Tandem domain structure determination based on a systematic enumeration of conformations

Thérèse Malliavin

scientific reports

OPEN

Tandem domain structure determination based on a systematic enumeration of conformations

Thérèse E. Malliavin

Protein structure determination is undergoing a change of perspective due to the larger importance taken in biology by the disordered regions of biomolecules. In such cases, the convergence criterion is more difficult to set up and the size of the conformational space is a obstacle to exhaustive exploration. A pipeline is proposed here to exhaustively sample protein conformations using backbone angle limits obtained by nuclear magnetic resonance (NMR), and then to determine the populations of conformations. The pipeline is applied to a tandem domain of the protein whirlin. An original approach, derived from a reformulation of the Distance Geometry Problem is used to enumerate the conformations of the linker connecting the two domains. Specifically designed procedure then permit to assemble the domains to the linker conformations and to optimize the tandem domain conformations with respect to two sets of NMR measurements: residual dipolar couplings and paramagnetic resonance enhancements. The relative populations of optimized conformations are finally determined by fitting small angle X-ray scattering (SAXS) data. The most populated conformation of the tandem domain is a semi-closed one, fully closed and more extended conformations being in minority, in agreement with previous observations. The SAXS and NMR data show different influences on the determination of populations.

During the last two decades, interest in the involvement of disordered regions of proteins in biological processes has steadily increased[1]. The influence of disordered regions of proteins has been demonstrated in many phenomena such as: moonlighting in the network of interactions between proteins[2] or in post-translational modifications[3], short linear patterns (SLiMS)[4,5] and molecular recognition functionalities (MoRF)[6]. In addition, proteins containing intrinsically disordered regions tend to act as hubs in protein interactome[7].

The study of disordered regions of proteins is made difficult because most of biophysical techniques concentrates on time-average or space-average data in order to obtain sufficient signal-to-noise ratio. In that frame, the observation of disordered molecules is hampered by the destruction of a part of the observable signal by averaging on heterogeneous sets of conformations. On the other hand, the signal produced by individual conformations is in most of the cases not strong enough to be recorded. Although nuclear magnetic resonance (NMR) is a method of choice for studying conformations of disordered regions of proteins, it faces problems similar to those of other biophysical approaches. Consequently, the approaches traditionally adopted to determine the structural set of proteins for disordered regions rely on two steps: firstly, the generation of a set of conformations, secondly the fitting of these conformations to various biophysical measurements to determine their relative populations. The conformations are classically generated from databases built in advance from the statistical analysis of protein structures[8–10]. Here, we propose a new point of view where all possible conformations are systematically enumerated within limits determined from measurements of chemical shifts.

This enumeration problem is addressed as an interval branch-and-prune (iBP) problem[11–16] in which the protein structure is described as a tree and the available experimental measurements permit tree branching and pruning. As the iBP description makes possible a discrete enumeration of all possible solutions, it strongly contrasts with most of the optimization approaches used for biomolecular structure calculations, as for example the simulated annealing procedure[17]. The approach iBP rejoins the work of Feldman and Hogue[18] who analyzed the exploration of the conformational space of proteins by systematic enumeration. The threading-augmented

[1]Unité de Bioinformatique Structurale, Institut Pasteur, UMR 3528, CNRS, Paris, France. [2]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, USR 3756, CNRS, Paris, France. email: therese.malliavin@pasteur.fr
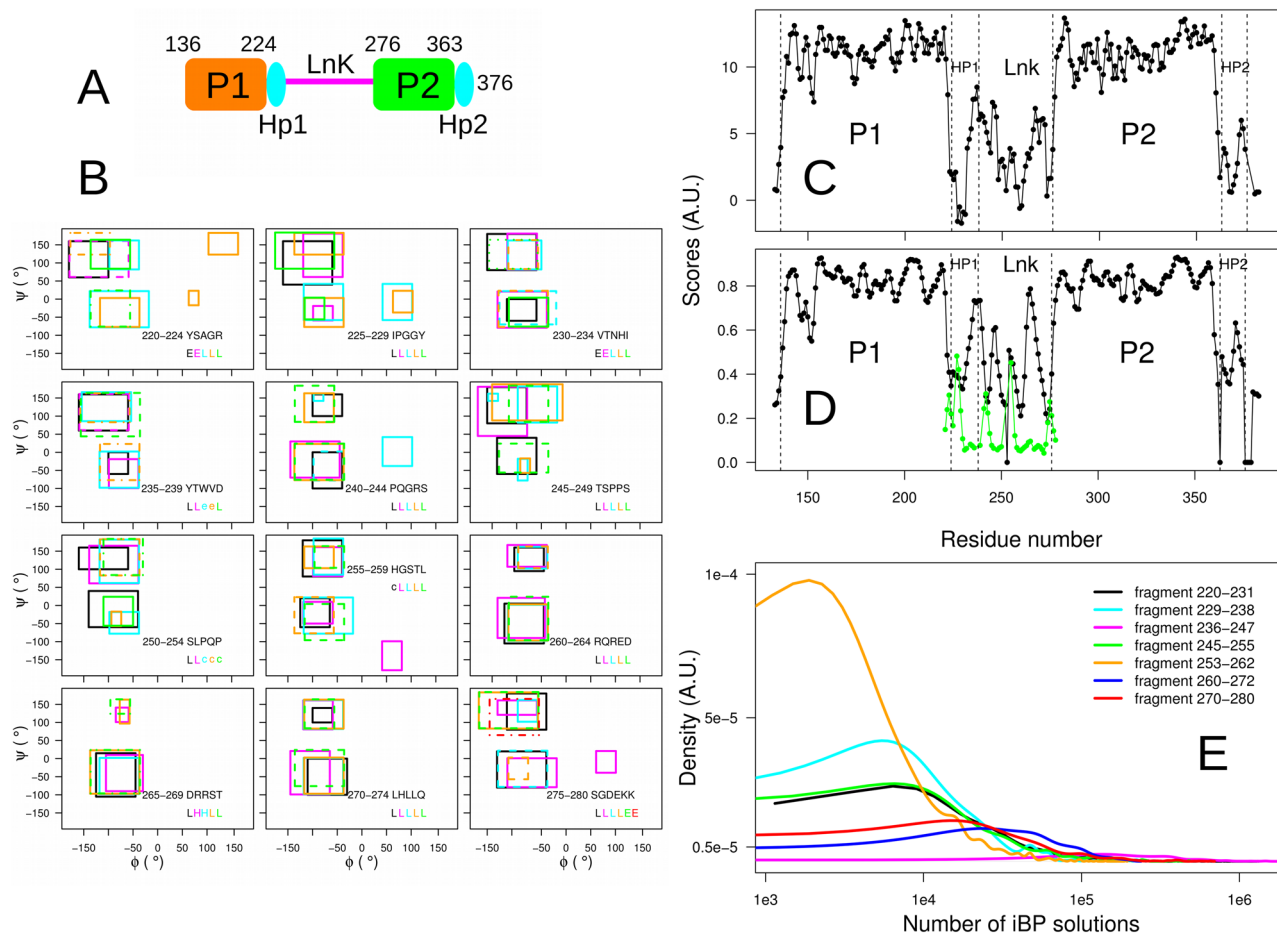
Figure 1. Exploration of the conformational space of Lnk. (**A**) Scheme of the primary structure of the tandem domain protein P1LnkP2 spanning residues 136–376 of protein whirlin. The β hairpins located after the PDZ domains P1 and P2 and corresponding to residues 225–238 and 363–376 are colored in cyan and labeled Hp1 and Hp2. (**B**) Boxes defining the ϕ and ψ intervals sampled during the iBP calculations. On each panel, for five consecutive residues of the linker (six residues 275–280 on the right bottom panel), the boxes are drawn in different colors for each residue and are plotted with different line types in order to reduce box overlays. The corresponding residue sequences are given on each panel along with the secondary structure prediction (H,E,L,e,c) proposed by TALOS-N[28], colored in the same way than the boxes. C-D. Disorder prediction along the residues of P1LnkP2 by Chemical shift Z-score for assessing Order/Disorder (CHEZOD)[29] (**C**), by random coil index (RCI)[30] calculated with TALOS-N[28] (**D**, black curve) and by 1-coil (**D**, green curve) where coil is the percentage of coil predicted by δ2D[31]. As the δ2D approach is specially designed for disordered protein regions, it was only applied on the Lnk region. (**E**) Distribution of the number of solutions for individual iBP calculations realized on a given peptide fragment. For the sake of clarity, the two axes are drawn in logarithmic scale. The panel (**A**) was prepared using LibreOffice 6.0.7.3[57] and the other panels using R 3.4.1[58].

interval Branch-and-Prune (TAiBP) approach was recently proposed[19,20] as a reformulation of iBP approach[11]. TAiBP provides a framework for the systematic enumeration of protein conformations, while allowing overcome the combinatorial barrier intrinsic to this enumeration, and will thus be used here for enumerating conformations in disordered protein regions.

Numerous proteins involved in cell signaling are formed of several folded domain connected by more or less disordered linkers. Such proteins, the so-called tandem domain proteins, play important role in various physiological processes[21–25]. The folded domains have usually well-known structural properties whereas the conformations of disordered linkers are less determined, but play a key role to understand the conformational landscapes of tandem domain proteins and thus their functions.

The approach for systematically exploring the conformational landscape of proteins focuses here is on the calculation of a tandem domain of the protein whirlin, a scaffold protein essential for the assembly of the USH2 protein complex[26]. A failure in the formation of this complex is observed in the Usher syndrome, a genetic disorder resulting in a combination of hearing loss and visual impairment. The tandem domain of whirlin spans the residues 136–376 and is formed of a PDZ domain, named hereafter P1 (residues 136–224), of a linker named hereafter Lnk (residues 225–280), and of a second PDZ domain (residues 281–362) followed by a β hairpin (residues 363–376), and named hereafter P2 (Fig. 1A). The structure of P1LnkP2 has been previously extensively

studied by various NMR approaches as well as SAXS measurements[27], but this enormous effort did not produce an exhaustive description of the protein structure including Lnk conformations.

Results

A several steps approach was here adopted for the calculations of P1LnkP2 conformations, including: (1) the use of nuclear magnetic resonance (NMR) chemical shifts measured on Lnk coupled to the TAiBP approach to systematically enumerate Lnk conformations; (2) the assembly of P1 and P2 folded structures to Lnk conformations; (3) the optimization of P1LnkP2 conformations with respect to measured NMR residual dipolar couplings (RDC) and paramagnetic resonance enhancements (PRE)[27]; (4) the determination of the populations of P1LnkP2 conformations using small angle X-ray scattering (SAXS) data.

Extraction of ϕ, ψ boxes used for enumerating the Lnk conformations. Using the chemical shifts measured for the nuclei Hα, HN, ^{15}N, ^{13}Cα, ^{13}Cβ of Lnk residues, the output of TALOS-N[28] was used to determine boxes of ϕ and ψ values, giving the limits in which the conformations of Lnk will be enumerated. TALOS-N uses an artificial neural network (ANN) to predict the 324-state ϕ/ψ distribution of residue i on the basis of the NMR chemical shifts and of the residues adjacent to i. The 324-state ϕ/ψ distribution corresponds to the likelihood that residue i adopts torsion angles that fall in any of the 324 voxels, of $20° \times 20°$ each, that make up the Ramachandran map[28]. Classically, as TALOS-N initially intends to predict a single region of ϕ, ψ values in order to build dihedral angle boxes used for optimization of protein structure, either the regions of the Ramachandran map with smallest likelihood have been classically assumed to correspond to statistically unlikely regions, either the ϕ, ψ angles were pointed as unpredictable in the case where several regions of the Ramachandran maps displayed comparable likelihood values. The present work proposes to take a new point of view in which all regions of Ramachandran map displaying likelihood larger than a given threshold may correspond to Ramachandran regions populated by some conformations present in solution for the studied protein. Using this point of view, a protocol is described to exhaustively enumerate all possible conformations of Lnk within the limits of the selected Ramachandran regions, as described in the Section "Enumeration of Lnk conformations" in Methods.

In order to obtain the ϕ and ψ boxes which will be used for sampling protein conformations, each ϕ/ψ distribution produced by TALOS-N for a given residue was normalized in order that the sum of all voxels values is equal to one. Then, all voxels for which the normalized value was larger than 0.01, were selected. Boxes were manually chosen overlying selected voxels. As for the residue P254, no ϕ/ψ distribution could be determined by TALOS-N, the same boxes than the ones chosen for H255 were used for this residue. The ϕ and ψ lower and upper limits defining the boxes are given in Table S1.

The boxes are drawn in the Ramachandran maps (Fig. 1B) for stretches of successive five or six residues in the linker. For most of the residues, two boxes have been extracted from the ϕ/ψ distributions, one being located in the negative ϕ and ψ values close to the α helix, and the other being located in the negative ϕ and positive ψ values close to the β region. For few residues, a box was also detected in the region of positive ϕ values. The predictions of secondary structures produced by TALOS-N, and also given in the panels of Fig. 1B, are in agreement with the boxes. Indeed, the residues predicted to be in "E" structure display a box containing the β strand region and the residues predicted to be in "H" structure display a box containing the α helix region. The residues predicted to be in the secondary structure "L" or "c" displays mostly at least two boxes located in the neighborhood of α helix or β strand and for some residues, additional boxes located in the positive ϕ region. This observation agrees with a description of "L" or "c" secondary structures corresponding to an equilibrium between conformations with various secondary structures.

From the chemical shift values, relative order has been predicted along the P1LnkP2 residues using several approaches: the Chemical shift Z-score for assessing Order/Disorder (CHEZOD)[29], the random coil index (RCI) score[30] implemented in TALOS-N and the method δ2D[31] (Fig. 1C,D). CHEZOD (Fig. 1C) and RCI (Fig. 1D, black curve) display similar profiles. For both methods, Lnk is more disordered than P1 and P2, but some Lnk regions display order peaks. At the contrary, the plot of 1-coil, where coil represents the percentage of coil value determined by δ2D[31] (Fig. 1D, green curve) displays smaller values and a flatter profile in the region Lnk, pointing to a greater disorder than the other methods, but nevertheless showing some peaks of order.

The analysis of likelihood maps produced by TALOS-N allowed us to determine ϕ, ψ boxes which agree with the secondary structure predictions made by TALOS-N. Relative order predictions indicate regions with residual order in Lnk.

TAiBP provides a complete enumeration of Lnk conformations. The TAiBP approach described in Methods (Section "Enumeration of Lnk conformations"), has been used for generating Lnk conformations. The cumulative numbers (N_{iBPrun}) of individual iBP runs for each fragment vary from 422 up to 1534, most of them being larger than 1000 (Table 1A). The distribution of the number of solutions for each individual iBP calculation was plotted (Fig. 1E) according to the analyzed fragment. An important feature of these distribution is that the number of solutions for individual iBP runs are all in the range 10^3–10^6, corresponding to values smaller than the maximum of requested solutions (10^9) meaning that all trees were completely explored. Most of the fragments display numbers of solutions in the range 10^3–10^5. But the fragment 253–262 (orange curve) displays somehow a narrower range (10^3–10^4) whereas the fragment 236–247 (magenta curve) displays a shifted range (10^4–10^6). The narrower range of fragment 253–262 may arise from smaller boxes used for residues 253, 255, 256, 259 and 260 (Table S1 and Fig. 1B).

The cumulative numbers of iBP solutions ($\log_{10}(N_{iBPconf})$) are located (Table 1A) in the $10^{7.2}$–$10^{7.8}$ range, except the fragment Pept3 (residues 236-247) which displays a cumulative number of solutions ten times larger than other fragments, in agreement with the largest range of numbers of solutions observed for the corresponding

A. iBP step	Lnk residue range	N_{iBPrun}	$\log_{10}(N_{iBPconf})$	$N_{clustiBP}$
Pept1	220–231	1534	7.8	477
Pept2	229–238	424	7.0	331
Pept3	236–247	1171	8.6	350
Pept4	245–255	768	7.4	182
Pept5	253–262	1534	7.2	303
Pept6	260–272	1024	7.8	361
Pept7	270–280	768	7.7	175
B. Assembly step	N_{trials}	$N_{clashes}$	N_{saved}	N_{clust}
Pept1-2	157,887	15,103	142,784	184
Pept1-2-3	64,400	7406	56,994	190
Pept1-2-3-4	34,580	3432	31,148	165
Pept1-2-3-4-5	49,995	4328	45,667	205
Pept1-2-3-4-5-6	74,005	12,615	61,390	234
Pept1-2-3-4-5-6-7	40,950	5953	34,997	215

Table 1. TAiBP calculation. A. Peptide fragments sampling with iBP. Lnk was divided in 7 fragments spanning the residues 220–231 (Pept1), 229–238 (Pept2), 236–247 (Pept3), 245–255 (Pept4), 253–262 (Pept5), 260–272 (Pept6) and 270-2-80 (Pept7). N_{iBPrun}: cumulative numbers of iBP runs for all sampled ϕ, ψ boxes on a given fragment. $N_{iBPconf}$: cumulative numbers of generated conformations in the iBP runs for all parsed restraints on a given fragment. N_{clust}: number of conformations after the clustering using self-organizing maps[32]. B. Assembling the 7 fragments Pept1 to Pept7 spanning Lnk. Each assembly step is labeled "Pept" followed by the index of assembled fragments and describes the addition of the last index fragment to the protein obtained at the previous assembly step. N_{trials}: number of trials during the assembly run. $N_{clashes}$: number of solutions removed because of Cα atoms closer than 1 Å. N_{saved}: number of solutions before clustering. N_{clust}: number of solutions after clustering using self-organizing maps[32].

individual iBP runs (magenta curve in Fig. 1E). A ratio of 2.7 is observed between the largest and smallest numbers of representative conformations ($N_{clustiBP}$) obtained after clustering by the self-organizing maps[32], whereas this ratio was of about 4 for the initial numbers of iBP solutions $N_{iBPconf}$. This proves that the clustering has a key influence on the combinatorial of the iBP calculation.

The various steps of fragment assembly were analyzed (Table 1B) by monitoring several parameters: N_{trials}, $N_{clashes}$, N_{saved} and N_{clust}. N_{trials} is the number of assembly attempts equal to the product of the number of conformations available for each assembled fragment and displays quite different values depending on these numbers. Interestingly, the number of cases where atoms Cα are closer than 1 Å ($N_{clashes}$) is for most cases about 10% of the number of assembly trials (N_{trials}), except for the assembly of the fragment Pept6 (residues 260–272), where the number of clashes is larger, about 15%. Interestingly, this fragment is the only one including two α helix predictions for R266 and R267. Discarding conformations for which Cα are closer than 1 Å, which represents a very simple geometric condition is thus quite efficient to reduce the number of solutions. A ratio of 1.4 is observed between the largest and smallest numbers of representative conformations (N_{clust}) obtained from self-organizing maps, while this ratio was 4.6 for the numbers of trials N_{trials}. The clustering has even a greater influence on the combinatorial during the assembly step than during the iBP step.

After having finished the assembly, the last clustering step provides 215 conformations, which will be starting point to build the tandem domain protein and optimizing it along other independent measurements: residual dipolar couplings (RDC), paramagnetic resonance enhancements (PRE), small angle X-ray scattering (SAXS).

Optimizing P1LnkP2 with respect to independent NMR parameters. Starting from the 215 Lnk conformations generated by TAiBP, the procedure to assemble P1 and P2 domains, described in Section "Assembling P1 and P2 to Lnk" of Methods, produced 83 conformations of P1LnkP2.

The fit to residual dipolar couplings (RDC) of the relative orientation of P1 and P2 domains within P1LnkP2 was evaluated by calculating theoretical RDCs using PALES 2.1[33], as well as their correlation factors R with experimental RDCs. Correlation factors in the 0.59-0.82 range were obtained, and XPLOR-NIH[34] was used to optimize with respect to RDCs the 33 conformations for which a correlation factor smaller than 0.7 was observed. The optimization along RDCs is described in the Section "Optimization of P1LnkP2 conformations along the RDCs" of Supplementary Materials. After optimization, the average correlation of the 33 conformations went up to 0.95 and all correlations were in the range 0.7- 0.98.

The conformations refined along RDCs were replaced among the 83 total conformations, and the 10 conformations (13, 20, 55, 145, 146, 150, 153, 160, 174, 186) displaying a distance between geometric centers of P1 and P2 smaller than 32 Å, were submitted to a second optimization round using XPLOR-NIH with respect to PREs as described in Supplementary Materials (Section "Refinement of P1LnkP2 conformations along the PREs"). This second refinement intended to improve the fit of PRE profiles by moving closer the domains P1 and P2.

A manual inspection of the fits of PRE on each of the conformations refined along PRE revealed that no single conformation was able to provide a satisfying fit of all experimental PRE profiles (data not shown). Thus, the
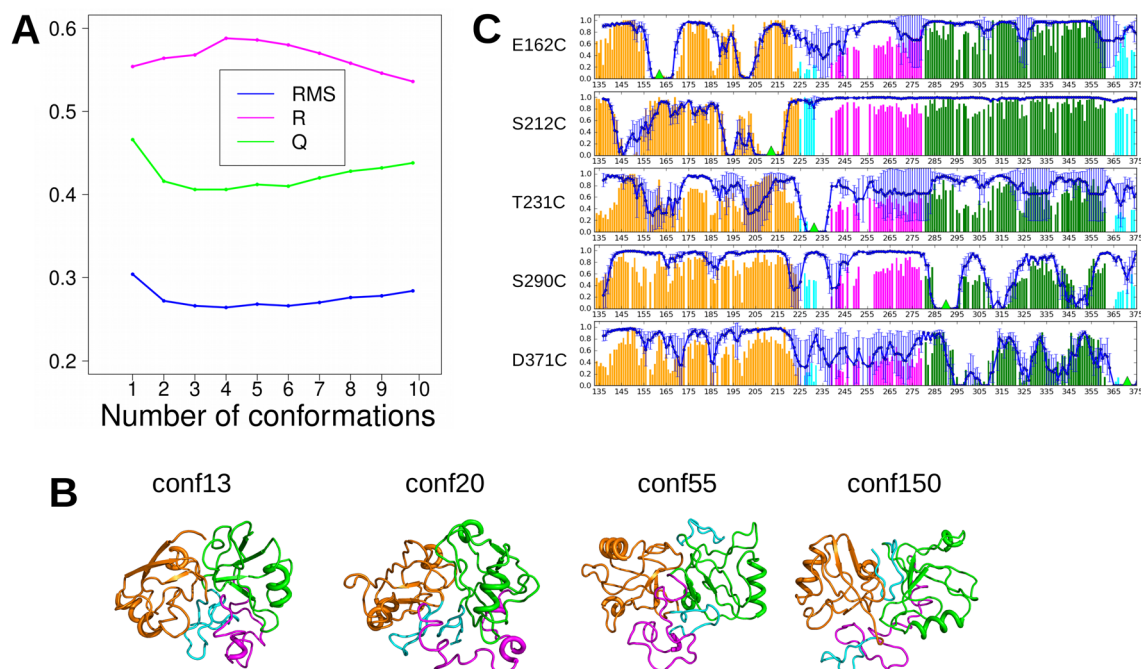
Figure 2. Selection of the P1LnkP2 conformations refined with respect to PRE data. A. Plot of the RMS, R and Q factors in function of the added conformation in the pool of selected closed conformations. The plot was prepared using R 3.4.1[58]. B. Conformations 13, 20, 55 and 150 detected in the group of best four conformations. They are drawn in cartoon with P1 colored in orange, P2 in green, Lnk in magenta and the two hairpins in cyan. The conformations were drawn using pymol 1.8.4.0[59]. C. Comparison of theoretical and experimental PRE profiles for conformations 13, 20 and 55. The theoretical profiles are plotted in blue with error bars and calculated assuming that the three conformations have the same populations. The experimental profiles are drawn as bars, colored with the same color-code than for the protein conformations. Each individual profile plot corresponds to one position of the MTSL probe from top to bottom: E162C, S212C, T231C, S290C, and D371C. The green triangles give the positions of the probe in the sequence. The plot was prepared using python script based on numpy 1.7.1[49] and Matplotlib 1.5.3[60] packages.

fit to PREs of all possible combinations of sub-groups of conformations has been analyzed. The set of P1LnkP2 conformations best suited to the PRE profile was built iteratively by choosing at each step to add the conformation such that the best correlation factor R was observed for the new set of conformations (Fig. 2A). In that way, a group of conformations was progressively built starting from conformation 13, then adding conformations 55, 20 and 150. RMS, R and Q values of 0.3, 0.6 and 0.4 were observed for this group. The plot of correlation factor R, of RMS and Q factors (Fig. 2A) along the size of conformation set shows that the minima of R, RMS and Q were observed for a group size of 4. Thus, the conformations 13, 20, 55 and 150 (Fig. 2B) were selected as closed conformations for further analysis.

Determining the populations of conformations by fitting the SAXS curve.
Theoretical SAXS curves were calculated on the P1LnkP2 conformation with the tool CRYSOL[35] available in the package ATSAS 3.0.3[36] using the procedure described in Section "Determination of the P1LnkP2 populations" of Supplementary Materials. The software BioEn 0.1.1[37] and Mesmer 1.0.0[38] were then used to determine the sets of populated conformations and their corresponding populations fitting the experimental SAXS curve. as described in the Section "Determination of the P1LnkP2 populations" in the Supplementary Materials. The set of processed conformations includes: (1) the $83 - 10 = 73$ P1LnkP2 conformations for which the distance between the geometric centers of P1 and P2 is larger than 32 Å; (2) various sets of conformations from the subset of closed conformations 13, 20, 55 and 150 previously selected after the PRE refinement step. Two series of BioEn calculations were performed using the whole SAXS curve (Table 2) or using the SAXS curve measured for scattering vectors q up to 3.5 nm^{-1} (Table S7).

The first conclusion which can be drawn from the BioEn calculations (Tables 2 and S7) is that whatever is the set of included closed conformations, only one of them displays a non-null population except for the calculation conf13,20,55,150 in Table S7. This population is most of the time around 10%, and is equal to 17% (Table 2) or 13% (Table S7) for conformation 55 in the case when the four closed conformations 13, 20, 55 and 150 are included in the calculation. The SAXS fitting is thus no sensitive to the presence of a specific closed conformation and a population of about 10% globally observed for all closed conformations has to be interpreted as the global population of all closed conformations.

The calculation performed using the four closed conformations (last right column of the Tables 2 and S7) displays a larger final χ^2 value than other calculations and shows a large variation in the populations. Indeed, the conformation 176 which displays a population around 60% for all other calculations, is discarded from the

	conf13	conf20	conf55	conf13,20,55	conf13,20,55,150
conf12	3.4 ± 1.2	–	$4.2 \pm 6.1e{-}3$	$4.2 \pm 3.4e{-}3$	$72.0 \pm 4.1e{-}4$
conf13	**9.9 ± 0.11**	**not incl**	**not incl**	**–**	**–**
conf20	**not incl**	**9.0 ± 0.16**	**not incl**	**–**	**–**
conf24	15 ± 0.38	18 ± 1.2	$16 \pm 2.9e{-}3$	$16 \pm 1.1e{-}03$	–
conf55	**not incl**	**not incl**	**$10 \pm 6.9e{-}4$**	**$10 \pm 4.0e{-}4$**	**$17 \pm 1.2e{-}4$**
conf140	8.6 ± 0.044	8.1 ± 0.22	$8.8 \pm 2.2e{-}4$	$8.8 \pm 3.8e{-}4$	$11.0 \pm 1.2e{-}4$
conf150	**not incl**	**not incl**	**not incl**	**not incl**	**–**
conf176	63 ± 0.94	65 ± 1.3	$60 \pm 4.6e{-}3$	$60 \pm 2.8e{-}3$	–
Final $\chi 2$	$1.7 \pm 2.5e{-}7$	$2.7 \pm 7.1e{-}7$	$2.0 \pm 1.0e{-}6$	$3.2 \pm 1.6e{-}6$	$5.4 \pm 1.5e{-}6$
Final S	$-2.4e{-}6 \pm 2.1e{-}10$	$-1.3e{-}5 \pm 6.0e{-}10$	$-2.8e{-}5 \pm 8.9e{-}10$	$-7.7e{-}05 \pm 1.3e{-}9$	$-4.7e{-}5 \pm 1.3e{-}9$

Table 2. Populations of conformations determined using BioEn 0.1.1[37] on the 73 P1LnkP2 conformations for which the distance between the geometric centers of P1 and P2 was larger than 32 Å and various closed conformations among 13, 20, 55 and 150, previously selected according to the fit of PRE data (Fig. 2). After ten runs starting from random values of populations and performed on the whole set of conformations, all conformations for which the sum of populations over the ten runs was larger than 0.01 were gathered, and a second run of ten additional BioEn calculations was performed on this reduced set of conformations. The average and standard deviation values of populations obtained for each selected conformation from the second set of BioEn runs, are given in the Table, along with the final average values of χ^2 and of entropy S. In each calculation, the conformations tagged as "not incl" have not been initially included in the calculation, whereas the conformations tagged as "–" were included in the calculation, but not selected by BioEn[37]. The conformations belonging to the set of closed conformations are written in bold.

final set of conformations. Similarly, the conformation 12, low populated in all other calculations, jumps up to 72%. Due to the inconsistency with other calculations observed when the conformation 150 is included, this conformation was discarded from the set of closed conformations.

In the first four columns of Tables 2 and S7, very similar population profiles are obtained, the biggest variations being observed between conformations 12 and 24. As the smallest average final value of χ^2 is observed for the calculations including only the conformation 13, the populations obtained in Table 2 for that case were kept as the result of the fitting of SAXS curve. The reduction of SAXS curve to the scattering vectors q up to $3.5 \, \mathrm{nm}^{-1}$ produce population results (Table S7) similar to those obtained with the full SAXS curve (Table 2). In addition, the superpositions of experimental and theoretical SAXS curves are quite similar in all Bioen calculations (Figure S2).

The software Mesmer[38] was used in parallel to produce an independent estimation of populations. Two sets of Mesmer calculation were performed using different numbers of randomly chosen initial conformations (the so-called components) equal to 40 and 60. The same sets of conformations were selected by the two calculations, with very similar populations (Table S2). Interestingly, these conformations are also the ones selected by BioEn, except that the closed selected conformation is 150. The populations of conformations determined by Mesmer are also similar to the ones obtained by BioEn. One can notice a slight decrease of the population of 176 from about 60% to about 50%, and a change of population balance between conformations 12 and 24.

The fitting of SAXS curve using the software BioEn 0.1.1[37] Mesmer 1.0.0[38] and ATSAS 3.0.3[36] select the same set of conformations and produce similar results for the populations of conformation. Nevertheless, the SAXS curve fitting is more reliable for small than for large scattering vectors. Unsurprisingly, it was not possible to determine the relative population of each closed conformations, as all closed conformations display similar global shapes.

The conformational landscape of P1LnkP2. The previous analyses allow us to propose the following set of conformations for describing the conformational landscape of P1LnkP2 in solution (Fig. 3). The most populated conformation 176, with semi-closed configuration of P1 and P2 displays a population of 60% or 50% depending on the algorithm used for population determination. Closed conformations display globally a population of 10%, but the SAXS fitting does not allow to determine the relative proportion of each closed conformation within this global value. Other conformations, 24 and 12, displays an open P1-P2 configuration with respective populations of 15 and 3% or 7 and 20% depending on the use of BioEn or Mesmer. The very extended conformation 140 displays a population around 10% whatever procedure used to determine the populations.

Using the individual gyration radii of the individual conformations along with their relative populations (Fig. 3), an average value of gyration value can be estimated to be 26 Å, in total agreement with the gyration radius of 27.5 Å previously[27] measured. The small discrepancy of 1.5Å between these two values could be assigned to solvation layer present in solution around the protein and not taken into account in the conformations determined in the present work.

The fit of PRE data by the three selected closed conformations 13, 20 and 55 can be evaluated by the comparison of experimental and theoretical PRE profiles (Fig. 2C) realized using the same relative populations for the three conformations. The theoretical profiles displays variations in reasonable agreement with the experimental
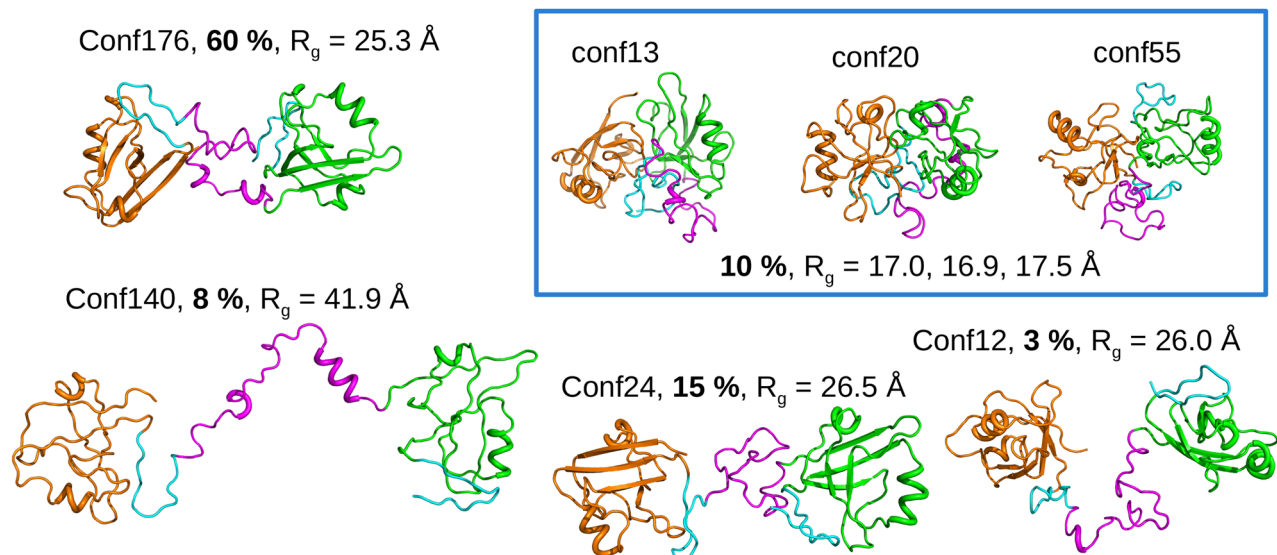
Conf176, **60 %**, R$_g$ = 25.3 Å

conf13   conf20   conf55

10 %, R$_g$ = 17.0, 16.9, 17.5 Å

Conf140, **8 %**, R$_g$ = 41.9 Å

Conf12, **3 %**, R$_g$ = 26.0 Å

Conf24, **15 %**, R$_g$ = 26.5 Å

Figure 3. Overview of the P1LnkP2 conformations selected by BioEn 0.1.1[37]. The conformation indexes are given along with the populations obtained by SAXS curve fitting in bold, and with the gyration radius **REV2NUMB6** (R$_g$). The most closed P1LnkP2 conformations, for which it was not possible to determine the individual populations from the fitting of SAXS data, are plotted within a frame. The conformations were drawn using pymol 1.8.4.0[59].
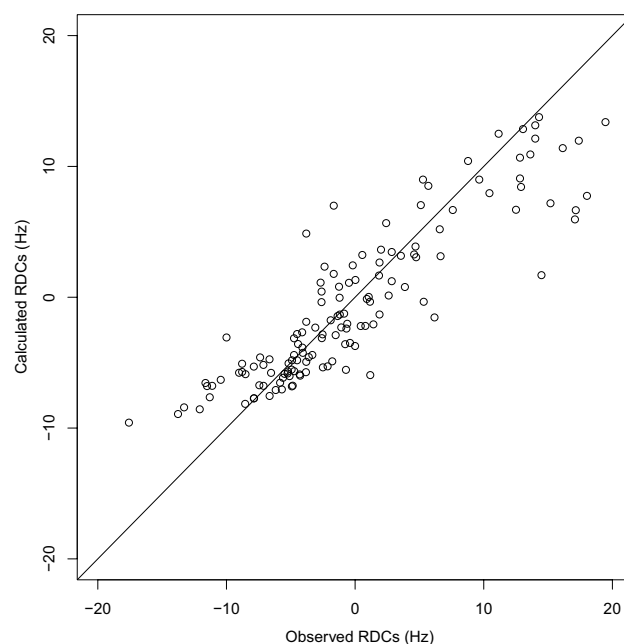

Figure 4. Comparison of the calculated and experimental residual dipolar coupling (RDC) values. The RDC values have been calculated with the software PALES 2.1[33] on the conformations 12, 13, 24, 140, 176 and were averaged using the populations determined in Fig. 3. The plot was prepared using R 3.4.1[58].

profiles. RMS, R and Q factors have been calculated for each MSTL spin label and each closed conformation (Table S3). For most spin labels and protein conformations, reasonable R correlations were obtained.

The RDC values have been recalculated on the conformations 12, 13, 24, 140, 176 using the software PALES and averaged according to populations displayed in Fig. 3. The comparison of calculated and observed RDC values (Fig. 4) reveals a reasonable agreement between the two sets of values, with a RMS of 4.1 Hz, a R factor of 0.8 and a Q factor of 0.5. The analysis of the agreement with RDCs for individual conformations (Table S4) reveals R correlation factors larger than 0.9 for all closed conformations and conformation 140 and larger than 0.8 for conformation 176.

Consistency between the observed and theoretical likelihood maps. The consistency of the obtained set of conformations and populations was probed with respect to the likelihood (ϕ, ψ) maps M$_{obs}$ predicted by TALOS-N from experimental chemical shifts and initially used to get the ϕ and ψ boxes (Fig. 1B). Theoretical chemical shifts have been calculated from the conformations 12, 13, 24, 140 and 176 for nuclei Hα, HN, ^{15}N, ^{13}Cα, ^{13}Cβ (when present in the residue) and ^{13}C' using the software Shiftx 1.0[39]. The chemical shifts were averaged according to the population of individual conformations (Fig. 3), assuming fast equilibrium between the conformations. The obtained averaged chemical shifts were then analyzed using TALOS-N and the obtained likelihood maps M$_{theo}$ compared to the initial maps M$_{obs}$.

The contour plots of maps M$_{theo}$ and M$_{obs}$ are superimposed for each Lnk residue using black color for M$_{obs}$ and red color for M$_{theo}$ (Figure S3). Inspection of the plots reveals a good agreement between M$_{obs}$ and M$_{theo}$ maps. Indeed, over the 61 residues considered, only 17 display discrepancies for the peak positions in the two maps. Two glycines, G228 and G256, display a displacement of the peak located in the region of positive ϕ values. As glycine residues display a tendency to populate regions of positive ϕ values, this may bias the TALOS-N database and consequently the predicted likelihood map. In addition, several residues display peak variations in the α and β regions of the maps. The residues Y235, T236, Q253 and E278 display on maps M$_{obs}$ (black) two peaks in the regions of α helix and β strand, and the α peak disappears in maps M$_{theo}$ (red). At the contrary, the residues S249, Q253, H255, as well as the residues 260-264 and 269-272, display also two peaks on maps M$_{obs}$ (black) in the regions of α helix and β strand, and the β peak disappears in maps M$_{theo}$ (red). One should notice that only one or two chemical shifts are assigned in H233, V238 and H255, and that the of residues W237, P247, and 252-254 are unassigned: the lack of measured chemical shifts makes the map M$_{obs}$ to rely mainly on sequence alignment with the database of TALOS-N and could hamper the precision of M$_{obs}$ prediction. We should also notice that for the residues displaying peak variations in regions α and β of maps M$_{theo}$ and M$_{obs}$, the 83 P1LnkP2 conformations displays backbone dihedral angles in the two regions. But, after determining populations by fitting of SAXS curve, the distribution of ψ angle values is shifted toward one of the α or β regions. Most of the discrepancies between M$_{theo}$ and M$_{obs}$ peaks in α and β Ramachandran regions is thus a consequence of the fitting of SAXS data.

Discussion

In the present work, a protocol for systematically enumerating the conformations of a tandem domain protein has been presented. This protocol makes use of the TAiBP approach[19,20] which is specifically designed for partially disordered protein regions. In TAiBP, a reformulation of the distance geometry approach permits an exhaustive sampling of protein conformations within a given set of distance and angle limits. The availability of all possible conformations of Lnk consequently provides the largest possible exploration of the relative positions of P1 and P2 in the space. The subsequent refinements of the P1LnkP2 conformations with respect to RDC and PRE measurements along with the population determination by fitting of SAXS curves produces a exhaustive set of conformations verifying all experimental measurements. The reliability of this set of conformations arises directly from the exhaustive exploration allowed by TAiBP. In this frame, one should point out that the convergence of calculation is not expected to validate the set of obtained conformations, as the approach proposed here for conformational space exploration is not based on a local optimization methods. The procedure proposed here requires at least the knowledge of heteronuclear NMR assignment and chemical shifts as well as the measurement of SAXS data.

The backbone angle limits used as inputs for Lnk conformational analysis are determined using a new point of view on the processing of likelihood maps outputs of the neural network TALOS-N[28]. These maps are transformed into probability maps and boxes of ϕ and ψ values are extracted from all probability regions larger than a given threshold. These boxes will be then used to sample Ramachandran multiple regions. As the bounds for TAiBP calculations are exclusively determined by geometric parameters, the Lnk conformations do not take into account the free energy aspects. The fitting of SAXS curve then permits to obtain the relative populations of the P1LnkP2 conformations, re-introducing the free energy information.

At the end of calculation, a limited number of discrepancies between observed and theoretical likelihood (ϕ, ψ) maps, concerns only 17 residues over 61. The discrepancies could arise from several reasons. An obvious one is that TALOS-N was designed on systems displaying a unique basin of conformations. Another one was pointed out above: the fitting of SAXS curve induce the selection of conformations from specific Ramachandran regions and their corresponding maps M$_{theo}$ are different from the M$_{obs}$ maps. Nevertheless, the small quantity of residues displaying discrepancies is of good hope for the quantitative analysis of TALOS-N likelihood maps. Beside, it should be noticed that cross-correlated spin relaxation was recently proposed[40,41] for directly determining probability distributions of ϕ and ψ angles: these distributions could be also used for determining ϕ and ψ boxes for enumeration of conformations. In addition, backbone angle information determined from J couplings could be used as inputs for the TAiBP calculation.

The refinement of P1LnkP2 conformations with respect to the NMR and SAXS measurements put in evidence the different influence of global (SAXS) and more local (RDC, PRE) parameters. In particular, the fitting from SAXS data is much less sensitive to closed conformations than the fitting to PRE data. Discrepancy of SAXS measurements with other biophysical techniques is a well-known aspect of studies on partially disordered proteins[42,43].

The obtained conformations and population gives a simplified representation of the conformational landscape of P1LnkP2. This landscape is in qualitative agreement with the conformational distribution determined in the previous structural study of the whirlin tandem domain[27], the mostly populated conformation being the one with P1 and P2 domain close to each other. The most closed conformations, in agreement with the PRE measurements are less populated than semi-closed ones. This may be related to a conformational entropy barrier. Indeed, as the Lnk region is more mobile than the P1 and P2 domains, Lnk contributes importantly to the conformational

8

entropy. The Lnk packing requires a decreasing of conformational entropy inducing consequently an entropy barrier, which could explain the lowest populations of closed conformations.

Methods

The P1LnkP2 conformations have been determined in two steps: first using the TAiBP approach for systematically enumerating Lnk conformations, then applying a specifically designed procedure to assemble P1 and P2 to the Lnk conformations. The scripts and programs required to run TAiBP are available at: https://github.com/tmalliavin/.

Enumeration of Lnk conformations.   The Lnk conformations have been enumerated using the threading-augmented interval branch-and-prune (TAiBP) approach recently proposed[20] to take advantage of the systematic enumeration permitted by the iBP approach[11–15] while overcoming the combinatorial barrier arising from the enormous space of protein conformations[44]. The approach efficiency was proved for proteins up to 60 residues[20] and TAiBP is thus relevant for the Lnk calculation, which contains 61 residues. TAiBP is composed two steps: (1) individual iBP calculations of peptide fragments spanning the studied protein; (2) enumeration of Lnk conformations by systematic assembly of fragment conformations.

The enumeration of Lnk conformations was performed using boxes of backbone angles ϕ and ψ. These boxes (Fig. 1B and Table S1) have been extracted from the likelihood Ramachandran maps obtained by TALOS-N[28] as described in Results. The enumeration of conformations in the iBP approach is based on the building of a tree, each node of the tree corresponding to a position of an atom. The tree building allows one to enumerate the various possibilities for atom positions (branching step) whereas additional geometric information is used to accept or reject each newly built branch (pruning step). The atoms are positioned from the position of previous atoms according to a specific ordering (Table S5) of the protein atoms[13] and using the iBP algorithm described by Worley et al[15,19]. The atom ordering ensures that there is a restricted and manageable locus for the spatial position of every atom. This is achieved using a "relaxed form" of trilateration with respect to the three preceding atoms in the order. More precisely, two out of three of the distances involved in trilateration must be known exactly, and one may be subject to uncertainty and represented by an interval[13,15]. As described in Section "Calculation of the backbone angles from interatomic distances" in the Supplementary Material, the backbone dihedral angles ϕ and ψ can be straightforwardly related to bond lengths and bond angles and respectively to distances between atoms C of residues $i-1$ and i and between atoms N of residues i and $i+1$. This equivalence between the backbone dihedral angles and interatomic distances permits to use the angles ϕ and ψ for the so-called branching step. Once atoms positions have been generated in the tree, each one is examined to check whether they verify pruning geometric information and is pruned in the case the information is not verified. In the present calculation, the pruning was applied by preventing atoms to be closer than the sum of their van der Waals radii and by checking that the improper angle values are correct. The bond lengths, bond angles, improper angles and van der Waals radii were taken from the force field protein-allhdg5-4 PARALLHDG (version 5.3)[45] (Table S6). The van der Waals radii were scaled by a factor of 0.7.

The ϕ/ψ boxes determined for each Lnk residue from TALOS output were systematically combined by permutation to prepare individual iBP calculations, realized in a way similar to those described in Ref.[20]. Lnk was divided in 7 fragments corresponding to residues 220-231, 229-238, 236-247, 245-255, 253-262, 260-272 and 270-280. For each fragment, two dummy residues were added at the N and C terminal extremities, and the ϕ and ψ dihedral angles of the inner peptide residues were sampled according to the box limits (Table S1). In order to avoid pruning due to slight discrepancy between distances, a tolerance of 0.05 Å has been added to the bounds of distance intervals. The maximum number of branches by interval was set to 4. The minimum discretization factor, which is the minimum ratio between each distance interval to the number of tree branches generated within the interval, was set to 0.1 Å, in order that the branching does not over-sample small intervals. The iBP algorithm was used to exhaustively enumerate the conformations verifying these limits by systematically exploring the tree defined by the boxes on ϕ and ψ angles. The number of saved conformations was reduced by enforcing that two successively saved conformations display a coordinate RMSD larger than 2 Å. A maximum number of 10^9 saved conformations was permitted for each iBP run. The solutions were stored in a multiframe dcd format[46].

The generated conformations of neighbouring peptide fragments in the protein sequence were then assembled by superimposing the three last and initial residues of the fragments successive in the sequence. The dummy N and C terminal residues added at the step of individual iBP calculations were not included in the superimposition and are removed after the assembly step. During the superimposition of backbone atoms located in the three superimposed residues, the atom number for which the smallest distance was observed between corresponding atoms in the two peptides was used to decide where to stop with the first peptide and to continue with the second one. The assembled conformation was then checked to verify whether there is no Cα atoms closer than 1 Å. The assembled conformations not verifying this condition were pruned from the calculation. The fragment assembly was implemented using python scripting based on the MDAnalysis 0.1[47,48] and numpy 1.7.1[49] packages.

To scale down the combinatorial explosion of the calculation, a clustering approach, the **S**elf-**O**rganizing **M**ap (SOM)[32,50–52], which is an artificial neural network (ANN) trained using unsupervised learning, was used to reduce the number of conformations. SOM displays the advantage with respect to the k-means clustering approach that it does not require the predetermined knowledge of the number of clusters. The SOM approach was used after each iBP calculation or assembly step as soon as the number of saved conformations was larger than 1000. The conformations are encoded from the distances d_{ij} calculated between the n Cα atoms by diagonalizing the covariance matrix C:

$$C_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{n}(d_{i,k}-\bar{d}_i)(d_{l,j}-\bar{d}_j) \tag{1}$$

where $\bar{d}_s = \frac{1}{n}\sum_{p=1}^{n} d_{s,p}$. The information contained in the matrix C is equivalent to its four largest eigenvalues along with the corresponding eigenvectors, corresponding to an input vector of length $4(n+1)$. The input vectors obtained on the saved conformations are used to train a periodic Euclidean 2D self-organizing map (SOM), which corresponds to a three-dimensional matrix. The first two matrix dimensions were chosen to be 100×100 and define the map size, the third dimension being equal to $4(n+1)$. Each vector along the third dimension defines a neuron of the map. The neurons of the self-organizing map are initialized with a random uniform distribution covering the range of values of the input vectors previously obtained from the covariance matrix C. At each step, an input vector is presented to the map, and the neuron closest to this input is updated. The maps were trained as previously described[32,53].

Once the SOM has been determined, representative conformations are extracted from the conventional **U**nified **d**istance **m**atrix (U-matrix) calculated from the final SOM neurons. For each neuron ν, the corresponding U-matrix element is calculated as the average Euclidean distance between the neuron μ and its eight immediate neighbors:

$$\text{U-matrix}\,(\mu) = \frac{1}{8}\sum_{\nu \in N(\mu)} d(\nu,\mu) \tag{2}$$

where $N(\mu)$ is the set of neighbors, and $d(\nu,\mu)$ is the Euclidean distance between the neurons μ and ν. The neurons corresponding to local minima of the U-matrix, and thus to local maxima of conformational homogeneity, are extracted and the protein conformation being first stored in that neuron is saved. The conformations generated during the iBP or assembly steps are finally replaced by the sets of representative conformations extracted from local minima of U-matrix.

Assembling P1 and P2 to Lnk. The PDZ domains P1 and P2 were added to each Lnk conformation according to the following procedure. The domain P1 extracted from the P1P2 HADDOCK complex[27], was assembled to Lnk conformation (Figure S1A) by superimposing the backbone atoms of the N-terminal segment 232-238 of Lnk to the backbone atoms of the segment 232–238 present at the C terminal extremity of the domain P1 (Figure S1A). The superimposition was realized using a 3D grid generated within the following limits: the maximum and minimum values extracted from the atomic coordinates of atoms Cα of Lnk, from which the gyration radius R_g of the domain P1 was added and subtracted (Figure S1B). The grid points which were closer than R_g from the atom Cα of residue 238 in Lnk and which were more apart than 2 Å from any atom Cα of Lnk, were selected for further processing.

Then, the geometric center of P1 was placed at each of the grid points previously generated. The two vectors V_{P1} and V_{Lnk} connecting the atoms Cα of residues 232 and 238 in P1 and in Lnk were calculated and the 3D rotation matrix allowing to transform V_{P1} into V_{Lnk} determined (Figure S1C). This rotation matrix was then applied on the atomic coordinates of P1. The vectors V_{232} (respectively V_{238}) connecting atom Cα of residue 232 (respectively 238) in Lnk and P1 were calculated and the P1 atomic coordinates were translated along the vector: $\mathbf{T} = \frac{1}{2}(\mathbf{V_{232}} + \mathbf{V_{238}})$ (Figure S1D). The translated position of P1 was accepted according to two conditions: (1) the number of atomic clashes between the P1 and Lnk, determined as the number of Cα pairs closer than 1 Å, was smaller than 3; (2) the average distances between atoms Cα of residues 232 in Lnk and P1 and between atoms Cα of residues 238 in Lnk and P1 were smaller than 6 Å. The P1-Lnk protein was then generated by merging the atomic coordinates of P1 for residues 136–238 and the atomic coordinates of Lnk for residue 239–280. Sidechains were generated for the linker in the selected solutions and molecular dynamics trajectories of 100 ps were recorded using the protocol described in Supplementary Material (Section "Molecular dynamics refinement in implicit solvent") with positional restraints on the backbone heavy atoms for residues 225–280. At the end, conformations for which the total energy was higher than − 2000 kcal/mol were discarded, which produced 153 conformations of P1-Lnk.

The domain P2 was then assembled to P1Lnk. During the previous structural work on P1LnkP2[27], a model of the complex P1/P2, in which the two PDZ domains were not covalently bound, was built using the HADDOCK server[54–56] from the RDC restraints and the previously determined structures of the human whirlin PDZ domains (PDB entries 1UEZ and 1UF1). This model was used here to impose a starting point for the relative orientations of the P1 and P2 domains inside P1LnkP2. The P1P2 HADDOCK complex and the P1-Lnk (Figure S1E) conformations were superimposed along the backbone atoms of the residues 136–238, corresponding to P1 and the hairpin Hp1 (Figures 1A and S1F). Second, the vector V_{276} relating the positions of the atom Cα of residue 276 in the HADDOCK P1P2 complex and in the P1-Lnk conformation was calculated. A translation defined by this vector was applied to the domain P2 present in the complex P1P2 and the translated P2 was concatenated to P1-Lnk (Figure S1G). The number of clashes was determined as the number of atoms closer than 1 Å between residues 136-275 (P1-Lnk) and 276–376 (P2). If less than 50 clashes were observed, the construction was kept and submitted to a 100ps molecular dynamics, following the procedure described in Supplementary Material (Section "Molecular dynamics refinement in implicit solvent"). During that run, positional restraints on all atoms except in residues 274–278 were applied. After this step, 87 conformations of P1LnkP2 for which the total energy was better than − 2000 kcal/mol were selected, and after manual inspection, four additional conformations displaying knots were removed from the set of solutions. Finally, 83 conformations were kept to perform further refinement.

10

References

1. Dunker, A. K. *et al.* What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord Proteins* **1**, e24157 (2013).
2. Csizmok, V., Follis, A. V., Kriwacki, R. W. & Forman-Kay, J. D. Dynamic protein interaction networks and new structural paradigms in signaling. *Chem. Rev.* **116**, 6424–6462 (2016).
3. Bah, A. & Forman-Kay, J. D. Modulation of intrinsically disordered protein function by post-translational modifications. *J. Biol. Chem.* **291**, 6696–6705 (2016).
4. Ali, M., Simonetti, L. & Ivarsson, Y. Screening intrinsically disordered regions for short linear binding motifs. *Methods Mol. Biol.* **2141**, 529–552 (2020).
5. Meyer, K. & Selbach, M. Peptide-based interaction proteomics. *Mol. Cell. Proteomics* **19**, 1070–1075 (2020).
6. Yan, J., Dunker, A. K., Uversky, V. N. & Kurgan, L. Molecular recognition features (MoRFs) in three domains of life. *Mol. BioSyst.* **12**, 697–710 (2016).
7. Cumberworth, A., Lamour, G., Babu, M. M. & Gsponer, J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochem. J.* **454**, 361–369 (2013).
8. Feldman, H. J. & Hogue, C. W. A fast method to sample real protein conformational space. *Proteins* **39**, 112–131 (2000).
9. Bernadó, P. *et al.* A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. *Proc. Natl. Acad. Sci. USA* **102**, 17002–17007 (2005).
10. Jha, A. K., Colubri, A., Freed, K. F. & Sosnick, T. R. Statistical coil model of the unfolded state: resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA* **102**, 13099–13104 (2005).
11. Lavor, C., Liberti, L., Maculan, N. & Mucherino, A. The Discretizable Molecular Distance Geometry Problem. *Comput. Optim. Appl.* **52**, 115–146 (2012).
12. Lavor, C., Liberti, L. & Mucherino, A. The interval Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances. *J. Glob. Optim.* **56**, 855–871 (2013).
13. Liberti, L., Lavor, C. & Mucherino, A. The discretizable molecular distance geometry problem seems easier on proteins. In *Distance Geometry: Theory, Methods and Applications* (eds Mucherino, L. *et al.*) 47–60 (Springer, Berlin, 2014).
14. Liberti, L., Lavor, C., Maculan, N. & Mucherino, A. Euclidean distance geometry and applications. *SIAM Rev.* **56**, 3–69 (2014).
15. Lavor, C., Alves, R., Figueiredo, W., Petraglia, A. & Maculan, N. Clifford algebra and the discretizable molecular distance geometry problem. *Adv. Appl. Clifford Algebras* **25**, 925–942 (2015).
16. Cassioli, A. *et al.* An algorithm to enumerate all possible protein conformations verifying a set of distance constraints. *BMC Bioinf.* **16**, 23–37 (2015).
17. Mareuil, F., Malliavin, T. E., Nilges, M. & Bardiaux, B. Improved reliability, accuracy and quality in automated NMR structure calculation with ARIA. *J. Biomol. NMR* **62**, 425–438 (2015).
18. Feldman, H. J. & Hogue, C. W. Probabilistic sampling of protein conformations: new hope for brute force?. *Proteins* **46**, 8–23 (2002).
19. Worley, B. *et al.* Tuning interval branch-and-prune for protein structure determination. *J. Global Optim.* **72**, 109–127 (2018).
20. Malliavin, T. E., Mucherino, A., Lavor, C. & Liberti, L. Systematic exploration of protein conformational space using a distance geometry approach. *J. Chem. Inf. Model.* **59**, 4486–4503 (2019).
21. Matsushima, N. *et al.* Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins. *Curr. Protein Pept. Sci.* **9**, 591–610 (2008).
22. Troncoso, M. F., Elola, M. T., Croci, D. O. & Rabinovich, G. A. Integrating structure and function of 'tandem-repeat' galectins. *Front. Biosci. (Schol Ed)* **4**, 864–887 (2012).
23. Sang, M. *et al.* LIM-domain-only proteins: multifunctional nuclear transcription coregulators that interacts with diverse proteins. *Mol. Biol. Rep.* **41**, 1067–1073 (2014).
24. Shaw, N. & Liu, Z. J. Role of the HIN domain in regulation of innate immune responses. *Mol. Cell. Biol.* **34**, 2–15 (2014).
25. Troncoso, M. F. *et al.* Galectin-8: a matricellular lectin with key roles in angiogenesis. *Glycobiology* **24**, 907–914 (2014).
26. Wang, L. *et al.* Identification of whirlin domains interacting with ESPIN: A study of the mechanism of Usher syndrome type II. *Mol. Med. Rep.* **20**, 5111–5117 (2019).
27. Delhommel, F. *et al.* Structural characterization of whirlin reveals an unexpected and dynamic supramodule conformation of its PDZ tandem. *Structure* **25**, 1645–1656 (2017).
28. Shen, Y. & Bax, A. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol. Biol.* **1260**, 17–32 (2015).
29. Nielsen, J. T. & Mulder, F. A. There is diversity in disorder—"In all chaos there is a cosmos, in all disorder a secret order". *Front Mol Biosci* **3**, 4 (2016).
30. Berjanskii, M. V. & Wishart, D. S. A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.* **127**, 14970–14971 (2005).
31. Camilloni, C., De Simone, A., Vranken, W. F. & Vendruscolo, M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* **51**, 2224–2231 (2012).
32. Bouvier, G. *et al.* Functional motions modulating VanA ligand binding unraveled by self-organizing maps. *J. Chem. Inf. Model.* **54**, 289–301 (2014).
33. Zweckstetter, M. & Bax, A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J. Am. Chem. Soc.* **122**, 3791–3792 (2000).
34. Schwieters, C. D., Bermejo, G. A. & Clore, G. M. Xplor-NIH for molecular structure determination from NMR and other data sources. *Protein Sci.* **27**, 26–40 (2018).
35. Svergun, D. I., Barberato, C. & Koch, M. CRYSOL—a Program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* **28**, 768–773 (1995).
36. Manalastas-Cantos, K. *et al.* ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* **54**, 343–355 (2021).
37. Köfinger, J. *et al.* Efficient ensemble refinement by reweighting. *J. Chem. Theory Comput.* **15**, 3390–3401 (2019).
38. Ihms, E. C. & Foster, M. P. MESMER: minimal ensemble solutions to multiple experimental restraints. *Bioinformatics* **31**, 1951–1958 (2015).
39. Neal, S., Nip, A. M., Zhang, H. & Wishart, D. S. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR* **26**, 215–240 (2003).
40. Kauffmann, C., Kazimierczuk, K., Schwarz, T. C., Konrat, R. & Zawadzka-Kazimierczuk, A. A novel high-dimensional NMR experiment for resolving protein backbone dihedral angle ambiguities. *J. Biomol. NMR* **74**, 257–265 (2020).
41. Kauffmann, C., Zawadzka-Kazimierczuk, A., Kontaxis, G. & Konrat, R. Using cross-correlated spin relaxation to characterize backbone dihedral angle distributions of flexible protein segments. *ChemPhysChem* **22**, 18–28 (2021).
42. Borgia, A. *et al.* Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).

43. Gomes, G. W. *et al.* Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* **142**, 15697–15710 (2020).
44. Levinthal, C. Are there pathways for protein folding?. *J. Chem. Phys.* **65**, 44–45 (1968).
45. Engh, R. & Huber, R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr. A* **47**, 392–400 (1991).
46. Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
47. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).
48. Gowers, R.J., *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In S. Benthall and S. Rostrup, editors, *Proceedings of the 15th Python in Science Conference*, 98–105 (2016).
49. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362. https://doi.org/10.1038/s41586-020-2649-2 (2020).
50. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
51. Kohonen, T. *Self-organizing Maps* (Springer, Heidelberg, 2001).
52. Miri, L. *et al.* Stabilization of the integrase-DNA complex by Mg2+ ions and prediction of key residues for binding HIV-1 integrase inhibitors. *Proteins* **82**, 466–478 (2014).
53. Spill, Y., Bouvier, G. & Nilges, M. A convective replica-exchange method for sampling new energy basins. *J. Comput. Chem.* **34**, 132–140 (2013).
54. Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
55. van Zundert, G. C. P. *et al.* The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
56. Saponaro, A., Maione, V., Bonvin, A. M. J. J. & Cantini, F. Understanding docking complexes of macromolecules using HADDOCK: the synergy between experimental data and computations. *Bio Protoc* **10**, e3793 (2020).
57. Foundation TD. LibreOffice [Internet]. 2020. Available from: https://www.libreoffice.org/.
58. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).
59. Warren L. DeLano. *The PyMOL Molecular Graphics System, Version 1.2r3pre*. Schrödinger, LLC.
60. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Acknowledgements

Author contributions

T.M. did the calculations, analyzed the results and wrote the manuscript.

Competing interests

The author declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-96370-z.

Correspondence and requests for materials should be addressed to T.E.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.