BIOINFORMATICS

Vol. 29 ISMB/ECCB 2013, pages i266–i273
doi:10.1093/bioinformatics/btt211

Predicting protein contact map using evolutionary and physical constraints by integer programming

Zhiyong Wang and Jinbo Xu*
Toyota Technological Institute at Chicago, 6045 S Kenwood, IL 60637, USA

ABSTRACT

Motivation: Protein contact map describes the pairwise spatial and functional relationship of residues in a protein and contains key information for protein 3D structure prediction. Although studied extensively, it remains challenging to predict contact map using only sequence information. Most existing methods predict the contact map matrix element-by-element, ignoring correlation among contacts and physical feasibility of the whole-contact map. A couple of recent methods predict contact map by using mutual information, taking into consideration contact correlation and enforcing a sparsity restraint, but these methods demand for a very large number of sequence homologs for the protein under consideration and the resultant contact map may be still physically infeasible.

Results: This article presents a novel method PhyCMAP for contact map prediction, integrating both evolutionary and physical restraints by machine learning and integer linear programming. The evolutionary restraints are much more informative than mutual information, and the physical restraints specify more concrete relationship among contacts than the sparsity restraint. As such, our method greatly reduces the solution space of the contact map matrix and, thus, significantly improves prediction accuracy. Experimental results confirm that PhyCMAP outperforms currently popular methods no matter how many sequence homologs are available for the protein under consideration.

Availability: http://raptorx.uchicago.edu.

Contact: jinboxu@gmail.com

1 INTRODUCTION

In this article, we say two residues of a protein are in contact if their Euclidean distance is <8 Å. The distance of two residues can be calculated using C_α or C_β atoms, corresponding to C_α- or C_β-based contacts. A protein contact map is a binary $L \times L$ matrix, where L is the protein length. In this matrix, an element with value 1 indicates the corresponding two residues are in contact; otherwise, they are not in contact. Protein contact map describes the pairwise spatial and functional relationship of residues in a protein. Predicting contact map using sequence information has been an active research topic in recent years partially because contact map is helpful for protein 3D structure prediction (Ortiz *et al.*, 1999; Vassura *et al.*, 2008; Vendruscolo *et al.*, 1997; Wu *et al.*, 2011) and protein model quality assessment (Zhou and Skolnick, 2007). Protein contact map has also been used to study protein structure alignment (Caprara *et al.*, 2004; Wang *et al.*, 2013; Xu *et al.*, 2007).

Many machine-learning methods have been developed for protein contact prediction in the past decades (Fariselli and Casadio, 1999; Göbel *et al.*, 2004; Olmea and Valencia, 1997; Punta and Rost, 2005; Vendruscolo and Domany, 1998; Vullo *et al.*, 2006). For example, SVMSEQ (Wu and Zhang, 2008) uses support vector machines for contact prediction; NNcon (Tegge *et al.*, 2009) uses a recursive neural network; SVMcon (Cheng and Baldi, 2007) also uses support vector machines plus features derived from sequence homologs; Distill (Baú *et al.*, 2006) uses a 2D recursive neural network. Recently, CMAPpro (Di Lena *et al.*, 2012) uses a multi-layer neural network. Although different, these methods are common in that they predict the contact map matrix element-by-element, ignoring the correlation among contacts and also physical feasibility of the whole-contact map (physical constraints are not totally independent of contact correlation). A special type of physical constraint is that a contact map matrix must be sparse, i.e. the number of contacts in a protein is only linear in its length.

Two recent methods [PSICOV (Jones *et al.*, 2012) and Evfold (Morcos *et al.*, 2011)] predict contacts by using only mutual information (MI) derived from sequence homologs and enforcing the aforementioned sparsity constraint. However, both of them demand for a large number (at least several hundreds) of sequence homologs for the protein under prediction. This makes the predicted contacts not useful in protein modeling, as a (globular) protein with many sequence homologs usually has similar templates in PDB; thus, template-based models are of good quality and hard to be further improved using predicted contacts. Conversely, a protein without close templates in PDB, which may require contact prediction, usually has few sequence homologs even if millions of protein sequences are now available. Further, these two methods enforce only a simple sparsity constraint (i.e. the total number of contacts in a protein is small), ignoring many more concrete constraints. To name a few, one residue can have only a small number of contacts, depending on its secondary structure and neighboring residues. The number of contacts between two β-strands is bounded by the strand length.

Astro-Fold (Klepeis and Floudas, 2003) possibly is the first method that applies physical constraints, which implicitly imply the sparsity constraint used by PSICOV and Evfold, to contact map prediction. However, some of the physical constraints are too restrictive and possibly unrealistic. For example, it requires that a residue in one β-strand can only be in contact with a residue in another β-strand. More importantly, Astro-Fold does not take into consideration evolutionary information; thus, it significantly reduces its prediction accuracy.

In this article, we present a novel method PhyCMAP for contact map prediction by integrating both evolutionary and physical constraints using machine learning [i.e. Random Forests (RF)] and integer linear programming (ILP). PhyCMAP first predicts the probability of any two residues forming a contact using evolutionary information (including MI), predicted

*To whom correspondence should be addressed.

secondary structure and distance-dependent statistical potential. PhyCMAP then infers a given number of top contacts based on predicted contact probabilities by enforcing a set of realistic physical constraints on the contact map. These restraints specify more concrete relationship among contacts and also imply the sparsity restraint used by PSICOV and Evfold. By combining both evolutionary and physical constraints, our method greatly reduces the solution space of contact map and leads to much better prediction accuracy. Experimental results confirm that PhyCMAP outperforms currently popular methods no matter how many sequence homologs are available for the protein under prediction.

2 METHODS

Overview. As shown in Figure 1, our method consists of several key components. First, we use RF to predict the contact probability of any two residues based on a few protein features related to these two residues. Then we use an ILP method to select a set of top contacts by maximizing their accumulative probabilities subject to a set of physical constraints. The resultant top contacts form a physically favorable contact map for the protein under consideration.

2.1 Predicting contact probability by Random Forests

We use RF to predict the probability of any two residues forming a contact using the following input features: EPAD (a context-specific distance-based statistical potential) (Zhao and Xu, 2012), PSIBLAST sequence profile (Altschul and Koonin, 1998), secondary structure predicted by PSIPRED (Jones, 1999), pairwise contact score and contrastive MI (CMI) derived from multiple sequence alignment (MSA) of the sequence homologs of the protein under prediction. The latter four features are calculated on the residues in a local window of size 5 centered at the residues under consideration. In total, there are ~300 features for each residue pair. We trained our RF model using the Random Forest package in R (Breiman, 2001; Liaw and Wiener, 2002) and selected the model parameters by 5-fold cross-validation.

EPAD. The context-specific interaction potential of the C_α or C_β atoms of two residues at all the possible distance bins is used as features. The atomic distance is discretized into some bins by 1 Å, and all the distance >15 Å is grouped into a single bin.

Sequence profile. The position-specific mutation scores at residues i and j and their neighboring residues are used.

In addition, a protein contact-based potential CCPC (Tan *et al.*, 2006) and amino acid physic-chemical properties are also used as features of our RF model.

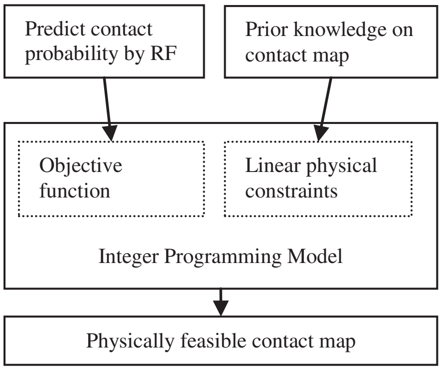
Predict contact probability by RF
Prior knowledge on contact map
Objective function
Linear physical constraints
Integer Programming Model
Physically feasible contact map

Fig. 1. The overview of our approach

Homologous pairwise contact score (HPS). Let i and j denote two residues of the protein under consideration. Let H denote the set of all the sequence homologs. Given an MSA of all the homologs in H, we calculate the homologous pairwise contact score HPS for two residues i and j as follows.

$$HPS(i,j) = \sum_{h\in H} PS_{beta}\left(a_i^h, a_j^h\right) + PS_{helix}\left(a_i^h, a_j^h\right)$$

where a_i^h (a_j^h) denotes the residue in a homolog h aligned to residue $i(j)$ in the query sequence. $PS_{beta}(a_i^h, a_j^h)$ is the probability of two amino acids a_i^h, a_j^h forming a contact in a β-sheet. $PS_{helix}(a_i^h, a_j^h)$ is the probability of two amino acids a_i^h, a_j^h forming a contact connecting two helices. The probability is calculated as follows.

$$PS_{beta}(A,B) = \frac{\text{The number of amino acids (A, B) forming a beta contact}}{\text{Total number of beta contact pairs in training set}}$$

$$PS_{helix}(A,B) = \frac{\text{The number of amino acids (A, B) forming a helix contact}}{\text{Total number of beta contact pairs in training set}}$$

The contrastive mutual information. Let $m_{i,j}$ denote the MI between these two residues i and j, which can be calculated from the MSA of all the sequence homologs. We define the CMI as the MI difference between one residue pair and all of its neighboring pairs, which can be calculated as follows.

$$CMI_{i,j} = \left(m_{i,j} - m_{i-1,j}\right)^2 + \left(m_{i,j} - m_{i+1,j}\right)^2 + \left(m_{i,j} - m_{i,j-1}\right)^2 + \left(m_{i,j} - m_{i,j+1}\right)^2$$

The CMI measures how the co-mutation strength of one residue pair differs from its neighboring pairs. By using CMI instead of MI, we can remove the background bias of MI in a local region, as shown in Figure 2. In the case that there are only a small number of sequence homologs available, some conserved positions, which usually have entropy <0.3, may have very low MI, which may result in artificially high CMI. To avoid this, we directly set the CMI of these positions to 0.

2.2 The integer linear programming method

The variables. Let i and j denote residue positions and L the protein length. Let u and v index secondary structure segments of a protein. Let $Begin(u)$ and $End(u)$ denote the first and last residues of the segment u and $SSeg(u)$ the set $\{i|Begin(u) \le i \le End(u)\}$. Let $SStype(u)$ denote the secondary structure type of one residue or one segment u. Let $Len(u)$ denote the length of the segment u. We use the binary variables in Table 1.

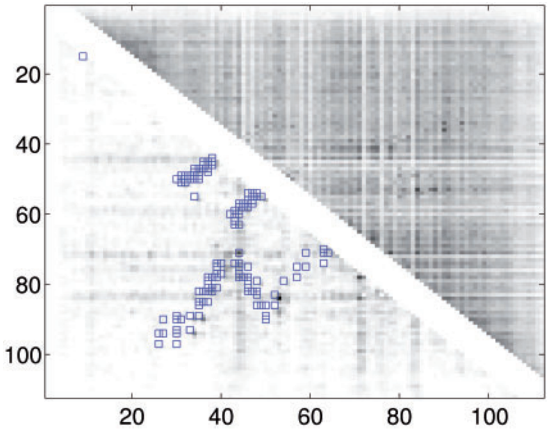

Fig. 2. The CMI (lower triangle) and MI (upper triangle) of protein 1j8bA.The native contact pairs are marked by boxes

Table 1. The binary variables used in the ILP formulation

Variables	Explanations
$X_{i,j}$	Equal to 1 if there is a contact between two residues i and j.
$AP_{u,v}$	Equal to 1 if two β-strands u and v form an anti-parallel β-sheet.
$P_{u,v}$	Equal to 1 if two β-strands u and v form a parallel β-sheet.
$S_{u,v}$	Equal to 1 if two β-strands u and v form a β-sheet.
$T_{u,v}$	Equal to 1 if there is an α-bridge between two helices u and v.
R_r	A non-negative integral relaxation variable of the r^{th} soft constraint.

Table 2. The empirical values of $b_{s1,s2}$ calculated from the training data

s1,s2	95%	Max
H,H	5	12
H,E	3	10
H,C	4	11
E,H	4	12
E,E	9	13
E,C	6	15
C,H	3	12
C,E	5	12
C,C	6	20

Note: The first column indicates the combination of two secondary structure types: H (α-helix), E (β-strand) or C (coil). Each row contains two statistical values for a pair of secondary structure types. Column '95%': 95% of the secondary structure pairs have the number of contacts smaller than the value in this column; column 'Max': the largest number of contacts.

The objective function. Intuitively, we shall choose those contacts with the highest probability predicted by our RF model, i.e. the objective function shall be the sum of predicted probabilities of the selected contacts. However, the selected contacts shall also minimize the violation of the physical constraints. To enforce this, we use a vector of relaxation variables R to measure the degree to which all the soft constraints are violated. Thus, our objective function is as follows.

$$\max_{X,R} \sum_{j-i\geq 6} (X_{i,j} \times P_{i,j}) - g(R)$$

where $P_{i,j}$ is the contact probability predicted by our RF model for two residues and $g(R) = \sum R_r$ is a linear penalty function where r runs over all the soft constraints. The relaxation variables will be further explained in the following section.

The constraints. We use both soft and hard constraints. There is a single relaxation variable for each group of soft constraint, but the hard constraints are strictly enforced. We penalize the violation of soft constraints by incorporating the relaxation variables to the objective function. The constraints in Groups 1, 2 and 6 are soft constraints. Those in Groups 3, 4, 5 and 7 are hard constraints, some of which are similar to what are used by Astro-Fold (Klepeis and Floudas, 2003).

Group 1. This group of soft constraints bound from above the total number of contacts that can be formed by a single residue i (in a secondary structure type $s1$) with all the other residues in a secondary structure type $s2$.

$$\forall i : SStype(i) = s1, \quad \sum_{j:SStype(j)=s2} X_{i,j} \leq R_1 + b_{s1,s2}$$

where $b_{s1,s2}$ is a constant empirically determined from our training data (Table 2), and R_1 is the relaxation variable.

Group 2. This group of constraints bound the total number of contacts between two strands sharing at least one contact. Let u and v denote two β-strands. We have

$$\sum_{i\in SSeg(v), j\in SSeg(u)} X_{i,j} + R_2 \geq 3 \times S_{u,v} \times \min(Len(u), Len(v))$$

$$\sum_{i\in SSeg(v), j\in SSeg(u)} X_{i,j} \leq 3.3 \times \max(Len(u), Len(v)) + R_3$$

The two constraints are explained in Figure 3 as follows. Figure 3A shows that the total number of contacts between two β-strands diverges into two groups when $\min(Len(u), Len(v)) \geq 9$. One group is due to β-strand pairs forming a β-sheet. The other may be due to incorrectly predicted β-strands or β-strand pairs not in a β-sheet. Figure 3B shows that the total number of contacts between a pair of β-strands has an upper bound proportional to the length of the longer β-strand. As there are points below the skew line in Figure 3A, which indicate that a β-strand pair may have fewer than $3 \times \min(Len(u), Len(v))$ contacts, we add a relaxation variable R_2 to the lower bound constraints in Group 2. Similarly, we use a relaxation variable R_3 for the upper bound constraints.

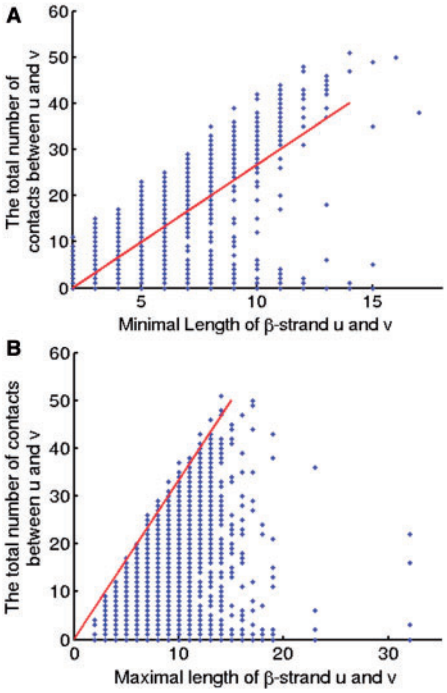

Fig. 3. The skew lines indicate the bounds for the total number of contacts between two β-strands. (**A**) Lower bound; (**B**) upper bound

Group 3. When two strands form an anti-parallel β-sheet, the contacts of neighboring residue pairs shall satisfy the following constraints.

$$X_{i,j} \geq X_{i-1,j+1} + X_{i+1,j-1} - 1$$

where $i, i \pm 1$ are residues in one strand, and $j, j \pm 1$ are residues in the other strand.

Group 4. When there are parallel contacts between two strands, the contacts of neighboring residue pairs should satisfy the following constraints.

$$X_{i,j} \geq X_{i-1,j-1} + X_{i+1,j+1} - 1$$

where $i, i \pm 1$ are residues in one strand, and $j, j \pm 1$ are residues in the other strand.

Group 5. One β-strand u can form β-sheets with up to two other β-strands.

$$\sum_{v:SStype(v)=\text{beta}} S_{u,v} \leq 2$$

Group 6. There is no contact between the start and end residues of a loop segment u.

$$X_{i,j} \leq 0 + R_4, i = Begin(u), j = End(u)$$

In our training set, there are totally ~8000 loop segments, and only 3.4% of them have a contact between the start and end residues. To allow the rare cases, we use a relaxation variable R_4 in the constraints.

Group 7. One residue i cannot have contacts with both j and $j+2$ when j and $j+2$ are in the same α helix.

$$X_{i,j} + X_{i,j+2} \leq 1$$

Group 8. This group of constraints models the relationship among different groups of variables.

$$\begin{aligned} &AP_{u,v} + P_{u,v} = S_{u,v} \\ &X_{i,j} \leq S_{u,v}, \forall i \in SSeg(u), j \in SSeg(v) \\ &\sum_{1 \leq i < j \leq L, j-i \geq 6} X_{i,j} = k \end{aligned}$$

where k is the number of top contacts we want to predict.

Our ILP model is solved by IBM CPLEX academic version V12.1 (CPLEX, 2009).

Training data. It consists of 900 non-redundant protein structures, any two of which share no >25% sequence identity. As there are far fewer contacting residue pairs than non-contacting pairs, we use all the contacting pairs and randomly sample only 20% of the non-contacting pairs as the training data. All the training proteins are selected before CASP10 (the 10th Critical Assessment of Structure Prediction) started in May 2012.

Test data I: CASP10. This set contains 123 CASP10 targets with publicly available native structures. Meanwhile, 36 of them are classified as hard targets because the top half of their submitted models have average TM-score <0.5. When they were just released, most of the CASP10 targets share low sequence identity (<25%) with the proteins in PDB. BLAST indicates that there are only five short CASP10 targets (~50 residues), which have sequence identity slightly >30% with our training proteins.

Test data II: Set600. This set contains 601 proteins randomly extracted from PDB25 (Brenner *et al.*, 2000) and was constructed before CASP10 started. The test proteins have the following properties: (i) they share <25% sequence identity with the training proteins; (iii) all proteins have at least 50 residues and an X-ray structure with resolution better than 1.9 Å; and (iii) all the proteins have at least five residues with predicted secondary structure being α-helix or β-strand.

Both the training set and Set600 are sampled from PDB25 (Wang and Dunbrack, 2003), in which any two proteins share <25% sequence identity. Sequence identity is calculated using the method in (Brenner *et al.*, 2000).

Programs to be compared. We compare our method, denoted as PhyCMAP, with four state-of-the-art methods: CMAPpro (Di Lena *et al.*, 2012), NNcon (Tegge *et al.*, 2009), PSICOV (Jones *et al.*, 2012) and Evfold (Morcos *et al.*, 2011). We run NNcon, PSICOV and Evfold locally and CMAPpro through its web server. We do not compare our method with Astro-Fold because it is not publicly available. Further, it does not perform well because of lack of evolutionary information.

Evaluation methods. Depending on the chain distance of the two residues, there are three kinds of contacts: short-range, medium-range and long-range. Short-range contacts are closely related to local conformation and are relatively easy to predict. Medium-range and long-range contacts determine the overall shape of a protein and are more challenging to predict. We evaluate prediction accuracy using the top 5, L/10, L/5 predicted contacts, where L is the protein length.

M$_{\text{eff}}$: the number of non-redundant sequence homologs. Given a target and the multiple sequence alignment of all of its homologs, we measure the number of non-redundant sequence homologs by M$_{\text{eff}}$ as follows.

$$\mathrm{M}_{\text{eff}} = \sum_i \frac{1}{\sum_j S_{i,j}} \tag{1}$$

where both i and j go over all the sequence homologs, and $S_{i,j}$ is a binary similarity value between two proteins. Following Evfold (Morcos *et al.*, 2011), we compute the similarity of two sequence homologs using their hamming distance. That is, $S_{i,j}$ is 1 if the normalized hamming distance is <0.3; 0 otherwise.

3 RESULTS

Performance on the CASP10 set. As shown in Table 3, on the whole-CASP10 set, our PhyCMAP significantly exceeds the second best method CMAPpro in terms of the accuracy of the top five, L/10 and L/5 predicted contacts. The advantage of PhyCMAP over CMAPpro becomes smaller but still substantial when short-range contacts are excluded from consideration. PhyCMAP significantly outperforms NNcon, PSICOV and Evfold no matter how the performance is evaluated.

Performance on the 36 hard CASP10 targets. As shown in Table 4, on the 36 hard CASP10 targets, our PhyCMAP exceeds the second best method CMAPpro by 5–7% in terms of the accuracy of the top five, L/10 and L/5 predicted contacts. The advantage of PhyCMAP over CMAPpro becomes smaller but still substantial when short-range contacts are excluded from consideration. PhyCMAP significantly outperforms NNcon, PSICOV and Evfold no matter how many predicted contacts are evaluated. PSICOV and Evfold almost fail on these hard CASP10 targets. By contrast, CMAPpro, NNcon and PhyCMAP still work, although they do not perform as well as on the whole CASP10 set.

Note that both PSICOV and Evfold, two recent methods receiving a lot of attentions from the community, do not perform well on the CASP10 set. This is partially because they require a large number of sequence homologs for the protein under prediction. Nevertheless, most of the CASP targets, especially the hard ones, do not have so many sequence homologs because a protein with so many homologs likely has similar templates in PDB and thus, were not used by CASP.

Relationship between prediction accuracy and the number of sequence homologs. We divide the 123 CASP10 targets into five groups according to their logM_{eff} values: (0,2), (2,4), (4,6), (6,8), (8,10), which contain 19, 17, 25, 36 and 26 targets, respectively. Meanwhile, M_{eff} is the number of non-redundant sequence homologs for the protein under consideration (see Section 2 for definition). Only medium- and long-range contacts are considered here. Figure 4 clearly shows that the prediction accuracy increases with respect to M_{eff}. The more non-redundant

Table 3. This table lists the prediction accuracy of PhyCMAP, PSICOV, NNcon, CMAPpro and Evfold on short-, medium- and long-range contacts, tested on CASP10 (123 targets)

Method	Short-range, sequence distance from 6 to 12			Medium- and long-range, sequence distance >12			Medium-range, sequence distance >12 and ≤24			Long-range, sequence distance >24		
	Top 5	L/10	L/5	Top 5	L/10	L/5	Top 5	L/10	L/5	Top 5	L/10	L/5
PhyCMAP (C_α)	0.623	0.555	0.459	0.650	0.584	0.523	0.585	0.518	0.448	0.421	0.363	0.320
PhyCMAP (C_β)	0.667	0.580	0.482	0.655	0.604	0.539	0.621	0.550	0.465	0.514	0.425	0.373
PSICOV (C_α)	0.294	0.225	0.179	0.397	0.345	0.306	0.384	0.303	0.255	0.350	0.277	0.226
PSICOV (C_β)	0.379	0.281	0.223	0.522	0.458	0.405	0.515	0.387	0.316	0.457	0.372	0.315
NNcon (C_α)	0.595	0.499	0.399	0.472	0.409	0.358	0.463	0.393	0.334	0.286	0.239	0.188
CMAPpro (C_α)	0.506	0.437	0.368	0.517	0.466	0.424	0.485	0.414	0.363	0.380	0.336	0.297
CMAPpro (C_β)	0.543	0.477	0.395	0.519	0.466	0.415	0.491	0.419	0.370	0.395	0.347	0.313
Evfold (C_α)	0.236	0.193	0.165	0.380	0.326	0.295	0.351	0.294	0.249	0.328	0.257	0.225

Table 4. Prediction accuracy on the 36 hard CASP10 targets

Method	Short-range, sequence distance from 6 to 12			Medium and long-range, sequence distance >12		
	Top 5	L/10	L/5	Top 5	L/10	L/5
PhyCMAP (C_α)	0.456	0.439	0.378	0.394	0.378	0.325
PhyCMAP (C_β)	0.478	0.469	0.414	0.444	0.409	0.363
PSICOV (C_α)	0.100	0.083	0.082	0.183	0.156	0.150
PSICOV (C_β)	0.144	0.113	0.103	0.239	0.196	0.180
NNcon (C_α)	0.400	0.372	0.320	0.367	0.317	0.289
CMAPpro (C_α)	0.383	0.347	0.314	0.328	0.322	0.292
CMAPpro (C_β)	0.433	0.398	0.344	0.394	0.362	0.308
Evfold (C_α)	0.100	0.095	0.094	0.194	0.179	0.156

Note: The C_β results are in gray rows.

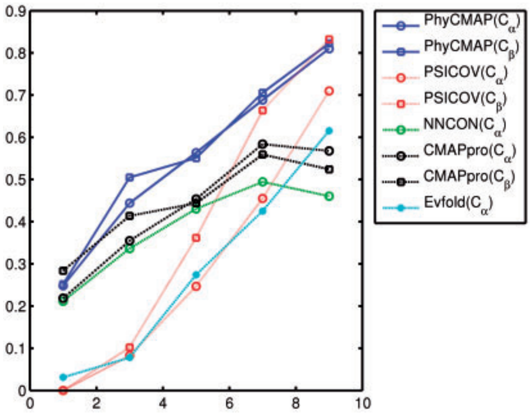

Fig. 4. The relationship between prediction accuracy and the number of non-redundant sequence homologs (M_{eff}). *x*-axis is logM_{eff} and *y*-axis is the mean accuracy of top L/10 predicted contacts on the corresponding CASP10 target group. Only medium- and long-range contacts are considered

homologs are available, the better prediction accuracy can be achieved by PhyCMAP, Evfold and PSICOV. However, CMAPpro and NNcon have decreased accuracy when logM_{eff} >8.

Figure 4 also shows that PhyCMAP outperforms Evfold, CMAPpro and NNcon across all M_{eff}. PhyCMAP greatly outperforms PSICOV in predicting C_α contacts regardless of M_{eff} and also in predicting C_β contacts when logM_{eff} ≤6. PhyCMAP has comparable performance as PSICOV in predicting C_β contacts when logM_{eff} >6.

Performance on Set600. To fairly compare our method with Evfold (Morcos *et al.*, 2011) and PSICOV (Jones *et al.*, 2012), both of which require a large number of sequence homologs, we divide Set600 into two subsets based on the amount of homologous information available for the protein under prediction. The first subset is relatively easier, containing 471 proteins with M_{eff}> 100 (see Section 2 for definition). All the proteins in this subset have >500 sequence homologs, which satisfies the requirement of PSICOV. The second subset is more challenging to predict, containing 130 proteins with M_{eff}≤ 100. As shown in Table 5, even on the first subset, PhyCMAP still exceeds PSICOV and Evfold, although the advantage over PSICOV is not substantial for C_β contacts prediction when short-range contacts are excluded from consideration. PhyCMAP also outperforms NNcon and CMAPpro on this set. As shown in Table 6, on the second subset, PhyCMAP significantly outperforms PSICOV and is slightly better than CMAPpro and NNcon. These results again confirm that our method applies to a protein without many sequence homologs, on which PSICOV and Evfold usually fail.

It should be noticed that CMAPpro used Astral 1.73 (Brenner *et al.*, 2000; Di Lena *et al.*, 2012) as its training set, which shares >90% sequence identity with 226 proteins in Set600 (180 with M_{eff}> 100 and 46 with M_{eff}≤ 100). To more fairly compare the prediction methods, we exclude the 226 proteins from Set600 that share >90% sequence identity with the CMAPpro training set. Here, the sequence identity is calculated using CD-HIT (Li and Godzik, 2006; Li *et al.*, 2001). This results in a set of 291 proteins with M_{eff}> 100 and 84 proteins M_{eff}≤ 100. Table 7 shows that PhyCMAP greatly outperforms CMAPpro and Evfold on the reduced dataset. PhyCMAP also outperforms

Table 5. Benchmark on the 471 proteins with $M_{eff} > 100$

Method	Short-range, sequence distance from 6 to 12			Medium- and long-range, sequence distance >12		
	Top 5	L/10	L/5	Top 5	L/10	L/5
PhyCMAP (C_α)	0.761	0.653	0.539	0.716	0.675	0.611
PhyCMAP (Cβ)	0.746	0.637	0.531	0.731	0.656	0.608
PSICOV (C_α)	0.457	0.341	0.257	0.528	0.465	0.411
PSICOV (C_β)	0.584	0.425	0.316	0.732	0.646	0.565
NNcon (C_α)	0.512	0.432	0.355	0.432	0.361	0.308
CMAPpro (C_α)	0.682	0.551	0.431	0.710	0.642	0.574
CMAPpro (C_β)	0.671	0.542	0.436	0.674	0.601	0.532
Evfold (C_α)	0.379	0.297	0.234	0.473	0.438	0.400

Table 6. Benchmark on the 130 proteins with $M_{eff} \leq 100$

Method	Short-range, sequence distance from 6 to 12			Medium- and long-range, sequence distance >12		
	Top 5	L/10	L/5	Top 5	L/10	L/5
PhyCMAP (C_α)	0.534	0.451	0.377	0.432	0.372	0.319
PhyCMAP (C_β)	0.505	0.435	0.365	0.418	0.364	0.314
PSICOV (C_α)	0.060	0.061	0.064	0.049	0.039	0.035
PSICOV (C_β)	0.077	0.070	0.073	0.069	0.050	0.045
NNcon (C_α)	0.442	0.363	0.309	0.368	0.339	0.301
CMAPpro (C_α)	0.435	0.365	0.314	0.368	0.331	0.300
CMAPpro (C_β)	0.532	0.434	0.353	0.358	0.331	0.280

Note: The result for Evfold is not shown, as it does not produce meaningful predictions for some proteins.

PSICOV in predicting C_α contacts, but it is slightly worse in predicting long-range C_β contacts.

3.1 Contribution of contrastive mutual information and pairwise contact scores

The CMI and HPS help improve the performance of our RF model. Table 8 shows their contribution to the prediction accuracy on the 471 proteins (with $M_{eff} > 100$) in Set600.

3.2 Contribution of physical constraints

Table 9 shows the improvement resulting from the physical constraints (i.e. the ILP method) over the RF method on Set600. On the 471 proteins with $M_{eff} > 100$, ILP improves medium and long-range contact prediction, but not short-range contact prediction. This result confirms that the physical constraints used by our ILP method indeed capture some global properties of a protein contact map. The improvement resulting from the physical constraints is larger on the 130 proteins with $M_{eff} \leq 100$. In particular, the improvement on short-range contacts is substantial. These results may imply that when homologous information is sufficient, we can predict short-range contacts accurately and thus, cannot further improve the prediction by using the physical constraints. When homologous information is insufficient for accurate contact prediction, we can improve the prediction using the physical constraints, which are complementary to evolutionary information.

3.3 Specific examples

We show the contact prediction results of two CASP10 targets T0693D2 and T0677D2 in Figures 5 and 6, respectively. T0693D2 has many sequence homologs with $M_{eff} = 2208.39$. As shown in Figure 5, PhyCMAP does well in predicting the long-distance contacts around the residue pair (20,100). For this target, PhyCMAP and Evfold obtain top L/10 prediction accuracy of 0.810 and 0.619, respectively, on medium- and long-range contacts. T0677D2 does not have many sequence homologs with $M_{eff} = 31.53$. As shown in Figure 6, our prediction matches well with the native contacts. PhyCMAP has top L/10 prediction accuracy 0.429 on medium- and long-range contacts, whereas Evfold cannot correctly predict any contacts.

4 CONCLUSION AND DISCUSSIONS

This article has presented a novel method for protein contact map prediction by integrating both evolutionary and physical constraints using machine learning and ILP. Our method differs from currently popular contact prediction methods in that we enforce a few physical constraints, which imply the sparsity constraint (used by PSICOV and Evfold), to the whole-contact map and take into consideration contact correlation. Our method also differs from the first-principle method (e.g. Astro-Fold) in that we exploit evolutionary information from several aspects (e.g. MI, context-specific distance potential and sequence profile) to significantly improve prediction accuracy. Experimental results confirm that our method outperforms existing popular machine-learning methods (e.g. CMAPpro and NNcon) and two recent co-mutation–based methods PSICOV and Evfold regardless of the number of sequence homologs available for the protein under consideration.

The study of our method indicates that the physical constraints are helpful for contact prediction, especially when the protein under consideration does not have many sequence homologs. Nevertheless, the physical constraints exploited by our current method do not cover all the properties of a protein contact map. To further improve prediction accuracy on medium- and long-range contact prediction, we may take into consideration global properties of a protein distance matrix. For example, the pairwise distances of any three residues shall satisfy the triangle inequality. Some residues also have correlated pairwise distances. To enforce this kind of distance constraints, we shall introduce distance variables and also define their relationship with contact variables. By introducing distance variables, we may also optimize the distance probability, as opposed to the contact probability used by our current ILP method. Further, we can also introduce variables of β-sheet (group of β-strands) to capture more global properties of a contact map.

One may ask how our approach compares with a model-based filtering strategy in which 3D models are built based on initial predicted contacts and then used to filter incorrect predictions. Our method differs from this general ‘model-based filtering’ strategy in a couple of aspects. First, it is time-consuming to build thousands or at least hundreds of 3D models with initial

Table 7. This table lists the prediction accuracy of PhyCMAP, PSICOV, NNcon, CMAPpro and Evfold on short-, medium- and long-range contacts, tested on Set600

Method	Short-range, sequence distance from 6 to 12			Medium- and long-range, sequence distance >12			Medium-range, sequence distance >12 and ≤24			Long-range, sequence distance >24		
	Top 5	L/10	L/5	Top 5	L/10	L/5	Top 5	L/10	L/5	Top 5	L/10	L/5
a) The 291 proteins in Set600 with Meff >100 and ≤90% sequence identify with Astral 1.73												
PhyCMAP(C_α)	0.759	0.653	0.536	0.713	0.680	0.622	0.639	0.570	0.496	0.582	0.528	0.461
PhyCMAP(C_β)	0.741	0.641	0.534	0.746	0.653	0.611	0.655	0.571	0.500	0.636	0.550	0.477
PSICOV(C_α)	0.459	0.343	0.258	0.528	0.469	0.417	0.462	0.363	0.282	0.483	0.418	0.358
PSICOV(C_β)	0.582	0.422	0.314	0.733	0.650	0.569	0.647	0.496	0.371	0.674	0.584	0.495
NNcon(C_α)	0.475	0.390	0.318	0.377	0.313	0.267	0.342	0.284	0.236	0.224	0.182	0.152
CMAPpro(C_α)	0.643	0.519	0.412	0.689	0.618	0.554	0.580	0.511	0.439	0.527	0.469	0.416
CMAPpro(C_β)	0.642	0.520	0.422	0.653	0.580	0.515	0.573	0.494	0.421	0.504	0.444	0.396
Evfold(C_α)	0.382	0.297	0.235	0.488	0.442	0.398	0.451	0.366	0.289	0.442	0.389	0.342
b) The 84 proteins in Set600 with Meff ≤100 and ≤90% sequence identity with Astral 1.73												
PhyCMAP(C_α)	0.580	0.488	0.404	0.481	0.430	0.357	0.476	0.417	0.335	0.204	0.179	0.159
PhyCMAP(C_β)	0.548	0.468	0.392	0.454	0.408	0.345	0.452	0.399	0.331	0.220	0.214	0.187
PSICOV(C_α)	0.070	0.071	0.072	0.065	0.050	0.044	0.074	0.055	0.049	0.063	0.043	0.035
PSICOV(C_β)	0.081	0.078	0.083	0.088	0.068	0.059	0.092	0.066	0.059	0.076	0.058	0.046
NNcon(C_α)	0.535	0.421	0.342	0.324	0.298	0.248	0.348	0.321	0.271	0.162	0.132	0.114
CMAPpro(C_α)	0.465	0.370	0.316	0.346	0.328	0.285	0.360	0.332	0.286	0.173	0.169	0.158
CMAPpro(C_β)	0.447	0.367	0.321	0.346	0.320	0.287	0.366	0.331	0.290	0.191	0.189	0.176
Evfold(C_α)	0.074	0.068	0.066	0.079	0.058	0.039	0.074	0.053	0.045	0.063	0.042	0.032

Table 8. The contribution of CMI and homologous pair contact scores to C_β contact prediction

Method	Short-range contacts			Medium- and long-range		
	Top 5	L/10	L/5	Top 5	L/10	L/5
with CMI and HPS	0.754	0.632	0.521	0.720	0.649	0.589
no CMI and HPS	0.600	0.570	0.487	0.538	0.560	0.506

Note: Similar results are observed for C_α contact prediction.

Table 9. The contribution of physical constraints

Method	Short-range contacts			Medium- and long-range		
	Top 5	L/10	L/5	Top 5	L/10	L/5
471 proteins in Set600 with $M_{eff} > 100$						
RF + ILP	0.746	0.637	0.531	0.731	0.656	0.608
RF	0.754	0.632	0.521	0.720	0.649	0.589
130 proteins in Set600 with $M_{eff} \leq 100$						
RF + ILP	0.505	0.435	0.365	0.418	0.364	0.314
RF	0.445	0.368	0.299	0.414	0.342	0.281

Note: The results are C_β contact prediction.

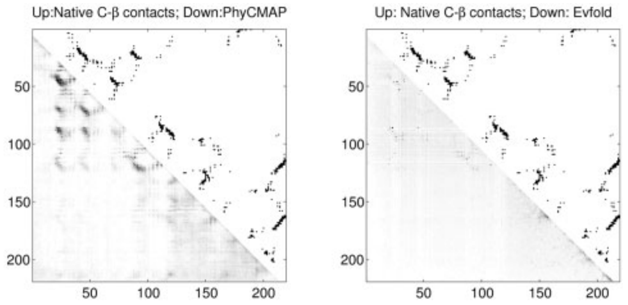

Fig. 5. The predicted medium- and long-range contacts for T0693D2. The upper triangles display the native C_β contacts. The lower triangles of the left and right plots display the contact probabilities predicted by PhyCMAP and Evfold, respectively

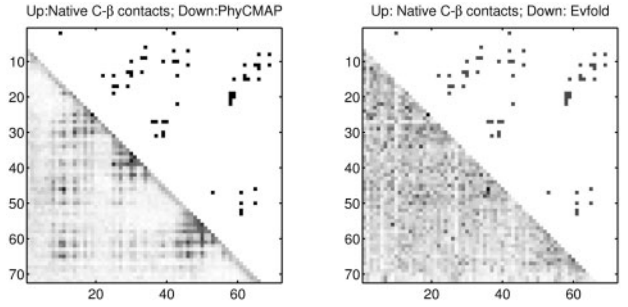

Fig. 6. The predicted medium- and long-range contacts for T0677D2. The upper triangles display the native C_β contacts. The lower triangles of the left and right plots display the contact probabilities predicted by PhyCMAP and Evfold, respectively

predicted contacts. In contrast, our method can do contact prediction (using physical constraints) within minutes. Second, the quality of the 3D models is also determined by other factors, such as energy function and energy optimization (or conformation sampling) methods, whereas our method is independent of these factors. Even if the energy function is accurate, the energy optimization algorithm often is trapped to local minima because the energy function is not rugged. That is, the 3D models resulting from energy minimization are biased toward a specific region of the conformation space, unless an extremely large-scale of conformation sampling is conducted. Therefore, the predicted contacts derived from these models may also suffer from this 'local minima' issue. By contrast, our integer programming method can search through the whole conformation space and find the global optimal solution; thus, it is not biased to any local minima region. By using our predicted contacts as constraints, we may pinpoint to the good region of a conformation space (without being biased by local minima), reduce the search space and significantly speed-up conformation search.

ACKNOWLEDGEMENTS

The authors are grateful to anonymous reviewers for their insightful comments and to the University of Chicago Beagle team and TeraGrid for their support of computational resources.

Funding: National Institutes of Health (R01GM0897532); National Science Foundation DBI-0960390; NSF CAREER award CCF-1149811; Alfred P. Sloan Research Fellowship.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.

Baú,D. *et al.* (2006) Distill: a suite of web servers for the prediction of one-, two-and three-dimensional structural features of proteins. *BMC Bioinformatics*, **7**, 402.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Brenner,S.E. *et al.* (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Caprara,A. *et al.* (2004) 1001 optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap. *J. Comput. Biol.*, **11**, 27–52.

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

International Business Machines Corporation. (2009) IBM ILOG CPLEX. *V12. 1: User's Manual for CPLEX*.

Di Lena,P. *et al.* (2012) Deep architectures for protein contact map prediction. *Bioinformatics*, **28**, 2449–2457.

Fariselli,P. and Casadio,R. (1999) A neural network based predictor of residue contacts in proteins. *Protein Eng.*, **12**, 15–21.

Göbel,U. *et al.* (2004) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Klepeis,J. and Floudas,C. (2003) ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.*, **85**, 2119–2146.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li,W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomforest. *R News*, **2**, 18–22.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA*, **108**, E1293–E1301.

Olmea,O. and Valencia,A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.*, **2**, S25–S32.

Ortiz,A.R. *et al.* (1999) Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, **37**, 177–185.

Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.

Tan,Y.H. *et al.* (2006) Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. *Proteins*, **64**, 587–600.

Tegge,A.N. *et al.* (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res.*, **37**, W515–W518.

Vassura,M. *et al.* (2008) Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 357–367.

Vendruscolo,M. and Domany,E. (1998) Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.*, **109**, 11101.

Vendruscolo,M. *et al.* (1997) Recovery of protein structure from contact maps. *Fold. Des.*, **2**, 295–306.

Vullo,A. *et al.* (2006) A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, **7**, 180.

Wang,G. and Dunbrack,R.L., Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,S. *et al.* (2013) Protein structure alignment beyond spatial proximity. *Sci. Rep.*, **3**, 1448.

Wu,S. *et al.* (2011) Improving protein structure prediction using multiple sequence-based contact predictions. *Structure*, **19**, 1182–1191.

Wu,S. and Zhang,Y. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, **24**, 924–931.

Xu,J. *et al.* (2007) A parameterized algorithm for protein structure alignment. *J. Comput. Biol.*, **14**, 564–577.

Zhao,F. and Xu,J. (2012) A position-specific distance-dependent statistical potential for protein structure and functional study. *Structure*, **20**, 1118–1126.

Zhou,H. and Skolnick,J. (2007) Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. *Proteins*, **71**, 1211–1218.