Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Theoretical Biology 260 (2009) 104–109

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Classification analysis of dual nucleotides using dimension reduction

Zhao-Hui Qi *, Jian-Min Wang, Xiao-Qin Qi

School of Computer and Information Engineering, Shijiazhuang Railway Institute, Shijiazhuang, Hebei 050043, People's Republic of China

ARTICLE INFO

Article history:
Received 31 March 2009
Received in revised form
7 May 2009
Accepted 16 May 2009
Available online 27 May 2009

Keywords:
Dual nucleotides compositions
Principal components analysis (PCA)
Graphic representation
Genome
Eubacteria

ABSTRACT

We introduce a new approach to investigate the dual nucleotides compositions of 11 Gram-positive and 12 Gram-negative eubacteria recently studied by Sorimachi and Okayasu. The approach firstly obtains a 16-dimension vector set of dual nucleotides by PN-curve from the complete genome of organism. Each vector of the set corresponds to a single gene of genome. Then we reduce the 16-dimension vector set to 2-dimension by principal components analysis (PCA). The reduction avoids possible loss of information averaging all 16-dimension vectors. Then we suggest a 2D graphical representation based on the 2-dimension vector to investigate the classification patters among different organisms.

1. Introduction

Recently, graphical techniques have emerged as a powerful tool for the visualization and analysis of complicated biological systems. These methods can provide an intuitive picture and help people gain useful insights. Many graphical approaches have also been used to deal with a wide variety of biological problems. For instance, various graphic schemes have been successfully used to study enzyme-catalyzed system (King and Altman, 1956; Chou et al., 1979; Chou, 1980; Chou and Forsen, 1980, 1981; Chou and Liu, 1981; Cornish-Bowden, 1979; Myers and Palmer, 1985; Zhou and Deng, 1984; Chou, 1989, 1990; Lin and Neet, 1990; Kuzmic et al., 1992; Andraos, 2008), protein folding kinetics (Chou, 1990, 1993), condon usage (Chou and Zhang, 1992; Zhang and Chou, 1994), HIV reverse transcriptase inhibition mechanisms (see Althaus et al., 1993 a–c, as well as a review article, Chou et al., 1994), and base frequency distribution in the anti-sense strands (Chou et al., 1996). Recently, the images of cellular automata were also used to represent biological sequences (Xiao et al., 2005a, b), predict protein subcellular location (Xiao et al., 2006a, b), investigate HBV virus gene missense mutation (Xiao et al., 2005a, b) and HBV viral infections (Xiao et al., 2006a, b), predicting protein structural classes (Xiao et al., 2008) and G-protein-coupled receptor functional classes (Xiao et al., 2009), as well as analyze the fingerprint of SARS coronavirus (Wang et al., 2005; Gao et al., 2006). Graphic approaches have been also used recently to examine the similarities/dissimilarities among the coding sequences of different species (Qi et al., 2007; Qi and Qi, 2007, 2009; Qi and Fan, 2007; Yao et al., 2006), analyze the network structure of the amino acid metabolism (Shikata et al., 2007), and study cellular signaling networks (Diao et al., 2007).

Another useful graphic method, radar chart, has been used to illustrate differences in amino acid compositions to predict protein subcellular localization (Chou and Elrod, 1999). Also, radar charts have been applied in a similar manner to classifying organisms (Sorimachi and Okayasu, 2004, 2008a, b; Okayasu and Sorimachi, 2009). Quite recently, Sorimachi reported some interesting results based on graphical analyses in Sorimachi and Okayasu (2008a, b) and Sorimachi (2009). In Sorimachi and Okayasu (2004), 23 eubacteria was classified (11 Gram-positive and 12 Gram-negative eubacteria) 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. The study shows that amino acid compositions are useful values to investigate genomic structures and biological evolution.

In this paper, we introduce a new approach to investigate the 23 eubacteria studied by Sorimachi and Okayasu (2004). The method consists of two parts: (i) PN-curve, a 3D graphical representation of DNA sequences presented in our earlier study (Qi and Fan, 2007) and (ii) principal components analysis (PCA), a projection method to analyze data set and reduce it from high dimensional space. Here, we firstly obtain a 16-dimension vector set of dual nucleotides by PN-curve from the complete genome of organism. Each vector of the set corresponds to a single gene of genome. Then the 16-dimension vector set is reduced to

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

0022-5193/$ - see front matter Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.05.011

2-dimension by PCA. The 2D graphical representation based on the 2-dimension vector set is proposed to investigate the classification patters among different organisms.

2. Methods

2.1. PN-curve and its applications

PN-curve is a 3D graphical representation of DNA sequences presented in our earlier study (Qi and Fan, 2007). It consider a consider a 4×4 matrix in which the rows and columns are assigned to pairs of nucleotides (PNs)

	A	T	G	C
A	AA	AT	AG	AC
T	TA	TT	TG	TC
G	GA	GT	GG	GC
C	CA	CT	CG	CC

Given an arbitrary DNA primary sequence, the PN-curve can be generated by the following map ϕ:

$$\phi(g_ig_{i+1})=\begin{cases}(1,(AA)_i,i) & \text{if } g_ig_{i+1}=AA\\ (2,(AT)_i,i) & \text{if } g_ig_{i+1}=AT\\ (3,(AG)_i,i) & \text{if } g_ig_{i+1}=AG\\ (4,(AC)_i,i) & \text{if } g_ig_{i+1}=AC\\ (5,(TA)_i,i) & \text{if } g_ig_{i+1}=TA\\ (6,(TT)_i,i) & \text{if } g_ig_{i+1}=TT\\ (7,(TG)_i,i) & \text{if } g_ig_{i+1}=TG\\ (8,(TC)_i,i) & \text{if } g_ig_{i+1}=TC\\ (9,(GA)_i,i) & \text{if } g_ig_{i+1}=GA\\ (10,(GT)_i,i) & \text{if } g_ig_{i+1}=GT\\ (11,(GG)_i,i) & \text{if } g_ig_{i+1}=GG\\ (12,(GC)_i,i) & \text{if } g_ig_{i+1}=GC\\ (13,(CA)_i,i) & \text{if } g_ig_{i+1}=CA\\ (14,(CT)_i,i) & \text{if } g_ig_{i+1}=CT\\ (15,(CG)_i,i) & \text{if } g_ig_{i+1}=CG\\ (16,(CC)_i,i) & \text{if } g_ig_{i+1}=CC\end{cases}$$

where $(AA)_i$, $(AT)_i$, $(AG)_i$, $(AC)_i$,…, $(CG)_i$ and $(CC)_i$ are the cumulative occurrence numbers of AA, AT, AG, AC,…, CG and CC, respectively, in the subsequence from the first base to the nth base in the sequence. Unlike important geometry curve Z (Zhang, 1997) and Z' (Zhang and Zhang, 2004), PN-curve is not unique because of the 16! different combinations about 16 kinds of PNs. However, we are only interested in PN-curve as numerical parameters that may extract characteristics of DNA sequences. Here, the different combinations attach no impact on the extracted numerical parameters.

For a given gene or genome, there is a cumulative PN-profile or PN-curve corresponding to it. The PN-curve or the cumulative PN-profile is used interchangeably in this paper. Note that the essence of cumulative PN-profile is to display the variations of the PN content along a gene or genome. The derivative of PN-curve with respect to the PN content is used to construct 16-component vectors related with the cumulative PN content. The 16-component vector consists of the percentage of 16 kinds of PNs: AA, AT, AG, AC,…, CG and CC.

For a given gene, there is a 16-component vector to reveal the patterns of PN compositions. As for genome of an organism, there are thousands of genes. It is not practical that the cluster tendency of patterns is only based on a single pattern derived from a single gene. There are two normal ways to know the tendency of genome: (i) all genes are linked each other into a very long sequence, and a 16-component vector by cumulative PN-profile is used to represent the cluster pattern of PN compositions; (ii) each gene is used to generate a corresponding 16-component vector by PN-profile, and the average of all vectors illustrates the cluster pattern. However, the two ways may hide some detail. For example, given an average 5, there are many possible chooses: 1 and 9, 2 and 8, or 5 and 5, 6 and 4. It was obvious that the average maybe hide the distinction among different chooses. So we use dimension reduction method to uncover the more detail hid in all vectors.

2.2. Dimension reduction method based on principal components analysis

Principal components analysis (Jackson and Wiley, 1991) is a projection method to analyze data set and reduce it from high dimensional space to few hidden variables while keeping information on its variability. It and its many expanded methods have been successfully applied to the resolution of some problems (Costa et al., 2009; Du et al., 2006). Since the patterns in 16-component vectors of genome of an organism can be hard to find in high dimension space, where graphical representation is not available, the possibility of grouping the variability in few variables is an important step to visualize and consequently uncover the information. In Wang et al. (2008), an effective dimension-reducing approach was introduced for predicting membrane protein types. Here, we reduce 16-dimension vector to 2-dimension by using PCA. Then we give the 2D graphical representation of the patterns of PN compositions and utilize the representation to intuitively observe the evolution patters among different organisms.

We now give the simple description of PCA. Assume that the mean of sample $X=\{x_i\}_{i=1}^n$ of space R^D is $\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i$. Write the singular value decomposition of covariance matrix Σ as $\Sigma=U\wedge U^T$, where $\Sigma=E(x-\bar{x})(x-\bar{x})^T$. Matrix U is orthogonal matrix. Diagonal matrix $\wedge$ is made up of the eigenvalues of Σ, where $\wedge=diag(\lambda_1,\ldots,\lambda_D)$ and $\lambda_1\geq\cdots\geq\lambda_D$. The principle component transformation is $U^T(X-\bar{X})$. Then a new data set $Y=\{y_i\}_{i=1}^n$ is obtained by $Y=U^T(X-\bar{X})$. The mean and covariance matrix of Y are 0 and diagonal matrix $\wedge$, respectively. Now we ignore the components of lesser significance and leave out some important components. Then the final data set will have fewer dimensions than the original. In this paper, the original data set is 16 dimensions. The final data set has only two dimensions by choosing only the first two eigenvectors.

2.3. Genomic data used for this study

Complete genome sequences were downloaded from NCBI GenBank (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Here, the analysis of genomes was performed by some bacteria consisting of 11 Gram-positive and 12 Gram-negative bacteria (Sorimachi and Okayasu, 2004). The genome sequences used for this study are summarized in Table 1.

3. Applications

3.1. Calculations

Dual nucleotide contents at various base positions were computationally calculated by PN-curve. As for a sequence, a 16-dimension vector related to dual nucleotide contents is

Table 1
Genomes used for this study.

Strain	Accession (GenBank)	RefSeq identifier	Total length (bp)	Genes
*Staphylococcus aureus Mu*50	BA000017.4	NC_002758	2,878,529	2775
*Streptococcus pyogenes M*1	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 J*99	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

generated by PN-curve. The complete genome of species consists of thousands of genes. Each gene corresponds to a vector. Then we can obtain a vector set corresponding to the complete genome of the specie. In order to visualize and uncover the information hidden in the vector set, we reduce 16-dimension vector to 2-dimension by using PCA. Then we give the 2D graphical representation of the patterns of PN compositions and utilize the representation to intuitively observe the evolution patters among different organisms. We develop two programs. A program named as ''GenomePNs.pl'' is designed to generate 16-dimension vector set of complete genome. The input of the perl program is file ''*.ffn'' from NCBI GenBank. Its output is a file called as ''percentage_PNs.txt''. The other program, ''DimReductionAnaly.m'', is a matlab program used to reduce 16-dimension vector set to 2-dimension by using PCA algorithm and visualize the 2-dimension vector set. Its input is ''percentage_PNs.txt'' and the output is a 2-dimension graphic representation.

3.2. Results

The patterns of dual nucleotide compositions based on the complete genomes of various eubacteria in Table 1 are 2-dimension dot-cluster graphs, as shown in Fig. 1.

To characterize the pattern of dual nucleotide compositions and to classify eubacteria, we focused on particular dot-cluster. A close look to Fig. 1 shows that dot-cluster in some graphs is mainly grouped into two clusters while those dots in other graphs is mainly grouped into one cluster. Now, we divide the grid coordinate system into two regions: I and II, as shown in Fig. 2. Concentrations of dot-cluster changed markedly two main groups: ''S-Type'' represented by *S. aureus* and ''E-Type'' represented by *E. coli*. The conception about ''S-Type'' and ''E-Type'' is presented by Sorimachi and Okayasu (2004). Here, concentrations of dot clusters are mainly inside Region I in ''S-Type'' whereas they are mainly inside Region II in ''E-Type''. The two groups are separated from each other by these dot-clusters.

By using PCA algorithm, eubacteria is classified into two groups, ''S-Type'' and ''E-Type'', based on the dual nucleotides compositions calculated from the complete genome. In ''S-Type'', the patterns of the dot-clusters also show much difference each other. According to the relative location between main clusters, ''S-Type'' can be classified into two subgroups: (i) one subgroup includes the bacteria, *S. aureus Mu*50, *Str. pyogenes M*1, *B. subtilis*, *R. prowazekii*, *C. perfringens* 13 and *B. burgdorferi*. Concentration of dots of them is mainly in the left dot-clusters and (ii) the other consists of the following, *L. monocytogenes*, *C. jejun*i, *M. pulmonis*, *H. pylori J*99, *H. influenzae*, *M. genitalium* and *H. pylori* 26695. Concentration of dots of these bacteria is mainly in the right dot-clusters. Similarly, ''E-Type'' can be also classified into two subgroups. The first subgroup includes those bacteria whose dots are classified into two clusters. They are *E. coli*, *S. typhi*, *V. cholerae* and *Y. pestis*. Concentration of the dots of the second is mainly one dot-cluster. They are *M. tuberculosis*, *M. leprae*, *N. meningitides*, *T. pallidum*, *M. pneumoniae* and *U. urealyticum*, respectively. The above results show that bacteria in Table 1 are grouped into two classes: *S. aureus* ''S-Type'' and *E. coli* ''E-Type'', based on their genomic structures. As Okayasu and Sorimachi (2009) reported both types ''S-Type'' and ''E-Type'', the above species were classified further into their subgroups based on amino acid compositions or codon usages. Similar results have also been obtained by Sorimachi and Okayasu (2004).

4. Discussion

By using data derived from dual nucleotides based on complete genomes, our studies are applicable to analyze genomic structures and provide their 2-dimension graphic representation by PCA algorithm. Then the method is used to investigate the dual

nucleotide compositions of 11 Gram-positive and 12 Gram-negative eubacteria in Table 1.

The amino acid compositions of the eubacteria of Table 1 have been studied by Sorimachi and Okayasu (2004). Their research results show that these eubacteria were classified into two groups, "S-Type" represented by *S. aureus* and "E-Type" represented by *E. coli*. Similarly, we also classified these eubacteria into two groups, "S-Type" and "E-Type", according to particular dot-cluster derived from complete genomes data by PCA. Compared with the research results of Sorimachi and Okayasu, our results show some diversity in three eubacteria: *H. influenzae*, *M. pneumoniae* and *U. urealyticum*. Here, *H. influenzae* belong to "S-Type" while *M. pneumoniae* and *U. urealyticum* are "E-Type". We do not think that the diversity imply some failure in our approach or

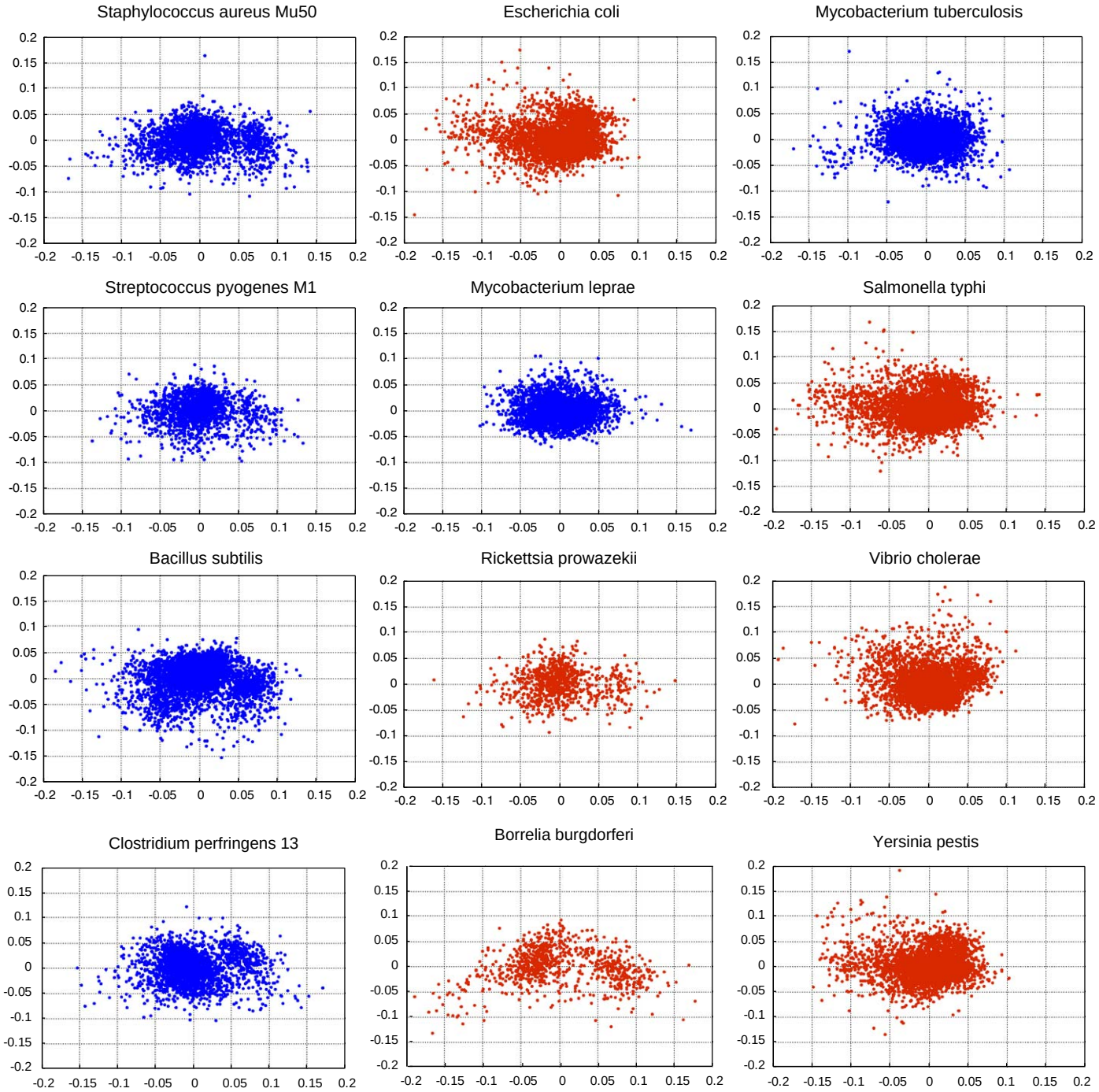

Fig. 1. 2-dimension dot-matrix graphs of dual nucleotides determined from the complete genomes of various eubacteria of Table 1. As shown in Fig. 1 of Sorimachi and Okayasu (2004), blue represents Gram-positive bacteria; red represents Gram-negative bacteria; green represents mycoplasmas, which lack a cell wall. (For interpretation of the references to the color in this figure legend, the reader is referred to the web version of this article.)

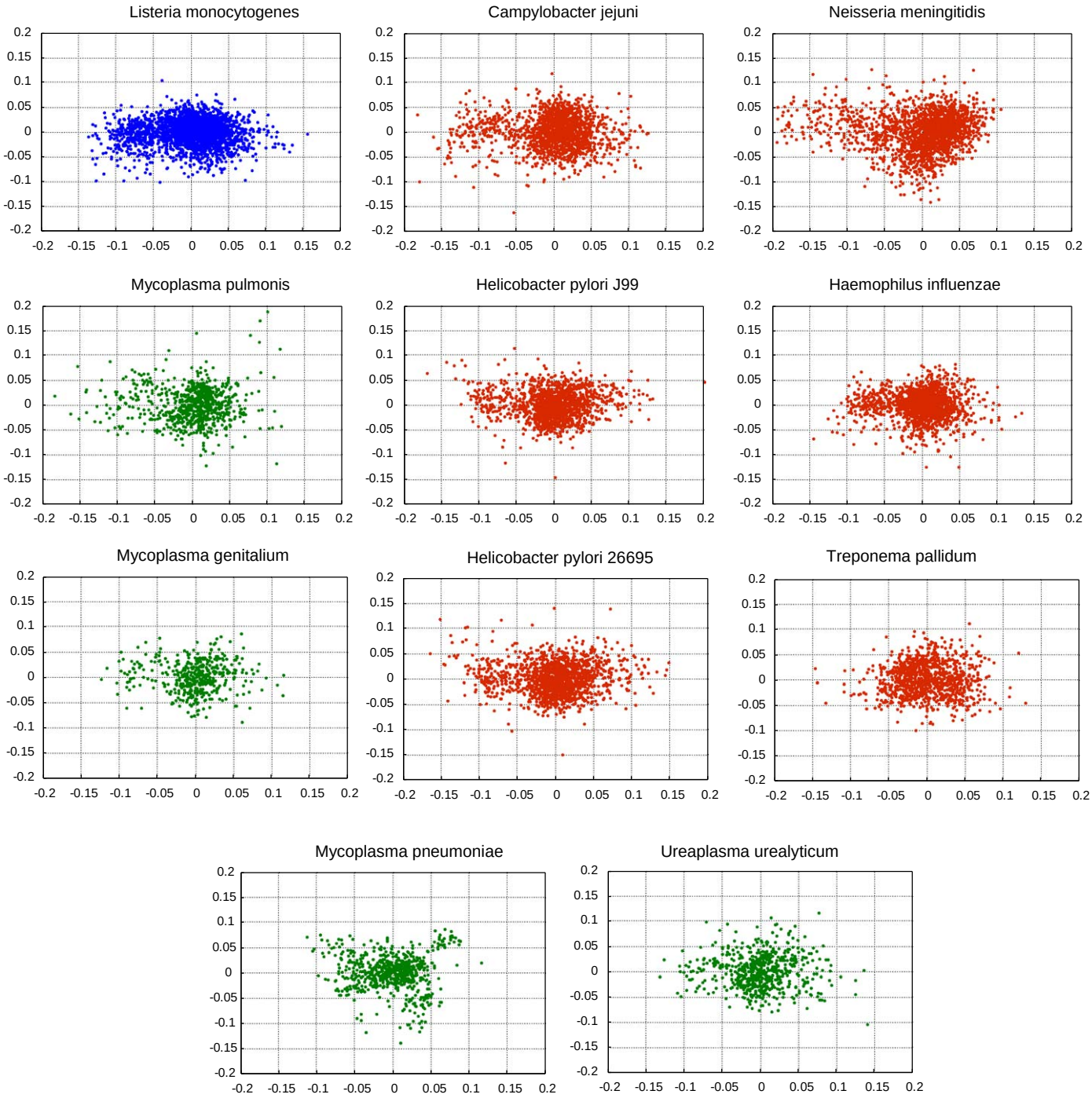

Fig. 1. (*Continued*)

Sorimachi and Okayasu's scheme. The proportion of diversity is low, only 13%. Looked by statistics viewpoint, the diversity is acceptable.

The present study demonstrates that dual nucleotide compositions are useful values to investigate genomic structures and biological evolution. The more similar the structures of dot-cluster are the more similar the organisms are. That is to say, the structures between evolutionary closely related species are more similar, while those between evolutionary disparate species are larger. Closely observing Fig. 1, we find that in "S-Type" the more similar species groups are the following: *S. aureus Mu*50, *S. pyogenes M*1 and *B. subtilis*; *L. monocytogenes*, *H. pylori J*99, *H. influenzae* and *H. pylori* 26695. In "E-Type" the more similar species groups are the following: *E. coli*, *S. typhi*, *Y. pestis*; *M. tuberculosis*, *M. leprae*, *T. pallidum* and *U. urealyticum*. Similar results can be found out in Fig. 1 of Sorimachi and Okayasu (2004).

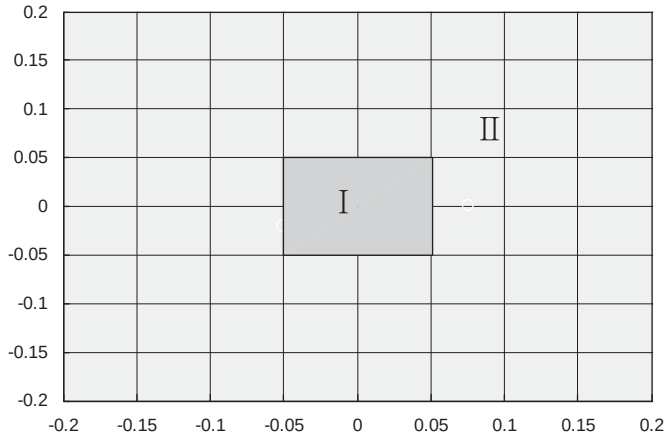

Fig. 2. Grid coordinate system is divided into two region: Region I ($x \in [-0.05, 0.05]$, $y \in [-0.05, 0.05]$) and Region II (outside Region I).

References

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., Reusser, F., 1993a. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548–6554.

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., Reusser, F., 1993b. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. Journal of Biological Chemistry 268, 6119–6124.

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., Reusser, F., 1993c. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. Journal of Biological Chemistry 268, 14875–14880.

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

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., 1993. Graphic rule for non-steady-state enzyme kinetics and protein folding kinetics. Journal of Mathematical Chemistry 12, 97–108.

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., Jiang, S.P., Liu, W.M., Fee, C.H., 1979. Graph theory of enzyme kinetics: 1. Steady-state reaction system. Scientia Sinica 22, 341–358.

Chou, K.C., Kezdy, F.J., Reusser, F., 1994. Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Analytical Biochemistry 221, 217–230.

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., Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Research and Human Retroviruses 8, 1967–1976.

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

Cornish-Bowden, A., 1979. Fundamentals of Enzyme Kinetics. Butterworths, London (Chapter 4).

Costa, J.C., Alves, M.M., Ferreira, E.C., 2009. Principal component analysis and quantitative image analysis to predict effects of toxics in anaerobic granular sludge. Bioresource Technology 100, 1180–1185.

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.

Du, Q.S., Jiang, Z.Q., He, W.Z., Li, D.P., Chou, K.C., 2006. Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. Journal of Biomolecular Structure and Dynamics 23, 635–640.

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.

Jackson, J.E., Wiley, J.W., 1991. A User's Guide to Principle Components. Wiley-Interscience, New York.

King, E.L., Altman, C., 1956. A schematic method of deriving the rate laws for enzyme-catalyzed reactions. Journal of Physical Chemistry 60, 1375–1378.

Kuzmic, P., Ng, K.Y., Heath, T.D., 1992. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. Analytical Biochemistry 200, 68–73.

Lin, S.X., Neet, K.E., 1990. Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. Journal of Biological Chemistry 265, 9670–9675.

Myers, D., Palmer, G., 1985. Microcomputer tools for steady-state enzyme kinetics. Bioinformatics 1, 105–110 (original: Computer Applied Bioscience).

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

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.

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

Sorimachi, K., 2009. A proposed solution to the historic puzzle of Chargaff's second parity rule. Open Genomics Journal 2, 12–14.

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

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

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

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.

Wang, T., Yang, J., Shen, H.B., Chou, K.C., 2008. Predicting membrane protein types by the LLDA algorithm. Protein & Peptide Letters 15, 915–921.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005a. Using cellular automata to generate image representation for biological sequences. Amino Acids 28, 29–35.

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., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005b. An application 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., 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., Wang, P., Chou, K.C., 2008. 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., 2009. GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. Journal of Computational Chemistry 30, 1414–1423.

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.

Zhang, C.T., 1997. A symmetrical theory of DNA sequences and its applications. Journal of Theoretical Biology 187, 297–306.

Zhang, C.T., Zhang, R., 2004. Isochore structures in the mouse genome. Genomics 83, 384–394.

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.

Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochemical Journal 222, 169–176.