Protein domain boundary prediction by combining support vector machine and domain guess by size algorithm[①]

Dong Qiwen (董启文)[②], Wang Xiaolong, Lin Lei
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, P. R. China)

Abstract

Successful prediction of protein domain boundaries provides valuable information not only for the computational structure prediction of multi-domain proteins but also for the experimental structure determination. A novel method for domain boundary prediction has been presented, which combines the support vector machine with domain guess by size algorithm. Since the evolutional information of multiple domains can be detected by position specific score matrix, the support vector machine method is trained and tested using the values of position specific score matrix generated by PSI-BLAST. The candidate domain boundaries are selected from the output of support vector machine, and are then inputted to domain guess by size algorithm to give the final results of domain boundary prediction. The experimental results show that the combined method outperforms the individual method of both support vector machine and domain guess by size.

Key words: domain boundary prediction, support vector machine, domain guess by size, position specific score matrix

0 Introduction

Domains are generally regarded as compact, semi-independent units[1] that could fold autonomously. The identification of structural domain boundaries is not only of theoretical interest but also of great practical importance. Successful domains boundary determination in proteins[2] is useful in structure analysis, function annotate[3], database searching[4], protein modeling[5], etc.

However, the domain annotation of a protein based solely on its sequence information, in the absence of the structural information, has remained as a difficult problem. A number of methods for predicting domain boundaries from the amino acid sequences have been developed. Early approaches to domain boundary prediction are based on information theory[6], statistical potentials[7] or similarity searches[8]. Subsequent methods relied on expert knowledge of protein families to construct models like hidden Markov models[9] and artificial neural networks[10] to identify other members of the family. However, the most useful and straightforward way to predict domains is using sequence homology or multiple sequence alignment. The ProDom[11] method generates a comprehensive set of protein domain families automatically from the SWISS-PROT and TrEMBL sequence databases[12]. DOMAINATION[13] delineates domains through analyzing position-specific iterative database search[14] alignments. Similarly, CHOP[15] identifies potential domain boundaries through hierarchical searches against databases of more or less well defined domains. PPRODO[16] uses the neural network with the position-specific scoring matrix (PSSM) generated by PSI-BLAST[14] to predict the domain boundary of two-domain proteins. Although all these methods provide valuable information about putative domains for proteins with similar sequences, they fail for small families or in the absence of homologous domain assignments. Recently, quite a few novel methods have been developed to predict domain boundaries directly from sequences. The 'Domain Guess by Size' (DGS) algorithm[17] 'guesses' domain boundaries solely based on observed domain size distributions. SnapDragon[18] predicts the domain by statistical analysis of the structure model generated by the *ab initio* protein structure method Dragon[19]. A hybrid learning system has been presented for domain boundaries prediction[20]. In general, most approaches predict the number of domains and only a few predict domain boundaries with reported sensitivity of between 50 % and 70 % for proteins with single domains and considerably less (< 30 %) for multi-domain proteins[21].

[①] Supported by the National Natural Science Foundation of China (No. 60435020).

[②] To whom correspondence should be addressed. E-mail: liuxuan@hit.edu.cn

Received on Dec. 22, 2005

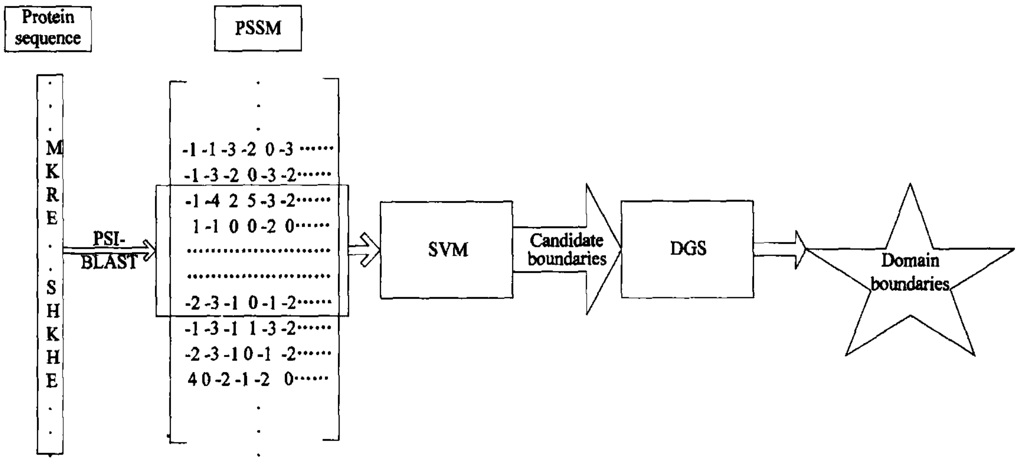

Fig. 1 The architecture of SVM + DGS method for domain boundary prediction. The position specific score matrix (PSSM) of a protein sequence is generated by PSI-BLAST. The neighboring elements of PSSM are inputted to SVM. The candidate boundaries are selected from the output of SVM, which are then inputted to DGS to give the results of domain boundaries.

Most domain boundary prediction methods generate a digital value for each position of the protein sequences indicating the likelihood that the position is a domain boundary. However, in most cases, the domain boundary doesn't correspond to the position with the optimal value. The DGS[17] algorithm introduces another way for domain boundary prediction, which is based solely on the domain size distribution. But the candidate domain boundaries are enumerated with a step size, which are too much and not enough precise. Here, a novel method for protein domain boundary prediction has been presented to overcome these limitations. The evolution information is inputted to Support Vector Machine (SVM)[22] in the form of Position-Specific Scoring Matrix (PSSM)[14]. The candidate domain boundaries are selected from the output of SVM, and is then inputted to DGS algorithm[17] to predict the domain boundaries. To the best of our knowledge, this is the first usage of SVM approach for domain boundary prediction. Experimental results show that such combined method outperforms the individual methods of both SVM and DGS.

1 Material and method

1.1 The protein data set

The standard evaluation data are taken from the Structural Classification of Proteins (SCOP) database[23] version 1.67. Sequences are selected using the ASTRAL Compendium[24] with sequence identity less than 40%. The proteins are then filtered by removing the proteins that contain unknown amino acids.

The resulting data set contains 6542 domains and is grouped into 5180 proteins. The multi-domain proteins are selected and divided into two dataset for usage. The SCOP-1 subset contains 600 proteins with two continuous domains and only one linker. The SCOP-2 subset contains 381 proteins with multi-domains, of which 181 proteins have two domains and 200 proteins have three or more domains.

1.2 SVM training

Support Vector Machine (SVM) is a class of supervised learning algorithms first introduced by Vapnik[22]. Given a set of labeled training vectors (positive and negative input samples), SVM can learn a decision boundary to discriminate between the two classes. The result is a classification rule that can be used to classify new test samples. SVM has exhibited excellent performance in practice and has strong theoretical foundation of statistical learning theory.

First, the PSSM of a protein sequence is generated by PSI-BLAST[14] with the default parameter values except that the number of iterations is set to 10. The search is performed on the nrdb90 database from EBI[25]. An extra residue is added to PSSM, indicating the vacancies at the N- and C-terminal ends. Then, the neighboring elements of PSSM are inputted to SVM by a sliding window. The output of SVM is the likelihood of the center residues that is a domain boundary. The window size is taken as 25, which is the optimal value of neural network for domain boundary prediction[16].

One of the technical difficulties in applying the SVM method to the domain boundary prediction is that the number of domain boundaries is extremely small in comparison with the total number of residues. To solve this problem, the boundary class is assigned in a less rigorous manner. In practice, all residues within ± 20 residues[16, 21, 26] from the true domain boundary residue are assigned into the boundary class and the rest into the non-boundary class. The total number of the residues assigned as the boundary class is 41.

In this study, the Gist SVM package implemented by Jaakkola et al.[27] is used for protein domain boundary prediction. The parameters of SVM are used by default of the Gist package except for the kernel that is the Radius Basis Function (RBF) kernel.

1.3 DGS filtering

The DGS algorithm[17] aims to predict the likelihood of domains within a given sequence based on probability distributions of chain and domain lengths. Supposing that there is a protein sequence with length n, having a set of d domains with individual lengths L and segment numbers S, the likelihood is:

$$L(d, L, S \mid n) = p(d \mid n)\, p(L, S \mid d, n) \quad (1)$$

where $p(d| n)$ is the probability that a chain of length n will have d domains, $p(L, S| d, n)$ is the probability that d domains will have lengths L and numbers of segments S.

The DGS algorithm enumerates a discrete list of possible domain boundaries. Across the list, $p(L, S| d, n)$ is estimated from empirical distributions for the length and segment number of individual domains, as observed in the training set:

$$p(L, S \mid d, n) = \prod_{L,S} p(l)\, p(s) / \sum [\prod_{L,S} p(l)\, p(s)] \quad (2)$$

where $p(l)$ is the probability of an individual domain with length l, $p(s)$ is the probability that an individual domain will be formed from s-chain continuous segments. The product is taken over all the d domains whose lengths and segment numbers are given by L and S. The summation in the denominator is taken across all the candidate domain boundaries.

In this study, ten maximum residues are selected as the candidate domain boundaries according to the SVM output. When a residue is selected, the neighboring residues (±10) are ignored. The candidate domain boundaries are inputted to DGS, which gives the final results of domain boundaries. Fig. 1 has given an outline of the SVM + DGS method.

1.4 Performance metrics

Domain linker predictions are correct when the predicted domain linker overlaps wholly or in part between the correct linker boundaries plus a ±20 residue margin of error added to each boundary as done by others[16, 21, 26]. Two performance measures including sensitivity and specificity are adopted. The sensitivity is defined as TP/ (TP + FN) where TP is the number of true positive boundary predictions and FN is the number of false negatives. The specificity is defined as TP/ (TP + FP), where FP is the number of false positives.

2 Results

2.1 Length distribution

The length distribution of chains and domains are re-estimated since the dataset in this study (SCOP) is different from the original one of DGS work (VAST database from NCBI)[28] and the domain assignments between the two databases are different[21]. Fig. 2 shows the chain length distribution, which is used to calculate the probability of $p(d| n)$ in Eq. (1). As chain length increases, the likelihood of the chain having a multi-domain conformation also increases. Fig. 3 shows the domain length distribution,

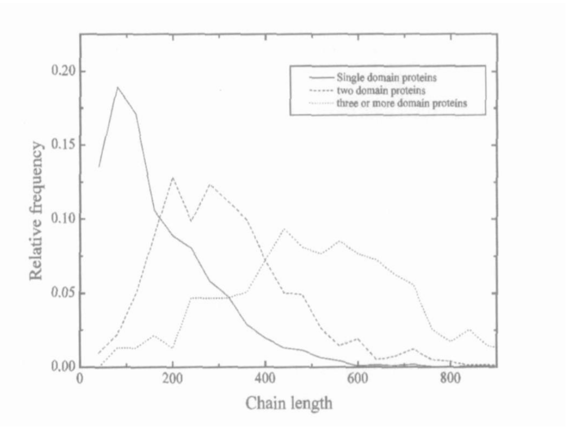

Fig. 2 Frequency of chain lengths of one, two, and three-or more domain chains in the SCOP40 dataset. A length-interval width of 40-residue is used to tabulate domain number frequency.

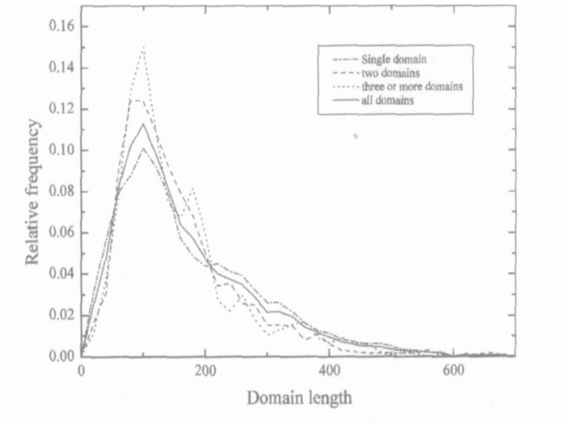

Fig. 3 Domain length distributions as observed in the SCOP representative set used in this study. An interval width of 20 residues is used to tabulate the number of domains.

from which the probability of $p(l)$ in Eq. (2) can be obtained. The mean lengths for domains in both single and multi-domain chains are similar (150 residues). The probability $p(s)$ in Eq. (2) is estimated from the fraction of domains in the training set formed from a single-chain continuous segment (0.97), two chain continuous segments (0.028), or three (or more) chain continuous segments (0.002).

2.2 Cross validation

The SCOP-1 dataset is randomly divided into ten groups. We use nine groups to train the SVM, and the other one to test. The above process is iterated until each group has been tested. The SVM + DGS method is compared with the individual methods of SVM and DGS. The results are shown in Table 1. Since the SCOP-1 dataset contains only two continuous domains and only one linker, the specificity and the sensitivity have the same value. The accuracy is used to denote the fraction of correctly identified domain boundaries. One can see from Table 1 that the combined method outperforms the individual methods of both SVM and DGS.

Table 1 Results of cross validation

	Top 1 accuracy	Top 5 accuracy
DGS	0.3415	0.7438
SVM	0.3672	0.7546
SVM + DGS	0.4026	0.7868

Notes: The testing process is performed on the SCOP-1 dataset with ten fold cross validation. Given in the table is the fraction of correctly identified domain boundaries of the top 1 and top 5 predictions.

Table 2 Comparative results on the multi-domain dataset

	Specificity	Sensitivity
DGS	0.2625	0.1924
SVM	0.2886	0.2257
SVM + DGS	0.3386	0.2755

Notes: The performance of top 1 prediction of various methods is given.

Table 3 Comparative results on the pdb25 dataset

	Specificity	Sensitivity
DGS	0.2537	0.1812
SVM	0.2663	0.2132
SVM + DGS	0.3248	0.2674

2.3 The multi-domain dataset

The secondary dataset used is the SCOP-2 dataset, in which the proteins contain two or more domains. The specificity and sensitivity are used as the evaluation criterions. For the SVM method, the residues whose prediction values are larger than a threshold are selected as the domain boundaries. The threshold is set as 0.1, which gives the best performance. The results are shown in Table 2. The performance of various methods on this dataset is lower than that on the two-domain proteins. The overall low success rate in SCOP-2 dataset may be caused by the inherent difficulty of domain prediction of multi-domain proteins. The SVM + DGS prediction is more sensitive and specific (5 %～8 %) than the individual methods.

Because the SCOP database with 40 % sequence identity may cause much redundancy, another experiment is made on the PDB subset with sequence identity less than 25 %. The results are shown in Table 3. As can be seen, the performance is lower than that on the SCOP dataset, because the low sequence identity is used. The SVM + DGS prediction still outperforms the individual methods.

2.4 Test on the CASP6 target proteins

We investigate the CASP6 targets for multi-domain proteins and find that only six proteins are available and have only two continuous domains. They are T0199, T0204, T0222, T0228, T0247 and T0260. Because the CASP6 targets are included in the latest SCOP database (version 1.69), none of the targets are contained in our training set that is derived from SCOP database version 1.67. The results are shown in Table 4. The performance of SVM + DGS method is reasonable.

Table 4 Domain linker prediction on CASP6 targets with two continuous domains

		T0199	T0204	T0222	T0228	T0247	T0260
	N_{res}	338	351	373	430	364	224
	True boundary	100	195	296	149	267	88
DGS	Prediction	138	192	250	290	250	120
	Error	38	3	46	141	17	32
SVM	Prediction	150	108	192	112	175	13
	Error	50	87	104	37	92	65
SVM + DGS	Prediction	139	200	215	155	201	113
	Error	39	5	81	6	66	25

Notes: N_{res} is the length of the protein chains. The true boundary is obtained by the domain annotation of SCOP. The error indicates the distance between the predicted boundary and the true boundary.

Conclusion

In this study, a novel method that combines support vector machine with domain guess by size algorithm for domain boundary prediction has been presented. The first usage of support vector machine for domain boundary prediction is introduced. Such method overcomes the shortcomings of both SVM and DGS and gets improved performance according to the experimental results. The advantages of our method are as follows:

1. The evolutional information has been used for domain boundary prediction.

2. The candidate domain boundaries are selected form the output of SVM, which are smaller and more accurate than that of the original DGS algorithm.

3. The predicted domain boundaries may not be the residues with the optimal values, whereas in SVM method they are always corresponding to the residues with the optimal values.

Reference

[1] Jaenicke R. Folding and association of proteins. *Prog Biophys Mol Biol*, 1987, 49(2-3): 117-237

[2] Saini H K, Fischer D. Meta-Dp: Domain prediction Meta-Server. *Bioinformatics*, 2005, 21(12): 2917-2920

[3] Wen Z N, Wang K L, Li M L, et al. Analyzing functional similarity of protein sequences with discrete wavelet transform. *Comput Biol Chem*, 2005, 29(3): 220-228

[4] Nikitin F, Lisacek F. Investigating protein domain combinations in complete proteomes. *Comput Biol Chem*, 2003, 27 (4-5): 481-495

[5] Xiao J F, Li Z S, Sun M, et al. Homology modeling and molecular dynamics study of Gsk3/ shaggy-like kinase. *Comput Biol Chem*, 2004, 28(3): 179-188

[6] Busetta B, Barrans Y. The prediction of protein domains. *Biochim Biophys Acta*, 1984, 790(2): 117-124

[7] Kikuchi T, Nemethy G, Scheraga H A. Prediction of the location of structural domains in globular proteins. *J Protein Chem*, 1988, 7(4): 427-471

[8] Gouzy J, Corpet F, Kahn D. Whole genome protein domain analysis using a new method for domain clustering. *Comput Chem*, 1999, 23(3-4): 333-440

[9] Ponting C P, Schultz J, Milpetz F, et al. Smart: Identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res*, 1999, 27(1): 229-232

[10] Miyazaki S, Kuroda Y, Yokoyama S. Characterization and prediction of linker sequences of multi-domain proteins by a neural network. *J Struct Funct Genomics*, 2002, 2(1): 37-51

[11] Corpet F, Servant F, Gouzy J, et al. Prodom and Prodom-Cg: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*, 2000, 28(1): 267-269

[12] Boeckmann B, Bairoch A, Apweiler R, et al. The Swiss-Prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, 2003, 31(1): 365-370

[13] George R A, Heringa J. Protein domain identification and improved sequence similarity searching using Psi-Blast. *Proteins*, 2002, 48(4): 672-681

[14] Altschul S F, Madden T L, Schaffer A A, et al. Gapped blast and Psi-Blast: a new generation of protein database search programs. *Nucleic Acids Research*, 1997, 25(17): 3389-3402

[15] Liu J F, Rost B. Chop proteins into structural domain-like fragments. *PROTEINS: Structure, Function, and Bioinformatics*, 2004, 55(3): 678-688

[16] Sim J, Kim S Y, Lee J. Pprodo: prediction of protein domain boundaries using neural networks. *Proteins*, 2005, 59(3): 627-632

[17] Wheelan S J, Marchler-Bauer A, Bryant S H. Domain size distributions can predict domain boundaries. *Bioinformatics*, 2000, 16(7): 613-618

[18] George R A, Snapdragon H J. A method to delineate protein structural domains from sequence data. *J Mol Biol*, 2002, 16 (3): 839-851

[19] Aszodi A, Gradwell M J, Taylor W R. Global fold determination from a small number of distance restraints. *J Mol Biol*, 1995, 251(2): 308-326

[20] Nagarajan N, Yona G. Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics*, 2004, 20(9): 1335-1360

[21] Dumontier M, Yao R, Feldman H J, et al. Armadillo: domain boundary prediction by amino acid composition. *J Mol Biol*, 2005, 350(5): 1061-1073

[22] Vapnik V N. Statistical Learning Theory. 1998, New York: Wiley, 1998

[23] Andreeva A, Howorth D, Brenner S E, et al. Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 2004, 32 (database issue): D226-D229

[24] Chandonia J M, Hon G, Walker N S, et al. The astral compendium in 2004. *Nucleic Acids Research*, 2004, 32(database issue): 189-192

[25] Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, 1998, 14(5): 423-429

[26] Marsden R L, McGuffin L J, Jones D T. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci*, 2002, 11(12): 2814-2824

[27] Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. *J Comput Biol*, 2000, 7(1-2): 95-114

[28] Matsuo Y, Bryant S H. Identification of homologous core structures. *Proteins*, 1999, 35(1): 70-79

Dong Qiwen, born in 1977. He is a Ph. D candidate in School of Computer Science and Technology in Harbin Institute of Technology. He received his B. S. and M. S. degrees from Harbin Institute of Technology in 2000 and 2002 respectively. His research interests include computational investigation of sequence-structure-function relationships in proteins and the language model of biological sequence.