Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

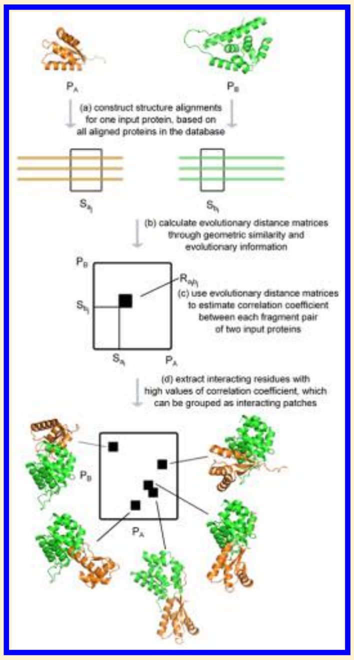
Identification of Protein−Protein Interactions by Detecting Correlated Mutation at the Interface

Fei Guo,* Yijie Ding, Zhao Li, and Jijun Tang

School of Computer Science and Technology, Tianjin University, 92 Weijin Road, Nankai District, Tianjin 300072, P.R. China

S *Supporting Information*

ABSTRACT: Protein−protein interactions play key roles in a multitude of biological processes, such as de novo drug design, immune response, and enzymatic activity. It is of great interest to understand how proteins in a complex interact with each other. Here, we present a novel method for identifying protein−protein interactions, based on typical co-evolutionary information. Correlated mutation analysis can be used to predict interface residues. In this paper, we propose a non-redundant database to detect correlated mutation at the interface. First, we construct structure alignments for one input protein, based on all aligned proteins in the database. Evolutionary distance matrices, one for each input protein, can be calculated through geometric similarity and evolutionary information. Then, we use evolutionary distance matrices to estimate correlation coefficient between each pair of fragments from two input proteins. Finally, we extract interacting residues with high values of correlation coefficient, which can be grouped as interacting patches. Experiments illustrate that our method achieves better results than some existing co-evolution-based methods. Applied to SK/RR interaction between sensor kinase and response regulator proteins, our method has accuracy and coverage values of 53% and 45%, which improves upon accuracy and coverage values of 50% and 30% for DCA method. We evaluate interface prediction on four protein families, and our method has overall accuracy and coverage values of 34% and 30%, which improves upon overall accuracy and coverage values of 27% and 21% for PIFPAM. Our method has overall accuracy and coverage values of 59% and 63% on Benchmark v4.0, and 50% and 49% on CAPRI targets. Comparing to existing methods, our method improves overall accuracy value by at least 2%.


P_A
P_B
(a) construct structure alignments for one input protein, based on all aligned proteins in the database
S_{A_i}
S_{B_j}
(b) calculate evolutionary distance matrices through geometric similarity and evolutionary information
P_B
S_{B_j}
$R_{a_i b_j}$
(c) use evolutionary distance matrices to estimate correlation coefficient between each fragment pair of two input proteins
S_{A_i}
P_A
(d) extract interacting residues with high values of correlation coefficient, which can be grouped as interacting patches
P_B
P_A

■ INTRODUCTION

Studies on protein−protein interactions are important in molecular biology research. How to build more effective models is a key technology for predicting interface residues. The protein−protein interface can be affected by sequence and structural changes at distant locations; it means that changes from different regions can together induce correlated mutation at the interface.[1,2] Typical co-evolutionary information can occur to reserve global stability, where amino acid change at one site may give rise to change at another site. The interface can be detected by correlated mutation analysis, because of sequence and structural changes taking place at distant locations in a cooperative way.[3,4]

A wide range of co-evolution-based methods have been used to study evolutionary conservation at the interface and designed to identify protein−protein interactions.[5−7] Residues with co-evolutionary information may share a certain function in a protein family. The crucial step is to distinguish co-evolutionary information at the interface from others.[8,9] There are three main approaches to identify interface residues, involving substitution correlation,[10] mutual information on amino acid frequency,[11] and global statistical model.[12,13]

First, some approaches use evolutionary correlation to distinguish interface residues from others. MirrorTree[14] is a well-known method to quantify correlation between two input proteins. They find an orthologue of each protein in multiple species and align sequences on common species to get a multiple sequence alignment. The phylogenetic tree can be generated to infer co-evolution on the basis of tree similarity. They create evolutionary distance matrix and use a linear correlation coefficient to predict possible protein−protein interactions. Tol-MirrorTree[15] takes advantage of global evolutionary relationship between species to correct background tree similarity. ContextMirror[16] uses a lot of trees to correct factors that affect observed tree similarity.

Second, mutual information methods examine rational protein−protein interactions. Mutual information, an information theory measure, has been extensively employed and modified to identify interface residues. They consider probability of each amino acid in different sequences for a site. In fact, it quantifies whether an amino acid in a given

Received: May 28, 2015
Published: August 18, 2015

DOI: 10.1021/acs.jcim.5b00320

sequence for one site is a good prediction of any given amino acid in the same sequence for another site.[17]

Third, other methods propose global statistical model to identify interface residues. Direct coupling analysis approaches[9,18,19] establish a global statistical model in terms of position-specific variabilities and inter-position contacts. It summarizes global structure of contact map, as a reliable guide to predict protein−protein interactions. PSICOV[20] introduces sparse inverse covariance estimation to interface prediction. It corrects phylogenetic correlation noise and allows accurate discrimination of direct contacts from indirectly mutation correlations. Statistical coupling analysis approaches characterize evolutionary patterns in a protein family and explore co-evolutionary information at the interface.[21]

Other kinds of efficient techniques have been developed for binding sites identification.[22−28] Some existing approaches are based on analyzing differences between interface residues and non-interface residues, through machine learning methods or statistical methods.[29−34] In addition, several structural algorithms have also been used to identify binding sites, through analyzing surface structures.[35−37] Meta-servers have also been constructed to combine strengths of some existing approaches. The program called meta-PPISP[38] combines three individual servers, namely cons-PPISP, ProMate, and PINUP; another program called metaPPI[39] combines five prediction methods, namely PPI-Pred, PINUP, PPISP, ProMate, and Sppider.

Identifying of protein−protein interface not only depends on sequence information, but also relies on structural information and other physicochemical properties. Some proteins have quite different sequences but high similar structures, and they may evolve from a common ancestor.[40] Structural information should be most effective features for predicting interface residues.[41] Hydrogen bonds are also known to be essential in identifying binding specificity.[42] Our method utilizes sequence and structural information to calculate evolutionary relationship between two input proteins; however, most existing technologies only analyze sequence information on current interacting residues, but cannot represent real situations well.[7,43] We also analyze correlated mutation of structural neighboring residues, thus are insufficient to predict binding sites with high accuracy. The underlying rational is that the more support from the neighboring structural information, the more likely that the interface residue is native-like.

Here, we present a novel method for identifying protein−protein interactions, based on typical co-evolutionary information. We propose a non-redundant database to detect correlated mutation at the interface. First, we use DeepAlign[44] to construct structure alignments for one input protein, based on all aligned proteins in the database. DeepAlign aligns protein three-dimensional structures using evolutionary information and beta strand orientation, in addition to geometric similarity. DeepAlign is much more consistent with manual alignments than other methods, such as DALI[45] and TMalign.[46] Evolutionary distance matrices, one for each input protein, can be calculated through geometric similarity and evolutionary information. Then, we use evolutionary distance matrices to estimate correlation coefficient between each pair of fragments from two input proteins. In correlated mutation analysis, correlation coefficient can be used to quantify co-evolutionary information at the interface. Finally, we extract interacting residues with high values of correlation coefficient, which can be grouped as interacting patches.

Experiments illustrate that our method achieves better results than some existing methods. Applied to SK/RR interaction between sensor kinase and response regulator proteins, our method has accuracy and coverage values of 53% and 45%, which improves upon accuracy and coverage values of 50% and 30% for DCA method. We evaluate interface prediction on four protein families, and our method has overall accuracy and coverage values of 34% and 30%, which improves upon overall accuracy and coverage values of 27% and 21% for PIFPAM. On Benchmark v4.0, our method has overall accuracy and coverage values of 59% and 63%. On CAPRI targets, our method has overall accuracy and coverage values of 50% and 49%. Comparing to existing methods, our method improves overall accuracy value by at least 2%.

■ METHODS

We present a novel method for identifying protein−protein interactions, based on typical co-evolutionary information. We propose a non-redundant database to detect correlated mutation at the interface. First, we use DeepAlign to construct structure alignments for one input protein, based on all aligned proteins in the database. Evolutionary distance matrices, one for each input protein, can be calculated through geometric similarity and evolutionary information. Then, we use evolutionary distance matrices to estimate correlation coefficient between each pair of fragments from two input proteins. Finally, we extract interacting residues with high values of correlation coefficient, which can be grouped as interacting patches. The entire process is illustrated in Figure 1.

A complex may contain several subunits and multiple interfaces. Each binding interface in a complex occurs in two subunits. Two residues in a pair of subunits are called *interface residues* if any two atoms, one from each residue, interact. By “interact”, here we mean that the distance between two atoms is less than 6 Å.

Database Construction. We construct structure alignments for one input protein, as defined by hidden Markov model in Pfam database.[47] Only data from unique species can be included to avoid over sampling of organisms. We prepare aligned proteins to construct structure alignments for each input protein. First, we use BLAST to retrieve a non-redundant database for each input protein.[48] We calculate pairwise identities between input protein and all proteins in the database and select aligned proteins with pairwise identities more than 40%. Second, all aligned proteins share no more than 90% identity to avoid local biases.

Protein Structure Alignment. We use DeepAlign[44] to construct structure alignments for one input protein, based on all aligned proteins in the database. DeepAlign aligns protein three-dimensional structures using evolutionary information and beta strand orientation, in addition to geometric similarity. DeepAlign identifies similar fragments by using amino acid and local structure mutation matrices and refines structure alignments via dynamic programming and gap elimination.

We construct structure alignments for one input protein. Each input protein and aligned proteins can be formed as a protein set $P = P_1, P_2, ..., P_N$, where P_1 is the input protein and P_2−P_N are all aligned proteins. A scoring function is used to determine how likely two fragments shall be aligned. It is composed of amino acid mutation score, local substructure substitution potential, hydrogen-bonding similarity, and geometric similarity. The equivalence of two residues P_k^i and P_l^i at the ith position of proteins P_k and P_l is estimated as

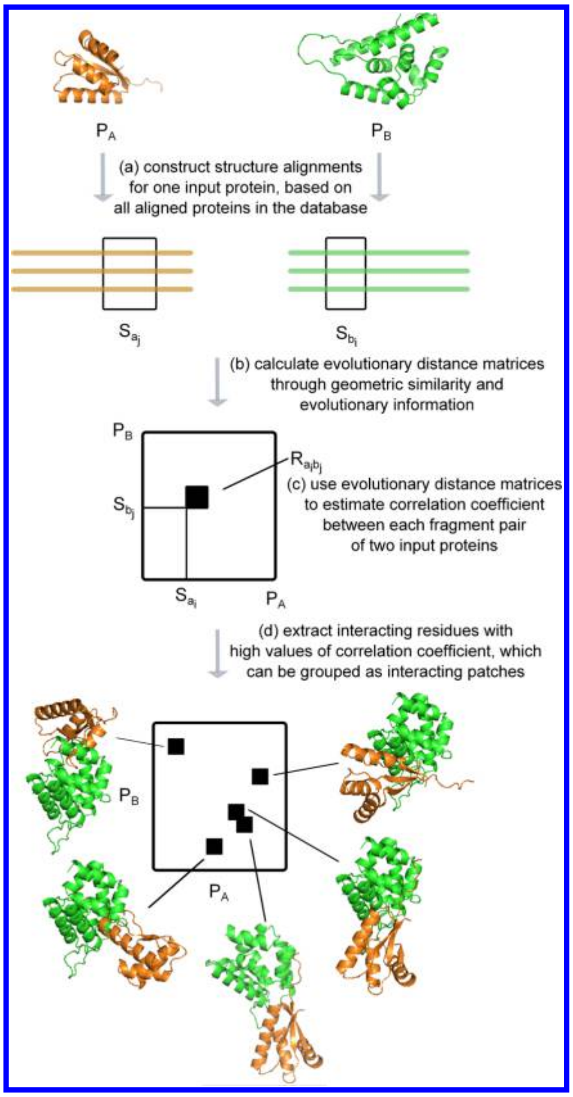

P_A
P_B
(a) construct structure alignments for one input protein, based on all aligned proteins in the database
S_{a_j}
S_{b_j}
(b) calculate evolutionary distance matrices through geometric similarity and evolutionary information
P_B
$R_{a_ib_j}$
(c) use evolutionary distance matrices to estimate correlation coefficient between each fragment pair of two input proteins
S_{b_j}
S_{a_i}
P_A
(d) extract interacting residues with high values of correlation coefficient, which can be grouped as interacting patches
P_B
P_A

Figure 1. Entire process of interface residue prediction.

$$S(P_k^i, P_l^i) = [\max(0, B(P_k^i, P_l^i)) + C(P_k^i, P_l^i)] \times \nu(P_k^i, P_l^i) \times d(P_k^i, P_l^i) \quad (1)$$

where $B(P_k^i,P_l^i)$ is widely used amino acid substitution matrix BLOSUM62,[49] $C(P_k^i,P_l^i)$ is local substructure substitution matrix CLESUM,[50] $\nu(P_k^i,P_l^i)$ measures hydrogen-bonding similarity,[44] and $d(P_k^i,P_l^i)$ is spatial proximity of two aligned residues. Here, we use TMscore[46] to calculate $d(P_k^i,P_l^i)$, defined as

$$d(P_k^i, P_l^i) = \frac{1}{1 + (|A_i - B_i|/d_0)^2} \quad (2)$$

where A_i and B_i are transformed 3D coordinates of two C_α atoms in residues P_k^i and P_l^i, and $d_0 = 1.24(L_s - 15)^{1/3} - 1.8$ is a length-dependent normalization factor (L_s being length of smaller protein), used to offset impact of protein length. TMscore is a widely used measure in protein structure analysis and demonstrates excellent performance on identifying similar structures.

Evolutionary Distance Matrix. Evolutionary distance matrices can be calculated through geometric similarity and evolutionary information. Each input protein is parsed into fragments through a sliding window of pre-determined length w. As windows slid along their corresponding alignments, all pairs of fragments can be obtained from two aligned proteins P_k and P_l, extracting from a protein set $P = P_1, P_2, ..., P_N$. For a structure alignment, we measure evolutionary distance of aligned fragments starting at mth position ($0 \le m \le L_1 - w$), using sequence and structural information.

$$D_m(P_k, P_l) = -\sum_{i=m}^{m+w-1} \ln[S(P_k^i, P_l^i)] \quad (3)$$

where P_k^i and P_l^i are two aligned residues at ith position, and aligned fragments contain w residues starting at mth position and ending at $(m + w - 1)$th position.

Calculating Evolutionary Relationship. We use evolutionary distance matrices to estimate correlation coefficient between each pair of fragments from two input proteins. We sort aligned proteins according to their species and re-organize evolutionary distance matrices. Two corresponding units, one from each input protein, represent evolutionary distances for identical or similar species.

In correlated mutation analysis, correlation coefficient can be used to quantify co-evolutionary information on interface residues. For two pairs of fragments starting at mth position and nth position, correlation coefficient can be calculated by two evolutionary distance matrices $D_{m,X}$ and $D_{n,Y}$.

$$r_{m,n} = \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} (D_{m,X}(P_k, P_l) - \bar{D}_{m,X}) \times (D_{n,Y}(P_k, P_l) - \bar{D}_{n,Y}) / \left[\sqrt{\sum_{k=1}^{N-1} \sum_{l=k+1}^{N} (D_{m,X}(P_k, P_l) - \bar{D}_{m,X})^2} \times \sqrt{\sum_{k=1}^{N-1} \sum_{l=k+1}^{N} (D_{n,Y}(P_k, P_l) - \bar{D}_{n,Y})^2}\right] \quad (4)$$

where $D_{m,X}(P_k,P_l)$ and $D_{n,Y}(P_k,P_l)$ are equivalent elements of $D_{m,X}$ and $D_{n,Y}$, and $\bar{D}_{m,X}$ and $\bar{D}_{n,Y}$ are means of $D_{m,X}(P_k,P_l)$ and $D_{n,Y}(P_k,P_l)$, respectively.

The correlation coefficient between two fragments must be translated into correlation coefficient between two residues. The initial $r_{i,j}$ value ($0 \le i \le L_1$ and $0 \le j \le L_2$) is zero. If each $r_{m,n}$ value is calculated, the corresponding $r_{i,j}$ value ($m \le i \le m + w - 1$ and $n \le j \le n + w - 1$) will be added by $r_{m,n}$ value. After all possible $r_{m,n}$ values are added to their corresponding $r_{i,j}$ values, each $r_{i,j}$ value is divided by the number of added times. We rescale all $r_{i,j}$ values to range [0,1].

Extracting Interface Residues. We extract interacting residues with high values of correlation coefficient, which can be grouped as interacting patches. The threshold value r_{th} is used to harvest all possible residue pairs between two input proteins. Surface residue pairs with $r \ge r_{\text{th}}$ are called interacting residues. We further investigate whether changing r_{th} value help to improve interface residue prediction.

Considering neighboring residues, we construct a sphere with radius of 10 Å for all interacting residues. Mutation correlation coefficient of each interacting residue pair is calculated as

$$r_{i,j}' = r_{i,j} + \sum_{\text{dis}(i,u) \le 10} \frac{1}{\text{dis}(i, u)} r_{u,j} + \sum_{\text{dis}(j,v) \le 10} \frac{1}{\text{dis}(j, v)} r_{i,v} \quad (5)$$

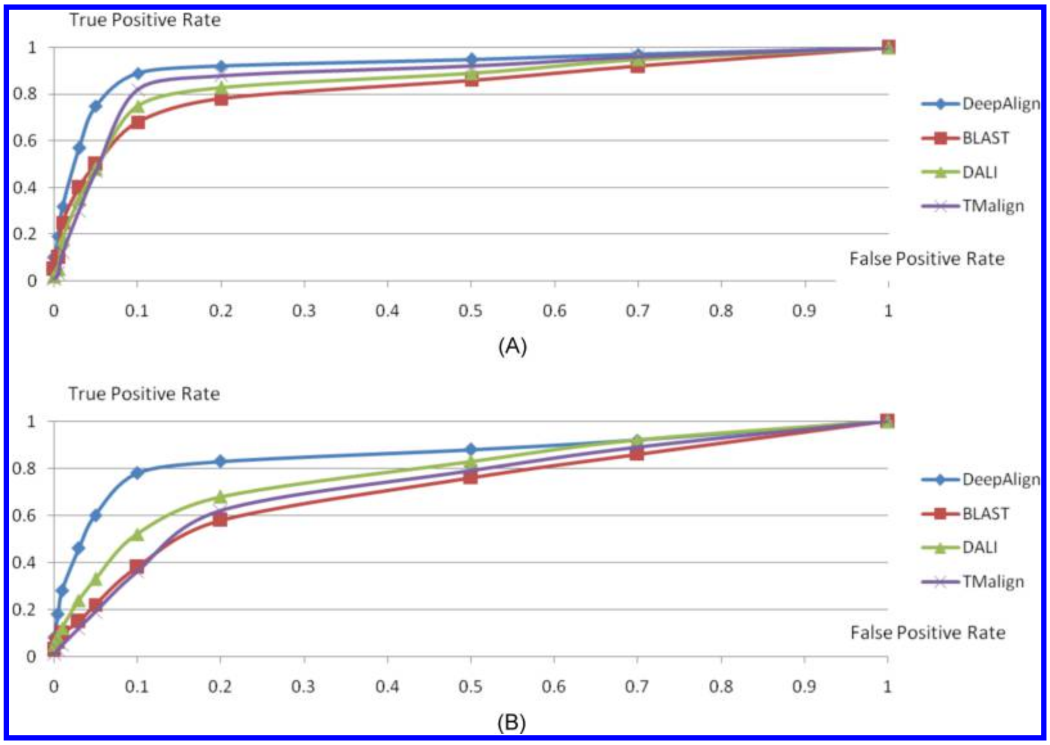


Figure 2. ROC curves by DeepAlign, BLAST, DALI, and TMalign: (A) SABmark-sup and (B) SABmark-twi.

where residues u are from the protein having residue i, and residues v are from the protein having residue j.

We rank interacting residues by using mutation correlation coefficient. Top interacting residues can be grouped into different regions. All interacting residues served as graph nodes, and each undirected edge is built when two nodes are within distance 10 Å. Strongly connected components are considered as interacting patches. One region, containing a very small number of interacting residues, indicates a weak signal and can be discarded. We cluster interacting patches as predicted interface residues.

Assessment of Interface Prediction. Protein−protein interface prediction must be tested at residue level. *Accuracy* is calculated as the number of correctly predicted interface residues divided by the total number of interface residues in predicted complex. *Coverage* is calculated as the number of correctly predicted interface residues divided by the total number of interface residues in native complex.

We also calculate P value for binding sites prediction. The calculation of P value should be probability of obtaining not less than n correctly predicted interface residues by randomly picking out N predicted interface residues. The probability that a random method obtains success in one trial is m/M, where M is the number of all surface residues, and m is the number of correctly interface residues among them. Therefore, P value for binding sites prediction is given by

$$P = \sum_{i=n}^{N} \frac{N!}{n!(N-n)!}\left(\frac{m}{M}\right)^{n}\left(1 - \frac{m}{M}\right)^{N-n} \quad (6)$$

RESULTS

In this section, we have done three kinds of experiments. First, we present statistical analysis of evolutionary information at the interface. Then, we compare our method to some existing techniques for identifying protein−protein interfaces. Finally, we evaluate our method on various sets of protein−protein complexes. Experiments illustrate that our method achieves better results than some existing methods.

Assessment of Evolutionary Information. We use DeepAlign[44] to construct structure alignments for one input protein, based on all aligned proteins in the database. DeepAlign aligns protein three-dimensional structures using evolutionary information and beta strand orientation, in addition to geometric similarity. We compare DeepAlign with sequence alignment method BLAST[48] and structure alignment methods DALI[45] and TMalign.[46] We use SABmark[51] to test performances of four alignment methods to identify distant homologues and structural analogues. SABmark-sup is a superfamily set in SABmark, and SABmark-twi is a twilight set in SABmark with low sequence identity. Given a protein, four methods rank all alignments between this protein and proteins in Benchmark. We examine if top-ranked proteins are in the same group as query protein or not. Results are evaluated by ROC curve, as shown in Figure 2. DeepAlign has best ROC curve, especially at high specificity area. DeepAlign is much more consistent with manual alignments than other methods. Evolutionary distance can be calculated through geometric similarity and evolutionary information.

Statistical Analysis. We use a sliding window of pre-determined length w, obtaining all pairs of fragments from two aligned proteins. In order to find most suitable pre-determined length, we produce value of length w from 5 to 10 with interval of 1. Here, we calculate correlation coefficient r on 100 pairs of interface residues and 100 pairs of non-interface residues, randomly extracted from Benchmark v4.0.[52] The relationship between interface prediction and value of length w is shown in Table 1. We can get the most accurate interface prediction, if pre-determined length w = 7.

We further investigate whether changing value of threshold r_{th} help to improve interface prediction. We produce value of

Table 1. Relationship between Interface Prediction and Value of Length w, Based on Correlation Coefficient r

	Acc	Cov
$w = 5$	43%	29%
$w = 6$	52%	38%
$w = 7$	57%	48%
$w = 8$	48%	36%
$w = 9$	46%	33%
$w = 10$	39%	27%

threshold r_{th} from 0.2 to 0.5 with interval of 0.05. Here, we calculate updated correlation coefficient r' on above 200 pairs of interface residues and non-interface residues. The relationship between interface prediction and value of threshold r_{th} is shown in Table 2. We can obtain the most accurate interface prediction, if threshold $r_{th} = 0.35$.

Table 2. Relationship between Interface Prediction and Value of Threshold r_{th}, Based on Updated Correlation Coefficient r'

	Acc	Cov
$r_{th} = 0.2$	39%	33%
$r_{th} = 0.25$	51%	46%
$r_{th} = 0.3$	62%	53%
$r_{th} = 0.35$	67%	59%
$r_{th} = 0.4$	59%	51%
$r_{th} = 0.45$	52%	47%
$r_{th} = 0.5$	49%	42%

Interface Prediction on SK/RR Interaction. We study HisKA domain of sensor histidine kinase (PF00512) and its partner response regulator domain (PF00072) in Pfam database.[47] Interface identification can be tested by using structural representatives of HisKA domain of SK (HK853; PDB ID code 2C2A, chain A) and of RR domain (Spo0F; PDB ID code 1PEY, chain A), as well as cocrystal structure of Spo0F in complex with Spo0B (PDB ID code 1F51, chain A:E).

Detecting Correlated Residues. We analyze 27 correlated residues with high values of mutation correlation coefficient, involving 15 SK positions and 12 RR positions. For HK853, predicted interface residues being part of interface are 267, 268, 271, 272, 275, 276, 291, 294, and 298, as indicated by red boxes in Figure 3. Predicted interface residues of SK belonging to non-interface are 245, 249, 253, 256, 305, and 308. For Spo0F, predicted interface residues being part of interface are 14, 15, 18, 19, 21, and 22, as indicated by red boxes in Figure 3. Predicted interface residues of RR belonging to non-interface are 56, 57, 86, 87, 90, and 91.

Identification of SK/RR Interaction. Our method predicts 32 interacting residues between Spo0B and Spo0F co-crystal structures. These residues are grouped into three interacting patches. We compare our method with DCA method[9,53] on SK/RR interaction. Applied to SK/RR interaction between sensor kinase and response regulator proteins, our method has accuracy and coverage values of 53% and 45%, which improves upon accuracy and coverage values of 50% and 30% for DCA method. For 300 pairs of interacting residues, scatter plot of distance between two interacting residues against their correlation coefficient is shown in Figure 4. Most interacting residues with high values of correlation coefficient correspond to short distance.

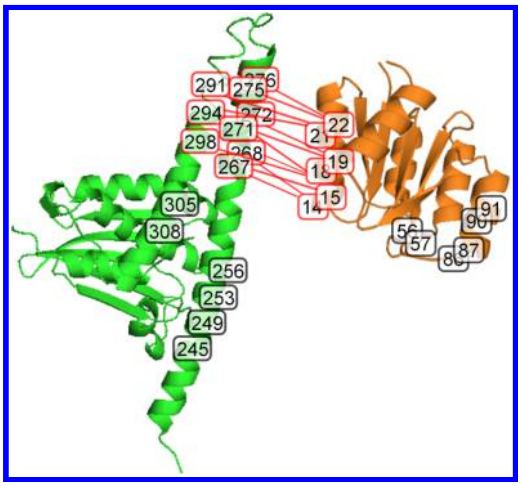


Figure 3. Our method detects correlated residues on SK/RR interaction. Interface residues are indicated in red boxes, and non-interface residues are indicated in black boxes.

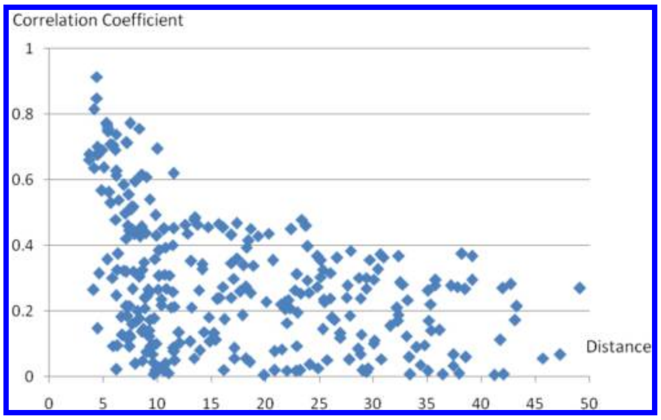


Figure 4. Scatter plot of distance between two interacting residues against their correlation coefficient in Spo0B/Spo0F co-crystal structure.

Interface Prediction on Four Protein Families. We evaluate interface prediction of our method on four protein families, covering four SCOP classes.[54] Our method predicts 31−68 interacting residues, which can be grouped into 3−8 interacting patches, as shown in Table 3. We compare our method with PIFPAM[55] on four protein families. On four protein families, our method has overall accuracy and coverage values of 34% and 30%, which improves upon overall accuracy and coverage values of 27% and 21% for PIFPAM.

Case of Spirulina platensis. We study *S. platensis* α-subunit (PDB ID code 1GH0, chain A) and β-subunit (PDB ID code 1GH0, chain B). Our method predicts 37 interacting residues between α-subunit and β-subunit structures. These residues are grouped into five interacting patches. On *S. platensis*, our method has accuracy and coverage values of 52% and 43%.

We analyze 32 correlated residues with high values of mutation correlation coefficient, involving 17 α-subunit positions and 15 β-subunit positions. For α-subunit, predicted interface residues being part of interface are 5, 6, 9, 10, 24, 27, 31, 38, and 42, as indicated by red boxes in Figure 5. Predicted interface residues of α-subunit belonging to non-interface are 78, 79, 82, 83, 117, 118, 121, and 122. For β-subunit, predicted interface residues being part of interface are 5, 6, 9, 10, 24, 27, 31, 38, and 42, as indicated by red boxes in Figure 5. Predicted

Table 3. Interface Prediction of Our Method and PIFPAM on Four Protein Families

		our method					PIFPAM		
protein family	PDB ID	$\mathrm{Int}_{\mathrm{res}}$[a]	$\mathrm{Int}_{\mathrm{pat}}$[b]	Acc	Cov	*P* value	Acc	Cov	*P* value
phycocyanin	1GH0	37	5	0.52	0.43	0.006	0.60	0.29	0.008
electron-transfer flavoprotein	1EFP	68	8	0.29	0.26	0.054	0.19	0.28	0.141
hydrogenase [Ni-Fe]	1FRF	52	7	0.33	0.37	0.025	0.21	0.15	0.106
dihydroorotate dehydrogenase B	1EP3	31	3	0.19	0.14	0.136	0.07	0.12	0.282

[a] $\mathrm{Int}_{\mathrm{res}}$ is the number of interacting residues in predicted complex. [b] $\mathrm{Int}_{\mathrm{pat}}$ is the number of interacting patches in predicted complex.

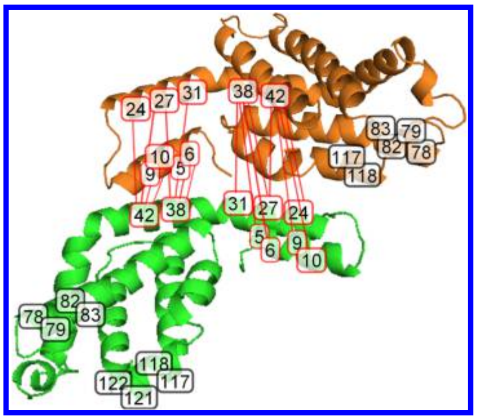


Figure 5. Our method detects correlated residues on *S. platensis*. Interface residues are indicated in red boxes, and non-interface residues are indicated in black boxes.

interface residues of β-subunit belonging to non-interface are 78, 79, 82, 83, 117, and 118.

Evaluation on Benchmark v4.0. To further evaluate our method, we perform tests on Benchmark v4.0,[52] as shown in Table 4. Our method has overall accuracy and coverage values of 59% and 63%. Complexes are classified into three categories, according to magnitude of conformational change after binding. In rigid-body group, our method has overall accuracy and coverage values of 63% and 67%. In medium difficulty group, our method has overall accuracy and coverage values of 55% and 56%. In difficulty group, our method has overall accuracy and coverage values of 45% and 49%.

Table 4. Prediction Results by Our Method on Benchmark v4.0

		our method	
subset[a]	no. of cases	Acc	Cov
rigid body	123	63%	67%
medium difficult	29	55%	56%
difficult	24	45%	49%
overall	176	59%	63%

[a] Subset is based on magnitude of conformational change after binding.

Evaluation on CAPRI. We also examine on CAPRI targets, as shown in Table 5. CAPRI[56] is a community-wide experiment to assess capacity of protein−protein interface prediction methods. Our method has overall accuracy and coverage values of 50% and 49%.

Binding Sites Identification. Some existing methods use machine learning and statistical approaches to predict binding sites. Results show that our method performs better than other existing methods in binding sites prediction.

Table 5. Prediction Results by Our Method on CAPRI Targets

	our method			our method	
target	Acc	Cov	target	Acc	Cov
T01	11%	14%	T21	82%	82%
T02	85%	72%	T22	87%	80%
T03	8%	13%	T23	64%	63%
T04	43%	31%	T24	54%	53%
T05	29%	12%	T25	62%	58%
T06	35%	11%	T26	84%	90%
T07	11%	25%	T27	65%	62%
T08	41%	20%	T29	76%	76%
T09	45%	49%	T30	26%	27%
T10	42%	39%	T32	39%	39%
T11	63%	63%	T35	41%	37%
T12	76%	75%	T36	41%	38%
T13	82%	81%	T37	45%	66%
T14	18%	19%	T39	77%	79%
T15	59%	60%	T40	44%	51%
T18	29%	41%	T41	79%	75%
T19	61%	52%	T42	23%	35%
T20	26%	19%			

Comparison to metaPPI, meta-PPISP, and PPI-Pred. In this experiment, we compare our method to metaPPI, meta-PPISP, and PPI-Pred, as shown in Table 6. Test data consists of 41 complexes by metaPPI,[39] divided into two categories: enzyme−inhibitor (E-I) and others. Our method has overall accuracy and coverage values of 63% and 61%, which improves upon overall accuracy and coverage values of 49% and 28% for metaPPI, 46% and 38% for meta-PPISP, and 36% and 38% for PPI-Pred.

Comparison to ProMate and PINUP. Our method is compared to ProMate and PINUP, as shown in Table 7. The test data are originally used by ProMate,[29] including 57 unbound proteins and their complexes. Our method has overall accuracy and coverage values of 55% and 60%, which improves upon overall accuracy and coverage values of 45% and 42% for PINUP, and 53% and 13% for ProMate.

Comparison to core-SVM. We compare our method to core-SVM with 50 dimers,[31] as shown in Table 8. Our method has overall accuracy and coverage values of 59% and 61%, which improves upon overall accuracy and coverage values of 56% and 61% for core-SVM.

■ CONCLUSION

In biological processes, a lot of proteins achieve biological functions through protein−protein interactions. How to build more effective models based on sequence and structural information is a key question for predicting interface residues. In this work, we present a novel method for identifying protein−protein interactions, based on typical co-evolutionary

DOI: 10.1021/acs.jcim.5b00320
J. Chem. Inf. Model. 2015, 55, 2042−2049

Table 6. Comparison to metaPPI, meta-PPISP, and PPI-Pred

	our method		metaPPI		meta-PPISP		PPI-Pred	
type	Acc	Cov	Acc	Cov	Acc	Cov	Acc	Cov
E-I[a]	73%	65%	61%	37%	56%	55%	46%	47%
others	55%	57%	41%	22%	39%	26%	29%	31%
overall	63%	61%	49%	28%	46%	38%	36%	38%

[a] E-I is type of enzyme−inhibitor.

Table 7. Comparison to PINUP and ProMate

	our method		PINUP		ProMate	
	Acc	Cov	Acc	Cov	Acc	Cov
overall	55%	60%	45%	42%	53%	13%

Table 8. Comparison to core-SVM

	our method		core-SVM	
	Acc	Cov	Acc	Cov
overall	59%	61%	56%	61%

information. We propose a non-redundant database to detect correlated mutation at the interface. First, we construct structure alignments for one input protein, based on all aligned proteins in the database. Evolutionary distance matrices, one for each input protein, can be calculated through geometric similarity and evolutionary information. Then, we use evolutionary distance matrices to estimate correlation coefficient between each pair of fragments from two input proteins. Finally, we extract interacting residues with high values of correlation coefficient, which can be grouped as interacting patches.

Our method utilizes sequence and structural information to calculate evolutionary relationship between two input proteins; however, most of existing technologies only analyze sequence information. Experiments illustrate that our method achieves better results than some existing co-evolution-based methods. Applied to SK/RR interaction between sensor kinase and response regulator proteins, our method has accuracy and coverage values of 53% and 45%, which improves upon accuracy and coverage values of 50% and 30% for DCA method. We evaluate interface prediction on four protein families, and our method has overall accuracy and coverage values of 34% and 30%, which improves upon overall accuracy and coverage values of 27% and 21% for PIFPAM. On Benchmark v4.0, our method has overall accuracy and coverage values of 59% and 63%. On CAPRI targets, our method has overall accuracy and coverage values of 50% and 49%. Comparing to existing methods, our method improves overall accuracy value by at least 2%.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00320.

> Information about prediction of testing complexes and software packages, which are available for download from https://sites.google.com/site/guofeics/correlation (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: fguo@tju.edu.cn.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by a grant from National Science Foundation of China (NSFC 61402326).

REFERENCES

(1) Skerker, J. M.; Perchuk, B. S.; Siryaporn, A.; Lubin, E. A.; Ashenberg, O.; Goulian, M.; Laub, M. T. Rewiring the specificity of two component signal transduction systems. *Cell* **2008**, *133*, 1043−1054.

(2) Al-Khayyal, F. Jointly constrained bilinear programs and related problems: an overview. *Computers and Mathematics with Applications* **1990**, *19*, 53−62.

(3) Bren, U.; Oostenbrink, C. Cytochrome P450 3A4 inhibition by ketoconazole: tackling the problem of ligand cooperativity using molecular dynamics simulations and free-energy calculations. *J. Chem. Inf. Model.* **2012**, *52*, 1573−1582.

(4) Bren, U.; Fuchs, J.; Oostenbrink, C. Cooperative binding of aflatoxin B1 by cytochrome P450 3A4: a computational study. *Chem. Res. Toxicol.* **2014**, *27*, 2136−2147.

(5) Casari, G.; Sander, C.; Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **1995**, *2*, 171−178.

(6) Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **2009**, *138*, 774−786.

(7) De Juan, D.; Pazos, F.; Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **2013**, *14*, 249−261.

(8) Hopf, T. A.; Colwell, L. J.; Sheridan, R.; Rost, B.; Sander, C.; Marks, D. S. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **2012**, *149*, 1607−1621.

(9) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 67−72.

(10) Pazos, F.; Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng., Des. Sel.* **2001**, *14*, 609−614.

(11) Tress, M. L.; Valencia, A. Predicted residue-residue contacts can help the scoring of 3D models. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1980−1991.

(12) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **2011**, *6*, e28766.

(13) Nugent, T.; Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1540−E1547.

(14) Edgar, R. S.; et al. Peroxiredoxins are conserved markers of circadian rhythms. *Nature* **2012**, *485*, 459−464.

(15) Ochoa, D.; Pazos, F. Studying the co-evolution of protein families with the MirrorTree web server. *Bioinformatics* **2010**, *26*, 1370−1371.

(16) Juan, D.; Pazos, F.; Valencia, A. High-confidence prediction of global interactomes based on genome-wide co-evolutionary networks. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 934−939.

(17) Korber, B. T.; Farber, R. M.; Wolpert, D. H.; Lapedes, A. S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 7176−7180.

(18) Schug, A.; Weigt, M.; Onuchic, J. N.; Hwa, T.; Szurmant, H. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 22124−22129.

(19) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-coupling analysis of residue co-evolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, E1293−E1301.

(20) Jones, D. T.; Buchan, D. W. A.; Cozzetto, D.; Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **2012**, *28*, 184−190.

(21) Reynolds, K. A.; McLaughlin, R. N.; Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **2011**, *147*, 1564−1575.

(22) Pierce, B.; Wiehe, K.; Hwang, H.; Kim, B.; Vreven, T.; Weng, Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **2014**, *30*, 1771−1773.

(23) Fernández-Recio, J.; Totrov, M.; Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.* **2004**, *335*, 843−865.

(24) Pierce, B.; Weng, Z. A combination of rescoring and refinement significantly improves protein docking performance. *Proteins: Struct., Funct., Genet.* **2008**, *72*, 270−279.

(25) Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **2004**, *20*, 45−50.

(26) Schueler-Furman, O.; Wang, C.; Baker, D. Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins: Struct., Funct., Genet.* **2005**, *60*, 187−194.

(27) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003**, *125*, 1731−1737.

(28) Carl, N.; Konc, J.; Vehar, B.; Janežič, D. Protein-protein binding site prediction by local structural alignment. *J. Chem. Inf. Model.* **2010**, *50*, 1906−1913.

(29) Neuvirth, H.; Raz, R.; Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **2004**, *338*, 181−199.

(30) Bradford, J. R.; Westhead, D. R. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* **2005**, *21*, 1487−1494.

(31) Li, N.; Sun, Z.; Jiang, F. Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC Bioinf.* **2008**, *9*, 553.

(32) Liang, S.; Zhang, C.; Liu, S.; Zhou, Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* **2006**, *34*, 3698−3707.

(33) Zhou, H.; Qin, S. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics* **2007**, *23*, 2203−2209.

(34) Wass, M. N.; David, A.; Sternberg, M. J. E. Challenges for the prediction of macromolecular interactions. *Curr. Opin. Struct. Biol.* **2011**, *21*, 382−390.

(35) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* **2005**, *33*, W337−W341.

(36) Konc, J.; Janežič, D. ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res.* **2010**, *38*, W436−W440.

(37) Konc, J.; Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160−1168.

(38) Qin, S.; Zhou, H. X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* **2007**, *23*, 3386−3387.

(39) Huang, B.; Schröder, M. Using protein binding site prediction to improve protein docking. *Gene* **2008**, *422*, 14−21.

(40) Wood, T.; Pearson, W. Evolution of protein sequences and structures. *J. Mol. Biol.* **1999**, *291*, 977−995.

(41) Cho, K.; Kim, D.; Lee, D. A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res.* **2009**, *37*, 2672−2687.

(42) Xu, D.; Tsai, C. J.; Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng., Des. Sel.* **1997**, *10*, 999−1012.

(43) Goh, C. S.; Bogan, A. A.; Joachimiak, M.; Walther, D.; Cohen, F. E. Coevolution of proteins with their interaction partners. *J. Mol. Biol.* **2000**, *299*, 283−293.

(44) Wang, S.; Ma, J.; Peng, J.; Xu, J. Protein structure alignment beyond spatial proximity. *Sci. Rep.* **2013**, *3*, 1−7.

(45) Holm, L.; Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **1993**, *233*, 123−123.

(46) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.

(47) Finn, R. D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E. L.; Tate, J.; Punta, M. The Pfam protein families database. *Nucleic Acids Res.* **2007**, *36*, D281−D288.

(48) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(49) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10915−10919.

(50) Zheng, W. M.; Liu, X. A protein structural alphabet and its substitution matrix CLESUM. *Transactions on Computational Systems Biology II* **2005**, 59−67.

(51) Van Walle, I.; Lasters, I.; Wyns, L. SABmark - a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* **2005**, *21*, 1267−1268.

(52) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-protein docking benchmark version 4.0. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 3111−3114.

(53) Baldassi, C.; Zamparo, M.; Feinauer, C.; Procaccini, A.; Zecchina, R.; Weigt, M.; Pagnani, A. Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* **2014**, *9*, e92721.

(54) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536−540.

(55) Xie, B.; Chen, X.; Zhang, X.; He, H.; Zhang, Y.; Zhou, B. Predicting protein interaction interfaces from protein sequences: case studies of subtilisin and phycocyanin. *Proteins: Struct., Funct., Genet.* **2008**, *71*, 1461−1474.

(56) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M.; Vajda, S.; Vakser, I.; Wodak, S. CAPRI: a critical assessment of predicted interactions. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 2−9.