In: P. Lemey, M. Salemi and A.-M. Vandamme (eds.). To appear in: The Phylogenetic Handbook. 2nd Edition. Cambridge University Press, UK.

(final version 21. 9. 2006)

Chapter 4. Nucleotide Substitution Models

THEORY

Korbinian Strimmer(1) and Arndt von Haeseler(2)

(1) Department of Statistics, University of Munich, Germany.

(2) Center for Integrative Bioinformatics Vienna, Max F. Perutz Laborarories, Austria.

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 DNA 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 Figure 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, 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*** provide a basis to infer the evolutionary tree relating the sequences. In particular, if sequences actually evolved according to a tree and if the correct ***genetic distances*** between all pairs of sequences are available, then it is computationally simple 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 difference per site.

The ***p-distance*** is a very intuitive measure. Unfortunately, it suffers from a severe shortcoming: if the rate of substitution is high it is 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, e.g. 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 ***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. 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*** can not directly be observed, statistical techniques are necessary to infer this quantity from the data. For example, using the relationship between d and p given in Figure 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 emphasise 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:

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

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 $1\text{-}\mu\delta t$. 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 $P_n(t)\cdot(1\text{-}\mu\delta t)$. 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:

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

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

When δt tends to zero the left part of equations 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 $P_0(t)$ 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 equation 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 from $f(t)=h(t)g(t)$:

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

Comparing equation 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 $P_1(t)=(\mu \text{ t}) \exp(-\mu t)$. If we compute the derivative, we reproduce equation (4.8). Induction leads to equation

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

The formula above 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 equation 4.10.

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 figure 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 parameters 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 summarised by the **Q** matrix in figure 4.3 belong to a general class of models known as ***time-homogeneous time-continuous stationary Markov models***. When applied to modelling 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 modelling 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 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*, **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 equation 4.6). When the probabilities **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 **P**(*t*) and the expected ***genetic distance*** in case of the simple Jukes and Cantor model of evolution (Jukes and 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), implies that the equilibrium frequencies of the four nucleotide 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 Figure 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 $\mathbf{P}(t)=\exp(\mathbf{Q}\mathrm{t})$ (equation 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 split t time units ago

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

and solving for $\mu\mathrm{t}$ we get

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

Thus the right hand side gives the number of substitutions occuring 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 Figure 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 equation 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

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

(Kimura and 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 12 will discuss a statistical test that can help selecting 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 modelled 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, 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

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

and

$$P_{ij}(t)>0 \text{ for } t>0. \qquad (4.17)$$

Moreover, it also fulfills the requirement that

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

known as Chapman-Kolmogorov equation, and the initial condition

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

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

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

$$\pi_i \mathrm{P}_{ij}(t) = \pi_j \mathrm{P}_{ji}(t) \qquad (4.20)$$

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 **Q** is called rate matrix. It provides an infinitesimal description of the substitution process. In order not to violate equation 4.16 the rate matrix **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 equation 4.20 the rate matrix is the first derivative of $\mathbf{P}(t)$ which is constant for all t in a time-homogenous ***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 .

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

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

From the Chapman-Kolmogorov equation 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 (equations 4.19a,b) to give

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

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

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

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 Σ

π_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_ι 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 figure 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 generalisation of equation 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 and 7 for more details). The likelihood is the probability to observe the two sequences given the distance *d*. It is defined as

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

where $x_{A(s)}$ is the state at site *s*=1, ..., *l* in sequence A and $P_{x_{A(s)}x_{B(s)}}\left(\frac{d}{\mu}\right)$ is the transition probability. A value for *d* that maximises *L*(*d*) is called a ***maximum-likelihood estimate*** **(MLE)** of the genetic distance. To find this estimate numerical optimisation 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 (equation 4.30) and the likelihood estimator (equation 4.31) using standard statistical techniques. The so-called delta method can be employed to compute the variance of an estimate obtained from equation 4.30, and 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 8 free parameters of a reversible nucleotide rate matrix **Q** are specified the general time reversible model (GTR) is derived (see figure 4.5). However, 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 Figure 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) Ts_R , and pyrimidine transitions (C ↔T) Ts_Y. When these constraints are imposed only two independent rate parameters (out of 5) 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 and Nei, 1993) which can be written as

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

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

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

If γ=1 and therefore the purine and pyrimidine transitions have the same rate this model reduces to the HKY85 model (Hasegawa, Kishino, and Yano, 1985)

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

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

If the base frequencies are uniform (π_i=1/4) the HKY85 model further reduces to the Kimura 2-parameters (K80) model (Kimura, 1980). For κ= 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 parameterisation. A single parameter τ generates the $\kappa$and γparameters of the TN93 model (see equations 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 figure 4.7.

4.6.1 Rate heterogeneity over 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 3rd codon positions mutate usually faster than 1st positions which, in turn, mutate faster than 2nd 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 first a plausible model for distribution of rates over sites is required. The common approach is to use a Γ-distribution with expectation 1.0 and variance $1/\alpha$.

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

By adjusting the shape parameter α the Γ-distribution accommodates for varying degree of rate heterogeneity (see figure 4.8). For $\alpha>1$ the distribution is bell-shaped and models weak rate heterogeneity over sites. The relative rates drawn from this distribution are all close to 1.0. For $\alpha<1$ 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, ..., 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 Pond and Frost (2005) who propose a two-stage hierarchical Beta-Gamma model for fitting the rate distribution across sites.

PRACTICE

Marco Salemi

Rega Institute for Medical Research, Katholieke Universiteit Leuven, Belgium

4.7. Software Packages

A large number of software packages exist for computing ***genetic distances*** from DNA sequences. An exhaustive list is maintained by Joe Felsenstein at the web address http://evolution.genetics.washington.edu/phylip/software.html. Among others, the programs PAUP* (see Chapter 6), PHYLIP (Felsenstein, 1993), TREE-PUZZLE (Strimmer and von Haeseler, 1996), PAL (Drummond and Strimmer, 2001), MEGA (Kumar *et al.*, 1993), TREECON (Van de Peer, 1994), DAMBE (Xia, 2001), and PAML (Yang, 2000) provide the possibility to infer maximum-likelihood distances. In the remaining part of the chapter we are going to use PHYLIP, DAMBE (see previous Chapter) and TREE-PUZZLE. PAUP* and TREECON will be discussed in Chapter 6 and Chapter 9, respectively.

PHYLIP, Phylogeny Inference Package, 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. Already-compiled executables files are available for Windows3.x/95/98, pre-386 and 386 DOS, Macintosh (non-PowerMac), and PowerMac. A complete description of the package including the instructions for installation on different machines can be found at http://evolution.genetics.washington.edu/phylip.html.

TREE-PUZZLE (Strimmer and von Haeseler, 1995) was originally developed to reconstruct phylogenetic trees from molecular sequence data by maximum likelihood with a fast tree-search algorithm called quartet puzzling (see Chapter 7). The program also computes pairwise maximum likelihood distances according to a number of models of nucleotide substitution. Versions of TREE-PUZZLE UNIX, MacOS PPC, and Windows95/98/NT compatible can be freely downloaded from the TREE-PUZZLE web page at http://www.tree-puzzle.de/.

As soon as the proper versions of these programs are installed, the PHYLIP folder and the TREE-PUZZLE folder should be visible on the local computer. These folders contains several files, including executable applications, documentation, and source codes. PHYLIP version 3.5 has three subdirectories: `doc`, `exe`, `src`; the executables are in the `exe` folder. In TREE-PUZZLE version 5.0 the executable `treepuzzle.exe` can be found in the TREE-PUZZLE folder. Any of the software modules within PHYLIP and TREE-PUZZLE works in the same basic way: it needs a file containing the input data, for example aligned DNA sequences in PHYLIP format, to be placed in the same directory where the program resides; it produces one or more output files, in text format, with the results of some kind of analysis. 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.

4.8. Jukes and Cantor (JC69) ***genetic distances*** (PHYLIP)

Let us see how to calculate d for the HIV data set: the aligned sequences can be downloaded from http://kuleuven.ac.be/aidslab/phylogenyBook/datasets.htm. Figure 4.9 shows a matrix with pairwise ***p-distances***, i.e. number of different sites between two sequences divided by the sequence length, for the HIV/SIV data. The matrix is written in lower-triangular form with the number on the first row indicating the number of sequences being compared. The ***genetic distance*** *d* between sequence L20571 and AF103818 according to the JC69 model can be obtained by substituting their ***observed distance*** p=0.3937 in equation 4.19 (see section 4.4.1), and it results 0.5582. Obviously, using p instead of d would grossly underestimate the genetic divergence between the two sequences.

Place the file hivALN in the directory `PHYLIP\exe` if you are working with PHYLIP v3.5, or in the same directory where the PHLYLIP software module `DNAdist` is if you are using a different version of the package. 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 3.5c

Settings for this run:
  D  Distance (Kimura, Jin/Nei, ML, J-C)?  Kimura 2-parameter
  T        Transition/transversion ratio?  2.0
  C   One category of substitution rates?  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)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes

Are these settings correct? (type Y or 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 Transition/transversion ratio?` is no longer present since under the JC69 model all nucleotide substitutions are equally likely (see section 4.4.1). Type `y` followed by the `enter` key to carry out the computation of the ***genetic distances***. The result is stored in a file called `outfile` which can be opened with any text editor (see Figure 4.10). 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 in Figure 4.10 can be obtained by replacing p in equation 4.19 with the ***observed distance*** given in Figure 4.9.

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

The Kimura 2-parameters (K80) model relaxes one of the main assumptions of the JC69 model allowing for a different instantaneous substitution rates between transitions and transversions (a=c=d=f=1 and b=e=κ in the **Q** matrix) (Kimura, 1980). Similarly to what has been done in section 4.4, by solving the exponential **P**(t)=exp(**Q**t) for **P**(t) we can obtain the K80 correction formula for the expected ***genetic distance*** between two DNA sequences:

$$d = log(1/(1\text{-}2P\text{-}Q)) + \log(1/(1\text{-}2Q)) \qquad (4.37a)$$

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 :

$$\mathrm{V}(d) = 1/n[(\mathrm{A}^2P + \mathrm{B}^2Q - (\mathrm{A}P + \mathrm{B}Q)^2] \qquad (4.37b)$$

with A=1/(1-2P-Q) and B=1/2[(1/1-2P-Q) + (1/1-2Q)]

K80-distances can be obtained with `DNAdist` by choosing `Kimura 2-parameter` within the `D` option. The user can also type an empirical Ts/Tv ratio by selecting option `T` from the main menu. The default value for Ti/Tv in `DNAdist` is 2.0. Considering that there are twice more possible transversions than transitions, a Ti/Tv=2.0 equals to assume that during evolution transitional changes occur four times faster 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 the next Chapter.

The ***genetic distance*** estimated with the K80 model (Ti/Tv=2.0) between the HIV group O strain L20571 and the SIV chimpanzee strain AF103818 results 0.6346, which is 1.6 times bigger than the ***p-distance***, but only 1.1 times bigger than the JC69 distance.

The K80 model still relies on very restricted assumptions such that of equal frequency of the four bases at equilibrium. The HKY85 (Hishino, Kasegawa and Yano, 1985) and F84 (Felsenstein, 1984; Kishino and 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 estimated from the data set). F84-distances can be computed with PHYLIP by selecting `Maximum Likelihood` within the `D` option. A new option, `F`, also appears in the main menu:

```
F   Use empirical base frequencies?   Yes
```

By default, `DNAdist` empirically estimates the frequencies for each sequence 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.

HKY85, as well as distances according to other models described in this chapter, can be estimated by TREE-PUZZLE (see following chapter).

4.10. More complex models

The TN93 (Tamura and Nei, 1993) model can be considered as a further extension of the F84 model, allowing different nucleotide substitution rates for purine (A↔G) and pyrimidine (C↔T) transition ($b \neq e$ 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 user can input from the menu empirical transition/transversion bias and pyrimidine/purine transition bias, otherwise the program will estimate those parameters from the data set (see next chapter and chapter 7).

4.10.1. Modelling rate heterogeneity over sites

The JC69 model assumes that each site in a sequence changes over time at uniform rate. 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 across different sites. In section 4.6.1 we pointed out that this assumption is not realistic and it is especially violated in coding regions where different codon positions usually evolve at different rate. Replacement at 2nd codon position are always ***nonsynonymous***, i.e. they change the encoded amino acid, (see chapter 11), whereas, because of the degeneracy of the genetic code, 65% of the possible replacements at 3rd codon position are ***synonymous***, i.e. no change in the encoded amino acid. Finally only 4% of the possible replacements at 1st codon position are ***synonymous***. Since mutations in a protein sequence are most of the time likely to reduce the ability of that protein to perform its biological function, they are rapidly removed from the population by purifying selection (see chapter 1 and 11). As a consequence, over time, mutations will be accumulated more rapidly at 3rd rather than at 2nd or 1st 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, 3rd codon positions mutate about eight times faster than 1st and sixteen times faster than 2nd positions (Salemi *et al.*, 2000). To model rate heterogeneity over sites `DNAdist` the user can select the option

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

in the main menu and choose up to nine different categories of substitution rate. The program then asks to input the relative substitution rate for each category as a nonnegative real number. Let say that we want to estimate the ***genetic distances*** for the hivALN data set with the JC69 model, but assuming that mutations at 3rd position accumulates 10 times faster than at 1st and 20 times faster than at 2nd codon position. Since what matters are the relative rates, one possibility is to set the rate at 1st codon position equal to 1, the rate at 2nd to 0.5, and the rate at 3rd to 10. It is also necessary to assign each site in the aligned data set to one of the three rate categories. This can be easily done by adding to the infile, after the initial line indicating the number of sequences and the number of nucleotides, one or more lines like the following:.

```
CATEGORIES 12312312311231231231[...]
```

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 forth position to rate category 1 again, and so forth. In the hivALN data set sequences are in frame, starting at 1st codon position and ending at a 3rd codon position, and there are 2352 positions. Thus we need to edit the input file in the following way:

```
14 2352
CATEGORIES 123123123 [and so forth up to 2352 digits]
L20571    ATGACAG [...]
[...]
```

A file already edited can be found at: http://kuleuven.ac.be/aidslab/phylogenyBook/datasets.htm:

i. Place the input file 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 we have no clue about the distribution and the extent of the relative substitution rates across sites, we can alternatively model rate heterogeneity using a Γ-distribution, as discussed in section 4.6.1, with a single parameter α describing the shape of the distribution: 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, it is usually not known. In the next chapter we will discuss how to estimate α

with TREE-PUZZLE and how to estimate genetic distances in case of Γ-distributed rates across sites.

It should look clear after a few exercises that ***genetic distances*** inferred according to different evolutionary models can lead to quite different results. Tree-building algorithms such as UPGMA and Neighbour-Joining (see next chapter) are based on pairwise distances among taxa: unreliable estimates could lead to the wrong tree topology and, certainly, to wrong branch lengths. It may look that the more complex the model, the more free parameters it allows for, the more accurate the inferred distances should be. However, this is not necessarily true. A model with less parameters will have a smaller variance (see for example equations 4.19b and 4.48b). Moreover, any evolutionary model share the underlying assumption that the number of sites compared between two given sequences is infinite, the violation of the assumption leading to sampling errors. Although for sequences at least 1000 nucleotides long the approximation usually holds well, it has been shown that models with more parameters produce a larger error than simpler ones (Gojobori *et al.*, 1992; Tajima and Nei, 1984; Zharkikh, 1994). In chapter 12 we will discuss a general strategy to select the best fitting evolutionary model for a given data set.

References

Baake, E. (1998). What can and what cannot be inferred from pairwise sequence comparison? *Mathematical Biosciences.* 154, 1-21.

Bailey, N.T.J. (1964). *The elements of Stochastic Processes with Application to the Natural Sciences*. New York, Wiley.

Drummond, A. & Strimmer K. (2001). PAL: An object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics* 17, 662-663.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17, 368-376.

Felsenstein, J. (1984). Distance methods for inferring phylogenies: a justification. *Evolution* 38, 16-24.

Felsenstein, J. (1993). *PHYLIP. Phylogenetic Inference Package, version 3.5c*. Seattle: Department of Genetics, University of Washington.

Hasegawa, M., Kishino, H. & Yano T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 21, 160-174.

Jukes, T. &Cantor, C.R. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism*, ed. Munro, H.N., pp. 21-132. Academic Press, New York.

Kimura, M. & Ohta T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution* 2, 87-90.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 15, 111-120.

Kishino, H. & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the Hominoidea. *Journal of Molecular Evolution* 29, 170-179.

Kosakovsky Pond, S. L. & Frost, S. D. (2005). A simple hierarchical approach to modeling distributions of substitution rate. *Molecular Biology and Evolution* 22, 223-234.

Lanave, C., Preparata, G., Saccone, C. & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 20, 86-93.

Lindgren, B.W. (1976). *Statistical Theory*. Macmillan, 3rd ed., New York.

Rodriguez, F., Oliver, J.F., Marin, A. & Medina, J.R. (1990). The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* 142, 485-501.

Strimmer, K. & Von Haeseler, A. (1996). Quartet-puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13, 964-969.

Swofford, D.L. (1998). *PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods). Version 4*. Sunderland (MA): Sinauer Associates.

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

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures in Mathematics and Life Sciences* 17, 57-86.

Thorne, J.L., Kishino, H. & Felsenstein J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution* 34, 3-16.

Uzzel, T. & Corbin, K.W. (1971). Fitting discrete probability distributions to evolutionary events. *Sciences* 172, 1089-1096.

Wakeley, J. (1993). Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *Journal of Molecular Evolution* 37, 613-623.

Yang, Z. (1994a). Estimating the pattern o nucleotide substitution. *Journal of Molecular Evolution* 39, 105-111.

Yang, Z. (1994b). Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39, 306-314.

Yang, Z. (2000). *Phylogenetic Analysis by Maximum Likelihood (PAML), version 3.0.* University College, London (UK).

FIGURE LEGENDS

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

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

Figure 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**). μ 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.

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

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

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

Figure 4.7. Hierarchy of nucleotide substitution models.

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

Figure 4.9. Pairwise ***p-distance*** matrix for the HIV/SIV example data set.

Figure 4.10. Pairwise ***genetic distances*** for the HIV/SIV example data set according to the Jukes and Cantor (JC69) model.