PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

CONFOLD: Residue-residue contact-guided *ab initio* protein folding

Badri Adhikari, Debswapna Bhattacharya, Renzhi Cao, and Jianlin Cheng*

Department of Computer Science, University of Missouri, Columbia Missouri 65211

ABSTRACT

Predicted protein residue–residue contacts can be used to build three-dimensional models and consequently to predict protein folds from scratch. A considerable amount of effort is currently being spent to improve contact prediction accuracy, whereas few methods are available to construct protein tertiary structures from predicted contacts. Here, we present an *ab initio* protein folding method to build three-dimensional models using predicted contacts and secondary structures. Our method first translates contacts and secondary structures into distance, dihedral angle, and hydrogen bond restraints according to a set of new conversion rules, and then provides these restraints as input for a distance geometry algorithm to build tertiary structure models. The initially reconstructed models are used to regenerate a set of physically realistic contact restraints and detect secondary structure patterns, which are then used to reconstruct final structural models. This unique two-stage modeling approach of integrating contacts and secondary structures improves the quality and accuracy of structural models and in particular generates better β-sheets than other algorithms. We validate our method on two standard benchmark datasets using true contacts and secondary structures. Our method improves TM-score of reconstructed protein models by 45% and 42% over the existing method on the two datasets, respectively. On the dataset for benchmarking reconstructions methods with predicted contacts and secondary structures, the average TM-score of best models reconstructed by our method is 0.59, 5.5% higher than the existing method. The CONFOLD web server is available at http://protein.rnet.missouri.edu/confold/.

Proteins 2015; 00:000–000.

Key words: protein residue-residue contacts; protein structure modeling; ab initio protein folding; contact assisted protein structure prediction; optimization.

INTRODUCTION

Emerging success of residue–residue contact predictions[1–16] and secondary structure predictions[17–23] demands more research on how predicted contacts and secondary structures may be directly used for predicting protein structures from scratch without using structural templates (template-free/*ab initio* modeling). Some experiments have been performed to study if accurate protein structures can be reconstructed using true contacts, providing strong evidences that contacts contain crucial information to reconstruct protein tertiary structures.[24–31] However, all of these reconstruction methods, including most recent ones, Reconstruct[25] based on Tinker[32] and C2S[33] based on FT-COMAR,[26] focus on using all true contacts rather than predicted, noisy, incomplete contacts, to construct three-dimensional structures. Thus, these methods generally cannot effectively use contacts predicted by practical contact prediction methods to build realistic protein structure models. Additionally, these reconstruction methods do not take into account secondary structure information, which is complementary with contacts and is very valuable for various protein structure prediction tasks. Therefore, robust reconstruction methods need to be developed to deal with real-world, predicted contacts and secondary structures to reconstruct protein structure models from scratch, which is still a largely unsolved problem.

Computational modeling tools like IMP[34] and Tinker[32] can accept different kinds of generic distance restraints, but they are not specifically designed to effectively handle noisy and incomplete contacts predicted from protein sequences and cannot build high-quality

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: NIH; Grant number: R01GM093123.

*Correspondence to: Jianlin Cheng, Department of Computer Science, University of Missouri, Columbia, MO 65211. E-mail:chengji@missouri.edu

Received 29 January 2015; Revised 11 April 2015; Accepted 2 May 2015
Published online 13 May 2015 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24829

secondary structures from these predicted information. The widely used modeling tool, Modeller,[35] can accept contacts and secondary structure information as restraints, and can be used for reconstruction, but its optimization process and energy function are primarily designed for template-based modeling and cannot best utilize incomplete, inaccurate, and predicted contacts for *ab initio* modeling. Most recent research[9,36] used the Crystallography & NMR System (CNS),[37,38] a method designed for building models from Nuclear magnetic resonance (NMR) experimental data, to reconstruct protein models from predicted contacts. However, the method does not reconstruct secondary structures well and cannot effectively handle noisy self-conflicting contacts.

To predict new protein folds using contact-guided protein modeling, we need an integrated reconstruction pipeline, which accepts contacts, secondary structure information and β-sheet pairing information as inputs and builds three-dimensional models. In this article, we develop a two-stage contact-guided protein folding method, CONFOLD, to synergistically integrate contacts, secondary structures, and β-sheet pairing information in order to improve *ab initio* protein modeling. Different from previous contact-based reconstruction method[32] that uses only distance restraints to encode secondary structures, we translate secondary structures into distance restraints, dihedral angles, and hydrogen bonds according to a set of new conversion rules, which leads to the improvement of overall topology and secondary structures in reconstructed models. In the first modeling stage, the initial contact-based distance restraints and secondary structure-based restraints are first used to reconstruct protein models. The reconstructed models are used to filter out unsatisfied contacts and detect beta-pairings. The remaining contacts realized in the models, beta-pairings detected in the models, and initial secondary structures are then used to regenerate restraints to build model in the second modeling stage. Reconstructing models in the second stage, not used by previous contact-based modeling methods, substantially improves the quality of modeling.

MATERIALS AND METHODS

Data sets and contact definitions

We used two standard protein data sets for our experiments: **(1)** 15 test proteins of different fold classes ranging from 48 to 248 residues used in EVFOLD,[9] and **(2)** 150 diverse globular proteins with average length of 145 residues used in FRAGFOLD.[39] For the 15 proteins in EVFOLD benchmark set, average precision of top 50 predicted contacts is 0.65 and within the range [0.38, 0.86] and predicted secondary structure have average accuracy (Q3 score) of 0.84 and within the range [0.56, 0.96]. Similarly, for the 150 globular proteins in FRAGFOLD benchmark set, the average precision of top L predicted contacts is 0.6 (with minimum 0.13 and maximum 0.93) and predicted secondary structure have average accuracy (Q3 score) of 0.84 (with minimum 0.63 and maximum 0.95). We also specifically tested our method's capability of reconstructing secondary structures on an antiparallel beta barrel protein 2QOM, and a classic beta-alpha-beta barrel protein 1YPI.

Consistent with the previous convention, we used three definitions of residue–residue contacts. On the EVFOLD benchmark dataset, two residues are considered in contact if the distance between their Cα atoms is less than or equal to 7 Å as defined in Ref. 9. On the FRAGFOLD data set, a residue pair is considered a contact if the distance between the Cβ-Cβ atoms of the two residues is at most 8 Å (Cα in case of Glycine) as defined in Ref. 39. For all reconstructions using true contacts, we define a pair of residues to be in contact if the two residues have sequence separation of at least 6 residues in the protein sequence and the distance between their Cβ atoms is less than or equal to 8 Å. To denote number of contacts ranked by prediction confidence that are used for reconstruction, we use the notation *x*L (*x* times L), where *x* ranges from 0.4 to 2.2 at step of 0.2 and L is length of a protein sequence. For example, for a protein having 100 residues, top-0.4 L contacts would refer to the top 40 (0.4 times 100) contacts.

Deriving restraints for building helices, strands, and β-sheets for contact-based modeling

One big challenge in contact-based protein modeling is to reconstruct realistic secondary structures since limited residue–residue contacts information is generally not sufficient and detailed enough for building all secondary structures. To do so, we derived dihedral angles (phi and psi), hydrogen bond distances, and various distances between backbone atoms (O, N, Cα, C) with upper and lower bounds for residues in different kinds of secondary structures from tertiary structures of the proteins in SABmark database[40] in order to use them to translate secondary structures into restraints. Since building helices with dihedral angles and hydrogen bond distance restraints (between ith and i + 4th residue) together with contact restraints did not guarantee to produce helices in the final models according to our experiment, especially when helices are long, we derived backbone atom restraints for helices as well. We also discovered that relative positions of backbone oxygen atoms in each residue along the strands was a key restraint in addition to the dihedral angle restraints to build parallel, antiparallel, and mixed β-sheets. Adding these relative oxygen positioning restraints substantially increases the chance of forming β-sheets in the models when contacts are used to drive protein model reconstruction. Another

important restraint for building β-sheets is the backbone atom to backbone atom distance between a residue on one side of the hydrogen bond and two neighboring residues on the other side. Interestingly, by deriving and using β-sheet restraints in this way, the right-handed twist property of β-sheets[41,42] is automatically preserved.

On the basis of the rationale and experiments described above and considering only ideally hydrogen-bonded helices and β-sheets in each tertiary structure in the SABmark database, we derived the following secondary structure restraints: (a) hydrogen bond distance between backbone atoms, O and H, (b) Cα—Cα, N—N, O—O, and C—C distances between the hydrogen bonded residues, (c) Cα—Cα, N—N, O—O, and C—C distances between hydrogen bonded residue on one side and two neighbor residues (±1 sequence separation) on the other side, (d) dihedral angles (phi and psi), and (e) O—O distance between the adjacent backbone oxygen atoms in strands. The symbols Cα, Cβ, N, O, and H are used to denote backbone carbon-alpha, carbon-beta, nitrogen, oxygen, and hydrogen atoms, respectively. On the basis of these restraints, in Table I, for a helix of 10 residues 107 restraints in total were derived, including 20 dihedral angle restraints, 7 hydrogen bond restraints, and 80 backbone atom restraints. Similarly, for a pair of strands, each 10 residues long, connected as antiparallel, 108 restraints were derived, including 20 dihedral restraints and 9 O—O backbone distance restraints for each strand, 10 hydrogen bond restraints, and 40 backbone atom restraints. Assuming these restraints measurements to be normally distributed, we tried various values of a scaling factor (λ) times the standard deviation (σ) to get different lower and upper bounds (range) of the measurements to build helices and β-sheets. When true contacts were used along with secondary structure information we set $\lambda = 1.0$ and when predicted information were used we set $\lambda = 0.5$. All the restraints were translated according to the exact values in Table I except for hydrogen bonds involving prolines. As proline's backbone nitrogen atom is not bound to any hydrogen, we translated all hydrogen bond restraints involving proline hydrogen atom to proline nitrogen atom and increased the distance by 1 Å.

Two-stage model building and contact filtering

Figure 1 shows our two-stage contact-guided protein modeling process (CONFOLD). In the first stage, secondary structures are converted into distance, dihedral angle, and hydrogen bond restraints as described in Section "Deriving restraints for building helices, strands, and β-sheets for contact-based modeling", and contacts into the range [3.5 Å, threshold]. One key issue is to decide how many contacts should be used to build models. To estimate the number of contacts needed for reconstruction, we scanned the structures in the Protein Data Bank (PDB)[43] and found that 99% of known 3D structures have <3 L true contacts, and more than 50% of them have less than 2 L (L: length of a protein) true contacts. And based on our test on 15 proteins in EVFOLD benchmark set, less than 1.6 L predicted contacts yielded best results. Therefore, for each protein, we built 20 models for each contact sets consisting of top 0.4 L, 0.6 L, 0.8 L, ... up to 2.2 L contacts. The models were constructed from these restraints by a customized distance geometry algorithm implemented in CNS ("Customization of distance geometry protocol for contact-based model generation" section). These models are used to filter out noisy contacts and detect strand pairings for the second round of modeling.

In the second-stage of model reconstruction (Fig. 1), we updated the contact information as well as the β-sheet information by analyzing the model having minimum restraints energy in the first stage. Specifically, we filter out contacts of which no two atoms of the two residues are within the contact distance threshold. We also identify the beta strands close to each other in the model, and then add β-strand pairing restraints ("Detection of β-sheets in structural models" section for details). The newly filtered contact restraints, the new strand pairing restraints, and the restraints derived from secondary structures are used to build tertiary structure models again. We experimented with two weighting schemes for residue contact restraints and secondary structure restraints (that is, the ratio between weights of contact restraints and secondary structures is either 1:5 or 1:0.5) to generate diverse models. Unlike existing methods[9,39] that weight the contacts considering the confidence of prediction to build models, we assign the same weight to all contact restraints or secondary structure restraints. Hence, for each of 10 sets of different contacts and each of two weighting schemes, 20 models were generated. In total, a pool of 400 models was reconstructed for a protein in each stage. The 400 models in the second stage were considered as final predictions.

Detection of β-sheets in structural models

For detecting strand-pairs in the models built in the first stage, we compute the distances between all the strands in the top model with the minimum restraint energy, and rank all pairs by the distances and select closest strands as pairs. To calculate the distance between a pair of strands of equal lengths, we consider ten antiparallel ideal hydrogen-bonding patterns and ten parallel hydrogen-bonding patterns (Fig. 2). We compute the distance between the strand pairs for all of these possible patterns and select the pattern with minimum distance. We define the distance between two equal-length strands (residues: a–b and residues: c–d) as the minimum of the

Table I

Upper bounds and Lower Bounds of Hydrogen Bond and Oxygen–Oxygen Distance, Dihedral Angle and Backbone Atom-Backbone Atom Distance Measurements Derived from the SABmark Database with $\lambda = 0.5$ for Reconstructing Alpha Helices, Strands and β-Sheets

Table (A)

Type	LB	UB
A	1.8	2.0
P	1.8	2.0
H	1.9	2.1

Table (B)

Type	LB	UB
A	4.5	4.7
P	4.5	4.7
U	4.5	4.7

Table (C)

Type	Angle	LB	UB
A	PSI	128.2	145.6
A	PHI	−131.9	−109.9
P	PSI	122.6	139.3
P	PHI	−125.2	−104.8
U	PSI	126.1	143.8
U	PHI	−129.8	−108.0
H	PSI	−46.4	−36.6
H	PHI	−68.1	−58.9

Table (D)

Type	A1-A2	Ref	N	LB	UB
A	O—O	O	1	7.4	8.0
A	O—O	O	−1	4.7	4.9
A	O—O	O	0	3.5	3.7
A	O—O	H	1	7.5	8.1
A	O—O	H	−1	4.7	5.1
A	O—O	H	0	3.4	3.8
A	C—C	O	1	7.4	8.0
A	C—C	O	−1	4.7	4.9
A	C—C	O	0	4.9	5.1
A	C—C	H	1	7.4	8.0
A	C—C	H	−1	4.7	5.1
A	C—C	H	0	4.9	5.1
A	N—N	O	1	4.9	5.3
A	N—N	O	−1	6.7	7.1
A	N—N	O	0	4.3	4.5
A	N—N	H	1	4.9	5.1
A	N—N	H	−1	6.7	7.1
A	N—N	H	0	4.3	4.5

Type	A1-A2	Ref	N	LB	UB
A	Cα-Cα	O	1	6.2	6.6
A	Cα—Cα	O	−1	5.6	5.8
A	Cα—Cα	O	0	5.2	5.4
A	Cα—Cα	H	1	6.2	6.6
A	Cα—Cα	H	−1	5.5	5.9
A	Cα—Cα	H	0	5.2	5.4
P	O—O	O	1	7.6	8.2
P	O—O	O	−1	4.8	5.0
P	O—O	O	0	3.6	4.0
P	O—O	H	1	4.7	5.1
P	O—O	H	−1	7.7	8.3
P	O—O	H	0	3.6	4.0
P	C—C	O	1	7.7	8.3
P	C—C	O	−1	4.7	4.9
P	C—C	O	0	5.1	5.3
P	C—C	H	1	7.7	8.1
P	C—C	H	−1	5.1	5.3
P	C—C	H	0	5.1	5.3

Type	A1-A2	Ref	N	LB	UB
P	N—N	O	1	7.9	8.3
P	N—N	O	−1	4.7	5.1
P	N—N	O	0	4.9	5.3
P	N—N	H	1	4.7	4.9
P	N—N	H	−1	7.2	7.8
P	N—N	H	0	5.0	5.2
P	Cα—Cα	O	1	8.4	8.8
P	Cα—Cα	O	−1	4.8	5.0
P	Cα—Cα	O	0	6.1	6.3
P	Cα—Cα	H	1	4.8	5.0
P	Cα—Cα	H	−1	7.2	7.8
P	Cα—Cα	H	0	6.1	6.3
H	O—O	O	1	8.3	8.5
H	O—O	O	−1	4.9	5.1
H	O—O	O	0	6.0	6.2
H	O—O	H	1	4.8	5.2
H	O—O	H	−1	8.2	8.6
H	O—O	H	0	6.0	6.2

Type	A1-A2	Ref	N	LB	UB
H	C—C	O	1	8.1	8.3
H	C—C	O	−1	4.8	5.0
H	C—C	O	0	6.0	6.2
H	C—C	H	1	4.8	5.0
H	C—C	H	−1	8.1	8.3
H	C—C	H	0	6.0	6.2
H	N—N	O	1	8.0	8.2
H	N—N	O	−1	4.7	4.9
H	N—N	O	0	6.0	6.2
H	N—N	H	1	4.7	4.9
H	N—N	H	−1	8.0	8.2
H	N—N	H	0	6.0	6.2
H	Cα—Cα	O	1	8.5	8.7
H	Cα—Cα	O	−1	5.0	5.2
H	Cα—Cα	O	0	6.1	6.3
H	Cα—Cα	H	1	5.0	5.2
H	Cα—Cα	H	−1	8.5	8.7
H	Cα—Cα	H	0	6.1	6.3

In all sub-tables, the first column defines secondary structure type: parallel (P) or antiparallel (A), generic strand (U), and helix (H). Measurements of upper and lower bounds of hydrogen bond distances for antiparallel and parallel β-sheets and helices (sub-Table A), adjacent oxygen–oxygen atom distances in strands (sub-Table B), dihedral angles (sub-Table C). Distance restraints for reconstructing helices and β-sheets are presented in sub-Table D. In sub-Table D, second column defines atom pair (atom of residue 1 – atom of residue 2), third column is the hydrogen bond reference atom (oxygen or hydrogen), and fourth column is the neighbor distance of the second residue. If strands a–b and c–d (a, b, c, and d being residue numbers) are antiparallel and have a hydrogen bond between residues b and c, with oxygen atom of b connected to hydrogen atom of c, then, referring to the first row from sub-Table D, we apply distance restraint of [7.4 Å, 8.0 Å] between oxygen of residue b and oxygen of residue (c + 1).

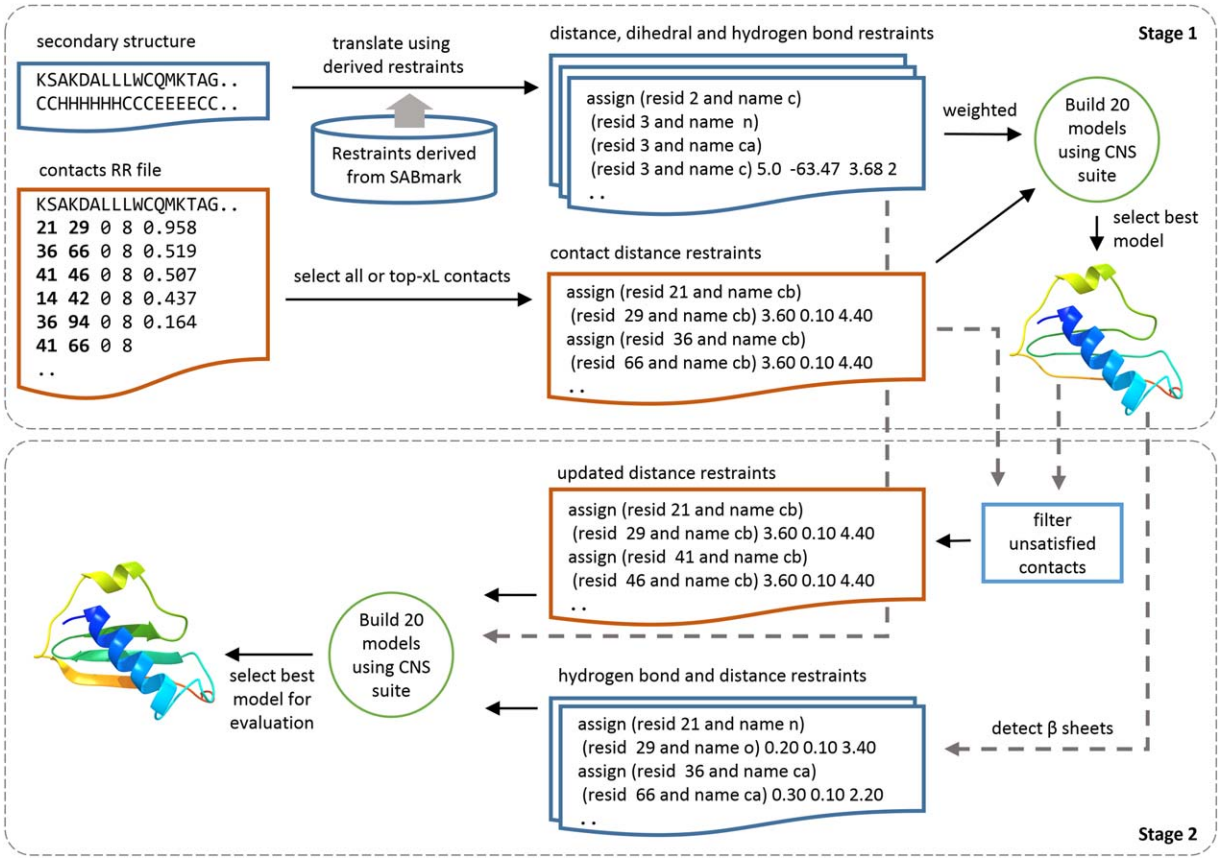

Figure 1

The CONFOLD method for building models with contacts and secondary structures in two stages. When true contacts are the input, all contacts are used to reconstruct models. For predicted contacts, top-*x*L contacts are used, where *x* ranges from 0.4 to 2.2 at a step of 0.2.

following two distances: the average of distance between the backbone nitrogen atom and oxygen atom of the residues that are supposed to be hydrogen bonded, and the average distance between the backbone C—C, Cα—Cα, N—N, and O—O atoms. For example, if residues numbered 15–20 and 30–35 are two strands, their parallel strand distance is the minimum of the average of distance between associated hydrogen bonded atoms 15N and 30O, 15O and 30N, 17N and 32O, 17O and 32N, 19N and 34O, and 19O and 34N, and the average of distance between Cα atoms of residues 15 and 30, 16, and 31, and so on, up to 20 and 35. In case that one of the strands in a pair is longer, we consider all possible ways of trimming the longer strand so that both strands in a pair are of the same length and use the minimum distance of the trimmed pairs as the distance of the two strands.

The rationale for having the two distance measurements between strands of equal size is to accommodate accurate as well as inaccurate contacts. When true (or very accurate) contacts are supplied, the strands are close enough and hydrogen bond associated distance measurement is much smaller and better for strand pairing detection, whereas when predicted contacts are used, the distance measurement based on backbone atoms, although higher, can detect strand pairings more accurately. After all strand pairs are sorted by their distances, we select the closest pair and add it to a list of detected pairs. The next pair in the rank that is not conflicting with hydrogen bonding residues of the previously selected pairs is also added into the list. The process is repeated until all pairs below a distance threshold are considered. Through trial and error, we set this distance threshold as 7 Å.

Customization of distance geometry protocol for contact-based model generation

All the distance, hydrogen bond, and dihedral angle restraints are passed as input to the distance geometry simulated annealing protocol implemented in a revised CNS suite[37,38] version 1.3. The initial suite is designed for experimental data and the parameter files are originally configured to make the van der Waals radii consistent with other NMR refinement programs. We changed the distance geometry simulated annealing protocol,

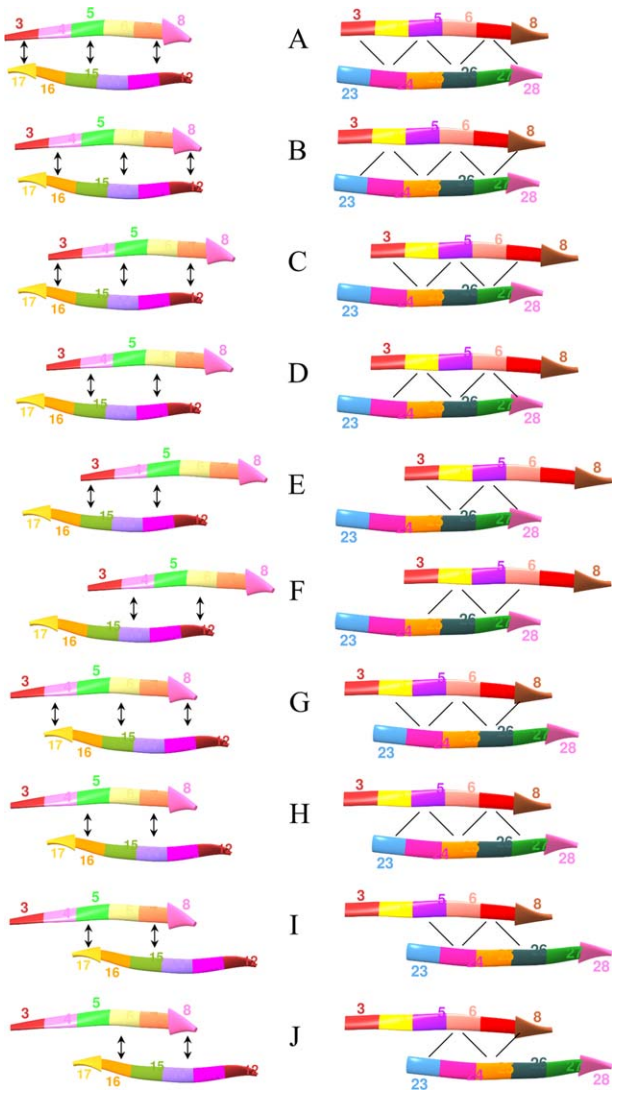

Figure 2

Ten alternate hydrogen-bonding patterns for antiparallel (left) and parallel (right) pairing for a pair of strands, each six residues long. First strand is from residues 3–8, and second strand is from residues 12–17 for antiparallel pairs and 23–28 for parallel pairs. The ideal hydrogen bonding pattern (**A**), alternate hydrogen bonding pattern (**B**), top strand right shifted by one residue (**C**), alternate pattern for C (**D**), top strand right shifted by 2 residues (**E**), alternate pattern for E (**F**), top strand left shifted by 1 residue (**G**), alternate pattern for G (**H**), top strand left shifted by 2 residues (**I**), and alternate pattern for I (**J**). In case of parallel pairing (right), although DSSP uses one more hydrogen bond to consider the strands to be in pair, we take a less strict approach and ignore the hydrogen bonding because we observed that this approach worked better when building models using predicted contacts. Black residue connecting lines show hydrogen bonding and double arrowed lines represent double hydrogen bonding.

"dg_sa.inp" script, by increasing the initial radius parameter "md.cool.init.rad" from 0.8 to 1.0, by increasing the number of minimization steps, and by augmenting the set of atoms used for distance geometry to the atoms we use for restraining, that is, backbone atoms N, Cα, C, O, and Cβ and H. We also updated the code of the subroutines "scalehot" and "scalecoolsetup" so that weighting of restraints could be implemented. A set of 20 three-dimensional models are generated for each execution of the distance geometry simulated annealing protocol.

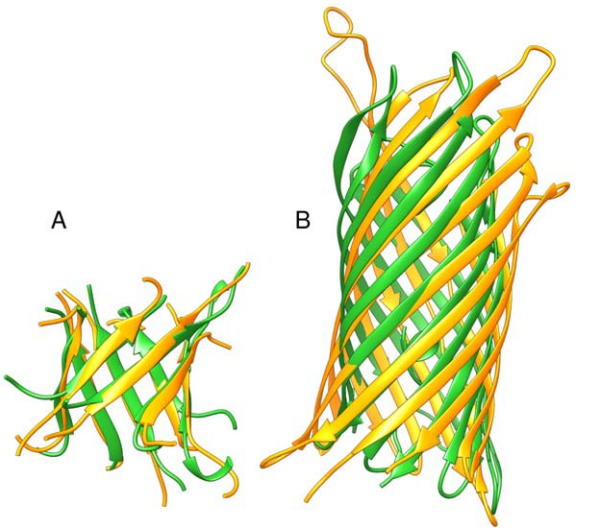

Figure 3

Top models reconstructed for the proteins 2QOM and 1YPI using true secondary structure information along with beta-pairing information but without using any residue contact information. Secondary structure restraints are computed using $\lambda = 0.5$. Superposition of crystal structure (green) and reconstructed top model (orange) of the beta-alpha-beta barrel protein 1YPI (**A**) and antiparallel beta barrel protein 2QOM (**B**).

RESULTS AND DISCUSSION

Optimization of secondary structure restraints

One challenge of contact-based protein structure modeling is to generate realistic secondary structures. We test the effectiveness of our derived secondary structure restraints by building β-sheets and helices for many kinds of proteins (e.g., Fig. 3). Furthermore, we build helix and β-sheet models (not complete fold) for 24 proteins in Tc category of the 11th Critical Assessment of Techniques for Protein Structure Prediction (CASP 11) using predicted helices, strands, and β-sheet topologies predicted by BETApro.[44] The top models successfully recover 33 out of 42 strand residues and 77 out of 79 for helix residues on average. The primary reason for a lower reconstruction rate of β-sheets than helices is the presence of proline in strands. Since proline acts as hydrogen-bond acceptor only and does not follow along with the typical Ramachandran plot, when it appears in strands, the hydrogen-bonding pattern is broken.[45]

We also investigate how the scaling factor (λ) controlling upper bound and lower bound of all secondary structure restraints (hydrogen bond, distance, and

Table II
Choice of λ, Controlling the Upper and Lower Bounds, Affecting the Reconstruction Quality of Secondary Structures for 15 Proteins in EVFOLD Dataset Reconstructed Using Top-L/2 Contacts Predicted by EVFOLD

	% of residues reconstructed	
λ	Strand	Helix
0.3	31	100
0.4	28	100
0.5	43	100
0.6	30	100
0.7	34	100
0.8	38	97
0.9	37	97
1.0	29	96
1.1	26	95
1.2	27	96

Percentages of helix and β-sheet residues reconstructed are listed against various values of λ.

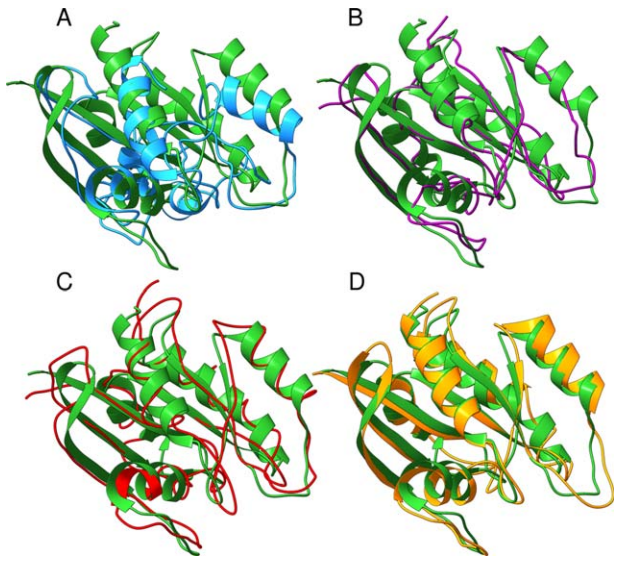

Figure 4
Best models reconstructed for the protein 5p21 using Modeler (**A**), reconstruct (**B**), customized CNS DGSA protocol (**C**), and CONFOLD (**D**). All models are superimposed with native structure (green). The TM-scores of Models A, B, C, and D are 0.53, 0.86, 0.88, and 0.94, respectively. Model D reconstructed by CONFOLD has higher TM-score and also much better secondary structure quality than the other models.

dihedral angle) affects the quality of reconstructed secondary structures. When true contacts are used for reconstruction, we find that the choice of λ does not heavily affect the quality of secondary structures; however, using restraints derived with the default value of λ, 1.0, can generate models of slightly higher quality. To determine the value of λ for generating restraints for predicted contacts, we test the values of λ ranging from 0.3 to 1.2 at step of 0.1. Using 15 proteins in the EVFOLD data set, we select top-L/2 predicted contacts, detect strand pairings from Stage 1 models, and build Stage 2 models, and record the number of helix residues and β-sheet residues realized in the final models. Table II illustrates the reconstruction quality affected by the choice of λ. Although, helix residues are reconstructed with almost all values of λ, β-sheet residues are reconstructed best with $\lambda = 0.5$. Moreover, in addition to the restraints derived from the SABmark database, we test the secondary structure restraints derived from other different sets of protein structures.[43,46] The secondary structures generated in these experiments are very similar, suggesting the restraints calculated from these datasets are equally effective and represent secondary structure patterns well.

Reconstruction of tertiary structural models using true contacts

We use CONFOLD to reconstruct the tertiary structures of all 15 proteins in the EVFOLD dataset and compare the results with those from Reconstruct[25] and Modeller.[35] From native tertiary structures of these proteins, we compute three-class secondary structure information using DSSP[47] and true Cβ—Cβ contacts at 8 Å threshold with sequence separation threshold of 6 residues. We experiment CONFOLD with contact restraints and secondary structure restraints (denoted as CONFOLD), CONFOLD without secondary structure restraints (denoted as CNS DGSA), Reconstruct with only contact restraints since it does not consider secondary structures, and Modeller with both contact restraints and secondary structure restraints. We generate 20 models using each method for each protein. The detailed results [e.g., TM-score[48] and Root Mean Square Deviation (RMSD) calculations] for all these proteins are reported in Table III. The average TM-score[48] of the best models constructed by CONFOLD with secondary structure restraints, CONFOLD without secondary structure restraints, Reconstruct and Modeler are 0.84, 0.77, 0.75, and 0.58, respectively. The accuracy of CONFOLD with secondary structure restraints is much higher than that of Modeler with the same input. All the methods perform better on single-domain proteins than on multi-domain proteins (e.g., 2O72 and 1G2E). Figure 4 shows the models reconstructed by these methods for the protein 5P21. For this protein of 166 residues, CONFOLD reconstructs a highly accurate model with a TM-score of 0.932 with 39 out of 44 β-sheet resides reconstructed. In contrast, the models reconstructed by CNS DGSA and Reconstruct have good global topology but poor secondary structures, whereas the model reconstructed by Modeler has poor global topology but better secondary structures.

Comparing the best models built using only contact restraints and those using both contact restraints and secondary structure restraints in Table III, we find that adding secondary structure restraints improves the

Table III

Comparison of Accuracy and Secondary Structure Quality of the Best of 20 Models Reconstructed for 15 Proteins in EVFOLD Benchmark Set Reconstructed Using CONFOLD With Secondary Structure Restraints, Our Customized CNS DGSA Protocol, Reconstruct, and Modeler

	Native				CNS DGSA				CONFOLD				Modeler				Reconstruct			
PDB code	N_c	L	H	E	TM-score	RMSD	H	E	TM-score	RMSD	H	E	TM-score	RMSD	H	E	TM-score	RMSD	H	E
5p21A	372	166	57	44	0.88	2.0	4	0	0.94	1.4	56	39	0.56	8.3	42	4	0.86	2.3	0	0
2o72A[a]	552	213	0	90	0.49	7.9	0	0	0.49	8.2	0	52	0.64	5.1	0	8	0.56	10.2	0	0
1oddA	160	100	31	23	0.83	1.8	0	6	0.90	1.4	29	20	0.64	4.5	29	0	0.79	2.1	0	6
5ptiA	113	58	8	14	0.71	2.1	0	0	0.82	1.5	7	14	0.53	6.2	6	0	0.70	2.1	0	0
1hzxA	606	340	201	12	0.63	6.7	4	0	0.93	2.3	200	8	0.55	8.8	181	0	-	-	-	-
1rqmA	209	105	34	25	0.84	1.8	4	4	0.91	1.2	33	19	0.46	7.8	26	0	0.80	2.3	0	0
2it6A	288	132	23	35	0.82	2.2	4	0	0.88	1.9	23	25	0.45	7.8	20	0	0.81	2.5	0	0
1wvnA	119	74	29	20	0.65	2.9	0	0	0.86	1.5	29	20	0.63	3.3	21	0	0.59	3.6	0	0
1f21A	344	152	54	44	0.82	2.7	0	4	0.90	1.8	52	36	0.48	9.8	47	0	0.79	2.9	0	0
2hdaA	133	59	0	25	0.81	1.5	0	0	0.78	1.6	0	21	0.67	3.5	0	0	0.74	1.8	0	8
1g2eA[a]	353	167	39	56	0.41	9.3	0	0	0.48	12.6	35	40	0.33	13.9	31	0	0.54	5.8	0	6
1r9hA	287	118	11	44	0.85	1.9	0	13	0.87	1.9	11	38	0.65	4.4	5	4	0.81	2.3	0	8
1e6kA	241	130	54	23	0.83	2.5	0	15	0.93	1.4	55	19	0.48	6.2	47	0	0.79	2.7	4	6
3tgiE	650	223	17	72	0.93	1.6	0	20	0.94	1.5	16	53	0.55	7.8	15	0	0.92	1.7	0	17
1bkrA	158	108	58	0	0.78	2.5	0	0	0.91	1.3	55	0	0.48	7.1	51	0	0.76	3.3	0	0
Avg	306	143	41	35	0.75	3.3	1	4	0.84	2.8	40	27	0.54	7.0	35	1	0.75	3.3	0	4

The column N_c refers to the number of contacts in the native structure, and the columns H and E are the number of helix and β-sheet residues computed using DSSP. Reconstruction results for the long protein 1 hzx using Reconstruct is not presented because Tinker failed to run because of memory requirement issues.

[a]Multidomain proteins.

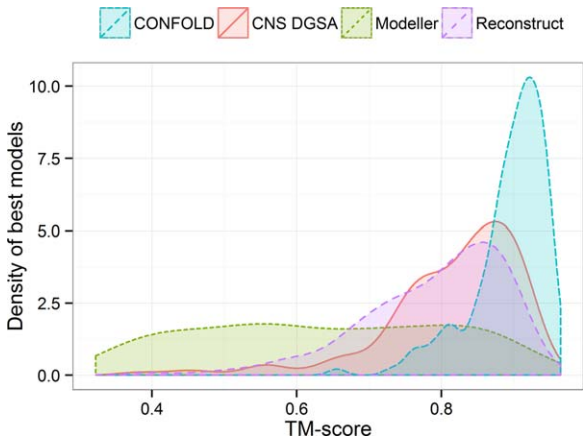

Figure 5

Distribution of TM-scores of the best models reconstructed by the four methods for 150 FRAGFOLD proteins.

quality of global topology of the models by increasing average TM-score from 0.75 to 0.84 as well as the quality of secondary structures in the models by recovering much more secondary structure residues. However, even though secondary restraints can help recover most helix residues, they can only help recover about 75% of β-sheet residues. The β-sheet detection technique seems to improve beta-sheet reconstruction; however, it does not remarkably improve the global quality of models when true contacts are used. For the 15 proteins, the models in Stage 2 have almost twice as many beta-sheet residues as in those in Stage 1, but they have almost the same TM-scores.

Furthermore, we compared CONFOLD, our customized CNS DGSA protocol, Reconstruct and Modeler on 150 proteins in FRAGFOLD benchmark set using the same protocol. Figure 5 shows the distribution of TM-Scores[48] of the models reconstructed by these methods. The average TM-score of the models are 0.89, 0.81, 0.79, and 0.63 for CONFOLD with secondary structures, customized CNS DGSA protocol, Reconstruct, and Modeler, respectively. The results show that when secondary structure restraints are considered, CONFOLD can reconstruct models from true contacts with substantially better quality than Modeler, and when only contact restraints are used for reconstruction our customized CNS DGSA protocol can reconstruct better than Reconstruct. Our customized CNS DGSA protocol performs better than Reconstruct, which uses Tinker for modeling, in 131 out of 150 cases, and the average improvement in TM-score on all 150 proteins is 3%. This may suggest our customized CNS DGSA protocol works better than the one implemented by Tinker in Reconstruct. Detailed comparison for all the 150 proteins is presented in Table I of Supporting Information.

Tertiary structure prediction using predicted contacts

Using predicted contacts and secondary structures available for 15 proteins in the EVFOLD benchmark set, we built 400 models for each protein using CONFOLD, and evaluated them against the same number of available EVFOLD models. The average TM-score of the best model predicted by CONFOLD is 0.59, 5.5% higher than the best models predicted by EVFOLD. CONFOLD produced models with higher TM-score for 12 out of 15 proteins. The average improvement in RMSD is 0.63 Å. Moreover, the best models reconstructed by CONFOLD have better secondary structure quality with 35 helix residues and 10 strand residues per model on average. Table IV presents the comparison of model accuracy and

Table IV

Comparison of Accuracy and Secondary Structure Quality of Best Models Built by CONFOLD and EVFOLD

	Native			EVFOLD				CONFOLD			
UNIPROT-NAME	L	H	E	TM-score	RMSD	H	E	TM-score	RMSD	H	E
YES_HUMAN	48	0	14	0.47	3.50	0	4	0.41	4.41	0	8
CHEY_ECOLI	114	47	20	0.69	3.28	46	10	0.77	2.50	53	10
SPTB2_HUMAN	106	58	0	0.51	6.65	21	0	0.67	3.39	52	0
OMPR_ECOLI	77	31	6	0.48	7.70	12	0	0.53	5.20	32	6
OPSD_BOVIN	248	165	8	0.56	8.05	116	0	0.59	7.07	176	0
O45418_CAEEL	95	11	31	0.53	5.77	0	0	0.56	4.76	0	12
RNH_ECOLI	140	53	44	0.57	6.99	23	4	0.63	6.03	54	0
PCBP1_HUMAN	63	25	17	0.40	6.33	13	0	0.46	4.73	19	4
ELAV4_HUMAN	71	20	23	0.60	3.21	18	0	0.62	3.10	20	0
THIO_ALIAC	103	30	25	0.59	3.99	25	8	0.64	3.70	31	16
CADH1_HUMAN	100	0	42	0.58	4.18	0	18	0.61	4.20	0	23
BPT1_BOVIN	53	8	14	0.56	2.95	5	0	0.50	3.27	5	8
RASH_HUMAN	161	57	40	0.76	3.15	49	10	0.78	3.01	61	27
A8MVQ9_HUMAN	107	23	24	0.53	5.57	18	0	0.56	6.17	22	4
TRY2_RAT	216	7	72	0.61	6.73	4	8	0.57	6.97	7	31
Average	**113**	**36**	**25**	**0.56**	**5.20**	**23**	**4**	**0.59**	**4.57**	**35**	**10**

Columns H and E are number of helix and β-sheet residues assigned by DSSP. RMSD values are in Å.

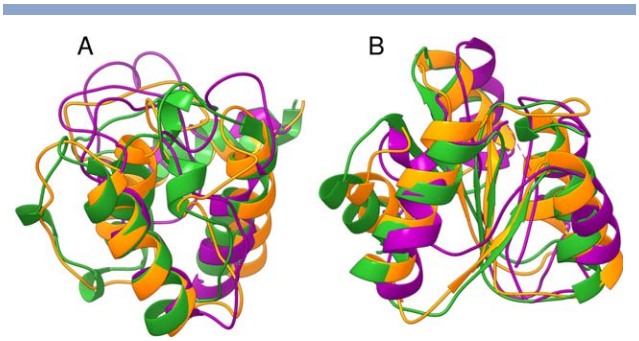

Figure 6

Best predicted models for the proteins RNH_ECOLI (**A**) and SPTB2_HUMAN (**B**) using EVFOLD (purple) and CONFOLD (orange) superimposed with native structures (green). The TM-scores of these models are reported in Table IV. CONFOLD models have higher TM-score and better secondary structure quality than EVAFOLD.

secondary structure quality for all 15 proteins. As an example, Figure 6 visualizes the best models reconstructed for proteins RNH_ECOLI and SPTB2_HUMAN.

In addition to comparing of best models, we also compare the quality of all models for all proteins (400 models for each of the 15 proteins) by EVFOLD with the models built by CONFOLD. The distribution of CONFOLD and EVFOLD models in Figure 7 shows that CONFOLD models are better in general. On average, the TM-score of all CONFOLD models is 0.42, 20% higher than EVFOLD model pool.

Besides comparing CONFOLD's final models with those of EVFOLD for the 15 proteins, we also compare the models in first and second stages of CONFOLD itself. Comparison of the best models in Stages 1 and 2 suggests a significant improvement in the accuracy and secondary structure quality of models from Stages 1 to 2. To analyze the improvement due to β-sheet detection and contact filtering in Stage 2, in Table V, we compare the best models in first stage, second stage with β-sheet detection only, and second stage with contact filtering only, and second stage with contact filtering and β-sheet detection (that is, CONFOLD). For 13 out of 15 proteins, the models in the second stage of CONFOLD have better accuracy than those in the first stage. For 12 proteins, models built by filtering contacts alone have better accuracy than the models of the first stage. For 8 proteins models built using β-sheet detection alone have better accuracy that the models of the first stage. On average, a 0.9 Å RMSD improvement is observed in CONFOLD second stage, and the number of strands in the second stage is more than three times that in the first stage on average. The main contributor of the higher accuracy of models in the second stage is contact filtering, with improvement of 0.5 Å RMSD on average. Figure 7 also shows that the second stage of CONFOLD improves the quality of reconstruction over its first stage and also over EVFOLD.

In addition to the EVFOLD data set, we test CONFOLD with predicted contacts on 150 proteins in FRAGFOLD benchmark dataset. Since predicted secondary structures are not available for these proteins, we predict secondary structure using PSIPRED, and then built models using CONFOLD. The best models predicted by FRAGFOLD have TM-score of 0.54,[39] and those by CONFOLD have TM-score of 0.55, on average. However, the comparison here should be only considered a qualitative understanding of the performance of CONFOLD because the models of the two methods were not generated in the exactly same conditions. The caveats are that: (a) FRAGFOLD's best models are best of 5 whereas CONFOLD's best models are best of 400 models, (b) FRAGFOLD used fragment information and CONFOLD did not, and (c) the secondary structures used by CONFOLD may not be same as the one used by FRAGFOLD. Besides comparing the quality of CONFOLD and

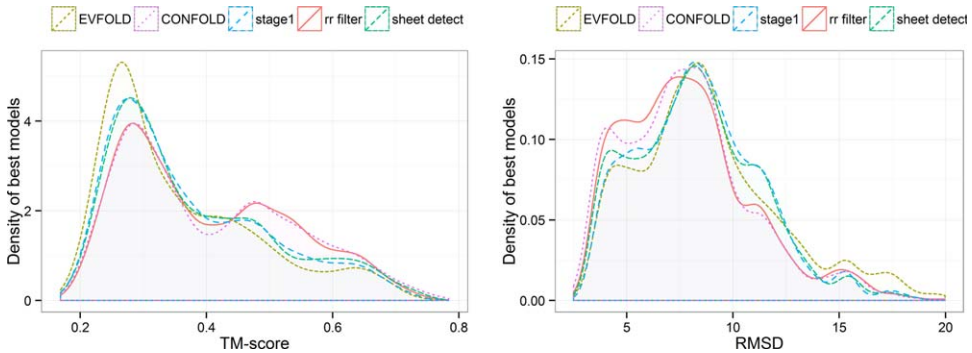

Figure 7

Distribution of model quality of the EVFOLD models and the models built by CONFOLD. Distribution of models built in first stage of CONFOLD (Stage 1), second stage with contact filtering only (rr filter), and second stage with β-sheet detection only (sheet detect) are also presented. Each curve represents the distribution of 400 times 15 models. Since some models in the EVFOLD model pool have RMSD 20 Å, all models with RMSD greater than 20 Å from all four model pools were filtered out.

Table V
Best Models Built in First Stage of CONFOLD, Second Stage of CONFOLD with Only β-Sheet Detection, the Second Stage of CONFOLD with Only Contact Filtering, and the Full Stage 2 of CONFOLD

	Stage1			Sheet detect			Contact filter			Stage 2		
UNIPROT-NAME	TM-score	H	E	TM-score	H	E	TM-score	H	E	TM-score	H	E
YES_HUMAN	0.42	0	9	0.44	0	6	0.45	0	0	0.41	0	8
CHEY_ECOLI	0.69	52	0	0.70	49	4	0.72	50	0	0.77	53	10
SPTB2_HUMAN	0.57	40	0	0.57	40	0	0.67	52	0	0.67	52	0
OMPR_ECOLI	0.52	31	0	0.53	31	0	0.49	36	0	0.53	32	6
OPSD_BOVIN	0.56	159	0	0.56	159	0	0.59	176	0	0.59	176	0
O45418_CAEEL	0.53	4	0	0.54	0	12	0.54	4	0	0.56	0	12
RNH_ECOLI	0.57	48	0	0.56	48	4	0.63	54	0	0.63	54	0
PCBP1_HUMAN	0.41	15	0	0.41	18	0	0.43	19	0	0.46	19	4
ELAV4_HUMAN	0.58	21	8	0.62	20	7	0.62	20	0	0.62	20	0
THIO_ALIAC	0.63	40	0	0.69	32	15	0.61	41	0	0.64	31	16
CADH1_HUMAN	0.52	0	6	0.51	0	12	0.56	0	18	0.61	0	23
BPT1_BOVIN	0.55	7	0	0.53	7	18	0.55	7	4	0.50	5	8
RASH_HUMAN	0.75	62	16	0.76	64	21	0.77	63	14	0.78	61	27
A8MVQ9_HUMAN	0.50	21	0	0.44	20	4	0.57	21	0	0.56	22	4
TRY2_RAT	0.56	6	4	0.57	6	25	0.57	7	8	0.57	7	31
Average	**0.56**	**34**	**3**	**0.56**	**33**	**9**	**0.58**	**37**	**3**	**0.59**	**35**	**10**

Columns H and E are the number of helix and β-sheet residues computed by DSSP.

FRAGFOLD models, we compare how well contacts are used to guide the model building process. For the 150 proteins, we calculated the Pearson's correlation between the precision of top-L/2 predicted contacts and the TM-scores of the best models for both FRAGFOLD and CONFOLD in order to find, which method is more contact driven. The correlation values for FRAGFOLD models and CONFOLD models are 0.53 and 0.70, respectively. This suggests that contacts played a more important role in the modeling process of CONFOLD than in FRAGFOLD. The detailed prediction results on FRAGFOLD dataset are presented in Table II in Supporting Information.

Comparing the models predicted for proteins in FRAGFOLD dataset in the two stages of CONFOLD, for 123 out of 150 proteins, we find the best models in the second stage of CONFOLD. The average TM-score of the best models in the second stage is 0.55, 6.1% higher than the best models in first stage. The change of TM-score of best models from the first stage to the second stage is in the range [−0.036, 0.1148]. The average number of beta sheet residues in a protein increases from 2 in Stage 1 to 9 in Stage 2. Furthermore, the average TM-score of all models for all proteins in Stage 2 is 0.38, 11% higher than that of Stage 1 models. The distribution of TM-score of the best models and all models in Stages 1 and 2 are shown in Figure 8.

In the second stage, CONFOLD tries to filter out noisy contacts through structure modeling in order to improve the quality of models. To check if CONFOLD's improvement in the second stage is biased toward high-accuracy contacts, we calculated the Pearson correlation between predicted confidence scores of top-L/2 original contacts and the TM-scores of the best models in Stages 1 and 2.

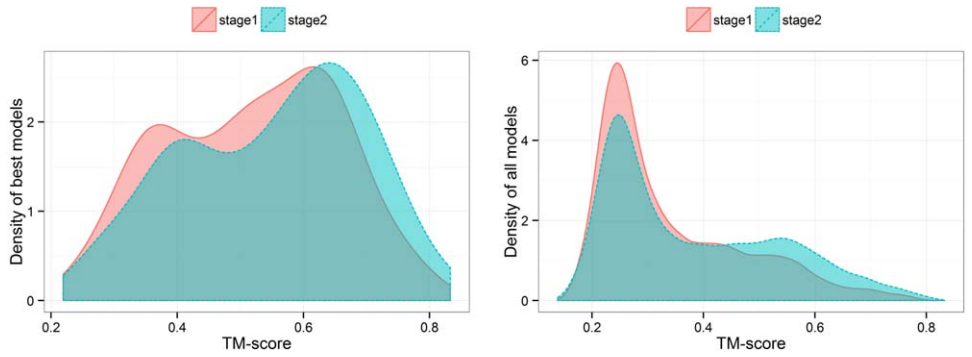

Figure 8
Improvement in the accuracy of best models (left) and all 400 models (right) in the second stage of CONFOLD over the first stage for 150 proteins in FRAGFOLD dataset.

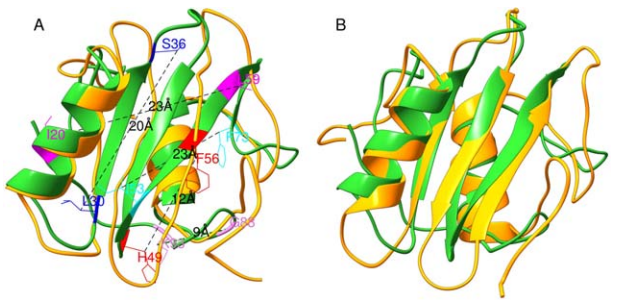

Figure 9

Contact filtering from Stages 1 to 2 for the protein 1NRV. (**A**) Superimposition of the best model in stage 1 reconstructed with top-0.6 L contacts by CONFOLD (orange) with the native structure (green). The model has TM-score of 0.50. Among the top-0.6 L (60) contacts, 5 out of 8 erroneous contacts that were removed in Stage 2 are visualized in the native structure along with the distance between their Cβ-Cβ atoms. The filtered, predicted contacts (20–59, 53–73, 30–36, 49–56, and 88–93) have Cβ-Cβ distances of 23, 23, 20, 12, and 9 Å, respectively, in the native structure. Each pair of residues predicted to be in contact is denoted by the same color. (**B**) Superimposition of the best model in Stage 2 reconstructed with reduced/filtered top-0.6 L contacts by CONFOLD (orange) with the native structure (green). TM-score of the model is 0.61.

The lower correlation score (0.2) suggests that CONFOLD improves the quality of the models even when the precision of contacts is not high. Interestingly, our experiment shows that in stage 2 CONFOLD mainly gets rid of the most inaccurate/noisy contacts. Figure 9 illustrates the models for protein 1NRV ($L = 100$) reconstructed with top-0.6 L contacts in Stages 1 and 2. Sixty contacts were used to construct the model in Stage 1, and 8 of them were removed in Stage 2. Five out of 8 removed contacts are separated by large distances in the native structure of this protein, which certainly would hinder the reconstruction process if they were kept. For this protein the best model in Stage 2 has TM-score of 0.61, 22% higher than the best model in Stage 1.

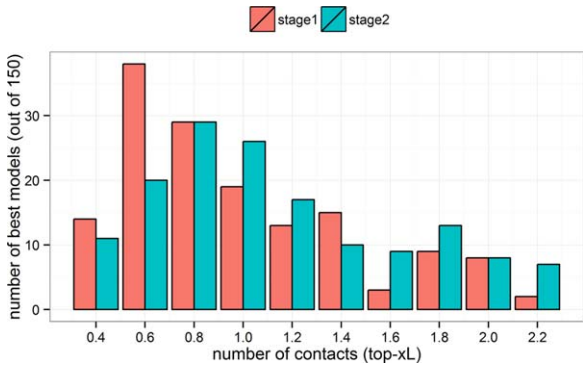

Figure 10

Number of best models and the number of contacts used to build the best models for 150 proteins in FRAGFOLD dataset.

Analysis of number of predicted contacts needed to obtain best fold

Although 99.9% of the proteins in PDB have less than 3 L contacts, much fewer true contacts are sufficient to fold the proteins accurately.[24,25] However, how many predicted contacts are needed to best fold proteins is still an open question. Using 150 proteins in FRAGFOLD dataset, we find that 60% of the best models are reconstructed with top 0.6 L, 0.8 L, 1.0 L, or 1.2 L contacts in both stages of CONFOLD (Fig. 10). The distribution shows that different proteins need different numbers of contacts to be folded well. Therefore, instead of fixing the number of contacts, predicting a range for the number of contacts will be useful for contact-based model reconstruction.

CONFOLD for *ab initio* protein structure prediction

Success of a complete *ab initio* protein structure prediction method based on predicted contacts and secondary structures primarily depends on (a) the precision of predicted contacts and the accuracy of predicted secondary structures, (b) selection of appropriate number of contacts, (c) how well noisy contacts are filtered, (d) reconstruction capability of the method, that is, how well models can be constructed using the predicted information, and (e) effectiveness of the model selection technique. Most contact prediction methods do not use any known homologs protein structure template and predict contacts purely based on sequences, and hence may be plugged into such a contact-based *ab initio* structure prediction method. For the 15 proteins in EVFOLD data set used in our experiments, the authors of the data set predicted secondary structures and contacts using sequence information only without using any known structural template or fragment information in order to fairly discuss their *ab initio* contact prediction approach. Therefore, the tertiary structure models reconstructed by CONFOLD for the proteins in EVFOLD data set are *ab initio* models. And the accuracy of the *ab initio* models is relatively high because the accuracy of contact predictions for most proteins in the data set is high due to the availability of a large number of homologs protein sequences. In real world, however, sequence-based contact prediction methods may make poor predictions for sequences that do not have sufficient number of sequences in the multiple sequence alignment, which may lead to less accurate tertiary structural models reconstructed from contacts. The minimum number of contacts needed for best reconstruction of a protein, although generally being around top-0.5 L to top-L predicted contacts, depends on the structure and should not be fixed for all proteins. Once number of contacts or a range for number of contacts is decided, a modeling approach like CONFOLD can make best use of contacts to build three-

dimensional models without using any template or fragment information, and therefore is a pure *ab initio* approach. Finally, for model selection, although we do not present any results in this work, Pcons[49] is suggested as one of the best clustering-based methods[36] to identify top-ranked models generated using a modeling approach like CONFOLD. Residue–residue contact predictions can also be combined with these model-ranking methods to select quality protein models.

CONCLUSION

We developed and evaluated a method that improved the reconstruction of protein structures from residue–residue contacts and secondary structures. Our method deterministically controls *ab initio* protein-folding process with restraints generated from a new, comprehensive set of parameters and rules for contacts and secondary structures. Our method optimizes protein structural models through a unique two-stage process and thus the models generated have high quality secondary structures. Our experiment demonstrates that the two-stage process filters noisy predicted contacts, enhances the quality of secondary structures, and improves the overall accuracy of models. Our work also shows that weighting contact restraints and secondary structure restraints appropriately is important for contact-guided structure modeling. Moreover, our analysis suggests that different proteins may need a different number of contacts in terms of sequence length to be folded well from residue–residue contacts.

REFERENCES

1. Monastyrskyy B, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue–residue contact predictions in casp9. Proteins: Struct Funct Bioinformatics 2011;79:119–125.
2. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue–residue contact prediction in casp10. Proteins: Struct Funct Bioinformatics 2014;82:138–153.
3. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 2007; 8:113.
4. Eickholt J, Cheng J. Predicting protein residue–residue contacts using deep networks and boosting. Bioinformatics 2012;28:3066–3072.
5. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
6. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28: 184–190.
7. Tegge AN, Wang Z, Eickholt J, Cheng J. NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Res 2009;37:W515–W518.
8. Wu S, Szilagyi A, Zhang Y. Improving protein structure prediction using multiple sequence-based contact predictions. Structure 2011; 19:1182–1191.
9. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PloS One 2011;6:e28766.
10. Taylor TJ, Bai H, Tai CH, Lee B. Assessment of casp10 contact-assisted predictions. Proteins: Struct Funct Bioinformatics 2014;82: 84–97.
11. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. Bioinformatics 2013;29:i266–i273.
12. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30:3128–3130.
13. Kaján L, Hopf TA, Marks DS, Rost B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 2014;15:85.
14. 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 2015;31:999–1006.
15. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 2014;3:e02030.
16. Skwark MJ, Raimondi D, Michel M, Elofsson A. Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput Biol 2014;10:e1003889.
17. Zhang H, Zhang T, Chen K, Kedarisetti KD, Mizianty MJ, Bao Q, Stach W, Kurgan L. Critical assessment of high-throughput standalone methods for secondary structure prediction. Brief Bioinformatics 2011;12:672–688.
18. Chen K, Kurgan L. Computational prediction of secondary and supersecondary structures. In Kister AE, editor. Protein supersecondary structures. New York: Springer; 2013. pp 63–86.
19. Pirovano W, Heringa J. Protein secondary structure prediction. In Carugo O, Eisenhaber F editors. Data mining techniques for the life sciences. New York: Springer; 2010. pp 327–348.
20. Cole C, Barber JD, Barton GJ. The jpred 3 secondary structure prediction server. Nucl Acids Res 2008;36:W197–W201.
21. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucl Acids Res 2005;33:W72–W76.
22. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chem 2012;33:259–267.
23. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
24. Sathyapriya R, Duarte JM, Stehr H, Filippis I, Lappe M. Defining an essence of structure determining residue contacts in proteins. PLoS Comput Biol 2009;5:e1000584.
25. Duarte JM, Sathyapriya R, Stehr H, Filippis I, Lappe M. Optimal contact definition for reconstruction of contact maps. BMC Bioinformatics 2010;11:283.
26. Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R. FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. Bioinformatics 2008;24:1313–1315.
27. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. Fold Des 1997;2:295–306.
28. Bohr J, Bohr H, Brunak S, Cotterill RM, Fredholm H, Lautrup B, Petersen SB. Protein structures from distance inequalities. J Mol Biol 1993;231:861–869.
29. Moré JJ, Wu Z. Distance geometry optimization for protein structures. J Global Optim 1999;15:219–234.
30. Di Lena P, Vassura M, Margara L, Fariselli P, Casadio R. On the reconstruction of three-dimensional protein structures from contact maps. Algorithms 2009;2:76–92.
31. Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R. Reconstruction of 3D structures from protein contact maps. IEEE/ACM Trans Comput Biol Bioinformatics (TCBB) 2008;5:357–367.

32. Ponder J, Richards F. TINKER molecular modeling package. J Comput Chem 1987;8:1016–1024.
33. Konopka BM, Ciombor M, Kurczynska M, Kotulska M. Automated procedure for contact-map-based protein structure reconstruction. J Membr Biol 2014;247:409–420.
34. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol 2012;10: e1001244.
35. Eswar N, Webb B, Marti-Renom MA, Madhusudhan M, Eramian D, Shen My, Pieper U, Sali A. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics 2007;Chapter 2:Unit 2.9..
36. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. PconsFold: improved contact predictions improve protein models. Bioinformatics 2014;30:i482–i488.
37. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang J-S, Kuszewski J, Nilges M, Pannu NS. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr Sect D: Biol Crystallogr 1998;54:905–921.
38. Brunger AT. Version 1.2 of the crystallography and NMR system. Nat Protoc 2007;2:2728–2733.
39. Kosciolek T, Jones DT. De novo structure prediction of globular proteins aided by sequence Variation-derived contacts. PloS One 2014;9:e92197.
40. Van Walle I, Lasters I, Wyns L. SABmark—a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 2005;21:1267–1268.
41. Salemme F. Structural properties of protein β-sheets. Prog Biophys Mol Biol 1983;42:95–133.
42. Salemme F, Weatherford D. Conformational geometrical properties of β-sheets in proteins: II. Antiparallel and mixed β-sheets. J Mol Biol 1981;146:119–141.
43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucl Acids Res 2000;28:235–242.
44. Cheng J, Baldi P. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Bioinformatics 2005;21:i75–i84.
45. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. J Mol Biol 1991;218:397–412.
46. Taylor TJ, Tai CH, Huang YJ, Block J, Bai H, Kryshtafovych A, Montelione GT, Lee B. Definition and classification of evaluation units for casp10. Proteins: Struct Funct Bioinformatics 2014;82:14–25.
47. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
48. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
49. Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network–based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.