ACADEMIC PRESS

Molecular Phylogenetics and Evolution 25 (2002) 101–111

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.academicpress.com

Proteome-wide analysis of protein function composition reveals the clustering and phylogenetic properties of organisms

Lunjiang Ling,[a,*] Jinhua Wang,[a,1] Yan Cui,[b,c] Wei Li,[a] and Runsheng Chen[a,*]

[a] *Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Road, Chaoyang District, Beijing 100101, China*
[b] *Department of Biostatistics, Harvard University, Boston, MA 02115, USA*
[c] *Dana–Farber Cancer Institute, Boston, MA 02115, USA*

Received 11 June 2001

Abstract

A 17-dimensional vector named the proteome vector is defined to represent an organism. The components of the vector reflect the relative contents of protein-encoding genes of the 17 cluster of orthologous groups of proteins (COGs) classes in the whole genome of the relevant organism. Based on the definition of this proteome vector, the fuzzy clustering of 36 completely sequenced organisms (8 archaea, 24 bacteria, and 4 eukarya) was performed and a proteome tree was constructed. Our results show that (1) the 36 organisms can be 100% correctly classified into three clusters corresponding to the three primary kingdoms, (2) our proteome tree is remarkably similar to that derived from 16S rRNA, and (3) the chromosomes and/or plasmids belonging to the same organism have very similar gene composition. Based on these results, we argue that the 17-dimensional proteome vector could be a good criterion for clustering approaches and to a large extent reveals the phylogenetic properties of organisms; the Three Primary Kingdoms Hypothesis is trustworthy although the existence of lateral gene transfer (LGT) brings controversy to the construction of the "universal tree of life."

1. Introduction

A phylogenetic tree is usually derived from the aligned sequences of a common protein or RNA of various organisms. This means that an organism is represented by only one of its proteins/RNAs. Unfortunately, the trees constructed in this way are rarely uniform over all of the genes conserved in the respective organisms. For example, archaea appear to be close to eukarya if the genes involved in transcription and translation are compared, but if metabolic genes are compared, they tend to be close to bacteria (Doolittle and Logsdon (1998)). This inconsistency was thought to be due to the existence of lateral gene transfer (LGT), unequal rates of nucleotide substitution, and gene displacement. Recently, more and more LGT events have been discovered and reported. It is estimated that 1.5–14.5% of genes in a genome are related to LGT (Garcia-Vallvé et al., 2000) and even rRNA molecules are involved in LGT (Yap et al., 1999). As a result, people have become suspicious of the reliability of phylogenetic trees constructed in this way. The Three Primary Kingdoms Hypothesis (Woese et al., 1990) has also been criticized (Gupta, 1998; Mayr, 1998). Moreover, some people argue that it is impossible to reconstruct a universal tree of life (Doolittle, 1998; Pennisi, 1999).

Is the Three Primary Kingdoms Hypothesis still trustworthy? To answer this question, clustering and phylogenetic approaches based on characteristics other than aligned sequences of common protein or RNA of the organisms should be explored. Because the tree of life should reflect the overall evolutionary relationships between the organisms, the information about the global properties of the genome/proteome should be taken into account to construct a more convincing tree of life. Nowadays, more and more organisms of different primary kingdoms have been completely sequenced; hence it has become possible to consider the clustering and phylogenetic characteristics of the organisms at the

* Corresponding author. Fax: +86-10-648-77-837.
E-mail addresses: ling@sun5.ibp.ac.cn (L. Ling), chenrs@sun5.ibp.ac.cn (R. Chen).
[1] Present address: Cold Spring Harbor Lab, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.

1055-7903/02/$ - see front matter
PII: S1055-7903(02)00354-8

genome level. Tekaia et al. (1999) constructed a genomic tree by comparing the full set of predicted gene products both within and between the completely sequenced organisms. Fitz-Gibbon and House (1999) constructed a tree of life based on presence/absence of families of protein-encoding genes in 11 complete genomes of free-living microorganisms. Snel et al. (1999) evaluated the similarity between two organisms by calculating the proportion of the shared genes in their genomes and accordingly constructed a genomic tree. Lin and Gerstein (2000) considered the phylogenetic problem at both protein sequence and three-dimensional structure levels. Montague and Hutchison (2000) used a 104-bit "01" sequence indicating the presence ("1") or absence ("0") of the 104 clusters of orthologous groups of proteins (COGs) in a genome and obtained the "gene content phylogeny" for the 13 completely sequenced herpesviruses by comparing such sequences. Most of these trees are basically consistent with that derived from 16S rRNA, but with some differences in inner bifurcations. Compared to the conventional methods of assessing molecular evolution, these "whole-genome-based" approaches took much more information into account, but still had their respective limitations. For example, some of these methods use only a subset of genes; some are not suitable for the study of organisms with large differences in their genome sizes. We believe that the impetus for evolution comes mainly from the structural changes of genomes, especially the loss or acquisition of genes. Therefore, this nature should be reflected by the changes in contents and functional properties of all genes. Here we present a different approach based on the "proteome vector" whose components reflect the relative contents of protein-encoding genes (simply referred to as "genes" hereafter) with different functions in the whole genome of an organism. The computer programs (all written in C) for this study and all resulting data can be accessed by anonymous FTP (site: 159.226.118.105, subdirectory: 17D-COG).

2. Methods

The COGs are classified into 17 COG classes (denoted as J, K, L, D, O, M, N, P, T, C, G, E, F, H, I, R, S) according to their potential biological functions. Each COG consists of individual proteins or groups of paralogs from at least three lineages and thus corresponds to an ancient conserved domain (Tatusov et al., 1997, 2001). Since the 17 functional classes cover almost all known functions needed for microbial survival and all of the organisms involved in the COG database construction are included in this study, it is reasonable to regard the relative content of genes corresponding to the 17 COG classes as a representation of a genome. We defined a 17-dimensional proteome vector and then performed clustering and phylogenetic approaches based on this definition. The steps are:

2.1. Step 1. Getting source data

Two sets of source data were used in our study. Set 1 includes the amino acid sequences of all gene products in genomes (that is, all proteins in the proteomes), and Set 2 includes the amino acid sequences of the COGs and their functional class attributes.

All of the source data were downloaded from NCBI's FTP site. Data of Set 1 were extracted from the "translation =" items of "FEATURE" regions in respective ".gbk" files located in subdirectories under "ncbi.nlm.nih.gov/genbank/genomes/." We list the serial numbers (assigned by us), whole names, short names, genome sizes, and revision date of the 36 organisms (8 archaea, 24 bacteria, and 4 eukarya) used in this study in Table 1. Data of Set 2 were downloaded from "ncbi.nlm.nih.gov/COG/" including 2885 COGs (53,753 protein sequences) from almost all of the organisms listed in Table 1, except Atha, Cele, and Dmel.

2.2. Step 2. Constructing proteome vectors $\{U_s\}$

(2a) Formalize the data of Set 2 to form a FASTA-formatted COG library file named "AllCogSq.fas."

(2b) Suppose that there are NP genes in a genome (comprising possibly several chromosomes and plasmids), and the relevant products are denoted as $P_1, P_2, \ldots, P_{NP}$. For each P_i $(i = 1, 2, \ldots, NP)$, the FASTA homology search was performed against the library "AllCogSq.fas" to find out the top NC P_i's homologous sequences (denoted as LS[1], LS[2], ..., LS[NC]).

(2c) Define a 17-dimensional protein vector $\{v_i\} = \{v_i[1], v_i[2], \ldots, v_i[17]\}$ to describe P_i's COG class attribute. The components of $\{v_i\}$ were determined by

$$\left\{ \begin{array}{l} v_i[m] = \sum_{j=1}^{NC} H_j \times L_j \times \delta_{m,C[j]} \\ \delta_{m,C[j]} = \begin{cases} 1, & m = C[j] \\ 0, & m \neq C[j] \end{cases} \end{array} \right\}, \quad (m = 1, 2, \ldots, 17), \tag{1}$$

where $C[j]$ is the COG class of LS[j], L_j is the length of the homologous region between P_i and LS[j], and H_j is the percentage identity value of the homologous region. So, the magnitude of $v_i[m]$ reflects how similar P_i is to the mth COG class. If $P_i \in$ AllCogSq.fas, then v_i would have a component much larger than those of all others (often, the values of all other components equal zero); if $P_i \notin$ AllCogSq.fas, then the corresponding v_i would have very small components. In our calculation, we let NC = 17.

Table 1
The 36 organisms

ID			Genome Size (Mb)	Revision date of the source data (.gbk files) used in this study
Serial numbers	Whole name	Short name		
1	*Archaeoglobus fulgidus*	Aful	2.18	17-DEC-1997
2	*Aeropyrum pernix*	Aper	1.67	21-JUN-1999
3	*Halobacterium* sp.	Halo	2.57	03-OCT-2000, 03-APR-2000
4	*Methanococcus jannaschii*	Mjan	1.66	10-MAY-1999
5	*Methanobacterium thermoautotrophicum*	Mthe	1.75	18-NOV-1997
6	*Pyrococcus abyssi*	Paby	1.77	14-JUL-1999
7	*Pyrococcus horikoshii*	Phor	1.74	22-JUN-1998
8	*Thermoplasma acidophilum*	Taci	1.56	17-OCT-2000
9	*Aquifex aeolicus*	Aaeo	1.55	27-MAR-1998
10	*Borrelia burgdorferi*	Bbur	1.52	17-DEC-1997
11	*Bacillus halodurans*	Bhal	4.2	03-AUG-2000
12	*Bacillus subtilis*	Bsub	4.21	10-MAY-1999
13	*Buchnera* sp.	Buch	0.64	12-SEP-2000, 09-MAY-2000
14	*Campylobacter jejuni*	Cjej	1.64	15-FEB-2000
15	*Chlamydia pneumoniae*	Cpne	1.23	10-MAR-1999
16	*Chlamydia trachomatis*	Ctra	1.04	15-DEC-1999
17	*Deinococcus radiodurans*	Drad	3.28	22-NOV-1999, 16-NOV-1999
18	*Escherichia coli.*	Ecol	4.64	18-NOV-1998
19	*Haemophilus influenzae*	Hinf	1.83	10-MAY-1999
20	*Helicobacter pylori*	Hpyl	1.67	16-APR-1999
21	*Mycoplasma genitalium*	Mgen	0.58	17-MAY-1999
22	*Mycoplasma pneumonia*	Mpne	0.82	24-NOV-2000
23	*Mycobacterium tuberculosis*	Mtub	4.41	07-JUL-1998
24	*Neisseria meningitidis*	Nmen	2.27	05 -APR-2000
25	*Pseudomonas aeruginosa*	Paer	6.3	30-AUG-2000
26	*Rickettsia prowazekii*	Rpro	1.11	10-MAY-1999
27	*Synechocystis* sp.	Syne	3.57	16-OCT-1997
28	*Thermotoga maritima*	Tmar	1.86	04-JUN-1999
29	*Treponema pallidum*	Tpal	1.14	16-JUN-1998
30	*Ureaplasma urealyticum*	Uure	0.75	07-FEB-2000
31	*Vibrio cholerae*	Vcho	4.03	03-AUG-2000
32	*Xylella fastidiosa*	Xfas	2. 68	JUL-2000
33	*Arabidopsis thaliana*	Atha	Incomplete data set	Chrom-I: 14-DEC-2000, Chrom-II: 21-DEC-1999, Chrom-IV: 17-DEC-1999
34	*Caenorhabditis elegans*	Cele	100	APR–MAY, 1999
35	*Drosophila melanogaster*	Dmel	120	OCT-2000
36	*Saccharomyces cerevisiae*	Scer	12.07	06-DEC-1999

(2d) For all P_i ($i = 1, 2, \ldots, \mathrm{NP}$), repeat the calculations of steps 2b and 2c and obtain a complete set of protein vectors $\{v_i, i = 1, 2, \ldots, \mathrm{NP}\}$. After that, sum them to form $V = \{V[1], V[2], \ldots, V[17]\}$:

$$V = \sum_{i=1}^{NP} v_i, \text{ or } V[m] = \sum_{i=1}^{NP} v_i[m], \qquad m = 1, 2, \ldots, 17. \tag{2}$$

Obviously, the length of V is proportional to the number of genes in the genome and the homologous degree of the products of these genes to the COG sequences, and the direction of V is determined by the relative contents of genes of different functions.

(2e) Suppose that the number of organisms concerned is NS; repeat the above steps 2b–d for all of the NS organisms and obtain a set of vectors: $\{V_s, s = 1, 2, \ldots, \mathrm{NS}\}$.

(2f) Because the numbers of genes of two genomes (for instance, genome i and genome j) are often greatly different, there will be large deviations and twists in our approaches if the distance between V_i and V_j is directly used to represent the distance between organism i and organism j. So, it is not the absolute amount but the relative content (or composition) of genes of different functions that are concerned in our approach. We normalized the vectors $\{V_s, s = 1, 2, \ldots, \mathrm{NS}\}$ to become $\{U_s, s = 1, 2, \ldots, \mathrm{NS}\}$ so that they had the same length,

$$U_s[m] = Q_s \times V_s[m] \cdot \qquad m = 1, 2, \ldots, 17, \quad s = 1, 2, \ldots, \mathrm{NS}, \tag{3}$$

where

$$Q_s = \frac{C}{\sqrt{\sum_{m=1}^{17} (V_s[m] \times V_s[m])}}, \tag{4}$$

so that

$$|U_s| = \sqrt{\sum_{m=1}^{17}(U_s[m])^2} = C.$$

We call vectors $\{U_s, s = 1, 2, \ldots, \mathrm{NS}\}$ proteome vectors. Now, we can calculate the Euclidean distance between organism x and organism t (denoted as $d[x,t]$) by the equation

$$d[x,t] = \sqrt{\sum_{m=1}^{17}(U_x[m] - U_t[m])^2}. \quad (5)$$

2.3. Step 3. Clustering approach

The clustering method used in this study is called Fuzzy c-means (FCM) (for mathematical details of the method, see Zhang et al. (1995) and Bezdek (1981).

Our goal is to classify the NS proteome vectors $\{U_1, U_2, \ldots, U_{\mathrm{NS}}\}$ into K classes ($k \leqslant \mathrm{NS}$). The centers of the K classes could be denoted as K centroid vectors $\{C_1, C_2, \ldots, C_k\}$. Take a particular U_s as a query vector; its K class attribute (or the similarities to the K centroid vectors) could be described by a K-dimensional weight vector $W_s = \{w_s[1], w_s[2], \ldots, w_s[K]\}$ which is a so-called "membership degree." W_s should satisfy the constraints

$$0 \leqslant w_s[k] \leqslant 1 \text{ and } \sum_{k=1}^{K} w_s[k] = 1. \quad (6)$$

2.4. Training

(3a) Select NT vectors from $\{U_1, U_2, \ldots, U_{\mathrm{NS}}\}$ to form the training group; redenote these NT vectors as $\{Y_1, Y_2, \ldots, Y_{\mathrm{NT}}\}$. Obviously, NT $\in$ NS, $\{Y_1, Y_2, \ldots, Y_{\mathrm{NT}}\} \in \{U_1, U_2, \ldots, U_{\mathrm{NS}}\}$.

(3b) Randomly generate K vectors $\{C_1, C_2, \ldots, C_k\}$, then use them as initial centroid vectors of $\{Y_1, Y_2, \ldots, Y_{\mathrm{NT}}\}$, and assign a very large value to the initial object function $J_q(0)$.

(3c) For each Y_t from $\{Y_1, Y_2, \ldots, Y_{\mathrm{NT}}\}$ and each C_k from $\{C_1, C_2, \ldots, C_K\}$, the distance between them is calculated by Eq. (5) and denoted as $\mathrm{d}[t,k]$.

(3d) Calculate the "membership degrees" $w_t[k]$ of the training group $\{Y_1, Y_2, \ldots, Y_{\mathrm{NT}}\}$ by

$$w_t[k] = \frac{[1/d^2[t,k]]^{1/(q-1)}}{\sum_{k=1}^{K}[1/d^2[t,k]]^{1/(q-1)}} \cdot t = 1, 2, \ldots, NT, \quad k = 1, 2, \ldots, K, \quad (7)$$

where $q \in [1, \infty]$ is the fuzzy index; $q = 1$ corresponds to the definite clustering. In this case, take Y_t as an example; there must be a component of weight vector W_t equal to 1; all other $K - 1$ components equal 0. As q becomes larger, these components of W_t become more even; hence the class attribute of Y_t becomes fuzzier.

(3e) Calculate new centroid vectors by

$$C_k = \frac{\sum_{t=1}^{NT}(w_t[k])^q U_t}{\sum_{t=1}^{NT}(w_t[k])^q}, \quad k = 1, 2, \ldots, K. \quad (8)$$

(3f) Calculate the objective function $J_q(1)$ by

$$J_q(1) = \sum_{k=1}^{K}\sum_{t=1}^{NT}(w_t[k])^q d^2[t,k]. \quad (9)$$

(3g) If

$$|J_q(1) - J_q(0)| < \varepsilon \ (\varepsilon \text{ is a very small positive fraction}) \quad (10)$$

stop the interaction, regard $C_1, C_2, \ldots, C_k$ as the final centroid vectors for the K classes and go to step 3h to start prediction steps. Otherwise, let $J_q(0) = J_q(1)$, and go to step 3c to start a new iteration.

Predicting

(3h) For a query proteome vector U_s, calculate its class membership degrees $w_s[1], w_s[2], \ldots, w_s[K]$ by Eq. (7) (simply replace the subscript t with s).

(3i) Find the maximum among $(w_s[1], w_s[2], \ldots, w_s[k])$, say $W_s[j]$; then classify U_s into class j. In our calculation, we let $K = 3$.

2.5. Step 4. Phylogenetic approach

By applying Eq. (5) to all proteome vectors corresponding to the organisms concerned (say, there are N_s organisms), we can obtain a distance matrix $D\{d[s,t], s = 1, 2, \ldots, \mathrm{NS}, \quad t = 1, 2, \ldots, \mathrm{NS}\}$. Based on the distance matrix D, we can construct a proteome tree by means of neighbor-joining (Saitou and Nei, 1987).

Because our proteome tree is based on the proteome vectors, no current software is available for evaluating its statistical reliability. We performed the reliability test for our proteome tree in the following manner.

(4a) Consider the mth components of all NS proteome vectors; they form a column vector: $(U_1[m], U_2[m], \ldots, U_{\mathrm{NS}}[m])$.

(4b) Transform this column vector to NS binary sequences. These sequences have the same length determined by the maximum of $U_1[m]$ to $U_{\mathrm{NS}}[m]$ (denoted as $U_{\max}$), and the proportion of "1" to "0" in the sth sequence is determined by the proportion of $U_s[m]$ to $U_{\max}$.

(4c) Use the method similar to "seqboot" in the PHYLIP package to generate 99 sets of such kind of sequences.

(4d) Transform the 99 sets of the binary sequences back to 99 column vectors.

(4e) Treat every column vector $\{(U_1[m], U_2[m], \ldots, U_{\mathrm{NS}}[m]), \ m = 1, 2, \ldots, 17\}$ in the same way men-

Table 2
The composition of proteins in a whole proteome belonging to 17 COG classes (proteome vector) of the 36 organisms

Organism	The 17 COG-classes																	CDS Number	1/Q
	J	K	L	D	O	M	N	P	T	G	C	E	F	H	I	R	S		
Aful	29.2	10.5	20.3	4.7	12.3	9.0	7.6	18.1	9.5	10.5	42.6	37.7	11.5	20.0	29.1	51.6	22.7	2407	7590.9
Aper	34.7	10.0	19.9	2.9	14.6	9.6	7.6	20.5	5.3	18.9	36.2	46.0	8.8	18.1	14.4	52.7	17.6	2694	5832.8
Halo	30.6	14.1	31.3	5.2	15.6	11.2	18.7	21.1	8.1	14.1	33.5	40.2	15.8	23.0	15.1	49.4	15.8	2234	6291.5
Mjan	38.3	10.7	30.7	7.3	12.9	10.6	8.1	17.0	2.5	13.7	32.7	39.7	16.1	23.1	4.5	50.5	24.2	1715	5454.8
Mthe	34.3	11.8	24.9	4.4	17.0	15.7	5.1	21.2	9.9	12.4	37.3	37.9	15.9	28.3	8.0	48.5	21.9	1869	5718.2
Paby	33.7	10.9	21.0	5.9	11.3	13.0	9.8	19.9	3.8	17.8	32.7	37.9	13.2	14.4	4.5	61.2	22.2	1765	7230.5
Phor	33.4	11.2	21.0	6.5	11.6	12.1	10.2	18.7	4.3	20.9	30.0	33.3	13.5	13.7	3.9	64.1	24.0	1979	6782.6
Taci	32.5	11.6	20.2	3.5	12.2	7.2	5.9	16.1	1.9	19.1	40.3	43.7	15.5	21.3	16.1	53.3	14.4	1478	5222.9
Aaeo	41.9	11.3	25.6	7.6	19.0	24.9	18.1	20.3	9.9	12.7	33.3	40.0	16.0	24.1	11.3	39.8	12.8	1522	6139.9
Bbur	54.1	14.8	45.9	16.3	21.0	20.8	28.7	14.0	6.7	23.4	12.3	14.6	8.6	5.7	4.6	37.0	11.8	850	3912.7
Bhal	24.1	22.2	26.7	6.0	12.8	15.3	13.6	22.3	12.2	30.9	29.2	42.8	12.0	16.5	14.5	49.3	15.9	4066	13409.1
Bsub	23.0	21.2	21.5	5.7	11.6	18.6	12.2	21.4	10.1	28.5	28.6	51.2	12.6	15.9	12.2	48.9	12.0	4100	14054.6
Buch	65.8	16.7	28.0	5.7	21.1	16.4	17.5	6.3	1.0	19.4	25.6	39.3	16.4	16.5	5.7	18.4	4.4	571	3729.5
Cjej	38.1	11.5	27.8	6.5	18.5	27.9	23.3	20.6	5.9	12.2	31.4	44.2	14.8	20.1	9.5	38.8	12.8	1645	6328.3
Cpne	57.9	15.9	40.2	7.1	21.6	19.6	18.9	14.9	5.4	22.6	24.8	24.9	9.6	16.8	12.6	30.7	7.3	1052	3825.2
Ctra	60.5	15.7	40.9	7.0	21.5	18.8	18.9	14.0	5.2	22.4	23.2	24.6	8.0	14.6	13.8	28.7	6.9	878	3707.7
Drad	28.3	14.6	30.2	5.1	16.0	16.4	9.6	22.1	14.2	20.0	29.9	46.4	14.0	17.0	17.4	47.9	14.3	3103	9724.2
Ecol	23.1	21.5	27.1	3.7	13.6	20.9	14.5	26.3	12.3	33.5	36.3	45.5	11.6	15.0	10.8	37.9	12.6	4289	16409.1
Hinf	41.9	17.4	33.5	5.8	19.2	23.6	10.9	19.2	4.8	26.9	26.5	43.8	16.0	16.5	10.5	34.2	14.0	1709	7615.2
Hpyl	40.7	10.6	41.6	8.0	20.1	26.7	23.5	21.3	4.3	13.5	26.7	34.1	12.9	20.2	11.6	37.7	10.7	1566	6052.1
Mgen	71.3	15.1	41.2	7.1	18.3	6.2	11.6	12.3	2.9	23.5	17.6	13.2	12.6	7.0	3.1	31.5	5.5	467	2409.3
Mpne	64.9	14.0	50.4	6.5	17.0	6.7	10.9	12.4	3.2	23.4	17.6	15.8	11.5	6.7	3.4	32.5	6.3	667	2720.2
Mtub	22.3	14.5	31.0	5.0	12.5	17.1	6.1	20.9	9.9	17.8	31.8	35.6	9.9	17.3	34.0	54.3	13.7	3918	12220.2
Nmen	40.9	13.6	39.2	6.2	15.8	28.6	12.5	19.1	4.8	16.3	31.0	40.4	14.1	19.9	10.0	35.5	12.3	2025	7820.3
Paer	19.2	28.7	16.3	3.4	13.8	18.7	18.8	30.8	18.2	16.7	31.0	49.3	8.8	15.4	18.4	43.0	13.2	5565	20511
Rpro	57.7	14.5	39.5	8.1	23.8	25.3	13.6	9.6	4.4	6.5	40.9	17.0	6.4	11.0	9.9	31.9	7.0	843	3850.6
Syne	29.3	12.6	36.8	4.3	18.8	24.2	11.1	30.9	20.9	20.5	26.5	36.0	11.1	19.9	7.9	44.8	12.2	3169	9905.8
Tmar	35.9	14.6	26.9	6.6	13.0	15.8	17.9	23.5	8.7	33.6	26.0	44.7	13.5	11.3	4.9	45.4	12.5	1846	7169.6
Tpal	54.4	16.1	41.1	9.4	25.1	24.6	28.0	10.8	6.8	23.1	16.0	14.8	8.8	9.6	6.1	38.9	10.1	1031	3623.4
Uure	67.9	13.1	51.7	5.9	15.0	4.3	13.0	16.7	2.6	13.9	13.5	13.6	10.7	8.9	3.5	31.4	7.4	611	2642.6
Vcho	28.8	22.4	27.1	5.0	16.1	18.5	26.2	25.0	20.6	25.2	29.0	42.4	12.6	18.9	10.5	37.6	16.1	5472	13033.5
Xfas	37.6	15.7	37.8	7.0	19.2	29.8	16.3	20.0	9.8	17.5	26.3	36.2	12.8	18.7	10.2	41.8	13.0	2831	7970.8
Atha	26.6	12.9	29.5	7.6	25.2	13.7	9.3	21.1	38.3	21.1	23.8	29.7	8.1	10.6	12.0	55.4	9.7	7857	18577.6
Cele	24.1	13.1	33.8	12.3	21.7	14.5	8.7	22.9	28.0	17.1	27.6	26.8	6.9	9.1	15.5	59.8	12.9	17085	19177.7
Dmel	29.8	14.5	32.2	12.0	26.4	12.5	7.7	19.0	25.9	18.6	29.6	28.4	8.3	10.0	13.4	57.3	10.7	14335	17918.6
Scer	37.7	19.8	35.7	7.1	26.1	6.9	4.6	20.5	21.5	22.4	21.6	38.0	10.8	12.3	8.3	49.3	7.0	6200	13947.5

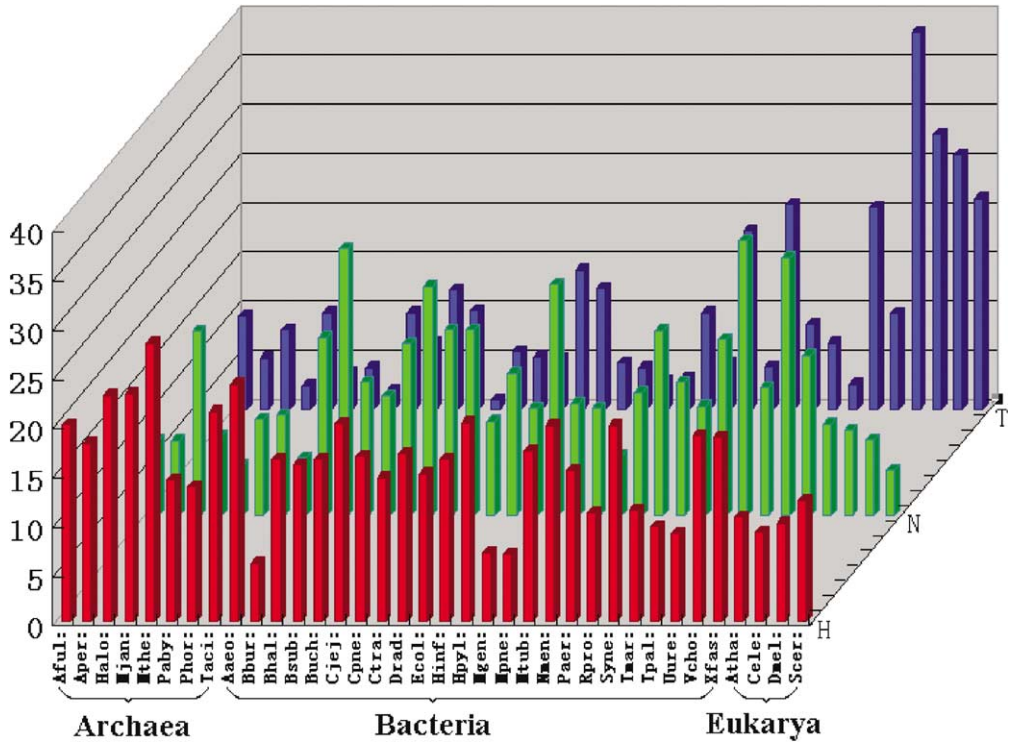

Fig. 1. The composition (relative contents) of protein genes of COG classes H, N, and T for the 36 organisms.

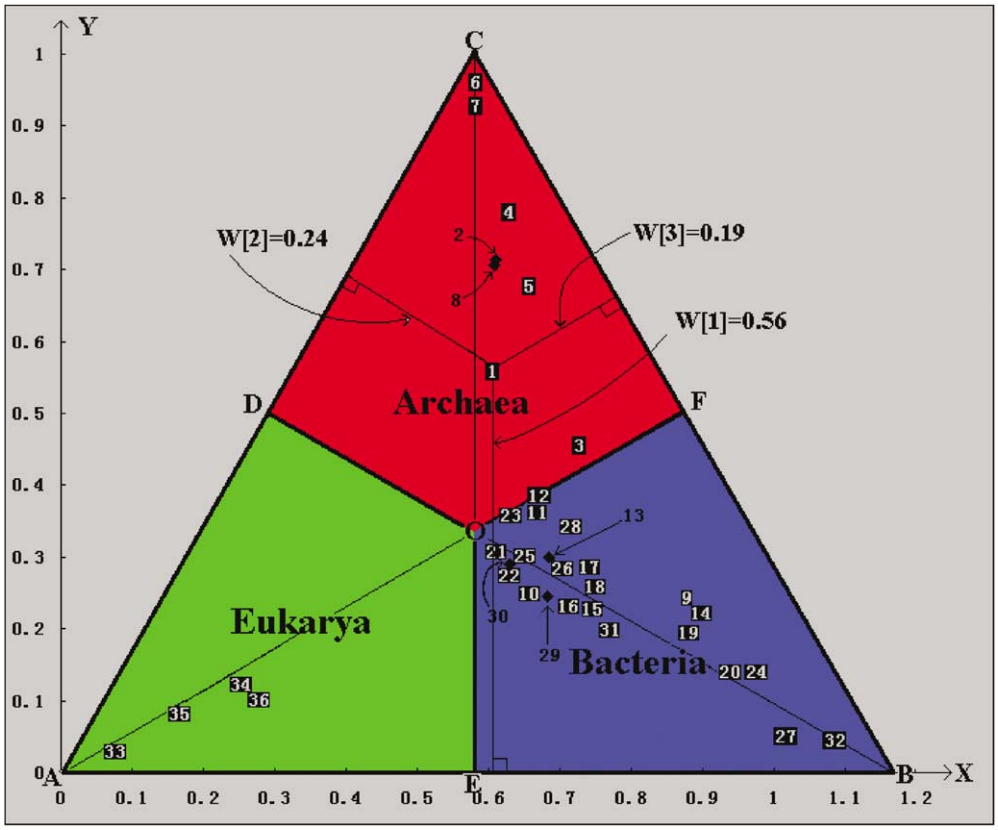

Fig. 2. Geometric representation of the fuzzy clustering result for the 36 organisms (denoted by their serial numbers listed in Table 1). The equilateral trigon ABC was equally divided into three areas, AEOD (green), BFOE (blue), and CDOF (red), representing the three primary kingdoms of Aukarya, Bacteria, and Archaea, respectively. The location for an organism is determined by $X = \sqrt{3}/3(W[1] + 2W[2]), Y = W[2]$.

tioned above (steps 4a–d) to obtain 99 sets of "resampled" proteome vectors. With the original one, we have 100 sets of proteome vectors.

(4f) Based on the 100 sets of proteome vectors, we can construct a consensus proteome tree with bootstrap-like values in nodes.

3. Results

Applying the steps of "Constructing Proteome Vectors $\{U_s\}$" described under Section 2 to the organisms listed in Table 1 (NS = 36), and letting the vector length $C = 100$, we obtained the 36 proteome vectors shown in the main frame of Table 2. The values of the last column in Table 2 are the $1/Q_s$ values [see Eq. (4)] and are proportional approximately to the respective genome sizes. Each row in the main frame forms a proteome vector representing an organism. The magnitude of a component of a proteome vector reflects the relative content of the corresponding functional class of proteins in the proteome, while each column in the main frame forms a column vector reflecting the different composition of a functional class of proteins among the organisms concerned.

Applying Eq. (5) to the proteome vectors shown in Table 2, we obtained Euclidean distances for all organism pairs and hence obtained a distance matrix for the 36 organisms (referred to as "D36" hereafter). According to D36, the distances between the organisms belonging to the same phylogenetic lineage are much smaller than the average. For example, the distance between Paby and Phor is 7.33, that between Mgen and Mpne is 11.8, and that between Ctra and Cpne is 4.92, while the average distance among all of the 36 organisms is 41.2. We also calculated the average distances within and between the three primary kingdoms. The results show that the average distances within kingdoms are 22.9 for Archaea, 40.73 for Bacteria, and 19.91 for Eukarya, indicating that organisms of both Archaea and Euharya have more similar gene compositions than those of Bacteria; the average distances between kingdoms are 43.23 for Archaea–Bacteria, 39.97 for Archaea–Eukarya, and 45.67 for Bacteria–Eukarya, indicating a closer relationship between Archaea and Eukarya, which is consistent with Woese et al.'s (1990). Three Primary Kingdoms Hypothesis.

By examining the column vectors in Table 2, we find that archaea have a higher content of genes that are related to metabolic functions such as classes C, E, F, and H, eukarya have a strikingly high content of genes that are related to "signal transduction mechanisms" (class T), and bacteria have a higher composition of genes that belong to classes M and N. Fig. 1 shows some of these characters more intuitively.

To investigate the clustering properties of our proteome vector, the FCM clustering method was used. Our practice shows that by appropriately choosing the training set and fuzzy index q, the clustering accuracy can reach as high as 100%. Table 3 shows a set of training conditions and the training results—three centroid vectors. Table 4 shows the clustering results based on the training conditions described in Table 3. We can see that the organisms of Archaea, Eukarya, and Bacteria are exactly classified into classes I, II, and III, respectively. Considering $\sum_{k=1}^{K} w_s[k] = 1$ and $K = 3$, we can associate it with the fact that "the height of an equilateral triangle equals the sum of the three distances of any inside point to the three rims," and hence we show our clustering result in an equilateral trigon, as shown in Fig. 2.

Finally, we constructed a proteome tree by using the program "neighbor" (neighbor-joining; Saitou and Nei, 1987) in the PHYLIP package with all default settings and using the distance matrix D36 as the input data. To check the reliability of this tree, we used a bootstrap-like method (as mentioned under Section 2) to generate 99 new sets of proteome vectors and accordingly obtained 99 new D36s. Then, using programs "neighbor" and "consense" in the PHYLIP package, we obtained a consensus proteome tree based on the 100 D36s (99 new

Table 3
The training of fuzzy clustering

Parameters		Q = 1.8, epsilon = 0.000001, 3 fuzzy classes																
17 COG	Classes	J	K	L	D	O	M	N	P	T	G	C	E	F	H	I	R	S
Training group	Mthe	34.3	11.8	24.9	4.4	17.0	15.7	5.1	21.2	9.9	12.4	37.3	37.9	15.9	28.3	8.0	48.5	21.9
	Phor	33.4	11.2	21.0	6.5	11.6	12.1	10.2	18.7	4.3	20.9	30.0	33.3	13.5	13.7	3.9	64.1	24.0
	Syne	29.3	12.6	36.8	4.3	18.8	24.2	11.1	30.9	20.9	20.5	26.5	36.0	11.1	19.9	7.9	44.8	12.2
	Xfas	37.6	15.7	37.8	7.0	19.2	29.8	16.3	20.0	9.8	17.5	26.3	36.2	12.8	18.7	10.2	41.8	13.0
	Atha	26.6	12.9	29.5	7.6	25.2	13.7	9.3	21.1	38.3	21.1	23.8	29.7	8.1	10.6	12.0	55.4	9.7
	Scer	37.7	19.8	35.7	7.1	26.1	6.9	4.6	20.5	21.5	22.4	21.6	38.0	10.8	12.3	8.3	49.3	7.0
Training result	Center_1	33.8	11.5	22.7	5.8	13.8	13.4	8.4	19.7	6.6	17.9	32.4	35.0	14.3	18.9	5.5	58.2	23.0
	Center_2	33.8	14.3	36.7	5.7	19.1	26.1	13.2	24.8	15.0	18.8	26.7	36.2	12.1	19.4	9.0	43.6	12.8
	Center_3	30.9	15.5	31.9	7.3	25.3	11.3	7.5	20.9	31.4	21.5	23.2	33.0	9.2	11.6	10.5	52.9	8.8

Table 4
Fuzzy clustering result and the geometry representation

Organisms			Fuzzy class membership				Geometry[a]	
Serial number	Name	Domain	W[1]	W[2]	W[3]	Class	X	Y
1	Aful	Archaea	0.5620	0.2441	0.1940	I	0.606	0.562
2	Aper	Archaea	0.7139	0.1713	0.1147	I	0.610	0.714
3	Halo	Archaea	0.4570	0.3977	0.1453	I	0.723	0.457
4	Mjan	Archaea	0.7815	0.1520	0.0665	I	0.627	0.781
5	Mthe	Archaea	0.6784	0.2277	0.0938	I	0.655	0.678
6	Paby	Archaea	0.9596	0.0241	0.0163	I	0.582	0.960
7	Phor	Archaea	0.9298	0.0389	0.0313	I	0.582	0.930
8	Taci	Archaea	0.7068	0.1728	0.1203	I	0.608	0.707
9	Aaeo	Bacteria	0.2461	0.6328	0.1211	II	0.873	0.246
10	Bbur	Bacteria	0.2499	0.4484	0.3017	II	0.662	0.250
11	Bhal	Bacteria	0.3614	0.3940	0.2445	II	0.664	0.361
12	Bsub	Bacteria	0.3816	0.3876	0.2307	II	0.668	0.382
13	Buch	Bacteria	0.2987	0.4432	0.2581	II	0.684	0.299
14	Cjej	Bacteria	0.2237	0.6609	0.1154	II	0.892	0.224
15	Cpne	Bacteria	0.2302	0.5287	0.2411	II	0.743	0.230
16	Ctra	Bacteria	0.2365	0.5039	0.2596	II	0.718	0.236
17	Drad	Bacteria	0.2871	0.4984	0.2145	II	0.741	0.287
18	Ecol	Bacteria	0.2621	0.5156	0.2223	II	0.747	0.262
19	Hinf	Bacteria	0.1966	0.6641	0.1393	II	0.880	0.197
20	Hpyl	Bacteria	0.1417	0.7478	0.1105	II	0.945	0.142
21	Mgen	Bacteria	0.3005	0.3839	0.3156	II	0.617	0.301
22	Mpne	Bacteria	0.2839	0.4015	0.3146	II	0.628	0.284
23	Mtub	Bacteria	0.3569	0.3667	0.2763	II	0.629	0.357
24	Nmen	Bacteria	0.1409	0.7701	0.0889	II	0.971	0.141
25	Paer	Bacteria	0.2971	0.4078	0.2951	II	0.642	0.297
26	Rpro	Bacteria	0.2841	0.4604	0.2556	II	0.696	0.284
27	Syne	Bacteria	0.0520	0.8527	0.0953	II	1.015	0.052
28	Tmar	Bacteria	0.3398	0.4498	0.2105	II	0.716	0.340
29	Tpal	Bacteria	0.2457	0.4682	0.2862	II	0.682	0.246
30	Uure	Bacteria	0.2902	0.4005	0.3093	II	0.630	0.290
31	Vcho	Bacteria	0.1977	0.5678	0.2344	II	0.770	0.198
32	Xfas	Bacteria				II	1.082	0.047
33	Atha	Eukarya				III	0.074	0.031
34	Cele	Eukarya				III	0.252	0.122
35	Dmel	Eukarya				III	0.165	0.081
36	Seer	Eukarya				III	0.272	0.102

[a] $X = \frac{\sqrt{3}}{3}(W[1] + 2W[2]),\ Y = W[2]$.

ones plus the original one) as shown (by program "TreeView;" see http://taxonomy.zoology.gla.ac.uk/rod/rod.html) in Fig. 3. The topology of this consensus tree is exactly the same as that derived from the original D36. Looking into this consensus tree, we can see that all the organisms belonging to the same phylogenetic lineage are located in the nearest treetops with high bootstrap-like values. If we regard the highlighted and circled node of the tree as the "root," then the 8 archaea plus Mtub and Drad form the first branch, the 4 eukarya plus Syne form the second branch, and the rest of the organisms (21 bacteria) form the third branch. Furthermore, the former two branches are closer to each other than to the third branch. These results are basically consistent with Woese et al.'s Three Primary Kingdoms Hypothesis, but there are some small differences. It seems that, although our proteome tree reflects more phenotypic than historical relationships among organisms, it does reveal phylogenetic information to a large extent.

4. Discussion

Some organisms involved in this study have more than one set of data available for different strains. These different strains are CpneA (Klenk *Chlamydia pneumoniae* AR39), CpneJ (*C. pneumoniae* J138), CtraM (*Chlamydia trachomatis* MoPn), Hp99 (*Helicobacter pylori* J99), and NmenA (*Neisseria meningitidis* serogroup A strain Z2491). To check the efficiency and validity of our method, we put all these different strains together with those listed in Table 1 and did the same calculations. The distances between the different strains of the same organisms are 0.79 for Cpne and CpneA, 0.93 for Cpne and CpneJ, 1.35 for CpneA and CpneJ, 2.84 for Ctra

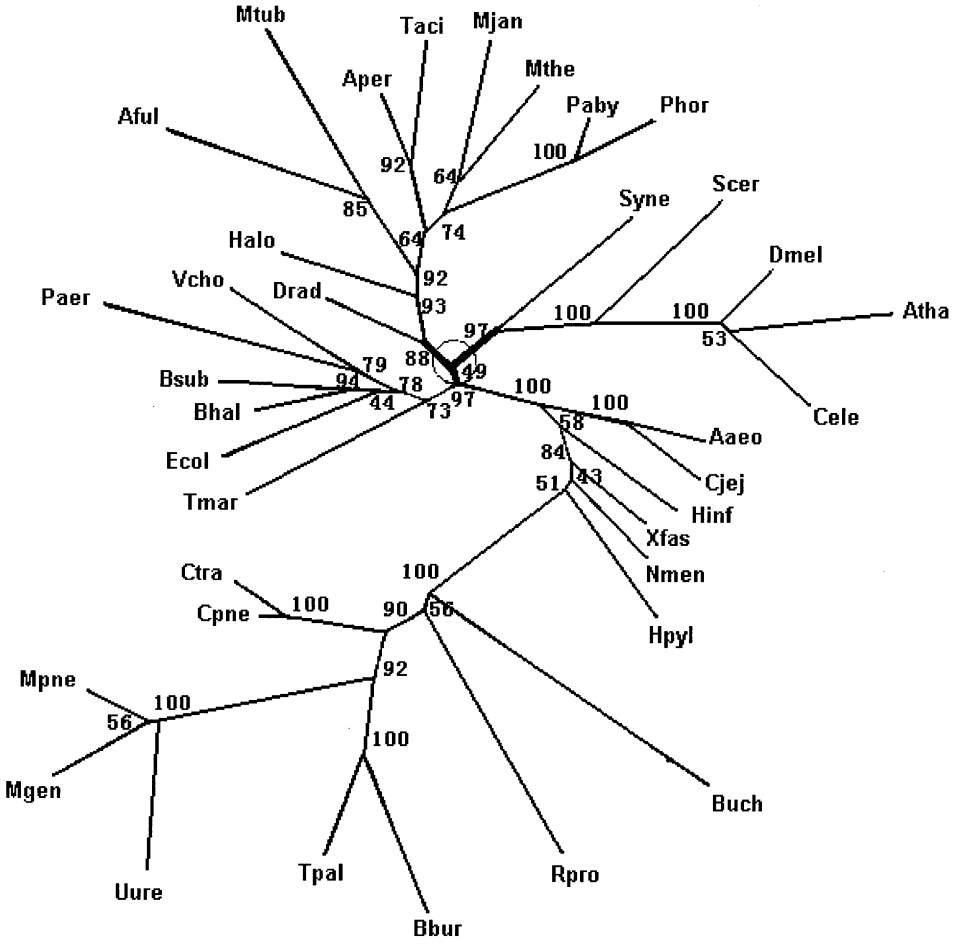

Fig. 3. The proteome tree for the 36 organisms derived from the proteome vector distances.

and CtraM, 4.31 for Hpy1 and Hp99, and 3.79 for Nmen and NmenA. Obviously, all of these distances are much less than the distances between two different organisms. We noticed that CpneA and CpneJ have almost the same genome size of about 1.23 Mb, but the numbers of predicted protein genes are quite different: 2220 for CpenA but only 1070 for CpneJ. No doubt the genes were either overpredicted for CpenA or insufficiently predicted for CpenJ or both. Even so, the distance between them is as small as 1.35. We think that this is because of the following two reasons. First, our proteome vector is a normalized vector. It is affected mainly by the functional attributes, instead of by the numbers of genes in the genome. Second, for those overpredicted genes, they should be quite different from any COG sequences, hence contributing little to the proteome vector. So, our proteome vectors can dispel the negative influences caused by gene overprediction.

But how would the insufficient gene prediction or incomplete source data affect our proteome vectors? We selected the genomes with more than one subunit (chromosome and plasmid), picked up all of those subunits containing more than 100 genes, and then treated them equally and calculated the corresponding vectors in the same way as described under Section 2 (Step 2). Based on these subunit vectors, we constructed a chromosome/plasmid tree, as shown in Fig. 4. Looking into this tree, we can find an interesting phenomenon—a chromosome (or plasmid) is generally closer to those from the same genome than to those from different genomes. Does this mean a general rule of self-similarity within genome chromosomes (or plasmids)? Further investigation based on a broader range of species and more comprehensive data should be performed to answer this question. But at least, this phenomenon implies that insufficient gene prediction or incomplete source data would not badly affect our proteome vector. This may be why Atha was still correctly clustered together with three other Eukarya, though only three of its chromosomes were available at the time and were used in our study.

Looking into our proteome tree (Fig. 3), we can find some noteworthy points. Mtub is located in the branch of Archaea, indicating that it has quite a different gene composition compared to other Bacteria. This is consistent with the facts reported by Cole et al. (1998). Drad

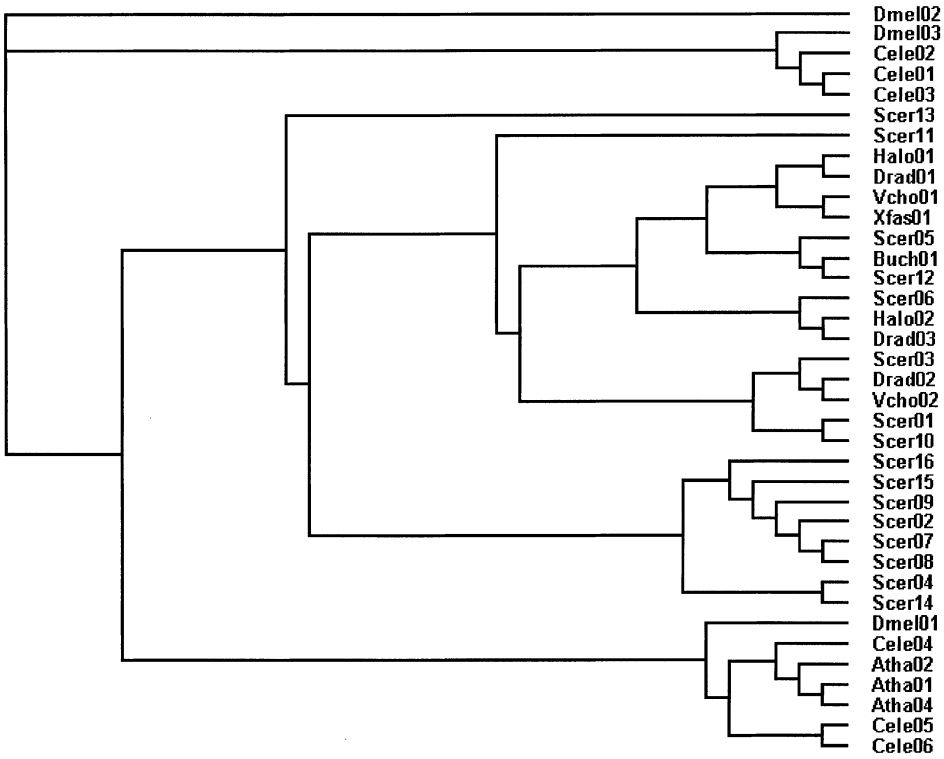

Fig. 4. The tree based on chromosome/plasmid vectors. Cele01, chromosome I of *Caenorhabditis elegans*; Scer02, chromosome II of *Saccharomyces cerevisiae*; and so on.

is located at the bottom of the Archaea branch, indicating that its radiation-resistant property (White et al., 1999) and archaea's ability to survive in extreme environments might result from their similar gene composition. Syne is located at the bottom of the Eukarya branch, which is perhaps because of its photosynthesis ability—the basic property of plants. We found that there are some organisms that are located in unexpected positions without obvious reasons, but have lower bootstrap-like values, such as *Escherichia coli*.

We had tried to exclude the COG classes of S and R from our study, but obtained a worse result. This fact indicates that although they are proteins of "general function prediction only" and "function unknown," they are still important for genomes and hence could not be neglected.

Our method involves all genes, including those that might be related to LGT. Even so, our results are still basically consistent with the Three Primary Kingdoms Hypothesis. This fact suggests that LGT might not be so "rampant" and affects the proteome vector only slightly.

In summary, our conclusions are

1. Our results support Woese et al.'s Three Primary Kingdoms Hypothesis.
2. The composition of the proteins with different functions in a proteome (our proteome vector) is a good criterion for clustering and to a large extent reveals phylogenetic properties of organisms. It has two advantages: first, it can dispel the negative influences brought by protein gene over prediction or insufficient prediction; second, it can be used to compare two organisms with great genome size differences.
3. For the organisms used in this study, self-similarities for within-genome chromosomes (or plasmids) might exist.

Acknowledgments

The study was supported by the Grants 39890070, 19890380, and 39993420 from the China National Foundation of Science (CNFS), by the Grants KSCX2-2-07 and KJCX1-08 from the Chinese Academy of Science, and by a special grant the from the Science and Technology Committee of Beijing.

References

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Cole, S.T., Brosch, R., Parkhill, J., et al., 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393, 537–544.

Doolittle, R.F., 1998. Microbial genomes opened up. Nature 392, 339–342.

Doolittle, R.F., Logsdon Jr., J.M., 1998. Archaeal genomics: do Archaea have a mixed heritage? Curr. Biol. 8, R209–R211.

Fitz-Gibbon, S.T., House, C.H., 1999. Whole genome-based phylogenetic analysis of free living microorganisms. Nucleic Acid Res. 27, 4218–4222.

Garcia-Vallvé, S., Romeu, A., Palau, J., 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 11, 1719–1725.

Gupta, R.S., 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes. Microbiol. Mol Biol. Rev. 62, 1435–1491.

Lin, J., Gerstein, M., 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10, 808–818.

Mayr, E., 1998. Two empires or three? Proc. Natl. Acad. Sci. USA 95, 9720–9723.

Montague, M.G., Hutchison III, C.A., 2000. Gene content phylogeny of herpesviruses. Proc. Natl. Acad. Sci. USA 97, 5334–5339.

Pennisi, E., 1999. Is it time to uproot the tree of life? Science 284, 1305–1307.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Snel, B., Bork, P., Huynen, M.A., 1999. Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637.

Tatusov, R.L. et al., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29, 22–28.

Tekaia, F., Lazcano, A., Dujon, B., 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. 9, 550–557.

White, O. et al., 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286, 1571–1577.

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl Acad. Sci. USA 87, 4576–4579.

Yap, W.H., Zhang, Z., Wang, Y., 1999. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. 181, 5201–5209.

Zhang, C.T., Chou, K.C., Maggiora, G.M., 1995. Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Engineer. 8, 425–435.