BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 92–98
doi:10.1093/bioinformatics/bts620

Systems biology

Advance Access publication October 18, 2012

Human protein–protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence

Chia Hsin Liu, Ker-Chau Li and Shinsheng Yuan*

Institute of Statistical Science, Academia Sinica, Nangang, Taipei 115, Taiwan

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Protein–protein interaction (PPI) plays an important role in understanding gene functions, and many computational PPI prediction methods have been proposed in recent years. Despite the extensive efforts, PPI prediction still has much room to improve. Sequence-based co-evolution methods include the substitution rate method and the mirror tree method, which compare sequence substitution rates and topological similarity of phylogenetic trees, respectively. Although they have been used to predict PPI in species with small genomes like *Escherichia coli*, such methods have not been tested in large scale proteome like *Homo sapiens*.

Result: In this study, we propose a novel sequence-based co-evolution method, co-evolutionary divergence (CD), for human PPI prediction. Built on the basic assumption that protein pairs with similar substitution rates are likely to interact with each other, the CD method converts the evolutionary information from 14 species of vertebrates into likelihood ratios and combined them together to infer PPI. We showed that the CD method outperformed the mirror tree method in three independent human PPI datasets by a large margin. With the arrival of more species genome information generated by next generation sequencing, the performance of the CD method can be further improved.

Availability: Source code and support are available at http://mib.stat.sinica.edu.tw/LAP/tmp/CD.rar.

Contact: syuan@stat.sinica.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 18, 2012; revised on October 7, 2012; accepted on October 16, 2012

1 INTRODUCTION

Protein–protein interactions (PPI) regulate many fundamental cellular processes (Alberts, 1998). One key step in understanding the function of a protein is to identify its potential interacting partners (Uetz, *et al.,* 2000; Walhout and Vidal, 2001). PPI identification and prediction have become an important topic in recent decades (Auerbach *et al.,* 2002; Bader and Hogue, 2003; Chen and Xu, 2003; Hart *et al.,* 2006; Jansen, 2003; Ramani and Marcotte, 2003; Rhodes *et al.,* 2005; Scott and Barton, 2007; von Mering *et al.,* 2002).

Up to date, many experimental methods to identify physical interaction between two proteins have been proposed, including the yeast two-hybrid system (Ito *et al.,* 2001), protein-fragment complementation assay (Michnick *et al.,* 2006), affinity purification-mass spectrometry (Bauer and Kuster, 2003), protein microarray and fluorescence resonance energy transfer (Werther *et al.,* 2008). The results from these experiments have been deposited in several public databases; for instance, HPRD (Keshava Prasad *et al.,* 2009), DIP (Salwinski *et al.,* 2004), IntAct (Aranda *et al.*, 2010), BioGRID (Stark *et al.,* 2006) and MINT (Ceol *et al.,* 2010). It has been suggested that there are ~300 000 PPI out of a total of ~300 000 000 protein pairs in the human proteome (Hart *et al.,* 2006). This estimate does not account for the numerous variations owing to post-translation modification or alternative splicing.

A number of computational techniques have been proposed to provide either complementary information or supporting evidence to experimental methods (Shoemaker and Panchenko, 2007); for example, phylogenetic profile (Pellegrini *et al.,* 1999), Rosetta stone (Enright *et al.,* 1999), conserved gene neighborhood analyses (Dandekar *et al.,* 1998), domain co-occurrence (Wuchty 2001; Wuchty and Almaas, 2005) and orthology information (Izarzugaza *et al.,* 2008; O'Brien *et al.,* 2005). These methods use domain, gene, functional pathway co-existence information or gene fusion event in related species to predict PPI. Sequence-based co-evolution is another approach, which has been proposed to predict PPI for many years. There are two types of methods in this approach: the site-specific method and the full-sequence method. The site-specific method detects sequence mutual changes in binding interfaces of interacting partners to infer PPI (Chen and Jeong, 2009; Wass *et al.,* 2011). But changes in such regions are hard to detect (Lovell and Robertson, 2010). Full-sequence methods, such as the mirror-tree method proposed by Pazos (Pazos and Valencia, 2001), compare a distance matrix between two proteins and use topological similarity of phylogenetic trees to predict PPI. In addition, Fraser and his collaborates (Fraser *et al.,* 2002) compared substitution rates between two proteins and concluded that interacting proteins evolve at a similar rate. However, both methods have not been applied to the whole large-scale PPI prediction in vertebrates.

In this article, we introduce a new method to exploit sequence-based co-evolution for PPI prediction in the human proteome. Our method is based on two basic assumptions. First, shared selection pressure may force two interacting proteins to co-evolve together. Changes in the sequence of one protein may induce appropriate changes in the sequence of its interacting partner to retain the binding affinity and maintain biological functions. Consequently, PPI pairs may have similar substitution rates. Second, protein interaction is more likely to

*To whom correspondence should be addressed.

conserve across related species. Thus, to study human proteins, we shall examine their orthologous proteins in all vertebrates whose genomes are already sequenced. Based on these two assumptions, we define the co-evolutionary divergence (CD) of a pair of proteins as the absolute value of the substitution rate difference between two proteins. The CD values of interacting protein pairs are expected to be smaller than those of non-interacting protein pairs. Therefore, we may use CD as a feature to distinguish interacting protein pairs from non-interacting protein pairs. We evaluate the performance of PPI prediction by the CD method and compare it with the mirror tree method. By examining three independent PPI datasets, we show that CD is a better method for human PPI prediction.

2 METHODS

2.1 Protein sequence and protein interaction dataset

In all, 16 229 proteins of *Homo sapiens* and their orthologous protein sequences in 13 vertebrates, *Canis lupus* (dog), *Bos taurus* (cow), *Equus caballus* (horse), *Gallus gallus* (chicken), *Macaca mulatta* (macaques), *Pan troglodytes* (chimpanzees), *Mus musculus* (mice), *Rattus norvegicus* (rat), *Monodelphis domestica* (opossums), *Oryzias latipes* (medaka), *Takifugu pardalis* (fugu), *Danio rerio* (zebrafish) and *Tetraodon nigrov*iridis, totaling 172 338 protein sequences, were obtained from Evola database (Matsuya *et al.,* 2008).

A protein interaction dataset was downloaded from the Human Protein Reference Database (Keshava Prasad *et al.,* 2009). Duplicate interactions and self-interactions were removed, resulting in 36 867 distinct human protein interactions for 8694 distinct human proteins. These interactions formed the golden standard positive dataset (GSP). The golden standard negative dataset (GSN) consisted of 2 750 990 protein pairs with one protein from the plasma membrane cellular component and the other from the nuclear cellular component. GSP and GSN were taken as our training dataset to construct the likelihood ratio table. Among 16 229 orthologous proteins, only 4116 orthologous proteins appear in all 14 vertebrates. Thus, only 3781 protein pairs of GSP and 160 160 protein pairs of GSN from HPRD database were used when comparing the CD method and the mirror tree method. This is because mirror tree method requires all orthologous proteins be in all species under consideration.

Our first independent test used the interacting protein dataset from GRID (Stark *et al.,* 2006). This dataset has 1497 interacting protein pairs with orthologous proteins in all 14 species. There are no overlapping pairs between this test dataset and the training dataset. Our second independent test dataset came from the protein complex database of CORUM (Ruepp *et al.,* 2010). There are 1846 protein complexes in this database. Complexes containing unspecific proteins, e.g. isoforms or splice variants whose sequences could not be specifically identified are eliminated. All proteins in the same complex are treated as having interaction with each other. There are 2738 distinct interacting protein pairs in our second testing dataset, and there are 389 distinct interacting protein pairs if we restrict protein complexes to sizes between 20 and 40 proteins.

2.2 Protein substitution rate estimation and CD of interacting protein pair

Pair-wise alignment of 172 338 orthologous proteins was made with software ClustalW2. For each protein with orthologs in n species, there are $n(n-1)/2$ pair-wise alignments. The number of amino acid substitutions per site, d, was used as substitution rate of protein. It was estimated by numerically solving the equation, $q = [\ln(1+2d)/2d]$, where q is the fraction of identical residues between two aligned sequences made by pairwise alignment (Grishin, 1995). More specifically, the pair-wise alignment by ClustalW outputs a conserved region of the sequence containing sites of 'identical', 'conserved substitution' and 'semi-conserved substitution'. 'Identical' means the amino acids of the two orthologous sequences at the site are identical. 'Conserved substitution' means the replacement of an amino acid residue by another one with similar properties, such as aspartate for glutamate. A semi-conserved amino acid replaces one residue with another one that has similar steric conformation, but does not share chemical properties. The value q is the fraction of identical amino acids in the conserved region. We use the absolute value of substitution rate difference between two proteins to measure the CD of interacting protein pair. The CD values of 36 867 protein pairs in GSP and 2 750 990 protein pairs in GSN were calculated. We used them to construct the likelihood ratio table of interacting protein pairs.

2.3 Likelihood ratio table construction of CD features

The CD values of 2 750 990 protein pairs in GSN were grouped into 10 bins of increasing substitution rate difference according to their quantiles (Supplementary Table S1). To calculate the likelihood ratio of protein interaction for a given bin, we divided the proportion of protein pairs of GSP with their CD values in the bin by the proportion of protein pairs of GSN with their CD values in the same bin. This yielded 91 CD likelihood ratios for describing the possibility of interaction for a protein pair (Supplementary Table S2).

2.4 The naive Bayes classifier

A naïve Bayes approach was used to combine the 91 individual likelihood ratios. The proportion of positive cases (P(pos)) and the proportion of negative cases ($P(neg) = 1 - P(pos)$) were used as the prior distribution. The 'posterior' odds after observing N features for prediction, $f1, \ldots, fN$, is $P(pos \mid f1, \ldots, fN) / P(neg \mid f1, \ldots, fN)$. The likelihood ratio L defined as $L(f1, \ldots, fN) = P(f1, \ldots, fN \mid pos) / P(f1, \ldots, fN \mid neg)$ relates the prior odds and posterior odds according to Bayes's rule. The naïve Bayes classifier assumes the conditional independence of N features to simplify the likelihood estimation. The likelihood L is simply the product of the N individual likelihood ratios obtained by considering each feature separately. This assumption is rarely true in most real world problems. However, if the dependence pattern of the joint density of N features under the positive set is similar to that under the negative set, then by cancelation, the true likelihood ratio will be approximately the same as the one calculated under the independence assumption. In Supplementary Figure S1, we compared each likelihood ratio for the pair-wise density conditional on the positive set over that conditional on the negative set. We showed that they were indeed very close to each other. Thus, in our application, the naïve Bayes classifier serves as a reasonable approximation to the optimal Bayes classifier.

2.5 Integration of 91 CD features to form the joint CD score by naïve Bayes

There were four steps involved. Given a protein pair, we first computed the CD value for each of the 91 combinations between 14 species. Then we used the 10-quantile (decile) table of CD distribution for GSN (Supplementary Table S1) to identify at which decile interval each of the 91CD values was located. The third step is to obtain the 91 likelihood ratios from the likelihood ratio table (Supplementary Table S2). If orthologous proteins were not present in a pair of species, we set the corresponding likelihood ratio to be 1. The final likelihood ratio, which we called the joint CD score, of each protein pairs was calculated by multiplying the 91 likelihood ratios. For example, if the CD of a protein pair is 0.02 between human and mouse, this value is between 10 and 20%, according to Supplementary Table S1. We could obtain the corresponding likelihood ratio is 1.4907 in Supplementary Table S2. Repeat the same

operation to obtain the other 90 CD values. Then we multiply all 91 likelihood ratio values to get the final score.

2.6 Mirror tree method: co-evolutionary constraint

The multiple alignment and the distance matrix for constructing phylogenetic tree of each orthologous protein were constructed by ClustalW. 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 certain order. When the distance 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 14. Therefore, the dimension of the phylogenetic vector is 91. The intensity of the co-evolutionary constraint between proteins A and B is evaluated by Pearson's correlation coefficient between phylogenetic vectors of A and B.

2.7 Accuracy measurement and receiver operating characteristic curve comparison

The receiver operating characteristic (ROC) curve was used to plot the fraction of true positives out of the positives (P) versus the fraction of false positives out of the negatives (N). The area under the receiver operating characteristic (AUROC) is a common summary statistic for the accuracy of a predictor in a binary classification task. Comparison of area under two ROC curves was made with 'pROC' package of R language (Robin *et al.,* 2011) using DeLong's non-parametric approach (DeLong *et al.,* 1988).

2.8 Null distribution of AUROC of the CD method

To obtain the null distribution of AUROC for the CD method, we randomly selected 30 774 protein pairs without self-interaction from all possible protein pairs, to form a null positive dataset. We used the 2 750 338 protein pairs in GSN as the negative dataset. We then calculated the AUROC value for this simulated null positive dataset versus the negative dataset. We repeated this process 1000 times to obtain the null distribution of AUROC. The result was shown in Figure 5b.

3 RESULTS

3.1 Protein substitution rate

The orthology relationships of proteins between human and 13 species were extracted from an ortholog information database: EVOLA (Matsuya *et al.,* 2008). The number of orthologous protein sequences between human and each species was shown in Figure 1. Chimpanzee has the largest number of human orthologs, whereas the numbers for chicken, medaka, tetradon, fugu and zebra fish are considerably lower than others. The distribution of protein substitution rate between human and other species was given in Figure 2. Chimpanzee and macaque showed the lowest average substitution rates, whereas chicken, medaka, tetradon, fugu, and zebra fish showed much higher rates.

3.2 Significant negative correlations between substitution rate and number of interacting partners in vertebrates

The negative correlation between substitution rate and the number of interacting partners of a protein was first reported in a study comparing *Saccharomyces cerevisiae* with *Caenorhabditis elegans*. (Fraser *et al.,* 2002) We investigated whether the same relationship still holds in vertebrates. We computed the Spearman correlation for each of the 91 combinations between 14 species. Remarkably, all of these correlations indeed turned out negative (Supplementary Table S3).

3.3 CD increases from chimpanzee to zebra fish

The CDs of interacting protein pairs between human and each of the other 13 vertebrates were computed. Figure 3a summarized the CD distribution for each human–vertebrate species pair. We found that the values increase from chimpanzee to zebra fish, in the same order as Figure 2. However, the distribution for non-interacting protein pairs (Fig. 3b) also showed the same pattern. This indicates the lack of power in using each CD feature separately for differentiating the interacting pairs from non-interacting pairs. We overcome the difficulty by combining all available CD features into a likelihood ratio.

3.4 CD score in gold standard positive dataset is significantly smaller than those in gold standard negative dataset

The joint CD score was computed for each protein pair in both GSP dataset and GSN dataset. In all, 6123 protein pairs in GSP and 652 protein pairs in GSN were eliminated because we could not find othologous protein pairs for them in any two species. This resulted in 30 744 protein pairs from the GSP dataset and 2 705 338 protein pairs from the GSN dataset. Figure 4 plotted the two distributions of joint CD scores. The Kolmogorov–Smirnov test was applied to test whether these two distributions were the same. We found the difference to be highly significant, and the distribution from gold standard positive is stochastically greater than that from gold standard negative. The accuracy of predicting PPI by joint CD scores, measured by the AUROC is 0.639 (Fig. 5a). This value is much larger than the AUROC values obtained from randomly generated null selected positive datasets, which fall between 0.510 and 0.530 based on 1000 simulations (Fig. 5b).

3.5 Comparison of the CD method and the mirror tree method

We compare the CD method and the mirror tree method. The mirror tree method requires that orthologous proteins must be present in all 14 species. Thus, both the positive set and the negative set used for comparison are subject to this constraint.

We first evaluated the performance on the HPRD dataset. Only 3781 interacting protein pairs with orthologous protein in all 14 species from GSP were used to compute the AUROC value for the CD method and that for the mirror tree method, respectively. The CD method had an AUROC value of 0.6078, significantly ($P < 2.2$ E–16) better than the AUROC value of 0.4352 for the mirror tree methods, Figure 6a.

As an independent testing dataset, we used 1497 interacting protein pairs from BIOGRID database. These pairs had no overlap with any of the 36 867 pairs in GSP. In this independent testing dataset, the CD method had an AUROC value of 0.6285, significantly ($P < 2.2$ E–16) better than the AUROC value of 0.4415 for the mirror tree methods, Figure 6b.

The mirror tree method has been applied to predict PPI with protein pairs from known protein complexes (Yang *et al.,* 2010). To see whether this method may work better than ours on

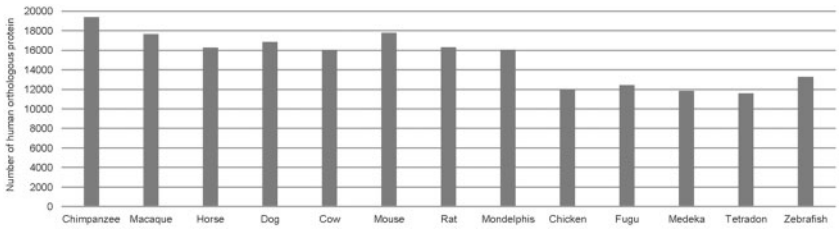

Fig. 1. Number of orthologous proteins in 13 vertebrates. The number of orthologous protein sequences between human and each species is shown. Chimpanzee has the largest number of human orthologs, whereas the numbers for chicken, medaka, tetradon, fugu and zebra fish are considerably lower than others

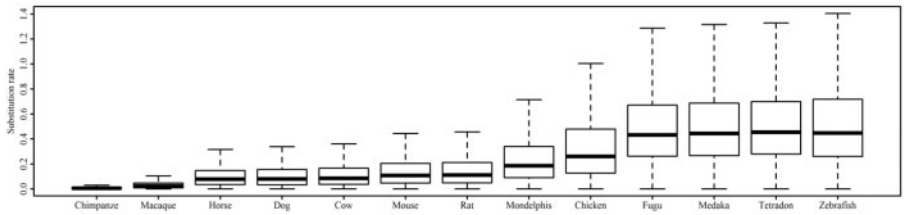

Fig. 2. Substitution rate of human protein calculated from the comparison with orthologs in 13 vertebrates. Distributions of the number of amino acid substitution per site (substitution rate) of orthologous proteins between human and each of the other 13 vertebrates are shown. Chimpanzee and zebra fish have the smallest and largest average substitution rates, respectively

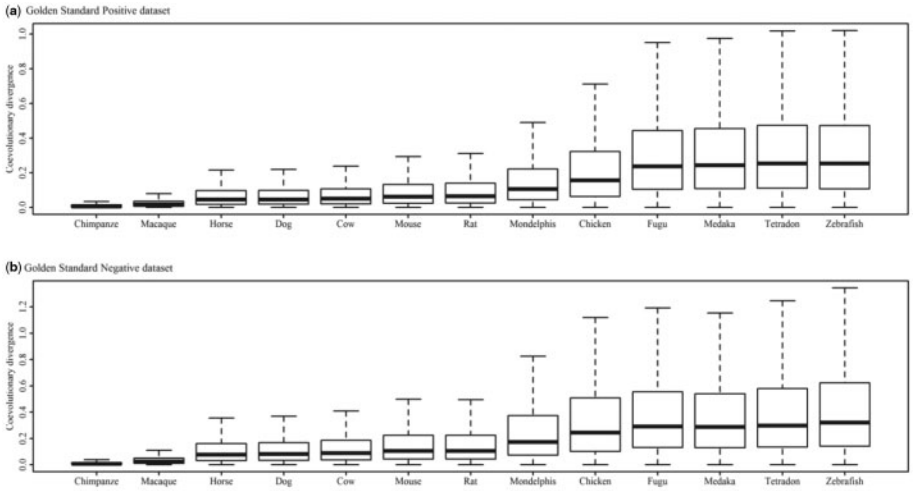

Fig. 3. CD of protein pairs in (**a**) GSP and (**b**) GSN dataset of 13 vertebrates. Chimpanzee and zebra fish have the smallest and largest average CD, respectively

protein complexes, our second independent testing dataset used protein complexes from CORUM as the positive PPI dataset. We found that the AUROC value of 0.714 for the CD method was still significantly ($P < 2.2\,\text{E}{-16}$) better than the AUROC value of 0.459 for the mirror tree method (Fig. 6c). The performance of mirror tree method is surprisingly low, although this method showed a good performance when applying to mitochondria respiratory chain proteins (Yang *et al.,* 2010). We also made another comparison by restricting protein complexes to sizes between 20 and 40 proteins. The result was similar. The AUROC values of our method and the mirror tree method were 0.783 and 0.426, respectively (Supplementary Fig. S2).

3.6 Protein pairs with high joint CD scores are likely to be functionally associated

We computed the joint CD score for each gene pair across the entire human genome and ranked all of the 131 682 106 pairs by their scores. Scanning the list from the top, we found that the highest score ($=6.8778\,\text{E}+13$) was obtained by 46 gene pairs (Supplementary Table S4a, rows 2–47). Several genes appeared multiple times in these pairs. The most visible one is a cluster of six subunits of the ribosomal protein complex (RPL27, RPL30, RPS14, RPS15A and RPS28), which appeared 23 times in total. In addition, there were proteasome subunits (PSMA4, PSMD7

and PSMD12) and DNA-directed RNA polymerase II subunit (POLR2C, POLR2H). It is also noteworthy to point out three genes involved in the histone modification, HDAC3 (histone deacetylase 3), RUVBL1 (Component of the NuA4 histone acetyltransferase complex) and RUNX1T1 (runt-related transcription factor 1; a gene well-known for the highly-frequent t(8;21)(q22,q22) translocation in acute myeloid leukemia). The protein–protein binding between RUNX1T1 and HDAC3 has been reported in the literature (Amann *et al.,* 2001). The CD score for this pair is remarkably high, ranking at the 154th place from the top. On the other hand, the alternation between histone deacetylation and histone acetylation underlines the dynamical change of chromatin structure, which is critical for regulating gene expression and other important nucleus biological events. Reaching the highest CD score by RUNBL1 and HDAC3 as well as RUNBL1 and RUNX1T1 reflects very well the importance of the co-evolution for genes that participate in two associated biological processes.

We set a score threshold of 10 E + 13 to investigate the functional sharing tendency between paired genes with high CD score. In all, 11 389 pairs involving 743 proteins scored higher than the threshold; see Supplementary Table S4 for the full list. The GO term system (http://go.princeton.edu/cgi-bin/GOTermMapper) was used to assign biological functions. By comparing with genes paired at random, our data showed an enrichment of GO term sharing for high CD score gene pairs; see Supplementary Figure S3. We concluded that the CD method is useful for inferring the functional association between genes.

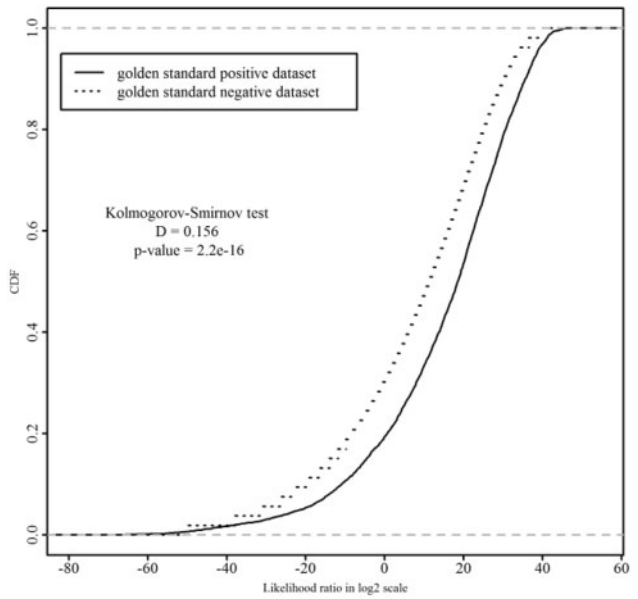

Fig. 4. The cumulative distribution of the likelihood ratio of protein pairs in the GSP dataset and that in the GSN dataset. The distribution curve from GSP set lies underneath that from GSN, indicating a significant difference between the two distributions (K-S test, *P*-value = 2.2 E–16)

4 DISCUSSION

Our CD method used the sequence-based co-evolution approach for PPI prediction. Although this method could rank the likelihood of interaction for a given pair of proteins, it did not infer specific features of interaction such as the interacting residues in the interfaces. Structure-based methods such as protein or domain docking methods emphasize the molecular interface between proteins or domains (Gray *et al.,* 2003; Heifetz *et al.,* 2002; Madaoui and Guerois, 2008; Norel *et al.,* 2001). However, the computing load of docking methods is too heavy to be applied for the large scale prediction in human genome. The CD method might provide a shorter list of candidate protein pairs for applying these methods.

One class of co-evolutionary models exploiting sequence covariations were used to predict interaction between residues within a protein sequence. For example, statistical coupling analysis (SCA) and mutual information successfully infer residues interactions in PDZ protein family and basic helix-loop-helix transcription factor, respectively (Atchley *et al.,* 2000; Lockless and Ranganathan, 1999). The assumptions of these two methods are similar to the CD method: if one protein interacts with the other protein in human, their orthologous proteins in other species are likely to have interactions too. The main difference is

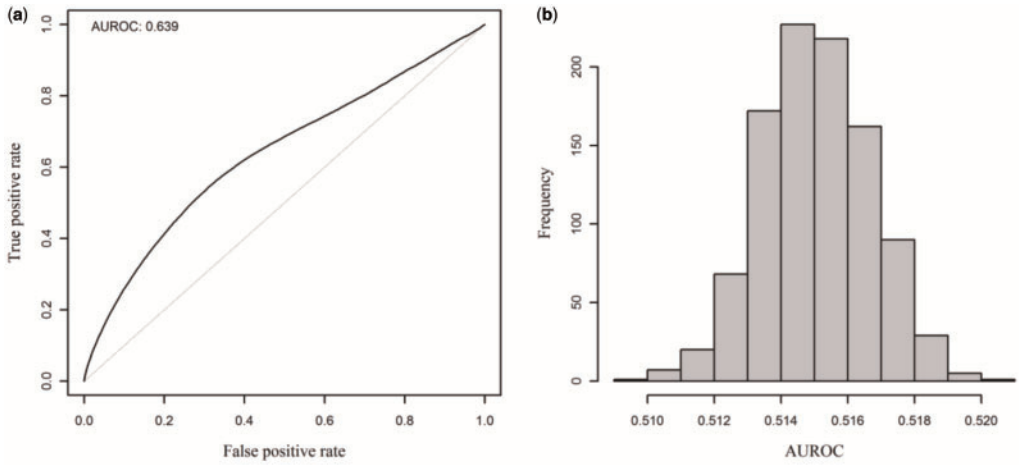

Fig. 5. (**a**) The ROC curve and the AUROC of training dataset. (**b**) The distribution of the AUROC values for randomly generated PPI pairs

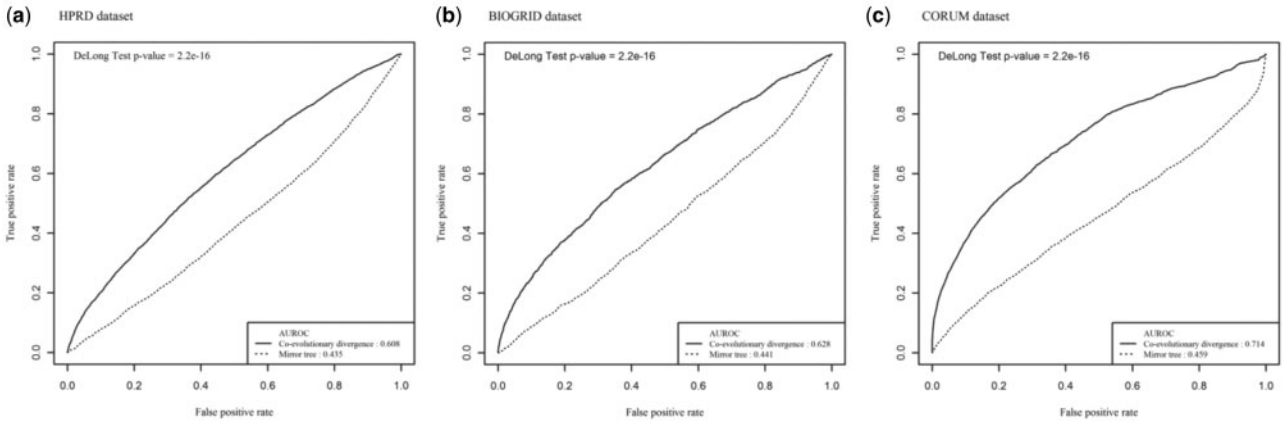

Fig. 6. ROC curves for the comparison of the CD method and the mirror tree method in three datasets. (**a**) HPRD dataset; (**b**) BIOGRID dataset; (**c**) CORUM dataset

that methods like SCA focus on the co-change between amino acids, whereas the CD method focuses on the rates of amino acid change between two sequences. Although methods like SCA provide computational methods to infer the residues interactions within protein sequence, they may be extended to infer residues interactions between protein sequences.

We demonstrated that our CD method performed much better than the mirror tree method in three independent datasets of interacting proteins in human. One major difference between the two methods is that the CD method is more sensitive to the absolute size difference of protein substitution rates. Because the mirror tree uses the correlation coefficient of the sequence substitution rate profiles across the selected species, it is only the relative size difference in the substitution rates that matters. Therefore, two protein families may have considerably different substitution rates across all branches of the evolution tree, whereas their profiles may still have high correlations. In such cases, they may not form protein–protein interactions, but are likely to be falsely predicted by the mirror tree method. The other difference is the alignment method used in comparing othologs. In the mirror tree method, the conserved regions identified by multiple alignment across all related species are used to generate the evolutionary profile for correlation analysis. The size of a conserved region would automatically get smaller when more species are used for alignment. This leads to the unfortunate consequence that less information could be used in estimating the rate of substitution between species. In the CD method, the conserved regions are identified separately by the pair-wise alignment for each species pair. The region sizes are larger, and the substitution rate estimation is more informative.

The negative correlation between protein substitution rate and number of interacting partner may be a potential source of false negatives in the CD method. Proteins with many interacting partners (hub protein) tend to have lower substitution rates. If a protein only interacts with one such hub protein and has no other interacting partner, its substitution rate may be higher. As a consequence, the CD of such protein pairs may have an upward bias. To investigate whether the bias is substantial, we compare the distribution of CD between PPI involving with hub proteins and not involving with hub proteins (Supplementary Figs S4 and S5). Although the CD of PPI involved with one hub protein is slightly greater than that of PPI not involving with any hub protein, the difference is not significant.

Our CD method used all 91 comparisons between pairs of 14 vertebrates. For human PPI prediction, an alternative version would be to construct the likelihood ratio from only the 13 comparisons involving human and other vertebrates. However, this did not yield a better performance (Supplementary Fig. S6). We also investigated the effect of bin size choice in constructing the likelihood ratio table. As shown in Supplementary Figure S7, the performance of our method is nearly unchanged when changing the bin size from 10% to 5%.

We conclude that the CD method may be a more direct way to combine co-evolutionary information of inter-acting protein pair from many species. The co-evolutionary diver-gence method does not use multiple alignment, thus taking less time than the mirror tree method. The mirror tree method is limited to those proteins with orthologous across all species under consideration. Thus, when more and more species genomes become available, less proteins could be applied. In contrast, increasing the number of species will provide more information to improve the accuracy of the co-evolutionary divergence method.

Conflict of Interest: none declared.

REFERENCES

Alberts,B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.

Amann,J.M. *et al.* (2001) ETO, a target of t(8;21) in acute leukemia, makes distinct contacts with multiple histone deacetylases and binds msin3a through its oligomerization domain. *Mol. Cell. Biol.*, **21**, 6470–6483.

Aranda,B. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

Atchley,W.R. *et al.* (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, **17**, 164–178.

Auerbach,D. *et al.* (2002) The post-genomic era of interactive proteomics: facts and perspectives. *Proteomics*, **2**, 611–623.

Bader,G. and Hogue,C. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Bauer,A. and Kuster,B. (2003) Affinity purification-mass spectrometry. *Eur. J. Biochem.*, **270**, 570–578.

Ceol,A. *et al.* (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
Chen,X.-w. and Jeong,J.C. (2009) Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, **25**, 585–591.
Chen,Y. and Xu,D. (2003) Computational analyses of high-throughput protein-protein interaction data. *Curr. Protein Pept. Sci.*, **4**, 159–181.
Dandekar,T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
DeLong,E.R. *et al.* (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.
Enright,A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
Fraser,H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
Gray,J.J. *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.
Grishin,N.V. (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.*, **41**, 675–679.
Hart,G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **7**, 120.
Heifetz,A. *et al.* (2002) Electrostatics in protein-protein docking. *Protein Sci.*, **11**, 571–587.
Ito,T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **10**, 4569–4574.
Izarzugaza,J. *et al.* (2008) Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, **9**, 35.
Jansen,R. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
Lovell,S.C. and Robertson,D.L. (2010) An integrated view of molecular coevolution in protein-protein interactions. *Mol. Biol. Evol.*, **27**, 2567–2575.
Madaoui,H. and Guerois,R. (2008) Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc. Natl Acad. Sci. USA*, **105**, 7708–7713.
Matsuya,A. *et al.* (2008) Evola: ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.*, **36** (**Suppl. 1**), D787–D792.
Michnick,S.W. *et al.* (2006) Chemical genetic strategies to delineate MAP kinase signaling pathways using protein-fragment complementation assays (PCA). *Methods*, **40**, 287–293.
Norel,R. *et al.* (2001) Electrostatic contributions to protein-protein interactions: fast energetic filters for docking and their physical basis. *Protein Sci.*, **10**, 2147–2161.
O'Brien,K.P. *et al.* (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D780.
Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, **14**, 609–614.
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.K. and Marcotte,E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**, 273–284.
Rhodes,D.R. *et al.* (2005) Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.*, **23**, 951–959.
Robin,X. *et al.* (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
Ruepp,A. *et al.* (2010) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **38** (**Suppl. 1**), D497–D501.
Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
Scott,M. and Barton,G. (2007) Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics*, **8**, 239.
Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein-protein interactions. part II. computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.
Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
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.
Walhout,A.J. and Vidal,M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell Biol.*, **2**, 55–62.
Wass,M.N. *et al.* (2011) Towards the prediction of protein interaction partners using physical docking. *Mol. Syst. Biol.*, **7**, 469.
Werther,M *et al.* (2008) Advanced technologies for studies on protein interactomes. In *Protein–Protein Interaction.* Vol. 110, Springer, Berlin/Heidelberg, pp. 1–24.
Wuchty,S. (2001) Scale-free behavior in protein domain networks. *Mol. Biol. Evol.*, **18**, 1694–1702.
Wuchty,S. and Almaas,E. (2005) Evolutionary cores of domain co-occurrence networks. *BMC Evol. Biol.*, **5**, 24.
Yang,M. *et al.* (2010) Coevolution study of mitochondria respiratory chain proteins: toward the understanding of protein-protein interaction. *J. Genet. Genomics*, **38**, 201–207.