Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 228 (2004) 551–557
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

Identify catalytic triads of serine hydrolases by support vector machines

Yu-dong Cai[a,*], Guo-Ping Zhou[b], Chin-Hung Jen[c], Shuo-Liang Lin[d], Kuo-Chen Chou[e,f,g]

[a] *Shanghai Research Center of Biotechnology, Chinese Academy of Sciences, Shanghai, 200233, China*
[b] *Harvard Medical School, Beth Israel Deaconess Medical Center, 41 Avenue Louis Pasteur, Research East Room 301, Boston, MA 02115, USA*
[c] *School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK*
[d] *Wyeth, Pearl River, New York 10965, USA*
[e] *Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, CA 92130, USA*
[f] *Tianjin Institute of Bioinformatics and Drug Discovery (TIBDD), Tianjin, China*
[g] *Shanghai Jiaotong University, Life Science Research Center, Shanghai 200030, China*

Received 13 October 2003; received in revised form 9 January 2004; accepted 13 February 2004

Abstract

The core of an enzyme molecule is its active site from the viewpoints of both academic research and industrial application. To reveal the structural and functional mechanism of an enzyme, one needs to know its active site; to conduct structure-based drug design by regulating the function of an enzyme, one needs to know the active site and its microenvironment as well. Given the atomic coordinates of an enzyme molecule, how can we predict its active site? To tackle such a problem, a distance group approach was proposed and the support vector machine algorithm applied to predict the catalytic triad of serine hydrolase family. The success rate by jackknife test for the 139 serine hydrolases was 85%, implying that the method is quite promising and may become a useful tool in structural bioinformatics.

Keywords: Distance-group; Support vector machine; Catalytic triad; Serine hydrolase; Structural bioinformatics

1. Introduction

After the structure of a protein has been solved, the next challenge is to extract functional information from its structural data (see, e.g., Chou et al., 1998, 1999a; Chou, 1992, 2004; Zhou and Troy, 2003). With the increasing number of protein 3D structures having been solved by X-ray crystallography and NMR spectroscopy techniques, it is highly desirable to develop a high throughput tool to identify their active sites, based on the structural information. For many years, it has been suggested that the active sites in proteins are straightforwardly related to their biochemical function and are better conserved during evolution than the other part of proteins (Chou and Howe, 2002; Chou et al., 1997, 2000; Zvelebil et al., 1987). Through the similarity searching of active sites, one could even identify proteins with the same function but almost without related sequences or global folds (Chou et al., 1998, 1999b; Zhang et al., 2002). Hence, the identification of protein active sites is the key element for the assignment of biochemical function to a new protein.

To detect an active site in proteins whose 3D structures are available, a straightforward approach is to search the structural analogues of the known active sites. Currently, there exist many protein structure comparison approaches based on the geometric hashing algorithm (Nussinov and Wolfson, 1991) or the graph-theoretic algorithm (Artymiuk et al., 1994) that can match the user-defined structure against a whole protein structure (Fischer et al., 1994; Wallace et al., 1996). With the enzyme active site templates provided by PROCAT database (Wallace et al., 1997) (http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html), one can use these approaches to compare the available active site templates with a target protein so as to deduce the possible active site for the query protein. However, the speed of structural comparison process is usually

*Corresponding author: Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, UK.
E-mail address: y.cai@umist.ac.uk (Yu-dong Cai).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.02.019

very slow. The present study was initiated in an attempt to develop a different approach that can be used for fast identification of active sites.

2. Materials and methods

In order to improve the speed for the active site identification and meanwhile keep the high success rate, the support vector machines (SVMs) approach has been introduced to detect the active site of proteins. Support vector machine (SVM) is a class of learning machines based on the statistical learning theory. The basic idea of applying SVM to pattern classification can be outlined as follows. First, map the input vectors into one feature space (possible with a higher dimension), either linearly or nonlinearly, which is relevant to the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e. construct a hyper-plane that separates the samples into two classes (this can be extended to multi-class). SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is given in the book by Vapnik (1998). SVMs have been used to deal with protein fold recognition (Ding and Dubchak, 2001), protein–protein interactions prediction (Bock and Gough, 2001), protein secondary structure prediction (Hua and Sun, 2001), protein subcellular location prediction (Chou and Cai, 2002), and membrane protein-type prediction (Cai et al., 2003). A brief introduction about SVMs and some relevant key equations can be found in some recent papers (see, e.g., Cai et al., 2003; Chou and Cai, 2002), and hence there is no need to repeat here.

The 139 serine hydrolase entries in enzyme class E.C. 3.4.21 were selected from Enzyme Database (Bairoch, 2000), and their atomic coordinates extracted from the corresponding PDB files. The catalytic triad of hydrolase family is formed by Ser, His, and Asp. Accordingly, the catalytic triad of hydrolase is characterized by the special positions of the three key residues, particularly by the 10 pair-wise distances between the following atoms: O^{γ} of Ser, N^{δ} and N^{ε} of His, and $O^{\delta 1}$ and $O^{\delta 2}$ of Asp (Fig. 1). Their values form a distance group to characterize the relative spatial position among the Ser, His, and Asp residues. Since the dihedral angles concerned are constrained by the 10 pair-wise distances (Chou et al., 1982), the effects of angles are automatically included in the distance group. For each of the 139 proteins, we can find all such distance groups with respect to different Ser, His, and Asp along the protein chain. Those distance groups that correspond to a virtual active site are called the "active distance group", while all the others the "non-active distance group".

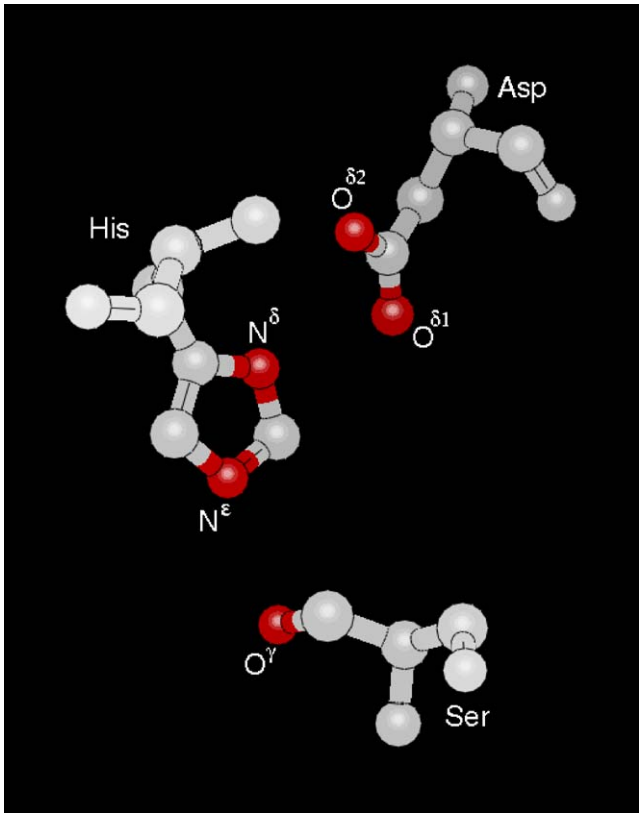

Fig. 1. The active sites (catalytic triad) of the serine hydrolase family are characterized by His–Asp–Ser, particularly the following five atoms (in red): N^{δ} and N^{ε} of His, $O^{\delta 1}$ and $O^{\delta 2}$ of Asp, as well as O^{γ} of Ser.

Thus, the following two data sets can be constructed: the positive data set S^{+} containing only the active distance groups, and the negative data set S^{-} with the negative distance groups only. The data in S^{+} and S^{-} can serve as the positive and negative benchmarks (Chou, 1993, 1996) to train the SVMs for the active site prediction.

3. Results and discussion

The computations were carried out on a Silicon Graphics IRIS Indigo workstation (Elan 4000). In the current research, the width of the Gaussian RBFs for the SVM was so selected as to minimize an estimate of the VC-dimension (Chou and Cai, 2002). The parameter C that controlled the error-margin trade-off was set at 1000 (Cai et al., 2003). After being trained, the hyper-plane output by the SVM was obtained, indicating that the trained model, i.e. the hyper-plane output, had contained the important information for identifying the catalytic triad.

Suppose n_i^{+} represents the number of active sites in the i-th protein, for which the number of distance groups investigated is n_i. From these groups, the number of correctly predicted active sites is m_i^{+} and that of

incorrectly predicted active sites is m_i^-. For hydrolase family, each enzyme has one active site, i.e. $n_i^+ = 1$. If $m_i^+ = 1$ and $m_i^- = 1$, i.e. the number of correctly predicted active sites for the i-th protein is 1 but the corresponding incorrectly predicted number is also 1, then the success prediction rate for the i-th protein would be $m_i^+/(n_i^+ + m_i^-) = \frac{1}{2}$; if $m_i^+ = 1$ and $m_i^- = 2$, then the corresponding rate would drop to $\frac{1}{3}$; and so forth. Accordingly, the overall success rate of prediction can be formulated by the following equation:

$$\lambda = \frac{\sum_{i=1}^{N}(m_i^+/(n_i^+ + m_i^-))}{\sum_{i=1}^{N} n_i^+}, \quad (1)$$

where N is the total number of proteins investigated. When each protein contains only one active site, the above equation can be further reduced to

$$\lambda = \frac{\sum_{i=1}^{N}(m_i^+/(1 + m_i^-))}{N}. \quad (2)$$

As is well known, the independent data set test, subsampling test and jackknife test are the three methods often used for cross-validation to examine the prediction quality in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one; see, e.g., Chou and Zhang (1995) for a comprehensive discussion about this, and Mardia et al. (1979) for the mathematical principle. Jackknife test is particularly useful for checking the cluster-tolerant capacity (Chou, 1999), and hence was often used for the case when the training data sets were far from complete yet (see, e.g., Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). During jackknifing, each protein in the data set is in turn singled out as a tested protein and all the rule-parameters are calculated based on the remaining proteins. The predicted results thus obtained are listed in Table 1. Substituting the data of columns 3 and 4 into Eq. (2), we obtained the overall success rate $\lambda = 118/139 = 84.9\%$. Also, in an average it took about 10 s of CPU time to identify the catalytic triad for a protein, much faster than the approach by the database search. Therefore, the current approach has provided an

Table 1
Predicted results for the active sites of the 139 proteins by the jackknife test

Index (i)	Protein code	Number of correctly predicted active sites (m_i^+)	Number of incorrectly predicted active sites (m_i^-)	Number of correctly predicted non-active sites	Number of distance groups investigated (n_i)
1	1A10	1	1	2337	2339
2	1A1Y	1	0	2518	2519
3	1A46	1	1	1437	1439
4	1A4W7	1	0	1278	1279
5	1A5G	1	0	1518	1519
6	1A61	1	0	1518	1519
7	1AB9	1	0	466	467
8	1ABI	1	0	1698	1699
9	1ABJ	1	0	1438	1439
10	1AD8	1	0	1518	1519
11	1AF4	1	1	1437	1439
12	1AFQ	1	0	484	485
13	1AHT	1	0	1518	1519
14	1AUT	1	0	6382	6383
15	1AY6	1	1	1517	1519
16	1B5G	1	1	1597	1599
17	1BA8	1	0	1528	1529
18	1BCU	1	0	1518	1519
19	1BH6	1	0	1678	1679
20	1BMA	1	0	922	923
21	1BRA	1	0	814	815
22	1BRB	1	0	862	863
23	1BRC	1	0	1062	1063
24	1BTU	1	1	921	923
25	1BTW	0	1	642	644
26	1BTX	0	0	793	794
27	1BTY	0	0	793	794
28	1BTZ	0	0	748	749
29	1C4U	1	0	1528	1529
30	1C4V	1	0	1698	1699
31	1C4Y	1	1	1517	1519
32	1CA8	1	1	1197	1199
33	1CSE	1	0	2290	2991

Table 1 (*continued*)

Index (i)	Protein code	Number of correctly predicted active sites (m_i^+)	Number of incorrectly predicted active sites (m_i^-)	Number of correctly predicted non-active sites	Number of distance groups investigated (n_i)
34	1DIT	1	1	1597	1599
35	1DWB	1	0	1598	1599
36	1DWC	1	0	1598	1599
37	1DWD	1	0	1598	1599
38	1DWE	1	0	1438	1439
39	1ELA	1	1	921	923
40	1ELB	1	1	921	923
41	1ELC	1	1	921	923
42	1ELD	1	0	922	923
43	1ELE	1	0	922	923
44	1ELF	1	0	922	923
45	1ELG	1	0	922	923
46	1ELV	1	0	1222	1223
47	1ESA	1	0	922	923
48	1ESB	1	1	921	923
49	1FAX	1	0	1258	1259
50	1FPC	1	0	1438	1439
51	1GCD	1	0	466	467
52	1GCT	1	0	466	467
53	1GMH	1	0	448	449
54	1H4W	1	0	1048	1049
55	1HAG	1	0	1698	1699
56	1HAH	1	0	1698	1699
57	1HAI	1	0	1613	1614
58	1HAO	1	0	1438	1439
59	1HAP	1	0	1438	1439
60	1HAX	1	0	922	923
61	1HAZ	1	1	921	923
62	1HB0	1	0	922	923
63	1HCG	1	0	1258	1259
64	1HGT	1	0	1698	1699
65	1HJA	1	0	988	989
66	1HPG	1	0	638	639
67	1HUT	1	0	1613	1614
68	1HXE	1	0	1518	1519
69	1HXF	1	0	1518	1519
70	1IHS	1	0	1438	1439
71	1IHT	1	0	1438	1439
72	1K1I	0	0	610	611
73	1K22	1	1	1597	1599
74	1MEE	1	0	4057	4058
75	1NRN	1	0	1698	1699
76	1NRO	0	0	1518	1519
77	1NRP	1	0	1698	1699
78	1NRQ	0	0	1358	1359
79	1NRR	1	0	1438	1439
80	1NRS	1	0	1598	1599
81	1PEK	1	0	1922	1923
82	1QGF	0	0	922	923
83	1QIX	1	0	922	923
84	1QJ1	1	0	1698	1699
85	1QL9	1	0	814	815
86	1S02	1	0	2218	2219
87	1SBC	1	0	1438	1439
88	1SBH	1	0	2506	2507
89	1SBI	1	0	2506	2507
90	1SBN	1	0	4210	4211
91	1SCA	1	0	1483	1484
92	1SCB	1	0	1438	1439
93	1SCD	1	0	1438	1439
94	1SCN	1	0	1438	1439
95	1SGP	1	0	673	674

Table 1 (*continued*)

Index (i)	Protein code	Number of correctly predicted active sites (m_i^+)	Number of incorrectly predicted active sites (m_i^-)	Number of correctly predicted non-active sites	Number of distance groups investigated (n_i)
96	1SGQ	1	0	673	674
97	1SGR	1	1	672	674
98	1SGT	1	0	124	125
99	1SIB	1	0	4561	4562
100	1SUE	1	0	2218	2219
101	1TBZ	1	0	1438	1439
102	1TGN	1	0	610	611
103	1THM	1	0	1558	1559
104	1THR	1	0	1888	1889
105	1THS	1	0	1613	1614
106	1TMB	1	0	1518	1519
107	1TMT	1	0	1438	1439
108	1TMU	1	0	1518	1519
109	1TOM	1	0	1518	1519
110	1TRY	1	1	627	629
111	1YJA	1	1	2505	2507
112	1YYY	1	0	610	611
113	2EST	1	0	922	923
114	2GCH	1	0	466	467
115	2GCT	1	0	466	467
116	2GMT	1	0	466	467
117	2HAT	1	1	1612	1614
118	2HGT	1	0	1698	1699
119	2HNT	1	0	1518	1519
120	2HPP	1	0	2728	2729
121	2HPQ	1	0	3190	3191
122	2SEC	1	0	2990	2991
123	2SGP	1	0	700	701
124	2SIC	1	0	5398	5399
125	2SNI	1	0	3190	3191
126	2TGD	0	1	609	611
127	3GCT	1	1	465	467
128	3HAT	1	0	1613	1614
129	3SIC	1	0	5398	5399
130	3TGI	1	0	1006	1007
131	3TGK	1	1	829	831
132	4HTC	1	1	2277	2279
133	5GDS	1	0	1698	1699
134	5SIC	1	1	5397	5399
135	6EST	1	0	922	923
136	7EST	1	0	922	923
137	7KME	1	0	1438	1439
138	8GCH	1	1	501	503
139	8KME	1	0	1438	1439

accurate and fast method for predicting the active sites of enzymes.

It is instructive to point out that, many of the failed-to-predict catalytic triads bear patterns that are far away from the putative standard one for the catalytic triad as shown in panel (a) of Fig. 2. These failed-to-predict catalytic triads contain some internal distance(s) greater than 15 Å, as shown in panels (b) of Fig. 2. According to the common sense in biochemistry it is highly unlikely for a catalytic triad to have such a large internal distance. The problem with unreasonably large internal distances for a catalytic triad might be due to some typographic error in annotating the His–Asp–Ser sequence positions. For illustration, let us take 1K1I as an example (Fig. 2b). According to the annotation of its PDB file, the catalytic triad is formed by His-40, Asp-102, and Ser-195. Based on such an annotation, of its 10 pair-wise distances, 5 are greater 10 Å (with 2 greater than 15 Å). Obviously, it is impossible for a structure with these internal distances to function as a catalytic triad. However, for the same PDB file, if only the His-40 was changed to His-57, most of the 10 pair-wise distances would be within the range of 2 Å and 5 Å and only one around 8 Å, suggesting that the structure

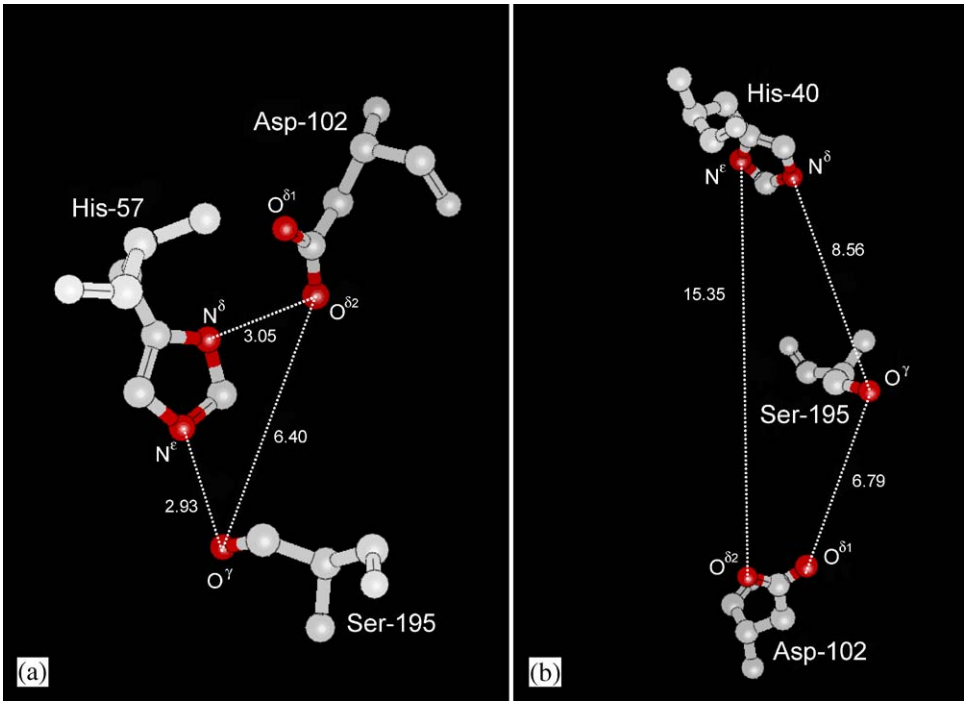

Fig. 2. Illustration to show the ten pair-wise internal distances for the His–Asp–Ser catalytic triad in (a) 1FAX (having a putative normal pattern with no internal distance greater than 10 Å), and (b) 1K1I (having a problematic pattern with some distance(s) greater than15 Å). For simplicity, only three of the ten internal pair-wise distances are shown.

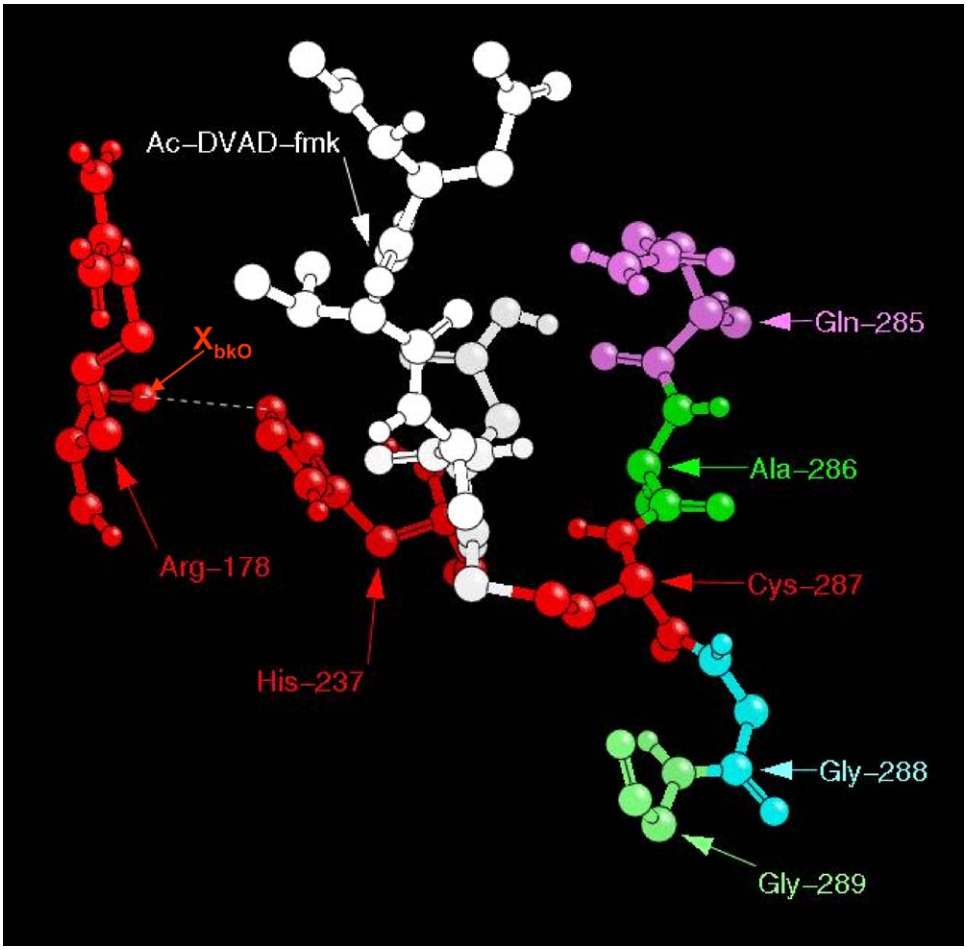

Fig. 3. Illustration to show the Cys–His–X_{bkO} catalytic triad (highlighted in red) in caspase family, where X_{bkO} represents the backbone carbonyl oxygen. Reproduced with permission from Chou et al. (2000). For further explanation, see Chou et al. (2000).

thus obtained would have the standard putative pattern for the His–Asp–Ser catalytic triad.

Finally, it should be pointed out that the catalytic triad for different enzyme family may be formed by different key residues and atoms. For example, the catalytic triad for caspase family has a catalytic Cys–His–X_{bkO} triad mechanism (Fig. 3), where X_{bkO} represents the backbone carbonyl oxygen of any residue at the position of the 3rd component of the triad, irrespective of the nature of the amino acid concerned: for caspase-1, it is Pro-177; for caspase-3, Thr-62; for caspase-8, Arg-258; and for capase-9, Arg-178 (Chou et al., 2000). Therefore, different key residues and atoms should be used to predict the catalytic triad for different enzyme families.

References

Artymiuk, P.J., Poirrette, A.R., Grindley, H.M., Rice, D.W., Willett, P., 1994. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J. Mol. Biol. 243, 327–344.

Bairoch, A., 2000. The ENZYME database in 2000. Nucleic Acids Res. 28, 304–305.

Bock, J.R., Gough, D.A., 2001. Predicting protein–protein interactions from primary structure. Bioinformatics 17, 455–460.

Cai, Y.D., Zhou, G.P., Chou, K.C., 2003. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys. J. 84, 3257–3263.

Chou, J.J., Li, H., Salvessen, G.S., Yuan, J., Wagner, G., 1999a. Solution structure of BID, an intracellular amplifier of apoptotic signalling. Cell 96, 615–624.

Chou, J.J., Matsuo, H., Duan, H., Wagner, G., 1998. Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. Cell 94, 171–180.

Chou, K.C., 1992. Energy-optimized structure of antifreeze protein and its binding mechanism. J. Mol. Biol. 223, 509–517.

Chou, K.C., 1993. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J. Biol. Chem. 268, 16938–16948.

Chou, K.C., 1996. Review: prediction of HIV protease cleavage sites in proteins. Anal. Biochem. 233, 1–14.

Chou, K.C., 1999. A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun. 264, 216–224.

Chou, K.C., 2004. Modelling extracellular domains of GABA-A receptors: subtypes 1,2,3, and 5. Biochem. Biophys. Res. Commun. 316, 636–642.

Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. 277, 45765–45769.

Chou, K.C., Howe, W.J., 2002. Prediction of the tertiary structure of the beta-secretase zymogen. Biochem. Biophys. Res. Commun. 292, 702–708.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Chou, K.C., Jones, D., Heinrikson, R.L., 1997. Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Lett. 419, 49–54.

Chou, K.C., Pottle, M., Nemethy, G., Ueda, Y., Scheraga, H.A., 1982. Structure of beta-sheets: Origin of the right-handed twist and of the increased stability of antiparallel over parallel sheets. J. Mol. Biol. 162, 89–112.

Chou, K.C., Watenpaugh, K.D., Heinrikson, R.L., 1999b. A model of the complex between cyclin-dependent kinase 5(Cdk5) and the activation domain of neuronal Cdk5 activator. Biochem. Biophys. Res. Commun. 259, 420–428.

Chou, K.C., Tomasselli, A.G., Heinrikson, R.L., 2000. Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Lett. 470, 249–256.

Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.

Fischer, D., Wolfson, H., Lin, S.L., Nussinov, R., 1994. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. Protein Sci. 3, 769–778.

Hua, S.J., Sun, Z.R., 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308, 397–407.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. In: *Multivariate Analysis*, Academic Press, London, pp. 322–381.

Nussinov, R., Wolfson, H.J., 1991. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc. Natl. Acad. Sci. USA 88, 10495–10499.

Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.

Wallace, A.C., Laskowski, R.A., Thornton, J.M., 1996. Derivation of 3D coordinate templates for searching structural databases: application to Ser–His–Asp catalytic triads in the serine proteinases and lipases. Protein Sci. 5, 1001–1013.

Wallace, A.C., Borkakoti, N., Thornton, J.M., 1997. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci. 6, 2308–2323.

Zhang, J., Luan, C.H., Chou, K.C., Johnson, G.V.W., 2002. Identification of the N-terminal functional domains of Cdk5 by molecular truncation and computer modeling. Proteins: Struct. Funct. Genetics 48, 447–453.

Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.

Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins: Struct. Funct. Genetics 44, 57–59.

Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct. Funct. Genetics 50, 44–48.

Zhou, G.P., Troy, F.A., 2003. Characterization by NMR and molecular modeling of the binding of polyisoprenols and polyisoprenyl recognition sequence peptides: 3D structure of the complexes reveals site of specific interactions. Glycobiology 13, 51–71.

Zvelebil, M.J., Barton, G.J., Taylor, W.R., Sternberg, M.J., 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. 195, 957–961.