frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 29 March 2023
DOI 10.3389/fgene.2023.1134509

Check for updates

OPEN ACCESS

EDITED BY
Tarun Jairaj Narwani,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

REVIEWED BY
Gauri Misra,
National Institute of Biologicals, India
Julien Diharce,
Université de Paris, France

*CORRESPONDENCE
R. Sowdhamini,
✉ mini@ncbs.res.in

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 30 December 2022
ACCEPTED 06 March 2023
PUBLISHED 29 March 2023

CITATION
Chauhan PK and Sowdhamini R (2023), Computational analysis of the flexibility in the disordered linker region connecting LIM domains in cysteine–glycine-rich protein.
Front. Genet. 14:1134509.
doi: 10.3389/fgene.2023.1134509

Computational analysis of the flexibility in the disordered linker region connecting LIM domains in cysteine–glycine-rich protein

Pankaj Kumar Chauhan[1] and R. Sowdhamini[1,2,3]*

[1]National Centre for Biological Sciences Tata Institute of Fundamental Research, Bangalore Karnataka, India, [2]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India, [3]Institute of Bioinformatics and Applied Biotechnology, Bangalore, India

One of the key proteins that are present in the Z-disc of cardiac tissues, CSRP3, has been implicated in dilated and hypertrophic cardiomyopathy leading to heart failure. Although multiple cardiomyopathy-related mutations have been reported to reside on the two LIM domains and the disordered regions connecting the domains in this protein, the exact role of the disordered linker region is not clear. The linker harbors a few post-translational modification sites and is expected to be a regulatory site. We have carried out evolutionary studies on 5614 homologs spanning across taxa. We also performed molecular dynamics simulations of full-length CSRP3 to show that the length variations and conformational flexibility of the disordered linker could provide additional levels of functional modulation. Finally, we show that the CSRP3 homologs with widely different lengths of the linker regions could display diversity in their functional specifications. The present study provides a useful perspective to our understanding of the evolution of the disordered region between CSRP3 LIM domains.

KEYWORDS

CSRP3, disordered region, cardiomyopathy, PPI, molecular dynamics

1 Introduction

The cysteine and glycine-rich protein (CSRP) gene belongs to a family of proteins, vitally involved in basic processes such as differentiation, growth, and gene regulation (Weiskirchen et al., 1995). Three types of this protein, CSRP1, CSRP2, and CSRP3, have been noted in three genomic locations in the human genome: chromosome 1, 10, and 11, respectively. CSRP1 is implicated in stress response by binding to zyxin (Schmeichel and Beckerle, 1998), while CSRP2 is implicated in leukemia (Wang et al., 2017).

The gene CSRP3, or muscle LIM protein (MLP) (HGNC:24722; chr. location: 11p15.1), is organized into six exons encoding a 194 amino acid long protein (Fung et al., 1996; Knöll et al., 2002). CRSP3 protein contains two LIM domains connected by a disordered region rich in glycyl and prolyl residues and is highly flexible. Each LIM domain, in turn, possesses two Zn-finger subdomains, each one having specific binding partners including transcription factors such as MyoD, MRF4, GATA4, and SRF (Buyandelger et al., 2011). The LIM domains are also involved in a distinct functional role such as the regulation of the expression of target genes. Alternatively spliced transcript variants with different 5′ UTR, but encoding the same protein, have been found for this gene.

CSRP3 is primarily expressed in the heart and to a lesser extent in prostate tissue of humans. The function of CSRP3 is attributed as a mechanical stretch sensor in the Z-disc complex of the heart (Buyandelger et al., 2011). A significant reduction in CSRP3 protein levels was reported in dilated and ischemic cardiomyopathy patients leading to heart failure (Zolk et al., 2000). In addition, W4R or C58G point mutations in this gene implicated hypertrophic cardiomyopathy and heart failure (Knöll et al., 2002; Knoll et al., 2010; Ehsan et al., 2018). Hence, mutations in this gene are thought to cause heritable forms of hypertrophic cardiomyopathy (HCM) and dilated cardiomyopathy (DCM) in humans. As of today, close to 20 mutations have been realized on CSRP3 and most of them are found to have deleterious effects. We had recently conducted a virtual saturation mutagenesis to understand the structural effects of amino acid residues and positions (Chauhan and Sowdhamini, 2022). Here, we also showed that C-terminal LIM exhibits higher conservation compared to N-terminal LIM. A study by Hoffmann *et al.* demonstrated that N-terminal LIM is involved in self-association of CSRP3 while C-terminal LIM exhibits direct interaction with actin filaments (AFs) and stabilizes AFs, cross-linking them into bundles (Hoffmann et al., 2014). This suggests that C-terminal LIM is relatively more important in AF stability. CSRP3 oligomerizes *via* the N-terminal LIM domain that seems to be initiated by post-translational modification, in particular the O-glycosylation, as shown in a study by Chiaki Nagai-Okatani and Naoto Minamino (Nagai-Okatani ¤a and Minamino, 2016). The study also postulates six O-glycosylation sites in the linker region between two LIM domains. The linker region may regulate the protein–protein interaction (PPI). It was demonstrated that the linker region between amino acids 94 to 105 interacts with cofilin 2 (CFL2) (Papalouka et al., 2009). It necessitates a thorough examination of the linker region in homologs to understand its role and evolution.

In this work, we have examined the evolutionary landscape of the CSRP3 gene in other species and report various aspects, such as length variations of the two LIM domains and the connecting linker region. To the best of our knowledge, this is the first study which explores the linker region variation and its taxonomic distribution. Few of the homologs which exhibit anomalous linker lengths have been discussed. All-atom long-length molecular dynamics simulations of the full-length CSRP3 protein reveal spontaneous interactions between the two LIM domains, facilitated by the flexible linker region. The backbone conformational flexibility of the linker region has been investigated. In order to understand the functional variations, we followed the post-translational and protein–protein interaction patterns of human CSRP3 and few homologs of variable lengths to show that such length variations could provide functional versatility in terms of their choice of partner proteins.

2 Materials and methods

2.1 Sequence retrieval and analysis

The query term “CSRP3” was searched in the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein/), and the human CSRP3 protein [accession number: NP_003467.1] FASTA sequence was selected. To obtain CSRP3 protein homologs from different organisms, the NCBI domain enhanced lookup time accelerated BLAST (DELTA-BLAST) (Boratyn et al., 2012) was performed where the previously obtained sequence was used as the query input and nr protein was selected as the database. The DELTA-BLAST parameters of Expect threshold = 0.00005, Matrix = BLOSUM62, gap costs = Existence:13 Gap: 1, compositional adjustments = composition-based statistics, PSI-BLAST threshold = 0.00005, DELTA-BLAST threshold = 0.00005, pseudocount = 0, and max output hits = 20000 were set, and it was run for one iteration. Next, the filter of query coverage >75% and percent identity >30% was applied to the output and sequence IDs were downloaded. The efetch mode of Entrez Programming Utilities (E-utilities) (Rédei, 2008) was utilized to retrieve protein sequences for these IDs programmatically. The domain prediction in these sequences was carried out using the HMMSCAN module of HMMER-3.1 (Eddy, 2008) on the all domains hmm of the PFAM dataset (Mistry et al., 2021). Sequences having query coverage <75% full-length human CSRP3 protein were filtered that were missed during blast filtering, and also, a list of dual-LIM, multi-LIM, and other domain containing IDs was created. The disordered region from dual-LIM sequences was extracted based on LIM1 and LIM2 boundaries. The histogram of length variability in LIM domains and the disordered region across sequences was plotted. The amino acid propensity of the disordered region was calculated as follows:

$$P_{AA_i} = \frac{\sum_{D=1}^{D=N} \mathrm{Count}\,(AA_i) \Big/ \sum_{j=1}^{N} \mathrm{Length}\big(D_j\big)}{\sum_{L_1+D+L_2=1}^{L_1+D+L_2=N} Count\,(AA_i) \Big/ \sum_{j=1}^{N} Length\big(L_{1j} + D_j + L_{2j}\big)}.$$

The P_{AA_i} is the propensity of an amino acid AA_i, N is total number of sequences, D_j is the disordered region of the j th sequence, L_{1j} is the first LIM of the j th sequence, and L_{2j} depicts the second LIM of the j th sequence.

2.2 Homology modeling of human CSRP3 protein and *Arabidopsis* and nematode representative homologs

The protein sequence of human CSRP3 was used for homology-based structure modeling. Next, in order to find a suitable template for modeling, protein BLAST (Camacho et al., 2009) was utilized to search for the nearest structural homolog in the Protein Data Bank (PDB) (Berman et al., 2003). Blast PDB hits, namely, 2O10 (human CSRP3 LIM1), 2O13 (human CSRP3 LIM2), and 1B8T (chicken CSRP1) NMR structures, were used for multi-template modeling of human CSRP3 full-length protein. The first conformer from each structure was used in the template selection. A homology-based structure modeling tool, MODELLER 9.12 (Eswar et al., 2006), was used for CSRP3 structure modeling. Ten homology models were generated and ranked according to the DOPE score. Then, the top five ranked models were assessed for structure validations using SAVES 5.0 (Laskowski et al., 1993) (https://servicesn.mbi.ucla.edu/SAVES/) and the ProSA server (Wiederstein and Sippl, 2007). The best-predicted model assessed based on the DOPE score, Ramachandran plot, and ProSA profile was selected for further structural analysis. In the case of homologs in *Arabidopsis* and nematode genomes, where CSRP3 contains short-length and long-length linker regions, respectively, a similar strategy was adopted. The details of query coverage, percent identity, and templates used in each sequence modeling are provided in Supplementary Table S1.

2.3 Molecular dynamics simulations of model structures

The best-predicted model was minimized at pH 7 using PROPKA from Protein Preparation Wizard in the Maestro package (Schrödinger Release 2020: Maestro, Schrödinger, LLC, New York, NY, 2020). Next, the structure was restrain-minimized using the OPLS3e force field (Harder et al., 2016) and solvated with the TIP3P water system using the System Builder from the Desmond module (Bowers et al., 2006). To account for periodic boundary conditions, the orthorhombic box was used and its volume was minimized by having a buffer distance of 10Å. The aforementioned system was neutralized with 15 CL^- ions, and additional 150 mM NaCl salt was added. The OPLS3e force field was chosen as the force field, and Molecular Dynamics (MD) run was carried out in the Molecular Dynamics package on the output generated by the System Builder. In the MD run, a short relaxation run was initiated with the default protocol on the solvated system generated from previous steps. After relaxation, the actual production MD simulation was accomplished under NPT constraint parameters and the OPLS3e force field. In the simulation, the default settings of the RESPA integrator (Humphreys et al., 1994) (2 femtoseconds time step for bonded or near non-bonded interactions and 6 femtoseconds for far non-bonded interactions) were incorporated. The temperature was kept fixed at 300 K using a nose-Hoover thermostat algorithm (Martyna et al., 1992). Similarly, the pressure (1 bar) was fixed using the Martyna–Tobias–Klein method (Martyna et al., 1994). The production MD simulation was executed for 200 nanoseconds in triple replicates. This protocol was followed for human CSRP3 and homologs from *Arabidopsis* and nematode model structures as representative for trimodal distribution of the linker region.

2.4 MD simulation event analysis

The simulation interaction diagram (SID) and simulation event analysis (SEA) modules of the Desmond package were explored for the MD trajectories analyses. The root mean square distance (RMSD) and root mean structure fluctuation (RMSF) of the protein backbone were calculated for the entire duration of simulation time by aligning them to the reference frame (0th frame) in the SID. The same protocol was followed for the replicates. Similarly, the Radius of Gyration (ROG) and energy change were calculated for each trajectory using the SEA module. The distance density map showing the distance between LIM1 and LIM2 throughout the simulation point was created using the proxy distance between Cys 10–Phe 176 for human CSRP3 and similarly for the *Arabidopsis* and nematode representative. Next, the phi–psi values corresponding to the disordered region were calculated for the structure at 0th frame (0 ns), 500th frame (100 ns), and 1000th frame (200 ns), and an arrow map was generated to highlight the deviation in phi–psi values at these stages. Furthermore, the time-dependent evolution of secondary structure elements (SSEs) along the MD simulation trajectory was generated by the STRIDE package (Frishman and Argos, 1995) in VMD software (Humphrey et al., 1996).

2.5 Post-translational modifications and taxonomy

The post-translational modifications (PTMs) in the disordered region of protein sequences were predicted using the MusiteDeep online server (https://www.musite.net/) (Wang et al., 2020). The PTMs such as phosphorylation, glycosylation, ubiquitination, SUMOylation, acetylation, methylation, palmitoylation, and hydroxylation were considered and score cut-off of 0.8 and 0. 9 were explored. The frequency of each PTM was calculated across the disordered sequences. The taxonomic distribution of sequences was analyzed at the phylum level using ETE toolkit (Huerta-Cepas et al., 2016).

2.6 Protein–protein interactions

The protein–protein interaction analysis was executed on the STRING database server (Szklarczyk et al., 2021). Protein sequences for protein IDs NP_001180500.1 (human CSRP1), NP_001400464.1 (human CSRP2), NP_003467.1 (human CSRP3), TWW71823.1 (sansaifugu CSRP2), KRX47685.1 (Trichinella CSRP2), CAB71053.1 (*Arabidopsis* LIM), and PIN15380.1 (*Handroanthus* MLP) were queried for protein–protein interactions. In case of no match, the next most similar (by % identity) sequence suggested by the server was used as a proxy for PPI.

3 Results

3.1 CSRP3 homolog sequence analysis

The human CSRP3 consists of two LIM domains connected by the 53-residue long disordered linker region (Figure 1). Starting from human CSRP3 (ID: NP_003467.1) as a query, a DELTA-BLAST run (Boratyn et al., 2012) was initiated against the non-redundant sequence database to yield homologs. Homologous sequences include other CSRP members (such as CSRP1 and CSRP2). Hits which are annotated as 'hypothetical' or 'unnamed' were also retained as long as they pass the thresholds on parameters. The sequence search was performed using a strict query coverage filter of 75% and a lower limit of 30% sequence identity such that only homologs that retain two LIM domains connected by a flexible linker region as query could be identified (please see Methods for details). Interestingly, 5614 homologs could be identified across a wide range of taxa, spanning 1404 species (Table 1; Supplementary Figure S1). As expected, the highest extent of homologs is observed in chordates (3211) but could be observed in lower-order organisms such as those who thrive in a freshwater habitat. Two hits were obtained from pathogenic bacterial genomes (MTV28691.1 in *Nitriliruptoraceae bacterium* ZYF776 and WP_254514402.1 in *Salmonella enterica*), which could be a contamination from hosts (please see Discussion).

We next analyzed the extent of length variations within each of the two LIM domains and the connecting flexible disordered region or linker. The distribution of the length of the LIM1 (N-terminal LIM) domain and the LIM2 (second LIM) domain was close to 57 residues, as observed in the human CSRP3. It is similar in other

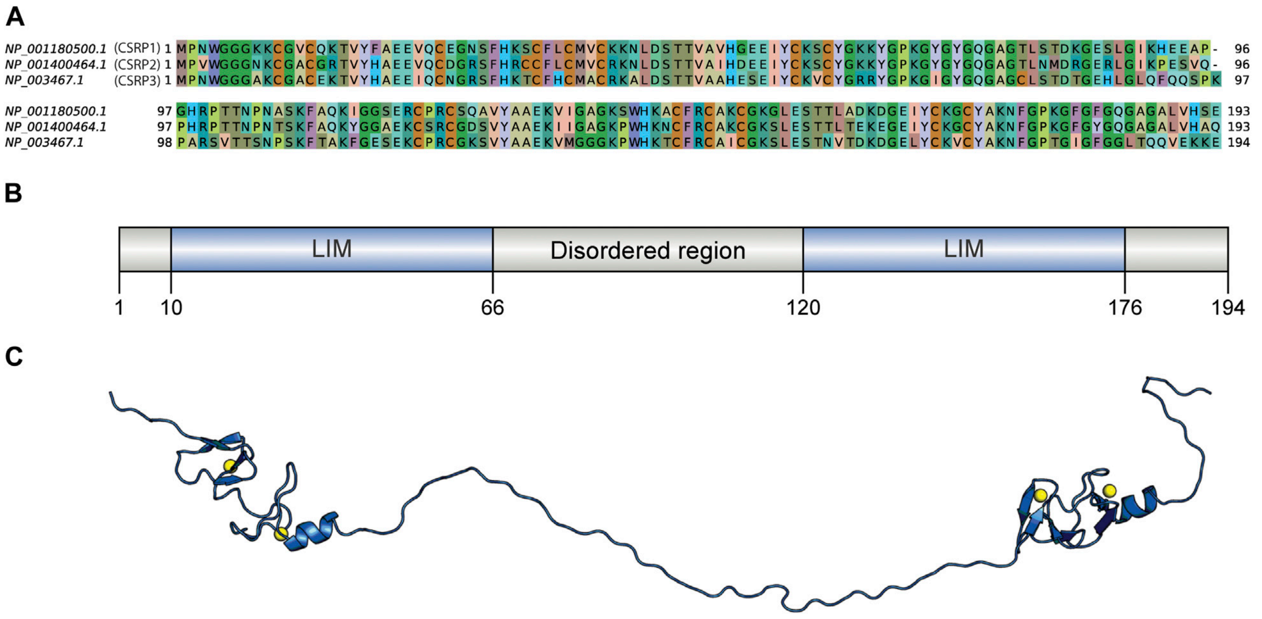

A
NP_001180500.1 (CSRP1) 1 MPNWGGGKKCGVCQKTVYFAEEVQCEGNSFHKSCFLCMVCKKNLDSTTVAVHGEEIYCKSCYGKKYGPKGYGYGQGAGTLSTDKGESLGIKHEEAP- 96
NP_001400464.1 (CSRP2) 1 MPVWGGGNKCGACGRTVYHAEEVQCDGRSFHRCCFLCMVCRKNLDSTTVAIHDEEIYCKSCYGKKYGPKGYGYGQGAGTLNMDRGERLGIKPESVQ- 96
NP_003467.1 (CSRP3) 1 MPNWGGGAKCGACEKTVYHAEEIQCNGRSFHKTCFHCMACRKALDSTTVAAHESEIYCKVCYGRRYGPKGIGYGQGAGCLSTDTGEHLGLQFQQSPK 97
NP_001180500.1 97 GHRPTTNPNASKFAQKIGGSERCPRCSQAVYAAEKVIGAGKSWHKACFRCAKCGKGLESTTLADKDGEIYCKGCYAKNFGPKGFGFGQGAGALVHSE 193
NP_001400464.1 97 PHRPTTNPNTSKFAQKYGGAEKCSRCGDSVYAAEKIIGAGKPWHKNCFRCAKCGKSLESTTLTEKEGEIYCKGCYAKNFGPKGFGYGQGAGALVHAQ 193
NP_003467.1 98 PARSVTTSNPSKFTAKFGESEKCPRCGKSVYAAEKVMGGGKPWHKTCFRCAICGKSLESTNVTDKDGELYCKVCYAKNFGPTGIGFGGLTQQVEKKE 194
B
LIM
Disordered region
LIM
1
10
66
120
176
194
C

FIGURE 1
Sequence comparison of human CSRPs homologs and the CSRP3 model structure. **(A)** Sequence alignment of CSRP1, CSRP2, and CSRP3 proteins. **(B)** Domain architecture of human CSRP3 as predicted by the PFAM database. It consists of two LIM domains connected by a disordered linker region. **(C)** Full length model structure of human CSRP3 modeled by a multi-template homology-based approach.

CRPs (CSRP1 and CSRP2) (please see Supplementary Table S2). We expected a good amount of deviation in the length of the linker region. Although the linker region is 53 amino acids long in human CSRP3, we find an interesting trimodal distribution of the linker length amongst homologs (Figure 2), with the highest peak around 52 residues. In particular, we examined those homologs which have either a longer or a shorter linker length (please see Methods for thresholds). These are in detail in Supplementary Table S3. In general, the number of homologs with unusual linker lengths is few—5 longer lengths and five shorter lengths (Supplementary Table S3). There are three CSRP sequences with longer flexible linkers from Bdelloidea (*Rotaria magna-calcarata* and *Rotaria socialis*) which thrive in freshwater. These are small microscopic animals first reported in mainland France and reproduce asexually. From our analysis, the other species where CSRP homologs retain longer-length linkers are from pufferfish and roundworm (please see Figure 3 for an alignment). From our dataset of CSRP homologs, we observe hits with appreciably shorter linkers from *Arabidopsis* and *Handroanthus impetiginosus* (pink trumpet tree; please see Figure 3 for alignment). Four others of this kind are from a variety of species—Hoatzin bird, Eurasian otter, ray-finned fish, and zebra shark. We mapped the homologs at the class level taxa and calculated the frequency of the trimodal category of a linker in each taxa. Our analysis shows that, in general, plants have a shorter linker (~1950 homologs) while nematodes have a longer linker (29 homologs). The linker region variation at the class taxa level is highlighted in Figure 4. It is not clear how the diversity in linker lengths is likely to affect their functions. In addition, one might expect the linker region to be relatively flexible in length variation and amino acid composition.

Indeed, we calculated the propensity of 20 different amino acids to be present in the linker regions of CSRP homologs in our dataset, in comparison to their presence in the three regions, namely, LIM1 and LIM2 domains and the linker region, using standard calculations (please see Methods). We observe that amino acids such as proline, glutamine, serine, asparagine, and threonine show higher propensity at the flexible linker region. Although glycyl residues are also high in this disordered region and play an important role in sampling conformational space (please see Figure 2D), they are prevalent throughout and frequently observed within the LIM domains as well.

3.2 Structural analysis of human CSRP3

To understand the dynamics of the disordered linker region and LIM domains, we carried out Molecular Dynamics simulations of the full-length CSRP3 protein structure. MD simulations helps us in understanding conformation changes and interaction properties in the protein over a period of time at a given temperature and pressure. In the absence of a full-length protein structure, a homology-based model of CSRP3 protein was modeled using MODELLER 9.12 (Eswar et al., 2006). A total of 10 models were generated, and the top five DOPE scoring models were validated using SAVES 5.0 (Laskowski et al., 1993) and the ProSA server (Wiederstein and Sippl, 2007). The best model was chosen for the MD analysis for 200 ns simulation interval in triplicates using the Schrödinger Desmond suite (see Material and Methods). The same approach was followed for the *Arabidopsis* and nematode representative homolog sequences.

TABLE 1 List of number of annotated and hypothetical protein IDs observed in the taxonomy at the phylum level.

S. no.	Phylum	Annotated sequences	Hypothetical sequences
1	Actinobacteria	0	1
2	Proteobacteria	0	1
3	Evosea	5	3
4	Chlorophyta	1	0
5	Streptophyta	1424	552
6	Basidiomycota	3	2
7	Zoopagomycota	1	1
8	Mucoromycota	15	11
9	Chytridiomycota	3	12
10	Porifera	2	0
11	Ctenophora	1	0
12	Cnidaria	27	0
13	Hemichordata	2	0
14	Echinodermata	14	0
15	Chordata	3021	190
16	Rotifera	10	2
17	Bryozoa	1	0
18	Annelida	1	4
19	Mollusca	45	5
20	Platyhelminthes	37	6
21	Priapulida	1	0
22	Arthropoda	148	10
23	Nematoda	32	1
24	Unclassified	12	1
	Total	**4806**	**802**

The MD trajectory analysis demonstrated that there is a large structural change in CSRP3 conformation with respect to initial conformer as evident from the statistical parameter such as the root mean square distance, root mean structure fluctuation, and Radius of Gyration analyses (Supplementary Figure S2). The ROG plot shows that the structure is stabilized in ~40 ns time interval; however, small fluctuations are observed in the RMSD plot that is eventually stabilized around 120 ns. A similar trend was observed in other two replicates (see Supplementary Figure S2 and S3). The *Arabidopsis* and nematode model structure MD simulation exhibited similar properties (Supplementary Figure S3). To further understand this behavior, we assessed the distance between two LIM domains as a function of time. The two LIM domains in human CSRP3 come together (from 120 Å to 35 Å) within ~75 ns time and remain in close proximity throughout 200 ns simulation time (Figure 5). We carried out this LIM domain distance measurement in a shorter linker representative sequence (*Arabidopsis*) model structure and longer linker representative sequence (nematode) model structure. Although the *Arabidopsis* model structure showed a similar pattern as to human CSRP3 over a period of simulation time, nematode was a bit deviant. In this, two LIM domains seem to move away at ~125 ns time. However, this trend is not consistent in all the replicates. The third replicate does not show movement of LIM domains once they have come together. This suggests that CSRP3 homologs with longer-length linkers might have lesser chance for the N- and C-terminal LIM domains to interact. This points to the potential role of the disordered linker region in allowing two domains to come together for a function or adapt according to interacting partners. We also investigated the secondary structural changes in the disordered region (phi–psi values). The arrow map shows that there is a dramatic change in phi–psi (φ-ψ) values from the 0th frame to 100 ns frame. The φ-ψ values were shifted from the β-sheet region to the helix region. However, there is a very small alteration from the 100 ns

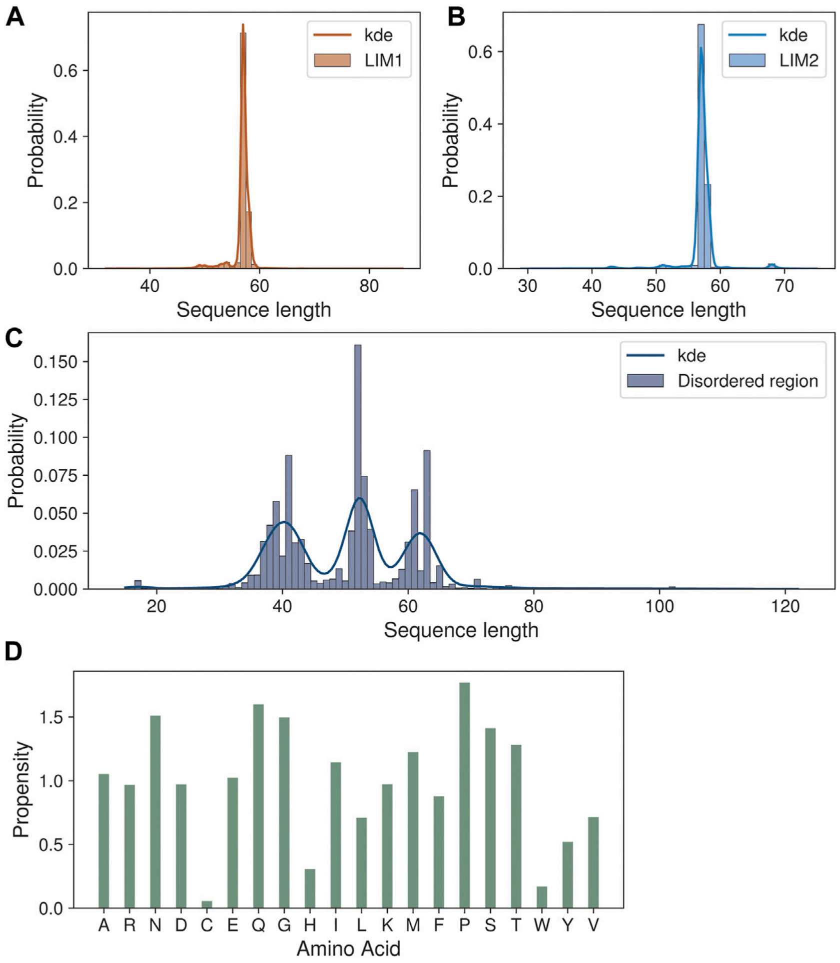


FIGURE 2
Sequence length variations in the CSRP3 homologs. **(A,B)** Histogram showing the length variation in LIM1 and LIM2 domains. **(C)** Trimodal distribution of disordered region length variation seen in the histogram plot. **(D)** Barplot highlighting the amino acid propensity in the disordered region.

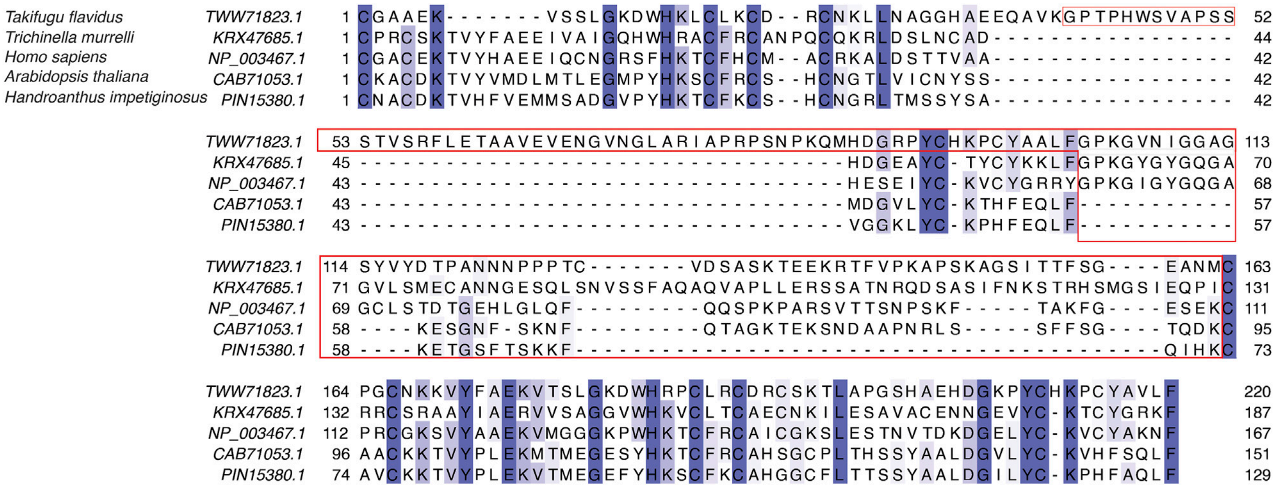

Takifugu flavidus TWW71823.1 1 CGAAEK-------VSSLGKDWHKLCLKCD--RCNKLLNAGGHAEEQAVKGPTPHWSVAPSS 52
Trichinella murrelli KRX47685.1 1 CPRCSKTVYFAEEIVAIGQHWHRACFRCANPQCQKRLDSLNCAD----------------- 44
Homo sapiens NP_003467.1 1 CGACEKTVYHAEEIQCNGRSFHKTCFHCM--ACRKALDSTTVAA----------------- 42
Arabidopsis thaliana CAB71053.1 1 CKACDKTVYVMDLMTLEGMPYHKSCFRCS--HCNGTLVICNYSS----------------- 42
Handroanthus impetiginosus PIN15380.1 1 CNACDKTVHFVEMMSADGVPYHKTCFKCS--HCNGRLTMSSYSA----------------- 42
TWW71823.1 53 STVSRFLETAAVEVENGVNGLARIAPRPSNPKQMHDGRPYCHKPCYAALFGPKGVNIGGAG 113
KRX47685.1 45 -----------------------------------HDGEAYC-TYCYKKLFGPKGYGYGQGA 70
NP_003467.1 43 -----------------------------------HESEIYC-KVCYGRRYGPKGIGYGQGA 68
CAB71053.1 43 -----------------------------------MDGVLYC-KTHFEQLF----------- 57
PIN15380.1 43 -----------------------------------VGGKLYC-KPHFEQLF----------- 57
TWW71823.1 114 SYVYDTPANNNPPPTC--------VDSASKTEEKRTFVPKAPSKAGSITTFSG----EANMC 163
KRX47685.1 71 GVLSMECANNGESQLSNVSSFAQAQVAPLLERSSATNRQDSASIFNKSTRHSMGSIEQPIC 131
NP_003467.1 69 GCLSTDTGEHLGLQF----------QQSPKPARSVTTSNPSKF-----TAKFG----ESEKC 111
CAB71053.1 58 ----KESGNF-SKNF----------QTAGKTEKSNDAAPNRLS-----SFFSG----TQDKC 95
PIN15380.1 58 ----KETGSFTSKKF---QIHKC 73
TWW71823.1 164 PGCNKKVYFAEKVTSLGKDWHRPCLRCDRCSKTLAPGSHAEHDGKPYCHKPCYAVLF 220
KRX47685.1 132 RRCSRAAYIAERVVSAGGVWHKVCLTCAECNKILESAVACENNGEVYC-KTCYGRKF 187
NP_003467.1 112 PRCGKSVYAAEKVMGGGKPWHKTCFRCAICGKSLESTNVTDKDGELYC-KVCYAKNF 167
CAB71053.1 96 AACKKTVYPLEKMTMEGESYHKTCFRCAHSGCPLTHSSYAALDGVLYC-KVHFSQLF 151
PIN15380.1 74 AVCKKTVYPLEKVTMEGEFYHKSCFKCAHGGCFLTTSSYAALDGILYC-KPHFAQLF 129

FIGURE 3
Sequence alignment of LIM domains and the disordered linker region in human CSRP3 (NP_003467.1), sansaifugu CSRP2 (TWW71823.1), nematode CSRP2 KRX47685.1, *Arabidopsis* LIM domain protein (CAB71053.1), and *Handroanthus* MLP (PIN15380.1).

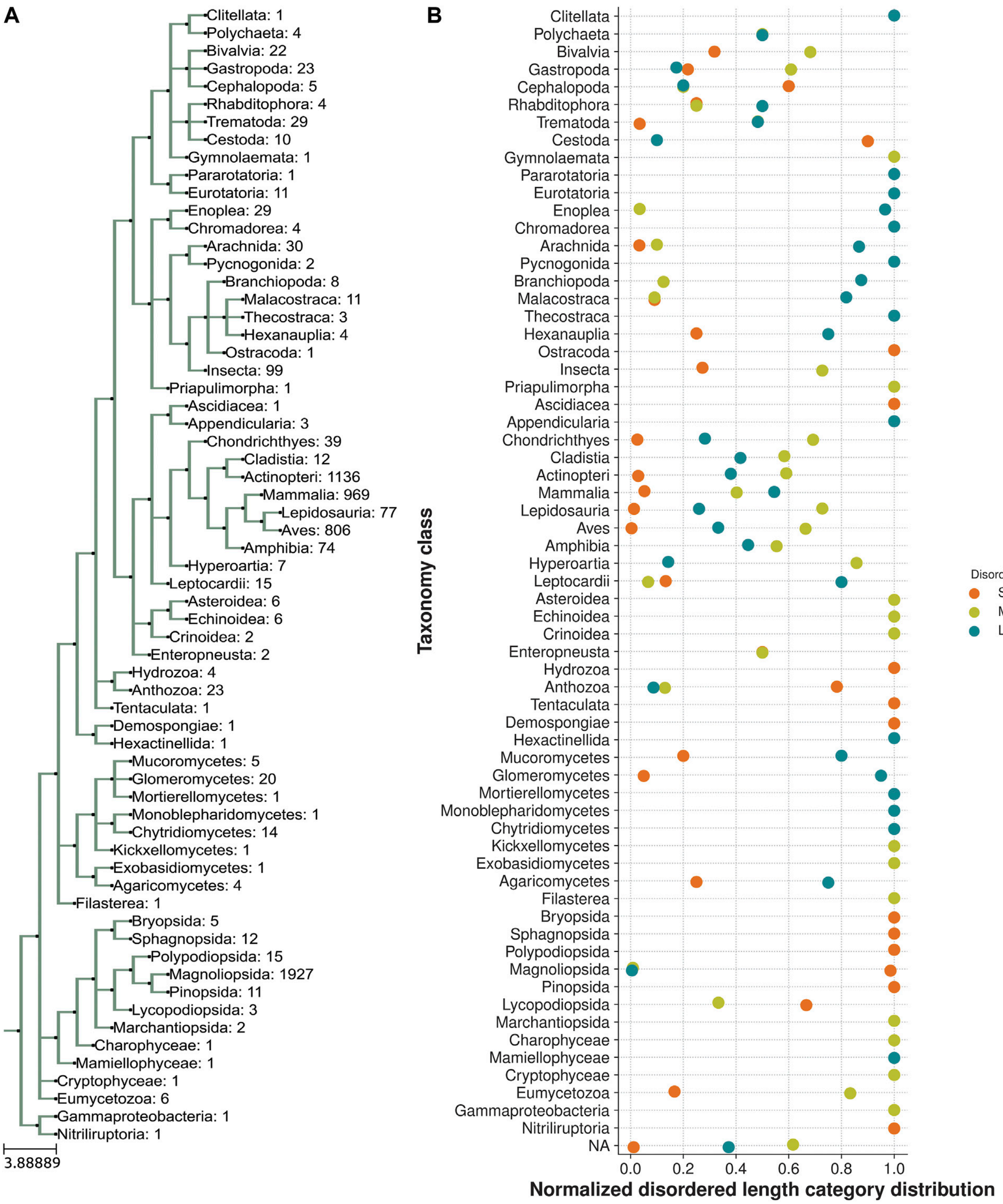


FIGURE 4
Taxonomic distribution of a trimodal disordered linker region. **(A)** Class taxa level phylogeny of CSRP3 homologs depicting the number of sequences in each class. **(B)** Normalized frequency of trimodal distribution of the linker length in each class.

frame to 200 ns frame (Supplementary Figure S4). Since the previous result point to difference in φ-ψ values in the disordered region, secondary structure analysis of full-length structure was carried out in the STRIDE package (Frishman and Argos, 1995) in VMD (Humphrey et al., 1996). The analysis revealed that LIM domain secondary structures are intact in throughout the 200 ns timeline. In addition, the disordered region remained as a coil–coil structure and there is no dramatic change due to φ-ψ changes seen in the previous result (Supplementary Figure S5).

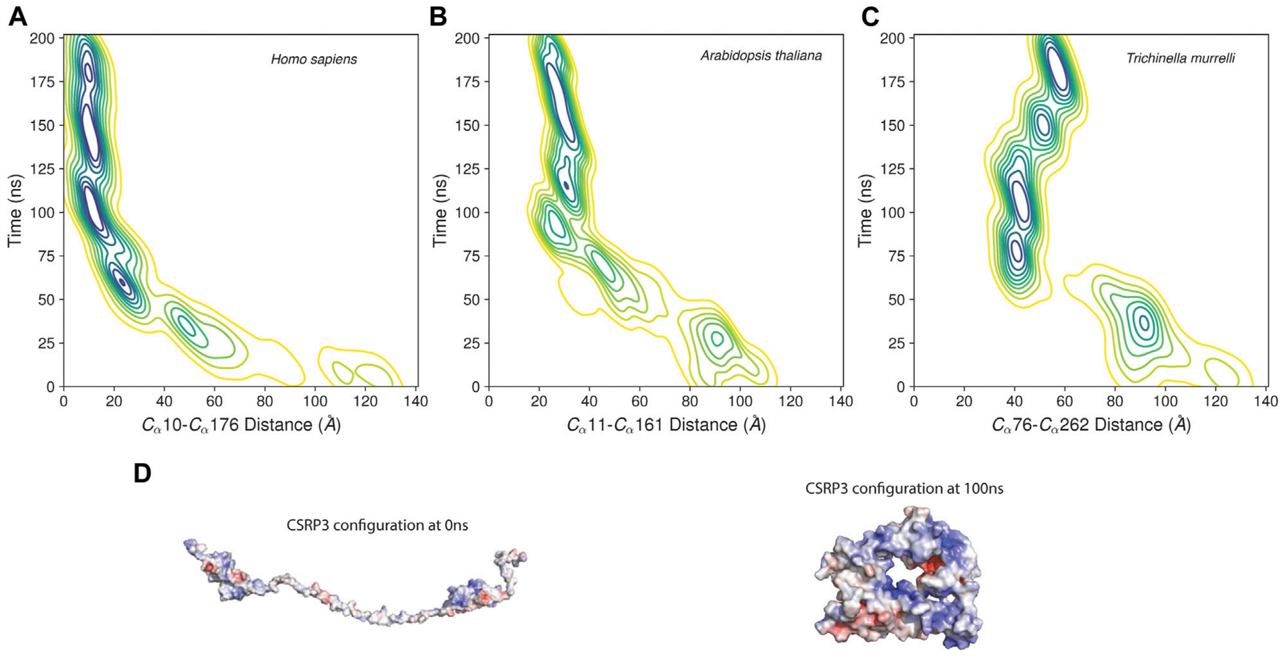
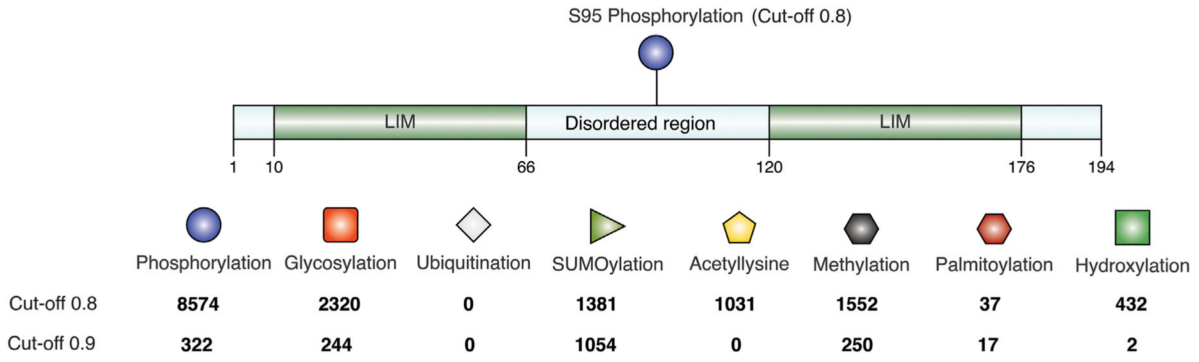


FIGURE 5
Distance between LIM1 and LIM2 as function of time in MD simulation. **(A–C)** Contour map of distance between two LIM domains of human CSRP3, *Arabidopsis* LIM domain protein, and nematode CSRP2. The distance is measured between alpha C atoms of Cys and Phe residues marking the beginning and end of N-terminal LIM and C-terminal LIM domains. **(D)** Electrostatic map of CSRP3 at the 0th frame and 100 ns frame.



	Phosphorylation	Glycosylation	Ubiquitination	SUMOylation	Acetyllysine	Methylation	Palmitoylation	Hydroxylation
Cut-off 0.8	**8574**	**2320**	**0**	**1381**	**1031**	**1552**	**37**	**432**
Cut-off 0.9	**322**	**244**	**0**	**1054**	**0**	**250**	**17**	**2**

FIGURE 6
Post-translational modifications in the disordered region. The frequency is calculated at score of 0.8 and 0.9 predicted by MusiteDeep. PTMs like phosphorylation, glycosylation, ubiquitination, SUMOylation, acetylation, methylation, palmitoylation, and hydroxylation are labeled.

3.3 CSRP3 homolog functional analysis

As seen in the previous sequence analysis results, the disordered linker region varies across CSRP3 homologs in species. Further MD analysis showed that the disordered region provides flexibility to LIM domains for dynamic adaptation. It is reported that post-translational modifications significantly change the conformer of intrinsically disordered proteins (Bah and Forman-Kay, 2016). Chiaki Nagai-Okatani and Naoto Minamino had demonstrated the functional role of O-glycosylation in the linker region of CSRP3 (Nagai-Okatani ¤a and Minamino, 2016). They highlighted that O-glycosylation plays an important role in oligomerization, and this ration is altered in a disease condition as compared to the normal condition. They postulated ~ six glycosylation sites in the human CSRP3 linker region. We carried out the TPMs prediction in the disordered region of homologs on the MusiteDeep online server (https://www.musite.net/) (Wang et al., 2020). All the major PTMs such as phosphorylation, glycosylation, ubiquitination, SUMOylation, acetylation, methylation, palmitoylation, and hydroxylation were chosen during the prediction. Two different score cut-off values (0.

8 and 0.9) were used to compare the frequency of these PTMs (Figure 6). In the human CSRP3 disordered region, only one PTM (phosphorylation of Ser 95) was predicted at the 0.8 cut-off value. Furthermore, at the cut-off value 0.9, it was absent. The O-linked_glycosylation was found to be insignificant and unreliable in human CSRP3 and needs to further investigated. In total, 8574 phosphorylation, 2320 glycosylation, 0 ubiquitination, 1381 SUMOylation, 1031 acetylation, 1552 methylation, 37 palmitoylation, and 432 hydroxylation PTMs were predicted across homologs at the score of 0.8. In contrast, 322 phosphorylation, 244 glycosylation, 0 ubiquitination, 1054 SUMOylation, 0 acetylation, 250 methylation, 17 palmitoylation, and two hydroxylation PTMs were seen at the score of 0.9. There was a significant drop in phosphorylation and glycosylation values, but the SUMOylation number remained high at the stringent value suggesting potential role of SUMOylation in the disordered region of CSRP3 homologs. The longest disordered sequence (ID = TWW71823.1 from *Takifugu flavidus*) showed similar PTMs (two phosphorylation sites) as human CSRP3 (ID = NP_003467.1) in the disordered region. Although nematode (*Trichinella murrelli*) CSRP2 (ID = KRX47685.1) was predicted to have N-linked as well as O-linked glycosylation and phosphorylation at multiple sites. Similarly, shorted disordered sequence in *Handroanthus impetiginosus* (ID = PIN15380.1) contained one methylation PTM at one site and *Arabidopsis* sequence (ID = CAB71053.1) exhibited N-linked glycosylation, hydroxylation, and phosphorylation PTMs (see Supplementary Table S4).

As reported previously, disordered sequences are advantageous for a varied range of protein–protein interactions (Fuxreiter et al., 2004; Bah and Forman-Kay, 2016), we searched a protein–protein interaction of highly varying disorder sequences KRX47685.1 (*Trichinella murrelli* CSRP2) and CAB71053.1 (*Arabidopsis* LIM domain protein) as well as human CSRP1, CSRP2, and CSRP3 protein sequences: NP_001180500.1 (human CSRP1), NP_001400464.1 (human CSRP2), and NP_003467.1 (human CSRP3). Human CSRP3 and CSRP1 show 62% sequence identity as the similar disorder region length variation. Similar observation is seen when CSRP3 is compared to CSRP2. The protein–protein interactions were searched using the sequence search option on the STRING database server (Szklarczyk et al., 2021). The similar approach was followed for the nematode sequence (ID = KRX47685.1) as well as *Arabidopsis* sequence (ID = KRX47685.1) three close homologs based on sequence similarity. Our analysis showed that even CSRP1, CSRP2, and CSRP3 have different PPIs (Supplementary Figure S6). In *Arabidopsis* and nematode close homologs, we were unable to mark similar PPIs due to annotation limitation. The full list of PPIs is provided as Supplementary Table S5.

4 Discussion and conclusion

CSRP proteins are important for detailed study since they are implicated in a wide variety of diseases including cardiomyopathies. Most structural characterization has been on the compact LIM domains. In this paper, we investigated the role of the disordered linker region in impacting structural and functional variations. Surprisingly, the length variations of this region show trimodal distribution, while comparing more than 5000 homologous sequences across many organisms. On the other hand, LIM domains are constrained in length variation. This suggests that there is considerable flexibility in the linker region and participates in the overall biological function of CSRP proteins. We see early evidence of CSRP3 homologs in microspecies in a freshwater habitat which do not have a circulatory system, where the precise role of CSRP is not clear. Our analysis points out that plants have a shorter linker region while nematodes seem to have a longer linker region. The mammalia taxa accommodates mixed distribution with (longer linker > medium linker >> smaller linker).

We observe that the disordered region displays conformational flexibility in a directed manner such that the two connected LIM domains could be spatially proximate. There has been limited and early evidence that LIM domains and the Zn-finger subdomains in particular facilitate dimerization of the related protein CSRP1 (Feuerstein et al., 1994). CSRP3 is present in two forms, oligomeric and monomeric (Boateng et al., 2007). These authors have proposed that the oligomeric forms are found in the cytoskeleton and monomeric forms in the nucleus. Further, the N-terminal LIM domain is implicated in dimerization/oligomerization, while the C-terminal LIM domain is involved in actin filaments (AFs) stability (Hoffmann et al., 2014). Hence, as observed by our Molecular Dynamics simulations, it is possible that the interdomain linker region may play an important role in LIM domain interaction. A longer linker tends to have relatively less domain interaction time and compactness compared to shorter and medium linker. Alternately, the open and closed conformations might influence the manner in which CSRP3 oligomerizes in the cell. The study by Papalouka et al. has previously shown that a linker region between two LIM domains of CSRP3 directly interacts with cofilin 2 (CFL2) emphasizing the role of the linker region in protein–protein interaction (Papalouka et al., 2009). Post-translational modifications are an important modulator of oligomerization. An earlier study reported that O-glycosylation influences the self-oligomerization of CSRP3 (Chiaki Nagai-Okatani and Naoto, PLoS One, 2016). Our analysis on deep learning-based prediction shows differential PTMs in CSRP3 homologs. However, O-glycosylation was not significant in the human CSRP3 linker region and more refined approach is required for finding PTMs in human CSRP3. To understand if linker region variation shows different protein–protein interactions, we also analyzed PPI for CSRP3 homologs. Human CSRPs (CSRP1, CSRP2, and CSRP3) showed shared and unique interactions (Figure 6). Other homologs exhibited different interactions that were not interpretable with limited data. Nevertheless, an extensive *in vitro* analysis is needed for a comprehensive conclusion.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.ncbi.nlm.nih.gov/protein/, https://www.rcsb.org/.

Author contributions

RS and PC conceptualized the study. PC integrated the data and performed data analysis and visualization. PC and RS wrote the manuscript, and RS wrote the final draft of the manuscript.

Funding

RS acknowledges funding and support provided by the JC Bose Fellowship (JBR/2021/000006) from the Science and Engineering Research Board, India, and the Bioinformatics Center Grant funded by the Department of Biotechnology, India (BT/PR40187/BTIS/137/9/2021). RS would also like to thank Institute of Bioinformatics and Applied Biotechnology for the funding through her Mazumdar-Shaw Chair in Computational Biology (IBAB/MSCB/182/2022).

Acknowledgments

The authors thank CAPS lab members for their technical help and discussions. They also thank NCBS (TIFR) for infrastructural facilities. PC would like to thank IBAB for the fellowship.

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1134509/full#supplementary-material

References

Bah, A., and Forman-Kay, J. D. (2016). Modulation of intrinsically disordered protein function by post-translational modifications. *J. Biol. Chem.* 291, 6696–6705. doi:10.1074/JBC.R115.695056

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein Data Bank. *Nat. Struct. Mol. Biol.* 10, 980. doi:10.1038/nsb1203-980

Boateng, S. Y., Belin, R. J., Geenen, D. L., Margulies, K. B., Martin, J. L., Hoshijima, M., et al. (2007). Cardiac dysfunction and heart failure are associated with abnormalities in the subcellular distribution and amounts of oligomeric muscle LIM protein. *Am. J. Physiol. Heart Circ. Physiol.* 292, 259–269. doi:10.1152/AJPHEART.00766.2006

Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., and Madden, T. L. (2012). Domain enhanced lookup time accelerated BLAST. *Biol. Direct* 7, 12–14. doi:10.1186/1745-6150-7-12

Bowers, K. J., Chow, E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters, Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC'06, Tampa, FL, USA, November 2006 (IEEE), 84. doi:10.1145/1188455.1188544

Buyandelger, B., Ng, K. E., Miocic, S., Piotrowska, I., Gunkel, S., Ku, C. H., et al. (2011). MLP (muscle LIM protein) as a stress sensor in the heart. *Pflugers Arch.* 462, 135–142. doi:10.1007/s00424-011-0961-2

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinforma.* 10, 421–429. doi:10.1186/1471-2105-10-421

Chauhan, P. K., and Sowdhamini, R. (2022). LIM domain-wide comprehensive virtual mutagenesis provides structural rationale for cardiomyopathy mutations in CSRP3. *Sci. Rep.* 12 (1 12), 1–11. doi:10.1038/s41598-022-07553-1

Eddy, S. R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* 4, e1000069. doi:10.1371/JOURNAL.PCBI.1000069

Ehsan, M., Kelly, M., Hooper, C., Yavari, A., Beglov, J., Bellahcene, M., et al. (2018). Mutant Muscle LIM Protein C58G causes cardiomyopathy through protein depletion. *J. Mol. Cell Cardiol.* 121, 287–296. doi:10.1016/J.YJMCC.2018.07.248

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., et al. (2006). Comparative protein structure modeling using modeller. *Curr. Protoc. Bioinforma.* 15, 5. doi:10.1002/0471250953.BI0506S15

Feuerstein, R., Wang, X., Song, D. C., Cooke, N. E., and Liebhaber, S. A. (1994). The LIM/double zinc-finger motif functions as a protein dimerization domain. *Proc. Natl. Acad. Sci.* 91, 10655–10659. doi:10.1073/PNAS.91.22.10655

Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Bioinforma.* 23, 566–579. doi:10.1002/prot.340230412

Fung, Y. W. W., Wang, R., and Liew, C. C. (1996). Characterization of a human cardiac gene which encodes for a LIM domain protein and is developmentally expressed in myocardial development. *J. Mol. Cell Cardiol.* 28, 1203–1210. doi:10.1006/JMCC.1996.0111

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

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., et al. (2016). OPLS3: A force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* 12, 281–296. doi:10.1021/acs.jctc.5b00864

Hoffmann, C., Moreau, F., Moes, M., Luthold, C., Dieterle, M., Goretti, E., et al. (2014). Human muscle LIM protein dimerizes along the actin cytoskeleton and cross-links actin filaments. *Mol. Cell Biol.* 34, 3053–3065. doi:10.1128/MCB.00651-14

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi:10.1093/MOLBEV/MSW046

Humphrey, W., Dalke, A., and Schulten, K. (1996). Vmd: Visual molecular dynamics. *J. Mol. Graph* 14, 33–38. doi:10.1016/0263-7855(96)00018-5

Humphreys, D. D., Friesner, R. A., and Berne', B. J. (1994). *A multiple-time-step molecular dynamics algorithm for macromolecules.*

Knöll, R., Hoshijima, M., Hoffman, H. M., Person, V., Lorenzen-Schmidt, I., Bang, M. L., et al. (2002). The cardiac mechanical stretch sensor machinery involves a Z disc complex that is defective in a subset of human dilated cardiomyopathy. *Cell* 111, 943–955. doi:10.1016/S0092-8674(02)01226-6

Knoll, R., Kostin, S., Klede, S., Savvatis, K., Klinge, L., Stehle, I., et al. (2010). A common MLP (Muscle LIM protein) variant is associated with cardiomyopathy. *Circ. Res.* 106, 695–704. doi:10.1161/CIRCRESAHA.109.206243

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). Procheck: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283–291. doi:10.1107/S0021889892009944

Martyna, G. J., Klein, M. L., and Tuckerman, M. (1992). Nosé-Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* 97, 2635–2643. doi:10.1063/1.463940

Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* 101, 4177–4189. doi:10.1063/1.467468

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi:10.1093/NAR/GKAA913

Nagai-Okatani ¤a, C., and Minamino, N. (2016). Aberrant glycosylation in the left ventricle and plasma of rats with cardiac hypertrophy and heart failure. *PLoS One* 11, e0150210. doi:10.1371/JOURNAL.PONE.0150210

Papalouka, V., Arvanitis, D. A., Vafiadaki, E., Mavroidis, M., Papadodima, S. A., Spiliopoulou, C. A., et al. (2009). Muscle Lim protein interacts with cofilin 2 and regulates F-actin dynamics in cardiac and skeletal muscle. *Mol. Cell Biol.* 29, 6046–6058. doi:10.1128/MCB.00654-09

Rédei, G. P. (2008). Entrez programming utilities (E-Utilities). *Encycl. Genet. Genomics, Proteomics Inf.*, 612. doi:10.1007/978-1-4020-6754-9_5383

Schmeichel, K. L., and Beckerle, M. C. (1998). LIM domains of cysteine-rich protein 1 (CRP1) are essential for its zyxin-binding function. *Biochem. J.* 331, 885–892. doi:10.1042/BJ3310885

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/NAR/GKAA1074

Wang, D., Liu, D., Yuchi, J., He, F., Jiang, Y., Cai, S., et al. (2020). MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.* 48, W140-W146–W146. doi:10.1093/NAR/GKAA275

Wang, S.-J., Wang, P.-Z., Gale, R. P., Qin, Y.-Z., Liu, Y.-R., Lai, Y.-Y., et al. (2017). Cysteine and glycine-rich protein 2 (CSRP2) transcript levels correlate with leukemia relapse and leukemia-free survival in adults with B-cell acute lymphoblastic leukemia and normal cytogenetics. *Oncotarget* 8, 35984–36000. doi:10.18632/ONCOTARGET.16416

Weiskirchen, R., Pino, J. D., Macalma, T., Bister, K., and Beckerle, M. C. (1995). The cysteine-rich protein family of highly related LIM domain proteins. *J. Biol. Chem.* 270, 28946–28954. doi:10.1074/JBC.270.48.28946

Wiederstein, M., and Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* 35, W407–W410. doi:10.1093/NAR/GKM290

Zolk, O., Caroni, P., and Böhm, M. (2000). Decreased expression of the cardiac LIM domain protein MLP in chronic human heart failure. *Circulation* 101, 2674–2677. doi:10.1161/01.CIR.101.23.2674