Commun. Korean Math. Soc. **0** (0), No. 0, pp. 1–0
https://doi.org/10.4134/CKMS.c180139
pISSN: 1225-1763 / eISSN: 2234-3024

SPECTRAL METHOD FOR RECONSTRUCTING PHYLOGENETIC TREE

Seong-Hun Paeng and Chunjae Park

Abstract. A new simple method is proposed for reconstructing phylogenetic trees, which we call the *spectral method*. The most common distance based method is the neighbor-joining method which is based on the minimum evolution principle. The spectral method shows similar performance to the neighbor-joining method for simulated data generated by `seq-gen`. For real data, the spectral method shows much better performance than the neighbor-joining method. Hence it can be a complementary method for reconstructing phylogenetic trees.

1. Introduction

Phylogenetics is the science which studies evolutionary relationship between species. In order to study the relationship, phylogenetic trees are constructed which link the species. Many methods for reconstructing phylogenetic tree have been proposed. The most common method based on genetic distances is the neighbor-joining (NJ) method developed by Saitou and Nei [4], which is based on the principle of minimum evolution. The standard algorithm based on this principle is to examine all possible topologies and to choose one which shows the smallest amount of total evolutionary changes. The neighbor-joining method approximately produces the minimum evolution tree. The neighbor-joining method is a recursive procedure taking the best candidate for a cherry and then peeling off the chosen one. Recall that a pair of leaves in a tree which are adjacent to the same internal node is called a cherry.

We will propose a new simple method for reconstructing phylogenetic trees which are trivalent. Our method is not based on the minimum evolution principle. Our purpose is to construct a tree whose tree metric approximates to distance data of DNA sequences. The tree metric is the distance function such that

Received April 5, 2018; Accepted May 31, 2018.
2010 *Mathematics Subject Classification.* 92D15.
Key words and phrases. spectrum and phylogenetic tree.
The first author was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A1B03931459). Also this paper was written as part of Konkuk University's research support program for its faculty on sabbatical leave in 2014.

the distance between two leaves (terminal nodes) is the sum of lengths of edges in the unique paths connecting two leaves. We will define a set $\Lambda_{XY} \subset \mathbb{R}$ called the *spectrum* for a pair of leaves (X, Y). Then we can show that the cardinality $|\Lambda_{XY}| = 1$ for a cherry (X, Y) if the distances are induced from a tree metric. So our basic idea is to take a pair (X, Y) as a cherry if $\max \Lambda_{XY} - \min \Lambda_{XY}$ is small. Hence our algorithm is based on finding minimal spectral range rather than minimal evolution.

In Section 2, we will define the spectrum and use the spectrum to reconstruct a tree whose metric is a tree metric. Actually, the neighbor-joining method also reconstructs a tree completely when the metric is a tree metric. We will show how to use the spectrum for reconstructing a tree. In general, the pairwise distances on a phylogenetic tree are not induced from a tree metric. In Section 3, we will propose the spectral method for a non tree metric. In Section 4, we will give experimental results and compare the performance of the spectral method with the performance of the neighbor-joining method.

2. Tree reconstruction for tree metric

Phylogenetic tree reconstruction is closely related to differential geometric problems, the boundary rigidity problem [2] and Plateau's problem. The boundary rigidity problem is to determine the metric of a compact Riemannian manifold with boundary, up to isometry, by knowing the boundary distance function between boundary points. Plateau's problem is concerned about the existence of a minimal surface with a given boundary.

Let T be a tree. We can consider the set of leaves as the boundary ∂T of T. Phylogenetic tree reconstruction is to construct a tree whose metric restricted to ∂T is distance data of DNA sequences in ∂T. Hence the tree reconstruction has similarity to the boundary rigidity problem. On the other hand, the minimum evolution tree can be considered as a minimal surface with given boundary data set. So the principle of minimum evolution is similar to Plateau's problem.

In this section, we define the spectrum and use it to reconstruct a tree for a tree metric. If distances between leaves are given by a tree metric, the boundary rigidity problem is completely solved, i.e., the metric of T is completely determined. Thus the tree is boundary rigid for a tree metric.

Let X, Y, Z be nodes in T. We denote by l_{XY} the path from X to Y. Let the paths l_{XY} and l_{XZ} be split at a node O in the interior of the path l_{XY} (Fig. 1). We denote the distance between the nodes A and B by $d(A, B)$. Let $x = d(O, X),\ y = d(O, Y),\ z = d(O, Z)$. Then it is well known that for a tree metric,

$$d(X, Y) = x + y,\ d(Y, Z) = y + z,\ d(Z, X) = x + z, \tag{1}$$

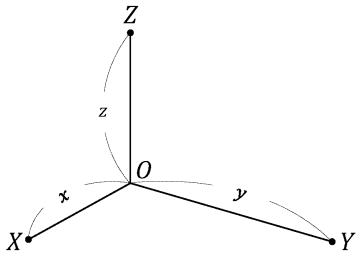


FIGURE 1. x is the spectrum $\lambda_{XY}(Z)$ of X, Y for Z

so

$$
\begin{aligned}
x &= \frac{1}{2}(d(X,Y) - d(Y,Z) + d(Z,X)) \\
y &= \frac{1}{2}(d(X,Y) + d(Z,X) - d(Y,Z)) \\
z &= \frac{1}{2}(d(Y,Z) + d(X,Y) - d(Z,X)).
\end{aligned} \tag{2}
$$

For the fixed nodes X, Y, we consider x as a function of Z. From the above observation, we define the spectrum as follows:

Definition 1. For an ordered pair of nodes (X, Y), the function $\lambda_{XY} : \partial T \to \mathbb{R}$ is defined as follows:

$$\lambda_{XY}(Z) = \frac{1}{2}(d(X,Y) - d(Y,Z) + d(Z,X)).$$

Then we define the spectrum for (X, Y) as the set

$$\Lambda_{XY} = \{\lambda_{XY}(Z) \mid Z \in \partial T\}.$$

Note that $\lambda_{XY} \neq \lambda_{YX}$ and $\lambda_{XY} \subset [0, d(X,Y)]$. The spectrum $\Lambda_{XY} = \{\lambda_{XY}(Z) \mid Z \in \partial T\}$ is the set of distances from X to the nodes in the path l_{XY}. Then we obtain the following observation immediately.

Proposition 2.1. *For a tree metric, if (X, Y) is a cherry, then the cardinality of Λ_{XY} satisfies that $|\Lambda_{XY}| = 1$.*

It will be used as the basic discriminant for a cherry in the next section.

Let $\Lambda_{XY} = \{\lambda_1, \ldots, \lambda_m\}$, where $\lambda_i < \lambda_j$ for $i < j$. Let V_j be the node in l_{XY} such that $d(X, V_j) = \lambda_j$. Let $S_j(X, Y)$ be the subset of ∂T defined as follows:

$$S_j(X,Y) = \{P \mid \lambda_{XY}(P) = \lambda_j\} \subset \partial T.$$

(See Fig. 2.) Then $d(V_j, P) = d(X, P) - \lambda_j$ for $P \in S_j(X, Y)$.

Take a terminal node Y_k in $S_k(X,Y)$ and follow the same process as above, i.e., find the spectrum for (V_k, Y_k)

$$\Lambda_{V_kY_k} = \{\lambda_{k1}, \dots, \lambda_{km}\}.$$

Then we can find the internal node V_{ki} corresponding to λ_{ki} in the branch connecting V_k and Y_k. Also we denote the set of terminal nodes connected to V_{ki} by $S_i(V_k, Y_k)$, i.e.,

$$S_i(V_k, Y_k) = \{P \mid \lambda_{V_kY_k}(P) = \lambda_{ki}\} \subset \partial T.$$

Take $Y_{ki} \in S_i(V_k, Y_k)$ and follow the same procedure with V_{ki}, Y_{ki} instead of V_k, Y_j. (See Fig. 2.)

Inductively, we can find every nodes and the distances between nodes.

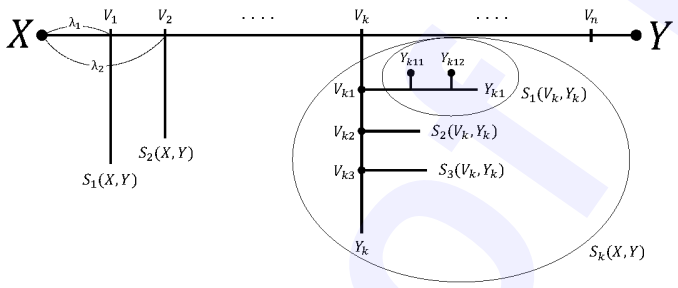


FIGURE 2. Spectrum of X, Y and reconstruction of tree

3. Spectral method

Based on Proposition 2.1, we propose the spectral method to reconstruct the phylogenetic tree. Similarly as the neighbor-joining method, the spectral method is a recursive procedure picking a cherry and then peeling off the chosen one. Hence we only need to explain the cherry picking algorithm.

If (X, Y) is a cherry, then there is exactly one internal node in a path l_{XY}, which implies that the spectrum Λ_{XY} is one point set with respect to tree metric. Hence the first part of our algorithm is to take a pair (X, Y) such that λ_{XY} is almost constant (Step 1, 2).

In Fig. 3, Λ_{XY} has two values whose difference is $\epsilon > 0$ if the metric is a tree metric. If ϵ is very small, then λ_{XY} could seem to be almost constant and (X, Y) could be taken as a cherry although (X, Y) is not a cherry, since the distances between leaves are not exactly induced from a tree metric. The

second part of our algorithm is an error correcting process (Step 3, 4). For $X, Y \in \partial T$, we define a function F as follows:

$$F(X,Y) = \sum_{W \in \partial T \setminus \{X,Y\}} \lambda_{XY}(W). \tag{3}$$

Then we obtain the following proposition immediately.

Proposition 3.1. *Let T be a tree with a tree metric such that $|\partial T| = n$. If (A, B) is a cherry, then $F(A, B) = (n-2)d$ when $\Lambda_{AB} = \{d\}$. Furthermore, if the second smallest spectrum of Λ_{AC} is $d + \epsilon$ for $C \in \partial T \setminus \{A, B\}$, then*

$$F(A,B) = (n-2)d < (n-2)d + (n-3)\epsilon \leq F(A,C). \tag{4}$$

Note that the smallest spectrum of Λ_{AC} is d, which is the distance from A to the closest internal node. Proposition 3.1 follows from the observation that $\lambda_{AC}(D) \geq d + \epsilon$ when $D \neq B$. Proposition 3.1 is our second discriminant for a cherry.

Step 3 is the process to verify if (X_0, Y_0) chosen in Step 2 is really a cherry by using F. In Fig. 3, (X, Y) is not a cherry and the spectral range $\max \Lambda_{XY} - \min \Lambda_{XY} = \epsilon$ for a tree metric. However, if $\epsilon > 0$ is very small, then λ_{XY} could seem to be almost constant. So the first part of algorithm may take (X, Y) as a cherry. In order to correct this error, we apply Proposition 3.1. First, consider the case that X is contained in a cherry (Fig. 3(A)). We have $F(X, Z) = (n-2)d$ and $F(X, Y) = (n-2)d + (n-3)\epsilon$ for a tree metric. Hence if $F(X, Y) > F(X, Z)$, then (X, Y) might not be a cherry in a high probability from (4). Note that $F(X, Y) - F(X, Z) = (n-3)\epsilon$ is much larger than ϵ, so F better discriminates whether (X, Y) is a cherry or not. So we do not pick (X, Y) as a cherry if $F(X, Z)$ is the minimum among $\{F(X, W) \mid W \in \partial T \setminus \{X\}\}$.

Second, we consider the case that X is not in a cherry as we see in Fig. 3(B). By the same reason as in Fig. 3(A), we could take (X, Y) as a cherry if $\epsilon > 0$ is very small. In such a case, we have $F(X, Y) = F(X, Z)$ both of which are the minimum for a fixed X. So only with the comparisons with $F(X, \cdot)$'s, we cannot exclude (X, Y) for a cherry even if (Y, Z) is a cherry. Hence, we also consider $F(Y, \cdot)$ together with $F(X, \cdot)$. Then we can exclude (Y, X).

Consequently, the cherry picking algorithm in the spectral method is as follows. We denote the average of the function $f : V \to \mathbb{R}$ by $\mathrm{mean} f$. Also we denote by $\mathrm{argmin}_{x \in V} f(x)$ a minimizer $x_0 \in V$ satisfying that

$$f(x_0) \leq f(x) \quad \text{for all } x \in V.$$

The initial set of ordered pairs of leaves is as follows:

$$\mathbf{L} = \{(X, Y) \in \partial T \times \partial T \mid X \neq Y\}.$$

Cherry picking algorithm in spectral method

Step 1: Compute Λ_{XY} for each ordered pair $(X, Y) \in \mathbf{L}$.

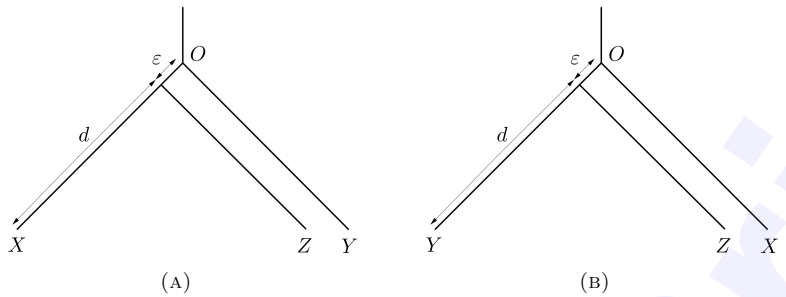


FIGURE 3. Failure of Step 1,2 by error of spectrum

Step 2: Find a cherry candidate

$$(X_0, Y_0) = \text{argmin}_{(X,Y)\in\mathbf{L}} \frac{\max \Lambda_{XY} - \min \Lambda_{XY}}{\text{mean}\Lambda_{XY}}.$$

Step 3: Check two additional criteria:

$$Y_0 = \text{argmin}_{Z\in\partial T\setminus\{X_0\}} F(X_0, Z), \quad X_0 = \text{argmin}_{Z\in\partial T\setminus\{Y_0\}} F(Y_0, Z). \tag{5}$$

If two criteria in (5) are all checked yes, go to Step 4. If not, go to Step 2 with the replaced $\mathbf{L}$ by

$$\mathbf{L} = \mathbf{L} \setminus \{(X_0, Y_0), (Y_0, X_0)\}.$$

Step 4: Take (X_0, Y_0) as a cherry.

If we meet a pathological case that $\mathbf{L} = \emptyset$ in Step 2, we could choose the first cherry candidate as a cherry.

Once a cherry $(\bar{X}, \bar{Y})$ is chosen in the above cherry picking algorithm, renew the terminal vertices by replacing $\bar{X}, \bar{Y}$ with the node $[\bar{X}, \bar{Y}]$ and repeat the same algorithm to pick a next cherry with the distance redefined as

$$d([\bar{X}, \bar{Y}], W) = d(\bar{X}, W) - \lambda_{\bar{X}\bar{Y}}(W), \quad W \neq X, Y.$$

Remark 3.2. In order to simplify the above algorithm, one may take (X, Y) as a cherry when $F(X, Y)$ is the minimum without the first part of algorithm (Step 1,2,3). Then even if T has a tree metric, the tree could not be reconstructed as follows: In Fig. 3(B), if $d(O, X) = \delta$, then $F(X, Y) = (n-2)\delta + 2\epsilon$. If δ is small compared to d, then (X, Y) can be chosen as a cherry instead of (Y, Z).

Example 1. At first glance, $Q(X, Y)$ in the neighbor-joining method [1] and $F(X, Y)$ in the spectral method have some similarity, where

$$Q(X, Y) = (n-2)d(X, Y) - \sum_Z d(X, Z) - \sum_Z d(Y, Z).$$

Simply, we consider a quartet (Fig. 4). Let (A, B) be a cherry. Then

$$Q(A,B) = -2(a+b+c+d) - 4\alpha,\ Q(A,C) = -2(a+b+c+d) - 2\alpha,$$

where α is the distance between the internal nodes. Hence we take (A, B) as a cherry if $Q(A, B)$ is the minimum.

On the other hand,

$$F(A,B) = 2a,\ F(A,C) = 2a + \alpha,$$

hence Q is a function concerned about a sum of distances between nodes and F is a function concerned about the distance from a leaf and the closest internal node, i.e., spectrum.

In the case of the neighbor-joining method, $Q(A, B) = Q(C, D)$, so it is necessary to determine which pair between (A, B) and (C, D) should be selected for a cherry. On the other hand, $F(A, B) \neq F(C, D)$ in our algorithm.

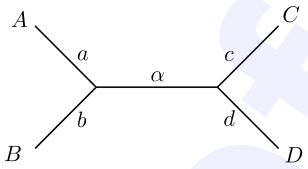
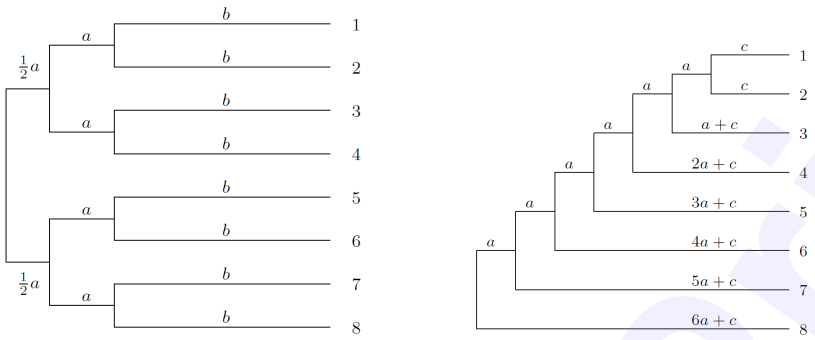


FIGURE 4. Lengths of edges for quartet

4. Performance analysis

4.1. Building phylogenetic trees with simulated data

We chose phylogenetic tree models as in Fig. 5 and simulated DNA sequence data on these trees using the program `seq-gen` with the GTR substitution model [3]. We compared the reconstruction percents for the neighbor-joining and the spectral method in Tables 1, 2. For each sequence length, 1000 DNA data sets were used. The spectral method has the edge over the neighbor-joining in case of the model A in Fig. 5(A), but loses in case of the model B in Fig. 5(B). The last columns in Tables 1, 2 show the numbers of correctly reconstructed data sets by either the neighbor-joining method or the spectral method. For example, in the case of $(a, b) = (0.01, 0.07)$ of Model A for 500bps, the spectral method succeeds in about $10\% \approx \frac{2.8}{27.2}$ among the failed data sets by the neighbor-joining method. In the case of $(a, c) = (0.01, 0.04)$ of Model B for 500bps, the spectral method succeeds in about $8\% \approx \frac{2.3}{30.1}$ among the failed data sets by the neighbor-joining method. Hence the spectral method can be considered as a complementary method for reconstructing tree.



(A) Model A with $(a, b) = (0.02, 0.19)$

(B) Model B with $(a, c) = (0.02, 0.13)$

FIGURE 5. Model trees

TABLE 1. Comparing results for algorithms for Model A

bps	NJ	SP	NJ ∪ SP
100	10.5	11.0	12.4
200	26.4	28.9	29.6
300	48.3	49.0	51.3
400	62.3	64.9	66.2
500	72.8	74.5	75.6
600	81.7	83.4	84.4
700	86.2	87.0	87.8
800	89.5	91.4	91.7
900	93.3	94.8	95.1
1000	94.8	95.4	95.8

$(a, b) = (0.01, 0.07)$

bps	NJ	SP	NJ ∪ SP
100	6.3	6.5	7.1
200	17.6	18.5	20.2
300	29.4	30.6	32.5
400	45.7	47.8	49.3
500	55.8	58.3	59.9
600	68.7	70.8	72.3
700	73.4	75.1	76.7
800	77.2	80.6	81.0
900	83.4	84.7	85.5
1000	89.4	90.6	90.9

$(a, b) = (0.02, 0.19)$

4.2. Building phylogenetic trees with real data

For real data, we use the September 2005 ENCODE Multi-Species Sequence Analysis sequence freeze with multiple sequence alignments [5], which are available at `http://hgdownload.soe.ucsc.edu/goldenPath/hg17/encode/alignments/SEP-2005`. We restrict our attention to the rodent problem to construct the phylogenetic tree for 8 species: human, chimp, galago, mouse, rat, cow, dog, and chicken. We process each of the 44 ENCODE regions to obtain data sets which have ungapped columns greater than 100 bps in length. We obtain 83 data sets in manually chosen 14 Enm regions and 397 data sets in all 44 Encode regions.

TABLE 2. Comparing results for algorithms for Model B

bps	NJ	SP	NJ ∪ SP
100	10.5	8.9	11.6
200	28.9	27.9	32.0
300	47.0	45.9	50.4
400	60.1	57.6	63.0
500	69.9	66.6	72.2
600	76.6	73.5	78.4
700	85.0	82.3	86.8
800	87.6	86.2	89.5
900	91.6	89.6	92.4
1000	94.7	93.9	96.1

$(a, c) = (0.01, 0.04)$

bps	NJ	SP	NJ ∪ SP
100	5.7	5.6	6.4
200	18.1	17.1	20.8
300	30.6	29.1	33.6
400	43.4	42.3	47.6
500	53.9	51.8	56.7
600	60.7	60.8	65.3
700	72.6	71.4	76.4
800	75.5	73.9	79.2
900	80.9	78.2	82.9
1000	84.8	84.9	87.0

$(a, c) = (0.02, 0.13)$

The proposed spectral method shows the better results than the neighbor joining method as given in Table 3. In Enm case, it is almost twice.

TABLE 3. Comparing results for algorithms on data from ENCODE alignments

data type	NJ	SP	NJ ∪ SP
83 data sets in Enm region	10.8	20.5	24.1
397 data sets in All region	8.1	12.6	16.1

5. Conclusion

The spectral method shows similar performance to the neighbor-joining method for simulated data. The spectral method could find significant amount of correct trees among the failed data sets by the neighbor-joining method. Furthermore, the spectral method shows much better performance than the neighbor-joining method for real data. Hence, it could be a quite good complementary method.

References

[1] J.-H. Cho, D. Joe, and Y. R. Kim, *Analysis of neighbor-joining based on box model,* J. Appl. Math. Computing **25** (2007), 150–169.

[2] C. B. Croke, *Rigidity for surfaces of nonpositive curvature*, Comment. Math. Helv. **65** (1990), no. 1, 150–169.

[3] A. Rambaut and N. Grassly, *Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees*, Comput. Appl. Biosci. **13** (1997), 235–238.

[4] N. Saitou and M. Nei, *The neighbor-joining method: A new method for reconstructing phylogenetic trees*, Molecular Biology and Evolution **4** (1987), 406–425.
[5] ENCODE Project Consortium, *The ENCODE (ENCyclopedia Of DNA Elements) Project*, Science **306** (2004), 636–40, DOI: 10.1126/science.110513.

Seong-Hun Paeng
Department of Mathematics
Konkuk University
Seoul 05029, Korea
Email address: shpaeng@konkuk.ac.kr

Chunjae Park
Department of Mathematics
Konkuk University
Seoul 05029, Korea
Email address: cjpark@konkuk.ac.kr