APPLYING FEATURE-BASED RESAMPLING TO PROTEIN STRUCTURE PREDICTION

Trent Higgs[1], Bela Stantic[1], Md Tamjidul Hoque[2] and Abdul Sattar[13]

[1]Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Australia.

[2]School of Informatics, Indiana Center for Computational Biology and Bioinformatics,
Indiana University Purdue University Indianapolis (IUPUI), USA.

[3] NICTA Queensland Research Laboratory.

{t.higgs, b.stantic, a.sattar}@griffith.edu.au, tamjidul.hoque@gmail.com

Abstract

Protein structure prediction (PSP) has been utilised in numerous biological areas (e.g. drug design) as a protein's shape dictates the function it performs within the cell. Therefore, in computer science we are trying to develop algorithms that accurately and quickly determine the three-dimensional conformation of a protein. To achieve this a lot of PSP algorithms use fragment libraries to limit the amount of conformations considered for a particular protein. The idea of fragments can be easily incorporated into feature-based resampling, with the fragments being considered as the features. In this paper we have analysed features from previous sampling rounds to gauge how similar they are compared to their corresponding native segments, and we have also investigated the effects different feature sets (i.e. decoys) have on feature-based resampling algorithms. From our experimentation we discovered that features/fragments contained within previous sampling rounds can be very similar to their native counterparts, and that our Genetic Algorithm (GA) feature-based resampling algorithm was able to refine structures with features generated by another state-of-the-art PSP suite by gaining an average 6.64% $RMSD$ improvement and an average 3.49% TM-*Score* improvement.

1 INTRODUCTION

Proteins are the most important macromolecules in all living organisms. Over half of the dry weight of a cell is made up of proteins of numerous shapes and sizes. A protein is formed by a string of amino acids folding into a specific three-dimensional shape, which determines the *biological task* it will perform. An example of this would be the haemoglobin, which performs the task of carrying oxygen to the blood stream. To elicit these three-dimensional shapes, a process known as *protein structure prediction* (PSP) is carried out.

To speed up the PSP process numerous computational methods have been introduced. These methods are usually grouped into three main categories, *comparative* or *homology modelling*, *threading* or *fold recognition*, and *ab initio* or *de nova*. A technique some of these methods use to help with the prediction process is fragments or fragment libraries. Fragments can help by limiting the amount of possible conformations considered for a particular segment of the protein's target sequence and highly-probable fragments allow an approximation of the populated areas of the local energy surface for the backbone of a particular protein structure. It does this by generating potential samples of the energy landscape, which allows the energy function to mainly focus on the global interactions of these fragments to select the best compact structures.

In [4] we developed a PSP approach that combined the concepts of fragments and feature-based resampling. Feature-based resamplings main goal is to take an already explored search space and find *native-like* features contained within it that may produce more accurate structures when combined together. For this we used a Genetic Algorithm (GA), which easily allowed us to crossover features from the previous search space to produce lower scoring models. In this work we have conducted extensive experiments that show how close some of the features/fragments are within the previous sampling round, and we have also conducted experiments to evaluate the successfulness of our approach by using features from a different PSP suite as the initial population. The results from these tests demonstrated that our GA feature-based resampling algorithm was able to refine structures, generated by a different state-of-the-art PSP suite, by gaining improvements in both $RMSD$ and TM-*Score*.

The rest of this paper is organised as follows. In Section 2 we discuss the general background, Section

Table 1: Results obtained from our feature analysis between decoys and their native conformation.

Rosetta			**I-Tasser**		
Protein	**Length**	**Avg $RMSD$**	**Protein**	**Length**	**Avg $RMSD$**
2ptl_	78	3.62 Å	1af7_	72	2.97 Å
1pgx_	83	3.36 Å	1fo5A	85	2.94 Å
1bds_	43	4.00 Å	1b72A	49	2.38 Å
1bm8_	99	5.74 Å	1fadA	92	2.02 Å
1emw_	88	4.96 Å	1tig_	88	3.62 Å
1aoy_	78	3.39 Å	1sro_	71	2.68 Å
1csp_	67	2.95 Å	1r69_	61	1.43 Å
2ppp_	107	5.18 Å	1egxA	115	1.80 Å
1kjs_	74	3.56 Å	1b4bA	71	2.79 Å
1vcc_	77	3.95 Å	1gjxA	77	4.03 Å

3 we will outline our methodology, Section 4 presents and analyses the results we gained from our experimentation, and finally in Section 5 we draw our conclusions.

2 BACKGROUND

Fragments or motifs are one or more secondary structure elements that make up a proteins three-dimensional structure. To this end a lot of protein folding software, Rosetta [2] [10], Fragfold [6], Undertaker [7], and Tasser [13] [15] [16], all use fragments or motifs within their search algorithms to build a proteins three-dimensional conformation, and to reduce the conformational search space (i.e. by limiting the amount of possible conformations a certain segment within the protein chain can adopt).

The concept of feature-based resampling can easily be combined with this fragment-based search technique by making the fragments the features of the search. Resampling techniques main goal is to refine a search space that has already been sampled. In PSP this can be looked at as taking already computed local minima and finding samples of conformation space within it that indicate regions containing consistently lower energy, and focusing further searching around those regions. There are two main types of resampling used in PSP: structure-based resampling, and feature-based resampling [3].

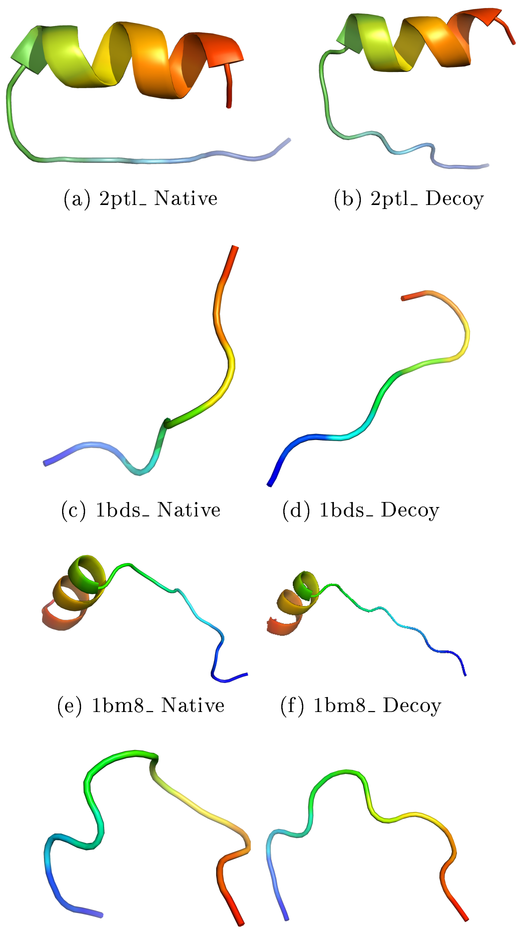
Structure-based resampling techniques either selects regions of conformation space or individual protein structures from the initial search, which showed a lot of promise. It then focuses further sampling around them. Feature-based resampling on the other hand is more concerned with *native-like* features from the previous sampling round. If no models from the previous round of sampling produces a structure close enough with the native structure, they still may contain various *native-like* features, which can be recombined to create new structures that are closer to the native conformation.

In [4] we developed a GA feature-based resampling algorithm, which combined the concepts of resampling and fragments. A GA search was the best option due to the crossover operation allowing global solutions to be built by the cooperative combination of numerous local sub-structures [12]. Another benefit can be seen that if a select local sub-structure is irrelevant to one solution there is a high probability that it will be relevant to another solution in the population. If this is the case the GA search can be looked at being driven by implicit parallelism, which means it will produce more successful descendants than when compared to a random search.

Due to having promising results in our previous work [4] [5] we have conducted several experiments to highlight how powerful using a feature-based resampling approach can be given the right combination of search algorithm, energy function, and feature set. In the next section we will discuss in detail how we achieved this.

3 METHODOLOGY

The idea of combining fragments and non-deterministic search for the PSP problem was first introduced by Baker and his group [11]. This approach to protein structure prediction performed very well, and other successful PSP suites (e.g. I-Tasser) have utilised similar methods. Most fragments that are used in PSP are picked based off of sequence alignment [1] and secondary structure prediction [9]. Instead of doing this step, our algorithm uses fragments generated from a previous sampling round. This means they should already be very close to their native structure; seeing that the PSP software used to generate them


(a) 2ptl_ Native (b) 2ptl_ Decoy

(c) 1bds_ Native (d) 1bds_ Decoy

(e) 1bm8_ Native (f) 1bm8_ Decoy

(g) 1gjxA Native (h) 1gjxA Decoy

Figure 1: In (b), (d), (f), and (h) we present some of the fragments/features produced from decoys in our feature analysis tests, and in (a), (c), (e), and (g) we show their corresponding features/fragments in their native structure.

would of already done some form of sequence alignment and secondary structure prediction.

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1} ||\, v_i - w_i \,||^2} \qquad (1)$$

To further highlight the potential of our approach developed in [4] we have analysed fragments/features from previous sampling rounds, and evaluated the effect of using decoys produced by a different state-of-the-art PSP suite. Analysing fragments or features from previous sampling rounds should allow us to show how close features found in local minima can be to their native conformation. If this is the case then it can bee seen that using a GA optimisation approach to reassemble these features more native conformations should be formed. Using decoys generated by a different PSP suit as the initial population will allow us to investigate how versatile our approach is. For details about our GA feature-based resampling approach please see [4] and [5].

To test how close fragments/features are from a previous sampling round we have created 1000 decoys (a decoy is a candidate solution for a target sequence), using the PSP tool Rosetta, for each protein. These 1000 structures are then read into our program, and for 500 iterations we perform the following steps:

1. Randomly select a decoy.
2. Randomly select a start and end point for a fragment of size n in the chosen decoy (*decoy_frag*).
3. Find the corresponding matching fragment in the native structure (*native_frag*).
4. Compare the structural similarity between *native_frag* and *decoy_frag*.

For the structural measure in our tests we have used a normalised version of the *root mean square de-*

Table 2: GA Results compared to Rosetta as presented in [4]. $RMSD$ Imp refers to the $RMSD$ improvement, and TM Imp refers to TM-$Score$ Improvement. The average $RMSD$ improvement was 9.5% and the average TM-$Score$ improvement was 17.36%.

Protein	**Length**	**Rosetta**			**GA**				
		f	$RMSD$	TM-$Score$	f	$RMSD$	TM-$Score$	$RMSD$ Imp	TM Imp
2ptl_	78	-124.10	8.764Å	0.4084	-129.46	3.888Å	0.5442	55.64%	33.25%
1pgx_	83	-120.12	4.385Å	0.6570	-88.82	3.980Å	0.6304	9.24%	-4.05%
1bds_	43	-21.97	6.151Å	0.2138	6.88	5.954Å	0.2406	3.20%	12.54%
1bm_8	99	-82.09	7.814Å	0.2751	-67.86	7.688Å	0.2973	1.61%	8.07%
1emw_	88	-48.36	8.415Å	0.2856	-51.69	8.591Å	0.3097	-2.09%	8.44%
1aoy_	78	-57.80	6.028Å	0.3854	-62.71	5.350Å	0.5425	11.25%	40.76%
1csp_	67	-84.95	2.576Å	0.7156	-79.62	2.527Å	0.7305	1.90%	2.08%
2ppp_	107	-27.43	9.632Å	0.2780	-48.82	9.681Å	0.4483	-0.51%	61.26%
1kjs_	74	-42.35	4.541Å	0.4706	-48.51	4.431Å	0.4999	2.42%	6.23%
1vcc_	77	-70.08	2.950Å	0.6800	-60.89	2.586Å	0.7140	12.34%	5.00%

viation ($RMSD$, see Equation 1) [8]. To cross validate these results we have run a similar experiment using I-Tasser decoys, another state-of-the-art PSP suite. For these tests our decoy sets range from 300-600 in size, due to the data made available at [14]. We have also chosen different proteins for these experiments for two reasons: (1) not all of the same proteins were available at [14], and (2) to allow for a more rigorous evaluation.

After each iteration the normalised $RMSD$ between the fragment/feature and its corresponding native segment is stored. Once 500 iterations has been reached these stored values are then averaged and that value is kept as the final result for a particular protein.

4 EMPIRICAL RESULTS

In Table 1 we present the results for each protein we tested in our feature analysis experiments. Here we list for both Rosetta and I-Tasser decoys: the PDB identifier for the protein, the length of the protein, and its average $RMSD$ for the entire 500 iterations. In Figure 1 we have shown visual comparisons of features found in previous sampling rounds compared to their corresponding native conformation.

Table 2 contains the results we obtained in [4], and Table 3 contains the results obtained from our GA approach using decoys generated by I-Tasser (applying the same experimental set up as described in [4] to allow for a fair comparison between the two). Both of these tables contain: the PDB identifier, the protein's length, the fitness, $RMSD$ and TM-$score$ for the highest ranked I-Tasser/Rosetta decoy, and the fitness, $RMSD$, TM-$Score$, $RMSD$ improvement in %, and TM-$Score$ improvement in % for the highest ranked GA decoy. The highest ranked decoy refers to the predicted protein structure, which had the best $RMSD$ result (i.e. closest to 0). Finally in Figure 2 we demonstrate how well our algorithm performed compared to the average $RMSD$ of the initial population (i.e. I-Tasser decoys).

4.1 Analysis and Discussion

From our results in Table 1 you can see on average fragments/features present in local minima from other sampling rounds can be very close to their native conformation. For Rosetta decoys after 500 iterations over 1000 decoys our worst average $RMSD$ was 5.74 Å for protein 1bm8. However, this is a very low similarity difference, which means these decoys will still contain a large amount of features that are close to their native structure. Overall you can see that we have an average 4.01 Å difference between fragments contained within Rosetta decoys from their actual native counterparts. This gives us empirical evidence that applying a feature-based resampling method should be able to produce more accurate protein models.

In regards to our I-Tasser tests (see Table 1) you can see that these decoys also have a large amount of features that are close to their corresponding native segments. The worst average $RMSD$ value we obtained in these experiments was 4.03 Å for protein 1gjxA, which is very similar to the worst average $RMSD$ produced by our Rosetta decoys (i.e. 5.74 Å). However, the overall average $RMSD$ between the two is quite different. As mentioned before the overall average $RMSD$ for our Rosetta decoys was 4.01 Å, whereas for the I-Tasser decoys we used it was 2.66 Å. This shows us that I-Tasser decoys contain more overall *native-like* features, which could significantly improve predictions made by feature-based resampling algorithms.

To further highlight our results we have presented

Table 3: GA Results compared to I-Tasser. $RMSD$ Imp refers to the $RMSD$ improvement, and TM Imp refers to *TM-Score* Improvement. The average $RMSD$ improvement was 6.64% and the average *TM-Score* improvement was 3.49%.

Protein	**Length**	**I-Tasser**			**GA**				
		f	$RMSD$	*TM-Score*	f	$RMSD$	*TM-Score*	$RMSD$ Imp	TM Imp
1af7_	72	-12.78	3.604Å	0.5090	-26.34	3.261Å	0.5237	9.52%	2.89%
1fo5A	85	-37.86	3.637Å	0.5599	-38.99	3.597Å	0.5744	1.09%	2.59%
1b72A	49	-17.73	3.167Å	0.6919	-23.10	2.694Å	0.7139	14.94%	3.18%
1fadA	92	-36.06	3.340Å	0.6006	-58.79	3.251Å	0.6184	2.66%	2.96%
1tig_	88	-10.13	3.522Å	0.5642	-27.42	3.508Å	0.5786	0.40%	2.55%
1sro_	71	15.02	3.075Å	0.7033	-68.24	2.359Å	0.6823	23.28%	-2.99%
1r69	61	-15.57	1.407Å	0.8301	-15.57	1.407Å	0.8301	0.00%	0.00%
1egxA	115	-11.29	2.052Å	0.8329	-106.34	1.895Å	0.8506	7.65%	2.13%
1b4bA	71	-11.33	4.408Å	0.4471	-25.87	4.204Å	0.5078	4.60%	13.58%
1gjxA	77	-11.71	5.704Å	0.4282	-29.40	5.574Å	0.4624	2.28%	7.99%

some visual samples (see Figure 1) of fragments created in our experiments and their corresponding native features/fragments. From Figure 1 we can see that features contained within decoys from a previous sampling round can be highly-similar to their native counterparts. By applying further optimisation through feature assembly and rotation we should be able to produce more *native-like* structures given an accurate energy function. Both of these operations can be efficiently incorporated in the GA search process due to its inbuilt crossover and mutation operators.

Our results from Table 1 show that I-Tasser decoys appear to have a large amount of native features contained within them. In our predictions using I-Tasser decoys (See Table 3) we obtained an average 6.64% $RMSD$ improvement, and a 3.49% *TM-Score* improvement. From this we can see that our $RMSD$ improvement was close to the results we obtained when using Rosetta decoys (e.g. 9.5%, See Table 2), however our *TM-score* improvement was considerably lower (3.49% compared to 17.36%) [4]. One of the main reasons why we gained better improvement percentages when using Rosetta decoys would be due to how close the best I-Tasser structures were to the native conformation. On average the best models produced by I-Tasser were 3.39Å away from the native, where as Rosetta decoys were 5.67Å, only leaving 2-3Å at most for our algorithm to improve upon.

In Figure 2 we show that even though our overall improvement was not that high the models our algorithm produced were considerably more accurate than the average $RMSD$ of the decoys contained in the initial population. From Figure 2 you can see only protein 1fo5A and 1fadA were still quite close to their average starting points. All other structures predicted by our algorithm improved substantially.

5 CONCLUSION

Protein structure prediction (PSP) suites that utilise protein fragments have proven to be very successful in producing accurate models. This has been demonstrated numerous times in Critical Assessment of Techniques for Protein Structure Prediction (CASP) where PSP programs like Rosetta and I-Tasser have both scored very well.

The concept of fragments and fragment libraries can be further extended into feature-based resampling techniques. The features used in the resampling phase can be viewed as the fragments. To show the potential of combining these two concepts together we have conducted several experiments that analyse how close features/fragments from previous sampling rounds can be to their corresponding native counterparts. From our results we have shown that they can be very similar by obtaining an overall average $RMSD$ of 4.01 Å for Rosetta decoys and 2.66 Å for I-Tasser decoys. This suggests that by applying further optimisation through feature assembly and rotation we should be able to produce more *native-like* structures.

To further highlight the successfulness of our GA approach we conducted experiments using features generated by I-Tasser. We achieved a 6.64% $RMSD$ improvement, and a 3.49% *TM-Score* improvement. Both of these totals were less than when we used Rosetta features, however the best I-Tasser models were already really close to the native conformation (an average 3.39Å), which meant our algorithm was limited to small improvements to begin with.

For future work it would be interesting to investigate mixing features together from various PSP suites, and combing different energy functions together to try and generate more accurate models. Seeing, that the I-Tasser models were already so close to the native con-

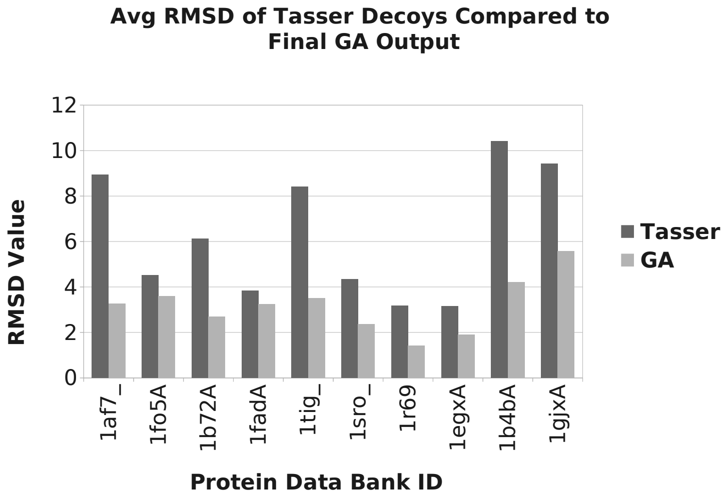


Figure 2: Here we show the average *RMSD* of all the I-Tasser decoys we used in our initial population, before our GA approach was started. To highlight our results we show the best model our GA outputted in comparison to this average.

formation more diverse structures may add features, which the search is missing. Also, a more strict energy function after a certain amount of generations would help limit the amount of moves that may degrade a particular structure from its native conformation.

References

[1] S. Altschul, T. Madden, A. Schaffer, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[2] D. Baker. Prediction and design of macromolecular structures and interactions. *Philosphical Transactions of the Royal Society B*, 361:459–463, Feb. 2006.

[3] B. Blum. *Resampling Methods for Protein Structure Prediction.* PhD thesis, Electrical Engineering and Computer Sciences University of California at Berkeley, Dec. 2008.

[4] T. Higgs, B. Stantic, T. Hoque, and A. Sattar. Genetic algorithm feature-based resampling for protein structure prediction. In *IEEE World Congress on Computational Intelligence*, pages 2665–2672, 2010.

[5] T. Higgs, B. Stantic, T. Hoque, and A. Sattar. Benefits of genetic algorithm feature-based resampling for protein structure prediction. In *Bioinformatics*, 2012.

[6] D. Jones and L. McGuffin. Assembling novel protein folds from supersecondary structural fragments. *Proteins*, 53:480–485, 2003.

[7] K. Karplus, R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, and R. Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins: Structure, Function, and Genetics*, 53:491–496, 2003.

[8] V. Maiorov and G. Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology*, 235:625–634, 1994.

[9] L. McGuffin, K. Bryson, and J. D.T. The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4):404–405, 2000.

[10] K. Simons, R. Bonneau, I. Ruczinski, and D. Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, pages 171–176, 1999.

[11] K. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268:209–225, 1997.

[12] R. Unger and J. Moult. On the applicability of genetic algorithms to protein folding. In *The 26th Hawaii International Conference on System Sciences*, pages 715–725, 1993.

[13] Y. Zhang. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*, 8:108–117, 2007.

[14] Y. Zhang. I-tasser online. http://zhanglab.ccmb.med.umich.edu/I-TASSER/, 2011.

[15] Y. Zhang and J. Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *PNAS*, 101(20):7594–7599, May 2004.

[16] Y. Zhang and J. Skolnick. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical Journal*, 87:2647–2655, Oct. 2004.