数理解析研究所講究録
1489 巻 2006 年 98-101

2005 年度冬の LA シンポジウム [15]

Robustness of Greedy Type Minimum Evolution Algorithms

繁住 健哉 *
Takeya Shigezumi

東京工業大学情報理工学研究科数理・計算科学専攻
Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology

概要

For a phylogeny reconstruction problem, Desper and Gascuel [2] proposed Greedy Minimum Evolution algorithm (in short, GME) and Balanced Minimum Evolution algorithm (in short, BME). Both of them are faster than the current major algorithm, Neighbor Joining (in short, NJ); however, less accurate when an input distance matrix has errors. In this paper, we prove that BME has the same optimal robustness to such errors as NJ but GME does not. Precisely, we prove that if the maximum distance error is less than a half of the minimum edge length of the target tree, then BME reconstruct it correctly. On the other hand, there is some distance matrix such that maximum distance error is less than $\frac{2}{\sqrt{n}}$ of the minimum edge length of the target tree, for which GME fails to reconstruct the target tree.

1 Problem and previous work

A phylogeny reconstruction problem is to determine an evolutionary tree representing relationships between a set of species from a distance matrix expressing the "closeness" of each pair of species. Recently, the phylogeny reconstruction is not only used for biology, but also for clustering various kinds of data, such as languages[7], music, etc.[8] To state the phylogeny construction problem formally, we need some notations. For any n, we use $L = \{1, \dots, n\}$ to denote a set of n species. Let $D = (d_{ij})$ over L be a distance matrix. An evolutionary tree over L is a tree such that its set of leaves is L and all internal vertices are of degree three. For any tree T and for any i and j in L, let d_{ij}^T denote the length of the unique path between i and j, i.e., the sum of the length of edges on the path.

Now the phylogeny construction problem is stated as follows.

Instance: A distance matrix $D = (d_{ij})$ over $L = \{1, \dots, n\}$.

*Email: Takeya.Shigezumi@is.titech.ac.jp

Question: Find a tree T^* minimizing the following square error:

$$\sum_{i<j} \left| d_{ij} - d_{ij}^{T^*} \right|^2 .$$

Day [6] showed that this problem is NP-hard in general. It is, however, possible to "reconstruct" a tree in polynomial time if a distance matrix is sufficiently close enough to the one for a given tree. Note that for any tree T, this T itself is the unique solution for the phylogeny construction problem on the distance matrix $D^T = (d_{ij}^T)$. Several polynomial time algorithms have been proposed for reconstructing the original tree T from D^T. Among them, Neighbor-Joining[4], its variants are most famous and they run in $O(n^3)$ for reconstructing a tree of n leaves. Though polynomial, this running time is not sufficient enough for constructing a large evolutionary trees, and several attempts have been made to improve the efficiency .

Recently, Desper and Gascuel [2] introduced Greedy Minimum Evolution (GME) algorithm based on the ordinary least squares (OLS) branch length estimation and Balanced Minimum Evolution (BME) algorithm based on a branch length estimation scheme of Pauplin [3]. BME runs in $O(n^2 \mathrm{diam}(T))$ where $\mathrm{diam}(T)$ is a diameter of the target tree, and GME runs in $O(n^2)$, which is optimal considering that the size of the input distance matrix is $O(n^2)$. Moreover, BME has an important property that the output tree has no edge with negative length [1]. Since edges with negative length are biologically meaningless, this property is also an advantage of BME.

2 Measuring errors

In practice, a distance matrix has errors, that is, some distance δ_{ij} differs from the original distance d_{ij}^T. For measuring such errors, Atteson [5] introduced the notation of "l_∞ radius." We say that a distance matrix $\Delta = (\delta_{ij})$ has *l_∞ radius α* to an original distance $D^T = (d_{ij}^T)$ if

$$\|D^T - \Delta\|_\infty = \max_{i<j} |d_{ij}^T - \delta_{ij}| < \alpha \min_{e \in E(T)} \{l(e)\},$$

where $E(T)$ is the set of edges of T and $l(e)$ denotes the length of an edge e. An algorithm has *l_∞ radius α* if it yields the original T for any distance matrix of l_∞ radius α to D^T. Note that no algorithm has l_∞ radius larger than $\frac{1}{2}$. Because if we allow errors larger than $\frac{1}{2}l(e)$, there exists one distance matrix made from two different trees with such errors. Atteson showed that NJ has l_∞ radius $\frac{1}{2}$.

3 Our result

Desper and Gascuel showed the relationship between BME length estimation and weighted least squares branch length estimation [1] as a theoretical foundation of BME. However, there is still no theoretical analysis about l_∞ radius for GME and BME. We analyze BME and GME, and obtain the following two theorems, which claim that BME has l_∞ radius $\frac{1}{2}$ but GME does not.

Theorem 1. *BME has l_∞ radius $\frac{1}{2}$.*

Theorem 2. *Let α_G be an l_∞ radius of GME. Then, $\alpha_G < \frac{2}{\sqrt{n}}$.*

Proof. See [9]. ■

4 Discussion

According to our theorems, BME has the same robustness as NJ, and it is faster than NJ. On the other hand, GME is faster than BME, it is less accurate in some cases. Our proof is obtained by a precise analysis of the proof of Desper and Gascuel [1] for the correctness of BME on error free distance matrix.

GME and BME algorithm chooses the position of the leaf k step by step. Let T_k^* be a tree over $\{1, \ldots, k\}$ induced from the target tree T^*. The proof of Theorem 1 is by the induction of k. In $k = 3$, there is only one topology T_3^*. Let us consider the k-th round. By the induction hypothesis, BME reconstructs the target tree topology T_{k-1}^*. BME chooses T_k minimizing the estimated length of T_k, where Δ_k is a distance matrix on $\{1, 2, \ldots, k\}$ induced from Δ. We analyzed the estimated length of T_k and an estimated length for a different topology T_k'. To prove Theorem 2, we presented an input matrix which GME fails to output the target tree.

Desper and Gascuel also proposed Nearest Neighborhood Interchange (FASTNNI) algorithm and Balanced Nearest Neighborhood Interchange (BNNI) algorithm to improve a phylogeny tree into minimal one. However, according to our theorem, we do not need to use FASTNNI or BNNI when GME or BME is used for a distance matrix Δ with sufficiently small $\|D^{T^*} - \Delta\|_\infty$ error. Moreover, our theorem 2 claims that if GME fails to put the last leaf, FASTNNI cannot correct such fault.

参考文献

[1] R. Desper, O. Gascuel. Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting. *J. of Mol. Biol. Evol.* 2004 21(3): 587–598.

[2] R. Desper, R. Gascuel. Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum Evolution Principle. *J. Comp. Biol.* 2002 9: 687–705.

[3] Y. Pauplin. Direct Calculation of a Tree Length Using a Distance Matrix. *J Mol Evol.* 2000 51(1): 41–7.

[4] N. Saitou, M. Nei. The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees. *Mol Biol Evol.* 1987 4(4): 406–25.

[5] K. Atteson. The Performance of Neighbor-Joining Algorithms of Phylogeny Reconstruction. *Proceedings of the Third Annual International Conference on Computing and Combinatorics, Lecture Notes In Computer Science Vol. 1276* 1997 :101 – 110.

[6] W. Day, D. Sankoff. Computational Complexity of Inferring Phylogenies by Compatibility. *Systematic Zoology,*1986 35(2):224–229.

[7] D. Benedetto, E. Caglioti and V. Loreto. Language Trees and Zipping. *Physical Review Letters* 2002 88(4):048702-1–04872-4.

[8] R. Cilibrasi, P. Vitányi. Clustering by Compression. *IEEE Transactions on Information Theory*, 2005 51(4):1523–1545.

[9] T. Shigezumi. Robustness of Greedy Type Minimum Evolution Algorithms. *Dept. of Math. and Comp. Sciences Tokyo Institute of Technology Research Reports (Series C)*, C-218, 2005.
http://www.is.titech.ac.jp/research/research-report/C/C-218.ps.gz