MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 469-476

ISSN 0340 - 6253

A Method for Constructing Phylogenetic Tree Based on the Minimum Spanning Tree of the Complete Graph

Jie Yang [1*], Zhi Cao[2], Huanwen Chen [1], Kai Long [1], Gangcheng Li[1], Li Zhao[1]

[1] *Hunan College of Information, Wangcheng Changsha 410200, China*
[2] *School of computer and communication, Hunan University, Changsha Hunan, 410082, China*

(Received April 6, 2010)

Abstract: The evolutionary history of various species can be represented by constructing the phylogenetic tree. In this paper, we proposed a novel method for constructing phylogenetic tree based on the minimal spanning tree of the complete graph, which is taken from the similarity matrix computed by 3D graphical representation of DNA sequences. This method didn't require sequence alignment and the computation was simple. The experiments proved its validity.

1. Introduction

With the development of molecular biology and bioinformatics, phylogenetic analysis and constructing the phylogenetic tree has become one of the major problems in computational biology. This is because the evolutionary relationship of species provides a great deal of information about their biochemical machinery. So many researchers focus on the research of constructing the phylogenetic tree [1].

A phylogenetic tree is a tree showing the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor. There are two main methods of constructing phylogenetic trees [2, 3]: (1) algorithm-based method such as UPGMA (unweighted pair group method with arithmetic mean) [4], Fitch-Margoliash [5], and NJ(Neighbor Joining)[6,7]. It is important to obtain a similarity matrix showing the

[*] Corresponding author email address: jt_yangjie@126.com (J. Yang)

relation of species. The computation of similarity matrix requires multiple sequence alignment. And the similarity matrix will be reconstructed constantly during the process of constructing the phylogenetic tree. Therefore, the time complexity is very high. (2) Optimal principles method such as maximum parisimony method (MP) [8] and maximum likelihood method (ML) [9]. It is important to get the best objective function based on the mathematical model. However, it is difficult to obtain the best objective function from the mathematical model. With more and more DNA and protein sequences have been obtained, the problem of time complexity has become one of the major problems of constructing the phylogenetic tree [10-16].

In this paper, we propose a minimum spanning tree method based on the complete graph, it does not also require multiple sequence alignment, and the computation is simple. The similarity matrix is used to do our experiment, which is computed by 3D graphical representation of DNA sequences based on dual nucleotides [17], and the experiment illustrates the utility of the approach.

2. Method

Obviously, we can obtain a complete graph based on the similarity matrix showing the relation of species. We propose a minimum spanning tree searching algorithm based on prim method and the complete graph, and in order to improve the quality of clustering, the depth-first search method is used. The main idea of our method is as follows:

1. Given a similarity matrix showing the relation of species.
2. Constructing a graph G<V, E>, where V denotes the set of vertices, E denotes the set of edges, each of which has an associated weight Wi. Denote Si as the sum of weights of all directly connected edges with vertex Vi. We add the vertex Vi with the smallest Si to a new vertex set U, where U={Vi}, and add all the edges connecting with Vi to a new edge set T(E).
3. Find the vertex Vj in U, which has the edge Ei of the minimum weight between Vi and Vj, and add Vj to U, where U={Vi,Vj}. While we join Ei to T(E). If edges in T(E) form a loop, we remove the edge which has the largest weight in the loop from T(E).

4. Depth-first search for another vertex Vk, which has the edge Ei of the minimum weight between Vj and Vk, and add Vk to U, where U={Vi,Vj,Vk}.And then we compare the weight between Vk and the vertex of the previous step adding to U before Vk to that of the two-step ,which will be a smaller edge ,we join it to T(E). If edges in T(E) form a loop, we remove the edge which has the largest weight in the loop from T(E).
5. Repeating the fourth step, until U=V.

There are n-1 edges in T(E), so T = (U, T(E)) is a minimum spanning tree.

The pseudo-code of algorithm is as follows:

```
I_primMLT(G)
Input：a[n][n];   // a[n][n] is stored edge weights for figure G change into a matrix
U[]←ϕ,TE[]←ϕ  //initialization, starting from the first vertex V0
for(i=1;i<=n;i++)//n is the number of vertices
for(j=1;j<=n;j++)
{
Weight_Sum[i] ←Weight_Sum +a[i][j]; //Weight_Sum[] is stored the sum of weights
                                    which connected every vertex to all the other
                                    vertices
}//endfor
If Min_Sum>Weight_Sum[i]   // selecting vertex tag of the minimum sum of weights
U[i]← vi, TE[i]←a[vi][1…n];
vnum←Findmin（）,U[i] ←vnum // starting from Vi, depth-first search for the smallest
                                    edge, returning another vertex vnum of the current
                                    minimum edge
 If the current minimum edge Ei more than the edge Elab connecting U[i] to U [i-1] ← a
[U[i]][U[i-1]],so joined Elab to TE[i]
Else joined Ei to TE[i]
If   edges in TE[i] form a loop, we remove the edge which has the largest weight in the loop
from TE[i]. }
//   endfor   then the cycle of depth-first search for a next vertex
```

3. Experiment

Obviously, we can obtain a complete graph based on a similarity matrix, and the weight of edge come from the similarity matrix. In this paper, we use the obtained similarity matrix in [17] to do the experiment (shown in table 1). Starting directly from the similarity matrix, and in accordance with the above given minimum spanning tree approach based on the complete graph, we can get a minimum spanning tree, as shown in figure 1.

Table 1 The symmetric similarity matrix for the coding sequences

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rabbit	Rat	Bovine	Gorilla	Chimpanzee
Human	0	0.1254	0.4844	0.3571	0.0425	0.0594	0.0108	0.0292	0.0589	0.0002	0.0117
Goat		0	0.1479	0.1584	0.0387	0.2746	0.0789	0.1754	0.0601	0.1172	0.0830
Gallus			0	0.1601	0.2749	0.7214	0.3785	0.5838	0.3635	0.4681	0.3797
Opossum				0	0.2023	0.5831	0.2969	0.3914	0.2528	0.3450	0.2865
Lemur					0	0.1815	0.0137	0.0899	0.0537	0.0368	0.0131
Mouse						0	0.1103	0.0559	0.1289	0.0663	0.1167
Rabbit							0	0.0568	0.0575	0.0080	0.0008
Rat								0	0.0468	0.0316	0.0604
Bovine									0	0.0570	0.0641
Gorilla										0	0.0087
Chimpanzee											0

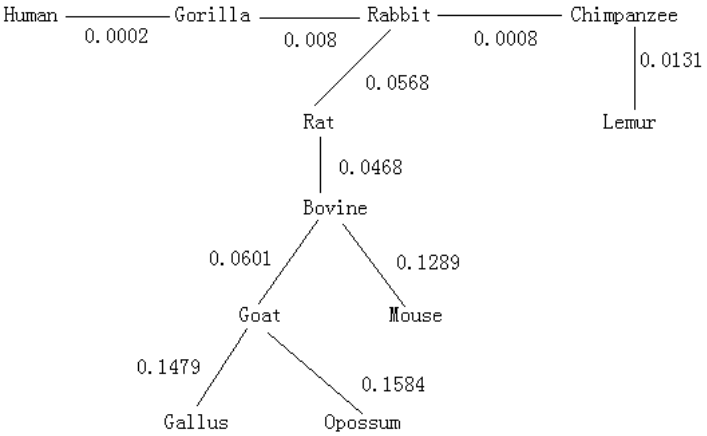

Figure 1 the minimum spanning tree using our method

Selecting $\lambda \in [0,1]$,and cutting off the branches of the weight below λ, we get a non-connected graph. So all connected branches constitute the horizontal classification of λ, we select it in turn : $\lambda \in \{0.1584, 0.1479, 0.1289, 0.0601, 0.0568, 0.0468, 0.0131, 0.008, 0.0002\}$.

For a fixed threshold $\lambda \in [0,1]$,pruning the branches with the weight less λ, we get a non-connected graph. So all connected branches constitute the horizontal classification of λ, we select it in turn: $\lambda \in \{0.1584, 0.1479, 0.1289, 0.0601, 0.0568, 0.0468, 0.0131, 0.008, 0.0002\}$.

Getting λ =0,11 species are divided into 11 categories: {human},{goat},{gallus},{opossum},{mouse},{rabbit},{rat},{bovine},{gorilla},{lemur},{chimpanzee};

Getting λ =0.0002, 11 species are divided into 10 categories: {human,gorilla},{goat}, {opossum },{gallus},{mouse},{rabbit},{rat},{bovine},{lemur},{chimpanzee};

Getting λ =0.0008, 11 species are divided into 9 categories: {human,gorilla},{goat}, {opossum },{gallus},{mouse},{rat},{bovine},{lemur},{ rabbit, chimpanzee};

Getting λ =0.0008, 11 species are divided into 8 categories: {human, gorilla, rabbit, chimpanzee},{goat}, {opossum },{gallus},{mouse},{rat},{bovine},{lemur};

Getting λ =0.0131, 11 species are divided into 7 categories: {human,gorilla, rabbit, chimpanzee, lemur},{goat}, {opossum },{gallus},{mouse},{rat},{bovine};

Getting λ =0.0468, 11 species are divided into 6 categories: {human,gorilla, rabbit, chimpanzee, lemur},{goat}, {opossum },{gallus},{mouse}, {rat,bovine};

Getting λ =0.0568, 11 species are divided into 5 categories: {human,gorilla, rabbit, chimpanzee, lemur, rat,bovine},{goat}, {opossum },{gallus},{mouse};

Getting λ =0.0601, 11 species are divided into 4 categories: {human,gorilla, rabbit, chimpanzee, lemur, rat,bovine, goat}, {opossum},{gallus},{mouse};

Getting λ =0.1289, 11 species are divided into 3 categories:

{human,gorilla, rabbit, chimpanzee, lemur, rat,bovine, goat, mouse}, {opossum},{gallus};

Getting λ=0.1479, 11 species are divided into 2 categories: {human,gorilla, rabbit, chimpanzee, lemur, rat,bovine, goat, mouse, gallus},{opossum };

Getting λ =0.1584, 11 species are divided into 1 categories: {rabbit,opossum,mouse,bovine,gorilla,lemur,chimpanzee,human,goat,rat,gallus}.

According to the minimum spanning tree method , we can obtain a dynamic clustering map, where the phylogenetic tree of 11 species, shown in figure 2, which is compared with the tree for the neighbor program of the construct software PHYLIP, shown in figure 3. Although the evolutionary trees are different each other, the effect of clustering basically is the same, even it is able to better reflect the evolutionary relationships between species. Leading to such differences that it may be not obvious to the similarity matrix differences. We compare with the βgene sequence of the first exon in 11 species. The largest characteristic of these sequences is stronger conservative, and less differences between sequences.

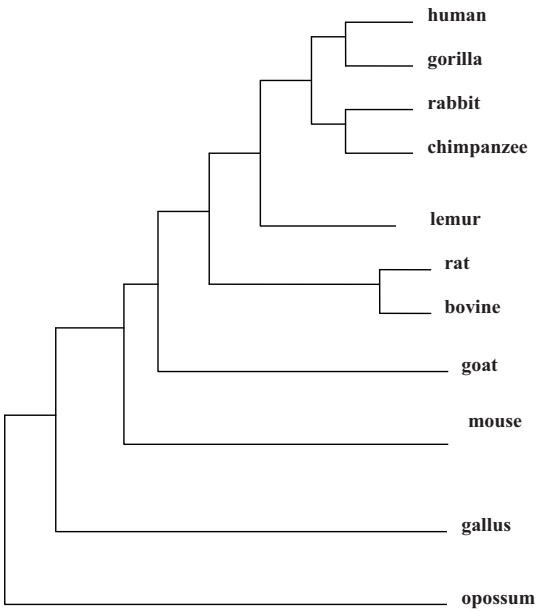

Figure 2: constructing the phylogenetic tree by the minimum spanning tree algorithm

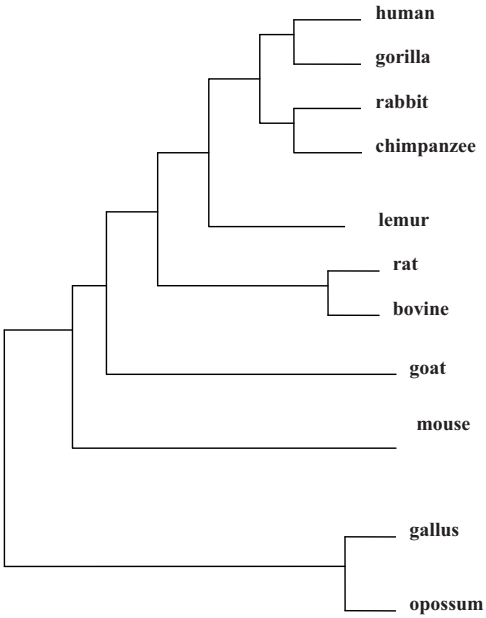

Figure 3: constructing the phylogenetic tree by software PHYLIP

4. Conclusion

In this paper, we introduce a minimum spanning tree method based on the complete graph to construct the phylogenetic tree, which is compared with algorithm-based method and based on the best principles method, the advantages of this method does not need for multiple sequence alignment and building the evolutionary model, the whole algorithm is calculated very simply. We use the minimum spanning tree method to construct phylogenetic tree,which is the same as the Neighbor program of software PHYLIP.

References

[1] M. J. Sanderson, A. C. Driskell, The challenge of constructing large phylogenetic trees, *Trends Plant Sci.* **8** (2004) 374–379.

[2] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, CSHL Press, New York, 2001, pp. 337–342.

[3] M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford Univ. Press, Oxford, 2000.

[4] R. R. Sokal, C. D. Michener, A statistical method for evaluating systematic relationnships, *Univ. Kansas Sci. Bull.* **28** (1958) 1409–1438.

[5] S. Kumar, K. Tamura, M. Nei, MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment, *Brief Bioinform.* **5** (2004) 150–163.

[6] N. Saitou, M. Nei, The neighbor-joining method : A new method for reconstructing phylogenetie trees, *Mol. Biol. Evol.* **4** (1987) 406–425.

[7] I. Elias, J. Lagergren, Fast neighbor joining, *Theor. Comput. Sci.* **410** (2009) 1993–2000.

[8] O. Satoshi, W. Li, NJML: A hybrid algorithm for the neighbor-joining and maximum-likelihood methods, *Mol. Biol. Evol.* **17** (2004) 1401–1409.

[9] J. Felsenstein, Evolutionary trees from DNA sequences:a maximum likelihood approach, *J. Mol. Evol.* **17** (1981) 368–376.

[10] B. Liao, X. Shan, W. Zhu, R. Li, Phylogenetic tree construction based on 2D graphical representation, *Chem. Phys. Lett.* **422** (2006) 282–288.

[11] W. Wang, B. Liao, T. Wang, W. Zhu, A graphical method to construct a phylogenetic tree, *Int. J. Quantum. Chem.* **106** (2006) 1998–2005.

[12] W. Zhu, B. Liao, R. Li, A method for constructing phylogenetic tree based on a dissimilarity matrix, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 483–492.

[13] B. Liao, Y. Liu, R. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, *Chem. Phys. Lett.* **421** (2006) 313–318.

[14] B. Liao, X. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, *J. Comput. Chem.* **27** (2006) 1196–1202.

[15] W. Chen, B. Liao, Y. Liu, W. Zhu, Z. Su, A numerical representation of DNA sequence and its applications, *MATCH Commun. Math. Comput. Chem.* **60** (2008) 291–300.

[16] B. Liao, L. Liao, G. Yue, R. Wu, W. Zhu, A Vertical and horizontal method for constructing phylogenetic tree, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 691–700.

[17] Z. Cao, B. Liao, R. Li, A group of 3D graphical representation of DNA sequences based on dual nucleotides, *Int. J. Quantum. Chem.* **108** (2008) 1485–1490.