On heuristic bias in fragment-assembly methods for protein structure prediction

Julia Handl[1], Mario Garza-Fabre[1], Shaun Kandathil[2] and Simon Lovell[2]

[1]Decision and Cognitive Sciences Research Centre, University of Manchester, Manchester, UK,
[2]School of Biological Sciences, University of Manchester, Manchester, UK
Email: julia.handl@manchester.ac.uk

Abstract

We discuss the issue of heuristic bias in fragment-assembly methods for protein structure prediction. We explain the importance of this issue, which has been paid insufficient attention by evolutionary computation researchers engaging with the structural biology community. We proceed by describing preliminary data that illustrates the significant (and expectable) impact that fragment library composition has on search performance, and discuss the challenges this poses for the development of improved fragment libraries.

1 Introduction

Heuristic optimization approaches are of increasing importance in identifying solutions to complex optimization problems that cannot be addressed using methods from exact optimization alone. Meta-heuristic optimizers, in particular, play a crucial role in identifying approximate solutions to problems that are challenging due to their scale, the presence of uncertainties and noise, and/or the existence of multiple conflicting criteria. Meta-heuristic optimizers are fundamentally designed as "off-the-shelf" methods that are sufficiently general to be useful for a diverse range of non-linear, global optimization problems. Nevertheless, to obtain competitive performance on difficult real-world problems, a careful design of representation, variation and initialization operators that introduce suitable heuristic bias, as well as rigorous tuning, is often essential, and can be, arguably, of more importance than the basic choice of meta-heuristic.

State-of-the-art methods for protein structure prediction typically employ a meta-heuristic optimizer, including methods such as evolutionary algorithms [1], EDAs [20] and simulated annealing [18]. There have been a number of recent papers by evolutionary computation researchers that consider the deployment and design of state-of-the-art meta-heuristics for this problem (see e.g. [19, 4, 7]). In terms of the choice of representation of candidate protein structures, there have been fewer concrete contributions from this community.

The class of fragment-assembly methods has remained the *de novo* prediction approach of choice for the past two decades. Fragment-assembly approaches typically employ an internal low-resolution representation of protein structure,

e.g. based on backbone torsion angles, and use insertions of short segments from known protein structures as their variation operator. More specifically, there are two aspects to this internal representation.

- The specific low-resolution representation used.
- The choice of the fragment library, which defines the values available for insertion at each position.

Arguably, the second of these is the most influential from a heuristic optimization perspective, as the composition of the fragment library restricts the available search space and may introduce significant heuristic bias towards certain regions of this space. This effective reduction of the search space is widely seen as the key strength of fragment-assembly, but it can also be seen as the Achilles heel of the approach: an unfavourable bias will clearly introduce problems for search algorithms, the severity of which will depend on the sensitivity of the search protocol to such bias, and aspects of the objective function.

While there has been extensive research on the development of improved fragment libraries in the structural biology community, this has focused on improving the biophysical plausibility of candidate fragments. As far as we are aware there is no published work that directly considers the impact of fragment library composition on search performance. The closest work that touches upon the issue is [3, 9], but this focuses on aspects of the variation operators, and the impact of their definition on the search space and / or search performance. In our current work, we are interested in defining the impact fragment selection will have on search, and to explain previous findings on prediction performance from this perspective.

The remainder of this paper is structured as follows. Section 2 describes the standard process of generating a fragment library, using the example of Rosetta's fragment picker. Section 3 discusses and formalizes the heuristic bias that the fragment library introduces into the search process. Section 4 summarizes current evidence for the existence, and the significant impact, of such bias. Section 5 highlights implications for future work and concludes.

2 Fragment-assembly and fragment library construction

2.1 Fragment-assembly methods

Predicting protein tertiary structure from sequence information remains an important unsolved problem. Techniques based on the principle of fragment-assembly [21] have emerged as the leading class of methods to tackle this problem, as evidenced by their performance in the CASP experiments [15, 13]. However, their accuracy is known to decrease for larger, more complex proteins [13].

In general, fragment-assembly techniques rely on the fact that secondary and tertiary structure can be strongly influenced by local amino acid sequence [21]. These local propensities are taken into account and exploited during model construction, by deriving fragments from known protein structures and using them as building blocks during the search. The search techniques employed are heuristic optimization algorithms that start from an initial structure (e.g. a

fully extended chain), and which iteratively apply randomly selected fragment insertions to generate novel candidate structures. An energy or scoring function is used to determine whether a particular candidate structure should be accepted. A key assumption behind the use of an optimization procedure is that near-native structures correspond to at least a local optimum in the energy landscape defined by this function. State-of-the-art fragment-based prediction pipelines typically employ many independent runs of a prediction technique (the random-restart strategy) to arrive at a pool of structures, from which a subset of promising predictions are chosen.

2.2 Fragment library generation - The example of Rosetta

The first step in applying any fragment-based prediction method is the selection of appropriate structural fragments for the target protein sequence. Fragments are typically identified based on sequence and structure profiles (obtained from multiple sequence alignments), on the basis of threading against known templates, or by using constant fragment sets selected from a non-redundant set of structures.

Like other methods, Rosetta's fragment generation process employs automated secondary structure prediction methods to inform the choice of fragments chosen for any given window of the sequence. A maximum of 3 three-state secondary structure predictors can be used: PSIPRED [10, 2], SAM (Sequence alignment and Modelling; [12]) and Porter [16, 14] are currently supported.

The fragment picking process identifies putative fragments through the application of a scoring function. Many different criteria can be used to score fragments [8], and some commonly used metrics include similarity scores based on PSI-BLAST sequence profiles, similarity between the predicted secondary structure for a local sequence and fragment secondary structure, and agreement with backbone torsion angles and solvent accessibility predictions from SPINE X [6]. Different scoring criteria can be assigned different priorities when selecting fragments; sequence profiles generally have the highest priority in deciding what fragments should be selected. If an insufficient number of candidate fragments are identified based on sequence profiles, the criterion with the next-highest priority value is used to select fragments (in this case, agreement with PSIPRED predictions), and so on. Other criteria may include agreement with experimental data (such as chemical shifts) or other distance- or angle-based constraint information.

Following the selection and scoring of putative fragments, the 200 highest-scoring fragments are returned in the fragment libraries, which can then be used for *de novo* structure prediction using Rosetta. Rosetta's fragment generation process is typically used to produce libraries of fragments that are 9 and 3 residues long (9-mers and 3-mers, used during different stages of Rosetta's *ab initio* protocol), although alternative lengths can be specified [9].

2.3 Diversity mechanisms during fragment picking

An interesting aspect of Rosetta's fragment generation process is the inclusion of a range of different criteria in the pipeline. This is testament to the fact that the definition of the selection criterion is difficult, that reliance on a single criterion

may be insufficient or risky, and that different criteria may gain importance in specific circumstances.

More fundamentally, we note that, in Rosetta's fragment picker, when more than one secondary structure predictor is used (see above), a quota system can be enabled, by which the fragment picker selects a certain percentage of fragments for each window based on each predictor (Gront et al., 2011). This quota mechanism is aimed at providing additional diversity in the fragment set, in situations when the secondary structure predictions produced by different methods do not agree.

Similarly, for any single predictor, fragments are chosen such that the predicted likelihood of the three secondary structural types (helix, strand or loop) for any residue are maintained as best as possible in the resulting fragment set.

The above features indicate that there is a clear appreciation that a poor quality fragment-library can be damaging to the fragment-assembly process and that there are two conflicting aspects to this. Firstly, the fragment library narrows down the search space. This facilitates the search process, and is the main driver behind the success of fragment-assembly methods. Secondly, where the fragment library for a particular position is inappropriate (i.e. it contains only non-native local structures), this will make it difficult, if not impossible, to identify a native structure at the tertiary level.

3 Heuristic bias and its presence in fragment-assembly

Raidl and Gottlieb [17] define heuristic bias as follows: "Heuristic bias concerns the mapping from search space to phenotype space... The efficacy of the search process is strongly influenced by the mapping between these spaces. Hence, using some heuristic in this mapping yields a certain distribution of phenotypes, which can help to increase performance if the distribution is biased towards phenotypes of higher fitness."

A differentiation between the genotype and the phenotype in a fragment-assembly method is not entirely straightforward. For a protein with N residues and a default use of Rosetta during its first three stages (use of 9mers, 25 fragments per position), one possible way to think about the genotype is to consider it a string of N integers, where each position can take up to 225 possible values[1], corresponding to an index of all possible angle triplets available for this position [9]. While this genotype is never explicitly encoded within the fragment-assembly method, this abstract definition allows us to think about the size of the search space independently of the choice of fragment library employed and the variation operator used.[2] The choice of variation operator can have the effect of eliminating access to portions of this search space, but this issue has been discussed in [9] and is not further considered here.

[1]The 225 values stem from nine overlapping insertion windows. There are less overlapping windows (and therefore values) for the first and the last eight positions of the string, see [9] for details.

[2]If we considered the genotype to correspond directly to the angle-based representation explicitly encoded in methods such as Rosetta, the set of possible genotypes/size of the search space would no longer be independent of the choice of fragment library. Furthermore, the likelihood of possible genotypes would be non-uniform in the sense that different instantiations arise with different probability.

Our focus here is on heuristic bias, i.e. the bias resulting from mapping the above genotype to the phenotype. Essentially, this can be thought to correspond to the mapping of each integer within our abstract genotype to the triplet of torsion angles that it indexes within the fragment-library, and the final decoding of the string of backbone angles into a tertiary structure. This dual mapping process is independent of the choice of variation operator, but, in itself, it clearly has the potential to introduce bias towards particular phenotypes. The vehicle controlling this bias is the choice of fragment library alone. We therefore take the view that, from an optimization perspective, the design of a fragment library fundamentally corresponds to the problem of defining a genotype-phenotype mapping with appropriate heuristic bias.

Considering the mechanisms discussed in Subsection 2.3, it is evident that some of these mechanisms have been designed to counter-balance the risks introduced through the fragment-picking process. Some of the procedures incorporated into existing methods implicitly reflect an understanding that, in regions where significant uncertainty remains regarding the local propensity towards particular types of secondary structure, fragment libraries need to remain diversified to allow for the balanced exploration of different types of solutions. In other words, this can be seen as preliminary attempts to control the amount of heuristic bias introduced for different parts of the protein chain. As fragment library composition has not usually been considered from a search perspective, it remains unclear to what extent these current ways of library construction are sufficient to ensure that access of the native structure does not become intractable for standard search heuristics.

In particular, it is unknown to what extent current fragment generation methods do indeed manage to achieve a suitable balance between helpful heuristic bias and a retention of unbiased options in those areas where uncertainty regarding structure propensity remains. This is due to a number of factors.

In Rosetta's fragment picker, diversity is implicitly defined at the level of the secondary structure type (i.e. three classes: alpha, beta, loops), but it is unclear whether this is appropriate and sufficient, e.g. as some types of local structure (helices) are significantly less diversified than others (especially loops, but also beta sheets).

Furthermore, estimates of the reliability of secondary structure predictions are taken from the secondary structure predictors, but the literature is unclear as to how accurate these estimates are (this is different to the actual estimates of prediction performance). This may be an issue when these estimates are used to inform the amount of diversity retained in the libraries, as is the case for Rosetta's fragment picker, see above.

Finally, we note that the variation operators used in most fragment-assembly protocols consist of full-fragment insertions. This removes access to some areas of the search space, and introduces interactions between the fragment libraries of neighboring positions. Together, this has the potential to further reinforce any bias introduced through the choice of fragment library.

4 Consequence of heuristic bias in fragment-assembly

In this manuscript, we aim to define the nature of heuristic bias in the context of fragment-assembly methods and to encourage the community to reconsider

the performance of current protocols in this context. To further emphasize this point, this section highlights recent results from the academic literature that, we believe, indicate the importance of the issue.

Recent research compared different fragment libraries in a setting that eliminated the confounding impact of imprecise energy functions and heuristic optimizers (through the use of a structure-based objective and a greedy construction heuristic) [22]. It was observed that fragment libraries constructed using sequence profiles alone allowed for a more accurate reconstruction of the native structure. However, when fragment selection considered secondary structures, this led to a pronounced reduction in the diversity of fragments. This goes some way to explain why state-of-the-art methods typically use both types of information. The search space reduction arising from the use of secondary structure information is likely to lead to "quick" wins on easy prediction targets, which will have contributed to the adoption of this approach in state-of-the-art pipelines. For future research, the finding does raise the question of whether more diverse libraries and improved search techniques may be a more fruitful avenue to scale prediction methods to more complex targets.

In another recent paper on fragment library construction [5], the authors found that a selection approach that applied scoring to a random sample outperformed the alternative of exhaustive scoring of all fragments. In particular, the resulting fragment libraries provided higher precision and coverage. This provides an additional indication that, given our reliance on imperfect fragment scoring criteria, a controlled diversification of fragment libraries may be desirable, even when the impact of heuristic search is not considered. It is currently unclear how this diversification is best approached to ensure inclusion of the most accurate fragments, and to appropriately moderate heuristic bias.

Our own experiments with iterated local search heuristics reveal significant differences in performance for different fragment libraries [11]. Strikingly, these sensitivities are significantly more pronounced for advanced search heuristics than for simple restart protocols such as Rosetta (see Figure 1), consistent with the increased sensitivity of such techniques to heuristic bias. Our observations also go some way to explain why the design of advanced sampling protocols has often led to limited success in the literature: the potential advantages arising from improved sampling may have been rendered insignificant by misleading heuristic bias, introduced through the use of inappropriate fragment libraries.

5 Conclusion

Moving forward, we believe that the subject of heuristic bias needs to be considered much more explicitly in the design and comparison of prediction protocols. Specifically, it can be challenging to draw conclusions regarding the performance of search techniques, where contestant techniques are tested in the context of different (customized) fragment-libraries, and are thus operating in search spaces with potentially different amounts of bias. Similarly, while methods are typically tested across a range of target proteins, deliberate testing across fragment libraries with different (known) levels of diversity / heuristic bias has not been considered. This would be desirable as such a setup appears to be more powerful at identifying differences in the performance of the search techniques. In our immediate future work, we will be developing strategies to explicitly under-

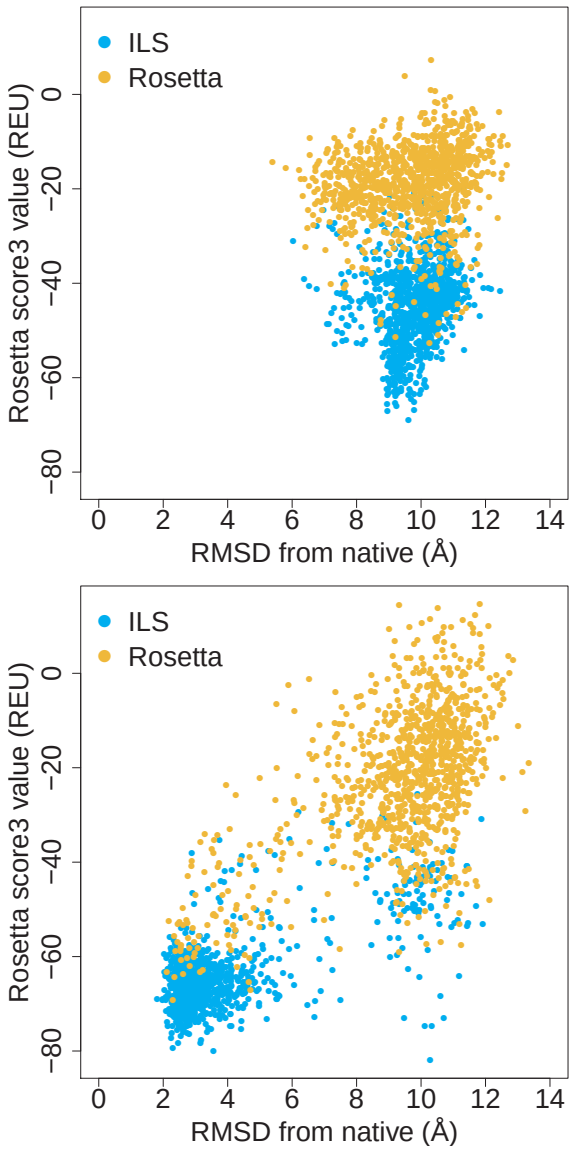

Figure 1: Results obtained by standard Rosetta (using restarts) and an iterated local search protocol [11] across two different fragment libraries for the same protein (1c8cA). The newer fragments (results in top figure) were generated using the fragment picker and the structure database supplied with Rosetta version 3.5 (weekly release 2014.16.56682). The older fragments (results in bottom figure) are taken from a previous study [9].

stand and control diversity of fragment libraries. This will feed into practical improvements of fragment libraries, but also the design of benchmark libraries that support the rigorous testing of new search protocols.

Acknowledgment

This work is funded by grant EP/M013766/1, Engineering and Physical Sciences Research Council, UK.

References

[1] James U Bowie and David Eisenberg. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proceedings of the National Academy of Sciences*, 91(10):4436–4440, 1994.

[2] Daniel WA Buchan, Federico Minneci, Tim CO Nugent, Kevin Bryson, and David T Jones. Scalable web services for the psipred protein analysis workbench. *Nucleic acids research*, 41(W1):W349–W357, 2013.

[3] George Chikenji, Yoshimi Fujitsuka, and Shoji Takada. A reversible fragment assembly method for de novo protein structure prediction. *The Journal of chemical physics*, 119(13):6895–6903, 2003.

[4] Rudy Clausen, Emmanuel Sapin, Kenneth A De Jong, and Amarda Shehu. Evolution strategies for exploring protein energy landscapes. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 217–224. ACM, 2015.

[5] Saulo HP de Oliveira, Jiye Shi, and Charlotte M Deane. Building a better fragment library for de novo protein structure prediction. *PloS one*, 10(4):e0123998, 2015.

[6] Eshel Faraggi, Tuo Zhang, Yuedong Yang, Lukasz Kurgan, and Yaoqi Zhou. Spine x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of computational chemistry*, 33(3):259–267, 2012.

[7] Mario Garza-Fabre, Shaun M Kandathil, Julia Handl, Joshua Knowles, and Simon C Lovell. Generating, maintaining, and exploiting diversity in a memetic algorithm for protein structure prediction. *Evolutionary computation*, 24(4):577–607, 2016.

[8] Dominik Gront, Daniel W Kulp, Robert M Vernon, Charlie EM Strauss, and David Baker. Generalized fragment picking in rosetta: design, protocols and applications. *PloS one*, 6(8):e23294, 2011.

[9] Julia Handl, Joshua Knowles, Robert Vernon, David Baker, and Simon C Lovell. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 80(2):490–504, 2012.

[10] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, 1999.

[11] Shaun M Kandathil, Mario Garza-Fabre, Julia Handl, and Simon C Lovell. Improved fragment-based protein structure prediction by redesign of search heuristics. *Under submission*, 2017.

[12] Sol Katzman, Christian Barrett, Grant Thiltgen, Rachel Karchin, and Kevin Karplus. Predict-2nd: a tool for generalized protein local structure prediction. *Bioinformatics*, 24(21):2453–2459, 2008.

[13] Andriy Kryshtafovych, Krzysztof Fidelis, and John Moult. Casp10 results compared to those of previous casp experiments. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):164–174, 2014.

[14] Claudio Mirabello and Gianluca Pollastri. Porter, paleale 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*, 29(16):2056–2058, 2013.

[15] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (casp)—round x. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):1–6, 2014.

[16] Gianluca Pollastri and Aoife Mclysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–1720, 2005.

[17] Günther R Raidl and Jens Gottlieb. Empirical analysis of locality, heritability and heuristic bias in evolutionary algorithms: A case study for the multidimensional knapsack problem. *Evolutionary Computation*, 13(4):441–475, 2005.

[18] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. *Methods in enzymology*, 383:66–93, 2004.

[19] Alena Shmygelska and Michael Levitt. Generalized ensemble methods for de novo structure prediction. *Proceedings of the National Academy of Sciences*, 106(5):1415–1420, 2009.

[20] David Simoncini, Francois Berenger, Rojan Shrestha, and Kam YJ Zhang. A probabilistic fragment-based protein structure prediction algorithm. *PloS one*, 7(7):e38799, 2012.

[21] Kim T Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of molecular biology*, 268(1):209–225, 1997.

[22] Raphael Trevizani, Fábio Lima Custódio, Karina Baptista dos Santos, and Laurent Emmanuel Dardenne. Critical features of fragment libraries for protein structure prediction. *PloS one*, 12(1):e0170131, 2017.