BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 17 2005, pages 3482–3489
doi:10.1093/bioinformatics/bti564

Sequence analysis

The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships

Tetsuya Sato[1,*], Yoshihiro Yamanishi[2], Minoru Kanehisa[1] and Hiroyuki Toh[3]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan, [2]Centre de Géostatistique, Ecole des Mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau cedex, France and [3]Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Fukuoka 812-8582, Japan

Received on April 23, 2005; revised on June 23, 2005; accepted on June 28, 2005
Advance Access publication June 30, 2005

ABSTRACT

Motivation: The prediction of protein–protein interactions is currently an important issue in bioinformatics. The mirror tree method uses evolutionary information to predict protein–protein interactions. However, it has been recognized that predictions by the mirror tree method lead to many false positives. The incentive of our study was to solve this problem by improving the method of extracting the co-evolutionary information regarding the protein pairs.

Results: We developed a novel method to predict protein–protein interactions from co-evolutionary information in the framework of the mirror tree method. The originality is the use of the projection operator to exclude the information about the phylogenetic relationships among the source organisms from the distance matrix. Each distance matrix was transformed into a vector for the operation. The vector is referred to as a 'phylogenetic vector'. We have proposed three ways to extract the phylogenetic information: (1) using the 16S rRNA from the same source organisms as the proteins under consideration, (2) averaging the phylogenetic vectors and (3) analyzing the principal components of the phylogenetic vectors. We examined the performance of the proposed methods to predict interacting protein pairs from *Escherichia coli*, using experimentally verified data. Our method was successful, and it drastically reduced the number of false positives in the prediction.

Availability: The R script for the prediction of protein–protein interactions reported in this manuscript is available at http://timpani.genome.ad.jp/~proj/

Contact: sato@kuicr.kyoto-u.ac.jp

Supplementary information: The information is also available at the same site as the R script.

1 INTRODUCTION

Information about protein–protein interactions in living cells provides deep insight into the biological functions of proteins and the behavior of cells. Genome-wide experimental analyses, such as the yeast 2-hybrid system (Ito *et al*., 2001; Uetz *et al*., 2000) and mass spectrometry (Gavin *et al*., 2002; Ho *et al*., 2002), have facilitated exhaustive investigations of protein–protein interactions in cells. However, such experimental methods have coverage and accuracy problems (Sprinzak *et al*., 2003; von Mering *et al*., 2002). Currently, the prediction of protein–protein interactions has become one of the major issues in bioinformatics. The predicted protein–protein interactions can provide complementary or supporting evidence to the genome-wide experimental studies on protein–protein interactions eventhough computational analyses also suffer from the same problems as experimental studies, such as low coverage and low accuracy.

Various methods to predict protein–protein interactions have been developed. One of these methods is the prediction through genome comparisons, which includes phylogenetic profile (Pellegrini *et al*., 1999), Rosetta stone (Enright *et al*., 1999) and conserved gene neighborhood analyses (Dandekar *et al*., 1998). Prediction by using information about the co-occurrence of domains in protein–protein interactions is another approach. Co-evolutionary behavior between interacting proteins is also useful information for predictions. There are two representative prediction methods that utilize co-evolutionary information, the mirror tree method (Pazos and Valencia, 2001) and the *in silico* 2-hybrid system method (Pazos and Valencia, 2002). In this paper, we focus on the mirror tree method.

Although there are several preceding works, such as Goh *et al*. (2000), the mirror tree method was developed by Pazos and Valencia (2001). The mirror tree method predicts protein–protein interactions under the assumption that the interacting proteins show similarity in the molecular phylogenetic tree because of the co-evolution through the interaction. However, it is difficult to evaluate the similarity directly between a pair of molecular phylogenetic trees. Instead, the mirror tree method compares a pair of distance matrices in order to evaluate the extent of co-evolutionary behavior between two proteins. We will explain the method briefly. Consider two proteins, say, proteins A and B. The orthologues of protein A are collected from n species. The n sequences of protein A are aligned and the distance matrix, D_{A}, is calculated. The size of D_{A} is $n \times n$, and each row or column of the matrix corresponds to a species under consideration. An element of the matrix, $D_{\mathrm{A}}(i, j)$, represents the genetic distance between species i and j, which is calculated by comparing the amino acid sequences of protein A between the two species. A distance matrix is symmetric, and only the upper or lower half of the matrix includes sufficient information for the prediction. Likewise, the orthologues of protein B are collected from the same n species,

*To whom correspondence should be addressed.

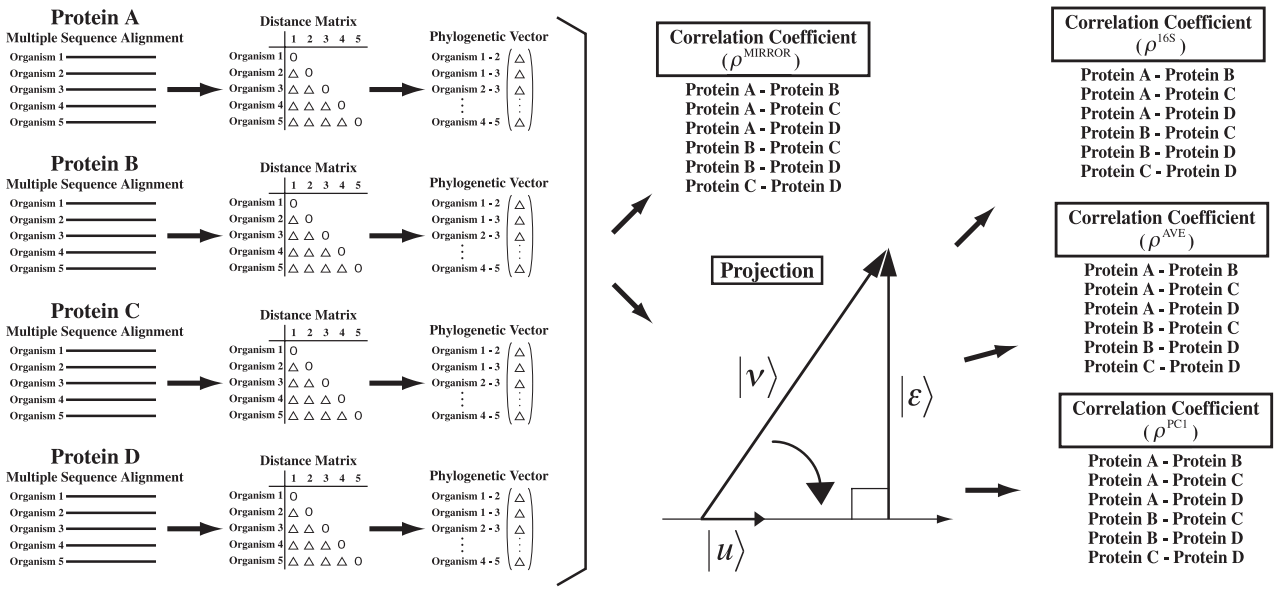

Fig. 1. A schematic representation of the procedures for predicting interaction partners from four types of correlation coefficients, ρ^{MIRROR}, ρ^{16S}, ρ^{AVE} and ρ^{PC1}. $|\nu\rangle$ represents a phylogenetic vector. $|u\rangle$ indicates a unit vector representing the phylogenetic relationship among source organisms. $|\varepsilon\rangle$ is a residual vector after excluding the phylogenetic relationship from $|\nu\rangle$, which is supposed to represent the co-evolutionary vector.

and the distance matrix, D_{B}, is calculated. The intensity of the co-evolutionary constraint between proteins A and B is evaluated as Pearson's correlation coefficient, ρ, between the distance matrices D_{A} and D_{B}, which is calculated as follows:

$$\rho_{\mathrm{AB}} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(D_{\mathrm{A}}(i,j) - \mathrm{Ave}(D_{\mathrm{A}}))(D_{\mathrm{B}}(i,j) - \mathrm{Ave}(D_{\mathrm{B}}))}{\sqrt{\mathrm{Var}(D_{\mathrm{A}})\mathrm{Var}(D_{\mathrm{B}})}}, \tag{1}$$

where Ave and Var represent the average and the variance of the upper (or lower) half elements of a distance matrix, respectively. When a pair of proteins shows a high correlation coefficient the proteins are regarded as interacting with each other. The mirror tree method evaluates the extent of the interaction between a pair of proteins. However, as seen in several ligand–receptor systems, such interactions are not always one-to-one. There are cases in which many homologous ligands interact with many homologous receptors. Gertz *et al*. (2003) and Ramani and Marcotte (2003) independently improved the mirror tree method in similar ways to allow the consideration of such multiple interaction cases. Tan *et al*. (2004) recently launched a web server, ADVICE, which automatically predicts protein–protein interactions by the mirror tree method upon client request.

One of the problems of the mirror tree method is the large number of false positives in the prediction. Even protein pairs that are known not to interact often show high correlation coefficients. The abundance of false positives in the mirror tree prediction reduces the reliability of the method in actual applications. The distance matrices of orthologous proteins from the same set of n source organisms are compared in the mirror tree method. Therefore, all of the distance matrices of the proteins are considered to include the information about the phylogenetic relationships among the same n sources, to some extent. The phylogenetic relationships among the identical set of sources behind the distance matrices would be the cause for such a high correlation between non-interacting proteins. If we can exclude the information about the phylogenetic relationships from the distance matrices then the performance of the mirror tree method may be improved.

In our method, we used a projection operator to exclude the information about the phylogenetic relationships of the sources, and then the residual information after this operation was used for the calculation of the correlation coefficient between proteins. The projection operator is a linear transformation in a vector space. A point in the vector space is projected to a subspace so that the difference vector between the original point and the image in the subspace is orthogonal to the subspace. The projection operator is widely used in various fields, such as multivariate analysis and quantum mechanics. One of the well-known examples of the use of the projection operator is spectral resolution. We applied our method to physically contacting proteins, to evaluate its performance. That is, in this manuscript a protein–protein interaction means physical contact. As discussed below, our method succeeded in drastically reducing the number of false positives in the predicted protein–protein interactions. The quality of the data needed to realize a correct prediction was also examined. We also found that the inclusion of distantly related orthologues in the data improves the performance. The benefits and limitations of our approach are discussed based on our observations.

2 METHODS

The method developed by us is outlined in Figure 1.

2.1 Data preparation

We selected 13 pairs of *Escherichia coli* proteins that are physically in contact, from the Database of Interacting Proteins (DIP) Version 01/02/2005 (Salwinski *et al*., 2004). The selected pairs are described in the legend for Table 1. Each pair was selected so that neither of the interacting proteins participated in the remaining 12 pairs of interacting proteins. Then, putative orthologues corresponding to the 26 proteins derived from *E.coli* were

Table 1. Comparison of the top 30 predicted interactive pairs on among the four methods

Rank	ρ^{MIRROR}		ρ^{16S}		ρ^{AVE}		ρ^{PC1}	
1	dnaN–rpoB	0.97680	sucD–sucC	0.92440*	sucD–sucC	0.90956*	sucD–sucC	0.91527*
2	dnaK–secY	0.96324	atpA–atpD	0.80301*	trpA–trpB	0.79218*	trpA–trpB	0.75387*
3	dnaK–rpoB	0.96269	carA–carB	0.80290*	rpoA–rpoB	0.65370*	carA–carB	0.64862*
4	sucD–sucC	0.96222*	dnaK–secY	0.80050	carA–carB	0.64216*	atpA–atpD	0.63991*
5	dnaN–dnaK	0.96019	trpA–trpB	0.79732*	dnaN–rpoB	0.63433	dnaN–rpoB	0.58736
6	atpA–atpD	0.95876*	dnaE–secA	0.78250	atpA–atpD	0.61494*	dnaK–atpD	0.55995
7	rpoA–rpoB	0.95755*	dnaK–atpD	0.77360	iscS–iscU	0.60684*	dnaK–secY	0.55993
8	rpoB–secY	0.95463	dnaN–rpoB	0.77236	grpE–clpP	0.55301	iscS–iscU	0.55503*
9	secY–secA	0.95449*	rpoA–rpoB	0.76777*	dnaK–carB	0.54073	grpE–clpP	0.54494
10	dnaK–atpD	0.95335	dnaN–carA	0.76090	grpE–tsf	0.54054	dnaK–carB	0.54207
11	dnaN–secY	0.95330	dnaN–dnaK	0.76043	dnaK–secY	0.51635	secY–carB	0.53090
12	dnaK–atpA	0.95193	dnaK–carB	0.75795	dnaK–atpD	0.51435	dnaK–rpoB	0.50026
13	dnaE–secA	0.94503	dnaN–secY	0.75778	ruvA–ruvB	0.51116*	rpoA–rpoB	0.49776*
14	dnaN–rpoA	0.94477	dnaE–secY	0.75508	secY–carB	0.50080	grpE–tsf	0.49741
15	dnaK–secA	0.94419	rpoB–secY	0.75349	rpoB–secY	0.49806	dnaK–atpA	0.49643
16	dnaE–secY	0.94359	dnaK–atpA	0.74688	tsf–trpB	0.47641	ruvA–ruvB	0.48901*
17	dnaN–clpX	0.94332	dnaE–dnaK	0.74302	dnaA–ruvB	0.46877	dnaE–secA	0.48040
18	dnaN–secA	0.93959	secY–carB	0.72709	secA–trpB	0.45282	dnaE–secY	0.47679
19	clpX–rpoB	0.93729	iscS–iscU	0.71725*	tufB–tsf	0.44526*	dnaA–ruvB	0.45486
20	dnaN–carA	0.93667	dnaK–carA	0.71550	dnaK–atpA	0.43830	secY–secA	0.43071*
21	carA–carB	0.93589*	secY–carA	0.71236	dnaA–ruvA	0.43710	grpE–sucC	0.42376
22	rpoB–atpA	0.93213	dnaK–rpoB	0.69761	dnaK–rpoB	0.42311	dnaE–carB	0.41792
23	dnaE–dnaK	0.93159	iscS–carA	0.69327	dnaE–secY	0.42211	dnaN–dnaK	0.41372
24	rpoB–secA	0.92895	dnaE–carB	0.68726	dnaE–carB	0.41183	tufB–tsf	0.40691*
25	rpoB–atpD	0.92884	dnaE–carA	0.66878	dnaN–rpoA	0.40868	dnaA–ruvA	0.40445
26	secY–carA	0.92852	dnaN–carB	0.65973	iscS–carA	0.40738	grpE–sucD	0.39991
27	dnaN–atpA	0.92801	dnaK–secA	0.65575	dnaE–secA	0.38987	rpoB–secY	0.38340
28	dnaK–carA	0.92795	grpE–clpP	0.65379	dnaN–secY	0.38495	dnaN–clpX	0.37297
29	clpX–secY	0.92630	rpoB–carA	0.63704	dnaN–dnaK	0.37687	tsf–trpB	0.36841
30	ruvA–ruvB	0.92608*	secA–carB	0.63347	iscU–atpD	0.35675	dnaE–dnaK	0.36360

The abbreviated names of the interacting proteins are as follows: sucC–sucD, succinyl–CoA synthetases alpha–beta; atpA–atpD, ATP synthases alpha–beta; rpoA–rpoB, DNA-directed RNA polymerases alpha–beta; secA–secY, preprotein translocase secA–secY; carA–carB, carbamoyl-phosphate synthases small–large; ruvA–ruvB, Holliday junction DNA helicases ruvA–ruvB; iscS–iscU, putative aminotransferase–NifU-like protein; dnaE–dnaN, DNA polymerases III alpha–beta; trpA–trpB, tryptophan synthases alpha–beta; tufB–tsf, elongation factors EF-Tu–EF-Ts; dnaA–dnaB, DNA helicase–dnaA; grpE–dnaK, heat shock protein grpE–dnaK protein; and clpX–clpP, ATP-dependent clp proteases ATP-binding subunit–protease proteolytic subunit.

collected from 40 different bacterial species, according to the description in the KEGG/KO database (Kanehisa *et al*., 2004). The sources are shown in the Supplemental Figure S1. Hereafter, the set of putative orthologues from the 41 bacterial sources is simply referred to as the orthologues. One of the important assumptions in this study is that a pair of proteins, which are orthologous to the interacting proteins of *E.coli*, are also physically in contact. The other assumption is that the interaction affects the co-evolution of the orthologues.

A multiple alignment of each set of orthologous proteins was made with the alignment software MAFFT (Katoh *et al*., 2005). A distance matrix for the orthologues was calculated from the multiple alignment. Then, a genetic distance between every pair of aligned sequences was calculated as a maximum likelihood estimate using the PROTDIST module in the PHYLIP package (Felsenstein, 2004). The score table by Jones *et al*. (1992) was used for the maximum likelihood estimation. A distance matrix for a set of orthologues was constructed with the genetic distances.

2.2 Transformation from distance matrix to phylogenetic vector

The distance matrix was transformed into a vector for easier formulation. The upper or lower half of the non-diagonal elements of the distance matrix was arranged as an array of the numerical values in a certain order. All of the matrices were transformed into vectors with the same order of the elements. When the matrix has a size of $n \times n$ the dimension of the vector is $n(n-1)/2$. The vector is hereafter referred to as a 'phylogenetic vector'. In this study, n is equal to 41. Therefore, the dimension of the phylogenetic vector is 820. Let us consider a pair of phylogenetic vectors $|\nu_i\rangle$ and $|\nu_j\rangle$, which are transformed from distance matrices D_i and D_j, where the subscripts i and j indicate different sets of orthologues. Then, we apply the normalization of the elements of each vector with the average and the standard deviation of the elements as follows:

$$|\nu_i^\star\rangle = \frac{|\nu_i\rangle - |\mu\rangle}{\sqrt{\mathrm{Var}(\nu_i)}},$$

where $|\mu\rangle$ is a vector with the same dimension as $|\nu_i\rangle$. All the elements of $|\mu\rangle$ are constant, and are equivalent to the arithmetic average over the elements of $|\nu_i\rangle$. $\mathrm{Var}(\nu_i)$ indicates the variance over all the elements of $|\nu_i\rangle$. The superscript $\star$ in $|\nu_i^\star\rangle$ indicates that the vector is normalized. Then, the inner product of a pair of normalized vectors is reduced to the Pearson's correlation coefficient used for the mirror tree method, which is defined by formula (1). Hereafter, the correlation coefficient will be denoted as $\rho_{ij}^{\mathrm{MIRROR}}$.

$$\rho_{ij}^{\mathrm{MIRROR}} = \langle \nu_i^\star | \nu_j^\star \rangle.$$

2.3 Projection operator

Consider a unit vector $|u\rangle$, which represents the phylogenetic relationship of the species under consideration. If such a vector is obtained, then the following projection operator P can be defined as

$$P = I - |u\rangle\langle u|, \tag{2}$$

where $|u\rangle\langle u|$ is also a projection operator onto the direction of the unit vector $|u\rangle$. The projection operator is a matrix with the size of $n(n-1)/2 \times n(n-1)/2$. The method to obtain $|u\rangle$ is explained below. I represents an identity matrix with the same size as $|u\rangle\langle u|$. By applying the projection operator (2) to a phylogenetic vector, say, $|v_i\rangle$, the component within $|v_i\rangle$, which is orthogonal to $|u\rangle$, is obtained as follows:

$$|\varepsilon_i\rangle = P|v_i\rangle = |v_i\rangle - |u\rangle\langle u|v_i\rangle. \tag{3}$$

That is, the projection operator can exclude the phylogenetic relationship from a phylogenetic vector. The same projection operator was applied to all of the phylogenetic vectors under consideration. Each of the residual vectors defined by formula (3) was normalized with the average and the standard deviation of the elements. Consider a pair of normalized vectors $|\varepsilon_i^\star\rangle$ and $|\varepsilon_j^\star\rangle$. Then, the inner product of the two vectors

$$\rho_{ij}^{\mathrm{PRJ}} = \langle\varepsilon_i^\star|\varepsilon_j^\star\rangle$$

represents Pearson's correlation coefficient between the residues, after excluding the phylogenetic relationship from the original phylogenetic vectors. ρ_{ij}^{PRJ} is a new measure to evaluate the co-evolutionary behavior between proteins i and j.

2.4 Unit vector in the projection operator

The remaining problem is how to obtain the unit vector $|u\rangle$ representing the phylogenetic relationship of the source organisms. We developed three different methods to design such a unit vector: (1) transformation of the distance matrix of 16S ribosomal RNA (rRNA) from the same source organisms as the proteins under consideration, (2) averaging the phylogenetic vectors and (3) analyzing the principal components of the phylogenetic vectors.

In the first method, 16S rRNA was used for the calculation. Basically, each organism has at least one copy of the 16S rRNA gene. Therefore, the distance matrix or the phylogenetic vector of the 16S rRNAs is considered to represent the phylogenetic relationship among the source organisms. The rRNA sequences from the same sources as the proteins under consideration were collected from the KEGG/GENES database (Kanehisa *et al*., 2004) and the Ribosomal Database Project-II Release 9 (Gustafson *et al*., 2005). The rRNA sequences thus collected were aligned, and the distance between every pair of the aligned RNA sequences was calculated by using the F84 scoring table (Kishino and Hasegawa, 1989) and the DNADIST module in the PHYLIP package (Felsenstein, 2004). The distance matrix was then transformed into a phylogenetic vector $|v_{16S}\rangle$. $\|v_{16S}\| = \sqrt{\langle v_{16S}|v_{16S}\rangle}$ indicates the size of the vector. Then, a unit vector $|u_{16S}\rangle$ was obtained as $|v_{16S}\rangle/\|v_{16S}\|$.

In the second method, all of the phylogenetic vectors under consideration were normalized so that the standard deviation of the elements in each protein was '1' at first. Then, they were averaged as

$$|v_{\mathrm{AVE}}\rangle = \frac{1}{m}\sum_{i=1}^{m}\frac{|v_i\rangle}{\|v_i\|},$$

where m is the number of proteins. In this study, m was equal to 26, as described above. The second unit vector $|u_{\mathrm{AVE}}\rangle$, was obtained as $|v_{\mathrm{AVE}}\rangle/\|v_{\mathrm{AVE}}\|$.

In the third method, the phylogenetic vectors were used again. Let X be a matrix in which the i-th column corresponds to a phylogenetic vector of protein i, normalized with the average and the standard deviation. The size of X is $n(n-1)/2 \times m$. Then, a correlation coefficient matrix Y was calculated as $X^{\mathrm{T}}X$. The superscript T indicates the transpose of a matrix. Therefore, the size of Y is $m \times m$. The principal component analysis for the data corresponding to X was carried out by solving the eigenvalue problem of Y. Then, $|v_{\mathrm{PC1}}\rangle$ was obtained as $|v_{\mathrm{PC1}}\rangle = X|z_1\rangle$, where $|z_1\rangle$ is the first principal component axis associated with the largest eigenvalue for the correlation coefficient matrix. $|v_{\mathrm{PC1}}\rangle$ thus obtained is expected to represent the most common features of the m phylogenetic vectors. Then, $|v_{\mathrm{PC1}}\rangle/\|v_{\mathrm{PC1}}\|$ generated the third unit vector, $|u_{\mathrm{PC1}}\rangle$.

In the second and third methods it is assumed that the information, except for the phylogenetic relationship of the sources, can be approximately canceled out by the average operation or principal component analysis. The first method requires the presence of 16S rRNA from the same sources as the proteins under consideration, whereas the latter two methods are feasible with only the phylogenetic vectors. The Pearson's correlation coefficients between the residues for two sets of orthologues i and j, which were projected out by the operators constructed with $|u_{16S}\rangle$, $|u_{\mathrm{AVE}}\rangle$ and $|u_{\mathrm{PC1}}\rangle$, were represented by ρ_{ij}^{16S}, ρ_{ij}^{AVE} and ρ_{ij}^{PC1}. When the subscripts, i and j, are omitted, ρ^* collectively represents the type of correlation coefficient indicated by the superscript.

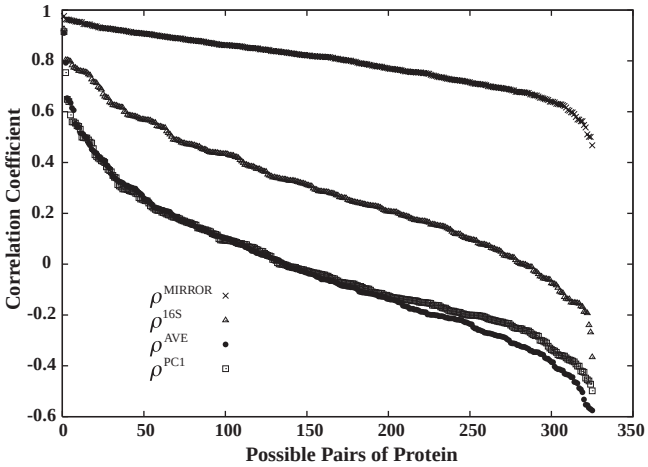

Fig. 2. The index plot of the sorted correlation coefficients of 325 possible pairs of proteins (x-axis). The y-axis indicates ρ^{MIRROR} (crosses), ρ^{16S} (open triangles), ρ^{AVE} (closed circles) and ρ^{PC1} (open squares).

3 RESULTS AND DISCUSSION

3.1 Prediction of protein–protein interactions by using ρ^{MIRROR}, ρ^{16S}, ρ^{AVE} and ρ^{PC1}

We calculated four types of correlation coefficients, ρ^{MIRROR}, ρ^{16S}, ρ^{AVE} and ρ^{PC1}, for all of the possible pairs of 26 proteins, that is, 325 pairs of proteins. The performance of each correlation coefficient was evaluated with the number of false positives. The correlation coefficients, sorted in decreasing order, are listed in the Supplemental Table S1, and only the top 30 members of the lists are shown in Table 1. Out of the 325 pairs, the interactions of 13 pairs have been experimentally identified and are highlighted with asterisks in the table. The top ranks of ρ^{16S}, ρ^{AVE} and ρ^{PC1} were occupied by pairs of actually interacting proteins. In contrast, non-interacting proteins were present within the top ranks of ρ^{MIRROR}. The decreasing patterns of the four correlation coefficients are shown in Figure 2, which shows that ρ^{MIRROR} decreased slowly, whereas ρ^{AVE} and ρ^{PC1} decreased rapidly. The rate of the ρ^{16S} decrease was rather moderate. Both Table 1 and Figure 2 clearly demonstrate the problem of the original mirror tree method. Even if a high value, say 0.9, is used as a threshold for the correlation coefficient to predict a protein–protein interaction,

Table 2. Prediction accuracy in terms of sensitivity and specificity

Method	0.9		0.8		0.7		0.6	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
ρ^{MIRROR}	61.54	13.79	84.62	6.21	100.00	4.96	100.00	4.17
ρ^{16S}	7.14	100.00	21.43	75.00	42.86	28.57	64.29	24.32
ρ^{AVE}	7.14	100.00	7.14	100.00	14.29	100.00	42.86	85.71
ρ^{PC1}	7.14	100.00	7.14	100.00	7.14	100.00	28.57	100.00

$$\text{Sensitivity} = \frac{\text{True positive}}{(\text{True positive} + \text{false negative})} \times 100\%; \quad \text{Specificity} = \frac{\text{True positive}}{(\text{True positive} + \text{false positive})} \times 100\%$$

ρ^{MIRROR} produces many pairs with high correlation, including non-interacting partners, and is likely to lead to many false positives in the prediction. However, the occupation of the top ranks by interacting proteins and the rapid decreases of ρ^{16S}, ρ^{AVE} and ρ^{PC1} guarantee the accuracy of prediction by the three correlation coefficients, if the threshold is set at a sufficiently high value.

The unit vector $|u\rangle$ seems to be a crucial factor for the prediction of a protein–protein interaction in the methods with a projection operator. Therefore, we examined the association among $|u_{16S}\rangle$, $|u_{\mathrm{AVE}}\rangle$ and $|u_{\mathrm{PC1}}\rangle$ by calculating Pearson's correlation coefficients, which is denoted as r as given below. We considered the absolute value of r because the sign of r does not make sense in this context. $|r|$ between $|u_{16S}\rangle$ and $|u_{\mathrm{AVE}}\rangle$ was 0.94697, whereas $|r|$ between $|u_{16S}\rangle$ and $|u_{\mathrm{PC1}}\rangle$ was 0.94597. The highest correlation, $|r| = 0.99805$, was observed between $|u_{\mathrm{AVE}}\rangle$ and $|u_{\mathrm{PC1}}\rangle$. The high correlation between $|u_{16S}\rangle$ and the other unit vectors suggests that one of our assumptions described above is correct. The information except for the phylogenetic relationship of sources can be approximately canceled out by the average operation or principal component analysis. The similarity in the patterns of the decreases in the correlation coefficients roughly corresponded to the similarity in the unit vectors. As shown in Figure 2, the two sets of plots of ρ^{AVE} and ρ^{PC1}, which were calculated with $|u_{\mathrm{AVE}}\rangle$ and $|u_{\mathrm{PC1}}\rangle$, overlapped each other. On the other hand, the plots of ρ^{16S}, which was related to $|u_{16S}\rangle$, slightly deviated from the plots of ρ^{AVE} and ρ^{PC1}.

The ρ^{16S}, ρ^{AVE} and ρ^{PC1} analyses seem to outperform the ρ^{MIRROR} analysis to a large extent. That is, the exclusion of the information about the phylogenetic relation among the source organisms from the distance matrices is effective to remove the false positives from the prediction by the mirror tree method. To investigate how different threshold values affect the accuracy of the prediction we introduced four thresholds for correlation coefficients, 0.9, 0.8, 0.7 and 0.6 (Table 2). The performances of the original mirror tree method and our proposed methods were evaluated with regard to sensitivity and specificity. When a pair of proteins had a correlation coefficient greater than the threshold the proteins were predicted to interact with each other. The advantage of ρ^{AVE} and ρ^{PC1} was the high specificity for any threshold. ρ^{16S} showed high specificity only for thresholds 0.9 and 0.8. In contrast, ρ^{MIRROR} showed high sensitivity in all of the cases, except for the threshold $= 0.9$. The high specificities of ρ^{16S}, ρ^{AVE} and ρ^{PC1} mean the drastic reduction of false positives, as compared with ρ^{MIRROR}. We will demonstrate how the number of false positives was reduced by our methods using a concrete example. For instance, we take proteins RpoB and SecY, which do not interact with each other. However, the ρ^{MIRROR} value of the pair was 0.95463, which occupies the 8th position of the list in Table 1. The same pair is presented at the 15th position in the sorted list of ρ^{16S}. As for ρ^{AVE} and ρ^{PC1}, the corresponding coefficients between the pair were 0.49806 and 0.38340, which are present at the 15th and 27th positions of the lists in Table 1.

Despite the improvement described above, the sensitivities of ρ^{16S}, ρ^{AVE} and ρ^{PC1} were lower than that of ρ^{MIRROR}. This means that a pair of proteins i and j, which interact with each other, will not always show high ρ_{ij}^{16S}, ρ_{ij}^{AVE} or ρ_{ij}^{PC1} coefficients. In other words, the number of false negatives increased when our methods were used, as compared with the original mirror tree method. In this study, we calculated the intensity of co-evolution between a pair of proteins as the correlation coefficient after the projection operation. However, the pairs may also interact with other proteins. If such proteins exist, the interaction with the pair would be difficult to detect, because the co-evolution with the other partners would interfere with the detection. To examine this hypothesis, we investigated the relationship between the multiplicity of the interaction and the correlation coefficient. The correlation coefficients, the multiplicities of interacting partners and the ranks in the sorted lists of the 13 pairs of interacting proteins are shown in Table 3. The multiplicity of interacting partners for proteins was evaluated with a modified Jaccard coefficient. The interacting partners were searched from the DIP database (Salwinski *et al*., 2004). Consider an interacting pair of proteins A and B. Let **M** and **N** be the sets of interacting partners of proteins A and B. Therefore, protein B belongs to **M**, whereas **N** includes protein A. The Jaccard coefficient is defined as $|\mathbf{M} \cap \mathbf{N}|/|\mathbf{M} \cup \mathbf{N}|$, where $|\mathbf{M}|$ is the size of the set **M** or the number of elements in the set. When the proteins A and B share many interacting partners the coefficient shows a value close to 1. However, it takes a low value close to 0 when protein A has many interacting partners which do not interact with protein B and *vice versa*. The deficiency of the original definition is that the coefficient is 0 when protein A interacts only with protein B. We modified the coefficient so that the coefficient between proteins A and B takes the value 1 when no other proteins interact with the pair. The modified Jaccard coefficient is defined as follows:

$$\text{Modified Jaccard coefficient} = \frac{|\mathbf{M} \cap \mathbf{N}| + 1}{|\mathbf{M} \cup \mathbf{N}| - 1}.$$

Table 3 clearly demonstrates the problem of the false negatives. At the same time, the table provides evidence to support our hypothesis. Roughly speaking, the correlation coefficient obtained after the projection operation, or the intensity of co-evolution, shows positive

Table 3. Ranking and correlation coefficient of protein pairs with experimentally identified interactions

Rank	Correlation coefficient	Interacting pair	Multiplicity of interacting partner
ρ^{MIRROR}			
4	0.96221	sucD–sucC	1.00 (1/1)
6	0.95876	atpA–atpD	0.80 (4/5)
7	0.95755	rpoA–rpoB	0.64 (9/14)
9	0.95449	secY–secA	0.67 (4/6)
21	0.93589	carA–carB	1.00 (1/1)
30	0.92608	ruvA–ruvB	1.00 (1/1)
46	0.91010	iscS–iscU	0.25 (1/4)
47	0.90963	dnaN–dnaE	0.57 (4/7)
69	0.88984	trpA–trpB	1.00 (1/1)
75	0.88481	tufB–tsf	0.50 (1/2)
156	0.81752	dnaA–dnaB	0.20 (1/5)
195	0.77738	grpE–dnaK	0.50 (3/6)
217	0.75471	clpX–clpP	0.33 (1/3)
$\rho^{16\mathrm{S}}$			
1	0.92439	sucD–sucC	1.00 (1/1)
2	0.80301	atpA–atpD	0.80 (4/5)
3	0.80290	carA–carB	1.00 (1/1)
5	0.79732	trpA–trpB	1.00 (1/1)
9	0.76777	rpoA–rpoB	0.64 (9/14)
19	0.71725	iscS–iscU	0.25 (1/4)
31	0.62588	secY–secA	0.67 (4/6)
32	0.62529	ruvA–ruvB	1.00 (1/1)
35	0.61821	dnaN–dnaE	0.57 (4/7)
62	0.53630	tufB–tsf	0.50 (1/2)
194	0.22533	grpE–dnaK	0.50 (3/6)
205	0.20446	dnaA–dnaB	0.20 (1/5)
218	0.17215	clpX–clpP	0.33 (1/3)
ρ^{AVE}			
1	0.90956	sucD–sucC	1.00 (1/1)
2	0.79218	trpA–trpB	1.00 (1/1)
3	0.65370	rpoA–rpoB	0.64 (9/14)
4	0.64216	carA–carB	1.00 (1/1)
6	0.61494	atpA–atpD	0.80 (4/5)
7	0.60684	iscS–iscU	0.25 (1/4)
13	0.51116	ruvA–ruvB	1.00 (1/1)
19	0.44526	tufB–tsf	0.50 (1/2)
71	0.17446	dnaA–dnaB	0.20 (1/5)
94	0.11203	secY–secA	0.67 (4/6)
149	−0.03236	dnaN–dnaE	0.57 (4/7)
218	−0.18071	clpX–clpP	0.33 (1/3)
223	−0.18909	grpE–dnaK	0.50 (3/6)
ρ^{PC1}			
1	0.91527	sucD–sucC	1.00 (1/1)
2	0.75387	trpA–trpB	1.00 (1/1)
3	0.64862	carA–carB	1.00 (1/1)
4	0.63991	atpA–atpD	0.80 (4/5)
8	0.55503	iscS–iscU	0.25 (1/4)
13	0.49776	rpoA–rpoB	0.64 (9/14)
16	0.48901	ruvA–ruvB	1.00 (1/1)
20	0.43071	secY–secA	0.67 (4/6)
24	0.40691	tufB–tsf	0.50 (1/2)
82	0.15353	dnaA–dnaB	0.20 (1/5)
151	−0.02804	dnaN–dnaE	0.57 (4/7)
162	−0.04645	grpE–dnaK	0.50 (3/6)
255	−0.20189	clpX–clpP	0.33 (1/3)

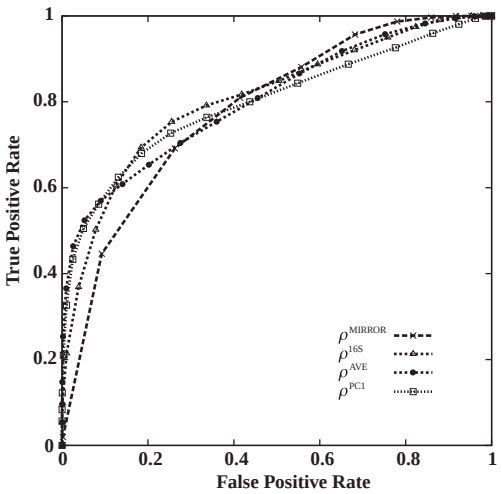

Fig. 3. The prediction accuracies of ρ^{MIRROR}, $\rho^{16\mathrm{S}}$, ρ^{AVE} and ρ^{PC1} were evaluated by the receiver operating characteristic (ROC) curves. The y-axis indicates the rate of the true positives and the x-axis indicates the rate of false positives. In the figure, crosses, open triangles, closed circles and open squares correspond to plots of ρ^{MIRROR}, $\rho^{16\mathrm{S}}$, ρ^{AVE} and ρ^{PC1}, respectively. The measure of the ROC curve shows that a curve in the upper area of the figure has higher accuracy, and the diagonal line corresponds to a random prediction accuracy.

correlation with the modified Jaccard coefficient. That is, the correlation coefficient obtained after the projection operation was high when proteins A and B formed a complex and no other proteins interacted with them. When the number of interacting partners increased, the intensity of co-evolution tended to be weak. However, when proteins A and B shared the interacting partners the intensity of co-evolution was high, in spite of the increase of the interacting partners. In such cases, all the interacting proteins may co-evolve each other. As shown in the table, there were several outliers from the tendency. One of the reasons for the deviation may be the lack of the experimental information. That is, all the protein–protein interactions have not been experimentally measured yet. In addition, it is suggested that some interaction have no functional meanings (Nooren and Thornton, 2003). Further accumulation of experimental knowledge is required to ascertain our hypothesis.

3.2 Assessment based on the ROC curve

The relationships between the true and false positives for the four correlation coefficients were also examined by drawing ROC curves (Fig. 3). As described above, proteins from 41 sources were used in this study. There is a possibility that the selection of source organisms may affect the accuracy of the prediction. In order to make the evaluation robust to the selection of source organisms, we took the following approach. Out of the 41 sources, 20 organisms were randomly selected. Then, ρ^{MIRROR}, $\rho^{16\mathrm{S}}$, ρ^{AVE} and ρ^{PC1} for every pair of 26 proteins were calculated using the randomly selected 20 organisms. The procedure was repeated 1000 times. The rates of true and false positives were calculated in each iteration step with 20 different threshold values. Based on the true and false positive rates averaged at each threshold value, ROC curves for ρ^{MIRROR}, $\rho^{16\mathrm{S}}$, ρ^{AVE} and ρ^{PC1} were drawn by connecting the points with 2D coordinates consisting of the two averaged rates. As shown in the figure, the ROC curves

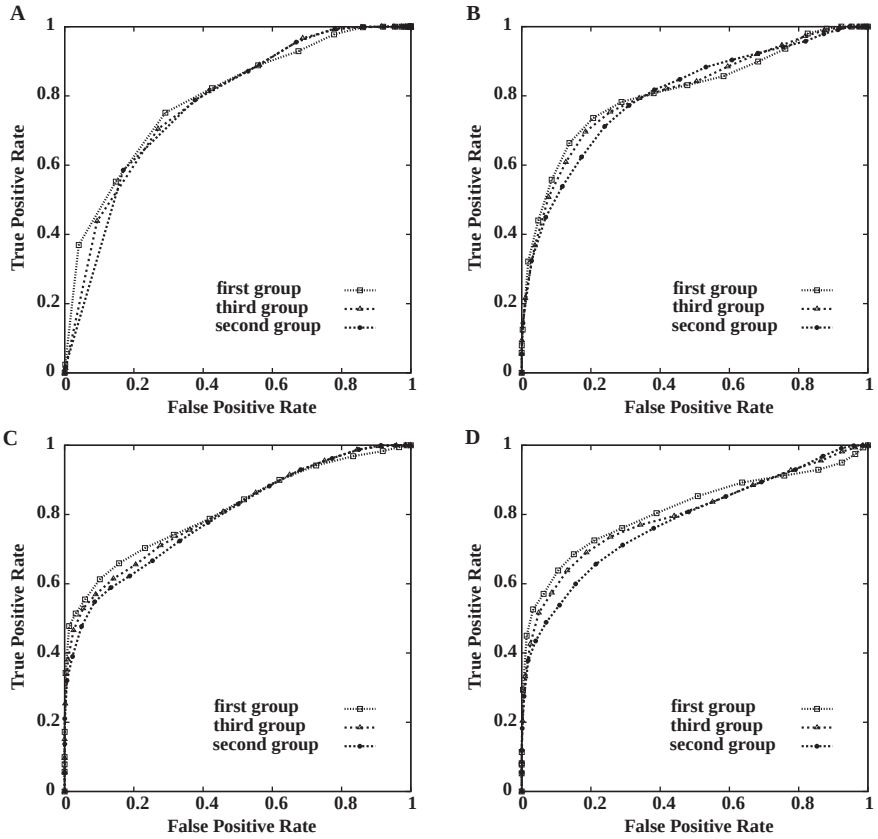

Fig. 4. Relationship between the prediction accuracy and the phylogenetic distance. The ROC curves for ρ^{MIRROR}, ρ^{16S}, ρ^{AVE} and ρ^{PC1} are respectively shown in (**A**)–(**D**). Open squares, closed circles and open triangles indicate the ROC curves of the first, second and third groups, respectively.

for ρ^{16S}, ρ^{AVE} and ρ^{PC1} deviated upward to that of ρ^{MIRROR} when the rates of false positives were small. However, when the rates of false positives increased the relationship was inverted and the curve of ρ^{MIRROR} was above those of ρ^{16S}, ρ^{AVE} and ρ^{PC1}. Considering actual applications, we are supposed to select pairs of proteins with high correlation coefficients as candidates for interacting partners. The result of the analysis with the ROC curve, together with the observation of the decreases in the patterns of correlation coefficients, suggests that our method realizes a high true positive rate and a low false positive rate for pairs of proteins showing high correlation. This would be a benefit of our prediction method, even when considering the deficiency of the higher ratio of false negatives than the original mirror tree method.

3.3 Prediction accuracy and distance between species

We finally examined how much the prediction accuracy is influenced by the closeness among the source organisms to be used in the data. Following is the procedure for the analysis.

(1) Randomly select 20 organisms from the 41 source organisms.

(2) Compute the average of the distances over all possible pairs of 20 organisms, based on the 16S rRNAs.

(3) Repeat (1) and (2) 10 000 times and generate the distribution of 10 000 average distances.

(4) Classify the sets into three groups based on the distribution: the first group (upper 5% of the distribution), the second group (lower 5% of the distribution) and the third group (the rest).

Note that the first group consisted of the sets of distantly related organisms, whereas the closely related organisms constituted the second group.

For each group, ρ^{MIRROR}, ρ^{16S}, ρ^{AVE} and ρ^{PC1} were calculated, and the corresponding ROC curves were drawn for the three groups from 20 different threshold values (Fig. 4). The rates of the false and true positives calculated at each threshold value were averaged and were then used to draw the ROC curve, as described above. As shown in the figure, the performance of the first group was better than those of the second and third groups, in terms of the false positive rates. This observation suggests that the inclusion of proteins from distantly related sources increases the reliability of the correlation coefficients for the detection of co-evolutionary behavior. The inclusion of distantly related sources would be required to accurately estimate the unit vector $|u\rangle$ used to construct the projection operator.

4 CONCLUSION

The mirror tree method is an outstanding approach for the prediction of protein–protein interactions. The approach with co-evolutionary information has introduced new perspectives into the computational

analyses of protein–protein interactions, which were mainly investigated by comparisons of genomic contexts. In this paper we presented several methods to improve the performance of the original mirror tree method by controlling for the phylogenetic relationships among the sources with the projection operator. In the experiment, we confirmed that our methods could drastically reduce the number of false positives in the prediction. We also showed that the inclusion of proteins from distantly related sources could improve the prediction accuracy.

Our method generated more false negatives than the original mirror tree method. As described above, we speculated that the number of interacting partners could be the reason for the increased number of false negatives. However, if we select protein pairs with a high correlation coefficient, say >0.8, by using our method, then we can predict with high reliability that the protein pair is interacting or is physically in contact.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, the Japan Society for the Promotion of Science and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Conflict of Interest: none declared.

REFERENCES

Dandekar,T. *et al*. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.

Enright,A. *et al*. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Felsenstein,J. (2004) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Gavin,A. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gertz,J. *et al*. (2003) Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, **19**, 2039–2045.

Goh,C. *et al*. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.

Gustafson,A. *et al*. (2005) ASRP: the Arabidopsis Small RNA Project Database. *Nucleic Acids Res.*, **33**, D637–D640.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jones,D. *et al*. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Kishino,H. and Hasegawa,M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.*, **29**, 170–179.

Nooren,I.M. and Thornton,J.M. (2003) Diversity of protein–protein interactions. *EMBO J.*, **22**, 3486–3492.

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.*, **14**, 609–614.

Pazos,F. and Valencia,A. (2002) *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227.

Pellegrini,M. *et al*. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Ramani,A. and Marcotte,E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**, 273–284.

Salwinski,L. *et al*. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Sprinzak,E. *et al*. (2003) How reliable are experimental protein–protein interaction data? *J. Mol. Biol.*, **327**, 919–923.

Tan,S. *et al*. (2004) ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. *Nucleic Acids Res.*, **32**, W69–W72.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.