Biochemical and Biophysical Research Communications 376 (2008) 321–325

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

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.

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, **pse**udo **p**osition-**s**pecific **s**coring **m**atrix; OET-KNN, **o**ptimized **e**vidence-**t**heoretic **K** **n**earest **n**eighbor.

* 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).

0006-291X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2008.08.125

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 $\geqslant$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.,

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

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

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

where $\mathbb{E}$ represents the prediction engine, $\mathbf{P}$ the query protein, $\triangleright$ is an identification operator, 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 training 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 evolutional 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 $\leqslant$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

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

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

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

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:

$$\mathbf{P}_{\text{PSSM}} = \begin{bmatrix} \mathbb{R}_{1\to1} & \mathbb{R}_{1\to2} & \cdots & \mathbb{R}_{1\to20} \\ \mathbb{R}_{2\to1} & \mathbb{R}_{2\to2} & \cdots & \mathbb{R}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{R}_{i\to1} & \mathbb{R}_{i\to2} & \vdots & \mathbb{R}_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{R}_{L\to1} & \mathbb{R}_{L\to2} & \cdots & \mathbb{R}_{L\to20} \end{bmatrix} \tag{5}$$