CHAPTER FOUR

Phylogenetic Analyses

The science of molecular phylogenetics consists of the estimation of the evolutionary past of a group of living organisms based on the comparison of DNA or protein sequences. Phylogenetic analyses have, therefore, become an important tool in evolutionary studies and are critical in eliciting the fundamental commonalities between living organisms in order to understand the genetic relationships of biologically complex systems. Phylogenetic relationships can be studied on most levels of classification of organisms, including humans, and can be presented as phylogenetic trees that are the representation of evolutionary ancestral relationships of sequence data, by using numerical calculations often incorporated in software algorithms (Nei, 1996).

Initially the principles of phylogenetics were only applied to the systematic organisation of animal and plant species and only in later years did it start to focus more on the molecular sequence data that became available as sequencing technologies developed (Morrison, 1996). The amount of information that is obtained from sequence variation data is much more powerful in terms of eliciting the evolutionary history of a group of living organisms than the previously used morphological data and fossil records (Cummings *et al.*, 1995). In this regard, the development of automated sequencing technologies has increased the number of DNA sequences that are available to be used for evolutionary studies and has, therefore, contributed greatly to the development of more accurate and informative evolutionary histories. Furthermore, DNA sequences provide various types of markers that can be used to establish ancestry between and among individuals, of which the most common type is the SNP, which refers to the substitution, insertion or deletion of nucleotides (Pollock *et al.*, 2000).

4.1 <u>THE ROLE OF GENETIC DIVERSITY IN PHYLOGENETIC ANALYSES</u>

Genetic distances between sequences are determined by evolutionary forces, as was discussed in Chapter 2. Sequences will over time become more varied and will eventually diverge from one another. Genetic diversity, therefore, plays an important role in the phylogenetic analysis of sequences and provides a measure of dissimilarity between sequences, which reflects the time since evolutionary change has taken place in situations where a molecular clock can be assumed (Salemi and Vandamme, 2003). Evolutionary relations between DNA sequences are estimated by observing the commonalities or variation of a specific nucleotide site among the DNA sequences of a group of individuals or between populations (Morrison, 1996). These simplistic observations of genetic diversity are then further explored under certain assumptions of evolutionary models to interpret the observed genetic signals in terms of their evolutionary past. It has been reported that highly conserved gene regions, as well as highly variable sequence regions, are both of value in phylogenetic analyses and often these regions are used in combination to provide an accurate and highly resolved phylogenetic tree of evolutionary relationships between or among populations of individuals (Suárez-Díaz and Anaya-Muñoz, 2008).

4.1.1 Homology and sequence alignment

Genetic distances and therefore genetic diversity can only be determined by comparing DNA sequence data - site by site - to determine the amount of variation. To ensure the inference of accurate ancestral relationships between the sequences, it is essential that the comparisons being made between characters should be made between characters with the same evolutionary origin (Suárez-Díaz and Anaya-Muñoz, 2008). Sequence homology, which refers to the ability to match nucleotides with other nucleotides in molecular sequences that have a similar evolutionary origin, is, therefore, important to the validity of the phylogenetic outcome. Although homology can be determined by using different kinds of phylogenetic data types, such as morphological characters and behavioural characteristics (Morrison, 1996), the current investigation will limit the use of this term to the evolutionary similarity between nucleotide sites of molecular sequences.

In practice, homology is, however, difficult to predict because it is not known whether two base pairs at a site are similar because of a homologous ancestral state or because of an independent homoplasious origin. Homoplasy can hide the number of true evolutionary events that have taken place at a specific site, which would be detrimental to the determination of ancestral phylogenetic relationships between molecular sequences and should, therefore, be prevented if possible. Regions of DNA sequences that exhibit high mutation rates are susceptible to homoplasy and the use of such sequence regions should for this reason preferably be avoided in phylogenetic analyses.

The genetic diversity between two or more sequences is determined by aligning the sequences with each other based on the evolutionary correspondence between the different sequences. Homologous nucleotides, in this context, therefore refer to the common ancestral origin of two nucleotides in different sequences (Page and Holmes, 1998). Multiple sequence alignment arranges sequences in such a way that each sequence will correspond to other sequences according to its evolutionary history and ancestral origins. By aligning the sequences in such a manner, it is possible to obtain a measure of homology and to quantify the genetic diversity between the sequences. Sequence alignment was initially used to indicate the degree of structural similarities between protein sequences and homology would be inferred if the protein structures were 30% similar or more. The concepts that underlie this theory are that the structure between proteins and also the amount of variation between molecular sequences would not exceed a certain point when the proteins or sequences are related (Suárez-Díaz and Anaya-Muñoz, 2008).

A valid scientific approach to alignments are to exclude parts of sequence data that cannot be aligned reliably rather than include non-homologous data in the phylogenetic analysis. If alignments are not correct, it will result in further errors in the subsequent phylogenetic analyses and should preferably be verified manually before embarking on further phylogenetic analysis. This can however become problematic when working with large quantities of data in high throughput automated systems and the alignment methods or software used for this purpose should be robust enough not to have to verify the alignments manually (Levasseur *et al.*, 2008). These quality-assessment measures of multiple alignments have, therefore, become critical to the accuracy of the subsequent phylogenetic analysis. The most commonly used objective functions used for this purpose are the sum-of-pairs score or the log-likelihood ratio, which both indicate the quality of the alignment by global scores and are used in large high-throughput studies where manual quality verification is not feasible (Levasseur, *et al.*, 2008).

Software programs are available to determine the alignment of large strings of sequences by using mathematical algorithms, of which ClustalX and ClustalW are the best known (Baldauf, 2003). Other widely used packages include T-Coffee, MAFFT and MUSCLE.

The alignment of homologous sequences through comparison of sequences nucleotide by nucleotide, might be between identical characters or between substitutions, where a character has mutated, or between sequence characters that have been deleted or inserted (indels). Indels are the most difficult to match appropriately (Kumar and Filipski, 2007). Software algorithms will introduce gaps in areas where indels are present to accommodate insertions or deletions in the other sequences under alignment that do not contain the indel (Kumar and Filipski, 2007). However, gaps are highly unlikely phenomena, and the insertion of gaps to improve alignment should be performed with care. For this reason, software programs usually appoint penalties to gaps according to the length, number and position of the gaps and cannot remove them, only enlarge or add to them (Suárez-Díaz and Anaya-Muñoz, 2008). Furthermore, software algorithms align sequences by seeking the optimal match criterion or match between sequences, which is referred to as the “cost” at which two (2) sequences are aligned (Needleman and Wunsch, 1970). Usually it is based on a pattern-matching process in which gaps will be introduced to accommodate the indels and dynamic programming inserts gaps at a cost or gap penalty. The final alignment is thus determined by the alignment with the lowest cost (Needleman and Wunsch, 1970). The mathematical equation that is often used to determine gap penalties is, therefore, incorporated into the algorithms of software programs such as ClustalX (Larkin *et al.*, 2007), as presented in Equation 4.1.

Equation 4.1 Gap penalty

$$GP = g + hl$$

GP = Gap penalty; g = gap-opening penalty; h = gap-extension penalty; l = length of the gap. From Larkin *et al.*, 2007.

As with phylogenetic methodologies, there are different philosophies that underlie different approaches to sequence alignment (Gu *et al.*, 1995; Qian and Goldstein, 2001; Rosenberg and Kumar, 2003). The software applications make use of different basic approaches to alignment. Global alignment algorithms involve all the characters of the sequences and are used for aligning closely related sequences in contrast to local alignment algorithms that build alignments only on areas with shared similarity (Kumar *et al.*, 2004). Multiple alignment algorithms align many sequences and semi-global alignments are a hybrid of all the methods (Suárez-Díaz and Anaya-Muñoz, 2008). Although these methods differ to a certain extent, they are all based on the same premise that alignments are scored in order of similarity (Kumar *et al.*, 2004). Examples of software that use both the global and local approach for alignment purposes are dbClustal, TCoffee, MAFFT, MUSCLE and Probcons (Levasseur *et al.*, 2008).

Sequences are commonly aligned by using the dot plot or dot-matrix representation algorithm, which is based on a simple dot plot that is constructed for a pair of sequences.

The nucleotide positions at which the two (2) sequences are identical are indicated with a dot and the alignment is determined by a high similarity region that runs diagonally across the plot. This technique is based on the concept that high similarity between sequences is indicative of sequence homology (Salemi and Vandamme, 2003).

Dynamic programming is another approach to sequence alignment, which aligns sequences by using a weighted sum of pairs value through which the similarity of sequences per column of the alignment is maximised and gap lengths are minimised. This method also uses the scoring approach by determining a score value for each pair of sequences, as well as a weight for each pair of sequences, giving an overall score according to which the alignment confidence is determined. This technique is computationally complex and therefore not commonly used (Salemi and Vandamme, 2003). Examples of software programs that use this approach are the MSA, DCA, PRPP and SAGA programs, which are all software applications that are adjusted to perform the alignments according to the sum of weight pairs principle.

The most commonly used algorithm for sequence alignment is the progressive alignment method, which uses the evolutionary relationships between homologous sequences to align them (Barton and Sternberg, 1987; Thompson *et al.*, 1994). This procedure uses the branching order of a phylogenetic-guide tree to build up pairwise alignments of multiple sequences by building the alignment with the most similar sequences first, and aligning the more divergent sequences last, until a global alignment of all sequences is achieved (Kumar *et al.*, 2004). Criticism against this approach is based on the fact that it does not construct a sequence alignment that is based on a global optimal alignment, but rather one based on an optimised pairwise sequence strategy (Brocchieri, 2001). It therefore differs from dynamic programming insofar as it does not guarantee an alignment that will necessarily have the best score and cannot guarantee the most optimal alignment. This is however not problematic in the alignment of closely related sequences and will only produce problem alignments in very complex cases (Kumar et al., 2004; Salemi and Vandamme, 2003).

Many software applications make use of the progressive alignment algorithm, of which Clustal W (Thompson *et al.*, 1994), Clustal X (Thompson *et al.*, 1994), and the updates Clustal W 2.0 and Clustal X 2.0 (Larkin *et al.*, 2007) are the most widely used. Other software applications include T-Coffee, which is used routinely, and MAFFT and MUSCLE that are extremely fast and accurate (Larkin *et al.*, 2007).

Another approach is called “motif finding” and selects ungapped stretches of sequence and aligns the most conserved areas. These methods are not as popular, since they only align extracted stretches of DNA or proteins and do not allow for gaps and insertions (Brocchieri, 2001).

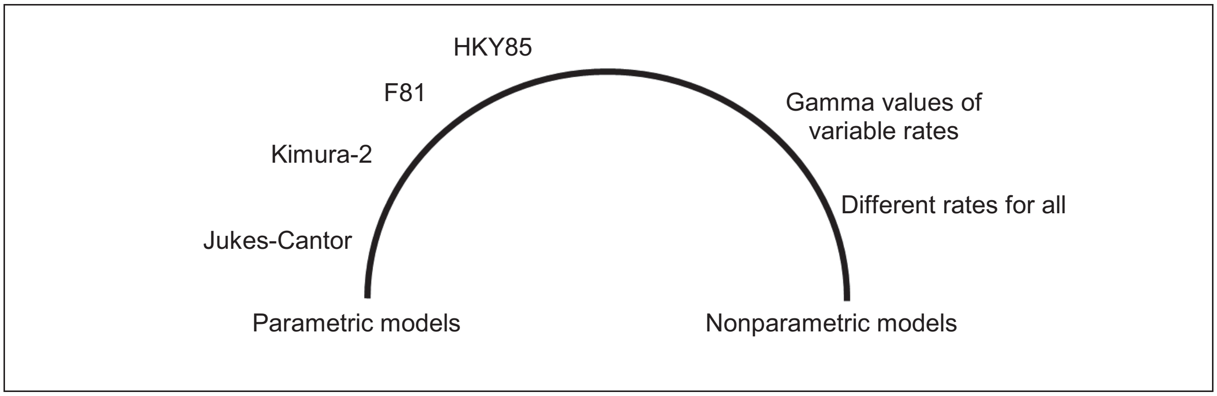
4.2 THE ROLE OF EVOLUTIONARY MODELS IN PHYLOGENETIC ANALYSES

Evolution is the process by which a mutation in a DNA sequence results in an advantageous or deleterious characteristic that would be fixed or lost in a population through evolutionary forces of natural selection and/or genetic drift (Nei, 1996). Since it is impossible to infer evolutionary history from studying DNA sequence variation i.e. mutations and genetic distance alone, it is necessary to incorporate assumptions about the evolutionary processes that gave rise to observed genetic distances. Some of these processes include the rate of nucleotide substitution, the ratio of transitions to transversions that occurred over time and whether different regions of the sequence under investigation were submitted to different rates of variation (Page and Holmes, 1998). Evolutionary models make it possible to reconstruct evolutionary events of the past and predict future events. They are based on the fundamental principles that the number of nucleotide changes between two genome sequences can be counted and studied and that the nucleotide changes can be placed within a time frame by assuming that mutations take place at regular intervals. Evolutionary events along the branches of a phylogenetic tree are, therefore, described by a model (Brocchieri, 2001; Baldauf, 2003).

4.2.1 Modelling evolution

The phylogenetic analyses of a group of sequences consist of a phylogenetic tree and a set of assumptions that explain how the observed nucleotide diversity that is represented by the branches of the tree originated. One method to determine evolutionary models is to follow an empirical approach by using parametric values obtained from studying large numbers of observed sequences and to apply these fixed values to other sets of sequences that are submitted to phylogenetic analysis. Models can also be built parametrically by using values based on biological properties of sequence divergence events, such as transition:transversion rates. These values are determined for each dataset and can differ between datasets (Whelan *et al.*, 2001).

Figure 4.1 Two approaches to construct evolutionary models



Parametric here refers to the determination of quantitative values that can be applied to model evolutionary processes either empirically or parametrically. Nonparametric here refers to the use of specific values for specific sequences and are, therefore, not applied to sets of sequences. Examples of parametric models: Jukes-Cantor = all substitutions are equally likely; Kimura-2 = does not assume that the rate of transitions and transversions are equal; F81 = Felsenstein's model that allows for different frequencies for different bases based on the base composition of sequences; HKY85 = combines Kimura-2 and F81. Examples of nonparametric models: Gamma = here refers to variable rates of substitution based on a gamma distribution. Adapted from Holmes, 2003.

Markov process models consist of either empirical or parametrical data that describe the relative rates of occurrence of all possible replacements and determine the probabilities of all nucleotide changes or non-changes at all sites by using a Q matrix to specify the relative rate of change of each nucleotide along the sequence (Whelan *et al.*, 2001). The Q matrix is a simple presentation of the base frequency and probability of substitution between two bases for all different types of bases. The base frequency and substitution probability parameters are the variables in these matrices that form the basis of the different evolutionary models.

The most commonly used models of evolution are based on the parametric approach. Examples are the Jukes-Cantor (JC) model (Jukes and Cantor, 1969) in which all bases have equal frequencies and all substitutions are equally likely, the Kimura-2 parameter (K2P) model (Kimura, 1980) which incorporates the fact that transitions and transversions do not occur at equal frequencies and thus uses a transition:transversion rate in the correction of genetic distances, the Felsenstein's (FEL/F81) model (1989) that allows for different frequencies for different bases based on the base composition of sequences, the Hasegawa, Kishino and Yano (HKY85) model (Hasegawa *et al.*, 1985) that combines the theories of K2P and FEL/F81, the general reversible model (REV) that allows probabilities for each substitution and the REV + ɪ (gamma distribution value) model that includes the random assignment of gamma distribution values (Yang, 1996). DNA substitution models are, therefore, based on parameters such as base frequency, base exchangeability and rate heterogeneity. The most efficient models use at least base frequency, the transition:transversion ratio parameter, all six base exchangeability parameters and the gamma rate heterogeneity values (Whelan *et al.*, 2001). Additional factors are taken into account when distances of protein coding DNA sequences are determined, such as the distinction between nonsynonymous and synonymous substitutions and by comparing codons rather than base pairs. Distance can also be determined by genotypic frequencies and restriction endonuclease data (Morrison, 1996).

4.2.2 Base composition parameter in evolutionary models

Base composition is determined by the frequencies of all the types of bases and averaged over all the sites of the sequences. Base composition varies among and within sequences because of G + C content, neighbour bases and differences in the efficiency of DNA repair in the heavy and light strands of the DNA (Liò and Goldman, 1998). Base composition differences can lead to problems in the branch length of phylogenetic trees and thus this needs to be compensated for (Hasegawa *et al.*, 1993). A parameter for base composition is used to represent the constraints on base frequencies and gives weight to the type of substitutions that are most likely to be observed in DNA sequences (Whelan *et al.*, 2001).

The most important reason for the variation in base composition is ascribed to large regions in the DNA with uniform G and C content, which varies from the G and C content in other isochores. Several theories exist about the reasons for these differences in G and C content. It has been proposed that the misincorporation errors during DNA replication and repair cause a directional mutation pressure that is responsible for these variations in G + C content (Sueoka, 2002). It is further believed that this phenomenon is an adaptation in animals and plants to protect the organisms against high temperatures because the G-C bonds protect the DNA against denaturation (Mooers and Holmes, 2000). Neutralists however claim that the isochores are the result of normal genetic mutation that takes place in all genomes and could be caused by DNA replication, DNA repair or recombination (Liò and Goldman, 1998). The LogDet transformation method accommodates datasets that violate the assumption of base composition equilibrium and is used to circumvent the problem of variable base composition between sequences in a dataset (Lockhart *et al.*, 1994).

4.2.3 Base substitution parameters in evolutionary models

Functional constraint in a gene is demonstrated by the probability that mutations occurring in a specific position within a coding region of a gene will be deleterious and result in a lower rate of substitution. The codon sequence positions demonstrate this theory by displaying higher functional constraints within the first and second position of the codons than in the third position of the codon. The non-coding regions of the DNA evolve faster than the coding regions and have more deletions and insertions because the functional constraints on these areas of the DNA are low and therefore substitution rates are high. (Page and Holmes, 1998) In mitochondrial DNA, for example, the mutation rate of the control area is known to be very high and therefore the control region is also referred to as hyper-variable areas (Page and Holmes, 1998).

Base substitution parameters therefore describe the probabilities of types of base substitutions and the rate of the type of change that could be expected. For example, transition substitutions are generally more probable than transversions; this phenomenon is referred to as the transition bias (Brown *et al.*, 1979; Strandberg and Salter, 2004). Mathematically, the rate for a transition substitution is given as κ in relation to a transversion rate of 1, in which case the transition rate κ would usually be more than 1 (Kimura, 1980). The ratio of transitions to transversions is an important parameter that needs to be incorporated into phylogenetic relationships because it is an indication of the measure of variability of DNA sequences over time and therefore gives an understanding of the patterns of evolution. It also plays a role in evolutionary distance correction methods. For this reason, transition:transversion ratios are calculated for the purpose of phylogenetic tree drawing by using distance-based methods that use pairwise distance measures, parsimony methods and likelihood function methods (Kimura, 1980; Strandberg and Salter, 2004).

The distance methods determine the transition:transversion ratio for each possible pair of sequences in the dataset and average the pairwise estimates into a single transition:transversion ratio estimate. In the parsimony approach, a maximum parsimony (MP) tree is determined and the transitional:transversional incidences are counted in each branch of the tree. The transition:transversion ratio is the total number of transitions to the total number of transversions. Certain transition:transversion ratio methods are adjusted to incorporate the time to divergence by plotting the transversion changes against the time since divergence. The F84 evolutionary models and HKY85 evolutionary model both include a transition:transversion ratio parameter and maximum likelihood (ML) methods are used to estimate the parameters in these models (Strandberg and Salter, 2004).

4.2.4 Rate heterogeneity parameters in evolutionary models

Base rate heterogeneity parameters describe the variation of evolution at different sites of a sequence and vary according to biochemical factors, functional constraints or natural selection (Whelan *et al.*, 2001). The main reason for substitution rate variation is not, as would be expected, the variation of mutation rate among sites, but rather the functional constraints on certain sites, resulting in specific sites with a high rate of change in contrast to other sites that display low rates of change (Yang, 1996).

Different models can be used to describe this phenomenon. Some models, such as the HKY85, denote a fraction of the sites of a sequence as one rate and the rest of the sites at an invariable rate (Hasegawa *et al*., 1985). The most widely used approaches, however, entail describing the rate of variation at each sequence site through a random draw from a gamma distribution (Nei and Gojoborit, 1986; Tamura and Nei, 1993; Yang, 1996). In addition, the "invariable-sites model" assigns a value of zero to sites that presumably do not have a high rate of change and assumes that other sites change at the same rate. However, based on biological expectations, this is not realistic and a continuous model is more suitable to describe rate variation (Steel *et al.*, 2000).

Rate variation can be described by using random values from a continuous gamma distribution. The distribution plots the proportion of sites (y-axis) against the substitution rates (x-axis) to represent the α parameter or range of rate variation among sites in a bell-shaped or L-shaped / reverse J-shaped curve (Page and Holmes, 1998). Under a gamma distribution, the number of substitutions at sites follows a negative binomial distribution and the variation of substitution rate (*λ*) is formulated, as is presented in Equation 4.2 (Yang, 1996; Gu and Zhang, 1997).

Equation 4.2 Variation of substitution rate (*λ*)

$$\emptyset(\lambda) = \frac{\beta}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

α = shape parameter; β = scalar; $\emptyset(\lambda)$ = the coefficient of variation ($1/\sqrt{\alpha}$); Γ = gamma shaped parameter. From Gu and Zhang,1997.

Large values of α parameters result in a bell-shaped curve, suggesting low rate heterogeneity, whereas small values of α result in a reverse-J shaped curve suggesting high rate heterogeneity (Gu and Chang., 1997). Two main methodologies for determining the α value are presented in the literature. The first is based on a ML framework with likelihood functions that are so time-consuming that they are difficult to apply to more than five sequences. The second is based on the principle of parsimony to estimate the number of substitutions, which has been determined to lead to an underestimation of the number of substitutions (Yang, 1996). Gu and Zhang (1997) developed a method that is also based on ML principles but is not as time-consuming. This method is used in this study to determine the α values of the datasets and consists of two steps. During the first step, the expected number of substitutions that is corrected for multiple hits is determined for each site by a likelihood approach and during the second step, the ML estimate of α is determined under a negative binomial distribution (Gu and Zhang, 1997).

Rate variation has been determined to have an important effect on phylogenetic analyses. When site rate variability is ignored, genetic distances and branch lengths are underestimated, with a bias for large distances and long branches. Since the lengths of especially long branches are underestimated, it results in an overestimation of the divergence time estimations. It also underestimates the transition:transversion rate ratio calculations because the transitions at the high rate variation sites will invariably also be ignored. It is, therefore, essential to apply an adequate rate variation model in instances where rate variation is high, as in the case of the mitochondrial genome (Yang, 1996).

Sequences with high rate variation tend to provide low levels of phylogenetic information because the sites that have low substitution rates contribute little information about evolutionary history and sites with high substitution rates will also provide little evolutionary information owing to mutational saturation. If rate variation is, therefore, present but ignored, the distance matrix and ML tree-building methods can be misleading. Even the parsimony-based methods have been determined to be misleading (Yang, 1996).

4.3 PHYLOGENETIC METHODS

Phylogenetic tree analysis is the most popular method to analyse genetic variation and was first proposed by Cavalli-Sforza and Edwards in 1964 (Cavalli-Sforza and Edwards, 1967).It not only represents the genetic diversity but also includes all the fissures that took place during the history of a population. Phylogenetic trees therefore offer a mathematical method for ordering large quantities of genetic information in a meaningful way (Nei, 1996). In response to the large amounts of sequencing data that have become available over the years, phylogenetic methods have developed dramatically and currently not only include information about the methodology, but also the complex computer software programs and computer hardware requirements required to perform these types of analyses successfully. The difficulty lies in the decision on which methodology and underlying principles of homology to follow to reconstruct the most objective and accurate history of evolutionary events (Suárez-Díaz and Anaya-Muñoz, 2008).

4.3.1 Basic principles of tree-building methods

The first phylogenetic tree was constructed according to the graphical representation methods of subsets of molecular characters by Fitch and Margoliash (1967). It was based on a distance matrix constructed from the amount of sequence variation observed in a sequence alignment from which the ancestral relationships were determined (Fitch, 1971). The development of computer technology resulted in software programs with algorithms that currently automatically perform these and even more complex tasks (Suárez-Díaz and Anaya-Muñoz, 2008).

Phylogenetic tree building is based on two basic processes. The first process entails the construction of the tree topology and the order of the branches and the second process is therefore, aimed at estimating the branch lengths. In cases where phylogenetic trees are, constructed from large sets of sequences (more than 50), thousands of different tree topologies are possible and the greatest difficulty lies in the selection of the most optimal tree topology. The current tree-building methods are classified into three major groups, namely methods based on determination of genetic distance, methods based on discrete data and methods based on ML (Nei, 1996).

4.3.1.1 Tree building by using distance or discrete data

The principle behind using distance data is to determine the evolutionary distances between a group of sequences based on the nucleotide differences between the sequences and to use these distances to build a phylogenetic tree to indicate the evolutionary relationships between them. Sequences that are identical between a pair will have a distance of zero, as opposed to a sequence that is totally different from the other having a distance of one. The sum of all the character state differences for all the characters are called the Manhattan distances or euclidian distances and they refer to the square root of the sum of the squared differences (Morrison, 1996). These methods convert the differences between aligned sequences into distance matrices and use the matrix values in tree-building methods (Baldauf, 2003).

The distances in the distance matrix are represented in the tree as branch lengths. Branches represent evolutionary divergence i.e. the more divergence in the data, the longer the tree branches will be (Baldauf, 2003). Tree distances are obtained from trees and observed distances are obtained from datasets. The observed distances are rarely represented in a tree with 100% accuracy and the fit can be regarded as a measure of the best tree representation (Page and Holmes, 1998).

Discrete methods use each nucleotide site for the tree-building method, thereby including information about ancestral states. It not only calculates the differences between nucleotides, but uses the type of change that has taken place to determine possible ancestral states by applying certain rules (Page and Holmes, 1998).

4.3.1.2 Tree building by clustering or searching

Tree-building methods are based on the principle of either clustering sequences stepwise by constructing a tree with initial data and then adding onto that tree as additional data are considered, or by searching through optimality criteria that select the best fitting tree between a set of possibilities. Clustering or constructive methods follow steps in an algorithm and have the advantage of producing a single tree that can be computed fast and accurately (Morrison, 1996). There is no way, however, to determine how well the tree fits the data and tree topology will be dependent on the order in which sequences are processed. So although the output of these methods is one tree, it is not necessarily the best tree for the data (Page and Holmes, 1998).

By using optimality criteria, a weight is determined for each tree indicating the fit to data suitability. Search methods allow for the comparison of different trees representing different hypotheses of the evolutionary relationship of a set of data (Morrison, 1996). They demand great computational power and can be very expensive to perform, especially when large datasets are analysed. In view of the high demand for computational ability, heuristic methods are often used to search for the most optimal tree in large datasets. Heuristic methods do not guarantee an optimal tree in the end (Nei, 1996). Heuristic methods search randomly for the best solution to a problem and should produce an outcome that is close to the most accurate outcome. These methods try to determine the most optimal tree by sequentially adding sequences to the most optimal branches of the growing tree and searching through the other trees to find a more optimal one, or making the tree more optimal by branch swapping (Morrison, 1996).

4.3.2 Distance methods

The goal of genetic distance methods is to match a distance matrix with a phylogenetic tree topology. *P*-distances are the fraction of positions at which a pair of sequences differ and from which a tree topology is inferred by generally using the least squares and minimum evolution (ME) principles (Nei, 1996). When using the least squares principle, the tree topology is determined by using the smallest minimum squared sum of differences between pairs of sequences (Bulmer, 1991). Tree topologies of phylogenetic trees that are constructed by using the minimum tree principle consist of the tree topology that displays the smallest sum of tree lengths of all the possible tree topologies for a set of sequences (Rzhetsky and Nei, 1992). Searching for the optimal ME tree is highly labour-intensive and nearly impossible for large datasets. For this reason some methods have built-in algorithms to search for the optimal tree while determining the topology, which produces the optimal tree automatically. Examples of these methods are the neighbour-joining (NJ) method of Saitou and Nei (1987), Wagner method, modified Farris method and the neighbourliness method (Nei, 1996).

The true number of substitutions between pairs of sequences is nearly impossible to determine, based only on the number of observed substitutions, and the accuracy of these additive tree topologies can therefore be fallible. The reason for this phenomenon is that distance methods are based on simple sequence difference calculations; they do not take into account reverse mutation events and can therefore underestimate the true evolutionary distances. Furthermore, it does not take substitution heterogeneity into account (Nei, 1996; Suárez-Díaz and Anaya-Muñoz, 2008). To overcome this problem it is critical to model the distance calculations against evolutionary models that make adjustments to correct for the less informative states that are used, as was discussed in Section 4.2 (Brocchieri, 2001). The fact that the distance methods are based on distance parameters only has also resulted in criticism of this approach and necessitates the use of statistical tests to estimate the confidence of the tree topologies constructed according to this method, such as the bootstrap method (Suárez-Díaz and Anaya-Muñoz, 2008).

4.3.2.1 Unweighted pair-group method

Methods such as the unweighted-pair group method with arithmetic means (UPGMA) or the weighted-pair group method with arithmetic means use sequential clustering techniques to construct phylogenetic trees. By these methods trees are built stepwise by grouping the most similar sequences together, from which point they are regarded as a single operational taxonomic unit (OTU), which again is grouped with the next most similar OTU. By continuing with this process all OTUs can be clustered into a tree. Clustering requires ultrametricity, which is the actual distances between sequences as represented in branch lengths, with a focus on molecular clock-like behaviour because it assumes that evolution takes place at a uniform rate over time. Many recent methods have been developed that deal with non-ultrametricity and non-clock-like behaviour and therefore clustering has become less popular (Morrison, 1996).

4.3.2.2 Neighbour-Joining method

Although the NJ method of Saitou and Nei (1987) is very similar in principle to the clustering method, it is an example of an ME distance method (Saitou and Nei, 1987), is computationally relatively fast and can handle large datasets (Suárez-Díaz and Anaya-Muñoz, 2008). The search for an optimal tree topology is embedded in the algorithms of the software used to construct NJ trees and therefore this method produces a single tree.

Neighbours in the context of this method can be defined as a pair of OTUs that are connected to a single node in an unrooted bifurcating tree. In this method the tree topology is determined by joining sets of neighbours that give the smallest branch length and represent the minimum evolutionary change, to become a new OTU that can be joined with a new neighbour with the smallest branch length and so on until the full tree topology is resolved. Since the true neighbours are not defined, the sum of branch lengths are computed for all pairs of OTUs and the pair with the smallest value is chosen as the next neighbour to become a single OTU. The procedure is applied again until there is only three OTUs left (Saitou and Nei, 1987; Nei, 1996). Criticism against the NJ method is that it uses a “greedy approach” to construct a tree topology, in which the clustering of sequences are progressive and does not always lead to the most optimal tree. The algorithm of NJ is similar to UPGMA, but differs in the fact that each pair of groups is adjusted according to its distance from all the other groups (Morrison, 1996). However, computer simulations performed by Saitou and Nei (1987) indicated that the NJ method is a reliable tree-building method and compares well with other methods in terms of accuracy (Saitou and Nei, 1987).

4.3.3 Discrete methods

These methods are also based on the principles of ME methods to determine tree topologies and use optimality criteria. Phylogenetic trees are constructed based on the smallest sum of all branch lengths (Morrison, 1996). In contrast to distance methods, discrete methods use the actual nucleotide site information to construct evolutionary trees and allow inference about the character content of the ancestor. Nucleotide site information could either be the type of substitution that took place or functional information of a site (Page and Holmes, 1998). The MP and ML methods are the two most common discrete methods used in phylogenetic analyses (Nei, 1996).

4.3.3.1 Maximum parsimony

This method seeks to reconstruct the ancestral sequences by assigning a hypothetical ancestral sequence to each internal node of the tree. These character states are predicted by using parsimony rules and are not always ambiguous. The ambiguous character sets imply that there is more than one possibility of an ancestral state, which is determined by the MP method as the least number of evolutionary changes required to satisfy an ancestral character set (Whelan *et al.*, 2001; Brocchieri, 2001; Nei, 1996).

Although this method can easily incorporate insertions and deletions into the phylogenetic analysis, the least number of evolutionary event estimates made by the MP method can lead to incorrect estimates of evolutionary rates (Suárez-Díaz and Anaya-Muñoz, 2008). The biggest criticism against the MP method is based on the heavy consistency constraints placed on substitution rates. The MP method is expected to determine accurate trees under conditions where there are no multiple substitutions at each site and when enough informative sites are examined. Therefore, backward and parallel substitutions will make the outcome of an MP tree uncertain (Nei, 1996). Kumar criticised it for not always constructing accurate branch lengths and Felsenstein argued that it did not increase in accuracy as the number of sequences increased (Brocchieri, 2001). Furthermore, there is no manner in which to estimate the means and variances of the minimal number of substitutions, making it difficult to treat the MP tree in a statistical framework (Nei, 1996).

In generalised parsimony, substitution models are used to allocate different "costs" to different types of substitutions. The smaller the probability of an evolutionary event, the higher the cost assigned. This relates especially to the transition bias and substitution rate heterogeneity present in the mitochondrial genome sequence, where a large weight is given to transversions or slowly evolving sites. In weighted parsimony, sites are weighted according to their phylogenetic value and can be determined *a priori* or *a posteriori* (Nei, 1996).

Examples of *a priori* weighting is the Dollo-parsimony where no parallelisms or convergences are included and the Camin-Sokal-parsimony where reverse mutations are not taken into account (Morrison, 1996). Many of these *a priori* types of methods are designed for use with molecular data because of the fact that nucleotides cannot be regarded as substituting at equal probabilities. Sites that mutate at a high rate will become saturated and have low phylogenetic value and will accordingly be appointed as having low parsimony weight. Functional inequalities along the sequences are indicated by using character weighting that allocates weight to different sequence positions, for example different codon positions and positions that play a critical role in the formation of protein or RNA structures. Character-state weighting is the allocation of different weights to the same sequence positions. This type of weighting refers to mutational biases, for example the probability of transversions versus transition-type substitutions, types of substitution frequencies, the base composition, and the presence of synonymous versus non-synonymous type substitutions (Morrison, 1996).

A major disadvantage of this method is that it can produce more than one most parsimonious tree and that the final tree must ultimately be chosen by the analyst (Suárez-Díaz and Anaya-Muñoz, 2008). Thus subjectivity is brought into the process, which is not optimal. Furthermore, the MP methods have been determined to be inconsistent because of "long branch attraction" (Nei, 1996). This can occur when rates of evolution differ dramatically between sequences under circumstances of rapid evolution, and leads to extremely long branches. Branches are joined when the same substitutions take place in two separate branches and the tree-building algorithm incorrectly joins it. This problem can be overcome by using the MP method in conjunction with other tree-building methods (Suárez-Díaz and Anaya-Muñoz, 2008).

4.3.3.2 Maximum likelihood

ML is based on the most likely tree to describe the observed dataset and this method, therefore, includes the evaluation of different hypotheses statistically. ML incorporates complex models of evolution in its inference of phylogeny and is regarded as a method that can make accurate inferences not only about the patterns of evolution but also about the processes of evolution. It is often regarded as the most powerful phylogenetic method (Whelan *et al.*, 2001).

The ML method is a model of sequence evolution that produces a tree from a set of observed data. It aims to determine the most likely tree to support the observed data by evaluating which tree topology will make the observed data most likely. Thus, what is important is not which tree is the most probable, but rather which makes the set of data most likely (Page and Holmes, 1998). It does this by using the probabilities for a specific nucleotide to be substituted to another nucleotide as described by an evolutionary model. Likelihood is determined by multiplying the probability of the nucleotide at a position with the probability of a specific transformation at that position. The product of the likelihoods at a position between two sequences equals the likelihood of divergence of those two sequences at that position and the product of all the likelihoods of all the branches equals the likelihood of the tree (Morrison, 1996). Therefore, given an observed nucleotide at a specific site, it needs to be calculated what ancestral state will have the highest probability of giving rise to that observed state. This probability will constitute the likelihood and is usually performed in a computationally demanding and time-consuming manner. As with parsimony, the method allows for the inclusion of weighting factors such as substitution rate, transition:transversion ratios and base composition. With ML, these values would constitute those that lead to ML and are easy to estimate (Suárez-Díaz and Anaya-Muñoz, 2008).

Another advantage of this method is that different hypotheses can be compared by using the likelihood ratio test. The use of a model compared to observed data can be tested by using the likelihood ratio statistic. Since the evolutionary model used will have a great influence on the branch lengths of the tree, it is sometimes necessary to reject a model if it does not fit the data. On the same basis, trees are compared to find the most likely tree to fit the observed data (Page and Holmes, 1998).

Edwards and Cavalli-Sforza postulated that a parsimonious tree can under some circumstances also be the most likely tree and can be used in cases where the computational power for ML is lacking (Cavalli-Sforza and Edwards, 1967). In cases where the substitution rate varies between branches, the ML and MP trees might, however, differ significantly (Page and Holmes, 1998). The biggest problem with the ML method is that it is computationally inefficient. The optimal tree is chosen from a large number of trees and using exact methods for this purpose is one of the mathematical problems for which no solution has been determined yet in terms of the time and power it takes to achieve this goal (Morrison, 1996).

4.3.4 Choosing a phylogenetic tree-building method

There are many different phylogenetic tree-building methods that are based on the basic principles discussed in the previous sections, and to decide upon the most appropriate method is not an easy task. Tree-building methods need to be computationally efficient, which refers to the speed at which a tree can be constructed (Morrison, 1996). This criterion is dependent on the type of algorithm used and it can therefore be expected that the optimality methods will score low on this criterion (Nei, 1996). The MP, ME and ML methods all search for optimal trees, making them computationally demanding. The NJ tree scores very high on this criterion, since it can handle large datasets and performing the bootstrapping is easy (Suárez-Díaz and Anaya-Muñoz, 2008).

The power of a method, which is the amount of data needed to deliver a single tree and is used to construct a tree, is another criterion to consider (Sanderson and Shaffer, 2002). The distance methods score low on this criterion owing to loss of informative data in the distance matrix (Nei, 1996).

Tree-building methods need to be consistent, which refers to the chance that if more data are added, the method will still deliver the same single accurate tree (Sanderson and Shaffer, 2002). The NJ, ME and ML methods produce consistent trees, provided that the correct nucleotide substitution model is used. The MP method, however, seems not to be as consistent (Nei, 1996). But according to Nei (1996), if the number of nucleotides is in the thousands, all of the methods seem to be problematic in terms of consistency.

It also needs to be robust, which is the most important criterion and refers to how consistent the method will be when the assumptions underlying it are changed (Sanderson and Shaffer, 2002). The ML method has proven to be the most robust method and the UPGMA and invariants methods the least robust (Morrison, 1996).

Tree-building methods need to be verified statistically and in this category the NJ and ME trees score highest (Sanderson and Shaffer, 2002). Statistical methods for testing these methods are well established. Statistical testing of the ML method however seems problematic and therefore scores low (Nei, 1996).

The probability of getting accurate tree topologies is the most important factor to take into account when choosing a tree-building method (Nei, 1996). This is, however, also the most difficult to determine. Computer simulations are the most effective way to study the accuracy of tree topology and have been performed many times in order to answer this question. But comparing results is difficult, since the type of simulations varied greatly and to such an extent that there is no obviously better or worse tree-building method currently in use. Each of these methods has strong and weak points and needs to be used accordingly (Sanderson and Shaffer, 2002).

Another important factor that goes with tree topology is the reliability of branch length estimations (Sanderson and Shaffer, 2002). In this category, it is likely that the ML, NJ and ME trees give more reliable branch length estimates than MP trees (Nei, 1996).

Phylogenetic analyses are always based on a degree of uncertainty because of the incompleteness of data and it is important to be able to compare alternative trees to decide on the most optimal topology (Morrison, 1996). For this reason, different tree-building methods are often used in conjunction to provide an objective estimate of the true tree topology of a set of sequences (Nei, 1996).