Chapter 1

Introduction

Evolution is the development of increasingly complex forms of life from simpler and earlier forms [Wikipedia, 2009c]. According to Charles Darwin [Darwin, 1859, Wikipedia, 2009a], Evolution is the descent with modifications. Most earlier evolutionary theories suffered the weakness of being based on a single factor and excluding all others [Mayr, 1970]. Darwin [1859] suggested four important features of evolution, namely, natural selection, sexual selection, artificial selection and the theory of pangenesis [Mayr, 1970, Wikipedia, 2009a]. The Natural selection theory explains about how nature selects the fittest species whereas sexual selection emphasizes the development of attractive characters in males. In contrast, the artificial selection focuses on how better breeds are artificially selected in domestic animals. According to the theory of pangenesis, characters are inherited from the parent to the offspring. When Darwin proposed his theory of evolution by natural selection he was ignorant of the source of variation and it is only after the discovery of Mendel’s laws and the demonstration that the genetic variations are generated by mutations, the Synthetic theory of Evolution or Neo-Darwinism emerged [Mayr, 1970, Graur and Li, 2000, Page and Holmes, 1998]. Nevertheless, it is remarkable that Darwin inadvertently used polygenic traits or characters to describe the hierarchy in complexity. Darwin's theory of evolution received remarkable support from various approaches, namely, palaeontological, genetic, biochemical and molecular [Kimura, 1983]. Furthermore, the distinction between qualitative and quantitative traits has also amplified our understanding of the process.
Palaeontological inference

Time-dependent evolutionary trends and rates are investigated by Paleontology, which sheds light on the origin of new classes, phyla and higher taxa. Fossil records have solved many evolutionary problems and were used to establish the age and the origin of different lineages [Mayr, 1970]. The oldest fossil for any given taxon tells us that the lineage must be at least as old and that an older fossils may still remain be discovered [Graur and Li, 2000, Page and Holmes, 1998]. Combining phylogenies and fossil record is a powerful approach to calibrate rates of molecular evolution. However, estimates related to extinct organisms must be treated with extreme care because most of the earth's earlier history is still not fully documented as even for the existing fossil records there exist many gaps. Consequently, the classification based on fossil records is biased in favor of more recent taxa because of their over-representation. These records also miss some lineages. In any case, fossils are important indicators of phylogenetic relationships because of their time-specified position between the ancestors and extant species. These ancestral groups can be identified from the available fossil records. At the same time, it is suggested [Page and Holmes, 1998] that fossils cannot be treated as ancestors of other taxa but may actually represent terminal taxa in phylogeny and, therefore, may not be direct ancestors of the modern living taxa. Nevertheless, fossils exhibiting various combinations of characters that are also found in modern groups can contribute to phylogenetic analysis [Kemp and Kemp, 1999, Smith and Szathmáry, 1997].

Genetic inference

Genotypes interact with the environment and produce phenotypes [Page and Holmes, 1998]. Genotypes are the result of combination, recombination and mutation of genes in the gene pool of a population. Genetic variation in a population is controlled by factors such as the
input of new genetic material through mutation, the erosion of this variation by selection and protection of the stored variability. All natural populations contain abundant genetic variation that serves as the raw material for evolutionary change [Mayr, 1970]. Genetic inferences can also be drawn from their identity and exclusivity through sexual reproduction. The rapidly expanding information on the structure and contents of genomes would further allow drawing specific genetic inferences.

Biochemical inference

There exists evidence on fossil bacteria that resemble modern bacteria. But, despite the existence of sufficient indicators on the pre-biotic and the present environment, a number of issues have to be resolved before a direct genealogical and evolutionary relationship can be derived among bacteria [Woese, 1987]. For example, it is not yet certain as to which among the key molecules, e.g., nucleic acids, proteins and carbohydrates were formed the first in response to the effect of sun’s radiation on hydrogen, ammonia, methane and water in Earth’s atmosphere. The temporal hierarchy among simple molecular is yet to be understood. Similarly, we do not understand the issue of the reiterative copying of such molecules, the process by which information became permanently bestowed on them and the perpetuation of their chemical identity through replication, all of which constitute a minimum requirement for the beginning of evolution [Lazcano et al., 1996]. In any case, the repetitive appearance in wide variety of taxa of the same or similar specific catalytic entities or enzymes positioned in discernible metabolic pathways, constitutes excellent evidence for evolutionary conservation [Nelson et al., 2004]. Similarly, adaptation of enzymes to varying physical conditions is suggestive of the way natural selection works. The discovery [Woese, 1967, Ridley, 2004, ] that RNA has some the catalytic activity to replicate is suggestive of RNA being the earliest replicating molecule. As RNA is the intermediate to DNA and protein in the Central Dogma,
it is likely that it also served as a link in transferring the replicating ability and information storage to DNA and the catalytic replicase activity to amino acids in proteins. It is well known that the energy from sunlight and chemical sources can drive biochemical reactions [Nelson et al., 2004]. In many conditions, biosynthesis is driven when coupled to process which releases energy by molecules like adenosine triphosphate (ATP). It is likely that the development of energy conversion processes and biosynthetic pathways were precursors for the evolution of a wide variety of unicellular organisms. The final stage of evolution would have begun when self-replicating functional entities like cells began adjusting to changing environments, became multi-cellular and then became subject of division of labor among different members of the community [Smith, 1998, Parsons et al., 1998, Smith et al., 1999, Wilson et al., 1977].

**Molecular inference**

Recent advances in sequences DNA, RNA and polypeptides has led to development of large data-bases containing sequences of same genes or polypeptides from a wide variety of organisms [Lesk, 2002, Mount, 2004]. This has created a huge opportunity to compare sequences of multiple molecular strings by aligning each monomer in the same order so as to examine the sequence identity or divergence for the same functional gene or phene. A variety of methods are available to objectively study similarity or dissimilarity among sequences by multiple sequence alignment of regions that are functionally important as well as those regions which ensure the structural backbone of a polypeptide (see, Chapter 6). Genome sequence data is mounting every day and sequences for more than 300 eukaryotes and 1500 prokaryotes are available [Genomes, 2009, NCBI, 2009] that can be compared to ascertain evolutionary distances among a variety of taxa.
Qualitative and quantitative traits

Many traits/characters are useful in phylogenetic studies. These characters are either qualitative or quantitative. The character state of a quantitative trait (e.g., height) is usually continuous and measurable on an interval scale where as the character state of qualitative trait is discrete. In molecular data the qualitative characters are usually considered as the presence or absence of a particular molecule [Graur and Li, 2000, Page and Holmes, 1998]. For example, nucleotide or amino acid sequence data allows examination of qualitative differences that lead to dramatic changes in the evolution of any given trait/s.

Phylogenetic Tree Building Methods

There exists several methods for building tree or relationship dendrograms and I have listed most of them (see Figure 1.1). These methods require comparison of molecular sequences (protein/nucleic acids) or distances to construct a phylogenetic tree. Furthermore, data from morphology, biochemical, biophysical can also be used to build trees.
### Tree Building Methods

<table>
<thead>
<tr>
<th>Data type</th>
<th>Optimality criterion</th>
<th>Clustering algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>Maximum likelihood</td>
<td>UPGMA</td>
</tr>
<tr>
<td></td>
<td>Maximum parsimony</td>
<td>Neighbor joining</td>
</tr>
<tr>
<td>Distances</td>
<td>Minimum evolution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Least squares</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.1: Different types of tree building methodology and their classification based on input data type and algorithms used to construct phylogenetic trees

**Unweighted Pair Group Method with Arithmetic mean (UPGMA)**

UPGMA is the simplest of all tree building (Sneath and Sokal, 1962, Wikipedia, 2009g) methods, based on 'all pairs distances' in a matrix format. It needs a distance matrix as input and it uses an algorithm (Sneath and Sokal, 1962) to construct a rooted phylogenetic tree. The algorithm examines the distance between all pairs and the minimum distances among all pairs of species from data. It then tabulates those with shortest distances as the terminal sets of pairs and assigns these as terminal nodes. The distance between the ancestor and all specie in the pair is the half of the minimum distance. Then the pair is made into a clade and a new distance matrix is formed in which the pair is merged. In the second round, it reexamines all thus reassigned terminal nodes to identify the sets internal nodes with shortest distances between each pair and assigns the second set of internal nodes. The process is reiterated till the distance matrix has only two entries (2 x 2 matrix). In this method it is assumed that...
the species evolve at the same evolutionary rate, i.e. the distance from root to all species are the same.

**Neighbor Joining (NJ)**

NJ is a clustering algorithm (Saitou and Nei, 1987, Wikipedia, 2009f] that resembles UPGMA except that the distance estimation is different. NJ requires a distance matrix to construct an unrooted phylogenetic tree by NJ algorithm (Saitou and Nei, 1987). In this method it is assumed that, unlike UPGMA, there is no uniform rate of evolution for parameters under consideration. Therefore, the distance from the root to all species may not be the same. Indeed, different species do not necessarily evolve at the same rate, thereby rendering this method as more accurate than UPGMA. NJ is a widely used distance based tree building method, but leads to an unrooted tree. In this tree a root can be tentatively assigned to the position of the outgroup.

**Minimum Evolution (ME)**

ME needs a distance matrix as input and uses a heuristic algorithm [Kidd and Sgaramella-Zonta, 1971, Rzhetsky and Nei, 1993] to construct an unrooted phylogenetic tree. A heuristic method uses the trial and error strategy for problem solving and leads to the best possible or optimal solution. ME assumes that the tree with the smallest sum of branch length is most likely to be the true one [Rzhetsky and Nei, 1993]. This is more accurate among distance based methods, but computationally time consuming. It is claimed that ME algorithm is slightly different from NJ, but the resulting trees are highly similar [Rzhetsky and Nei, 1993]. ME tree is also unrooted.
Maximum Likelihood (ML)

Maximum likelihood estimation (ML) is a popular statistical method, developed by R. A. Fisher to obtain best fits of data to a mathematical model. Here you first assume that the data set exhibits a normal distribution but with unknown mean and variance, so that the mean and variance exhibit the maximum likelihood for the total data set. For a fixed set of data and underlying probability model, maximum likelihood gives the most probable values [Wikipedia, 2009d]. ML uses nucleotide or amino acid sequences to construct unrooted phylogenetic trees by heuristic algorithm [Wikipedia, 2009b]. Sometimes ML yields more than one tree that have equal likelihood. In addition to the sequences it also requires substitution model to construct a phylogenetic tree. ML is useful to analyze both closely related and distantly related species. ML is probably the most accurate method, but computationally expensive [Hasegawa and Yano, 1984, Hasegawa et al., 1985, Kishino and Hasegawa, 1989].

Maximum Parsimony (MP)

Parsimony is a character-based non-parametric statistical method to construct a phylogenetic tree based on the least evolutionary change and uses a matrix of discrete phylogenetic characters to infer one or more optimal phylogenetic trees for a set of taxa. Using an optimality criterion the tree with the most favorable score is taken as the best estimate of the phylogenetic relationships of the included taxa. MP also requires comparison by alignment of nucleic acid or amino acid sequences to construct unrooted phylogenetic trees by heuristic algorithm [Wikipedia, 2009e]. It may also yield many resulting trees which are equally parsimonious. MP is independent of substitution model [Eck and Dayhoff, 1966, Kluge and Farris, 1969, Fitch, 1971]. This method is also accurate but computationally expensive.
Important issues in Phylogenetic analyses

Most methodologies used to construct phylogenetic/evolutionary relationships compare the similarities or dissimilarities among a set of species based on one trait at a time [Milner et al., 2003, Milner et al., 2004]. The resultant trees may not represent a consensus representation. Indeed, another procedure was introduced recently [Hedges, 2001] in which more than one type of strings of sequences are linearly joined to obtain a combined effect of as set of multiple gene sequences. While this represents a definite step forward towards obtaining a consensus on the evolutionary relationships, there still exist a number of unresolved issues that plague the reasonableness and consequent interpretation of results obtained by this method. Here, I have developed a totally novel approach towards building consensus phylogenetic trees and developed a series of algorithms that allow creating a reasonable tree topology in a 3D space and measure the fidelity or the extent of similarity between the resultant tree with the benchmark such as Linnaean taxonomy.

There exist a few interesting phenomena classified as Homoplasy that include Convergent evolution, Parallel evolution, and Evolutionary reversals which pose serious problems in estimating phylogenetic relationships. Furthermore, as all existing tree building methods yield two-dimensional dendrograms or trees, it is virtually impossible to pinpoint the effect of homoplasy on the topology of the phylogenetic trees. For a number of reasons, two organisms from widely placed taxa may express the same/similar trait due to convergent evolution and thereby distort the topology if one used that trait to examine evolutionary relationships. At present, unfortunately, most methods used to build phylogenetic trees concentrate one trait at a time with reference to a best known trait as a benchmark. Therefore, one ends up with widely varied and trait-specific phylogenetic relationships. Even the trees based on multiple...
traits [Gadagkar et al., 2005] do not resolve the issues faced due to homoplasy. In this thesis, I have addressed some of these problems.