Constraint Guided Beta-Sheet Refinement for Protein Structure Prediction

Constraint Guided Beta-Sheet Refinement for Protein Structure Prediction

M A Hakim Newton[a,d], Rianon Zaman[b,d], Fereshteh Mataeimoghadam[b], Julia Rahman[b], Abdul Sattar[b,c]

[a] *School of Information and Physical Sciences, The University of Newcastle, Australia*
[b] *School of Information and Communication Technology, Griffith University, Australia*
[c] *Institute for Integrated and Intelligent Systems, Griffith University, Australia*
[d] *These two authors have contributed equally to this paper and are joint first-authors*

Abstract

Protein structure prediction (PSP) is a crucial issue in Bioinformatics. PSP has its important use in many vital research areas that include drug discovery. One of the important intermediate steps in PSP is predicting a protein's beta-sheet structures. Because of non-local interactions among numerous irregular areas in beta-sheets, their highly accurate prediction is challenging. The challenge is compounded when a given protein's structure has a large number of beta-sheets. In this paper, we specifically refine the beta-sheets of a protein structure by using a local search method. Then, we use another local search method to refine the full structure. Our search methods analyse residue-residue distance-based scores and apply geometric restrictions gained from deep learning models. Moreover, our search methods recognise the regions of the current conformations prompting the nether scores and generate neighbouring conformations focusing on that identified regions and making alterations there. On a set of standard 88 proteins of various sizes between 46 and 450 residues, our method successfully outperforms state-of-the-art PSP search algorithms. The improvements are more than 12% in average root mean squared distance (RMSD), template modeling score (TM-score), and global distance test (GDT) values.

Keywords: Protein Structure Prediction, Backbone Angles, Machine Learning, Search-Based Optimisation, Neighbour Generation

1. Introduction

Protein structure prediction (PSP) is a very demanding and challenging problem [1]. Proteins are made up of amino acid (AA) sequences and fold into three-dimensional native structures. The native structure of a protein has the lowest minimum Gibb's free energy and as per Anfinsen's dogma [2] is largely determined by the AA sequence. There are 20 types of AAs that occur in any order in any protein, subject to stoichiometric constraints [3]. Amino acids have N, C^{α}, C, O, C^{β}, and other atoms. Instances of AAs in proteins are called *residues*. A peptide bond is created when the C atom of one residue connects to the N atom of the following residue. As a result, we achieve

Email addresses: mahakim.newton@newcastle.edu.au (M A Hakim Newton), rianon.zaman@griffithuni.edu.au (Rianon Zaman), fereshteh.mataeimoghadam@griffithuni.edu.au (Fereshteh Mataeimoghadam), julia.rahman@griffithuni.edu.au (Julia Rahman), a.sattar@griffith.edu.au (Abdul Sattar)

the main chain of a protein. Aside from the protein's main chain, there is a side chain for every AA (except Glycine). The side chain begins from the C^{α} atom and the first atom in this chain is C^{β}. Usually three dihedral angles ϕ, ψ, and ω are used in representing the main chain of a protein assuming conventional bond distances and bond angles. Every four consecutive atoms in the order C_{i-1}, N_i, C_i^{α}, C_i, N_{i+1}, C_{i+1}^{α} define these three angles. However, ω is 180° [4], for most proteins, while ϕ and ψ are any values between $-180°$ to $+180°$. There are dihedral angles also for every individual AA side chains, but we are primarily interested in the main chain's ϕ and ψ angles.

Protein structures exhibit three types of local structures called secondary structures (SS) while the amino acid residue sequence is called the primary structure and the three dimensional structures are called the tertiary structures. The three SS types are alpha-helices, beta-sheets, and coils or loops. Coils are the flexible ones while the helices and sheets are rigid. PSP methods often try to build the rigid structures first by using machine learning approaches before going into consideration of the global characteristic of the protein structure and thus fixing the coil areas [5]. However, between the rigid structures, sheets are more difficult to predict than helices [6, 7]. Nevertheless, correct prediction of beta-sheet is very important in reducing the PSP search space. More accurate beta-sheet prediction has further important implications since beta-sheet inter-linkages have been linked to the generation of protein accession in a variety of human disorders that include cancer and Alzheimer's disease [8]. So, in this work, we focus on improving prediction of beta-sheet structures. Note that beta-sheets are made up of strand pairs that are kept together in parallel and anti-parallel configurations by beta-residue contact mapping [9]. The major hindrance is in taking into account the long-range intercommunication between abandoned beta-strands that are consecutively distant yet geographically adjacent in the structure [10]. Another major challenge with the prediction of beta-sheet structures is the large number of potential sheet structures, which makes determining the precise arrangement of strands inside sheets more difficult [11].

Effective approaches for predicting the beta-sheet structures are needed to address these challenges. To record beta-residue interactions and to enhance beta-sheet structure predictions, BetaDL [12] uses deep learning models. BetaDL also uses a graph-based approach to model the conformational space of beta-sheets and a scoring function to evaluate beta-sheets. BetaTop [11] rapidly searches the conformational space with the purpose of lowering the problem's time complexity. To simulate the search space, BetaTop uses two tree structures named sheet-tree and grouping-tree. In reality, BetaTop divides the problem down into small problems before proposing a dynamic programming technique for determining the best conformation. This enables BetaTop to save intermediate findings and reuse them instead of doing brute-force calculations. Moreover, BetaTop builds the trees in a manner that eliminates many nodes that will not be repeated at higher levels, reducing the problem's space needs. In this paper, we use a constraint-guided local search approach to refine beta sheet structures constructed by using deep learning predictions.

Overall, PSP has an astronomically large search space in terms of possible fractional values for ϕ and ψ. Various methods have been proposed for PSP search so far [13, 14]. These consist of evolutionary algorithms [15, 16, 17], Monte Carlo algorithms [18, 19], multi-objective optimisation [20, 21], resampling [22], sequential search [23], differential evolution [24, 25, 26, 27, 28, 29], and memetic algorithms [30, 31, 32]. A hybrid search framework embedding a tabu-based local search within a population based genetic algorithm has been proposed by [33]. Again, a random-walk based stagnation recovery approach has been proposed [34]. Constraint-based approaches have been used in PSP in the past but in simplified PSP [35, 36]. Fragment assembly techniques are popular in conformational search [37, 18, 38, 29]. These techniques use fragments from known protein

structures to explore conformational search space and thus can be restricted by the limitation of the fragment library particularly in dealing with flexible loop regions or coils in the proteins. Loop sampling methods [39, 40, 41, 42, 43, 44, 45, 46] have been used to overcome the loop related issues. Among recent methods, AlphaFold [47] uses a gradient descent method with deep learning predicted inter-residue distances. Later, AlphaFold2 [48] made significant progress in PSP using end-to-end deep learning models. However, AlphaFold2 not only does require a lot of computing power, but it also trains on practically all known proteins. The algorithmic intricacies of the system are currently not very well known. AlphaFold2 has extremely restricted access through the Google Collab interface. As a result, the PSP researchers face the problem of obtaining alphafold precision utilising a simplified and more effective PSP approach. Among further PSP methods, DeepAccNet [49] uses ROSETTA search method along with its deep learning based estimation of signed errors in residue-residue distances. A recent PSP method CGNP [5] proposes an intelligent neighbour selection approach for structure refinement. However, like other methods, CGNP focuses on refining the coil residues only. In this work, we use a search framework similar to CGNP. Besides using CGNP type approach to refine coil areas, we do also use constraint guided search to refine beta sheet structures. In search based PSP methods, scoring functions are needed to measure the quality of the conformations. While previous PSP methods used physical, chemical, and force-field based equations as scoring functions (e.g. CHARMM [50] and ROSETTA [51]), recent PSP methods use residue-residue *distance maps* and *contact maps* (two residues are in contact if their distance is within 8Å). In distance and contact maps, residues are usually represented by C^{β}, except for Glycine, which is represented by C^{α}. CONFOLD [52, 53], MULTICOM [54], and CGLFOLD [29] are recent contact map based PSP search methods. On the other hand, RaptorX [55, 56, 57] and AlphaFold [47] are among the recent methods that use distance map based scoring functions. In this work, we use distance map based scoring functions to refine coil areas as well as beta sheet structures. Overall, from artificial intelligence perspectives, generating neighbour conformations based on informed decisions is one of our key focus in this work. As such we first focus on the beta-sheets to refine them separately in isolation using constraint guided search. Then, we move to the refinement of whole structure focusing on the coil areas. In more abstract sense, we identify troublesome parts of the current conformation in the beta-sheet region first and make improvements primarily in those areas. Then we start improving the problematic parts of the coil areas. We employ constraint-guided techniques to find the problematic regions and these techniques aid in the analysis of unfulfilled geometric constraints that result in lower scores for the existing conformations. Our method is straightforward and it can largely explain the selections made in the neighbourhood generation procedure. On a set of standard 88 proteins having beta sheets and of various sizes between 46 and 450 residues, our method successfully outperforms state-of-the-art PSP search algorithms. The improvements are more than 12% in average root mean squared distance (RMSD), template modeling score (TM-score), and global distance test (GDT) values.

The rest of the paper details our methodology and implementation, and presents our experimental results and conclusions.

2. Methodology

Figure 1 shows the main steps in our proposed PSP workflow. We start from an initial conformation. The initial conformation is obtained by using predicted dihedral angles of the main chains of the given protein. In the next step, using predicted secondary structures, we explant each beta sheet structure out of the conformation and refine it in isolation using a local search method. The

local search method takes into account the predicted inter-residue distances among the residues that are in the respective beta sheet and makes changes to the dihedral angles of the beta sheet residues. Once a beta sheet refinement is done, we implant it back to the conformation. When all the beta sheets are refined in this way, another local search method works on the whole structure. The local search method, in this case, makes changes to the dihedral angles of the coil residues taking into account the predicted inter-residue distance among coil residues and also among residue pairs with one residue from one rigid structure (helix or sheet) and another residue from another rigid structure (helix or sheet). In both cases, the local search algorithms follow the constraint guided approach used in CGNP [5]. We name our proposed PSP method as constraint guided beta sheet refinement for PSP (CGSR). While CGSR refines beta sheets in isolation CGNP does not have any beta sheet refinement steps. We describe further details of CGSR along with a brief description of CGNP's search method and the machine learning methods that we use to predict main-chain dihedral angles, secondary structures, and inter-residue distances.

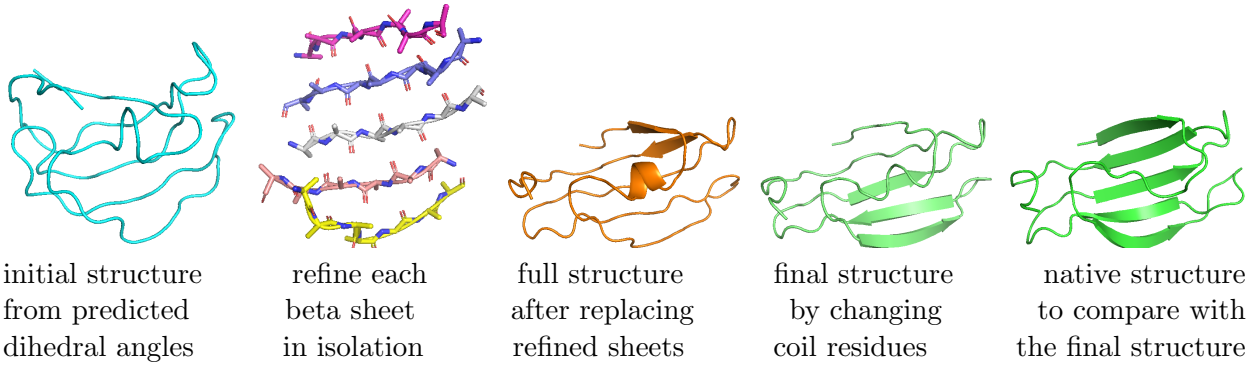


Figure 1: Steps in our proposed approach from left to right while the second and third steps are not in CGNP [5]

2.1. Employing Machine Learning Algorithms

We describe existing state-of-the-art machine learning algorithms that we use to predict main-chain angles, distance maps, and secondary structures of a given protein.

Main-Chain Angle Prediction. Recent main-chain angle prediction methods are SAP [58], OPUS-TASS [59], and SPOT-1D [60]. In this work, after some preliminary experiments, we use OPUS-TASS, which predicts ϕ and ψ angles with mean-absolute errors (MAE) of about $\phi_{\mathsf{MAE}} = 16.28$ and $\psi_{\mathsf{MAE}} = 21.98$ degrees respectively. OPUS-TASS uses convolutional Neural Network module, a bidirectional Long-Short-Term-Memory module, a modified Transformer module and a fully connection module. OPUS-TASS has used more than 10000 proteins in its training set.

Secondary Structure Prediction. Three main secondary structure (SS) types are helices, sheets, and coils, but considering their variants, there are 8 such types [61]. Helices and sheets are rigid and most residues in them have certain ranges of about length 20 degrees for their ϕ and ψ values. Coils or loops are flexible and residues in coils have any values from $-180°$ to $180°$ for their ϕ and ψ angles. We use SSpro8 [62] to predict SS types. SSpro8 uses an ensemble of Bidirectional Recursive Neural Networks (BRNNs) trained on 5772 proteins which are found before August 20, 2013 in the Protein Data Bank (PDB). SSpro8 acquires 92% accuracy on proteins for which homologs can be found in the PDB and 79% accuracy for proteins with no homologs.

Distance Map Prediction. A real-valued distance predictor SDP [63] predicts both short and long-range distances among inter-residue distances of a target protein. SDP uses 9055 training proteins that are subsets of the datasets of MapPred [64] and SPOT-1D [60]. We also use inter-residue distances predicted by trRosettaX [65], which uses 15,051 non-redundant protein chains. We run separate experiments to show that our CGSR method does not depend on the inter-residue distance prediction method and also for comparison with state-of-the-art PSP methods in terms of tertiary structures produced, we use the predicted distances that give us the best performance.

Algorithm 1 Constraint Guided Sheet Refinement for Protein Structure Prediction

function CGNP(c, cond) // see [5] for details
  while not termination criterion **do**
    select residues i, j, k from c where condition cond on i, j, k holds and
      the distance between i and j has the maximum deviation from
      the predicted values and tabu conditions on i, j, k are not violated
    generate a number of neighbour conformations changing the dihedral angles of k,
      compute the evaluation scores for each generated conformation,
      and select the best conformation as c for the next iteration
  end while
end function
function CGSR // use CGNP to refine beta sheets and the whole structure
  $c \leftarrow$ construct initial conformation from predicted ϕ and ψ values
  for $l = 1$ to the number of beta sheets in conformation c **do**
    $b \leftarrow$ explant ith beta sheet out from conformation c
    $b \leftarrow$ CGNP(b, select sheet residues for i, j, k during search)
    $c \leftarrow$ implant beta sheet b in its position in conformation c
  end for
  $c \leftarrow$ CGNP(c, selection condition cond on i, j, k)
    where cond is k is a coil residue between i and j,
      and both i and j cannot be from the same sheet or helix
  return the best 5 conformations explored by CGNP
end function

2.2. Proposed Optimisation Search Algorithm

As mentioned before, we develop our CGSR method on top of the local search algorithm used in CGNP [5]. Below we review CGNP first and then describe how we adapt it for CGSR.

Original CGNP Method. Protein conformations in CGNP are represented by the main-chain dihedral angles. The initial conformation is obtained by using the main-chain dihedral angles predicted by machine learning algorithms employed. The search algorithm changes the dihedral angles of the coil residues only. The dihedral angles of the helix and sheet residues, after initialisation, do not undergo changes by the search algorithm. The resultant conformations after every potential or actual change are evaluated using scoring functions that involve inter-residue distances. The CGNP method uses a constraint guided neighbour generation process in which a coil residue k is selected from the coil residues that are in between two other residues i and j with their distances not satisfying the predicted distance constraint. Residues i and j could be coil residues or could be

helix or sheet residues but with the condition that both i and j cannot be from the same helix or sheet. From a current conformation, CGNP then generates a number of neighbour conformation by changing the dihedral angles of the residue k and select the best neighbour in terms of the scoring functions as the next current conformation. To avoid revisitation of the same k or the same (i, j) pair within close temporal proximity, CGNP uses the tabu metaheuristic [66].

Adapting CGNP for CGSR. Algorithm 1 shows the pseudocode for CGSR. In CGSR algorithm, the original CGNP method is represented as a function CGNP with two parameters: a conformation and a condition to select residues i, j, k during neighbouring conformation generation. When the CGNP function is called in function CGSR for each isolated beta sheet, the beta sheet is actually considered like an independent conformation in which all residues belong to the beta sheet. Then when the CGNP function is called for the entire conformation, it is exactly like the original CGNP method [5]. The CGSR function returns the best 5 conformations explored in its entire execution time. The CGSR implementation is done on a recent python-based PSP platform named Koala that borrows features from a constraint-based local search system named Kangaroo [67].

3. Experiments

To evaluate CGSR, we use 88 proteins having 42 to 450 residues. These include 32 β type and 56 α/β type proteins. The β type proteins have beta sheets and coils only while the α/β type proteins have alpha helices, beta sheets, and coils. We do not use α type proteins that have only alpha helices and coils since our contribution in this work is on beta sheet refinement. Nevertheless, the 88 proteins used in our experiments have been used in existing PSP optimisation search algorithms such as CGNP [5], QUARK [19], MODE-K [28], and MODCSA/CA [68] or a machine learning algorithm such as SPOT-1D [60]. We have also used CASP13 and CAMEO 144 standard dataset. Overall, with the variations in lengths and types, these proteins represent a benchmark dataset, that could be used in evaluating search-based PSP optimisation methods. Note that all of our evaluation proteins have been checked to ensure that they have less than 25% sequence similarity with the training proteins of the machine learning algorithms employed in our PSP pipeline. As mentioned before, we use OPSUS-TASS [59] to predict main-chain angles, SSpro8 [62] to predict secondary structures, and SDP [63] and trRosettaX [65] to predict inter-residue distances.

We compare CGSR with CGNP [5] and trRosettaX [65]. Each of CGSR, SDP and trRosettaX is separately run with inter-residue distances predicted by SDP [63] and trRosettaX [65]. So we have in total 6 versions of the algorithms. For convenience, we denote trRosettaX by TRX and each solver version by X-Y where X is a search algorithm in {CGSR, CGNP, TRX} and Y is a distance predictor in {TRX,SDP}. We run each algorithm version five times on each protein and from each run we select the best five conformations based on the scoring functions used in the respective algorithm. The generation and evaluation of 320,000 conformations were used as the termination condition for each run of each algorithm. Nevertheless, from the 5 runs and 5 conformations from each run, we get 25 conformations in total for each protein and each algorithm version. We then compute RMSD, TM-score and GDT values for each conformation and then compute mean values over the 25 values. Note that lower RMSD values, higher TM-scores, and higher GDT values denote better performances. For TM-scores and GDT values, we use a 0-1 scale.

The following tables in the appendix shows our detailed results for the 6 algorithm versions.

Table A1 : Mean RMSD values obtained for β type proteins.

Table A2 : Mean RMSD values obtained for α/β type proteins.

Table A3 : Mean TM-socre values obtained for β type proteins.

Table A4 : Mean TM-score values obtained for α/β type proteins.

Table A5 : Mean GDT values obtained for β type proteins.

Table A6 : Mean GDT values obtained for α/β type proteins.

We show summarised results in Table 1. We see that CGSR-TRX, CGSR-SDP, and TRX-TRX are respectively the best, second best and third best performing algorithm versions. Moreover, both CGSR algorithm versions outperform both CGNP algorithm versions. This effectively shows performance difference contributed by the beta sheet refinement. Between distance predictors TRX and SDP, TRX appears to be resulting into better performance for all three search algorithms. The reason could be that SDP predicts larger distances better but all three algorithms mainly use smaller distances in which TRX predicts better. Among β and α/β type proteins, all algorithm versions appear to be performing better in α/β proteins than in β proteins. This again shows the difficulty in more accurate prediction of sheet structures over that of helix structures.

Table 1: Average RMSD, TM-score, and GDT values over β and α/β proteins as obtained by various solver versions. The best and the second best values are respectively emboldened and underlined.

Score	Proteins	CGSR-TRX	CGNP-TRX	TRX-TRX	CGSR-SDP	CGNP-SDP	TRX-SDP
RMSD	β	**10.89**	15.56	12.31	<u>11.35</u>	15.15	15.56
	α/β	**9.92**	13.49	11.15	<u>10.21</u>	14.78	13.80
TM-score	β	**0.47**	0.32	0.40	<u>0.44</u>	0.29	0.32
	α/β	**0.50**	0.32	0.42	<u>0.44</u>	0.29	0.32
GDT	β	**0.40**	0.29	0.37	<u>0.39</u>	0.26	0.26
	α/β	**0.44**	0.29	0.38	<u>0.40</u>	0.26	0.27

Figures 2, 3, and 4 show numbers of proteins with final conformations having mean RMSD, TM-score, and GDT values below or above certain thresholds. From the figures, CGSR-TRX and CGSR-SDP appear to be close although CGSR-TRX has slightly better performance. Basically, results in these figures are fully consistent with that presented in the summary in Table 1.

Table 2: Nemenyi test p values for various algorithm versions with $p \geq 0.05$ are emboldened.

	CGNP-TRX	TRX-TRX	CGSR-SDP	CGNP-SDP	TRX-SDP
CGSR-TRX	0.00	0.00	**0.90**	0.00	0.00
CGNP-TRX		0.00	0.00	**0.73**	**0.45**
TRX-TRX			0.00	0.00	0.00
CGSR-SDP				0.00	0.00
CGNP-SDP					**0.90**
CGSR-TRX	0.00	0.00	0.00	0.03	0.00
CGNP-TRX		0.00	0.00	**0.40**	**0.90**
TRX-TRX			0.02	0.00	0.00
CGSR-SDP				0.00	0.00
CGNP-SDP					**0.67**
CGSR-TRX	0.00	0.00	**0.16**	0.00	0.00
CGNP-TRX		0.00	0.00	**0.07**	**0.18**
TRX-TRX			0.03	0.00	0.00
CGSR-SDP				0.00	0.00
CGNP-SDP					**0.90**

Table 2 shows the p values obtained by Nemenyi test and we consider 95% confidence level. We see that the differences in RMSD and GDT between CGSR-TRX and CGSR-SDP are not

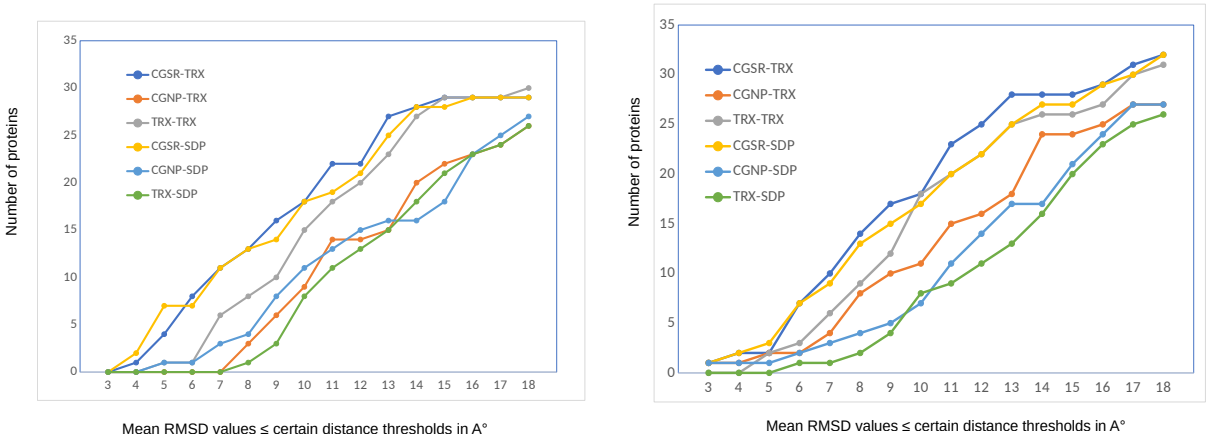


Figure 2: Performance comparison of solver versions in numbers of proteins (y-axis) with final conformations having mean RMSD values $\leq$ certain distance thresholds in Å (x-axis) for β type (left) and α/β (right) type proteins

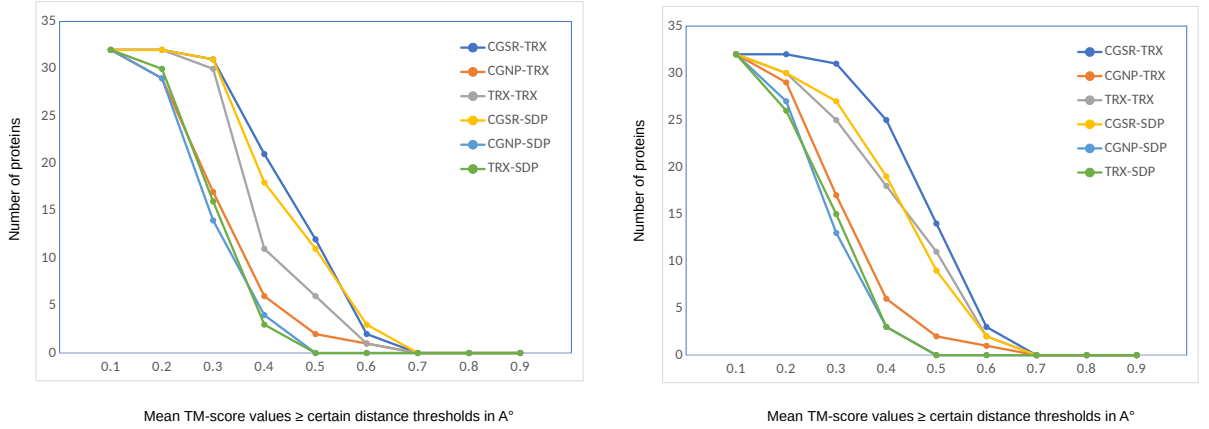


Figure 3: Performance comparison of solver versions in numbers of proteins (y-axis) with final conformations having mean TM-score values $\geq$ certain threshold values (x-axis) for β type (left) and α/β (right) type proteins

statistically significant while that in TM-score is statistically significant. Moreover, the differences in all three scores between TRX-TRX and TRX-SDP are statistically significant while the three scores between CGNP-TRX and CGNP-SDP are statistically not significant. So, SDP predicted distances make differences only with TRX but not with CGSR and CGNP. Nevertheless, CGSR-TRX and CGSR-SDP statistically significantly outperforms CGNP-TRX and CGNP-SDP showing the importance of beta sheet refinement. Furthermore, the performance differences between CGSR-TRX and TRX-TRX are statistically significant and so are the differences between CGSR-SDP and TRX-TRX. These shows the prominence of CGSR-TRX and CGSR-SDP over other solvers.

Figure 5 shows the performances only when beta sheet regions are considered since beta sheet refined is our main focus. For this we use CGSR-TRX, CGNP-TRX, and TRX-TRX as TRX predicted distances appear to help better than SDP predicted distances. We also use all 88 proteins of both β and α/β types together since only beta sheet regions are considered here. We see from the figure that CGSR performances are better than CGNP and TRX in most proteins.

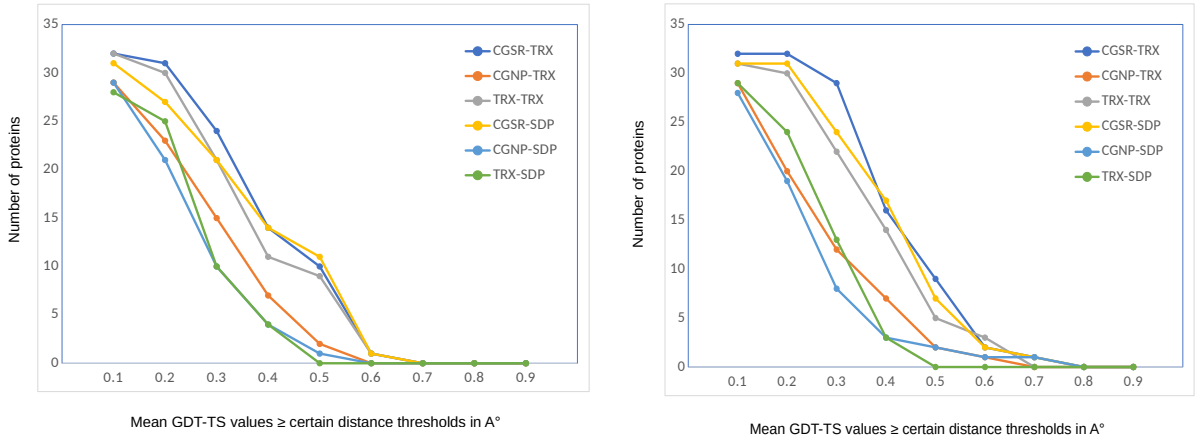


Figure 4: Performance comparison of solver versions in numbers of proteins (y-axis) with final conformations having mean GDT values $\geq$ certain threshold values (x-axis) for β type (left) and α/β (right) type proteins

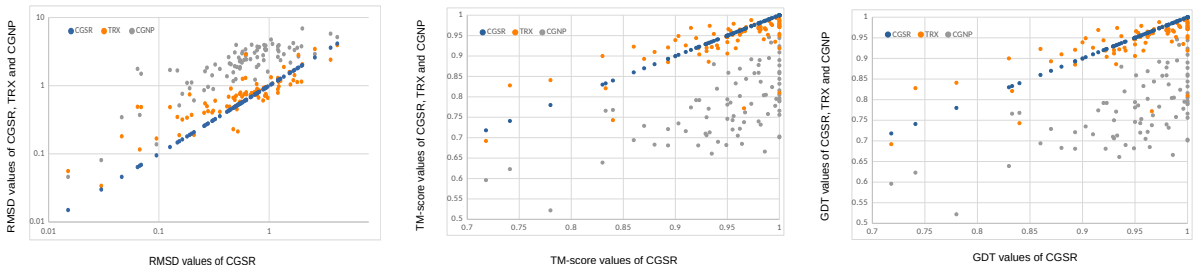


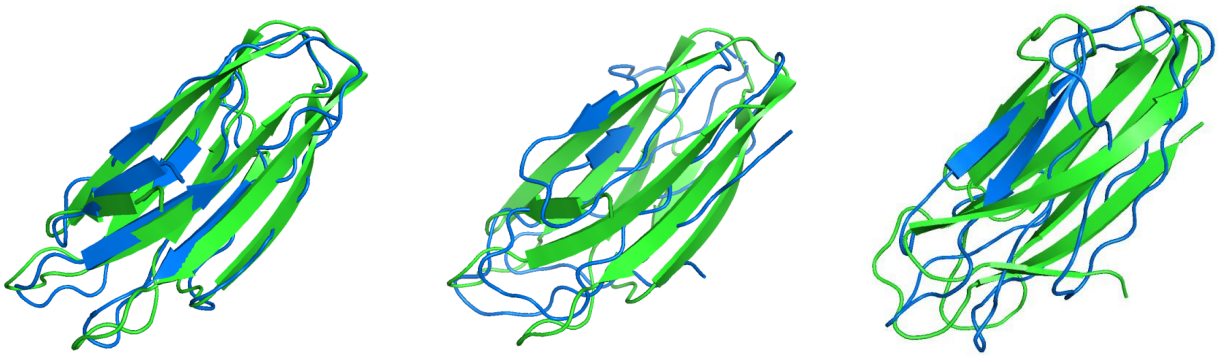
Figure 5: Sctter plots of RMSD (left), TM-score (middle), and GDT (right) values of CGSR, TRX, and CGNP but only for beta sheet regions and using TRX predicted distances for all 88 proteins of both β and α/β types together

Figure 6 shows the best conformations attained by CGSR, TRX and CGNP for a sample protein 5AZW when using TRX predicted inter-residue distance. It is very visible that CGSR is more successful in producing the better structure including beta-sheet regions.

Table 3 nominally shows running times of CGSR, CGNP, and TRX on only two proteins of two different types. We have executed all the algorithms in a Linux 64-bit system with Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz and 8GB memory. Actually, CGSR and CGNP have been implemented on Python, which is by design a very slow programming language and platform. On the other hand, TRX has been implemented on C/C++ programming language and programs written in C/C++ are typically very fast. With these differences in the programming languages and the platforms, comparison of running times of the three PSP methods is actually not meaningful. Note that in our comparison, we have used the generation and evaluation of the same numbers of conformations as the termination criterion.

4. Conclusions

Beta-sheets are the one of the most problematic parts of the protein structures. Better prediction of the Beta-sheet can lead to more accurate tertiary structure prediction. Protein structure



CGSR	TRX	CGNP
RMSD 2.154	RMSD 7.634	RMSD 9.140
TM-score 0.814	TM-score 0.332	TM-score 0.3250
GDT 0.845	GDT 0.321	GDT 0.313

Figure 6: Best conformations obtained by CGSR, TRX and CGNP (all blue) with respect to the native conformations (green) for a sample protein 5AZW of type β when using TRX predicted inter-residue distances

Table 3: Running time analysis

Protein	Type	Length	Method	Time
5AEJ	β	113	CGSR	3 hrs 45 mins
			CGNP	3 hrs 20 mins
			trRosettaX	30 mins 20s
2O42	α β	138	CGSR	2 hrs 30 mins
			CGNP	2 hrs 10 mins
			trRosettaX	25 mins

prediction (PSP) is a challenging problem that needs efforts from both machine learning and search based optimisation approaches. Since the scoring functions that are to be minimised to obtain the native protein structures are not known precisely, machine learning algorithms have been used to learn such functions from known protein structures. Recently, residue-residue distance prediction algorithms have shown a considerable promise in obtaining scoring functions that could effectively approximate such energy functions. However, such scoring functions have been mostly used in evaluation and ranking of generated protein conformations. In this paper, we first try to predict the beta-sheet regions accurately. Then, we focus on the full structure. We show that various constraints learnt by machine learning algorithms and scoring functions that are built from the constraints learnt can not only be used in conformation evaluation but can also be exploited to detect problematic regions of a conformation and then neighbouring conformations could be generated from the identified regions. We evaluate our proposed algorithm on a set of 88 benchmark proteins of different sizes and types. Compared to the state-of-the-art protein structure search algorithms, our algorithms obtain better results. Our algorithm is conceptually simple.

References

[1] K. D. Gibson, H. A. Scheraga, Minimization of polypeptide energy. i. preliminary structures of bovine pancreatic ribonuclease s-peptide., Proceedings of the National Academy of Sciences of the United States of America 58 (2)

(1967) 420.
[2] C. B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (4096) (1973) 223–230.
[3] A. Mittal, B. Jayaram, S. Shenoy, B. Tejdeep Singh, A stoichiometry driven universal spatial organization of backbones of folded proteins: are there chargaff's rules for protein folding?, Journal of Biomolecular Structure and Dynamics 28 (2) (2010) 133–142.
[4] V. Cutello, G. Narzisi, G. Nicosia, A multi-objective evolutionary approach to the protein structure prediction problem, Journal of The Royal Society Interface 3 (6) (2005) 139–151.
[5] R. Zaman, M. A. H. Newton, F. Mataeimoghadam, A. Sattar, Constraint guided neighbour generation for protein structure prediction, IEEE Access 10 (2022) 54991–55001.
[6] W. Qu, H. Sui, B. Yang, Qian, W., Improving protein secondary structure prediction using a multi-modal bp method, Computational Biology and Medicine 41 (2011).
[7] W. Mao, T. Wang, W. Zhang, H. Gong, Identification of residue pairing in interacting β-strands from a predicted residue contact map, BMC Bioinformatics 146 (2018).
[8] K. E., K. T., C. H. s., Mean curvature as a major determinant of β-sheet propensity, Bioinformatics 22 (2006) 297–302.
[9] W. Kabsch, Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[10] I. Ruczinski, C. Kooperberg, R. Bonneau, D. Baker, Distributions of beta sheets in proteins with application to structure prediction,, Proteins 48 (2002) 85–97.
[11] M. Sabzekar, M. Naghibzadeh, M. Eghdami, Protein β-sheet prediction using an efficient dynamic programming algorithm,, Comput. Biol. Chem 20 (2017) 142–155.
[12] T. Dehghani, M. Naghibzadeh, M. Eghdami, BetaDL: A protein beta-sheet predictor utilizing a deep learning model and independent set solution, Computers in Biology and Medicine 104 (2019) 241–249.
[13] D. B. Kc, Recent advances in sequence-based protein structure prediction, Briefings in bioinformatics 18 (6) (2017) 1021–1032.
[14] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (casp)—round xii, Proteins: Structure, Function, and Bioinformatics 86 (2018) 7–15.
[15] F. L. Custódio, H. J. Barbosa, L. E. Dardenne, A multiple minima genetic algorithm for protein structure prediction, Applied Soft Computing 15 (2014) 88–99.
[16] G.-J. Zhang, X.-G. Zhou, X.-F. Yu, X.-H. Hao, L. Yu, Enhancing protein conformational space sampling using distance profile-guided differential evolution, IEEE/ACM transactions on computational biology and bioinformatics 14 (6) (2016) 1288–1301.
[17] X.-G. Zhou, C.-X. Peng, J. Liu, Y. Zhang, G.-J. Zhang, Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction, IEEE Transactions on Evolutionary Computation 24 (3) (2019) 536–550.
[18] C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, Protein structure prediction using rosetta, Methods in enzymology 383 (2004) 66–93.
[19] D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field, Proteins: Structure, Function, and Bioinformatics 80 (7) (2012) 1715–1735.
[20] B. Olson, A. Shehu, Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction, in: Intl Conf on Bioinf and Comp Biol (BICoB). Las Vegas, NV, 2014, pp. 143–148.
[21] S. Song, S. Gao, X. Chen, D. Jia, X. Qian, Y. Todo, AIMOES: Archive information assisted multi-objective evolutionary strategy for ab initio protein structure prediction, Knowledge-Based Systems 146 (2018) 58–72.
[22] R. Shrestha, K. Y. Zhang, Improving fragment quality for de novo structure prediction, Proteins: Structure, Function, and Bioinformatics 82 (9) (2014) 2240–2252.
[23] d. Oliveira, S. H. Law, J. Shi, Eleanor C, M. D. Charlotte, Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction, Bioinformatics 34 (7) (2018) 1132–1140.
[24] X.-g. Zhou, G.-j. Zhang, X.-h. Hao, L. Yu, A novel differential evolution algorithm using local abstract convex underestimate strategy for global optimization, Computers & Operations Research 75 (2016) 132–149.
[25] X.-g. Zhou, G.-j. Zhang, X.-h. Hao, D.-w. Xu, L. Yu, Enhanced differential evolution using local lipschitz underestimate strategy for computationally expensive optimization problems, Applied Soft Computing 48 (2016) 169–181.
[26] X.-G. Zhou, G.-J. Zhang, X.-H. Hao, L. Yu, D.-W. Xu, Differential evolution with multi-stage strategies for global optimization, in: 2016 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2016, pp. 2550–2557.
[27] G.-J. Zhang, L.-F. Ma, X.-Q. Wang, X.-G. Zhou, Secondary structure and contact guided differential evolution for protein structure prediction, IEEE/ACM transactions on computational biology and bioinformatics 17 (3)

(2018) 1068–1081.
[28] X. Chen, S. Song, J. Ji, Z. Tang, Y. Todo, Incorporating a multiobjective knowledge-based energy function into differential evolution for protein structure prediction, Information Sciences 540 (2020) 69–88.
[29] J. Liu, X.-G. Zhou, Y. Zhang, G.-J. Zhang, CGLFold: a contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm, Bioinformatics 36 (8) (2020) 2443–2450.
[30] L. Correa, B. Borguesan, C. Farfán, M. Inostroza-Ponta, M. Dorn, A memetic algorithm for 3D protein structure prediction problem, IEEE/ACM transactions on computational biology and bioinformatics 15 (3) (2016) 690–704.
[31] L. de Lima Correa, B. Borguesan, M. J. Krause, M. Dorn, Three-dimensional protein structure prediction based on memetic algorithms, Computers & Operations Research 91 (2018) 160–177.
[32] L. de Lima Correa, M. Dorn, A multi-population memetic algorithm for the 3-D protein structure prediction problem, Swarm and Evolutionary Computation 55 (2020) 100677.
[33] M. Rashid, M. A. H. Newton, M. T. Hoque, A. Sattar, A local search embedded genetic algorithm for simplified protein structure prediction, in: 2013 IEEE Congress on Evolutionary Computation, 2013, pp. 1091–1098. doi:10.1109/CEC.2013.6557688.
[34] M. Rashid, S. Shatabda, M. A. H. Newton, M. T. Hoque, D. N. Pham, A. Sattar, Random-walk: A stagnation recovery technique for simplified protein structure prediction, 2012, pp. 620–622. doi:10.1145/2382936.2383043.
[35] S. Shatabda, M. A. H. Newton, A. Sattar, Neighborhood selection in constraint-based local search for protein structure predictions, Australasian Joint Conference on Artificial Intelligence 8272 (2013) 44–55.
[36] S. Shatabda, M. A. H. Newton, D. N. Pham, A. Sattar, Memory-based local search for simplified protein structure prediction, BCB '12: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine 8272 (2012) 345–352.
[37] J. Handl, J. Knowles, R. Vernon, D. Baker, S. C. Lovell, The dual role of fragments in fragment-assembly methods for de novo protein structure prediction, Proteins: structure, function, and bioinformatics 80 (2) (2012) 490–504.
[38] D. Xu, Y. Zhang, Toward optimal fragment generations for ab initio protein structure assembly, Proteins: Structure, Function, and Bioinformatics 81 (2) (2013) 229–239.
[39] Y. A. Arnautova, R. A. Abagyan, M. Totrov, Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling, Proteins: Structure, Function, and Bioinformatics 79 (2) (2011) 477–498.
[40] S. Liang, C. Zhang, Y. Zhou, Leap: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains, Journal of computational chemistry 35 (4) (2014) 335–341.
[41] V. Z. Spassov, P. K. Flook, L. Yan, Looper: a molecular mechanics-based algorithm for protein loop prediction, Protein Engineering, Design & Selection 21 (2) (2008) 91–100.
[42] M. Garza-Fabre, S. M. Kandathil, J. Handl, J. Knowles, S. C. Lovell, Generating, maintaining, and exploiting diversity in a memetic algorithm for protein structure prediction, Evolutionary computation 24 (4) (2016) 577–607.
[43] S. Heo, J. Lee, K. Joo, H.-C. Shin, J. Lee, Protein loop structure prediction using conformational space annealing, Journal of chemical information and modeling 57 (5) (2017) 1068–1078.
[44] C. Marks, J. Nowak, S. Klostermann, G. Georges, J. Dunbar, J. Shi, S. Kelm, C. M. Deane, Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction, Bioinformatics 33 (9) (2017) 1346–1353.
[45] C. S. Soto, M. Fasnacht, J. Zhu, L. Forrest, B. Honig, Loop modeling: Sampling, filtering, and scoring, Proteins: Structure, Function, and Bioinformatics 70 (3) (2008) 834–843.
[46] C. Marks, C. M. Deane, Increasing the accuracy of protein loop structure prediction with evolutionary constraints, Bioinformatics 35 (15) (2019) 2585–2592.
[47] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (7792) (2020) 706–710.
[48] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature (2021) 1.
[49] N. Hiranuma, H. Park, M. Baek, I. Anishchenko, J. Dauparas, D. Baker, Improved protein structure refinement guided by deep learning based accuracy estimation, Nature communications 12 (1) (2021) 1–11.
[50] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. t. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al., CHARMM: the biomolecular simulation program, Journal of computational

chemistry 30 (10) (2009) 1545–1614.
[51] T. M. Leaver-Fay A, S. Lewis, ROSETTA3: an object- oriented software suite for the simulation and design of macromolecules, Methods Enzymol 487 (2011) 545–574.
[52] B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng, CONFOLD: residue-residue contact-guided ab initio protein folding, Proteins: Structure, Function, and Bioinformatics 83 (8) (2015) 1436–1449.
[53] B. Adhikari, J. Cheng, CONFOLD2: improved contact-driven ab initio protein structure modeling, BMC bioinformatics 19 (1) (2018) 1–5.
[54] J. Hou, T. Wu, R. Cao, J. Cheng, Protein tertiary structure modeling driven by deep learning and contact distance prediction in casp13, Proteins: Structure, Function, and Bioinformatics 87 (12) (2019) 1165–1178.
[55] J. Ma, S. Wang, Z. Wang, J. Xu, Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning, Bioinformatics 31 (21) (2015) 3506–3513.
[56] S. Wang, W. Li, R. Zhang, S. Liu, J. Xu, COINFOLD: a web server for protein contact prediction and contact-assisted protein folding, Nucleic acids research 44 (W1) (2016) W361–W366.
[57] S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, Accurate de novo prediction of protein contact map by ultra-deep learning model, PLoS computational biology 13 (1) (2017) e1005324.
[58] F. Mataeimoghadam, M. H. Newton, A. Dehzangi, A. Karim, B. Jayaram, S. Ranganathan, A. Sattar, Enhancing protein backbone angle prediction by using simpler models of deep neural networks, Scientific Reports 10 (1) (2020) 1–12.
[59] X. Gang, Q. Wang, J. Ma, Opus-tass: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks, Bioinformatics 36 (20) (2020) 5021–502.
[60] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, Y. Zhou, Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks, Bioinformatics 35 (14) (2018) 2403–2410.
[61] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers: Original Research on Biomolecules 22 (12) (1983) 2577–2637.
[62] C. N. Magnan, P. Baldi, SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity, Bioinformatics 30 (18) (2014) 2592–2597.
[63] J. Rahman, M. A. H. Newton, M. K. B. Islam, A. Sattar, Enhancing protein inter-residue real distance prediction by scrutinising deep learning models, Scientific Reports 12 (87) (2022).
[64] Q. Wu, Q. Cong, D. Baker, J. Yang, Protein contact prediction using metagenome sequence data and residual neural networks, Bioinformatics 787 (1967) 420.
[65] H. Su, W. Wang, Z. yang Du, Z. Peng, S.-H. Gao, M.-M. Cheng, J. Yang, Improved protein structure prediction using a new multi-scale network and homologous templates, Advanced Science 8 (24) (2021).
[66] F. Glover, Tabu search methods in artificial intelligence and operations research, ORSA Artificial Intelligence 1 (2) (1987) 6.
[67] M. A. H. Newton, D. N. Pham, A. Sattar, M. Maher, Kangaroo: An efficient constraint-based local search system using lazy propagation, in: International Conference on Principles and Practice of Constraint Programming, Springer, 2011, pp. 645–659.
[68] D. Ramyachitra, A. Ajeeth, MODCSA-CA: a multi objective diversity controlled self adaptive cuckoo algorithm for protein structure prediction, Gene Reports 8 (2017) 100–106.

Appendix

Table A1: Mean RMSD values for β type proteins as obtained by CGSR, CGNP and TrRosettaX using inter-residue distances predicted by trRoesttaX and SDP. Best values are emboldened and second best values are underlined

Protein		Distances by trRosettaX			Distances by SDP		
ID	Length	CGSR	CGNP	trRosettaX	CGNP	CGSR	trRosettaX
1R75	110	6.066	7.124	6.518	**3.998**	9.498	7.537
1OK0	74	4.711	7.032	8.486	**4.519**	6.608	8.029
2BT9	90	**3.779**	8.54	7.028	6.686	8.248	9.743
2CHH	113	8.154	10.9	11.348	**7.194**	11.593	10.957
2V33	91	6.882	8.22	9.488	**6.514**	8.028	9.798
5AEJ	113	8.116	10.03	**7.496**	9.336	15.804	12.027
5AOT	102	9.228	10.771	9.642	**9.015**	11.075	11.768
5EZU	67	**6.216**	7.91	6.921	6.246	6.451	8.913
5FUI	124	**10.868**	12.31	11.974	11.814	12.738	12.085
5HDW	131	**9.012**	11.59	10.31	9.769	10.381	11.594
5KEW	125	**10.515**	16.9	12.636	11.515	15.115	16.903
2AYD	76	**5.618**	8.69	9.429	6.358	7.403	9.692
5AZW	94	4.475	13.16	4.645	**4.411**	8.448	13.761
5DHD	98	**4.03**	9.26	6.719	4.335	9.571	9.456
6WES	158	12.489	13.88	**10.747**	13.358	15.019	14.878
6ZGQ	154	**14.125**	17.23	14.133	15.265	16.622	17.425
7BQF	84	**13.288**	14.495	13.615	13.535	15.662	14.875
7BQG	85	**12.352**	13.89	13.318	12.882	14.034	13.894
7C28	65	5.669	9.36	6.512	**4.185**	4.604	9.86
T0960	374	49.11	51.43	**48.773**	51.054	70.786	51.454
T0968S2	114	**7.151**	10.72	13.028	8.316	15.812	10.965
T0990	134	34.671	39.98	35.81	34.711	**31.451**	39.785
T0992	107	5.112	14.06	6.676	**4.469**	8.352	15.055
T0981	203	**10.168**	19.675	10.595	10.232	18.265	19.61
T0987	381	18.411	23.14	**17.659**	21.022	30.812	23.742
T1010	210	**12.1**	22.65	14.128	13.85	17.209	24.949
5JOE	92	5.276	13.81	9.933	**3.634**	9.018	13.979
6QPN	200	**12.477**	20.68	12.743	12.491	18.629	22.67
6QPO	200	12.587	17.12	12.903	**12.326**	16.188	17.479
7CCB	164	**8.369**	15.39	9.355	9.778	17.107	15.789
3X29	166	**7.321**	9.12	8.162	7.744	14.172	10.458
7BZZ	109	10.202	13.67	13.234	12.591	**10.073**	14.673

Acknowledgements

This research is partially supported by Australian Research Council Grant DP180102727.

Author contributions statement

M.A.H.N. and R.Z contributed equally and in all parts of the work. F.M. and J.R. helped to run experiments. A.S. took part in discussions and reviewed the manuscript.

Additional information

The author(s) declare no competing interests.

Table A2: Mean RMSD values for α/β type proteins as obtained by CGSR, CGNP and TrRosettaX using inter-residue distances predicted by trRoesttaX and SDP. Best values are emboldened and second best values are underlined

Protein		Distances by trRosettaX			Distances by SDP		
ID	Length	CGSR	CGNP	trRosettaX	CGNP	CGSR	trRosettaX
1CRN	46	**2.148**	2.704	4.602	2.488	2.775	7.15
1CF7	82	5.92	7.88	8.938	**5.207**	6.528	8.449
1IS7	84	6.226	8.649	6.603	**4.612**	7.106	5.46
1KA8	100	8.868	**6.953**	9.604	7.651	11.881	9.656
1MC2	122	5.091	7.311	**5.065**	5.073	10.161	11.043
1T1J	125	**5.044**	10.32	10.612	5.999	12.076	9.425
1Y71	112	**7.041**	8.616	9.962	8.814	12.955	10.247
2BSE	107	**6.001**	10.215	7.48	7.122	10.175	9.772
3BJO	100	7.997	7.44	8.921	**6.336**	9.847	8.978
3CHB	103	9.151	10.63	9.334	**9.041**	11.046	12.23
6I1M	93	**7.111**	10.884	7.602	7.207	12.016	9.861
6KYF	137	10.163	13.485	**9.276**	10.475	16.982	14.528
6L7Q	145	12.191	**12.123**	12.986	12.672	15.104	13.23
6LXG	90	3.522	4.136	4.201	**3.381**	5.288	13.622
6TZN	115	**5.021**	11.346	9.839	9.264	11.159	14.887
6UXC	172	**10.123**	13.892	10.503	10.488	14.418	15.047
6XFU	70	6.252	7.328	6.905	**6.131**	8.679	13.727
6M1J	205	**11.215**	13.611	11.434	11.342	14.903	19.616
T0949	183	**10.703**	18.71	19.655	15.865	18.685	17.735
T0957S1	157	12.925	16.964	13.013	**12.514**	18.166	19.275
T0958	84	5.669	6.14	6.393	**5.623**	9.783	12.866
T0968S1	116	**8.082**	13.831	15.031	13.425	14.554	16.652
T0969	354	**12.185**	21.536	12.546	12.354	23.356	16.199
T0970	97	**7.304**	9.464	7.772	7.669	10.709	14.588
T0978	413	15.923	24.773	16.843	**15.468**	25.722	18.358
T0980S1	98	11.446	12.3	12.282	11.773	**10.692**	15.535
T0986S1	89	**10.686**	13.8	11.81	13.634	14.423	14.512
T0986S2	150	10.111	13.976	**9.641**	10.032	15.638	11.423
T0991	118	**16.294**	18.177	16.732	17.511	16.825	18.445
T0997	185	**8.611**	15.366	8.926	8.912	16.706	15.765
T0998	166	17.595	16.496	17.032	17.013	**15.915**	19.118
T1000	450	16.111	27.406	**16.092**	16.713	25.867	18.139
T1001	139	10.305	13.732	10.845	**10.022**	14.002	11.1
T1008	80	**6.016**	8.492	6.91	6.392	6.142	10.295
T1015S1	88	8.519	7.721	9.728	**7.129**	11.381	9.884
T1017S2	125	**10.111**	12.645	11.792	10.264	14.404	14.784
T1019S1	58	**2.906**	3.768	4.895	3.233	3.251	6.742
5CX7	147	10.946	10.837	11.567	**7.975**	13.532	9.535
5ECD	132	4.519	6.56	5.594	**3.482**	8.894	5.4
5J4A	112	**6.327**	6.35	8.387	7.373	7.449	9.456
5SV2	135	8.335	10.958	11.193	**7.422**	11.683	14.475
5T6J	92	4.594	7.53	6.131	**4.388**	7.867	9.357
5LSI	76	7.094	10.992	**6.549**	7.841	13.323	10.477
1VF5	168	18.252	21.102	18.282	**18.231**	22.067	21.574
3VWI	177	10.002	14.788	12.928	**8.294**	16.036	9.458
6YYL	147	**8.133**	9.159	8.726	8.521	13.407	10.833
6TTS	176	**7.202**	11.225	7.799	9.322	14.817	10.235
6Z92	151	**9.757**	11.055	17.03	12.902	13.794	14.845
7JH1	64	**3.114**	5.028	5.508	3.224	3.987	8.571
6ZSO	96	**8.151**	9.323	8.496	8.293	10.397	14.14
T0953S2	241	13.444	23.1	**12.388**	**12.245**	25.881	17.559
T0963	364	**45.351**	69.759	46.487	46.94	68.773	51.437
T0989	134	14.221	**14.128**	14.473	14.166	15.348	16.739
T1005	319	12.181	21.116	12.075	**11.841**	23.358	15.838
T1021S3	275	18.621	30.521	**14.428**	15.834	35.509	18.198
T1022S1	223	14.411	18.915	14.668	**14.342**	22.139	16.504

Table A3: Mean TM-score values for β type proteins as obtained by CGSR, CGNP and TrRosettaX using inter-residue distances predicted by trRoesttaX and SDP. Best values are emboldened and second best values are underlined

Protein		Distances by trRosettaX			Distances by SDP		
ID	Length	CGSR	CGNP	trRosettaX	CGNP	CGSR	trRosettaX
1R75	110	**0.635**	0.371	0.51	0.621	0.352	0.375
1OK0	74	**0.534**	0.391	0.357	0.533	0.408	0.476
2BT9	90	**0.589**	0.374	0.568	0.574	0.354	0.381
2CHH	113	**0.546**	0.273	0.337	0.413	0.336	0.387
2V33	91	**0.457**	0.364	0.328	0.433	0.369	0.377
5AEJ	113	0.535	0.281	**0.589**	0.534	0.275	0.445
5AOT	102	**0.512**	0.338	0.342	0.394	0.295	0.275
5EZU	67	0.376	**0.405**	0.353	0.369	0.367	0.264
5FUI	124	**0.369**	0.286	0.301	0.341	0.264	0.273
5HDW	131	**0.367**	0.339	0.354	0.357	0.311	0.266
5KEW	125	**0.368**	0.294	0.319	0.331	0.229	0.263
2AYD	76	**0.587**	0.385	0.327	0.562	0.409	0.276
5AZW	94	0.585	0.358	**0.656**	0.575	0.368	0.374
5DHD	98	**0.555**	0.351	0.542	0.559	0.303	0.39
6WES	158	**0.429**	0.228	0.311	0.411	0.252	0.354
6ZGQ	154	**0.486**	0.208	0.334	0.422	0.213	0.359
7BQF	84	**0.476**	0.274	0.382	0.396	0.259	0.362
7BQG	85	**0.395**	0.278	0.386	0.323	0.267	0.275
7C28	65	0.474	0.419	0.458	**0.506**	0.445	0.369
T0960	374	**0.378**	0.113	0.215	0.327	0.112	0.226
T0968S2	114	**0.428**	0.266	0.352	0.407	0.215	0.372
T0990	134	**0.391**	0.173	0.242	0.271	0.167	0.161
T0992	107	0.495	0.46	0.485	**0.598**	0.451	0.278
T0981	203	0.221	0.231	**0.394**	0.373	0.208	0.281
T0987	381	**0.489**	0.138	0.33	0.421	0.124	0.283
T1010	210	**0.491**	0.211	0.461	0.322	0.218	0.154
5JOE	92	0.519	0.591	0.319	**0.615**	0.364	0.473
6QPN	200	0.361	0.223	**0.457**	0.349	0.215	0.269
6QPO	200	0.342	0.311	0.319	**0.376**	0.234	0.272
7CCB	164	**0.551**	0.431	0.466	0.491	0.208	0.395
3X29	166	**0.692**	0.631	0.55	0.673	0.302	0.384
7BZZ	109	**0.391**	0.342	0.315	0.352	0.291	0.282

Table A4: Mean TM-score values for α/β type proteins as obtained by CGSR, CGNP and TrRosettaX using inter-residue distances predicted by trRoesttaX and SDP. Best values are emboldened and second best values are underlined

Protein		Distances by trRosettaX			Distances by SDP		
ID	Length	CGSR	CGNP	trRosettaX	CGNP	CGSR	trRosettaX
1CRN	46	**0.623**	0.371	0.508	0.573	0.345	0.476
1CF7	82	**0.592**	0.391	0.5	0.578	0.412	0.393
1IS7	84	**0.545**	0.374	0.49	0.512	0.499	0.37
1KA8	100	**0.481**	0.273	0.471	0.474	0.297	0.19
1MC2	122	**0.587**	0.364	0.572	0.585	0.329	0.295
1T1J	125	**0.631**	0.281	0.551	0.601	0.301	0.492
1Y71	112	**0.485**	0.338	0.46	0.473	0.325	0.386
2BSE	107	**0.486**	0.405	**0.468**	0.475	0.269	0.375
3BJO	100	**0.493**	0.286	0.289	0.463	0.379	0.286
3CHB	103	**0.493**	0.339	0.391	0.422	0.314	0.38
6I1M	93	**0.535**	0.294	0.492	0.513	0.255	0.478
6KYF	137	**0.397**	0.385	0.27	0.393	0.223	0.291
6L7Q	145	0.472	0.358	0.343	0.438	0.235	0.277
6LXG	90	**0.697**	0.351	0.68	0.679	0.489	0.279
6TZN	115	**0.457**	0.228	0.324	0.421	0.276	0.391
6UXC	172	**0.551**	0.208	0.513	0.352	0.237	0.375
6XFU	70	**0.482**	0.274	0.404	0.348	0.322	0.276
6M1J	205	**0.543**	0.278	0.53	0.421	0.301	0.382
T0949	183	0.368	**0.419**	0.123	0.234	0.121	0.133
T0957S1	157	**0.393**	0.113	0.272	0.262	0.188	0.162
T0958	84	**0.493**	0.266	0.381	0.491	0.263	0.379
T0968S1	116	**0.414**	0.173	0.169	0.324	0.173	0.362
T0969	354	0.552	0.461	**0.619**	0.394	0.201	0.298
T0970	97	**0.556**	0.231	0.544	0.365	0.335	0.279
T0978	413	**0.493**	0.138	0.48	0.259	0.182	0.204
T0980S1	98	**0.581**	0.211	0.346	0.374	0.265	0.265
T0986S1	89	0.318	**0.591**	0.202	0.374	0.312	0.276
T0986S2	150	**0.381**	0.223	0.362	**0.439**	0.204	0.171
T0991	118	0.256	**0.312**	0.244	0.191	0.202	0.141
T0997	185	**0.585**	0.441	0.565	0.571	0.231	0.303
T0998	166	0.378	**0.631**	0.351	0.194	0.218	0.143
T1000	450	**0.567**	0.342	0.5	0.523	0.191	0.305
T1001	139	**0.379**	0.371	0.346	0.335	0.216	0.216
T1008	80	**0.567**	0.391	0.513	0.373	0.478	0.371
T1015S1	88	0.356	**0.374**	0.308	0.346	0.255	0.266
T1017S2	125	**0.485**	0.273	0.472	0.376	0.231	0.187
T1019S1	58	0.596	0.364	0.516	**0.572**	0.552	0.486
5CX7	147	**0.465**	0.281	0.33	0.435	0.265	0.309
5ECD	132	**0.782**	0.338	0.761	0.681	0.398	0.494
5J4A	112	**0.594**	0.405	0.541	0.495	0.409	0.396
5SV2	135	**0.563**	0.286	0.55	0.522	0.325	0.302
5T6J	92	**0.581**	0.339	0.487	0.523	0.473	0.389
5LSI	76	**0.583**	0.294	0.34	0.534	0.343	0.334
1VF5	168	0.436	0.385	**0.441**	0.293	0.248	0.17
3VWI	177	**0.578**	0.358	0.238	0.467	0.246	0.288
6YYL	147	**0.595**	0.351	0.588	0.475	0.292	0.384
6TTS	176	**0.529**	0.228	0.51	0.517	0.283	0.489
6Z92	151	**0.463**	0.208	0.168	0.276	0.253	0.186
7JH1	64	**0.704**	0.274	0.508	0.594	0.585	0.378
6ZSO	96	**0.59**	0.278	0.516	0.556	0.321	0.468
T0953S2	241	**0.465**	0.419	0.398	0.393	0.159	0.285
T0963	364	**0.328**	0.113	0.208	0.317	0.116	0.231
T0989	134	0.323	0.266	0.226	**0.472**	0.214	0.368
T1005	319	**0.587**	0.173	0.513	0.522	0.174	0.303
T1021S3	275	0.427	0.461	0.407	**0.506**	0.188	0.347
T1022S1	223	0.403	0.231	0.384	**0.474**	0.199	0.361

Table A5: Mean GDT values for β type proteins as obtained by CGSR, CGNP and TrRosettaX using inter-residue distances predicted by trRoesttaX and SDP. Best values are emboldened and second best values are underlined

Protein		Distances by trRosettaX			Distances by SDP		
ID	Length	CGSR	CGNP	trRosettaX	CGNP	CGSR	trRosettaX
1R75	110	**0.558**	0.355	0.535	0.557	0.318	0.444
1OK0	74	**0.563**	0.364	0.385	0.555	0.446	0.393
2BT9	90	**0.576**	0.461	0.569	0.571	0.368	0.373
2CHH	113	**0.432**	0.243	0.296	0.372	0.294	0.248
2V33	91	0.343	0.382	0.321	**0.424**	0.355	0.266
5AEJ	113	0.489	0.266	**0.553**	0.446	0.244	0.226
5AOT	102	**0.382**	0.317	0.314	0.372	0.285	0.253
5EZU	67	**0.544**	0.481	0.53	0.541	0.452	0.3
5FUI	124	**0.323**	0.238	0.25	0.281	0.226	0.26
5HDW	131	**0.395**	0.284	0.29	0.293	0.257	0.221
5KEW	125	**0.289**	0.246	0.29	0.252	0.192	0.218
2AYD	76	0.539	0.408	0.345	**0.576**	0.434	0.492
5AZW	94	**0.675**	0.359	0.639	0.645	0.359	0.457
5DHD	98	0.361	0.338	**0.532**	**0.537**	0.298	0.368
6WES	158	**0.321**	0.161	0.236	0.242	0.182	0.2
6ZGQ	154	0.271	0.151	**0.287**	0.141	0.163	0.104
7BQF	84	**0.379**	0.291	0.365	0.373	0.281	0.273
7BQG	85	**0.432**	0.303	0.406	0.394	0.293	0.281
7C28	65	0.527	0.511	0.511	**0.577**	0.504	0.407
T0960	374	**0.212**	0.061	0.137	0.164	0.059	0.049
T0968S2	114	**0.393**	0.243	0.319	0.388	0.239	0.234
T0990	134	**0.178**	0.063	0.118	0.145	0.045	0.043
T0992	107	**0.578**	0.411	0.54	0.562	0.392	0.247
T0981	203	0.227	0.133	**0.362**	0.163	0.125	0.091
T0987	381	0.217	0.071	**0.251**	0.084	0.067	0.057
T1010	210	**0.378**	0.129	0.347	0.359	0.132	0.279
5JOE	92	0.507	0.571	0.315	**0.595**	0.362	0.262
6QPN	200	0.332	0.141	**0.372**	0.249	0.132	0.189
6QPO	200	**0.245**	0.211	0.2	0.221	0.143	0.122
7CCB	164	0.422	0.198	**0.463**	0.454	0.151	0.315
3X29	166	0.547	0.481	**0.561**	0.531	0.227	0.319
7BZZ	109	0.292	**0.365**	0.273	0.363	0.261	0.22

Table A6: Mean GDT values for α/β type proteins as obtained by CGSR, CGNP and TrRosettaX using inter-residue distances predicted by trRoesttaX and SDP. Best values are emboldened and second best values are underlined

Protein		Distances by trRosettaX			Distances by SDP		
ID	Length	CGSR	CGNP	trRosettaX	CGNP	CGSR	trRosettaX
1CRN	46	**0.758**	0.692	0.61	0.722	0.714	0.496
1CF7	82	**0.551**	0.411	0.547	0.518	0.434	0.406
1IS7	84	**0.58**	0.366	0.49	0.572	0.305	0.378
1KA8	100	**0.529**	0.421	0.445	0.464	0.283	0.371
1MC2	122	0.58	0.388	**0.605**	0.595	0.292	0.351
1T1J	125	**0.516**	0.338	0.398	0.487	0.258	0.25
1Y71	112	**0.576**	0.408	0.525	0.537	0.292	0.358
2BSE	107	**0.491**	0.291	0.434	0.475	0.255	0.343
3BJO	100	0.392	0.404	0.267	**0.432**	0.371	0.37
3CHB	103	0.376	0.278	0.36	**0.382**	0.295	0.26
6I1M	93	**0.521**	0.283	0.477	0.493	0.269	0.371
6KYF	137	**0.391**	0.212	0.323	0.321	0.186	0.236
6L7Q	145	**0.395**	0.184	0.289	0.289	0.186	0.222
6LXG	90	0.681	0.564	0.67	**0.691**	0.513	0.493
6TZN	115	**0.331**	0.269	0.278	0.325	0.246	0.258
6UXC	172	**0.431**	0.197	0.413	0.415	0.179	0.208
6XFU	70	**0.448**	0.396	0.435	0.439	0.378	0.31
6M1J	205	**0.492**	0.216	0.409	0.476	0.203	0.24
T0949	183	**0.302**	0.082	0.078	0.083	0.078	0.079
T0957S1	157	0.231	0.144	0.212	**0.254**	0.138	0.105
T0958	84	0.498	0.321	0.39	**0.512**	0.276	0.171
T0968S1	116	0.325	0.167	0.166	**0.394**	0.147	0.226
T0969	354	**0.432**	0.112	0.42	0.245	0.091	0.068
T0970	97	0.342	0.297	0.34	**0.352**	0.322	0.263
T0978	413	**0.352**	0.027	0.344	0.209	0.091	0.061
T0980S1	98	**0.377**	0.261	0.324	0.312	0.273	0.151
T0986S1	89	0.325	0.412	0.2	**0.415**	0.383	0.375
T0986S2	150	**0.282**	0.171	0.267	0.252	0.143	0.208
T0991	118	**0.292**	0.187	0.219	0.221	0.188	0.213
T0997	185	**0.481**	0.158	0.446	0.452	0.149	0.31
T0998	166	**0.343**	0.162	0.275	0.279	0.149	0.197
T1000	450	**0.321**	0.074	0.308	0.312	0.069	0.119
T1001	139	**0.389**	0.179	0.367	0.355	0.172	0.132
T1008	80	**0.621**	0.401	0.615	0.523	0.494	0.287
T1015S1	88	**0.385**	0.353	0.324	0.374	0.283	0.27
T1017S2	125	**0.422**	0.221	0.413	0.334	0.223	0.142
T1019S1	58	**0.771**	0.607	0.591	0.642	0.639	0.255
5CX7	147	0.256	0.273	0.255	**0.357**	0.231	0.246
5ECD	132	0.611	0.484	0.613	**0.621**	0.343	0.54
5J4A	112	**0.488**	0.435	0.487	0.461	0.354	0.36
5SV2	135	**0.489**	0.426	0.487	0.454	0.273	0.346
5T6J	92	**0.572**	0.437	**0.572**	0.564	0.456	0.485
5LSI	76	**0.526**	0.51	0.299	0.517	0.377	0.405
1VF5	168	**0.311**	0.181	0.295	0.272	0.183	0.105
3VWI	177	**0.372**	0.201	0.216	0.328	0.169	0.202
6YYL	147	**0.471**	0.313	0.45	0.461	0.227	0.321
6TTS	176	**0.527**	0.283	0.51	0.517	0.198	0.311
6Z92	151	0.268	**0.286**	0.117	0.212	0.193	0.123
7JH1	64	**0.662**	0.544	0.626	0.660	0.635	0.524
6ZSO	96	**0.513**	0.418	0.494	0.495	0.317	0.355
T0953S2	241	**0.296**	0.104	0.24	0.282	0.092	0.183
T0963	364	**0.212**	0.062	0.132	0.112	0.066	0.051
T0989	134	**0.383**	0.189	0.272	0.213	0.176	0.121
T1005	319	**0.35**	0.092	0.319	0.337	0.021	0.229
T1021S3	275	**0.334**	0.117	0.272	0.319	0.103	0.267
T1022S1	223	**0.289**	0.134	0.277	0.288	0.123	0.181