BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 20 2006, pages 2488–2492
doi:10.1093/bioinformatics/btl419

Sequence analysis

Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions

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

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan, [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan and [3]Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, 3-1-1, Maidashi, Higashi-ku, Fukuoka 812-8582, Japan

Received on March 3, 2006; revised and accepted on July 27, 2006
Advance Access publication July 31, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: The computational prediction of protein–protein interactions is currently a major issue in bioinformatics. Recently, a variety of co-evolution-based methods have been investigated toward this goal. In this study, we introduced a partial correlation coefficient as a new measure for the degree of co-evolution between proteins, and proposed its use to predict protein–protein interactions.

Results: The accuracy of the prediction by the proposed method was compared with those of the original mirror tree method and the projection method previously developed by our group. We found that the partial correlation coefficient effectively reduces the number of false positives, as compared with other methods, although the number of false negatives increased in the prediction by the partial correlation coefficient.

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

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

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

1 INTRODUCTION

The use of co-evolutionary information is one of the popular approaches for predicting a protein–protein interaction (PPI). This approach is based on the assumption that interacting protein pairs are likely to evolve in a correlated fashion, and that the phylogenetic trees of the interacting proteins are similar to each other, owing to the co-evolution. Based on several previous reports, such as that by Goh *et al*. (2000), Pazos and Valencia (2001) developed a method to evaluate the intensity of co-evolution by comparing a pair of distance matrices, or a set of genetic distances, instead of phylogenetic trees. The intensity of co-evolution is calculated with the Pearson's correlation coefficient between the distance matrices (for details, see Methods). This is referred to as the 'mirror tree' method, according to the assumption. In recent years, some variants of the mirror tree method have been proposed (Gertz *et al*., 2003; Goh and Cohen, 2002; Kim *et al*., 2004; Ramani and Marcotte, 2003; Tan *et al*., 2004). One of the latest improvements in the mirror tree framework is the reduction of false positives, which is facilitated by removing the information about the evolutionary relationships of the organism sources from the distance matrices. It has been pointed out that predictions by the mirror tree methods tend to introduce many false positives. One explanation for this phenomenon is that, in the mirror tree method, the distance matrices of all of the proteins are constructed with a set of orthologues from the same organism sources. As a result, all of the distance matrices share the information about the evolutionary relationship of the sources, which is suspected of being the cause of the false positives. Sato *et al*. (2005) and Pazos *et al*. (2005) independently developed methods to remove such evolutionary information from the distance matrices, and demonstrated that the approaches can dramatically reduce the number of false positives in the prediction.

In this manuscript, we propose the use of the partial correlation coefficient as a novel measure for the intensity of co-evolution, to improve the original mirror tree approach. From the definition (see Methods), the partial correlation coefficient is expected to be useful to remove the information about the evolutionary relationship of the sources. The performance of the partial correlation coefficient, in terms of reducing the false positives from the prediction, was improved, as compared with that reported by Sato *et al*. (2005). We also examined the relationship between the accuracy of the PPI prediction by the co-evolutionary analysis and the number of proteins used in the study.

2 METHODS

2.1 Dataset

We selected 821 interacting pairs of *Escherichia coli* proteins that were experimentally determined, from the Database of Interacting Proteins (DIP) Version 01/16/2006 (Salwinski *et al*., 2004). The selected pairs did not include homo-interactions, and comprised 179 different *E.coli* proteins. All of the proteins are registered in KEGG/KO (Kanehisa *et al*., 2006), a database of putative orthologous sequences identified from fully sequenced genome data. Then, putative orthologues corresponding to 179 proteins derived from *E.coli* were collected from 65 different bacterial species, according to the description in KEGG/KO. Hereafter, the putative orthologues are simply referred to as the orthologues. The lists of PPI pairs, protein names and source organisms are provided in the Supplementary

*To Whom correspondence should be addressed.

data. One of the assumptions in this study is that a pair of proteins from a bacterium, which are each orthologous to the interacting proteins of *E.coli*, also interacts in the bacterial species. The other assumption is that the interaction affects the co-evolution of the orthologues.

2.2 Multiple sequence alignment and distance matrix

A multiple alignment of each set of selected orthologous proteins was performed with the alignment software, MAFFT (Katoh *et al*., 2005). A distance matrix for the orthologues was calculated from the multiple sequence alignment. Suppose that we want to obtain a distance matrix, D_{A}, for protein A. The orthologues of protein A are collected from n species. 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, $D_{\mathrm{A}}(x, y)$, represents the genetic distance of protein A between species x and y. The distance was calculated as a maximum likelihood estimate, with the PROTDIST module in the PHYLIP package (Felsenstein, 2004). The score table by Jones *et al*. (1992) was used for the maximum likelihood estimation. The distance matrix is symmetric, and only the upper or lower triangular part of the matrix is used in the prediction process.

2.3 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. This operation was applied to all of the proteins, and the corresponding distance 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'. As described below, n was set to 66 in this study. Let us consider the pair of protein i and protein j, and their phylogenetic vectors $|\nu_i\rangle$ and $|\nu_j\rangle$, transformed from the corresponding distance matrices D_i and D_j, where the subscripts i and j indicate different sets of orthologues. Then, we applied the normalization of the elements of each vector with the average and the standard deviation of the elements as follows:

$$|\nu_i^{\star}\rangle = \frac{|\nu_i\rangle - |\mu\rangle}{\sqrt{\mathrm{Var}(\nu_i)}}, \tag{1}$$

where $|\mu\rangle$ is a vector with the same dimension as $|\nu_i\rangle$. All of 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 of the elements of $|\nu_i\rangle$. The inner product of a pair of normalized vectors is reduced to the Pearson's correlation coefficient used for the mirror tree method. Hereafter, the correlation coefficient is denoted as $\rho_{ij}^{\mathrm{MIRROR}}$.

$$\rho_{ij}^{\mathrm{MIRROR}} = \langle \nu_i^{\star} \mid \nu_j^{\star} \rangle. \tag{2}$$

2.4 Projection operator

Here we will briefly review the projection method proposed in our previous manuscript (Sato *et al*., 2005). In this method, we use a unit vector $|u\rangle$, which represents the evolutionary relationship of the species. We designed the unit vector in three different ways: (1) transforming the distance matrix of 16S ribosomal RNA (rRNA) from the same source organisms as for the proteins under consideration, (2) averaging the phylogenetic vectors and (3) extracting the principal components of the phylogenetic vectors. The unit vectors obtained in these three ways were respectively designated as $|u_{16S}\rangle$, $|u_{\mathrm{AVE}}\rangle$ and $|u_{\mathrm{PC1}}\rangle$.

Using the unit vector $|u\rangle$, a projection operator P is obtained as

$$P = I - |u\rangle\langle u|, \tag{3}$$

where $|u\rangle\langle u|$ is also a projection operator onto the direction of the unit vector $|u\rangle$. I represents an identity matrix with the same size as $|u\rangle\langle u|$. By applying the projection operator (3) to a phylogenetic vector, say, $|\nu_i\rangle$, the component orthogonal to $|u\rangle$ is obtained as follows:

$$|\varepsilon_i\rangle = P|\nu_i\rangle = |\nu_i\rangle - |u\rangle\langle u|\nu_i\rangle. \tag{4}$$

The projection operator can exclude the information about the evolutionary relationship among the source organisms 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 (4) 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

$$\rho_{ij}^{\mathrm{PROJECTION}} = \langle \varepsilon_i^{\star} | \varepsilon_j^{\star} \rangle \tag{5}$$

represents the Pearson's correlation coefficient between the residues, after excluding the information about the evolutionary relationship from the original phylogenetic vectors. The Pearson's correlation coefficients between the residual vectors for proteins i and j, based on the unit vectors $|u_{16S}\rangle$, $|u_{\mathrm{AVE}}\rangle$ and $|u_{\mathrm{PC1}}\rangle$ in the construction of the projection operator P, are represented by ρ_{ij}^{16S}, ρ_{ij}^{AVE} and ρ_{ij}^{PC1}, respectively.

2.5 Partial correlation coefficient

Suppose that we have m proteins and we want to detect interacting pairs among them. Let us consider multiple regressions of $|\nu_i\rangle$ and $|\nu_j\rangle$ with $(m-2)$ phylogenetic vectors as follows:

$$|\nu_i\rangle = \alpha_0 + \sum_{k \neq i,j}^{m} \alpha_k |\nu_k\rangle + |\delta_i\rangle, \tag{6}$$

$$|\nu_j\rangle = \beta_0 + \sum_{l \neq i,j}^{m} \beta_l |\nu_l\rangle + |\delta_j\rangle, \tag{7}$$

where α_i and β_j are scalar parameters. The linear combination of $(m-2)$ phylogenetic vectors is expected to represent the evolutionary relationship of the source organisms, since the principal component analysis of a number of phylogenetic vectors suggested that the first principal component vector, with a cumulative rate of contribution >80%, represents the evolutionary relationship (data not shown). Therefore, the residual vectors, $|\delta_i\rangle$ and $|\delta_j\rangle$, essentially lack the evolutionary information of the source organisms. $|\nu_j\rangle$ is excluded from the summation on the right side of the first equation. If protein i, represented by $|\nu_i\rangle$, co-evolves with protein j, represented by $|\nu_j\rangle$, through a PPI, then the effect of the co-evolution with protein j is expected to be present in the residual vector $|\delta_i\rangle$. Likewise, the effect of co-evolution with protein i, if any, is expected to be present in $|\delta_j\rangle$. Therefore, the similarity between the two residual vectors is considered to indicate the extent of co-evolution between proteins i and j. To evaluate the similarity between the residual vectors, we normalized the vectors with the averages and the standard deviations, and we expressed them as $|\delta_i^{\star}\rangle$ and $|\delta_j^{\star}\rangle$. The inner product between the normalized residual vectors is a measure to indicate the similarity between them, and is called the partial correlation coefficient. That is, the partial correlation coefficient $\rho_{ij}^{\mathrm{PARTIAL}}$ between $|\nu_i\rangle$ and $|\nu_j\rangle$ is expressed as follows:

$$\rho_{ij}^{\mathrm{PARTIAL}} = \langle \delta_i^{\star} | \delta_j^{\star} \rangle. \tag{8}$$

Instead of constructing the above multiple regression models in each computation, the following formula was used to obtain the partial correlation coefficient.

$$\rho_{ij}^{\mathrm{PARTIAL}} = \frac{-(R^{-1})_{ij}}{\sqrt{(R^{-1})_{ii}(R^{-1})_{jj}}}, \tag{9}$$

where R is the correlation coefficient matrix whose (i, j)-th element is $\rho_{ij}^{\mathrm{MIRROR}}$, and the superscript -1 indicates the inverse of the matrix. For the derivation of Equation (9), see Supplementary data. When the subscripts, i and j, are omitted, $\rho^{\#}$ collectively represents the type of correlation coefficient indicated by the superscript.

Table 1. Comparison of the top 20 protein pairs sorted in decreasing order of the correlation coefficients

Rank	ρ^{MIRROR}		ρ^{16S}		ρ^{AVE}		ρ^{PC1}		$\rho^{PARTIAL}$	
1	rpoC - rpoB	0.9883*	sdhB - sdhA	0.9659*	sdhB - sdhA	0.9562*	sdhB - sdhA	0.9586*	nrdB – nrdA	0.9360*
2	rpoC - dnaN	0.9831	priA - ileS	0.9423	priA - ileS	0.9291	nrdB - nrdA	0.9268*	sdhB – sdhA	0.9103*
3	sdhB - sdhA	0.9820*	carB - murC	0.9398	nrdB - nrdA	0.9271*	priA - ileS	0.9242	trpB - trpA	0.6719*
4	rpoC - polA	0.9814*	rpoC - rpoB	0.9288*	rpoD - rpoC	0.9257*	carB - murC	0.9158	gltD - gltB	0.6708*
5	groL - rpoC	0.9794	nrdB - nrdA	0.9207*	carB - murC	0.9187	rpoD - rpoC	0.8953*	pstS - pstB	0.5679
6	priA - ileS	0.9790	rpoD - rpoC	0.9103*	rpoC - rpoB	0.8860*	murC - trxB	0.8685	atpE - atpB	0.5531*
7	rpoB - nusA	0.9786*	murC - trxB	0.9075	murC - trxB	0.8737	carB - trxB	0.8542	valS - gltA	0.5263
8	rpoB - rpoA	0.9783*	rpoC - dnaN	0.8976	malG - malF	0.8722*	rpoC - rpoB	0.8464*	groL - dnaK	0.5217*
9	groL - polA	0.9780	carB - trxB	0.8961	carB - trxB	0.8605	malG - malF	0.8386*	hemC - murA	0.5106
10	carB - murC	0.9777	hflB - rho	0.8925	rpoD - gyrB	0.8506	rpoD - gyrB	0.8375	gltA - pyrD	0.4598
11	uvrC - uvrB	0.9771*	rpoC - nusG	0.8889*	priA - serS	0.8438	priA - serS	0.8375	malE - malG	0.4474*
12	dnaX - dnaG	0.9765	rpoB - nusA	0.8752*	rpoD - rpoB	0.8408*	serS - ileS	0.8330	sucC - sucD	0.4373*
13	rpoB - dnaN	0.9763	rpoD - rpoB	0.8716*	serS - ileS	0.8398	ileS - clpP	0.8310	rplK - valS	0.4215
14	rpoC - nusG	0.9760*	priA - serS	0.8711	rpoB - nusA	0.8297*	gltD - gltB	0.8245*	valS - leuS	0.4039
15	uvrA - dnaX	0.9759	serS - ileS	0.8689	ileS - clpP	0.8293	ileS - sucB	0.8155	leuS - purB	0.3769
16	rpoC - gyrB	0.9756	malG - malF	0.8678*	rpoC - dnaN	0.8230	sucC - sucD	0.8043*	hflB - rho	0.3603
17	hflB - rho	0.9754	prlA - polA	0.8626	ileS - sucB	0.8229	hflB - rho	0.8030	dnaK - serS	0.3601
18	rpoB - polA	0.9746*	ileS - clpP	0.8555	rpoD - dnaN	0.8187	rpoB - nusA	0.7999*	gltX - leuA	0.3578
19	rpoA - polA	0.9745*	rpoB - rpoA	0.8550*	gltD - gltB	0.8157*	priA - clpP	0.7838	infA - murA	0.3441
20	uvrA - lepA	0.9745	murC - glmS	0.8548	hflB - rho	0.8141	greA - pnp	0.7838	rpsN - ilvB	0.3418*

3 RESULTS AND DISCUSSION

3.1 Prediction of protein–protein interaction by co-evolutionary analysis

To evaluate the intensity of the co-evolution between proteins, we calculated five types of correlation coefficients, ρ^{MIRROR}, ρ^{16S}, ρ^{AVE}, ρ^{PC1} and $\rho^{PARTIAL}$, for all of the possible pairs of 179 proteins, i.e. 15 931 pairs of proteins. In this analysis, the unit vectors required for the calculations of ρ^{AVE} and ρ^{PC1} were obtained by the average operation or the principal component analysis of the phylogenetic vectors of the 179 proteins, whereas 177 proteins were used as the explanatory variables of multiple regression for the calculation of the partial correlation coefficient between the remaining two proteins. The correlation coefficients, sorted in decreasing order, are listed in the Supplementary data, and only the top 20 members of the lists are shown in Table 1. Interacting pairs registered in DIP are highlighted by asterisks in the table. As shown in the table, the top 20 positions for either method were occupied by 8 or 9 interacting protein pairs. In the case of $\rho^{PARTIAL}$, however, the interacting pairs were the most abundant near the top of the list, as compared to the other methods. The fifth position of the list for $\rho^{PARTIAL}$ is occupied by the pair, pstS and pstB. There is no description about the interaction between these proteins in DIP. However, both proteins are involved in the phosphate transport system, according to their KEGG/KO annotations. Therefore, the proteins may interact with each other. Even if we use a high value, say 0.9, as a threshold for the correlation coefficient to predict a PPI, ρ^{MIRROR} will produce many pairs with high scores, including many non-interacting proteins, leading to more false positives in the actual prediction. This observation agrees with the results of our previous work (Sato *et al*., 2005). In contrast, the presence of interacting proteins in the top 20 list and the rapid decreases of ρ^{16S}, ρ^{AVE}, ρ^{PC1} and $\rho^{PARTIAL}$ guarantee the accuracy of prediction with the correlation coefficients, if the threshold is set at a sufficiently high value. Out of the four correlation coefficients, the decrease of $\rho^{PARTIAL}$ was the steepest.

Table 2. Sensitivity and Specificity of the prediction

Method	Sensitivity				Specificity			
	0.9	0.8	0.7	0.6	0.9	0.8	0.7	0.6
ρ^{MIRROR}	23.02	60.29	80.15	87.09	7.54	6.72	6.32	5.81
ρ^{16S}	0.49	2.68	7.31	13.28	57.14	22.68	13.57	8.94
ρ^{AVE}	0.37	1.34	3.29	6.58	60.00	45.83	27.27	18.31
ρ^{PC1}	0.24	0.85	2.31	5.24	50.00	41.18	26.03	18.38
$\rho^{PARTIAL}$	0.24	0.24	0.24	0.49	100.00	100.00	100.00	100.00

Sensitivity $=$ (True positive/(True positive $+$ False negative)) $\times 100\%$, Specificity $=$ (True positive/(True positive $+$ False positive)) $\times 100\%$.

To examine the effect of the threshold values on the prediction accuracy, we evaluated the performance by introducing four different thresholds, 0.9, 0.8, 0.7 and 0.6, for the correlation coefficient in each method. The performances of the original mirror tree method and our four proposed methods were compared in terms of the sensitivity and the specificity, and the results are summarized in Table 2. In each case, when a pair of proteins had a correlation coefficient larger than a threshold, the pairs of proteins were predicted to interact with each other. Table 2 shows that $\rho^{PARTIAL}$ provided extremely high specificity at any threshold, while ρ^{MIRROR} provided high sensitivity in all of the cases, except for the threshold $= 0.9$. The high specificities of $\rho^{PARTIAL}$ mean that the use of a partial correlation coefficient helps to drastically reduce the false positives, as compared to those of ρ^{MIRROR}. This is an important feature for practical use in actual applications, because we would start the prediction with the part with high confidence; that is, from the top of the ranked list of predictions.

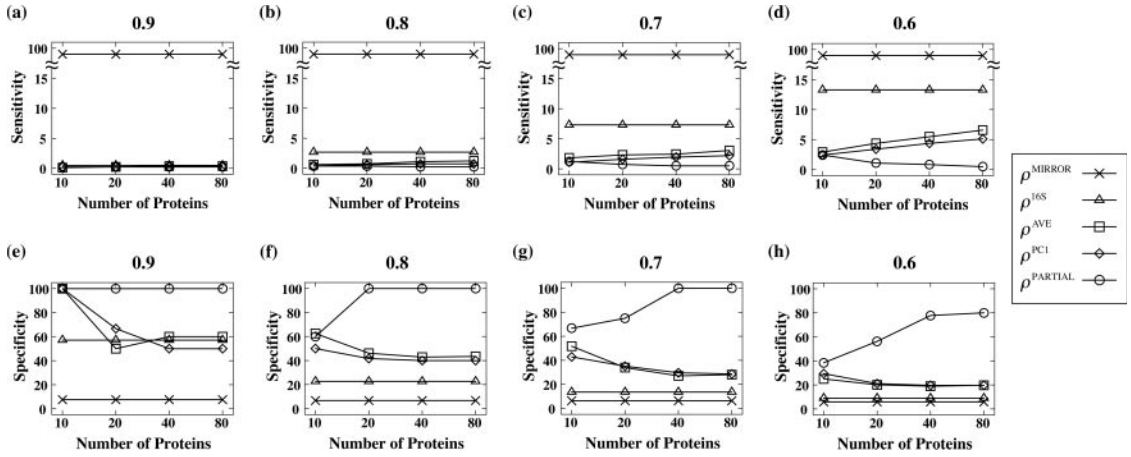

Fig. 1. The relationship between the prediction accuracy and the number of proteins. The *x*-axis indicates the number of proteins used in the analyses, and the *y*-axis indicates the sensitivity (**a**)–(**d**) or specificity (**e**)–(**h**). The performance of the five types of correlation coefficients was evaluated based on several threshold values, which are denoted at the top of each panel.

In other words, we would not be interested in the part of the prediction with lower confidence; that is, the lower ranking members of the prediction, in a practical situation. Table 2 also shows that the specificities of ρ^{16S}, ρ^{AVE} and ρ^{PC1} were much higher than those of ρ^{MIRROR}, but less than those of ρ^{PARTIAL}. On the other hand, the sensitivities of ρ^{16S}, ρ^{AVE} and ρ^{PC1} were higher than those of ρ^{PARTIAL}, but less than those of ρ^{MIRROR}. These observations about ρ^{16S}, ρ^{AVE} and ρ^{PC1} agree with those in our previous report with regard to the sensitivity and the specificity of the correlation coefficients (Sato *et al*., 2005). Thus, the partial correlation coefficient was more effective for the reduction of false positives from the prediction of PPI by the analysis of co-evolution than the previously proposed methods.

3.2 Prediction accuracy and the number of proteins

Out of the five types of correlation coefficients used in this study, it is clear from the definitions that ρ^{AVE}, ρ^{PC1} and ρ^{PARTIAL} are dependent on the number of proteins used for the calculations. In contrast, ρ^{MIRROR} and ρ^{16S} are independent of the number of proteins. We examined the relationship between the prediction accuracy and the number of proteins by the following procedure.

(1) Select m proteins randomly from the 179 proteins.

(2) Compute the five types of correlation coefficients for every possible pair of the m proteins. As for ρ^{AVE} and ρ^{PC1}, the unit vector $|u\rangle$ was estimated with the phylogenetic vectors of the m proteins, whereas ρ^{PARTIAL} for a pair was calculated with the remaining $(m-2)$ proteins as explanatory variables.

(3) Store the correlation coefficients in the five bins corresponding to the different types of correlation coefficients for each pair.

(4) Repeat (1–3) until each bin for all of the possible pairs from the 179 proteins includes 100 correlation coefficients.

The mean value of 100 correlation coefficients in a bin was used as the corresponding type of correlation coefficient between a pair of proteins. When the mean correlation coefficient between two proteins was larger than a given threshold value, the proteins were predicted to interact with each other. We examined four cases of m, 10, 20, 30 and 40, in this study. For each case, the sensitivity and specificity for the five types of correlation coefficients were calculated. The sensitivity and the specificity are shown as the functions of the number of proteins in Figure 1.

The specificity for ρ^{PARTIAL} increases and those for ρ^{AVE} and ρ^{PC1} decrease, as the number of proteins increases. In contrast, ρ^{MIRROR} and ρ^{16S} have constant values, since they are independent from m, as described above. In any case, however, the specificity of ρ^{PARTIAL} is basically larger than or equal to those of other types of correlation coefficients. The difference in the dependence on the number of proteins between ρ^{PARTIAL} and other methods suggests that the prediction with ρ^{PARTIAL} is more useful for the reduction of false positives, when a large number of proteins is available for the prediction. At the same time, ρ^{PARTIAL} is usually effective in reducing the false positives, even when the number of proteins is small.

The sensitivities for ρ^{16S}, ρ^{AVE}, ρ^{PC1} and ρ^{PARTIAL} were quite low, as compared to that of ρ^{MIRROR}. The sensitivities for ρ^{MIRROR} and ρ^{16S} were constant. In contrast, the sensitivities for the remaining three types of correlation coefficients showed a dependence on the number of proteins. As the number of proteins decreases, the sensitivity for ρ^{PARTIAL} increases, whereas those for ρ^{AVE} and ρ^{PC1} decrease. In any case, the sensitivity of ρ^{PARTIAL} was quite small and lower than those of the other types of correlation coefficients. Although the analyses suggested that the prediction with ρ^{PARTIAL} includes more false negatives than other methods, the high specificity for ρ^{PARTIAL} is considered to compensate for such deficits.

4 CONCLUSION

In this report, we demonstrated the potential of the partial correlation coefficient as a new measure for the intensity of co-evolution between proteins. In the numerical experiment, we showed that the performance of the partial correlation coefficient is better than or comparable with that of the projection method developed by Sato *et al*. (2005) and the original mirror tree method. We also examined the relationship between the prediction accuracy and the number of

proteins used for the prediction analysis. The specificity for the prediction with ρ^{PARTIAL} was the highest among the five types of correlation coefficients, when a large number of sequences was available. Even when the number of proteins was small, the specificity for ρ^{PARTIAL} was better than those by the other methods. However, the prediction with ρ^{PARTIAL} included many false negatives, like the prediction by the projection methods. This problem should be examined for further improvement of PPI prediction by co-evolutionary analysis.

ACKNOWLEDGEMENTS

This work was supported by Grants-in-Aid for Scientific Research on Priority Areas 'Systems Genomics' (K.H.), 'Comprehensive Genomics' (M.K.) and 'Membrane Interface' (H.T.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Conflict of Interest: none declared.

REFERENCES

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

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.

Goh,C. and Cohen,F.E. (2002) Co-evolutionary analysis reveals insights into protein–protein interactions. *J. Mol. Biol*., **324**, 177–192.

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*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.

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

Kim,W. *et al*. (2004) Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, **20**, 1138–1150.

Pazos,F. *et al*. (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol*., **352**, 1002–1015.

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

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.

Sato,T. *et al*. (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, **21**, 3482–3489.

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