Characterization of the conformational space of intrinsically disordered proteins

Georg Daniel Förster, Jérôme Idier, Leo Liberti, Antonio Mucherino, Jung-Hsin Lin, Thérèse Malliavin

XIIth International Conference NMR: a tool for biology, Institut Pasteur, May 2nd-4th 2022

Characterization of the conformational space of intrinsically disordered proteins

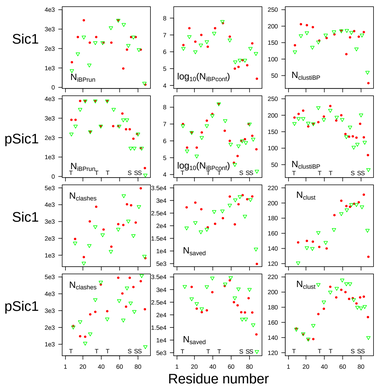
Daniel Förster[a], Jérôme Idier[b], Leo Liberti[c], Antonio Mucherino[d], Jung-Hsin Lin[e] & Thérèse E. Malliavin[f,g,h]

[a] UMR7374 Interfaces, Confinement, Matériaux et Nanostructures, Université d'Orléans, France; [b] UMR6004 Laboratoire des Sciences du Numérique de Nantes, France; jerome.idier@ls2n.fr [c] LIX UMR 7161 CNRS École Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France; [d] IRISA, University of Rennes 1, France; [e] Biomedical Translation Research Center, Academia Sinica, Taiwan; [f] Institut Pasteur, Université Paris Cité, CNRS UMR3528, Unité de Bioinformatique Structurale, F-75015 Paris, France; [g] Laboratoire de Physique et Chimie Théoriques (LPCT), University of Lorraine, Vandoeuvre-lès-Nancy, France; [h] Laboratoire International Associé, CNRS and University of Illinois at Urbana-Champaign, Vandoeuvre-lès-Nancy, France; therese.malliavin@univ-lorraine.fr

Abstract

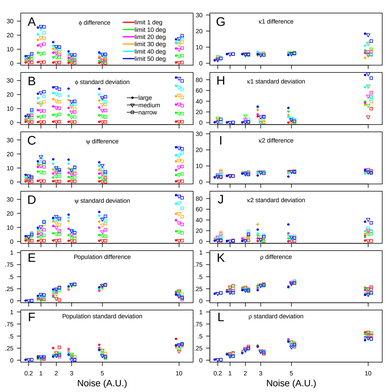
The TALOS-N neural network has been developed [7] for inferring information about the backbone dihedral angles from NMR chemical shifts. The chemical shifts can be measured with an equal precision in folded proteins as well as in intrinsically disordered proteins. For folded proteins, TALOS-N was initially proposed in order to provide additional restraints for structure calculation. Nevertheless, the approach TAiBP was recently proposed [5, 6] to use the TALOS-N likelihood maps as distributions of the backbone angle values in the context of disordered protein regions. This was made possible thanks to the availability of a branch-and-prune algorithm, iBP [3], allowing a systematic enumeration of protein conformations. We are presenting here the results obtained by the TAiBP algorithm on two intrinsically disordered proteins described in the Protein Ensemble Database [4], along with an original finite mixture model allowing the determination of the relative populations of conformations from the TALOS-N likelihood maps. The populations obtained using this mixture model will be compared to those determined from SAXS measurements [2]. The use of conformations obtained from TAiBP along with the selection of the most populated conformations permit a low-resolution description of the conformational space of the studied proteins.

Enumeration of protein conformations



The TAiBP protocol [5] have been applied on Sic1 and pSic1, corresponding to unphosphorylated and phosphorylated states of an intrinsically disordered protein (IDP), using as input the probability maps outputs of the prediction neural network TALOS-N [7]. Using a threshold on the probability maps, boxes largely encompassing the most probable regions have been manually determined. Restraints on backbone angles ϕ and ψ were derived from box definitions and used as input of TaiBP calculation. For the iBP and assembly steps forming the TAiBP approach, two duplicate runs (Sic1^1, Sic1^2, pSic1^1, pSic1^2) marked in colors red and green, produces parameter values similar in most of the protein sequence. For the iBP steps, three parameters were compared (first and second lines) along the residue number located at the middle of each fragment: the number of individual iBP runs (N_{iBPrun}) displays the largest observed values (3888) around the positions of phosphorylated Threonines in agreement with the larger generic boxes used in these protein regions. For every calculation, the number of saved conformations $N_{iBPconf}$ is smaller than 10^9, which is the maximum possible number of solutions: all individual iBP trees have been thus completely parsed. The numbers of clustered conformations ($N_{clustiBP}$) display for Sic1 a larger reduction of conformations in the region of residues 60-90 which is the sign of these conformations are more diverse in Sic1 than in pSic1. For the assembly step, three parameters are plotted (third and fourth lines): the number of conformations rejected due to Cα atoms closer than 1Å ($N_{clashes}$) correspond between 10% and 15% of the number of saved conformations (N_{saved}). The profiles of the number of clustered conformations (N_{clust}) are different for Sic1 and pSic1 and agrees with the smaller gyration radii observed by pSic1 (data not shown).

Validation of the finite mixture model on synthetic data



The populations of conformations were determined from the Ramachandran probability maps using a finite mixture model named RamaMix. The efficiency of this approach was validated on synthetic data, formed by 15 couples of ϕ, ψ values, more or less scattered, (●, ▽, □) as well as randomly chosen populations and using different noise levels. During each RamaMix run, several upper limits were imposed to the drift of the backbone angles during the optimization, with values of: 1°, 10°, 20°, 30°, 40° and 50°. For each Ramachandran synthetic map, each noise level and each drifting limit value, one hundred runs are performed producing sets of backbone angles (ϕ_0 and ψ_0), von Mises parameters describing probability densities on a torus ($\kappa1$, $\kappa2$ and ρ) and populations γ_q. Over the 12600 individual RamaMix runs, only 275 runs were terminated without convergence of the optimization. Averages and standard deviations were calculated from the sets of obtained parameters. The differences between the averaged and the input values, as well as the standard deviations are used to evaluate RamaMix. The differences between average and initial populations (panel E) as well as the standard deviations of populations reveals that the determination of populations is not much influenced by the level of noise, but the population values are rather qualitative. The efficiency of the determination of backbone angles (panels A-D) is much influenced by the drifting limit imposed on the ϕ, ψ values: this would support not allowing large drift for the calculations. The parameters describing the von Mises distribution (panels G-L) display contrasted results: the differences are larger for ρ than for $\kappa1$ and $\kappa2$. For $\kappa1$ and $\kappa2$, the standard deviations are much larger than the differences whereas they are similar for ρ.

Conformations and populations selected by fitting of the Ramachandran maps using RamaMix are displayed in the Table. For each set of protein conformations, 100 runs were performed starting from random values for the populations. The backbone angles ϕ and ψ were allowed to move up to 15°. The populations of conformations for the converged runs were averaged and these mean values are given as percentages in the Table along with the corresponding standard deviation values. The labels of conformations also selected by fitting of SAXS curves are written in bold.

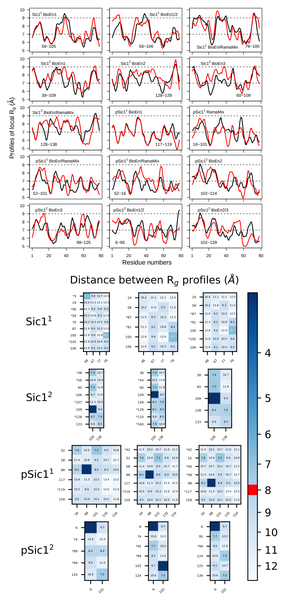
A. Sic1[1] conformation numbers	populations percentages	C. pSic1[1] conformation numbers	populations percentages
79	44.7 ± 0.5	**98**	23.2 ± 1.4
77	23.4 ± 0.6	154	22.7 ± 2.0
67	21.7 ± 0.4	101	21.2 ± 0.7
46	10.2 ± 0.4	135	19.2 ± 3.3
		16	13.7 ± 1.0
B. Sic1[2] conformation numbers	populations percentages	D. pSic1[2] conformation numbers	populations percentages
109	67.8 ± 2.9	**6**	59.2 ± 3.7
138	32.1 ± 0.8	**102**	40.7 ± 3.0

Low resolution model of the conformations

By analogy to the cross-sectional gyration radius, we propose here the profiles of local gyration radii to describe the local variation in the shape of conformations. These profiles P_q of local gyration radii are calculated along residue number n for each conformation q in the following way:

$$P_q(n) = \sqrt{\frac{1}{N_n} \sum_{i=n-N_{win}}^{n+N_{win}} (\mathbf{X}_i - \mathbf{X}_n^{ave})^2} \qquad (1)$$

where $\mathbf{X}_i$ represents the vector of atomic coordinates for the backbone atoms of residue i in the range $n - N_{win}$, $n + N_{win}$, and N_{win}=5 is the residue window around n on which a local gyration radii is calculated, N_n being the number of backbone atoms located in this window. $\mathbf{X}_n^{ave}$ is the coordinate vector of the centroid of the atomic coordinates of the backbone atoms of residues in the range $n - N_{win}$, $n + N_{win}$.



Examples of profiles P_q superimposition (Figure top) have been picked up for distances between them in the 4.0-7.9 Å range. They give an estimation of the connection between the information related to atomic coordinates and the distance between the profiles. The examination of Figure reveals that the profile peaks are mostly located at similar places in the protein sequence. This gives a qualitative description of the conformations separated in extended regions (profile maxima) and in aggregated regions (profile minima).

The comparison between BioEn and RamaMix fitting (Figure bottom) displays contrasted behaviors between the duplicated TAiBP runs. For Sic1^2 and pSic1^2, all RamaMix conformations display profiles closer than 8 Å to the profiles of BioEn conformations. For pSic1^1, this is also the case for three RamaMix conformations (16, 98, 101) over five. For Sic1^1, only the conformation 79 display profile distances smaller than 8 Å for the three comparisons.

References

[1] Gomes et al. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J Am Chem Soc*, 142:15697, 2020.

[2] Köfinger et al. Efficient Ensemble Refinement by Reweighting. *J Chem Theory Comput*, 15:3390, 2019.

[3] Lavor et al. The interval Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances. *J Glob Optim*, 56:855, 2013.

[4] Lazar et al. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res*, 49:D404, 2021.

[5] Malliavin et al. Systematic Exploration of Protein Conformational Space Using a Distance Geometry Approach. *J Chem Inf Model*, 59:4486, 2019.

[6] Malliavin. Tandem domain structure determination based on a systematic enumeration of conformations. *Sci Rep*, 11:16925, 2021.

[7] Shen and Bax. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol Biol*, 1260:17, 2015.

Acknowledgements

Funding: ANR-19-CE45-0019 (multiBioStruct); Institut Pasteur; CNRS; Ecole Polytechnique and University of Rennes. Tanja Mittag is acknowledged for providing SAXS data recorded in the conditions described in Ref. [1]. Cyprien Bertran is acknowledged for fruitful discussions.