4

Genetic distances and nucleotide substitution models

THEORY

Korbinian Strimmer and Arndt von Haeseler

4.1 Introduction

One of the first steps in the analysis of aligned nucleotide or amino acid sequences typically is the computation of the matrix of ***genetic distances*** (or *evolutionary distances*) between all pairs of sequences. In the present chapter we discuss two questions that arise in this context. First, what is a reasonable definition of a genetic distance, and second, how to estimate it using statistical models of the substitution process.

It is well known that a variety of evolutionary forces act on DNA sequences (see Chapter 1). As a result, sequences change in the course of time. Therefore, any two sequences derived from a common ancestor that evolve independently of each other eventually diverge (see Fig. 4.1). A measure of this divergence is called a genetic distance. Not surprisingly, this quantity plays an important role in many aspects of sequence analysis. First, by definition it provides a measure of the similarity between sequences. Second, if a ***molecular clock*** is assumed (see Chapter 11), then the genetic distance is linearly proportional to the time elapsed. Third, for sequences related by an evolutionary tree, the branch lengths represent the distance between the nodes (sequences) in the tree. Therefore, if the exact amount of sequence divergence between all pairs of sequences from a set of n sequences is known, the genetic distance provides a basis to infer the evolutionary tree relating the sequences. In particular, if sequences actually evolved according to a tree and if

The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing,
Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme (eds.). Published by Cambridge
University Press. © Cambridge University Press 2009.

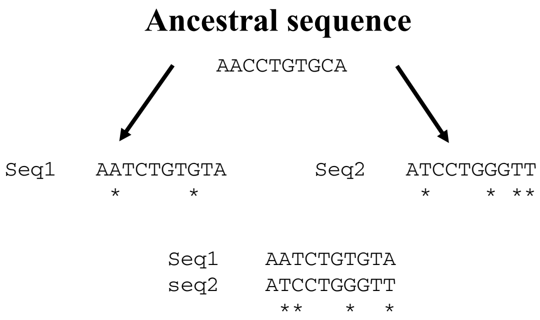

Fig. 4.1 Two sequences derived from the same common ancestral sequence mutate and diverge.

the correct genetic distances between all pairs of sequences are available, then it is computationally straightforward to reconstruct this tree (see next chapter).

The substitution of nucleotides or amino acids in a sequence is usually modeled as a random event. Consequently, an important prerequisite for computing genetic distances is the prior specification of a ***model of substitution***, which provides a statistical description of this stochastic process. Once a mathematical model of substitution is assumed, then straightforward procedures exist to infer genetic distances from the data.

In this chapter we describe the mathematical framework to model the process of nucleotide substitution. We discuss the most widely used classes of models, and provide an overview of how genetic distances are estimated using these models, focusing especially on those designed for the analysis of nucleotide sequences.

4.2 Observed and expected distances

The simplest approach to measure the divergence between two strands of aligned DNA sequences is to count the number of sites where they differ. The proportion of different homologous sites is called ***observed distance***, sometimes also called ***p-distance***, and it is expressed as the number of nucleotide differences per site.

The p-distance is a very intuitive measure. Unfortunately, it suffers from a severe shortcoming: if the degree of divergence is high, p-distances are generally not very informative with regard to the number of substitutions that actually occurred. This is due to the following effect. Assume that two or more mutations take place consecutively at the same site in the sequence, for example, suppose an A is being replaced by a C, and then by a G. As result, even though two replacements have occurred, only one difference is observed (A to G). Moreover, in case of a back-mutation (A to C to A) we would not even detect a single replacement. As a consequence, the observed distance p underestimates the true

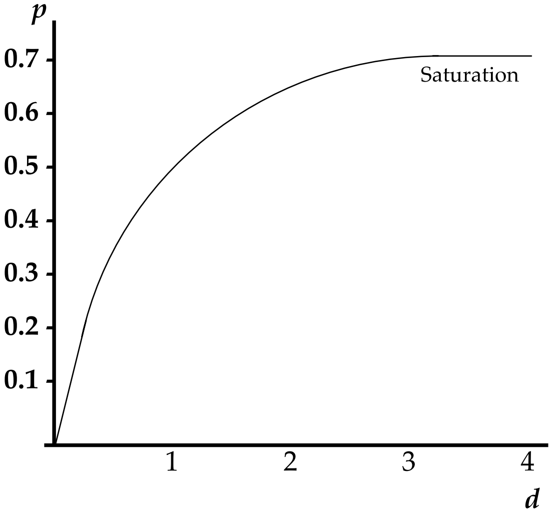

Fig. 4.2 Relationships between expected ***genetic distance*** *d* and observed ***p-distance***.

genetic distance *d*, i.e. the actual number of substitutions per site that occurred. Figure 4.2 illustrates the general relationship between *d* and *p*. As evolutionary time goes by, multiple substitutions per site will accumulate and, ultimately, sequences will become random or ***saturated*** (see Chapter 20). The precise shape of this curve depends on the details of the ***substitution model*** used. We will calculate this function later.

Since the genetic distance cannot be observed directly, statistical techniques are necessary to infer this quantity from the data. For example, using the relationship between *d* and *p* given in Fig. 4.2, it is possible to map an observed distance *p* to the corresponding genetic distance *d*. This transformation is generally non-linear. On the other hand, *d* can also be inferred directly from the sequences using ***maximum likelihood*** methods.

In the next sections we will give an intuitive description of the substitution process as a stochastic process. Later we will emphasize the "mathematical" mechanics of nucleotide substitution and also outline how ***maximum likelihood estimators*** (***MLEs***) are derived.

4.3 Number of mutations in a given time interval *(*optional*)

To count the number of mutations X(t) that occurred during the time t, we introduce the so-called *Poisson process* which is well suited to model processes like radioactive decay, phone calls, spread of epidemics, population growth, and so on. The structure of all these phenomena is as follows: at any point in time an event, i.e. a mutation, can take place. That is to say, per unit of time a mutation occurs with intensity or rate μ. The number of events that can take place is an integer number.

Let $P_n(t)$ denote the probability that exactly n mutations occurred during the time t:

$$P_n(t) = \mathrm{P}(\mathrm{X}(t) = n) \tag{4.1}$$

If t is changed, this probability will change.

Let us consider a time interval δt. It is reasonable to assume that the occurrence of a new mutation in this interval is independent of the number of mutations that happened so far. When δt is small compared to the rate μ, $\mu\delta t$ equals the probability that exactly one mutation happens during δt. The probability of no mutation during δt is obviously $1 - \mu\delta t$. In other words, we are assuming that, at the time $t + \delta t$, the number of mutations either remains unchanged or increases by one. More formally

$$P_0(t + \delta t) = P_0(t) \cdot (1 - \mu\delta t) \tag{4.2}$$

That is the probability of no mutation up to time $t + \delta t$ is equal to the probability of no mutation up to time t multiplied by the probability that no mutation took place during the interval $(t, t + \delta t)$. If we observe exactly n mutations during this period, two possible scenarios have to be considered. In the first scenario, $n - 1$ mutations occurred up to time t and exactly one mutation occurred during δt, with the probability of observing n mutations given by $P_{n-1}(t) \cdot \mu\delta t$. In the second scenario, n mutations already occurred at time t and no further mutation takes place during δt, with the probability of observing n mutations given by $P_n(t) \cdot (1 - \mu\delta t)$. Thus, the total probability of observing n mutations at time $t + \delta t$ is given by the sum of the probabilities of the two possible scenarios:

$$P_n(t + \delta t) = P_{n-1}(t) \cdot \mu\delta t + P_n(t) \cdot (1 - \mu\delta t) \tag{4.3}$$

Equations (4.2) and (4.3) can be rewritten as:

$$[P_0(t + \delta t) - P_0(t)]/\delta t = -\mu P_0(t) \tag{4.4a}$$

$$[P_n(t + \delta t) - P_n(t)]/\delta t = \mu[P_{n-1}(t) - P_n(t)] \tag{4.4b}$$

When δt tends to zero, the left part of (4.4a, b) can be rewritten (ignoring certain regularity conditions) as the first derivative of $P(t)$ with respect to t

$$P_0'(t) = -\mu \cdot P_0(t) \tag{4.5a}$$

$$P_n'(t) = \mu \cdot [P_{n-1}(t) - P_n(t)] \tag{4.5b}$$

These are typical differential equations which can be solved to compute the probability $P_0(t)$ that no mutation has occurred at time t. In fact, we are looking for a function of $P_0(t)$ such that its derivative equals $P_0(t)$ itself multiplied by the rate μ. An obvious solution is the exponential function:

$$P_0(t) = \exp(-\mu t) \tag{4.6}$$

That is, with probability $\exp(-\mu t)$ no mutation occurred in the time interval $(0, t)$. Alternatively, we could say that probability that the first mutation occurred at time $x \geq t$ is given by:

$$F(x) = 1 - \exp(-\mu t) \tag{4.7}$$

This is exactly the density function of the *exponential distribution* with parameter μ. In other words, the time to the first mutation is exponentially distributed: the longer the time, the higher the probability that a mutation occurs. Incidentally, the times between any two mutations are also exponentially distributed with parameter μ. This is the result of our underlying assumption that the mutation process "does not know" how many mutations already occurred.

Let us now compute the probability that a single mutation occurred at time t: $P_1(t)$. Recalling (4.5b), we have that:

$$P_1'(t) = \mu \cdot [P_0(t) - P_1(t)] \tag{4.8}$$

From elementary calculus, we remember the well-known rule of products to compute the derivative of a function $f(t)$, when $f(t)$ is of the form $f(t) = h(t)g(t)$:

$$f''(t) = g'(t)h(t) + g(t)h'(t) \tag{4.9}$$

Comparing (4.9) with (4.8), we get the idea that $P_1(t)$ can be written as the product of two functions, i.e. $P_1(t) = h(t)\ g(t)$ where $h(t) = P_0(t) = \exp(-\mu t)$ and $g(t) = \mu t$. Thus $P_1(t) = (\mu t) \exp(-\mu t)$. If we compute the derivative, we reproduce (4.8). Induction leads to (4.10):

$$P_n(t) = [(\mu t)^n \exp(-\mu t)]/n! \tag{4.10}$$

This formula describes the *Poisson distribution*, that is, the number of mutations up to time t is Poisson distributed with parameter μt. On average, we expect μt mutations with variance μt. It is important to note that the parameters μ, nucleotide substitutions per site per unit time, and t, the time, are confounded, meaning that we cannot estimate them separately but only through their product μt (number of mutations per site up to time t). We will show in the practical part of the chapter an example from literature on how to use (4.10).

$$Q=\left(\begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \\ -\mu(a\pi_C+b\pi_G+c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ g\mu\pi_A & -\mu(g\pi_A+d\pi_G+e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ h\mu\pi_A & i\mu\pi_C & -\mu(h\pi_A+j\pi_C+f\pi_T) & f\mu\pi_T \\ j\mu\pi_A & k\mu\pi_C & l\mu\pi_G & -\mu(i\pi_A+k\pi_C+l\pi_G) \end{array}\right)$$

Fig. 4.3 Instantaneous rate matrix **Q**. Each entry in the matrix represents the instantaneous substitution rate form nucleotide *i* to nucleotide *j* (rows, and columns, follow the order ***A***, ***C***, ***G***, ***T***). m is the mean instantaneous substitution rate; *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *j*, *k*, *l*, are relative rate parameters describing the relative rate of each nucleotide substitution to any other. π_A, π_C, π_T, π_G, are frequency parameters corresponding to the nucleotide frequencies (Yang, 1994). Diagonal elements are chosen so that the sum of each row is equal to zero.

4.4 Nucleotide substitutions as a *homogeneous Markov process*

The nucleotide substitution process of DNA sequences outlined in the previous section (i.e. the Poisson process) can be generalized to a so-called ***Markov process*** which uses a ***Q matrix*** that specifies the relative rates of change of each nucleotide along the sequence (see next section for the mathematical details). The most general form of the **Q** matrix is shown in Fig. 4.3. Rows follow the order A, C, G, and T, so that, for example, the second term of the first row is the instantaneous rate of change from base A to base C. This rate is given by the product of μ, the mean instantaneous substitution rate, times the frequency of base A, times *a*, a relative rate parameter describing, in this case, how often the substitution A to C occurs during evolution with respect to the other possible substitutions. In other words, each non-diagonal entry in the matrix represents the flow from nucleotide *i* to *j*, while the diagonal elements are chosen in order to make the sum of each row equal to zero since they represent the total flow that leaves nucleotide *i*.

Nucleotide substitution models like the ones summarized by the **Q** matrix in Fig. 4.3 belong to a general class of models known as *time-homogeneous time-continuous stationary* ***Markov models***. When applied to modeling nucleotide substitutions, they all share the following set of underlying assumptions:

(1) At any given site in a sequence, the rate of change from base *i* to base *j* is independent from the base that occupied that site prior *i* (*Markov property*).
(2) Substitution rates do not change over time (***homogeneity***).
(3) The relative frequencies of A, C, G, and T (π_A, π_C, π_G, π_T) are at equilibrium (***stationarity***).

These assumptions are not necessarily biologically plausible. They are the consequence of modeling substitutions as a stochastic process. Within this general framework, we can still develop several sub-models. In this book, however, we will examine only the so-called time-reversible models, i.e. those ones assuming for any

$$Q = \begin{pmatrix} -\tfrac{3}{4}\mu & \tfrac{1}{4}\mu & \tfrac{1}{4}\mu & \tfrac{1}{4}\mu \\ \tfrac{1}{4}\mu & -\tfrac{3}{4}\mu & \tfrac{1}{4}\mu & \tfrac{1}{4}\mu \\ \tfrac{1}{4}\mu & \tfrac{1}{4}\mu & -\tfrac{3}{4}\mu & \tfrac{1}{4}\mu \\ \tfrac{1}{4}\mu & \tfrac{1}{4}\mu & \tfrac{1}{4}\mu & -\tfrac{3}{4}\mu \end{pmatrix}$$

Fig. 4.4 Instantaneous rate matrix **Q** for the Jukes and Cantor model (JC69).

two nucleotides that the rate of change from i to j is always the same than from j to i ($a = g$, $b = h$, $c = i$, $d = j$, $e = k$, $f = g$ in the **Q** matrix). As soon as the **Q** matrix, and thus the evolutionary model, is specified, it is possible to calculate the probabilities of change from any base to any other during the evolutionary time t, $\mathbf{P}(t)$, by computing the matrix exponential

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) \tag{4.11}$$

(for an intuitive explanation of why, consider in analogy the result that led us to (4.6)). When the probabilities $\mathbf{P}(t)$ are known, this equation can also be used to compute the expected genetic distance between two sequences according to the evolutionary models specified by the **Q** matrix. In the next section we will show how to calculate $\mathbf{P}(t)$ and the expected genetic distance in case of the simple Jukes and Cantor model of evolution (Jukes & Cantor, 1969), whereas for more complex models only the main results will be discussed.

4.4.1 The Jukes and Cantor (JC69) model

The simplest possible nucleotide substitution model, introduced by Jukes and Cantor in 1969 (JC69), specifies that the equilibrium frequencies of the four nucleotides are 25% each, and that during evolution any nucleotide has the same probability to be replaced by any other. These assumptions correspond to a **Q** matrix with $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$, and $a = b = c = g = e = f = 1$ (see Fig. 4.4). The matrix fully specifies the rates of change between pairs of nucleotides in the JC69 model. In order to obtain an analytical expression for p we need to know how to compute $P_{ii}(t)$, the probability of a nucleotide to remain the same during the evolutionary time t, and $P_{ij}(t)$, the probability of replacement. This can be done by solving the exponential $\mathbf{P}(t) = \exp(\mathbf{Q}t)$ (4.11), with **Q** as the instantaneous rate matrix for the JC69 model. The detailed solution requires the use of matrix algebra (see next section for the relevant mathematics), but the result is quite straightforward:

$$P_{ii}(t) = 1/4 + 3/4\exp(-\mu t) \tag{4.12a}$$

$$P_{ij}(t) = 1/4 - 1/4\exp(-\mu t) \tag{4.12b}$$

From these equations, we obtain for two sequences that diverged t time units ago:

$$p = 3/4[1 - \exp(-2\mu t)] \tag{4.13}$$

and solving for μt we get:

$$\mu t = -1/2\, log(1 - 4/3\, p) \tag{4.14}$$

Thus the right-hand side gives the number of substitutions occurring in both of the lines leading to the shared ancestral sequence. The interpretation of the above formula is very simple. Under the JC69 model $3/4\mu t$ is the number of substitutions that actually occurred per site (see **Q** matrix in Fig. 4.4). Therefore, $d = 2$ ($3/4\ \mu t$) is the genetic distance between two sequences sharing a common ancestor. On the other hand, p is interpreted as the observed distance or p-distance, i.e. the observed proportion of different nucleotides between the two sequences (see Section 4.4). Substituting μt with $2/3d$ in (4.14) and re-arranging a bit, we finally obtain the Jukes and Cantor correction formula for the genetic distance d between two sequences:

$$d = -3/4 \ln(1 - 4/3\, p) \tag{4.15a}$$

It can also be demonstrated that the variance $V(d)$ will be given by:

$$V(d) = 9\, p(1 - p)/(3 - 4\, p)^2 n \tag{4.15b}$$

(Kimura & Ohta, 1972). More complex nucleotide substitution models can be implemented depending on which parameters of the **Q** matrix we decide to estimate (see Section 4.6 below). In the practical part of this chapter we will see how to calculate pairwise genetic distances for the example data sets according to different models. Chapter 10 will discuss a statistical test that can help select the best-fitting nucleotide substitution model for a given data set.

4.5 Derivation of Markov Process *(*optional*)

In this section we show how the stochastic process for nucleotide substitution can be derived from first principles such as detailed balance and the Chapman–Kolmogorov equations. To model the substitution process on the DNA level, it is commonly assumed that a replacement of one nucleotide by another occurs randomly and independently, and that nucleotide frequencies π_i in the data do not change over time and from sequence to sequence in an alignment. Under these assumptions the mutation process can be modeled by a *time-homogeneous stationary* ***Markov process***.

In this model, essentially each site in the DNA sequence is treated as a random variable with a discrete number n of possible states. For nucleotides there are four states ($n = 4$), which correspond to the four nucleotide bases A, C, G, and T. The ***Markov process*** specifies the transition probabilities from one state to the other, i.e. it gives the probability of the replacement of nucleotide i by nucleotide j after a certain period of time t. These probabilities are collected in the transition probability matrix $\mathbf{P}(t)$. Its components $P_{ij}(t)$ satisfy the conditions:

$$\sum_{j=1}^{n} P_{ij}(t) = 1 \tag{4.16}$$

and

$$P_{ij}(t) > 0 \quad \text{for } t > 0 \tag{4.17}$$

Moreover, it also fulfills the requirement that

$$\mathbf{P}(t + s) = \mathbf{P}(t) + \mathbf{P}(s) \tag{4.18}$$

known as the Chapman–Kolmogorov equation, and the initial condition

$$P_{ij}(0) = 1, \quad \text{for } i = j \tag{4.19a}$$

$$P_{ij}(0) = 0, \quad \text{for } i \neq j \tag{4.19b}$$

For simplicity it is also often assumed that the substitution process is reversible, i.e. that

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \tag{4.20}$$

holds. This additional condition on the substitution process, known as detailed balance, implies that the substitution process has no preferred direction. For small t the transition probability matrix $\mathbf{P}(t)$ can be linearly approximated (Taylor expansion) by:

$$\mathbf{P}(t) \approx \mathbf{P}(0) + t\mathbf{Q} \tag{4.21}$$

where $\mathbf{Q}$ is called rate matrix. It provides an infinitesimal description of the substitution process. In order not to violate (4.16) the rate matrix $\mathbf{Q}$ satisfies

$$\sum_{i=1}^{n} Q_{ij} = 0 \tag{4.22}$$

which can be achieved by defining

$$Q_{ii} = -\sum_{i \neq j}^{n} Q_{ij} \tag{4.23}$$

Note that $Q_{ij} > 0$, since we can interpret them as the flow from nucleotide i to j, $Q_{ii} < 0$ is then the total flow that leaves nucleotide i, hence it is less than zero. In contrast to **P**, the rate matrix **Q** does not comprise probabilities. Rather, it describes the amount of change of the substitution probabilities per unit time. As can be seen from (4.21) the rate matrix is the first derivative of $\mathbf{P}(t)$, which is constant for all t in a time-homogeneous Markov process. The total number of substitutions per unit time, i.e. the total rate μ, is

$$\mu = -\sum_{i=1}^{n} \pi_i Q_{ii} \tag{4.24}$$

so that the number of substitutions during time t equals $d = \mu t$. Note that, in this equation, μ and t are confounded. As a result, the rate matrix can be arbitrarily scaled, i.e. all entries can be multiplied with the same factor without changing the overall substitution pattern, only the unit in which time t is measured will be affected. For a reversible process **P**, the rate matrix **Q** can be decomposed into rate parameters R_{ij} and nucleotide frequencies π_i.

$$Q_{ij} = R_{ij}, \quad \pi_j, \text{ for } i \neq j \tag{4.25}$$

The matrix $\mathbf{R} = R_{ij}$ is symmetric, $R_{ij} = R_{ji}$, and has vanishing diagonal entries, $R_{ii} = 0$.

From the Chapman–Kolmogorov (4.18) we get the forward and backward differential equations:

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q} = \mathbf{Q}\mathbf{P}(t) \tag{4.26}$$

which can be solved under the initial condition (4.19a,b) to give

$$\mathbf{P}(t) = \exp(t\mathbf{Q}). \tag{4.27}$$

For a reversible rate matrix **Q** (4.20) this quantity can be computed by spectral decomposition (Bailey, 1964)

$$P_{ij}(t) = \sum_{m=1}^{n} \exp(\lambda_m t) U_{mi} U_{jm}^{-1} \tag{4.28}$$

where the λ_i are the eigenvalues of **Q**, $\mathbf{U} = (U_{ij})$ is the matrix with the corresponding eigenvectors, and $\mathbf{U}^{-1}$ is the inverse of **U**.

Choosing a model of nucleotide substitution in the framework of a reversible rate matrix amounts to specifying explicit values for the matrix **R** and for the frequencies π_i. Assuming n different states, the model has $n-1$ independent frequency parameters π_i (as $\sum \pi_i = 1$) and $[n(n-1)/2]-1$ independent rate parameters (as the scaling of the rate matrix is irrelevant, and $R_{ij} = R_{ji}$ and $R_{ii} = 0$). Thus, in the case of nucleotides ($n = 4$) the substitution process is governed by 3 independent frequency parameters π_i and 5 independent rate parameters R_{ij}.

4.5.1 Inferring the expected distances

Once the rate matrix **Q** or, equivalently, the parameters π_{i} and R_{ij}, are fixed, the substitution model provides the basis to statistically infer the genetic distance d between two DNA sequences. Two different techniques exist, both of which are widely used. The first approach relies on computing the exact relationship between d and p for the given model (see Fig. 4.2). The probability that a substitution is observed after time t is

$$p = 1 - \sum_{i=1}^{n} \pi_i P_{ii}(t) \tag{4.29}$$

With the definition of μ (equation 4.24) and $t = d/\mu$ we obtain

$$p = 1 - \sum_{i=1}^{n} \pi_i P_{ii} \left(- \frac{d}{\sum_{i=1}^{n} \pi_i Q_{ii}} \right) \tag{4.30}$$

This equation can then be used to construct a *method of moments estimator* of the expected distance by solving for d and estimating p (observed proportion of different sites) from the data. This formula is a generalization of (4.13).

Another way to infer the expected distance between two sequences is to use a maximum-likelihood approach. This requires the introduction of a ***likelihood function*** $L(d)$ (see Chapter 6 for more details). The likelihood is the probability to observe the two sequences given the distance d. It is defined as

$$L(d) = \prod_{s=1}^{l} \pi_{x_{\mathrm{A}(s)}} P_{x_{\mathrm{A}(s)} x_{\mathrm{B}(s)}} \left(\frac{d}{\mu} \right) \tag{4.31}$$

where $x_{\mathrm{A}(s)}$ is the state at site $s = 1, \ldots, l$ in sequence A and $P_{x_{\mathrm{A}(s)} x_{\mathrm{B}(s)}}(\frac{d}{\mu})$ is the transition probability. A value for d that maximizes $L(d)$ is called a ***maximum likelihood estimate*** (MLE) of the genetic distance. To find this estimate, numerical optimization routines are employed, as analytical results are generally not available. Estimates of error of the inferred genetic distance can be computed for both the methods of moments estimator (4.30) and the likelihood estimator (4.31) using standard statistical techniques. The so-called "delta" method can be employed to compute the variance of an estimate obtained from (4.30), and the *Fisher information criterion* is helpful to estimate the asymptotic variance of maximum likelihood estimates. For details we refer to standard statistics textbooks.

4.6 Nucleotide substitution models

If all of the eight free parameters of a reversible nucleotide rate matrix **Q** are specified, the *general time reversible* model (GTR) is derived (see Fig. 4.5). However,

$$Q=\begin{array}{c}\begin{array}{cccc}\mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T}\end{array}\\ \begin{pmatrix} -\mu(a\pi_C+b\pi_G+c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & -\mu(a\pi_A+d\pi_G+e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & -\mu(b\pi_A+d\pi_C+f\pi_T) & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & -\mu(c\pi_A+e\pi_C+f\pi_G) \end{pmatrix}\end{array}$$

Fig. 4.5 **Q** matrix of the general time reversible (GTR) model of nucleotide substitutions.

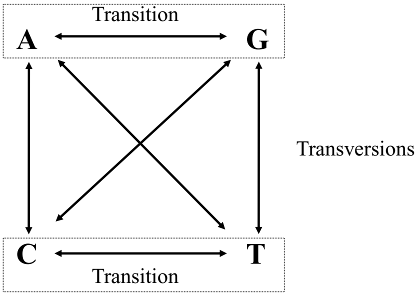

Fig. 4.6 The six possible substitution patterns for nucleotide data.

it is often desirable to reduce the number of free parameters, in particular when parameters are unknown (and hence need to be estimated from the data). This can be achieved by introducing constraints reflecting some (approximate) symmetries of the underlying substitution process. For example, nucleotide exchanges all fall into two major groups (see Fig. 4.6). Substitutions where a purine is exchanged by a pyrimidine or vice versa (A↔C, A↔T, C↔G, G↔T) are called ***transversions*** (***Tv***), all other substitutions are ***transitions*** (***Ts***). Additionally, one may wish to distinguish between substitutions among purine and pyrimidines, i.e. purine transitions (A↔G) $\mathrm{Ts_R}$, and pyrimidine transitions (C↔T) $\mathrm{Ts_Y}$. When these constraints are imposed, only two independent rate parameters (out of five) remain, the ratio κ of the Ts and Tv rates and the ratio γ of the two types of transition rates. This defines the Tamura–Nei (TN93) model (Tamura & Nei, 1993) which can be written as

$$R_{ij}^{\mathrm{TN}} = \kappa\left(\frac{2\gamma}{\gamma+1}\right) \quad \text{for } \mathrm{Ts_Y} \tag{4.32a}$$

$$R_{ij}^{\mathrm{TN}} = \kappa\left(\frac{2}{\gamma+1}\right) \quad \text{for } \mathrm{Ts_R} \tag{4.32b}$$

$$R_{ij}^{\mathrm{TN}} = 1 \quad \text{for Tv} \tag{4.32c}$$

If $\gamma = 1$ and therefore the purine and pyrimidine transitions have the same rate, this model reduces to the HKY85 model (Hasegawa *et al.*, 1985)

$$R_{ij}^{HKY} = \kappa \quad \text{for Ts} \tag{4.33a}$$

$$R_{ij}^{HKY} = 1 \quad \text{for Tv} \tag{4.33b}$$

If the base frequencies are uniform ($p_i = 1/4$), the HKY85 model further reduces to the Kimura 2-parameter (K80) model (Kimura, 1980). For $\kappa = 1$, the HKY85 model is called F81 model (Felsenstein, 1981) and the K80 model degenerates to the Jukes and Cantor (JC69) model. The F84 model (Thorne *et al.*, 1992; Felsenstein, 1993) is also a special case of the TN93 model. It is similar to the HKY85 model but uses a slightly different parameterization. A single parameter τ generates the κ and γ parameters of the TN93 model (4.32a,b,c) in the following fashion. First, the quantity

$$\rho = \frac{\pi_R\pi_Y[\pi_R\pi_Y\tau - (\pi_A\pi_G + \pi_C\pi_T)]}{(\pi_A\pi_G\pi_Y + \pi_C\pi_T\pi_R)} \tag{4.34}$$

is computed which then determines both

$$\kappa = 1 + \frac{1}{2}\rho\left(\frac{1}{\pi_R} + \frac{1}{\pi_Y}\right) \tag{4.35}$$

and

$$\gamma = \frac{\pi_Y + \rho}{\pi_Y}\frac{\pi_R}{\pi_R + \rho} \tag{4.36}$$

of the TN93 model, where π_A, π_C, etc. are the base frequencies, π_R and π_Y are the frequency of purines and pyrimidines.

The hierarchy of the substitution models discussed above is shown in Fig. 4.7.

4.6.1 Rate heterogeneity among sites

It is a well-known phenomenon that the rate of nucleotide substitution can vary substantially for different positions in a sequence. For example, in protein coding genes third codon positions mutate usually faster than first positions, which, in turn, mutate faster than second positions. Such a pattern of evolution is commonly explained by the presence of different evolutionary forces for the sites in question. In the previous sections we have ignored this problem and silently assumed rate homogeneity over sites, but rate heterogeneity can play a crucial part in the inference of genetic distances. To account for the site-dependent rate variation, a plausible

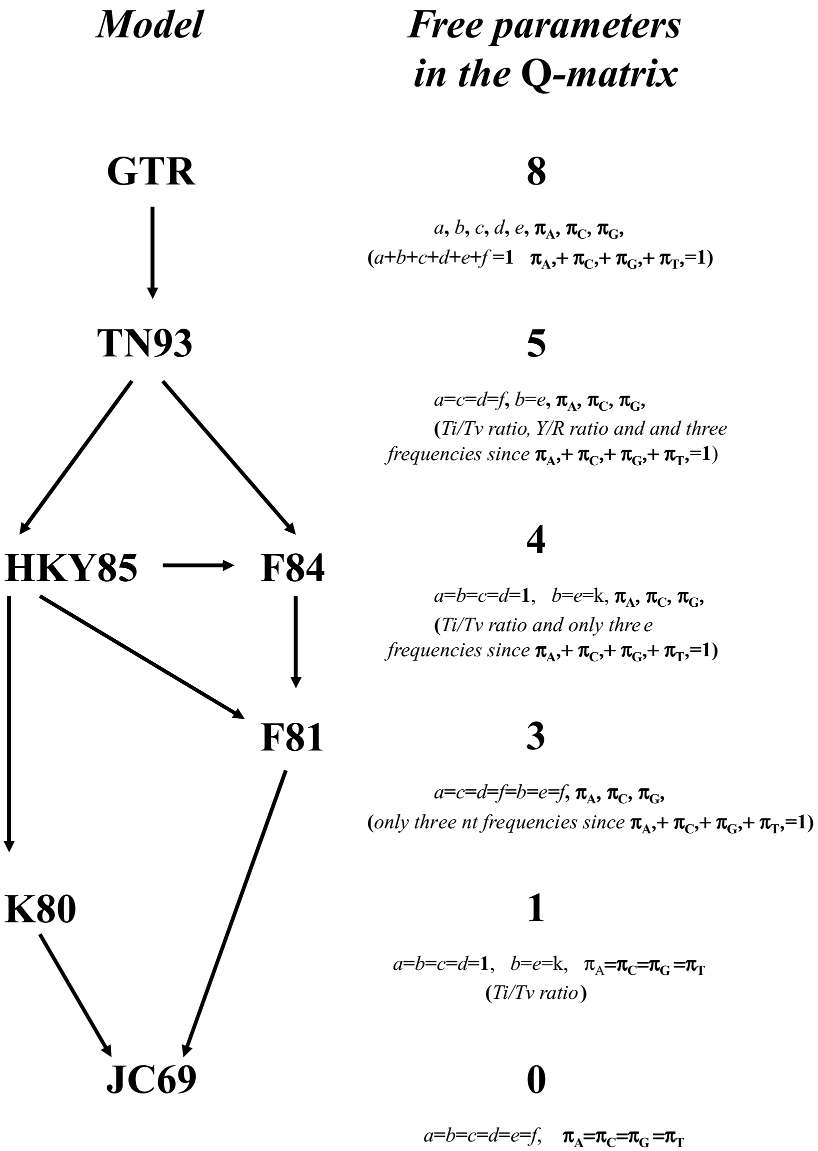
Model
Free parameters in the Q-matrix
GTR
8
$a, b, c, d, e, \pi_A, \pi_C, \pi_G,$
$(a+b+c+d+e+f=1 \quad \pi_A,+\pi_C,+\pi_G,+\pi_T,=1)$
TN93
5
$a=c=d=f, b=e, \pi_A, \pi_C, \pi_G,$
(*Ti/Tv ratio, Y/R ratio and and three frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)
HKY85
F84
4
$a=b=c=d=1, \quad b=e=\mathrm{k}, \pi_A, \pi_C, \pi_G,$
(*Ti/Tv ratio and only thre e frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)
F81
3
$a=c=d=f=b=e=f, \pi_A, \pi_C, \pi_G,$
(*only three nt frequencies since* $\pi_A,+\pi_C,+\pi_G,+\pi_T,=1$)
K80
1
$a=b=c=d=1, \quad b=e=\mathrm{k}, \quad \pi_A=\pi_C=\pi_G=\pi_T$
(*Ti/Tv ratio*)
JC69
0
$a=b=c=d=e=f, \quad \pi_A=\pi_C=\pi_G=\pi_T$

Fig. 4.7 Hierarchy of nucleotide substitution models.

model for distribution of rates over sites is required. The common approach is to use a gamma (Γ) distribution with expectation 1.0 and variance $1/\alpha$.

$$Pdf(r) = \alpha^{\alpha} r^{\alpha-1}/\exp(\alpha r)\Gamma(\alpha) \tag{4.37}$$

By adjusting the shape parameter α, the Γ-distribution accommodates for varying degree of rate heterogeneity (see Fig. 4.8). For $\alpha > 1$, the distribution is bell-shaped

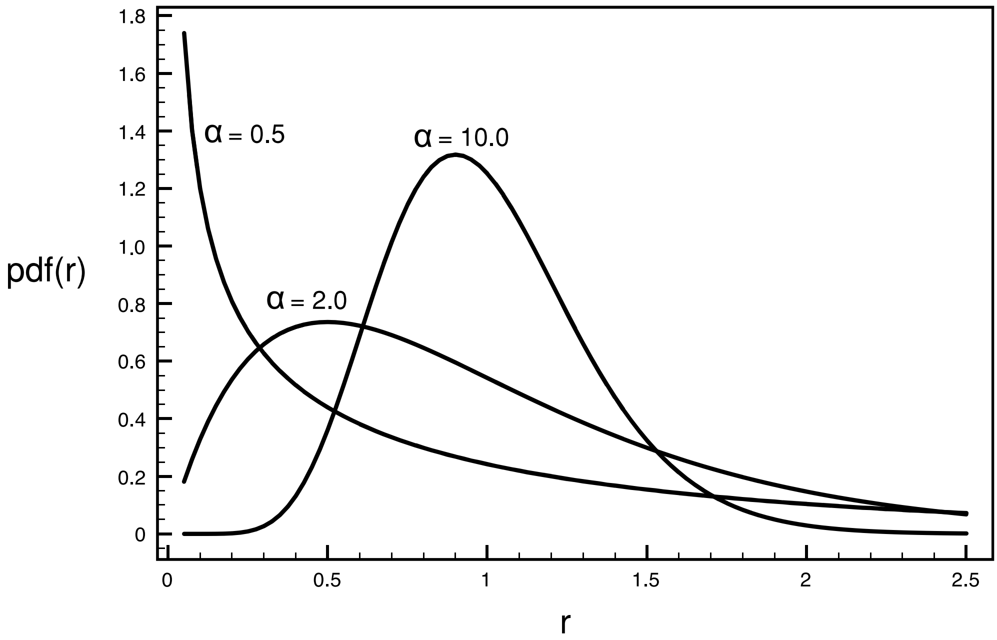

Fig. 4.8 Different shapes of the Γ-distribution depending on the α-shape parameter.

and models weak rate heterogeneity over sites. The relative rates drawn from this distribution are all close to 1.0. For $\pi < 1$, the Γ-distribution takes on its characteristic L-shape, which describes situations of strong rate heterogeneity, i.e. some positions have very large substitution rates but most other sites are practically invariable.

Rather than using the continuous Γ-distribution it is computationally more efficient to assume a discrete Γ-distribution with a finite number *c* of equally probable rates $q_1, q_2, \ldots, q_c$. Usually, 4–8 discrete categories are enough to obtain a good approximation of the continuous function (Yang, 1994b). A further generalization is provided by the approach of Kosakovsky *et al.* (2005) who propose a two-stage hierarchical Beta–Gamma model for fitting the rate distribution across sites.

PRACTICE

Marco Salemi

4.7 Software packages

A large number of software packages are available to compute genetic distances from DNA sequences. An exhaustive list is maintained by Joe Felsenstein at *http://evolution.genetics.washington.edu/PHYLIP/software.html*. Among others, the programs PHYLIP, PAUP* (see Chapter 8), TREE-PUZZLE (Schmidt *et al.*, 2002; see Chapter 6), MEGA4 (Kumar *et al.*, 1993), DAMBE (Xia, 2000; see Chapter 20), and PAML (Yang, 2000; see Chapter 11) provide the possibility to infer genetic distances and will be discussed in this book.

Phylogeny Inference Package (PHYLIP), was one of the first freeware phylogeny software to be developed (Felsenstein, 1993). It is a package consisting of several programs for calculating genetic distances and inferring phylogenetic trees according to different algorithms. Pre-compiled executables files are available for Windows3.x/95/98, Windows Vista, pre-386 and 386 DOS, Macintosh (non-PowerMac), and MacOSX. A complete description of the package including the instructions for installation on different machines can be found at *http://evolution.gs.washington.edu/phylip.html*. The PHYLIP software modules that will be discussed throughout the book are briefly summarized in Box 4.1.

TREE-PUZZLE was originally developed to reconstruct phylogenetic trees from molecular sequence using maximum likelihood with a fast tree-search algorithm called ***quartet puzzling*** (Strimmer & von Haeseler, 1995; see Chapter 6). The program also computes pairwise maximum likelihood distances according to a number of models of nucleotide substitution. Versions of TREE-PUZZLE for UNIX, MacOSX, and Windows95/98/NT can be freely downloaded from the TREE-PUZZLE web page at *http://www.TREE–PUZZLE.de/*. The quartet-puzzling algorithm to infer phylogenetic trees will be described in detail in Chapter 6. In what follows, it is shown how to compute genetic distances according to different evolutionary models using TREE-PUZZLE.

Installation of these programs should make the PHYLIP folder and the TREE-PUZZLE folder visible on your local computer. These folders contain several files, including executable applications, documentation, and source codes. PHYLIP version 3.66 has three subdirectories: doc, exe, src; the executables are in the exe folder. The doc directory contains an extensive documentation, whereas the source codes are in src. In TREE-PUZZLE version 5.3 the program executable can be found in the src folder within the TREE-PUZZLE folder. Any of the software modules within PHYLIP and TREE-PUZZLE works in the same basic way: they need a

Box 4.1 The Phylip software package

Phylip contains several executables for analyzing molecular as well as morphological data and drawing phylogenetic trees. However, only some of the software modules included in the package will be discussed in this book; that is, those dealing with DNA and amino acid sequences analysis. These modules are summarized herein. Information about the other modules can be found in the Phylip documentation available with the program at *http://evolution.genetics.washington.edu/PHYLIP/software.html.*

Phylip executable	input data	type of analysis	book chapters
`DNAdist.exe`	aligned DNA sequences	calculates genetic distances using different nucleotide substitution models	5
`ProtDist.exe`	aligned protein sequences	calculates genetic distances using different amino acid substitution matrices	See also 9
`Neighbor.exe`	genetic distances	calculates NJ or UPGMA trees	5
`Fitch.exe`	genetic distances	calculates Fitch–Margoliash trees	5
`Kitch.exe`	genetic distances	calculates Fitch–Margoliash trees assuming a molecular clock	5
`DNAML.exe`	aligned DNA sequences	calculates maximum likelihood trees	See also 6
`ProtPars.exe`	aligned protein sequences	calculates maximum parsimony trees	See also 8
`SeqBoot.exe`	aligned DNA or protein sequences	generates bootstrap replicates of aligned DNA or protein sequences	5
`Consense.exe`	phylogenetic trees (usually obtained from bootstrap replicates)	generates a consensus tree	5

file containing the input data, for example, aligned DNA sequences in Phylip format (see Box 2.2 in Chapter 2 for different file formats), to be placed in the same directory where the program resides; it produces one or more output files in text format (usually called outfile and outtree), containing the analysis result. By default, any application reads the data from a file named infile (no extension type!) if such a file is present in the same directory, otherwise the user is asked to enter the name of the input file. Other details about Phylip modules are summarized in Box 4.1.

Molecular Evolutionary Genetics Analysis (MEGA4) is a sophisticated program originally developed to carry out a number of distance and parsimony based analysis of both nucleotide and amino acid sequences (Kumar *et al.*, 2004). One of the advantages of MEGA4 is the possibility to calculate standard errors of distance estimates either using analytical formulas derived for a specific evolutionary model or using ***bootstrap*** analysis. The latest version of the program also includes an excellent data editor that allows multiple sequences to be aligned using a native implementation of the Clustal algorithm (see Chapter 3) and to manually edit aligned and unaligned sequences. The software is freeware and can be downloaded from *http://www.megasoftware.net/overview.html*. The website also contains a detailed overview of the program capabilities, installation instructions, and extensive online documentation. MEGA4 only works under Windows, but it can be run in Mac by using a PC emulator (which, unfortunately, makes the program very slow) or under the Windows partition installed in the new Macs with Intel processors.

The aligned sequences in PHYLIP format (required for the analysis with PHYLIP or TREE-PUZZLE) or FASTA format (required for the analysis in MEGA4) can be downloaded from *www.thephylogenetichandbook.org*.

4.8 Observed vs. estimated genetic distances: the JC69 model

In what follows, we will use as an example the primate Trim5α sequences also used in Chapter 3 and Chapter 11 (‘`Primates.phy`’). Figure 4.9a shows a matrix with pairwise p-distances, i.e. number of different sites between two sequences divided by the sequence length, for the primates data. The matrix is written in lower-triangular form. Comparison of Human and Chimp sequences reveals 13 point mutations over 1500 nucleotides giving an observed distance $p = 13/1500 = 0.008667$. The estimated genetic distance according to the JC69 model, obtained by substituting the observed distance in (4.15a) (see Section 4.4.1), is 0.008717. The two measures agree up to the fourth digit. This is not surprising because, as we have seen in the first part of the chapter, the relationship between observed and estimated genetic distance is approximately linear when the evolutionary rate is low and sequences share a relatively recent common ancestor. If mutations occur according to a Poisson process (Section 4.3), we expect to see only a few nucleotide substitutions between the two sequences, and no more than one substitution per site. Simply counting the observed differences between two aligned sequences is going to give an accurate measure of the genetic distance. Human and Chimp split about five million year ago (MYA) and the evolutionary rate (μ) of cellular genes is approximately 10^{-9} nucleotide substitutions per site per year (Britten, 1986). According to (4.10) the number of mutations between the Human and Chimp lineage up to time $t = 5 \times 10^6$ is Poisson distributed with parameter

$\mu = 10^{-9}$. On average, we expect to see $2\mu t = 0.01$ mutations per nucleotide site along the two phylogenetic lineages since their divergence from the common ancestor with variance 0.01. For two sequences 1500 nucleotides long we should observe a mean of 15 mutations with a 95% confidence interval of 15 ± 7.59. Indeed, the observed genetic distance between Human and Chimp in our example falls within the expected interval.

As discussed in Section 4.2, however, using the *p*-distance to compare more distantly related species is likely to lead to systematic underestimation of the genetic distance. By comparing the *p*-distance matrix in Fig. 4.9a with the JC69 distance matrix in Fig. 4.9b, we can see that the larger the observed distance the larger the discrepancy with the JC69 distance. In our alignment there are 234 mutations between Human and Squirrel (p-distance $= 0.156$). **Assuming that the JC69 model correctly describes the evolutionary process** (which, as we will see, is actually not true) the *p*-distance underestimates the actual genetic distance by about 11% ($d = 0.1749$). JC69 distances, as well as distances according to more complex evolutionary models, can be easily calculated with the DNADIST program of the PHYLIP software package.

Place the file `Primates.phy`, containing the primate nucleotide alignment in PHYLIP format, in the directory PHYLIP\EXE, or in the same directory as where the PHYLIP software module DNADIST is on your computer. Rename the file infile and start DNADIST by double clicking on its icon. A new window will appear with the following menu:

```
Nucleic acid sequence Distance Matrix program, version 3.66

Settings for this run:
  D  Distance (F84, Kimura, Jukes-Cantor, LogDet)?  F84
  G  Gamma distributed rates across sites?          No
  T  Transition/transversion ratio?                 2.0
  C  One category of substitution rates?            Yes
  W  Use weights for sites?                         No
  F  Use empirical base frequencies?                Yes
  L  Form of distance matrix?                       Square
  M  Analyze multiple data sets?                    No
  I  Input sequences interleaved?                   Yes
  0  Terminal type (IBM PC, VT52, ANSI)?            IBM PC
  1  Print out the data at start of run?            No
  2  Print indications of progress of run?          Yes

Y  to accept these or type letter for one to change
```

Type `D` followed by the `enter` key and again until the model selected is Jukes–Cantor. In the new menu the option `T` and option `F` will no longer be present,

(a)

```
[Human]
[Chimp]       0.0087
[Gorilla]     0.0133 0.0087
[Orangutan]   0.0267 0.0233 0.0253
[Gibbon]      0.0380 0.0347 0.0353 0.0340
[Rhes_cDNA]   0.0680 0.0653 0.0673 0.0687 0.0773
[Baboon]      0.0653 0.0627 0.0647 0.0660 0.0747 0.0087
[AGM_cDNA]    0.0660 0.0633 0.0640 0.0660 0.0753 0.0153 0.0127
[Tant_cDNA]   0.0660 0.0633 0.0640 0.0653 0.0760 0.0153 0.0127 0.0027
[Patas]       0.0667 0.0633 0.0653 0.0653 0.0753 0.0227 0.0200 0.0180 0.0180
[Colobus]     0.0540 0.0513 0.0533 0.0553 0.0640 0.0253 0.0240 0.0233 0.0240 0.0267
[DLangur]     0.0560 0.0533 0.0553 0.0587 0.0660 0.0260 0.0247 0.0240 0.0247 0.0273 0.0087
[PMarmoset]   0.1400 0.1380 0.1420 0.1440 0.1480 0.1587 0.1573 0.1533 0.1533 0.1593 0.1467 0.1500
[Tamarin]     0.1407 0.1393 0.1433 0.1440 0.1487 0.1613 0.1600 0.1553 0.1553 0.1613 0.1487 0.1520 0.0347
[Squirrel]    0.1560 0.1520 0.1560 0.1560 0.1593 0.1740 0.1713 0.1707 0.1707 0.1753 0.1627 0.1640 0.0727 0.0713
[Titi]        0.1453 0.1440 0.1467 0.1473 0.1493 0.1620 0.1613 0.1573 0.1587 0.1647 0.1500 0.1520 0.0600 0.0600 0.0740
[Saki]        0.1367 0.1360 0.1387 0.1407 0.1440 0.1560 0.1547 0.1527 0.1533 0.1580 0.1440 0.1473 0.0567 0.0573 0.0727 0.0393
[Howler]      0.1533 0.1487 0.1533 0.1527 0.1573 0.1733 0.1727 0.1700 0.1700 0.1720 0.1620 0.1640 0.0800 0.0800 0.0907 0.0740 0.0787
[Spider]      0.1420 0.1400 0.1420 0.1400 0.1467 0.1567 0.1553 0.1527 0.1520 0.1567 0.1447 0.1480 0.0633 0.0640 0.0740 0.0607 0.0633 0.0547
[Woolly]      0.1453 0.1447 0.1467 0.1440 0.1540 0.1607 0.1593 0.1547 0.1540 0.1627 0.1500 0.1520 0.0727 0.0707 0.0800 0.0653 0.0687 0.0627 0.0300
```

(b)

```
[Human]
[Chimp]      0.0087
[Gorilla]    0.0135 0.0087
[Orangutan]  0.0272 0.0237 0.0258
[Gibbon]     0.0390 0.0355 0.0362 0.0348
[Rhes_cDNA]  0.0713 0.0684 0.0705 0.0720 0.0816
[Baboon]     0.0684 0.0654 0.0676 0.0691 0.0787 0.0087
[AGM_cDNA]   0.0691 0.0662 0.0669 0.0691 0.0794 0.0155 0.0128
[Tant_cDNA]  0.0691 0.0662 0.0669 0.0684 0.0801 0.0155 0.0128 0.0027
[Patas]      0.0698 0.0662 0.0684 0.0684 0.0794 0.0230 0.0203 0.0182 0.0182
[Colobus]    0.0560 0.0532 0.0553 0.0575 0.0669 0.0258 0.0244 0.0237 0.0244 0.0272
[DLangur]    0.0582 0.0553 0.0575 0.0611 0.0691 0.0265 0.0251 0.0244 0.0251 0.0278 0.0087
[PMarmoset]  0.1550 0.1525 0.1574 0.1599 0.1649 0.1783 0.1766 0.1715 0.1715 0.1791 0.1632 0.1674
[Tamarin]    0.1558 0.1541 0.1591 0.1599 0.1657 0.1817 0.1800 0.1741 0.1741 0.1817 0.1657 0.1699 0.0355
[Squirrel]   0.1749 0.1699 0.1749 0.1749 0.1791 0.1980 0.1945 0.1936 0.1936 0.1997 0.1834 0.1851 0.0764 0.0750
[Titi]       0.1615 0.1599 0.1632 0.1640 0.1665 0.1825 0.1817 0.1766 0.1783 0.1859 0.1674 0.1699 0.0625 0.0625 0.0779
[Saki]       0.1509 0.1501 0.1533 0.1558 0.1599 0.1749 0.1732 0.1707 0.1715 0.1774 0.1599 0.1640 0.0589 0.0596 0.0764 0.0404
[Howler]     0.1715 0.1657 0.1715 0.1707 0.1766 0.1971 0.1962 0.1928 0.1928 0.1954 0.1825 0.1851 0.0846 0.0846 0.0966 0.0779 0.0831
[Spider]     0.1574 0.1550 0.1574 0.1550 0.1632 0.1757 0.1741 0.1707 0.1699 0.1757 0.1607 0.1649 0.0662 0.0669 0.0779 0.0633 0.0662 0.0568
[Woolly]     0.1615 0.1607 0.1632 0.1599 0.1724 0.1808 0.1791 0.1732 0.1724 0.1834 0.1674 0.1699 0.0764 0.0742 0.0846 0.0684 0.0720 0.0654 0.0306
```

Fig. 4.9 (a) Pairwise *p*-distance and (b) Jukes-and-Cantor matrix for the primate TRIM5α sequences.

since under the JC69 model all nucleotide substitutions are equally likely and base frequency is assumed to be 0.25 for each base (see Section 4.4.1). Type `y` followed by the `enter` key to carry out the computation of genetic distances. The result is stored in a file called `outfile`, which can be opened with any text editor. The format of the output matrix, square or lower triangular, can be chosen before starting the computation by selecting option `L`. Of course, each pairwise distance can be obtained by replacing p in (4.15a) with the observed distance given in Fig. 4.9. That is exactly what the program DNAdist with the current settings has done: first it calculates p-distances, and then it uses the JC69 formula to convert them in genetic distances.

4.9 Kimura 2-parameter (K80) and F84 *genetic distances*

The K80 model relaxes one of the main assumptions of the JC69 model allowing for a different instantaneous substitution rate between transitions and transversions ($a = c = d = f = 1$ and $b = e = \kappa$ in the $\mathbf{Q}$ matrix) (Kimura, 1980). Similarly to what has been done in Section 4.4, by solving the exponential $\mathbf{P}(t) = \exp(\mathbf{Q}t)$ for $\mathbf{P}(t)$ the K80 correction formula for the expected genetic distance between two DNA sequences is obtained:

$$d = 1/2\ln(1/(1 - 2P - Q)) + 1/4\ln(1/(1 - 2Q)) \tag{4.38a}$$

where P and Q are the proportion of the transitional and transversional differences between the two sequences, respectively. The variance of the K80 distances is calculated by:

$$V(d) = 1/n[(\mathrm{A}^2 P + \mathrm{B}^2 Q - (\mathrm{A}P + \mathrm{B}Q)^2] \tag{4.38b}$$

with $\mathrm{A} = 1/(1\text{–}2P\text{-}Q)$ and $\mathrm{B} = 1/2[(1/1\text{–}2P\text{–}Q) + (1/1\text{–}2Q)]$.

K80-distances can be obtained with `DNAdist` by choosing `Kimura 2-parameter` within the D option. The user can input an empirical ***transition/transversion ratio*** (***Ti/Tv***) by selecting option T from the main menu. *Ti*/*Tv* is the probability of *any* transition (over a single unit of time) divided by the probability of *any* transversion (over a single unit of time), which can be obtained by dividing the sum of the probabilities of transitions (four terms) by the sum of the probabilities of transversions (eight terms). The default value for *Ti*/*Tv* in `DNAdist` is 2.0. Considering that there are twice more possible transversions than transitions, the default value of *Ti*/*Tv* = 2.0 in DNAdist assumes that, during evolution, transitional changes are about four times more likely than transversional ones. When an empirical *Ti*/*Tv* value for the set of organisms under investigation is not known from the literature, it is good practice to estimate it directly from the data. A general strategy to estimate the *Ti*/*Tv* ratio of aligned DNA sequences will be discussed in Chapter 6. Note that some programs use the transition/transversion *rate* ratio (κ) instead of the expected *Ti*/*Tv* ratio, which is the instantaneous rate of transitions divided by the instantaneous rate of transversions and does not involve the equilibrium base frequencies. Depending on the equilibrium base frequencies, this rate ratio will be about twice the *Ti*/*Tv* ratio.

The genetic distance estimated with the K80 model ($Ti/Tv = 2.0$) between Human and Chimp (0.008722), is still not significantly different from the *p*-distance for the same reasons discussed above. The K80 distance between Squirrel and Human is 0.180, which is slightly larger than the one estimated by the JC69. However, even small changes in the distance can influence the topology of phylogenetic trees inferred with distance-based methods. The K80 model still relies on very restricted assumptions such as that of equal frequency of the four bases at equilibrium. The HKY85 (Hishino *et al.*, 1985) and F84 (Felsenstein, 1984; Kishino & Hasegawa, 1989) models relax that assumption allowing for unequal frequencies; their **Q** matrices are slightly different, but both models essentially share the same set of assumptions: a bias in the rate of transitional with respect to the rate of transversional substitutions and unequal base frequencies (which are usually set to the empirical frequencies). F84 is the default model in PHYLIP version 3.66. Since the F84 model assumes unequal base frequencies, **DNADIST** empirically estimates the frequencies for each sequence (option F) and it uses the average value over all sequences to compute pairwise distances. When no is selected in option F, the program asks the user to input the base frequencies in order A, C, G, T/U separated by blank spaces. As with the K80 model, F84 scores transitional and transversional substitutions differently and it is possible to input an empirical *Ti*/*Tv* ratio with the option T.

4.10 More complex models

The TN93 model (Tamura & Nei, 1993), an extension of the F84 model, allows different nucleotide substitution rates for purine (A↔G) and pyrimidine (C↔T) transitions ($b \neq e$ in the correspondent **Q** matrix). TN93 genetic distances can be computed with **TREE-PUZZLE** by selecting from the menu: `Pairwise distances only (no tree)` in option k, and TN (Tamura & Nei, 1993) in option `m`. The menu allows the user to input empirical *Ti*/*Tv* bias and pyrimidine/purine transition bias, otherwise the program will estimate those parameters from the data set (see Chapter 6 for the details). Genetic distances can also be obtained according to simpler models. For example, by selecting HKY in option m, **TREE-PUZZLE** computes HKY85 distances. Since the JC69 model is a further simplification of the HKY85 model where equilibrium nucleotide frequencies are equal and there is no nucleotide substitution bias (see above), JC69 distances can be calculated with TREE-PUZZLE by selecting the HKY85 model and setting nucleotide frequencies equal to 0.25 each (option f in the menu) and the T*i*/T*v* ratio equal to 0.5 (option t). Note that since there are twice more transversions than transitions (see Fig. 4.6) the T*i*/T*v* ratio needs to be set to 0.5 and not to 1 in order to reduce the HKY85 model to the JC69! The distance matrix in square format is written to the `outdist` file and can be opened with any text editor. The program also outputs an `oufile` with several statistics about the data set (the file is mostly self-explanatory, but see also Section 4.13 and Chapter 6).

4.10.1 Modeling rate heterogeneity among sites

The JC69 model assumes that all sites in a sequence change at a uniform rate over time. More complex models allow particular substitutions, for example, transitions, to occur at different rate than others, for example, transversions, but any particular substitution rate between nucleotide *i* and nucleotide *j* is the same among different sites. Section 4.6.1 pointed out that such an assumption is not realistic, and it is especially violated in coding regions where different codon positions usually evolve at different rates. Replacements at the second codon position are always ***non-synonymous***, i.e. they change the encoded amino acid (see Chapter 1), whereas, because of the degeneracy of the genetic code, 65% of the possible replacements at the third codon position are ***synonymous***, i.e. no change in the encoded amino acid. Finally only 4% of the possible replacements at the first codon position are synonymous. Since mutations in a protein sequence will most of the time reduce the ability of the protein to perform its biological function, they are rapidly removed from the population by purifying selection (see Chapter 1). As a consequence, over time, mutations will accumulate more rapidly at the third rather than at the second or the first codon position. It has been shown, for example, that in each coding region of the human T-cell lymphotropic viruses (HTLVs), a group of human oncogenic retroviruses, the third codon positions evolve about eight times faster than the first and 16 times faster than the second positions (Salemi *et al.*, 2000). It is possible to model rate heterogeneity over sites by selecting the option:

```
B One category of substitution rates? Yes
```

in the main menu of DNADIST (which toggles this option to `No`) and to choose up to nine different categories of substitution rates. The program then asks for input of the relative substitution rate for each category as a non-negative real number. For example, consider how to estimate the genetic distances for the primates data set using the JC69 model, but assuming that mutations at the third position accumulate ten times faster than at the first and 20 times faster than at the second codon position. Since only the relative rates are considered, one possibility is to set the rate at the first codon position equal to 1, the rate at the second to 0.5, and the rate at the third to 10. It is also necessary to assign each site in the aligned data set to one of the three rate categories. Phylip assigns rates to sites by reading an additional input file with default name "categories" containing a string of digits (a new line or a blank can occur after any character in this string) representing the rate category of each site in the alignment. For example, to perform the calculation with the primates data set, we need to prepare a text file called categories (with no extension) containing the following string:

```
1231231231123123123l[ ...]
```

Each number in the line above represents a nucleotide position in the aligned data set: for example, the first four numbers, 1231, refer to the first four positions in the alignment and they assign the first position to rate category 1, the second position to rate category 2, the third position to rate category 3, the fourth position to rate category 1 again, and so forth. In the primates data set, sequences are in the correct reading frame, starting at the first codon position and ending at a third codon position, and there are 1500 positions. Thus the `categories` file has to be prepared in the following way:

`123123123` (and so forth up to 1500 digits)

An appropriately edited file (`Primates_cdp_categories.phy`) can be found at *www.thephylogenetichandbook.org*. After renaming this file as "`categories`," the following exercise can be carried out:

(i) Place the input files (`primates.phy` and `categories`) in the **Phylip** folder and run **DNAdist**
(ii) Select option C and type 3 to choose three different rate `categories`
(iii) At the prompt of the program asking to specify the relative rate for each category type: `1 0.5 10` and press `enter`
(iv) Choose the desired evolutionary model as usual and run the calculation.

If there is no information about the distribution and the extent of the relative substitution rates across sites, rate heterogeneity can be modeled using a G-distribution (Yang, 1994b), a negative binomial distribution (Xia, 2000) or a two-stage hierarchical Beta–Gamma model (Kosakovsky *et al.*, 2005). As discussed in Section 4.6.1, a single parameter α describes the shape of the Γ-distribution (Fig. 4.8): L-shaped for $\alpha < 1$ (strong rate heterogeneity), or bell-shaped for $\alpha > 1$ (weak rate heterogeneity). Which value of α is the most appropriate for a given data set, however, is usually not known. The next few chapters will discuss how to estimate a with different approaches and how to estimate genetic distances with Γ-distributed rates across sites (in particular, see Chapter 6). However, it is important to keep in mind that, even though different sites across the genome do change at different rates (Li, 1997), the use of a *discrete* Γ-distribution to model rate heterogeneity over sites has no biological justification. It merely reflects our ignorance about the underlying distribution of rates. It is widely used because it allows both low and high rate heterogeneity among sites to be modeled easily and flexibly by varying the *a* parameter.

Chapter 10 will show how to compare different evolutionary models. In such a way it is also possible to test whether a nucleotide substitution model implementing Γ-distributed rates across sites usually fits the data significantly better than a model assuming uniform rates. Genetic distances with Γ-models can be estimated by selecting the option `G Gamma distributed rates across sites?` in `DNAdist`. For example, to estimate F84+Γ distances for the primates data set, just run `DNAdist` as before, type G followed by the enter key (the menu will change to display `G Gamma distributed rates across sites?` Yes), and type y followed again by the enter key. Notice that, before running the analysis, the program will ask to enter the coefficient of variation CV, which is required for the specific computational implementation of G-models in Phylip. The relationship between α and CV is $CV = 1/\sqrt{a}$. Therefore, to use $a = 0.5$ we digit the value 1.414 and press the enter key. As usual, the calculated distances will be written to outfile.

To illustrate the effect of model complexity and rate heterogeneity among sites on distance estimation, the genetic distances for Human–Squirrel are shown for different substitution models in Fig. 4.10. When correcting for "multiple hits," increasingly complex models have only a marginal effect on evolutionary distances for this data set, whereas modeling rate heterogeneity (shown using circles) has a profound effect on distance estimation. The practice section of the next chapter demonstrates that this can also have an important impact on phylogenetic inference.

4.11 Estimating standard errors using Mega4

The program Mega4 can estimate genetic distance estimates using most of the nucleotide substitution models (with and without Γ-distribution) discussed above. In addition, it is possible to use Mega4 to calculate the standard errors of the estimated distances either analytically or by bootstrapping (a statistical technique that is often used to assess the robustness of phylogenetic inference, see Chapter 5). Standard errors for JC69 or K80 models are calculated in Mega4 by employing the same variance formulas given in (4.15b) and (4.37), and can be useful to perform statistical tests comparing different distance estimates (for example, to decide whether or not two sets of distances are significantly different). The idea behind bootstrapping, on the other hand, is to generate a large number of random replicates (usually 1000–10 000) by randomly resampling with replacement from

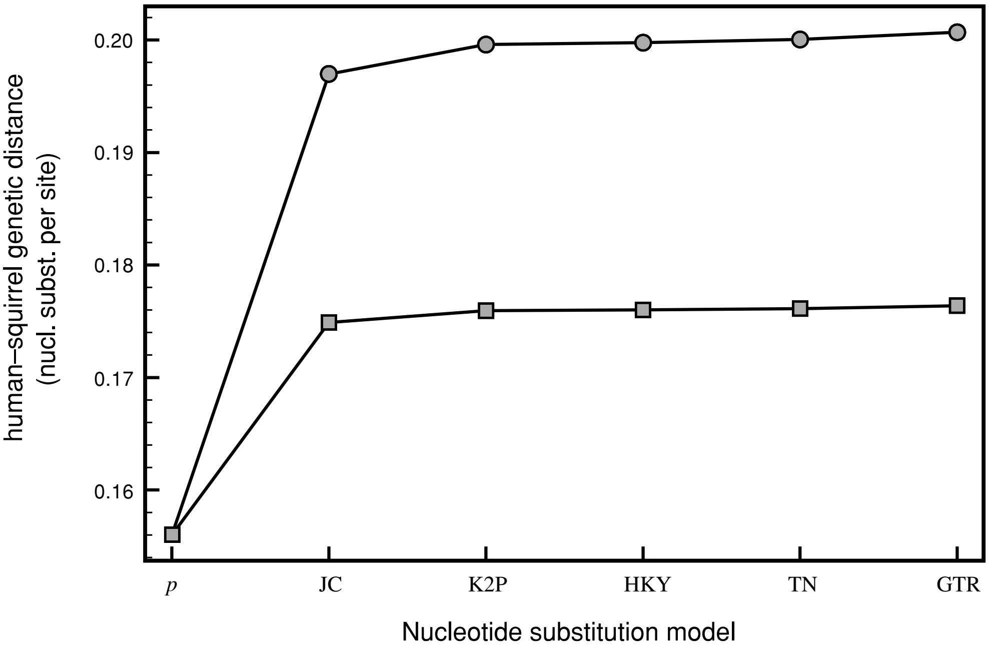

Fig. 4.10 Pairwise genetic distance for Human–Squirrel TRIM5α sequences using different evolutionary models: p = p-distance, JC = Jukes and Cantor, K2P = Kimura 2-parameter, HKY = Hasegawa–Kishino–Yano (85), TN = Tamura–Nei, GTR = General Time-Reversible. Distances computed using gamma distributed rate among sites (α = 1.0) are indicated using circles.

the original data points (for example, each column in a given alignment) so that each replicate contains exactly the same number of data points as the original set (for example, the same number of columns, i.e. sites, of the original alignment). As a consequence, some of the original data points may be absent in particular replicates, whereas others may be present more than once. The variance of each parameter (for example, the genetic distance between two sequences) is then calculated using the distribution of the parameter values obtained for each random replicate. The underlying idea of this technique is to evaluate how robust the parameter values are with respect to small changes in the original data.

To perform analyses in MEGA4, the sequence data must first be imported in the program. Sequences in FASTA format with the extension .fas under Windows should be associated automatically with MEGA4 as soon as the program is installed. By double clicking on the file "Primates.fas" the nucleotide sequences appear in the `Alignment Explorer` window of MEGA4 where they can be aligned, translated to amino acid and edited in different ways. By closing the `Alignment Explorer` window the program asks whether the user wants to save the data in MEGA format. After choosing yes, you can save the file in MEGA on your computer. The user is now asked whether to open the data file in MEGA4. By selecting yes again, the sequences are displayed in a new window called `Sequence Data Explorer`. To perform specific analyses, the user needs to close the `Sequence Data Explorer` window and select the appropriate menu from the main MEGA4 window. As an example, we will obtain JC69 genetic distances with Γ-distributed rates across sites (parameter $\alpha = 0.5$) and standard errors calculated by 1000 bootstrap replicates:

(1) From the `Distances` menu, select `Choose model...`
(2) >Click on the green square to the right of the row saying `->Model` and select `Nucleotide>Jukes and Cantor`
(3) Click on the green square to the right of the row `->Rates among sites` and select `Different(Gamma Distributed)`
(4) Click on the button to the right of the row `->Gamma Parameter`, select 0.5, click the `Ok` button at the bottom of the window.
(5) From the `Distances` menu select `Compute Pairwise`
(6) Click on the green square to the right of the row `->Compute` and select `Distance & Std. Err.`
(7) Click the button to the right of the row `Std. Err. Computation by`
(8) In the new window select `Bootstrap` and `1000 Replications`.
(9) Select the `Option Summary` tab and run the analysis by clicking the `Compute` button at the bottom of the window.

Estimated distances and standard errors appear in a new window with a square matrix providing pairwise distances in the lower triangular part (in gray) and standard errors in the upper triangular part (in blue). The matrix can be printed and/or exported in text format from the `File` menu of the `Pairwise Distances` window.

4.12 The problem of substitution saturation

It can be demonstrated that two randomly chosen, aligned DNA sequences of the same length and similar base composition would have, on average, 25% identical residues. Moreover, if gaps are allowed, as much as 50% of the residues can be identical. This is the reason why the curve in Fig. 4.2, showing the relationship between p-distance and genetic distance, reaches a plateau for p between 0.5 and 0.75 (i.e. for similarity scores between 50% and 25%). Beyond that point, it is not possible anymore to infer the expected genetic distance from the observed one, and the sequences are said to be saturated or to have reached ***substitution saturation***. The similarities between them are likely to be the result of chance alone rather than ***homology*** (common ancestry). **In other words, when full saturation is reached, the phylogenetic signal is lost, i.e. the sequences are no longer informative about the underlying evolutionary processes**. In such a situation any estimate of genetic distances or phylogenetic trees, no matter which method is used (parsimony, distance or even maximum likelihood methods!), **is going to be meaningless** since gene sequences will tend to cluster according to the degree of similarity in their base (or amino acid) composition, irrespectively of their true genealogy. The problem of saturation is often overlooked in phylogeny reconstruction, when in fact it is rather crucial (see Chapter 20). For example, in coding regions, third codon positions, which usually evolve much faster than first and second (see above), are likely to be saturated especially when distantly related taxa are compared. One way to avoid this problem is to exclude third positions from the analysis, or to analyze the corresponding amino acid sequence (see Chapter 9).

The program DAMBE implements different methods to check for saturation in a data set of aligned nucleotide sequences (see Chapter 20). Here, a graphical exploration tool is introduced. The method takes advantage of the empirical observation that, in most data sets, transitional substitutions happen more frequently than transversional ones. Therefore, by plotting the observed number of transitions and transversions against the corrected genetic distance for the $n(n-1)/2$ pairwise comparison in an alignment of n taxa, transitions and transversions should both increase linearly with the genetic distance, with transitions being higher than transversions. However, as the genetic distance (the evolutionary time) increases, i.e. more divergent sequences are compared, saturation is reached and transversions will eventually outnumber transitions. This is because by chance alone there are eight possible transversions but only four transitions (see Fig. 4.6). In coding sequences, saturation will be more pronounced in the rapidly evolving third codon position.

Figure 4.11 (a) and (b) show the result for the Primates data set (Primates.fas file) analyzed using DAMBE. To analyze first and second codon positions (Fig. 4.11a) or third codon position (Fig. 4.11b) separately, select the item `Work on codon position 1 and 2` or `Work on codon position 3` from the `Sequences` menu before starting any analysis (the original sequences can be restored by choosing `Restore sequences` from the same menu). The transition and transversion vs. divergence plot can be obtained by selecting the item from the `Graphics` menu in DAMBE.

The plots show that both transitions and transversions grow approximately linear with the genetic distance indicating no saturation in the Primates data set. Figure 4.11c, on the other hand, is an example of substitution saturation at the third codon position in envelope gp120 HIV-1 sequences aligned with simian immunodeficiency virus (SIVcpz) isolated from Chimpanzees (the data set was previously used in Salemi *et al.*, 2001). Saturation becomes evident for sequence pairs with F84 distances greater than 0.70. Therefore, in spite of SIVcpz and HIV-1 belonging to the same phylogenetic lineage and sharing a common ancestor

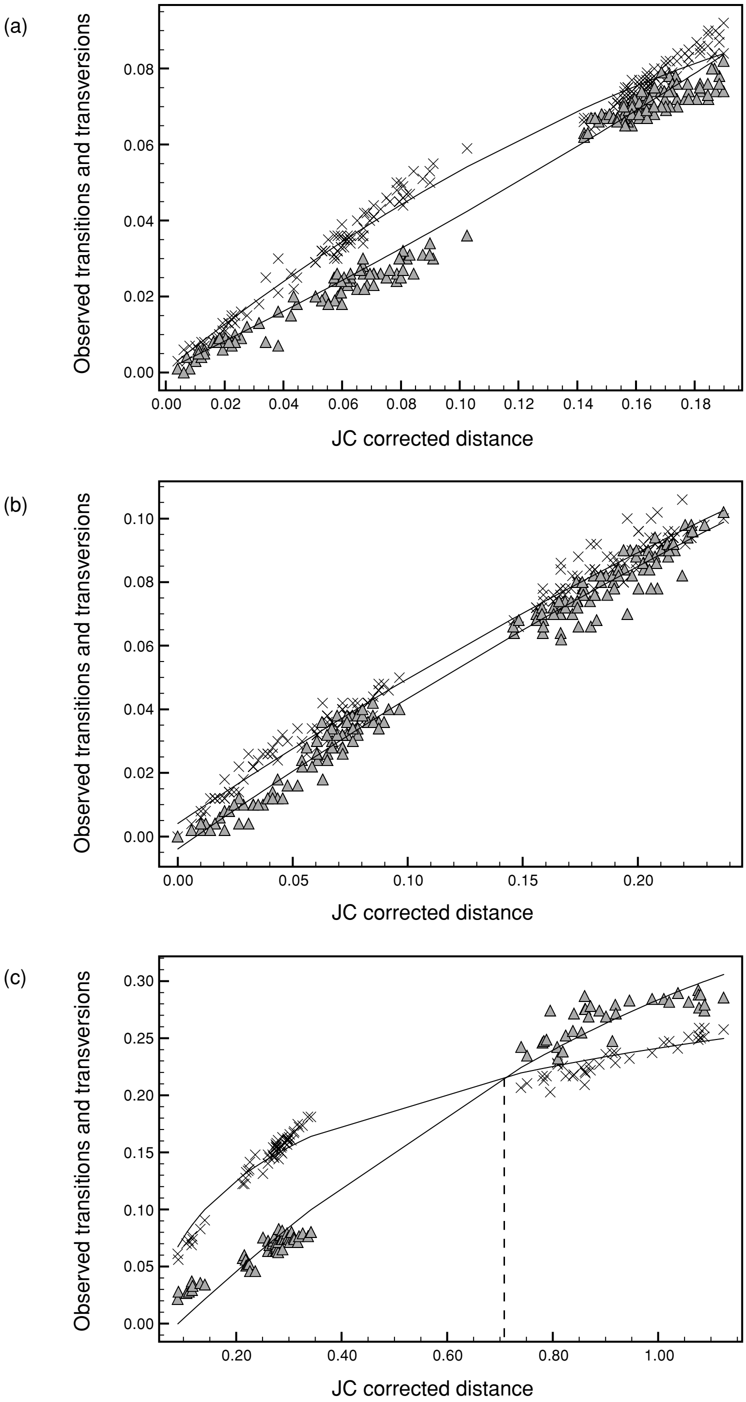

Fig. 4.11 Plotting the observed transitions and transversions against a corrected genetic distance using Dambe. (a) First and second codon position of the primate data set. (b) Third codon position of the primate data set. (c) Third codon position of an HIV/SIV envelope data set.

within the last 300 years (Salemi *et al.*, 2001) any phylogenetic inference based on the signal present at the third codon position has to be considered unreliable. Substitution saturation in this case is due to the extremely fast evolutionary rate of HIV-1, around 10^{-3} nucleotide substitutions per site per year. In addition to the graphical tool introduced here, DAMBE also implements different statistical tests to assess substitution saturation (see Chapter 20).

4.13 Choosing among different evolutionary models

After a few exercises, it should become clear that genetic distances inferred according to different evolutionary models can lead to rather different results. Tree-building algorithms such as ***UPGMA*** and ***Neighbor-Joining*** (see next chapter) are based on pairwise distances among taxa: unreliable estimates will lead to inaccurate branch lengths and, in some cases, to the wrong tree topology. When confronted with model choice, the most complex model with the largest number of parameters is not necessarily the most appropriate. A model with fewer parameters will produce estimates with smaller variances. Since we always analyze a finite data sample, our parameter estimates will be associated with sampling errors. Although for sequences of at least 1000 nucleotides, these may be reasonably small, it has been shown that models with more parameters still produce a larger error than simpler ones (Tajima & Nei, 1984; Gojobori *et al.*, 1992; Zharkikh, 1994). When a simple evolutionary model (for example, JC or F84) fits the data not significantly worse than a more complex model, the former should be preferred (see Chapter 10). Finally, complex models can be computationally daunting for analyzing large data sets, even using relatively fast computers.

Another basic assumption of *Time-homogeneous time-continuous stationary* ***Markov models***, the class of nucleotide substitution models discussed in the present chapter (Section 4.4), is that the base composition among the sequences being analyzed is at equilibrium, i.e. each sequence in the data set is supposed to have similar base composition, which does not change over time. Such an assumption usually holds when closely related species are compared, but it may be violated for very divergent *taxa* risking flawed estimates of genetic distances. TREE-PUZZLE implements by default a chi-square test comparing whether the nucleotide composition of each sequence is significantly different with respect to the frequency distribution assumed in the model selected for the analysis. The result is written in the outfile at the end of any computation. As an example, Fig. 4.12 shows the results of the chi-square test for a set of mtDNA sequences from different organisms. There are significant differences in base composition for 8 of the 17 taxa included in the data: a severe violation of the equal frequency assumption in the homogeneous Markov process. In this case, a more reliable estimate of the genetic distance can be

```
SEQUENCE COMPOSITION (SEQUENCES IN INPUT ORDER)

              5% chi-square test     p-value
 Lungfish Au         passed           6.20%
 Lungfish SA         failed           0.62%
 Lungfish Af         failed           1.60%
 Frog                passed          58.01%
 Turtle              passed          44.25%
 Sphenodon           passed          59.78%
 Lizard              passed          38.67%
 Crocodile           failed           2.51%
 Bird                failed           0.00%
 Human               failed           0.85%
 Seal                passed          68.93%
 Cow                 passed          59.11%
 Whale               passed          97.83%
 Mouse               failed           1.43%
 Rat                 passed          39.69%
 Platypus            failed           3.46%
 Opossum             failed           0.01%


The chi-square test compares the nucleotide composition of each
sequence to the frequency distribution assumed in the maximum
likelihood model.
```

Fig. 4.12 Comparing nucleotide composition using a chi-squared test for a mitochondrial DNA data set as outputted by TREE-PUZZLE.

obtained with the LogDet method, which has been developed to deal specifically with this kind of problem (Steel, 1994; Lockart *et al.*, 1994). The method estimates the distance *d* between two aligned sequences by calculating:

$$d = -\ln[\det F] \qquad (4.39)$$

Det F is the determinant of a 4 × 4 matrix where each entry represents the proportion of sites having any possible nucleotide pair within the two aligned sequences. A mathematical justification for (4.39) is beyond the scope of this book. An intuitive introduction to the LogDet method can be found in Pages and Holmes (1998), whereas a more detailed discussion is given by Swofford *et al.* (1996). LogDet distances can be calculated using the program PAUP* and their application to the mtDNA data set is discussed in the practical part of Chapter 8. Chapter 10 will focus on statistical tests for selecting the best evolutionary model for the data under investigation.