Vol.1, No.2, 63-92 (2009)
http://dx.doi.org/10.4236/ns.2009.12011

Natural Science

REVIEW

Recent advances in developing web-servers for predicting protein attributes*

Kuo-Chen Chou[1,2], Hong-Bin Shen[1,2]

[1]Gordon Life Science Institute, San Diego, California 92130, USA; kcchou@gordonlifescience.org
[2]Institute of Image Process & Pattern Recognition, Shanghai Jiaotong University, Shanghai, China

Received 7 August 2009; revised 25 August 2009; accepted 28 August 2009.

ABSTRACT

Recent advance in large-scale genome sequencing has generated a huge volume of protein sequences. In order to timely utilize the information hidden in these newly discovered sequences, it is highly desired to develop computational methods for efficiently identifying their various attributes because the information thus obtained will be very useful for both basic research and drug development. Particularly, it would be even more useful and welcome if a user-friendly web-server could be provided for each of these methods. In this minireview, a systematic introduction is presented to highlight the development of these web-servers by our group during the last three years.

Keywords: Cell-PLoc; Signal-CF; Signal-3L; MemType-2L; EzyPred; HIVcleave; GPCR-CA; ProtIdent; QuatIdent; FoldRate

1. INTRODUCTION

Proteomics, or "protein-based genomics", is the large-scale study of proteins. It was born due to the explosion of protein sequences generated in the post genomic era [1] as well as the necessity to understand the biological process at the cellular or system level.

To effectively conduct studies in proteomics, it is highly desired to develop high throughput tools by which one can timely identify various attributes of proteins in a large-scale manner.

For instance, given an uncharacterized protein sequence, how can we identify which subcellular location site it resides at? Does the protein stay in a single subcellular location or can it simultaneously exist in or move between two and more subcellular locations? Which part of the protein is its signal sequence? Is it a membrane protein or non-membrane protein? If it is the former, to which membrane protein type does it belong? Is it an enzyme or non-enzyme? If the former, to which main functional class and sub-functional class does it belongs to? Is it a protease on non-protease? If it is the former, to which protease type does it belong? Which sites of the protein can be cleaved by proteases such as HIV protease and SARS enzyme? Is it a GPCR (G-protein coupled receptor) or non-GPCR? If it is the former, to which type of GPCR does it belongs to? What kind of quaternary structure does it belong to? What kind of fold pattern does it assume? How can we estimate its folding rate? The list of questions is vast.

Although the answers to these questions can be determined by conducting various biochemical experiments, the approach of purely doing experiments is both time-consuming and costly. Consequently, the gap between the number of newly discovered protein sequences and the knowledge of their attributes is becoming increasingly wide.

For instance, in 1986 the Swiss-Prot databank contained merely 3,939 protein sequence entries **(Table 1)**, but the number has since jumped to 428,650 according to version 57.0 of 24-Mar-2009 (www.ebi.ac.uk/swiss-prot), meaning that the number of protein sequence entries now is more than 108 times the number from about 23 years ago. The rapid increase in protein sequence entries is also shown by the **Figure 1**, where a statistical illustration to show the growth of the UniProtKB/ TrEMBL Protein Database (http://www.ebi.ac.uk/uniprot/TrEMBLstats/) is given.

In order to use these newly found proteins for basic research and drug discovery in a timely manner, it is highly desired to bridge such a gap by developing effective computational methods to predict their 3D (three-dimensional) structures [2,3] as well as various function-related attributes based on their sequence information alone.

* Part of the contents in this article was presented in Shanghai University in June of 2009.

In this mini-review, we are to systematically introduce the recent progresses in addressing the aforementioned problems, particularly, for those prediction methods with web-servers available.

Table 1. The growth of protein sequences in SWISS-PROT data bank[a].

Release	Date	Number of sequence entries	Number of amino acids	Average length per sequenceb
2.0	09/86	3,939	900,163	229
5.0	09/87	5,205	1,327,683	236
9.0	11/88	8,702	2,498,140	287
12.0	10/89	12,305	3,797,482	309
16.0	11/90	18,364	5,986,949	326
20.0	11/91	22,654	7,500,130	331
24.0	12/92	28,154	9,545,427	339
27.0	10/93	33,329	11,484,420	345
30.0	10/94	40,292	14,147,368	351
32.0	11/95	49,340	17,385,503	352
34.0	10/96	59,021	21,210,389	359
35.0	11/97	69,113	25,083,768	363
37.0	12/98	77,977	28,268,293	363
38.0	07/99	80,000	29,085,965	364
39.0	05/00	86,593	31,411,114	363
40.0	10/01	101,602	37,315,215	367
42.0	10/03	135,850	50,046,799	368
45.0	10/04	163,235	59,631,787	365
48.0	09/05	194,317	70,391,852	362
51.0	10/06	241,242	88,541,632	367
56.0	07/08	392,667	141,217,034	360
57.0	03/09	428,650	154,416,236	360

a. From http://www.ebi.ac.uk/swissprot/.
b. The average length per sequence is defined as the total number of amino acids divided by the total number of sequences. The quotient is rounded to an integer.

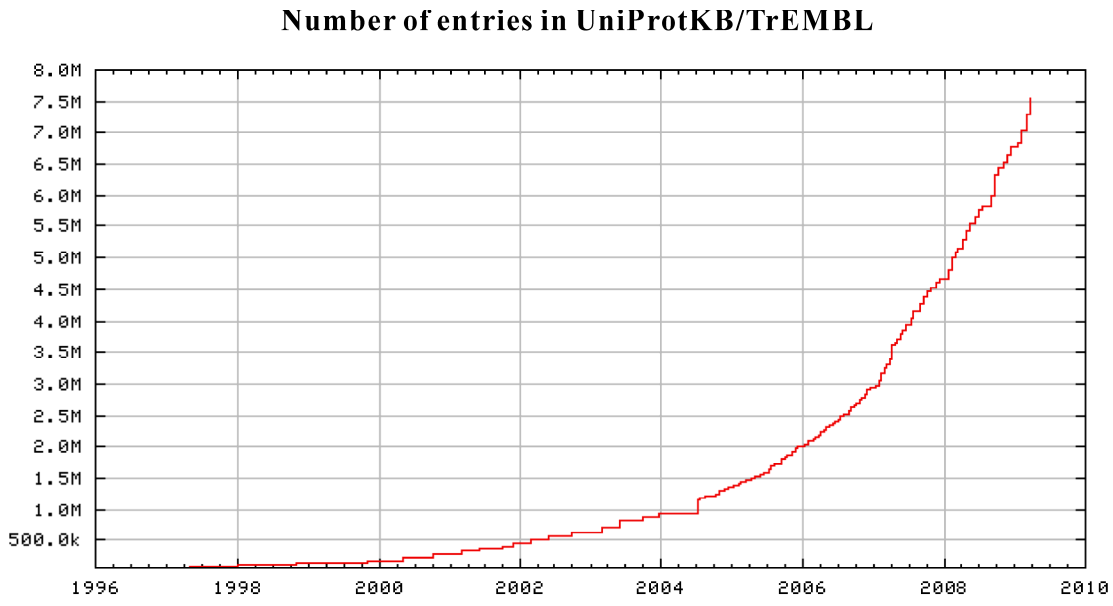

Figure 1. A statistical illustration to show the growth of the UniProtKB/TrEMBL Protein Database (http://www.ebi.ac.uk/uniprot/TrEMBLstats/).

2. WEB-SERVERS

Recently, a series of web-servers have been developed in our group, as described below.

2.1. Cell-PLoc

Thought by many as the most basic structural and functional unit of all living organisms, a cell is constituted by many different components, compartments or organelles **(Figure 2)**, and they are specialized to perform different tasks. For instance: cytoplasm, a jelly-like material, takes up most of the cell volume, filling the cell and serving as a "molecular soup" in which all of the cell's organelles are suspended; cell membrane functions as a boundary layer to contain the cytoplasm, while cell wall provides protection from physical injury; the cell nucleus contains the genetic material (DNA) governing all functions of the cell; the cytoskeleton functions as a cell's scaffold, organizing and maintaining the cell's shape, as well as anchoring organelles in place; mitochondrion is the "power generator" playing a critical role in generating energy in the eukaryotic cell; and so forth. However, most of these functions, which are critical to the cell's survival, are performed by the proteins in a cell [4,5]. Divided by many different compartments or organelles usually termed as "subcellular locations" **(Figure 2)**, a cell typically contains approximately one billion or 10^9 protein molecules each having its own location (for a single-location protein) or locations (for a multiple-location or multiplex protein). Therefore, one of the fundamental goals in proteomics and cell biology is to identify the subcellular localization of proteins and their functions.

During the past 18 years, varieties of predictors have been developed to address this problem (see, e.g., [6-48] and the relevant references cited in a recent review paper [49].

Developed recently, the **Cell-PLoc** [50] package contains a set of six web-servers for predicting subcellular localization of proteins in six different organisms. The six web servers and their coverage scopes can be summarized by the following formulation

$$\textbf{Cell-PLoc} = \begin{cases} \textbf{Euk-mPLoc}, & \text{for eukaryotic proteins covering 22 sites} \\ \textbf{Hum-mPLoc}, & \text{for human proteins covering 14 sites} \\ \textbf{Plant-PLoc}, & \text{for plant proteins covering 11 sites} \\ \textbf{Gpos-PLoc}, & \text{for Gram positive proteins covering 5 sites} \\ \textbf{Gneg-PLoc}, & \text{for Gram negative proteins covering 8 sites} \\ \textbf{Virus-PLoc}, & \text{for virus proteins covering 7 sites} \end{cases} \quad (1)$$

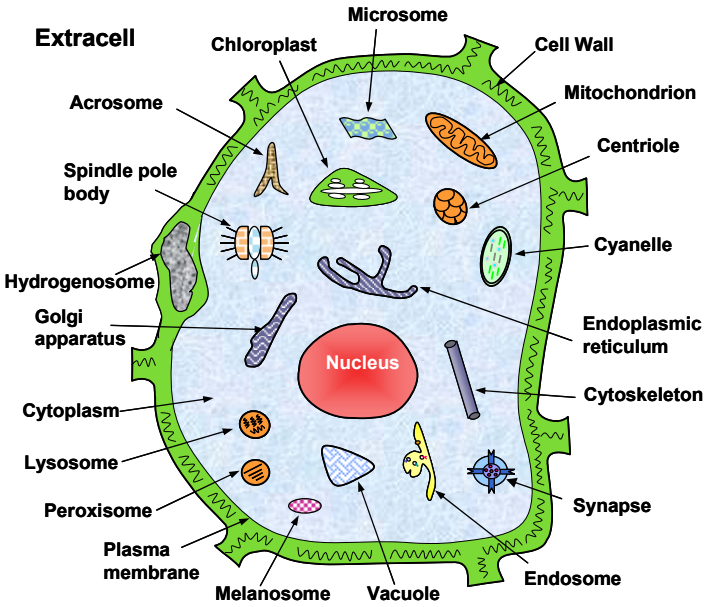

Figure 2. Schematic illustration to show many different components or organelles in a eukaryotic cell. Reproduced from [51] with permission.

where the character "m" in front of "PLoc" stands for "multiple", meaning that the corresponding predictor can be used to deal with both single-location and multiple-location proteins.

To use the web-server package, just do the following procedures. **(1)** Open the webpage http://chou.med.harvard.edu/bioinf/Cell-PLoc/, and you will see the top page of the **Cell-PLoc** package [50] on your computer screen, as shown in **Figure 3. (2)** To predict the subcellular localization of eukaryotic proteins, click the "Euk-mPLoc" button; to predict the subcellular localization of human proteins, click the "Hum-mPLoc" button; to predict the subcellular localization of plant proteins, click the "Plant-PLoc" button; and so forth. **(3)** Now, you can follow the procedures (3) – (11) as described in [50] to get the desired results for the query proteins in the six different organisms.

To maximize the convenience for the people working in the relevant areas, each of the six predictors in the **Cell-PLoc** package has been used to identify all the protein entries in the corresponding organism (except those annotated with "fragment" or those with less than 50 amino acids) in the Swiss-Prot database that do not have subcellular location annotations or are annotated with uncertain terms such as "probable", "potential", "likely", or "by similarity". These large-scale predicted results can be directly downloaded by clicking the Download button after getting on the top page of each of the six web-servers. These results can serve two purposes: one is that they can be directly used by those who need the information immediately; the other is to set a preceding mark to examine the accuracy of these web-server predictors by the future experimental results.

For example, listed in **Appendix A** are 334 eukaryotic proteins. Their experimental annotated subcellular locations were not available before Swiss-Prot 53.2 was released on 26-June-2007. However, according to the large-scale predicted results by **Euk-mPLoc** that were submitted for publication on November-12-2006 as Supporting Information B in [51] and were also at the same time placed in the downloadable file called **Tab_Euk-mPLoc** at http://chou.med.harvard.edu/bioinf/euk-multi/ [50] or http://202.120.37.186/bioinf/euk-multi/ [51], the predicted subcellular locations of the 334 eukaryotic proteins are given in column 4 of **Appendix A**, where for facilitating comparison the corresponding experimental results available about seven months later are also listed in column 5. From the table we can see the following: of the 334 eukaryotic proteins, 309 are with single location site and 25 with multiple location sites. Of the 309 single location proteins, only 22 were incorrectly predicted; of the 25 multiple location proteins, 2 (i.e., No.104 and No.322) were incorrectly predicted. It is interesting to see that the predicted result for No.104 was "Centriole; Nucleus" while the experimental observation "Cytoplasm; Nucleus", meaning only one of its two location sites was incorrectly predicted; and that the predicted result for No.322 was "Centriole; Cytoplasm; Nucleus" while the experimental observation "Nucleus; Cytoplasm", meaning both of its observed location sites were correctly predicted although the site of "Centriole" was over-predicted. Accordingly, the overall success rate for the 334 proteins is over 93% as proved later by experiments.

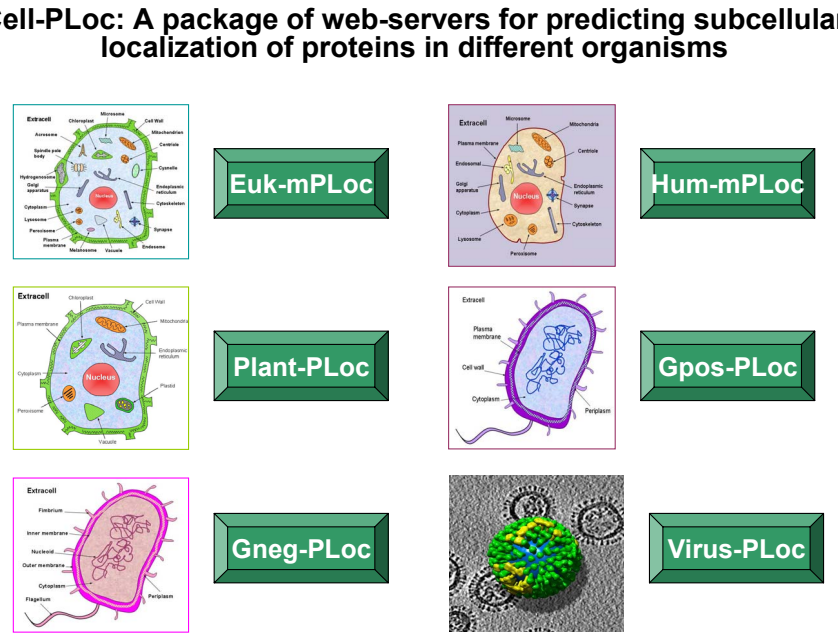

Figure 3. A semi-screenshot to show the Cell-PLoc web-page at (http://chou.med.harvard.edu/bioinf/Cell-PLoc/).

Although the predictors in the **Cell-PLoc** package [50] are very powerful, they have the following shortcomings. **(1)** In order for taking the advantage of Gene Ontology (GO) [52] approach [49], the input for a query protein must include its accession number. However, many proteins, such as synthetic and hypothetical proteins, as well as those newly-discovered proteins that have not been deposited into databanks yet, do not have accession numbers, and hence their subcellular locations cannot be predicted via the GO approach. **(2)** Since the current GO database is far from complete yet, many proteins cannot be meaningfully formulated in a GO space even if their accession numbers are available. **(3)** Although the PseAA (pseudo amino acid) composition [18,53] or PseAAC approach, a complement to the GO approach in **Cell-PLoc**, can take into account some partial sequence order effects, the original PseAAC [18] missed the functional domain (FunD) [23] and sequential evolution (SeqE) information [54,55]. To improve the aforementioned shortcomings, the **Cell-PLoc** package is currently under developing to be a new version, the **Cell-PLoc 2.0**. At this stage, some of the predictors therein, such as **Hum-mPLoc**2.0[56], **Plant-mPLoc** [56], **Gpos-mPLoc** [57], and **Gneg-mPLoc** [58], have been completed, as will be briefed below.

To show the difference of **Hum-mPLoc 2.0** with the original **Hum-mPLoc** [44] in the **Cell-PLoc** package [55], let us see the following demonstration steps.

Step1. Open the webpage http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/, and you will see its top page on your computer screen [50], as shown in **Figure 4a**.

Step 2. Either type or copy and past the query protein sequence into the input box (depicted by the box at the center of **Figure 4a**). The input sequence should be in FASTA format (http://en.wikipedia.org/wiki/Fasta_format), as shown by clicking on the Example button right above the input box. For example, if you use the 1st query protein sequence in the Example window, the input screen should look like the illustration in **Figure 4b**.

Step 3. After clicking the Submit button, you will see "Cell membrane; Cytoplasm; Nucleus" shown on the screen **(Figure 4c)** after 15 seconds or so, indicating that the query protein is a multiplex protein that may simultaneously exist in the three subcellular location sites, fully in agreement with experimental observations.

Step 4. If using the 2nd query protein sequence in the Example window as an input, after clicking the Submit

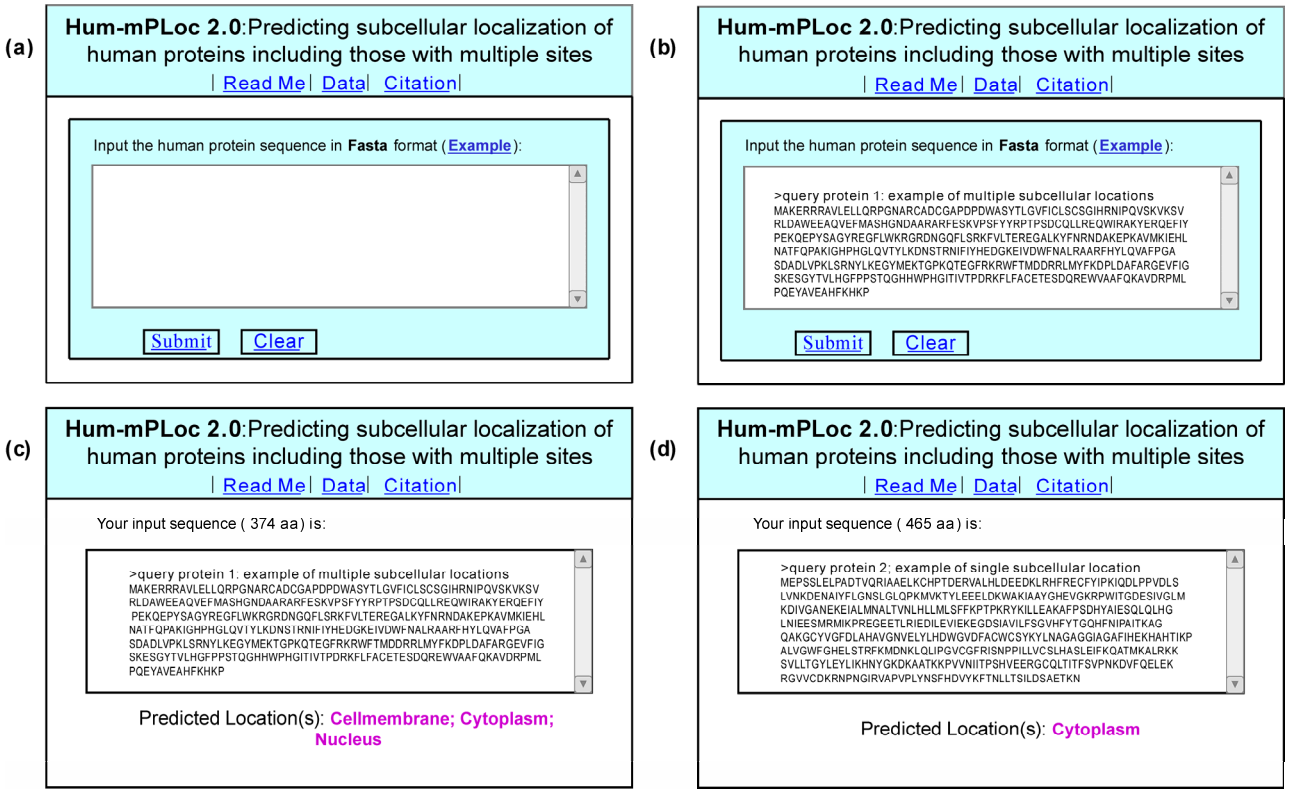

Figure 4. A semi-screenshot to show **(a)** the top page of the web-server Hum-mPLoc 2.0 at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/, **(b)** the input in FASTA format taken from the 1st query protein sequence in the Example window, **(c)** the output generated by clicking the Submit button in panel **b**, and **(d)** the output generated through the similar procedure but using the input taken from the 2nd query protein sequence in the Example window.

button, you will see "Cytoplasm" shown on the screen **(Figure 4d)**, indicating the query protein is a single-location protein residing at the cytoplasm compartment or organelle, also fully in agreement with experimental observations.

As we can see from the above steps, no accession numbers whatsoever are needed for the input data. This is quite different with the cases when using the original **Hum-mPLoc** in [55] to conduct prediction. Furthermore, the success rate expectancy has also been enhanced owing to taking into account the FunD and SeqE information.

Besides the improvements mentioned above, the developments from **Plant-PLoc** [43] in the **Cell-PLoc** package [50] to **Plant-mPLoc** [59], from **Gpos-PLoc** [60] to **Gpos-mPLoc** [57], and from **Gneg-PLoc** [61] to **Gneg-mPLoc** [58], have made it possible to deal with the multiple-location problem for plant proteins, Gram-positive bacterial proteins, and Gram-negative bacterial proteins, respectively, as well.

2.2. Nuc-PLoc

The nucleus exists only in eukaryotic cells. Located at the center of a cell like its kernel, the nucleus is the most prominent and largest cellular organelle [5], with the diameter from 11 to 22 micrometers (μm) and occupying about 10% of the total volume of a typical animal cell [62]. The life processes of a eukaryotic cell are guided by its nucleus. In addition to the genetic material, the cellular nucleus contains many proteins located at its different compartments, called subnuclear locations. Therefore, the information of protein subnuclear localization is not only equally important to that of protein subcellular localization but also possesses the sense at a deeper level.

By fusing the SeqE approach and PseAAC approach [63], a web-server called **Nuc-PLoc** was developed that is accessible to the public via the website http://chou.med.harvard.edu/bioinf/Nuc-PLoc/. It can be used to identify nuclear proteins among the following nine subnuclear locations: (1) chromatin, (2) heterochromatin, (3) nuclear envelope, (4) nuclear matrix, (5) nuclear pore complex, (6) nuclear speckle, (7) nucleolus, (8) nucleoplasm, (9) nuclear PML body **(Figure 5)**.

2.3. Signal-CF

Functioning as a "zip code" or "address tag" in guiding proteins to the cellular locations where they are supposed to be **(Figure 6)**, signal peptides control the entry of virtually all secretory proteins to the pathway, both in eukaryotes and prokaryotes [64-66]. If the signal peptide for a nascent protein was changed, the protein could end in a wrong cellular location causing a variety of strange diseases. Accordingly, knowledge of signal peptides can be utilized to reprogram cells in a desired way for future cell and gene therapy. However, to realize this, an indispensable thing is to identify the signal peptide for a

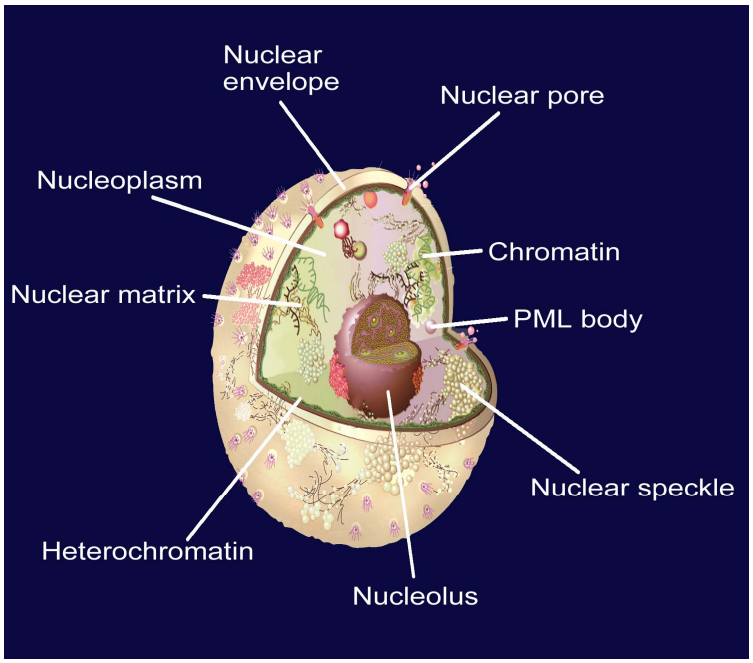

Figure 5. Schematic drawing to show the nine subnuclear locations: (1) chromatin, (2) heterochromatin, (3) nuclear envelope, (4) nuclear matrix, (5) nuclear pore complex, (6) nuclear speckle, (7) nucleolus, (8) nucleoplasm, (9) nuclear PML body. Adapted from [252] with permission.

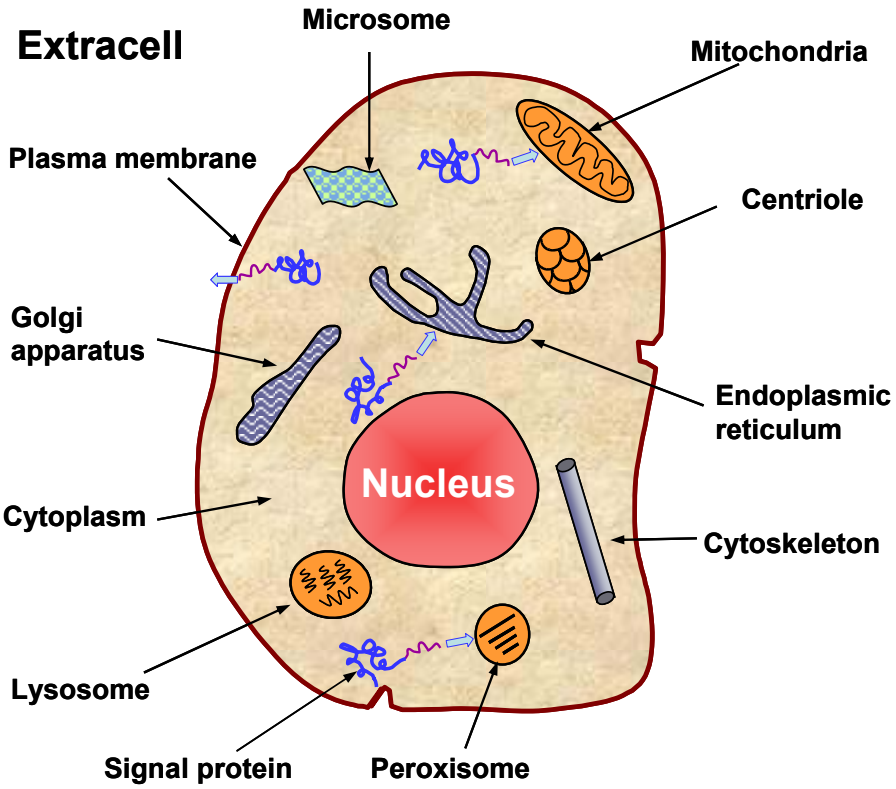

Figure 6. A schematic drawing to show: how the signal peptides of secretory proteins function as an "address tag" in directing the proteins to their proper cellular and extracellular locations. The signal peptide sequence is colored in puple, and the mature protein sequence in blue.

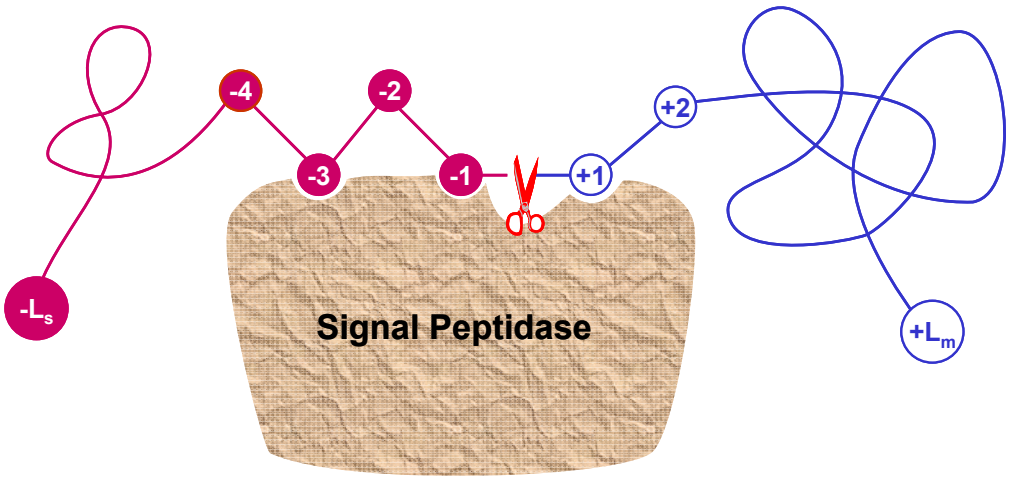

Figure 7. A schematic drawing to show the signal sequence of a protein and how it is cleaved by the signal peptidase. An amino acid in the signal part is depicted as a red circle with a white number to indicate its sequential position, while that in the mature protein depicted as an open circle with a blue number. The signal sequence contains L_s residues and the mature protein L_m residues. The cleavage site is at the position (-1, +1), i.e., between the last residue of the signal sequence and the first residue of the mature protein.

nascent protein. Many efforts have been made in this regards (see, e.g., [67-76] as well as the relevant references listed in a review article [77]).

The signal peptide of a secretory protein is usually located at its N-terminal, and it will be cleaved off by a signal peptidase once the protein is translocated through

a membrane **(Figure 7)**, where the cleavage site is commonly symbolized by $(-1, +1)$, namely the position between the last residue of the signal peptide and the first residue of the mature protein. It can also be seen from **Figure 7** that once the cleavage site is identified, the corresponding signal peptide is automatically known; and vice versa.

The difficulty in predicting signal peptides is that for different secretory proteins, their signal peptides are quite different not only in sequence components and sequence orders but also in sequence lengths. Also, many previous methods were lacking of considering the coupling effects of the subsites around the cleavage sites, as analyzed in [78].

To address the above two problems, the web-server predictor called **Signal-CF** [79] was developed recently. Its features are reflected by its name, where "C" stands for "Coupling" and "F" for "Fusion", meaning that **Signal-CF** is formed by incorporating the subsite coupling effects along a protein sequence and by fusing the results derived from many width-different scaled windows through a voting system.

Signal-CF is a 2-layer predictor: the 1st-layer prediction engine is to identify a query protein as secretory or non-secretory; if it is secretory, the process will be automatically continued with the 2nd-layer prediction engine to further identify the cleavage site of its signal peptide. The predictor is also featured by high success prediction rates with short computational time, and hence is particularly useful for the analysis of large-scale datasets. **Signal-CF** is freely accessible at http://chou.med.harvard.edu/bioinf/Signal-CF/.

2.4. Signal-3L

This is a 3-layer predictor developed for identifying the signal peptides of human, plant, animal, eukaryotic, Gram-positive, and Gram-negative proteins. The target of the 1st-layer is to identify a query protein as secretory or non-secretory. If the protein is identified as secretory, the process will be automatically continued by the 2nd-layer prediction engine to identify the potential cleavage sites **(Figure 7)** along its sequence. The 3rd-layer is to finally determine the unique cleavage site through a global sequence alignment operation. **Signal-3L** is accessible to the public as a web-server at http://chou.med.harvard.edu/bioinf/Signal-3L/. Compared with **Signal- CF**, it might take a little longer computational time but yield a little higher accuracy.

Table 2. List of examples showing that signal peptides miss-predicted by SignalP-NN and/or SignalP-HMM are corrected by Signal-3L.

Protein[a]	Experimentally verified signal peptide[a]	SignalP 3.0-NN	SignalP 3.0-HMM	Signal-3L
AAF91396.1	**1-40**	1-37	1-37	**1-40**
DKK1_HUMAN	**1-31**	1-22	1-28	**1-31**
MIME_HUMAN	**1-20**	1-19	1-19	**1-20**
NP_057466.1	**1-21**	1-19	1-19	**1-21**
NP_057663.1	**1-35**	1-30	1-46	**1-35**
NP_443122.2	**1-21**	1-22	1-22	**1-21**
NP_443164.1	**1-26**	1-33	1-33	**1-26**
Q6UXL0	**1-28**	1-29	1-29	**1-28**
STC1_HUMAN	**1-17**	1-21	1-18	**1-17**
TRLT_HUMAN	**1-25**	1-24	1-27	**1-25**
CD5L_HUMAN	**1-19**	1-18	**1-19**	**1-19**
EDAR_HUMAN	**1-26**	1-28	**1-26**	**1-26**
FZD3_HUMAN	**1-22**	1-17	**1-22**	**1-22**
IBP7_HUMAN	**1-26**	**1-26**	1-29	**1-26**
KLK3_HUMAN	**1-17**	**1-17**	1-23	**1-17**
NMA_HUMAN	**1-20**	**1-20**	1-26	**1-20**
NP_064510.1	**1-22**	**1-22**	1-23	**1-22**
NP_068742.1	**1-24**	**1-24**	1-25	**1-24**
NTRI_HUMAN	**1-33**	1-30	1-33	**1-33**
SY01_HUMAN	**1-23**	**1-23**	1-18	**1-23**
TIE1_HUMAN	**1-21**	**1-21**	1-22	**1-21**
TL19_HUMAN	**1-26**	1-23	**1-26**	**1-26**
TR14_HUMAN	**1-38**	1-36	**1-38**	**1-38**
TR19_HUMAN	**1-29**	**1-29**	1-25	**1-29**
XP_166856	**1-17**	**1-17**	1-20	**1-17**
XP_209141	**1-22**	1-23	**1-22**	**1-22**

[a] Data taken from [251]. The signal peptides experimentally verified and correctly predicted are in bold-face type colored in blue; those incorrectly predicted in red. (For interpretation of the references to color in this table caption, the reader is referred to the web version of this paper.)

Both **Signal-CF** and **Signal-3L** can be used to refine the results by other predictors in this area. For instance, listed in **Table 2** are the signal peptides that were miss-predicted by **SignalP-NN** and/or **SignalP-HMM** in the **SignalP** package [75] but corrected by **Signal-3L**.

Also, according to a recent report (see Table 1 of [80]) **Signal-CF** performed the best in predicting the long signal peptides, among the following eight web-server predictors: **SignalP-NN** [75], **SignalP-HMM** [75], **SignalP-NN** or **SignalP-HMM** [75], **Phobius** [81], **PrediSi** [76], **Signal-CF** [79], **Signal-3L** [82], and **Philius** [83].

2.5. MemType-2L

Given a protein sequence, how can one identify whether it is a membrane protein or not? If it is, which membrane protein type it belongs to? It is important to address these problems because they are closely relevant to the biological function of the protein concerned and to its interaction process with other molecules in a biological system. Most functional units or organelles in a cell are "enveloped" by one or more membranes, which are the structural basis for many important biological functions. Although the basic structure of membranes is lipid bilayer, many specific functions of the cell membrane are performed by the membrane proteins (see, e.g., [4,5]). For example, it is through membrane proteins that various chemical messages such as nerve impulses and hormone activity can be passed between cells (see, e.g., [84]); that cells can be attached to an extracellular matrix in grouping cells together to form tissues; that parts of the cytoskeleton can be attached to the cell membrane in order to provide shape; that the metabolism process and body's defense mechanisms can be completed; as well as that molecules can be transported into and out of cells by such methods as proton pumps (see, e.g., [85-87]) and ion pumps (see, e.g., [88,89]), channel proteins [90-92] and carrier proteins (see, e.g., [93]).

Membrane proteins possess different types, which are closely correlated with their functions. For instance, the transmembrane proteins can transport molecules across the membrane or function on both its sides, whereas proteins functioning on only one side of the lipid bilayer are often associated exclusively with either the lipid monolayer or a protein domain on that side. Therefore, information about membrane protein type can provide useful hints for determining the function of an uncharacterized membrane protein. Furthermore, because of the fluid nature of their infrastructure, membrane proteins can move around the cell membrane so as to reach where their function is required. Therefore, it will certainly expedite the pace in determining the function and action process of uncharacterized membrane proteins if we can timely acquire the knowledge of their type. With the avalanche of protein sequences generated in the post genomic age and the fact that membrane proteins are encoded by 20-35% of genes [94], it is self-evident why it is so important to develop a sequence-based automated method for fast and effectively addressing the two problems posed at the beginning of this Section.

Stimulated by the encouraging results in predicting the structural classification of proteins based on their amino acid (AA) composition or AAC [95-103], the covariant discriminant algorithm was introduced to identify the types of membrane proteins also based on their AA composition in 1999 [104]. However, the AA composition does not contain any sequence order information. To avoid completely losing the sequence order information, the PseAA composition or PseAAC was introduced [18]. Since then, various prediction methods have been proposed in this area [53,105-118].

Recently, a user-friendly web-server predictor called "**MemType-2L**" was developed [54]. Compared with the other predictors which only cover 5-6 membrane types, **MemType-2L** can cover 8 membrane types **(Figure 8)**. **MemType-2L** is a 2-layer predictor: the 1st layer prediction engine is to identify a query protein as membrane or non-membrane; if it is membrane, the process will be automatically continued with the 2nd-layer prediction engine to further identify its type among the following eight categories **(Figure 8)**: (1) type I, (2) type II, (3) type III, (4) type IV, (5) multipass, (6) lipid-chain-anchored, (7) GPI-anchored, and (8) peripheral.

MemType-2L is accessible to the public via the website at http://chou.med.harvard.edu/bioinf/MemType/.

2.6. EzyPred

Nearly all known enzymes are proteins that catalyze chemical reactions and are vitally important in the metabolic process. Given a protein sequence, how can we identify whether it is an enzyme or non-enzyme? If it is, which main functional class it belongs to? What about its sub functional class? These problems are closely correlated with the biological function of an uncharacterized protein and its acting object and process [119]. Although their answers can be found by conducting various biochemical experiments, it is both time-consuming and costly to do so solely by experimental approaches. During the last six years, a number of predictors have been developed to address these problems [53,120-125].

Recently, a top-down automated method called "**EzyPred**" was developed [126]. It not only covers all the six enzyme main-functional classes [127], but also many of their sub-functional classes (see **Figure 9). EzyPred** is a 3-layer predictor: the 1st layer prediction engine is for identifying a query protein as enzyme or non-enzyme;

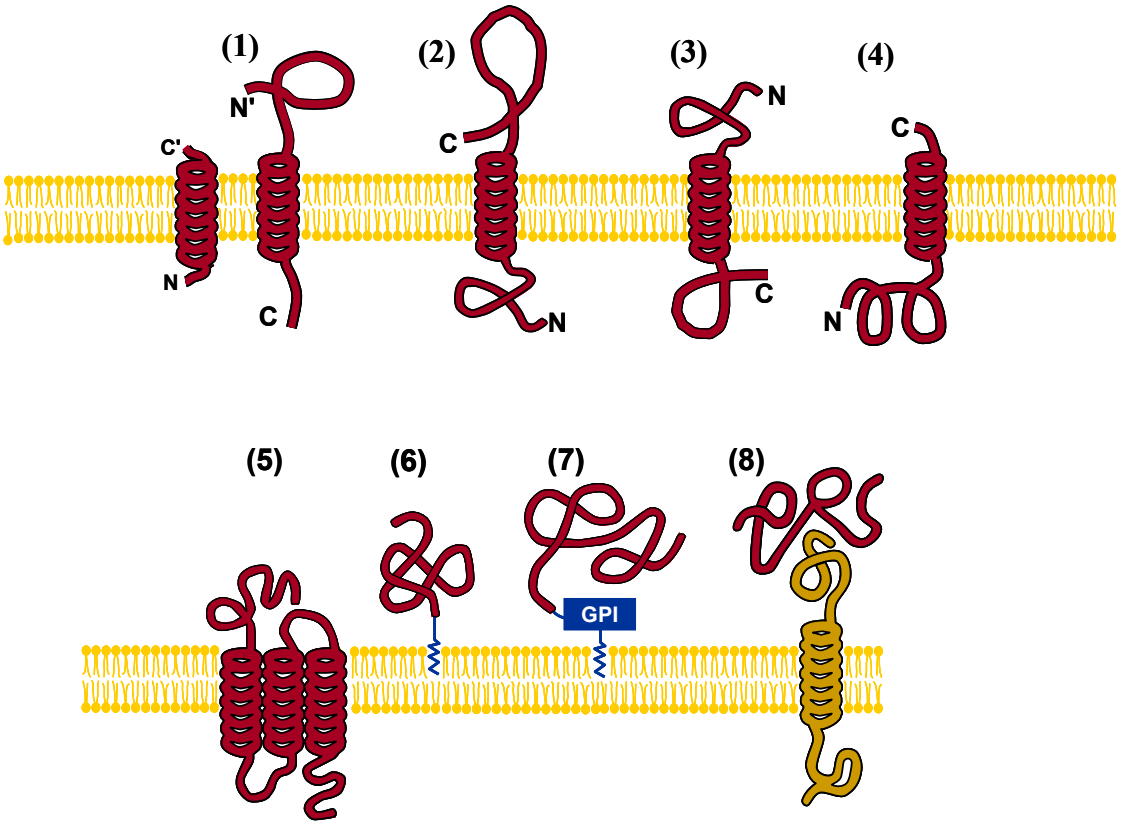

Figure 8. Schematic illustration showing the 8 types of membrane proteins: (1) type I transmembrane, (2) type II, (3) type III, (4) type IV, (5) multipass transmembrane, (6) lipid-chain-anchored membrane, (7) GPI-anchored membrane, and (8) peripheral membrane. As shown in the figure, types I, II, III, and IV are all of single-pass transmembrane proteins; see [253] for a detailed description about their difference. Reproduced from [54] with permission.

the 2nd layer for the main functional class; and the 3rd layer for the sub functional class. Within 90 seconds of submitting the sequence of a query protein into its input box, **EzyPred** will identify whether the query protein is enzyme or non-enzyme and, if it is an enzyme, to which main-functional class and sub-functional class it belongs.

EzyPred is accessible to the public as a web-server at http://chou.med.harvard.edu/bioinf/EzyPred/.

2.7. ProtIdent

Called by many as the biology's version of Swiss army knives, proteases cut long sequences of amino acids into fragments and regulate most physiological processes. They are vitally important in life cycle and have become a main target for drug design (see, e.g., [2,128-134]).

The actions of proteases are exquisitely selective (see, e.g. [135-139]), 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, proteases are classified the following six types: (1) aspartic, (2) cysteine, (3) glutamic, (4) metallo, (5) serine, and (6) threonine [140]. 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?

During the last three years, some efforts have been made in this regard [141,142]. However, none of these methods provided a web-server that can be easily used by the majority of experimental and pharmaceutical scientists to obtain the desired data.

Very recently, a web-server called "**ProtIdent**" was developed [55] by fusing the FunD (functional domain) and SeqE (sequential evolution) information **(Figure 10a)**. **ProtIdent** is a 2-layer predictor: the 1st layer is for identifying a 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 different mechanistic types **(Figure 10b)**.

Furthermore, a step-by-step protocol guide [143] was

provided for demonstrating how to use the **ProtIdent** web-server, by which one can get the desired 2-level results for a query protein sequence in around 25 seconds.

ProtIdent is freely accessible to the public via the site at http://www.csbio.sjtu.edu.cn/bioinf/Protease.

2.8. GPCR-CA

One of the largest families in the human genome is the one encoding the G-protein-coupled receptors (GPCRs), which are cell surface receptors. Owing to their characteristic transmembrane topology, GPCRs are also known as 7-transmembrane receptors, 7TM receptors, heptahelical receptors, and serpentine receptors that "snake"

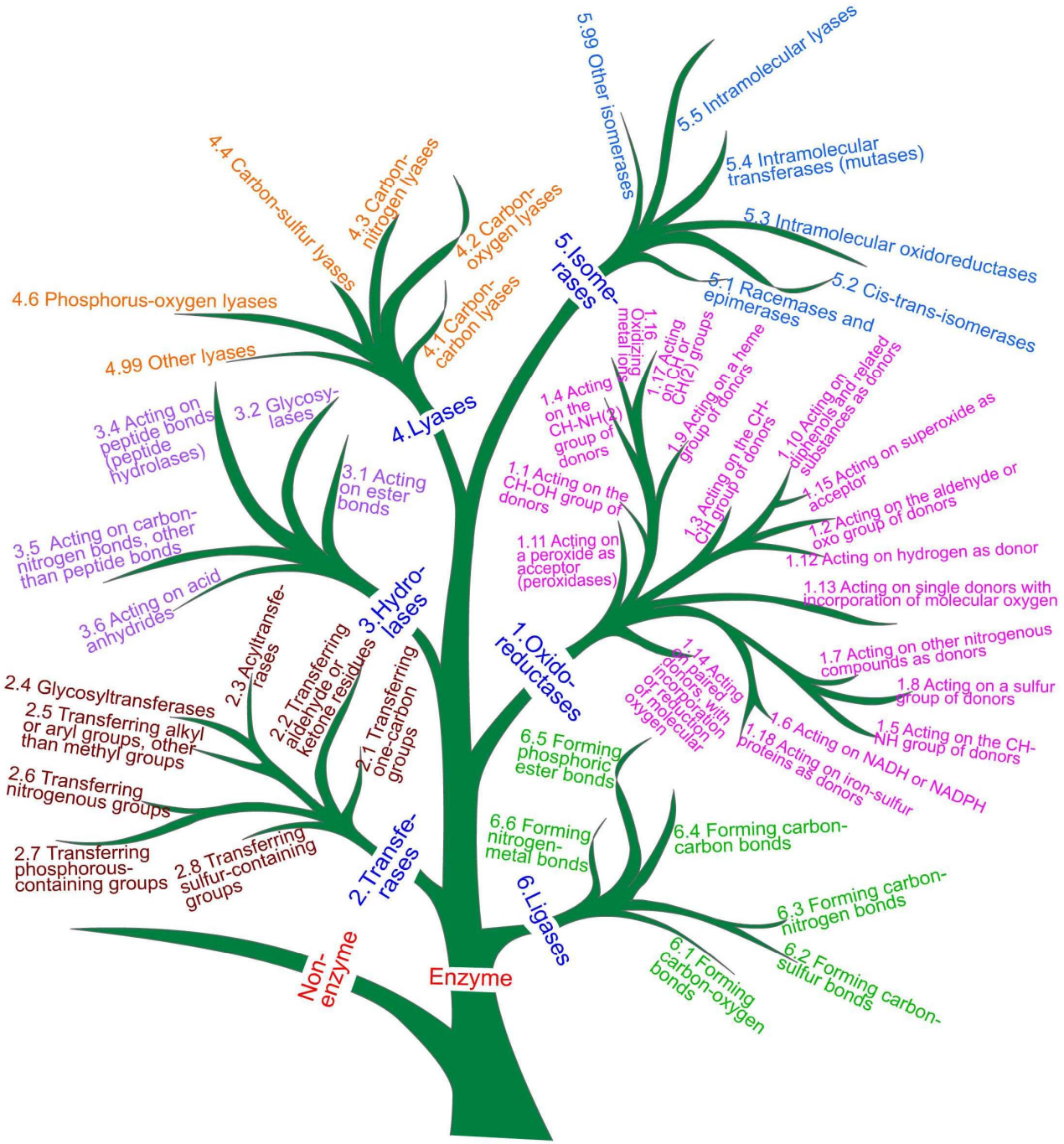

Figure 9. A schematic drawing to use tree branches to classify enzyme and non-enzyme as well as the six main functional classes of enzymes and their subclasses.

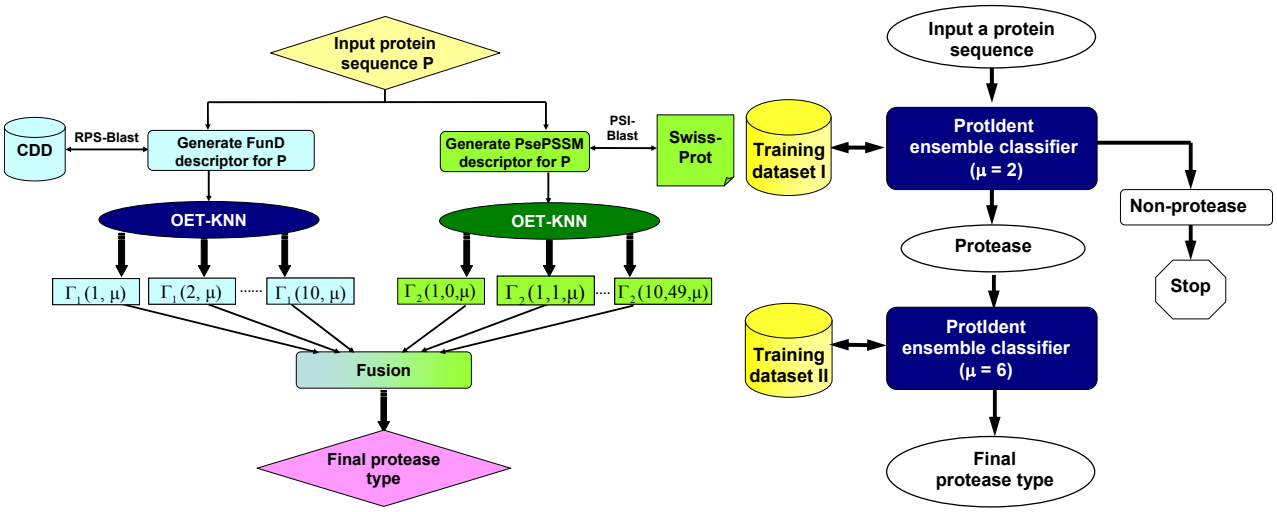

Figure 10. A flowchart to show **(a)** how to fuse the FunD approach and PsePSSM approach, and **(b)** how the two-layer Prot-Ident ensemble classifier works in identifying proteases and their functional types. See [55] for further explanation.

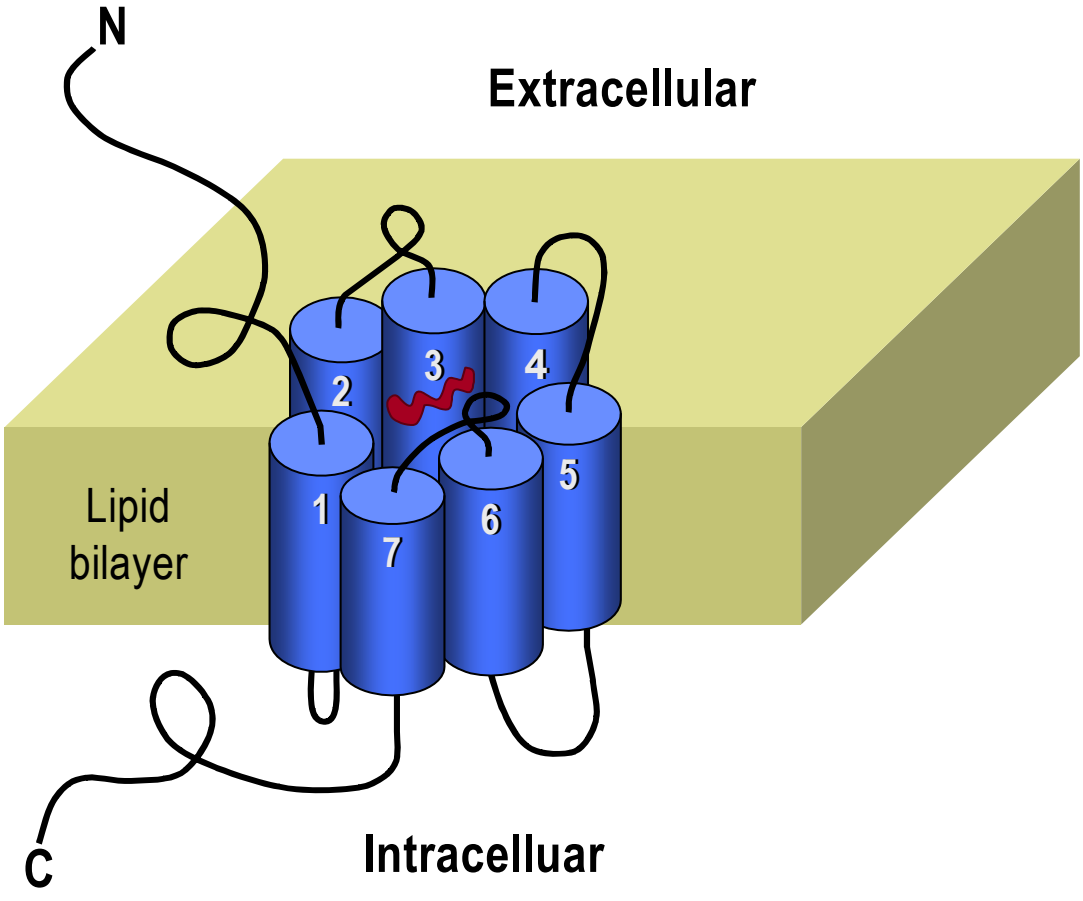

Figure 11. Schematic representation of a GPCR with a trademark of seven-transmembrane helices, depicted as cylinders and connected by alternating cytoplasmic and extracellular hydrophilic loops. The 7-helix bundle thus formed has a central pore on its extracellular surface. The red entity located in the central pore represents a ligand messenger.

across a cell membrane seven times **(Figure 11)**. The major role of GPCRs is to transmit signals into the cell. GPCR-associated proteins may play at least the following four distinct roles in receptor signaling [144-147]: (1) directly mediate receptor signaling, as in the case of G proteins; (2) regulate receptor signaling through controlling receptor localization and/or trafficking; (3) act as a scaffold, physically linking the receptor to various effectors; (4) act as an allosteric modulator of receptor conformation, altering receptor pharmacology and/or other aspects of receptor function.

Much effort has been invested for studying GPCRs by both academic institutions and pharmaceutical industries. Today, approximately one third of the world small molecule drug markets are GPCR agonists and antagonists.

The functions of many of GPCRs are still unknown, and it is both time-consuming and costly to determine their ligands and signaling pathways. Particularly, as membrane proteins, GPCRs are very difficult to crystallize and most of them will not dissolve in normal solvents. Accordingly, so far very few crystal GPCR structures have been determined. Although the recently developed state-of-the-art NMR technique is a very powerful in determining the 3D structures of membrane proteins [87,92-94,148], it is time-consuming and costly. In order to timely obtain the protein 3D structures for rational drug design, the approach of structural bioinformatics has been often adopted (see, e.g., [84,149-153]). Unfortunately, such an approach fails to work in most GPCR-related cases because very few GPCRs have sufficiently high sequence similarity with existing structure-known proteins, an indispensable condition for developing a reasonable starting structure via structural bioinformatics [2,3]. Consequently, it is highly desired to develop automated methods that can fast and effectively identify the functional families of GPCRs according to their sequence information because the information thus obtained can help classifying drugs, a technique called "evolutionary pharmacology" quite useful for drug development.

During the last 7 years or so, a number of methods were proposed in this regard [154-159]. Some of them were developed for identifying the main functional classes of GPCRs (see, e.g., [157]) and some for the sub- functional classes (see, e.g., [155]). None of these methods has provided a web-server for the public usage, and hence their practical application value is quite limited.

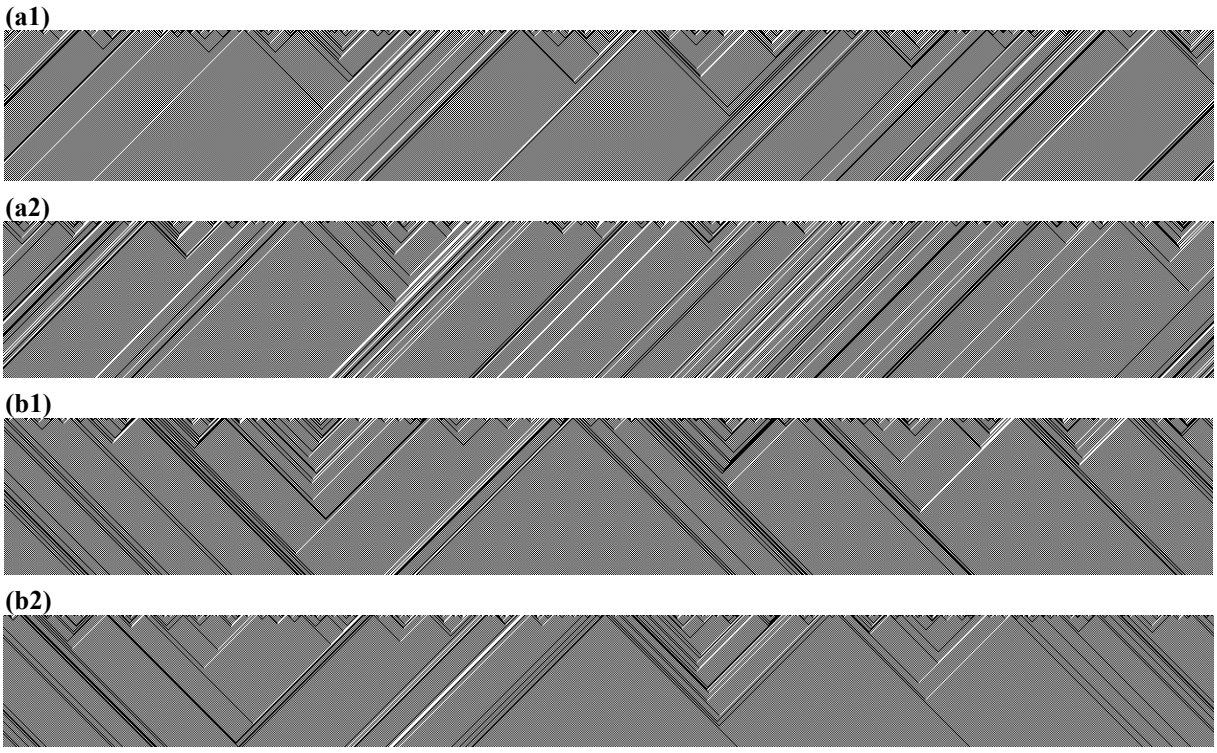

Figure 12. The cellular automaton image generated according to Eqs.2-5 for (**a1**) the rhodopsin like family member with accession number P41595; (**a2**) the rhodopsin like family member with accession number P18599; (**b1**) the secretin like family member with accession number O95838; and (**b2**) the secretin like family member with accession number Q02644. Panels (a1) and (a2) share a quite similar texture because the protein sequences from which the cellular automaton images were derived belong to a same GPCR family. And the same is true for panels (b1) and (b2).

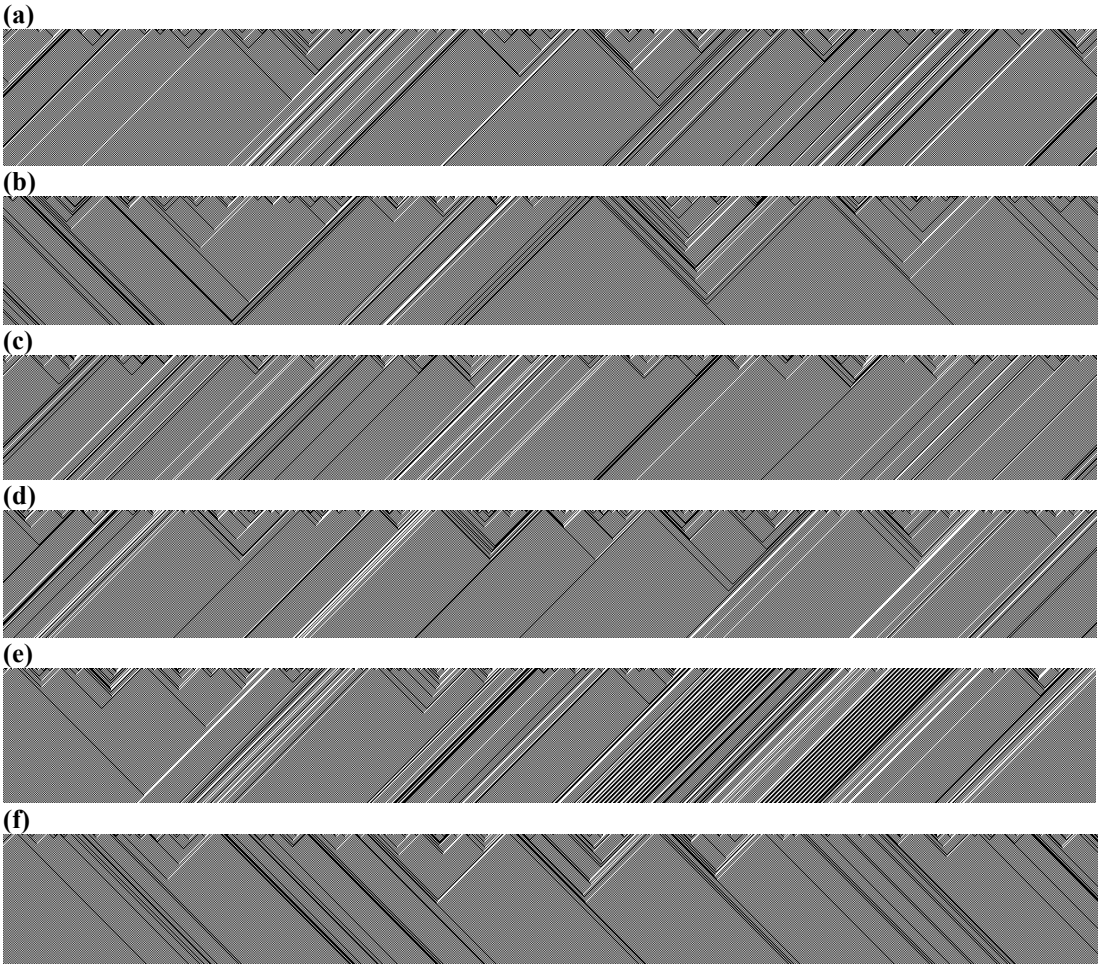

Figure 13. The cellular automaton image generated according to Eqs.2-5 for a protein taken from **(a)** A-rhodopsin like family, **(b)** B-secretin like family, **(c)** C-metabotrophic/glutamate/pheromone family; **(d)** D-fungal pheromone family, **(e)** E-cAMP receptor family, and **(f)** F-Frizzled/Smoothemed family, respectively. The six panels have completely different textures because they represent six different GPCR family members.

Recently, a web-server predictor was developed [160] with the name as **GPCR-CA**, where "CA" stands for "Cellular Automaton" [161], meaning that the cellular automaton images have been utilized to reveal the pattern features hidden in piles of long and complicated protein sequences. Cellular automata are discrete dynamical systems whose behavior is completely specified in terms of a local relation. A cellular automaton can be thought of as a stylized universe consisting of a regular grid of cells, each of which is in one of a finite number of possible states, updated synchronously in discrete time steps according to a local, identical interaction rule [162].

The procedures of generating the cellular automaton images for protein sequences can be briefed as follows. As a first step, each of the 20 native amino acids in a protein sequence is represented by a 5-digit strain according to the binary coding as defined in [163]. Thus, a protein consisting of N amino acids can be converted to a sequence with $5N$ digits (or grids); i.e,,

$$\mathrm{g}_1(t)\mathrm{g}_2(t)\cdots\mathrm{g}_N(t)\cdots\mathrm{g}_{5N}(t), \quad (t=0) \tag{2}$$

where $\mathrm{g}_i(t)=0$ or 1 $(i=1,\ 2,\ \cdots,\ 5N)$ as defined in [163]. Suppose the time for each updated step is consecutively expressed by $t=0,\ 1,\ 2,\ \cdots,\ \Omega$, we have

$$\begin{cases} \mathrm{g}_1(0)\ \mathrm{g}_2(0)\ \cdots\mathrm{g}_N(0)\ \cdots\ \mathrm{g}_{5N}(0) \\ \Downarrow \\ \mathrm{g}_1(1)\ \mathrm{g}_2(1)\ \cdots\ \mathrm{g}_N(1)\ \cdots\ \mathrm{g}_{5N}(1) \\ \Downarrow \\ \mathrm{g}_1(2)\ \mathrm{g}_2(2)\ \cdots\ \mathrm{g}_N(2)\ \cdots\ \mathrm{g}_{5N}(2) \\ \Downarrow \\ \vdots \\ \Downarrow \\ \mathrm{g}_1(\Omega)\mathrm{g}_2(\Omega)\cdots\mathrm{g}_N(\Omega)\ \cdots\mathrm{g}_{5N}(\Omega) \end{cases} \tag{3}$$

where

$$g_i(t+1)=\begin{cases}0, & \text{if } g_{i-1}(t)=0,\ g_i(t)=0,\ g_{i+1}(t)=0\\ 0, & \text{if } g_{i-1}(t)=0,\ g_i(t)=0,\ g_{i+1}(t)=1\\ 1, & \text{if } g_{i-1}(t)=0,\ g_i(t)=1,\ g_{i+1}(t)=0\\ 0, & \text{if } g_{i-1}(t)=0,\ g_i(t)=1,\ g_{i+1}(t)=1\\ 1, & \text{if } g_{i-1}(t)=1,\ g_i(t)=0,\ g_{i+1}(t)=0\\ 0, & \text{if } g_{i-1}(t)=1,\ g_i(t)=0,\ g_{i+1}(t)=1\\ 1, & \text{if } g_{i-1}(t)=1,\ g_i(t)=1,\ g_{i+1}(t)=0\\ 0, & \text{if } g_{i-1}(t)=1,\ g_i(t)=1,\ g_{i+1}(t)=1\end{cases} \quad (t=0,\ 1,\ \cdots,\ \Omega) \qquad \textbf{(4)}$$

with the spatially periodic boundary conditions; i.e.,

$$g_0(t)=g_{5N}(t) \quad \text{and} \quad g_{5N+1}(t)=g_1(t) \qquad \textbf{(5)}$$

Suppose: $g_i(t)$, the ith grid at t, is filled with white color if $g_i(t)=0$ and black if $g_i(t)=1$. Accordingly, each row of **Eq.3** corresponds to a narrow ribbon mixed with white and black colors. Scanning these ribbons successively on to a screen or sheet will generate a 2D (2-dimensional) black-and-white image. It has been observed that the image texture is basically steady after $t=\Omega=100$. The image thus evolved is called the cellular automaton image for the protein sequence concerned. The advantage of using the cellular automaton image to represent the protein is that it can help us visualize some special features hidden in its long and complex sequence [163]. For instance, the cellular automata images for proteins from a same GPCR family share a similar texture pattern **(Figure 12)**, while those from different GPCR families have different texture patterns **(Figure 13)**.

Subsequently, the gray-level co-occurrence matrix factors extracted from the cellular automaton images were used to represent the samples of proteins through their pseudo amino acid composition [18,53], followed by utilizing the augmented covariant-discriminant classifier [12,164] to operate the prediction of **GPCR-CA.**

GPCR-CA is a 2-layer predictor: the 1st layer prediction engine is for identifying a query protein as GPCR on non-GPCR; if it is a GPCR protein, the process will be automatically continued with the 2nd-layer prediction engine to further identify its type among the following six functional classes: (1) rhodopsin-like, (2) secretin-like, (3) metabotrophic/glutamate/pheromone; (4) fungal pheromone, (5) cAMP receptor, and (6) Frizzled/Smoothened family. **GPCR-CA** is freely accessible at http://218.65.61.89:8080/bioinfo/GPCR-CA, by which one can get the desired 2-layer results for a query protein sequence within about 20 seconds.

2.9. HIVcleave

During the past 17 years, the following two strategies have often been utilized to find drugs against AIDS (acquired immunodeficiency syndrome). One is to target the HIV (human immunodeficiency virus) reverse transcriptase (see, e.g., [165-171]); the other is to design HIV protease inhibitors [128,136,138,139,172-174].

Functioning as a dimer, the HIV protease is made up of two identical subunits, each having 99 residues, but with only one active site [136,174]. The essential function of HIV protease is to cleave the precursor polyprotens; loss of the cleavage-ability will stop the life cycle of infectious HIV, the culprit [175,176] of AIDS.

To find the effective inhibitors against HIV protease, it is very helpful to understand the mechanism of how it cleaves the polyproteins and utilize the "distorted key" theory [136] to approach the problem, as illustrated below. HIV protease is a member of the aspartyl proteases that is highly substrate-selective and cleavage-specific. The HIV protease-susceptible sites in a given protein extend to an octapeptide region [177], with its amino acid residues sequentially symbolized by eight subsites R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, $R_{4'}$ [178], as shown in **Figure 14**. The scissile bond is located between the subsites R_1 and $R_{1'}$. Occasionally, the susceptible sites in some proteins may contain one subsite less or one subsite more, corresponding to the case of a heptapeptide or nonapeptide, respectively. However, in investigating the cleavability of peptide sequences by HIV proteases, heptapeptides and nonapeptides need to be considered very rarely. This might be the result of a compromise between the following two factors. On one hand, according to the "rack mechanism" [179], the active site of HIV protease can be likened to a "rack" during the peptide cleaving process. Thus, it appears that the more residues that are bound to the rack of enzyme, the more strained the peptide, and hence the more efficient the cleavage process. On the other hand, however, the active site of an HIV protease can hardly accommodate more than 8 residues. Consequently, for most cases, the protease-susceptible sites in proteins are strings of octapeptides as observed [135].

Thus, according to the "lock-and-key" mechanism in enzymology, an HIV protease-cleavable peptide must satisfy the substrate specificity, i.e., a good fit for binding to the active site. However, such a peptide, after a modification of its scissile bond with some chemical procedure, will completely lose its cleavability but it can

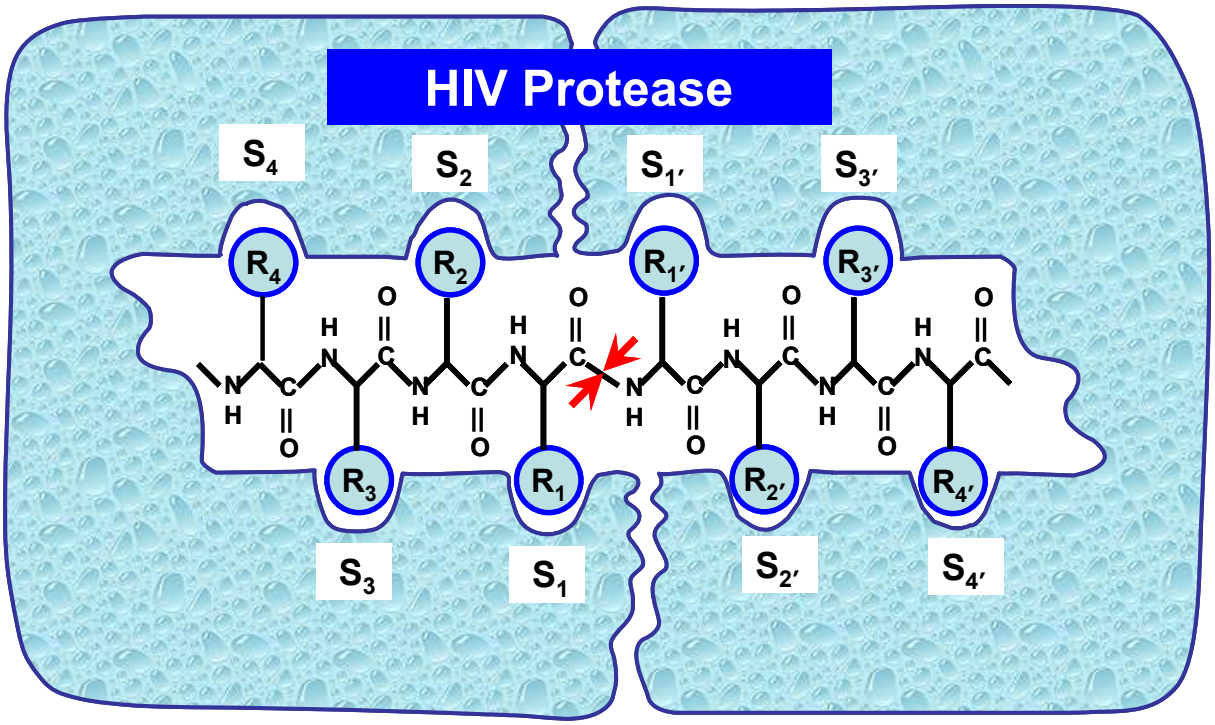

Figure 14. Schematic representation of substrate bound to HIV protease based an analysis of protease-inhibitor crystal structures. The active site of enzyme is composed of eight extended "subsites", S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, $S_{4'}$, and their counterparts in a substrate extend to an octapeptide region, sequentially symbolized by R_4, R_3, R_2, R_1, $R_{1'}$, $R_{1'}$, $R_{2'}$, $R_{3'}$, $R_{4'}$, respectively. The scissile bond is located between the subsites R_1 and $R_{1'}$. Reproduced with permission from Figure 3 of K.C. Chou [136].

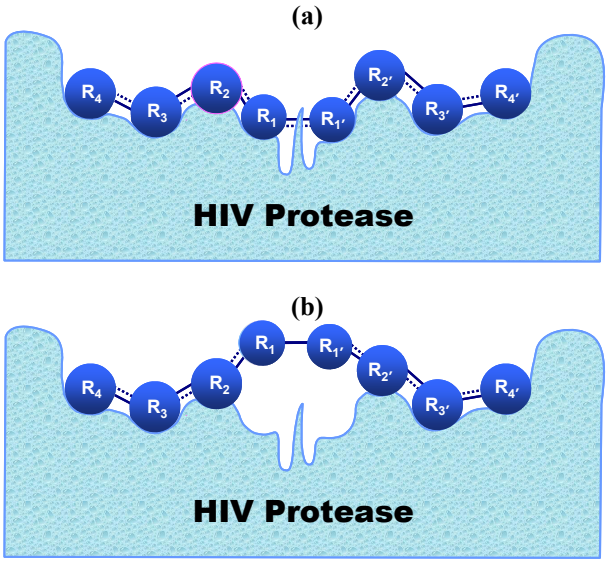

Figure 15. Schematic illustration to show (a) a cleavable octapeptide is chemically effectively bound to the active site of HIV protease, and (b) although still bound to the active site, the peptide has lost its cleavability after its scissile bond is modified from a hybrid peptide bond [254] to a single bond by some simple routine procedure. The eight residues of the peptide is sequentially symbolized R_4, R_3, R_2, R_1, $R_{1'}$, $R_{1'}$, $R_{2'}$, $R_{3'}$, $R_{4'}$. The scissile bond is located between R_1 and $R_{1'}$. Adapted from [136] with permission.

still bind to the active site of an enzyme. Actually, the molecule thus modified can be deemed as a "distorted key", which can be inserted into a lock but can neither open the lock nor be pulled out from it. That is why a molecule modified from a cleavable peptide can spontaneously become a competitive inhibitor against the enzyme. An illustration about such a concept is given in **Figure 15**, where panel **(a)** shows an effective binding of a cleavable peptide to the active site of HIV protease, while panel **(b)** shows that the peptide has become a non-cleavable one after its scissile bond is modified although it can still tightly bind to the active site. Such a modified peptide, or "distorted key", will automatically become an inhibitor candidate of HIV protease. Even for non-peptide inhibitors, it can also provide useful insights about the key binding groups, hydrophobic or hydrophilic environment, fitting conformation, et al. Accordingly, in search for the potential inhibitors, a matter of paramount importance is to discern what kind of peptides can be cleaved by HIV protease and what kind cannot be. Even if limited in the range of an octapeptide, it is by no means easy to address the question. This is because the number of possible octapeptides formed from 20 amino acids runs into $20^8 = 10^{8\log 20} \cong 2.56\times 10^{10}$. It would be exhausting to experimentally test out such an astronomical number of octapeptides. However, if one could find an effective computational method for predicting the cleavage sites in proteins by HIV protease,

the pace in search for the proper inhibitors of HIV protease would be significantly expedited. Actually, during the last decade or so, various prediction methods have been developed in this regard [128,135,137-139,180-186].

Recently, based on the discriminant function algorithm [136], a web server called **HIVcleave** [187] was established at the website http://chou.med.harvard.edu/bioinf/HIV/. For a given protein sequence, one can use **HIVcleave** to predict its cleavage sites by HIV-1 and HIV-2 proteases, respectively.

2.10. QuatIdent

As the chief actors of various biological processes in a cell, proteins have the following four different structural levels: primary, secondary, tertiary, and quaternary [188]. The primary structure refers to the constituent amino acid sequence; the secondary, to the local spatial arrangement of a polypeptide's backbone without regard to the conformations of its side chains; the tertiary, to the three-dimensional structure of an entire polypeptide; and the quaternary, to how many polypeptide chains (subunits) involved in forming a protein and the spatial arrangement of its subunits. The concept of quaternary structure is derived from the fact that many proteins are composed of two or more subunits which associate with each other through non-covalent interactions and, in some cases, disulfide bonds. According to the number of subunits aggregated together in an oligomeric complex, protein quaternary structures can be classified into: monomer, dimer, trimer, tetramer, pentamer, and so forth [189]. A statistical distribution of different quaternary structural types is shown in **Figure 16**, from which we can see that the nature prefers those oligomers with even and/or small number of subunits, fully consistent with the findings by the previous investigators [190,191]. If the subunits in a complex are identical, then the complex is called homo-oligomer; otherwise hetero-oligomer. For example, the sodium channel is formed by a monomer [192] while the potassium channel by a homo-tetramer [88]; the phospholamban is formed by homo-pentamer [93,193] while the Gamma-aminobutyric acid type A (GABAA) receptor by a hetero-pentamer [84,194]; the M2 proton channel is formed by a homo-tetramer [87] while hemoglobin by a hetero-tetramer [195].

Facing the explosion of newly generated protein sequences, we are challenged to develop an automated method for rapidly and reliably identify the quaternary structural attributes of uncharacterized proteins because they are closely relevant to the functions and mechanisms of proteins (see, e.g., [87,195]. Besides, the information thus obtained is very useful in screening the candidates of proteins for their 3D structure determination. It is known that many functionally important proteins exist in vivo as oligomers rather than single individual chains. For example, hemoglobin is a hetero-tetramer of two α chains and two β chains, and the four chains must be aggregated into one construct to perform its cooperative function during the oxygen-transporting process [195]. Also, the novel allosteric drug-inhibition mechanism for the M2 proton channel was recently revealed by the NMR observations [87,92]. It has been found through an in-depth analysis that such a subtle mechanism is closely correlated with a unique packing arrangement of four transmembrane helices from four identical protein chains [90,91,196]. For this kind of proteins, determination of their individual chains independently would be less interesting or should be avoided. Therefore, developing an effective method to predict the quaternary structural attributes of proteins based on their sequence information alone would provide useful clues for both basic research and drug development.

To address the challenge, the web-server predictor called "**QuatIdent**" [197] was developed recently by fusing the functional domain and sequential evolution information. **QuatIdent** is a 2-layer predictor. The 1st layer is for identifying a query protein as belonging to which one of the following ten main quaternary structural attributes: (1) monomer, (2) dimer, (3) trimer, (4) tetramer, (5) pentamer, (6) hexamer, (7) heptamer, (8) octamer, (9) decamer, and (10) dodecamer. If the result thus obtained turns out to be anything but monomer, the process will be automatically continued to further identify it belonging to a homo-oligomer or hetero-oligomer. **QuatIdent** is freely accessible to the public as a web server via the site at http://www.csbio.sjtu.edu.cn/bioinf/Quaternary/, by which one can get the desired 2-level results for a query protein sequence in around 25 seconds. And the longer the sequence is, the more time that is needed.

2.11. PQSA-Pred

This is another web-server predictor [198] developed by hybridizing the functional domain composition approach and pseudo amino acid composition approach for predicting protein quaternary structural attribute based on the sequence information alone. **PQSA-Pred** can be used to predict a query protein among the following three quaternary attributes according to its sequence information: monomer, homo-oligomer, and heterooligomer. As a useful tool for crystallographic scientists in screening for their targets, **PQSA-Pred** is freely accessible to the public via the website at http://218.65.61.89:8080/bioinfo/pqsa-pred .

Besides QuatIdent [197] and PQSA-Pred [198], some other efforts were also made in this regard [189,199,200]. However, none of these methods provide a web-server that can be easily used by the public.

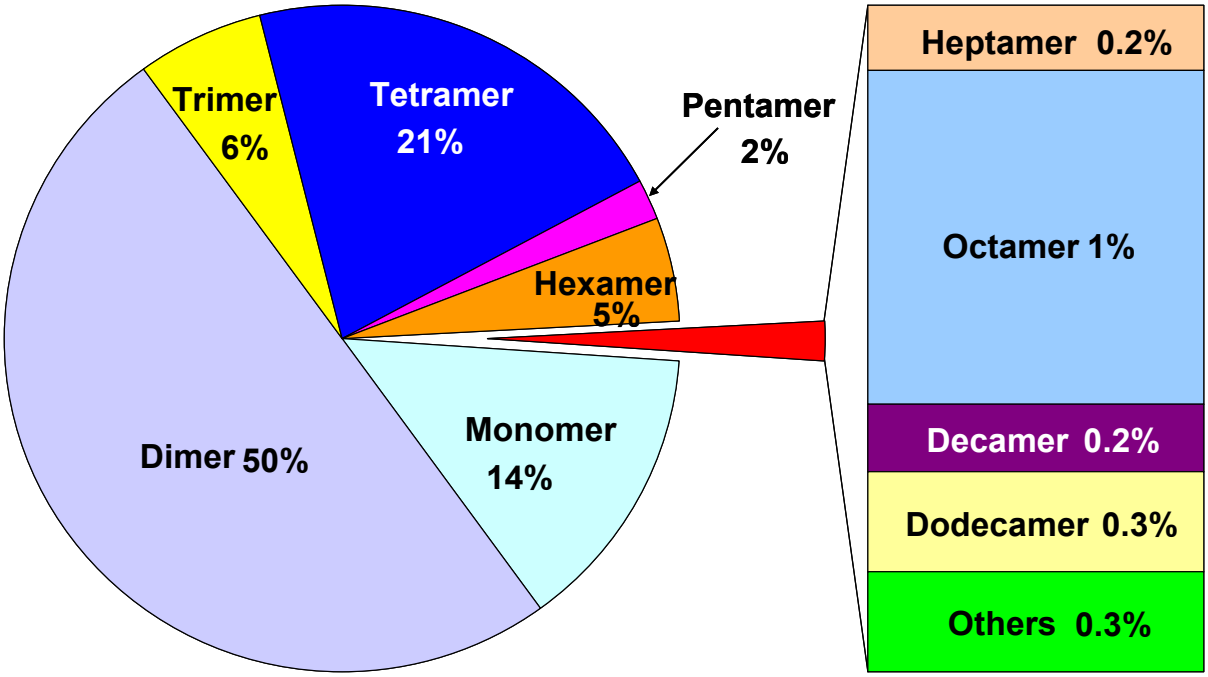

Figure 16. A pie chart to show the statistical distribution of different quaternary structural types in the nature derived from version 55.3 of Swiss-Prot database released 29-April-2008. Reproduced with permission from [197].

2.12. PFP-Pred

A protein can function properly only if it is folded into a very special and individual shape or conformation, i.e., has the correct secondary, tertiary and quaternary structure [201]. Failure to fold into the intended 3D structure usually produces inactive proteins or misfolded proteins [202] that may cause cell death and tissue damage [203] and be implicated in prion diseases such as bovine spongiform encephalopathy (BSE, also known as "mad cow disease") in cattle and Creutzfeldt-Jakob disease (CJD) in humans. All prion diseases are currently untreatable and are always fatal [204].

Although the X-ray crystallography is a powerful tool in determining protein 3D structures, it usually takes months or even years to determine the structure of a single protein. Also, the determination might fail for those proteins (particularly membrane proteins) that are difficult to crystallize. Although the nuclear magnetic resonance (NMR) technique is very powerful in determining membrane protein structures [87,93,94,148], it requires expensive equipments and take equally long or even longer time. The avalanche of protein sequences generated in the Post Genomic Age has challenged us for developing computational methods by which the structural information can be timely extracted from sequence databases. Although the direct prediction of the 3D structure of a protein from its sequence based on the least free energy principle [201,205] is scientifically quite sound and some encouraging results already obtained in elucidating the handedness problems and packing arrangements in proteins (see, e.g., [206-211]), it is far from successful yet for predicting its 3D structure owing to the notorious local minimum problem except for some very special cases or by utilizing some additional information from experiments (see, e.g., [212,213]). Actually, it is even not successful yet for simply predicting the overall fold of a query protein based on its sequence alone. For further information about protein folding, refer to a recent review [214] and the references cited therein. Again, although it is quite successful to predict the 3D structure of a protein according to the homology modeling approach [2,215] as reflected by a series of homology-modeled proteins for drug development [84,147,149-151,153,216-226], a hurdle exists when the query protein does not have any structure-known homologous protein in the existing databases [3].

Facing this kind of situation, a different strategy, the so-called taxonomic approach [227] was developed to address the problem. According to such a strategy, predicting the 3D structure of a protein may be first converted to a problem of classification; i.e., identifying which fold pattern it belongs to. Its underpinning is based on the assumption that the number of protein folds is limited [228-231].

The fold pattern of a protein is one level deeper than its structural classification [98,99,229], and hence is more challenging and complicated for prediction.

PFP-Pred [232] is one of these kinds of predictors. It was formed by a set of basic classifiers, with each trained in different parameter systems, such as predicted secondary structure, hydrophobicity, van der Waals volume, polarity, polarizability, as well as different dimensions of pseudo amino acid composition, that were extracted from a training dataset. The operation engine for the constituent individual classifiers was OET-KNN (Optimized Evidence-Theoretic K-Nearest Neighbors) rule [32,113,233]. Their outcomes were combined thru a weighted voting to give a final determination for classifying a query protein. The recognition was to find the true fold among the 27 possible patterns. The web-server of **PFP-Pred** is available to the public via the site http://chou.med.harvard.edu/bioinf/PFP-Pred/.

2.13. PFP-FunDSeqE

This is an improved version of **PFP-Pred** by combining the functional domain information and the sequential evolution information through a fusion ensemble classifier [234], as reflected by parts of its name where "FunD" stands for "functional domain" while "SeqE" for "sequential evolution". Compared with the other existing methods for predicting the protein fold patterns, **PFP-FunDSeqE** can usually yield better results [234]. Its web-server is available at http://www.csbio.sjtu.edu.cn/bioinf/PFP-FunDSeqE/.

2.14. Pred-PFR

Since each protein begins as a polypeptide translated from a sequence of mRNA as a linear chain of amino acids, it is interesting to study the folding rates of proteins from their primary sequences. Actually, protein chains can fold into the functional 3D structures with quite different rates, varying from several microseconds [235] to even an hour [236]. Since the 3D structure of a protein is determined by its primary sequence, we can assume the same is true for its folding rate. In view of this, we are challenged by an interesting question: Given a protein sequence, can we find its folding rate? Although the answer can be found by conducting various biochemical experiments, doing so is both time- consuming and expensive. Also, although a number of prediction methods were proposed [237-242], they need the input from the 3D structure of the protein concerned, and hence the prediction is feasible only after its 3D structure has been determined. However, according to data released on5-May-2009 by the RCSB Protein Data Bank (http://www.rcsb.org/pdb), the number of proteins with 3D structure known is only about 1.34% of the number of sequence-known proteins. Therefore, it is highly desired to develop an automated method that can rapidly and approximately predict the folding rates of proteins according to their sequence information alone. Some efforts have been made in this regard (see, e.g., [243,244]).

Since the experimentally observed folding rate for a protein chain usually represents the "apparent folding rate constant" [245] as denoted by K_{f}, it is instructive to unravel its relationship with the detailed rate constants, as given below.

The apparent folding rate constant K_{f} for a protein chain is defined via the following differential equation

$$\begin{cases} \dfrac{\mathrm{dP}_{\mathrm{unfold}}(t)}{\mathrm{d}t} = -K_{\mathrm{f}}\mathrm{P}_{\mathrm{unfold}}(t) \\ \dfrac{\mathrm{dP}_{\mathrm{fold}}(t)}{\mathrm{d}t} = K_{\mathrm{f}}\mathrm{P}_{\mathrm{unfold}}(t) \end{cases} \tag{6}$$

where $\mathrm{P}_{\mathrm{unfold}}(t)$ and $\mathrm{P}_{\mathrm{fold}}(t)$ represent the concentrations of its unfolded state and folded state, respectively. Suppose the total protein concentration is C_0, and initially only the unfolded protein is present; i.e., $\mathrm{P}_{\mathrm{unfold}}(t) = C_0$ and $\mathrm{P}_{\mathrm{fold}}(t) = 0$ when $t = 0$. Subsequently, the protein system is subjected to a sudden change in temperature, solvent, or any other factor that causes the protein to fold. Obviously, the solution for **Eq.6** is

$$\begin{cases} \mathrm{P}_{\mathrm{unfold}}(t) = C_0 \exp\left(-K_{\mathrm{f}}t\right) \\ \mathrm{P}_{\mathrm{fold}}(t) = C_0\left[1-\exp\left(-K_{\mathrm{f}}t\right)\right] \end{cases} \tag{7}$$

It can be seen from the above equation that the larger the K_{f}, the faster the folding rate will be. Given the value of K_{f}, the half-life of an unfolded protein chain can be expressed by

$$T_{1/2} = -\frac{\ln(1/2)}{K_{\mathrm{f}}} \cong 0.693/K_{\mathrm{f}} \tag{8}$$

which can also be used to reflect the time that is needed for a protein chain to be half folded. However, the actual folding process is much more complicated than the one as described by **Eq.6** even if the reverse rate for the folding system concerned can be ignored. As an illustration, let us consider the following three-state folding mechanism

$$\mathrm{P}_{\mathrm{unfold}} \xrightarrow{k_{12}} \mathrm{P}_{\mathrm{inter}} \xrightarrow{k_{23}} \mathrm{P}_{\mathrm{fold}} \tag{9}$$

where $\mathrm{P}_{\mathrm{inter}}(t)$ represents the concentration of an intermediate state between the unfolded and folded states, k_{12} is the rate constant for $\mathrm{P}_{\mathrm{unfold}}$ converting to $\mathrm{P}_{\mathrm{inter}}$, and k_{23} the rate constant for $\mathrm{P}_{\mathrm{inter}}$ converting to $\mathrm{P}_{\mathrm{fold}}$. Thus we have the following kinetic equation

$$\begin{cases} \dfrac{\mathrm{d}\mathrm{P}_{\mathrm{unfold}}(t)}{\mathrm{d}t} = -k_{12}\mathrm{P}_{\mathrm{unfold}}(t) \\ \dfrac{\mathrm{d}\mathrm{P}_{\mathrm{inter}}(t)}{\mathrm{d}t} = k_{12}\mathrm{P}_{\mathrm{unfold}}(t) - k_{23}\mathrm{P}_{\mathrm{inter}}(t) \\ \dfrac{\mathrm{d}\mathrm{P}_{\mathrm{fold}}(t)}{\mathrm{d}t} = k_{23}\mathrm{P}_{\mathrm{inter}}(t) \end{cases} \quad \textbf{(10)}$$

To get the solution of **Eq.10**, let us use an intuitive diagram called "directed graph" or "digraph" $\mathbb{G}$ **(Figure 17a)** [245,246] to represent **Eq.9**. To reflect the variation of the concentrations of the three protein states with time, the digraph $\mathbb{G}$ is further transformed to the phase digraph $\tilde{\mathbb{G}}$ [245,246] as shown in **Figure 17b**, where s is an interim parameter associated with the Laplace transform as shown in **Eq.11**.

$$\begin{cases} \tilde{\mathrm{P}}_{\mathrm{unfold}}(s) = \int_0^\infty \mathrm{P}_{\mathrm{unfold}}(t)\exp(-ts)\,\mathrm{d}t \\ \tilde{\mathrm{P}}_{\mathrm{inter}}(s) = \int_0^\infty \mathrm{P}_{\mathrm{inter}}(t)\exp(-ts)\,\mathrm{d}t \\ \tilde{\mathrm{P}}_{\mathrm{fold}}(s) = \int_0^\infty \mathrm{P}_{\mathrm{fold}}(t)\exp(-ts)\,\mathrm{d}t \end{cases} \quad \textbf{(11)}$$

where $\tilde{\mathrm{P}}_{\mathrm{unfold}}$, $\tilde{\mathrm{P}}_{\mathrm{inter}}$ and $\tilde{\mathrm{P}}_{\mathrm{fold}}$ are the phase concentrations of $\mathrm{P}_{\mathrm{unfold}}$, $\mathrm{P}_{\mathrm{inter}}$ and $\mathrm{P}_{\mathrm{fold}}$, respectively. Thus, according to the phase digraph $\tilde{\mathbb{G}}$ of **Figure 17b** and using the graphic rule 4 [245,246], which is also called the graphic rule for non-steady-state kinetics" in literatures (see, e.g., [247]), we can directly write out the following phase concentrations:

$$\tilde{\mathrm{P}}_{\mathrm{unfold}}(s) = \frac{(s+k_{23})sC_0}{s\left[(s+k_{23})s+k_{12}s+k_{12}k_{23}\right]} = \frac{(s+k_{23})C_0}{(s+k_{12})(s+k_{23})} = \frac{C_0}{s+k_{12}} \quad \textbf{(12.1)}$$

$$\tilde{\mathrm{P}}_{\mathrm{inter}}(s) = \frac{k_{12}s\ C_0}{s\left[(s+k_{23})s+k_{12}s+k_{12}k_{23}\right]} = \frac{k_{12}C_0}{(s+k_{12})(s+k_{23})} \quad \textbf{(12.2)}$$

$$\tilde{\mathrm{P}}_{\mathrm{fold}}(s) = \frac{k_{12}k_{23}C_0}{s\left[(s+k_{23})s+k_{12}s+k_{12}k_{23}\right]} = \frac{k_{12}k_{23}C_0}{s(s+k_{12})(s+k_{23})} \quad \textbf{(12.3)}$$

Through the above phase concentrations and using Laplace transform table (see, e.g., [248] or any standard mathematical tables), we can immediately obtain the desired concentrations for $\mathrm{P}_{\mathrm{unfold}}$, $\mathrm{P}_{\mathrm{inter}}$ and $\mathrm{P}_{\mathrm{fold}}$ of **Eq.10**, as given by **Eq.13**.

$$\begin{cases} \mathrm{P}_{\mathrm{unfold}}(t) = C_0\,\mathrm{e}^{-k_{12}t} \\ \mathrm{P}_{\mathrm{inter}}(t) = \dfrac{k_{12}C_0}{k_{23}-k_{12}}\left(\mathrm{e}^{-k_{12}t} - \mathrm{e}^{-k_{23}t}\right) \\ \mathrm{P}_{\mathrm{fold}}(t) = \dfrac{C_0}{k_{23}-k_{12}}\left(k_{12}\,\mathrm{e}^{-k_{23}t} - k_{23}\,\mathrm{e}^{-k_{12}t}\right) + C_0 \end{cases} \quad \textbf{(13)}$$

Accordingly, it follows from **Eq.13** that

$$\frac{\mathrm{d}\mathrm{P}_{\mathrm{fold}}(t)}{\mathrm{d}t} = \frac{k_{12}k_{23}C_0}{k_{23}-k_{12}}\left(\mathrm{e}^{-k_{12}t} - \mathrm{e}^{-k_{23}t}\right) = \frac{k_{12}k_{23}}{k_{23}-k_{12}}\left[1-\mathrm{e}^{-\left(k_{23}-k_{12}\right)t}\right]\mathrm{P}_{\mathrm{unfold}} \quad \textbf{(14)}$$

Comparing **Eq.14** with **Eq.6**, we obtain the following equivalent relation

$$K_{\mathrm{f}} \Leftrightarrow \frac{k_{12}k_{23}}{k_{23}-k_{12}}\left[1-\mathrm{e}^{-\left(k_{23}-k_{12}\right)t}\right] \quad \textbf{(15)}$$

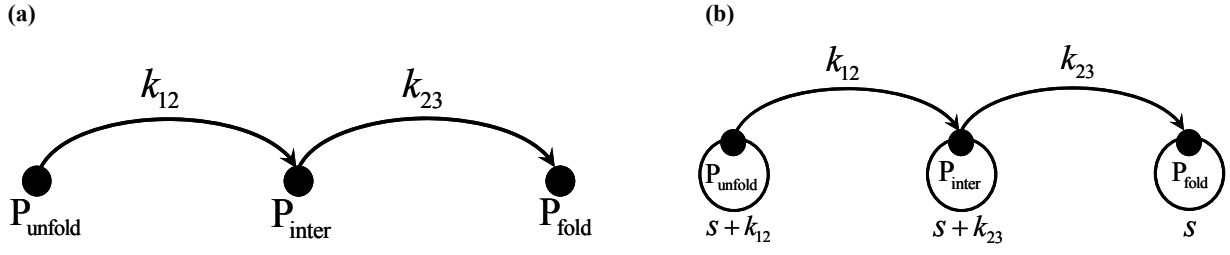

Figure 17. (a) The directed graph or digraph $\mathbb{G}$ [245,246] for the three-state protein folding mechanism as schematically expressed by **Eq.9** and formulated by **Eq.10**. **(b)** The phase digraph $\tilde{\mathbb{G}}$ obtained from $\mathbb{G}$ of panel (a) according to graphic rule 4 for enzyme and protein folding kinetics [245,246], where s is an interim parameter (see the text for further explanation).

meaning that the apparent folding rate constant K_f is a function of not only the detailed rate constants, but also t. Accordingly, K_f is actually not a constant but will change with time. Only when $k_{23} >> k_{12}$ and $k_{23} >> 1$, can **Eq.15** be reduced to $K_f \approx k_{12}$ and **Eq.14** to

$$\frac{dP_{folded}(t)}{dt} \approx k_{12}P_{unfold}(t) = K_f P_{unfold}(t) \quad (16)$$

and K_f be treated as a constant.

Even for a two-state protein folding system when the reverse effect needs to be considered, i.e., the system described by the following scheme and equation

$$P_{unfold} \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} P_{fold} \quad (17)$$

$$\begin{cases} \dfrac{dP_{unfold}(t)}{dt} = -k_{12}P_{unfold}(t) + k_{21}P_{fold}(t) \\ \dfrac{dP_{fold}(t)}{dt} = k_{12}P_{unfold}(t) - k_{21}P_{fold}(t) \end{cases} \quad (18)$$

where k_{21} represents the reverse rate constant converting P_{fold} back to P_{unfold}. With the similar derivation by using the non-steady state graphic rule [245,246] as described above, we can get the following equivalent relation [249]

$$K_f \Leftrightarrow \left\{ \frac{k_{12}(k_{12}+k_{21})}{k_{21}+k_{12}\exp\left[-(k_{12}+k_{21})t\right]} \exp\left[-(k_{12}+k_{21})t\right] \right\} \quad (19)$$

indicating that, even for the two-state folding system of **Eq.17**, the apparent folding rate constant K_f can be treated as a constant only when $k_{12} >> k_{21}$ and $k_{12} >> 1$.

It can be imagined that for a general multi-state folding system, K_f will be much more complicated. Consequently, all the experimental apparent folding rate constants were actually measured under some special conditions.

Recently, a web-server, called "**Pred-PFR**" (Predicting Protein Folding Rate), was developed for predicting the folding rate of a protein [249]. The predictor is featured by fusing multiple individual predictors, each of which is established based on one special feature derived from the protein sequence. As a user-friendly web-server, Pred-PFR is freely accessible to the public at www.csbio.sjtu.edu.cn/bioinf/FoldingRate/.

2.15. FoldRate

This is a different kind of protein folding rate predictor developed by fusing the folding-correlated features that can be either directly obtained or easily derived from the sequences of proteins [250]. **FoldRate** is freely accessible to the public at www.csbio.sjtu.edu.cn/bioinf/FoldRate/.

Both **Pred-PFR** and **FoldRate** can be used to predict the folding rate of a protein according to its sequence alone. The time by using the two web-server predictors to get the desired result for a query protein sequence is around 30 seconds. And the results obtained thus obtained are usually at least comparable with or even better than the existing methods that, however, need both the sequence and 3D structure information for prediction.

3. LIST OF WEB SERVERS

For reader's convenience, a brief description of each of the 15 web servers introduced in this article as well as its website address is given in **Table 3**.

4. CONCLUSION

Web-server is a newly emerging thing in the Internet Age. Technically speaking, a web-server means a computer program that is responsible for accepting HTTP (Hypertext Transfer Protocol) requests from clients. By means of web-servers, many computational prediction methods, regardless how difficult their mathematics or how complicated their algorithms are, can be easily used by the vast majority of scientists without the need to understand the mathematical details. Written as a laboratory protocol with a "recipe" style, the web-servers introduced here are user friendly and can be very easily used. Therefore, they are particularly useful for bench scientists to generate various data or information in a timely manner that they may need for their research projects.

It is anticipated that all these web-servers are constantly evolving with continuously improving the training datasets and prediction algorithms. To keep the users timely informed of the development, a short note will be published or an announcement will be placed in the relevant website.

Table 3. List of the 15 web servers introduced in this paper as well as their website addresses and targets.

No.	Name	Website address	Target
1	**Cell-PLoc package**	http://chou.med.harvard.edu/bioinf/Cell-PLoc/	Protein subcellular localization [49]
2	**Nuc-PLoc**	http://chou.med.harvard.edu/bioinf/Nuc-PLoc/	Protein subnuclear localization [63]
3	**Signal-CF**	http://chou.med.harvard.edu/bioinf/Signal-CF/	Protein signal peptide [79]
4	**Signal-3L**	http://chou.med.harvard.edu/bioinf/Signal-3L/	Protein signal peptide [82]
5	**MemType-2L**	http://chou.med.harvard.edu/bioinf/MemType/	Membrane protein type [54]
6	**EzyPred**	http://chou.med.harvard.edu/bioinf/EzyPred/	Enzyme functional class [126]
7	**ProtIdent**	http://www.csbio.sjtu.edu.cn/bioinf/Protease/	Protease type [55]
8	**GPCR-CA**	http://218.65.61.89:8080/bioinfo/GPCR-CA	GPCR type [160]
9	**HIVcleave**	http://chou.med.harvard.edu/bioinf/HIV/	HIV protease cleavage site [187]
10	**QuatIdent**	www.csbio.sjtu.edu.cn/bioinf/Quaternary/	Protein quaternary structural attribute [197]
11	**PQSA-Pred**	http://218.65.61.89:8080/bioinfo/pqsa-pred	Protein quaternary structural attribute [198]
12	**PFP-Pred**	http://www.csbio.sjtu.edu.cn/bioinf/PFP-Pred/	Protein fold pattern [232]
13	**PFP-FunDSeqE**	www.csbio.sjtu.edu.cn/bioinf/PFP-FunDSeqE/	Protein fold pattern [234]
14	**Pred-PFR**	www.csbio.sjtu.edu.cn/bioinf/FoldingRate/	Protein folding rate [249]
15	**FoldRate**	www.csbio.sjtu.edu.cn/bioinf/FoldRate/	Protein folding rate [250]

REFERENCES

[1] Chou, K.C. (2002) A new branch of proteomics: prediction of protein cellular attributes. In Weinrer, P. W. and Lu, Q. (eds.), *Gene Cloning & Expression Technologies, Chapter 4*. Eaton Publishing, Westborough, MA, pp. 57-70.

[2] Chou, K.C. (2004) Review: Structural bioinformatics and its impact to biomedical science. *Current Medicinal Chemistry*, **11,** 2105-2134.

[3] Chou, K.C. (2006) Structural bioinformatics and its impact to biomedical science and drug discovery. *Frontiers in Medicinal Chemistry*, **3,** 455-502.

[4] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994) *Molecular Biology of the Cell, chap.1*. 3rd ed. Garland Publishing, New York & London.

[5] Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaira, P. and Darnell, J. (1995) *Molecular Cell Biology , Chap.3*. 3rd ed. Scientific American Books, New York.

[6] Nakai, K. and Kanehisa, M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins: Structure, Function and Genetics*, **11,** 95-110.

[7] Nakashima, H. and Nishikawa, K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol*, **238,** 54-61.

[8] Cedano, J., Aloy, P., P'erez-Pons, J.A. and Querol, E. (1997) Relation between amino acid composition and cellular location of proteins. *J Mol Biol*, **266,** 594-600.

[9] Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Science*, **24,** 34-36.

[10] Chou, K.C. and Elrod, D.W. (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. *BBRC*, **252,** 63-68.

[11] Reinhardt, A. and Hubbard, T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research*, **26,** 2230-2236.

[12] Chou, K.C. and Elrod, D.W. (1999) Protein subcellular location prediction. *Protein Engineering*, **12,** 107-118.

[13] Yuan, Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Letters*, **451,** 23-26.

[14] Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry*, **54,** 277-344.

[15] Murphy, R.F., Boland, M.V. and Velliste, M. (2000) Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc Int Conf Intell Syst Mol Biol*, **8,** 251-259.

[16] Chou, K.C. (2000) Review: Prediction of protein structural classes and subcellular locations. *Current Protein and Peptide Science*, **1,** 171-208.

[17] Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300,** 1005-1016.

[18] Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol44, 60)*, **43,** 246-255.

[19] Feng, Z.P. (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, **58,** 491-499.

[20] Hua, S. and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17,** 721-728.

[21] Feng, Z.P. and Zhang, C.T. (2001) Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *Int J Biol Macromol*, **28,** 255-261.

[22] Feng, Z.P. (2002) An overview on predicting the subcellular location of a protein. *In Silico Biol*, **2,** 291-303.

[23] Chou, K.C. and Cai, Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem*, **277,** 45765-45769.

[24] Zhou, G.P. and Doctor, K. (2003) Subcellular location prediction of apoptosis proteins. *PROTEINS: Structure, Function, and Genetics*, **50,** 44-48.

[25] Pan, Y.X., Zhang, Z.Z., Guo, Z.M., Feng, G.Y., Huang, Z.D. and He, L. (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *Journal of Protein Chemistry*, **22,** 395-402.

[26] Park, K.J. and Kanehisa, M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. *Bioinformatics*, **19,** 1656-1663.

[27] Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K. *et al.* (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research*, **31,** 3613-3617.

[28] Huang, Y. and Li, Y. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, **20,** 21-28.

[29] Xiao, X., Shao, S., Ding, Y., Huang, Z., Huang, Y. and Chou, K.C. (2005) Using complexity measure factor to predict protein subcellular location. *Amino Acids*, **28,** 57-61.

[30] Gao, Y., Shao, S.H., Xiao, X., Ding, Y.S., Huang, Y.S., Huang, Z.D. and Chou, K.C. (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. *Amino Acids*, **28,** 373-376.

[31] Lei, Z. and Dai, Y. (2005) An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, **6,** 291.

[32] Shen, H.B. and Chou, K.C. (2005) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem Biophys Res Comm*, **337,** 752-756.

[33] Garg, A., Bhasin, M. and Raghava, G.P. (2005) Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J Biol Chem*, **280,** 14427-14432.

[34] Matsuda, S., Vert, J.P., Saigo, H., Ueda, N., Toh, H. and Akutsu, T. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci*, **14,** 2804-2813.

[35] Gao, Q.B., Wang, Z.Z., Yan, C. and Du, Y.H. (2005) Prediction of protein subcellular location using a combined feature of sequence. *FEBS Lett*, **579,** 3444-3448.

[36] Chou, K.C. and Shen, H.B. (2006) Predicting protein subcellular location by fusing multiple classifiers. *Journal of Cellular Biochemistry*, **99,** 517-527.

[37] Guo, J., Lin, Y. and Liu, X. (2006) GNBSL: A new integrative system to predict the subcellular location for Gram-negative bacteria proteins. *Proteomics*, **6,** 5099-5105.

[38] Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D. and Chou, K.C. (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids*, **30,** 49-54.

[39] Hoglund, A., Donnes, P., Blum, T., Adolph, H.W. and Kohlbacher, O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22,** 1158-1165.

[40] Lee, K., Kim, D.W., Na, D., Lee, K.H. and Lee, D. (2006) PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic Acids Res*, **34,** 4655-4666.

[41] Zhang, Z.H., Wang, Z.H., Zhang, Z.R. and Wang, Y.X. (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on

grouped weight and support vector machine. *FEBS Lett*, **580,** 6169-6174.
[42] Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.-M. and Xie, J. (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids*, **33,** 69-74.
[43] Chou, K.C. and Shen, H.B. (2007) Large-scale plant protein subcellular location prediction. *Journal of Cellular Biochemistry*, **100,** 665-678.
[44] Shen, H.B. and Chou, K.C. (2007) Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem Biophys Res Commun*, **355,** 1006-1011.
[45] Shen, H.B., Yang, J. and Chou, K.C. (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, **33,** 57-67.
[46] Chen, Y.L. and Li, Q.Z. (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *Journal of Theoretical Biology*, **248,** 377–381.
[47] Chen, Y.L. and Li, Q.Z. (2007) Prediction of the subcellular location of apoptosis proteins. *Journal of Theoretical Biology*, **245,** 775-783.
[48] Mundra, P., Kumar, M., Kumar, K.K., Jayaraman, V.K. and Kulkarni, B.D. (2007) Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recognition Letters*, **28,** 1610-1615.
[49] Chou, K.C. and Shen, H.B. (2007) Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry*, **370,** 1-16.
[50] Chou, K.C. and Shen, H.B. (2008) Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols*, **3,** 153-162.
[51] Chou, K.C. and Shen, H.B. (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *Journal of Proteome Research*, **6,** 1728-1734.
[52] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25,** 25-29.
[53] Chou, K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21,** 10-19.
[54] Chou, K.C. and Shen, H.B. (2007) MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Comm*, **360,** 339-345.
[55] Chou, K.C. and Shen, H.B. (2008) ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem Biophys Res Comm*, **376,** 321-325.
[56] Shen, H.B. and Chou, K.C. (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Analytical Biochemistry*, in press.
[57] Shen, H.B. and Chou, K.c. (2009) Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. *Protein & Peptide Letters*, submitted.
[58] Shen, H.B. and Chou, K.C. (2009) Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins, to be submitted.
[59] Chou, K.C. and Shen, H.B. (2009) Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization, to be submitted.
[60] Shen, H.B. and Chou, K.C. (2007) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Engineering, Design, and Selection*, **20,** 39-46.
[61] Chou, K.C. and Shen, H.B. (2006) Large-scale predictions of Gram-negative bacterial protein subcellular locations. *Journal of Proteome Research*, **5,** 3420-3428.
[62] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) *Molecular biology of the cell, 4th edition*. Garland Science, New York.
[63] Shen, H.B. and Chou, K.C. (2007) Nuc-PLoc: A new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Engineering, Design & Selection*, **20,** 561-567.
[64] Rapoport, T.A. (1992) Transport of proteins across the endoplasmic reticulum membrane. *Science*, **258,** 931-936.
[65] Zheng, N. and Gierasch, L.M. (1996) Signal sequences: the same yet different. *Cell*, **86,** 849-852.
[66] Chou, K.C. (2001) Prediction of signal peptides using scaled window. *Peptides*, **22,** 1973-1979.
[67] McGeoch, D.J. (1985) On the predictive recognition of signal peptide sequences. *Virus Res*, **3,** 271-286.
[68] von Heijne, G. (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research*, **14,** 4683-4690.
[69] Folz, R.J. and Gordon, J.I. (1987) Computer-assisted predictions of signal peptidase processing sites. *Biochem Biophys Res Comm*, **146,** 870-877.
[70] Ladunga, I., Czako, F., Csabai, I. and Geszti, T. (1991) Improving signal peptide prediction accuracy by simulated neural network. *Comput Appl Biosci*, **7,** 485-487.
[71] Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A. and Damiani, G. (1991) Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. *Comput Appl Biosci*, **7,** 353-357.
[72] Schneider, G. and Wrede, P. (1993) Signal analysis of protein targeting sequences. *Protein Seq Data Anal*, **5,** 227-236.
[73] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, **10,** 1-6.
[74] Emanuelsson, O., Nielsen, H. and von Heijne, G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, **8,** 978-984.
[75] Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340,** 783-795.
[76] Hiller, K., Grote, A., Scheer, M., Munch, R. and Jahn, D. (2004) PrediSi: prediction of signal peptides and their

cleavage positions. *Nucleic Acids Res*, **32,** W375-379.
[77] Chou, K.C. (2002) Review: Prediction of protein signal sequences. *Current Protein and Peptide Science*, **3,** 615-622.
[78] Chou, K.C. (2001) Using subsite coupling to predict signal peptides. *Protein Engineering*, **14,** 75-79.
[79] Chou, K.C. and Shen, H.B. (2007) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Comm*, **357,** 633-640.
[80] Hiss, J.A. and Schneider, G. (2009) Architecture, function and prediction of long signal peptides. *Brief Bioinform*, **10**, 569-578.
[81] Kall, L., Krogh, A. and Sonnhammer, E.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res*, **35,** W429-432.
[82] Shen, H.B. and Chou, K.C. (2007) Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem Biophys Res Comm*, **363,** 297-303.
[83] Reynolds, S.M., Kall, L., Riffle, M.E., Bilmes, J.A. and Noble, W.S. (2008) Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol*, **4,** e1000213.
[84] Chou, K.C. (2004) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochemical and Biophysical Research Communications*, **316,** 636-642.
[85] Chou, K.C. (1993) Conformational change during photocycle of bacteriorhodopsin and its proton-pumping mechanism. *Journal of Protein Chemistry*, **12,** 337-350.
[86] Chou, K.C. (1994) Mini Review: A molecular piston mechanism of pumping protons by bacteriorhodopsin. *Amino Acids*, **7,** 1-17.
[87] Schnell, J.R. and Chou, J.J. (2008) Structure and mechanism of the M2 proton channel of influenza A virus. *Nature*, **451,** 591-595.
[88] Doyle, D.A., Morais, C.J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T. and MacKinnon, R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280,** 69-77.
[89] Chou, K.C. (2004) Insights from modelling three-dimensional structures of the human potassium and sodium channels. *Journal of Proteome Research*, **3,** 856-861.
[90] Huang, R.B., Du, Q.S., Wang, C.H. and Chou, K.C. (2008) An in-depth analysis of the biological functional studies based on the NMR M2 channel structure of influenza A virus. *Biochem Biophys Res Comm*, **377,** 1243-1247.
[91] Du, Q.S., Huang, R.B., Wang, C.H., Li, X.M. and Chou, K.C. (2009) Energetic analysis of the two controversial drug binding sites of the M2 proton channel in influenza A virus. *Journal of Theoretical Biology*, **259,** 159-164.
[92] Pielak, R.M., Jason R. Schnell, J.R. and Chou, J.J. (2009) Mechanism of drug inhibition and drug resistance of influenza A M2 channel. *Proceedings of National Academy of Science, USA*, **106,** 7379-7384.
[93] Oxenoid, K. and Chou, J.J. (2005) The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc Natl Acad Sci U S A*, **102,** 10870-10875.
[94] Douglas, S.M., Chou, J.J. and Shih, W.M. (2007) DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. *Proc Natl Acad Sci U S A*, **104,** 6644-6648.
[95] Nakashima, H., Nishikawa, K. and Ooi, T. (1986) The folding type of a protein is relevant to the amino acid composition. *J Biochem*, **99,** 152-162.
[96] Klein, P. and Delisi, C. (1986) Prediction of protein structural class from amino acid sequence. *Biopolymers*, **25,** 1659-1672.
[97] Klein, P. (1986) Prediction of protein structural class by discriminant analysis. *Biochim Biophys Acta*, **874,** 205-215.
[98] Chou, K.C. and Zhang, C.T. (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J Biol Chem*, **269,** 22014-22020.
[99] Chou, K.C. (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins: Structure, Function & Genetics*, **21,** 319-344.
[100] Liu, W. and Chou, K.C. (1998) Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. *Journal of Protein Chemistry*, **17,** 209-217.
[101] Chou, K.C., Liu, W., Maggiora, G.M. and Zhang, C.T. (1998) Prediction and classification of domain structural classes. *PROTEINS: Structure, Function, and Genetics*, **31,** 97-103.
[102] Chou, K.C. and Maggiora, G.M. (1998) Domain structural class prediction. *Protein Engineering*, **11,** 523-538.
[103] Chou, K.C. (1999) A key driving force in determination of protein structural classes. *Biochemical and Biophysical Research Communications*, **264,** 216-224.
[104] Chou, K.C. and Elrod, D.W. (1999) Prediction of membrane protein types and subcellular locations. *PROTEINS: Structure, Function, and Genetics*, **34,** 137-153.
[105] Cai, Y.D., Liu, X.J. and Chou, K.C. (2001) Artificial neural network model for predicting membrane protein types. *Journal of Biomolecular Structure and Dynamics*, **18,** 607-610.
[106] Guo, Z.M. (2002) Prediction of Membrane protein types by using pattern recognition method based on pseudo amino acid composition. *Master Thesis, Bio-X Life Science Research Center, Shanghai Jiaotong University*.
[107] Cai, Y.D., Zhou, G.P. and Chou, K.C. (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical Journal*, **84,** 3257-3263.
[108] Cai, Y.D., Pong-Wong, R., Feng, K., Jen, J.C.H. and Chou, K.C. (2004) Application of SVM to predict membrane protein types. *Journal of Theoretical Biology*, **226,** 373-376.
[109] Wang, M., Yang, J., Liu, G.P., Xu, Z.J. and Chou, K.C. (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. *Protein Engineering, Design, and Selection*, **17,** 509-516.
[110] Chou, K.C. and Cai, Y.D. (2005) Prediction of membrane protein types by incorporating amphipathic effects. *Journal of Chemical Information and Modeling*, **45,** 407-413.

[111] Liu, H., Wang, M. and Chou, K.C. (2005) Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem Biophys Res Commun*, **336,** 737-739.
[112] Wang, M., Yang, J., Xu, Z.J. and Chou, K.C. (2005) SLLE for predicting membrane protein types. *Journal of Theoretical Biology*, **232,** 7-15.
[113] Shen, H.B. and Chou, K.C. (2005) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochemical & Biophysical Research Communications*, **334,** 288-292.
[114] Shen, H.B., Yang, J. and Chou, K.C. (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *Journal of Theoretical Biology*, **240,** 9-13.
[115] Wang, S.Q., Yang, J. and Chou, K.C. (2006) Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. *Journal of Theoretical Biology*, **242,** 941-946.
[116] Shen, H.B. and Chou, K.C. (2007) Using ensemble classifier to identify membrane protein types. *Amino Acids*, **32,** 483-488.
[117] Yang, X.G., Luo, R.Y. and Feng, Z.P. (2007) Using amino acid and peptide composition to predict membrane protein types. *Biochem Biophys Res Commun*, **353,** 164-169.
[118] Pu, X., Guo, J., Leung, H. and Lin, Y. (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. *J Theor Biol*, **247,** 259–265.
[119] Afjehi-Sadat, L. and Lubec, G. (2007) Identification of enzymes and activity from two-dimensional gel electrophoresis. *Nature Protocols*, **2,** 2318-2324.
[120] Chou, K.C. and Elrod, D.W. (2003) Prediction of enzyme family classes. *Journal of Proteome Research*, **2,** 183-190.
[121] Chou, K.C. and Cai, Y.D. (2004) Predicting enzyme family class in a hybridization space. *Protein Science*, **13,** 2857-2863.
[122] Cai, C.Z., Han, L.Y., Ji, Z.L. and Chen, Y.Z. (2004) Enzyme family classification by support vector machines. *PROTEINS: Structure, Function, and Bioinformatics*, **55,** 66-76.
[123] 1Cai, Y.D. and Chou, K.C. (2005) Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *Journal of Proteome Research*, **4,** 967-971.
[124] Huang, W.L., Chen, H.M., Hwang, S.F. and Ho, S.Y. (2006) Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. *Biosystems*, **90,** 405-413.
[125] Zhou, X.B., Chen, C., Li, Z.C. and Zou, X.Y. (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *Journal of Theoretical Biology*, **248,** 546–551.
[126] Shen, H.B. and Chou, K.C. (2007) EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem Biophys Res Comm*, **364,** 53-59.
[127] Bairoch, A. (2000) The ENZYME Database in 2000. *Nucleic Acids Research*, **28,** 304-305.
[128] Poorman, R.A., Tomasselli, A.G., Heinrikson, R.L. and Kezdy, F.J. (1991) 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,** 14554-14561.
[129] Qin, H., Srinvasula, S.M., Wu, G., Fernandes-Alnemri, T., Alnemri, E.S., and Shi, Y. (1999) Structural basis of procaspase-9 recruitment by the apoptotic protease-activating factor 1. *Nature*, **399,** 549-557.
[130] Chou, J.J., Li, H., Salvessen, G.S., Yuan, J. and Wagner, G. (1999) Solution structure of BID, an intracellular amplifier of apoptotic signalling. *Cell*, **96,** 615-624.
[131] Watt, W., Koeplinger, K.A., Mildner, A.M., Heinrikson, R.L., Tomasselli, A.G. and Watenpaugh, K.D. (1999) The atomic resolution structure of human caspase-8, a key activator of apoptosis. *Structure*, **7,** 1135-1143.
[132] Chou, K.C., Wei, D.Q. and Zhong, W.Z. (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: ibid., 2003, Vol.310, 675). *Biochem Biophys Res Comm*, **308,** 148-151.
[133] Puente, X.S., Sanchez, L.M., Overall, C.M. and Lopez-Otin, C. (2003) Human and mouse proteases: a comparative genomic approach. *Nat Rev Genet*, **4,** 544-558.
[134] Chou, K.C., Wei, D.Q., Du, Q.S., Sirois, S., Shen, H.B. and Zhong, W.Z. (2009) Study of inhibitors against SARS coronavirus by computational approaches. In Lendeckel, U. and Hooper, N. (eds.), Viral proteases and antiviral protease inhibitor therapy. *Proteases in Biology and Disease*, Springer Publishing, **8.**
[135] Chou, K.C. (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J Biol Chem*, **268,** 16938-16948.
[136] Chou, K.C. (1996) Review: Prediction of HIV protease cleavage sites in proteins. *Analytical Biochemistry*, **233,** 1-14.
[137] You, L., Garwicz, D. and Rognvaldsson, T. (2005) Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease. *J Virol*, **79,** 12477-12486.
[138] Rognvaldsson, T., You, L. and Garwicz, D. (2007) Bioinformatic approaches for modeling the substrate specificity of HIV-1 protease: an overview. *Expert Rev Mol Diagn*, **7,** 435-451.
[139] Liang, G.Z. and Li, S.Z. (2007) A new sequence representation as applied in better specificity elucidation for human immunodeficiency virus type 1 protease. *Biopolymers*, **88,** 401-412.
[140] Rawlings, N.D., Tolle, D.P. and Barrett, A.J. (2004) MEROPS: the peptidase database. *Nucleic Acids Research*, **32,** D160-D164.
[141] Chou, K.C. and Cai, Y.D. (2006) Prediction of protease types in a hybridization space. *Biochem Biophys Res Comm*, **339,** 1015-1020.
[142] Zhou, G.P. and Cai, Y.D. (2006) Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. *PROTEINS: Structure, Function, and Bioinformatics*, **63,** 681-684.
[143] Shen, H.B. and Chou, K.C. (2009) Identification of proteases and their types. *Analytical Biochemistry*, **385,** 153-160.
[144] Heuss, C. and Gerber, U. (2000) G-protein-independent

signaling by G-protein-coupled receptors. *Trends Neurosci*, **23,** 469-475.
[145] Milligan, G. and White, J.H. (2001) Protein-protein interactions at G-protein-coupled receptors. *Trends Pharmacol Sci*, **22,** 513-518.
[146] Hall, R.A. and Lefkowitz, R.J. (2002) Regulation of G protein-coupled receptor signaling by scaffold proteins. *Circ Res*, **91,** 672-680.
[147] Chou, K.C. (2005) Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. *Journal of Proteome Research*, **4,** 1681-1686.
[148] Call, M.E., Schnell, J.R., Xu, C., Lutz, R.A., Chou, J.J. and Wucherpfennig, K.W. (2006) The structure of the zetazeta transmembrane dimer reveals features essential for its assembly with the T cell receptor. *Cell*, **127,** 355-368.
[149] Chou, K.C. (2004) Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. *Biochemical and Biophysical Research Communication*, **319,** 433-438.
[150] Chou, K.C. (2004) Molecular therapeutic target for type-2 diabetes. *Journal of Proteome Research*, **3,** 1284-1288.
[151] Wei, D.Q., Du, Q.S., Sun, H. and Chou, K.C. (2006) Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. *Biochem Biophys Res Comm*, **344,** 1048-1055.
[152] Wang, S.Q., Du, Q.S. and Chou, K.C. (2007) Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases. *Biochem Biophys Res Comm*, **354,** 634-640.
[153] Wang, S.Q., Du, Q.S., Huang, R.B., Zhang, D.W. and Chou, K.C. (2009) Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus. *Biochem Biophys Res Commun*, **386,** 432-436.
[154] Elrod, D.W. and Chou, K.C. (2002) A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Engineering*, **15,** 713-715.
[155] Chou, K.C. and Elrod, D.W. (2002) Bioinformatical analysis of G-protein-coupled receptors. *Journal of Proteome Research*, **1,** 429-433.
[156] Bhasin, M. and Raghava, G.P. (2005) GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. *Nucleic Acids Research*, **33,** W143-147.
[157] Chou, K.C. (2005) Prediction of G-protein-coupled receptor classes. *Journal of Proteome Research*, **4,** 1413-1418.
[158] Wen, Z., Li, M., Li, Y., Guo, Y. and Wang, K. (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. *Amino Acids*, **32,** 277-283.
[159] Gao, Q.B. and Wang, Z.Z. (2006) Classification of G-protein coupled receptors at four levels. *Protein Eng Des Sel*, **19,** 511-516.
[160] Xiao, X., Wang, P. and Chou, K.C. (2009) GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *Journal of Computational Chemistry*, **30,** 1414-1423.
[161] Wolfram, S. (1984) Cellular automation as models of complexity. *Nature*, **311,** 419-424.
[162] Wolfram, S. (2002) *A New Kind of Science*. Wolfram Media Inc., Champaign, IL.
[163] Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X. and Chou, K.C. (2005) Using cellular automata to generate Image representation for biological sequences. *Amino Acids*, **28,** 29-35.
[164] Chou, K.C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochemical & Biophysical Research Communications*, **278,** 477-483.
[165] Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J Biol Chem*, **268,** 6119-6124.
[166] Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *J Biol Chem*, **268,** 14875-14880.
[167] Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry*, **32,** 6548-6554.
[168] Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1994) Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. *Experientia*, **50,** 23-28.
[169] Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G. *et al.* (1994) Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-90152E. *Biochemical Pharmacology*, **47,** 2017-2028.
[170] Althaus, I.W., Chou, K.C., Franks, K.M., Diebel, M.R., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1996) The benzylthio-pyrididine U-31,355 is a potent inhibitor of HIV-1 reverse transcriptase. *Biochemical Pharmacology*, **51,** 743-750.
[171] Chou, K.C., Kezdy, F.J. and Reusser, F. (1994) Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Analytical Biochemistry*, **221,** 217-230.
[172] McQuade, T.J., Tomasselli, A.G., Liu, L., Karacostas, V., Moss, B., Sawyer, T.K., Heinrikson, R.L. and Tarpley, W.G. (1990) A synthetic HIV-1 protease inhibitor with antiviral activity arrests HIV-like particle maturation. *Science*, **247,** 454-456.
[173] Meek, T.D., Lambert, D.M., Dreyer, G.B., Carr, T.J., Tomaszek, T.A., Jr., Moore, M.L., Strickler, J.E., Debouck, C., Hyland, L.J., Matthews, T.J. *et al.* (1990) Inhibition of HIV-1 protease in infected T-lymphocytes by synthetic peptide analogues. *Nature*, **343,** 90-92.
[174] Wlodawer, A. and Erickson, J.W. (1993) Structure-based inhibitors of HIV-1 protease. *Annu Rev Biochem*, **62,** 543-585.

[175] Barre-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C. *et al.* (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, **220,** 868-871.

[176] Gallo, R.C., Salahuddin, S.Z., Popovic, M., Shearer, G.M., Kaplan, M., Haynes, B.F., Palker, T.J., Redfield, R., Oleske, J., Safai, B. *et al.* (1984) Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, **224,** 500-503.

[177] Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B. and Wlodawer, A. (1989) Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. *Science*, **246,** 1149-1152.

[178] Schechter, I. and Berger, A. (1967) On the size of the active site in protease. I. Papain. *Biochem Biophys Res Comm*, **27,** 157-162.

[179] Chou, K.C., Chen, N.Y. and Forsen, S. (1981) The biological functions of low-frequency phonons: 2. Cooperative effects. *Chemica Scripta*, **18,** 126-132.

[180] Chou, K.C., Zhang, C.T. and Kezdy, F.J. (1993) A vector approach to predicting HIV protease cleavage sites in proteins. *Proteins: Structure, Function, and Genetics*, **16,** 195-204.

[181] Chou, J.J. (1993) Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. *Journal of Protein Chemistry*, **12,** 291-302.

[182] Chou, K.C. and Zhang, C.T. (1993) Studies on the specificity of HIV protease: an application of Markov chain theory. *Journal of Protein Chemistry*, **12,** 709-724.

[183] Chou, J.J. (1993) A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. *Biopolymers*, **33,** 1405-1414.

[184] Zhang, C.T. and Chou, K.C. (1993) An alternate-subsite-coupled model for predicting HIV protease cleavage sites in proteins. *Protein Engineering*, **7,** 65-73.

[185] Thompson, T.B., Chou, K.C. and Zheng, C. (1995) Neural network prediction of the HIV-1 protease cleavage sites. *Journal of Theoretical Biology 177***,** 369-379.

[186] Chou, K.C., Tomasselli, A.L., Reardon, I.M. and Heinrikson, R.L. (1996) Predicting HIV protease cleavage sites in proteins by a discriminant function method. *PROTEINS: Structure, Function, and Genetics*, **24,** 51-72.

[187] Shen, H.B. and Chou, K.C. (2008) HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. *Analytical Biochemistry*, **375,** 388-390.

[188] Klotz, I.M., Darnell, D.W. and Langerman, N.R. (1975) Quaternary structure of proteins. In Neurath, H. and Hill, R. L. (eds.), *The Proteins (3rd ed)*. Academic Press, New York, **1**, 226-241.

[189] Chou, K.C. and Cai, Y.D. (2003) Predicting protein quaternary structure by pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics*, **53,** 282-289.

[190] Goodsell, D.S. and Olson, A.J. (2000) Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*, **29,** 105-153.

[191] Levy, E.D., Boeri Erba, E., Robinson, C.V. and Teichmann, S.A. (2008) Assembly reflects evolution of protein complexes. *Nature*, **453,** 1262-1265.

[192] Chen, Z., Alcayaga, C., Suarez-Isla, B.A., O'Rourke, B., Tomaselli, G. and Marban, E. (2002) A "minimal" sodium channel construct consisting of ligated S5-P-S6 segments forms a toxin-activatable ionophore. *J Biol Chem*, **277,** 24653-24658.

[193] Oxenoid, K., Rice, A.J. and Chou, J.J. (2007) Comparing the structure and dynamics of phospholamban pentamer in its unphosphorylated and pseudo-phosphorylated states. *Protein Sci*, **16,** 1977-1983.

[194] Tretter, V., Ehya, N., Fuchs, K. and Sieghart, W. (1997) Stoichiometry and assembly of a recombinant GABAA receptor subtype. *Journal of Neuroscience*, **17,** 2728-2737.

[195] Perutz, M.F. (1964) The Hemoglobin Molecule. *Scientific American*, **211,** 65-76.

[196] Wei, H., Wang, C.H., Du, Q.S., Meng, J. and Chou, K.C. (2009) Investigation into adamantane-based M2 inhibitors with FB-QSAR. *Medicinal Chemistry*, **5,** 305-317.

[197] Shen, H.B. and Chou, K.C. (2009) QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *Journal of Proteome Research*, **8,** 1577–1584.

[198] Xiao, X., Wang, P. and Chou, K.C. (2009) Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. *Journal of Applied Crystallography*, **42,** 169-173.

[199] Garian, R. (2001) Prediction of quaternary structure from primary structure. *Bioinformatics*, **17,** 551-556.

[200] Zhang, S.W., Chen, W., Yang, F. and Pan, Q. (2008) Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. *Amino Acids*, **35,** 591-598.

[201] Anfinsen, C.B. and Scheraga, H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv Protein Chem*, **29,** 205-300.

[202] Aguzzi, A. (2008) Unraveling prion strains with cell biology and organic chemistry. *Proc Natl Acad Sci U S A*, **105,** 11-12.

[203] Dobson, C.M. (2001) The structural basis of protein folding and its links with human disease. *Philos Trans R Soc Lond B Biol Sci*, **356,** 133-145.

[204] Prusiner, S.B. (1998) Prions. *Proc Natl Acad Sci U S A*, **95,** 13363-13383.

[205] Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181,** 223-230.

[206] Chou, K.C. and Scheraga, H.A. (1982) Origin of the right-handed twist of beta-sheets of poly-L-valine chains. *Proceedings of National Academy of Sciences, USA*, **79,** 7047-7051.

[207] Chou, K.C., Maggiora, G.M., Nemethy, G. and Scheraga, H.A. (1988) Energetics of the structure of the four-alpha-helix bundle in proteins. *Proceedings of National Academy of Sciences, USA*, **85,** 4295-4299.

[208] Chou, K.C., Nemethy, G. and Scheraga, H.A. (1990) Review: Energetics of interactions of regular structural elements in proteins. *Accounts of Chemical Research*, **23,** 134-141.

[209] Chou, K.C., Nemethy, G. and Scheraga, H.A. (1984) Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. *Journal of American Chemical Society*, **106,** 3161-3170.
[210] Chou, K.C., Nemethy, G., Pottle, M. and Scheraga, H.A. (1989) Energy of stabilization of the right-handed beta-alpha-beta crossover in proteins. *Journal of Molecular Biology*, **205,** 241-249.
[211] Chou, K.C. and Carlacci, L. (1991) Energetic approach to the folding of alpha/beta barrels. *Proteins: Structure, Function, and Genetics*, **9,** 280-295.
[212] Chou, K.C. (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. *Journal of Molecular Biology*, **223,** 509-517.
[213] Carlacci, L., Chou, K.C. and Maggiora, G.M. (1991) A heuristic approach to predicting the tertiary structure of bovine somatotropin. *Biochemistry*, **30,** 4389-4398.
[214] Scheraga, H.A., Khalili, M. and Liwo, A. (2007) Protein-folding dynamics: overview of molecular simulation techniques. *Annu Rev Phys Chem*, **58,** 57-83.
[215] Holm, L. and Sander, C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Research*, **27,** 244-247.
[216] Chou, K.C. (1995) The convergence-divergence duality in lectin domains of the selectin family and its implications. *FEBS Letters*, **363,** 123-126.
[217] Chou, K.C., Jones, D. and Heinrikson, R.L. (1997) Prediction of the tertiary structure and substrate binding site of caspase-8. *FEBS Letters*, **419,** 49-54.
[218] Chou, J.J., Matsuo, H., Duan, H. and 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.
[219] Chou, K.C., Tomasselli, A.G. and Heinrikson, R.L. (2000) Prediction of the Tertiary Structure of a Caspase-9/Inhibitor Complex. *FEBS Letters*, **470,** 249-256.
[220] Chou, K.C. and Howe, W.J. (2002) Prediction of the tertiary structure of the beta-secretase zymogen. *BBRC*, **292,** 702-708.
[221] Du, Q.S., Wang, S., Wei, D.Q., Sirois, S. and Chou, K.C. (2005) Molecular modelling and chemical modification for finding peptide inhibitor against SARS CoV Mpro. *Analytical Biochemistry*, **337,** 262-270.
[222] Zhang, R., Wei, D.Q., Du, Q.S. and Chou, K.C. (2006) Molecular modeling studies of peptide drug candidates against SARS. *Medicinal Chemistry*, **2,** 309-314.
[223] Wang, J.F., Wei, D.Q., Li, L., Zheng, S.Y., Li, Y.X. and Chou, K.C. (2007) 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. *Biochem Biophys Res Commun (Corrigendum: ibid, 2007, Vol357, 330)*, **355,** 513-519.
[224] Wang, J.F., Wei, D.Q., Lin, Y., Wang, Y.H., Du, H.L., Li, Y.X. and Chou, K.C. (2007) Insights from modeling the 3D structure of NAD(P)H-dependent D-xylose reductase of Pichia stipitis and its binding interactions with NAD and NADP. *Biochem Biophys Res Comm*, **359,** 323-329.
[225] Wang, J.F., Wei, D.Q., Chen, C., Li, Y. and Chou, K.C. (2008) Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. *Protein & Peptide Letters*, **15,** 27-32.
[226] Wang, J.F., Wei, D.Q., Du, H.L., Li, Y.X. and Chou, K.C. (2008) Molecular modeling studies on NADP-dependence of Candida tropicalis strain xylose reductase. *The Open Bioinformatics Journal*, **2,** 72-79.
[227] Ding, C.H. and Dubchak, I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17,** 349-358.
[228] Finkelstein, A.V. and Ptitsyn, O.B. (1987) Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol*, **50,** 171-190.
[229] Chou, K.C. and Zhang, C.T. (1995) Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, **30,** 275-349.
[230] Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. and Kim, S.H. (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *PROTEINS: Structure, Function, and Genetics*, **35,** 401-407.
[231] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of protein database for the investigation of sequence and structures. *Journal of Molecular Biology*, **247,** 536-540.
[232] Shen, H.B. and Chou, K.C. (2006) Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **22,** 1717-1722.
[233] Chou, K.C. and Shen, H.B. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *Journal of Proteome Research*, **5,** 1888-1897.
[234] Shen, H.B. and Chou, K.C. (2009) Predicting protein fold pattern with functional domain and sequential evolution information. *Journal of Theoretical Biology*, **256,** 441-446.
[235] Qiu, L.L., Pabit, S.A., Roitberg, A.E. and Hagen, S.J. (2002) Smaller and faster: The 20-residue Trp-cage protein folds in 4 microseconds. *Journal of American Chemical Society*, **124,** 12952-12953.
[236] Goldberg, M.E., Semisotnov, G.V., Friguet, B., Kuwajima, K., Ptitsyn, O.B. and Sugai, S. (1990) An early immunoreactive folding intermediate of the tryptophan synthease beta 2 subunit is a 'molten globule'. *FEBS Lett*, **263,** 51-56.
[237] Plaxco, K.W., Simons, K.T. and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, **277,** 985-994.
[238] Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D. and Finkelstein, A.V. (2003) Contact order revisited: influence of protein size on the folding rate. *Protein Science*, **12,** 2057-2062.
[239] Zhou, H. and Zhou, Y. (2002) Folding rate prediction using total contact distance. *Biophys Journal*, **82,** 458-463.
[240] Gromiha, M.M. and Selvaraj, S. (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol Biol*, **310,** 27-32.
[241] Nolting, B., Schalike, W., Hampel, P., Grundig, F., Gantert, S., Sips, N., Bandlow, W. and Qi, P.X. (2003) Structural determinants of the rate of protein folding. *J Theor Biol*, **223,** 299-307.
[242] Ouyang, Z. and Liang, J. (2008) Predicting protein folding rates from geometric contact and amino acid se-

quence. *Protein Science*, **17,** 1256-1263.
[243] Ivankov, D.N. and Finkelstein, A.V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc Natl Acad Sci USA*, **101,** 8942-8944.
[244] Gromiha, M.M., Thangakani, A.M. and Selvaraj, S. (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res*, **34,** W70-74.
[245] Chou, K.C. (1990) Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophysical Chemistry*, **35,** 1-24.
[246] Chou, K.C. (1989) Graphical rules in steady and non-steady enzyme kinetics. *J Biol Chem*, **264,** 12074-12079.
[247] Lin, S.X. and Neet, K.E. (1990) Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. *J Biol Chem*, **265,** 9670-9675.
[248] Beyer, W.H. (1988) *CRC Handbook of Mathematical Science, (6th Edition), Chapter 10, page 544*. CRC Press, Inc., Boca Raton, Florida.
[249] Shen, H.B., Song, J.N. and Chou, K.C. (2009) Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *Journal of Biomedical Science and Engineering (JBiSE)*, **2,** 136-143 (open accessible at http://www.srpublishing.org/journal/jbise/).
[250] Chou, K.C. and Shen, H.B. (2009) FoldRate: A web-server for predicting protein folding rates from primary sequence. *The Open Bioinformatics Journal*, **3,** 31-50 (open accessible at http://www.bentham.org/open/tobioij/).
[251] Zhang, Z. and Henzel, W.J. (2004) Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci*, **13,** 2819-2824.
[252] Spector, D.L. (2001) Nuclear domains. *J Cell Sci*, **114,** 2891-2893.
[253] Spiess, M. (1995) Heads or tails - what determines the orientation of proteins in the membrane. *FEBS Lett*, **369,** 76-79.
[254] Schulz, G.E. and Schirmer, R.H. (1985) *Principles of Protein Structure, Chapter 2,* Springer-Verlag, New York. 17-18.