Predicting protein–protein interactions based only on sequences information

Juwen Shen[†], Jian Zhang[†], Xiaomin Luo[†], Weiliang Zhu[†‡], Kunqian Yu[†], Kaixian Chen[†], Yixue Li[§], and Hualiang Jiang[†‡¶]

[†]Center for Drug Discovery and Design, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, and Graduate School of Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China; [‡]School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China; and [§]Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved December 28, 2006 (received for review September 8, 2006)

Protein–protein interactions (PPIs) are central to most biological processes. Although efforts have been devoted to the development of methodology for predicting PPIs and protein interaction networks, the application of most existing methods is limited because they need information about protein homology or the interaction marks of the protein partners. In the present work, we propose a method for PPI prediction using only the information of protein sequences. This method was developed based on a learning algorithm-support vector machine combined with a kernel function and a conjoint triad feature for describing amino acids. More than 16,000 diverse PPI pairs were used to construct the universal model. The prediction ability of our approach is better than that of other sequence-based PPI prediction methods because it is able to predict PPI networks. Different types of PPI networks have been effectively mapped with our method, suggesting that, even with only sequence information, this method could be applied to the exploration of networks for any newly discovered protein with unknown biological relativity. In addition, such supplementary experimental information can enhance the prediction ability of the method.

conjoint triad | support vector machine

The molecular bases of cellular operations are sustained largely by different types of interactions among proteins. Thus, a major goal of functional genomics is to determine protein interaction networks for whole organisms (1). However, only recently has it become possible to combine the traditional study of proteins as isolated entities with the analysis of large protein interaction networks by using microarray and proteomic approaches (2, 3). Such kinds of studies are significantly important because many of the functions of complex systems seem to be more closely determined by their interactions rather than by the characteristics of their individual components (4). For example, metabolic pathways, signaling cascades, and transcription control processes involve complicated interaction networks (5). Recently, interaction networks have begun to be appreciated because it is necessary to address the general principles of biological systems by means of systems biology (6). Moreover, the study of protein interaction networks has been driven by potentially practical applications in drug discovery, because it might provide great insights into mechanisms of human diseases. This study may revolutionize the pipeline of drug discovery, because drugs discovered based on the protein interaction network may specifically modulate the disease-related pathway rather than simply inhibit or activate the functions of an individual target protein (7, 8). Determining accurate cellular protein interaction networks with experimental methods in combination with computational approaches therefore has become a major theme of functional genomics and proteomics efforts (9).

An impressive set of experimental techniques has been developed for the systematic analysis of protein–protein interactions (PPIs), including yeast two-hybrid-based methods (10), mass spectrometry (11), and protein chips (12) and hybrid approaches (13). Several binding reaction-detected methods, based on the presumption that the binding of one protein to another provokes a variety of biophysical changes, have been developed (14). These technologies recently identified hundreds of potentially interacting proteins and complexes in several species such as yeast (15), *Drosophila* (16), and *Helicobacter pylori* (17). Ulrich *et al.* (18) presented a large-scale two-hybrid map of >3,000 putative human PPIs. These data will serve as an important source of information regarding individual protein partners and offer preliminary insight into the global molecular organization of human cells.

However, current PPI pairs obtained with experimental methods cover only a fraction of the complete PPI networks (19). Therefore, computational methods for the prediction of PPIs have an important role (20). A number of computational methods have been developed for the prediction of PPIs. Computational methods based on genomic information, such as phylogenetic profiles, predict PPIs by accounting for the pattern of the presence or absence of a given gene in a set of genomes (21, 22). The main limitation of these approaches is that they can be applied only to completely sequenced genomes, which is the precondition to rule out the absence of a given gene. Similarly, they cannot be used with the essential proteins that are common to most organisms (23). The prediction of functional relationships between two proteins according to their corresponding adjacency of genes is another popular approach. This method is directly applicable only to bacteria, in which the genome order is relatively more relevant (24). Park *et al.* (25) tried to find protein interaction partners by viewing interactions between protein domains in terms of the interactions between structural families of evolutionarily related domains. Sprinzak and Margalit (26) put forward another indirect interaction prediction method, digging out signature feature related to interactions rather than domain interaction information from the protein sequences via protein classification. However, these methods are not universal, because the accuracy and reliability of these methods depend on the information of protein homology or interaction marks of the protein partners.

It is virtually axiomatic that "sequence specifies structure," which gives rise to an assumption that knowledge of the amino acid sequence alone might be sufficient to estimate the interacting propensity between two proteins for a specific biological function (27). Accordingly, prediction of PPIs based only on

Author contributions: J.S. and J.Z. contributed equally to this work; H.J. designed research; J.S., J.Z., and X.L. performed research; J.S., J.Z., X.L., W.Z., K.Y., K.C., Y.L., and H.J. analyzed data; and J.S., J.Z., W.Z., K.C., Y.L., and H.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviations: PPI, protein–protein interaction; SVM, support vector machine.

¶To whom correspondence should be addressed. E-mail: hljiang@mail.shcnc.ac.cn.

This article contains supporting information online at www.pnas.org/cgi/content/full/0607879104/DC1.

sequence information is an ideal approach for both the computational and experimental senses. The advantage of such a method is that it is much more universal. However, it is a major challenge in computational biology, and only a few groups have engaged in the development of methodology for such a prediction approach. Joel and David (28) endeavored to solve this problem by using a machine learning method with several physiochemical descriptors. Loris (29) developed a fusion technique of classifiers to predict PPIs. Nevertheless, these methods are not robust and reliable because they have not adequately considered the local environments of the residues in the sequences. Moreover, the prediction models were constructed based on limited PPI pairs (<3,000 pairs) but with hundreds of variants. Therefore, on one hand, they are apt to encounter the problem of overfitting and results that are data-dependent; on the other hand, these methods have not been used to predict PPI networks among a great many proteins.

In the present work, a machine learning method based on a support vector machine (SVM) combined with a kernel function and a conjoint triad feature abstract was developed for the prediction of PPIs based only on the primary sequences of proteins. To reduce the problem of overfitting, >16,000 PPI pairs were used to generate the prediction models. The prediction results of our method are more robust than those of recently published sequence-based PPI prediction models (28, 29). Notably, different levels of networks of PPIs have been effectively reproduced with this approach, indicating that, even with only sequence information, this sequence-based approach could be applied to explore the networks for newly discovered proteins with unknown biological functions.

Results

Our method for PPI prediction was developed based on an SVM. The detailed procedure of an SVM can be found in refs. 30–34. In our approach, each protein sequence is represented by a vector space consisting of features of amino acids [supporting information (SI) Fig. 3], and the PPI pair is characterized by concatenating the two vector spaces of two individual proteins. To reduce the dimensions of vector space and suit synonymous mutation, the 20 amino acids were clustered into several classes according to their dipoles and volumes of the side chains. The conjoint triad method abstracts the features of protein pairs based on the classification of amino acids. A kernel function that is especially designed propitious to the symmetrical property of PPI has been adopted for binary classification on a large data set.

Classification of Amino Acids. Electrostatic (including hydrogen bonding) and hydrophobic interactions dominate PPIs. These two kinds of interactions may be reflected by the dipoles and volumes of the side chains of amino acids, respectively. Accordingly, these two parameters were calculated, respectively, by using the density-functional theory method B3LYP/6–31G* and molecular modeling approach. The result is listed in SI Table 2. Based on the dipoles and volumes of the side chains, the 20 amino acids could be clustered into seven classes. Amino acids within the same class likely involve synonymous mutations because of their similar characteristics.

Conjoint Triad Method. For predicting PPI by sequences, one of the main computational challenges is to find a suitable way to fully describe the important information of PPI. To solve this problem, we proposed a descriptor named conjoint triad, which considered the properties of one amino acid and its vicinal amino acids and regarded any three continuous amino acids as a unit. Thus, the triads can be differentiated according to the classes of amino acids, i.e., triads composed by three amino acids belonging to the same classes, such as ART and VKS, could be treated identically, because they may be considered to play similar roles while processing PPI. The PPI information of protein sequences can be projected into a homogeneous vector space by counting the frequencies of each triad type. The process of generating descriptor vectors is described as follows.

First, we use a binary space (**V**, **F**) to represent a protein sequence. Here, **V** is the vector space of the sequence features, and each feature ($\mathbf{v}_i$) represents a sort of triad type; **F** is the frequency vector corresponding to **V**, and the value of the ith dimension of **F** ($\mathbf{f}_i$) is the frequency of type $\mathbf{v}_i$ appearing in the protein sequence. For the amino acids that have been catalogued into seven classes, the size of **V** should be $7 \times 7 \times 7$; thus, $i = 1, 2, \ldots 343$. The detailed definition and description for (**V**, **F**) are illustrated in SI Fig. 3. Clearly, each protein has a corresponding **F** vector. However, the value of $\mathbf{f}_i$ correlates to the length (number of amino acids) of protein. In general, a long protein would have a large value of $\mathbf{f}_i$, which complicates the comparison between two heterogeneous proteins. To solve this problem, we defined a new parameter, $\mathbf{d}_i$, by normalizing $\mathbf{f}_i$ with Eq. **1**:

$$\mathbf{d_i} = (\mathbf{f_i} - \min\{\mathbf{f}_1, \mathbf{f}_2, \ldots\ldots, \mathbf{f}_{343}\})/\max\{\mathbf{f}_1, \mathbf{f}_2, \ldots\ldots, \mathbf{f}_{343}\}. \quad [1]$$

The numerical value of $\mathbf{d}_i$ of each protein ranges from 0 to 1, which thereby enables the comparison between proteins. Accordingly, we obtain another vector space (designated **D**) consisting of $\mathbf{d}_i$ to represent protein. Next, we concatenate the vector spaces of two proteins ($\mathbf{D_A}$ and $\mathbf{D_B}$) to represent their interaction features ($\mathbf{D_{AB}}$) (Eq. **2**):

$$\{\mathbf{D_{AB}}\} = \{\mathbf{D_A}\} \oplus \{\mathbf{D_B}\}. \quad [2]$$

Thus, a 686-dimensional vector [343 (for one protein) plus 343 (for another protein)] has been built to represent each protein pair.

Kernel Function. The kernel function $\mathbf{K}(\cdot,\cdot)$ dominates the learning capability of the SVM. Considering the fact that PPI is symmetrical, i.e., $\{\mathbf{D_{AB}}\}$ and $\{\mathbf{D_{BA}}\}$ represent the same interaction pairs between proteins A and B, we designed a kernel function, $\mathbf{K}(\mathbf{D_{AB}}, \mathbf{D_{EF}})$ (Eq. **3**), which is denoted as a S-kernel function in the following:

$$\mathbf{K}(\mathbf{D_{AB}}, \mathbf{D_{EF}}) = \exp(-\gamma\|\mathbf{s}\|^2)s = \min\{(\|\mathbf{D_A} - \mathbf{D_E}\|^2 + \|\mathbf{D_B} - \mathbf{D_F}\|^2), (\|\mathbf{D_A} - \mathbf{D_F}\|^2 + \|\mathbf{D_B} - \mathbf{D_E}\|^2)\}. \quad [3]$$

SVM Parameter Optimization. As in other multivariate statistical models, the performances of the SVM for classification depend on the combination of several parameters. In general, the SVM involves two classes of parameters: the capacity parameter C and kernel type K. C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. The kernel type K is another important parameter. In the S-kernel function used in this study (Eq. **3**), γ is an important parameter to dominate the generalization ability of SVM by regulating the amplitude of the kernel function. Accordingly, two parameters, C and γ, should be optimized. The parameter optimization was performed by using a grid search approach within a limited range. To minimize the overfitting of the prediction model, 3-fold crossover validation was used to investigate the training set. Predict accuracy defined by Eq. **4** that is associated with mean-square-error was used to select the parameters:

$$\text{Predict accuracy} = 1 - \text{MSE}/(1 - (-1))^2. \quad [4]$$

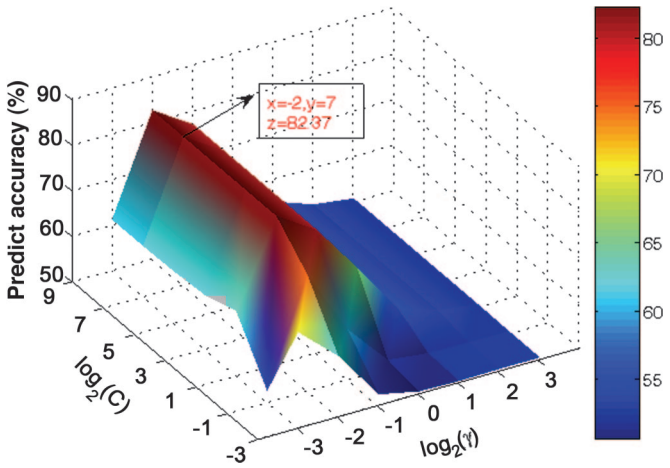

Fig. 1. Accuracy surface of threefold crossover validation on training set versus the variations of parameters C and γ.

During SVM classification, each data point represents a protein pair ($\mathbf{D}_{AB}$, y); if the protein pair is experimentally interactive, y is assigned 1, otherwise y is -1. Fig. 1 shows the profile of predicting the accuracy of the threefold crossover validation on the training set versus the variations of parameters C and γ. Obviously, the prediction accuracy profile has a maximum peak at (C, γ) = (128, 0.25), indicating that the optimal values of C and γ for constructing SVM models are 128 and 0.25, respectively.

Prediction Ability. Using the optimal values of C and γ, the PPI prediction model was constructed based on the training set by using the SVM learning algorithm with the S-kernel function. To minimize data dependence on the prediction model, five training sets and five test sets were prepared by the sampling method described in *Materials and Methods*. Each training set consisted of 32,486 protein pairs; half of the protein pairs were randomly selected from the data of positive PPI pairs, and the other half were randomly selected from the negative protein pairs. Each test set was constructed with another 400 protein pairs. Thus, five prediction models were generated for the five sets of data. The prediction results are listed in Table 1. For all five models, the precisions are >82.23%, the sensitivities are >84.00%, and the prediction accuracies are >82.75%. On average, our method may produce a PPI prediction model with an accuracy of 83.90 ± 1.29%. To test the reliability of our S-kernel function, we also constructed PPI prediction models by using four other kernel functions, namely, radial basis, polynomial, sigmoid, and linear functions, on the same PPI data sets. The prediction accuracies of those four kernel functions were 80.5%, 72.9%, 50.0%, and 62%, respectively (SI Table 3). The results indicate that the prediction model constructed with our kernel function is more accurate than the models constructed with the other kernel functions. Considering the potential errors of gathering a large amount of PPI data from different sources and experimental errors, it can be concluded that the prediction ability of our method may compare with that of experiments like the yeast two-hybrid method. Moreover, the prediction models constructed by our method may be extended to predict PPIs encoded in a pairwise PPI network, as we demonstrate below.

Table 1. Prediction results of the test sets

Test set	Accuracy, %	Precision, %	Sensitivity, %	Mean square error
1	84.25	84.42	84.50	0.63
2	82.75	83.59	84.00	0.69
3	83.25	84.82	85.50	0.67
4	83.25	82.23	84.00	0.67
5	86.00	86.00	86.00	0.56
Sum*	83.90 ± 1.29	84.21 ± 1.41	84.80 ± 0.91	0.64 ± 0.05

Precision is the true positive/(true positive + false positive). Sensitivity is the true positive/(true positive + false negative).
*Mean and variance are averaged by the results of five test sets.

Network Prediction. The most useful application of a PPI prediction method is its capability of predicting PPI networks. To our knowledge, PPI prediction methods based on protein sequences have not been reported in the application of network prediction. So, we extended our method to predict PPI networks assembled by pairwise PPIs. Three kinds of PPI networks have been predicted by our method: the one-core network, which is constructed by a core protein interacting with numerous other proteins (Fig. 2*A*); the multiple-core network consisting of an interacting pathway of several core proteins, which interact with other proteins (Fig. 2*B*); and the crossover network, which consists of several multiple-core networks and/or one-core networks with complicated interaction among these networks (Fig. 2*C*).

The one-core network is the simplest because one protein radially interacts with other proteins. CD9, an important tetraspanin protein, interacts with many associated factors, forming a typical one-core network (35). The prediction result revealed that 13 of the 16 PPI pairs could be addressed by our method (Fig. 2*A*), indicating that this method is capable of digging out partners of a protein encoded in a network composed of pairwise PPIs.

The Ras-Raf-Mek-Erk-Elk-Srf pathway is a currently accepted consensus network that has been implicated in a variety of cellular processes (36). So far, >100 cytoplasmic proteins have been reported to be involved in this pathway connected by means of a typical multicore network. Ras, Raf, Mek, Erk, Elk, and Srf serve as core proteins, which determine the signal transduction. Of the 189 PPI pairs in this network, 161 PPI pairs were predicted correctly by our method (Fig. 2*B*). The distribution of false frequency for each core protein is listed in SI Table 4. This result suggests that our method also can be used to predict PPIs encoded in a more complicated network.

Biologically, general PPI networks are crossover networks. If a computational method can predict such networks, it should be useful in practical applications. The Wnt-related pathway is essential in signal transduction, and the related network is a typical crossover network, of which the protein interaction topology has recently been demonstrated by Ulrich *et al.* (18) using yeast two-hybrid experiments (18). To illustrate the capability of our method in predicting such networks, we tried to reproduce the network of Wnt. The prediction result showed that 73 interactions among the 96 PPI pairs in the network were covered by our method (Fig. 2*C*), suggesting that our method can be applied in the prediction of general PPI networks.

In practice, primary experimental information may be used in PPI network prediction. To test whether the additional experimental clues can enhance the prediction ability of our method, we repredicted the network of Ras-Raf-Mek-Erk-Elk-Srf pathway by adding the existing interaction information of 30% interaction pairs. Indeed, the additional experimental information can increase the prediction ability; the accuracy increased from 84% to 90% (SI Fig. 4).

Discussion

We have presented a computational approach for PPI prediction. The SVM learning algorithm was used to develop meth-

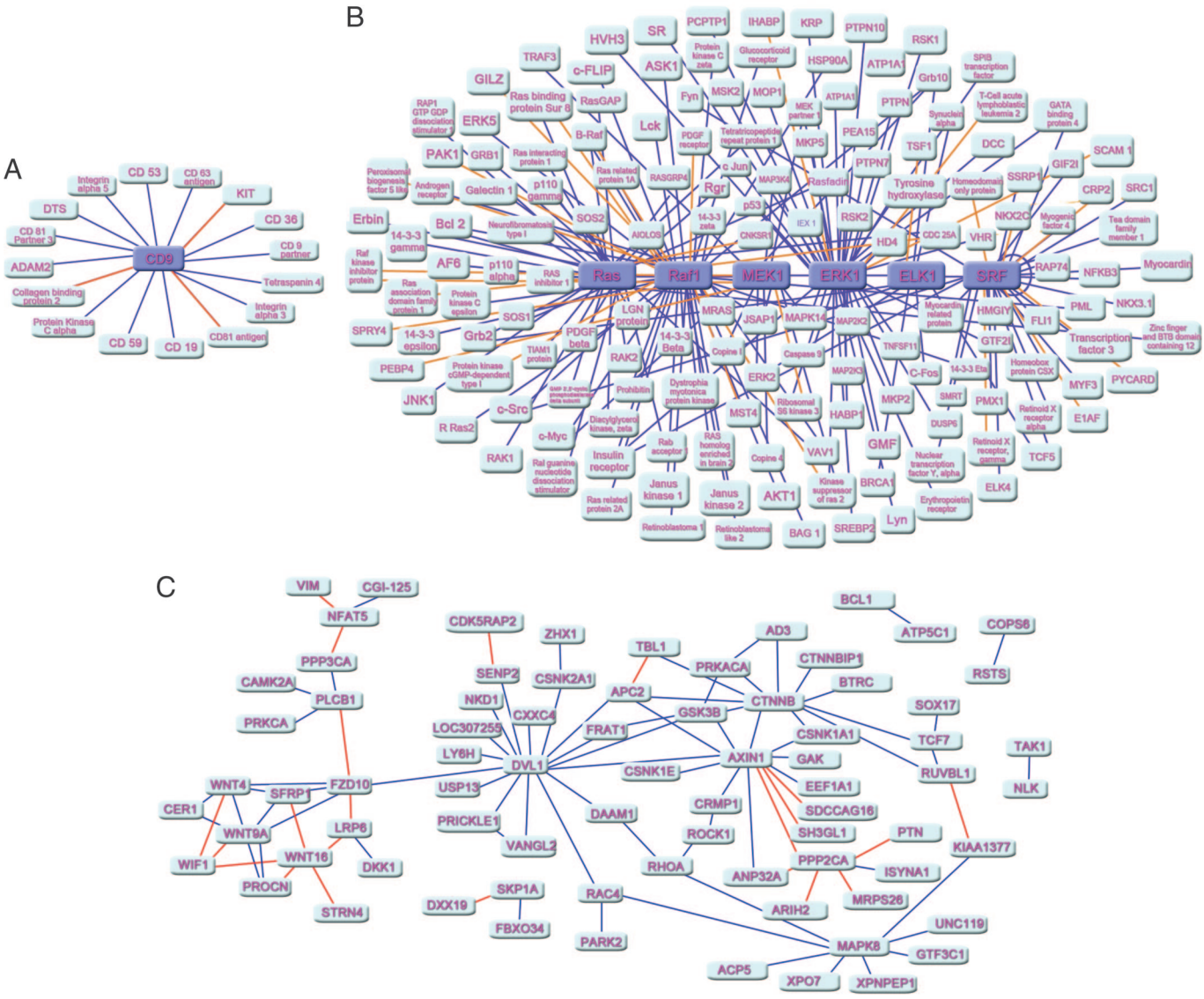

Fig. 2. Predicted results of PPI networks of a one-core network for CD9 (*A*), a multicore network for the Ras-Raf-Mek-Erk-Elk-Srf pathway (*B*), and a crossover network for the Wnt-related pathway (*C*). Core and satellite proteins are colored indigo blue and light blue, respectively. Lines connecting core proteins and satellite proteins are divided into two classes: dark blue, true prediction; orange, false prediction.

odology. Notably, we advanced a strategy to construct the vector space (**V**) for representing the PPI information. The protein sequence was characterized by a conjoint triad descriptor, which considered the properties of each amino acid and its vicinal amino acids. The PPI information encoded in the protein sequences was projected into the vector space by calculating the frequencies of the conjoint triads (SI Fig. 3). Because the 20 amino acids have been catalogued into seven classes according to their physicochemical properties (SI Table 2), the dimensions of vector space could be dramatically reduced, from $20 \times 20 \times 20$ to $7 \times 7 \times 7$. Accordingly, the reduced variance dimension may partially overcome the overfitting problem, by which a good prediction model can be constructed based on even a limited amount of data (e.g., currently available data for PPI pairs). In addition, the conjoint triad descriptor implicitly includes the information of synonymous mutation for PPI, which widens the prediction range.

Threefold crossover validation was used to increase the prediction accuracy, and the two important parameters for SVM learning, C and γ, were optimized as 128 and 0.25 (Fig. 1). The prediction ability of our method was first evaluated five times on five test sets. The result indicates that this method may produce a stable prediction model with higher accuracy than other similar methods (Table 1) (26, 27). Moreover, the data set for constructing prediction models is much larger than others, suggesting that our prediction models are more reliable. Considering the symmetrical nature of PPI, i.e., A–B interaction equals B–A interaction, we designed a kernel function-symmetrical kernel function (S-kernel function) for the SVM (Eq. **3**). The advantage of S-kernel function is obvious when comparing the prediction accuracy with other kernel functions (SI Table 3).

Accurately predicting PPI networks is the most important issue for PPI prediction methods. So far, PPI prediction methods based only on protein sequence information have not been applied to the prediction of PPI networks. We used such a method to predict

networks composed of pairwise PPIs. The results indicate that our method may reproduce most of the PPIs encoded in three typical networks (Fig. 2). Most importantly, this method can be used to predict the most complicate PPI network, the crossover network (Fig. 2*C*).

Computation Environment. All calculation programs implementing SVM were written in C++ based on the core of the libsvm 2.8 package (www.csie.ntu.edu.tw/~cjlin/libsvm) and run on a 128-CPU Origin 3800 server (Silicon Graphics, Mountain View, CA).

Materials and Methods

Data Set Preparation. PPI information was from the Human Protein References Database (HPRD), version 2005_0913 (www.hprd.org). This version of HPRD contains 16,443 nonredundant entries of experimentally verified PPIs obtained from a manual search of the literature. More than 95% of the interactions in the database are based on individual *in vivo* (e.g., coimmunoprecipitation) or *in vitro* (e.g., GST pull down) experiments (37). The data quality of this database is high enough for the construction of PPI prediction models. All of these PPI pairs were used in preparing the positive data set. The selection of a negative data set is essential to the reliability of the prediction model. However, it is difficult to generate such a data set because we have limited information about proteins that are really noninteractive. Unlike the random way for selecting a negative data set (38), we used a relatively rational strategy to select the negative data set, which was composed by the proteins appearing in the positive data set. Any protein pair appearing in the positive data set could be selected as a candidate for a negative pair in an exclusive way, for example, AB and IJ are positive interaction pairs, thus AI, AJ, BI, or BJ could be the negative pairs. Additionally, other requirements were considered: (*i*) the total number of negative pairs should equal that of the positive pairs (16,443 in this study) and (*ii*) the contribution of the proteins composing the negative set should be as harmonious as possible.

The training set consisted of 32,486 protein pairs, half from the positive data set and half from the negative date set. A test set was constructed with another 400 protein pairs. Both the positive and negative pairs were randomly selected.

Dipole and Volume Calculations. The structures of the 20 amino acids were extracted from the standard fragment library of Insight2005 (Accelrys, San Diego, CA). Dipoles of the side chains of the amino acids were calculated by using the density-functional theory method of B3LYP/6–31G* encoded in Gaussian03 (39), and the volumes of the side chains were calculated by using Sybyl6.8 (Tripos, St. Louis, MO).

Prediction for PPI Networks. The core protein in a one-core network was removed from the PPI data set, and the rest of the PPI proteins were used to build a prediction model with the optimal parameters. Afterward, PPIs between the core protein and satellite proteins were predicted by the prediction model. A similar method was used to construct prediction model for a multicore network by removing the core proteins from the PPI data set, and PPIs between the core proteins and satellite proteins were predicted by the model. For a crossover network, all proteins in the network were removed from the PPI data set during construction of the prediction model, and then PPIs in the network were predicted by the model.

The Shanghai Supercomputing Center and Computer Network Information Center of Chinese Academy of Sciences are acknowledged for allocation of computing time. This study was supported by Special Fund for the Major State Basic Research Project of China Grants 2002CB512802 and 2006CB0D1204 and Shanghai Science and Technology Commission Grant 03DZ19228.

1. Chen L, Wu LY, Wang Y, Zhang XS (2006) *Proteins* 62:833–837.
2. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) *Science* 308:523–529.
3. Plavec I, Sirenko O, Privat S, Wang Y, Dajee M, Melrose J, Nakao B, Hytopoulos E, Berg EL, Butcher EC (2004) *Proc Natl Acad Sci USA* 101:1223–1228.
4. Pawson T (2004) *Cell* 116:191–203.
5. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1989) *Molecular Biology of the Cell* (Garland, New York), 2nd Ed.
6. Ge H, Walhout AJ, Vidal M (2003) *Trends Genet* 19:551–560.
7. Ryan DP, Matthews JM (2005) *Curr Opin Struct Biol* 15:441–446.
8. John MP, Srdjan A, Robert RB, Cindy LC, Yew-Seng JH, Vladimir K, Shuping L, Tahmina M, Mike P, Paul BR, *et al.* (2004) *Int J Mass Spectrom* 238:119–130.
9. Vittoria C, Alessandro F, Amos M, Alessandro V (2005) *Phys A Stat Mech Appl* 352:1–27.
10. Fields S, Song OK (1989) *Nature* 340:245–246.
11. Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.* (2002) *Nature* 415:141–147.
12. Heng Z, Metin B, Rhonda B, David H, Antonic C, Paul B, Ning L, Ronald J, Scott B, Thomas H, *et al.* (2001) *Science* 293:2101–2105.
13. Tong AHY, Becky D, Giuliano N, Gary DB, Barbara B, Luisa C, Marie E, Silvia F, Bryce N, Serena P, *et al.* (2002) *Science* 295:321–324.
14. Lakey JH, Raggett EM (1998) *Curr Opin Struct Biol* 8:119–123.
15. Peter U, Loic G, Gerard C, Traci AM, Richard SJ, James RK, Daniel L, Vaibhav N, Maithreyan S, Pascale P, *et al.* (2000) *Nature* 403:623–631.
16. Sarah EB, Xin XT, Kathleen SM (2006) *Mol Cell Proteomics* 5:824–834.
17. Jean CR, Luc S, Hilde DR, Veronique B, Celine R, Stephane S, Gerlinde L, Fabien P, Jerome W, Vincent S, *et al.* (2001) *Nature* 409:211–215.
18. Ulrich S, Uwe W, Maciej L, Christian H, Felix HB, Heike G, Martin S, Martina Z, Anke S, Susanne K, *et al.* (2005) *Cell* 122:957–968.
19. Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M (2005) *Nat Biotechnol* 23:839–844.
20. Wodak SJ, Mendez R (2004) *Curr Opin Struct Biol* 14:242–249.
21. Thomas D, Berend S, Martijn H, Peer B (1998) *Trends Biochem Sci* 23:324–328.
22. Matteo P, Edward MM, Michael JT, David E, Todd OY (1999) *Biochemistry* 96:4285–4288.
23. Valencia A, Pazos F (2002) *Curr Opin Struct Biol* 12:368–373.
24. Wojcik J, Boneca IG, Legrain P (2002) *J Mol Biol* 323:763–770.
25. Park J, Lappe M, Teichmann SA (2001) *J Mol Biol* 307:929–938.
26. Sprinzak E, Margalit H (2001) *J Mol Biol* 311:681–692.
27. Christian BA (1973) *Science* 81:223–230.
28. Joel RB, David AG (2001) *Bioinformatics* 17:455–460.
29. Loris N (2005) *Neurocomputing* 68:289–296.
30. Vapnik V (2005) *The Nature of Statistical Learning Theory* (Springer, New York), pp 96–99.
31. Wang WJ, Xu ZB, Lu WZ, Zhang XY (2003) *Neurocomputing* 55:643–663.
32. Francis EET, Cao LJ (2002) *Neurocomputing* 48:847–861.
33. Burbidge R, Trotter M, Buxton B, Holden S (2001) *Comput Chem* 26:5–14.
34. Chris HQD, Inna D (2001) *Bioinformatics* 17:349–358.
35. Yang XH, Oleg VK, Tatiana VK, Milena MA, Eric R, Jack LS, Martin EH (2006) *J Biol Chem* 281:12976–12985.
36. Davis RJ (1995) *Mol Reprod* 42:459–467.
37. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, *et al.* (2006) *Nat Genet* 38:285–293.
38. Chen XW, Liu M (2005) *Bioinformatics* 21:4394–4400.
39. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, *et al.* (2003) GAUSSIAN 03 (Gaussian, Pittsburgh, PA), Revision C.02.