₹ 200 ISSN - 2249-555X
Volume : 1 Issue : 7 April 2012

Indian Journal Of Applied Research

Journal for All Subjects

www.ijar.in
Listed in International ISSN Directory, Paris.

ISSN - 2249-555X

Indian Journal of Applied Research

Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College, New Congres Bhavan, Paldi,
Ahmedabad-380006, Gujarat, INDIA
Contact.: +91-9824097643 E-mail : editor@ijar.in

INDEX

Volume : 1 | Issue : 7 | April 2012 ISSN - 2249-555X

Research Paper Engineering

Analyzing the Phylogenetic Trees with Tree- building Methods

* Jasmine **Pankaj Bhambri ***Dr. O.P. Gupta

* Research Scholar, Department of C.S.E, G.N.D.E.C., Ludhiana

** Assistant Prof., Department of IT., G.N.D.E.C., Ludhiana

*** Dy. Director, School of Info. Tech, P.A.U., Ludhiana

Keywords :

1. INTRODUCTION

1.1 Introduction to Bio-informatics

Bioinformatics is a merger of different disciplines including statistics, mathematics, computer science, molecular biology etc. Over the past few decades, major advancements in this field have led to an explosive growth in the biological information. The computerized databases are used to organize, store and index the data. The scope of bio-informatics lies in the development of computational tools and databases.

1.2 Phylogenetic Trees

The Development of a biological form from other pre-existing forms or its origin to the current existing forms through some modifications is known as evolution. The study of evolutionary history of some organisms using tree-like diagrams is known as Phylogenetic tree construction or phylogenetic analysis. These constructions are being used to visualize similarities and divergence among related biological sequences which is done through sequence alignment. The evolutionary divergence is called as phyogeny which is being represented by tree branching patterns. Each time a branch divides into a smaller branch, it shows the emergence of a new group of organisms.

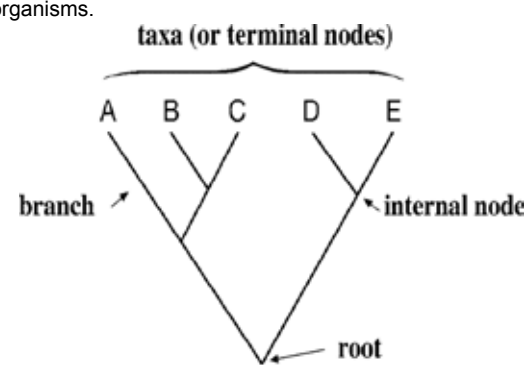

Figure1.1: A typical bifurcating phylogenetic tree showing root, internal nodes, terminal nodes and branches.

2. PROBLEM FORMULATION

The analysis of phylogenetic trees of different species is being formulated by having comparisons among the different tree building methods, thereby obtaining an optimal tree.

2.1 Phylogenetic Analysis

The most popular distance-based methods used for the analysis are:

- The Un-weighted pair group method with arithmetic mean (UPGMA)
- Neighbor joining (NJ)

UPGMA Method: It follows a clustering procedure where initially each species is a cluster on its own. The closest 2 clusters are joined and distance of the joint pair is calculated by taking the average. This process is being repeated until all species are connected in a single cluster. This method is simple and fast. However, it behaves poorly at most cases where the above presumptions are not met.

Neighbor Joining Method (NJ): It begins with an unresolved star-like tree. Each pair is evaluated for being joined and the sum of all branches length is calculated of the resultant tree. The pair that yields the smallest sum is considered the closest neighbors and is thus joined. A new branch is inserted between them and the rest of the tree and the branch length is recalculated. This process is repeated until only one terminal is present. It generally gives better results than UPGMA method. Under some condition, this method will yield a biased tree.

3. MODEL OF COMPUTATION

The distance-based approaches are used for the comparison of trees. The methods used are:

- UPGMA tree-building method
- Neighbor joining method.

3.1 Implementation of UPGMA Method

The simplest clustering method is UPGMA, which builds a tree by a sequential clustering method. UPGMA clusters sequences based on a distance matrix. As the clusters grow, a tree is being assembled.

To measure the distances between the sequences, the following steps are being followed:

- A distance matrix is considered to identify the least dissimilar groups. All nodes are given equal weights. If there are several equidistant minimal pairs, one is picked randomly.
- Combine two nodes to form a new group node.
- This results in the formation of a new, clustered distance matrix having one fewer row and column than the initial matrix.
- Dissimilarities that are not involved in the formation of the new cluster remain unchanged.
- Identify the next smallest dissimilarity and combine those nodes to generate a second clustered dissimilarity matrix. In this process a new distance matrix is formed, and the tree continues to be constructed again and again.
- Continue until there are only two remaining groups, and join those.

The UPGMA method is a commonly used distance method in a variety of applications.

3.2 Implementation of Neighbor Joining Method
It needs the construction of a distance metric.
A distance metric is distinguished by three properties:

- The distance from a point to itself must be zero, that is, D (x, x) = 0;
- The distance from point x to y must equal the distance from y to x, that is, D (x, y) = D (y, x);
- The triangle inequality must apply in that D (x, y) ≤ D (x, z) + D (z, y).

The similarities are also useful in this case. But the distances (which differ from differences when they obey the above properties) offer appealing properties for describing the relationships between objects.

Given the observed distances between any two sequences i, j can be denoted dij. The sum of the branch lengths of the tree from node i and j can be denoted d'ij. Ideally, these two distance measures are the same, but phenomena such as the occurrence of multiple substitutions at a single position typically cause dij and d'ij to differ. The branch lengths can be estimated as:

$$\sum_i \sum_j w_{ij}\left(d_{ij}-d_{ij}^{'}\right)^2 \quad (3.1)$$

The goal is to minimize this value; it is zero when the branch lengths of a tree match the distance matrix exactly.

However, the NJ method does not assume the nodes to be equidistant from the root. It corrects for unequal evolutionary rates between sequences by using a conversion step. This conversion requires the calculations of "r-values" and "transformed r-values".

The values are described as:

- d'_{AB} = converted distance between A and B
- d_{AB} = actual evolutionary distance between A and B

The formula for conversion is:

$$d'_{\mathrm{AB}} = d_{\mathrm{AB}} - 1/2 \times (r_{\mathrm{A}} + r_{\mathrm{B}}) \quad (3.2)$$

The value of $r_{\mathrm{A}}\left(\mathrm{or} r_{\mathrm{B}}\right)$ is the sum of distances of A (or B) to all other nodes. A generalized expression of the r-value, that is, r_{i}, is calculated based on the following formula:

$$r_i = \sum d_{\mathrm{ij}} \quad (3.3)$$

where i and j are two different nodes. The r-values here are needed to create a modified distance matrix for the algorithm to perform. The transformed r-values (r') are used to determine the distances of an individual taxon to the nearest node.

$$r'_i = r_i / \ n-2 \quad (3.4)$$

Here n is the total no. of nodes.

4. RESULTS AND DISCUSSION

4.1 Purpose of Analysis

The purpose of phylogenetic analysis is to detect evolutionary changes in species and thereby calculating structural and functional relations among various species.

4.1.1 Methods and Assumptions

Phylogenetic analysis is very subjective in nature. Various assumptions are also needed to obtain the results. One of the most important assumptions is the rate or probability at which one nucleotide substitutes for another. This is called as substitution model. The first main step is to evaluate whether the rates are equal or not. This is done by two methods. Those are:

- Jukes-Cantor Method: No biasing is there. Substitutions occur more randomly. Equal probability of mutation for all nucleotides is there. For deriving evolutionary distances that include hidden changes is calculated by using a logarithmic function.

$$d_{\mathrm{AB}} = -(3/4)\ln\left[\ 1-(4/3)\, p_{\mathrm{AB}}\right] \quad (4.1)$$

where d_{AB} is the evolutionary distance between sequences A and B and p_{AB} is the observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.

- Kimura Model: The transitions are more frequent than transversions. It is also called as the Kimura two-parameter model. This is a more sophisticated model than the above model as in this model, mutation rates for transitions and transversion are assumed to be different, which is more realistic. Therefore, it provides a more realistic estimate of evolutionary distances. It uses the following formula:

$$d_{\mathrm{AB}} = -(1/2)\ \ln\left(\ 1-2p_{\mathrm{ti}}-p_{\mathrm{tv}}\right) - (1/4)\ln\left(\ 1-2p_{\mathrm{tv}}\right) \quad (4.2)$$

where d_{AB} is the evolutionary distance between sequences A and B, p_{ti} is the observed frequency for transition and ptv is the frequency of transversion.

4.2 Results Regarding Analysis

The methods for determining the phylogenetic trees shows variations among the trees build. These results are being carried out by taking five species of the same family of organisms. As per the data, the phylogenetic tree by UPGMA method is being computed. The results shown are outcome of the distance metrics which are being applied. The second tree which is build by the neighbor joining method is also being computed. The changes in the trees are due to the evolutionary distances between the branches of the trees.

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The model is constructed for deciding that which method among the distance-based tree building methods yield an optimal tree

UPGMA method had a critical assumption that the rate of nucleotide or amino acid substitution is constant for all the branches in the tree. The branch lengths can be used to estimate the dates of divergence. The assumptions when violated with unequal substitution rates along different branches of the tree, produces an incorrect tree. The un-weighted distances also made an impact that a better tree must be produced for the species given.

The neighbour-joining method gave an optimal tree for the input data and information given because it did not use the un-weighted distances. Also, this algorithm is especially useful when studying large numbers of taxa.

The overall computational results are being shown with respect to the evolutionary distances along both the distance-based methods. Thus, the user can analyze the different computation steps involved for the final assessment.

5.2 Future Scope

Following improvements and further enhancements regarding the developed analysis of phylogenetic trees can be made:

- The model can be extended to analyze and to take a comparison with one tree building method with others.
- Phylogenetic structure prediction can be incorporated to further enhance the model.
- A new tool can be proposed for the phylogenetic tree to give one and only one optimal result without any error.
- Phylogenetic networks can be constructed from the new tool designed.

- Neighbor Joining can be extended to enhance further to give optimal result.

6. SCREEN SHOTS

Figure 6.1: Input of accession no's of five species and specifying the type of distance method to compute distances

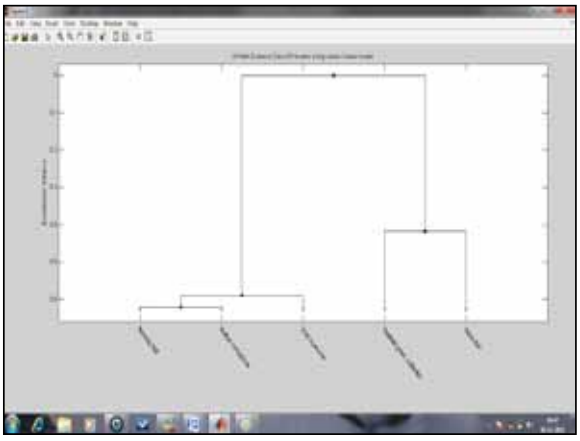

Figure 6.2: Displaying the tree when clicked to UPGMA method.

Figure 6.3: Specifying the second method to build a phylogenetic tree.

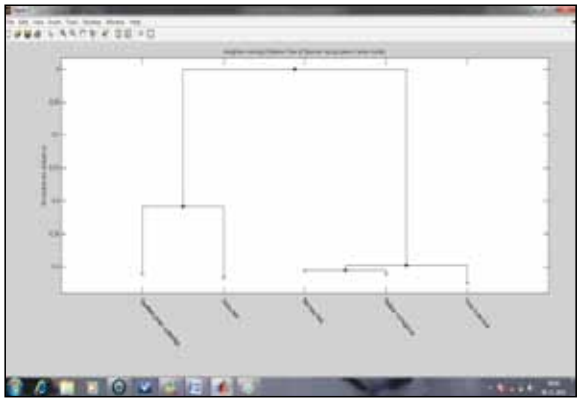

Figure 6.4: Displaying the tree when clicked to Neighbor-Joining method.

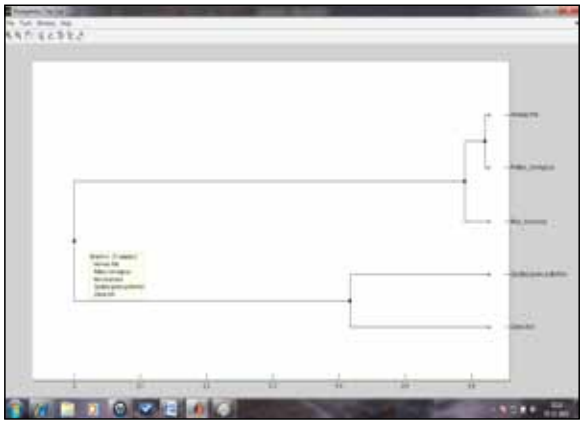

Figure 6.7: Displaying the tree with information

REFERENCES

[1] Nakhleh, L. (2010), "A Metric on the Space of Reduced Phylogenetic Networks", IEEE/ACM Transactions on computational biology and bioinformatics, vol. 7, no. 2, pp 1-5 | [2] Huson, D., H. (2009), "Drawing Rooted Phylogenetic Networks", IEEE/ACM Transactions on computational biology and bioinformatics, vol. 6, no. 1, pp 1-7. | [3] Cardona, G., Llabre´s, M., Rossello´, F. and Valiente, G. (2009), "On Nakhleh's Metric for Reduced Phylogenetic Networks", IEEE/ACM Transactions on computational biology and bioinformatics, vol. 6, no. 4, pp 1-10. | [4] Pevsner J., (2009),"Bioinformatics and Functional Genomics", A john wiley & sons, Inc., Publication, pp 215-221. [5] Xiong J., (2006), "Essential Bioinformatics", United States Of America, Cambridge University Press, New York., pp 3-8, 31-32, 63, 127-133. | [6] Molad, O., Altman Y. (2001), "Bioinformatics Databases and Tools – Introduction", Algorithms for Molecular Biology, Tel Aviv Univ., pp 1-31. | [7] Rastogi, S., C., Mendiratta, N., Rastogi, P. (2007), "Allignment of Multiple Sequences and Phylogenetic Analysis", Bioinformatics- Methods and Applications, New Delhi, pp 93-120. | [8] Chen, X., Wang, M. and Zhang, H. (2011), "The use for classification trees for bioinformatics", John Wiley & Sons, Inc. WIREs Data Mining Knowl Discov , pp 55–63. | [9] Stoye, J. and Wittler, R. (2009), "A unified approach for reconstructing ancient gene clusters", IEEE Transactions on computational biology and bioinformatics, vol. 6, no. 3, pp 1-14. [10] Mandoiu, I., Narasimhan, G., Pan, Y. and Zhang, Y. (2010), "Guest Editors' Introduction to the Special Section on Bioinformatics Research and Applications", IEEE Transactions on computational biology and bioinformatics, vol. 7, no. 4, pp 1-14.

Sara Publishing Academy
Indian Journal Of Applied Research
Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College,
New Congres Bhavan, Paldi, Ahmedabad-380006.
Contact.: +91-9824097643 E-mail : editor@ijar.in

Printed at Unique Offset, Novatsing Rupam Estate, Opp. Abhay Estate, Tavdipura, Shahibaug, Ahmedabad