Reconstructing Phylogenetic Trees of Prokaryote Genomes by Randomly Sampling Oligopeptides

Osamu Maruyama¹, Akiko Matsuda², and Satoru Kuhara³

¹ Faculty of Mathematics, Kyushu University, Hakozaki 6-10-1,
Higashi-ku, Fukuoka 812-8581, Japan
`om@math.kyushu-u.ac.jp`

² Graduate School of Systems Life Sciences, Kyushu University,
Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan
`3s103012s@sls.kyushu-u.ac.jp`

³ Graduate School of Genetic Resources Technology, Kyushu University,
Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan
`kuhara@grt.kyushu-u.ac.jp`

Abstract. In this paper, we propose a method for reconstructing phylogenetic trees of a given set of prokaryote organisms by randomly sampling relatively small oligopeptides of a fixed length from their complete proteomes. For each of the organisms, a vector of frequencies of those sampled oligopeptides is generated and used as a building block in reconstructing phylogenetic trees. By this procedure, phylogenetic trees are generated independently, and a consensus tree of the resulting trees is obtained. We have applied our method to a set of 109 organisms, including 16 Archaea, 87 Bacteria, and 6 Eukarya, using less 10 % of all the 3,200,000 oligopeptides of length 5. Our consensus tree agrees with the tree of Bergey's Manual in most of the basic taxa. In addition, they have almost the same quality as the trees of the same organisms reconstructed using *all* the 20^K oligopeptides of length $K = 5$ and 6 given by Qi *et al.* Thus we can conclude that, the frequencies of a relatively small number of oligopeptides of length 5, even if those oligopeptides are determined in a random method, has phylogenetic information almost equivalent to the frequencies of *all* the oligopeptides of length 5 or 6.

1 Introduction

One of the interesting problems in computational biology is that of reconstructing phylogenetic trees of microbial organisms from their genomes and/or proteomes. Furthermore, more available genomes and proteomes are making this problem more challenging. A merit of studying this problem is that we can gain an insight into evolution. In most of the previous works on this problem, a distance-matrix between organisms is calculated and then a phylogenetic tree is generated from the matrix using a distance-based phylogenetic tree generator like UPGMA[14], Neighbor-Joining (NJ)[12], BIONJ[6], and NEIGHBORNET[1]. Almost all the definitions of “distance” between two organisms are formulated

based on *sequence alignment*, including gene content [13, 15], presence-absence of families of genes [4, 8], presence-absence of genes in clusters of orthologous genes [16], supertree[2], and all-against-all genome-wide pairwise comparison [7].

Recently, Qi *et al.* [11] proposed a method for this problem, in which a vector of attributes based on frequencies of all the oligopeptides of a fixed length $K = 5$ or 6 in the complete proteome is used, in order to resolve the major branchings with Archaea and Bacteria more accurately than the results of the previous works and to avoid the inherent faults in sequence alignment. The distance between two arbitrary organisms is determined based on the angle between the corresponding vectors of the organisms. The matrix of those distances is calculated, and is taken as input to a phylogenetic tree generator, Neighbor-Joining [12]. They have applied their method to a set of 109 organisms, including 16 Archaea, 87 Bacteria, and 6 Eukarya, using a supercomputer, and reported that their phylogenetic trees constructed from the attribute vectors agree with the biologists' "tree of life" [5] in a majority of basic branchings, and especially, in all lower taxa.

Some natural questions of the work done by Qi *et al.* [11] are raised. *All* the frequencies of 20^K oligopeptides were used in their method. This requirement imposes a huge computation task, which is almost impossible in a current personal computer. For example, if $K = 6$, for each of the 109 organisms, the frequencies of 64,000,000 oligopeptides in the proteome should be calculated. Actually, a supercomputer was used to carry out their work. Our first question is whether the dimension of those vectors of oligopeptide frequencies can be reduced. Even if this is possible, how oligopeptides effective in reconstructing phylogenetic trees are selected? The second question is whether there are other sequence attributes of oligopeptides that are effective in reconstructing phylogenetic trees. As various formulations of species-species distance based on sequence alignment were discussed, various sequence attributes based on oligopeptide frequency should be considered and compared in order to identify better ones.

In this paper, to gain some insight into these questions, we have proposed a novel method for reconstructing phylogenetic trees using *random sampling* of oligopeptides. In our method, a relatively small number N of oligopeptides $r_1, r_2, \dots, r_N$ of a fixed length $k - 1$ are randomly sampled from the complete proteomes of all the given organisms. Let $\Sigma = \{a_1, \dots, a_{20}\}$ be an alphabet of symbols representing the 20 amino acids. For all the length- k oligopeptides $r_i a_j$ with $1 \leq i \leq N$ and $1 \leq j \leq 20$, the frequencies of the oligopeptides in the proteomes of an organism are calculated. The frequencies constitute an attribute vector of the organism. An additional advantage of this strategy is that we can afford to set k to be larger, e.g., $k = 12$, even if the computation is carried out on a personal computer. The attributes that we have applied in this work are the attribute of oligopeptide frequency used in [11] and another oligopeptide frequency attribute based on a k -th order Markov chain over a proteome. It should be noted here that the framework of our method is general such that various kinds of attributes based on oligopeptides are applicable. This random sampling procedure for making a phylogenetic tree is repeated several

times independently. A consensus tree of the resulting trees is calculated using CONSENSE [10], whose software is provided as a part of PHYLIP [3].

In our preliminary computational experiments, we have applied our method to the set of 109 organisms, including 16 Archaea, 87 Bacteria, and 6 Eukarya, that is the same set of species as [11]. Our resulting trees agree with the biologists' "tree of life" [5] in most of the basic taxa. In addition, they have almost the same quality as the trees of the same organisms reconstructed using all the 20^K oligopeptides of length $K = 5$ or 6, given by Qi *et al.* [11]. Thus we can conclude that, with respect to the attribute of oligopeptide frequency in [11], a relatively small number of oligopeptides of length 5 that are derived from randomly sampled oligopeptides of length 4 has information almost equivalent in reconstructing phylogenetic trees to *all* the oligopeptides of length 5 or 6.

2 Materials and Methods

In this section, the materials and methods we use here are described.

The set of organisms we use in this work is the same as the set of the organisms used in [11], which is 109 organisms, including 16 Archaea, 87 Bacteria, and 6 Eukarya, to compare our phylogenetic trees with their ones. The set of the protein sequences of an organisms is downloaded from "Genome Assembly/Annotation Projects" at the NCBI ftp site [17].

An overview of our method is given in Fig.1. Let n be the number of organisms. For $1 \leq i \leq n$, we denote the set of all the protein sequences of the i -th organism by S_i . Let $\mathcal{S} = \{S_1, \dots, S_n\}$. For a positive integer k , the length of oligopeptides which are randomly samples from their complete proteomes is set to be $k - 1$. Let N be the number of different oligopeptides which are randomly sampled. An *attribute function* α is a function $\alpha : \Sigma^\ell \times 2^{\mathcal{S}^*} \rightarrow \mathcal{R}$, where $\Sigma = \{a_1, \dots, a_{20}\}$ is an alphabet of symbols representing the 20 amino acids. We call the first argument of α , which is an oligopeptide of length ℓ , a *seed* of α . The second argument of α is supposed to be the set of all the protein sequences of an organism. The details of each step in Fig. 1 are described as follows.

Input: positive integers k , N , and R .

Output: a phylogenetic tree.

Procedure:

1. Repeat the followings R times independently.
 - (a) A set of N different oligopeptides of length $k - 1$ is randomly chosen from $\mathcal{S}$. The set is denoted by $P = \{p_1, \dots, p_N\}$.
 - (b) For $i = 1, \dots, n$, a vector $u_i = (\alpha(p_1 a_1, S_i) \cdots \alpha(p_1 a_{20}, S_i) \cdots \alpha(p_N a_1, S_i) \cdots \alpha(p_N a_{20}, S_i))$ is calculated.
 - (c) A distance matrix M among the n organisms is calculated from $u_1, \dots, u_n$.
 - (d) A phylogenetic trees is constructed from M .
2. A consensus tree of all the resulting trees is generated by CONSENSE[10].

Fig. 1. An overview of our method

The determination of seeds of an attribute function α is not completely random, but the prefixes of length $k - 1$ of seeds are chosen randomly from the proteomes of the given organisms. For each randomly chosen prefix r , the strings ra of length k are used as seeds, where a is a symbol representing an amino acid. In this sense, the set of the resulting seeds is *biased*, which is more effective in reconstructing phylogenetic trees than a set of completely randomly sampled oligopeptides of length k (data not shown). The procedure of the determination of seeds is as follows. First, the following things are, in sequence, chosen uniformly at randomly, respectively: (i) a species i , (ii) a protein sequence s in S_i , (iii) a position $p \in \{1, \dots, |s| - k + 2\}$. Then, the substring of s of length $k - 1$ starting at p is added into a set P . This procedure is repeated until $|P| = N$. Next, from each sampled oligopeptide r , 20 different oligopeptides ra are generated, which are used as seeds of an attribute function. Thus, in this case, the parameter ℓ of α turns to be k . The size of the resulting set of seeds is $20N$, denoted by N' . Thus, the vector of α for the i -th organism is $(\alpha(p_1 a_1, S_i) \dots \alpha(p_N a_{20}, S_i))$, whose dimension is N' .

Here, the attribute function of oligopeptide frequency used in [11], which we denote by α_1 , is described briefly. Let $x = x_1 \dots x_\ell$ be a string of length ℓ over Σ and $S \subseteq \Sigma^*$. We denote the number of occurrences of x in $s \in S$ by $c(x, s)$.

$$p(x, S) = \sum_{s \in S} c(x, s) \bigg/ \sum_{s \in S} (|s| - \ell + 1)$$

$$p^0(x, S) = p(x_1 \dots x_{\ell-1}, S) p(x_2 \dots x_\ell, S) / p(x_2 \dots x_{\ell-1}, S)$$

$$\alpha_1(x, S) = \begin{cases} \frac{p(x, S) - p^0(x, S)}{p^0(x, S)} & \text{if } p^0(x, S) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

A feature of this attribute function is that the expectation $p^0(x, S)$ of $p(x, S)$ is subtracted from $p(x, S)$. This expectation is derived from a Markov model. Note that the attribute function is a natural generalization of “dinucleotide frequency” [9]. Another attribute function we consider in this work, denoted by α_2 , is described as follows. This is designed based on a partial $(\ell - 1)$ -th order Markov chain.

$$\alpha_2(x, S) = \sum_{s \in S} c(x, s) \bigg/ \sum_{s \in S} \sum_{a' \in \Sigma} c(x_1 \dots x_{\ell-1} a', s),$$

Thus, $\alpha_2(x, S)$ represents the probability that the amino acid x_ℓ occurs just after $x_1 \dots x_{\ell-1}$ occurs in the sequences in S . A feature of this attribute is that the attribute is much easier to calculate rather than that of [11].

The distance between two arbitrary organisms is determined by the angle between the corresponding two attribute vectors. Let X and Y be attribute vectors corresponding to two organisms. The distance between X and Y is defined as

$$D(X, Y) = \frac{1 - C(X, Y)}{2} \quad \text{where } C(X, Y) = \sum_{k=1}^{N'} X_k Y_k \bigg/ \sqrt{\sum_{k=1}^{N'} X_k^2 \cdot \sum_{k=1}^{N'} Y_k^2}.$$

Note that the lengths of the vectors are not considered in this definition.

A phylogenetic tree is generated from n attribute vectors of the n organisms using Neighbor-Joining [12], whose software is available as a part of a software package PHYLIP [3]. Note that an output tree of the method is unrooted. In the final step of our procedure, a consensus tree of the resulting trees calculated using R different sets of N oligopeptides of length k is generated by the CONSENSE[10] algorithm, whose software is also available as a part of PHYLIP[3].

The difference between our procedure with the attribute function α_1 and that of [11] is that in our procedure, a relative small N' oligopeptides are used as seeds of α_1 , on the other hand, all the 20^K oligopeptides of length K are used in [11]. The remaining parts, including the definition of the distance between two organisms, the phylogenetic tree generator, and the consensus tree generator are the same as those of [11].

We define a measure to evaluate how a reconstructed phylogenetic tree agrees with the biologists' "tree of life" shown in [5]. Let T be a unrooted phylogenetic tree with the trifurcation of the three main domains, Archaea, Bacteria, and Eukarya. Thus this measure is defined for a tree with the trifurcation. The rooted tree derived from T by making the trifurcation node rooted is denoted by T' . Let E be the set of the branches in T' . For a branch $e \in E$, by L_e we denote the set of all the organisms assigned to the leaves that are descendants of the child node of e . For a set C of organisms, we can determine the unique edge e_C such that $C \subseteq L_{e_C}$ and $|L_{e_C}|$ is minimized. We call e_C the *separator* for C in T . The *unity* of C in T is defined as $u(T, C) = |C|/|L_{e_C}|$. The following taxa are considered to be C : *Archaea*(A), *Bacteria*(B), *Eukaryotes*(E), *Crenarchaeota*(A1), *Euryarchaeota*(A2), *Cyanobacteria*(B10), *Proteobacteria*(B12), *Alphaproteobacteria*(B12.1), *Betaproteobacteria*(B12.2), *Gammaproteobacteria*(B12.3), *Epsilonproteobacteria*(B12.5), *Firmicutes*(B13), *Actinobacteria*(B14), *Chlamydiae*(B16), and *Spirochaetes*(B17). Each taxon can be denoted by the code for it in Bergey's Manual if exists, which is given in parentheses. Since Eukaryotes are not included in Bergey's Manual, 'E' is assigned to it temporarily.

3 Results

Tab. 1 shows how the consensus phylogenetic trees constructed by the method in Fig. 1 with $k = 5$ and the trees in Fig. 1 and Fig. 2 in [11] agree with the biologists' "tree of life" shown in [5]. Recall that the tree in Fig. 2 and Fig. 1 in [11] are generated using *all* the oligopeptides of length 5 and 6, respectively, and that the trees agree with the biologists' "tree of life" [5] in a majority of basic branchings, and especially, in all lower taxa. As we can see in Tab. 1, the consensus trees constructed by the random sampling method with $k = 5$ have achieved almost the same quality as the tree in Fig. 2 and Fig. 1 in [11]. Surprisingly, the unities in the consensus tree with $N' = 280,000$ are all better or equal to those in Fig. 2 in [11]. It should be noted here that the number of seeds of α_1 used in constructing these trees are $N' = 280,000$ and $3,200,000$, respectively, so the former is only 8.75% of the latter. For Fig. 1 in [11], the

Table 1. A summary of the consensus trees constructed with the parameter sets (k, N', R) with $k = 5$, $R = 10$, and $N' = (16 + 4j) \times 10^4$ ($0 \leq j \leq 7$). Each column labeled with number N'' , except the last two columns, corresponds to the consensus tree constructed with the parameter $N' = N'' \times 10^3$. Note that all those trees have the trifurcations (data not shown). The column labeled with “Qi2” is given for the consensus tree of Fig. 2 in [11], which is constructed by bootstrap with all the oligopeptides of length 5. The column labeled with “Qi1” is given for the tree of Fig. 1 in [11], which is constructed by a deterministic way with all the oligopeptides of length 6. The columns labeled with “u” shows the unities for the taxa shown in the first column, and the columns labeled with “f” shows the frequencies of appearances of the separators for the taxa, which are scaled between 0 and 1

taxon	160		200		240		280		320		360		400		440		Qi2		Qi1	
	u _i	f	u _i	f	u _i	f	u _i	f	u _i	f	u _i	f	u _i	f	u _i	f	u _i	f	u _i	f
E	1.00, 0.5	1.00, 0.6	1.00, 0.7	1.00, 0.8	1.00, 0.7	1.00, 0.8	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 0.8	1.00, 0.95	~	1.00	1.00	1.00	1.00	1.00
A	1.00, 0.9	1.00, 0.8	1.00, 0.8	1.00, 0.8	1.00, 0.8	1.00, 0.8	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 1.0	1.00, 1.0	1.00, 0.95	~	1.00	1.00	1.00	1.00	1.00
A1	1.00, 0.9	1.00, 0.7	1.00, 0.7	1.00, 0.6	1.00, 0.6	1.00, 0.8	1.00, 0.7	1.00, 1.0	1.00, 1.0	1.00, 0.7	1.00, 0.9	1.00, 0.9	1.00, 0.9	1.00, 0.95	~	1.00	1.00	1.00	1.00	1.00
A2	0.75, 0.9	0.75, 0.8	0.75, 0.8	0.75, 0.8	0.75, 0.8	0.75, 0.8	0.75, 0.9	0.75, 0.9	0.75, 0.9	0.75, 0.9	0.75, 1.0	0.75, 1.0	0.75, 1.0	0.75, 0.95	~	1.00	1.00	1.00	1.00	1.00
B	1.00, 0.8	1.00, 0.7	1.00, 0.8	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 0.9	1.00, 0.9	1.00, 0.95	~	1.00	1.00	1.00	1.00	1.00
B10	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 0.95	~	1.00	1.00	1.00	1.00	1.00
B12	0.50, 0.2	0.59, 0.1	0.43, 0.8	0.49, 0.1	0.43, 1.0	1.00, 0.1	0.50, 0.4	0.60, 0.2	0.49,						~	0.85	1.00	1.00	1.00	1.00
B12.1	0.26, 0.2	0.26, 0.3	0.10, 0.8	1.00, 0.5	0.10, 1.0	0.26, 0.5	0.26, 0.7	1.00, 0.5	1.00,						~	0.85	0.26	1.00	1.00	1.00
B12.2	0.60, 0.5	1.00, 0.6	1.00, 0.7	0.60, 0.4	1.00, 0.5	0.60, 0.4	0.60, 0.5	0.60, 0.7	0.13, 0.85						~	0.90	0.13	1.00	1.00	1.00
B12.3	0.88, 0.5	0.65, 0.3	0.25, 0.8	0.88, 0.6	0.25, 1.0	0.88, 0.4	0.65, 0.7	0.88, 0.7	0.88, 0.85						~	0.90	0.64	1.00	1.00	1.00
B12.5	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 0.95						~	1.00	1.00	1.00	1.00	1.00
B13	0.68, 0.1	0.31, 0.7	0.68, 0.2	0.31, 1.0	0.60, 0.3	0.36, 0.1	0.68, 0.2	0.31, 0.9	0.31, 0.95						~	1.00	0.31	1.00	1.00	1.00
B14	1.00, 0.7	1.00, 0.9	1.00, 1.0	1.00, 0.9	1.00, 1.0	1.00, 0.9	1.00, 0.9	1.00, 0.8	1.00, 0.95						~	1.00	1.00	1.00	1.00	1.00
B16	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 1.0	1.00, 0.95						~	1.00	1.00	1.00	1.00	1.00
B17	0.03, 0.8	0.03, 0.7	0.03, 0.8	0.04, 0.2	0.03, 1.0	0.04, 0.1	0.03, 1.0	0.03, 0.9	0.03, 0.95						~	1.00	0.03	1.00	1.00	1.00

consensus tree with $N' = 360,000$, given in Fig. 2, is quite a similar to the tree with respect to the unity scores. Our tree wins the unities of B12.2, B12.3, and B13 and loses only the unity of A2. A drawback of the consensus tree in Fig. 2 is that *E. coli* (B12.C3.O13.F1.G1) and *S. flexneri* 2a (B12.C3.O13.F1.G37) are mixed, although the other genera come together as they should.

These results imply that the set of all the oligopeptides of a fixed length are redundant with respect to reconstructing phylogenetic trees through the attribute function α_1 . In addition, we can say that if a certain amount of oligopeptides are used to be seeds of α_1 , it is not important how to select oligopeptides effective in reconstructing phylogenetic trees.

4 Discussion

For the attribute function α_1 , we have also constructed phylogenetic trees with the different values of k , the length of oligopeptides sampled randomly. For $k = 6, \dots, 15$, we have carried out our method with $N' = 100,000$, 200,000 and 400,000, and $R = 6$. Interestingly, all the consensus trees are inferior to those of $k = 5$ so some amount of oligopeptides of length 5 would have more phylogenetically valuable information on microbial organisms.

We have also applied our method with another attribute function α_2 , which is an attribute based on a k -th order Markov chain over a proteome. We have

References

1. D. Bryant and V. Moulton. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.*, 21:255–265, 2004.
2. V. Daubin, M. Gouy, and G. Perrière. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, 12:1080–1090, 2002.
3. J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, 1993. <http://evolution.genetics.washington.edu/phylip.html>.
4. S.T. Fitz-Gibbon and C.H. House. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.*, 27:4218–4222, 1999.
5. G.M. Garrity, J.A. Bell, and T.G. Lilburn. *Taxonomic Outline of the Prokaryotes Release 5.0 Bergey's Manual of Systematic Bacteriology, 2nd edition*. Springer-Verlag, New York, May 2004.
6. O. Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14:685–695, 1997.
7. S.R. Henz, D.H. Huson, A.F. Auch, K. Nieselt-Struwe, and S.C. Schuster. Whole-genome prokaryotic phylogeny. *Bioinformatics*, 2004.
8. C.H. House and S.T. Fitz-Gibbon. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J. Mol. Evol.*, 54:539–547, 2002.
9. S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, 11:283–290, 1995.
10. T. Margush and F.R. McMorris. Consensus n-trees. *Bulletin of Mathematical Biology*, 43:239–244, 1981.
11. J. Qi, B. Wang, and B.-I. Hao. Whole proteome prokaryote phylogeny without sequence alignment: a k -string composition approach. *J. Mol. Evol.*, 58:1–11, 2004.
12. N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
13. B. Snel, P. Bork, and M.A. Huynen. Genome phylogeny based on gene content. *Nat. Genet.*, 21:108–110, 1999.
14. R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.
15. F. Tekaiia, A. Lazcano, and B. Dujon. The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, 9:550–557, 1999.
16. Y.I. Wolf, I.B. Rogozin, N.V. Grishin, R.L. Tatusov, and E.V. Koonin. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, 1, 2001.
17. <ftp://ftp.ncbi.nih.gov/genomes/>.