Adhikari and Cheng *BMC Bioinformatics* (2018) 19:22
DOI 10.1186/s12859-018-2032-6

BMC Bioinformatics

SOFTWARE

Open Access

CONFOLD2: improved contact-driven ab initio protein structure modeling

Badri Adhikari[1] and Jianlin Cheng[2*]

Abstract

Background: Contact-guided protein structure prediction methods are becoming more and more successful because of the latest advances in residue-residue contact prediction. To support contact-driven structure prediction, effective tools that can quickly build tertiary structural models of good quality from predicted contacts need to be developed.

Results: We develop an improved contact-driven protein modelling method, CONFOLD2, and study how it may be effectively used for ab initio protein structure prediction with predicted contacts as input. It builds models using various subsets of input contacts to explore the fold space under the guidance of a soft square energy function, and then clusters the models to obtain the top five models. CONFOLD2 obtains an average reconstruction accuracy of 0.57 TM-score for the 150 proteins in the PSICOV contact prediction dataset. When benchmarked on the CASP11 contacts predicted using CONSIP2 and CASP12 contacts predicted using Raptor-X, CONFOLD2 achieves a mean TM-score of 0.41 on both datasets.

Conclusion: CONFOLD2 allows to quickly generate top five structural models for a protein sequence when its secondary structures and contacts predictions at hand. The source code of CONFOLD2 is publicly available at https://github.com/multicom-toolbox/CONFOLD2/.

Keywords: Contacts, Protein folding, CONFOLD, Model selection

Background

The most successful ab initio protein structure methods, i.e. fragment-assembly based methods, require generating a lot of decoys to deliver accurate predictions. Methods that can build models faster and are more residue contact sensitive are needed to realize the promise of ab initio protein structure prediction driven by the recent advances in contact prediction [1, 2]. The CONFOLD method [3] can build high quality secondary structures (also pairing beta-strands to form beta-sheets) and correct tertiary structures when predicted contacts are accurate. It is integrated into other protein structure prediction methods like CoinFold [4] and PconsFold2 [2]. In this paper, we develop an improved version of CONFOLD by incorporating a soft-square energy function into CONFOLD, building models using multiple sub-sets of contacts, adding model selection capability, and rigorously testing it on various datasets including the Critical Assessment of protein Structure Prediction (CASP) 11 and 12 datasets. CONFOLD2 also addresses a major limitation of the CONFOLD method, i.e. generating a decoy of 200 models and not producing top one or top five models. Compared to fragment-assembly methods that need to generate thousands of model decoys [5], CONFOLD2 explores the fold space by generating just a few hundred model decoys, and hence it runs relatively fast.

Implementation

Recently, it is found that energy functions that do not penalize unsatisfied predicted contacts after certain distance threshold yield more accurate model reconstruction [5–7]. Different contact energy functions like FADE [5], square-well function with exponential decay [6], and modified Lorentz potential [7] applied to contact-guided

*Correspondence: chengji@missouri.edu
[2]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
Full list of author information is available at the end of the article

protein folding have been found to work best for various folding algorithms, mostly fragment-assembly based methods. When distance geometry based approaches are used to fold proteins with restraints, it has been shown that soft-square function performs best, with the 'rswitch' parameter to be tuned [8].

$$E_{\text{contact}} = min(ceil, w) \begin{cases} a + \frac{b}{\Delta^{\text{softexp}}}, & \text{R} \geq d + d_{\text{plus}} + r_{\text{sw}} \\ \Delta^{\text{exp}}, & \text{R} < d + d_{\text{plus}} + r_{\text{sw}} \end{cases} \quad (1)$$

$$Error(\Delta) = \begin{cases} R - (d + d_{\text{plus}}), & \text{R} \geq d + d_{\text{plus}} \\ (d - d_{\text{minus}}) - R, & \text{R} < d - d_{\text{minus}} \\ 0, & otherwise \end{cases} \quad (2)$$

We replaced CONFOLD's [3] soft-square asymptotic energy function (designed originally for the experimental NOE restraints) with the soft-square function (Eq. (1)), where the error is defined in Eq. (2). The parameter d, d_{minus}, and d_{plus} define the interval $[\text{d-d}_{\text{minus}}, \text{d+d}_{\text{plus}}]$, where the error is zero. For contacts predicted to be less than 8 Å distance, we set d, d_{minus}, and d_{plus} to 3.6, 0.1, and 4.4 respectively. The switching parameter r_{sw} defines the boundary where the square error function starts to taper into a constant error (see Fig. 1). R is the actual distance between Cβ atoms of the predicted contact residue pair in the model. The exponents, 'exp' and 'softexp' are both set to 2. Since the contact weight multiplies the energy term, the maximum weight (ceil) that any pair of predicted contacts can have is set to 1000, and 'w' is the weight of each contact pair and is set to 1. The most important parameter affecting the quality of reconstruction is r_{sw} and we optimized it to be 1.8. 'a' and 'b' are constants determined at run-time such that the function is smooth at r_{sw} equal to 1.8. Our soft-square contact energy term is calculated either using a square error function or approximately constant error function based on a switching parameter - r_{sw}. It defines a threshold until which the error increases as a square error function and beyond which the error tapers to a constant error. Figure 1 demonstrates how the switching parameter affects the overall energy calculations.

Using the soft-square function as contact energy term, CONFOLD2 initially predicts 200 models using various subsets of input contacts, and selects five top models by clustering them. To effectively explore the fold space captured by the predicted contacts, we prepare 40 different subsets of input contacts by selecting top xL contacts, where x = 0.1, 0.2, 0.3, ..., 4.0 and L is length of the protein, and build 20 models for each subset. For each subset of contacts, top-five models in the second stage of CONFOLD modeling are selected based on the contact energy score, resulting in a total of 200 models. Next, to filter out unfolded models, we rank these 200 models by calculating their contact satisfaction score using top L/5 long-range contacts, and filter out the bottom 150 models. The remaining 50 models are clustered into five clusters by calculating their pairwise structural similarity measured by TM-score. We select the five models closest to the centroids of these five clusters as the top five predictions with the rank determined by the satisfaction score of the top L/5 long-range contacts. SCRATCH suite [9] is used to predict three-state secondary structure and Maxcluster [10] to compute pairwise model similarity for clustering.

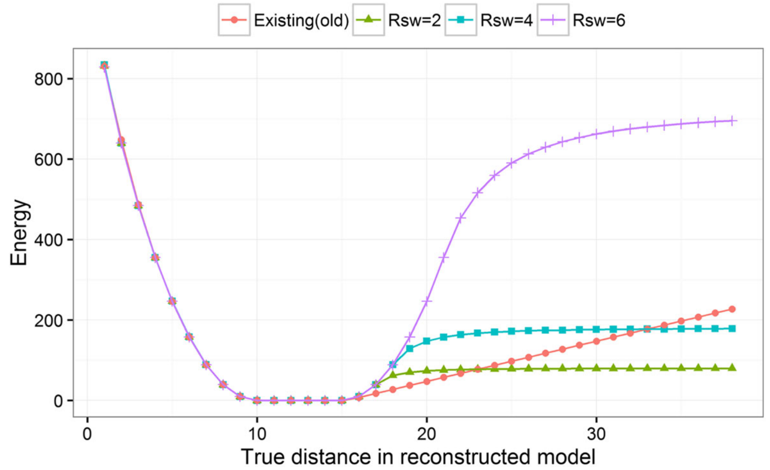

Fig. 1 Behavior of the contact energy term for various r_{sw} values. For this demonstration desired distance is set to 10 Å with a lower-bound of 0 Å and upper-bound of 5 Å, i.e. the desired distance between the pair of restrained residues is 10.0 Å and 15.0 Å. The "Existing" energy calculations refers to the old energy term implemented in CONFOLD method. The plot shows that depending upon the switching parameter, r_{sw}, the energy calculations can taper early at around 1 or 2 Å for $\text{r}_{\text{sw}} = 2$ or at more than 25 Å for $\text{r}_{\text{sw}} = 6$

Results

As the first benchmark, we compared the performance of CONFOLD2 with the original CONFOLD method [3] on the 150 proteins in the PSICOV dataset [11] using the contacts predicted by the PSICOV method [11] (see Table 1). The original CONFOLD method generates top 200 models and provides no ranking of the reconstructed models, so we compare the two methods using best-of-200 models. On the PSICOV dataset, when best of 200 models are evaluated, CONFOLD2 achieves a mean TM-score of 0.57 compared to 0.55 of CONFOLD. This improvement in CONFOLD2 is statistically significant per paired t-test with a p-value of 4×10^{-8} (see Additional file 1: Table S1 for a detailed comparison).

Next, to evaluate our model selection technique (selecting the top five models from 200) we compared our approach of model selection using clustering with the model ranking using contact satisfaction score only. On the same dataset, when we selected the top five models using contact satisfaction score of top L/5 or L/2 long-range contacts, we achieved best-of-top-five TM-score of 0.50. The rationale for using top L/5 or L/2 contacts (instead of L or more) is that these subsets are found to best reflect the accuracy of the predicted contacts [12]. In contrast, when we filter out the bottom 150 models, cluster the remaining 50 into five clusters, and select the cluster centroids, we obtain best-of-top-five TM-score of 0.52, suggesting that the clustering approach is effective in selecting models built from contacts. As summarized in Table 1, we also reconstructed models for the PSICOV-150 dataset using contacts predicted by MetaPSICOV [13] and obtained a mean TM-score of 0.62 when best of top-five models are evaluated (see Additional file 1: Table S1 for detailed results), indicating that the improved contact prediction leads to the better tertiary structure reconstruction.

For a more rigorous evaluation, on the PSICOV-150 dataset, we compared CONFOLD2's performance with two other state-of-the-art modeling methods that use structural template fragments. First, we evaluated the top-one models built using CONFOLD2 and compared against the models built using Rosetta in the PconsFold method [14]. When top one models are evaluated, CONFOLD2's average TM-score is 0.48 compared to 0.55 using the PconsFold method. Second, upon comparing the performance of CONFOLD2's best-of-top-five models against the best-of-top-five models built using another fragment-based method, FRAGFOLD [15], we find that CONFOLD2's mean TM-score is slightly higher (0.57 vs 0.54). Compared to the FRAGFOLD method which could recover the correct fold (with TM-score ≥ 0.5) of 100 out of 150 proteins [15], CONFOLD2 recovered the correct fold of 107 proteins. This further supports the improved performance of CONFOLD2 over the FRAGFOLD method. It is worth noting that CONFOLD2 does not use any structural template fragment information.

Finally, using CONFOLD2, we predicted models for the protein sequence targets in the CASP11 and CASP12 datasets with contacts predicted by the most accurate predictor in each of the CASP experiments - CONSIP2 [16] in CASP11 and Raptor-X [17] in CASP12 (see Table 1). The average TM-score of the reconstructed models for both datasets (CASP 11 and 12) is 0.46 when best-of-200 models are evaluated and 0.41 when best-of-five models are evaluated. We obtained the predicted contacts from the official CASP website www.predictioncenter.org. It is worth noting that the latest results of RaptorX on the CASP12 targets are better which can be found in [17].

Discussion

Observing the lower reconstruction accuracy for the CASP datasets compared to the PSICOV dataset, we

Table 1 Summary of the performance of CONFOLD2 on PSICOV, CASP11, and CASP12 datasets

Dataset	Contact Precision (L/5)			TM-score of Models	
	Method	$P_{SR+MR+LR}$	P_{LR}	Best-of-200	Best-of-5
PSICOV-150	PSICOV	72.6	64.0	0.57	0.52
PSICOV-150	MetaPSICOV	88.4	77.2	0.65	0.62
CASP12 all	Raptor-X	71.3	58.6	0.46	0.41
CASP12 single domain	Raptor-X	70.6	58.6	0.49	0.44
CASP12 multi-domain	Raptor-X	72.0	58.7	0.44	0.38
CASP11 all	CONSIP2	71.8	50.2	0.46	0.41
CASP11 single domain	CONSIP2	75.8	57.4	0.52	0.48
CASP11 multi-domain	CONSIP2	67.7	42.4	0.40	0.34

Mean contact precision of top L/5 for (i) all (short-range, medium-range, and long-range: $P_{SR+MR+LR}$) contacts, and (ii) long-range contacts (P_{LR}) is reported for all the datasets. The TM-score of the best-of-200 and best-of-5 models reconstructed by CONFOLD2 are also presented. Results for single-domain and multi-domain subsets of the CASP11 and CASP12 datasets are also reported separately investigated if the performance was affected by multi-domain proteins because we build models for the whole targets first and evaluated them at domain level. As shown in Table 1, the reconstruction accuracy is higher for single domain proteins than multi-domain proteins (see Additional file 1: Table S2 and S3 for details). Yet, the reconstruction accuracy for single domain proteins is still lower than that of the PSICOV dataset. For the further investigation, from the single domain proteins in both CASP11 and 12 datasets, we removed some proteins with low accuracy contact predictions so that both datasets have the mean contact precision of top L/5 long-range contacts the same as that of the PSICOV dataset, i.e. precision = 64%. On such reduced datasets, the average TM-scores of the best-of-200 models for the CASP11 and 12 proteins are 0.55 and 0.52 respectively, which are slightly lower than the mean TM-score for PSICOV dataset (0.57). Since TM-score of 0.5 is the threshold if the topology of a protein structure is correctly predicted, for all three datasets, it can be concluded that the fold of single domain proteins can be reconstructed correctly (TM-score $\geq$ 0.5) on average if the precision of predicted long-range contacts is at least 64%. Although the sequence lengths of the domains in the CASP datasets are much higher than the PSICOV-150 dataset, which have up to 500 residues, we did not find any substantial correlation between the domain length and the reconstruction accuracy. For a discussion on relationship between the precision of predicted contacts and the accuracy of reconstructed models see our recent work at [12].

A head-to-head comparison of CONFOLD2 and CONFOLD1 shows that for some proteins in the PSICOV dataset of 150 proteins, CONFOLD2's reconstruction accuracy is slightly worse than that of CONFOLD. For instance, for the protein ID '1g2r' the reconstruction TM-score of CONFOLD2's best of 200 model is 0.49 whereas one reconstructed using CONFOLD is 0.58. To investigate the possible reasons of poor performance for these proteins we checked if the improvement from two-stage modeling in CONFOLD2 is not as much pronounced as in CONFOLD due to the implementation of the new energy function in CONFOLD2. For this, we analyzed the cases in which CONFOLD2 outperforms/underperforms CONFOLD and the cases where second stage models are better/worse than the first stage models. Overall, we did not observe any meaningful correlations. In summary, our results suggest that CONFOLD2 is better than CONFOLD, in general. However, if the purpose of modeling is to explore the fold space (at the expense of computing resources) then running both versions of CONFOLD may be slightly helpful. On the other extreme, if CONFOLD2 is being used for large-scale modeling, skipping the second stage modeling can save 50% of computing resources at the expense of around 0.5 lower TM-score, on average.

Conclusions

We have developed CONFOLD2, a method for building three-dimensional protein models using predicted contacts and secondary structures. It explores the fold space captured in predicted contacts by creating various subsets of predicted contacts and builds decoy sets, and then clusters the decoys to obtain the top five models. CONFOLD2 is significantly better than the original CONFOLD method. Structure predictions using some recently available contact prediction datasets, show that the for most protein sequences CONFOLD2 is able to capture the structural fold of the protein.

Availability and requirements

Project name: CONFOLD2
Project home page: https://github.com/multicom-toolbox/CONFOLD2
Operating systems: Platform independent
Programming language: Perl
Other requirements: Perl interpreter, CNS suite, TM-score (included), MaxCluster (included), and DSSP (included)
License: GNU GPL
Any restrictions to use by non-academics: None

Additional file

Additional file 1: Supplementary Tables. Docx file containing various tables with detailed results. (PDF 167 kb)

Abbreviations

CASP: Critical assessment of protein structure prediction; Cβ: Carbon-β atom of a residue; L: Length of protein sequence; NOE: Nuclear overhauser effect

Acknowledgements

Some of the computation for this work was performed on the high performance computing infrastructure provided by Research Computing Support Services and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri, Columbia MO). We would like to thank the RCSS team for their infrastructure and technical support.

Funding

This work has been supported in part by the US National Institutes of Health (NIH) grant (R01GM093123) to JC.

Availability of data and materials

The datasets generated results obtained during the current study are available at http://sysbio.rnet.missouri.edu/bdm_download/confold2/.

Authors' contributions

JC supervised the project. BA designed and implemented the method, performed the experiments and analysed the result. JC revised the manuscript. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1] Department of Mathematics and Computer Science, University of Missouri-St. Louis, St. Louis, MO 63121, USA. [2] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA.

Received: 19 November 2017 Accepted: 17 January 2018
Published online: 25 January 2018

References

1. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. Proteins Struct Funct Bioinforma. 2016;84(S1):131–44.
2. Michel M, Hurtado DM, Uziela K, Elofsson A. Large-scale structure prediction by improved contact predictions and model quality assessment. bioRxiv. 2017;128231.
3. Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins. 2015;83(8):1436–49.
4. Nilges M, Gronenborn AM, Brünger AT, Clore GM. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. Protein Eng Des Sel. 1988;2(1):27–38.
5. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. Pconsfold: improved contact predictions improve protein models. Bioinformatics. 2014;30(17):482–8.
6. Kosciolek T, Jones DT. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS ONE. 2014;9(3): e92197.
7. Mabrouk M, Werner T, Schneider M, Putz I, Brock O. Analysis of free modeling predictions by rbo aleph in casp11. Proteins Struct Funct Bioinforma. 2016;84(S1):87–104.
8. Nilges M, Gronenborn AM, Brünger AT, Clore GM. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. Protein Eng Des Sel. 1988;2(1):27–38.
9. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. 2005;33(Web Server):72–6.
10. Siew N, Elofsson A, Rychlewski L, Fischer D. Maxsub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 2000;16(9):776–85.
11. Jones DT, Buchan DWA, Cozzetto D, Pontil M. Psicov: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012;28(2):184–90.
12. Adhikari B, Nowotny J, Bhattacharya D, Hou J, Cheng J. Coneva: a toolbox for comprehensive assessment of protein contacts. BMC Bioinformatics. 2016;17(1):517.
13. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 2014;31(7):791.
14. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. Pconsfold: improved contact predictions improve protein models. Bioinformatics. 2014;30(17):482–8.
15. Kosciolek T, Jones DT. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLOS ONE. 2014;9(3):1–15.
16. Kosciolek T, Jones DT. Accurate contact predictions using covariation techniques and machine learning. Proteins Struct Funct Bioinforma. 2016;84(S1):145–51.
17. Wang S, Sun S, Xu J. Analysis of deep learning methods for blind protein contact prediction in casp12. Proteins Struct Funct Bioinforma. 2017;00(00):000–000.