead to new highly selective kinase drugs. What is missing?

Old Fashioned?

A surprising breakthrough came from an unexpected direction. A new method of molecular time-travel back to the origin of these kinases and resurrection of their evolutionary trajectories into the modern kinases delivered the mechanism of Gleevec selectivity. Ironically, not only the resurrected enzymes that provided the understanding were old, so was the technique that yielded the answer. Stopped-flow kinetics, first described in the 1940s (Chance, 1940; Gibson et al., 1955) and often perceived as old-fashioned, has enormous potential when it comes to characterizing enzyme–drug interactions.

However the first hint for a new and unanticipated model came from following Gleevec binding to human Abl and Src by NMR, which revealed a slow conformational transition after drug binding that was different for the two kinases. Moreover, binding was sensed by residues far from the binding pocket indicating propagated conformational changes (Agafonov et al., 2014). Stopped-flow fluorescence experiments with modern Abl and Src (Agafonov et al., 2014) delivered quantification of the steps observed in the NMR experiments.

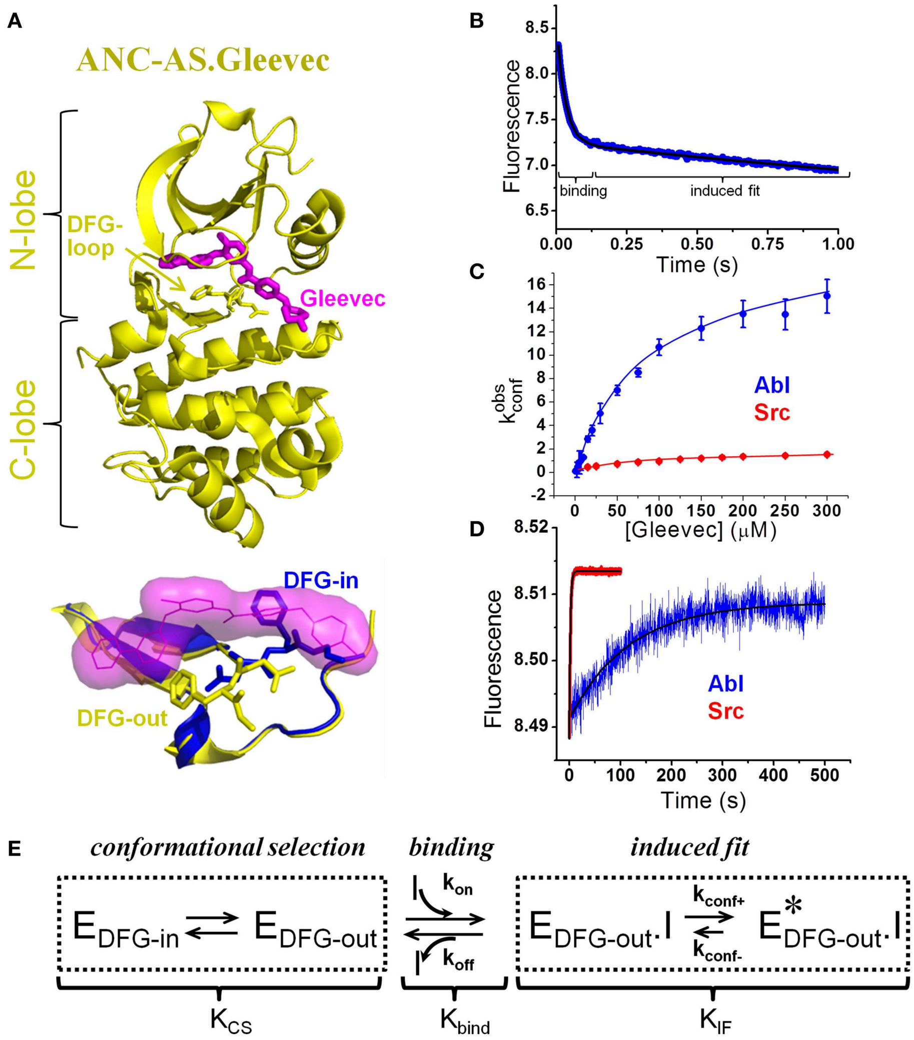


FIGURE 1 | Novel model of Gleevec binding to tyrosine kinases with quantification of individual steps. (A) Top: Crystal structure (4CSV) (Wilson et al., 2015) of last common ancestor of Src and Abl (ANC-AS) bound to Gleevec (magenta); the DFG loop is shown in stick. Bottom: DFG-loop in the -in (2SRC) and -out (4CSV) conformation is shown with Gleevec bound (magenta surface). Only the DFG-out conformation is compatible with Gleevec binding. **(B–D)** Binding and dissociation kinetics of Gleevec to Abl and Src measured by stopped-flow fluorescence (for details see Agafonov et al., 2014). **(B)** Gleevec binding to Abl at 5°C is biphasic with the fast phase corresponding to the physical binding step and slow phase corresponding to the induced fit step. Blue – experimental data, black – double-exponential fit. **(C)** Dependence of k_{conf}^{obs} [observed rate of the induced fit step, see scheme in **(E)**] on Gleevec concentration. **(D)** Dissociation kinetics of Gleevec from Abl and Src measured by dilution of enzyme-Gleevec complexes, which determines the k_{conf-} rate constant [see scheme in **(E)**]. **(E)** Gleevec binding scheme showing three distinct steps: conformational selection step, physical binding of the drug to the binding competent state, and the following conformational transition (induced fit). Equilibrium constants corresponding to each step (K_{CS}, K_{bind}, and K_{IF}) determine the overall binding affinity (K_D): $K_D = \frac{(K_{CS}+1) \bullet K_{bind} \bullet K_{IF}}{(1+K_{IF})}$.

Contrary to the previously explored models, the dominant role in Gleevec's selectivity belongs to the conformational transitions in the kinase-drug complex (induced fit, **Figure 1E**), and not to the DFG-loop conformational selection or the physical binding step. These induced fit transitions are the slowest steps with the forward rate (k_{conf+}) roughly 10 times faster in Abl compared to Src (**Figure 1C**). The rate of the reverse step, k_{conf-}, measured by dilution experiments, is 70-fold slower in Abl (**Figure 1D**), leading to a 700-fold difference in the overall equilibrium (K_{IF}) (**Figure 1E**). Because of simple principles of coupled equilibria, this 700-fold shift of the induced fit step equilibrium results in a 700-fold increase in the overall affinity for Gleevec, therefore accounting for most of the observed 3000-fold difference [the remaining four-fold difference comes from the DFG-loop conformational selection (see below)]. The actual binding step to the two kinases is nearly identical highlighting the limited usefulness of docking studies that play a prominent role in the current computational efforts in drug design. This "numbers-game" from the stopped-flow experiments delivered a new mechanism that quantitatively accounts for the long-known difference in kinase affinities for Gleevec and hence answers the long-standing question of specificity (Agafonov et al., 2014).

Inspired by the new findings for Gleevec we advocate that the full energy profiles need to be considered, since the differences between kinases are rooted in the differences of the free energies of all states along the binding trajectory. The role of induced fit in substrate binding to enzymes for better substrate positioning for catalysis has been appreciated, however its experimental quantification is still not a commonly applied practice. Possible roles of induced fit for drug binding was also nicely discussed (Copeland, 2011), but its role in inhibitor affinity and selectivity remains undervalued. Notably, only the local rearrangements around the drug-binding pocket instead of long-range conformational transitions are often considered in rational drug design. Such long-range dynamics is, in fact, in play for the Gleevec specificity, as exposed by the ancestor resurrection.

The Devil is in the [Atomistic] Details

While a physical chemist might be satisfied having figured out the kinetic scheme with hard numbers that rationalize the different drug affinities, the structural biologist will ask: which residues are responsible for the different energy landscapes? This might appear easy—just start mutating residues in the weak binding Src to mimic Abl. However, in spite of a large number of tested substitutions, such efforts were not successful indicating that the underling mechanism for Gleevec selectivity is more complex than anticipated (Seeliger et al., 2007). This approach although tempting has the following unavoidable drawbacks. Many differences accumulated during divergent evolution result from neutral drift (substitutions that are neutral for function and thus are not under selective pressure), and basically represent noise, from which one needs to fish out the sequence changes linked to the property of interest. To make the mater worse, some amino acid changes only come into play in the background of other mutations – a phenomenon called epistasis (Depristo et al., 2005; Harms and Thornton, 2013; Boucher et al., 2014). As a consequence, simple sequence swaps between two modern enzymes don't work because they miss the effect of the corresponding evolution of the amino acid background.

As illustrated in Wilson et al. (2015), ancestral sequence reconstruction (ASR) can be a powerful tool to overcome this challenge. ASR is a rapidly developing method that allows the inference of now nonexistent ancestral sequences using the growing amount of sequence information available. This approach was already formulated more than 50 years ago by Pauling and Zuckerkandl (1963). Modern enzymes (even the ones close in structure) still often differ from each other by 100+ residues. Such divergence in combination with neutral drift and epistasis makes it virtually impossible to rationally analyze the sequence differences. Ancestral reconstruction kills two birds with one stone. First, the sequence differences between two ancestors (or an ancestor and a modern protein) are smaller than those between the two modern enzymes, which makes a productive analysis of sequences more probable. Second, swaps between ancestor and its "grand-grand-children" can indeed shed light into atomistic mechanisms since epistasis is naturally accounted for.

In the work of Wilson et al. (2015) a phylogenetic tree of 76 modern kinases from different families and organisms of non-receptor tyrosine kinases was reconstructed, and protein sequences corresponding to key evolutionary branching points were resurrected (**Figure 2A**). Remarkably, all reconstructed ancient enzymes, differing by up to 100 amino acids from anything you can find today in nature, are fully active! The common ancestor of Src and Abl (called ANC-AS) had an intermediate affinity for Gleevec that increased along the evolutionary branch leading to Abl and decreased along the Src branch (**Figure 2B**).

Combining ancestral reconstruction with their Gleevec binding kinetics and structure illustrates the evolution of divergent energy landscapes (**Figure 2C**). Of interest to drug designers, it indeed delivered the atomistic mechanism responsible for Gleevec selectivity. Fifteen amino acid differences (out of 146) were identified to encode Gleevec specificity for Abl (**Figure 2D**) (Wilson et al., 2015). Their role in the induced fit step can now be rationalized structurally including stabilizing effects on drug–protein interaction and tuning differential flexibility via H-bonds remote from the drug-binding site (**Figure 2D**) (Wilson et al., 2015). So indeed long-range dynamics and epistasis are in play for Gleevec binding as first seen in the NMR studies (Agafonov et al., 2014) and hinted by the unsuccessful early swop approach (Seeliger et al., 2007).

Interestingly, the same residues correlated well with several resistant mutations found in patients who developed Gleevec resistance (Wilson et al., 2015). In other words, current evolution appears in these "dynamic hotspots," and the rationalization of the underlying atomistic mechanism for Gleevec resistance might help in designing drugs that overcome this detrimental evolution of cancer cells.

New Tool in Biophysics—Ancestral Sequence Reconstruction (ASR)

The reader should wonder why an evolutionary approach is useful to solve a mechanism of a modern-day, man-made molecule? Obviously Abl did not evolve to bind Gleevec and be "strangled" by it! Rather, Gleevec accidently took advantage of differences in kinase regulation created by divergent evolution. While kinases are similar in their turnover rates upon activation, they vary drastically in their regulatory mechanisms. Such evolution of regulation became necessary with the developing of multicellularity and increasingly complex signaling cascades. Although in the case of Gleevec phylogenetic considerations were not part of the design, and overlap between Gleevec's selectivity and evolution of regulation was coincidental, we propose that targeting the unique energy landscapes underlying the regulatory features of a kinase of interest can be a powerful strategy for developing new selective inhibitors.

Evolution is rooted in the most fundamental process of random mutations, and driven by selection for better fitness. In light of this, the weird link between Gleevec selectivity and

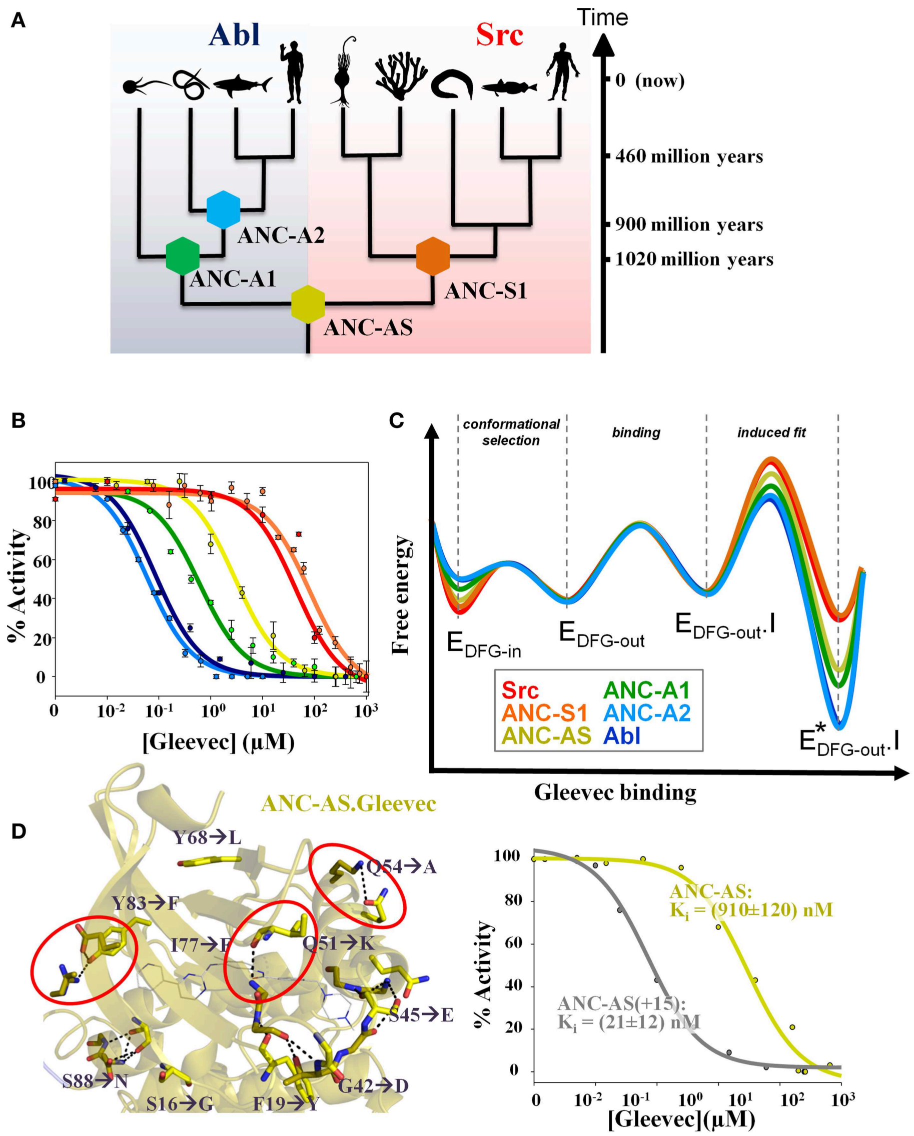


FIGURE 2 | Ancestral sequence reconstruction reveals the evolution of the energy landscape for Gleevec binding and identifies the residues responsible for Gleevec selectivity. (A) Phylogenetic tree of Abl and Src families showing the reconstructed nodes. Timeline indicates approximate age of the reconstructed ancestors. The corresponding sequences including the alignment are given in Wilson et al. (2015) **(B)** Inhibition constants K_i for each kinase were determined from the activity versus drug concentration profiles showing a gradual change in Gleevec affinity from the weak binder Src to the tight binder Abl via the intermediate binder ANC-AS. Same colors are used as defined in **(A)**. **(C)** Schematic representation of the evolution of the Gleevec binding energy landscape based on data from Wilson et al. (2015). The major difference between kinases is in the induced fit step. Conformational selection step provides a minor contribution and physical binding step is nearly identical in all kinases. **(D)** Substitution of only 15 residues in the N-terminal lobe of ANC-AS (resulting in ANC-AS(+15)) guided by ancestral sequence reconstruction, structure, and biochemical analysis (Wilson et al., 2015) results in dramatic increase in Gleevec affinity (right panel). Ten of the amino acid changes from ANC-AS into the corresponding residues in Abl are indicated by arrows. A subset of these identified mutations disrupt hydrogen bonds (shown as dotted lines) that are present in weak binders (some highlighted by red circles) leading to an increase in kinase flexibility for the strong binders thereby enabling an efficient induced fit step. Some panels in **Figures 1**, **2** are adapted from Agafonov et al. (2014) and Wilson et al. (2015).

evolution is actually not so far-fetched. Evolution as a result of chance shows itself in this story as a friend and foe: it led to the development of humans, but also to cancer and drug resistance. Using ASR to solve a modern cancer drugs mechanism is unorthodox, since until recently this method has been applied to recapitulate nature's paths to modern proteins with differential functions. Arguably the most famous ASR story has come from the Thornton lab in their successful inference of ancient corticoid receptors (Thornton et al., 2003; Ortlund et al., 2007; Bridgham et al., 2009). A story spanning over a half a dozen research papers not only shed light on the understanding of the different selectivity of modern steroid receptors for their corresponding hormones, but also answered some long standing questions in the field, including the role of epistasis in macromolecular evolution. Recently, ASR has leaped over to successfully recreate ancestral enzymes with reaction efficiencies near that of modern day enzymes (Perez-Jimenez et al., 2011; Hobbs et al., 2012; Ingles-Prieto et al., 2013; Risso et al., 2013; Boucher et al., 2014). Resurrection of enzymes has been an important step in validating the accuracy of ASR because of the need to maintain enzymatic activity, which is extremely sensitive to mutational change. These studies have largely focused on understanding changes in the enzyme's melting temperatures or underlying structural changes. The Gleevec story takes it to the next step to characterize the evolution of energy landscapes that ultimately underlies function.

"Tell Us What the Future Holds, So We May Know That You Are Gods" (Isaiah 41, 23)

How can understanding of Gleevec selectivity and the differential *evolution* of kinases guide the *creation* of better cancer drugs? We are not god, we do not have the ultimate answer. However the door to intelligent design of successful cancer therapeutics may have opened a little wider with recent advances in genome information including ASR and personal genomic profiling, characterization of free energy landscapes of the drug binding process to targets, advances in medicinal chemistry and computation.

The history of Gleevec research teaches us a number of lessons: First, the correct microscopic binding model (meaning the correct scheme), ideally with quantification of each step, is crucial. Slow progress in understanding Gleevec's selectivity was in large extent due to the overwhelming attention to the DFG-loop conformational selection model (Cowan-Jacob et al., 2005; Dar et al., 2008; Shan et al., 2009; Aleksandrov and Simonson, 2010; Lovera et al., 2012; Lin and Roux, 2013; Lin et al., 2014). Second, the physical binding step that has been the major focus in docking simulations is only one piece of the puzzle, and conformational changes are crucially linked to both affinity and selectivity (**Figure 2C**). Therefore, experimental and computational efforts should be more centered on the dynamics of the target and drug/target complex. Third, the trivial (simple laws of thermodynamics) but at the same time profound recognition that conformational change after binding (an induced fit step) delivers two essential components of a good drug: increased affinity and long drug residence times on the target (**Figure 2C**). In addition, it can provide excellent specificity particularly when such conformational changes involve elements remote from the binding site as seen in Abl-Gleevec. In contrast, conformational selection (ability of the apo protein to sample multiple conformations) by definition weakens the overall drug affinity by the fraction of the protein in the binding-incompetent states. While such a step can offer drug specificity, the new results suggest that DFG-loop conformational selection seems to play only a minor role for kinase selectivity due to the fact that the DFG-loop readily interconverts between states. We propose that induced fit steps are in play in many successful drugs leading to very tight binding and long on-target residence times. Finally, molecular dynamics simulations will play an increasing role in rational drug design, but such simulations need to be based on the solid foundation of biochemical research. In the case of Gleevec and other kinase inhibitors, future computational emphasize should be centered on dynamics of the enzyme/drug complex characterizing the induced fit step and not on the DFG-loop dynamics. Having the correct binding scheme established with corresponding structural information available, MD can sample the conformational space identifying new local minima and potentially cryptic or allosteric sites that are hard to trap experimentally if they are low-populated. If such states are unique for a particular kinase, they can be excellent targets for new specific inhibitors.

We are excited about the future prospect of a happy marriage between experiments and computation, and between basic academic research and pharmaceutical industry to tackle the very challenging but rewarding goal of designing perfect weapons against deadly diseases.

Acknowledgments

This work was supported by the Howard Hughes Medical Institute, the Office of Basic Energy Sciences, Catalysis Science Program, U.S. Dept. of Energy, award DE-FG02-05ER15699, and NIH (GM100966-01) to DK.

References

Acker, M. G., and Auld, D. S. (2014). Considerations for the design and reporting of enzyme assays in high-throughput screening applications. *Perspect. Sci.* 1, 56–73. doi: 10.1016/j.pisc.2013.12.001

ACS Chemical Biology. (2015). Special section: new frontiers in kinases reviews. *ACS Chem. Biol.* 10, 175–256

Agafonov, R. V., Wilson, C., Otten, R., Buosi, V., and Kern, D. (2014). Energetic dissection of Gleevec's selectivity toward human tyrosine kinases. *Nat. Struct. Mol. Biol.* 21, 848–853. doi: 10.1038/nsmb.2891

Aleksandrov, A., and Simonson, T. (2010). Molecular dynamics simulations show that conformational selection governs the binding preferences of imatinib for several tyrosine kinases. *J. Biol. Chem.* 285, 13807–13815. doi: 10.1074/jbc.M110.109660

Biophysical Chemistry. (2014). Special issue: conformational selection. *Biophys. Chem.* 186, 1–54.

Boucher, J. I., Jacobowitz, J. R., Beckett, B. C., Classen, S., and Theobald, D. L. (2014). An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *Elife* 3. doi: 10.7554/eLife.02304

Bridgham, J. T., Ortlund, E. A., and Thornton, J. W. (2009). An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* 461, 515–519. doi: 10.1038/nature08249

Capdeville, R., Buchdunger, E., Zimmermann, J., and Matter, A. (2002). Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat. Rev. Drug Discov.* 1, 493–502. doi: 10.1038/nrd839

Chance, B. (1940). The accelerated flow method for rapid reactions. *J. Franklin Inst.* 229, 737–766. doi: 10.1016/S0016-0032(40)90963-2

Cohen, M. H., Williams, G., Johnson, J. R., Duan, J., Gobburu, J., Rahman, A., et al. (2002). Approval summary for imatinib mesylate capsules in the treatment of chronic myelogenous leukemia. *Clin. Cancer Res.* 8, 935–942.

Cohen, P. (2002). Protein kinases–the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* 1, 309–315. doi: 10.1038/nrd773

Cohen, P., and Alessi, D. R. (2013). Kinase drug discovery–what's next in the field? *ACS Chem. Biol.* 8, 96–104. doi: 10.1021/cb300610s

Copeland, R. A. (2011). Conformational adaptation in drug-target interactions and residence time. *Future Med. Chem.* 3, 1491–1501. doi: 10.4155/fmc.11.112

Cowan-Jacob, S. W., Fendrich, G., Manley, P. W., Jahnke, W., Fabbro, D., Liebetanz, J., et al. (2005). The crystal structure of a c-Src complex in an active conformation suggests possible steps in c-Src activation. *Structure* 13, 861–871. doi: 10.1016/j.str.2005.03.012

Daley, G. Q., and Baltimore, D. (1988). Transformation of an interleukin 3-dependent hematopoietic cell line by the chronic myelogenous leukemia-specific P210bcr/abl protein. *Proc. Natl. Acad. Sci. U.S.A.* 85, 9312–9316. doi: 10.1073/pnas.85.23.9312

Dar, A. C., Lopez, M. S., and Shokat, K. M. (2008). Small molecule recognition of c-Src via the Imatinib-binding conformation. *Chem. Biol.* 15, 1015–1022. doi: 10.1016/j.chembiol.2008.09.007

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051. doi: 10.1038/nbt.1990

Depristo, M. A., Weinreich, D. M., and Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* 6, 678–687. doi: 10.1038/nrg1672

Gibson, Q. H., Kreuzer, F., Meda, E., and Roughton, F. J. W. (1955). The kinetics of human haemoglobin in solution and in the red cell at 37-Degrees-C. *J. Physiol.* 129, 65–89. doi: 10.1113/jphysiol.1955.sp005339

Harms, M. J., and Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* 14, 559–571. doi: 10.1038/nrg3540

Hobbs, J. K., Shepherd, C., Saul, D. J., Demetras, N. J., Haaning, S., Monk, C. R., et al. (2012). On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus. *Mol. Biol. Evol.* 29, 825–835. doi: 10.1093/molbev/msr253

Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J. M., Gaucher, E. A., et al. (2013). Conservation of protein structure over four billion years. *Structure* 21, 1690–1697. doi: 10.1016/j.str.2013.06.020

Kerns, S. J., Agafonov, R. V., Cho, Y. J., Pontiggia, F., Otten, R., Pachov, D. V., et al. (2015). The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* 22, 124–131. doi: 10.1038/nsmb.2941

Lemonick, M. D., and Park, A. (2001, May 28). New Hope For Cancer. *Time*.

Levinson, N. M., Kuchment, O., Shen, K., Young, M. A., Koldobskiy, M., Karplus, M., et al. (2006). A Src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol.* 4:e144. doi: 10.1371/journal.pbio.0040144

Lin, Y. L., Meng, Y., Huang, L., and Roux, B. (2014). Computational study of Gleevec and G6G reveals molecular determinants of kinase inhibitor selectivity. *J. Am. Chem. Soc.* 136, 14753–14762. doi: 10.1021/ja504146x

Lin, Y. L., and Roux, B. (2013). Computational analysis of the binding specificity of Gleevec to Abl, c-Kit, Lck, and c-Src tyrosine kinases. *J. Am. Chem. Soc.* 135, 14741–14753. doi: 10.1021/ja405939x

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Systematic validation of protein force fields against experimental data. *PLoS ONE* 7:e32131. doi: 10.1371/journal.pone.0032131

Lovera, S., Sutto, L., Boubeva, R., Scapozza, L., Dolker, N., and Gervasio, F. L. (2012). The different flexibility of c-Src and c-Abl kinases regulates the accessibility of a druggable inactive conformation. *J. Am. Chem. Soc.* 134, 2496–2499. doi: 10.1021/ja210751t

Meng, Y., Lin, Y. L., and Roux, B. (2015). Computational Study of the "DFG-Flip" conformational transition in c-Abl and c-Src tyrosine kinases. *J. Phys. Chem. B* 119, 1443–1456. doi: 10.1021/jp511792a

Nagar, B., Hantschel, O., Young, M. A., Scheffzek, K., Veach, D., Bornmann, W., et al. (2003). Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell* 112, 859–871. doi: 10.1016/S0092-8674(03)00194-6

Newsweek. (2001, May 27). A Cure For Cancer? *Newsweek*.

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., and Thornton, J. W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 317, 1544–1548. doi: 10.1126/science.1142819

Pauling, L., and Zuckerkandl, E. (1963). Chemical paleogenetics molecular restoration studies of extinct forms of life. *Acta Chem. Scand.* 17:9. doi: 10.3891/acta.chem.scand.17s-0009

Perez-Jimenez, R., Ingles-Prieto, A., Zhao, Z. M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., et al. (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* 18, 592–596. doi: 10.1038/nsmb.2020

Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* 100, L47–L49. doi: 10.1016/j.bpj.2011.03.051

Richter, D. J., and King, N. (2013). The genomic and cellular foundations of animal origins. *Annu. Rev. Genet.* 47, 509–537. doi: 10.1146/annurev-genet-111212-133456

Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A., and Sanchez-Ruiz, J. M. (2013). Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian beta-lactamases. *J. Am. Chem. Soc.* 135, 2899–2902. doi: 10.1021/ja311630a

Rowley, J. D. (1973). Letter: a new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* 243, 290–293. doi: 10.1038/243290a0

Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B., and Kuriyan, J. (2000). Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science* 289, 1938–1942. doi: 10.1126/science.289.5486.1938

Seeliger, M. A., Nagar, B., Frank, F., Cao, X., Henderson, M. N., and Kuriyan, J. (2007). c-Src binds to the cancer drug imatinib with an inactive Abl/c-Kit conformation and a distributed thermodynamic penalty. *Structure* 15, 299–311. doi: 10.1016/j.str.2007.01.015

Shan, Y., Seeliger, M. A., Eastwood, M. P., Frank, F., Xu, H., Jensen, M. O., et al. (2009). A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proc. Natl. Acad. Sci. U.S.A.* 106, 139–144. doi: 10.1073/pnas.0811223106

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

Stockbridge, R. B., and Wolfenden, R. (2009). The intrinsic reactivity of ATP and the catalytic proficiencies of kinases acting on glucose, N-acetylgalactosamine, and homoserine: a thermodynamic analysis. *J. Biol. Chem.* 284, 22747–22757. doi: 10.1074/jbc.M109.017806

Thornton, J. W., Need, E., and Crews, D. (2003). Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301, 1714–1717. doi: 10.1126/science.1086185

Treiber, D. K., and Shah, N. P. (2013). Ins and outs of kinase DFG motifs. *Chem. Biol.* 20, 745–746. doi: 10.1016/j.chembiol.2013.06.001

Vajpai, N., Strauss, A., Fendrich, G., Cowan-Jacob, S. W., Manley, P. W., Grzesiek, S., et al. (2008). Solution conformations and dynamics of ABL kinase-inhibitor complexes determined by NMR substantiate the different binding modes of imatinib/nilotinib and dasatinib. *J. Biol. Chem.* 283, 18292–18302. doi: 10.1074/jbc.M801337200

Vogtherr, M., Saxena, K., Hoelder, S., Grimme, S., Betz, M., Schieborr, U., et al. (2006). NMR characterization of kinase p38 dynamics in free and ligand-bound forms. *Angew. Chem. Int. Ed Engl.* 45, 993–997. doi: 10.1002/anie.200502770

Wade, N. (2001, May 8). Powerful anti-cancer drug emerges from basic biology. *The New York Times.*

Wang, Q., Zorn, J. A., and Kuriyan, J. (2014). A structural atlas of kinases inhibited by clinically approved drugs. *Methods Enzymol.* 548, 23–67. doi: 10.1016/B978-0-12-397918-6.00002-1

Wilson, C., Agafonov, R. V., Hoemberger, M., Kutter, S., Zorba, A., Halpin, J., et al. (2015). Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* 347, 882–886. doi: 10.1126/science.aaa1823

Xu, W., Harrison, S. C., and Eck, M. J. (1997). Three-dimensional structure of the tyrosine kinase c-Src. *Nature* 385, 595–602. doi: 10.1038/385595a0

Conflict of Interest Statement: DK is the inventor on a patent applied for by Brandeis University that describes a biophysical platform for drug development based on energy landscapes. 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.

REVIEW
published: 19 May 2015
doi: 10.3389/fmolb.2015.00023

Fairy “tails”: flexibility and function of intrinsically disordered extensions in the photosynthetic world

Gabriel Thieulin-Pardo, Luisana Avilan, Mila Kojadinovic and Brigitte Gontero *

UMR 7281, Centre National de la Recherche Scientifique, Aix-Marseille Université, Marseille, France

Intrinsically Disordered Proteins (IDPs), or protein fragments also called Intrinsically Disordered Regions (IDRs), display high flexibility as the result of their amino acid composition. They can adopt multiple roles. In globular proteins, IDRs are usually found as loops and linkers between secondary structure elements. However, not all disordered fragments are loops: some proteins bear an intrinsically disordered extension at their C- or N-terminus, and this flexibility can affect the protein as a whole. In this review, we focus on the disordered N- and C-terminal extensions of globular proteins from photosynthetic organisms. Using the examples of the A_2B_2-GAPDH and the α Rubisco activase isoform, we show that intrinsically disordered extensions can help regulate their “host” protein in response to changes in light, thereby participating in photosynthesis regulation. As IDPs are famous for their large number of protein partners, we used the examples of the NAC, bZIP, TCP, and GRAS transcription factor families to illustrate the fact that intrinsically disordered extremities can allow a protein to have an increased number of partners, which directly affects its regulation. Finally, for proteins from the cryptochrome light receptor family, we describe how a new role for the photolyase proteins may emerge by the addition of an intrinsically disordered extension, while still allowing the protein to absorb blue light. This review has highlighted the diverse repercussions of the disordered extension on the regulation and function of their host protein and outlined possible future research avenues.

Keywords: intrinsically disordered proteins, GAPDH, CP12, Rubisco activase, cryptochromes, transcription factors

OPEN ACCESS

Edited by:
Peter Tompa,
Flanders Institute of Biotechnology (VIB), Belgium

Reviewed by:
Anastassios Papageorgiou,
University of Turku, Finland
Esko Oksanen,
European Spallation Source, Sweden

Correspondence:
Brigitte Gontero,
Centre National de la Recherche Scientifique-BIP, 31 Chemin Joseph Aiguier, 13 402 Marseille, France
bmeunier@imm.cnrs.fr

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

Received: *25 February 2015*
Accepted: *04 May 2015*
Published: *19 May 2015*

Citation:
Thieulin-Pardo G, Avilan L, Kojadinovic M and Gontero B (2015) Fairy “tails”: flexibility and function of intrinsically disordered extensions in the photosynthetic world. Front. Mol. Biosci. 2:23. doi: 10.3389/fmolb.2015.00023

Introduction

Proteins occupy a central position in the architecture and functioning of living matter. A major objective of protein biochemistry consists in explaining the physiological functions of these molecules by means of structural studies, also known as the “structure-function” relationship. Among others, X-ray crystallography is a powerful tool to solve macromolecular three-dimensional 3D structures. However, some proteins cannot be crystallized because they are fully disordered or possess disordered parts that are missing in the electron density map of the crystals. In the 1990s, Sedzik and Kirschner (1992) attempted to crystallize the myelin basic protein (MBP), the predominant extrinsic protein in both central and nervous system myelins. After several attempts, the authors concluded that MBP adopts a random coil conformation and that as long as its flexibility was not suppressed, it was not possible to obtain crystals (Sedzik and Kirschner, 1992). MBP was one of the first examples of many other “un-crystallizable” proteins. These proteins, originally named Intrinsically Unstructured Proteins (IUPs), are nowadays termed Intrinsically Disordered Proteins (IDPs) (Wright and Dyson, 1999; Dunker et al., 2001, 2008a,b). In 1998, Romero et al. showed that 15 000 proteins from the Swiss-Prot database contain one or more Intrinsically disordered regions (IDRs) comprising more than 40 amino acid residues (Romero et al., 1998). It was shown later that despite their lack of well-defined 3D structure, many partially or completely disordered proteins are functional (Wright and Dyson, 1999; Dunker et al., 2001, 2008a,b; Tompa, 2002). In the late 1990s, studies of disordered yet functional proteins emerged as a new research field, extending the traditional paradigm to include a more comprehensive view of protein structure-function (Wright and Dyson, 1999; Dunker et al., 2001; Tompa, 2002; Dunker et al., 2008a,b). In the past, different models have been proposed to explain protein functioning, and protein flexibility has appeared as a key point (Fersht, 1998). Among these models, the "induced-fit" model (Koshland et al., 1966) introduced the idea that protein conformational changes could be triggered upon ligand binding. These notions were applied to IDPs, and many of them were shown to undergo an "induced-folding" upon binding to their partners (Dunker et al., 2002). Short motifs called MoREs (Molecular Recognition Elements) are often involved in the interaction, involving disorder to order transitions (Fuxreiter et al., 2004, 2007; Oldfield et al., 2005; Mohan et al., 2006; Vacic et al., 2007; Hazy and Tompa, 2009). However, the idea that preformed binding elements exist before the binding, and even in the absence of a partner, led to the "conformational selection" model. In some cases, the IDP is not fully structured in the presence of its partner and the term "fuzziness" was coined by Fuxreiter and Tompa to describe such complexes (Tompa and Fuxreiter, 2008; Hazy and Tompa, 2009; Fuxreiter and Tompa, 2012). The flexibility of IDPs increases the chance of their polypeptide chains adopting the right conformations in the presence of their partners. Furthermore, the high ratio of hydrophilic residues in IDPs facilitates initial contacts with their partners. The interactions are also stronger with IDPs: their lack of structure or absence of rigidity increasing association constants (Dunker et al., 2001; Meszaros et al., 2007; Chouard, 2011). The ability of the IDPs to adopt multiple conformations allows the same region to adapt to different binding sites in many "induced-fits" and thus to have multiple partners (Uversky et al., 2000; Tompa, 2002; Uversky, 2002, 2011; Meszaros et al., 2007; Carmo-Silva and Salvucci, 2013). The discovery of IDPs and their singular lack of definite structure brought nuances to the "structure-function" dogma, showing that the same structure, or lack of one, could have multiple partners and thus multiple functions (Wright and Dyson, 1999; Dunker et al., 2001; Tompa, 2002; Uversky, 2002, 2011; Meszaros et al., 2007; Sun et al., 2013). In this regard, IDPs could be seen as the "master keys" of the protein-protein interaction network.

The ability of IDPs to bind to multiple partners makes them naturally good regulators, as they can modulate the activity of several proteins in a coordinated way (Dunker et al., 2001, 2002, 2005; Gavin et al., 2002, 2006; Haynes et al., 2006; Patil and Nakamura, 2006; Mittag et al., 2010; Uversky, 2010; Pancsa and Tompa, 2012). Therefore, multiple areas of the cellular response can be affected by a single signal allowing IDPs to play a major role in regulatory pathways (Tompa, 2002; Dunker et al., 2005; Haynes et al., 2006; Uversky, 2010; Pancsa and Tompa, 2012). The flexibility of IDPs can also be modified by the cellular environment or by post-translational modifications. IDPs and IDRs are often targets of different post-translational modifications (the most common being phosphorylation, methylation and ubiquitination) which can radically affect their affinity for their partners and their stability, thus multiplying the possibilities for a fine-tuned regulation (Tompa, 2002; Dunker et al., 2005; Haynes et al., 2006; Uversky, 2010; Pancsa and Tompa, 2012). These particularities make IDPs the hubs in a vast net of protein-protein interactions (Gavin et al., 2002, 2006). They carry out basic functions such as regulation of metabolic pathways, transcription, translation or cellular signal transduction; they can act as scavengers of toxic molecules and they play a key role in the assembly of multi-protein complexes (Uversky, 2011). Moreover, their roles in several diseases of major medical interest, such as cancer (Castillo et al., 2014; Saha et al., 2014; Xue et al., 2014), Alzheimer's disease (Uversky, 2009; Kovacech and Novak, 2010; Salminen et al., 2011; Karagoz and Rudiger, 2015) prion disease (Tompa, 2009; Uversky, 2009; Breydo and Uversky, 2011) or Parkinson's disease (Uversky, 2009; Hazy et al., 2011; Breydo et al., 2012; Alderson and Markley, 2013) have been extensively studied (Babu et al., 2011).

While the discovery and characterization of IDPs and IDRs is a rapidly growing, and an increasingly recognized, area of protein science, (Tompa, 2002; Uversky, 2010; Uversky and Dunker, 2010; Chouard, 2011), little information is available photosynthetic organisms, where IDPs have been described as central players in many responses such as biotic and abiotic stress, development, metabolism regulation, or adaptation to oxic atmosphere (Kragelund et al., 2012; Pancsa and Tompa, 2012; Yruela and Contreras-Moreira, 2012, 2013; Pietrosemoli et al., 2013; Sun et al., 2013; Panda and Ghosh, 2014). Published data mainly concern *Arabidopsis thaliana*, a higher plant model with one of the best-annotated sequenced genomes (Arabidopsis Genome Initiative, 2000). Yet, the recent analysis of 12 plant genomes revealed that the occurrence of disorder in plants is similar to the one found in other eukaryotes (Bracken et al., 1999; Yruela and Contreras-Moreira, 2012, 2013; Sun et al., 2013). An *in silico* analysis of plant nuclear proteomes suggested a higher disorder in the internal part of nuclear-encoded plant proteins rather than at their extremities, in contrast to the chloroplast- and mitochondrion-encoded proteomes (Yruela and Contreras-Moreira, 2012). This is also pointed by studies on prokaryotes showing that the IDRs may be more frequent at the extremities of the proteins that act as "molecular shields" such as chaperones (Krisko et al., 2010; Chakrabortee et al., 2012).

In this review, we describe several globular proteins with N- or C-terminal IDR extensions in photsynthetic organisms, as opposed to entirely disordered proteins or globular proteins containing one or more IDRs in the middle of their sequences. The aim of this work is not to give an exhaustive list of the roles undertaken by such disordered extensions, as this has recently been reviewed (Uversky, 2013). Instead, we focus on globular proteins or domains that acquired their disordered tails during evolution, using examples from photosynthetic organisms. The addition of a disordered extension to a globular protein created new regulation opportunities, making these proteins responsive to environmental factors through self-regulation, post-translational modifications or new protein-protein interactions. We illustrate the impact of disordered extensions by describing proteins involved in photosynthetic metabolism and regulation of gene expression (**Table 1**).

GAPDH and CP12

As knowledge about proteins progressed, new results appeared showing that some enzymes were able to carry out more than

TABLE 1 | Summary of the characteristics of the Intrinsically disordered extensions presented in this review.

	Protein/Protein family	Protein role	Position of the extension	Role (or possible role) of the disordered extension	Particularities of the disordered extension*	References
Metabolism regulation	GAPDH GapB	Enzyme from the Benson-Calvin cycle	C-terminus	Autonomous redox regulation of the GAPDH activity	Pair of redox-sensitive cysteine residues	Cerff, 1979; Brinkmann et al., 1989; Baalmann et al., 1996; Li and Anderson, 1997; Scagliarini et al., 1998; Sparla et al., 2002; Petersen et al., 2006
	Rubisco ActIvase (α isoform)	Activator of the Rubisco enzyme	C-terminus	Possible redox regulation of the RCA activity	Pair of redox-sensitive cysteine residues	Werneke et al., 1989; Shen and Ogren, 1992; Zhang and Portis, 1999; Portis, 2003; Henderson et al., 2011; Stotz et al., 2011; Carmo-Silva and Salvucci, 2013; Gontero and Salvucci, 2014
Gene regulation	NAC family (No Apical Meristem, ATAF, Cup Shaped Cotyledon)	Transcription factors	C-terminus	Regulation of the NAC transcription factor domain through protein-protein interactions	Presence of multiple MoREs (conserved within a subfamily)	Ooka et al., 2003; Olsen et al., 2005; Jensen et al., 2010a,b; Kjaersgaard et al., 2011
	bZIP family (basic Leucine Zipper)	Transcription factors	N-terminus	Regulation of the bZIP transcription factor domain and its stability through protein-protein interactions and post-translational modifications	Presence of multiple MoREs (conserved within a subfamily) Phosphorylation sites	Ang et al., 1998; Campbell et al., 2000; Hardtke et al., 2000; Moreau et al., 2004; Yoon et al., 2006; Sun et al., 2013
	TCP family [Teosinte branched 1 (tbl), Cycloidea (cyc) and Proliferating Cell Factor	Transcription factors	N-terminus and C-terminus	N-terminus: Binding of target DNA C-tenninus: TCP self-association and regulation	N-terminus: Induced-fit binding of DNA C-terminus: Coiled-coil self association	Viola et al., 2011, 2012; Steiner et al., 2012; Valsecchi et al., 2013
Signaling	GRAS family [Gibberellic Acid Insensitive (GAI). Repressor of Gai (RGA) and Scarecrow (SCR)]	Transcriptional co-activator	N-terminus	Regulation of the GRAS activator domain and its stability through protein-protein interactions and post-translational modifications	Presence of multiple MoREs (conserved within a subfamily) Phosphorylation sites	Triezenberg, 1995; Czikkel and Maxwell, 2007; Sun et al., 2010, 2011, 2012, 2013
	Cryptochiomes	Light-signaling Control of circadian and annual cyles	C-terminus	Protein-protein interaction upon captation of blue light: Initiation of developpemental responses	Presence of MoREs Multiple phosphorylation sites	Lin and Shalitin, 2003; Green, 2004; Partch et al., 2005; Yu et al., 2010; Chaves et al., 2011; Liu et al., 2011a,b

All the extensions present features of "disorder": enrichment in hydrophylic charged amino acids and few hydrophobic residues.

one function in a single polypeptide chain and were classified as multifunctional proteins (Kirschner and Bisswanger, 1976). In many cases, the dual function resulted from the fusion of two genes that initially encoded different proteins. Later on, the term "moonlighting" (Jeffery, 1999) categorized proteins that have different functions. The glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a well-known moonlighting enzyme and has at least ten distinct, confirmed non-enzymatic activities apart from its enzymatic function (Sirover, 1999, 2011; Hildebrandt et al., 2015). The GAPDHs constitute a large and diverse family of dehydrogenases universally represented in living organisms. They catalyze the reductive dephosphorylation of 1, 3-bisphosphoglyceric acid (BPGA) producing glyceraldehyde-3-phosphate (GAP) and inorganic phosphate using NAD(P)H as a cofactor (Trost et al., 2006). Glycolytic GAPDHs(also named GapC) are NAD-specific and mainly found in the cytosol, but in land plants a second type of glycolytic GAPDH named GapCp is targeted to the chloroplast(Petersen et al., 2003; Marri et al., 2005; Munoz-Bertomeu et al., 2010). Both GapC and GapCp are NAD-specific and form homotetramers *in vivo* that are not subject to complex regulatory mechanisms. However, in photosynthetic organisms, another GAPDH catalyzes the unique reductive step of the Benson-Calvin cycle is present and uses both NADH and NADPH with a marked preference for NADPH (Falini et al., 2003).

Like all GAPDHs, the NADPH-dependent GAPDH is made up of two functional domains, one corresponding to the catalytic domain (residues 148–313 in spinach GAPDH) and the other one being the cofactor-binding domain, or Rossman fold (residues 1–147 and 313–334, respectively). The latter contains a flexible, arginine-rich region called the S-loop (Fermani et al., 2001). In higher plants, this GAPDH exists in different forms such as a heterotetramer made up of two GapA and two GapB subunits (A_2B_2), a homotetramer made up of four GapA subunits, and as a hexadecamer (A_8B_8) (Baalmann et al., 1994, 1995; Scheibe et al., 2002; Howard et al., 2011a,b). The GapB subunit is similar to the GapA subunit but bears a C-terminal extension which has a regulatory function (Cerff, 1979; Brinkmann et al., 1989; Baalmann et al., 1996; Li and Anderson, 1997; Scagliarini et al., 1998; Sparla et al., 2002) (**Figure 1A**). This subunit is thought to derive from a gene duplication

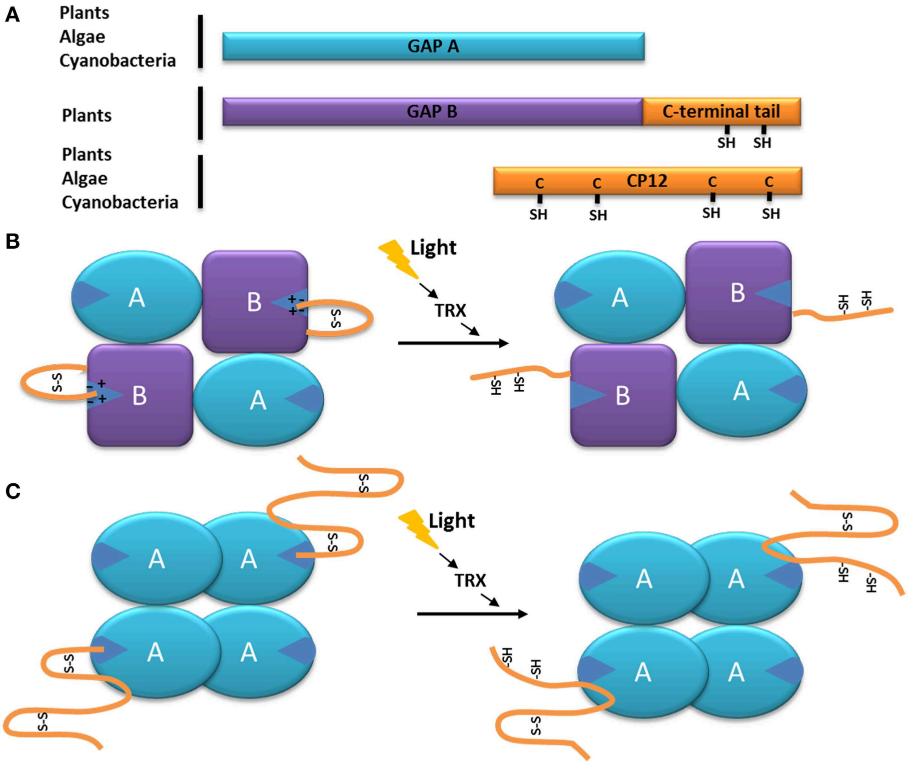


FIGURE 1 | Model of the function of the C-terminal extension of the GAPDH subunit GapB. (A) Schematic organization of the GapB from higher plants and *O. tauri* (Robbens et al., 2007) subunit compared to the GapA and CP12 proteins. The C-terminal extension of GapB is homologous to the C-terminal of CP12 and present two regulatory cysteine residues. **(B)** Schematic representation of the autonomous redox regulation of the A_2B_2-GAPDH. When oxidized, the C-terminal extension of the GapB subunit presents a disulfide bridge, which places the C-terminal amino acids inside the active site of the enzyme, resulting in its inhibition. The disulfide bridge can be reduced by the thioredoxin *f* (TRX), and the enzyme becomes active. **(C)** Schematic representation of the autonomous redox regulation of the A_4-GAPDH by CP12. When oxidized, the C-terminal part of the CP12 protein presents a disulfide bridge, which places its C-terminal amino acids inside the active site of GAPDH, resulting in its inhibition. The disulfide bridge can be reduced by the thoioredoxin *f* (TRX) or DTT and the enzyme becomes active.

event which occurred near the origin of Streptophyta (which include charophytes and land plants) (Petersen et al., 2006) and might be a construct of a GapA moiety fused at the C-terminus with the C-terminal half of the CP12, a 8.2–8.5 kDa Chloroplast Protein (Pohlmeyer et al., 1996). The portion of CP12 acquired by GapB subunits confers redox properties to GapB-containing GAPDH (Pupillo and Piccari, 1973; Pupillo and Giuliani Piccari, 1975; Wolosiuk and Buchanan, 1976, 1978; Trost and Pupillo, 1993; Baalmann et al., 1994, 1995; Sparla et al., 2002). Cyanobacteria, primitive photosynthetic eukaryotes (including the glaucophyte *Cyanophora paradoxa*), and red and green algae (except charophytes and sister lineages) seem to contain exclusively the GapA subunit (Petersen et al., 2006; Trost et al., 2006). However, the small prasinophyte green alga *Ostreococcus tauri* has been shown to possess GapA and GapB (Robbens et al., 2007).

CP12 is a protein of about 80 amino acid residues that was originally described by Pohlmeyer et al. (1996) and has been found in most photosynthetic organisms (Groben et al., 2010; Gontero and Maberly, 2012; Gontero and Salvucci, 2014; Lopez-Calcagno et al., 2014). The CP12 proteins show high primary sequence variability, in particular at the N-terminus. However, they share some remarkable common features.CP12 proteins have an amino acid composition poor in order-promoting residues although they contain cysteine residues (Groben et al., 2010), and behave abnormally under gel electrophoresis and size exclusion chromatographies (Gontero and Maberly, 2012) suggesting that they are IDPs. Moreover, recent data from fluorescence correlation spectroscopy (FCS) show that the hydrodynamic radius of CP12 from the green alga *Chlamydomonas reinhardtii* is large compared to that expected for globular proteins of this molecular mass (Moparthi et al., 2014). The cysteine residues are involved in the formation of disulfide bridges and peptide loops and are found as pairs at the C-terminus and/or at the N-terminus. When oxidized, CP12 proteins may present α helices maintained by the N-terminal disulfide bridge (Graciet et al., 2003a; Gardebien et al., 2006). The algal CP12 is a key component of a supra-molecular complex controlling the activity of the Benson-Calvin cycle by regrouping several key enzymes of the cycle, including GAPDH, phosphoribulokinase (PRK) and fructose 1,6-bisphosphate aldolase (FBP aldolase). Within the ternary GAPDH/CP12/PRK complex, both enzymes are strongly inhibited (Avilan et al., 1997; Lebreton et al., 1997; Graciet et al., 2003a; Erales et al., 2008; Marri et al., 2008). CP12 forms a fuzzy complex with the green alga *C. reinhardtii* A_4-GAPDH, as revealed by EPR studies (Mileo et al., 2013) (see the Minireview by Lebreton et al. in this Topic Research). The ternary complex has also been found in the cyanobacterium *Synechococcus elongatus* and in the higher plant *A. thaliana*. The A_4-GAPDH-CP12 sub-complex from these organisms have been crystallized but in both complexes, the first 50 amino acid residues were not visible in the density map, consistent with a high flexibility of this region in the crystal (Avilan et al., 2000; Matsumura et al., 2011; Fermani et al., 2012). Recently, it was also observed using FCS and FRET (Förster Resonance Energy Transfer) that the algal CP12 flexibility is not abolished by its interactions with either GAPDH or with PRK (Moparthi et al., 2014, 2015).

In the case of the GapB subunit of GAPDH from higher plants, the C-terminal end of the protein (*ca* 30 residues)is strongly homologous to the C-terminal part of CP12 (Pohlmeyer et al., 1996; Trost et al., 2006; Groben et al., 2010) (**Figure 1A**).The regulation of the *A. thaliana* A_2B_2-GAPDH activity by the C-terminal extension of the GapB subunit is now very well understood: upon oxidation (which happens during the day-night transition) the two cysteine residues of the CP12-like tail form a disulfide bridge that places the C-terminal ultimate glutamate residue (E362) inside the active site (stabilized by the electrostatic interactions with an arginine residue R77 involved in the NADP cofactor binding). Consequently, the NADPH cofactor is not able to enter the catalytic site and thus NAPDH-dependant A_2B_2-GAPDH activity is inhibited (Sparla et al., 2005; Fermani et al., 2007) (**Figure 1B**). In contrast, during the night-day transition, the disulfide bridge maintaining the C-terminal extension into the active site is reduced by thioredoxin *f*, thereby releasing the CP12-like tail and resulting in A_2B_2-GAPDH activity (**Figure 1B**) (Sparla et al., 2002; Trost et al., 2006; Fermani et al., 2007). This mechanism is very similar to the one observed in *C. reinhardtii* between the homotetrameric A_4-GAPDH and free CP12, where the penultimate glutamate (E79) of the CP12 interacts with the arginine residue R82 of A_4-GAPDH (**Figure 1C**) (Trost et al., 2006; Erales et al., 2011; Avilan et al., 2012). The reduction of the GAPDH-CP12 by dithiothreitol (DTT) in the alga results in a more active NADPH-GAPDH as a consequence of the rupture of disulfide bridges on CP12. Of interest, DTT, *in vitro* mimicks thioredoxins *in vivo* and it has been shown that CP12 can be reduced by thioredoxin *f* in the light (Marri et al., 2009).

In the higher plant, *A. thaliana* and in the green alga, *C. reinhardtii*, the stoichiometry of the oxidized A_4-GAPDH-CP12 sub-complex is two CP12 molecules for one A_4-GAPDH (Marri et al., 2008; Kaaki et al., 2013), while four CP12 molecules interact with each GAPDH tetramer in the cyanobacterium, *S. elongatus* (Matsumura et al., 2011). When interacting with CP12, A_4-GAPDH activity decreased by two-fold (in the case of the *C. reinhardtii* proteins, the catalytic constant k_{cat} of the free enzyme was $430 \pm 17\ s^{-1}$, and became $251 \pm 9\ s^{-1}$ in the presence of CP12), suggesting that only two of the four active sites were blocked (Graciet et al., 2003b) (**Figure 1C**). The same observation was made for the *A. thaliana* A_2B_2-GAPDH: upon oxidation, the A_2B_2-GAPDH activity decreased by 2-fold (its k_{cat} changed from $59 \pm 19\ s^{-1}$ to $27 \pm 10\ s^{-1}$) although its catalytic constant in a reduced state ($k_{cat} = 59 \pm 19\ s^{-1}$) was comparable to the one of the free A_4-GAPDH ($k_{cat} = 61 \pm 4\ s^{-1}$) (Sparla et al., 2004). The regulation of the plant A_2B_2-GAPDH and of the algal A_4-GAPDH-CP12 complex is thus very similar. With the addition of the C-terminal extension within the GapB subunit, the A_2B_2-GAPDH has become autonomously redox-regulated, a property that was previously provided through interaction with CP12.

Although the appearance of the GapB subunit represents an important step in the evolution of the redox control of the Calvin-Benson cycle enzymes, this new autonomous regulation co-exists with the CP12-based one in higher plants (Scheibe et al., 2002), and a A_2B_2-GAPDH-PRK complex entirely devoid of CP12 has yet to be identified. The presence of CP12 is likely to be required for the assembly of larger supramolecular complex, and in *C. reinhardtii*, A_4-GAPDH-CP12-PRK was shown to interact with the aldolase (Erales et al., 2008). In this regard, one may wonder how this system will continue to evolve, and if more enzymes of the Benson-Calvin cycle will also acquire similar CP12-like disordered extensions, possibly meaning that the CP12 protein will become redundant. However, CP12 seems to be a part of numerous other processes in photosynthetic organisms (Singh et al., 2008; Howard et al., 2011a,c; Stanley et al., 2013), so it is unlikely to disappear completely from higher plants genomes in the future.

Rubisco Activase

Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) is the enzyme that catalyzes the formation of two molecules of phosphoglyceric acid using one molecule of ribulose 1,5-bisphosphate (RuBP) and one of carbon dioxide (CO_2). As the primary CO_2 acceptor of most photoautotrophic organisms, Rubisco can represent up to half of soluble proteins in higher plants, and is believed to be the most abundant protein on Earth (Ellis, 1979; Losh et al., 2013; Raven, 2013). Rubisco from most photosynthetic organisms, including plants and cyanobacteria, is a very large protein (550 kDa), composed of large (L, 52 kDa) and small (S, 12 kDa) subunits arranged as a L_8S_8 hexadecamer. For this enzyme to be active, a lysine residue inside the Rubisco active site (K201 in *Nicotiana tobacum*) must be carbamylated and bind a Mg^{2+} ion (Lorimer et al., 1976; Andersson and Backlund, 2008). The addition of the non-catalytic CO_2 molecule to the active site is a spontaneous process, but the presence of RuBP or other sugar-phosphate at the active site decreases the carbamylation efficiency of Rubisco and thus its activity (Lorimer et al., 1976; Cleland et al., 1998). The Rubisco activases (RCAs) exhibit ATPase activity and were first characterized for their ability to promote the carbamylation of such RuBP-inhibited Rubisco (Portis et al., 1986). With time, it became clear that the RCAs allowed the CO_2 to enter the active site of Rubisco by removing the hindering RuBP or its analog, carboxyarabinitol bisphosphate (CABP) (**Figure 2A**) (Portis et al., 2008). The presence of RCAs allows Rubisco to function at its maximal

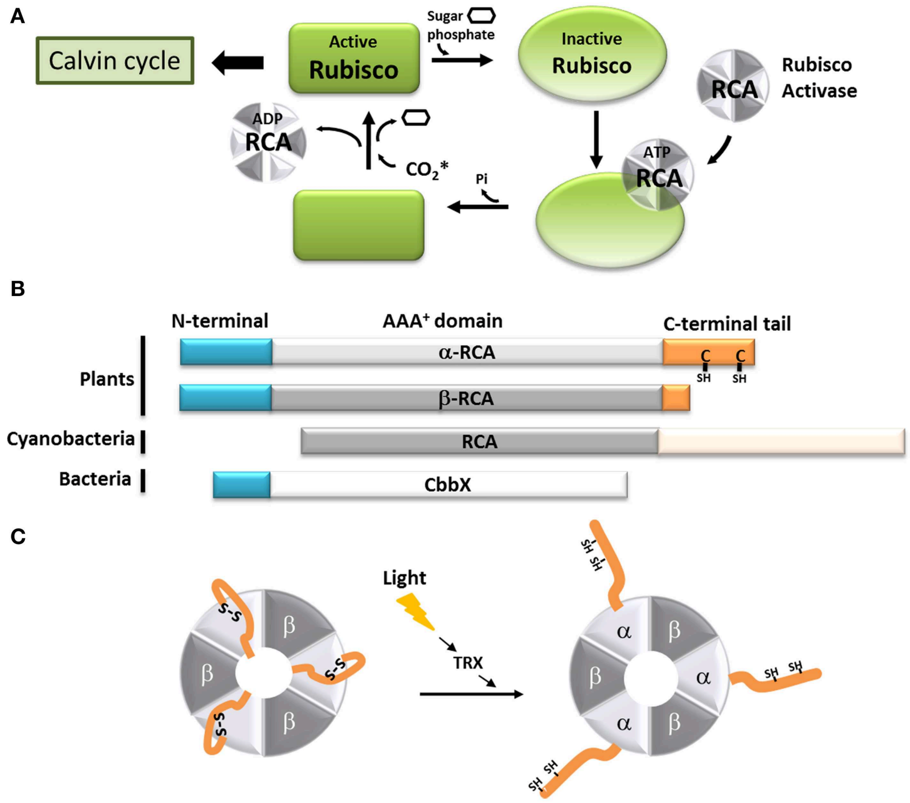


FIGURE 2 | Model of the function of the C-terminal extension of the α-Rubisco activase. (A) Schematic representation of the Rubisco activase reaction with Rubisco. Only the carbamylated Rubisco is active and can participate to the Calvin-Benson cycle. However, the presence of sugar phosphate in the Rubisco active site prevents is carbamylation. Using ATP, the Rubisco activase facilitates the departure of the inhibiting sugar phosphate and promotes the Rubisco carbamylation. The CO_2* represents the non-substrate molecule carbamylating Rubisco. **(B)** Schematic organization of the α and β Rubisco activases subunits from plants, β-cyanobacteria and CbbX. **(C)** Schematic representation of the light-dependent redox regulation of the $\alpha_3\beta_3$ Rubisco activase. When oxidized, the C-terminal extension of the α-RCA subunit bears a disulfide bridge, which places the C-terminal amino acids inside the nucleotide-binding site of the protein, resulting in its inhibition. The disulfide bridge can be reduced by the thioredoxin *f* (TRX), and Rubisco becomes active.

capacity in sub-optimal CO_2 concentration that would normally not permit carbamylation *in vivo* (Portis et al., 1986). In higher plants, RCAs, as expected, are mostly present in the parts of plants involved in photosynthesis (Watillon et al., 1993; Liu et al., 1996), and their expression follows a daily cycle that is regulated by external factors like light and temperature (Martino-Catt and Ort, 1992; Watillon et al., 1993; Liu et al., 1996; To et al., 1999).

In most organisms from the green lineage, two isoforms of RCA are found: an α isoform of 45–46 kDa and a β isoform of 41–43 kDa (Werneke et al., 1989; Rundle and Zielinski, 1991; To et al., 1999; Gontero and Salvucci, 2014). The only difference between the two RCA isoforms is the presence of a short C-terminal extension (*ca* 30 amino acid residues depending on the species) on the α isoform (**Figure 2B**). Both the α and β isoforms were found in *Arabidopsis thaliana*, spinach, and rice, although only one RCA gene is present (Werneke et al., 1988; To et al., 1999); in these species, the presence of the two RCA isoforms is the result of alternate splicing (Werneke et al., 1989; Rundle and Zielinski, 1991; To et al., 1999). On the other hand, other species like barley, cotton, maize and tobacco have multiple RCA genes (Rundle and Zielinski, 1991; Qian and Rodermel, 1993; Salvucci et al., 2003; Yin et al., 2010). In most cases, these organisms have separate genes coding for α and β RCAs without alternative splicing (Rundle and Zielinski, 1991), although all the genes identified in tobacco and cucumber appear to only encode the β isoform (Portis, 2003). To the best of our knowledge, the C-terminus of the α and β RCA isoforms were never tested for intrinsic disorder. Using several disorder predictors, including MeDor (Lieutaud et al., 2008) and MFDp2 (Mizianty et al., 2014), we were able to determine that the end of both the C-terminal part of the α and β RCAs (*ca* 50 residues for the α isoform and 20 residues for the β isoform) seem to be intrinsically disordered, including the entire C-terminal extension of the α RCA (**Figure 3**). The most remarkable features of this disordered tail are the two cysteine residues (C392 and C411 in the *A. thaliana* protein), that are highly conserved among the α RCA isoforms (Zhang and Portis, 1999).

The crystal structure of the tobacco β RCA was recently solved at 3 Å (Stotz et al., 2011), showing that RCA proteins are functional doughnut-shaped hexamers displaying an AAA^+ fold, as was predicted using other AAA^+ proteins (ATPases involved in a multitude of processes, Neuwald et al., 1999 as templates Portis, 2003). Interestingly, the last 23 residues of the protein were absent from the structure, indicating that this part of the molecule is very flexible and can adopt different conformations. Moreover, the substrate recognition site of the RCA from the creosote bush, *Larrea tridentata*, was solved at the atomic level (Henderson et al., 2011). Unfortunately, the structural studies were performed only on the β RCA isoform. As the α RCA core is identical, its structure should not be different from the β RCA, but the α RCA was shown to form functional $\alpha_n\beta_n$ heteromers rather than α_n homomers (Crafts-Brandner et al., 1997; Zhang et al., 2001). In the light of these new structural data, we can suppose that α RCA can form heterohexamers $\alpha_3\beta_3$ (**Figure 2C**). These structural data also show the presence of three loops containing hydrophilic amino acid residues lining the surface of a central pore. Site-directed mutations introduced in this part of the proteins severely diminished the Rubisco activation and the ATP hydrolysis by the RCA proteins (Stotz et al., 2011), confirming that this region is implicated in the binding of ATP (Salvucci et al., 1993; Li et al., 2006). Based on this information, the model that has been proposed for RCA interaction with Rubisco includes one face of the flat hexamer interacting with the surface of the Rubisco, while an exposed loop of the Rubisco protein could fit into the central hole. Minor conformational changes of this Rubisco loop, allowed by the ATP hydrolysis, would then be transmitted to Rubisco allowing the inhibiting RuBP

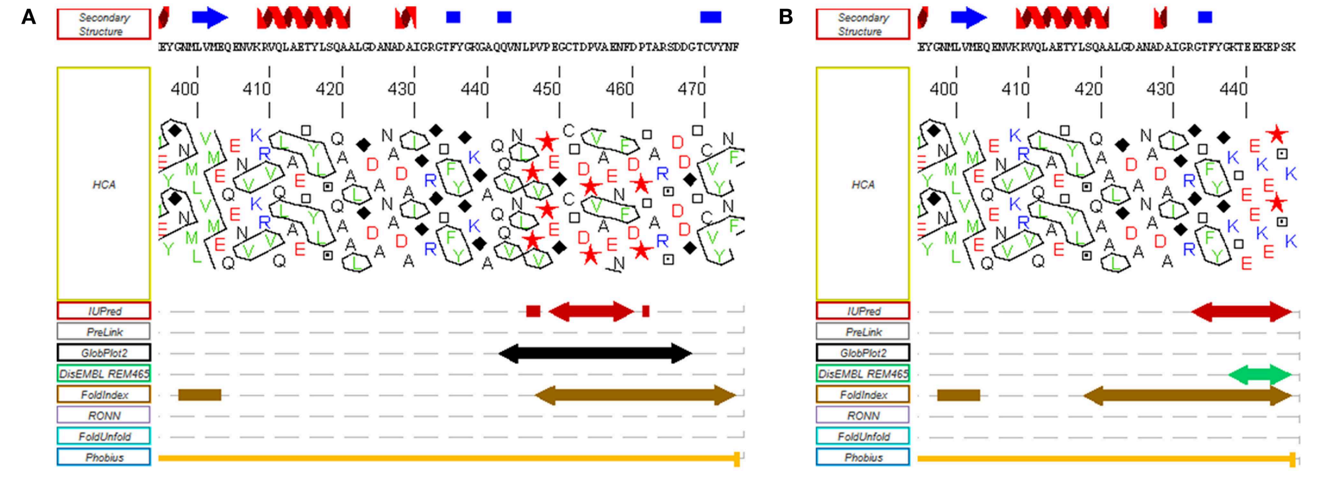


FIGURE 3 | Disorder predictions of the C-terminal regions of the α and β RCA proteins from *A. thaliana*. MeDor (http://www.vazymolo.org/MeDor/) graphical output of the C-terminal part of the α **(A)** and β **(B)** Rubisco activase isoforms from *A. thaliana*. Predicted secondary structure elements and the HCA plot, are shown above and below the amino acid sequence, respectively. Arrows below the HCA plot correspond to regions of predicted disorder (Lieutaud et al., 2008).

to be released and Rubisco to be carbamylated (Stotz et al., 2011).

The activity of the α and β RCAs is classically described to be dependent on the ATP:ADP ratio (Zhang and Portis, 1999; Carmo-Silva and Salvucci, 2013), and is extremely sensitive to high temperatures (Portis, 2003; Salvucci, 2004). Moreover, the activity of the α RCA is regulated by light (Mächler and Nösberger, 1980; Perchorowicz et al., 1981). This observation is linked to the action of thioredoxin *f* on the two cysteine residues present on its C-terminal extension (Shen and Ogren, 1992; Zhang and Portis, 1999; Zhang et al., 2001, 2002; Portis, 2003; Wang and Portis, 2006).A site-directed mutagenesis study (Shen and Ogren, 1992) showed that the substitution of only one of the two cysteine residues was enough to abolish the light regulation of α RCA, implicating the involvement of a disulfide bridge. Several studies showed that the mechanism of inhibition involves the blocking of the ATP-binding region by the C-terminal extension upon oxidation. This self-inhibition would be stabilized by strong electrostatic forces between the negatively-charged tail and the positively charged nucleotide site (Shen and Ogren, 1992; Zhang and Portis, 1999; Zhang et al., 2001, 2002; Wang and Portis, 2006; Carmo-Silva and Salvucci, 2013) (**Figure 2C**). It was also observed that the β RCA, although devoid of regulatory cysteine residues, could be light-regulated in the presence of the α isoform (Zhang and Portis, 1999; Zhang et al., 2001). In the hypothesis that RCAs form $\alpha_3\beta_3$ heterohexamers, we can assume that the combined bulk of C-terminal extensions efficiently inhibit the whole complex in dark conditions (**Figure 2C**). The RCA activity can be restored consequently by the reduction of the C-terminal disulfide bridge by the thioredoxin *f*, which occurs upon dark-light transitions (Carmo-Silva and Salvucci, 2013). In this case, the acquisition of a C-terminal tail, originally by alternate splicing, has allowed the RCA protein to fine-tune the activity of Rubisco in function of the light availability in addition to the energetic state of the cell.

In other photosynthetic organisms, the Rubisco activase system is different or works in a different way. In β-cyanobacteria, the RCA protein has the same main domains as plant RCAs, but lacks the N-terminal domain necessary for Rubisco activation found in plants and green algae (Van De Loo and Salvucci, 1996; Li et al., 1999; Stotz et al., 2011; Gontero and Salvucci, 2014; Mueller-Cajar et al., 2014). This could explain why no Rubisco activation has been observed using cyanobacterial RCAs (Li et al., 1999; Pearce, 2006). The latter also possess a very long (180 residues) intrinsically disordered C-terminal extension that seems to target the protein to the carboxysomes (Zarzycki et al., 2013) (**Figure 2B**). Organisms from the red lineage (α-proteobacteria, rhodophyta, heterokontophyta, etc.) do not have exactly the same Rubisco as the green lineage, and the so-called “Red Rubisco” has a slightly longer large subunit. These organisms do not have RCA genes, but the same Rubisco activase function is carried out by another protein, CbbX (Pearce, 2006; Gontero and Salvucci, 2014) (**Figure 2B**). The crystal structure of CbbX has recently been solved, showing that this protein is organized in hexamers arranged in a very comparable manner to green RCAs (Mueller-Cajar et al., 2011). It was also suggested that CbbX mechanisms are based on the same principles as the one of RCA, with the C-terminus of the large Rubisco subunit inserted into the central hole of CbbX (Mueller-Cajar et al., 2011). It should be noted that CbbX seems to have an IDR at its C-terminus, but its implications in CbbX activity is yet to be studied.

Rubisco activase is not the only “friendly” protein involved in the regulation of Rubisco, since other proteins are needed during its assembly and folding, including the cpn60 chaperone, which also has a disordered C-terminal tail (Goloubinoff et al., 1989; Cloney et al., 1992; Libich et al., 2013).

Three Transcription Factor Families (NAC, bZIP and TCP)

Disordered regions are ideal for proteins coordinating regulatory events and as such, transcription factors participating in regulation and signaling functions are enriched in IDRs.

The NAC family (named after No Apical Meristem, ATAF, Cup-Shaped Cotyledon) is one of the largest families of plant-specific transcription factors (Ooka et al., 2003; Olsen et al., 2005; Rushton et al., 2008; Sun et al., 2013). These family members are involved in a very large variety of processes, including plant development (Olsen et al., 2005), biotic and abiotic stress responses (Jensen et al., 2010b; Seo and Park, 2010; Seo et al., 2010) and leaf senescence (Kjaersgaard et al., 2011). The NAC transcription factors usually contain two domains: the N-terminal NAC domain and the C-terminal extremity domain (**Figure 4A**). The NAC domain is mainly conserved and well-ordered, displaying a typical structure comprising α helices flanking one β strand (Ernst et al., 2004). This domain binds the consensus DNA sequence CGT(GA) (Olsen et al., 2005). The C-terminal domain of the NAC proteins is highly variable within the family; however, some motifs in the C-terminus may display a sub-family-specific conservation (Jensen et al., 2010a). The C-terminal domains composition reveals a very high percentage of hydrophilic (Asp, Glu, Ser, Thr) and proline (Pro) residues, whereas the proportion of hydrophobic and aromatic residues is very low (Olsen et al., 2005; Jensen et al., 2010a). These specificities are typical of IDRs, and the C-terminal domain of some NAC proteins was experimentally characterized as an IDR (Jensen et al., 2010a,b). Despite this IDR feature, some hydrophobic and/or aromatic residues are present in this domain; interestingly, these amino acid residues are often conserved among a subfamily (Jensen et al., 2010a,b; Kjaersgaard et al., 2011). The IDR C-terminal domains of the NAC proteins are predicted to contain MoREs that are conserved in sub-families (Jensen et al., 2010a). It has been experimentally confirmed that these particular residues are very important to the specific function of each sub-group in the NAC family, and are essential to activation mechanisms often involving many different partners (Ooka et al., 2003; Ernst et al., 2004; Taoka et al., 2004; Olsen et al., 2005; Ko et al., 2007; Jensen et al., 2010a) (**Figure 4B**).

The bZIP (basic Leucine Zipper) transcription factors family is ubiquitous and is one of the largest families of transcription factors in eukaryotes. bZIP transcription factors take part

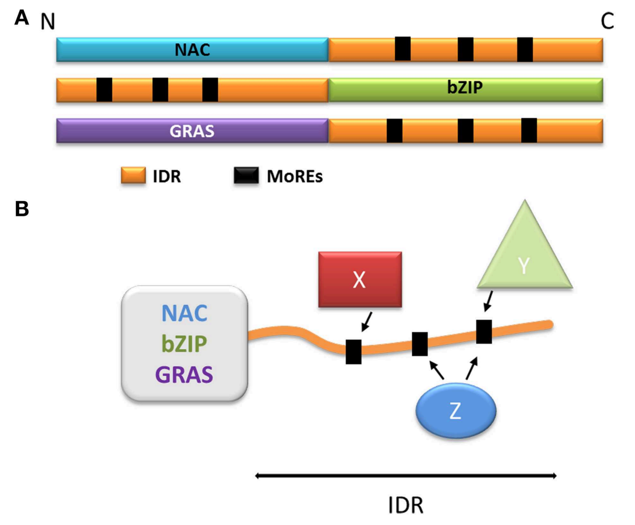


FIGURE 4 | Model of the function of the disordered extremities of the NAC, bZIP, and GRAS proteins. (A) Schematic organization of the NAC, bZIP, and GRAS protein families. The disordered parts are schematized in orange, the MoREs are represented as black squares. **(B)** Schematization of the multiple protein-protein interactions involving the disordered extremity. X, Y, and Z represent different protein partners capable of interacting with one or more MoREs and regulate the behavior of the globular domain.

in a multitude of regulatory pathways such as development, metabolism, circadian rhythm and response to stress (Sun et al., 2013). The bZIP proteins are composed of two domains: a C-terminal bZIP domain and a N-terminal activation domain (**Figure 4A**). The C-terminal bZIP domain gives its name to the family and displays large patches of basic residues and leucine zipper motifs (Ellenberger et al., 1992; Vinson et al., 1993). The leucine zipper regions are organized in α helices and are responsible for the dimerization of the proteins through the formation of a coiled-coil structure (Vinson et al., 1993; Yoon et al., 2006), while the basic regions bind to the DNA molecule (Ellenberger et al., 1992). Interestingly, the basic regions have been described either as fully ordered, very flexible or intrinsically disordered depending on the protein (Bracken et al., 1999; Podust et al., 2001; Moreau et al., 2004; Yoon et al., 2006). When bound to DNA, the basic regions have however been observed as α helices, suggesting that the interaction triggers folding in response to a specific DNA motif (Hollenbeck et al., 2002), illustrating once more the disorder to order transition (induced-fit). The N-terminal regions of bZIP proteins act as regulators (Ang et al., 1998; Sun et al., 2013), and are mostly intrinsically disordered (Campbell et al., 2000; Moreau et al., 2004; Yoon et al., 2006; Sun et al., 2013). These regions typically contain different MoREs, and their flexibility allows the interaction with multiple partners, again by adopting different secondary structures (Ang et al., 1998; Campbell et al., 2000; Oldfield et al., 2005; Yoon et al., 2006) (**Figure 4B**). Through these activating or inhibiting interactions, transcription of the genes targeted by bZIP proteins is effectively modulated in response to several signals. The N-terminal disordered domain also modulates the activity of bZIP transcription factors through post-translational modifications, and phosphorylation in particular. In plants, bZIP transcription factors can be phosphorylated in response to illumination, which disrupts the interactions between the bZIP proteins and their activating partners (Ciceri et al., 1997; Hardtke et al., 2000). The phosphorylated proteins also have lower affinity for their DNA targets, resulting in a decrease of gene activation (Ciceri et al., 1997; Hardtke et al., 2000). Interestingly, some bZIP proteins also display IDRs in their C-terminal domain. In the case of bZIP28 (initially a transmembrane protein), these IDRs are exposed to the lumen of the endoplasmic reticulum and allow the interaction, through MOREs with BIP, the majority reticulum chaperone. In response to stress, bZIP28 is relocated to the Golgi and the cytoplasmic domain is detached, allowing it to enter the nucleus and to control gene expression (Srivastava et al., 2013, 2014).

A recent study on TCP8, a transcription factor belonging to the TCP [Teosinte branched 1 (tb1), Cycloidea (cyc) and Proliferating Cell Factor (PCF)] family, showed the presence of three IDRs, two of them at the N- and C-terminal extremities (Valsecchi et al., 2013). While the N-terminus binds DNA in an induced-fit mechanism, the C-terminal region is involved in the TCP protein self-association in a coiled-coil structure (Valsecchi et al., 2013). Furthermore, it seems that different transcription factors from the TCP family can interact, modulating the response of different pathways to multiple stimuli (Baier and Latzko, 1975; Viola et al., 2011, 2012; Steiner et al., 2012; Valsecchi et al., 2013).

As illustrated in these examples, the disordered tails of transcription factors have an essential role in modulating their activities through protein-protein interactions with a wide range of activators and inhibitors. Moreover, these extensions are often prone to phosphorylation and constitute another level of regulation. Together, these IDRs form a complex signaling web, turning the transcription factors into hubs and allowing the genes involved in adaptive responses to be finely regulated.

GRAS Family

The GRAS family comprises proteins involved in numerous aspects of plant development and growth. This large family is named after its first members, Gibberellic Acid Insensitive (GAI), Repressor of Gai (RGA) and Scarecrow (SCR), and its members are mostly related to signaling in response to phytohormones [gibberellic acid (GA), auxin, brassinosteroids] and biotic and abiotic stress (Bolle, 2004; Sun et al., 2011). The GRAS family proteins are composed of one variable N-terminal region and a commonly conserved C-terminal GRAS domain (**Figure 4A**), and are divided into ten subfamilies based on phylogeny (Bolle, 2004; Tian et al., 2004; Lim et al., 2005; Sanchez et al., 2007; Sun et al., 2011). The conserved GRAS domain (*ca* 380 residues depending on the subfamilies) acts as a transcriptional co-activator (Heery et al., 1997) through leucine-rich motifs. GRAS domains typically contain two leucine-rich motifs, which are needed for specific protein-protein interactions(Cui et al., 2007; Vacic et al., 2007; Fode et al., 2008; Hirsch and Oldroyd, 2009; Hirsch et al., 2009; Hou et al., 2010). The GRAS proteins interact with a large number of nuclear proteins, most of which are transcription factors, thereby modulating their target activity (Hirsch and Oldroyd, 2009; Hirsch et al., 2009; Hou et al., 2010; Sun et al., 2012).

In contrast to the highly conserved GRAS domain, the N-terminal domains of the GRAS family proteins display a rich diversity at the sequence level, although the N-terminus is conserved within subfamilies (Sun et al., 2010, 2011, 2012, 2013). Moreover, these N-terminal domains have recently been identified as intrinsically disordered (Sun et al., 2010, 2011, 2012, 2013). Interestingly, patches of repeated hydrophobic and/or aromatic residues are found in the N-terminal region (Triezenberg, 1995; Sun et al., 2011). These patches are arranged in conserved motifs within subfamilies (Triezenberg, 1995; Sun et al., 2011), and are involved in specific multiple protein-protein interactions (Sun et al., 2010, 2012). In the case of the DELLA subfamily which has been intensively studied, the N-terminal domain can interact with the gibberellic acid receptor GIB1, but only when GIB1 has bound its ligand (Murase et al., 2008; Hirano et al., 2010; Sun et al., 2010, 2012). Moreover, each DELLA protein domain (N-terminal and C-terminal domains) can interact with several partners, making these proteins a hub at the center of the gibberellic acid response pathway. Other examples of GRAS proteins are important in other regulatory pathways, although subfamilies are always specialized in a precise type of stimulus (phytohormones, biotic and abiotic stress, etc…) (Sun et al., 2010, 2011, 2012, 2013).

A common feature of the GRAS proteins is their ability to acquire a structure when bound to a partner, unlike the fuzzy GAPDH/CP12 complex (Mileo et al., 2013). As mentioned above, MoREs are present in GRAS proteins; each one was predicted to occur within the N-terminal domains, and more specifically in the elements conserved within subfamilies, strengthening the idea that these motifs are the key to the specificity of GRAS proteins (Sun et al., 2011, 2012). In the case of the DELLA subfamily, the presence of the MoREs has been verified experimentally (Sun et al., 2010, 2011, 2012). Interestingly, the N-terminal domain of the GRAS proteins is also the target of phosphorylation, which again introduces another way to fine-tune the regulation of these proteins (Fu et al., 2002; Iakoucheva et al., 2004; Hussain et al., 2007; Mittag et al., 2010). Phosphorylation of the N-terminal domain is directly linked to the activity of the GRAS proteins, modulating the affinity of the N-terminus for its partners, and having a direct effect on the GRAS proteins stability through the control of their degradation (Day et al., 2004; Hussain et al., 2005; Itoh et al., 2005; Czikkel and Maxwell, 2007).

When considering the GRAS family as a whole, it is remarkable how conserved the GRAS domains and patterns are, while the N-terminal domains are highly variable. It seems that the addition of a disordered protein segment to the GRAS domain has increased its number of partners, and thus turned it into a signal-integration hub involved in many different pathways. On the other hand, one could consider that the addition of GRAS domains to pre-existing IDPs involved in the phytohormonal and/or stress responses has allowed these IDPs to control, even more directly, the cellular responses by acting on gene expression.

Cryptochrome

Cryptochromes are a group of proteins in which most members have an intrinsically disordered C-terminal tail that can have a profound impact on their overall function. Together with the photolyases, these proteins belong to the photolyase/cryptochrome family (Lin and Shalitin, 2003; Sancar, 2004; Chaves et al., 2006; Ozturk et al., 2007; Fortunato et al., 2015).

Photolyases are ancient enzymes that use blue light to catalyze the repair of DNA lesions caused by ultraviolet light. Lesions such as cyclobutane pyrimidine dimers (CPD) and pyrimidine-pyrimidone photoproducts are repaired by photolyases CPD and by photolyases 6–4, respectively. Photolyase capacity to use blue light is due to the presence of two chromophores: a photoantenna pterin (5,10-methenyltetrahydrofolateor a-hydroxy-5-deazaflavin) and flavin adenine dinucleotide (FAD). During the DNA repair, the two chromophores cofactors absorb blue photons and initiate splitting of the cyclobutane ring by a mechanism involving reactive radicals (Liu et al., 2011b).

Cryptochromes, the other group of proteins in the photolyase/cryptochrome family, have a photolyase homologous region (called PHR) and a C-terminal tail (**Figure 5A**) (Yu et al., 2010). Cryptochromes are able to absorb blue light in a very similar way to the photolyases. Another group within this family includes DASH-type cryptochromes named after the *Drosophila*, *Arabidopsis*, *Synechocystis* and Human. Members of this group are closer to photolyases than to cryptochromes, and are able to repair single-stranded DNA (Chaves et al., 2011) and may also have N-terminal and C-terminal disordered extensions.

In contrast to photolyases, cryptochromes do not have the ability to repair DNA. However, in many organisms, the absorption of photons by the chromophores in the photolyase homologous region of these proteins, induces conformational change (through electron transfer and subsequent phosphorylation), which in turn trigger specialized signaling events through protein-protein interactions (Liu et al., 2011b). It has been shown that the function of cryptochromes resides mainly within their C-terminal tails (Yang et al., 2000; Green, 2004; Chaves et al., 2006, 2011; Yu et al., 2010). Interestingly, this tail is poorly conserved among groups of organisms. In *Arabidopsis*, two cryptochromes are present, CRY1 and CRY2, that have different C-terminal extensions although a DAS motif is found in both (Lin and Shalitin, 2003). The length of the C-terminal tail in cryptochromes of animals, plants and some unicellular organisms varies from 30 to 250 residues and, as mentioned above, is intrinsically disordered. This characteristic has been established by sequence analysis, biochemical methods such as analysis of the sensitivity to protease cleavage, and physical methods such as circular dichroism and nuclear magnetic resonance (NMR)on recombinant C-terminal extensions of both *Arabidopsis* and human cryptochromes (Partch et al., 2005). Comparison of the proteolysis susceptibility between full-length cryptochromes and their C-terminal tail showed that this tail interacts with the photolyase domain, causing it to adopt a tertiary structure. The susceptibility to proteolysis of the C-terminal tail of the CRY1

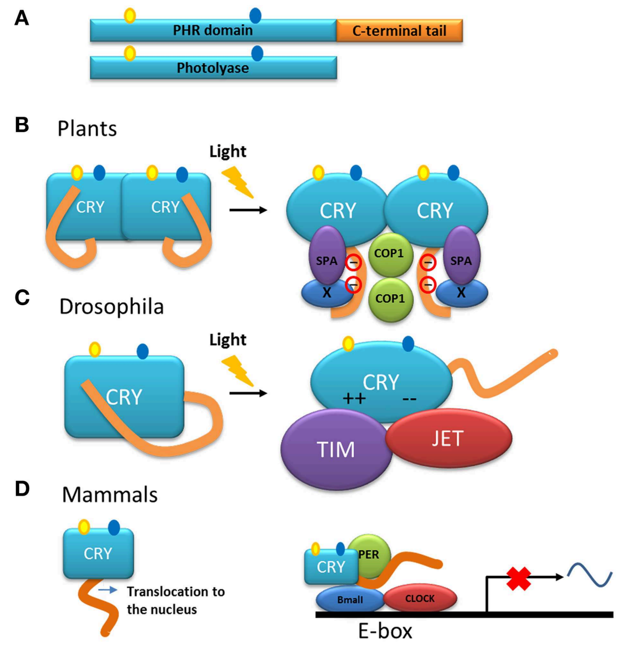


FIGURE 5 | Models of the function of the C-terminal tail of cryptochromes. (A) Representation of cryptochrome (CRY) and photolyase. Cryptochromes have a photolyase-homologous region (PHR) and a C-terminal tail. The chromophore molecules of the PHR are shown. **(B)** Model of the action mechanism of cryptochromes from Arabidopsis. After absorption of light, the C-terminal tail is phosphorylated and a change in conformation is triggered in the entire molecule. The C-terminal tail is exposed at the surface of the protein and as a consequence interactions with partner proteins such as COP1 and SPA are induced (Liu et al., 2011a,b). **(C)** In darkness, the C-terminal tail of the cryptochrome from *Drosophila* inhibits the binding of the proteins involved in the circadian rhythm. After illumination, the inhibition by the tail is released and the PHR domain interacts through electrostatic interaction with the protein partners TIM and JET (Green, 2004; Czarna et al., 2013). **(D)** In mammals, cryptochrome is necessary for the translocation of the protein into the nucleus in which it is part of the core of the transcription/translation feedback that controls the circadian clock together with the proteins PER, BMAL, and CLOCK.

from *A. thaliana* increases after illumination, which is consistent with a conformational change (Partch et al., 2005). Indeed, the crystal structure of the complete cryptochrome from *Drosophila* confirmed that the C-terminal tail stays in a groove of the photolyase domain and mimics the recognition of photolyases with DNA (Zoltowski et al., 2011; Czarna et al., 2013).

In plants, cryptochromes play a role, together with other photoreceptors, in a variety of functions. In general, the cryptochromes of plants are involved in mechanisms that respond to blue light and their action has been explored in the inhibition of the elongation of hypocotyls, in the photoperiodic induction of flowering, in the circadian clock as in animals, and in other functions (Yu et al., 2010; Chaves et al., 2011; Liu et al., 2011b).These studies have been mainly performed in the model plant *A. thaliana*. Studies using transgenic plants overexpressing the C-terminal tail of CRY1 or CRY2, fused with β-glucuronidase (GUS) showed a constitutive morphogenic phenotype similar to that produced by blue light (Yang et al., 2000), indicating that, in the cryptochrome molecule, the C-terminal tail is responsible for the light-induced function. Moreover, NC80, an 80-residues segment present in the *Arabidopsis* protein, is responsible for the function of the C-terminal tail of CRY2 (Yu et al., 2007). The C-terminal tail of these proteins interacts with other proteins such as COP1 (constitutive photomorphogenic 1) (Wang et al., 2001; Yang et al., 2001), a multifunctional E3 ubiquitin ligase, and SPA1 (suppressor of phytochrome A 1) (Zuo et al., 2011; Liu et al., 2011a). This interaction is part of the initial steps for the light signaling and mechanisms to modulate the developmental process in the plant either by: (1) modulation of gene transcription or (2) suppression of proteolysis of regulators involved in development (i.e., flowering) (Liu et al., 2011a,b). Models have been proposed to explain the mode of action of plant cryptochromes (Lin and Shalitin, 2003; Partch et al., 2005; Yu et al., 2007; Liu et al., 2011a). In general, in these models, the photolyase domain and the C-terminal tail form a closed conformation in the dark. Upon illumination, an open and active conformation is adopted and, in this new conformation, the C-terminal tail is exposed allowing its interaction with other proteins to initiate signaling (**Figure 5B**). A model of action that includes dimerization and light dependent-phosphorylation that explains the exposure of the C-terminal tail as a result of charge repelling has also been proposed (**Figure 5B**) (Lin and Shalitin, 2003; Yu et al., 2007).

Although cryptochromes of plants are involved as photoreceptors in the circadian cycle, the molecular role of cryptochromes in relation to this cycle has been more elucidated in *Drosophila*. In this organism, the cryptochrome modulates the central oscillator, or clock, through the light-dependent interaction with the protein Timeless (TIM) (Busza et al., 2004), one of the components of the clock core. This interaction favors the degradation of both TIM and the cryptochrome itself, thus triggering the light/dark cycle each day by synchronization of the clock with the environment. The protein Jetlag (JET), an E3 ligase, also binds to the cryptochrome in a light-dependent manner and is responsible for the ubiquitination and subsequent proteolysis of both the cryptochrome and TIM (Peschel et al., 2009). In this case, and in contrast with the cryptochromes in *Arabidopsis*, the binding of the cryptochrome from *Drosophila* to its partners is performed by the photolyase domain of the protein (**Figure 5C**), whereas in the dark, the C-terminal tail inhibits this binding determining thus the photosensitivity of the circadian clock (Busza et al., 2004; Green, 2004).

In contrast to their homologs from plants and *Drosophila*, where the disordered C-terminal tail is used for light signaling, mammalian cryptochromes are light–independent transcriptional repressors in the core of the circadian clock (**Figure 5D**). Mammalian cryptochromes repress transcription processes that are dependent on the protein complex BMAL/CLOCK (Sancar, 2004; Chaves et al., 2011). In the case of these cryptochromes, the function of the C-terminal tail is more complex: (i) it is involved in the nuclear localization of the protein and (ii) with the photolyase domain, it also has a role in the interaction with other components of the clock such as BMAL (Chaves et al., 2006). Interestingly, the C-terminal tail also contributes to the circadian period length, since its phosphorylation affects the level of the protein, either promoting its own degradation in the case of CRY 2 (Harada et al., 2005) or stabilizing the protein as for CRY 1 (Gao et al., 2013).

It has been proposed that cryptochromes have evolved several times independently as an example of convergent evolution (Green, 2004). Only small changes have occurred in the photolyase domain, this part of the protein being conserved among cryptochromes and photolyases. One possible mechanism to explain the acquisition of C-terminal extensions in existing proteins would be through gene fusion (Marsh and Teichmann, 2010). If this mechanism had taken place at the origin of cryptochromes, it would suggest that proteins related to the C-terminal tail of cryptochromes already existed independently and had a function of their own. These independent domains became later associated to a photolyase domain providing them with the capacity to detect light. As mentioned above, the plant cryptochromes C-terminal domain is active and has the information needed to achieve signaling (Yang et al., 2000). During evolution, this protein could have fused with a duplicate of photolyase. In this hypothesis, the addition of the light-dependent photolyase module might be a way to adjust the physiology of the organisms to their environment through light perception. This could therefore be seen as an IDP having acquired a globular extension. Since in plants, a motif (DAS) within the C-terminal tail is conserved, it has been proposed that the ancestral plant cryptochrome emerged from a fusion of a photolyase with a protein containing the DAS motif (Lin and Shalitin, 2003). Another hypothesis that could explain the acquisition of the C-terminal tail in cryptochromes is by gene extension into a non-coding region (Marsh and Teichmann, 2010). The photolyase gene could thus have been extended through junk DNA. Analysis of phylogenetic relationships of gene families in animals showed that extension of an existing gene by "exonization" of a previous non-coding region seems to be an important evolutionary strategy to add a C-terminal disordered extension to proteins (Buljan et al., 2010). The high variability and different functions of the C-terminal tail of cryptochromes among plants and animals are in accordance with this hypothesis. Studies on the origin and evolution of the C-terminal tail of cryptochromes will give insights into the adaptation of organisms to light.

Conclusion

Within the present review, we tried to demonstrate the central and multiple roles of intrinsically disordered tails carried by certain globular proteins. Describing several examples of proteins displaying IDRs in photosynthetic organisms, we discussed how IDRs impact on both the functions and mechanisms of action of their "host" proteins. The examples of the A_2B_2-GAPDH and the α-Rubisco activase isoform show that their C-terminal disordered extensions participate in the light-dependent redox regulation of the photosynthetic metabolism. The cases of the multiple transcription factors with a disordered tail are very similar yet very different. In the few examples listed here, the disordered region plays a major role in the regulation of the DNA-binding domain through protein-protein interactions or post-translational modifications. Their sensitivity to a large number of signals allows the activity of the transcription factors to be modulated according to many factors (one to many), turning these proteins into hubs in a large signaling web. Lastly, the cryptochrome family is a prime example of a disordered extension changing the fundamental function of the initial photolyase into a light-dependent signaling protein, conserving the ability to absorb blue light and repurposing it.

The examples presented here are but a few of the multitude of proteins that have acquired a disordered extension (Uversky, 2013), although most examples do not usually come from the photosynthetic world. We can expect that in the years to come, an increasing number of these proteins will be identified. A great question that remains is how these proteins originated. While in some cases, the addition of an IDR seems to be quite recent like the GapB subunit. In other cases, this addition might be very ancient as in the NAC, bZIP, and GRAS families, in which there are multiple disordered extensions families that may derive from multiple fusion events, or a long succession of duplications followed by diverging evolution of the subfamilies. We hope that the expansion of the IDP field in general and specifically, the one involved in "green" biochemistry, will 1 day answer these questions.

Acknowledgments

Studentship (GT) was given by Ecole Normale Supérieure, Cachan. This work (BG) was supported by the Centre National de la Recherche Scientifique (CNRS) and Aix-Marseille Université. LA got a fellowship from ANR-13-BIME-0001-02. We would like to thank Pr. Carla Green for her kind help. We also thank Pr. Stephen Maberly for editing the english.

References

Alderson, T. R., and Markley, J. L. (2013). Biophysical characterization of alpha-synuclein and its controversial structure. *Intrinsically Disord. Proteins* 1, 18–39. doi: 10.4161/idp.26255

Andersson, I., and Backlund, A. (2008). Structure and function of Rubisco. *Plant Physiol. Biochem.* 46, 275–291. doi: 10.1016/j.plaphy.2008.01.001

Ang, L. H., Chattopadhyay, S., Wei, N., Oyama, T., Okada, K., Batschauer, A., et al. (1998). Molecular interaction between COP1 and HY5 defines a regulatory switch for light control of *Arabidopsis* development. *Mol. Cell* 1, 213–222. doi: 10.1016/S1097-2765(00)80022-2

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. doi: 10.1038/35048692

Avilan, L., Gontero, B., Lebreton, S., and Ricard, J. (1997). Memory and imprinting effects in multienzyme complexes–I. Isolation, dissociation, and reassociation of a phosphoribulokinase-glyceraldehyde-3-phosphate dehydrogenase complex from *Chlamydomonas reinhardtii* chloroplasts. *Eur. J. Biochem.* 246, 78–84. doi: 10.1111/j.1432-1033.1997.00078.x

Avilan, L., Lebreton, S., and Gontero, B. (2000). Thioredoxin activation of phosphoribulokinase in a bi-enzyme complex from *Chlamydomonas* *reinhardtii* chloroplasts. *J. Biol. Chem.* 275, 9447–9451. doi: 10.1074/jbc.275.13.9447

Avilan, L., Puppo, C., Erales, J., Woudstra, M., Lebrun, R., and Gontero, B. (2012). CP12 residues involved in the formation and regulation of the glyceraldehyde-3-phosphate dehydrogenase-CP12-phosphoribulokinase complex in *Chlamydomonas reinhardtii*. *Mol. Biosyst.* 8, 2994–3002. doi: 10.1039/c2mb25244a

Baalmann, E., Backhausen, J. E., Kitzmann, C., and Scheibe, R. (1994). Regulation of NADP-dependent glyceraldehyde 3-phosphate dehydrogenase activity in spinach chloroplast. *Botanica Acta* 107, 313–320. doi: 10.1111/j.1438-8677.1994.tb00801.x

Baalmann, E., Backhausen, J. E., Rak, C., Vetter, S., and Scheibe, R. (1995). Reductive modification and nonreductive activation of purified spinach chloroplast NADP-dependent glyceraldehyde-3-phosphate dehydrogenase. *Arch. Biochem. Biophys.* 324, 201–208. doi: 10.1006/abbi.1995.0031

Baalmann, E., Scheibe, R., Cerff, R., and Martin, W. (1996). Functional studies of chloroplast glyceraldehyde-3-phosphate dehydrogenase subunits A and B expressed in *Escherichia coli*: formation of highly active A_4 and B_4 homotetramers and evidence that aggregation of the B_4 complex is mediated by the B subunit carboxy terminus. *Plant Mol. Biol.* 32, 505–513. doi: 10.1007/BF00019102

Babu, M. M., Van Der Lee, R., De Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* 21, 432–440. doi: 10.1016/j.sbi.2011.03.011

Baier, D., and Latzko, E. (1975). Properties and regulation of C-1-fructose-1,6-diphosphatase from spinach chloroplasts. *Biochim. Biophys. Acta* 396, 141–148. doi: 10.1016/0005-2728(75)90197-8

Bolle, C. (2004). The role of GRAS proteins in plant signal transduction and development. *Planta* 218, 683–692. doi: 10.1007/s00425-004-1203-z

Bracken, C., Carr, P. A., Cavanagh, J., and Palmer, A. G. III. (1999). Temperature dependence of intramolecular dynamics of the basic leucine zipper of GCN4: implications for the entropy of association with DNA. *J. Mol. Biol.* 285, 2133–2146. doi: 10.1006/jmbi.1998.2429

Breydo, L., and Uversky, V. N. (2011). Role of metal ions in aggregation of intrinsically disordered proteins in neurodegenerative diseases. *Metallomics* 3, 1163–1180. doi: 10.1039/c1mt00106j

Breydo, L., Wu, J. W., and Uversky, V. N. (2012). Alpha-synuclein misfolding and Parkinson's disease. *Biochim. Biophys. Acta* 1822, 261–285. doi: 10.1016/j.bbadis.2011.10.002

Brinkmann, H., Cerff, R., Salomon, M., and Soll, J. (1989). Cloning and sequence analysis of cDNAs encoding the cytosolic precursors of subunits GapA and GapB of chloroplast glyceraldehyde-3- phosphate dehydrogenase from pea and spinach. *Plant Mol. Biol.* 13, 81–94. doi: 10.1007/BF00027337

Buljan, M., Frankish, A., and Bateman, A. (2010). Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.* 11:R74. doi: 10.1186/gb-2010-11-7-r74

Busza, A., Emery-Le, M., Rosbash, M., and Emery, P. (2004). Roles of the two *Drosophila* CRYPTOCHROME structural domains in circadian photoreception. *Science* 304, 1503–1506. doi: 10.1126/science.1096973

Campbell, K. M., Terrell, A. R., Laybourn, P. J., and Lumb, K. J. (2000). Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. *Biochemistry* 39, 2708–2713. doi: 10.1021/bi9923555

Carmo-Silva, A. E., and Salvucci, M. E. (2013). The regulatory properties of Rubisco activase differ among species and affect photosynthetic induction during light transitions. *Plant Physiol.* 161, 1645–1655. doi: 10.1104/pp.112.213348

Castillo, P., Cetina, A. F., Mendez-Tenorio, A., Espinoza-Fonseca, L. M., and Barron, B. L. (2014). Papillomavirus binding factor (PBF) is an intrinsically disordered protein with potential participation in osteosarcoma genesis, in silico evidence. *Theor. Biol. Med. Model.* 11:51. doi: 10.1186/1742-4682-11-51

Cerff, R. (1979). Quaternary structure of higher plant glyceraldehyde-3-phosphate dehydrogenases. *Eur. J. Biochem.* 94, 243–247. doi: 10.1111/j.1432-1033.1979.tb12891.x

Chakrabortee, S., Tripathi, R., Watson, M., Schierle, G. S., Kurniawan, D. P., Kaminski, C. F., et al. (2012). Intrinsically disordered proteins as molecular shields. *Mol. Biosyst.* 8, 210–219. doi: 10.1039/C1MB05263B

Chaves, I., Pokorny, R., Byrdin, M., Hoang, N., Ritz, T., Brettel, K., et al. (2011). The cryptochromes: blue light photoreceptors in plants and animals. *Annu. Rev. Plant Biol.* 62, 335–364. doi: 10.1146/annurev-arplant-042110-103759

Chaves, I., Yagita, K., Barnhoorn, S., Okamura, H., Van Der Horst, G. T., and Tamanini, F. (2006). Functional evolution of the photolyase/cryptochrome protein family: importance of the C-terminus of mammalian CRY1 for circadian core oscillator performance. *Mol. Cell. Biol.* 26, 1743–1753. doi: 10.1128/MCB.26.5.1743-1753.2006

Chouard, T. (2011). Structural biology: breaking the protein rules. *Nature* 471, 151–153. doi: 10.1038/471151a

Ciceri, P., Gianazza, E., Lazzari, B., Lippoli, G., Genga, A., Hoscheck, G., et al. (1997). Phosphorylation of Opaque2 changes diurnally and impacts its DNA binding activity. *Plant Cell* 9, 97–108. doi: 10.1105/tpc.9.1.97

Cleland, W. W., Andrews, T. J., Gutteridge, S., Hartman, F. C., and Lorimer, G. H. (1998). Mechanism of Rubisco: the carbamate as general base. *Chem. Rev.* 98, 549–562. doi: 10.1021/cr970010r

Cloney, L. P., Bekkaoui, D. R., Wood, M. G., and Hemmingsen, S. M. (1992). Assessment of plant chaperonin-60 gene function in Escherichia coli. *J. Biol. Chem.* 267, 23333–23336.

Crafts-Brandner, S. J., Van De Loo, F. J., and Salvucci, M. E. (1997). The two forms of ribulose-1,5-bisphosphate carboxylase/oxygenase activase differ in sensitivity to elevated temperature. *Plant Physiol.* 114, 439–444.

Cui, H., Levesque, M. P., Vernoux, T., Jung, J. W., Paquette, A. J., Gallagher, K. L., et al. (2007). An evolutionarily conserved mechanism delimiting SHR movement defines a single layer of endodermis in plants. *Science* 316, 421–425. doi: 10.1126/science.1139531

Czarna, A., Berndt, A., Singh, H. R., Grudziecki, A., Ladurner, A. G., Timinszky, G., et al. (2013). Structures of *Drosophila* cryptochrome and mouse cryptochrome1 provide insight into circadian function. *Cell* 153, 1394–1405. doi: 10.1016/j.cell.2013.05.011

Czikkel, B. E., and Maxwell, D. P. (2007). NtGRAS1, a novel stress-induced member of the GRAS family in tobacco, localizes to the nucleus. *J. Plant Physiol.* 164, 1220–1230. doi: 10.1016/j.jplph.2006.07.010

Day, R. B., Tanabe, S., Koshioka, M., Mitsui, T., Itoh, H., Ueguchi-Tanaka, M., et al. (2004). Two rice GRAS family genes responsive to N - acetylchitooligosaccharide elicitor are induced by phytoactive gibberellins: evidence for cross-talk between elicitor and gibberellin signaling in rice cells. *Plant Mol. Biol.* 54, 261–272. doi: 10.1023/B:PLAN.0000028792.72343.ee

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry* 41, 6573–6582. doi: 10.1021/bi012159+

Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* 272, 5129–5148. doi: 10.1111/j.1742-4658.2005.04948.x

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59. doi: 10.1016/S1093-3263(00)00138-8

Dunker, A. K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., et al. (2008a). The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* 9(Suppl. 2):S1. doi: 10.1186/1471-2164-9-S2-S1

Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008b). Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 18, 756–764. doi: 10.1016/j.sbi.2008.10.002

Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell* 71, 1223–1237. doi: 10.1016/S0092-8674(05)80070-4

Ellis, R. J. (1979). The most abundant protein in the world. *Trends Biochem. Sci.* 4, 241–244. doi: 10.1016/0968-0004(79)90212-3

Erales, J., Avilan, L., Lebreton, S., and Gontero, B. (2008). Exploring CP12 binding proteins revealed aldolase as a new partner for the phosphoribulokinase/glyceraldehyde 3-phosphate dehydrogenase/CP12 complex - purification and kinetic characterization of this enzyme from *Chlamydomonas reinhardtii*. *FEBS J.* 275, 1248–1259. doi: 10.1111/j.1742-4658.2008.06284.x

Erales, J., Mekhalfi, M., Woudstra, M., and Gontero, B. (2011). Molecular mechanism of NADPH-glyceraldehyde-3-phosphate dehydrogenase regulation through the C-terminus of CP12 in *Chlamydomonas reinhardtii*. *Biochemistry* 50, 2881–2888. doi: 10.1021/bi1020259

Ernst, H. A., Olsen, A. N., Larsen, S., and Lo Leggio, L. (2004). Structure of the conserved domain of ANAC, a member of the NAC family of transcription factors. *EMBO Rep.* 5, 297–303. doi: 10.1038/sj.embor.7400093

Falini, G., Fermani, S., Ripamonti, A., Sabatino, P., Sparla, F., Pupillo, P., et al. (2003). Dual coenzyme specificity of photosynthetic glyceraldehyde 3-phosphate dehydrogenase interpreted by the crystal structure of A_4 isoform complexed with NAD. *Biochemistry* 42, 4631–4639. doi: 10.1021/bi0272149

Fermani, S., Ripamonti, A., Sabatino, P., Zanotti, G., Scagliarini, S., Sparla, F., et al. (2001). Crystal structure of the non-regulatory A_4 isoform of spinach chloroplast glyceraldehyde-3-phosphate dehydrogenase complexed with NADP. *J. Mol. Biol.* 314, 527–542. doi: 10.1006/jmbi.2001.5172

Fermani, S., Sparla, F., Falini, G., Martelli, P. L., Casadio, R., Pupillo, P., et al. (2007). Molecular mechanism of thioredoxin regulation in photosynthetic A_2B_2-glyceraldehyde-3-phosphate dehydrogenase. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11109–11114. doi: 10.1073/pnas.0611636104

Fermani, S., Trivelli, X., Sparla, F., Thumiger, A., Calvaresi, M., Marri, L., et al. (2012). Conformational selection and folding-upon-binding of the intrinsically disordered protein CP12 regulate photosynthetic enzymes assembly. *J. Biol. Chem.* 287, 21372–21383. doi: 10.1074/jbc.M112.350355

Fersht, A. (1998). *Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding*. New York: W. H. Freeman and Company.

Fode, B., Siemsen, T., Thurow, C., Weigel, R., and Gatz, C. (2008). The *Arabidopsis* GRAS protein SCL14 interacts with class II TGA transcription factors and is essential for the activation of stress-inducible promoters. *Plant Cell* 20, 3122–3135. doi: 10.1105/tpc.108.058974

Fortunato, A. E., Annunziata, R., Jaubert, M., Bouly, J. P., and Falciatore, A. (2015). Dealing with light: the widespread and multitasking cryptochrome/photolyase family in photosynthetic organisms. *J. Plant Physiol.* 172, 42–54. doi: 10.1016/j.jplph.2014.06.011

Fu, X., Richards, D. E., Ait-Ali, T., Hynes, L. W., Ougham, H., Peng, J., et al. (2002). Gibberellin-mediated proteasome-dependent degradation of the barley DELLA protein SLN1 repressor. *Plant Cell* 14, 3191–3200. doi: 10.1105/tpc.006197

Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. (2004). Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* 338, 1015–1026. doi: 10.1016/j.jmb.2004.03.017

Fuxreiter, M., and Tompa, P. (2012). Fuzzy complexes: a more stochastic view of protein function. *Adv. Exp. Med. Biol.* 725, 1–14. doi: 10.1007/978-1-4614-0659-4_1

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23, 950–956. doi: 10.1093/bioinformatics/btm035

Gao, P., Yoo, S. H., Lee, K. J., Rosensweig, C., Takahashi, J. S., Chen, B. P., et al. (2013). Phosphorylation of the cryptochrome 1 C-terminal tail regulates circadian period length. *J. Biol. Chem.* 288, 35277–35286. doi: 10.1074/jbc.M113.509604

Gardebien, F., Thangudu, R. R., Gontero, B., and Offmann, B. (2006). Construction of a 3D model of CP12, a protein linker. *J. Mol. Graph. Model.* 25, 186–195. doi: 10.1016/j.jmgm.2005.12.003

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636. doi: 10.1038/nature04532

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147. doi: 10.1038/415141a

Goloubinoff, P., Christeller, J. T., Gatenby, A. A., and Lorimer, G. H. (1989). Reconstitution of active dimeric ribulose bisphosphate carboxylase from an unfoleded state depends on two chaperonin proteins and Mg-ATP. *Nature* 342, 884–889. doi: 10.1038/342884a0

Gontero, B., and Maberly, S. C. (2012). An intrinsically disordered protein, CP12: jack of all trades and master of the Calvin cycle. *Biochem. Soc. Trans.* 40, 995–999. doi: 10.1042/BST20120097

Gontero, B., and Salvucci, M. E. (2014). Regulation of photosynthetic carbon metabolism in aquatic and terrestrial organisms by Rubisco activase, redox-modulation and CP12. *Aquat. Bot.* 118, 14–23. doi: 10.1016/j.aquabot.2014.05.011

Graciet, E., Gans, P., Wedel, N., Lebreton, S., Camadro, J. M., and Gontero, B. (2003a). The small protein CP12: a protein linker for supramolecular assembly. *Biochemistry* 42, 8163–8170. doi: 10.1021/bi034474x

Graciet, E., Lebreton, S., Camadro, J. M., and Gontero, B. (2003b). Characterization of native and recombinant A_4 glyceraldehyde 3- phosphate dehydrogenase. *Eur. J. Biochem.* 270, 129–136. doi: 10.1046/j.1432-1033.2003.03372.x

Green, C. B. (2004). Cryptochromes: tail-ored for distinct functions. *Curr. Biol.* 14, R847–R849. doi: 10.1016/j.cub.2004.09.040

Groben, R., Kaloudas, D., Raines, C. A., Offmann, B., Maberly, S. C., and Gontero, B. (2010). Comparative sequence analysis of CP12, a small protein involved in the formation of a Calvin cycle complex in photosynthetic organisms. *Photosynth. Res.* 103, 183–194. doi: 10.1007/s11120-010-9542-z

Harada, Y., Sakai, M., Kurabayashi, N., Hirota, T., and Fukada, Y. (2005). Ser-557-phosphorylated mCRY2 is degraded upon synergistic phosphorylation by glycogen synthase kinase-3 beta. *J. Biol. Chem.* 280, 31714–31721. doi: 10.1074/jbc.M506225200

Hardtke, C. S., Gohda, K., Osterlund, M. T., Oyama, T., Okada, K., and Deng, X. W. (2000). HY5 stability and activity in *Arabidopsis* is regulated by phosphorylation in its COP1 binding domain. *EMBO J.* 19, 4997–5006. doi: 10.1093/emboj/19.18.4997

Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., et al. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.* 2:e100. doi: 10.1371/journal.pcbi.0020100

Hazy, E., Bokor, M., Kalmar, L., Gelencser, A., Kamasa, P., Han, K. H., et al. (2011). Distinct hydration properties of wild-type and familial point mutant A53T of alpha-synuclein associated with Parkinson's disease. *Biophys. J.* 101, 2260–2266. doi: 10.1016/j.bpj.2011.08.052

Hazy, E., and Tompa, P. (2009). Limitations of induced folding in molecular recognition by intrinsically disordered proteins. *Chemphyschem* 10, 1415–1419. doi: 10.1002/cphc.200900205

Heery, D. M., Kalkhoven, E., Hoare, S., and Parker, M. G. (1997). A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature* 387, 733–736. doi: 10.1038/42750

Henderson, J. N., Kuriata, A. M., Fromme, R., Salvucci, M. E., and Wachter, R. M. (2011). Atomic resolution x-ray structure of the substrate recognition domain of higher plant ribulose-bisphosphate carboxylase/oxygenase (Rubisco) activase. *J. Biol. Chem.* 286, 35683–35688. doi: 10.1074/jbc.C111.289595

Hildebrandt, T., Knuesting, J., Berndt, C., Morgan, B., and Scheibe, R. (2015). Cytosolic thiol switches regulating basic cellular functions: GAPDH as an information hub? *Biol. Chem.* 396, 523–537. doi: 10.1515/hsz-2014-0295

Hirano, K., Asano, K., Tsuji, H., Kawamura, M., Mori, H., Kitano, H., et al. (2010). Characterization of the molecular mechanism underlying gibberellin perception complex formation in rice. *Plant Cell* 22, 2680–2696. doi: 10.1105/tpc.110.075549

Hirsch, S., Kim, J., Munoz, A., Heckmann, A. B., Downie, J. A., and Oldroyd, G. E. (2009). GRAS proteins form a DNA binding complex to induce gene expression during nodulation signaling in *Medicago truncatula*. *Plant Cell* 21, 545–557. doi: 10.1105/tpc.108.064501

Hirsch, S., and Oldroyd, G. E. (2009). GRAS-domain transcription factors that regulate plant development. *Plant Signal. Behav.* 4, 698–700. doi: 10.4161/psb.4.8.9176

Hollenbeck, J. J., Mcclain, D. L., and Oakley, M. G. (2002). The role of helix stabilizing residues in GCN4 basic region folding and DNA binding. *Protein Sci.* 11, 2740–2747. doi: 10.1110/ps.0211102

Hou, X., Lee, L. Y., Xia, K., Yan, Y., and Yu, H. (2010). DELLAs modulate jasmonate signaling via competitive binding to JAZs. *Dev. Cell* 19, 884–894. doi: 10.1016/j.devcel.2010.10.024

Howard, T. P., Fryer, M. J., Singh, P., Metodiev, M., Lytovchenko, A., Obata, T., et al. (2011a). Antisense suppression of the small chloroplast protein CP12 in tobacco alters carbon partitioning and severely restricts growth. *Plant Physiol.* 157, 620–631. doi: 10.1104/pp.111.183806

Howard, T. P., Lloyd, J. C., and Raines, C. A. (2011b). Inter-species variation in the oligomeric states of the higher plant Calvin cycle enzymes glyceraldehyde-3-phosphate dehydrogenase and phosphoribulokinase. *J. Exp. Bot.* 62, 3799–3805. doi: 10.1093/jxb/err057

Howard, T. P., Upton, G. J., Lloyd, J. C., and Raines, C. A. (2011c). Antisense suppression of the small chloroplast protein CP12 in tobacco: a transcriptional viewpoint. *Plant Signal. Behav.* 6, 2026–2030. doi: 10.4161/psb.6.12.18055

Hussain, A., Cao, D., Cheng, H., Wen, Z., and Peng, J. (2005). Identification of the conserved serine/threonine residues important for gibberellin-sensitivity of *Arabidopsis* RGL2 protein. *Plant J.* 44, 88–99. doi: 10.1111/j.1365-313X.2005.02512.x

Hussain, A., Cao, D., and Peng, J. (2007). Identification of conserved tyrosine residues important for gibberellin sensitivity of *Arabidopsis* RGL2 protein. *Planta* 226, 475–483. doi: 10.1007/s00425-007-0497-z

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'connor, T. R., Sikes, J. G., Obradovic, Z., et al. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32, 1037–1049. doi: 10.1093/nar/gkh253

Itoh, H., Sasaki, A., Ueguchi-Tanaka, M., Ishiyama, K., Kobayashi, M., Hasegawa, Y., et al. (2005). Dissection of the phosphorylation of rice DELLA protein, SLENDER RICE1. *Plant Cell Physiol.* 46, 1392–1399. doi: 10.1093/pcp/pci152

Jeffery, C. J. (1999). Moonlighting proteins. *Trends Biochem. Sci.* 24, 8–11. doi: 10.1016/S0968-0004(98)01335-8

Jensen, M. K., Kjaersgaard, T., Nielsen, M. M., Galberg, P., Petersen, K., O'shea, C., et al. (2010a). The *Arabidopsis thaliana* NAC transcription factor family: structure-function relationships and determinants of ANAC019 stress signalling. *Biochem. J.* 426, 183–196. doi: 10.1042/BJ20091234

Jensen, M. K., Kjaersgaard, T., Petersen, K., and Skriver, K. (2010b). NAC genes: time-specific regulators of hormonal signaling in *Arabidopsis*. *Plant Signal. Behav.* 5, 907–910. doi: 10.4161/psb.5.7.12099

Kaaki, W., Woudstra, M., Gontero, B., and Halgand, F. (2013). Exploration of CP12 conformational changes and of quaternary structural properties using electrospray ionization traveling wave ion mobility mass spectrometry. *Rapid. Commun. Mass Spectrom.* 27, 179–186. doi: 10.1002/rcm.6442

Karagoz, G. E., and Rudiger, S. G. (2015). Hsp90 interaction with clients. *Trends Biochem. Sci.* 40, 117–125. doi: 10.1016/j.tibs.2014.12.002

Kirschner, K., and Bisswanger, H. (1976). Multifunctional proteins. *Annu. Rev. Biochem.* 45, 143–166. doi: 10.1146/annurev.bi.45.070176.001043

Kjaersgaard, T., Jensen, M. K., Christiansen, M. W., Gregersen, P., Kragelund, B. B., and Skriver, K. (2011). Senescence-associated barley NAC (NAM, ATAF1,2, CUC) transcription factor interacts with radical-induced cell death 1 through a disordered regulatory domain. *J. Biol. Chem.* 286, 35418–35429. doi: 10.1074/jbc.M111.247221

Ko, J. H., Yang, S. H., Park, A. H., Lerouxel, O., and Han, K. H. (2007). ANAC012, a member of the plant-specific NAC transcription factor family, negatively regulates xylary fiber development in *Arabidopsis thaliana*. *Plant J.* 50, 1035–1048. doi: 10.1111/j.1365-313X.2007.03109.x

Koshland, D. E. Jr., Nemethy, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 5, 365–385. doi: 10.1021/bi00865a047

Kovacech, B., and Novak, M. (2010). Tau truncation is a productive posttranslational modification of neurofibrillary degeneration in Alzheimer's disease. *Curr. Alzheimer Res.* 7, 708–716. doi: 10.2174/156720510793611556

Kragelund, B. B., Jensen, M. K., and Skriver, K. (2012). Order by disorder in plant signaling. *Trends Plant Sci.* 17, 625–632. doi: 10.1016/j.tplants.2012.06.010

Krisko, A., Smole, Z., Debret, G., Nikolic, N., and Radman, M. (2010). Unstructured hydrophilic sequences in prokaryotic proteomes correlate with dehydration tolerance and host association. *J. Mol. Biol.* 402, 775–782. doi: 10.1016/j.jmb.2010.08.012

Lebreton, S., Gontero, B., Avilan, L., and Ricard, J. (1997). Memory and imprinting effects in multienzyme complexes–II. Kinetics of the bienzyme complex from *Chlamydomonas reinhardtii* and hysteretic activation of chloroplast oxidized phosphoribulokinase. *Eur. J. Biochem.* 246, 85–91. doi: 10.1111/j.1432-1033.1997.t01-2-00085.x

Li, A. D., and Anderson, L. E. (1997). Expression and characterization of pea chloroplastic glyceraldehyde-3-phosphate dehydrogenase composed of only the B-subunit. *Plant Physiol.* 115, 1201–1209. doi: 10.1104/pp.115.3.1201

Li, C., Wang, D., and Portis, A. R. Jr. (2006). Identification of critical arginine residues in the functioning of Rubisco activase. *Arch. Biochem. Biophys.* 450, 176–182. doi: 10.1016/j.abb.2006.04.002

Li, L. A., Zianni, M. R., and Tabita, F. R. (1999). Inactivation of the monocistronic *rca* gene in *Anabaena variabilis* suggests a physiological ribulose bisphosphate carboxylase/oxygenase activase-like function in heterocystous cyanobacteria. *Plant Mol. Biol.* 40, 467–478. doi: 10.1023/A:1006251808625

Libich, D. S., Fawzi, N. L., Ying, J., and Clore, G. M. (2013). Probing the transient dark state of substrate binding to GroEL by relaxation-based solution NMR. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11361–11366. doi: 10.1073/pnas.1305715110

Lieutaud, P., Canard, B., and Longhi, S. (2008). MeDor: a metaserver for predicting protein disorder. *BMC Genomics* 9(Suppl. 2):S25. doi: 10.1186/1471-2164-9-S2-S25

Lim, J., Jung, J. W., Lim, C. E., Lee, M. H., Kim, B. J., Kim, M., et al. (2005). Conservation and diversification of SCARECROW in maize. *Plant Mol. Biol.* 59, 619–630. doi: 10.1007/s11103-005-0578-y

Lin, C., and Shalitin, D. (2003). Cryptochrome structure and signal transduction. *Annu. Rev. Plant Biol.* 54, 469–496. doi: 10.1146/annurev.arplant.54.110901.160901

Liu, B., Zuo, Z., Liu, H., Liu, X., and Lin, C. (2011a). *Arabidopsis* cryptochrome 1 interacts with SPA1 to suppress COP1 activity in response to blue light. *Genes Dev.* 25, 1029–1034. doi: 10.1101/gad.2025011

Liu, H., Liu, B., Zhao, C., Pepper, M., and Lin, C. (2011b). The action mechanisms of plant cryptochromes. *Trends Plant Sci.* 16, 684–691. doi: 10.1016/j.tplants.2011.09.002

Liu, Z., Taub, C. C., and Mcclung, C. R. (1996). Identification of an *Arabidopsis thaliana* ribulose-1,5-bisphosphate carboxylase/oxygenase activase (RCA) minimal promoter regulated by light and the circadian clock. *Plant Physiol.* 112, 43–51. doi: 10.1104/pp.112.1.43

Lopez-Calcagno, P. E., Howard, T. P., and Raines, C. A. (2014). The CP12 protein family: a thioredoxin-mediated metabolic switch? *Front. Plant. Sci.* 5:9. doi: 10.3389/fpls.2014.00009

Lorimer, G. H., Badger, M. R., and Andrews, T. J. (1976). The activation of ribulose-1,5-bisphosphate carboxylase by carbon dioxide and magnesium ions. Equilibria, kinetics, a suggested mechanism, and physiological implications. *Biochemistry* 15, 529–536. doi: 10.1021/bi00648a012

Losh, J. L., Young, J. N., and Morel, F. M. (2013). Rubisco is a small fraction of total protein in marine phytoplankton. *New Phytol.* 198, 52–58. doi: 10.1111/nph.12143

Mächler, F., and Nösberger, J. (1980). Regulation of ribulose bisphosphate carboxylase activity in intact wheat leaves by light, CO_2, and temperature. *J. Exp. Bot.* 31, 1485–1491. doi: 10.1093/jxb/31.6.1485

Marri, L., Sparla, F., Pupillo, P., and Trost, P. (2005). Co-ordinated gene expression of photosynthetic glyceraldehyde-3-phosphate dehydrogenase, phosphoribulokinase, and CP12 in *Arabidopsis thaliana*. *J. Exp. Bot.* 56, 73–80. doi: 10.1093/jxb/eri020

Marri, L., Trost, P., Trivelli, X., Gonnelli, L., Pupillo, P., and Sparla, F. (2008). Spontaneous assembly of photosynthetic supramolecular complexes as mediated by the intrinsically unstructured protein CP12. *J. Biol. Chem.* 283, 1831–1838. doi: 10.1074/jbc.M705650200

Marri, L., Zaffagnini, M., Collin, V., Issakidis-Bourguet, E., Lemaire, S. D., Pupillo, P., et al. (2009). Prompt and easy activation by specific thioredoxins of Calvin cycle enzymes of *Arabidopsis thaliana* associated in the GAPDH/CP12/PRK supramolecular complex. *Mol. Plant* 2, 259–269. doi: 10.1093/mp/ssn061

Marsh, J. A., and Teichmann, S. A. (2010). How do proteins gain new domains? *Genome Biol.* 11, 126. doi: 10.1186/gb-2010-11-7-126

Martino-Catt, S., and Ort, D. R. (1992). Low temperature interrupts circadian regulation of transcriptional activity in chilling-sensitive plants. *Proc. Natl. Acad. Sci. U.S.A.* 89, 3731–3735. doi: 10.1073/pnas.89.9.3731

Matsumura, H., Kai, A., Maeda, T., Tamoi, M., Satoh, A., Tamura, H., et al. (2011). Structure basis for the regulation of glyceraldehyde-3-phosphate dehydrogenase activity via the intrinsically disordered protein CP12. *Structure* 19, 1846–1854. doi: 10.1016/j.str.2011.08.016

Meszaros, B., Tompa, P., Simon, I., and Dosztanyi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.* 372, 549–561. doi: 10.1016/j.jmb.2007.07.004

Mileo, E., Lorenzi, M., Erales, J., Lignon, S., Puppo, C., Le Breton, N., et al. (2013). Dynamics of the intrinsically disordered protein CP12 in its association with GAPDH in the green alga *Chlamydomonas reinhardtii*: a fuzzy complex. *Mol. Biosyst.* 9, 2869–2876. doi: 10.1039/c3mb70190e

Mittag, T., Kay, L. E., and Forman-Kay, J. D. (2010). Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* 23, 105–116. doi: 10.1002/jmr.961

Mizianty, M. J., Uversky, V., and Kurgan, L. (2014). Prediction of intrinsic disorder in proteins using MFDp2. *Methods Mol. Biol.* 1137, 147–162. doi: 10.1007/978-1-4939-0366-5_11

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., et al. (2006). Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* 362, 1043–1059. doi: 10.1016/j.jmb.2006.07.087

Moparthi, S. B., Thieulin-Pardo, G., De Torres, J., Ghenuche, P., Gontero, B., and Wenger, J. (2015). FRET analysis of CP12 structural interplay by GAPDH and PRK. *Biochem. Biophys. Res. Commun.* doi: 10.1016/j.bbrc.2015.01.135

Moparthi, S. B., Thieulin-Pardo, G., Mansuelle, P., Rigneault, H., Gontero, B., and Wenger, J. (2014). Conformational modulation and hydrodynamic radii of CP12 protein and its complexes probed by fluorescence correlation spectroscopy. *FEBS J.* 281, 3206–3217. doi: 10.1111/febs.12854

Moreau, V. H., Da Silva, A. C., Siloto, R. M., Valente, A. P., Leite, A., and Almeida, F. C. (2004). The bZIP region of the plant transcription factor opaque-2 forms stable homodimers in solution and retains its helical structure upon subunit dissociation. *Biochemistry* 43, 4862–4868. doi: 10.1021/bi035905e

Mueller-Cajar, O., Stotz, M., and Bracher, A. (2014). Maintaining photosynthetic CO_2 fixation via protein remodelling: the Rubisco activases. *Photosyn. Res.* 119, 191–201. doi: 10.1007/s11120-013-9819-0

Mueller-Cajar, O., Stotz, M., Wendler, P., Hartl, F. U., Bracher, A., and Hayer-Hartl, M. (2011). Structure and function of the AAA^+ protein CbbX, a red-type Rubisco activase. *Nature* 479, 194–199. doi: 10.1038/nature10568

Munoz-Bertomeu, J., Cascales-Minana, B., Alaiz, M., Segura, J., and Ros, R. (2010). A critical role of plastidial glycolytic glyceraldehyde-3-phosphate dehydrogenase in the control of plant metabolism and development. *Plant Signal. Behav.* 5, 67–69. doi: 10.4161/psb.5.1.10200

Murase, K., Hirano, Y., Sun, T. P., and Hakoshima, T. (2008). Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. *Nature* 456, 459–463. doi: 10.1038/nature07519

Neuwald, A. F., Aravind, L., Spouge, J. L., and Koonin, E. V. (1999). AAA^+: a class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.* 9, 27–43.

Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., and Dunker, A. K. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 44, 12454–12470. doi: 10.1021/bi050736e

Olsen, A. N., Ernst, H. A., Leggio, L. L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. *Trends Plant Sci.* 10, 79–87. doi: 10.1016/j.tplants.2004.12.010

Ooka, H., Satoh, K., Doi, K., Nagata, T., Otomo, Y., Murakami, K., et al. (2003). Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Res.* 10, 239–247. doi: 10.1093/dnares/10.6.239

Ozturk, N., Song, S. H., Ozgur, S., Selby, C. P., Morrison, L., Partch, C., et al. (2007). Structure and function of animal cryptochromes. *Cold Spring Harb. Symp. Quant. Biol.* 72, 119–131. doi: 10.1101/sqb.2007.72.015

Pancsa, R., and Tompa, P. (2012). Structural disorder in eukaryotes. *PLoS ONE* 7:e34687. doi: 10.1371/journal.pone.0034687

Panda, A., and Ghosh, T. C. (2014). Prevalent structural disorder carries signature of prokaryotic adaptation to oxic atmosphere. *Gene* 548, 134–141. doi: 10.1016/j.gene.2014.07.002

Partch, C. L., Clarkson, M. W., Ozgur, S., Lee, A. L., and Sancar, A. (2005). Role of structural plasticity in signal transduction by the cryptochrome blue-light photoreceptor. *Biochemistry* 44, 3795–3805. doi: 10.1021/bi047545g

Patil, A., and Nakamura, H. (2006). Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.* 580, 2041–2045. doi: 10.1016/j.febslet.2006.03.003

Pearce, F. G. (2006). Catalytic by-product formation and ligand binding by ribulose bisphosphate carboxylases from different phylogenies. *Biochem. J.* 399, 525–534. doi: 10.1042/BJ20060430

Perchorowicz, J. T., Raynes, D. A., and Jensen, R. G. (1981). Light limitation of photosynthesis and activation of ribulose bisphosphate carboxylase in wheat seedlings. *Proc. Natl. Acad. Sci. U.S.A.* 78, 2985–2989. doi: 10.1073/pnas.78.5.2985

Peschel, N., Chen, K. F., Szabo, G., and Stanewsky, R. (2009). Light-dependent interactions between the *Drosophila* circadian clock factors cryptochrome, jetlag, and timeless. *Curr. Biol.* 19, 241–247. doi: 10.1016/j.cub.2008.12.042

Petersen, J., Brinkmann, H., and Cerff, R. (2003). Origin, evolution, and metabolic role of a novel glycolytic GAPDH enzyme recruited by land plant plastids. *J. Mol. Evol.* 57, 16–26. doi: 10.1007/s00239-002-2441-y

Petersen, J., Teich, R., Becker, B., Cerff, R., and Brinkmann, H. (2006). The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). *Mol. Biol. Evol.* 23, 1109–1118. doi: 10.1093/molbev/msj123

Pietrosemoli, N., Garcia-Martin, J. A., Solano, R., and Pazos, F. (2013). Genome-wide analysis of protein disorder in *Arabidopsis thaliana*: implications for plant environmental adaptation. *PLoS ONE* 8:e55524. doi: 10.1371/journal.pone.0055524

Podust, L. M., Krezel, A. M., and Kim, Y. (2001). Crystal structure of the CCAAT box/enhancer-binding protein beta activating transcription factor-4 basic leucine zipper heterodimer in the absence of DNA. *J. Biol. Chem.* 276, 505–513. doi: 10.1074/jbc.M005594200

Pohlmeyer, K., Paap, B. K., Soll, J., and Wedel, N. (1996). CP12: a small nuclear-encoded chloroplast protein provides novel insights into higher-plant GAPDH evolution. *Plant Mol. Biol.* 32, 969–978. doi: 10.1007/BF00020493

Portis, A. R., Jr. (2003). Rubisco activase - Rubisco's catalytic chaperone. *Photosyn. Res.* 75, 11–27. doi: 10.1023/A:1022458108678

Portis, A. R. Jr., Li, C., Wang, D., and Salvucci, M. E. (2008). Regulation of Rubisco activase and its interaction with Rubisco. *J. Exp. Bot.* 59, 1597–1604. doi: 10.1093/jxb/erm240

Portis, A. R., Salvucci, M. E., and Ogren, W. L. (1986). Activation of ribulosebisphosphate carboxylase/oxygenase at physiological CO_2 and ribulosebisphosphate concentrations by Rubisco activase. *Plant Physiol.* 82, 967–971. doi: 10.1104/pp.82.4.967

Pupillo, P., and Giuliani Piccari, G. (1975). The reversible depolymerization of spinach chloroplast glyceraldehyde-phosphate dehydrogenase. Interaction with nucleotides and dithiothreitol. *Eur. J. Biochem.* 51, 475–482. doi: 10.1111/j.1432-1033.1975.tb03947.x

Pupillo, P., and Piccari, G. G. (1973). The effect of NADP on the subunit structure and activity of spinach chloroplast glyceraldehyde-3-phosphate dehydrogenase. *Arch. Biochem. Biophys.* 154, 324–331. doi: 10.1016/0003-9861(73)90064-7

Qian, J., and Rodermel, S. R. (1993). Ribulose-1,5-bisphosphate carboxylase/oxygenase activase cDNAs from *Nicotiana tabacum*. *Plant Physiol.* 102, 683–684. doi: 10.1104/pp.102.2.683

Raven, J. A. (2013). Rubisco: still the most abundant protein of Earth? *New Phytol.* 198, 1–3. doi: 10.1111/nph.12197

Robbens, S., Petersen, J., Brinkmann, H., Rouze, P., and Van De Peer, Y. (2007). Unique regulation of the Calvin cycle in the ultrasmall green alga *Ostreococcus*. *J. Mol. Evol.* 64, 601–604. doi: 10.1007/s00239-006-0159-y

Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S., et al. (1998). Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* 3, 437–448. Available online at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.4594

Rundle, S. J., and Zielinski, R. E. (1991). Organization and expression of two tandemly oriented genes encoding ribulosebisphosphate carboxylase/oxygenase activase in barley. *J. Biol. Chem.* 266, 4677–4685.

Rushton, P. J., Bokowiec, M. T., Han, S., Zhang, H., Brannock, J. F., Chen, X., et al. (2008). Tobacco transcription factors: novel insights into transcriptional regulation in the *Solanaceae*. *Plant Physiol.* 147, 280–295. doi: 10.1104/pp.107.114041

Saha, T., Kar, R. K., and Sa, G. (2014). Structural and sequential context of p53: a review of experimental and theoretical evidence. *Prog. Biophys. Mol. Biol.* 117, 250–263. doi: 10.1016/j.pbiomolbio.2014.12.002

Salminen, A., Ojala, J., Kaarniranta, K., Hiltunen, M., and Soininen, H. (2011). Hsp90 regulates tau pathology through co-chaperone complexes in Alzheimer's disease. *Prog. Neurobiol.* 93, 99–110. doi: 10.1016/j.pneurobio.2010.10.006

Salvucci, M. E. (2004). Potential for interactions between the carboxy- and amino-termini of Rubisco activase subunits. *FEBS Lett.* 560, 205–209. doi: 10.1016/S0014-5793(04)00111-5

Salvucci, M. E., Rajagopalan, K., Sievert, G., Haley, B. E., and Watt, D. S. (1993). Photoaffinity labeling of ribulose-1,5-bisphosphate carboxylase/oxygenase activase with ATP gamma-benzophenone. Identification of the ATP gamma-phosphate binding domain. *J. Biol. Chem.* 268, 14239–14244.

Salvucci, M. E., Van De Loo, F. J., and Stecher, D. (2003). Two isoforms of Rubisco activase in cotton, the products of separate genes not alternative splicing. *Planta* 216, 736–744. doi: 10.1007/s00425-002-0923-1

Sancar, A. (2004). Photolyase and cryptochrome blue-light photoreceptors. *Adv. Protein Chem.* 69, 73–100. doi: 10.1016/S0065-3233(04)69003-6

Sanchez, C., Vielba, J. M., Ferro, E., Covelo, G., Sole, A., Abarca, D., et al. (2007). Two SCARECROW-LIKE genes are induced in response to exogenous auxin in rooting-competent cuttings of distantly related forest species. *Tree Physiol.* 27, 1459–1470. doi: 10.1093/treephys/27.10.1459

Scagliarini, S., Trost, P., and Pupillo, P. (1998). The non-regulatory isoform of NADP(H)-glyceraldehyde 3-phosphate dehydrogenase from spinach chloroplasts. *J. Exp. Bot.* 49, 1307–1315. doi: 10.1093/jxb/49.325.1307

Scheibe, R., Wedel, N., Vetter, S., Emmerlich, V., and Sauermann, S. M. (2002). Co-existence of two regulatory NADP-glyceraldehyde 3-P dehydrogenase complexes in higher plant chloroplasts. *Eur. J. Biochem.* 269, 5617–5624. doi: 10.1046/j.1432-1033.2002.03269.x

Sedzik, J., and Kirschner, D. A. (1992). Is myelin basic protein crystallizable? *Neurochem. Res.* 17, 157–166. doi: 10.1007/BF00966794

Seo, P. J., Kim, M. J., Song, J. S., Kim, Y. S., Kim, H. J., and Park, C. M. (2010). Proteolytic processing of an *Arabidopsis* membrane-bound NAC transcription factor is triggered by cold-induced changes in membrane fluidity. *Biochem. J.* 427, 359–367. doi: 10.1042/BJ20091762

Seo, P. J., and Park, C. M. (2010). A membrane-bound NAC transcription factor as an integrator of biotic and abiotic stress signals. *Plant Signal. Behav.* 5, 481–483. doi: 10.4161/psb.11083

Shen, J. B., and Ogren, W. L. (1992). Alteration of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase activase activities by site-directed mutagenesis. *Plant Physiol.* 99, 1201–1207. doi: 10.1104/pp.99.3.1201

Singh, P., Kaloudas, D., and Raines, C. A. (2008). Expression analysis of the *Arabidopsis* CP12 gene family suggests novel roles for these proteins in roots and floral tissues. *J. Exp. Bot.* 59, 3975–3985. doi: 10.1093/jxb/ern236

Sirover, M. A. (1999). New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase. *Biochim. Biophys. Acta* 1432, 159–184. doi: 10.1016/S0167-4838(99)00119-3

Sirover, M. A. (2011). On the functional diversity of glyceraldehyde-3-phosphate dehydrogenase: biochemical mechanisms and regulatory control. *Biochim. Biophys. Acta* 1810, 741–751. doi: 10.1016/j.bbagen.2011.05.010

Sparla, F., Fermani, S., Falini, G., Zaffagnini, M., Ripamonti, A., Sabatino, P., et al. (2004). Coenzyme site-directed mutants of photosynthetic A_4-GAPDH show selectively reduced NADPH-dependent catalysis, similar to regulatory A_2B_2-GAPDH inhibited by oxidized thioredoxin. *J. Mol. Biol.* 340, 1025–1037. doi: 10.1016/j.jmb.2004.06.005

Sparla, F., Pupillo, P., and Trost, P. (2002). The C-terminal extension of glyceraldehyde-3-phosphate dehydrogenase subunit B acts as an autoinhibitory domain regulated by thioredoxins and nicotinamide adenine dinucleotide. *J. Biol. Chem.* 277, 44946–44952. doi: 10.1074/jbc.M206873200

Sparla, F., Zaffagnini, M., Wedel, N., Scheibe, R., Pupillo, P., and Trost, P. (2005). Regulation of photosynthetic GAPDH dissected by mutants. *Plant Physiol.* 138, 2210–2219. doi: 10.1104/pp.105.062117

Srivastava, R., Deng, Y., and Howell, S. H. (2014). Stress sensing in plants by an ER stress sensor/transducer, bZIP28. *Front. Plant Sci.* 5:59. doi: 10.3389/fpls.2014.00059

Srivastava, R., Deng, Y., Shah, S., Rao, A. G., and Howell, S. H. (2013). BINDING PROTEIN is a master regulator of the endoplasmic reticulum stress sensor/transducer bZIP28 in Arabidopsis. *Plant Cell* 25, 1416–1429. doi: 10.1105/tpc.113.110684

Stanley, D. N., Raines, C. A., and Kerfeld, C. A. (2013). Comparative analysis of 126 cyanobacterial genomes reveals evidence of functional diversity among homologs of the redox-regulated CP12 protein. *Plant Physiol.* 161, 824–835. doi: 10.1104/pp.112.210542

Steiner, E., Efroni, I., Gopalraj, M., Saathoff, K., Tseng, T. S., Kieffer, M., et al. (2012). The *Arabidopsis* O-linked N-acetylglucosamine transferase SPINDLY interacts with class I TCPs to facilitate cytokinin responses in leaves and flowers. *Plant Cell* 24, 96–108. doi: 10.1105/tpc.111.093518

Stotz, M., Mueller-Cajar, O., Ciniawsky, S., Wendler, P., Hartl, F. U., Bracher, A., et al. (2011). Structure of green-type Rubisco activase from tobacco. *Nat. Struct. Mol. Biol.* 18, 1366–1370. doi: 10.1038/nsmb.2171

Sun, X., Jones, W. T., Harvey, D., Edwards, P. J., Pascal, S. M., Kirk, C., et al. (2010). N-terminal domains of DELLA proteins are intrinsically unstructured in the absence of interaction with GID1/gibberellic acid receptors. *J. Biol. Chem.* 285, 11557–11571. doi: 10.1074/jbc.M109.027011

Sun, X., Jones, W. T., and Rikkerink, E. H. (2012). GRAS proteins: the versatile roles of intrinsically disordered proteins in plant signalling. *Biochem. J.* 442, 1–12. doi: 10.1042/BJ20111766

Sun, X., Rikkerink, E. H., Jones, W. T., and Uversky, V. N. (2013). Multifarious roles of intrinsic disorder in proteins illustrate its broad impact on plant biology. *Plant Cell* 25, 38–55. doi: 10.1105/tpc.112.106062

Sun, X., Xue, B., Jones, W. T., Rikkerink, E., Dunker, A. K., and Uversky, V. N. (2011). A functionally required unfoldome from the plant kingdom: intrinsically disordered N-terminal domains of GRAS proteins are involved in molecular recognition during plant development. *Plant Mol. Biol.* 77, 205–223. doi: 10.1007/s11103-011-9803-z

Taoka, K., Yanagimoto, Y., Daimon, Y., Hibara, K., Aida, M., and Tasaka, M. (2004). The NAC domain mediates functional specificity of CUP-SHAPED COTYLEDON proteins. *Plant J.* 40, 462–473. doi: 10.1111/j.1365-313X.2004.02238.x

Tian, C., Wan, P., Sun, S., Li, J., and Chen, M. (2004). Genome-wide analysis of the GRAS gene family in rice and *Arabidopsis*. *Plant Mol. Biol.* 54, 519–532. doi: 10.1023/B:PLAN.0000038256.89809.57

To, K. Y., Suen, D. F., and Chen, S. C. (1999). Molecular characterization of ribulose-1,5-bisphosphate carboxylase/oxygenase activase in rice leaves. *Planta* 209, 66–76. doi: 10.1007/s004250050607

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* 27, 527–533. doi: 10.1016/S0968-0004(02)02169-2

Tompa, P. (2009). Structural disorder in amyloid fibrils: its implication in dynamic interactions of proteins. *FEBS J.* 276, 5406–5415. doi: 10.1111/j.1742-4658.2009.07250.x

Tompa, P., and Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 33, 2–8. doi: 10.1016/j.tibs.2007.10.003

Triezenberg, S. J. (1995). Structure and function of transcriptional activation domains. *Curr. Opin. Genet. Dev.* 5, 190–196. doi: 10.1016/0959-437X(95)80007-7

Trost, P., Fermani, S., Marri, L., Zaffagnini, M., Falini, G., Scagliarini, S., et al. (2006). Thioredoxin-dependent regulation of photosynthetic glyceraldehyde-3-phosphate dehydrogenase: autonomous vs. CP12-dependent mechanisms. *Photosyn. Res.* 89, 1–13. doi: 10.1007/s11120-006-9099-z

Trost, P., and Pupillo, P. (1993). Inhibition of spinach D-glyceraldehyde 3-phosphate: NADP+ oxidoreductase (nonphosphorylating) by adenylate compounds: the effect of dead-end inhibitors on a steady state random reaction mechanism. *Arch. Biochem. Biophys.* 306, 76–82. doi: 10.1006/abbi.1993.1483

Uversky, V. N. (2002). What does it mean to be natively unfolded? *Eur. J. Biochem.* 269, 2–12. doi: 10.1046/j.0014-2956.2001.02649.x

Uversky, V. N. (2009). Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front. Biosci.* 14, 5188–5238. doi: 10.2741/3594

Uversky, V. N. (2010). The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J. Biomed. Biotechnol.* 2010:568068. doi: 10.1155/2010/568068

Uversky, V. N. (2011). Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. *Chem. Soc. Rev.* 40, 1623–1634. doi: 10.1039/C0CS00057D

Uversky, V. N. (2013). The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. *FEBS Lett.* 587, 1891–1901. doi: 10.1016/j.febslet.2013.04.042

Uversky, V. N., and Dunker, A. K. (2010). Understanding protein non-folding. *Biochim. Biophys. Acta* 1804, 1231–1264. doi: 10.1016/j.bbapap.2010.01.017

Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41, 415–427. doi: 10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N., et al. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* 6, 2351–2366. doi: 10.1021/pr0701411

Valsecchi, I., Guittard-Crilat, E., Maldiney, R., Habricot, Y., Lignon, S., Lebrun, R., et al. (2013). The intrinsically disordered C-terminal region of *Arabidopsis thaliana* TCP8 transcription factor acts both as a transactivation and self-assembly domain. *Mol. Biosyst.* 9, 2282–2295. doi: 10.1039/c3mb70128j

Van De Loo, F. J., and Salvucci, M. E. (1996). Activation of ribulose-1,5-biphosphate carboxylase/oxygenase (Rubisco) involves Rubisco activase Trp16. *Biochemistry* 35, 8143–8148. doi: 10.1021/bi9604901

Vinson, C. R., Hai, T., and Boyd, S. M. (1993). Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design. *Genes Dev.* 7, 1047–1058. doi: 10.1101/gad.7.6.1047

Viola, I. L., Reinheimer, R., Ripoll, R., Manassero, N. G., and Gonzalez, D. H. (2012). Determinants of the DNA binding specificity of class I and class II TCP transcription factors. *J. Biol. Chem.* 287, 347–356. doi: 10.1074/jbc.M111.256271

Viola, I. L., Uberti Manassero, N. G., Ripoll, R., and Gonzalez, D. H. (2011). The *Arabidopsis* class I TCP transcription factor AtTCP11 is a developmental regulator with distinct DNA-binding properties due to the presence of a threonine residue at position 15 of the TCP domain. *Biochem. J.* 435, 143–155. doi: 10.1042/BJ20101019

Wang, D., and Portis, A. R. Jr. (2006). Increased sensitivity of oxidized large isoform of ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) activase to ADP inhibition is due to an interaction between its carboxyl extension and nucleotide-binding pocket. *J. Biol. Chem.* 281, 25241–25249. doi: 10.1074/jbc.M604756200

Wang, H., Ma, L. G., Li, J. M., Zhao, H. Y., and Deng, X. W. (2001). Direct interaction of *Arabidopsis* cryptochromes with COP1 in light control development. *Science* 294, 154–158. doi: 10.1126/science.1063630

Watillon, B., Kettmann, R., Boxus, P., and Burny, A. (1993). Developmental and circadian pattern of rubisco activase mRNA accumulation in apple plants. *Plant Mol. Biol.* 23, 501–509. doi: 10.1007/BF00019298

Werneke, J. M., Chatfield, J. M., and Ogren, W. L. (1988). Catalysis of ribulosebisphosphate carboxylase/oxygenase activation by the product of a Rubisco activase cDNA clone expressed in *Escherichia coli*. *Plant Physiol.* 87, 917–920. doi: 10.1104/pp.87.4.917

Werneke, J. M., Chatfield, J. M., and Ogren, W. L. (1989). Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and *Arabidopsis*. *Plant Cell* 1, 815–825. doi: 10.1105/tpc.1.8.815

Wolosiuk, R. A., and Buchanan, B. B. (1976). Studies on the regulation of chloroplast NADP-linked glyceraldehyde-3-phosphate dehydrogenase. *J. Biol. Chem.* 251, 6456–6461.

Wolosiuk, R. A., and Buchanan, B. B. (1978). Regulation of chloroplast phosphoribulokinase by the ferredoxin/thioredoxin system. *Arch. Biochem. Biophys.* 189, 97–101. doi: 10.1016/0003-9861(78)90119-4

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

Xue, B., Ganti, K., Rabionet, A., Banks, L., and Uversky, V. N. (2014). Disordered interactome of human papillomavirus. *Curr. Pharm. Des.* 20, 1274–1292. doi: 10.2174/13816128113199990072

Yang, H. Q., Tang, R. H., and Cashmore, A. R. (2001). The signaling mechanism of *Arabidopsis* CRY1 involves direct interaction with COP1. *Plant Cell* 13, 2573–2587. doi: 10.1105/tpc.13.12.2573

Yang, H. Q., Wu, Y. J., Tang, R. H., Liu, D., Liu, Y., and Cashmore, A. R. (2000). The C-termini of *Arabidopsis* cryptochromes mediate a constitutive light response. *Cell* 103, 815–827. doi: 10.1016/S0092-8674(00)00184-7

Yin, Z., Meng, F., Song, H., Wang, X., Xu, X., and Yu, D. (2010). Expression quantitative trait loci analysis of two genes encoding rubisco activase in soybean. *Plant Physiol.* 152, 1625–1637. doi: 10.1104/pp.109.148312

Yoon, M. K., Shin, J., Choi, G., and Choi, B. S. (2006). Intrinsically unstructured N-terminal domain of bZIP transcription factor HY5. *Proteins* 65, 856–866. doi: 10.1002/prot.21089

Yruela, I., and Contreras-Moreira, B. (2012). Protein disorder in plants: a view from the chloroplast. *BMC Plant Biol.* 12:165. doi: 10.1186/1471-2229-12-165

Yruela, I., and Contreras-Moreira, B. (2013). Genetic recombination is associated with intrinsic disorder in plant proteomes. *BMC Genomics* 14:772. doi: 10.1186/1471-2164-14-772

Yu, X., Liu, H., Klejnot, J., and Lin, C. (2010). The cryptochrome blue light receptors. *Arabidopsis Book* 8:e0135. doi: 10.1199/tab.0135

Yu, X., Shalitin, D., Liu, X., Maymon, M., Klejnot, J., Yang, H., et al. (2007). Derepression of the NC80 motif is critical for the photoactivation of *Arabidopsis* CRY2. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7289–7294. doi: 10.1073/pnas.0701912104

Zarzycki, J., Axen, S. D., Kinney, J. N., and Kerfeld, C. A. (2013). Cyanobacterial-based approaches to improving photosynthesis in plants. *J. Exp. Bot.* 64, 787–798. doi: 10.1093/jxb/ers294

Zhang, N., Kallis, R. P., Ewy, R. G., and Portis, A. R. Jr. (2002). Light modulation of Rubisco in *Arabidopsis* requires a capacity for redox regulation of the larger Rubisco activase isoform. *Proc. Natl. Acad. Sci. U.S.A.* 99, 3330–3334. doi: 10.1073/pnas.042529999

Zhang, N., and Portis, A. R. Jr. (1999). Mechanism of light regulation of Rubisco: a specific role for the larger Rubisco activase isoform involving reductive activation by thioredoxin-*f*. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9438–9443. doi: 10.1073/pnas.96.16.9438

Zhang, N., Schürmann, P., and Portis, A. Jr. (2001). Characterization of the regulatory function of the 46-kDa isoform of Rubisco activase from *Arabidopsis*. *Photosyn. Res.* 68, 29–37. doi: 10.1023/A:1011845506196

Zoltowski, B. D., Vaidya, A. T., Top, D., Widom, J., Young, M. W., and Crane, B. R. (2011). Structure of full-length *Drosophila* cryptochrome. *Nature* 480, 396–399. doi: 10.1038/nature10618

Zuo, Z., Liu, H., Liu, B., Liu, X., and Lin, C. (2011). Blue light-dependent interaction of CRY2 with SPA1 regulates COP1 activity and floral initiation in *Arabidopsis*. *Curr. Biol.* 21, 841–847. doi: 10.1016/j.cub.2011.03.048

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.

PERSPECTIVE
published: 09 July 2015
doi: 10.3389/fmolb.2015.00040

Globular and disordered—the non-identical twins in protein-protein interactions

Kaare Teilum *, Johan G. Olsen and Birthe B. Kragelund* *

Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Copenhagen, Denmark

In biology proteins from different structural classes interact across and within classes in ways that are optimized to achieve balanced functional outputs. The interactions between intrinsically disordered proteins (IDPs) and other proteins rely on changes in flexibility and this is seen as a strong determinant for their function. This has fostered the notion that IDP's bind with low affinity but high specificity. Here we have analyzed available detailed thermodynamic data for protein-protein interactions to put to the test if the thermodynamic profiles of IDP interactions differ from those of other protein-protein interactions. We find that ordered proteins and the disordered ones act as non-identical twins operating by similar principles but where the disordered proteins complexes are on average less stable by 2.5 kcal mol^{-1}.

Keywords: ITC, IDP, intrinsically disordered, entropy, enthalpy, stability

Edited by:
Kris Pauwels,
Vrije Universiteit Brussel, Belgium

Reviewed by:
Alfonso De Simone,
Imperial College London, UK
Daniel E. Otzen,
Aarhus University, Denmark

***Correspondence:**
Kaare Teilum and Birthe B. Kragelund,
Department of Biology, Ole Maaløes
Vej 5, 2200 Copenhagen, Denmark
kaare.teilum@bio.ku.dk;
bbk@bio.ku.dk

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

Received: *30 April 2015*
Accepted: *29 June 2015*
Published: *09 July 2015*

Citation:
Teilum K, Olsen JG and Kragelund BB (2015) Globular and disordered—the non-identical twins in protein-protein interactions. Front. Mol. Biosci. 2:40. doi: 10.3389/fmolb.2015.00040

Introduction

Proteins function though the action and communication with other molecules and the intricate interplay among residues within every binding site results in diagnostic thermodynamic profiles implicit to the particular molecular pair. In protein-protein interaction the majority of the binding energy comes from a few critical hot-spot interactions (Clackson and Wells, 1995), but the binding energy also depends on other factors such as interface size, residue composition, flexibility of the interacting partners as well as on environmental cues. The discovery of a large fraction of the proteome being intrinsically disordered (ID) means that a substantial fraction of protein-protein interactions involves proteins or parts of proteins, which do not adopt a well-defined three-dimensional structure in the unbound state. These proteins, or regions in proteins, originate from the class of *intrinsically disordered proteins* (IDPs) (Dunker et al., 2000; Tompa, 2002; Nilsson et al., 2011). They are central to a plethora of key biological processes, are multi-specific and possess a versatile interaction potential placing many of them centrally in cellular hubs (Han et al., 2004). The prevailing notion is that IDPs are able to bind with high specificity, but low affinity, although recent kinetic studies suggest that this concept may not be straightforward (Dogan et al., 2014; Iesmantavicius et al., 2014; Krieger et al., 2014). IDPs contain very few hydrophobic residues (Dunker et al., 2001), which suggests that their interaction energies may be comparatively low, substantiated by the entropy loss of ordered complex formation from a disordered peptide chain. Specificity, on the other hand, arises when the polypeptide chain adopts the correct conformation in which the distribution of side chains match electrostatic and hydrogen bonding donors and acceptors as well as hydrophobic patches on the target. This paradigm of lower affinity of IDPs compared to globular proteins has been suggested but never challenged by a large-scale thermodynamic assessment, which is the aim of the present paper.

Results and Discussion

Based on previous collections of data (Stites, 1997; Huang and Liu, 2013) and including several additional data from the literature found by searching PubMed for "ITC protein-protein interactions," "ITC intrinsically disordered protein," "thermodynamics protein-protein interactions," and "thermodynamics intrinsically disordered protein," we have compiled thermodynamic parameters from close to 200 different protein-protein interaction studies (Supplementary Table 1). The data were standardized to 298 K assuming that $\Delta C_p = 0$, as ΔC_p has only been estimated for very few of the complexes. We have estimated that the error introduced in ΔG^0 is less than 0.2 kcal mol^{-1} in the most extreme cases where the data were measured at 281 K. For most cases where there is less than 5 K difference the error is less than 0.05 kcal mol^{-1}. We subsequently compared and correlated the parameters for interactions that involve only globular proteins (91 complexes), to the parameters for interactions, where one partner is an IDP (106 complexes). To avoid over-representing a single protein-protein complex we exclusively compared wild-type proteins so that protein specific irregularities will be averaged out. In the cases where a structure of the complex has been determined, we have calculated the interaction surface area using PISA (Krissinel and Henrick, 2007) (Supplementary Table 1), and determined the amino acid composition of the interface using NCONT from the CCP4i suite (Winn et al., 2011). The amino acids were divided into four classes for analysis (FWY, CILMV, AGPST, and DEHKNQR) based on the BLOSSUM50 substitution matrix as defined by Weathers et al. (2004). The interfaces of the all ordered (ORD-ORD) complexes and the ordered-IDP (ORD-IDP) complexes were then compared in this context (**Figure 1**).

Complexes from the two groups were almost equally represented (ORD-ORD: 52 structures average interface size of 886 ± 46 $Å^2$; ORD-IDP: 41 structures, average interface size of 905 ± 80 $Å^2$) in the protein data bank (Berman et al., 2000). The sizes of the binding interface areas in the two groups of proteins were not significantly different (*t*-test, $P > 0.05$), (Supplementary Table 1). This is perhaps not unexpected, although one might have anticipated the IDP-complexes to have—on average—smaller interfaces, as many of their interactions are mediated by small linear motifs (SLiMs) (Dinkel et al., 2014), and short molecular recognition motifs (MoRFs) (Mohan et al., 2006). These motifs are typically peptide regions that fold into regular secondary structure on binding. Thus, one conclusion is that in the globular complexes analyzed here, there are equally many small interfaces, matching those of SLiMs and MoRFs of IDPs.

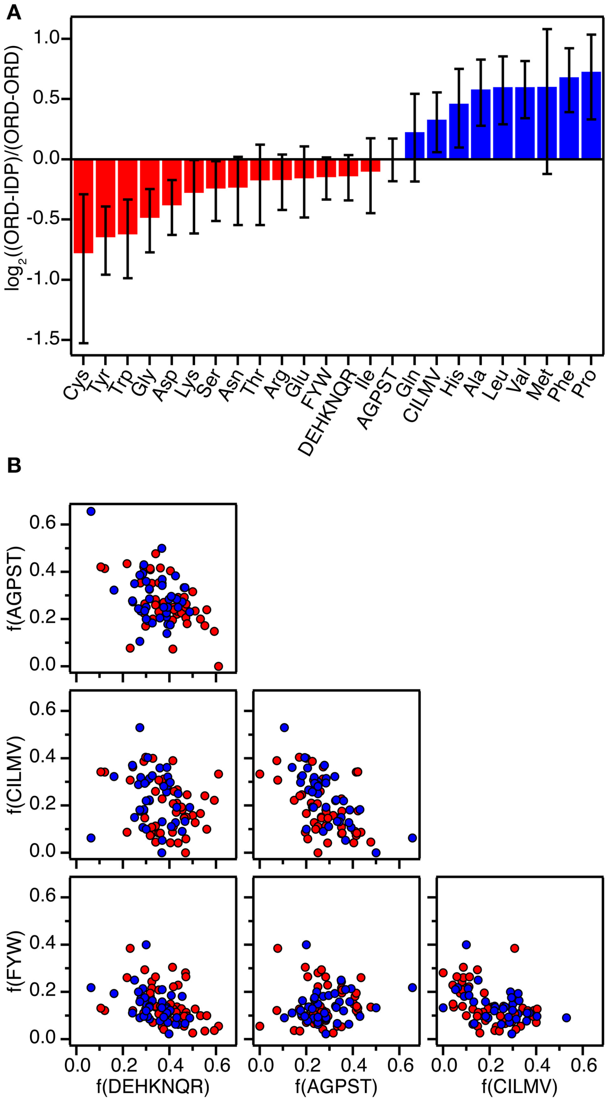


FIGURE 1 | Amino acid composition of protein-protein interfaces extracted from 87 high-resolution structures of protein-protein complexes. (A) Fractional overrepresentation of each amino acid residue type and of the four amino acid residue classes (FWY, CILMV, AGPST, and DEHKNQR) in ORD-IDP complexes relative to ORD-ORD complexes. $\log_2$ of the ratios are plotted with positive values indicating overrepresentation in ORD-IDP complexes. **(B)** Correlation plots of the fractions of the four amino acid residue classes (FWY, CILMV, AGPST, and DEHKNQR) in protein-protein interfaces. Each point represents a protein-protein complex and is colored either red (ORD-ORD) or blue (ORD-IDP).

The second result of the structural analysis is that the intermolecular interactions, as reflected in the distribution of the four groups of amino acids, is the same (**Figure 1**). This observation is perhaps more surprising since the amino acid composition of IDPs is very distinct and different from that of globular proteins (Weathers et al., 2004; Uversky et al., 2005; Han et al., 2006; Hansen et al., 2006) with the low content of hydrophobic residues as the underlying reason for IDPs not forming globular structures. However, the amino acid composition on the surface of globular proteins seems to resemble that of IDPs more than the overall composition (Fukuchi and Nishikawa, 2001; Tompa, 2002; Levy, 2010). Moreover, it differs significantly from the composition of interfaces in obligate oligomers that are typically much more hydrophobic (Janin et al., 2008). In a previous study the residue composition of extended binding surfaces of IDPs bound to an ordered partner was investigated (Wong et al., 2013). Compared to interfaces between two ordered proteins, the IDPs in complex with an ordered partner had in that work an overrepresentation of hydrophobic residues as leucine and isoleucine in the core of the interface, and the ordered binding partner had an increased number of charged residues. Thus, this apparent counter balance is in full accordance with the overall sum of the interface we report here. A decomposition of the distribution into individual residues within the current set supports previous findings, although the effect is small (the largest difference is for Cys which is 41% less abundant in the ORD-IDP complexes) (**Figure 1A**). Therefore, if specificity is embedded in interactions between charged and polar side-chains in the interface (Eaton et al., 1995; Wong et al., 2013), we find no indication to suggest that the IDPs bind to globular proteins with higher specificity than globular proteins do.

Recall the basic thermodynamic relation, $\Delta G^0 = \Delta H^0 - T\Delta S^0$ in which the entropy-enthalpy compensation infers that ΔH^0 and $T\Delta S^0$ are highly correlated (Brady and Sharp, 1997; Williams et al., 2004; Teilum et al., 2009). Thus, ΔG^0 for the complexes in the selected sets covers a narrow range from $-19.8\,\text{kcal mol}^{-1}$ to $-4.2\,\text{kcal mol}^{-1}$ (corresponding to K_d from 3 fM to 830 μM) compared to ΔH^0 and $T\Delta S^0$ that are found in the ranges from -66.7 to $19.9\,\text{kcal mol}^{-1}$ and from -56.1 to $28.5\,\text{kcal mol}^{-1}$, respectively. The analysis of the thermodynamic parameters shows that the enthalpy (ΔH°) and the entropy (ΔS°) for binding are *not* significantly different between the two groups of proteins (t-test, $P > 0.1$). However, the average entropic contribution (−TΔS°) to the binding free energy for interactions between two ordered proteins is $2.5 \pm 1.6\,\text{kcal mol}^{-1}$ smaller (more stabilizing) than for interactions between an ordered and a disordered protein. Within both groups there is a linear correlation between TΔS° and ΔH° (ORD-ORD: slope $= 1.09 \pm 0.03$, $r = 0.97$; ORD-IDP: slope $= 1.06 \pm 0.02$, $r = 0.98$), which demonstrates a similar entropy-enthalpy compensation (**Figure 2A**). Thus, the same underlying thermodynamic principles are true for both groups.

In contrast to ΔS° and ΔH°, there is a significant difference in ΔG° between the groups (t-test, $P < 0.0001$). For the ORD-ORD complexes $<\Delta G^\circ> = -11.1 \pm 0.4\,\text{kcal mol}^{-1}$, and for the ORD-IDP complexes $<\Delta G^\circ> = -8.5 \pm 0.2\,\text{kcal mol}^{-1}$. The difference in $<\Delta G^\circ>$ is $2.5 \pm 0.4\,\text{kcal mol}^{-1}$, which is primarily accounted for by the difference in TΔS° (*vide supra*). This number is close to the $2.6\,\text{kcal mol}^{-1}$ recently published from a much smaller dataset based on mutation studies (Huang and Liu, 2013). Note that the distribution of ΔG° among the complexes for which a structure is available is similar to the distribution in the full dataset, and that this is true for the difference in <ΔG°> too. As we see no differences in the sizes of the binding interfaces or the amount of hydrophobic residues in the interfaces, and since the disordered proteins in the ORD-IDP complexes rarely form extended hydrophobic cores in their folded conformations, the hydrophobic surface area buried in ligand binding process must be similar in the two classes of protein complexes. Consequently, the difference in TΔS° is unlikely to arise from significant differences in the desolvation entropy contribution. This conclusion is in contrast to a computational study of complexes involving extended IDPs, which were selected based on a radius-of-gyration criterion of the three-dimensional structure of the complex (Wong et al., 2013). However, in that work the energetic terms were not decomposed into enthalpic and entropic contributions. Nevertheless, the experimental data for the large group of complexes that we have compiled suggest to us that the less favorable entropic contribution for the ORD-IDP complexes primarily originates from loss in conformational entropy. Indeed, it agrees with the mechanistic difference between binding an ordered and a disordered ligand. The disordered polypeptide has to fold to form

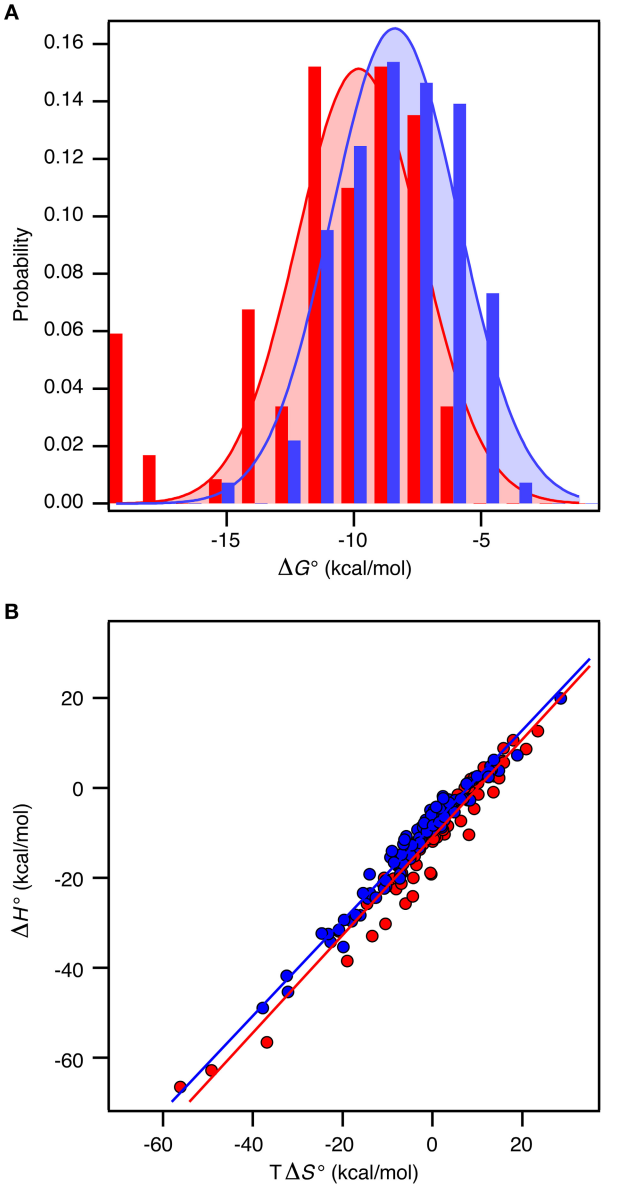


FIGURE 2 | Thermodynamics of 196 protein-protein complexes. (A) Histogram of the binding free energy, ΔG°, for complexes between two ordered proteins (red) and one ordered and one disordered protein (blue). Both distributions were fit to a Gaussian distribution (solid lines). **(B)** Plot of ΔH° versus TΔS° for the same protein–protein complexes with the same color code as in **(A)**. The solid lines represent the best linear fits to the data.

the final complex, which inherently will be associated with a relative large loss in conformational entropy. It is important to note that it is not possible to conclude from equilibrium ITC data *when* the folding of the ligand occurs during the binding process. It is highly likely that for some of the protein complexes the IDP folds and then binds in a conformer selection process while for others the IDP folds upon binding in an induced fit process. The difference in <ΔG°> may still, however, be explained by the required folding of the disordered ligand in the ORD-IDP complexes.

We next analyzed the distribution of the ΔG°-values for the ORD-ORD and ORD-IDP complexes (**Figure 2B**). Interestingly, the most stable complexes ($\Delta G^\circ < -15\,\text{kcal mol}^{-1}$) are exclusively formed between two ordered proteins, and the least stable complexes ($\Delta G^\circ \sim -5\,\text{kcal mol}^{-1}$) are exclusively formed between an ordered and a disordered protein. Among the most stable complexes we find several enzyme: inhibitor complexes, such as the bacterial DNAses in complex with bacterial immunity proteins (Keeble et al., 2006). These DNAses form both very stable cognate complexes and less stable non-cognate complexes with immunity proteins. All these complexes are formed with similar on-rates in the order of $10^7\ \text{M}^{-1}\ \text{s}^{-1}$, and the stronger binding is achieved by slower off rates (Keeble and Kleanthous, 2005; Keeble et al., 2006). Another strong binding complex is that of barnase and barstar which has become a classical example where the electrostatic surfaces of the proteins have evolved to enhance the on-rate ($k_{\text{ass}} = 10^8 - 10^9\ \text{M}^{-1}\ \text{s}^{-1}$) (Schreiber and Fersht, 1996). Similar fast on-rates are reported for IDPs (Arai et al., 2012; Rogers et al., 2013; Dogan et al., 2014) and similar on-rate dependence on electrostatics has been noted (Rogers et al., 2013). Consequently, the main difference between ORD-IDP and ORD-ORD complexes seems not to reside in on-rate differences, but may therefore reside in off-rates, noted earlier in comparative kinetic studies (Huang and Liu, 2009; Shammas et al., 2012; Dogan et al., 2014). It is still possible that the electrostatic influence from a globular binding partner will cause an induction of a binding-competent conformation within the ensemble distribution of the IDP. A result of this is that it can potentially influence the on-rate and subsequently the binding energy. Alternatively, the binding-competent conformation of the IDP may be required to guide it into the electrostatic field of the globular partner. We do not currently have any data to elaborate further on these scenarios.

One of the hall-marks of IDPs is their ability to interact with many different proteins, for instance in cellular hubs (Oldfield et al., 2008; Cumberworth et al., 2013). Based on computational analyses of structures from a large set of both ordered and disordered hub-complexes from yeast, it was suggested that the binding energies become weaker as the number of interacting proteins increases (Carbonell et al., 2009). Thus proteins with only one binding partner bound with higher affinity than promiscuous proteins with more than on binding partner. This difference may possibly be caused by a broader distribution of hot spots in the promiscuous proteins (Carbonell et al., 2009). It is possible that the difference in average binding affinity <ΔG°> observed in the current set is related to an increased number of interacting partners. We have no data on the number of alternative binding partners for complexes in our analysis. Still, it is interesting to note that the difference in binding energy between specific-to-specific complexes and specific-to-promiscuous was $0.08 \pm 0.01\,\text{kcal mol}^{-1}\ \text{residue}^{-1}$ (Carbonell et al., 2009), which with an average of 47 residues in the interfaces provided in our data set, amounts to $3.8 \pm 0.5\,\text{kcal mol}^{-1}$, close to the average difference of $2.5 \pm 0.4\,\text{kcal mol}^{-1}$ that we found between the ORD-ORD and the ORD-IDP complexes.

One alternative explanation for the lower average stability of IDP-ORD complexes may be purely technical and unrelated to any *de facto* differences between globular proteins and IDPs. The vast majority of the experimental studies in our set are conducted on recombinant proteins, typically expressed in *Escherichia coli*. Since phosphorylations and other post translational modifications are widespread in IDPs and is a way of regulating their activity (Iakoucheva et al., 2004) the $2.5\,\text{kcal mol}^{-1}$ displacement of the average ΔG° could reflect the fact that some of the IDPs examined lack certain post-translational modifications that would be stabilizing to the interaction. However, the same argument may hold also for globular proteins and a phosphorylation may even destabilize a complex. The lack of other factors (chaperones, carrier proteins, methyl-groups, carbohydrates), which may alter the energy of binding *in vivo* cannot be excluded as origin for the displacement either, but again we see no reason why this should not be an even more pronounced effect for the ORD-ORD complexes.

Based on the data collected, we have reached the—perhaps—counterintuitive conclusion that interfaces formed between globular proteins and IDPs are not overall significantly different from the interfaces between two globular proteins, although the contribution of residues within the binding interface is slightly skewed. We find instead that there is a small but significant difference in the average binding free energy in favor of the ORD-ORD complexes. We suggest that this difference is primarily caused by the loss of conformational free energy upon IDP binding, which affects the off-rate of the complex, although other reasons may exists such as an increased number of binding partners for the IDP.

Finally, we would like to add that the present analysis almost exclusively involves binary complexes. It has been suggested that IDPs are particularly well suited as scaffolds for large complexes or as hubs for signaling assemblies. Therefore, we may have missed thermodynamic fingerprints that stand out and reveal IDPs that diverge more from their globular twins than the ones analyzed in the present paper. Allostery in IDP-interactions where more binding sites are in play is an emerging subject (Ferreon et al., 2013; Shammas et al., 2014) and in the ensemble view of allostery, IDP-linked negative and positive allostery is possible (Motlagh et al., 2014). This aspect is not decomposed in the present set of data and allostery may be one underlying cause of the observed differences. Also, highly fuzzy complexes acting e.g. as electrostatic clouds (Mittag et al., 2010; Fuxreiter and Tompa, 2012) are most likely not captured by the methods available for measuring the thermodynamics of protein interactions and are most definitely not targets for structure determination and hence do not contribute to the current analyses. Although the concept of fuzziness has emerged from studies on IDPs we cannot exclude that they also exist for complexes of two ordered proteins.

The reason, if any, for the evolution of protein intrinsic disorder remains to be disclosed. The present paper hints strongly that the answer does not lie directly in differences in thermodynamic parameters or the energetic principles of ligand binding.

Author Contributions

All authors contributed equally to the work and wrote the manuscript in collaboration.

Acknowledgments

This work was supported by a grant from the Danish Natural Research councils (grant number 12-12880 3 to BK).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmolb.2015.00040

References

Arai, M., Ferreon, J. C., and Wright, P. E. (2012). Quantitative analysis of multisite protein–ligand interactions by NMR: binding of intrinsically disordered p53 transactivation subdomains with the TAZ2 domain of CBP. *J. Am. Chem. Soc.* 134, 3792–3803. doi: 10.1021/ja209936u

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

Brady, G. P., and Sharp, K. A. (1997). Entropy in protein folding and in protein-protein interactions. *Curr. Opin. Struct. Biol.* 7, 215–221. doi: 10.1016/S0959-440X(97)80028-0

Carbonell, P., Nussinov, R., and del Sol, A. (2009). Energetic determinants of protein binding specificity: insights into protein interaction networks. *Proteomics* 9, 1744–1753. doi: 10.1002/pmic.200800425

Clackson, T., and Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* 267, 383–386. doi: 10.1126/science.7529940

Cumberworth, A., Lamour, G., Babu, M. M., and Gsponer, J. (2013). Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochem. J.* 454, 361–369. doi: 10.1042/BJ20130545

Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., et al. (2014). The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.* 42, D259–D266. doi: 10.1093/nar/gkt1047

Dogan, J., Gianni, S., and Jemth, P. (2014). The binding mechanisms of intrinsically disordered proteins. *Phys. Chem. Chem. Phys.* 16, 6323–6331. doi: 10.1039/c3cp54226b

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59. doi: 10.1016/S1093-3263(00)00138-8

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* 11, 161–171.

Eaton, B. E., Gold, L., and Zichi, D. A. (1995). Let's get specific: the relationship between specificity and affinity. *Chem. Biol.* 2, 633–638. doi: 10.1016/1074-5521(95)90023-3

Ferreon, A. C. M., Ferreon, J. C., Wright, P. E., and Deniz, A. A. (2013). Modulation of allostery by protein intrinsic disorder. *Nature* 498, 390–394. doi: 10.1038/nature12294

Fukuchi, S., and Nishikawa, K. (2001). Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J. Mol. Biol.* 309, 835–843. doi: 10.1006/jmbi.2001.4718

Fuxreiter, M., and Tompa, P. (2012). Fuzzy complexes: a more stochastic view of protein function. *Adv. Exp. Med. Biol.* 725, 1–14. doi: 10.1007/978-1-4614-0659-4_1

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93. doi: 10.1038/nature02555

Han, P., Zhang, X., Norton, R. S., and Feng, Z.-P. (2006). Predicting disordered regions in proteins based on decision trees of reduced amino acid composition. *J. Comput. Biol.* 13, 1723–1734. doi: 10.1089/cmb.2006.13.1723

Hansen, J. C., Lu, X., Ross, E. D., and Woody, R. W. (2006). Intrinsic protein disorder, amino acid composition, and histone terminal domains. *J. Biol. Chem.* 281, 1853–1856. doi: 10.1074/jbc.R500022200

Huang, Y., and Liu, Z. (2013). Do intrinsically disordered proteins possess high specificity in protein-protein interactions? *Chemistry* 19, 4462–4467. doi: 10.1002/chem.201203100

Huang, Y., and Liu, Z. (2009). Kinetic advantage of intrinsically disordered proteins in coupled folding–binding process: a critical assessment of the "Fly-Casting" mechanism. *J. Mol. Biol.* 393, 1143–1159. doi: 10.1016/j.jmb.2009.09.010

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., et al. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32, 1037–1049. doi: 10.1093/nar/gkh253

Iesmantavicius, V., Dogan, J., Jemth, P., Teilum, K., and Kjaergaard, M. (2014). Helical propensity in an intrinsically disordered protein accelerates ligand binding. *Angew. Chem. Int. Ed. Engl.* 53, 1548–1551. doi: 10.1002/anie.201307712

Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.* 41, 133–180. doi: 10.1017/S0033583508004708

Keeble, A. H., and Kleanthous, C. (2005). The kinetic basis for dual recognition in colicin endonuclease-immunity protein complexes. *J. Mol. Biol.* 352, 656–671. doi: 10.1016/j.jmb.2005.07.035

Keeble, A. H., Kirkpatrick, N., Shimizu, S., and Kleanthous, C. (2006). Calorimetric dissection of colicin DNase–immunity protein complex specificity. *Biochemistry* 45, 3243–3254. doi: 10.1021/bi052373o

Krieger, J. M., Fusco, G., Lewitzky, M., Simister, P. C., Marchant, J., Camilloni, C., et al. (2014). Conformational recognition of an intrinsically disordered protein. *Biophys. J.* 106, 1771–1779. doi: 10.1016/j.bpj.2014.03.004

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

Levy, E. D. (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* 403, 660–670. doi: 10.1016/j.jmb.2010.09.028

Mittag, T., Kay, L. E., and Forman-Kay, J. D. (2010). Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* 23, 105–116. doi: 10.1002/jmr.961

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., et al. (2006). Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* 362, 1043–1059. doi: 10.1016/j.jmb.2006.07.087

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

Nilsson, J., Grahn, M., and Wright, A. P. H. (2011). Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome Biol.* 12, R65. doi: 10.1186/gb-2011-12-7-r65

Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., and Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9(Suppl. 1), S1. doi: 10.1186/1471-2164-9-S1-S1

Rogers, J. M., Steward, A., and Clarke, J. (2013). Folding and binding of an intrinsically disordered protein: fast, but not 'diffusion-limited'. *J. Am. Chem. Soc.* 135, 1415–1422. doi: 10.1021/ja309527h

Schreiber, G., and Fersht, A. R. (1996). Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* 3, 427–431. doi: 10.1038/nsb0596-427

Shammas, S. L., Rogers, J. M., Hill, S. A., and Clarke, J. (2012). Slow, reversible, coupled folding and binding of the spectrin tetramerization domain. *Biophys. J.* 103, 2203–2214. doi: 10.1016/j.bpj.2012.10.012

Shammas, S. L., Travis, A. J., and Clarke, J. (2014). Allostery within a transcription coactivator is predominantly mediated through dissociation rate constants. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12055–12060. doi: 10.1073/pnas.1405815111

Stites, W. E. (1997). Protein-protein interactions: interface structure, binding thermodynamics, and mutational analysis. *Chem. Rev.* 97, 1233–1250. doi: 10.1073/pnas.0608582104

Teilum, K., Olsen, J. G., and Kragelund, B. B. (2009). Functional aspects of protein flexibility. *Cell. Mol. Life. Sci.* 66, 2231–2247. doi: 10.1007/s00018-009-0014-6

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* 27, 527–533. doi: 10.1016/S0968-0004(02)02169-2

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 18, 343–384. doi: 10.1002/jmr.747

Weathers, E. A., Paulaitis, M. E., Woolf, T. B., and Hoh, J. H. (2004). Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett.* 576, 348–352. doi: 10.1016/j.febslet.2004.09.036

Williams, D. H., Stephens, E., O'brien, D. P., and Zhou, M. (2004). Understanding noncovalent interactions: ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angew. Chem. Int. Ed. Engl.* 43, 6596–6616. doi: 10.1002/anie.200300644

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* 67, 235–242. doi: 10.1107/S0907444910045749

Wong, E. T. C., Na, D., and Gsponer, J. (2013). On the importance of polar interactions for complexes containing intrinsically disordered proteins. *PLoS Comput. Biol.* 9:e1003192. doi: 10.1371/journal.pcbi.1003192

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.

PERSPECTIVE
published: 12 August 2015
doi: 10.3389/fmolb.2015.00047

Anatomy of protein disorder, flexibility and disease-related mutations

Hui-Chun Lu [1], *Sun Sook Chung* [1, 2], *Arianna Fornili* [1, 3] and *Franca Fraternali* [1]*

[1] Randall Division of Cell and Molecular Biophysics, King's College London, London, UK, [2] Department of Haematological Medicine, King's College London, London, UK, [3] School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

Integration of protein structural information with human genetic variation and pathogenic mutations is essential to understand molecular mechanisms associated with the effects of polymorphisms on protein interactions and cellular processes. We investigate occurrences of non-synonymous SNPs in ordered and disordered protein regions by systematic mapping of common variants and disease-related SNPs onto these regions. We show that common variants accumulate in disordered regions; conversely pathogenic variants are significantly depleted in disordered regions. These different occurrences of pathogenic and common SNPs can be attributed to a negative selection on random mutations in structurally highly constrained regions. New approaches in the study of quantitative effects of pathogenic-related mutations should effectively account for all the possible contexts and relative functional constraints in which the sequence variation occurs.

Keywords: non-synonymous SNPs, protein disorder, order-disorder propensity, disease-related mutations, protein flexibility

Edited by:
Kris Pauwels,
Vrije Universiteit Brussel, Belgium

Reviewed by:
Elena Papaleo,
University of Copenhagen, Denmark
Sreenivas Chavali,
MRC Laboratory of Molecular Biology, UK

***Correspondence:**
Franca Fraternali,
Randall Division of Cell and Molecular Biophysics, King's College London,
New Hunt's House, Guy's Campus,
London SE1 1UL, UK
franca.fraternali@kcl.ac.uk

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

Received: *31 May 2015*
Accepted: *29 July 2015*
Published: *12 August 2015*

Citation:
Lu H-C, Chung SS, Fornili A and Fraternali F (2015) Anatomy of protein disorder, flexibility and disease-related mutations. Front. Mol. Biosci. 2:47. doi: 10.3389/fmolb.2015.00047

Introduction

Because of the intrinsic complexity of biological systems, reductionist approaches have traditionally been used that concentrate on carefully chosen sub-systems. The availability of complete genome sequences and large (but incomplete) collections of biomolecular structures at atomic resolution favors large-scale computational approaches to investigate multiple components and their interactions (Lu et al., 2013). The undisputed relationship between protein-coding elements and their protein products has dominated the field of genomics/proteomics research in the past and the relationship between structure and function has been widely investigated.

Large-scale studies have been performed on how disease-related mutations may disrupt protein functions and ultimately regulate the function of biological systems (Studer et al., 2013). Mutations are classified as "loss of function," "gain of function," or "neutral" according to their effect on protein function. These effects can be mediated by alterations of the protein stability induced by the mutation (Yue et al., 2005; Studer et al., 2013). The impact of SNPs on protein function and structural stability has been extensively studied at the level of the single protein (Yue et al., 2005; Schuster-Bockler and Bateman, 2008; Wang et al., 2012; Nishi et al., 2013; Studer et al., 2013; Yates and Sternberg, 2013; Scharner et al., 2014) and a number of predictors have been developed to evaluate the impact of SNPs on individual proteins (Thomas and Kejariwal, 2004; Capriotti et al., 2005; Bromberg and Rost, 2007; Adzhubei et al., 2010;

Reva et al., 2011; Al-Numair and Martin, 2013; Shihab et al., 2013; Pires et al., 2014; Yates et al., 2014). With the intention to expand the single protein structure-function paradigm, the interplay between Protein Protein Interactions (PPI) networks, structures, and disease mutations has been explored by several groups (see reviews Lu et al., 2013; Yates and Sternberg, 2013) and reference therein, Kelley et al., 2015; Mosca et al., 2015). Particularly the crucial role of interfaces in modulating the effects of pathogenic variation in binding and signaling (Stefl et al., 2013; Yates and Sternberg, 2013) has been generally accepted. In recent years additional findings have contributed to further expanding classical structure-function approaches: firstly, the widely recognized importance of non-coding elements (Necsulea and Kaessmann, 2014; Ling et al., 2015) (not discussed here) and the enrichment of SNPs in these (Consortium, 2012; Kircher et al., 2014); secondly, the role of unstructured regions, intrinsically disordered elements and flexibility in protein function versatility (Uversky, 2013; Dunker et al., 2015; Wright and Dyson, 2015). Even in the absence of intrinsic disorder, there is growing evidence that conformational flexibility is important in regulating protein-protein interactions (Dobbins et al., 2008; Stefl et al., 2013; Uversky, 2013). This effect has also been shown for proteins that have multiple partners (hubs) and are essential in protein-protein communication and signaling. Hubs' promiscuous binding sites have been demonstrated to display specific dynamical properties, pre-existing in the isolated state of the individual protein, allowing for polyvalent partner binding (Fornili et al., 2013). In any case, quantification of the occurrence of SNPs in disordered and flexible protein regions is a complex task, because different shades of disorder have been identified as playing a role in protein function stability and binding (Uversky et al., 2014; Wright and Dyson, 2015 and references therein). One particularly interesting case is represented by mutations related to disorder-to-order (D-O) transitions; there are often associated to post-translational modifications or with defense mechanisms to protect proteins from toxic aggregation and oxidative stress (Winter et al., 2008) and therefore may result in a stronger impact on the protein functional role. Consequently, order/disorder-sensitive descriptors of the specific chemico-physical environment in the vicinity of the observed variant are needed to evaluate rigorously the relationship between disorder and disease-related mutations.

We aim to contribute to this debate by exploring and quantifying in a systematic way the relationship between order/disorder and the occurrence of common variants (dbSNP: common variations from the 1000 Genomes project, Sherry et al., 2001), disease-related SNPs (OMIM: Mendelian genetic diseases, Hamosh et al., 2005) and cancer-related SNPs (COSMIC, Forbes et al., 2011). To this end we decompose the protein sequences anatomically in folded domain regions, unfolded-disordered (intra-domain) regions and inter-domain disordered regions and calculate the enrichment/depletion of SNPs in each of these regions. These comparisons based on mapping SNPs on static, crystallographic structures, represent a first step in quantifying the different roles played by the two (ordered vs. disordered) environments in which a common or pathogenic mutation may occur. We also explored scenarios of mutual effects of mutations in ordered regions on the disorder content within that domain. We discuss two cases of hubs that are strongly involved in cancer: BRAF (Haling et al., 2014; Thevakumaran et al., 2015) and JAK2 (Bandaranayake et al., 2012), both with phenotypic pathogenic mutations occurring in ordered regions and affecting the disorder content of distal sites in the domain.

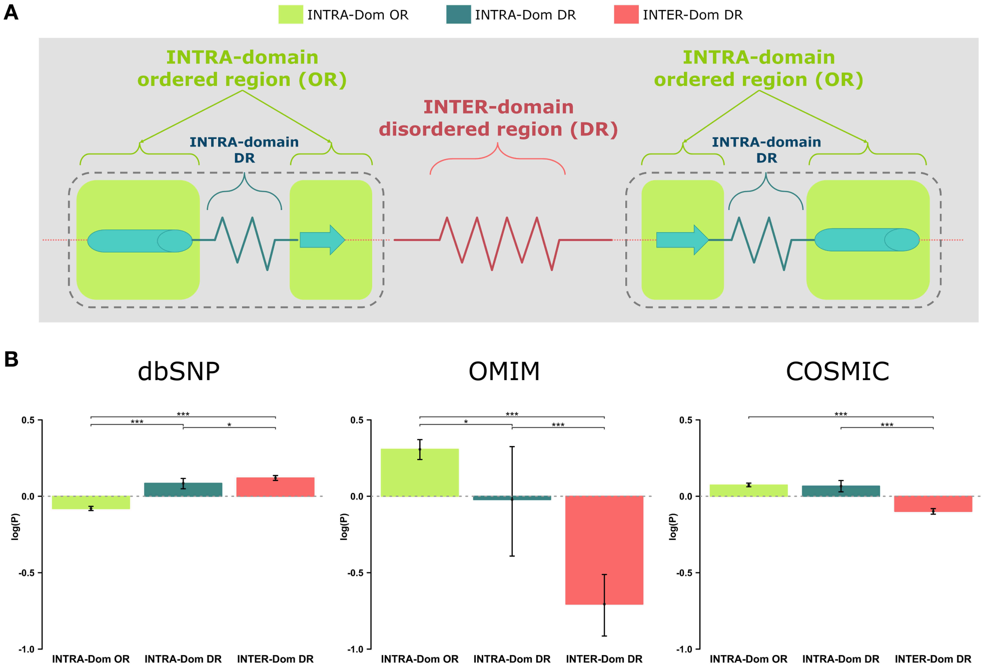
Results

Dearth of Disease-Related SNPs in Inter-domain Disordered Regions

The relative enrichment of SNPs in the dissected disordered regions of the protein have been analyzed by comparing three different classes of SNPs: (a) the common variants from the 1000 Genomes project (dbSNP) (Sherry et al., 2001), (b) the genetic-disease variants from OMIM (Hamosh et al., 2005), and (c) the COSMIC cancer-related SNPs (Forbes et al., 2011). Details on the enrichment/depletion measures are given in the Section "Strategy for the investigation of disordered regions and SNPs occurrence."

The outcome of our analysis is presented **Figure 1B** and the barplots relative to each region are colored according to the scheme in **Figure 1A**. The results for dbSNP data are reported as a comparison of the observed human variation in the analyzed regions vs. the pathogenic mutations observed for OMIM and COSMIC data. To our knowledge, this is the first time that such a comparison is presented. The results have been statistically tested (see Strategy section) and the *p*-values of the comparison tests between the distributions are annotated with stars to show their significance (see **Figure 1** legend for clarification).

The most striking difference amongst all data lies in the opposite behavior observed for INTER-domain disordered regions (INTER-Dom DRs, red) in dbSNP vs. OMIM and COSMIC data (enrichment vs. depletion, respectively). The reasons for such trend can be ascribed to the fact that common variations usually do not occur in structurally and functionally constrained regions but rather accumulate in disordered regions, particularly inter-domain ones. These are usually more flexible to allow the orientation of protein domains and binding multiplicity (Fong and Panchenko, 2010). Conversely, an opposite trend is observed for the INTRA-domain ordered regions (INTRA-Dom OR, light green) of dbSNP vs. the disease-related INTRA-Dom OR plots. For both the disease-related OMIM (**Figure 1B**, center) and COSMIC (**Figure 1B**, right) datasets, there is clear evidence that pathogenic mutations are enriched in ordered domain regions. These are the fragile sites that once mutated can cause a functional impairment of the protein either by destabilizing the fold (Studer et al., 2013), or by affecting structurally important regions for partner binding and consequent signaling activity (Yates and Sternberg, 2013). The enrichment in INTER-Dom disordered regions vs. INTRA-Dom ordered regions is particularly pronounced for the OMIM dataset, but also significantly important for the COSMIC data. The difference in the relative order/disorder populations of the two datasets might be related to the fact that mutations with Mendelian inheritance



C

	N(Proteins)	N(SNPs)	N(SNPs) INTRA-Dom OR	N(SNPs) INTRA-Dom DR	N(SNPs) INTER-Dom DR
dbSNP subset	5,504	148,828	58,486	9,255	48,402
<L>(region) ± SEM			268±3	36±1	182±4
OMIM subset	700	3138	2,013	213	344
<L>(region) ± SEM			355±13	52±5	167±10
COSMIC subset	5,422	102,972	46,870	6,253	26,458
<L>(region) ± SEM			268±4	36±1	180±4

FIGURE 1 | Analyses of non-synonymous single nucleotide polymorphisms (SNPs) in intra-domain ordered regions, intra-domain disordered regions and inter-domain disordered regions. (A) Scheme of protein regions. A protein contains (intra−)domain regions (dashed boundary line) and inter-domain regions. Domain regions contain ordered regions (INTRA-Dom ORs; light-green squares) and disordered regions (INTRA-Dom DRs; dark green zigzag line). Inter-domain regions are predominantly disordered (INTER-Dom DR; red zigzag line). **(B)** SNP frequency analysis. The propensity of SNPs P(SNP) to occur in each region was calculated using Equation 1. Average propensity values are reported as relative entropies log(P(SNP)). Error bars were estimated using bootstrap re-sampling with 10,000 replicates. Stars denote the alpha levels of the test statistics (*$p < 0.05$; ***$p < 0.001$). **(C)** Number of SNPs mapped onto different protein regions. The number of nsSNPs in each class and the average lengths of the protein regions are listed together with the standard error of the mean (SEM). The column "N(proteins)" contains the number of proteins selected for the study of a SNP class, while column "N(SNPs)" reports the total number of SNPs mapped onto the reference proteins.

are potentially more harmful to the protein than some of the passenger mutations observed in cancer.

Our results support previous studies that compared differences in "natural" mutations from dbSNP and disease-associated OMIM data (De Beer et al., 2013). The difference in order vs. disorder propensities observed in our study is therefore an additional discriminant in evaluating the mutability of proteins.

Examples of Intra-domain Mutations and Effects on Disorder Occurrence

In a number of recent studies it has been reported that disordered regions harbor pathogenic mutations (Iakoucheva et al., 2002; Uversky et al., 2008; Babu et al., 2011; Hu et al., 2011; Pajkos et al., 2012; Vacic and Iakoucheva, 2012; Vacic et al., 2012). Some of these observations referred to SNPs in segments involved in D-O transitions, but as we observed a clear dearth of pathogenic mutations in INTER-domain disordered regions (INTER-Dom DRs), we decided to investigate the occurrence of SNPs in INTRA-Dom DRs in more detail. A particularly interesting case is the mutual effect of intra-domain pathogenic mutations and disorder observed within the domain, even at sites distant from the original mutation. We found such examples in BRAF and JAK2 kinases, which are involved in cancer pathologies (Vogelstein and Kinzler, 2004).

We previously studied the BRAF V600E mutation that destabilizes the inactive conformation of the BRAF kinase and consequently induces ERK activation (Satoh et al., 2012; Lu et al., 2013). The V600 residue is in a cluster of hydrophobic residues with Phe468, therefore the presence of a negative charge (residue E) will be disruptive for this cluster, resulting in destabilization of the inactive conformation. Interestingly, introducing the V600E mutation in the BRAF protein kinase domain increases the INTRA-Dom DRs prediction, as shown in the table (**Figure 2A**) and the plot (**Figure 2B**). By running the DISOPRED2 predictor for the V600E mutant, one can observe an increase in the span of the predicted disordered region found in a distal site (607–611). Notably, the predicted disorder region span was not affected by mutations found within the INTRA-Dom DRs (yellow residues in **Figure 2A** for BRAF). These findings suggest that, besides destabilizing the hydrophobic cluster, the V-E substitution in the kinase domain (Pkinase_Tyr(PF07714)) might also have an effect on the INTRA-Dom disorder content by unwinding the downstream loop, as shown in the wild type 3D structure (4MNE_B) (Haling et al., 2014) (**Figure 2B** and Figure S1). This could in turn affect the ligand-binding region (structure 4WO5_A) (Thevakumaran et al., 2015), with a possible impact on the binding affinity.

The mutation V600E has been studied in detail by sophisticated enhanced sampling methods (Marino et al., 2015) and one of the main consequences of the pathogenic variant highlighted in this study is reflected the enhancement of the active-to-inactive state barrier and the increased flexibility (disorder) of the activation loop (region 602–612). These combined effects result in keeping the kinase in an active state and therefore favor phosphorylation to occur. This study supports the idea that an accurate descriptions of the structure, dynamics, and energetics of the protein and its mutated states are necessary to extract molecular fingerprints that rationalize the impact of pathogenic vs. commonly occurring mutations. Interestingly, in recent times the tendency of BRAF in adopting permanently an active state not detectable by current structure has been highlighted as one of the paradigmatic cases for which the currently adopted strategies for structure-based drug discovery may be ineffective (Holderfield et al., 2014).

Long-range effects of mutations on domain-disorder content are partially observed also for the V617F SNP of the JAK2 kinase, a mutation mostly observed in leukaemias. Our predictions indicate that this mutation leads to an extension of the INTRA-Dom DRs, (Bandaranayake et al., 2012) as shown in the table (Figure S2D) and the plot (Figure S2E).

The changes of disorder probability between the wild type sequences (BRAF and JAK2) and those with the cancer driver SNPs (V600E and V617F, respectively) have been predicted by five different methods which include highly ranked methods in CASP10 (Monastyrskyy et al., 2014) such as DISOPRED3, PrDOS, Biomine_MFDp and a recent method using backbone dynamics, DynaMine (Cilia et al., 2014) (Figures S2, S4). The results do not show a strict consensus in the boundaries and in the absolute differences of disorder content, this can be ascribed to the different algorithms used. However, most of the methods predict an increase of the disorder probability in the mutation distal regions we observe for BRAF that the cancer driver mutation is at the periphery of the kinase binding site and in an ordered region, while the non-driver mutations mostly occur in the disordered regions. The two locations seem to be correlated in the sense that the observed change in the driver mutation alters the disorder content of the other mutation loci. This may be ascribed to correlated dynamical couplings between disordered and ordered regions within the same protein domain that may lead to an enrichment of pathogenic variants in flexible and less structured regions. This long-range coupling is an indirect and probably down-tuned mutational effect on the protein function, which may result in a higher acceptance of the mutation in these regions.

Strategy for the Investigation of Disordered Regions and SNP Frequencies

Data Set Preparation

A data set of human proteins, using UniProt accession identifiers as reference, was generated by mapping SNPs onto experimentally resolved 3D structures of proteins. Native and homologous structures were identified by running NCBI-BLAST (version 2.2.29+) (Camacho et al., 2009) against the PDB sequence library. Homologues were accepted above the 30% sequence identity threshold. Non-synonymous SNPs were retrieved from the dbSNP database (build 141) (Sherry et al., 2001), germ-line disease-related SNPs were extracted from the "Online Mendelian Inheritance in Man" (OMIM) database (version July 2014) (Hamosh et al., 2005) and somatic cancer-related SNPs were taken from the "Catalog of Somatic Mutations in Cancer" (COSMIC) database (version July 2014) (Forbes et al., 2011). Only the proteins having a native/homologous structure and SNP information were selected, yielding a reference data set comprising 5587 proteins.

Definition of Protein Domains and Disordered Regions

The selected proteins were assigned with domain definitions and disordered region predictions. For each protein sequence of the reference data set as query, the HMM sequence aligner HMMER3 (Finn et al., 2011) was used to search against the Pfam domain sequence library (Pfam-A.hmm version 26.0) (Punta et al., 2012) and to assign the matched PFAM domain definition to the query protein, given the alignment *E*-value was smaller than 1e-3. Disordered regions of the selected proteins were predicted using the DISOPRED program (Ward et al., 2004). The combination of domain definitions and disordered region predictions leads to three distinct regional classes (**Figure 1A**): (1) intra-domain ordered region (INTRA-Dom OR), (2) intra-domain disordered region (INTRA-Dom DR), and (3) inter-domain disordered region (INTER-Dom DR).

SNPs Enrichment Analysis

We computed the regional enrichment/depletion of SNPs as propensities P(SNPregion) by normalizing the relative regional

A

Position	BRAF_WT	R444Q	V600E	S607P	H608R	H608L	Q609R	F610S	E611D
600	V . 0.008 (0.018)	V . 0.009 (0.019)	E . 0.008 (0.025)	V . 0.012 (0.020)	V . 0.006 (0.021)	V . 0.013 (0.026)	V . 0.003 (0.015)	V . 0.008 (0.016)	V . 0.007 (0.019)
601	K . 0.012 (0.025)	K . 0.014 (0.026)	K . 0.014 (0.031)	K . 0.017 (0.028)	K . 0.011 (0.026)	K . 0.016 (0.030)	K . 0.006 (0.021)	K . 0.013 (0.022)	K . 0.013 (0.027)
602	S . 0.018 (0.033)	S . 0.020 (0.035)	S . 0.019 (0.037)	S . 0.019 (0.037)	S . 0.016 (0.034)	S . 0.024 (0.038)	S . 0.010 (0.028)	S . 0.017 (0.030)	S . 0.019 (0.037)
603	R . 0.047 (0.038)	R . 0.049 (0.039)	R . 0.083 (0.042)	R . 0.061 (0.040)	R . 0.063 (0.039)	R . 0.070 (0.043)	R . 0.032 (0.032)	R . 0.039 (0.034)	R . 0.040 (0.042)
604	W . 0.043 (0.041)	W . 0.046 (0.042)	W . 0.057 (0.048)	W . 0.060 (0.042)	W . 0.066 (0.045)	W . 0.080 (0.048)	W . 0.033 (0.035)	W . 0.041 (0.037)	W . 0.046 (0.043)
605	S . 0.070 (0.045)	S . 0.076 (0.047)	S * 0.151 (0.055)	S . 0.090 (0.046)	S . 0.133 (0.049)	S . 0.161 (0.051)	S . 0.072 (0.039)	S . 0.067 (0.041)	S . 0.081 (0.047)
606	G . 0.137 (0.050)	G * 0.141 (0.053)	G * 0.235 (0.058)	G * 0.136 (0.052)	G * 0.194 (0.054)	G * 0.214 (0.057)	G . 0.139 (0.045)	G . 0.125 (0.046)	G * 0.141 (0.054)
607	S * 0.187 (0.060)	S * 0.202 (0.062)	S * 0.222 (0.068)	P * 0.208 (0.061)	S * 0.196 (0.063)	S * 0.216 (0.065)	S * 0.167 (0.054)	S * 0.154 (0.054)	S * 0.198 (0.066)
608	H * 0.242 (0.065)	H * 0.259 (0.067)	H * 0.271 (0.066)	H * 0.269 (0.065)	R * 0.223 (0.063)	L * 0.242 (0.065)	H * 0.223 (0.057)	H * 0.218 (0.057)	H * 0.290 (0.069)
609	Q * 0.328 (0.065)	Q * 0.326 (0.067)	Q * 0.267 (0.064)	Q * 0.331 (0.065)	Q * 0.284 (0.061)	Q * 0.305 (0.06)	R * 0.283 (0.055)	Q * 0.287 (0.057)	Q * 0.374 (0.069)
610	F * 0.257 (0.064)	F * 0.261 (0.065)	F * 0.211 (0.062)	F * 0.236 (0.063)	F * 0.219 (0.061)	F * 0.188 (0.061)	F * 0.209 (0.055)	S * 0.223 (0.058)	F * 0.275 (0.069)
611	E * 0.146 (0.062)	E * 0.155 (0.062)	E * 0.127 (0.059)	E * 0.123 (0.058)	E * 0.122 (0.061)	E * 0.099 (0.058)	E * 0.124 (0.053)	E * 0.136 (0.056)	D * 0.142 (0.065)
612	Q . 0.172 (0.051)	Q . 0.166 (0.052)	Q * 0.145 (0.053)	Q . 0.125 (0.046)	Q . 0.155 (0.051)	Q . 0.115 (0.048)	Q . 0.139 (0.044)	Q . 0.162 (0.046)	Q . 0.173 (0.051)
613	L . 0.113 (0.048)	L . 0.113 (0.049)	L . 0.092 (0.049)	L . 0.085 (0.044)	L . 0.100 (0.048)	L . 0.082 (0.045)	L . 0.096 (0.043)	L . 0.106 (0.044)	L . 0.102 (0.048)
614	S . 0.067 (0.049)	S . 0.067 (0.050)	S . 0.049 (0.046)	S . 0.055 (0.046)	S . 0.069 (0.046)	S . 0.048 (0.044)	S . 0.059 (0.043)	S . 0.062 (0.045)	S . 0.059 (0.051)
615	G . 0.042 (0.045)	G . 0.039 (0.045)	G . 0.031 (0.040)	G . 0.033 (0.043)	G . 0.039 (0.041)	G . 0.035 (0.039)	G . 0.036 (0.039)	G . 0.042 (0.041)	G . 0.031 (0.047)

B

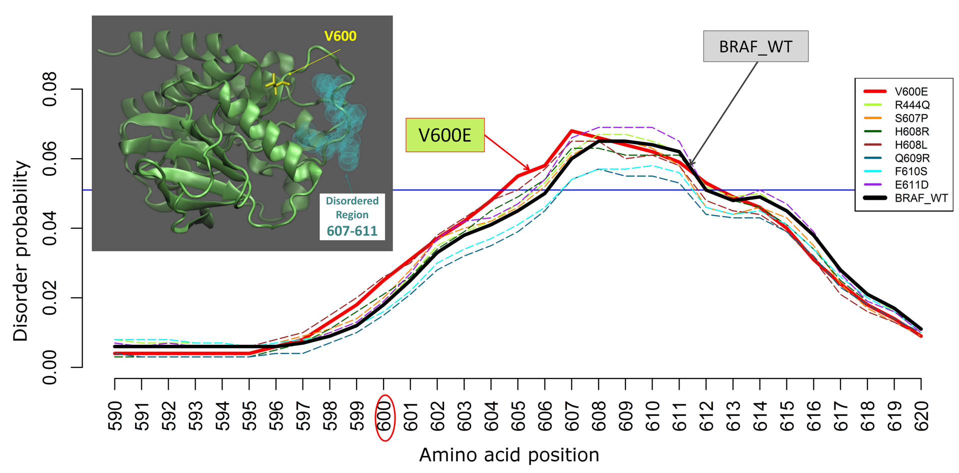


FIGURE 2 | Example of changes in disordered regions (DRs) conferred by SNPs in distant ordered regions. (A) Disorder prediction by DISOPRED2 of wild type (WT) and mutated sequence segments (600–615) of BRAF. Each column is labeled with the specific SNP used for DR prediction and contains the confidence scores of the DISOPRED2 prediction involving raw scores of disorder probability and their filtered scores with parentheses. The residues in DRs are annotated with (*) asterisks and colored in blue. SNPs within the sequence segment 600–615 are colored in yellow. **(B)** Plot of the DISOPRED2 filtered confidence scores of the BRAF WT and mutated sequences. The predicted behavior of V600E (red line) is distinct from that of the BRAF WT sequence (thick black line). The horizontal blue line indicates 5% of filter threshold of the method. The inset shows the 3D structure of the BRAF kinase domain (4MNE_B, cyan cartoon), the location of residue V600 (yellow licorice) and the predicted disordered positions (light green spheres).

frequency with the relative frequency over the total protein length (Equation 1).

$$P\left(SNP_{region}\right) = \frac{(N(SNPs)_{region}/length_{region})}{(N(SNPs)_{protein}/length_{protein})} \tag{1}$$

These propensities are plotted in **Figure 1B** as relative entropies log(P(SNPregion)). A relative entropy of zero indicates a regional frequency equal to the background frequency (denominator), positive values indicate relative enrichment and negative values correspond to relative depletion. All SNPs (from dbSNP, OMIM, and COSMIC) were mapped onto the protein sequences: 5504 of 5587 proteins were mapped with SNPs from dbSNP, 700 of 5587 with SNPs from OMIM and 5422 of 5587 with SNPs from COSMIC. SNPs from each database were further classified into different classes by mapping their positions onto the corresponding protein regions (INTRA-Dom OR SNPs, INTRA-Dom DR SNPs, and INTER-Dom DR SNPs). The number of SNPs in the different classes and the mean lengths of the regions are given in **Figure 1C**.

Statistical Evaluation

To obtain an estimate of the uncertainty associated with the propensity calculations and to reduce biases incurred by the protein selection procedure, we used a bootstrapping method (R function *boot*()) to create random re-sampled subsets of the reference data set. 10,000 independent subsets were generated of the SNPs propensities within each pre-defined protein regional class (INTRA-Dom OR SNPs, INTRA-Dom DR SNPs, and INTER-Dom DR SNPs) and the mean of each subset was computed. The distributions of the resampled means are by normally distributed, as expected (Figure S4). Confidence intervals at the 95% level were calculated from the bootstrap distributions. The statistical significance of differences between propensity distributions was calculated by Student's *t*-test on the confidence intervals (Wolfe and Hanley, 2002). The statistical analyses were performed using R (R Core Team, 2014).

Conclusions and Perspectives

We performed a large-scale statistical analysis of the relationship between protein disorder and disease-related mutations. We report that both genetic-disease variants from OMIM and cancer-related SNPs from COSMIC are depleted in disordered regions compared to common human variation. This is in line with the fact that mutations in highly constrained regions of the protein are more likely to be disruptive or deleterious. This is why mutations in ordered states of proteins (domains, ligand-binding sites, PPI sites) have been investigated quite in detail in the last years.

We offer here a starting and objective point to discriminate between completely ordered regions, disordered regions occurring in ordered domains, and inter domain predicted disordered segments. We observe and quantify the result of the mapping of available SNPs data onto a large set of human proteins and their close homologs. From this study a number of interesting cases can be extracted for functional validation and close investigation of the dynamical role played by the disorder content.

New perspectives in the field can be explored from this starting point, as the more complicate cases in which flexibility and/or disorder play a direct role in the protein function have not yet been fully elucidated. Particularly complex are the cases where flexible residues modulate protein binding and promiscuity (Fornili et al., 2013) and disorder-to-order causing mutations (Vacic and Iakoucheva, 2012; Dunker et al., 2015). These more "dynamically" driven processes are difficult to parametrise and the restraints playing a role in selecting the actual functional states are not always quantifiable. At this purpose, systematic studies collecting critical examples of experimentally proved correlations between flexibility, presence of disordered states, D-O transitions, and functional studies are needed in the field for the benchmarking and validation of predictive tools for the impact of pathogenic variation on proteins and their partners. Most recent development in disorder prediction methods exploits successfully the mutual interplay between backbone and side-chain dynamics (Cilia et al., 2014; Kosciolek and Jones, 2015).

Nevertheless, more sophisticated methods are needed to quantify these observations, like large-scale molecular simulations, [so far performed for isolated cases Vacic et al., 2012; Marino et al., 2015] and measurements of conformational signal transduction within protein structures (Pandini et al., 2012). The correlated dynamical couplings between disordered and ordered regions may be exploited in the design of drugs targeting distal sites from the dominant mutation, and by fine-tuning the effects on the overall protein function. Additionally, the possibility to predict the "allosteric" modulation of mutations occurring in regions with a different level of order/disorder and possibly correlated with the same or different pathogenic manifestation can open new avenues to investigate the underlying molecular mechanisms and rectify current strategies for drug-discovery.

Acknowledgments

We thank Dr. Jens Kleinjung and Dr. Nicholas Shaun B. Thomas for their critical reading of the manuscript. This research was supported by the Biotechnology and Biological Sciences Research Council (BB/H018409/1 to FF), the British Heart Foundation (FS/12/41/29724 to AF and FF) and the Leukaemia & Lymphoma Research (to FF). SSC is funded by a Leukaemia & Lymphoma Research Gordon Piller PhD Studentship.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmolb.2015.00047

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Al-Numair, N. S., and Martin, A. C. (2013). The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics* 14(Suppl. 3):S4. doi: 10.1186/1471-2164-14-S3-S4

Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* 21, 432–440. doi: 10.1016/j.sbi.2011.03.011

Bandaranayake, R. M., Ungureanu, D., Shan, Y., Shaw, D. E., Silvennoinen, O., and Hubbard, S. R. (2012). Crystal structures of the JAK2 pseudokinase domain and the pathogenic mutant V617F. *Nat. Struct. Mol. Biol.* 19, 754–759. doi: 10.1038/nsmb.2348

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35, 3823–3835. doi: 10.1093/nar/gkm238

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310. doi: 10.1093/nar/gki375

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.* 42, W264–W270. doi: 10.1093/nar/gku270

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247 de Beer, T. A., Laskowski, R. A., Parks, S. L., Sipos, B., Goldman, N., and Thornton, J. M. (2013). Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput. Biol.* 9:e1003382. doi: 10.1371/journal.pcbi.1003382

Dobbins, S. E., Lesk, V. I., and Sternberg, M. J. (2008). Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10390–10395. doi: 10.1073/pnas.0802496105

Dunker, A. K., Bondos, S. E., Huang, F., and Oldfield, C. J. (2015). Intrinsically disordered proteins and multicellular organisms. *Semin. Cell Dev. Biol.* 37, 44–55. doi: 10.1016/j.semcdb.2014.09.025

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Fong, J. H., and Panchenko, A. R. (2010). Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Mol. Biosyst.* 6, 1821–1828. doi: 10.1039/c005144f

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., et al. (2011). COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 39, D945–D950. doi: 10.1093/nar/gkq929

Fornili, A., Pandini, A., Lu, H.-C., and Fraternali, F. (2013). Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. *J. Chem. Theory Comput.* 9, 5127–5147. doi: 10.1021/ct400486p

Haling, J. R., Sudhamsu, J., Yen, I., Sideris, S., Sandoval, W., Phung, W., et al. (2014). Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer Cell* 26, 402–413. doi: 10.1016/j.ccr.2014.07.007

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi: 10.1093/nar/gki033

Holderfield, M., Deuker, M. M., McCormick, F., and McMahon, M. (2014). Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat. Rev. Cancer* 14, 455–467. doi: 10.1038/nrc3760

Hu, Y., Liu, Y., Jung, J., Dunker, A. K., and Wang, Y. (2011). Changes in predicted protein disorder tendency may contribute to disease risk. *BMC Genomics* 12(Suppl. 5):S2. doi: 10.1186/1471-2164-12-S5-S2

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 323, 573–584. doi: 10.1016/S0022-2836(02)00969-5

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892

Kosciolek, T., and Jones, D. T. (2015). "Investigations of structural ensembles and disorder-to-order transitions in intrinsically disordered proteins," in *3DSIG An ISMB Satellite Meeting*: *3DSIG Structural Bioinformatics and Computational Biophysics* (Dublin), 30.

Ling, H., Vincent, K., Pichler, M., Fodde, R., Berindan-Neagoe, I., Slack, F. J., et al. (2015). Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*. doi: 10.1038/onc.2014.456. [Epub ahead of print].

Lu, H. C., Fornili, A., and Fraternali, F. (2013). Protein-protein interaction networks studies and importance of 3D structure knowledge. *Expert Rev. Proteomics* 10, 511–520. doi: 10.1586/14789450.2013.856764

Marino, K. A., Sutto, L., and Gervasio, F. L. (2015). The effect of a widespread cancer-causing mutation on the inactive to active dynamics of the B-Raf kinase. *J. Am. Chem. Soc.* 137, 5280–5283. doi: 10.1021/jacs.5b01421

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., and Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins* 82(Suppl. 2), 127–137. doi: 10.1002/prot.24391

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Céol, A., Soler-López, M., et al. (2015). dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods* 12, 167–168. doi: 10.1038/nmeth.3289

Necsulea, A., and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* 15, 734–748. doi: 10.1038/nrg3802

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., et al. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS ONE* 8:e66273. doi: 10.1371/journal.pone.0066273

Pajkos, M., Meszaros, B., Simon, I., and Dosztanyi, Z. (2012). Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol. Biosyst.* 8, 296–307. doi: 10.1039/C1MB05246B

Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. (2012). Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* 26, 868–881. doi: 10.1096/fj.11-190868

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342. doi: 10.1093/bioinformatics/btt691

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39, e118. doi: 10.1093/nar/gkr407

Satoh, T., Smith, A., Sarde, A., Lu, H. C., Mian, S., Trouillet, C., et al. (2012). B-RAF mutant alleles associated with Langerhans cell histiocytosis, a granulomatous pediatric disease. *PLoS ONE* 7:e33891. doi: 10.1371/annotation/74a67f4e-a536-4b3f-a350-9a4c1e6bebbd

Scharner, J., Lu, H. C., Fraternali, F., Ellis, J. A., and Zammit, P. S. (2014). Mapping disease-related missense mutations in the immunoglobulin-like fold domain of lamin A/C reveals novel genotype-phenotype associations for laminopathies. *Proteins* 82, 904–915. doi: 10.1002/prot.24465

Schuster-Böckler, B., and Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biol.* 9:R9. doi: 10.1186/gb-2008-9-1-r9

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65. doi: 10.1002/humu.22225

Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* 425, 3919–3936. doi: 10.1016/j.jmb.2013.07.014

Studer, R. A., Dessailly, B. H., and Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.* 449, 581–594. doi: 10.1042/BJ20121221

Thevakumaran, N., Lavoie, H., Critton, D. A., Tebben, A., Marinier, A., Sicheri, F., et al. (2015). Crystal structure of a BRAF kinase domain monomer explains basis for allosteric regulation. *Nat. Struct. Mol. Biol.* 22, 37–43. doi: 10.1038/nsmb.2924

Thomas, P. D., and Kejariwal, A. (2004). Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15398–15403. doi: 10.1073/pnas.0404380101

Uversky, V. N. (2013). Under-folded proteins: conformational ensembles and their roles in protein folding, function, and pathogenesis. *Biopolymers* 99, 870–887. doi: 10.1002/bip.22298

Uversky, V. N., Dave, V., Iakoucheva, L. M., Malaney, P., Metallo, S. J., Pathak, R. R., et al. (2014). Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.* 114, 6844–6879. doi: 10.1021/cr400713r

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* 37, 215–246. doi: 10.1146/annurev.biophys.37.032807.125924

Vacic, V., and Iakoucheva, L. M. (2012). Disease mutations in disordered regions–exception to the rule? *Mol. Biosyst.* 8, 27–32. doi: 10.1039/C1MB05251A

Vacic, V., Markwick, P. R., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., et al. (2012). Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput. Biol.* 8:e1002709. doi: 10.1371/journal.pcbi.1002709

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* 10, 789–799. doi: 10.1038/nm1087

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164. doi: 10.1038/nbt.2106

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20, 2138–2139. doi: 10.1093/bioinformatics/bth195

Winter, J., Ilbert, M., Graf, P. C., Ozcelik, D., and Jakob, U. (2008). Bleach activates a redox-regulated chaperone by oxidative protein unfolding. *Cell* 135, 691–701. doi: 10.1016/j.cell.2008.09.024

Wolfe, R., and Hanley, J. (2002). If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *CMAJ* 166, 65–66. Available online at: http://www.cmaj.ca/content/166/1/65.full

Wright, P. E., and Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16, 18–29. doi: 10.1038/nrm3920

Yates, C. M., Filippis, I., Kelley, L. A., and Sternberg, M. J. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* 426, 2692–2701. doi: 10.1016/j.jmb.2014.04.026

Yates, C. M., and Sternberg, M. J. (2013). The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J. Mol. Biol.* 425, 3949–3963. doi: 10.1016/j.jmb.2013.07.012

Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353, 459–473. doi: 10.1016/j.jmb.2005.08.020

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: 27 May 2015
doi: 10.3389/fmolb.2015.00020

Protein flexibility in the light of structural alphabets

Pierrick Craveur [1, 2, 3, 4], Agnel P. Joseph [5], Jeremy Esque [6], Tarun J. Narwani [1, 2, 3, 4], Floriane Noël [1, 2, 3, 4], Nicolas Shinada [1, 2, 3, 4], Matthieu Goguet [1, 2, 3, 4], Sylvain Leonard [1, 2, 3, 4], Pierre Poulain [1, 2, 3, 4, 7], Olivier Bertrand [1, 3, 4], Guilhem Faure [8], Joseph Rebehmed [9], Amine Ghozlane [10], Lakshmipuram S. Swapna [11, 12], Ramachandra M. Bhaskara [11, 13], Jonathan Barnoud [1, 2, 3, 4, 14], Stéphane Téletchéa [1, 2, 3, 4, 15], Vincent Jallu [16], Jiri Cerny [17], Bohdan Schneider [17], Catherine Etchebest [1, 2, 3, 4], Narayanaswamy Srinivasan [11], Jean-Christophe Gelly [1, 2, 3, 4] and Alexandre G. de Brevern [1, 2, 3, 4]*

[1] Institut National de la Santé et de la Recherche Médicale U 1134, Paris, France, [2] UMR_S 1134, DSIMB, Université Paris Diderot, Sorbonne Paris Cite, Paris, France, [3] Institut National de la Transfusion Sanguine, DSIMB, Paris, France, [4] UMR_S 1134, DSIMB, Laboratory of Excellence GR-Ex, Paris, France, [5] Rutherford Appleton Laboratory, Science and Technology Facilities Council, Didcot, UK, [6] Institut National de la Santé et de la Recherche Médicale U964,7 UMR Centre National de la Recherche Scientifique 7104, IGBMC, Université de Strasbourg, Illkirch, France, [7] Ets Poulain, Pointe-Noire, Congo, [8] National Library of Medicine, National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA, [9] Centre National de la Recherche Scientifique UMR7590, Sorbonne Universités, Université Pierre et Marie Curie – MNHN – IRD – IUC, Paris, France, [10] Metagenopolis, INRA, Jouy-en-Josas, France, [11] Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Bangalore, India, [12] Hospital for Sick Children, and Departments of Biochemistry and Molecular Genetics, University of Toronto, Toronto, ON, Canada, [13] Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Frankfurt, Germany, [14] Laboratoire de Physique, École Normale Supérieure de Lyon, Université de Lyon, Centre National de la Recherche Scientifique UMR 5672, Lyon, France, [15] Faculté des Sciences et Techniques, Université de Nantes, Unité Fonctionnalité et Ingénierie des Protéines, Centre National de la Recherche Scientifique UMR 6286, Université Nantes, Nantes, France, [16] Platelet Unit, Institut National de la Transfusion Sanguine, Paris, France, [17] Institute of Biotechnology, The Czech Academy of Sciences, Prague, Czech Republic

OPEN ACCESS

Edited by:
Kris Pauwels,
Vrije Universiteit Brussel, Belgium

Reviewed by:
Mark Pfuhl,
King's College London, UK
Elena Papaleo,
University of Copenhagen, Denmark

***Correspondence:**
Alexandre G. de Brevern,
Institut National de la Santé et de la
Recherche Médicale U 1134, 6 Rue
Alexandre Cabanel, 75015 Paris,
France
alexandre.debrevern@
univ-paris-diderot.fr

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

Received: 28 February 2015
Accepted: 30 April 2015
Published: 27 May 2015

Citation:
Craveur P, Joseph AP, Esque J, Narwani TJ, Noël F, Shinada N, Goguet M, Leonard S, Poulain P, Bertrand O, Faure G, Rebehmed J, Ghozlane A, Swapna LS, Bhaskara RM, Barnoud J, Téletchéa S, Jallu V, Cerny J, Schneider B, Etchebest C, Srinivasan N, Gelly J-C and de Brevern AG (2015) Protein flexibility in the light of structural alphabets. Front. Mol. Biosci. 2:20. doi: 10.3389/fmolb.2015.00020

Protein structures are valuable tools to understand protein function. Nonetheless, proteins are often considered as rigid macromolecules while their structures exhibit specific flexibility, which is essential to complete their functions. Analyses of protein structures and dynamics are often performed with a simplified three-state description, i.e., the classical secondary structures. More precise and complete description of protein backbone conformation can be obtained using libraries of small protein fragments that are able to approximate every part of protein structures. These libraries, called structural alphabets (SAs), have been widely used in structure analysis field, from definition of ligand binding sites to superimposition of protein structures. SAs are also well suited to analyze the dynamics of protein structures. Here, we review innovative approaches that investigate protein flexibility based on SAs description. Coupled to various sources of experimental data (e.g., B-factor) and computational methodology (e.g., Molecular Dynamic simulation), SAs turn out to be powerful tools to analyze protein dynamics, e.g., to examine allosteric mechanisms in large set of structures in complexes, to identify order/disorder transition. SAs were also shown to be quite efficient to predict protein flexibility from amino-acid sequence. Finally, in this review, we exemplify the interest of SAs for studying flexibility with different cases of proteins implicated in pathologies and diseases.

Keywords: protein structures, disorder, secondary structure, structural alphabet, protein folding, allostery, protein complexes, protein—DNA interactions

Introduction

Analysis of protein structures is crucial to understand protein dynamics and functions. X-ray crystallography, the gold-standard method for solving 3D structures at atomic resolution, is impeded by protein dynamics. Hence, tricks are frequently used to restrict motions. It is why proteins have been often considered as static macromolecules, composed of *rigid* repetitive secondary structures and *less rigid* random coils. However, more and more emerging evidences show that protein structures are more complex with their internal dynamics being a key determinant of their function. Analyses of protein structures are often performed with a simplified three-state description known as α-helix, β-strand and coil which constitutes the classical secondary structures (Corey and Pauling, 1953; Kabsch and Sander, 1983). A more precise and complete description of protein backbone conformation exists based on the definition of libraries of small protein fragments, namely the structural alphabets (SAs) (Unger et al., 1989; Fetrow et al., 1997; Camproux et al., 1999; Offmann et al., 2007; Tyagi et al., 2007; Joseph et al., 2010a,b). SAs are designed to approximate every part of the local protein structures providing conformational detail. They have performed remarkably well spanning various problems in structural bioinformatics, from the characterization of ligand binding sites to the superimposition of protein structures (Joseph et al., 2010b). Furthermore, SAs are also very well suited to analyze the internal dynamics of protein structures. SAs have been used at three different levels to comprehend protein flexibility: (i) for studying specific fundamental biological and biomedical problems, (ii) to analyze changes associated with protein complexation and allostery, and (iii) to predict protein flexibility.

Here, we present state-of-the-art of developments in the study of protein flexibility using SAs based approximation. The backbone conformational variations can be described as changes in the pattern of SAs, which acts as fingerprints of the dynamics involved. These innovative approaches are useful, customizable, and deal with specific proteins involved in pathologies and diseases. They are also powerful to evaluate generalized principles from large biological complex structures. Thus, SAs provide new vision for detailed analysis and prediction flexibility of proteins.

The Different Views of Protein Structures

The primary sequence of the protein—the succession of amino acids—is assumed to encompass all the information necessary for its function. The protein structures resolved from X-ray crystallography or Nuclear Magnetic Resonance (NMR) (see **Figures 1A,B**) can be obtained in the Protein DataBank format (PDB, Bernstein et al., 1977; Berman et al., 2000). From the very beginning, theoreticians or experimentalists have described local protein structures by using three states (see **Figure 1C**, Corey and Pauling, 1953; Kabsch and Sander, 1983; Eisenberg, 2003). Two of them are repetitive structures stabilized by hydrogen bond patterns, namely the α-helices and the β-sheets (composed of β-strands). These structures are connected with more variable structures, i.e., random coil or loops. Later studies have identified spotted small repetitive and regular structures such as the β-hairpins or different kinds of turns in several protein structures (Richardson, 1981). These simplified descriptions were nicely represented with 3D visualization software (e.g., arrows for β-sheets, springs for α-helix) and accompanying the emergence of macromolecular crystallography. However these simplistic representations also contributed to the static and rigid views of these structures (Chavent et al., 2011).

In fact, growing evidence shows that proteins are highly dynamic macromolecules and that this dynamics is crucial in many biological processes. Thus, recent studies have demonstrated that conformational transitions in folded states of many proteins are essential to accomplish their functions, e.g., enzyme catalysis, activity regulation (Goh et al., 2004; Grunberg et al., 2004; Lensink and Mendez, 2008). Flexibility also allows interactions with different partners, with ligands by induced-fit interaction, with other proteins, or nucleic acids to form complex structures. NMR based methods and computational experiments such as Molecular Dynamic (MD) simulations, have largely contributed to gain valuable insights into the observation, understanding, and analyses of flexibility (Hirst et al., 2014). Flexibility can be versatile and covers a large range of timescales and amplitudes of structural modifications. It encompasses different kinds of conformational changes corresponding to (i) mobility of rigid part of the protein, e.g., domain motions (ii) deformability of the protein backbone, e.g., crankshaft motions or (iii) both. These different transitions are shown by analyzing and comparing protein structures (see **Figure 1D**). At a local level, the flexibility can be identified by the information contained in diffraction images of X-ray crystallography experiments and quantified along the refinement process through the Debye-Waller factors (expressed as surface units) also known as “B-factors” or temperature (displacement) factors. These so-called B-factors reflect atom mobility due to thermal vibration and measure the static disorder. They allow quantifying different levels of flexibility in proteins (see **Figure 1E**, Marsh, 2013). This criterion is also used by majority of flexibility prediction methods (from the sequence) (Schlessinger and Rost, 2005).

In this context, missing coordinates of whole residues in X-ray protein structures (usually labeled as missing residues, see **Figure 1F**) and several dedicated biochemical analyses have suggested these protein segments should be considered as disordered regions (see **Figure 1G**, Uversky et al., 2000; Dunker et al., 2001). From few years, beside the paradigm of a well-defined 3D folded state, new visions of protein structure and dynamics have emerged, namely the Intrinsically Disordered Proteins (IDP) or disordered regions. IDP may exhibit large structural rearrangements like the formation (then the loss) of secondary structures depending on the environment or the interacting partners. The impressive amount of research in this field is motivated by the implication of IDP in multiple crucial biological functions (Dunker et al., 2000; Dunker and Obradovic, 2001), for e.g., 14-3-3 proteins (Uhart and Bustos, 2014) or the Innate Antiviral Immunity (Xue and Uversky, 2014). Nevertheless, the regions with missing residues can be found resolved in other PDB structures of the same (or highly homologous) protein (see **Figure 1I**) (see

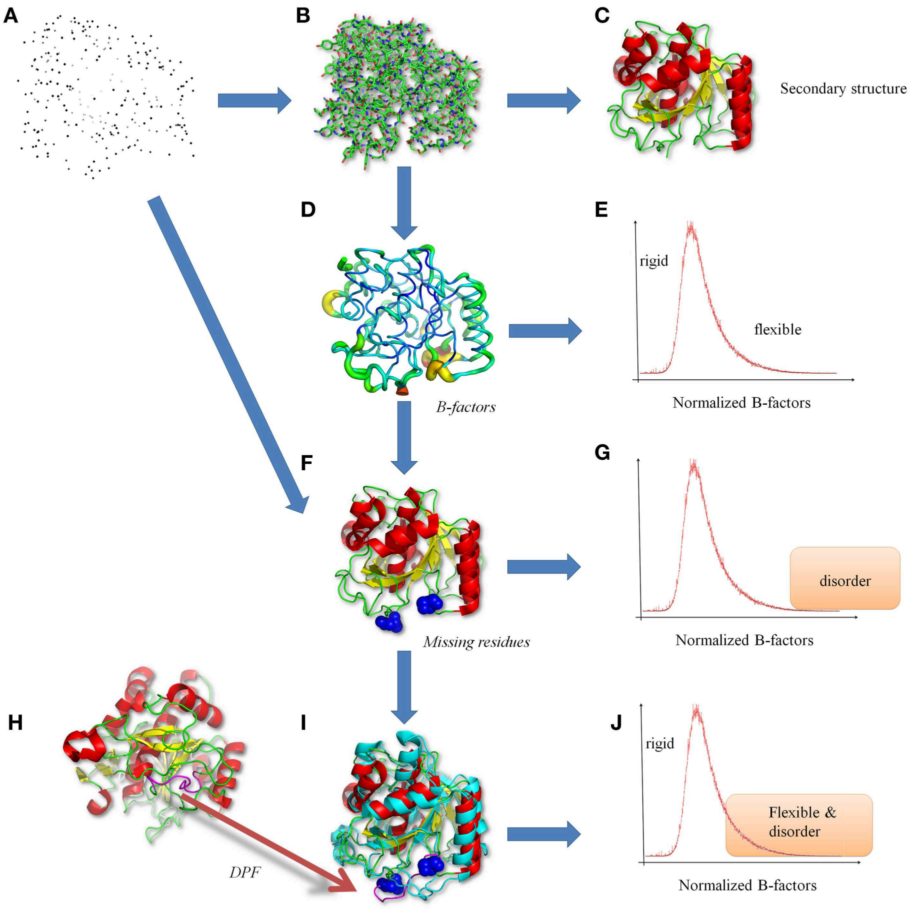


FIGURE 1 | Classical views of protein structures. (A) The protein structure is a file in PDB format (Bernstein et al., 1977; Berman et al., 2000), containing the 3D atomic coordinates. **(B)** The atoms are bound to build the protein backbone and side-chain residues. **(C)** From this information, secondary structures are performed (Kabsch and Sander, 1983). **(D)** From crystallographic data, B-factors are analyzed underlining **(E)** rigid to flexible residues. **(F)** More precise analyses shows missing residues revealing. **(G)** disorder regions (Uversky et al., 2000; Dunker et al., 2001). **(H)** Interestingly same or similar proteins can be found in the PDB (Berman et al., 2000), and **(I)** in numerous times with the missing regions resolved, leading to **(J)** a more complex definition and an ambiguity between flexibility and disorder. Protein visualization was created by the program PyMOL (http://www.pymol.org, Delano, 2013). The proteins used are two proteases (PDB codes 1dbi chain A, and 1wmd chain A, for this last only residues 1–306 are shown for more clarity).

Figure 1H, Berman et al., 2000). These ambiguous regions, termed Dual Personality Fragments (DPFs, see Dunker, 2007; Zhang et al., 2007), complicate the distinction and *per se* the definition of disorder versus flexibility (see **Figure 1J**). In **Figure 1**, we show a protease (PDB code 1dbi chain A), the corresponding DPF found (with a good resolution) in another protease (PDB code 1wmd chain A). Correlation between B-factors (representing flexibility) and disorder predictor outputs has been explored and shows a good agreement (Jin and Dunbrack, 2005; Schlessinger et al., 2009).

In the light of the above observations, the classic representation of protein structure as a succession of repetitive ordered secondary structures and random coil does not allow understanding of the complexity associated with structural flexibility. Actually, the coarseness of the secondary structure assignment may prevent from identifying conformational changes. Therefore distinction between flexible loops and rigid loops, for example, cannot be made on the sole basis of a three-state secondary structure assignment. A more precise and local description of protein structure is needed. In this regard, Structural Alphabet (SAs), allow to investigate primarily the complexity of the protein conformations, and consequently of their associated dynamics.

A SA is a library of *N* structural prototypes (the letters). Each prototype is representative of a backbone local structure of *l*-residues length. The combination of those structural prototypes is assumed to approximate any given protein structure. Many different libraries have been developed, (e.g., Unger et al., 1989; Fetrow et al., 1997; Camproux et al., 1999; Tung et al., 2007). Depending on the targeted accuracy, the length *l* and the number *N* can vary significantly. The length *l* typically ranges between 4

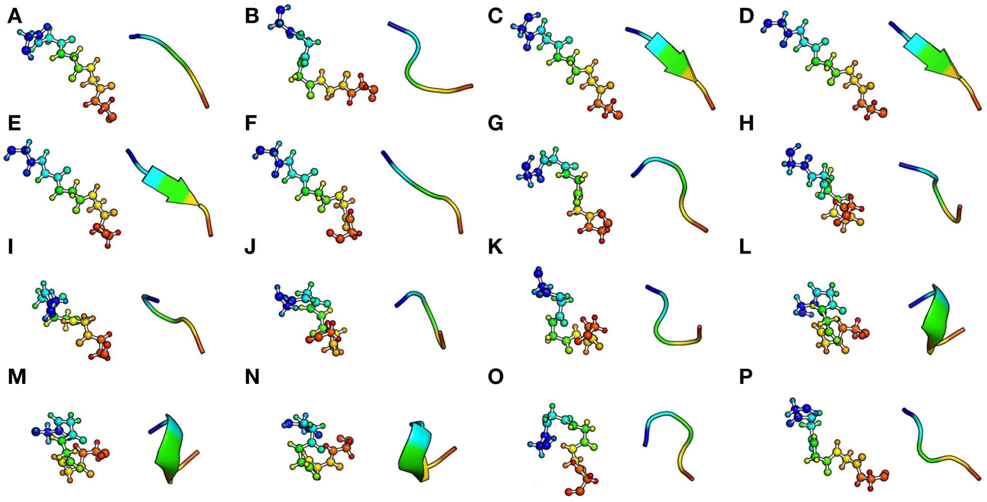


FIGURE 2 | The Protein Blocks structural alphabet. The conformation of each 16 pentapeptides is presented with ball and stick (left) and cartoon representation (right). The N-ter and C-ter extremity are respectively colored in blue and red. Visualization was created by the program PyMOL (http://www.pymol.org, Delano, 2013).

and 9 while can vary, the most frequent value being close to 20 (see Offmann et al., 2007; Joseph et al., 2010a,b for more details). The various structural alphabets also differ by the description parameters of the protein backbone. The description can be based on Cα coordinates, Cα-Cα distances, α or dihedral angles. The classification and learning methods that were used, are also various, e.g., hierarchical clustering, empirical function, Kohonen Maps, neural network or Hidden Markov Model Besides their interest to provide a finer description, They SA have been also designed for prediction purpose, which requires to decipher the sequence—structure relationship.

As example, in their respective work, Park and Levitt (1995) and Kolodny et al. (2002) aimed at finding representations based on smallest libraries of protein fragments to accurately construct protein structures. Fragments of four to seven residues long were considered in a library of 25–300 fragments. Micheletti et al. (2000) did similar studies and constructed a library that encompassed from 28 to 2561 recurrent local structures.

To date, one of the most developed and comprehensive SA is the Protein Blocks approach (PBs, de Brevern et al., 2000). This SA is composed by 16 local structure prototypes of 5 residues fragments (see **Figure 2**). It was shown to efficiently approximate every part of the protein structure. The PBs *m* and *d* can be roughly described as prototypes for the central region of α-helix and β-strand, respectively. PBs *a-c* primarily represent the N-cap of β-strand while *e* and *f* correspond to C-caps; PBs *g -j* are specific to coils, PBs *k* and *l* correspond to N cap of α -helix while PBs *n-p* to C-caps. PBs have been used to address various problems, including protein superimposition (Gelly et al., 2011; Joseph et al., 2012), general analyses of flexibility (Dudev and Lim, 2007; Wu et al., 2010) or and prediction of structure and flexibility (Zimmermann and Hansmann, 2008; Rangwala et al., 2009; Suresh et al., 2013; Joseph and de Brevern, 2014).

The assignment algorithm (see **Figure 3A**, de Brevern et al., 2000) runs through the 3D structure of the target protein, from the N to the C-ter of the sequence. The algorithm is iterative and uses 5 residues long overlapping windows over the entire sequence to assign a PB to every position. For each "n^{th}" position of the structure, 8 dihedrals $\psi\ (n-2)$, $\varphi\ (n-1)$, $\psi\ (n-1)$, $\varphi\ (n)$, $\psi\ (n)$, $\varphi\ (n+1)$, $\psi\ (n+1)$, $\varphi\ (n+2)$ are compared to each of the 16 PBs. The comparison is made by a least squares approach to match the RMSDA criteria (*Root mean square Deviation on Angular Values*) (Schuchhardt et al., 1996):

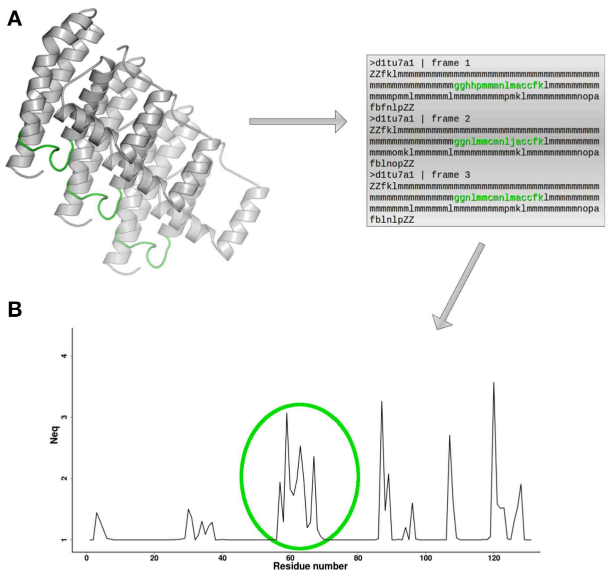
$$RMSDA\,(V_1, V_2) = \sqrt{\frac{1}{2(M-1)} \sum_{i=1}^{i=M-1} [\psi_i(V_1) - \psi_i(V_2)]^2 + [\varphi_{i+1}(V_1) - \varphi_{i+1}(V_2)]^2} \quad (1)$$

RMSDA formula

where V_1 is the vector of 8 dihedral angles extracted from the 5 residues long window, and V_2 is the 8 vector of dihedral corresponding to the individual PB type. The PB with the lowest RMSDA, is assigned to the corresponding position for that window. This PB captures the overall local conformation and approximates the transition along the main-chain smoothly.

PB assignments can be done using the Python PBxplore tool (https://github.com/pierrepo/PBxplore, in preparation). The result is a translation of a 3D structure into a 1D sequence of PBs.

Interestingly, the subtle differences between protein conformations can be captured by the assignment of the PB sequences. By analyzing the variation of PBs assigned at a given position for multiple conformers, the local conformational properties and corresponding changes can be easily identified. Moreover, a quantification of the flexibility at a given position *n* can be obtained by calculating, the average number of PBs across a set of conformers in this position or the "equivalent number" of PBs (N_{eq}). N_{eq} is based on a statistical metric similar



```
>d1tu7a1 | frame 1
ZZfklmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
mmmmmmmmmmmmmmmgghhpmmmnlmaccfklmmmmmmmmm
mmmmpmmlmmmmmmlmmmmmmmmmpmklmmmmmmmmmnopa
fbfnlpZZ
>d1tu7a1 | frame 2
ZZfklmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
mmmmmmmmmmmmmmmggnlmmcmnljaccfklmmmmmmmmm
mmmmomklmmmmmmlmmmmmmmmmmmklmmmmmmmmmnopa
fblnopZZ
>d1tu7a1 | frame 3
ZZfklmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
mmmmmmmmmmmmmmmggnlmmcmnlmaccfklmmmmmmmmm
mmmmmmmlmmmmmmlmmmmmmmmmpmklmmmmmmmmmnopa
fblnlpZZ
```

FIGURE 3 | N_{eq} and local flexibility. (A) For each conformation extracted from MD simulation, a PB sequence is assigned. **(B)** N_{eq} profile provides direct identification of protein fragments in which local conformational change is observed. Here, in green, is indicated a flexible loop. The protein 3D structure representation is generated using PyMOL software (http://www.pymol.org, Delano, 2013).

to Shannon entropy (de Brevern et al., 2000) and is calculated as follows:

$$N_{eq} = \exp(-\sum_{x=1}^{16} f_x \, ln \, (f_x)) \qquad (2)$$

N_{eq}formula

where f_x is the frequency of PB x (x takes values from a to p). A N_{eq} value of 1 indicates that only one type of PB is observed, while a value of 16 is equivalent to a random distribution. For example N_{eq} value equal to 6, could mean that 6 different PBs are observed in equal proportions (1/6), or that more than 6 PBs are observed in different proportions. By plotting the computed value for each residue position (see **Figure 3B**), it is possible to easily localize which protein regions present local conformation change, or in other words, which regions represent local flexibility.

This PB derived-entropy index is an interesting feature of PBs, which can be used to analyze PB prediction (de Brevern et al., 2000) or an ensemble of structures, corresponding to the same protein solved in different experiments, or to several structures extracted from MD simulation (Jallu et al., 2012). Note that PBxplore can be used to calculate N_{eq}, and to visualize in various ways the PB variation for each position from a collection of models or through a MD trajectory (de Brevern et al., 2005).

Other interesting SAs used in the flexibility context. We have proposed an extension of our SA through a novel library consisting of 120 overlapping structural classes of 11-residues fragments, firstly defined as PBs series (Benros et al., 2006). This library was constructed with an original unsupervised structural clustering method called the Hybrid Protein Model (de Brevern and Hazout, 2003). For each class, a mean representative fragment, or "local structure prototype" (LSP), correctly approximate the local structures with an average Cα RMSD of 1.61 Å. LSPs capture both the continuity between the identified recurrent local structures and long-range interactions. From this description, two methodologies were developed to predict flexibility. The first one was based on simple logistic functions and supervised with a system of experts (Benros et al., 2006). The second one was a combination of Support Vector Machines (SVMs) and evolutionary information (Bornot et al., 2009).

Pandini and co-workers developed their own SA; it is derived from the notion of attractors in conformational space, a more complex approach than PBs (Pandini et al., 2010). Pandini and co-workers developed their own SA; it is derived from the notion of attractors in conformational space, a more complex approach than PBs (Pandini et al., 2010). They focused on four-residue long fragments, the conformation of each being defined by internal angles between Cα atoms, i.e., *two* pseudo-bond angles and one pseudo torsion angle. All protein fragments were mapped as points in a three-dimensional space of these internal angles. The optimal number of clusters, i.e., structural prototypes, was assessed by the quality of the reconstructed protein structures and by information content. They ended with an alphabet of 25 letters, called M32K25. The alphabet starts from extended structures (e.g., A letter) and ends with turns (e.g., Y letter), passing through loops (e.g., P letter) and helical structures (e.g., U letter). The authors compared their approach with other SAs of four-residue fragments and showed the superiority of their method (Camproux et al., 2004; Tung et al., 2007). An interesting point was the analysis of the correlation between local flexibility and variability in the assignment. Thereafter, they have developed GSATools, (http://mathbio.nimr.mrc.ac.uk/wiki/GSATools, Pandini et al., 2013), composed of a set of programs, that encode ensembles of protein conformations into alignments of structural strings using their Structural Alphabet. This software package is particularly well suited for the investigation of the conformational dynamics of local structures, the analysis of functional correlations between local and global motions, and the mechanisms of allosteric communication. It performs a wide range of statistical analyses using a various set of external tools, mainly from R (Ihaka and Gentleman, 1996) and Python (Python Software Foundation, 2015). The software has been integrated into the GROMACS environment (Lindahl et al., 2001; Van Der Spoel et al., 2005). The user must compile it specifically.

GSATools was used to finely analyse the NtrC receiver domain and its homologs CheY and FixJ. For this purpose, different conformations of the protein extracted from a MDs simulation were encoded. The distributions of SA strings were used to compute different mutual information matrices using information theory. Remarkably, they were able to detect allosteric signal transmission from protein dynamics (Pandini et al., 2012). They also applied this methodology to a larger set of related proteins to show how evolutionary conservation and binding promiscuity have opposite effects on intrinsic protein dynamics (Fornili et al., 2013). Other examples are provided in Section 4.

These innovative approaches have been useful to study specific proteins implicated in pathologies and diseases. They are also sufficiently powerful to analyze large datasets of protein structures using automated pipelines. To summarize, SAs provide new visions for the analyses and prediction of protein structure flexibility. Different examples will be detailed in the following sections.

Duffy Antigen/Chemokine Receptor (DARC) Protein

Using the approaches described above, we analyzed conformations of different proteins implicated in pathologies. A very first study was done on predicting flexibility of loops in the Duffy antigen/receptor for chemokine (DARC) protein (Cutbush and Mollison, 1950; Compton and Haber, 1960). DARC is a transmembrane protein localized in the plasma membrane of red blood cells. It is a non-specific receptor for several chemokines (Allen et al., 2007); it is also named atypical chemokine receptor 1, Fy glycoprotein (FY), or CD234 (Cluster of Differentiation 234). The transmembrane chemokine receptors comprise two main families, defined by differences in their ligands. Indeed, chemokines can contain either two consecutive Cysteines (the CC chemokines) or two adjacent Cysteines with one amino acid in-between (the CXC chemokines). Furthermore, the two families of chemokine receptors have a specific linear sequence motif in their C-terminus region that enables signal transduction. In contrast, DARC lacks the specific motif, thus showing a specific difference coming probably from a distinct evolution.

This protein is also known as the receptor for the human malarial parasites *Plasmodium vivax* and *Plasmodium knowlesi* (Miller et al., 1975, 1976). Polymorphisms of DARC are the basis of the Duffy blood group system. While malaria is the most important sickness associated with DARC (Guerra et al., 2006; Cutts et al., 2014), DARC plays also a role in numerous other diseases, such as HIV and cancer, and risk factor associated with many other diseases is emerging (Liu et al., 1999; Horne and Woolley, 2009).

Like most transmembrane proteins, no experimental structure of DARC is currently available (de Brevern et al., 2005). We designed a structural model based on a comparative modeling approach. Using rhodopsin (the only available related structure at this time) as a structural template (a simple alignment showed a very low sequence identity value of 12%, e.g., close to a random value), we carefully built different structural models, based on a hierarchical and iterative procedure. A first step was to predict using more than 10 methods the positions of the 7 transmembrane helices along the sequence. From this initial and rough model, helices of DARC were aligned with rhodopsin helices assigned from the 3D structure. The same methodology was used for the loops, a complete alignment was generated using helices and connecting loops. A specific treatment was done for N- and C-termini region, combining Protein Blocks prediction (de Brevern et al., 2004; Etchebest et al., 2005) with threading approaches.

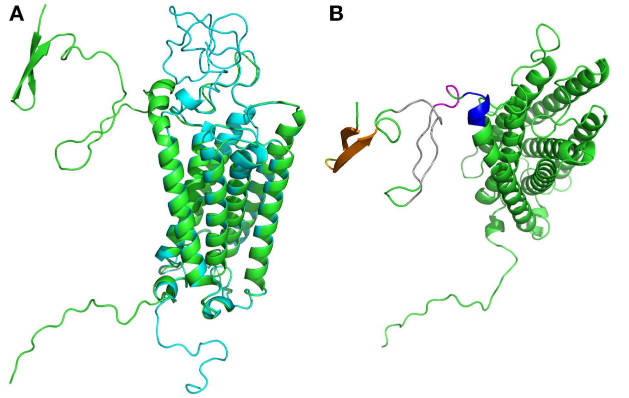


FIGURE 4 | Duffy Antigen/Receptor for Chemokines structural models. **(A)** Are shown the two best structural models obtained (de Brevern et al., 2005), in blue the compact one, in green the located far away. They are near identical on transmembrane domains. **(B)** The different regions with significant local conformational tendencies are shown in other colors. The most important ones are within the first 15 residues (predicted as disordered), with two β-stands (in orange) connected by a short turn (in yellow), later two long regions show some local conformational tendencies to be in extended conformation (in gray color) while the end of this N-terminus is packed in helical conformation (in pink and in violet). Visualization was created by the program PyMOL (http://www.pymol.org, Delano, 2013).

Experimentally, 40 Alanine mutants had been produced and associations binding constants with CXC-L8 were evaluated (Tournamille et al., 2003, 2005). We used these experiments to assess the quality of our best refined models. From the results, we generated new models by manually changing the positions of helices (and the alignments). Building and refinements were done 10 times until a proper set of characteristics were obtained. In regards to these experiments, *in silico* analysis of protein flexibility has underlined specific characteristics of different epitopes and interaction regions.

Interestingly, we obtained two different conformations (see **Figure 4A**) that were both as compatible with experimental data and similarly scored by the few assessment approaches available for transmembrane structural models. Interestingly five years later, an attempt to generate better models with the best available methods was not crowned with success (de Brevern et al., 2009; Smolarek et al., 2010).

It took us one year to build such models (models are available at Model Archive website (http://modelarchive.org/, Schwede et al., 2009). The $N_{terminus}$ is particularly important in the infection by *Plasmodium vivax* (Batchelor et al., 2014). It is nearly 55 residues long and different disorder prediction methods (i.e., DisEMBL, Linding et al., 2003 or PrDOS, Ishida and Kinoshita, 2007), predicted as partially disordered, with the beginning of the sequence as fully disordered.

To evaluate the different conformational states of the Duffy protein, we carried out numerous MDs simulated annealing simulations with the GROMACS software (Lindahl et al., 2001; Van Der Spoel et al., 2005). MD simulated annealing allows a harsh sampling of the conformational space by crossing energetic barriers in an efficient and fast way. Many runs were performed and the different conformations obtained at room temperature were analyzed using Protein Blocks. In practice, we encoded each 3D protein structural model conformation into a 1D string (the length of the protein sequence) using Protein Blocks. Then, we computed, the number of times each PB was observed for each position. Positions with a high frequency of a single PB exhibit no local change, while some others positions exhibit local deformations that require a more in-depth analyses. Few variations could be observed in the helical regions (PB *m* and encompassing PBs) that were weakly restrained with harmonic forces. Instead, loops sampled large regions of the conformational space. A very interesting result was observed for the $N_{terminus}$ region, and especially the distal region. In contrast to what was suggested by disorder predictors, this region was not a random coil region, but in fact a small β-sheet composed of two β-strands (PBs *d* and encompassing PBs, seen in orange on **Figure 4B**), connected by a short turns (in yellow). In the β-sheets, some positions, e.g., 12 and 13, were invariant. Likewise, the closest region to the first helix was more constrained than expected and not disordered (in pink and violet). Even the central regions (in gray) showed some tendencies to be structured. It was a striking example of a complex series of conformations which cannot be analyzed for instance through classical secondary structure (Kabsch and Sander, 1983).

A second example on DARC loops was the last extra-cellular loops for which a specific and constrained loop conformation was observed. Remarkably, this unexpected conformation explains a "lethal" mutation for the binding of CXCL8. It was the first time a structural alphabet was used to analyze the dynamics of a protein structures or structural model.

Human Integrin α2bβ3

In another project, we were interested in integrins, a large family of cell surface receptors involved in cell—cell or cell—matrix adhesion. Integrins are type I membrane glycoproteins composed of two distinct α and β subunits. Each subunit has a large extracellular region (composed of multiple structural domains), a trans-membrane segment and a short intracellular domain. Integrins interact with cell cytoskeleton and mediate bi-directional trans-membrane signal transduction. These receptors are expressed in vertebrate, but also in lower metazoans including sponges, nematode *Caeorhabditis elegans* and fruitfly *Drosophila Melanogaster*. In mammals, 18 α and 8 β subunits assemble in 24 distinct integrin complexes. Integrins play critical roles in many physiological processes like hemostasis, immune response, leukocyte trafficking, development and angiogenesis or in pathology like cancer. In human, they are responsible for many diseases from genetic or immune origins. They also make effective targets for drug therapies in thrombosis and inflammation. Furthermore, integrins are binding sites for many viruses and bacteria (Hynes, 2002; Takada et al., 2007).

In regard to these various characteristics, integrins have been extensively studied over the past decades. Especially, structural analyses have provided substantial insights to explain functional mechanism(s). In 2004, the first structure of the extracellular domain of αVβ3 integrin, a vitronectin receptor found in platelets, was proposed (Xiao et al., 2004). Then, several structures of αVβ3 but also of αIIbβ3 integrin (Zhu et al., 2008), a fibrinogen receptor involved in platelet aggregation, were resolved in different activation states. Molecular models for both trans-membrane and cytoplasmic domains were also proposed. Thus, it opens the way to investigate impact of mutant using *in silico* mutagenesis.

Hence, we examined the effect of the β3-Leu253Met substitution of αIIbβ3 complex in patients with Glanzmann thrombasthenia (Jallu et al., 2010), a rare bleeding disorder characterized by an impaired platelet aggregation (George et al., 1990). For the first time, we showed that residue Leu253—localized at the interface of the complex—is playing a major role in the stability of αIIbβ3. Nonetheless, structural models reflecting static specific states do not depict structural dynamics accompanying the various aspects of integrin functions. For instance, when integrins are activated by substrates, large conformational changes are observed. Analyses of static structures (e.g., B-factor, electrostatics), give only a limited view of the protein complex behavior, contrary to MDs simulations which are able to some extent, to reproduce the inner dynamics of protein structures.

α and β subunits of integrins are associated to rigid, flexible and even disorder properties (such as Duffy protein presented in the section above). We ran independent MDs simulations on different systems, i.e., the wild type but also variants and mutants, using GROMACS MDs package (Van Der Spoel et al., 2005) to examine specific regions of αIIbβ3. We observed different opposite behaviors depending on the region and mutants studied.

Hence, we studied the $Cab3^{a+}$ alloantigen resulting from a Leu841Met substitution in the αIIb chain. This polymorphism might result in severe life-threatening thrombocytopenias. $Cab3^{a+}$ corresponds to a Leu841Met mutation. We evaluated the flexibility by using N_{eq} index and found that this polymorphism locates in a very flexible sequence in the wild type (with a $N_{eq} >$ 4), but the mutation did not modify the N_{eq} behaviors (Jallu et al., 2013). Moreover, no change in the secondary structure content, neither the PBs adopted by residues of encompassing sequences change. Hence, intriguingly, this substitution would have little effect, if any, on the backbone structure of the peptide 829–853. It must be noticed that disorder prediction does not show this region has flexible property, i.e., prediction with IUPred (Dosztanyi et al., 2005) or DisEMBL (Linding et al., 2003).

In Caucasian population, the Human Platelet Alloantigenic (HPA) system 1 is involved in most neonatal thrombocytopenias (NAITP) and post-transfusion purpura (PTP) (Espinoza et al., 2013). The HPA-1 system results from a Leucine to Proline substitution in position 33 of the β3 chain (alleles HPA-1a and HPA-1b, respectively) in platelet αIIbβ3 integrin (Jallu et al., 2012). Alloantibodies to the HPA-1a variant can induce very severe immune thrombocytopenia (Espinoza et al., 2013). Furthermore, the Pro33 allelic variant of β3 is considered as a risk factor of thrombosis in patients with cardiovascular diseases.

To compare the HPA-1a and -1b variants, we have proposed for the first time to use a combination of standard analysis of flexibility (namely Root Mean Square Fluctuation, RMSF) and Protein Blocks analyses. MD simulations have revealed that (i) the Leu33Pro substitution of the β3 knee (a domain of β3 integrin chain) leads to adverse structural effects not highlighted by static models; and (ii) that these alterations can explain the increased adhesion potential of HPA-1b platelets to fibrinogen and the possible thrombotic risk associated with the HPA-1b phenotype (Jallu et al., 2012). These molecular simulations also support a novel structural explanation for the epitope complexity of the HPA-1 antigen (Jallu et al., 2012).

Although not yet known to be involved in an alloimmune response, a third variant discovered more recently and characterized by a Valine in position 33 of β3, was also examined. Analyses of the protein flexibility properties can mainly explain the variable reactivity of anti-HPA-1a alloantibodies. This result suggests that dynamics plays a key role in the binding of these alloantibodies. Unlike the L33P substitution which increases the local structure flexibility, the L33V transition would not affect the local structure flexibility, and consequently the functions of αIIbβ3 (Jallu et al., 2014). Although, this region is considered as rigid by disorder prediction, both RMSF and PBs analysis shows a high mobility. This behavior may be explained by a local rigidity, surrounded by deformable regions.

Figure 5 represents another MDs simulation focusing here only on the Calf-1 domain (a domain of α2b integrin chain), using same parameters as before. Simulations were analyzed through PB approaches underlining its interest for flexibility studies using PBxplore. **Figures 5A,B** show the superimposition of two distinct snapshots (in red and in yellow) extracted from the MDs simulation. **Figure 5C** shows the frequency of PBs at each position, calculated along the MD trajectory, and represented as a WebLogo graphic (Crooks et al., 2004) obtained with PBxplore. WebLogo (Crooks et al., 2004) summarizes this information with an entropy of every PBs at each position. **Figure 5D** is the superimposition of N_{eq} and RMSF. Interestingly, even though some regions show similar tendencies, namely large RMSF associated with large N_{eq}, other regions exhibit different and even opposite tendencies. For example, focusing on the residues near position 66 of Calf-1, the RMSF given **Figure 5G**, is the highest one (in blue on **Figure 5E**) as highly flexible, but it is not the case as the N_{eq}-values for this residue is not high. Therefore, these residues appear to be a mobile region between two deformable regions. This example confirms the interest to examine N_{eq} index beside RMSF because each measure brings related but different information on flexibility.

Figure 6 shows the structural alphabet distribution during the simulation obtained with GSATools (see Section The Different Views of Protein Structures). The most frequent letters seen (in black) are from the beginning of the alphabet, underlining its all-β composition (**Figure 6A**). The decomposition by this SA shows a large number of conformational changes at each position of the sequence. Only few positions, e.g., 10, 131, and 44 represented by B, H, and X letters, respectively, remained unchanged during the Calf-1 simulations. The transition probability matrix calculated between SA letters (**Figure 6B**) reflects how the local structure changes occur. Along the diagonal, high values are found, the highest ones being for letters N and X while the lowest ones being for letters U and Y The Mutual Information (MI) matrix presented in **Figure 6C** describes the correlation of local conformational changes among the protein fragments. Significant off-diagonal values are found but actually they correspond to strands forming β-sheets. Hence, in contrast to the examples detailed in (Pandini et al., 2012, 2013), the all-β conformation of the protein impedes to enlighten long-range correlations, except between β–strands close in 3D and found all along the protein sequences. The Shannon entropy per position shows quite similar profile between β –strands (mainly between 1.0 and 2.5 bits). All the lowest values correspond to residues inside loops. One of the most interesting features of GSATools is the graph representation of the correlated local motions from the MI matrix; it describes the relative importance of the nodes in the network useful to analyze allosteric behaviors. **Figure 6F** is a visualization of the two most important peaks underlined, they are found far away from the rigid β -sheet region.

Protein Complexes and Allostery

It is well documented that protein–protein interactions are often guided by flexibility (Jones and Thornton, 1996; Salwinski et al., 2004) and that alternative conformations can have a significant influence on the binding process. It is why predicting the structure of a complex using the unbound structures of the partners remains highly challenging, despite a scrutinizing examination of the amino acid composition of the interface (Janin et al., 2008). Thus, in most cases, protein structures change during the formation of the complex. The changes can be limited to few side chains motions but can also correspond to major reorganization in the fold. Therefore, we undertook the analysis of the protein–protein complexes in the light of structural alphabet. We compared proteins 3D structures in free form, and as part of larger macromolecular complexes.

The building of the protein dataset was quite strict leaving only 76 high quality complexes representing very different configurations with free and bound forms (Swapna et al., 2012). Accordingly, structural changes occurring between the free and bound forms of the protein were analyzed using three different measures: the Cα root mean square deviation, the percentage of PB change and a specific PB substitution score. This last score relies on a PB structural substitution matrix that quantifies the cost to replace a given PB by another PB. The more similar the PBs, the more favorable the substitution score. Consequently, this score permits to quantify the conformational change by distinguishing similar PBs from to the most distinct ones. Comparison between unbound and bound forms shows that significant structural rearrangement occurs at the interface but also in regions away from the interface upon the formation of a highly specific, stable and functional complex. For 50% of them, which correspond to signaling proteins, the major changes correspond to allosteric ones, localized far away from the interface. These sites could be associated to mutations known to be involved in multiple diseases such as cancer. PB allows distinguishing here also between large movements, from mobility to deformability or flexibility. Normal Mode Analysis was also performed to gain deeper insights (Swapna et al.,

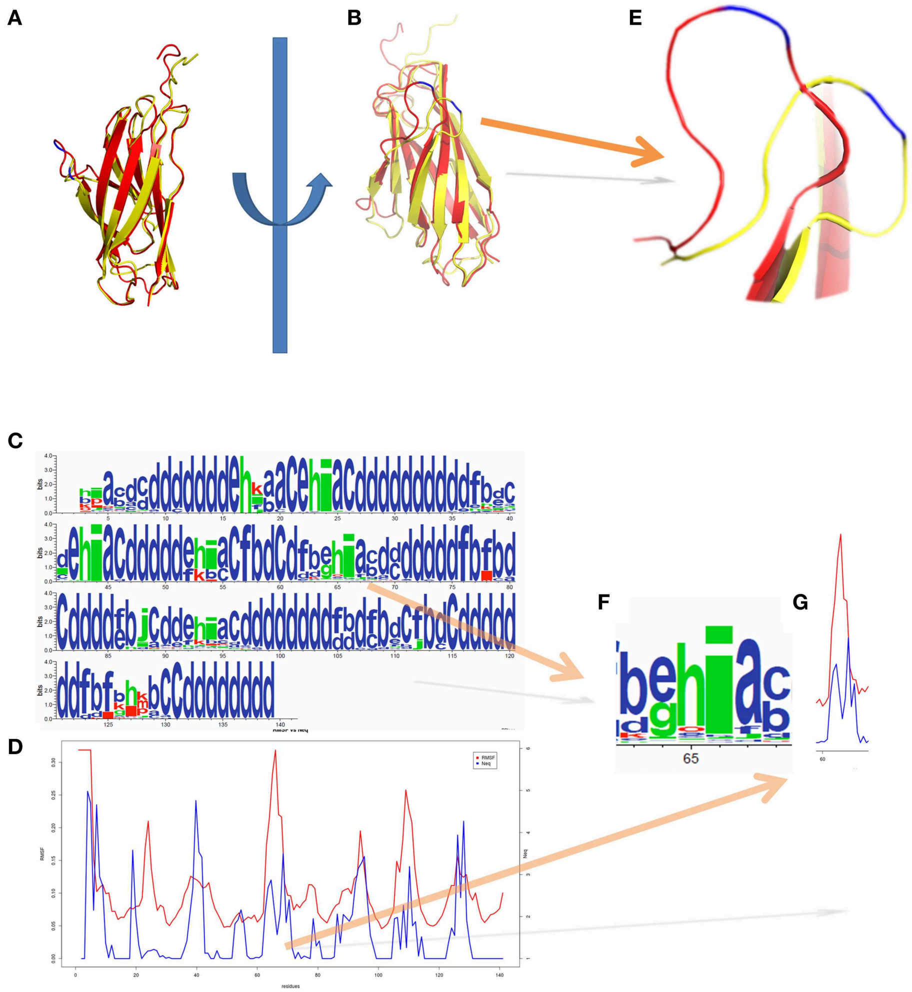


FIGURE 5 | Integrin Calf-1 domain simulation using PBxplore. (A,B) show the superimposition of two distinct snapshots (in red and in yellow) extracted from a molecular dynamics. **(C)** Shows the PB distribution in terms of WebLogo (Crooks et al., 2004) obtained thanks to PBxplore, and **(D)** represents the superimposition of N_{eq} and RMSF for the whole domain. **(E)** Zoom on the loop containing the residue 66 of Cab-1 domain (blue) which shows a dedicated **(F)** PB pattern *i* with a **(G)** low N_{eq} and high RMSF, i.e., a mobile position in a "flexible" region. Visualization was created by the program PyMOL (http://www.pymol.org, Delano, 2013).

2012). The results obtained for signaling complexes underline the importance of allostery-like structural changes much more than appreciated before (see **Figure 7**).

Flexibility becomes a critical issue in complexes especially the ones involving intrinsically disordered protein. Fine analyses have shown that disordered proteins can also adopt well-defined conformations in their bound form; their inherently dynamic nature is cast into their complexes (Meszaros et al., 2011). Protein families with more diverse interactions exhibit less average disorder over all members of the family (Fong and Panchenko, 2010). Inter-domain linkers are evolutionarily well conserved and are constrained by the domain-domain interface interactions (Bhaskara et al., 2013). An interesting resource is the ComSin database which provides a collection of structures of proteins solved in unbound and bound form, targeted toward disorder–order transitions (Lobanov et al., 2010).

Protein/DNA Interfaces

Beside protein-protein interactions, which govern many biological functions, fundamental biological processes like transcription also require complex formation, i.e., between protein and DNA. As for protein-protein interaction, complexation can change structures of both partners, but most studies focused on the protein side. Most of protein/DNA interfaces only extend the classical approaches to analyze protein/protein interfaces or protein/ligands interface. For instance, in PDIdb (Ferrada and Melo, 2009) or Biswas and coworkers studies, the interface is classified into core and rim regions, the first one being more sequentially conserved. Biswas and coworkers proposed a new classification scheme for the interfaces based on the composition of secondary structures (Biswas et al., 2009). Beyond this description in terms of

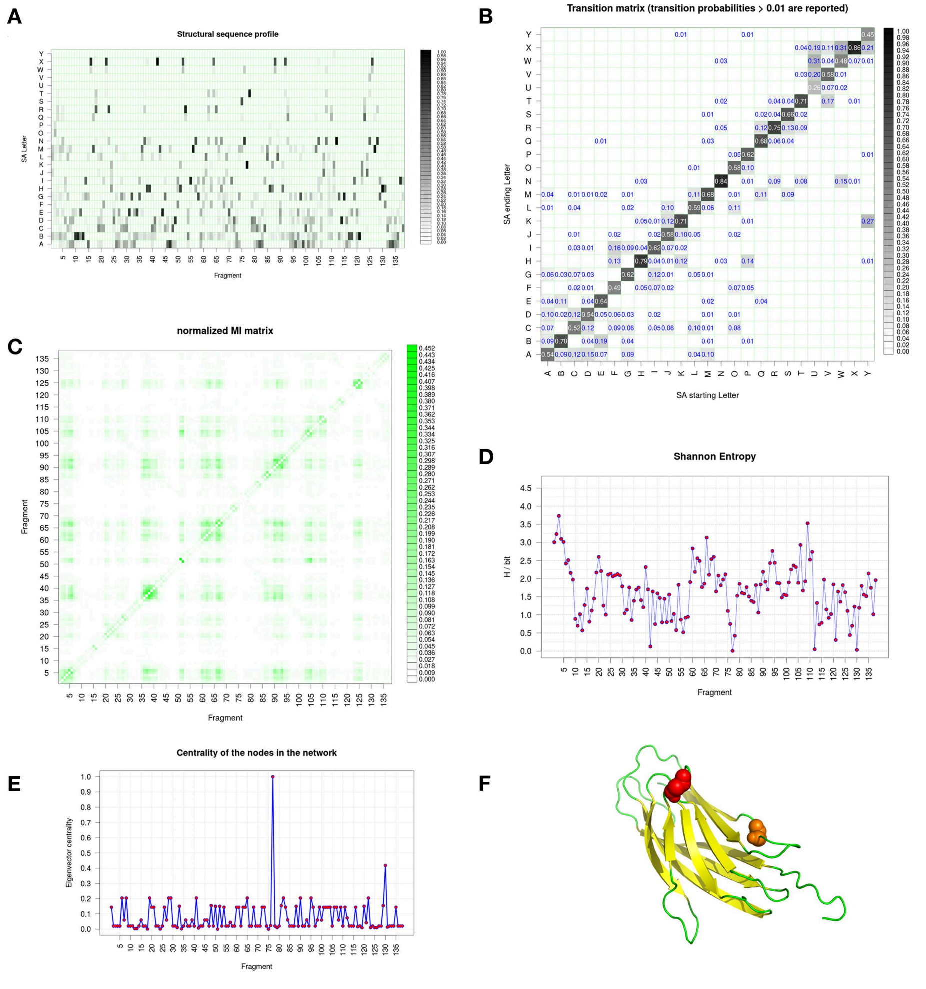


FIGURE 6 | Integrin Calf-1 domain simulation analyses using GSATools. (A) shows the structural alphabet distribution during the simulation, i.e., the sequence profile of the alignment, **(B)** the transition probability matrix for the transitions between SA letters, which reflect local conformational changes, **(C)** the Mutual Information (MI) matrix describing the correlation of local conformational changes among the protein's fragments, **(D)** the Shannon entropy per position, and **(E)** the graph representation of the correlated local motions from the MI matrix, with the eigenvector centrality, which describes the relative importance of the nodes in the network. **(F)** Is a visualization created by the program PyMOL (http://www.pymol.org, Delano, 2013) of the two most important peaks underlined by **(E)**.

regular local structures, Sunami and Kono (2013) conducted a quantitative analysis to understand the conformational changes in proteins when they bind to DNA. They compared DNA-free and DNA-bound forms of proteins and used structural alphabets to describe conformational changes in 4-residue fragments. They found that (i) three specific alphabets appeared in the DNA interfaces, (ii) conformational changes in DNA interfaces are more frequent than in non-interfaces and importantly, (iii) regions involved in DNA interfaces have more conformational variations in the DNA-free form. This study underlines also the importance of intrinsic flexibility of interacting regions to fit into DNA structure.

Another recent analysis has explored an extensive set of protein/DNA complexes and looked at conformational changes occurring in proteins but also in DNA. Importantly, for both molecules, structural alphabets were used. The alphabet used for describing protein backbone is the Protein Blocks. For DNA, a structural alphabet was obtained using a new approach of registering torsion angles of a dinucleotide unit combined with Fourier averaging and clustering (http://www.dnatco.org/, Svozil et al., 2008; Cech et al., 2013). These structural alphabets describe biopolymer conformations at greater detail than the 3-state protein secondary structure and basic DNA structural types such as A, BI and BII. **Figure 8** shows an example of different conformations. This study compared structural features of the protein/DNA interface with the features of non-interacting parts of protein and DNA molecules. Clear differences in preferences for occurrences of local protein and DNA conformations were observed. Specific preferences were underlined between complexes containing various types of proteins such as transcription factors and nucleases. Minor DNA conformers are often significantly enriched at the interface so that

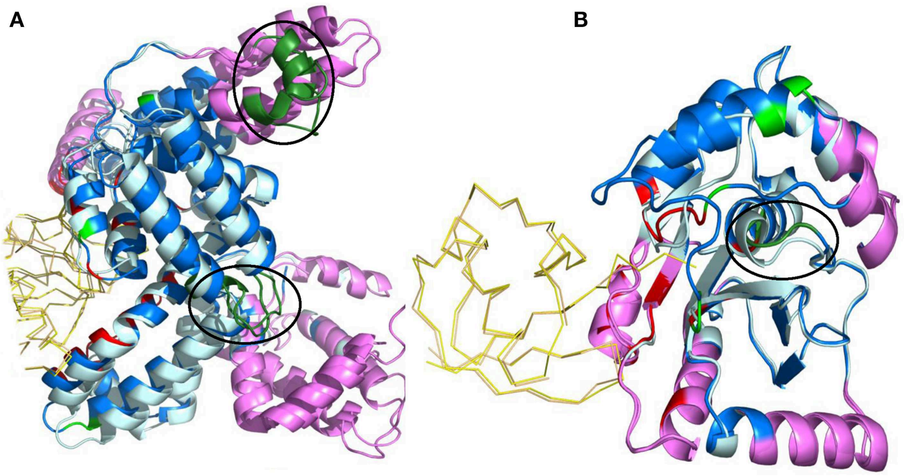


FIGURE 7 | Normal mode analysis of structural changes in regions of low B-factor far from interface. The protein containing the region of interest is depicted as cartoon and the interface of the other protein in ribbon. Unbound and bound forms of the protein of interest are in pale cyan and marine blue, respectively. The partner protein's unbound and bound forms are in light orange and yellow, respectively. Interacting residues are in red and non-interacting residues with PB change in green. All regions of interest are marked with a black circle, irrespective of whether they are intrinsically mobile or rigid. Regions identified to be intrinsically mobile according to NMA are in violet. Regions of interest occurring within the intrinsically mobile segments are in dark green. The complexes shown are **(A)** α-actin and Vitamin D - binding protein (PDB code 1KXP, Otterbein et al., 2002) **(B)** Ubiquitin Carboxyl-terminal esterase L3 protein and Ubiquitin complex (PDB code 1XD3, Misaghi et al., 2005). These figures show that non-interacting regions observed to undergo conformational changes upon complexation are usually intrinsically mobile, which is a characteristic of a functional site. Visualization was created by the program PyMOL (http://www.pymol.org, Delano, 2013).

the ability of DNA to adopt non-canonical conformers, rare in naked DNA, is clearly essential for the recognition by proteins. Rare DNA conformations introduce significant deformations to the DNA regular structure. The occurrence of these rare forms was estimated and characterized enabling a better understanding of the role of non-B-DNA structures. A critical feature was the distinct interaction patterns for the DNA minor groove relative to the major groove and phosphate, and the importance of water-mediated contacts. Indeed, water molecules mediate a proportionally largest number of contacts in the minor groove and form the largest proportion of contacts in complexes of transcription factors (Schneider et al., 2014). It corroborates to previous researches on the importance of mobility of such water molecules (Luo et al., 2011; Russo et al., 2011).

The above-discussed analyses pointed to some remarkable features about the protein/DNA interfaces, so that we performed a more specific analysis of the protein and DNA dynamics based on crystal structures. The analysis of B-factors (Schneider et al., 2014) showed that the dynamics of biopolymer residues, amino acids and nucleotides, as well as ordered water molecules is first of all a function of their neighborhood: amino acids in the interior of proteins have the tightest distribution of their displacements, residues forming the biopolymer interfaces (protein/protein or protein/DNA) intermediate, and residues exposed to the solvent the widest distribution (**Figure 9**). This general picture is best pronounced for structures with the highest crystallographic resolution since discrimination of different types of residues in structures becomes unclear with lower crystallographic resolution. Besides, amino acid residues in the protein core display a unique feature: their backbone and side chain atoms have virtually identical B-factor distributions. The protein core is therefore extremely well packed leaving minimum free space for atomic movements. B-factors of water molecules bridging protein and DNA molecules were surprisingly significantly lower than B-factors of DNA phosphates; in opposite, solvent-accessible phosphates were extremely flexible. An unexpected conclusion of this analysis is that a part of the observed trends could be due to improper refinement protocols that may need slight modifications (Schneider et al., 2014). Hence, the B-factors of high-resolution structures reflect the expected dynamics of residues in protein–DNA complexes but the B factors of lower resolution structures should be treated cautiously. Based on such kinds of ideas, Vriend proposed a dedicated dataset of refined B-factors (http://www.cmbi.umcn.nl/bdb/, Touw and Vriend, 2014).

PTMs

As seen in the previous sections, protein flexibility is essential for interactions between proteins and ligand, nucleic acid, or protein partners. Apart from interaction with partners, chemical modifications like formation or breaking of covalent bonds, can impact structural and dynamics properties. One of the most spectacular examples is depicted by the serpin family members when they interact with the protease (see **Figure 10**, Huntington et al., 2000; Kim et al., 2001). An initial large conformational change, consecutive to the cleavage of the reactive center of the serpin by the protease, occurs. The loop involved in the cleavage moves, folds as a β-strand that inserts between the other strands of the β-sheet composing the serpin protein core. The two

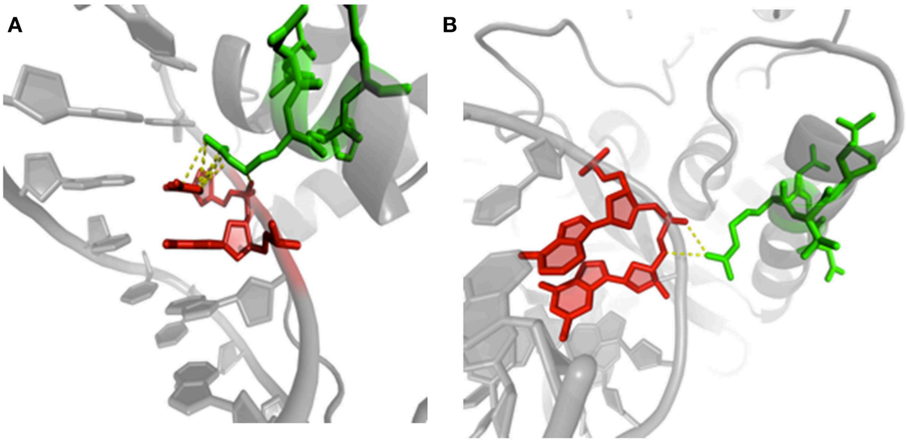


FIGURE 8 | Examples of protein/DNA interactions. (A) Structure of human centromere protein B (CENP-B) binding to DNA CENP-B box (PDB code 1HLV, Tanaka et al., 2001). The image highlights contacts between arginine 125 (chain A, green) in PB *m* (regular helix) and cytosine 15 (chain B, red) in ntC 41. **(B)** Details of methionine repressor protein (MetJ) binding to DNA metbox (PDB code 1MJQ, Garvie and Phillips, 2000). The same PB *m* and amino acid residue (arginine 40 in chain H, green) is in contact with guanine 2 (chain K, red) in NtC 13. Visualization was created by the program PyMOL (http://www.pymol.org, Delano, 2013).

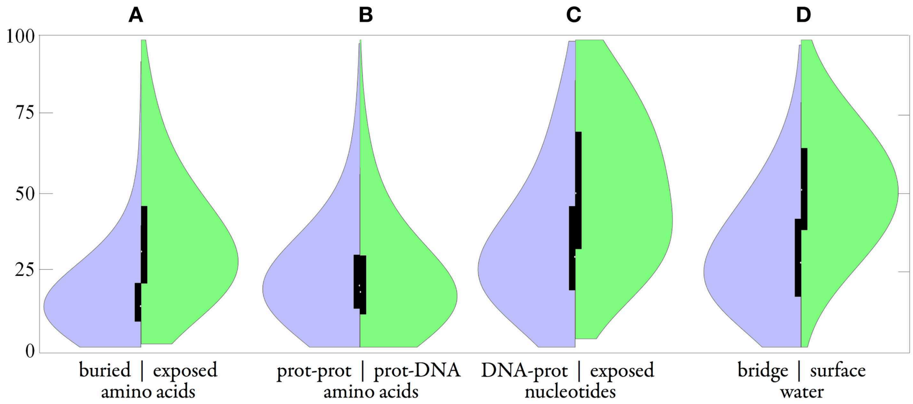


FIGURE 9 | Distributions of B-factors in the group of protein-DNA complexes (165 structures with crystallographic resolution 1.9 Å and better). Smooth plot **(A)** compared buried amino acid (left in purple) vs. exposed aa (right in green); **(B)** protein-protein aa vs. protein-DNA aa; **(C)** DNA-protein nucleotide vs. exposed nt; **(D)** bridge water vs. surface w. Black boxes show the second and third quartiles; the white spot indicates the median.

proteins are tightly linked, which significantly affects the protease that looses more than 30% of its structure.

Among chemical modifications, post-translational modifications (PTMs), like phosphorylation, play a major role in many biology processes. Integrins, for example, can be activated consecutive to phosphorylation. The impact of these modifications on the structure and the dynamics of proteins is thus of particular interest.

Recent studies have shown that PTMs have significant effects on the protein conformations and on their flexibility. Hence Xin and Radivojac used 3D structures from the PDB and studied the conformational heterogeneity of protein structures corresponding to identical sequences in their unmodified and modified forms (Xin and Radivojac, 2012). They demonstrated that PTMs induce conformational changes at both local and global level, but with a limited impact. Accordingly PTMs would affect regulatory and signaling pathways (Nussinov et al., 2012; Xin and Radivojac, 2012) by subtle but common mechanisms of allostery. Some prediction approaches and are included into dedicated databases (Matlock et al., 2015), but few analyzed precisely the whole PTMome.

This led us to conduct a deep analysis of structures of the same protein with or without PTMs. As an example, we selected 157 PDB chains of the human Cyclin-dependent kinase 2 (*UniProt AC: P24941*) in complex form, and 222 PDB chains of unbound monomer. Based on data from PTM-SD (Craveur et al., 2014), a database of structurally solved and annotated post-translational modifications, 112 chains among the 157

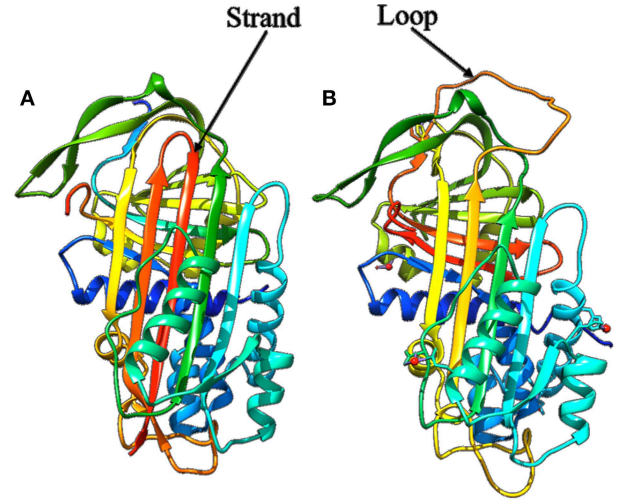


FIGURE 10 | Structure of the alpha 1 antitrypsin. (A) The cleaved form after complexation with protease (PDB code 1EZX, Huntington et al., 2000) showing the strand inserted in the β-sheet after cleavage, **(B)** the uncleaved form (PDB code 1HP7, Kim et al., 2001) showing the wild whole loop.

complexes, present a phosphorylated threonine at position 160 in the structure of the kinase. As described in **Table 1**, we compared the backbone flexibility of three different cases: unbound kinase, kinase complex, and phospho-Thr160 kinase in complex.

Comparison of the three N_{eq} profiles, shown in **Figure 11**, highlights significant differences in local flexibility of the kinase structures. **Figure 11A** shows that, when kinase is in unbound form, the polypeptide chain presents a flexible fragment (colored in green), which corresponds to a large loop. When complex is formed (**Figure 11B**), this loop is placed at the interface and leads to stiffening of its edges and higher flexibility in the neighborhood of Thr-160. This change is characterized by a diminution and an increase of N_{eq}-values, respectively. Finally, when the Thr-160 is phosphorylated (**Figure 11C**), the green region becomes comparatively rigid, which results to limited flexibility ($N_{eq} \leq 3.16$). However, another region in kinase (position 8 to 18) is associated with increasing flexibility. When the complex is forming, the N_{eq} range in this area increases from (1;2.77) to (1;3.76), and secondly, when the phosphorylation is in place, the range increases to (1;5.91). Interestingly, this region corresponds to the neighboring positions of two other phosphorylation sites, at Thr-14 and Tyr-15. It is important to note that these phosphorylations were absent in the structures used here for the N_{eq} computation.

In a functional point of view, the phosphorylation in position 160 is known to promote the activation of the kinase, while the phosphorylation of position 14 and 15 slightly reduce its activity (Gu et al., 1992). Thereby, the changes in flexibility observed at these 3 phosphorylation sites, could reflect that the activity of the kinase is regulated by a mechanism of complementary rigidity/flexibility of local protein backbone, which could be related to allosteric effects.

The red line plotted in **Figure 11** represents the number of available structural data for each position. Interestingly, the green region in **Figure 11** is proportionally less resolved when kinase is in monomer than when it is in complex, and even more solved when the Thr-160 is phosphorylated. This observation emphasizes that the decrease of flexibility in this region facilitates the resolution of the structures. Several structures of the same protein present specific regions that are disordered in some crystals and ordered in others. These regions were defined by Zhang and collaborators as “Dual Personality Fragments” (Zhang et al., 2007), and the corresponding fragment of the green region in Cyclin dependent kinase was the emblematic example used by Zhang et al. (2007) to defined DPF. In the same way, the region between positions 35 to 45 were also identify as DP fragments.

Prediction of Protein Flexibility

The growing gap between the number of protein sequences and the number of atomic structures imposes to resort to alternative approaches to gain structural and dynamics information. They are mainly based on crystallographic B-factor analyses. It is often seen that crystallographic B-factors are a mix of properties, dynamics being one of them. Recent approaches show that NMR spectroscopy provides an ever increasing amount of dynamics data, going well beyond the simple thermal vibrations (Powers et al., 1993; Palmer, 2001; Olsson et al., 2014). None of them can describe all the important flexible movement or even disorder. Hence, it must be taken into account that everything in protein dynamics cannot be assessed based on a single view Prediction methods are therefore of particular importance. Flexibility prediction from sequences started as a Boolean prediction, i.e., rigid or flexible, using simple statistical analyses of B-factor values (Karplus and Schulz, 1985; Vihinen et al., 1994). Following developments combined evolutionary information to different machine learning methods, such as Artificial Neural Networks (Schlessinger et al., 2006), support vector regression coupled with random forest (Pan and Shen, 2009), and support vector machines (Kuznetsov, 2008; Kuznetsov and McDuffie, 2008). Additional sources of information were progressively take into account, rather than X-ray B-factors, as Nuclear magnetic resonance data (NMR) (Trott et al., 2008; Zhang et al., 2010), dihedral angles and accessibility (Hwang et al., 2011), or computational data from Normal Mode Analysis (Hirose et al., 2010). At last, some methodology, dedicated to predict protein disorder were also developed and designed to high flexibility prediction (Galzitskaya et al., 2006; Mamonova et al., 2010; Jones and Cozzetto, 2015). Recent approaches are quite complex like (i) the DynaMine webserver (http://dynamine.ibsquare.be/, Cilia et al., 2013, 2014), DynaMine predicts backbone flexibility at the residue-level in the form of backbone N-H S^2 order parameter values learnt from NMR data, or (ii) as a predictor which used relative solvent accessibility (RSA) and custom-derived amino acid (AA) alphabets. The prediction is done in two-stage linear regression model that uses RSA-based space in a local sequence window in the first stage and a reduced AA pair-based space in the second stage as the inputs (Zhang and Kurgan, 2014).

TABLE 1 | Composition of structural complexes involving the Cyclin-dependent kinase 2.

	UniProt AC in complex	Nbr of complex	Protein name and organism
	COMPLEX WITH PHOSPHORYLATION ON Thr 160 IN P24941		
P24941 Cyclin-dependent kinase 2 Homo sapiens (Human)	P14635	1	G2/mitotic-specific cyclin-B1 Homo sapiens (Humain)
	P20248	74	Cyclin-A2 Homo sapiens (Humain)
	P24864	1	G1/S-specific cyclin-E1 Homo sapiens (Humain)
	P30274	22	Cyclin-A2 Bos taurus (Bovin)
	P51943	8	Cyclin-A2 Mus musculus (Souris)
	Q16667	1	Cyclin-dependent kinase inhibitor 3 Homo sapiens (Humain)
	P20248 + Q99741	4	Cyclin-A2 Homo sapiens (Humain) + Cell division control protein 6 homolog Homo sapiens (Humain)
	P20248 + P46527	1	Cyclin-A2 Homo sapiens (Humain) + Cyclin-dependent kinase inhibitor 1B Homo sapiens (Humain)
	Total	112	
	COMPLEX WITHOUT PHOSPHORYLATION ON Thr 160 IN P24941		
	P20248	42	Cyclin-A2 Homo sapiens (Humain)
	P61024	1	Cyclin-dependent kinases regulatory subunit 1 Homo sapiens (Humain)
	P89883	2	V-cyclin of Murid herpesvirus 4
	Total	45	

We also proposed prediction of protein flexibility of an amino acid sequence using the potentialities of SA prediction. The approach is not only innovative through the use of local protein conformations, but also with specific definition of flexibility. Flexibility is often defined based on α-carbon B-factor values obtained from X-ray experiments. As mentioned above, these data reflect protein flexibility, but may also be prone to experimental and systematic biases. Hence, flexibility was considered with X-ray B-factor descriptors and the RMSF observed in MDs simulations, which is calculated from the amplitude of atom motions during simulation. Both descriptors were combined to define and to examine flexibility classes of SA.

This dedicated prediction method is divided in two steps: first an SA prediction from sequence, and second a flexibility prediction from the SA predicted. The SA used in this method is the LSP (see Section The Different Views of Protein Structures). They consist of 120 overlapping structural classes of 11-residue long fragments (Benros et al., 2006), which encompass all known local protein structures and ensure good quality 3D local approximation. The major advantage of this library is its capacity to capture the continuity between the identified recurrent local structures (Benros et al., 2009). We can notice that is quite difficult to have a good correlation between theoretical results to actual experiments. With LSPs, we have shown that they have on average a correlation > 0.9 with B-factors.

Relevant sequence–structure relationships were also observed and further used for prediction. Briefly, LSP prediction is based on SVM training. With the LSP prediction, a Confidence Index (CI) that is based on the discriminative power of the SVMs is provided. The higher CI, the better the prediction rate is. The prediction rate reaches 63.1%, a rather high value given the high number of structural classes (Bornot et al., 2009).

In a second step, we considered the two descriptors for quantifying protein dynamics, X-ray B-factors and RMSF. They were combined to define 3 flexibility classes of LSPs: rigid, intermediate and flexible. Then for each 11-residue long target sequences, the SA prediction provided a list of five possible LSP candidates. Based on the previously defined flexibility classes of these structural candidates, the prediction of target flexibility is made. Interestingly, the prediction rate is slightly better than the one of PROFbval (Schlessinger et al., 2006) that was optimized for only two classes.

Hence, the originality of the method lies (i) in the use of a combination of B-factors and RMSF for quantifying protein dynamics, (ii) in prediction of flexibility through SA prediction of LSPs, and (iii) in prediction of three classes of flexibility, which are usually limited to two. The method is implemented in a web server named PredyFlexy (http://www.dsimb.inserm.fr/dsimb_tools/predyflexy, de Brevern et al., 2012), in which the users have access to a confidence index (CI) for assessing the quality of the prediction rate.

Conclusion

The protein structure organization is characterized by a conformational arrangement of repetitive structures (secondary structures, i.e., α-helices, β-sheets and coils/loops). Static observation of protein organization has revealed some of their essential properties, i.e., active sites are generally found at the protein core in which residues are well packed and mainly hydrophobic, while the surface residues, exposed to solvent or to another partner(s) (protein, DNA), are more flexible because less constrained than the core. The function of proteins and their interaction mechanism need some flexible properties that are considerably more complex than this simplistic binary view.

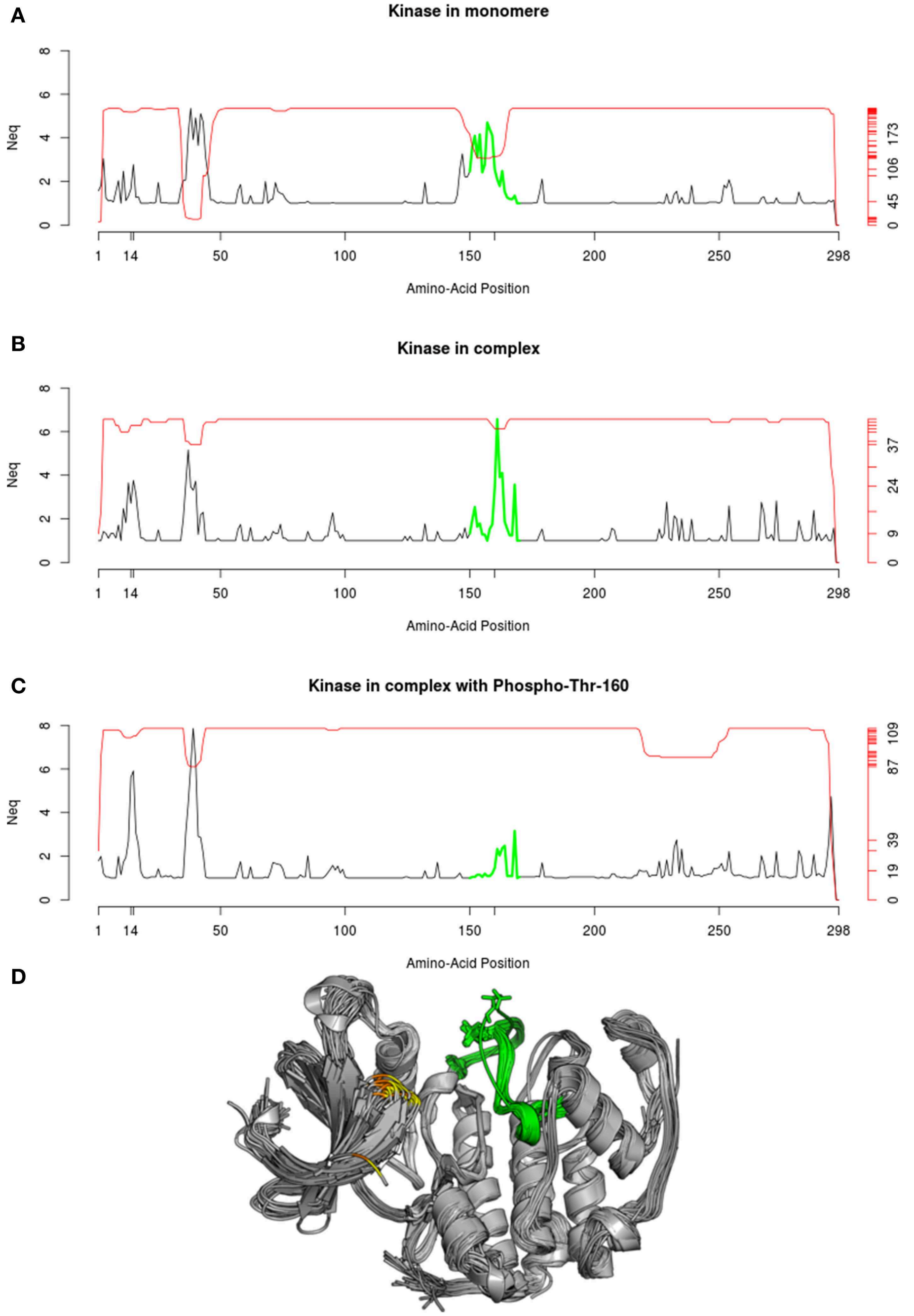


FIGURE 11 | N_{eq} profile of Cyclin-dependent kinase 2. The N_{eq} profile is given in each case: **(A)** for structure of kinase found in monomer, **(B)** found as part of a complex, **(C)** found in complex with a phosphorylation solved in the Thr-160. **(D)** The superposition of the 112 PDB chains used to compute the N_{eq} profile in **(C)** is shown. The corresponding green region is highlighted, and positions 14 and 15 are, respectively, indicated in yellow and orange. Protein visualization was created by the program PyMOL (http://www.pymol.org, Delano, 2013).

By exploiting various structural data sources and by developing different computational methods (B-factor, NMR data, MDs Simulation, NMA, …) dynamics of proteins turn out to cover a large spectrum of conformational changes (combined by mobility of rigid fragment and deformability of backbone), by the existence of intrinsic disorder region, by allosteric effect… Some of these flexible mechanisms need structural reorganization at a local level. Thus investigation of protein flexibility requires a more local and complex description of protein structures than the classic representation.

In this review we have illustrated using numerous examples (DARC protein, Human integrins, Protein Complexes, Protein/DNA interfaces, Proteins with Post-Translational Modifications) how the approaches, based on Structural Alphabets, are a valuable tool to study flexibility at this level.

From our experiences with these examples, we can state that the use of SAs allows to tackle and address the important problem of the comparison of an ensemble of protein conformations. Indeed, in a recent paper, Scott and Strauss (Scott and Straus, 2015) underlines the bias related to the use of RMSD, which needs beforehand an optimally superimposed approach often remains as rigid bodies. They proposed an elegant method, fleximatch, of protein structure comparison that tries to take flexibility into account. As it was done for protein superimposition methods (Yang and Tung, 2006; Tung et al., 2007; Tung and Yang, 2007; Le et al., 2009; Budowski-Tal et al., 2010; Gelly et al., 2011; Leonard et al., 2014), SA is an efficient approach, not considering proteins as rigid bodies. We underline the interest of our approach based on Protein Blocks with the PBxplore tools (https://github.com/pierrepo/PBxplore, in preparation) or GSAtools (http://mathbio.nimr.mrc.ac.uk/wiki/GSATools, Pandini et al., 2013) in other cases. The use of SAs and the development of associated metrics such as N_{eq} is required to study the details and begin to understand the complexity of protein flexibility. It allows discriminating flexibility from mobility and deformability, which is not currently considered by other available methods. Nonetheless, it also had drawbacks as no simple threshold will guide the researcher to point out that certain segment is THE highly flexible part and not the other, same as for RMSF. In the same way, use of information theory with GSATools also requires expertise. Moreover, as SA represents a simplification of the 3D description, its results can be compared to the Normal Mode Analysis based on Elastic Network Model (Suhre and Sanejouand, 2004; Tiwari et al., 2014; Eyal et al., 2015) that are efficient to define large movement. However, changes at a finer level such as side chain rotameric states or minor changes in the backbone (but essential for the biological functions) are more difficult to handle. Here as always, a good knowledge of the biological system is essential as a correct definition of the scientific question and its scale (Buehler and Yung, 2009).

To conclude, we can find that all these approaches are suitable for highlighting both flexible and rigid parts of a protein from structures derived from NMR, X-ray diffraction or molecular simulation.

Author Contributions

Section Introduction: PC, APJ, JE, TJN, FN, JR, CE, NS, JCG. Section The Different Views of Protein Structures: PC, APJ, JE, TJN, MG, SL, PP, GF, JR, AG, LSS, RMB, JB, ST, JC, BS, CE, NS, JCG. Section Duffy Antigen/Chemokine Receptor Protein: NS, OB, CE. Section Human Integrin α2bβ3: APJ, NS, MG, PP, JR, JB, ST, VJ. Section Protein Complexes and Allostery: PC, APJ, JE, LSS, RMB, NS. Section Protein/DNA Interfaces: JC, BS, JCG. Section PTMs: PC, JE, TJN, FN, SL, GF, JR. Section Prediction of Protein Flexibility: PC, APJ, JE, SL, GF, AG, CE, JCG. Section Conclusion: PC, APJ, GF, CE, NS. AdB conceived the review and participated in all the different sections.

Acknowledgments

PC, APJ, JE, TJN, FL, NS, MG, SL, PP, OB, GF, JR, AG, JB, ST, CE, JCG, and AdB are supported by the Ministry of Research (France), University Paris Diderot, Sorbonne Paris Cité (France), National Institute of Blood Transfusion (INTS, France), National Institute of Health and Medical Research (INSERM, France), and labex GR-Ex (France). The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program "Investissements d'avenir" of the French National Research Agency, reference ANR-11-IDEX-0005-02. This work was supported by the Czech-France collaboration Partenariat Hubert Curien Barrande (MEB021032) for JCG, JC, BS, and ADB. This study was also supported by BIOCEV CZ.1.05/1.1.00/02.0109 from the ERDF and RVO: 86652036 for BS and JC, by University of Toronto (Canada) for LSS, by Max Planck (Germany) for BMR, by INTS (France) for VJ, by CNRS (France) and Univ Lyon (France) for JB, by CNRS (France), INSERM (France) and Univ Strasbourg (France) for JE, by CNRS (France) and Univ Nantes (France) for ST, by HFSP to APJ; by NIH (USA) for GF, by Department of Biotechnology (India) to NS, and by CNRS (France), Univ Pierre & Marie Curie (France) and ANR NaturaDyRe (France, ANR-2010-CD2I-014-04) to JR. APJ was supported by CEFIPRA number 3903-E and TJN is supported by CEFIPRA number 5203-2. NS and AdB also acknowledge to CEFIPRA for collaborative grant (numbers 3903-E and 5203-2).

References

Allen, S. J., Crown, S. E., and Handel, T. M. (2007). Chemokine: receptor structure, interactions, and antagonism. *Annu. Rev. Immunol.* 25, 787–820. doi: 10.1146/annurev.immunol.24.021605.090529

Batchelor, J. D., Malpede, B. M., Omattage, N. S., DeKoster, G. T., Henzler-Wildman, K. A., and Tolia, N. H. (2014). Red blood cell invasion by Plasmodium vivax: structural basis for DBP engagement of DARC. *PLoS Pathog.* 10:e1003869. doi: 10.1371/journal.ppat.1003869

Benros, C., de Brevern, A. G., Etchebest, C., and Hazout, S. (2006). Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins* 62, 865–880. doi: 10.1002/prot.20815

Benros, C., de Brevern, A. G., and Hazout, S. (2009). Analyzing the sequence-structure relationship of a library of local structural prototypes. *J. Theor. Biol.* 256, 215–226. doi: 10.1016/j.jtbi.2008.08.032

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

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. J.r,., Brice, M. D., Rodgers, J. R., et al. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542. doi: 10.1016/S0022-2836(77)80200-3

Bhaskara, R. M., de Brevern, A. G., and Srinivasan, N. (2013). Understanding the role of domain-domain linkers in the spatial orientation of domains in multi-domain proteins. *J. Biomol. Struct. Dyn.* 31, 1467–1480. doi: 10.1080/07391102.2012.743438

Biswas, S., Guharoy, M., and Chakrabarti, P. (2009). Dissection, residue conservation, and structural classification of protein-DNA interfaces. *Proteins* 74, 643–654. doi: 10.1002/prot.22180

Bornot, A., Etchebest, C., and de Brevern, A. G. (2009). A new prediction strategy for long local protein structures using an original description. *Proteins* 76, 570–587. doi: 10.1002/prot.22370

Budowski-Tal, I., Nov, Y., and Kolodny, R. (2010). FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc. Natl. Acad. Sci. U.S.A.* 107, 3481–3486. doi: 10.1073/pnas.0914097107

Buehler, M. J., and Yung, Y. C. (2009). Deformation and failure of protein materials in physiologically extreme conditions and disease. *Nat. Mater.* 8, 175–188. doi: 10.1038/nmat2387

Camproux, A. C., Gautier, R., and Tuffery, P. (2004). A hidden markov model derived structural alphabet for proteins. *J. Mol. Biol.* 339, 591–605. doi: 10.1016/j.jmb.2004.04.005

Camproux, A. C., Tuffery, P., Chevrolat, J. P., Boisvieux, J. F., and Hazout, S. (1999). Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng.* 12, 1063–1073. doi: 10.1093/protein/12.12.1063

Cech, P., Kukal, J., Cerny, J., Schneider, B., and Svozil, D. (2013). Automatic workflow for the classification of local DNA conformations. *BMC Bioinformatics* 14, 205. doi: 10.1186/1471-2105-14-205

Chavent, M., Levy, B., Krone, M., Bidmon, K., Nomine, J. P., Ertl, T., et al. (2011). GPU-powered tools boost molecular visualization. *Brief. Bioinformatics* 12, 689–701. doi: 10.1093/bib/bbq089

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2013). From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.* 4, 2741. doi: 10.1038/ncomms3741

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.* 42, W264–W270. doi: 10.1093/nar/gku270

Compton, A., and Haber, J. M. (1960). The duffy blood group system in transfusion reactions: a reviw of the literature and report of four cases. *Blood* 15, 186–191.

Corey, R. B., and Pauling, L. (1953). Fundamental dimensions of polypeptide chains. *Proc. R. Soc. Lond. B Biol. Sci.* 141, 10–20. doi: 10.1098/rspb.1953.0011

Craveur, P., Rebehmed, J., and de Brevern, A. G. (2014). PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. *Database* 2014:bau041. doi: 10.1093/database/bau041

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

Cutbush, M., and Mollison, P. L. (1950). The Duffy blood group system. *Heredity (Edinb.)* 4, 383–389. doi: 10.1038/hdy.1950.31

Cutts, J. C., Powell, R., Agius, P. A., Beeson, J. G., Simpson, J. A., and Fowkes, F. J. (2014). Immunological markers of Plasmodium vivax exposure and immunity: a systematic review and meta-analysis. *BMC Med.* 12:150. doi: 10.1186/s12916-014-0150-1

de Brevern, A. G., Autin, L., Colin, Y., Bertrand, O., and Etchebest, C. (2009). *In silico* studies on DARC. *Infect. Disord. Drug Targets* 9, 289–303. doi: 10.2174/1871526510909030289

de Brevern, A. G., Benros, C., Gautier, R., Valadie, H., Hazout, S., and Etchebest, C. (2004). Local backbone structure prediction of proteins. *In Silico Biol.* 4, 381–386.

de Brevern, A. G., Bornot, A., Craveur, P., Etchebest, C., and Gelly, J. C. (2012). PredyFlexy: flexibility and local structure prediction from sequence. *Nucleic Acids Res.* 40, W317–W322. doi: 10.1093/nar/gks482

de Brevern, A. G., Etchebest, C., and Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* 41, 271–287. doi: 10.1002/1097-0134(20001115)41:3<271::AID-PROT10>3.0.CO;2-Z

de Brevern, A. G., and Hazout, S. (2003). 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics* 19, 345–353. doi: 10.1093/bioinformatics/btf859

de Brevern, A. G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C., and Etchebest, C. (2005). A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). *Biochim. Biophys. Acta* 1724, 288–306. doi: 10.1016/j.bbagen.2005.05.016

Delano, W. L. (2013). *The PyMOL Molecular Graphics System on World Wide Web.* Available online at: http://www.pymol.org

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433–3434. doi: 10.1093/bioinformatics/bti541

Dudev, M., and Lim, C. (2007). Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. *BMC Bioinformatics* 8:106. doi: 10.1186/1471-2105-8-106

Dunker, A. K. (2007). Another window into disordered protein function. *Structure* 15, 1026–1028. doi: 10.1016/j.str.2007.08.001

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59. doi: 10.1016/S1093-3263(00)00138-8

Dunker, A. K., and Obradovic, Z. (2001). The protein trinity–linking function and disorder. *Nat. Biotechnol.* 19, 805–806. doi: 10.1038/nbt0901-805

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* 11, 161–171.

Eisenberg, D. (2003). The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11207–11210. doi: 10.1073/pnas.2034522100

Espinoza, J. P., Caradeux, J., Norwitz, E. R., and Illanes, S. E. (2013). Fetal and neonatal alloimmune thrombocytopenia. *Rev. Obstet. Gynecol.* 6, e15–21. doi: 10.1097/AOG.0b013e31823403f4

Etchebest, C., Benros, C., Hazout, S., and de Brevern, A. G. (2005). A structural alphabet for local protein structures: improved prediction methods. *Proteins* 59, 810–827. doi: 10.1002/prot.20458

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

Ferrada, E., and Melo, F. (2009). Effective knowledge-based potentials. *Protein Sci.* 18, 1469–1485. doi: 10.1002/pro.166

Fetrow, J. S., Palumbo, M. J., and Berg, G. (1997). Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins* 27, 249–271.

Fong, J. H., and Panchenko, A. R. (2010). Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Mol. Biosyst.* 6, 1821–1828. doi: 10.1039/c005144f

Fornili, A., Pandini, A., Lu, H. C., and Fraternali, F. (2013). Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. *J. Chem. Theory Comput.* 9, 5127–5147. doi: 10.1021/ct400486p

Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006). FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 22, 2948–2949. doi: 10.1093/bioinformatics/btl504

Garvie, C. W., and Phillips, S. E. (2000). Direct and indirect readout in mutant Met repressor-operator complexes. *Structure* 8, 905–914. doi: 10.1016/S0969-2126(00)00182-9

Gelly, J. C., Joseph, A. P., Srinivasan, N., and de Brevern, A. G. (2011). iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res.* 39, W18–W23. doi: 10.1093/nar/gkr333

George, J. N., Caen, J. P., and Nurden, A. T. (1990). Glanzmann's thrombasthenia: the spectrum of clinical disease. *Blood* 75, 1383–1395.

Goh, C. S., Milburn, D., and Gerstein, M. (2004). Conformational changes associated with protein-protein interactions. *Curr. Opin. Struct. Biol.* 14, 104–109. doi: 10.1016/j.sbi.2004.01.005

Grunberg, R., Leckner, J., and Nilges, M. (2004). Complementarity of structure ensembles in protein-protein binding. *Structure* 12, 2125–2136. doi: 10.1016/j.str.2004.09.014

Gu, Y., Rosenblatt, J., and Morgan, D. O. (1992). Cell cycle regulation of CDK2 activity by phosphorylation of Thr160 and Tyr15. *EMBO J.* 11, 3995–4005.

Guerra, C. A., Snow, R. W., and Hay, S. I. (2006). Mapping the global extent of malaria in 2005. *Trends Parasitol.* 22, 353–358. doi: 10.1016/j.pt.2006.06.006

Hirose, S., Yokota, K., Kuroda, Y., Wako, H., Endo, S., Kanai, S., et al. (2010). Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. *BMC Struct. Biol.* 10:20. doi: 10.1186/1472-6807-10-20

Hirst, J. D., Glowacki, D. R., and Baaden, M. (2014). Molecular simulations and visualization: introduction and overview. *Faraday Discuss.* 169, 9–22. doi: 10.1039/C4FD90024C

Horne, K., and Woolley, I. J. (2009). Shedding light on DARC: the role of the Duffy antigen/receptor for chemokines in inflammation, infection and malignancy. *Inflamm. Res.* 58, 431–435. doi: 10.1007/s00011-009-0023-9

Huntington, J. A., Read, R. J., and Carrell, R. W. (2000). Structure of a serpin-protease complex shows inhibition by deformation. *Nature* 407, 923–926. doi: 10.1038/35038119

Hwang, H., Vreven, T., Whitfield, T. W., Wiehe, K., and Weng, Z. (2011). A machine learning approach for the prediction of protein surface loop flexibility. *Proteins* 79, 2467–2474. doi: 10.1002/prot.23070

Hynes, R. O. (2002). Integrins: bidirectional, allosteric signaling machines. *Cell* 110, 673–687. doi: 10.1016/S0092-8674(02)00971-6

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* 5, 299–314. doi: 10.2307/1390807

Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35, W460–W464. doi: 10.1093/nar/gkm363

Jallu, V., Bertrand, G., Bianchi, F., Chenet, C., Poulain, P., and Kaplan, C. (2013). The alphaIIb p.Leu841Met (Cab3(a+)) polymorphism results in a new human platelet alloantigen involved in neonatal alloimmune thrombocytopenia. *Transfusion* 53, 554–563. doi: 10.1111/j.1537-2995.2012.03762.x

Jallu, V., Dusseaux, M., Panzer, S., Torchet, M. F., Hezard, N., Goudemand, J., et al. (2010). AlphaIIbbeta3 integrin: new allelic variants in Glanzmann thrombasthenia, effects on ITGA2B and ITGB3 mRNA splicing, expression, and structure-function. *Hum. Mutat.* 31, 237–246. doi: 10.1002/humu.21179

Jallu, V., Poulain, P., Fuchs, P. F., Kaplan, C., and de Brevern, A. G. (2012). Modeling and molecular dynamics of HPA-1a and -1b polymorphisms: effects on the structure of the beta3 subunit of the alphaIIbbeta3 integrin. *PLoS ONE* 7:e47304. doi: 10.1371/journal.pone.0047304

Jallu, V., Poulain, P., Fuchs, P. F., Kaplan, C., and de Brevern, A. G. (2014). Modeling and molecular dynamics simulations of the V33 variant of the integrin subunit beta3: structural comparison with the L33 (HPA-1a) and P33 (HPA-1b) variants. *Biochimie* 105, 84–90. doi: 10.1016/j.biochi.2014.06.017

Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.* 41, 133–180. doi: 10.1017/S0033583508004708

Jin, Y., and Dunbrack, R. L. J.r,. (2005). Assessment of disorder predictions in CASP6. *Proteins* 61(Suppl. 7), 167–175. doi: 10.1002/prot.20734

Jones, D. T., and Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31, 857–863. doi: 10.1093/bioinformatics/btu744

Jones, S., and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13–20. doi: 10.1073/pnas.93.1.13

Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J. C., Swapna, L. S., Offmann, B., et al. (2010a). A short survey on protein blocks. *Biophys. Rev.* 2, 137–145. doi: 10.1007/s12551-010-0036-1

Joseph, A. P., Bornot, A., and de Brevern, A. G. (2010b). "Local structural alphabet," in *Protein Structure Methods and Algorithms*, eds H. Rangwala and G. Karypis (Hoboken, NJ: Wiley), 75–106. doi: 10.1002/9780470882207.ch5

Joseph, A. P., and de Brevern, A. G. (2014). From local structure to a global framework: recognition of protein folds. *J. R. Soc. Interface* 11:20131147. doi: 10.1098/rsif.2013.1147

Joseph, A. P., Srinivasan, N., and de Brevern, A. G. (2012). Progressive structure-based alignment of homologous proteins: adopting sequence comparison strategies. *Biochimie* 94, 2025–2034. doi: 10.1016/j.biochi.2012.05.028

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

Karplus, P., and Schulz, G. (1985). Prediction of chain flexibility in proteins. A tool for the selection of peptide antigens. *Naturwissenschaften* 72, 212–213. doi: 10.1007/BF01195768

Kim, S., Woo, J., Seo, E. J., Yu, M., and Ryu, S. (2001). A 2.1 A resolution structure of an uncleaved alpha(1)-antitrypsin shows variability of the reactive center and other loops. *J. Mol. Biol.* 306, 109–119. doi: 10.1006/jmbi.2000.4357

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.* 323, 297–307. doi: 10.1016/S0022-2836(02)00942-7

Kuznetsov, I. B. (2008). Ordered conformational change in the protein backbone: prediction of conformationally variable positions from sequence and low-resolution structural data. *Proteins* 72, 74–87. doi: 10.1002/prot.21899

Kuznetsov, I. B., and McDuffie, M. (2008). FlexPred: a web-server for predicting residue positions involved in conformational switches in proteins. *Bioinformation* 3, 134–136. doi: 10.6026/97320630003134

Le, Q., Pollastri, G., and Koehl, P. (2009). Structural alphabets for protein structure classification: a comparison study. *J. Mol. Biol.* 387, 431–450. doi: 10.1016/j.jmb.2008.12.044

Lensink, M. F., and Mendez, R. (2008). Recognition-induced conformational changes in protein-protein docking. *Curr. Pharm. Biotechnol.* 9, 77–86. doi: 10.2174/138920108783955173

Leonard, S., Joseph, A. P., Srinivasan, N., Gelly, J. C., and de Brevern, A. G. (2014). mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet. *J. Biomol. Struct. Dyn.* 32, 661–668. doi: 10.1080/07391102.2013.787026

Lindahl, E., Hess, B., and van der Spoel, D. (2001). GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.* 7, 306–317. doi: 10.1007/s008940100045

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure* 11, 1453–1459. doi: 10.1016/j.str.2003.10.002

Liu, X. H., Hadley, T. J., Xu, L., Peiper, S. C., and Ray, P. E. (1999). Up-regulation of Duffy antigen receptor expression in children with renal disease. *Kidney Int.* 55, 1491–1500. doi: 10.1046/j.1523-1755.1999.00385.x

Lobanov, M. Y., Shoemaker, B. A., Garbuzynskiy, S. O., Fong, J. H., Panchenko, A. R., and Galzitskaya, O. V. (2010). ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder. *Nucleic Acids Res.* 38, D283–D287. doi: 10.1093/nar/gkp963

Luo, X., Lv, F., Pan, Y., Kong, X., Li, Y., and Yang, Q. (2011). Structure-based prediction of the mobility and disorder of water molecules at protein-DNA interface. *Protein Pept. Lett.* 18, 203–209. doi: 10.2174/092986611794475066

Mamonova, T. B., Glyakina, A. V., Kurnikova, M. G., and Galzitskaya, O. V. (2010). Flexibility and mobility in mesophilic and thermophilic homologous proteins from molecular dynamics and FoldUnfold method. *J. Bioinform. Comput. Biol.* 8, 377–394. doi: 10.1142/S0219720010004690

Marsh, J. A. (2013). Buried and accessible surface area control intrinsic protein flexibility. *J. Mol. Biol.* 425, 3250–3263. doi: 10.1016/j.jmb.2013.06.019

Matlock, M. K., Holehouse, A. S., and Naegle, K. M. (2015). ProteomeScout: a repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Res.* 43, D521–D530. doi: 10.1093/nar/gku1154

Meszaros, B., Simon, I., and Dosztanyi, Z. (2011). The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Phys. Biol.* 8:035003. doi: 10.1088/1478-3975/8/3/035003

Micheletti, C., Seno, F., and Maritan, A. (2000). Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* 40, 662–674. doi: 10.1002/1097-0134(20000901)40:4<662::AID-PROT90>3.0.CO;2-F

Miller, L. H., Mason, S. J., Clyde, D. F., and McGinniss, M. H. (1976). The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* 295, 302–304. doi: 10.1056/NEJM197608052950602

Miller, L. H., Mason, S. J., Dvorak, J. A., McGinniss, M. H., and Rothman, I. K. (1975). Erythrocyte receptors for (*Plasmodium knowlesi*) malaria: Duffy blood group determinants. *Science* 189, 561–563. doi: 10.1126/science.1145213

Misaghi, S., Galardy, P. J., Meester, W. J., Ovaa, H., Ploegh, H. L., and Gaudet, R. (2005). Structure of the ubiquitin hydrolase UCH-L3 complexed with a suicide substrate. *J. Biol. Chem.* 280, 1512–1520. doi: 10.1074/jbc.M410770200

Nussinov, R., Tsai, C. J., Xin, F., and Radivojac, P. (2012). Allosteric post-translational modification codes. *Trends Biochem. Sci.* 37, 447–455. doi: 10.1016/j.tibs.2012.07.001

Offmann, B., Tyagi, M., and de Brevern, A. G. (2007). Local protein structures. *Curr. Bioinform.* 3, 165–202. doi: 10.2174/157489307781662105

Olsson, S., Vögeli, B., Cavalli, A., Boomsma, W., Ferkinghoff-Borg, J., Lindorff-Larsen, K., et al. (2014). Probabilistic determination of native state ensembles of proteins. *J. Chem. Theory Comput.* 10, 3484–3491. doi: 10.1021/ct5001236

Otterbein, L. R., Cosio, C., Graceffa, P., and Dominguez, R. (2002). Crystal structures of the vitamin D-binding protein and its complex with actin: structural basis of the actin-scavenger system. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8003–8008. doi: 10.1073/pnas.122126299

Palmer, A. G. 3rd. (2001). Nmr probes of molecular dynamics: overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* 30, 129–155. doi: 10.1146/annurev.biophys.30.1.129

Pan, X. Y., and Shen, H. B. (2009). Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein Pept. Lett.* 16, 1447–1454. doi: 10.2174/092986609789839250

Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. (2012). Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* 26, 868–881. doi: 10.1096/fj.11-190868

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

Pandini, A., Fornili, A., and Kleinjung, J. (2010). Structural alphabets derived from attractors in conformational space. *BMC Bioinformatics* 11:97. doi: 10.1186/1471-2105-11-97

Park, B. H., and Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* 249, 493–507. doi: 10.1006/jmbi.1995.0311

Powers, R., Clore, G., Garrett, D., and Gronenborn, A. (1993). Relationships between the precision of high-resolution protein NMR structures, solution-order parameters, and crystallographic B factors. *J. Magn. Reson. B* 101, 325–327. doi: 10.1006/jmrb.1993.1051

Python Software Foundation. (2015). *Python Language Reference, Version 2.7.* Available online at: http://www.python.org

Rangwala, H., Kauffman, C., and Karypis, G. (2009). svmPRAT: SVM-based protein residue annotation toolkit. *BMC Bioinformatics* 10:439. doi: 10.1186/1471-2105-10-439

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 34, 167–339. doi: 10.1016/S0065-3233(08)60520-3

Russo, D., Teixeira, J., Kneller, L., Copley, J. R., Ollivier, J., Perticaroli, S., et al. (2011). Vibrational density of states of hydration water at biomolecular sites: hydrophobicity promotes low density amorphous ice behavior. *J. Am. Chem. Soc.* 133, 4882–4888. doi: 10.1021/ja109610f

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451. doi: 10.1093/nar/gkh086

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., and Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE* 4:e4433. doi: 10.1371/journal.pone.0004433

Schlessinger, A., and Rost, B. (2005). Protein flexibility and rigidity predicted from sequence. *Proteins* 61, 115–126. doi: 10.1002/prot.20587

Schlessinger, A., Yachdav, G., and Rost, B. (2006). PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics* 22, 891–893. doi: 10.1093/bioinformatics/btl032

Schneider, B., Gelly, J. C., de Brevern, A. G., and Cerny, J. (2014). Local dynamics of proteins and DNA evaluated from crystallographic B factors. *Acta Crystallogr. D Biol. Crystallogr.* 70(Pt 9), 2413–2419. doi: 10.1107/S13990047140 14631

Schuchhardt, J., Schneider, G., Reichelt, J., Schomburg, D., and Wrede, P. (1996). Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng.* 9, 833–842. doi: 10.1093/protein/9. 10.833

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H. M., Jones, D., et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. *Structure* 17, 151–159. doi: 10.1016/j.str.2008.12.014

Scott, W. R., and Straus, S. K. (2015). Determining and visualizing flexibility in protein structures. *Proteins* 83, 820–826. doi: 10.1002/prot.24776

Smolarek, D., Bertrand, O., Czerwinski, M., Colin, Y., Etchebest, C., and de Brevern, A. G. (2010). Multiple interests in structural models of DARC transmembrane protein. *Transfus. Clin. Biol.* 17, 184–196. doi: 10.1016/j.tracli.2010.05.003

Suhre, K., and Sanejouand, Y. H. (2004). ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.* 32, W610–W614. doi: 10.1093/nar/gkh368

Sunami, T., and Kono, H. (2013). Local conformational changes in the DNA interfaces of proteins. *PLoS ONE* 8:e56080. doi: 10.1371/journal.pone.0056080

Suresh, V., Ganesan, K., and Parthasarathy, S. (2013). A protein block based fold recognition method for the annotation of twilight zone sequences. *Protein Pept. Lett.* 20, 249–254. doi: 10.2174/0929866138049 10617

Svozil, D., Kalina, J., Omelka, M., and Schneider, B. (2008). DNA conformations and their sequence preferences. *Nucleic Acids Res.* 36, 3690–3706. doi: 10.1093/nar/gkn260

Swapna, L. S., Mahajan, S., de Brevern, A. G., and Srinivasan, N. (2012). Comparison of tertiary structures of proteins in protein-protein complexes with unbound forms suggests prevalence of allostery in signalling proteins. *BMC Struct. Biol.* 12:6. doi: 10.1186/1472-6807-12-6

Takada, Y., Ye, X., and Simon, S. (2007). The integrins. *Genome Biol.* 8:215. doi: 10.1186/gb-2007-8-5-215

Tanaka, Y., Nureki, O., Kurumizaka, H., Fukai, S., Kawaguchi, S., Ikuta, M., et al. (2001). Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. *EMBO J.* 20, 6612–6618. doi: 10.1093/emboj/20.23.6612

Tiwari, S. P., Fuglebakk, E., Hollup, S. M., Skjaerven, L., Cragnolini, T., Grindhaug, S. H., et al. (2014). *WEBnm@ v2.0: Web Server and Services for Comparing Protein Flexibility. BMC Bioinformatics* 15:6597. doi: 10.1186/s12859-014-0427-6

Tournamille, C., Filipe, A., Badaut, C., Riottot, M. M., Longacre, S., Cartron, J. P., et al. (2005). Fine mapping of the Duffy antigen binding site for the *Plasmodium vivax* Duffy-binding protein. *Mol. Biochem. Parasitol.* 144, 100–103. doi: 10.1016/j.molbiopara.2005.04.016

Tournamille, C., Filipe, A., Wasniowska, K., Gane, P., Lisowska, E., Cartron, J. P., et al. (2003). Structure-function analysis of the extracellular domains of the Duffy antigen/receptor for chemokines: characterization of antibody and chemokine binding sites. *Br. J. Haematol.* 122, 1014–1023. doi: 10.1046/j.1365-2141.2003.04533.x

Touw, W. G., and Vriend, G. (2014). BDB: databank of PDB files with consistent B-factors. *Protein Eng. Des. Sel.* 27, 457–462. doi: 10.1093/protein/gzu044

Trott, O., Siggers, K., Rost, B., and Palmer, A. G. 3rd. (2008). Protein conformational flexibility prediction using machine learning. *J. Magn. Reson.* 192, 37–47. doi: 10.1016/j.jmr.2008.01.011

Tung, C. H., Huang, J. W., and Yang, J. M. (2007). Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for fast protein structure database search. *Genome Biol.* 8:R31. doi: 10.1186/gb-2007-8-3-r31

Tung, C. H., and Yang, J. M. (2007). fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies. *Nucleic Acids Res.* 35, W438–W443. doi: 10.1093/nar/gkm288

Tyagi, M., Benros, C., Martin, J., and de Brevern, A. G. (2007). "Description of the local protein structure II. Novel approaches," in *Recent Research Developments in Protein Engineering*, ed A. G. de Brevern (Trivandrum: Research Signpost), 23–36.

Uhart, M., and Bustos, D. M. (2014). Protein intrinsic disorder and network connectivity. The case of 14-3-3 proteins. *Front. Genet.* 5:10. doi: 10.3389/fgene.2014.00010

Unger, R., Harel, D., Wherland, S., and Sussman, J. L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 5, 355–373. doi: 10.1002/prot.340050410

Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41, 415–427. doi: 10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi: 10.1002/jcc.20291

Vihinen, M., Torkkila, E., and Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins* 19, 141–149. doi: 10.1002/prot.340190207

Wu, C. Y., Chen, Y. C., and Lim, C. (2010). A structural-alphabet-based strategy for finding structural motifs across protein families. *Nucleic Acids Res.* 38:e150. doi: 10.1093/nar/gkq478

Xiao, T., Takagi, J., Coller, B. S., Wang, J. H., and Springer, T. A. (2004). Structural basis for allostery in integrins and binding to fibrinogen-mimetic therapeutics. *Nature* 432, 59–67. doi: 10.1038/nature02976

Xin, F., and Radivojac, P. (2012). Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics* 28, 2905–2913. doi: 10.1093/bioinformatics/bts541

Xue, B., and Uversky, V. N. (2014). Intrinsic disorder in proteins involved in the innate antiviral immunity: another flexible side of a molecular arms race. *J. Mol. Biol.* 426, 1322–1350. doi: 10.1016/j.jmb.2013.10.030

Yang, J. M., and Tung, C. H. (2006). Protein structure database search and evolutionary classification. *Nucleic Acids Res.* 34, 3646–3659. doi: 10.1093/nar/gkl395

Zhang, H., and Kurgan, L. (2014). Improved prediction of residue flexibility by embedding optimized amino acid grouping into RSA-based linear models. *Amino Acids* 46, 2665–2680. doi: 10.1007/s00726-014-1817-9

Zhang, T., Faraggi, E., and Zhou, Y. (2010). Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins* 78, 3353–3362. doi: 10.1002/prot.22842

Zhang, Y., Stec, B., and Godzik, A. (2007). Between order and disorder in protein structures: analysis of “dual personality” fragments in proteins. *Structure* 15, 1141–1147. doi: 10.1016/j.str.2007.07.012

Zhu, J., Luo, B. H., Xiao, T., Zhang, C., Nishida, N., and Springer, T. A. (2008). Structure of a complete integrin ectodomain in a physiologic resting state and activation and deactivation by applied forces. *Mol. Cell* 32, 849–861. doi: 10.1016/j.molcel.2008.11.018

Zimmermann, O., and Hansmann, U. H. (2008). LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. *J. Chem. Inf. Model.* 48, 1903–1908. doi: 10.1021/ci800178a

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

MINI REVIEW
published: 05 August 2015
doi: 10.3389/fmolb.2015.00045

Computational approaches for inferring the functions of intrinsically disordered proteins

Mihaly Varadi [1, 2], *Wim Vranken* [1, 2, 3], *Mainak Guharoy* [1, 2] *and Peter Tompa* [1, 2]*

[1] Flemish Institute of Biotechnology, Brussels, Belgium, [2] Department of Structural Biology, VIB, Vrije Universiteit Brussels, Brussels, Belgium, [3] ULB-VUB - Interuniversity Institute of Bioinformatics in Brussels (IB)^2, Brussels, Belgium

Intrinsically disordered proteins (IDPs) are ubiquitously involved in cellular processes and often implicated in human pathological conditions. The critical biological roles of these proteins, despite not adopting a well-defined fold, encouraged structural biologists to revisit their views on the protein structure-function paradigm. Unfortunately, investigating the characteristics and describing the structural behavior of IDPs is far from trivial, and inferring the function(s) of a disordered protein region remains a major challenge. Computational methods have proven particularly relevant for studying IDPs: on the sequence level their dependence on distinct characteristics determined by the local amino acid context makes sequence-based prediction algorithms viable and reliable tools for large scale analyses, while on the structure level the *in silico* integration of fundamentally different experimental data types is essential to describe the behavior of a flexible protein chain. Here, we offer an overview of the latest developments and computational techniques that aim to uncover how protein function is connected to intrinsic disorder.

Keywords: intrinsically disordered proteins, IDP ensembles, IDP function, disorder prediction, protein ensemble database

OPEN ACCESS

Edited by:
*Piero Andrea Temussi,
Università di Napoli Federico II, Italy*

Reviewed by:
*Alfonso De Simone,
Imperial College London, UK
Henriette Molinari,
Istituto di Chimica delle Macromolecole ISMAC CNR, Italy
Tobias Madl,
Medical University Graz, Austria*

Correspondence:
*Peter Tompa,
Department of Structural Biology,
Flemish Institute of Biotechnology (VIB), Structural Biology Research Center, Vrije Universiteit Brussels,
Pleinlaan 2, Brussels 1050, Belgium
ptompa@vub.ac.be*

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

Received: *29 May 2015*
Accepted: *21 July 2015*
Published: *05 August 2015*

Citation:
Varadi M, Vranken W, Guharoy M and Tompa P (2015) Computational approaches for inferring the functions of intrinsically disordered proteins. Front. Mol. Biosci. 2:45. doi: 10.3389/fmolb.2015.00045

Introduction

The traditional goal of protein structural biology is to relate the well-defined three-dimensional structure(s) of a protein to its biological function. This structure-function paradigm continues to facilitate many important discoveries, but has largely ignored the possible roles of conformational flexibility on function (Forman-Kay and Mittag, 2013). Yet, in the recent years it became apparent that structural disorder is ubiquitously present in diverse cellular processes, and has a particularly prominent role in regulation and signaling events occurring in the complex cellular environment (Tompa et al., 2006; Dunker et al., 2015). Proteins or protein regions that are enriched in conformational flexibility are referred to as intrinsically disordered proteins (IDPs) or protein regions (IDRs) (Dyson and Wright, 2005; Fink, 2005; Tompa, 2005). IDPs and IDRs lack a well-defined, stable three-dimensional fold, and therefore they populate ensembles of dynamically exchanging conformations, separated by low energy barriers. This dynamic behavior challenges the traditional structure-function paradigm (Wright and Dyson, 1999; Chouard, 2011), since it is far from trivial to describe the structural behavior of proteins that adopt such an extensive range of conformations, let alone infer their biological role (Tompa, 2011). Even though intrinsic disorder is occurring ubiquitously—more than 30% of the proteins in the known eukaryotic proteomes have disordered segments of 30 or more consecutive disordered residues

(Dunker et al., 2000)—we are only beginning to understand how protein function arises from the disordered state (Tompa, 2011). Here, we provide an overview of the recent developments in terms of computational methods and data resources that facilitate the understanding of intrinsic disorder and its connection to protein function.

Functional Consequences of Intrinsic Disorder

IDPs and IDRs can be viewed as having complementary functions to those of their folded counterparts. While the latter are often involved in enzymatic activities, molecular transportation or binding short peptides and small molecules, IDPs are mainly involved in signaling, regulation and enzymatic activity inhibition (Xie et al., 2007; Dunker et al., 2015), for example in cell cycle regulation (Yoon et al., 2012), cell division and differentiation (Ward et al., 2004; Xie et al., 2007).

There are a number of possible ways in which an IDP/IDR can realize its function. In perhaps the simplest scenarios, they serve as entropic chains, effectively influencing the orientation and distance between folded domains (Chong et al., 2010), and organizing the super-tertiary structure of the protein (Tompa, 2012). In some cases they are entropic springs or even timers, where the length and flexibility of the linker can determine stochastically how often two folded domains may encounter each other (Bentrop et al., 2001; Smagghe et al., 2010).

Another important role of conformational flexibility is in binding protein or nucleic acid partners. IDPs excel in establishing specific, but transient interactions (Dunker et al., 1998). From an energetic point of view, the reason behind their weaker binding affinities is that the entropic cost of stabilizing a single conformation from the dynamic ensemble that the IDP/IDR is sampling is relatively high (Dyson and Wright, 2005). However, in some cases the fine-tuning of favorable interactions is known to yield surprisingly strong affinities (Ferreon et al., 2013; Follis et al., 2013). An additional advantage of the high degree of conformational freedom is that an IDR can bind very diverse partners because it can easily adopt different conformations (Wang et al., 2011; Hsu et al., 2013) and is often enriched in short binding- and recognition motifs. It is therefore no surprise that IDPs are often hub- (Kim et al., 2008) or scaffold proteins (Dyson and Wright, 2005; Kim et al., 2008; Mittag et al., 2010a) that play essential roles in the cell by integrating signals (Lobley et al., 2007), so increasing the complexity of cellular networks (Dunker et al., 2005; Oldfield et al., 2008). Consequently, IDPs are often implicated in pathological conditions where loss of regulation is the major issue, such as different types of cancer (Andresen et al., 2012). Their involvement in diseases has recently turned IDPs into potential drug targets by either targeting the IDP, or its protein-protein interactions (Funk and Galloway, 1998; Metallo, 2010; Rezaei-Ghaleh et al., 2012).

Conformational flexibility implies high accessibility for potential binding partners and/or enzymes. Consequently, post-translational modification (PTM) sites are often found to be enriched in intrinsically disordered regions (Iakoucheva et al., 2004), with especially phosphorylation sites being prevalent (Gao et al., 2010). While IDPs often go through disorder-to-order transitions upon binding to their partners (Mohan et al., 2006; Wright and Dyson, 2009), in many cases they remain partially or fully flexible in their bound state, forming fuzzy complexes (Tompa and Fuxreiter, 2008; Fuxreiter and Tompa, 2012). One of the advantages of this fuzziness is that PTM sites within the chain can remain relatively accessible, allowing easier regulation of the IDP by modification enzymes (Mittag et al., 2010a). Such regulation by PTM sites is not limited to activation/deactivation of the protein; the modification of the surface of the IDR may also be the prerequisite of binding to a different partner (Oldfield et al., 2008), or even to the same partner, but with increased affinity (Mittag et al., 2010b).

However, intrinsic disorder also has a dark side. In particular, the amino acid compositional bias of IDPs coupled with relatively high propensities to form β-sheets and turns leads to elevated aggregation potentials, and the formation of amyloid-type beta-structures (Levine et al., 2015). Indeed, IDPs have been implicated in aggregation-based diseases, such as Alzheimer's and Parkinson's (Huang and Stultz, 2009; Uversky, 2010).

Sequence-based Investigation of IDPs

There are a number of experimental techniques currently available for identifying and characterizing intrinsic disorder, such as circular dichroism (CD) (Weinreb et al., 1996), protease digestion (Johnson et al., 2012), Förster resonance energy transfer (FRET) (Haas, 2012), Electron Paramagnetic Resonance (EPR) spectroscopy (Drescher, 2012), small-angle X-ray and neutron scattering (SAXS and SANS) (Bernado and Svergun, 2012; Gabel, 2012) and nuclear magnetic resonance spectroscopy (NMR) (Kosol et al., 2013; Konrat, 2014). For initial and for high-throughput investigations, computational methods are however a very popular choice (Ward et al., 2004; Ishida and Kinoshita, 2007). Intrinsic disorder is associated with distinct sequence characteristics; IDPs/IDRs are enriched in "disorder promoting" amino acids, such as charged or polar residues, glycines and prolines, while hydrophobic residues are underrepresented (Uversky et al., 2000).Their conformational flexibility also implies that the local sequence context predominantly dictates the amino acid interactions that can take place, making IDPs more amendable to prediction of their characteristics from sequence. Throughout the last decade many disorder prediction algorithms were designed to exploit the information contained within the amino acid sequence of an IDP; there are more than 50 disorder predictors worldwide (He et al., 2009). The first disorder predictors, such as DisEMBL (Linding et al., 2003) were primarily based on the distinct compositional bias of IDPs. They were followed by faster and more reliable algorithms, such as IUPred (Dosztanyi et al., 2005), RONN (Yang et al., 2005), and Espritz (Walsh et al., 2012). Some of these more advanced methods rely on machine learning techniques (Bellay et al., 2012), or combine the results of several algorithms, such as the meta-predictor metaPrDOS (Ishida and Kinoshita, 2008). Overall, the accuracy of most predictors is consistently above 80%, with the best methods currently peaking around 85% (Monastyrskyy et al., 2014). Alternatively, the novel Dynamine approach predicts backbone dynamics, which correlates (negatively) with intrinsic disorder (Cilia et al., 2014); interestingly, this approach is trained on estimations directly from NMR data and avoids structure-based information, complex machine-learning and evolutionary information (Cilia et al., 2013). The distribution of charged amino acids in the sequence of an IDP can also offer information on whether the protein chain is extended or collapsed (Das and Pappu, 2013).

Often it is unnecessary even to predict disorder, since there are a number of openly accessible online resources that store information of the disorder content of specific proteins. The Disordered Protein Database (DisProt) is the primary one of these sequence-based resources (Sickmeier et al., 2007). DisProt is manually curated, and stores information on proteins for which intrinsic disorder was experimentally determined. Where available, the proteins are also annotated with their known functions. However, DisProt houses data for 694 disordered proteins, which is only a minor fraction of the expected number of IDPs. MobiDB (Potenza et al., 2015) and D2P2 (Oates et al., 2013) on the other hand are online resources that store IDPs identified using prediction algorithms from the whole UniProt in addition to experimentally determined ones.

Sequence information on intrinsic disorder can be exploited for more than merely the prediction of disordered residues. In fact, there are many recent algorithms that aim at predicting functional sites and/or the functional role of IDRs. For example, larger hydrophobic residues such as tryptophan and leucine are often found within peptide motifs that act as recognition units located within IDR segments, called molecular recognition features (MoRFs) (Mohan et al., 2006; Fuxreiter et al., 2007; Brown et al., 2010). Disordered motifs are generally short, 3-15 residue long segments; therefore identifying them poses a computational challenge (Gould et al., 2010). Consequently, predicting functional sites in IDRs is not straightforward, and prone to high false positive rates (Tompa, 2011). Additional layers of information can enhance the performance, for example MoRFpred, which uses order/disorder patterns (Cheng et al., 2007), ANCHOR, which estimates the interaction of a segment with a general partner (Meszaros et al., 2009) or DisCons, which takes into consideration the evolutionary conservation of both the amino acid sequence and of the disorder as a feature (Varadi et al., 2015).

Structural Representation of IDPs

Ideally it would be possible to describe the structure of an IDP/IDR in full atomic detail. Indeed, the set of conformations IDPs sample is often not completely random, for example both p21 and p27 are known to sample distinct secondary structural elements that are biologically relevant and involved in conformation selection (Kriwacki et al., 1996; Sivakolundu et al., 2005). Due to their inherent conformational flexibility, however, the structure of an IDP cannot be described with a single, static conformation (Tompa and Varadi, 2014). This conformational diversity of IDPs precludes crystallization and examination by X-ray crystallography is therefore not a viable option. NMR spectroscopy, while more attuned to conformational diversity, remains hindered by distinct difficulties such as peak overlap (Bellay et al., 2012), while traditional structure calculation protocols do not properly account for multiple conformations (Vranken, 2014).

In response to this challenge, a number of approaches were developed that combine experimental data with computational methodology with the aim to accurately describe the full conformational ensemble adopted by IDPs. Experimental data from techniques that rely on measurements performed in solution are particularly well suited for studying the dynamic structure of an IDP, even though they often represent an average over the different conformations that are adopted by the IDP. These experimental measurements predominantly include NMR-derived parameters, such as chemical shifts (CSs) (Jensen et al., 2011), residual dipolar couplings (RDCs) (Mittag et al., 2010b), paramagnetic relaxation enhancements (PREs) (Mittag et al., 2010b), and J-couplings (Mittag et al., 2010b), as well as scattering intensities from small-angle X-ray scattering (SAXS) (Allison et al., 2009) and probe distances from Forster resonance energy transfer (FRET) (Haas, 2012). These experimental data are then combined with computational methods to determine an ensemble of conformations for an IDP, with two main approaches being used; the first approach is referred to as pool-based modeling, while the second one is based on molecular dynamics (MD) simulations (Tompa and Varadi, 2014) (**Figure 1**).

When using a pool-based approach, such as the ensemble optimization method (EOM) (Bernado et al., 2007) which was designed to model ensembles based on SAXS data, the initial step is to generate a very large random or semi-random pool of conformations based on the amino acid sequence of the IDP/IDR with algorithms such as Flexible-Meccano (Ozenne et al., 2012). The sampling of conformations can be biased using experimental data, such as secondary structure propensities derived from chemical shifts. Next, theoretical parameters are estimated for each conformer in the pool, for example the theoretical scattering intensities calculated by software such as CRYSOL (Bernado and Svergun, 2012). Selection algorithms are then deployed in order to select subsets of the pool in a way that when the theoretical parameters are averaged over a subset, they are in excellent agreement with the experimental data (Sibille and Bernado, 2012). Another example is the ENSEMBLE methodology, which uses as input a large set of conformations together with relevant experimental data, and then prunes the ensemble of conformations to a smaller subset. During the filtering step, conformations are assigned weights so that the resulting ensemble average values fit the experimental input values. Structures that do not contribute to this fitting are discarded (Krzeminski et al., 2013).

In contrast, ensemble modeling procedures are based on molecular dynamics (MD) simulations and begin with random conformations in parallel. Multiple "replica" simulations are initiated using these initial conformations, and constraints are applied over multiple models based on the experimental data, i.e., sets of conformations are required to satisfy experimentally determined constraints, such as pair-wise

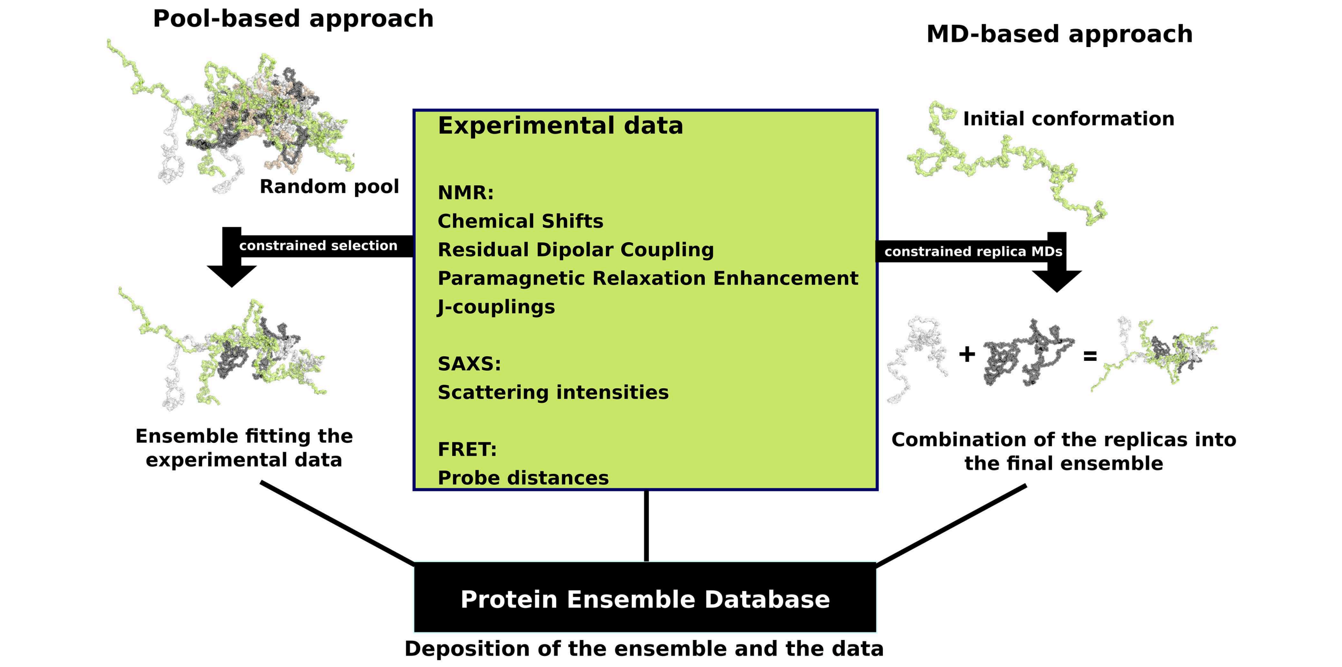

Pool-based approach
Random pool
constrained selection
Ensemble fitting the experimental data
Experimental data
NMR:
Chemical Shifts
Residual Dipolar Coupling
Paramagnetic Relaxation Enhancement
J-couplings
SAXS:
Scattering intensities
FRET:
Probe distances
MD-based approach
Initial conformation
constrained replica MDs
+
=
Combination of the replicas into the final ensemble
Protein Ensemble Database
Deposition of the ensemble and the data

FIGURE 1 | Schematic view of the two main ensemble modeling approaches. Pool-based ensemble modeling (left) starts by generating a pool of random or semi-random conformations based on the protein sequence. Subsets of conformations are selected iteratively from the pool and theoretical parameters are calculated for each conformer in the subset. The final ensemble consists of conformations for which the theoretical parameters are in agreement with the experimental data. MD-based approaches start by initiating short replica MD simulations in parallel using an initial conformation. The MD replicas are constrained with the experimental data. The final ensemble is a combination of the resulting replica runs.

distances or secondary structure propensities (Cavalli et al., 2013). Given the conformational heterogeneity of IDPs, which sample an extensive range of conformations during their biological lifetime, extensive simulations are required to ensure adequate sampling of relevant regions of the conformational space. This significantly increases the computational costs associated with IDP simulations and makes it difficult to achieve even for systems of modest size. Recent developments to address these issues include techniques such as multi-scale enhanced sampling (MSES) (Lee and Chen, 2015) and replica exchange with guided annealing (RE-GA) (Zhang and Chen, 2014). The MSES protocol combines coarse-grained, topology-based models with atomistic force fields to enhance sampling and was recently optimized for simulating IDP conformational ensembles, where it could capture reversible helix-coil transitions (Lee and Chen, 2015). RE-GA has been suggested to be suitable for systems with small conformational transition barriers (as is the case for IDPs), and helped the disordered kinase inducible domain (KID) protein to efficiently escape non-specific compact states while requiring less computation (Zhang and Chen, 2014).

Therefore, the main differences between these two classes of modeling techniques are that the pool-based approach is much faster; conformations can be easily generated, but the final results strongly depend on the quality and diversity of this initial pool of conformations. The ensemble modeling MD approaches are in contrast very slow; conformational sampling depends on the MD simulation, but they have the advantage that the experimental data are continuously applied, a timeline of conformational changes is available, and due to their more rigorous simulation of physical reality, they should give a better representation of the thermally accessible structural ensemble.

These two modeling approaches (i.e., pool and MD-based) are currently the state-of-the-art and have been applied to generate the structural ensemble of many IDPs (**Table 1**). These ensemble models are, however, not straightforward to interpret (Tompa and Varadi, 2014). The key issue is that the experimental information that is available to either filter or constrain during the calculation is very sparse compared to the immense degree of conformational freedom the IDP experiences in solution, resulting in a hugely underdetermined problem. As a direct consequence, many ensembles of models can describe the experimental data equally well, allowing multiple, ambiguous solutions that strongly depend on the calculation approach and the amount and type of experimental data available. In fact, one can model the ensemble of an IDP with an excellent fit to the data, then discard the ensemble and remodel another, unique and different ensemble with an equally good fit. In this sense, ensembles should be considered as a whole and their structural characteristics analyzed as the average over the ensemble, and over-interpretation of single conformations should be avoided. However, if certain characteristics or pre-formed secondary structural elements are consistently modeled in multiple ensembles, then such structural features might be functionally relevant.

TABLE 1 | Recently published ensemble models from the Protein Ensemble Database.

Protein	Data type	Protocol	PED ID	References
Sic1/Cdc4	NMR and SAXS	Pool-based	PED9AAA	Mittag et al., 2010b
p15 PAF	NMR and SAXS	Pool-based	PED6AAA	De Biasio et al., 2014
MKK7	NMR	Pool-based	PED5AAB	Kragelj et al., 2015
Beta-synuclein	NMR	MD-based	PED1AAD	Allison et al., 2014
P27 KID	NMR	MD-based	PED2AAA	Sivakolundu et al., 2005

The field of ensemble modeling therefore still presents exciting opportunities for further development, and several important issues will have to be addressed before the techniques become more reliable. In our view, the first is increasing the number of experimentally derived constraints, which will lead to higher quality models, or cross-validation with new types of experimental data, which will also increase the power of ensembles. The second is the further incorporation of knowledge based information into the calculations, such as the results from reliable predictions or improved force fields. The third is that specific validation and evaluation approaches are required for these ensembles, likely starting from the currently well-developed NMR validation field (Rosato et al., 2013; Vuister et al., 2014) but with better accounting for multiple conformations (Vranken, 2014). Especially NMR CS values, whenever available, are very useful for the estimation of residue-level backbone and side-chain dynamics (Berjanskii and Wishart, 2005, 2013) as well as secondary structure populations (Shen and Sali, 2006; Camilloni et al., 2012), with new methods providing reference chemical shift values for IDPs (Tamiola et al., 2010). They are already effectively used to generate pools with predetermined conformations and as restraints in molecular dynamics simulations (Krieger et al., 2014), but have immense potential for the validation and evaluation of ensembles. Finally, the overwhelming majority of the already generated ensemble models were previously unavailable to the scientific community, impeding the establishment of standardized validation and evaluation protocol. The Protein Ensemble Database (PED) is an international initiative launched to address this issue, effectively making the experimental data and the ensemble models available to the scientific community (Varadi et al., 2014). This is expected to facilitate the development of the next generation of ensemble modeling techniques, and should provide a basis for defining standards of validation and evaluation.

Toward the Functional Interpretation of IDP Ensembles

While ensemble models do not yet possess a predictive power comparable to that of the structure of folded proteins/domains, these models can already offer insights regarding the function of an IDP. Through the integration of experimental data into an ensemble model, functionally important segments might be inferred. For example, transient secondary structural elements in the ensemble of an IDP are often important in terms of function. Such pre-formed elements are often molecular recognition units, playing major roles in binding to various partners. For example, thep27 protein samples transient helices are consistent with the secondary structure of the bound state p27-Cdk2-cyclin (Sivakolundu et al., 2005). Therefore, in accord with the notion of conformational selection, if a certain secondary structural element is sampled in the ensemble, it might be functionally relevant in the bound form as well (Yoon et al., 2012). Again, such interpretations have to be treated with caution given that the ensembles are based on lower resolution, averaged experimental observations, and the ensemble models should therefore be accurate on average, not by single conformations.

Conclusions

IDPs are involved ubiquitously in biological processes, and play essential roles in the regulation of complex cellular systems (Tompa et al., 2006; Dunker et al., 2015). These multi-purpose proteins combine conformational flexibility with an enrichment of binding motifs and post-translational modification sites (Iakoucheva et al., 2004; Wang et al., 2011; Hsu et al., 2013). Due to their biological importance, it is imperative to characterize them and attempt to relate their sequence and structure with their physiological roles (Tompa, 2011). Such an endeavor has the potential to offer valuable insights that can be translated into new drugs and therapies (Funk and Galloway, 1998; Metallo, 2010; Rezaei-Ghaleh et al., 2012). Sequence-based *in silico* techniques such as disorder prediction algorithms are already comparable in terms of accuracy to that of the secondary structure prediction algorithms of folded proteins (Monastyrskyy et al., 2014), and functional prediction algorithms are also widely available (Cheng et al., 2007; Meszaros et al., 2009; Varadi et al., 2015). Yet, a major breakthrough is expected when ensemble models based on diverse experimental data prove to be biologically relevant, so that we can confidently infer specific protein function from the structural representation of an IDP (Tompa and Varadi, 2014). However, in order to realize this goal a number of challenges need to be tackled first (Tompa, 2011). Chief among these issues are improving the amount and available types of experimental data and establishing standardized protocols for the validation and evaluation of the ensemble modeling procedures. Only then can the field advance in terms of increasing the predictive power of ensemble models (Tompa and Varadi, 2014).

Acknowledgments

This work was supported by the Odysseus grant G.0029.12 (FWO, Research Foundation Flanders) to PT and by a VIB international postdoctoral (omics@VIB) Marie-Curie COFUND fellowship for MG. WV acknowledges the Brussels Institute for Research and Innovation (Innoviris) grant BB2B 2010-1-12.

References

Allison, J. R., Rivers, R. C., Christodoulou, J. C., Vendruscolo, M., and Dobson, C. M. (2014). A relationship between the transient structure in the monomeric state and the aggregation propensities of alpha-synuclein and beta-synuclein. *Biochemistry* 53, 7170–7183. doi: 10.1021/bi5009326

Allison, J. R., Varnai, P., Dobson, C. M., and Vendruscolo, M. (2009). Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements. *J. Am. Chem. Soc.* 131, 18314–18326. doi: 10.1021/ja904716h

Andresen, C., Helander, S., Lemak, A., Farès, C., Csizmok, V., Carlsson, J., et al. (2012). Transient structure and dynamics in the disordered c-Myc transactivation domain affect Bin1 binding. *Nucleic Acids Res.* 40, 6353–6366. doi: 10.1093/nar/gks263

Bellay, J., Michaut, M., Kim, T., Han, S., Colak, R., Myers, C. L., et al. (2012). An omics perspective of protein disorder. *Mol. Biosyst.* 8, 185–193. doi: 10.1039/C1MB05235G

Bentrop, D., Beyermann, M., Wissmann, R., and Fakler, B. (2001). NMR structure of the "ball-and-chain" domain of KCNMB2, the beta 2-subunit of large conductance Ca2+- and voltage-activated potassium channels. *J. Biol. Chem.* 276, 42116–42121. doi: 10.1074/jbc.M107118200

Berjanskii, M. V., and Wishart, D. S. (2005). A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.* 127, 14970–14971. doi: 10.1021/ja054842f

Berjanskii, M. V., and Wishart, D. S. (2013). A simple method to measure protein side-chain mobility using NMR chemical shifts. *J. Am. Chem. Soc.* 135, 14536–14539. doi: 10.1021/ja407509z

Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., and Svergun, D. I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* 129, 5656–5664. doi: 10.1021/ja069124n

Bernadó, P., and Svergun, D. I. (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.* 8, 151–167. doi: 10.1039/C1MB05275F

Brown, C. J., Johnson, A. K., and Daughdrill, G. W. (2010). Comparing models of evolution for ordered and disordered proteins. *Mol. Biol. Evol.* 27, 609–621. doi: 10.1093/molbev/msp277

Camilloni, C., De Simone, A., Vranken, W. F., and Vendruscolo, M. (2012). Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* 51, 2224–2231. doi: 10.1021/bi3001825

Cavalli, A., Camilloni, C., and Vendruscolo, M. (2013). Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* 138, 094112. doi: 10.1063/1.4793625

Cheng, Y., Oldfield, C. J., Meng, J., Romero, P., Uversky, V. N., and Dunker, A. K. (2007). Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 46, 13468–13477. doi: 10.1021/bi7012273

Chong, P. A., Lin, H., Wrana, J. L., and Forman-Kay, J. D. (2010). Coupling of tandem Smad ubiquitination regulatory factor (Smurf) WW domains modulates target specificity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18404–18409. doi: 10.1073/pnas.1003023107

Chouard, T. (2011). Structural biology: breaking the protein rules. *Nature* 471, 151–153. doi: 10.1038/471151a

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2013). From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.* 4, 2741. doi: 10.1038/ncomms3741

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.* 42, W264–W270. doi: 10.1093/nar/gku270

Das, R. K., and Pappu, R. V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13392–13397. doi: 10.1073/pnas.1304749110

De Biasio, A., Ibáñez de Opakua, A., Cordeiro, T. N., Villate, M., Merino, N., Sibille, N., et al. (2014). p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. *Biophys. J.* 106, 865–874. doi: 10.1016/j.bpj.2013.12.046

Dosztányi, Z., Csizmák, V., Tompa, P., and Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* 347, 827–839. doi: 10.1016/j.jmb.2005.01.071

Drescher, M. (2012). EPR in protein science: intrinsically disordered proteins. *Top. Curr. Chem.* 321, 91–119. doi: 10.1007/128_2011_235

Dunker, A. K., Bondos, S. E., Huang, F., and Oldfield, C. J. (2015). Intrinsically disordered proteins and multicellular organisms. *Semin. Cell Dev. Biol.* 37, 44–55. doi: 10.1016/j.semcdb.2014.09.025

Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* 272, 5129–5148. doi: 10.1111/j.1742-4658.2005.04948.x

Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., et al. (1998). Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.* 473–484.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* 11, 161–171.

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. Nature reviews. *Mol. Cell Biol.* 6, 197–208. doi: 10.1038/nrm1589

Ferreon, A. C., Ferreon, J. C., Wright, P. E., and Deniz, A. A. (2013). Modulation of allostery by protein intrinsic disorder. *Nature* 498, 390–394. doi: 10.1038/nature12294

Fink, A. L. (2005). Natively unfolded proteins. *Curr. Opin. Struct. Biol.* 15, 35–41. doi: 10.1016/j.sbi.2005.01.002

Follis, A. V., Chipuk, J. E., Fisher, J. C., Yun, M. K., Grace, C. R., Nourse, A., et al. (2013). PUMA binding induces partial unfolding within BCL-xL to disrupt p53 binding and promote apoptosis. *Nat. Chem. Biol.* 9, 163–168. doi: 10.1038/nchembio.1166

Forman-Kay, J. D., and Mittag, T. (2013). From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* 21, 1492–1499. doi: 10.1016/j.str.2013.08.001

Funk, J. O., and Galloway, D. A. (1998). Inhibiting CDK inhibitors: new lessons from DNA tumor viruses. *Trends Biochem. Sci.* 23, 337–341. doi: 10.1016/S0968-0004(98)01242-0

Fuxreiter, M., and Tompa, P. (2012). Fuzzy complexes: a more stochastic view of protein function. *Adv. Exp. Med. Biol.* 725, 1–14. doi: 10.1007/978-1-4614-0659-4_1

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23, 950–956. doi: 10.1093/bioinformatics/btm035

Gabel, F. (2012). Small angle neutron scattering for the structural study of intrinsically disordered proteins in solution: a practical guide. *Methods Mol. Biol.* 896, 123–135. doi: 10.1007/978-1-4614-3704-8_8

Gao, J., Thelen, J. J., Dunker, A. K., and Xu, D. (2010). Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell. Proteom.* 9, 2586–2600. doi: 10.1074/mcp.M110.001388

Gould, C. M., Diella, F., Via, A., Puntervoll, P., Gemund, C., Chabanis-Davidson, S., et al. (2010). ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.* 38, D167–D180. doi: 10.1093/nar/gkp1016

Haas, E. (2012). Ensemble FRET methods in studies of intrinsically disordered proteins. *Methods Mol. Biol.* 895, 467–498. doi: 10.1007/978-1-61779-927-3_28

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. N., and Dunker, A. K. (2009). Predicting intrinsic disorder in proteins: an overview. *Cell Res.* 19, 929–949. doi: 10.1038/cr.2009.87

Hsu, W. L., Oldfield, C. J., Xue, B., Meng, J., Huang, F., Romero, P., et al. (2013). Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Prot. Sci.* 22, 258–273. doi: 10.1002/pro.2207

Huang, A., and Stultz, C. M. (2009). Finding order within disorder: elucidating the structure of proteins associated with neurodegenerative disease. *Future Med. Chem.* 1, 467–482. doi: 10.4155/fmc.09.40

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., et al. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32, 1037–1049. doi: 10.1093/nar/gkh253

Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35, W460–W464. doi: 10.1093/nar/gkm363

Ishida, T., and Kinoshita, K. (2008). Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* 24, 1344–1348. doi: 10.1093/bioinformatics/btn195

Jensen, M. R., Communie, G., Ribeiro, E. A. Jr., Martinez, N., Desfosses, A., Salmon, L., et al. (2011). Intrinsic disorder in measles virus nucleocapsids. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9839–9844. doi: 10.1073/pnas.1103270108

Johnson, D. E., Xue, B., Sickmeier, M. D., Meng, J., Cortese, M. S., Oldfield, C. J., et al. (2012). High-throughput characterization of intrinsic disorder in proteins from the Protein Structure Initiative. *J. Struct. Biol.* 180, 201–215. doi: 10.1016/j.jsb.2012.05.013

Kim, P. M., Sboner, A., Xia, Y., and Gerstein, M. (2008). The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.* 4, 179. doi: 10.1038/msb.2008.16

Konrat, R. (2014). NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J. Magn. Reson.* 241, 74–85. doi: 10.1016/j.jmr.2013.11.011

Kosol, S., Contreras-Martos, S., Cedeño, C., and Tompa, P. (2013). Structural characterization of intrinsically disordered proteins by NMR spectroscopy. *Molecules* 18, 10802–10828. doi: 10.3390/molecules180910802

Kragelj, J., Palencia, A., Nanao, M. H., Maurin, D., Bouvignies, G., Blackledge, M., et al. (2015). Structure and dynamics of the MKK7-JNK signaling complex. *Proc. Natl. Acad. Sci. U.S.A.* 112, 3409–3414. doi: 10.1073/pnas.1419528112

Krieger, J. M., Fusco, G., Lewitzky, M., Simister, P. C., Marchant, J., Camilloni, C., et al. (2014). Conformational recognition of an intrinsically disordered protein. *Biophys. J.* 106, 1771–1779. doi: 10.1016/j.bpj.2014.03.004

Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., and Wright, P. E. (1996). Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. U.S.A.* 93, 11504–11509. doi: 10.1073/pnas.93.21.11504

Krzeminski, M., Marsh, J. A., Neale, C., Choy, W. Y., and Forman-Kay, J. D. (2013). Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* 29, 398–399. doi: 10.1093/bioinformatics/bts701

Lee, K. H., and Chen, J. (2015). Multiscale enhanced sampling of intrinsically disordered protein conformations. *J. Comp. Chem.* doi: 10.1002/jcc.23957. [Epub ahead of print].

Levine, Z. A., Larini, L., LaPointe, N. E., Feinstein, S. C., and Shea, J. E. (2015). Regulation and aggregation of intrinsically disordered peptides. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2758–2763. doi: 10.1073/pnas.1418155112

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure* 11, 1453–1459. doi: 10.1016/j.str.2003.10.002

Lobley, A., Swindells, M. B., Orengo, C. A., and Jones, D. T. (2007). Inferring function using patterns of native disorder in proteins. *PLoS Comput. Biol.* 3:e162. doi: 10.1371/journal.pcbi.0030162

Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* 5:e1000376. doi: 10.1371/journal.pcbi.1000376

Metallo, S. J. (2010). Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.* 14, 481–488. doi: 10.1016/j.cbpa.2010.06.169

Mittag, T., Kay, L. E., and Forman-Kay, J. D. (2010a). Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recogn.* 23, 105–116. doi: 10.1002/jmr.961

Mittag, T., Marsh, J., Grishaev, A., Orlicky, S., Lin, H., Sicheri, F., et al. (2010b). Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* 18, 494–506. doi: 10.1016/j.str.2010.01.020

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., et al. (2006). Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* 362, 1043–1059. doi: 10.1016/j.jmb.2006.07.087

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., and Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins* 82(Suppl. 2), 127–137. doi: 10.1002/prot.24391

Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., et al. (2013). D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.* 41, D508–D516. doi: 10.1093/nar/gks1226

Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., and Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9(Suppl. 1):S1. doi: 10.1186/1471-2164-9-S1-S1

Ozenne, V., Bauer, F., Salmon, L., Huang, J. R., Jensen, M. R., Segard, S., et al. (2012). Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 28, 1463–1470. doi: 10.1093/bioinformatics/bts172

Potenza, E., Di Domenico, T., Walsh, I., and Tosatto, S. C. (2015). MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* 43, D315–D320. doi: 10.1093/nar/gku982

Rezaei-Ghaleh, N., Blackledge, M., and Zweckstetter, M. (2012). Intrinsically disordered proteins: from sequence and conformational properties toward drug discovery. *Chembiochem* 13, 930–950. doi: 10.1002/cbic.201200093

Rosato, A., Tejero, R., and Montelione, G. T. (2013). Quality assessment of protein NMR structures. *Curr. Opin. Struct. Biol.* 23, 715–724. doi: 10.1016/j.sbi.2013.08.005

Shen, M. Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Prot. Sci.* 15, 2507–2524. doi: 10.1110/ps.062416606

Sibille, N., and Bernadó, P. (2012). Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. *Biochem. Soc. Trans.* 40, 955–962. doi: 10.1042/BST20120149

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., et al. (2007). DisProt: the database of disordered proteins. *Nucleic Acids Res.* 35, D786–D793. doi: 10.1093/nar/gkl893

Sivakolundu, S. G., Bashford, D., and Kriwacki, R. W. (2005). Disordered p27Kip1 exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation. *J. Mol. Biol.* 353, 1118–1128. doi: 10.1016/j.jmb.2005.08.074

Smagghe, B. J., Huang, P. S., Ban, Y. E., Baker, D., and Springer, T. A. (2010). Modulation of integrin activation by an entropic spring in the {beta}-knee. *J. Biol. Chem.* 285, 32954–32966. doi: 10.1074/jbc.M110.145177

Tamiola, K., Acar, B., and Mulder, F. A. (2010). Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J. Am. Chem. Soc.* 132, 18000–18003. doi: 10.1021/ja105656t

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* 579, 3346–3354. doi: 10.1016/j.febslet.2005.03.072

Tompa, P. (2011). Unstructural biology coming of age. *Curr. Opin. Struct. Biol.* 21, 419–425. doi: 10.1016/j.sbi.2011.03.012

Tompa, P. (2012). On the supertertiary structure of proteins. *Nat. Chem. Biol.* 8, 597–600. doi: 10.1038/nchembio.1009

Tompa, P., Dosztanyi, Z., and Simon, I. (2006). Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J. Proteom. Res.* 5, 1996–2000. doi: 10.1021/pr0600881

Tompa, P., and Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 33, 2–8. doi: 10.1016/j.tibs.2007.10.003

Tompa, P., and Varadi, M. (2014). Predicting the predictive power of IDP ensembles. *Structure* 22, 177–178. doi: 10.1016/j.str.2014.01.003

Uversky, V. N. (2010). Targeting intrinsically disordered proteins in neurodegenerative and protein dysfunction diseases: another illustration of the D(2) concept. *Exp. Rev. Proteom.* 7, 543–564. doi: 10.1586/epr.10.36

Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41, 415–427. doi: 10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Varadi, M., Guharoy, M., Zsolyomi, F., and Tompa, P. (2015). DisCons: a novel tool to quantify and classify evolutionary conservation of intrinsic protein disorder. *BMC Bioinform.* 16:153. doi: 10.1186/s12859-015-0592-2

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A. K., et al. (2014). pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* 42, D326–D335. doi: 10.1093/nar/gkt960

Vranken, W. F. (2014). NMR structure validation in relation to dynamics and structure determination. *Prog. Nuclear Magn. Reson. Spectr.* 82, 27–38. doi: 10.1016/j.pnmrs.2014.08.001

Vuister, G. W., Fogh, R. H., Hendrickx, P. M., Doreleijers, J. F., and Gutmanas, A. (2014). An overview of tools for the validation of protein NMR structures. *J. Biomol. NMR* 58, 259–285. doi: 10.1007/s10858-013-9750-x

Walsh, I., Martin, A. J., Di Domenico, T., and Tosatto, S. C. (2012). ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* 28, 503–509. doi: 10.1093/bioinformatics/btr682

Wang, Y., Fisher, J. C., Mathew, R., Ou, L., Otieno, S., Sublet, J., et al. (2011). Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21. *Nat. Chem. Biol.* 7, 214–221. doi: 10.1038/nchembio.536

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337, 635–645. doi: 10.1016/j.jmb.2004.02.002

Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T. Jr. (1996). NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* 35, 13709–13715. doi: 10.1021/bi961799n

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. (2009). Linking folding and binding. *Curr. Opin. Struct. Biol.* 19, 31–38. doi: 10.1016/j.sbi.2008.12.003

Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N., et al. (2007). Functional anthology of intrinsic disorder. 1.Biological processes and functions of proteins with long disordered regions. *J. Proteom. Res.* 6, 1882–1898. doi: 10.1021/pr060392u

Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21, 3369–3376. doi: 10.1093/bioinformatics/bti534

Yoon, M. K., Mitrea, D. M., Ou, L., and Kriwacki, R. W. (2012). Cell cycle regulation by the intrinsically disordered proteins p21 and p27. *Biochem. Soc. Trans.* 40, 981–988. doi: 10.1042/BST20120092

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 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 July 2015
doi: 10.3389/fmolb.2015.00039

The use of ion mobility mass spectrometry to probe modulation of the structure of p53 and of MDM2 by small molecule inhibitors

Eleanor R. Dickinson[1], *Ewa Jurneczko*[2], *Judith Nicholson*[2, 3], *Ted R. Hupp*[3], *Joanna Zawacka-Pankau*[4], *Galina Selivanova*[4] and *Perdita E. Barran*[1*]

[1] The Michael Barber Centre for Collaborative Mass Spectrometry, Manchester Institute of Biotechnology, School of Chemistry, University of Manchester, Manchester, UK, [2] School of Chemistry, University of Edinburgh, Edinburgh, UK, [3] Institute of Genetics and Molecular Medicine, CRUK Cancer Research Centre, University of Edinburgh, Edinburgh UK, [4] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

OPEN ACCESS

Edited by:
Kris Pauwels,
Vrije Universiteit Brussel, Belgium

Reviewed by:
Tom Edwards,
University of Leeds, UK
Manuel M. Sánchez Del Pino,
Universidad de Valencia, Spain

***Correspondence:**
Perdita E. Barran,
The Michael Barber Centre for Collaborative Mass Spectrometry,
Manchester Institute of Biotechnology,
School of Chemistry, University of Manchester, 131 Princess Street,
Manchester M1 7DN, UK
perdita.barran@manchester.ac.uk

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

Received: 27 April 2015
Accepted: 22 June 2015
Published: 10 July 2015

Citation:
Dickinson ER, Jurneczko E, Nicholson J, Hupp TR, Zawacka-Pankau J, Selivanova G and Barran PE (2015) The use of ion mobility mass spectrometry to probe modulation of the structure of p53 and of MDM2 by small molecule inhibitors.
Front. Mol. Biosci. 2:39.
doi: 10.3389/fmolb.2015.00039

Developing drug-like molecules to inhibit the interactions formed by disordered proteins is desirable due to the high correlation of disorder with protein implicated in disease, but is challenging due in part to the lack of atomistically resolved and resolvable structures from conformationally dynamic systems. Ion mobility mass spectrometry (IM-MS) is well-positioned to assess protein ligand interactions along with the effect of a given inhibitor on conformation. Here we demonstrate the use of IM-MS to characterize the effect of two inhibitors RITA and Nutlin-3 on their respective binding partners: p53 and MDM2. RITA binds N-terminal transactivation domain of p53 (Np53) weakly, preventing direct observation of the complex in the gas phase. Nonetheless, upon incubation with RITA, we observe an alteration in the charge state distribution and in the conformational distributions adopted by Np53 in the gas phase. This finding supports the hypothesis that RITAs mode of action proceeds *via* a conformational change in p53. Circular dichroism corroborates our gas phase findings, showing a slight increase in secondary structure content on ligand incubation, and HDX-MS experiments also highlight the dynamic properties of this protein. Using the same approach we present data to show the effect of Nutlin-3 binding to the N-terminal domain of MDM2 (N-MDM2), N-MDM2 presents as at least two conformational families in the absence of Nutlin-3. Upon Nutlin-3 binding, the protein undergoes a compaction event similar to that exhibited by RITA on Np53. This multi-technique approach highlights the inherent disorder in these systems; and in particular exemplifies the power of IM-MS as a technique to study transient interactions between small molecule inhibitors and intrinsically disordered proteins.

Keywords: conformational dynamics, ion mobility mass spectrometry, p53, MDM2, small molecule modulation

Introduction

The transcription factor p53, dubbed the Death Star (Vousden, 2000), is a multi-domain, intrinsically disordered protein (IDP) (Bell et al., 2002; Dawson et al., 2003). The protein comprises the disordered N-terminal domain (Np53) (Joerger and Fersht, 2008) containing the transactivation domain (residues 1–61) and the proline-rich domain (residues 62–94), the central DNA binding domain (residues 94–292), the tetramerization domain (residues 325–355) and the C-terminal regulatory domain (residues 363–393). It is strongly implicated in tumor suppression pathways, where it functions to block tumor development by triggering cellular senescence or apoptosis upon signals indicating DNA damage, oncogene activation, or telomere erosion (Vousden and Prives, 2009). Under non-stressed conditions, low p53 levels are tightly maintained by MDM2 (murine double minute 2). MDM2 is a ~55 KDa IDP with roles as an Ubiquitin E3 ligase, as a molecular chaperone and also in translational control. MDM2 comprises the disordered "lid" mini-domain (residues 1–24), (Uhrinova et al., 2005) the N-terminal domain (residues 25–109), the disordered central acidic domain (residues 221–276), the zinc finger domain (residues 299–331), and the C-terminal RING (really interesting new gene) domain (residues 430–480). MDM2 down regulates p53 activity in a negative autoregulatory feedback loop via three mechanisms; firstly, MDM2 blocks the transcription ability of p53 by direct binding through their respective N-terminal domains (Wu et al., 1993; Haupt et al., 1997). Secondly, MDM2 exports p53 from the nucleus and thirdly, targets p53 by Ubiquitination for degradation via the proteasome (Freedman and Levine, 1998; Tao and Levine, 1999). p53 N-terminal domain binds into the MDM2 N-terminal domain hydrophobic pocket as an amphipathic helix, with residues Ph19, Trp23, and Leu26 comprising a triad of required contacts which insert into the MDM2 binding cleft (Kussie et al., 1996).

Alteration of the p53 pathway is an almost universal hallmark of human cancers, with 22 million cancer patients living with abrogation of the p53 pathway, half of which display suppressed p53 function (Brown et al., 2009) and half of which exhibit p53 mutations. Cellular overexpression of MDM2 effectively abolishes p53 function, allowing unregulated cell cycle events in tumor cells. Inhibition of the p53:MDM2 complex is therefore a highly desirable therapeutic strategy; releasing, reactivating and stabilizing p53 levels, thus providing an attractive cancer therapy drug target. To date, numerous p53:MDM2 protein–protein interaction (PPI) antagonists have been identified, including cis-imidazolines (Vassilev et al., 2004; Vu et al., 2013), "stapled" peptides (Brown et al., 2012; Chang et al., 2013), terphenyls (Yin et al., 2005), oligobenzamides (Lu et al., 2006), spiro-oxindoles (Ding et al., 2006), chromenotriazolopyrimidine (Rew et al., 2012), Benzodiazepinedione (Grasberger et al., 2005), and Chromenotriazolopyrimidines (Allen et al., 2009). The cis-imidazoline Nutlin-3 is composed of enantiomers a and b, of which enantiomer a is 150 times more potent, and binds MDM2 in the p53 peptide groove, mimicking the three p53 residues responsible for the bulk of binding interactions (Vassilev et al., 2004). Nutlin-3 is effective in numerous cell lines, and is able to arrest or induce apoptosis in proliferating cancer cells with micromolar concentrations (Tovar et al., 2006). The drug candidate RITA (reactivation of p53 and induction of tumor cell apoptosis, NSC 652287) has been shown to restore wild-type p53 function in tumor cells by preventing the p53:MDM2 interaction (Issaeva et al., 2004). In contrast to the Nutlins, which bind MDM2 in its N-terminal hydrophobic pocket (Vassilev et al., 2004), RITA binds to p53 N-terminal domain with estimated $K_D = 1.5$ nM. It is hypothesized that RITA binds outside of the p53/MDM2 binding cleft, allosterically exerting its effect via a conformational change in the highly disordered N-terminus of p53 (Np53) (Issaeva et al., 2004).

Since its advent in the 1970's (Hogg and Kebarle, 1965; Kebarle and Hogg, 1965), the hybrid gas phase technique Ion Mobility-Mass Spectrometry (IM-MS) has gained credibility as a tool to study the conformations adopted by proteins and peptides in the gas phase. IM-MS is especially effective in its use for studying IDPs (Bernstein et al., 2004; Harvey et al., 2012; Pagel et al., 2013) due to its ability to observe conformations adopted by analytes on a millisecond time scale (Wyttenbach et al., 2001; McCullough et al., 2008). IM-MS provides information regarding charge, mass and shape of an analyte. The simplest setup of IM-MS is that of drift time IM-MS (DT IM-MS) (McAfee and Edelson, 1963). Ions are separated by their mobility (K) as they traverse a drift cell of known length filled with buffer gas to a known pressure and temperature. Ions travel down a weak electric field (5–50 V cm^{-1}) colliding with buffer gas molecules which counter their progress until an equilibrium drift velocity, proportional to the electric field, is reached. The mobility (K) of an ion is the ratio between the drift velocity (v_d) and the applied electric field (E). The mobility of an ion can be used to calculate the rotationally averaged collision cross section (CCS, Ω, Å^2) using Equation (1) (Mason and McDaniel, 2005):

$$K_0 = \frac{3ze}{16N}\left(\frac{2\pi}{\mu k_B T}\right)^{0.5}\frac{1}{\Omega} \tag{1}$$

Where K_0 is the reduced mobility; z is the ion charge state; e is the elementary charge; N is the gas number density; μ is the reduced mass of the ion-neutral pair; k_B is the Boltzmann constant, and T is the gas temperature.

Here we employ native mass spectrometry, DT IM-MS, circular dichroism (CD) and hydrogen-deuterium exchange coupled to mass spectrometry (HDX-MS) to observe the conformations of N-terminal p53 domain (Np53) and the N-terminal domain of MDM2 (N-MDM2) both in the gas phase and in solution. We also probe the binding and conformational changes conferred by small molecule inhibitors; Nutlin-3 for N-MDM2, and RITA for Np53. Further information about DT IM-MS, CD and HDX-MS methodology can be found in the Supporting Information.

Materials and Methods

Expression and purification of both Np53 (residues 1–100) (Szekely et al., 1993; Bakalkin et al., 1995) and N-MDM2 (residues 1–126) (Worrall et al., 2010) have been previously described. Before the analysis reported here, the protein samples were thawed and dialysed in 50 mM ammonium acetate using Bio-RAD micro bio-spin chromatography columns (Bio-Rad Laboratories, Inc.). Concentrations of purified proteins were measured by the Thermo Scientific NanoDrop Spectrophotometer ND 1000 (Thermo Scientific, USA). Small molecule RITA [2,5-bis(5-hydroxymethyl-2-thienyl) furan, NSC 652287] was reconstituted in 100% IPA and stored at −20°C. Before analysis, RITA was thawed and diluted to 100 μM and an IPA concentration of 5% using 50 mM ammonium acetate. Nutlin-3 was reconstituted in 100% DMSO and stored at −80°C. Before analysis, Nutlin-3 was thawed and diluted to 500 μM and a DMSO concentration of 1% using 50 mM ammonium acetate.

MS and IM-MS experiments were performed on Np53 and N-MDM2 from solutions buffered with ammonium acetate (pH 6.8). Np53 samples were incubated with 5% IPA for 30 min at 37°C to account for the solvent present in the RITA sample. N-MDM2 samples were incubated with 0.5% DMSO for 30 min at room temperature to account for the solvent present in the Nutlin-3 sample. Binding experiments were performed on Np53 with RITA in a 1:2 protein:ligand ratio, samples were incubated for 30 min at 37°C. Binding experiments were performed on N-MDM2 and Nutlin-3 in a 1:10 protein:ligand ratio, samples were incubated for 30 min at room temperature. All MS and DT IM-MS data were acquired on an in-house modified quadropole time-of-flight mass spectrometer (Waters, Manchester, UK) (McCullough et al., 2008) containing a copper drift cell of length 5.1 cm. Ions were produced by positive nano-electrospray ionization (nESI) with a spray voltage of 1.3–1.62 kV. Helium was used as the buffer gas, its pressure measured using a baratron (MKS Instruments, UK). Buffer gas temperature and pressure readings (294.31–303.69 K and 3.518–3.898 Torr, respectively) were taken at each drift voltage and used in the analysis of drift time measurements. The drift voltage across the cell was varied by decreasing the cell body potential from 60 to 15 V, with arrival time measurements taken at a minimum of five distinct voltages. Instrument parameters were kept as constant as possible and are as follows: cone voltage: 114–119 V, source temperature: 80°C.

nESI tips were prepared in-house using a micropipette puller (Fleming/Brown model P-97, Sutter Instruments Co., USA) using 4″ 1.2 mm thin wall glass capillaries (World Precision Instruments, Inc., USA) and filled with 10–20 μL of sample.

Data was analyzed using MassLynx v4.1 software (Waters, Manchester, UK), Origin v9.0 (OriginLab Corporation, USA) and Microsoft Excel. Experiments were carried out in triplicate. Ion arrival time distributions were recorded by synchronization of the release of ions into the drift cell with mass spectral acquisition. The collision cross section distributions (CCSD) are derived from arrival time data using Equation (2) (Mason and McDaniel, 2005):

$$\Omega_{\text{avg}} = \frac{(18\pi)^{1/2}}{16}\left[\frac{1}{m_b} + \frac{1}{m}\right]^{1/2} \frac{ze}{(K_B T)^{1/2}} \frac{1}{\rho} \frac{t_d V}{L^2} \qquad (2)$$

Where Ω is the collision cross section (Å^2); m and m_b are the masses of the ion and buffer gas, respectively; z is the ion charge state; e is the elementary charge; k_B is the Boltzmann constant; T is the gas temperature; ρ is the buffer gas density; L is the drift tube length; V is the voltage across the drift tube; and t_d is the drift time. For these experiments where the CCS has been evaluated experimentally with helium as the buffer gas and using a drift tube with a linear field we use the convention $^{\text{DT}}\text{CCS}_{\text{He}}$ to report our collision cross section values in the context of the mobility technique employed as well as the buffer gas used.

HDX-MS experiments were carried out using a fully automated LEAP autosampler system (HTS PAL, Leap Technologies, Carrboro, NC, USA) previously described (Chalmers et al., 2006; Zhang et al., 2009) and an online Acquity UPLC M-class HDX System (Waters Inc., Manchester, UK). Np53 and RITA were mixed at a 1:2 protein:ligand ratio and incubated for 30 min at 37°C before analysis. Stock protein solutions (50 μM Np53 ± 100 μM RITA, with 5% IPA) were diluted to 10 μM with equilibration buffer. 3.8 μl protein solution was incubated with D_2O (54.2 μl labeling buffer) and incubated at 18°C for 15, 30, 60, or 120 s. Following deuterium on-exchange, 50 μl of the labeled protein solution was quenched by adding 50 μl of quench buffer at 1°C, and samples were passed across an immobilized pepsin column (enzymate BEH pepsin column, Waters Inc., Manchester, UK) at 100 μL min^{-1} (H_2O + 0.1% formic acid, 20°C). The resulting peptides were trapped on a UPLC BEH C_{18} Van-Guard Pre-column (Waters Inc., Manchester, UK) and then gradient eluted (1 min loading time, 8–85% ACN + 0.1% formic acid gradient, 40 μl min^{-1}, 1°C) across a UPLC BEH C_{18} column (Waters Inc., Manchester, UK) before undergoing electrospray ionization and analysis using a Synapt G2Si mass spectrometer (Waters Inc., Manchester, UK). Data was analyzed using ProteinLynx Global Server (PLGS) (Waters, Manchester, UK), Dynamx v1.0 (Waters, Manchester, UK) and Origin v9.0 (OriginLab Corporation, USA).

Hundred percentage of sequence coverage was obtained for Np53 ± RITA. Selected peptides were restricted to be present in all three repeats of 0 s incubation time experiments.

Results

Modulation of N-terminal p53 by RITA

In the absence of RITA and under near neutral solution pH conditions, the mass spectra of Np53 (**Figure 1A**) presents a broad monomeric charge state distribution (CSD) range $5 \leq z \leq 12$, with three major signals corresponding to the ions $[M+6H]^{6+}$, $[M+7H]^{7+}$, and $[M+8H]^{8+}$, of which the $[M+6H]^{6+}$ species is most intense. Upon incubation of Np53 with RITA we observe a shift in the CSD toward lower charge states. Specifically, Np53 in the presence of RITA (**Figure 1B**) exhibits a significant decrease in intensity of the $[M+7H]^{7+}$ and $[M+8H]^{8+}$ species, along with an increase in the intensity of the $[M+5H]^{5+}$ species, an appearance of the $[M+4H]^{4+}$ species and a loss of the high charge states $z > 10$. Although source conditions were carefully controlled to give gentle ionization of the sample, the Np53:RITA complex was not strong enough to be retained during desolvation at any protein:ligand ratio (data not shown).

DT IM-MS was performed on Np53 both in the absence and presence of RITA. The collision cross section distribution

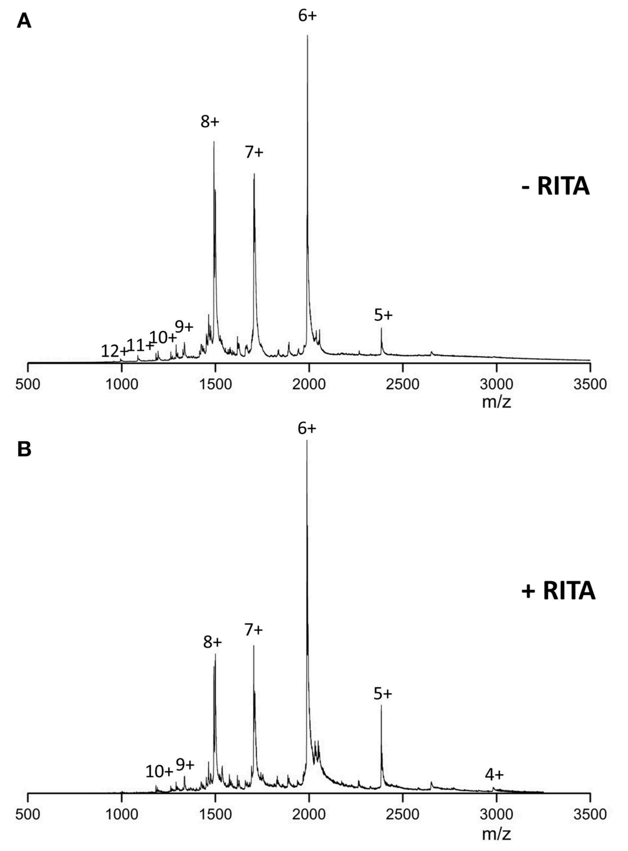


FIGURE 1 | n-ESI mass spectra recorded for (A) wild-type Np53 and (B) wild-type Np53 in the presence of RITA. Np53 was incubated for 30 min at 37°C with addition of 5% IPA. Binding studies were carried out at a 1:2 Np53:RITA ratio, incubated for 30 min at 37°C with 5% IPA content.

($^{DT}CCSD_{He}$) (**Figure 2A** top panel) shows the Np53 $[M+6H]^{6+}$ charge state presents as two conformational families; a more populated compact form (denoted C_1, blue Gaussian distribution) centered at ~1250 Å^2 and a low intensity extended form (denoted X, green Gaussian distribution) centered at ~1500 Å^2. Two conformations are also observed for $[M+7H]^{7+}$ (**Figure 2B**), which are assigned to X and a more intense larger distribution, centered at ~1750 Å^2, which is assigned to an unfolded form of the protein (U, purple Gaussian distribution). $[M+8H]^{8+}$ (**Figure 2C**), is also made up of U, along with low intensity signal from a still more extended form (U_2, gold Gaussian distribution), although this latter distribution is poorly resolved. $[M+9H]^{9+}$ (Figure S2) presents in three conformational families; X, U, and U2, of which the most extended U2 is most populated.

Upon incubation with RITA the $^{DT}CCSD_{He}$ for $[M+6H]^{6+}$ is significantly altered (**Figure 2A** bottom panel); we no longer observe the extended conformer X, observe a reduction in the population of compact conformer C_1, and the induction of a highly populated novel conformational family centered at ~950 Å^2, C_0 (red Gaussian distribution). The $[M+7H]^{7+}$ CCSD (**Figure 2B**) is also altered by the presence of RITA, with loss of conformer U, and induction of both conformers C_1 and C_0. This change is accompanied by a decrease in intensity of this charge state. $[M+8H]^{8+}$ (**Figure 2C**) behaves similarly to $[M+7H]^{7+}$, with loss of conformer U_2, and induction of conformers X and C_1. We observe an increase in the intensity of the $[M+5H]^{5+}$ species (Figure S1) along with the appearance of a highly compact form of the protein C_0. This compaction is evident in all charge states, for example $[M+9H]^{9+}$ (Figure S2) has lost the population of the unfolded conformer U_2 upon incubation with RITA, alongside a reduction in intensity of conformers X and U and induction of highly populated C_1 conformational family. This alteration of the conformational spread as shown by the $^{DT}CCSD_{He}$ is supported in solution by CD. Figure S3 (Supporting Information) shows the secondary structure content of Np53 increases upon incubation with RITA, supporting the hypothesis that RITA induces a novel conformer of Np53. Structural analysis using DiChroWeb (Lobley et al., 2002) using CONTILL algorithm (Provencher and Gloeckner, 1981; Sreerama and Woody, 2000) predicted that Np53 is 32% disordered, and upon incubation with RITA the level of disorder reduced to 28%.

Hydrogen–deuterium exchange coupled to mass spectrometry (HDX-MS) was used to ascertain if the conformational changes induced by RITA could be mapped in the solution phase. Np53 was incubated for varying time points in deuterated buffer, and the mass shift of peptides was determined. Np53 shows a significant uptake of deuterium at the shortest experimental time point of 15 s for a large proportion of peptides detected (**Figures 3A–D**). From the mas spectrometry data of each peptide, we observe no significant difference between deuterium uptake in the absence or presence of RITA, as shown by the deuterium uptake curves for selected representative peptides residues 23–30 and 53–63 (**Figures 3E,F**, respectively). This indicates that we cannot sample the interconverting solution conformations for this highly dynamic protein over the longer timescale of the HDX-MS experiment. The butterfly plot in **Figure 3G** depicts the overall deuterium uptake differences between Np53 in the absence and presence of RITA. Each set of points along the x-axis represent a peptide, with time points denoted by different colored points and lines [15 (yellow), 30 (red), 60 (blue), and 120 (black) s incubation time]. Gray bands indicate the error in the uptake level and vertical lines indicate the sum of uptake differences for each time point. Several peptides show deuterium uptake differences slightly above the error, but all at <1 Da, indicating that this protein is highly dynamic with or without RITA, for example, the greatest deuterium uptake difference in Np53 in the absence and presence of RITA being 0.271 Da, for a peptide spanning residues 23–39 with a $[M+2H]^{2+}$ charge state.

Modulation of N-terminal MDM2 by Nutlin-3

Mass Spectra for MDM2 (**Figure 4A**) sprayed from native conditions with 50 mM ammonium acetate and 0.5% DMSO show a broad bimodal CSD spanning charge states $5 \leq z \leq 14$. The most intense species is $[M+10H]^{10+}$ with significant intensity also in $[M+7H]^{7+}$ and $[M+6H]^{6+}$. We observe low intensity $[D+11H]^{11+}$, $[D+13H]^{13+}$, and $[D+15H]^{15+}$ dimers,

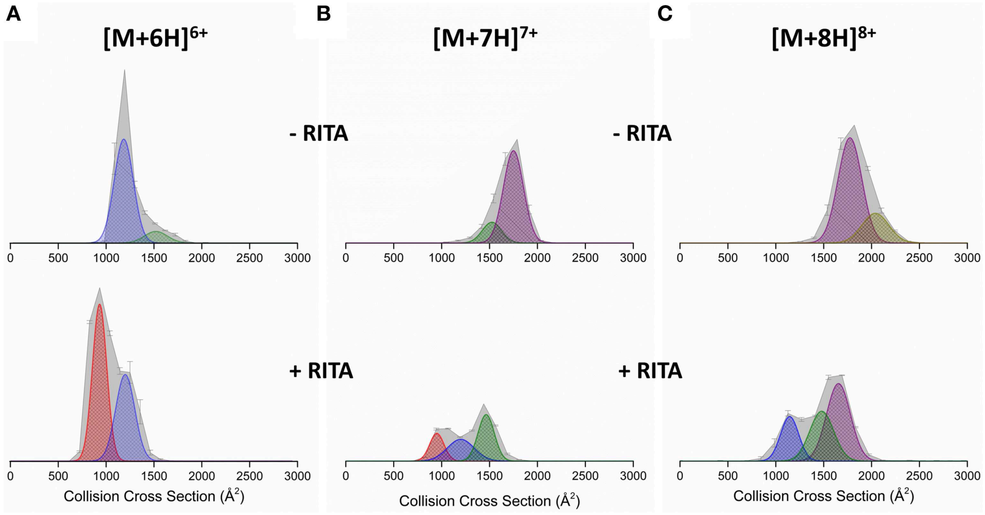


FIGURE 2 | Collision cross section distributions ($^{DT}CCSD_{He}$) arising from arrival time distributions (ATDs) at drift voltage 35 V for Np53 in the absence (top panels) and presence (bottom panels) of RITA. Distributions shown for the **(A)** $[M+6H]^{6+}$ **(B)** $[M+7H]^{7+}$, and **(C)** $[M+8H]^{8+}$ species. In the absence of RITA, Np53 was incubated with 5% IPA for 30 min at 37°C. Binding studies were carried out at a 1:2 Np53:RITA ratio incubated at 37°C for 30 min with 5% IPA. $^{DT}CCSD_{He}$ are normalized to the intensity of the ion peak in the corresponding mass spectrum. Conformational families are denoted by hatched Gaussian curves showing novel compact conformational family C_0 in red, compact conformational family C_1 in blue, extended conformational family X in green, and unfolded conformational family U in purple.

which means that the species attributed to $[M+5H]^{5+}$ will also contain some $[D+10H]^{10+}$ (and the $[M+6H]^{6+}$ some $[D+12H]^{12+}$ etc.) but since the flanking unique m/z dimers are of an intensity of <5% we ignore this contribution. Upon incubation with Nutlin-3 (**Figure 4B**) we see a CSD shift toward the lower charge states, with the $[M+6H]^{6+}$ species most intense, although the CSD range is retained. Binding of one Nutlin-3 molecule to MDM2 is observed at the $[M+5H]^{5+}$, $[M+6H]^{6+}$, and $[M+7H]^{7+}$ charge states. The shift in the N-MDM2 CSD upon incubation with Nutlin-3 is substantially greater than that caused by DMSO alone (Figure S4, Supporting Information) suggesting that Nutlin-3 confers a structural or conformational change in N-MDM2.

DT IM-MS analysis reveals that N-MDM2 in the absence of Nutlin-3 presents as at least two conformational families at all charge states (Figure S5, Supporting Information). The $[M+5H]^{5+}$ charge state (**Figure 5A**) presents as two conformers centered at ~1000 and ~1250 Å^2, referred to as C_1 (black Gaussian curve) and C_2 (red Gaussian curve), respectively. The $[M+6H]^{6+}$ charge state (**Figure 5B**) presents as three conformers, the compact C_1 and C_2 families and a more extended family, X (blue Gaussian curve) centered at ~1400 Å^2. The $[M+7H]^{7+}$ charge state (**Figure 5C**) exhibits conformational family C_1 and X and also presents as a large conformer, U (green Gaussian curve) centered at ~1700 Å^2. Upon binding to Nutlin-3, we see a change in the $^{DT}CCSD_{He}$ of N-MDM2 for each charge state. The $[M+5H]^{5+}$ charge state (**Figure 5A**, middle panel) shows retention of the compact conformer C_1, but a significant decrease in the intensity of C_2. The $[M+6H]^{6+}$ charge state, when bound to Nutlin-3, undergoes a compaction event to produce a single conformational family centered at ~1250 Å^2, corresponding to conformer C_2 (**Figure 5B**, middle panel). This effect is again seen for the $[M+7H]^{7+}$ charge state (**Figure 5C**, middle panel), which presents as a single conformer corresponding to conformer X when bound to Nutlin-3. These altered conformations remain, even when Nutlin-3 is not bound to N-MDM2. **Figure 5** bottom panels show the $^{DT}CCSD_{He}$ of N-MDM2 in the presence of Nutlin-3, but not bound in a complex. The $[M+5H]^{5+}$ species undergoes a minor change in the $^{DT}CCSD_{He}$ with an increase in conformational family C_2 compared with the bound complex. Charge states $[M+6H]^{6+}$ and $[M+7H]^{7+}$ remain in the single conformational families C_2 and X, respectively, even after the ligand is no longer bound.

Discussion

The MS spectra for Np53 in the absence of RITA (**Figure 1A**) corroborates previous reports of disorder for the N-terminus of p53 (Bell et al., 2002; Dawson et al., 2003), a broad charge state range $5 \leq z \leq 12$, indicative of a disordered system (Testa et al., 2011; Beveridge et al., 2014) with numerous residues available for protonation in solution. We are unable to preserve the binding of small molecule RITA to Np53, suggesting the binding is lower affinity than that reported previously ($K_D = 1.5$ nM, Issaeva et al., 2004) or that it proceeds principally by hydrophobic interactions that are significantly diminished in the absence of solvent, resulting in loss of ligand during desolvation. We observe a narrowing of the CSD for Np53 on ligand incubation, we also observe the isolated ligand (data not shown), which also supports our assertion of ligand dissociation during desolvation. This CSD shift toward lower charge states suggests conformational tightening induced by RITA, a hypothesis that is supported

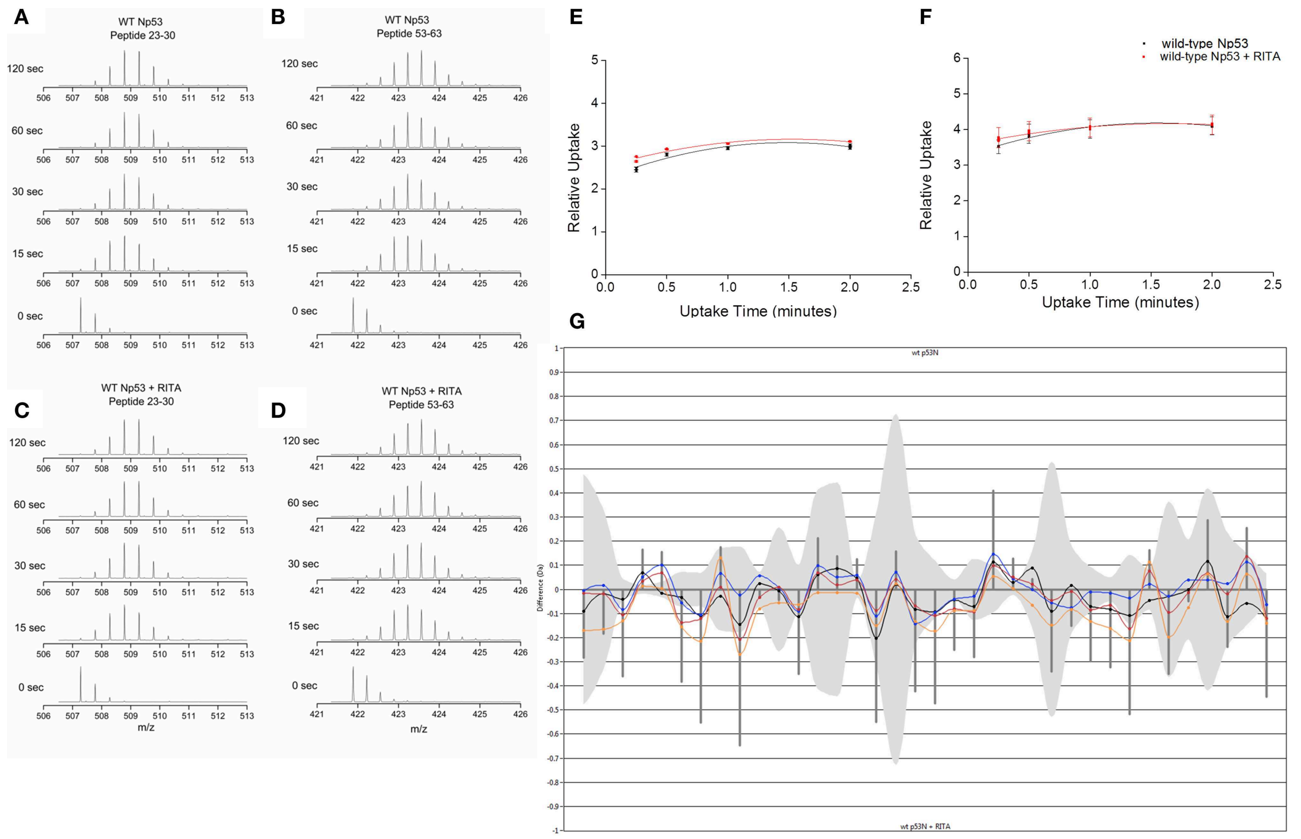


FIGURE 3 | (A,C) Deuterium uptake mass spectra for Np53 peptide residues 23–30 at time points 0, 15, 30, 60, and 120 s in the absence **(A)** and presence **(C)** of RITA. **(B,D)** Deuterium uptake charts for Np53 peptide residues 53–63 at time points 0, 15, 30, 60, and 120 s in the absence **(B)** and presence **(D)** of RITA. **(E,F)** Relative deuterium uptake plots for Np53 in the absence (black points) and presence (red points) of RITA for peptide residues 23–30 and 53–63, respectively. **(G)** Butterfly plot of uptake difference comparison for Np53 in the absence and presence of RITA. X-axis denotes each peptide detected, in ascending residue number. Y-axis denotes uptake difference (Daltons). Colored lines show uptake difference at 15 (yellow), 30 (red), 60 (blue), and 120 (black) s incubation time. Gray band indicates the error associated with each peptide. Vertical lines are the summed difference of all time points for each peptide.

by DT IM-MS data. The $^{DT}CCSD_{He}$ for Np53 is significantly altered in the presence of RITA at all charge states present, with loss of larger conformational families and induction of more compact conformers. We observe a compact conformer C_0 for $[M+5H]^{5+}$, $[M+6H]^{6+}$, and $[M+7H]^{7+}$ charge states, which is not present in the absence of RITA. Whilst the $[M+8H]^{8+}$species does not contain any of the C_0, it no longer contains conformer U, rather is populated by the more compact conformers C_1 and X, although conformer X is poorly resolved. $[M+9H]^{9+}$ undergoes loss of conformer U_2 with induction of compact conformer C_1 at a much lower $^{DT}CCS_{He}$. The use of IM-MS to discern conformational tightening due to ligand binding has been previously reported, (Harvey et al., 2012) and along with these findings provides an exciting prospect as a method for screening inhibitors to conformationally dynamic systems. As RITA is predicted to bind outside of the p53:MDM2 hydrophobic binding pocket (Issaeva et al., 2004), it has been asserted that the observed inhibition proceeds *via* a conformational change, which in turn will allosterically prevent the binding of MDM2. Our IM-MS data is evidence for the conformational modulation of Np53 by RITA. The induction of a smaller conformation is corroborated by CD results, which show an increase in secondary structure, and a decrease in the disordered content when analyzed using the CONTILL algorithm. We do note, however, that the calculated differences in structural content predictions for Np53 are minimal, with only a 4% decrease in disordered content, this is less informative than the clear conformational change provided by IM-MS.

The use of HDX-MS reinforces the view that Np53 is conformationally dynamic in solution; high levels of deuterium uptake are observed after 15 s incubation, with very little further uptake at longer incubation times. This suggests that backbone amides are solvent exposed and free to exchange with deuterium. When uptake was compared after RITA incubation we observe no significant changes in deuterium uptake for Np53 (**Figure 3G**). While there are several peptides which exhibit deuterium uptake differences outside of the error, the greatest difference is 0.271 Da for a $[M+2H]^{2+}$ peptide. As we see no difference greater than 1 Da, the mass difference between a hydrogen and deuterium atom, we can infer that there is no

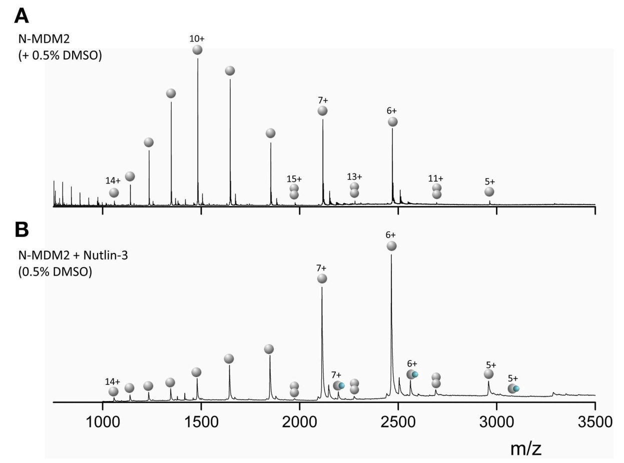


FIGURE 4 | nESI mass spectra of (A) 50 μM N-terminal MDM2 (residues 1–126) sprayed from 50 mM ammonium acetate + 0.5% DMSO. (B) 50 μM N-MDM2 + Nutlin-3 in a 1:10 protein:ligand ratio, incubated for 30 min at room temperature. Final DMSO concentration 0.5%. Monomeric species are denoted by single gray spheres, dimeric species by two gray spheres and N-MDM2 bound to Nutlin-3 by a single gray sphere and small blue sphere.

significant structural difference between Np53 in the absence and presence of RITA on the timescale of these experiments. Our shortest time step (15 s) is insufficient to observe the conformational changes occurring as the protein has enough time to rearrange back to its original conformations. In contrast, the isolated gas phase conformers exiting the electrosprayed droplets appear trapped in distinct conformers at least over the time scale of our IM-MS experiments. We estimate this time to be ~15 ms including the transmission of ions to our drift cell (McCullough et al., 2008), which is short enough to retain the conformational changes induced by RITA such that they can be observed. Both of the solution approaches *indicate* conformational flexibility and some slight change in structural content in the presence of the ligand, IM-MS provides a more definitive readout of the modulation of conformation to Np53 in the presence of RITA.

We can contrast the results observed for the RITA interaction with Np53 with that for the well-studied drug candidate Nutlin-3 with MDM2. N-MDM2 presents with a wide CSD ($5 \leq z \leq 14$), again suggesting a disordered protein. DT IM-MS shows that N-MDM2 presents in the gas-phase in at least two conformational families, potentially assignable to the previously reported “open” and “closed” position of the lid mini-domain (Uhrinova et al., 2005; Worrall et al., 2009, 2010). When incubated with Nutlin-3, we observe a substantial CSD shift toward the lower charge states which cannot be attributed to the effect of DMSO alone (Figure S4), again suggesting some conformational effect conferred by Nutlin-3 binding. We observe binding of Nutlin-3 to N-MDM2 over three charge states, $[M+5H]^{5+}$, $[M+6H]^{6+}$, and $[M+7H]^{7+}$. As there is no binding to the more extended high charge states, Nutlin-3 may only be able to bind N-MDM2 in a compact conformation, which is transferred to the gas phase as low charge state complex. DT IM-MS analysis showed that Nutlin-3 configures N-MDM2 into a more compact and inflexible conformer. The $[M+5H]^{5+}$ charge state retains both conformational families upon Nutlin-3 binding, however the larger conformer at $^{DT}CCS_{He}$ ~1250 Å^2 was greatly reduced. For the $[M+6H]^{6+}$ and $[M+7H]^{7+}$ ions, Nutlin-3 binding configures the protein into a single conformer with a narrow $^{DT}CCSD_{He}$, indicating less dynamics. This single conformational family was centered at a $^{DT}CCS_{He}$ ~1250 Å^2 for $[M+6H]^{6+}$, corresponding to conformational family C_2, and ~1400 Å^2 for $[M+7H]^{7+}$ corresponding to family X. We see loss of both the C_1 and X families for $[M+6H]^{6+}$ and loss of C_2 and U for $[M+7H]^{7+}$, suggesting much lower flexibility of the protein when bound to Nutlin-3.

Interestingly, it appears as for Np53, that the ligand free N-MDM2 in the IM-MS experiments also retains a “memory” of its in solution Nutlin-3 bound state. $^{DT}CCSD_{He}$ of N-MDM2, incubated with Nutlin-3 but in its apo-form, show similar conformers than those which retain binding of Nutlin-3 (**Figure 5**, bottom panels). This suggests that Nutlin-3 binds a higher proportion of analyte molecules than we observe, but is not retained fully during desolvation. The apo $[M+5H]^{5+}$ species is not only compact, suggesting that it rearranges back to the free N-MDM2 conformer, or that some of it arises from a conformer in solution that is incapable of binding Nutlin-3. The apo $[M+6H]^{6+}$ and $[M+7H]^{7+}$ remain in tight, single conformational families, and a much lower proportion of the Nutlin-bound N-MDM2 presents in the $[M+5H]^{5+}$ charge state (**Figure 4**) supporting our hypothesis that Nutlin-3 is unable to bind as well to the very compact conformer C_1. For the larger conformational families, N-MDM2 seems unable to rearrange back to its original conformations within the timescale of desolvation and analysis.

Conclusions

Multiple techniques have been used to probe the binding of small molecule inhibitors RITA and Nutlin-3 to N-terminal p53 (Np53) and N-terminal MDM2 (N-MDM2), respectively. Native mass spectrometry of Np53 shows a shift in the CSD toward the lower charge states and loss of the more extended charge states upon incubation with RITA. IM-MS of Np53 reveals two conformational families in the absence of RITA. Upon incubation with RITA, Np53 is configured into a novel, more compact conformer C_0 with loss of the more extended conformational family. We are able to retain this conformational tightening in the gas-phase on the time scale of our DT IM-MS experiments, even though we are unable to preserve the RITA:Np53 complex in the gas phase. HDX-MS data highlights the disordered nature of Np53, with no discernible conformations visible on a longer timescale. Very little differences are noted between the deuterium on-exchange of Np53 in the absence and presence of RITA, and we are unable to locate RITA induced conformational changes.

The nESI mass spectrum of N-terminal MDM2 shows a wide range of charge states ($5 \leq z \leq 14$) indicative of a disordered protein (Testa et al., 2011; Beveridge et al., 2014). The bimodal distribution suggests the protein may possess a more compact and more extended conformer. Indeed, DT IM-MS results show the protein presents as at least two conformational families at all

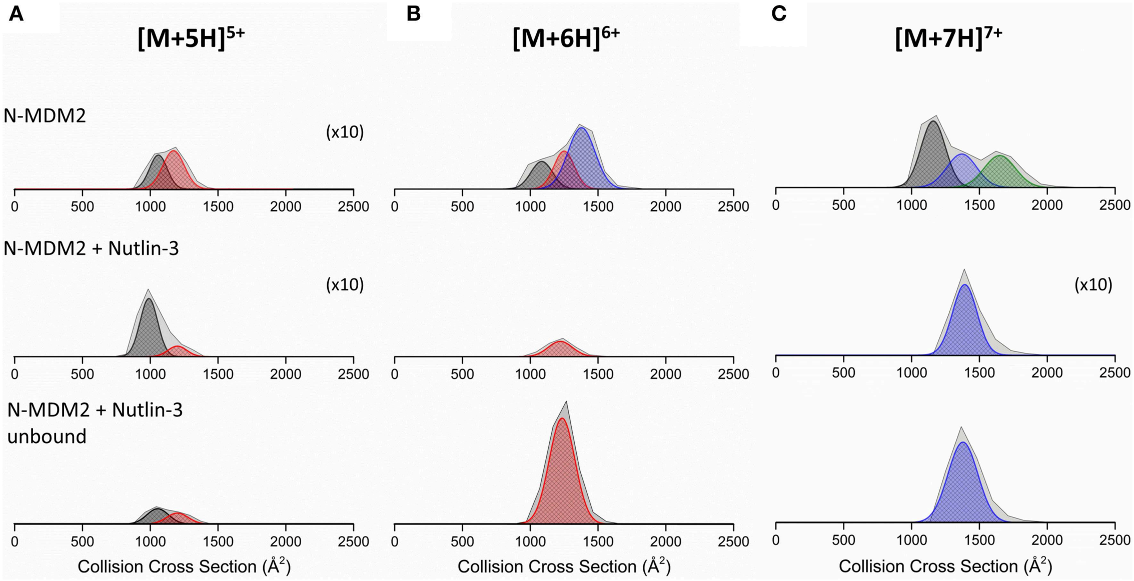


FIGURE 5 | Collision cross section distributions ($^{DT}CCSD_{He}$) derived from arrival time distributions (ATDs) for N-MDM2 sprayed from 50 mM ammonium acetate (+ 0.5% DMSO) for the (A) $[M+5H]^{5+}$, (B) $[M+6H]^{6+}$, and (C) $[M+7H]^{7+}$ charge states. Top panel represents N-MDM2, middle panel represents N-MDM2 bound to Nutlin-3 and bottom panel represents N-MDM2 incubated with Nutlin-3 but in apo-form with no small molecule bound. Proteins were incubated with Nutlin-3 in a 1:10 protein:ligand ratio, for 30 min at room temperature. CCSDs were taken at a drift voltage of 35 V and are normalized to the intensity of the ion species in the corresponding mass spectrum however to allow greater visibility, the $[M+5H]^{5+}$, $[M+5H]^{5+}$ bound to Nutlin-3 and $[M+7H]^{7+}$ bound to Nutlin-3 $^{DT}CCSD_{He}$ have been magnified X10. Conformational families are depicted by colored Gaussian curves; C_1 (black), C_2 (red), X (blue), and U (green).

charge states. Upon incubation with Nutlin-3, we observe ligand binding to the forms of the protein that present to the gas phase with low charge states $[M+5H]^{5+}$, $[M+6H]^{6+}$, and $[M+7H]^{7+}$, suggesting selective binding to a compact conformer of MDM2, or possibly that more extended forms lose Nutlin-3 on transfer to the gas phase. The bound species of MDM2 are compact at all three charge states, with $[M+6H]^{6+}$ and $[M+7H]^{7+}$ forming a single conformational family centered at a $^{DT}CCS_{He}$ of the middle conformational family exhibited by apo-N-MDM2. These conformational changes are likely retained by ions which lose the bound Nutlin-3 molecule during desolvation, indicating that the protein is unable to rearrange during the experiment. IM-MS is presented as a promising technique able to track conformational changes in unstructured proteins on a millisecond timescale.

Acknowledgments

This work has been funded by the award of a BBSRC (Biological and Biotechnological Science Research Council) Industrial Case Studentship to EJ in collaboration with Waters Corp. and by the BBSRC in awards BB/L015048/1 and BB/H013636/1. We are also grateful for the support of the Universities of Edinburgh and Manchester.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmolb.2015.00039

References

Allen, J. G., Bourbeau, M. P., Wohlhieter, G. E., Bartberger, M. D., Michelsen, K., Hungate, R., et al. (2009). Discovery and optimization of chromenotriazolopyrimidines as potent inhibitors of the mouse double minute 2-tumor protein 53 protein-protein interaction. *J. Med. Chem.* 52, 7044–7053. doi: 10.1021/jm900681h

Bakalkin, G., Selivanova, G., Yakovleva, T., Kiseleva, E., Kashuba, E., Magnusson, K. P., et al. (1995). p53 binds single-stranded DNA ends through the C-terminal domain and internal DNA segments via the middle domain. *Nucleic Acids Res.* 23, 362–369. doi: 10.1093/nar/23.3.362

Bell, S., Klein, C., Müller, L., Hansen, S., and Buchner, J. (2002). p53 contains large unstructured regions in its native state. *J. Mol. Biol.* 322, 917–927. doi: 10.1016/S0022-2836(02)00848-3

Bernstein, S. L., Liu, D., Wyttenbach, T., Bowers, M. T., Lee, J. C., Gray, H. B., et al. (2004). α-Synuclein: stable compact and extended monomeric structures and pH dependence of dimer formation. *J. Am. Soc. Mass Spectrom.* 15, 1435–1443. doi: 10.1016/j.jasms.2004.08.003

Beveridge, R., Covill, S., Pacholarz, K. J., Kalapothakis, J. M., Macphee, C. E., and Barran, P. E. (2014). A mass spectrometry based framework to define the extent of disorder in proteins. *Anal. Chem.* 86, 10979–10991. doi: 10.1021/ac5027435

Brown, C. J., Lain, S., Verma, C. S., Fersht, A. R., and Lane, D. P. (2009). Awakening guardian angels: drugging the p53 pathway. *Nat. Rev. Cancer* 9, 862–873. doi: 10.1038/nrc2763

Brown, C. J., Quah, S. T., Jong, J., Goh, A. M., Chiam, P. C., Khoo, K. H., et al. (2012). Stapled peptides with improved potency and specificity that activate p53. *ACS Chem. Biol.* 8, 506–512. doi: 10.1021/cb3005148

Chalmers, M. J., Busby, S. A., Pascal, B. D., He, Y., Hendrickson, C. L., Marshall, A. G., et al. (2006). Probing protein ligand interactions by automated hydrogen/deuterium exchange mass spectrometry. *Anal. Chem.* 78, 1005–1014. doi: 10.1021/ac051294f

Chang, Y. S., Graves, B., Guerlavais, V., Tovar, C., Packman, K., To, K.-H., et al. (2013). Stapled α-helical peptide drug development: a potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proc. Natl. Acad. Sci.U.S.A.* 110, E3445–E3454. doi: 10.1073/pnas.1303002110

Dawson, R., Müller, L., Dehner, A., Klein, C., Kessler, H., and Buchner, J. (2003). The N-terminal domain of p53 is natively unfolded. *J. Mol. Biol.* 332, 1131–1141. doi: 10.1016/j.jmb.2003.08.008

Ding, K., Lu, Y., Nikolovska-Coleska, Z., Wang, G., Qiu, S., Shangary, S., et al. (2006). Structure-based design of spiro-oxindoles as potent, specific small-molecule inhibitors of the MDM2-p53 interaction. *J. Med. Chem.* 49, 3432–3435. doi: 10.1021/jm051122a

Freedman, D. A., and Levine, A. J. (1998). Nuclear export is required for degradation of endogenous p53 by MDM2 and human papillomavirus E6. *Mol. Cell. Biol.* 18, 7288–7293.

Grasberger, B. L., Lu, T., Schubert, C., Parks, D. J., Carver, T. E., Koblish, H. K., et al. (2005). Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. *J. Med. Chem.* 48, 909–912. doi: 10.1021/jm049137g

Harvey, S. R., Porrini, M., Stachl, C., Macmillan, D., Zinzalla, G., and Barran, P. E. (2012). Small-molecule inhibition of c-MYC:MAX leucine zipper formation is revealed by ion mobility mass spectrometry. *J. Am. Chem. Soc.* 134, 19384–19392. doi: 10.1021/ja306519h

Haupt, Y., Maya, R., Kazaz, A., and Oren, M. (1997). Mdm2 promotes the rapid degradation of p53. *Nature* 387, 296–299. doi: 10.1038/387296a0

Hogg, A., and Kebarle, P. (1965). Mass-spectrometric study of ions at near-atmospheric pressure. II. Ammonium ions produced by the alpha radiolysis of ammonia and their solvation in the gas phase by ammonia and water molecules. *J. Chem. Phys.* 43, 449–456. doi: 10.1063/1.1696762

Issaeva, N., Bozko, P., Enge, M., Protopopova, M., Verhoef, L. G., Masucci, M., et al. (2004). Small molecule RITA binds to p53, blocks p53–HDM-2 interaction and activates p53 function in tumors. *Nat. Med.* 10, 1321–1328. doi: 10.1038/nm1146

Joerger, A. C., and Fersht, A. R. (2008). Structural biology of the tumor suppressor p53. *Annu. Rev. Biochem.* 77, 557–582. doi: 10.1146/annurev.biochem.77.060806.091238

Kebarle, P., and Hogg, A. (1965). Mass-spectrometric study of ions at near atmospheric pressures. I. The ionic polymerization of ethylene. *J. Chem. Phys.* 42, 668–674. doi: 10.1063/1.1695987

Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J., et al. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* 274, 948–953. doi: 10.1126/science.274.5289.948

Lobley, A., Whitmore, L., and Wallace, B. (2002). DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics* 18, 211–212. doi: 10.1093/bioinformatics/18.1.211

Lu, F., Chi, S. W., Kim, D. H., Han, K. H., Kuntz, I. D., and Guy, R. K. (2006). Proteomimetic libraries: design, synthesis, and evaluation of p53-MDM2 interaction inhibitors. *J. Comb. Chem.* 8, 315–325. doi: 10.1021/cc050142v

Mason, E. A., and McDaniel, E. W. (2005). *Transport Properties of Ions in Gases.* Berlin: Wiley-VCH Verlag GmbH & Co. KGaA.

McAfee, K. Jr., and Edelson, D. (1963). Identification and mobility of ions in a Townsend discharge by time-resolved mass spectrometry. *Proc. Phys. Soc.* 81, 382. doi: 10.1088/0370-1328/81/2/125

McCullough, B. J., Kalapothakis, J., Eastwood, H., Kemper, P., Macmillan, D., Taylor, K., et al. (2008). Development of an ion mobility quadrupole time of flight mass spectrometer. *Anal. Chem.* 80, 6336–6344. doi: 10.1021/ac800651b

Pagel, K., Natan, E., Hall, Z., Fersht, A. R., and Robinson, C. V. (2013). Intrinsically disordered p53 and its complexes populate compact conformations in the gas phase. *Angew. Chem. Int. Ed.* 52, 361–365. doi: 10.1002/anie.201203047

Provencher, S. W., and Gloeckner, J. (1981). Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* 20, 33–37. doi: 10.1021/bi00504a006

Rew, Y., Sun, D., Gonzalez-Lopez De Turiso, F., Bartberger, M. D., Beck, H. P., Canon, J., et al. (2012). Structure-based design of novel inhibitors of the MDM2–p53 interaction. *J. Med. Chem.* 55, 4936–4954. doi: 10.1021/jm300354j

Sreerama, N., and Woody, R. W. (2000). Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal. Biochem.* 287, 252–260. doi: 10.1006/abio.2000.4880

Szekely, L., Selivanova, G., Magnusson, K. P., Klein, G., and Wiman, K. G. (1993). EBNA-5, an Epstein-Barr virus-encoded nuclear antigen, binds to the retinoblastoma and p53 proteins. *Proc. Natl. Acad. Sci. U.S.A.* 90, 5455–5459. doi: 10.1073/pnas.90.12.5455

Tao, W., and Levine, A. J. (1999). Nucleocytoplasmic shuttling of oncoprotein Hdm2 is required for Hdm2-mediated degradation of p53. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3077–3080. doi: 10.1073/pnas.96.6.3077

Testa, L., Brocca, S., Šamalikova, M., Santambrogio, C., Alberghina, L., and Grandori, R. (2011). Electrospray ionization-mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein. *Biotechnol. J.* 6, 96–100. doi: 10.1002/biot.201000253

Tovar, C., Rosinski, J., Filipovic, Z., Higgins, B., Kolinsky, K., Hilton, H., et al. (2006). Small-molecule MDM2 antagonists reveal aberrant p53 signaling in cancer: implications for therapy. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1888–1893. doi: 10.1073/pnas.0507493103

Uhrinova, S., Uhrin, D., Powers, H., Watt, K., Zheleva, D., Fischer, P., et al. (2005). Structure of free MDM2 N-terminal domain reveals conformational adjustments that accompany p53-binding. *J. Mol. Biol.* 350, 587–598. doi: 10.1016/j.jmb.2005.05.010

Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., et al. (2004). *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* 303, 844–848. doi: 10.1126/science.1092472

Vousden, K. H. (2000). p53: death star. *Cell* 103, 691–694. doi: 10.1016/S0092-8674(00)00171-9

Vousden, K. H., and Prives, C. (2009). Blinded by the light: the growing complexity of p53. *Cell* 137, 413–431. doi: 10.1016/j.cell.2009.04.037

Vu, B., Wovkulich, P., Pizzolato, G., Lovey, A., Ding, Q., Jiang, N., et al. (2013). Discovery of RG7112: A small-molecule MDM2 inhibitor in clinical development. *ACS Med. Chem. Lett.* 4, 466–469. doi: 10.1021/ml4000657

Worrall, E. G., Worrall, L., Blackburn, E., Walkinshaw, M., and Hupp, T. R. (2010). The effects of phosphomimetic lid mutation on the thermostability of the N-terminal domain of MDM2. *J. Mol. Biol.* 398, 414–428. doi: 10.1016/j.jmb.2010.03.023

Worrall, E., Wawrzynow, B., Worrall, L., Walkinshaw, M., Ball, K., and Hupp, T. (2009). Regulation of the E3 ubiquitin ligase activity of MDM2 by an N-terminal pseudo-substrate motif. *J. Chem. Biol.* 2, 113–129. doi: 10.1007/s12154-009-0019-5

Wu, X., Bayle, J. H., Olson, D., and Levine, A. J. (1993). The p53-mdm-2 autoregulatory feedback loop. *Genes Dev.* 7, 1126–1132. doi: 10.1101/gad.7.7a.1126

Wyttenbach, T., Kemper, P., and Bowers, M. T. (2001). Design of a new electrospray ion mobility mass spectrometer. *Int. J. Mass Spectrom.* 212, 13–23. doi: 10.1016/S1387-3806(01)00517-6

Yin, H., Lee, G.-I., Park, H. S., Payne, G. A., Rodriguez, J. M., Sebti, S. M., et al. (2005). Terphenyl-based helical mimetics that disrupt the p53/HDM2 interaction. *Angew. Chem. Int. Ed.* 44, 2704–2707. doi: 10.1002/anie.200462316

Zhang, H.-M., Bou-Assaf, G. M., Emmett, M. R., and Marshall, A. G. (2009). Fast reversed-phase liquid chromatography to reduce back exchange and increase throughput in H/D exchange monitored by FT-ICR mass spectrometry. *J. Am. Soc. Mass Spectrom.* 20, 520–524. doi: 10.1016/j.jasms.2008.11.010

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

MINI REVIEW
published: 19 May 2015
doi: 10.3389/fmolb.2015.00021

Exploring intrinsically disordered proteins using site-directed spin labeling electron paramagnetic resonance spectroscopy

Nolwenn Le Breton, Marlène Martinho, Elisabetta Mileo, Emilien Etienne, Guillaume Gerbaud, Bruno Guigliarelli and Valérie Belle *

Bioénergétique et Ingénierie des Protéines Laboratory, UMR 7281, Aix-Marseille Université and Centre National de la Recherche Scientifique, Marseille, France

OPEN ACCESS

Edited by:
Peter Tompa,
Flanders Institute of Biotechnology (VIB), Belgium

Reviewed by:
Patrick Senet,
Université de Bourgogne, France
Steffen P. Graether,
University of Guelph, Canada

****Correspondence:***
Valérie Belle,
Bioénergétique et Ingénierie des Protéines Laboratory UMR 7281, Aix-Marseille Université and Centre National de la Recherche Scientifique, 31 chemin J. Aiguier, 13402 Marseille, France
belle@imm.cnrs.fr

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

Received: *26 March 2015*
Accepted: *03 May 2015*
Published: *19 May 2015*

Citation:
Le Breton N, Martinho M, Mileo E, Etienne E, Gerbaud G, Guigliarelli B and Belle V (2015) Exploring intrinsically disordered proteins using site-directed spin labeling electron paramagnetic resonance spectroscopy. Front. Mol. Biosci. 2:21. doi: 10.3389/fmolb.2015.00021

Proteins are highly variable biological systems, not only in their structures but also in their dynamics. The most extreme example of dynamics is encountered within the family of Intrinsically Disordered Proteins (IDPs), which are proteins lacking a well-defined 3D structure under physiological conditions. Among the biophysical techniques well-suited to study such highly flexible proteins, Site-Directed Spin Labeling combined with EPR spectroscopy (SDSL-EPR) is one of the most powerful, being able to reveal, at the residue level, structural transitions such as folding events. SDSL-EPR is based on selective grafting of a paramagnetic label on the protein under study and is limited neither by the size nor by the complexity of the system. The objective of this mini-review is to describe the basic strategy of SDSL-EPR and to illustrate how it can be successfully applied to characterize the structural behavior of IDPs. Recent developments aimed at enlarging the panoply of SDSL-EPR approaches are presented in particular newly synthesized spin labels that allow the limitations of the classical ones to be overcome. The potentialities of these new spin labels will be demonstrated on different examples of IDPs.

Keywords: intrinsically disordered proteins, structural transitions, induced folding, fuzzy complex, nitroxide spin labels, site-directed spin labeling EPR spectroscopy

Introduction

Characterizing dynamics of macromolecules is a complex task requiring the use of appropriate techniques. In parallel to the development of methods leading to structure resolution of proteins, there is also an increasing need to develop biophysical techniques able to describe structural flexibility, as this dynamical aspect is closely related to protein function (Dyson and Wright, 2002; Salmon et al., 2010). The most highly dynamic biological systems are the so-called Intrinsically Disordered Proteins (IDPs) or Regions (IDRs), which lack a well-defined 3D structure under physiological conditions while being associated to key functions such as regulation, molecular assembly, signaling…(for recent reviews see Uversky and Dunker, 2010; Babu et al., 2011; Chouard, 2011; Habchi et al., 2014). Among the various techniques able to give access to dynamic properties of biomolecules is Site-Directed Spin Labeling combined with Electron Paramagnetic Resonance (SDSL-EPR), a technique that was pioneered about 20 years ago by (Hubbell et al., 1996). SDSL-EPR is very sensitive for probing flexible regions of proteins and revealing dynamic changes or interaction sites in protein-protein interactions

(Belle et al., 2007; Lorenzi et al., 2012). It can also be used to determine accessibility profiles, a powerful approach to determine the topology of membrane proteins (Gross et al., 1999; Kaplan et al., 2000). More recently, the development of pulse EPR, in particular Double Electron Electron Resonance (DEER) techniques allowed the measurement of distances between spin labeled sites in the range of 1.8–6.0 nm (Pannier et al., 2000), thus covering a wide range of interest for the study of large conformational transitions and biomolecule associations. Excellent reviews describing all of these approaches, as well as applications of SDSL-EPR on proteins, have been recently published (Klare and Steinhoff, 2009; Bordignon, 2011; Mchaourab et al., 2011; Drescher, 2012; Hubbell et al., 2013). The present contribution will focus on dynamic analyses of extremely flexible biological systems and recent synthesis of new spin labels designed to enlarge the potentialities of the technique.

General Principles and Development of New Spin Labels

SDSL-EPR relies on selective grafting of a paramagnetic species, usually a nitroxide derivative, at a desired position of a protein and subsequent analysis by EPR spectroscopy. Nitroxides are stable radicals in which the unpaired electron is delocalized on the N-O group, leading to a 3-line spectrum arising from the hyperfine interaction between the electron spin and the nuclear spin of the ^{14}N atom. Nitroxide spin labels are classically functionalized to react specifically with cysteine residues. The technique thus requires the construction of cysteine mutants in order to target specific sites. The most frequently used nitroxide spin label is the MTSL (1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl) methanethiosulfonate leading to the formation of a disulphide bridge between the side-chain of the cysteine and the label (**Figure 1A**, side-chain R_1). Its relatively small size minimizes the potential perturbation of the biological system. Other commercial spin labels are available, such as the maleimide-functionalized ones having the advantage of forming a thio-ether bond with the protein side-chain preventing the label release in the presence of reducing agents. One example of such spin label is the 3-maleimido-2,2,5,5-tetramethyl-pyrrolidinyloxy referred to as Proxyl or P depicted in **Figure 1A** (spin label N_2). These labels are, however, more sterically demanding than MTSL, and they can react with amines at high pH (Hideg et al., 2005). As for all labeling techniques, control experiments are essential for comparing wild-type protein and labeled mutants to check the non-invasiveness of the label with respect to protein structural and functional properties. By investigating the relationship between the mobility of the MTSL nitroxide side-chain and the structural elements of the T4 lysozyme taken as a model protein, Mchaourab and co-workers established the basis for the interpretation of EPR spectral shapes of spin labels (Mchaourab et al., 1996). The power of the technique relies on the sensitivity of the EPR spectral shapes to the mobility of the label in the nanosecond time window described by the rotational correlation time τ_c (**Figure 1B**), a parameter that can be obtained by spectral simulation. Indeed the magnetic hyperfine interaction between the electron spin and the ^{14}N nucleus is highly anisotropic. This anisotropy is fully averaged when the radical is highly mobile and the spectrum displays three narrow lines. When the mobility decreases, the averaging becomes partial and the lines broaden progressively until reaching a limit of a fully anisotropic spectrum corresponding to an immobilized spin label. A spectral modification thus represents a change in the environment of the label affecting its mobility and thus indicates a structural transition (**Figure 1C**).

The technique has still some limitations, in particular due to the chemical nature of the spin labels, which our recent works aim to overcome. One limitation comes from the poor diversity of EPR spectral shapes of nitroxide labels. The similarity of the 3-line spectral shapes precludes the simultaneous investigation of two different regions of a protein or two interacting proteins, a situation that can be encountered in allosteric mechanism in which ligand binding at one site influences binding at another site through a propagated structural change within the protein. To overcome this limitation, we designed and synthesized a new spin label based on a maleimido-functionalized β-phosphorylated nitroxide: the {2-(diethoxyphosphoryl)-5-[(2,5-dioxo-2,5-dihydro-1H-pyrrol-1-yl)methyl]-2,5 dimeth ylpyrrolidin-1-yl}oxidanyl, referred to as Phosphorylated Proxyl or PP (**Figure 1A**, N_3). Thanks to the supplementary high magnetic coupling between the electron spin and the nuclear spin of ^{31}P, this label gives a well-resolved 6-line spectrum composed of a doublet of triplets (Le Breton et al., 2014). Another limitation of conventional SDSL concerns the fact that only cysteines are specifically targeted by commercial nitroxides spin labels which is unsuitable when cysteines play roles either in the function and/or in structural elements (active sites, disulfide bridges) of the protein under study. We designed and synthesized various nitroxides to react with the phenol group of the tyrosine *via* a three-components Mannich type reaction (Lorenzi et al., 2011; Mileo et al., 2013a). The best results were obtained with an isoindoline-based nitroxide: 5-amino-1,1,3,3-tetramethyl-isoindolin-2-yloxyl, referred to as Nox (**Figure 1A**, N_4), which has proved to be a good reporter of structural modifications (Mileo et al., 2013a). The successful applications of these two new nitroxides to report on structural transitions on IDPs will be presented in the next section.

Applications

Several IDPs have already been investigated by SDSL-EPR. This approach has been successfully used to reveal various structural states and higher-order organizations of flexible proteins involved in neurodegenerative diseases such as α-synuclein (Chen et al., 2007), amyloid-β peptide (Torok et al., 2002) and tau (Siddiqua and Margittai, 2010). Another example of IDP studied by SDSL-EPR is the small acid protein IA3 that acts as an inhibitor of the yeast proteinase A (Casey et al., 2014). In the following, results with two highly flexible proteins are reviewed to illustrate the potential of SDSL-EPR in the field of IDPs.

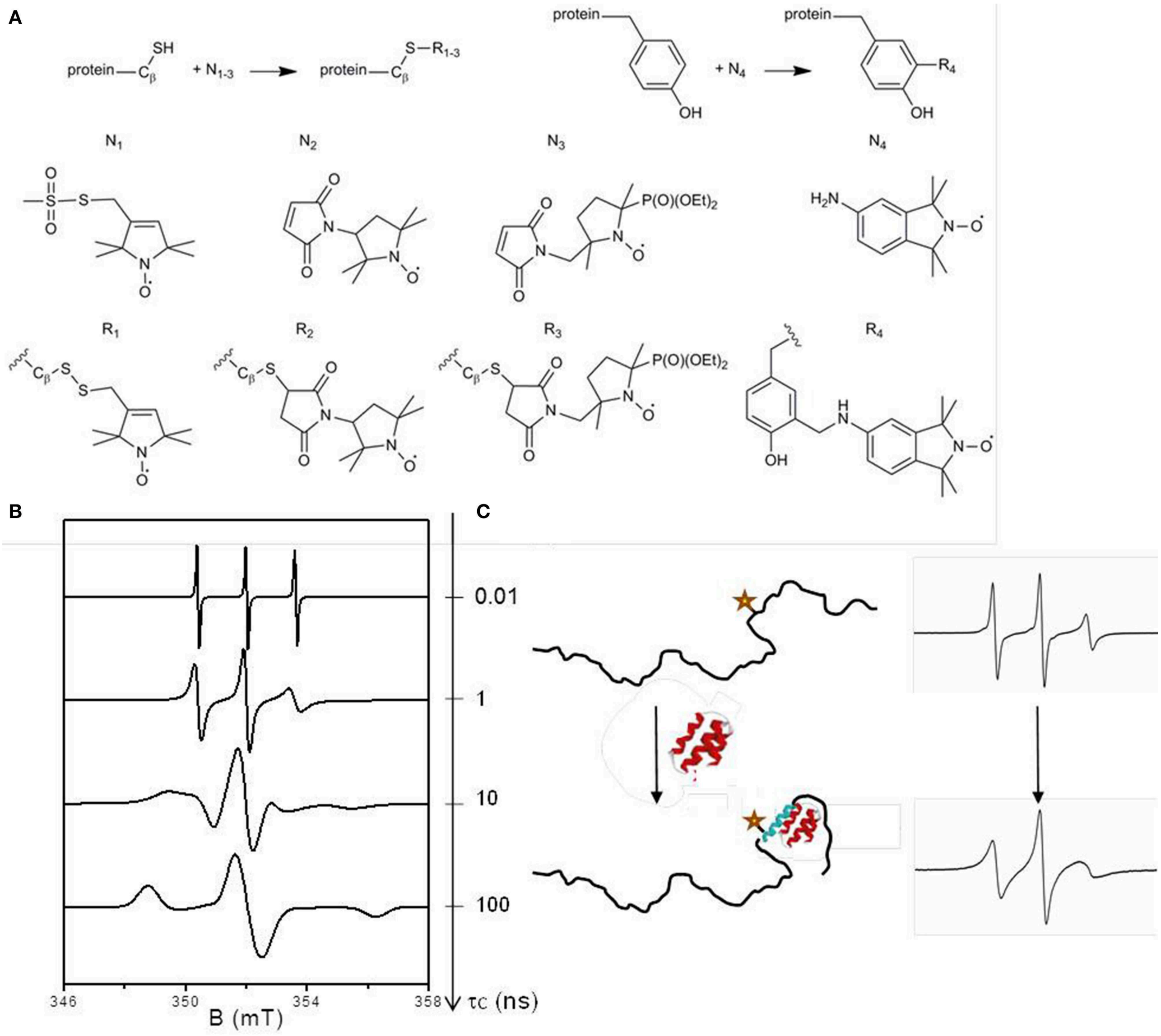


FIGURE 1 | (A) Labeling reaction scheme for cysteine and tyrosine residues. N_{1-3} are spin labels for grafting on cysteine residues: N_1 is the methanethiosulfonate spin label (MTSL), N_2 the maleimido proxyl (P) and N_3 the phosphorylated maleimido proxyl (PP). N_4 is the an isoindoline-based nitroxide (Nox) dedicated to tyrosine labeling. R_{1-4} are the corresponding spin label side-chains. **(B)** EPR spectral shape modifications as a function of the mobility of the spin label described by its rotational correlation time τ_C. The spectra were simulated using the EasySpin software (Stoll and Schweiger, 2006) for different values of τ_C: 0.01, 1, 10, and 100 ns. **(C)** Illustration of an EPR spectral shape broadening in the case of a disorder-to-order transition due to an induced folding mechanism. The star represents the spin label.

Cartography of Induced Folding

Most IDPs undergo an induced folding in presence of a partner protein i.e., a disorder-to-order transition that can be limited to a particular region. An illustrative example concerns the cartography of induced folding of nucleoproteins (N) from three viruses belonging to the *Paramyxoviridae* family, namely Measles (MeV), Nipah (NiV) and Hendra (HeV) viruses. Multiple copies of N are structurally organized to encapsidate the viral genome. The particular case of MeV nucleoprotein has been the most extensively studied by complementary biophysical approaches (Habchi and Longhi, 2012). MeV N consists of two regions: a N-terminal globular one (aa 1–400) and a C-terminal domain N_{TAIL} (aa 401–525) that is fully disordered. This disordered part is essential for transcription and replication of the virus *via* interaction with the phosphoprotein of the viral polymerase complex. Our study was focused on the interaction between MeV N_{TAIL} and the C-terminal part (X Domain) of the phosphoprotein P_{XD} (aa 459–507 of P), which is constituted of 3 α-helices (Longhi et al., 2003). In order to precisely localize the regional folding that MeV N_{TAIL} undergoes in the

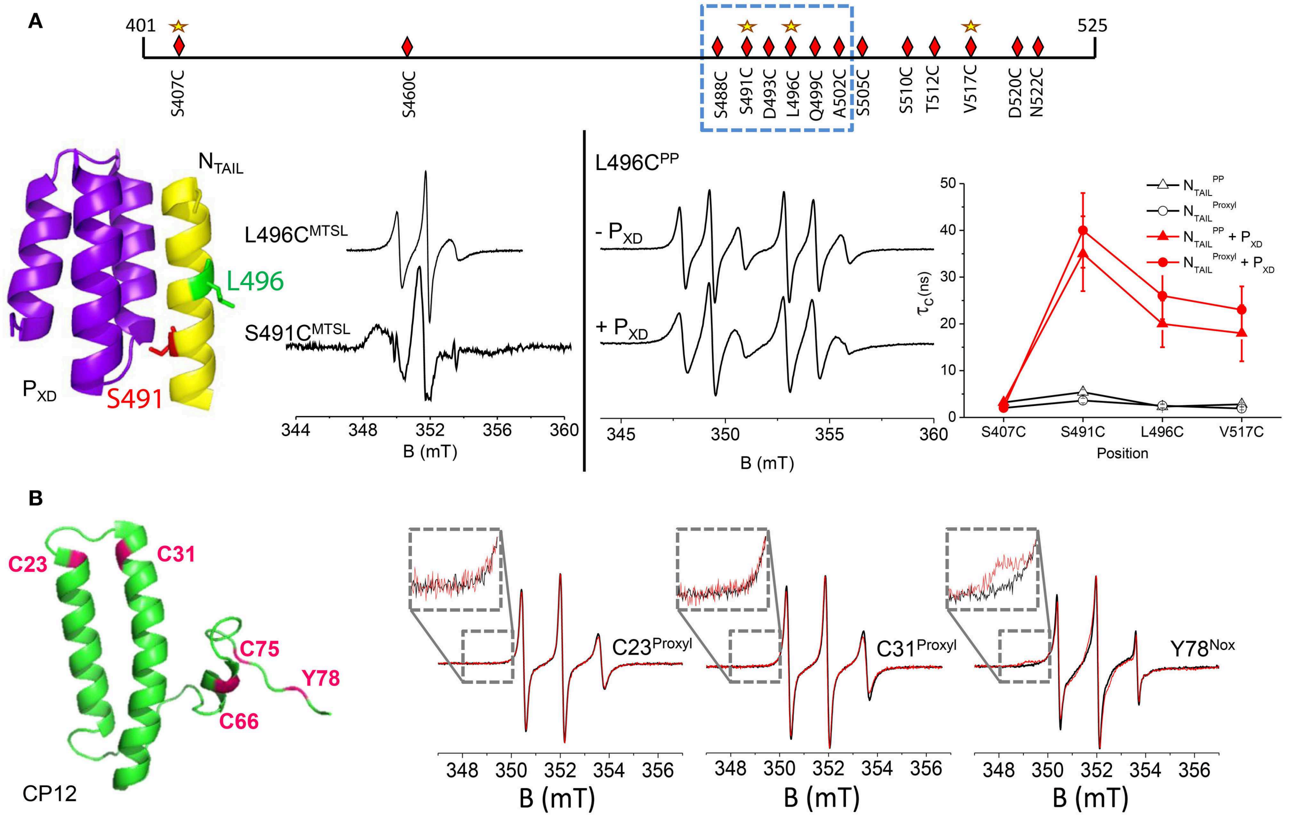


FIGURE 2 | (A) upper panel: schematic representation of positions targeted for cysteine substitution and spin labeling of N_{TAIL} (aa 401–525) with MSTL spin label (diamonds) and PP spin label (stars). The dotted frame indicates the region that undergoes an induced folding in the presence of the partner protein P_{XD}. Left panel: Illustration of two EPR spectral shapes obtained in two positions of the MTSL within the induced folding region of N_{TAIL}: positions 491 and 496. These two positions are highlighted in the crystal structure of the chimera construct between P_{XD} and the N_{TAIL} region encompassing residues 486–504 (pdb code 1T60). Right panel: EPR spectra of the phosphorylated proxyl grafted at position 496 in the absence and in the presence of P_{XD}. Variation of the rotational correlation time τ_C of Proxyl (circles) and Phosphorylated Proxyl PP (triangles) spin labeled N_{TAIL} variants without (black open symbols) and with (red filled symbols) saturating amounts of P_{XD} as a function of spin label position. τ_C values have been obtained by simulating the EPR spectra using the program ROKY (Rockenbauer and Korecz, 1996). **(B)** Left panel: 3D structural model of *C. reinhardtii* CP12 (pdb 2DDN) in which the positions of the four cysteines and the unique tyrosine are highlighted in pink. Right panel: Superimposition of amplitude-normalized EPR spectra of CP12 C23^{Proxyl}, C31^{Proxyl}, and Y78^{Nox} in absence (black) and presence of GAPDH in equimolar ratio (red). In inset: a zoom on the low-field region.

presence of P_{XD}, we targeted 14 sites for spin-labeling (with MTSL) within N_{TAIL}, 12 of which being concentrated in the C-terminal region (aa 488–525) that was known to be specifically involved in the interaction with MeV P_{XD} (Longhi et al., 2003) (**Figure 2A**, upper panel). For all labeled variants of N_{TAIL}, room temperature EPR spectra were recorded in the absence and presence of P_{XD} (Belle et al., 2008). An important decrease of the mobility of the spin labels was observed in the 488–502 region that we have been able to attribute to the formation of an α-helix using complementary circular dichroism analyses. At one particular position (aa 491) a broad shape component was detected, indicating a very restricted environment of the spin label (**Figure 2A**, left panel). This observation allowed us to confirm the structural model of a chimera construct between MeV P_{XD} and a small N_{TAIL} region (aa 486–504) in which the amino acid side-chain 491 points toward P_{XD} whereas the other ones are solvent-exposed (Kingston et al., 2004). Interestingly, the mobility of the induced folding region was found to be slightly but significantly restrained even in the absence of the partner protein, a behavior that could indicate the existence of a pre-structuration of this region. This observation has been further confirmed combining SDSL-EPR data and modeling of local rotation conformational space (Kavalenka et al., 2010) and also by NMR studies and modeling of this region as a dynamic equilibrium between a completely unfolded state and different partially helical conformations (Jensen et al., 2011). Concerning the C-terminal end of N_{TAIL}, EPR spectral shapes of the labels grafted in the 505–525 region showed a moderate decrease of mobility that has been attributed to a gain of rigidity arising from α-helical folding of the neighboring 488–502 region (Belle et al., 2008). This observation was consistent with further analyses where the 505–525 region was found to conserve a significant degree of freedom even in the bound form (Kavalenka et al., 2010). Using the same strategy based on multiple individual labeling sites, we also mapped the induced folding of N_{TAIL} in association with P_{XD} for HeV and NiV viruses (Martinho et al., 2013) and validated previously proposed structural models obtained by homology modeling (Habchi et al., 2011).

Thanks to the well-characterized MeV N_{TAIL}-P_{XD} interaction by conventional spin labels, this biological system was used as a model to characterize the new label PP, having a ^{31}P atom in the vicinity of the nitroxide leading to a 6-line spectrum (**Figure 1A**, N_3) and to probe its ability to report structural transitions. Four grafting sites on N_{TAIL} were judiciously chosen (within and outside the induced folding region) (**Figure 2A**, upper panel). For comparative purposes, proxyl P (**Figure 1A**, side-chain R_2) was grafted at the same positions. The ability of PP to report on structural transitions was evaluated by analyzing the spectral shape modifications induced either by a secondary structure stabilizer (trifluoroethanol) or by the presence of the partner protein P_{XD} (**Figure 2A**, right panel). All the spectra were simulated to extract the dynamic parameter τ_c that represents the mobility of the spin labels. The modification of this parameter according to the position of the spin label and the condition (free or bound to P_{XD}) was very similar for both the classical proxyl and the new phosphorylated spin labels (**Figure 2B**, right panel). Taken together the results demonstrated that PP is able to monitor from subtle to larger structural transitions, as efficiently as the classical spin label. Molecular dynamics (MD) calculations were performed to gain further insights into the binding process between the labeled N_{TAIL} and P_{XD}. MD calculations revealed that the new phosphorylated label does not perturb the interaction between the two partner proteins and reinforced the conclusion on its ability to probe different local environments in a protein (Le Breton et al., 2014). Thanks to its new EPR spectral signature, the combination of PP with classical spin labels opens the way to study two protein sites simultaneously.

Revealing Fuzziness in a Supramolecular Complex

In this section we will show how SDSL-EPR provides a unique tool to reveal regions of an IDP remaining highly flexible after complex formation, information that often escapes to detection by other biostructural techniques. This study concerns the structural flexibility of a small chloroplastic protein called CP12 (80 aa) and its association with the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) from the green alga *Chlamydomonas reinhardtii*. This association is a key step in the formation of a ternary supramolecular complex involved in the regulation of the Calvin cycle in many photosynthetic organisms (for details see the review Tieulin-Pardo et al in the present journal). In its oxidized state the green alga CP12 contains four cysteine residues engaged in two disulfide bridges (C23–C31 and C66–C75) and presents some α-helical structural elements modeled from each side of the N-terminal disulfide bridge, whereas the C-terminal part appears mainly disordered (**Figure 2B**) (Gardebien et al., 2006). In contrast, in its reduced state the algal CP12 is fully disordered (Graciet et al., 2003). In a first study, we used MTSL (**Figure 1A**, N_1), which led to the labeling of the only C-terminal cysteine residues leaving the N-terminal unmodified as revealed by mass spectrometry analyses. Surprisingly, the partner protein GAPDH induced the cleavage of the disulfide bridge between the cysteine and the label, resulting in the full release of the label. This result showed the existence of a transitory interaction between both proteins and we proposed a mechanism based on a thiol-disulfide exchange reaction involving cysteines C21 and C291 of the *C. reinhardtii* GAPDH (Erales et al., 2009). Even if this observation led us to propose a new role for the algal GAPDH, it however doesn't allow us to study the GAPDH-CP12 complex. As the central region of CP12 has been proposed to be involved in the interaction with GAPDH, we used two cysteine-to-serine mutants (C23S and C31S) that were known to be fully disordered, due to the loss of the N-terminal disulfide bridge, but still able to interact with GAPDH (Lebreton et al., 2006). In order to individually target the two N-terminal cysteine residues, proxyl P was chosen as spin label (**Figure 1A**, N_2) to prevent label release induced by GAPDH. EPR analyses revealed that this part of the protein remains fully disordered after association with GAPDH as no spectral modification was detected (**Figure 2B**) (Mileo et al., 2013b). The association between the two partner proteins was checked by probing the accessibility of the labels using a reducing agent. Indeed, in the absence of GAPDH, spin labels grafted on CP12 are reduced by an excess of DTT and become EPR-silent with a characteristic time of about 25 min. On the contrary, if GAPDH is present in the sample, the same amount of DTT has almost no effect as the EPR signal remains stable for at least 2 h confirming that the association really takes place and that CP12 remains highly dynamic in its bound form (Mileo et al., 2013b).

The C-terminal region has been demonstrated to be mainly responsible for the redox regulation of GAPDH (Lebreton et al., 2006). To gain further insights into the dynamics of this region of CP12 while keeping the 2 disulfide bridges intact, tyrosine 78 was chosen as an alternative labeling site. The location of tyrosine 78, highly conserved in CP12s from different organisms, is particularly interesting as it is close to residues that play a crucial role in the activity modulation of GAPDH either in the binary complex (Erales et al., 2011) or in the ternary complex with PRK (Avilan et al., 2012). The new isoindoline-based nitroxide Nox (**Figure 1A**, N4) was used to selectively target this unique tyrosine. The presence of GAPDH led to a slight modification of the EPR spectral shape (**Figure 2B**) indicating that the interaction of the C-tail with the partner protein is not tight. This result showed however that this region is close to the interaction site without being directly involved in it (Mileo et al., 2013b). All together the analyses of the three labeling sites were in good agreement with the partial view of the CP12-GAPDH complexes from other organisms given by recent crystallographic and NMR studies where only the 20 last amino acids of CP12 were detected (Matsumura et al., 2011; Fermani et al., 2012). Finally, the use of different spin labels is this peculiar biological system allowed us to conclude that GAPDH-CP12 from *C. reinhardtii* is a new example of a fuzzy complex, a concept introduced by Tompa and Fuxreiter (2008), in which the IDP keeps most of its disorder and dynamics upon complex formation. This fuzziness could be one of the keys to facilitate the formation of a ternary complex (CP12, GAPDH and phosphoribulokinase, PRK) as well as functional actions of the machinery which necessitate dynamic assembly and disassembly processes controlled by dark/light transitions.

Conclusion

The examples described above, along with numerous other studies, illustrate the power of SDSL-EPR to access information on protein dynamics. Characterized by a remarkable conformational flexibility, IDPs form a unique protein category that is particularly suited for SDSL-EPR applications. SDSL-EPR is a rapidly growing field, in particular with recent developments focused on the detection of labeled protein in intact cells, an area that promises to be very interesting for IDPs applications. We can anticipate that these recent developments will create further need for the design of new labels with greater stability toward bio-reduction, an environment that is encountered inside the cell.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (ANR SPINFOLD n°09-BLAN-0100), the Centre National de la Recherche Scientifique (CNRS) and Aix-Marseille Université. The authors are grateful to the EPR facilities available at the national TGE RPE facilities (IR 3443). We thank our collaborators and their teams: R. Lebrun (Proteomic center of IMM, Marseille, France), S. Longhi (AFMB, CNRS-AMU UMR 7257, Marseille, France), J. Golebiowsky (ICN, CNRS Université de Nice UMR 7272, Nice, France), B. Gontero (BIP, CNRS-AMU UMR 7281, Marseille, France), S.R.A Marque (ICR, CNRS-AMU UMR 7273, Marseille, France) and A. Rockenbauer (Research Centre of Natural Sciences, Budapest, Hungary). NL is grateful to Aix-Marseille Université and PACA Region for her PhD fellowship.

References

Avilan, L., Puppo, C., Erales, J., Woudstra, M., Lebrun, R., and Gontero, B. (2012). CP12 residues involved in the formation and regulation of the glyceraldehyde-3-phosphate dehydrogenase-CP12-phosphoribulokinase complex in Chlamydomonas reinhardtii. *Mol. Biosys.* 8, 2994–3002. doi: 10.1039/c2mb25244a

Babu, M. M., Van Der Lee, R., De Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struc. Biol.* 21, 432–440. doi: 10.1016/j.sbi.2011.03.011

Belle, V., Fournel, A., Woudstra, M., Ranaldi, S., Prieri, F., Thomé, V., et al. (2007). Probing the opening of the pancreatic lipase lid using site-directed spin labeling and EPR spectroscopy. *Biochemistry* 46, 2205–2214. doi: 10.1021/bi0616089

Belle, V., Rouger, S., Costanzo, S., Liquière, E., Strancar, J., Guigliarelli, B., et al. (2008). Mapping alpha-helical induced folding within the intrinsically disordered C-terminal domain of the measles virus nucleoprotein by site-directed spin-labeling EPR spectroscopy. *Proteins* 73, 973–988. doi: 10.1002/prot.22125

Bordignon, E. (2011). Site-directed spin labeling of membrane proteins. *Top. Curr. Chem.* 321, 121–157. doi: 10.1007/128_2011_243

Casey, T. M., Liu, Z. L., Esquiaqui, J. M., Pirman, N. L., Milshteyn, E., and Fanucci, G. E. (2014). Continuous wave W- and D-Band EPR spectroscopy offer "sweet-spots" for characterizing conformational changes and dynamics in intrinsically disordered proteins. *Biochem. Bioph. Res. Com*. 450, 723–728. doi: 10.1016/j.bbrc.2014.06.045

Chen, M., Margittai, M., Chen, J., and Langen, R. (2007). Investigation of alpha-synuclein fibril structure by site-directed spin labeling. *J. Biol. Chem.* 282, 24970–24979. doi: 10.1074/jbc.M700368200

Chouard, T. (2011). Breaking the protein rules. *Nature* 471, 151–153. doi: 10.1038/471151a

Drescher, M. (2012). EPR in protein science intrinsically disordered proteins. *EPR Spectrosc. Appl. Chem. Biol.* 321, 91–119. doi: 10.1007/128_2011_235

Dyson, H. J., and Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 12, 54–60. doi: 10.1016/S0959-440X(02)00289-0

Erales, J., Lorenzi, M., Lebrun, R., Fournel, A., Etienne, E., Courcelle, C., et al. (2009). A new function of GAPDH from *Chlamydomonas reinhardtii*: a thiol/disulfide exchange reaction with CP12. *Biochemistry* 48, 6034–6040. doi: 10.1021/bi900569h

Erales, J., Mekhalfi, M., Woudstra, M., and Gontero, B. (2011). Molecular mechanism of NADPH-Glyceraldehyde-3-phosphate dehydrogenase regulation through the C-Terminus of CP12 in Chlamydomonas reinhardtii. *Biochemistry* 50, 2881–2888. doi: 10.1021/bi1020259

Fermani, S., Trivelli, X., Sparla, F., Thumiger, A., Calvaresi, M., Marri, L., et al. (2012). Conformational selection and folding-upon-binding of intrinsically disordered protein CP12 regulate photosynthetic enzymes assembly. *J. Biol. Chem.* 287, 21372–21383. doi: 10.1074/jbc.M112.350355

Gardebien, F., Thangudu, R. R., Gontero, B., and Offmann, B. (2006). Construction of a 3D model of CP12, a protein linker. *J. Mol. Graph. Model* 25, 186–195. doi: 10.1016/j.jmgm.2005.12.003

Graciet, E., Gans, P., Wedel, N., Lebreton, S., Camadro, J. M., and Gontero, B. (2003). The small protein CP12: a protein linker for supramolecular complex assembly. *Biochemistry* 42, 8163–8170. doi: 10.1021/bi034474x

Gross, A., Columbus, L., Hideg, K., Altenbach, C., and Hubbell, W. L. (1999). Structure of the KcsA potassium channel from Streptomyces lividans: a site-directed spin labeling study of the second transmembrane segment. *Biochemistry* 38, 10324–10335. doi: 10.1021/bi990856k

Habchi, J., Blangy, S., Mamelli, L., Jensen, M. R., Blackledge, M., Darbon, H., et al. (2011). Characterization of the interactions between the nucleoprotein and the phosphoprotein of henipavirus. *J. Biol. Chem.* 286, 13583–13602. doi: 10.1074/jbc.M111.219857

Habchi, J., and Longhi, S. (2012). Structural disorder within paramyxovirus nucleoproteins and phosphoproteins. *Mol. Biosys.* 8, 69–81. doi: 10.1039/C1MB05204G

Habchi, J., Tompa, P., Longhi, S., and Uversky, V. N. (2014). Introducing protein intrinsic disorder. *Chem. Rev.* 114, 6561–6588. doi: 10.1021/cr400514h

Hideg, K., Kalai, T., and Sar, C. P. (2005). Recent results in chemistry and biology of nitroxides. *J. Heterocyclic. Chem.* 42, 437–450. doi: 10.1002/jhet.5570420311

Hubbell, W. L., Lopez, C. J., Altenbach, C., and Yang, Z. Y. (2013). Technological advances in site-directed spin labeling of proteins. *Curr. Opin. Struc. Biol.* 23, 725–733. doi: 10.1016/j.sbi.2013.06.008

Hubbell, W. L., Mchaourab, H. S., Altenbach, C., and Lietzow, M. A. (1996). Watching proteins move using site-directed spin labeling. *Structure* 4, 779–783. doi: 10.1016/S0969-2126(96)00085-8

Jensen, M. R., Communie, G., Ribeiro, E. A., Martinez, N., Desfosses, A., Salmon, L., et al. (2011). Intrinsic disorder in measles virus nucleocapsids. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9839–9844. doi: 10.1073/pnas.1103270108

Kaplan, R. S., Mayor, J. A., Kotaria, R., Walters, D. E., and Mchaourab, H. S. (2000). The yeast mitochondrial citrate transport protein: determination of secondary structure and solvent accessibility of transmembrane domain IV using site-directed spin labeling. *Biochemistry* 39, 9157–9163. doi: 10.1021/bi000433e

Kavalenka, A., Urbancic, I., Belle, V., Rouger, S., Costanzo, S., Kure, S., et al. (2010). Conformational analysis of the partially disordered measles virus NTAIL-XD complex by SDSL EPR spectroscopy. *Biophys. J.* 98, 1055–1064. doi: 10.1016/j.bpj.2009.11.036

Kingston, R. L., Hamel, D. J., Gay, L. S., Dahlquist, F. W., and Matthews, B. W. (2004). Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8301–8306. doi: 10.1073/pnas.0402690101

Klare, J. P., and Steinhoff, H. J. (2009). Spin labeling EPR. *Photosynth. Res.* 102, 377–390. doi: 10.1007/s11120-009-9490-7

Le Breton, N., Martinho, M., Kabytaev, K., Topin, J., Mileo, E., Blocquel, D., et al. (2014). Diversification of EPR signatures in site directed spin labeling using a beta-phosphorylated nitroxide. *Phys. Chem. Chem. Phys.* 16, 4202–4209. doi: 10.1039/c3cp54816c

Lebreton, S., Andreescu, S., Graciet, E., and Gontero, B. (2006). Mapping of the interaction site of CP12 with glyceraldehyde-3-phosphate dehydrogenase from Chlamydomonas reinhardtii. Functional consequences for glyceraldehyde-3-phosphate dehydrogenase. *FEBS J.* 273, 3358–3369. doi: 10.1111/j.1742-4658.2006.05342.x

Longhi, S., Receveur-Brechot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., et al. (2003). The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J. Biol. Chem.* 278, 18638–18648. doi: 10.1074/jbc.M300518200

Lorenzi, M., Puppo, C., Lebrun, R., Lignon, S., Roubaud, V., Martinho, M., et al. (2011). Tyrosine-targeted spin labeling and EPR spectroscopy: an alternative strategy for studying structural transitions in proteins. *Angew Chem. Int. Edit.* 50, 9108–9111. doi: 10.1002/anie.201102539

Lorenzi, M., Sylvi, L., Gerbaud, G., Mileo, E., Halgand, F., Walburger, A., et al. (2012). Conformational selection underlies recognition of a molybdoenzyme by its dedicated chaperone. *PLoS ONE* 7:e49523. doi: 10.1371/journal.pone.0049523

Martinho, M., Habchi, J., El Habre, Z., Nesme, L., Guigliarelli, B., Belle, V., et al. (2013). Assessing induced folding within the intrinsically disordered C-terminal domain of the Henipavirus nucleoproteins by site directed spin labeling EPR spectroscopy. *J. Biomol. Struc. Dyn.* 31, 453–471. doi: 10.1080/07391102.2012.706068

Matsumura, H., Kai, A., Maeda, T., Tamoi, M., Satoh, A., Tamura, H., et al. (2011). Structure basis for the regulation of Glyceraldehyde-3-Phosphate dehydrogenase activity via the intrinsically disordered protein CP12. *Structure* 19, 1846–1854. doi: 10.1016/j.str.2011.08.016

Mchaourab, H. S., Lietzow, M. A., Hideg, K., and Hubbell, W. L. (1996). Motion of spin-labeled side chains in T4 lysozyme. Correlation with protein structure and dynamics. *Biochemistry* 35, 7692–7704. doi: 10.1021/bi960482k

Mchaourab, H. S., Steed, P. R., and Kazmier, K. (2011). Toward the fourth dimension of membrane protein structure: insight into dynamics from spin-labeling EPR spectroscopy. *Structure* 19, 1549–1561. doi: 10.1016/j.str.2011.10.009

Mileo, E., Etienne, E., Martinho, M., Lebrun, R., Roubaud, V., Tordo, P., et al. (2013a). Enlarging the panoply of site-directed spin labeling electron paramagnetic resonance (SDSL-EPR): sensitive and selective spin-labeling of tyrosine using an isoindoline-based nitroxide. *Bioconj. Chem.* 24, 1110–1117. doi: 10.1021/bc4000542

Mileo, E., Lorenzi, M., Erales, J., Lignon, S., Puppo, C., Le Breton, N., et al. (2013b). Dynamics of the intrinsically disordered protein CP12 in its association with GAPDH in the green alga Chlamydomonas reinhardtii: a fuzzy complex. *Mol. Biosys.* 9, 2869–2876. doi: 10.1039/c3mb70190e

Pannier, M., Veit, S., Godt, A., Jeschke, G., and Spiess, H. W. (2000). Dead-time free measurement of dipole-dipole interactions between electron spins. *J. Magn. Reson.* 142, 331–340. doi: 10.1006/jmre.1999.1944

Rockenbauer, A., and Korecz, L. (1996). Automatic computer simulations of ESR spectra. *Appl. Magn. Reson.* 10, 29–43. doi: 10.1007/BF03163097

Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M. R., Zweckstetter, M., et al. (2010). NMR characterization of long-range order in intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* 132, 8407–8418. doi: 10.1021/ja101645g

Siddiqua, A., and Margittai, M. (2010). Three- and four-repeat tau coassemble into heterogeneous filaments: an implication for Alzheimer disease. *J. Biol. Chem.* 285, 37920–37926. doi: 10.1074/jbc.M110.185728

Stoll, S., and Schweiger, A. (2006). EasySpin, a comprehensive software package for spectral simulation and analysis in EPR. *J. Magn. Reson.* 178, 42–55. doi: 10.1016/j.jmr.2005.08.013

Tompa, P., and Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 33, 2–8. doi: 10.1016/j.tibs.2007.10.003

Torok, M., Milton, S., Kayed, R., Wu, P., Mcintire, T., Glabe, C. G., et al. (2002). Structural and dynamic features of Alzheimer's A beta peptide in amyloid fibrils studied by site-directed spin labeling. *J. Biol. Chem.* 277, 40810–40815. doi: 10.1074/jbc.M205659200

Uversky, V. N., and Dunker, A. K. (2010). Understanding protein non-folding. *BBA Proteom* 1804, 1231–1264. doi: 10.1016/j.bbapap.2010.01.017

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: 11 February 2016
doi: 10.3389/fmolb.2016.00004

There is Diversity in Disorder—"In all Chaos there is a Cosmos, in all Disorder a Secret Order"

Jakob T. Nielsen* and Frans A. A. Mulder

Department of Chemistry and Interdisciplinary Nanoscience Center, University of Aarhus, Aarhus, Denmark

The protein universe consists of a continuum of structures ranging from full order to complete disorder. As the structured part of the proteome has been intensively studied, stably folded proteins are increasingly well documented and understood. However, proteins that are fully, or in large part, disordered are much less well characterized. Here we collected NMR chemical shifts in a small database for 117 protein sequences that are known to contain disorder. We demonstrate that NMR chemical shift data can be brought to bear as an exquisite judge of protein disorder at the residue level, and help in validation. With the help of secondary chemical shift analysis we demonstrate that the proteins in the database span the full spectrum of disorder, but still, largely segregate into two classes; disordered with small segments of order scattered along the sequence, and structured with small segments of disorder inserted between the different structured regions. A detailed analysis reveals that the distribution of order/disorder along the sequence shows a complex and asymmetric distribution, that is highly protein-dependent. Access to ratified training data further suggests an avenue to improving prediction of disorder from sequence.

Keywords: intrinsically disordered proteins, NMR spectroscopy, data interpretation, statistical, chemical shift, protein conformation

OPEN ACCESS

Edited by:
Kris Pauwels,
Vrije Universiteit Brussel, Belgium

Reviewed by:
Wim Vranken,
Vrije Universiteit Brussel, Belgium
Piero Andrea Temussi,
Università di Napoli Federico II, Italy
Miquel Adrover,
Universitat de les Illes Balears, Spain

****Correspondence:***
Jakob T. Nielsen
jtn@chem.au.dk
Quote from: Carl G. Jung in "Die Archetypen und das kollektive Unbewusste", 4th ed., Walter, 1980

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

Received: *27 November 2015*
Accepted: *25 January 2016*
Published: *11 February 2016*

Citation:
Nielsen JT and Mulder FAA (2016) There is Diversity in Disorder—"In all Chaos there is a Cosmos, in all Disorder a Secret Order". Front. Mol. Biosci. 3:4. doi: 10.3389/fmolb.2016.00004

INTRODUCTION

A systematic re-examination of the protein structure–function paradigm is required to accommodate intrinsically unfolded/disordered proteins (IDPs). There are two major reasons for this reappraisal: (1) the results of bioinformatics analyses of the genomic codes for protein amino acid sequences, and (2) the accumulation of experimental evidence for the existence of a rather large number of protein domains and even entire proteins, lacking ordered structure under physiological conditions (Dyson and Wright, 2001, 2004; Uversky, 2003; Vucetic et al., 2003). Analysis of sequence data for complete genomes indicates that intrinsically disordered proteins are highly prevalent, and that the proportion of proteins that contain such segments increases with the increasing complexity of an organism; Putative long (>30 residue) disordered segments are found to occur in 2.0% of archaean, 4.2% of eubacterial, and 33% of eukaryotic proteins (Ward et al., 2004). Protein disorder roughly segregates into three major classes, depending on whether disorder serves a primary functional role, or serves permanent or transient interactions (van der Lee et al., 2014). Moreover, disorder maps to proteins with important functions, such as signal transduction and control of transcription, and IDPs are involved in all major classes of disease (Gregersen et al., 2006; Uversky et al., 2008). Although proteins may adopt different structures inside a cellular milieu, this paper is concerned solely with proteins studied under *in vitro* conditions. For the biological relevance of disorder for the selected proteins, the interested reader is referred to the original papers that described the proteins considered herein.

This paper draws on methodological advances in NMR spectroscopy to study IDPs, and systematic analysis of chemical shift data for the prediction of disorder from sequence. The aim of this work is to develop an experimentally calibrated "ruler" to detect and quantify sequence-specific protein disorder. NMR chemical shifts offer highly reliable and redundant residue-specific information on positional disorder, and this information is easy and unambiguous to get, using recently developed approaches (Jensen et al., 2013; Kragelj et al., 2013; Felli and Pierattelli, 2014; Konrat, 2014). In addition, the growing amount of NMR chemical shift assignment data now allows for rigorous and comprehensive analysis of protein disorder, and to employ this ruler to gauge the types of variation of protein disorder.

In this paper, we search for any potential trends or variations in order/disorder in an assorted set of proteins. To this end, we constructed a comprehensive collection of proteins with varying degrees of (partial) disorder, for which assigned NMR chemical shifts are available. We subsequently asked: "Is disorder similar in proteins, or are there different patterns to be discerned?," "What is a *typical* variation between order/disorder?," and "Are there proteins that deserve the label *super unfolded*, and are they representative of the general class of IDPs?." We demonstrate that it is possible to answer these questions, with the methods discussed herein. It is hoped that this initial experimental evaluation of residue-specific protein positional disorder will spark the further evolution of assessment tools for predicting disorder with greater detail and accuracy. In addition, our analysis paints a validated picture of protein disorder for a diverse subset of 117 example proteins that are either classified as IDPs, or possess long intrinsically disordered regions (IDRs), showing a highly abounding sequence context dependence.

METHODS

Generation of a Curated Database with Available NMR Chemical Shift Data

A set of proteins with different degrees of structural disorder, for which assigned chemical shifts were available, was generated in two steps. First, a set of proteins was generated from a keyword search in the BioMagResBank (BMRB) database (http://www.bmrb.wisc.edu; Ulrich et al., 2008). Second, we augmented this database with sequences from the Database of Protein Disorder (DisProt) (http://www.disprot.org; Vucetic et al., 2005; Sickmeier et al., 2007) of disordered regions, for which data were also present in the BMRB. In the first step, the BMRB database was searched for entries according to the BMRB database tag for the physical state of the protein; Entries were selected where _Entity_assembly.Physical_state = "denatured," "intrinsically disordered," "molten globule," "partially disordered," and "unfolded." In addition, entries with the words "disordered" or "unstructured" in the entry title were also included. In the second case, the BMRB database was searched for matches of all SwissProt identifiers present in the DisProt database. The sequences from DisProt and BMRB were aligned using the EMBOSS (http://www.ebi.ac.uk/Tools/psa/emboss_needle/) implementation of the Needleman–Wunsch alignment algorithm (Needleman and Wunsch, 1970) and BMRB entries with >20% of the aligned residues classified as disordered by DisProt were retained. In addition, an entry with in-house assigned chemical shift for the N-terminal heavy metal binding domain, residues 1–84, of a cupper-binding ATP-ase (Gourdon et al., 2011) in an unstructured state was added to the database.

Subsequently, the database was curated. First, conditions that deliberately destabilize folded proteins, such as extremes of pH, extremes of temperature, and denaturants were selected against, in order to avoid the artificial inclusion of disorder under non-native conditions. This step also removes any digressions that such conditions have on the chemical shifts: Only proteins with near-neutral pH ($4 < \mathrm{pH} < 9$), in weak salt buffer, not in complex with other molecules, and without modified amino acids were kept. Typical conditions of excluded entries were: pH < 4.0, bicelle/micelle/SDS present in buffer, TFE/DMSO/GuHCl or other denaturants added, presence of co-factors, and phosphorylation. Next, entries having fewer than 40 amino acid residues or fewer than 50 assigned chemical shifts were removed. Third, the collection of entries was culled in two steps to remove redundant data. To find highly similar entries, the EMBOSS Needle program was used to calculate pairwise sequence identity between the sequences. In the first step, sets of proteins families, defined as sets of chains with >90% mutual sequence identity, were identified and only the entry with the most native-like sequence/condition from each set was kept in the database. For example, wild type sequences were kept and mutant sequences were discarded when both were present, and in case data were available at different acidities, the pH closest to neutral was preferred. In the second step, groups of homologous chains, defined by having >50% sequence identity between at least one other sequence in the set, were identified. In each set, the subset of sequences that yielded the maximum total number of chemical shifts, and where all pairs had < 50% sequence identity, were kept. Finally, it was required that at least one residue could be defined as disordered based on NMR chemical shift data (see below, Equation 5). This procedure resulted in the construction of a set of 117 proteins with assigned chemical shifts, with varying degrees of structure/disorder. The list of 117 BMRB identifiers is provided as Table S1.

Calculation of Order/Disorder Metrics from Experimental NMR Data

Weighted sum of squared differences between observed and predicted shifts were calculated as follows

$$\chi^2(i) = \sum_{n} \sum_{j=i-1,i,i+1} \min(\Delta_{jn}{}^2, 16) \tag{1}$$

where the difference was truncated to four standard deviations using:

$$\Delta_{jn} = \frac{\delta_{obs}\left(j, n\right) - \delta_{pred}\left(j, n\right)}{\sigma\left(n\right)} \tag{2}$$

Here $\delta_{obs}\left(j, n\right)$ is the observed (offset corrected) chemical shift of atom type n for residue j, and $\delta_{pred}\left(j, n\right)$ is the neighbor corrected random coil chemical shift predicted based on the tripeptide centered at residue j (Tamiola et al., 2010). The chemical shift difference is scaled with the expected difference for residue j in an IDP using $\sigma(n) = 0.627$, 0.310, 0.219, 0.281, 0.102, 0.0566, 0.0546 ppm for $n =$ N, C′, Cβ, Cα, H_N, Hα, and Hβ, respectively. The supposed chi-square distributed number χ^2 is transformed to an approximately normal distributed number L, by using linear combinations of fractional powers of χ^2(Canal, 2005).

$$L = \rho^{1/6} - \frac{1}{2}\rho^{1/3} + \frac{1}{3}\rho^{1/2},\ \rho = \frac{\chi^2}{N} \tag{3}$$

where N is the number of assigned chemical shifts for the triplet. L is converted to a standard normal distributed number, Z_{IDR}, by correcting with the known mean and standard deviation for L (Canal, 2005).

$$Z_{IDR} = \frac{L - \mu_L\left(N\right)}{\sigma_L\left(N\right)} \tag{4}$$

We assign local sequence specific residue states as either *disordered* (D) according to the definition:

$$Z_{IDR} < 3 \tag{5}$$

or alternatively as *ordered* (O) if $Z_{IDR} \geq 3$. A protein sequence can be thought of as consisting of alternating segments of disordered and ordered residues with lengths s_i. The *sequence disorder complexity*, C_{SD}, is now defined as

$$C_{SD} = \frac{e^H - 1}{N},\ H = -\sum_i \frac{s_i}{N}\ln\left(\frac{s_i}{N}\right), \tag{6}$$

where N is the number of residues in the protein. Note that H is closely related to the Shannon entropy of a statistical distribution. If a protein is built of n segments of equal length then $e^H = n$ and if the segments have different lengths $e^H < n$. In particular, a protein with exclusively disordered/ordered residues (pure state) will have $H = 0$, $e^H = 1$, and $C_{SD} = 0$ whereas a protein where order and disorder continuously alternate along the sequence has a maximal value of $C_{SD} = 1$. If order and disorder are independently randomly distributed with a probability of 0.5, we simulated with a random number generator that the sequence disorder complexity would be ca. 0.41 on average. For a general probability, p, and random outcome: $C_{SD} \approx p$. Therefore, it make sense to compare the sequence disorder complexity to the fraction of disordered residues $f_D = N_D/N$ where N_D is the number of disordered residues ($Z_{IDR} < 3$). As such, the *relative sequence disorder complexity*, C_{SD}/f_D is expected to have a smaller variation.

Procedure for Re-Referencing Assigned Chemical Shifts

Chemical shifts are deposited using different referencing procedures at different conditions such as temperature, added salt, and pH, and hence, it is likely that in some cases the observed chemical shift would be slightly, yet systematically, offset from the random chemical shift derived from the sequence. However, since even small deviations from random coil shifts are indicative of structure ordering, we estimated an offset correction for each entry in our database. The chemical shifts were re-referenced for each atom type independently using the following procedure: First, the neighbor corrected random coil chemical shifts were calculated for all residues following the procedure of Tamiola et al. as implemented in the program ncIDP (http://www.protein-nmr.org/; Tamiola et al., 2010), and the deviations from random coil chemical shifts, Δ, were calculated using Equations (1) and (2) above. Assuming that the NMR data is correctly referenced, this procedure identifies small deviations due to deviations in pH and temperature of the experimental data relative to the reference database. Next, the standard deviation of Δ was calculated for nine consecutive residues, and the sequence position with the smallest standard deviation was identified. The average of Δ for the nine residues was then used as candidate offset correction. The average value of Z_{IDR} was evaluated using (i) the candidate offset correction as described above and (ii) no offset correction. The scenario leading to the smallest average Z_{IDR} was chosen as the initial offset estimation (i.e., either using the candidate offset or no offset correction). because chemical shift distributions in IDRs show a distribution (Tamiola et al., 2010), much narrower than those in structured parts (Wang and Jardetzky, 2002; Zhang et al., 2003), owing to the many additional contributions to the chemical shifts in structured proteins (Wishart and Case, 2001; Shen and Bax, 2010; Nielsen et al., 2012). Finally, using the chosen initial offset estimation, the average Δ using all the chemical shifts for the particular atom type in disordered residues only (Equation 5) was calculated and this number was used as the final offset correction. To avoid eliminating true deviations originating from structure, in both cases, the revised offset was only used if there was a significant reduction in the chemical shift standard deviation, $\sigma(\delta_{\text{off}})$, relative to the uncorrected equivalent, $\sigma(\delta_0)$, considering the number of chemical shift observations, N, by the application of Akaikes Information Criterion (Akaike, 1974, 1985) using a variance inflation factor of 10.0. i.e., the offset correction was only accepted if: $N^*\ln(\sigma(\delta_0)/\sigma(\delta_{\text{off}})) > 20.0$ and $N > 3$.

RESULTS

Analysis of Chemical Shift Dispersion for Seven Case Story Proteins

The database of proteins containing disordered regions was constructed as described in Methods. This carefully curated database contains 117 unique proteins chains and 13,069 residues with 65,574 assigned backbone and Cβ/Hβ chemical shifts in total (excluding terminal residues). The database contains many well-characterized IDPs such as α-synuclein (αSyn, bmrbID = 6968) (Bermel et al., 2006), small heat shock protein (Hsp12, bmrbID = 17483; Singarapu et al., 2011), the CD79a cytosolic domain (bmrbID 18867; Isaksson et al., 2013), apo-IscU (bmrbID = 17836; Kim et al., 2012), and the cytoplasmic domain of human neuroligin-3 (bmrbID = 17290; Wood et al., 2012). For each protein in the database we calculated the scaled difference, Δ, from random coil shifts (Equation 2). This difference is plotted as a function of residue number for seven representative proteins from the database in **Figure 1**. It is seen that for the two proteins, known to be intrinsically unstructured, αSyn (bmrbID = 6968) and Hsp12 (bmrbID = 17483), fluctuations away from random coil shifts are very small throughout the sequence. This is also borne out by the fraction of disordered residues being 97% in both cases (see Methods Equations 1–5 and below). Conversely, two other proteins reported to be IDPs, the 18.5 kDa isoform of murine myelin basic protein (bmrbID = 15131; Libich et al., 2007), and the Cholera Toxin Enzymatic Domain (1–167) (bmrbID = 15162; Ampapathi et al., 2008) display larger scatter. Values for the fraction of disordered residues f_D are 0.45 and 0.28, respectively, which indicates that these order/disorder transitions of the cholera toxin enzymatic domain are essential for function, the protein is, in fact, mostly folded. For three further proteins, we note yet another pattern in their chemical shift residuals along the sequence, containing distinct separate segments with larger local spread of the chemical shifts for all nuclei. The examples chosen here are human cardiac troponin I, residues 1–73 (Hwang et al., 2014), inhibitor-2 involved in protein phosphatase 1 regulation (Kelker et al., 2007), and the small VCP/p97-interacting protein (Wu et al., 2013; BMRB ids: 25118, 15179, and 19485, respectively). In these examples, the larger local scatter is biased in the direction of downfield shifted C' and Cα chemical shifts and upfield shifted Cβ, Hα, H_N, and N chemical shifts (**Figures 1C–E**). This observation is consistent with the presence of helical structure formation. In addition, in these three examples, we see a variation in helix size within the same protein, as well as amongst different proteins. A variation in the amplitude of the chemical shift residuals is also observed. Since chemical shifts are time-averaged observables, smaller amplitude of the chemical shift residuals correspond to fractional occupancies of helical states as discussed previously (Marsh et al., 2006).

Local Disordered Residues

Inspired by the observations of local disorder/order in small segments in a protein, we develop here a formalism, where we state that a residue can be in one of two situations: either an (intrinsically) disordered state (D) corresponding to a non-biased mixture of conformations dictated by the Boltzmann distribution with resulting population-averaged chemical shifts, or in a completely ordered state (O) with a fixed structure. Here we use the simplest probabilistic model, a normal distribution, for the chemical shift in a disordered state. The mean can be estimated from the primary structure as described in Tamiola et al. (2010) and the standard deviation can be derived from statistics (see Methods). Conversely, for an ordered residue, the chemical shifts would have much larger deviation from the mean, corresponding to a bias in the Boltzmann distribution of conformations. Outliers from the normal distribution are indicative of residual order, and can be identified and analyzed for each atom specific chemical shift, but when analyzed in combination the evidence is much more reliable. Therefore, we introduce the *Chemical shift Z-score for assessing Order/Disorder* (the *CheZOD score*), Z_{IDR}, derived from the rmsd of all chemical shift residuals within a residue triplet and linear combinations of fractional powers of this rmsd as described in Methods (Equations 1–4). Assuming that the chemical shifts are normally distributed in disordered residues, Z_{IDR} is standard normal distributed. This is valid independent of the number of backbone chemical shifts available. Hence, we define the distinction between disordered and ordered residues based on outliers from the normal distribution for Z_{IDR}; a residue is said to be *disordered* if the CheZOD score, $Z_{IDR} < 3$, and protein is said to have *local disorder* at this position in the sequence. Furthermore, the CheZOD score is not only a binary classifier of order/disorder, but provides a scale for the "degree of disorder," which can both classify partially formed/fractionally occupied structure (ca. $3 < Z_{IDR} < 8$) and fully formed structures ($Z_{IDR} > 8$).

Disorder Profiles of All Proteins in the Database

The CheZOD score, Z_{IDR}, was calculated for all residues in all the 117 proteins in the database, henceforth the CheZOD database, and the fraction of the disordered residues was calculated for each protein. We observe that no protein has 100% disordered residues according to our definition. However, five proteins have 95% disordered residues or more, among these are αSyn and Hsp12 (discussed above) and also p15(PAF) (bmrbID = 19332) (De Biasio et al., 2014), the CD3e cytosolic domain of Eukaryota Metazoa Homo sapiens T-cells (bmrbID = 18889; Isaksson et al., 2013), and recombinant murine BG21 isoform of Golli myelin basic protein (bmrbID = 7358; Ahmed et al., 2007). Furthermore, the CheZOD database contains 15, 34, and 70 proteins that have >90, 80, and 50% disordered residues, respectively, for which NMR chemical shifts are available, spanning the full range from disordered to ordered.

The CheZOD score along the sequence (disorder profile) is visualized in **Figure 2** for all proteins in the CheZOD database. It is seen that the database roughly separates into proteins that are mostly disordered, with small segments of order scattered along the sequence, and mostly structured proteins, with small segments of disorder bridging between ordered domains, and in particular at the termini. Conversely, cases with roughly equal amounts of disordered and ordered residues are relatively rare (see also **Figures 3A, 4**). A second conclusion from our analysis is, that we did not identify proteins where the fluctuations between ordered and disordered residues were completely random. Rather, all proteins have distinct medium-length segments of ordered/disordered residues (with a typical length of 10–30 residues) indicative of short-range correlated behavior, irrespective of the average state of the protein. Careful inspection of **Figure 2** reveals that most residues are either fully disordered or fully ordered, whereas fewer are partly ordered (ca. $3 < Z_{IDR} < 8$; see also **Figure 3B**). Hence, proteins in "the Twilight zone" which are not completely ordered or disordered,

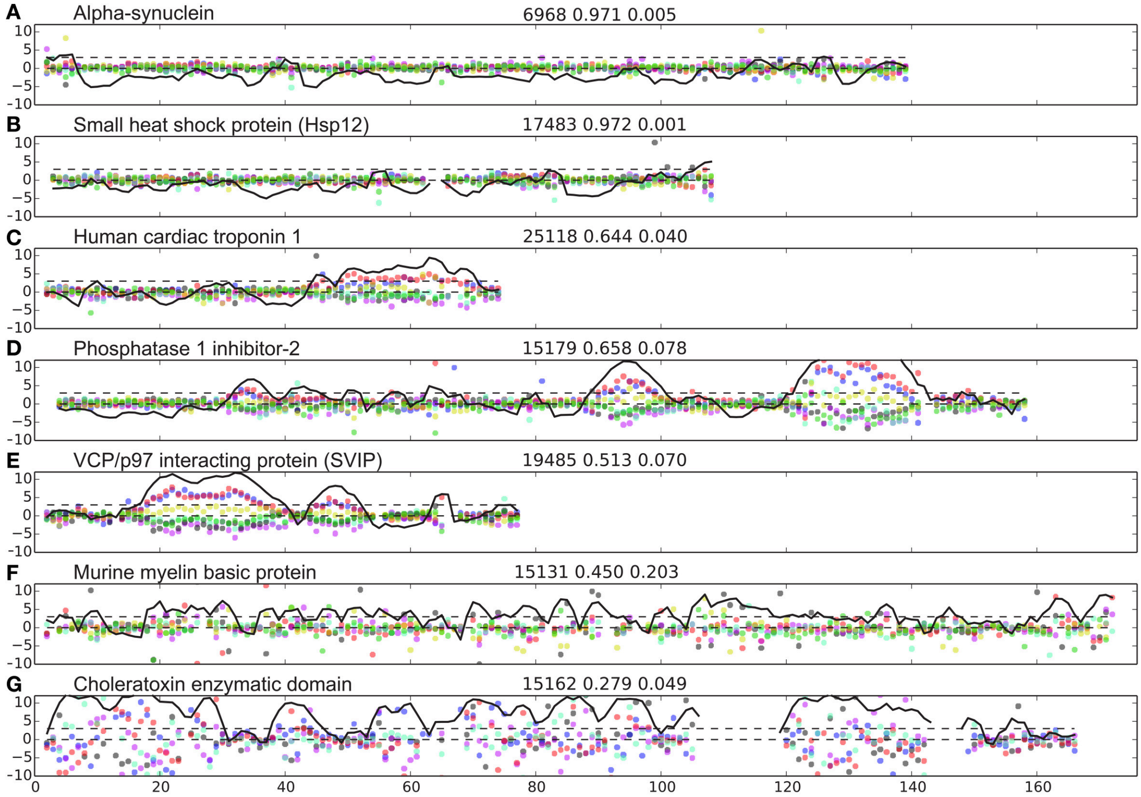


FIGURE 1 | Examples of IDP disorder profiles. (A–G) weighted difference between observed and predicted shifts, Δ_{jn} (Equation 2) shown with blue, red, black, green, cyan, magenta, and yellow dots for C′, Cα, Cβ, Hα, H_N, N, and Hβ, respectively. The Z-score (Equation 4) is shown as a black line, showing the lines, $Z = 0$ and $Z = 3$ for reference with black broken lines. The name of the protein analyzed is indicated at the top of each panel to the left. Three numbers are provided on top of each panel (middle) referring to, the BMRB id, the fraction of disordered residues, and the sequence disorder complexity, respectively.

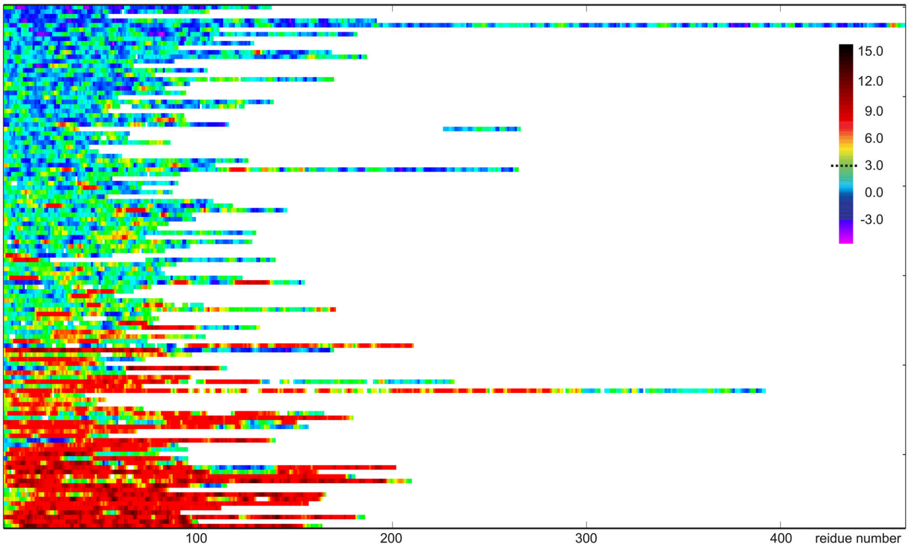


FIGURE 2 | Visualization of disorder profiles for the 117 proteins sequences in the CheZOD database. Each row represents a single protein where disordered residues ($Z_{IDR} < 3$) are shown in blue, ordered residues ($Z_{IDR} > 3$) are shown in red, and residues with average order shown in green/yellow. The proteins are depicted from top to bottom sorted according to the average of Z_{IDR} for the protein.

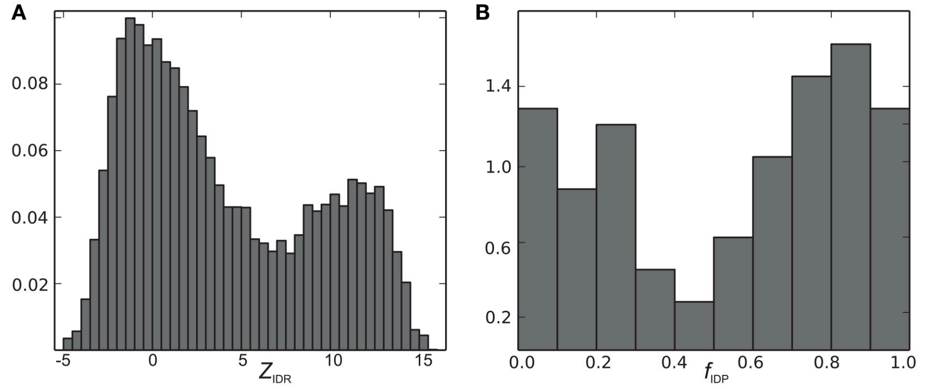


FIGURE 3 | Histograms of (A) Z_{IDR} and (B) f_{IDP} (normalized to a total area under the curve of 1.0).

both at the local and global level, appear to be under-represented. An example of such a rare protein is the 18.5 kDa isoform of murine myelin basic protein, shown in **Figure 1F**.

The Sequence Disorder Complexity of a Protein

We now define a measure of the extent of alternations between ordered and disordered segments, the *sequence disorder complexity*, C_{SD}, which is defined as the Shannon entropy of $Z_{IDR} < 3$ (along the sequence), scaled by the length of the sequence (see Equation 6). C_{SD} is 0 for a 100% disordered or ordered protein, and largest for a protein with many alternations between ordered/disordered segments (showing the protein with the maximum sequence disorder complexity in the CheZOD database in **Figure 1F**). In **Figure 4** the *sequence disorder complexity*, C_{SD}, is shown as a function of the fraction of disordered residues, f_D. For comparison, the relative complexities C_{SD}/f_D and $C_{SD}/(1-f_D)$, which have a more even variation, are shown in **Figure 5**. It is seen that (i) the distribution of order/disorder along the sequence is not completely random (i.e., the complexity is always much less than the maximum value for C_{SD}) (ii) The distribution is also not close to minimum complexity, which would correspond to two separate regions of order/disorder. (iii) The relative complexities C_{SD}/f_D and $C_{SD}/(1-f_D)$, as viewed in **Figure 5**, are asymmetric around $f_D = 0.5$. It is seen that mostly disordered proteins ($f_D > 0.5$) are more complex compared to their mostly structured counterparts, i.e., they appear to have a relative larger number of scattered smaller segments with local order. However, this apparent asymmetry is mostly due to the specific definition of complexity; i.e., if we choose a different cut-off for a unstructured residue, say $Z_{IDR} > 6$ (rather than $Z_{IDR} < 3$, Equation 5), the corresponding correlation becomes more symmetric (data not shown) and barely reflects that structured regions are formed internally in the sequence, whereas disordered regions are more typically formed at the termini of the sequence.

Data Extraction and Fraction of Disordered Residues: BMRB vs. Disprot

In **Figures 4**, **5** there are some trends visible in the fraction of disordered residues, f_D, related to the procedure for data extraction and the physical_state tag. In particular, the proteins that have physical_state = "intrinsically disordered" are indeed mostly disordered ($f_D \geq 0.5$) in 15 cases, except one. Likewise, proteins corresponding to physical state tags "denatured" and "unfolded" describe only 1 of 6 and 1 of 7 proteins, respectively, that are mostly ordered. This is in some contrast to the group of proteins found from text searches of "disordered" or "unstructured" (see Methods), which were also labeled as "native" (green points in **Figures 4**, **5**). For this group, 10 out of 42 are actually mostly ordered, yet all except one of these proteins still have some degree of disorder with $f_D > 0.1$. This bias was even more pronounced when considering the group of proteins identified by searching the DisProt database for corresponding entries in the BMRB database (see Methods). The proteins in this group were labeled with a "native" physical state (blue points in **Figures 4**, **5**) or had no label (2 cases, cyan). The search in the

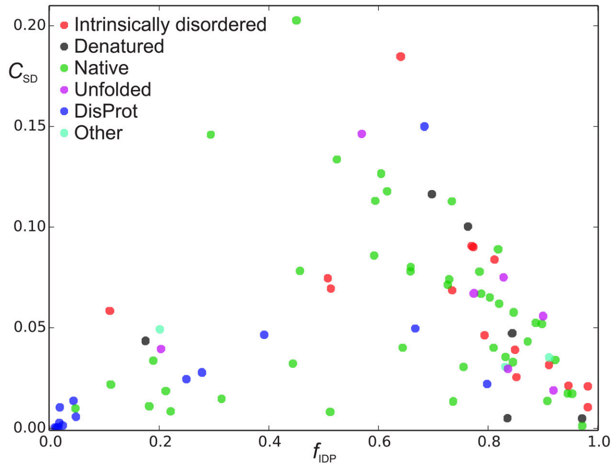


FIGURE 4 | Sequence disorder complexity, C_{SD} (Equation 6), as a function of the fraction of disordered residues, f_D. Each protein is shown with a different color according to the "physical_state" tag provided in the BMRB database. The proteins shown in blue were found in the DisProt database search, where proteins shown in blue had "physical_state" = "native" and cyan points refer to as "unknown" or had missing specification of "physical_state" in the BMRB file. The proteins shown with other colors than blue and gray were found in the BMRB database key word search.

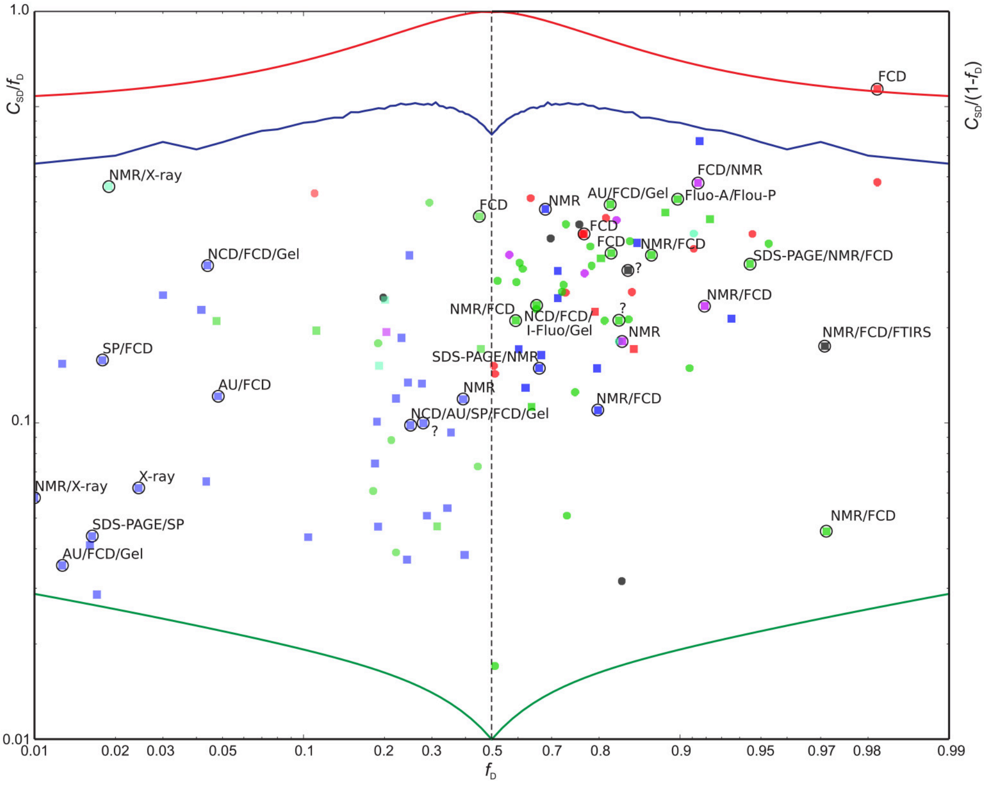


FIGURE 5 | Scaled sequence disorder complexity C_{SD}/f_D as function of f_D (left) and $C_{SD}/(1-f_D)$ vs. $(1-f_D)$ (right). Extreme values are shown for reference: The minimum possible C_{SD} for a protein of length L = 200 (green), the maximum C_{SD} (L = 200) (red), and C_{SD} for randomly distributed outcomes with a probability f_D for a disordered residue (blue). Entries highlighted with circles were also found in the DisProt database with confirmed > 90% sequence identity relative to the BRMB database amino acid sequence and with 90% of the aligned residues classified as disordered in DisProt (31 entries). See **Figure 4** above for color-coding. The proteins found by searching the DisProt database are shown as squares. The methods used by the DisProt depositors for assessing disorder are provided as annotations near the highlighting circles for the 31 validated entries above. An "?" indicates that no method was given. The following abbreviations were used: AU, Analytical ultracentrifugation; DLS, Dynamic light scattering; EMSA, Electrophoretic Mobility Shift Assay; FCD, Circular dichroism (CD) spectroscopy, far-UV; Fluo-A/Flou-P, Fluorescence polarization/anisotropy; FTIRS, Fourier transform infrared spectroscopy; Gel, Size exclusion/gel filtration chromatography; I-Fluo, Fluorescence, intrinsic; SAXS, Small-angle X-ray scattering (SAXS); SDS-PAGE, Polyacrylamide gel electrophoresis in sodium dodecyl sulfate; SP, Sensitivity to proteolysis.

DisProt database also resulted in re-identification of some of the proteins already present in the database from the BMRB physical state tag and text search (shown as colored squares in **Figure 5**, and not included in the counts below).

Since the DisProt database contains sequences with various contents of disorder, we contend that a protein from DisProt/BMRB is *validated* as disordered if at least 90% of the residues (aligned between the BMRB and DisProt entries) were defined as disordered in the DisProt database. Only the validated Disprot entries are shown in **Figure 4** (together with all the BMRB entries), whereas all DisProt entries are shown in **Figure 5** highlighting the validated ones by circles. It is seen in **Figure 4** that the validated group of DisProt entries contained 11 of 14 mostly ordered proteins, and seven of these were structured proteins, with $f_D < 0.1$. It appears that the classification "*mostly disordered*" in DisProt does not assert that a protein, under the conditions corresponding to the BRMB entry, will be disordered, when judged by chemical shift dispersion. However, if the validated entries from DisProt—which correspond to entries already found by the BMRB physical state searches (**Figure 5**, square points surrounded by circles)—are included in this analysis, the percentage increases, as now 18 of 31 proteins are mostly disordered. Next, we inspected the methods that were used for the disorder classification for the 31 validated DisProt entries (as provided in the DisProt database). A significant correlation is revealed regarding the use of NMR, which was only used in two of the eight cases with almost completely structured proteins, in 3/12 cases for mostly structured, and in 11/19 cases for the mostly disordered proteins (see annotations in **Figure 5**) suggesting that NMR is likely one of the most accurate methods for assessing disorder in proteins.

The above observations together show that the physical_state tag corresponds well to the content of structure in the protein, and, in particular, using this tag to search for intrinsically disordered proteins is a reliable procedure for identifying IDPs. Although the tags "intrinsically disordered," "unfolded," and "denatured" all consistently yield disordered proteins, one could suspect that in the latter two cases the protein might be in a somewhat non-native state, biased by conditions that could induce unfolding/denaturation of the protein (although it was specifically tried to avoid this, by excluding entries where it was specified that denaturants or co-factors were added, see Methods). This impression is supported by the fact that all six "denatured" proteins in the CheZOD database, and four of seven of the "unfolded" proteins, also have a structure available in the PDB database (as defined by the cross-ref PDB match in the BMRB entry), indicating that a folded version of the protein exists under some condition. For comparison, only 2 of 15 proteins with "intrinsically disordered" physical state have a folded structure in the PDB. An alternative procedure to identify "truly disordered" proteins, as demonstrated here, would be to search in the BMRB title for the words "disordered" or "unstructured" and analyze the proteins with a "native" physical state (green points in **Figures 4**, **5**). In this case, only 15 out of 42 had a folded structure available in the PDB database. In general, there is a close relationship between the fraction of disordered residues and the presence of a folded structure in the PDB; 26 of 70 of the mostly disordered proteins have a determined structure, whereas a vast excess of 44 of 47 from the mostly ordered proteins have a structure present in the PDB.

DISCUSSION

We have constructed here the CheZOD database that contains detailed information about the extent of disorder in 117 carefully curated protein data entries. Following our bottom-up approach to include all available data for potentially disordered proteins in the BMRB database, it was intended that the CheZOD database should be representative for the full range of intrinsically disordered proteins. In the quest for this we used both tag searches in BMRB for the physical state, as well as searches to find chemical shifts for entries in the DisProt database. The CheZOD database is relatively small compared to other databases with 117 proteins, but one can imagine that the CheZOD database could be expanded even further in the future by including searches for entries in other databases such as MobiDB (Potenza et al., 2014), IDEAL (Fukuchi et al., 2014), and D^2P^2 (Oates et al., 2013). We also expect further entries with chemical shifts to be available for inclusion in the CheZOD database in the near future following the recent growing focus on IDP research and advances in NMR analysis of IDPs. Despite the efforts made here, we do not claim here that our database is complete in the sense that it would cover all possible classes of proteins or types of disorder. We also concede that our database is slightly "biased" as a whole, in the sense that all entries correspond to proteins amenable to NMR spectroscopy, such as soluble proteins with small/medium size at relatively high concentrations. Notwithstanding these subtle objections, we still argue that the CheZOD database is *representative* for protein disorder. All entries in the CheZOD database is summarized in Table S1, and the full database including all the CheZOD Z-scores and backbone chemical shifts for each residues is available from www.protein-nmr.org.

We have used here our method for chemical shift referencing based on recalibration of the distribution of chemical shift for the disordered residues, and the random neighbor corrected chemical shifts of Tamiola et al. (2010). We now compare with the LACS method for re-referencing chemical shifts (Wang et al., 2005), Kjaergaard et al. random coil chemical shifts (Kjaergaard and Poulsen, 2011), and the RCI method for estimating local dynamics based on the chemical shifts (Berjanskii and Wishart, 2005).

Our method for re-referencing the chemical shifts identifies the disordered residues and calculates the average secondary chemical shifts, Δ_{ave}, for these residues. In theory, Δ_{ave} must be very close to zero, but it will deviate from zero for several possible reasons: (i) incorrect referencing by the authors of the entry, (ii) influence by sample conditions such as isotope effects, solvent, buffer, temperature, and pH, (iii) systematic bias in the estimation of neighbor corrected random coil shifts, (iv) small "true deviation" due to residual structure bias. Our method aims at addressing (i–iii) with a phenomenological offset correction by subtracting the average value, Δ_{ave}, from the chemical shifts. While some of the effects might be rather small, they could still have a dramatic impact on the interpretation of disorder due to the small variation in secondary shifts for IDPs. Unfortunately, it is of course not possible to separate effects of (i–iii) from "true deviation" which could mask structural signatures and thereby the disordered residues would appear slightly more disordered. We note, however, that this re-referencing does not lead to a larger number of residues being classified as disordered but only effects the amplitude of the disorder, merely leading to a slight skewing of the Z-score scale in the low-value end of the scale. Following the rationale that the effect of (i–iii) would often be much larger than the "true deviation" correction, this is why we applied the offset correction exclusively in cases where it was significantly different from zero as judged by Akaikes information criterion.

Analysis of all the applied offset corrections for the full CheZOD database by each atom type reveals that the offset correction was used in 20.6% of the chemical shift sets when analyzing the individual atom types separately (see Figure S1 in the Supplementary Material). In most cases the offset corrections were small (<0.5 ppm for ^{13}C and ^{15}N and <0.15 ppm for ^{1}H) but for a few entries, a large offset correction (>2 ppm) was needed for the carbon atom types. In these cases with large corrections, the offset for the different carbon atom types were correlated, underscoring the credibility of our method for offset corrections although these chemical shifts are sometimes assigned from different spectra. Some systematic trends in the sign of the offset correction can be learned, in particular the Hα and H_N correction were often negative and positive, respectively, by ca. 0.1 ppm (discussed in more detail below). Furthermore, it is seen that offsets were more frequently used for more disordered proteins, which was to be expected, since the offset is calculated using only the disordered residues and hence, a deviation from the expected mean would be more statistically significant according to Akaikes information criterion, which is proportional to the number of data points.

In order to provide independent validation of our method, we compared our offset corrections to offset corrections estimated by LACS (Wang et al., 2005), and found that the two methods agree exceedingly well (see Figure S2 in the Supplementary Material), with the exception of the Hα shift, for which there was a small systematic deviation, which is, however, accounted for when using our phenomenological offset correction method. These Hα offsets predicted by our method were ca. −0.1 ppm on average whereas the offsets for the same entries were ca. +0.1 ppm although still apparently linearly related. This observation suggests small differences in the reference values used for Hα random coil chemical shifts for the two methods. There were some notable differences between our method and LACS that implicate that it cannot be applied in all cases; in particular LACS requires assigned Cα and Cβ chemical shifts and only provides offset correction estimates for Cα, Cβ, Hα, and C′.

To address the possibility that observed chemical shifts could be affected by sample conditions, we compared our secondary chemical shifts and Z-scores with the corresponding results obtained from derivations using the random coil chemical shifts from Kjaergaard et al. (Kjaergaard and Poulsen, 2011), which includes an estimate of the effect due to variations in temperature and pH. Comparing the results for our seven test case entries in **Figure 1**, there were only some small differences for the disordered residues where the secondary chemical shifts were small compared to the difference between the random coil shift estimates, and elsewhere largely very similar results for the two methods (see Figure S3 in the Supplementary Material). This observation suggests that our method already includes the effects on chemical shifts from pH and temperature implicitly. However, since it was not possible to account for all effects of the sample conditions on the chemical shifts or for cooperative/non-linear effects of the different parameters, we still argue that our phenomenological correction is the most appropriate.

NMR is very sensitive to ensemble averaging of local conformations, which is measured very accurately from the chemical shifts. Therefore, chemical-shift based methods for assessing order/disorder in proteins are both position-specific, and also provide a scale for the extent of disorder/order. These properties contrast with other biophysical techniques such as circular dichroism (CD), and infra-red (IR) spectroscopies, aimed at estimating the content of secondary structure, or techniques for estimating the protein size, density, hydrodynamic drag, or diffusion properties, using for example, size exclusion chromatography, ultracentrifugation, SAXS, or limited proteolysis (Vucetic et al., 2005; Tompa, 2009; Uversky and Longi, 2010). Missing density in X-ray crystallography data may also be indicative of local disorder (though it doesn't provide a scale for the disorder), but X-ray diffraction cannot be applied to proteins that do not contain any structured elements. Hence, with the chemical shift analysis proposed here, it is possible to "zoom-in" to look for finer details, and get a more detailed picture of the diversity in disorder.

Some other methods also provide residue based estimates for the local dynamics including δ2D (Camilloni et al., 2012) and the Random Coil Index (RCI) method (Berjanskii and Wishart, 2005). We have compared our Z-score for our seven test case proteins to the RCI estimates of the S^2 local order parameter in Figure S3. There was qualitatively agreement between the RCI estimated S^2 and our Z-score, i.e., low order for the disordered residue and high order (S^2 close to 1) for the ordered residues. However, there were still some more subtle differences between the two scores. Firstly, the largest differences were for the disordered residues where the secondary chemical shifts were small compared to the difference between the random coil shift estimates. This can be understood, since RCI uses a floor value for the absolute chemical shifts meaning that very small deviations from the chemical shifts are not captured and hence also very disordered residues are not distinguished. Secondly, the full range of orders appears to be less well described by the RCI estimates yielding only small differences for fully and partly formed structure (see e.g., Figure S3d). Thirdly, the RCI method gives a smoother trajectory along the sequence, which is because it uses both a three-point averaging of the absolute secondary chemical shifts as well as for the actual RCI index values. Finally, we note that whereas the RCI method is heuristic, based on a weighted sum of the absolute value of secondary chemical shifts for all atom types, our method is statistical, based on the chi square distribution of squared secondary chemical shifts, and due to this formulation it provides adequate estimates for the local dynamics also in cases were only a subset of the chemical shifts are available such as only proton or carbon chemical shifts.

Inspection of our CheZOD database of IDPs reveals that the proteins span a broad range of fractions of disorder and extent of disorder. Proteins are seen with both many fluctuations between segments of order/disorder, characterized by a high sequence disorder entropy, and with separation into larger completely ordered/disordered domains. This great diversity in disorder reflects the broad class of IDPs known to form various functions or interaction with different targets and malleability at different conditions.

Proteins that are either almost completely disordered, or completely structured, are most abundant in the CheZOD database, whereas cases with roughly equally mixed states are rare. This observation likely reflects the cooperative behavior in order/disorder transitions, where formation of ordered segments promotes the formation of other ordered segments and vice versa. Following this line of thought, proteins in the "Twilight Zone" might reflect folding intermediates on a transition path from unfolded to folded. The seemingly cooperative nature of disorder also provides a clue to why it has been so difficult to construct predictive models for disorder from local amino acid composition alone. It is at current difficult to address whether the formation of medium length segments of alternating order/disorder are due to a cooperative transition together with nearest neighbors or due to properties of the amino acids in the segment.

The analysis of the CheZOD database revealed that entries found in the DisProt database were often mostly structured. The higher degree of structure for these protein entries could be due to differences between the sample conditions during analysis corresponding to the BMRB and DisProt entries. For example, the NMR analysis corresponding to the BMRB entry could be performed under conditions that favor structure determination. Alternatively, one could speculate that the identification of disordered regions, leading to the inclusion in the DisProt database, could be based on methods that are less strict compared to NMR or more "coarse-grained," in the sense that they only provide a classification for the full protein and not at a local level. In support of this, we found that DisProt entries, which used

NMR to assess disorder, were also more often confirmed to be disordered in our analysis. Hence, a more critical assessment of disorder based on a more reliable and uniform criterion, such as can be derived with NMR, is recommended.

Implications for the Bioinformatics Analysis of Protein Disorder

The last years have seen a tremendous increase in the number of bioinformatics tools that try to predict protein (dis)order (Schlessinger et al., 2009; Dosztányi et al., 2010) and several prediction methods participated in recent rounds of the Critical Assessment of Structure Predictions (CASP; Noivirt-Brik et al., 2009; Monastyrskyy et al., 2011). Unfortunately, current datasets of experimentally classified ordered and disordered regions (Sickmeier et al., 2007) contain many misclassified segments: In X-ray crystal structures regions may appear ordered due to binding partners or crystal packing forces, and could be disordered in isolation. In disorder databases segments may even be more prone to misclassification, since many longer disordered regions are characterized by semi-quantitative experiments that lack position specific information. This suspicion of misclassification was confirmed in our analysis of the validated entries from the DisProt database where several entries where almost completely structured in the CheZOD database. Furthermore, the order/disorder status is also sensitive to environmental conditions, and this fact is not considered. The lack of sufficiently reliable datasets and the noise in the assignment of order and disorder represent a serious limitation in developing accurate prediction methods for protein disorder (Dosztányi et al., 2010). The final, serious shortcoming, of current prediction methods is their inaccuracy when going down to shorter stretches (preliminary analysis). This study reveals a preponderance of proteins with mixed ordered and disordered segments and high sequence disorder complexity—typically, for proteins with mixed order, $C_{SD} \approx 0.1$ corresponding to an average segment length of ca. 10 residues.

The CheZOD database presented here was carefully manually curated to exclude any entries with biasing conditions and strictly contains proteins under native conditions. Here the analysis and processing of chemical shifts provides a unique experimentally validated local and quantitative measure of order/disorder. Furthermore, our database covers a broad range of proteins ranging from completely ordered to almost completely structureless. Since predictions in CASP9 were no better than those in CASP8 (Monastyrskyy et al., 2011), and only a small improvement was noted for CASP10 (Monastyrskyy et al., 2014), we hope that our curated CheZOD database can help form the basis for the development of even more accurate and sophisticated predictive models of order/disorder. This will enable us to ask more detailed questions and provide answers to complex biologically relevant problems related to intrinsically disordered proteins.

CONCLUSIONS

We have used a systematic analysis of NMR chemical shifts to build a database of experimentally validated disordered proteins, the CheZOD database. In contrast to other methods for order classification, our procedure provides a reliable position-specific quantitative measure of order/disorder through our *Chemical shift Z-score for assessing Order/Disorder* (the *CheZOD score*). Examples were observed of both maximum CheZOD score in completely ordered segments, intermediate values in loops and fractionally populated structure, and small values in completely disordered regions. Careful inspection and systematic analysis of the entries in the CheZOD database revealed interesting trends and variations. In particular, it was discussed here how proteins can be completely disordered, partially disordered, or only disordered in a small segment. Some proteins can indeed be classified as *super unfolded*, like human α-synuclein, indicating that this protein is not an archetypal IDP. Through the introduction of the sequence disorder complexity we found diverse patterns of disorder, e.g., that proteins can be segregated into two distinct parts of an ordered and a disordered domain, but also be composed of smaller segments varying alternatingly between order and disorder. A typical segment length of ordered/disordered residues was estimated to be ca. 10 residues. We foresee that further systematic analysis of our CheZOD database will contribute to a more detailed understanding of the relationship between primary sequence and disorder/structure and function.

AUTHOR CONTRIBUTIONS

Both authors JTN and FM contributed to designing and developing the project and writing the paper. JTN performed the data collection, processing, and analysis by building the database of (intrinsically disordered) proteins with assigned chemical shifts, calculating the chemical shift Z-score indicative of disorder, and analyzing the trends in the database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmolb.2016.00004

REFERENCES

Ahmed, M. A., Bamm, V. V., Harauz, G., and Ladizhansky, V. (2007). The BG21 isoform of Golli myelin basic protein is intrinsically disordered with a highly flexible amino-terminal domain. *Biochemistry* 46, 9700–9712. doi: 10.1021/bi700632x

Akaike, H. (1974). A new look at the statistical model identification. *Automat. Control IEEE Trans.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Akaike, H. (1985). “Prediction and entropy,” in *A Celebration of Statistics,* eds A. Atkinson and S. Fienberg (New York, NY: Springer), 1–24.

Ampapathi, R. S., Creath, A. L., Lou, D. I., Craft, J. W. Jr., Blanke, S. R., and Legge, G. B. (2008). Order-disorder-order transitions mediate the activation of cholera toxin. *J. Mol. Biol.* 377, 748–760. doi: 10.1016/j.jmb.2007.12.075

Berjanskii, M. V., and Wishart, D. S. (2005). A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.* 127, 14970–14971. doi: 10.1021/ja054842f

Bermel, W., Bertini, I., Felli, I. C., Lee, Y. M., Luchinat, C., and Pierattelli, R. (2006). Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J. Am. Chem. Soc.* 128, 3918–3919. doi: 10.1021/ja0582206

Camilloni, C., De Simone, A., Vranken, W. F., and Vendruscolo, M. (2012). Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* 51, 2224–2231. doi: 10.1021/bi3001825

Canal, L. (2005). A normal approximation for the chi-square distribution. *Comput. Stat. Data Anal.* 48, 803–808. doi: 10.1016/j.csda.2004.04.001

De Biasio, A., Ibáñez de Opakua, A., Cordeiro, T. N., Villate, M., Merino, N., Sibille, N., et al. (2014). p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. *Biophys. J.* 106, 865–874. doi: 10.1016/j.bpj.2013.12.046

Dosztányi, Z., Mészáros, B., and Simon, I. (2010). Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief. Bioinformatics* 11, 225–243. doi: 10.1093/bib/bbp061

Dyson, H. J., and Wright, P. E. (2001). Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Meth. Enzymol.* 339, 258–270. doi: 10.1016/S0076-6879(01)39317-5

Dyson, H. J., and Wright, P. E. (2004). Unfolded proteins and protein folding studied by NMR. *Chem. Rev.* 104, 3607–3622. doi: 10.1021/cr030403s

Felli, I. C., and Pierattelli, R. (2014). Novel methods based on (13)C detection to study intrinsically disordered proteins. *J. Magn. Reson.* 241, 115–125. doi: 10.1016/j.jmr.2013.10.020

Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., et al. (2014). IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.* 42, D320–D325. doi: 10.1093/nar/gkt1010

Gourdon, P., Liu, X. Y., Skjørringe, T., Morth, J. P., Møller, L. B., Pedersen, B. P., et al. (2011). Crystal structure of a copper-transporting PIB-type ATPase. *Nature* 475, 59–64. doi: 10.1038/nature10191

Gregersen, N., Bross, P., Vang, S., and Christensen, J. H. (2006). Protein misfolding and human disease. *Annu. Rev. Genomics Hum. Genet.* 7, 103–124. doi: 10.1146/annurev.genom.7.080505.115737

Hwang, P. M., Cai, F., Pineda-Sanabria, S. E., Corson, D. C., and Sykes, B. D. (2014). The cardiac-specific N-terminal region of troponin I positions the regulatory domain of troponin C. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14412–14417. doi: 10.1073/pnas.1410775111

Isaksson, L., Mayzel, M., Saline, M., Pedersen, A., Rosenlöw, J., Brutscher, B., et al. (2013). Highly efficient NMR assignment of intrinsically disordered proteins: application to B- and T cell receptor domains. *PLoS ONE* 8:e62947. doi: 10.1371/journal.pone.0062947

Jensen, M. R., Ruigrok, R. W., and Blackledge, M. (2013). Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.* 23, 426–435. doi: 10.1016/j.sbi.2013.02.007

Kelker, M. S., Dancheck, B., Ju, T., Kessler, R. P., Hudak, J., Nairn, A. C., et al. (2007). Structural basis for spinophilin-neurabin receptor interaction. *Biochemistry* 46, 2333–2344. doi: 10.1021/bi602341c

Kim, J. H., Tonelli, M., and Markley, J. L. (2012). Disordered form of the scaffold protein IscU is the substrate for iron-sulfur cluster assembly on cysteine desulfurase. *Proc. Natl. Acad. Sci. U.S.A.* 109, 454–459. doi: 10.1073/pnas.1114372109

Kjaergaard, M., and Poulsen, F. M. (2011). Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. *J. Biomol. NMR* 50, 157–165. doi: 10.1007/s10858-011-9508-2

Konrat, R. (2014). NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J. Magn. Reson.* 241, 74–85. doi: 10.1016/j.jmr.2013.11.011

Kragelj, J., Ozenne, V., Blackledge, M., and Jensen, M. R. (2013). Conformational propensities of intrinsically disordered proteins from NMR chemical shifts. *Chemphyschem* 14, 3034–3045. doi: 10.1002/cphc.201300387

Libich, D. S., Monette, M. M., Robertson, V. J., and Harauz, G. (2007). NMR assignment of an intrinsically disordered protein under physiological conditions: the 18.5 kDa isoform of murine myelin basic protein. *Biomol. NMR Assign.* 1, 61–63. doi: 10.1007/s12104-007-9016-1

Marsh, J. A., Singh, V. K., Jia, Z., and Forman-Kay, J. D. (2006). Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci.* 15, 2795–2804. doi: 10.1110/ps.062465306

Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., and Kryshtafovych, A. (2011). Evaluation of disorder predictions in CASP9. *Proteins* 79(Suppl. 10), 107–118. doi: 10.1002/prot.23161

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., and Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. *Proteins* 82, 127–137. doi: 10.1002/prot.24391

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Nielsen, J. T., Eghbalnia, H. R., and Nielsen, N. C. (2012). Chemical shift prediction for protein structure calculation and quality assessment using an optimally parameterized force field. *Prog. Nucl. Magn. Reson. Spectrosc.* 60, 1–28. doi: 10.1016/j.pnmrs.2011.05.002

Noivirt-Brik, O., Prilusky, J., and Sussman, J. L. (2009). Assessment of disorder predictions in CASP8. *Proteins* 77(Suppl. 9), 210–216. doi: 10.1002/prot.22586

Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., et al. (2013). D2P2: database of disordered protein predictions. *Nucleic Acids Res.* 41, D508–D516. doi: 10.1093/nar/gks1226

Potenza, E., Domenico, T. D., Walsh, I., and Tosatto, S. C. E. (2014). MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* 43, D315–D320. doi: 10.1093/nar/gku982

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., and Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE* 4:e4433. doi: 10.1371/journal.pone.0004433

Shen, Y., and Bax, A. (2010). SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* 48, 13–22. doi: 10.1007/s10858-010-9433-9

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., et al. (2007). DisProt: the database of disordered proteins. *Nucleic Acids Res.* 35, D786–D793. doi: 10.1093/nar/gkl893

Singarapu, K. K., Tonelli, M., Chow, D. C., Frederick, R. O., Westler, W. M., and Markley, J. L. (2011). Structural characterization of Hsp12, the heat shock protein from *Saccharomyces cerevisiae*, in aqueous solution where it is intrinsically disordered and in detergent micelles where it is locally alpha-helical. *J. Biol. Chem.* 286, 43447–43453. doi: 10.1074/jbc.M111.306464

Tamiola, K., Acar, B., and Mulder, F. A. A. (2010). Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J. Am. Chem. Soc.* 132, 18000–18003. doi: 10.1021/ja105656t

Tompa, P. (2009). *Structure and Function of Intrinsically Disordered Proteins*. Boca Raton, FL: CRC Press.

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., et al. (2008). BioMagResBank. *Nucleic Acids Res.* 36, D402–D408. doi: 10.1093/nar/gkm957

Uversky, V. N. (2003). Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cell. Mol. Life Sci.* 60, 1852–1871. doi: 10.1007/s00018-003-3096-6

Uversky, V. N., and Longi, S. (2010). *Instrumental Analysis of Intrinsically Disordered Proteins: Assessing Structure and Conformation*. New York, NY: John Wiley & Sons.

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* 37, 215–246. doi: 10.1146/annurev.biophys.37.032807.125924

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., et al. (2014). Classification of intrinsically disordered regions and proteins. *Chem. Rev.* 114, 6589–6631. doi: 10.1021/cr400525m

Vucetic, S., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2003). Flavors of protein disorder. *Proteins Struct. Funct. Genet.* 52, 573–584. doi: 10.1002/prot.10437

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., et al. (2005). DisProt: a database of protein disorder. *Bioinformatics* 21, 137–140. doi: 10.1093/bioinformatics/bth476

Wang, L., Eghbalnia, H. R., Bahrami, A., and Markley, J. L. (2005). Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. *J. Biomol. NMR* 32, 13–22. doi: 10.1007/s10858-005-1717-0

Wang, Y., and Jardetzky, O. (2002). Investigation of the neighboring residue effects on protein chemical shifts. *J. Am. Chem. Soc.* 124, 14075–14084. doi: 10.1021/ja026811f

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337, 635–645. doi: 10.1016/j.jmb.2004.02.002

Wishart, D. S., and Case, D. A. (2001). Use of chemical shifts in macromolecular structure determination. *Meth. Enzymol.* 338, 3–34. doi: 10.1016/S0076-6879(02)38214-4

Wood, K., Paz, A., Dijkstra, K., Scheek, R. M., Otten, R., Silman, I., et al. (2012). Backbone and side chain NMR assignments for the intrinsically disordered cytoplasmic domain of human neuroligin-3. *Biomol. NMR Assign.* 6, 15–18. doi: 10.1007/s12104-011-9315-4

Wu, J., Peng, D., Voehler, M., Sanders, C. R., and Li, J. (2013). Structure and expression of a novel compact myelin protein - small VCP-interacting protein (SVIP). *Biochem. Biophys. Res. Commun.* 440, 173–178. doi: 10.1016/j.bbrc.2013.09.056

Zhang, H., Neal, S., and Wishart, D. S. (2003). RefDB: a database of uniformly referenced protein chemical shifts. *J. Biomol. NMR* 25, 173–195. doi: 10.1023/A:1022836027055

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.

The reviewer WV and Handling Editor declared their shared affiliation, and the Handling Editor states that the process nevertheless met the standards of a fair and objective review.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on