International Journal of Scientific Research in Multidisciplinary Studies
Vol.3, Issue.6, pp. pp.54-60, June (2017)

Review Paper
E-ISSN: 2454-9312
P-ISSN: 2454-6143

Optimal Multiple Sequence Alignment

Pankaj Bhanbri[1*], O.P. Gupta[2]

[1]Department of Computer Science and Engineering, I.K.G. Punjab Technical University, Kapurthala, India
[2]School of Elect. Engineering and Information Technology, Punjab Agriculture University, Ludhiana, India

Corresponding Author e-mail: pkbhambri@gmail.com

Available online at: www.isroset.org

Received 16th May 2017, Revised 28th May 2017, Accepted 24th Jun 2017, Online 30th Jun 2017

Abstract--The research in bioinformatics has accumulated large amount of data. The biological data is available in different formats and is comparatively more complex. The term bioinformatics is related to study of bio-molecules information. The informatics techniques are applied to understand and organize the information associated with these molecules. Bioinformatics offers different knowledge discovery concepts for molecular biology and has many practical applications. DNA sequence alignment is one of the applications of the bioinformatics. Multiple Sequence Alignment is used to align the biological sequences along a column. As the process generates distances of multiple alignments among the pairs of different species, phylogenetic tree is being formulated. Multiple sequence alignment arranges the sequences in such a way that evolutionarily equivalent positions across all sequences are matched. Alignment of Substitutions made into two categories: Jukes Cantor Method and Kimura's Method.

Keywords-- Bioinformatics, Phylogenetic tree, Juke's Cantor model, Kimura's 2-parameter model

I. Introduction

With the help of computer tools, biological information is gathered and analyzed. It is the science of managing, mining and interpreting information from biological sequences and structures. It deals with algorithms, databases and information systems, data mining, image processing and improving & discovering new models of computation. This scientific field deals with the computational management of all kinds of biological information. This information can be on genes and their products, whole organisms or even ecological systems. Mainly bioinformatics involves merger of different applications of mathematical, statistical, computational or molecular biological tools to gather different types of information and by analyzing them, researches can be carried out. 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. Java, XML, Perl, C, C++, SQL and MATLAB are the programming languages popularly used in this field. The tools of bioinformatics include computer programs that help to reveal fundamental mechanisms. The biological problems related to the structure and function of macromolecules, disease processes, and evolution are contained in the tools. The applications of the tools is being categorised into sequence analysis, structure analysis, and function analysis. These three aspects of bioinformatics often interact to produce integrated and good results. Bioinformatics includes the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins. Extract DNA and Protein Sequences from Database. Whole of the database is being searched to compare the DNA's in a pair-wise fashion. DNA is transcribed to RNA which is further translated to proteins. This make possible to analyze the behaviour of the cell. [1,2,3,4,5,6].

II. Applications, Goals, Scope and Challenges of Bioinformatics

The goals of bioinformatics are: 1) To understand a living cell and it's functioning at the molecular level. Analysis of Molecular sequences and structural data that can be used for the functioning of the cell. 2) To solve formal and practical problems arising in analysis of biological data. 3) Mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences, compare them and creating and visualizing 3-D models of protein structures.

Two subfields consists of Bioinformatics, that is, 1) The development of computational tools and databases. 2) The application of these tools and databases in generating biological knowledge to better understand living systems. These two subfields are complementary to each other. The application of the tools is being used in construction and crating of biological databases. The analyses of biological data often generate new problems that use to develop new and better computational tools.

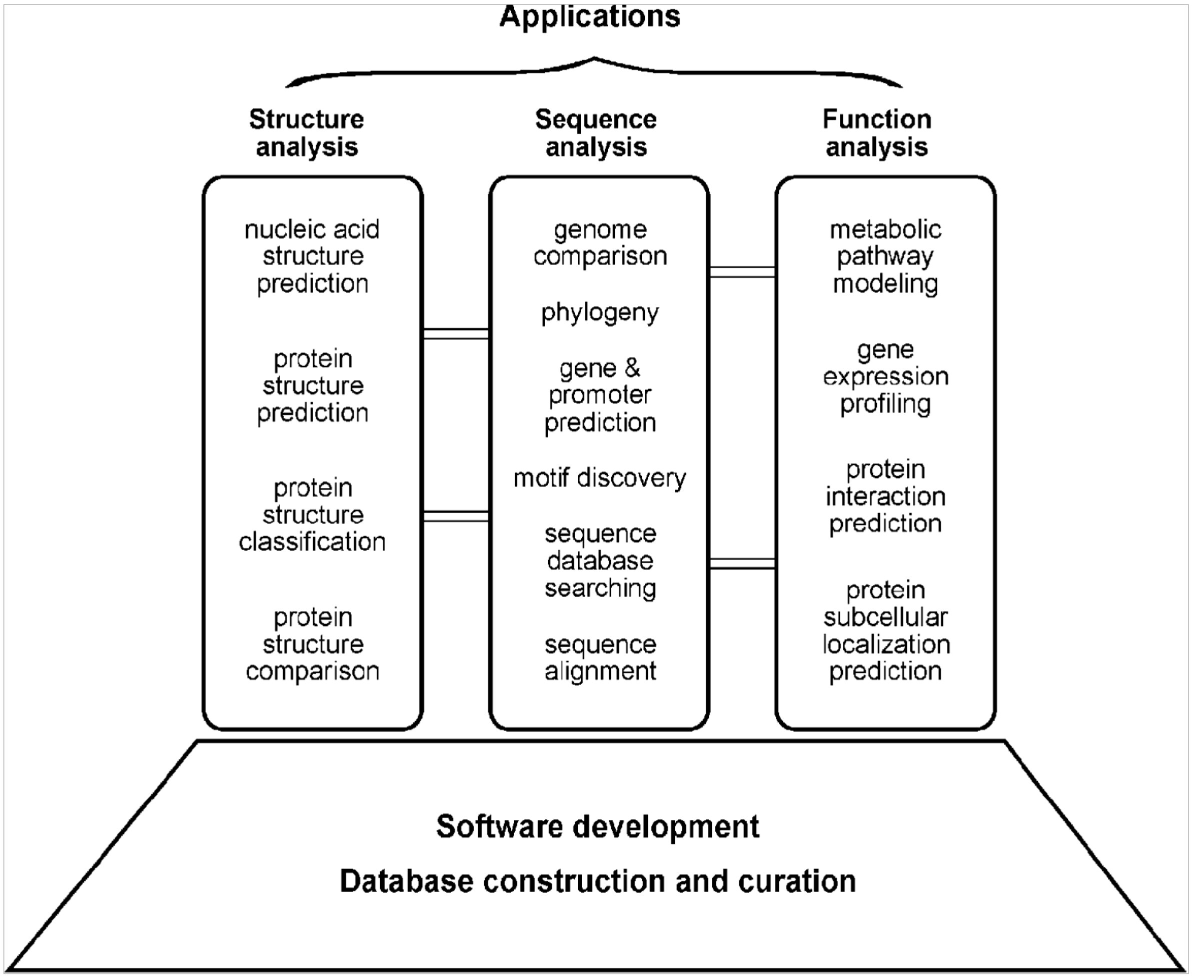

Applications
Structure analysis
Sequence analysis
Function analysis
nucleic acid structure prediction
protein structure prediction
protein structure classification
protein structure comparison
genome comparison
phylogeny
gene & promoter prediction
motif discovery
sequence database searching
sequence alignment
metabolic pathway modeling
gene expression profiling
protein interaction prediction
protein subcellular localization prediction
Software development
Database construction and curation

Figure 1. Overview of various subfields of bioinformatics [1,2,3,4,5,6]

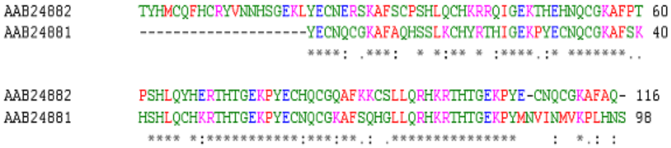
The areas shown in Figure 1 consists of sequence analysis that include sequence alignment, sequence database searching, motif and pattern discovery, gene and promoter finding, reconstruction of evolutionary relationships, and genome assembly and comparison. Structural analyses include protein and nucleic acid structure analysis, comparison, classification, and prediction. The functional analyses include gene expression profiling, protein–protein interaction prediction, protein sub-cellular localization prediction, metabolic pathway reconstruction, and simulation.

The challenges in bioinformatics include the basic requirements is to meet the computation of results. With the enormous amounts of data, the challenge of bioinformatics is to store, manage, analyze and interpret the sequence data. For beneficial of researchers there should be an easy access to the information needed. Also, there should be a method for extracting only the information needed to answer a specific biological question. With advancing the technology of hardware, the cost of data storage is decreasing. The result is overflow of data but not underflow of information from raw data. Thus there is need for new techniques and tools to be developed. These tools are useful for handling intelligently and automatically transferring the data into useful knowledge.

III. Alignment of Sequences

In bio-informatics, sequence alignment is a way of arranging the primary sequences of DNA, RNA and proteins to identify regions of similarity. These regions of similarity may be a consequence of functional, structural or evolutionary relationships between the sequences. This is used to find the best-matching sequences. A sequence alignment is a scheme of writing one sequence on top of another where the residues in one position are deemed to have a common evolutionary origin. Aligned sequences of nucleotides or amino acid residues are represented as rows within a matrix. A letter or a stretch of letters may be paired up with dashes in the other sequence to signify such an insertion or deletion. Gaps are inserted between the residues so that residues with identical or similar characters are aligned in successive columns. If the same letter occurs in both sequences then the position is conserved in evolution. If the letters differ then take one residue or neither from two derives from an ancestral letter.

Homologous sequences may have different lengths. *Homologous sequences* mean two or more sequences have a common ancestor.



```
AAB24882      TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881      --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                  ****: .***:  * *:** * :****.:* *******..

AAB24882      PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881      HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
               **** *:***********:***:**.: .**************    :  *.: :
```

A sequence alignment identified on the left by GenBank accession number. Key: Single Letters: amino acids. Red: small, hydrophobic, aromatic, not Y. Blue: acidic. Magenta: basic. Green: hydroxyl, amine, amide, basic. Gray: others. “*”: identical. “:”: conserved substitutions (same colour group). “.”: semi-conserved substitution (similar shapes). The absence of substitutions, or the presence of only very conservation substitutions in a particular region of the sequence, suggest that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than amino acids, the conservation of base pairs can indicate a similar functional or structural role.

Sequence alignment can be used for non-biological sequences, such as natural language or in financial data. Alignment can be done locally or globally. Global alignment need to use gaps while local alignments can avoid them, aligning regions between gaps.

Aligning two sequences can allow one to detect their overlap or to notice that one sequence is a part of the other or that the two sequences share a subsequence. Instead of two sequences, one can also align many sequences or match a sequence against a DNA, RNA, or protein database. Multiple alignments of RNA or amino acid sequences in proteins allow one to infer their secondary and tertiary structures as well as functionally important sites in proteins.

IV. Alignment Methods

Very short or very similar sequences can be aligned by hand. However, problem occur in alignment of lengthy, highly, variable or extremely numerous sequences that cannot be aligned by human effort and required computational approaches to align the sequences. Two categories are: *global alignments* and *local alignments*.

Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set and of roughly equal size. Global alignment get the maximum match between the sequences as it assume that the two sequences are similar. This alignment attempts to match the two sequences from the end to the end even though if they are different in some parts. Calculating a global alignment is a form of global optimization that “forces” the alignment to span the entire length of all query sequences. A general alignment technique is called the Needleman-Wunsch algorithm and is based on dynamic programming.

```
 NLGPSTKDFGKISESREFDNQ
   |        | | | |      |
QLNQLERSFGKINMRLEDALV
```

By contrast, Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. Local alignment searches for the part of the two sequences that match well. There is no attempt to “force” entire sequences into an alignment, just those parts that appear to have good similarity, according to some criterion are considered. The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments.

```
NLGPSTKDDFGKILGPSTKDDQ
| | | |
QNQLERSSNFGKINQLERSSNN
```

V. Types of Alignments

Sequence Alignment considers two types of alignments, *Pair-wise Sequence Alignment* and *Multiple Sequence Alignment [1,2,3,4,5,6]*.

Pairwise sequence alignment is concerned with comparing two DNA or amino acid sequences to find the best matching by global and local “optimum alignment” of the two sequences. Pairwise alignment can only be used between two sequences at a time and they calculate the sequences effectively. Based on differences between the two sequences, one can calculate the "cost" of aligning the two sequences by using replacements, deletions and insertions, and assign a similarity score. A particular application of pairwise sequence alignment is quickly searching large DNA and protein databases for matches to a query sequence. There are three methods of producing pairwise alignments. These are 1) Dot Matrix Methods, 2) Dynamic Programming and 3) Word Methods

Although each method has its individual strengths and weaknesses, all three methods have difficulty with highly repetitive sequences of low information content (low meaning of information) – especially the number of repetitions differ in the two sequences to be aligned. The best utility of given pairwise alignment has ‘maximum unique match’ (MUM) or longer subsequence occur in both query sequence. A longer MUM sequence reflects closer relatedness. Popular heuristic algorithms, such as *FASTA* or

BLAST families are much faster than algorithms based on dynamic programming. Heuristic method may not be as accurate as dynamic programming, but it is fast and effective computational method.

Multiple sequence alignment is the extension of pair-wise sequence alignment. It is used to align multiple related sequences to figure out the optimal matching of the sequences. Multiple sequence alignments are very powerful as two sequences that may not align well to each other can be aligned via their relationship to a third sequence of any family. Multiple sequence alignment aims to find similarities between many sequences. Multiple alignments can use more than two sequences at a time. A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequence, generally protein, DNA or RNA. In a multiple sequence alignment, homologous residues among a set of sequences are aligned together in columns. These homologous residues can be defined in both the structural and evolutionary sense. It also gives more biological information than many pair-wise sequence alignments. Like pair-wise alignment, multiple sequence alignment can be categorized in two ways locally or globally. It has basically an essential feature to do phylogenetic analysis of sequence families.

The most successful MSA solutions are heuristic algorithms with approximate approaches, such as the *CLUSTAL* and *Profile Hidden Markov Models* (*HMMs*). There are three methods of producing multiple alignments. These are as follows, 1) Progressive Alignment Method, 2) Iterative Alignment Method and 3) Block-based Alignment

Both pairwise and multiple sequence alignment algorithms use substitution matrices to score the sequence alignment. In substitution matrices each possible residue substitution is given a score reflecting the probability of such a change. There are two popular protein substitution matrix models: *P*ercent *A*ccepted *M*utation (*PAM*) and *Blo*cks *Su*bstitution *M*atrix (*BLOSUM*).

VI. Phylogenetic Tree Construction

To begin the phylogenetic, we need to understand the term evolution. Evolution means to development of biological form from pre-existing and current existing form through some modifications. The study of evolutionary history of some organisms using tree-like diagrams is known as phylogenetic tree construction or phylogenetic analysis. Molecular data are available in the form of DNA or protein sequences. Molecular data are numerous than other records and easier to obtain. More clear-cut and robust phylogenetic trees can be constructed with the molecular data.

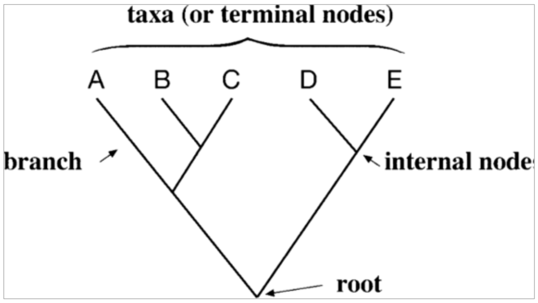


Figure 2. A typical phylogenetic tree showing root, internal nodes, terminal nodes and branches

Phylogenetic trees may be built on the basis of different approaches, which might be divided into *data-oriented* and *model-oriented* methods. Examples of the data-oriented methods are the *distance methods*: the tree is constructed by joining sequences with a small distance between them. Another example is the *maximum parsimony method*: the tree that explains the observed data using the smallest number of substitutions is accepted.

Distance methods are based on creating a distance matrix. Starting from an alignment, pairwise distances are calculated between DNA sequences. From the obtained distance matrix, a phylogenetic tree is calculated with clustering algorithms.

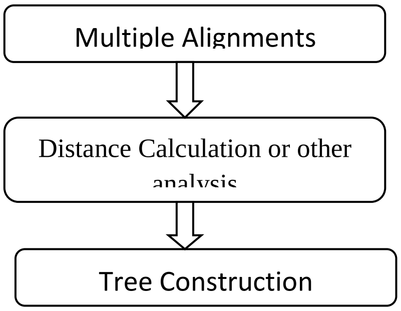


Figure 3. Steps in Phylogenetic Analysis

The *model-based approaches* include the *maximum likelihood method*. In maximum likelihood methods, a probabilistic model of evolution is assumed and its fit to the sequence data maximizes the likelihood of all possible trees. Calculating the likelihoods is computationally intensive, but the method can be extended in several directions, including evolution under selective pressure, which may be helpful in the identification of active sites in proteins.

A. Stages of Phylogenetic Analysis

Molecular phylogenetic analysis can be divided into five stages:

(1) Selection of sequences for analysis,

(2) Multiple sequence alignment of homologous protein or nucleic acid sequences,

(3) Specification of a statistical model of nucleotide or amino acid evolution,
(4) Tree building, and
(5) Tree evaluation.

The first step in phylogenetic analysis is to acquire the Sequence. There are number of choice of DNA, RNA, or protein sequences for molecular phylogeny. We can acquire the sequences from many sources, including the NCBI includes thousands of eurkaryotic protein families. These HomoloGene entries can be viewed as sequences in the fasta format. The results from the BLAST family of proteins can be selected, viewed in Entrez Protein or Entrez Nucleotide, and formatted in the fasta format. The sequences from a large variety of databases can be output in the fasta format.

The second step in phylogenetic analysis is to construct sequence alignment. This is probably the most critical step in the procedure because it establishes positional correspondence in evolution. Only the correct alignment produces correct phylogenetic inference. Incorrect alignment leads to systematic errors in the final tree or even a completely wrong tree. For that reason, it is essential that the sequences are correctly aligned. In addition, there is an automatic approach in improving alignment quality. Rascal and NorMD can help to improve alignment by correcting alignment errors and removing potentially unrelated or highly divergent sequences.

The third step is to count the number of substitutions in an alignment. The proportion of substitutions defines the observed distance between the two sequences. It is used to defining the relatedness of a group of nucleotide (or amino acid) sequences is to align pairs of sequences and count the number of differences. When a mutation is observed as A replaced by C, the nucleotide may have actually undergone a number of intermediate steps to become C, such as A→T→G→C. Similarly, a back mutation could have occurred when a mutated nucleotide reverted back to the original nucleotide. This means that when the same nucleotide is observed, mutations like G→C→G may have actually occurred. Such multiple substitutions and convergence at individual positions obscure the estimation of the true evolutionary distances between sequences. This effect is known as *homoplasy*, which, if not corrected, can lead to the generation of incorrect trees. To correct homoplasy, statistical models are used. These model are called *substitution models* or *evolutionary models.* The substitution models define the methods for correcting multiple substitutions in molecular sequences. The methods are: Juke's Cantor model and Kimura's 2-parameter model.

Jukes–Cantor model is the simplest nucleotide substitution model. For DNA sequences, the model assumes that all nucleotides are substituted with an equal rate. It is also called the *one-parameter model.*

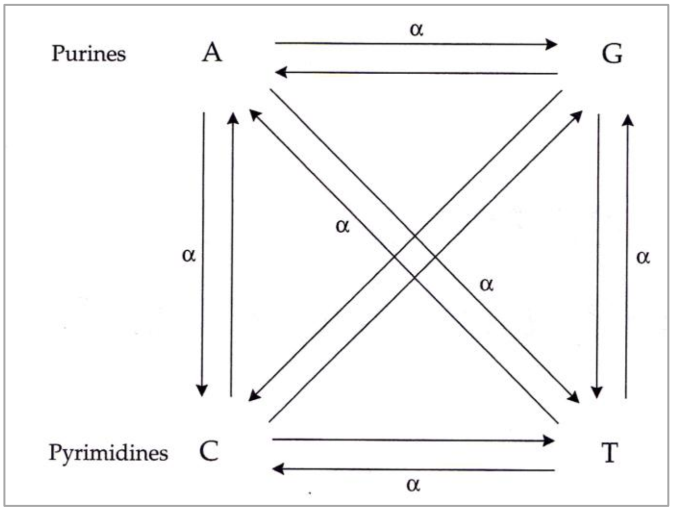


Figure 4. Jukes-Cantor model

This model estimates the evolutionary distance between two sequences. It is used in sufficient statistic for calculated the Jukes-Cantor distance correction, but is not sufficient for the calculation of the evolutionary distance under the more complex models that follow. JC is limited in modeling transitions/transversions variation. Its limitation was overcome by using Kimura's model. The overall rate of substitution for any nucleotides was 3α. The initial probability (P) of site C and estimated substitution (K) was defined by the following equation:

$$P_{c(t)} = 1/4 + (3/4)e^{-4\alpha t}\ K= -3/4 \ln[1-(4/3)(p)]$$

p is fraction of nucleotides.

Kimura's 2-parameter model is a more sophisticated model. For DNA sequences, the model assumes that there are two different substitution rates, one for transition (α) and the other for transversion (β). It is also called the *two-parameter model*. The Kimura model adds one parameter to the Jukes-Cantor model in order to allow the rate of change between purines and pyr-midines (transversions) to be different from changes within purines or within pyr-midines (transitions). This is a rough-and-ready distance formula for approximating PAM distance by simply measuring the fraction of amino acids, that differs between two sequences and computing the distance. According to this model, transitions occur more frequently than transversions.

The rates of substitution in transition were assumed at a uniform rate α and transversions at a different, uniform rate of β. The initial probability (P) of site C and estimated substitution (K) was defined by the following equation:

$$P_{cc(t)} = 1/4 + (1/4)e^{-4\beta t} + (1/2)e^{-2(\alpha+\beta)t}$$

$$K = (1/2)\ln[1/(1 - 2P - Q)] + (1/4)\ln[1/(1 - 2Q)]$$

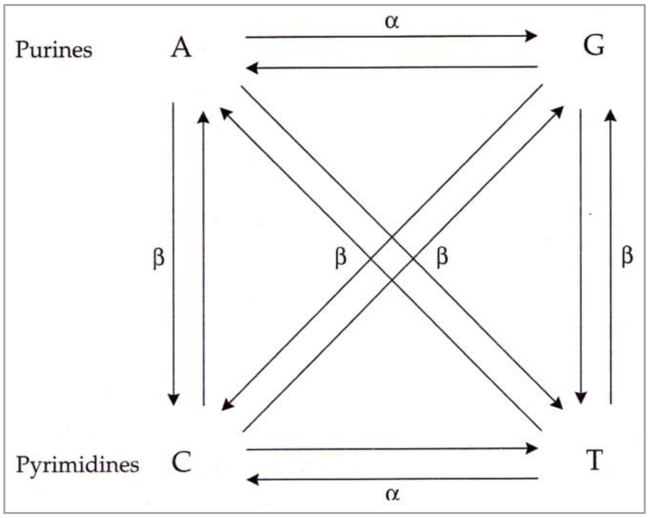


Figure 5. Kimura's model

The next step is to build a phylogenetic tree. We will consider four principal methods of making trees: distance-based methods, maximum parsimony, maximum likelihood, and Bayesian inference. Distance-based methods begin by analyzing pairwise alignments of the sequences and using those distances to infer the relatedness between all the taxa. Maximum parsimony is a character-based method in which columns of residues are analyzed in a multiple sequence alignment to identify the tree with the shortest overall branch lengths that can account for the observed character differences. Maximum likelihood and Bayesian inference are model-based statistical approaches in which the best tree is inferred that can account for the observed data.

After you have constructed a phylogenetic tree, our next and last step is to assess the accuracy. The main criteria are consistency, efficiency, and robustness. The most common approach is bootstrap analysis. Bootstrapping is not a technique to assess the accuracy of a tree. Instead, it describes the robustness of the tree topology.

VII. Problem Formulation

The formulation of the problem reaches its end when the finished product is being obtained. The major problem being faced by today's researchers and biologists is huge amount of raw data and how to store and effectively use this data. The substitutional models are used to count the number of substitutions in an alignment and here presented different methods to find the substitutions. These different methods are used to observe the distances between the pair of sequences of different species. After calculating the distances from the pair of sequences, phylogenetic tree is being formulated. These trees are being formulated from different tree building methods so that obtaining an optimal tree.

VIII. Substitution Model

In bioinformatics, a substitutional model is a method that used to count the number of distances between the sequences of different species. This method occurs after aligning the sequences and alignments are of two types, that is, pairwise sequence alignment and multiple sequence alignment. Here, multiple sequence alignment is used to make a substitution. It is essential that the sequences are correctly aligned. Only the correct alignment produces correct substitutions and produces correct phylogenetic tree. Incorrect alignment leads to incorrect substitution which shows systematic errors in the final tree or even a completely wrong tree. Substitutional model is used to visualize the distances of evolutionary relationships between species. It defines the relatedness of a group of nucleotide sequences. The nucleotide substitution rate matrix is a key parameter of molecular evolution. The substitution models define the methods for correcting multiple substitutions in molecular sequences. For distance-based methods, substitutional models are used to estimate the number of DNA or amino acid changes that occurred in a series of pairwise comparisons of sequences. The distance based correction methods that used in estimation of substitution are Juke's Cantor model and Kimura's 2-parameter and proposed Kimura model.

IX. Phylogenetic Analysis

The development of a biological form from other pre-existing forms or current existing form 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. Each time a branch divides into a smaller branch, it shows the emergence of a new group of organisms. The most popular distance-based methods which are being used for the comparison are the Un-weighted pair group method with arithmetic mean (UPGMA) and Neighbor joining (NJ).

UPGMA Method follows a procedure of clustering 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, fast and has been extensively used in literature. However, it behaves poorly at most cases where the above presumptions are not met.

Neighbor Joining Method (NJ) 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.

References

[1]. S. Vijan and R. Mehra, "Biological Sequence Alignment for Bioinformatics Applications Using MATLAB", Vol. 2, Issue.5, pp.310-315, 2011.

[2]. J. Xiong, "*Essential Bioinformatics",*Cambridge University Press, New York, pp.88-113, 2006.

[3]. S. C. Rastogi, N. Mendiratta, P. Rastogi, "*Allignment of Multiple Sequences and Phylogenetic Analysis*, Bioinformatics- Methods and Applications", Prentice Hall of India, India, pp.1-220, 2007.

[4]. D. E. Krane, M. L. Raymer, "*Fundamental concepts of bioinformatics"*, Pearson Education Publishers, India, pp.23-56, 2006.

[5]. J. Pevsner, *"Bioinformatics and Functional Genomics"*, Wiley-Blackwell Publication, US, pp.107-131, 2009.

[6]. D. W. Mount, "*Bioinformatics- Sequence and Genome Analysis"*, Cold Spring Harbor Laboratory Press, USA, pp.28-73, 2001.

AUTHOR'S PROFILE

Mr. Pankaj Bhambri is Research Scholar at I.K.G. Punjab Technical University Ludhiana. He is pursuing his Ph.D. programme. He is working as Assistant Professor at Guru Nanak Dev Engineering College Ludhiana since 2004. His area of interest is Machine Learning and Bioinformatics.

Dr. O.P. Gupta is working as Associate Prof. cum Head of Information Technology Department of Punjab Agriculture University. He has vast experience while working with Distributed Computing and Bioinformatics. He has published many articles in renowned journal and conferences.