



ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information

Kuo-Chen Chou*, Hong-Bin Shen*

Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai, 200240, China
Gordon Life Science Institute, Bioinformatics and Drug Delivery, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

ARTICLE INFO

Article history:

Received 20 August 2008

Available online 5 September 2008

Keywords:

Protease type
 Evolution
 Functional domain
 Fusion approach
 OET-KNN
 ProtIdent
 Web-server

ABSTRACT

Proteases are vitally important to life cycles and have become a main target in drug development. According to their action mechanisms, proteases are classified into six types: (1) aspartic, (2) cysteine, (3) glutamic, (4) metallo, (5) serine, and (6) threonine. Given the sequence of an uncharacterized protein, can we identify whether it is a protease or non-protease? If it is, what type does it belong to? To address these problems, a 2-layer predictor, called "ProtIdent", is developed by fusing the functional domain and sequential evolution information: the first layer is for identifying the query protein as protease or non-protease; if it is a protease, the process will automatically go to the second layer to further identify it among the six types. The overall success rates in both cases by rigorous cross-validation tests were higher than 92%. ProtIdent is freely accessible to the public as a web server at <http://www.csbio.sjtu.edu.cn/bioinf/Protease>.

© 2008 Elsevier Inc. All rights reserved.

Proteases, also termed proteinases or peptidases [1], are proteolytic enzymes. They are biomolecular version of "Swiss army knives" cutting long amino acid sequences into fragments [2], which is essential for the synthesis of all proteins, controlling their size, composition, shape, turnover, and ultimate destruction.

Proteases account for about 2% of the human genome and 1–5% of genomes of infectious organisms [3]. Actually, according to the recent inference by Rawlings et al. [4], the number of proteases might be at least twice as much. Regulating most physiological processes by controlling the activation, synthesis, and turnover of proteins, proteases play pivotal regulatory roles in conception, birth, digestion, growth, maturation, aging, and even death of all organisms (see, e.g., [5–11]). Proteases are also essential in viruses, bacteria, and parasites for their replication and the spread of infectious diseases, in all insects, organisms, and animals for effective transmission of disease, and in human and animal hosts for the mediation and sustenance of diseases. Because of their important

regulatory roles in life cycle, proteases represent important potential targets for medical intervention.

The actions of proteases are exquisitely selective (see, e.g., [12–16]), with each protease being responsible for splitting very specific sequences of amino acids under a preferred set of environmental conditions.

According to their catalytic mechanisms, proteinases are classified into the following six types: (1) aspartic, (2) cysteine, (3) glutamic, (4) metallo, (5) serine, and (6) threonine [4]. Different types of proteases have different action mechanisms and biological processes.

Therefore, it is important for both basic research and drug discovery to consider the following two problems. Given the sequence of a protein, can we identify whether it is a protease or non-protease? If it is, what protease type does it belong to? Although the answers to the above two questions can be found through biochemical experiments, the approach by purely doing experiments is both time-consuming and costly. Particularly, the number of newly-found protein sequences has increased explosively in the Post Genomic Age. For example, in 1986 the SWISS-PROT databank [17] contained only 3939 entries of protein sequences; recently, the number jumped to 392,667 according to the version 56.0 released on 22-July-2008 at <http://www.ebi.ac.uk/swissprot/>, meaning that the number of the entries now is more than 99 times the number of 1986! Facing such an avalanche of protein sequences, the challenge to address these questions has become even more critical and urgent.

Abbreviations: FunD, functional domain; PSSM, position-specific scoring matrix; PsePSSM, pseudo position-specific scoring matrix; OET-KNN, optimized evidence-theoretic K nearest neighbor.

* Corresponding authors. Addresses: Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA (K.-C. Chou); Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 800 Dongchun Road, Shanghai 200240, China (H.-B. Shen). Fax: +1 858 380 4623 (K.-C. Chou).

E-mail addresses: kcchou@gordonlifescience.org (K.-C. Chou), hbshen@sjtu.edu.cn (H.-B. Shen).

Actually, some efforts have been made in this regard [18,19]. However, the topic is worthy of further investigation due to the following reasons. First, none of the methods developed in [18,19] provided a web server that can be easily used by the majority of experimental and pharmaceutical scientists to obtain the desired data. Second, with more data entering into data bank, the benchmark dataset used to train and test the predictor needs to be updated. Third, none of the aforementioned methods took into account the sequential evolution information, which may play an important role in identifying protease types. The present study was initiated in an attempt to reconsider this topic from the above three points.

Materials

To develop a powerful statistical predictor, the first important thing is to construct a high quality benchmark dataset [20]. To realize this, the data were taken from the “Peptidase Protein Sequences” in the MEROPS database [4] at <http://merops.sanger.ac.uk/> (version 8.1, released on 05-May-2008) and screened strictly according to the following procedures. (1) To avoid fragment data, those proteins whose sequences were annotated with “fragment” or had less than 50 amino acids were excluded. (2) Sequences which contain two or more consecutive uncertain residues (i.e., “XX”, “XXX”, and so forth) were removed. (3) To reduce the homology bias, a redundancy cutoff was operated by an in-house program to winnow those sequences which have $\geq 25\%$ pairwise sequence identity to any other in a same subset or type. Thus, a total of 3051 protease sequences were collected that consist of 258 aspartic proteases, 589 cysteine, 39 glutamic, 1040 metallo, 1063 serine, and 62 threonine.

Meanwhile, by following the same screening procedures, a total of 3278 non-protease protein sequences were randomly taken from the SWISS-PROT databank (version 55.3 released on 29-April-2008) at <http://www.ebi.ac.uk/swissprot/>.

The 3051 protease sequences classified into six subsets and the 3278 non-protease protein sequences are provided in the [Online supporting information A](#) and [Online supporting information B](#), respectively, which constitute the benchmark dataset for the current study.

Methods

Once the benchmark dataset is established, the subsequent problem is how to find an effective prediction engine and use what kind of descriptor to represent the protein samples for training the engine and conducting the prediction.

For the convenience of later formulation, let us suppose the benchmark dataset constructed in the above section is denoted by \mathbb{S} , which consists of the protease dataset \mathbb{S}^+ and the non-protease dataset \mathbb{S}^- ; i.e.,

$$\begin{cases} \mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \\ \mathbb{S}^+ = \mathbb{S}_1^+ \cup \mathbb{S}_2^+ \cup \mathbb{S}_3^+ \cup \mathbb{S}_4^+ \cup \mathbb{S}_5^+ \cup \mathbb{S}_6^+ \end{cases} \quad (1)$$

where \cup is the symbol for the union in the set theory, \mathbb{S}_1^+ the subset containing the aspartic proteases only, \mathbb{S}_2^+ the subset containing the cysteine proteases only, and so forth.

Now, the problem can be formulated as

$$\mathbb{E} \triangleright \mathbf{P} = \mathbf{C} \in \begin{cases} \mathbb{S}^+ \cup \mathbb{S}^-, & \text{if among protease and non-protease} \\ \mathbb{S}^+, & \text{if among six protease types} \end{cases} \quad (2)$$

where \mathbb{E} represents the prediction engine, \mathbf{P} the query protein, \triangleright is an identification operator, \mathbf{C} the predicted result, \in is a symbol in the set theory meaning “member of”, and \mathbb{S} is defined by Eq. (1). Before the prediction engine can be used, it must be trained by a train-

ing dataset where all the proteins must have the same descriptor as that of the query protein \mathbf{P} .

In the area of predicting protein attributes, two kinds of descriptors are often used to represent protein samples. One is the sequential model, and the other the discrete model. In the sequential model, the sample of a protein is represented by its amino acid sequence, and the sequence similarity search-based tools such as BLAST [21] are used to conduct prediction. However, this approach failed to work when a query protein did not have significant homology to character-known proteins. Thus, various discrete models were introduced by representing the sample of a protein with a set of discrete numbers. The simplest discrete model is to represent the sample of a protein with its amino acid (AA) composition or AAC (see, e.g., [22]). However, in the AAC model, all the sequence-order effects are lost. To avoid completely lose the sequence-order information, the pseudo amino acid (PseAA) composition or PseAAC was introduced [23]. The PseAAC model can be used to represent a protein sequence with a discrete model yet without completely losing its sequence-order information, and have been widely used to deal with varieties of problems in proteins and protein-related systems (see, e.g., [24–26]).

Here, we shall introduce a novel discrete model to represent protein samples by fusing the functional domain information and sequential evolutionary information.

Functional domain (FunD) composition

Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. Based on such a fact, several FunD databases were developed, such as SMART [27], Pfam [28], COG [29], KOG [29], and CDD [30]. CDD database defines conserved domains based on recurring sequence patterns or motifs and it contains domains imported from SMART, Pfam and COG databases. Therefore, CDD is a much more complete domain database [30]; the version 2.11 of CDD contains 17,402 common protein domains and families. With each of the 17,402 domain sequences as a vector-base [31], a given protein sample can be defined as a 17402-D (dimensional) vector according to the following procedures. (1) Use RPS-BLAST (Reverse PSI-BLAST) program [32] to compare the protein sequence with each of the 17,402 domain sequences in the CDD database. (2) If the significance threshold value (expect value) is ≤ 0.001 for the i -th profile meaning a “hit” is found, then the i -th component of the protein in the 17402-D space is assigned 1; otherwise, 0. (3) The protein sample \mathbf{P} in the FunD space can thus be formulated as

$$\mathbf{P}_{\text{FunD}} = [\mathbb{D}_1 \quad \mathbb{D}_2 \quad \cdots \quad \mathbb{D}_i \quad \cdots \quad \mathbb{D}_{17402}]^T \quad (3)$$

where \mathbf{T} is the transpose operator, and

$$\mathbb{D}_i = \begin{cases} 1, & \text{when a hit is found for } \mathbf{P} \text{ in CDD database} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Pseudo position-specific scoring matrix (PsePSSM)

To incorporate the evolution information of proteins, the PSSM (position-specific scoring matrix) [32] was used; i.e., according to the concept of PSSM, the sample of a protein \mathbf{P} can be represented by:

$$\mathbf{P}_{\text{PSSM}} = \begin{bmatrix} \mathbb{R}_{1 \rightarrow 1} & \mathbb{R}_{1 \rightarrow 2} & \cdots & \mathbb{R}_{1 \rightarrow 20} \\ \mathbb{R}_{2 \rightarrow 1} & \mathbb{R}_{2 \rightarrow 2} & \cdots & \mathbb{R}_{2 \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{R}_{i \rightarrow 1} & \mathbb{R}_{i \rightarrow 2} & \vdots & \mathbb{R}_{i \rightarrow 20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{R}_{L \rightarrow 1} & \mathbb{R}_{L \rightarrow 2} & \cdots & \mathbb{R}_{L \rightarrow 20} \end{bmatrix} \quad (5)$$

where \mathbb{R}_{i-j} represents the score of the amino acid residue in the i -th position of the protein sequence being changed to amino acid type j during the evolution process. Here, the numerical codes 1, 2, ..., 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes. The $L \times 20$ scores in Eq. (5) were generated by using PSI-BLAST [32] to search the SWISS-PROT database (version 52.0 released on 6-March-2007) through three iterations with 0.001 as the E -value cutoff for multiple sequence alignment against the sequence of the protein \mathbf{P} , followed by a standard conversion given below:

$$\mathbb{R}_{i-j} = \frac{\mathbb{R}_{i-j}^0 - \langle \mathbb{R}_i^0 \rangle}{SD(\mathbb{R}_i^0)} \quad (i = 1, 2, \dots, L; j = 1, 2, \dots, 20) \quad (6)$$

where \mathbb{R}_{i-j}^0 represent the original scores directly created by PSI-BLAST [32] that are generally shown as positive or negative integers, the symbol $\langle \rangle$ means taking the average of the quantity therein over 20 native amino acids, and SD means the corresponding standard deviation. The converted values obtained by Eq. (6) will have a zero mean value over the 20 amino acids and will remain unchanged if going through the same conversion procedure again. The positive score means that the corresponding mutation occurs more frequently in the alignment than expected by chance, while the negative one means just the opposite. However, according to the PSSM descriptor (Eq. (5)), proteins with different lengths will correspond to row-different matrices. To make the PSSM descriptor become a size-uniform matrix, one possible approach is to represent a protein sample \mathbf{P} by

$$\bar{\mathbf{P}}_{\text{PSSM}} = [\bar{\mathbb{R}}_1 \quad \bar{\mathbb{R}}_2 \quad \dots \quad \bar{\mathbb{R}}_{20}]^T \quad (7)$$

where

$$\bar{\mathbb{R}}_j = \frac{1}{L} \sum_{i=1}^L \mathbb{R}_{i-j} \quad (j = 1, 2, \dots, 20) \quad (8)$$

where $\bar{\mathbb{R}}_j$ represents the average score of the amino acid residues in the protein \mathbf{P} being changed to amino acid type j during the evolution process. However, if $\bar{\mathbf{P}}_{\text{PSSM}}$ of Eq. (7) was used to represent the protein \mathbf{P} , all the sequence-order information during the evolution process would be lost. To avoid complete loss of the sequence-order information, the concept of the pseudo amino acid composition (PseAAC) as originally proposed in [23] was adopted; i.e., instead of Eq. (7), let us use the pseudo position-specific scoring matrix (PsePSSM) as given by

$$\mathbf{P}_{\text{PsePSSM}}^\lambda = [\bar{\mathbb{R}}_1 \quad \bar{\mathbb{R}}_2 \quad \dots \quad \bar{\mathbb{R}}_{20} \quad \mathbb{R}_1^\lambda \quad \mathbb{R}_2^\lambda \quad \dots \quad \mathbb{R}_{20}^\lambda]^T \quad (9)$$

to represent the protein \mathbf{P} , where

$$\mathbb{R}_j^\lambda = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} [\mathbb{R}_{i-j} - \mathbb{R}_{(i+\lambda)-j}]^2 \quad (j = 1, 2, \dots, 20; \lambda < L) \quad (10)$$

meaning that \mathbb{R}_j^1 is the correlation factor by coupling the most contiguous PSSM scores along the protein chain for the amino acid type j ; \mathbb{R}_j^2 that by coupling the second-most contiguous PSSM scores; and so forth. Note that, as mentioned in the Material section, the length of the shortest protein sequence in the benchmark dataset is $L = 50$, and hence the value allowed for λ in Eq. (10) must be smaller than 50. When $\lambda = 0$, \mathbb{R}_j^λ becomes a naught element and Eq. (9) is degenerated to Eq. (7).

Prediction engine and fusion approach

Once the descriptor for protein samples is set up, we need a prediction engine to operate the prediction. The prediction engine in the previous work [18,19] was based on the NN (Nearest Neighbor) algorithm [33]. In this study, the OET-KNN (optimized evidence-theoretic K nearest neighbor) classifier [34] was utilized to identify

the proteases and their types. The OET-KNN classifier is a very powerful classification engine as demonstrated by its role in enhancing the success rates of protein subcellular localization [35], where a detailed mathematical formulation for OET-KNN was also provided in its Appendix B. Below, we shall give a brief description of how to use it to identify protease and its type.

Let us first consider the top level problem, i.e., how to identify a protein as protease or non-protease based on the benchmark dataset $\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^-$ (Eq. (1)). Suppose $\mu = 1$ represents the protease class, and $\mu = 2$ the non-protease class. Thus, according to Eq. (2), it follows

$$\begin{aligned} \mathbb{E} \triangleright \mathbf{P} &= \text{OET-KNN} \triangleright \mathbf{P} \\ &= \begin{cases} \text{OET-KNN} \triangleright \mathbf{P}_{\text{FunD}} = \Gamma_1(K, \mu), & \text{for FunD frame} \\ \text{OET-KNN} \triangleright \mathbf{P}_{\text{PsePSSM}}^\lambda = \Gamma_2(K, \lambda, \mu), & \text{for PsePSSM} \end{cases} \quad (\mu = 1, 2) \end{aligned} \quad (11)$$

where $\Gamma_1(K, \mu)$ is the creditability score for the query protein believed in the μ -th class when it is defined in the FunD frame (Eq. (3)), K the parameter selected for the OET-KNN classifier [35], $\Gamma_2(K, \lambda, \mu)$ the corresponding creditability score when the prediction is operated in the PsePSSM frame (Eq. (9)), and λ the parameter selected for defining $\mathbf{P}_{\text{PsePSSM}}^\lambda$ (Eqs. (9) and (10)). Accordingly, using different descriptors to represent protein samples may lead to different results; even if the same descriptor is adopted, selecting different parameters may lead to different results as well. In order to get a unique result, the fusion approach is introduced as formulated below.

The parameter K in Eq. (11) is the number of the nearest proteases counted against the query protein during the prediction process [35]. Generally speaking, for most training datasets, when $K > 10$ the success rate drops down remarkably and hence we can narrow the scope of K from 1 to 10. Also, the parameter λ must be smaller than 50 (see Eq. (10)), the number of amino acids for the shortest protein sequence in the benchmark dataset. Therefore, the final predicted result should be determined by a fusion approach through a voting mechanism [20]. According to Eq. (11), the voting score for the query protein \mathbf{P} belonging to the μ -th class is given by

$$\Omega_\mu = \sum_{K=1}^{10} w_K^1 \Gamma_1(K, \mu) + \sum_{K=1}^{10} \sum_{\lambda=0}^{49} w_{K,\lambda}^2 \Gamma_2(K, \lambda, \mu), \quad (\mu = 1, 2) \quad (12)$$

where w_K^1 and $w_{K,\lambda}^2$ are the weight factors and were set at 1 for simplicity, thus the query protein \mathbf{P} is predicted belonging to the class or subset for which the score of Eq. (12) is the highest; i.e., the predicted class should be

$$m = \text{argmax}_\mu \{\Omega_\mu\}, \quad (\mu = 1, 2) \quad (13)$$

where m is the argument of μ that maximize Ω_μ . If there is a tie among the two classes, then the final predicted result will be randomly assigned among the two although this kind of tie case rarely happens and actually was not observed in the current study.

By changing $(\mu = 1, 2)$ to $(\mu = 1, 2, \dots, 6)$ and working on the benchmark dataset $\mathbb{S}^+ = \mathbb{S}_1^+ \cup \mathbb{S}_2^+ \cup \mathbb{S}_3^+ \cup \mathbb{S}_4^+ \cup \mathbb{S}_5^+ \cup \mathbb{S}_6^+$, Eqs. (11)–(13) can be automatically used to solve the next level problem, i.e. identify the proteases among their six different types.

The above fusion approach can not only incorporate both the functional domain information and the protein evolution information but also automatically solve the problem caused by the incompleteness of the FunD database. For example, if a query protein had no hit whatsoever when searching the CDD database [30], it would correspond to a naught vector according to Eq. (3). The creditability score for a naught vector is zero [35]; i.e., $\Gamma_1(K, \mu) = 0$ according to Eq. (11). Thus, the creditability score will be solely determined by $\Gamma_2(K, \lambda, \mu)$ derived from the PsePSSM frame.

The entire ensemble classifier thus established is called ProtIdent. To provide an intuitive picture, a flowchart to show how to

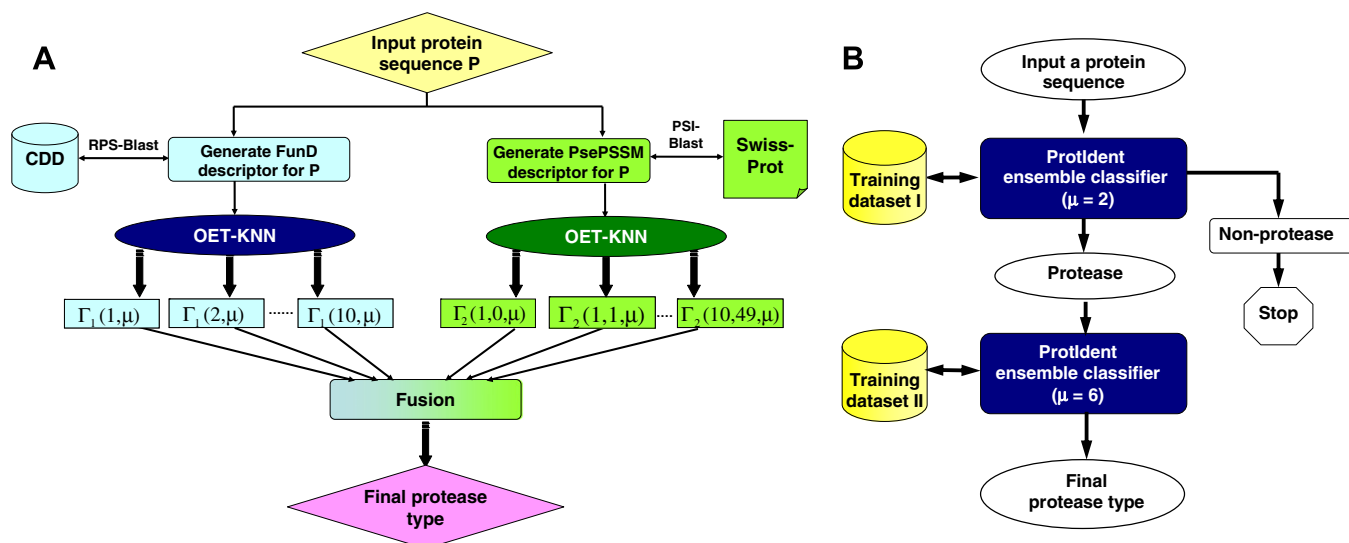


Fig. 1. A flowchart to show (A) how to fuse the FunD approach and PsePSSM approach, and (B) how the two-layer ProtIdent ensemble classifier works in identifying proteases and their functional types, where the training dataset I means $\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^-$, and the training dataset II means $\mathcal{S}^+ = \mathcal{S}_1^+ \cup \mathcal{S}_2^+ \cup \mathcal{S}_3^+ \cup \mathcal{S}_4^+ \cup \mathcal{S}_5^+ \cup \mathcal{S}_6^+$ (see Eq. (1)).

Table 1

Success rates by the jackknife cross-validation test in identifying proteins as proteases or non-proteases

Protein type	Number of proteins	Number of correct prediction	Success rate (%)
Protease	3051 ^a	2804	91.9
Non-protease	3278 ^b	3016	92.0
Overall	6329	5820	92.0

^a The sequences of the 3051 proteases are given in the Online supporting information A.

^b The sequences of the 3278 non-protease proteins are given in the Online supporting information B.

Table 2

Success rates by the jackknife cross-validation test in identifying proteases among their six functional types

Functional type	Number of proteases ^a	Number of correct prediction	Success rate (%)
Aspartic	258	224	86.8
Cysteine	589	572	97.1
Glutamic	39	38	97.4
Metallo	1040	1003	96.4
Serine	1063	1024	96.3
Threonine	62	60	96.3
Overall	3051	2921	95.7

^a The sequences of proteases for each of their six types are given in the Online supporting information A.

fuse the FunD approach and PsePSSM approach is given in Fig. 1A, and that to show the classification process of the predictor is given in Fig. 1B.

Results and discussion

In statistical prediction the independent dataset test, sub-sampling test, and jackknife test are often used in literatures for examining the accuracy of a predictor. However, as elucidated in [36] and demonstrated by Eq. (50) of [20], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark data-

set, and hence has been increasingly used by investigators to examine the accuracy of various predictors.

The jackknife cross-validation results by ProtIdent on the datasets \mathcal{S} and \mathcal{S}^+ (cf. Eq. (1) and Online supporting information A and Online supporting information B) are given in Tables 1 and 2, respectively, from which we can see that the overall success rate in identifying the proteins as proteases or non-proteases is 92.0%, and that the overall success rate in identifying the proteases among their six functional types is 96.0%.

Conclusions

The reason why ProtIdent predictor can yield so high success rates is because it operates by fusing the FunD approach and PsePSSM approach. The former is closely related to the functions of proteins, while the latter can incorporate their sequential evolution information. Moreover, the OET-KNN classifier is also very powerful in dealing with these kinds of problems. As a web server, ProtIdent is freely available to the public at <http://www.csbio.sjtu.edu.cn/bioinf/Protease>.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 60704047) and sponsored by Shanghai Pujiang Program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.bbrc.2008.08.125](https://doi.org/10.1016/j.bbrc.2008.08.125).

References

- [1] A.J. Barrett, J.K. McDonald, Nomenclature: protease, proteinase and peptidase, *Biochem. J.* 237 (1986) 935.
- [2] C. Seife, Blunting nature's Swiss army knife, *Science* 277 (1997) 1602–1603.
- [3] X.S. Puente, L.M. Sanchez, C.M. Overall, C. Lopez-Otin, Human and mouse proteases: a comparative genomic approach, *Nat. Rev. Genet.* 4 (2003) 544–548.
- [4] N.D. Rawlings, D.P. Tolle, A.J. Barrett, MEROPS: the peptidase database, *Nucleic Acids Res.* 32 (2004) D160–D164.
- [5] R.A. Poorman, A.G. Tomasselli, R.L. Heinrichson, F.J. Kezdy, A cumulative specificity model for proteases from human immunodeficiency virus types 1

- and 2 inferred from statistical analysis of an extended substrate data base, *J. Biol. Chem.* 266 (1991) 14554–14561.
- [6] H. Qin, S.M. Srinivasula, G. Wu, T. Fernandes-Alnemri, E.S. Alnemri, Y. Shi, Structural basis of procaspase-9 recruitment by the apoptotic protease-activating factor 1, *Nature* 399 (1999) 549–557.
- [7] J.J. Chou, H. Li, G.S. Salvessen, J. Yuan, G. Wagner, Solution structure of BID an intracellular amplifier of apoptotic signalling, *Cell* 96 (1999) 615–624.
- [8] W. Watt, K.A. Koeplinger, A.M. Mildner, R.L. Heinrichson, A.G. Tomasselli, K.D. Watenpaugh, The atomic resolution structure of human caspase-8, a key activator of apoptosis, *Structure* 7 (1999) 1135–1143.
- [9] K.C. Chou, D.Q. Wei, W.Z. Zhong, Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS (Erratum: *Biochem. Biophys. Res. Commun.* 310 (2003) 675), *Biochem. Biophys. Res. Commun.* 308 (2003) 148–151.
- [10] X.S. Puente, L.M. Sanchez, C.M. Overall, C. Lopez-Otin, Human and mouse proteases: a comparative genomic approach, *Nat. Rev. Genet.* 4 (2003) 544–558.
- [11] K.C. Chou, Review: structural bioinformatics and its impact to biomedical science, *Curr. Med. Chem.* 11 (2004) 2105–2134.
- [12] K.C. Chou, A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins, *J. Biol. Chem.* 268 (1993) 16938–16948.
- [13] K.C. Chou, Review: prediction of HIV protease cleavage sites in proteins, *Anal. Biochem.* 233 (1996) 1–14.
- [14] L. You, D. Garwicz, T. Rognvaldsson, Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease, *J. Virol.* 79 (2005) 12477–12486.
- [15] T. Rognvaldsson, L. You, D. Garwicz, Bioinformatic approaches for modeling the substrate specificity of HIV-1 protease: an overview, *Expert Rev. Mol. Diagn.* 7 (2007) 435–451.
- [16] G.Z. Liang, S.Z. Li, A new sequence representation as applied in better specificity elucidation for human immunodeficiency virus type 1 protease, *Biopolymers* 88 (2007) 401–412.
- [17] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL, *Nucleic Acids Res.* 25 (2000) 31–36.
- [18] K.C. Chou, Y.D. Cai, Prediction of protease types in a hybridization space, *Biochem. Biophys. Res. Commun.* 339 (2006) 1015–1020.
- [19] G.P. Zhou, Y.D. Cai, Predicting protease types by hybridizing gene ontology and pseudo amino acid composition, *Proteins: Struct. Funct. Bioinform.* 63 (2006) 681–684.
- [20] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction, *Anal. Biochem.* 370 (2007) 1–16.
- [21] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402.
- [22] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, *J. Biochem.* 99 (1986) 152–162.
- [23] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition (Erratum: *Proteins: Struct. Funct. Bioinform.* 44 (2001) 60), *Proteins: Struct. Funct. Bioinform.* 43 (2001) 246–255.
- [24] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, *J. Theor. Biol.* 248 (2007) 546–551.
- [25] H. Lin, The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition, *J. Theor. Biol.* 252 (2008) 350–356.
- [26] H. Gonzalez-Diaz, Y. Gonzalez-Díaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics networks and connectivity indices, *Proteomics* 8 (2008) 750–778.
- [27] I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz, P. Bork, SMART 5: domains in the context of genomes and networks, *Nucleic Acids Res.* 34 (2006) D257–D260.
- [28] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer, A. Bateman, Pfam: clans web tools and services, *Nucleic Acids Res.* 34 (2006) D247–D251.
- [29] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics* 4 (2003) 41.
- [30] A. Marchler-Bauer, J.B. Anderson, M.K. Derbyshire, C. DeWeese-Scott, N.R. Gonzales, M. Gwadz, L. Hao, S. He, D.I. Hurwitz, J.D. Jackson, Z. Ke, D. Krylov, C.J. Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, G.H. Marchler, M. Mullokandov, J.S. Song, N. Thanki, R.A. Yamashita, J.J. Yin, D. Zhang, S.H. Bryant, CDD: a conserved domain database for interactive domain family analysis, *Nucleic Acids Res.* 35 (2007) D237–D240.
- [31] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, *J. Biol. Chem.* 277 (2002) 45765–45769.
- [32] A.A. Schaffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res.* 29 (2001) 2994–3005.
- [33] J.H. Friedman, F. Baskett, L.J. Shustek, An algorithm for finding nearest neighbors, *IEEE Trans. Info. Theor.* C-24 (1975) 1000–1006.
- [34] T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, *IEEE Trans. Syst. Man Cybern.* 25 (1995) 804–813.
- [35] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers, *J. Proteome Res.* 5 (2006) 1888–1897.
- [36] K.C. Chou, H.B. Shen, Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms, *Nat. Protoc.* 3 (2008) 153–162.