W158–W163 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl331

KemaDom: a web server for domain prediction using kernel machine with local context

Lusheng Chen[1,2], Wei Wang[3], Shaoping Ling[4], Caiyan Jia[1,2] and Fei Wang[1,2,*]

[1]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, PR China, [2]Department of Computer Science and Engineering and and [3]Institute of Genetics, School of Life Science, Fudan University, Shanghai, PR China [4]College of Information Engineering, Xiangtan University, Xiangtan, Hunan, PR China

Received February 13, 2006; Revised March 1, 2006; Accepted April 14, 2006

ABSTRACT

Predicting domains of proteins is an important and challenging problem in computational biology because of its significant role in understanding the complexity of proteomes. Although many template-based prediction servers have been developed, *ab initio* methods should be designed and further improved to be the complementarity of the template-based methods. In this paper, we present a novel domain prediction system KemaDom by ensembling three kernel machines with the local context information among neighboring amino acids. KemaDom, an alternative *ab initio* predictor, can achieve high performance in predicting the number of domains in proteins. It is freely accessible at http://www.iipl.fudan.edu.cn/lschen/kemadom.htm and http://www.iipl.fudan.edu.cn/~lschen/kemadom.htm.

INTRODUCTION

Domains are the structural, functional and evolutionary units of proteins. Most multidomain proteins are formed by duplication, divergence and recombination of domains in the history of evolution (1). Thus domains are a key to understand the evolution of proteomes and their complexities. It is therefore of great importance to predict domains in proteins. The importance of this task has been emphasized by the CASP 6 (http://predictioncenter.org/) and the CAFASP 4 (http://www.cs.bgu.ac.il/dfischer/CAFASP4/ and http://www.cs.bgu.ac.il/~dfischer/CAFASP4/) protein structure prediction experiments. However, predicting domains from sequence remains an open problem.

Previous works exhibit great successes in domain prediction. Most of them are online web servers which can be publicly accessed from Internet. All these methods can be classified into two classes: template-based methods (scoring the sequence against domain templates or secondary structure elements) and *ab initio* methods (non-template methods). The template-based methods include Robetta-Ginzu (2), http://ekhidna.biocenter.helsinki.fi:9801/sqgraph/pairsdb,ADDA (3); http://bioinf.cs.ucl.ac.uk/dompred/DomPredform.html, Dompred-Domssea (4); Dopro (5); http://www.ebi.ac.uk/InterProScan, InterProScan (6); and http://www.bio.ifi.lmu.de/SSEP/, SSEP-Domain (7). And the *ab initio* methods include http://biozon.org/tools/domain/, Biozon(8); CHOPnet (9); Armadillo (http://armadillo.blueprint.org/; http://www.ics.uci.edu/baldig/dompro.html), DOMpro (10); http://bioinf.cs.ucl.ac.uk/dompred/DomPredform.html, Dompred-DPS (11); http://globplot.embl.de/, Globplot (12); and http://bioinformatics.cribi.unipd.it/cgi-bin/primex_client.cgi, Mateo (13). Additionally, http://meta-dp.cse.buffalo.edu/ Meta-DP (14) is an integrated domain prediction server which ensembles various template-based and *ab initio* methods with a 'majority voting' strategy.

Template-based methods become less effective when a potential domain shares low similarity with the identified domains. Thus, with the availability of domain databases such as CATH (15), SCOP (16) and FSSP-Dali Domain Dictionary (17), the effective *ab initio* methods using machine learning techniques have been developed (8–10). These methods using different artificial neural networks with various features have made important contributions to this task. Biozon (8) is a hybrid learning system for domain prediction and adopts a feed-forward network using back-propagation algorithm. In this system, the input units consist of sequence termination, correlation, contact profile, class and amino acid entropy, secondary structure, and physio-chemical properties. CHOPnet (9) also uses a three-layer feed-forward neural network but with different features, including secondary structure, solvent accessibility, HSSP conservation weight, the profile of six critical residues {P, H, D, Y, V, C},

*To whom correspondence should be addressed. Tel: +86 21 5566 4712; Fax: +86 21 6565 4253; Email: wangfei@fudan.edu.cn.
These authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authours

secondary structure difference and flexibility of a five-residue segment. These features are proved to be important to the performance of the network. DOMpro (10) applies the 1D-recursive neural network that leverages evolutionary profiles, predicted secondary structure and relative solvent accessibility. It is ranked among the top *ab initio* domain predictors in the CAFASP 4 evaluation.

Since the most important step of the *ab initio* methods for domain prediction is to discriminate boundary residues from domain residues, the prediction can be viewed as a two-class classification problem. As to the classifier, support vector machine (SVM), a classical kernel machine, not only is well-founded theoretically, but also has satisfactory abilities of generalization and avoiding over-fitting (18). Encouraged by the successful applications of SVM in computational biology, including remote protein homology detection (19,20), secondary structure prediction (21,22), and the like, we developed a novel predictor, KemaDom abbreviated from 'kernel machine for domain prediction', by ensembling three SVM classifiers, KemaSelf, KemaNeiOne and KemaNeiTwo, with different feature subspaces. The SVM classifiers with different feature subspaces improve the diversity of the result. It makes the ensemble work though SVM is a stable classifier and simply ensembling this kind of classifier with same features is not a good choice. The empirical study has shown that KemaDom has good performance in practice for predicting domains in proteins.

MATERIALS AND METHODS

Training and testing data

Liu *et al*. (9) have curated a dataset from multiple sources and Cheng *et al*. (10) have curated another dataset from CATH (15) to avoid the data conflict. In this paper, the latter is used to develop and test the algorithm. In this dataset, a total of 354 multi-domain chains and 963 single-domain chains are retrieved. Among these chains, no pair of sequences share sequence similarity above 25% in a global alignment of length 250. The sequences and the information of secondary structure and solvent accessibility can be obtained from Cheng's website (http://contact.ics.uci.edu/download.html).

In the prediction procedure, we focus on discriminating boundary residues from domain residues. Thus, multi-domain chains are used for training and testing, and single-domain chains are only for testing against the model trained by multi-domain chains. Additionally, a blind set from CAFASP 4 is used as the testing set.

Feature extraction

Feature extraction for training and testing is crucial to the model. In our method, we obtain amino acid entropy and physio-chemical properties according to the profile of amino acids. Amino acid entropy measuring the conservation of an alignment can be computed by information entropy. Ferran *et al*. clustered the 20 residues into 6 classes according to similarity scores of their physio-chemical property (23). One measurement for physio-chemical property is class entropy defined in Ref. (8). Alternatively, we only choose the value of the representative residue from each class to denote physio-chemical property. The six residues are {D, H, C, P, Y, V} because they are most different between domain residues and boundary residues (9). The difference of average profile of critical residues and the difference of average profile of six physio-chemical classes between boundary residues and domain residue (Figure 1) indicate that the latter is more proper as feature units. Secondary structure and relative solvent accessibility can be predicted by widely accepted tools.

According to the above analysis, three sub-models with different input units are designed (Table 1). For KemaSelf, 32 U are extracted as the inputs: 6 U represent physio-chemical information, 1 U represents amino acid entropy, 5×3 U are secondary structure of five-residue segment (a center resiude, two left neighborhoods and two right neighborhoods), 5×2 U represent solvent accessibility of the segment. For KemaNeiOne (or KemaNeiTwo), 26 U are extracted as the inputs: 2×3 U denote secondary structure of the residues with distance $d = 1$ (or $d = 2$) from the center residue, 2×2 U encode solvent accessibility of those residues, 2 U are amino acid entropy, 2×6 U denote physio-chemical properties and the last 2 U allow the exceeding of the N-terminus or C-terminus of the chain.

The model and post-processing

Figure 2 shows the architecture of KemaDom which integrates three binary classification sub-models, KemaSelf, KemaNeiOne and KemaNeiTwo. SVM with probability estimates is used to work out the probability of a residue belonging to boundary residue class, P^{KemaSelf}, $P^{\text{KemaNeiOne}}$ and $P^{\text{KemaNeiTwo}}$. The free online tool, libsvm (http://www.csie.ntu.edu.tw/cjlin/libsvmtools/ http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/), is modified for domain prediction purpose. Among the classical kernels, the radial basic function (RBF) is adopted because of its superior performance in generalization ability and convergence speed (18). After the kernel selection, the parameters C and γ are determined as $C = 4$ and $\gamma = 2$, separately.

A residue can be assigned into boundary residue class with the probability $P = \max\{P^{\text{KemaSelf}}, P^{\text{KemaNeiOne}}, P^{\text{KemaNeiTwo}}\}$ and non-boundary residue class with $1 - P$. As we know, the output of the learning model is quite noisy. So we smooth the result by averaging the probabilities of three consecutive residues. To reduce the influence of false signals, we regard that any two boundary residues with distance $d \leq 10$ belong to the same domain boundary region. This assumption is reasonable because the reliable domain boundaries can be accepted within 20 residues of the true domain boundary annotated in the CATH database (4,9–11). In addition, boundary residues with no neighboring boundary residues or with the distance <10 from the start position of a chain are ignored while computing the number of domains.

RESULTS AND DISCUSSION

In this section, we test our model and compare its performance with other methods. The measurements of sensitivity (S_N) and specificity (S_P) are the same with the classical one used in CASP 6 and CAFASP 4. The overall accuracy *Acc* is the number of correctly predicted chains over the

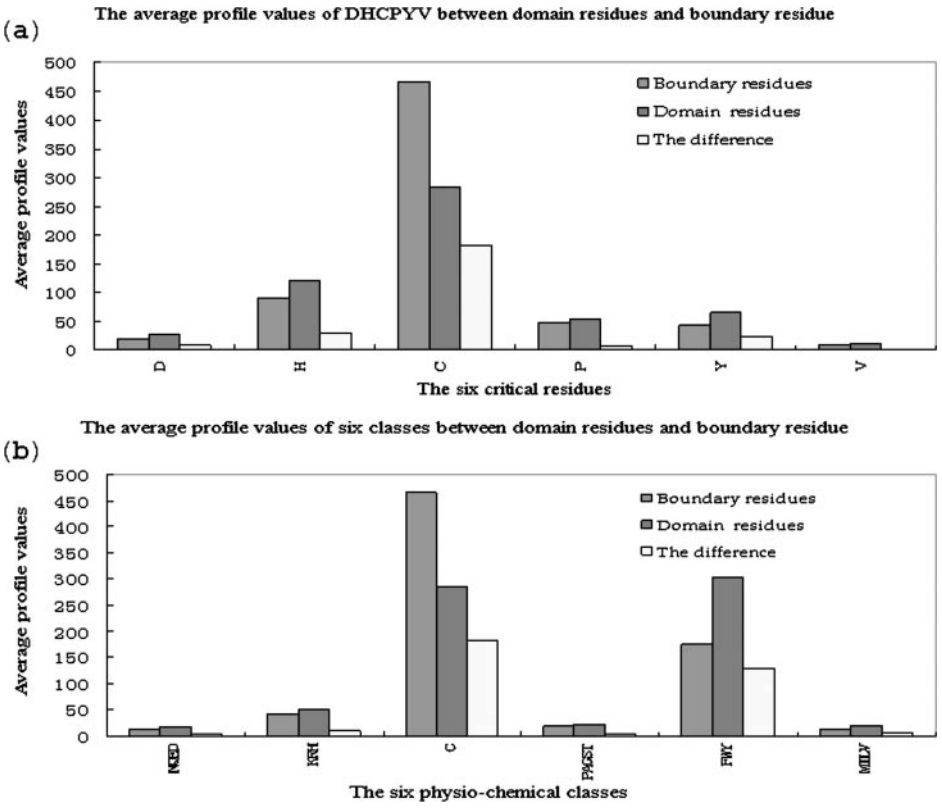

Figure 1. Comparison between average profile of critical residues (**a**) and the average profile of six physio-chemical classes (**b**).

Table 1. Features of the sub-models

Model	Unit position	Description
KemaSelf	1–5	Secondary structure and solvent accessibility of a center residue;
	6–11	Physio-chemical properties of a center residue;
	12–31	Secondary structure and solvent accessibility of residues with $0 < d \leq 2$;
	32	Amino acid entropy of a center residue;
KemaNeiOne	1–6	Secondary structure of the residues with $d = 1$;
	7–10	Solvent accessibility of the residues with $d = 1$;
	11–22	Physio-chemical properties of the residues with $d = 1$;
	23–24	Amino acid entropy of the neighboring residues with $d = 1$;
	25–26	Labels to denote the exceeding of the N-terminus or C-terminus of the chain.
KemaNeiTwo	1–6	Secondary structure of the left residues with $d = 2$;
	7–10	Solvent accessibility of the left residues with $d = 2$;
	11–22	Physio-chemical properties of the left residues with $d = 2$;
	23–24	Amino acid entropy of the neighboring residues with $d = 2$;
	25–26	Labels to denote the exceeding of the N-terminus or C-terminus of the chain.

total number of chains. Eightfold cross validation is used to measure the performance.

To provide a baseline to compare the result of KemaDom, we run the random control prediction algorithm as in Ref. (9)

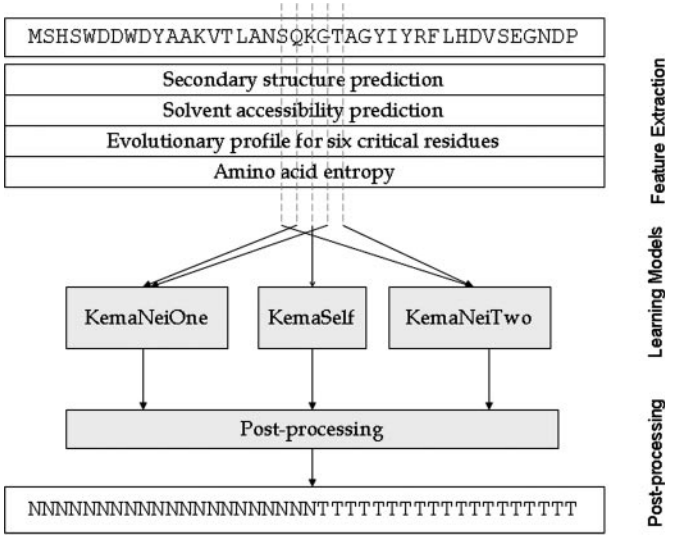

Figure 2. The architecture of KemaDom for domain prediction.

on the same dataset. First, the dataset is randomly divided into eight subsets. Then, the number of domains for proteins in each subset are predicted according to the composition of domain numbers in remaining subsets. We repeat this test 100 times and average over the results.

Performance of KemaDom and its sub-models

The results are shown in Table 2, where 1D denotes single-domain chains and 2D denotes two-domain chains. KemaSelf

Table 2. Performance of KemaDom and sub-models

Model/Sub-model	1D S_N	1D S_P	2D S_N	2D S_P	*Acc*
KemaDom	0.88	0.83	0.41	0.57	0.76
KemaSelf	0.89	0.81	0.36	0.55	0.74
KemaNeiOne	0.90	0.79	0.23	0.44	0.71
KemaNeiTwo	0.90	0.79	0.22	0.42	0.71
Baseline	0.74	0.72	0.26	0.23	0.60

achieves 3% higher *Acc*, 13% (14%) higher 2D S_N and 11% (13%) higher 2D S_P than KemaNeiOne (KemaNeiTwo). KemaNeiOne has 1% higher 2D S_N and 2% higher 2D S_P than KemaNeiTwo. This implies that the 1-neighboring residue information contribute more to identifying boundary residues than the 2-neighboring residue information does. After combining these three sub-models, KemaDom improves *Acc* up to 76%. And the sensitivity and specificity for single-domain chains are 88 and 83%, respectively. Those for two-domain chains increase to 41 and 57%, separately. In contrary, random control prediction method correctly predicts only 74% single-domain chains and 26% two-domain chains. These results show that the neighboring residue information can be used to improve the domain prediction and KemaDom is more effective than random control prediction method.

We also use an individual SVM with a combined feature map of three sub-models to predict domains. The results show that this strategy fails in prediction because only two two-domain chains are correctly predicted and others are all inferred to be single-domain chains. Although no well-established theory of this ensemble technique with different features has been given, the subspace ensemble for supervised learning has been successfully applied in bioinfomatics with a satisfactory result (24).

While predicting domain boundary position, KemaDom only correctly predicts 15% of the two-domain chains and 12% of the multi-domain chains; they are both lower than those of DOMpro, 25 and 20%, respectively. It should be pointed out that the reliable domain boundaries are acceptable within 20 residues of the true domain boundary annotated in CATH and predicting domain boundary locations is more difficult than predicting domain numbers.

Objectively, in order to evaluate the performance of KemaDom, we also test KemaDom against CAFASP 4 dataset, in which there are 41 single-domain chains and 17 two-domain chains. In these chains, KemaDom shows 95% 1D S_N, 77% 1D S_P, 24% 2D S_N and 57% 2D S_P. The *Acc* is 74% and the average overlap score of the two-domain chains is 64.18.

Performance comparison with other predictors

The performance of available *ab initio* systems can be taken from the previous publications and the website of CAFASP 4 (Table 3). It is easy to see that predicting two-domain or multi-domain chains is more difficult than predicting single-domain chains. The 2D S_N varies from 12% (Mateo) to 59% (DOMpro), and the 2D S_P ranges from 15% (Mateo) to 60% (Globplot) while the *Acc* lies between 17% (Biozon) and 76% (KemaDom). Moreover, the selection of training and testing datasets influences the performance of the predictors significantly.

Table 3. Performance of *ab initio* predictors[a]

Predictor name	1D S_N	1D S_P	2D S_N	2D S_P	*Acc*	Dataset
KemaDom	0.88	0.83	0.41	0.57	0.76	(10)
DOMpro	0.76	0.85	0.59	0.38	0.69	(10)
CHOPnet[b]	0.42–0.73	N/A	0.40–0.59	N/A	0.69	(9)
KemaDom	0.95	0.77	0.24	0.57	0.74	CAFASP 4
DOMpro	0.85	0.76	0.35	0.50	0.70	CAFASP 4
Biozon	0.10	10.00	0.35	0.19	0.17	CAFASP 4
Globplot	0.83	0.71	0.18	0.60	0.64	CAFASP 4
Dompred-DPS	0.68	0.78	0.47	0.50	0.62	CAFASP 4
Mateo	0.51	0.78	0.12	0.15	0.40	CAFASP 4

[a]The values taken from the previous publications and the website of CAFASP 4.
[b]The performance of CHOPnet is tested against multiple datasets with cross validation of networks; S_P values are not shown in their paper and are denoted by N/A in this table.

Compared with DOMpro, KemaDom achieves 19% higher 2D S_P and 7% higher *Acc* on the CATH dataset though it has 18% lower 2D S_N. Similarly, on CAFASP 4 dataset, KemaDom has 11% lower 2D S_N but 7% higher 2D S_P than DOMpro. Obviously, KemaDom achieves a good *Acc* because of its high 1D S_N. On this point, we can not conclude that our method is better or worse than the other methods because the knowledge is still not sufficient for discriminating the boundary residues exactly.

WEB SERVER: KemaDom

The web server can be accessed from http://www.iipl.fudan.edu.cn/lschen/kemadom.htm. and http://www.iipl.fudan.edu.cn/~lschen/kemadom.htm. This system is mainly composed of two subsystems, the background system and the interface system.

The background system is implemented by Perl including package BioPerl and CGI script. The whole processing flowchart of this system can be summarized as the following steps: (i) a remote user submits a target sequence to the server; (ii) a PSSM profile for the sequence is generated by PSI-blast (25) against the non-redundant (nr) database; and (iii) secondary structure prediction and solvent accessibility prediction are performed by SSpro (26) and ACCpro (27), respectively; (iv) a Perl script generates the feature vectors for all the residues of the input sequence; (v) boundary residues prediction is executed with the feature vectors against the trained model. (vi) post-processing is done for the raw output; and (vii) KemaDom sends the result to the user.

The interface system is written with HTML language. KemaDom provides a friendly interface (Figure 3). Users should submit sequences with the format which BioPerl (Bio::SeqIO) can recognize. Also, the email address and the customized job name are required in submission. The only constraint is that protein sequence to be predicted should contain >30 residues.

CONCLUSION

In this paper, we have presented a novel domain prediction server, KemaDom, modeling the local context information.

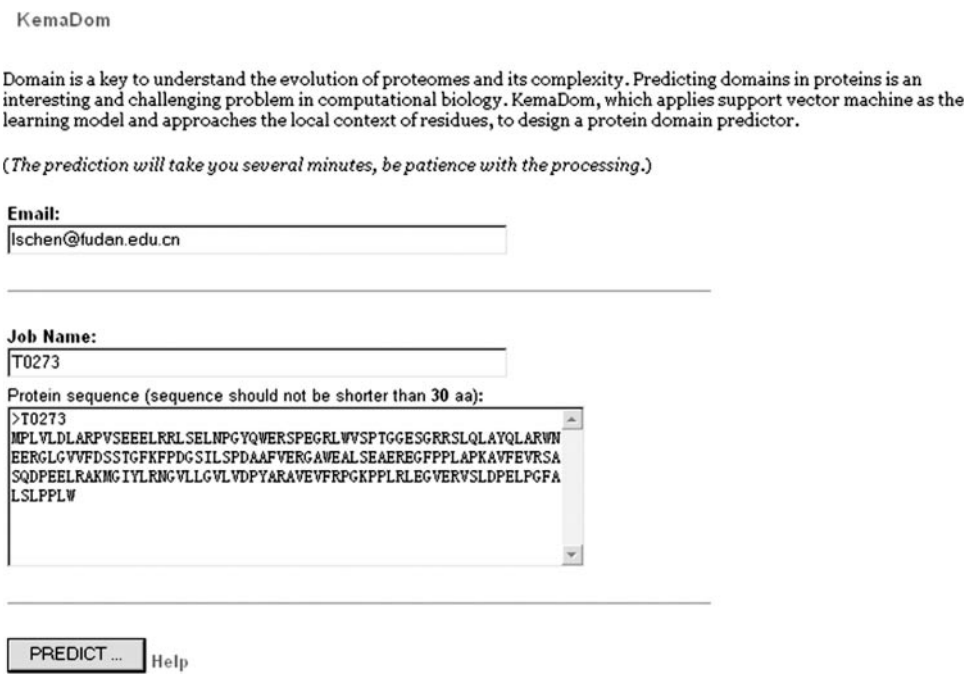

Figure 3. The interface of KemaDom. The Email address and the customized job name are required. The target sequence should be input with format which BioPerl (Bio::SeqIO) can recognize.

As a domain prediction server, it is powerful and easy to use. This method is a good option for domain prediction compared with the existing methods.

ACKNOWLEDGEMENTS

The authors would like to thank both anonymous reviewers for their constructive comments and Yanqiu Chen for improving the presentation of the manuscript. The authors also would like to thank Yu Xue from USTC, People Republic of China for his valuable discussion. This work is supported by the Major Research Program of the National Natural Science Foundation of China (No. 60496324) and the National Natural Science Foundation of China (No. 60303009). The authors would also like to acknowledge the grant of the Open Program of Beijing Municipal Key laboratory (No. KP701200372). Funding to pay the Open Access publication charges for this article was provided by Shanghai Key Laboratory of Intelligent Information Processing, Fudan University.

Conflict of interest statement. None declared.

REFERENCES

1. Vogel,C., Teichmann,S.A. and Pereira-Leal,J. (2004) The relationship between domain duplication and recombination. *J. Mol. Biol.*, **346**, 355–365.
2. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bon-neau,R., Rohl,C.A. and Baker,D. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, **53**, 524–533.
3. Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, **328**, 749–776.
4. Marsden,R.L., McGuffin,L.J. and Jones,D.T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.*, **11**, 2814–2824.
5. von Ohsen,N., Sommer,I., Zimmer,R. and Lengauer,T. (2004) Arby: automatic protein structure prediction using profile–profile alignment and confidence measures. *Bioinformatics*, **20**, 2228–2235.
6. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan: an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
7. Gewehr,J.E. and Zimmer,R. (2006) SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics*, **22**, 181–187.
8. Nagarajan,N. and Yona,G. (2004) Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics*, **20**, 1335–1360.
9. Liu,J. and Rost,B. (2004) Sequence-based prediction of protein domains. *Nucleic Acids Res.*, **32**, 3522–3530.
10. Cheng,J., Sweredoski,M.J. and Baldi,P. (2005) DOMpro: protein domain prediction using profiles,secondary structure, relative solvent accessibility, and recursive neural networks. *Data Mining and Knowledge Discovery*, in press.
11. Bryson,K., McGuffin,L.J., Marsden,R.L., Ward,J.J., Sodhi,J.S. and Jones,D.T. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res.*, **33**, W36–W38.
12. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
13. Lexa,M. and Valle,G. (2003) PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics*, **19**, 2486–2488.
14. Saini,H.K. and Fischer,D. (2005) Meta-DP: domain prediction meta server. *Bioinformatics*, **21**, 2917–2920.
15. Orengo,C.A., Bray,J.E., Buchan,D.W., Harrison,A., Lee,D., Perl,F.M., Sillitoe,I., Todd,A.E. and Thornton,J.M. (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics*, **2**, 11–21.
16. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
17. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

18. Vapnik,V. (1998) *Statistical Learning Theory*. John Wiley and Sons. Ins., NY.
19. Busuttil,S., Abela,J. and Pace,G.J. (2004) Support vector machines with profile-based kernels for remote protein homology detection. *Genome Inform. Ser. Workshop Genome Inform.*, **15**, 191–200.
20. Saigo,H., Vert,J.P., Ueda,N. and Akutsu,T. (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.
21. 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.
22. Guo,J., Chen,H., Sun,Z. and Lin,Y. (2004) a novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins*, **54**, 738–743.
23. Ferran,E.A., Pflugfelder,B. and Ferrara,P. (1994) Self-organized neural maps of human protein sequences. *Protein Sci.*, **3**, 507–521.
24. Bertoni,A., Folgieri,R. and Valentini,G. (2005) Bio-molecular cancer prediction with random subspace ensembles of Support Vector Machines. *Neurocomputing*, **63C**, 535–539.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem. *J. Mach. Learn. Res.*, **4**, 575–602.
27. Pollastri,G., Baldi,P., Fariselli,P. and Casadio,R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.