http://www.jstage.jst.go.jp/browse/biophysics

BIOPHYSICS

Vol. 3, pp. 13–26 (2007)
doi: 10.2142/biophysics.3.13

Prediction of interacting proteins from homology-modeled complex structures using sequence and structure scores

Naoshi Fukuhara[1], Nobuhiro Go[1,2] and Takeshi Kawabata[1,3]

[1]*Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan*
[2]*Neutron Biology Research Center, Quantum Beam Science Directorate, Japan Atomic Energy Agency, 8-1 Umemidai, Kizu, Soraku, Kyoto, 619-0215, Japan*
[3]*CREST, JST*

Received 15 February, 2007; accepted 31 May, 2007

Protein-protein interactions support most biological processes, and it is important to find specifically interacting partner proteins among homologous proteins in order to elucidate cellular functions such as signal transduction systems. Various high-throughput experimental methods for identifying these interactions have been invented, and used to generate a huge amount of data. Because these experiments have been applied to only a few organisms, and their accuracy is believed to be limited, it would be valuable to develop computational methods for predicting protein-protein interactions from their amino acid sequences or tertiary structural information. In this study, we describe a prediction method of interacting proteins based on homology-modeled complex structures. We employed the statistical residue-residue contact energy used in a previous study, and two types of new scores, simple electrostatic energy and sequence similarity between target sequences and template structures. The validity of each protein-protein complex model was measured using their single and combined scores. We applied our method to all the protein heterodimers of *Saccharomyces cerevisiae*. To evaluate the prediction performance of our method, we prepared two types of protein-protein interaction dataset: a complete dataset and high confidence dataset. The complete dataset (10,325 protein dimer models) contains all the yeast protein heterodimers whose complex structures can be modeled. Among them, pairs registered in the DIP database are defined as interacting pairs, and those not registered are defined as non-interacting protein pairs. The high confidence dataset (3,219 protein dimer models) is a more reliable subset of the complete dataset extracted using the criteria of the common subcellular localization. Both datasets show that sequence similarity has a much higher discrimination power than the other structure-based scores, but that the inclusion of contact energy results in significant improvement over predictions using sequence similarity alone. These results suggest that the sequence similarity is indispensable for the prediction, whereas structure scores can play supporting roles.

Key words: protein-protein interaction,
homology modeling, binding specificity,
sequence similarity, contact energy

Protein-protein interactions support most important cellular functions, such as signal transduction, enzymatic activities, replication and translation. Recently, high-throughput screening methods, such as yeast-two-hybrid (Y2H) and tandem affinity purification (TAP), have generated large datasets of protein-protein interactions[1–6]. These interaction data are compiled in databases such as DIP, MIPS and BIND, which also contain data obtained by classical "low-throughput" methods[7–9].

The high-throughput genome-wide screening experiments provide us with rich information about cellular processes.

Corresponding author: Takeshi Kawabata, Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan.
e-mail: takawaba@is.naist.jp

Because these techniques are costly and labor-intensive, however, these experiments have been performed only for a few organisms (e.g., *Saccharomyces cerevisiae*), even though complete genome sequences for more than two hundred organisms have been determined to date. To fill the gap between the vast amount of genome sequence data and the relatively smaller scope of interaction data, many researchers have worked to develop methods for computational prediction of protein-protein interactions from their amino acid sequences[10,11].

Various approaches have been proposed to predict protein-protein interactions, such as gene fusion methods[12,13], phylogenetic profiling methods[14], co-evolution methods[15,16], and homologous interaction methods[17–21]. Recently, several researchers proposed prediction methods based on 3D structures of protein-protein complexes[22–26]. These studies employed a common standard procedure. First, a structure of the two target proteins in complex is generated by comparative-modeling methods. For example, Alloy and Russell employed BLAST to find template structures for homology modeling; Lu *et al.* used a threading program developed by them for modeling multimers. In contrast to the residue-level coarse-grained models in these studies, Davis *et al.* used full-atomic models obtained from MODBASE[27]. Second, the validity of modeled structures is evaluated by interaction energies. Knowledge-based residue-residue contact energies were employed by each of these three studies cited above. Third, interaction energies are evaluated by applying various statistical scores. Alloy and Russell and Davis *et al.* employed the Z-score, using randomly shuffled sequences as the reference. Lu *et al.* also used the Z-score, but their reference state was a set of scores of all the template structures in a library. The prediction accuracies of all these studies were mainly confirmed by the overlaps with experimentally determined interactions. False predicted interactions have not been evaluated as extensively.

In this study, we also employed a structure-based approach, but we evaluated our predictions by discriminating between interacting and non-interacting protein pairs. In other words, we mainly focused on the interaction specificities among homologous protein pairs. We chose to do so because the specific interactions among similar homologous proteins are important for many cellular functions. There are many paralogous protein domains in eukaryotic genomes, and each has its own set of specific interacting partners. Proteins working in signal transduction pathways, especially protein kinases, G-proteins and transcription factors, have many similar homologues within genomes[28]. Binding specificities of these proteins are the basis of a complicated and robust signal transduction systems within the cell[29].

One of the problems for evaluating reliability and coverage of predictions is that there is no gold standard for discriminating interacting and non-interacting protein pairs. This problem arises in part because high-throughput experiments of protein-protein interaction are believed to contain unreliable or inaccurate data[30–32]. Specifically, there is no gold standard for unambiguously defining non-interacting protein pairs. In this study, we prepared two types of dataset comprising interacting and non-interacting protein pairs: the “complete” dataset and the “high confidence” dataset. The complete dataset contains all the protein heterodimers whose complex structure can be modeled. Protein pairs registered in the DIP database are defined as interacting pairs, while those not registered are defined as non-interacting protein pairs. We expect these assumptions are safe for *Saccharomyces cerevisiae*, because the yeast is the most popular model organism for protein-protein interactions, and a huge amount of experimental data has been accumulated to date. However, the DIP database may contain both false positive data (i.e., protein pairs registered as interacting that do not, in fact, interact) and false negative data (unregistered protein pairs that actually interact in the cell). To evaluate our method more accurately, we therefore prepared the high confidence dataset, which is a more reliable subset of the complete dataset extracted using subcellular localization data. Recently, genome-wide analyses determining subcellular localization of yeast have been published[33–35]. We used data from these analyses to determine whether the proteins in each registered interacting pair share a common localization; if so, we regarded as interacting pair as reliable and included it in the high confidence dataset[32,36,37]. The performance of our method was evaluated by discriminating interacting and non-interacting protein pairs, using both the complete and high confidence dataset.

The outline of our prediction method is as follows. First, we predict the dimer structure of two target proteins by a homology modeling method. Sequence homology searches for the two target protein sequences are run against the sequence library of the component proteins of known dimer structures. If we find a dimer template structure that is composed of two proteins homologous to each target protein, a complex structure of the target proteins are modeled based on the template. To evaluate the validity of structure models, we employed three kinds of scores. First, we used knowledge-based residue-residue contact energy, which is used in each of the three previous studies discussed above. Second, because we expected long-range interaction between protein pairs and binding specificities to be provided by electrostatic interactions, we introduced a simple electrostatic energy. Third, we also employed a score based on sequence similarity between target and template proteins. The sequence similarity for interacting protein pairs has often been used in sequence-based predictions[17-21]; to date, however, it has not been used in combination with structural features. All three scores were transformed to a Z-score using randomized sequences as a reference. In contrast to previous studies, we analytically estimated the average and variance of energies. The performance using each of the three scores, both individually and in combination, was evaluated by recall-precision plots and maximum F-measure using the complete and high confidence datasets.

Materials and methods

Datasets of heterodimer structures

Datasets of heterodimer structures are required for the library of template structures and for estimating values for statistical contact energies. We excluded the homodimers (pairs of identical proteins) because homodimeric crystal structures of single proteins are less reliable due to artificial crystal packing[38]. Heterodimers were defined as the proteins whose sequence identity is smaller than 50%. These sets comprise non-redundant representative tertiary structural data of heterodimers obtained from the PQS server[39]. The PQS server contains putative biological units of quarternary structures determined by X-ray crystallography, which are automatically chosen among the candidate complex structures generated by crystallographic symmetry operations of PDB data. The heterodimers datasets were generated by the following procedures: First, all the multimers included in the PQS server were separated into dimers. Dimers with fewer than five interacting residues (defined as a residue that has at least one Cβ atom located within 7 Å of Cβ atoms of another protein chain) were removed. Second, these dimers were clustered by single-linkage clustering algorithm[40] according to similarities between dimers, defined as the lower sequence of the two sequence similarities between corresponding proteins. One representative dimer with the largest number of interacting residues was extracted from each cluster. We used the structural data from PQS (version of April 14, 2006). Two types of representative dataset were prepared using different threshold values of similarity of complexes. The former set comprises 1,687 heterodimers generated by the threshold of 40% similarity and is used as the dataset for calculation of contact energy; the latter comprises 2,635 heterodimers generated by the threshold of 95% similarity and used as the template structure library for the homology modeling.

Building complex structure models of yeast hetero protein pairs

From the UniProt ver. 49.4 database[41], which is a curated protein sequence database with a high level of annotation, we extracted 5,314 *Saccharomyces cerevisiae* amino acid sequences. All the hetero pairs of the 5,314 yeast protein sequences were subjected to interaction prediction. To construct the sequence profile of each yeast amino acid sequence, PSI-BLAST was run against the nr database (version of September 22, 2006). The threshold for E-value (expected hits) was set to 0.001, and the number of iterations was set to three. Using the generated sequence profiles, we ran PSI-BLAST[42] against the template structure library described above. For each target protein pair, we checked whether a dimer template structure consisting of two homologous proteins of each target protein exists in the database. If a dimer template structure was found for the target protein pairs, we required that the following conditions be met: (1) In the two alignments between target protein sequences and template complex, ratios of aligned interacting residues must not be smaller than 50%. (2) The numbers of aligned interacting residues must not be smaller than 10. If several template dimer structures were found, we selected the template whose lowest sequence identity is the highest among the template dimers. In this study, because a fast modeling method is necessary in order to allow us to deal with a large number of protein pairs, we use the conformation of aligned residues from the template structures, ignoring inserted residues, and did not build in side chain atoms for substituted residue.

Interacting and non-interacting protein pairs

The generated complex structure models were labeled either as “interacting” or “non-interacting” protein pairs. We prepared two types of the dataset using different criteria of interaction. In the “complete” dataset, if a protein pair of complex model is registered in protein-protein interaction databases, the pair is considered as an interacting pair. If it is not registered, it is considered as a non-interacting pair. Among many available protein-protein interaction databases, we chose the DIP database, because it contains data obtained via a wide range of experimental methods, such as yeast two hybrid, tandem affinity purification, affinity chromatography, *in vitro* binding, copurification, complex structures by X-ray crytallography. We used the dataset version of January 16, 2006. Although the DIP database contains a huge number of protein-protein interaction data, several latest experimental results are not yet registered. If we found complex template structures of almost identical (more than 95% sequence identity) proteins to target protein pairs, we relabeled these pairs as “interacting” pairs even if they are not registered in the DIP database, considering these experimentally determined complex structures as sufficiently well-supported to justify registration in the DIP database.

The complete dataset assumes all the interactions are already registered; however, high-throughput experiments of protein-protein interaction are believed to contain unreliable or inaccurate data, and protein pairs not registered in the DIP database may interact in the cell. To increase the reliability of the dataset, we prepared a “high confidence” dataset, a more reliable subset of the complete dataset extracted using subcellular localization information. Subcellular localization data was downloaded from the MIPS database (version of November 14, 2005), where one or more localized compartment types are assigned to each yeast protein. Localized compartment types consist of 19 types: extracellular, bud, cell wall, cell surface, plasma membrane, inner membrane, cytoplasm, cytoskeleton, endoplasmic reticulum, golgi body, transport vesicle, nuclear, mitochondria, peroxisome, endosome, vacuole, microsome, lipid particle, and other subcellular localization. The ratios of localized compartment types for the 5,211 registered yeast proteins are 34% for nuclear, 18% for cytoplasm, 18% for mitochondria, 5% for vacuole, 5% for endoplasmic reticulum, 4% for unknown, 2% for transport vesicle and 14% for other localized compartment types.

Protein pairs registered in the DIP database sharing at least one localizing compartment are selected for the high confident interacting protein pairs, those not registered in DIP not sharing any localizing compartments are selected for the high confident non-interacting protein pairs. The assumption of the high confidence dataset is that two proteins having different subcellular localizations do not interact each other, whereas reported pairs with similar localizations certainly interact in the cell.

Residue-residue statistical contact energy for protein-protein interaction

Residue-residue statistical contact energies were originally developed for coarse-grained models of protein folding and threading[43–45]. Recently, similar approaches were applied for evaluating protein-protein interaction[46,47]. In this study, we employed a typical log-odds formula for extracting the value of contact energies. A statistical contact energy $e_{con}(a, b)$ for contacting residues a and b in different polypeptide chains is estimated by the form of the log-odds score:

$$e_{con}(a, b) = -\log \frac{Q(a, b)}{P(a)P(b)} \qquad (1)$$

where $P(a)$ and $P(b)$ are the probabilities that amino acids a and b appear on the surface, $Q(a, b)$ is the probability that amino acids a and b on the surface contact each other in the protein-protein interface. Surface residues of a protein are defined as those residues whose relative accessible surface areas are larger than 35%. Contacting residue pairs are defined as the residues in different chains, whose Cβ atoms are located within 7 Å of one another. Both probabilities are estimated using the dataset for calculation of contact energy (see Datasets of heterodimer structures). If the interface contacts between residues a and b are often found in the interface, the value of $e_{con}(a, b)$ is large and negative.

The estimated energy values are summarized in Figure 1. Hydrophobic residues are attractive to each other, especially in the case of the cysteine-cysteine pair. Hydrophilic residues, however, are generally repulsive even for differently charged residue pairs, such as the arginine-glutamic acid pair. These features are similar to those employed in previous studies[46,47].

The total contact energy E_{con} is the sum of the e_{con} for all the contacted residue pairs including both surface and buried residues:

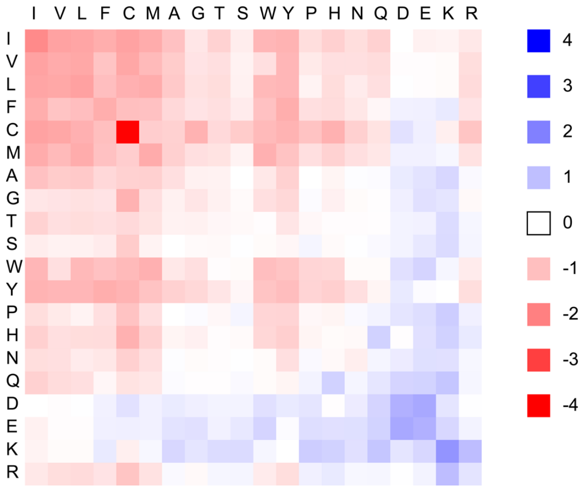

Figure 1 Residue-residue statistical contact energy in protein-protein interfaces. In the horizontal and vertical axes, 20 amino acids are arranged in descending order of hydrophobicity. Energy values are represented from red (low energy) to blue (high energy).

$$E_{con} = \sum_{i,j\ (i \text{ contacts with } j)}^{N,M} e_{con}(a_i, a_j) \qquad (2)$$

where N and M are the total number of the residues of proteins, and a_i and a_j are the amino acids of residues i and j.

Electrostatic energy for protein-protein interaction

Electrostatic interactions also play an important role in protein-protein interactions[48]. To validate our dimer models, we employed simplified electrostatic energies as proposed by Shaul and Schreiber[49]. An electrostatic energy e_{ele} between charges q_1 and q_2 is calculated by the following equation based on the Debye-Huckel theory:

$$e_{ele}(r, q_1, q_2) = \frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{q_1 q_2}{r} \frac{e^{-\kappa(r-a)}}{1+\kappa a} \qquad (3)$$

where ε_r is the relative permittivity of water (=80). The variable r is a distance between the charges q_1 and q_2, and κ is Debye-Huckel screening parameter (=0.488 Å^{-1}). The parameter a is set to 6 Å.

The total electrostatic energy E_{ele} is the sum of the e_{ele} for all of the charged atom pairs:

$$E_{ele} = \sum_{i}^{N} \sum_{j}^{M} \sum_{s \in Q_i} \sum_{t \in Q_j} e_{ele}(r_{st}, q_s(a_i), q_t(a_j)) \qquad (4)$$

where i and j are residues included in different proteins. The numbers N and M are the total number of residues, and Q_i and Q_j are the sets of charged atoms belonging to the residue i and j. The variable r_{st} is a distance between atom s and t. The variable $q_s(a_i)$ is the charge of the atom s of amino acid a_i.

Formal charges are assigned to the atoms in the modeled complex structure: charge = −1 for aspartic acid and glutamic acid, and charge = +1 for lysine and arginine. To

Table 1 Atoms of amino acids where charges can be assigned

Residue	Atom	Residue	Atom	Residue	Atom
GLU	OE1	TRP	CE3	SER	OG
GLU	OE2	TYR	OH	ILE	CD1
ASP	OD1	PHE	CZ	MET	CE
ASP	OD2	GLN	OE1	LEU	CD1
ARG	NH1	GLN	NE2	LEU	CD2
ARG	NH2	ASN	OD1	VAL	CG1
LYS	NZ	ASN	ND2	VAL	CG2
HIS	ND1	CYS	SG	ALA	CB
HIS	NE2	THR	OG1	PRO	CG
TRP	CE2	THR	CG2	GLY	CA

PDB atomic names are shown. These atoms are mainly taken from Shaul and Schreiber's charge rules (Shaul and Schreiber, 2005). Atoms of proline and glycine have been added; OXT and the N-terminus atom have been removed.

assign the charges for the model structures, we employed the charge rule proposed by Shaul and Schreiber[49]. For a substituted residue of the target sequence, total charges of the residue are equally assigned to the position of the selected atoms of the corresponding residue on the template structure. The location of the pseudo-charge on the amino acids is given in Table 1. For example, if the amino acid of the target protein is glutamic acid, and the corresponding amino acid of the template structure is threonine, a charge −0.5 is assigned to both OG1 and CG2 atoms of the threonine residue.

Normalization of the energies

A Z-score is introduced to normalize the contact and electrostatic energies, and to remove biases of amino acid compositions of target proteins[22–25]. The Z-score for energy E is defined as follows:

$$Z(E) = \frac{E - Mean\,[E]}{\sqrt{Var\,[E]}} \tag{5}$$

where $Mean[E]$ and $Var[E]$ are the average and variance of E respectively for randomly shuffled amino acids sequences of the same composition. Z-score shows how many units of the standard deviation an energy of a protein pair is above or below the average by the random shuffling. Calculation of the averages and variances of the contact energy and electrostatic energy are described in the following sections. In contrast to studies by other groups, we analytically estimated the average and variance of energies without explicitly generating randomly shuffled sequences.

Mean and variance of contact energy for randomly shuffled sequences

We assume that random contacting amino acid pairs are generated by picking up two amino acids randomly from the surfaces of different proteins. For this random set of contacting amino acids, the average μ_{con} and variance σ^2_{con} of the contact energy are calculated as follows:

$$\mu_{con} = \sum_{a\in A}\sum_{b\in A}\{e_{con}(a, b)\cdot P(a)\cdot P(b)\}, \tag{6}$$

$$\sigma^2_{con} = \sum_{a\in A}\sum_{b\in A}\{e^2_{con}(a, b)\cdot P(a)\cdot P(b)\} - \mu^2_{con} \tag{7}$$

where $P(a)$ and $P(b)$ are the proportions of amino acid a and b in surface residues for each protein, and A is the set of 20 genetically encoded amino acids. If we assume that the all the contacting protein pairs are independent in the shuffling process, the average and variance of the total contact energy E_{con} are calculated as follows:

$$Mean\,[E_{con}] = \mu_{con}\cdot N_{contact}, \tag{8}$$

$$Var\,[E_{con}] = \sigma^2_{con}\cdot N_{contact} \tag{9}$$

where $N_{contact}$ is the total number of the contacting residues.

Mean and variance of electrostatic energy for randomly shuffled sequences

The average and variance of the electrostatic energy can be calculated in a similar way to that of the contact energy. We assume that random contacting amino acid pairs on the i-th and j-th positions of proteins are generated by picking up two amino acids randomly from the surfaces of different proteins. The average $\mu_{ele}(i, j)$ and variance $\sigma^2_{ele}(i, j)$ values of the electrostatic energy for the random sets are calculated as follows:

$$\mu_{ele}(i, j) = \sum_{a\in A}\sum_{b\in A}P(a)P(b)\sum_{s\in Q_i}\sum_{t\in Q_j}e_{ele}(r_{st}, q_s(a), q_t(b)), \tag{10}$$

$$\sigma^2_{ele}(i, j) = \left(\sum_{a\in A}\sum_{b\in A}P(a)P(b)\sum_{s\in Q_i}\sum_{t\in Q_j}e^2_{ele}(r_{st}, q_s(a), q_t(b))\right) - \mu^2_{ele}(i, j) \tag{11}$$

where the variable r_{st} is the distance between atom s and t. The variables $q_s(a)$ and $q_t(b)$ are the charges of the atoms s and t of the i-th and j-th residues when they are replaced by amino acids a and b. $P(a)$ and $P(b)$ are the frequencies of amino acids a and b in surface residues for each protein. Q_i and Q_j are the set of charged atoms belonging to the residues i and j. If we assume that the all the protein pairs are independent in the shuffling process, the average and variance of the total electrostatic energy E_{ele} are calculated as the sum of average and variance of each amino acids pairs:

$$Mean\ [E_{ele}] = \sum_{i}^{N} \sum_{j}^{M} \mu_{ele}(i, j), \tag{12}$$

$$Var\ [E_{ele}] = \sum_{i}^{N} \sum_{j}^{M} \sigma^2_{ele}(i, j) \tag{13}$$

where N and M are the total numbers of residues in each protein.

Sequence similarity between target and template

We employed sequence similarity between target protein and template protein as another feature for finding interacting proteins. We expected that two proteins will interact with each other if they have close homologues whose dimer structures have been experimentally determined. Here a Z-score is also introduced to measure sequence similarities. In this case, the number of identical residues N_{iden} in the alignment is normalized by average and variance values for randomly shuffled sequences:

$$Z(N_{iden}, N_{comp}) = -\frac{N_{iden} - N_{comp}p}{\sqrt{N_{comp}p(1-p)}} \tag{14}$$

where N_{iden} is the number of the identical residues, and N_{comp} is the number of compared residues in the alignment with gaps removed. We assume that random shuffling is applied using the uniform distribution of amino acids (p is set to 1/20), and that the number of identical residue N_{iden} obeys the binominal distribution. Because the other two Z-scores of energies have negative value for probable interfaces, the Z-score for sequence similarity was multiplied by minus one to facilitate comparison. Because we are modeling dimer structures, two different sequence similarities are obtained for one protein complex. We employed the higher score (in other words, the lower sequence similarity) for the purposes of discrimination.

The random shuffling process for sequence similarity is subtly different from that of contact and electrostatic energy. For contact and electrostatic energy, two amino acids on the surface are randomly chosen. In the case of the sequence similarity, the sequence of the template protein is fixed, and the sequence of the target protein is randomly generated using a uniform distribution of amino acids.

Evaluation by recall-precision plots

To evaluate the discriminating powers between the interacting and non-interacting protein pairs, recall-precision plots were generated. Recall and precision are defined as follows:

$$Recall(S) = \frac{N_{tp}(S)}{N_t}, \tag{15}$$

$$Pecision(S) = \frac{N_{tp}(S)}{N_p(S)} \tag{16}$$

where $N_{tp}(S)$ is the number of interacting protein pairs with a score better than S, N_t is the number of interacting protein pairs and $N_p(S)$ is the number of pairs with a score better than S. Recall shows how many correct interactions are covered by the prediction, precision shows how reliable the prediction is. Recall and precision were calculated against all of the observed scores and plotted as a line on the plane. The line plotted more towards the upper right has larger Recall and Precision values than those toward the lower left. Generally speaking, predictions with high Recall value tend to have a low value of Precision. Thus, the maximum F-measure is introduced to find a good balance point between recall and precision. F-measure $F(S)$ is defined as the harmonic mean of recall and precision, and the maximum F-measure $F_{\max}$ is the largest F-measure among all of the observed scores:

$$\begin{aligned} F(S) &= 2\left(\frac{1}{Recall(S)} + \frac{1}{Precision(S)}\right)^{-1} \\ &= \frac{2N_{tp}(S)}{N_t + N_p(S)}, \end{aligned} \tag{17}$$

$$F_{\max} = \max_{S}\ [F(S)] \tag{18}$$

Results and Discussion

Homology-modeled dimer structures of the interacting and non-interacting protein pairs

We modeled dimer structures of hetero protein pairs of *Saccharomyces cerevisiae* by the homology-modeling method. 10,325 models of protein pairs were generated; among them, 417 pairs were regarded as interacting, and 9,908 pairs were regarded as non-interacting. We call these pairs the complete dataset of protein-protein interaction. To select reliable data, the complete dataset is classified into three types of protein pairs: (i) Two proteins share at least one common localized compartment type. (ii) Subcellular localization of at least one protein is unknown. (iii) Two proteins do not share any localized compartment type. The classification is shown in Table 2. The interacting pairs in the complete dataset sharing at least one localized compartment are selected for the high confidence interacting pairs (380 pairs), and the non-interacting pairs not sharing any localized compartments are selected for the high confidence non-interacting pairs (2,839 pairs). Notably, the high confidence dataset contains only 37 fewer interacting pairs than the complete dataset, but 7,069 fewer non-interacting pairs. In other words, most of the protein pairs registered in DIP database have a similar localization, but there are many protein pairs that have a similar localization but nonetheless are not reported.

Network of the protein-protein interaction in the complete dataset

In order to have a full picture of these protein pairs, we

Table 2 The classification of interacting and non-interacting protein pairs included in the complete dataset by subcellular localization

		Interacting pairs	Non-interacting pairs
(i)	Two proteins share at least one common localized compartment	**380**	5,631
(ii)	Subcellular localization of at least one protein is unknown	10	1,438
(iii)	Two proteins do not share any localized compartments	27	**2,839**
	Total	417	9,908

The underlined numbers are for the complete dataset; bold numbers are for the high confidence dataset.

drew a network of protein-protein interaction in the complete dataset (Fig. 2). In this network, nodes correspond to target proteins and edges correspond to target protein pairs whose dimer structure can be modeled. There are 1,036 nodes and 10,325 edges in the network. As there are approximately twenty-four times more non-interacting than interacting pairs, most of the edges are colored in blue. The network was separated into 64 clusters by single linkage clustering. Our network was more sparse than those appearing in previous experimental studies[3,6], probably because we more stringently restricted the protein pairs that are able to be homology-modeled.

The largest cluster (Cluster A) has 573 proteins, and the second and third largest cluster (Cluster B and Cluster C) have 41 and 30 proteins, respectively. We focused on the target proteins included in Cluster A, and colored the nodes in the network according to the major domains included in Cluster A. Cluster A contains proteins involved in the signal transduction system. The numbers of the target proteins which include the domain of protein kinases catalytic subunit (green), WD40-repeat (cyan), G proteins (red), canonical RBD (yellow), ankyrin repeat (gray), cyclin (black) are 119, 97, 55, 50, 18 and 16, respectively. Cluster B contains proteins associated with ubiquitination, and consists of two major families: 17 domains of RING finger domain C3HC4 and 14 domains of ubiquitin conjugating enzyme UBC. Cluster C contains proteins involved in the DNA replication, and there were 23 domains of extended AAA ATPase, and 7 domains of DNA polymerase III clamp loader subunits C-terminal.

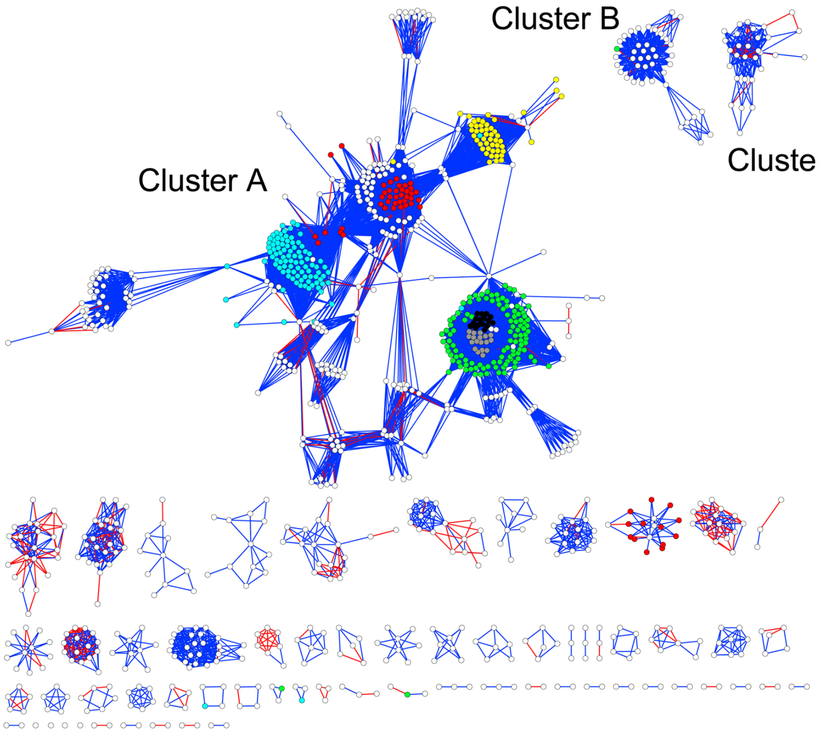

Figure 2 The protein-protein interaction network of the interacting and non-interacting protein pairs included in the complete dataset. The graph was visualized by Cytoscape[50]. The nodes correspond to the target proteins; edges correspond to interactions. The interacting protein pairs are shown in red, the non-interacting ones in blue. The proteins including the domains of protein kinase catalytic subunit, WD40-repeat, G proteins, canonical RBD, ankyrin repeat, cyclin are colored green, cyan, red, yellow, gray and black, respectively. If the target protein includes more than two domains from the six types of domains, the node is colored according to the domain nearest to the N-terminus. The SCOP, which is the structural classification database of proteins, was used for identifying the domains[51].

Table 3 Family pairs frequently appearing in template complexes

Family pairs of the template structures[a]	PDB[b]	Complete dataset		High confidence dataset	
		Inter[c]	Non-inter[d]	Inter[c]	Non-inter[d]
Top 10 family pairs of the interacting protein pairs					
1. b.38.1.1/b.38.1.1	1b34AB	33	24	33	0
2. d.153.1.4/d.153.1.4	1g65JK	30	44	30	0
3. h.1.15.1/h.1.15.1	1gl2BC	20	80	10	45
4. c.37.1.20-a.80.1.1/c.37.1.20-a.80.1.1	1sxjBC	19	95	19	15
5. d.144.1.7/a.74.1.1-a.74.1.1	1finAB	18	1662	14	559
6. c.3.1.3-d.16.1.6-c.3.1.3/c.37.1.8	1ukvGY	13	61	12	13
7. d.144.1.7/d.211.1.1	1bi7AB	10	1912	10	381
8. a.22.1.1/a.22.1.1	1id3AF	9	12	9	0
9. i.1.1.1/i.1.1.1	1s1hJN	8	16	8	9
10. a.116.1.1/c.37.1.8	1ow3AB	6	342	5	99
Top 10 family pairs of the non-interacting protein pairs					
1. d.144.1.7/d.211.1.1	1g3nAB	10	1912	10	381
2. a.74.1.1-a.74.1.1/d.144.1.7	1oiuBC	18	1662	14	559
3. c.37.1.8-a.66.1.1-c.37.1.8/b.69.4.1	1gotAB	1	530	1	321
4. a.116.1.1/c.37.1.8	1ow3AB	6	342	5	99
5. j.66.1.1/d.144.1.7	1f3mAC	1	319	1	112
6. c.10.2.4/d.58.7.1	1a9nAB	2	257	1	109
7. c.37.1.8/c.10.1.2	1k5dAC	4	239	3	87
8. c.45.1.1/d.144.1.7	1fq1AB	0	204	0	59
9. a.48.1.1-a.39.1.7-d.93.1.1-g.44.1.1/d.20.1.1	1fbvAC	3	189	3	38
10. a.118.1.1/c.37.1.8	1qbkBC	4	184	4	44

[a] SCOP ID included in the table are following; a.22.1.1:Nucleosome core histones, a.39.1.7:EF-hand modules in multidomain proteins, a.48.1.1:N-terminal domain of cbl (N-cbl), a.66.1.1:Transducin (alpha subunit) insertion domain, a.74.1.1:Cyclin, a.80.1.1:DNA polymerase III clamp loader subunits C-terminal domain, a.116.1.1:BCR-homology GTPase activation domain (BH-domain), a.118.1.1:Armadillo repeat, b.38.1.1:Sm motif of small nuclear ribonucleoproteins SNRNP, b.69.4.1:WD40-repeat, c.3.1.3:GDI-like N domain, c.10.1.2:Rna1p (RanGAP1) N-terminal domain, c.10.2.4:U2A′-like, c.37.1.8:G proteins, c.37.1.20:Extended AAA-ATPase domain, c.45.1.1:Dual specificity phosphatase-like, d.16.1.6:GDI-like, d.20.1.1:Ubiquitin conjugating enzyme UBC, d.58.7.1:Canonical RBD, d.93.1.1:SH2 domain, d.144.1.7:Protein kinases catalytic subunit, d.153.1.4:Proteasome subunits, d.211.1.1:Ankyrin repeat, g.44.1.1:RING finger domain C3HC4, h.1.15.1:SNARE fusion complex, i.1.1.1:Ribosome complexes, j.66.1.1:pak1 autoregulatory domain.

[b] PDB code of the template complexes.

[c] Number of interacting protein pairs.

[d] Number of non-interacting protein pairs.

To show frequently appearing families in the network more precisely, we show statistics for the family pairs of the template complexes according to the interacting and non-interacting pairs included in the complete and high confidence dataset (Table 3). The family pairs of the non-interacting protein pairs are more biased than those of the interacting pairs, and the biases are mostly caused by the six major families colored in the network of Figure 2. For example, in case of the complete dataset, protein kinase catalytic subunit domains and ankyrin repeat domains form as many as 1,912 non-interacting protein pairs. Similar biases were also observed in the high confidence dataset, although its observed numbers of family pairs are smaller.

Recently, researchers report that a protein-protein interaction network is a small world network, which is a network in which the length of the shortest path between any protein pairs tends to be small, but also has densely connected local neighborhood, and the number of interactions per proteins (degree) appears to follow a power law distribution[52,53]. Our non-interacting protein network was not a small world network, because its average length of the shortest path was not small (proteins are clustered into the 64 clusters), and number of interaction per proteins of our network did not follow a power law distribution (number of proteins with degree = 12 was larger than that of degree = 1). The deviation from the power law distribution was caused by the biased family distribution of non-interacting network.

Score distributions of the complete dataset for each feature

The Z-score distributions of three features (contact energy, electrostatic energy and sequence similarity between target and template) of the complete dataset are shown in Figure 3–5. As we assume that similar random surface amino acid pairs are generated in Z-score calculations of both contact and electrostatic energy, these Z-scores are comparable to each other. Z-scores for the contact energy ranged lower, and were distributed more widely, than Z-score for the electrostatic energy. The averages of Z-score of the contact energy for interacting and non-interacting protein pairs were −4.6 and −2.2, respectively, whereas those for the electrostatic energy were −0.77 and −0.15. The variances of the contact energies are 7.6 (interacting) and 4.6 (non-interacting) and those of the electrostatic energies are 0.99 (interacting) and 0.67 (non-interacting). As the differences of the averages between the interacting and non-interacting

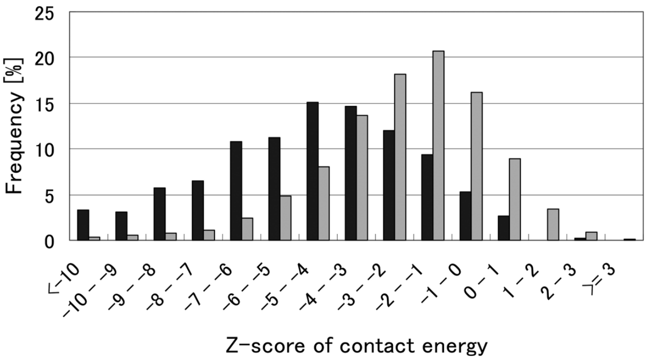

Figure 3 Distributions of Z-scores of contact energy calculated for protein pairs included in the complete dataset. Black and gray bars correspond to interacting and non-interacting protein pairs respectively.

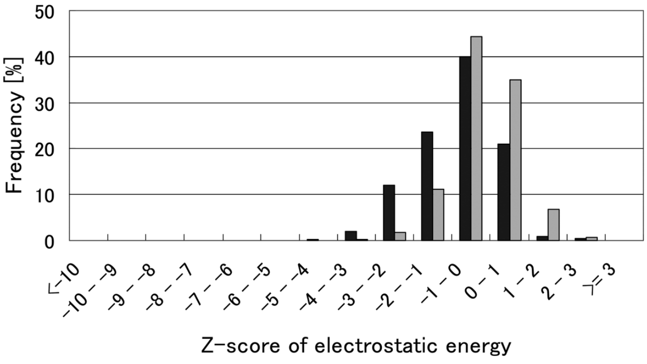

Figure 4 Distributions of Z-score of electrostatic energy calculated for the protein pairs included in the complete dataset.

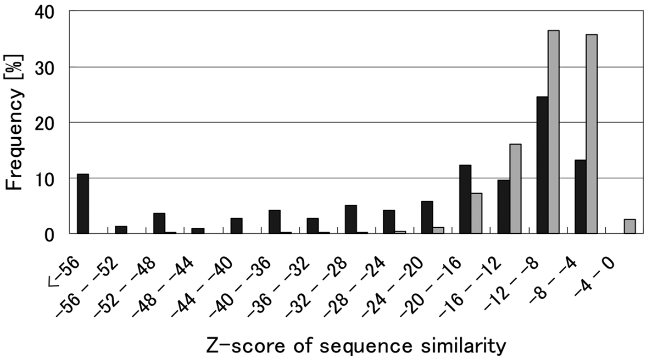

Figure 5 Distributions of Z-score of sequence similarity calculated for the protein pairs included in the complete dataset.

interacting protein pairs were 2.4 (contact energy) and 0.62 (electrostatic energy), the discrimination power of the contact energy seemed to be better than that of the electrostatic energy. The distribution of sequence similarities for the interacting protein pairs was not bell-shaped (as was the case for the contact and electrostatic energies), and was skewed toward the left. The distribution of the interacting

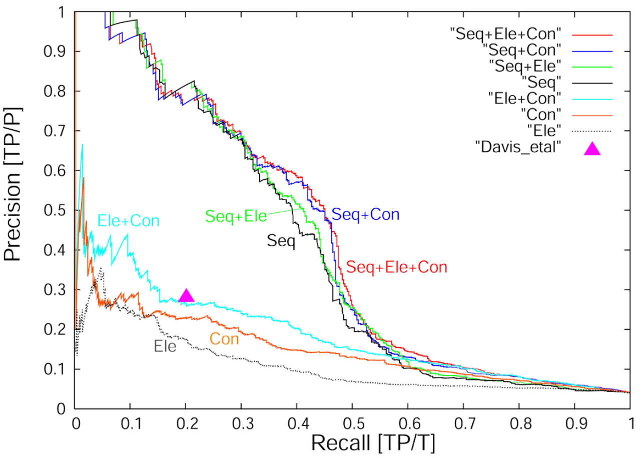

Figure 6 Recall-precision plots for discrimination between interacting and non-interacting protein pairs using single and combined scores in the complete dataset. "Con": contact energy, "Ele": electrostatic energy, "Seq": sequence similarity. "Ele+Con", "Seq+Con", "Seq+Ele" and "Seq+Ele+Con" correspond to the plots using combined Z-scores. The purple triangle shows the performance of the method of Davis *et al.*[25]

pairs was broader than that of the non-interacting pairs; the variances of the Z-score distribution of sequence similarity are 394.2 (interacting) and 20.7 (non-interacting). The high confidence dataset also yields similar distributions (data not shown).

Recall-precision plots

To evaluate the discrimination more strictly, we generated recall-precision plots for all three Z-scores, both individually and in combination. To generate combined scores, two or three Z-scores were added without any weights. Recall-precision plots are shown in Figure 6 (complete dataset) and Figure 7 (high confidence dataset); maximum F-measures of the recall-precision plot are summarized in Figure 8 (complete dataset) and Figure 9 (high confidence dataset). We also tested various weights such as Fischer's discriminant method, but performance was not significantly improved. The basic characteristics of plots using the complete and high confidence dataset are similar, except that precision values and maximum F-measure of the high confidence dataset were generally higher than those of the complete dataset, probably because the number of non-interacting protein pairs (2,839 pairs) in the high confidence dataset was about one forth of that in the complete set (9,908 pairs). Similar biased results using co-localization datasets are reported in previous studies[36,37].

In both datasets, the discriminating power of sequence similarity alone was much higher than that of the contact and electrostatic energies. This high performance was consistent with other studies based on sequence similarities[17–21]. However, when the contact energy and the electrostatic energy were combined with the sequence similarity, the maximum F-measure was improved by 0.038 for the com-

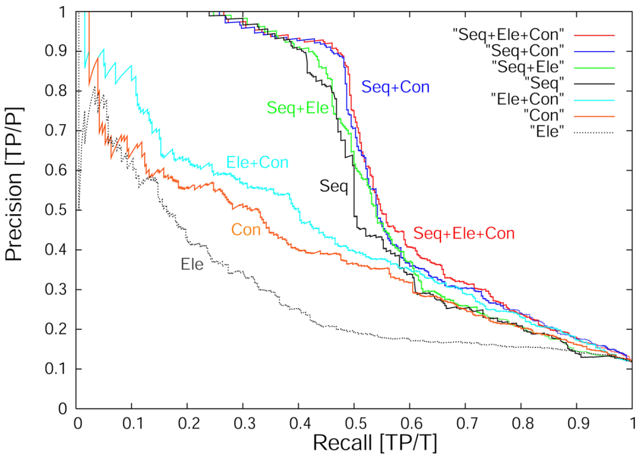

Figure 7 Recall-precision plots for discrimination between interacting and non-interacting protein pairs using single and combined scores in the high confidence dataset. Abbreviations as in Figure 6.

plete dataset. Similar improvements were observed for the high confidence dataset. This indicates that while sequence information is the most effective feature for detecting interacting protein pairs, structural information is able to improve prediction performance.

To validate the statistical significance of these improvements, we performed bootstrap sampling tests. The maximum F-measure was recalculated using protein pairs bootstrap-sampled from the all protein complex models. The sampling was repeated 1,000 times to generate 1,000 different maximum F-measures. In both datasets, among the 1,000 F-measures, all of the 1,000 F-measures of sequence similarity and contact energy (Seq+Con), and of all the three scores combined (Seq+Ele+Con) were larger than those of only sequence similarity (Seq). However, only 984 F-measures of sequence similarity and electrostatic energy (Seq+Ele) were larger than those of sequence similarity for the complete dataset. For the high confidence dataset, only 809 F-measures of Seq+Ele were larger than those of Seq. Thus, in both datasets, the improvement in discrimination after incorporation of contact energy was statistically significant ($p<0.01$), whereas, the improvement after incorporation of electrostatic energy was not. That is to say, sequence similarity has a much higher discriminating power than the other structure-based scores, but using contact energy results in significant improvement over predictions using sequence similarity alone.

The level of prediction accuracy practically required by users depends on their purposes. If a researcher needs to know interacting protein pairs without any confirming experiments, we would recommend the prediction with high Precision and low Recall. In contrast, if a researcher plans to perform a number of experiments to confirm protein-protein interactions, and needs candidates of interacting protein pairs, we would recommend the prediction with high Recall and low Precision. The improvement by our contact energy can contribute to the latter case, because Figure 6 indicates that the difference between the sequence similarity and the combined score is the largest in the region where Recall is high (0.4–0.5) and Precision is low (0.3–0.6).

Performance comparison with the previously published method

Generally speaking, it is difficult to quantitatively compare the protein-protein interaction prediction methods, because the criteria for interacting protein pairs and the libraries of complex structures can both differ. We compare the performance of our method with the latest related method proposed by Davis *et al.*[25], by checking overlaps of their predictions with our complete dataset. Their method was based on the statistical contact energy in conjunction with functional annotation and subcellular localization data. The contact energy metric employed in their study was

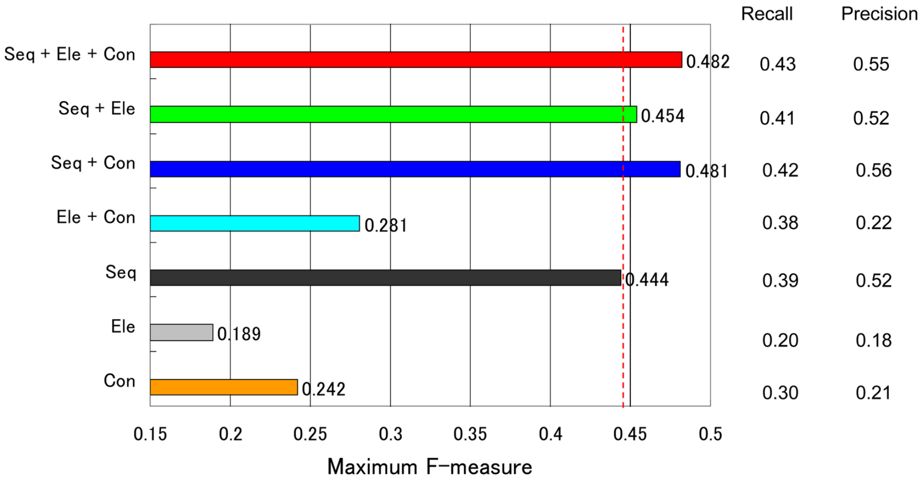

Figure 8 The maximum F-measures with their recall and precision values for each recall-precision plot using single and combined Z-scores in the complete dataset. Abbreviations as in Figure 6. Dotted line: maximum F-measure of sequence similarity alone.

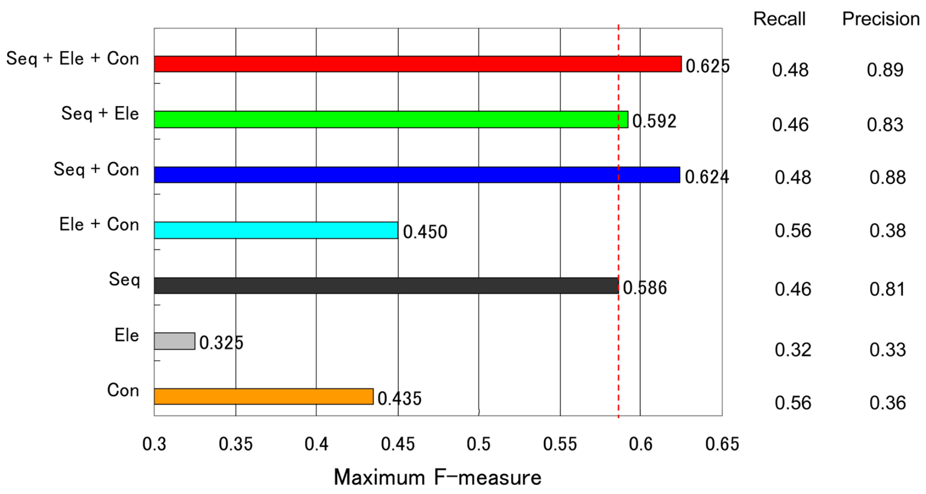

Figure 9 The maximum F-measures with their recall and precision values for each recall-precision plot using single and combined Z-scores in case of the high confidence dataset. Abbreviations are the same as those used in Figure 6.

similar to ours, except that it was weighted by the ratio of contacting atoms to total atoms, and its contacting atomic types and threshold distance of contacts were deliberately chosen. Because their complex models were generated by structural alignments of monomer models to template complex structures, the number of model complex structure could be larger than ours if we employed the same structural library. Davis *et al.* applied their method to all the protein pairs of yeast, finally predicted 3,387 interacting protein pairs. Among the 3,387 predictions, 2,520 predictions are hetero (sequence identity is smaller than 50%) protein pairs, and only 300 pairs are included in our complete dataset; 84 pairs are interacting, and 216 are non-interacting pairs. The remaining 2,220 pairs are modeled by Davis *et al.*, but not modeled by our method. This last difference was caused by the difference of the template structure library; we did not use homodimer templates to avoid artificial crystal packing, whereas they used all kinds of complex structures. We found that most of the remaining 2,220 hetero protein pairs were modeled using homodimer templates. Thus, by the equations (15) and (16), the values of recall and precision of the method of Davis *et al.* are,

$$Recall(S) = \frac{N_{tp}(S)}{N_t} = \frac{84}{417} = 0.201,$$

$$Precision(S) = \frac{N_{tp}(S)}{N_p(S)} = \frac{84}{84+216} = 0.280.$$

Their values are plotted in Figure 6 (purple triangles). The performance of their method is better than that of our contact energy, and slightly better than that of our contact energy combined with electrostatic energy. This is probably due to their different estimation of contact energy and their filter by co-localization and co-functional annotation. However, the predictive performance of Davis *et al.*'s method is plotted under the line of sequence similarity (Seq in Fig. 6). Although the comparisons in the two studies were not performed on identical structural libraries and the assumption of our complete dataset is not absolutely correct, our results suggest that methods incorporating sequence similarity will yield more accurate predictions than methods incorporating only structure-based scores along with functional and localization data.

Conclusions

In this study, we developed a method for predicting protein-protein interaction based on dimer structure models, using two structural scores and sequence similarity. Because we restricted the protein pairs whose complex can be modeled by homology, the essence of our approach is the discrimination of specific interaction among similar homologous sequences. Previous structure-based prediction studies of protein-protein interaction have evaluated overlaps of predicted and experimentally observed interacting pairs, but have not checked as carefully the overlaps of non-interacting pairs. Because we believe that non-interacting protein pairs should be also evaluated, we prepared two kinds of datasets containing interacting and non-interacting protein pairs. The complete dataset contains all the hetero protein pairs whose complex structure can be modeled, and the high confidence dataset is the reliable subset using subcellular localization data. The two datasets have both assets and liabilities. On the one hand, reliability of interactions of the high confidence dataset should be higher than that of the complete dataset. On the other hand, precision values estimated from the high confidence dataset are biased to large values, because that set ignores co-localized protein pairs not registered in the DIP database.

Both datasets showed that the performance of a sequence similarity-based score was much greater than scores based on contact and electrostatic energies. Nonetheless, scores related to contact energy, as calculated from structural models, can contribute to improvements over the performance of sequence similarity alone. These results suggest that sequence similarity is indispensable for the prediction, whereas structure scores can play supporting roles. Our preliminary calculation showed that a score only using number of aligned interface residues had a high discrimination power, although it was smaller than that of contact energy. We suggest that the contact energy may indirectly check whether a modeled structure has a sufficient size of interface.

Electrostatic energy showed the worst performance, and did not significantly improve the performance of sequence similarity alone. There are several possible reasons for this poor performance. We employed the simplified electrostatic energy proposed by Shaul and Schreiber[49]. They reported that this energy successfully predicted the change of association rate k_{on}, however, it may be insufficient to predict binding free energy. This energy ignores partial charges on polar atoms, it can not consider any polar interactions such as hydrogen bonds. Another reason is the inaccuracy of complex models of interacting protein pairs, which may more affect the performance of the electrostatic energy than that of the contact energy. It is because the electrostatic energy depends on sidechain conformations, whereas the contact energy does not. The omission of charges on binding ligands such as nucleotides and metal ions may be a serious problem. Many protein interactions of signal transduction systems are regulated by bindings of charged ligands, such as GTP and GDP.

Our results showed that combined score using sequence similarity and contact energy is the currently most accurately predictive score. Using the combined score, we now plan to apply our method to different organisms, and we hope to obtain new biological findings through our predicted interactions. We also plan to build a WWW server in order to make our prediction service freely available to other researchers.

Acknowledgments

We are grateful to Drs. Kei Yura, Hidetoshi Kono, Kensuke Nakamura, and Gautam Basu for stimulating discussions and encouragements. We thank Drs. Toshio Hakoshima and Naotake Ogasawara for supports and suggestions throughout the work. We also thank Drs. Andrej Sali and Fred P. Davis for providing us the list of their predicted interacting protein pairs to validate our prediction performance. This work was supported by the Special Coordination Funds Promoting Science and Technology and a Grant-in-Aid for Scientific Research on Priority Area (C), Genome Information Science, from MEXT (the Ministry of Education, Culture, Sports, Science and Technology of Japan). N. Fukuhara was supported by a Grand-in-Aid for the 21st Century COE Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J. M. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
2. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
3. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
4. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D. & Tyers, M. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
5. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B. & Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
6. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., Onge, P. S., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A. & Greenblatt, J. F. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
7. Bader, G. D., Betel, D. & Hogue, C. W. V. BIND: the Biomolecular interaction network database. *Nucleic Acids Res.* **31**, 248–250 (2003).

8. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
9. Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H. W. & Stumpflen, V. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.* **34**, D436–D441 (2006).
10. Salwinski, L. & Eisenberg, D. Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.* **13**, 377–382 (2003).
11. Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I. & Marcotte, E. M. Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292–299 (2004).
12. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
13. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
14. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
15. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614 (2001).
16. Sato, T., Yamanishi, Y., Kanehisa, M. & Toh, H. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* **21**, 3482–3489 (2005).
17. Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S. & Vidal, M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions of "Interologs". *Genome Res.* **11**, 2120–2126 (2001).
18. Wojcik, J. & Schachter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* **17**, S296–S305 (2001).
19. Wojcik, J., Boneca, I. G. & Legrain, P. Prediction, assessment and validation of protein interaction maps in bacteria. *J. Mol. Biol.* **323**, 763–770 (2002).
20. McDermott, J. & Samudrala, R. Enhanced functional information from predicted protein networks. *Trends Biotechnol.* **22**, 60–62 (2004).
21. Patil, A. & Nakamura, H. HINT: a database of annotated protein-protein interactions and their homologs. *Biophysics* **1**, 21–24 (2005).
22. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA* **99**, 5896–5901 (2002).
23. Lu, L., Lu, H. & Skolnick, J. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**, 350–364 (2002).
24. Lu, L., Arakaki, A. K., Lu, H. & Skolnick, J. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.* **13**, 1146–1154 (2003).
25. Davis, F. P., Braberg, H., Shen, M. Y., Pieper, U., Sali, A. & Madhusudhan, M. S. Protein complex compositions predicted by structural similarity. *Nucleic Acids Res.* **34**, 2943–2952 (2006).
26. Grigoryan, G. & Keating, A. E. Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.* **355**, 1125–1142 (2006).
27. Pieper, U., Eswar, N., Davis, F. P., Braberg, H., Madhusudhan, M. S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B. M., Eramian, D., Shen, M. Y., Kelly, L., Melo, F. & Sali, A. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **34**, D291–D295 (2006).
28. Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J. M., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H., Zhong, F., Zhong, W., Gibbs, R., Venter, J. C., Adams, M. D. & Lewis, S. Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
29. Gomperts, B. D., Kramer, I. M. & Tatham, P. E. R. *Protein domains and signal transduction. In:Signal Transduction.* pp. 393–410 (Academic Press, San Diego, 2002).
30. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein Interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Prot.* **1**, 349–356 (2002).
31. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
32. Sprinzak, E., Sattath, S. & Margalit, H. How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327**, 919–923 (2003).
33. Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K. H., Miller, P., Gerstein, M., Roeder, G. S. & Snyder, M. Subcellular localization of the yeast proteome. *Genes & Dev.* **16**, 707–719 (2002).
34. Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. & Weissman, J. S. Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
35. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. & O'Shea, E. K. Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
36. Ben-Hur, A. & Noble, W. S. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* **7**, S2 (2006).
37. Li, M. H., Wang, X. L., Lin, L. & Liu, T. Effect of example weights on prediction of protein-protein interactions. *Computat. Biol. Chem.* **30**, 386–392 (2006).
38. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* **332**, 989–998 (2003).
39. Henrick, K. & Thornton, J. M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361 (1998).
40. Johnson, R. A. & Wichern, D. W. *Applied multivariate statistical analysis.* pp. 740 (Prentice-Hall, London, 1998).
41. Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N. & Suzek, B. The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191 (2006).
42. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J.,

Zhang, Z., Miller, W. & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

43. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).
44. Sippl, M. J. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883 (1990).
45. Jones, D. T., Taylor, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* **358**, 86–89 (1992).
46. Moont, G., Gabb, H. A. & Sternberg, M. J. E. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35**, 364–373 (1999).
47. Lu, H., Lu, L. & Skolnick, J. Development of unified statistical potentials describing protein-protein interactions. *Biophys J.* **84**, 1895–1901 (2003).
48. Sheinerman, F. B., Norel, R. & Honig, B. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159 (2000).
49. Shaul, Y. & Schreiber, G. Exploring the charge space of protein-protein association: A proteomic study. *Proteins* **60**, 341–352 (2005).
50. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
51. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. & Murzin, A. G. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**, D226–D229 (2004).
52. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
53. Wagner, A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292 (2001).