Using Machine Learning to Explore the Relevance of Local and Global Features During Conformational Search in Rosetta

Mario Garza-Fabre
Faculty of Life Sciences
University of Manchester
mario.garza-fabre@manchester.ac.uk

Shaun M. Kandathil
Faculty of Life Sciences
University of Manchester

Julia Handl
Decision and Cognitive Sciences Research Centre
University of Manchester

Joshua Knowles
School of Computer Sciences
University of Manchester

Simon C. Lovell
Faculty of Life Sciences
University of Manchester

ABSTRACT

Our ongoing work focuses on improvements to the exploration behaviour of heuristic search techniques in fragment-assembly methods for protein structure prediction. Analysing and improving exploration in fragment-assembly can be difficult due to the complexity of measuring diversity between decoys in a meaningful manner. Here, we define a set of local and global features of decoy structures generated by Rosetta, and we use Machine Learning to explore the extent to which these are predictive of the final prediction results achieved by individual runs. The aim is to identify those feature subsets that show a significant correlation with final prediction outcome, and identify when they become fixed during the search. It is thought that such features may help in the formulation of new diversity measures that can be utilized in the context of explicit diversity mechanisms such as crowding, external archives etc. The time of fixture can help in deciding at what stage of the search the implementation of diversity mechanisms may be the most relevant.

Keywords

Protein structure prediction, Fragment assembly, Features, Machine Learning

1. INTRODUCTION

Fragment-assembly techniques employ heuristic search approaches to explore the space of possible conformations of a protein. Analysis of the search trajectories of current methods shows that the balance between exploration and exploitation breaks down with an increase in the length and the complexity of protein structures. Attempts at incorporating more advanced search mechanisms (such as estimation of distribution algorithms or landscape learning [12, 2]) into the search have typically been disappointing. One of the most successful approaches to fragment-assembly continues to be the Rosetta ab initio [9], which uses large numbers of independent restarts of a local search heuristic.

Rosetta ab initio consists of two protocols, a low-resolution part and a full-atom component. The low-resolution protocol is responsible for conformational search and is the component that employs the fragment-assembly paradigm. In other words, this phase assembles candidate ("decoy") structures through the iterative insertion of small structural segments that are obtained from other known protein structures in the Protein Database (PDB). Sampling is guided by the Metropolis Monte Carlo algorithm using a default temperature of T=2 (with a possibility to re-heat when insertion attempts become unsuccessful). More specifically, this low-resolution protocol consists of four stages that differ in the size of the fragments and the energy functions used. The first three stages use a fragment size of 9 residues ("9mer"), while the last stage uses just 3 residues ("3mers"). The energy function consists of a sum of ten separate energy terms, which are progressively switched on and / or weighted more heavily over the course of the protocol (the first stage uses a van der Waals term only, while all energy terms are active in the last stage of the search). The decoys generated at this low-resolution stage are passed forward to the full-atom protocol, which serves the primary purpose of (locally) refining structures and identifying the most promising predictions.

The standard approach to running Rosetta is to use thousands or tens of thousands of restarts, and this has been observed to outperform the use of fewer but longer runs. Unfortunately, this brute-force breaks down as protein size (and complexity) increases, as evidenced by the deterioration of prediction performance [8] as well as the analysis of sampling trajectories, which indicates a rapid convergence of each trajectory and an inability to access relevant parts of the search space [4]. For this reason, our current work focuses on the development of sampling approaches that can improve the exploratory behaviour of the search.

One of the difficulties in this lies in measuring meaningful diversity in the conformational search space. Measures of diversity are important both for assessing sampling behaviour and for implementing mechanisms that encourage

GECCO '15, July 11 - 15, 2015, Madrid, Spain

DOI: http://dx.doi.org/10.1145/2739482.2768441

exploration [10, 7]. Furthermore, as decoy structures undergo a progressive folding progress (as detailed above), we need to understand at which stage of the search exploratory behaviour is of primary importance. Recent work [11, 5] suggests that certain features of a structure become "fixed" early on during the search, and can then present decisive factors for the success of a run. More accurate insight into this behaviour may help in the design of search protocols that strategically encourage diversity at specific stages of the search. Here, we aim to foster our understanding of these aspects by employing a machine learning approach.

2. FEATURES

We implemented a set of different features to describe various aspects of decoy structures that have or may be considered in the development of advanced search protocols. These measures can be broadly classified either as *global features*, accounting for general characteristics of a structure as a whole, or as *residue-level features*, accounting for specific properties of amino acid residues in the folded conformation.

2.1 Global features

- **Energies**
  - Total energy score. The energy of a given conformation is evaluated using a knowledge-based energy function, which consists of a linear weighted sum of ten energy terms. Different terms and weights are used at different stages of the search process. For the purposes of this study, the energy function as used in stage 4 was adopted.
  - The 10 individual energy terms of the energy function. These terms involve descriptions of steric repulsion and compactness, as well as a statistical potential and interactions between specific elements of secondary structure.
- **Contacts**
  - Total number of contacts in the folded conformation. In this study, two amino acid residues are said to be in contact if the distance between them is at most 8 Å.
  - Number of local contacts and non-local contacts. A given contact can be classified as local or non-local depending on whether the sequence distance between the interacting amino acids is within a given cutoff. A cutoff of 5 Åwas used in this study.
- **HP model-based features** The Hydrophobic-Polar (HP) model is a simplified version of the protein structure prediction problem [3]. This model is based on the fact that the hydrophobicity of amino acids is an important force determining the folded state of a protein chain. Based on this model, some features describing a folded conformation can be investigated:
  - Number of H-H, H-P, and P-P contacts. Contacts (as defined above) can be further classified based on the hydrophobic properties of the interacting amino acid residues.
  - H-exposure. The radius of gyration (RG), which accounts for the compactness of a protein conformation, is computed separately for H and P residues. If H residues are more exposed than P ones (higher RG), the difference in RG values is reported, 0 otherwise. A similar measure has also been investigated in [6].
- **Distance-based features**
  - Distance between the first and last amino acid residues in the protein chain, commonly referred to as the N (left) and C (right) terminus in the specialized literature.
  - Distance between each pair of secondary structure chunks. The central amino acid in the corresponding sequence segment is selected as the representative reference element for each secondary structure chunk. By using these representative elements, it is possible to capture the relative position between each pair of secondary structure chunks using a reduced set of features.

2.2 Residue-level features

- **Torsion angles**. The ψ, ϕ and ω torsion angles describing the backbone configuration for each amino acid residue.
- **ABEGO classification**. The configuration of a torsion angle triplet (as described above) can be classified either as "A", "B", "E", "G", or "O", based on its corresponding region in the Ramachandran plot.
- **Fragment identifiers**. These identifiers encode information related to which specific fragment is a given residue taking its torsion angle values from. Numeric, consecutive identifiers have been assigned to all available fragments in the fragment libraries.

  The use of residue-level information considering all amino acid residues results in very high-dimensional data. In order to reduce the cardinality of the considered feature sets, we therefore decided to focus our attention on only those residues that, based on secondary structure predictions, are believed to be part of loop regions in the protein chain; we consider these loop regions to be of crucial importance for describing the overall folded state of a protein conformation.

3. EXPERIMENT DESIGN

In our experiments, we generated decoy structures for six different proteins and used a machine learning technique, *random forests*, to ask questions regarding the importance of particular features at different stages of the search. Random forest works by growing an ensemble of classification trees. Then, the classification provided by all the constructed trees is taken into consideration in order to make a final decision at the moment of classifying a new instance [1].

Below we describe the most relevant details of the conducted experiments:

- **Data.** Mapping from feature space (structures at the end of stages 1, 2, 3, or 4) to final (stage 4) energy

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Energies	•		•	•	•		•					
Contacts	•	•	•	•	•	•						
HP model	•	•		•		•						
Distances	•	•	•	•	•	•		•				
Torsion	•	•	•						•	•		
ABEGO	•	•	•						•		•	
Fragments	•	•	•						•			•

Table 1: Summary of the specific features considered in the different investigated feature subsets.

or RMSD value. Target classes were defined for "successful" or "unsuccessful" Rosetta runs, depending on whether the final prediction quality (energy or RMSD) of each trajectory was below or above the median of a sample of 10,000 runs.

- **Test proteins.** A total of 6 different proteins of varying size and structural properties were studied: 1acf (125, α+β), 1bk2 (57, β), 1enh (54, α), 1fna (91, β), 1lis (125, α), 1pgx (55, α+β).[1]
- **Features.** A total of 12 different subsets of features were considered (summarized in Table 1).
- **Parameters of learning process/random forest technique.** The prediction performance of random forest is sensitive to some important parameters. We therefore investigated different settings by varying the size of the set of training data, the number of trees in the forest, the number of features considered when looking for the best split at each node of the tree, as well as the number of samples required to split an internal node. In all the cases, a test set of size 1,000 was used, and several repetitions of the training/testing process for each of the investigated parameter configurations were performed.

4. RESULTS

Results obtained for the prediction of energy are shown in Figure 1, with the results for the prediction of RMSD in Figure 2. Each plot in these figures presents results for a particular protein with regard to the prediction of the success of individual Rosetta trajectories (refer to Section 3 for further details). Prediction accuracy is computed as the fraction of predictions that were correctly classified. Unsurprisingly, those feature sets that include energy (e.g. the feature set F7 which includes energy only) outperform other feature sets during the prediction of final energy, and this is consistent across all four stages. Also unsurprisingly, the prediction of final energy becomes easier (close to 100% accuracy) as the features of decoys from later stages are considered. In general, the relative performance advantage of those features that contain energy also becomes more pronounced for later stages; interesting exceptions to this are the two alpha helical proteins 1pgx and 1acf, and the beta protein 1bk2, for which the energy values after Stage 1 already contain some information regarding the final outcome of the search.

As the low-energy function in Rosetta is known to be inaccurate, the results related to the prediction of RMSD are potentially more relevant. Interestingly, energy is seen to be

[1] α, β, and α+β, stand respectively for the alpha helical, beta, and alpha-beta structural classes of proteins.

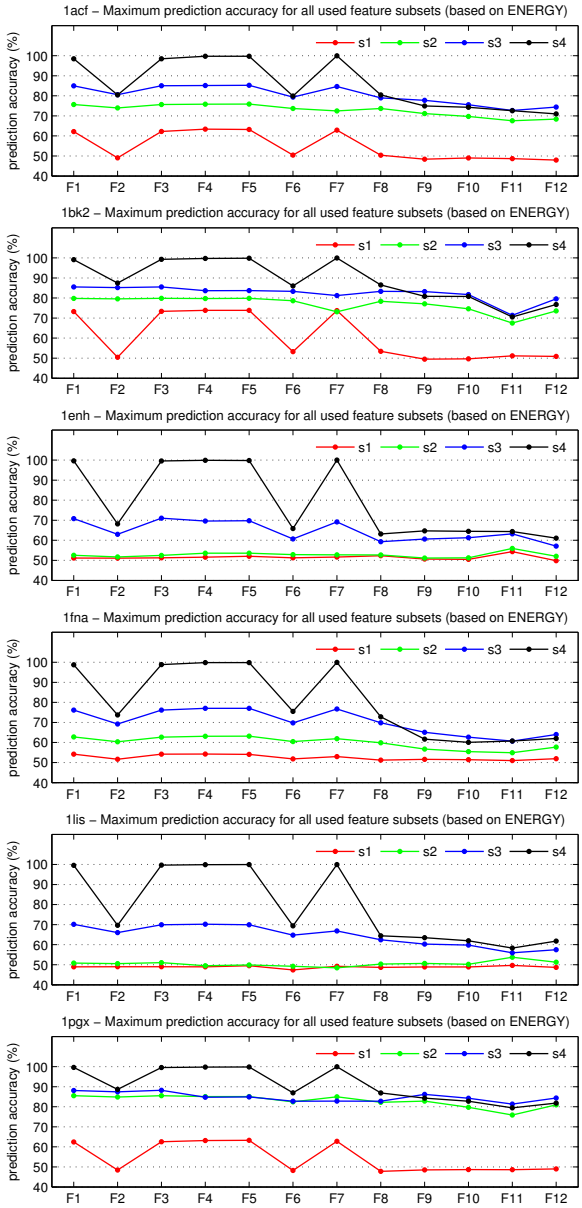

Figure 1: Average prediction accuracy obtained, where classification is based on energy values.

a fairly poor predictor of final distance to the native. For most proteins (the exception to this is 1pgx), some of the worst prediction accuracies are seen for the feature set F7, and this effect is present throughout all stages. One of the most relevant feature sets for final RMSD appears to be the global distance-based features, with the feature set F8 showing a robust performance across all six proteins. For alpha-helical proteins this feature becomes relevant from Stage 3 onwards only, while, for alpha-beta and beta proteins, it carries information from Stage 2 onwards. This confirms two different aspects about Rosetta's search protocol: the fact that the relative arrangement of secondary structure is indicative of different folds and may provide a meaningful descriptor of structural diversity. In addition, these results also confirm our experience that Rosetta trajectories tend

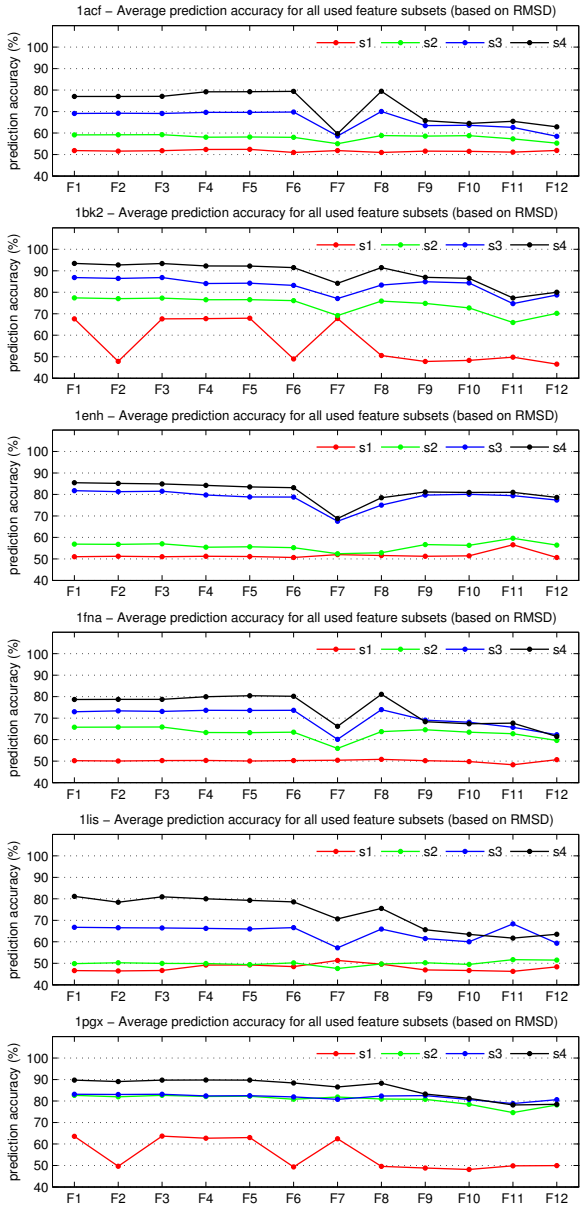

Figure 2: Average prediction accuracy obtained, where classification is based on RMSD values.

to collapse quickly, with key aspects of the fold becoming fixed from Stage 2 onwards.

The results regarding local feature sets are somewhat disappointing, as we expected (following the work of [5]) that certain key loop regions may play a more significant role in determining the final success of Rosetta trajectories. Our results indicate that, especially for the larger proteins (1acf, 1fna and 1lis), these features have relatively little influence on the final quality of the structure.

4.1 Conclusions

The results from our preliminary analysis are consistent with previous findings regarding the dynamics of conformational search in Rosetta, although we were expecting a more significant contribution of local "linchpin" features. In our future work, we will expand this analysis to integrate additional features, consider the ranking returned by the classifier, and consider results across a larger number of proteins. We are also exploring the possibility that a more meaningful formulation of this classification problem may require the separation between near-native and non-native structures, rather than our current differentiation between above-average and below-average performance.

5. REFERENCES

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] T. Brunette and O. Brock. Guiding conformation space search with an all-atom energy potential. *Proteins: Structure, Function, and Bioinformatics*, 73(4):958–972, 2008.

[3] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509, 1985.

[4] S. M. Kandathil, S. C. Lovell, and J. Handl. Towards a detailed understanding of search trajectories in fragment assembly approaches to protein structure prediction. Technical report, University of Manchester, 2014.

[5] D. E. Kim, B. Blum, P. Bradley, and D. Baker. Sampling bottlenecks in de novo protein structure prediction. *Journal of molecular biology*, 393(1):249–260, 2009.

[6] H. Lopes and M. Scapin. An enhanced genetic algorithm for protein structure prediction using the 2d hydrophobic-polar model. In E.-G. Talbi, P. Liardet, P. Collet, E. Lutton, and M. Schoenauer, editors, *Artificial Evolution*, volume 3871 of *Lecture Notes in Computer Science*, pages 238–246. Springer Berlin Heidelberg, 2006.

[7] K. Molloy, S. Saleh, and A. Shehu. Probabilistic search and energy guidance for biased decoy sampling in ab initio protein structure prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 10(5):1162–1175, Sept. 2013.

[8] S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, et al. Structure prediction for casp8 with all-atom refinement using rosetta. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):89–99, 2009.

[9] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods in enzymology*, 383:66–93, 2004.

[10] A. Shehu and B. Olson. Guiding the search for native-like protein conformations with an ab-initio tree-based exploration. *The International Journal of Robotics Research*, 29(8):1106–1127, 2010.

[11] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, et al. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences*, 105(12):4685–4690, 2008.

[12] D. Simoncini, F. Berenger, R. Shrestha, and K. Y. Zhang. A probabilistic fragment-based protein structure prediction algorithm. *PloS one*, 7(7):e38799, 2012.