RESDIUE-RESIDUE CONTACT DRIVEN PROTEIN STRUCTURE PREDICTION USING OPTIMIZATION AND MACHINE LEARNING

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

BADRI ADHIKARI

Professor Jianlin Cheng, Dissertation Supervisor

JULY 2017

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

RESDIUE-RESIDUE CONTACT DRIVEN PROTEIN STRUCTURE PREDICTION USING OPTIMIZATION AND MACHINE LEARNING

presented by Badri Adhikari, a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Jianlin Cheng

Professor Dong Xu

Professor Jeffrey Uhlmann

Professor Liscum Mannie

DEDICATION

I dedicate this dissertation to my younger brother Kedar Adhikari.

"Only a brother can love like a father, care like a mother, and support like a friend."

- Unknown

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisor and committee chair Professor Jianlin Cheng. Without his guidance and persistent help this dissertation would not have been possible. I also like to thank Professor Dong Xu and Professor Jeffrey Uhlmann in the EECS department, Professor Liscum Mannie in the Division of Biological Sciences, and Dr. Robin Walker in the Office of Research and Graduate School, for guidance, encouragement and advice throughout my time as a student.

I must express my gratitude to Nilima, my wife, who experienced many of the ups and downs of my research. My special thanks go my father-in-law, Laxman Kafle, for his support during his stay in the United States.

Finally, I would like to thank my friends Dr. Renzhi Cao, Dr. Debswapna Bhattacharya, Dr. Jilong Li, Dr. Jesse Eickholt, Dr. Zheng Wang, Tuan A. Trieu and Jie Hou in our research group for listening, offering me advice, and supporting me through this entire process.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

Significant improvements in the prediction of protein residue-residue contacts are observed in the recent years. These contacts, predicted using a variety of coevolution-based and machine learning methods, are the key contributors to the recent progress in ab initio protein structure prediction, as demonstrated in the recent CASP experiments. Continuing the development of new methods to reliably predict contact maps, tools to assess the utility of predicted contacts, and methods to construct protein tertiary structures from predicted contacts, are essential to further improve ab initio structure prediction. In this dissertation, three contributions are described – (a) DNCON2, a two-level convolutional neural network-based method for protein contact prediction, (b) ConEVA, a toolkit for contact assessment and evaluation, and (c) CONFOLD, a method of building protein 3D structures from predicted contacts and secondary structures. Additional related contributions on protein contact prediction and structure reconstruction are also described. DNCON2 and CONFOLD demonstrate state-of-the-art performance on contact prediction and structure reconstruction from scratch. All three protein structure methods are available as software or web server which are freely available to the scientific community.

Chapter 1

Introduction

Given around a hundred thousand protein amino acid sequences and their correct three-dimensional structures, can we predict the structures for other protein sequences that do not have solved structures? This protein structure prediction problem, although appears tempting to solve, has been vexing bioinformaticians since half a century [1]. Solving this problem can save millions of dollars of wet-lab experimental research, can lead to the cure of thousands of diseases through drug design and will push humanity closer to understand more about life processes. As many research groups have demonstrated, that structures can be predicted with high accuracy when there are structural templates available. Because templates cannot always be found, it is important to study and develop protein models using ab initio or de novo techniques. Residue-residue contacts, in the recent years, have been found extremely useful for the same. The application of contacts, however, extends much beyond structural bioinformatics can be valuable to researchers working in X-ray crystallography, cryo-EM or NMR [2].

A major motivation for protein contact prediction and contact-guided protein structure prediction comes from the general finding that accurate contacts lead to accurate tertiary structural models. Studies like FT-COMAR [3] and Reconstruct [4] on protein structure reconstruction using true contacts have shown that in general three-dimensional protein structures can be recovered using two-dimensional contact maps. For instance, using true Cα contact maps derived with a distance threshold of 9Å, a study reconstructed 19 proteins with accuracy of 1Å RMSD [5]. Similarly, deriving true contacts at distance cut-offs higher than 9Å, Vassura et al. reconstructed Cα models for 1,760 proteins of different fold classes with RMSD of around 2Å using the FT-COMAR method [3, 6]. In another study, authors have shown that the quality of 3D reconstruction is unaffected by deleting

up to an average 75% of the real contacts [7]. Likewise, in a different study, it is demonstrated that the number of contacts needed for reconstruction can be decreased using a cone-peeling method and a reconstruction accuracy of ≤ 4Å can be achieved with just around 20 to 30% of true contacts on a data set of 12 proteins [8]. Most recently, it is also shown that a distance cut-off of 9Å to 11Å delivers accurate reconstructions using Cβ atoms for defining contacts on a data set of 60 proteins [4].

Realizing that the contacting residues which are far apart in the protein sequence but close together in the three-dimensional space are important for protein folding [9], contacts are widely categorized as short-range, medium-range and long-range. Short-range contacts are those separated by 6 to 11 residues in the sequence; medium-range contacts are those separated by 12 to 23 residues, and long-range contacts are those separated by at least 24 residues. Most contact prediction assessment methods evaluate long range contacts separately as they are the most important of the three and also the hardest to predict [10, 11, 12]. Depending upon the three-dimensional shape (fold), some proteins have a lot of short range contacts while others have more long range contacts, as shown in **Figure 1.1**. Besides the three categories of contacts, the total number of contacts in a protein is also important if we are to utilize the contacts to reconstruct three-dimensional models for the protein. Certain proteins, such as those having long tail like structures, have fewer contacts and are difficult to reconstruct even using true contacts while others, for example compact globular proteins, have a lot of contacts and can be reconstructed with high accuracy. Another important element of predicted contacts is the coverage of contacts, i.e., how well the contacts are distributed over the structure of a protein. A set of contacts having low coverage will have most of the contacts clustered in a specific region of the structure, which means that even if all predicted contacts are correct, we may still need additional information to reconstruct the protein with high accuracy.

Among all the contacts, long-range contacts, which are generally harder to predict [13, 14, 15, 16], are relatively more useful for structure reconstruction [17]. Hence, recent contact prediction methods focus on the prediction and evaluation of long-range contacts, and so do the CASP experiments (http://www.predictioncenter.org). When the contact prediction category was introduced in the CASP experiments, in the initial rounds, methods like SVMcon [14] and DNcon [13] that use support vector machines and deep learning networks with traditional features such as sequence profile, secondary structure and solvent accessibility, were often the top performers demonstrating that machine learning techniques were useful for contact prediction. Recent methods like PconsC2 [18], MetaPSICOV [16] and RaptorX method [19] show that including contact predictions from coevolution-based methods like CCMpred [20], PSICOV [21], and FreeContact [22] as additional

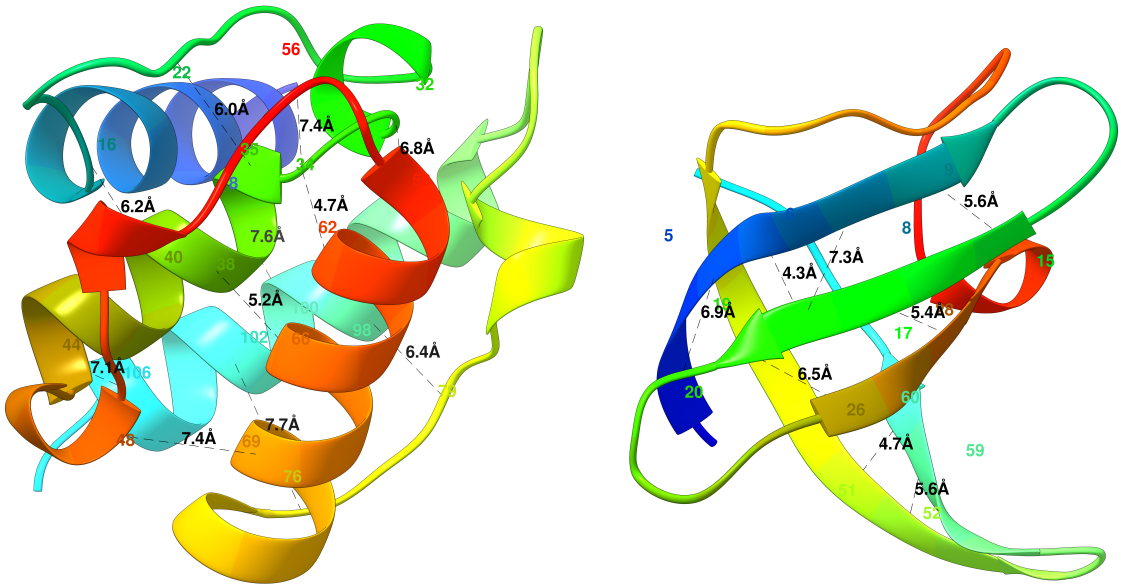

Figure 1.1: Two globular proteins with some contacts in them shown in black dotted lines along with the contact distance in Armstrong. The alpha helical protein 1bkr (left) has many long range contacts and the beta sheet protein 1c9o (right) has more short and medium range contacts.

features can significantly improve the performance, if at least a few hundred homologous sequences can be found for an input sequence. Often, when sufficient homologous sequences can be found, these 'meta' methods can predict top L/5 or L/10 long-range contacts with pretty high precision [16, 19, 20], where L is the length of the protein sequence. All these recently successful methods highlight that, besides machine learning techniques, coevolution-based features are important for accurate contact prediction.

Predicted contacts are evaluated using precision, i.e., the number of contacts that are correct out of all predicted contacts. For a lot of proteins as few as 8% of native contacts are sufficient to reconstruct the fold of proteins [23]. Moreover, all proteins do not have their number of contacts proportional to the sequence length. Hence, it is common to evaluate the top L/2 or just the top L/5 predicted contacts using precision, with L being the sequence length of the protein. Since short/medium range contacts are relatively easier to predict (especially for proteins having beta-sheets), the CASP competition focuses on evaluating predicted long-range contacts. The evaluation of contact prediction using precision is simple and is currently being used widely, but it does not cover two important aspects: coverage and number of contacts. Predicted top L/5 contacts may be highly precise, but can have a low coverage, such that they only cover a part of the protein and, thus, cannot capture the overall fold of the protein. Debora et al. attempted to qualitatively assess the coverage of contacts and Eickholt et al. discussed evaluating coverage using the idea of omitting

neighboring contacts [12, 24], and yet, the question of how many predicted contacts are needed to fold a protein remains unanswered. Although some authors have suggested that the number of contacts needed can be specific to prediction methods, the top 1L contacts have shown to produce good results [25, 26].

In this thesis, three contributions are described – (a) DNCON2, a two-level convolutional neural network-based method for protein contact prediction, (b) ConEVA [27], a toolkit for contact assessment and evaluation, and (c) CONFOLD [25], a method of building protein 3D structures from predicted contacts and secondary structures. Chapter 2 of this dissertation describes the DNCON2 method. Chapter 3 is on the ConEVA toolkit published in the BMC Bioinformatics journal and Chapter 4 is on the CONFOLD method published in the Proteins journal. In addition, Chapter 5 discusses a contact prediction method we developed and was ranked among the top predictors in the CASP12 contact prediction experiment. Chapter 6 discusses the reconstruction of protein structures using true contacts and secondary structure. Finally, in Chapter 7, we describe how the methods we developed can be utilized for ab initio protein structure prediction. Finally, in the last chapter we discuss summary and possible future works. The contents for the chapter 3 and 4 are from the following publications:

***Adhikari B**, Bhattacharya D, Cao R, Cheng J. CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins. 2015;83:1436–49. [25]*

***Adhikari B**, Nowotny J, Bhattacharya D, Hou J, Cheng J. ConEVA: a toolbox for comprehensive assessment of protein contacts. BMC Bioinformatics. 2016;17:517. [27]*

Chapter 2

DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks

2.1 Abstract

Significant improvements in the prediction of protein residue-residue contacts are observed in the recent years. These contacts, predicted using a variety of coevolution-based and machine learning methods, are the key contributors to the recent progress in ab initio protein structure prediction, as demonstrated in the recent CASP experiments. Continuing the development of new methods to reliably predict contact maps is essential to further improve ab initio structure prediction. In this paper, we discuss DNCON2, an improved protein contact map predictor based on two-level deep convolutional neural networks. It consists of six convolutional neural networks – the first five predict contacts at 6, 7.5, 8, 8.5, and 10 Å distance thresholds, and the last one uses these five predictions as additional features to predict final contact maps. On the free-modeling datasets in CASP10, 11, and 12 experiments, DNCON2 achieves mean precisions of 35%, 50%, and 53.4%, respectively, higher than 30.6% by MetaPSICOV on CASP10 dataset, 34% by MetaPSICOV on CASP11 dataset, and 46.3% by Raptor-X on CASP12 dataset, when top L/5 long-range contacts are evaluated. We attribute the improved performance of DNCON2 to the inclusion of short- and medium-range contacts into training, two-level approach to prediction, use of the state-of-the-art optimization and activation functions, and a novel deep learning architecture that allows each filter

in a convolutional layer to access all the input features. DNCON2 is currently available as a web-server at http://sysbio.rnet.missouri.edu/dncon2 where the predictions for CASP10, 11, and 12 free-modeling datasets can also be downloaded.

2.2 Introduction

In recent years, protein residue-residue contacts have been identified as a key feature for accurate de novo protein structure prediction [26, 28, 29, 30, 31]. Successful de novo structure prediction methods, in the recent CASP experiments, have attributed much of their performance to the incorporation of predicted contacts [17, 32, 33]. In terms of usefulness, contacts with sequence separation of at least 24 residues, i.e. long-range contacts, have been found more useful in structure modeling and are usually the primary evaluation target for evaluating and comparing contact-prediction methods. While long-range contacts are most useful for folding proteins using fragment-based methods like FRAGFOLD [34], Rosetta [17] and Quark [33], for fragment-free methods like CONFOLD [25] and GDFuzz3D [35], other two types of contacts - short- and medium-range contacts - are also important. While successful contact prediction methods like DNCON [13] find it effective to predict these separately, a more recent trend of predicting all contacts with a single machine learning architecture appears promising [16, 19].

Much of the recent improvement in the performance of contact prediction is from detecting coevolving residue pairs in a multiple sequence alignment and from the machine learning techniques used to integrate these predictions as features along with other standard features. Coevolution-based contact predictors can generally predict accurate contacts in presence of at least a few hundred effective sequences in the input alignment [36]. However, recent state-of-the-art methods demonstrate that integrating these co-evolution-based predictions with other features and using a machine learning method to make final predictions, can almost performs better than using coevolution information alone. These integrative contact predictors have used neural networks [16], random forests [37], and convolutional neural networks [19] to combine co-evolutionary features with other common features like secondary structures and position specific scoring matrices.

As a successor of our deep belief network based contact predictor, DNCON [12, 13], which was ranked as the top method in the CASP10 experiment [11], in this paper, we present our improved contact prediction method - DNCON2. The primary enhancements of DNCON2 are (a) inclusion of coevolution-based features, (b) new deep convolutional neural networks to predict full contact maps, and (c) addition of new features at multiple distance thresholds, which further improve the

performance. In DNCON2, we transform all 27 input features, e.g., scalar features like protein length, one-dimensional (1D) features like secondary structure prediction, and two-dimensional (2D) features like coevolution-based predictions, into 56 two-dimensional features. As the first step of our two-level prediction approach, we train five convolutional neural networks (CNNs) which accept these 56 2D features and predict contact maps at distance thresholds of 6, 7.5, 8, 8.5, and 10 Å. In the second level, a separate CNN is trained with these five sets of predictions as additional 2D features, to make final short-, medium-, and long-range predictions in one contact map all at once. Finally, we test our method using the free-modeling datasets of CASP10, 11, and 12 and compare it with other state-of-the-art methods, and, also discuss how the various training hyperparameters influence the performance.

2.3 Methods

2.3.1 Datasets and evaluation metrics

We used the original DNCON dataset consisting of 1426 proteins having length between 30 and 300 residues curated before the CASP10 experiment to train and test DNCON2. The protein structures in the dataset were obtained from the Protein Data Bank (PDB), had 0-2 Å resolution and were filtered by 30% sequence identity. 1230 proteins from the dataset are used for training and 196 as the validation set, and the two sets have less than 25 percent sequence identity. In addition to the validation dataset, we benchmarked our method using (a) 37 free-modeling domains in the CASP12 experiment, (b) 30 free-modeling domains in the CASP11 experiment [38], and (c) 15 free-modeling domains in the CASP10 experiment [11]. These CASP free-modeling datasets have zero or very little identity with the training dataset.

In this study, we define a pair of residues in a protein to be in contact if their carbon beta atoms (carbon alpha for glycine), are closer than 8 Å in the native structure. We consider contacts as long-range when the paring residues are separated by at least 24 residues in the protein sequence. Similarly, medium-range contacts are pairs which have sequence separation between 12 and 23 residues and short-range contacts are pairs with sequence separation between 6 and 11 residues. These definitions are consistent with the common standards used in the field [39].

As a primary evaluation metric of contact prediction accuracy, we use the precision of top L/5 or L/2 predicted long-range contacts, where L is the length of the predicted contacts. The metric has also been the main measure in the recent CASP evaluations [10, 11, 39] and some recent studies

[27]. When evaluating the predictions for the proteins in the CASP datasets, we evaluate them at the domain level to be consistent with the past CASP assessments, although all predictions were made on the full target sequences without any knowledge of domains. We used the ConEVA tool to carry out all our evaluations [27].

2.3.2 Input features

In addition to the existing features used in the original DNCON, we used new features derived from multiple sequence alignments, coevolution-based predictions, and three-state secondary structure predictions from PSIPRED [40]. The original DNCON feature set includes length of the protein, secondary structure and solvent accessibility predicted using the SCRATCH suite [41], position specific scoring matrix (PSSM) based features (e.g. PSSM sums and PSSM sum cosines), Atchley factors, and several pre-computed statistical potentials. During our experiments, we found PSSM and amino acid composition from the original DNCON feature set were not very useful and hence removed them from the feature list. Besides the DNCON features, the new features include coevolutionary contact probabilities/scores predicted using CCMpred [20], FreeContact [22], PSICOV [21], and alignment statistics such as number of effective sequences, Shannon entropy sum, mean contact potential, normalized mutual information, and mutual information generated using the alignment statistics tool 'alnstat' [16]. During our experiments, often, PSICOV did not converge when there are too many or too few alignments, especially if the target sequence is long. To guarantee to get some results, we set a time limit of 24 hours, and run PSICOV with three convergence parameters ('d = 0.03', 'r = 0.001', and 'r = 0.01') in parallel. If the first prediction (with option d = 0.03) finishes within 24 hours, we use the prediction, and if not, we use the second prediction and so on. Using all these features above as input, we predict contact maps at 6, 7.5, 8, 8.5, and 10 Å distance thresholds at first, and then use these five contact-map predictions as additional features to make a second round of prediction. Contact predictions at lower distance thresholds are relatively sparse and include only the residue pairs that are very close in the structure, whereas, contact predictions at higher distance thresholds are denser and provide more positive cases for the deep convolution neural network to learn.

2.3.3 Generating multiple sequence alignments

Generating a diverse/informative multiple sequence alignment with a sufficient number of sequences is critical for generating quality coevolution-based features for contact prediction. On one hand,

having too few sequences in the alignment, even though they may be highly diverse, can lead to low contact prediction accuracy. On the other hand, having too many sequences can slow down the process of co-evolution feature generation, creating a bottleneck for an overall structure prediction pipeline. To reliably generate multiple sequence alignments, an alignment method should produce at least some sequences in alignment whenever possible, and does not generate too many more sequences than necessary. Following a similar procedure in [17] and [42], we first run HHblits [43] with 60% coverage thresholds, and if a certain number of alignments are not found (usually around 2L), then we run JackHMMER [44] with e-value thresholds of $1E^{-20}$, $1E^{-10}$, $1E^{-4}$ and 1 until we find some alignments. JackHMMER is not run if HHblits can find at least 5000 sequences. These alignments are used by the three coevolution-based methods (CCMpred, FreeContact, and PSICOV) to predict contact probabilities / scores, which are used as two-dimensional features and to generate alignment statistics related features for deep convolutional neural network.

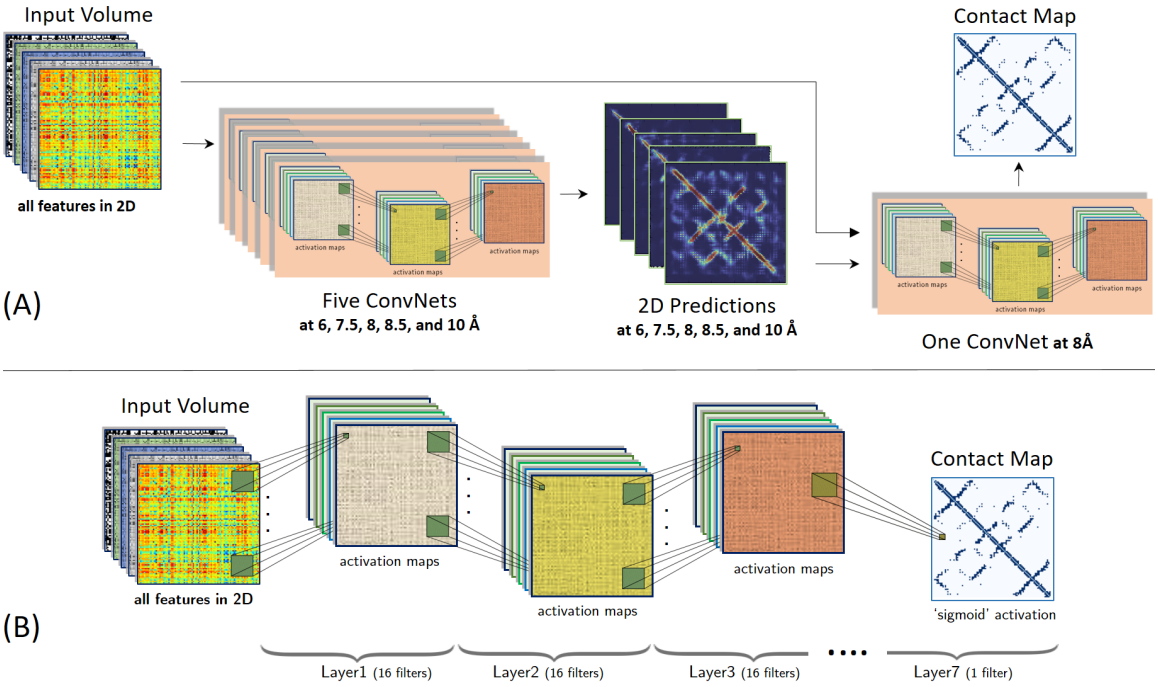

Figure 2.1: (A) The block diagram of DNCON2s overall architecture. The 2D volumes representing a proteins features are used by five convolution neural networks to predict preliminary contact probabilities at 6, 7.5, 8, 8.5 and 10 thresholds at the first level. The preliminary 2D predictions and the input volume are used by a convolutional neural network to predict final contact probability map at the second level. (B) The structure of one deep convolutional neural network in DNCON2 consisting of six hidden convolutional layers with 16 5x5 filters and an output layer consisting of one 5x5 filter to predict a contact probability map.

2.3.4 Deep convolutional neural network architecture

Convolutional neural networks (CNNs) are widely applied to recognize images with each input image translated into an input volume such that the size of the image are length and width of the volume, and the three channels (hue, saturation, and value) represent the depth. Based on such ideas, to build an input volume for each protein, we translate all scalar and one-dimensional input features into two-dimensional features (channels) so that all features (including the ones already in 2D) are in two-dimension and can be viewed as separate channels. While scalar features like sequence length are duplicated to form a two-dimensional matrix (one channel), each one-dimensional feature like solvent accessibility prediction is duplicated across the row and across the column to generate two channels. The size of the channels for a protein is decided by the length of the protein. By having all features in separate input channels in the input volume, each filter in a convolutional layer convolving through the input volume, has access to all the input features, and can learn the relationships across the channels. Compared to the input volumes of images that have three channels, our input volumes have 56 channels.

We use a total of six CNNs, i.e. five in the first level to predict preliminary contact probabilities at 6, 7.5, 8, 8.5, and 10 Å distance thresholds separately by using an input volume of a protein as input, and one in the second level that take both the input volume and the 2D contact probabilities predicted in the first level to make final predictions (**Figure 2.1 (A)**). Each of the six CNN networks have the same architecture, which has six hidden convolutional layers and one output layer consisting of 16 filters of 5 by 5 size and one output layer (**Figure 2.1 (B)**). In the hidden layers, the batch normalization is applied, and 'Rectified Linear Unit' [45] is used as the activation function. The last output layer consists of one 5 by 5 filter with 'sigmoid' as the activation function to predict final contact probabilities. Hence, our deep network can accept a protein of any length and predict a contact map of the same size. We use the Nesterov Adam (nadam) method [46] as the optimization function to train the network.

We train each CNN for a total of 1600 epochs with each epoch of training taking around 2 minutes. After training, we rank and select best model using the mean precision of top L/5 long-range contacts calculated on the validation dataset of 196 proteins. Our raw feature files for all 1426 proteins use 8 GigaBytes (GB) of disk space and are expanded to around 35 GB when all features are translated to 2D format. To minimize disk input/output, we translate our scalar features and 1D features into 2D, at runtime, in CPU memory. We used the Keras library (http://keras.io/) along with Tensorflow (www.tensorflow.org) to implement our deep CNN networks. Our trainings

were conducted on Tesla K20 Nvidia GPUs each having 5 GB of GPU memory, on which, training one model took around 12 hours. Finally, we use an ensemble of 20 trained deep models to make final predictions for testing.

2.4 Results

2.4.1 Using contact predictions at 6, 7.5, 8, 8.5, and 10 Å distance thresholds as features improves precision

A contact map is a binary version of the distance map of a protein structure according to a distance threshold, which usually is 8 Å. This threshold of 8 Å, although widely used, can be viewed as an arbitrary and rigid criterion to decide if a pair of residue is a contact or non-contact. For instance, a pair separated by 8.1 Å distance is, by definition, a non-contact, but by 7.9 Å is a contact. And using one distance threshold causes the loss of some distance information. In order to account for uncertainty and ambiguity in residue-residue distance, in a first round of prediction, using all the features and true contact maps at 6, 7.5, 8, 8.5, and 10 Å distance thresholds, we trained five CNN models to predict contact probabilities at these five distance thresholds. Then, in the second round of prediction, we added these predictions as new 2D features into the feature list and trained a sixth CNN model to predict contacts at 8 Å distance threshold.

On the 196 proteins in the validation dataset, the CNN model in the second level achieves a precision of up to 73.5 % higher than 70.7 % in the first level, when top L/5 long-range contacts are evaluated. To verify if the improvement comes from the predictions at different distance thresholds or from the iterative two-level training, in a separate experiment, we trained a second level model with only the contact prediction at 8 Å distance threshold as additional feature. In this case, a precision of 72.2 % is achieved, higher than 70.7% of using one level prediction, but lower than 73.5% of using both two-level prediction and multiple thresholds, indicating that both two-level training and multiple thresholds contribute to the improvement. The results summarized in **Figure 2.2** show similar results when top L/2 contacts are evaluated. In addition to these experiments, we tested adding more predictions at higher distance thresholds of 12, 14, 16, and 18 Å as features, and found that they did not significantly improve the performance. As an additional validation, we used the ensemble of the models trained with five distance thresholds in the first level to predict the contacts for the proteins in the validation dataset, similarly as a traditional neural network ensemble in [16]. Such a multi-distance ensemble has a precision of 72.8% slightly higher than the 72.6% precision

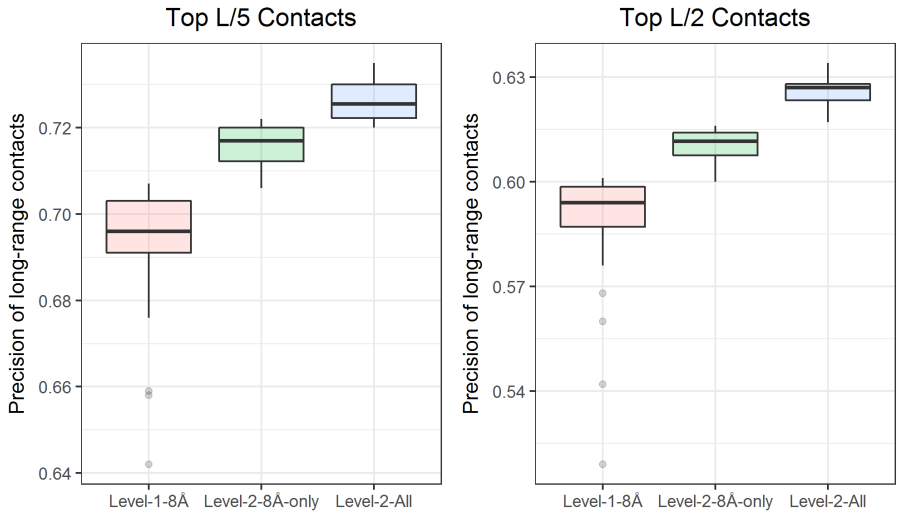

Figure 2.2: The improvement from inclusion of predictions at distance thresholds of 6, 7.5, 8, 8.5, and 10 as additional features, measured using the precision of top L/5 (left) and top L/2 (right) long-range contacts on the validation dataset. Box plot of precision for best 30 of 40 models for the level one model trained only using the original features (pink), the level-two model trained using only 8 prediction as additional feature (green), and the level-two model trained by adding all five predictions at multiple thresholds as additional features (blue).

achieved by an ensemble of all five models trained at the same 8 Å distance threshold, but lower than 73.5% of using predictions of multiple thresholds with two-level networks in DNCON2.

2.4.2 Comparison between deep belief network in DNCON 1.0 and deep convolutional neural networks in DNCON2

DNCON 1.0 used an ensemble of deep belief networks (DBN) trained with windows of seven different fixed sizes and boosting to predict contacts and achieved an accuracy of 34% on the 196 proteins in the validation dataset. For a fair comparison with DNCON, we trained one CNN using the same features that DNCON 1.0 used (excluding new coevolution-based features). Different from the DNCON 1.0 of using different networks to predict contacts at different ranges, DNCON2 uses a CNN network to predict short-, medium-, and long-range contacts of a protein of arbitrary length. With the same features as input, a CNN network trained with all contacts and non-contacts achieves a slightly better precision of 35.4 % on top L/5 long-range contacts than DNCON 1.0. So, a single CNN model performs better than a boosted and ensembled of deep belief networks, suggesting that the deep convolutional neural network (CNN) is more suitable for contact prediction than the deep belief network (DBN). Moreover, it is more convenient to train and test CNN than DBN because CNN can take a full input matrix of arbitrary size as input to predict full contact maps without generating

the features for each pair of residues, separating contacts at different ranges, and balancing the ratio of contacts and non-contacts as required by DBN based on fixed size windows.

2.4.3 Performance of DNCON2 on the validation and CASP datasets

On the 196 proteins in the validation dataset, compared to the 35.4% precision of one CNN without using coevolution-based features, by adding coevolution-based features and using multiple CNNs in the two levels CNNs, DNCON2 yields a mean precision of 74%, when top L/5 long-range contacts are evaluated. As summarized in **Table 2.1**, the average length, number of sequences in the alignment, and the number of effective sequences for these proteins are 190, 5,351, and 1,718 respectively. On this dataset, the three individual coevolution-based features generated by CCMpred, FreeContact, and PSICOV can predict contacts with precisions of 51.0%, 43.1%, 42.1% respectively, for top L/5 long-range contacts, which is much lower than 74% of DNCON2. And for 96% of these proteins, DNCON2 performs better than any of the individual coevolution-based features. These results indicate that integrating all the 2D coevolution-based features with the other features can drastically improve the accuracy of contact prediction.

Table 2.1: Performance of DNCON2 on the 196 proteins in the validation dataset when top L/5 and top L/2 long-range contacts are evaluated. L, N, and N_{eff} stand for length of a protein, number of sequences in the alignment, and the number of effective sequences in the alignment. $P_{L/5}$ and $P_{L/2}$ are the precisions of top L/5 and L/2 long-range contacts.

	L	**N**	$\mathbf{N}_{eff}$	$\mathbf{P}_{L/5}$	$\mathbf{P}_{L/2}$
Average	190	5351	1718	74.0%	64.2%
Median	188	1607	412	88.2%	74.4%
Maximum	299	62889	29547	100.0%	100.0%
Minimum	50	1	1	0.0%	0.0%

Since predicted contacts are most useful for ab initio folding of proteins whose structures cannot be predicted by template-based modeling, we evaluated our method on the free-modeling protein datasets in the CASP10, 11, and 12 experiments and compared it with top CASP methods and a standard coevolution-based method MetaPSICOV [16] (see **Table 2.2**). Since our training and validation datasets were curated before the CASP10 experiment, the CASP datasets are independent test data. For evaluating our method on the most recent CASP12 dataset, we generated all features using all programs and databases released before the CASP12 experiment started, making our results not influenced by the new releases of protein structures and sequences thereafter. On the 37 free-modeling CASP12 domains, for which native structures were available for us to perform the evaluation, DNCON2 outperforms all the top methods participating in the CASP12

experiment such as Raptor-X [19], MetaPSICOV [16], iFold_1, and our own method MULTICOM-CONSTRUCT as well as the baseline method DNCON 1.0. When top L/5 long-range contacts are evaluated, DNCON2 achieves an average precision of 53.4% compared to 46.3%, 42.9%, and 45.7% by Raptor-X, MetaPSICOV, and iFold_1, respectively. A similar performance is observed when top L/2 contacts are evaluated instead of L/5. The 24.9% precision of DNCON 1.0, which does not use any coevolution-based features, is a benchmark for other methods, and the difference between its accuracy with the other methods highlights the improvement gained from the inclusion of the coevolution-based features into the input.

Table 2.2: Summary of the performance of DNCON2 on the 15 CASP10, 30 CASP11, and 37 CASP12 free-modeling (FM) structural domains, measured using the precision of top L/5 long-range contacts. The precision of the top method in each CASP experiment and a standard method MetaPSICOV (run locally) is also included as a reference.

FM Dataset	**Domain Count**	**Precision of top L/5 long-range contacts (%)**		
		Top CASP Group	**MetaPSICOV**	**DNCON2**
CASP10	15	18.1 (DNCON 1.0)	30.6	35.0
CASP11	30	29.7 (CONSIP2)	34.4	50.0
CASP12	37	46.3 (Raptor-X)	42.9	53.4

For evaluating our method on CASP11 and CASP10 free-modeling datasets, we ran MetaPSICOV locally to use as a benchmark. For a fair comparison, we use the same sequence databases for DNCON2 and MetaPSICOV. For completeness, we also compared DNCON2 with the best performing groups in the two CASP experiments - CONSIP2 in the CASP11 experiment [39], and DNCON 1.0 in the CASP10 experiment [11] (**Table 2.2**). On the 30 free-modeling domains in the CASP11 experiment, DNCON2 achieves an average precision of 50% compared to 34.4% by MetaPSICOV and 29.7% by the best performing method CONSIP2 [42] in CASP11, when top L/5 long-range contacts are evaluated. Similarly, on the 15 free-modeling structural domains in the CASP10 experiment, DNCON2 achieves a mean precision of 35%, compared to 21.1% by MetaPSICOV, and 19.4% by the best performing method DNCON. For both datasets, similar results are observed when medium-range and short-range contacts are evaluated.

2.4.4 Hyper-parameters optimization

To obtain best performance on the validation dataset, we fine-tuned our network by investigating a range of values/options for the following hyper-parameters: (a) depth of the network, (b) filter sizes in each layer, (c) number of filters in each layer, (d) batch normalization, (e) batch size, (f) optimization function, and (g) activation function. After many rounds of iterative hyper-parameter

selection, we found that the optimal parameters for number of layers was seven, filter size was five, number of filters was 16, batch size was 30, and chosen ReLU as the activation function in hidden layers, applied batch normalization at each layer, and used NAdam as the optimization function. With the performance of the CNN in a setting as a reference, we tuned each parameter, one-by-one, to study how they influenced the performance on the training data and validation data. For the depth of the network we tested networks with two to nine layers. Similarly, for filter size and number of filters, we tested filter sizes of 1, 3, 5, 7, 9 and 11, and number of filters as 1, 4, 8, 16, and 24. On the validation dataset, the networks with filter sizes more than 3, with 8 or more filters and 5 or more hidden layers deliver around the top performance. When the filter size is increased beyond 9, or the number of filters is increased beyond 24, or the depth of the network is increased beyond 9, the training was very slow and often failed because of insufficient GPU memory. In addition, through trials, we found that keeping the filter size in all seven layers and number of filters in the six hidden layers the same performs better than having different filter sizes or numbers of filters in different layers.

Batch normalization is important for training deep CNN to deal with the covariate shift problem [47]. To test how batch normalization affects the training performance, we tried applying batch normalization after each layer (a), after every alternate layer (b), or only on the first layer (c), and not using batch-normalization at all (d). We found that applying batch normalization at each layer delivers the best performance compared to any of the three other settings. While the full batch normalization applied after each layer delivers a mean precision of 70.8% and the batch normalization at every alternate layer results in a mean precision of 68.7%, for top L/5 predicted long-range contacts on the validation dataset. When batch normalization is not used at all, or is applied only to the first layer, the precision drops to 65.7%. Similarly, after testing various batch sizes, we found batch sizes of around 30 delivered the best performance on the validation dataset. For optimization methods, we tested (a) ADADELTA, (b) Adagrad, (c) Adam, (d) Nesterov Adam, (e) RMSprop, and (f) stochastic gradient descent optimizers. The results show that three optimization functions Adam, Nesterov Adam, and RMSprop deliver better performance than the others, with Nesterov Adam performing best among all. Finally, the activation functions sigmoid, tanh, and ReLU can achieve the precisions of 70.4%, 69.4%, and 70.9%, respectively, when top L/5 long-range contacts are evaluated.

Besides the machine learning hyperparameters, we also tested if training using only long-range contacts improves the precision of long-range contact prediction. Interestingly, we find that including medium-range contacts and short-range contacts into training improves the performance even when

only long-range contacts are evaluated. However, when all the local contacts are included, i.e. contacts with sequence separation less than five, we observed a slight decrease in performance. In summary, including all except the local contacts during training yields better precision. In addition, to test how ensembling improves the performance, we first ranked the trained models using the average precision on the validation dataset. Then, we calculated precision of the averaged predictions of top x models, where x is an integer in the range [1, 50]. The precision of ensembling increases initially and then saturates after more than four models are used.

2.4.5 Importance of features

We removed one or more features at a time and trained the CNN using the remaining features, to study the contribution of the removed features towards the overall performance of DNCON2. We tested by removing (a) multiple sequence alignment (MSA) statistics related features comprising of Shannon entropy sum, mean contact potential, normalized mutual information, and mutual information, (b) CCMpred coevolution feature, (c) FreeContact coevolution feature, (d) PSICOV coevolution feature, (e) several pre-computed statistical potentials, (f) number of sequences in the alignment and the number of effective sequences in the alignment, (g) PSIPRED and PSISOLV predictions of secondary structures and solvent accessibility, (h) PSSM related features comprising of PSSM sums and PSSM sum cosines, (i) SCRATCH secondary structure and solvent accessibility predictions, (j) relative counts of helical residues, strand residues, and buried residues, (k) sequence separation related features, and (l) length of the protein. Our results, summarized in **Figure 2.3**, show that the features from multiple sequence alignment related statistics are more important than any single coevolution-based features (CCMpred, FreeContact, or PSICOV). We find the length feature unimportant. Sequence separation related features and relative counts of helical, strand, and buried residues, also do not contribute much to the performance. Secondary structure predictions from both methods SCARTCH and PSIPRED are useful, and complement each other to improve the overall performance. Among the three coevolution-based features CCMpred, FreeContact, and PSICOV, the first two (CCMpred and FreeContact) contribute equally to the overall performance. If all three coevolution-based predictions (CCMpred, FreeContact, and PSICOV) are removed, the precision drops from 60% to 38%, when top L/2 long-range contacts were evaluated, suggesting that the three coevolution based features (combined) are the most important features.

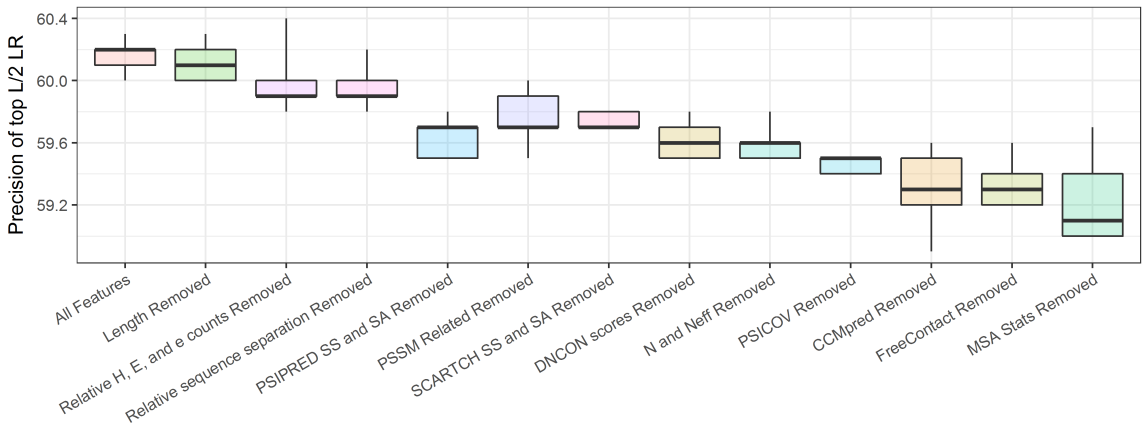

Figure 2.3: Importance of features measured by calculating the best of five precisions of top L/2 long-range contacts on the validation dataset after removing a feature or a set of features. MSA Stats features are multiple sequence alignment (MSA) statistics related features comprising of Shannon entropy sum, mean contact potential, normalized mutual information, and mutual information, DNCON scores are set of several pre-computed statistical potentials, N and Neff are number of se-quences and effective number of sequences. If all three coevolution-based predictions (CCMpred, FreeContact, and PSICOV) are removed (not shown in the plot), the precision drops from 60% to 38%, when top L/2 long-range contacts were evaluated.

2.5 Conclusion

We developed DNCON2 - a new two-level deep convolutional neural network method - to predict the contact map of a protein of any length by integrating both residue-residue coevolution features and other features such as secondary structures, solvent accessibility, and pairwise contact potentials. The method can predict all the contacts in a protein at once from the entire input information of a protein, which is more effective and easier to train and use than local fixed window-based approaches such as deep belief networks. By including new co-evolution features, using CNNs of multiple-distance thresholds, integrating all the features of all the residues through 2D-convolution in a two-level architecture, and adopting the latest optimization and training techniques, DNCON2's accuracy is more than double of that of DNCON 1.0 on the same validation dataset. On the three independent CASP datasets, DNCON2 outperforms the top methods in CASP10, CASP11, and CASP12 experiments. The results demonstrate DNCON2 and its deep convolutional neural network architecture is useful for protein contact prediction.

Chapter 3

ConEVA: a toolbox for comprehensive assessment of protein contacts

3.1 Abstract

In recent years, successful contact prediction methods and contact-guided ab initio protein structure prediction methods have highlighted the importance of incorporating contact information into protein structure prediction methods. It is also observed that for almost all globular proteins, the quality of contact prediction dictates the accuracy of structure prediction. Hence, like many existing evaluation measures for evaluating 3D protein models, various measures are currently used to evaluate predicted contacts, with the most popular ones being precision, coverage and distance distribution score (X_d). We have built a web application and a downloadable tool, ConEVA, for comprehensive assessment and detailed comparison of predicted contacts. Besides implementing existing measures for contact evaluation we have implemented new and useful methods of contact visualization using chord diagrams and comparison using Jaccard similarity computations. For a set (or sets) of predicted contacts, the web application runs even when a native structure is not available, visualizing the contact coverage and similarity between predicted contacts. We applied the tool on various contact prediction data sets and present our findings and insights we obtained from the evaluation of effective contact assessments. ConEVA is useful for a range of contact related analysis and evaluations including predicted contact comparison, investigation of individual protein

folding using predicted contacts, and analysis of contacts in a structure of interest. ConEVA is publicly available at http://iris.rnet.missouri.edu/coneva/. The source code of a light-weight downloadable version of ConEVA and the source code of the web-server are all hosted in GitHub at http://github.com/multicom-toolbox/ConEVA.

3.2 Background

The success of many protein residue contact prediction methods, in the recent years, has kindled a new hope to solve the long standing problem of ab initio protein structure prediction [13, 14, 16, 20, 21, 22]. Consequently, contact-guided ab initio structure prediction has emerged as an important field. When accurately predicted contacts are supplied as input to structure prediction or reconstruction methods, accurate folds can be predicted consistently [16, 25, 29, 34]. In general, accurate contacts lead to accurate structural models. However, for predicting folds of sequences which do not have homologous templates (hard sequences), the optimal way of utilizing predicted contacts is still an ongoing research. For instance, experiments on true contact reconstruction have suggested that 9 Å or more distance threshold delivers best reconstruction with Cβ atom [4, 6], but the Critical Assessment of Protein Structure Prediction (CASP)'s definition of 8 Å threshold is still widely used to predict contacts [11, 13, 16, 20, 22]. Marks et al. have even demonstrated successful structure predictions using Cα atoms and 7 Å threshold for defining contacts [24]. Similarly, it is widely accepted that long-range contacts [10, 11, 48] are the most useful of the three contact types (short-, medium-, and long-range), but some structural domains introduced in CASP like T0765-D1, T0709-D1, T0711-D1, T0756-D2, T0700-D1 have very few or no long-range contacts at all. In addition, Michel et al. discuss some examples of proteins that could not be accurately reconstructed despite high accuracy of predicted contacts in their PconsFold method [26]. Using the protein 1JWQ, Vassura et al. show how some structures cannot be folded with distance thresholds below 16 Å [6]. Zhang et al. report folding 90 transmembrane proteins at 14 Å cut-off [49]. Furthermore, in these works, no common agreement is found on the optimal number of contacts (or a range) needed for accurate reconstruction.

Hence, a tool to study the relationship between contact parameters and structure types is deemed necessary. Currently, for evaluating predicted contacts, the three most widely used evaluation measures are precision, coverage and distance distribution score (X_d) [10, 11, 13, 15, 50, 51, 52, 53]. In addition, other measures like 'mean false positive error', 'distance in contact map' or 'spread' [24], F-score and Matthews correlation coefficient (MCC) [11] are also used for a more rigorous

evaluation of the predicted contacts. Osvaldo et al. [54] had published EVAcon in 2005 that could calculate some of these measures, which no longer seems accessible. On the other hand, existing tools like CMView [55] and CoeViz [56] only enable contact map visualization and multiple sequence visualization.

In this paper, we present ConEVA, a fast web application (along with a downloadable tool) for protein contact evaluation and comparison. Besides the server, we also report some of our observations obtained through the application of our tool on larger data sets. We discuss how the length of a protein can influence various evaluation measures, the minimum number of contacts to evaluate, and the range of the evaluation measure values associated with the determination of the correct fold of a protein.

3.3 Methods

3.3.1 Datasets

In this work, we often refer to the dataset of 150 diverse proteins with average length of 150 residues introduced by Jones et al. in the PSICOV paper [21]. This data set along with other examples, including many CASP data sets, are provided as pre-curated data sets available through the "All Examples" link in the web server homepage.

3.3.2 Contact definition

Other than the places where we explicitly mention, in this work we primarily use the CASP definition of contacts, which is – a pair of residues separated by at least 6 residues are said to be in contact if their Cβ atoms (Cα in case of Glycine) are closer than 8 Å.

3.3.3 Input and interface

The primary input to ConEVA is residue-residue contacts in CASP's RR file format, whose description is available at http://predictioncenter.org/casprol/index.cgi?page=format#RR. A single RR file or multiple RR files zipped into a single zip file can be supplied. Along with predicted contacts, a native structure in PDB file format [57], may be supplied for contact evaluation. For domain based evaluations, as performed in CASP evaluations, the domain structure may be supplied as native PDB file instead of the full target structure. Besides these data inputs, the server also allows

to specify if the input contacts are between Cα or Cβ atoms. In addition, a user can choose to evaluate short-, medium-, long-range, or all contacts by defining the sequence separation distances. **Figure 3.1** shows a screenshot of ConEVA input interface. Besides allowing users to supply contact RR files, many pre-curated data sets are available through the "All Examples" link in the homepage for users to test.

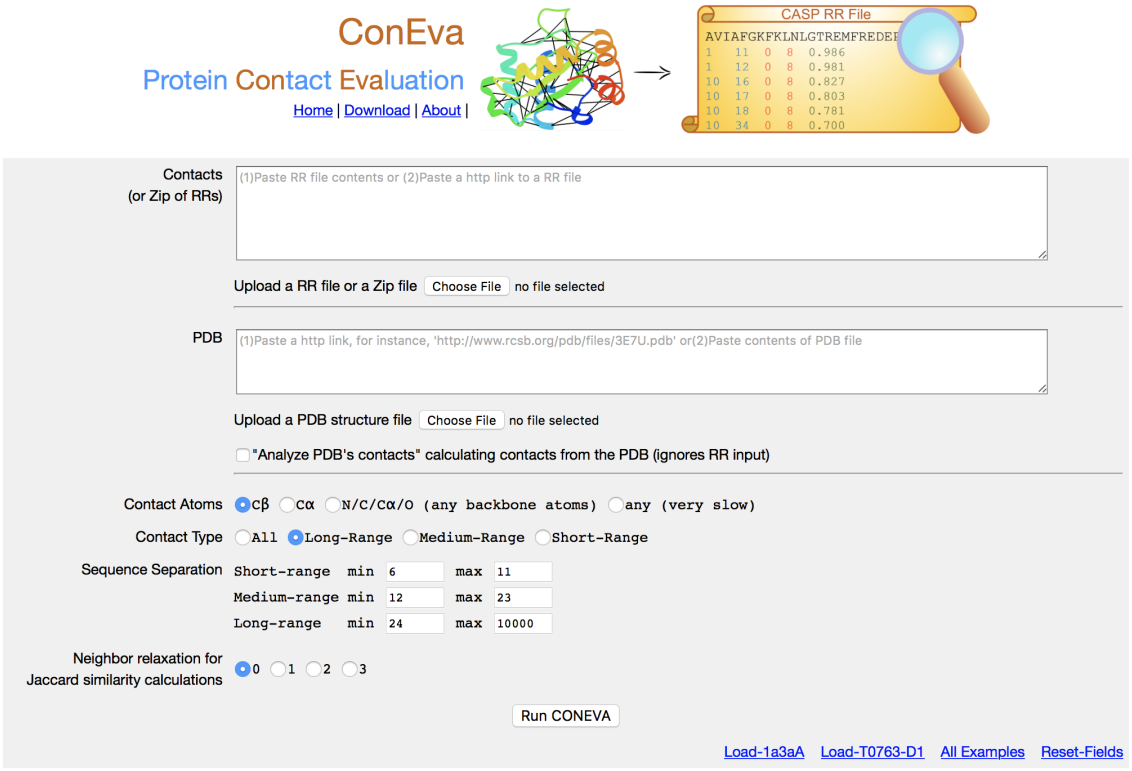

Figure 3.1: A screenshot of ConEVA homepage showing all input fields.

3.3.4 Server description

Input contacts are first sorted using the confidence column in the contact rows. Using the minimum and maximum sequence separation thresholds supplied for defining short-, medium- and long-range contacts, and the choice made for contact type (all/short-range/medium-range/long-range) contact rows that are not of a user's interest are filtered out. If a native structure is also supplied, contact residue pairs that do not exist in the native structure are filtered out. Then, the top-5, L/10, L/5, L/2, L, and top-2L contacts are selected and grouped for assessment. L is the length of the native chain when supplied, and otherwise, it is the length of the sequence for which contacts are predicted. Perl and Perl CGI is used for server development, and we use 'heatmap.2' function in the 'gplots' package [58] in R for visualizing Jaccard similarity matrix, and 'plotrix' package [59] for drawing

chord diagrams.

3.3.5 Sever Output

The web-server output is organized in various sections. The first section summarizes the input files, contacts computed from the native structure in EVACon format [54], sequence length of contacts file and native structure with a link to the sequence comparison, and a description of the definition of contact used for all following results. The next section tabularizes contact counts for short-, medium-, and long-range contacts, and for top-5, top-L/10, etc. up to top-2L contacts. Number of contacts that are not in native structure is also shown. In addition, if a native structure is provided as input, all numbers appear as hyperlinks to UCSF Chimera command line scripts, which can be downloaded and opened in UCSF Chimera to directly visualize the selected number of contacts within the native structure. The next section, visualizes Jaccard similarity matrices in the form of 'heatmap' and 'dendrogram' plots. The dendrogram shows similar contact sets in closer branches. Each plot has a link below it which links to the actual similarity matrix. The next section visualizes Chord diagrams. Contact maps appear in the next section, with native contact map shown in background. The subsequent sections present calculations and plots for precision, mean false positive error, coverage, X_d, and spread. ROC curves with calculations for Area Under the Curve (AUC) are displayed next, followed by precision-recall curves. The last two sections present calculations for Matthew's correlation coefficient and 1D visualization of coordination numbers. In the absence of a native structure, only the first five sections and the last section are reported, and further, if only a single contact prediction file is supplied, the section for Jaccard similarity calculations is skipped.

3.3.6 Measures computed on contacts

For each group of selected top contacts, coordination numbers [60] and contact maps are shown as 1D and 2D visualizations. Coordination number defines the number of contacts that a residue is involved in. Realizing the importance of contact assessment in the absence of a native structure, we introduce visualization and comparison using chord diagrams. See Discussion section for illustrations.

3.3.7 Quality measures with respect to native structure

For each group of these selected contacts the following evaluation measures are calculated: precision, coverage, mean false positive error, distance distribution score (X_d) [10, 11, 13, 15, 50, 51, 52, 53],

Spread [24], MCC [11], AUC_PR [61]. Precision is defined as the percentage of correctly predicted contacts, calculated as the ratio of the number of predicted contacts that are correct and the number of predicted contacts selected for evaluation, $Precision = \frac{TP}{TP+FP}$. The true positives (TP) and false positives (FP) are the number of correctly and incorrectly predicted contacts. For instance, when we select top five contacts for evaluation, TP+FP is fixed at five and TP can range from 0 to 5. Coverage is the percentage of true contacts contained in a predicted list of contacts, calculated as the ratio of the number of correctly predicted contacts and the total number of contacts in the native structure, $Coverage = \frac{TP}{N_c}$, where N_c is the number of true contacts in the native structure. Mean false positive error is calculated as the mean of absolute deviation of all the incorrectly predicted contacts, $Mean\ FP\ Error = \frac{1}{FP}\sum(d_{ij} - d)$, where d is the distance threshold for the contact definition (usually 8 Å) and d_{ij} is the actual distance of a false positive pair of predicted contacts in the native structure.

The distance distribution score (X_d) measures the weighted harmonic average difference between the predicted contacts distance distribution and the all-pairs distance distribution. While predicted contact distance distribution refers to the distribution of actual distances for the predicted contacts, all-pairs distance distribution is the distribution of distances for all the true contacts in the native structure. X_d is calculated as,

$$X_d = \sum_{i=1}^{15} \frac{PiP - PiA}{d_i * 15}$$

where the sum runs for 15 distance bins covering the range from 0 to 60 Å. d_i is the distance representing each bin, its upper limit (normalized to 60). PiP is the percentage of predicted pairs whose distance is included in bin i. PiA is the same for all the pairs and is zero for all bins with $d_i >$ 8 Å, such that the value of X_d increases heavily because of the contacts that are very incorrect, i.e. the contacts whose true distance is very large. Defined in this way, although the harmonic average reflects the difference between the real and predicted distances of residues, interpreting the meaning of a particular valued of X_d can be difficult. In general, for a given set of predicted contacts, $X_d >$ 0 indicates the positive cases where at least some contacts in the set are correct, whereas when $X_d$is closer to 0, the set can be considered random contacts.Spread [24] is computed using contact maps. For a given set of predicted contacts, it is the mean of the distances from every true contact to the nearest predicted contact in 2D contact map.

$$Spread = \frac{1}{N_c}\sum_{i=1}^{Nc} min\{dist\,(T_i - P)\}$$

where N_c is the number of true contacts, T_i is a true contact in the native structure, and $min\{dist\,(T_i - P)\}$ is the minimum Euclidean distance between the true pair T_i and all predicted residue pairs in the 2D contact map where every residue sequence separation is considered a unit.

3.3.8 Measures of similarity between predicted sets

In addition, for computing similarity between predicted contacts in the absence of native structure we introduce Jaccard similarity matrix [62] computations with neighborhood relaxation. For each pair of input contact sets, say A and B, we compute the Jaccard similarity score between A and B, J_{AB} as $J_{AB}=\frac{|A \cap B|}{|A \cup B|}$ where |A∩B| is the number of common contacts (intersection) between sets A and B, and |A⋃B| is the count of contacts in the set A union B. This similarity computation can evaluate to very small percentages in case of hard predictions because two sets must have precisely the same residue pair to be common, especially when we are evaluating top five or top L/10 contacts. For this reason, we introduce the idea of relaxing the similarity computation by considering contacts with ±N residue number deviation as same contact (N may be selected as 0, 1, 2 or 3). For instance, if set A has a pair 3-15 and set B has a pair 3-16, they may be considered as the same contact at N equal to 1. However, high similarity observed with N more than 1 in helical proteins can be sometimes misleading because shifts of two or more residues can have dramatic effect on the quality of the models generated using the contacts.

Besides these "reduced list" metrics [39] that only evaluate selected top contacts, ConEVA also presents "full list" metrics including Matthew's correlation coefficient (MCC), area under the precision-recall curve (AUC_PR) [61], and Receiver Operating Characteristic (ROC) curve. To calculate MCC for a set of predicted contacts, all contacts having confidence more than 0.5 are considered as predicted contacts to calculate true positive (TP), true negative (TN), false positive (FP) and false negative (FN) so that

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3.3.9 Contact prediction and model generation

Throughout this work, we use the publicly available contacts predicted by PSICOV [21]. In addition, we also installed a local copy of the tools coevolution based tool CCMpred [20], pure machine-learning based method DNcon [13], and a hybrid method MetaPSICOV [16] to make contact predictions

for various data sets including the PSICOV data set of 150 proteins. These contacts along with secondary structures predicted using PSIPRED [63] were used for building models using CONFOLD [25], a fragment-free ab initio method that we recently developed to build 3D models from scratch. As discussed in the CONFOLD paper, for each protein, we selected various top predicted contacts (top-5, L/10, L/5, L/2, L, and 2L) and built models using subsets, resulting in a total of 400 models for each protein. We selected the best model out of 400 for our analysis. To study how various evaluation measure correlate to the final quality of models reconstructed using the predicted contacts, we build 3D models with CONFOLD using the contacts predicted for the 150 proteins in the PSICOV dataset. We argue that the TM-score [64] of the best model can be used as a score that suggests the best utility of the predicted contacts.

3.4 Results

3.4.1 Dependence of evaluation measures on L

The length of the sequence may be ignored when we are evaluating and comparing contacts predicted for a single protein sequence. However, when we are comparing contact prediction methods on more than one protein sequence and the sequences are not of same length, sequence length can bias the comparisons. For instance, if the evaluation measures we choose to make the comparison is influenced by the length of the sequence and penalizes longer sequences more, then the methods that perform poorly particularly on longer sequences can be ranked lower than they should. This is also the reason why evaluation measures like TM-score were introduced to address the limitations of measures like RMSD. Thus, it is important to study how various contact evaluation measures are correlated to the length of the protein sequence.

Table 3.1: Spearmans rank correlation coefficient between the length of a protein (L) and evaluation measures for PSICOV predicted long-range contacts in the PSICOV data set. It shows that spread, coverage and X_d are more correlated to L and Nc than precision and mean false positive error, especially below top-L contact selection. For this da-taset, the lengths are distributed in the range [50, 266] with mean and standard deviation of 145 and 52 respectively.

Contact-Selection	Top-5	Top-L/10	Top-L/5	Top-L/2	Top-L	Top-2L
L vs Precision	-0.01	-0.07	0.06	0.24	0.26	0.27
L vs Coverage	-0.88	-0.59	-0.51	-0.34	-0.30	-0.31
L vs X_d	0.31	0.35	0.46	0.49	0.51	0.55
L vs FP-Error	-0.05	0.04	-0.06	-0.16	-0.01	0.21
L vs Spread	0.88	0.66	0.60	0.58	0.55	0.57

To study the relationship between length of the protein (L) and the quality of contacts suggested

by the various contact evaluation measures, we computed Spearman's rank correlation coefficient between the length of the protein and the evaluation measures – precision, coverage, X_d, mean false positive error, and spread – for the long-range contacts (with sequence separation more than 23) predicted in the PSICOV dataset. In **Table 3.1** we show that mean false positive error is the measure most uncorrelated with the length of a protein, followed by precision values for all contact selections (top-5 to top-2L). Spread and coverage are more correlated with the length at lesser contact selections (top-5, top-L/10 and top-L/5) whereas X_d is more correlated with L when we select more contacts for evaluation (top-L/2, top-L, and top-2L). Similar correlation values were obtained for the number of contacts in a protein (N_c). In summary, these observations lead us to argue that precision and mean false positive error are the most reliable measures when comparing contact predictions.

3.4.2 Number of contacts to evaluate

How many contacts should we evaluate, top-5 or top-L or top-2L? On one hand, reconstruction studies using true contacts focus on the minimum number of contacts needed to recover the fold of a protein. For instance, DE et al. suggest that 1 contact in every 12 residues is sufficient to robustly fold a protein at topology level [65]. This translates to L/12 predicted contacts if we assume that the contacts are spread out without any overlaps. In a similar study, introducing a novel cone-peeling algorithm, Sathyapriya et al. suggest that as little as 8% of the native contacts are sufficient to determine the tertiary structure [8]. On the other hand, contacts are currently evaluated on a wide range of contact selections. It is a common practice for CASP assessors to evaluate top-5, top-L/10, and top-L/5 predicted long-range contacts. Similarly, recent contact prediction methods that utilize the predicted contacts to build three dimensional models discuss evaluating top-L/10, L/5, L/2, up to top-L contacts [20, 21, 22].

We argue that the minimum set of contacts for which there is a high correlation between the quality of contacts and the quality of the reconstructed models, is the optimal number of contacts we can evaluate. To test this, in the PSICOV data set, we calculated the Spearman's rank correlation coefficients between the evaluation measures (precision, coverage, X_d, spread, and mean false positive error) and the TM-score of the best CONFOLD reconstructed model, for various contact selections. The plot of correlation against top contact selections in **Figure 3.2**, shows that correlation for the three important measures precision, X_d, and mean false positive error, is high for at least top-L/5 contacts. In summary, we find that top-L/5 is the minimum number of long-range contacts to

evaluate.

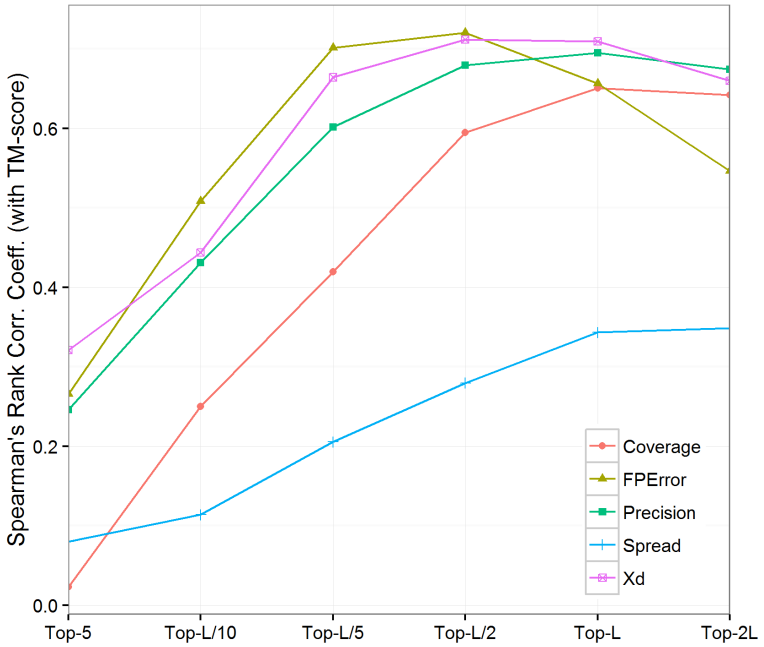

Figure 3.2: Spearman's rank correlation coefficient between the evaluation measures (coverage, mean false positive error, precision, spread, and X_d) and TM-score of the reconstructed models against various contact selections (top-5, top-L/10, etc.), for long-range contacts in the 150 proteins in PSICOV data set. The correlation values for mean false positive error and spread are negated to show all measures in the same quadrant.

3.4.3 Expected TM-score for values of evaluation measures

For a given protein, what values of precision, coverage, X_d, or mean false positive error of predicted contacts may fold the protein accurately (with TM-score > 0.5)? For the contacts predicted using PSICOV [21] for the 150 proteins in the PSICOV data set we classified top-L/5 long-range contacts into 3 bins for each measure. We binned predicted contacts into three precision bins – 0 to 40%, 40% to 60% and 60+ %, three X_d bins – 0 to 20, 20 to 28, and 28+, three mean false positive error bins – 0 to 1, 1 to 4, and 4+, and three coverage bins – 0-10, 10-15, 15+, and observed the distribution of TM-score values in each bins. The thresholds for these bins were selected by clustering the TM-scores into three clusters. We find that on average at least 40-60% precision is required to get a TM-score of 0.5 when folding using predicted contacts only; see **Figure 3.3**. We also find that to get similar TM-score, X_d should be more than 20, mean false positive error should be less than 4 and coverage should be more than 10. It is important to also note that coverage and X_d are also dependent upon the length of the protein unlike precision and mean false positive error.

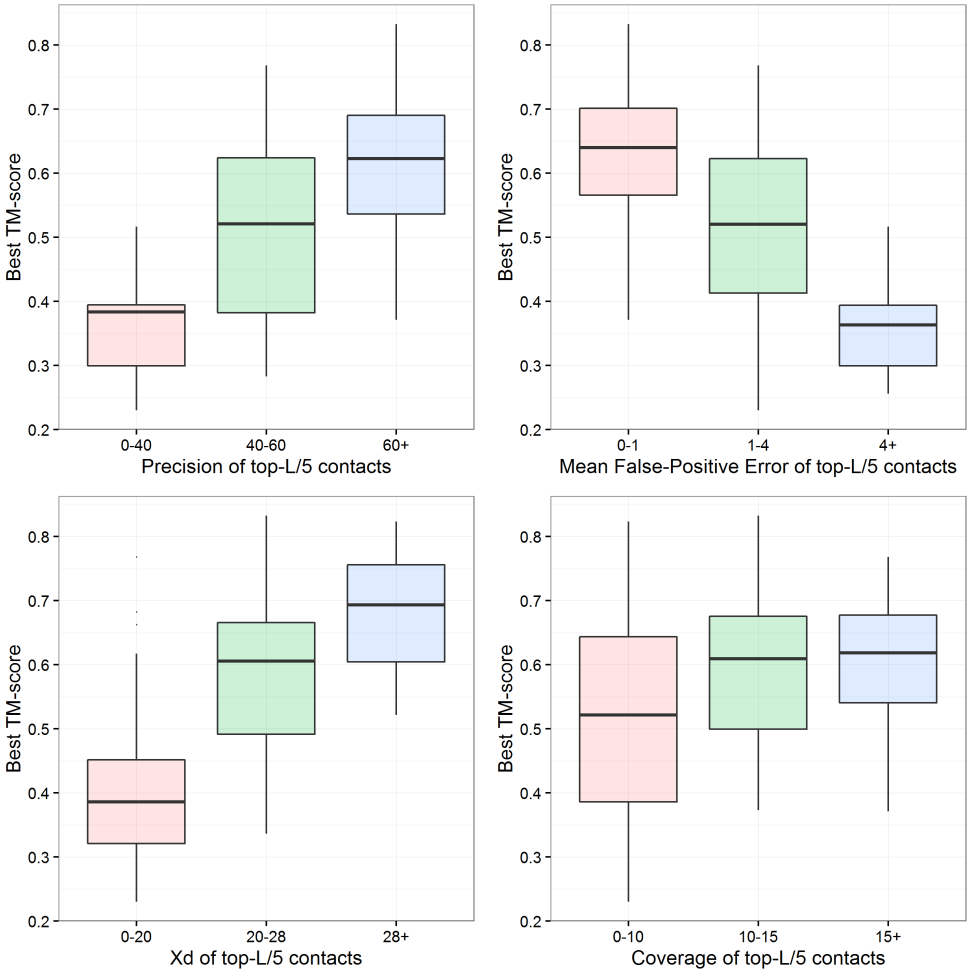

Figure 3.3: Expected TM-score of the best model reconstructed using CONFOLD against precision, mean false positive error, X_d, and coverage bins. Top-L/5 contacts predicted by PSICOV for the 150 proteins in the PSICOV data set were used as input for the calculations.

3.4.4 Protein types and evaluation measures

Using ConEVA we studied how the evaluation of predicted long-range contacts vary for the various protein folds (α, α+β, α/β, β) in the PSICOV data set. We find that mean false positive error has the highest correlation with the TM–score of the models for all protein folds, except for β proteins. For α proteins, mean false positive error and spread have the highest correlation with TM-score suggesting that α proteins are better evaluated using these two measures than others. For α+β and α/β proteins we observed that coverage has much lower correlation than other measures (X_d, precision, and mean-false-positive-error). All correlations are presented in **Table 3.2** and visualized in **Figure 3.4**. Similar statistics were observed when we selected "all" contacts instead of long-range.

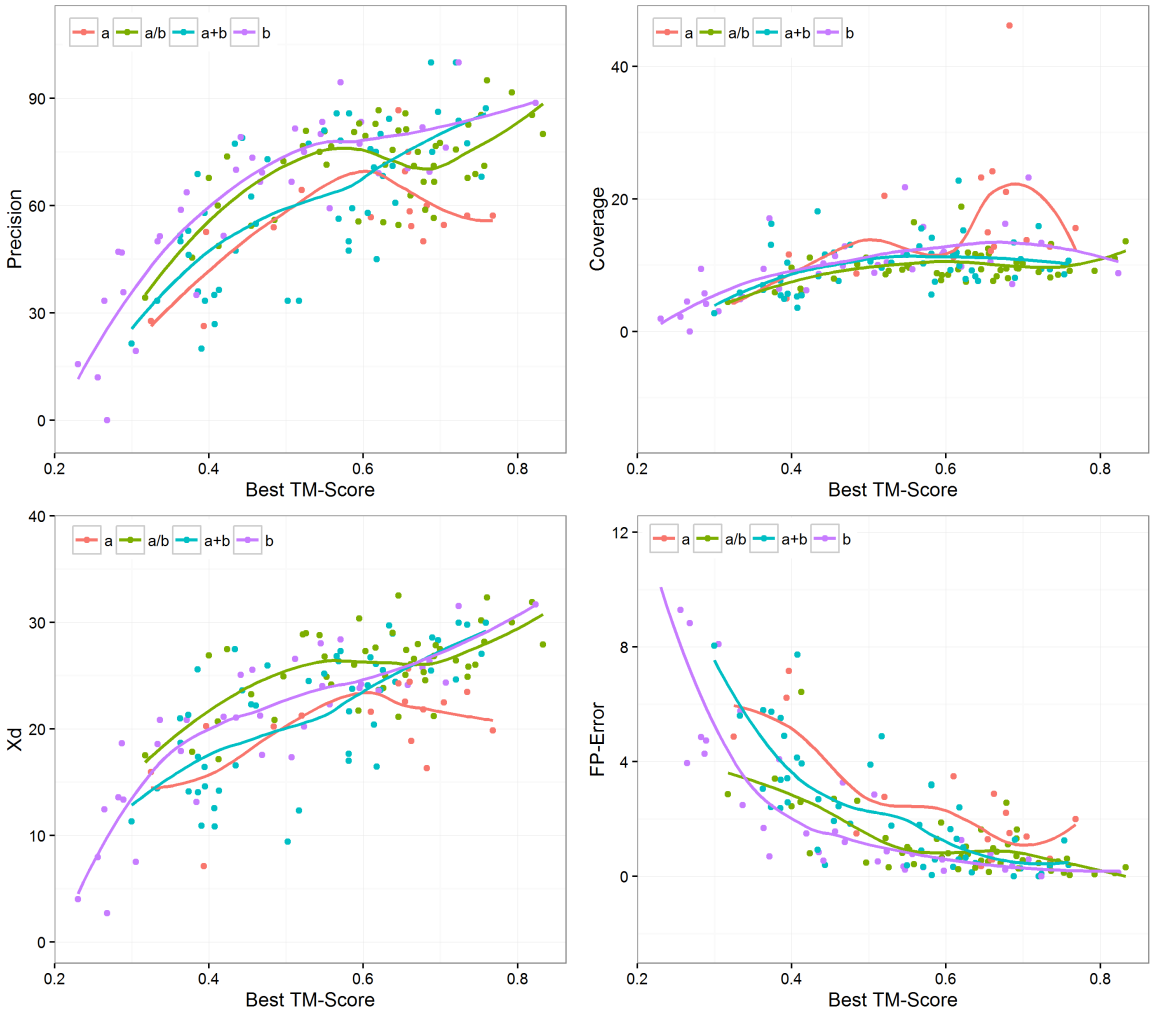

Figure 3.4: Relationship between precision, coverage, mean false positive error, and X_d with the best TM-score for various protein folds. It shows that β proteins are best evaluated using precision and X_d and coverage is relatively most important for α proteins. Evaluations are performed on top L/5 long-range contacts predicted by PSICOV and TM-score is that of the best model built using CONFOLD.

Table 3.2: Spearmans rank correlation coefficient calculations of L, Nc, and various evaluation measures with TM-score of the best CONFOLD built model for various protein fold types. Top-L/5 PSICOV predicted contacts are evaluated.

	α	$\alpha+\beta$	α/β	β
L	-0.34	0.21	0.05	-0.13
Nc	-0.47	0.27	0.08	-0.14
Precision	0.33	0.67	0.38	0.85
Coverage	0.6	0.33	0.28	0.7
X_d	0.31	0.69	0.44	0.84
Mean false positive error	-0.48	-0.78	-0.63	-0.86
Spread	-0.48	0.02	-0.3	-0.29

3.4.5 Similarity between predicted contacts

No methods currently exist for assessing the quality of predicted contacts in the absence of native structures. Since Jaccard similarity score provides a quantitative comparison of contact sets, we hypothesized that when there is larger agreement between multiple sets of predicted contacts, the confidence of the contact prediction for the protein is higher. Using the same PSICOV data set, we first computed the Jaccard similarity between the PSICOV predicted contacts and CCMpred predicted contacts, and then calculated the Spearman's rank correlation coefficient between this similarity and the precision of the predicted contacts (see **Figure 3.5**). High correlation coefficients of 0.63, 0.64, and 0.57 for N (neighborhood relaxation for computing Jaccard similarity) equal to 0, 1, and 2 respectively validates our hypothesis. These findings, although obvious (i.e., accurate contacts will be correlated), can have interesting applications. For instance, a very wide range of features are used for developing protein model quality assessment (QA) methods, including many contact related scores [66, 67, 68]. Jaccard similarity score is a potentially useful feature for developing QA methods. In addition, this similarity score can even be integrated into model building methods like FUSION [69], UniCon3D [70], and FRAGFOLD [28] to decide the weight of the contact energy term.

3.5 Discussion

ConEVA allows a user to choose from various contact types, distance thresholds, and sequence separation thresholds for defining contacts and enables study of how the various measures change over various numbers of top contacts. It accepts contacts in Critical Assessment of protein Structure Prediction (CASP) RR file format. We verified ConEVA evaluations by comparing against the CASP evaluations available at http://predictioncenter.org. A downloadable version is also available that calculates all the quantitative measures without any visualizations. Below we outline some of its

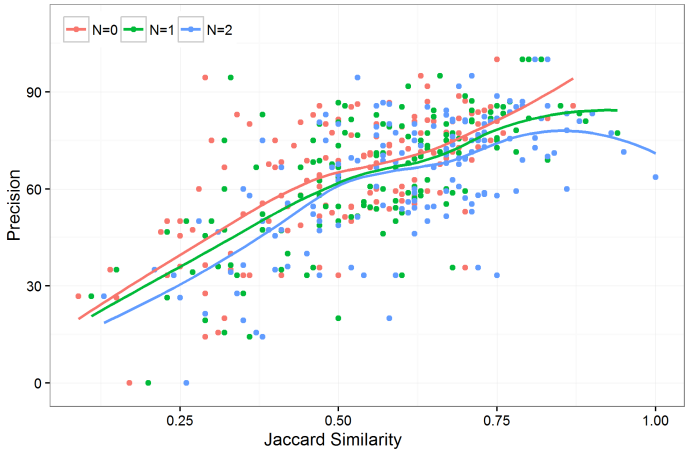

Figure 3.5: Precision of top-L/5 PSICOV predicted contacts versus the Jaccard similarity score between PSICOV contacts and CCMpred predicted contacts for the 150 proteins in PSICOV data set. N corresponds to the neighborhood size in computing Jaccard similarity.

features with the evaluation of predicted long-range Cβ contacts for the protein '1aa3' (chain A) in the PSICOV data set and protein domain T0763-D1 in the CASP11 data set as reference examples.

3.5.1 Contact evaluation

For predicted contacts, ConEVA evaluates the top five, L/10, L/5, L/2, L and top 2L contacts against a native structure using precision, coverage, X_d, mean false positive error, spread, MCC, AUC_PR, and ROC curves (see **Figure 3.6**). For analysis and comparison, it also produces neat plots of two dimensional contact maps. For convenient comparison, in the presence of a native structure, contact maps are displayed with the native structure's contact maps in the background (see **Figure 3.7**). For visualizing predicted contacts in the native structure, UCSF Chimera command scripts [71] are provided to download and run locally (see **Figure 3.8**).

3.5.2 Contact assessment in the absence of a native structure

When only predicted contacts (or multiple set of contacts) are submitted, two-dimensional contact maps and one dimensional coordination numbers are presented along with counts for short-, medium-, and long-range contacts and visualizations using contact maps, chord diagrams and Jaccard similarity matrixes along with dendrograms (see **Figure 3.9**). The visualization of coordination numbers serves as a detailed analysis of the residue location of predicted contacts (see **Figure 3.10**). When analyzed along with predicted three-state secondary structures (helix, strand, and coil), coordination numbers can show the contrast or agreement between predicted secondary structures and

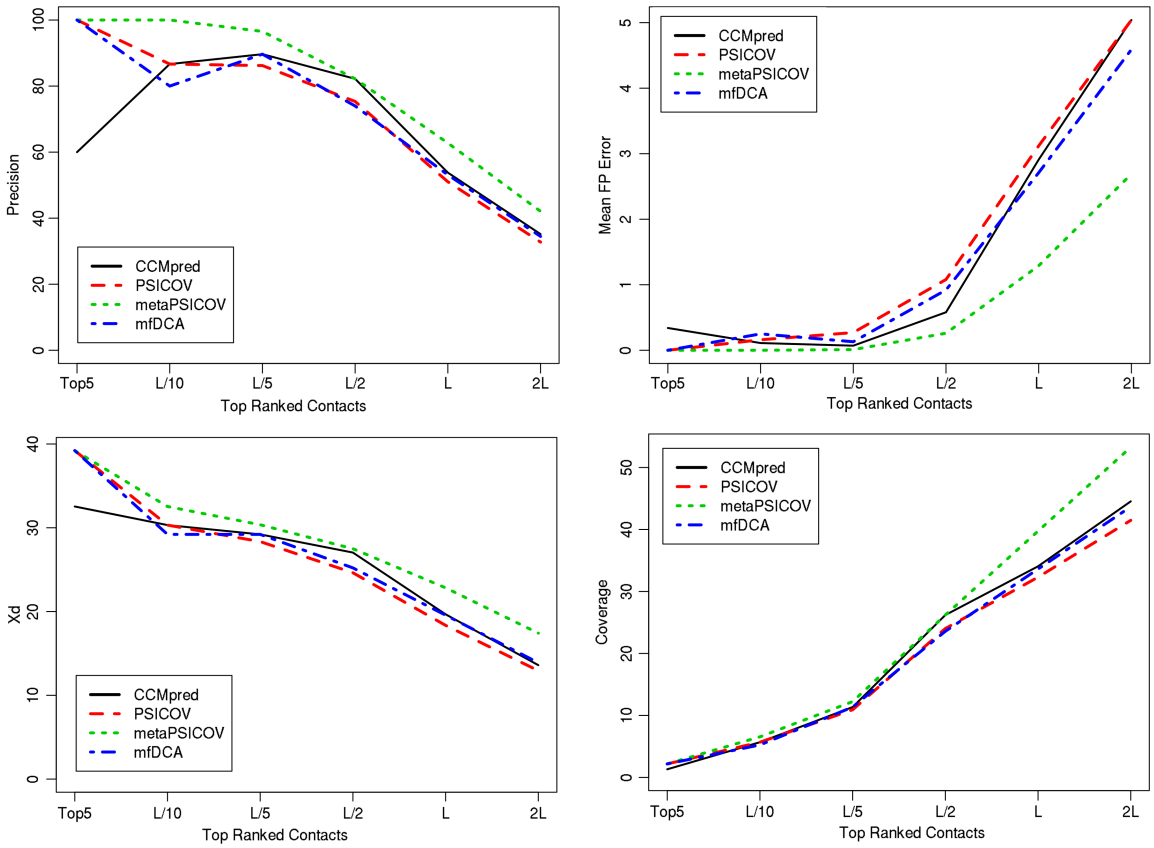

Figure 3.6: A screenshot of ConEVA evaluation of contacts predicted for the protein '1a3aA' showing calculations for precision (top left), mean false positive error (top right), X_d (bottom left), and coverage (bottom right). For this protein, MetaPSICOV has shown slightly better performance than CCMpred, PSICOV, and mfDCA in every evaluation measure.

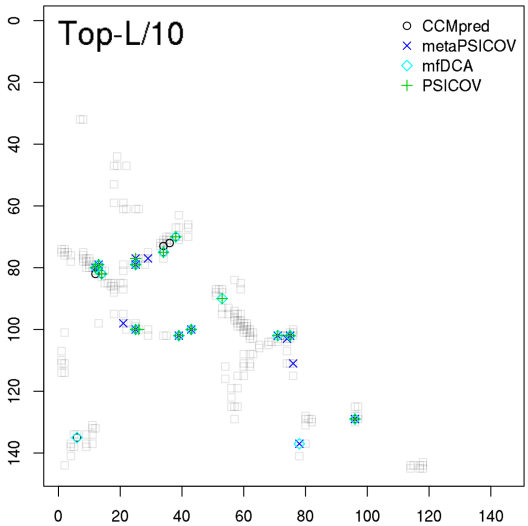

Figure 3.7: A screenshot of contact map showing long-range contacts for top-L/10 predicted contacts for the protein '1a3a' with the native contacts shown in gray in background.

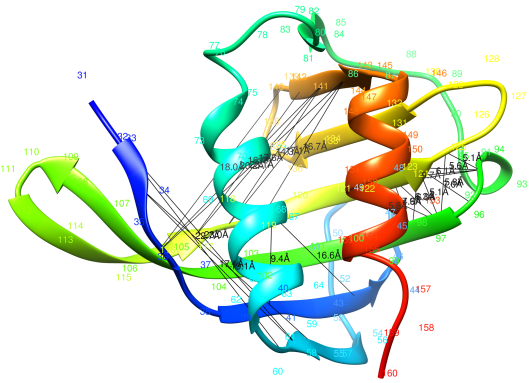

Figure 3.8: Top-L/5 CONSIP2 predicted long-range contacts (total 26 contacts) shown in the native structure domain of T0763-D1 as an example of visualizing the contacts in UCSF Chimera using ConEVA downloaded scripts. This visualization shows the clustering of the predicted CONSIP2 contacts in three regions and mostly between the beta strands, where one cluster (on the right) is correct and two other clusters are mostly wrong (with long black lines showing the distance between predicted contacts).

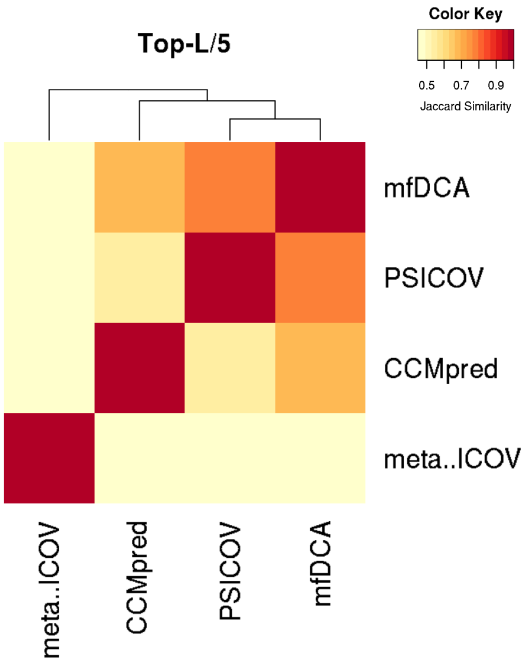

Figure 3.9: A screenshot of Jaccard similarity matrix visualization of contacts predicted for the protein 1a3a chain A. The Jaccard similarity matrix with N equals 0 (right) shows that contacts predicted by mfDCA and PSICOV are most similar and MetaPSICOV contacts are equally similar to all other predictions.

```
         10        20        30        40        50        60        70        80        90       100
         |         |         |         |         |         |         |         |         |         |
LFKLGAENIFLGRKAATKEEAIRFAGEQLVKGGYVEPEYVQAMLDREKLTPTYLGESIAVPHGTVEAKDRVLKTGVVFCQYPEGVRFGEEEDDIARLVIG [sequence]
-----1-----211----------2--------2-1-11---1-------------------------1-11-2---21-2------------1---2 [CCMpred: 15 pairs 22 residues]
----------111----------31-------1---11---1---------1----------------11---2-1-21-1-------1-----1---3 [PSICOV: 15 pairs 22 residues]
----------11-------1---3---1---------1---1---------------------------1--1112121---------------1-1-2 [metaPSICOV: 15 pairs 23 residues]
-----1-----111----------2--------1---11---1---------1----------------11---2--121-1-------1-----1---2 [mfDCA: 15 pairs 24 residues]
```

Figure 3.10: A screenshot of 1D visualization of coordination numbers within the first 100 residues of the protein '1aa3'. Each row represents the contacts predicted by a single method, with number of contacts and number of residues involved in the contacts shown at the end. From this visualization, three clusters of contacts can be observed as common between the four methods.

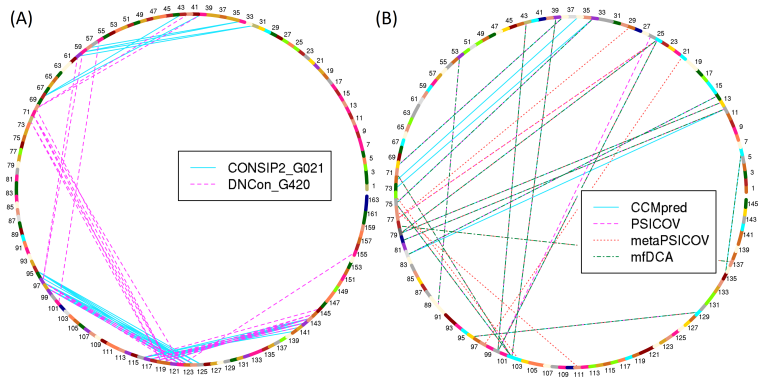

Figure 3.11: Chord diagrams for top-L/10 contacts for T0763-D1 **(A)** and for top-L/10 contacts predicted for '1aa3' **(B)**. The diagrams show that contacts predicted for T0763-D1 are clustered with no contacts predicted for the first 30 residues (which is in fact a disordered region with no native coordinates), whereas, predicted contacts have high overlaps between methods and are well spread for '1aa3'.

contacts. For instance, clusters of predicted contacts are expected in the strand regions. Similarly, Chord diagrams can be useful to observe contact clusters, similarities in predicted contacts and even to predict disordered regions (see **Figure 3.11**). Both, coordination numbers and Chord diagrams can also be useful to detect predicted contacts that have extremely low coverage, i.e. highly clustered contact predictions. Identifying such predictions and prediction methods can help us make decisions on using more contacts from the same source or resort to other methods of contact prediction. These results can be useful for predictive analysis of contacts to study how the contacts may be selected and/or combined for building models.

3.5.3 Analysis of a structure's contacts

A three-dimensional protein data bank (PDB) structure [57] file or a 'pdb id' may be provided as input to study its true contacts for a chosen definition of contacts. This feature is useful not only to study the reconstruction of a protein but also to understand the maximum and minimum values of measures like X_d for a structure, also allowing us to investigate what contact definitions yield a desired set of contacts for a structure of interest. This is sometimes important to investigate whether some protein structure has too few or no long range contacts at all.

3.6 Conclusion

Contacts are becoming increasingly useful not just for ab initio protein structure prediction but also for being integrated into experimental methods, and we are finding many more applications of

contacts with the increasing research on contacts. We hope that ConEVA will be useful not only to contact prediction developers but also to general public who need to predict structures for their sequences that do not have a good template.

Chapter 4

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

4.1 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 reconstruction

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/.

4.2 Introduction

Emerging success of residue-residue contact predictions [10, 11, 13, 14, 16, 18, 20, 21, 22, 24, 53, 72, 73, 74, 75, 76] and secondary structure predictions [40, 41, 77, 78, 79, 80, 81] 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 [3, 4, 5, 6, 8, 82, 83, 84]. However, all of these reconstruction methods, including most recent ones, Reconstruct [4] based on Tinker [85] and C2S [86] based on FT-COMAR [3], 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 [87] and Tinker [85] 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 secondary structures from these predicted information. The widely used modeling tool, Modeller [88], 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 [24, 26] used the Crystallography & NMR System (CNS) [89, 90], 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 paper, 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 [85] 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 re-generate 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.

4.3 Materials and Methods

4.3.1 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 [24], and (2) 150 diverse globular proteins with average length of 145 residues used in FRAGFOLD [34]. 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 reside-reside 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 [24]. 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 [34]. 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 xL (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-0.4L contacts would refer to the top 40 (0.4 times 100) contacts.

4.3.2 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 [91] 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, anti-parallel 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 [92, 93] is automatically preserved.

Based on 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. Based on these restraints, in **Table 4.1** 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 (o) 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 4.1** 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.

4.3.3 Two-stage model building and contact filtering

Figure 4.1 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 2.2, and contacts into the range [3.5 Å, threshold]. One key issue is to decide how many contacts should be used to build models. In order to estimate the number of contacts needed for reconstruction, we scanned the structures in the Protein Data Bank (PDB) [57] and found that more than 99% of known 3D structures have less than 3L true contacts, and more than 50% of them have less than 2L (L: length of a protein) true contacts. And based on our test on 15 proteins in EVFOLD benchmark set, less than 1.6L predicted contacts yielded best results. Therefore, for each protein, we built 20 models for each contact sets consisting of top 0.4L, 0.6L, 0.8L, ... up to 2.2L contacts. The models were constructed from these restraints by a customized distance geometry algorithm implemented in CNS (see Section 2.5). 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 (**Figure 4.1**), we updated the contact information as well as the β-sheet information by analyzing the model having minimum restraints energy in the

Table 4.1: 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. In all sub-tables, the first column defines secondary structure type: parallel (P) or anti-parallel (A), generic strand (U), and helix (H). Measurements of upper and lower bounds of hydrogen bond distances for anti-parallel 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 [7.4Å, 8.0Å] between oxygen of residue b and oxygen of residue (c+1).

Table (A)

Type	LB	UB
A	1.8	2
P	1.8	2
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
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
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
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
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
A	C-C	O	1	6.2	6.6
A	C-C	O	-1	5.6	5.8
A	C-C	O	0	5.2	5.4
A	C-C	H	1	6.2	6.6
A	C-C	H	-1	5.5	5.9
A	C-C	H	0	5.2	5.4
P	O-O	O	1	7.6	8.2
P	O-O	O	-1	4.8	5
P	O-O	O	0	3.6	4
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
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	4.7	5.1
P	C-C	H	-1	7.7	8.1
P	C-C	H	0	5.1	5.3
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	5.2
P	C-C	O	1	8.4	8.8
P	C-C	O	-1	4.8	5
P	C-C	O	0	6.1	6.3
P	C-C	H	1	4.8	5
P	C-C	H	-1	7.2	7.8
P	C-C	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	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	6.2
H	C-C	O	1	8.1	8.3
H	C-C	O	-1	4.8	5
H	C-C	O	0	6	6.2
H	C-C	H	1	4.8	5
H	C-C	H	-1	8.1	8.3
H	C-C	H	0	6	6.2
H	N-N	O	1	8	8.2
H	N-N	O	-1	4.7	4.9
H	N-N	O	0	6	6.2
H	N-N	H	1	4.7	4.9
H	N-N	H	-1	8	8.2
H	N-N	H	0	6	6.2
H	C-C	O	1	8.5	8.7
H	C-C	O	-1	5	5.2
H	C-C	O	0	6.1	6.3
H	C-C	H	1	5	5.2
H	C-C	H	-1	8.5	8.7
H	C-C	H	0	6.1	6.3

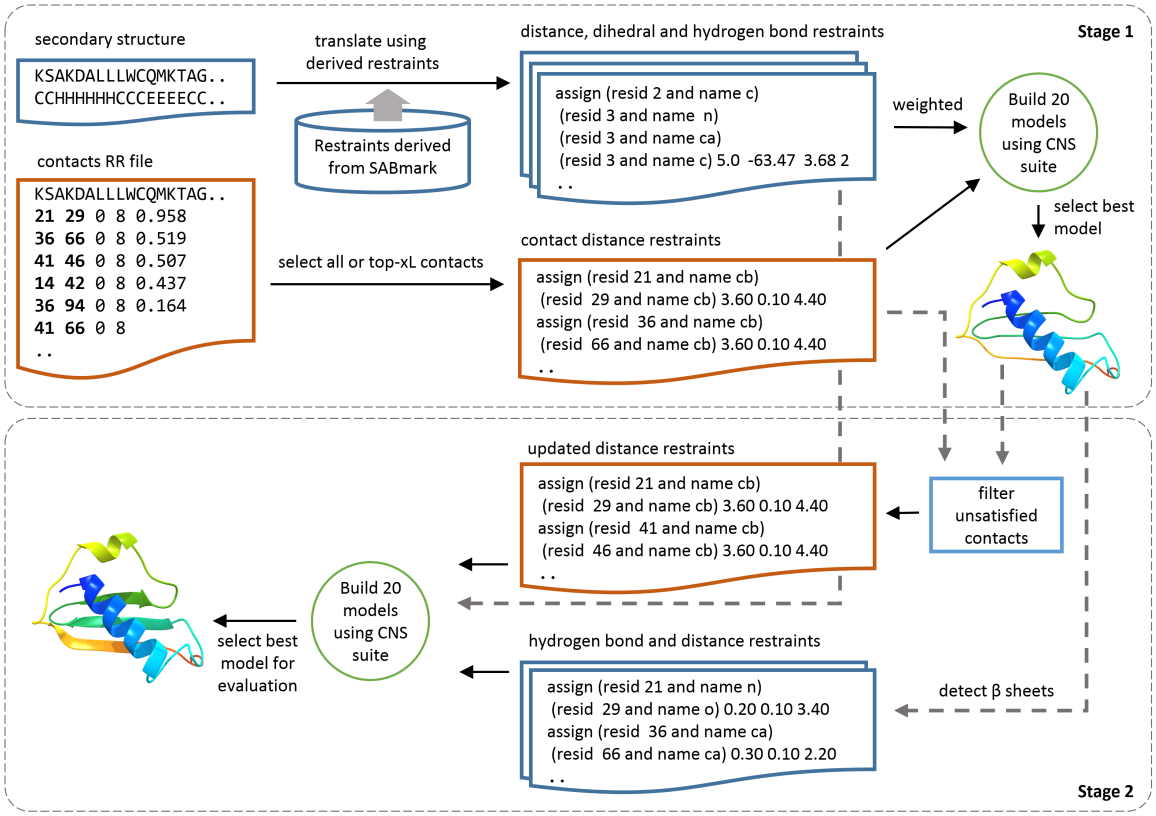

Figure 4.1: The CONFOLD method for building models with contacts and secondary structures in two stages. When true contacts are the input, all contacts are used to reconstruct models. For predicted contacts, top-xL contacts are used, where x ranges from 0.4 to 2.2 at a step of 0.2.

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 (see Section 2.4 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 (i.e., 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 [24, 34] 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.

4.3.4 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 anti-parallel ideal hydrogen-bonding patterns and ten parallel hydrogen-bonding patterns (see **Figure 4.2**). 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 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

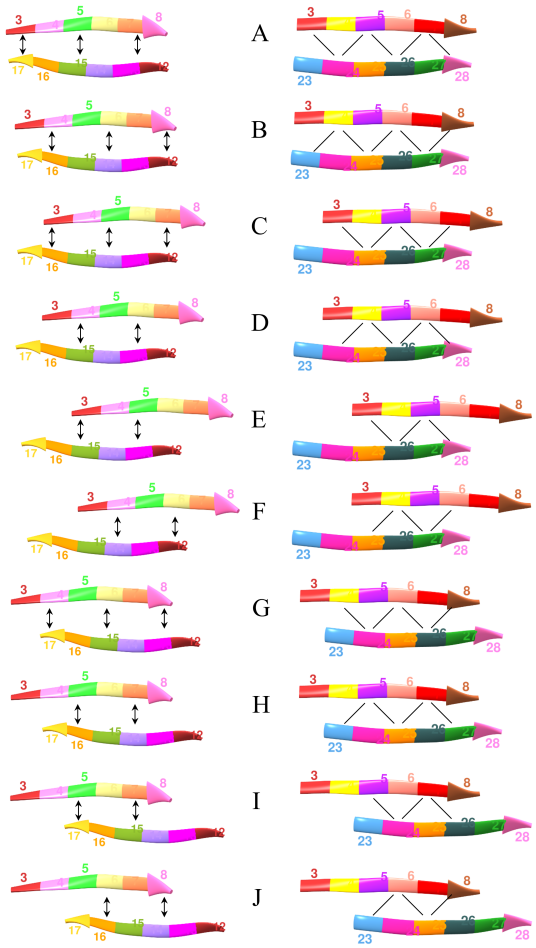

Figure 4.2: Ten alternate hydrogen-bonding patterns for antiparallel (left) and parallel (right) pairing for a pair of strands, each six residues long. First strand is from residues 3 to 8, and second strand is from residues 12 to 17 for antiparallel pairs and 23 to 28 for parallel pairs. The ideal hydrogen 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.

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 7Å.

4.3.5 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 [89, 90] 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, '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, i.e., 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.

4.4 Results and Discussion

4.4.1 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 (see **Figure 4.3** for examples). 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 [94]. 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 [95].

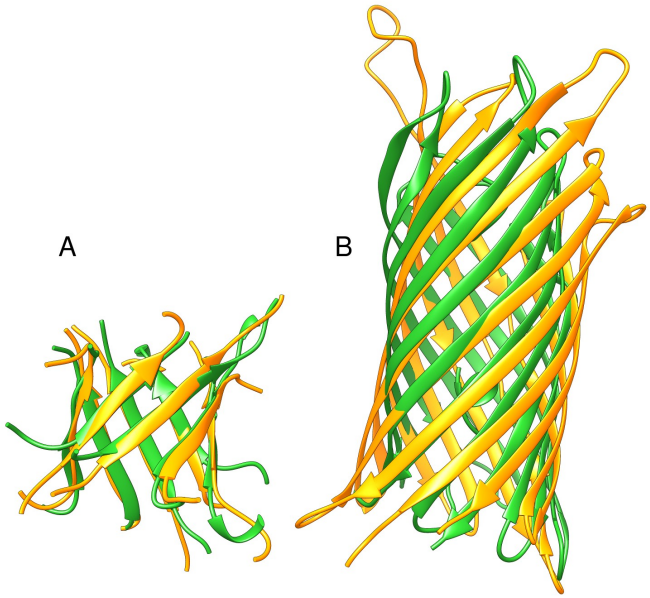

Figure 4.3: Top models reconstructed for the proteins 2QOM and 1YPI using true secondary structure information along with beta-pairing information but without using any residue contact information. Secondary structure restraints are computed using $\lambda = 0.5$. Superposition of crystal structure (green) and reconstructed top model (orange) of the beta-alpha-beta barrel protein 1YPI (A) and antiparallel beta barrel protein 2QOM (B).

We also investigate how the scaling factor (λ) controlling upper bound and lower bound of all secondary structure restraints (hydrogen bond, distance, and 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 4.2** 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 [57, 96]. 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.

Table 4.2: 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. Percentage of helix and β-sheet residues reconstructed are listed against various values of λ.

λ	**% reconstructed**	
	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

4.4.2 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 [4] and Modeller [88]. From native tertiary structures of these proteins, we compute three-class secondary structure information using DSSP [97] and true Cβ-Cβ contacts at 8Å 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 [98] and Root Mean Square Deviation (RMSD)) for all these proteins are reported in **Table 4.3**. The average TM-score [98] of the best models constructed by CONFOLD with secondary structure restraints, CONFOLD without secondary structure restraints, Reconstruct and Modeller 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 Modeller with the same input. All the methods perform better on single-domain proteins than on multi-domain proteins (e.g. 2O72 and 1G2E). **Figure 4.4** 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 Modeller has poor global topology but better secondary structures.

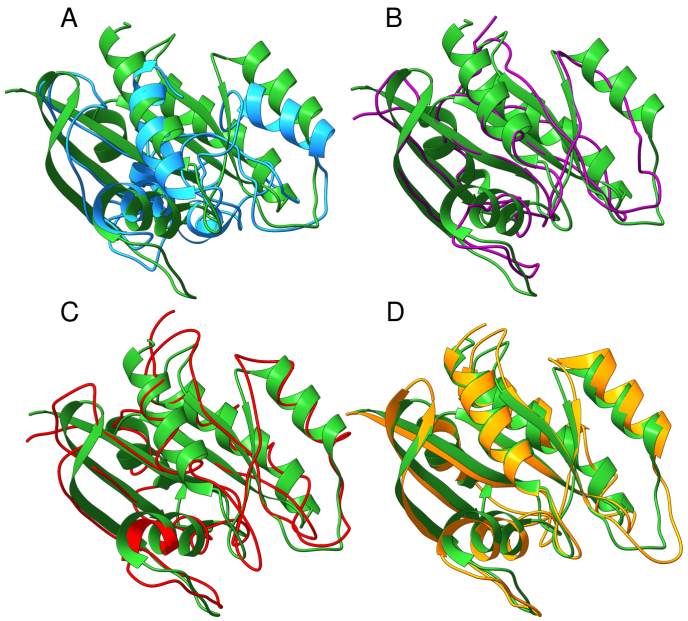

Figure 4.4: Best models reconstructed for the protein 5p21 using Modeller (A), Reconstruct (B), customized CNS DGSA protocol (C), and CONFOLD (D). All models are superimposed with native structure (green). The TM-scores of Models A, B, C, and D are 0.53, 0.86, 0.88, and 0.94, respectively. Model D reconstructed by CONFOLD has higher TM-score and also much better secondary structure quality than the other models.

Comparing the best models built using only contact restraints and those using both contact restraints and secondary structure restraints in **Table 4.3**, we find that adding secondary structure restraints improves the 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 Modeller on 150 proteins in FRAGFOLD benchmark set using the same protocol. **Figure 4.5** shows the distribution of TM-Scores [98] 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 Modeller, respectively. The results show that when secondary structure restraints are considered, CONFOLD can reconstruct models from true

Table 4.3: 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 Modeller. 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 1hzx using Reconstruct is not presented because Tinker failed to run because of memory requirement issues.

PDB Code	Native				CNS DGSA				CONFOLD				Modeller				Reconstruct			
	Nc	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*	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*	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	35	1	0.75	3.3	0	4

* multi-domain proteins

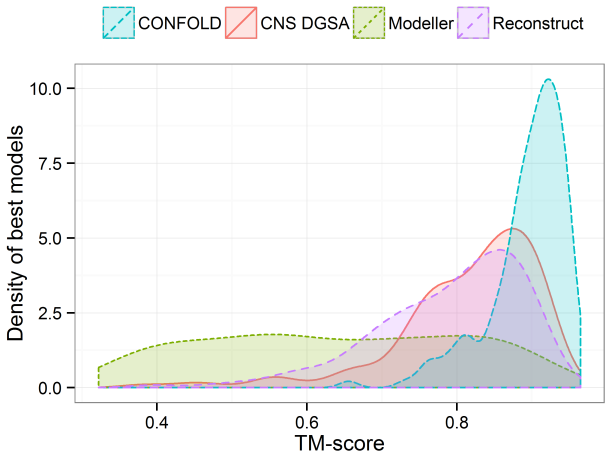

Figure 4.5: Distribution of TM-scores of the best models reconstructed by the four methods for 150 FRAGFOLD proteins.

contacts with substantially better quality than Modeller, 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 modelling, 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.

4.4.3 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 4.4** presents the comparison of model accuracy and secondary structure quality for all 15 proteins. As an example, **Figure 4.6** 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 4.7** shows that CONFOLD models are better in general. On average, the TM-score of all CONFOLD models is 0.42, 20% higher than

Table 4.4: Comparison of accuracy and secondary structure quality of best models built by CONFOLD and EVFOLD. Columns H and E are number of helix and β-sheet residues assigned by DSSP. RMSD values are in Å.

UNIPROT-NAME	Native			EVFOLD				CONFOLD			
	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	2.5	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	5.2	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	3.1	20	0
THIO_ALIAC	103	30	25	0.59	3.99	25	8	0.64	3.7	31	16
CADH1_HUMAN	100	0	42	0.58	4.18	0	18	0.61	4.2	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**

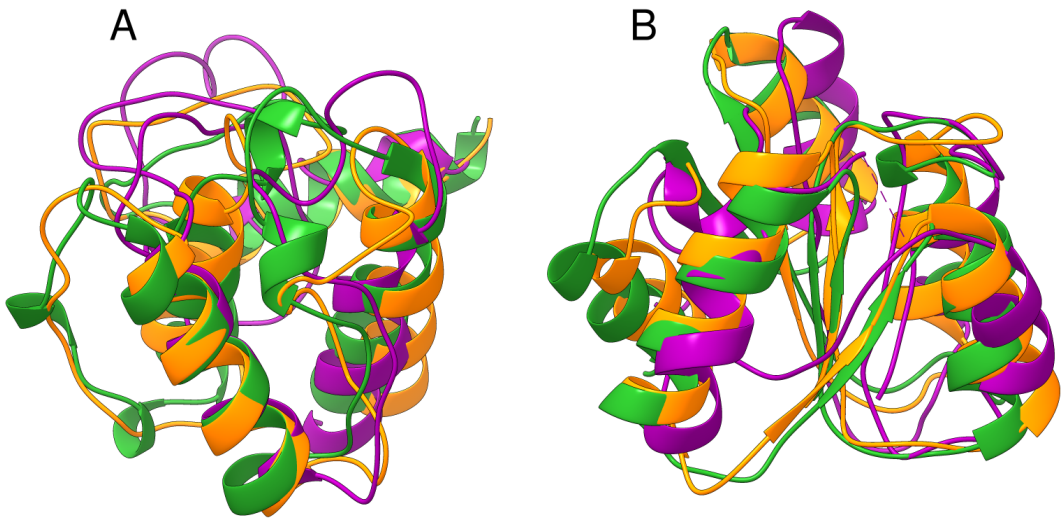

Figure 4.6: Best predicted models for the proteins RNH_ECOLI (A) and SPTB2_HUMAN (B) using EVFOLD (purple) and CONFOLD (orange) superimposed with native structures (green). The TM-scores of these models are reported in Table 4.4. CONFOLD models have higher TM-score and better secondary structure quality than EVAFOLD.

EVFOLD model pool.

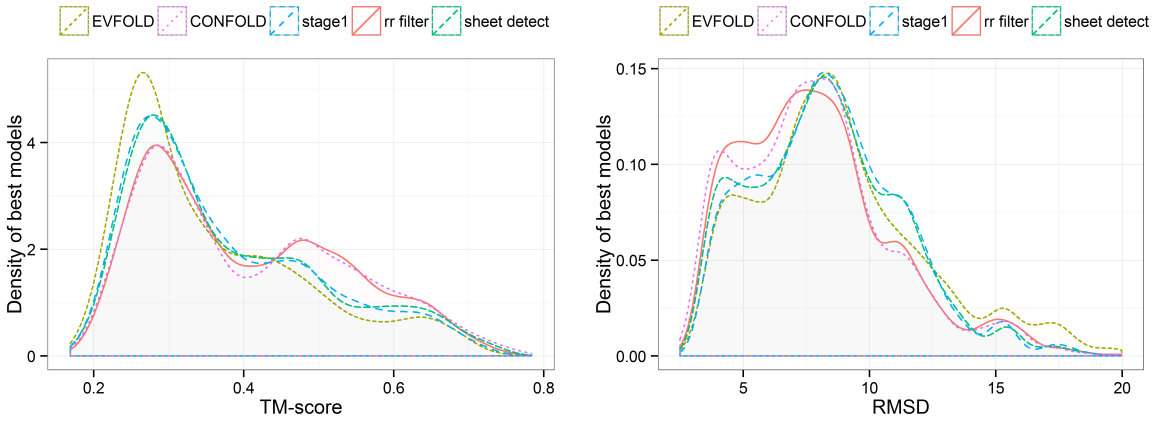

Figure 4.7: Distribution of model quality of the EVFOLD models and the models built by CONFOLD. Distribution of models built in first stage of CONFOLD (stage1), second stage with contact filtering only (rr filter), and second stage with β-sheet detection only (sheet detect) are also presented. Each curve represents the distribution of 400 times 15 models. Since some models in the EVFOLD model pool have RMSD greater than 20 Å, all models with RMSD greater than 20 Å from all four model pools were filtered out.

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 stage 1 and stage 2 suggests a significant improvement in the accuracy and secondary structure quality of models from stage 1 to stage 2. To analyze the improvement due to β-sheet detection and contact filtering in stage 2, in **Table 4.5**, 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 (i.e. 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 0.9Å RMSD improvement is observed in CONFOLD second stage, and the number of strands in the second stage is more than 3 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 0.5Å RMSD on average. **Figure 22** 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 [34], 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 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.

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 stage 1 and stage 2 are shown in **Figure 4.8**.

Table 4.5: 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. Columns H and E are the number of helix and β-sheet residues computed by DSSP.

UNIPROT-NAME	**stage1**			**sheet detect**			**contact filter**			**stage 2**		
	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**

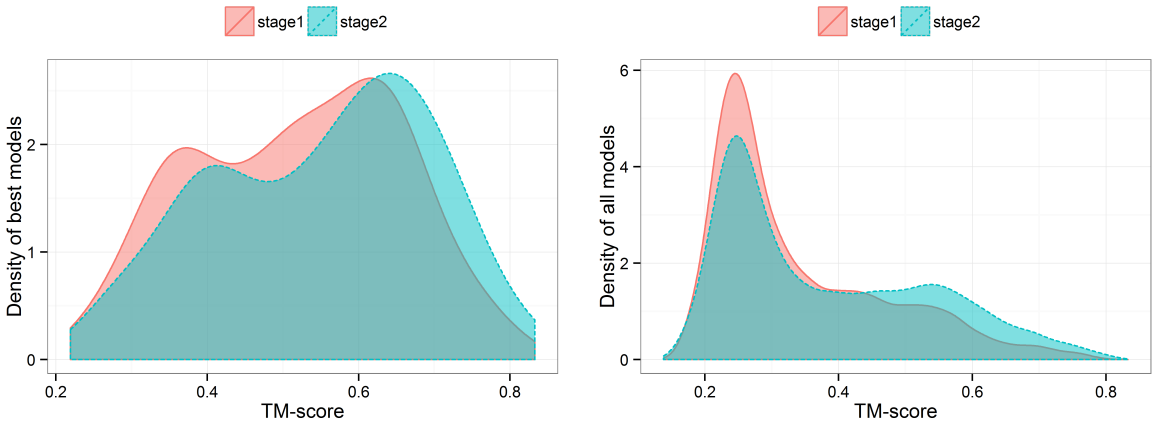

Figure 4.8: Improvement in the accuracy of best models (left) and all 400 models (right) in the second stage of CONFOLD over the first stage for 150 proteins in FRAGFOLD dataset.

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 towards 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 stage 1 and stage 2. 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 4.9** illustrates the models for protein 1NRV (L = 100) reconstructed with top-0.6L contacts in stage 1 and stage 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.

4.4.4 Analysis of number of predicted contacts needed to obtain best fold

Although 99.9% of the proteins in PDB have less than 3L contacts, much fewer true contacts are sufficient to fold the proteins accurately [4, 8]. 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 0.6L, 0.8L, 1.0L, or 1.2L contacts in both stages of CONFOLD (**Figure 4.10**). 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.

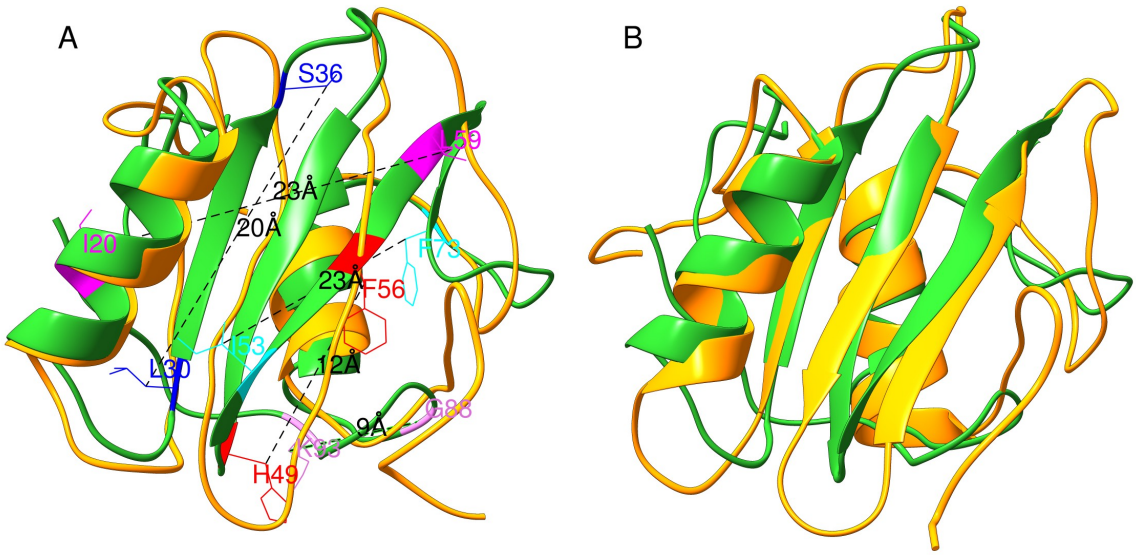

Figure 4.9: Contact filtering from stage 1 to stage 2 for the protein 1NRV. (A) Superimposition of the best model in stage 1 reconstructed with top-0.6L contacts by CONFOLD (orange) with the native structure (green). The model has TM-score of 0.50. Among the top-0.6L (60) contacts, 5 out of 8 erroneous 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-0.6L contacts by CONFOLD (orange) with the native structure (green). TM-score of the model is 0.61.

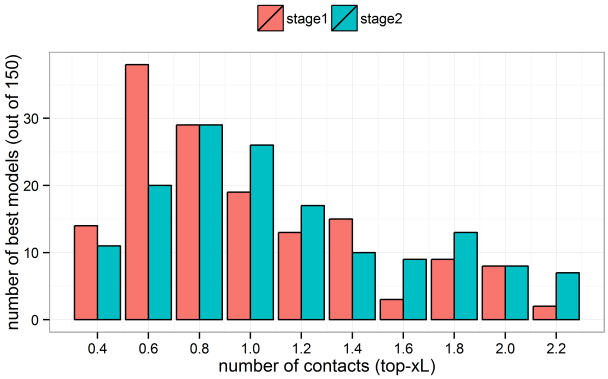

Figure 4.10: Number of best models and the number of contacts used to build the best models for 150 proteins in FRAGFOLD dataset.

4.4.5 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, i.e., 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 homologous 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 homologous 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-0.5L 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 [99] is suggested as one of the best clustering-based methods [26] 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.

4.5 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.

Chapter 5

Improved protein structure reconstruction using secondary structures, contacts at higher distance thresholds, and non-contacts

5.1 Abstract

Residue-residue contacts are key features for accurate ab initio protein structure prediction. For the optimal utilization of these predicted contacts in folding proteins accurately, it is important to study the challenges of reconstructing protein structures using true contacts. Because contact-guided protein modeling approach is valuable for predicting the folds of proteins that do not have structural templates, it is necessary for reconstruction studies to focus on hard-to-predict protein structures. Using a data set consisting of 496 structural domains released in recent CASP experiments and a dataset of 150 representative protein structures, in this work, we discuss three techniques to improve the reconstruction accuracy using true contacts – adding secondary structures, increasing contact distance thresholds, and adding non-contacts. We find that reconstruction using secondary structures and contacts can deliver accuracy higher than using full contact maps. Similarly, we demonstrate that non-contacts can improve reconstruction accuracy not only when the used non-contacts are true but also when they are predicted. On the dataset consisting of 150 proteins, we find

that by simply using low ranked predicted contacts as non-contacts and adding them as additional restraints, can increase the reconstruction accuracy by 5% when the reconstructed models are evaluated using TM-score. Our findings suggest that secondary structures are invaluable companions of contacts for accurate reconstruction. Confirming some earlier findings, we also find that larger distance thresholds are useful for folding many protein structures which cannot be folded using the standard definition of contacts. Our findings also suggest that for more accurate reconstruction using predicted contacts it is useful to predict contacts at higher distance thresholds (beyond 8 Å) and predict non-contacts.

5.2 Background

A major motivation for protein contact prediction and contact-guided protein structure prediction comes from the general finding that accurate contacts lead to accurate tertiary structural models. Studies like FT-COMAR [3] and Reconstruct [4] on protein structure reconstruction using true contacts have shown that in general three-dimensional protein structures can be recovered using two-dimensional contact maps. For instance, using true Cαup contact maps derived with a distance threshold of 9Å, a study reconstructed 19 proteins with accuracy of 1Å RMSD [5]. Similarly, deriving true contacts at distance cut-offs higher than 9Å, Vassura et al. reconstructed Cαup models for 1,760 proteins of different fold classes with RMSD of around 2Å using the FT-COMAR method [3, 6]. In another study, authors have shown that the quality of 3D reconstruction is unaffected by deleting up to an average 75% of the real contacts [7]. Likewise, in a different study, it is demonstrated that the number of contacts needed for reconstruction can be decreased using a cone-peeling method and a reconstruction accuracy of $\leq$ 4Å can be achieved with just around 20 to 30% of true contacts on a data set of 12 proteins [8]. Most recently, it is also shown that a distance cut-off of 9Å to 11Å delivers accurate reconstructions using Cβup atoms for defining contacts on a data set of 60 proteins [4].

These studies on reconstruction present many invaluable insights for utilizing contacts to fold proteins. However, in the context of reconstruction studies being useful for ab initio protein structure prediction, they have some limitations. Firstly, these studies use complete contact maps to reconstruct protein structures, whereas, recent practice for most model building methods has been to use much lesser predicted contacts. Consequently, these reconstruction studies also do not comply with the widely-used contact definition, i.e., the Critical Assessment of Protein Structure Prediction's (CASP) definition of contacts where 8 Å distance threshold is used with minimum sequence

separation of 6 residues. Secondly, these studies cover the issues related to the reconstruction of all types of proteins, and do not focus on the proteins that demand ab initio protein structure modeling. Since contact-guided protein modeling approaches are mostly useful when significant homologous templates are not found, it is important for reconstruction studies to focus on the proteins for which structural templates are hard to find. Lastly, none of these studies consider secondary structure information during reconstruction. Since secondary structure prediction has reached an accuracy higher than 80% [41, 63], it is meaningful to study how the knowledge of secondary structures can influence the quality of reconstructed models.

In this study, we investigate how accurately we can reconstruct 'hard' proteins (like the proteins categorized as 'free-modeling' in the CASP competitions) using true contacts and discuss various techniques to fold the ones whose structures cannot be accurately built in conventional ways. These techniques include, adjusting contact definitions, adding non-contacts into reconstruction, and incorporating secondary structure. Using our fragment-free ab initio reconstruction method CONFOLD [25] to carry out the experiments, we show that these techniques are useful to improve contact-based protein structure reconstruction.

Table 5.1: Comparison of the best of 20 models reconstructed using CONFOLD with the best of 20 models reconstructed using Reconstruct on the 12 benchmark proteins. Models are evaluated using TM-score, RMSD (in Å), and GDT-TS scores. Proteins are identified by their PDB ID followed by the chain ID. L is the length of the protein chain.

PDB code - chain ID	**SCOP class**	**L**	**Reconstruct**			**CONFOLD**		
			TM-score	**RMSD**	**GDT-TS**	**TM-score**	**RMSD**	**GDT-TS**
1bkr-A	all-α	109	0.88	1.54	81.02	0.89	1.61	85.42
1odd-A	all-α	118	0.85	1.62	78.75	0.87	1.56	83.75
1cem-A	all-α	363	0.81	2.2	63.91	0.96	1.53	80.79
1pzc-A	all-β	123	0.91	1.38	85.04	0.91	1.28	84.84
1onl-A	all-β	128	0.91	1.42	83.86	0.91	1.39	84.65
1eur-A	all-β	365	0.83	2.04	68.98	0.96	1.42	83.38
1e6k-A	α/β	130	0.89	1.75	82.5	0.91	1.42	82.69
1o8w-A	α/β	146	0.9	1.65	79.72	0.91	1.5	82.52
1ede-A	α/β	310	0.95	1.61	82.26	0.96	1.4	82.58
1r9h-A	$\alpha+\beta$	135	0.85	1.83	78.6	0.87	1.75	81.14
1ugm-A	$\alpha+\beta$	125	0.85	1.88	77.21	0.87	1.71	80.53
1iu4-A	$\alpha+\beta$	331	0.83	4.19	63.29	0.93	1.93	77.04
Average		**199**	**0.87**	**1.93**	**77.1**	**0.91**	**1.54**	**82.44**

5.3 Results

As the first step of testing our reconstruction pipeline, we reconstructed the 12 protein structures used by Duarte et al. [4] as benchmark dataset and compared our results with their tool Reconstruct.

For the comparison, we ran the Reconstruct tool locally to generate 20 models for each protein and the CONFOLD method to generate 20 models. Then, we considered best of the 20 models, by each method, for evaluation. **Table 5.1** shows that our method reconstructs more accurate models (20% improvement in RMSD) than Reconstruct when we compare the best models reconstructed by the two methods. Evaluation and comparison using other standard metrics like TM-score and GDT-TS score [98] also confirms that CONFOLD reconstructs better models. In summary, we observe that our method can reconstruct full atom tertiary structures of various folds with accuracy at least as good as the state-of-the art method Reconstruct.

Table 5.2: Reconstruction accuracy of 496 free-modeling (FM), template-based modeling (TBM), and hard template-based modeling (TBM-HA) domains in CASP 8, 9, 10 and 11 as measured by TM-score and RMSD. Three domains in CASP11, which are not classified into any of the three groups are categorized in the 'Other' group.

Group	**Domain Count**	**TM-score**	**RMSD**
FM	72	0.69	4.57
TBM-HA	71	0.78	3.24
TBM	350	0.8	2.88
Other	3	0.87	2.33
All	**496**	**0.78**	**3.18**

5.3.1 Reconstruction of CASP 8, 9, 10 and 11 domains using contacts

We reconstructed the structures for a total of 496 structural domains of the proteins released as regular targets in CASP 8, 9, 10 and 11 experiments using CONFOLD method with the true contacts derived from their native structures. The accuracy of reconstructing these structural domains, summarized in **Table 5.2**, shows that the mean TM-score [98] and RMSD of the reconstructed models is 0.78 and 3.2 Å. Our mean RMSD (3.2 Å) appears much higher than the expected mean RMSD of 2 Å as suggested in [6] because we did not consider local contacts (residue pairs closer than 6 residues in sequence) in order to comply with the currently widely accepted CASP's definition of contacts. CASP defines that residues must be separated by at least 6 residues to be in contact. In other words, we used all short-, medium-, and long-range contacts but not the complete contact map. To validate our assumption that the decrease in accuracy is because of the exclusion of the local contacts, we repeated our reconstruction experiments by including the contacts with sequence separation less than 6 residues and obtained mean TM-score and RMSD of 0.86 and 2.2 Å respectively. In addition, for each of the 496 domains, we also reconstructed 20 models using another reconstruction method FT-COMAR [3]. FT-COMAR's average reconstruction accuracy for these

domains is 4.9 Å when measured using RMSD and 0.68 when measured using TM-score, when best of 20 models are evaluated, much lower than the accuracy of CONFOLD's models. These results confirm existing findings that in general, local contacts are useful for reconstructing high-resolution models.

From our reconstruction using the standard CASP's definition of contacts, we find that the mean reconstruction accuracy for free-modeling (FM) targets is much lower than their template-based modeling (TBM) counterparts (see **Table 5.2** and **Figure 5.1**), indicating that the structures of hard targets are more difficult to reconstruct than easy targets. We also find that 28 out of the 496 domains were reconstructed with less than 0.5 TM-score, i.e. incorrect topology. In **Table 5.3** we list these 'hard-to-reconstruct' domains. To ensure that the low TM-score for these domains is not due to the method's ability to satisfy contacts, we calculated the sum of deviation (error) for all input contacts for each of the best model and found that in all cases this deviation is either zero or close to zero. This shows that the contacts restraints have been satisfied well and the low accuracy is due to the insufficiency of the input information. Almost all of these proteins are primarily helical, having 51% helix residues for the 13 FM domains and 65% for the 15 TBM domains, on average. This suggests that contact information alone (including all short-, medium-, and long-range contacts) cannot accurately guide the assembly of helices in many protein structures, and that knowing secondary structure (particularly helices) may improve the reconstruction accuracy. In the next section, we discuss the reconstruction results when secondary structures are included.

5.3.2 Reconstruction using contacts and secondary structures

In addition to reconstruction using contacts only, we reran our experiments by adding true 3-state secondary structures restraints (coil, helix and strand). On the same data set of 496 CASP structural domains, we obtained a mean TM-score of 0.88 and RMSD of 2.0 Å. This accuracy is slightly higher than the accuracy (TM-score = 0.86 and RMSD = 2.2 Å) when using complete contact maps (i.e., including contact pairs closer than 6 residues). The slightly higher TM-score and lower RMSD due to the use of secondary structure information suggests that aiding contacts with secondary structures is more useful than including the local contacts without secondary structure information. The improvement from using secondary structures and true contacts is significant according to paired t-test of TM-scores between the models reconstructed with contacts and secondary structures and the models reconstructed using the whole contact map without secondary structures (p-value = 2.2 x 10^{-16}). We also observed that out of the 28 protein domains that had less than 0.5 TM-score when

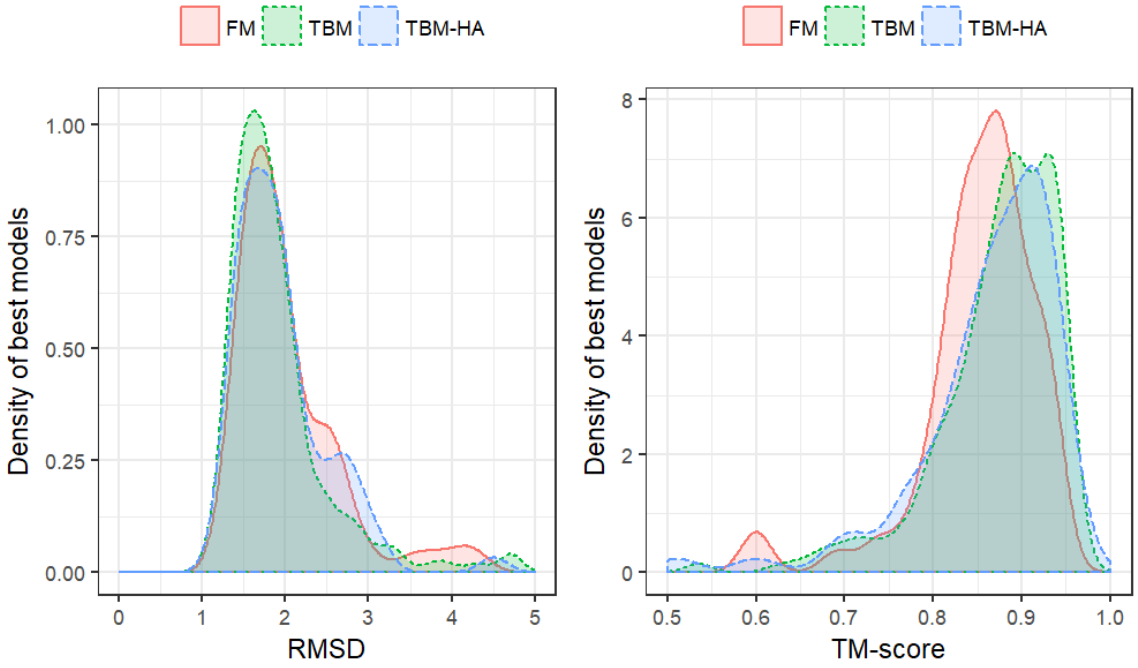

Figure 5.1: Distribution of the RMSD (left) and TM-score (right) of the best reconstructed models for the free-modeling (FM), template-based modeling hard (TBM-HA), and template-based modeling (TBM) domains in CASP 8, 9, 10, 11.

reconstructed with contacts only, 24 of them have TM-score higher than 0.5 after adding secondary structures. The remaining 4 domains (out of 28) listed in **Table 5.4** could not be reconstructed accurately (with TM-score > 0.5) using true contacts despite being supplemented by true secondary structures. Among these domains, T0629-D2 is a domain in a long tail needle-shaped receptor-binding tip protein 2XGF, T0693-D1 is a small helical region in the alpha-beta protein 4P7C, T0741-D1 is a V-shaped protein with two long beta hair-pins, and T0756-D2 is a helix bundle domain in the alpha-beta protein 4G6Q.

To investigate why helical proteins have much higher reconstruction accuracy with secondary structure input, we calculated the correlation between the percentage of helical residues in the proteins and reconstruction accuracies. For this, we selected all structural domains having at least one helix residue and computed the correlation between the percentage of helical residues in the proteins against the RMSD of the best models reconstructed with and without secondary structure input. When the reconstruction was carried out without secondary structures, we observed a Spearman's rank correlation coefficient of 0.58, between the percentage of helical residues and RMSD, suggesting that having more helical residues in a structure is likely to make the reconstruction more difficult. Then, we re-computed the correlations by adding secondary structures. When the reconstructions

Table 5.3: List of all domains with reconstruction accuracy below 0.5 TM-score. The models were reconstructed with contacts only. L, H, E, and N_c refer to length of the protein, number of helical residues, strand residues, and number of native contacts in the native structures, respectively. TM-score, RMSD, and GDT-TS of the best-of-20 models for each domain are presented. The last column (Energy) is the sum of the distance deviation from 8 Å for all the contacts supplied as distance restraints.

CASP	Domain	L	Type	H	E	Nc	TM-score	RMSD	GDT-TS	Energy
8	T0393-D2	99	TBM	74	0	50	0.29	10.6	27.8	0
8	T0405-D1	72	FM	58	0	67	0.42	7.2	45.5	0
8	T0443-D1	66	FM	41	0	42	0.45	6.6	50	0
8	T0443-D3	66	TBM	35	6	67	0.41	5.9	48.1	0.5
8	T0454-D2	140	TBM	94	0	141	0.49	6.6	40	1
8	T0470-D2	77	TBM	45	0	71	0.34	7.7	35.7	0
8	T0482-D1	67	FM	17	32	119	0.4	8	44	5.9
9	T0548-D2	60	TBM	43	0	45	0.42	7.4	49.6	0.2
9	T0553-D2	71	FM	46	0	59	0.49	4.6	52.8	0
9	T0575-D2	127	TBM	100	0	128	0.45	6.4	37	2.5
9	T0589-D2	82	TBM	58	0	74	0.48	5.2	48.8	0
9	T0598-D1	127	TBM	64	11	141	0.48	6.5	41.9	1
9	T0616-D1	97	FM	41	0	84	0.32	12.3	28.6	0.7
9	T0617-D1	136	TBM	96	8	143	0.49	11.8	43.2	8.1
9	T0629-D2	159	FM	0	4	31	0.16	25.2	12.1	0
9	T0637-D1	135	FM	109	0	75	0.33	16.1	24.4	0.3
9	T0639-D1	124	FM	76	4	133	0.36	8.3	30.9	2.7
10	T0680-D1	96	TBM	79	0	108	0.36	7.2	33.9	7.7
10	T0685-D1	72	TBM	54	0	42	0.27	8.5	31.3	0
10	T0693-D1	100	FM	47	12	101	0.38	14.7	34.5	1.3
10	T0724-D1	119	TBM	38	40	133	0.3	13.3	26.3	1.4
10	T0732-D2	91	TBM	48	0	91	0.44	5.8	46.2	1.5
10	T0741-D1	125	FM	0	73	218	0.45	17.1	39	5.8
10	T0756-D2	86	FM	45	0	15	0.25	12	25.9	0
11	T0820-D1	90	FM	65	0	72	0.4	7.3	41.9	0
11	T0821-D1	255	TBM	195	0	378	0.46	8.6	26.9	35.6
11	T0831-D1	155	TBM	114	0	141	0.44	15.8	34.8	1.5
11	T0836-D1	204	FM	157	0	198	0.38	12.9	22.8	7.2

were aided by secondary structures, the Spearman's rank correlation coefficient dropped to -0.14 (see **Figure 5.2**). This suggests that adding secondary structure information makes reconstruction accuracy nearly independent of the composition of helices in a protein. To check if a similar pattern is observed in beta proteins, we selected all domains having at least one beta strand, and calculated the Spearman's rank correlation coefficient between the best models' RMSD and the percentage of beta strand residues. In case of the beta proteins we found the correlation coefficient to be 0.15 when no secondary structures are used, suggesting no such correlation between difficulty of reconstruction and the number of strand residues in structures.

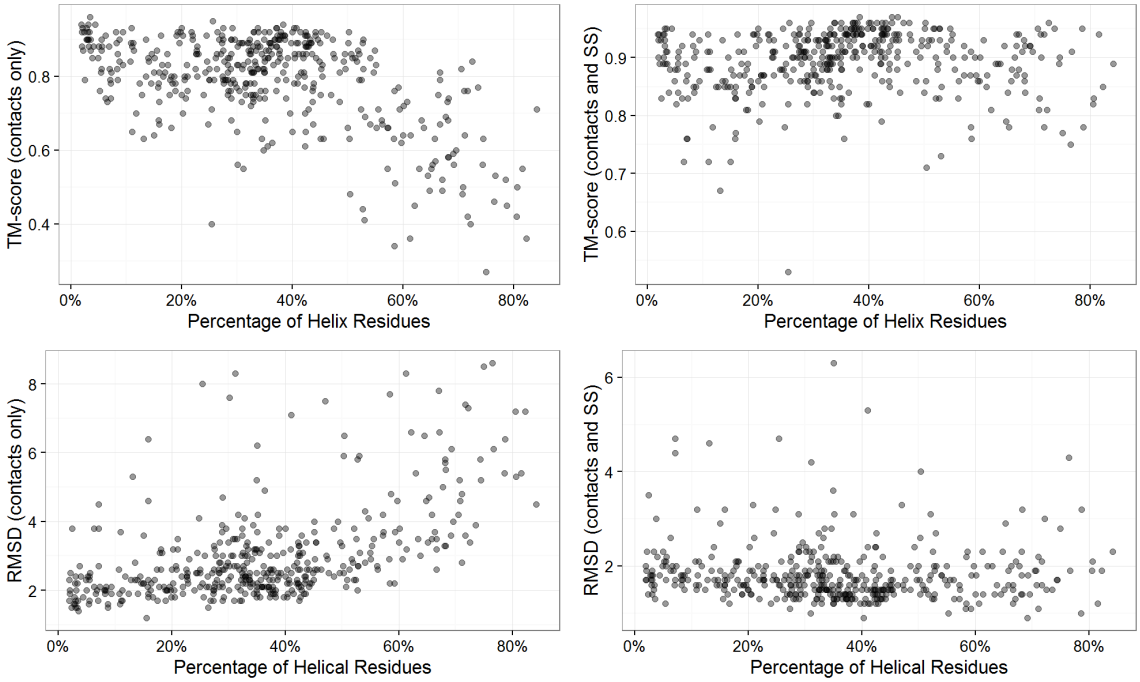

Figure 5.2: Analysis of the impact of the presence and absence of helix information on reconstruction. TM-score (plots in top row) and RMSD (plots in bottom row) of the best models when reconstructed without secondary structures (left two plots) and with secondary structures (right two plots).

Table 5.4: List of CASP domains for which reconstruction could not recover the fold (a) using contacts only or (b) using contacts and secondary structures. TM-score, RMSD, and GDT-TS of the best-of-20 models for each domain are presented. L, H, and E, refer to the length of the protein, number of helical residues, and number of strand residues, respectively.

CASP	Domain	L	H	E	Without SS			With SS		
					TM-score	RMSD	GDT-TS	TM-score	RMSD	GDT-TS
9	T0629-D2	159	0	4	0.16	25.2	12.1	0.16	21.4	12.4
10	T0693-D1	100	76	4	0.38	14.7	34.5	0.44	12	41.8
10	T0741-D1	125	0	73	0.45	17.1	39	0.39	13.1	32.8
10	T0756-D2	86	45	0	0.25	12	25.9	0.38	15.4	39.5

5.3.3 Reconstruction at higher distance thresholds for defining contacts

It is known that some structures are difficult to fold with some distance thresholds of defining contact. For instance, Human Myeloperoxidase Isoform C (1cxp chain B, 104 residues, all-alpha) could only be folded at a distance threshold of 16 Å instead of the more widely used 8 Å threshold [6]. For this protein structure, the authors showed that the RMSD drops from 41 Å to 4.9 Å when the contact distance threshold is increased from 7 Å to 16 Å. Similarly, in another work, authors found 14 Å distance threshold useful and reconstructed 87 protein chains using the same definition [49]. In this spirit, we tried to reconstruct the four 'hard-to-reconstruct' domains (T0629-D2, T0693-D1, T0741-D1, and T0756-D2) using various distance thresholds ranging from 8 Å to 20 Å. By testing these various distance thresholds along with secondary structure restraints, 3 out of the 4 structure domains could be correctly folded (TM-score > 0.5) with at least one of the distance thresholds (see **Figure 5.3A)**. These observations lead us to conclude that the reconstruction at higher distance thresholds can be useful for at least some structural folds. We find that the primary reason for more accurate reconstruction at the higher distance thresholds, is that increasing distance thresholds increases the number of contact restraints (see **Figure 5.3B**), thereby increasing the coverage of contacts and being particularly useful for many structural folds. The challenge, however, is that not all structures can be equally accurately folded at one distance threshold.

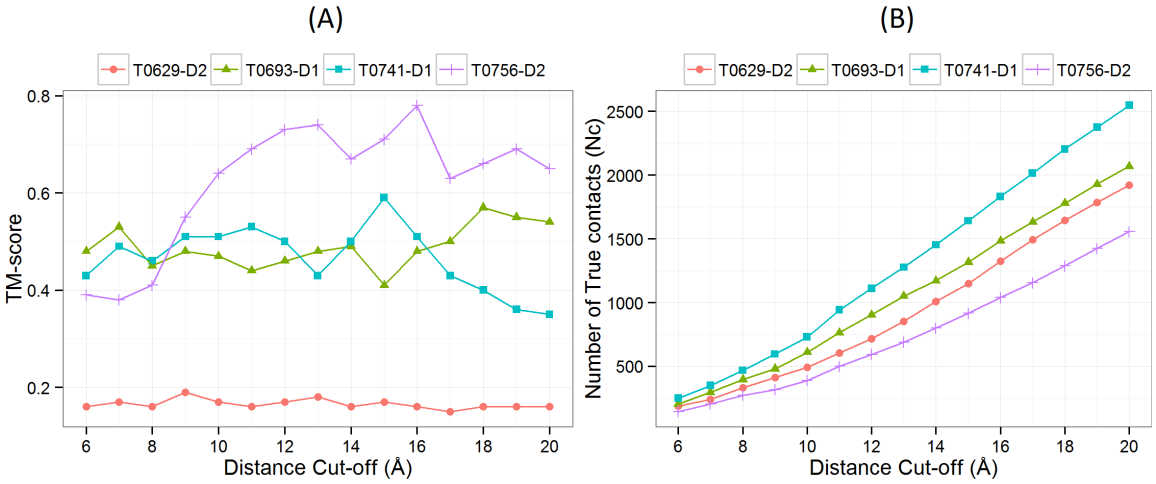

Figure 5.3: Improvement in reconstruction of 'hard to reconstruct' protein domains in CASP versus the increase distance cut-off thresholds (left) and the increase in number of contacts versus the increase of distance thresholds (right).

Absence of secondary structure elements in the structure, we find, is one reason for low reconstruction accuracy for these hard-to-fold proteins. One of these four structures, 159-residue domain T0629-D2, was the most difficult to reconstruct primarily because of its lack of secondary structure.

In fact, among all 496 CASP domains, this domain has the minimum percentage of secondary structure elements, i.e. 3%. Among the domains having minimum percentage of secondary structure elements, the next one is T0650-D1 with 20% of the residues forming secondary structures. The best model for this domain has GDT-TS of 0.5. **Figure 5.4** visualizes these four proteins showing how their non-globular structures impose challenges on reconstruction.

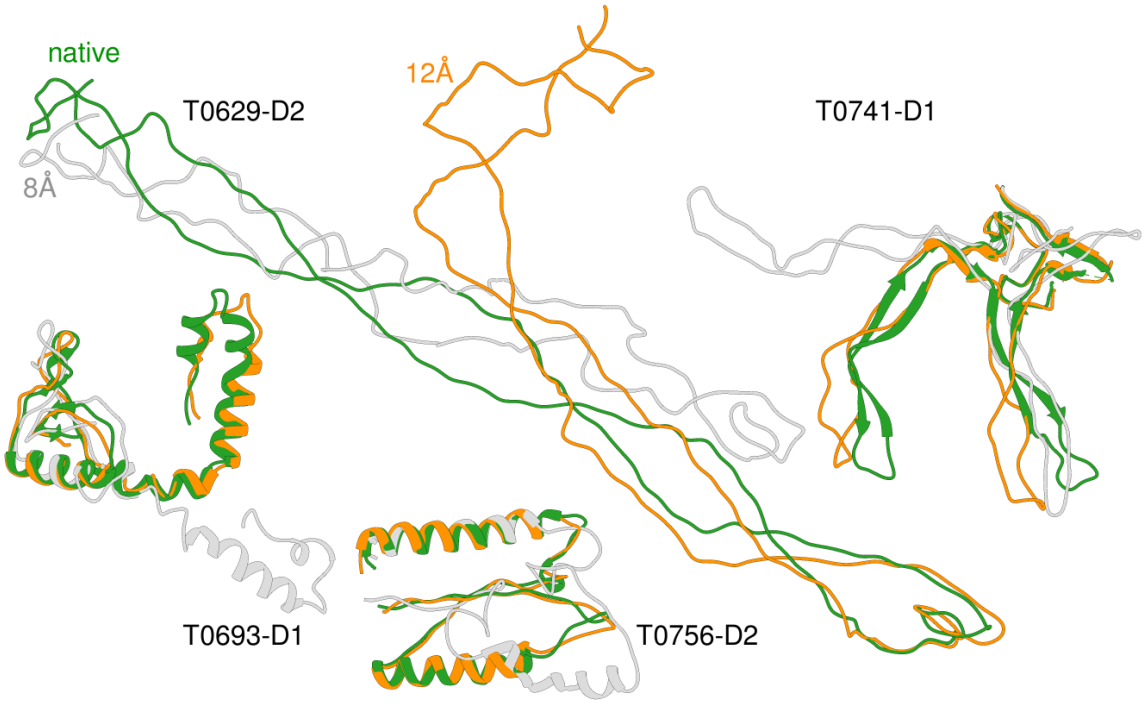

Figure 5.4: The true (native) structures of the domains T0629-D2, T0693-D1, T0741-D1, and T0756-D2 shown in green superimposed with structures reconstructed at distance cut-off of 8 Å (shown in grey), and at 12 Å (shown in orange).

5.3.4 Reconstruction with non-contacts

Different from all existing methods that use only contact information for reconstruction, we tested if adding non-contact information (a pair of residues whose distance is greater than a defined distance threshold) can increase the accuracy of reconstruction. To begin, we selected the same four hard-to-reconstruct proteins and reconstructed their models using both contacts and non-contact as restraints at various distance thresholds. **Figure 5.5** shows that at higher distance thresholds, non-contact information is surprisingly informative for reconstructing high-quality structures for three out of these four proteins. For at least one of the many distance thresholds, two of the four domains (T0693-D1 and T0756-D2) were reconstructed with around 1 Å RMSD and the third one (T0741-D1) with 2 Å RMSD. The hardest structure, T0629-D2, although showing some improvement with non-

contacts, still could not be folded, suggesting, again, that (a) some folds are hard to reconstruct, and (b) structures without secondary structure elements are among the most challenging structures to be reconstructed. For this domain (T0629-D2), to test if the knowledge of the quaternary structure of the domain could be useful for the reconstruction of the domain, we reconstructed the whole protein, with PDB ID 2XGF having 648 residues. Best-of-20 model, from such a reconstruction, had a TM-score of 0.32, suggesting that the knowledge of quaternary structure could not recover the fold of the domain.

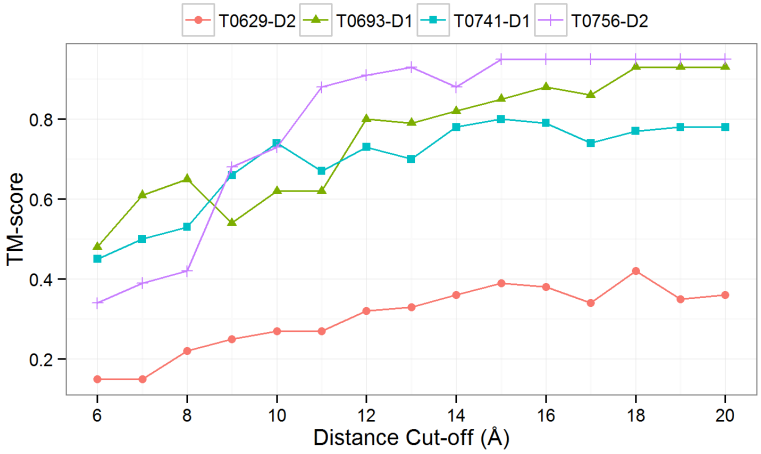

Figure 5.5: Reconstruction of the four hard-to-reconstruct CASP domains T0629-D2, T0693-D1, T0741-D1, and T0756-D2 using contacts and non-contacts at various contact thresholds.

For a more rigorous testing, we repeated our reconstruction tasks for all the 496 CASP domains using contacts defined at 8 Å threshold and the corresponding non-contacts. Specifically, we supplied the residue pairs not defined in true contacts list as non-contact restraints to CONFOLD, and observed around 2.5% improvement in TM-score on average. **Figure 5.6** shows that for 479 out of 496 structures, the accuracy either stays same or improves, suggesting that adding non-contact restraints improves the model reconstruction accuracy in most cases. This improvement from the addition of non-contacts is significant according to paired t-test of TM-scores between the models reconstructed with contacts and non-contacts and the models reconstructed using contacts only (p-value = 2.2 x 10^{-16}).

5.3.5 Shape of the structures and reconstruction difficulty

Using our largest dataset of 1901 proteins in the SCOP classification dataset, we reconstructed the structures using true contacts derived from the structures, to investigate the difficulty of reconstruction across various SCOP classes, and how this difficulty varies after inclusion of non-contacts. Our

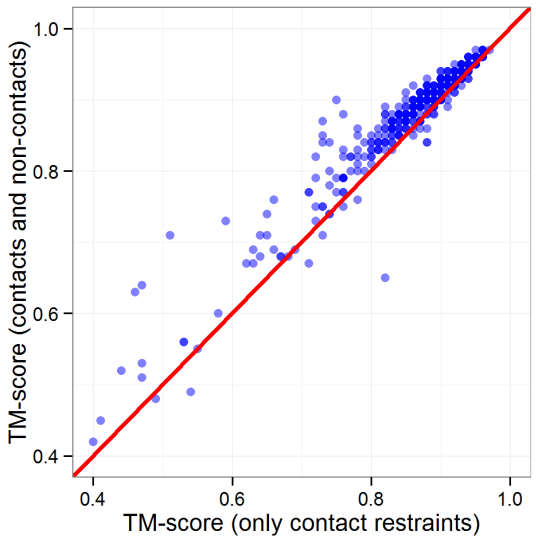

Figure 5.6: Improvement of adding non-contacts as restraints for CASP 8, 9, 10 and 11 target domains. (a) using contacts and secondary structure, and (b) using contacts and non-contacts together with secondary structures.

reconstruction results summarized in **Table 5.5**, which agree with the findings of [3], show that the average TM-score of the reconstructed models for class C proteins (alpha and beta (a/b) proteins) is 0.923 and are the easiest to reconstruct, followed by the class A (all alpha), B (all beta), and D (alpha and beta a+b). Similarly, the average TM-scores for membrane and cell surface proteins (class F) is 0.72, suggesting that the class is hardest to reconstruct. The smaller average TM-score of 0.68 for small proteins (class G) does not necessarily suggest that they are hardest proteins to reconstruct because the TM-score evaluation is not expected to perform well for short proteins [98]. This conclusion is supported by our observation that the average RMSD for the small proteins (3 Å) is much lower than the average RMSD for membrane and cell surface proteins (4.5 Å).

Furthermore, as shown in **Table 5.5**, on this large dataset, adding non-contacts improves the average TM-score of the reconstructed models to 0.84 from 0.816. **Figure 5.7** shows that the improvement from adding non-contacts is observed in all fold classes – all alpha proteins (class A), all beta proteins (class B), alpha and beta proteins (class C), alpha and beta proteins (class D), multi-domain proteins (class E), membrane and cell surface proteins (class F), and small proteins (class G). The addition of non-contacts, on average, improves the reconstruction accuracy for all protein classes but does not alter the relative difficulty of the classes.

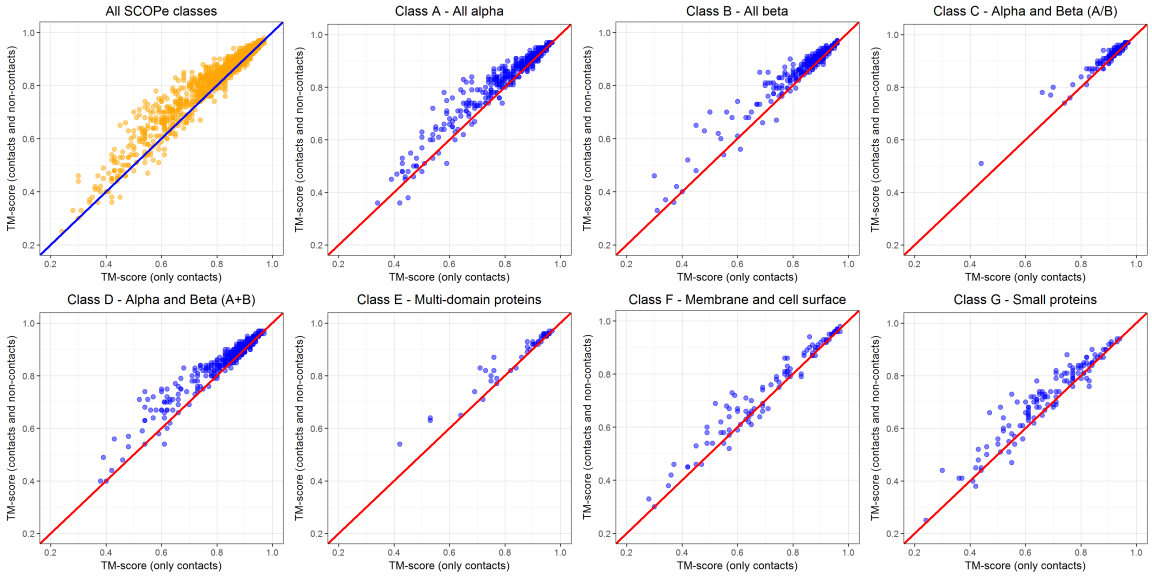

Figure 5.7: Improvement in reconstruction accuracy by using non-contacts together with the true contacts for all the 1901 proteins in the SCOP dataset and the seven classes (subsets). TM-scores of the best models reconstructed with contacts only are plotted against the TM-scores of the best models reconstructed with contacts and non-contacts.

Table 5.5: Reconstruction summary of the 1901 structural domains in SCOP dataset showing the reconstruction accuracy when only contacts are used and when non-contacts are added along with contacts. Best of 20 reconstructed models are reported.

SCOPe Class	**Class Description**	**Number of Domains**	**Using Contacts Only**		**Using Contacts and Non-Contacts**	
			TM-score	**RMSD**	**TM-score**	**RMSD**
A	All alpha proteins	500	0.829	2.74	0.854	2.46
B	All beta proteins	349	0.851	2.43	0.873	2.19
C	Alpha and beta proteins (a/b)	232	0.923	1.84	0.932	1.68
D	Alpha and beta proteins (a+b)	538	0.856	2.46	0.878	2.22
E	Multi-domain proteins (alpha and beta)	49	0.853	3.47	0.878	3.14
F	Membrane and cell surface proteins	102	0.719	4.54	0.745	4.08
G	Small proteins	131	0.680	3.02	0.717	2.50
	Total/Average	**1901**	**0.816**	**2.93**	**0.840**	**2.61**

5.3.6 Reconstruction at various sequence separation thresholds

It is widely understood that long range contacts (sequence separation of at least 24 residues) are the most important of the three contact types – short-, medium-, and long-range. To study how sequence separation affects the reconstruction accuracy of proteins, we reconstructed all the 496 CASP domains by removing contacts at various sequence separation thresholds, with and without the knowledge of secondary structure. Specifically, for each CASP structural domain, we removed all contacts closer than x residues in the corresponding sequence, where x = {0, 3, 6, ..., 51}, and reconstructed models using CONFOLD, with and without three-state secondary structure information. **Figure 5.8** shows that when secondary structures are used in reconstruction, the gain in accuracy from the use of local contacts (with sequence separation less than 6) is much lower. On average, when models are reconstructed using contacts, the mean reconstruction TM-scores at minimum sequence separation threshold of 6, 12, and 24 residues are 0.78, 0.74, and 0.55, respectively. Similarly, when secondary structures are added, the mean reconstruction TM-scores at minimum sequence separation threshold of 6, 12, and 24 residues are 0.88, 0.85, and 0.75, respectively. Setting sequence separation thresholds to 6, 12, and 24 correspond to removing local contacts, short-range contacts, and medium-range contacts, respectively. The relatively large drop in the accuracy at the sequence separation threshold of 24 residues suggests that compared to local contacts and short-range contacts, medium-range contacts are very important for reconstruction.

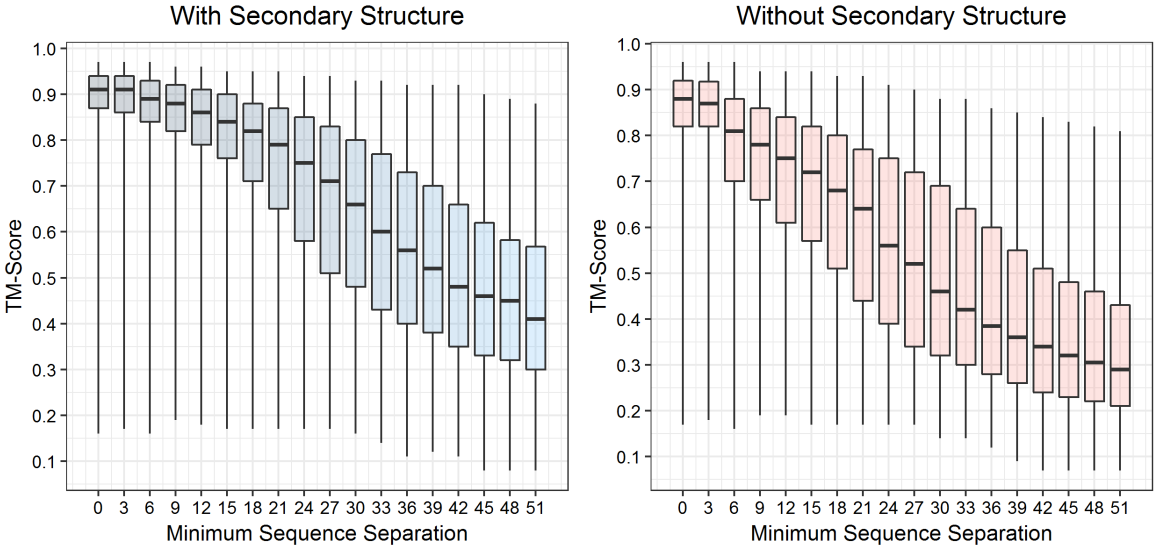

Figure 5.8: Reconstruction accuracy against various thresholds for sequence separation (for selecting contacts) on the 496 proteins in the CASP dataset.

5.4 Discussion

Realizing the importance of contact definition at higher distance thresholds, tools like NNcon [53] predict contacts at both distance thresholds – 8 Å and 12 Å. There are, however, challenges in predicting contacts at higher distance thresholds and utilizing them to build models. The first challenge is that the number of contacts increases rapidly as the distance threshold increases, making it harder for reconstruction methods to decide the number of contacts to consider for modeling. The second challenge is deciding the threshold that works for all proteins. Although the threshold of 8 Å between Cβup atoms is widely used, many studies demonstrate otherwise. For instance, Vassura et al., using a large data set of 1,760 proteins, found that increasing the distance threshold up to 18 Å improves the reconstruction accuracy monotonically. Similarly, Duarte et al., using a data set of 60 proteins, found that the best reconstruction accuracies were obtained with distance thresholds between 9 and 11 Å. Although these studies do not agree on the optimal cut-off distance, all of them demonstrate that contact restraints at higher distance thresholds are useful.

Following our finding that true non-contacts can help structure reconstruction, as the next step, we studied if predicted non-contact information can improve ab initio contact-guided modeling. For this we chose the contacts predicted by PSICOV for the 150 proteins [21] and built models with predicted contacts and compared with the models built using predicted contacts as well as predicted non-contacts. For predicting non-contact information, we did not use any additional method. Instead, in the same set of contacts predicted by PSICOV, we considered the contacts predicted with lowest confidence score (those having negative confidence values) as predicted non-contacts. Specifically, we selected top L predicted pairs as contacts and selected all pairs with predicted confidence less than -1 as predicted non-contacts. While the predicted contacts were translated into distance restraints of 3.5 Å to 8 Å between corresponding Cβup atoms, non-contacts were translated to distance restraints of 10 Å to 200 Å between corresponding Cβup atoms. We found that setting a slightly higher distance threshold of 10 Å instead of 8 Å yields better reconstruction accuracy. With these contacts and non-contacts, we reconstructed 20 models using CONFOLD and selected best model generated at reconstruction stages 1 and 2 for analysis. **Figure 5.9** shows that adding non-contact information improves the accuracy of the best reconstructed models for most proteins. When we selected residue pairs with confidence less than -1 as non-contacts, we observed 5% improvement in the TM-score on average; and 1.5% improvement with -2 as the threshold. This improvement from adding non-contacts is significant according to the paired t-test of TM-scores between the models in the second stage reconstructed with both contacts and non-contacts (selected

with contact prediction confidence less than -1) and the models in the second stage reconstructed with contacts only (p-value = 4 x 10^{-5}). Similar significant difference was observed when we compared the models in the first stage (p-value = 7 x 10^{-14}). We believe that better non-contact selection techniques can improve the reconstruction accuracy to much higher ranges.

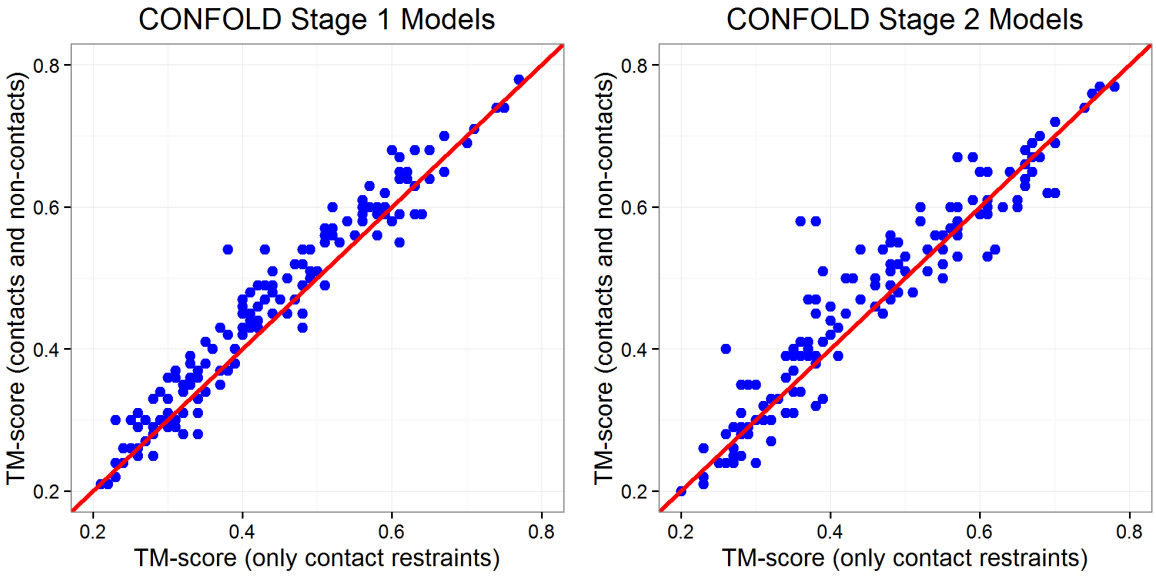

Figure 5.9: Improvement in reconstruction accuracy by using predicted non-contacts together with the predicted contacts for the 150 proteins in the PSICOV dataset in reconstruction stage 1 (left) and reconstruction stage 2 (right) of CONFOLD.

Finally, using the contacts predicted by MetaPSICOV [16] for the 496 structural domains in the CASP dataset, for each input sequence, we built models using CONFOLD. Our results, summarized in **Figure 5.10**, show that the accuracy of the reconstructed model (model having highest TM-score) is highly correlated to the precision of the predicted contacts, and the Pearson's correlation coefficient between the TM-score of the best predicted model and the precision of top L long-range contacts is 0.74. Compared to the average TM-score of 0.69, 0.78, and 0.80 for free-modeling (FM), template-based modeling hard (TBM-HA), and template-based modeling (TBM) domains when true contacts and secondary structures are used, when predicted contacts and secondary structures were used, we obtained average TM-scores of 0.40, 0.48, and 0.50 for FM, TBM-HA, and TBM domains, respectively. As expected, the relative difficulty of reconstruction between free-modeling domains and template-based domains is also pronounced when predicted contacts are used.

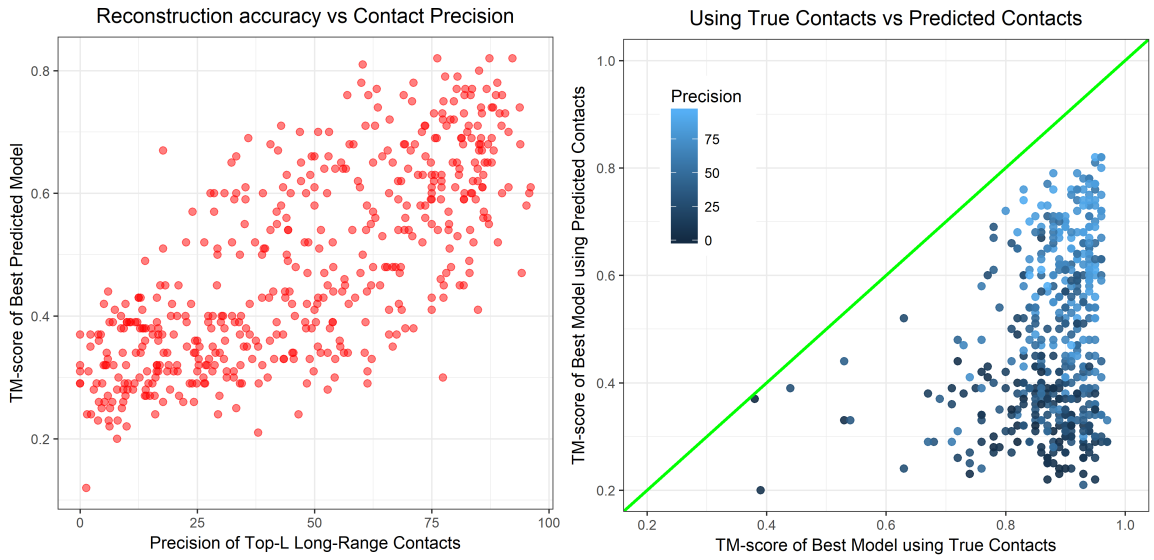

Figure 5.10: TM-scores of CONFOLD's best predicted model plotted against the precisions of top-L long-range contacts (left) and TM-scores of the best models reconstructed using true contacts plotted against the TM-scores of the best model reconstructed using predicted contacts (right) on the CASP domains dataset.

5.5 Methods

5.5.1 Contact definition

In this work, we define a pair of residues to be in contact if the distance between their Cβup atoms (Cαup in glycine) is less than 8 Å. Contacts separated by 6 to 11 residues in the corresponding sequence are categorized as short-range, contacts separated by 12 to 23 residues are categorized as medium-range, and those separated by 24 or more residues are defined as long-range contacts. In addition, we define contacting pairs, which are closer than 6 residues in the sequence as 'local' contacts. Local, short-range, medium-range, and long-range contacts all together make the complete contact map of a protein.

5.5.2 Data sets

For comparison with Reconstruct [4], we used the data set of 12 proteins used to benchmark it (see **Table 5.2** for the list of proteins). Similarly, for our analysis involving CASP's data sets, we considered all regular target domains released in CASP 8, 9, 10 and 11 having at least 60 residues. Domains like T0605-D1 that have no native contacts were also excluded from our data set. Our final data set consisted of 496 structural domains consisting of 72 free-modeling (FM) domains, 71 hard template-based modeling (TBM-HA) domains, 350 template-based (TBM) domains, and 3 'other'

domains (see **Table 5.6**).

In addition to the two datasets, for studying the reconstruction difficulty of various protein shapes (fold classes), we curated a structure dataset by selecting one protein from each superfamily within each fold of the seven classes (class A through G) of SCOP 2.04 database [100]. Since some of the proteins have many domains and are relatively very long, we removed all the proteins longer than 450 residues from our set. Our final set consisting of total 1901 proteins, has 500 all alpha proteins (class A), 349 all beta proteins (class B), 232 alpha and beta proteins (a/b) (class C), 538 alpha and beta proteins (a+b) (class D), 49 multi-domain proteins (class E), 102 membrane and cell surface proteins (class F), and 131 small proteins (class G).

Table 5.6: Number of free-modeling (FM) and template-based modeling (TBM) domains in CASP 8, 9, 10 and 11 competitions.

	FM	TBM-HA	TBM	Other	**Total**
CASP-8	8	48	93	0	**149**
CASP-9	23	3	106	0	**132**
CASP-10	11	12	89	0	**112**
CASP-11	30	8	62	3	**103**
Total	**72**	**71**	**350**	**3**	**496**

5.5.3 Reconstruction using true contacts

In order for our study not to be influenced by additional information (like information about structural fragments), we used our CONFOLD [25, 101] method to build models, which uses purely contacts (and secondary structure information when supplied) to build models. For reconstruction tests that involve using contacts only, we obtained contacts from the native structures/domains, and used them as input to CONFOLD to build 20 models. For evaluating the reconstructed models we use Template-Modeling score (TM-score), RMSD, and Global Distance Test (GDT-TS) score [98] and used the best of the 20 models for each target for assessment.

Following this protocol, we reconstructed the structural models of 12 proteins in the Reconstruct [4] dataset, as a benchmark for our reconstruction pipeline. Then we reconstructed models for the 496 proteins in the CASP 8, 9, 10, and 11 datasets using true contacts derived from the native structure. In addition, to study the relationship between the shape of the proteins and the difficulty of reconstruction, we reconstructed models for the 1901 proteins from the SCOP 2.04 [100] classification belonging to the seven classes (class A through G).

5.5.4 Reconstruction using contacts and secondary structures

In all the reconstruction experiments where we use true contacts and secondary structures, we derived secondary structures from the corresponding native structure using DSSP [97]. From the various DSSP assignments to each residue (strand, turn, alpha-helix, etc.), we translate all assignments except stand (E) and alpha-helix (H) to coil (C), such that our true secondary structures are in the same 3-state format as predicted contacts. For reconstruction, CONFOLD translates the input contacts into distance restraints, and secondary structures into distance restraints, dihedral angle restraints, and hydrogen-bond restraints (see the CONFOLD paper [25] for details). Following this protocol, we derived true contacts and secondary structures for two datasets (a) 496 proteins in the CASP dataset, and (b) 1901 proteins in the SCOP dataset. We generated 20 models for each protein and used the best model for our analysis and comparison with the models reconstructed using contacts only (without secondary structures).

5.5.5 Reconstruction using non-contacts and contacts at higher distance thresholds

From the dataset of 496 CASP structural domains, for the domains whose fold could not be recovered from reconstruction (i.e. TM-score of the best model is less than 0.5), we considered (a) increasing the threshold to define contacts, and (b) adding non-contacts along with contacts as restraints. Specifically, for each domain, we derived contacts between the carbon-atoms ($C\beta up$) of the residues from the native structure with minimum distance thresholds ranging from 8 Å to 20 Å and reconstructed models using these contacts. In addition, for such proteins, we also tested by providing non-contacts as an additional information (along with contacts) for reconstruction.

5.5.6 Contact prediction and reconstruction

In addition to the reconstructions using true contacts, for all the 496 CASP structural domains, instead of using true contacts and secondary structures, using the domains' sequence as input we predicted contacts and secondary structures and built models, to study the relationship between the models built using predicted and true contacts, and to study the relationship between predicted contact precision and reconstruction accuracy. For this, we predicted contacts using the state-of-the-art contact prediction method MetaPSICOV [16] and 3-state secondary structures using PSIPRED [40]. Many of the features needed by MetaPSICOV rely on the quality of multiple sequence alignments generated from the input sequence. For generating input multiple sequence alignments we used

HHblits [43] and JackHMMER [44] as discussed in [42]. Using MetaPSICOV's second stage contact predictions as input, we build 5 models with top xL contacts as input to CONFOLD, where x = {0.1, 0.2, 0.3, ..., 4.0} generating a total of 200 models for each protein. For our evaluation, we considered the best of these 200 predicted models.

5.6 Conclusions

In this study, we revisited the problem of protein structure reconstruction using true contacts focusing on the proteins whose structures are hard to predict. We show that increasing the distance threshold for defining contacts, using secondary structures, and adding non-contacts can improve the reconstruction accuracy of protein structures, particularly the ones that are hard to fold. Our findings provide useful insights to improve existing contact prediction and structure reconstruction/-folding methods.

Chapter 6

Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning

6.1 Abstract

In this work, we report the evaluation of the residue-residue contacts predicted by our three different methods in the CASP12 experiment, focusing on studying the impact of multiple sequence alignment, residue coevolution and machine learning on contact prediction. The first method (MULTICOM-NOVEL) uses only traditional features (sequence profile, secondary structure and solvent accessibility) with deep learning to predict contacts and serves as a baseline. The second method (MULTICOM-CONSTRUCT) uses our new alignment algorithm to generate deep multiple sequence alignment to derive coevolution-based features, which are integrated by a neural network method to predict contacts. The third method (MULTICOM-CLUSTER) is a consensus combination of the predictions of the first two methods. We evaluated our methods on 94 CASP12 domains. On a subset of 38 free-modeling domains, our methods achieved an average precision of up to 41.7% for top L/5 long-range contact predictions. The comparison of the three methods shows that the quality and effective depth of multiple sequence alignments, coevolution-based features, and machine learning integration of coevolution-based features and traditional features drive the quality of predicted protein contacts. On the full CASP12 dataset, the coevolution-based features alone

can improve the average precision from 28.4% to 41.6%, and the machine learning integration of all the features further raises the precision to 56.3%, when top L/5 predicted long-range contacts are evaluated. And the correlation between the precision of contact prediction and the logarithm of the number of effective sequences in alignments is 0.66.

6.2 Introduction

In the absence of homologous structural templates, a key input for successful *ab initio* protein structure prediction is residue-residue contacts [32, 17]. If a sufficient number of contacts can be predicted accurately, they alone can be used to reconstruct near native models for most proteins with accuracy of 2 Å RMSD [3, 4]. Among all the contacts, long-range contacts, which are generally harder to predict [13, 14, 15, 16], but much more useful for structure reconstruction [17]. Hence, recent contact prediction methods focus on the prediction and evaluation of long-range contacts, and so do the CASP experiments. When the contact prediction category was introduced in the CASP experiments, in the initial rounds, methods like SVMcon [14] and DNcon [13] that use support vector machines and deep learning networks with traditional features such as sequence profile, secondary structure and solvent accessibility, were often the top performers demonstrating that machine learning techniques were useful for contact prediction. Recent methods like PconsC2 [18], MetaPSICOV [16] and RaptorX method [19] show that including contact predictions from coevolution-based methods like CCMpred [20], PSICOV [21], and FreeContact [22] as additional features can significantly improve the performance, if at least a few hundred homologous sequences can be found for an input sequence. Often, when sufficient homologous sequences can be found, these 'meta' methods can predict top L/5 or L/10 long-range contacts with pretty high precision [16, 19, 20], where L is the length of the protein sequence. All these recently successful methods highlight that, besides machine learning techniques, coevolution-based features are important for accurate contact prediction.

Realizing the importance of coevolution-based features, which are entirely dependent upon the availability of homologous sequences, we developed a method for reliably generating deep multiple sequence alignments and coevolution-based features for accurate contact prediction, and participated in the recent CASP 12 experiment with three automated contact prediction methods - MULTICOM-NOVEL, MULTICOM-CONSTRUCT, and MULTICOM-CLUSTER. Our first method, MULTICOM-NOVEL, predicts contacts based on a deep learning contact prediction method - DNcon [13] that uses only traditional features such as sequence profile, secondary structure, and solvent accessibility. Our second method, MULTICOM-CONSTRUCT, relies on our deep multiple

sequence alignment generation algorithm to predict coevolution-based features, which are used by a consensus method MetaPSICOV [16] as input to make contact prediction. Our third method, MULTICOM-CLUSTER, combines the predictions from the first two methods by choosing their common highly ranked contacts. Our second and third predictors mainly rely on our deep alignment generation algorithm to make predictions. In this paper, we discuss the performance of our methods in the CASP12 experiment, primarily focusing on identifying the major factors influencing contact prediction accuracy. Since predicted contacts are most useful for protein sequences for which homologous structural templates cannot be found, we emphasize our analysis on free modeling (template-free) targets, although we also include our analysis for all CASP12 targets to assess the benefits of combining traditional features and coevolution-based features with machine learning.

Overall, our contact prediction methods were successful mainly because of our deep alignment generation algorithm, which generates high-quality alignments when sufficient homologous alignments can be found, and at least some alignments (if possible) when homologous sequences are hard to find. We find that multiple sequence alignments, coevolution-based features, and machine learning integration are the key factors for successful protein contact prediction. In addition to the analysis on predicted contacts, we also discuss some findings of building 3D structural models using the CONFOLD method [25] with our predicted contacts as input.

6.3 Materials and Methods

6.3.1 Generating deep multiple sequence alignments to derive coevolution-based features

Multiple sequence alignments (MSAs) play a central role for the success of a protein contact prediction method because the quality of multiple sequence alignment (MSA) entirely decides the accuracy of the coevolution-based contact prediction features, which largely determines the accuracy of overall contact prediction. Hence, it is crucial to have a reliable algorithm for producing high quality multiple sequence alignments. For reliability, it is important that the algorithm generates at least some sequences when homologous sequences are hard to find in sequence databases, and generates smaller but more useful alignments when an excessively large number of homologous sequences is available. On one hand, in the absence of any homologous sequences in the multiple sequence alignments or when there are just a few sequences, coevolution-based methods fail to make any predictions. On the other hand, when the size of alignment is too large (e.g. more than fifty thousand) and the

input protein sequence is long, some methods like PSICOV [21] may take too long to converge and sometimes do not produce any results even in a few days. Based on this understanding, we designed an alignment generation algorithm that attempts to generate high coverage alignments at first, and when sufficient homologous sequences are not found, relies on various sequence similarity cut-off thresholds to increase the depth of search to generate at least some sequences whenever possible.

For generating MSAs, we start by assuming sufficient homologous sequences covering most of our input sequence are available. Then we gradually switch towards choosing the settings that allow us to search deeply to generate at least some sequences. Using HHblits [43], we first generate alignments that cover 75% of a target sequence and check if the alignment has at least 2.5L sequences, where L is the length of the query sequence. If at least 2.5L sequences are not obtained, the coverage threshold is lowered, at first to 68% and then to 60% if needed. If none of these coverage thresholds deliver at least 2.5L sequences, we switch to using JackHMMER [44] to find remotely homologous sequences. Once again, we assume that sufficient significant hits can be found and start alignment search with a very stringent e-value cut-off threshold of $1E^{-40}$ to find homologous sequences. If this threshold fails to generate at least 2.5L sequences, we increase the e-value threshold to $1E^{-30}$, $1E^{-20}$, $1E^{-10}$, $1E^{-4}$, and 1, step by step, and conclude when more than 2.5L sequences are generated. If none of the thresholds leads to an alignment with more than 2.5L sequences, the alignments generated with high e-value threshold of 1 are used as the final alignment. A range of e-value thresholds is required because, for some input protein sequences, a stringent e-value criterion (like $1E^{-40}$) produces too few sequences (just a hundred or so) whereas a looser criterion (like $1E^{-4}$) generates many sequences. We used the 'UNIPROT20-2016' and 'UNIREF90' sequence databases for HHblits and JackHMMER search respectively.

6.3.2 MULTICOM contact prediction methods

Our first method, MULTICOM-NOVEL, is based on our method DNcon, an *ab initio* contact prediction method trained using deep belief networks and boosting [13, 12]. Unlike recent contact prediction methods that use co-evolutionary features as key features, it does not use any co-evolutionary information. Boosting and 'ensemble' are the key techniques that contribute to the performance of DNcon. Results from the training and testing experiments show that an ensemble of models trained at 7 window sizes (window sizes of 7, 9, 11, 13, 15, 17, and 19) delivers an accuracy of 34%, compared to 24% to 28% of individual models, on a test dataset of 196 proteins, when top L/5 long-range contacts are evaluated. DNcon was the top performer in the CASP10 contact prediction category

[12] and therefore serves as a good benchmark (baseline) to study where the improvement of contact prediction comes from in CASP12.

Our second method, MULTICOM-CONSTRUCT, primarily relies on our deep alignment generation algorithm to generate multiple sequence alignments, which are supplied as input to the three standard coevolution-based methods, PSICOV [21], CCMpred [22], and FreeContact [23] to generate two-dimensional co-evolution features to be combined by MetaPSICOV [16] with traditional features to make contact prediction. During the development of the method, we found that some coevolution-based methods like PSICOV sometime could not converge within a reasonable time limit when there were too many or too few sequences in alignments. To guarantee to generate predictions from such methods within a certain time limit, tweaking their convergence parameters is needed. Specifically, to get around the convergence issue of PSICOV, we run it with three convergence parameters ('d = 0.03', 'r = 0.001', and 'r = 0.01') in parallel and wait for a maximum of five hours. The 'd' parameter selects the glasso exact algorithm and is expected to produce more accurate results but is slow. The 'rho' parameter (r) controls how quickly the programs converges and higher values tend to speed up the convergence but at the loss of prediction accuracy. We pick the job that finishes within the five-hour time limit according to the order ('d = 0.03', 'r = 0.001', and 'r = 0.01'). In this way we are always able to have some prediction produced within the limited time. Such a shorter time limit was used during the CASP 12 experiment because our *ab initio* structure prediction methods used these predicted contacts as input to build 3D models, which themselves needed up to two days to build models.

Our third method, MULTICOM-CLUSTER, is a meta-predictor that combines contacts predicted by the first two methods. When at least 50 homologous sequences are found, this method uses the predictions made by MULTICOM-CONSTRUCT, otherwise combines the predictions of MULITICOM-CONSTRUCT and MULTICOM-NOVEL. As the first step for contact combination, we select two sets of up to 5L long-range contacts - one set from MULTICOM-NOVEL and another set from MULTICOM-CONSTRUCT. For each target, we first select top 5L contacts predicted by MULTICOM-NOVEL filtering out all the contacts not predicted by MULTICOM-CONSTRUCT, and then select up to top 5L contacts predicted by MULTICOM-CONSTRUCT which are present in the top 5L contacts from MULTICOM-NOVEL. This new set of contacts by MULTICOM-CONSTRUCT (having at most 5L contacts), and the set of contacts by MULTICOM-NOVEL are then updated by replacing their confidence values with the ranks, i.e. integer numbers starting from '5L' for the most confident contact prediction and ending at '1' for the least confident one. At this point, both sets have same contacts but different rankings. Then, the ranks for the MULTICOM-

CONSTRUCT's set are updated as the sum of the ranks in the two sets and are normalized by 10L. This new rank scores are then used to sort the contacts, as confidence scores, and used as input for MULTICOM-CLUSTER predictions. The final step is to scale the confidence values into a meaningful range between 0 and 1. Ideally, if we knew the number of long-range contacts in the target structure (N_c), we would normalize the confidence values such that the top $N_c$predictions have confidence more than 0.5. In the absence of such knowledge in reality, we normalize the confidence scores such that top L predicted contacts have confidence values more than 0.5, and submitted these contacts as MULTICOM-CLUSTER predictions.

6.3.3 Datasets and evaluation metrics

Out of the 90 targets released during the CASP 12 season, CASP12's official contact evaluations released at CASP's website were carried out on 70 targets (i.e. corresponding to 94 domains), excluding domains 'T0865-D1' and 'T0880-D1' because they do not have any long-range contacts. In this work, we consider all these 94 structural domains and its subset of 38 free-modeling domains for evaluation and comparison of our three methods, MULTICOM-NOVEL, MULTICOM-CONSTRUCT, and MULTICOM-CLUSTER. In this set of 94 domains, the native structures of 84 of them were available for our assessments. Hence, for some of our own evaluations, like evaluating the precision of co-evolutionary features, we use these 84 domains only. And to maintain consistency with the CASP released evaluations, we focus our analysis and evaluation at the domain level, although all predictions were made for the whole targets during the CASP12 experiment. Finally, before the CASP12 experiment, we used the dataset of CASP11 free-modeling domains to benchmark our methods. The results on CASP11 are also reported as a comparison with those on CASP12.

In addition to using our ConEVA contact evaluation toolkit [27] to do evaluation, we also referred to the evaluations published by CASP (released at http://predictioncenter.org/). We focus our evaluations on top L/5 and L/2 predicted long-range contacts and use precision as the primary evaluation metric, which is the fraction (ratio) of correct predictions in top predicted contacts. One important factor influencing the precision of contact prediction is the number of effective sequences in multiple sequence alignment, N_{eff}, which is calculated at the domain or target level using the following equation:

$$N_{eff} = \sum_{i=0}^{N} \frac{1}{n_i}$$

where N is the number of sequences in the multiple sequence alignment and n_i is the number of sequences which have at least 62% sequence identity with the i^{th} sequence. If all sequences in

the alignment are very different, n_i is 1 for each sequence and hence N_{eff} sums to N, and on the contrary, if all sequences are very similar, n_i is equal to N for all sequences and the sum of 1/N for N sequences gives 1, i.e. the N_{eff} is just 1. For calculating N_{eff} at the domain level, we trim the multiple sequence alignment column-wise, removing all the columns for which the reference native structure of a domain does not have any residues defined, so that the width of the alignment (number of columns) is same as the number of residues in the native structure of the domain.

6.4 Results and Discussion

6.4.1 Initial benchmark on CASP11 free-modeling dataset before CASP12 experiment

Prior to the CASP 12 experiment, we evaluated MULTICOM-CONSTRUCT that uses our new deep alignment generation algorithm to generate co-evolution features for contact prediction, on the dataset of 30 free-modeling structural domains of the CASP 11 experiment [102]. Following MULTICOM-CONSTRUCT's pipeline, we generated alignments and coevolution-based features for the 24 protein targets (with full targets as input) containing the 30 free-modeling domains, predicted contacts for the targets, and evaluated the predictions at the domain level. For comparison, on the same dataset, we also predicted contacts using the publicly available MetaPSICOV method with default options, where alignments were generated using HHblits [43]] with the coverage threshold parameter set to 60%. Moreover, we compared our results with the best performing group in the CASP11 contact prediction category, CONSIP2 [42], on the same dataset. The mean precisions of top L/5 long-range contacts predicted by MetaPSICOV, CONSIP2, and MULTICOM-CONSTRUCT are 29%, 29%, and 34.4% respectively (see **Table 6.1**). The improvement of our method is significant according to paired t-test of the difference in precision (p-value = 0.03). It is important to note that the same protein sequence database was used with MetaPSICOV and our method for a fair comparison. On average, our method can increase the number of sequences (N) in the alignment to 1546 (from 152), and the number of effective sequences (N_{eff}) to 222 (from 69), which is probably the primary contributor for the improvement (**Table 6.1**). For these free-modeling domains, the Pearson's correlation coefficient between the precision of top L/5 long-range contacts predicted by MULTICOM-CONSTRUCT and the logarithm of the number of effective sequences ($\log(N_{eff})$) in alignments is 0.60, which highlights the importance of the depth of multiple sequence alignments for contact quality. It is also important to note that the number of effective sequences was calculated at

the domain level. Pearson's correlation, when calculated using the number of effective sequences for the whole target alignment, gives much lower coefficients. This is because a high effective sequence number at whole target level does not guarantee a high number of effective sequences for each domain of a multi-domain target, as a sequence in an alignment may only cover a portion of the target.

Table 6.1: Top L/5 long-range contacts predicted by MULTICOM-CONSTRUCT method compared with the top L/5 contacts predicted using the default MetaPSICOV method and the CONSIP2 method, on the 30 CASP11 free-modeling domains. N_{target} is the number of sequence in the alignment which is generated with the target sequence as input. $Neff_{domain}$ is number of effective sequences in the alignment when alignments are trimmed to match the residues of the native structural domain. $P_{L/5}$ refers to the precision of top L/5 long-range contacts.

Domain	**MetaPSICOV**			**CONSTRUCT**			**CONSIP2**
	$\mathbf{N}_{target}$	$\mathbf{Neff}_{domain}$	$\mathbf{P}_{L/5}$	$\mathbf{N}_{target}$	$\mathbf{Neff}_{domain}$	$\mathbf{P}_{L/5}$	$\mathbf{P}_{L/5}$
T0761-D1	1	1	0	4	2	0	5.6
T0761-D2	1	1	13	4	2	13	8.7
T0763-D1	3	2	30.8	7	3	15.4	46.2
T0767-D2	109	58	66.7	774	88	66.7	58.3
T0771-D1	9	4	26.7	32	11	16.7	10
T0777-D1	55	25	15.9	747	41	18.8	23.2
T0781-D1	2	2	10	40	15	2.5	5
T0785-D1	1	1	4.6	6	2	4.6	18.2
T0789-D1	274	133	44.8	2465	484	62.1	51.7
T0789-D2	274	139	44	2465	522	60	28
T0790-D1	276	140	44.4	1829	440	59.3	44.4
T0790-D2	276	136	26.9	1829	455	69.2	26.9
T0791-D1	265	109	63.3	2488	401	66.7	53.3
T0791-D2	265	118	35.7	2488	481	75	42.9
T0794-D2	258	121	52.9	1653	176	38.2	26.5
T0806-D1	766	369	62.8	1130	306	70.6	84.3
T0808-D2	121	29	27.8	1257	92	27.8	35.2
T0810-D1	49	24	21.7	8669	1147	21.7	17.4
T0814-D1	118	106	25.9	1404	145	48.2	37
T0814-D2	118	107	69.6	1404	174	73.9	82.6
T0820-D1	1	1	5.6	1	1	5.6	5.6
T0824-D1	79	32	36.4	1257	254	72.7	45.5
T0827-D2	680	229	26.7	3164	558	20	10
T0831-D2	189	100	10.3	4659	242	7.7	7.7
T0832-D1	5	2	2.4	83	26	4.8	2.4
T0834-D1	42	21	0	269	48	0	5
T0834-D2	42	16	5.9	269	45	5.9	17.7
T0836-D1	223	26	36.6	2627	167	68.3	43.9
T0837-D1	32	8	37.5	132	10	37.5	29.2
T0855-D1	11	7	21.7	3234	322	0	17.4
Average	**152**	**69**	**29**	**1546**	**222**	**34.4**	**29.7**

6.4.2 Performance on CASP12 dataset

Table 6.2 summarizes the performance of our three methods on the subset of 38 CASP12 domains classified as free modeling. The mean precision of top L/5 long-range contacts predicted by our

three methods MULTICOM-NOVEL, MULTICOM-CONSTRUCT, and MULTICOM-CLUSTER are 25.4%, 41.6% and 41.7% respectively. MULTICOM-CONSTRUCT and MULTICOM-CLUSTER, which rely on our deep multiple sequence alignment generation algorithm and coevolution-based features, have much higher mean precision compared to the baseline sequence-based machine learning method MULTICOM-NOVEL without using coevolution features, suggesting the enhanced coevolution features is a major contributor to the improved precision. On this free-modeling dataset, our contact combination method, MULTICOM-CLUSTER, has improved performance on two domains T0869-D1 and T0923-D1, although, on average, its performance is similar to the MULTICOM-CONSTRUCT method. For 23 out of these 38 domains, our deep alignment generation algorithm concluded with alignments generated by JackHMMER at high e-value threshold of 1, suggesting that most of the domains in the free-modeling dataset did not have sufficient significantly homologous sequences with high coverage. The low-quality alignments, generated by JackHMMER at e-value threshold of 1, have the number of effective sequences ranging from 1 to 1331 (with mean as 107 and median as 31), and the precision of MULTICOM-CONSTRUCT's contact predictions for these domains ranges from 3% to 95%. This suggests that high e-value thresholds do not always necessarily generate poor alignments, but rather lead to alignments of variable quality, some of which are useful for contact prediction.

On the full dataset consisting of all 94 CASP12 domains, the mean precision of top L/5 long-range contacts for MULTICOM-NOVEL, MULTICOM-CONSTRUCT, and MULTICOM-CLUSTER are 25.8%, 50.3% and 50.1% respectively. Higher precisions on the complete dataset is due to the fact that the mean N_{eff} for all the targets is 1619, greater than 253 for the free-modeling targets. Finally, the same as on CASP11 free-modeling dataset, we observed a Pearson's correlation coefficient of 0.66 between the precision of top L/5 long-range contacts predicted by MULTICOM-CONSTRUCT and the logarithm of the number of effective sequences (N_{eff}) on the CASP12 full dataset. Since it is relevant to compare the performance of all three methods on the target domains for which no sufficient number of sequences in alignments were found, we selected six free-modeling domains for which our method generated less than 20 sequences in the alignments. For these targets, while MULTICOM-NOVEL and MULTICOM-CONSTRUCT have average precision of 15% and 15.9% respectively, the contact combination made by MULTICOM-CLUSTER has average precision of 16.7%, showing a slight improvement.

Table 6.2: Comparison of top L/5 long-range contacts predicted by our three methods, MULTICOM-NOVEL, MULTICOM-CONSTRUCT, and MULTICOM-CLUSTER for the 38 free-modeling structural domains using precision measure. L, N_{target}, and $Neff_{domain}$ stand for the length of the target sequence, number of sequence in the alignment for the whole target sequence, and the number of effective sequences in the alignment when alignments are trimmed to match the residues of the native structural domain, respectively. The last three columns show the precision of top L/5 long-range contacts for the three methods. The 'Alignment' column shows the method and parameter used to generate the alignment, where 'jhm' stands for JackHMMER and 'hhb' stands for HHblits.

Target	**FM Domain**	**L**	**Alignment**	**N_{target}**	**$Neff_{domain}$**	**CONSTRUCT**	**CLUSTER**	**NOVEL**
T0859	T0859-D1	133	jhm-e-0	2	1	4.4	4.4	0
T0862	T0862-D1	239	jhm-e-0	163	31	26.3	26.3	21.1
T0863	T0863-D1	670	jhm-e-0	453	73	2.6	2.6	5.1
T0863	T0863-D2	670	jhm-e-0	453	54	4.2	4.2	4.2
T0864	T0864-D1	246	jhm-e-0	526	134	64	64	32
T0866	T0866-D1	183	hhb-cov75	1388	560	100	100	14.3
T0869	T0869-D1	120	jhm-e-0	17	12	42.9	52.4	47.6
T0870	T0870-D1	138	jhm-e-0	137	81	16	16	40
T0878	T0878-D1	358	jhm-e-0	856	250	42	42	26.1
T0880	T0880-D2	193	jhm-e-0	2	1	25	21.9	18.8
T0886	T0886-D1	346	jhm-1e-40	3013	1182	78.6	78.6	7.1
T0886	T0886-D2	346	jhm-1e-40	3013	1837	88.5	88.5	23.1
T0888	T0888-D1	121	jhm-e-0	2	1	8	0	0
T0890	T0890-D2	191	jhm-e-0	70	17	13.6	13.6	9.1
T0892	T0892-D2	193	jhm-e-0	579	202	54.6	54.6	63.6
T0894	T0894-D1	324	jhm-e-0	438	61	11.1	11.1	55.6
T0896	T0896-D3	486	jhm-e-0	2295	7	12.1	12.1	9.1
T0897	T0897-D1	285	jhm-e-0	130	10	7.1	7.1	17.9
T0897	T0897-D2	285	jhm-e-0	130	57	52	52	20
T0898	T0898-D1	169	jhm-1e-4	50000	389	4.6	4.6	13.6
T0899	T0899-D1	423	jhm-1e-10	6580	125	71.2	71.2	40.4
T0899	T0899-D2	423	jhm-1e-10	6580	31	44.4	44.4	33.3
T0900	T0900-D1	106	jhm-e-0	16243	1331	95.2	95.2	71.4
T0901	T0901-D2	328	hhb-cov50	5167	127	64.3	64.3	42.9
T0904	T0904-D1	341	jhm-1e-10	23741	609	72.6	72.6	29.4
T0905	T0905-D1	353	jhm-1e-10	8623	346	79.6	79.6	63.3
T0905	T0905-D2	353	jhm-1e-10	8623	88	42.9	42.9	42.9
T0907	T0907-D3	315	jhm-e-0	219	1	79.2	79.2	41.7
T0912	T0912-D3	624	jhm-1e-20	7240	426	42.9	42.9	4.8
T0914	T0914-D1	337	jhm-e-0	325	70	6.3	6.3	31.3
T0914	T0914-D2	337	jhm-e-0	325	33	6.1	6.1	15.2
T0915	T0915-D1	161	jhm-e-0	34	21	48.4	45.2	29
T0918	T0918-D1	546	jhm-1e-20	3517	356	77.3	77.3	40.9
T0918	T0918-D2	546	jhm-1e-20	3517	487	88	88	20
T0918	T0918-D3	546	jhm-1e-20	3517	513	66.7	66.7	0
T0923	T0923-D1	409	jhm-e-0	10	7	12.1	19	22.4
T0941	T0941-D1	470	jhm-e-0	3	1	2.9	2.9	1.5
T0946	T0946-D1	292	hhb-cov50	3170	80	25	25	6.3
Average					**253**	**41.6**	**41.7**	**25.4**

6.4.3 Significance of coevolution-based features and machine learning integration

If reliable and deep multiple sequence alignments are available, two-dimensional pairwise features (contact probabilities or scores) predicted by coevolution-based methods are a key factor for high accuracy in final contact prediction. To study the significance of these features, we evaluated the precision of the coevolution-based contacts predicted by PSICOV, CCMpred and FreeContact separately, and compared them with the final prediction made by MULTICOM-CONSTRUCT. Excluding some targets for which PSICOV failed to converge within the five-hour time limit and some additional targets for which no more than 5 homologous sequences could be found, on the remaining 70 structural domains, CCMpred, FreeContact, and PSICOV have mean precision of 41.6%, 36.3%, and 34.1% respectively, for top L/5 long-range contacts. When top L/2 contacts are evaluated, the similar trend is observed for the three methods with the mean precision of 32.6% for CCMpred, 28.3% for FreeContact, and 25.4% for PSICOV, suggesting that the most accurate single coevolution-based predictor is CCMpred followed by FreeContact and PSICOV (see **Table 6.3**). These precisions are much higher than the average precision (25.4%) of our baseline MULTICOM-NOVEL method that does not use any coevolution-based features as input. These results indicate that the coevolution-based features are crucial for accurate contact prediction.

In **Table 6.3** MULTICOM-CONSTRUCT has much higher mean precision of 56.3% for top L/5 long-range predictions (46.2% for top L/2), compared to each of the three individual coevolution-based features above. If we selected the best contact predictions made by the three coevolution-based predictions to evaluate for each domain, the mean precision (called maximum in **Table 6.3**) is 44.2% for top L/5 contacts, which is only slightly (2.6%) better than the performance of the best individual coevolution-based feature predictor CCMpred, but is still much lower than the mean precision 56.3% of MULTICOM-CONSTRUCT. These results indicate that, in addition to coevolution-based features being important, the machine learning approaches of integrating these coevolution-based features with the traditional sequence-based features are also very important. Analyzing the predictions made by MULTICOM-CONSTRUCT, we only found 2 out of 70 domains (T0918-D3 and T0912-D2) for which the machine learning integration had failed to perform better than an individual coevolution-based feature. Upon inspecting the three-dimensional structures of these two domains, however, we find both of them have the middle region of the structure missing, which might cause the failure of the machine learning integration. Generally speaking, in MULTICOM-CONSTRUCT, the neural network-based combination of the multiple co-evolution features and traditional features almost

Table 6.3: Precision of top L/5 and L/2 contacts predicted for CASP12 structural domains using PSICOV, FreeContact, and CCMpred, the maximum precision of the three methods, and the MULTICOM-CONSTRUCT (MULTICOM) method of using machine learning to integrate multiple co-evolution features. This dataset excludes the cases in where PSICOV failed to generate any results within the time limit.

Domain	PSICOV		FreeContact		CCMpred		Maximum		MULTICOM	
	L/5	L/2	L/5	L/2	L/5	L/2	L/5	L/2	L/5	L/2
T0861-D1	79	54.5	79	66	83.9	77.6	83.9	77.6	85.5	81.4
T0862-D1	0	0	0	0	15.8	6.4	15.8	6.4	26.3	12.8
T0863-D1	2.6	3.1	0	0	2.6	2.1	2.6	3.1	2.6	7.2
T0863-D2	1.4	1.7	0	0	0	0.6	1.4	1.7	4.2	3.4
T0864-D1	20.4	14.6	42.9	25.2	55.1	26.8	55.1	26.8	65.3	45.5
T0866-D1	95.2	63.5	81	63.5	95.2	71.2	95.2	71.2	100	78.9
T0868-D1	13	10.3	4.4	5.2	17.4	13.8	17.4	13.8	82.6	60.3
T0869-D1	14.3	12.5	0	3.9	19.1	13.5	19.1	13.5	42.9	36.5
T0870-D1	16	9.7	12	6.5	16	9.7	16	9.7	16	8.1
T0871-D1	73.4	50	57.8	38.8	81.3	61.3	81.3	61.3	93.8	79.4
T0872-D1	27.8	18.2	33.3	18.2	33.3	22.7	33.3	22.7	66.7	31.8
T0873-D1	43.5	26.8	66.3	55.4	66.3	58.9	66.3	58.9	82.6	70.6
T0877-D1	10.7	7	7.1	5.6	10.7	8.5	10.7	8.5	17.9	21.1
T0878-D1	27.5	18.6	39.1	21.5	36.2	20.4	39.1	21.5	42	29.7
T0879-D1	81.8	70	75	70.9	77.3	73.6	81.8	73.6	97.7	85.5
T0881-D1	5	4	2.5	5	5	5	5	5	0	3
T0882-D1	6.3	5	12.5	7.5	18.8	10	18.8	10	6.3	10
T0884-D1	14.3	13.9	7.1	5.6	7.1	8.3	14.3	13.9	7.1	13.9
T0885-D1	56.5	35.1	39.1	33.3	47.8	33.3	56.5	35.1	95.7	61.4
T0886-D1	71.4	57.1	78.6	68.6	78.6	77.1	78.6	77.1	78.6	77.1
T0886-D2	80	50	88	60.9	92	60.9	92	60.9	88	82.8
T0889-D1	89.6	78.3	87.5	80.8	87.5	80.8	89.6	80.8	95.8	90.8
T0890-D1	25	24.4	12.5	9.8	12.5	7.3	25	24.4	43.8	22
T0890-D2	19.1	11.3	0	0	0	5.7	19.1	11.3	14.3	11.3
T0891-D1	63.6	41.1	59.1	42.9	68.2	46.4	68.2	46.4	90.9	87.5
T0892-D1	21.4	14.3	42.9	22.9	50	25.7	50	25.7	35.7	28.6
T0892-D2	22.7	16.4	31.8	18.2	18.2	14.6	31.8	18.2	54.6	49.1
T0893-D1	0	2.7	0	5.4	6.7	8.1	6.7	8.1	6.7	8.1
T0893-D2	91.2	80	91.2	80	94.1	83.5	94.1	83.5	97.1	89.4
T0894-D1	11.1	11.1	27.8	15.6	22.2	13.3	27.8	15.6	11.1	13.3
T0894-D2	18.2	18.5	27.3	14.8	36.4	18.5	36.4	18.5	54.6	33.3
T0895-D1	4.2	5	4.2	5	12.5	5	12.5	5	33.3	30
T0897-D1	0	0	0	1.5	0	0	0	1.5	7.1	7.3
T0897-D2	24	22.6	20	14.5	32	24.2	32	24.2	52	25.8
T0898-D1	0	0	0	0	0	0	0	0	4.8	7.6
T0898-D2	0	0	9.1	14.3	9.1	3.6	9.1	14.3	9.1	10.7
T0899-D1	26.9	20.8	26.9	19.2	28.9	15.4	28.9	20.8	71.2	49.2
T0899-D2	16.7	13.6	5.6	6.8	11.1	11.4	16.7	13.6	44.4	36.4
T0900-D1	50	43.1	50	45.1	65	56.9	65	56.9	95	80.4
T0901-D1	62.2	46.4	62.2	47.3	60	48.2	62.2	48.2	84.4	67
T0901-D2	14.3	11.4	7.1	5.7	0	2.9	14.3	11.4	64.3	40
T0902-D1	67.4	56	73.9	60.3	67.4	65.5	73.9	65.5	93.5	88.8
T0903-D1	27.7	22.8	1.5	4.3	27.7	18.5	27.7	22.8	98.5	80.9
T0904-D1	50	31.8	24	16.7	70	44.4	70	44.4	74	48.4
T0905-D1	33.3	23.1	41.7	26.5	41.7	26.5	41.7	26.5	79.2	54.6
T0905-D2	30.8	24.2	0	6.1	7.7	12.1	30.8	24.2	46.2	48.5
T0909-D1	27	15.1	20.3	19.7	44.3	28.4	44.3	28.4	43.1	32.7
T0911-D1	65.9	50.5	78.1	64.7	72	69.1	78.1	69.1	86.6	76
T0912-D1	20.5	20.2	84.3	69.1	78.3	63.8	84.3	69.1	91.6	82.1
T0912-D2	29.4	18.5	58.8	42.9	58.8	47.6	58.8	47.6	41.2	33.3
T0912-D3	0	0	14.3	15.4	23.8	17.3	23.8	17.3	42.9	28.9
T0913-D1	48.5	34.3	64.7	48.5	79.4	57.4	79.4	57.4	69.1	62.1
T0914-D1	6.3	5.1	3.1	2.5	6.3	2.5	6.3	5.1	6.3	8.9
T0914-D2	9.4	7.4	3.1	2.5	6.3	2.5	9.4	7.4	6.3	4.9
T0915-D1	6.5	6.5	6.5	2.6	0	2.6	6.5	6.5	48.4	27.3
T0917-D1	82.1	70.4	76.9	66.3	89.7	79.6	89.7	79.6	97.4	84.7
T0918-D1	40.9	27.8	50	40.7	59.1	50	59.1	50	77.3	59.3
T0918-D2	48	29	60	48.4	68	58.1	68	58.1	88	71
T0918-D3	25	15.3	75	47.5	75	52.5	75	52.5	66.7	45.8
T0920-D1	85.9	65.8	87.5	75.8	89.1	78.3	89.1	78.3	93.8	88.8
T0920-D2	0	0	2.3	1.8	4.6	2.7	4.6	2.7	22.7	18.2
T0921-D1	64.3	34.8	60.7	46.4	57.1	39.1	64.3	46.4	96.4	76.8
T0922-D1	26.7	27	33.3	29.7	33.3	32.4	33.3	32.4	53.3	46
T0928-D1	52.9	36.3	72.1	43.3	66.2	51.5	72.1	51.5	79.4	63.2
T0944-D1	70.6	44.1	62.8	49.6	68.6	58.3	70.6	58.3	88.2	66.9
T0945-D1	29.3	25	46.7	28.2	73.3	53.2	73.3	53.2	86.7	64.9
T0946-D1	12.5	10	0	2.5	25	17.5	25	17.5	25	30
T0946-D2	66.7	52.8	66.7	50.9	61.9	57.6	66.7	57.6	81	73.6
T0947-D1	57.1	36.4	65.7	46.6	77.1	50	77.1	50	80	67.1
T0948-D1	3.3	4	16.7	9.3	6.7	6.7	16.7	9.3	6.7	10.7
Mean	**34.1**	**25.4**	**36.3**	**28.3**	**41.6**	**32.6**	**44.2**	**34.1**	**56.3**	**46.2**

always performs better than individual coevolution-based features. Taking domain T0868-D1 as an example, when top L/5 long-range contacts are evaluated, the predictions by PSICOV, CCMpred, and FreeContact have precision of 13%, 4.4%, and 17.4% respectively, the final prediction made by MULTICOM-CONSTRUCT, however, boosts the precision to 82.6%. As shown **Figure 6.1**, the contacts predicted by MULTICOM-CONSTRUCT (**Figure 6.1(A)**) are much more near-native compared to the individual coevolution-based predictions.

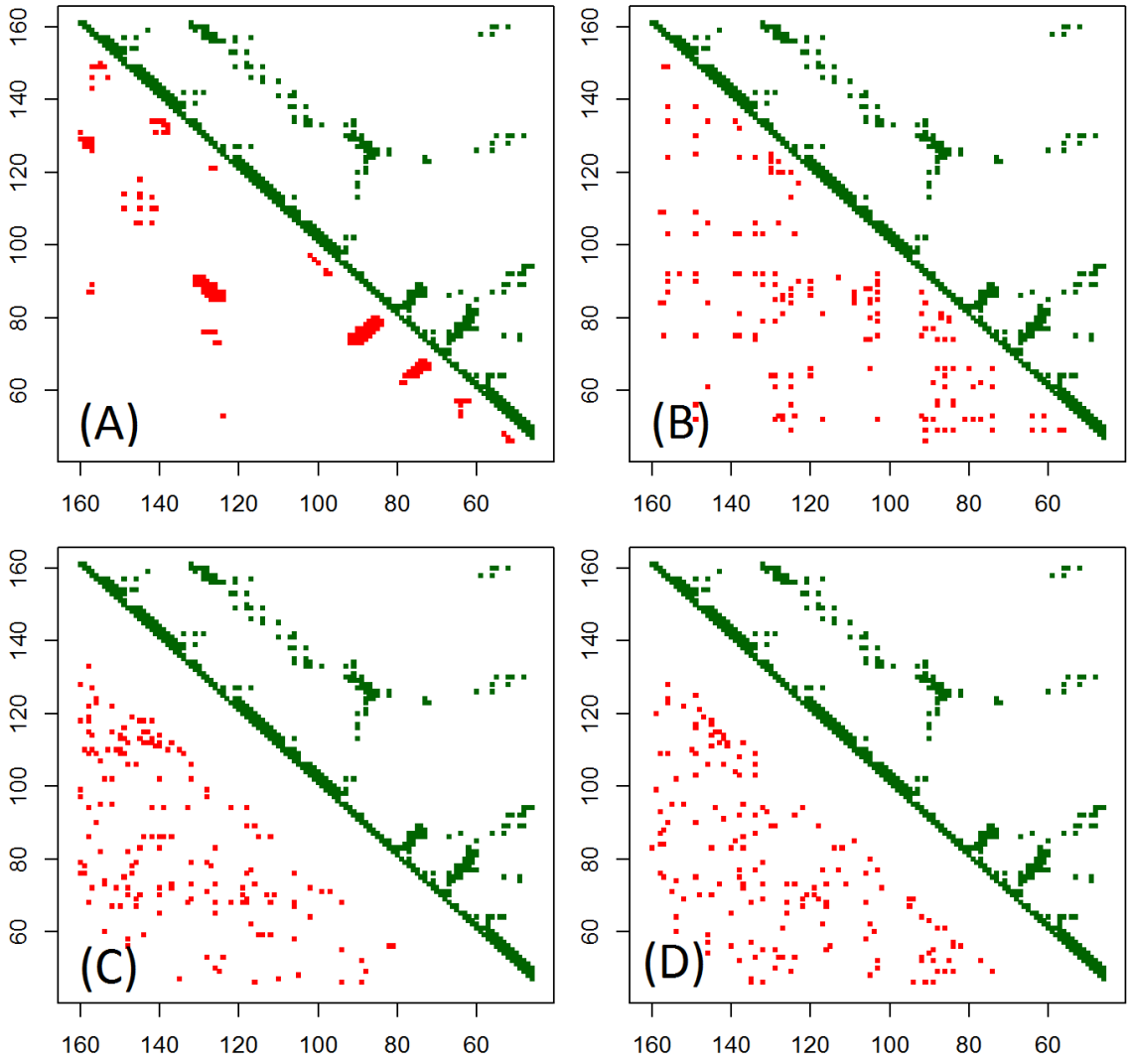

Figure 6.1: Contact map visualization of top L contacts predicted by MULTICOM-CONSTRUCT **(A)**, PSICOV **(B)**, FreeContact **(C)**, and CCMpred **(D)** for the target domain T0868-D1. Green dots in upper triangles represent contacts in the native structure and red dots in lower triangles denote the contact predictions.

6.4.4 Relationship between number of effective sequences and precision of contact prediction

Study of the relationship between the number of effective sequences (N_{eff}) in the alignment and the precision of the predicted contacts can provide useful insights on estimating the accuracy of the predicted contacts. A direct comparison between N_{eff} and precision is less meaningful if N_{eff} is calculated for the whole target sequence and the contact precision are evaluated at the domain level. Hence, we also calculated N_{eff} using our N_{eff} calculation method at the domain level. **Figure 6.2** plots the precisions of top L/5 contact predictions of the domains in CASP12 dataset against the logarithm of their number of sequences (N) in the alignments generated for the whole targets and the logarithm of the number of effective sequences (N_{eff}) at the domain level, respectively. The Pearson's correlation between the precision and log(N) is 0.47, lower than 0.66 between the precision and log(N_{eff}) at the domain level. According to the plot between contact prediction precision and N_{eff} in **Figure 6.2**, it can be inferred that multiple sequence alignments with at least around 100 effective sequences at domain level has a good chance to produce 50% precise contact predictions, whereas, when the N_{eff} is more than 1000, the precision has a high chance to reach above 70 to 80%, for L/5 long-range contacts.

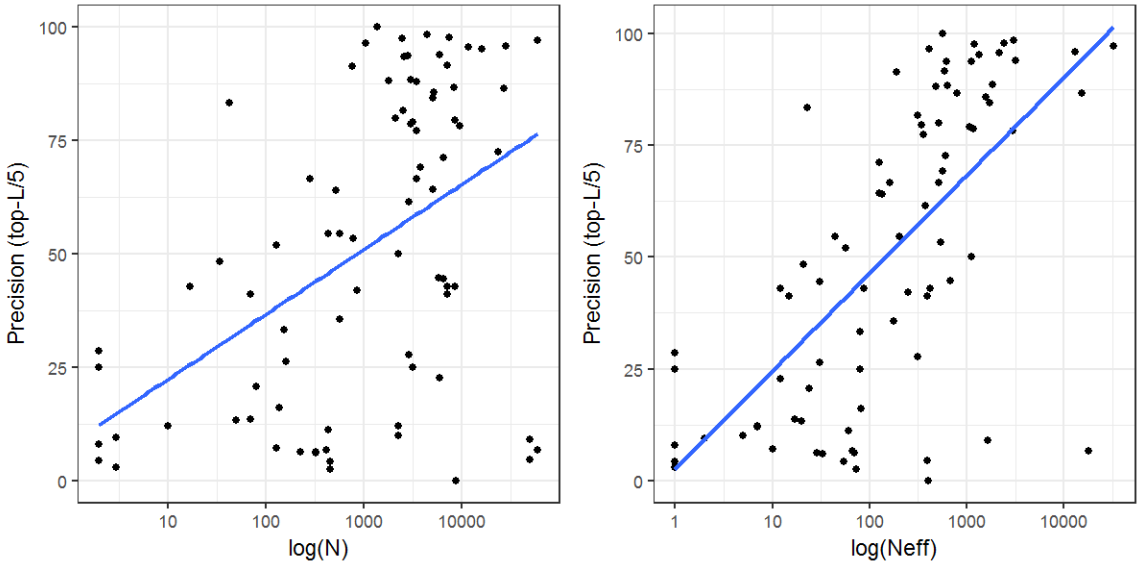

Figure 6.2: The precision of top L/5 long-range contacts predicted by MULITCOM-CONSTRUCT is plotted against the logarithm of number of sequences (N) in the alignments generated for the whole targets (left) and the logarithm of number of effective sequences (N_{eff}) calculated for the domains (right) on the CASP12 dataset. The Pearson's correlation coefficients of the precision with log(N) and log(N_{eff}) are 0.47 and 0.66, respectively.

However, there are some exceptional cases where the contact prediction precision is very low even though the number of sequences in the multiple sequence alignment is high. MULTICOM-

CONSTRUCT's contact precision for top L/5 long-range contacts is only 4.6% for the domain T0898-D1, whose number of sequences in the alignment is very high (∼50K) and the number of effective sequences is 389. When checking the quality of coevolution-based features of this domain, we observed that all individual coevolution-based features also had low-quality contact predictions. However, in general, 389 effective sequences should be sufficient to produce coevolution-based features and final contact predictions of decent quality. After checking the sequence alignment of this domain, we find that most of the sequences have gaps for the first domain (i.e. T0898-D1) and cover only the second domain of the target, such that the N_{eff} for the second domain is much higher, 1648. Moreover, although the multiple sequence alignment has many sequences, most sequences are extremely short, having only around 30 valid residues (non-gaps), and are not useful for predicting long-range contacts with sequence separation >= 24. To verify our observation through N_{eff} calculations, we modified our program to calculate N_{eff} so that aligned gaps were also considered as a match (gap was considered as 21st amino acid) and calculated new N_{eff}. For this domain, such a gap-considered N_{eff} is just 2, suggesting that the poor coverage is the cause of the poor contact prediction. Another exceptional case is MULTICOM-CONSTRUCT's precision for top L/5 long-range contacts is only 7% for the domain T0893-D1, although the multiple sequence alignment generated with 75% coverage threshold has 63,308 sequences with N_{eff} of 17,939. For this domain, all standard coevolution-based features also have poor predictions. We suspect one reason for the low contact precision is the unusual shape of the domain as its tertiary structure consists of just two long helices side by side, whereas the other domain (T0893-D2) of regular shape in the same target has a much higher N_{eff} resulting in long-range contact predictions of 97% precision. These exceptions suggest that, sometimes, coevolution-based contact prediction methods can fail to produce accurate contacts even in the presence of a large number of sequences in the alignments, possibly because many of the sequences in the alignment are false positive homologous sequences or do not align well with target domains. Therefore, in addition to alignment depth as measured by number of (effective) sequences, alignment quality needs to be considered for assessing the accuracy of co-evolution-based contact prediction.

6.4.5 Impact of alignment parameters on the quality and depth of multiple sequence alignments

Our alignment generation algorithm gradually switches to pick lower quality multiple sequence alignments when high-coverage and highly homologous sequences cannot be found. For deciding when to

use a lower quality alignment, we set a threshold of minimum 2.5L sequences in the alignment. We run HHblits with three pre-specified coverage options and JackHMMER with six different e-value thresholds. For example, when HHblits search with 75% coverage option produces an alignment having less than 2.5L sequences, we check the output of the search with 68% coverage, and so on. To analyze if these parameters were well tuned, we studied two subsets - (a) all the targets where we used the results of HHblits search with 75% coverage, and (b) all the targets where we used JackHMMER with e-value threshold of $1E^{-40}$. For these two sets of targets, to study how the various parameters influence the quality of the multiple sequence alignment (and ultimately the quality of contact prediction), we generated multiple sequence alignment with all kinds of parameter settings. In other words, for the first subset where we had chosen HHblits alignments with 75% coverage in CASP12 experiment, we regenerated the alignments with all three coverage options (60%, 68%, and 75%) and predicted contacts using the coevolution-based method CCMpred, respectively. For this set, surprisingly, the precision of contacts predicted using the alignments generated with coverage parameter of 60% is slightly higher, on average, than the ones predicted using the coverage parameter of 75%. The average precisions of top L/2 long-range contacts for the three coverage thresholds (60%, 68%, and 75%) are 61.8%, 60.7%, and 58.1% respectively. This is true for both multi-domain and single-domain targets in the dataset, suggesting that only one HHblits search with coverage option of 60% is generally sufficient to generate good results. Similarly, for the second set of targets where we had used JackHMMER with e-value threshold of $1E^{-40}$, we regenerated the alignments with all six e-value thresholds (1, $1E^{-4}$, $1E^{-10}$, $1E^{-20}$, $1E^{-30}$, and $1E^{-40}$) and predicted contacts using the coevolution-based method CCMpred. On this dataset, the best precision is obtained when alignments are selected with less stringent criteria of $1E^{-10}$ or $1E^{-20}$ e-value threshold. While the mean precision for these domains is 61.8% and 61.7% at e-value threshold of $1E^{-30}$and $1E^{-40}$, the precision increases to 63.5% at the threshold of $1E^{-10}$ and $1E^{-20}$. These results suggest that JackHMMER searches with e-value threshold of $1E^{-30}$ and $1E^{-40}$ need not to be run. In addition to these analyses on the contact predictions of CCMpred, we also predicted contacts using FreeContact method and observed similar results confirming our conclusion.

6.4.6 Impact of the convergence of coevolution methods on contact prediction

During our experiment, the coevolution-based tool PSICOV sometime could not converge within several hours, either because there were too few sequences or too many sequences in the alignment

or because the input sequence was long. Hence, we ran three PSICOV jobs with different parameters in parallel and picked the one that finished within the waiting time limit, based on a preferred order. The preferred order for selecting PSICOV predictions was 'd = 0.03' followed by 'r = 0.001' and 'r = 0.01'. To verify if this preference order was effective, from the dataset of all the targets for which native structures were available for us, we selected the targets for which a multiple sequence alignment with at least 5 sequences could be generated and for which all three PSICOV jobs converged without any time limit constraint, resulting in a dataset of 60 domains. On this dataset, the mean precision of top L/5 long-range contacts for the options 'd=0.03', 'r=0.001', and 'r=0.01' are 35.4%, 33.3%, and 18.1%, respectively. The relatively higher precision of the option 'd=0.03' and much lower precision of the option 'r=0.01' validates that our preference order is fine.

Further, to check how much accuracy was lost due to the five-hour time limit, from the above set of 60 domains, we selected the domains for which we could not select the first PSICOV job (with d=0.03 option) because of the time limit and had instead selected the second PSICOV job (with r = 0.001 option). This resulted in a set of 10 domains for which the mean precision of top L/5 and L/2 long-range contacts were 57.5% and 41% when the contacts were predicted with the 'r=0.001' option. However, had we waited for long enough to let the first set of jobs finish for these targets, the mean precision would have increased to 64.9% and 46.9% for top L/5 and L/2 contacts respectively.

Overall, the experiments show that generating reliable multiple sequence alignments is not a straightforward process. The definition of 'a useful alignment' also depends upon the coevolution-based method used to predict contacts from the alignment. While some of these methods are resource expensive and take longer to run, other methods are relatively fast and are almost independent of the alignment size and length of the protein sequence. Hardware resources and the waiting time limit available for co-evolution feature generation can influence the decision to generate and pick the best alignments. In general, coevolution-based methods take longer to run if the size of alignment (number of sequences in alignment) is big. In some case, CCMpred can run on CPUs for more than a day and PSICOV can run for days. If the hardware resources are limited, it is appropriate to attempt to obtain a reasonable, but less extensive alignment before running these tools. For instance, if HHBlits coverage option of 75% produces 90K sequences, it may be appropriate to increase the coverage threshold to a higher value like 80% to obtain an alignment of smaller size for which the coevolution-based methods can make predictions within a time limit.

6.4.7 Three-dimensional model reconstruction using the predicted contacts

The primary objective of predicting contacts is to use them for three-dimensional structure prediction. In this context, with the contacts predicted by MULTICOM-CONSTRUCT, we built three-dimensional models using our fragment-free *ab initio* folding tool CONFOLD 1.0 [25] to study the usefulness of the predicted contacts. CONFOLD is guided by predicted contacts and secondary structures only, and hence is a good method to build models to study the independent value of the predicted contacts. Using CONFOLD, we built five models for each target in the CASP12 dataset with five sets of contacts - top 0.8L, 1.0L, 2.0L, 3.0L, and 4.0L contacts, without removing short-range or medium-range contacts. To be consistent with other similar works, we built models for the whole target sequence first, without using any knowledge of domains, and then evaluated the predicted models against structural domains. Furthermore, since the number of contacts selected to build models greatly influences the quality of the reconstructed models, we selected 'best of five' models for our analysis. Our reconstruction results, shows that in general, predicted contacts and secondary structures alone could recover the folds of 15 out of the 87 domains, i.e. with TM-score [98] greater than 0.5. We investigated structural domains for which the accuracy of the models was low, and found that many of them are from multi-domain proteins, which are hard for all *ab initio* methods to fold as whole. This suggests that dividing multi-domain proteins into individual domains before folding them with predicted contacts is desirable. For each of the structural domains, we also studied the relationship between the best reconstructed models and the quality of the contact sets selected for the reconstruction. The Pearson's correlation coefficient between the TM-score of the reconstructed models and precision of long-range, medium-range, and short-range contacts are 0.60, 0.42, and 0.34, respectively, indicating long-range contacts are most useful for tertiary structure modeling. We also find that the proportion of the number of long-range, medium-range, and short-range contacts in the native structures is more similar to the proportion of the contacts that were used to build the best models, suggesting that contact-selection i.e. the number of short-range, medium-range, and long-range contacts to select for building models, is important for accurate reconstruction.

As an example, we discuss the reconstruction of a free-modeling domain T0900-D1. T0900-D1 consisting of 102 residues is a complicated beta-sheet domain having 194 long-range, 31 medium-range, and 27 short-range contacts. Of the five sets of contacts selected for reconstruction (0.8L, 1L, 2L, 3L, and 4L), the second set of top 1L contacts generated best models for this domain. This

top 1L set of 60 long-range, 30 medium-range, and 13 short-range contacts generated the top model with 0.43 TM-score, almost recovering the fold of the protein. Despite predicted contacts being very precise (i.e. top L/5 precision of 95% and top L precision of 60%) for this domain, the less accurate reconstruction can be attributed to the poor distribution of predicted contacts used to build the models (see **Figure 6.3(A)** and **(B)**). The correctly predicted contacts only cover a portion of the structure of this domain. In a different experiment, we reconstructed this domain using all true contacts and obtained a model with 0.9 TM-score and 1.4 Å RMSD, which is near native. These examples suggest that the gap between the reconstruction accuracy of using true contacts and that of using only predicted contacts alone (i.e. without using other information like structural templates or fragments), is still wide and the contact-based protein folding requires more research.

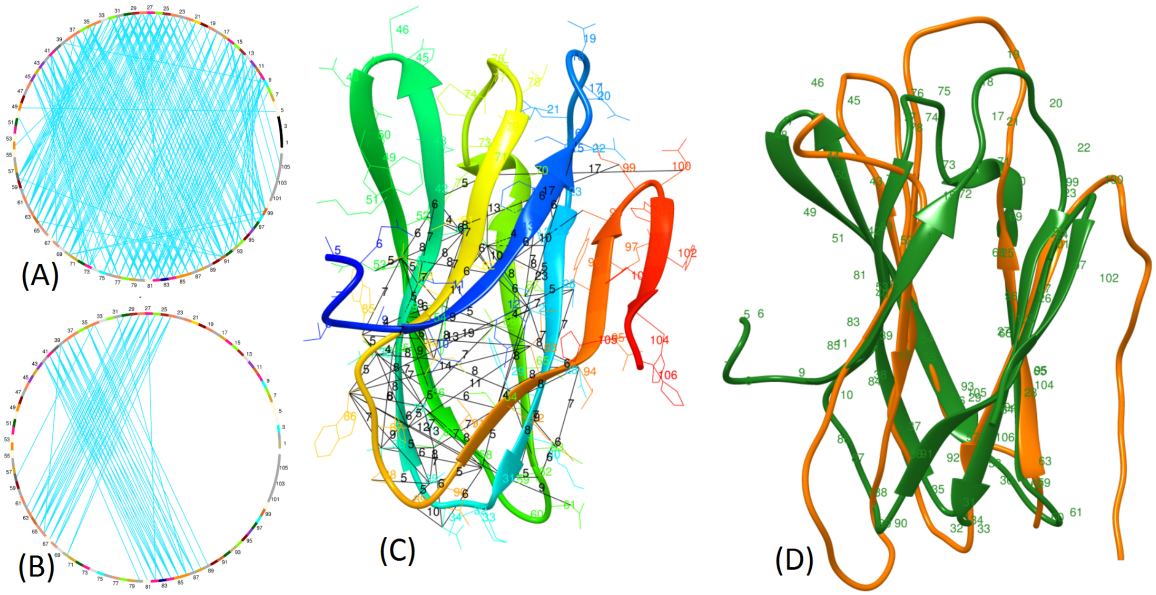

Figure 6.3: Visualization of the top L contacts predicted using MULTICOM-CONSTRUCT and reconstructed model for the domain T0900-D1. Chord diagram for the long-range contacts in the native structure are shown in **(A)** and the top L contacts predicted by MULTICOM-CONSTRUCT shown in **(B)**. MULTICOM-CONSTRUCT predicted contacts are highlighted in the native structure with actual distances between the residues shown in black **(C)** and the reconstructed structure (in orange) superimposed with the native structure (in green) is shown in **(D)**.

Chapter 7

Ab initio protein structure prediction using DNCON2, ConEVA, and CONFOLD

7.1 Introduction

The first step in predicting three-dimensional models for a protein sequence is to check if there are homologous structural templates, by searching the input sequence against existing structural template databases. If we are lucky, we will find at least one good homologous template, which is not usually the case. If we do not find homologous structural templates for our input sequence, ab initio structure prediction will be the default choice. The first step for ab initio protein structure prediction is to predict contacts. DNCON2 can be a default choice for contact prediction as it is demonstrated to outperform other state-of-the-art methods like MetaPSICOV [16] and Raptor-X [19] (see **Chapter 2**). The second step is contact assessment. During assessment, predicted contacts may be compared using Jaccard similarity matrices and visualized using methods like chord diagrams. Our ConEVA web-server toolkit [27] is ideal for such assessments (see **Chapter 3**). The final step is to build three-dimensional models using the predicted contacts, for which the CONFOLD method [25] can be used. CONFOLD accepts predicted contacts and predicted secondary structures as input and delivers top five predicted models (see **Chapter 3**). In this chapter, we discuss how DNCON2, ConEVA, and CONFOLD can be utilized for ab initio protein structure prediction.

7.2 DNCON2 for contact prediction

The DNCON2 web-server at http://sysbio.rnet.missouri.edu/dncon2/ requires as input (a) the sequence, (b) email address where the results are to be sent to, and (c) job name (optional). These three parameters can be supplied as input to the homepage of DNCON2 (see **Figure 7.1**). Once prediction is complete, the web-server sends out an email with the predicted contacts in CASP RR format in the body of the email along with a contact-map visualization attached along. The description of RR format is at http://predictioncenter.org/casprol/index.cgi?page=format#RR. In each contact row, the first two numbers are residue numbers of the pair of residues predicted as a contact, and the last number in the fifth column is the confidence score of prediction with a score of 1.0 being the most confident prediction. Among the few header rows beginning with 'REMARK', three rows are of special interest – (a) Number of sequences in the alignment, (b) Effective number of sequences in the alignment, and (c) Alignment generated. The remark row on the number of sequences in the alignment informs the size of the multiple sequence alignment generated, which is used by the coevolution-based feature generation tools. Similarly, the remark row on the number of effective sequences informs the effective size of the multiple sequence alignment. Empirically, if this number is at least a thousand, the fold of the protein structure can generally be recovered from the predicted contacts. The last of the three remark rows of interest (alignment method) informs if HHblits (hhb) [43] or JackHMMER (jhm) [44] was used to generate alignments. While alignments generated with HHblits with high coverage thresholds generally deliver accurate contacts, contacts predicted using higher e-value thresholds using JackHMMER do not guarantee accurate contacts. Finally, besides these numerical assessments, visual appearance of the contact map (attached in the email) also provides an intuition regarding the quality of predicted contacts. Contact maps that have patterns are generally more accurate than those that do not.

7.3 ConEVA for contact assessment

To access predicted contacts using ConEVA, the predicted contacts RR file can either be uploaded or copied into the text field at the ConEVA homepage at http://iris.rnet.missouri.edu/coneva/ (see **Figure 7.2**). ConEVA then presents various contact counts – proportions of short-, medium-, and long-range contacts, and number of top L/5 and L/2 contacts. L is the length of the sequence. These proportions are useful to study if the predicted contacts are well distributed. For instance, if a method predicts only short-range and medium-range contacts with high confidence, the counts

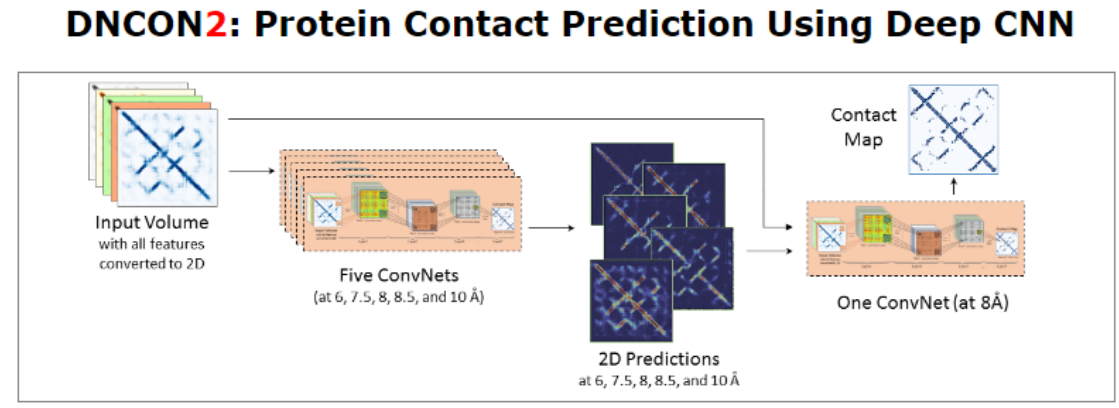

Figure 7.1: A screenshot of DNCON2 web-server at http://iris.rnet.missouri.edu/dncon2/.

of long-range contacts will be very low. When no true structure exists, the only analysis we can perform is visualizations to check the proportion of contact types and ensure a good coverage. Visualizing three-dimensional information in lower dimensions is challenging, but if we are interested in a particular aspect of the data, simpler visualizations in lower dimensions can be easy and yet effective. Next, the visualization using chord diagrams, contact maps, and coordination numbers qualitatively present the distribution of contacts. Generally, for accurate model reconstruction using predicted contacts, we desire contacts to be well distributed. Similarly, when multiple methods are used to predict contacts, ConEVA plots Jaccard similarity matrices showing the similarity between the various predicted contacts. High similarity between predicted contacts from multiple state-of-the-art prediction methods usually suggest accurate contacts (see **Chapter 3**).

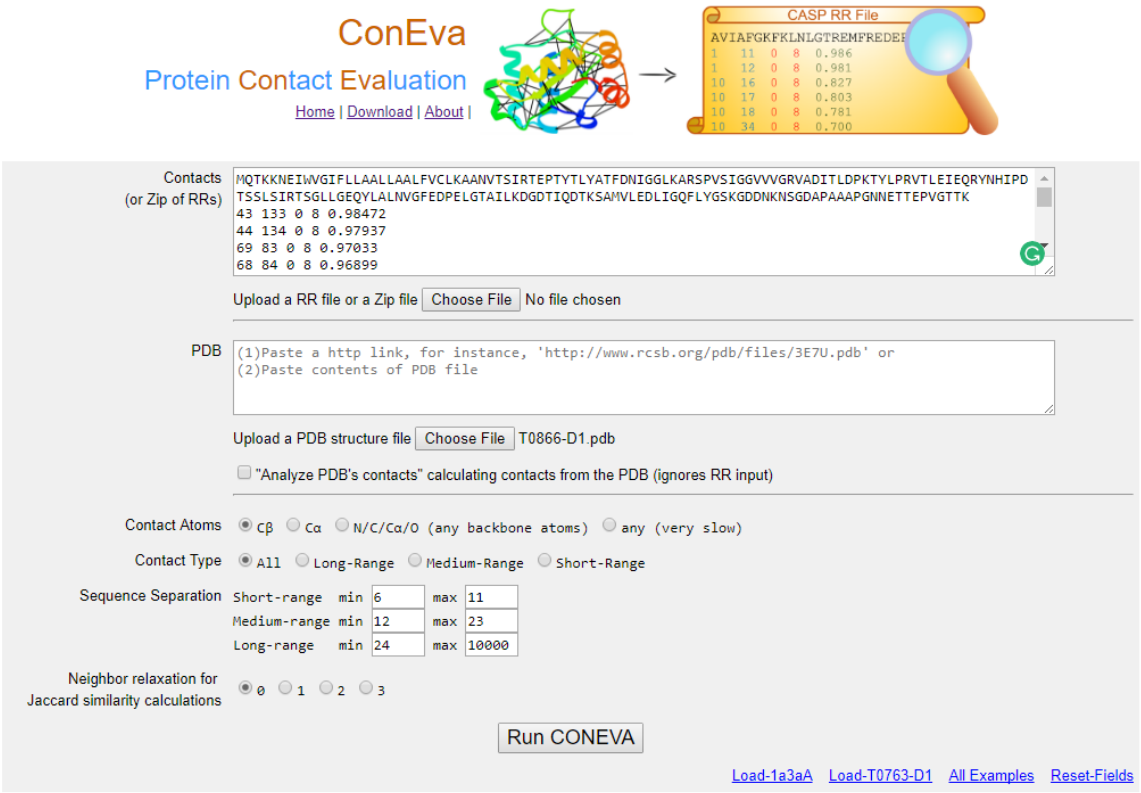

Figure 7.2: A screenshot of ConEVA web-server at http://iris.rnet.missouri.edu/coneva/.

7.4 CONFOLD for building models

To build 3D models for a given input sequence, we need to decide the number of contacts to use. When reconstructing using true contacts, we know that this number must be at least 8% of the native contacts. For predicted contacts, however, the number of contacts that may be used (to

build models) depends on various factors - (a) contact prediction method, (b) model building tool, (c) whether or not additional information is used for modeling, and also (d) the protein structure's reconstruct-ability. Generally, it is found that when short-, medium-, and long-range contacts are used, top L to top 2L contacts can be used for building models.

After predicting secondary structures (in FASTA format) using SCRATCH suite [41], the next step for ab initio structure prediction is to supply these predictions for reconstruction using the CONFOLD web-server at http://protein.rnet.missouri.edu/confold/. The input sequence, predicted contacts and secondary structures can be pasted into the text fields in the CONFOLD web-server along with email address and job name (see **Figure 7.3**). Before submitting the job, it is important to select the appropriate number of contacts to be used; 2L contacts can be a default choice. Since the CONFOLD web-server, by default, does not remove local contacts, i.e. contacts closer than six residues, it is important to remove these local contacts and sort the predicted contacts based on confidence score (highest confidence on top) before submitting to CONFOLD. This can be done through scripting or using Microsoft Excel. Next, running the second stage of CONFOLD with 'sheet detect and contact filter' should be the default choice for second stage modeling. With these default parameters, once CONFOLD receives the inputs, it builds 20 models and sends out the top five models through the email address supplied as input.

7.5 Example

As an example, here we provide the steps to predict the structure for the CASP12 free-modeling protein target T0866. It is a single domain protein with 183 residues. The corresponding PDB entry for this target is a beta-barrel shaped membrane protein '5UW2'.

1. Supply the input sequence to the DNCON2 web-server with your email address and job id, as shown in **Figure 7.1**. Wait for a few hours to receive an email from the DNCON2 web-server.

2. Once the contact prediction email is received, observe the number of sequences in the alignment, number of effective sequences, and the method used to generate the alignment. For this sequence, the number of sequences and the effective number of sequences is 4016 and 908 respectively, suggesting that the quality of contacts is good. Also, the remark row on alignment method inform that HHBlits with 60% coverage was used to generate the alignment. This further assures us that the accuracy of predicted contacts is possibly high. Furthermore, the contact map image (received as an attachment in the email) has visible patterns.

3. Save the body of the email to a file, say 'contacts.rr' and upload it to the ConEVA server for assessment (see **Figure 7.2**). Observing the chord diagrams, we can guess that there are a lot of contacting pairs clustered in various groups suggesting that the structure has many beta sheets. Similarly, from the coordination number visualization of top 2L contacts, it can be observed that no contacts are predicted for the last 30 residues.

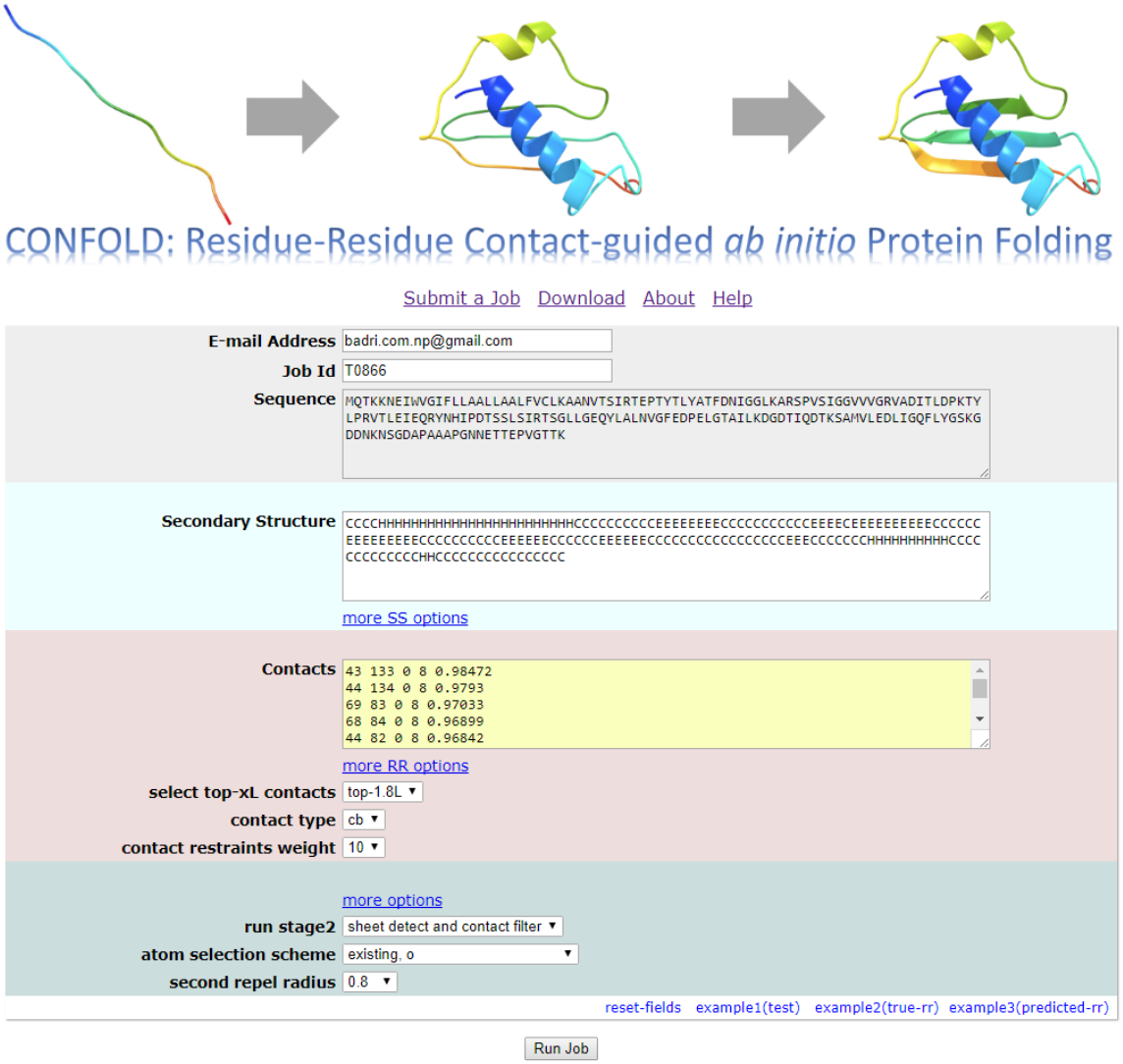

Figure 7.3: A screenshot of CONFOLD web-server at http://protein.rnet.missouri.edu/confold/.

4. Next, predict the secondary structures for this sequence using the SCRATCH suite available for download at http://download.igb.uci.edu/#sspro. The secondary structure prediction (see below), shows that this protein has 36 helical residues and 45 strand residues.

```
>T0866
CCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCEEEEEEEECCCC
CCCCCCCEEEECEEEEEEEEECCCCCCEEEEEEEECCCCCCCCCCEEE
```

```
EEECCCCCCCEEEEEECCCCCCCCCCCCCCCCCEEECCCCCCCCHHHHHHHH
HHCCCCCCCCCCCCCHHCCCCCCCCCCCCCCCC
```

5. **(Option 1).** Next, we supply the predicted contacts and the predicted secondary structure to the CONFOLD web-server for structure modeling. Using Microsoft Excel or any scripting language, remove the local contacts, i.e. pairs with sequence separation less than size residues. Then, sort the contact rows by confidence values, with most confident contacts on top. Paste the input sequence, predicted contact rows, and predicted secondary structure into the CONFOLD text fields and submit the job (see **Figure 7.3**). For this protein, we can select top-1.8L contacts to build models. After modeling (which takes around two hours), CONFOLD sends out the top five models to the email address supplied.

```
PFRMAT RR
TARGET 20170712135143web T0866 Test
AUTHOR DNCON2
METHOD DNCON2
REMARK Number of sequences in the alignment = 4016
REMARK Effective number of sequences in the alignment = 907.7
REMARK Alignment generated using hhb cov60.aln
MODEL  1
MQTKKNEIWVGIFLLAALLAALFVCLKAANVTSIRTEPTYTLYATFDNIG
GLKARSPVSIGGVVVGRVADITLDPKTYLPRVTLEIEQRYNHIPDTSSLS
IRTSGLLGEQYLALNVGFEDPELGTAILKDGDTIQDTKSAMVLEDLIGQF
LYGSKGDDNKNSGDAPAAAPGNNETTEPVGTTK
1 2 0 8 0.34929
1 3 0 8 0.31871
1 4 0 8 0.09939
1 5 0 8 0.05254
1 6 0 8 0.04175
1 7 0 8 0.02262
1 8 0 8 0.01744
1 9 0 8 0.00914
1 10 0 8 0.00503
1 11 0 8 0.00335
...
```

6. **(Option 2).** The key challenge in using the CONFOLD web-server to build protein models is deciding the number of contacts to use. In the newer version, CONFOLD2.0, available at https://github.com/multicom-toolbox/CONFOLD2, this is not an issue. CONFOLD2.0 explores the fold space captured by contacts by building models for various contact selections, clusters the predictions and automatically selects top five models. Since, CONFOLD2.0 is resource intensive (it needs to explore the fold space) it is currently available only as a downloadable tool. If 40 CPUs are used, CONFOLD2.0 runs as fast as original CONFOLD method. For building models using CONFOLD2.0, download the tool and follow the instructions to in-

stall it. For running CONFOLD2.0, the contacts predicted by DNCON2 need not be filtered or sorted. It takes three parameters as input – (a) predicted contacts, (b) predicted secondary structures in SCRATCH suite's FASTA format, and (c) output directory. CONFOLD2.0 delivers top five predicted models in the output directory after all jobs are complete.

7. Since the native structure for this target is available at http://predictioncenter.org, the next step is to evaluate predicted contacts and predicted models. For evaluating predicted contacts, the native structure and the predicted contacts can be uploaded to the ConEVA web-server, we can find that the precision of top L/10 and top L/5 contacts is 100% and 90.5% respectively. Alternately, the local version of ConEVA can be downloaded from https://github.com/multicom-toolbox/ConEVA/ to evaluate the contacts. Similarly, upon evaluating the predicted models using the TM-score program [98], we find that the TM-score of the top-one model is 0.6, suggesting that the fold is recovered using the predicted contacts.

Chapter 8

Conclusion and future work

8.1 Introduction

DNCON2, CONEVA, and CONFOLD, deliver a novel and promising framework to solve the three most important sub-problems in contact-driven ab initio protein structure prediction - contact prediction, contact assessment, and accurate three-dimensional modeling. The performance of our contact prediction method, DNCON2, on the hard datasets of CASP 10, 11, and 12 free-modeling targets shows that the existing state-of-the-art for protein contact prediction can be significantly improved. Our study also shows that convolutional neural networks (CNNs) are well-suited for the protein contact prediction problem. Similarly, for contact assessment, our CONEVA web-server allows us to assess predicted contacts even in the absence of native structure and to comprehensively evaluate contacts when a native structure is available. The CONEVA web-server also allows one to study the contacts in a protein structure to find how many and what types of contacts it has. Finally, for the third sub-problem, our CONFOLD method has demonstrated state-of-the-art reconstruction accuracy. The precision of DNCON2 and the reconstruction accuracy of CONFOLD along with its speed are the backbones of this novel framework towards solving the long-standing protein structure prediction problem.

8.2 Protein Contact Prediction

The improved performance of DNCON2 can be attributed to the following – (a) high quality multiple sequence alignments, (b) inclusion of short- and medium-range contacts into training, (c) two-level approach to prediction, (d) use of the state-of-the-art optimization and activation functions, and (e) a novel deep learning architecture that allows each filter in a convolutional layer to access all the input features. Out of these, the two most of important factors are the quality of multiple sequence alignments and the use of convolutional neural networks. The following are some future works that will possibly improve the precision of DNCON2 predictions.

8.2.1 Improving the quality of multiple sequence alignments

The two major criteria for including sequences into a multiple sequence alignment are coverage and similarity (e-value threshold). Currently, DNCON2 uses two tools, HHblits and JackHMMER, to obtain optimal number of sequences in the alignment. A major limitation of the current algorithm is that it, by default, does not consider the fact that the proteins can be multi-domain. An approach as simple as running the whole alignment generation algorithm two times, with the second run focusing on the region for which alignments could not be generated, could improve the overall performance for multi-domain proteins. One challenge, however, can be to merge the alignments in the two stages. Finally, from our experience of generating alignments, we believe that there must be a way to use only one tool, say JackHMMER, with appropriate parameters integrated into the alignment search process to generate appropriate size of alignments. Developing such a method will require significant amount of study of the multiple sequence alignments and the algorithms used in HHblits and JackHMMER.

8.2.2 Improving the CNN block diagram and architecture

We believe that two technologies that will significantly improve the current performance of DNCON2 are – boosting and residual neural networks. Boosting, in particular, could be highly effective because DNCON2 uses one protein as one dataset, which allows boosting to easily separate proteins into various ranges of prediction difficulty. Using residual neural networks will also possibly improve the overall performance, provided that we have sufficient GPU resources to train deeper CNNs.

8.2.3 Improving overall contact prediction

Training DNCON2 using much larger data set, we believe, will improve the performance of DNCON2. A challenge in increasing the size of the dataset, in current implementation, is memory requirement. Currently, DNCON2 splits the training data into three groups and performs training using one group at a time. To train using larger datasets, the current implementation needs to be changed to a more general version, independent of the training data size. In addition, studying the interplay of CPU and GPU memory and effectively utilizing the GPU resources should improve the training time.

8.3 Protein 3D modeling

The improved version of CONFOLD, at https://github.com/multicom-toolbox/CONFOLD2, addresses the major limitation, i.e. generating top five models instead of just predicting a pool of models. The reconstruction accuracy of CONFOLD may not be easy to improve further. However, one interesting direction for improving CONFOLD is to integrate template-modeling and contact-driven modeling. The idea we propose, as a future work, is to develop a mechanism to feed to CONFOLD, templates (or template fragments) along with predicted contacts, so that it already knows the structure for a region in the input sequence. In other words, if we have already found a structural template for a region in the sequence and have contacts predicted for the rest of the sequence (with overlap between template and predicted contacts), we may supply both to CONFOLD and let CONFOLD use the template and the predicted contacts to generate a final model. This may be achieved by extracting pairwise distance restraints from the template and using them as restraints along with contact restraints.

8.4 Ab initio structure prediction

The current settings of our three tools – DNCON2, CONEVA, and CONFOLD – perform best for single domain proteins. This is because both the methods DNCON2 and CONFOLD deliver better performance on single domain proteins. One obvious improvement to improve overall ab initio structure prediction is to use domain boundary prediction methods to split the input sequence into domains. The next improvement we propose, although it may require quite some study, is to predict contacts at distance thresholds other than 8Å (say 9Å or 10Å) and use them as additional restraints to build models.

Bibliography

[1] K. A. Dill and J. L. MacCallum. The Protein-Folding Problem, 50 Years On. *Science*, 338(6110):1042–1046, 2012.

[2] Felix Simkovic, Sergey Ovchinnikov, David Baker, Daniel J. Rigden, Dror R. O., Kruse A. C., Blacklow S. C., Russel D., Sali A., Zeth K., Sheldrick G. M., Usón I., Montelione G. T., Baker D., and Baker D. Applications of contact predictions to structural biology. *IUCrJ*, 4(3):291–300, may 2017.

[3] Marco Vassura, Luciano Margara, Pietro Di lena, Filippo Medri, Piero Fariselli, and Rita Casadio. FT-COMAR: Fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics*, 24(10):1313–1315, may 2008.

[4] Jose M Duarte, Rajagopal Sathyapriya, Henning Stehr, Ioannis Filippis, and Michael Lappe. Optimal contact definition for reconstruction of contact maps. *BMC bioinformatics*, 11(1):283, jan 2010.

[5] Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997.

[6] Marco Vassura, Luciano Margara, Pietro Di Lena, Filippo Medri, Piero Fariselli, and Rita Casadio. Reconstruction of 3D structures from protein contact maps. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 5(3):357–67, 2008.

[7] Marco Vassura, Pietro Di Lena, Luciano Margara, Maria Mirto, Giovanni Aloisio, Piero Fariselli, Rita Casadio, Giovanni Alosio, Piero Fariselli, and Rita Casadio. Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3d structure. *BioData mining*, 2011.

[8] R. Sathyapriya, Jose M. Duarte, Henning Stehr, Ioannis Filippis, and Michael Lappe. Defining an essence of structure determining residue contacts in proteins. *PLoS computational biology*, 5(12):e1000584, dec 2009.

[9] M Niggemann and B Steipe. Exploring local and non-local interactions for protein stability by structural motif engineering. *Journal of molecular biology*, 296(1):181–95, 2000.

[10] Bohdan Monastyrskyy, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych. Evaluation of residue-residue contact predictions in CASP9. *Proteins*, 79 Suppl 1(Suppl 10):119–25, jan 2011.

[11] Bohdan Monastyrskyy, Daniel D'Andrea, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych. Evaluation of residue-residue contact prediction in CASP10. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):138–153, jun 2014.

[12] Jesse Eickholt and Jianlin Cheng. A study and benchmark of DNcon: a method for protein residue-residue contact prediction using deep networks. *BMC bioinformatics*, 14 Suppl 1(Suppl 14):S12, jan 2013.

[13] Jesse Eickholt and Jianlin Cheng. Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics (Oxford, England)*, 28(23):3066–72, dec 2012.

[14] Jianlin Cheng and Pierre Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC bioinformatics*, 8(1):113, jan 2007.

[15] Pietro Di lena, Ken Nagata, and Pierre Baldi. Deep architectures for protein contact map prediction. *Bioinformatics*, 28(19):2449–2457, 2012.

[16] David T. Jones, Tanya Singh, Tomasz Kosciolek, and Stuart Tetchner. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):btu791, 2014.

[17] Sergey Ovchinnikov, David E. Kim, Ray Yu-Ruei Ruei Wang, Yuan Liu, Frank Dimaio, and David Baker. Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta, sep 2016.

[18] Marcin J. Skwark, Daniele Raimondi, Mirco Michel, and Arne Elofsson. Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Computational Biology*, 10(11), 2014.

[19] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*, 13(1):1–34, 01 2017.

[20] Stefan Seemayer, Markus Gruber, and Johannes Söding. CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, 30(21):3128–3130, 2014.

[21] David T. Jones, Daniel W A Buchan, Domenico Cozzetto, and Massimiliano Pontil. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–90, jan 2012.

[22] László Kaján, Thomas a Hopf, Matúš Kalaš, Debora S Marks, and Burkhard Rost. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC bioinformatics*, 15:85, 2014.

[23] Carol A Rohl, Charlie E M Strauss, Kira M S Misura, and David Baker. Protein structure prediction using Rosetta. *Methods in enzymology*, 383(null):66–93, jan 2004.

[24] Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas a Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, jan 2011.

[25] Badri Adhikari, Debswapna Bhattacharya, Renzhi Cao, and Jianlin Cheng. CONFOLD: Residue-residue contact-guided ab initio protein folding. *Proteins*, 83(8):1436–49, 2015.

[26] Mirco Michel, Sikander Hayat, Marcin J Skwark, Chris Sander, Debora S Marks, and Arne Elofsson. Pconsfold: improved contact predictions improve protein models. *Bioinformatics*, 30(17):i482–i488, 2014.

[27] Badri Adhikari, Jackson Nowotny, Debswapna Bhattacharya, Jie Hou, and Jianlin Cheng. Coneva: a toolbox for comprehensive assessment of protein contacts. *BMC Bioinformatics*, 17(1):517, 2016.

[28] David T. Jones. Predicting novel protein folds by using FRAGFOLD. *Proteins: Structure, Function and Genetics*, 45(SUPPL. 5):127–132, 2001.

[29] Debora S Marks, Thomas a Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–80, nov 2012.

[30] Mirco Michel, David Menendez Hurtado, Karolis Uziela, and Arne Elofsson. Large-scale structure prediction by improved contact predictions and model quality assessment. *bioRxiv*, page 128231, 2017.

[31] Mahmoud Mabrouk, Tim Werner, Michael Schneider, Ines Putz, and Oliver Brock. Analysis of free modeling predictions by rbo aleph in casp11. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):87–104, 2016.

[32] Wenxuan Zhang, Jianyi Yang, Baoji He, Sara Elizabeth Walker, Hongjiu Zhang, Brandon Govindarajoo, Jouko Virtanen, Zhidong Xue, Hong-Bin Shen, and Yang Zhang. Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):76–86, sep 2016.

[33] Dong Xu and Yang Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, 80(7):1715–35, jul 2012.

[34] Tomasz Kosciolek and David T. Jones. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS ONE*, 9(3), 2014.

[35] Michal J. Pietal, Janusz M. Bujnicki, and Lukasz P. Kozlowski. GDFuzz3D: A method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function. *Bioinformatics*, 2014.

[36] Mirco Michel, Marcin J Skwark, David Menéndez Hurtado, Magnus Ekeberg, and Arne Elofsson. Predicting accurate contacts in thousands of pfam domain families using pconsc3. *Bioinformatics*, 2017.

[37] Marcin J. Skwark, Abbi Abdel-Rehim, and Arne Elofsson. PconsC: Combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, 29(14):1815–1816, 2013.

[38] Lisa N. Kinch, Wenlin Li, R. Dustin Schaeffer, Roland L. Dunbrack, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick V. Grishin. CASP 11 target classification. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):20–33, sep 2016.

[39] Bohdan Monastyrskyy, Daniel D'Andrea, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych. New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins: Structure, Function and Bioinformatics*, 2015.

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

[41] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33(Web Server):W72–W76, jul 2005.

[42] Tomasz Kosciolek and David T. Jones. Accurate contact predictions using covariation techniques and machine learning. *Proteins: Structure, Function and Bioinformatics*, 2015.

[43] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9(2):173–5, feb 2011.

[44] L Steven Johnson, Sean R Eddy, and Elon Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1):431, 2010.

[45] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[46] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[47] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[48] Jianlin Cheng, Zheng Wang, Allison N. Tegge, and Jesse Eickholt. Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):181–184, 2009.

[49] Huiling Zhang, Qingsheng Huang, Zhendong Bei, Yanjie Wei, and Christodoulos A. Floudas. COMSAT: Residue contact prediction of transmembrane proteins based on support vector machines and mixed integer linear programming. *Proteins: Structure, Function and Bioinformatics*, 84(3):332–348, 2016.

[50] Iakes Ezkurdia, Osvaldo Graña, José M G Izarzugaza, Michael L. Tress, Lakes Ezkurdia, Osvaldo Grana, José M G Izarzugaza, and Michael L. Tress. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins*, 77(SUPPL. 9):196–209, jan 2009.

[51] Osvaldo Graña, David Baker, Robert M. MacCallum, Jens Meiler, Marco Punta, Burkhard Rost, Michael L. Tress, and Alfonso Valencia. CASP6 assessment of contact prediction. *Proteins: Structure, Function and Genetics*, 61 Suppl 7(SUPPL. 7):214–24, jan 2005.

[52] José M G Izarzugaza, Osvaldo Graña, Michael L. Tress, Alfonso Valencia, and Neil D. Clarke. Assessment of intramolecular contact predictions for CASP7. *Proteins: Structure, Function and Genetics*, 69 Suppl 8(SUPPL. 8):152–8, jan 2007.

[53] Allison N. Tegge, Zheng Wang, Jesse Eickholt, and Jianlin Cheng. NNcon: Improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Research*, 37(SUPPL. 2):W515–W518, jul 2009.

[54] Osvaldo Graña, Volker A. A. Eyrich, Florencio Pazos, Burkhard Rost, and Alfonso Valencia. EVAcon: A protein contact prediction evaluation service. *Nucleic Acids Research*, 33(SUPPL. 2), 2005.

[55] Corinna Vehlow, Henning Stehr, Matthias Winkelmann, José M. Duarte, Lars Petzold, Juliane Dinse, and Michael Lappe. CMView: Interactive contact map visualization and analysis. *Bioinformatics*, 27(11):1573–1574, 2011.

[56] Frazier N. Baker and Aleksey Porollo. CoeViz: a web-based tool for coevolution analysis of protein residues. *BMC Bioinformatics*, 17(1):119, 2016.

[57] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, jan 2000.

[58] Gregory R Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang H A Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, and Bill Venables. gplots: Various R Programming Tools for Plotting Data. *R package version 2.17.0.*, page 2015, 2015.

[59] Lemon J. Plotrix: a package in the red light district of R. *R-News*, 6(4):8–12, 2006.

[60] Gianluca Pollastri, Pierre Baldi, Pietro Fariselli, and Rita Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins: Structure, Function and Genetics*, 47(2):142–153, 2002.

[61] Jesse Davis and Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine learning – ICML'06*, pages 233–240, 2006.

[62] G. Gilbert. Distance between Sets. *Nature*, 239(5368):174–174, 1972.

[63] L J McGuffin, K Bryson, and D T Jones. The PSIPRED protein structure prediction server. *Bioinformatics (Oxford, England)*, 16(4):404–405, 2000.

[64] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, dec 2004.

[65] David E. E. Kim, Frank Dimaio, Ray Yu-Ruei Wang, Yifan Song, and David Baker. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):208–218, 2014.

[66] Renzhi Cao and Jianlin Cheng. Protein single-model quality assessment by feature-based probability density functions. *Scientific reports*, 6(April):23990, 2016.

[67] Renzhi Cao, Debswapna Bhattacharya, Badri Adhikari, Jilong Li, and Jianlin Cheng. Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics*, 31(12):i116–i123, jun 2015.

[68] Renzhi Cao, Zheng Wang, Yiheng Wang, and Jianlin Cheng. SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC bioinformatics*, 15(1):120, 2014.

[69] Debswapna Bhattacharya and Jianlin Cheng. De novo protein conformational sampling using a probabilistic graphical model. *Scientific reports*, 5:16332, 2015.

[70] Debswapna Bhattacharya, Renzhi Cao, and Jianlin Cheng. UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics (Oxford, England)*, 32(18):btw316, sep 2016.

[71] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, oct 2004.

[72] Piero Fariselli, Osvaldo Olmea, Alfonso Valencia, and Rita Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein engineering*, 14(11):835–843, 2001.

[73] Sitao Wu, Andras Szilagyi, and Yang Zhang. Improving Protein Structure Prediction Using Multiple Sequence-Based Contact Predictions. *Structure*, 19(8):1182–1191, 2011.

[74] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residueresidue interactions across protein interfaces using evolutionary information. *eLife*, 3, 2014.

[75] Zhiyong Wang and Jinbo Xu. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics (Oxford, England)*, 29(13):i266–73, jul 2013.

[76] Todd J Taylor, Hongjun Bai, Chin-Hsien Tai, and Byungkook Lee. Assessment of CASP10 contact-assisted predictions. *Proteins*, 82 Suppl 2:84–97, feb 2014.

[77] Hua Zhang, Tuo Zhang, Ke Chen, Kanaka Durga Kedarisetti, Marcin J Mizianty, Qingbo Bao, Wojciech Stach, and Lukasz Kurgan. Critical assessment of high-throughput standalone methods for secondary structure prediction. *Briefings in bioinformatics*, 12(6):672–688, 2011.

[78] Ke Chen and Lukasz Kurgan. Computational prediction of secondary and supersecondary structures. In *Protein Supersecondary Structures*, pages 63–86. Springer, 2013.

[79] Vladimir Golkov, Marcin J Skwark, Antonij Golkov, Alexey Dosovitskiy, Thomas Brox, Jens Meiler, and Daniel Cremers. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. In *Advances in Neural Information Processing Systems*, pages 4222–4230, 2016.

[80] Christian Cole, Jonathan D Barber, and Geoffrey J Barton. The Jpred 3 secondary structure prediction server. *Nucleic acids research*, 36(suppl 2):W197–W201, 2008.

[81] Eshel Faraggi, Tuo Zhang, Yuedong Yang, Lukasz Kurgan, and Yaoqi Zhou. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of computational chemistry*, 33(3):259–267, 2012.

[82] Jakob Bohr, Henrik Bohr, Søren Brunak, Rodney M J Cotterill, Henrik Fredholm, Benny Lautrup, and Steffen B Petersen. Protein structures from distance inequalities. *Journal of molecular biology*, 231(3):861–869, 1993.

[83] Jorge J Moré and Zhijun Wu. Distance geometry optimization for protein structures. *Journal of Global Optimization*, 15(3):219–234, 1999.

[84] Pietro Di Lena, Marco Vassura, Luciano Margara, Piero Fariselli, and Rita Casadio. On the reconstruction of three-dimensional protein structures from contact maps. *Algorithms*, 2(1):76–92, 2009.

[85] J W Ponder and F M Richards. TINKER molecular modeling package. *J. Comput. Chem*, 8:1016–1024, 1987.

[86] Bogumil M Konopka, Marika Ciombor, Monika Kurczynska, and Malgorzata Kotulska. Automated Procedure for Contact-Map-Based Protein Structure Reconstruction. *The Journal of membrane biology*, 247(5):409–420, 2014.

[87] Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson, and Andrej Sali. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS biology*, 10(1):e1001244, 2012.

[88] Narayanan Eswar, Ben Webb, Marc A Marti-Renom, M S Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*, pages 5.6. 1–5.6. 30, 2006.

[89] AXEL T. Brunger, Paul D Adams, G Marius Clore, Warren L DeLano, Piet Gros, Ralf W Grosse-Kunstleve, J-S S Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, A T Brünger, Paul D Adams, G Marius Clore, Warren L DeLano, Piet Gros, Ralf W Grosse-Kunstleve, J-S S Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, R J Read, L M Rice, T Simonson, G L Warren, AXEL T. Brunger, A T Brünger, Paul D Adams, G Marius Clore, Warren L DeLano, Piet Gros, Ralf W Grosse-Kunstleve, J-S S Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, R J Read, L M Rice, T Simonson, and G L Warren. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta crystallographica. Section D, Biological crystallography*, 54(Pt 5):905–921, sep 1998.

[90] Axel T Brunger. Version 1.2 of the Crystallography and NMR system. *Nature protocols*, 2(11):2728–33, jan 2007.

[91] Ivo Van Walle, Ignace Lasters, and Lode Wyns. SABmarka benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, 21(7):1267–1268, 2005.

[92] F R Salemme. Structural properties of protein β-sheets. *Progress in biophysics and molecular biology*, 42:95–133, 1983.

[93] F.R. Salemme and D.W. Weatherford. Conformational and geometrical properties of β-sheets in proteins. *Journal of Molecular Biology*, 146(1):101–117, feb 1981.

[94] J. Cheng and P. Baldi. Three-stage prediction of protein -sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, 21(Suppl 1):i75–i84, jun 2005.

[95] Malcolm W MacArthur and Janet M Thornton. Influence of proline residues on protein conformation. *Journal of molecular biology*, 218(2):397–412, 1991.

[96] Todd J Taylor, Chin-Hsien Tai, Yuanpeng J Huang, Jeremy Block, Hongjun Bai, Andriy Kryshtafovych, Gaetano T Montelione, and Byungkook Lee. Definition and classification of evaluation units for CASP10. *Proteins*, 82 Suppl 2:14–25, feb 2014.

[97] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[98] Y. Zhang. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, apr 2005.

[99] Jesper Lundström, Leszek Rychlewski, Janusz Bujnicki, and Arne Elofsson. Pcons: A neural-networkbased consensus predictor that improves fold recognition. *Protein Science*, 10(11):2354–2362, 2001.

[100] Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. Scope: Structural classification of proteinsextended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2013.

[101] Jilong Li, Badri Adhikari, and Jianlin Cheng. An improved integration of template-based and template-free protein structure modeling methods and its assessment in casp11. *Protein and peptide letters*, 22(7):586–593, 2015.

[102] Lisa N. Kinch, Wenlin Li, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick V. Grishin. Evaluation of free modeling targets in CASP11 and ROLL, 2016.

VITA

Badri Adhikari is the author of the protein 3D modeling tool, CONFOLD, which has been integrated into many other protein structure prediction methods. Born and raised in Nepal, he obtained his undergraduate degree in Computer Engineering from Advanced College of Engineering and Management, Tribhuvan University. With research interests in machine learning, data mining and bioinformatics, he has published ten papers in journals like Bioinformatics, PLOS ONE, BMC Bioinformatics, and BMC Genomics. He also actively serves as a reviewer for these journals. He has three years of industry experience at the data mining and warehousing companies Verisk Health Inc. and Yomari Inc. He has also taught fundamental computer science courses like Java programming and data structures at Tribhuvan University in Nepal and at Westminster College in Fulton, Missouri.