Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 17 February 2017
Original Paper

Structural Bioinformatics

Sequential search leads to faster, more efficient fragment-based *de novo* protein structure prediction.

Saulo H. P. de Oliveira [1,*], Eleanor C. Law [1], Jiye Shi [2,3] and Charlotte M. Deane [1]

[1] Department of Statiscts, University of Oxford, Oxford, OX1 3LB, United Kingdom and
[2] Department of Informatics, UCB Pharma, Slough, SL1 3WE, United Kingdom and
[3] Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai, 201800, China

* To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Most current *de novo* structure prediction methods randomly sample protein conformations and thus require large amounts of computational resource. Here, we consider a sequential sampling strategy, building on ideas from recent experimental work which shows that many proteins fold cotranslationally.

Results: We have investigated whether a pseudo-greedy search approach, which begins sequentially from one of the termini, can improve the performance and accuracy of *de novo* protein structure prediction. We observed that our sequential approach converges when fewer than 20,000 decoys have been produced, fewer than commonly expected. Using our software, SAINT2, we also compared the run time and quality of models produced in a sequential fashion against a standard, non-sequential approach. Sequential prediction produces an individual decoy 1.5 to 2.5 times faster than non-sequential prediction. When considering the quality of the best model, sequential prediction led to a better model being produced for 31 out of 41 soluble protein validation cases and for 18 out of 24 transmembrane protein cases. Correct models (TM-Score > 0.5) were produced for 29 of these cases by the sequential mode and for only 22 by the non-sequential mode. Our comparison reveals that a sequential search strategy can be used to drastically reduce computational time of *de novo* protein structure prediction and improve accuracy.

Availability: Data is available for download from: http://opig.stats.ox.ac.uk/resources. SAINT2 is available for download from: https://github.com/sauloho/SAINT2

Contact: saulo.deoliveira@dtc.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A standard *de novo* protein structure prediction pipeline consists of randomly sampling the conformational space to identify minimum-energy conformations. This sampling is usually carried out via a Monte-Carlo search (Raman, S. et al , 2009); by causing perturbations to a fully-elongated protein chain and accepting/rejecting the resulting conformations based on an acceptance probability. This probability is defined in terms of a scoring function that combines physical and statistical terms. After many successive perturbations, a conformation is output. The model generation protocol is repeated via multiple independent runs to produce a large number of candidate models (decoys). This process tends to be computationally intensive; one estimate suggests that it takes

approximately 150 CPU days to accurately predict a protein's structure (Abbass, J. et al , 2015).

There has been significant effort to test different sampling strategies to improve both the efficiency and performance of *de novo* protein structure prediction. Replica Exchange Monte Carlo has been used as an extension to the traditional Monte Carlo protocol in different implementations (Kosciolek, T., and Jones, D.T. , 2014; Blaszczyk, M. et al , 2013; Xu, D., and Zhang, Y. , 2012) and it has been suggested as a more efficient sampling technique. Evolutionary algorithms have also been applied to structure prediction in order to detect multiple candidate energy minima conformations (e.g. Custodio, F. L. et al , 2014; Zhang, G. et al , 2016; Garza-Fabre, M. et al , 2016). Other search strategies include the optimisation of a multi-objective function (Olson, B., and Shehu, A. , 2014), or approaches based on molecular dynamics (Perez, A. et al , 2016).

Deviating from the random-restart strategy used in conventional protocols, search algorithms have also been implemented to extract information from decoys that have been produced to improve subsequent modelling runs (re-sampling) (e.g. Mabrouk, M. et al , 2015; Brunette, T. J., and Brock, O. , 2008; Shrestha, R., and Zhang, K. Y. , 2014). In several of its implementations, re-sampling has been shown to improve the results of the Monte Carlo search implemented in Rosetta (Simoncini, D. et al , 2012, 2017). Another perspective is to explore probabilistic frameworks such as Hidden Markov Model sub-optimal sampling (Lamiable, A. et al , 2016) and conditional sampling from a united-residue probabilistic model (Bhattacharya, D. et al , 2016). The latter was based on experimental evidence supporting the notion of foldon units. These probabilistic frameworks aim to break down the problem of folding into smaller local folding problems. Here we propose a similar reductionist effect by performing the Monte Carlo search in a sequential fashion, reducing the global folding problem to a more tractable, local conformational search.

Sequential search strategies have been previously explored. A modified version of ROSETTA (Raman, S. et al , 2009) was used to perform a comparison between predictions generated using a fully-elongated protein chain and predictions performed sequentially (Ellis, J.J. et al, 2010). Predictions performed sequentially using the modified ROSETTA were shown to be better than predictions generated non-sequentially for approximately half of the cases. Sequential protein structure prediction has also been used in the *ab initio* transmembrane protein protocol of ROSETTA (Yarov-Yarovoy, V. et al , 2006). The strategy starts with a helix in the middle of the protein, then adds further transmembrane helices randomly at either the C- or N-terminal end.

Regardless of the search strategy used to sample conformations, modelling success is highly dependent on the accuracy of the scoring function. Improvement of existing scoring potentials has been the focus of several articles published recently (Chae, M. H., et al , 2015; Ovchinnikov, S. et al , 2015; Yang, J., and Zhang, Y. , 2015; O'Meara, M. J. et al , 2015). In particular, pairwise potentials based on distance restraints inferred from co-evolution information have made consistent and accurate template-free structure prediction possible (Marks, D. S. et al , 2011; Jones, D. T. et al , 2012; Kamisetty, H. et al , 2013; de Oliveira, S.H.P. et al , 2016, e.g.), when a sufficient number of homologue sequences is available. Metagenomics from microbial DNA has been used to complement this sequence information, further broadening the applicability of such approaches (Ovchinnikov, S. et al , 2017). Contact predictions have been shown to be critical for modelling success. As co-evolution methods have only recently become a standard part of *de novo* protein structure prediction, most search strategies (including the sequential implementations of ROSETTA) have not incorporated these distance restraints in their tests. One exception is the work described in (Jones, D. T. et al , 2012), in which sequential incorporation of distance restraints let to better modelling results when the contact order was sufficiently high.

One of the main limitations that has not been addressed by contact prediction is the fact that *de novo* structure predictors still require a large amount of computational resources for accurate and consistent modelling. This relates to the large number of decoys that need to be produced during model generation and to the large number of moves performed to generate a single decoy. There is no consistency in terms of the number of decoys that need to be produced across different prediction software (S1 Table). Three recent studies using the software ROSETTA describe the use of 10,000 (Ovchinnikov, S. et al , 2015), 20,000 (Ovchinnikov, S. et al , 2017), and 20,000-900,000 (Kim, D. E. et al , 2014) decoys per target, meaning that the consensus is not clear even for the same structure predictor. Furthermore, no rationale as to how many decoys should be produced is presented in articles describing different methods, and for some cases the choice appears arbitrary.

Here, we investigate whether a sequential search heuristic could be used to improve both the efficiency and the accuracy of template-free protein structure prediction. To do so, we developed SAINT2, a completely independent fragment-assembly structure predictor. SAINT2 differs from conventional fragment-assembly approaches as it is able to perform predictions either sequentially, starting from either terminus, or non-sequentially, similar to traditional structure prediction software such as ROSETTA. Both sequential and non-sequential protocols use exactly the same parameters and input to facilitate unbiased comparison between the two modes. Given that successful *de novo* modelling is reliant on accurate contact-prediction, SAINT2 incorporates predicted protein contacts into its modelling routine.

First, we present a rationale for the number of decoys that SAINT2 needs to generate in order for a correct answer to be produced. We then compare the run time and the modelling results of SAINT2's sequential and non-sequential approaches on validation sets of 41 soluble proteins and 24 transmembrane proteins. Our results show that sequential protein structure prediction requires fewer decoys to be produced, produces individual decoys significantly faster, and is capable of consistently generating better models.

2 Methods

We have implemented a sequence-to-structure pipeline to perform *de novo* protein structure prediction (see SI Figure 1). Our pipeline takes as input a target sequence for which we generate secondary structure predictions using PSIPRED (Jones, D. T. , 1999), torsion angle predictions using SPINE-X (Faraggi, E. et al , 2009, 2012), a fragment library using Flib (de Oliveira, S.H.P. et al , 2015), and, when possible, residue-residue contact predictions using metaPSICOV (Jones, D. T. et al , 2014) as it was shown to produce the most accurate predictions (de Oliveira, S.H.P. et al , 2016) (for full details see SI). The final step in our pipeline is to generate structure predictions using SAINT2. SAINT2 requires the output files of steps one to four to generate models.

Fragment Library

Flib (de Oliveira, S.H.P. et al , 2015) is used to generate the fragment libraries for SAINT2. Flib extracts fragments from a curated database of known structures. This database is a non-redundant (sequence identity $< 90\%$), high quality (resolution < 2.5 Å) subset of the PDB (Berman, H. M. et al , 2000). HHSearch (Söding, J. , 2005) is used to identify and remove homologs to the target from this database in order to represent a realistic *de novo* structure prediction scenario. Fragments are selected from structures in this database based on the target's sequence, predicted secondary structure and predicted torsion angles. On average, Flib generates $\sim$30 fragments per target position that are 6 to 20 residues long. The same fragment library is used for a target in all three modes of SAINT2.

SAINT2

SAINT2 is based on a heuristic that treats protein structure prediction as a global optimization problem. Its energy function is a combination of different knowledge-based and physical potentials (see SI for more details). The conformational space is sampled using a library of fragments of known structures, performing successive fragment replacements on an existing peptide conformation (SI Figure 2).

SAINT2 builds models for a given target using the 3D Cartesian coordinates of the five main backbone atoms (C, N, O, C-α and C-β). These coordinates are calculated using the fragment library (see SI for more details) and completed using ideal bond lengths. SAINT2 does not consider side-chains explicitly.

Three different modes have been implemented within SAINT2: Forward, Non-sequential, and Reverse. SAINT2 Forward is initialized by selecting a fragment from the fragment library corresponding to the N-terminal residues of the target protein. In this mode, the peptide will grow as the simulation is executed. The direction of peptide extrusion is N-terminal to C-terminal. The Reverse mode is analogous to the Forward mode, but the initialisation occurs at the C-terminus. In the Reverse mode, the peptide will also grow as the simulation is executed, but the direction of peptide extrusion is reversed (C-terminal to N-terminal). SAINT2 can also perform fragment-assembly in a similar fashion to traditional approaches such as ROSETTA. We refer to this as Non-sequential structure prediction. In the Non-sequential mode, SAINT2 is initialized with a fully extruded protein conformation where the torsion angles are set to 180 degrees and ideal bond lengths and angles are used. In the analyses described in this manuscript, an identical number of moves is used for each of the three modes of SAINT2.

Model Generation

The Forward mode of SAINT2 is outlined in SI Figure 2. Here, we outline each of the stages of our model generation routine.

Fragment replacement step: fragments are selected at random from the fragment library. The probability of selecting a given fragment is proportional to the fragment score assigned by Flib, which is based on the predicted Torsion angle score (SI Figure 2).

Extrusion steps are a specific type of fragment replacement that always takes place at the end of the existing conformation, growing the peptide by one residue. For the Forward mode, an extrusion always occurs at the C-terminal end of the peptide, in which a fragment representing the C-terminal is randomly selected from the fragment library (see SI Figure 2). The extrusion replacement always adds a new residue to the existing peptide conformation. For the Reverse mode, extrusion occurs in an analogous fashion, but at the N-terminal end of the peptide. The new conformation resulting from an extrusion step is always accepted. No extrusions are performed for the Non-sequential mode, as in this mode the initial conformation already contains all the residues of the target.

Different increments of up to ten residues were tested for extrusion steps and produced comparable results. We have chosen to use an increment size of one residue. Extrusion steps use a different fragment as opposed to extending the existing fragment by one residue. However, this choice does not affect the results as extrusion steps are always accepted and a significant number of move steps is performed between extrusions.

Move steps are fragment replacements that take place at random positions in the existing peptide conformation. Unlike extrusion steps, move steps do not append new residues at the end of the sequence.

The Mover is responsible for swapping between move and extrusion steps in the Sequential and Reverse modes of SAINT2 (for more details, see SI).

Score: SAINT2 uses a combined knowledge-based and physical potential that consists of five different components: RAPDF, Lennard-Jones, solvation, predicted secondary structure, and predicted inter-residue contacts. The score is a weighted sum of each of its five components (refer to SI for more details).

Decoy Selection: SAINT2 samples the conformational space by generating thousands of decoys for each target (see Determining the Number of Decoys section). Decoys are ranked according to our combined knowledge-based and physical potential (see Score section of SI).

Data Sets

SAINT2 was trained using a set of 43 structurally diverse proteins extracted from the PDB (Berman, H. M. et al , 2000).

A full list of these proteins is given in Supplementary Table 1. These proteins are all single chain, single domain proteins proportionally distributed among the four SCOP (Murzin, A. G. et al , 1995) protein classes: all α, all β, α/β, and $\alpha+\beta$. They are also evenly spread in terms of length, ranging from 59 to 508 residues. Each of the proteins in our dataset belongs to a different Pfam family (Punta, M. et al , 2011).

Our sequential comparison analyses were carried out on two validation data sets: a soluble set of 41 structurally diverse proteins and a transmembrane set of 24 α-helical bundles extracted from the PDB (Berman, H. M. et al , 2000). A full list of these proteins is given in SI Table 2. For the soluble set, analogous to the Training data set, the proteins are all single chain, single domain proteins proportionally distributed into the four SCOP (Murzin, A. G. et al , 1995) protein classes. They are also evenly spread in terms of length, ranging from 54 to 504 residues. For the transmembrane set, a set of polytopic α-helical transmembrane chains was taken from the Orientations of Proteins in Membranes (OPM) datbase (Lomize M.A. et al , 2006). Taking only unbroken chains, the set was culled to keep no more than one member of each family, as defined by the OPM database, and also culled by PISCES (Wang G. and Dunbrack R.L. , 2003) so that the maximum sequence identity between chains was 20%. By manual inspection, chains were selected which had no soluble domain, consisted of at least four helices, and formed a single transmembrane domain. We used only the 24 shortest proteins in the resulting set, ranging from 132 to 385 residues in length. There is no overlap between the Pfam families in the training and validation sets.

CASP12 Data set

We used SAINT2 to generate models for 23 free-modelling domains from CASP12. For this comparison, we considered only free-modelling targets and included all the domains for which structural data was available on the CASP12 website.

Only sequence and structure information available before the beginning of CASP12 was used in this analysis (we excluded all structures and sequences published after April 2016 from our databases).

Validation

TM-Score (Zhang, Y. and Skolnick, J. , 2004; Xu, J., and Zhang, Y. , 2010) was used to evaluate the quality of the decoys generated by SAINT2. Three different TM-Score based measures were defined to help assess our results:

- TM-Score of the best decoy (TM-Score Best): computed by selecting the decoy from among all decoys generated with the highest TM-Score compared to the target's native structure.
- TM-Score of the Top 5 decoys (TM-Score Top-5): the TM-Score of the best decoy among the five top decoys output by our sequence-to-structure pipeline.

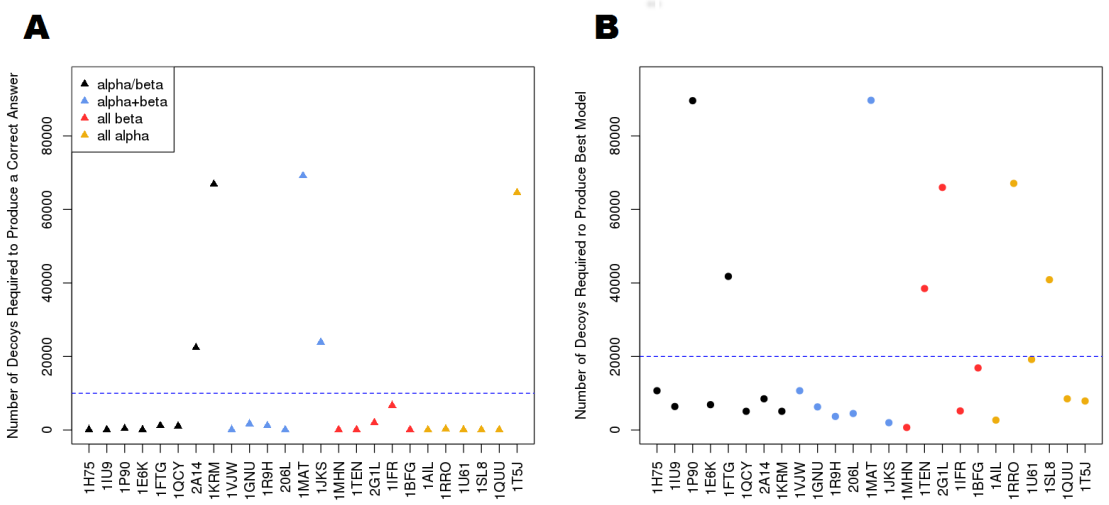

Fig. 1. Number of decoys required by SAINT2 to produce a correct answer or a "best" model. We generated 100,000 decoys for each target in the training data set and have estimated both the number of decoys required to produce a correct answer (A) and to produce a "best" model (B). A correct answer is one with TM-Score to the native structure greater than 0.5 and a "best" model is a decoy within 0.05 TM-Score units of the best possible solution produced in the 100,000 decoy ensemble (see SI Figure 3 for more details). Proteins are coloured according to their SCOP classes.

3 Results

Determining the Number of Decoys Required by SAINT2

Successful *de novo* protein structure prediction methods tend to rely on brute force approaches that generate hundreds of thousands of conformations (Kandathil, S. M. et al , 2016; Moult, J. et al , 2014). Therefore, accurate template-free modelling is heavily dependent on the availability of large computational resources. As seen in S1 Table, there is not a consensus as to the number of decoys that need to be produced across different methods or even for a single method. It is hard to draw a comparison across different predictors as some perform significantly higher numbers of moves for a single decoy and produce a smaller number of decoys.

Little analysis has been done to assess how many decoys are actually needed in order to obtain a good answer. The only common result is the suggestion that the longer the protein, the larger the number of decoys needed (Moult, J. et al , 2014; Xu, D., and Zhang, Y. , 2012; Simoncini, D. and Zhang, K. Y. , 2013; Kim, D. E. et al , 2014). Given the recent improvements obtained by incorporation of co-evolution restraints into prediction pipeline, it is possible that more efficient search heuristics could be used to reduce the number of decoys used.

SAINT2 Forward, which performs prediction sequentially starting with a fragment representing the N-terminus and gradually growing this peptide as the conformational space is sampled, was used to generate 100,000 decoys for each of the 43 proteins in our training data set. Correct answers (TM-Score to native structure > 0.5 (Xu, J., and Zhang, Y. , 2010)) were generated for 25 targets. These 25 cases were used to estimate how many decoys are necessary to obtain a correct answer and a "best" model in the 100,000 decoy ensemble (Figure 1). We define a "best" model as a decoy within 0.05 TM-Score units of the best possible solution in the 100,000 ensemble. In order to identify the number of decoys required to produce either a "best" model or a correct answer, we sampled decoys from the ensemble. A hundred samples of each size were taken and we noted the sample size (number of decoys) needed to observe at least one correct answer (SI Figure 3) or "best" model in over 95% of samples of a given size.

Our results show that when fewer than 10,000 decoys are generated, SAINT2 consistently produces a correct answer for 20 targets and a "best model" (as good as any in the 100,000 ensemble) for 14 targets (Figure 1).

We analysed whether sequence features could be used to estimate the number of decoys required to consistently produce a correct answer (Figure 2). For the proteins shorter than 250 residues where SAINT2 has produced a correct answer, this answer can consistently be produced when fewer than 10,000 decoys have been generated. For proteins longer than 250 residues, a larger number of decoys had to be output to achieve this consistency. Other than this binary behaviour, no correlation was observed between length and the calculated required number of decoys. When considering the number of loop positions, similar results were observed. Results for the SCOP classes show that SAINT2 can consistently generate a correct answer for most All α and All β proteins in our data set when fewer than 5,000 decoys are produced. However, no correct answer was produced for any All β proteins longer than 250 residues.

Sequence features were also compared against the number of decoys required to consistently produce a "best" model for each protein in our Training Data Set (SI Figure 4). We observe no correlation between our estimate for the number of decoys required to produce a "best" model and protein length or number of loop positions. Results for the SCOP classes show that $\alpha + \beta$ proteins tend to require fewer decoys compared to other SCOP classes in order for a "best" model to be generated. We also tested for a relationship between the number of decoys required to produce a correct answer or a "best" model with the fragment library precision and the precision of predicted residue-residue contacts and observed no correlations (SI Figures 5-8).

We have assessed the number of decoys that need to be generated by SAINT2 so that a correct answer or a "best" model is produced. Our results reveal that generating 10,000 decoys is sufficient to ensure that a correct answer is produced for most cases. We have also shown that the number of decoys required to produce a "best" model shows little correlation to protein length and is lower for proteins belonging to α+β SCOP classes. These results allow us to estimate the number of decoys that SAINT2 should generate for a given target.

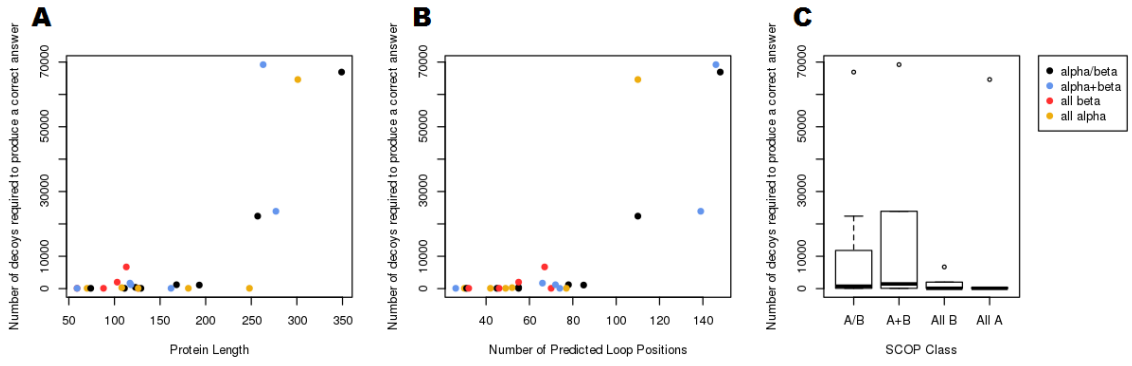

Fig. 2. Correlation between the number of decoys required by SAINT2 to produce a correct answer and three sequence-based features. The x-axis represents a feature, protein length (A), number of predicted loop positions (B), and SCOP class (C). The y-axis is the number of decoys required to generate a correct answer for the 25 targets in our Training data set where SAINT2 produced a correct answer.

Impact of Sequentiality on the Quality of Soluble Protein Models

The main concern when employing a pseudo-greedy search heuristic is local minimum entrapment, especially when considering a rugged objective function. We have, therefore, investigated the impact of sequentiality on the quality of models. We performed a comparison between SAINT2 Forward and SAINT2 Non-sequential to evaluate performance on a validation set of 41 soluble proteins.

We generated 10,000 decoys for each of the 41 targets in our soluble validation set using SAINT2 Forward and SAINT2 Non-sequential. For this comparison, identical fragment libraries, contact predictions and number of moves to generate a decoy were used. In that sense, both approaches were nearly identical. No aspect of our pipeline was designed to favour sequential prediction over its non-sequential counterpart. In fact, the knowledge-based potentials in SAINT2 were developed for a non-sequential prediction pipeline.

Decoys generated using SAINT2 Forward tend to present a higher TM-Score Best when compared to SAINT2 Non-sequential (Figure 3 - left). SAINT2 Forward produced a correct answer (Best TM-Score > 0.5) for 18 cases whereas SAINT2 Non-sequential produced a correct answer for only 13 cases. There are no cases where a correct answer was generated using the SAINT2 Non-sequential that was not also produced sequentially. SAINT2 Forward predictions were better in 10 out of the 13 cases for which SAINT2 Non-sequential generated a correct answer.

These trends are reproduced if we look at the TM-Score Top-5 (SI Figure 9). SAINT2 Forward presents a higher TM-Score Top-5 than SAINT2 Non-sequential in 33 of the 41 cases, generating 11 correct answers (TM-Score Top-5 > 0.5). SAINT2 Non-sequential produced a correct answer amongst its top five scoring decoys for only in 3 of the 41 cases.

It has previously been suggested that proteins belonging to the α/β SCOP class and longer proteins are more likely to fold in a sequential fashion (Deane, C. M. et al , 2007; Saunders, R. et al , 2011). We used our SAINT2 modes to assess the relationship between length/SCOP class and the improvements observed by performing predictions sequentially. SAINT2 Cotranslational presents a higher TM-Score Best for 9 out of 10 α/β proteins in our soluble validation set (as seen on Figure 3). Furthermore, the effect seems to be stronger (i.e. the differences between the TM-Score Best of SAINT2 Forward and Non-sequential are larger) for α/β proteins. We observe no relationship between model improvement and protein length.

We have also compared the run time of our different modes of SAINT2. The time complexity of our fragment assembly approach is quadratic on the number of atoms. Given that SAINT2 Forward performs many moves on a reduced number of atoms (a portion of the full protein chain), it can generate decoys at least 1.5 times faster than SAINT2 Non-sequential (SI Table 3).

Overall, our results show that SAINT2 Forward employs a more efficient search approach and produces better models for a majority of modelling cases. There are no cases where a conventional, non-sequential search strategy is capable of producing a model that is significantly better than the ones generated sequentially. The improvement in modelling results was observed across all protein lengths and across all SCOP classes represented in our validation set.

Impact of Sequentiality on the Quality of Transmembrane Models

We also used SAINT2 to test whether a sequential approach is a more efficient way to sample the conformational space and generate accurate decoys than a standard, non-sequential approach for transmembrane α-helical bundles. These are known to be inserted cotranslationally into the membrane, so a fragment assembly protocol which imitates this process may succeed by following the natural folding pathway.

We generated 10,000 decoys for each of the 24 targets in our transmembrane set using SAINT2 Forward and SAINT2 Non-sequential. For this comparison, identical fragment libraries, residue-residue contacts and number of moves to generate a decoy were used.

We compared the best TM-score of all decoys (TM-score Best) produced by each of SAINT2 Forward and SAINT2 Non-sequential (Figure 3). For 18 out of 24 proteins, SAINT2 Forward produces a more accurate decoy. In two cases, the improvement in TM-score Best for sequential over non-sequential is > 0.15. There are five cases where SAINT2 Forward produces a correct answer (TM-score > 0.5) and SAINT2 Non-sequential does not. Two of these correspond to the longest proteins in the set (292 and 324 residues) for which a correct answer was produced. There were just two cases where a correct answer was only produced non-sequentially.

CASP12 Results

We have also compared our sequential approach to state-of-the-art prediction software. We used SAINT2 Forward to produce 10,000 decoys for the 23 free-modelling domains from CASP12 for which structural data was available. We compared the results obtained by SAINT2 Forward against the most successful predictor in CASP12, the Baker Group (SI Figure 10). As our models are postdiction rather than prediction, this comparison should only be used to see if SAINT2 Forward can produce results of similar quality.

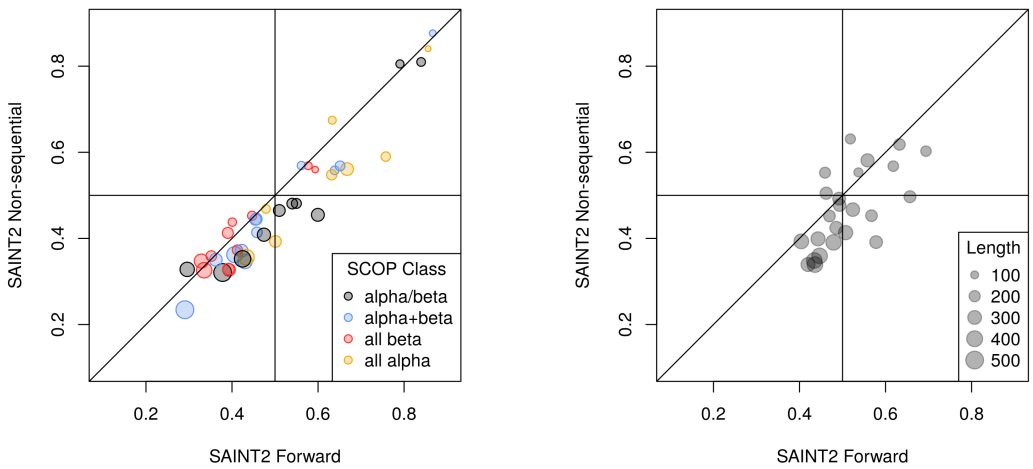

Fig. 3. Comparison of the TM-Score Best for a validation set of 41 soluble proteins (left) and 24 transmembrane proteins (right) obtained using SAINT2 Forward (x-axis) against SAINT2 Non-sequential (y-axis). Points below the diagonal indicate cases where sequential prediction performs better than non-sequential prediction. Point size indicates protein length and point colour indicates the protein SCOP class.

We compared the best model submitted by the Baker Group to CASP12 (best-of-five) to the five highest scoring models produced by SAINT2 Forward (best-of-five). Given that we currently do not have a model selection protocol for SAINT2, we have used the scoring function developed for model generation to select these (SI Figure 10A). We have also provided results for the best model produced by SAINT2 Forward (SI Figure 10B). The Baker group submitted models with correct topology (TM-Score > 0.5) for 4 out of the 23 free-modelling targets. SAINT2 Forward produced a model with correct topology for 8 targets, but only 3 of these cases were amongst the five highest scoring models. Our results show that models produced by SAINT2 Forward are of comparable quality to the state of the art.

Investigating the Effect of Directionality During Model Generation

The ROSETTA *ab initio* membrane protocol uses an incremental but bi-directional method to build decoys (Yarov-Yarovoy, V. et al , 2006). It is therefore using sequential sampling, but not in the direction of protein synthesis. We assessed how directionality may affect conformational sampling by comparing SAINT2 Forward to SAINT2 Reverse, which performs the same sequential protocol, but in the non-biological C- to N-terminal direction.

We generated 10,000 decoys for each of the 41 proteins in our soluble set and for the 24 targets in our transmembrane set using SAINT2 Forward and SAINT2 Reverse (SI Figure 11). For both modes, identical fragment libraries, residue-residue contacts and number of moves to generate a decoy were used.

We observed small differences between the TM-Score Best of models generated for the soluble set by SAINT2 Forward and SAINT2 Reverse. For the transmembrane set, little difference was observed.

4 Discussion

In this paper, we have investigated the behaviour of a sequential search heuristic for fragment-based *de novo* structure prediction. Our aim was to test whether a pseudo-greedy search strategy could reduce the computational cost of accurate *de novo* modelling.

Our initial study assessed how efficiently our sequential predictor SAINT2 can produce a model of similar quality to its best possible model. There is a general perception in the literature that hundreds of thousands of decoys are required for a correct model to be produced (Xu, D., and Zhang, Y. , 2012; Simoncini, D. and Zhang, K. Y. , 2013; Kim, D. E. et al , 2014), whereas little evidence is presented as to how many decoys are required to produce a sufficiently good answer. SAINT2 Forward, in the majority of cases where a correct answer is produced, is able to produce it when less than 10,000 decoys are generated. This suggests that SAINT2 is either more efficient at sampling the conformational space or that other methods are generating an excessive number of decoys. It is possible that the number of decoys could be further reduced by optimisation of the sequential search in terms of satisfying the distance constraints. As the number of available homologue sequences increases, so does the precision of contact prediction (de Oliveira, S.H.P. et al , 2016), which may enable greedier strategies.

Traditional structure predictors always perform moves on and score a full protein conformation. SAINT2 Forward optimizes performance by performing moves on and scoring a peptide that is shorter than the target protein. In the analyses in this manuscript, where we use the same number of moves for the two methods, SAINT2 Forward is capable of generating an individual decoy between 1.5 and 2.5 times faster than SAINT2 Non-sequential. This means that regardless of the number of decoys produced, a sequential approach can significantly reduce the computational cost of accurate structure modelling.

One possible issue with using a sequential protein structure predictor is the idea of local entrapment, that by folding the N-terminal residues before the rest of the protein, they could become trapped in a local minimum that is not relevant for the global fold. This type of entrapment does not appear to influence our methodology as, on our soluble and transmembrane validation sets, SAINT2 Forward generates more correct models and better models than SAINT2 Non-sequential. Considering a threshold of 0.02 TM-score units to establish model similarity (Li, W. et al , 2016; Kryshtafovych, A. et al , 2015), there are no cases across all soluble and membrane cases where SAINT2 Non-sequential predictions are significantly better than SAINT2 Forward.

It is arguable that the improvement in modelling results described in this work could be a consequence of the particular implementation used in SAINT2. The same implementation was used in all modes of SAINT2 (the same potentials, scoring functions, heuristics, and parameters were used). Furthermore, we have used potentials that have been developed, trained, and used in non-sequential protocols. Sequentiality had also been previously explored in transmembrane protein structure prediction by ROSETTA (Yarov-Yarovoy, V. et al , 2006). Therefore, it seems that the improvement observed for sequential predictions is unlikely to be a consequence of our implementation.

We carried out a comparison of SAINT2 Forward's performance on the CASP12 targets in order to establish whether its results are equivalent to those of the best performers within CASP. However, as our models are postdiction (though every care was taken to remove information available post CASP12 - see Methods) we see these results as indicative rather than definitive. Models with a TM-score above 0.5 may not be useful for a large number of biological applications. Nonetheless, results from the most recent iteration of CASP show that, in the absence of a reliable template, protein structure predictors rarely achieve TM-Scores greater than 0.8. The most successful template-free predictor (Baker Group) produced a model with a TM-Score greater than 0.8 for only one free-modelling domain in our CASP12 data set. SAINT2 was able to produce models with TM-Score greater than 0.8 for approximately 10% of its soluble targets (4 out of 41). However, no model of this quality was produced for the targets in our transmembrane and CASP12 sets.

Currently, our structure prediction pipeline does not have a model selection protocol implemented. Therefore, for our CASP12 comparison, we considered both the best-of-five (as selected by SAINT2 score, see Results) and the best model out of all 10,000 decoys generated by SAINT2 Forward. Our score has not been optimised for ranking and this approach is unlikely to outperform any clustering selection protocol (Kryshtafovych, A. et al , 2015). To make a fairer comparison, it would be ideal to replicate the Baker Group's decoy selection protocol, but unfortunately their method is not reproducible due to use of human intervention. When considering the best-of-five models output by SAINT2, we were able to predict the correct topology for three cases, one fewer than the Baker Group. When considering the best model output by SAINT2, the correct topology was predicted for eight of these cases. However, we do not know how many cases had at least one model with correct topology across all the decoys produced by the Baker group as this data is not publicly available. Even if we were to consider the best model produced by the Baker Group, the number of decoys produced by their protocol is in the order of hundreds of thousands which far exceeds the 10,000 decoys produced by SAINT2. It may be that a comparison using identical computational time for SAINT2 as that used by the Baker group in CASP12 would be most appropriate.

The results in SI Figure 10 establish that SAINT2 Forward is capable of producing models of comparable quality to those produced by the state of the art. Our findings highlight the importance of re-evaluating search strategies with the advent of increasingly more accurate scoring functions.

The way by which predicted contacts are introduced during sampling has an impact on which conformations are sampled. For instance, it has been suggested that using only short-range contacts during the earlier stages of sampling can lead to modelling improvements for some proteins (Kosciolek, T., and Jones, D.T. , 2014). Due to the sequential nature of our algorithm, N-terminal contacts are introduced earlier and more moves can be dedicated to satisfying these constraints than for the C-terminal contacts. Our approach paves the way for considering different ways in which predicted contacts can be incorporated into structure prediction protocols.

Existing structure prediction software can, at times, produce correct models without the use of predicted contacts. We assessed the role of these contacts in the quality of modelling as performed by SAINT2 Forward by testing the protocol without predicted contacts (SI Figure 12). We find that correct models were produced by SAINT2 Forward without contacts for only 10 of the 18 cases where a correct model was produced by SAINT2 Forward with contacts. These results highlight the importance of accurate contact prediction for successful modelling (de Oliveira, S.H.P. et al , 2017).

We observed comparable results when predictions were generated in a biological direction and its reverse. This is consistent with the notion that protein folding is a series of small optimisation problems where segments of the chain fold independently (foldons) and then collapse to the complete structure (Maity, H. et al , 2005; Hu, W. et al , 2013). Given the amount of experimental evidence to support the notion that proteins are folding as they are being translated (Fedorov, A.N. and Baldwin, T.O. , 1997; Basharov, M. , 2000; Kolb, V. , 2001; Giglione, C. et al , 2009; Holtkamp W., et al , 2015; Puglisi, J.D. , 2015), we have opted to maintain the biological direction as the standard approach in SAINT2

We have demonstrated the validity and applicability of a sequential, pseudo-greedy search heuristic to perform *de novo* model generation. When drawing an unbiased comparison, sequential prediction requires fewer decoys to produce good answers, can generate individual decoys faster, and improves the overall modelling results.

Acknowledgements

The authors would like to acknowledge the Oxford Protein Informatics Group for their contribution and interesting discussions.

Funding

SHPO was funded by: Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPQ) - Ciencia Sem Fronteiras - Grant Number: 237656/2012-4 (http://www.cienciasemfronteiras.gov.br/web/csf). SHPO, ECL and CMD were funded by: Engineering and Physical Sciences Research Council grant to System Approaches to Biomedical Sciences-Centre for Doctoral Training (http://www.epsrc.ac.uk/) Grant Number: EP/G037280/1. These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. UCB Pharma provided support in the form of salaries for authors [JS], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Abbass, J., and Nebel, J. C., (2015), Customised fragments libraries for protein structure prediction based on structural class annotations, *BMC bioinformatics*, **16(1)**, 136.

Basharov, M. (2000). Cotranslational folding of proteins. *Biochemistry (Moscow)*, 65(12):1380–1384.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

Bhattacharya, D., Cao, R., and Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *B*ioinformatics, **32** (18), 2791-2799.

Blaszczyk, M., Jamroz, M., Kmiecik, S., and Kolinski, A. (2013). CABS-fold: server for the de novo and consensus-based prediction of protein structure. *N*ucleic acids research, **41**(W1), W406-W411.

Brunette, T. J., and Brock, O. (2008). Guiding conformation space search with an all-atom energy potential. *P*roteins: Structure, Function, and Bioinformatics,**73**(4), 958-972.

Chae, M. H., Krull, F., and Knapp, E. W. (2015). Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, **83(5)**, 881-890.

Custodio, F. L., Barbosa, H. J., and Dardenne, L. E. (2014). A multiple minima genetic algorithm for protein structure prediction. *A*pplied Soft Computing, **15**, 88-99.

de Oliveira, S.H.P., Shi, J., and Deane, C. M. (2015). Building a better fragment library for de novo protein structure prediction. *PloS one*, **10**(4), e0123998.

de Oliveira, S.H.P., Shi, J., and Deane, C. M. (2016). Comparing co-evolution methods and their application to template-free protein structure prediction *Bioinformatics*, btw618.

de Oliveira, S.H.P., and Deane, C. M. (2017). Co-evolution techniques are reshaping the way we do structural bioinformatics. *F1000Research*, **6**.

Deane, C. M., Dong, M., Huard, F. P., Lance, B. K., and Wood, G. R. (2007). Cotranslational protein folding – fact or fiction? *Bioinformatics*, 23(13):i142–i148.

Ellis, J.J., Huard, F.P., Deane, C.M., Srivastava, S., and Wood, G.R. (2010). Directionality in protein fold prediction. *BMC bioinformatics*, **11(1)**:172.

Faraggi, E., Yang, Y., Zhang, S., and Zhou, Y. (2009). Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure*, 17(11):1515–1527.

Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., and Zhou, Y. (2012). 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.

Fedorov, A.N. and Baldwin, T.O. (1997). Cotranslational protein folding. *Journal of Biological Chemistry*, 272(52):32715–32718.

Garza-Fabre, M., Kandathil, S. M., Handl, J., Knowles, J., and Lovell, S. C. (2016). Generating, Maintaining, and Exploiting Diversity in a Memetic Algorithm for Protein Structure Prediction. *E*volutionary computation, **24**(4), 577-607.

Giglione, C., Fieulaine, S., and Meinnel, T. (2009). Cotranslational processing mechanisms: towards a dynamic 3d model. *Trends in biochemical sciences*, 34(8):417–426.

Holtkamp W., Kokic G., Jäger M., Mittelstaet J., Komar A.A., and Rodnina M. V. Cotranslational protein folding on the ribosome monitored in real time. *Science*, 350(6264):1104–1107, 2015.

Hu, W., Walters, B. T., Kan, Z. Y., Mayne, L., Rosen, L. E., Marqusee, S., and Englander, S. W. (2013). Stepwise protein folding at near amino acid resolution by hydrogen exchange and mass spectrometry. *Proceedings of the National Academy of Sciences*, **110(19)**, 7684-7689.

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

Jones, D. T., Buchan, D. W., Cozzetto, D., and Pontil, M. (2012). Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**(2), 184–190.

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2014). Metapsicov: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, page btu791.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, **110**(39), 15674–15679.

Kandathil, S. M., Handl, J., and Lovell, S. C. (2016). Toward a detailed understanding of search trajectories in fragment assembly approaches to protein structure prediction. *Proteins: Structure, Function, and Bioinformatics.*, **84**, 411-426.

Kim, D. E., DiMaio, F., Yu-Ruei Wang, R., Song, Y., and Baker, D. (2014). One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, **82**(S2), 208–218.

Kolb, V. (2001). Cotranslational protein folding. *Molecular Biology*, 35(4):584–590.

Kosciolek, T., and Jones, D.T. (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PloS one*, 9(3):e92197, 2014.

Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2015). Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. *Proteins: Structure, Function, and Bioinformatics.*, 84(Suppl 1):349-369, 2016.

Lamiable, A., Thevenet, P., and Tufféry, P. (2016). A critical assessment of hidden markov model sub-optimal sampling strategies applied to the generation of peptide 3D models. *J*ournal of Computational Chemistry, **37**(21), 2006-2016.

Li, W., Schaeffer, R.D., Otwinowski, Z., and Grishin, N.V. (2016) Estimation of Uncertainties in the Global Distance Test (GDT_TS) for CASP Models. *PloS one*, 11(5),e0154786, 2016.

Lomize M.A., Lomize A.L., Pogozheva I.D., and Mosberg H.I. (2006). Opm: orientations of proteins in membranes database. *Bioinformatics*, 22(5):623–625, 2006.

Mabrouk, M., Putz, I., Werner, T., Schneider, M., Neeb, M., Bartels, P., and Brock, O. (2015). RBO Aleph: leveraging novel information sources for protein structure prediction. *N*ucleic acids research, **43** (W1): W343-W348.

Maity, H., Maity, M., Krishna, M. M., Mayne, L., and Englander, S. W. (2005). Protein folding: the stepwise assembly of foldon units. *Proceedings of the National Academy of Sciences of the United States of America*, **102(13)**, 4741-4746.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PLoS one*, **6**(12), e28766.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (casp) - round x. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):1–6.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540.

Olson, B., and Shehu, A. (2014, March). Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction. *I*ntl Conf on Bioinf and Comp Biol (BICoB), Las Vegas, NV..

Puglisi J.D. (2015) The delicate dance of translation and folding. *Science*, 348(6233):399–400, 2015.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al, (2011), The pfam protein families database, *Nucleic acids research*, **40**, D290-D301.

O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., DiMaio, F., and Kuhlman, B. (2015). Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *Journal of chemical theory and computation*, **11(2)**, 609-622.

Ovchinnikov, S., Kim, D. E., Wang, R. Y. R., Liu, Y., DiMaio, F., and Baker, D. (2015). Improved de novo structure prediction in CASP11 by incorporating Co-evolution information into rosetta. *Proteins: Structure, Function, and Bioinformatics.*, **84**, 67-75.

Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D. E., Kamisetty, H., Grishin, N.V. and Baker, D. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *E*life, **4**, e09248.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P., Pavlopoulos, P.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C., Baker, D., (2017) Protein structure determination using metagenome sequence data,*Science*, **355**, 294-298.

Perez, A., Morrone, J. A., Brini, E., MacCallum, J. L., and Dill, K. A. (2016). Blind protein structure prediction using accelerated free-energy simulations. *S*cience advances, **2**(11), e1601274.

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

Saunders, R., Mann, M., and Deane, C. M. (2011). Signatures of co-translational folding. *Biotechnology journal*, 6(6):742–751.

Shrestha, R., and Zhang, K. Y. (2014). Improving fragment quality for de novo structure prediction. *P*roteins: Structure, Function, and Bioinformatics, **82**(9), 2240-2252.

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

Simoncini, D., Schiex, T., and Zhang, K. Y. (2017). Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. *P*roteins: Structure, Function, and Bioinformatics.

Simoncini, D. and Zhang, K. Y. (2013). Efficient sampling in fragment-based protein structure prediction using an estimation of distribution algorithm. *PloS one*, 8(7).

Söding, J. (2005) Protein homology detection by HMM-HMM comparison, *Bioinformatics*, **21**, 951-960. doi:10.1093/bioinformatics/bti125pmid:15531603

Wang G. and Dunbrack R.L. (2003) Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

Xu, D., and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *P*roteins: Structure, Function, and Bioinformatics, **80**(7), 1715-1735.

Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895.

Yang, J., and Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic acids research*, **43(W1)**, W174-W181.

Yarov-Yarovoy, V., Schonbrun, J., and Baker, D., (2006) Multipass membrane protein structure prediction using rosetta. *Proteins: Structure, Function, and Bioinformatics*, **62(4)**:1010–1025.

Zhang, G., Zhou, X. G., Yu, X. F., Hao, X. H., and Yu, L. (2016). Enhancing Protein Conformational Space Sampling Using Distance Profile-Guided Differential Evolution. *I*EEE/ACM Transactions on Computational Biology and Bioinformatics..

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710.