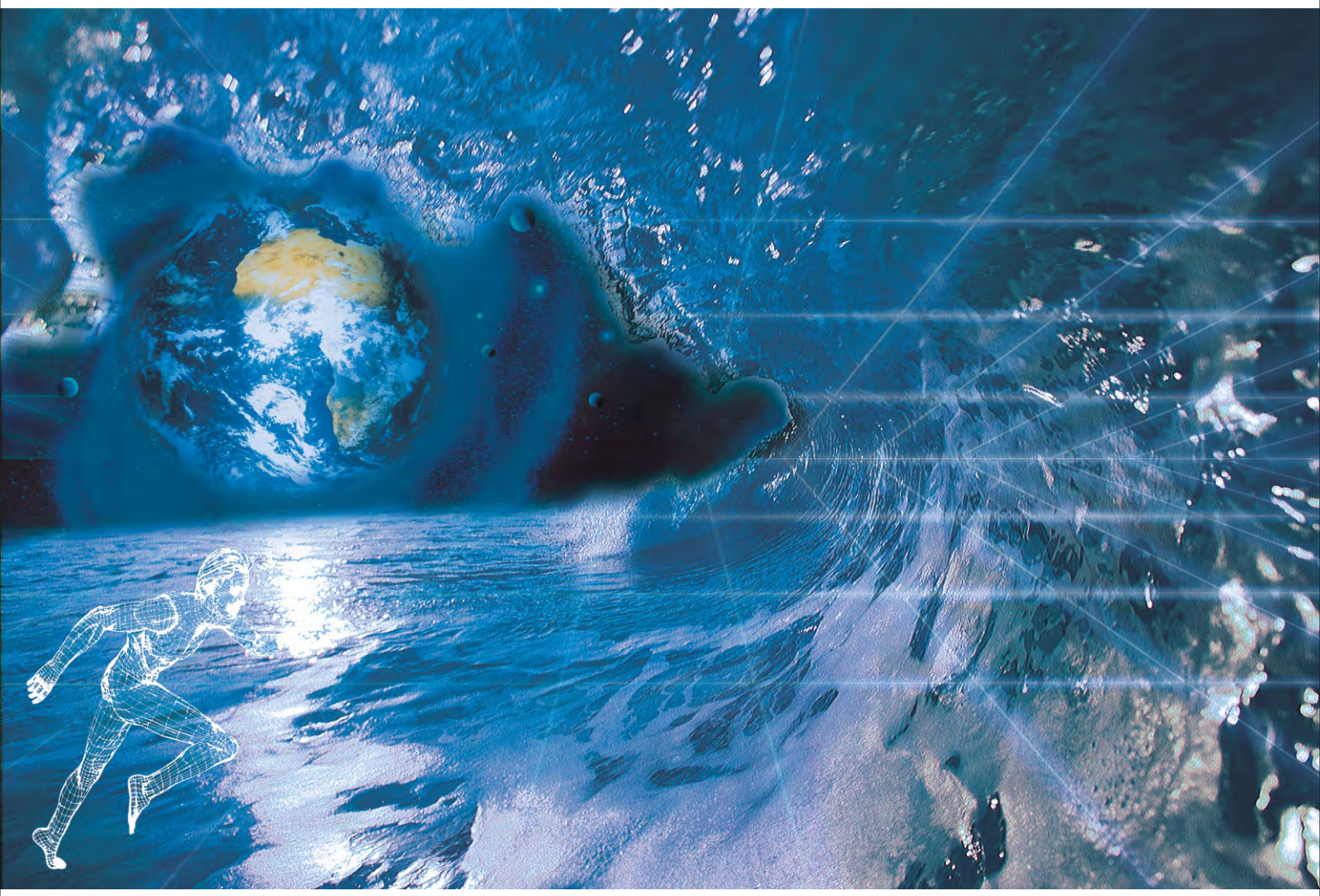
ISSN: 2150-4091 (Print), 2150-4105 (Online) Volume 1, Number 2, September 2009

Scientific Research

Natural Science



www. scirp. org/journal/ns/

Journal Editorial Board

ISSN: 2150-4091 (Print) ISSN: 2150-4105 (Online)
http://www.scirp.org/journal/ns/

TABLE OF CONTENTS

Volume 1, Number 2, Septemebr 2009

Natural Science

Journal Information

SUBSCRIPTIONS

The *Natual Science* (Online at Scientific Research Publishing, www.SciRP.org) is published quarterly by Scientific Research Publishing, Inc.,USA.

E-mail:service@scirp.org

Subscription rates: Volume 1 2009

Print: $50 per copy.

Electronic: free, available on www.SciRP.org.

To subscribe, please contact Journals Subscriptions Department, E-mail: service@scirp.org

Sample copies: If you are interested in subscribing, you may obtain a free sample copy by contacting Scientific Research Publishing, Inc at the above address.

SERVICES

Advertisements

Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)

Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.

E-mail: service@scirp.org

COPYRIGHT

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:

E-mail: ns@scirp.org

Vol.1, No.2, 63-92 (2009)
doi: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.

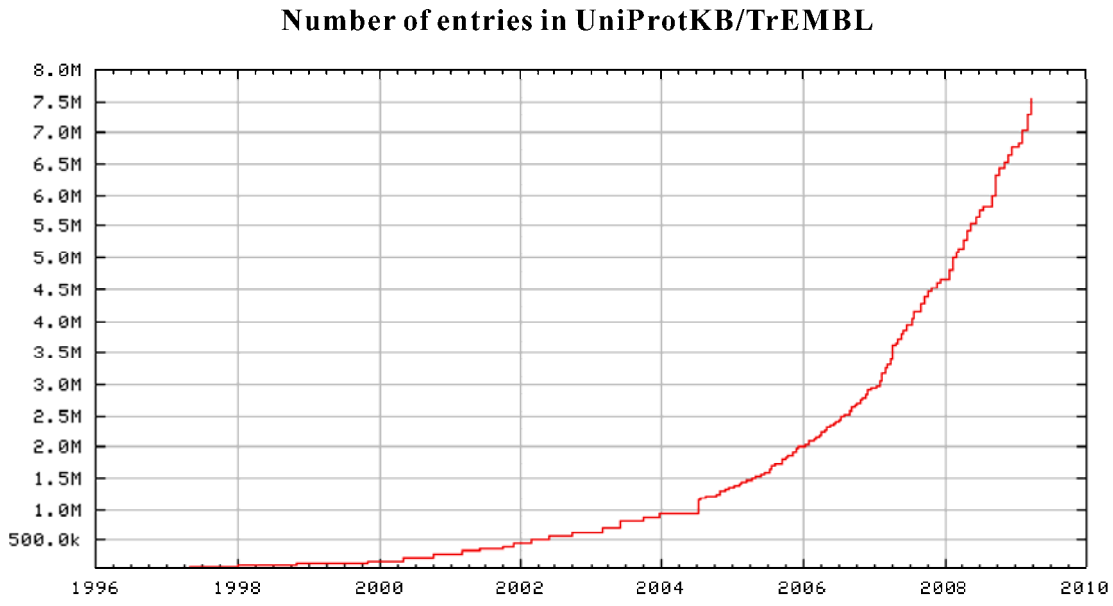


Figure 1. A statistical illustration to show the growth of the UniProtKB/TrEMBL Protein Database (http://www.ebi.ac.uk/uniprot/TrEMBLstats/).

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

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

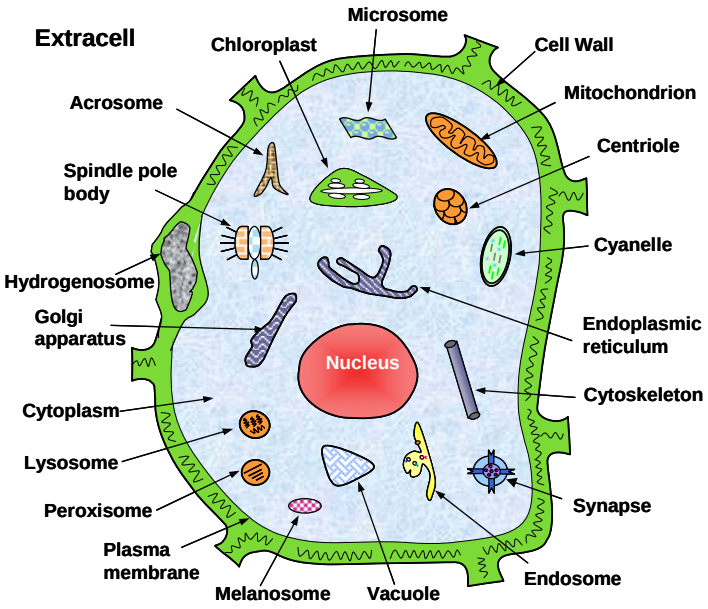


Figure 2. Schematic illustration to show many different components or organelles in a eukaryotic cell. Reproduced from [51] with permission.

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.

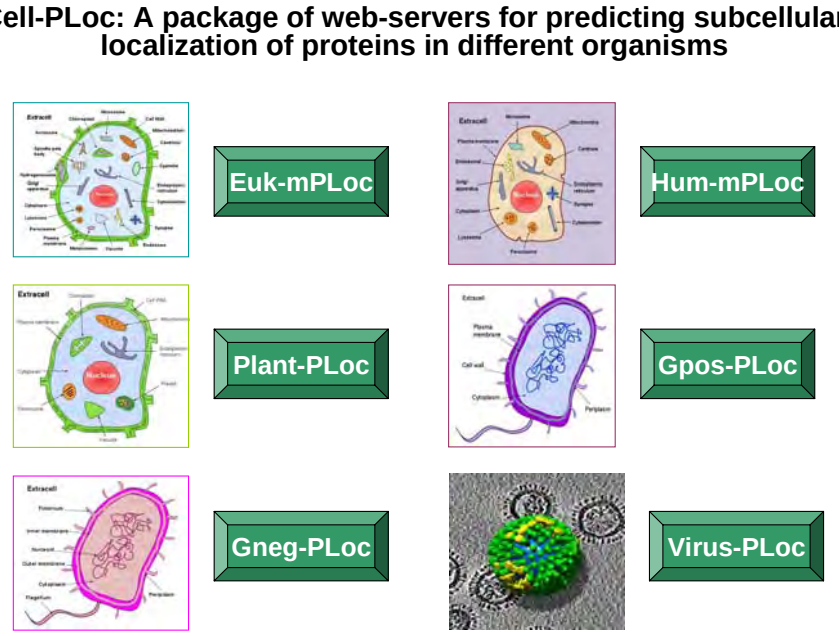


Figure 3. A semi-screenshot to show the Cell-PLoc web-page at (http://chou.med.harvard.edu/bioinf/Cell-PLoc/).

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

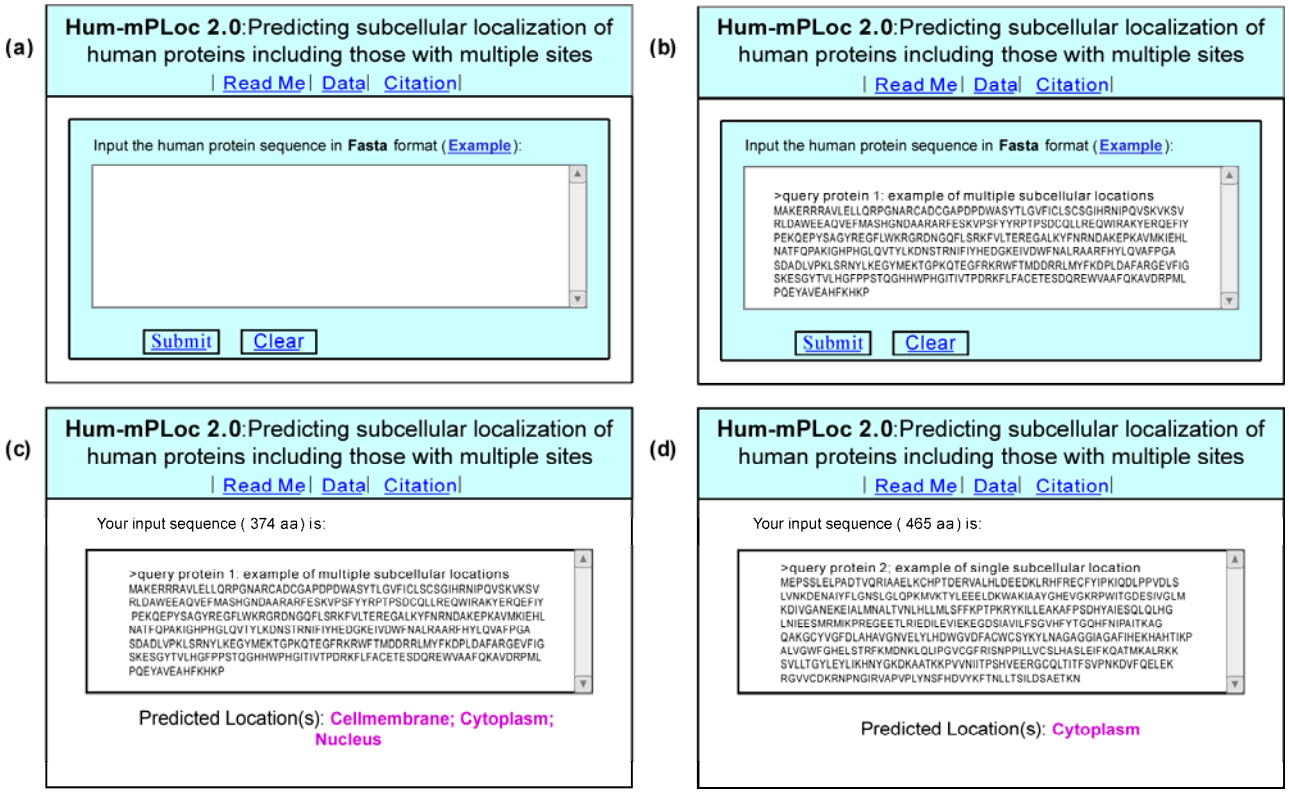


Figure 4. A semi-screenshot to show **(a)** the top page of the web-server Hum-mPLoc 2.0 at http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/, **(b)** the input in FASTA format taken from the 1st query protein sequence in the Example window, **(c)** the output generated by clicking the Submit button in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

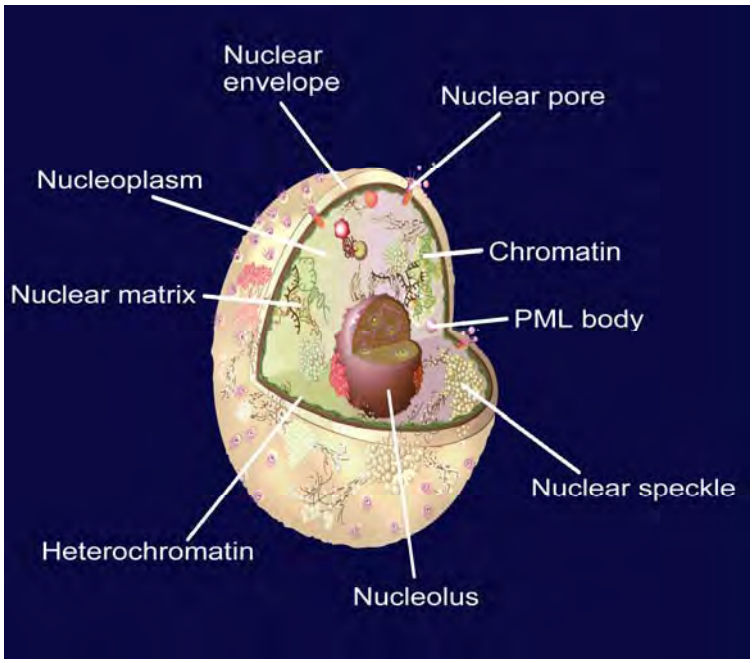


Figure 5. Schematic drawing to show the nine subnuclear locations: (1) chromatin, (2) heterochromatin, (3) nuclear envelope, (4) nuclear matrix, (5) nuclear pore complex, (6) nuclear speckle, (7) nucleolus, (8) nucleoplasm, (9) nuclear PML body. Adapted from [252] with permission.

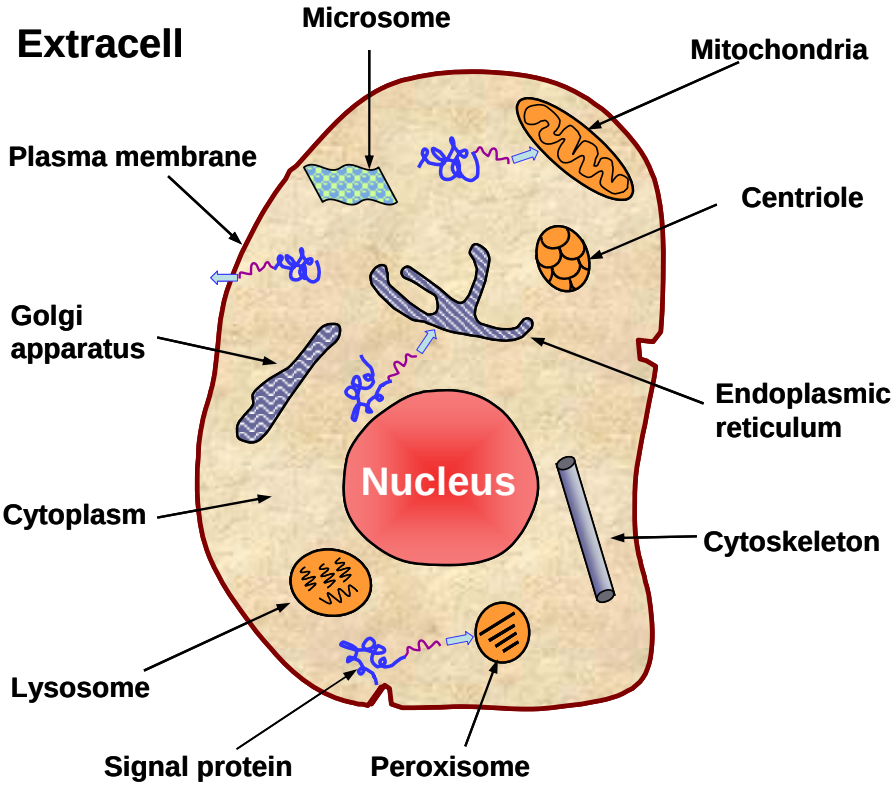


Figure 6. A schematic drawing to show: how the signal peptides of secretory proteins function as an "address tag" in directing the proteins to their proper cellular and extracellular locations. The signal peptide sequence is colored in puple, and the mature protein sequence in blue.

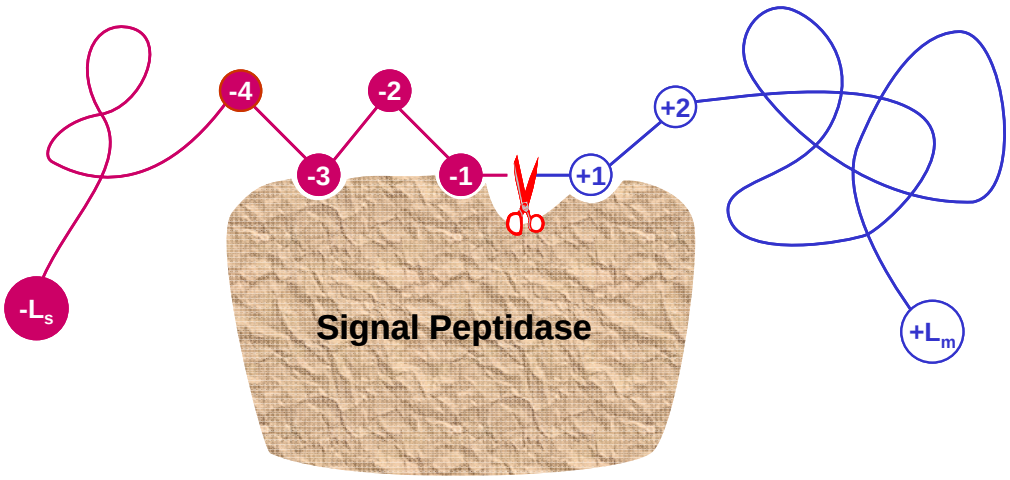


Figure 7. A schematic drawing to show the signal sequence of a protein and how it is cleaved by the signal peptidase. An amino acid in the signal part is depicted as a red circle with a white number to indicate its sequential position, while that in the mature protein depicted as an open circle with 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;

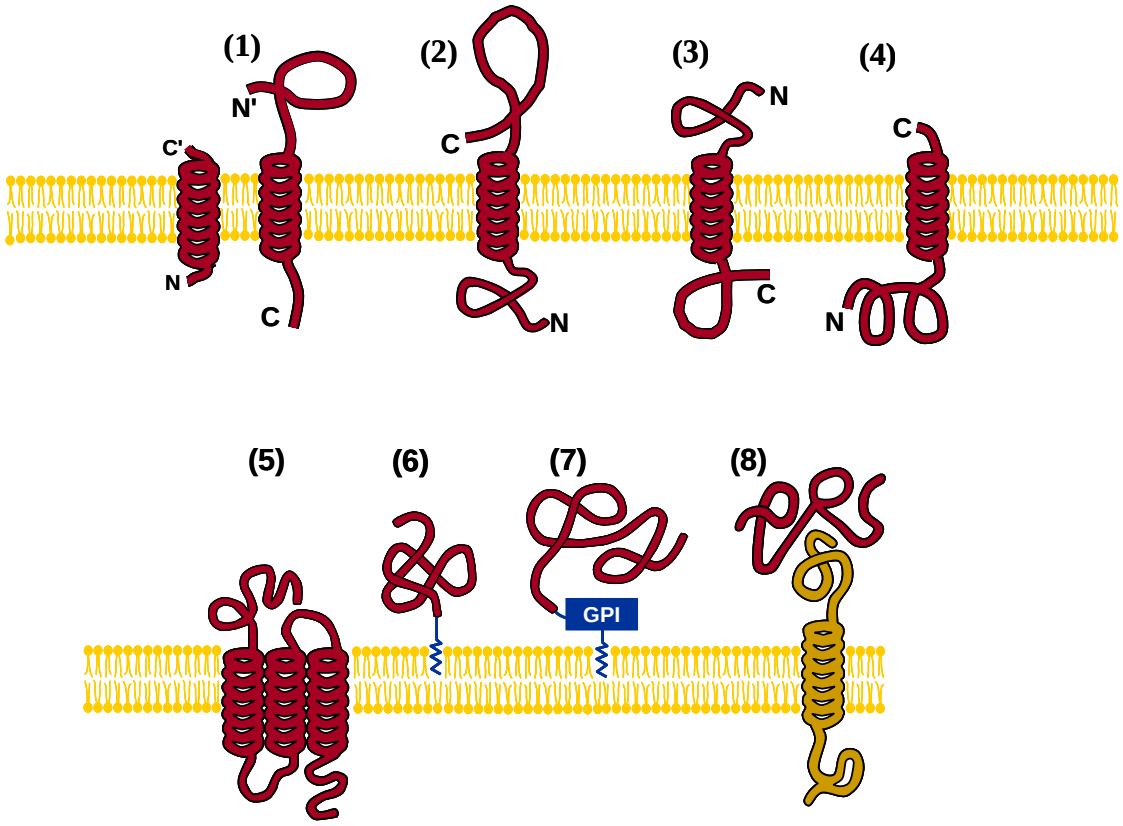


Figure 8. Schematic illustration showing the 8 types of membrane proteins: (1) type I transmembrane, (2) type II, (3) type III, (4) type IV, (5) multipass transmembrane, (6) lipid-chain-anchored membrane, (7) GPI-anchored membrane, and (8) peripheral membrane. As shown in the figure, types I, II, III, 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"

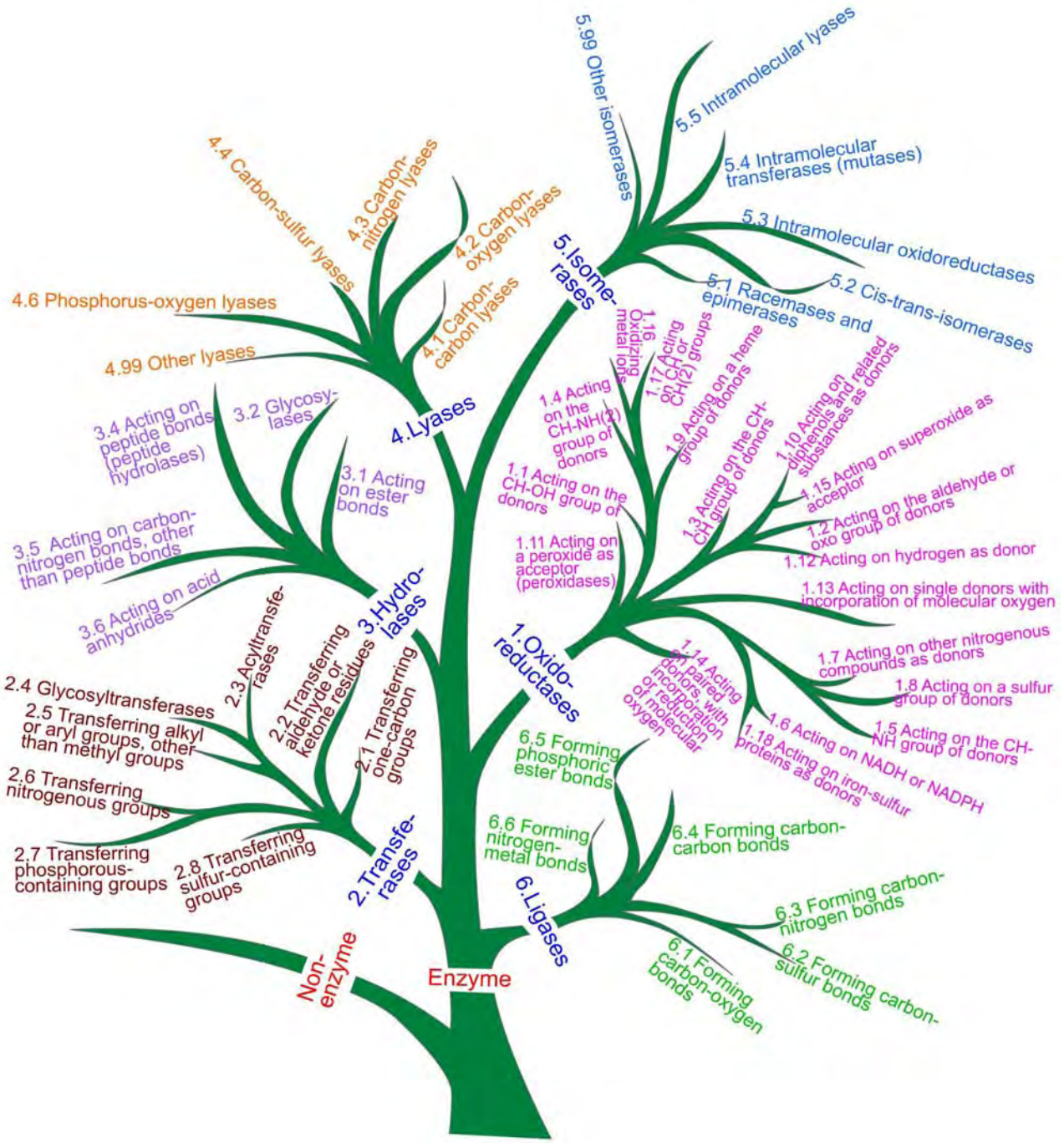


Figure 9. A schematic drawing to use tree branches to classify enzyme and non-enzyme as well as the six main functional classes of enzymes and their subclasses.

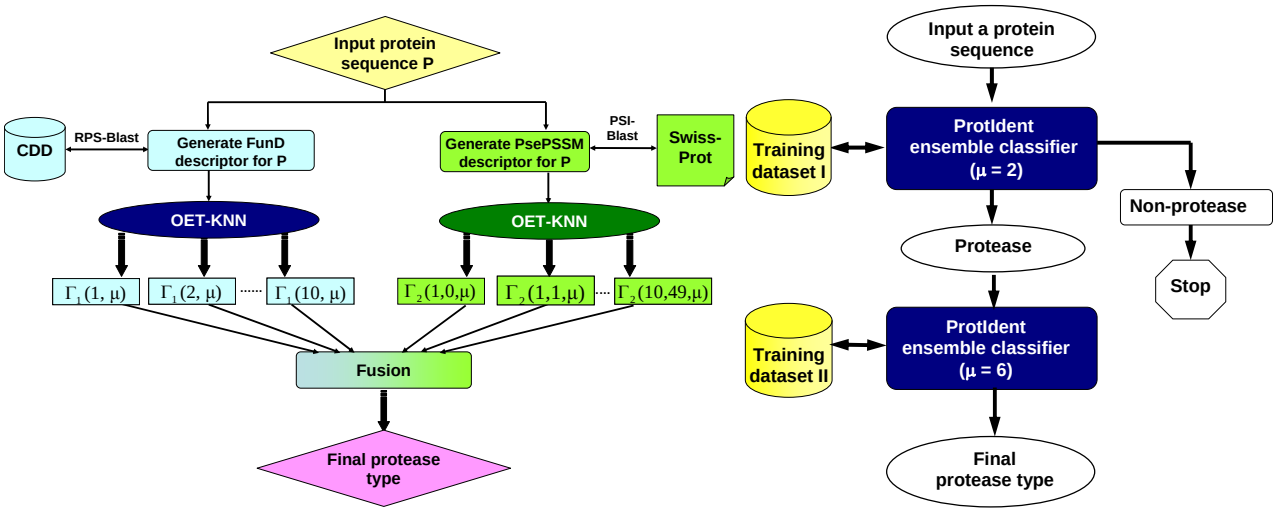


Figure 10. A flowchart to show **(a)** how to fuse the FunD approach and PsePSSM approach, and **(b)** how the two-layer Prot-Ident ensemble classifier works in identifying proteases and their functional types. See [55] for further explanation.

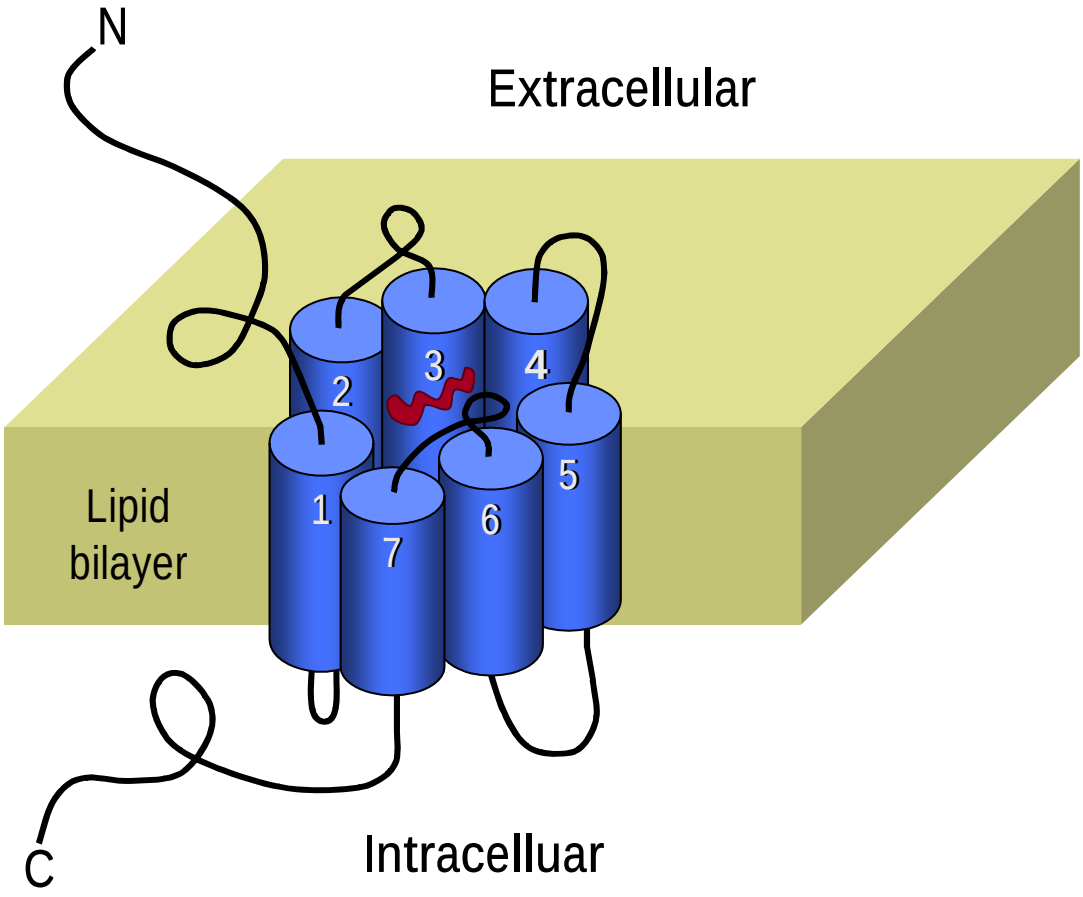


Figure 11. Schematic representation of a GPCR with a trademark of seven-transmembrane helices, depicted as cylinders and connected by alternating cytoplasmic and extracellular hydrophilic loops. The 7-helix bundle thus formed has a central pore on its extracellular surface. The red entity located in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

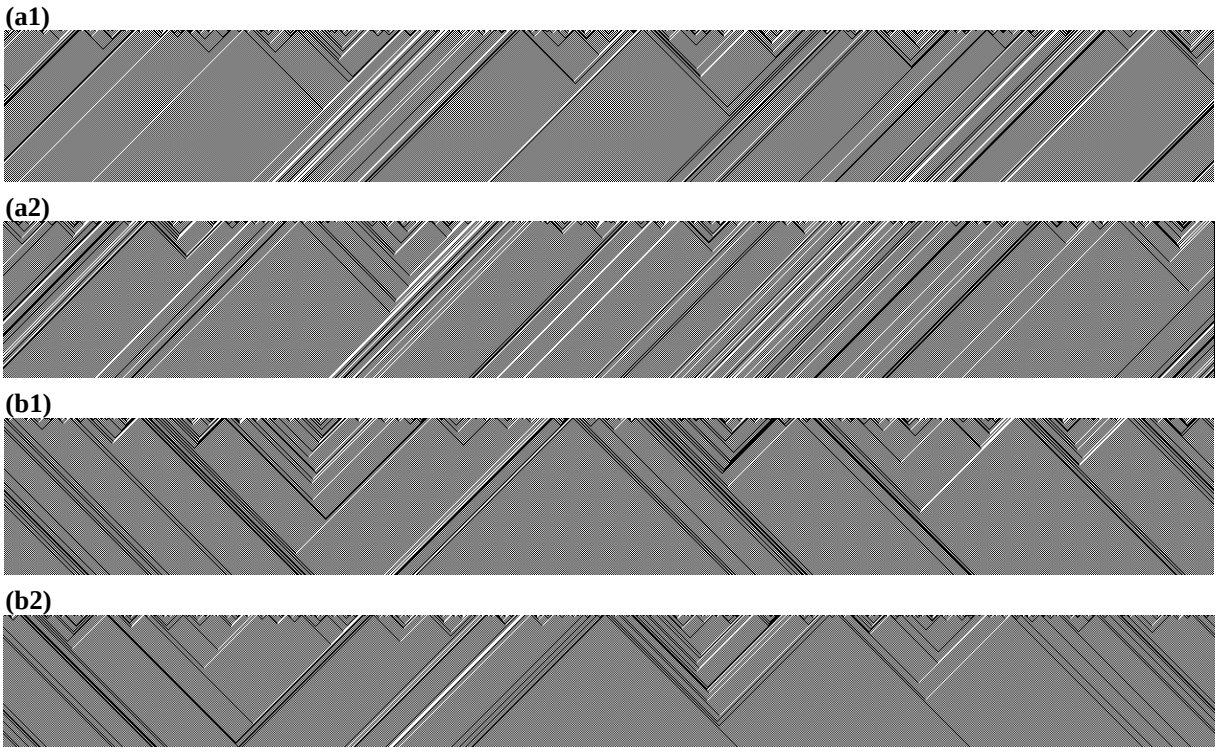


Figure 12. The cellular automaton image generated according to Eqs.2-5 for (**a1)** the rhodopsin like family member with accession number P41595; **(a2**) the rhodopsin like family member with accession number P18599; **(b1)** the secretin like family member with accession number O95838; and **(b2**) 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).

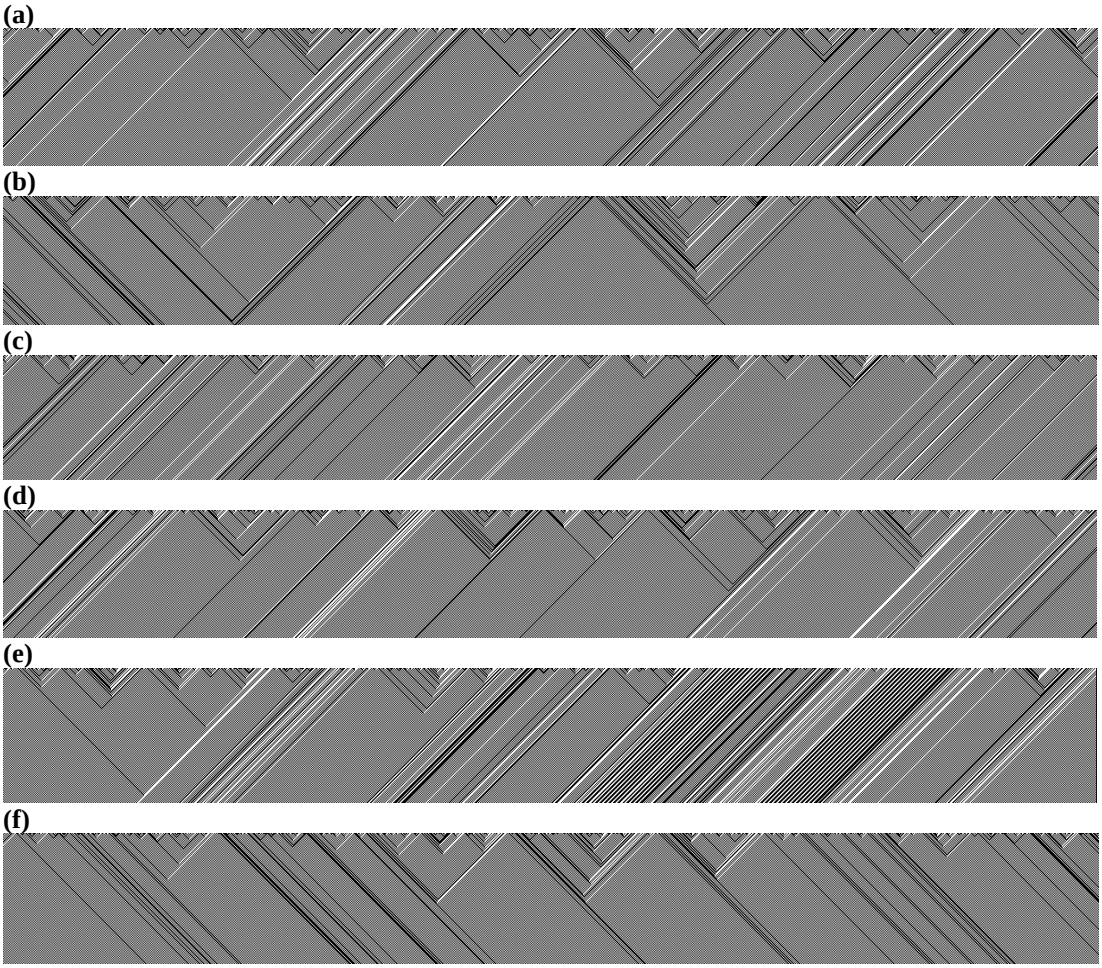


Figure 13. The cellular automaton image generated according to Eqs.2-5 for a protein taken from **(a)** A-rhodopsin like family, **(b)** B-secretin like family, **(c)** C-metabotrophic/glutamate/pheromone family; **(d)** D-fungal pheromone family, **(e)** E-cAMP receptor family, and **(f)** F-Frizzled/Smoothemed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

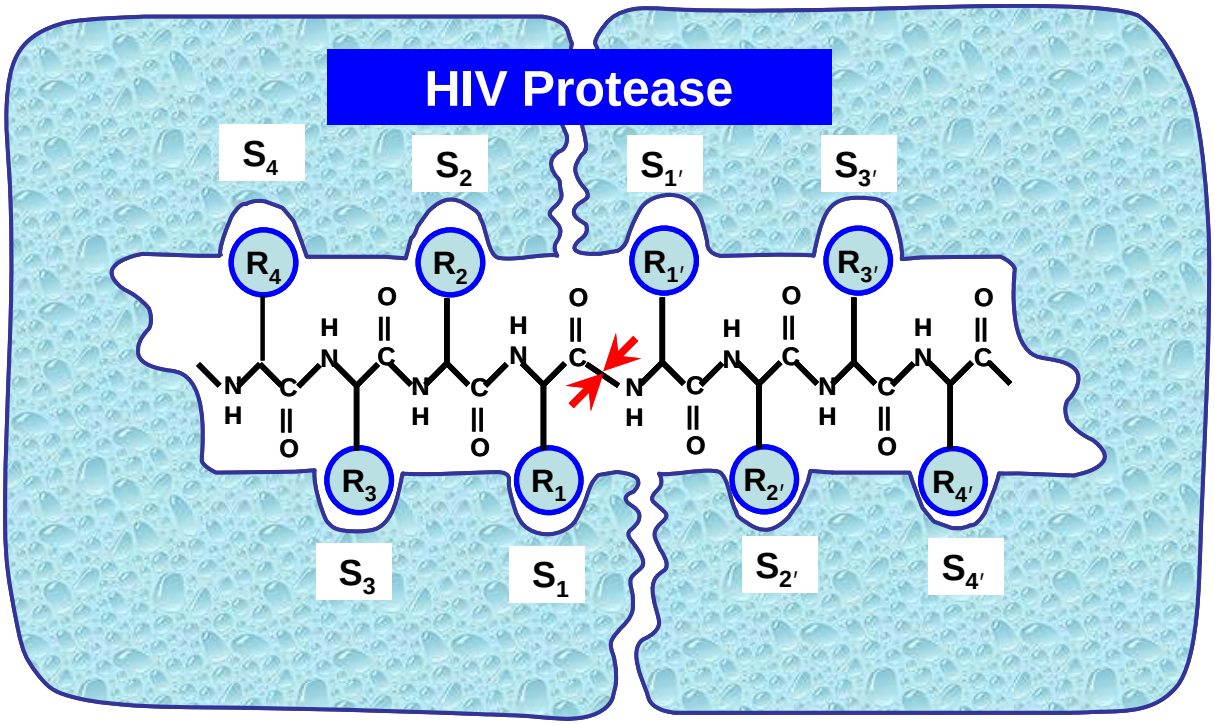


Figure 14. Schematic representation of substrate bound to HIV protease based an analysis of protease-inhibitor crystal structures. The active site of enzyme is composed of eight extended "subsites", S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, $S_{4'}$, and their counterparts in a substrate extend 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].

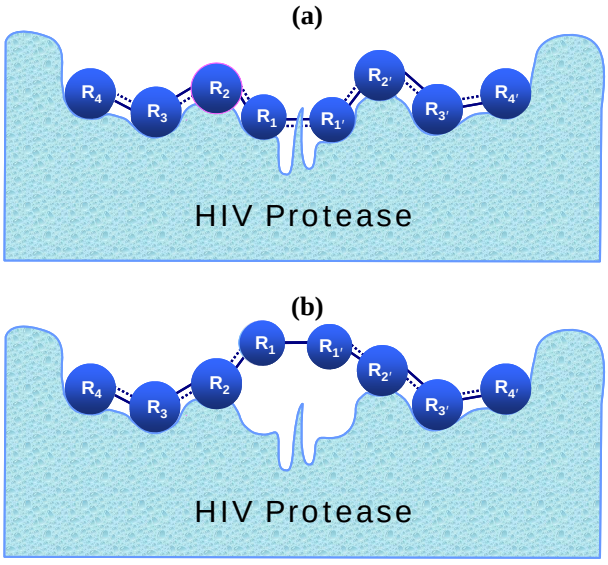


Figure 15. Schematic illustration to show (a) a cleavable octapeptide is chemically effectively bound to the active site of HIV protease, and (b) although still bound to the active site, the peptide has lost its cleavability after its scissile bond is modified from a hybrid peptide bond [254] to a single 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.

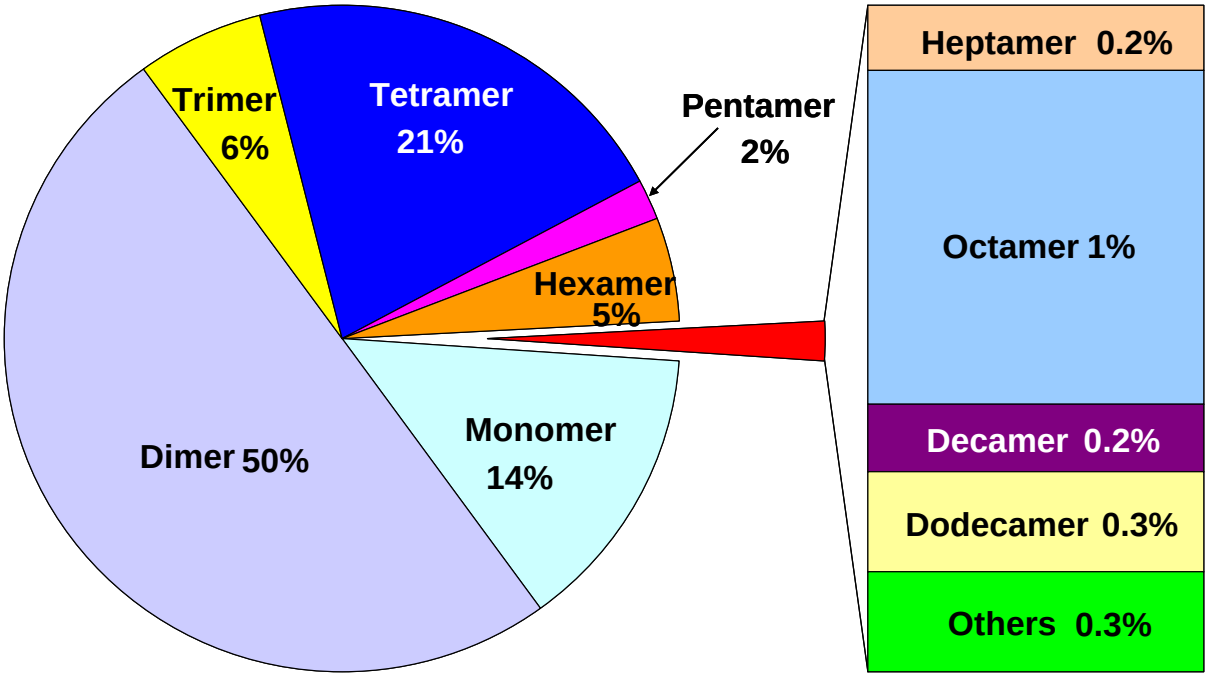


Figure 16. A pie chart to show the statistical distribution of different quaternary structural types in the nature derived from version 55.3 of Swiss-Prot database released 29-April-2008. Reproduced with permission from [197].

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

$$\begin{cases} \dfrac{\mathrm{d}P_{unfold}(t)}{\mathrm{d}t} = -K_f P_{unfold}(t) \\ \dfrac{\mathrm{d}P_{fold}(t)}{\mathrm{d}t} = K_f P_{unfold}(t) \end{cases} \quad (6)$$

where $P_{unfold}(t)$ and $P_{fold}(t)$ 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., $P_{unfold}(t) = C_0$ and $P_{fold}(t) = 0$ 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

$$\begin{cases} P_{unfold}(t) = C_0 \exp(-K_f t) \\ P_{fold}(t) = C_0 \left[1 - \exp(-K_f t)\right] \end{cases} \quad (7)$$

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

$$T_{1/2} = -\frac{\ln(1/2)}{K_f} \cong 0.693/K_f \quad (8)$$

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

$$P_{unfold} \xrightarrow{k_{12}} P_{inter} \xrightarrow{k_{23}} P_{fold} \quad (9)$$

where $P_{inter}(t)$ represents the concentration of an intermediate state between the unfolded and folded states, k_{12} is the rate constant for P_{unfold} converting to P_{inter}, and k_{23} the rate constant for P_{inter} converting to P_{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" 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 G is further transformed to the phase digraph $\tilde{\mathrm{G}}$ [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 $\tilde{\mathrm{G}}$ 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

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

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

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

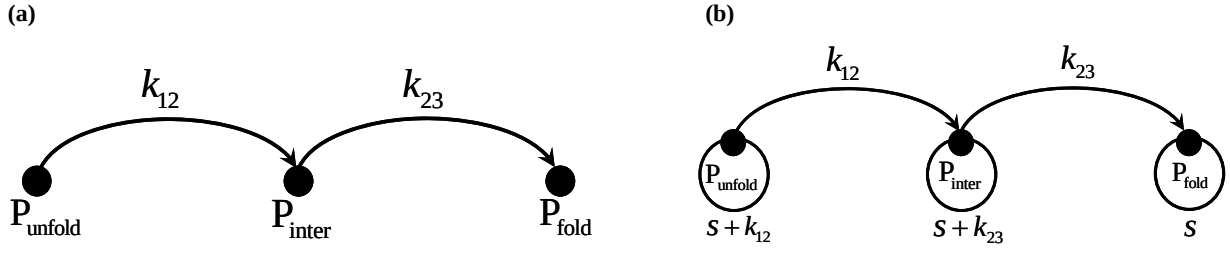


Figure 17. (a) The directed graph or digraph G [245,246] for the three-state protein folding mechanism as schematically expressed by **Eq.9** and formulated by **Eq.10**. **(b)** The phase digraph $\tilde{\mathrm{G}}$ obtained from G of panel (a) according to graphic rule 4 for enzyme and protein folding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

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

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

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

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

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]

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

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.

Appendix A. A comparison between the predicted results by Euk-mPLoc and the experimental results reported latter. Listed in column 4 are the predicted results (marked in blue); those in column 5 are the experimental results. The comments in column 6 indicate whether the proteins concerned are with single location or multiple locations; the comment content is colored in red when the prediction is inconsistent or partly inconsistent with observation. See the text for further explanation. (For interpretation of the references to color in this caption, the reader is referred to the web version of this paper.)

No.	Accession Number	Subcellular location annotated in Swiss-Prot 50.7 released on 19-Sept-2006	Subcellular location predicted prior to experimental reports by **Euk-mPLoc** before November 2006	Subcellular location observed by experiments later and annotated in Swiss-Prot 53.2 released on 26-June-2007	Comment
1	O13674	Unknown	Cytoplasm	Cytoplasm	Single
2	O13699	Unknown	Mitochondrion	Mitochondrion	Single
3	O13715	Unknown	Cytoplasm	Cytoplasm	Single
4	O13795	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
5	O13826	Unknown	Nucleus	Nucleus	Single
6	O13859	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
7	O13894	Unknown	Nucleus	Nucleus (nucleolus)	Single
8	O14013	Unknown	Nucleus	Nucleus (nucleolus)	Single
9	O14015	Unknown	Cytoplasm; Nucleus.	Cytoplasm; Nucleus	Multiple
10	O14019	Unknown	Cytoplasm	Cytoplasm	Single
11	O14077	Unknown	Cytoplasm	Cytoplasm	Single
12	O14140	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
13	O14183	Unknown	Cytoplasm	Cytoplasm	Single
14	O14185	Unknown	Cytoplasm	Cytoplasm (localizes to the barrier septum)	Single
15	O14202	Unknown	Mitochondrion	Mitochondrion (mitochondrial inner membrane; single-pass membrane protein)	Single
16	O14216	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
17	O14235	Unknown	Mitochondrion	Mitochondrion	Single
18	O14455	Unknown	Cytoplasm	Cytoplasm	Single
19	O42654	Unknown	Cytoplasm	Cytoplasm (cell cortex)	Single
20	O42980	Unknown	Cytoplasm	Cytoplasm	Single
21	O43541	Unknown	Nucleus	Nucleus	Single
22	O47950	Unknown	Mitochondrion	Mitochondrion	Single
23	O60094	Unknown	Nucleus	Nucleus	Single
24	O74317	Unknown	Mitochondrion	Mitochondrion	Single
25	O74381	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
26	O74405	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
27	O74531	Unknown	Cytoplasm	Cytoplasm	Single
28	O74783	Unknown	Mitochondrion	Mitochondrion	Single

29	O74854	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus (constantly expressed throughout the cell cycle; expressed in nucleus except the nucleolus and is localized at cell tips on both sides of the septum in septated cells)	Multiple
30	O74910	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
31	O75251	Unknown	Mitochondrion	Mitochondrion	Single
32	O80448	Unknown	Cytoplasm	Cytoplasm	Single
33	O94334	Unknown	Cytoplasm	Cytoplasm	Single
34	O94435	Unknown	Cytoplasm	Cytoplasm	Single
35	O94661	Unknown	Endoplasmic reticulum	Endoplasmic reticulum (endoplasmic reticulum membrane; single-pass membrane protein)	Single
36	O94665	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
37	O94668	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
38	P01014	Unknown	Secreted protein	Secreted protein	Single
39	P01023	Unknown	Secreted protein.	Secreted protein	Single
40	P01025	Unknown	Secreted protein	Secreted protein	Single
41	P01026	Unknown	Secreted protein	Secreted protein	Single
42	P01029	Unknown	Secreted protein	Secreted protein	Single
43	P01031	Unknown	Secreted protein	Secreted protein	Single
44	P01032	Unknown	Secreted protein	Secreted protein	Single
45	P01034	Unknown	Secreted protein	Secreted protein	Single
46	P01035	Unknown	Secreted protein	Secreted protein	Single
47	P01036	Unknown	Secreted protein	Secreted protein	Single
48	P01037	Unknown	Secreted protein	Secreted protein	Single
49	P01038	Unknown	Secreted protein	Secreted protein	Single
50	P01127	Unknown	Secreted protein	Secreted protein	Single
51	P01356	Unknown	Secreted protein	Secreted protein	Single
52	P02400	Unknown	Cytoplasm	Cytoplasm	Single
53	P02405	Unknown	Cytoplasm	Cytoplasm	Single
54	P02407	Unknown	Cytoplasm	Cytoplasm	Single
55	P02735	Unknown	Secreted protein	Secreted protein	Single
56	P02738	Unknown	Secreted protein	Secreted protein	Single
57	P02739	Unknown	Secreted protein	Secreted protein	Single
58	P02740	Unknown	Secreted protein	Secreted protein	Single
59	P03952	Unknown	Secreted protein	Secreted protein	Single
60	P04003	Unknown	Secreted protein	Secreted protein	Single
61	P04085	Unknown	Secreted protein	Secreted protein	Single
62	P04449	Unknown	Cytoplasm	Cytoplasm	Single
63	P04551	Unknown	Cytoplasm	Cytoplasm	Single
64	P05318	Unknown	Cytoplasm	Cytoplasm	Single
65	P05319	Unknown	Cytoplasm	Cytoplasm	Single

66	P05735	Unknown	Cytoplasm	Cytoplasm	Single
67	P05736	Unknown	Cytoplasm	Cytoplasm	Single
68	P05737	Unknown	Cytoplasm	Cytoplasm	Single
69	P05738	Unknown	Cytoplasm	Cytoplasm	Single
70	P05745	Unknown	Cytoplasm	Cytoplasm	Single
71	P05747	Unknown	Cytoplasm	Cytoplasm	Single
72	P05749	Unknown	Cytoplasm	Cytoplasm	Single
73	P05753	Unknown	Cytoplasm	Cytoplasm	Single
74	P06307	Unknown	Secreted protein	Secreted protein	Single
75	P06684	Unknown	Secreted protein	Secreted protein	Single
76	P06911	Unknown	Secreted protein	Secreted protein	Single
77	P07279	Unknown	Cytoplasm	Cytoplasm	Single
78	P07280	Unknown	Cytoplasm	Cytoplasm	Single
79	P07281	Unknown	Cytoplasm	Cytoplasm	Single
80	P08607	Unknown	Secreted protein	Secreted protein	Single
81	P08621	Unknown	Nucleus	Nucleus	Single
82	P08649	Unknown	Secreted protein	Secreted protein	Single
83	P09040	Unknown	Secreted protein	Secreted protein	Single
84	P09240	Unknown	Secreted protein	Secreted protein	Single
85	P09859	Unknown	Secreted protein	Secreted protein	Single
86	P09932	Unknown	Nucleus	Nucleus	Single
87	P0C0L4	Unknown	Secreted protein	Secreted protein	Single
88	P0C0L5	Unknown	Secreted protein	Secreted protein	Single
89	P0C0T4	Unknown	Cytoplasm;	Cytoplasm	Single
90	P0C0V8	Unknown	Cytoplasm	Cytoplasm	Single
91	P0C0W9	Unknown	Cytoplasm	Cytoplasm	Single
92	P0C0X0	Unknown	Cytoplasm	Cytoplasm	Single
93	P10622	Unknown	Cytoplasm	Cytoplasm	Single
94	P10664	Unknown	Cytoplasm	Cytoplasm	Single
95	P12082	Unknown	Secreted protein	Secreted protein	Single
96	P14127	Unknown	Cytoplasm	Cytoplasm	Single
97	P14272	Unknown	Secreted protein	Secreted protein	Single
98	P14605	Unknown	Cytoplasm; Membrane	Cytoplasm	Single
99	P14796	Unknown	Cytoplasm	Cytoplasm	Single
100	P14841	Unknown	Secreted protein	Secreted protein	Single
101	P15638	Unknown	Secreted protein	Secreted protein	Single
102	P17076	Unknown	Cytoplasm	Cytoplasm	Single
103	P17079	Unknown	Chloroplast; Cytoplasm	Cytoplasm	Single
104	P17157	Unknown	Centriole; Nucleus	Cytoplasm; Nucleus	Multiple
105	P17248	Unknown	Cytoplasm	Cytoplasm	Single
106	P17629	Unknown	Nucleus	Nucleus	Single
107	P19313	Unknown	Secreted protein	Secreted protein	Single
108	P19707	Unknown	Secreted protein	Secreted protein	Single
109	P19708	Unknown	Secreted protein	Secreted protein	Single
110	P19823	Unknown	Secreted protein	Secreted protein	Single

111	P19827	Unknown	Secreted protein	Secreted protein	Single
112	P20033	Unknown	Secreted protein	Secreted protein	Single
113	P20851	Unknown	Secreted protein	Secreted protein	Single
114	P21651	Unknown	Nucleus	Nucleus (localizes to	Single
115	P22227	Unknown	Nucleus	Nucleus	Single
116	P22298	Unknown	Secreted protein	Secreted protein	Single
117	P23023	Unknown	Nucleus	Nucleus	Single
118	P23248	Unknown	Cytoplasm	Cytoplasm	Single
119	P23362	Unknown	Secreted protein	Secreted protein	Single
120	P23381	Unknown	Cytoplasm	Cytoplasm.	Single
121	P23699	Unknown	Acrosome	Secreted protein	Single
122	P24000	Unknown	Cytoplasm	Cytoplasm	Single
123	P25328	Unknown	Cytoplasm	Cytoplasm (the virus has no extracellular transmission pathway; it exists as a ribonucleoprotein viral particle in the host cytoplasm)	Single
124	P25355	Unknown	Cytoplasm	Cytoplasm	Single
125	P25454	Unknown	Nucleus	Nucleus (localizes as foci on meiotic chromosomes)	Single
126	P25574	Unknown	Endoplasmic reticulum	Endoplasmic reticulum (Endoplasmic reticulum membrane; single-pass type 1 membrane protein)	Single
127	P25586	Unknown	Nucleus	Nucleus (nucleolus)	Single
128	P26262	Unknown	Secreted protein	Secreted protein	Single
129	P26781	Unknown	Cytoplasm	Cytoplasm	Single
130	P26782	Unknown	Cytoplasm; Mitochondrion	Cytoplasm	Single
131	P28325	Unknown	Secreted protein	Secreted protein	Single
132	P28576	Unknown	Secreted protein	Secreted protein	Single
133	P29453	Unknown	Cytoplasm	Cytoplasm	Single
134	P30183	Unknown	Centriole; Nucleus	Nucleus	Single
135	P31532	Unknown	Secreted protein	Secreted protein	Single
136	P32344	Unknown	Mitochondrion	Mitochondrion	Single
137	P32452	Unknown	Chloroplast; Cytoplasm	Cytoplasm	Single
138	P32769	Unknown	Cytoplasm	Cytoplasm	Single
139	P32827	Unknown	Cytoplasm	Cytoplasm	Single
140	P32841	Unknown	Cytoplasm; Nucleus	Nucleus (localizes to chromosomes)	Single
141	P32921	Unknown	Cytoplasm	Cytoplasm	Single
142	P33420	Unknown	Cytoplasm	Cytoplasm (localizes to spindle poles throughout the cell cycle)	Single
143	P33442	Unknown	Cytoplasm	Cytoplasm	Single

144	P34007	Unknown	Secreted protein	Secreted protein	Single
145	P34217	Unknown	Cytoplasm	Cytoplasm	Single
146	P34241	Unknown	Nucleus	Nucleus (accumulates in the immediate vicinity of the dense fibrillar component of nucleolus)	Single
147	P34544	Unknown	Nucleus	Nucleus	Single
148	P35481	Unknown	Secreted protein	Secreted protein	Single
149	P35541	Unknown	Secreted protein	Secreted protein	Single
150	P35542	Unknown	Secreted protein	Secreted protein	Single
151	P35735	Unknown	Membrane	Cell membrane (multi-pass membrane protein)	Single
152	P35997	Unknown	Cytoplasm	Cytoplasm	Single
153	P36013	Unknown	Mitochondrion	Mitochondrion (mitochondrial matrix)	Single
154	P36038	Unknown	Mitochondrion	Mitochondrion	Single
155	P36056	Unknown	Mitochondrion	Mitochondrion	Single
156	P36105	Unknown	Cytoplasm	Cytoplasm	Single
157	P36138	Unknown	Cytoplasm	Cytoplasm	Single
158	P36141	Unknown	Mitochondrion	Mitochondrion	Single
159	P38175	Unknown	Mitochondrion	Mitochondrion	Single
160	P38212	Unknown	Endoplasmic reticulum; Golgi	Endoplasmic reticulum (endoplasmic reticulum membrane; single-pass type 1 membrane protein)	Single
161	P38260	Unknown	Cytoplasm	Cytoplasm	Single
162	P38289	Unknown	Mitochondrion	Mitochondrion	Single
163	P38324	Unknown	Nucleus	Nucleus	Single
164	P38334	Unknown	Golgi	Golgi apparatus (cis-Golgi network)	Single
165	P38339	Unknown	Cytoplasm	Cytoplasm (bud and bud neck)	Single
166	P38344	Unknown	Cytoplasm	Cytoplasm	Single
167	P38711	Unknown	Cytoplasm	Cytoplasm	Single
168	P38754	Unknown	Cytoplasm	Cytoplasm	Single
169	P38779	Unknown	Nucleus	Nucleus (nucleolus)	Single
170	P38783	Unknown	Mitochondrion	Mitochondrion	Single
171	P38813	Unknown	Endoplasmic reticulum; Golgi	Endoplasmic reticulum (endoplasmic reticulum membrane; single-pass type 1 membrane protein)	Single
172	P38844	Unknown	Cell wall; Secreted protein	Cell wall (lipid-anchor; GPI-anchored cell wall protein)	Single
173	P38961	Unknown	Nucleus	Nucleus (nucleolus)	Single

174	P39016	Unknown	Cytoplasm	Cytoplasm	Single
175	P39729	Unknown	Cytoplasm	Cytoplasm	Single
176	P39732	Unknown	Cytoplasm	Cytoplasm	Single
177	P39741	Unknown	Cytoplasm	Cytoplasm	Single
178	P39939	Unknown	Cytoplasm	Cytoplasm	Single
179	P40005	Unknown	Cytoplasm	Cytoplasm	Single
180	P40033	Unknown	Mitochondrion	Mitochondrion	Single
181	P40048	Unknown	Cytoplasm	Cytoplasm	Single
182	P40096	Unknown	Nucleus	Nucleus	Single
183	P40186	Unknown	Cytoplasm	Cytoplasm	Single
184	P40212	Unknown	Cytoplasm	Cytoplasm	Single
185	P40213	Unknown	Cytoplasm	Cytoplasm	Single
186	P40215	Unknown	Mitochondrion	Mitochondrion (mitochondrial intermembrane space)	Single
187	P40453	Unknown	Cytoplasm; Nucleus	Cytoplasm	Single
188	P40496	Unknown	Mitochondrion	Mitochondrion	Single
189	P40525	Unknown	Cytoplasm	Cytoplasm	Single
190	P40530	Unknown	Mitochondrion	Mitochondrion (mitochondrial matrix)	Single
191	P40558	Unknown	Cytoplasm	Cytoplasm	Single
192	P40976	Unknown	Chloroplast; Cytoplasm	Cytoplasm	Single
193	P41056	Unknown	Cytoplasm	Cytoplasm	Single
194	P41057	Unknown	Cytoplasm	Cytoplasm	Single
195	P41058	Unknown	Cytoplasm	Cytoplasm	Single
196	P41229	Unknown	Nucleus	Nucleus	Single
197	P41520	Unknown	Secreted protein	Secreted protein	Single
198	P42027	Unknown	Mitochondrion	Mitochondrion	Single
199	P42028	Unknown	Mitochondrion	Mitochondrion	Single
200	P42819	Unknown	Secreted protein	Secreted protein	Single
201	P42846	Unknown	Nucleus	Nucleus (nucleolus)	Single
202	P43565	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
203	P46784	Unknown	Cytoplasm; Mitochondrion	Cytoplasm	Single
204	P46990	Unknown	Chloroplast; Cytoplasm	Cytoplasm	Single
205	P46995	Unknown	Nucleus	Nucleus	Single
206	P47006	Unknown	Nucleus	Nucleus (nucleolus)	Single
207	P47025	Unknown	Mitochondrion	Mitochondrion (mitochondrial outer membrane; cytoplasmic side)	Single
208	P47076	Unknown	Nucleus	Nucleus	Single
209	P47108	Unknown	Nucleus	Nucleus (nucleolus)	Single
210	P47122	Unknown	Cytoplasm	Cytoplasm	Single
211	P47141	Unknown	Mitochondrion	Mitochondrion	Single
212	P47150	Unknown	Mitochondrion	Mitochondrion	Single

213	P48524	Unknown	Cytoplasm	Cytoplasm	Single
214	P49166	Unknown	Cytoplasm	Cytoplasm	Single
215	P49167	Unknown	Cytoplasm	Cytoplasm	Single
216	P49591	Unknown	Cytoplasm	Cytoplasm	Single
217	P49626	Unknown	Cytoplasm	Cytoplasm	Single
218	P49631	Unknown	Cytoplasm	Cytoplasm	Single
219	P50109	Unknown	Cytoplasm	Cytoplasm	Single
220	P51401	Unknown	Cytoplasm	Cytoplasm	Single
221	P51402	Unknown	Cytoplasm	Cytoplasm	Single
222	P53030	Unknown	Cytoplasm	Cytoplasm	Single
223	P53080	Unknown	Cytoplasm	Cytoplasm	Single
224	P53088	Unknown	Mitochondrion	Mitochondrion	Single
225	P53124	Unknown	Cytoplasm	Cytoplasm	Single
226	P53188	Unknown	Nucleus	Nucleus (nucleolus)	Single
227	P53292	Unknown	Mitochondrion	Mitochondrion	Single
228	P53305	Unknown	Mitochondrion	Mitochondrion	Single
229	P53552	Unknown	Cytoplasm; Nucleus	Nucleus	Single
230	P53743	Unknown	Cytoplasm; Nucleus	Nucleus (nucleolus)	Single
231	P53890	Unknown	Cytoplasm	Cytoplasm (arrives at the bud site approximately coincident with bud emergence and dissociates from the septin scaffold before cytokinesis)	Single
232	P53908	Unknown	Chloroplast; Membrane	Membrane (multi-pass membrane protein)	Single
233	P53964	Unknown	Cytoplasm; Membrane	Membrane (single-pass membrane protein)	Single
234	P54005	Unknown	Cytoplasm	Cytoplasm	Single
235	P54867	Unknown	Membrane.	Cell membrane (single-pass type 1 membrane protein)	Single
236	P56628	Unknown	Cytoplasm	Cytoplasm	Single
237	P79263	Unknown	Secreted protein	Secreted protein	Single
238	P80110	Unknown	Secreted protein	Secreted protein	Single
239	P80111	Unknown	Secreted protein	Secreted protein	Single
240	P80344	Unknown	Secreted protein	Secreted protein	Single
241	P81061	Unknown	Secreted protein	Secreted protein	Single
242	P87054	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
243	P87133	Unknown	Mitochondrion	Mitochondrion	Single
244	P87262	Unknown	Cytoplasm	Cytoplasm	Single
245	P87299	Unknown	Cytoplasm	Cytoplasm	Single
246	P97278	Unknown	Secreted protein	Secreted protein	Single
247	P97279	Unknown	Secreted protein	Secreted protein	Single
248	P97280	Unknown	Secreted protein	Secreted protein	Single

249	P97430	Unknown	Secreted protein	Secreted protein	Single
250	P98119	Unknown	Secreted protein	Secreted protein	Single
251	P98121	Unknown	Secreted protein	Secreted protein	Single
252	Q00420	Unknown	Nucleus	Nucleus	Single
253	Q01163	Unknown	Mitochondrion	Mitochondrion	Single
254	Q01448	Unknown	Nucleus	Nucleus	Single
255	Q02326	Unknown	Chloroplast; Cytoplasm	Cytoplasm	Single
256	Q02753	Unknown	Cytoplasm	Cytoplasm	Single
257	Q03213	Unknown	Nucleus	Nucleus	Single
258	Q03337	Unknown	Golgi	Golgi apparatus (cis-Golgi network)	Single
259	Q03758	Unknown	Cytoplasm	Cytoplasm	Single
260	Q03784	Unknown	Golgi	Golgi apparatus (cis-Golgi network)	Single
261	Q04231	Unknown	Centriole; Cytoplasm	Nucleus	Single
262	Q04235	Unknown	Cytoplasm	Cytoplasm	Single
263	Q04264	Unknown	Nucleus	Nucleus	Single
264	Q04806	Unknown	Cytoplasm	Cytoplasm	Single
265	Q04949	Unknown	Cytoplasm	Cytoplasm (concentrates at motile dots in the cytoplasm corresponding to the plus ends of cytoplasmic microtubules)	Single
266	Q06033	Unknown	Secreted protein	Secreted protein	Single
267	Q06078	Unknown	Nucleus	Nucleus (nucleolus)	Single
268	Q06547	Unknown	Nucleus	Nucleus	Single
269	Q07092	Unknown	Secreted protein	Secreted protein (extracellular space; extracellular matrix)	Single
270	Q09094	Unknown	Centriole; Nucleus	Nucleus	Single
271	Q09792	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
272	Q09796	Unknown	Nucleus	Nucleus (nuclear rim)	Single
273	Q09815	Unknown	Cytoplasm	Cytoplasm (septum)	Single
274	Q09855	Unknown	Cytoplasm	Cytoplasm	Single
275	Q09868	Unknown	Cytoplasm	Cytoplasm	Single
276	Q09884	Unknown	Chloroplast; Cytoplasm	Cytoplasm; Nucleus	Multiple
277	Q09902	Unknown	Cytoplasm; Nucleus.	Cytoplasm; Nucleus	Multiple
278	Q10168	Unknown	Nucleus	Nucleus (nuclear pore complex; cytoplasmic side. Nucleus; nuclear pore complex; nucleoplasmic side)	Single
279	Q10180	Unknown	Cytoplasm	Cytoplasm (localizes to the barrier septum and cell tip)	Single
280	Q10223	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus (associated with vesicle-like and endoplasmic reticulum structures)	Multiple

281	Q10225	Unknown	Cytoplasm	Cytoplasm	Single
282	Q10253	Unknown	Cytoplasm	Cytoplasm	Single
283	Q10257	Unknown	Nucleus	Nucleus (nucleolus)	Single
284	Q10271	Unknown	Cytoplasm; Nucleus	Nucleus; Cytoplasm (localizes to a large number of foci in both the nucleus and cytoplasm)	Multiple
285	Q10274	Unknown	Cytoplasm; Nucleus	Nucleus	Single
286	Q10308	Unknown	Mitochondrion	Mitochondrion	Single
287	Q10326	Unknown	Cytoplasm	Cytoplasm (localizes to the barrier septum)	Single
288	Q10432	Unknown	Nucleus	Nucleus; Nucleoplasm	Single
289	Q10434	Unknown	Cytoplasm	Cytoplasm (localizes to cell tips during interphase)	Single
290	Q10447	Unknown	Cytoplasm	Cytoplasm (located at the cell tip)	Single
291	Q10474	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
292	Q12087	Unknown	Cytoplasm	Cytoplasm	Single
293	Q12213	Unknown	Chloroplast; Cytoplasm	Cytoplasm	Single
294	Q12215	Unknown	Membrane	Membrane (multipass membrane protein)	Single
294	Q12263	Unknown	Cytoplasm	Cytoplasm (bud neck)	Single
295	Q12690	Unknown	Cytoplasm	Cytoplasm	Single
296	Q14624	Unknown	Secreted protein	Secreted protein	Single
297	Q20347	Unknown	Cytoplasm	Cytoplasm	Single
298	Q22866	Unknown	Cytoplasm	Cytoplasm	Single
299	Q28065	Unknown	Secreted protein	Secreted protein	Single
300	Q28066	Unknown	Secreted protein	Secreted protein	Single
301	Q3E754	Unknown	Cytoplasm	Cytoplasm	Single
302	Q3E757	Unknown	Cytoplasm	Cytoplasm	Single
303	Q3E792	Unknown	Cytoplasm; Mitochondrion	Cytoplasm	Single
304	Q3E7X9	Unknown	Cytoplasm	Cytoplasm	Single
305	Q42577	Unknown	Mitochondrion	Mitochondrion	Single
306	Q43844	Unknown	Mitochondrion	Mitochondrion	Single
307	Q61702	Unknown	Secreted protein	Secreted protein	Single
308	Q61703	Unknown	Secreted protein	Secreted protein	Single
309	Q61704	Unknown	Secreted protein	Secreted protein	Single
310	Q62261	Unknown	Membrane	Cell membrane (peripheral membrane protein; cytoplasmic side)	Single
311	Q63514	Unknown	Secreted protein	Secreted protein	Single
312	Q63515	Unknown	Secreted protein	Secreted protein	Single
313	Q6NS38	Unknown	Cytoplasm; Nucleus	Nucleus (detected in replication foci during s-phase)	Single

314	Q86XK2	Unknown	Nucleus	Nucleus	Single
315	Q8TCJ0	Unknown	Cytoplasm; Nucleus	Nucleus	Single
316	Q96Q83	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
317	Q96RK4	Unknown	Centriole; Cytoplasm	Centrosome (localizes to the pericentriolar region throughout the cell cycle)	Single
318	Q9C0U3	Unknown	Mitochondrion	Mitochondrion	Single
319	Q9C0W0	Unknown	Nucleus	Nucleus	Single
320	Q9C104	Unknown	Cytoplasm	Cytoplasm	Single
321	Q9C110	Unknown	Cytoplasm	Cytoplasm	Single
322	Q9DB96	Unknown	Centriole; Cytoplasm; Nucleus	Nucleus; Cytoplasm (detected in axons, dendrites and filopodia)	Multiple
323	Q9H7D7	Unknown	Cytoplasm	Cytoplasm	Single
324	Q9NR20	Unknown	Cytoplasm	Cytoplasm	Single
325	Q9P7N0	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
326	Q9UPN7	Unknown	Cytoplasm	Cytoplasm	Single
327	Q9US49	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
328	Q9USR9	Unknown	Nucleus	Nucleus (nucleoplasm)	Single
329	Q9USV4	Unknown	Cytoplasm	Cytoplasm	Single
330	Q9UT31	Unknown	Mitochondrion	Mitochondrion	Single
331	Q9UTR7	Unknown	Cytoplasm	Cytoplasm	Single
332	Q9UU87	Unknown	Cytoplasm; Nucleus	Cytoplasm; Nucleus	Multiple
333	Q9Y7V0	Unknown	Endoplasmic reticulum	Endoplasmic reticulum (endoplasmic reticulum membrane; single-pass type 2 membrane protein)	Single
334	Q9ZNR6	Unknown	Cytoplasm	Cytoplasm	Single

Vol.1, No.2, 93-106 (2009)
doi:10.4236/ns.2009.12012

Natural Science

Sequence-Based Protein Crystallization Propensity Prediction for Structural Genomics: Review and Comparative Analysis

Lukasz Kurgan*, Marcin J. Mizianty

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada.
*University of Alberta, ECERF, 9107 116 Street, Edmonton, Alberta, Canada; lkurgan@ece.ualberta.ca

Received 6 August 2009; revised 28 August 2009; accepted 30 August 2009.

ABSTRACT

Structural genomics (SG) is an international effort that aims at solving three-dimensional shapes of important biological macro-molecules with primary focus on proteins. One of the main bottlenecks in SG is the ability to produce diffraction quality crystals for X-ray crystallography based protein structure determination. SG pipelines allow for certain flexibility in target selection which motivates development of in-silico methods for sequence-based prediction/assessment of the protein crystallization propensity. We overview existing SG databanks that are used to derive these predictive models and we discuss analytical results concerning protein sequence properties that were discovered to correlate with the ability to form crystals. We also contrast and empirically compare modern sequence-based predictors of crystallization propensity including OB-Score, ParCrys, XtalPred and CRYSTALP2. Our analysis shows that these methods provide useful and complimentary predictions. Although their average accuracy is similar at around 70%, we show that application of a simple majority-vote based ensemble improves accuracy to almost 74%. The best improvements are achieved by combining XtalPred with CRYSTALP2 while OB-Score and ParCrys methods overlap to a larger extend, although they still complement the other two predictors. We also demonstrate that 90% of the protein chains can be correctly predicted by at least one of these methods, which suggests that more accurate ensembles could be built in the future. We believe that current protein crystallization propensity predictors could provide useful input for the target selection procedures utilized by the SG centers.

Keywords: Structural Genomics; X-Ray Crystallography; Crystallization Propensity Prediction; Protein Structure; Protein Crystallization

1. INTRODUCTION

Proteins are organic compounds composed of amino acids arranged in a linear chain polymer with the help of peptide bonds. Proteins implement a wide variety of functions such as transportation, signalling, catalysis of chemical reactions, formation of the cell cytoskeleton, immune responses, regulation of cell processes, etc. etc. They are so versatile due to their ability to adopt an immense variety of shapes. Knowledge of the tertiary (three-dimensional) structure of proteins is vitally important for understanding and manipulating their biochemical and cellular functions. For instance, this knowledge is exploited in rational drug design via virtual screening [1-3], provides insights into various diseases [4] and it is used in deciphering interactions of proteins with other macro molecules and smaller ligands [5-7].

1.1. Structural Genomics

As of July 2009 we know close to 8.2 million nonredundant protein chains which can be found in SeqRef database [8] but the corresponding structure is know for "only" about 55 thousand proteins which are deposited into the Protein Data Bank (PDB) database [9]. This wide and continually widening sequence-structure gap calls for new and efficient efforts that would help in acquiring protein structures. This resulted in creation of structural genomics (SG) which is an international effort to find the three-dimensional shapes of important biological macro-molecules, primarily focusing on proteins [10]. In contrast to a traditional approach used by structural biologists who often work with a given protein that they try to solve for many years, the structural genomics efforts frequently concern "unknown" proteins. Moreover, SG focuses on development and usage of high

throughput and cost-effective methods for protein production and determination of the corresponding structure which are implemented with the help of dedicated SG centers. In the United States one of the first SG efforts, which was undertaken around year 2000, was the creation of a multi-center, including four large-scale centers and six specialized centers, Protein Structure Initiative. Similar SG projects were also carried out in Canada, Israel, Japan, and Europe. For example, Structural Genomics Consortium which was formed in 2004 spans centers at the Oxford University, University of Toronto and Karolinska Institute. Analysis shows that in 2004/2005 about half of protein structures were solved at a SG centers rather than in the traditional laboratory [11]. Also, at that time the cost of solving a structure at the most efficient SG center in the United States was equal to about 25% of the estimated cost when using the traditional methods [11]. Another more recent study shows that the production-line approach taken at the Protein Structure Initiative centers reduced the cost of solving structures from ~$250,000 apiece in 2000 to ~$66,000 in 2008 [12]. Most importantly, from our point of view, these SG initiatives shifted the focus from one-by-one determination of individual protein structures, which is being pursued by structural biologists, to protein family-directed structure analyses in which a group of proteins is targeted and structure(s) of representative members are determined and used to represent the entire group [13]. The corresponding process of choosing representative proteins is known as target selection and it encompasses a computational process of restricting candidate proteins to those that are tractable and of unknown structure and prioritizing them according to expected interest and accessibility [14]. In the case of the Protein Structure Initiative, the target selection concentrates on representatives from large, structurally uncharacterized protein domain families, and from structurally uncharacterized subfamilies in very large and diverse families with incomplete structural coverage [15]. We note that this approach allows for some flexibility in the selection of the targets.

1.2. X-ray Crystallography and Protein Crystallization

The protein structures are being determined with the help of experimental methods including X-ray crystallography [16], NMR spectroscopy [17], electron microscopy [18], and (more recently) by application of computational approaches such as homology modelling [19, 20]. The most popular method, which accounts for approximately 86% of the solved and deposited protein structures, is the X-ray crystallography; see **Figure 1**. At the same time, the other approaches play a strong complementary role for some protein types, such as membrane proteins [21, 22].

One of the main challenges the SG initiative faces it that only about 2-10% of protein targets pursued in the context of the second step of the Protein Structure Initiative yield high-resolution protein structures [23]. We further investigated these estimates based on data published in the TargetDB database [24] in July 2009. TargetDB is a world-wide database that provides information on the experimental progress and status of targets

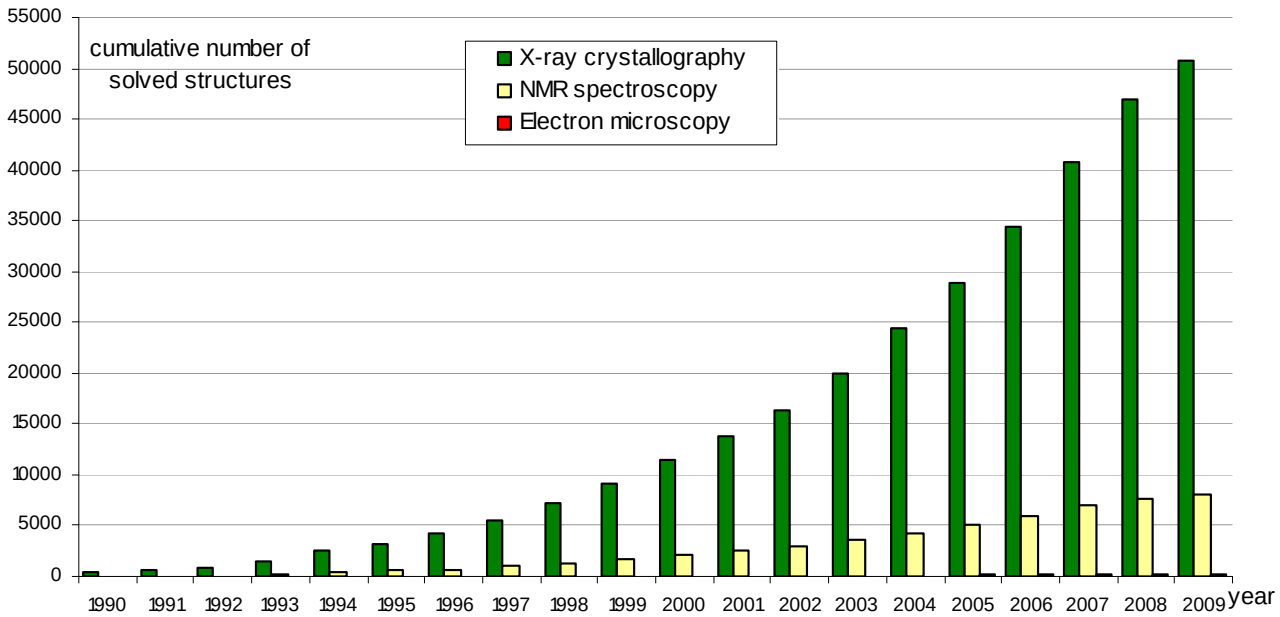


Figure 1. The growth in the number of protein structures deposited into PDB by that were solved by X-ray crystallography, NMR spectroscopy and electron microscopy (source http://www.rcsb.org/).

selected for structure determination. Among 150,727 cloned targets that were deposited into TargetDB, only 37,398 (24.8%) were reported to be successfully purified, 12,923 (8.6%) to be successfully crystallized, and 6,942 (4.6%) gave diffraction quality crystals. Moreover, some estimates show that more than 60% of the cost of structure determination is consumed by the failed attempts [25] while crystallization is characterized by a significant rate of attrition and is among the most complex and least understood problems in structural biology [26]. The above provides a strong motivation for further research and development in this area. Several strategies have been proposed to improve the success rate including obtaining one representative structure per protein family and working with multiple orthologues [14, 26, 27, 28]. In spite of advances made in the context of protein crystallization [29], the above numbers and insights from some researchers [30-32] demonstrate that the production of high-quality crystals is one of the major bottlenecks in the protein structure determination. The crystals should be sufficiently large (> 50 micrometres), pure in composition, regular in structure and with no significant internal imperfections. The problem of production of diffraction-quality crystals is usually tackled using an empirical approach based mainly on trial and error (also called the "art" of crystallisation), in which a large number of experiments is brute-forced to find a suitable setup, and through understanding of the fundamental principles that govern crystallisation [30]. The latter is used to design new (and improved) experimental methodologies that would produce high-quality crystals.

1.3. Databases

One of the early steps taken to alleviate the abovementioned difficulties in resolving the structures via X-ray crystallography was to create databanks that record information concerning both successful and failed attempts to produce the structures. The importance of these efforts was advocated in 2000 by Raymond Stevens who said that "industrial-scale efforts will lead to the generation of knowledge bases that will be mined to expand our understanding of the techniques used in protein crystallography. These efforts will act as 'learning factories', in which successes and failures will be used to continually improve the technology for high-throughput protein crystallography" [33].

These words were echoed in 2003 by Rodrigues and Hubbard who said "as structural genomics projects evolve, valuable experimental data will be accumulated, thus presenting researchers with a unique opportunity to establish improved predictive methods for a protein's chemical and physical behaviour based on its amino acid sequence. It is essential for laboratories producing such data to keep track of both 'successful' and 'unsuccessful' results, so that these can be fed back into the structural determination pipeline through the improvement of the target selection procedures" [34]. The development of the databases was fuelled by generation of large and well annotated experiments by SG centers, such as one for the *Thermotoga maritime* proteome [35]. To the best of our knowledge, the first such initiative was the PRESAGE database which included annotations indicating current experimental status, structural predictions and suggestions [36]. Some of the SG consortia have established on-line progress reports which contain details and current experimental status of their targets. Examples include Integrated Consortium Experimental Database [37], ZebaView (http://www-nmr.cabm.rutgers.edu/bioinformatics/ZebaView/), ReportDB (http://www.secsg.org/cgi-bin/report.pl) and SPINE (Structural Proteomics in the NorthEast) [38, 39]. SPINE, which was developed in early 2000 and reengineered in 2003, integrates a tracking database and a data mining method for identifying feasible targets. Each protein deposited in this database is described with information related to the experimental progress (e.g., expression level, solubility, ability to crystallize) and 42 descriptors of the underlying protein sequence (amino acid composition, secondary structure, etc.). The largest and most comprehensive TargetDB [24] was launched July 2001 and it builds upon the work on the PRESAGE database. TargetDB serves as a primary target registration database for structural SG project worldwide. It consolidates data from 28 SG centers in USA, Canada, Germany, Isreal, Japan, France and UK, including 9 Protein Structure Initiative centers. PepcDB (Protein Expression Purification and Crystallization DataBase), which was created around 2004, was established as an extension to TargetDB to collect more detailed status information and the experimental details of each step in the protein structure production pipeline [40]. This database stores a complete history of the experimental steps in each production trial besides describing the current target production status. PepcDB records status history, stop conditions, reusable text protocols and contact information collected from 15 SG centers in USA. The interested readers are directed to two recent articles by Helen Berman that introduce a wealth of resources concerning the SG initiative [41] and a knowledgebase developed by the Protein Structure Initiative [42].

1.4. Computational Models in Protein Crystallization

The problems with the protein crystallization and the availability of the suitable databases motivated the development of analytical and predictive models that can be used to either support or directly predict protein crystallization [43]. These models were often developed by researchers at certain SG centers who used their own data to draw conclusions. In one of the first attempts, a

decision tree that predicts solubility from protein sequence was developed [44]. The SPINE system, which was developed at the Northeast Structural Genomics Consortium, incorporates decision tree-based classifiers for solubility and crystallization propensity. This system was used to extract a few interesting rules such that soluble proteins tend to have more acidic residues and fewer hydrophobic segments [38]. The SG project on *Plasmodium falciparum* has lead to an analysis of protein characteristics, such as the presence of transmembrane helices, low-complexity regions, and coiled-coil regions, in the context of the crystallization propensity [34]. Another decision tree-based predictive model developed by Goh and colleagues in 2004 using data from TargetDB has revealed several protein features that influence the feasibility of using a given target protein chain for a high-throughput structure determination [45]. They include conservation of the sequence across organisms, composition of charged residues, occurrence of hydrophobic patches in the sequence, number of binding partners, and chain length. Based on the data from the *Thermotoga maritime* proteome [35], the researchers at the Joint Center for Structural Genomics discovered a few features, which include isoelectric point, sequence length, average hydropathy, existence of low complexity regions, presence of signal peptides and trans-membrane helices, that correlate with crystallization [46]. The isoelectric point calculated from the protein sequence was also used to develop a method that suggests optimal pH ranges for crystallization screening [47, 48]. Experimental work by Derewenda's group shows that crystallization can be improved by application of surface entropy reduction approach in which clusters of two or three exposed amino acids with high conformational entropy side chains (such as Lys, Glu and Gln) are replaced with lower-entropy residues (like Ala) [49-54]. One drawback of this method is that it may decrease protein solubility which hinders crystallization screening [50, 52]. The surface entropy reduction approach was recently implemented as a web server [55]. This server utilizes information concerning conformational entropy and solvent exposure indices, predicted secondary structure, residues conservation scores, and close homologues to propose crystallization enhancing mutations for a given protein sequence. Another study, which was conducted at the Center for Eukaryotic Structural Genomics, used disorder prediction algorithms to analyze the impact of intrinsic protein disorder on crystallization efficiency [56]. The Berkeley Structural Genomics Center has utilized several protein features including length of the sequence and predicted transmembrane helices, coiled coils, and low-complexity regions to eliminate targets predicted to be intractable for the high-throughput structure determination [57]. The most recent study that was performed at the Northeast Structural Genomics Consortium shows that crystallization propensity depends primarily on the prevalence of well-ordered surface epitopes [58]. More specifically, the authors show that crystallization propensity can be computed from the knowledge of predicted disordered regions, side-chain entropy of predicted exposed residues, the amount of predicted buried Gly and the fraction of Phe in the input sequence.

2. SEQUENCE-BASED METHODS FOR PREDICTION OF PROTIEN CRYSTALLIZATIONPROPENSITY

The SG efforts allow for certain flexibility in selection of the chains for the crystallization and the subsequent structure determination and this motivates development of methods that aim at the prediction/assessment of the crystallization propensity for a given input sequence. Such methods could be incorporated into target selection pipelines that are utilized by SG centers. Their development is often supported and motivated by the described above computational analyses/models. We also note that numerous studies have already demonstrated that sequence-based prediction approaches, which may address a variety of structural and functional properties of proteins, provide useful information and insights for both basic research and drug design and hence are widely welcome by the scientific community [59-63].

Crystallization propensity prediction methods incorporate predictive models that are extracted from larger datasets that span data coming from multiple SG centers and they take the protein sequence as their only input. The underlying principle is that the predictive models summarize/describe patterns (similarities) hidden in the data from databases such as TargetDB. This is done by generating a set of patterns that describe sequences that can be crystallized (crystallizable proteins) and another set of patterns for sequences that were shown to be impossible to crystallize (noncrystallizable proteins). The two sets of patterns should describe the two corresponding sets of protein chain and, at the same time, each of them should exclude sequences from the other set. The existing crystallization propensity predictors include SECRET [64] that was developed by Frishman's group, OB-Score [65] and ParCrys [66] that were produced by the Barton's group, XtalPred [67, 68] that came from Godzik's group, and CRYSTALP [69] and most recent CRYSTALP2 [70] that were developed by Kurgan's group. These methods perform the prediction in two steps: (1) the input sequence is converted into a set of numerical features that describe certain characteristics of the sequence; and (2) the feature values are fed into a predictive model that outputs the outcome that quantifies propensity for crystallization. The predictive model encapsulates the patterns that are computed from the information encoded by the features. **Table 1** shows a

Table 1. A side-by-side comparison of existing methods for sequence-based protein crystallization propensity prediction.

Methods [reference]	Source of data	Input features		Predictive model	Web server/page	Notes
		description	#			
SECRET [64]	Deposition from PDB assuming that NMR only solved protein are difficult/impossible to crystallize; Depositions in TargetDB	Content of mono-, di-, and tripeptides represented by 20-letter amino acid alphabet and by several reduced alphabets grouped by physicochemical and structural properties of amino acids	103	Two-layered structure where output of several support vector machine classifiers are combined by a second-level Naive Bayes classifier	http://mips.helmholtz-muenchen.de/secret/	Limited to sequences between 46 and 200 amino acids
OB-Score [65]	Depositions in TargetDB	Isoelectric point and average hydrophobicity	2	Z-score (two-dimensional lookup-table)	http://www.compbio.dundee.ac.uk/xtal/	
CRYSTALP [69]	Deposition from PDB assuming that NMR only solved protein are difficult/impossible to crystallize;	Content of selected mono-, di- and collocated dipeptides	46	Naive Bayes	N/A	Limited to sequences between 46 and 200 amino acids
XtalPred [67, 68]	Depositions in TargetDB	Protein length, molecular mass, gravy and instability indices, extinction coefficient, isoelectric point, content of Cys, Met, Trp, Tyr, and Phe residues, insertions in the alignment compared to homologs in non-redundant protein sequences database, predicted secondary structure, predicted disordered, low-complexity and coiled-coil regions, predicted transmembrane helices and signal peptides.	9	Normalized product	http://ffas.burnham.org/XtalPred/	Outputs 1 of 5 crystallization classes: optimal, suboptimal, average, difficult, and very difficult
ParCrys [66]	Depositions in TargetDB and PepcDB	Isoelectric point and average hydrophobicity, content of Ser, Cys, Gly, Phe, Tyr, and Met residues	8	Kernel-based classifier using Parzen window	http://www.compbio.dundee.ac.uk/xtal/	
CRYSTALP2 [70]	Depositions in TargetDB and PepcDB	Isoelectric point, average hydrophobicity, content of selected mono-, di- and collocated di- and tripeptides	88	Normalized Gaussian radial basis function network	http://biomine.ece.ualberta.ca/CRYSTALP2/CRYSTALP2.html	

side-by-side comparison of the six existing methods based on the data source that was used to generate predictive model and the applied input features and predictive models. It also provides URLs of the corresponding web servers or web pages.

Two early methods, namely SECRET and CRYSTALP, accept only sequences between 46 and 200 amino acids in length. This limitation is due to the composition of datasets used to generate these prediction models. Although OB-Score predictor does not impose a limit on sequence size, it considers only two predictive features, i.e., isoelectric point and hydrophobicity. This method was developed for the Scottish Structural Proteomics Facility [65]. The ParCrys method extends OB-score by using an advanced kernel-based classification algorithm and by adding information concerning content of several amino acids including Ser, Cys, Gly, Phe, Tyr, and Met to the set of predictive features. Similarly, CRYSTALP2 improves upon CRYSTALP by applying a more advanced kernel-based classifier and by introducing new predictive features that are based on the collocation of amino acids in the sequence, isoelectric point and hydrophobicity. The motivation for the application of the collocation based features comes from their applications in related fields [71-74] and the fact that they consider local neighbourhood information in the protein chain, which was also utilized in a recent method for surface entropy reduction based design of crystallizable protein variants [55]. A significant majority of the collocations used by CRYSTALP2 incorporate residues with high conformational entropy, or with low entropy and high potential to mediate crystal contacts, and these residues are utilized by the surface entropy reduction methods [51, 52].

The above five methods are built using black-box (not readable by a human) classification models, which are inductively learned from a set of protein chains which are annotated as crystallizable and noncrystallizable. By contrast, the XtalPred is a white-box (human readable) approach that combines probabilities of successful crystallization calculated from several protein features. This method, which was developed based on experiences at the Joint Center for Structural Genomics, which is one of the large centers in the Protein Structure Initiative, mimics the work performed by structural biologists. XtalPred utilizes nine biochemical and biophysical features of an input protein with probability distributions estimated from data from TargetDB. The individual probabilities concerning each input feature are combined into a single crystallization score which is used to assign one of five crystallization classes: optimal, suboptimal, average, difficult, and very difficult. The design of XtalPred shows that medium sequence length and hydrophobicity combined with acidic character improve the success in protein production. It also demonstrates that very short, very long, or very hydrophobic proteins are more difficult to crystallize under standard experimental setups. This method also confirms the utility of predicted structural disorder, presence of transmembrane helices, instability, and high content of predicted loops, insertions, and coiled-coil structures for the prediction of the crystallization propensity [67]. Several methods, including XtalPred, OB-Score, ParCrys and CRYSTALP2, utilize information concerning isoelectric point which is estimated from protein sequence. This agrees with prior finding that indicate important role of this feature [46-48].

We note that all investigated crystallization propensity predictors take into account only intra-molecular factors that are encoded in the protein chain. This means that they may not provide reliable predictions when inter-molecular factors such as protein-protein and/or protein-precipitant interactions, buffer composition, precipitant diffusion method, etc. must be considered. Also, they are limited to predictions for non-redundant chains and should not be used when assessing crystallization of homologues. In the latter case we recommend the use of the surface entropy reduction server [55].

3. COMPARATIVE ANALYSIS

Following we perform empirical comparison of the quality of predictions offered by the sequence-based protein crystallization propensity predictors. Our analysis excludes CRYSTALP and SECRET methods since they are limited to only relatively small chains and since their quality was show to be inferior when compared with other methods [66,70]. Our comparative analysis is performed based on predictions performed for a dataset of relatively recent depositions to TargetDB and PepcDB. We analyze predictive power of individual methods and we also investigate their complementarity.

3.1. Dataset

We use a dataset composed of 2000 protein chains (hereafter TEST-NEW), which was originally introduced in [70] and which was developed using procedure proposed in [66]. The crystallizable proteins were extracted from sequences deposited in TargetDB and they include the last 1000 depositions as of December 2008. The noncrystallizable sequences, which correspond to the actual construct sequences used, were extracted from the trial sequences stored in PepcDB. As in the case of crystallizable chains, they include the last 1000 depositions as of December, 2008. The selected sequences were also processed to remove the N-terminal hexaHis tag and LEHHHHHH tag at the C-terminus, which are introduced to ease the purification. Duplicate sequences were removed and thus the resulting dataset consists of non-redundant chains. It can be freely downloaded from

Table 2. Summary of results for predictions performed with OB-Score, ParCrys, XtalPred and CRYSTALP2 methods on the TEST-NEW dataset.

	Accuracy	MCC	TPR	TNR	AROC
OB-Score[1]	69.8	0.42	0.86	0.54	0.74
ParCrys[1]	70.6	0.43	0.83	0.58	0.75
XtalPred[2]	70.0	0.40	0.76	0.64	0.76
CRYSTALP2[3]	69.3	0.39	0.76	0.63	0.74

[1]Results computed using the ParCrys/OB-Score server at http://www.compbio.dundee.ac.uk/xtal/
[2]Results computed using the XtalPred server at http://ffas.burnham.org/XtalPred/
[3]Results based on [70]

http://biomine.ece.ualberta.ca/CRYSTALP2/CRYSTALP2.html.

3.2. Quality Measures

The annotations from TargetDB were stripped from the input sequences, which in turn were inputted into the corresponding predictors. The prediction outputs were compared with the original annotations to assess the prediction quality. Four potential prediction outcomes are possible: TP (true positive) which corresponds to crystallizable chains that were correctly predicted as crystallizable, FN (false negative) which corresponds to crystallizable chains that were incorrectly predicted as noncrystallizable, FP (false positive) which indicates that noncrystallizable chains were incorrectly predicted as crystallizable, and TN (true negative) which denotes cases where noncrystallizable chains were correctly predicted as noncrystallizable. The predictions were assessed based on the following quality indices:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

The accuracy measures the fraction of correct predictions among all predictions. The Matthews Correlation Coefficient (MCC) is confined to <-1,1> interval. If the MCC value is close to 0 then the prediction method is not better than a random classification. Higher MCC value corresponds to better performance of the prediction method. These two measures provide an evaluation of the prediction quality over the entire dataset. In contrast, TPR (true positive rate) and TNR (true negative rate) evaluate the quality separately for crystallizable (positive) and noncrystallizable (negative) proteins. TPR/TNR quantifies the fraction of correctly predicted crystallizable/noncrystallizable proteins. We also report receiver-operator characteristics (ROC) curves that present a graphical plot of the TP rate = TP / (TP + FN) against FP rate = FP / (FP + TN). This is performed by thresholding the confidence values (probabilities) that are generated together with the predicted classes (crystallizable vs. noncrystallizable). These plots are also used to compute the area under the ROC curve (AROC). The higher the AROC value is the better the predictive power of the corresponding method.

3.3. Comparison of Existing Prediction Methods

Results of application of the four crystallization propensity predictors on the TEST-NEW dataset are summarized in **Table 2**. In the case of XtalPred we assume a prediction assignment in which optimal, suboptimal, and average outcomes are categorized as crystallizable proteins and difficult and very difficult as noncrystallizable. The same assignment was used in [70] since it leads to optimal results.

The comparison shows that the four methods are characterized by relatively similar prediction quality with MCC and accuracy values ranging between 0.39 and 0.43 and between 69.3 and 70.6%, respectively. We note that since the dataset is balanced a random assignment of the prediction outcomes would give accuracy of 50%. This means that the accuracy of the existing methods is better by about 20% than the random coin-toss approach. At the same time we observe a considerable space for improvement although we caution the reader that the upper limit of the prediction accuracy should not be assumed at 100%. This is since the input data likely includes mislabeled proteins. In particular, since data comes from multiple SG centers, some proteins that could not be crystallized in one center could be potentially crystallized by another center that uses different protocols and equipment. This means that some of the proteins could be mislabeled as noncrystallizable, i.e., some of the FPs are in fact TPs. At this time we are not able to estimate their number. We observe that OB-Score and ParCrys are both strongly biased towards prediction of crystallizable proteins, i.e., their TPR values are much higher than TNR values and the TNR values are relatively low. The XtalPred and CRYSTALP2 provide a

more balanced prediction for the two classes of proteins and their TNR values are above 0.63. All four methods provide better predictions for crystallizable proteins, i.e., they correctly predict a bigger fraction of crystallizable proteins, when compared with the noncrystallizable proteins. In other words, they are more likely to succeed in confirming that a crystallizable chain can be crystallized rather than in showing that a chain difficult to crystallize cannot be crystallized; although in both cases all of the considered methods work better than the coin-toss. **Figure 2** shows the ROC curves for the four predictors. We again observe that all considered methods behave similarly, i.e., they provide comparable TP rates for the same FP rates.

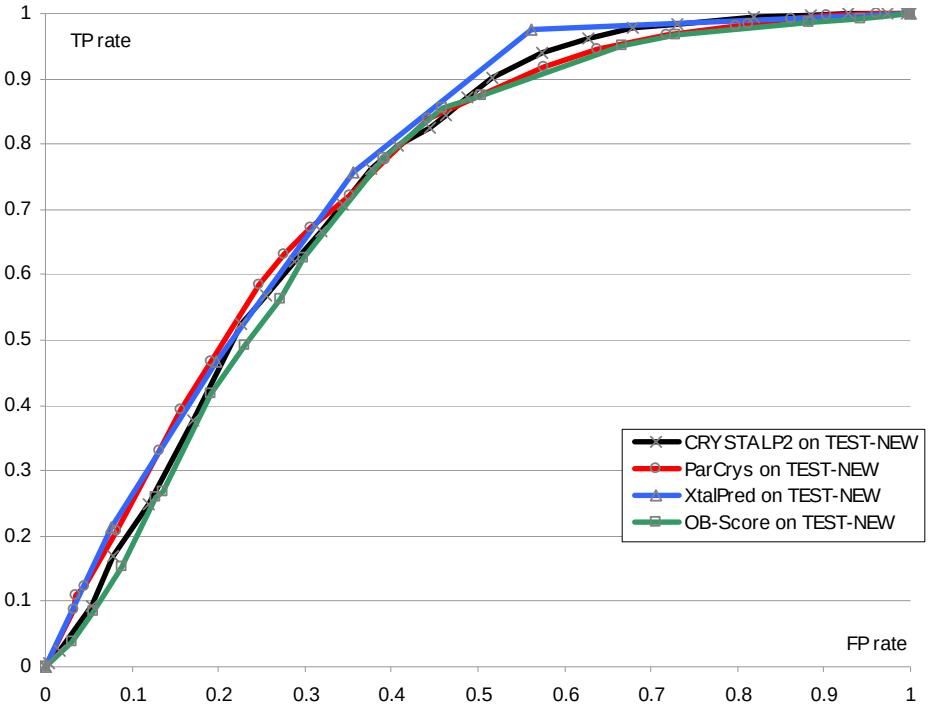


Figure 2. ROC curves for the tests performed with OB-Score, ParCrys, XtalPred and CRYSTALP2 methods on the TEST-NEW dataset.

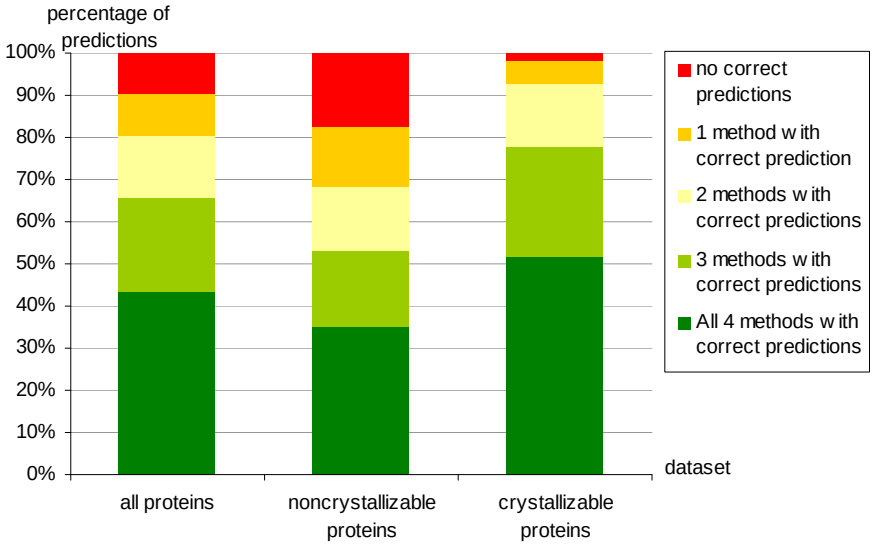


Figure 3. Analysis of the number of correct predictions produced by OB-Score, ParCrys, XtalPred and CRYSTALP2 methods on the all proteins, only crystallizable and only noncrystallizable proteins from the TEST-NEW dataset.

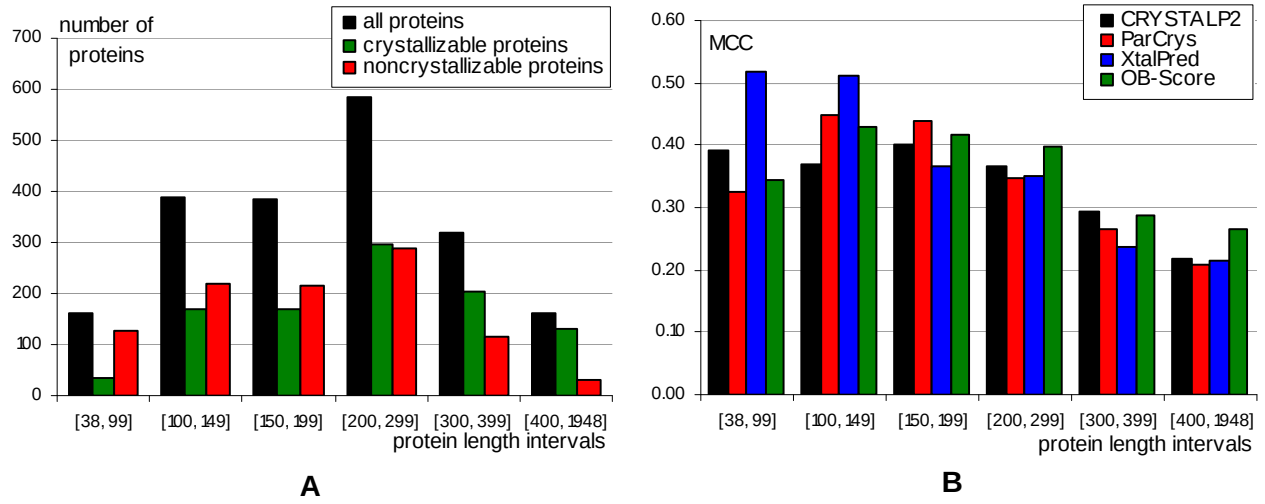


Figure 4. Analysis the predictions and characteristics of the TEST-NEW dataset with respect to the input protein chain length. A) Distribution of number of proteins (black bars), number of crystallizable (green bars) and noncrystallizable (red bars) proteins in the considered protein length intervals. B) Prediction quality measured using MCC for OB-Score, ParCrys, XtalPred and CRYSTALP2 methods for each of the protein size intervals.

Figure 3 analyzes the predictions with respect to the number of correct predictions produced by the four methods for each input protein. Analysis of the results obtained on the entire TEST-NEW set indicates that at least three methods provide correct predictions simultaneously for two thirds of the test proteins. It also shows that only 9.6% of the proteins cannot be correctly predicted by any of the considered methods. We again observe that predictions for crystallizable proteins are characterized by higher quality than for the noncrystallizable proteins. In particular, only 1.6% of crystallizable proteins are never correctly predicted and 78.0% are correctly predicted by at least 3 methods. In contrast, the same numbers for the noncrystallizable proteins are 17.6% and 53.2%, respectively.

3.4. Analysis of Predictions for Varying Protein Sizes

The protein chain length was indicated as one of the important factors related to the protein crystallization propensity [45, 46, 57, 67]. It is also correlated with the quality of the secondary structure prediction [75], which is utilized in the prediction of protein crystallization [55, 67]. To this end, **Figure 4** summarizes results that are organized by binning the input protein chains into six size-based intervals. **Figure 4A** shows, as expected [67], uneven distribution of the crystallizable and noncrystallizable proteins against the protein chain length. We observe that majority of short chains with less than 100 amino acids are difficult to crystallize while the crystallization is more successful for longer chains. More importantly, the XtalPred method stands out from the competition as it provides better performing predictions for short sequences of up to 150 amino acids. On the other hand, a slight improvement over the competition is observed for the OB-Score method when predicting long chains with above 400 amino acids. Finally, the CRYSTALP2 method is characterized by the most even quality. We also observe a generic trend that best results are on average obtained for the average sized protein chains between 100 and 200 amino acids.

3.5. Complementarity of Existing Methods

Although the above results indicate that the existing methods are characterized by comparable prediction quality, substantial differences in their underlying design and results shown in **Figures 3** and 4B suggest that their results could be complementary with each other. In other words, although on average they provide the same number of correct predictions, these prediction likely concern different input proteins.

We investigate the complementarity by combining multiple methods using OR operator, i.e., a given prediction is assumed correct if at least one of the methods in an ensemble provides a correct prediction. This approach allows quantifying the amount of overlap in predictions and it also estimates the upper boundary of a potential meta-predictor that combines predictions from the individual methods. **Figure 5** shows summary of results, in terms of achieved TPR, TNR and MCC values for all combinations of two, three, and four predictors as well as for the individual methods. We observe that certain ensembles obtain higher quality of predictions indicating a stronger complementarity. In particular combining either OB-Score and XtalPred or CRYSTALP2 with XtalPred gives better results than any other combination

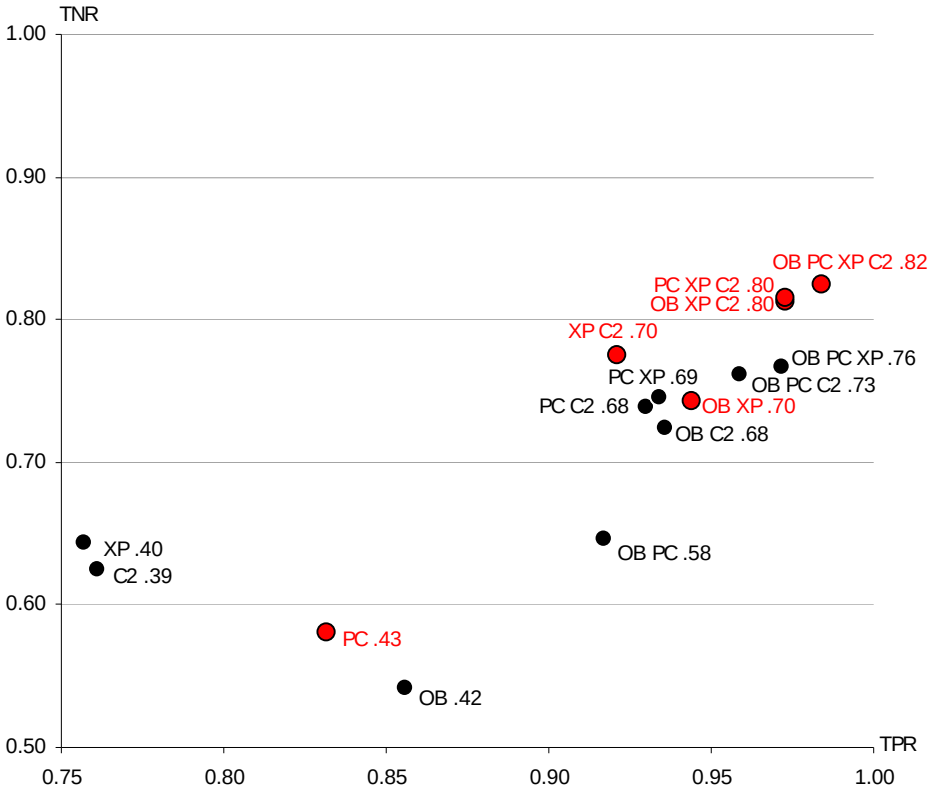


Figure 5. Analysis the complementarity of predictions for OB-Score (OB), ParCrys (PC), XtalPred (XP) and CRYSTALP2 (C2) methods on the TEST-NEW dataset. Each combination of 1, 2, 3, and 4 methods was applied using OR operator, i.e., a given prediction was assumed correct if at least one of the predictors predicted it correctly. The x-axis/y-axis shows TPR/TNR values (TPR values are scaled between 0.75 and 1 while TNR values are scaled between 0.5 and 1), and the labels next to markers denote a particular combination of applied predictions together with the MCC value (e.g., "PC XP C2 .80" means that combination of ParCrys, XtalPRed and CRYSTALP2 obtained MCC of 0.8). Markers and labels in red denote the best results for a given number of applied methods.

of two methods. Among the ensembles of three methods, the combination of XtalPred and CRYSTALP2 with either ParCrys or OB-Score works best. This observation and the fact that OB-Score and ParCrys are the least complimentary among all pairs of predictors indicate that these two methods provide relatively overlapping outputs. Finally, an ensemble of all four methods obtains MCC of 0.82 which is not much higher than 0.80 achieved with just three methods, showing that addition of the fourth predictor brings relatively minor improvements. Finally, we again observe that results indicate that both individual and ensemble-based predictions are characterized by higher quality for crystallizable rather than noncrystallizable proteins.

We also investigate a possibility of implementing a simple, majority-vote based meta-predictor. Such method generates predictions which correspond to the most frequent prediction of its member methods. We apply a simple majority vote for the three members based meta-predictors, while for ensemble of four methods we resolve the tie-break (2 vs 2 split decisions from the member methods) by applying the prediction of one selected method. This leads to eight potential configurations, i.e., three combinations of three out of four methods and four configurations with four member methods each time using a different method as a tie-breaker. The corresponding results are presented in **Figure 6**. The results demonstrate that the best ensemble includes XtalPred, CRYSTALP2 and OB-Score. The runner-up configurations include an ensemble of XtalPred, CRYSTALP2 and ParCrys and two ensembles of four methods with tie-breakers as XtalPred and CRYSTALP2. These results are consistent with the above complementarity analysis and indicate beneficial overlap between XtalPred and CRYSTALP2. We also observe that application of a majority-vote mechanism provides only moderate improvements. More specifically, the best vote-based ensemble obtains MCC of 0.49 while the MCC of best individual method equals 0.43 and the MCC of best combination of methods from **Figure 5** gives MCC

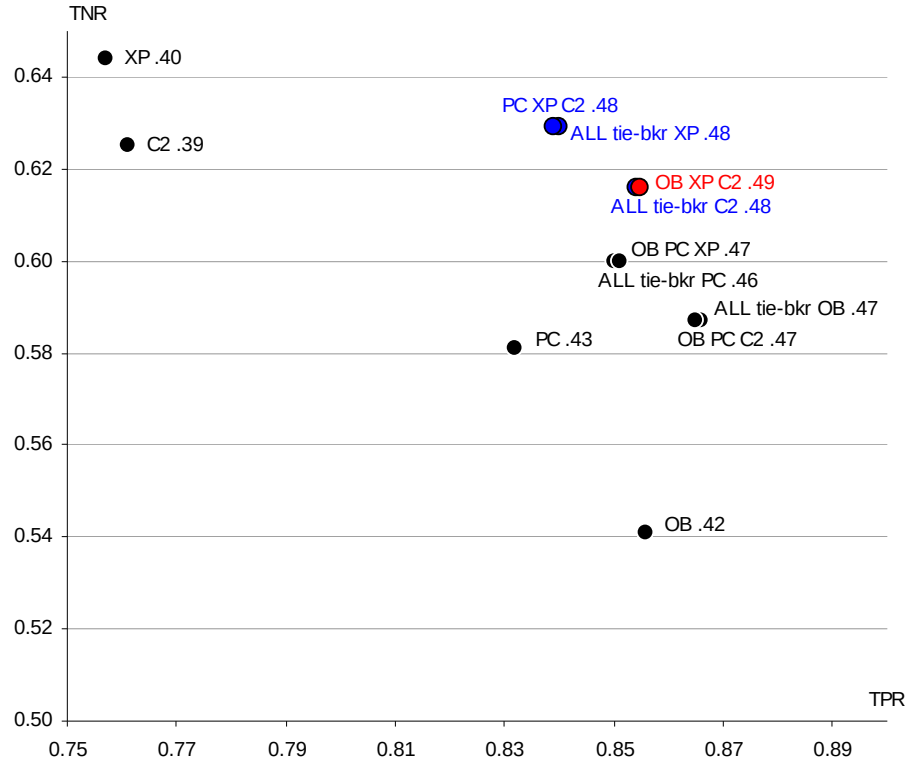


Figure 6. Analysis the performance of majority-vote based ensembles of OB-Score (OB), ParCrys (PC), XtalPred (XP) and CRYSTALP2 (C2) methods on the TEST-NEW dataset. The x-axis/y-axis shows TPR/TNR values (TPR values are scaled between 0.75 and 0.9 while TNR values are scaled between 0.5 and 0.65), and the labels next to markers denote a particular ensemble together with the MCC value (e.g., "OB XP C2 .49" means that ensemble composed of OB-Score, XtalPRed and CRYSTALP2 obtained MCC of 0.49). The prediction of the ensemble corresponds to the most frequent prediction of its members. The tie-breaker for ensembles of 4 methods is chosen as the prediction of one specific method, i.e., "ALL tie-brk XP" corresponds to an ensemble of all four methods in which a split 2 vs 2 decision is decided by the prediction of XtalPred. Markers and labels in red/blue denote the best/second best results.

equal to 0.82. In terms of the corresponding accuracies, this means that although the considered four methods can correctly predict up to 90.4% of proteins, the simple voting provides only 73.6% of correct predictions.

Overall, the analysis shows that the best improvements, when compared with using individual predictors, are achieved by combining XtalPred with CRYSTALP2. The OB-Score and PareCrys methods overlap to a larger extend although they also complement the other two predictors. This can be explained by the use of very similar input features in ParCrys and OB-Score and use of larger numbers of more complementary features in CRYSTALP2 and XtalPred. Finally, a simple voting based meta-predictor is shown to provide some improvements although more complex designs should be considered to better exploit complementarity between the existing prediction methods. Such advanced heterogeneous (using diverse types of member methods) meta-predictors were already successfully used in sequence-based prediction of other protein properties such as fold type [76, 77], subcellular localization [78-80], structural class [81], and solvent accessibility [82].

4. SUMMARY AND CONCLUSIONS

Structural genomics efforts have entered a mature stage when a wealth of data that could be analyzed to build useful supporting tools has been already accumulated. One of most significant bottlenecks in the protein structure determination pipelines implemented by SG centers is the ability to generate diffraction quality crystals. Although some mechanisms were already implemented to improve the corresponding success rates, our analysis shows a significant room for further improvements. In this context we have overviewed existing databases, analytical results and predictive methods that aim at supporting the protein crystallization task.

We show that analysis of data from certain SG centers and community-wide databases such as TargetBD re-

vealed that certain factors, such as protein size, isoelectric point, disorder regions, presence of transmembrane helices, etc. were found to correlate with the ability to produce quality protein crystals. We also contrasted and compared several modern sequence-based predictors of crystallization propensity including OB-Score, ParCrys, XtalPred and CRYSTALP2. We demonstrate that these methods provide useful predictions which are complementary to each other. Although their average success rate is similar and at about 70%, we show that usage of a simple majority-vote based combination of these methods can improve the success rate to almost 74%. Our work also reveals that close to 90% of the protein chains can be correctly predicted by at least one of these methods, which motivates development of more advanced meta-predictors. The best predictions for short, under 100 amino acids, chains are produced by XtalPred and the most accurate predictions, on average, are generated for medium-sized chains of 100 to 200 amino acids. We believe that these crystallization propensity predictors could provide useful input for current SG efforts that could be incorporated into the target selection procedure.

REFERENCES

[1] Guido, R.V., Oliva, G. and Andricopulo, A.D. (2008) Virtual screening and its integration with modern drug design technologies. *Current Medicinal Chemistry*, **15(1)**, 37-46.

[2] Norin, M. and Sundström, M. (2001) Protein models in drug discovery. *Current Opinion in Drug Discovery & Development*, **4,** 284-290.

[3] Klebe, G. (2000) Recent developments in structure-based drug design. *Journal of Molecular Medicine*, **78(5),** 269-281.

[4] Fernàndez-Busquets, X., de Groot, N.S., Fernandez, D. and Ventura, S. (2008) Recent structural and computational insights into conformational diseases. *Current Medicinal Chemistry*, **15,** 1336-1349.

[5] Luscombe, N.M., Laskowski, R.A. and Thornton, J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research*, **29,** 2860-2874.

[6] Ellis, J.J., Broom, M. and Jones, S. (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins*, **66,** 903-911.

[7] Chen, K. and Kurgan, L. (2009) Investigation of atomic level patterns in protein - small ligand interactions. *PLoS ONE*, **4(2),** e4473.

[8] Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **35(Database issue),** D61-65.

[9] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28,** 235-242.

[10] Brenner, S.E. (2001) A tour of structural genomics. *Nature Reviews Genetics*, **2(10),** 801-809.

[11] Chandonia, J.M. and Brenner, S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311,** 347-351.

[12] Service, R.F. (2008) Protein Structure Initiative: Phase 3 or Phase Out. *Science*, **319(5870),** 1610-1613.

[13] Terwilliger, T.C., Waldo, G., Peat, T.S., Newman, J.M., Chu, K. and Berendzen, J. (1998) Class-directed structure determination: Foundation for a protein structure initiative. *Protein Science*, **7(9),** 1851-1856.

[14] Brenner, S.E. (2000) Target selection for structural genomics. *Nature Structural Biology*, **7,** 967-969.

[15] Dessailly, B.H., Nair, R., Jaroszewski, L., Fajardo, J.E., Kouranov, A., Lee, D., Fiser, A., Godzik, A., Rost, B. and Orengo, C. (2009) PSI-2: structural genomics to cover protein domain family space. *Structure*, **17(6)**, 869-881.

[16] Ilari, A. and Savino, C. (2008) Protein structure determination by x-ray crystallography. *Methods in Molecular Biology*, **452,** 63-87.

[17] Wishart, D. (2005) NMR spectroscopy and protein structure determination: applications to drug discovery and development. *Current Pharmaceutical Biotechnology*, **6(2),** 105-120.

[18] Hite, R.K., Raunser, S. and Walz, T. (2007) Revival of electron crystallography. *Current Opinion in Structural Biology*, **17(4),** 389-395.

[19] Fischer, D. (2006) Servers for protein structure prediction. *Current Opinion in Structural Biology*, **16(2),** 178-182.

[20] Xiang, Z. (2006) Advances in homology protein structure modeling. *Current Protein & Peptide Science*, **7(3),** 217-227.

[21] Lacapère, J.J., Pebay-Peyroula, E., Neumann, J.M. and Etchebest, C. (2007) Determining membrane protein structures: still a challenge! *Trends in Biochemical Sciences*, **32(6),** 259-270.

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

[23] Service, R. (2005) Structural genomics, round 2. *Science*, **307,** 1554-1558.

[24] Chen, L., Oughtred, R., Berman, H.M. and Westbrook, J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **20(16),** 2860-2862.

[25] Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007) XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics*, **23(24),** 3403-3405.

[26] Hui, R. and Edwards, A. (2003) High-throughput protein crystallization. *Journal of Structural Biology*, **142,** 154-161.

[27] Savchenko, A., Yee, A., Khachatryan, A., Skarina, T., Evdokimova, E., Pavlova, M., Semesi, A., Northey, J., Beasley, S., Lan, N., Das, R., Gerstein, M., Arrowmith, C.H. and Edwards, A.M. (2003) Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and x-ray crystallography approaches. *Proteins*, **50,** 392-399.

[28] Chandonia, J.M. and Brenner, S.E. (2005) Implications of structural genomics target selection strategies:

Pfam5000, whole genome, and random approaches. *Proteins*, **58,** 166-179.

[29] McPherson, A. (2004) Protein crystallization in the structural genomics era. *Journal of Structural and Functional Genomics*, **5(1-2),** 3-12.

[30] Chayen, N.E. (2004) Turning protein crystallisation from an art into a science. *Current Opinion in Structural Biology*, **14(5),** 577-583.

[31] Biertumpfel, C., Basquin, J. and Suck, D. (2005) Practical implementations for improving the throughput in a manual crystallization setup. *Journal of Applied Crystallography*, **38,** 568-570.

[32] Puesy, M., Liu, Z.J., Tempel, W., Praissman, J., Lin, D., Wang, B.C., Gavira, J.A. and Ng, J.D. (2005) Life in the fast lane for protein crystallization and X-ray crystallography. *Progress in Biophysics and Molecular Biology*, **88,** 359-386.

[33] Stevens, R.C. (2000) High-throughput protein crystallization. *Current Opinion in Structural Biology*, **10(5),** 558-63.

[34] Rodrigues, A. and Hubbard, R.E. (2003) Making decisions for structural genomics. *Briefings in Bioinformatics*, **4,** 150-167.

[35] Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L.S., Miller, M.D., McPhillips, T.M., Miller, M.A., Scheibe, D., Canaves, J.M., Guda, C., Jaroszewski, L., Selby, T.L., Elsliger, M.A., Wooley, J., Taylor, S.S., Hodgson, K.O., Wilson, I.A., Schultz, P.G., Stevens, R.C. (2002) Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proceedings of the National Academy of Sciences of USA*, **99,** 11664-11669.

[36] Brenner, S.E., Barken, D. and Levitt, M. (1999) The PRESAGE database for structural genomics. *Nucleic Acids Research*, **27(1),** 251-253.

[37] Chance, M.R., Bresnick, A.R. Burley, S.K., Jiang, J.S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M., Wang, L.K. (2002) Structural genomics: pipeline for providing structures for the biologist, *Protein Science*, **11(4),** 723-738.

[38] Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T. and Gerstein, M. (2001) SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Research*, **29,** 2884-2898.

[39] Goh, C.S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G.T. and Gerstein, M. (2003) SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Research*, **31,** 2833-2838.

[40] Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E. and Berman, H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Research*, **4(Database issue),** D302-305.

[41] Berman, H.M. (2008) Harnessing knowledge from structural genomics. *Structure*, **16,** 16-18.

[42] Berman, H.M., Westbrook, J.D., Gabanyi, M.J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L., Kopp, J., Podvinec, M., Adams, P.D., Carter, L.G., Minor, W., Nair, R. and La Baer, J. (2008) The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Research*, **37(Database issue),** D365-368.

[43] Rupp, B. and Wang, J.W. (2004) Predictive models for protein crystallization. *Methods*, **34,** 391-408.

[44] Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M., Arrowsmith, C.H. (2000) Structural proteomics of an archaeon. *Nature Structural Biology*, **7,** 903-909.

[45] Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H. and Gerstein, M. (2004) Mining the structural genomics pipeline: Identification of protein properties that affect high-throughput experimental analysis. *Journal of Molecular Biology*, **336,** 115-130.

[46] Canaves, J.M., Page, R., Wilson, I.A. and Stevens, R.C. (2004) Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: Maximum clustering strategy for structural genomics. *Journal of Molecular Biology*, **344,** 977-991.

[47] Kantardjieff, K.A. and Rupp, B. (2004) Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics*, **20,** 2162-2168.

[48] Kantardjieff, K.A., Jamshidian, M. and Rupp, B. (2004) Distributions of pI vs pH provide strong prior information for the design of crystallization screening experiments. *Bioinformatics*, **20,** 2171-2174.

[49] Longenecker, K.L., Garrard, S.M., Sheffield, P.J. and Derewenda, Z.S. (2001) Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. *Acta Crystallographica Section D: Biological Crystallography*, **57,** 679-688.

[50] Mateja, A., Devedjiev, Y., Krowarsch, D., Longenecker, K., Dauter, Z., Otlewski, J., Derewenda, Z.S. (2002) The impact of Glu-Ala and Glu-Asp mutations on the crystallization properties of RhoGDI: the structure of RhoGDI at 1.3 A resolution. *Acta Crystallographica Section D: Biological Crystallography*, **58,** 1983-1991.

[51] Derewenda, Z.S. (2004) The use of recombinant methods and molecular engineering in protein crystallization. *Methods*, **34,** 354-363.

[52] Derewenda, Z.S. (2004) Rational protein crystallization by mutational surface engineering. *Structure*, **12,** 529-535.

[53] Derewenda, Z.S. and Vekilov, P.G. (2006) Entropy and surface engineering in protein crystallization. *Acta Crystallographica Section D: Biological Crystallography*, **62,** 116-124.

[54] Cooper, D.R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M. and Derewenda, Z. (2007) Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallographica Section D: Biological Crystallography*, **63,** 636-645.

[55] Goldschmidt, L., Cooper, D.R., Derewenda, Z. and Eisenberg, D. (2007) Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein Science*, **16,** 1569-1576.

[56] Oldfield, C.J., Ulrich, E.L., Cheng, Y., Dunker, A.K. and Markley, J.L. (2005) Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins*, **59,** 444-453.

[57] Chandonia, J.M., Kim, S.H. and Brenner, S.E. (2006) Target selection and deselection at the Berkeley Structural Genomics Center. Proteins, **62,** 356-370.

[58] Price, W.N. 2nd, Chen, Y., Handelman, S.K., Neely, H., Manor, P., Karlin, R., Nair, R., Liu, J., Baran, M., Everett, J., Tong, S.N., Forouhar, F., Swaminathan, S.S., Acton, T., Xiao, R., Luft, J.R., Lauricella, A., DeTitta, G.T., Rost, B., Montelione, G.T. and Hunt, J.F.. (2009) Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nature Biotechnology*, **27(1),** 51-57.

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

[60] Chou, K.C. (2005) Progress in protein structural class prediction and its impact to bioinformatics and proteomics. *Current Protein & Peptide Science*, **6,** 423-436.

[61] Yang, Z. R., Wang, L., Young, N. and Chou, K.C. (2005) Pattern recognition methods for protein functional site prediction. *Current Protein & Peptide Science*, **6,** 479-491.

[62] Chou, K.C. and Shen, H.B. (2007) Recent progresses in protein subcellular location prediction. *Analytical Biochemistry*, **370,** 1-16.

[63] Kurgan, L., Cios, K.J., Zhang, H., Zhang, T., Chen, K., Shen, S. and Ruan, J. (2008) Sequence-based methods for real value predictions of protein structure. *Current Bioinformatics*, **3(3),** 183-196.

[64] Smialowski, P., Schmidt, T., Cox, J., Kirschner, A. and Frishman, D. (2006) Will my protein crystallize? A sequence-based predictor. *Proteins*, **62,** 343-355.

[65] Overton, I.M. and Barton, G.J. (2006) A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Letters*, **580,** 4005-4009.

[66] Overton, I.M., Padovani, G., Girolami, M.A. and Barton, G.J. (2008) ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics*, **24,** 901-907.

[67] Slabinski, L., Jaroszewski, L., Rodrigues, A.P.C., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007) The challenge of protein structure determination - lessons from structural genomics. *Protein Science*, **16(11),** 2472-2482.

[68] Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007) XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics*, **23(24),** 3403-3405.

[69] Chen, K., Kurgan, L. and Rahbari, M. (2007) Prediction of protein crystallization using collocation of amino acid pairs. *Biochemical and Biophysical Research Communications*, **355,** 764-769.

[70] Kurgan, L., Razib, A.A., Aghakhani, S., Dick, S., Mizianty, M.J. and Jahandideh, S. (2009) CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Structural Biology*, **9**, 50.

[71] Campbell, K. and Kurgan, L. (2008) Sequence-only based prediction of ß-turn location and type using collocation of amino acid pairs. *Open Bioinformatics Journal*, **2,** 37-49.

[72] Chen, K., Kurgan, L. and Ruan, J. (2007) Prediction of flexible/rigid regions in proteins from sequences using collocated amino acid pairs. *BMC Structural Biology*, **7,** 25.

[73] Chen, Y.Z., Tang, Y.R., Sheng, Z.Y. and Zhang, Z. (2008) Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinformatics*, **9,** 101.

[74] Chen, K., Jiang, Y., Du, L. and Kurgan, L. (2009) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *Journal of Computational Chemistry*, **30(1),** 163-172.

[75] Kurgan L. (2008) On the relation between the predicted secondary structure and the protein size. *The Protein Journal*, **24(4),** 234-239.

[76] Shen, H.B. and Chou, K.C. (2009) Predicting protein fold pattern with functional domain and sequential evolution information. *Journal of Theoretical Biology*, **256(3),** 441-446.

[77] Chen, K. and Kurgan, L. (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics*, **23(21),** 2843-2850.

[78] Assfalg, J., Gong, J., Kriegel, H.P., Pryakhin, A., Wei, T. and Zimek, A. (2009) Supervised ensembles of prediction methods for subcellular localization. *Journal of Bioinformatics and Computational Biology*, **7(2),** 269-285.

[79] 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. *Biochemical and Biophysical Research Communications*, **355(4),** 1006-1011.

[80] Chou, K.C. and Shen, H. B. (2006) Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochemical and Biophysical Research Communications*, **347,** 150-157.

[81] Kedarisetti, K.D., Kurgan, L. and Dick, S. (2006) Classifier ensembles for protein structural class prediction with varying homology. *Biochemical and Biophysical Research Communications*, **348(3),** 981-988.

[82] Chen, H. and Zhou, H.X. (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Research*, **33(10),** 3193-3199.

Vol.1, No.2, 107-119 (2009)
doi:10.4236/ns.2009.12013

Natural Science

Evolution from Primitive Life to *Homo sapiens* Based on Visible Genome Structures: The Amino Acid World

Kenji Sorimachi

Educational Support Center, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan

Received 4 August 2009; revised 23 August 2009; accepted 26 August 2009.

ABSTRACT

It is not too much to say that molecular biology, including genome research, has progressed based on the determination of nucleotide or amino acid sequences. However, these approaches are limited to the analysis of relatively small numbers of the same genes among species. On the other hand, by graphical presentation of the ratios of the numbers of amino acids present to the total numbers of amino acids presumed from the target gene(s) or genome or those of the numbers of nucleotides present to the total numbers of nucleotides calculated from the target gene(s) or genome, we can readily draw conclusions from extraordinarily huge data sets integrated by human intelligence.

1) Assuming polymerization of amino acids or nucleotides in a simulation analysis based on a random choice, proteins were formed by simple amino acid polymerization, while nucleotide polymerization to form nucleic acids encoding specific proteins needed certain specific control. These results proposed that protein formation chronologically preceded codon formation during the establishment of primitive life forms. In the prebiotic phase, amino acid composition was a dominant factor that determined protein characteristics; the "Amino Acid World".

2) The genome is constructed homogeneously from putative small units displaying similar codon usages and coding for similar amino acid compositions; the unit is a gene assembly encoding 3,000 - 7,000 amino acid residues and this unit size is independent not only of genome size, but also of species.

3) In codon evolution, all nucleotide alternations are correlated, not only in coding regions, but also in non-coding regions; the correlations can be expressed by linear formulas; y = ax + b, where "y" and "x" represent nucleotide contents, and "a" and "b" are constant.

4) The basic pattern of cellular amino acid compositions obtained from whole cell lysates is conserved from bacteria to *Homo sapiens*, and resembles that calculated from complete genomes. This basic pattern is characterized by a "star-shape" that changes slightly among species, and changes in amino acid composition seem to reflect biological evolution.

5) Organisms can essentially be classified according to two codon patterns.

Biological evolution due to nucleotide substitutions can be expressed by simple linear formulas based on mathematical principles, while natural selection must affect species preservation after nucleotide alternations. Therefore, although Darwin's natural selection is not directly involved in nucleotide alternations, it contributes obviously to the selection of nucleotide alternations. Thus, Darwin's natural selection is doubtless an important factor in biological evolution.

Keywords: Evolution; Primitive Life Form; Genome; Nucleotide Content; Chargaff's Parity Rules; Codon; Amino Acids; Linear Formula; Classification

1. INTRODUCTION

It is well known that Alfred R. Wallace's theory based on the geographical distribution of animal species, represented by the Wallace line, and the voyage on HMS Beagle, contributed to the development of Darwin's theory.

Molecular biology has progressed with the purification of proteins and the cloning of the genes encoding them, accompanied by sequencing of nucleotides and amino acid residues to understand complicated metabolic pathways. Therefore, the contributions of Frederich Sanger, who developed methods of amino acid [1,2] and nucleotide [3] sequence analyses, and that of Allan Maxam and Walter Gilbert who also developed nucleo-

tide sequence analyses [4], to the development of molecular biology, are inestimable. An approach using nucleotide sequences has a merit that excludes standard errors. Changes in nucleotide or amino acid sequences in a single gene have been applied to evolutionary research based on the assumption that amino acid sequence changes are linked to biological evolution – a "molecular clock" [5]. In general, it is possible to compare sequences among the same kinds of genes or proteins, but it is hard to compare different kinds of genes or their products. Thus, the approach using nucleotide sequences seems not to be suitable for genome research handling genomes consisting of different kinds and numbers of genes among species. On the other hand, focusing on constitutional differences in proteins, the ratios of the numbers of amino acids present to the total numbers of amino acids presumed from the target gene(s) or genome and those of the numbers of nucleotides present to the total numbers of nucleotides in the target gene(s) or genome are applicable for the comparison not only of the same kinds of genes, but also for the comparison of different kinds of genes and different genomes. Ratios based on amino acid or nucleotide sequences can exclude deviations, and the combinations of 20 amino acid or four nucleotide distributions can characterize genomes including a huge amount of data. Therefore, these ratios are a useful tool for genome research, which handles enormously huge data sets. In addition, using certain graphical presentations, huge data sets on genomes can be easily recognized as simple patterns representing complicated organisms.

Graphic representation or a diagram approach to the study of complicated biological systems can provide an intuitive picture and provide useful insights. The historic puzzle of Chargaff's second parity rule in molecular biology has recently been solved using a simple graphic DNA model [6]. Various graphical approaches have been successfully used, for example, to study codon usage [7-12], enzyme catalyzed systems [13-18], and HIV reverse transcriptase inhibition mechanisms [19,20]. Graphical approaches have also been used recently to represent DNA sequences [21].

1.1. Biological Evolution Based on Cellular Amino Acid Compositions

Microorganism fossils were found in 2,500 – 2,800 million year-old rocks [22-24]. Evidence for the existence of microorganisms in ancient rocks indicates that these microorganisms were closed to primitive life forms on earth. *Australopithecus*, the forebears of *Homo sapiens afarensis*, are thought to have appeared about 4 million years ago in Africa, based on the fossil record [25], strongly supporting Darwin's theory and the existence of many extinct species, such as dinosaurs.

The scientific discovery that explained hereditary characteristics was made by James D. Watson and Francis Crick, namely, the double helix structure of DNA [26]. The pairs of A versus T and G versus C in the double helix structure of DNA produce hereditary characteristics in the replication system and transcription system. According to the transcription system, where U is used instead of T in RNA, cellular proteins are the products of DNA, including various genes, which are responsible for genetic characteristics. Thus, cellular proteins naturally reflect genetic characteristics, even though the amount of each protein may differ. Cellular amino acid analysis was first carried out in bacteria by Noboru Sueoka [27]. Then, my group investigated the cellular amino acid composition not only of bacteria, but also of archaea and eukaryotes, and found by graphical presentation of data on radar charts that the basic pattern of cellular amino acid compositions is conserved from bacteria to mammalian cells [28]. This basic pattern, the "star-shape", is formed with high concentrations of Asp, Glu, Gly, Ala, Val, Ile, Leu and Lys, and with low concentrations of Ser, His, Arg, Pro, Tyr, Met, Cys and Phe (**Figure 1**). In archaea [29] and plants [30], similar basic patterns of cellular

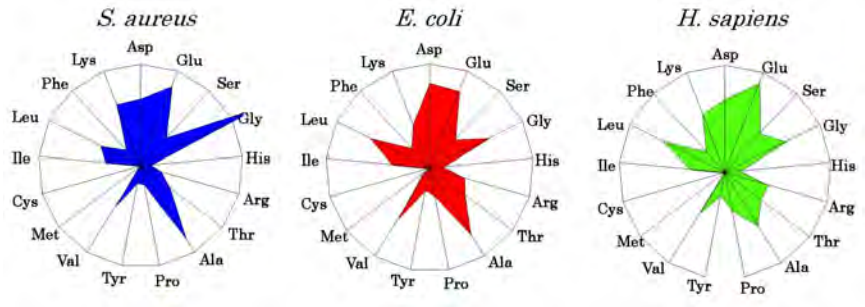


Figure 1. Cellular amino acid compositions on radar charts. The value is expressed as the percentage of total amino acids and in the mean of 3 or 4 independent experiments. Gln and Asn were incorporated into Glu and Asp, respectively, because the former two are converted to the latter two during acidic hydrolysis (Sorimachi 1999). In addition, Try was omitted because of higher decomposition during acidic hydrolysis.

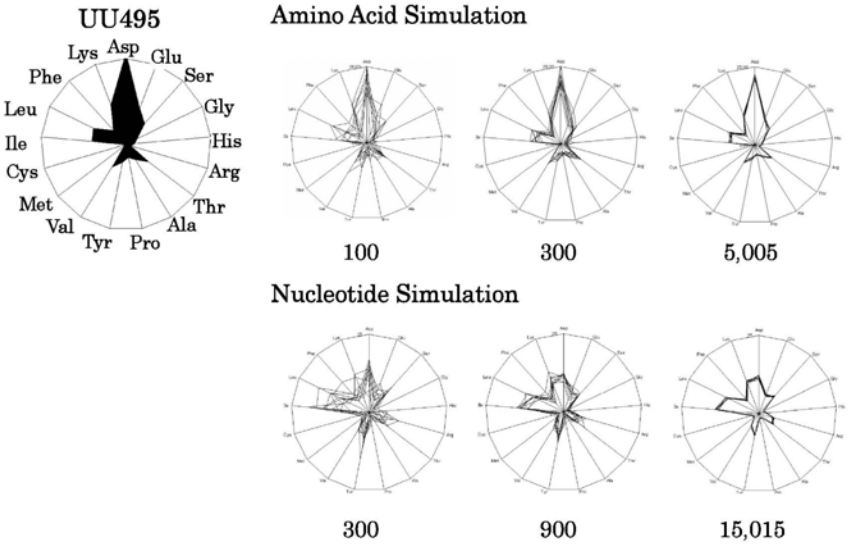


Figure 2. Computational amino acid compositions of *Ureaplasma urealyticum* gene. Upper panel; random choice of amino acid was carried out in the original gene (5,005 amino acid pool). Lower; random choice of nucleotide was carried out in the original gene (15,018 nucleotides). In the simulation using nucleotides, the stop codon and Trp were discarded from the calculation of amino acid compositions, and a triplet formed was immediately counted as an amino acid. This figure was reproduced from Kenji Sorimachi and Teiji Okayasu. (2007) Mathematical proof of the chronological precedence of protein formation over codon formation. Curr. Top. in Pep. Prot. Res. **8,** 25-34.

amino acid compositions are obtained. The fact that the basic pattern, the "star-shape", is conserved from bacteria to *Homo sapiens*, suggests that the pattern is extremely important for organisms on earth. Each amino acid composition changes slightly accompanied with conservation of the basic pattern, and these minor changes seem to reflect biological evolution. Intra-cellular free amino acid compositions also show species-specific patterns [31].

Whole cell lysates consist of many different proteins, the quantities of which show similar amino acid compositions among various organisms; however, species differences are observed. It would be quite interesting to evaluate whether this "star-shape" is conserved on other planets with life in the future, if any are found.

1.2. Primitive Life Formation

Based on the principles of molecular biology, the parental genetic information is transferred to daughter cells by the replication system. The fact that the basic pattern of cellular amino acid composition appears to be conserved from bacteria to *Homo sapiens* suggests that the presumed amino acid composition of primitive life forms might resemble the cellular amino acid composition obtained from modern organisms, because the original pattern could have been maintained by the replication system after codon establishment.

1.3. Chronological Precedence of Protein Formation over Codon Formation

We can easily understand that proteins are translated from codons within genes in modern organisms. However, it is unclear if codon formation really preceded protein formation. Although there have been several reports explaining the mechanisms of codon formation [32-34], no one theory has become established. At present, we cannot experimentally make life in the laboratory, because there are too many unknown factors. On the other hand, computational analysis is an ideal method for solving problems that cannot be solved experimentally. On the basis of molecular biological research, we cannot deny that codons are linked to the determination of the amino acid residues in proteins. Assuming that a structure can sometimes reveal its formation process, it is possible to investigate the relationship between protein and codon formation based on the amino acid compositions presumed from codon usages.

Before establishing the well-known protein synthesis pathway in the presence of codons, protein formation occurred via the polymerization of amino acids, the monomers of proteins. Indeed, amino acid polymerization occurred by heat without enzymes in clay [35]. Proteins can be synthesized computationally by selecting a random order of amino acids from an amino acid pool presumed from a protein. When more than 300 amino

acid residues are chosen at random, the amino acid composition resembles that of the original protein, and amino acid compositions with reduced similarities are obtained by even the first 100 amino acid residues chosen (**Figure 2**). On the other hand, the amino acid composition presumed from more than 900 randomly selected nucleotides, equal to 300 amino acid residues, cannot show the same pattern of amino acid composition. The amino acid composition based on fewer than 300 nucleotides also can not show the specific pattern. These results clearly indicate that mere polymerization of nucleotides, assumed by random choice of nucleotides, can not produce a specific protein. Eventually, the amino acid compositions of proteins obtained from freely polymerized nucleotides depend on both the concentrations of all four nucleotides and the genetic code, and proteins with specific amino acid compositions can not be obtained from nucleic acids formed by free nucleotide polymerization (**Figure 2**). When codon conversion is neglected, the nucleotide composition of polynucleotides can be expressed by a simple quadrangle based on the concentrations of the four nucleotides on radar charts. A consistent result was obtained when various genes were analyzed [36]. In a gene encoding 5,005 amino acid residues, the amino acid compositions of small segments encoding 100 amino acid residues resemble that of the complete gene, and the gene is constructed homogeneously from putative small units encoding similar amino acid compositions [36]. This result, based on gene segments, is consistent with that based on selecting a random order of amino acids or nucleotides. Thus, the initial codon formation might be surely controlled by certain factors to form specific proteins. On the contrary, protein formation could occur via simple polymerization of free amino acids without codons.

1.4. A Hypothesis Based on Simulation Analysis

Although it is difficult for us to envisage an inverse mechanism in which the information within polypeptides is transferred to nucleotide polymerization, this is the mathematical conclusion based on simple simulation analysis using a random choice, which assumes free amino acid or nucleotide polymerizations. In Miller's experiments, which assumed an atmosphere on primitive Earth, certain amino acids were formed by electrical discharges [37]. Amino acids have also been identified in meteorites [38,39]. Thus, proteins might be formed even without codons in prebiotic states, and then polynucleotides, including codons, might be formed under conditions that enabled the transfer of protein information.

Based on this assumption, primitive life forms might have consisted of proteins reflecting the concentrations of free amino acids that existed on primitive Earth. The concentrations of amino acids would have been controlled by various factors, such as gamma rays, UV light and heat, like the natural selection. These effects must have induced homogeneous amino acid concentrations and, eventually, the proteins formed must have had similar amino acid compositions. Indeed, considering the concentrations of each amino acid in cells, the concentrations of those with a benzene ring, Tyr, Phe and His, in their side chains are comparatively very low (**Figure 1**); UV light induces photo-decomposition of organic compounds. For example, the thyroid hormone, thyroxine, an amino acid derivative having two benzene rings its structure, is easily decomposed by UV light irradiation [40,41]. Sometimes, though, this irradiation produces new compounds from certain organic compounds [42,43]. Trp is heat sensitive and is decomposed during cell hydrolysis. On the other hand, the concentrations of amino acids such as Ala, Ile and Leu, with high hydrophobicity, are comparatively high on radar charts. This must have contributed to self-protein assembly from relatively low concentrations of proteins on primitive earth. The hydrophobic interaction must have been an important factor forming the "coacervates" proposed by Aleksandr Ivanovich Oparin. In addition, Gly and Ala were formed in Miller's experiments using electrical charges [37]. In the prebiotic world, amino acid concentration was a dominant factor in the formation of primitive life forms. Therefore, I propose here an existence of the "Amino Acid World" during the prebiotic world based on both experimental and genomic data as a hypothesis of primitive life forms.

A "RNA world" has been proposed as a hypothesis of primitive life forms, as certain RNAs have an enzymatic activity for self replication – "ribozyme" [44]. Even in this case, it is hard to image that free nucleotides formed primitive RNA molecules possessing template characteristics that would induce codon formations. In addition, nucleic acids are very sensitive to UV light, with this light irradiation commonly used for pasteurization. Thus, RNA might not have played a crucial role in primitive life formation on primitive Earth which would have been exposed to strong UV light and gamma rays.

1.5. Homogeneity of Genome Structures

Simulations based on a random choice of amino acids or nucleotides suggest that primitive life forms consisted of proteins formed with the same amino acid compositions, because the amino acid polymerization of proteins occurred in the presence of the same amino acid composition, as mentioned above. Therefore, the genomes of primitive life forms must have been homogeneous in terms of amino acid composition, and this characteristic must have been conserved in the genomes of modern organisms by a late-established replication system. In addition, the basic pattern of cellular amino acid compo-

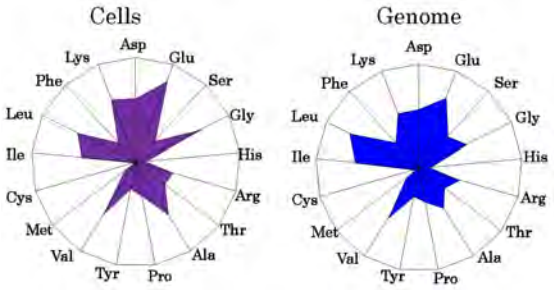


Figure 3. Cellular and genomic amino acid compositions on radar charts. The value is expressed as the percentage of total amino acids. *Methanobacte©rium thermoautotrophicum* was examined. The cellular amino acid composition was obtained from 3 independent analyses. In genomic calculations, Gln and Asn were also incorporated into Glun and Asp, respectively, to compare with data based on amino acid analysis.

sition is conserved from bacteria to *Homo sapiens*, even though the cells are constructed from many different kinds of proteins in different quantities [28]. This measurement of cellular amino acids is experimentally possible at present. However, we cannot evaluate the degree of gene expression of each gene in live cells. To overcome this problem, calculation of gene expression levels was carried out assuming conveniently that each gene is expressed equally [29]; this assumption equally means that the genome is constructed apparently from a single large coding region consisting of many genes, and another single non-coding region. The relationship between nucleotide contents can be expressed by different linear formulas for coding and non-coding regions [11]. This suggests that the two regions were formed at different stages during the establishment of primitive life forms. Surprisingly, the amino acid composition calculated from the complete genome is extremely similar to that obtained from amino acid analysis of cell lysates, as shown in **Figure 3**.

This puzzle was solved as follows. I proposed that a genome may be constructed from putative small units encoding similar amino acid compositions [45]. On the other hand, each gene has a different amino acid sequence and different amino acid composition, although some genes show a similar amino acid composition to the whole group. Thus, a gene assembly containing certain genes can show a similar amino acid composition to the whole group. Similarly, as proteins are gene products, it is possible to assume that cell lysates consist of assemblies of proteins. Therefore, the cellular amino acid composition based on amino acid analysis resembles that based on genomic calculation.

To prove this, the complete genome of the archaeon *Methanobacterium thermoautotrophicum* was examined. Both one-tenth segments (encoding 30,000 – 60,000 amino acid residues) and one-twentieth segments (encoding 20,000 – 30,000 amino acid residues) showed almost the same amino acid composition, and small units encoding 3,000 – 7,000 amino acid residues obtained from genome division showed similar amino acid compositions (**Figure 4**). In *Saccharomyces cerevisiae*, chromosomes of different sizes showed almost the same amino acid composition. As shown in **Fig. 4**, it is clear that the genome is constructed homogeneously from putative small units having almost the same amino acid compositions, not only in bacteria, but also in eukaryotes. The putative unit size is independent of its location in the genome. Obviously, this fact led naturally to synchronous mutations across the genome during biological evolution; and as a result, genome structure is homogeneous based on codon usage [9] and amino acid composition [45].

1.6. Mathematical Proof of the Unit Size

In general, natural proteins are polymers of 20 kinds of amino acid residues. To clarify the reason why a gene assembly encoding 3,000 – 7,000 amino acid residues represents a total population of amino acids based on the complete genome, a multinomial distribution analysis [46] was carried out. In this analysis, 17 amino acid residues were chosen at random from the amino acid pool based on the complete genome to compare the amino acid composition with those calculated from gene assemblies on the complete genome, because Glu and Asp were converted to Gln and Asn, respectively, and Trp was decomposed, during our amino acid analyses using cell lysates [28]. Mathematical analysis clearly showed that the 17-amino acid composition based on a random choice of 3,000 -7,000 amino acid residues represents an amino acid composition with 95% level simultaneous confidence intervals for all amino acid probabilities in the sample [47]. Reducing the level of simultaneous confidence intervals or sample size decreases the similarity of the amino acid composition.

1.7. Bacterial Classification Based on Complete Genomes

Bacteria can be classified by Gram staining into two groups, Gram-positive and Gram-negative bacteria, and both biochemical and morphological characteristics contribute to precise classification [48]. At the end of the 20th century, the methodology for genomic research was established, and the genomes of several hundred bacteria have been completely analyzed to date. The first complete genome analysis of a free-living organism was carried out in *Haemophilus influenzae* in 1995 [49], and the complete human genome was analyzed at the beginning of the 21st century [50,51].

Bacteria seem worthy of classification based on genome sequence, because using the ratios of the numbers of amino acids present to the total numbers of amino

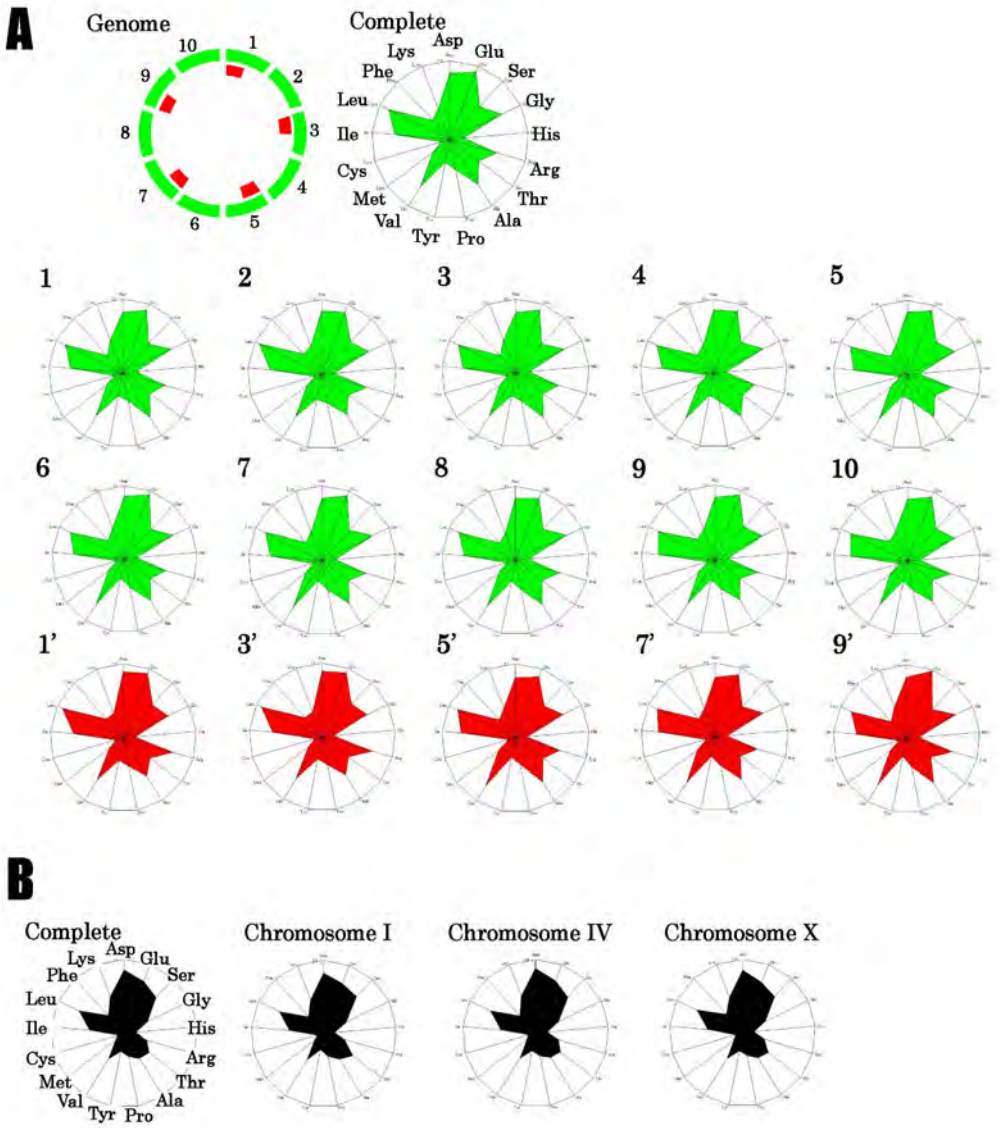


Figure 4. Amino acid compositions calculated from various units of the complete genome of *Methanobacterium autotrophicum* and *Saccharomyces cerevisiae* on radar charts. A, the compete *M. thermoautotrophicum* genome consisting of 1,869 protein genes (Smithe et al. 1997) was divided into 10 (9 units consisting of 186 genes and one units consisting of 195 genes) or 20 (5 units consisting of 93 genes). *B, Scaachromyces cerevisiae*. This figure was reproduced from Kenji Sorimachi and Teiji Okayasu. (2005) Genomic structure consisting of putative units coding similar amino acid composition: synchronous mutations in biological evolution. Dokkyo J. Med. Sci. **32,** 101-106.

acids presumed from the target gene(s) or whole genome, or those of the numbers of nucleotides present to the total numbers of nucleotides in the target gene(s) or whole genome makes it possible to directly compare different genes or genomes, as mentioned above. As the genome is constructed homogeneously from putative small units encoding almost the same amino acid composition, the factor of genome size is irrelevant to comparisons of amino acid compositions.

The patterns of amino acid compositions based on the complete genomes of various bacteria, 11 Gram-positive and 12 Gram-negative bacteria, are star shaped, as mentioned above. According to differences in concentrations of Ala, Arg or Lys, bacteria are classified into two groups, “S-type”, represented by *Staphylococcus aureus*, and “E-type”, represented by *Escherichia coli*; this classification is independent of Gram staining [52]. Differences in Gram staining based on structural differences in cell walls are not detected in genomic structures, while precise changes in amino acid composition, expressed by

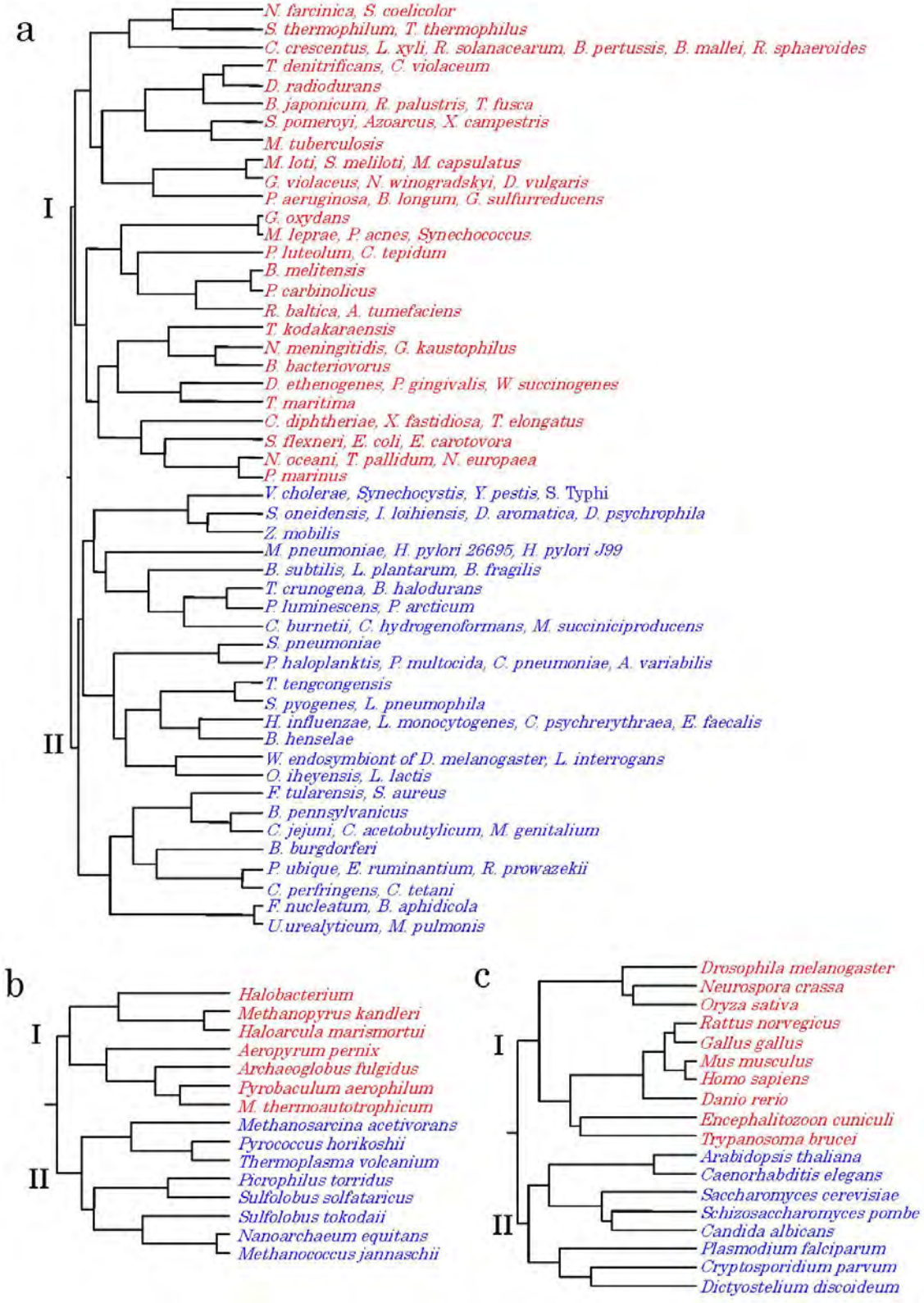


Figure 5. Dendrograms of organism classifications obtained utilizing the Ward method. As traits, GC contents at the three codon positions were used. "a" 112 bacteria, "b" 15 archaea, "c" 18 eukaryotes. Blue charcters represent "AT-type" equal to "S-type" and red represent "GC-type" equal to "E-type". This figure was reproduced from Teiji Okayasu and Kenji Sorimachi. (2009) Organisms can essentially be classified according to two codon patterns, Amino Acids, **36,** 261-271.

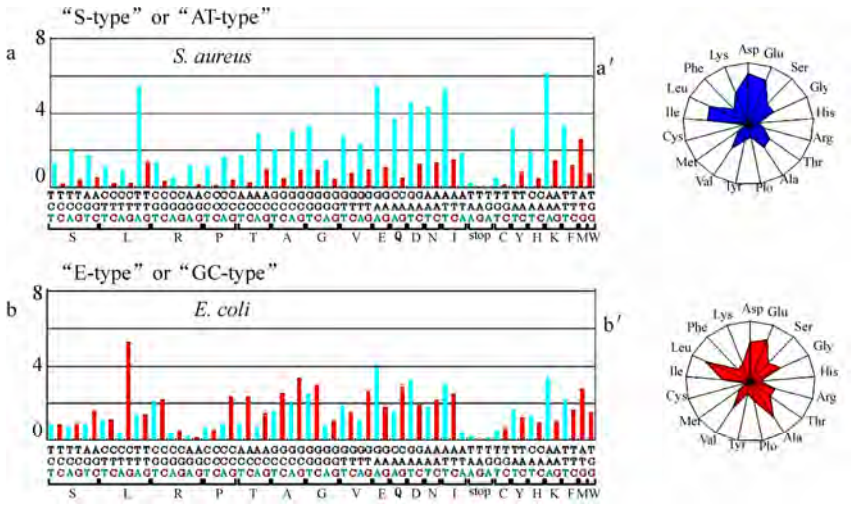


Figure 6. Codon usage patterns and amino acid compositions of *Staphylococcus aureus* and *Escherichia coli*. Codon usage (bar) and amino acid composition (radar chart) were expressed by percent of total codons and amino acids, respectively. These figures were reproduced from Kenji Sorimachi and Teiji Okayasu. (2008) Codon evolution is governed by linear formulas, Amino Acids, **34**, 661-668.

the "star-shape", seem to reflect biological evolution.

1.8. Classification of Organisms into Dendrograms

Changes in nucleotide or amino acid sequences have been applied to evolutionary research and their results are expressed by phylogenic trees on the assumption that these changes are linked to biological evolution [53-58]. This analytical method is applicable to genes for which amino acid or nucleotide sequences have been determined, but it is not suitable for genome research handling extremely huge data sets. In addition, we cannot examine organisms that lack a certain target gene. Using the ratios of the numbers of amino acids present to the total numbers of amino acids presumed from the whole genome or those of the numbers of nucleotides to the total numbers of nucleotides in the whole genome, organisms consisting of numerous different genes can be examined. Indeed, a small number of 23 bacteria has been classified into two groups on the basis of only one amino acid, Arg, Ala or Lys [52]. To quantitatively examine a large number of organisms, multivariate analysis using many factors is applicable to cluster analysis [59]. Organisms consisting of 112 bacteria, 15 archaea and 18 eukaryotes were classified into two major groups by multivariate analysis using GC contents at the three different codon positions, calculated from complete genomes (**Figure 5**). When 20 amino acid concentrations or 64 codon usages are used as traits instead of GC content, similar dendrograms are obtained [59].

The 145 organisms were classified into "GC-type equal to E-type" and "AT-type equal to S-type" represented by high G or C (low T or A, and high A or T (low G or C) contents, respectively, at every third codon position. The organism that has the highest GC content at the third codon position is *Streptomyces coelicolor* [60], and that which has the lowest GC content at the third codon position is *Ureaplasma urealyticum* [61]. Reciprocal changes between G or C and A or T contents at the third codon position occurred synchronously in every codon among the organisms, as shown in **Figure 6**. Thus, all organisms can basically be classified into two groups according to their characteristic codon patterns with low GC and high AT contents at the third codon position, and the opposite. A similar conclusion was obtained from research that examined the content of G + C in a large number of genes [62]. These facts indicate that codon alternations occur synchronously, not only within three codon positions, but also among codons to form new species, as codon alternations occur synchronously over the genome [9,10,45]. This principle is independent of genome size as well as species, from bacteria to *Homo sapiens*.

1.9. Biological Evolution Can Be Expressed by Linear Formulas

A half century ago, two great scientific concepts regarding DNA structures were discovered. One of them is the helical double-stranded structure of DNA [26], which can explain characteristic heredity. Another is Chargaff's parity rules obtained experimentally; Chargaff's first parity rule [63] in which C/G, T/A and (C + T)/(A + G) ratios are one in the DNA extracted from organisms ;

and Chargaff's second parity rule [64] in which these ratios are nearly one in single stranded DNA isolated from double stranded DNA. The first parity rule is entirely based on physicochemical and intra-strand characteristics of nucleotides. Thus, the rule is independent of biological and intra-molecular influences, while biological divergences are excluded from this rule. The relationships between the contents of two nucleotides are expressed by linear lines whose regression coefficients are one based on the first rule. The second rule has historically been a puzzle in molecular biology, because we can not image that the pairings G to C and A to T are formed in the single stranded DNA. This is an intra-molecular rule governing single stranded DNA. Quite recently, however, I was able to solve the puzzle, based on our results that genome structure is homogeneous [6], and that the sizes of the coding regions are nearly equal between the forward and reverse strands [11]. Thus, mitochondrial genome in which coding sizes differ between the forward and reverse strands appears not to be subject to the second parity rule [65,66]. It has been indicated that the double stranded DNA structure is important for biological evolution and that the double strand might be established during primitive life formation [6]. This second parity rule has recently been applied to complete genomes derived from double stranded DNA [67]. Chargaff's rules are universal for all replicating organisms, but they cannot reflect evolutionary differences based on different kingdoms. The findings of certain rules that govern biological evolution will help us to understand scientifically the evolutionary process over an extremely long time and based on unknown factors.

Fortunately, a huge amount of data regarding genomes has been accumulated by a large number of scientists. The present state could not be imagined in Darwin's Age. When nucleotide (G, C, T and A) contents based on complete genomes are plotted against the content each nucleotide among various organisms, their relationships can clearly be expressed by a linear formula, $y = ax + b$, where y and x represent nucleotide contents, and "a" and "b" are constants. These constant values differ between the coding and non-coding regions. This linear relationship is obtained from the complete single-stranded DNA forming the nuclear genome [11,67]. The values of "a" and "b" in either coding or non-coding region differ slightly among kingdoms, such as bacteria, archaea and eukaryotes [11]. Thus, nucleotide alternations are governed by slightly different rules among different kingdoms. Among these linear regression lines, the constant value "b" has never been zero, and the regression coefficients have never been one. This confirms that the formulas differ from Chargaff's formulas, while differences in regression lines among different kingdoms are the results of biological divergence.

As the relationships between two nucleotide contents are expressed by linear experimental formulas among various organisms, the determination of any one nucleotide content can essentially allow the estimation of all four nucleotide contents. In addition, because the relationships between nucleotide content and 64 codon usages are also governed by linear formulas, the 64 codons in the coding region can be estimated from the content of just one nucleotide (**Figure 7**).

In mitochondria and chloroplasts, nucleotide alternations are also expressed by similar linear formulas with

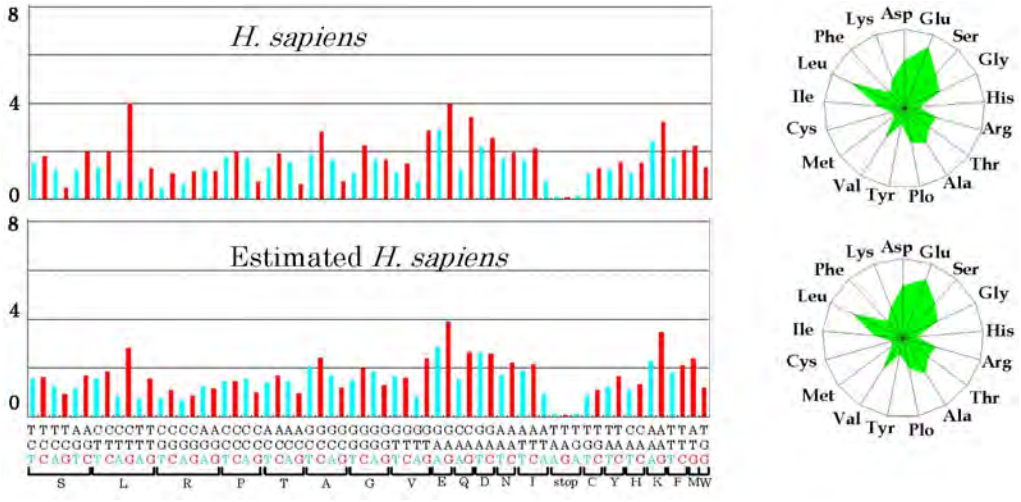


Figure 7. Codon usage patterns and amino acid compositions of *Homo sapience*. Codon usage (bar) and amino acid composition (radar chart) were expressed by percent of total codons and amino acids, respectively. Upper and lower panels represent genomic and estimated data, respectively. These figures were reproduced from Kenji Sorimachi and Teiji Okayasu. (2008) Codon evolution is governed by linear formulas, Amino Acids, **34,** 661-668.

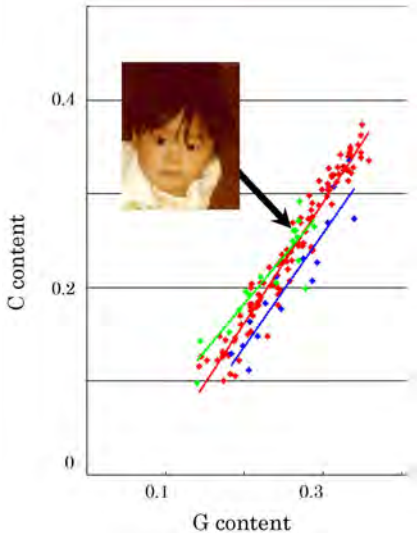


Figure 8. Correlation of G content to C content in various organisms based on their complete genomes. Red, blue and green symbols represent 112 bacter©ia, 15 archaea and 18 eukaryotes, respectively. Each line was drawn computationally. This figure was reproduced from Kenji Sorimachi and Teiji Okayasu. (2008) Codon evolution is governed by linear formulas, Amino Acids, **34,** 661-668

slightly different constant values representing the slope and its intercept [12]. All nucleotide alternations in nuclei, mitochondria and chloroplasts are expressed by linear formulas with different constant values resulting from organelle characteristics among various organisms. Namely, a certain nucleotide content "y" can be expressed inter-species by linear formulas, y = ax + b, based on a single nucleotide content "x". Among four equations presenting four nucleotide contents after normalization, the summation of the value of the slope, "a", is zero and that of the value of constant, "b", is one [11]. This relationship is mathematically definitive and independent of the co-relationships among four nucleotide contents. Chargaff's parity rules, G/C = 1, A/T = 1, (A + G)/(C + T) =1, are alternated as follows: G = G, C = G, T = – G + 0.5, and A = – G + 0.5. Thus, Chargaff's parity rules, even those governing single species DNA, are derived from the general formula, y = ax + b, when slope, "a" of the two equations' is 1 or – 1, and when the intercept, "b", is 0.5 or 0 in the equation with – 1 and 1, respectively, as the "a". On the other hand, the values of "a" and "b" in both codon evolution [11] and organelle evolution [68] shifted from 1 or – 1 and 0.5 or 0, respectively because of biological divergences, and the regression coefficient also shifted from one. The shift of the regression coefficient from one represents biological divergence.

It has been thought that cellular organelle such as mitochondria [68] and chloroplasts [69] were derived during biological evolution from protobacteria and cyanobacteria, respectively, and that their evolutionary processes appear different from nuclear genome evolution、as mentioned above. In addition, it is known that mutation rate is remarkably high in mitochondrial DNA [70]. In our study, amino acid compositions of chloroplast and plant mitochondria resemble those of nuclear DNA, whereas those of vertebrate mitochondria differ from those of other organelle [12]. Particularly, the content of Leu was extremely high in animal mitochondria [12]. Comparing the shapes of the radar charts based on amino acid compositions, that of the ancient fish, the coelacanth (*Latimeria chalumnae*), more closely resembles those of salamanders and birds compared than those of other fish (*Diodon holocunthus*) [12]. In further study, using multivariate analysis based on amino acid compositions, lung fish (*Neoceratodus forsteri*) and coelacanth were both found to belong to the cluster representing a reptile; a cluster separated from that one representing other fish (carp, rainbow trout and killifish). These results are consistent with the already established phylogenic concept.

The apparent great divergence of *Homo sapiens* from bacteria can be expressed by linear formulas with small turbulences based on the complete genome in biological evolution. Thus, biological evolution seems to be observed as a result of mere nucleotide substitutions based on simple mathematical principles, while natural selection affects species preservation after nucleotide alternations. This conclusion is consistent with the idea that evolution is based on neutral mutation [71,72]. Therefore, natural selection does not directly regulate nucleotide substitutions, but is indirectly involved in biological evolution.

2. PERSPECTIVES

The present paper reveals that the analytical method using the ratios of the numbers of amino acids present to the total numbers of amino acids presumed from the whole genome, or those of the numbers of nucleotides present to the total numbers of nucleotides in the whole genome is useful for genome research, as well as methods using the sequences of amino acids or nucleotides. These ratios based on nucleotide sequences can exclude deviations in certain calculations. The fact that genome structures regarding amino acid compositions or codon usages are homogeneous makes it possible for us to compare various genomes with different sizes and genes. Namely, a large data set obtained from the complete genome can be expressed by just a simple point on a graph. Thus, using the ratios of amino acids or nucleotides to their total numbers seems to be an excellent method for genome research based on extremely huge data sets. In

addition, even a certain size of gene assembly can be used instead of the complete genome for limited purposes.

In prebiotic evolution, amino acid composition might have been the strongest factor determining the characteristics of biopolymers used for the establishment of primitive life forms, whereas since the establishment of the codon system, biological evolution has been carried out by nucleotide alternations expressed by linear formulas based on nucleotide contents, as shown in **Figure 8**. Thus, 64 codon usages can be estimated from just one nucleotide content (**Figure 7**), and the characteristic amino acid composition is expressed by the "star-shape" (**Figures 1-7**), not only in cell analysis, but also in genome analysis. This fact strongly suggests that this "star-shape" may be conserved in both primitive life forms and future organisms, because all organisms must be governed by universal rules on earth, without exception. Thus, this amino acid composition represented by the "star-shape" may reflect the "Amino Acid World".

We, *Homo sapiens*, stand merely in the middle of a line (**Figure 8**). We are not the end of line, nor do we have an "ultimate" status. Therefore, we have been and will be exposed to natural selection without exception.

3. ACKNOWLEDGMENTS

The author expresses his thanks to Professor Kuo-Chen Chou, Chief-in-Editor of Natural Science, for the opportunity to write this review; to Professor Hiroto Naora, Research School of Biological Sciences, Australian National University; Professor Makoto Miyaji, Chiba University, and Dr. Emiko Furuta, Institute of Comparative Immunology, for encouragement given in respect of the author's genome research, to Dr. Teiji Okayasu, Dokkyo Medical University, for help with computer analysis of genomic data, and to Dr. Kazumi Akimoto, Dokkyo Medical University for taking care of cell cultures.

REFERENCES

[1] Sanger, F. and Thompson, E.O. (1953) The amino acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem. J.*, **53,** 353-366.

[2] Sanger F. and Thompson, E.O. (1953) The amino acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem. J* , **53,** 366-374.

[3] Sanger, F. and Coulson, A.R. (1975) A rapid method for determing sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94,** 441-446.

[4] Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci.*, USA **74,** 560-564.

[5] Zuckerkandl, E. and Pauling, L.B. (1962) Molecular disease, evolution, and genetic heterogeneity in Kasha M and Pullman B (editors). *Horizons in Biochemistry*, Academic Press, New York, 189-225.

[6] Sorimachi, K. (2009) A proposed solution to the historic puzzle of Chargaff's second parity rule. *Open Genom. J.*, **2,** 12-14.

[7] Chou, K-C. and Zhang, C.T. (1992) Diagrammatization of codon usage in 339 HIV proteins and its biological implication. *AIDS Research and Human Retroviruses*, **8,** 1967-1976.

[8] Zhang, C-T. and Chou, K-C. (1993) Graphic analysis of codon usage strategy in 1490 human proteins. *J. Prot. Chem.*, **12,** 329-335.

[9] Sorimachi, K. and Okayasu, T. (2004) An evolutionary theories based on genomic structures in *Saccharomyces cerevisiae* and *Enchephalitozoon cuniculi*. *Mycoscience* , **45,** 345-350.

[10] Sorimachi, K. and Okayasu, T. (2007) Genomic structure is homogeneous based on codon usages. *Curr. Top. Pep. Protein Res.*, **8,** 19-24.

[11] Sorimachi, K. and Okayasu, T. (2008) Codon evolution is governed by linear formulas. *Amino Acids*, **34,** 661-668.

[12] Sorimachi, K. and Okayasu, T. (2008) Universal rules governing genome evolution expressed by linear formulas. *Open Genom. J.*, **1,** 33-43.

[13] Chou, K-C. (1983) Advances in graphical methods of enzyme kinetics. *Biophys. Chem.*, **17,** 51-55.

[14] Chou, K-C. (1989) Graphical rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.*, **264,** 12074-12079.

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

[16] Chou, K-C. (1993) Graphic rule for non-steady-state enzyme kinetics and protein folding kinetics. *J. Math. Chem.*, **12,** 97-108.

[17] Lin, S.X. and Neet, K.E. (1990) Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. *J. Biol. Chem*., **265,** 9670-5.

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

[19] Althaus, I.W., Chou, J.J., Gonzales, A.J. et al. (1993) Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry*, **32,** 6548-6554.

[20] Chou, K-C., Kezdy, F.J. and Reusser, F. (1994) Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal. Biochem.*, **221,** 217-230.

[21] Qi, X.Q., Wen, J. and Qi, Z.H. (2007) New 3D graphical representation of DNA sequence based on dual nucleotides. *J. Theoret. Biol*., **249,** 681–690.

[22] MacGregor, I.M., Truswell, J.F. and Eriksson, K.A. (1974) Filamentous alga from the 2,300 m.y. old Transvaal Dolomite. *Nature*, **247,** 538-539.

[23] Nagy, L.A. and Zumberge, J.E. (1976) Fossil microorganisms from the approximately 2800 to 2500 million-year-old Bulawayan stromatolite: Application of ultrami-

crochemical analyses. *Proc. Natl. Acad. Sci. USA*, **73,** 2973-2976.
[24] Schopf, J.W., Barghoorn, E.S., Maser, M.D. and Gordon, R.O. (1965) Electron microscopy of fossil bacteria two billion years old. *Science*, **149,** 1365-1367.
[25] Johanson, D.C. and Taieb, M. (1976) Plio-Pleistocene hominid discoveries in Hadar, Ethiopia. *Nature*, **260,** 293-297.
[26] Watson, J.D. and Crick, F.H.C. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171,** 964-967.
[27] Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition in proteins. *Proc. Natl. Acad. Sci. USA*, **47,** 1141-1149.
[28] Sorimachi, K. (1999) Evolutionary changes reflected by the cellular amino acid composition. *Amino Acids*, **17,** 207-226.
[29] Sorimachi, K., Itoh, T., Kawarabayasi, Y., Okayasu, T., Akimoto, K. and Niwa, A. (2001) Conservation of the basic pattern of cellular amino acid composition during biological evolution and the putative amino acid composition of primitive life forms. *Amino Acids*, **21,** 393-399.
[30] Sorimachi, K., Okayasu, T., Akimoto, K. and Niwa, A. (2000) Conservation of the basic pattern of cellular amino acid composition during biological evolution in plants. *Amino Acids*, **18,** 193-196.
[31] Sorimachi, K. (2002) The classification of various organisms according to the free amino acid composition change as the result of biological evolution. *Amino Acids*, **22,** 55-69.
[32] Woese, C.R. (1965) Order in the genetic code. *Proc. Natl. Acad. Sci. USA*, **54,** 71-75.
[33] Crick, F.H.C. (1968) The origin of genetic code. *J. Mol. Biol.*, **38,** 367-379.
[34] Wong, J.T-F. (1975) A co-evolutionary theory of the genetic code. *Proc. Natl. Acad. Sci. USA*, **72,** 1909-1912.
[35] Lahav, N., White, D. and Chang, S. (1978) Peptide formation in the prebiotic era: thermal condensation of glycine in fluctuating clay environments. *Science*, **201,** 67-69.
[36] Sorimachi, K. and Okayasu, T. (2007) Mathematical proof of the chronological precedence of protein formation over codon formation. *Curr. Top. Pep. Protein Res.*, **8,** 25-34.
[37] Miller, S.L. (1953) A production of amino acids under possible primitive earth conditions. *Science*, **117,** 528-529.
[38] Kvenvolden, K., Lawless, J., Pering, K., Peterson, E., Flores, J., Ponnamperuma, C., Kaplan, I.R. and Moore, C. (1970) Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. *Nature*, **228,** 923-926.
[39] Wolman, Y., Haverland, W. and Miller, S.L. (1972) Nonprotein amino acids from spark discharges and their comparison with the Muchison meteorite amino acids. *Proc. Natl. Acad. Sci. USA*, **69,** 809-811.
[40] Sorimachi, K. and Ui, N. (1975) Ion-exchange chromatographic analysis of iodothyronines. *Anal. Biochem.*, **67,** 157-165.
[41] van der Walt, B, Cahnmann, H.J. (1982) Synthesis of thyroid hormone metabolites by photolysis of thyroxine and thyroxine analogs in the near UV. *Proc. Natl. Acad. Sci. USA*, **79,** 1492-1496.
[42] Shizuka, H., Sorimachi, K., Morita, T., Nishiyama, K. and Sato, T. (1971) Photochemical oxidation of 4, 5, 9, 10tetrahydropyrenes. *Bull. Chem. Soc. Japan*, **44,** 1983-1984.
[43] Sorimachi, K., Morita, T. and Shizuka, H. (1974) Photocyclization of 〔2,2〕metacyclophane at 2537 A.. *Bull. Chem. Soc. Japan*, **47,** 987-990.
[44] Gilbert, W. (1986) The RNA World. *Nature*, **319,** 618.
[45] Sorimachi, K. and Okayasu, T. (2003) Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. *Mycoscience*, **44,** 415-417.
[46] Hochberg, Y. and Tamhane, A.C. (1987) Multiple comparison procedures, In Probability and Mathematical Statistics (eds. Y. Hochberg and A.C. Tamhane), John Wiley & Sons, New York, 274-309.
[47] Sorimachi, K., Okayasu, T., Ebara, Y. and Nakagawa, T. (2005) Mathematical proof of genomic amino acid composition homogeneity based on putative small unts. *Dokkyo J. Med. Sci*., **32,** 99-100.
[48] Bergey's Mmanual of Systemic Bacteriology.
[49] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R. *et al.* (1995) Whole -genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269, 496-512.
[50] International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409:** 860-921.
[51] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291,** 1304-1351.
[52] Sorimachi, K. and Okayasu, T. (2004). Classification of eubacteria based on their complete genome: where does Mycoplasmataceae belong? *Proc. R. Soc. Lond. B (Suppl.)*, **271,** S127-S130.
[53] Dayhoff, M.O., Park, C.M. and McLaughlin, P.J. (1977) Building a phylogenetic trees: cytochrome C. In: Atlas of protein sequence and structure. National Biomedical Foundation, Washington, D.C., **5,** 7-16.
[54] Sogin, M.L., Elwood, H.J. and Gunderson, J.H. (1986) Evolutionary diversity of eukaryotic small subunit rRNA genes. *Proc Natl Acad Sci USA*, **83,** 1383-1387.
[55] DePouplana, L., Turner, R.J., Steer, B.A. *et al*. (1998) Genetic code origins: tRNAs older than their synthetases? *Proc Natl Acad Sci USA*, **95,** 11295-11300.
[56] Doolittle, W.F. and Brown, J.R. (1994) Tempo, mode, the progenote, and the universal root. *Proc Natl Acad Sci USA*, **91,** 6721-6728.
[57] Maizels, N. and Weiner, A.M. (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc Natl Acad Sci USA*, **91,** 6729-6734.
[58] Sakagami, M., Nakayama, T., Hashimoto, T. *et al*. (2006) Phylogeny of the centrohelida inferred from SSU rRNA, tubulin, and actin genes. *J. Mol. Evol.*, **61,** 765-775.
[59] Okayasu, T. and Sorimachi, K. (2008) Organisms can essentially be classified according to two codon patterns. *Amino Acids*, **36,** 261-271.
[60] Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, M.A., Challis, G.L., Thompson, N.R., James, K.D., Harris, D.E.,

Quail, M.A., Kieser, H., Harper, D. *et al*. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417,** 141-147.

[61] Glass, J.I., Lefkowitz, E.J., Glass, J.S., Heiner, C.R., Chen, E.Y. and Cassell, G.H. (2000) The complete sequence of the mucosal pathogen Ureaplasma urealyticum. *Nature*, **407,** 757-762.

[62] Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA*, **85,** 2653-2657.

[63] Chargaff, E. (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, **VI,** 201-209.

[64] Rundner, R., Karkas, J.D., and Chargaff, E. (1968) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. USA*, **60,** 921-922.

[65] Nikolaou, C. and Almirantis, Y. (2006) Deviations from Chargaff's second parity rule in organelle DNA insights into the evolution of organelle genomes. *Gene*, **381,** 34-41.

[66] Bell, S.J. and Forsdyke, D.R. (1999) Deviations from Chargaff's second parity rule with direction of transcription. *J. Theor. Biol*., **197,** 63-76.

[67] Mitchell, D. and Bridge, R. (2006) A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun*., **340,** 90-94.

[68] Gray, M.W., Burger, G. and Lang, B.F. (1999) Mitochondrial evolution. *Science*, **283,** 1476-1481.

[69] Raven, J.A. and Allen, J.F. (2003) Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genom. Biol*., **4,** 209-215.

[70] Brown, W.M., George, Jr.M. and Wilson, A.C. (1979) Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA*, **76,** 1967-1971.

[71] Kimura M. (1983) The neutral theory of molecular evolution. Cambridge, Cambridge Univ. Press.

[72] Van Nimwegen, E., Crutchfield, J.P. and Huynen, M. (1999) Neutral evolution of mutational robustsness. *Proc. Natl. Acad. Sci. USA*, **96,** 9716-9720.

Vol.1, No.2, 120-123 (2009)
doi:10.4236/ns.2009.12014

Natural Science

An Improved Model for Bending of Thin Viscoelastic Plate on Elastic Foundation

Zhi-Da Li[1,2], Ting-Qing Yang[1], Wen-Bo Luo[3]

[1]Department of Mechanics, Huazhong University of Science and Technology, Wuhan, China; zhidali@163.com
[2]School of transportation, Wuhan University of Technology, Wuhan, China
[3]College of Civil Engineering and Mechanics, Xiangtan University, Xiangtan, China

Received 10 June 2009; revised 16 July 2009; accepted 20 July 2009.

ABSTRACT

An improved model for bending of thin viscoelastic plate resting on Winkler foundation is presented. The thin plate is linear viscoelastic and subjected to normal distributed loading, the effect of normal stress along the plate thickness on the deflection and internal forces is taken into account. The basic equations for internal forces and stress distribution are derived based on the general viscoelastic theory under small deformation condition. The reduced equations for elastic case are given as well. It is shown that the proposed model reveals a larger flexural rigidity compared to that in classic models, in which the normal stress along the plate thickness is neglected.

Keywords: Thin Viscoelastic Plate; Deformable Foundation; Flexural Rigidity; Winkler Foundation

1. INTRODUCTION

The analysis of soil-structure interaction has a wide range of applications in structural and geotechnical engineering, for instance, in highway asphalt pavement engineering, the pavement is usually treated as thin elastic/viscoelastic plate structure resting on elastic/viscoelastic foundation. Due to the complexity of the actual behavior of foundations, many idealized foundation models have appeared in the literature [1]. The simplest of those models, which was proposed in 1867 by Winkler, assumes that the soil medium consists of a system of mutually independent spring elements. There are many papers dealing with the elastic beam or plates resting on the Winkler foundation in the literature [2,3]. As computational power has developed, more realistic modeling of soil-structure interaction has become possible. Because of the importance of viscoelastic nature of the materials used for structures, e.g. asphalt layer of pavement structure, many works have been done to deal with the bending behaviour of thin viscoelastic plate on elastic/viscoelastic foundation. Most of such works utilized the models similar to those for bending of elastic plate. Mase [4] directly offered the fundamental equations for bending of viscoelastic plate by replacing the flexural rigidity of elastic plate with the rigidity of viscoelastic plate. Radovskii [5] discussed the problem of treating highway and airport pavement as thin viscoelastic plate. Pister [6], Robertson [7] and Hewitt and Mazumdar [8] applied the elasticity-viscoelasticity correspondence principle to get the solution of the bending problem of viscoelastic plate. In contrast to dealing with the viscoelastic plate on elastic foundation, some attempts have also been made to solve the bending problem of elastic plate resting on viscoelastic foundation. Sonoda et al [9,10] studied the circular and rectangular plate on linear viscoelastic foundation. Lin [11] and Yang *et al.* [12] analyzed the dynamic response of circular plate resting on viscoelastic half space. All of the above studies followed the classic model and traditional flexural rigidity for thin plate bending, in which the Kirchhoff hypothesis was used and σ_{zz} was neglected [13]. However, in the case of large lateral load subjecting to thin plate resting on a deformable foundation with relatively large rigidity, the bearing stresses along the plate thickness, σ_{zz}, may not be ignored. Furthermore, it is the bearing stress of the plate that transfer the active lateral load to the foundation, the boundary condition on the main surfaces of the plate should be satisfied. Therefore it is necessary to develop a method to consider the effect of lateral normal stress. In this paper, we first seek to develop a modified Kirchhoff theory for thin viscoelastic plate resting on Winkler foundation, in which the effect of the lateral normal stress is considered. Then we reduce the obtained results to the problem of thin elastic plate resting on elastic foundation, and a different elastic flexural rigidity is obtained.

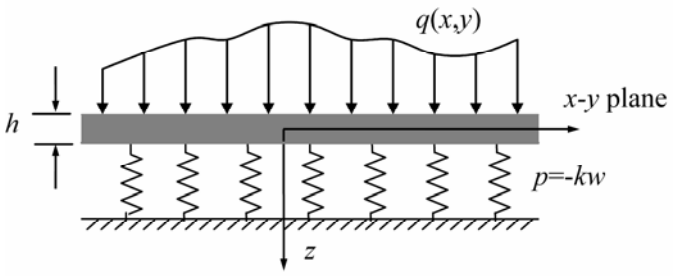


Figure 1. Thin viscoelastic plate resting on a Winkler foundation.

2. VISCOELASTIC PLATE ON WINKLER FOUNDATION: AN IMPROVED MODEL

Figure 1 illustrates a thin viscoelastic plate of a thickness h, its middle-plane coincides with the x-y plane of the reference coordinate system. Let the upper surface ($z=-h/2$) be subjected to a normal distributed loading with intensity of $q(x,y)$, while the lower surface ($z=h/2$) rests on a Winkler type foundation. In the Winkler foundation model, the foundation for the plate is assumed to act like a set of springs. Thus the foundation reaction force can be written as $p=-kw$, where k denotes the elastic stiffness of the foundation and w is the lateral deflection of the plate.

For a linear viscoelastic plate, the equilibrium equations, the strain-displacement relations and the constitutive equations are given, respectively, by

$$\sigma_{ij;j}=0 \tag{1}$$

$$\varepsilon_{ij}=\frac{1}{2}\left(u_{i;j}+u_{j;i}\right) \tag{2}$$

$$s_{ij}=2G_1*de_{ij},\quad \sigma_{kk}=3K_1*d\varepsilon_{jj} \tag{3}$$

where the body forces are neglected, $G_1(t)$ and $K_1(t)$ are the shear modulus function and volume modulus function, respectively, the * denotes convolution product. The deviatoric components of stress and strain tensors are

$$s_{ij}=\sigma_{ij}-\frac{1}{3}\delta_{ij}\sigma_{kk}\quad \text{and}\quad e_{ij}=\varepsilon_{ij}-\frac{1}{3}\delta_{ij}\varepsilon_{kk} \tag{4}$$

For simplicity, the displacement components along *x*, *y* and *z* axis are denoted by *u*, *v* and *w* respectively. Assuming $\varepsilon_{zz}=0$, $\varepsilon_{zx}=0$, $\varepsilon_{zy}=0$, from **Eq.2** we have

$$w=w(x,y,t),\quad u=-zw_{;x},\quad v=-zw_{;y} \tag{5}$$

and

$$\varepsilon_{xx}=-zw_{;xx},\quad \varepsilon_{yy}=-zw_{;yy},\quad \varepsilon_{xy}=-zw_{;xy},\quad \varepsilon_{kk}=-z\nabla^2 w \tag{6}$$

Combining **Eq.4** and **Eq.3**, along with substitution of **Eq.6** into it, the stress components can be written as

$$\sigma_{xx}=-2zG_1*dw_{;xx}-z\left(K_1-\frac{2}{3}G_1\right)*d\left(\nabla^2 w\right) \tag{7}$$

$$\sigma_{yy}=-2zG_1*dw_{;yy}-z\left(K_1-\frac{2}{3}G_1\right)*d\left(\nabla^2 w\right) \tag{8}$$

$$\sigma_{xy}=-2zG_1*dw_{;xy} \tag{9}$$

The other three components of the stress tensor can be obtained by using the equilibrium equations, **Eq.1**. Integrating the first and second representations of **Eq.1** over *z* and considering the boundary condition $\sigma_{zx}=\sigma_{zy}=0$ for $z=\pm h/2$ (i.e. there is no shear stress on the plate surface and no friction between plate and foundation), we get

$$\sigma_{zx}=\left(z^2-\frac{h^2}{4}\right)\left(\frac{K_1}{2}+\frac{2}{3}G_1\right)*d\left(\nabla^2 w_{;x}\right) \tag{10}$$

$$\sigma_{zy}=\left(z^2-\frac{h^2}{4}\right)\left(\frac{K_1}{2}+\frac{2}{3}G_1\right)*d\left(\nabla^2 w_{;y}\right) \tag{11}$$

The third representation of **Eq.1** is

$$\frac{\partial\sigma_{zz}}{\partial z}=-\frac{\partial\sigma_{zx}}{\partial x}-\frac{\partial\sigma_{zy}}{\partial y}$$

Substituting **Eqs.10** and **11** into the above equation and integrating it over *z* yields

$$\sigma_{zz}=\left[\frac{h^2}{4}\left(z-\frac{h}{2}\right)-\frac{1}{3}\left(z^3-\frac{h^3}{8}\right)\right]\left(\frac{K_1}{2}+\frac{2}{3}G_1\right)*d\left(\nabla^4 w\right)-kw \tag{12}$$

It should be noted here that the stress boundary condition $\sigma_{zz}=-kw$ at the lower surface ($z=h/2$) of the plate is used to determine the integral constant. Moreover, $\sigma_{zz}=-q$ at the upper surface ($z=-h/2$) of the plate, thus

$$\frac{h^3}{12}\left(K_1+\frac{4}{3}G_1\right)*d\left(\nabla^4 w\right)=q-kw \tag{13}$$

This is a differential-integral equation interrelating the materials' properties, the applied normal loads and the corresponding lateral deflection.

We denote by M_{ij} the bending moments and the twisting moments, and Q_j the shear forces

$$M_{ij}=\int_{-h/2}^{h/2}\sigma_{ij}z\,dz,\quad Q_j=\int_{-h/2}^{h/2}\sigma_{zj}dz$$

Table 1. Variation of D_1/D with Poisson ratio.

μ	0	0.2	0.25	0.30	0.35	0.40	0.45	0.49	0.5
D_1/D	1	1.067	1.125	1.225	1.408	1.80	3.025	13.005	∞

Substituting **Eqs.7-11** into the above expressions, we get

$$M_{xx} = -\frac{h^3}{12}\left[2G_1 * dw_{;xx} + \left(K_1 - \frac{2}{3}G_1\right) * d\left(\nabla^2 w\right)\right] \quad \textbf{(14)}$$

$$M_{yy} = -\frac{h^3}{12}\left[2G_1 * dw_{;yy} + \left(K_1 - \frac{2}{3}G_1\right) * d\left(\nabla^2 w\right)\right] \quad \textbf{(15)}$$

$$M_{xy} = -\frac{h^3}{6}G_1 * dw_{;xy} \quad \textbf{(16)}$$

$$Q_x = -\frac{h^3}{12}\left(K_1 + \frac{4}{3}G_1\right) * d\left(\nabla^2 w_{;x}\right) \quad \textbf{(17)}$$

$$Q_y = -\frac{h^3}{12}\left(K_1 + \frac{4}{3}G_1\right) * d\left(\nabla^2 w_{;y}\right) \quad \textbf{(18)}$$

Furthermore, the stress components can be expressed in the form of M_{ij} and Q_j as follows:

$$\sigma_{ij} = \frac{12}{h^3} z M_{ij}, \quad (i, j = x, y) \quad \textbf{(19)}$$

$$\sigma_{zj} = \frac{3}{2h}\left[1 - \left(\frac{2z}{h}\right)^2\right] Q_j, \quad (j = x, y) \quad \textbf{(20)}$$

$$\sigma_{zz} = -2(q - kw)\left(\frac{z}{h} - \frac{1}{2}\right)^2\left(1 + \frac{z}{h}\right) - kw \quad \textbf{(21)}$$

The forms of **Eqs.19-21** are as same as those for elastic bending plate [13,14], however it should be noted that M_{ij} and Q_j involved are the moments and the shear forces in viscoelastic cases.

3. ELASTIC PLATE ON WINKLER FOUNDATION

In the case of thin elastic plate with material constants of G and K, resting on a Winkler foundation, the **Eq.13** reduces to the following equation

$$D_1 \nabla^4 w = q - kw \quad \textbf{(22)}$$

in which

$$D_1 = \frac{h^3}{12}\left(K + \frac{4}{3}G\right) = \frac{Eh^3}{12\left(1-\mu^2\right)} \cdot \frac{(1-\mu)^2}{1-2\mu} \quad \textbf{(23)}$$

E and μ are the elastic modulus and Poisson ratio of the elastic plate, respectively.

Again, it can be seen that the form of the control equation for elastic plate bending is similar to that derived from classic theory. However the flexural rigidity D_1 is different from that in classic theory, which gives $D = Eh^3/12\left(1-\mu^2\right)$. Obviously, $D_1/D = \left(1-\mu^2\right)/(1-2\mu)$. It can be seen that D_1/D increases with the Poisson's ratio. Several results are listed in **Table 1**. For most engineering structural materials, μ is about 0.3 [13], thus the flexural rigidity given in the present model is about 22% larger than that in the classic model.

In elastic cases, **Eqs.14-18** reduces to the followings in terms of D_1:

$$M_{xx} = -D_1\left[w_{;xx} + \frac{\mu}{1-\mu} w_{;yy}\right] \quad \textbf{(24)}$$

$$M_{yy} = -D_1\left[w_{;yy} + \frac{\mu}{1-\mu} w_{;xx}\right] \quad \textbf{(25)}$$

$$M_{xy} = -D_1\left[\frac{1-2\mu}{1-\mu}\right] w_{;xy} \quad \textbf{(26)}$$

$$Q_x = -D_1 \nabla^2 w_{;x} \quad \textbf{(27)}$$

$$Q_y = -D_1 \nabla^2 w_{;y} \quad \textbf{(28)}$$

Moreover, the stress distribution in the elastic plate can be read in the same forms as **Eqs.19-21** if the moments and shear forces are given by **Eqs.24-28**.

4. CONCLUDING REMARKS

We have derived in this paper the basic equations for bending of viscoelastic thin plate on Winkler foundation. In classic treatment, the normal stress σ_{zz} is assumed to be zero, which is the main characteristic of the plane stress state in classic elastic theory, and thus the corresponding classic bending model of thin elastic plate can be considered as a plane stress model. In contrast with the classic treatment, in the proposed model in this paper, σ_{zz} is taken into account and the general viscoelastic constitutive equation is used to represent the thin plate behavior, and the plate thickness is further set to be constant, i.e. $\varepsilon_{zz} = 0$, which is the main characteristic of the plane strain state in classic elastic theory. Thus we may consider this improved model as a quasi-plane strain model.

5. ACKNOWLEDGMENTS

The study was supported by the National Natural Sci-

ence Foundation of China (No.10372074).

REFERENCES

[1] Selvadurai, P.S. (1979) Elastic analysis of soil-foundation interaction. Amsterdam: Elsevier.

[2] Kaschiev, M.S. and Mikhajlov, K. (1995) A beam resting on a tensionless Winkler foundation. *Computer and Structures*, **55,** 261-264.

[3] Kim, S.M. and Roesset J.M. (1998) Moving loads on a plate on elastic foundation. *Journal of Engineering Mechanics*, **124,** 1010-1017.

[4] Mase, G.E. (1960) Behavior of viscoelastic plates in bending. *Journal of Engineering Mechanics*, **86,** 25-39.

[5] Radovskii, S. (1980) Application of the calculation scheme for a layered viscoelastic medium to the estimation of the stressed state of highways and airport pavements with moving loads. *Soviet Applied Mechanics*, **15,** 940-946.

[6] Pister, K.S. (1961) Viscoelastic plate on a viscoelastic foundation. *Journal of Engineering Mechanics*, **87,** 43-54.

[7] Robertson, S.R. (1971) Solving the problem of forced motion of viscoelastic plates by Valanis' method with an application to a circular plate. *Journal of Sound and Vibration*, **l4,** 263-278.

[8] Hewitt, J.S. and Mazumdar, J. (1974) Vibration of viscoelastic plates under transverse load by the method of constant deflection contours. *Journal of Sound and Vibration*, **33,** 319-333.

[9] Sonoda, K., Ishio, T. and Kobayashi, H. (1978) Circular plates on linear viscoelastic foundations. *Journal of Engineering Mechanics*, **104,** 819-828.

[10] Sonoda, K. and Kobayashi, H. (1980) Rectangular plates on linear viscoelastic foundations. *Journal of Engineering Mechanics*, **106,** 323-338.

[11] Lin, Y.J. (1978) Dynamic response of circular plates resting on viscoelastic half space. *Journal of Applied Mechanics*, **45,** 379-384.

[12] Yang, T.Q., Wang, R. and Yang, Z.W. (1991) Dynamic response of a viscoelastic circular plate on a viscoelastic half space foundation. *In: Zyczkowski M (ed.) Creep in Structures*, Berlin: Springer-Verlag, 685-692.

[13] Timoshenko, S. and Woinowsky-Krieger, S. (1959) Theory of Plates and Shells. New York: McGraw-Hill.

[14] Yang, T.Q., Lu, H.B. and Huang, Y.Y. (1987) Quasi-static bending of a thin viscoelastic plate on foundations. *Journal of Huazhong University of Science and Technology*, **15,** 1-6 (in Chinese).

Vol.1, No.2, 124-128 (2009)
doi:10.4236/ns.2009.12015

Natural Science

Studies of Uni-Univalent Ion Exchange Reactions Using Strongly Acidic Cation Exchange Resin Amberlite IR-120

Pravin Singare[1], Ram Lokhande[2], Neelima Samant[3]

[1]Department of Chemistry, Bhavan's College, Andheri, Mumbai, India
[2]Department of Chemistry, University of Mumbai, Vidyanagari, Santacruz, Mumbai, India
[3]Department of Chemistry, Sathaye College, Vile Parle, Mumbai, India

Received 15 June 2009; revised 22 July 2009; accepted 25 July 2009.

ABSTRACT

The selectivity behaviour of ion exchange resin Amberlite IR-120 for inorganic cations like sodium and potassium was predicted on the basis of thermodynamic data. The equilibrium constant *K* values calculated for uni-univalent ion exchange reaction systems were observed to increase with rise in temperature, indicating endothermic ion exchange reactions. From the *K* values calculated at different temperatures the enthalpy values were calculated. The low enthalpy and higher *K* values for K^+ ion exchange reaction indicates more affinity of the resin for potassium ions as compared to that for sodium ions also in the solution. The technique used in the present experimental work will be useful in understanding the selectivity behaviour of different ion exchange resins for ions in the solution. Although the ionic selectivity data for the ion exchange resins is readily available in the literature, it is expected that the information obtained from the actual experimental trials will be more helpful. The technique used in the present experimental work when applied to different ion exchange resins will help in there characterization.

Keywords: Ion Exchange Equilibrium; Equilibrium Constant; Enthalpy; Endothermic Reaction; Amberlite IR-120

1. INTRODUCTION

There are number of liquid processes waste streams at chemical processing, nuclear power plants, nuclear fuel reprocessing plants and nuclear research centers that requires treatment for removal of various contaminants. One of the most common treatment methods for such aqueous streams is the use of ion exchange, which is a well developed technique that has been employed for many years in chemical as well as nuclear industries. While designing an ion exchange liquid waste processing system it is desirable to have an adequate knowledge about the distribution coefficient values and the selectivity behaviour of these ion exchange resin towards different ions present in liquid waste. Generally the selected ion exchange materials must be compatible with the chemical nature of the waste such as type and concentration of ionic species present as well as the operating parameters notably temperature. Considerable work was done by previous researchers to study the properties of the ion exchange resins, to generate thermodynamic data related to various uni-univalent and heterovalent ion exchange systems [1-6]. A number of researchers carried out equilibrium studies, extending over a wide range of composition of solution and resin phase [7-27]. Attempts were also made to study the temperature effect on anion exchange systems [10,22-29] for computing the thermodynamic equilibrium constants. However, very little work was carried out to study the equilibrium of cation exchange systems [7-21]. Therefore in the present investigation attempts were made to study the thermodynamics of uni-univalent cation exchange equilibrium, the results of which will be of considerable use in explaining the selectivity of ion exchanger for various univalent ions in solution.

2. EXPERIMENTAL

The ion exchange resin Amberlite IR-120 as supplied by the manufacturer (Rohm and Hass Co.,USA) was a strongly acidic cation exchange resins in H^+ form containing 8% S-DVB of 16-50 mesh size, having moisture content of ~45%. The maximum operating temperature is 120^0C, operating in the pH range of 0-14. The exchange capacity of the resin is 1.9 meq/mL on wetted bed volume basis. For present investigation, the resin grains of 30-40 mesh size were used. The conditioning of the resins was done by usual methods [23-26].

The ion exchange resins in H^+ form weighing 0.500g were equilibrated with Na^+ ion solution of different concentrations at a constant temperature of 30.0℃ for 3 h. From the results of kinetics study reported earlier [30-41]; it was observed that this duration was adequate to attain the ion exchange equilibrium. After 3 h the different Na^+ ion solutions in equilibrium with ion exchange resins were analyzed for their H^+ ion concentration by potentiometric titration with standard 0.1N NaOH solution. From the results ***K*** value for the reaction

$$R\text{-}H + Na^{+}_{(aq.)} \rightleftharpoons R\text{-}Na + H^{+}_{(aq.)} \quad \textbf{(1)}$$

was determined at 30.0 ^{0}C. Similar values of ***K*** for the above H^+ / Na^+ system was determined for different temperatures in the range of 30.0 ^{0}C to 40.0 ^{0}C.

The study was also carried out as explained above for H^+ / K^+ system in the same temperature range, to study the ***K*** values for the reaction

$$R\text{-}H + K^{+}_{(aq.)} \rightleftharpoons R\text{-}K + H^{+}_{(aq.)} \quad \textbf{(2)}$$

The sodium and potassium ion solutions used in the entire experimental work, where prepared by dissolving potassium and sodium chloride salts (Analytical grade) in distilled deionised water. In the present study, a semi-micro burette having an accuracy of 0.05 mL was used in the titrations and the titration readings were accurate to ± 0.05 mL. Considering the magnitude of the titer values, the average equilibrium constants reported in the experiment are accurate to ± 3 %.

3. RESULTS AND DISCUSSION

The equilibrium constants for the uni-univalent ion exchange reactions (1 and 2) would be given by the expression

$$K = \frac{C_{RX} \cdot C_H^+}{(A - C_{RX}) \cdot C_X^+} \quad \textbf{(3)}$$

here A is the ion exchange capacity of the resin, x^+ represents Na^+ or K^+ ions.

For different concentrations of x^+ ions in solution at a given temperature, ***K*** values was calculated from which average value of ***K*** for that set of experiment was calculated. Similar values of ***K*** were calculated for both H^+/Na^+ and H^+/K^+ systems for different temperatures **(Table 1)**.

Earlier researchers have expressed the concentration of ions in the solution in terms of molality and concentration of ions in resin in terms of mole fraction [21]. In view of above, the experimental results obtained in the present study have been substituted in the following equation by Bonner *et al.* [14, 18] and the equilibrium constant ***K*** ' was calculated **(Table 2)**.

$$K' = \frac{[N_x^+][m_H^+]}{[N_H^+][m_x^+]} \quad \textbf{(4)}$$

here N_X^+ = mole fraction of K^+ or Na^+ ions exchanged on the resin

m_H^+ = molality of H^+ ions exchanged in the solution

N_H^+ = mole fraction of H^+ ions remained on the resin

m_x^+ = molality of K^+ or Na^+ ions remained in the solution at equilibrium.

Since in the present study the solution was dilute, the molality and molarity of the ions in the solution were almost the same, with negligible error. Therefore the molality of the ions can be easily replaced by molarity. The equilibrium constant ***K***' was calculated by **Eq.4** and the average value of ***K***' is reported **(Table 2)**. Such ***K***' values were calculated for different temperatures and the values were in good agreement with ***K*** values calculated by **Eq.3 (Table 1 and Table 2)**.This justifies that the choice of units for the concentration in the present study is insignificant. The enthalpy value for the ion exchange reactions 1 and 2 were calculated by plotting the graph of log ***K*** against 1/T **(Figure 1)**. Bonner and Pruett [14] studied the temperature effect on uni–univalent exchanges involving some bivalent ions. In all bivalent exchanges the equilibrium constant decreases with rise in temperature resulting in exothermic reactions. However in the present investigation, the equilibrium constant values increases with rise in temperature **(Table 1 and Table 2)**, indicating the endothermic ion exchange reactions [23-26]. The low enthalpy and higher ***K*** values for H^+/ K^+ exchange as compared to that for H^+/Na^+ exchange **(Table 1 and Table 2)**, indicate that the resins in H^+ form are having more affinity for larger ionic size K^+ ions in solution as compared to that for Na^+ ions also in the solution.

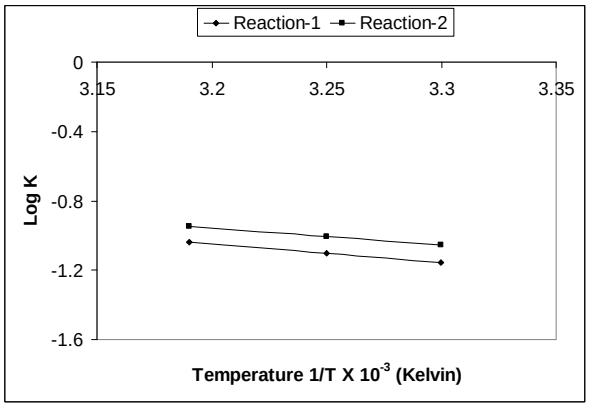


Figure 1. Variation of Equilibrium Constant with Temperature for the Ion Exchange Reactions (1) And (2) Using Ion Exchange Resin Amberlite IR-120. Amount of the ion exchange resin in H^+ form = 0.500g, Temperature Range = 30.0^{0}C – 40.0^{0}C.

Table 1. Equilibrium constant for the uni-univalent ion exchange reactions using ion exchange resin Amberlite IR-120.

Amount of the ion exchange resin in H^+ form = 0.500g, Ion exchange capacit y = 3.3 meq./g , Volume of external ionic solution= 80.0 mL

Reaction 1					*Reaction 2*				
Temperature =30.0 ^{0}C									
Initial conc. of Na^+ ion solution (M)	Final conc.of Na^+ ions (M) C_{Na^+}	Conc. of H^+ ions Exchanged in solution (M) C_{H^+}	Amount of Na^+ ions Exchanged on the resin meq./ 0.5g C_{RNa}	Equilibrium constant K	Initial Conc. of K^+ ion solution (M)	Final conc. of K^+ ions (M) C_{K^+}	Conc. of H^+ ions exchanged in solution (M) C_{H^+}	Amount of K^+ ions Exchanged on the resin meq./0.5 g C_{RK}	Equilibrium constant K
0.01	0.0052	0.0048	0.24	0.072	0.01	0.0048	0.0052	0.26	0.099
0.02	0.0126	0.0074	0.37	0.074	0.02	0.0120	0.0080	0.40	0.099
0.025	0.0166	0.0084	0.42	0.074	0.025	0.0160	0.0090	0.45	0.096
0.03	0.0208	0.0108	0.46	0.072	0.03	0.0200	0.0100	0.50	0.096
0.04	0.0292	0.0112	0.54	0.072	0.04	0.0282	0.0118	0.59	0.098
Average equilibrium constant (*K*) = 0.073 Standard deviation (s) = ± 0.001					Average equilibrium constant (*K*) = 0.098 Standard deviation (s) = ± 0.001				

Temperature ℃	30.0	35.0	40.0	Enthalpy (kJ/mol)	30.0	35.0	40.0	Enthalpy (kJ/mol)
Equilibrium Constant (*K*)	0.073	0.079	0.094	19.9	0.098	0.107	0.120	15.0

Table 2. Equilibrium constant for the uni-univalent ion exchange reactions using ion exchange resin Amberlite IR-120 calculated by Bonner et.al. Equation.

Amount of the ion exchange resin in H^+ form = 0.500g, Ion exchange capacity = 3.3meq./g, Volume of external ionic solution =80.0mL

Reaction 1					*Reaction 2*				
Temperature =30.0 ^{0}C									
Initial conc. of Na^+ ion solution (M)	Final conc. of H^+ ions (M) C_{Na^+}	Mole fraction of Na^+ ions exchanged on the resin N_{Na^+}	Mole fraction of H^+ ions remained on the resin N_{H^+}	Equilibrium constant K'	Initial Conc. of K^+ ion solution (M)	Final conc. of K^+ Ions (M) C_{K^+}	Mole fraction of K^+ ions Exchanged on the resin N_{K^+}	Mole fraction of H^+ ions Remained on theresin N_{H^+}	Equilibrium constant K'
0.01	0.0052	0.073	0.927	0.072	0.01	0.0048	0.079	0.921	0.093
0.02	0.0126	0.112	0.888	0.074	0.02	0.0120	0.012	0.879	0.092
0.025	0.0166	0.127	0.873	0.074	0.025	0.0160	0.136	0.864	0.089
0.03	0.0208	0.164	0.836	0.081	0.03	0.0200	0.152	0.849	0.089
0.04	0.0292	0.170	0.830	0.075	0.04	0.0282	0.179	0.821	0.091
Average equilibrium constant (*K*) = 0.075 Standard deviation (s) = ± 0.004					Average equilibrium constant (*K*) = 0.091 Standard deviation (s) = ± 0.002				

Temperature ℃	30.0	35.0	40.0	Enthalpy (kJ/mol)	30.0	35.0	40.0	Enthalpy (kJ/mol)
Equilibrium Constant (*K'*)	0.075	0.079	0.094	19.9	0.091	0.099	0.111	15.0

4. CONCLUSION

Efforts to develop new ion exchangers for specific applications are continuing. In spite of their advanced stage of development, various aspects of ion exchange technologies have been continuously studied to improve the efficiency and economy in various technical applications. The selection of an appropriate ion exchange material is possible on the basis of information provided by the manufacturer. However, it is expected that the data obtained from the actual experimental trials will prove to be more helpful. The thermodynamic data obtained in the present experimental work will be useful to understand the selectivity behaviour of ion exchange resins for various ions in solution thereby helping in characterization of resins.

REFERENCES

[1] Bhargava, A. and Janardanan, C. (1997) Ion exchange properties of bismuth antimonite. *Indian J. Chem.*, **36A(7),** 624-625.

[2] Muraviev, D., Gonzalo, A. and Valiente, M. (1995) Ion exchange on resins with temperature-responsive selectivity. 1. Ion-Exchange Equilibrium of Cu^{2+} and Zn^{2+} on Iminodiacetic and Aminomethylphosphonic Resins. *Anal. Chem.*, **67(17),** 3028-3035.

[3] Boyd, G.E., Vaslow, F. and Lindenbaum, S. (1967) Thermodynamic quantities in the exchange of zinc with sodium ions in variously cross-linked polystyrene sulfonate cation exchangers at 25. degree. *J. Phys. Chem.*, **71(7),** 2214-2219.

[4] Duncan, J.F. (1955) Enthalpies and entropies of ion-exchange reactions. *Australian Journal of Chemistry*, **8(1),** 1-20.

[5] Boyd, G.E., Vaslow, F. and Lindenbaum, S. (1964) Calorimetric determinations of the heats of ion-exchange reactions. I. Heats of exchange of the alkali metal cations in variously cross-linked polystyrene sulfonates. *J. Phys. Chem.*, **68(3),** 590-597.

[6] Schwarz, A. and Boyd, G.E. (1965) Thermodynamics of the exchange of tetramethylammonium with sodium ions in cross-linked polystyrene sulfonates at 25°. *J. Phys. Chem.*, **69(12),** 4268-4275.

[7] Myers, G.E. and Boyd, G.E. (1956) A thermodynamic calculation of cation exchange selectivities. *J. Phys. Chem.*, **60(5),** 521-529.

[8] Bonner, O.D. (1955) Ion-exchange equilibria involving rubidium, cesium and thallous ions. *J. Phys. Chem.*, **59(8),** 719-721.

[9] Bonner, O.D. (1954) A selectivity scale for some monovalent cations on Dowex 50. *J. Phys. Chem.*, **58(4),** 318-320.

[10] Lindenbaum, S., Jumper, C.F. and Boyd, G.E. (1959) Selectivity coefficient measurements with variable capacity cation and anion exchangers. *J. Phys. Chem.*, **63(11)**, 1924-1929.

[11] Kraus, K.A. and Raridon, R.J. (1959) Temperature dependence of some cation exchange equilibria in the range 0 to 200°C . *J. Phys. Chem.*, **63(11),** 1901-1907.

[12] Bonner, O.D. and Payne, W.H. (1954) Equilibrium studies of some monovalent ions on Dowex 50. *J. Phys. Chem.*, **58(2),** 183-185.

[13] Argersinger, W.J. and Davidson, A.W. (1952) Experimental factors and activity coefficients in ion exchange equilibria. *J. Phys. Chem.*, **56(1)**, 92-96.

[14] Bonner, O.D. and Pruett, R.R. (1959) The effect of temperature on ion exchange equilibria. III. Exchanges Involving Some Divalent Ions, *J. Phys. Chem.*, **63(9)**, 1420-1423.

[15] Bonner, O.D. and Livingston, F.L. (1956) Cation-exchange equilibria involving some divalent ions. *J. Phys. Chem.*, **60(5)**, 530-532.

[16] Bonner, O.D. and Smith, L.L. (1957) A selectivity scale for some divalent cations on Dowex 50. *J. Phys. Chem.*, **61(3)**, 326-329.

[17] Bonner, O.D., Jumper, C.F. and Rogers, O.C. (1958) Some cation-exchange equilibria on Dowex 50 at 25°. *J. Phys. Chem.*, **62(2)**, 250-252.

[18] Bonner O.D. and Smith L.L. (1957) The effect of temperature on ion-exchange equilibria. I. The Sodium–Hydrogen and Cupric–Hydrogen Exchanges, *J. Phys. Chem.*, **61(12)**, 1614-1617.

[19] Kielland J. (1935) Thermodynamics of base-exchange equilibria of some different kinds of clays. *J. Soc. Chem. Ind.*, (London) **54,** 232-234.

[20] Vanselow A.P. (1932) The utilization of the base-exchange reaction for the determination of activity coefficients in mixed electrolytes. *J. Am. Chem. Soc.*, **54(4)**, 1307-1311.

[21] Gaines, G.L. (Jr.) and Thomas, H.C. (1953) Adsorption studies on clay minerals. II. A Formulation of the Thermodynamics of Exchange Adsorption. *J. Chem. Phys.*, **21(4),** 714-718.

[22] Kraus, K.A., Raridon, R.J. and Holcomb, D.L. (1960) Anion exchange studies: A column method for measurement of ion exchange equilibria at high temperature. Temperature coefficient of the Br^--Cl^- exchange reaction, *Chromatogr. J.*, **3(1)**,178-179.

[23] Lokhande, R.S., Singare, P.U. and Patil, A.B. (2007) A study of ion exchange equilibrium for some uni-univalent and uni-divalent reaction systems using strongly basic anion exchange resin Indion-830 (Type 1). *Russ. J. Phys. Chem. A*, **81(12),** 2059-2063.

[24] Singare, P.U., Lokhande, R.S. and Prabhavalkar, T.S. (2008) Thermodynamics of ion exchange equilibrium for some uni-univalent and divalent reaction systems using strongly basic anion exchange resin Indion FF-IP. *Bull. Chem. Soc. Ethiop.*, **22(3),** 415-421.

[25] Lokhande R.S. and Singare P.U. (2007) Study on ion exchange equilibrium for some uni-univalent reaction systems using strongly basic anion exchange resin Amberlite IRA-400. *J. Ind. Council Chem.*, **24(2),** 73-77.

[26] Lokhande R.S., Singare P.U. and Kolte A.R. (2008) Ion exchange equilibrium for some uni-univalent and uni-divalent reaction systems using strongly basic anion exchange resin Duolite A-102 D. *Bull. Chem. Soc. Ethiop.*, **22(1)**, 107-114.

[27] Heumann, K.G. and Baier, K. (1982) Chloride distribution coefficient on strongly basic anion-exchange resin: Dependence on co-ion in alkali fluoride solutions.

Chromatographia, **15(11),** 701-703.

[28] Lokhande, R.S., Singare, P.U. and Karthikeyan, P. (2008) Study on ion exchange equilibrium using strongly basic anion exchange resin Duolite A-162. *J. Ind. Council Chem.*, **25(2),** 117-121.

[29] Singare, P.U., Lokhande, R.S., Kolte, A.R., Dole, M.H., Karthikeyan, P. and Parab, S.A. (2008) Studies on ion exchange equilibrium using some anion exchange resins, Int. *J. Chem. Sci*., **6(4)**, 2172-2181.

[30] Lokhande R.S., Singare P.U. and Patil A.B. (2007) Application of radioactive tracer technique on Industrial-grade ion exchange resins Indion-830 (Type-1) and Indion-N-IP (Type-2). *Radiochim. Acta*, **95(1),** 111-114.

[31] Lokhande, R.S. and Singare, P.U. (2007) Comparative study on ion-isotopic exchange reaction kinetics by application of tracer technique, *Radiochim. Acta*, **95(3)**, 173-176.

[32] Lokhande, R.S., Singare, P.U. and Kolte, A.R. (2007) Study on kinetics and mechanism of ion-isotopic exchange reaction using strongly basic anion exchange resins Duolite A-101 D and Duolite A-102 D. *Radiochim. Acta*, **95(10)**, 595-600.

[33] Lokhande, R.S., Singare, P.U. and Dole, M.H. (2006) Comparative study on bromide and iodide ion-Isotopic exchange reactions using strongly basic anion exchange resin Duolite A-113, *J. Nucl. Radiochem. Sci.*, **7(2),** 29-32.

[34] Lokhande, R.S. and Singare, P.U. (2008) Comparative study on iodide and bromide ion-isotopic exchange reactions by application of radioactive tracer technique, *J .Porous Mater*, **15(3),** 253-258.

[35] Lokhande, R.S., Singare, P.U. and Karthikeyan, P. (2007) The kinetics and mechanism of bromide ion isotope exchange reaction in strongly basic anion-exchange resin Duolite A-162 determined by the radioactive tracer technique, *Russ. J. Phys. Chem. A*, **81(11),** 1768-1773.

[36] Lokhande, R.S., Singare, P.U. and Dole, M.H. (2007) Application of radiotracer technique to study the ion isotope exchange reactions using a strongly basic anion-exchange resin Duolite A-113. *Radiochemistry*, **49(5)**, 519-522.

[37] Singare, P.U., Lokhande, R.S. and Patil, A.B. (2008) Application of radioactive tracer technique for characterization of some strongly basic anion exchange resins, *Radiochim. Acta*, **96(2),** 99-104.

[38] Lokhande, R.S., Singare, P.U. and Tiwari, S.R.D. (2008) Study of bromide ion-isotopic exchange reaction kinetics using a weakly basic macro porous resin Indion-860. *Radiochemistry*, **50(6),** 633-637.

[39] Lokhande, R.S., Singare, P.U. and Prabhavalkar, T.S. (2008) The application of the radioactive tracer technique to study the kinetics of bromide isotope exchange reaction with the participation of strongly basic anion exchange resin Indion FF-IP. Russ. *J. Phys. Chem. A*, **82(9)**, 1589-1595.

[40] Lokhande, R.S., Singare, P.U. and Parab, S.A. (2008) Application of radioactive tracer technique to study the kinetics of iodide ion-isotopic exchange reaction using strongly basic anion exchange resin Duolite A-116, *Radiochemistry*, **50(6),** 642-644.

[41] Lokhande, R.S., Singare, P.U. and Patil, V.V. (2008) Application of radioactive tracer technique to study the kinetics and mechanism of reversible ion-isotopic exchange reaction using strongly basic anion exchange resin Indion-850, *Radiochemistry*, **50(6),** 638-641.

Vol.1, No.2, 129-135 (2009)
doi:10.4236/ns.2009.12016

Natural Science

ZnO Nanoparticles: Synthesis and Adsorption Study

K. Prasad[1], Anal K. Jha[2]

[1]University Department of Physics, T.M. Bhagalpur University, Bhagalpur - 812 007, India; *k.prasad65@gmail.com
[2]University Department of Chemistry, T.M. Bhagalpur University, Bhagalpur - 812 007, India

Received 21 July 2009; revised 27 July 2009; accepted 30 July 2009.

ABSTRACT

A low-cost, green and reproducible probiotic microbe (*Lactobacillus sporogens*) mediated biosynthesis of ZnO nanoparticles is reported. The synthesis is performed akin to room temperature in five replicate samples. X-ray and transmission electron microscopy analyses are performed to ascertain the formation of ZnO nanoparticles. Rietveld analysis to the X-ray data indicated that ZnO nanoparticles have hexagonal unit cell structure. Individual nanoparticles having the size of 5-15 nm are found. A possible involved mechanism for the synthesis of ZnO nanoparticles has been proposed. The H_2S adsorption characteristic of ZnO nanoparticles has also been assayed.

Keywords: *ZnO Nanoparticle; Biosynthesis; Nanobiotechnology; Eco-friendly; H_2S Adsorption*

1. INTRODUCTION

Nature by dint of its diversity provides exponential possibilities in terms of endearing adaptability of its constituent cohorts. Both bacteria and fungi make such an exciting category of microorganisms having naturally bestowed property of reducing/oxidizing metal ions into metallic/oxide nanoparticles thereby functioning as 'mini' nano-factories. [1,2] It is indeed their chemical constitutions (or metabolic status) which provide them strength to withstand such environmentally diverse habitats. The non-pathogenic, gram positive, mesophilic facultative anaerobe *Lactobacillus*, commonly used for curdling of milk forms part of the beneficial community of microbes present in the human intestinal tract.

Zinc oxide (ZnO) is considered to be a technologically prodigious material having a wide spectrum of applications such as that of a semiconductor (E_g = 3.37 eV), magnetic material, electroluminescent material, UV-absorber, piezoelectric sensor and actuator, nanostructure varistor, field emission displaying material, thermoelectric material, gas sensor, constituent of cosmetics etc. [3-9] There are several synthesis procedures for the preparation of ultrafine oxide nanoparticles such as sol-gel, hydrothermal, solvothermal, flame combustion, emulsion precipitation, fungus mediated biosynthesis, *etc*. [10-16] Each method has its own merits and demerits. They are time consuming, capital intensive and require trained manpower. Besides, the development of eco-friendly, 'green' synthesis protocols is in line with the recent RoHS and WEEE legislation stipulated by the EU. Therefore, an urge to develop green synthesis protocols which goes in consonance with the above mentioned stipulations is need of the hour. Microbes exhibit a natural capability to adapt to changes in their environment. Recent research devoted towards the study of interaction between inorganic substances and biological systems has highlighted its potential application for the production of nanomaterials with interesting technological properties. [1,2,17-20] Numerous recent publications have highlighted the potential for microbes, particularly bacteria (including thermophilies) and fungi, to synthesize metallic and/or oxide nanoparticles. [21-35] The facultative nature of *Lactobacilli*, offers the potential to produce nanoparticles under both oxidizing and reducing conditions. [2,18,35]

No work to the best of author's acquaintance has so far been reported regarding the synthesis of ZnO nanoparticles employing *Lactobacilli*. *Lactobacilli* strain, cultured from spores an effort has been taken for synthesizing ZnO nanoparticles (ZnO NPs) in the present work. We have tried to explore a cost effective, green and readily reproducible approach for the purpose of scaling up and subsequent downstream processing. An effort to understand the nano-transformation mechanism of biosynthesis has also been made. It is well established that Hydrogen sulfide (H_2S) is a colorless, corrosive and highly toxic gas, a low concentration of which in air, brings smell of rotten eggs and it substantially contributes towards air pollution. [36,37] The potential of ZnO NPs towards H_2S adsorption has also been assayed in the present study.

2. MATERIALS AND METHODS

2.1. Biosynthesis of ZnO Nanoparticles

Pharmaceutical grade Lactic acid Bacillus spore tablets (SporeLac DS, Sanyko Pharmaceuticals, Japan) were procured and two tablets were dissolved in 50 mL sterile distilled water containing standard carbon and nitrogen source. As per specification, each tablet was capable of producing 120 million spores of the bacterium. The culture solution was allowed to incubate on room temperature overnight. Next day, the presence of *Lactobacillus* was confirmed under an optical microscope. The pH of this source culture solution was observed to be equal to 3. Now, 10 mL of this source culture was doubled in volume by mixing equal volume of sterile distilled water containing nutrients in five different hard glass test tubes. In yet another tube instead of adding the source culture solution, sterile distilled water containing nutrients was pooled and this was treated as control. All these culture tubes were gently heated on a steam bath and were allowed to incubate overnight in laboratory ambience for another 24 hours on orbital shaker. Next day, the pH was taken and found to be in the range of 4-5 in case of culture solution and 7 in case of control. Small quantity of $NaHCO_3$ was added in culture solution until it attains pH 6. It was brought to this pH as a lower value delays the process of transformation. [2] Similarly, a small volume of distilled water along with carbon and nitrogen source and $NaHCO_3$ were pipetted in the control tube and the pH = 8.5 were recorded. Analytical reagent grade Zinc Chloride ($ZnCl_2$) was taken into use for preparing a solution of 0.25(M) strength at room temperature. Control solution was prepared by adding 100 mL sterile distilled water, carbon and nitrogen containing nutrients and the mild base in known quantitative ratio (5:1:1). To each of these tubes, 20 mL of Zinc Chloride solution was added. The pH of the control tube was noted to be 8-9 in 5 different set of experiments. Culture solution containing tubes including control tube were heated on the steam bath up to 80°C for 5 to 10 minutes. An appearance of starch like haziness in solution and white deposition at the bottom of the tube was perceived as an indication of commencement of transformation. No such deposition or haziness was observed in control tube. The tubes were allowed to incubate in the laboratory ambience for another 9 hours, after which distinctly markable coalescent white clusters deposited at the bottom of all the tubes except in control. A remarkable change in pH was observed at this stage (6.0 to 7.5) excluding control (8 to 9). The chemical reactions which proceed in the culture medium may be as follows:

$$\underset{\text{(Glucose)}}{C_6H_{12}O_6} \rightarrow \underset{\text{(Pyruvate)}}{CH_3-\overset{\overset{O}{\|}}{C}-COOH} \Leftrightarrow \underset{\text{(Lactate)}}{CH_3.CH(OH).COOH}$$

$$NaHCO_3 \Leftrightarrow Na^+ + HCO_3^-$$

$$HCO_3^- \Leftrightarrow OH^- + CO_2$$

$$ZnCl_2 \Leftrightarrow Zn^{2+} + 2Cl^-$$

$$Zn^{2+} + 2OH^- \rightarrow Zn(OH)_2$$

$$Zn(OH)_2 \rightarrow ZnO + H_2O$$

2.2. Characterization

The formation of ZnO NPs was checked by X-ray diffraction (XRD) technique using an X-ray diffractometer (XPERT-PRO, Pan Analytical) with CuK_α radiation (λ = 1.5406Å) over a wide range of Bragg angles ($10° \le 2\theta \le 80°$). The *XY* (2θ *vs*. intensity) data obtained from this experiment were plotted with the WinPLOTR program and the angular positions of the peaks were obtained with the same program. [38] The dimensions of the unit cell, hkl values and space group of ZnO NPs were obtained using the DICVOL program in the FullProf 2000 software package and then refinement was carried out through the profile matching routine of FullProf. [39] The Bragg peaks were modeled with pseudo-Voigt function and the background was estimated by linear interpolation between selected background points. The crystallite size (*D*) and the lattice strain of ZnO NPs were estimated by analyzing the broadening of X-ray diffraction peaks, using Williamson-Hall approach. [40]

$$\eta \cos\theta = (K\lambda / D) + 2(\Delta\xi / \xi)\sin\theta \qquad \textbf{(1)}$$

where η is diffraction peak width at half intensity (FWHM) and $\Delta\xi/\xi$ is the lattice strain and *K* is the Scherrer constant (0.89). The term $K\lambda/D$ represents the Scherrer particle size distribution. TEM micrograph of ZnO NPs was obtained using Hitachi H-7500 transmission electron microscope. The specimen was suspended in distilled water, dispersed ultrasonically to separate individual particles, and two drops of the suspension deposited onto holey-carbon coated copper grids.

3. RESULTS

3.1. Structural and Microstructural Studies

Rietveld refinements on the X-ray (XRD) data were done on ZnO NPs, selecting the space group *P*6/*mmm*. **Figure 1** depicts the observed, calculated and difference XRD profiles for ZnO NPs after final cycle of refinement. It can be seen that the profiles for observed and calculated one are perfectly matching. The value of χ^2 comes out to be equal to 3.16, which may be considered to be very good for estimations. The profile fitting procedure adopted was minimizing the χ^2 function. [41] The XRD analyses indicated that ZnO NPs has a hexagonal unit cell. The crystal data and refinement factors of ZnO

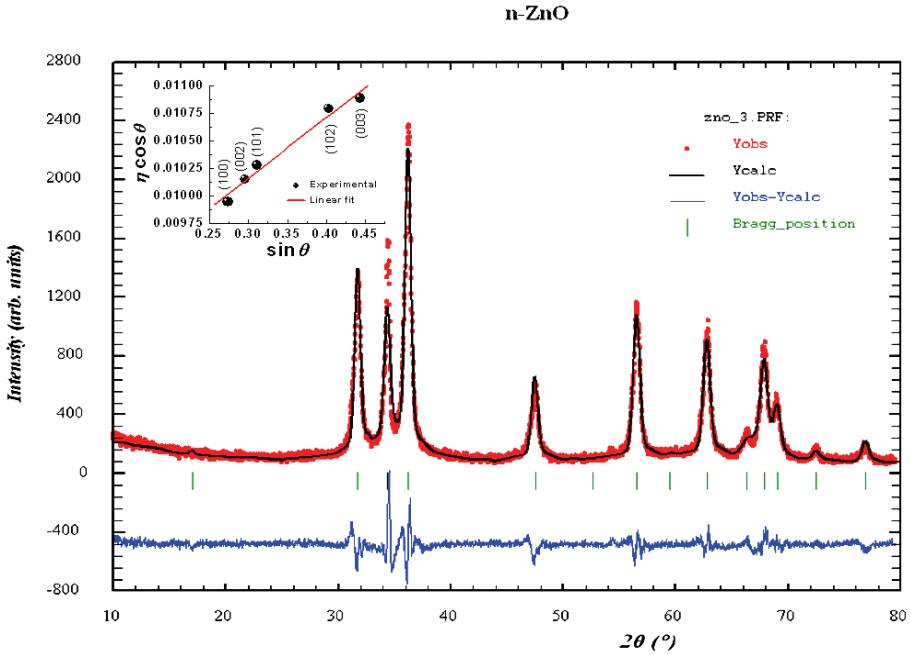


Figure 1. Rietveld refined pattern of ZnO NPs in the space group *P*6/*mmm*. Symbols represent the observed data points and the solid lines their Rietveld fit. Inset: Williamson-Hall plot for ZnO NPs.

Table 1. The crystal data and refinement factors of ZnO NPs obtained from X-ray powder diffraction data.

Parameters	Results	Description of parameters
Crystal System	Hexagonal	R_p (profile factor) = $100[\Sigma\|y_i-y_{ic}\|/\Sigma\|y_i\|]$, where y_i is the observed intensity and y_{ic} is the calculated intensity at the i^{th} step.
Space group	*P*6/*mmm*	
a (Å)	3.2524	R_{wp} (weighted profile factor) = $100[\Sigma\omega_i\|y_i-y_{ic}\|^2/\Sigma\omega_i(y_i)^2]^{1/2}$, where $\omega_i = 1/\sigma_i^2$ and σ_i^2 is variance of the observation.
b (Å)	3.2524	
c (Å)	5.2120	
α (°)	90.000	R_{exp} (expected weighted profile factor) = $100[(n-p)/\Sigma\omega_i(y_i)^2]^{1/2}$, where *n* and *p* are the number of profile points and refined parameters, respectively.
β (°)	90.000	
γ (°)	120.000	
V (Å^3)	47.7463	R_B (Bragg factor) = $100[\Sigma\|I_{obs}-I_{calc}\|/\Sigma\|I_{obs}\|]$, where I_{obs} is the observed integrated intensity and I_{calc} is the calculated integrated intensity.
R_p	25.2	
R_{wp}	23.8	R_F (crystallographic R_F factor) = $100[\Sigma\|F_{obs}-F_{calc}\|/\Sigma\|F_{obs}\|]$, where *F* is the structure factor, $F = \sqrt{(I/L)}$, where *L* is Lorentz polarization factor.
R_{exp}	13.4	
R_B	0.175E-3	$\chi^2 = \Sigma\omega_i(y_i-y_{ic})^2$.
R_F	0.133E-3	
χ^2	3.16	*d* (Durbin–Watson statistics) = $\Sigma\{[\omega_i(y_i-y_{ic})-\omega_{i-1}(y_{i-1}-y_{ic-1})]^2\}/\Sigma[\omega_i(y_i-y_{ic})]^2$.
d	0.6844	Q_D = expected *d*.
Q_D	1.9059	
S	1.776	*S* (goodness of fit) = (R_{wp}/R_{exp}).

NPs obtained from XRD data are depicted in **Table 1**. The lattice parameter as obtained for ZnO NPs is in good agreement with the literature report (PCPDF No. #89-0510). Inset **Figure 1** illustrates the Williamson-Hall plot for ZnO NPs. A linear least square fitting to $\eta\cos\theta$–$\sin\theta$ data yielded the values of average crystallite size and lattice strain respectively to be 11 nm and 0.0035. The low value of lattice strain might be due to the fact that the procedure adopted in the synthesis of nanoparticles is natural (biosynthetic) one.

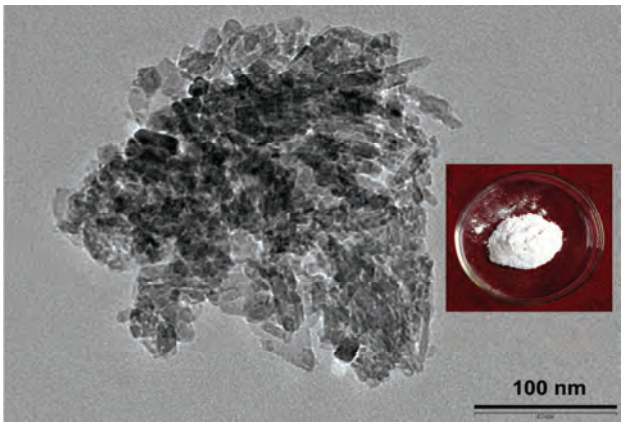


Figure 2. TEM photograph of ZnO NPs. Inset: ZnO NPs.

Figure 2 shows the TEM micrograph of ZnO NPs (inset **Figure** 2) being formed using *Lactobacillus* strain. The micrograph clearly illustrates the nanoparticles with tubules and other irregular forms having the sizes of 5-15 nm. The measurement of size was carried along the diameter of the particles. The difference in particle size is possibly due to the fact that the nanoparticles are being formed at different times. It is found that the size of the ZnO NPs estimated using TEM analysis to be in fairly good agreement with the size estimated by the Williamson-Hall approach.

3.2. Adsorption Study

Figure 3 shows the experimental setup to assess the adsorption capacity of synthesized ZnO NPs as well as bulk ZnO. Freshly prepared H_2S was allowed to pass through the equal quantities of bulk ZnO and ZnO NPs (5 gm each) for a fixed span of time (30 min.) and flow of gas was suitably regulated. The degree of absorption was assessed directly through the change in colour of lead acetate solution (from clear solution to black). It was observed that the presence of bulk ZnO blackens the solution (due to formation of lead sulfide) within 5 minutes, while ZnO NPs does the same in 25 min. The experiment was pursued as five replicates and each gave approximately the same result. This happens due to the fact that nanoparticles have large surface to volume ratio and hence a high surface activity and these features might have led to better degree of adsorption of H_2S in comparison to its bulk counterpart. H_2S absorption by ZnO proceeds according to the reaction: $ZnO + H_2S \rightarrow ZnS + H_2O$ that results into formation of inert Zinc sulfide.

4. DISCUSSION

Lactobacilli cells are prokaryotes in terms of cellular organization. They are gram positive (a thick peptidoglycan cell wall) bacteria showing facultative anaerobic properties, which probably make them suitable candidate microorganism for biosynthesis of metal as well as oxide nanoparticle. Like most of the bacteria, they have a negative electro-kinetic potential; which readily attracts the cations and this step probably acts as a trigger of the procedure of biosynthesis. Earlier, such a possibility of biosorption and bioreduction had been reported in case of silver iodide by the *Lactobacillus* sp. A09*. [42] The mesophilic, non-pathogenic and facultatively anaerobic microbe like *Lactobacillus* has robust metabolic capabilities. Addition of simple carbohydrates into the culture medium tends to lower the value of oxidation-reduction potential (or the Eh value). The oxidation-reduction potential expresses the quantitative character of degree of aerobiosis having a designated unit expressed as rH_2 (the negative logarithm of the partial pressure of gaseous hydrogen). By controlling rH_2 of the nutrient medium, conditions can be engineered for the growth of anaerobes in the presence of oxygen by lowering the rH_2, and also by cultivating the aerobes in anaerobic conditions by increasing the rH_2 of the medium.

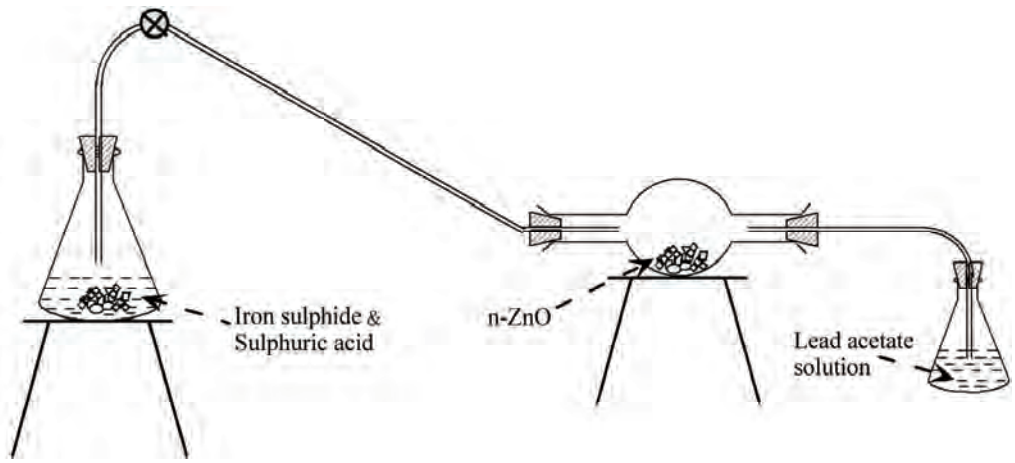


Figure 3. Experimental set up to study adsorption of H_2S by ZnO NPs.

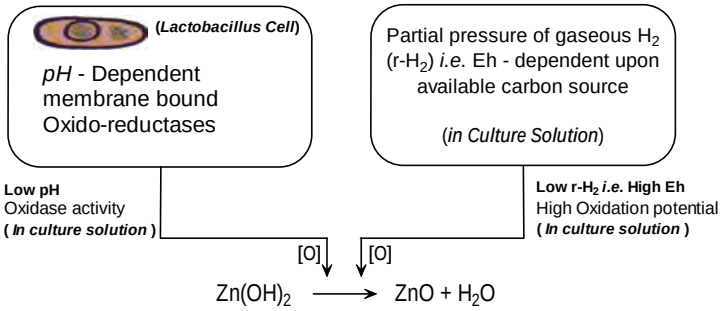


Figure 4. Schematic showing the mechanism for the biosynthesis of ZnO NPs.

Composition of nutrient media, therefore; plays a pivotal role in biosynthesis of metallic and/or oxide nanoparticles which is done in the present investigation. Energy yielding material – suitable carbohydrate (which controls the value of rH_2), the ionic status of the medium pH and overall oxidation-reduction potential (Eh) of the culture medium, all these factors cumulatively negotiate the synthesis of ZnO nanoparticles in the presence of *Lactobacillus* strain. Taking use of the above mentioned facts, our group had earlier reported synthesis of metallic cadmium [18], silver [17,34,43] as well as antimony oxide [1,44] and titanium dioxide. [2] A mildly acidic pH also activates the membrane bound oxidoreductases and makes the requisite ambience for an oxide nanoparticle synthesis as illustrated in **Figure 4**. Therefore, compared to other techniques, the present procedure is less expensive more reproducible, emphatically non-toxic and a truly green approach.

A large quantity of hydrogen sulfide is liberated in gas and petroleum industries and has been considered as a major pollutant. Besides, according to the international environmental regulations, H_2S contained in the acid gases should be effectively removed before its release to atmosphere. Its characteristic odor could easily be perceived in a dilution of 0.002 mg /L in ambient air. Intake of higher concentrations, could lead to the collapse from respiratory failure. For the purpose of protection, the concentration should be reduced to less than 15 ppm. [45] Pollution of underground aquifers has been a prevalent problem in the areas adjoining oil and gas reserves, which miserably affects the health of nearby inhabitants. Use of ZnO NPs produced using present green and low cost protocol based devices could prove to be an effective step towards mitigation of the menace.

5. CONCLUSIONS

The present biosynthesis method is a green low cost approach, capable of producing ZnO NPs nearby room temperature. The synthesis of ZnO NPs might have resulted due to variation in the level of rH_2 or pH, which activates the pH sensitive oxido-reductases. ZnO nanoparticles could be effective in controlling the pollution generated due to H_2S in air as well as underground aquifers. However, bright possibility exists with regard to the development of different products/devices in order to get rid of the menace of different forms of air and/or water pollution such as face masks, water filters, de-odorizing cakes and screens. Cosmetic industries can bank upon this product in order to synthesize sunscreen lotions etc., which would be done in the immediate future.

REFERENCES

[1] Jha, A.K., Prasad, K. and Prasad, K. (2009) A green low-cost biosynthesis of Sb_2O_3 nanoparticles. *Biochem Engg J*, **43,** 303-306.

[2] Jha, A.K., Prasad, K. and Kulkarni, A.R. (2009) Synthesis of TiO_2 nanoparticles using microorganisms. *Colloid Surf B: Bioint*, **71,** 226–229.

[3] Park, S., Lee, J.–H., Kim, H.–S., Park, H.–J. and Lee, J. C. (2009) Effects of ZnO nanopowder dispersion on photocatalytic reactions for the removal of Ag^+ ions from aqueous solution. *J Electroceram*, **22,** 105–109.

[4] Wang, Z.L. (2008) Energy harvesting for self-powered nanosystems. *Nano Res*, **1,** 1-8.

[5] Botello-Méndez, A.R., López-Urías, F., Terrones, M. and Terrones, H. (2008) Enhanced ferromagnetism in ZnO nanoribbons and clusters passivated with sulfur. *Nano Res*, **1,** 420-426.

[6] Grigorjeva, L., Millers, D., Grabis, J., Monty, C., Kalinko, A., Smits, K., Pankratov, V. and Lojkowski, W. (2008) Luminescence properties of ZnO nanocrystals and ceramics. *IEEE Trans Nucl Sci*, **55,** 1551-1555.

[7] Daneshvar, N., Aber, S., Seyed Dorraji, M.S., Khataee, A.R. and Rasoulifard, M.H. (2008) Preparation and investigation of photocatalytic properties of ZnO nanocrystals: effect of operational parameters and kinetic study. *Int J Chem Biomol Engg*, **1,** 24-29.

[8] Lee, C.–Y., Haung, Y.–T., Su, W.–F. and Lin, C.-F. (2006) Electroluminescence from ZnO nanoparticles/organic nanocomposites. *Appl Phys Lett*, **89**, 231116-231118.

[9] Tong, Y.H., Liu, Y.C., Lu, S.X., Dong, L., Chen, S.J. and Xiao, Z.Y. (2004) The optical properties of ZnO nanoparticles capped with polyvinyl butyral. *J Sol-Gel Sci Tech*, **30**, 157-61.

[10] Moghaddam, A.B., Nazari, T., Badraghi, J. and Kazemzad, M. (2009) Synthesis of ZnO nanoparticles and electrodeposition of polypyrrole/ZnO nanocomposite film. *Int J Electrochem Sci*, **4,** 247–257.
[11] Shokuhfar, T., Vaezi, M.R., Sadrnezhad, S.K. and Shokuhfar, A. (2008) Synthesis of zinc oxide nanopowder and nanolayer via chemical processing. *Int J Nanomanufacturing*, **2,** 149-162.
[12] Kim, S.–J. and Park, D.-W. (2007) Synthesis of ZnO nanopowder by thermal plasma and characterization of photocatalytic property. *Appl Chem*, **11,** 377-380.
[13] Vaezi, M.R. and Sadrnezhaad, S. (2007) Nanopowder synthesis of zinc oxide via solochemical processing. *Mater Design*, **28,** 515–519.
[14] Ge, M.Y., Wu, H.P., Niu, L., Liu, J.F., Chen, S.Y., Shen, P.Y., Zeng, Y.W., Wang, Y.W., Zhang, G.Q. and Jiang, J.Z. (2007) Nanostructured ZnO: from monodisperse nanoparticles to nanorods. *J Cryst Growth*, **305,** 162–166.
[15] Hambrock, J., Rabe, S., Merz, K., Birkner, A., Wohlfart, A., Fischer, R.A. and Driess, M. (2003) Low-temperature approach to high surface ZnO nanopowders and a non-aqueous synthesis of ZnO colloids using the single-source precursor $[MeZnOSiMe_3]_4$ and related zinc siloxides. *J Mater Chem*, **13,** 1731–1736.
[16] Kwon, Y.J., Kim, K.H., Lim, C.S. and Shim, K.B. (2002) Characterization of ZnO nanopowders synthesized by the polymerized complex method via an organochemical route. *J Ceram Process Res*, **3,** 146-149.
[17] Prasad, K., Jha, A.K. and Kulkarni, A.R. (2008) Yeast mediated synthesis of silver nanoparticles. *Int J Nanosci Nanotech*, in press.
[18] Prasad, K., Jha, A.K. and Kulkarni, A.R. (2007) Microbe mediated nano transformation: cadmium. *NANO: Brief Rep Rev*, **2,** 239-242.
[19] Shahverdi, A.R., Minaeian, S., Shahverdi, H.R., Jamalifar, H. and Nohi, A.-A. (2007) Rapid synthesis of silver nanoparticles using culture supernatants of entrobacteria: a novel biological approach. *Process Biochem*, **42,** 919-923.
[20] Husseiny, M.I., Abd El-Aziz, M., Badr, Y. and Mahmoud, M.A. (2007) Biosynthesis of gold nanoparticles using *Pseudomonas aeruginosa*. *Spectrochim Acta Part A*, **67,** 1003-1006.
[21] Klaus, T., Joerger, R., Olsson, E. and Granqvist, C.G. (2001) Bacteria as workers in the living factory: metal accumulating bacteria and their potential for materials science. *Trends Biotechnol*, **19**, 15-20.
[22] Gericke, M. and Pinches, A. (2006) Biological synthesis of metal nanoparticles. *Hydrometallurgy*, **83**, 132-140.
[23] Senapati, S., Ahmad, A., Khan, M.I., Sastry, M. and Kumar, R. (2005) Extracellular biosynthesis of bimetallic Au-Ag alloy nanoparticles. *Small*, **1**, 517-520.
[24] Vigneshwaran, N., Ashtaputre, N.M., Varadarajan, P.V., Nachane, R.P., Paralizar, K.M. and Balasubramanya, R.H. (2007) Biological synthesis of silver nanoparticles using the fungus *Aspergillus flavus*. *Mater Lett*, **61,** 1413-1418.
[25] Mohanpuria, P., Rana, N.K. and Yadav, S.K. (2008) Biosynthesis of nanoparticles: technological concepts and future applications. *J Nanopart Res*, **10**, 507-517.
[26] Mandal, D., Bolander, M.E., Mukhopadhyay, D., Sarkar, G. and Mukherjee, P. (2006) The use of microorganisms for the formation of metal nanoparticles and their application. *Appl Microbiol Biotechnol*, **69**, 485-492.
[27] Bansal, V., Rautaray, D., Barred, A., Ahire, K., Sanyal, A. and Ahmad, A. (2005) Fungus-mediated biosynthesis of silica and titania particles. *J Mater Chem*, **15**, 2583-2589.
[28] Sadowski, Z., Maliszewska, I.H., Grochowalska, B., Polowczyk, I. and Koźlecki, T. (2008) Synthesis of silver nanoparticles using microorganisms. *Mater Sci-Poland*, **26**, 419-424.
[29] Joerger, R., Klaus, T. and Granqvist, C.G. (2001) Biologically produced silver-carbon composite materials for optically functional thin-film coating. *Adv Mater*, **12**, 407-409.
[30] Mukherjee, P., Ahmad, A., Mandal, D., Senapati, S., Sainkar, S.R., Khan, M.I., Parischa, R., Ajaykumar, P.V., Alam, M., Kumar, R. and Sastry, M. (2001) Fungus–mediated synthesis of silver nanoparticles and their immobilization in the mycelial matrix: A novel biological approach to nanoparticle synthesis. *Nano Lett*, **1**, 515-519.
[31] Ankamwar, B., Damle, C., Absar, A. and Sastry, M. (2005) Biosynthesis of gold and silver nanoparticles using *Emblica Officinalis* fruit extract, their phase transfer and trans-metallation in an organic solution. *J Nanosci Nanotechnol*, **10**, 1665-1671.
[32] Armendariz, V., Herrera, I., Peralta-Videa, J.R., Jose-Yacaman, M., Toroiani, H., Santiago, P. and Gardea-Torresdey, J.L. (2004) Size controlled gold nanoparticles formation by *Avena sativa* biomass: use of plants in nanobiotechnology. *J Nanopart Res*, **6**, 377-382.
[33] Shankar, S.S., Rai, A., Ahmad, A. and Sastry, M. (2004) Rapid synthesis of Au, Ag and bimetallic Au core-Ag shell nanoparticles using Neem (*Azadirachta indica*) leaf broth. *J Colloid Interface Sci*, **275**, 496-502.
[34] Jha, A. K., Prasad, K., Kumar, V. and Prasad, K. (2009) Biosynthesis of silver nanoparticles using *Eclipta* leaf. *Biotechnol Prog*, in print.
[35] Nair, B. and Pradeep, T. (2002) Coalescence of nanoclusters and formation of submicron crystallites assisted by *Lactobacillus* strains. *Cryst Growth Design*, **2**, 293-298.
[36] Haimour, N., El-Bishtawi, R. and Ail-Wahbi, A. (2005) Equilibrium adsorption of hydrogen sulfide onto CuO and ZnO. *Desalination*, **181**, 145-152.
[37] Duan, Z., Sun, R., Liu, R. and Zhu, C. (2007) Accurate thermodynamic model for the calculation of H_2S solubility in pure water and brines. *Energ Fuel*, **21**, 2056-2065.
[38] Roisnel, J. and Rodrıguez-Carvajal, J. (2000) WinPLOTR; laboratoire leon brillouin (CEA-CNRS) centre d'Etudes de saclay: gif sur yvette cedex. France.
[39] Rodriguez-Carvajal, J. (2000) FullProf: A Rietveld Refinement and Pattern Matching Analysis Program, (Version: April 2008). *Laboratoire Léon Brillouin (CEA-CNRS)*, France.
[40] Williamson, G.K. and Hall, W.H. (1953) X-ray line broadening from filed aluminum and wolfram. Acta Metall, **1,** 22-31.
[41] McCusker, L. B., Von Dreele, R. B., Cox, D. E., Louër, D. and Scardi, P. (1999) Rietveld refinement guidelines. J Appl Cryst, **32,** 36-50.
[42] Fu, J.K., Liu, Y.Y., Gu, P.Y., Tang, D.L., Lin, Z.Y., Yao, B.X. and Wen, S.Z. (2000) Spectroscopic characterization on the biosorption and bioreduction of Ag(I) by

Lactobacillus sp A09*. *Acta Physico-Chimica Sinica*, **16**, 779-782.

[43] Jha, A.K., Prasad, K., Prasad, K. and Kulkarni, A.R. (2009) Plant system: nature's nanofactory. *Colloid Surf B: Bioint*, **73**, 219-223.

[44] Jha, A.K., Prasad, K. and Prasad, K. (2009) Biosynthesis of Sb_2O_3 nanoparticles: A low cost green approach. *Biotechnol J*, in press.

[45] Davidson, E. (2004) Method and composition for scavenging sulphide in drilling fluids. US Patent: 6476611.

Vol.1, No.2,136-141 (2009)
doi:10.4236/ns.2009.12017

Investigation on Third-Order Optical Nonlinearities of Two Organometallic Dmit^{2-} Complexes Using Z-Scan Technique

He-Liang Fan[1], Quan Ren[2,*], Xin-Qiang Wang[1], Ting-Bin Li[3], Jing Sun[1], Guang-Hui Zhang[1], Dong Xu[1,*], Gang Yu[1], Zhi-Hua Sun[1]

[1]State Key Laboratory of Crystal Materials and Institute of Crystal Materials, Shandong University, Jinan 250100, China; qren@sdu.edu.cn, xdoffice@sdu.edu.cn
[2]Department of Optics, Shandong University, Jinan 250100, China
[3]Department of Materials and Chemical Engineering, Taishan University, Taian 271021, China

Received 9 August 2009; revised 27 August 2009; accepted 29 August 2009.

ABSTRACT

The third-order nonlinear optical properties of two dmit organometallic complexes, $[(CH_3)_4N][Au(C_3S_5)_2]$ (MeAu) and $[(CH_3)_4N][Ni(C_3S_5)_2]$ (MeNi) in acetone solutions, were characterized using a short pulse Z-scan technique at 1064 nm wavelength. Self-defocusing effects were found in both samples and stronger saturable absorption was observed in MeNi solution comparing with that of MeAu. The origins were analyzed for the differences between the results. Two figures of merit *W* and *T* were also calculated to evaluate the suitability of two materials for all-optical integrated devices. The results of *W*=22.84 and *T*≈0 of MeAu make it an excellent candidate for the all-optical applications.

Keywords: Z-scan Technique; Third-order Nonlinearity; Metal-dmit Complexes; Figure of Merit.

1. INTRODUCTION

As the development of optical communication networks progress, the demand for ultrafast optical switching with femtosecond or picosecond response time operation is rising. materials with large third-order nonlinear optical (NLO) properties and ultrafast response time have arose great interest for its widespread applications in optical switching, signal processing, ultrafast optical communications and optical limiting [1-4]. In recent years, π-conjugated organometallic complexes have emerged as a promising class of third-order nonlinear optical (NLO) materials because of their architectural flexibility with a variety of combinations of central metals and ligands as well as the charge-transfer nature of the metal-ligand bonds, which can further enhance the nonlinearity [5-6].

Special π-conjugated electron systems, like 4,5–dithiolato-1,3-dithiole-2-thione (dmit) complexes, have been used as building blocks for organic, organometallic and coordination-complex electrical conductors and superconductors [7-10]. Currently, more attention has been paid to the third-order NLO properties of these materials [11-13]. These structures that contain transition metal ions may exhibit new properties due to the richness of various excited states present in these systems in addition to the tailorability of metal-organic ligand interactions. Also the π-electron delocalization and the transfer of electron densities between metal atom and the ligands make this kind of compounds exhibit a large molecular hyperpolarizability which can contribute to ultrafast optical response capability and larger third NLO effects. Usually, assessing the suitability of a material for all-optical switching devices is evaluated through two figures of merit: $W = n_2 I/\alpha_0 \lambda$ and $T = \beta\lambda/n_2$ (n_2 is the nonlinear refractive index, I is incident light intensity, α_0 is the linear absorption coefficient, λ is wavelength and β is the nonlinear absorption coefficient) [2,14]. In order to satisfy the requirement, it is necessary to achieve $W>>1$ and $T<<1$. In this paper, the third-order optical nonlinearity of two dmit organometallic complexes, MeAu and MeNi, were reported using a Z-scan technique at 1064 nm with 20 ps pulse duration and 10 Hz repetition rate. Additionally, *W* and *T* of these samples were obtained which were used to evaluate their feasibility of to be applied in all-optical device field.

The Z-scan technique which was firstly reported by M. Sheik-Bahae *et al.* [15], is a simple and sensitive single beam method for measurement of third-order nonlinear optical coefficients. It is based on the self-focusing or

defocusing of a distorted beam of known spatial structure induced by moving a nonlinear sample along the light-propagation direction (Z-axis). Using this method, the magnitude and sign of both the real (nonlinear refraction, NLR) and imaginary (nonlinear absorption, NLA) parts of the nonlinearity of transparent mediums can be immediately obtained based on the relationship of the variation of transmittance in the far field and the sample position. The Z-scan technique can be signed two types: closed-aperture Z-scan and open-aperture Z-scan. For closed-aperture Z-scan both NLR and NLA can be measured simultaneously while for another, the NLA can be independently measured more accurately. To date the Z-scan technique is becoming an increasingly popular approach on measurement of nonlinear optical responses for its convenient operation, higher sensitivity and simple apparatus comparing with other method such as degenerated four-wave mixing, optical Kerr gate, nonlinear interference and so on.

2. EXPERIMENTAL

The molecular structures of MeAu and MeNi were illustrated in **Figure 1**. The synthesized procedures were respectively referred by the literatures [14,16]. The linear UV-Vis-NIR absorption spectra of 1×10^{-4} mol/L solution of two materials in acetone were recorded using a scanning spectrophotometer (Hitachi U-4100, Japan). **Figure 2** shows the results with the wavelength region 330-1500 nm at room temperature. Both MeAu and MeNi represent several absorptive peaks in the UV-Vis region which can be regarded as attributing to the *n*-π transition and the *d*-*p* interaction [16,17]. In another words, for MeAu, there was so wider a transparent window with no absorption in the wavelength region longer than 500 nm. While for MeNi, it also exhibits a strong absorption band with the peak at about 1137 nm in the NIR region (800-1500 nm) which may be assigned to the low-energy π-π^* transition [18].

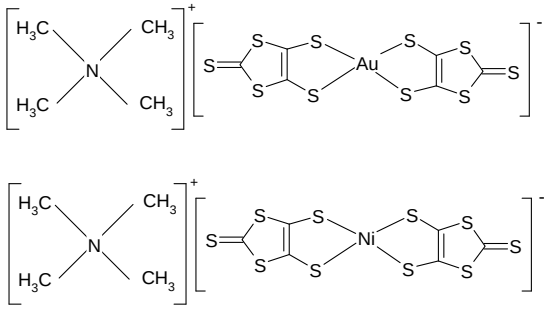


Figure 1. Molecular structures of MeAu and MeNi.

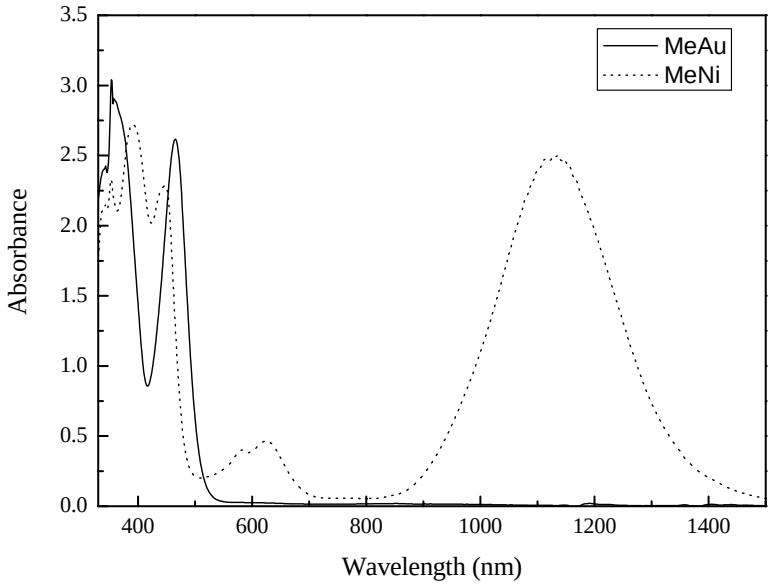


Figure 2. UV-Vis-NIR absorption spectra of acetone solutions of MeAu (solid line) and MeNi (dot line) with a concentration of 1×10^{-4} mol/L at room temperature.

The Z-scan technique was used to characterize the third-order nonlinear optical response of the acetone solutions of two materials. In our measurement, a mode-locked Nd:YAG laser (Leopard-10, Continuum) was employed as the Gaussian light source with a repetition rate of 10 Hz, pulse width of 20 ps and wavelength of 1064 nm. The sample is moved along the optic axis (the Z-direction) through the focus of the lens, which has a focal length of 150 mm, while the energy transmitted through an aperture in the far-field is recorded as a function of the sample position. The radius of the beam waist (w_0) was determined to be 39 μm. Accordingly, the Rayleigh length (z_0) was calculated to be 4.5 mm, much larger than either the thickness of a quartz cell with the optical length of 1 mm. The incident beam intensity (I) performed on the samples was set to be 5.3(±0.3) GW/cm^2. Before measuring the samples, the system was calibrated using a standard CS_2 solution in quartz cell as reference. Measurements on the pure solvent (acetone) in the cell were also performed under the same measuring condition to verify that the peak-valley configuration in the Z-scan curves all originated from the material, but not from the solvent or the quartz cell [19].

3. RESULTS AND DISCUSSION

The concentration of samples used in Z-scan measurements is 1×10^{-4} mol/L. **Figure 3** exhibits the closed-aperture (CA) and open-aperture (OA) data for MeAu in acetone solution. The symmetrical valley-to-peak configuration of the CA curve and the horizontal straight line of the OA curve reveal that the sample shows obvious self-defocusing effect and tiny NLA, which is considered a potential feature for all-optical switching. Then the Z-scan curves for MeNi are shown in **Figure 4 (a)** and **(b)**. The configurations of CA and OA curves in **Figure 4 (a)** both demonstrate single peak,

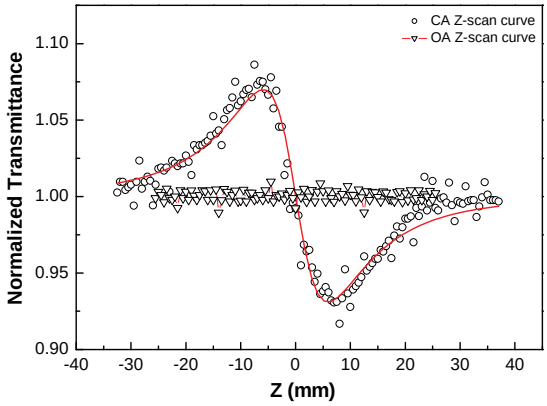
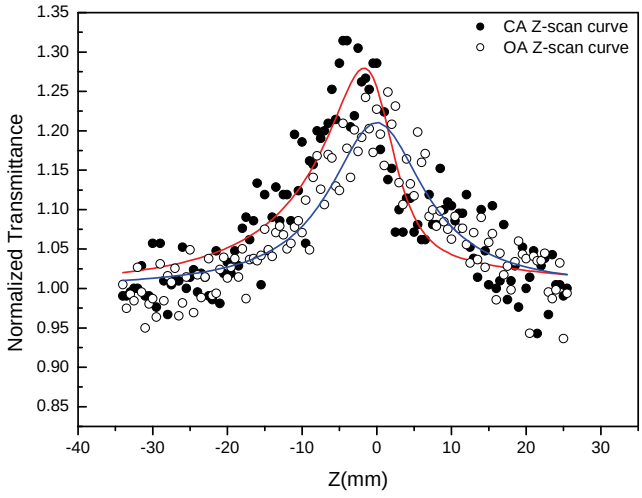
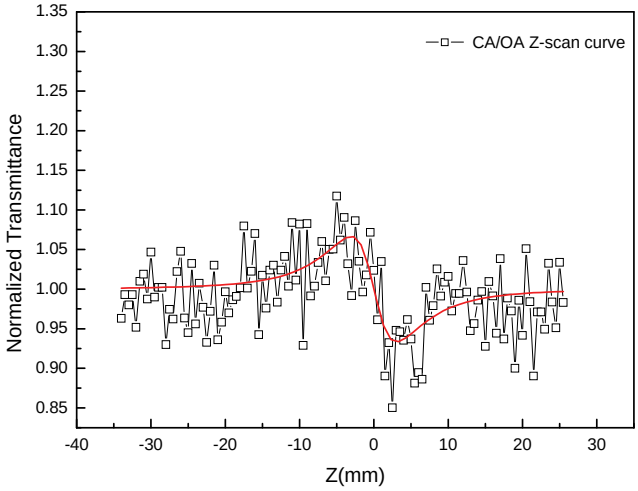


Figure 3. Normalized Z-scan transmittance curves of MeAu in acetone solution with concentration of 1×10^{-4} mol/L. The solid line is the theoretical fitting curve. Z is the sample position away from the focus.



(a)



(b)

Figure 4. Normalized Z-scan transmittance curves of MeNi in acetone solution with concentration of 1×10^{-4} mol/L. The solid line is the theoretical fitting curve. Z is the sample position away from the focus.

Table 1. Nonlinear optical parameters of MeAu and MeNi at 1064 nm.

Sample	α_0 (mm^{-1})	n_2 (10^{-19} m^2/W)	β (10^{-12} m/W)	$\chi^{(3)}$ (10^{-13} esu)	γ (10^{-30} esu)	W	T
MeAu	0.0021	-9.50	-9.17×10^{-3}	4.58	2.65	22.84	≈0
MeNi	4.60	-5.68	-10.34	5.03	2.91	0.013	1.94

suggestion that the NLA which is regarded as the saturable absorption, covers the contribution of the NLR and plays a dominant role in the third-order nonlinear process of the sample, which is a very attractive feature for laser mode-locking, laser Q-switching and optical bistability applications [20-22]. **Figure 4 (b)** shows the division curve of the CA by OA data of MeNi, which also reveals a self-defocusing effect for the sample.

In general, to distinguish the NLR and NLA, the normalized transmittance dependence can be presented as follows [23-24]:

$$T = 1 + \frac{4x}{\left(x^2+9\right)\left(x^2+1\right)}\Delta\Phi - \frac{2\left(x^2+3\right)}{\left(x^2+9\right)\left(x^2+1\right)}\Delta\Psi \quad \textbf{(1)}$$

with

$$\Delta\Phi = kn_2 I_0 L_{eff} \quad \textbf{(2)}$$

$$\Delta\Psi = \beta I_0 L_{eff}/2 \quad \textbf{(3)}$$

where T is the normalized transmittance of the sample, $\Delta\Phi$ and $\Delta\Psi$ is the nonlinear phase shifts due to the NLR and NLA, respectively. Here $x=z/z_0$, indicates the dimensionless relative position from the waist, k is wave vector, $L_{eff} = \left[1-\exp\left(-\alpha_0 L\right)\right]/\alpha_0$ is the effective thickness (L denotes its real thickness).

The real and imaginary parts of the third-order nonlinear susceptibility $\chi^{(3)}$ are related to the NLR and NLA coefficients by [15, 25]:

$$\chi^{(3)} = \chi_R^{(3)} + \mathrm{i}\chi_I^{(3)} \quad \textbf{(4)}$$

$$\chi_R^{(3)}(esu) = \frac{cn_0^2}{120\pi^2} n_2(\mathrm{m}^2/\mathrm{W}) \quad \textbf{(5)}$$

$$\chi_I^{(3)}(esu) = \frac{c^2 n_0^2}{240\pi^2\omega}\beta(\mathrm{m}/\mathrm{W}) \quad \textbf{(6)}$$

where ω is the angular frequency of the light field and c is the velocity of the light in vacuum.

Forwards, the second-order hyperpolarizability γ of the sample molecule can be estimated through the equation [26]

$$\gamma = \frac{\chi^{(3)}}{N_c L_c} \quad \textbf{(7)}$$

where N_c is the number density of molecules and L_c is the local field correction factor which equals $\left[\left(n_0^2+2\right)/3\right]^4$.

According to above-mentioned procedure, the nonlinear parameters of the two sample MeAu and MeNi n_2, β, $\chi^{(3)}$, and γ can be obtained in succession. Additionally, the results of two figures of merit W and T were also calculated basing on the NLO parameters. All the parameters were listed in **Table 1**. We can see that both

MeAu and MeNi show larger third-order nonlinear optical properties because of the delocalized electronic states formed by the overlapping between *p*-*p* and *d* orbits [10]. But the nonlinear absorption coefficient of MeNi is rapidly larger than that of MeAu. The resonant wavelength of 1064 nm of MeNi gives to the stronger saturable absorption comparing with the weaker nonlinear absorption of MeAu at 1064 nm which locates on the off-resonant field of linear absorption [27]. The figures of merit of MeAu were calculated to be W=22.84 and T≈0, which finely satisfy the requirement of suitability for all-optical switching devices W>>1 and T<<1. So the material can be considered to be an excellent candidate to be applied in integrated optics field as all-optical switching devices. While for MeNi, W=0.013 and T=1.94, the values of two figures of merit don't satisfy the requirement of all-optical devices but may be applied in laser mode-locking, laser Q-switching and optical bistability fields because of its saturable absorption properties.

4. CONCLUSIONS

The third-order nonlinear properties of two metal-dmit complexes MeAu and MeNi were investigated using a Z-scan technique at 1064 nm with 20 ps pulse width and 10 Hz repetition rate. Z-scan curves indicated that both MeAu and MeNi show negative nonlinear refraction which are regarded as self-defocusing effects. Meanwhile, tiny nonlinear absorption and stronger saturable absorption was found in MeAu and MeNi, respectively. The figures of merit W and T of two materials were calculated to judge the suitability as all-optical switching devices. The values of MeAu W=22.84 and T≈0 were considered to be appropriate for applications in all-optical integrated field. While for MeNi, the stronger saturable absorption comparing with nonlinear refraction makes it a fine material to be applied in laser mode-locking, laser Q-switching, optical bistability field and so on.

5. ACKNOWLEDGEMENTS

This research work is supported by the grants (Nos. 50772059, 60778037, 60608010 and 50872067) of the National Natural Science Foundation of China (NNSFC) and the Foundation for the Author of National Excellent Doctoral Dissertation of P. R. China (No. 200539).

REFERENCES

[1] Chen, Q.Y., Sargent, E.H., Leclerc, N. and Attias, A.–J. (2003) Ultrafast nonresonant third-order optical nonlinearity of a conjugated 3,3'-bipyridine derivative from 1150 to 1600 nm. *Applied Physics Letters*, **82,** 4420-4422.

[2] Stegeman, G.I., Wright, E.M., Finlayson, N., Zanoni, R. and Seaton, C.T. (1988) Third order nonlinear integrated optics, *Journal of Lightwave Technology*, **6,** 953-970.

[3] Kuang, L., Chen, Q.Y., Sargent, E.H. and Wang, Z.Y. (2003) [60]fullerene-containing polyurethane films with large ultrafast nonresonant third-order nonlinearity at telecommunication wavelengths. *Journal of the American Chemical Society*, **125,** 13648-13649.

[4] He, G.S., Bhawalkar, J.D., Zhao, C.F. and Prasad, P.N. (1995) Optical limiting effect in a two-photon absorption dye doped solid matrix. *Applied Physics Letters*, **67,** 2433-2435.

[5] Tao, S., Miyagoe, T., Maeda, A., Matsuzaki, H. *et al*. (2007) Ultrafast optical switching by using nanocrystals of a halogen-bridged nickel-chain compound dispersed in an optical polymer. *Advanced Materials*, **19,** 2707-2710.

[6] Nalwa, H.S. and Miyata, S. (1997) Nonlinear optics of organic molecules and polymers (Boca Raton: CRC Press, Inc.) Chap. 11.

[7] Takahashi, K., Cui, H.B., Okano, Y., Kobayashi, H., Einaga, Y. and Sato, O. (2006) Electrical conductivity modulation coupled to a high-spin-low-spin conversion in the molecular system [Fe-III(qsal)(2)][Ni(dmit)(2)] (3)center dot CH3CN center dot H_2O. *Inorganic Chemistry*, **45,** 5739-5741.

[8] De Caro, D., Fraxedas, J., Faulmann, C., Malfant, I., Milon, J., Lamere, J.F., Colliere, V. and Valade, L. (2004) Metallic thin films of TTF[Ni(dmit)(2)](2) by electrodeposition on (001)-oriented silicon substrates. *Advanced Materials*, **16,** 835-838.

[9] Tamura, M. and Kato, R. (2002) Magnetic susceptibility of beta '-[Pd(dmit)(2)] salts (dmit=1,3-dithiol-2-thione-4, 5-dithiolate, C_3S_5): evidence for frustration in spin-1/2 Heisenberg antiferromagnets on a triangular lattice. *Journal of Physics-Condensed Matter*, **14,** L729-L734.

[10] Wang, S.F., Huang, W.T., Zhang, T.Q., Gong, Q.H., Okuma, Y., Horikiri, M. and Miura, Y.F. (1999) Third-order nonlinear optical properties of didodecyldimethylammonium-Au(dmit)(2). *Applied Physics Letters*, **75,** 1845-1847.

[11] Dai, J., Bian, C.Q., Wang, X., Xu, Q.F., Zhou, M.Y., Munakata, M., Maekawa, M., Tong, M.H., Sun, Z.R. and Zeng, H.P. (2000) A new method to synthesize unsymmetrical dithiolene metal complexes of 1,3-dithiole-2-thione-4,5-dithiolate for third-order nonlinear optical applications. *Journal of the American Chemical Society*, **122,** 11007-11008.

[12] Ji, Y., Zuo, J.L., Chen, L.X., Tian, Y.Q., Song, Y.L., Li, Y.Z. and You, X.Z. (2005) Synthesis, characterization and optical limiting effect of nickel complexes of multi-sulfur 1,2-dithiolene. *Journal of Physics and Chemistry of Solids*, **66,** 207-212.

[13] Cheng, Y.G., Mao, Y.L., Liu, J.H., Feng, S.Y. and He, T.C. (2007) Third-order nonlinear optical properties of a dmit(2-) salt by Z-scan technique, *Journal of Modern Optics*. **54,** 2763-2768.

[14] Sun, J., Guo, W.F., Wang, X.Q., Zhang, G.H., Sun, X.B., Zhu, L.Y., Ren, Q. and Xu, D. (2007) Nonlinear optical absorption studies of an organo-metallic complex by Z-scan technique. *Optics Communications*, **280,** 183-187.

[15] Sheik-Bahae, M., Said, A.A., Wei, T.H., Hagan, D.J. and Van Stryland, E.W. (1990) Sensitive measurement of optical nonlinearities using a single beam. *IEEE Journal of Quantum Electronics*, **26,** 760-769.

[16] Sun, X.B., Ren, Q., Wang, X.Q., Zhang, G.H., Yang, H.L., Feng, L., Xu, D. and Liu, Z.B. (2006) Nonlinear optical properties of [(CH3)(4)N]Au(dmit)(2) using Z-scan technique. *Chinese Physics*, **15,** 2618-2622.

[17] Liu, C.M., Zhang, D.Q., Song, Y.L., Zhan, C.L., Li, Y.L. and Zhu, D.B. (2002) Synthesis, crystal structure and third-order nonlinear optical behavior of a novel dimeric mixed-ligand zinc(II) complex of 1,3-dithiole-2-thione-4,5-dithiolate. *European Journal of Inorganic Chemistry*, **7,** 1591-1594.

[18] Herman, Z.S., Kirchner, R.F., Loew, G.H., Mueller-Westerhoff, U.H., Nazzal, A. and Zerner, M.C. (1982) Electronic spectra and structure of bis(ethylene-1, 2-dithiolato)nickel and bis-(propene-3-thione-1-thiolato) nickel, *Inorganic Chemistry*, **21,** 46-56.

[19] Ren, Q., Sun, X.B., Wang, X.Q., Zhang, G.H., Yang, X.D., Zhang, F.J., Yang, H.L., Chow, Y.T. and Xu, D. (2008) Short pulse Z-scan investigations of optical nonlinearities of a novel organometallic complex: [(C2H5)(4)N](2)[Cu(dmit)(2)] at 532 and 1064nm. *Applied Physics A-Materials Science & Processing*, **90,** 685-688.

[20] Yoshita, M., Kuramoto, M., Ikeda, M. and Yokoyama, H. (2009) Mode locking of a GaInN semiconductor laser with an internal saturable absorber. *Applied Physics Letters*, **94,** 061104.

[21] Zhang, S.M., Lu, F.Y. and Wang, J. (2006) Self-Q-switching and mode-locking in an all-fiber Er/Yb co-doped fiber ring laser. *Optics Communications*, **263,** 47-51.

[22] Mao, Q.H. and Lit, J.W.Y. (2002) Optical bistability in an L-band dual-wavelength erbium-doped fiber laser with overlapping cavities. *IEEE Photonics Technology Letters*, **14,** 1252-1254.

[23] Yin, M., Li, H.P., Tang, S.H. and Ji, W. (2000) Determination of nonlinear absorption and refraction by single Z-scan method. *Applied Physics B-Lasers and Optics*, **70,** 587-591.

[24] Liu, X.D., Guo, S.L., Wang, H.T. and Hou, L.T. (2001) Theoretical study on the closed-aperture Z-scan curves in the materials with nonlinear refraction and strong nonlinear absorption. *Optics Communications*, **197,** 431-437.

[25] Chapple, P.B., Staromlynska, J., Hermann, J.A., Mckay, T.J. and McDuff, R.G. (1997) Single-beam Z-scan: measurement techniques and analysis. *Journal of Nonlinear Optical Physics and Materials*, **6,** 253-293.

[26] Gong, Q.H., Sun, Y.X. Xia, Z.J., Zou, Y.H., Gu, Z.N., Zhou, X.H. and Qiang, D. (1992) Nonresonant third- order optical nonlinearity of all-carbon molecules C_{60}. *Journal of Applied Physics*, **71,** 3025-3026.

[27] Li, C.F., Zhang, L., Yang, M., Wang, H. and Wang, Y.X. (1994) Dynamic and steady-state behaviors of reverse saturable absorption in metallophthalocyanine. *Physical Review A*, **49,** 1149-1157.

Vol.1, No.2,142-150 (2009)
doi:10.4236/ns.2009.12018

Natural Science

Electric-Jet Assisted Layer-by-Layer Deposition of Gold Nanoparticles to Prepare Conducting Tracks

S. R. Samarasinghe[1], Isabel Pastoriza-Santos[3], M. J. Edirisinghe[1*], M. J. Reece[2], Luis M. Liz-Marzán[3], M. R. Nangrejo[1], Z. Ahmad

[1]Department of Mechanical Engineering, University College of London, Torrington Place, London. WC1E 7JE, UK; m.edirisinghe@ucl.ac.uk
[2]Centre for Materials Research, Queen Mary, University of London, Mile End Road, London. E1 4NS, UK
[3]Department of Physical Chemistry, University of Vigo, Vigo 36310, Spain

Received 18 July 2009; revised 30 July 2009; accepted 3 August 2009.

ABSTRACT

A suspension of 15nm diameter gold nanoparticles has been deposited along a line on a silicon substrate with the assistance of a jet generated in an electric field. In order to control the evaporation of the solvent used to suspend the gold nanoparticles, a heating device was used to change the substrate temperature. Layer-by-layer deposition enabled the direct writing of gold tracks having an electrical resistivity of 1.8 × 10^{-7} Ωm, only about an order of magnitude above the electrical resistivity of bulk gold.

Keywords: Gold; Electrohydrodynamic; Jet; Direct Write; Track; Electrical Conductivity

1. INTRODUCTION

The forming of fine metallic patterns from colloids and suspensions is gaining tremendous interest because it is a potential fabrication route for the next generation electronic devices. Techniques, such as electron-beam lithography and photo-lithography are the most popular patterning techniques, and are at the heart of modern day microfabrication, nanotechnology and molecular electronics. Lithography techniques require a mask or resistive film to pattern microstructures on substrates and thereafter harsh chemical etching is needed to produce the final pattern. This makes them unsuitable for patterning nanoparticles or molecules with organic or biological functionalities, since it impairs the organic molecules and biological entities [1]. In addition, these lithographic techniques are not only time consuming but also quite complicated. For these reasons, the development of convenient and fast processing techniques to fabricate conductive lines has attracted more attention in recent years [2].

Recently developed techniques, for example, micro-contact printing, also require an elastomer stamp for patterning, which can deform due to its elastomeric nature, resulting in distorted patterns [3]. Dip-pen lithography is also a recently developed patterning technique, which allows direct transporting and patterning of particles and molecules at nanometer scale (30-100 nm) onto a substrate from the tip of an atomic force microscope. However, this technique can usually convey only a small amount of materials, since the transfer efficiency is relatively low [1].

Direct write technologies have been explored recently for fabricating fine patterns whose line widths range from the meso to the nanoscale. The term direct write refers to any technique or process that is capable of depositing, dispensing or processing different types of materials on various surfaces following a preset pattern or layout. The main advantages of the direct write approach is that patterns or structures can be obtained without the use of moulds or pre-fabrication processes, masks, and liquids for etching. Direct write technologies are therefore low cost, high speed, non-contact and environmentally friendly processes [4,5].

As a non-contact patterning technique, ink-jet printing (IJP) has been used in the last decade for a number of new applications, such as the fabrication of organic light-emitting diodes, transistors and integrated circuits, conducting polymer devices, structural polymers, ceramics and biomolecular arrays [6]. **Table 1** shows recent publications on fabricating conducting tracks by ink-jet printing and the concentration and resistivities achieved by other researchers. The metal concentration used in most of the studies varies from 20 – 48 wt.%.

Printing using a jet generated in an electric field is an emerging direct write patterning technology. In this method, the medium is made to jet and disperse into fine droplets, which are deposited on a substrate using computer control to form a pre-designed pattern. Deposition

Table 1. Summary of recent works on fabricating conducting tracks by IJP.

Metal	Particle Size (nm)	Width (μm)	Conc. (wt.%)	Curing Condition	Resistivity (Ωm)	Ref.
Ag	10 – 20	60	34.5	Electrical	2.7×10^{-8}	[10]
Ag	21	90	20	150 ^{0}C	3.2×10^{-8}	[11]
Ag	1 – 10	125	30	300 ^{0}C	3.5×10^{-7}	[12]
Ag	20	65	20	300 ^{0}C	3.5×10^{-8}	[13]
Ag	-	100	16	150 ^{0}C	4.8×10^{-8}	[14]
Ag	-	750	48	300 ^{0}C	1.5×10^{-7}	[15]
Ag	10 – 50	130	25	260 ^{0}C	1.6×10^{-7}	[2]
Ag	5 – 10	160	60	Microwave	1.6×10^{-7}	[16]
Ag	5 – 7	1500	10	320 ^{0}C	1.1×10^{-6}	[6]
Au	2 – 4	17	30	Laser	1.4×10^{-7}	[17]
Au	2 – 5	20	30	Laser	1.4×10^{-7}	[18]
Au	-	360	31	500 ^{0}C	2.7×10^{-7}	[19]
Au	2 – 4	125	34	Laser	6.2×10^{-8}	[20]
Cu	40 – 50	65	20	325 ^{0}C	1.72×10^{-7}	[21]

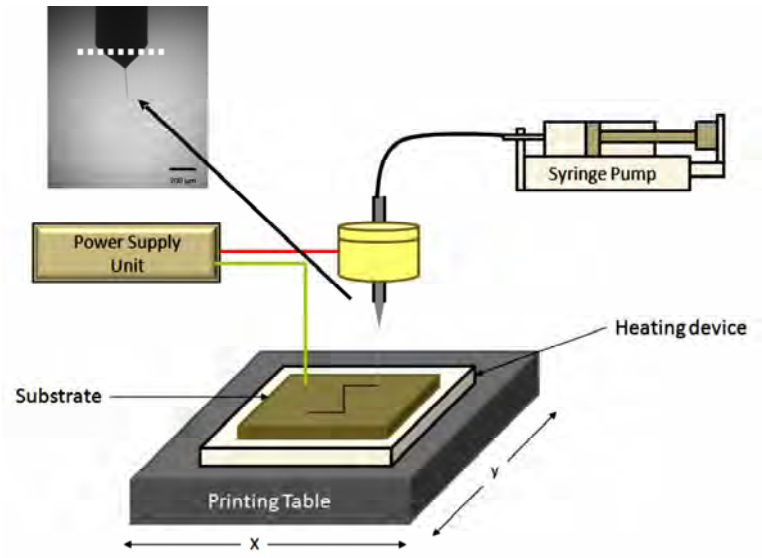


Figure 1. Schematic diagram of the electrohydrodynamic jet printing process.

of materials using electric field assisted jet printing offers some advantages over other non-contact printing techniques. Firstly, by careful tailoring of the physical properties of the medium and by suitably fixing its flow rate and the applied voltage controlling the electric field it is possible to produce very fine droplets in the range of 40nm - 1.8μm [7,8]. Second, the diameter of the capillaries (needles, nozzles) used in this method is much coarser (> 100μm inner diameter) than the capillaries used in ink-jet printing (usually around 20-60μm). The use of larger capillaries reduces the possibilities of blockages and allows viscous suspension containing high volume loading (above 30 vol.%) of solid particles to be processed [9].

Electric-field assisted jet printing has been used to produce conducting tracks from silver nanoparticles [22,23], the silver concentrations used in their experiments varied from 20 – 30 wt. %. In contrast, in this

paper, a low concentration gold suspension (~0.1 wt. %) containing 15nm diameter particles was used to deposit conducting tracks layer-by-layer with the aid of a jet generated in an electric field.

2. EXPERIMENTAL DETAILS

2.1. Preparation of Gold Suspension

Tetrachloroauric acid and trisodium citrate were purchased from Aldrich. Poly (vinylpyrrolidone) (PVP, molecular weight 10,000) was supplied by Fluka. All of the chemicals were used as received and milli-Q water was used to make up all solutions (R > 18.2MΩ cm). Gold nanoparticles with an average diameter of 15nm and 10% polydispersity were synthesized according to the standard sodium citrate reduction method [24] by boiling 5×10^{-4} M $HAuCl_4$ in the presence of 1.6×10^{-3} M sodium citrate for 900s. After cooling down, the particles were transferred into ethanol upon functionalization with PVP [25]. Briefly, an amount of PVP sufficient to coat the particles with 60 PVP monomers per nm^2 was dissolved by ultrasonication for 900s in water and added to the gold colloid. The polymer was allowed to adsorb to the gold particles overnight while stirring. The particles were subsequently centrifuged (3500 r.p.m.) to remove the unbound PVP and redispersed in ethanol.

2.2. Electric-field Assisted Jetting and Deposition

The apparatus (**Figure 1**) consists of a programmable syringe pump (Harvard Apparatus Ltd., Edenbridge, UK) supplying the gold suspension to the stainless steel nozzle (internal diameter of ~200μm and external diameter of ~400μm) held in epoxy resin. The electrical power supply unit consisted of a high voltage power supply (Glassman Europe Ltd., Tadley, UK) capable of supplying up to 30kV between the electrodes.

A custom built printing device was used to pattern microstructures on substrates. It consists of a stepper motor driven 2D system and the X and Y tables (**Figure 1**) are mounted on one another keeping the 2-axis profile very low and the system is computer-driven using a programmable motion controller. A datum and an end of travel limit sensor are fitted on each of the tables to trigger the controller when a respective carriage reaches a limit. A perspex sheet was mounted firmly on the 2-axis system in order to accommodate the heating device, which was used to control the temperature of the silicon wafer substrate. A power supply was used to form an electric field between the nozzle and the heating device. Using motion planner software X and Y coordinates can be created and downloaded to the 2-axis controller, allowing the 2-axis system to write the path described by the co-ordinates provided.

Printing was carried out under various conditions (see text below) and tracks were sintered by heating to 400°C at 2°C min^{-1} and held for 1800s before cooling down to the ambient temperature.

2.3. Microscopical Characterisation

Samples were coated with carbon before examination by Scanning Electron Microscopy (SEM) and Energy Dispersive X-ray (EDX) analysis. The structure and the surface morphology of the sintered films were investigated using a JEOL JSM-6301F field emission scanning electron microscope operating in the secondary electron mode with an accelerating voltage of 10 kV and with a working distance up to 15mm. The EDX analysis was performed with an Oxford INCA Energy 200 X-ray energy dispersive spectrometer system.

2.4. Electrical Testing

In order to measure the resistance of the tracks, silver (Silver Ink-P6100, Johnson Matthey Catalysts, Enfield, UK) electrodes were placed along the sintered track. Firstly, a multimeter (FLUKE 189 TRUE RMS) was used to measure the resistance of the tracks and later a four-point method was used to obtain a more accurate result by eliminating the resistance of the electrodes and the equipment used.

3. RESULTS AND DISCUSSION

When the electric field is applied, the gold suspension jets **(Figure 1**) and droplets from the jet break-up were deposited on silicon wafers in order to produce a continuous track. The processing parameters such as flow rate, applied voltage and the distance between the substrate and needle exit were varied to find the optimum pattering conditions. Thereafter in all of the patterning experiments the flow rate and applied voltage were set to 5×10^{-11} m^3s^{-1} and 1.4 kV, respectively, and the distance between the needle exit and substrate was ~0.4mm. For the layer-by-layer deposition approach, the printing table was moved at 5mm s^{-1} and a new layer was deposited every 15s. **Figure 2** shows an optical micrograph of a single layer track. Due to the well known Marangoni effect, immediately after deposition, most of the particles can be seen clustering together at the edges of the line and a random distribution of particles can be seen at the centre of the track.

Deegan *et al* [26]., reported that when the contact angle of the droplet on the substrate is $< 90^0$ as in this case, solvent evaporation plays a critical role in drying-mediated self assembly from a dilute colloidal droplet on a wetted surface. In order to observe the gold nanoparticle distribution along the track at different solvent evaporation rates, the substrate was heated at different temperatures (35 – 85 ^{0}C) before patterning. When

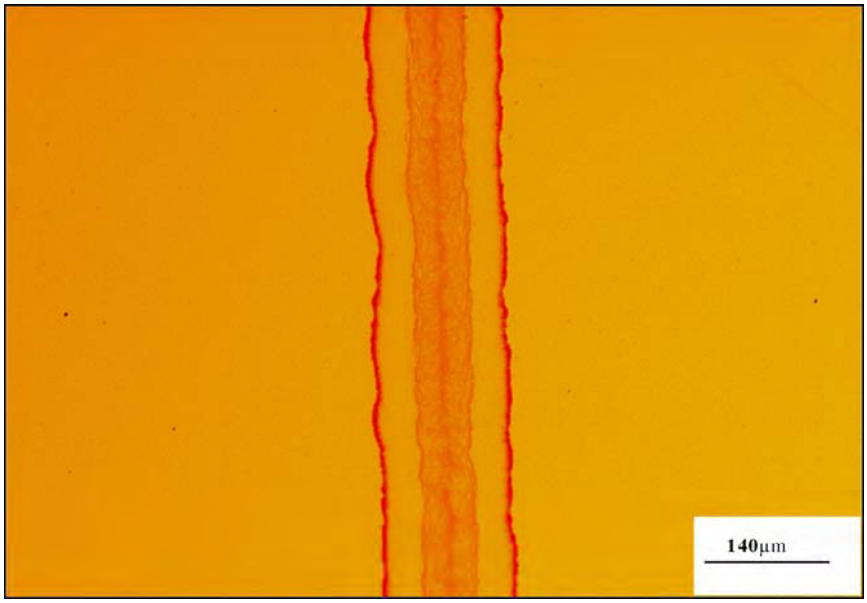


Figure 2. Single layer gold track deposited at 5 x $10^{-11}m^3s^{-1}$ and 1.4 kV.

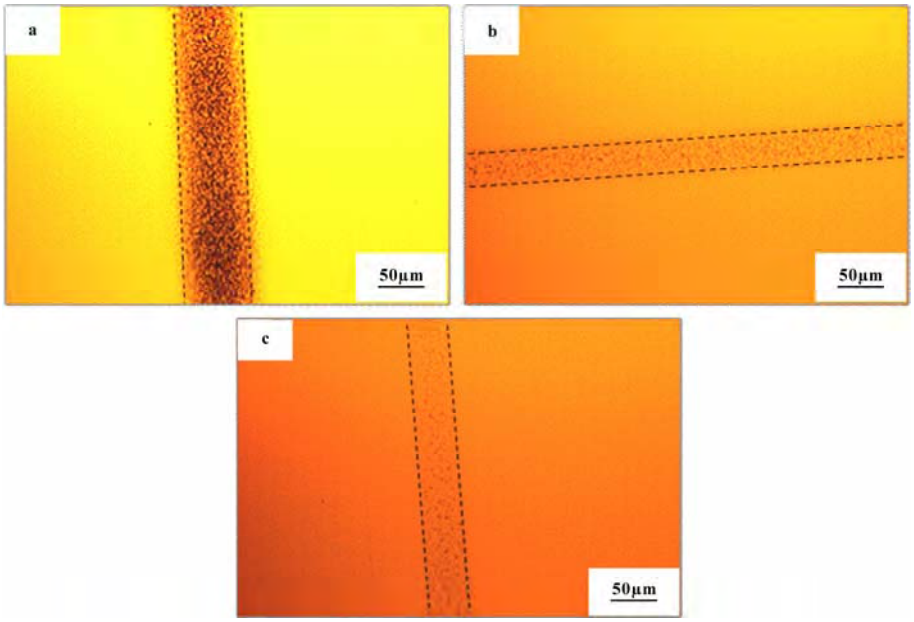


Figure 3. Optical micrographs of the single layer tracks deposited at 85 ^{0}C with increasing table speeds. a) 3 mms^{-1}, b) 6 mms^{-1} and c) 9 mms^{-1}. Dotted lines indicates the edge of the track.

the substrate temperature was increased to >85 ^{0}C patterning was not possible as a stable–jet could not be achieved due to rapid solvent evaporation from the exit of the needle. In order to find out the effect of patterning speed, the table speed was varied from 3 – 9mm s^{-1}. **Figure 3** shows optical micrographs of single layer of track deposited at 85 ^{0}C and with varying table speeds, although a higher speed generates a narrower track, the number of gold particles in the track after a single printing pass is higher at the lower speed (3mms^{-1}), therefore, this speed was selected for the multi-layer printing deposition work described below.

Although the particle spreading during solvent evaporation can be reduced significantly by increasing the substrate temperature, due to the low concentration of Au nanoparticles in the suspension it was not possible to produce dense tracks by a single deposition run, thus a layer-by-layer deposition technique was employed.

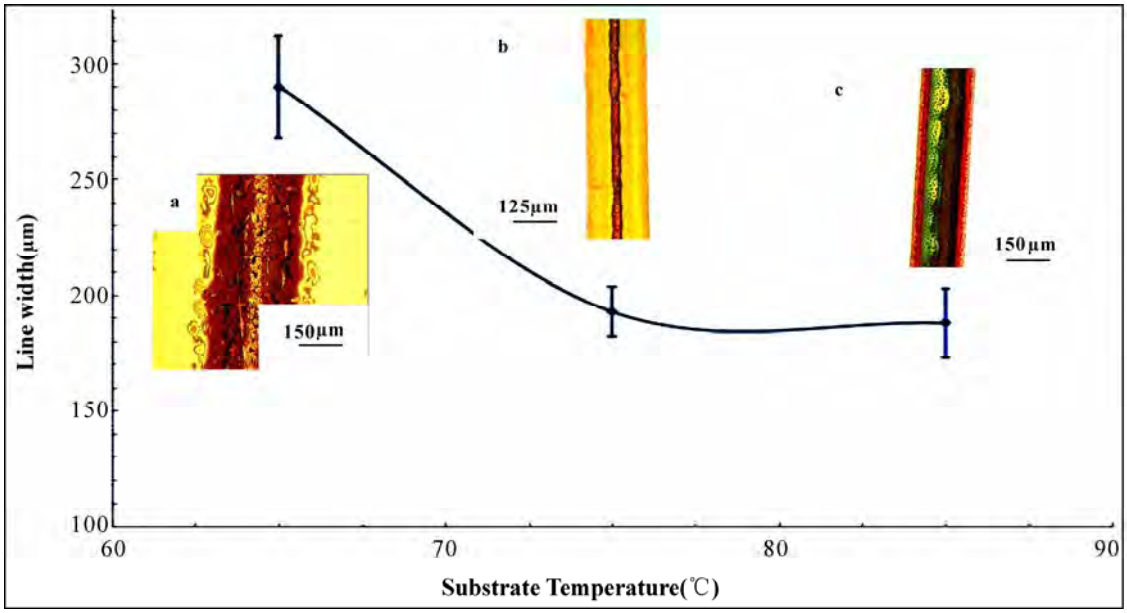


Figure 4. A graph depicting the track width variation with increasing substrate temperature a) 65 ^{0}C, b) 75 ^{0}C and c) 85 ^{0}C . The flow rate was 5 x 10^{-11} m^3s^1, the applied voltage was 1.4 kV and 50 layers were prepared.

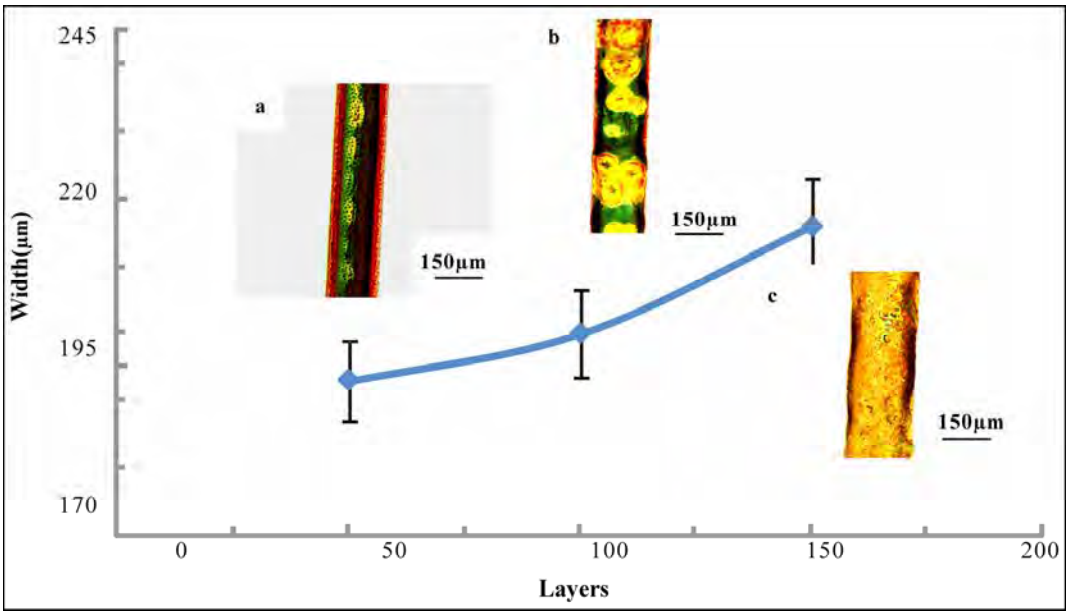


Figure 5. A graph depicting the line width variation with increasing layers at 85°C, a) 50, b) 100 and c) 150. The flow rate was 5 x 10^{-11} m^3s^{-1} and applied voltage was 1.4 kV.

Three different substrate temperatures were investigated in conjunction with layer-by-layer deposition, **Figure 4** shows the effect of increasing substrate temperature on the track width prepared using 50 layers. The minimum width is achieved at 75°C. **Figure 5** shows the effect on track width due to different numbers of layers deposited at 85 ^{0}C. The line width increased with increasing number of layers due to the suttle oscillatory motion of the jet and its digression from the centre line [23,27].

Figure 6 shows a macro image and scanning electron micrographs of a sintered-layered track prepared using a substrate temperature of 85°C. Although the films appear uniform at a low magnifications, at higher magnifications **(Figure 7)** they reveal that the films contain "hillocks" (small Au hills that rise above the film). The formation of hillocks is due to preferential landing of some droplets on the substrate and is a characteristic of the fabrication route used and have been explained in

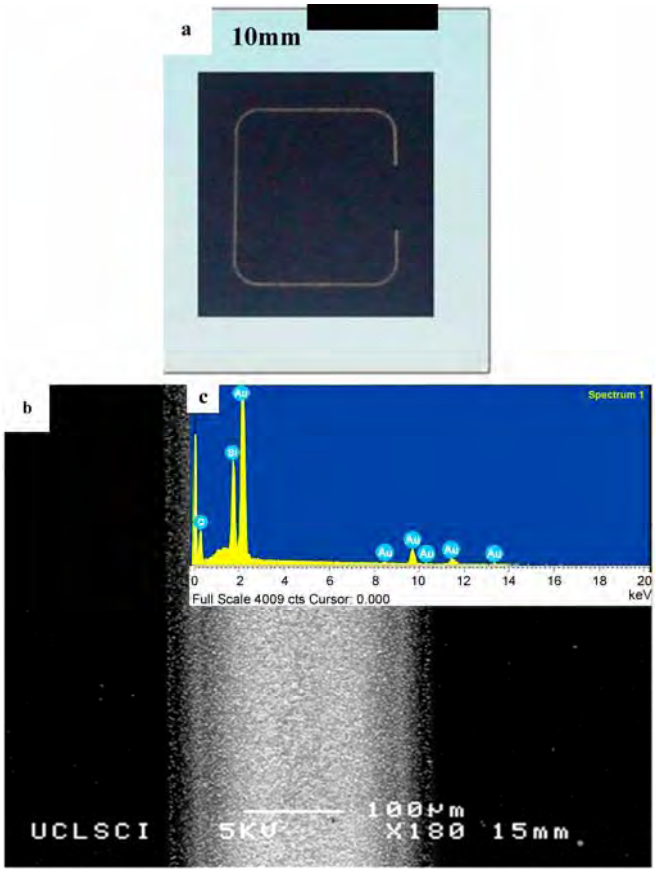


Figure 6. a) Sintered track image, b) scanning electron micrograph of sintered track and c) spectra analysis of *(b)*.

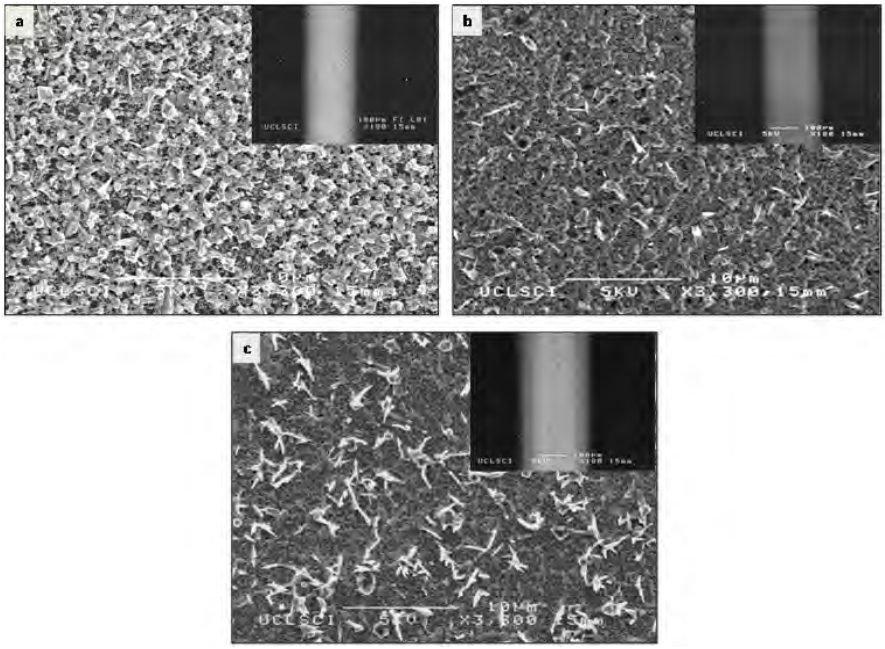


Figure 7. Scanning electron micrographs of the centre of the track with a) 50 layers, b) 100 layers and c) 150 layers. The substrate temperature was set at 85 ^{0}C. Bright spots indicate the top of hillocks.

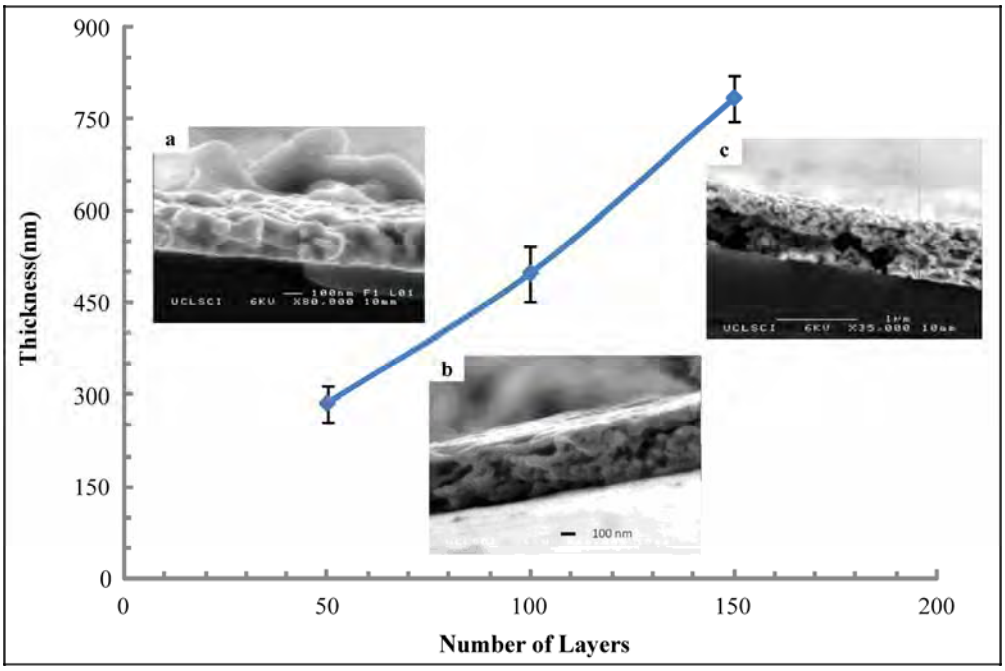


Figure 8. Graph depicting the thickness variation with increasing layers a) 50, b) 100 and c) 150.

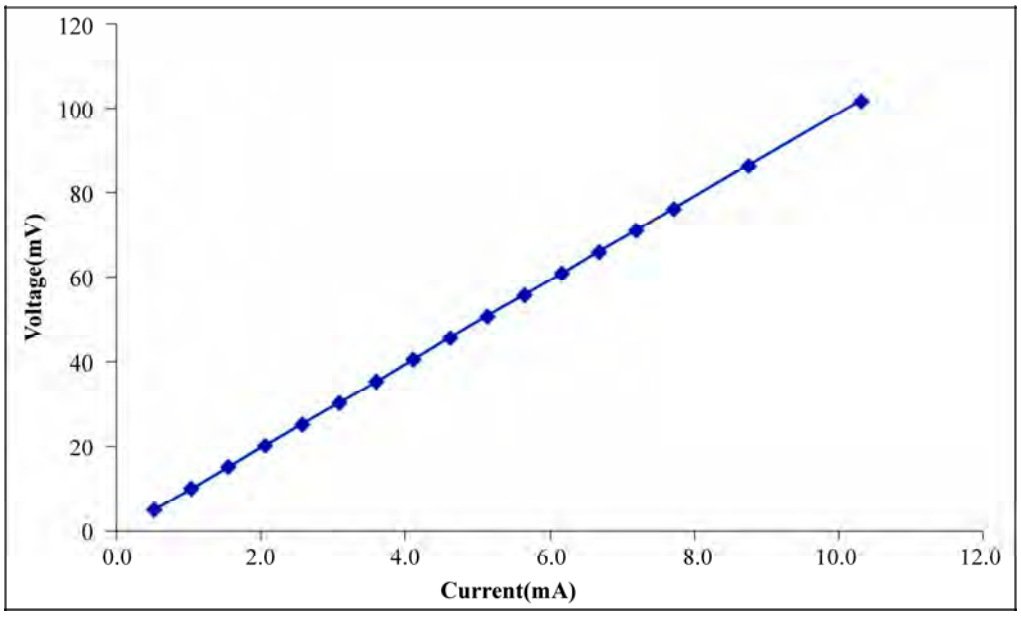


Figure 9. Voltage (V) *vs* current (I) relationship.

more detail in Samarasinghe *et al* [28,29]. The melting point of 15nm diameter gold particles is ~950 ^{0}C [30,31]. Therefore, at 400 ^{0}C appreciable sintering and growth of the particles can be expected. Cross-section images of the tracks and the variation of the thickness of the tracks deposited due to increasing the number of layers from 50 to 150 are illustrated in **Figure 8**. The thickness is an essential parameters for the electrical measurements discussed below.

Figure 9 shows that voltage-current (*V-I)* characteristics of the tracks showed a linear Ohmic behaviour. The specific electrical resistivity ρ of the produced tracks were calculated by the formula $\rho = RA/L$, where R is the electrical resistance of the line, L is the length of line and A is the cross section area of the line. The resistance of the track was measured using the *V-I* curve. The cross section area of the track was taken as $A = wt$ where w is the width of the track and t is the thickness of the track.

Table 2. Resistivity of the printed tracks at different deposition parameters. The resistivity of bulk gold is 2.4 x 10^{-8} Ω m. Substrate temperature is 85°C.

Layers	Length (mm)	Width (μm)	Thickness (nm)	Resistance (Ω)	Resistivity (Ωm)
50	9.5	193	286	47.6	2.8 x 10^{-7}
100	9.5	200	497	24.8	2.6 x 10^{-7}
150	9.5	216	784	9.9	1.8 x 10^{-7}

The electrical resistivity of the printed tracks (**Table 2**) using a layer-by-layer approach was in the range of 1.8 x 10^{-7} – 2.8 x 10^{-7} Ωm. Although this value is higher than the resistivity of bulk gold, it is deemed satisfactory especially if one considers the fact that the initial concentration of the gold suspension used in this study was 0.1 wt. % and no specific processing steps were performed to target the reduction of electrical resistivity. The resistivity compares well with the values of other ink-jet based methods listed in **Table 1.**

4. CONCLUSIONS

This paper demonstrates that gold nanoparticles in dilute suspensions have been successfully assembled to direct write conducting tracks using a simple, economical electric-field assisted printing method. A printing speed of 3mms^{-1}, a suspension flow rate of 5 × 10^{-11} m^{3}s^{-1}, an applied voltage of 1.4kV, and the distance between the needle exit and substrate kept at 0.4m were found to be optimum. However, the control of the number of layers deposited and the substrate temperature are crucial parameters to control the track geometry. A track containing of 50 layers deposited with the substrate held at 85°C provided a continuous track with a resistivity of 2.8×10^{-7}Ω, but this was one order of magnitude above the resistivity of bulk gold.

5. ACKNOWLEDGEMENTS

The authors would like to thank the Leverhulme Trust (Grant: F/07 134/BL), EPSRC, UK (platform grant: EP/E045839) for supporting this work. Dr. K.B. Chong at QMUL is thanked for his assistance with electrical measurements; SRS acknowledges the PhD scholarship awarded by UCL, which initiated this work. The help of the Archaeology Department of the University College London is acknowledged for technical assistance with the microscopy.

REFERNCES

[1] Xu, J., Drelich, J. and Nadgorny, E.M. (2004) Laser-based patterning of gold nanoparticles into microstructures. *Langmuir*, **20(4),** 1021-1025.

[2] Lee, H.H., Chou, K.S. and Huang, K.C. (2005) Inkjet printing of nanosized silver colloids. *Nanotechnology*, **16 (10)**, 2436-2441.

[3] Rogers, J.A., Paul, K.E. and Whitesides, G.M. (1998) Quantifying distortions in soft lithography. *Journal of Vacuum Science & Technology B*, **16(1)**, 88-97.

[4] Chrisey, D.B. (2000) Materials processing - The power of direct writing. *Science*, **289(5481),** 879-881.

[5] Yu, J.H., Kim, S.Y. and Hwang, J. (2007) Effect of viscosity of silver nanoparticle suspension on conductive line patterned by electrohydrodynamic jet printing. *Applied Physics A-Materials Science & Processing*, **89(1),** 157-159.

[6] Kamyshny, A., Ben-Moshe, M., Aviezer, S. and Magdassi, S. (2005) Ink-jet printing of metallic nanoparticles and microemulsions. *Macromolecular Rapid Communications*, **26(4),** 281-288.

[7] Chen, D.R., Pui, D.Y.H. and Kaufman, S.L. (1995) Electrospraying of conducting liquids for monodisperse aerosol generation in the 4 Nm to 1.8 Mu-M diameter range. *Journal of Aerosol Science*, **26(6),** 963-977.

[8] Sullivan, A.C., Scott, K. and Jayasinghe, S.N. (2007) Nanofabrication by electrohydrodynamic jetting of a tailor-made living siloxane sol. *Macromolecular Chemistry and Physics*, **208,** 2032-2038.

[9] Jayasinghe, S.N., Edirisinghe, M.J. and Wang, D.Z. (2004) Controlled deposition of nanoparticle clusters by electrohydrodynamic atomization. *Nanotechnology*, **15(11),** 1519-1523.

[10] Allen, M.L., Aronniemi, M., Mattila, T., Alastalo, A., Ojanpera, K., Suhonen, M. and Seppa, H. (2008) Electrical sintering of nanoparticle structures. *Nanotechnology*, **19(17),** Article Number: 175201.

[11] Kim, D., and Moon, J. (2005) Highly conductive ink jet printed films of nanosilver particles for printable electronics. *Electrochemical and Solid State Letters*, **8(11),** J30-J33.

[12] Szczech, J.B., Megaridis, C.M., Gamota, D.R. and Zhang, J. (2002) Fine-line conductor manufacturing using drop-on-demand PZT printing technology. *IEEE Transactions on Electronics Packaging Manufacturing*, **25,** 26-33.

[13] Kim, D., Jeong, S., Park, B.K. and Moon, J. (2006) Direct writing of silver conductive patterns: Improvement of film morphology and conductance by controlling solvent compositions. *Applied Physics Letters*, **89(26),** article number: 264101.

[14] Dearden, A.L., Smith, P.J., Shin, D.Y., Reis, N., Derby,

B. and O'Brien, P. (2005) A low curing temperature silver ink for use in ink-jet printing and subsequent production of conductive tracks. *Macromolecular Rapid Communications*, **26(4)**, 315-318.

[15] Liu, Z.C., Su, Y. and Varahramyan, K. (2005) Ink-jet-printed silver conductors using silver nitrate ink and their electrical contacts with conducting polymers. *Thin Solid Films*, **478(1-2),** 275-279.

[16] Perelaer, J., De Gans, B.J. and Schubert, U.S. (2006) Ink-jet printing and microwave sintering of conductive silver tracks. *Advanced Materials*, **18(16),** 2101-2104.

[17] Bieri, N.R., Chung, J., Haferl, S.E., Poulikakos, D. and Grigoropoulos, C.P. (2003) Microstructuring by printing and laser curing of nanoparticle solutions. *Applied Physics Letters*, **82(20),** 3529-3531.

[18] Bieri, N.R., Chung, J., Poulikakos, D. and Grigoropoulos, C.P. (2004) Manufacturing of nanoscale thickness gold lines by laser curing of a discretely deposited nanoparticle suspension. *Superlattices and Microstructures*, **35(3-6),** 437-444.

[19] Nur, H.M., Song, J.H., Evans, J.R.G. and Edirisinghe, M.J. (2002) Ink-jet printing of gold conductive tracks. *Journal of Materials Science-Materials in Electronics*, **13,** 213-219.

[20] Chung, J.W., Ko, S.W., Bieri, N.R., Grigoropoulos, C.P. and Poulikakos, D. (2004) Conductor microstructures by laser curing of printed gold nanoparticle ink. *Applied Physics Letters*, **84(5),** 801-803.

[21] Park, B.K., Kim, D., Jeong, S., Moon, J. and Kim, J.S. (2007) Direct writing of copper conductive patterns by ink-jet printing. *Thin Solid Films*, **515(19),** 7706-7711.

[22] Lee, D.Y., Hwang, E.S., Yu, T.U., Kim, Y.J. and Hwang, J. (2006) Structuring of micro line conductor using electro-hydrodynamic printing of a silver nanoparticle suspension. *Applied Physics A-Materials Science & Processing*, **82(4),** 671-674.

[23] Lee, D.Y., Shin, Y.S., Park, S.E., Yu, T.U. and Hwang, J. (2007) Electrohydrodynamic printing of silver nanoparticles by using a focused nanocolloid jet. *Applied Physics Letters*, **90(8),** article number: 081905.

[24] Enustun, B.V. and Turkevich, J. (1963) Coagulation of colloidal gold. *Journal of the American Chemical Society*, **85(21),** 3317-3328.

[25] Graf, C., Vossen, D.L.J., Imhof, A. and Van Blaaderen, A. (2003) A general method to coat colloidal particles with silica. *Langmuir*, **19(17),** 6693-6700.

[26] Deegan, R.D., Bakajin, O., Dupont, T.F., Huber, G., Nagel, S.R. and Witten, T.A. (1997) Capillary flow as the cause of ring stains from dried liquid drops. *Nature*, **389(6653),** 827-829.

[27] Reneker, D.H., Yarin, A.L., Fong, H. and Koombhongse, S. (2000) Bending instability of electrically charged liquid jets of polymer solutions in electrospinning. *Journal of Applied Physics*, **87(9),** 4531-4547.

[28] Samarasinghe, S.R., Pastoriza-Santos, I., Edirisinghe, M. J., Reece, M.J. and Liz-Marzan, L.M. (2006) Printing gold nanoparticles with an electrohydrodynamic direct-write device. *Gold Bulletin*, **39(2),** 48-53.

[29] Samarasinghe, S.R., Pastoriza-Santos, I., Edirisinghe, M. J. and Liz-Marzan, L.M. (2008) Fabrication of nanostructured gold films by electrohydrodynamic atomisation. *Applied Physics A -Materials Science & Processing*, **91(1),** 141-147.

[30] Cortie, M.B. (2004) The weird world of nanoscale gold. Gold Bulletin, **37(1-2),** 12-19.

[31] Shim, J.H., Lee, B.J. and Cho, Y.W. (2002) Thermal stability of unsupported gold nanoparticle: a molecular dynamics study. *Surface Science*, **512 (3),** 262-268.

Vol.1, No.2, 151-155 (2009) **Natural Science**
doi:10.4236/ns.2009.12019

A Modified Particle Swarm Optimization Algorithm

Ai-Qin Mu[1,2], De-Xin Cao[1], Xiao-Hua Wang[2]

[1]College of Science, China University of Mining & Technology, XuZhou, China; muaqin@126.com, caodx@cumt.edu.cn
[2]Foundation Departments, Xuzhou Air Force Academy, XuZhou, China

Received 17 August 2009; revised 28 August 2009; accepted 30 August 2009.

ABSTRACT

Particle Swarm Optimization (PSO) is a new optimization algorithm, which is applied in many fields widely. But the original PSO is likely to cause the local optimization with premature convergence phenomenon. By using the idea of simulated annealing algorithm, we propose a modified algorithm which makes the most optimal particle of every time of iteration evolving continuously, and assign the worst particle with a new value to increase its disturbance. By the testing of three classic testing functions, we conclude the modified PSO algorithm has the better performance of convergence and global searching than the original PSO.

Keywords: PSO; Simulated Annealing Algorithm; Global Searching

1. INTRODUCTION

PSO algorithm is a new intelligent optimization algorithm intimating the bird swarm behaviors, which was proposed by psychologist Kennedy and Dr. Eberhart in 1995 [1]. Compared with other optimization algorithms, the PSO is more objective and easily to perform well, it is applied in many fields such as the function optimization, the neural network training, the fuzzy system control, etc.

In PSO algorithm, each individual is called "particle", which represents a potential solution. The algorithm achieves the best solution by the variability of some particles in the tracing space. The particles search in the solution space following the best particle by changing their positions and the fitness frequently, the flying direction and velocity are determined by the objective function.

For improving the convergence performance of PSO, the inertia factor w is used by Shi and Eberhart [2] to control the impact on current particle by former particle's velocity. PSO algorithm has preferred global searching ability when w is relatively large. On the contrary, its local searching ability becomes better when w is smaller. Now the PSO algorithm with inertia weight factor was called standard PSO.

However, in PSO algorithm, particles would lost the ability to explore new domains when they are searching in solution space, that is to say it will entrap in local optimization and causes the premature phenomenon. Therefore, it is very import for PSO algorithm to be guaranteed to converge to the global optimal solution, and many modify PSO algorithms were researched in recent ten years. For example, linearly decreasing inertia weight technique was studied in [3].

In order to solve the premature phenomenon, many modified algorithms based on Simulated Annealing Algorithm are proposed. For example, the new location of all particles is selected according to the probability [4, 5]; the PSO and simulated annealing algorithm are iterated alternatively [6,7]; Gao Ying and Xie Shengli [8] add hybridization and Gaussian mutation to alternative iterations; in [9] particles are divided into two groups, PSO and simulated annealing algorithm are iterated to them respectively and then mixed two algorithms. This paper proposed a new modify PSO algorithm. The arrangement of this paper is as follows. In section 2, the principle of standard PSO is introduced. In section 3, the modified PSO algorithm is described. In section 4, three benchmark functions are used to evaluate the performance of algorithm, and the conclusions are given in section 5.

2. STANDARD PSO ALGORITHM

Assuming $X_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$ is the position of i-th particle in D-dimension, $V_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$ is its velocity which represents its direction of searching. In iteration process, each particle keeps the best position *pbest* found by itself, besides, it also knows the best position *gbest* searched by the group particles, and changes its velocity according two best positions. The standard

formula of PSO is as follow:

$$v_{id}^{k+1} = wv_{id}^{k} + c_1 r_1 (p_{id} - x_{id}^{k}) + c_2 r_2 (p_{gd} - x_{id}^{k}) \quad \textbf{(1)}$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \quad \textbf{(2)}$$

In which: $i = 1, 2, \cdots N$; *N*-the population of the group particles; $d = 1, 2, \cdots, D$; k -the maximum number of iteration; r_1, r_2 -the random values between [0,1], which are used to keep the diversity of the group particles; c_1, c_2 -the learning coefficients, also are called acceleration coefficients; v_{id}^{k} -the number d component of the velocity of particle *i* in *k*-th iterating; x_{id}^{k} -the number d component of the position of particle *i* in *k*-th iterating; p_{id} -the number d component of the best position particle i has ever found; p_{gd} -the number d component of the best position the group particles have ever found.

The procedure of standard PSO is as following:

1) Initialize the original position and velocity of particle swarm;

2) Calculate the fitness value of each particle;

3) For each particle, compare the fitness value with the fitness value of *pbest*, if current value is better, then renew the position with current position, and update the fitness value simultaneously;

4) Determine the best particle of group with the best fitness value, if the fitness value is better than the fitness value of *gbest*, then update the *gbest* and its fitness value with the position;

5) Check the finalizing criterion, if it has been satisfied, quit the iteration; otherwise, return to step 2).

3. THE MODIFIED PSO

In standard PSO, because the particle has the ability to know the best position of the group particles have been searched, we need one particle to find the global best position rather than all particles to find it, and other particles should search more domains to make sure the best position is global best position not the local one. Based on these ideas, we propose some modifications with the standard PSO algorithm. Firstly, the modified algorithm chooses the particle with maximum fitness when it is iterating, initializes its position randomly for increasing the chaos ability of particles. By this means, the particle can search more domains. Secondly, by referring to ideas of the simulated annealing algorithm and using neighborhoods to achieve the guaranteed convergence PSO in [10], it is hoped that the fitness of the particle which has the best value in last iteration would be smaller than last times, and it is acceptable the fitness is worse in a limited extent α . We calculate the change of fitness value of two positions Δf , and accept the new position if Δf is smaller than α . Otherwise, a new position is assigned to the particle randomly from its neighborhood with radius r.

The procedure of modified PSO is as following:

1) Initialize the position and velocity of each particle;

2) Calculate the fitness of each particle;

3) Concern the particle with the biggest fitness value, reinitialize its position; and evaluate the particle with the smallest fitness value whether its new position is acceptable, if the answer is yes, update its position, otherwise, a new position is assigned to the particle randomly in its neighborhood with radius r; then renew the position and velocity of other particles according to **Formula (1)** and **(2)**;

4) For each particle, compare its current fitness value with the fitness of its *pbest*, if the current value is better, then update *pbest* and its fitness value;

5) Determine the best particle of group with the best fitness value, if the current fitness value is better than the fitness value of *gbest*, then update the *gbest* and its fitness value with the position;

6) Check the finalizing criterion, if it has been satisfied, quit the iteration; otherwise, return to step 3).

4. NUMERICAL SIMULATION

For investigating the modified PSO's convergence and searching performance, three benchmark functions are used to compare with standard PSO in this section. The basic information of three functions is described in **Table 1**.

Benchmark function 1 is non-linear single-peak function. It is relatively simple, and mainly used to test the accuracy of searching optimization.

Table 1. Benchmark functions used in experiment.

expression	minimum point	optimal solution
$F_1 = (x_1 - x_2)^2 + ((x_1 + x_2 - 10)/3)^2$	(5,5)	0
$F_2 = 100(x_2 - x_1^{\,2})^2 + (1 - x_1)^2$	(1,1)	0
$F_3 = \sum_{i=1}^{2} [x_i^{\,2} - 10\cos(2\pi x_i) + 10]$	(1,1)	0

Table 2. Results of experiment.

Benchmark function	Algorithm	Total number of iterations	Mean of optimal solution	Minimum of optimal solution	Minimum times of iteration
F_1	standard PSO	20795	4.3708e-6	6.9446e-9	130
	modified PSO	10595	2.1498e-6	1.26e-9	68
F_2	standard PSO	23836	1.7674e-4	1.9403e-8	205
	modified PSO	21989	8.8366e-6	2.52012e-8	350
F_3	standard PSO	24990	0.0667	8.9467e-8	237
	modified PSO	29611	7.7294e-6	9.5000e-9	853

Benchmark function 2 is typical pathological quadratic function which is difficult to be minimized. There is a narrow valley between its global optimum and the reachable local optimum, the chance of finding the global optimal point is hardly. It is typically used to evaluate the implementation of the performance optimization.

Base on sphere function, benchmark function 3 uses cosine function to produce a mounts of local minimum, it is a typical complex multi-peak function which has the massive local optimal point. This function is very easy to make the algorithm into a local optimum not the global optimal solution.

In experiment, the population of group particle is 40; c_1 and c_2 are set to 2; the maximum time of iteration is 10000. It is acceptable if the difference between the best solution obtained by the optimization algorithm and the true solution is less then 1e-6. In standard PSO and modified PSO, the inertia weight is linear decreasing inertia all, which is determined by the following equation:

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{iter_{\max}} \times k$$

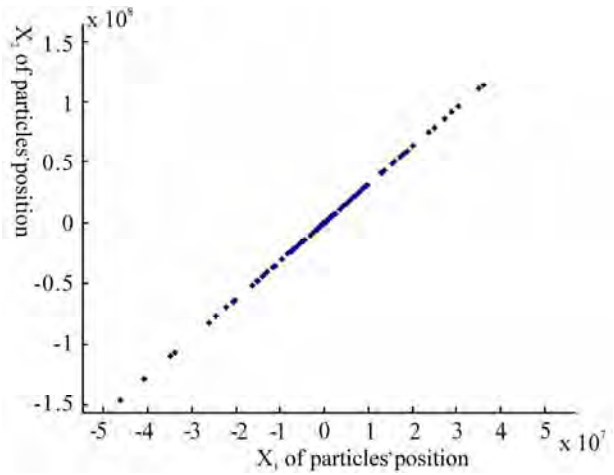


Figure 1. Path of standard PSO's particle.

Where $w_{\max}$ is the start of inertia weight which is set to 0.9, and $w_{\min}$, the end of inertia weight, is set to 0.05. $iter_{\max}$ is the maximum times of iteration; k is the current iteration times. In order to reflect the equity of experiment, two algorithms all use the same original position and velocity randomly generated.

Parameter α which represents the acceptable limited extent of the fitness value is set to 0.5 in modified PSO. For using less parameter, the dynamic neighborhood is used and its radius is set to w. Each experiment is Executed 30 times, and takes their total iteration times and mean optimal solution for comparing. **Table 2** presents the results of experiment.

From **Table 2**, it is easy to find that the modified PSO takes half time as standard PSO to achieve the best solution of function 1. Although the modified PSO has not remarkable improvement in convergence rate from function 2, its mean optimal solution is better than standard PSO, which implies the modified PSO has the better performance in global searching, the conclusion is proved in function 3. Though the total number of iteration of standard PSO is less than modified PSO, but its mean optimal solution is 0.0667, that indicate the rapid convergence of standard PSO is built on running into local optimal. On the contrary, the modified PSO can jump from local optimal successfully, that enhances the stability of algorithm greatly. Observe the results of modified PSO concretely; it can be found that the worst optimal solution of all iterations is 8.5442e-5, which indicates the convergence rate is 100%. The details of 30 loops will no longer run them out.

For observing the movement of particles from benchmark function 3, the standard PSO and the modified PSO are run again by giving the same position and velocity of each particle initialized randomly. The result is that standard PSO has iterated 800 times for the optimal solution to be 1.3781e-6, and the modified PSO has iterated 1216 times for the optimal solution to be 6.3346e-7. A particle is randomly chosen to observe its trace. **Figure 1** and **Figure 2** present the result.

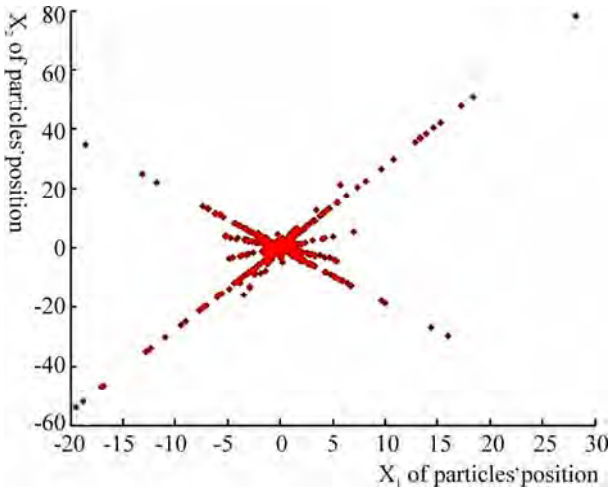


Figure 2. Path of modified PSO's particle.

From **Figure 1** and **Figure 2**, it is easily to find out that the particle of standard PSO was vibrating nearby the optimal position until converging at the optimal position, otherwise, the particle of modified PSO has searched more domains then jumped from the local optimal solution, which ensured the algorithm to converge to the global optimal solution stably.

Generally, the improvement of the modified PSO based on simulated annealing algorithm is applied to all particles. In order to compare the performance of algorithms, we proposed the second improvement to all particles based on the ideas that mentioned before in this article, the main idea is that it is acceptable for all particles when their fitness would be worse in a limited extent α at the next iteration, otherwise, new positions are assigned to the particles randomly from their neighborhood with radius r.

Next, the total iteration times and mean optimal solution are compared between modified PSO and the second improvement from three benchmark functions. The parameters are set as following: the original velocity is 0; other parameters are just same as former case. The experiment is executed 100 times and same original settings are assigned randomly. The results are shown in **Table 3**. If the maximum and minimum velocity of particles are limited to 1 and -1, **Table 4** shows the results.

From **Table 3** and **Table 4**, it is obviously that both modified PSO and the second improvement can jump from local optimal convergence, which means they have the better global searching performance. For function 2 and function 3, the convergence rate of modified PSO is faster than the second improvement. It implied that although the modified PSO do some modifications to two particles' movement, the results is not worse than do the same modifications to all particles', sometimes, it has the better convergence performance. For function 1, when the velocity of particle is not limited, the performance of modified PSO can not compare with the second improvement, but they have the same performance when $V_{\max}$ and $V_{\min}$ are limited. If compare vertically, it can be found that whether the modified PSO or the second improvement have better convergence rate when the velocity of particles is limited. Especially, the second improvement has half times iteration of the modified

Table 3. Performance comparison between modified PSO and the second improvement without limited velocity.

benchmark function	total iteration times		mean optimal solution	
	modified PSO	second improvement	modified PSO	second improvement
F_1	33036	13063	4.3640e-6	1.5878e-6
F_2	72438	86025	1.3661e-5	2.3398e-5
F_3	93506	155573	1.1128e-5	2.7690e-4

Table 4. Performance comparison between modified PSO and the second improvement with limited velocity.

benchmark function	total iteration times		mean optimal solution	
	modified PSO	second improvement	modified PSO	second improvement
F_1	12381	12034	1.7625e-6	9.1580e-7
F_2	37917	60139	8.1302e-6	9.1149e-5
F_3	53453	131291	1.3804e-5	2.4989e-4

PSO. This proved that it is important for PSO to limit the velocity of particles.

5. CONCLUSIONS

In this paper, a modified PSO is proposed based on the simulated annealing algorithm. Through the results achieved in experiments, we can draw following conclusions:

1) The modified PSO has a better performance in stability and global convergence; it is the most important conclusion.

2) Although the modified PSO do some modifications to two particles' position and velocity, but its convergence rate for the multi-peak function is much faster as compared with the second improvement.

3) In modified PSO, the maximum and minimum velocity of particles have obvious impact on the convergence rate. How to choose the appropriate velocity limitation is the next step in our research.

REFERENCES

[1] Kennedy, J. and Eberhart, R.C. (1995) P article swarm optimization. *IEEE International Conference on Neural Network*, 1942-1948.

[2] Shi, Y. and Eberhart, R.C. (1998) A modified particle swarm optimizer. *Proceedings of Congress on Evolutionary Computation*, 79-73.

[3] Shi, Y. and Eberhart, R.C. (1999) Empirical study of particle swarm optimization. Proceedings of the 1999 Congress on Evolutionary Computation, 1945-1950.

[4] Wang, L.Z. (2006) Optimization of the solution to the problems simulated annealing to improve particle swarm algorithm. *Journal of Liuzhou Teachers College*, **21(3),** 101-103.

[5] Gao, S., Yang, J.Y., Wu, X.J. and Liu, T.M. (2005) Particle swarm optimization based on the ideal of simulated annealing algorithm. *Computer Applications and Software*, **22(1),** 103-104.

[6] Wang, Z.S., Li, L.C. and Li, B. (2008) Reactive power optimization based on particle swarm optimization and simulated annealing cooperative algorithm. *Journal of Shandong University (Engineering Science)*, **38(6),**15-20.

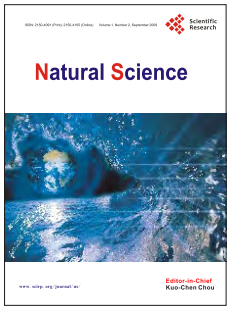
[7] Wang, L.G., Hong, Y., Zhao, F.Q. and Yu, D.M. (2008) A hybrid algorithm of simulated annealing and particle swarm optimization. *Computer Simulation*, **25(11),** 179-182.

[8] Gao, Y. and Xie, S.L. (2004) Particle swarm optimization algorithms based on simulated annealing. *Computer Engineering and Applications*, **40(1),** 47-50.

[9] Pan, Q.K., Wang, W.H. and Zhu, J.Y. (2006) Effective hybrid heuristics based on particle swarm optimization and simulated annealing algorithm for job shop scheduling. Chinese Journal of Mechanical Engineering, **17(10),** 1044-1046.

[10] Peer, E.S., Van den Bergh, F. and Engelbrecht A.P. (2003) Using neighbourhoods with the guaranteed convergence PSO. *Proceeding of the IEEE Swarm Intelligence Symposium*, 235-242.

Call for Papers



Natural Science

A Journal Published by Scientific Research Publishing, USA
www.scirp.org/journal/ns

Natural Science is an international journal dedicated to the latest advancement of natural sciences. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in different areas of natural sciences. All manuscripts must be prepared in English, and are subject to a rigorous and fair peer-review process. Accepted papers will immediately appear online followed by printed hard copy. The journal publishes original papers including but not limited to the following fields:

- **Astronomy & Space Sciences**
  - Astronomy
  - Astrophysics
  - Atmospheric Science
  - Space Physics
- **Earth Science**
  - Geography
  - Geology
  - Geophysics/Geochemistry
  - Oceanography
- **Chemistry**
  - Analytical Chemistry
  - Biochemistry
  - Computational Chemistry
  - Inorganic Chemistry
  - Organic Chemistry
  - Physical Chemistry
- **Life Science**
  - Cell Biology
  - Computational Biology
  - Genetics
  - Immunology
  - Medicine/Diseases
  - Microbiology
  - Molecular Biology
  - Neuroscience
  - Pharmacology/Toxicology
  - Physiology
  - Psychology
  - Virology
- **Physics**
  - Applied Physics
  - Atomic, Molecular, and Optical Physics
  - Biophysics
  - High Energy/Particle Physics
  - Material Science
  - Plasma Physics
- **Others**
  - Education
  - History of Science
  - Science and Innovations

We are also interested in: 1) Short Reports—2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book Reviews—Comments and critiques.

Notes for Intending Authors

Submitted papers should not be previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. For more details, please access the website.

Website and E-Mail

http://www.scirp.org/journal/ns ns@scirp.org

Natural Science, 2009, 1, 63-155
Published Online September 2009 in SciRes (http://www.scirp.org/journal/ns).

TABLE OF CONTENTS

Volume 1, Number 2, September 2009

ISSN: 2150-4091
9772150409001 02