A Novel Genetic Algorithm based Approach for Optimization of Distance Matrix for Phylogenetic Tree Construction

Mridu Gupta,
Department Of Computer Science and Engineering
PEC University of Technology, Chandigarh

Shailendra Singh
Department Of Computer Science and Engineering
PEC University of Technology, Chandigarh

ABSTRACT

Phylogenies are useful for organizing knowledge of biological diversity, for structuring classifications, and for providing knowledge of events that occurred during evolution. Different phylogenetic reconstruction techniques are available. In this paper Distance based technique is used. Distance measure is an important issue in phylogenetic analysis. Traditional approaches are time-consuming due to the fact that they require multiple sequence alignment, while the K-tuple distance is easy to compute and has been used in phylogenetic tree reconstruction. Based on this K- tuple distance, a genetic algorithm is proposed to find a new F-tuple distance measure which takes into account the position of occurrence of tuples and instead of considering difference, similarities between the sequences are considered. The K-tuple distance approach is not effective for set of sequences which are almost identical where as F-tuple distance is useful for constructing phylogenetic tree for set of identical sequences. This novel approach is capable of efficiently building phylogenetic trees and is less computational intensive.

Index Terms

Distance Method, Distance Matrix, Phylogenetics, Phylogenetic Tree, K-tuple distance, Genetic Algorithm, F-tuple distance.

1. INTRODUCTION

Bioinformatics is blend of biology and computers, means it is an applied science where computational theories and technologies are used in order to process, relate and derive predictions and inferences from data obtained in molecular biology. Bioinformatics' goal is to understand and analyze the information control and flow within different organisms. There is a synergic interaction between computer science and biology, each with its own richness and limitations [1].With advances in gene bank database inferring phylogenetic have become one of major task in bioinformatics in recent era [2].

Phylogenetics is the field at the interface of biology, mathematics, and computer science, which studies the reconstruction of plausible evolutionary scenarios of group of organisms [3]. Constructing phylogenetic trees is a crucial step for biologists to find out how today's species are related to one another in terms of common ancestors [4]. Phylogenetics is useful for organizing knowledge of biological diversity, for structuring classifications, and for providing knowledge of events that occurred during evolution [2].

Methods for construction of phylogenetic tree is basically divided in two categories distance based methods and character based methods. Character-based methods are based on discrete characters from molecular sequences from individual taxa. On the other hand, distance-based methods are based on the distance, the degrees of differences between pairs of sequences. Character based methods include Parsimony, Maximum Likelihood and Bayesian methods [5] [6], and Distance based methods include UPGMA, NJ, FM and Minimum Evolution. Distance based methods are faster than character based methods and is more useful for analyzing large data sets for thousands of sequences and for bootstrapping. Distance methods are less computational, optimality criteria can be used to calculate distance matrix so as to build a much reliable tree.

The paper presents a novel technique to optimize the distance matrix to construct a phylogenetic tree using genetic algorithms and is divided into two sections. The first section includes basics of phylogenetics tree, followed by the details on Distance based method, and genetic algorithm. The last section includes the use of genetic algorithm approach to find optimized distance matrix for the construction of phylogenetic tree.

2. PHYLOGENETICS BASICS

The similarity of biological functions and molecular mechanisms in living organisms strongly suggests that species descended from a common ancestor [7]. Phylogenetics uses the structure and function of molecules and how they change over time to infer these evolutionary relationships and construct phylogenetic tree. Phylogenetic trees (Figure 1) are composed of branches, also known as edges that connect and terminate at nodes.

Branches and nodes can be internal or external (terminal). The terminal nodes at the tips of trees represent operational taxonomic units (OTUs). OTUs correspond to the molecular sequences or taxa (species) from which the tree was inferred. Internal nodes represent the last common ancestor (LCA) to

all nodes that arise from that point. Trees can be made of a single gene from many taxa (a species tree) or multi-gene families (gene trees). [8][10].

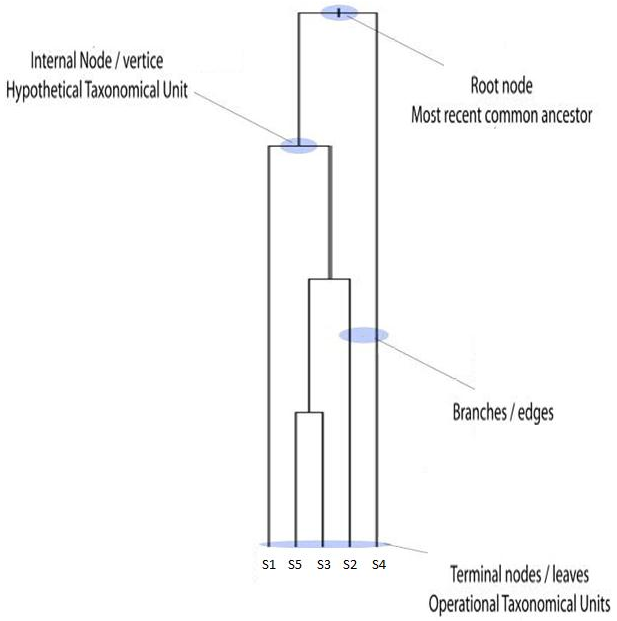

Figure 1Phylogenetic Tree

3. DISTANCE METHOD

Distance-based methods are based on the distance, the degrees of differences between pairs of sequences. Such distance will be used to construct the distance matrix that is the matrix of pair wise "distances" between sequences that approximate evolutionary distance [11][2]. One of challenges' in using distance matrices with distance methods to build phylogenetic tree is the building of the matrix [9].Distance matrix is an n*n matrix where n is the no Sequences. Each row corresponds to a single sequence and every column contains distance between two sequences. Given a collection of n sequences, with distance d defined between any pair of sequences, the following matrix

$$\begin{bmatrix} d_{11} & \cdots & d_{1n} \\ d_{21} & d_{22} & d_{2n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{bmatrix}$$

is called the distance matrix, where d_{ij} is the distance between the i_{th} and j_{th} sequences.

There are various models for calculating distance between sequences like p-Distance, Jukes-Cantor [11], alignment-score. One commonly used distance in bioinformatics is to define the distance between any two nodes in an evolutionary phylogenetic tree based on observation as the sum of all weights along the path from one node to the other. Since there exists no circuit in a tree, this distance is well-defined and it is additive. In general, for any pair A,B of sequences, other than giving the evolutionary distance between two species based on observation, it is defined the distance D_{AB} between the two sequences as the fraction f of sites u where residuesX_u^A andX_u^B differ.[9] There are also some other ways to find distance like Snel et al.[12] defined gene contents a distance measure in genome phylogeny, in which the similarity between two species is defined as the number of genes they have in common divided by their total number of genes. The general idea was further extended to identify evolutionary history and protein functionality [14].Fitz-Gibbon and House brought out a similar approach [13]. Lin and Gerstein [15] constructed phylogenetic trees based on the occurrence of particular molecular features: presence or absence of either folds or orthologs throughout the whole genome. Takaiaetal. [16] used whole proteome comparisons in deriving genome phylogeny, considering the overall similarity and the predicted gene product content of each organism.

Distance-based methods tend to be in polynomial time and are quite fast in practice [17]. There are two different algorithms in distance based method, the cluster-based and the optimality-based. The cluster-based method algorithms build a phylogenetic tree based on a distance matrix starting from the most similar sequence pairs. The algorithms of cluster-based include unweighted pair group method using arithmetic average (UPGMA), neighbor joining and weighbor[18]. The optimality-based method algorithms compare numerous different tree topologies and select the one which is believed to best fit between computed distances in the trees and the desired evolutionary distances which often referred as actual evolutionary distances. Algorithms of optimality based include Fitch-Margoliash and minimum evolution. [9] These approaches have potential for computational simplicity and therefore speed [19].

4. GENETIC ALGORITHM

Genetic algorithm [21] [22] is based on the principle of evolution and natural genetics for the optimization and randomized search techniques or it can be said that genetic algorithm is a biologically inspired technology. GAs is efficient, adaptive, and robust search processes, producing near optimal solutions, and has a large degree of implicit parallelism. Therefore, the application of GAs for solving certain problems of bioinformatics, which need optimization of computation requirements, and robust, fast and close approximate solutions, appears to be appropriate and natural. Moreover, the errors generated in experiments with bioinformatics data can be handled with the robust characteristics of GAs. To some extent, such errors may be regarded as contributing to genetic diversity, a desirable property. The problem of integrating GAs and bioinformatics constitutes a new research area.

Genetic Algorithm is executed iteratively on a set of coded solutions, called population, with three basic operators: selection/reproduction, crossover, and mutation. They use

only the payoff (objective function) information and probabilistic transition rules for moving to the next iteration. Of all the evolutionarily inspired approaches, GAs seem particularly suited to implementation using DNA, protein, and other bioinformatics tasks [23]. This is because GA is generally based on manipulating populations of bit-strings using both crossover and point-wise mutation.

Advantages

1. Several tasks in bioinformatics involve optimization of different criteria (such as energy, alignment score, and overlap strength), thereby making the application of GAs more natural and appropriate.

2. Problems of bioinformatics seldom need the exact optimum solution; rather, they require robust, fast, and close approximate solutions, which GAs are known to provide efficiently.

3. GAs can process, in parallel, populations billions times larger than is usual for conventional computation. The usual expectation is that larger populations can sustain larger ranges of genetic variation, and thus can generate high-fitness individuals in fewer generations.

4. Laboratory operations on DNA inherently involve errors. These are more tolerable in executing evolutionary algorithms than in executing deterministic algorithms. (To some extent, errors may be regarded as contributing to genetic diversity—a desirable property.)

There are various applications of Genetic Algorithm such as alignment and comparison of DNA, RNA, and protein sequences [24][25][26],Gene mappings in chromosomes, Gene finding and promoter identification from DNA sequences[27], Interpretation of gene expression and micro array data [28], Gene regulatory network identification[29], RNA structure prediction[30], DNA structure prediction[31], Protein structure prediction and clustering [32],Molecular design and molecular docking [17] etc.

5. PROPOSED METHODOLOGY

In this section it is proposed to use Genetic algorithm for generating an optimized distance matrix. The current paradigm in phylogenetic tree reconstruction is to start from a set of sequences, build multiple sequence alignment (MSA), and then based on the MSA score distance matrix is generated and then build a tree using one or several methods[20]. In proposed method there is no need of multiple sequence alignment; distance between the sequences is calculated as F-tuple distance which is an improvement of k-tuple distance. The k-tuple distance refers to the total of the differences, over all possible k tuples, between the sequences, for any given length k [20].k-tuple distance measure was useful in the reconstruction of trees from highly divergent sequences but the k-tuple distance may have low resolution when a set of sequences are almost identical. In this paper it is suggested that position of the tuples can also be considered and rather than calculating the frequency of occurrence of different tuples, frequency of similar tuples should be calculated and there evolution can be considered with the help of genetic algorithm. The process of calculating the F-tuple distance and use of genetic algorithm is for generating optimized distance matrix to construct a phylogenetic tree is divided in three sections as shown in Figure 2.

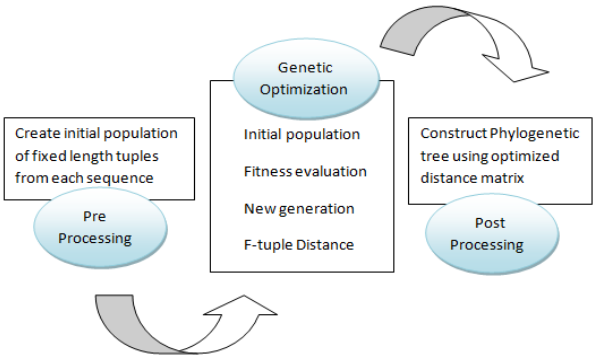

Figure 2 Optimization Process

5.1. Pre-Processing

In the Pre-Processing phase sequences from the database are selected that can be of various lengths. A sequence, S, of length l, is defined as a linear succession of l symbols from a finite alphabet A, of length n. A segment of F symbols, with F≤ l, is designated an F-tuple or F-word. Obviously, there are a total of n*F possible F-tuples for the alphabet A. Each sequence from database is divided into fixed length tuples to form a pool of initial tuples.

This initial population has to be encoded first before sent for genetic optimization so that it is easily processed further. For bitstring and doublestring population type, there is need to encode the population. In the current scenario, the population is bitstring, so it needs encoding. ASCII encoding can be done like ‘a’ is encoded as 97;’t’ is encoded as 74; ‘g’ is encoded as 103; ‘c’ is encoded as 119.

5.2. Genetic Optimization

First step in genetic optimization is to select individuals from initial population and then calculate fitness of each tuple depending upon the frequency and position of the tuple.The number of occurrences of an F-tuple w i.e.$N_w$is counted by moving a sliding window of length F over the sequence with 1 basepair overlapping step size. The frequency, f_w,of this word is obtained from N_w by dividing the total number of tuples.

Then the fittest tuples are selected and rest is rejected. New population is generated by crossover and mutation functions. This new generation is again evaluated for fitness and again reproduction phase occur until the fitness value is maximum or generation is maximum.

At the end of this phase the fittest tuple is obtained. The frequency and position information of this tuple in every sequence is used to find distance that is F-tuple distance

between every pair of sequences. These obtained distances are then used to construct distance matrix.

5.3. Post Processing

In this phase many different methods can be used to build a tree from the optimized distance matrix for a set of DNA sequences, such as Minimum Evolution (ME), Neighbor-Joining (NJ), UPGMA and Fitch and Margoliash (FM). These methods used the distance matrix obtained in the genetic optimization phase.In the last step phylogenetic tree for the database is displayed.

The proposed methodology is applied on mtDNA sequences of 20 different mammals. The sequences are obtained from the database of NCBI site. The trees obtained for these sequences are more reliable as it is obtained from an optimized distance matrix. This tree can be compared from the tree obtained by traditional methods and then reliability of tree can be evaluated using different approaches.

6. CONCLUSIONS

The proposed methodology gives a new distance measure for sequences based on K-tuple, a simple three step algorithm is proposed for finding new F-tuple distance that is to find an optimized distance matrix for the construction of phylogenetic tree. This approach takes into account similar tuples as well as the position of occurrence and it effectively highlights the unseen information behind the relatedness of sequences. It is also effective for constructing phylogenetic tree for a set of sequences which are almost identical. The genetic programming uses the idea of natural selection and global optimization so the distance obtained does not suffer from some evolutionary events. The new approach can discriminate between different species and is effective for building phylogenetic trees from genome sequences. In addition F-tuple distance has low computational complexity and also genetic algorithms are very fast and robust.

7. REFERENCES

[1] AnkitaJiwan, Shailendra Singh, 2012. "A Review on RNA Pseudoknot Structure Prediction Techniques, International Conference on Computing",Electronics and Electrical Technologies [ICCEET], IEEE sponsored, 975-978.

[2] Mridu Gupta, Shailendra Singh, 2012. "Computational Approaches for Phylogenetic Tree Construction: A Review", International Conference on Recent Technology 2012 [ICORT].

[3] Leo van Iersel, Judith Keijsper, Steven Kelk, LeenStougie, Ferry Hagen, and TeunBoekhout, "Constructing Level-2 Phylogenetic Networksfrom Triplets", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6(4),2009.

[4] Chen Yang and Sami Khuri, "PTC: An Interactive Tool for Phylogenetic Tree Construction", Journal of Computational Systems Bioinformatics (CSB'03), 2003.

[5] B. Larget and D. Simon, "Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees", Journal of Mol Biology Evolution 16, 1999, 750-759.

[6] Anupam Bhattacharjee, Kazi Zakia Sultana and Zalia Shams, 2006. "Dynamic and Parallel Approaches to Optimal Evolutionary Tree Construction", Ieee/Ccece/Ccgei, Ottawa.

[7] Andreas D. Baxevanis and B. F. Francis Ouellette, 2001. Bioinformatics a Practical Guide to the Analysis of Genes and Proteins. Willey inter science publication, second edition.

[8] Baldauf SL, "Phylogeny for the faint of heart: a tutorial", Journal of Trends in Genetics 19 (6), 2003, 345-351.

[9] Chuang Peng, 2007. Distance Based Methods in Phylogenetic Tree Construction.

[10] Hall, BG, 2004. Phylogenetic Trees Made Easy: A How-To Manual, 2nd ed. Sinauer Associates, Inc.: Sunderland.

[11] T. Jukes and C. Cantor, 1969. Evolution of protein molecules, Mammalian Protein Metabolism (ed.HN Munro), New York: Academic Press, 21-32.

[12] B. Snel, P. Bork and M. A. Huynen, "Genome phylogeny based on gene content", Journal of Nal.Genet. 21, 1999, 108-110.

[13] S. T. Fitz-Gibbon and C. H. House, "Whole genome-based phylogenetic analysis of free-living microorganism", Journal of Nucleic Acidy Res. 27, 1999, 4218-4222, (1999).

[14] B. Snel. P. Bork and M. A. Huynen, "Genomes in flux: the evolution of archival and proteobacterial gene content", Journal of Genome Res. 12, 2002, 17-25.

[15] J. Lin and M. Gerstein, "Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels", Journal of Genome Res. 10, 2000, 808-818.

[16] F. Tekaia, A. Lazacano and **B.**Dujon, **"**The genomic tree as revealed from whole protein comparisons", Journal of Genome Res 9, 1999, 550-557.

[17] J. M. Yang and C. Y. Kao, "A family competition evolutionary Algorithm for automated docking of flexible ligands to proteins", IEEE Trans. Inf. Technol. Biomedecine. 4(3), 2000, 225–237.

[18] N. Saitou and M. Nei, "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees", Journal of Mol. Biology Evolution 4(4), 1987, 406-425.

[19] C. Jill Harrison and Jane A. Langdale, "A Step By Step Guide to Phylogeny Reconstruction", Plant Journal 45, 2006, 561-572.

[20] Kuan Yang and Liqing Zhang,, " Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction", Journal of Nucleic Acids Research 36(5), 2008.

[21] D. Bhandari, C. A. Murthy, and S. K. Pal, "Genetic algorithm with elitist model and its convergence", Int. J. Pattern Rcognit. Artif. Intell. 10(6), 1996, 731–747.

[22] L. B. Booker, D. E. Goldberg and J. H. Holland, "Classifier systems and genetic algorithms", Artif. Intell, 40(1–3), 1989, 235–282.

[23] S. B. Needleman and C. D. Wunsch,"A general method applicable to the search for similarities in the amino acid sequence of two proteins", J. Mol. Biology 48,1970, 443–453.

[24] T. F. Smith and M. S. Waterman, 2001.Identification of common Informatics: Edmonton. AB, Canada: IMIA, 83– 100.

[25] T. Murata and H. Ishibuchi, "Positive and negative combination effects of crossover and mutation operators in sequencing problems", Journal of Evol. Computation 20, 1996, 170–175.

[26] H. D. Nguyen, I. Yoshihara, K. Yamamori and M. Yasunaga, 2002. A parallel hybrid genetic algorithm for multiple protein sequence alignment, In Proc. Congress Evolutionary Computation, 309–314.

[27] V. G. Levitsky and A. V. Katokhin, "Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis", Journal of Silico Biology 3(1–2), 2003, 81–87.

[28] H. K. Tsai, J. M. Yang, Y. F. Tsai and C. Y. Kao ,"An evolutionary approach for gene expression patterns", IEEE Trans. Inf. Technol. Biomedicine 8(2), 2004, 69–78.

[29] Leping Li, Yu Liang and Robert L. Bass, "GAPWM: a genetic algorithm method for optimizing a position weight matrix", Journal of Bioinformatics 23(10), 2007, 1188-1194.

[30] K. C. Wiese and E. Glen, "A permutation-based genetic algorithm for the RNA folding problem: A critical look at selection strategies, crossover operators, and representation issues", Journal of Biosystems 72(1–2), 2003, 29– 41.

[31] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Savino and A. Scipioni, "A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability", Journal of Biophysics 79(2), 2000, 601–613.

[32] S. Schulze-Kremer, "Genetic algorithms and protein folding. Methods in molecular biology", Journal of Protein Structure Prediction: Methods and Protocols 143, 2000, 175–222.