Cellular Automata and Its Applications in Protein Bioinformatics

Xuan Xiao[1,2,*], Pu Wang[1], and Kuo-Chen Chou[2]

[1]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China;* [2]*Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, USA*

Abstract: With the explosion of protein sequences generated in the postgenomic era, it is highly desirable to develop high-throughput tools for rapidly and reliably identifying various attributes of uncharacterized proteins based on their sequence information alone. The knowledge thus obtained can help us timely utilize these newly found protein sequences for both basic research and drug discovery. Many bioinformatics tools have been developed by means of machine learning methods. This review is focused on the applications of a new kind of science (cellular automata) in protein bioinformatics. A cellular automaton (CA) is an open, flexible and discrete dynamic model that holds enormous potentials in modeling complex systems, in spite of the simplicity of the model itself. Researchers, scientists and practitioners from different fields have utilized cellular automata for visualizing protein sequences, investigating their evolution processes, and predicting their various attributes. Owing to its impressive power, intuitiveness and relative simplicity, the CA approach has great potential for use as a tool for bioinformatics.

Keywords: Protein bioinformatics, cellular automata, sequences visualization, protein attributes prediction, pseudo amino acid composition, system biology, DNA sequence model, data mining.

1. INTRODUCTION

The advent of genomics and post-genomics technologies is leading to the generation of vast amounts of biological data from catalogs of genes, their products and expression profiles, to their interactions with other molecules, kinetics and cellular locations. For instance, according to GenBank statistics, there were 680,338 bases in 606 sequence records in 1982, while 99,116,431,942 bases in 98,868,465 sequence records were found in 2008 Fig. (**1**). With the same trend, in 1986 the Swiss-Prot databank contained merely 3,939 protein sequence entries Fig. (**2**), but the number has since jumped to 516,081 according to UniProt release of 19-Mar-2010 (www.uniprot.org), meaning that the number of protein sequence entries now is more than 131 times the number from about 24 years ago.

In order to use these newly found protein sequences in a timely way for basic research and drug development, it is highly desirable to determine their structures and functions in a large-scale manner, as well as understand the protein-protein interactions and other complex interactions at a cellular level or in a living system. As a consequence, biologists have increasingly relied on the information and data derived from bioinformatics tools for their ongoing researches [1-27].

The topics of protein bioinformatics are very wide and include many areas, such as prediction and characterization of protein function, localization, secondary structure, tertiary structure, quaternary structure, solvent accessibility, sequential/structural motifs, folding kinetics, binding interaction

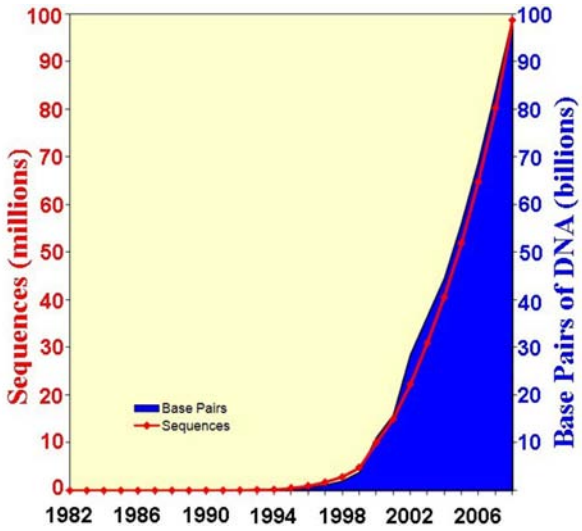

Fig. (1). Growth of GenBank (1982 - 2008).

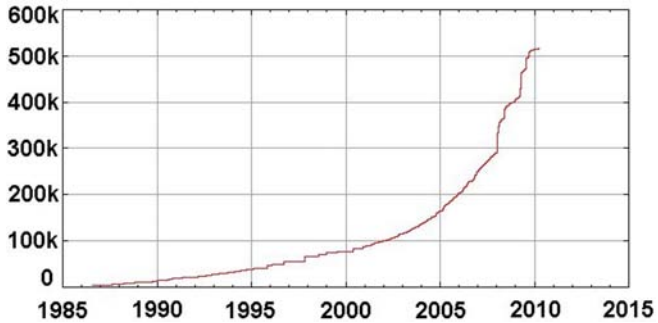

Fig. (2). Number of protein entries in UniProtKB/Swiss-Prot.

*Address correspondence to this author at the Computer Department, Jing-De-Zhen Ceramic Institute, China; Tel:/ Fax: 086-079-8849-9229;
E-mail: xxiao@gordonlifescience.org

1389-2037/11 $58.00+.00

with ligand, DNA, RNA, as well as protein-protein interactions. All of these are far more challenging than the topics in the classical molecular biology because they will involve not only very large structural components but also extremely complicated dynamic processes. Although the research above can be done by conducting a variety of biochemical experiments, the straightforward approach of performing experiments is not only time-consuming but also very costly. Scientists cross-trained in the areas of computer science and molecular biology have become essential partners in the drug discovery process. They have modeled many three-dimensional structures of proteins important for drug development (see, e.g., [28-30]), and helped to understand some subtle action mechanisms (see, e.g., [31,32]). Many studies from various research laboratories around the world have indicated that developing computational prediction tools, conducting computational modeling, and introducing novel physical concept to solve important problems in biology and medicine, such as protein subcellular location prediction [33-35], protein structural class prediction [36-41], modeling 3D structures of targeted proteins for drug design [42,43], diffusion-controlled reaction simulation [44-47], cellular responding kinetics [48,49], bio-macromolecular internal collective motion simulation [50-52], membrane protein type prediction [53], protein cleavage site prediction [54,55], and signal peptide prediction [56], can provide very useful information and insights for both basic research and drug design and hence are widely welcome by science community. Particularly, the recent outbreak of influenza A virus (subtype H1N1) has posed global concerns. Although very important progress relevant to the topic of influenza A virus have been achieved by experiments (see, e.g., [57,58] and a recent comprehensive review [59]), a lot of detailed information in this regard has been acquired by resorting to computational approaches (see, e.g.,[31,32,60-62]).

Most successful protein bioinformatics tools are built utilizing machine learning methods, such as artificial neural network (ANN) [63-65], covariant discriminant (CD) algorithm [5,66], K nearest neighbor (KNN) algorithm [5,67], and Support Vector Machine (SVM) [68-70]. The same is true for some basic tools such as protein sequence alignment method [71-73], Hidden Markov Model (HMM) [74-76], Genetic Algorithm (GA) [77,78]. Meanwhile, a new kind of science with the core of "cellular automata" [79] has been extensively studied for many applications. The concept of cellular automaton (CA) was initiated in the early 1950's by J. von Neumann and Stan Ulam [80]. Since then cellular automata have been reinvented several times under various names such as "cellular spaces", "tessellation automata", "cellular structures", and "iterative arrays". S. Wolfram has performed comprehensive studies of cellular automata since 1980s, and Wolfram's fundamental research in this field has culminated in the publication of his book with the title of "A New Kind of Science" [79] in which Wolfram presents a gigantic collection of results concerning automata, among which are a number of groundbreaking new discoveries.

The theory of cellular automata is immensely rich, with simple rules and structures being capable of producing a great variety of unexpected behaviors. There is no fixed mathematical formula for constructing cellular automata, so the design is open and flexible. Cellular automata have been used to model biological systems from the level of intracellular activity to the levels of clusters of cells, and population of organisms [81,82]. Cellular automata have been used to model the kinetics of molecular systems and crystal growth in chemistry [83]. In physics, the applications cover the study of dynamical systems starting from the interaction of particles to the clustering of galaxies [84]. In the field of computer science, cellular automata based methods have been employed to model the von Neumann (self-reproducing) machines as well as the parallel processing architecture [82]. Beyond the domain of natural science, cellular automata have also been used to study other diverse fields [85].

In this review, we shall give a brief and systematic introduction of various modes of cellular automata and their applications in the field of protein bioinformatics.

2. CELLULAR AUTOMATA

Cellular automata are a discrete model studied in computability theory, mathematics, physics, complexity science, theoretical biology and microstructure modeling. It consists of discrete agents or cells, which occupy some or all sites of a regular lattice. These cells have one or more internal state variables and a set of rules describing the evolution of their state and position. Both the movement and change of state of particles depend on the current state of the cell and those of neighboring cells. These rules may either be discrete or continuous, deterministic or probabilistic.

According to spatial levels, cellular automata may be divided into one dimension, two dimension and high dimension model. The simplest type of cellular automata is a binary, nearest-neighbor, one-dimensional automaton. Such automata were called "elementary cellular automata" by S. Wolfram, who has extensively studied their amazing properties [79]. The elementary cellular automata consist of identical cells, …,*i-3*,*i-2*,*i-1*,*i*,*i+1*,*i+2*,*i+3*,…, and the corresponding states of these cells are $\ldots, S_{i-3}, S_{i-2}, S_{i-1}, S_i, S_{i+1}, S_{i+2}, S_{i+3}, \ldots$; the state of the *i*th cell takes values form a predefined discrete set: $S_i \in \{0, 1\}$. As visualization is considered in a two-state automaton, each of the cells can be either black or white. A cellular automaton's rule F [79] can be expressed as a lookup table that lists, for each local neighborhood, the state that is taken on by the neighborhood's central cell at the next step. A neighborhood comprises a cell and its r neighbors on either side, where r is called the cellular automaton's radius. The neighborhood is defined as follows:

$$N(i,r) = \{S_{i-r}, \ldots, S_{i-1}, S_i, S_{i+1}, \ldots, S_{i+r}\}, \quad r = 0,1,2,\cdots \quad (1)$$

If $r = 1$, the neighborhood of the *i*th cell consists of the same cell and its left and right immediate neighbors; i.e.,

$$N(i,1) = \{S_{i-1}, S_i, S_{i+1}\} \quad (2)$$

The state of the *i*th cell at time step $t+1$ is affected by the states of its neighbors at the previous time step t, namely, the state of the *i*th cell at a time step is function of

the states of its neighbors at the previous time step. The course of state evolving can be represented as:

$$S_i^{t+1} = F(S_{i-r}^t, \ldots, S_i^t, \ldots, S_{i+r}^t) \quad (3)$$

The upper index in the state symbol denotes the time step. S_i^{t+1} is the state of the *i*th cell at time step $t+1$. If $r=1$, Eq. (3) becomes

$$S_i^{t+1} = F(S_{i-1}^t, S_i^t, S_{i+1}^t) \quad (4)$$

Shown in Fig. (**3**) is the evolution of a one-dimensional cellular automaton. The horizontal axis is space and the vertical axis is time. Each row represents the cellular automaton at each time step and each column represents the state of the same cell at various time steps. In elementary cellular automata, each cell can be either black or white, then this will allows $2^3 = 8$ possible color combinations along the $(S_{i-1}^t, S_i^t, S_{i+1}^t)$. Because each of these combinations will cause S_i^{t+1} to be either black or white and there are eight possible upper color combinations then there will be $2^8 = 256$ possibilities in total. We can easily utilize a binary byte to encode these rule sets into decimal numbers between the numbers 0 and 255. For example, rule number 184 would correspond to Fig. (**4**).

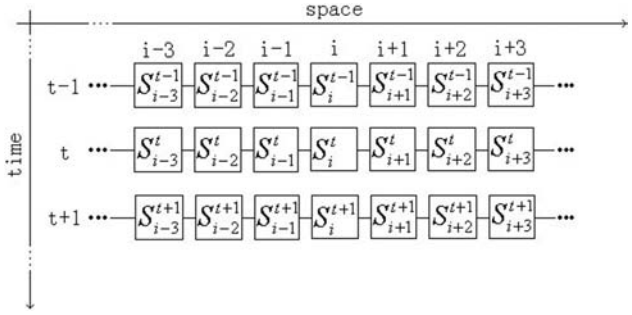

Fig. (3). Illustration to show the evolution of a one-dimensional cellular automaton.

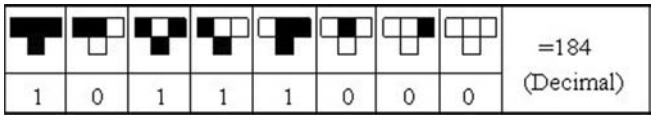

Fig. (4). Rule number 184 [79]. The string of eight "0"s and "1"s create one binary byte, which can represent a decimal number between 0 and 255.

The best-known two dimension cellular automaton is Conway's game of life devised by the British mathematician John Horton Conway in 1970 [86], its rules are as follows: If a black cell has 2 or 3 black neighbors, it stays black. If a black cell has less than 2 or more than 3 black neighbors it becomes white. If a white cell has 3 black neighbors, it becomes black Fig. (**5**). The initial pattern constitutes the seed of the system. The first generation is created by applying the above rules simultaneously to every cell in the seed—births and deaths happen simultaneously, and the discrete moment at which this happens is sometimes called a tick (in other words, each generation is a pure function of the one before). The rules continue to be applied repeatedly to create further generations. Life provides an example of emergence and self-organization. It is interesting for physicists, biologists, economists, mathematicians, philosophers, generative scientists and others to observe the way that complex patterns can emerge from the implementation of very simple rules.

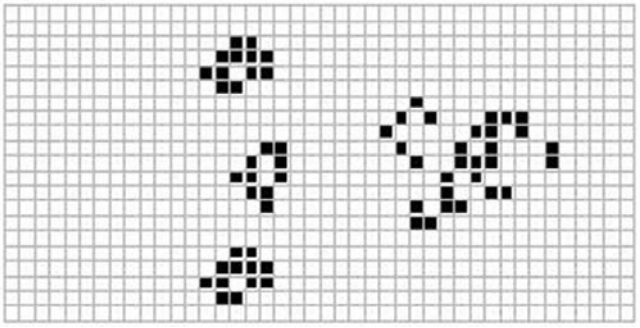

Fig. (5). Two-dimensional cellular automaton Conway's game of life.

3. APPLICATION IN PROTEIN BIOINFORMATICS

3.1. Protein Sequences Visualization

In general, gene or protein sequences are stored in the computer database system in the form of long character strings. It would act like a snail's pace for human beings to read these sequences with the naked eyes. Also, it is very hard to extract any key features by directly reading these sequences. However, if they can be converted to some diagrams [87,88], some important features can automatically manifested and become easily visible.

The question of how to visualize biology sequence is an important topic today [89-94], and much effort has been put into it [89-100]. The benefit to do so is quite like the case that introduction of graph or diagram approaches to study biological systems can provide an intuitive picture or useful insights, helping us to catch the essence or core of the problem as demonstrated in a series of previous studies (see, e.g., [101-105]). However, there is a common characteristic in the general visual methods, i.e., the point of the special curve corresponding to a certain nucleic acid is colligated only with the base prior to it, while the effects of all the bases behind it are totally ignored. This is inconsistent with the fact that all the bases in a gene are coupled with each other as an entity in nature. Protein sequence visualization has the same problem. In view of this, a completely new and different method is introduced to image the molecular sequences. The novel method is based on cellular automata.

As is well known, a DNA molecule is formed by nucleic acids, also called bases. The 4 nucleic acids are adenine (A), cytosine (C), guanine (G), and uracil (U). To deal with it in a computer, the four bases in a nucleotide sequence is coded as follows [106]

$$A=00,\ C=01,\ G=10,\ U=11 \quad (5)$$

Proteins are represented by sequences of amino acids, also called residues. There are 20 native amino acids. By means of the similarity rule, complementarity rule, molecular recognition theory and information theory, a set of digital codes are formulated to represent amino acids, as shown in

Table **1**. The representation can better reflect the chemical physical properties of amino acids, as well as their structure and degeneracy. Through the above encoding procedure, a protein sequence is transformed to a series of digital signals. For example, the sequence "MASAA…" is accordingly transformed to "10011110010100111001110011001…".

Suppose a protein $\mathbf{P}$ consists of N amino acids; i.e.,

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_N \tag{6}$$

where R_1 represents the 1st residue of the protein, R_2 the 2nd residue, and so forth. According to Table **1**, the residue chain of **Eq.6** is initially converted to a sequence with $5N$ digits; i.e.,

$$\mathrm{P}(t) = \mathrm{g}_1(t)\mathrm{g}_2(t)\cdots\mathrm{g}_N(t)\cdots\mathrm{g}_{5N}(t), \quad (t=0) \tag{7}$$

where $\mathrm{g}_i(t) = 0$ or 1 $(i = 1, 2, \cdots, 5N)$ as defined by Table **1**. Suppose the time for each updated step is consecutively expressed by $t = 0, 1, 2, \cdots, \Omega$, it follows [107]

$$\mathrm{P} = \begin{bmatrix} \mathrm{P}(1) \\ \mathrm{P}(2) \\ \vdots \\ \mathrm{P}(\Omega) \end{bmatrix} = \begin{bmatrix} \mathrm{g}_1(0)\,\mathrm{g}_2(0)\,\cdots\mathrm{g}_N(0)\,\cdots\,\mathrm{g}_{5N}(0) \\ \mathrm{g}_1(1)\,\mathrm{g}_2(1)\,\cdots\,\mathrm{g}_N(1)\,\cdots\,\mathrm{g}_{5N}(1) \\ \vdots \\ \mathrm{g}_1(\Omega)\mathrm{g}_2(\Omega)\cdots\mathrm{g}_N(\Omega)\cdots\mathrm{g}_{5N}(\Omega) \end{bmatrix} \tag{8}$$

Thus, for rule 84 in [79], we have (cf. Eq.4 of [24]):

$$\mathrm{g}_i(t+1) = \begin{cases} 0, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=1 \\ 1, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=0,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=1 \\ 1, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=0,\ \mathrm{g}_{i+1}(t)=1 \\ 1, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=0 \\ 0, & \text{if } \mathrm{g}_{i-1}(t)=1,\ \mathrm{g}_i(t)=1,\ \mathrm{g}_{i+1}(t)=1 \end{cases} \quad (t = 0, 1, \cdots, \Omega) \tag{9}$$

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

$$\mathrm{g}_0(t) = \mathrm{g}_{5N}(t) \text{ and } \mathrm{g}_{5N+1}(t) = \mathrm{g}_1(t) \tag{10}$$

In case the ith grid at t is filled with white color if $\mathrm{g}_i(t) = 0$ and black if $\mathrm{g}_i(t) = 1$. Accordingly, each $\mathrm{P}(t)$ $(t = 0, 1, 2, \cdots, \Omega)$ in Eq.8 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 (dimensional) black-and-white image. The image thus evolved is called the cellular automaton image (CAI). Its advantage is that it can help us visualize some special features hidden in a long and complex sequence [108].

If the rule and time for the evolution are all changeless, the molecular sequence and image thus produced will be one-to-one correspondence. Because digital coding for amino acid and nucleotide are of degeneracy, the images will appear in different cells for the first row at least. Fig. (**6**) shows the comparative image between SARS-CoVs and other coronavirus. For the figure, we can clearly see the difference that the images of SARS-CoVs are mainly with the V-shaped cross-lines pattern, whereas that of non-SARS virus sequence is mainly with the parallel slash-lines pattern [108-111].

3.2. Protein Attributes Prediction

Many typical topics in sequential bioinformatics are relevant to prediction of protein attributes [112,113]. For example, given an uncharacterized protein sequence, does it locate in nucleus or other subcellular location? What kind of structure characteristics does it have? Is it a G-protein-coupled receptor (GPCR) or a non-GPCR? If it is the former, to which GPCR functional type does it belong? The list of questions is vast.

Although biochemical experiments may put forward the solutions to these problems, however it is not only time-consuming but also very costly. Then it would be highly desirable to develop effective bioinformatics tools for predicting various kinds of attributes for uncharacterized proteins based on their sequence information alone [113].

Table 1. Binary Notation of Amino Acid Coding Language [108]

Codon	Amino acid	Binary notation	Codon	Amino acid	Binary notation
ccu ccc cca ccg	P	00001	cuu cuc cua cug uua uug	L	00011
caa cag	Q	00100	cau cac	H	00101
cgu cgc cga cgg aga agg	R	00110	ucu ucc uca ucg agu agg	S	01001
uuu uuc	F	01011	uau uac	Y	01100
ugg	W	01110	ugu ugc	C	01111
acu acc aca acg	T	10000	auu auc aua	I	10010
aug	M	10011	gcu gcc gca gcg	D	11100
gaa gag	E	11101	ggu ggc gga ggg	G	11110
uaa uag uga	End	11111			

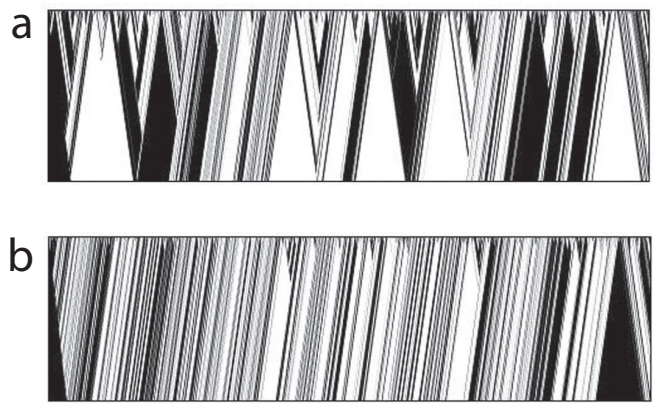

Fig. (6). Sample images obtained by applying the Rule 184 on (a) the SARS coronal virus BJ01 (AY278488), and (b) non-SARS coronavirus AF208066_Murine. The abscissa stands for the length of digital encoding sequence, while the ordinate for the evolving times. Here the time of evolving was 2400. The compression ratio is 14:2. The SARS image is with a V-shaped cross-lines pattern, a token for SARS coronal viruses; and the non-SARS coronavirus image is with a parallel slash-lines pattern, a remarkable distinction with the SARS coronal virus [109].

To develop an effective method for predicting protein attributes, one of the most important things is to find an effective mathematical expression for the protein samples that can truly reflect their intrinsic correlation with the object to be predicted. There are many different ways to formulate protein sequence samples. However, they can be basically categorized into two different kinds of representations: the sequential representation and the discrete model [114].

The simplest discrete model used to represent a protein sample is its amino acid (AA) composition or AAC. For a protein sequence formulated in **Eq. 6**, its AAC can be expressed as a vector given by [115]

$$\mathrm{P}=\left[\begin{array}{cccc} f_1 & f_2 & \cdots & f_{20} \end{array}\right]^{\mathrm{T}} \tag{11}$$

where $f_i \ (i=1,2,\cdots,20)$ are the normalized occurrence frequencies of the 20 native amino acids [116,117] in **P**, and **T** the transposing operator. However, as one can see from **Eq. 11**, all the sequence-order effects would be lost by using the AAC-discrete model. To avoid completely losing the sequence-order information, the concept of pseudo amino acid composition (PseAAC) was proposed [118]. According to the concept of PseAAC, a protein **P** of **Eq.6** can be represented by

$$\mathrm{P}=\left[\begin{array}{cccccccc} p_1, & p_2, & \cdots, & p_{20}, & p_{20+1}, & \cdots, & p_{20+\lambda} \end{array}\right]^{\mathrm{T}} \tag{12}$$

where

$$p_k=\begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i+\sum_{j=1}^{\lambda} w_j \Phi_j}, & (1\le k\le 20) \\ \dfrac{w_{(k-20)}\Phi_{(k-20)}}{\sum_{i=1}^{20} f_i+\sum_{j=1}^{\lambda} w_j \Phi_j}, & (21\le k\le 20+\lambda) \end{cases} \tag{13}$$

where $f_i \ (i=1,2,\cdots,20)$ have the meaning as those in **Eq. 7**, λ is a integer but must be smaller than N [118], Φ_j the j-th tier correlation factor that reflects the sequence order correlation between all the j-th most contiguous residues, and $w_j \ (j=1,2,\cdots,\lambda)$ are the weight factors. For a detailed formulation and elaboration about PseAAC, refer to [118,119]. For a summary about its development and applications, see a recent comprehensive review [114]. Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems (see, e.g., [120-147]). Meanwhile, using the concept of PseAAC, the cellular automata approaches were used to formulate various different modes of PseAAC for predicting the following attributes.

3.2.1. Protein Subcellular Localization

The cellular automaton images were introduced to reflect protein sequence via PseAAC approach for improving the prediction quality of protein subcellular localization, as proposed by Xiao *et al.* [148].

3.2.2. Protein Structural Classes

The geometric moments of cellular automaton images Fig. (**7**) were used via the PseAAC formulation [118] for predicting protein structural classes [149].

3.2.3. GPCR Families

The gray-level co-occurrence matrix factors [150] extracted from the cellular automaton images Fig. (**8**) were used to represent the samples of proteins through the PseAAC formulation for predicting GPCR (G-protein coupled receptor) families, as elaborated by Xiao *et al.* [107]. Meanwhile, a publicly accessible web-server called GPCR-CA has been provided at http://icpr.jci.jx.cn/bioinfo/GPCR-CA, by which one can easily identify whether it belongs to GPCR for a given protein sequence, and, if it does, which GPCR family it belongs to.

3.2.4. Predicting Transmembrane Regions in Protein

By fusing the information of cellular automata and Lempel-Ziv complexity into the PseAAC formulation [118], Diao *et al.* [151] proposed a method for predicting the TM (transmembrane) regions of integral membrane proteins including both α- helical and β- barrel membrane proteins

32.5. Protein Quaternary Structural Attributes

Recently, as proposed by Xiao *et al.* [152], the cellular automaton image and complexity measurement factor [153] were used to formulate protein samples via the PseAAC approach for predicting protein quaternary structural attributes. Moreover, a user-friendly web-server called Quat-2L has been provided at via http://icpr.jci.jx.cn/bioinfo/Quat-2L. Quat-2L is a 2-layer predictor: The 1st layer is for identifying the query protein as monomer, homo oligomer, or hetero-oligomer. If the result thus obtained turns out to be homo-oligomer or hetero-oligomer, then the prediction will be automatically continued to further identify it as belonging to

which one of the following six subtypes: (1) dimer, (2) trimer, (3) tetramer, (4) pentamer, (5) hexamer, and (6) octamer.

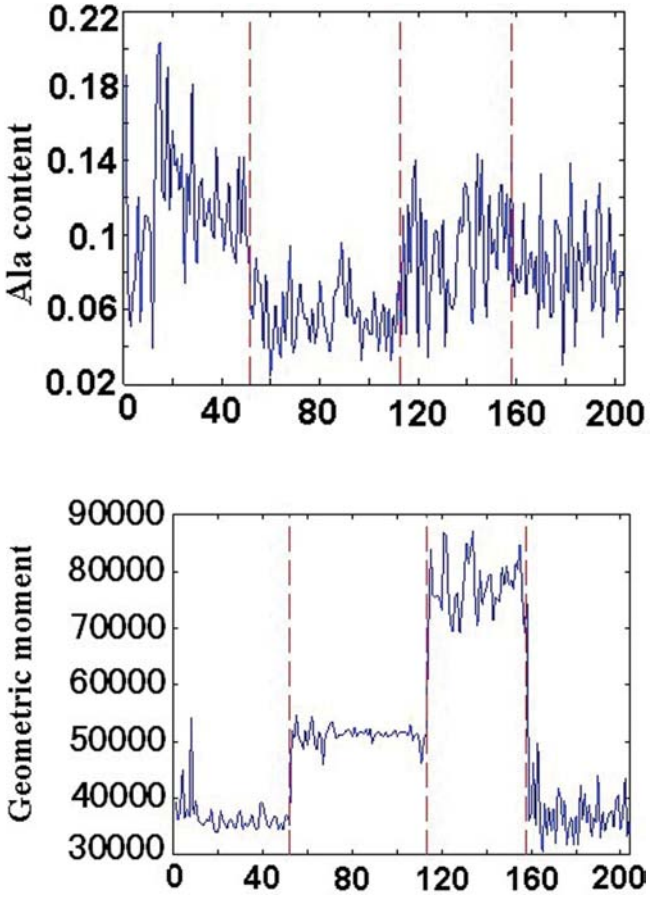

Fig. (7). The content of Ala for the 204 proteins taken from Table 2 of [175]: **(a)** the 1st column as marked by the red dashed line is for the 52 proteins of all-α structural class; the 2nd column is for the 61 proteins of all-β structural class; the 3rd column for the 45 proteins of the α/β structural class; and the 4th column for the 46 proteins of the α+β structural class. **(b)** The corresponding geometric moment values for the 204 proteins. The red dashed lines are used to define the aforementioned four columns. The chart shows that the geomantic moment from the cellular automaton images can well separate proteins according to their secondary structure topology.

3.3. Modeling Systems Biology

3.3.1. Virus Infection Dynamics

In-depth understanding virus infection mechanism is quite important to human health and medical science (see, e.g., [31,32,57,58]). Many mathematical models have been developed trying to understand the dynamics of the virus infection. Most of them have used the ordinary (or partial) differential equations to describe different aspects of the dynamics of the host-parasite interaction. Although these models have made contributions in helping understand the development of the disease, they failed to describe the following two time scales observed during the course of infection: the short time scale associated with the primary response and the long time scale associated with the clinical latency period. These mean-field-like models as such were actually based on a simplification of the biology of virus infection; they did not take into account the cell proliferation, the local interactions, and the spatial inhomogeneities. These features are of central importance. In view of this, researchers have resorted to the cellular automaton approach. It could take a spatial element into consideration that was not evident in most of the previous models of differential equations, and hence provide us with a feasible way to model complex dynamical phenomena by reformulating the macroscopic behavior in terms of the microscopic and mesoscopic rules that are discrete in space and time.

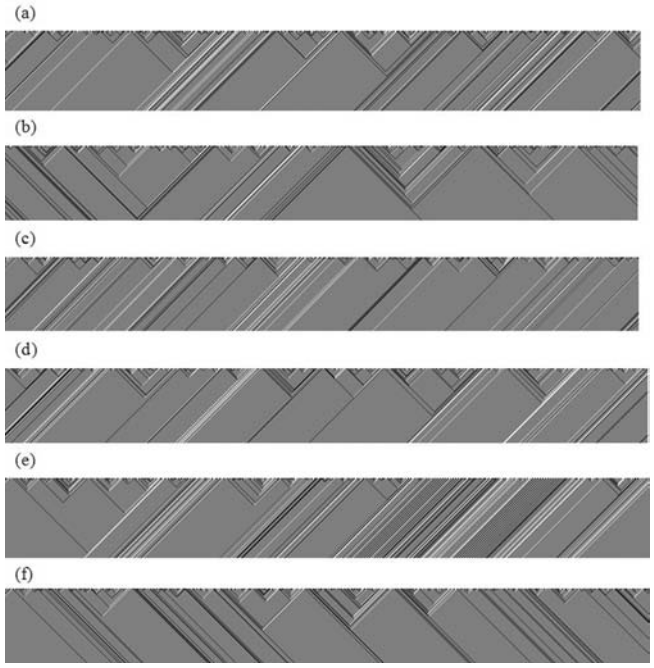

Fig. (8). The cellular automaton image for a protein taken from **(a)** A-rhodopsin like family, **(b)** B-secretin like family, **(c)** C-metabotrophic/glutamate/pheromone family; **(d)** D-fungal pheromone family, **(e)** E-cAMP receptor family, and **(f)** F-Frizzled/Smoothemed family, respectively. As we can see, GPCR members from different families have quite different textures in their cellular automation images [107].

Using a cellular automata model, Zorzenon dos Santos and Coutinho [154] investigated the evolution of human immunodeficiency virus (HIV) infection and the onset of acquired immune deficiency syndrome (AIDS). Their model took into account the global features of the immune response to any pathogen, the fast mutation rate of the HIV, and a fair amount of spatial localization, which might occur in the lymph nodes. The results thus obtained by them could reproduce the three-phase pattern as observed in T cell, i.e., the primary response, the clinical latency period, and the onset of AIDS [154].

To study the dynamic process of Hepatitis B virus (HBV) infection, Xiao *et al.* [155] introduced a simple 2D (dimensional) probability cellular automaton model. The model took into account the existence of different types of HBV infectious and non-infectious particles. The simulation results thus obtained showed that the cellular automaton model

could successfully elucidate some important features of the disease, such as its wide variety in manifestation and its age dependency. Meanwhile, the effects of the model's parameters on the dynamical process of the infection were also investigated, as reflected by Fig. (**9**).

In order to investigate the influence of spatial heterogeneities on viral spread, a simple 2-D cellular automaton model of a viral infection was developed by Beauchemin *et al.* [156]. In this initial phase of the investigation, their cellular automaton model was validated against clinical immunological data for uncomplicated influenza A infections.

With a stochastic cellular automaton model, Castiglione *et al.* [157] conducted a study on the infection process of the human herpesvirus EBV. Their study has successfully addressed the issues such as spatial and temporal heterogeneity, as well as the simulation produces kinetics of virus-infected cells and virion production in the acute phase that resemble that seen in patients.

3.3.2. Simulating Protein Biochemistry Processes

Kier *et al.* [158] have developed a cellular automata model for a biochemical system of enzyme reaction with a substrate in water. The model can fairly produce Michaelis-Menten kinetics with good Lineweaver-Burk plots. With the variation in affinity parameters, they further found that hydrophobic substrates are generally more reactive with enzymes than their other features.

As a simulation tool, Siehs *et al.* [159] presented the lattice molecular automaton model, which could be used to represent complex biomolecular dynamics at different levels of granularity.

Sanford *et al.* [160] have utilized the cellular automata engine to simulate biochemical pathways. They studied the potential impact of spatial organization in the evolution and engineering of signaling and metabolic pathways, demonstrating the features consistent with the phenomenon of metabolic channeling.

Kier and Chen [161] have created a cellular automata model of a semi-permeable membrane [161]. The system based on their model could respond to the presence of a solute on one side of the membrane by reducing the passage of water from this side into the membrane. It could also be used to study the effect of solute lipophilicity on the concentration of solute passing through the membrane.

Most of the existing mathematical models for tumour growth and tumour-induced angiogenesis neglect blood flow. Actually, the latter is an important factor on which both nutrient and metabolite supply depend. To address such a shortcoming, Alarcon *et al.* [162] have developed a cellular automaton model, by which it can be shown how blood flow and red blood cell heterogeneity affect the growth of systems of normal and cancerous cells.

3.3.3. Modeling Immune System

The development of the immune repertoire during neonatal life involves a strong selection process among different clones. In view of this, de Boer and Perelson [163] have investigated the hypothesis that repertoire selection is carried out during the early life by the immune network. There are at least two processes in repertoire selection: clonal expansion and recruitment of clones by the bone marrow. Because both occur on the time scales of a few days, they proposed a 2D cellular automaton model to investigate the two processes. The outcome derived from their model confirmed the hypothesis, implying that the cellular automaton model is indeed quite helpful for us to understand the pattern formation in the immune network systems due to its straightforward visualization.

3.4. Protein Sequence Evolution Model

Early in 1980s, a number of cellular automaton applications have been reported in the fields of DNA sequences [164-166]. Researchers have sought to describe a starting point for modeling the evolution and role of DNA sequences within the framework of cellular automata.

Sirakoulis *et al.* [167] designed a cellular automata model for DNA structure, function and evolution. In their model, DNA was treated as a one-dimensional cellular automaton with four states per cell. These states are the four bases in DNA: A, C, T and G. The four states are represented by numbers of the quaternary number system, and the linear evolution rules are considered as represented by square matrices. Based on this model, a simulator of DNA evolution

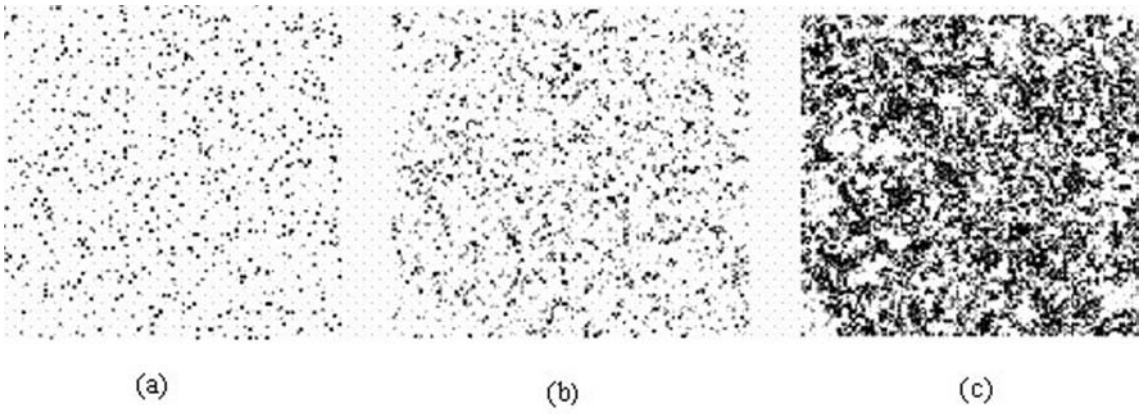

Fig. (9). Snapshots to show the effects of the parameter of time period on the lattice configuration: (a) one week, (b) two weeks, and (c) 30 weeks. The codes for the different states of the cell are the following: healthy ⇔ white, dead ⇔ grey, and infected cells ⇔ black [155].

was developed and simulation results were presented. Lately, they also studied the reconstruction of DNA sequences using genetic algorithms (GA) and cellular automata [168]. In their study, DNA was also modeled as a one-dimensional cellular automaton, where the sugar-phosphate backbone corresponds to the cellular automaton lattice and the organic bases to the cellular automaton cell. The genetic algorithms thus developed could efficiently extract the cellular automaton rule for governing the sequence evolution.

3.5. Biomedicine Data Mining

Moore and Hahn [169] have introduced cellular automata as a novel computational approach for identifying combinations of single-nucleotide polymorphisms (SNPs) associated with clinical endpoints. With the simulated data, they demonstrated that their approach has good power for identifying high-order nonlinear interactions among four SNPs in the absence of independent main effects [170].

For common multifactorial diseases such as hypertension, interactions between genetic variations are likely to be more important than the independent effects of any single genetic variation. Attribute interaction is a well-known problem in data mining and is a complicating factor in genetic data analysis. Moore and Hahn [169,170] have addressed this problem by developing a parallel approach where the one-dimensional cellular automata were utilized for knowledge representation while the genetic algorithms for optimization. They also evaluated the power of such a parallel approach by simulating gene-gene interactions, as well as by adding noise from several common real-world sources. These kinds of simulation studies have further documented the strengths of the cellular automaton approach [170].

Identifying the promoter regions play a vital role in understanding the human genes. In view of this, Kiran and Ramesh [171] proposed a new cellular automata based text clustering algorithm for identifying these promoter regions in genomic DNA.

4. CONCLUSION AND PERSPECTIVES

The huge amount of protein sequences generated in the postgenomic era has stimulated the development of protein bioinformatics. In order to understand and timely use these data for basic research and drug development, it has been highly desired to develop various powerful computational methods.

A cellular automaton is a discrete dynamical system providing an excellent platform for performing complex computation with the help of only local information. Researchers, scientists and practitioners from different fields have exploited the cellular automaton paradigm of local information, decentralized control and universal computation for modeling different applications and investigating various complicated problems, including many fundamental issues in biology.

Protein sequences stored in databases are often strings of characters, and how to read or compare them is one of the basic problems we are often encountered with. Visualization may be a good choice. By encoding a protein sequence into digital format with genetic and physical chemistry information, followed by using cellular automaton to evolve a 2-dimensional image by taking into account the interaction between bases or amino acids, many important features, which are originally hidden in the biomolecule sequence, can be clearly revealed thru its cellular automaton image. According to Wolfram's theory, each protein sequence is corresponding to a cellular automaton image with its own textural feature. Accordingly, those proteins that belong to a same attribute must have some similar textures in their cellular automaton images. Thus, the features extracted from their cellular automaton images can be used to cluster or distinguish various attributes of proteins.

Systems biology involves the usage of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes [172] which comprise metabolism [173], signal transduction pathways and gene regulatory networks) for both analyzing and visualizing the complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms. Cellular automata were introduced as a possible idealization of biological systems, in order to model biological self-reproduction. Despite the simple construction of cellular automata, they are capable of simulating highly complex behaviors. The advantage of such a technique is twofold. First, the components or systems thus generated for simulations usually bear more biological characters than those by the traditional mathematical approaches. Second, it is easy to modify the complexity of the interactions without introducing any new qualitative difficulties in solving the model. Nowadays, cellular automata have been used to model not only biological systems from the level of intracellular activity to the levels of clusters of cells, but also the population of organisms, i.e., virus infection, immune system, molecular kinetics, biochemical pathways, tumor development and other biochemistry processes. All of these would help the understanding of biological mechanisms at the level of systems biology and stimulate new strategies for developing novel drugs [174].

In elementary cellular automata, the evolution rule can be extracted from a given number of cellular automaton patterns. This method can also be applied to the cellular automata for modeling protein sequences. The activities of cellular automata can mediate biological regulation and information processing via nonlinear electrodynamic effects in cytoskeletal lattice arrays. Some work has already been devoted to provide a framework for modeling protein sequences as automata.

Data reduction and pattern recognition approaches are good at identifying complex relationships in data. Since the cellular automaton approach is nonparametric, model-free and simple in calculation, it is quite useful in this regard, particularly in mining and classifying biomedicine data, such as in identifying and characterizing susceptibility genes for common complex multifactorial human diseases. Moreover, cellular automata can also be used for identifying the promoter regions and finding protein-coding regions in genomic DNA.

It is expected that, as a new kind of science, cellular automata hold enormous potential for investigating compli-

cated systems, including many challenging problems in bioinformatics and system biology.

ACKNOWLEDGEMENTS

The authors wish to thank the three anonymous reviewers, whose constructive comments are very helpful for improving the presentation of this paper. This work was supported by the grants from the National Natural Science Foundation of China (No. 60961003), the Key Project of Chinese Ministry of Education (No. 210116), the Province National Natural Science Foundation of JiangXi (No. 2010GZS0122 and 2009GZS0064), the Department of Education of JiangXi Province (No.GJJ11557), and the plan for training youth scientists (stars of Jing-Gang) of Jiangxi Province.

REFERENCES

[1] Gonzalez-Diaz, H.; Gonzalez-Diaz, Y.; Santana, L.; Ubeira, F.M.; Uriarte, E. Proteomics, networks, and connectivity indices. *Proteomics*, **2008**, *8*, 750-778.

[2] Garcia, I.; Diop, Y.F.; Gomez, G. QSAR & complex network study of the HMGR inhibitors structural diversity. *Curr. Drug Metab.*, **2010**, *11*, 307-314.

[3] Chou, K.C.; Shen, H.B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Comm.*, **2007**, *357*, 633-640.

[4] Gonzalez-Diaz, H. Network topological indices, drug metabolism, and distribution. *Curr. Drug Metab.*, **2010**, *11*, 283-284.

[5] Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[6] Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F.M.; Prado-Prado, F.; Perez-Montoto, L.G.; Concu, R.; Podda, G.; Shen, B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curr. Drug Metab.*, **2010**, *11*, 379-406.

[7] Khan, M.T. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr. Drug Metab.*, **2010**, *11*, 285-295.

[8] Chou, K.C.; Shen, H.B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Comm.*, **2007**, *360*, 339-345.

[9] Martinez-Romero, M.; Vazquez-Naya, J.M.; Rabunal, J.R.; Pita-Fernandez, S.; Macenlle, R.; Castro-Alvarino, J.; Lopez-Roses, L.; Ulla, J.L.; Martinez-Calvo, A.V.; Vazquez, S.; Pereira, J.; Porto-Pazos, A.B.; Dorado, J.; Pazos, A.; Munteanu, C.R. Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. *Curr. Drug Metab.*, **2010**, *11*, 347-368.

[10] Chou, K.C.; Shen, H.B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protocols*, **2008**, *3*, 153-162.

[11] Mrabet, Y.; Semmar, N. Mathematical methods to analysis of topology, functional variability and evolution of metabolic systems based on different decomposition concepts. *Curr. Drug Metab.*, **2010**, *11*, 315-341.

[12] Wang, J.F.; Chou, K.C. Molecular modeling of cytochrome P450 and drug metabolism. *Curr. Drug Metab.*, **2010**, *11*, 342-346.

[13] Caballero, J.; Fernandez, M. Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). *Curr. Top. Med. Chem.*, **2008**, *8*, 1580-1605.

[14] Duardo-Sanchez, A.; Patlewicz, G.; Lopez-Diaz, A. Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues. *Curr. Top. Med. Chem.*, **2008**, *8*, 1666-1675.

[15] Chou, K.C.; Shen, H.B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Comm.*, **2008**, *376*, 321-325.

[16] Gonzalez, M.P.; Teran, C.; Saiz-Urra, L.; Teijeira, M. Variable selection methods in QSAR: an overview. *Curr. Top. Med. Chem.*, **2008**, *8*, 1606-1627.

[17] Gonzalez-Diaz, H. Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR). *Curr. Top. Med. Chem.*, **2008**, *8*, 1554.

[18] Gonzalez-Diaz, H.; Prado-Prado, F.; Ubeira, F. .M. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr. Top. Med. Chem.*, **2008**, *8*, 1676-1690.

[19] Helguera, A.M.; Combes, R.D.; Gonzalez, M.P.; Cordeiro, M.N. Applications of 2D descriptors in drug design: a DRAGON tale. *Curr. Top. Med. Chem.*, **2008**, *8*, 1628-1655.

[20] Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr. Top. Med. Chem.*, **2008**, *8*, 1691-1709.

[21] Vilar, S.; Cozza, G.; Moro, S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr. Top. Med. Chem.*, **2008**, *8*, 1555-1572.

[22] Wang, J.F.; Wei, D.Q.; Chou, K.C. Drug candidates from traditional chinese medicines. *Curr. Top. Med. Chem.*, **2008**, *8*, 1656-1665.

[23] Wang, J.F.; Wei, D.Q.; Chou, K.C. Pharmacogenomics and personalized use of drugs. *Curr. Top. Med. Chem.*, **2008**, *8*, 1573-1579.

[24] Chou, K.C.; Shen, H.B. Review: recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.*, **2009**, *2*, 63-92 (openly accessible at http://www.scirp.org/journal/NS/).

[25] González-Díaz, H.; Prado-Prado, F.; Pérez-Montoto, L.G.; Duardo-Sánchez, A.; López-Díaz, A. QSAR Models for Proteins of Parasitic Organisms, Plants and Human Guests: Theory, Applications, Legal Protection, Taxes, and Regulatory Issues. *Curr. Proteomics*, **2009**, *6*, 214-227.

[26] Reimel, B.A.; Pan, S.; May, D.H.; Shaffer, S.A.; Goodlett, D.R.; McIntosh, M.W.; Yerian, L. M.; Bronner, M.P.; Chen, R.; Brentnall, T.A. Proteomics on Fixed Tissue Specimens - A Review. *Curr. Proteomics*, **2009**, *6*, 63-69.

[27] Torrens, F.; Castellano, G. Topological Charge-Transfer Indices: From Small Molecules to Proteins. *Curr. Proteomics*, **2009**, *6*, 204-213.

[28] Chou, K.C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **2004**, *11*, 2105-2134.

[29] Chou, K.C. Molecular therapeutic target for type-2 diabetes. *J. Proteome Res.*, **2004**, *3*, 1284-1288.

[30] Chou, K.C. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochem. Biophys. Res. Comm.*, **2004**, *316*, 636-642.

[31] Huang, R.B.; Du, Q.S.; Wang, C.H.; Chou, K.C. An in-depth analysis of the biological functional studies based on the NMR M2 channel structure of influenza A virus. *Biochem. Biophys. Res. Comm.*, **2008**, *377*, 1243-1247.

[32] Wei, H.; Wang, C.H.; Du, Q.S.; Meng, J.; Chou, K.C. Investigation into adamantane-based M2 inhibitors with FB-QSAR. *Med. Chem.*, **2009**, *5*, 305-317.

[33] Zhou, G.P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *PROTEINS: Struct., Funct., Genet.*, **2003**, *50*, 44-48.

[34] Chou, K.C.; Shen, H.B. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS ONE*, **2010**, *5*, e11335.

[35] Chou, K.C.; Shen, H.B. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE*, **2010**, *5*, e9931.

[36] Zhou, G.P. An intriguing controversy over protein structural class prediction. *J. Prot. Chem.*, **1998**, *17*, 729-738.

[37] Zhou, G.P.; Assa-Munt, N. Some insights into protein structural class prediction. *PROTEINS: Struct., Funct., Genet.*, **2001**, *44*, 57-59.

[38] Chou, K.C.; Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*, 275-349.

[39] Kedarisetti, K.D.; Kurgan, L.A.; Dick, S. Classifier ensembles for protein structural class prediction with varying homology. *Biochem. Biophys. Res. Commun.*, **2006**, *348*, 981-988.

[40] Chen, K.; Kurgan, L.A.; Ruan, J. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J. Comput. Chem.*, **2008**, *29*, 1596-1604.

[41] Kurgan, L.; Homaeian, L. Prediction of Structural Classes for Protein Sequences and Domains - Impact of Prediction Algorithms, Sequence Representation and Homology, and Test Procedures on Accuracy. *Pattern Recognition Lett.*, **2006**, *39*, 2323-2343.

[42] Zhou, G.P.; Troy, F.A. NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. *Curr. Prot. Pept. Sci.*, **2005**, *6*, 399-411.

[43] Chou, K.C.; Wei, D.Q.; Zhong, W.Z. 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.*, **2003**, *308*, 148-151.

[44] Zhou, G.Z.; Wong, M.T.; Zhou, G.Q. Diffusion-controlled reactions of enzymes. An approximate analytic solution of Chou's model. *Biophys. Chem.*, **1983**, *18*, 125-132.

[45] Zhou, G.Q.; Zhong, W.Z. Diffusion-controlled reactions of enzymes. A comparison between Chou's model and Alberty-Hammes-Eigen's model. *Eur. J. Biochem.*, **1982**, *128*, 383-387.

[46] Chou, K.C.; Zhou, G.P. Role of the protein outside active site on the diffusion-controlled reaction of enzyme. *J. Am. Chem. Society*, **1982**, *104*, 1409-1413.

[47] Zhou, G.P.; Li, T.T.; Chou, K.C. The flexibility during the juxtaposition of reacting groups and the upper limits of enzyme reactions. *Biophys. Chem.*, **1981**, *14*, 277-281.

[48] Qi, J.P.; Shao, S.H.; Li, D.D.; Zhou, G.P. A dynamic model for the p53 stress response networks under ion radiation. *Amino Acids*, **2007**, *33*, 75-83.

[49] Qi, J.P.; Ding, Y. S.; Shao, S.H.; Zeng, X.H.; Chou, K.C. Cellular responding kinetics based on a model of gene regulatory networks under radiotherapy. *Health*, **2010**, *2*, 137-146.

[50] Zhou, G.P. Biological functions of soliton and extra electron motion in DNA structure. *Physica Scripta*, **1989**, *40*, 698-701.

[51] Chou, K.C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophysic. Chem.*, **1988**, *30*, 3-48.

[52] Chou, K.C. The biological functions of low-frequency phonons: 6. A possible dynamic mechanism of allosteric transition in antibody molecules. *Biopolymers*, **1987**, *26*, 285-295.

[53] Cai, Y.D.; Zhou, G.P.; Chou, K.C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical J.*, **2003**, *84*, 3257-3263.

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

[55] Chou, K.C. Review: Prediction of HIV protease cleavage sites in proteins. *Analytical Biochem.*, **1996**, *233*, 1-14.

[56] Chou, K.C. Review: Prediction of protein signal sequences. *Curr. Prot. Pept. Sci.*, **2002**, *3*, 615-622.

[57] Schnell, J.R.; Chou, J.J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature*, **2008**, *451*, 591-595.

[58] Wang, J.; Pielak, R.M.; McClintock, M.A.; Chou, J.J. Solution structure and functional analysis of the influenza B proton channel. *Nat. Struct. Mol. Biol.*, **2009**, *16*, 1267-1271.

[59] Pielak, R.M.; Chou, J.J. Flu channel drug resistance: a tale of two sites *Prot. Cell*, **2010**, *1*, 246-258.

[60] Du, Q.S.; Huang, R.B.; Wang, S. Q.; Chou, K.C. Designing inhibitors of M2 proton channel against H1N1 swine influenza virus. *PLoS ONE*, **2010**, *5*, e9388.

[61] Wang, J.F.; Chou, K.C. Insights from studying the mutation-induced allostery in the M2 proton channel by molecular dynamics. *Prot. Eng. Des. Sel.*, **2010**, *23*, 663-666.

[62] Du, Q.S.; Huang, R.B.; Wang, C.H.; Li, X.M.; Chou, K.C. Energetic analysis of the two controversial drug binding sites of the M2 proton channel in influenza A virus. *J. Theor. Biol.*, **2009**, *259*, 159-164.

[63] Qian, N.; Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **1988**, *202*, 865-884.

[64] Stormo, G.D.; Schneider, T.D.; Gold, L.; Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **1982**, *10*, 2997-3011.

[65] Wu, C.H. Artificial neural networks for molecular sequence analysis. *Comp. Chem.*, **1997**, *21*, 237-256.

[66] Chou, K.C.; Elrod, D.W. Protein subcellular location prediction. *Prot. Eng.*, **1999**, *12*, 107-118.

[67] Denoeux, T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transac. Sys., Man Cybern.*, **1995**, *25*., 804-813.

[68] Chen, K.; Kurgan, M.; Kurgan, L. Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values. *J. Biomed. Sci. Eng. (JBiSE)*, **2008**, *1*, 1-9 (open accessible at http://www.srpublishing.org/journal/jbise/).

[69] Ben-Hur, A.; Ong, C.S.; Sonnenburg, S.; Scholkopf, B.; Ratsch, G. Support Vector Machines and Kernels for Computational Biology. *PLoS Comput. Biol.*, **2008**, *4*, e1000173.

[70] Byvatov, E.; Schneider, G. Support vector machine applications in bioinformatics. *Applied Bioinform.*, **2003**, *2*, 67-77.

[71] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.*, **1990**, *215*, 403-410.

[72] Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **1970**, *48*, 443-453.

[73] Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.*, **1981**, *147*, 195-197.

[74] Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press, 1999.

[75] Eddy, S. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. *PLoS Comput. Biol.*, **2008**, *5*.

[76] Krogh, A.; Brown, M.; Mian, I.S.; Sjolander, K.; Haussler, D. Hidden Markov Models in Computational Biology : Applications to Protein Modeling. *J. Mol. Biol.*, **1994**, *235*, 1501-1531.

[77] Ooi, C.H.; Tan, P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data . *Bioinformatics*, **2003**, *19*, 37-44.

[78] Unger, R.; Moult, J. Genetic Algorithms for Protein Folding Simulations. *J. Mol. Biol.*, **1993**, *231*, 75-81.

[79] Wolfram, S. *A new kind of science*; Wolfram Media Inc., **2002.**

[80] Neumann, J.V. *Theory of Self-Reproducing Automata*; University of Illinois Press, 1966.

[81] Savill, N.J.; Hogeweg, P. Modelling Morphogenesis: From Single Cells to Crawling Slugs. *J. Theor. Biol.*, **1997**, *184*, 229-235.

[82] High-speed computing: scientific applications and algorithm design; Wilhelmson, R.B., Ed.; University of Illinois Press, 1988.

[83] Packard, N.H. In First International Symosium for Science on From, **1986**.

[84] Schonfisch, B. Propagation of fronts in cellular automata. *Physica D: Nonlinear Phenomena*, **1995**, *80*, 433-450.

[85] Springer-VerlagSerge, G. In Proceedings of the 5th International Conference on Cellular Automata for Research and Industry; Springer-Verlag, 2002.

[86] Gardner, M. The fantastic combinations of John Conway's new solitaire game ``life". *Scientific American*, **1970**, *223*, 120-123.

[87] Chou, K.-C.; Zhang, C.-T. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. *AIDS Res. Hum. Retroviruses*, **1992**, *8*, 1967-1976.

[88] Zhang, C.-T.; Chou, K.-C. A Graphic Approach to Analyzing Codon Usage in 1562 Escherichia coli Protein Coding Sequences. *J. Mol. Biol.*, **1994**, *238*, 1-8.

[89] Hu, Z.; Frith, M.; Niu, T.; Weng, Z. SeqVISTA: a graphical tool for sequence feature visualization and comparison. *BMC Bioinform.*, **2003**, *4*, 1.

[90] Kashuk; Carl; Sengupta; Sanghamitra; Eichler; Evan; Chakravarti; Aravinda. ViewGene: a graphical tool for polymorphism visualization and characterization. *Genome Res.*, **2002**, *12*, 333-338.

[91] Liu, Y.; Guo, X.; Xu, J.; Pan, L.; Wang, S. Some Notes on 2-D Graphical Representation of DNA Sequence. *J. Chem. Inform. Comp. Sci.*, **2002**, *42*, 529-533.

[92] Mayor, C.; Brudno, M.; Schwartz, J.R.; Poliakov, A.; Rubin, E.M.; Frazer, K.A.; Pachter, L. S.; Dubchak, I. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **2000**, *16*, 1046-1047.

[93] Nandy, A. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comp. appl. Biosci.*, **1996**, *12*.

[94] Randic, M.; Vracko, M.; Nandy, A.; Basak, S.C. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J. Chem. Inform. Comp. Sci.*, **2000**, *40*, 1235-1244.

[95] Chou, K. C.; Zhang, C.T.; Maggiora, G.M. Disposition of amphiphilic helices in heteropolar environments. *Proteins: Struct., Funct., Genet.*, **1997**, *28*, 99-108.

[96] Hamori, E. Novel DNA sequence representations. *Nature*, **1985**, *314*, 585-586.

[97] Jeffrey, J. Chaos game representation of gene structure. *Nucleic Acids Res.*, **1990**, *18*, 2163-2170.

[98] Zhou, G.P. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into protein-protein interaction mechanism. *J. Theor. Biol.*, **2011**, *284*, 142-148.

[99] Yau, S.S.T.; Wang, J.S.; Niknejad, A.; Lu, C.X.; Jin, N.; Ho, Y.K. DNA sequence representation with degeneracy. *Nucleic Acids Res.*, **2003**, *31*, 3078-3080.

[100] Wu, Z.C.; Xiao, X.; Chou, K.C. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J. Theor. Biol.*, **2010**, in press.

[101] Chou, K.C. Graphic rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.*, **1989**, *264*, 12074-12079.

[102] Zhou, G.P.; Deng, M.H. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. *Biochem. J.*, **1984**, *222*, 169-176.

[103] Chou, K.C. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.*, **1990**, *35*, 1-24.

[104] Andraos, J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. *Can. J. Chem.*, **2008**, *86*, 342-357.

[105] Chou, K.C. Graphic rule for drug metabolism systems. *Curr. Drug Metab.*, **2010**, *11*, 369-378.

[106] Xiao, X.; Shao, S.-H.; Ding, Y.-S.; Chen, X.-J. In *Proceedings of the IEEE International Conference on Systems, Man Cybernetics:* The Hague, Netherlands, 2004.

[107] Xiao, X.; Wang, P.; Chou, K.C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J. Comput. Chem.*, **2009**, *30*, 1414-1423.

[108] Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K.C. Using cellular automata to generate image representation for biological sequences. *Amino Acids*, **2005**, *28*, 29-35.

[109] Wang, M.; Yao, J.S.; Huang, Z.D.; Xu, Z.J.; Liu, G.P.; Zhao, H.Y.; Wang, X.Y.; Yang, J.; Zhu, Y.S.; Chou, K.C. A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. *Med. Chem.*, **2005**, *1*, 39-47.

[110] Chou, K.C.; Wei, D.Q.; Du, Q.S.; Sirois, S.; Zhong, W.Z. Review: Progress in computational approach to drug development against SARS. *Curr. Med. Chem.*, **2006**, *13*, 3263-3270.

[111] Chou, K.C.; Wei, D.Q.; Du, Q.S.; Sirois, S.; Shen, H. B.; Zhong, W. z. In *Proteases in Biology and Disease: Viral proteases and antiviral protease inhibitor therapy*; Lendeckel, U., Hooper, N. M., Eds.; Springer Science: Media B.V., 2009; Vol. 8.

[112] Chou, K.C. In In Gene Cloning & Expression Technologies, Weinrer, P. W., Lu, Q., Eds.; Eaton Publishing: Westborough, MA, Chp. 4, pp. 57-70.; Eaton Publishing: Westborough, MA, 2002.

[113] Chou, K.C. In Automation in Proteomics and Genomics: An Engineering Case-Based Approach (Harvard-MIT interdisciplinary special studies courses), Alterovitz, G., enson, R., and Ramoni, M.F., Eds. Wiley & Sons, Ltd.: West Sussex, UK, 2009; Chap. 5, pp. 97-143. **2009**.

[114] Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics*, **2009**, *6*, 262-274.

[115] Chou, K.C.; Zhang, C.T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.*, **1994**, *269*, 22014-22020.

[116] Nakashima, H.; Nishikawa, K.; Ooi, T. The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **1986**, *99*, 153-162.

[117] Chou, K.C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins: Struct., Funct., Genet.*, **1995**, *21*, 319-344.

[118] Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct. Funct. Genet.*, **2001**, *43*, 246-255.

[119] Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **2005**, *21*, 10-19.

[120] Pan, Y.X.; Zhang, Z.Z.; Guo, Z.M.; Feng, G.Y.; Huang, Z.D.; He, L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Prot. Chem.*, **2003**, *22*, 395-402.

[121] Chen, Y.L.; Li, Q.Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *J. Theor. Biol.*, **2007**, *248*, 377-381.

[122] Kurgan, L.A.; Stach, W.; Ruan, J. Novel scales based on hydrophobicity indices for secondary protein structure. *J. Theor. Biol.*, **2007**, *248*, 354-366.

[123] Perez-Bello, A.; Munteanu, C.R.; Ubeira, F.M.; De Magalhaes, A.L.; Uriarte, E.; Gonzalez-Diaz, H. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. *J. Theor. Biol.*, **2009**, *256*, 458-466.

[124] Lin, H.; Wang, H.; Ding, H.; Chen, Y.L.; Li, Q.Z. Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. *Acta. Biotheoretica*, **2009**, *57*, 321-330.

[125] Nanni, L.; Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **2008**, *34*, 653-660.

[126] Lin, H.; Li, Q.Z. Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components. *J. Comput. Chem.*, **2007**, *28*, 1463-1466.

[127] Qiu, J.D.; Huang, J.H.; Liang, R.P.; Lu, X.Q. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. *Analytical Biochem.*, **2009**, *390*, 68-73.

[128] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.*, **2010**, *263*, 203-209.

[129] Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2009**, *257*, 17-26.

[130] Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol.*, **2008**, *252*, 350-356.

[131] Zeng, Y.H.; Guo, Y.Z.; Xiao, R.Q.; Yang, L.; Yu, L.Z.; Li, M.L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.*, **2009**, *259*, 366–372.

[132] Aguero-Chapin, G.; Varona-Santos, J.; de la Riva, G.A.; Antunes, A.; Gonzalez-Vlla, T.; Uriarte, E.; Gonzalez-Diaz, H. Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from Coffea arabica and prediction of a new sequence. *J. Proteome Res.*, **2009**, *8*, 2122-2128.

[133] Chen, C.; Tian, Y.X.; Zou, X.Y.; Cai, P.X.; Mo, J.Y. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J. Theor. Biol.*, **2006**, *243*, 444-448.

[134] Zhang, G.Y.; Fang, B.S. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. *J. Theor. Biol.*, **2008**, *253*, 310-315.

[135] Mundra, P.; Kumar, M.; Kumar, K.K.; Jayaraman, V.K.; Kulkarni, B.D. Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recognition Lett.*, **2007**, *28*, 1610-1615.

[136] Zhou, G.P.; Cai, Y.D. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. *PROTEINS: Struct., Funct., Bioinform.*, **2006**, *63*, 681-684.

[137] Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.*, **2007**, *248*, 546–551.

[138] Ding, Y.S.; Zhang, T.L. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognition Lett.*, **2008**, *29*, 1887-1892.

[139] Mondal, S.; Bhavna, R.; Mohan Babu, R.; Ramakumar, S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.*, **2006**, *243*, 252-260.

[140] Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Prot. Pept. Lett.*, **2009**, *16*, 27-31.

[141] Ding, H.; Luo, L.; Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Prot. Pept. Lett.*, **2009**, *16*, 351-355.

[142] Jiang, X.; Wei, R.; Zhang, T.L.; Gu, Q. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Prot. Pept. Lett.*, **2008**, *15*, 392-396.

[143] Li, F.M.; Li, Q.Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Prot. Pept. Lett.*, **2008**, *15*, 612-616.

[144] Lin, H.; Ding, H.; Feng-Biao Guo, F.B.; Zhang, A.Y.; Huang, J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Prot. Pept. Lett.*, **2008**, *15*, 739-744.

[145] Qiu, J.D.; Huang, J.H.; Shi, S.P.; Liang, R.P. Using the Concept of Chou's Pseudo Amino Acid Composition to Predict Enzyme Family Classes: An Approach with Support Vector Machine Based on Discrete Wavelet Transform. *Prot. Pept. Lett.*, **2010**, *17*, 715-712.

[146] Gu, Q.; Ding, Y.S.; Zhang, T.L. Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. *Prot. Pept. Lett.*, **2010**, *17*, 559-567.

[147] Mohabatkar, H. Prediction of Cyclin Proteins Using Chou's Pseudo Amino Acid Composition. *Prot. Pept. Lett.*, **2010**.

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

[149] Xiao, X.; Wang, P.; Chou, K.C. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. *J. Theor. Biol.*, **2008**, *254*, 691-696.

[150] Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural features for image classification. *IEEE Transac. Sys., Man, Cybern.*, **1973**, *3*, 610-621.

[151] Diao, Y.; Ma, D.; Wen, Z.; Yin, J.; Xiang, J.; Li, M. Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids*, **2008**, *34*, 111-117.

[152] Xiao, X.; Wang, P.; Chou, K.C. Quat-2L: a web-server for predicting protein quaternary structural attributes. *Mol. Divers.*, **2010**.

[153] Ziv, J.; Lempel, A. On the complexity of finite sequences. *IEEE Trans. Inf. Theory*, **1976**, *IT-22*, 75-81.

[154] Zorzenon dos Santos, R.M.; Coutinho, S. Dynamics of HIV Infection: A Cellular Automata Approach. *Phys. Rev.Lett.*, **2001**, *87*, 168102.

[155] Xiao, X.; Shao, S.-H.; Chou, K.-C. A probability cellular automaton model for hepatitis B viral infections. *Biochem. Biophys. Res. Comm.*, **2006**, *342*, 605-610.

[156] Beauchemin, C.; Samuel, J.; Tuszynski, J. A simple cellular automaton model for influenza A viral infections. *J. Theor. Biol.*, **2005**, *232*, 223-234.

[157] Castiglione, F.; Duca, K.; Jarrah, A.; Laubenbacher, R.; Hochberg, D.; Thorley-Lawson, D. Simulating Epstein-Barr virus infection with C-ImmSim. *Bioinformatics*, **2007**, *23*, 1371-1377.

[158] Kier, L.B.; Cheng, C.K.; Testa, B.; Carrupt, P.-A. A cellular automata model of enzyme kinetics. *J. Mol. Graph.*, **1996**, *14*, 227-231.

[159] Siehs, C.; Oberbauer, R.; Mayer, G.; Lukas, A.; Mayer, B. Discrete simulation of regulatory homo- and heterodimerization in the apoptosis effector phase *Bioinformatics*, **2002**, *18*, 67-76.

[160] Sanford, C.; Yip, M.L.; White, C.; Parkinson, J. Cell++--simulating biochemical pathways. *Bioinformatics*, **2006**, *22*, 2918-2925.

[161] Kier, L.B.; Cheng, C.-K. Cellular Automata Model of Membrane Permeability. *J. Theor. Biol.*, **1997**, *186*, 75-80.

[162] Alarcon, T.; Byrne, H.M.; Maini, P.K. A cellular automaton model for tumour growth in inhomogeneous environment. *J. Theor. Biol.*, **2003**, *225*, 257-274.

[163] de Boer, R.J.; Perelson, A.S. Size and connectivity as emergent properties of a developing immune network. *J. Theor. Biol.*, **1991**, *149*, 381-424.

[164] Burks, C.; Farmer, D. Towards modeling DNA sequences as automata. *Physica D: Nonlinear Phenomena*, **1984**, *10*, 157-167.

[165] Smith, S.A.; Watt, R.C.; Hameroff, S.R. Cellular automata in cytoskeletal lattices. *Physica D: Nonlinear Phenomena*, **1984**, *10*, 168-174.

[166] Young, D.A. A local activator-inhibitor model of vertebrate skin patterns. *Mathematical Biosciences*, **1984**, *72*, 51-58.

[167] Sirakoulis, G.C.; Karafyllidis, I.; Mizas, C.; Mardiris, V.; Thanailakis, A.; Tsalides, P. A cellular automaton model for the study of DNA sequence evolution. *Comput. Biol. Med.*, **2003**, *33*, 439-453.

[168] Mizas, C.; Sirakoulis, G.; Mardiris, V.; Karafyllidis, I.; Glykos, N.; Sandaltzopoulos, R. Reconstruction of DNA sequences using genetic algorithms and cellular automata: towards mutation prediction? *Biosystems*, **2008**, *92*, 61-68.

[169] Moore, J.H.; Hahn, L.W. A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. *Pacific Symposium on Biocomputing*, **2002**, *7*, 53-64.

[170] Moore, J.H.; Hahn, L.W. Cellular automata and genetic algorithms for parallel problem solving in human genetics; Springer: Berlin, ALLEMAGNE, 2002.

[171] Kiran, S.; Ramesh, B. Identification of Promoter Region in Genomic DNA Using Cellular Automata Based Text Clustering. *Int. Arab J. Information Technol.*, **2010**, *7*, 75-78.

[172] Chen, L.; Feng, K.Y.; Cai, Y.D.; Chou, K.C.; Li, H.P. Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. *BMC Bioinformatics*, **2010**, *11*, 293.

[173] Huang, T.; Shi, X.H.; Wang, P.; He, Z.; Feng, K.Y.; Hu, L.; Kong, X.; Li, Y.X.; Cai, Y.D.; Chou, K.C. Analysis and Prediction of the Metabolic Stability of Proteins Based on Their Sequential Features, Subcellular Locations and Interaction Networks. *PLoS ONE*, **2010**, *5*, e10972.

[174] He, Z.S.; Zhang, J.; Shi, X.H.; Hu, L.L.; Kong, X.G.; Cai, Y.D.; Chou, K.C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, **2010**, *5*, e9603.

[175] Chou, K.C. A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Comm.*, **1999**, *264*, 216-224.

Received: April 01, 2011 **Revised: April 01, 2011** **Accepted: May 04, 2011**