*doi:10.1093/nar/gkm254*

# CytoSVM: an advanced server for identification of cytokine-receptor interactions

**Jin-Rui Xu[1], Jing-Xian Zhang[1], Bu-Cong Han[1], Liang Liang[1] and Zhi-Liang Ji[1,2,*]**

[1]Key Laboratory for Cell Biology & Tumor Cell Engineering, the Ministry of Education of China, School of Life Sciences and [2]The Key Laboratory for Chemical Biology of Fujian Province, Xiamen University, Xiamen 361005, FuJian Province, P R China

Received January 23, 2007; Revised March 24, 2007; Accepted April 8, 2007

## ABSTRACT

**The interactions between cytokines and their complementary receptors are the gateways to properly understand a large variety of cytokine-specific cellular activities such as immunological responses and cell differentiation. To discover novel cytokine-receptor interactions, an advanced support vector machines (SVMs) model, CytoSVM, was constructed in this study. This model was iteratively trained using 449 mammal (except rat) cytokine-receptor interactions and about 1 million virtually generated positive and negative vectors in an enriched way. Final independent evaluation by rat's data received sensitivity of 97.4%, specificity of 99.2% and the Matthews correlation coefficient (MCC) of 0.89. This performance is better than normal SVM-based models. Upon this well-optimized model, a web-based server was created to accept primary protein sequence and present its probabilities to interact with one or several cytokines. Moreover, this model was applied to identify putative cytokine-receptor pairs in the whole genomes of human and mouse. Excluding currently known cytokine-receptor interactions, total 1609 novel cytokine-receptor pairs were discovered from human genome with probability ~80% after further transmembrane analysis. These cover 220 novel receptors (excluding their isoforms) for 126 human cytokines. The screening results have been deposited in a database. Both the server and the database can be freely accessed at http://bioinf.xmu.edu.cn/software/cytosvm/cytosvm.php.**

## INTRODUCTION

The binding of cytokines to their receptors on cell membranes triggers the cellular activities such as immunological regulation, cell growth, differentiation, apoptosis and migration in vertebrates (1). Therefore, characterization of novel cytokine-receptor pairs becomes the shortcut to understand these cytokine-mediated signal pathways.

The traditional isolation and characterization methods for identification of cytokine-receptor pairs are significantly limited by their characteristics of short half life, low plasma concentrations, pleiotropy and redundancy. It has been improved by the applications of modern molecular technologies such as cloning technology. Furthermore, as a complementary solution to experimental approaches, searches for new members of cytokines or their receptors are now often conducted by identifying genes highly homologous to known cytokine/receptor genes. Currently, 203 human cytokine-receptor pairs have been characterized as presented in KEGG pathway database (2). Unfortunately, it has become more and more difficult to discover new partners of cytokine and receptor if no new sequence features were identified. Especially for those peptides without significant sequence similarity to known cytokines/receptors, their functions are difficult to be probed on the basis of homologous or clustering methods.

Various alternative methods for describing protein interactions have been developed in recent years. These include evolutionary analysis (3,4), Hidden Markov Models (5), structural consideration (6–8), protein/gene fusion (9,10), motifs recognition (11), family classification by sequence clustering (12) and functional family prediction by statistical learning methods (13,14). Support vector machines (SVMs) is a two-class classifier, which has been previously used in the classification of cytokine families (http://www.bioinfo.tsinghua.edu.cn/%7Ehn/CTKPred/index.html) (14). In this study, we constructed an improved SVM model, CytoSVM, for the identification of cytokine-receptor interactions on the basis of protein

*To whom correspondence should be addressed. Tel: 86-0592-2182897; Fax: 86-0592-2181015;
Email: appo@bioinf.xmu.edu.cn; zhiliang.ji@gmail.com;

The authors wish it to be known that, in their opinion, the second and third authors contributed equally to this work primary sequences. This model was further applied to screen the whole genomes of human and mouse for novel cytokine-receptor pairs.

## CONSTRUCTION OF CytoSVM MODEL

CytoSVM is a model based on the statistical learning algorithm, SVM. This algorithm has been well-studied and implemented to solve a variety of protein classification problems including protein functional class (13,15), fold recognition (16), analysis of solvent accessibility (17), prediction of secondary structures (18) and protein–protein interactions (19). As a method that uses sequence-derived physicochemical properties of proteins as the basis for classification, SVM may be particularly useful for functional classification of distantly related proteins and homologous proteins of different functions (13). Such a feature makes SVM a potentially attractive method for probing the novel cytokine receptors, especially when the diversity of cytokine receptors in sequence cannot be properly handled by sequence homology-based approaches.

### The data sets

*The positive data pool.* The positive data (the true cytokine-receptor interactions) were collected from the KEGG pathway database (2) and the literatures. These interaction pairs cover 449 distinct known cytokine-receptor interactions in mammals except rat. To be eligible for model construction, every sequence was represented by specific feature vector assembled from encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility for each residue in the sequence (13,15–19). A positive vector of interaction pair was formed by joining the vectors of the cytokine and its complementary receptor. To enlarge the positive data pool, four virtual vectors were generated around each positive vector by slightly (about 1/1000 folds) increasing/decreasing the value of vector elements in multi-dimension space. As a result, total 2243 positive data (449 true positives and 1794 virtual positives) were prepared for model training.

*The negative data pool.* The negative data pool includes both the true and the virtual data. The true negatives are literature-reported 126 non-cytokine–protein interactions, which are very limited in the representation of sequential and structural features of non-cytokine–receptor interactions. To cover all possible negative conditions, a large number of virtual negative interaction pairs were generated as follows: 7816 seed sequences representing diverse domain families, excluding those containing any known cytokine or its receptor, were extracted from Pfam protein families database (20). These Pfam seeds were paired with, covering all possible combinations, mammal cytokines to form the virtual negative interactions. Same transformations from sequences to vectors were demonstrated to these negative interaction pairs as described earlier. Totally, about 1 million negative data were ready in negative data pool.

### The SVM algorithm

The theory of SVM has been well described in literature (21,22). The structural and physicochemical features of a protein interaction are represented by a feature vector quantified from its primary sequence as described earlier. The vector is projected into a hyperspace wherein a hyperplane is used to classify this protein interaction pair as either positive (cytokine–receptor interaction) or negative (non-cytokine–receptor interaction) depending on the side of the hyperplane the vector is located. In this study, an RBF kernel function $K(x_i, x_j)$ was adopted to map the input vector into a high dimensional feature space:

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2), \gamma > 0. \quad \mathbf{1}$$

The output of SVM model is the respective class of input, directly associated with the posterior probability by fitting a sigmoid (23):

$$P(y = 1|f(x)) = \frac{1}{1 + \exp(Af(x) + B)} \quad \mathbf{2}$$

where $f(x)$ is the output of SVM, and the parameters $A$ and $B$ are estimated from the negative log likelihood of the training data. A higher probability indicates the higher confidence of positive prediction.

### Evaluation and performance measure

As in the case of all discriminative methods (24), the performance of SVM classification can be measured by: the quantity of true positives $TP$, true negatives $TN$, false positives $FP$, false negatives $FN$, sensitivity $SE = TP/(TP + FN)$ which is the accuracy of cytokine–protein interaction prediction and specificity $SP = TN/(TN + FP)$ which is the accuracy of non-cytokine–protein interaction prediction. The overall performance of the model can be measured both using the Matthews correlation coefficient (MCC) below:

$$MCC = \frac{TP \cdot TN + FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad \mathbf{3}$$

and a receiver operating characteristic (ROC) plot (25). ROC plot is a plot of the true positive rate against the false positive rate for the different possible thresholds of a model test. The area under the ROC curve (AUC) is usually adopted as a scalar measure that gauges one facet of performance (25). In this study, ROC plot (Please refer to http://bioinf.xmu.edu.cn/software/cytosvm/help.htm#roccurve) and AUC were used to compare the performance of different SVM models (Table 1). It is shown that the enriched-SVM model with virtual positives (model M1) has the best performance, which was chosen to classify the cytokine–receptor interactions.

### The enriched-SVM model

The model construction adopted all 2243 real and virtual positive vectors in positive data pool and about 1 million negative vectors in negative data pool. To represent all negative sequential and structural features and at the same time reduce the very unbalance between positive and negative data, the vectors in negative data pool were randomly divided into 230 groups of about 4200 negative vectors. These 230 negative data groups were combined with the 2243 positive data respectively to form totally 230 data sets for the construction of model. These data sets were arranged in the way of: 229 groups were used for independent trainings, while the remaining one was left for testing.

The SVM model was initialized by 229 independent model trainings and optimized through several rounds of training in an enriched way. The negative support vectors (vectors close to the hyperplane on negative side) that decide the hyperplanes of the 229 independent models were extracted to form a new negative data pool. This pool was further arranged into groups for next round of learning process. The iterative learning process, or enriched selection of support vectors, was continued to seek the global optimally separating hyperplane (OSH) until the positive and negative data come to a near balance, of which the ratio is about 1:3 in this case. The optimized model was first tested by the remaining data set to assess its theoretical performance, which achieved sensitivity of 100%, specificity of 99.98% and MCC of 0.99. Considering the 'overfitting' problem due to the overtraining on the same data set, the model was further independently evaluated by 79 real cytokine–receptor interactions and 2360 generated negative data in rat, achieving sensitivity of 97.4%, specificity of 99.2% and MCC of 0.89 (Table 2). Such performance is comparable to other computational approaches in protein–protein interactions.

**Table 1.** The descriptions of different SVM models

| Model | Enriched training[a] | Virtual Positives[b] | Ratio of Positive/Negative[c] | AUC[d] |
|---|---|---|---|---|
| M1 | Yes | Yes | 1: 2.83 | 0.9692 |
| M2 | Yes | No | 1: 3.37 | 0.9204 |
| M3 | No | No | 1: 2.83 | 0.8856 |
| M4 | No | Yes | 1: 3.31 | 0.8897 |
| M5 | No | Yes | 1: 1.08 | 0.8353 |
| M6 | No | Yes | 1: 6.02 | 0.8834 |

[a]Yes' means all virtual negative vectors are adopted for model training in an iterative manner (the enriched training). 'No' means only certain portion of random selection of negative vectors are used for model training.

[b]'Yes' means virtual positive vectors are adopted for model training. 'No' means no virtual positive vectors are used.

[c]The ratio of positive vectors against negative vectors in the training data sets. The ratio for Models M1-4 is about 1:3, M5 is about 1:1 and M6 is about 1:6.

[d]The area under receiver operating characteristic (ROC) plot. AUC is often used to measure the performance of models; the higher value indicates better performance.

## THE ACCESS OF SERVER AND DATABASE

### The descriptions of server

The web-based server upon the optimized CytoSVM model can be freely accessed at http://bioinf.xmu.edu.cn/software/cytosvm/PredictReceptor.php (Figure 1). The server runs under Linux environment that allows user to submit the query through a PHP-coded dynamic interface. The default input of the server is the protein primary sequence of putative receptor/cytokine in standard FASTA format or raw data format. The server is case insensitive, however, wild characters like '*,-' and non-amino acids characters will be removed from sequence automatically. An optional function of prediction by protein names is also provided. To initialize the prediction, user is required to choose a cytokine/receptor or cytokine/receptor families as well. The output of the server is the list of cytokines/receptors which are able to interact with query sequence with certain probabilities. Clicking on the name of a cytokine/receptor will lead user to the detailed information page, where user may find links to search for other putative receptors interacting with this cytokine in human or mouse genomes.

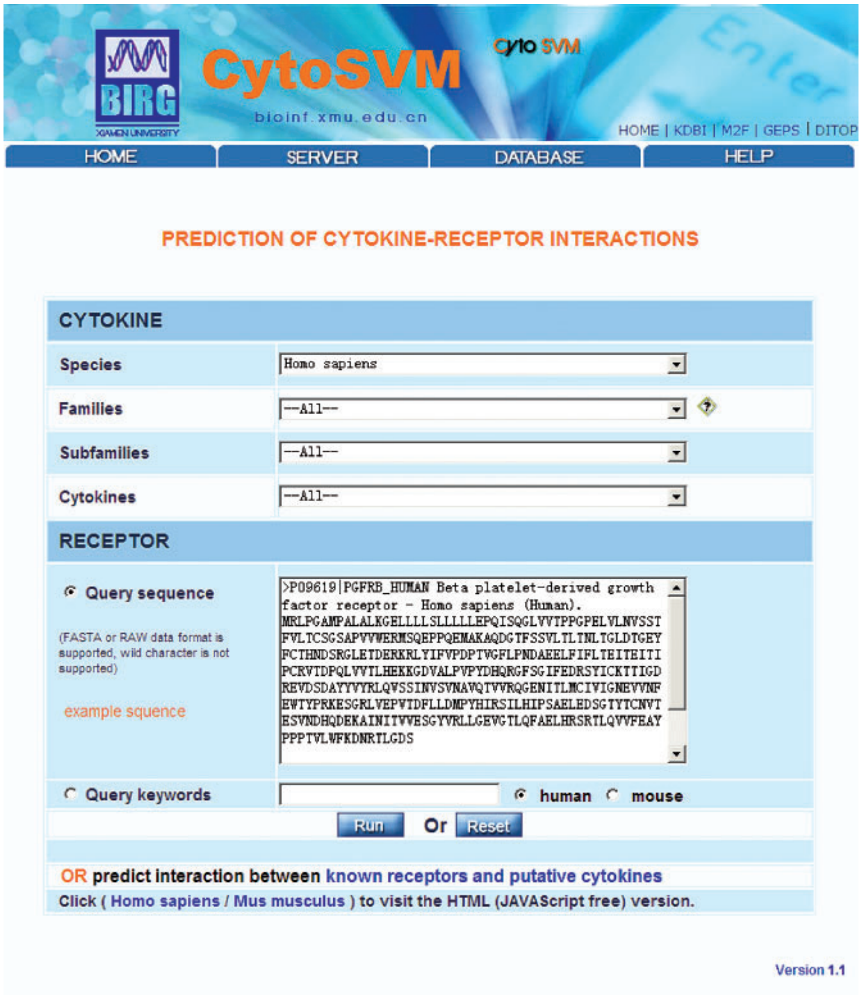
### The descriptions of database

In this study, the well-optimized CytoSVM model was also applied to screen putative cytokine–receptor interactions in whole genomes of human and mouse. Finally, 1609 novel cytokine-receptor pairs with probability >80% (3346 pairs with probability >50%), covering 220 novel receptors (excluding their isoforms) for 126 human cytokines were identified in human genome after further transmembrane analysis (http://bioinf.xmu.edu.cn/software/cytosvm/statistics.php). These predicted results were deposited in a database at http://bioinf.xmu.edu.cn/software/cytosvm/BrowseSearch.php. The database is running upon Linux/Apache/PHP platform and maintained by RDBMS system of *Oracle 9i*, which enables multiple accesses simultaneously. User is allowed to search the putative receptors of a definite cytokine by selecting the item from the cytokine classification list (Figure 2). Quick search by keywords is also supported to find putative interactions of cytokines or receptors. Only interactions with probability value >50%

**Table 2.** The evaluation of CytoSVM model

| Testing set | | | | | | | Independent evaluation set | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Positive | | | Negative | | | MCC | Positive | | | Negative | | | MCC |
| TP | FN | SE (%) | TN | FP | SP (%) | | TP | FN | SE (%) | TN | FP | SP (%) | |
| 2343 | 0 | 100 | 4445 | 1 | 99.98 | 0.99 | 77 | 2 | 97.4 | 2343 | 17 | 99.2 | 0.89 |

TP: true positives; FN: false negatives; TN: true negatives; FP: false positives; SE: sensitivity SE = TP/(TP + FN); SP: specificity SP = TN/(TN + FP); MCC: Matthews correlation coefficient.



Species: Homo sapiens

Families: --All--

Subfamilies: --All--

Cytokines: --All--

● Query sequence

(FASTA or RAW data format is supported, wild character is not supported)

example squence

```
>P09619|PGFRB_HUMAN Beta platelet-derived growth
factor receptor - Homo sapiens (Human).
MRLPGAMPALALKGELLLLSLLLLLEPQISQGLVVTPPGPELVLNVSST
FVLTCSGSAPVVWERMSQEPPQEMAKAQDGTFSSVLTLTNLTGLDTGEY
FCTHNDSRGLETDERKRLYIFVPDPTVGFLPNDAEELFIFLTEITEITI
PCRVTDPQLVVTLHEKKGDVALPVPYDHQRGFSGIFEDRSYICKTTIGD
REVDSDAYYVYRLQVSSINVSVNAVQTVVRQGENITLMCIVIGNEVVNF
EWTYPRKESGRLVEPVTDFLLDMPYHIRSILHIPSAELEDSGTYTCNVT
ESVNDHQDEKAINITVVESGYVRLLGEVGTLQFAELHRSRTLQVVFEAY
PPPTVLWFKDNRTLGDS
```

○ Query keywords ● human ○ mouse

OR predict interaction between known receptors and putative cytokines

Click ( Homo sapiens / Mus musculus ) to visit the HTML (JAVAScript free) version.

**Figure 1.** The interface of CytoSVM server.

will be responded for each single search. Clicking on the name of a cytokine or receptor will guide user into the detailed information page, where the general properties of the interactive partners are shown. Statistic of putative cytokine-receptor pairs in human genome and the help documents are also provided to aid database and server access.

## CONCLUSION

In conclusion, a web-based enriched-SVM model, CytoSVM, was successfully constructed in this study to predict the putative cytokine–receptor interactions. As a complementary method to homologous-based methods and other computational approaches in protein–protein interaction prediction, CytoSVM shows its capability in functionally annotating those proteins that possess poor sequence similarity to known proteins. The application of CytoSVM in the discovery of novel cytokine–receptor interactions in genome scale broadens the understanding of cytokines' physiological activities in the systematic level. Via these predicted interactants, the identification of novel cytokine-involved cellular processes is possible. Furthermore, it prompts the identification of new therapeutic targets for the treatment of various diseases. It is thus expected that experimental verifications could be demonstrated according to the clues provided by our study in the future.

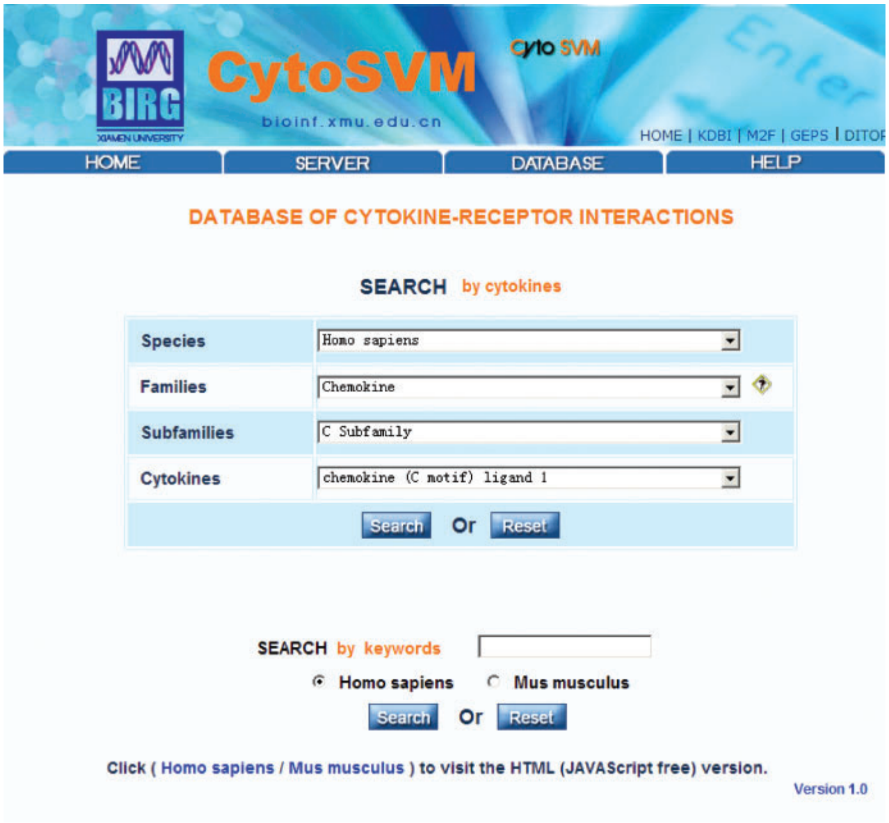
## ACKNOWLEDGMENTS

The support from the National Natural Science Foundation of China (#30400573) and the Program for New Century Excellent Talents in University (NCET) of MOE and Xiamen University are gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by NCET 2006 to ZL Ji.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Oppenheim,J.J. (2001) Cytokines: past, present, and future. *Int. J. Hematol.*, **74**, 3–8.
2. Altermann,E. and Klaenhammer,T.R. (2005) PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics*, **6**, 60–66.
3. Pazos,F., Ranea,J.A., Juan,D. and Sternberg,M.J. (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.*, **352**, 1002–1015.



DATABASE OF CYTOKINE-RECEPTOR INTERACTIONS

SEARCH by cytokines

| | |
|---|---|
| Species | Homo sapiens |
| Families | Chemokine |
| Subfamilies | C Subfamily |
| Cytokines | chemokine (C motif) ligand 1 |

SEARCH by keywords

● Homo sapiens ○ Mus musculus

Click ( Homo sapiens / Mus musculus ) to visit the HTML (JAVAScript free) version.

**Figure 2.** The interface of CytoSVM database.

4. Goh,C.S., Bogan,A.A., Joachimiak,M., Walther,D. and Cohen,F.E. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.
5. Fujiwara,Y. and Asogawa,M. (2002) Protein function prediction using hidden Markov models and neural networks. *NEC. Res. Dev.*, **43**, 238–241.
6. Di Gennaro,J.A., Siew,N., Hoffman,B.T., Zhang,L., Skolnick,J., Neilson,L.I. and Fetrow,J.S. (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.*, **134**, 232–245.
7. Teichmann,S.A., Murzin,A.G. and Chothia,C. (2001) Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol.*, **11**, 354–363.
8. Chen,R. and Weng,Z. (2002) Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, **47**, 281–294.
9. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
10. Marcotte,E.M., 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.
11. Hodges,H.C. and Tsai,J.W. (2002) 3D-Motifs: an informatics approach to protein function prediction. *FASEB. J.*, **16**, A543–A543.
12. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) an efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
13. Cai,C.Z., Han,L.Y., Ji,Z.L., Chen,X. and Chen,Y.Z. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**, 3692–3697.
14. Huang,N., Chen,H. and Sun,Z. (2005) CTKPred: an SVM-based method for the prediction and classification of the cytokine superfamily. *Protein Eng. Des. Sel.*, **18**, 365–368.
15. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
16. Ding,C.H. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.
17. Yuan,Z., Burrage,K. and Mattick,J.S. (2002) Prediction of protein solvent accessibility using support vector machines. *Proteins*, **48**, 566–570.
18. Hua,S. and Sun,Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.
19. Bock,J.R. and Gough,D.A. (2001) Predicting protein—protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.
20. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
21. Burges,C. (1998) A Tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.*, **2**, 121–167.
22. Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory* Springer, New York.
23. Platt,J.C. (2000) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola,A.J., Bartlett,P.L., Schölkopf,B. and Schuurnabs,D. (eds), *Advances in Large Margin Classiers*. MIT Press, Cambridge, MA, pp. 61–74.
24. 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.
25. Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.