Reduced Fragment Diversity for Alpha and Alpha-Beta Protein Structure Prediction using Rosetta

Jad Abbass[1]* and Jean-Christophe Nebel[1]

[1]Faculty of Science, Engineering and Computing, Kingston University, London, KT1 2EE, UK

**Correspondence to: Jad Abbass, Faculty of Science, Engineering and Computing, Kingston University, London, KT1 2EE, UK. T: +44 (0) 208 417 2740. F: +44 (0)208 417 2972. E-mail: k1064285@kingston.ac.uk.*

Abstract: Protein structure prediction is considered a main challenge in computational biology. The biannual international competition, Critical Assessment of protein Structure Prediction (CASP), has shown in its eleventh experiment that free modelling target predictions are still beyond reliable accuracy, therefore, much effort should be made to improve ab initio methods. Arguably, Rosetta is considered as the most competitive method when it comes to targets with no homologues. Relying on fragments of length 9 and 3 from known structures, Rosetta creates putative structures by assembling candidate fragments. Generally, the structure with the lowest energy score, also known as first model, is chosen to be the "predicted one".

A thorough study has been conducted on the role and diversity of 3-mers involved in Rosetta's model "refinement" phase. Usage of the standard number of 3-mers – i.e. 200 – has been shown to degrade alpha and alpha-beta protein conformations initially achieved by assembling 9-mers. Therefore, a new prediction pipeline is proposed for Rosetta where the "refinement" phase is customised according to a target's structural class prediction. Over 8% improvement in terms of first model structure accuracy is reported for alpha and alpha-beta classes when decreasing the number of 3-mers.

Keywords: Rosetta; ab initio protein structure prediction; fragment-based protein structure prediction; CATH; protein structural class; 9-mers; 3-mers.

1. INTRODUCTION

From a drug design perspective, determination of a protein's native structure represents a crucial step since it allows gaining important insights of molecular mechanisms involved in many diseases [1]. Despite the advancements achieved in both wet laboratories and computational techniques, protein structure determination still faces many challenges. Bioinformatics is usually considered as "the last chance" when neither X-ray crystallography nor nuclear magnetic resonance (NMR) can be used due to time, cost or/and experimental constraints. Whereas performing protein folding simulation conforming to Newton's second law may appear as an attractive approach, it is only practical when applied on very small targets while using state-of-the-art supercomputers and grid computing [2] [3]. Consequently, many current computational methods rely on Monte Carlo simulations and heuristic search techniques besides reduction of amino acids and energy functions' representations [4] [5].

In order to assess fairly accuracy and advancements in the field of computational techniques for protein structure prediction (PSP), the Critical Assessment of protein Structure Prediction (CASP) was launched in 1994 [6]. Every two years, a set of protein sequences are released gradually across a couple of months during which research groups from around the world attempt to predict their 3D structures by submitting putative models (up to 5 per target). Once a target's submission deadline has passed, determination of its native structure is conducted in vitro. If successful, a thorough evaluation is performed on the submitted models. Targets released by CASP have usually been classified into two categories: Free modelling (FM) and template-based modelling (TBM). Whereas the TBM category comprises "easy targets" for which structures of homologous proteins have already been deposited in the

protein data bank (PDB) [7], FM targets represent the greatest challenge in the competition since only groups that rely on ab initio methods can contribute. Due to the complexity of the task, any minor improvement regarding accuracy of FM targets amongst competing groups is considered worthwhile. Note that in the ongoing CASP12 experiments, FM and TBM terms have been replaced by "high accuracy models" and "topology" respectively. Whilst homology modelling methods perform well on template-based targets, they cannot deal with FM ones. On the other hand, although ab initio approaches that mimic Anfinsen's principle (also known as physics-based modelling) [8][9][10] by looking for the conformation of lowest possible energy score produce predictions of increasing accuracy, their expensive running time has limited usage to around 50 amino acid length [11]. As an alternative, "coarse grained" ab initio protein structure predictors, also called fragment-based, have been developed. Methods such as FRAGFOLD [12], Rosetta [13], I-TASSER [14], and QUARK [15] have demonstrated the strength of such approach. Regardless of the fragments' length used by those methods, their popularity is supported by three main points: (1) since the "smallest element" considered in computation is a set of amino acids instead of a single one, entropy in conformational search space is decreased in a dramatic way, (2) short sub-sequences converge towards a relatively limited number of sub-structures and (3) usage of Monte Carlo simulations instead of Molecular Dynamics ones has allowed making those methods much faster than pure ab initio ones. Nevertheless, fragment-based methods continue to fail reaching reasonable accuracy for many CASP's targets, which has motivated further improvements, amendments and tuning [16] [17] [18] [19] [20] [21] [22] [23].

Performance at various CASP events has led to the conclusion that, currently, Rosetta may be considered as the most competitive method in the FM target category [24] [25] [26] [27]: its capability to produce new folds has been clearly demonstrated [28] [29] [30]. This may be explained by the fact that, unlike FRAGFOLD, I-TASSER and QUARK which use relatively long fragments, Rosetta conforms to the original study of fragment assembly PSP conducted by Bowie and Eisenberg by taking into consideration short fragments [31]. Adding to its popularity, Rosetta offers a complete and open-source software suite for PSP which allows integrating refinements and parameter optimisation.

Rosetta, developed at the University of Washington at Seattle, relies on an energy function that combines physics-based and knowledge-based terms [32]. Selection of fragments depends on several criteria that constitute a weighted-function score. Besides sequence similarity, sequence profile, and Ramachandran map propensities, secondary structure predictions taken from three resources represent a crucial factor [33]. In order to generate a conformation's backbone along with its side chain centroids, Rosetta operates in two main steps: first, 9-mer fragments are inserted within the initial fully extended conformation; second, insertions of 3-mer fragments are used to refine the structure previously generated. 9-mers and 3-mers are protein fragments extracted for each amino acid - except for the protein C-terminus - of the protein of interest from a template database according to some similarity criteria. Note that Rosetta keeps fragments rigid during simulation [34]. Eventually, Rosetta converts the coarse-grained conformation into an all-atom representation by adding all missing atoms using knowledge-based information extracted from known structures [35].

This work, first, presents a thorough study on the effects of the selection of 3-mers on the quality of conformations generated by Rosetta and, second, introduces a new pipeline for Rosetta protein structure prediction dedicated to alpha and alpha beta proteins. The proposed approach relies on a limitation of 3-mer fragment diversity so that conformations generated during the 9-mer phase can be refined in more depth than with the standard Rosetta settings. Following a description of the evaluation framework, this paper justifies both theoretically and experimentally the principles of the proposed method. Then, a variety of experiments are conducted to validate them and results are discussed in light of other relevant studies.

2. MATERIALS AND METHODS

2.1 Data sets and evaluation framework.

The evaluation dataset covers the three main protein structural classes, i.e. mainly alpha, mainly beta and alpha beta [36]. It comprises 33 targets the length of which ranges from 33 to 141 amino acids. They were selected from models of the Free Modelling and Template-Based Modelling categories assessed during CASP8, 9, 10, 11 and CASP ROLL. In order not to take advantage of any homology, all homologues, defined by an E-value lower than 0.05 on PSI-BLAST [37], were removed from Rosetta's fragment libraries. This typical threshold is used to evaluate Rosetta's ability to infer new folds, which is its raison d'être [35]. Since targets are taken from CASP, the formal evaluation metric used in that worldwide and community-wide competition is adopted, i.e. Global Distance Test – Total Score (GDT-TS) (GDT in the text) [38].

The assessment framework is based on the first model, i.e. the structure with the lowest energy amongst 20,000 decoys, sample which is considered to be high enough so that meaningful conclusions could be drawn [35]. Not only does such an evaluation comply with CASP's formal ranking of competitive groups in the Tertiary Structure category, but it is also directly relevant to life scientists. Moreover, correlation between best model and first model is calculated to evaluate aspects of the energy-structure correlation across different experiments.

2.2 Principles

Rosetta's first phase with its 28,000 9-mer insertions is considered the essential part of the process since it builds the general shape or fold of the structure guided by secondary structure predictions. Those insertions are divided into several sub-phases where terms of energy functions are increasingly added to tighten the acceptance criterion of a fragment replacement. 9-mer insertions can be seen as relatively coarse scale as each insertion may change dramatically the structure being built. Although this allows escaping local minima, this coarse modelling phase is unlikely to reach a near-native conformation.

As a consequence, Rosetta includes a 3-mer phase to improve that initial conformation by performing 8,000 additional insertions. Although the 3-mer insertion phase is generally seen as structure refinement, the fact that Rosetta uses by default 200 fragments means that they can be quite diverse and insertions may in occasion lead to dramatic structural corrections. Whereas those corrections are certainly beneficial to conformations which failed to adopt the correct fold during the first phase, they may be detrimental to those which only needed some fine tuning. In this work, it is proposed to investigate and exploit that hypothesis by adapting the number of 3-mer fragments according to the perceived structural complexity of the protein target.

First, 'correction' abilities of the 200 3-mer fragments are demonstrated by generating 20,000 decoys using Rosetta without the initial 9-mer insertion phase. Although, as expected, performance is generally below that of the standard 2-phase Rosetta (-21.8% in terms of GDT of the model with the lowest energy, or first model), 3-mer only Rosetta was still able to generate a better first model in 9 out of the 33 tested targets, see Figure 1. This experiment clearly demonstrates that usage of 200 3-mer fragments goes well beyond refinement, but has some abilities of conformation generation. Figure 2 illustrates this where 3-mer only insertions are able to generate a good quality first model (74.7 GDT) for a target of length 94. Structures are visualised using PyMol [39].

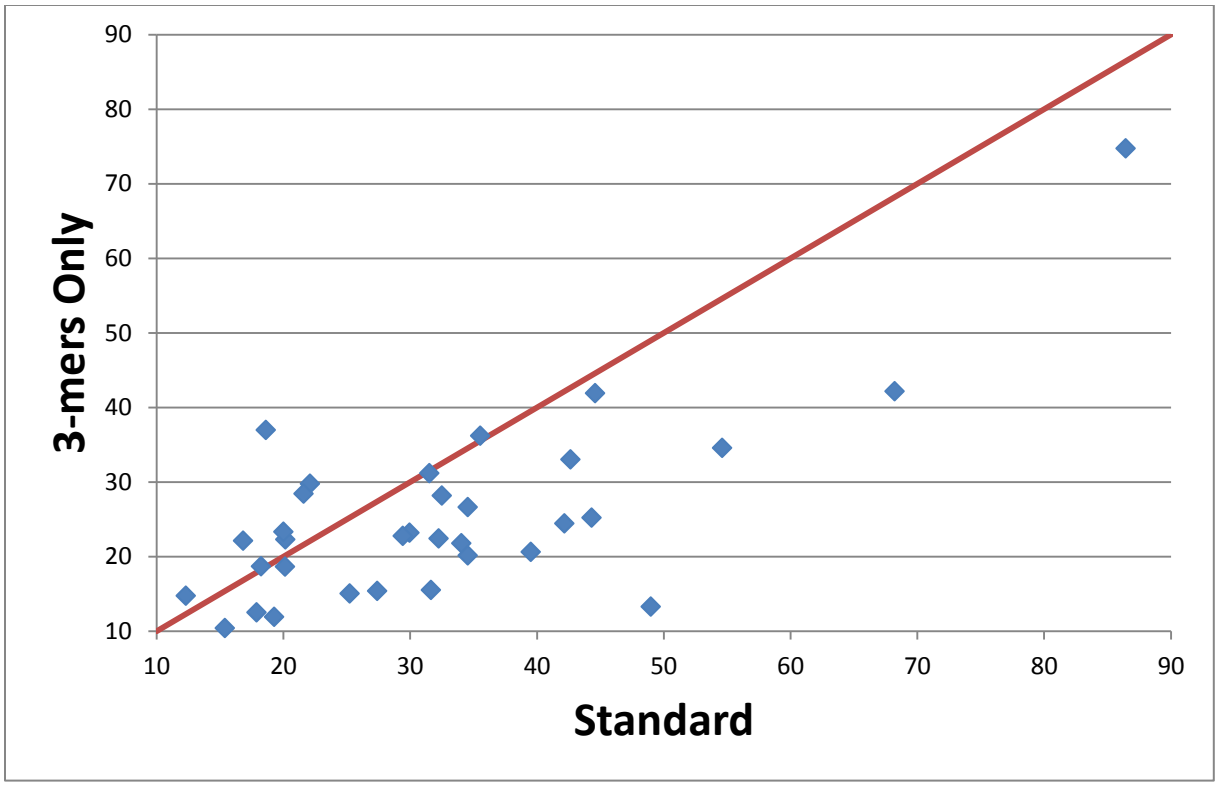

Figure 1: First Model's GDT out of 20,000 decoys

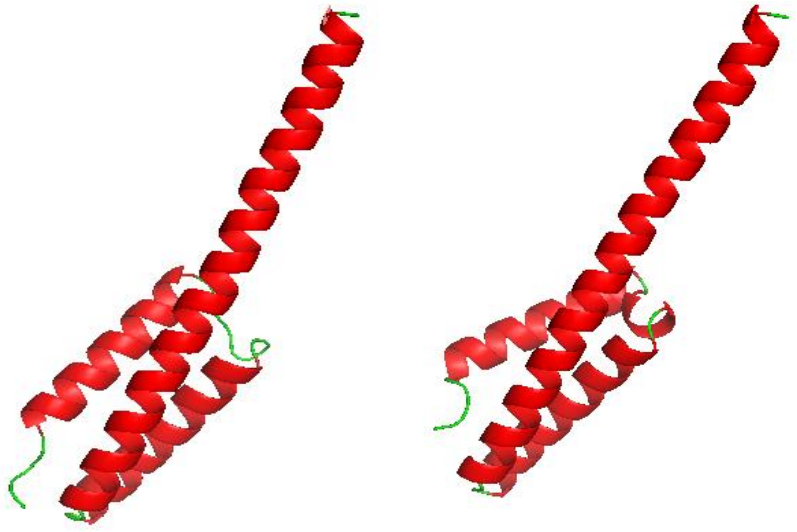

Figure 2: Structures of the native model (PDB ID: 4FM3) and first model's conformation using 3-mer only Rosetta

Second, a previous study has shown that performance of standard Rosetta depends on the structural class of a protein target [23]. Such classes may be annotated using CATH (Class, Architecture, Topology and Homology) [36], a hierarchical database for classifying and annotating proteins (mainly domains) in terms of their structures and functions. The top level, or structural class, is primarily based on a protein's secondary structure content. The three

main classes in CATH are mainly alpha, mainly beta and alpha beta. As experimental results generated for proteins belonging to the three main CATH classes show, see Table 1, alpha and alpha-beta protein conformations are better predicted than mainly beta proteins.

Table 1: Results of a thorough study on 67 targets [23]

	Number of targets	Average of the first model's GDT for standard Rosetta	Average of the first model's GDT for CATH-based Rosetta [Ab15]	Improvements
Mainly Alpha	16	39.5	46.5	17.7%
Mainly Beta	18	23.4	25.7	9.6%
Alpha Beta	33	27.4	31.5	15.0%

Moreover, class prediction using the sequence alone is highly accurate as some methods were able show accuracy exceeding 90% [40]. Table 2 shows the confusion matrix of the results of comparing the correct annotation versus the predicted ones using MODAS (MODular Approach to Structural class prediction); a free web-based piece of software that can generate results in seconds [41]. On this specific dataset, a prediction accuracy of 80.6% is achieved. Furthermore, CATH-based Rosetta [23] has shown that exploiting that information leads to better model predictions: GDT increased by over 10%, especially for proteins classified in alpha and alpha-beta classes, see Table 1.

Table 2: Confusion matrix showing CATH classes versus MODAS predicted ones [23]

Predicted	A	AB	B	Other
A (16)	**15**	1	0	0
AB (33)	2	**25**	3	3
B (18)	0	4	**14**	0

In view of those experimental results, it is proposed to adapt Rosetta's 3-mer phase according to a target's structural class. Since structure prediction of proteins belonging to alpha and alpha-beta structural classes is more accurate, one may infer that the 'correction' behaviour of the 3-mer phase is less needed whereas additional refinement could lead to the generation of better models. Here, it is demonstrated this behaviour can be achieved by reducing 3-mer diversity. Figure 3 show a new processing pipeline describing optimisation of Rosetta's 3-mer phase according to a target's predicted structural class.

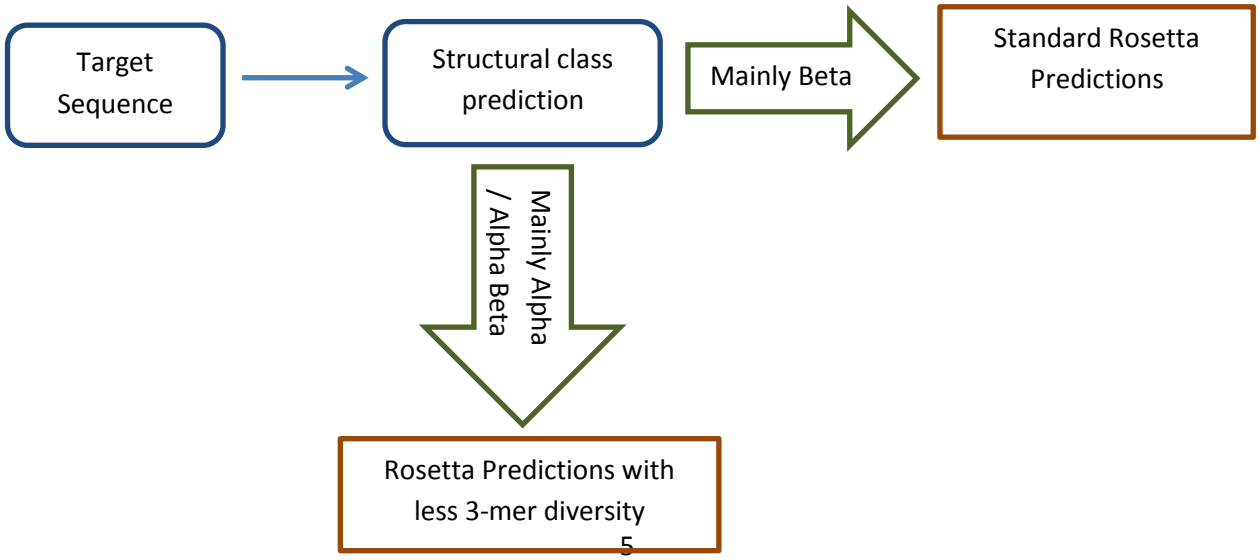

Figure 3: New pipeline for optimisation of Rosetta for mainly-alpha and alpha-beta protein structure predictions

2.3 Experimental setup

For each target, three experiments were conducted using 200 (denoted later on as "Standard"), 100 and 25 3-mers, where reduction of 3-mers is performed by excluding the most diverse fragments. For each of them, 20,000 decoys were generated. CATH annotations, when available, were used to allocate a structural class to each target. However, whenever a protein did not have any formal CATH annotation in the PDB, CATH's standard thresholds of 15% helix and 10% strand [42] were used to infer structural classification.

3. RESULTS

As Figure 4 shows, 14 out of the 23 targets belonging to the alpha and alpha-beta structural classes achieved a higher first model's GDT when the number of 3-mer fragments was reduced to 100. Moreover, GDT scores were higher on average by 8.7%. However, when all 33 targets are considered, this reduction of the number of 3-mers does not affect performance significantly, +0.4%, since the GDT scores of mainly beta targets fell in average by 18.0%. Those results confirm that removal of 'correction' fragments improves predictions of alpha and alpha-beta structures, while degrades the generation of beta structures, see Table 3. Additional experiment, where the number of 3-mers was further decreased to 25, reveals that such a dramatic reduction of 3-mer diversity leads to poorer performance when all targets are considered. That outcome is in line with the results of the comprehensive study on fragments conducted by Zhang and Xu, I-TASSER's pioneers, where they showed that at least a set of 100 fragments is needed for each amino acid position to achieve a native-like conformation [43]. However, when only alpha and alpha-beta's targets are considered, usage of 25 3-mers delivers still slightly better performance than the standard approach (note that Rosetta uses an additional 25 9-mers in the first phase of the prediction process). Table 3 displays a summary of this first model study.

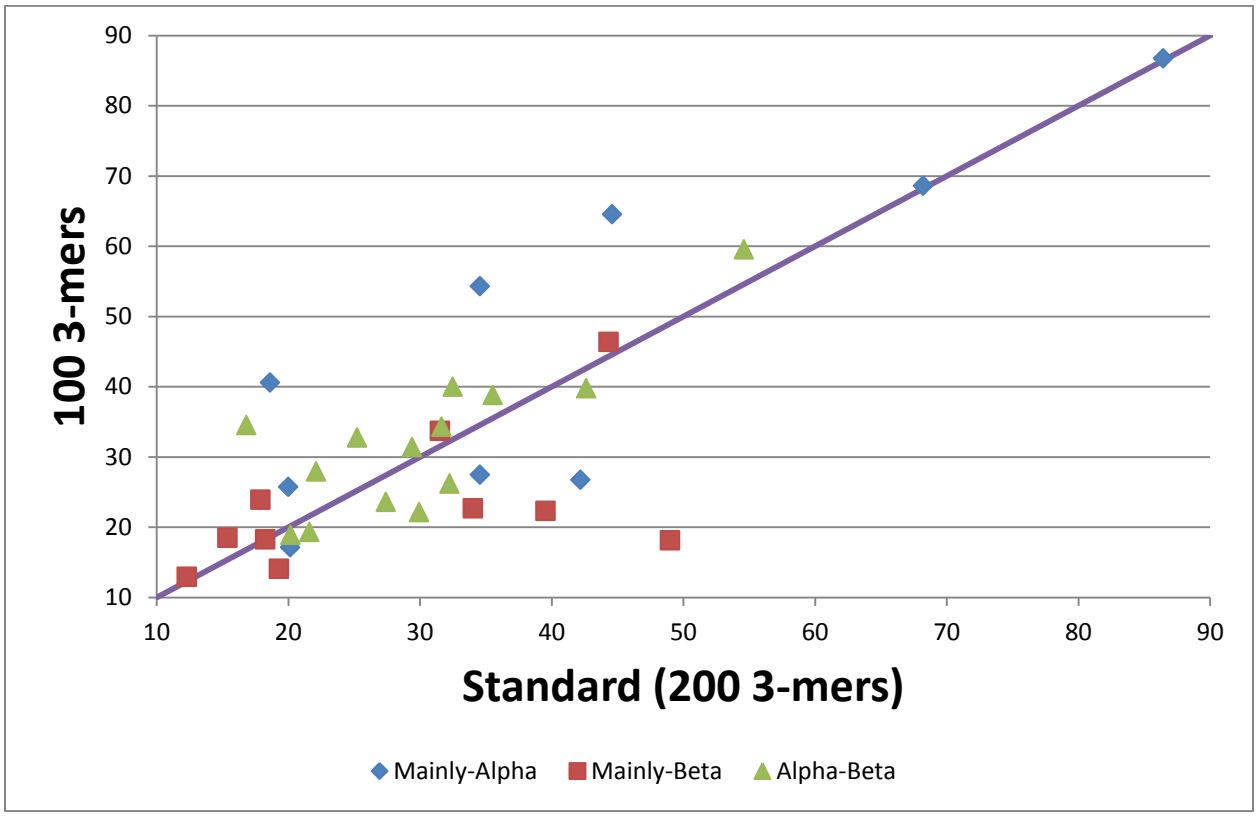

Figure 4: GDT of standard predictions versus predictions using 100 3-mer fragments only

Table 3: Comparison of first model's quality according to 3-mer reduction strategy with standard approach

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%

A pictorial evidence showing the accuracy of 100 3-mers for alpha targets over standard's is displayed in Figure 5: except for the N and C-terminus coil regions the conformations of which are predicted incorrectly, the structure of the first model generated using 100 3-mers is very close to the native one in terms of fold and alpha helix topology; on the other hand the standard first model is much less accurate due to the incorrect orientation of the third alpha helix.

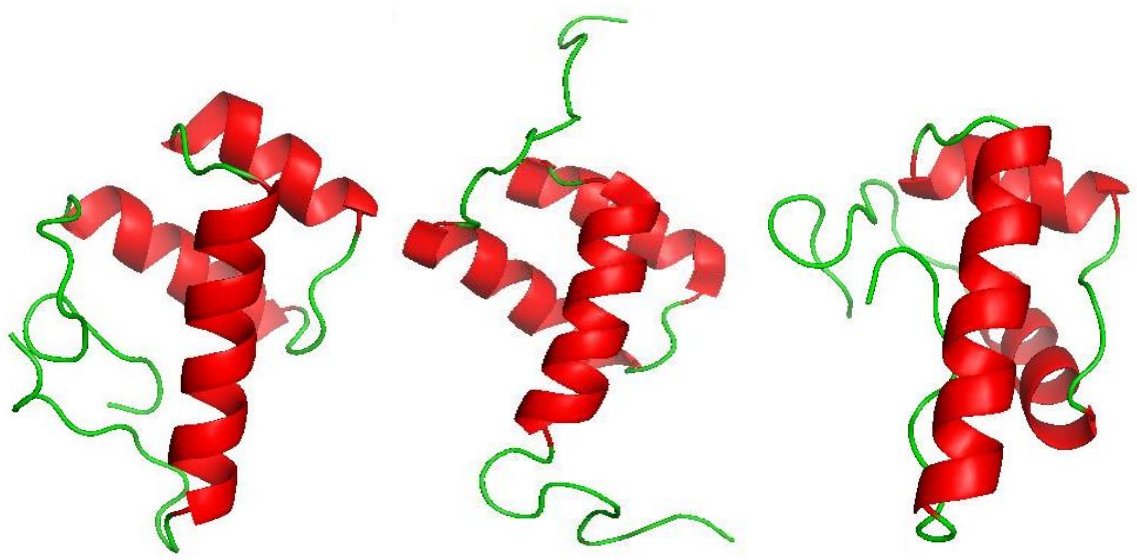

Figure 5: Structures of 100 3-mer approach's first model (GDT = 64.5), native (PDB ID: 2LY9) and standard approach's first model (GDT = 44.5) of that 74 amino acid protein, respectively.

The evaluation of the structure-energy correlation amongst the three experiments is performed by calculating the percentage of the best model's GDT achieved by the first model's. As shown in Table 5, for the mainly alpha and alpha beta classes, the 100 3-mer approach delivers first models which are significantly closer to the best model, +8.0%, than the standard approach.

Table 5: Comparison of structure-energy correlation in terms of GDT.

Average of percentage of the best model's GDT achieved by the first model's		
	All three classes	Mainly alpha and alpha beta classes only
Standard	62.2%	60.7%
100 3-mers	62.3%	65.7%
25 3-mers	59.0%	61.9%

4. DISCUSSIONS

Although Rosetta generates all-atom models, it relies on moderate coarse-grained protein modelling, where each amino acid is represented using C-alpha, C-Beta and side chain's centroid. As a consequence, the energy landscape that Rosetta explores is expected to be quite smooth as illustrated in Figure 6 [44].

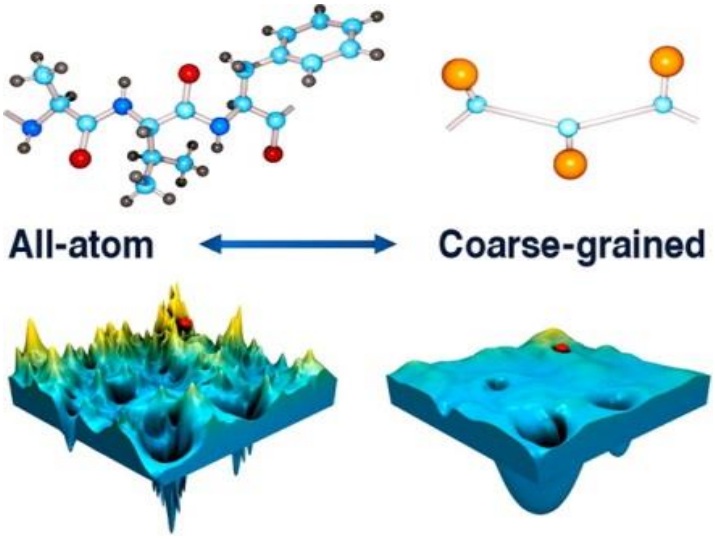

Figure 6: Energy landscape of all-atom versus coarse-grained protein modelling. Taken from: Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A.E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. Chemical Reviews, 2016, 116, 7898–7936, http://pubs.acs.org/doi/full/10.1021/acs.chemrev.6b00163 (with permission)

When the 9-mer insertions phase is performed, the energy landscape is explored through relatively "big jumps" corresponding to 9-mer substitutions. Consequently, local minimum of a given funnel (position A in Figure 7-a) may not be reached leading to a locally suboptimal conformation (position B in Figure 7-a). Then, during the 3-mer insertion phase, the dual role of correction and refinement is played. Whereas refinement allows moving a conformation deeper in the current funnel, correction permits investigating neighbouring funnels. On one hand, the more diverse the 3-mer library (e.g. 200), the larger and the farther the set of explored funnels is likely to be. On the other hand, less diversity (e.g. 100) is likely to reduce the size of the searched space allowing deeper exploration of the initially selected funnel (Figure 7-b).

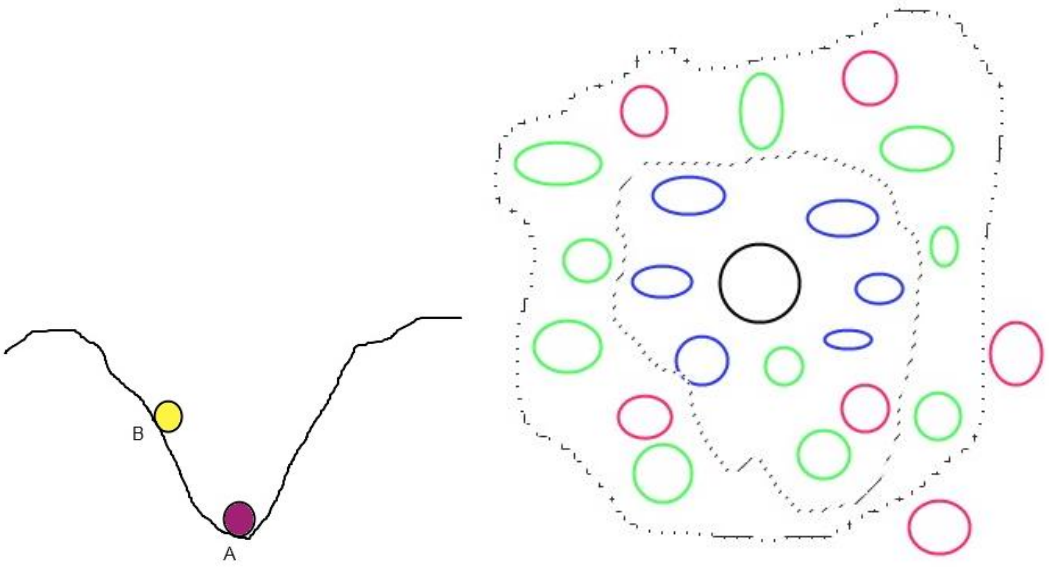

Figure 7: (a) Positions A and B illustrate the energy levels of the conformations resulting from the 3-mer and 9-mer insertion phases respectively. (b) The black circle represents the funnel which contains the conformation produced by the 9-mer insertion phase. Blue ellipses represent funnels that contain structures with good accuracy, whereas green and purple ones have worse accuracy. The inner, respectively outer, dashed contour denotes the limit of the search space created by less, respectively more, diverse 3-mer insertions.

This suggests that, when dealing with the easier targets, i.e. from alpha and alpha-beta classes, the 9-mer insertion phase tends to succeed in identifying a funnel close to the native area. As a consequence, usage of 3-mers with relatively low diversity, e.g. 100 fragments, is beneficial allowing exploration of that zone more in depth and eventually producing a more optimal conformation. Alternatively, for the harder targets, i.e. from the beta class, where the 9-mer insertion phase is less likely to have generated a conformation close to the native one, keeping a larger search space by using quite diverse fragments, e.g. 200, increases the probability of converging towards an acceptable conformation.

Usage of a reduced set of 3-mers for the easier targets is further supported by studies which demonstrated that native and native-like structures are likely to be found in the largest cluster/ broader funnel of decoys [45] [46], see figure 8. Those observations have resulted in development of quality assessment prediction techniques, known as decoy clustering [47], to identify the “best model” produced by ab initio methods that, like Rosetta, generate a large number of candidate structures known as decoys. Their strategy relies on, first, clustering those decoys according to some threshold similarity threshold, typically 3 to 4 Å, and, second, selecting the conformation with the lowest energy score from the largest cluster.

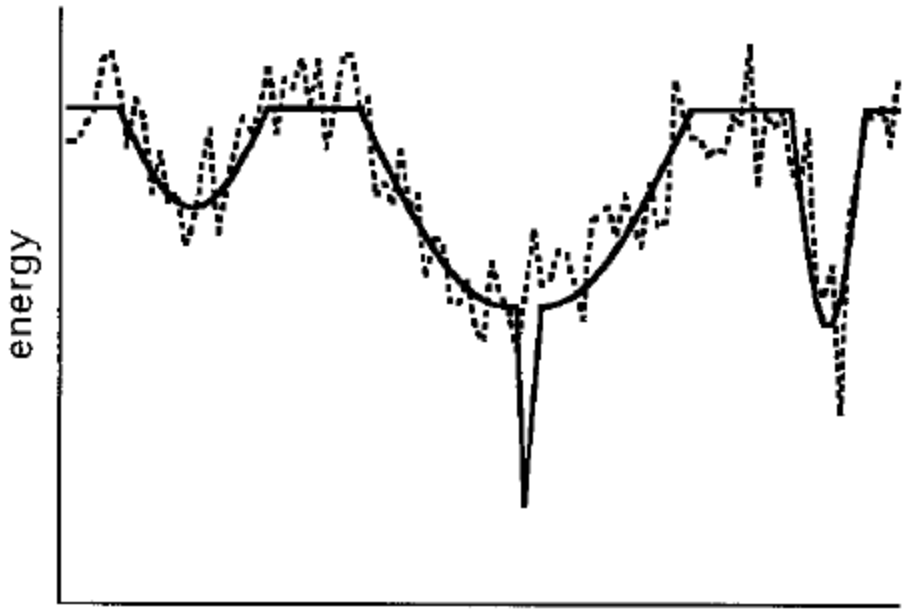

Figure 8: Hypothetical folding energy landscape. The solid and dashed lines represent the "real" energy and force field scores respectively (y axis) according to a generalised structure coordinate. This shows clearly that the broader funnel is the one that comprises the "best candidates" as they neighbour the native one. Taken from: Shortle, D.; Simons, K.T.; Baker, D. Clustering of low-energy conformations near the native structures of small proteins. Proc. Natl. Acad. Sci. U. S. A., 1998, 95, 11158–11162. Copyright (1998) National Academy of Sciences, U.S.A. (with permission).

CONCLUSION

This paper has presented a comprehensive study on the importance, role and effects of the fragments of size 3 in Rosetta protein structure prediction for the three main structural classes. Usage of the standard number of 3-mers for each position – i.e. 200 – has been shown to degrade alpha and alpha-beta protein conformations initially achieved by assembling 9-mers. Owing to the high accuracy of structural class prediction from sequence, a new Rosetta's pipeline dedicated to alpha and alpha beta proteins has been proposed where 3-mer diversity is reduced. Experimental results has confirmed that a smaller number, namely 100, of less diverse 3-mers is more appropriate when predicting alpha and alpha-beta targets since it allows Rosetta focusing on the refinement of the initially generated conformations. In addition to produce better quality "first models", those models delivered by the proposed pipeline prove to be significantly closer to the actual "best model", which is directly relevant to life scientists.

ACKNOWLEDGMENTS

The authors would like to thank the Faculty of Science, Engineering and Computing for grating us 128 processors on the Kingston University High Performance Cluster (KUHPC) to perform all predictions involved in this study.

REFERENCES

[1] Ramirez-Alvarado, M.; Kelly, J.W.; Dobson, C.M. Protein Misfolding Diseases: Current and Emerging Principles and Therapies; John Wiley and Sons, **2010**.

[2] Baker, D. Protein folding, structure prediction and design. *Biochem. Soc. Trans.*, **2014**, *42*, 225–229.

[3] Huang, P.-S.; Boyken, S.E.; Baker, D. The coming of age of de novo protein design. *Nature*, **2016**, *537*, 320–327.

[4] Abbass, J.; Nebel, J.-C.; Mansour, N. In Biological Knowledge Discovery Handbook; Elloumi, M.; Zomaya, A. Y., Eds.; John Wiley & Sons, Inc.: Hoboken, New Jersey, **2013**; pp. 703–724.

[5] Dill, K. a; MacCallum, J.L. The protein-folding problem, 50 years on. *Science*, **2012**, *338*, 1042–1046.

[6] Moult, J.; Pedersen, J.T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Bioinforma.*, **1995**, *23*, ii--iv.

[7] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.*, **2000**, *28*, 235–242.

[8] Anfinsen, C.B.; Haber, E.; Sela, M.; White, F.H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **1961**, *47*, 1309–1314.

[9] Epstein, C.J.; Goldberger, R.F.; Anfinsen, C.B. The Genetic Control of Tertiary Protein Structure: Studies With Model Systems. *Cold Spring Harb Symp Quant Biol*, **1963**, *28*, 439–449.

[10] Sun, S. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.*, **1993**, *2*, 762–785.

[11] Shaw, D.E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R.O.; Eastwood, M.P.; Bank, J.A.; Jumper, J.M.; Salmon, J.K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science*, **2010**, *330*, 341–346.

[12] Kosciolek, T.; Jones, D.T. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One*, **2014**, *9*, e92197.

[13] Leaver-Fay, A.; Tyka, M.; Lewis, S.M.; Lange, O.F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P.D.; Smith, C.A.; Sheffler, W.; Davis, I.W.; Cooper, S.; Treuille, A.; Mandell, D.J.; Richter, F.; Ban, Y.-E.A.; Fleishman, S.J.; Corn, J.E.; Kim, D.E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J.J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J.J.; Kuhlman, B.; Baker, D.; Bradley, P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.*, **2011**, *487*, 545–574.

[14] Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: protein structure and function prediction. *Nat Meth*, **2015**, *12*, 7–8.

[15] Xu, D.; Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, **2012**, *80*, 1715–1735.

[16] Wang, T.; Yang, Y.; Zhou, Y.; Gong, H. LRFragLib: an effective algorithm to identify fragments for de novo protein structure prediction. *Bioinformatics*, **2016**, btw668.

[17] Guyon, F.; Tufféry, P. Assessing 3D scores for protein structure fragment mining. *Open Access Bioinformatics*, **2010**, *2*, 67–77.

[18] Olson, B.; Molloy, K.; Hendi, S.F.; Shehu, A. Guiding probabilistic search of the protein conformational space with structural profiles. *J. Bioinform. Comput. Biol.*, **2012**, *10*, 1242005.

[19] Simoncini, D.; Berenger, F.; Shrestha, R.; Zhang, K.Y.J. A probabilistic fragment-based protein structure prediction algorithm. *PLoS One*, **2012**, *7*, e38799.

[20] Baeten, L.; Reumers, J.; Tur, V.; Stricher, F.; Lenaerts, T.; Serrano, L.; Rousseau, F.; Schymkowitz, J.

Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLoS Comput. Biol.*, **2008**, *4*, e1000083.

[21] Li, S.C.; Bu, D.; Xu, J.; Li, M. Fragment-HMM: a new approach to protein structure prediction. *Protein Sci.*, **2008**, *17*, 1925–1934.

[22] Uziela, K.; Wallner, B. ProQ2: Estimation of model accuracy implemented in Rosetta. *Bioinformatics*, **2016**, *32*, 1411–1413.

[23] Abbass, J.; Nebel, J.-C. Customised fragments libraries for protein structure prediction based on structural class annotations. *BMC Bioinformatics*, **2015**, *16*, 136.

[24] Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B.-H.; Das, R.; Grishin, N. V; Baker, D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, **2009**, *77 Suppl 9*, 89–99.

[25] Tai, C.H.; Bai, H.; Taylor, T.J.; Lee, B. Assessment of template-free modeling in CASP10 and ROLL. *Proteins Struct. Funct. Bioinforma.*, **2014**, *82*, 57–83.

[26] Vincent, J.J.; Tai, C.-H.; Sathyanarayana, B.K.; Lee, B. Assessment of CASP6 predictions for new and nearly new fold targets. *Proteins*, **2005**, *61 Suppl 7*, 67–83.

[27] Jauch, R.; Yeo, H.C.; Kolatkar, P.R.; Clarke, N.D. Assessment of CASP7 structure predictions for template free targets. *Proteins Struct. Funct. Bioinforma.*, **2007**, *69*, 57–67.

[28] Ovchinnikov, S.; Park, H.; Kim, D.E.; Liu, Y.; Wang, R.Y.R.; Baker, D. Structure prediction using sparse simulated NOE restraints with Rosetta in CASP11. *Proteins: Structure, Function and Bioinformatics*, **2016**.

[29] Ovchinnikov, S.; Kim, D.E.; Wang, R.Y.R.; Liu, Y.; Dimaio, F.; Baker, D. Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins: Structure, Function and Bioinformatics*, **2016**.

[30] Kinch, L.; Yong Shi, S.; Cong, Q.; Cheng, H.; Liao, Y.; Grishin, N. V. CASP9 assessment of free modeling target predictions. *Proteins Struct. Funct. Bioinforma.*, **2011**, *79*, 59–73.

[31] Bowie, J.U.; Eisenberg, D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. U. S. A.*, **1994**, *91*, 4436–4440.

[32] Rohl, C.A.; Strauss, C.E.M.; Misura, K.M.S.; Baker, D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology*, **2004**, *383*, 66–93.

[33] Gront, D.; Kulp, D.W.; Vernon, R.M.; Strauss, C.E.M.; Baker, D. Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One*, **2011**, *6*, e23294.

[34] Simons, K.T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **1997**, *268*, 209–225.

[35] Bradley, P.; Misura, K.M.S.; Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science (80-.).*, **2005**, *309*, 1868–1871.

[36] Sillitoe, I.; Lewis, T.E.; Cuff, A.; Das, S.; Ashford, P.; Dawson, N.L.; Furnham, N.; Laskowski, R.A.; Lee, D.; Lees, J.G.; Lehtinen, S.; Studer, R.A.; Thornton, J.; Orengo, C.A. CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.*, **2015**, *43*, D376–D381.

[37] Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*,

1997, *25*, 3389–3402.

[38] Zemla, a. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **2003**, *31*, 3370–3374.

[39] Schrödinger, LLC. The {PyMOL} Molecular Graphics System, Version~1.8; **2015**.

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

[41] Mizianty, M.J.; Kurgan, L. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC Bioinformatics*, **2009**, *10*, 414.

[42] Michie, A.D.; Orengo, C.A.; Thornton, J.M. Analysis of domain structural class using an automated class assignment protocol. *J. Mol. Biol.*, **1996**, *262*, 168–185.

[43] Xu, D.; Zhang, Y. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins*, **2013**, *81*, 229–239.

[44] Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A.E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, **2016**, *116*, 7898–7936.

[45] Shortle, D.; Simons, K.T.; Baker, D. Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **1998**, *95*, 11158–11162.

[46] Betancourt, M.R.; Skolnick, J. Finding the Needle in a Haystack: Educing Native Folds from Ambiguous Ab Initio Protein Structure Predictions. *J. Comput. Chem.*, **2001**, *22*, 339–353.

[47] Perez, A.; Yang, Z.; Bahar, I.; Dill, K.A.; Maccallum, J.L.; Olechnovič, K.; Kulberkytė, E.; Venclovas, C.; Mrozek, D.; Li, S.C.; Ng, Y.K.; Jamroz, M.; Kolinski, A.; Kihara, D.; Guyon, F.; Tufféry, P.; Giorgetti, A.; Raimondo, D.; Miele, A.E.; Tramontano, A.; Galiez, C.; Coste, F.; Cristobal, S.; Zemla, a; Fischer, D.; Rychlewski, L.; Elofsson, a. Calibur: a tool for clustering large numbers of protein decoys. *Bioinformatics*, **2014**, *30*, 256.