J. Microbiol. Biotechnol. (2009), **19**(11), 1301–1305
doi: 10.4014/jmb.0905.05017
First published online 24 August 2009

JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY

Elucidation of Multifaceted Evolutionary Processes of Microorganisms by Comparative Genome-Based Analysis

Nguyen, Thuy Vu An[1], Soon Ho Hong[1]*, and Sang Yup Lee[2,3]

[1]*School of Chemical Engineering and Bioengineering, University of Ulsan, Ulsan 680-749, Korea*
[2]*Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Program), Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea*
[3]*BioProcess Engineering Research Center, Bioinformatics Research Center, Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea*

Received: May 12, 2009 / Revised: June 18, 2009 / Accepted: June 19, 2009

The evolution of living organisms occurs *via* a combination of highly complicated processes that involve modification of various features such as appearance, metabolism and sensing systems. To understand the evolution of life, it is necessary to understand how each biological feature has been optimized in response to new environmental conditions and interrelated with other features through evolution. To accomplish this, we constructed contents-based trees for a two-component system (TCS) and metabolic network to determine how the environmental communication mechanism and the intracellular metabolism have evolved, respectively. We then conducted a comparative analysis of the two trees using ARACNE to evaluate the evolutionary and functional relationship between TCS and metabolism. The results showed that such integrated analysis can give new insight into the study of bacterial evolution.

Keywords: Integrated evolutionary analysis, genome sequences, metabolic networks, two-component systems, ARACNE

Biological activities of living organisms are represented by several features such as substrate consumption, metabolism, metabolite excretion, reproduction, cellular communication, motility, and interaction with the environment. When a living organism is exposed to a new environment, it does not modify a single gene such as 16S rRNA, but it instead modifies multiple features until their metabolic and physiological characteristics are optimized for the new environmental condition. Therefore, the evolution of living organisms is achieved through integrated, highly sophisticated, and complex processes [8, 22, 24–26]. For this reason, it has been suggested that traditional single gene-based evolutionary trees are not sophisticated enough to represent the complicated and integrated evolutionary process of living organisms, even though 16S and 23S rRNA sequences-based analyses produce reliable and reproducible results [9, 29, 30].

Systems level analysis of the complete genome sequences of living organisms has become a new paradigm of biological research [1, 12, 14, 16, 18, 19]. Various complete genome-based analyses of evolution have been conducted to overcome the drawbacks associated with the single gene-based analysis [2, 3, 5, 10, 13, 17, 20, 27, 28]. However, to understand such processes of the evolution, it is important to determine how living organisms have modified their various features and how the evolutionary processes associated with these features are interconnected with one another.

Here, we report the results of an integrated microbial evolutionary analysis that was conducted based on the quantitative structural analysis of multiple trees and Algorithm for the Reconstruction of Accurate Cellular Networks or ARACNE-mediated connectivity analysis [21]. To accomplish this, we constructed two trees based on the metabolic network content and the two-component system (TCS) content, which reflect the evolutionary histories of intracellular metabolism and the environmental communication mechanism, respectively. Next, the evolutionary relationships among strains were quantified by measuring the relative distances among strains in each tree. Then, comparative analysis was conducted to estimate the evolutionary characteristics of the TCS system and metabolism. ARACNE has been used to reveal functional connections among genes from microarray data [6, 11, 21]. In this study, ARACNE analysis was employed to determine how metabolic networks and the TCS have become interconnected during evolution.

*Corresponding author
Phone: +82-52-259-1293; Fax: +82-52-259-1689;
E-mail: shhong@mail.ulsan.ac.kr

Two evolutionary trees were constructed based on the TCS content and metabolic network content of microbial genomic data present in the KEGG database and the literature [13, 15, 17]. To accomplish this, the TCS genes and metabolic network genes from 77 microorganisms were collected to construct a TCS content matrix and metabolic network content matrix, respectively (Supplementary Tables S1 and S2). The TCS content matrix and metabolic network content matrix describe the distribution of the TCS and metabolic network-related genes in each strain (Fig. 1).

Hierarchical clustering was performed with the Cluster program package using the complete-linkage hierarchical clustering algorithm [7] to construct the TCS content and metabolic network content-based trees. The trees were then visualized using the TreeView software package [7]. The distance between two strains in each tree was then quantified by measuring the length from the root to the separating branch, after which the individual distances were divided by the longest distance to determine the relative distances (Fig. 1).

ARACNE generates a putative transcriptional network in two primary steps. First, ARACNE computes mutual information (MI) using the Gaussian kernel density. Each Mij represents the relatedness for a pair of i and j in the data set. The key elements in this step are similar to those of the Relevance Networks method and include determination of the parameters that are used to compute the MI and determination of the MI threshold for statistical independence [4]. In the second step, ARACNE attempts to eliminate indirect relationships in which two data points are co-regulated through one or more intermediaries using a well-known property of MI called the data processing inequality (DPI). Hence, the relationships included in the final reconstructed network have a high probability of direct interactions.

ARACNE was employed to construct relationship networks among the TCS and metabolic network genes. Briefly, the TCS genes and metabolic network-related genes were collected from the KEGG database and the literature. The metabolic network-related genes were then grouped into 163 subpathways according to their functions, after which the 291 TCS-related genes and the 163 metabolic subpathway groups were arranged into one input matrix. It was necessary to define several parameters to obtain good network results. The first required parameter was the kernel width of the Gaussian estimator used for the MI estimation. The optimal choices of the kernel width depend on the sample size and statistics of the data set; for the present study, a kernel width of 0.15 was used. The second parameter was the *p*-value, which represents the significance threshold. The choice of significance threshold depends on the desired trade-off between false-positives and false-negatives. For this study, the *p*-value was set to 1E-7. Finally, the reconstructed networks were viewed using the Network Browser [6].

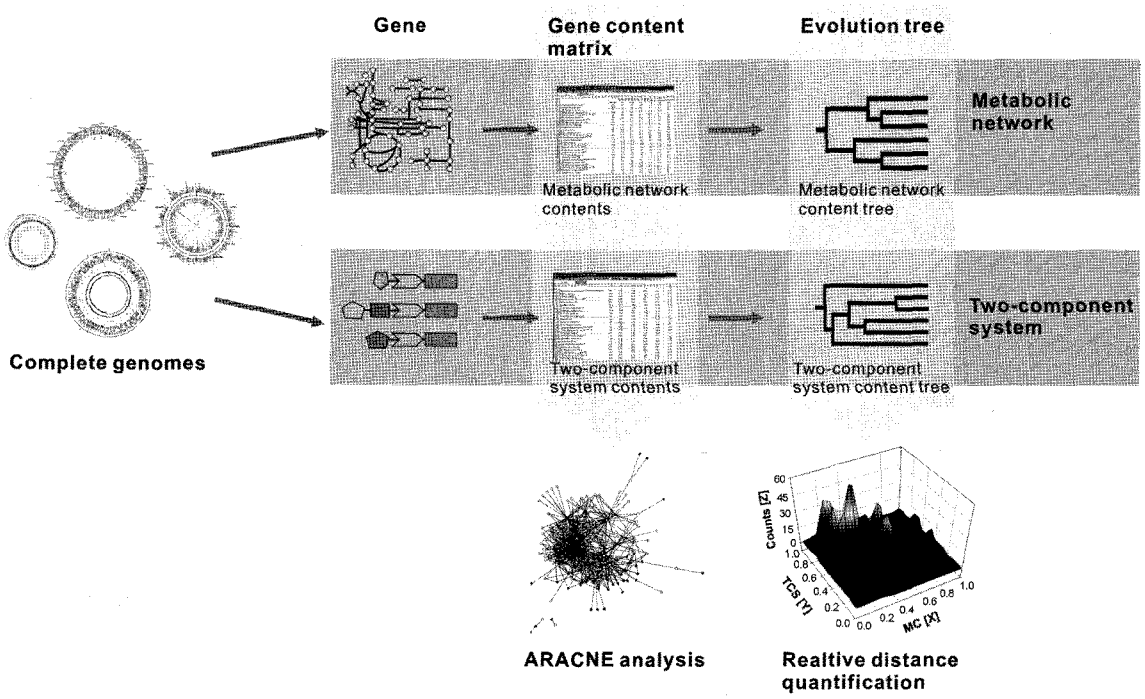

Fig. 1. Schematic diagram of the experimental procedure.
From the complete genome, metabolic network and two-component system (TCS)-related genes were extracted and summarized in gene content matrixes. The metabolic network matrix and TCS matrix were analyzed by ARACNE. Comparative analyzis of metabolic network- and TCS-based trees were also conducted.

To gain a better understanding of the complex evolutionary process, we conducted an integrated analysis of the bacterial evolutionary processes. Specifically, two different evolutionary trees were constructed based on the TCS content and the metabolic network content of 77 microorganisms. The generated trees were expected to reflect the evolutionary histories of the environmental communication mechanism and intracellular metabolism, respectively (Fig. 1; Supplementary Fig. S1 and S2). For the tree that was based on the contents of the TCS, 2,926 relative distances among tested strains were measured. Those relative distances represent the distance between strains in the TCS tree as well as their evolutionary relationships (*e.g.*, the long distance in the tree suggests that there is less evolutionary relationship between the strains). A relative distance matrix of the metabolic network content-based tree was also constructed (Fig. 1; Supplementary Table S3 and S4).

The constructed relative distance matrixes were then integrated and visualized in three-dimensional space to estimate multiple aspects of the evolutionary processes (Fig. 2). The X value represents the distance between two strains in the metabolic network-based tree, whereas the Y value represents the evolutionary distances between two strains in the TCS-based tree (Fig. 2). The Z axis shows the number of relationships (or strain pairs) with the same (X, Y) coordinates. The integrated graph of the TCS–metabolic network contents revealed that the profile was not evenly distributed, but contained several groups at specific points. These findings suggest that the TCS and metabolic networks have evolved with certain tendencies, rather than through a series of random mutations (Fig. 2).

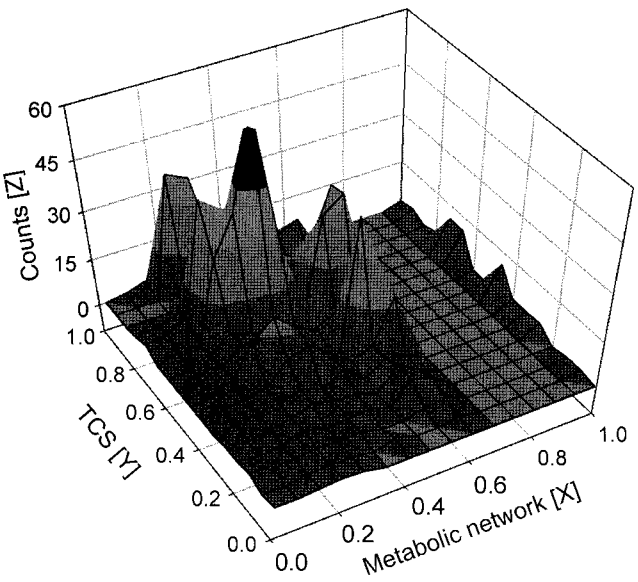

Fig. 2. Integrated evolutionary graphs.
The X and Y values represent the distances between two strains in the metabolic network-based tree and the TCS-based tree, respectively. The Z axis represents the number of strain pairs with the same (X, Y) coordinates.

The peaks were generally located near the Y axis, which indicated that the majority of relationships showed a larger Y value than X value. In addition, the metabolic network content-based relative distances were generally distributed in the 0.2–0.4 region and had an average value of 0.36. The distances between the organisms in the TCS content-based trees were mostly distributed in the 0.6–0.8 region and had an average value of 0.65. These findings indicated that the TCS contents varied much more than the metabolic network contents during the evolution, and suggest that the contents of the TCS can be more easily altered for the organisms to adapt to new environmental conditions. This is supported by the finding that 2,473 strain pairs (84.5% of the total related pairs) had Y values that were greater than their corresponding X values.

The *Escherichia coli* K-12 and *Salmonella typhimurium* pair represents a good example of the versatility of the TCS. The (X, Y) coordinate of the *E. coli* K-12 and *S. typhimurium* pair was (0.06, 0.29), which indicates that the contents of their metabolic network were almost identical to the metabolic network content, but that their TCS contents differed (Supplementary Table S3 and S4). These findings suggest that *S. typhimurium* became pathogenic through the modification of their TCS, whereas their metabolic networks remained almost the same as those of *E. coli* K-12.

Integrated evolutionary analysis of *Bacillus subtilis, Streptococcus pyogenes*, and *Lactococcus lactis* also provided interesting results. Although it is well known that these strains were closely related evolutionarily, they have completely different phenotypic characteristics. Specifically, *B. subtilis* is a free-living Gram-positive bacterium, whereas *S. pyogenes* is a Gram-positive parasitic pathogen and *L. lactis* is a nonpathogenic lactic acid producing bacterium. The coordinate of the *B. subtilis* and *L. lactis* pair was (0.22, 1.0), which suggests that intensive modification of the two-component systems has occurred throughout the evolutionary process. However, the results of the *B. subtilis* and *S. pyogenes* pair were different. Specifically, the coordinate of the *B. subtilis* and *S. pyogenes* pair was (0.44, 0.65), which indicates that the pathogen *S. pyogenes* modified or lost its unnecessary metabolic networks to become a parasitic strain while also undergoing modification of its TCS. These findings demonstrate that concomitant alteration of metabolic networks and the TCS are required if mere modification of the TCS does not enable an organism to survive in new extreme conditions.

Archaea exhibited different evolutionary characteristics. Eubacteria represented by the previous examples clearly showed that the TCS network underwent adaptive evolution more readily than the metabolic network. However, among archaea, the relative distances between two species in the tree generated based on the metabolic contents were greater than those in the TCS tree, which indicates that the metabolic contents underwent adaptive evolution more significantly.

For example, the coordinates of the *Methanosarcina mazei*–*Methanosarcina acetivorans* pair and the *M. mazei*–*Halobacterium* sp. pair were (0.37, 0.13) and (0.56, 0.01), respectively (Tables S3 and S4). In combination with the result that archaea have only 3 or 4 TCS genes whereas bacteria have 20 TCS genes in average, it can be deduced that TCS was a newly evolved apparatus that was not fully developed in early evolved archaea.

To evaluate the evolutionary relationship between the TCS and metabolic networks, putative relationship networks were constructed using ARACNE based on the TCS content matrix and the metabolic network content matrix (Fig. 1). ARACNE analysis resulted in the construction of a relationship network that consisted of 142 nodes (93 metabolic subpathways and 49 TCS genes) and 563 edges (Fig. 3). The TCS–TCS or metabolic network–metabolic network relationships were not considered, and only the TCS–metabolic network relationships were established. Only 31% of the tested data (163 subpathways and 291 TCS genes) was incorporated in the produced relationship network. When the TCS genes were evaluated, only 49 of 291 genes (17%) were included in the relationship network, whereas more than half of the metabolic subpathways (93 out of 163) were included. These findings indicate that most of the TCS genes have evolved independently of other TCS or metabolic pathways, which is supported by the results of previously conducted studies [23].

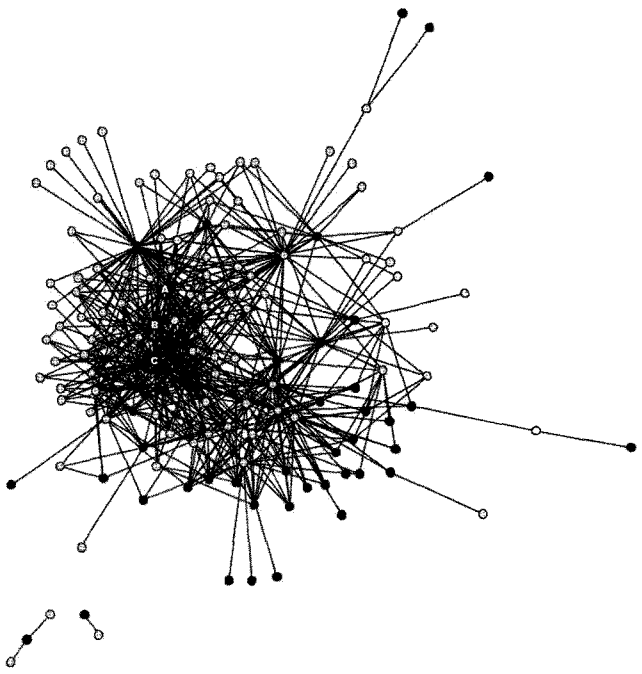

Fig. 3. The relationship network describing two-component system genes and metabolic subpathways.
This network includes 49 TCS genes (represented by grey circles), which are directly connected to 93 metabolic subpathways (represented by white circles). The busiest nodes are labeled: A, *glnG*; B, *glnL*; C, *narL*.

When the constructed relationship network was evaluated, each node was found to have an average of 7.9 relationships. Specifically, each TCS gene node was connected with an average of 11.5 metabolic network nodes, and the busiest TCS nodes (*glnG* and *glnL*) were connected with 49 metabolic network nodes (Supplementary Table S5). Of the 49 TCS nodes, 16 of them had greater than 10 connections with metabolic subpathways, and these nodes were fairly well distributed (Supplementary Table S5). In the case of the metabolic subpathways, only a small number of nodes (13 out of 93) had more than 10 connections and most of them (86%) had less than 10 connections. Furthermore, only 3 nodes were found to have a relationship with more than 20 TCS nodes (Supplementary Table S5). Several TCS genes were found to be related to approximately 50 metabolic subpathways, and only a small number of the TCS genes were related to metabolic subpathways. Taken together, these results indicate that some TCS genes exert a global effect on cellular metabolism, but most TCS genes are not related to metabolism.

Analysis of the hub or core nodes in the ARACNE-generated relation network enabled identification of the factors that exert global effects on evolution. The *glnG* and *glnL* gene, which are composed of nitrogen-related TCS genes, act as highly connected hubs and have relationships with 49 metabolic subpathways. The nitrate-related *narL* gene and the phosphate-related *phoB* and *phoR* genes share more than 30 connections. In addition, the *envZ* (osmotic pressure), and *cheA* and *cheB* (chemotaxis) genes share more than 20 connections. These results suggest that the nitrogen-, phosphate-, osmotic pressure-, and chemotaxis-related TCS genes were subjected to global environmental pressure during the evolutionary process.

The functional relationship between the TCS and metabolism can also be deduced from ARACNE analysis. In the case of the nitrate responsive TCS (encoded by the *narL* gene), 12 out of 45 connected metabolic subpathways were found to be related to amino acid metabolism. This high connectivity between *narL* and amino acid metabolism indicates that there is a functional relationship between nitrate availability and amino acid biosynthesis. This can be explained by the well-known fact that atmospheric nitrogen is initially fixed as inorganic nitrate and ultimately as amino groups in amino acids.

Although ARACNE analysis indicated that the majority of TCS genes do not share an evolutionary relationship with cellular metabolism, these findings do not mean that the TCS is not related to other functional systems such as signal networking and quorum sensing circuits. Additional studies are under way to decipher these relationships.

In this study, we reported the results of an integrated analysis that was designed to evaluate complex and multifaceted evolutionary processes. This strategy provided us with greater insight regarding the evolutionary process

in microorganisms. Because life is a combination of various biological functions, an integrated and multidimensional analysis is required to reveal the relationships among these functions that have developed with time as a result of evolution. We believe that this study provides a good example of integrated evolutionary analysis based on cellular functions represented by the TCS and metabolism. Expansion of this study to include further multidimensional integrated evolutionary analyses will enable us to decipher more valuable biological information during the course of evolution.

Acknowledgments

This study was supported by the Korean Systems Biology Program from the Ministry of Education, Science and Technology through the Korea Science and Engineering Foundation (No. M10503020001-07N0302-00112), and a Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-211-D00026). Further supports by the LG Chem Chair Professorship and Microsoft (S.Y.L.) are appreciated.

REFERENCES

1. Asenjo, J. A., P. Ramirez, I. Rapaport, J. Aracena, E. Goles, and B. A. Andrews. 2007. A discrete mathematical model applied to genetic regulation and metabolic networks. *J. Microbiol. Biotechnol.* **17:** 496–510.
2. Bansal, A. K. 1999. An automated comparative analysis of 17 complete microbial genomes. *Bioinfomatics* **15:** 900–908.
3. Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope. 2001. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28:** 281–285.
4. Butte, A. J. and I. S. Kohane. 2000. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* **5:** 418–429.
5. Daubin, V., M. Gouy, and G. Perriere. 2002. A phylogenic approach to bacterial phylogeny: Evidence of core genes sharing a common history. *Genome Res.* **12:** 1080–1090.
6. Duarte, N. C., S. A. Becker, N. Jamshidi, I. Thiele, M. I. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.* **104:** 1777–1782.
7. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 14863–14868.
8. Feng, D. F., G. Cho, and R. F. Doolittle. 1997. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci. U.S.A.* **94:** 13028–13033.
9. Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. *Science* **155:** 279–284.
10. Fitz-Gibbon, S. T. and C. H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27:** 4218–4222.
11. Franckea, C., R. J. Siezena, and B. Teusink. 2005. Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* **13:** 550–558.
12. Hong, S. H. 2007. Systems approaches to succinic acid-producing microorganisms. *Biotechnol. Bioprocess Eng.* **12:** 73–79.
13. Hong, S. H., T. Y. Kim, and S. Y. Lee. 2004. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Appl. Microbiol. Biotechnol.* **65:** 203–210.
14. Ideker, T., T. Galitski, and L. Hood. 2001. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* **2:** 343–372.
15. Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30:** 42–46.
16. Kitano, H. 2002. Systems biology: A brief overview. *Science* **295:** 1662–1664.
17. Kim, J. S. and S. Y. Lee. 2006. Genomic tree of gene contents based on the functional groups of KEGG orthology. *J. Microbiol. Biotechnol.* **16:** 748–756.
18. Kim, T. Y. and S. Y. Lee. 2006. Accurate metabolic flux analysis through data reconciliation of isotope balance-based data. *J. Microbiol. Biotechnol.* **16:** 1139–1143.
19. Lee, S. Y., H. M. Woo, D-Y. Lee, H. S. Choi, T. Y. Kim, and H. Yun. 2005. Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol. Bioprocess Eng.* **10:** 425–431.
20. Ma, H. W. and A. P. Zeng. 2004. Phylogenetic comparison of metabolic capacities of organisms at genome level. *Mol. Phylogenet. Evol.* **31:** 204–213.
21. Margolin, A. A., K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano. 2006. Reverse engineering cellular networks. *Nat. Protocol* **1:** 663–672.
22. Meyer, T. E., M. A. Cusanovich, and M. D. Kamen. 1986. Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* **83:** 217–220.
23. Nguyen, T. V. A. and S. H. Hong. 2008. Whole genome-based phylogenetic analysis of bacterial two-component systems. *Biotechnol. Bioprocess Eng.* **13:** 288–292.
24. Olsen, G. J., C. R. Woese, and R. Overbeek. 1994. The wind of (evolutionary) change: Breathing new life into microbiology. *J. Bacteriol.* **176:** 1–6.
25. Ribeiro, S. and G. B. Golding. 1998. The mosaic nature of the eukaryotic nucleus. *Mol. Biol. Evol.* **15:** 779–788.
26. Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake. 1998. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 6239–6244.
27. Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.
28. Tekaia, F., A. Lazcano, and B. Dujon. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9:** 550–557.
29. Woese, C. R. 1987. Bacterial evolution. *Microbiol. Rev.* **51:** 221–271.
30. Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8:** 357–366.