Journal of Biological Physics **28:** 439–447, 2002.

Phylogeny Based on Whole Genome as inferred from Complete Information Set Analysis

W. LI[1,3], W. FANG[2,3], L. LING[1], J. WANG[1], Z. XUAN[1] and R. CHEN[1,*]
[1]*Laboratory of Bioinformatics, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*
[2]*Academy of Mathematical and Systemic Sciences, Chinese Academy of Sciences, Beijing 100080, China*
[3]*These authors contributed equally to this work*
(*Author for correspondence, e-mail: crs@sun5.ibp.ac.cn)

Abstract. Previous molecular phylogeny algorithms mainly rely on multi-sequence alignments of cautiously selected characteristic sequences, thus not directly appropriate for whole genome phylogeny where events such as rearrangements make full-length alignments impossible. We introduce here the concept of Complete Information Set (CIS) and its measurement implementation as evolution distance without reference to sizes. As method proof-test, the 16s rRNA sequences of 22 completely sequenced Bacteria and Archaea species are used to reconstruct a phylogenetic tree, which is generally consistent with the commonly accepted one. Based on whole genome, our further efforts yield a highly robust whole genome phylogenetic tree, supporting separate monophyletic cluster of species with similar phenotype as well as the early evolution of thermophilic Bacteria and late diverging of Eukarya. The purpose of this work is not to contradict or confirm previous phylogeny standards but rather to bring a brand-new algorithm and tool to the phylogeny research community. The software to estimate the sequence distance and materials used in this study are available upon request to corresponding author.

Key words: comparative genomics, information discrepancy, molecular evolution, sequence analysis

1. Introduction

The fast advance of worldwide genome sequencing projects affords unprecedented opportunities and perspectives for dissecting evolutionary relationships. One of the major open problems concerns the whole-genome phylogeny [1]. Traditional phylogenetic reconstruction and classifications have relied on phenotypic (morphological, physiological, behavioral) characters and paleontological records. The established molecular phylogenies often have been based on cautiously selected single characteristic sequence with emphasis on 16S rRNA, which led to the proposal of the three primary kingdoms or domains (Eukarya, Bacteria, and Archaea) [2]. However, often one gets varied conclusions using different characteristic sequences [3], which not only reflect classical problems due to horizontal gene transfer [4], unrecognized paralogy and highly variable rates of evolution but also

highlight the fact that the evolution distance between genes and not between entire genomes. Therefore, it is not exactly to equate the history of whole genome with the history of only a portion or even the majority of its contents. Moreover, experimental evidence is now available indicating that rRNA itself can be horizontally transferred between organisms [5]. Hence, people begin to doubt all trees based on a single characteristic sequence, even further, one doubt the dogma of 'three kingdom', some even indicated it may be impossible to construct a 'universal tree' [6].

Previously existing algorithms such as multiple alignment and various sequence evolutionary models can not directly apply to complete genomes where events such as rearrangements make full-length alignments impossible. The computational complexity will boost exponentially when the sequence number or size increases, thus, impossible to deal with huge genome data; furthermore, different sequence order and alternative options such as the score matrix often led to varied phylogenetic trees. Fortunately, there have already been proposals of whole genome phylogeny using gene order [7], gene content [8], folds [9] and large combined protein sequences [10]. Such methods are time consuming with human intervention and only use not whole but partial information extracted from genome in the opinion of respective considerations.

Because biological sequences encode information, and the occurrence of evolutionary events (such as insertions, deletions, point mutations, inversions and rearrangements) separating two sequences sharing a common ancestor will result in the loss of their shared information. Meanwhile, sequences which do not have a common ancestor will not share more information than would be expected at random. Generally, most information-theoretic attempts are unlikely to provide enough information to distinguish closely related species [11], and some even deny the utility of information theory [12]. In this work, we introduce the new concept of Complete Information Set (CIS), containing all primary information of a sequence. Any two different sequences are bound to have different CIS (and vice versa), so a reasonable approach to measurement of sequence distance is the CIS comparison. Here we present a fully automated and accurate software based on such distance to compare two sequences and demonstrate that both whole genome phylogeny and 16s rRNA phylogeny, which is used for method proof-test, can be reconstructed automatically from 24 completely sequenced species.

2. Materials and Method

Genome sequences as well as corresponding 16s rRNA data were obtained directly from genebank website (http;//www.ncbi.nlm.nih.gov), including 16 Bacteria, 6 Archaea and 2 Eukarya as shown in Table I.

We begin by what we call Complete Information Set (CIS). Let $\Sigma = \{a_1, a_2, \ldots, a_m\}$ be an *alphabet* of m symbols, and suppose $S = \{S_1, S_2, \ldots, S_s\}$ is a set of sequences formed from the symbol set Σ. We denote the set of all different sequences

Table I. This table lists the completely sequenced genome and corresponding 16s rRNA data in the text, from genebank website (http://www.ncbi.nlm.nih.gov). The first column lists the abbreviations that are used for the figures in this paper, corresponding to the genome name in the second column. Column three shows the top three levels of the phylogenetic lineage (B: Bacteria; A: Archaea; E: Eukarya). The size of the complete genome and 16s rRNA of the particular species is shown in the fourth and fifth columns respectively. Obviously, there are no 16s rRNAs in the two Eukarya genomes, thus those two organisms are not include in our 16s rRNA phylogeny. The final column lists the original publication citation for each of the genomes

Abbrev.	Organism	Domain	Genome size (bp)	16s rRNA size (bp)	Reference
hinf	*Haemophilus influenzae*	B	1,830,138	1,442	Fleischmann et al., 1995
mgen	*Mycoplasma genitalium*	B	580,073	1,490	Fraser et al., 1995
synecho	*Synechocystis sp.*	B	3,573,470	1,489	Kaneko et al., 1996
mpneu	*Mycoplasma pneumoniae*	B	816,394	1,463	Himmelreich et al., 1996
ecoli	*Escherichia coli*	B	4,639,221	1,526	Blattner et al., 1997
bsub	*Bacillus subtilis*	B	4,214,814	1,497	Kunst et al., 1997
bbur	*Borrelia burgdorferi*	B	910,724	1,515	Fraser et al., 1997
aquae	*Aquifex aeolicus*	B	1,551,335	1,587	Deckert et al., 1998
mtub	*Mycobacterium tuberculosis*	B	4,411,529	1,464	Cole et al., 1998
tpal	*Treponema pallidum Nichols*	B	1,138,011	1,537	Fraser et al., 1998
ctra	*Chlamydia trachomatis*	B	1,042,519	1,548	Stephens et al., 1998
rpxx	*Rickettsia prowazekii*	B	1,111,523	1,508	Andersson et al., 1998
cpneu	*Chlamydia pneumoniae CWL029*	B	1,230,230	1,554	Kalman et al., 1999
tmar	*Thermotoga maritima*	B	1,860,725	1,562	Nelson et al., 1999
hpyl	*Helicobacter pylori*	B	1,667,867	1,765	Tomb et al., 1997
hpyl99	*Helicobacter pylori J99*	B	1,643,831	1,763	Alm et al., 1999
mjan	*Methanococcus jannaschii*	A	1,664,970	1,478	Bult et al., 1996
mthe	*Methanobacterium thermoautotrophicum*	A	1,751,377	1,501	Smith et al., 1997
aful	*Archaeoglobus fulgidus*	A	2,178,400	1,492	Klenk et al., 1997
pyro	*Pyrococcus horikoshii*	A	1,738,505	1,463	Kawarabayasi et al., 1998
aero	*Aeropyrum pernix*	A	1,551,335	1,444	Kawarabayasi et al., 1999
pabyssi	*Pyrococcus abyssi*	A	1,765,118	1,429	Heilig et al., unpublished
yeast	*Saccharomyces cerevisiae*	E	12,069,247		Goffeau et al., 1997
cegans	*Caenorhabditis elegans*	E	97,000,000		Ainscough et al., 1998

formed from Σ with length l by Θ^l, then the number $m(l)$ of all sequences of Θ^l equals m^l. For a sequence $S_k \in S$, let L_k be its length and n_{ik}^l denote the number of contiguous subsequences in S_k which match the $i-th$ sequence of Θ^l, $l \leq L_k$. It is easy to see that

$$\sum_{i=1}^{m(l)} n_{ik}^l = L_k - l + 1 \text{ for each } l \leq L_k \text{ and } k. \tag{1}$$

Letting $p_{ik}^l = n_{ik}^l/(L_k - l + 1)$, we obtain a distribution

$$U_k^l := (p_{1k}^l, p_{2k}^l, \ldots, p_{m(l)k}^l)^T \text{ where } \sum_{i=1}^{m(l)} p_{ik}^l = 1 \tag{2}$$

Let Γ^l denote the set of all distributions satisfying $\sum_{i=1}^{m(l)} p_{ik}^l = 1$, i.e.,

$$\Gamma^l := \{(p_{1k}^l, p_{2k}^l, \ldots, p_{m(l)k}^l)^T \mid \sum_{i=1}^{m(l)} p_{ik}^l = 1 \text{ and } p_{ik}^l \geq 0\}, (l = 2, 3, \ldots). \tag{3}$$

Thus, for each sequence S_k, we can get a unique set of distributions

$$(U_k^1, U_k^2, \ldots, U_k^{L_k}) \text{ where } U_k^1 \in \Gamma^1, U_k^2 \in \Gamma^2, \ldots, U_k^{L_k} \in \Gamma^{L_k} \tag{4}$$

This set contains all primary information of a sequence: in particular, $U_k^{L_k}$ uniquely determines the original sequence, so we call this set *a complete information set (CIS) of the sequence* S_k. Discrepancies among these sets mean discrepancies of all primary information among sequences, and due to the augment and heredity property of biological sequence, it is enough to efficiently discriminate different sequences as evolution distance just by comparing the beginning information subsets with window size l. Besides, for a measurement R to be a normalized 'distance' such that $0 \leq R(x, y) \leq 1$ for all the sequences x and y, it had better satisfy the following properties: Non-negative, Boundedness, Symmetry, Continuity, Monotonicity, $R(x, x) = 0$ and $R(x, y) \leq R(x, z) + R(y, z)$. A FDOD [13, 14] measurement successfully satisfying those conditions is introduced here with proofs in the references.

$$R(U_1^l, \ldots, U_s^l) = \frac{\sum_{k=1}^{s} \sum_{i=1}^{m(l)} p_{ik}^l \log(p_{ik}^l/(\sum_{ik}^{s} p_{ik}^l/s))}{s \log s} \leq 1 \tag{5}$$

where $0 \cdot \log \frac{0}{0}$ is defined as 0 as in the Kullback-Leiber entropy [15]; s denotes the number of the sequences; l denotes the window size.

For the appropriately selection of window size l, an empirical formula is

$$l \leq a + Int[\log L_{\max}/\log m] \quad (6)$$

where $l_{\max}$ denotes the largest length of sequences; $a = 2$ if $l_{\max} \leq 1000$, otherwise $a = 0$. For example, $l = 12$ for the approximate genome size 100 Mb.

Then, from the distance matrix, we can reconstruct phylogenetic trees using the neighboring-joining [16] algorithm with bootstrap [17] value calculated by selecting with replacements the random subsets of l-parameter CIS.

3. Results and Discussion

Comparative sequence analysis of 16s rRNA currently is the most widely used approach for the reconstruction of microbial phylogeny. For the purpose of method proof-test, we apply our new method as well as a popular multi-alignment based program CLUSTALW [18], to the same 16s rRNA data sets of 22 completely sequenced Bacteria and Archaea species in Table I. An exception is that the 16s rRNA sequence of *D. radiodurans* (abbreviation as drad, White et al. 1999) is included in this phylogeny instead of that of *H. pylori J99*. Topologies of these two phylogenetic trees are remarkably similar as shown in Figure 1. The two major lineages of cellular life, Archaea and Bacteria, are all monophyletic with maximal bootstrap values. In the Archaea branch, the only unmatched position of the six Euryarchaeota organisms is the relative branching order of *M. thermoautotrophicum* and *A. pernix*. As for the Bacteria part, *T. maritima* and *A. aeolicus* both appear at the root. In addition, *Purple bacteria, Spirochaetales, Chlamydiae, Mycoplasma* subbranches are all monophyletic, which has been confirmed by observed morphological features. Nevertheless, the rest Bacterial species, *Synechocystis, B. subtilis, M. tuberculosis* and*D. radiodurans* are placed in different positions, which probably due to the somewhat distrusted [19] neighbor-joining algorithm [16] and the small branch lengths in the Bacteria part. Here we show, at least, our new method is generally consistent with the commonly accepted one when applying to the same data sets.

Based on the whole genome of 24 completely sequenced Bacteria, Archaea and Eukarya species in Table I, our further efforts yield a highly robust whole genome phylogenetic tree, as shown in Figure 2. The different species are not distributed at random, but the overall topology is somewhat similar to the three-domain distribution with relatively high bootstrap values, for example, at least eighty percent of the clades are of maximum bootstrap value 100, which seems to imply that huge sequence data could remarkably increase the stability of phylogeny [20]. In the Archaea branch, to our surprise, *M. jannaschii* deviates from the Archaea and clusters with *B. burgdorferi* in the Bacteria part, but the remaining five Archaea organisms form a monophyletic branch. Interestingly, two completely sequenced Eukarya organisms in this study, *S. cerevisiae* and *C. elegans*, are closely clustered together with maximum bootstrap value and minimum distance value 0.348 (distance matrix data not shown), implying the 'late diverging', although their genome sizes are totally different (approximately 100 Mb vs. 10 Mb). Meanwhile, the Eu-

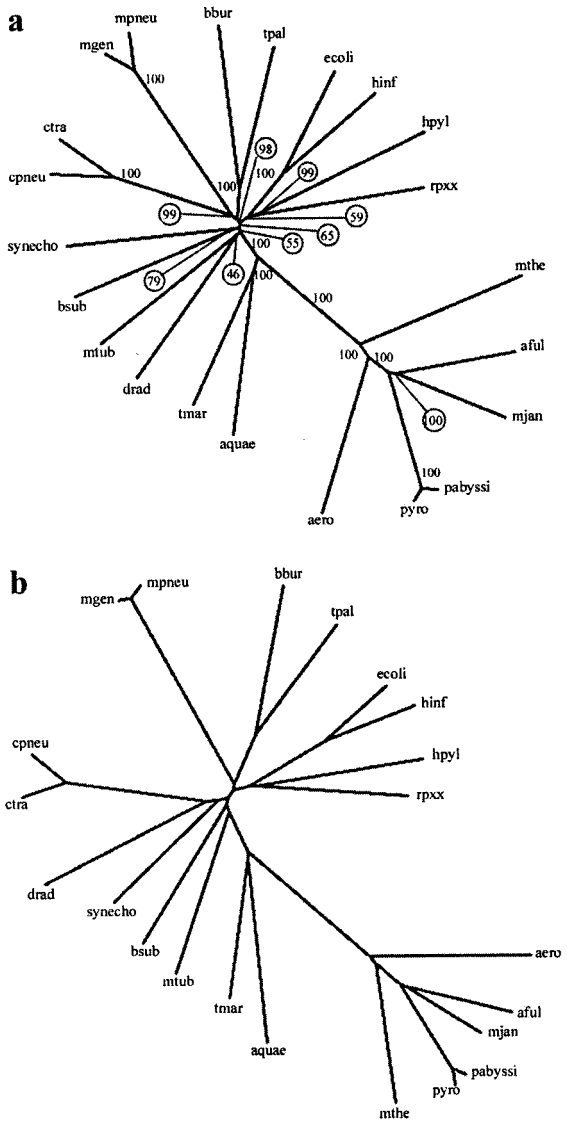

Figure 1. Phylogeny of 22 completely sequenced Bacteria and Archaea species in Table I on the basis of 16s rRNA. An exception is that the 16s rRNA sequence of *D. radiodurans R1* (abbreviation as drad, White et al. 1999) is included in this phylogeny instead of that of *H. pylori J99*. **a**. Phylogenetic tree reconstucted from our new method. The evolution distances between any pair of sequences are calculated by the FDOD function on the basis of Complete Information Set (CIS) with parameter equal to 7. The phylogeny is a neighbor-joining clustering of the resulting distance matrix, with the bootstrap values counted by selecting randomly from the -parameter CIS with replacements and recalculating the trees. The bootstrap values beside the tree clades represent the number of times (out of 100) a specific cluster was present. **b**. Phylogenetic tree built by CLUSTALW program with default parameters. Bootstrap values are not shown.

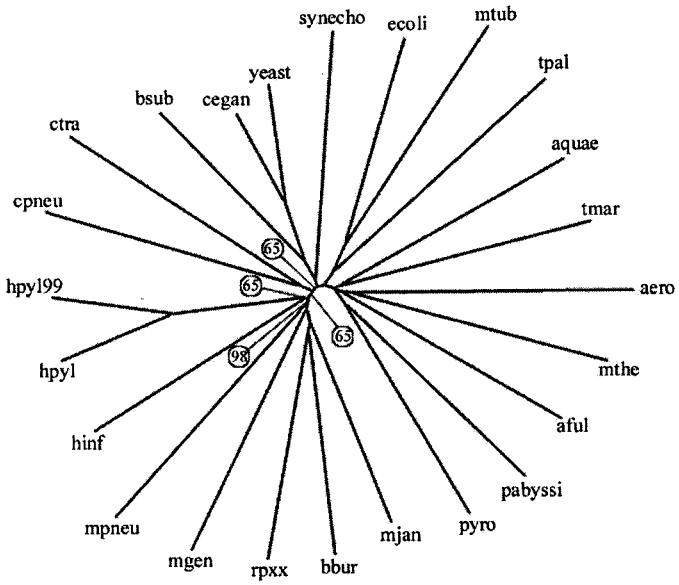

Figure 2. Phylogenetic tree based on whole genome of 24 completely sequenced Bacteria, Archaea and Eukarya species in Table I, as inferred from Complete Information Set (CIS). The whole procedures are as in the legend of Figure 1 with parameter l equal to 12. Bootstrap values at the tree clades indicate the number of times (out of 100) a specific cluster was present. Only bootstrap values less than 100 are shown.

karya branch does not lie between the Bacteria and Archaea but seems to merge into the Bacteria branch. For the Bacteria species, *A. aeolicus* and *T. maritima* group with Archaea as the earliest evolved bacterial lineage, which increase our confidence that both organisms are 'early diverging'. Moreover, those species with similar phenotype are closely clustered, such as *H. pylor* and *H. pylori J99; M. genitalium* and *M. pneumoniae; C. trachomatis* and *C. pneumoniae* monophyletic subbranches.

The purpose of this work is not to contradict or confirm existing phylogeny standards but rather to bring a brand-new algorithm and tool to the phylogeney research community. Our new method is fully automatic with no need of gene identification or human intervention, utilizing all the information embedded in whole genome including both coding and non-coding region. One predominant feature, when comparing with previous work [11], is the sensitivity of our distance estimation; even discrepancy of very similar sequences can be successfully measured. As for computational complexity, it is approximately proportional to the largest sequence size in the data sets, which is a great advantage to handle tremendous genome data. Furthermore, the changeable parameter l ensures that, given different sequence size, one can always find the best efficient measurement of the characteristic of sequences. Although, the chief criticism to our method may be that

it is mainly depended on information theory rather than a meaningful biological model such as homology, it is worth emphasizing that the alignment algorithms that biologist use most also depend on information theory with a slight difference. Our preliminary experiments have shown that our method is fruitful, and its main possible use is as an evaluator, when individual gene trees or phylogenetic trees from different algorithms do not agree well [6]. A more thorough assessment could be done with the explosion of genome data in the public efforts.

4. Acknowledgements

This work is supported by grants 39392900, 39830070 and 19890380 of Chinese National Scientific Foundation.

References

1. Koonin, E.V.: The Emerging Paradigm and Open Problems in Comparative Genomics, *Bioinformatics* **15** (1999), 265–266.
2. Woese, C.R., Kandler, O. and Wheelis, M.L.: Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. USA* **87** (1990), 4576–4579.
3. Doolittle, W.F. and Logsdon, J.M., Jr.: Archaeal Genomics: Do Archaea have a Mixed Heritage? *Curr. Biol.* **8** (1998), R209–211.
4. Woese, C.: The Universal Ancestor, *Proc. Natl. Acad. Sci. USA* **95** (1998), 6854–6859.
5. Nomura, M.: Engineering of Bacterial Ribosomes: Replacement of all Seven *Escherichia coli* rRNA Operons by a Single Plasmid-Encoded Operon, *Proc. Natl. Acad. Sci. USA* **96** (1999), 1820–1822.
6. Pennisi, E.: Is it Time to Uproot the Tree of Life? *Science* **284** (1999), 1305–1307.
7. Boore, J.L. and Brown, W.M.: Big Trees from Little Genomes: Mitochondrial Gene Order as a Phylogenetic Tool, *Curr. Opin. Genet. Dev.* **8** (1998), 668–674.
8. Snel, B., Bork, P. and Huynen, M.A.: Genome Phylogeny Based on Gene Content, *Nat. Genet.* **21** (1999), 108–110.
9. Lin, J. and Gerstein, M.: Whole-Genome Trees based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels, *Genome Res.* **10** (2000), 808–818.
10. Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. and Stanhope, M.J.: Universal Trees based on Large Combined Protein Sequence Data Sets, *Nat. Genet.* **28** (2001), 281–285.
11. Li, M. et al.: An Information-Based Sequence Distance and its Application to Whole Mitochondrial Genome Phylogeny, *Bioinformatics* **17** (2001), 149–154.
12. Hariri, A., Weber, B. and Olmsted, J.: 3rd. On the Validity of Shannon-Information Calculations for Molecular Biological Sequences, *J. Theor. Biol.* **147** (1990), 235–254.
13. Fang, W.W.: The Characterization of a Measure of Information Discrepancy, *Information* **125** (2000), 207–252.
14. Fang, W.W.: On a Global Optimization Problem in the Study of Information Discrepancy, *J. Global Optimization* **11** (1997), 387–408.
15. Kullback, S.: *Information Theory and Statistics*, Wiley, New York, 1959.
16. Saitou, N. and Nei, M.: The Neighbor-Joining Method: A new Method for Reconstructing Phylogenetic Trees, *Mol. Biol. Evol.* **4** (1987), 406–425.

17. Efron, B., Halloran, E. and Holmes, S.: Bootstrap Confidence Levels for Phylogenetic Trees, *Proc. Natl. Acad. Sci. USA* **93** (1996), 13429–13434.
18. Thompson, J.D., Higgins, D.G. and Gibson, T.J.: CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice, *Nucleic Acids Res.* **22** (1994), 4673–4680.
19. Hillis, D.M., Huelsenbeck, J.P. and Swofford, D.L.: Hobgoblin of Phylogenetics? *Nature* **369** (1994), 363–364.
20. Russo, C.A., Takezaki, N. and Nei, M.: Efficiencies of Different Genes and Different Tree-Building Methods in Recovering a Known Vertebrate Phylogeny, *Mol. Biol. Evol.* **13** (1996), 525–536.