Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 677–681
DOI: 10.15439/2024F8235
ISSN 2300-5963 ACSIS, Vol. 39

Impact of Local Geometry on Methods for Constructing Protein Conformations

W. Da Rocha,* T.E. Malliavin,† A. Mucherino,‡ L. Liberti*

*LIX, École Polytechnique, Palaiseau, France.
{ wagner.rocha, liberti } @lix.polytechnique.fr

†LPCT-CNRS and University of Lorraine, Vandoeuvre-lès-Nancy, France.
therese.malliavin@univ-lorraine.fr

‡IRISA, University of Rennes, Rennes, France.
antonio.mucherino@irisa.fr

***Abstract*—The prediction of protein structures is an important problem in molecular biology. In spite of the large efforts from the research community, and of the recent development of artificial intelligence tools specifically designed for this problem, a complete and definitive solution to the problem has not been found yet. This work is based on the observation that many tools for the prediction of protein conformations rely on both local and non-local geometrical information, even though the non-local information can be very hard to identify within the desired precision in some particular situations. For this reason, we explore in this work the effect of local geometry on methods capable of constructing protein conformations. This initial study has the final aim of devising new alternative methods where the predictions may be guided mainly by the local geometry of proteins.**

I. Introduction

THE prediction of suitable conformations for a given protein is of fundamental importance in science, in particular in the context of drug design. Since several years the research community has been working on this topic, with the final goal of understanding how these three-dimensional conformations can be "predicted" by using some of the information that can be obtained through experimental techniques. When the predictions rely solely on the protein sequence (i.e. on the list of *amino acids* forming the main protein chain), then it is common to refer to the problem of identifying these possible conformations as the "protein folding" problem [5].

In spite of the large efforts in this scientific domain, the protein folding problem remained for several years, as long as general instances are concerned, among the practically intractable problems. Experimental techniques that are able of providing additional information about the molecules (and not only its amino acid sequence) were meanwhile developed, and methods and algorithms were thus proposed that are capable of determining protein conformations from these experimental data. One example, on which we have been working in the past 15 years, is given by the experimental technique based on Nuclear Magnetic Resonance (NMR) [2], where a Distance Geometry Problem (DGP) [15], [17] is formulated for the determination of the protein conformations.

These methods exploit both local structural information, as well as long-range proximity measures [12]. Local structural information can for example be deduced from the study of the chemical structure of each amino acid: if two atoms are chemically bonded, then it is possible to *guess*, in a rather precise way, the distance separating the two atoms. Force fields such as AMBER [4] and PARALLHDG [6] collect a certain number of parameters which also comprise this kind of local proximity information. Naturally, the given values for such parameters are not wholly satisfied in all proteins. In fact, terms of energy functions given by such force fields are actually able to give a measure on the variations of these values in protein conformations.

It is common to talk about long-range proximity measures when we can obtain estimates on the distances between two atoms belonging to two amino acids that may be separated by several other amino acids in the protein sequence. The experiments based on NMR techniques (already mentioned above), for example, are able to give estimates on such long-range distances, and most commonly between pairs of hydrogen atoms [8]. Alternatively, methods based on multiple sequence alignments [21] can also provide long-range distance information, but they are likely to give imprecise results for particular cases of proteins [20]. A real challenge for both NMR experiments and methods based on protein sequence alignments are the so-called Intrinsically Disordered Proteins (IDPs) [19].

Nevertheless, AlphaFold (the release of the version 3 is very recent, see [1]), the very well-known Artificial Intelligence (AI) tool for protein folding, strongly relies on long-range proximity information, normally obtained from protein sequence alignments. The success of AlphaFold is therefore strongly dependent on the availability of such long-range restraints, and its actual success rate can therefore depend on the number of alignments that it is possible to exploit in order to derive the long-range proximity measures.

In this work, we intend investigating the impact of local geometry in the determination of protein folds. Our work is motivated by some previous analysis performed by some of us [9] where we have identified some particular situations in which the local geometry seems to have a larger impact on the protein folds than long-range distances (which were identified though NMR experiments in that work). Some initial investigations in line with the present work were already conducted and published in [10]. Our work extends those initial studies and uses a larger subset of protein conformations in the computational experiments.

The remainder of the paper is organized as follows. In Section II, we will describe in more detail what we intend by local geometry of protein conformations, and we will explain how to define DGP instances for protein conformations carrying specific local geometry information. In Section III, we will briefly describe the Branch-and-Prune (BP) [14], used for solving the artificially generated DGP instances. Finally, Section IV will present our preliminary computational experiments, and Section V will briefly conclude the paper.

II. Local geometry of proteins

Proteins are defined by one or more sequences of amino acids. In this work, we focus our attention on proteins defined by only one amino acid chain. Amino acids are the building blocks of proteins. There are 20 different amino acids that can be involved in the protein synthesis, and they all have a common part, while another, named the *side chain* of the amino acid, makes each amino acid different from one another. The subset of atoms forming the protein which are not included in the side chains is generally referred to as the *protein backbone*.

The chemical composition for every of the 20 amino acids is a priori known, and therefore the chemical composition of the entire protein can be simply obtained from the amino acid sequence. Part of the local geometry can be as a consequence derived from a simple analysis of the atomic bonds that are present in the structure. As already mentioned, bonded atoms satisfy a relative distance (a "bond length") that is generally considered to depend solely on the type of the two involved atoms.

Similarly, we can extend the same idea to the angle that every triplet of bonded atoms can form (say the atoms are A, B and C, where A is bonded to B, B is bonded to C, and we are interested in the angle in B formed by the segments AB and BC). In this case, we rather talk about "bond angles", and it is again generally supposed that these angles basically depend on only the type of involved atoms. We point out, however, that in the case of bond angles, a larger variation of the angles can be observed around their average value.

The situation is a little more complex when quadruplets of consecutive atoms are defined. They allow us to define the so-called *torsion angles*. Torsion angles exhibit larger variations over the protein conformations, and they are not as regular as bond lengths and angles. However, there are some special cases where we can constrain the values of these angles. One example is given by the torsion angle ω crossing a peptide bond (from the C_α Carbon of the amino acid i to the C_α Carbon of the amino acid $i+1$ in the sequence), which is generally fixed and set to $178°$. Another example is given by the protein secondary structures, which strongly restrict the ranges of the torsion angles for every quadruplet of atoms that we can define on the protein backbones. This is a very important result for protein conformations, which was studied for the first time in the well-known Ramachandran map [18]. In the following, we will employ the typical notations ϕ, ψ and ω for the torsion angles that we can define on the protein backbones.

In this work, we investigate the dependence of local geometry in proteins (bond lengths and angles, as well as torsion angles) on parameters other than the simple atom type, as it was instead supposed in the seminal works of Engh and Huber [6]. In particular, we will compare the three following **setup**s:

1. local geometry is unique for every protein and cannot be predicted by the analysis of the protein sequence;
2. local geometry can be predicted by using the information about the secondary structures related to the protein chain, together with the atom type;
3. local geometry can be predicted by the simple analysis of the atom types (as in Engh and Huber's works).

In order to study and compare these three setups, we will artificially generate three different sets of DGP instances, which we will solve by the BP algorithm briefly summarized in Section III. Our computational experiments will be then presented and commented in Section IV.

III. An implementation of the BP algorithm

In this section, we will first of all introduce the DGP in formal terms, and we will briefly describe a well-known algorithm for the solution of DGP instances that can be *discretized*, and finally mention to the specific implementation of the algorithm that we will use in our computational experiments.

Let $G = (V, E, d)$ be a simple weighted undirected graph, where vertices represent the atoms of our proteins, and the existence of an edge between two atoms indicate that their relative distance is known [15]. The weight function d associates the numerical value of the distance to every edge of E. This numerical value $d(u, v)$ can be either exact (i.e. very precise), so that it can represented by a singleton, or rather imprecise and hence represented by a real-valued interval $[\underline{d}(u, v), \bar{d}(u, v)]$. Let E' be the subset of the edge set E containing only the exact distances.

Given a simple weighted undirected graph $G = (V, E, d)$, the Distance Geometry Problem (DGP) in dimension 3 asks whether a graph embedding

$$x : v \in V \longrightarrow x_v \in \mathbb{R}^3$$

exists such that

$$\forall \{u, v\} \in E, \quad ||x_u - x_v|| \in d(u, v), \tag{1}$$

where $||\cdot||$ represents the Euclidean norm. We say that the graph embedding x is a *realization* of the graph when it satisfies all the constraints in Eq. (1).

In the past years, some of us have been focusing on a special class of DGP instances where the search space can be discretized and reduced to a tree [13]. Let $G[\cdot]$ be the subgraph of G induced by a subset of vertices of V. In formal terms, a given DGP instance can be discretized (so that it represents an instance of the Discretizable DGP, or DDGP) when there exists a vertex ordering on V such that the following two assumptions are satisfied:

(a) $G[\{1,2,3\}]$ is a clique whose edges are in E';
(b) $\forall v \in \{4,\ldots,|V|\}$, there exist $u_1,u_2,u_3 \in V$ such that

- **(b.1)** $u_1 < v$, $u_2 < v$, $u_3 < v$;
- **(b.2)** $\{\{u_1,v\},\{u_2,v\}\} \subset E'$, $\{u_3,v\} \in E$;
- **(b.3)** $d(u_1,u_3) < d(u_1,u_2) + d(u_2,u_3)$.

When the two assumptions **(a)** and **(b)** are satisfied, we can construct a search tree where the candidate positions for every atom are collected on a common tree layer [15]. In our experiments, we consider the vertex ordering defined in [22], which makes an extensive use of repeated vertices in order to achieve a direct branching on the torsion angles ϕ, ψ and ω that we can define in the protein backbones. For strict enough values for these torsion angles, it is in fact possible to avoid branching and hence locally reduce, *a priori*, the tree width.

When the edge $\{u_3,v\}$ is not in E' (see assumption **(b.2)**), the distance $d(u_3,v)$ is represented by an interval. In this situation, the set of possible positions for the atom v is actually continuous, but in some particular conditions (which are satisfied by the instances we use in this work) we can consider to take sample distance values from the original intervals, and to have a dedicated branch in the tree for every extracted sample distance [13]. This methodology introduces an additional factor (given by the number of samples taken from every interval distance) in the combinatorics, but it has the advance to make us deal with more complex instances by using an approach that was initially designed to work in simpler conditions (i.e. with distances that are not affected by uncertainty).

The Branch-and-Prune (BP) algorithm performs a systematic exploration of this search tree. It uses the additional distances, which are not necessary for the construction of the tree, to verify the "feasibility" of the generated atomic positions [14]. This is the so-called *pruning phase* of the algorithm, which is actually very important in BP, because it allows the algorithm to focus the search over the tree branches that contain no infeasibilities.

For more information about the BP algorithm and its previous uses in the context of structural biology, the reader is mainly referred to [7], [16]. In our computational experiments, we will use the implementation of the BP algorithm available on the following GitHub repository:

`https://github.com/tmalliavin/ibp-ng-fullchain`

IV. Computational experiments

In our computational experiments, we have selected a subset of protein conformations from the Protein Data Bank (PDB) [3]. The conformations have been selected in order to satisfy the following properties:

- the conformations are obtained through techniques that are based on X-ray crystallography, with resolution of at least 1.6 Å and crystallographic R factor larger than 0.25;
- the protein sequences (only one chain) are not longer than 100 amino acids;
- the similarity between the amino acid sequences of any pair of proteins is smaller than 20%;
- the molecules do not contain *cis* peptide bonds;
- at least two secondary structure elements (α-helix or β-strand) are present in the protein.

Our subset finally contains 308 protein conformations, and we consider the three main setups listed in Section II for the generation of our artificial instances. When we use **setup 1**, we suppose that the local geometry is unique for every protein, and hence we extract the information from the original PDB conformations. When we use **setup 2**, the distances and angles that we use to define our instances are averaged over the Hollingsworth's regions [11], which provide a finer-grained partitioning of initial Ramachandran regions. Finally, we use the values proposed by Engh and Huber's works [6] under the hypothesis that they can only depend on the atom types (our **setup 3**). Notice that these setups can be mixed so that a different one can be considered for a different kind of local information. Details for each set of performed experiments are given in the caption of Fig. 1.

The protein conformation have been reconstructed using "one-shot" BP runs. The branching phase is performed with a discretization factor allowing to have a variation on the torsion angles ϕ and ψ of magnitude about 5°. The ω values are instead used in the pruning phase. A scaling factor of 0.8 is applied to van der Waals radii in order to introduce lower bounds on unknown distances (atoms cannot be closer than a certain threshold when they are not chemically bonded), and the error tolerance ϵ is set to 0.1. The runs are stopped as soon as the first solution is constructed.

Fig. 1 displays the distributions of the root-mean-square deviation (RMSD, Å) between the atomic coordinates related to the solutions found by the BP algorithm and the original PDB conformations. As expected, the best results in terms of RMSD are obtained when the local geometry is extracted from the original PDB conformations (Fig. 1**(a)**). Interestingly, when we consider the bond lengths from Engh and Huber's works, a similar distribution is obtained (data not shown), implying therefore that the bond lengths have actually little impact on the reconstruction process for the conformations. The influence of the bond angles appears instead to be more important.

When the torsion angles ω are imposed to 178° degrees, we can notice a large increase of the RMSD values, reaching values of 10 or even 12 Å (see Fig. 1**(b)**). This result shows

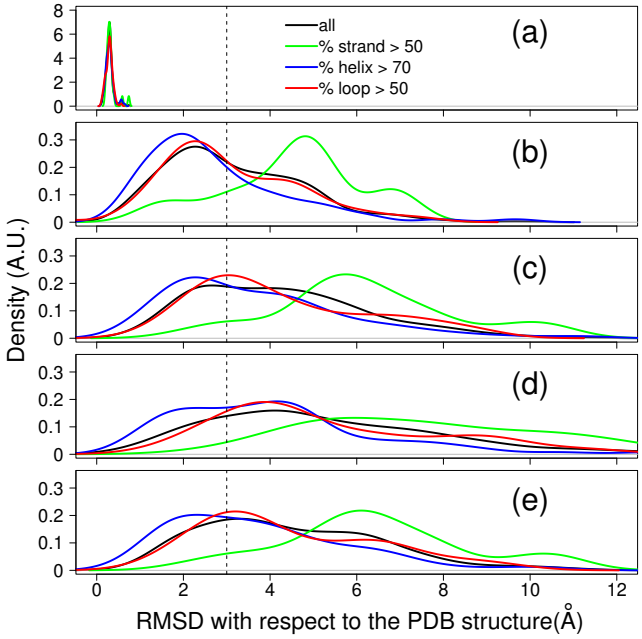

Fig. 1. Distribution of the root-mean-square deviation (Å) between the original PDB conformations and those reconstructed using the BP algorithm. From up to down: **(a)** setup 3 for bond lengths, setup 1 for the remaining local geometry; **(b)** as previous one but with ω angles fixed to 178°; **(c)** setup 2 for the entire local information; **(d)** as previous one but with ω angles set to 178°; **(e)** setup 3 for the entire local geometry. The dashed vertical line is placed at 3 Å.

the importance of the local variability in the peptide bond geometry.

We can remark moreover that some RMSD distributions strongly depend on the secondary structures that are actually contained in the conformations. Proteins with a large percentage of α-helices (blue lines) or mostly containing loops (red lines) are shifted towards smaller RMSD values. By contrast, the structures containing mostly β-strands (green lines) exhibit RMSD values larger than 3 Å. The calculations of conformations based on Ramachandran regions with ω values extracted from the PDB conformations (see Fig. 1**(c)**) display slightly better RMSD values than the ones using the fixed value of 178° (see Fig. 1**(d)**).

Finally, when it is supposed that local geometry only depends on the atom types (see Fig. 1**(e)**), then the observed RMSD values are even larger. These results thus underline the negative impact of uniform local geometry on the reconstructed conformations.

V. Conclusions

We have presented a study on the local geometrical information of protein conformations. Even though we are aware that the local and global geometry are likely to be highly entangled in proteins, this work consisted in investigating how much the local information can have an impact on the protein folds. This study, together with others that we plan to perform in the near future, can potentially help us in developing new methods and algorithms for the construction of protein conformations which mainly (or even solely) use information about the local geometry of proteins.

Acknowledgments

This work was partially supported by the CNRS (ITINERANCE and IRP projects), Lorraine University, IRISA, ANR PRCI multiBioStruct (ANR-19-CE45-0019). High Performance Computing resources were provided by the EXPLOR center at Lorraine University (2022CPMXX2687).

References

[1] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, S.W. Bodenstein, D.A. Evans, C-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Zemgulyté, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A.I. Cowen-Rivers, A. Cowie, M. Figurnov, F.B. Fuchs, H. Gladman, R. Jain, Y.A. Khan, C.M.R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E.D. Zhong, M. Zielinski, A. Zídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J.M. Jumper, *Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3*, to appear in Nature, accelerated preview on `Nature.com` published on May 8, 2024.

[2] F.C.L. Almeida, A.H. Moraes, F. Gomes-Neto, *An Overview on Protein Structure Determination by NMR, Historical and Future Perspectives of the Use of Distance Geometry Methods*. In: [17], 377–412, 2013.

[3] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne, *The Protein Data Bank*, Nucleic Acids Research **28**, 235–242, 2000.

[4] D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, P.A. Kollman, *AMBER 2016*, University of California, San Francisco, 2016.

[5] K.A. Dill, S. Banu Ozkan, M. Scott Shell, T.R. Weikl, *The Protein Folding Problem*, Annual Review of Biophysics **37**, 289–316, 2008.

[6] R Engh, R Huber, *Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement*, Acta Crystallographica A **47**, 392–400, 1991.

[7] D. Förster , J. Idier, L. Liberti, A. Mucherino, J-H. Lin, T.E. Malliavin, *Low-Resolution Description of the Conformational Space for Intrinsically Disordered Proteins*, Scientific Reports **12**, 19057, 16 pages, 2022.

[8] P. Güntert, L. Buchner, *Combined Automated NOE Assignment and Structure Calculation with CYANA*, Journal of Biomolecular NMR **62**, 453–471, 2015.

[9] S.B. Hengeveld, T. Malliavin, J.H. Lin, L. Liberti, A. Mucherino, *A Study on the Impact of the Distance Types Involved in Protein Structure Determination by NMR*, IEEE Conference Proceedings, Computational Structural Bioinformatics Workshop (CSBW21), International Conference on Bioinformatics & Biomedicine (BIBM21), online event, 9 pages, 2021.

[10] S.B Hengeveld, M. Merabti, F. Pascale, T.E. Malliavin, *A Study on the Covalent Geometry of Proteins and Its Impact on Distance Geometry*, Lecture Notes in Computer Science **14072** (part 2), F. Nielsen, F. Barbaresco (Eds.), Proceedings of Geometric Science of Information (GSI23), Saint Malo, France, 520–530, 2023.

[11] S.A. Hollingsworth, M.C. Lewis, D.S. Berkholz, W.K. Wong, P.A. Karplus, *(phi,psi) Motifs: a Purely Conformation-based Fine-Grained Enumeration of Protein Parts at the Two-Residue Level*. Journal of Molecular Biology **416**(1), 78–93, 2012.

[12] B. Kuhlman and P. Bradley, *Advances in protein structure prediction and design*, Nature Reviews Molecular Cell Biology **20**, 681–697, 2019.

[13] C. Lavor, L. Liberti, A. Mucherino, *The interval Branch-and-Prune Algorithm for the Discretizable Molecular Distance Geometry Problem with Inexact Distances*, Journal of Global Optimization **56**(3), 855–871, 2013.

[14] L. Liberti, C. Lavor, N. Maculan, *A Branch-and-Prune Algorithm for the Molecular Distance Geometry Problem*, International Transactions in Operational Research **15**, 1–17, 2008.

[15] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean Distance Geometry and Applications*, SIAM Review **56**(1), 3–69, 2014.

[16] T.E. Malliavin, *Tandem Domain Structure Determination based on a Systematic Enumeration of Conformations*, Scientific Reports **11**, 16925, 2021.

[17] A. Mucherino, C. Lavor, L. Liberti, N. Maculan (Eds.), *Distance Geometry: Theory, Methods and Applications*, 410 pages, Springer, 2013.

[18] G.N.T. Ramachandran, V. Sasisekharan, *Conformation of Polypeptides and Proteins*, Advances in Protein Chemistry **23**, 283–437, 1968.

[19] P. Tompa, *Intrinsically Disordered Proteins: a 10-Year Recap*, Trends in Biochemical Sciences **37**(12), 509–516, 2012.

[20] T. Warnow. *Revisiting Evaluation of Multiple Sequence Alignment Methods*, Methods in Molecular Biology **2231**, 299–317, 2021.

[21] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, T. Hwa. *Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing*. PNAS **106**, 67–72, 2009.

[22] B. Worley, F. Delhommel, F. Cordier, T.E. Malliavin, B. Bardiaux, N. Wolff, M. Nilges, C. Lavor, L. Liberti, *Tuning Interval Branch-and-Prune for Protein Structure Determination*, Journal of Global Optimization **72**, 109–127, 2018.