Adjusting Local Conformational Sampling For Fragment Assembly Protein Structure Prediction Based On Secondary Structure Complexity

Jad Abbass[1,2]
[1]Department of Computer Science
Lebanese International University
Bekaa, Lebanon
Jad.abbas@liu.edu.lb

Jean-Christophe Nebel[2]
[2]Faculty of Science, Engineering and Computing
Kingston University
London, UK
j.nebel@kingston.ac.uk

Abstract— **Fragment assembly protein structure prediction is one of the most successful methods whenever reliable templates (for homology-based approaches) and/or massive computational resources (for physics-based approaches) are not available. However, it suffers from important limitations: tremendous search space, energy scores inaccuracy, and consequently the large number of decoys which are needed to be generated. Taking advantage of the different protein sequence-structure complexity shown by the various types of secondary structure, - using Rosetta - we propose to customize the diversity of fragments for each region of the conformation being built. By eventually reducing the size of search space, this approach permits better exploitation of promising areas. Experiments demonstrate the value of the proposed strategy: compared to standard Rosetta's performance in terms of *first model*, accuracy improves significantly (~6%), respectively dramatically (~24%), when using 20,000, resp. 2,000, decoy-based predictions. Furthermore, performance using 2,000 decoys is equivalent to that of standard Rosetta using 20,000 decoys, which means that predictions can be executed on a standard PC instead of a high-performance computing system.**

Keywords— fragment assembly protein structure prediction; Rosetta; 3-mers; 9-mers; protein secondary strucutre;

I. INTRODUCTION

In all living organisms, once a protein's correct folded structure is attained, it is able – in principle - to perform its critical and diverse biochemical reactions. The folding process of a protein starting from a linear chain of amino acids into its native 3D shape takes places *in vivo* on a millisecond timescale. After more than half a century of research, neither full understanding of the folding process nor availability of a 'robust' computerized prediction tool of the native structure for all types of proteins has been achieved yet. Despite enormous efforts in biochemistry and bioinformatics fields, both *in vitro* and *in silico* methods to determine and predict the final structure remain flawed. Despite the high cost in terms of time and money of wet laboratory techniques such as X-ray crystallography, Nuclear Magnetic Resonance (NMR), and Electron Microscopy (EM), they are still prone to errors. On the computational biology side, the recent exploitation of machine learning has led to a series of breakthroughs in protein structure prediction (PSP)[1]. The latest one was recorded very recently by DeepMind in their contribution in the 14th round of the biannual worldwide competition CASP (Critical Assessment of Structure Prediction) [2], where its end-to-end deep learning techniques were able to reach an unprecedented accuracy in the prediction of single-domain proteins. Nonetheless, challenges still exist for a range of proteins including *de novo* designed proteins, quaternary complexes, protein-ligand complexes, and multi-domain proteins [3].

Since knowledge of the spatial data of a protein's tertiary native structure is invaluable to biochemists, for about three decades, bioinformaticians have tackled the challenge of designing computer software that predict a protein's correct native structure from its amino acid sequence. Those computational approaches can be classified within two large categories: Template-Based Modelling (TBM) and template Free Modelling (FM) techniques – also known as *ab initio*, which are the only ones able to solve targets for which no reliable template can be found.

Ab initio methods attempt to build conformations from scratch relying on Anfinsen's theories [4], [5]. Those two theories state that (i) the amino acid sequence is sufficient to infer the corresponding structure since the folding process is a result of biophysical forces and (ii) the native structure corresponds to the lowest free energy. This has given birth to "physics-based methods", which, as the name suggests, rely on the laws of physics as an attempt to mimic the actual folding process. In principle, such simulations are able to reach the native structure, however, their enormous computational needs have restricted their usage to very short proteins and/or on massive distributed computing systems [6].

The alternative to physics-based approaches was introduced more than two decades ago under the category of fragment-based methods, which, until recently, dominated the FM CASP competition. Whilst these methods are able to handle FM targets, they require much less computational resources than 'classical' *ab initio*. In a nutshell, they are based on two main observations. First, as the sequence-structure correlation is stronger for short sequences, the search space at the local level can be narrowed - it is quite important to note that such correlation's strength varies depending on the type of the secondary structure (this point will be elaborated further since it represents the basis of the work presented in this paper). Second, any protein conformation can successfully be built by simply assembling short fragment structures from other proteins whose global shape – known as architecture or fold – could be totally different. The second point is the reason behind

fragment-based methods' ability to infer FM targets. Amongst the most successful fragment assembly PSP, some use short fragments, e.g., Rosetta [7] and Quark [8], while others rely on long fragments, e.g., I-TASSER [9] and Robetta [10]. One should note that adopting long fragments has triggered an endless debate whether to continue to consider such methods in the *ab initio* category. In terms of fragment size, Rosetta – the tool we used to conduct our experiments - relies on two sets of fixed size, i.e., 9 and 3, which are referred to as 9-mers and 3-mers respectively.

Related to the concept of fragments, secondary structures are the main constituents of any protein's tertiary structure. Alpha helices and beta sheets are not only the main secondary structures, but they are also regular, in the sense that, as their overall shape is 'fixed', their degree of flexibility is limited when compared to the irregular secondary structures called coils. Accordingly, coils represent the highest complexity in terms of sequence–to-structure mapping, even for short sequences. In addition, a study has shown that alpha helix fragments' diversity is much lower than beta sheets' [11]. Herein, we take advantage of the different complexity of the various secondary structures to adjust the number/diversity of fragments being inserted during fragment-based PSP. Whilst standard Rosetta uses a fixed number of candidate fragments from its fragments library, i.e., 25 for 9-mers and 200 for 3-mers, we propose to refine this process by selecting a number of candidates reflecting the typical complexity of the predicted local secondary structure.

II. Fragment-based methods and Rosetta

While the term *ab initio* suggests that the amino acid sequence is the main unit of construction and the only source of data, fragment-based methods use short structural fragments as building unit and a set of structure templates from which fragments are extracted as an 'indirect' additional source. As a preliminary step, such methods create a set of non-redundant high resolution template structures extracted from the database of proteins of known structures – the Protein Data Bank (PDB) [12]. First, there is a search process for structure fragments by identifying those matching sequence fragments of the protein of interest. Candidate fragments are then collected to form the fragments library. Second, the fragment assembly stage concatenates relevant fragments by using randomized techniques such as simulated annealing. Third, the free energy of each of these produced conformations is assessed. Finally, the most promising ones are further refined. Due to the randomness of search trajectories in fragment-assembly methods, they usually produce a large number of candidate structures, called decoys, from which the most native-like conformation(s) has (have) to be selected ultimately. While identifying the 'best model' within a pool of decoys may be performed using clustering or quality assessment techniques, typically, the conformation with the lowest energy score is often elected. Such model is known as the *first model*.

Fragment selection is critical to the success of fragment-based methods. One of the best criteria has been to quantify the similarity between the secondary structure of a structure fragment and the predicted secondary structure of the sequence fragment that it is supposed to model [13]. Such similarity is not employed extensively in TBM approaches when choosing (a) template(s). As secondary structures play a crucial role in fragments selection, this study will go further investigating their sequence-structure complexity to exploit it to adjust the local conformational sampling used during the assembly stage.

Rosetta has been continuously contributing in CASP since its third round, being at the forefront in most of them and showing particularly promising results for hard targets [14]. Rosetta's fragment selection strategy for ab initio PSP relies on its "fragment picker" tool and its customized "quota protocol" [15]: fragments are excised from a set of high resolution template structures to form a fragment library of 25 9-mers and 200 3-mers for each position of the target protein's conformation. The score upon which candidate fragments compete to be selected is calculated using a weighted function taking into account the sequence profile, the secondary structure similarity and the Ramachandran map probability. In our experiments, the secondary structure predictors come from three sources PsiPred [16], Jufo [17] and SAM [18] and all parameters in "quota protocol" have been kept at their default values [15]. The fragment-assembly phase comprises two steps: 9-mers insertion (up to 28,000 attempts) and 3-mers insertion (up to 8,000 attempts). Note that those large numbers of insertions attempt, as well as the large number of decoys, to cover the largest possible area of the configuration space. Since all fragments were extracted from native structures, which implies that they correspond to local energy minima, they are kept rigid when inserted so that their integration within a model being built decreases the conformation's total entropy. The choice of the location of insertion, and the selection of which 9-mer, resp. 3-mer, among the 25, resp. 200, fragments available is inserted are made randomly.

III. Proposed Methodology And Preliminary Tests

Despite the significant amount of efforts that has been dedicated to improve Rosetta, - to our knowledge – it has never been proposed to vary the number of possible insertion fragments according the section of a target's sequence. Taking advantage of the relative 'straightforward' sequence-structure mapping for alpha helices, the more complex one for beta strands, and the much more obscure one for coils, a novel approach is proposed to customized the conformational sampling at local level according to the associated predicted secondary structure by amending the number of available insertion fragments per position.

Before implementation, preliminary tests were conducted to estimate what number of 9-mers and 3-mers for each of the three different secondary structures would work best. This was performed by selecting a 'representative' protein, i.e., a target which displays relative evenly distributed secondary structures and whose prediction complexity could be qualified as average. Within this study's dataset – which will be detailed later – the protein, whose PDBID is 1CC8, was chosen as, in addition to a balanced distribution of secondary structures, its best predicted model by standard Rosetta scored a Global

Distance Test (GDT) of 55 (this metric ranges from 0 to 100 where 100 means the predicted structure is undistinguishable from the native structure).

In the 1CC8's structure there are 9 positions where there are 9 consecutive amino acids that belong to a pure alpha helical structure. For each of those 9 positions, the Root Mean Square Deviations (RMSD) was calculated between each of the 25 possible substitution 9-mers and the native structure. Figure 1 displays for those positions the first, average, lowest (best) and highest (worst) obtained RMSD. Before analyzing this figure is worth remembering that (i) the lower the RMSD is, the more similar the fragments are, (ii) Rosetta's fragment picker sorts fragments according to their similarity to the sequence fragment as described in the previous section. Figure 1 shows that the first fragment is not only much better than the worst and better than the average one but is also very close to the best one. Since the first fragment appears to be a 'very good' choice, replacing the whole set of 25 9-mers by the first fragment is unlikely to lead to any significative degradation of accuracy. Similarly, we pursued this study by investigating the quality of the 21 and 17 positions in 1CC8 where ideal 3-mers are supposed to belong to a pure alpha helical and a pure beta strand respectively. In order to set an adequate minimum number of fragments at those positions, the average RMSD gain when using the best fragment out of a set of 5, 10, 15, 20, 25, 30, 35 and 40 were plotted in Figure 2.

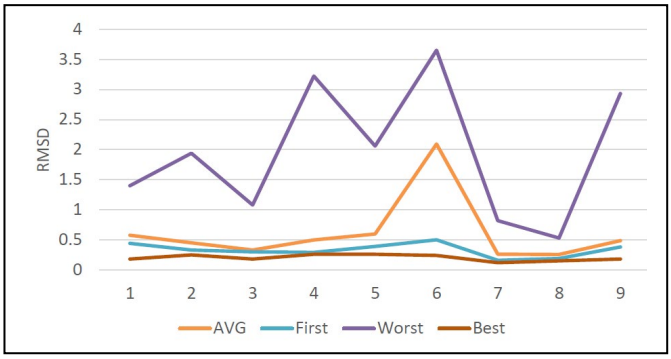

Figure 1. The averages, first, lowest, and highest RMSD of 225 9-mers

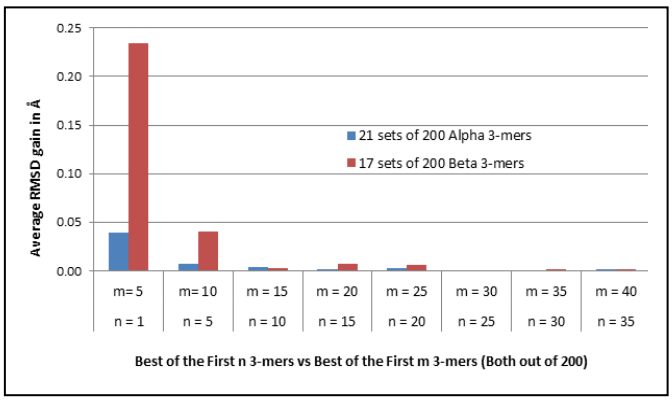

Figure 2. The average of RMSD gain when selecting the best fragment from a set of size m instead of a smaller set of size n.

Figure 2 shows that replicating the strategy selected for 9-mers, i.e., picking the first fragment only, would not be an adequate choice since tangible improvements (~11% and 27 % for alpha helix and beta sheet 3-mers respectively) were measured when the number of fragments was increased to 5. On the other hand, one notices that beyond the first 25 fragments, improvements become negligeable (~0.15%) for both types of fragments. In the case of alpha helical 3-mers, considering more than 5 fragments leads to an improvement of less than 1.5%, thus, 5 could be chosen as the number of fragments for that category. Regarding the beta 3-mers, quality still improves (+4%) with 10 fragments and keeps increasing (an additional 1.5%) until reaching 25. Thus, adopting 25 3-mers in this case seems more suitable.

In conclusion, application of the proposed methodology to both fragment files yields to a significant reduction of the overall number of candidate fragments in relatively low complexity regions. One could imagine that during the sampling process, some regions would be 'frozen' or at least much less often substituted than other parts. In summary, the proposed process is as follows: in the 9-mer file, whenever 9 consecutive amino acids are predicted to belong to a helix, only the first structural fragment is made available for insertion, otherwise, the standard number (25) can be selected. Similarly, when building the 3-mers file, whenever 3 consecutive amino acids are predicted to belong to a helix, resp. a beta strand, only 5, resp. 25, structural fragments can be inserted respectively. In all remaining positions, the standard number of available fragments (200) is used.

IV. Data Sets, Experiments, Results and Discussion

In order to validate the proposed methodology, the performance of the Secondary Structure-based Rosetta (denoted as SS-Rosetta) approach - where the number of fragments is customized according to the position - is compared to that of the standard version of Rosetta (denoted as Standard) for a set of targets. Our methodology requires choosing a secondary structure predictor and feed it with the target's sequence to obtain the required secondary structure information. Since the assessment of our proposed approach should be independent from the choice of a secondary structure predictor and its accuracy, we have used the secondary structure annotations associated to each target instead. Usage of actual secondary structure information will not affect the conclusions of this study since state-of-the-art predictors provide predictions with very high accuracy (>80%) [19]. Note that those correct annotations are used only when choosing the positions where the number of fragments should be reduced; the fragment libraries are built using the secondary structures predictors mentioned earlier.

Whenever parallel computing facilities are available, researchers tend to produce several thousands of decoys for each target to assess Rosetta's performance; 20,000 is considered an appropriate number in this regard [20], [21]. In some experiments, the production of 2,000 decoys has also been adopted since this allows them to be conducted on a typical PC [22]. Here, we have considered both numbers of

decoys, i.e., 20,000 and 2,000, for each experiment and for each target. This has allowed (i) investigating the effects of the number of decoys along with the number of fragments on the exploration/exploitation compromise of the search space and (ii) evaluating the performance of our new methodology when executed on either a cluster or a standard PC.

Since many previous studies aiming at improving fragment choice in Rosetta were evaluated by using a diverse 20-protein dataset [20], [23], [24], the same dataset has been used in this work. However, that dataset suffers from two minor issues: (i) there are only 3 targets that belong to all-alpha structural class and (ii) the largest target has a size of 128. Since fair evaluation of Rosetta's performance usually involves processing targets whose length spans up to 150 amino acids [20] and evaluation of our approach requires targets with a variety of secondary structures, 5 additional targets (2KDL, 2LR8, 4HLB, 2K4V and 2KY4) – all annotated as FM targets in previous CASP experiments – were included. Note that for each target, any homolog – if present - was removed from the fragment libraries. Finally, one of the targets (1BQ9) in the 20-protein dataset was excluded because, as it is composed mostly of coils, it could not take advantage of the proposed approach and in practice would be processed using the standard Rosetta parameters. Eventually, an updated dataset of 24 proteins whose sequence length ranges from 56 to 149 was created.

In our blind assessment, CASP's rules were followed: up to 5 models are submitted and one of them is defined as the *first model*. As in CASP, two main ranking scores were adopted: the GDTs of both the *first model* and the *Best model* among the5 submitted structures. Figures 3, resp. 4, compares the *First models* and *Best models* produced by standard Rosetta and SS-Rosetta for the 20,000-, resp. 2,000-, decoy experiment. When 20,000 decoys are considered (See Figure 3), SS-Rosetta produced better *First models* for 15 out of the 24 targets, leading to a total improvement of 6.3%, and 12 better and 3 equal *Best models* with a total improvement of 5.0%. In the 2,000-decoy experiment (see Figure 4), SS-Rosetta delivered dramatic improvements: 24.2% and 11.5% for the *First models* and *best models* respectively, where higher accuracy higher than standard Rosetta's was achieved for 16 and 18 targets respectively.

A possible explanation of the much higher improvement provided by our proposed methodology in 2,000-decoy experiment over the 20,000-decoy is as follows: When the number of decoys is small, further exploration is a more dominant process than additional exploitation as randomly generated trajectories from a large number of available fragments are likely to show high diversity. On the other hand, when the number of available fragments is much smaller, search trajectories tend to focus on exploiting limited regions in the search space (known as funnels) rather than exploring new regions using new fragments. As a result, such further exploitation yields discovering a larger number of local minima in funnels. Taking into consideration the known inaccuracy of energy functions, enhanced performance provided by our approach - which discards unhelpful fragments - suggests that the production of those larger numbers of local minima within a relatively narrower search space could benefit particularly the quality of *First Model* for the 2,000-decoy predictions.

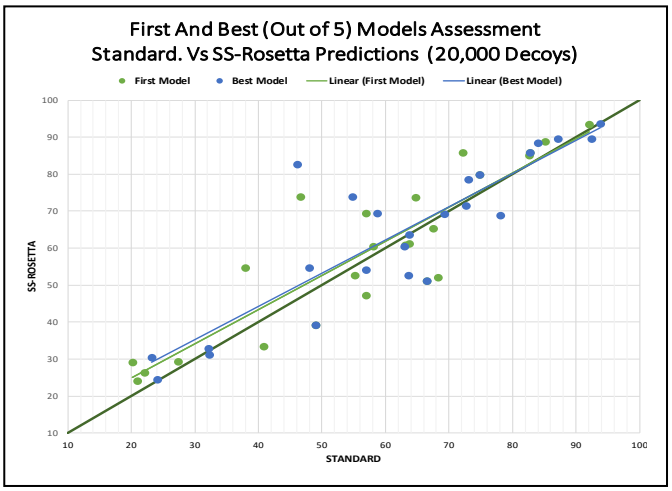

Figure 3. GDT of the First and Best (out of 5) models for standard Versus SS-Rosetta predictions using 20,000 decoys. Linear regression lines are shown for both scores

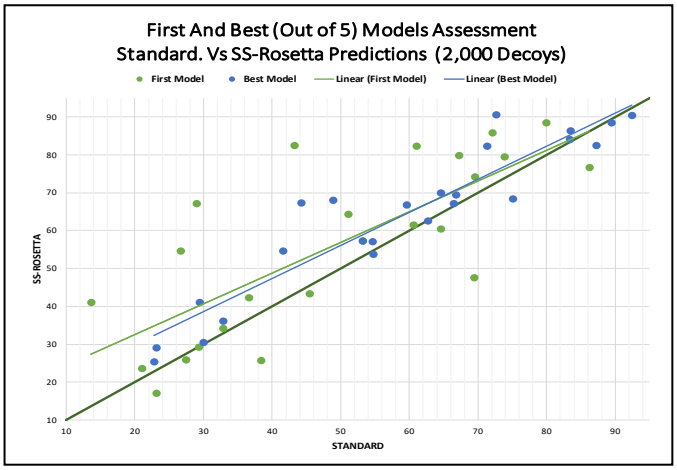

Figure 4. GDT of the First and Best (out of 5) models for standard Versus SS-Rosetta predictions using 2,000 decoys. Linear regression lines are shown for both scores

The comparison presented in Table I reveals another outcome, possibly the most important one of this study: SS-Rosetta with 2,000 decoys delivers similar accuracy to standard Rosetta with 20,000 decoys! Consequently, the novel SS-Rosetta democratizes PSP as performance, which was only possible by running standard Rosetta on high-performance computing systems, can now be achieved using a typical PC.

V. CONCLUSION

We have shown in this study that the evenly-distributed fragment insertion process across a target's whole sequence that Rosetta uses leads to the exploration of regions that are unlikely to be relevant. Instead, inspired by findings related to the proteins' sequence-structure mapping complexity, we have amended Rosetta to exploit funnels further by limiting the

number of fragments for low complexity regions, focusing more on the challenging parts of the protein target. Actually, the paradigm we have proposed in this paper is applicable to any fragment-based PSP tool regardless of the fragments size it uses. In cases where the length of the fragment size is so long that a single secondary structure is unlikely to span over it, cardinalities could be customized based on the dominant type of secondary structure instead.

Our novel methodology has revealed tangible (~6%) and significant (~24%) improvement in terms of *first model* accuracy when adopting 20,000 decoys and 2,000 decoys respectively. Furthermore, performance of the proposed approach using 2,000 decoys proved to be as accurate as that of standard Rosetta using 20,000 decoys, making PSP available to all.

TABLE I. Results of standard predictions using 20,000 decoys against SS-based predictions using 2,000 decoys.

	First model	**Best model**
20,000 Decoys (Standard Predictions) Vs 2,000 Decoys (SS-based Predictions)	12/24 + 0.4%	13/24 + 4.6%

ACKNOWLEDGMENT

The authors would like to thank Kingston University for allowing them to use the Kingston University High Performance Cluster (KUHPC) to generate more than 1 million decoys and the Lebanese International University for funding this paper. Part of the work presented in this paper was previously published [25].

REFERENCES

[1] A. W. Senior *et al.*, "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)," *Proteins*, vol. 87, no. 12, pp. 1141–1148, 2019.

[2] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold.," *Nature*, 2021.

[3] M. AlQuraishi, "Machine learning in protein structure prediction," *Curr. Opin. Chem. Biol.*, vol. 65, pp. 1–8, 2021.

[4] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 47, no. 9, pp. 1309–1314, 1961.

[5] C. B. Anfinsen, "Principles that govern the folding of protein chains.," *Science*, vol. 181, no. 96, pp. 223–230, 1973.

[6] D. Baker, "Protein folding, structure prediction and design.," *Biochem. Soc. Trans.*, vol. 42, no. 2, pp. 225–9, 2014.

[7] A. Leaver-Fay *et al.*, "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.," *Methods Enzymol.*, vol. 487, pp. 545–74, Jan. 2011.

[8] D. Xu and Y. Zhang, "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.," *Proteins*, vol. 80, no. 7, pp. 1715–35, Jul. 2012.

[9] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER Suite: protein structure and function prediction," *Yang, J., Yan, R., Roy, A., Xu, D., Poisson. J., Zhang, Y. (2015). I-TASSER Suite protein Struct. Funct. Predict. Nat Meth, 12, 7–8.Nat Meth*, vol. 12, no. 1, pp. 7–8, 2015.

[10] D. E. Kim, D. Chivian, and D. Baker, "Protein structure prediction and analysis using the Robetta server.," *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W526–W531, 2004.

[11] S. H. P. de Oliveira, J. Shi, and C. M. Deane, "Building a Better Fragment Library for De Novo Protein Structure Prediction," *PLoS One*, vol. 10, no. 4, p. e0123998, 2015.

[12] P. W. Rose *et al.*, "The RCSB protein data bank: integrative view of protein, gene and 3D structural information," *Nucleic Acids Res.*, vol. 45, no. D1, p. D271, 2017.

[13] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker, "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.," *Proteins*, vol. 34, no. 1, pp. 82–95, 1999.

[14] J. Abbass and J.-C. Nebel, "Rosetta and the Journey to Predict Proteins' Structures, 20 Years On," *Current Bioinformatics*, vol. 15, no. 6. pp. 611–629, 2020.

[15] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, and D. Baker, "Generalized fragment picking in Rosetta: design, protocols and applications.," *PLoS One*, vol. 6, no. 8, p. e23294, Jan. 2011.

[16] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–5., 2000.

[17] J. K. Leman, R. Mueller, M. Karakas, N. Woetzel, and J. Meiler, "Simultaneous prediction of protein secondary structure and transmembrane spans," *Proteins Struct. Funct. Bioinforma.*, vol. 81, no. 7, pp. 1127–1140, Jul. 2013.

[18] K. Karplus, "SAM-T08, HMM-based protein structure prediction," *Nucleic Acids Res.*, vol. 37, no. Web Server issue, pp. W492–W497, 2009.

[19] Y. Yang *et al.*, "Sixty-five years of the long march in protein secondary structure prediction: The final stretch?," *Brief. Bioinform.*, vol. 19, no. 3, pp. 482–494, Dec. 2018.

[20] D. Simoncini, F. Berenger, R. Shrestha, and K. Y. J. Zhang, "A probabilistic fragment-based protein structure prediction algorithm.," *PLoS One*, vol. 7, no. 7, p. e38799, Jan. 2012.

[21] K. W. Kaufmann, G. H. Lemmon, S. L. Deluca, J. H. Sheehan, and J. Meiler, "Practically useful: What the R osetta protein modeling suite can do for you," *Biochemistry*, vol. 49, no. 14. American Chemical Society, pp. 2987–2998, 13-Apr-2010.

[22] M. Michel, D. Menéndez Hurtado, K. Uziela, and A. Elofsson, "Large-scale structure prediction by improved contact predictions and model quality assessment," *Bioinformatics*, vol. 33, no. 14, pp. i23–i29, Jul. 2017.

[23] D. Simoncini and K. Y. J. Zhang, "Efficient Sampling in Fragment-Based Protein Structure Prediction Using an Estimation of Distribution Algorithm," *PLoS One*, vol. 8, no. 7, pp. 1–10, 2013.

[24] D. Simoncini, T. Schiex, and K. Y. J. Zhang, "Balancing

exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction," *Proteins Struct. Funct. Bioinforma.*, vol. 85, no. 5, pp. 852–858, 2017.

[25] J. Abbass and J.-C. Nebel, "Enhancing fragment-based protein structure prediction by customising fragment cardinality according to local secondary structure," *BMC Bioinformatics*, vol. 21, no. 1, p. 170, 2020.