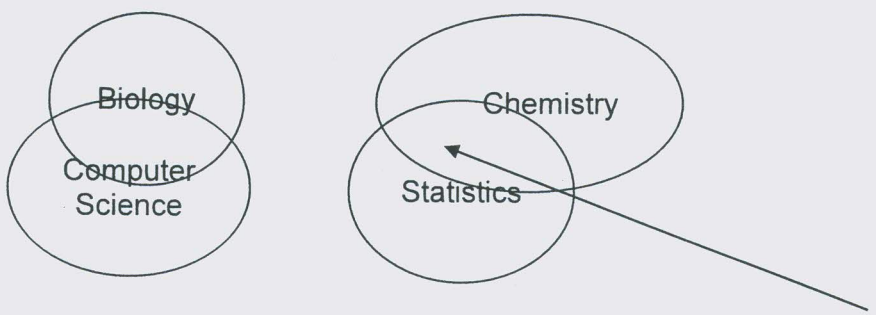
Bioinformatics in Microbial systematics

Murugadas.V and M. M. Prasad

ICAR- Central Institute of Fisheries Technology,
Cochin -29.
murugadascift81@gmail.com

Introduction:

Bioinformatics is a new science that uses computational approaches to answer biological questions. Bioinformatics is a new scientific discipline created from the interaction of biology and computer. Biological questions raised from the researchers will be investigated with the large & complex data sets available in public as well as generated by the own laboratory in private to arrive at a valid biological conclusions.



The National Center for Biotechnology Information (NCBI) defines bioinformatics as: "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline"

Broad Areas in Bioinformatics

- Genomics - Finding Genes
- Proteomics
- Phylogenetic
- Medical Implications

Some of the bioinformatics applicable are

⊙*Similarity search*

⊙*Sequence comparison: Alignment, multiple alignment, retrieval*

⊙*Sequences analysis: Signal peptide, transmembrane domain,…*

⊙*Protein folding: secondary structure from sequence*

⊙*Sequence evolution: phylogenetic trees*

Important terms in Bioinformatics

Fasta sequences

The FASTA format is used in a variety of molecular biology software suites. In its simplest incarnation (as shown above) the "greater than" character (>) designates the beginning of a new file. An identifier (L04459 in the first of the preceding examples) is followed by the DNA sequence in lowercase or uppercase letters, usually with 60 characters per line. Users and databases can then, if they wish, add a certain degree of complexity to this format. For example, without breaking any of the rules just outlined, one could add more information to the FASTA definition line, making the simple format a little more informative, as follows:

>gi|171361|gb|L04459|YSCCYS3A Saccharomyces cerevisiae cystathionine gamma-lyase (CYS3) gene, complete cds.
GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCG
AGTACCGCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAA
CGAGATTAGCG
Similarly the protein record in fasta as follows
>P31373
MTLQESDKFATKAIHAGEHVDVHGSVIEPISLSTTFKQSSPANPIGTYEYSRSQNPNRENLERAVAALENAQ
YGLAFSSGSATTATILQSLPQGSHAVSIGDVYGGTHRYFTKVANAHGVETSFTNDLLNDLPQLIKENTKLV
W
Majority of the procedure analysing either DNA or Protein sequences involves the use of fasta format

Application of Bioinformatics in Microbiology

Microbial systematics

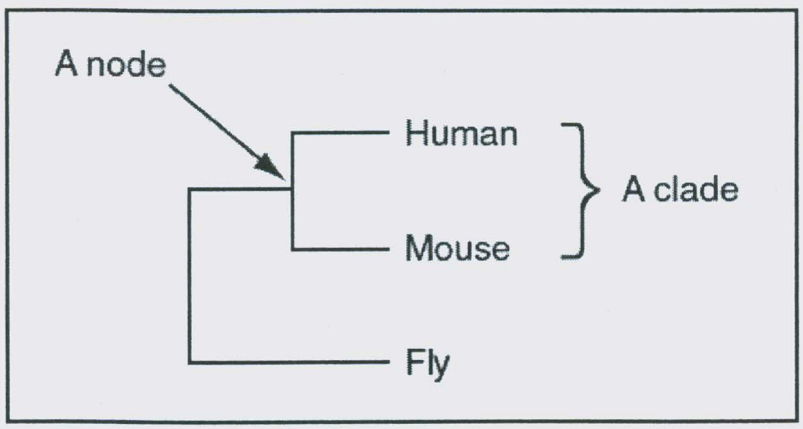
Identification

Phylogenetic tree construction

Phylogenetic

- The taxonomical system reflects evolutionary relationships
- Phylogenetics trees are things which reflect the evolutionary relationship through a picture/graph
- Rooted trees where there is only one ancestor
- Un rooted trees just showing the relationship
- Phylogenetic tree reconstruction algorithms are also an area of research

Phylogenetics is the study of evolutionary relationships. Phylogenetic analysis is the means of inferring or estimating these relationships. The evolutionary history inferred from phylogenetic analysis is usually depicted as branching, treelike diagrams that represent an estimated pedigree of the inherited relationships among molecules ("gene trees"), organisms, or both. Phylogenetics is sometimes called cladistics because the word "clade," a set of descendants from a single ancestor, is derived from the Greek word for branch. However, cladistics is a particular method of hypothesizing about evolutionary relationships.



• There are three basic assumptions in cladistics: Any group of organisms is related by descent from a common ancestor (fundamental tenet of evolutionary theory).

• There is a bifurcating pattern of cladogenesis. This assumption is controversial.

• Change in characteristics occurs in lineages over time. This is a necessary condition for cladistics to work. The resulting relationships from cladistic analysis are most commonly represented by a phylogenetic tree: A node even with this simple tree, a number of terms that are used frequently in phylogenetic analysis can be introduced:

• A **clade** is a monophyletic taxon. Clades are groups of organisms or genes that include the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor. Clade is derived from the Greek word "klados," meaning branch or twig.

• A **taxon** is any named group of organisms but not necessarily a clade. • In some analyses, **branch** lengths correspond to divergence (e.g., in the aboveexample, mouse is slightly more related to fly than human is to fly).

• A **node** is a bifurcating branch point. Macromolecules, especially sequences, have surpassed morphological and other organismal characters as the most popular form of data for phylogenetic or cladistic analysis. It is this molecular phylogenetic analysis that we will introduce here. It is unrealistic to believe that an all-purpose phylogenetic analysis recipe can be delineated. Although numerous phylogenetic algorithms, procedures, and computer programs have been devised, their reliability and practicality are, in all cases, dependent on the structure and size of the data. The merits and pitfalls of various methods are the subject of often acrimonious debates in taxonomic and phylogenetic journals. Some of these debates are summarized in a series of useful reviews of phylogenetics (Saitou, 1996; Li, 1997; Swofford et al., 1996). An especially concise introduction to molecular phylogenetics is provided by Hillis et al.(1993). The danger of generating incorrect results is inherently greater in computational phylogenetics than in many other fields of science. The events yielding a phylogeny happened in the past and can only be inferred or estimated. Despite the well-documented limitations of available phylogenetic procedures, current biological literature is replete with examples of conclusions derived from the results of analyses in which data had been simply run through one or another phylogeny program. Occasionally, the limiting factor in phylogenetic analysis is not so much the computational method used; more often than not, the limiting factor is the users' understanding of what the method is actually doing with the data.

This brief guide to phylogenetic analysis has several objectives. First, a conceptual approach that describes some of the most important principles underlying the most widely and easily applied methods of phylogenetic analyses of biological sequences and their interpretation will be introduced. The aim is to show that practical phylogenetic analysis should be conceived as a search for a correct model, as much as a search for the correct tree. In this context, some of the particular models assumed by various popular methods and how these models might affect analysis of particular data sets will be discussed. Finally, some examples of the application of particular methods to the inferences of evolutionary history are provided. Note that the principles for DNA analysis will be initially discussed, although most also apply to protein sequences (except where further description of protein sequences is indicated).

FUNDAMENTAL ELEMENTS OF PHYLOGENETIC MODELS

Phylogenetic tree-building methods presume particular evolutionary models. For a given data set, these models can be violated because of occurrences such as the transfer of genetic material between organisms. Thus, when interpreting a given analysis, one should always consider the model used and its assumptions and entertain other possible explanations for the observed results. As an example, consider the tree in Figure 14.1. An investigation of organismal relationships in the tree suggests the eukaryote 1 is more related to the bacteria than to the other eukaryotes. Because the vast majority of other cladistic

analyses, including those based on morphological features, suggest that eukaryote 1 is more related to the other eukaryotes than to bacteria, we suspect that for this analysis the assumptions of a bifurcating pattern of evolution are incorrect. We suspect that horizontal gene transfer from an ancestor of the bacteria 1, 2, and 3 to the ancestor of eukaryote 1 occurred because this would most simply explain the results.

Models inherent in phylogenetics methods make additional "default" assumptions:

1. The sequence is correct and originates from the specified source.
2. The sequences are homologous (i.e., are all descended in some way from a shared ancestral sequence).
3. Each position in a sequence alignment is homologous with every other in that alignment.
4. Each of the multiple sequences included in a common analysis has a common phylogenetic history
5. The sampling of taxa is adequate to resolve the problem of interest.
6. Sequence variation among the samples is representative of the broader group of interest.
7. The sequence variability in the sample contains phylogenetic signal adequate to resolve the problem of interest.

Example for the phylogenetic tree constructed out

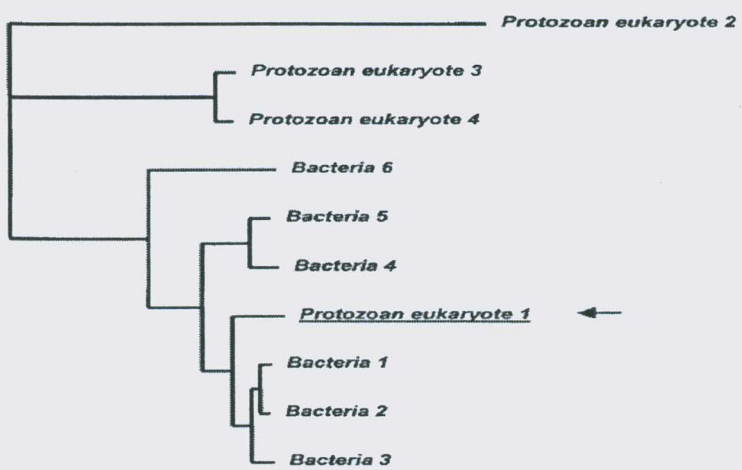


Figure 14.1. Example of a phylogenetic tree based on genes that do not match organismal phylogeny, suggesting horizontal gene transfer has occurred. The ancestor of protozoan eukaryote 1 (underlined and marked with an arrow) appears to have obtained the gene from the ancestor of Bacteria 1, 2, and 3, as this is the simplest explanation for the results. This unexpected result is not without precedent: there have been a number of reported phylogenetic analyses that suggest that protozoa have taken up genes from bacteria, most likely from bacteria that they have ingested.

PHYLOGENETIC DATA ANALYSIS: THE FOUR STEPS

A straightforward phylogenetic analysis consists of four steps:

1. Alignment (both building the data model and extracting a phylogenetic dataset)
2. Determining the substitution model
3. Tree building
4. Tree evaluation

Homologs are most commonly either orthologs, paralogs, or xenologs.

• *Orthologs* are homologs produced by speciation. They represent genes derived from a common ancestor that diverged due to divergence of the organisms they are associated with. *They tend to have similar function.*

• *Paralogs* are homologs produced by gene duplication. They represent genes derived from a common ancestral gene that duplicated within an organism and then subsequently diverged. *They tend to have different functions.*

• *Xenologs* are homologs resulting from horizontal gene transfer between two organisms. The determination of whether a gene of interest was recently transferred into the current host by horizontal gene transfer is often difficult. Occasionally, the %(G _ C) content may be so vastly different from the average gene in the current host that a conclusion of external origin is nearly inescapable, however often it is unclear whether a gene has horizontal origins. Function of xenologs can be variable depending on how significant the change in context was for the horizontally moving gene; however, in general, the function tends to be similar.

ALIGNMENT: BUILDING THE DATA MODEL

Phylogenetic sequence data usually consist of multiple sequence alignments; the individual, aligned-base positions are commonly referred to as "sites." These sites are equivalent to "characters" in theoretical phylogenetic discussions, and the actual base (or gap) occupying a site is the "character state." Multiple alignment methods are reviewed in Chapter 9. This chapter reviews similar alignment methods in the context of phylogenetic analysis. Aligned sequence positions subjected to phylogenetic analysis represent a priori phylogenetic conclusions because the sites themselves (not the actual bases) are effectively assumed to be genealogically related, or homologous. Sites at which one is confident of homology and that contain changes in character states useful for the given phylogenetic analysis are often referred to as "informative sites." Steps in building the alignment include selection of the alignment procedure(s) and extraction of a phylogenetic data set from the alignment. The latter procedure requires determination of how ambiguously aligned regions and insertion/deletions (referred to as *indels*, or gaps) will be treated in the tree-building procedure. A typical alignment procedure involves the application of a program such as CLUSTAL W, followed by manual alignment editing and submission to a treebuilding program.

ALIGNMENT: EXTRACTION OF A PHYLOGENETIC DATA SET

In alignments that include length variation, the phylogenetic data set is usually not identical to the alignment. Even in alignments of length-invariable sequences, the data set can be different—for example, when only first and second codon positions are to be analyzed to avoid the strong G _ C bias in the third codon position from affecting the final results. In summary, the following points should be considered when constructing a multiple sequence alignment for a phylogenetic analysis:

• The alignment step in phylogenetic analysis is one of the most important because it produces the data set on which models of evolution are used.

• It is not uncommon to edit the alignment, deleting unambiguously aligned regions and inserting or deleting gaps to more accurately reflect probable evolutionary processes that led to the divergence between sequences.

• It is useful to perform phylogenetic analyses based on a series of slightly modified alignments to determine how ambiguous regions in the alignment affect the results and what aspects of the results one may have more or less confidence in.

TREE-BUILDING METHODS

Tree-building methods implemented in available software are discussed in detail in the literature (Saitou, 1996; Swofford et al., 1996; Li, 1997) and described on the Internet. This section briefly describes some of the most popular methods. Treebuilding methods can be sorted into distance-based vs. character-based methods. Much of the discussion in molecular phylogenetics dwells on the utility of distanceand character-based methods (e.g., Saitou, 1996; Li, 1997). Distance methods compute pairwise distances according to some measure and then discard the actual data, using only the fixed distances to derive trees. Character-based methods derive trees that optimize the distribution of the actual data patterns for each character. Pairwise distances are, therefore, not fixed, as they are determined by the tree topology. The most commonly applied distance-based methods include neighbor-joining and the Fitch-Margoliash method, and the most common character-based methods include maximum parsimony and maximum likelihood.

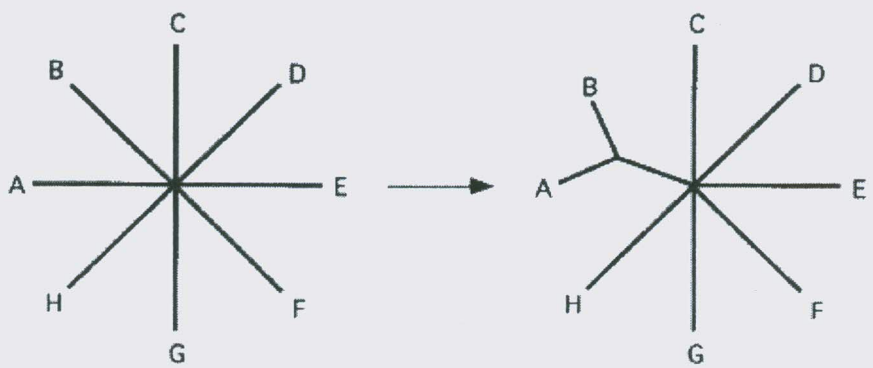


Figure 14.8. Star decomposition. This is how tree-building algorithms such as neighbor-joining work. The most similar terminals are joined, and a branch is inserted between them and the remainder of the star. Subsequently, the new branch is consolidated so that its value is a mean of the two original values, yielding a star tree with n-1 terminals. The process is repeated until only one terminal remains.

Distance-Based Methods

Distance-based methods use the amount of dissimilarity (the distance) between two aligned sequences to derive trees. A distance method would reconstruct the true tree if all genetic divergence events were accurately recorded in the sequence (Swofford et al., 1996). However, divergence encounters an upper limit as sequences become mutationally saturated. After one sequence of a diverging pair has mutated at a particular site, subsequent mutations in either sequence cannot render the sites any more "different." In fact, subsequent mutations can make them again equal (for example, if a valine mutates to an isoleucine, which mutates back to a valine).

Therefore, most distance-based methods correct for such "unseen" substitutions. In practice, application of the rate matrix effectively presumes that some proportion of observed pairwise base identities actually represents multiple mutations and that this proportion increases with increasing overall sequence divergence. Some programs implement, at least optionally, calculation of uncorrected distances, whereas, for example, the MEGA program (Kumar et al., 1994) implements only uncorrected distances for codon and amino acid data. Unless overall divergences are very low, the latter approach is virtually guaranteed to give inaccurate results. Pairwise distance is calculated using maximum-likelihood estimators of substitution rates. The most popular distance tree-building programs have a limited number of substitution models, but PAUP 4.0 implements a number of models, including the actual model estimated from the data using maximum likelihood, as well as the logdet distance method.

Distance methods are much less computationally intensive than maximum likelihood but can employ the same models of sequence evolution. This is their biggest advantage. The disadvantage is that the actual character data are discarded. The most commonly applied distance-based methods are the unweighted pair group method with arithmetic mean (UPGMA), neighbor joining (NJ), and methods that optimize the additivity of a distance tree, including the minimum evolution (ME) method. Several methods are available in more than one phylogenetics software package but not all implementations allow the same parameter specifications and/or tree optimization features (e.g., branch swapping; see below).

Unweighted Pair Group Method with Arithmetic Mean (UPGMA).

UPGMA is a clustering or phenetic algorithm—it joins tree branches based on the criterion of greatest similarity among pairs and averages of joined pairs. It is not strictly an evolutionary distance method (Li, 1997). UPGMA is expected to generate an accurate topology with true branch lengths only when the divergence is according to a molecular clock (ultrametric; Swofford et al., 1996) or approximately equal to raw sequence dissimilarity. As mentioned earlier, these conditions are rarely met in practice

Neighbor Joining (NJ). The neighbor-joining algorithm is commonly applied with distance tree building, regardless of the optimization criterion. The fully resolved tree is "decomposed" from a fully unresolved "star" tree by successively inserting branches between a pair of closest (actually, most isolated) neighbors and the remaining terminals in the tree (Fig. 14.8). The closest neighbor pair is then consolidated, effectively reforming a star tree, and the process is repeated. The method is comparatively rapid.

Fitch-Margoliash (FM). The Fitch-Margoliash (FM) method seeks to maximize the fit of the observed pairwise distances to a tree by minimizing the squared deviation of all possible observed distances relative to all possible path lengths on the tree (Felsenstein, 1997). There are several variations that differ in how the error is weighted. The variance estimates are not completely independent because errors in all the internal tree branches are counted at least twice (Rzhetsky and Nei, 1992).

Minimum Evolution (ME). Minimum evolution seeks to find the shortest tree that is consistent with the path lengths measured in a manner similar to FM; that is, ME works by minimizing the squared deviation of observed to tree-based distances (Rzhetsky and Nei, 1992; Swofford et al., 1996; Felsenstein, 1997). Unlike FM, ME does not use all possible pairwise distances and all possible associated tree path lengths. Rather, it fixes the location of internal tree nodes based on the distance to external nodes and then optimizes the internal branch length according to the minimum measured error between these "observed" points. It thus purports to eliminate the non-independence of FM measurements.

Which Distance-Based Tree-Building Procedure Is Best?

ME and FM appear to be the best procedures, and they perform nearly identically in simulation studies (Huelsenbeck, 1995). ME is becoming more widely implemented in computer programs, including METREE (Rzhetsky and Nei, 1994) and PAUP. For protein data, the FM procedure in PHYLIP offers the greatest range of substitution models but no correction for among-site rate heterogeneity. The MEGA (Kumar et al., 1994) and METREE packages include a gamma correction for proteins, but only in conjunction with a raw ("*p*-distance") divergence model (no distance or bias correction), which is unreliable except for small divergences (Rzhetsky and Nei, 1994). MEGA also computes separate distances for synonymous and nonsynonymous sites, but this method is valid only in the absence of substitution or base frequency bias and when there is no correction for among-site rate heterogeneity. Thus, for most data sets, using the nucleotide data under a more realistic model might be preferable to MEGA's methods.

Simulation studies indicate that UPGMA performs poorly over a broad range of tree shape space (Huelsenbeck, 1995). The use of this method is not recommended; it is mentioned here only because its

application seems to persist, as evidenced by UPGMA gene trees appearing in publications (Huelsenbeck, 1995). NJ is clearly the fastest procedure and generally yields a tree close to the ME tree. (Rzhetsky and Nei, 1992; Li, 1997). However, it yields only one tree. Depending on the structure of the data, numerous different trees might be as good or significantly better than the NJ tree (Swofford et al., 1996).

Character-Based Methods

The character-based methods have little in common with each other, besides the use of the character data at all steps in the analysis. This allows the assessment of the reliability of each base position in an alignment on the basis of all other base positions.

Maximum Parsimony (MP).

Maximum parsimony is an optimization criterion that adheres to the principle that the best explanation of the data is the simplest, which in turn is the one requiring the fewest ad hoc assumptions. In practical terms,the MP tree is the shortest—the one with the fewest changes—which, by definition, is also the one with the fewest parallel changes. There are several variants of MP that differ with regard to the permitted directionality of character state change (Swofford et al., 1996). To accommodate substitution bias, MP is amenable to weighting; for example,the transformation of a transversion can be weighted relative to a transition (see above). The easiest way to do this is to create a weighting step matrix in which theweights are the reciprocal of the rates estimated using ML as described above. However, step-matrix weighting can greatly slow MP computation.The MP method performs poorly when there is substantial among-site rate heterogeneity (Huelsenbeck, 1995). There are few good fixes for this problem. One approach is to modify the data set to include only sites that exhibit little or no heterogeneity as determined by likelihood estimation (see above). Another approach is to recursively reweight positions according to their propensity to change as observed in preliminary trees. This "successive approximations" approach is automatically facilitated in PAUP, but it is prone to error to the degree that the preliminary trees are incorrect.

MP analyses tend to yield numerous (and sometimes many thousands of) trees that have the same score. Because each is held to be as optimal as any other, only groupings present in the strict consensus of all trees are considered to be supported by the data. The reason that distance and ML tree methods tend to arrive at a single best tree is that their calculations involve division and decimals, whereas MP merely counts discrete steps. For a given data set, a strict consensus of all ME or ML trees that are not significantly worse than optimal probably would yield resolution more or less comparable to the MP consensus. Unfortunately, whereas MP users conventionally present strict consensus (and sometimes consensus of trees one or two steps worse), ME and ML users typically do not. Simulation studies have shown that MP performs no better than ME and worsethan ML when the amount of sequence evolution since lineages diverged is much greater than the amount of divergence that occurred between lineage splits (i.e., in a tree with very long terminal branches and short internal internodes) (Huelsenbeck, 1995). This condition produces "long branch attraction"—the long branches become artificially connected because the number of on homologous similarities the sequences have accumulated exceeds the number of homologous similarities they have retained with their true closest relatives (Swofford et al., 1996). Character weighting improves the performance of MP under these conditions (Huelsenbeck, 1995).

Maximum Likelihood (ML).

ML turns the phylogenetic problem inside out. ML searches for the evolutionary model, including the tree itself, that has the highest likelihood of producing the observed data. In practice, ML is derived for each base position in an alignment. The likelihoodis calculated in terms of the probability that the pattern of variation at a site would be produced by a particular substitution process, given a particular tree and the

overall observed base frequencies. The likelihood becomes the sum of the probabilities of each possible reconstruction of substitutions under a particular substitution process. The likelihoods for all the sites are multiplied to give an overall "likelihood of the tree" (i.e., the probability of the data given the tree and the substitution process). As one can imagine, for one particular tree, the likelihood of the data is low at some sites and high at others. For a "good" tree, many sites will have higher likelihood, so the product of likelihoods is high. For a "poor" tree, the reverse will be true. The substitution model should be optimized to fit the observed data. For example, if there is a transition bias, evident by an inordinate number of sites that include only purines or pyrimidines, the likelihood of the data under a model that assumes no bias will never be as good as one that does. Likewise, if a substantial proportion of the sites are occupied by a single base and another substantial proportion have equal base frequencies, the likelihood of the data under a model that assumes that all sites evolve equally will be less than that of a model that allows rate heterogeneity. Modifying the substitution parameters, however, modifies the likelihood of the data associated with particular trees. Thus, the tree yielding the highest likelihood under one substitution model might yield much lower likelihood under another. Because ML uses great amounts of computational time, it is usually impractical to perform a complete search that simultaneously optimizes the substitution model and the tree for a given data set. An economical, heuristic approach is recommended (Adachi and Hasegawa, 1996; Swofford et al., 1996). Perhaps the best time saver in this regard is preliminary ML estimation of the substitution model (as can be performed using PAUP). This procedure can be applied iteratively, searching for better ML trees, then reestimating the parameters, and then searching for better trees. As algorithms, computers, and phylogenetic understanding have improved, the ML criterion has become more popular for molecular phylogenetic analysis. In simulation studies, ML has consistently outperformed ME and MP when the data analysis proceeds according to the same model that generates the data (Huelsenbeck, 1995). ML will always be the most computationally intensive method of all, however, so there will always be situations in which it is not practical.

DISTANCE, PARSIMONY, AND MAXIMUM LIKELIHOOD: WHAT'S THE DIFFERENCE?

Distance matrix methods simply count the number of differences between two sequences. This number is referred to as the evolutionary distance, and its exact size depends on the evolutionary model used. The actual tree is then computed from the matrix of distance values by running a clustering algorithm that starts with the most similar sequences (i.e., those that have the shortest distance between them) or by trying to minimize the total branch length of the tree. The principle of maximum parsimony searches for a tree that requires the smallest number of changes to explain the differences observed among the taxa under study. A maximum-likelihood approach to phylogenetic inference evaluates the probability that the chosen evolutionary model has generated the observed data. The evolutionary model could simply mean that one assumes that changes between all nucleotides (or amino acids) are equally probable. The program will then assign all possible nucleotides to the internal nodes of the tree in turn and calculate the probability that each such sequence would have generated the data (if two sister taxa have the nucleotide "A," a reconstruction that assumes derivation from a "C" would be assigned a low probability compared with a derivation that assumes there already was an "A"). The probabilities for all possible reconstructions (not just the more probable one) are summed up to yield the likelihood for one particular site. The likelihood for the tree is the product of the likelihoods for all alignment positions in the data set.

Searching for Trees

The number of unique phylogenetic trees increases exponentially with the number of taxa, becoming astronomical even for, say, 50 sequences (Swofford et al., 1996;Li, 1997). In most cases,

computational limitations permit exploration of only a small fraction of possible trees. The exact number will depend mainly on the number of taxa, the optimality criterion (e.g., MP is much faster than ML), the parameters (e.g., unweighted MP is much faster than weighted; ML with fewer preset parameters is much faster than with more and/or simultaneously optimized parameters), computer hardware, and computer software (some algorithms are faster than others; some software allows multiprocessing; some software limits the number and kind of trees that can be stored in memory). The search procedure is also affected by data structure: poorly resolvable data produce more "nearly optimal" trees that must be evaluated to find the most optimal. Branch-swapping algorithms successively modify existing trees built by an initial step (Swofford et al., 1996). The algorithms range from those that generate all possible unique trees (exhaustive algorithms) to those that evaluate only minor modifications. Quartet puzzling is a relatively rapid tree-searching algorithm available for ML tree building (Strimmer and von Haeseler, 1996) and is available in PUZZLE. One of the best ways to economize the search effort is to prune the data set. For example, it might be apparent from the data alone or from preliminary searching hat a particular cluster of five terminals is unresolvable, that the arrangement of these terminals does not impact the remainder of the topology, and/or that resolution of these terminals is not the objective of the analysis. Removing four of the terminals from the analysis simplifies the search by several orders of magnitude. Every analysis is unique. The elements that influence the choice of optimal search strategy (amount of data, structure of data, amount of time, hardware, objective of analysis) are too variable to suggest a foolproof recipe. Thus, researchers must be familiar with their data; they must also have specific objectives in mind, understanding the various search procedures as well as the capabilities of their hardware and software.

Rooting Trees

The methods described above produce unrooted trees (i.e., trees having no evolutionary polarity). To evaluate evolutionary hypotheses, it is often necessary to locate the root of the tree. Rooting phylogenetic trees is not a trivial problem (Nixon and Carpenter, 1993). If one accepts a molecular clock, then the root will alw ays be at the midpoint of the longest span across the tree (Weston, 1994). Whether molecular evolution is indeed clocklike generally remains a contentious issue (Li, 1997), but most gene trees exhibit unclocklike behavior regardless of where the root is placed. Thus, rooting is generally evaluated by extrinsic evidence, that is, by means of determining where the tree would attach to an "outgroup," which can be any organism/sequence not descended from the nearest common ancestor of the organisms/sequences analyzed (for example, a bird sequence could be used to root an analysis of mammals). Outgroup rooting, however, creates a dilemma: an outgroup that is closely related to the ingroup might be simply an erroneously excluded member of the ingroup. A clearly distant outgroup (e.g., a fungus for an analysis of plants) can have a sequence so diverged that its attachment to the ingroup is subject to the long-branch attraction problem mentioned above. It is wise to examine the results obtained for trees both with and without an outgroup. Another means of rooting involves analysis of a duplicated gene or gene with an internal duplication (Lawson et al., 1996). If all the paralogs from most or all of the organisms are included in the analysis, then one can logically root the tree exactly where the paralog gene trees converge, assuming that there are not long branch problems in all trees.

TREE EVALUATION

Several procedures are available that evaluate the phylogenetic signal in the data and the robustness of trees (Swofford et al., 1996; Li, 1997). The most popular of the former class are tests of data signal versus randomized data (skewness and permutation tests). The latter class includes tests of tree support from resampling of observed data (nonparametric bootstrap). The likelihood ratio test provides a means of evaluating both the substitution model and the tree.

Randomized Trees (Skewness Test)

Simulation studies indicate that the distribution of random MP tree lengths generated using random data sets will be symmetrical, whereas those using data sets with phylogenetic signal will be skewed. The critical value of the g1 statistic of skewness will vary with the number of taxa and variable sites in the sequence. The test does not estimate the reliability of a particular topology, and it is sensitive to even very small amounts of signal present in an otherwise random data set. If taxa from groups that are obviously well supported by the data are selectively deleted, the test can be used to determine whether a phylogenetic signal remains, provided at least 10 variable characters and 5 taxa are examined. The procedure is implemented in PAUP.

Randomized Character Data (Permutation Tests)

The randomized data approach determines whether an MP tree or portion of it derived from the actual data could have arisen by chance. The data are not truly randomized but permuted within each aligned column, so that covariation in the initial data is removed. The result is an alignment of sequences that are not random sequences; rather, the base at each site in these sequences is randomly drawn from the population of bases occupying that site in the overall alignment. The permutation tail probability test (PTP) compares the score for the MP tree with trees generated by numerous permutations of the data at each site, determining only whether there is a phylogenetic signal in the original data. A topology-dependent test (T-PTP) compares the scores for specific trees to determine whether the difference can be attributed to chance. This method does not evaluate whether the tree or any portion of it is correct (Swofford et al., 1996). In particular, the T-PTP test will appear to corroborate groups that are in trees close to the MP tree but not in it. This is because the method detects the collective signal that places a taxon even approximately, if not actually, in its correct position. The results can be fine-tuned, however, by additional applications using relevant subsets of the data (Faith and Trueman, 1996). The procedure is implemented in PAUP.

Bootstrap

Bootstrapping is a resampling tree evaluation method that works with distance, parsimony, likelihood, and just about any other tree derivation method. It was invented in 1979 (Efron, 1979) and introduced as a tree evaluation method in phylogenetic analysis by Felsenstein (1985). The result of bootstrap analysis is typically a number associated with a particular branch in the phylogenetic tree that gives the proportion of bootstrap replicates that supports the monophyly of the clade. How is this done practically? Bootstrapping can be considered a two-step process comprising the generation of (many) new data sets from the original set and the computation of a number that gives the proportion of times that a particular branch (e.g., a taxon) appeared in the tree. That number is commonly referred to as the bootstrap value. New data sets are created from the original data set by sampling columns of characters at random from the original data set with replacement. "With replacement" means that each site can be sampled again with the same probability as any of the other sites. As a consequence, each of the newly created data sets has the same number of total positions as the original data set, but some positions are duplicated or triplicated and others are missing. It is therefore possible that some of the newly created data sets are completely identical to the original set—or, on the other extreme, that only one of the sites is replicated, say, 500 times, whereas the remaining 499 positions in the original data set are dropped. Although it has become common practice to include bootstrapping as part of a thorough phylogenetic analysis, there is some discussion on what exactly is measured by this method. It was originally suggested that the bootstrap value is a measure of repeatability (Felsenstein, 1985). In more recent interpretations, it has been considered to be a measure of accuracy—a biologically more relevant parameter that gives the

probability that the true phylogeny has been recovered. On the basis of simulation studies, it has been suggested that, under favorable conditions (roughly equal rates of change, symmetric branches), bootstrap values greater than 70% correspond to a probability of greater than 95% that the true phylogeny has been found (Hillis and Bull, 1993). By the same token, under less favorable conditions, bootstrap values greater than 50% will be overestimates of accuracy (Hillis and Bull, 1993). Simply put, under certain conditions, high bootstrap values can make the wrong phylogeny look good; therefore, the conditions of the analysis must be considered. Bootstrapping can be used in experiments in which trees are recomputed after internal branches are deleted one at a time. The results provide information on branching orders that are ambiguous in the full data set (cf. Leipe et al., 1994).

Parametric Bootstrap

The parametric bootstrap differs from the nonparametric in that it uses simulated (yet actual) replicates rather than pseudoreplicates. In the case of phylogenetic sequence analysis, replicate data sets of the same size as the original data set are generated according to a specified model of sequence evolution, including the optimal tree topology determined according to that model (Huelsenbeck et al., 1996). Each data set is then analyzed according to the method of interest. Support for the branches in the test tree can be determined in much the same way as in the nonparametric bootstrap.

Likelihood Ratio Tests

As the name implies, likelihood ratio tests are applicable to ML analyses. A suboptimal likelihood value is evaluated for significance against a normal distribution of the error in the optimal model. In ideal applications, the error curve is presumed to be a _ 2 distribution. Thus, the test statistic is twice the difference between the optimal and test values, and the degrees of freedom is the number of parameter differences. Application of the _ 2 test to alternative phylogenetic trees is problematic, especially because of the "irregularity of [the] parameter space" (Yang et al., 1995), but its use has been advocated for evaluating optimality of the substitution model when the number of parameters between models is known.

Conclusion

Bioinformatics is an important field which is now useful in microbiology for identification of bacteria, identification of gene seqencing having mutations, comparing few microbes based on expression of certain genes and so on. The field has grown upto the stage of whole genome sequence analysis of a microbe, metagenomic analysis of microbial population at a given niche, microbiome of given niche, microbial community profiling, fingerprinting of bacterial strains for understanding their evolutionary relationship, transcriptome analysis of bacteria in different physiological conditions etc.

Reference

1. Andreas D. Baxevanis and B. F. Francis Ouellette. BIOINFORMATICS-A Practical Guide to the Analysis of Genes and Proteins. SECOND EDITION. Wileys Inerscience, A JOHN WILEY & SONS, INC., PUBLICATION. 2001.
2. Des Higgins and Willie Taylor. Bioinformatics-Sequences, structure, and databanks- A practical approach. Oxford University Press. 2000.
3. Andrzej Polanski and Marek Kimmel. Bioinformatics. Springer-Verlag Berlin Heidelberg 2007