Tailoring Contact Based Scoring Functions for Protein Structure Prediction

Tailoring Contact Based Scoring Functions for Protein Structure Prediction

Rianon Zaman[1], M.A. Hakim Newton[2],
Fereshteh Mataeimoghadam[1], and Abdul Sattar[1,2]

[1] School of ICT, Griffith University, Australia
[2] IIIS, Griffith University, Australia
{rianon.zaman,fereshteh.mataeimoghadam}@griffithuni.edu.au
{mahakim.newton,a.sattar}@griffith.edu.au

Abstract. Protein structure prediction (PSP) is a challenging problem in Bioinformatics. Given a protein's amino acid sequence, PSP involves finding its three dimensional native structure having the minimum free energy. Unfortunately, the search space is astronomical and the energy function is not known. Many PSP search algorithms develop their own proxy energy functions known as scoring functions using predicted contacts between amino acid residue pairs where two residues are said to be in contact if their distance in the native structure is within a given threshold. Scoring functions are crucial for search guidance since they allow evaluation of the generated structures. Unfortunately, existing contact based scoring functions have not been directly compared and which one among them is the best is not known. In this paper, we evaluate a number of existing contact based scoring functions within the same PSP search framework on the same set of benchmark proteins. Moreover, we also propose a number of contact based scoring function variants. Our proposed contact based scoring functions help our search algorithm to significantly outperform existing state-of-the-art PSP search algorithms that use contact based scoring functions.

Keywords: Protein Structure Prediction, Search-Based Optimisation, Contact-Based Energy Function

1 Introduction

Protein structure prediction (PSP) is a challenging problem in Bioinformatics. Proteins comprise amino acid (AA) sequences and fold into three dimensional structures to perform their functions. Given a protein's AA sequence, PSP involves finding its native structure that has the minimum free energy. Unfortunately, the search space is astronomical and the energy function is not known.

Energy functions have been developed based on molecular dynamics e.g. CHARMM [3]. Unfortunately, such energy functions involve all atomic details and so are computationally very expensive. Rosetta [8] is a popular energy function but involves 18 different energy components. Consequently, various other proxy energy functions known as scoring functions have been designed. In this context, contact based scoring functions have been used by many recent PSP search algorithms. Two amino acid residues of a protein are in contact if their

distance in the native structure of the protein is at most 8Å. Machine learning algorithms are normally used in predicting potential contacts between residue pairs. Contact based scoring functions are then developed to evaluate protein structures based on the deviations in the distances between residue pairs that are supposed to be in contact. Search algorithms then use the scoring functions to rank generated protein structures or conformations.

Machine learning algorithms such as SPOT-Contact [5], Restriplet [9], and TripletRes[10] predict contacts among residues. Search algorithms such as Pconsfold [17], CONFOLD [1], RBO Aleph [12], Unicon3D [2] and CGLFOLD[11] use contact based scoring functions. In this context, recent contact based scoring functions include modified Lorentz potential [12], soft square [1], square well [7, 2], bounded potential [7], and cglfold [11]. However, these scoring functions have not been directly compared and which one among them is the best is not known.

In this paper, we evaluate the aforementioned five contact based scoring functions within the same PSP search framework on the same single set of benchmark proteins. Based on the results, we also propose four contact based scoring function variants. Our proposed contact based scoring functions help our search algorithm to significantly outperform existing state-of-the-art PSP search algorithm CGLFOLD [11] that uses a contact based scoring function.

The rest of the papers is organized as follows: Section 2 provide preliminaries of protein structures and our search framework; Section 3 describes existing contact based scoring functions as well as our proposed ones; Section 4 provides our experimental results and analyses; and Section 5 presents our conclusions.

2 Preliminaries

We briefly describe protein structure preliminaries and our search framework.

Protein Structures. Proteins comprise 20 types of AA and the AAs can appear in any order any number of times. Moreover, AAs all have N, C^α, and C atoms in their main chains. Two successive AA residues in a protein are joined by a non-rotatable peptide bond formed between the C atom of the previous residue and the N atom of the next residue. The bond between N and C^α in an AA is rotatable and the rotation angle is denoted by ϕ. Similarly, the bond between C^α and C in an AA is also rotatable and the rotation angle is denoted by ψ. Both ϕ and ψ can take any value from $[-180°, +180°]$. The rotatable bonds are essentially responsible for the three dimensional folding of a protien. Proteins exhibit certain local structures comprising successive residues. These local structures known as secondary structures are of three major types: helices, sheets, and coils. Among these, helices and sheets are rigid and normally have narrow ranges of ϕ and ψ values but coils are very flexible; hence, in this work, we mainly search for ϕ and ψ angles of the coil reidues. Nevertheless, besides main chains, AAs have unique side chains (Glycine has no side chain) starting from C^α and having C^β as the first atom. Side chains have dihedral angles, too, but they are out of scope of this work. Nevertheless, in the definition of contacts between residues, typically distances are measured between the C^β atoms (C^α for Glycine) of the two residues so that side chains are counted to some extent.

Search Framework. We use a constraint based local search (CBLS) framework to evaluate the existing and the proposed contact based energy functions. The search algorithm is implemented on top of a new python library named Koala, which draws concepts from a constraint based local search system named Kangaroo [15]. We briefly describe the steps of our search algorithm below:

1. Generate one initial conformation c using ϕ, ψ angles predicted for each residue of the protein by a machine learning algorithm.
2. Evaluate the conformation c using a contact based scoring function σ.
3. Select the residue pair $\langle i, j\rangle$ from c such that residues i and j are supposed to be in contact (as predicted by a machine learning algorithm) but their distance is the maximum in c among all such candidate residue pairs.
4. Select a residue k randomly from any coil (not helices and sheets since they are rigid) in between the selected residues i and j. Changing ϕ and ψ of the selected residue k might essentially bring residues i and j in contact.
5. Generate a number (e.g. 20) of neighbouring conformations by changing ϕ and ψ angles of the residue selected residue k. Consider up to $\pm\Delta$ with interval $\delta = 3$ for ϕ and ψ values where Δ is the mean absolute error of the machine learning algorithm used in Step 1 for the respective ϕ or ψ angle.
6. Evaluate the generated neigbouring conformations using the same contact based scoring function σ used in Step 2.
7. Accept the neighbouring conformation having the minimum score as the current conformation for the next iteration.
8. Return the best conformatoin found so far (in terms of scores) if the termination criterion is satisfied; otherwise, move to Step 3.

3 Scoring Functions

Assume d_{ij} is the distance and σ_{ij} is the score for a residue pair $\langle i, j\rangle$ in a conformation c having the score $\sigma = \sum \sigma_{ij}$. Also, assume p_{ij} be the probability that residues i and j are in contact in the native conformation.

3.1 Existing scoring functions

Fig. 1 shows five existing contact based scoring functions. These functions are square well (sw) [7, 2], bounded potential (bp) [7], modified Lorentz potential (mlp) [12], soft square (ss) [1], cglfold (cf) [11]. The parameter values used in the functions are as suggested by the respective methods using them.

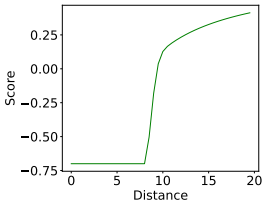
From the charts of the five scoring functions in Fig. 1, notice that most of the scoring functions have a d_{ij} range with some least penalty value. Any d_{ij} below ≈ 3.8Å is highly penalised to avoid steric clash between residues. Also, any d_{ij} above ≈ 8Å is penalised to avoid not having contact while a contact is rather expected. The square well function does not penalise for steric clash ($d_{ij} \leq d_0 = 3.8$) but the other functions do. The modified Lorentz potential and the soft square functions become flat in large d_{ij} values. So these functions perhaps would not be able to provide effective search guidance when d_{ij} values are large since such values cannot be differentiated. The square well and the cglfold functions are very similar for large d_{ij} values but are different for small d_{ij}

Square Well

$$\sigma_{ij} = -p_{ij} \quad \text{if } p_{ij} \leq d_0$$
$$= -p_{ij} e^{-(d-d_0)^2} + p_{ij}\left(\frac{d-d_0}{d}\right) \text{ if } d > d_0$$

where $d_0 = 8$

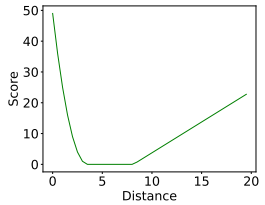
$p_{ij} = 0.7$ in the chart



Bounded Potential

$$\sigma_{ij} = \left(\frac{d_{ij}-l}{s}\right)^2 \quad \text{if } d_{ij} < l$$
$$= 0 \quad \text{if } l < d_{ij} \leq u$$
$$= \left(\frac{d_{ij}-u}{s}\right)^2 \quad \text{if } u < d_{ij} \leq u + 0.5s$$
$$= \left(\frac{d_{ij}-u-0.5s}{s}\right) + 0.25 \text{ if } d_{ij} > u + 0.5s$$

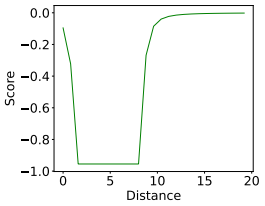
where $l = 3.5, u = 8, s = 0.5$



Modified Lorentz Potential

$$\sigma_{ij} = \frac{c_{ij}}{\pi} \times \frac{\frac{w}{2}}{(d_{ij}-l)^2+(\frac{w}{2})^2} \text{ if } d_{ij} < l$$
$$= \frac{c_{ij}}{\pi} \times \frac{\frac{w}{2}}{(\frac{w}{2})^2} \quad \text{if } l < d_{ij} \leq u$$
$$= \frac{c_{ij}}{\pi} \times \frac{\frac{w}{2}}{(d_{ij}-u)^2+(\frac{w}{2})^2} \text{ if } u < d_{ij}$$

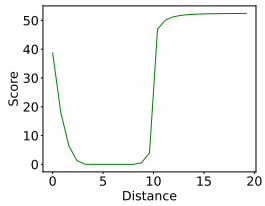
where $l = 1.5, u = 8, w = 1.0, c_{ij} = 1.5$



Soft Square

$$\sigma_{ij} = \min(\bar{w}, w) \times a + \frac{b}{\Delta^s} \text{ if } d_{ij} \geq d^0 + d^+ + r$$
$$= \Delta^e \quad \text{if } d_{ij} < d^0 + d^+ + r$$
$$\Delta = d_{ij} - (d + d^+) \quad \text{if } d_{ij} \geq d + d^+$$
$$= (d - d^-) - d_{ij} \quad \text{if } d_{ij} < d - d^-$$
$$= 0 \quad \text{if } d^- \leq d_{ij} < d^+$$

where $d^0 = 3.6, d^- = 0.1, d^+ = 4.4$
$a = 52.488, b = -75.58, s = 2.92, e = 3$
$\bar{w} = 1000, w = 1, r = 1.8$



CGLFOLD

$$\sigma_{ij} = 8^{p_{ij}}(l - d_{ij}) \quad \text{if } d_{ij} \leq l$$
$$= 8^{p_{ij}} \quad \text{if } l < d_{ij} \leq u$$
$$= 8^{p_{ij}} \ln(d_{ij} - u + 1) \text{otherwise}$$
$$= \text{where } l = 3.8, u = 8$$

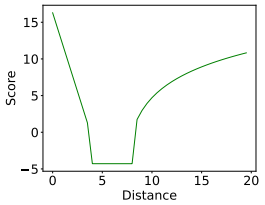


Fig. 1. Five existing contact based scoring functions

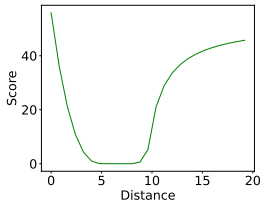
values. Moreover, both functions somewhat keep growing in the large d_{ij} values and so could provide some search guidance. The bounded potential function grows steadily in the large d_{ij} values and is expected to provide effective search guidance as it will be able to distinguish large d_{ij} values from each other.

3.2 Proposed Scoring Functions

Considering the qualitative similarity of the existing scoring functions, we choose the soft square and the bounded potential functions and create their variants. The variants will be mainly created more based on qualitative considerations than quantitative ones, particularly changing the steepness of the transition of

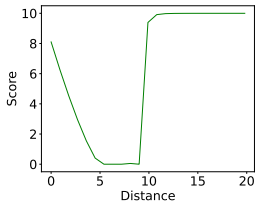
Soft Square Moderated

$$\sigma_{ij} = a + \frac{b}{\Delta} \text{ if } d_{ij} > u + r$$
$$= \Delta^{2.5} \text{ if } d_{ij} \leq u + r$$
$$\Delta = d_{ij} - u \text{ if } d_{ij} \geq u$$
$$= l - d_{ij} \text{ if } d_{ij} < l$$
where $l = 5, u = 8, r = 1$
$a = 52.488, b = -75.58$



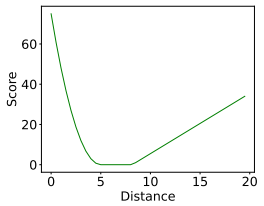
Soft Square Steepened

$$\sigma_{ij} = a + \frac{b}{\Delta^5} \text{ if } d_{ij} > u + r$$
$$= \Delta^{1.3} \text{ if } d_{ij} \leq u + r$$
$$\Delta = d_{ij} - u \text{ if } d_{ij} \geq u$$
$$= l - d_{ij} \text{ if } d_{ij} < l$$
where $l = 5, u = 8, r = 1$
$a = 10, b = -15$



Bounded Potential Moderated

$$\sigma_{ij} = m(l - d_{ij})^2 \text{ if } d_{ij} < l$$
$$= s(d_{ij} - u - 0.25) + 0.25 \text{ if } d_{ij} > u$$
$$= 0 \text{ otherwise}$$
where $l = 5, u = 8, m = 3, s = 3$



Bound Potential Steepened

$$\sigma_{ij} = m(l - d_{ij})^2 \text{ if } d_{ij} < l$$
$$= s(d_{ij} - u - 0.25) + 0.25 \text{ if } d_{ij} > u$$
$$= 0 \text{ otherwise}$$
where $l = 5, u = 8, m = 3, s = 7$

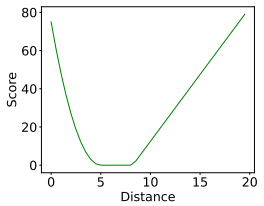


Fig. 2. Four proposed contact based scoring functions

the function from low to high for as the d_{ij} grows. The motif behind creating these variants is to study the effect of the slope of the curve on the progress of the search towards the region with the least function values. Fig. 2 shows the four proposed scoring function variants: soft square moderated (ssm), soft square steepened (sss), bounded potential moderated (bpm), and bounded potential steepened (bps). Also, these functions are somewhat simplified in their expressions compared to the original versions. The two soft square variants differ on the values of a and b and in the power of Δ while the two bounded potential variants differ on the value of m. From our intuition, we expect the bounded potential steepened variant to perform better than the other variants.

4 Experiments

We describe the experimental setup, compare the contact based scoring functions, and compare our best results with the results obtained by a recent state-of-the-art PSP search algorithm that uses a contact based scoring function.

4.1 Experimental Setup

To obtain the ϕ and ψ values for the initial conformation construction, among the available backbone angle predictor methods SAP[14], OPUS-TASS [19], and SPOT-1D [6], we run SPOT-1D, since in our pilot runs, SPOT-1D predicted values lead to better results. For SPOT-1D, the mean absolute error Δ is 16° for ϕ and 23° for ψ. To obtain secondary structure prediction of the residues, we run SSpro8 [13] and get 8-state predictions but we convert them into three states such as helices, sheets, and coils. Note that once initial conformation is obtained, ϕ and ψ angles of the coil residues get changed during search while the helix and sheet regions remain unchanged.

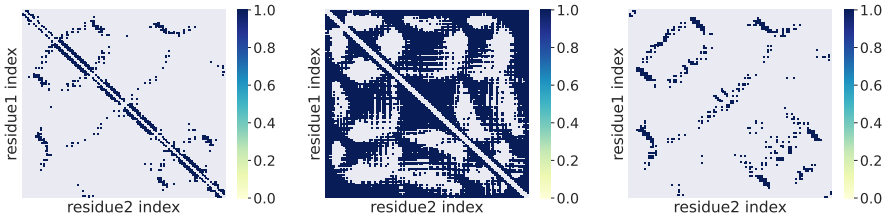


Fig. 3. Actual contact map (left), predicted contact map before filtering (middle), and predicted contact map after filtering (right) for protein `1T1J`

To obtain predicted contact for the residue pairs, we run SPOT-Contact [5]. A *contact map* is a two dimensional array showing the contact probability for each residue pair. We filter the contact map discarding contacts with probabilities below 30% and also the contacts between residues that are within the same helices or sheets and so are not changed during search. Fig. 3 shows the actual contact map for one protein `1T1J` and the predicted one before and after filtering.

To evaluate the contact based scoring functions, we use 39 proteins that have 42 to 181 residues. Out of them, 15 are α type, 13 are β type, and 11 are α/β type. These proteins have been obtained from QUARK [18], MODE-K [4], and MODCSA/CA [16] or SPOT-1D [6]. We have used CD-HIT to check for 25% sequence similarity of these proteins with the training proteins of the machine learning algorithms SPOT-1D [6], SSPro8 [13], and SPOT-Contact [5].

4.2 Comparison of Scoring Functions

Table 1 shows the mean of root mean square deviation (RMSD) values for the 39 proteins as obtained by running each of the scoring functions with our search framework 5 times. Note each run explores 160000 conformations.

As we see the results in Fig. 1, among the existing 5 scoring functions, as expected before in their descriptions, bp achieves the best results. Among all 9 scoring functions, bps function obtains the best results. Notice that bps obtains the best mean RMSD in 14 out of 39 proteins and the second best in 12 proteins. The second best scoring function among all 9 scoring functions is bpm with the best performance in 10 and the second best performance in 9 proteins.

Since bp, bpm, and bps have no flat region for the undesired d_{ij} values, they do not loose search direction and essentially perform better than other

Table 1. Top: comparison of mean RMSD values obtained by existing and proposed scoring functions; Bottom: the numbers of proteins for which scoring functions obtained mean RMSD values $\leq$ various threshold levels. The emboldened numbers are the best ones, while the underlined ones are the second-best ones among the versions.

Type	Protein	Length	sw	bp	mlp	ss	cf	ssm	sss	bpm	bps
α	5AON	48	4.12	2.91	3.97	4.06	3.56	2.79	3.37	**2.32**	4.23
	5B1A	58	8.66	7.88	10.04	9.03	9.63	7.57	7.85	6.51	**6.52**
	1SXD	91	11.13	8.66	8.81	8.13	8.55	10.98	10.47	8.42	**7.76**
	5B1N	59	5.14	5.41	4.31	4.21	4.51	4.38	4.33	4.41	**3.76**
	5COS	56	3.56	4.26	4.42	3.26	3.92	3.73	4.07	**3.00**	4.03
	5E5Y	61	10.15	9.24	**8.04**	9.03	9.83	8.21	8.45	8.80	8.10
	5FVK	82	5.40	5.26	5.27	5.75	5.96	6.61	8.24	6.09	**3.47**
	5EMX	54	5.94	6.08	5.08	6.07	5.03	5.65	6.03	**4.67**	5.37
	5TDY	42	7.18	8.61	6.81	8.61	9.92	7.71	6.94	7.44	**6.34**
	5HE9	56	6.34	6.44	6.39	6.58	6.29	**5.98**	6.19	6.06	6.68
	2O4T	90	9.38	**7.83**	10.11	9.50	7.87	9.25	9.11	9.41	9.07
	2O42	138	20.67	26.95	13.59	13.92	13.68	15.42	13.89	**11.51**	13.52
	5B5I	67	9.5	10.08	9.22	9.37	8.91	10.18	**8.40**	9.48	9.63
	5DIC	115	10.25	**7.19**	7.84	9.97	8.41	6.80	9.94	9.18	9.47
	5CKL	181	17.83	16.41	16.29	17.73	15.67	15.68	18.57	**12.83**	14.63
β	1R75	110	9.69	8.09	10.86	11.56	10.09	8.49	**7.49**	9.70	7.57
	1OKO	74	9.56	9.62	7.23	6.67	7.72	7.99	9.08	6.7	**6.43**
	2AXW	134	13.02	13.49	14.94	15.83	**11.79**	14.20	16.40	12.80	12.22
	2BT9	90	8.74	8.83	9.75	10.02	7.92	8.18	8.11	**6.23**	6.22
	2CHH	113	19.34	15.33	21.82	21.33	18.75	17.00	20.59	**13.74**	14.42
	2V33	91	9.58	7.28	13.76	9.34	8.59	10.59	11.82	8.02	**6.54**
	5AEJ	113	17.32	14.22	18.15	14.26	15.54	14.57	14.43	14.36	**14.09**
	5AOT	102	17.68	17.46	19.02	**15.16**	17.15	17.26	18.35	17.2	17.25
	5EZU	67	9.61	7.58	9.46	7.48	8.57	**6.65**	7.93	7.21	7.48
	5FUI	124	12.32	**9.08**	14.25	13.94	11.89	12.07	14.48	10.13	11.33
	5HDW	131	13.35	11.19	13.61	13.05	11.52	11.89	13.21	11.19	**10.47**
	7C28	58	7.74	8.04	8.66	8.20	8.19	6.96	6.72	6.70	**6.55**
	6WES	158	23.18	**21.15**	22.13	21.72	21.73	21.83	23.80	22.83	22.01
α/β	1CRN	46	5.1	**4.47**	5.42	4.53	5.03	5.87	4.99	5.08	5.15
	1CF7	82	8.40	7.85	8.37	5.51	5.41	4.6	7.38	8.48	**4.30**
	1IS7	84	8.70	7.32	6.85	8.37	8.30	**6.5**	8.10	8.56	7.43
	1KA8	100	12.10	11.97	11.86	10.77	11.31	10.49	11.38	**8.00**	8.1
	1MC2	122	10.46	10.49	11.01	12.33	12.05	12.19	10.01	**8.69**	9.05
	1T1J	125	9.91	7.54	9.84	7.52	8.13	7.83	6.23	6.15	**5.74**
	1Y71	112	8.47	9.68	11.76	9.10	13.54	9.49	10.98	7.11	**7.08**
	2BSE	107	14.31	10.1	14.63	11.04	11.78	11.64	9.96	**9.57**	9.97
	3BJO	100	10.30	9.86	**7.53**	12.42	9.92	11.47	10.47	8.78	8.72
	3CHB	103	12.47	**9.06**	16.49	12.85	10.61	10.58	11.93	10.43	10.58
	6CP8	163	13.67	11.61	13.81	11.84	12.21	12.22	12.28	12.69	**11.18**
	Average RMSD		10.78	9.96	10.79	10.26	10.01	9.84	10.36	9.13	**8.78**
mean RMSD $\leq$ 6Å			6	5	6	6	5	7	4	5	**8**
mean RMSD $\leq$ 9Å			14	20	16	15	19	19	18	**23**	**23**
mean RMSD $\leq$ 12Å			26	**32**	26	27	30	29	29	**32**	**32**

functions. Moreover, bps is steeper than bpm which is steeper than bp for large d_{ij} values. Arguably, greater slopes essentially push the search more towards the minimum regions of the functions. Nevertheless, ssm performs better than ss but sss performs worse than ss. The reason is sss is more flat than ss which is more

flat than ssm for large d_{ij} values. The more flat the function, the more loss of direction for the search. These are the explanations behind the performances.

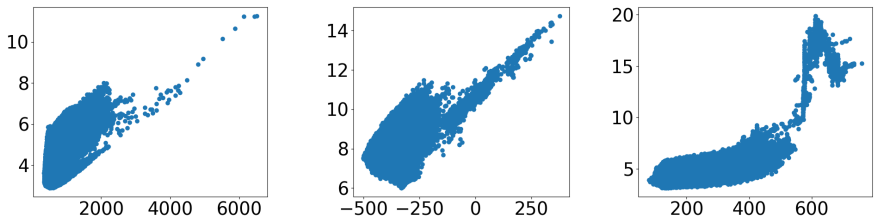
To determine the statistical significance of the performance differences of the scoring functions at 95% confidence level, we perform Friedman test and get 5.78×10^{-11} as the p value. Then, we perform Nemenyi test and show the p values in Table 2. Notice that existing functions sw and bp are significantly different but all other pairs are not significantly different from each other. On the other hand, the proposed functions are signficantly different from one another. Among other pairs, bps is significantly different from all other while bpm is not significantly different from cf. Both sss and ssm show mixed performance.

Table 2. Nemenyi test results for the scoring functions where $p \geq 0.05$ are emboldened

	bp	mlp	ss	cf	ssm	sss	bpm	bps
sw	0.04	**0.90**	**0.62**	**0.11**	0.01	**0.53**	0.00	0.00
bp		**0.20**	**0.90**	**0.90**	**0.09**	0.01	0.04	0.04
mlp			**0.90**	**0.39**	0.03	0.01	0.00	0.00
ss				**0.90**	**0.09**	**0.90**	0.02	0.00
cf					**0.09**	0.01	**0.22**	0.02
ssm						0.01	0.03	0.04
sss							0.02	0.00
bpm								0.00

Among the 9 contact based scoring functions studied, since the bps function performs the best in RMSD values, we provide its further analysis.

Fig. 4 shows the correlations between the bps scores and the RMSD values of the conformations generated during search for three proteins 5FVK, 2V33, and 1T1J. The Pearson correlation coefficients for these three proteins are 0.647, 0.454, and 0.661 respectively. These results give the evidence that improving the bps scores lead us to better conformations in terms of the RMSD values.



5FVK α type 2V33 β type 1T1J α/β type

Fig. 4. Scatter plots of bps contact based scores (x-axis) vs RMSD values (y-axis)

Fig. 5 depicts the mean RMSD values of the intial and final conformations obtained for all proteins by using the bps function during search. Clearly, the bps function, improves the quality of the conformations.

Fig. 6 shows samples of the initial and the final conformations obtained for three proteins when the bps function is used in search.

4.3 Comparison with Existing Methods

We finalize the bps function along with our search framework as our final algorithm named Contact Guided PSP Search (CGPSPS). We then compare its

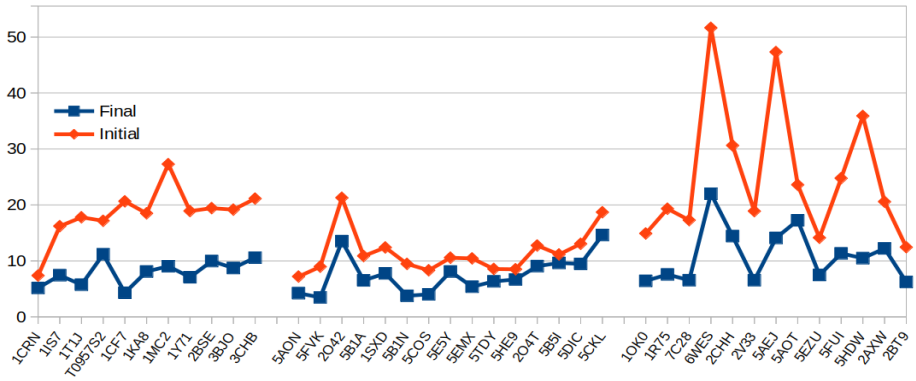
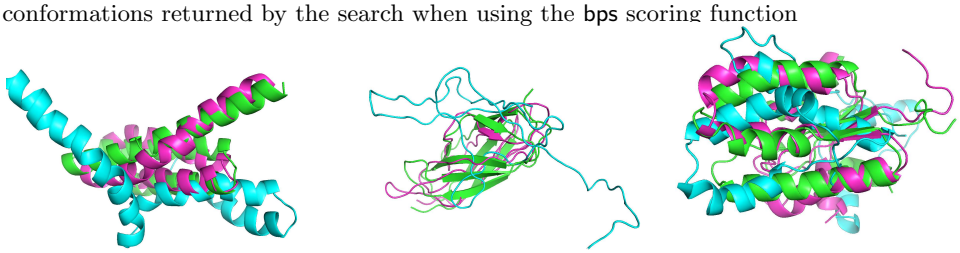


Fig. 5. Deviation in mean RMSD values of the initial conformations and the final conformations returned by the search when using the bps scoring function



5FVK Type α Length 82
RMSD Init 15.188 Final 3.377

2V33 Type β Length 91
RMSD Init 18.995 Final 3.602

1T1J Type α/β Length 125
RMSD Init 14.064 Final 4.982

Fig. 6. Sample final conformations (magenta) obtained by scoring function bps from initial ones (cyan) w.r.t. native ones (green)

performance with CGLFOLD [11] the most relevant state-of-the-art algorithm for PSP. We choose CGLFOLD becuase it uses a similar type of contact map based scoring function like ours. Also, CGLFOLD uses loop sampling and makes change to the angles in the loop residues; which is quite similar to ours.

We run both CGPSPS and CGLFOLD 5 times on each protein. Each run explores 160000 conformations before termination. This is the same termination criterion used in evaluation of CGLFOLD [11]. We take the mean RMSD and Global Distance Test (GDT) scores over the 5 runs. Note that the smaller the RMSD value, the better the performance, while the larger the GDT score, the better the performance. Also, note GDT scores are in a 0-1 scale.

Table 3 depicts that in terms of RMSD values, CGPSPS outperforms CGLFold in 31 out 39 proteins. CGLFold along with its contact based scoring function, also uses rosetta energy function. However, CGPSPS using only the contact based scoring function outperforms it. Table 3 also shows that in terms of mean GDT values, in 22 out of 39 proteins, CGPSPS performs better than CGLFold. Considering protein types, CGPSPS is better than CGLFold in 9 in RMSD values and 4 in GDT values, 12 in RMSD values and 9 in GDT values, and 10 in RMSD value and 9 in GDT values in 11 α/β, 15 α and 13 β type proteins respectively. CGPSPS obtains the best performance both in RMSD and GDT

Table 3. Mean RMSD and GDT values obtained by our algorithm and state-of-the-art CGLFOLD algorithm. The emboldened numbers are the best ones while the underlined ones are the very close second best ones.

Type	Protein	Length	Mean RMSD		Mean GDT	
			CGPSPS	CGLFOLD	CGPSPS	CGLFOLD
α	5AON	48	**4.23**	6.41	**0.63**	0.54
	5B1A	58	**6.52**	17.14	**0.49**	0.35
	1SXD	91	**7.76**	9.06	**0.44**	0.40
	5B1N	59	**3.76**	4.43	**0.6**	**0.60**
	5COS	56	4.03	**3.13**	0.60	**0.72**
	5E5Y	61	8.10	**6.03**	0.39	**0.41**
	5FVK	82	**3.47**	3.57	0.59	**0.72**
	5EMX	54	**5.37**	5.54	0.56	**0.64**
	5TDY	42	**6.34**	10	**0.50**	0.35
	5HE9	56	**6.68**	8.25	0.54	**0.59**
	2O4T	90	**9.07**	10.68	**0.39**	0.24
	2O42	138	**13.52**	13.62	**0.4**	0.27
	5B5I	67	**9.63**	9.86	**0.45**	0.34
	5DIC	115	9.47	**3.33**	**0.44**	0.38
	5CKL	181	**14.63**	14.74	**0.3**	0.19
β	1R75	110	**7.57**	13.08	**0.39**	0.18
	1OK0	74	**6.43**	7.85	**0.51**	0.38
	2AXW	134	**12.22**	15.47	**0.25**	0.19
	2BT9	90	**6.22**	6.57	**0.48**	0.47
	2CHH	113	14.42	**8.57**	0.24	**0.35**
	2V33	91	**6.54**	7.38	**0.49**	0.36
	5AEJ	113	**14.09**	17.07	**0.27**	0.23
	5AOT	102	17.25	**12.23**	0.31	**0.31**
	5EZU	67	**7.48**	7.53	0.41	**0.45**
	5FUI	124	**11.33**	11.38	**0.30**	0.23
	5HDW	131	**10.47**	12.01	0.26	**0.26**
	7C28	58	**6.55**	9.26	**0.45**	0.29
	6WES	158	22.01	**19.43**	**0.12**	0.12
α/β	1CRN	46	5.15	**4.84**	0.60	**0.65**
	1CF7	82	**4.3**	4.60	0.50	**0.60**
	1IS7	84	**7.43**	7.50	0.39	**0.51**
	1KA8	100	**8.1**	8.72	0.29	**0.40**
	1MC2	122	**9.05**	10.29	0.39	**0.47**
	1T1J	125	**5.74**	6.47	**0.48**	0.47
	1Y71	112	**7.08**	7.78	**0.42**	0.42
	2BSE	107	**9.97**	10.26	0.30	**0.34**
	3BJO	100	**8.72**	9.02	**0.40**	0.33
	3CHB	103	10.58	**8.96**	0.28	**0.35**
	6CP8	163	**11.18**	13.52	**0.20**	0.15
Mean over all proteins			**8.78**	9.45	**0.41**	0.40

values in 4, 9 and 9 proteins, respectively, in total 22 out of 39 proteins. At the bottom of Table 3, we observe that about 0.77Å average RMSD and 0.01 average GDT values improvement than CGLFOLD. We perform the Wilcoxon signed rank test with 95% confidence level and found the difference in GDT is not significant with p value 0.44 but is significant in RMSD with p value 0.02.

Table 4 shows the number of proteins in which two algorithms obtain mean RMSD values less than or equal to and mean GDT values greater than certain threshold values. In most of the protein types, CGPSPS outperforms CGLFold.

Table 4. Numbers of proteins with mean RMSD values $\leq$ various threshold values.

Algorithm	mean RMSD $\leq$ 6Å				mean RMSD $\leq$ 9Å				mean RMSD $\leq$ 12Å			
Name	α	β	α/β	all	α	β	α/β	all	α	β	α/β	all
CGPSPS	5	**0**	**3**	**8**	**10**	**6**	**7**	**23**	**13**	**8**	**11**	**32**
CGLFOLD	**5**	**0**	2	7	8	5	**7**	20	11	7	10	28

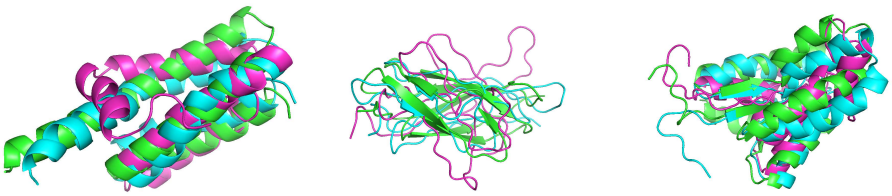


Fig. 7. Sample best conformations obtained by CGPSPS (cyan) and CGLFOLD (magenta) w.r.t. native conformations (green)

5 Conclusions

Scoring functions are crucial in protein structure prediction. Contacts between residues in given proteins are predicted by machine learning algorithms. Search algorithms then design scoring functions using the predicted contacts and compare conformations generated during search using the scoring functions. There exists a number of contact based scoring functions but they have not been compared within the same search framework on the same set of benchmark proteins. We evaluate five existing and four proposed contact based scoring functions. One of our proposed scoring function along with our search framework performs the best and significantly outperforms a similar state-of-the-art PSP search method in average root mean square distance and global distance test scores.

Acknowledgements

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

References

1. Adhikari, B., Cheng, J.: CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC bioinformatics **19**(1), 1–5 (2018)

2. Bhattacharya, D., Cao, Renzhi, C., Jianlin: UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics **32**(18), 2791–2799 (2016)
3. Brooks, B.R., Brooks III, C.L., Mackerell Jr, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al.: CHARMM: the biomolecular simulation program. Journal of computational chemistry **30**(10), 1545–1614 (2009)
4. Chen, X., Song, S., Ji, J., Tang, Z., Todo, Y.: Incorporating a multiobjective knowledge-based energy function into differential evolution for protein structure prediction. Information Sciences **540**, 69–88 (2020)
5. Hanson, J., Paliwal, K., Litfin, T., Yang, Y., Zhou, Y.: Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics **34**(23), 4039–4045 (2018)
6. Hanson, J., Paliwal, K., Litfin, T., Yang, Y., Zhou, Y.: 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), 2403–2410 (2018)
7. Hou, J., Wu, T., Cao, R., Cheng, J.: Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins: Structure, Function, and Bioinformatics **87**(12), 1165–1178 (2019)
8. Leaver-Fay A, T.M., SM, L.: ROSETTA3: an object- oriented software suite for the simulation and design of macromolecules. Methods Enzymol **487**, 545–574 (2011)
9. Li, Y., Zhang, Bell2, Chengxin Eric W., Y., Dong-Jun, Zhang, Y.: Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins **87**, 1082–1091 (2019)
10. Li, Y., Zhang, C., Bell, E.W., Zheng, W., Zhou, X., Yu, D.J., Zhang, Y.: Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLoS Computational Biology **17**, 1–19 (2021)
11. Liu, J., Zhou, X.G., Zhang, Y., Zhang, G.J.: CGLFold: a contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm. Bioinformatics **36**(8), 2443–2450 (2020)
12. Mabrouk, M., Werner, T., Schneider, T., Putz, I., Brock, O.: Analysis of free modelling predictions by RBO aleph in CASP11. Proteins (84), 87–104 (2015)
13. Magnan, C.N., Baldi, P.: SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics **30**(18), 2592–2597 (2014)
14. Mataeimoghadam, F., Newton, M.H., Dehzangi, A., Karim, A., Jayaram, B., Ranganathan, S., Sattar, A.: Enhancing protein backbone angle prediction by using simpler models of deep neural networks. Scientific Reports **10**(1), 1–12 (2020)
15. Newton, M.H., Pham, D.N., Sattar, A., Maher, M.: Kangaroo: An efficient constraint-based local search system using lazy propagation. In: International Conference on Principles and Practice of Constraint Programming. pp. 645–659. Springer (2011)
16. Ramyachitra, D., Ajeeth, A.: MODCSA-CA: a multi objective diversity controlled self adaptive cuckoo algorithm for protein structure prediction. Gene Reports **8**, 100–106 (2017)
17. Skwark, M.J., Abdel-Rehim, A., Elofsson, A.: PconsC: combination of direct information methods and alignments improves contact prediction. Bioinformatics **29**(14), 1815–1816 (2013)

18. Xu, D., Zhang, Y.: Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Structure, Function, and Bioinformatics **80**(7), 1715–1735 (2012)
19. Xu, G., Wang, Q., Ma, J.: OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. Bioinformatics **36**(20), 5021–5026 (2020)