Journal of Biological Physics **28:** 55–62, 2002.

Gene's Functional Arrangement as a Measure of the Phylogenetic Relationships of Microorganisms

JINHUA WANG[1], WEIWU FANG[2], LUNJIANG LING[1] and RUNSHENG 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*
(**Author for correspondence, e-mail: crs@sun5.ibp.ac.cn; Fax: 0086-10-64877837; Tel: 0086-10-64888546*)

Abstract. With the development of genome sequencing more whole genomes of microorganisms were completed, many methods were introduced to reconstruct the phylogenetic tree of those microorganisms with the information extracted from the whole genomes through various ways of transforming or mapping the whole genome sequences into other forms which can describe the evolutionary distance in a new way. We think it might be possible that there exists information buried in the whole genome transferred along lineage, which remains stable and is more essential than sequence conservation of individual genes or the arrangement of some genes of a selected set. We need to find one measurement that can involve as many phylogenetic features as possible that are beyond the genome sequence itself. We converted each genome sequence of the microorganisms into another linear sequence to represent the functional structure of the sequence, and we used a new information function to calculate the discrepancy of sequences and to get one distance matrix of the genomes, and built one phylogenetic tree with a neighbor joining method. The resulting tree shows that the major lineages are consistent with the result based on their 16srRNA sequences. Our method discovered one phylogenetic feature derived from the genome sequences and the encoded genes that can rebuild the phylogenetic tree correctly. The mapping of one genome sequence to its new form representing the relative positions of the functional genes provides a new way to measure the phylogenetic relationships, and with the more specific classification of gene functions the result could be more sensitive.

Key words: information theory, phylogenetic trees, sinteny, whole genome analysis

1. Introduction

Traditional methods use sequence alignment to determine the phylogenetic relationships of microorganisms. Some universally conserved nucleic acid (in particular the small subunit rRNA gene) or protein sequences were analysed based on point mutations. However, the horizontal transfer of genes from one species to another resulted in different independent phylogenetic trees with each gene. In particular, misallignment and the variance in sequence length can also can lead to phylogenetic trees with the wrong topology. So it is interesting to find other useful measurements which can describe the history of the species more accurately.

Availability of complete nucleotide sequence of microorganisms suggests the possibility of inferring useful information to rebuild the phylogenetic trees. There have been a number of studies focusing on the analysis of information extracted in different ways via different methods from the whole genome sequence. Berenet described a distance-based phylogeny constructed on the basis of gene content of 13 completely sequenced genomes of unicellular species, by counting the numbers of the orthologous genes that each genome has in common and by defining the evolutionay distance as the acquisition and loss of genes, they got a tree correlate with the standard reference of prokarytic phylogeny based on sequence similarity of 16srRNA [1]. David used the complete mitochondrial sequences and constructed the 16 mitochondrial gene orders, by analysing the distance of this genes arrangement order they infered the phylogenetic distances among the microorganisms, also their results generally agree with evolutionary relationships inffered from gene sequences [2]. Sorel selected 11 complete genomes of free-living microorganisms and reconstructed the evolutionary relationships of them by the observed presence and absence of families of protein-coding genes [3]. Their research shows that there is a strong signal within the genomes reflecting the evolutionary histories of the organisms.

These studies are in contrast to the traditional notions that a robust phylogenetic reconstruction of microorganisms is impossible due to their genomes being composed of an incomprehensible amalgam of genes with complicated histories, actually there exists in the sequence of the whole genome different kinds of information, the problem is to use an efficient method to get them out.

In this paper, we propose a new method to compare whole genomes, each genome sequence is mapping to a new sequence composed of functional code for each gene, so the comparison will concentrate on how each gene of different function is arranged, then we calculate the discrepancy for each genome pair, and the resulting tree gives a very interesting result.

2. Materials and Methods

2.1. DATA PREPARATION

All the data we used in this analysis were obtained from genebank database directly, 32 completely sequenced organisms were used in this work, including 24 Bacteria and 8 Archaea (Table 1).

2.2. GENE'S FUNCTIONAL ARRANGMENT ORDER SEQUENCE ACQUISITION

We extracted the coding sequence in each genome, recording their position coordinates, and assigned each CDS (coding sequence) one function code by analysing the original annotation information, also we did a sequence homology search

Table 1. The names and abbreviations of the 32 microorganisms with classifications

Genome	abbreviation
Bacteria	
Aquificales	
Aquifex aeolicus	Aaeo
Thermotogales	
Thermotoga maritime	Tmar
Thermus/Deinococcusgroup	
Deinococcus radiodurans	Drad
Spirochaetales	
Borrelia burgdorferi	Bbur
Treponema pallidum	Tpal
Chlamydiales	
Chlamydia pneumoniae	Cpne
Chlamydia trachomatis	Ctra
Firmicutes	
Bacillus/Clostridium	
Mycoplasma pneumoniae	Mpne
Mycoplasma genitalium	Mgen
Bacillus subtilis	
Ureaplasma urealyticum	Uure
Bacillus halodurans	Bhal
Actinobacteria	
Mycobacterium tuberculosis	Mtub
Cyanobacteria	
Symechocystis sp.	Syme
Proteobacteria	
alpha subdivision	
Rickettsia prowazekii	Rpro
beta subdivision	
Neisseria meningitides	Nmen
gamma subdivision	
Escherichia coli	Ecoli
Haemophilus influenzae	Hinf
Buchnera sp.	Buch
Pseudomonas aeruginosa	Paer
Rickettsia prowazekii	Rpro
Xylella fastidiosa	Xfas
epsilon subdivision	
Helicobacter pylory	Hpyl
Campylobacter jejuni	Cjej

Table 1. The names and abbreviations of the 32 microorganisms with classifications

Genome	abbreviation
Archaea	
Euryarchaeota	
Archaeoglobus fulgidus	Aful
Halobacterium sp.	Halo
Methanococcus jannaschii	Mjan
Methanothermobacter thermotrophicus	Mthe
Pyrococcus abyssi	Paby
Pyrococcus horikoshii	Phor
Thermoplasma acidophilum	Taci
Crenarchaeota	
Desulfurococcales	
Aeropyrum pernix	Aper

with FASTA against the COG database to modify the discrepancies in the previous annotation. If there is no apparent similarity the CDS will be considered as function unknown, also we used W to stand for the tRNA and V for rRNA. Then all the CDS fall into one functional catalog and arranging them according to their position coordinates, we get each genome one sequence carrying the information of in what order different functional genes are arranged. The function code is shown in the next pharagraph, we show one such sequence of Thermotoga maritime (Tmar) in the following.

```
>tmar
XSJJJGGTLSNPSCCTKIXJNXJJRNJJJLFTJLRXRRWXFLMGCLELXRTTKH
DCCSMHCWNKJGNXKPPPPIMXLLOHHMXXKJISJGXXNNRHJJJKJOIRN
R IXJLSEOOEOXRSCXOOGJIIJJMFEEXJRJXRJXESXWWEEXTLTTRJRRX
IILIIXPRXLFXLMPSEEOXOEOODNEWWXMFLGGLFLLXPJPSROEEEEE
KPGRGWWMJHSSGPGERRHHSWWRLDCGRXPCRPSRTIRIILMXMXCC
CGRLSMXSSRRETXLSSRIGLMMMTMRRLSLSICCEXIHORXEGGFIPGX
KOHDTJLCCCCCXRTJGKKJJJJKNWJWJXRXWEGLXHGLLSGGXSCOJOO
CRLRDNRRJJXRWXPXRXEEEEREEEEEOXSERCSGRREEIFFXGSSEXJJK
OOKSMCCNJRHSTNEJMIEEWPPPRJJWESRTMHSHSEOEHRPJJJJCMXX
XJXLNXIIFIJMRSRJLRIJRMWSTTWXLRWSSXJTLEEEGLSRHELRGRKR
LNGIWLDHFLLGWIGJKJJNJJJJJJJJJJJJJJJJJNJMIMMILRLWOJOJJGLXR
DTJMOJEKXCHNXNNNNEXXCNNNNEELRRKXRJDXXJLGTGXCCCRO
NNDNNNCOWOPFLLKISHHHKJJMXXKMXRLJSHRTMWSHXCSWEML
LGCLRSXGGHLRLRXEWFRJSLLHXSDRMGCLDNTNWNXWWJFJJWXM
RRRREKEESPGXXOLRIRNGRJOOOWKDCPEXCXKCLNCEGJXHDPEJHH
```

HPPESMRDWVVVHSRLGHHHLJVVVGFJLROWMMMNDMMSMTJHRSIL
CEPIIHLXOJJOLJJJXNSLLJLXJIGWJJJJSPOXJTJIJMXPGMEERNCCORSIF
FRWRMKJJJJWRRRROJJFLRRXXGJSSPCPPMMXIROWLCGXDXXGXXM

2.3. FUNCTION CODE

According to the COG database [4], each function code stands for one specific biological function as indicated in the following.

J	Translation, ribosomal structure and biogenesis
K	Transcription
L	DNA replication, recombination and repair Cellular processes
D	Cell division and chromosome partitioning
O	Posttranslational modification, protein turnover, chaperones
M	Cell envelope biogenesis, outer membrane
N	Cell motility and secretion
P	Inorganic ion transport and metabolism
T	Signal transduction mechanisms Metabolism
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme metabolism
I	Lipid metabolism
X	function unknown

3. Phylogenetic Analysis

Since we focus on the arrangement of the function character, not the content of the sequence, the traditional method of sequence alignment is not appropriate[5], we need a new method to evaluate the discrepancy for each genome pair with the derived functional chracter sequences, in doing this we are interested in how the functional characters in each sequence are arranged, what its upward and downward flanking short fragments for each functional character is, and how the distribution of such fragments around one specific element character in the whole genome is, with this information we can tell roughly the organization of the whole genome in the level of special function clusters of genes. We restricted the flanking fragment length to $4 \sim 7$, we denoting this as window size l, and calculate the discrepancies for each sequence pair using a function described in the following paragraph.

Let $\sum = \{a_1, a_2, \ldots, a_m\}$ 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 $\sum$. We denote the set of all different sequences formed from $\sum$ 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^l_{ik} denote the number of subsequences in S$_k$ with a step-length of 1, which match the i-th sequence of Θ^l, $l \leq$L$_k$. It is easy to see that

$$\sum_{i=1}^{m(l)} n^l_{ik} = L_k - l + 1$$

for each $l \leq$ L$_k$ and k.

Letting $p^l_{ik} = n^l_{ik}/(L_k - 1 + 1)$, we obtain a distribution

$$U^l_k := \left(p^l_{1k}, p^l_{2k}, \ldots, p^l_{m(l)k}\right)^T$$

where

$$\sum_{i=1}^{m(l)} p^l_{ik} = 1$$

Thus, for each sequence S_k, we can get a unique set of distributions $(U^1_k, U^2_k, \ldots, U^{L_k}_k)$. This set contains all primary information of a sequence: in particular, $U^{L_k}_k$ uniquely determines the original sequence, so we call this set a complete information set of the sequence S_k.

A function of measuring of information discrepancy has been introduced (abbreviated as FDOD) [7, 8]. To develop a discrepancy measure of sequences, a measure based on the FDOD function [9] is as follows:

$$R(U^l_1, \ldots, U^l_s) = \frac{\sum_{k=1}^{s} \sum_{i=1}^{m(l)} p^l_{ik} \log\left(p^l_{ik} / \left(\sum_{k=1}^{s} p^l_{ik}/s\right)\right)}{s \log s} \leq 1$$

where $0 \cdot \log\frac{0}{0}$ is defined as 0 as in the Kullback-Liebler entropy [10]; s denotes the number of the sequences; l denotes the window size. The FDOD function is characterized by a axiom set similar to Shannon's axioms: non-negativity, symmetry, continuity, identity and symmetric recursiveness. For s distributions ($s \geq 2$), this FDOD function also has the following properties: boundedness, maximum, convexity, monotonicity, and so on. Meanwhile, it's easy to see that, using this measure, the maximum discrepancy between any two sequences is less than or equal to 1, while the minimum one is equal to 0.

4. Results and Discussion

We convert the discrepancies for each 'sequence' pair to a matrix, and apply the neighbor joining method to draw the phylogenic tree [6] (Figure 1). The resulting

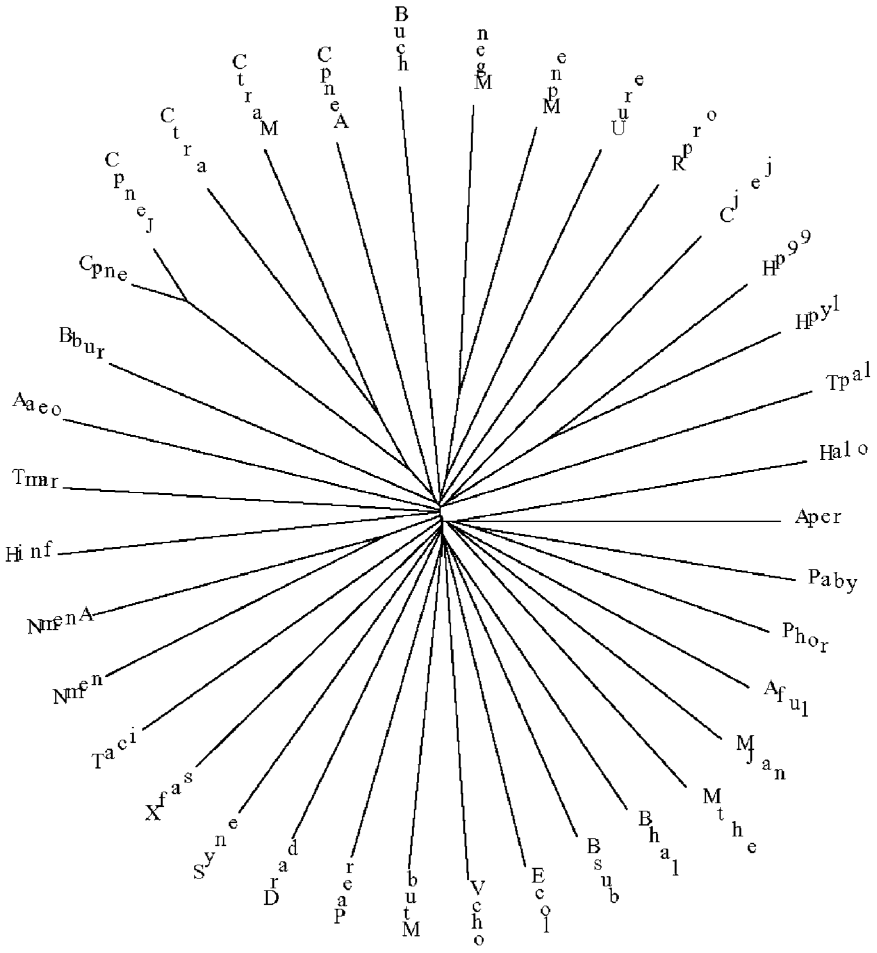

Figure 1. The phylogenic tree created with the matrix computed by the new algorithm with the window size $l = 4$.

tree has many interesting features, the two major lineages of cellular life, the Archaea and the Bacteria are separated, the seven archaeas formed one monophyletic branch, in this branch the phor and paby cluster together, the mthe and mjan come together. In the bacterial branch the chlamydiae (ctra,cpne), and the 'low G+C' Gram positive bacterial (mpne,mgen) are monophyletic. This indicates that during the process of evolution the arrangement of genes of different functional categories remains a kind of conservation, from Archae to Bacterial there is a clear boundary, and among the Archae the sub-branch correlates well with the results based on their 16srRNA sequences. In the Bacterial, except for several closely related species, the conservation between species is not very clear, this might be caused by genome segment translocation or horizontal gene transfer.

To justify whether our method exactly extracts the information of the order, we used the multi-sequence alignment to make the distance matrix of the 32 genomes, the resulting phylogeny is very poor in giving the exact topology, this confirms that the information we draw out with the FDOD function is unique, buried in each distinct genome sequence which cannot be discovered with sequence alignment. In this analysis we only use the 16 functional classes to define each gene, if we can get more specific classfication of the genes, then we might have a larger set of characters to transform the genome sequence and therefore the result tree should be capable to resolve deeper branches of the phylogeny tree.

Acknowledgements

This work is supported by Chinese National Scientific Foundation in fund-numbers of 39890070, 19890380, and 39993420.

References

1. Snel, B., Bork, P. and Huynen, M.A.: Genome Phylogeny based on Gene Content, *Nature genetics* **21** (1999), 108–110.
2. Sankoff, D., Bryant, D., Deneault, M., Lang, B.F. and Burger, G.: Early Eukaryote Evolution based on Mitochondrial Gene Order Breakpoints, *J. Comput. Biol.* **7**(2000), 521–535.
3. Fitz-Gibbon, S.T. and House, C.H.: Whole Genome-Based Phylogenetic Analysis of Free Living Microorganisms, *Nucleic Acid Res.* **27** (1999), 4218–4222.
4. Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V.: The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution, *Nucleic Acid Res.* **28** (2000), 33–36.
5. Higgins, D.G., Thompson J.D. and Gibson T.J.: Using CLUSTAL for Multiple Sequence Alignments, *Methods Enzymol.* **266** (1996), 383–402.
6. Saitou, N. and Nei, M.: The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees, *Mol. Biol. Evol.* **4** (1987), 406–425.
7. Fang, W.W., Roberts, F.S. and Ma, Z.: An Approach of Information Theory to Multiple Sequence Comparison, *DIMACS Tec. Rep.* **58** (1999).
8. Fang, W.W.: The Characterization of a Measure of Information Discrepancy, to appear in *Information Science* **125** (2000), 207–252.
9. Fang, W.W.: On a Global Optimization Problem in the Study of Information Discrepancy, *J. Global Optimization* **11** (1997), 387–408.
10. Kullback, S.: *Information Theory in Statistics*, Wiley, New York, 1959.