This is the accepted version of this paper. The version of record is available at https://doi.org/10.1109/MECBME47393.2020.9265175

Enhanced Rosetta-based Protein Structure Prediction For non-Beta Sheet Dominated Targets

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— **Protein structure prediction has been one of the most challenging tasks undertaken in bioinformatics. Fragment assembly methodologies have emerged as the most accurate approaches to predict protein conformations without the need of homologues. Rosetta – a fragment-based tool – has consistently been at the forefront for two decades. Rosetta assembles candidate conformations using fragments of length 9 and 3 extracted from a pool of high-resolution proteins. Herein, an extensive study has been conducted highlighting the importance of the size 3 fragments – 3-mers - the role of which is both refining and correcting. Reduction of the number of those fragments from 200 to 100 revealed that Rosetta was able to produce *first models* of improved accuracy (+8%) for alpha and alpha-beta targets by 8%. Accordingly, an amended pipeline was proposed: it involves adjusting the number of 3-mers according to sequence-based structural class prediction of the protein target.**

Keywords— fragment assembly protein structure prediction; Rosetta; 3-mers; protein structural class; CATH;

I. INTRODUCTION

Among all the biochemical reactions taking place in the living cell, many relies on the action of proteins. For instance, it has been estimated that absence of some enzymes – the largest set of proteins – can make one biological reaction lasts up to 1.1 billion years [1]; unsurprisingly, they are referred to as "the eyes, arms and legs of living cells" [2]. Following the translation process, the ribosome generates a protein as a linear sequence of amino acids adopting the shape of an extended random coil. Then after a set of biophysical interactions, the very quick folding process occurs spontaneously where the linear chain adopts a compact and rigid structure. This final conformation, also known as the native structure, is believed to be unique amongst all possible conformations as it corresponds to the shape with the lowest free energy.

From a drug design perspective, knowledge of the exact coordinates of the tertiary native structure of proteins is considered invaluable [3]. To this end, scientists and bioengineers have designed *in vitro* environments where they are able to observe proteins folding and/or reaching their final thermodynamically stable conformations, gaining insight about the natural folding mechanism that takes place inside the cell, i.e. *in vivo*. The main laboratory techniques are X-ray crystallography, Nuclear Magnetic Resonance (NMR), and Electron Microscopy (EM). Eventually, the 3D conformations that they managed to resolve are deposited in the Protein Data Bank (PDB) – the world's largest repository of proteins with known structures [4]. Despite enormous advancements in those techniques, there is still enormous gap between the numbers of proteins with known structures and known proteins: the current ratio is estimated at about 0.01%[*]. This is due to, not only the high cost of those laboratory techniques in terms of time and money, but also technical limitations when dealing, in particular, with membrane and large proteins.

To address this, for more than two decades, bioinformaticians have undertaken the challenge to design computer software that predict a protein's correct native structure from its amino acid sequence, i.e. *in silico* prediction. Although production of such tool has remained an ambitious goal, the research field of Protein Structure Prediction (PSP) has made significant progress. In order to access those advancements, a biannual competition called Critical Assessment of techniques for protein Structure Prediction (CASP) was established in 1994. In its latest version, CASP13 – 2018, the average structure accuracy when dealing with the most challenging targets reached nearly 66%. Although, thanks to mainly the exploitation of deep learning techniques, this is a dramatic improvement on performance delivered 2 years earlier at CASP12 (+13%) [5], further improvements and original ideas are still needed. Indeed, none of the available PSP tools has so far been able to deliver 'acceptable' conformations, i.e. structures whose knowledge support drug design, for all targets.

Rosetta is considered to be one of the most accurate PSP tools when dealing with challenging targets. Conceptually, it is based on two steps: the assembly of fragments of 9 amino acids to predict a target's general shape and its refinement using fragments of size 3. Herein, we examine the role of size 3 fragments exploring if their capability goes further than ameliorating an already correct fold. Then, we propose a strategy relying on the protein target's structural class to ensure that Rosetta improves models during its second step.

[*] http://www.ebi.ac.uk/uniprot/TrEMBLstats/

II. Computational Techniques for Protein Strucutre Prediction

A. Overview

Despite the variety of computational approaches developed for protein structure prediction, they can be classified within two categories: Template-Based Modelling (TBM) and template-Free Modelling (FM). As its name implies, the TBM category relies on known protein structure(s), i.e. template(s), available in the PDB, to build a putative structure from the sequence of a protein of interest. Generally, methods belonging to this category rely on two steps. First, appropriate template(s) need to be identified, and then they are exploited to build the protein's conformation. Those methods can be further divided into two major classes: homology modelling and fold recognition. Whilst the first one relies on sequence-based homology for template selection, the second one requires sequence-structure compatibility. In homology modelling, also known as comparative modelling, sequence similarity exceeding 30% suggests that the target and the template share a similar 3D structure [6]. However, when sequence similarity is lower, the more demanding fold recognition approach, also called threading, may be a better alternative. Sequence-structure alignment is conducted to find templates the conformation of which are compatible with the amino acid sequence of the target. In practice, some carefully crafted fitness function is used to assess if any of the templates' conformations is likely to be adopted by the target's sequence.

The alternative to TBM, i.e. template free modelling – also known as ab initio –, is to build conformation 'from scratch' relying on Anfinsen's theories [7], [8]. Those theories state that, not only is the amino acid sequence alone sufficient to dictate the corresponding structure as the folding process is purely physical, but also the native structure is the one associated to the lowest free energy. Consequently, ab initio approaches employ energy functions known as force fields (FF) to evaluate the forces amongst atoms and amino acids so that the conformation with the lowest energy score can be discovered. Although ab initio methods are considered the most challenging PSP approaches, they are able, in principle, to predict structures of targets for which no template can be identified. Those techniques, also known as physics-based since they use laws of physics to simulate the folding mechanism, have enormous computational needs, which limits their usage to state-of-the-art supercomputers and large grid computing systems. Such processing requirements led to the development of distributed computing initiatives such as, Folding@home and Rosetta@home[9], where the general public is asked to participate by giving access to their personal computers, which, when idle, perform protein folding tasks.

B. Fragment-based methods

Limitations due to the computational complexity of 'pure ab initio' methods have prompted the need of a more practical approaches, yet still able to predict Free Modelling targets. The concept behind a new type of approach, i.e. fragment-based assembly, was inspired from two main observations. First, the sequence-structure correlation is stronger when short sequences are considered, although its strength varies depending on the secondary structure. Second, any protein conformation can successfully be replicated by simply assembling short substructures from other proteins the global shape, i.e. fold, of which may be totally different. Consequently, FM targets could be built by assembling short substructures from diverse protein templates without homology requirements regarding the selection of those templates. Although fragment-based PSP approaches only rely loosely on physics, they have delivered remarkable performance in recent CASP events. As a result, Rosetta [10], I-TASSER [11], and QUARK [12], all fragment-based frameworks, are now the leading approaches when dealing with FM targets. Such success has been driven by two main features of those approaches: i) instead of employing resource hungry molecular dynamics simulations to simulate the actual folding path, they concentrate on finding the resulting conformation using typically Monte Carlo simulations; and ii) instead of using individual atoms or amino acids as basic building blocks, sets of amino acids, or fragments, are considered as rigid units of construction, making much faster the optimization process towards a near-native conformation. A consequence of those strategies is that fragment-based PSP tools do not require high performance facilities and can often be executed on standard personal computers.

A characteristic of fragment-assembly methods is that, due to the random nature of search trajectories, they usually produce a large number of putative structures, called decoys, from which the most native-like conformation has to be identified. While selecting the *best model* within a pool - sometimes thousands of them - of decoys may be performed using clustering or quality assessment techniques, often the model with the lowest energy score is elected, it is then known as the *first model*.

C. Rosetta

Among fragment-based methods, Rosetta - described for the first time twenty years ago contributing to CASP3 [13] - has consistently been at the forefront of template free modelling being updated regularly. Nowadays, it is an open source package that comprises many macromolecular modelling tools and applications [10]. The execution of the fragment assembly phase – Rosetta's core – relies on the target's sequence, its profile, and typically 25 fragments of size 9 and 200 of size 3 for each amino acid position. They are extracted from a fragment library built in advance using the 'fragment picker tool' [14] and a dataset of non-homologous high-resolution protein templates (~16k). Besides sequence similarity, secondary structure predictions are used to select suitable fragments of both size 9 – also known as 9-mers – and size 3 – 3-mers.

The fragment-assembly task is divided into two subsequent phases: 9-mers and then 3-mers insertions. As each fragment, in principle, corresponds to a local energy minimum, it is kept rigid so that its integration within the model being build decrease the conformation's entropy. Not only is the choice of the location of insertion chosen randomly, but also the choice of a 9-mer among the 25 available and a 3-mer out of 200 is also random. Once the fragment insertion process finishes, a coarse-grained conformation is generated where side chains are represented by centroid atoms. All bond angles, bond lengths, and side chains atoms are added afterwards based on statistical data that suggest their most likely values [15].

III. PRELIMINARY EXPERIMENTS

As mentioned earlier, Rosetta operates using two stages of fragment insertions/substitutions: 28,000 insertion attempts of 9-mers, followed by 8,000 3-mer insertion attempts. The first stage is considered as the crucial one since it leads to the overall shape of the conformation; it is divided into several sub-stages where terms of the force field are subsequently added to tighten the acceptance criterion of a fragment replacement. In comparison with a 3-mer substitution, not only may a 9-mer one dramatically changes the substructure where the replacement takes place, but will also affect the general shape of the model being built. While usage of such large 'jumps' provides a good strategy to escape local minima, they are unlikely to allow reaching a near-native conformation. This is achieved by the 3-mer phase, the purpose of which is to refine the conformation built in the initial stage by performing 8,000 additional insertions, i.e. small 'jumps' are attempted to get deeper in the funnels of the search area rather than skipping them.

Since the number of available fragments of size 3 for each position is 200, they may offer quite a lot of structural diversity. A consequence could be that those insertions may go well beyond structure refinement leading to relatively dramatic structural corrections. Whilst such corrections can be advantageous to the putative models that did not converge towards the correct fold following the 9-mer insertions phase, they may be destructive to those that only lack some refinement.

To highlight the modelling abilities of the 3-mer insertion phase, structure predictions were conducted using standard Rosetta and disabling Rosetta's 9-mer insertion phase (3-mer only Rosetta): in each scenario, 20,000 decoys were generated for each of the 33 protein targets defined in Section IV. Figure 1 shows their result comparison in terms of the *first model* using the GDT-TS metric. As expected, standard predictions clearly outperform 3-mer only Rosetta (quality average: -21.8%), yet, it was able to attain a better *first model* in 9 out of the 33 targets. This confirms that correction and refinement are not the sole capabilities of the 200 3-mer fragments insertion phase, since it can also produce competitive conformation for some targets. Figure 2 shows one of those successful 3-mer only predictions (74.7 GDT) using PyMol for their visualisation [16].

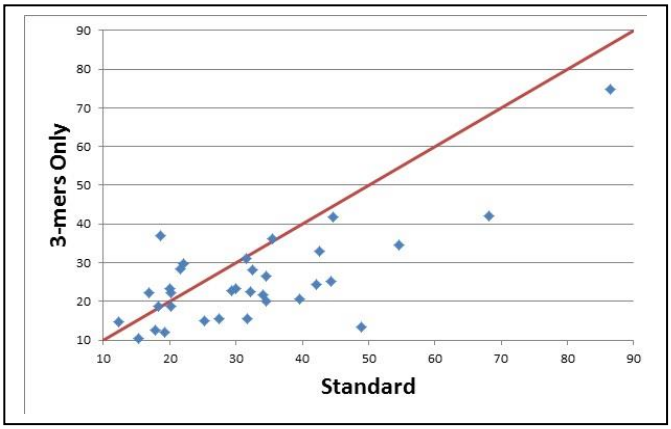

Figure 1. GDT of the First Model out of 20,000 decoys. The score used is the GDT-TS, where a high value means a high similarity.

A previous study demonstrated that standard Rosetta's performance varies according to the structural class of the protein target [17]. It relies on the three main classes defined by CATH – a database that classifies proteins based on structure and functions [18]: mainly alpha, mainly beta and alpha beta. Table I reports the results of that study showing that the higher the amount of beta sheets in a target, the lower is the accuracy of the first model generated by Rosetta. As sequence-based structural class prediction is highly accurate, many tools having reported an accuracy exceeding 90% [19], this technology is mature enough to exploit such prediction in a sequence-based PSP pipeline.

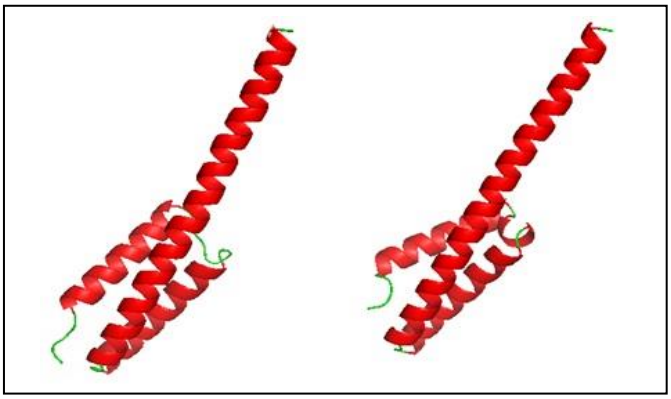

Figure 2. Structures of the native model (left) and first model's conformation (right) using 3-mer only Rosetta (PDB ID: 4FM3)

TABLE I. Performance of Rosetta according to a target's structural class

	Number of targets	**Average of the first model's GDT for Rosetta**
Mainly Alpha	16	39.5
Mainly Beta	18	23.4
Alpha Beta	33	27.4

Based on the above, one may suggest that the role of 3-mers in Rosetta may be optimised according to a protein's structural class. Herein, using empirical studies, this hypothesis is investigated by amending the number of 3-mer fragments based on the predicted structural complexity of the protein target. In the next section, two sets of experiments are presented. According to the target's structural class, they reveal when the standard 3-mer insertion phase is either useful or detrimental.

IV. DATA SETS, EXPERIMENTS AND RESULTS

The evaluation dataset contains 33 targets selected from previous CASP competitions. It is diverse in terms of sequence lengths and protein structural classes: lengths range from 33 to 141 amino acids and proteins belong to the three main structural classes, i.e. mainly alpha, mainly beta and alpha beta. Since the raison d'être of ab initio protein structure prediction is the ability to infer new folds, all proteins homologous to those of the evaluation dataset were removed from the data set from which fragments are built (based on E-value lower than 0.05 on PSIBLAST). The evaluation metric used to evaluate the accuracy of predicted models against the native structures is the

Global Distance Test – Total Score (GDT-TS) (GDT in the text) – that is the formal assessment criterion used in CASP. Its value ranges between 0 and 100, where 0 means no structural similarity and 100 represents a perfect superimposition [20].

Conforming to the blind assessment of PSP methodologies that involve a large number of decoys, the *first model*, i.e. the structure with the lowest energy amongst the set of decoys, is used to compare Rosetta's standard predictions with those of the amended pipeline.

For each protein target, three experiments were performed, each using a different number of available 3-mers: standard predictions, i.e., 200 3-mers, 100 3-mer predictions (denoted as '100 3-mers'), and 25 3-mer predictions (denoted as '25 3-mers'). The 100 and 25 3-mers were chosen using the top scores that Rosetta's fragment-picker associated to each fragment. For each of the three experiments, 20,000 decoys were generated. When available, CATH annotations were used to associate the targets to structural classes, otherwise CATH's standard thresholds of 15% helix and 10% strand [21] were adopted to manually annotate targets.

As Figure 3 and Table II show, when the number of 3-mer fragments was reduced to 100, improvement in terms of overall performance – in terms of *first model* - was negligible, i.e. +0.4%. However, when only alpha and alpha-beta targets are considered, a different picture emerges: for 14 out of those 23 targets, a better *first model* was produced with an average improvement of 8.7%. Conversely, the average of GDT degradation of the 10 mainly beta targets reaches 18.0%. The results of this experiment support our hypothesis: not only do alpha and alpha-beta structures not need a large amount of correction, but also it may also affect negatively the conformation produced during the 9-mer insertion phase.

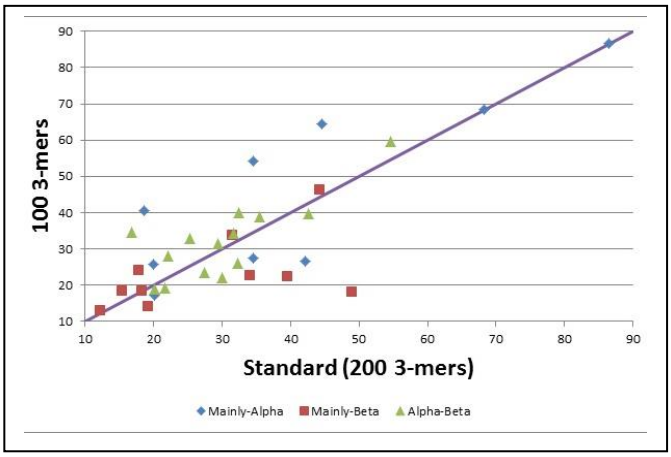

Figure 3. GDT of the first models of standard predictions versus predictions made using 100 3-mers

This conclusion is confirmed by the second set of experiments where the number of 3-mers was further decreased to 25. However, as only minor improvements are achieved for alpha and alpha beta targets (see Table II), this suggests that the low diversity of the 25 fragments does not allow exploring sufficiently the conformation space around the model produced during the first phase. Those results are in line with those of a thorough study by Zhang and Xu - I-TASSER's pioneers - where they concluded that no less than 100 fragments is needed for each position in the amino acid sequence to attain a native-like conformation [22]. Figure 4 illustrates the improved quality of the *first model* obtained for an alpha target when using 100 3-mers instead of standard Rosetta.

TABLE II. Detailed comparison between the first model's quality of 100 3-mers and 25 3-mers with that of the standard predictions

Average GDT change of First Models compared to standard approach					
	All three classes	**Mainly alpha**	**Mainly beta**	**Alpha beta**	**Mainly alpha and alpha beta classes only**
100 3-mers	+0.4%	+11.4%	-18.0%	+6.4%	+8.7%
25 3-mers	-4.8%	+3.3%	-27.8%	+1.5%	+2.4%

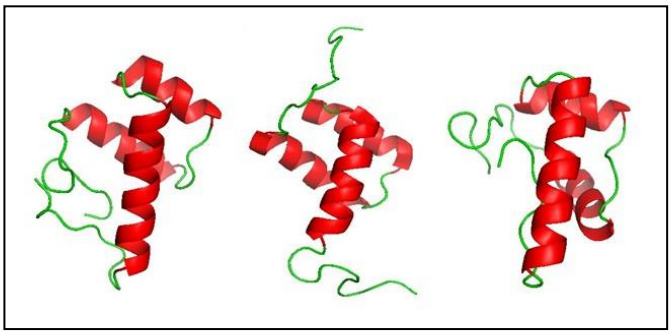

Figure 4. From left to right. Structures of first model produced using 100 3-mers (GDT = 64.5), native structure and first model produced using standard Rosetta (GDT = 44.5) for a 74 amino acid protein (PDB ID: 2LY9)

V. CONCLUSION

This investigation clarifies the role that 3-mer insertions have while conducting protein structure prediction using Rosetta. It unveils that usage of 3-mers is not limited to minor corrections, but that fragments of size 3 may be sufficient to generate a conformation of reasonable accuracy. Moreover, experiments reveal that the standard number of 3-mers, i.e. 200, is not optimal when dealing with targets composed of more than 15% helices, i.e. alpha and alpha beta targets, as excessive fragment diversity tends to degrade the quality of the generated conformation. Accordingly, a new pipeline for Rosetta-based protein structure prediction is proposed - see Figure 5. This study proposes that usage of 100 3mers offers probably a good compromise in terms of fragment variety for non-beta dominated targets.

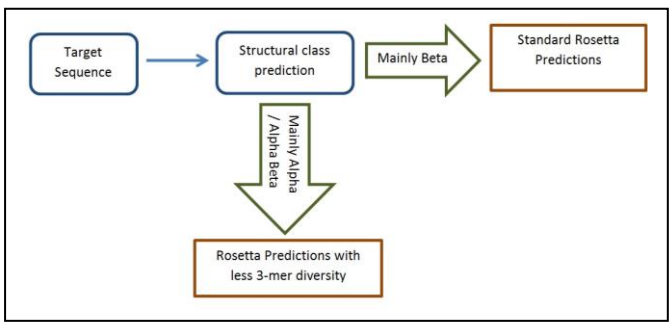

Figure 5. Proposed Rosetta-based protein structure prediction pipeline.

ACKNOWLEDGMENT

The authors would like to thank Kingston University for allowing them to use up to 128 processors of the Kingston University High Performance Cluster (KUHPC) for more than two months. More than 2 million decoys were generated to perform the experiments presented in this study. Part of the work presented in this paper was previously published [18].

REFERENCES

[1] R. Wolfenden and M. J. Snider, "The depth of chemical time and the power of enzymes as catalysts," *Acc. Chem. Res.*, vol. 34, no. 12, pp. 938–945, 2001.

[2] A. K. Dunker and R. W. Kriwacki, "The orderly chaos of proteins.," *Sci. Am.*, vol. 304, no. 4, pp. 68–73, 2011.

[3] J. Abbass, J.-C. Nebel, and N. Mansour, "Ab Initio Protein Structure Prediction: Methods and challenges," in *Biological Knowledge Discovery Handbook*, M. Elloumi and A. Y. Zomaya, Eds. Hoboken, New Jersey: John Wiley & Sons, Inc., 2013, pp. 703–724.

[4] H. M. Berman *et al.*, "The Protein Data Bank.," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[5] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (CASP)-Round XIII," *Proteins Struct. Funct. Bioinforma.*, Oct. 2019.

[6] S. D. Lam, S. Das, I. Sillitoe, and C. A. Orengo, "An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences," *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 73, no. 8, pp. 628–640, 2017.

[7] C. J. Epstein, R. F. Goldberger, and C. B. Anfinsen, "The Genetic Control of Tertiary Protein Structure: Studies With Model Systems," *Cold Spring Harb Symp Quant Biol*, vol. 28, no. 0, pp. 439–449, 1963.

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

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

[10] 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.

[11] S. Wu, J. Skolnick, and Y. Zhang, "Ab initio modeling of small proteins by iterative TASSER simulations.," *BMC Biol.*, vol. 5, p. 17, Jan. 2007.

[12] C. H. Tai, H. Bai, T. J. Taylor, and B. Lee, "Assessment of template-free modeling in CASP10 and ROLL," *Proteins Struct. Funct. Bioinforma.*, vol. 82, no. SUPPL.2, pp. 57–83, 2014.

[13] K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, "Ab Initio Protein Structure Prediction of CASP III Targets Using ROSETTA," *Proteins Struct. Funct. Genet.*, vol. 37, no. May, pp. 171–176, 1999.

[14] 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.

[15] P. Barth, J. Schonbrun, and D. Baker, "Toward high-resolution prediction and design of transmembrane helical protein structures," *Proc. Natl. Acad. Sci.*, vol. 104, no. 40, pp. 15682–15687, Oct. 2007.

[16] Schrödinger, LLC, "The {PyMOL} Molecular Graphics System, Version~1.8," Nov. 2015.

[17] J. Abbass and J.-C. Nebel, "Customised fragments libraries for protein structure prediction based on structural class annotations.," *BMC Bioinformatics*, vol. 16, no. 1, p. 136, Apr. 2015.

[18] J. Abbass and J.-C. Nebel, "Reduced Fragment Diversity for Alpha and Alpha-Beta Protein Structure Prediction using Rosetta," *Protein Pept. Lett.*, vol. 24, no. 3, pp. 215–222, 2017.

[19] Z. X. Liu, S. lei Liu, H. Q. Yang, and L. H. Bao, "Using protein granularity to extract the protein sequence features," *J. Theor. Biol.*, vol. 331, pp. 48–53, 2013.

[20] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3370–3374, Jul. 2003.

[21] A. D. Michie, C. A. Orengo, and J. M. Thornton, "Analysis of domain structural class using an automated class assignment protocol.," *J. Mol. Biol.*, vol. 262, no. 2, pp. 168–185, 1996.

[22] D. Xu and Y. Zhang, "Toward optimal fragment generations for ab initio protein structure assembly.," *Proteins*, vol. 81, no. 2, pp. 229–39, Feb. 2013.