Journal of Theoretical Biology 280 (2011) 10–18

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Self-similarity analysis of eubacteria genome based on weighted graph

Zhao-Hui Qi [a,*], Ling Li [b], Zhi-Meng Zhang [a], Xiao-Qin Qi [a]

[a] College of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, Hebei 050043, People's Republic of China
[b] Basic Courses Department, Zhejiang Shuren University, Hangzhou, Zhejiang 310015, People's Republic of China

ARTICLE INFO

Article history:
Received 18 December 2010
Received in revised form
8 March 2011
Accepted 26 March 2011
Available online 8 April 2011

Keywords:
Biological sequence
Evolution pattern
Social structure
Graphic approach

ABSTRACT

We introduce a weighted graph model to investigate the self-similarity characteristics of eubacteria genomes. The regular treating in similarity comparison about genome is to discover the evolution distance among different genomes. Few people focus their attention on the overall statistical characteristics of each gene compared with other genes in the same genome. In our model, each genome is attributed to a weighted graph, whose topology describes the similarity relationship among genes in the same genome. Based on the related weighted graph theory, we extract some quantified statistical variables from the topology, and give the distribution of some variables derived from the largest social structure in the topology. The 23 eubacteria recently studied by Sorimachi and Okayasu are markedly classified into two different groups by their double logarithmic point-plots describing the similarity relationship among genes of the largest social structure in genome. The results show that the proposed model may provide us with some new sights to understand the structures and evolution patterns determined from the complete genomes.

1. Introduction

The success of human genome project has pushed newly discovered biological sequences to grow in an explosive rate (Chou, 2009). Facing the explosive growth of DNA and protein sequences, experimental, mathematical and graphical approaches have been employed to study the structure, function, evolution and attribution (Chou, 2011) of these sequences. As a powerful tool, the graphical methods can help people gain useful insights in an intuitive or visual manner. Many graphic approaches have been successfully used to study complicated biological topics, such as enzyme-catalyzed reactions (Chou, 1980, 1989, 1990; Chou and Forsen, 1980, 1981; Chou and Liu, 1981; Andraos, 2008), protein folding kinetics (Chou, 1990; Chou and Shen, 2009; Shen et al., 2009), drug metabolism kinetics (Chou, 2010), analysis of codon usage (Zhang and Chou, 1993, 1994), analysis of base frequencies in the anti-sense strands (Chou et al., 1996), prediction of protein subcellular location (Xiao et al., 2006a, b), HBV virus gene missense mutation (Xiao et al., 2005), hepatitis B viral infections (Xiao et al., 2006a, b), G-protein coupled receptors (Lin et al., 2009; Xiao et al., 2009a, 2011) and prediction of protein structure (Xiao et al., 2008a,b, 2009b,c, 2010). Recently, several graphical method applications were also used to examine the similarities/dissimilarities among the coding sequences of different species (Randić et al., 2003a, b; Randić, 2006; Yao et al., 2006, 2008, 2010; Qi et al., 2007; Qi and Qi, 2007, 2009; Qi and Fan, 2007), study cellular signaling networks (Diao et al., 2007), analyze the network structure of the amino acid metabolism (Shikata et al., 2007), provide web-server for protein sequences (Wu et al., 2010), study the fingerprint of SARS coronavirus (Wang et al. 2005; Gao et al., 2006) and discover codon position patterns of eubacteria (Qi and Wei, 2011).

Another graphical method, radar chart, has been used to predict protein subcellular localization (Chou and Elrod, 1999). In addition, radar charts have been applied in a similar manner to classifying organisms (Sorimachi and Okayasu, 2004, 2008a, b; Okayasu and Sorimachi, 2009). In Sorimachi and Okayasu (2004), the amino acid compositions of 11 Gram-positive and 12 Gram-negative eubacteria were determined from their complete genomes. They were classified into two groups, 'S-type' represented by *Staphylococcus aureus* and 'E-type' represented by *Escherichia coli*, based on their patterns of amino acid compositions by radar charts determined from the complete genome. Then, in Qi et al. (2009), the 23 eubacteria were also classified into "S-Type" and "E-Type" by dimensionality reduction method. These results determined by different methods show that the classification about the eubacteria is a reasonable conclusion.

In the present study, we analyze self-similarity characteristics of the 23-eubacteria genomes based on weighted graph theory. Similarity is a concept from signal similarity theory. It describes the similarity degree when one signal is compared with another signal. Here, the concept about similarity degree of signals is used to describe the similarity degree between two genes. Then we obtain the similarity information between any pair of genes in a genome when every gene of the genome is compared with other genes in the same genome. The overall statistical results about similarity degree are called as self-similarity characteristic of genome. The self-similarity

* Corresponding author.
E-mail address: zhqi_yh2004@yahoo.com.cn (Z.-H. Qi).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.03.033

analysis of genome can discover some interesting biological information hidden in genome. Generally, gene evolution includes three main approaches: gene point mutation, gene recombination and horizontal gene transfer (Syvanen and Kado, 2002). The same result from different evolution approaches is that every gene of genome perhaps has its similar genes within the same genome. If one gene has many similar genes within genome, it is relatively active in its evolution pathway. Otherwise, this gene is comparatively conservative. Then, how many active genes does a genome have? What differences do there exist among different genomes? Here, we introduce weighted graph method to investigate these problems about the 23 eubacteria genome studied in Sorimachi and Okayasu (2004) and Qi et al. (2009). The rest of the paper is organized as follows. Section 2 presents the self-similarity analysis model of genome based on weighted graph and some statistical information of this model. Section 3 mainly discusses some quantified statistical variables about the self-similarity characteristic of different genomes and several distributions about the statistical variables. Section 4 gives the conclusion of this paper.

2. Materials and methods

2.1. Materials

Complete genome sequences were downloaded from NCBI GenBank (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). In Sorimachi and Okayasu (2004), these bacteria are grouped into two main classes: *S. aureus* ‘S-Type’ and *E. coli* ‘E-type’, based on their genomic structures through different classifying schemes. In this paper, we will investigate the self-similarity characteristics of these bacteria genomes to reveal the possible evolution patterns about genome itself. The genome sequences used for this study are summarized in Table 1.

2.2. Methods

2.2.1. Weighted graph theory

In the past decades, the idea of graph has turned out to be widespread in computer science, and physical, biological, social and man-made systems. In graphs, each element is treated as a vertex (or node, or point). The links (or connections) between vertexes denote their interactions and correlations. Recently, two seminal models about graph theory, “small-world” network (Watts and Strogatz, 1998) and “BA” network (Barabási and Albert, 1999) were proposed. The models brought about the naissance of complex networks theory as a new branch about graph theory. In the past ten years, the new branch received an increasing attention. Here, we introduce weighted graph theory to investigate the self-similarity characteristics of the 23 eubacteria genomes. In a weighted graph, a specific value is associated with each connection and the value describes the strength of the connection. An adjacency matrix corresponding to a weighted graph is a weight matrix. The weight is used to describe the strength of each different links. Of course, 0 is set to stand for nothing if there is no connection.

2.2.2. Self-similarity analysis model of genome based on weighted graph

As for a genome, there are hundreds of genes. Large genome even has tens of thousands of gene sequences. These sequences such as nucleotide or amino acid from GenBank are commonly indexed by their sequence IDs. In the proposed model, each gene of a genome is used to construct the weighted graph corresponding to the genome.

Firstly, in a graph a node associated with a sequence ID is used to identify a specific gene. The connection (or link) between any pair of nodes is represented as their interaction and correlation. As for a genome with n genes, the number of the vertexes and the links are n and $C_n^2=(n(n-1)/2)$, respectively. Fig. 1 shows how a graph is constructed according to a genome, where G_i denotes the ith gene, $i=1,2,\ldots,n$. Fig. 1 is a full-connected graph because there is a connection between any pair of genes. Then a specific value is assigned to each connection to describe the similarity degree between two genes. There are several methods such as BLAST (Altschul et al., 1997; http://blast.ncbi.nlm.nih.gov/bl2seq/wblast2.cgi) and Smith–Waterman algorithm-needle (Needleman and Wunsch, 1970; http://www.ebi.ac.uk/Tools/emboss/align/index.html) to calculate the similarity of two genes. The schemes ensure the optimal local alignment by exploring all possible alignments and choosing the best. The gap insertion penalty, gap extension penalty and substitution matrix used to calculate the alignments are specified. We do not choose the global alignment algorithms because they perhaps

Table 1
Genomes used for this study.

Strain	*Accession (GenBank)*	*RefSeq identifier*	*Total length (bp)*	*Genes*
Staphylococcus aureus Mu50	BA000017.4	NC_002758	2,878,529	2775
Streptococcus pyogenes M1	AE004092.1	NC_002737	1,852,441	1811
Bacillus subtilis	AL009126.2	NC_000964	4,214,630	4225
Clostridium perfringens 13	BA000016.3	NC_003366	3,031,430	2786
Listeria monocytogenes	AL591824.1	NC_003210	2,944,528	2940
Mycoplasma pulmonis	AL445566.1	NC_002771	963,879	815
Mycoplasma genitalium	L43967.2	NC_000908	580,076	525
Mycoplasma pneumoniae	U00089.2	NC_000912	816,394	733
Ureaplasma urealyticum	CP001184.1	NC_011374	874,478	692
Mycobacterium tuberculosis	AE000516.2	NC_002755	4,403,837	4293
Mycobacterium leprae	AL450380.1	NC_002677	3,268,203	2770
Rickettsia prowazekii	AJ235269.1	NC_000963	1,111,523	886
Borrelia burgdorferi	AE000783.1	NC_001318	910,724	875
Campylobacter jejuni	CP000538.1	NC_008787	1,616,554	1707
Helicobacter pylori 26695	AE000511.1	NC_000915	1,667,867	1630
Helicobacter pylori J99	AE001439.1	NC_000921	1,643,831	1535
Escherichia coli	U00096.2	NC_000913	4,639,675	4467
Salmonella typhi	AL513382.1	NC_003198	4,809,037	4711
Vibrio cholerae	AE003852.1	NC_002505	2,961,149	2889
	AE003853.1	NC_002506	1,072,315	1119
Yersinia pestis	AL590842.1	NC_003143	4,653,728	4103
Neisseria meningitidis	AL157959.1	NC_003116	2,184,406	2065
Haemophilus influenzae	L42023.1	NC_000907	1,830,138	1789
Treponema pallidum	AE000520.1	NC_000919	1,138,011	1095

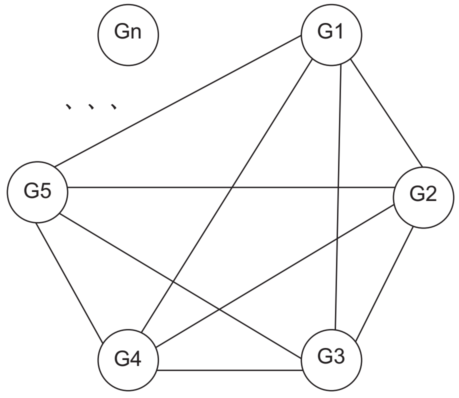

Fig. 1. A full-connected and weighted graph.

produce some results but much of the alignments may have little or no biological significance. Here, we use BLAST, a local alignment algorithm, to calculate the similarity degree of two genes by the consideration for speed and efficiency.

The similarity degree of any pair of genes in genome is determined by BLAST (bl2seq) (http://blast.ncbi.nlm.nih.gov/bl2seq/wblast2.cgi) for pairwise protein–protein sequence comparison. The detailed view for each segment of alignment includes the statistics with the percentage of identities, positives and gaps, schematic view, and the text alignment view. We can obtain the detailed statistics information for two closely related sequences by BLAST (bl2seq) with a given expectation value (E-value). As for two distantly related sequences, we cannot possibly get any statistics information. There are no similar areas between the two sequences. In order to simplify the graph structure, we remove the links with zero value from the weighted graph. Similarly, the node is also removed from the graph if a node has no any links connected to other nodes. Then in the graph, every link is associated with a real value derived from the detailed statistics information by BLAST (bl2seq) with a given E-value. The graph becomes a weighted graph related to a genome. For example, Fig. 2(a) shows all links of the gene "gi|3322290" of *Treponema pallidum* connected to other genes by BLAST (bl2seq) with E-value 0.001. We extract the E-value from the detailed statistics information as the weighted value. Fig. 2(b) illustrates the overall weighted graph related to *Treponema pallidum* by Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). In order to clarify the figure, in Fig. 2(b) we do not illustrate the E-values related to links.

A close look at Fig. 2 reveals that the main characteristics among genes show clustering features. Some small social structures only have several genes. However, most of all genes are clustered in a very large social structure. From the biological point of view, these clustering features reveal some insights about gene evolution. Gene evolution includes three main approaches: gene point mutation, gene recombination and horizontal gene transfer (Syvanen and Kado, 2002). Obviously, the accumulation of point mutation and gene recombination are the important reasons leading to large social structure. By the accumulation of point mutation and gene recombination, the genes in structure obtain similarity relation with others.

However, all genes in Fig. 2 only are a part of *Treponema pallidum* genome. There are 609 genes in Fig. 2. The total number of genes in the genome is 1095. There approximately exist 44.4% of the genes unrelated to other genes in similarity. These genes are relatively conservative in genome. They rarely interact with others in the evolution pathway, or do not generate new genes by mutation. Next, we will continue to discuss the following problems. How many relatively conservative genes does every different genome have? How do we determine the degree of similarity among those relatively active genes in the evolution pathway?

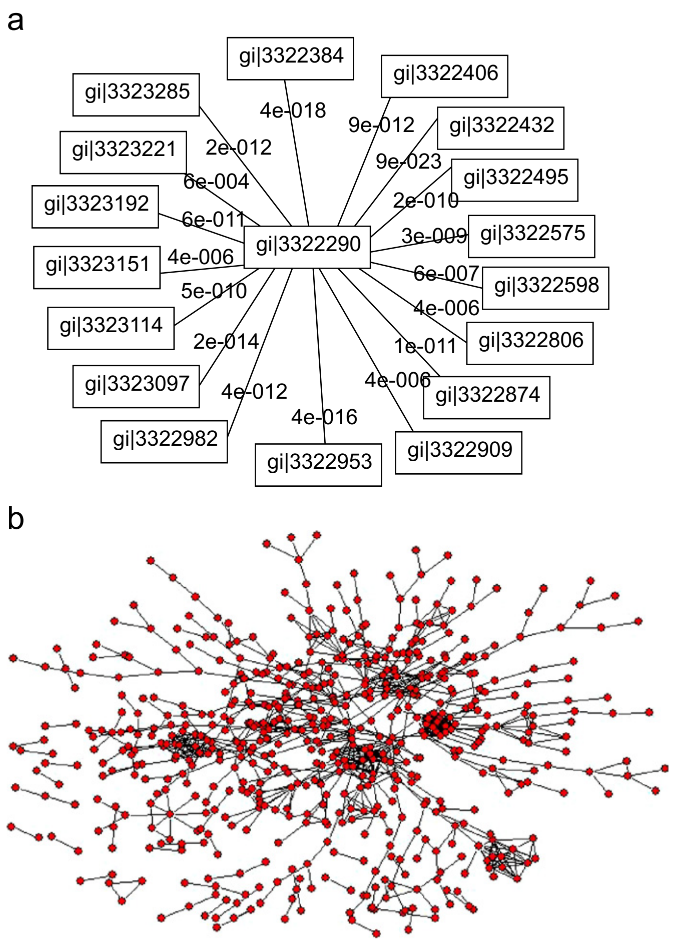

Fig. 2. An example for a weighted graph related to *Treponema pallidum*. (a) All links of the gene "gi|3322290" of *Treponema pallidum* connected to other genes. The E-value form the detailed statistics information is used as the weighted value. (b) The overall weighted graph related to *Treponema pallidum*.

2.2.3. Some statistical information about self-similarity analysis model of genome

In Fig. 2, we illustrate the graph structure of *Treponema pallidum* genome. The clustering feature is the main characteristic. The E-value derived from the detailed statistics information by BLAST (bl2seq) is used to describe the similarity degree between genes. It is well known that E-value is a statistical calculation based on the score that gives the number of hits of this score that this search would return by chance using a database of this size. According to BLAST tutorial, if the E-value is less than 1×10^{-50}, the hit is very similar to the query sequence and is very likely to be evolutionarily related. If the E-value is between 1×10^{-50} and 1×10^{-2}, the hit has some similarity to the query sequence and may be related. The E-values in the range can indicate that the sample sequence is in the same family as the hit or it may have closely related functional domains. Here we set the E-value less than 1×10^{-3} to obtain results that are more reasonable.

Now, we obtain the detailed statistics information for two sequences by BLAST (bl2seq) with a given E-value threshold. We choose the best High-scoring segment pair (HSP) when there are many hits in an alignment. The reported results still include other outcomes besides E-value such as Score, Identities and Positives. Obviously, only the E-value does not fully reflect the similarity degree between genes. Here, we analyze the reported results by BLAST (bl2seq) in order to obtain a weighted value (*w*-value) to describe the similarity degree. Table 2 lists some alignment results of the gene "gi|3322290" of *Treponema pallidum* by BLAST (bl2seq) with E-value threshold 0.001. In Table 2 L_1 and L_2 are the length of Gene1 and

Table 2
Some alignment results of the gene "gi|3322290" of *Treponema pallidum* by BLAST (bl2seq) with E-value 0.001.

Gene1 ID	*Gene2 ID*	*L1*	*L2*	*Score*	*Expect*	*Identities*	*Positives*
gi|3322290	gi|3322384	238	269	73.6	4e−018	49/192(0.25)	86/192(0.44)
gi|3322290	gi|3322406	238	220	52.0	9e−012	47/194(0.24)	85/194(0.43)
gi|3322290	gi|3322432	238	266	89.0	9e−023	66/203(0.32)	106/203(0.52)
gi|3322290	gi|3322495	238	255	48.1	2e−010	47/200(0.23)	85/200(0.42)
gi|3322290	gi|3322575	238	516	45.1	3e−009	31/92(0.33)	49/92(0.53)
gi|3322290	gi|3322598	238	533	37.4	6e−007	31/104(0.29)	54/104(0.51)
gi|3322290	gi|3322806	238	960	35.4	4e−006	27/84(0.32)	40/84(0.47)
gi|3322290	gi|3322874	238	226	51.6	1e−011	52/207(0.25)	87/207(0.42)

Gene2, respectively. The denominator of Identities or Positives is the length of similar region in the hit.

Then we define several new parameters. Let L_{region} denote the length of similar region of two genes. Let S_{score}, $P_{positive}$ and E_{value} be the score, the positive and E-value, respectively. As for the default choice of BLAST, the E-value 1×10^{-10} is used as the E-value threshold. Here, the new parameter E_{value} is defined as 0.8 when the E-value ranges from 0.001 to 1×10^{-10}. If the E-value is more than 1×10^{-10}, the value is set to "1". There are two reasons: (i) When the E-value of BLAST program is small, the similarity degree between genes is better. This is not consistent with other parameters and our customary understanding; (ii) The E-value is the probability due to chance, that there is another alignment with a similarity greater than the given score S_{score}. Here, our definition is a normalization process. The value "1" denotes the probability is very satisfactory. The value "0.8" means the probability is also acceptable.

Then we propose a new parameter W_{value} to describe the similarity degree of two genes according to the main statistics parameters in the reported results by BLAST (bl2seq). The W_{value} is $S_{score}P_{positive}E_{value}(L_{region}/\min\{L_1,L_2\})$. The parameter comprehensively integrated with various statistics information can fully reflect the reported results by BLAST. Of course, people often use the E-value derived from the detailed statistics information to describe the similarity degree of genes. Here, the proposed parameter W_{value} includes the various statistics information and shows the overall result of an alignment. This consideration possibly provides a directly quantitative comparison for the similarity degree of genes. It can also avoid the inconvenience in describing the quantitative similarity comparison by the E-value. Then by the parameter W_{value} we can obtain a weighted graph to describe closely the self-similarity characteristics of a genome.

In the next section, we will discuss in detail some statistical information related to weighted graph to describe the self-similarity characteristic of different genomes.

3. Results and discussion

3.1. Some quantified statistical variables about the self-similarity characteristic of different genomes

In the above section, Fig. 2 reveals that the main characteristics among genes show clustering features. The genes are clustered into many social structures. These social structures have different scales. Here, as for every genome of Table 1 we calculate some quantified statistical variables. To simplify the description in Table 3, let N_S be the number of social structures. Let $N_{S\text{-}genes}$ and $P_{S\text{-}genes}$ be the total number of genes in all social structures and their proportion in genome, respectively. Let $N_{LS\text{-}genes}$ and $P_{LS\text{-}genes}$ be the total number of genes in the largest social structure and their proportion in genome, respectively. Let $N_{LS\text{-}links}$ and $A_{LS\text{-}links}$ be the total number of links in the largest social structure and the average of links of every gene.

From Table 3 we can get some interested conclusions:

(i) From the fourth and the fifth column of the table, we find out that the genes clustered into social structures occupy the majority proportion of complete genome. The proportion $P_{S\text{-}genes}$ of clustering genes of 17 species in the table are more than 70%. The proportion in *Clostridium perfringens* 13 has the highest percentage 92%. There are only three species with a percentage less than 60%. These data indicate that most of genes in genome are relatively active in their evolution pathways. As for a genome, it is impossible for point mutation to reach so large social structures. The statistic data give a possible explanation that some genes continuously exchange some segments each other. Perhaps segment reassortment (or recombination) and association in genome, that is to say, are the main evolutionary approaches.

(ii) In the sixth, seventh, eighth and ninth column of the table, we calculate the statistic information about the largest one in all social structures. The statistic data shows that the genes in the largest social structure occupy the majority proportion of complete genome. The other social structures only occupy a small percentage in genome except for the largest social structure. The statistic data shows that most of genes in genome are associated with each other. The ninth column in the table illustrates the average number of links of every gene in the largest social structure. Obviously, the genes are more active when their links are high. The species with the largest number of connections is *Clostridium perfringens* 13 in which the average number of links reaches 6.96. The average links can discover the comprehensive trend about relation among different genes. In the following section, we will further reveal more details about the largest social structure such as degree distribution and edge-weight distribution.

3.2. Several distributions about statistical variables in the largest social structure

In this section, we use several statistical variables in weighted graph (Watts and Strogatz, 1998; Barabási and Albert, 1999), to analyze the self-similarity characteristics of genome. Here we only discuss the statistical characteristics in the largest social structure instead of all social structures because the genes in the largest social structure occupy the majority proportion of complete genome as shown in Table 3.

3.2.1. Degree distribution

A gene's degree in genome is a connection degree granted to the individual gene that it denotes the number of other genes associated with the gene by BLAST (bl2seq) with E-value 1×10^{-3}. A gene with a large degree means that it may be a more active gene. Some segments in the gene may be transferred to other genes, or come

Table 3
Some quantified statistical variables; N_S: the number of social structures; $N_{S\text{-}genes}$: the total number of genes in all social structures; $P_{S\text{-}genes}$: the proportion of $N_{S\text{-}genes}$ in genome; $N_{LS\text{-}genes}$: the total number of genes in the largest social structure; $P_{LS\text{-}genes}$: the proportion of $N_{LS\text{-}genes}$ in genome; $N_{LS\text{-}links}$: the total number of links in the largest social structure; $A_{LS\text{-}links}$: the average of links of every gene in the largest social structure.

Strain	*Genes*	N_S	$N_{S\text{-}genes}$	$P_{S\text{-}genes}$	$N_{LS\text{-}genes}$	$P_{LS\text{-}genes}$	$N_{LS\text{-}links}$	$A_{LS\text{-}links}$
Staphylococcus aureus Mu50	2775	12	2533	0.91	2505	0.90	10822	4.32
Streptococcus pyogenes M1	1811	51	1379	0.76	1264	0.70	3615	2.86
Bacillus subtilis	4225	16	3845	0.91	3812	0.90	18850	4.94
Clostridium perfringens 13	2786	4	2577	0.92	2571	0.92	17891	6.96
Listeria monocytogenes	2940	12	2679	0.91	2657	0.90	11665	4.39
Mycoplasma pulmonis	815	3	702	0.86	698	0.86	4493	6.44
Mycoplasma genitalium	525	20	335	0.64	295	0.56	646	2.19
Mycoplasma pneumoniae	733	32	493	0.67	403	0.55	1555	3.86
Ureaplasma urealyticum	692	5	560	0.81	550	0.79	2647	4.81
Mycobacterium tuberculosis	4293	12	3901	0.91	3879	0.90	22217	5.73
Mycobacterium leprae	2770	44	1253	0.45	1149	0.41	2290	1.99
Rickettsia prowazekii	886	20	622	0.70	573	0.65	1074	1.87
Borrelia burgdorferi	875	4	754	0.86	748	0.85	2979	3.92
Campylobacter jejuni	1707	9	1497	0.88	1481	0.87	4860	3.28
Helicobacter pylori 26695	1630	16	1323	0.81	1291	0.79	3634	2.81
Helicobacter pylori J99	1535	20	1265	0.82	1225	0.80	3473	2.84
Escherichia coli	4467	34	3793	0.85	3712	0.83	16643	4.48
Salmonella typhi	4711	45	3987	0.85	3892	0.83	15688	4.03
Vibrio cholerae	4008	59	2238	0.56	2101	0.52	6849	3.26
Yersinia pestis	4103	54	3486	0.85	3351	0.82	20108	6.00
Neisseria meningitidis	2065	92	1483	0.72	1243	0.60	2126	1.03
Haemophilus influenzae	1789	68	1316	0.74	1145	0.64	2546	2.22
Treponema pallidum	1095	82	609	0.56	352	0.32	695	1.97

from other genes by gene recombination. Let k_i be the degree of node (gene) i in the largest social structure. Then we can use distribution function $P(k)$ to describe the degree distribution of node in the graph. The distribution function $P(k)$ denotes that the probability of the node degree is just k as for a randomly chosen node. The patterns of the distribution function $P(k)$ based on the complete genomes of various eubacteria are 'double logarithmic point-plot', as shown in Fig. 3.

Double logarithmic point-plot: The logarithm of a number to a given base is the exponent to which the base must be raised in order to produce that number. For example, the logarithm of 1000 to base 10 is 3, because three factors of 10 must be multiplied to yield a thousand: $10 \times 10 \times 10$, also written as 10^3, equals 1000. The double logarithmic point-plot is a figure in which all data points are plotted on the logarithmic scales on both axes. Here, the distribution function $P(k)$ possibly ranges from 10^{-4} to 1. The parameter k ranges from 1 to 10^3. The data-points plotted on linear scales (Cartesian coordinate system) would be very close to the axes. Therefore, for sake of comparison, all data points are plotted on the logarithmic scales on both axes, as shown in Fig. 3.

According to the scale-free network BA model (Barabási and Albert, 1999), the power law phenomenon about the degree distribution of node in double logarithmic coordinates is one of the most important characteristics of the model. Based on the rule eubacteria can be mainly classified into two groups by the double logarithmic point-plots calculated from their complete genome. The pattern of double logarithmic point-plot of *Bacillus subtilis* resembles that of *Escherichia coli*, compared with that of *Campylobacter jejuni*. The pattern of the double logarithmic point-plot of *Campylobacter jejuni* is similar to the power law distribution. The eubacteria similar to *Campylobacter jejuni* include *Helicobacter pylori J99*, *Mycobacterium tuberculosis*, *Rickettsia prowazekii* and so on, as shown in Fig. 3. However, the other pattern of the double logarithmic point-plot has a different appearance, an upward tail at the end of the point-plot. The eubacteria with this special pattern include *Bacillus subtilis*, *Escherichia coli*, *Listeria monocytogenes*, *Salmonella typhi* and so forth. In order to give a clear comparison among different point-plots in Fig. 3, we draw a fitting power-law curve ck^{-b} without considering the upward tail, where c and b are constant and power law exponent, respectively. Then we give the power law exponent b of each figure.

In fact, in Sorimachi and Okayasu (2004) and Qi et al. (2009) the 23 eubacteria are markedly classified into two main groups: "S-Type" represented by *Staphylococcus aureus Mu50* and "E-Type" represented by *Escherichia coli*. The classification into two groups by different approaches shows the evolutionary distance among different eubacteria. Here we draw a different conclusion. Our classification into two groups by the double logarithmic point-plot cannot show their evolution distances. However, we can discover the evolution patterns among the different eubacteria. The following are the details:

(i) If ignoring the upward tails of some curves, all double logarithmic point-plots in Fig. 3 are similar to the power law distribution. This shows that the relations among genes of the largest social structure are neither random nor homogeneous. They show some features similar to "small world". From a biological view, this reflects on a possible evolution pattern. Most of genes are relatively conservative and merely have a little chance to interact with others in their evolution pathways. However, a few genes with very high degree perhaps exchanged gene pieces with quite a number of other ones.

(ii) Some double logarithmic point-plots of Fig. 3, such as *Bacillus subtilis* and *Escherichia coli*, have upward tails at the end of the curve. This shows that there is a high increase in the number of genes with very high degree. This increase is so high that the corresponding dot-set of the double logarithmic point-plot departs from the power law distribution. These genes are quite active and perhaps exchanged gene pieces with quite a number of other ones in their evolution pathways.

3.2.2. Edge-weight distribution

A gene's edge-weight in genome is the value of the parameter W_{value}. The parameter comprehensively integrated with various statistics information can fully reflect the similarity degree between genes. An edge with a large weight means that the pair of genes is

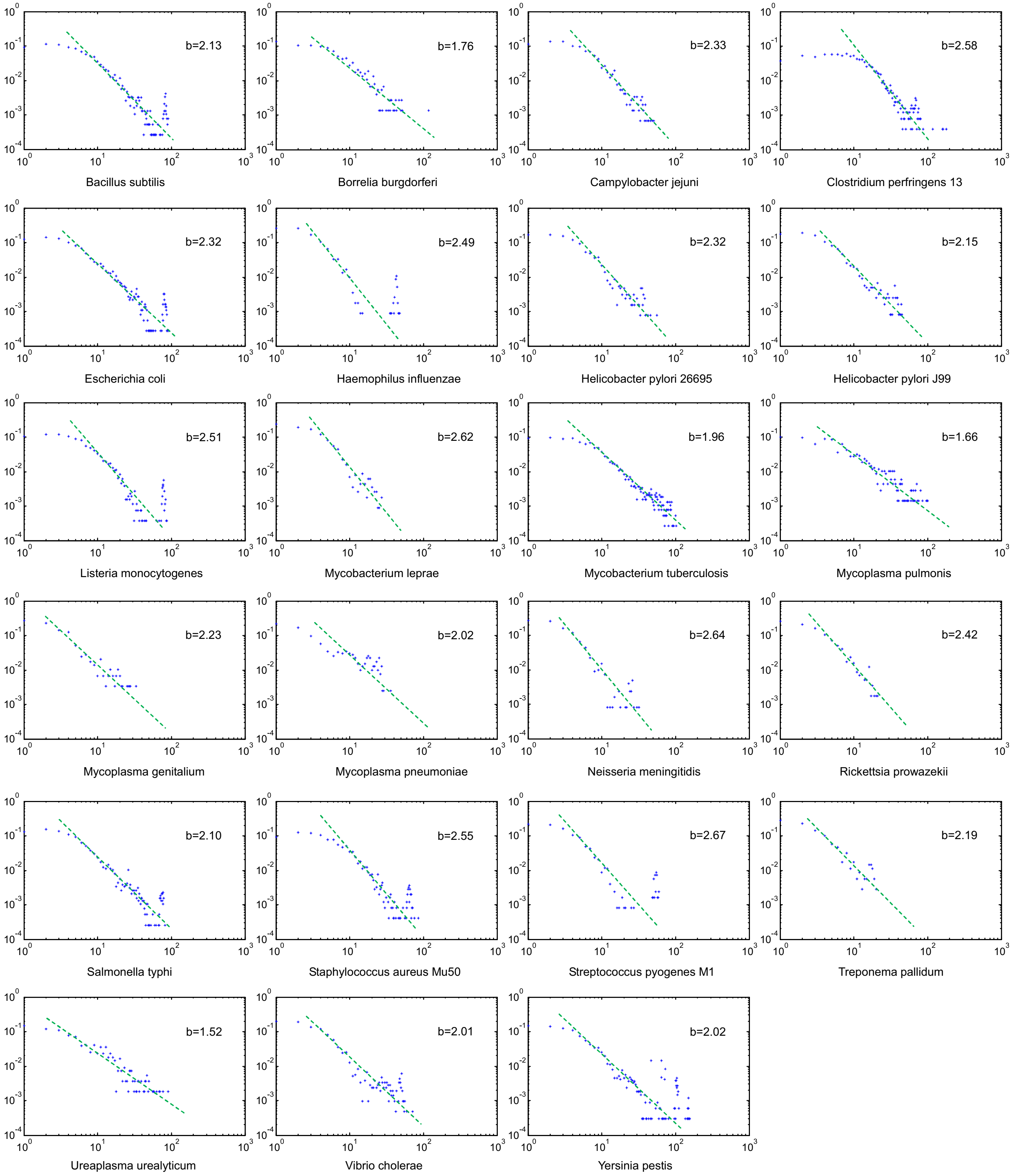

Fig. 3. Double logarithmic point-plot of degree distribution function $P(k)$ in the largest social structure of genome and the fitting power-law curve ck^{-b}.

more similar to each other. Here, let w_i be the weight of edge (between two genes) i in the largest social structure. Like the analysis of degree distribution, we also use distribution function $P(w)$ to describe the edge-weight distribution in the network. The distribution function $P(w)$ denotes that the probability of the edge-weight is just w as for a randomly chosen edge. The patterns of the distribution function $P(w)$ based on the complete genomes in Table 1 are ‘double logarithmic point-plot’, as shown in Fig. 4. At the same time, we still give the fitting power-law curve ck^{-b} and the power law exponent b of each figure.

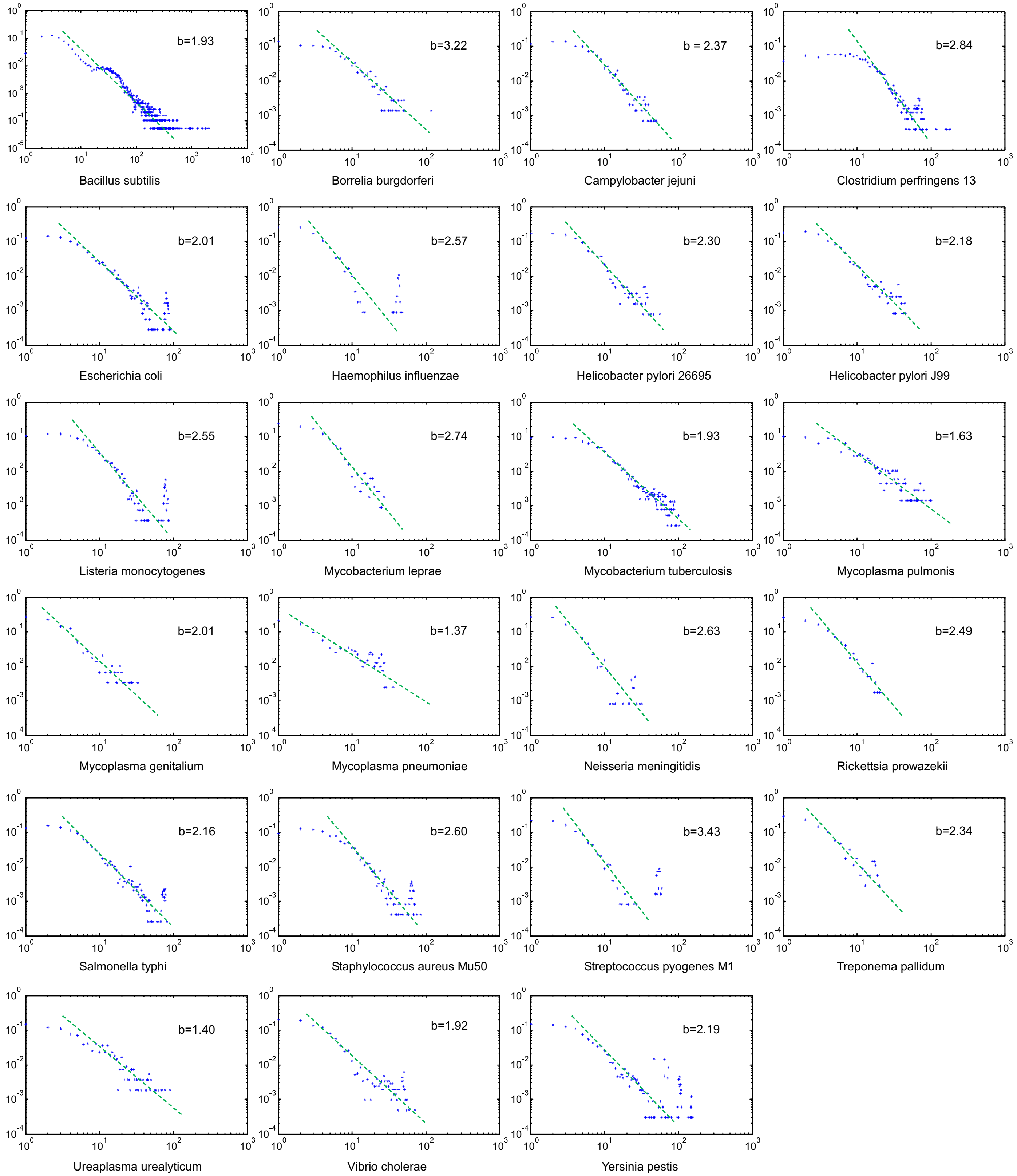

Fig. 4. Double logarithmic point-plot of edge-weight distribution function $P(w)$ in the largest social structure of genome and the fitting power-law curve ck^{-b}.

Like Fig. 3, the eubacteria can be also mainly classified into the same two groups: the pattern similar to the power law distribution and the pattern with an upward tail at the end of the curve. The eubacteria similar to the power law distribution include *Campylobacter jejuni Helicobacter pylori J99*, *Mycobacterium tuberculosis*, *Rickettsia prowazekii* and so on, as shown in Fig. 4. The eubacteria with an upward tail at the end of the curve include *Escherichia coli*, *Listeria monocytogenes*, *Salmonella typhi* and so forth. A close look at Figs. 3 and 4 shows that only one species, *Bacillus subtilis*, shows different pattern in the two figures.

In Fig. 3 *Bacillus subtilis* has a pattern with an upward tail at the end of the curve. However, in Fig. 4 the species has a pattern with an approximation of the power distribution instead of the upward tail in Fig. 3. According to the pattern shown in Fig. 3, we know that there are many genes with very high degree. This shows that these genes are relatively active and perhaps exchanged gene pieces with quite a number of other ones in their evolution pathways. Yet the pattern as shown in Fig. 4 means that the similarities between the genes are low. There are two possible reasons to explain the special phenomenon about *Bacillus subtilis*. One is that the scale of the exchanged gene pieces between different genes is relatively small. The other is that it perhaps takes so much time that the further evolution reduces the similarity degree among the exchanged gene pieces.

4. Conclusions

Here we propose a self-similarity analysis model of genome based on weighted graph. In this model, a genome is attributed to a weighted graph to express the relationship among similar genes in the same genome. Then the weighted graphs are used to investigate the self-similarity characteristics of different genomes from 23 eubacteria. Unlike the studies of the 23 eubacteria in Sorimachi and Okayasu (2004) and Qi et al. (2009), here they are markedly classified into two different groups by their double logarithmic point-plots describing the similarity relations among genes of the largest social structure in genome. One group is the eubacteria similar to the power law distribution, such as *Campylobacter jejuni* and *Helicobacter pylori J99*. The other includes the eubacteria with an upward tail at the end of the double logarithmic point-plot, such as *Escherichia coli* and *Staphylococcus aureus Mu50*. The results show that the proposed model may provide us with some new sights to understand the structures and evolution patterns determined from the complete genomes.

Acknowledgment

We thank the anonymous reviewers for their valuable comments to improve this paper. This work is supported by Natural Science Foundation of Hebei Province Education Department (Project No. 2010276), and the Zhejiang Provincial Natural Science Foundation of China (No. Y6100339)

The authors declare that they have no conflict of interest.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402.

Andraos, J., 2008. Kinetic plasticity and the determination of product ratios forkinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Canadian Journal of Chemistry 86, 342–357.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

Chou, K.C., 1980. A new schematic method in enzyme kinetics. European Journal of Biochemistry 113, 195–198.

Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics. Journal of Biological Chemistry 264, 12074–12079.

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.

Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6, 262–274.

Chou, K.C., 2010. Graphic rule for drug metabolism systems. Current Drug Metabolism 11 (4), 369–378.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273 (1), 236–247.

Chou, K.C., Elrod, D.W., 1999. Protein subcellular location prediction. Protein Engineering 12, 107–118.

Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochemical Journal 187, 829–835.

Chou, K.C., Forsen, S., 1981. Graphical rules of steady-state reaction systems. Canadian Journal of Chemistry 59, 737–755.

Chou, K.C., Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. Journal of Theoretical Biology 91, 637–654.

Chou, K.C., Shen, H.B., 2009. FoldRate: a web-server for predicting protein folding rates from primary sequence. The Open Bioinformatics Journal 3, 31–50 Accessible at : 〈http://www.bentham.org/open/tobioij/〉.

Chou, K.C., Zhang, C.T., Elrod, D.W., 1996. Do antisense proteins exist? Journal of Protein Chemistry 15, 59–61.

Diao, Y., Li, M., Feng, Z., Yin, J., Pan, Y., 2007. The community structure of human cellular signaling network. Journal of Theoretical Biology 247, 608–615.

Gao, L., Ding, Y.S., Dai, H., Shao, S.H., Huang, Z.D., Chou, K.C., 2006. A novel fingerprint map for detecting SARS-CoV. Journal of Pharmaceutical and Biomedical Analysis 41, 246–250.

Lin, W.Z., Xiao, X., Chou, K.C., 2009. GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. Protein Engineering, Design and Selection 22, 699–705.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48, 443–453.

Needleman S.B., Wunsch C.D., EMBOSS - Needle, 〈http://www.ebi.ac.uk/Tools/emboss/align/index.html〉.

Okayasu, T., Sorimachi, K., 2009. Organisms can essentially be classified according to two codon patterns. Amino Acids 36 (2), 261–271.

Pajek Batagelj, V., 2010. 〈http://vlado.fmf.uni-lj.si/pub/networks/pajek/〉.

Qi, X.Q., Wen, J., Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. Journal of Theoretical Biology 249, 681–690.

Qi, Z.H., Qi, X.Q., 2007. Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chemical Physics Letters 440, 139–144.

Qi, Z.H., Fan, T.R., 2007. PN-curve: a 3D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters 442, 434–440.

Qi, Z.H., Qi, X.Q., 2009. Numerical characterization of DNA sequences based on digital signal method. Computers in Biology and Medicine 39, 388–391.

Qi, Z.H., Wang, J.M., Qi, X.Q., 2009. Classification analysis of dual nucleotides using dimension reduction. Journal of Theoretical Biology 260, 104–109.

Qi, Z.H., Wei, R.Y., 2011. A combination dimensionality reduction approach to codon position patterns of eubacteria based on their complete genomes. Journal of Theoretical Biology 272, 26–34.

Randić, M., Vracko, M., Lers, N., Plavsic, D., 2003a. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chemical Physics Letters 371, 202–207.

Randić, M., Vracko, M., Lers, N., Plavsic, D., 2003b. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters 368, 1–6.

Randić, M., 2006. Spectrum-like graphical representation of DNA based on codons. Acta Chimica Slovenica 53, 477–485.

Shen, H.B., Song, J.N., 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 Accessible at: 〈http://www.srpublishing.org/journal/jbise/〉.

Shikata, N., Maki, Y., Noguchi, Y., et al., 2007. Multi-layered network structure of amino acid (AA) metabolism characterized by each essential AA-deficient condition. Amino Acids 33, 113–121.

Sorimachi, K., Okayasu, T., 2004. Classification of eubacteria based on their omplete genome: where does Mycoplasmataceae belong? Biology Letters 271, S127–S130.

Sorimachi, K., Okayasu, T., 2008a. Universal rules governing genome evolution expressed by linear formulas. Open Genomics Journal 1, 33–43.

Sorimachi, K., Okayasu, T., 2008b. Codon evolution is governed by linear formulas. Amino Acids 34 (4), 661–668.

Syvanen, M., Kado, C.I., 2002. Horizontal Gene Transfer second ed. Academic Press, New York.

Tatiana, A.T. Thomas, L.M., BLAST (bl2seq), 〈http://blast.ncbi.nlm.nih.gov/bl2seq/wblast2.cgi〉.

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

Wu, Z.C., Xiao, X., Chou, K.C., 2010. 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. Journal of Theoretical Biology 267, 29–34.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684), 409–410.

Xiao, X., Lin, W.Z., Chou, K.C., 2008a. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. Journal of Computational Chemistry 29, 2018–2024.

Xiao, X., Lin, W.Z., 2009c. Application of protein grey incidence degree measure to predict protein quaternary structural types. Amino Acids 37 (4), 741–749.

Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C., 2006a. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30, 49–54.

Xiao, X., Shao, S.H., Chou, K.C., 2006b. A probability cellular automaton model for hepatitis B viral infections. Biochemical and Biophysical Research Communications 342, 605–610.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005. Anapplication of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. Journal of Theoretical Biology 235, 555–565.

Xiao, X., Wang, P., Chou, K.C., 2008b. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. Journal of Theoretical Biology 254, 691–696.

Xiao, X., Wang, P., Chou, K.C., 2009a. GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. Journal of Computational Chemistry 30, 1414–1423.

Xiao, X., Wang, P., Chou, K.C., 2009b. Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. Journal of Applied Crystallography 42, 169–173.

Xiao, X., Wang, P., Chou, K.C., 2011. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Molecular BioSystems 7 (3), 911–919.

Xiao, X., Wang, P., and Chou, K.C., 2010. Quat-2L: a web-server for predicting protein quaternary structural attributes. Molecular Diversity, 10.1007/s11030-010-9227-8.

Yao, Y.H., Nan, X.Y., Wang, T.M., 2006. A new 2D graphical representation classification curve and the analysis of similarity/dissimilarity of DNA sequences. Journal of Molecular Structure: THEOCHEM 764, 101–108.

Yao, Y.H., Dai, Q., Nan, X.Y., et al., 2008. Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation. Journal of Computational Chemistry 2, 1632–1639.

Yao, Y.H., Dai, Q., Li, L., et al., 2010. Similarity/Dissimilarity studies of protein sequences based on a new 2D graphical representation. Journal of Computational Chemistry 31, 1045–1052.

Zhang, C.T., Chou, K.C., 1993. Graphic analysis of codon usage strategy in1490 human proteins. Journal of Protein Chemistry 12, 329–335.

Zhang, C.T., Chou, K.C., 1994. Analysis of codon usage in 1562 *E. coli* protein coding sequences. Journal of Molecular Biology 238, 1–8.