1

Molecular Evolution & Phylogeny: What, When, Why & How?

Pandurang Kolekar[1], Mohan Kale[2] and Urmila Kulkarni-Kale[1]
[1]*Bioinformatics Centre, University of Pune*
[2]*Department of Statistics, University of Pune*
India

1. Introduction

The endeavour for the classification and study of evolution of organisms, pioneered by Linneaus and Darwin on the basis of morphological and behavioural features of organisms, is now being propelled by the availability of molecular data. The field of evolutionary biology has experienced a paradigm shift with the advent of sequencing technologies and availability of molecular sequence data in the public domain databases. The post-genomic era provides unprecedented opportunities to study the process of molecular evolution, which is marked with the changes organisms acquire and inherit. The species are continuously subjected to evolutionary pressures and evolve suitably. These changes are observed in terms of variations in the sequence data that are collected over a period of time. Thus, the molecular sequence data archived in various databases are the snapshots of the evolutionary process and help to decipher the evolutionary relationships of genes/proteins and genomes/proteomes for a group of organisms. It is known that the individual genes may evolve with varying rates and the evolutionary history of a gene may or may not coincide with the evolution of the species as a whole. One should always refrain from discussing the evolutionary relationship between organisms when analyses are performed using limited/partial data. Thorough understanding of the principles and methods of phylogeny help the users not only to use the available software packages in an efficient manner, but also to make appropriate choices of methods of analysis and parameters so that attempts can be made to maximize the gain on huge amount of available sequence data.

As compared to classical phylogeny based on morphological data, molecular phylogeny has distinct advantages, for instance, it is based on sequences (as descrete characters) unlike the morphological data, which is qualitative in nature. While the tree of life is depicted to have three major branches as bacteria, archaea and eukaryotes (it excludes viruses), the trees based on molecular data accounts for the process of evolution of bio-macromolecules (DNA, RNA and protein). The trees generated using molecular data are thus referred to as 'inferred trees', which present a hypothesized version of what might have happened in the process of evolution using the available data and a model. Therefore, many trees can be generated using a dataset and each tree conveys a story of evolution. The two main types of information inherent in any phylogenetic tree are the topology (branching pattern) and the branch lengths.

Before getting into the actual process of molecular phylogeny analysis (MPA), it will be helpful to get familiar with the concepts and terminologies frequently used in MPA.

Phylogenetic tree: A two-dimensional graph depicting nodes and branches that illustrates evolutionary relationships between molecules or organisms.

Nodes: The points that connect branches and usually represent the taxonomic units.

Branches: A branch (also called an edge) connects any two nodes. It is an evolutionary lineage between or at the end of nodes. Branch length represents the number of evolutionary changes that have occurred in between or at the end of nodes. Trees with uniform branch length (cladograms), branch lengths proportional to the changes or distance (phylograms) are derived based on the purpose of analysis.

Operational taxonomic units (OTUs): The known external/terminal nodes in the phylogenetic tree are termed as OTU.

Hypothetical taxonomic units (HTUs): The internal nodes in the phylogenetic tree that are treated as common ancestors to OTUs. An internal node is said to be bifurcating if it has only two immediate descendant lineages or branches. Such trees are also called binary or dichotomous as any dividing branch splits into two daughter branches. A tree is called a 'multifurcating' or 'polytomous' if any of its nodes splits into more than two immediate descendants.

Monophyletic: A group of OTUs that are derived from a single common ancestor containing all the descendents of single common ancestor.

Polyphyletic: A group of OTUs that are derived from more than one common ancestor.

Paraphyletic: A group of OTUs that are derived from a common ancestor but the group doesn't include all the descendents of the most recent common ancestor.

Clade: A monophyletic group of related OTUs containing all the descendants of the common ancestor along with the ancestor itself.

Ingroup: A monophyletic group of all the OTUs that are of primary interest in the phylogenetic study.

Outgroup: One or more OTUs that are phylogenetically outside the ingroup and known to have branched off prior to the taxa included in a study.

Cladogram: The phylogenetic tree with branches having uniform lengths. It only depicts the relationship between OTUs and does not help estimate the extent of divergence.

Phylogram: The phylogenetic tree with branches having variable lengths that are proportional to evolutionary changes.

Species tree: The phylogenetic tree representing the evolutionary pathways of species.

Gene tree: The phylogenetic tree reconstructed using a single gene from each species. The topology of the gene tree may differ from 'species tree' and it may be difficult to reconstruct a species tree from a gene tree.

Unrooted tree: It illustrates the network of relationship of OTUs without the assumption of common ancestry. Most trees generated using molecular data are unrooted and they can be rooted subsequently by identifying an outgroup. Total number of bifurcating unrooted trees can be derived using the equation: $Nu= (2n-5)!/2^{n-3} (n-3)!$

Rooted tree: An unrooted phylogenetic tree can be rooted with outgroup species, as a common ancestor of all ingroup species. It has a defined origin with a unique path to each ingroup species from the root. The total number of bifurcating rooted trees can be calculated using the formula, $Nr= (2n-3)!/2^{n-2} (n-2)!$ (Cavalli-Sforza & Edwards, 1967). Concept of unrooted and rooted trees is illustrated in Fig. 1.

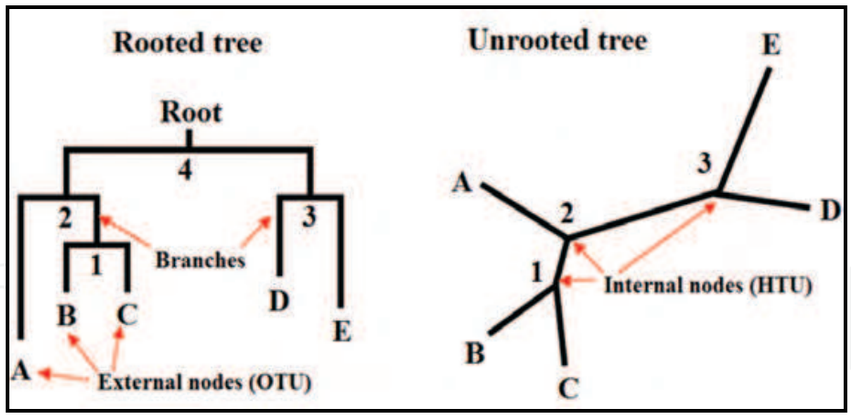


Fig. 1. Sample rooted and unrooted phylogenetic trees drawn using 5 OTUs . The external and internal nodes are labelled with alphabets and Arabic numbers respectively. Note that the rooted and unrooted trees shown here are one of the many possible trees (105 rooted and 15 unrooted) that can be obtained for 5 OTUs.

The MPA typically involves following steps

- Definition of problem and motivation to carry out MPA
- Compilation and curation of homologous sequences of nucleic acids or proteins
- Multiple sequence alignments (MSA)
- Selection of suitable model(s) of evolution
- Reconstruction of phylogenetic tree(s)
- Evaluation of tree topology

A brief account of each of these steps is provided below.

2. Definition of problem and motivation to carry out MPA

Just like any scientific experiment, it is necessary to define the objective of MPA to be carried out using a set of molecular sequences. MPA has found diverse applications, which include classification of organisms, DNA barcoding, subtyping of viruses, study the co-evolution of genes and proteins, estimation of divergence time of species, study of the development of pandemics and pattern of disease transmission, parasite-vector-host relationships etc. The biological investigations where MPA constitute a major part of analyses are listed here. A virus is isolated during an epidemic. Is it a new virus or an isolate of a known one? Can a genotype/serotype be assigned to this isolate just by using the molecular sequence data? A few strains of a bacterium are resistant to a drug and a few are sensitive. What and where are the changes that are responsible for such a property? How do I choose the attenuated strains, amongst available, such that protection will be offered against most of the wild type strains of a given virus? Thus, in short, the objective of the MPA plays a vital role in deciding the strategy for the selection of candidate sequences and adoption of the appropriate phylogenetic methods.

3. Compilation and curation of homologous sequences

The compilation of nucleic acid or protein sequences, appropriate to undertake validation of hypothesis using MPA, from the available resources of sequences is the next step in MPA.

At this stage, it is necessary to collate the dataset consisting of homologous sequences with the appropriate coverage of OTUs and outgroup sequences, if needed. Care should be taken to select the equivalent regions of sequences having comparable lengths (± 30 bases or amino acids) to avoid the subsequent errors associated with incorrect alignments leading to incorrect sampling of dataset, which may result in erroneous tree topology. Length differences of >30 might result in insertion of gaps by the alignment programs, unless the gap opening penalty is suitably modified. Many comprehensive primary and derived databases of nucleic acid and protein sequences are available in public domain, some of which are listed in Table 1. The database issue published by the journal 'Nucleic Acids research' (NAR) in the month of January every year is a useful resource for existing as well as upcoming databases. These databases can be queried using the 'text-based' or 'sequence-based' database searches.

Database	URL	Reference
Nucleotide		
GenBank	http://www.ncbi.nlm.nih.gov/genbank/	Benson et al., 2011
EMBL	http://www.ebi.ac.uk/embl/	Leinonen et al., 2011
DDBJ	http://www.ddbj.nig.ac.jp/	Kaminuma et al., 2011
Protein		
GenPept	http://www.ncbi.nlm.nih.gov/protein	Sayers et al., 2011
Swiss-Prot	http://expasy.org/sprot/	The UniProt Consortium (2011)
UniProt	http://www.uniprot.org/	The UniProt Consortium (2011)
Derived		
RDP	http://rdp.cme.msu.edu/	Cole et al., 2009
HIV	http://www.hiv.lanl.gov/content/index	Kuiken et al., 2009
HCV	http://www.hcvdb.org/	http://www.hcvdb.org/

Table 1. List of some of the commonly used nucleotide, protein and molecule-/species-specific databases.

Text-based queries are supported using search engines viz., Entrez and SRS, which are available at NCBI and EBI respectively. The list of hits returned after the searches needs to be curated very carefully to ensure that the data corresponds to the gene/protein of interest and is devoid of partial sequences. It is advisable to refer to the feature-table section of every entry to ensure that the data is extracted correctly and corresponds to the region of interest. The sequence-based searches involve querying the databases using sequence as a probe and are routinely used to compile a set of homologous sequences. Once the sequences are compiled in FASTA or another format, as per the input requirements of MPA software, the sequences are usually assigned with unique identifiers to facilitate their identification and comparison in the phylogenetic trees. If the sequences posses any ambiguous characters or low complexity regions, they could be carefully removed from sequences as they don't contribute to evolutionary analysis. The presence of such regions might create problems in alignment, as it could lead to equiprobable alternate solutions to 'local alignment' as part of

a global alignment. Such regions possess 'low' information content to favour a tree topology over the other. The inferiority of input dataset interferes with the analysis and interpretation of the MPA. Thus, compilation of well-curated sequences, for the problem at hand, plays a crucial role in MPA.

The concept of homology is central to MPA. Sequences are said to be homologous if they share a common ancestor and are evolutionarily related. Thus, homology is a qualitative description of the relationship and the term %homology has no meaning. However, supporting data for deducing homology comes from the extent of sequence identity and similarity, both of which are quantitative terms and are expressed in terms of percentage.

The homologous sequences are grouped into three types, viz., orthologs (same gene in different species), paralogs (the genes that originated from duplication of an ancestral gene within a species) and xenologs (the genes that have horizontally transferred between the species). The orthologous protein sequences are known to fold into similar three-dimensional shapes and are known to carry out similar functions. For example, haemoglobin alpha in horse and human. The paralogous sequences are copies of the ancestral genes evolving within the species such that nature can implement a modified function. For example haemoglobin alpha and beta in horse. The xenologs and horizontal transfer events are extremely difficult to be proved only on the basis of sequence comparison and additional experimental evidence to support and validate the hypothesis is needed. The concepts of sequence alignments, similarity and homology are extensively reviewed by Phillips (2006).

4. Multiple sequence alignments (MSA)

MSA is one of the most common and critical steps of classical MPA. The objective of MSA is to juxtapose the nucleotide or amino acid residues in the selected dataset of homologous sequences such that residues in the column of MSA could be used to derive the sequence of the common ancestor. The MSA algorithms try to maximize the matching residues in the given set of sequences with a pre-defined scoring scheme. The MSA produces a matrix of characters with species in the rows and character sites in columns. It also introduces the gaps, simulating the events of insertions and deletions (also called as indels). Insertion of gaps also helps in making the lengths of all sequences same for the sake of comparison. All the MSA algorithms are guaranteed to produce optimal alignment above a threshold value of detectable sequence similarity. The alignment accuracy is observed to decrease when sequence similarity drops below 35% towards the twilight (<35% but > 25%) and moonlight zones (<25%) of similarity. The character matrix obtained in MSA reveals the pattern of conservation and variability across the species, which in turn reveals the motifs and the signature sequences shared by species to retain the fold and function. The analysis of variations can be gainfully used to identify the changes that explain functional and phenotypic variability, if any, across OTUs.

Many algorithms have been specially developed for MSA and subsequently improved to achieve higher accuracy. One of the popular heuristics-based MSA approach follows progressive alignment procedure, in which sequences are compared in a pair wise fashion to build a distance matrix containing percent identity values. A clustering algorithm is then applied to distance matrix to generate a guide tree. The algorithm then follows a guide tree to add the pair wise alignments together starting from the leaf to root. This ensures the sequences with higher similarity are aligned initially and distantly related sequences are progressively added to the alignment of aligned sequences. Thus, the gaps inserted are always retained. A suitable scoring function, sum-of-pairs, consensus, consistency-based etc.

is employed to derive the optimum MSA (Nicholas et al., 2002; Batzoglou, 2005). Most of the MSA packages use Needleman and Wunsch (1970) algorithm to compute pair wise sequence similarity. The ClustalW is the widely used MSA package (Thompson et al., 1994). Recently many alternative MSA algorithms are also being developed, which are enlisted in Table 2. The standard benchmark datasets are used for comparative assessment of the alternative approaches (Aniba et al., 2010; Thompson et al., 2011). Irrespective of the proven performance of MSA methods for individual genes and proteins, some of the challenges and issues regarding computational aspects involved in handling genomic data are still the causes of concern (Kemena & Notredame, 2009).

Alignment programs	**Algorithm description**	**Available at / Reference**
ClustalW	Progressive	http://www.ebi.ac.uk/Tools/msa/clustalw2/ ; Thompson et al., 1994
MUSCLE	Progressive/iterative	http://www.ebi.ac.uk/Tools/msa/muscle/ ; Edgar, 2004
T-COFFEE	Progressive	http://www.ebi.ac.uk/Tools/msa/tcoffee/ ; Notredame et al., 2000
DIALIGN2	Segment-based	http://bibiserv.techfak.uni-bielefeld.de/dialign/ ; Morgenstern et al., 1998
MAFFT	Progressive/iterative	http://mafft.cbrc.jp/alignment/software/ ; Katoh et al., 2005
Alignment visualization programs		
*BioEdit		http://www.mbio.ncsu.edu/bioedit/bioedit.html ; Hall, 1999
MEGA5		http://www.megasoftware.net/ ; Kumar et al., 2008
DAMBE		http://dambe.bio.uottawa.ca/dambe.asp ; Xia & Xie, 2001
CINEMA5		http://aig.cs.man.ac.uk/research/utopia/cinema ; Parry-Smith et al., 1998

*: Not updated since 2008, but the last version is available for use.

Table 2. List of commonly used multiple sequence alignment programs and visualization tools.

The MSA output can also be visualized and edited, if required, with the software like BioEdit, DAMBE etc. Multiple alignment output shows the conserved and variable sites, usually residues are colour coded for the ease of visualisation, identification and analysis. The character sites in MSA can be divided as conserved (all the sequences have same residue or base), variable-non-informative (singleton site) and variable-informative sites. The sites containing gaps in all or majority of the species are of no importance from the evolutionary point of view and are usually removed from MSA while converting MSA data to input data for MPA. A sample MSA is shown in Fig. 2. The sequences of surface hydrophobic (SH) protein from various genotypes (A to M) of Mumps virus, are aligned. A careful visual inspection of MSA allows us to locate the patterns and motifs (LLLXIL) in a given set of sequences. Apart from MPA, the MSA data in turn can be used for the construction of position specific scoring matrix (PSSM), generation of consensus sequence,

sequence logos, identification and prioritisation of potential B- and T-cell epitopes etc. Nowadays the databases of curated, pre-computed alignments of reference species are also being made available, which can be used for the benchmark comparison, evaluation purpose (Thompson et al., 2011) and it also helps to keep the track of changes that get accumulated in the species over a period of time. For example, in case of viruses, observed changes are correlated with emergence of new genotypes (Kulkarni-Kale et al., 2004; Kuiken et al., 2005).

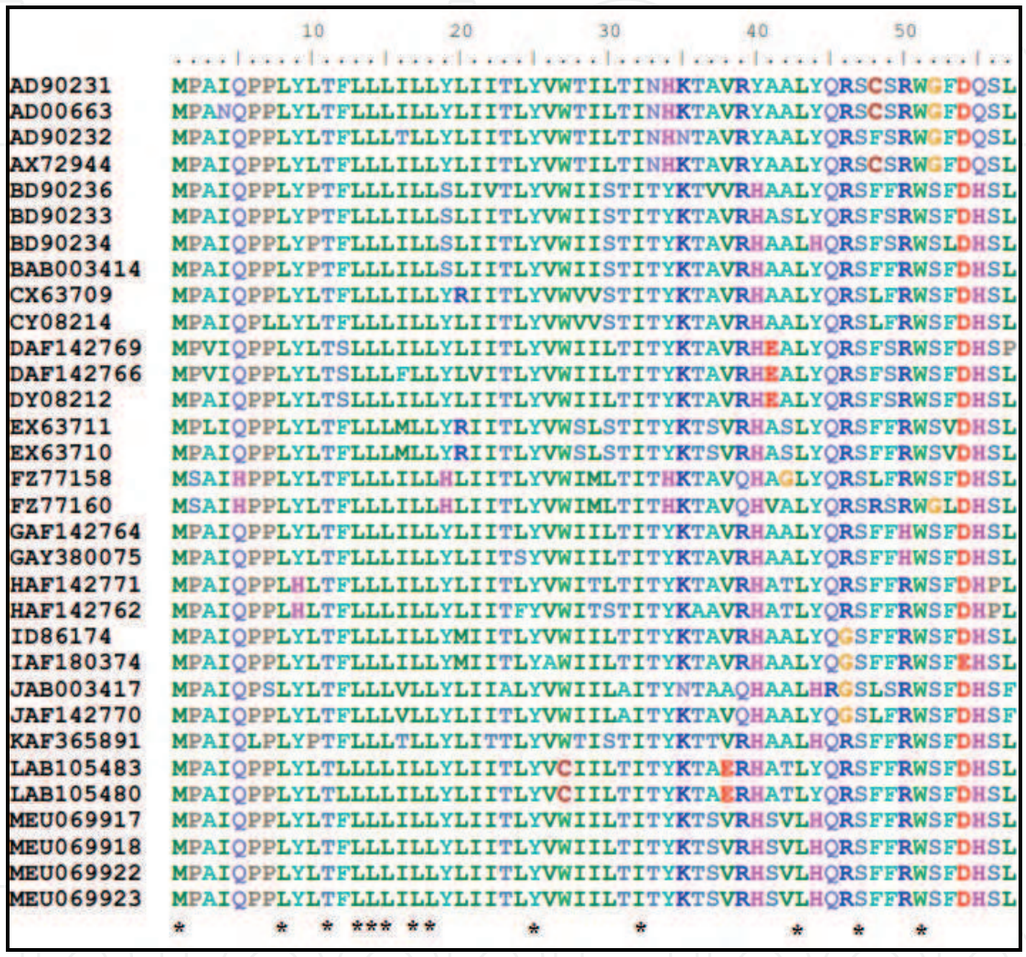


Fig. 2. The complete multiple sequence alignment of the surface hydrophobic (SH) proteins of Mumps virus genotypes (A to M) carried out using ClustalW. The MSA is viewed using BioEdit. The species labels in the leftmost column begin with genotype letter (A-M) followed by GenBank accession numbers. The scale for the position in alignment is given at the top of the alignment. The columns with conserved residues are marked with an "*" in the last row.

5. Selection of a suitable model of evolution

The existing MPA methods utilize the mathematical models to describe the evolution of sequence by incorporating the biological, biochemical and evolutionary considerations. These mathematical models are used to compute genetic distances between sequences. The use of appropriate model of evolution and statistical tests help us to infer maximum evolutionary information out of sequence data. Thus, the selection of the right model of

sequence evolution becomes important as a part of effective MPA. Two types of approaches are adapted for the building of models, first one is empirical i.e. using the properties revealed through comparative studies of large datasets of observed sequences, and the other is parametrical, which uses biological and biochemical knowledge about the nucleic acid and protein sequences, for example the favoured substitution patterns of residues. Parametric models obtain the parameters from the MSA dataset under study. Both types of approaches result in the models based on the Markov process, in the form of matrix representing the rate of all possible transitions between the types of residues (4 nucleotides in nucleic acids and 20 amino acids in proteins). According to the type of sequence (nucleic acid or protein), two categories of models have been developed.

5.1 Models of nucleotide substitution

The nucleotide substitution models are based on the parametric approach with the use of mainly three parameters i) nucleotides frequencies, ii) rate of nucleotide substitutions and iii) rate heterogeneity. Nucleotide frequencies, account for the compositional sequence constraints such as GC content. These are subsequently used in a model to allow the substitutions of a certain type to occur more likely than others. The nucleotide substitution parameter is used to represent a measure of biochemical similarity. Higher the similarity between the nucleotide bases, the more is the rate of substitution between them, for example, the transitions are more frequent than transversions. A parameter of rate heterogeneity accounts for the unequal rates of substitution across the variable sites, which can be correlated with the constraints of genetic code, selection for the gene function etc. The site variability is modelled by gamma distribution of rates across sites. The shape parameter of gamma distribution determines amount of heterogeneity among sites, larger values of shape parameter gives a bell shaped distribution suggesting little or no rate variation across the sites whereas small values of it gives J-shaped distribution indicating high rate variation among sites along with low rates of evolution at many sites.

Varieties of nucleotide substitution models have been developed with a set of assumptions and parameters described as above. Some of the well-known models of nucleotide substitutions include Jukes-Cantor (JC) one-parameter model (Jukes & Cantor, 1969), Kimura two-parameter model (K2P) (Kimura, 1980), Tamura's model (Tamura, 1992), Tamura and Nei model (Tamura & Nei, 1993) etc. These models make use of different biological properties such as, transitions, transversions, G+C content etc. to compute distances between nucleotide sequences. The substitution patterns of nucleotides for some of these models are shown in Fig. 3.

5.2 Models of amino acid replacement

In contrast to nucleotide substitution models, amino acid replacement models are developed using empirical approach. Schwarz and Dayhoff (1979) developed the most widely used model of protein evolution in which, the replacement matrix was obtained from the alignment of globular protein sequences with 15% divergence. The Dayhoff matrices, known as PAM matrices, are also used by database searching methods. The similar methodology was adopted by other model developers but with specialized databases. Jones et al., (1994) have derived a replacement matrix specifically for membrane proteins, which has values significantly different from Dayhoff matrix suggesting the remarkably different pattern of amino acid replacements observed in the membrane proteins. Thus, such a matrix will be more

appropriate for the phylogenetic study of membrane proteins. On the other hand, Adachi and Hasegawa (1996) obtained a replacement matrix using mitochondrial proteins across 20 vertebrate species and can be effectively used for mitochondrial protein phylogeny. Henikoff and Henikoff (1992) derived the series of BLOSUM matrices using local, ungapped alignments of distantly related sequences. The BLOSUM matrices are widely used in similarity searches against databases than for phylogenetic analyses.

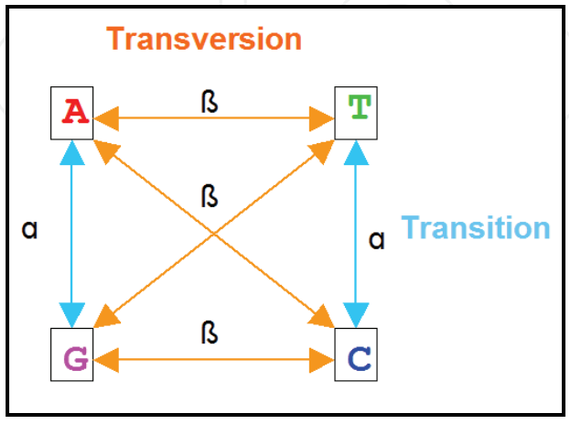


Fig. 3. The types of substitutions in nucleotides. α denotes the rate of transitions and β denotes the rate of transversions. For example, in the case of JC model α=β while in the case of K2P model α>β.

Recently, structural constraints of the nucleic acids and proteins are also being incorporated in the building of models of evolution. For example, Rzhetsky (1995) contributed a model to estimate the substitution patterns in ribosomal RNA genes with the account of secondary structure elements like stem-loops in ribosomal RNAs. Another approach introduced a model with the combination of protein secondary structures and amino acid replacement (Lio & Goldman, 1998; Thorne et al., 1996). The overview of different models of evolution and the criteria for the selection of models is also provided by Lio & Goldman (1998); Luo et al. (2010).

6. Reconstruction of a phylogenetic tree

The phylogeny reconstruction methods result in a phylogenetic tree, which may or may not corroborate with the true phylogenetic tree. There are various methods of phylogeny reconstruction that are divided into two major groups viz. character-based and distance-based.

Character-based methods use a set of discrete characters, for example, in case of MSA data of nucleotide sequences, each position in alignment is reffered as "character" and nucleotide (A, T, G or C) present at that position is called as the "state" of that "character". All such characters are assumed to evolve independent of each other and analysed separately. Distance-based methods on other hand use some form of distance measure to compute the dissimilarity between pairs of OTUs, which subsequently results in derivation of distance matrix that is given as an input to clustering methods like Neighbor-Joining (N-J) and Unweighted Pair Group Method with Arithmetic mean (UPGMA) to infer phylogenetic tree. The character-based and distance-based methods follow exhaustive search and/or stepwise clustering approach to arrive at an optimum phylogenetic tree, which explains the

evolutionary pattern of the OTUs under study. The exhaustive search method examines theoretically all possible tree topologies for a chosen number of species and derives the best tree topology using a set of certain criteria. Table 3 shows the possible number of rooted and unrooted trees for n number of species/OTUs.

Number of OTUs	Number of unrooted trees	Number of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
10	2027025	34459425

Table 3. The number of possible rooted and unrooted trees for a given number of OTUs. The number of possible unrooted trees for n OTUs is given by $(2n-5)!/[2^{n-3}(n-3)!]$; and rooted trees is given by $(2n-3)!/[2^{n-2}(n-2)!]$

Whereas, stepwise clustering methods employ an algorithm, which begins with the clustering of highly similar OTUs. It then combines the clustered OTUs such that it can be treated as a single OTU representing the ancestor of combined OTUs. This step reduces the complexity of data by one OTU. This process is repeated and in a stepwise manner adding the remaining OTUs until all OTUs are clustered together. The stepwise clustering approach is faster and computationally less intensive than the exhaustive search method.

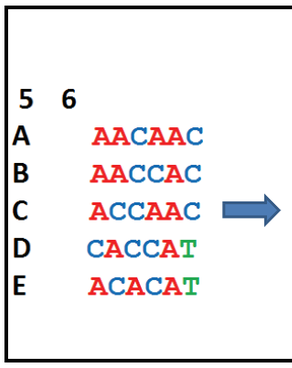
The most widely used distance-based methods include N-J & UPGMA and character-based methods include Maximum Parsimony (MP) and Maximum Likelihood (ML) methods (Felsenstein, 1996). All of these methods make particular assumptions regarding evolutionary process, which may or may not be applicable to the actual data. Thus, before selection of a phylogeny reconstruction method, it is recommended to take into account the assumptions made by the method to infer the best phylogenetic tree. The list of widely used phylogeny inference packages is given in Table 4.

Package	Available from / Reference
PHYLIP	http://evolution.genetics.washington.edu/phylip.html ; Felsenstein, 1989
PAUP	http://paup.csit.fsu.edu/ ; Wilgenbusch & Swofford, 2003
MEGA5	http://www.megasoftware.net/ ; Kumar et al., 2008
MrBayes	http://mrbayes.csit.fsu.edu/ ; Ronquist & Huelsenbeck, 2003
TREE-PUZZLE	http://www.tree-puzzle.de/ ; Schmidt et al., 2002

Table 4. The list of widely used packages for molecular phylogeny.

6.1 Distance-based methods of phylogeny reconstruction

The distance-based phylogeny reconstruction begins with the computation of pair wise genetic distances between molecular sequences with the use of appropriate substitution model, which is built on the basis of evolutionary assumptions, discussed in section 4. This step results in derivation of a distance matrix, which is subsequently used to infer a tree topology using the clustering method. Fig. 4 shows the distance matrix computed for a sample sequence dataset of 5 OTUs with 6 sites using Jukes-Cantor distance measure. A distance measure possesses three properties, (a) a distance of OTU from itself is zero, D(i, i) = 0; (b) the distance of OTU i from another OTU j must be equal to the distance of OTU j from OTU i, D(i, j) = D(j, i); and (c) the distance measure should follow the triangle inequality rule i.e. D(i, j) ≤ D(i, k) + D(k, j). The accurate estimation of genetic distances is a crucial requirement for the inference of correct phylogenetic tree, thus choice of the right model of evolution is as important as the choice of clustering method. The popular methods used for clustering are UPGMA and N-J.



	A	B	C	D	E
A	0.000000				
B	0.188486	0.000000			
C	0.188486	0.440840	0.000000		
D	0.823959	0.440840	1.647918	0.000000	
E	1.647918	0.823959	0.823959	0.823959	0.000000

Fig. 4. The distance matrix obtained for a sample nucleotide sequence dataset using Jukes-Cantor model. Dataset contains 5 OTUs (A-E) and 6 sites shown in Phylip format. Dnadist program in PHYLIP package is used to compute distance matrix.

6.1.1 UPGMA method for tree building

The UPGMA method was developed by Sokal and Michener (1958) and is the most widely used clustering methodology. The method is based on the assumptions that the rate of substitution for all branches in the tree is constant (which may not hold true for all data) and branch lengths are additive. It employs hierarchical agglomerative clustering algorithm, which produces ultrametric tree in such a way that every OTU is equidistant from the root.

The clustering process begins with the identification of the highly similar pair of OTUs (i & j) as decided from the distance value D(i, j) in distance matrix. The OTUs i and j are clustered together and combined to form a composite OTU ij. This gives rise to new distance matrix shorter by one row and column than initial distance matrix. The distances of un-clustered OTUs remain unchanged. The distances of remaining OTUs (for e.g. k) from composite OTUs are represented as the average of the initial distances of that OTU from the individual members of composite OTU (i.e. D(ij, k) = [D(i, k) + D(j, k)]/2). In this way a new distance matrix is calculated and in the next round, the OTUs with least dissimilarity are clustered together to form another composite OTU. The remaining steps are same as discussed in the first round. This process of clustering is repeated until all the OTUs are clustered.

The sample calculations and steps involved in UPGMA clustering algorithm using distance matrix shown in Fig. 4 are given below.

Iteration 1: OTU A is minimally equidistant from OTUs B and C. Randomly we select the OTUs A and B to form one composite OTU (AB). A and B are clustered together. Compute new distances of OTUs C, D and E from composite OTU (AB). The distances between unclustered OTUs will be retained. See Fig. 4 for initial distance matrix and Fig. 5 for updated matrix after first iteration of UPGMA.

d(AB,C) = [d(A,C) + d(B,C)]/2 = [0.188486 + 0.440840]/2 = 0.314633
d(AB,D) = [d(A,D) + d(B,D)]/2 = [0.823959 + 0.440840]/2 = 0.632399
d(AB,E) = [d(A,E) + d(B,E)]/2 = [1.647918 + 0.823959]/2 = 1.235938

	AB	C	D	E
AB	0.000000			
C	0.314633	0.000000		
D	0.632399	1.647918	0.000000	
E	1.235938	0.823959	1.647918	0.000000

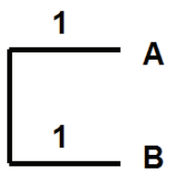


Fig. 5. The updated distance matrix and clustering of A and B after the 1st iteration of UPGMA.

Iteration 2: OTUs (AB) and C are minimally distant. We select these OTUs to form one composite OTU (ABC). AB and C are clustered together. We then compute new distances of OTUs D and E from composite OTU (ABC). See Fig. 5 for distance matrix obtained in iteration 1 and Fig. 6 for updated matrix after the second iteration of UPGMA.

d(ABC,D) = [d(AB,D) + d(C,D)]/2 = [0.632399 + 1.647918]/2 = 1.140158
d(ABC,E) = [d(AB,E) + d(C,E)]/2 = [1.235938 + 0.823959]/2 = 1.029948

	ABC	D	E
ABC	0.000000		
D	1.140158	0.000000	
E	1.029948	0.823959	0.000000

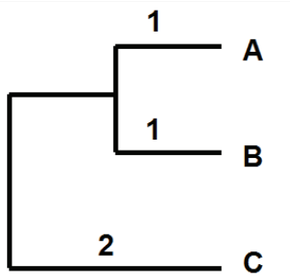


Fig. 6. The updated distance matrix and clustering of A, B and C after the 2nd iteration of UPGMA.

Iteration 3: OTUs D and E are minimally distant. We select these OTUs to form one composite OTU (DE). D and E are clustered together. Compute new distances of OTUs (ABC) and (DE) from each other. Finally the remaining two OTUs are clustered together. See Fig. 6 for distance matrix obtained in iteration 2 and Fig. 7 for updated matrix after third iteration of UPGMA.

d(ABC,DE) = [d(ABC,D) + d(ABC,E)]/2 = [1.140158 + 1.029948]/2 = 1.085053

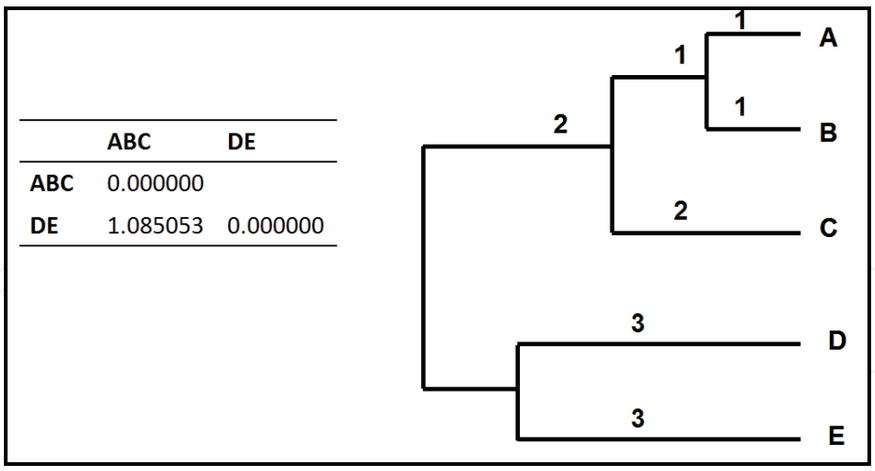


Fig. 7. The updated distance matrix and clustering of OTUs after the 3rd iteration of UPGMA. Numbers on the branches indicate branch lengths, which are additive.

6.1.2 N-J method for tree building

The N-J method for clustering was developed by Saitou and Nei (1987). It reconstructs the unrooted phylogenetic tree with branch lengths using minimum evolution criterion that minimizes the lengths of tree. It does not assume the constancy of substitution rates across sites and does not require the data to be ultrametric, unlike UPGMA. Hence, this method is more appropriate for the sites with variable rates of evolution.

N-J method is known to be a special case of the star decomposition method. The initial tree topology is a star. The input distance matrix is modified such that the distance between every pair of OTUs is adjusted using their average divergence from all remaining OTUs. The least dissimilar pair of OTUs is identified from the modified distance matrix and is combined together to form single composite OTU. The branch lengths of individual members, clustered in composite OTU, are computed from internal node of composite OTU. Now the distances of remaining OTUs from composite OTU are redefined to give a new distance matrix shorter by one OTU than the initial matrix. This process is repeated till all the OTUs are grouped together, while keeping track of nodes, which results in a final unrooted tree topology with minimized branch lengths. The unrooted phylogenetic tree, thus obtained can be rooted using an outgroup species. The BIONJ (Gascuel 1997), generalized N-J (Pearson et al., 1999) and Weighbor (Bruno et al., 2000) are some of the recently proposed alternative versions of N-J algorithm. The sample calculation and steps involved in N-J clustering algorithm, using distance matrix shown in Fig. 4, are given below.

Iteration 1: Before starting the actual process of clustering the vector r is calculated as following with N=5, refer to the initial distance matrix given in Fig. 4 for reference values.

r(A) = [d(A,B)+ d(A,C)+ d(A,D)+ d(A,E)]/(N-2) = 0.949616
r(B) = [d(B,A)+ d(B,C)+ d(B,D)+ d(B,E)]/(N-2) = 0.631375
r(C) = [d(C,A)+ d(C,B)+ d(C,D)+ d(C,E)]/(N-2) = 1.033755
r(D) = [d(D,A)+ d(D,B)+ d(D,C)+ d(D,E)]/(N-2) = 1.245558
r(E) = [d(E,A)+ d(E,B)+ d(E,C)+ d(E,D)]/(N-2) = 1.373265

Using these r values, we construct a modified distance matrix, Md, such that
MD(i,j) = d(i,j) – (ri + rj).
See Fig. 8 for Md.

Md	A	B	C	D	E
A	0.000000				
B	-1.392505	0.000000			
C	-1.794885	-1.224290	0.000000		
D	-1.371215	-1.436093	-0.631395	0.000000	
E	-0.674963	-1.180681	-1.583061	-1.794864	0.000000

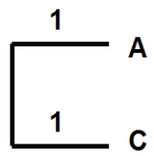


Fig. 8. The modified distance matrix Md and clustering for iteration 1 of N-J.

As can be seen from Md in Fig. 8, OTUs A and C are minimally distant. We select the OTUs A and C to form one composite OTU (AC). A and C are clustered together.

Iteration 2: Compute new distances of OTUs B, D and E from composite OTU (AC). Distances between unclustered OTUs will be retained from the previous step.

d(AC,B) = [d(A,B) + d(C,B)-d(A,C)]/2 = 0.22042

d(AC,D) = [d(A,D) + d(C,D) -d(A,C)]/2 = 1.141695

d(AC,E) = [d(A,E) + d(C,E) -d(A,C)]/2 = 1.141695

Compute r as in the previous step with N=4. See Fig. 9 for new distance matrix and r vector.

d	AC	B	D	E
AC	0.000000			
B	0.220420	0.000000		
D	1.141695	1.647918	0.000000	
E	1.141695	0.823959	1.647918	0.000000

r	
AC	1.251905
B	1.346148
D	2.218765
E	1.806786

Fig. 9. The new distance matrix D and vector r obtained for NJ algorithm iteration 2.

Now, we compute the modified distance matrix, Md as in the previous step and cluster the minimally distant OTUs. See Fig. 10

Md	AC	B	D	E
AC	0.000000			
B	-2.377633	0.000000		
D	-2.327975	-1.916995	0.000000	
E	-1.916996	-2.328975	-2.377633	0.000000

Fig. 10. The modified distance matrix Md, obtained during N-J algorithm iteration 2.

In this step, AC & B and D & E are minimally distant, so we cluster AC with B and D with E. Repeating the above steps we will finally get the following phylogenetic tree, Fig. 11.

Both the distance-based methods, UPGMA and N-J, are computationally faster and hence suited for the phylogeny of large datasets. N-J is the most widely used distance-based method for phylogenetic analysis. The results of these methods are highly dependent on the model of evolution selected a priori.

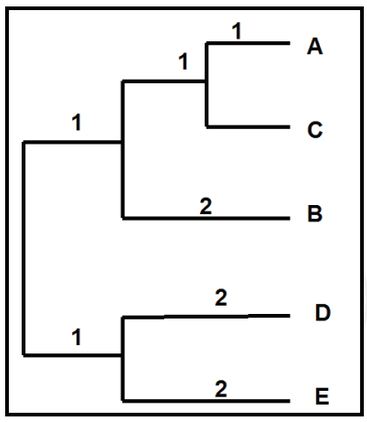


Fig. 11. The phylogenetic tree obtained using N-J algorithm for distance matrix in Fig 4. Numbers on the branches indicate branch length.

6.2 Character-based methods of phylogeny reconstruction

The most commonly used character-based methods in molecular phylogenetics are Maximum parsimony and Maximum likelihood. Unlike the distance-based MPA, character-based methods use character information in alignment data as an input for tree building. The aligned data is in the form of character-state matrix where the nucleotide or amino acid symbols represent the states of characters. These character-based methods employ optimality criterion with the explicit definition of objective function to score the tree topology in order to infer the optimum tree. Hence, these methods are comparatively slower than distance-based clustering algorithms, which are simply based on a set of rules and operations for clustering. But character based methods are advantageous in the sense that they provide a precise mathematical background to prefer one tree over another unlike in distance-based clustering algorithms.

6.2.1 Maximum parsimony

The Maximum parsimony (MP) method is based on the simple principle of searching the tree or collection of trees that minimizes the number of evolutionary changes in the form of change of one character state into other, which are able to describe observed differences in the informative sites of OTUs. There are two problems under the parsimony criterion, a) determining the length of the tree i.e. estimating the number of changes in character states, b) searching overall possible tree topologies to find the tree that involves minimum number of changes. Finally all the trees with minimum number of changes are identified for each of the informative sites. Fitch's algorithm is used for the calculation of changes for a fixed tree topology (Fitch, 1971). If the number of OTUs, N is moderate, this algorithm can be used to calculate the changes for all possible tree topologies and then the most parsimonious rooted tree with minimum number of changes is inferred. However, if N is very large it becomes computationally expensive to calculate the changes for the large number of possible rooted trees. In such cases, a branch and bound algorithm is used to restrict the search space of tree topologies in accordance with Fitch's algorithm to arrive at parsimonious tree (Hendy & Penny, 1982). However, this approach may miss some parsimonious topologies in order to reduce the search space.

An illustrative example of phylogeny analysis using Maximum parsimony is shown in Table 5 and Fig. 12. Table 5 shows a snapshot of MSA of 4 sequences where 5 columns show the

aligned nucleotides. Since there are four taxa (A, B, C & D), three possible unrooted trees can be obtained for each site. Out of 5 character sites, only two sites, viz., 4 & 5 are informative i.e. sites having at least two different types of characters (nucleotides/amino acids) with a minimum frequency 2. In the Maximum parsimony method, only informative sites are analysed. Fig. 12 shows the Maximum parsimony phylogenetic analysis of site 5 shown in Table 5. Three possible unrooted trees are shown for site 5 and the tree length is calculated in terms of number of substitutions. Tree II is favoured over trees I and III as it can explain the observed changes in the sequences just with a single substitution. In the same way unrooted trees can be obtained for other informative sites such as site 4. The most parsimonious tree among them will be selected as the final phylogenetic tree. If two or more trees are found and no unique tree can be inferred, trees are said to be equally parsimonious.

OTUs	Character sites				
	1	2	3	4	5
A	A	A	C	A	A
B	A	A	C	C	A
C	A	C	T	A	G
D	C	A	C	C	G

Table 5. Example of phylogenetic analysis from 5 aligned character sites in 4 OTUs using Maximum parsimony method.

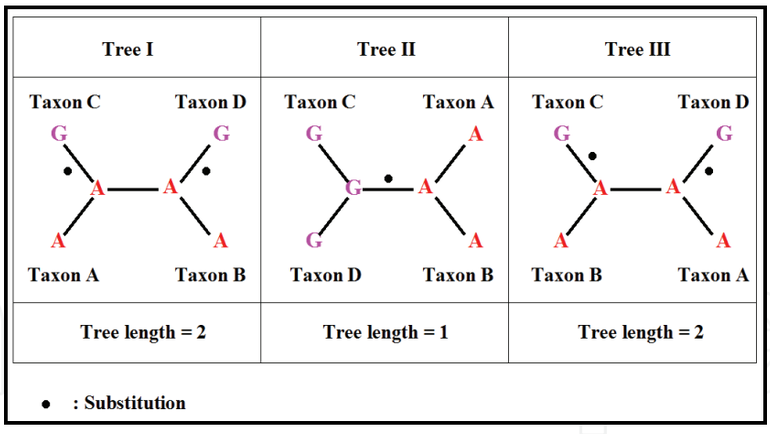


Fig. 12. Example showing various tree topologies based on site 5 in Table 5 using the Maximum parsimony method.

This method is suitable for a small number of sequences with higher similarity and was originally developed for protein sequences. Since this method examines the number of evolutionary changes in all possible trees it is computationally intensive and time consuming. Thus, it is not the method of choice for large sized genome sequences with high variation. The unequal rates of variation in different sites can lead to erroneous parsimony tree with some branches having longer lengths than others as parsimony method assumes the rate of change across all sites to be equal.

6.2.2 Maximum likelihood

As mentioned in the beginning, another character based method for the MPA is the Maximum likelihood method. This method is based on probabilistic approach to phylogeny. This approach is different from the methods discussed earlier. In this method probabilistic models for phylogeny are developed and the tree would be reconstructed using Maximum likelihood method or by sampling method for the given set of sequences. The main difference between this method and some of the available methods discussed before is that it ranks various possible tree topologies according to their likelihood. The same can be obtained by either using the frequentist approach (using the probability (data|tree)) or by using the Baysian approach (likelihood based on the posterior probabilities i.e. by using probability (tree|data)). This method also facilitates computing the likelihood of a sub-tree topology along the branch.

To make the method operative, one must know how to compute P(x*|T,t*) probability of set of data given tree topology T and set of branch length t*. The tree having maximum probability or the one, which maximizes the likelihood would be chosen as the best tree. The maximization can also be based on the posterior probability P(tree|data) and can be carried out by obtaining required probability using P(x*|T,t*)=P(data|tree) and by applying the Baye's theorem.

The exercise of maximization involves two steps:

a. A search over all possible tree topologies with order of assignment of sequences at the leaves specified.
b. For each topology, a search over all possible lengths of edges in t*

As mentioned in the chapter earlier, the number of rooted trees for given number of sequences (N) grows very rapidly even as N increases to 10. An efficient search procedure for these tasks is required, which was proposed by Felsenstein (1981) and is extensively being used in the MPA. The maximization of likelihood of edge lengths can be carried out using various optimization techniques.

An alternative method is to search stochastically over trees by sampling from posterior distribution P(T,t*|x*). This method uses techniques such as Monte Carlo method, Gibb's sampling etc. The results of this method are very promising and are often recommended.

Having briefly reviewed the principles, merits and limitations of various methods available for reconstruction of phylogenetic trees using molecular data, it becomes evident that the choice of method for MPA is very crucial. The flowchart shown in Fig. 13 is intended to serve as a guideline to choose a method based on extent of similarity between the sequences. However, it is recommended that one uses multiple methods (at least two) to derive the trees. A few programs have also been developed to superimpose trees to find out similarities in the branching pattern and tree topologies.

7. Assessing the reliability of phylogenetic tree

The assessment of the reliability of phylogenetic tree is an important part of MPA as it helps to decide the relationships of OTUs with a certain degree of confidence assigned by statistical measures. Bootstrap and Jackknife analyses are the major statistical procedures to evaluate the topology of phylogenetic tree (Efron, 1979; Felsenstein, 1985).

In bootstrap technique, the original aligned dataset of sequences is used to generate the finite population of pseudo-datasets by "sampling with replacement" protocol. Each pseudo-dataset is generated by sampling n character sites (columns in the alignment)

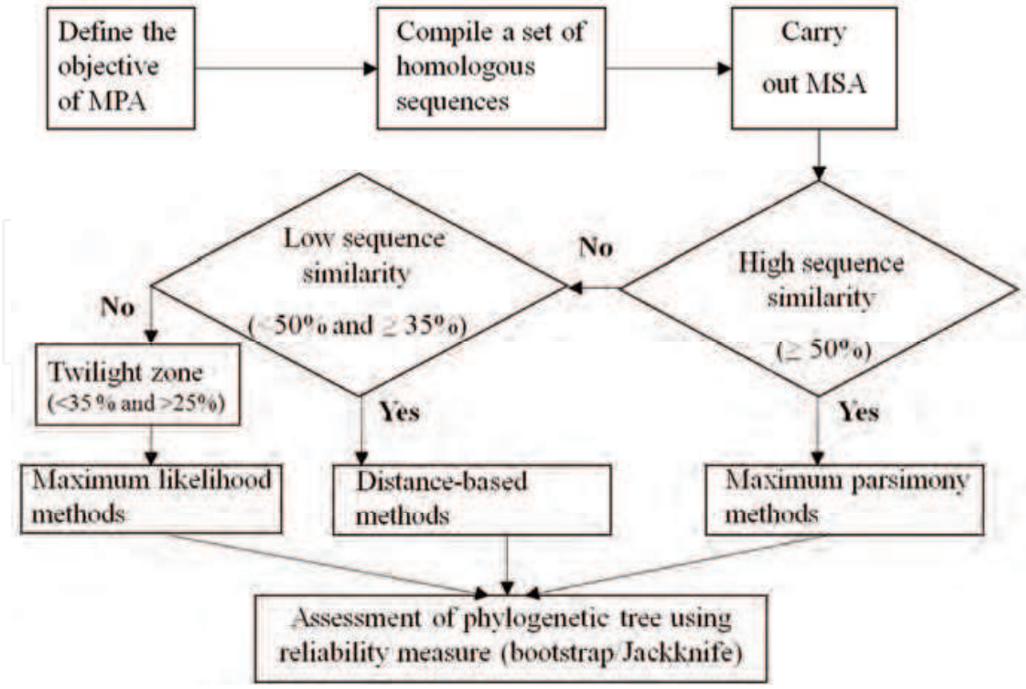


Fig. 13. Flowchart showing the analysis steps involved in phylogenetic reconstruction.

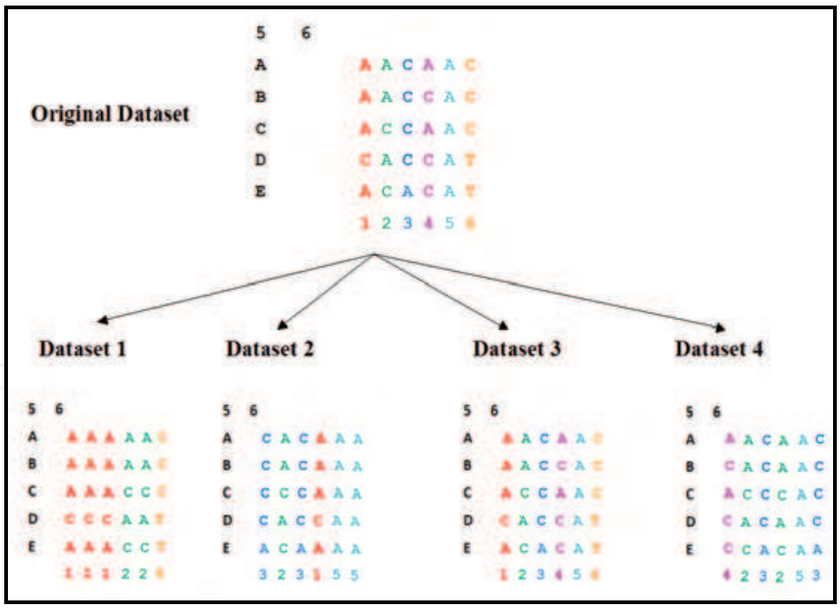


Fig. 14. The procedure to generate pseudo-replicate datasets of original dataset using bootstrap is shown above. The character sites are shown in colour codes at the bottom of datasets to visualize "sampling with replacement protocol".

randomly from original dataset with a possibility of sampling the same site repeatedly, in the process of regular bootstrap. This leads to generation of population of datasets, which are given as an input to tree building methods thus giving rise to population of phylogenetic

trees. The consensus phylogenetic tree is then inferred by the majority rule that groups those OTUs, which are found to cluster most of the times in the population of trees. The branches in consensus phylogenetic tree are labelled with bootstrap support values enabling the significance of the relationship of OTUs as depicted using a branching pattern. The procedure for regular bootstrap is illustrated in the Fig. 14. It shows the original dataset along with four pseudo-replicate datasets.

The sites in the original dataset are colour coded to visualize the "sampling with replacement protocol" used in generation of pseudo-replicate datasets 1-4. Seqboot program in PHYLIP package was used for this purpose with choice of regular bootstrap. For example, pseudo-replicate dataset 1 contains the site 1 (red) from original dataset sampled 3 times. In the general practice, usually 100 to 1000 datasets are generated and for each of the datasets phylogenetic tree is obtained. The consensus phylogenetic tree is then obtained by majority rule. The reliability of the consensus tree is assessed from the "branch times" value displayed along the branches of tree.

In Jackknife procedure, the pseudo-datasets are generated by "sampling without replacement" protocol. In this process, sampling (<n) character sites randomly from original dataset generates each pseudo dataset. This leads to generation of population of datasets, which are given as an input to tree building methods thus giving rise to population of phylogenetic trees. The consensus phylogenetic tree is inferred by the majority rule that groups those OTUs, which are found to be clustered most of the times in the population of trees.

8. The case study of Mumps virus phylogeny

We have chosen a case study of Mumps virus (MuV) phylogeny using the amino acid sequences of surface hydrophobic (SH) proteins. There are 12 different known genotypes of MuV, which are designated through A to L, based on the sequence similarity of SH gene sequences. Recently a new genotype of MuV, designated as M, has been identified during parotitis epidemic 2006-2007 in the state of São Paulo, Brazil (Santos et al., 2008). Extensive phylogenetic analysis of newly discovered genotype with existing genotypes of reference strains (A-L) has been used for the confirmation of new genotype using character-based Maximum likelihood method (Santos et al., 2008). In the case study to be presented here, we have used distance-based Neighbor-Joining method with an **objective to re-confirm the presence of new MuV genotype M.** The dataset reported in Santos et al., (2008) is used for the re-confirmation analysis. The steps followed in the MPA are listed below.

a. Compilation and curation of sequences: The sequences of SH protein of the strains of reference genotypes (A to L) as well as newly discovered genotype (M) of MuV were retrieved using GenBank accession numbers as given in Santos et al., (2008). Sequences were saved in Fasta format.
b. Multiple sequence alignment (MSA): SH proteins were aligned using ClustalW (See Fig. 2). MSA was saved in Phylip or interleaved (.phy) format.
c. Bootstrap analysis: 100 pseudo-replicate datasets of the original MSA data (obtained in step b) were generated using regular bootstrap methods in Seqboot program of PHYLIP package.
d. Derivation of distance: The distances between sequences in each dataset were calculated using Dayhoff PAM model assuming uniform rate of variation at all sites. The 'outfile' generated by Seqboot program was used as an input to Protdist program in PHYLIP package.

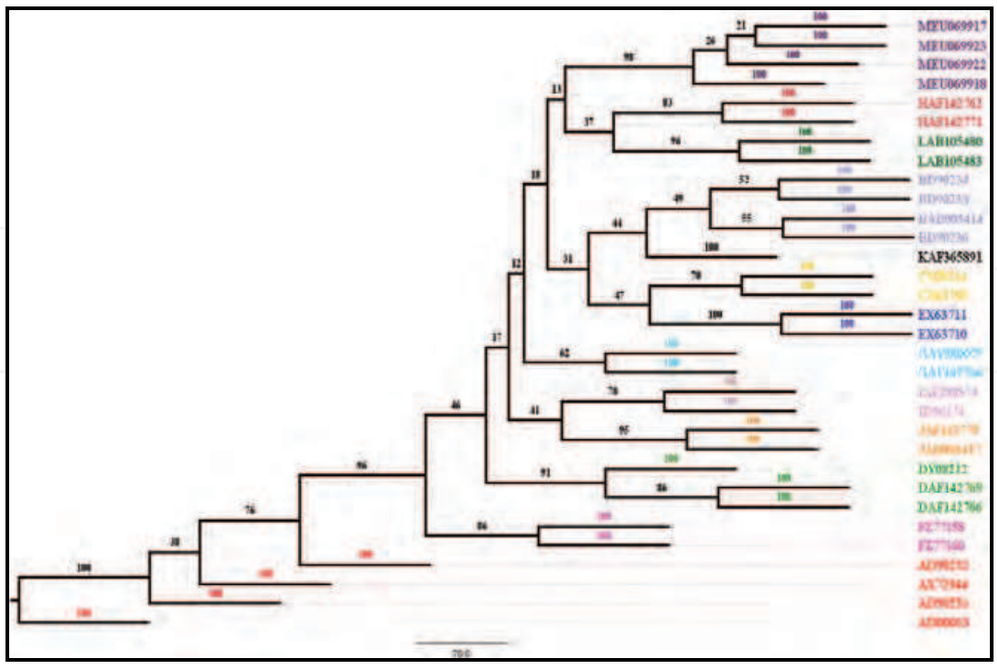


Fig. 15. The unrooted consensus phylogenetic tree obtained for Mumps virus genotypes using Neighbor-Joining method. The first letter in OTU labels indicates the genotype (A-M), which is followed by the GenBank accession numbers for the sequences. The OTUs are also colour coded according to genotypes as following, A: red; B: light blue; C: yellow; D: light green; E: dark blue; F: magenta; G: cyan; H: brick; I: pink; J: orange; K: black; L: dark green; M: purple. All of the genotypes have formed monophyletic clades with high bootstrap support values shown along the branches. The monophyletic clade of M genotypes (with 98 bootstrap support at its base) separated from the individual monophyletic clades of other genotypes (A-L) re-confirms the detection of new genotype M.

e. Building phylogenetic tree: The distance matrices obtained in the previous step were given as an input to N-J method to build phylogenetic trees. The 'outfile' generated by Protdist program containing distance matrices was given as an input to Neighbor program in PHYLIP package.
f. The consensus phylogenetic tree was then obtained using Consense program. For this purpose the 'outtree' file (in Newick format) generated by Neighbor program was given as an input to Consense program.
g. The consensus phylogenetic tree was visualized using FigTree software (available from http://tree.bio.ed.ac.uk/software/figtree/). The consensus unrooted phylogenetic tree is shown in Fig. 15.

The phylogenetic tree for the same dataset was also obtained by using Maximum parsimony method, implemented as the Protpars program in PHYLIP by carrying out MSA and bootstrap as detailed above. The consensus phylogenetic tree is shown in Fig. 16.

Comparison of the trees shown in Fig. 15 & Fig. 16 with that of the published tree re-confirms the emergence of new MuV genotype M during the epidemic in São Paulo, Brazil (Santos et al., 2008), as the members of genotype M have formed a distinct monophyletic clade similar to the known genotypes (A-L). But, a keen observer would note the differences in ordering of clades in the two phylograms obtained using two different methods viz., N-J

and MP. For example, the clade of genotype J is close to the clade of genotype I in the N-J phylogram (see Fig. 15) whereas in the MP phylogram (Fig. 16) the clade of genotype J is shown to cluster with the clade of genotype F. Such differences in the ordering of clades are observed some times as these methods (N-J & MP) employ different assumptions and models of evolution. The user can interpret the results with reasonable confidence where the similar clustering pattern of clades is observed in trees drawn using multiple methods. The user, on the other hand, should refrain from over interpretation of sub-tree topolologies, where branching order doesn't match in the trees drawn using different methods. Similarly, a lot of case studies pertaining to the emergence of new species as well as evolution of individual genes/proteins have been published. It is advisable to re-run through a few case studies, which are published, to understand the way in which the respective authors have interpreted the results on the basis of phylogenetic analyses.

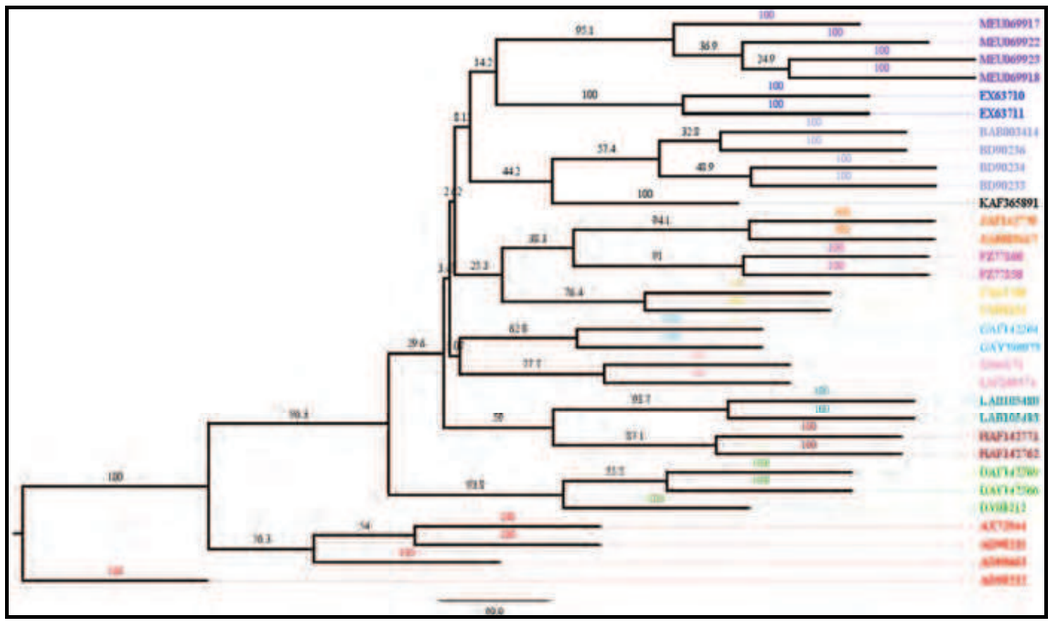


Fig. 16. The unrooted consensus phylogenetic tree obtained for Mumps virus genotypes using Maximum parsimony method. The labelling of OTUs and colour coding is same as in Fig. 15.

9. Challenges and opportunities in phylogenomics

The introduction of next-generation sequencing technology has totally revived the pace of genome sequencing. It has inevitably posed challenges on traditional ways of molecular phylogeny analysis based on single gene, set of genes or markers. Currently the phylogeny based on molecular markers such as 16S rRNA, mitochondrial, nuclear genes etc. provide the taxonomic backbone for Tree of Life (http://tolweb.org/tree/phylogeny.html). But the single gene based phylogeny does not necessarily reflect the phylogenetic history among the genomes of organisms from which these genes are derived. Also the types of evolutionary events such as lateral gene transfer, recombination etc. may not be revealed through the phylogeny of single gene. Thus whole genome based phylogeny analyses become important for deeper understanding of the evolutionary pattern in the organisms (Konstantinidis &

Tiedje, 2007). But whole genome based phylogeny poses many challenges to the traditional methods of MPA, major concerns of them being the size, memory and computational complexity involved in alignment of genomes (Liu et al., 2010).

The methods of MSA developed so far are adequate to handle the requirements of limited amount of data viz. individual gene or protein sequences from various organisms. The increased size of data in terms of the whole genome sequences, however, poses constrains on use and applicability of currently available methods of MSA as they become computationally intensive with requirement of higher memory. The uncertainty associated with alignment procedures, which leads to variations in the inferred phylogeny, has also been pointed out to be the cause of concern (Wong et al., 2008). The benchmark datasets are made available to validate performance of multiple sequence alignment methods (Kemena & Notredame, 2009). These challenges have opened up opportunities for development of alternative approaches for MPA with emergence of alignment-free methods for the same (Kolekar et al., 2010; Sims et al., 2009; Vinga & Almeida, 2003). The field of MPA is also evolving with attempts to develop novel methods based on various data mining techniques viz. Hidden Markov Model (HMM) (Snir & Tuller, 2009), Chaos game theory (Deschavanne et al., 1999), Return Time Distributions (Kolekar et al., 2010) etc. The recent approaches are undergoing refinement and will have to be evaluated with the benchmark datasets before they are routinely used. However, sheer dimensionality of genomic data demands their application. These approaches along with the conventional approaches are extensively reviewed elsewhere (Blair & Murphy, 2010; Wong & Nielsen, 2007).

10. Conclusion

The chapter provides excursion of molecular phylogeny analyses for potential users. It gives an account of available resources and tools. The fundamental principles and salient features of various methods viz. distance-based and character-based are explained with worked out examples. The purpose of the chapter will be served if it enables the reader to develop overall understanding, which is critical to perform such analyses involving real data.

11. Acknowledgment

PSK acknowledges the DBT-BINC junior research fellowship awarded by Department of Biotechnology (DBT), Government of India. UKK acknowledges infrastructural facilities and financial support under the Centre of Excellence (CoE) grant of DBT, Government of India.

12. References

Adachi, J. & Hasegawa, M. (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution* 42(4):459-468.

Aniba, M.; Poch, O. & Thompson J. (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Research* 38(21):7353-7363.

Batzoglou, S. (2005) The many faces of sequence alignment. *Briefings in Bioinformatics* 6(1):6-22.

Benson, D.; Karsch-Mizrachi, I.; Lipman, D.; Ostell, J. & Sayers, E. (2011) GenBank. *Nucleic Acids Research* 39(suppl 1):D32-D37.

Blair, C. & Murphy, R. (2010) Recent Trends in Molecular Phylogenetic Analysis: Where to Next? *Journal of Heredity* 102(1):130.

Bruno, W.; Socci, N. & Halpern A. (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution* 17(1):189.

Cavalli-Sforza, L. & Edwards, A. (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* 19(3 Pt 1):233.

Cole, J.; Wang, Q.; Cardenas, E.; Fish, J.; Chai, B.; Farris, R.; Kulam-Syed-Mohideen, A.; McGarrell, D.; Marsh, T.; Garrity, G. & others. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* 37(suppl 1):D141-D145.

Deschavanne, P.; Giron, A.; Vilain, J.; Fagot, G. & Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* 16(10):1391-9.

Edgar, R.(2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113

Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* 7:1-26.

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

Felsenstein, J. (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* 39 783-791.

Felsenstein, J.(1989). PHYLIP-phylogeny inference package (version 3.2). *Cladistics* 5:164-166

Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* 266:418-27.

Fitch, W. (1971) Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology* 20(4):406-416.

Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* 14(7):685-695.

Hall, T. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95-98.

Hendy, M. & Penny, D. (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* 59(2):277-290.

Henikoff, S. & Henikoff, J. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* 89(22):10915.

Jones, D.; Taylor, W. & Thornton, J. (1994) A mutation data matrix for transmembrane proteins. *FEBS Letters* 339(3):269-275.

Jukes, T. & Cantor, C. (1969) Evolution of protein molecules. In "*Mammalian Protein Metabolism*"(HN Munro, Ed.). Academic Press, New York.

Kaminuma, E.; Kosuge, T.; Kodama, Y.; Aono, H.; Mashima, J.; Gojobori, T.; Sugawara, H.; Ogasawara, O; Takagi, T.; Okubo, K. & others. (2011). DDBJ progress report. Nucleic Acids Research 39(suppl 1):D22-D27

Katoh, K.; Kuma, K.; Toh, H. & Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511 - 518

Kemena, C. & Notredame, C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25(19):2455-2465.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16(2):111-120.

Kolekar, P.; Kale, M. & Kulkarni-Kale, U. (2010) `Inter-Arrival Time' Inspired Algorithm and its Application in Clustering and Molecular Phylogeny. *AIP Conference Proceedings* 1298(1):307-312.

Konstantinidis, K. & Tiedje, J. (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology* 10(5):504-509.

Kuiken, C.; Leitner, T.; Foley, B.; Hahn, B.; Marx, P.; McCutchan, F.; Wolinsky, S.; Korber, B.; Bansal, G. & Abfalterer, W. (2009) HIV sequence compendium 2009. *Document LA-UR*:06-0680

Kuiken, C.; Yusim, K.; Boykin, L. & Richardson, R. (2005) The Los Alamos hepatitis C sequence database. *Bioinformatics* 21(3):379.

Kulkarni-Kale, U.; Bhosle, S.; Manjari, G. & Kolaskar, A. (2004) VirGen: a comprehensive viral genome resource. *Nucleic Acids Research* 32(suppl 1):D289.

Kumar, S.; Nei, M.; Dudley, J. & Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform*, 9(4):299-306.

Leinonen, R.; Akhtar, R.; Birney, E.; Bower, L.; Cerdeno-TÃ¡rraga, A.; Cheng, Y.; Cleland, I.; Faruque, N.; Goodgame, N.; Gibson, R. & others. (2011) The European Nucleotide Archive. *Nucleic Acids Research* 39(suppl 1):D28-D31.

Lio, P. & Goldman, N. (1998) Models of molecular evolution and phylogeny. *Genome Res* 8(12):1233-44.

Liu, K.; Linder, C. & Warnow, T. (2010) Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Currents* 2.

Luo, A.; Qiao, H.; Zhang, Y.; Shi, W.; Ho, S.; Xu, W.; Zhang, A. & Zhu, C. (2010) Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evolutionary Biology* 10(1):242.

Morgenstern, B.; French, K.; Dress, A. & Werner, T. (1998). DIALIGN: finding local similarities by multiple sequence alignment. *Bionformatics* 14:290 - 294

Needleman, S. & Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3):443-453.

Nicholas, H.; Ropelewski, A. & Deerfield DW. (2002) Strategies for multiple sequence alignment. *Biotechniques* 32(3):572-591.

Notredame, C.; Higgins, D. & Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302:205 - 217

Parry-Smith, D.; Payne, A.; Michie, A.& Attwood, T. (1998). CINEMA--a novel colour INteractive editor for multiple alignments. *Gene* 221(1):GC57-GC63

Pearson, W.; Robins, G. & Zhang, T. (1999) Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. *Molecular Biology and Evolution* 16(6):806.

Phillips, A. (2006) Homology assessment and molecular sequence alignment. *Journal of Biomedical Informatics* 39(1):18-33.

Ronquist, F. & Huelsenbeck, J. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574

Rzhetsky, A. (1995) Estimating substitution rates in ribosomal RNA genes. *Genetics* 141(2):771.

Saitou, N. & Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406-25.

Santos, C.; Ishida, M.; Foster, P.; Sallum, M.; Benega, M.; Borges, D.; Corrêa, K.; Constantino, C.; Afzal, M. & Paiva, T. (2008) Detection of a new mumps virus genotype during parotitis epidemic of 2006–2007 in the State of São Paulo, Brazil. *Journal of Medical Virology* 80(2):323-329.

Sayers, E.; Barrett, T.; Benson, D.; Bolton, E.; Bryant, S.; Canese, K.; Chetvernin, V.; Church, D.; DiCuccio, M.; Federhen, S. & others. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 39(suppl 1):D38-D51.

Schwartz, R. & Dayhoff, M. (1979) Matrices for detecting distant relationships. M. 0. Dayhoff (ed.), *Atlas of protein sequence and structure* 5:353-358.

Schmidt, H.; Strimmer, K.; Vingron, M. & Von Haeseler A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18(3):502.

Sims, G.; Jun, S.; Wu, G. & Kim, S. (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci U S A* 106(8):2677-82.

Snir, S. & Tuller, T. (2009) The net-hmm approach: phylogenetic network inference by combining maximum likelihood and hidden Markov models. *Journal of bioinformatics and computational biology* 7(4):625-644.

Sokal, R. & Michener, C. (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 38:1409-1438.

Tamura, K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution* 9(4):678-687.

Tamura, K. & Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10(3):512-526.

The UniProt Consortium. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research* 39(suppl 1):D214-D219.

Thompson, J.; Higgins, D. & Gibson, T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22(22):4673-80.

Thompson, J.; Linard, B.; Lecompte, O. & Poch, O. (2011) A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PLoS ONE* 6(3):e18093.

Thorne, J.; Goldman, N. & Jones, D. (1996) Combining protein evolution and secondary structure. *Molecular Biology and Evolution* 13(5):666-673.

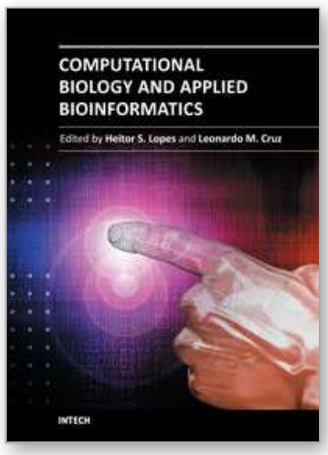
Vinga, S. & Almeida, J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics* 19(4):513-23.

Wilgenbusch, J. & Swofford, D. (2003). Inferring Evolutionary Trees with PAUP*. *Current Protocols in Bioinformatics*. 6.4.1–6.4.28

Wong, K.; Suchard, M. & Huelsenbeck, J. (2008) Alignment Uncertainty and Genomic Analysis. *Science* 319(5862):473-476.

Wong, W. & Nielsen, R. (2007) Finding cis-regulatory modules in Drosophila using phylogenetic hidden Markov models. *Bioinformatics* 23(16):2031-2037.

Xia, X. & Xie, Z. (2001). DAMBE: software package for data analysis in molecular biology and evolution. *Journal of Heredity* 92(4):371



Computational Biology and Applied Bioinformatics

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4
Hard cover, 442 pages
Publisher InTech
Published online 02, September, 2011
Published in print edition September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen