BIOINFORMATICS

Vol. 17 no. 5 2001
Pages 455–460

Predicting protein–protein interactions from primary structure

*Joel R. Bock and David A. Gough**

Department of Bioengineering, 9500 Gilman Drive, University of California, San Diego, La Jolla, CA 92093-0412, USA

Received on August 22, 2000; revised on November 22, 2000; accepted on January 4, 2001

ABSTRACT

Motivation: An ambitious goal of proteomics is to elucidate the structure, interactions and functions of all proteins within cells and organisms. The expectation is that this will provide a fuller appreciation of cellular processes and networks at the protein level, ultimately leading to a better understanding of disease mechanisms and suggesting new means for intervention. This paper addresses the question: can protein–protein interactions be predicted directly from primary structure and associated data? Using a diverse database of known protein interactions, a Support Vector Machine (SVM) learning system was trained to recognize and predict interactions based solely on primary structure and associated physicochemical properties.

Results: Inductive accuracy of the trained system, defined here as the percentage of correct protein interaction predictions for previously unseen test sets, averaged 80% for the ensemble of statistical experiments. Future proteomics studies may benefit from this research by proceeding directly from the automated identification of a cell's gene products to prediction of protein interaction pairs.

Contact: dgough@bioeng.ucsd.edu

INTRODUCTION

The interaction between proteins is fundamental to a broad spectrum of biological functions, including regulation of metabolic pathways, immunologic recognition, DNA replication, progression through the cell cycle, and protein synthesis (Alberts *et al.*, 1989). Whether or not two proteins will bind to form a stable complex that is prerequisite to biological function is dependent on the three-dimensional conformations of the proteins (Jones and Thornton, 1996). For a given conformation, the chemical reactivity of an individual protein is defined by the type and spatial orientation of surface-accessible amino acid side chains. Conformation therefore determines protein–ligand binding. In biology, it is virtually axiomatic that 'sequence specifies conformation' (Anfinsen, 1973), suggesting an intriguing postulate: knowledge of the amino acid sequence alone might be sufficient to estimate the propensity for two proteins to interact and effect useful biological function.

The science of proteomics endeavors to elucidate the structures, interactions and functions of all of a cell's or organism's proteins (Anonymous, 1999), with the objective of understanding cellular processes and networks and, ultimately, disease processes at the protein level (Blackstock and Weir, 1999). Current technology for cataloging the proteins contained within a cell involves separation via two-dimensional gel electrophoresis, followed by identification using tandem mass spectrometry (Dove, 1999). Experimental techniques such as two-hybrid screens (Fields and Song, 1989) are often employed to study dynamic interactions between the identified cellular proteins (Bartel and Fields, 1997; Uetz *et al.*, 2000). As such techniques are 'tedious, labor-intensive and potentially inaccurate' (Enright *et al.*, 1999), investigators have recently been prompted to seek computational methods to predict whether or not two proteins will interact. Previous research groups have presented predictive methodologies based on various principles, including correlated changes in amino acid sequence between interacting protein domains (Pazos *et al.*, 1997); using genomic context to infer functional protein interactions between the gene products (Huynen *et al.*, 2000); or inference from genome sequences, given observed homologies in other organisms, where interacting proteins have fused into a single protein chain (Marcotte *et al.*, 1999; Enright *et al.*, 1999).

Earlier prediction techniques were focused on estimating the *site* of interaction, without reference to specific binding partners. These methods utilized features and properties related to interface topology, solvent Accessible Surface Area (ASA) and hydrophobicity (Jones and Thornton, 1997), or the recognition of specific residue or geometric motifs (Kini and Evans, 1996; Nissinka *et al.*, 2000). Antigenic determinant sites in proteins were predicted using hydrophilicity profiling methods presented in Hopp and Woods (1981) and Welling *et al.* (1985).

*To whom correspondence should be addressed.

In contrast to the cited investigations, the methodology reported herein takes an entirely different approach to computational prediction of protein interactions. Given a database of known protein–protein interaction pairs, a machine learning system is trained to recognize interactions based *solely on primary structure and associated physicochemical properties*. Generalization of results obtained by the system upon introduction of unseen testing sequences is encouraging, given the volume of the dataset. Future proteomics studies may benefit from this research by proceeding directly from the automated identification of a cell's gene products to prediction of the protein interaction pairs.

The success of the new methodology is based on the automatic recognition of correlated patterns of sequence and substructure in the interacting pairs. These patterns typically comprise a small number of functional residues in each protein (Casari *et al.*, 1995).

Complete proteomic functional assignment requires the identification and quantitation of all contributors to dynamic multi-protein complexes. Many molecular signal transduction processes are regulated by the intermediary characteristics of discrete protein recognition 'domains', evolutionarily-conserved modules of amino acid sequence found in catalytic proteins, as well as on scaffold, anchoring or adaptor proteins (Pawson and Scott, 1997). Protein interactions are frequently mediated by these domains, each of which bind to specific peptides. Such interactions form the basis for structural and functional organization within cells (Pawson, 1995).

Protein domains are often observed across genomes of multiple species (Bateman *et al.*, 2000). While certain discrete enzymatic signaling domain families are common to all three divisions of cellular life, many non-enzymatic eukaryotic signaling domains with prokaryotic homologs have been identified (Ponting *et al.*, 1999). Important examples include the SH3 (50 a.a) and PDZ (90 a.a) domains (Fanning and Anderson, 1996; Pawson and Scott, 1997). Other domains organize into larger structural domains or families, subsequently facilitating the assembly and interaction of other proteins. For example: the tetratricopeptide repeat domain (TPR; 34 a.a.) forms a superhelical structure with an amphipathic groove for binding protein targets, and mediates protein–protein interactions (Das *et al.*, 1998). β-propeller superstructures are a common motif, comprising, e.g. NHL repeat domains (45 a.a) found on proteins involved in mediating activity of lentiviral Tat proteins *in vivo* (Fridell *et al.*, 1995), and WD40 repeat domains (40 a.a.) on G-proteins, important regulators of a host of cellular functions (Neer *et al.*, 1994).

Table 1. Organism representation by proteins found in the DIP database. Frequency expressed as fraction of total number of occurrences of each organism. The top 95% most frequent organisms are listed

Organism	Superkingdom	Frequency
S.cerevisiae	Eukaryota	0.639
H.sapiens	Eukaryota	0.184
Mus musculus	Eukaryota	0.049
D.melanogaster	Eukaryota	0.033
R.norvegicus	Eukaryota	0.020
E.coli	Bacteria	0.013
Bos taurus	Eukaryota	0.012

SYSTEM AND METHODS

The protein–protein interaction prediction method is described in this section.

Database of interacting proteins

Protein interaction data were obtained from the Database of Interacting Proteins (DIP; http://www.dip.doe-mbi.ucla.edu/). At the time of writing, the database comprises 2664 entries representing pairs of proteins known to mutually bind, giving rise to a specific biological function. Each interaction pair contains fields representing accession codes linking to other public protein databases, protein name identification and references to experimental literature underlying the interactions. Alternative fields include protein interaction domains, superfamily identification, interacting residue ranges, and protein–protein complex dissociation constants.

The representation of the various biological superkingdoms in the DIP database is heavily biased towards the Eukaryotes. In Table 1, the top 95% most-frequently occurring organisms and their kingdom membership are summarized. Note that the budding yeast *Saccharomyces cerevisiae* accounts for 64% of the interactions, which are readily accessible online (Uetz *et al.*, 2000). The bacterium *Escherichia coli* constitutes the most frequent non-eukaryote proteome, yet accounts for only 1.3% of the proteins found in the database.

On the molecular level, the protein interaction substructural domain coverage within DIP is diverse. Submitting the protein sequences to the Protein Families Database (Bateman *et al.*, 2000) of protein domains and profile hidden Markov models (Pfam v. 5.5; http://www.pfam.wustl.edu/), we estimated that at least 1394 distinct domains are represented. Table 2 lists the most frequent protein domains found in DIP, using a sequence E-value cutoff level of 1.0. A histogram portraying the distribution of all protein sequence lengths within the database is presented in Figure 1. The mean and standard deviation of amino acid chain lengths are 481 and 386 residues, respectively.

Table 2. Most frequent protein domains in the interaction dataset. Frequency expressed as fraction of total occurrences of each domain. Prediction using the Protein Families Database (Pfam v 5.5; Bateman *et al.*, 2000)

No.	Domain	Frequency
1	WD40	0.056
2	pkinase	0.030
3	TPR	0.028
4	zf-C2H2	0.018
5	Armadillo_seg	0.016
6	EGF	0.016
7	HLH	0.013
8	spectrin	0.013
9	bZIP	0.011
10	ank	0.011
11	rrm	0.009
12	SH2	0.008
13	SH3	0.008
14	Sm	0.007
15	ras	0.007
16	fn3	0.007
17	PHD	0.006
18	efhand	0.006
19	myb_DNA-binding	0.006
20	arf	0.006

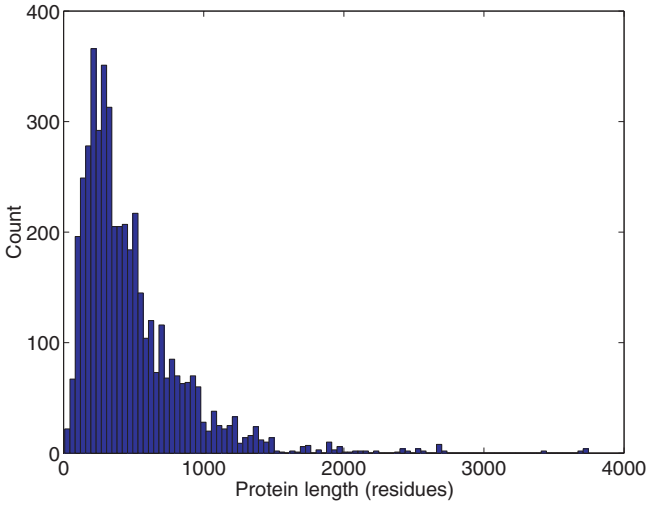

Fig. 1. Distribution of protein sequence lengths in database. At least 1394 distinct interacting domains are represented. $\mu = 481 \pm 386$ residues.

Support vector machine learning

The new protein–protein interaction estimator utilizes the technique of 'support vector' learning, an area of statistical learning theory subject to extensive recent research (Vapnik, 1995; Schölkopf *et al.*, 1999). A selection of recent bioinformatic investigations utilizing Support Vector Machine (SVM) learning includes Brown *et al.* (1999), Jaakkola *et al.* (2000) and Zien *et al.* (1999). Useful for function approximation, signal processing and regression, SVM has several advantages as applied in the present context:

(1) SVM generates a representation of the non-linear mapping from residue sequence to protein fold space (Baldi and Brunak, 1998) using relatively few adjustable model parameters.

(2) Based on the principle of *structural risk minimization*, SVM provides a principled means to estimate generalization performance via an *analytic* upper bound on the generalization error. This means that a confidence level may be assigned to the prediction, and alleviates problems with overfitting inherent in neural network function approximation (Hecht-Nielsen, 1989).

(3) Computationally efficient (Joachims, 1999), SVM is characterized by fast training which is essential for high-throughput screening of large protein datasets.

(4) SVM is readily adaptable to new data, allowing for continuous model updates in parallel with the continuing growth of biological databases.

In the present research, we train an SVM to recognize pairs of interacting proteins culled from the DIP database. The decision rules developed by the system are then used to generate a discrete, binary decision (1 = interaction, 0 = no interaction) upon introduction of a new feature set based on primary structure of the putative protein interaction pair.

Feature representation

For each amino acid sequence of a protein–protein complex, feature vectors were assembled from encoded representations of tabulated residue properties including charge, hydrophobicity, and surface tension for each residue in sequence. This set of features was motivated by the previous demonstration of sequential hydrophilicity profiles as sensitive descriptors of local interaction sites (Hopp and Woods, 1981). Here, this concept was extended to integrate sequential charge and surface tension, as water molecules influence atomic packing for shape complementarity, and mediate polar interactions at protein–protein recognition sites (Conte *et al.*, 1999). Our postulate is that since sequentially-proximal protein secondary structure elements are often co-located in three-dimensional conformation (Levitt and Chotia, 1976), the sequential profile of these additional features (charge, surface tension) must similarly 'co-locate' upon folding.

Let the vector of numbers $\{\mathbf{v}_j\}^i,\ i \in 1, \ldots, M$ in L-dimensional real space $\mathbb{R}^L$ denote feature i for a given amino acid sequence of length L residues, where M different features are considered. Lengths of the individual

feature vectors **v** were normalized by mapping onto a fixed-length interval K, via $\{\mathbf{y}_k\}^i = f\ (\{\mathbf{v}_j\}^i)$, where the function f is defined by $f : \mathbb{R}^L \rightarrow \mathbb{R}^K$. In this transformed space, the arc length coordinate ξ along the peptide sequence now varies as $\xi \in [0, 1]$. This is an essential step for representing proteins of widely varying native length (Figure 1). The full feature vector for a particular protein A is constructed by concatenation of each feature sequence **y**. This is written as $\{\varphi_A^+\} = \{\mathbf{y}_k\}^1 \oplus \{\mathbf{y}_k\}^2 \oplus \cdots \oplus \{\mathbf{y}_k\}^M$, where $\mathbf{a} \oplus \mathbf{b}$ indicates simple concatenation of vectors **a** and **b**. Finally, a representation of an interaction pair, $\{\varphi_{AB}^+\}$ is formed by concatenating the feature vectors for proteins A and B, i.e. $\{\varphi_{AB}^+\} = \{\varphi_A^+\} \oplus \{\varphi_B^+\}$. The vector $\{\varphi_{AB}^+\}$ becomes a positive training example for the SVM.

Negative examples (putative non-interacting protein pairs) must also be presented to the SVM. In this context, it may be insufficient to merely randomize the residues, a practice commonly carried out to estimate the statistical significance of biological sequence alignments as contrasted against a random control (Needleman and Wunsch, 1970; Fitch, 1983). Since a database of non-interacting proteins was not readily available, we chose to create negative controls by randomizing amino acids sequences sampled from DIP, while preserving both (1) amino acid composition and (2) di- and tri-peptide 'k-let' frequencies (Coward, 1999; Kandel *et al.*, 1996). Presumably, where $k > 1$, this procedure provides more native-like artificial proteins by conserving higher-order biases. Without performing exhaustive wet experiments to prove the biological inertness of proteins encoded by negative exemplars $\{\varphi_{CD}^-\}$, thereby proving that in fact proteins C and D do *not* interact, this must suffice to design and implement the numerical experiments. Randomized amino acid sequences were generated using `Shufflet` (http://www.genetique.uvsq.fr/eivind/shufflet.html) (Coward, 1999).

Data partitioning

In the experiments reported here, the DIP database entries were sampled at random, and data were partitioned into training and testing sets, at approximately a $1 : 1$ ratio. Feature vectors constructed as described in the previous section were used as examples for training and testing the prediction system. Testing examples were not exposed to the system during SVM learning. The database is robust in the sense that it represents a compendium of protein interaction data collected from diverse experiments. As noted above, 1394 different protein domains are represented. There is a negligible probability that the learning system will 'learn its own input' (see Baldi *et al.*, 2000) on a narrow, highly self-similar set of data examples. This enhances the generalization potential of the trained SVM.

IMPLEMENTATION

Software methods for parsing the DIP database, control of randomization and sampling of records and sequences, and feature vector creation were developed in Java. A new database was constructed by augmenting the original DIP records. Additional fields added included amino acid sequence data and associated residue features, generated as described in the section Feature representation.

SVM learning was implemented using SVM^{light} (Joachims, 1999), available at the following URL: http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/svm_light.eng.html.

Training and testing exemplar data files were developed using a prescribed k-let frequency ($k \in [1, 2, 3]$) and ensemble sampling size as input parameters to the data preparation software. Each member of the statistical ensemble involved a random sampling of the DIP interacting proteins and newly-created 'shuffled' amino acid sequences. A different SVM was trained for each k-let correlation frequency and experimental trial. The results of these trials were averaged to eliminate potential biases due to chance sampling of the dataset.

The performance of each SVM was evaluated using the inductive accuracy on the previously unseen test examples as the performance metric. 'Inductive accuracy' is defined here as the percentage of correct protein interaction predictions on the test set, consisting of nearly equal numbers of positive and negative interaction examples.

The main results of the protein–protein interaction predictions are summarized in Table 3. Each row in the table corresponds to a constant k-let frequency used to generate the negative training and testing examples. Data in the column headed '# Examples' indicate the average total number of each type of examples for each case. These data have been averaged over an ensemble of $N = 10$ trials, a sufficient sample as indicated by the low variance shown in Column 3.

DISCUSSION AND CONCLUSION

The inductive accuracy of the learning machines as summarized in Table 3 is encouraging, given the depth of the DIP database. For each statistical background comprising k-let orders 1–3, about four out of five potential protein interactions are correctly estimated by the system. It bears reiteration here that only primary structure data have been used to train the SVM. We submit that some implicit information regarding structural, chemical and biological affinity has been represented and learned by virtue of the affirmative labeling of protein interaction pairs. The implications of the results shown in Table 3 for future proteomics research are intriguing.

While the methodology presented here is generally applicable, the proteins in the interaction database

Table 3. System generalization accuracy summary. 'Inductive accuracy' is the percentage of correct protein interaction predictions on test data not previously seen by the system. $N = 10$ trials

k-let frequency	# Examples (Train, Test)	Inductive accuracy (%)
1	(2190, 2189)	80.96 ± 1.42
2	(2192, 2192)	80.19 ± 0.86
3	(2203, 2195)	80.13 ± 0.89

predominantly represent eukaryotes, as summarized in Table 1. This bias may also be manifested in the trained SVM, which may not immediately generalize to bacteria or archea, although prokaryotic homologues of many non-enzymatic eukaryotic signaling domains associated with protein–protein interactions have been identified (Ponting *et al.*, 1999). To identify conserved interactions across species, additional training based on more kingdom-diverse proteomes may be required.

With reference to the first row of Table 3, we observe that good predictive accuracy is achieved when amino acid composition alone is preserved during randomization ($k = 1$). System performance is not degraded relative to cases $k = 2, 3$. If the results indicated a predictive performance deficit where $k = 1$, one might have conjectured that the SVM had merely learned to discriminate native interactions from random, non-native proteins here. It is unclear whether this observation is an artifact of the particular bias toward *S.cerevisiae* in the database. This question should be addressed in future research.

In conclusion, the prediction methodology reported in this paper generates a binary decision about potential protein–protein interactions, based only on primary structure and associated physicochemical properties. This suggests the possibility of proceeding directly from the automated identification of a cell's gene products to inference of the protein interaction pairs, facilitating protein function and cellular signaling pathway identification.

This research represents only an initial step in the automated prediction of protein interactions. The discovery of patterns within respective primary structures of known protein interaction pairs may be subsequently enhanced by using other features (secondary and tertiary structure, binding affinities, etc.) in the learning machine.

With experimental validation, further development along these lines may produce a robust computational screening technique that narrows the range of putative candidate proteins to those exceeding a prescribed threshold probability of interaction.

ACKNOWLEDGEMENTS

The authors thank Eivind Coward of DKFZ Heidelberg for making the Shufflet sequence-randomizing code available. Constructive input from the anonymous referees was extremely helpful in refining the original manuscript. JRB would like to acknowledge stimulating discussions with Akhilesh Maewal during this research.

REFERENCES

Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1989) *Molecular Biology of the Cell*. 2nd edn, Garland, New York.

Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Anonymous (1999) The promise of proteomics. *Nature*, **402**, 703.

Baldi,P. and Brunak,S. (1998) Bioinformatics: the machine learning approach. In *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.

Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bartel,P.L. and Fields,S. (eds) (1997) The yeast two-hybrid system. In *Advances in Molecular Biology*. Oxford University Press, New York.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Blackstock,W.P. and Weir,M.P. (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.*, **17**, 121–127.

Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C., Furey,T.S., Ares,M. and Haussler,D. (1999) Support vector machine classification of microarray gene expression data. Technical Report UCSC-CRL-99-09, University of California, Santa Cruz, Santa Cruz, CA, June 1999.

Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

Coward,E. (1999) Shufflet: shuffling sequences while conserving the k-let counts. *Bioinformatics*, **15**, 1058–1059.

Das,A.K., Cohen,P.W. and Barford,D. (1998) The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein–protein interactions. *EMBO J.*, **17**, 1192–1199.

Dove,A. (1999) Proteomics: translating genes into products? *Nat. Biotechnol.*, **17**, 233–236.

Enright,A.J., Ililopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Fanning,A.S. and Anderson,J.M. (1996) Protein–protein interactions: PDZ domain networks. *Curr. Biol.*, **6**, 1385–1388.

Fields,S. and Song,O.-K. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.

Fitch,W.M. (1983) Random sequences. *J. Mol. Biol.*, **163**, 171–176.

Fridell,R.A., Harding,L.S., Bogerd,H.P. and Cullen,B.R. (1995) Identification of a novel human zinc finger protein that specif-

ically interacts with the activation domain of lentiviral tat proteins. *Virology*, **209**, 347–357.

Hecht-Nielsen,R. (1989) *Neurocomputing*. Addison-Wesley, Reading, MA.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.

Huynen,M., Snel,B., Lathe,W. and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.

Joachims,T. (1999) Making large-scale support vector machine learning practical. In *Advances in Kernel Methods—Support Vector Learning,* chapter 11, MIT Press, Cambridge, MA, pp. 169–184.

Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Jones,S. and Thornton,J.M. (1997) Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.

Kandel,D., Mathias,Y., Unger,R. and Winkler,P. (1996) Shuffling biological sequences. *Discrete Appl. Math.*, **71**, 171–185.

Kini,R.M. and Evans,J.H. (1996) Prediction of potential protein–protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site. *FEBS Lett.*, **385**, 81–86.

Levitt,M. and Chotia,C. (1976) Structural patterns in globular proteins. *Nature*, **261**, 552–558.

Marcotte,E., Pellegrini,M., Ng,H.-L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Neer,E.J., Schmidt,C.J., Nambudripad,R. and Smith,T.F. (1994) The ancient regulatory-protein family of wd-repeat proteins. *Nature*, **371**, 297–300.

Nissinka,J.W., Verdonk,M.L. and Klebe,G. (2000) Knowledge-based descriptors to predict protein–ligand interactions. *Proceedings of the 13th European Symposium on Quantitative Structure-Activity Relationships.* Heinrich-Heine Universität, Düsseldorf, Germany.

Pawson,T. (1995) Protein modules and signalling networks. *Nature*, **373**, 573–580.

Pawson,T. and Scott,J.D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science*, **278**, 2075–2080.

Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.*, **1**, 511–523.

Ponting,C.P., Aravind,L., Schultz,J., Bork,P. and Koonin,E.V. (1999) Eukaryotic signalling domain homologues in Archaea and Bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.*, **289**, 729–745.

Schölkopf,B., Burges,C.J. and Smola,A.J. (eds) (1999) *Advances in Kernel Methods*. MIT Press, Cambridge, MA.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P., Qureshi-Emili,A., Li,Y., Godwin,B., Conover,D., Kalbfleisch,T., Vijayadamodar,G., Yang,M., Johnston,M., Fields,S. and Rothberg,J.M. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York, NY.

Welling,G.W., Weijer,W.J., van der Zee,R. and Welling-Wester,S. (1985) Prediction of sequential antigenic regions in proteins. *FEBS Lett.*, **188**, 215–218.

Zien,A., Rätsch,G., Mika,S., Schölkopf,B., Lemmen,C., Smola,A., Lengauer,T. and Müller,K.R. (1999) Engineering support vector machine kernels that recognize translation initiation sites. *Proceedings of the German Conference on Bioinformatics '99.* Hannover, Germany, pp. 37–43. GBF.