STOCHASTIC MODELS IN BIOLOGICAL SCIENCES
BANACH CENTER PUBLICATIONS, VOLUME 80
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSZAWA 2008

LINKING POPULATION GENETICS TO PHYLOGENETICS

PAUL G. HIGGS

Department of Physics and Astronomy, McMaster University
Hamilton, Ontario L8S 4M1, Canada
E-mail: higgsp@mcmaster.ca

Abstract. Population geneticists study the variability of gene sequences within a species, whereas phylogeneticists compare gene sequences between species and usually have only one representative sequence per species. Stochastic models in population genetics are used to determine probability distributions for gene frequencies and to predict the probability that a new mutation will become fixed in a population. Stochastic models in phylogenetics describe the substitution process in the single sequence that represents a species. These latter models are of great practical use in inferring evolutionary trees, but they ignore population genetics. Here, we consider the way that the two fields are linked. In a finite population with a low mutation rate, most sites in a gene are monomorphic; therefore it is a reasonable approximation to represent a species by a single sequence and to ignore within population variability. We show that population genetics theory can be used to predict the substitution rates that will be seen in phylogenetic models. However, a substitution in the phylogenetic sense is really the fixation of a new mutation in the population genetics sense. Substitution rates are dependent on mutation, selection and random drift, not just on mutation. An appreciation of the relationship with population genetics is important when designing new empirical models for phylogenetics, for example, paired-site models for RNA sequences with secondary structure and codon-based models for protein-coding regions of DNA.

1. Introduction. This paper arises from the school on “Stochastic Models in Biological Sciences” held at the Banach Centre in 2006. It is intended as a pedagogical introduction to the population genetics models of allele frequencies and the phylogenetics models of substitution rates. The aim is to show that these two types of models can be brought together, and that an understanding of population genetics is useful when developing new models in phylogenetics.

2000 *Mathematics Subject Classification*: 92D15.

Key words and phrases: population genetics, phylogenetics.

This research was supported by the Canada Research Chairs organization and by the Natural Sciences and Engineering Research Council of Canada.

The paper is in final form and no version of it will be published elsewhere.

Population genetics is the study of the variability of genes within populations. This is a well-established discipline with a strong theoretical foundation dating back to the first half of the twentieth century. A classic problem in population genetics is the calculation of the frequency distribution of alleles at a genetic locus. (A locus is a site on a chromosome where a particular type of gene is found and an allele is one of a number of alternative versions of the gene that can be present at that locus.) By taking a sample of individuals from a population, it is possible to estimate the frequency of each of the alternative alleles in the population. The frequency of an allele will vary in time due to mutation, selection and random drift. Stochastic models are used to study the way allele frequencies change, and to predict the distribution of frequencies that is to be expected in different circumstances.

Molecular phylogenetics is the use of nucleic acid or protein sequence data from different species to construct evolutionary trees. Typically, a single sequence from each species is used. The variability within each species is neglected, because differences in gene sequences between individuals in a species are assumed to be smaller than differences between species. Stochastic models are also central to molecular phylogenetics. Phylogenetic models of DNA sequence evolution assume that substitutions of one nucleotide by another occur as random events at points in a gene sequence. The sequences of different species diverge over time due to the accumulation of many of these random substitutions independently in different lineages. The models are defined in terms of a substitution rate matrix in which the rates of different types of substitution depend on a number of parameters whose values can be estimated by fitting the model to sequence data.

Population genetics was founded before the era of molecular biology. It is possible to discuss the segregation of alleles for round and wrinkled peas or brown and blue eyes in a meaningful way without even knowing that these genes are encoded by DNA. The advent of experimental techniques for detecting alternative alleles, such as allozymes and restriction fragment length polymorphisms, led to much more detailed molecular input to population genetics, but these techniques still did not give the full sequence. More recently, it has become possible to sequence DNA from many individuals in a population, and to detect the specific sites in a gene where variability exists (known as single nucleotide polymorphisms - SNPs). This type of information is useful for medical studies, and in some cases, diseases can be linked to specific mutations. Because of the medical relevance, intraspecific sequence data is now becoming more common for humans. In most other species, we are still limited to only one sequence of any given gene, and we must presume that this is a typical representative of the species. Nevertheless, variability at the sequence level must exist in other species, as it does in humans.

The models of population genetics are defined from a mathematic point of view and are as close as one comes in biology to first-principles theory. On the other hand, the models of phylogenetics are empirical and have been built up to describe phenomena seen in real sequence data. The important results regarding the probabilities of fixation of new alleles as a function of mutation, selection and drift have been understood for a long time. However, until recently, these things have largely been ignored in practical phylogenetic studies. Substitution rate parameters in phylogenetic models are usually estimated merely with the goal of obtaining the phylogeny, and little attempt has been made to interpret the substitution rates in terms of population genetics. Currently, with the increasing availability of both inter- and intra-specific sequence data, the two fields are beginning to touch on one another more closely than was formerly the case. Phylogenetic models are becoming increasingly more complex, and this allows a closer link to population genetics in the way the models are defined. Section 2 of this paper gives a brief introduction to the ideas of gene frequency distributions and fixation probabilities that are key aspects of population genetics. In Sections 3, 4 and 5, simple models of DNA sequence evolution from phylogenetics are introduced, and simple simulations are presented as illustrations of the way the two types of model are related. Sections 6 and 7 discuss more recent models that deal with sequences coding for RNAs and proteins. We consider the way that population genetics issues, like selection, fixation, and compensatory mutations, are relevant to evolution of these molecules and the way that these effects can be built into more realistic models of sequence evolution for use in molecular phylogenetics.

2. Stationary gene frequency distributions. We will consider a standard model in population genetics for the gene frequency distribution at a locus with four alleles. This applies to a single site in a DNA sequence that can be either A, C, G or T, under the assumption that the site is neutral (i.e. there is no difference in fitness between the nucleotides). Let $n_i(t)$ be the number of gene copies in the population at generation t that have base i at that site. We will follow the convention that the total number of copies of the gene is $2N$. This applies to a diploid splecies where N is the population size, or to a haploid species with population $2N$. The frequencies of the bases are $x_i(t) = n_i(t)/2N$. The Wright-Fisher model deals with constant population size and non-overlapping generations. Each gene copy at time $t+1$ is descended from one randomly chosen copy at time t. Let there be a probability u that there is a mutation from any base to any other base in one generation. The probability, a_i, that a new gene copy has i at this site is the probability that its parent was i and there was no mutation plus the probability that its parent was something other than i and there was a mutation to i.

$$a_i = x_i(t)(1-3u) + \sum_{j \neq i} x_j(t)u \tag{1}$$

Each gene copy is sampled independently by copying from the previous generation. Therefore, the probability distribution for the numbers of copies of each allele at the next generation (time $t+1$) is a multinomial distribution:

$$p(n_A, n_C, n_G, n_T) = \frac{(2N)!}{n_A! n_C! n_G! n_T!} a_A^{n_A} a_C^{n_C} a_G^{n_G} a_T^{n_T}. \tag{2}$$

For clarity, we have omitted the $t+1$ in all the $n_i(t+1)$ in equation 2. Using equations 1 and 2 repeatedly allows us to simulate a population history over many generations. If we keep track of the value of $n_i(t)$ at each time for any one of the four bases, we can calculate the probability, $\Phi(n)$, that $n_i(t) = n$. As $2N \gg 1$, this can be viewed as a continuous probability distribution $\phi(x)$, where $x = n/2N$. After a long time, $\phi(x)$ converges to the stationary distribution in equation 3, which can be calculated using diffusion theory (see

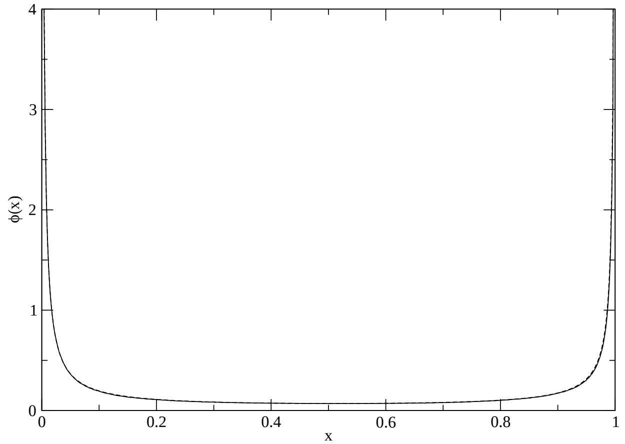

Fig. 1. Stationary frequency distribution for the case $\theta = 0.025$. The solid line is calculated from equation 3. The result from simulation of the the model for 10^8 generations with $2N = 500$ and $u = 2.5 \times 10^{-5}$ is superimposed as a dashed line, and is almost indistinguishable.

Wright [1], chapter 13, Crow and Kimura [2], chapter 9, or Durrett [3], chapter 1).

$$\phi(x) = \frac{\Gamma(4\theta)}{\Gamma(\theta)\Gamma(3\theta)} x^{\theta-1}(1-x)^{3\theta-1} \tag{3}$$

The parameter $\theta = 4Nu$ controls the shape of this distribution. The mutation rate, u, is very small in real cases (often quoted as of order 10^{-9} per site per generation), whereas the effective population size might be only around 10^4. Usually, therefore, $\theta \ll 1$, and both x and $(1-x)$ are raised to negative powers in equation 3. Hence, $\phi(x)$ has a U-shaped distribution that diverges as x tends to 1 or 0. The gamma functions in equation 3 are just constants that ensure the probability distribution is normalized. Figure 1 shows an example of the shape of the distribution for $\theta = 0.025$.

When $\theta \ll 1$, most of the weight of the probability distribution is close to 0 or to 1. As the four bases are equivalent in this model, there is a probability of 1/4 that any base will have frequency close to 1, and a probability of 3/4 that it will have frequency close to 0. A site is described as monomorphic if almost all the population has the same base at this site. For concreteness, we will say it is monomorphic if the most common base has a frequency greater than 95%. The site is polymorphic if its most common base has frequency less than 95%. If θ is small, most sites will be monomorphic most of the time, but they will occasionally pass through periods of polymorphism as they flip from being monomorphic for one base to being monomorphic for some other base.

The process of flipping from one state to another is called fixation. Consider a site that is completely monomorphic - let's say 100% A. If a mutation to a C occurs, there is just one gene copy that has C at this site. The frequency of C in the population will vary in time due to random drift (i.e. the multinomial sampling described by equation 2). The C allele has two possible fates. With a very small probability, p_{fix}, it will spread through the population until x_C is close to 1. In this case, we say that the C mutation has become fixed in the population. However, with a probability $1 - p_{fix}$, random drift leads to the disappearance of C, and the population remains dominated by the A allele.

The probability of fixation of a neutral mutation, given that it begins with just one copy out of $2N$, can be shown to be $p_{fix} = 1/2N$ (Crow and Kimura [2] chapter 8). This can be calculated from diffusion theory, but it can also be obtained with the following simple argument. If we trace the lines of any two gene copies in the current population backwards in time, there comes a point where they coalesce, i.e. the two gene copies have the same ancestor. If we trace back the lines of descent of all genes in the current generation, they eventually coalesce to the a single gene copy that is the last common ancestor (LCA) of the whole of the current population. Now suppose a mutation occurred in the LCA generation. If it occurred in the gene copy that is the LCA itself, then it becomes fixed in the population at the current generation. If it occurs in any of the $2N - 1$ gene copies that is not the LCA, then it dies out before the current generation. Hence $p_{fix} = 1/2N$.

We can now calculate the probability that the population flips between the four alternative monomorphic states. For the population in the A state, the mean number of C mutations per generation in the whole population is $2Nu$. Each of these has a probability $1/2N$ of being fixed. Thus, the rate of flipping from state A to state C (or between any other two bases) is $r = 2Nu \times 1/2N = u$. This is a fundamental property of neutral evolution: the rate of evolution of the population is equal to the rate at which mutations occur in a single individual. It should be borne in mind that this is only true for a neutral mutation. For an advantageous mutation, $r > u$, and for a deleterious mutation, $r < u$.

3. Can the population be represented by a single sequence? We now begin to see the potential link between the population genetics and phylogenetics viewpoint. In phylogenetics, it is assumed that a population can be represented by a single gene sequence and that variation within the population is not important. If most sites in a gene are monomorphic, then this assumption seems reasonable at first sight. A single DNA site can be in states i = A, C, G or T. A phylogenetic model describing the evolution of the site is defined by a rate matrix r_{ij} in which each off-diagonal element is the rate of substitution from state i to state j. The diagonal elements are defined as $r_{ii} = -\sum_{j \neq i} r_{ij}$. It is necessary to insist on the distinction between substitutions and mutations. In a population genetics model, a mutation is a change in a single gene copy (due to replication error or chemical damage) that can be passed on to the next generation. In a phylogenetics model, a substitution is a change in the sequence that is being used to represent the population. A substitution arises as the result of the fixation of a mutation. Substitution rates are influenced by random drift and selection, as well as mutation.

The fundamental quantities to be calculated in a phylogenetic model are the transition probabilities $P_{ij}(t)$, defined as the probability that the site will be in state j after time t given that it was initially in state i. These probabilities satisfy the rate equation:

$$\frac{dP_{ij}}{dt} = \sum_{k} P_{ik} r_{kj}. \tag{4}$$

For simple choices of the r_{ij} matrix, this equation can be solved analytically. For more general cases, it can be solved numerically.

The simplest model of this type is the Jukes-Cantor (JC) model [4], in which $r_{ij} = u$, for all $i \neq j$ and $r_{ii} = -3u$. In this section, u is a substitution rate, i.e. it is the rate at which the sequence representing the population changes. In the previous section, u was a mutation rate. However, we saw that fixation and mutation rates are equal for neutral mutations; therefore, it should be possible to link the two models directly and have u be the same in both models. We now show that this *is* possible, but it is necessary to be careful.

If we simulate the evolution of a population evolving according to the Wright-Fisher model (as in section 2), we can use it to generate a time series, such as GGGGGTTTAAAAATTTTCCCCCCAA..., which lists the state the population is in at each generation. We will consider three different ways of choosing the base to represent the state of the population below, but the application of the JC model to the time series is the same in each case. The JC model is a Markov model where it is assumed that the probability of being in a given state at any point in time depends only on the state at the previous generation. The transition probabilities in a fixed period of 1 generation are $P_{ij}(1)$, obtained from solution of equation 4. The time series can be summarized by the matrix of observed numbers of substitutions S_{ij}, which is simply the count of the number of times that state j follows state i in the time series. It is straightforward to calculate $P_{ij}(t)$ by solving equation 4 for the JC model (detailed solutions are given by Li [5], Durrett [3] and Higgs and Attwood [6]). The probability that the base at time t is the same as the initial base is $P_{ii}(t)$, and the probability that it is different from the initial base, irrespective of the which two bases these are, is $D(t)$, where:

$$D(t) = 1 - P_{ii}(t) = \frac{3}{4}(1 - \exp(-4ut)). \tag{5}$$

The total number of steps in the time series is $S_{tot} = \sum_i \sum_j S_{ij}$. The number of times that i and j are different is $S_{dif} = \sum_i \sum_{j \neq i} S_{ij}$. Therefore, the observed probability that the states are different at two successive steps is $\hat{D} = S_{dif}/S_{tot}$. The 'hat' denotes the value estimated from the data. By substituting $\hat{D}$ into equation 5 and inverting it, we obtain an estimate of ut.

$$\widehat{ut} = -\frac{1}{4} \ln(1 - 4\hat{D}/3). \tag{6}$$

It is not possible to estimate u and t separately from the observed transition matrix because transition probabilities are dimensionless, and they always depend on products of rates and times.

We simulated the Wright-Fisher model for 10^8 generations with parameters as in Figure 1 and used this to generate time series, from which we estimated ut by the method above. The most obvious way to generate the time series is to select one random individual from the population at each generation and to use the nucleotide from this individual to represent the population. This corresponds to the usual case in phylogenetics, where only one sequence from a species is available. A second method is to use the consensus state of the population, i.e. the nucleotide that is most frequent in the population at

Table 1. Alternative methods of estimating the substitution rate u from simulation data

method	interval t	S_{tot}	S_{dif}	$\hat{D}$	ut	u
random individual	1	10^8	6803246	0.0680	2.38×10^{-2}	2.38×10^{-2}
consensus state	1	10^8	254002	2.54×10^{-3}	8.48×10^{-4}	8.48×10^{-4}
counting fixations	1	10^8	7384	7.38×10^{-5}	2.46×10^{-5}	2.46×10^{-5}
random individual	10^3	10^5	13311	0.1331	0.0488	4.88×10^{-5}
consensus state	10^3	10^5	9416	0.0942	0.0335	3.35×10^{-5}
counting fixations	10^3	10^5	7154	0.0715	0.0251	2.51×10^{-5}
random individual	10^4	10^4	4952	0.4952	0.270	2.70×10^{-5}
consensus state	10^4	10^4	4772	0.4772	0.253	2.53×10^{-5}
counting fixations	10^4	10^4	4705	0.4705	0.247	2.47×10^{-5}

each generation. A third method is by counting the fixations, which can be done in the following way. If a population is monomorphic for base i (i.e. $x_i > 0.95$), we use i as the base in the time series. We then continue to use i in the time series until such point as the population becomes monomorphic for another base j. In this method, the population is still considered to be in state i in the period during which it is polymorphic, and the transition to state j is only deemed to have occurred when the population becomes monomorphic for j.

These three methods generate different time series from the same simulation. The estimates of ut from the three time series are shown in the top three lines of Table 1. We know in the simulation that the samples were taken at an interval of $t = 1$ generation. Thus $ut = u$ in the table. The true mutation rate is $u = 2.5 \times 10^{-5}$. The result obtained from counting the fixations is consistent with this, as we expected. However, the time series from the consensus sequence overestimates u by a factor of over thirty, and the random individual time series overestimates by a factor of a thousand. It is clear that the first two methods are inconsistent.

To see why these estimates are wrong, consider Figure 2(a), which shows the genealogical tree generated by tracing the ancestors of each gene copy back through time until the LCA. Consider, initially, two randomly chosen gene copies in the *same* generation. The probability that they have the same parent in the previous generation is $1/2N$, and the probability that they have different parents is $1 - 1/2N$. Hence, the probability that their common ancestor existed T generations in the past is

$$F(T) = \left(1 - \frac{1}{2N}\right)^{T-1} \frac{1}{2N} \approx \frac{1}{2N} \exp(-T/2N), \tag{7}$$

where the approximation assumes that $N \gg 1$. From this, the mean time back to the common ancestor is $2N$ generations. Now consider two randomly chosen gene copies at two successive generations, as in the time series method. These are the black dots in Figure 2(a). These are usually not direct ancestors of one another. The length of the chain connecting them is $2T + 1$ generations, and the 1 is negligible compared to the $2T$.

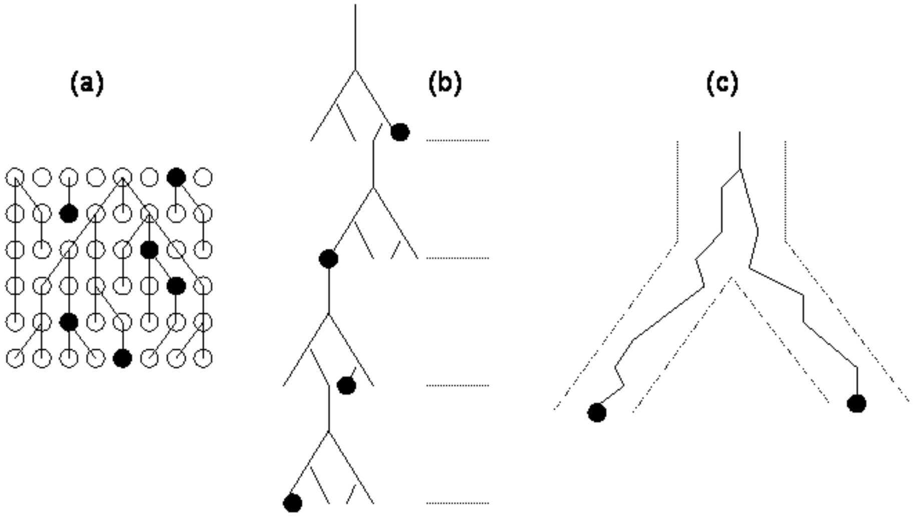

Fig. 2. (a) The lines of descent of gene copies in a population over a few generations. The black dots indicate a randomly sampled individual at each generation. (b) The relationship between gene copies sampled at large intervals from a population. (c) Illustration of a gene tree within a species tree.

For this reason, the probability that they are different is much larger than if they were direct ancestors. By combining equations 5 and 7, the probability that these two gene copies are different is

$$\int_0^{\infty} D(2T)F(T)dT = \frac{3}{4}\frac{16Nu}{1+16Nu}. \tag{8}$$

In our case, $16Nu = 0.1$, and this probability is 0.068, which is exactly what we obtained for the observed $\hat{D}$ in line 1 of Table 1. If this value of $\hat{D}$ is used inappropriately in equation 6, an incorrect estimate of u is obtained.

The estimate of u from the consensus sequence is also very inaccurate. This is because, when a neutral mutation spreads through a population, its frequency does not necessarily increase in a steady sweep. The frequency of the new base may wander up and down past the 50% mark several times during the transition period before it eventially goes to fixation. Therefore, the number of changes of the consensus base is more than the number of fixations. If the consensus base changed in the recent past, it is probable that the population is polymorphic, and it is quite likely that the consensus will change back to the old consensus base in the near future. On the other hand, if the consensus base did not change for a long time, the population is probably monomorphic and the consensus base is much less likely to change in the near future. Thus, the probability of change of the consensus base depends on past history, so the consensus base changes are not well described by a Markov model.

The situation is much better, however, if we only consider samples taken at broadly spaced intervals of time, rather than at every generation. Table 1 shows what happens if we sample at intervals of $t = 10^3$ and 10^4 generations from the same simulation as before. S_{tot} is now the simulation length divided by the interval. $\hat{D}$ is larger than before because the time between each sample is larger; hence the estimates of ut are also larger. However, after dividing by t, the estimates of u are better than before. When $t = 10^3$, the estimate from the random invidiual is only out by a factor of 2, and when $t = 10^4$, it is out by less than 10%. The estimate from counting fixations is consistent, whatever the sampling interval. The estimate from the consensus sequence is between that of the other two methods.

The typical time to coalescence of all the lines of descent in a population of $2N$ gene copies can be shown to be $4N$. The interval of 10^3 in Table 1 is equal to $4N$, and the interval is 10^4 is much larger than this. When the interval is large, the relationship between the sampled sequences looks like Figure 2(b). The genealogies of the populations at each of the sampling times are now separate from one another. The LCA of each population is descended from a random individual in the previous population. Only the branches of the tree that link to individuals at the times of sampling are drawn in the figure. The population is constant in size: the figure does not imply that the population passes through repeated bottlenecks. If the sampling interval is much larger than the coalescence time, most of the changes between two successive samples occur on the line separating the populations rather than within the genealogy of each population. For this reason the different estimates of the substitution rate all converge.

Let us recap: we asked whether it is valid to represent the whole population by a single sequence and treat model it as though it were a Markov substitution process, such as the JC model. If we use either the random individual method or the consensus sequence method to generate the time series, the straightforward answer is ‘no’. It is not possible to tell the probability of the nucleotide state at time $t + 1$ simply from the state at time t. Further information about the nucleotide frequencies in the population and the genealogical structure of the population is required. Therefore the assumptions of the Markov model do not hold. However, the fixation process is fairly well described by a Markov process when the allele frequency distributions are U-shaped. Hence, the method of counting fixations gives a very good estimate. When dealing with real data, we do not have a complete history of the fixation process through time; nevertheless, if we sample the population at intervals wider than the coalescence time of the population, using either a random individual or the consensus sequence, the process is quite closely described by a Markov model and the estimate of the substitution rate is very close to the true rate of fixations. Therefore, it seems that approximating the population by a single sequence might not be such a bad idea after all, which is comforting, because phylogeneticists have been doing this all along!

4. The simplest tree. Of course, the sequences used in phylogenetics do not consist of samples taken at broad intervals of time from a single lineage. Instead they consist of samples from separate species at the current time. Furthermore, we have many sites in a sequence, not just one. However, if a Markov substitution model applies for a single lineage, it will also apply for species on a tree. We will show this using the simplest possible tree: just two species.

We begin with a population of $2N = 500$, each of which has a sequence of length 2000. Evolution follows the neutral Wright-Fisher model. Each sequence is copied from the sequence of a randomly chosen parent sequence at the previous generation. Mutations occur independently at each site when the sequence is copied. After 10000 generations, a speciation event occurs, as in Figure 2c. The population is split randomly into two halves of size $2N = 250$, and these then evolve independently of one another for a period of $T = 20000$ generations. On the first generation after the split, the two populations are allowed to immediately double back to $2N = 500$, after which they remain constant in size. At the end of the simulation, one random individual is chosen from each of the two populations and the matrix of observed substitutions is calculated by comparing these two sequences.

In this example, we will use a more complex model for the mutation rates due to Tamura and Nei (TN) [7], which is defined by the matrix

$$\mathbf{r} = \begin{array}{c} \\ A \\ G \\ T \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & G & T & C \end{array} \\ \left(\begin{array}{cccc} * & \kappa_1 u\pi_G & u\pi_T & u\pi_C \\ \kappa_1 u\pi_A & * & u\pi_T & u\pi_C \\ u\pi_A & u\pi_G & * & \kappa_2 u\pi_C \\ u\pi_A & u\pi_G & \kappa_2 u\pi_T & * \end{array} \right) \end{array} . \tag{9}$$

In the TN model, π_i is the frequency of base i. The four frequencies may be all different, as long as they sum to 1. The parameter u controls the rate of mutations between purines and pyrimidines (these are known as transversions). Transitions between purines (A and G) occur a factor κ_1 times faster than transversions, while transitions between pyrimidines (T and C) occur a factor κ_2 times faster. The $*$ denotes the fact that the matrix elements on the diagonal must be equal to minus the sum of the other elements on the row. The TN model is one of a class of phylogenetic models that are time reversible, i.e. they satisfy the rule $\pi_i r_{ij} = \pi_j r_{ji}$ for all i and j. For more examples of such models, see Higgs and Attwood [6].

The TN model is usually used as a substitution rate model in phylogenetics, but here we are using it to describe the mutation process. The probability that state i in the parent mutates to state j in the offspring is $r_{ij}\delta t$, where δt is a small time step, in this case, one generation. The probability that no mutation occurs and i remains unchanged is 1 - $\sum_{j \neq i} r_{ij}\delta t$. We used the following parameters in this example: $\pi_A = \pi_T = 0.2$, $\pi_C = \pi_G = 0.3$, $\kappa_1 = 3$, $\kappa_2 = 2$, and $u = 2 \times 10^{-5}$. The observed number of sites with i in species 1 and j in species 2 is shown below.

$$\begin{array}{c} \\ A \\ G \\ T \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & G & T & C \end{array} \\ \left(\begin{array}{cccc} 147 & 118 & 40 & 58 \\ 126 & 307 & 69 & 93 \\ 42 & 68 & 168 & 89 \\ 77 & 101 & 111 & 386 \end{array} \right) \end{array} . \tag{10}$$

We will now use the TN model in its usual way as a substitution model for fitting the data. The object is to show that the parameters estimated from the data are consistent with the known values used in the simulation. An analytical solution of the TN model is possible [7], by first solving equation 4 to obtain the substituion probabilities $P_{ij}(t)$, as a function of the rate parameters in equation 9, and then inverting these equations to give estimates of the rate parameters. Parameter estimates depend on the observed numbers of substitutions of three types: $\hat{D}_1$, the fraction of sites that differ by a purine transition; $\hat{D}_2$, the fraction that differ by a pyrimidine transition; and $\hat{D}_3$, the fraction that differ by a transversion. In section 3, we used $\hat{D}$ to denote the fraction of times that two successive states in the time series were different. Here, we are using $\hat{D}$ to denote the fraction of homologous sites in the two species that are different. These two quantities are analogous. It is necessary to separate the observed substitutions into three types because they depend on the rate parameters in different ways. After some algebra, the parameter estimates are:

$$\widehat{\kappa_1 ut} = -\frac{1}{2\pi_R}\ln\left(1 - \frac{\pi_R \hat{D}_1}{2\pi_A\pi_G} - \frac{\hat{D}_3}{2\pi_R}\right), \tag{11}$$

$$\widehat{\kappa_2 ut} = -\frac{1}{2\pi_Y}\ln\left(1 - \frac{\pi_Y \hat{D}_2}{2\pi_T\pi_C} - \frac{\hat{D}_3}{2\pi_Y}\right), \tag{12}$$

$$\widehat{ut} = -\frac{1}{2}\left(1 - \frac{\pi_A\pi_G}{\pi_R^2} - \frac{\pi_C\pi_T}{\pi_Y^2}\right)\ln\left(1 - \frac{\hat{D}_3}{2\pi_R\pi_Y}\right). \tag{13}$$

In the above equations, $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_T + \pi_C$. Table 2 shows the estimates calculated from the results in matrix 10. In a real phylogenetic problem we do not know rates and times separately, so we must be satisfied with the relative rates in column 3 of Table 2. However, in the simulation, we know that the time since the speciation is 20000 generations. The time separating the species is, therefore, approximately twice this. However, the point of coalescence of the two individuals from which the sequences were taken is slightly before the speciation, as shown in Figure 2c. We kept track of the times of common ancestors in this simulation, and it happens that the point of coalescence is 97 generations before the speciation. Therefore, the time separating the sequences is $2 \times 20097 = 40194$. Using this value of t, we obtain the absolute values of the rates in the final column of the table. These are close to the real values given above, within statistical error.

Table 2. Estimates of rate parameters in a simulation of the TN model

Type of substitution	Observed frequency	Relative rate	Absolute rate
Transitions between purines	$\hat{D}_1 = 244/2000 = 0.122$	$\kappa_1 ut = 2.58$	$\kappa_1 u = 6.4 \times 10^{-5}$
Transitions between pyrimidines	$\hat{D}_2 = 200/2000 = 0.100$	$\kappa_2 ut = 1.50$	$\kappa_2 u = 3.7 \times 10^{-5}$
Transversions	$\hat{D}_3 = 548/2000 = 0.274$	$ut = 0.795$	$u = 2.0 \times 10^{-5}$

In this case, sampling a random member of each population leads to a consistent estimation of the rate matrix. In hindsight, this should not surprise us. This is a simulation of *neutral evoluton* where each gene copy has the same fitness. The lineages of the two sequences we used for data fitting are two typical lineages in the population. Mutations occur on these lineages independently of any mutations occurring in other members of the population. The matrix we used for mutations is the TN matrix; therefore it is not surprising that the TN model fits the data correctly. In fact, the population makes no difference in this simulation, and we could have gotten the same answer by simulating just two sequences instead of two populations. This simulation therefore demonstrates what we knew in advance from neutral evolution theory. The other point illustrated by this simulation is that the coalescence point of the two genes is slightly before the point where the species split. In cases where two speciation events occur shortly after one another, this can lead to an effect known as lineage sorting in which the branching order of the genes is not the same as the branching order of the species (Li [5], chapter 5). However, this is not a serious issue in most phylogenetic studies as long as the time between speciation events is large compared to N generations.

5. Adding natural selection. So far, things look good for the phylogeneticists: it seems to be consistent to treat the population as though it were a single sequence. However, all the models considered up to now are neutral. If selection is included, the fixation rate is no longer the same as the mutation rate. In any model with selection, the ancestors of the surviving individuals are those that were fitter than average at the time they existed. The substitutions in the lineages leading to the sampled individuals are weighted by selection, and they are not equivalent to the underlying mutations.

Suppose that the mutation process follows the JC model, so that the four bases are equivalent under mutation, but let the bases have different fitnesses, w_i. The multinomial sampling formula still applies (equation 2), but the probabilities of the four states are now dependent on the fitnesses of each state and on the mean fitness $\overline{w}$.

$$a_i = x_i(t)(1-3u)w_i/\overline{w} + \sum_{j \neq i} x_j(t)uw_j/\overline{w}. \tag{14}$$

As a specific example, consider the case where the A base is under positive selection: $w_A = 1 + s$, for some small positive selection coefficient s, and $w_i = 1$ for the other three bases. There are now three types of substitutions in this model. Substitutions between any two bases other than A should be neutral, and these should occur at rate $r = u$ as in section 2. Substitutions to A from any other base are advantageous, and these should occur at rate $r = \lambda_1 u$, where λ_1 is $2N\times$ the probability of fixation of an advantageous mutation. Substitutions from A to any other base are deleterious, and these should occur at rate $r = \lambda_2 u$, where λ_2 is $2N\times$ the probability of fixation of a deleterious mutation. These fixation probabilities are known from diffusion theory (Crow and Kimura [2], Durrett [3]):

$$\lambda_1 = \frac{2N(1-e^{-2s})}{(1-e^{-4Ns})}, \tag{15}$$

$$\lambda_2 = \frac{2N(e^{2s}-1)}{(e^{4Ns}-1)}. \tag{16}$$

From consideration of the fixation process, our expected substitution rate matrix looks like this:

$$
\mathbf{r} = \begin{array}{c} \\ A \\ G \\ T \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & G & T & C \end{array} \\ \left(\begin{array}{cccc} * & \lambda_2 u & \lambda_2 u & \lambda_2 u \\ \lambda_1 u & * & u & u \\ \lambda_1 u & u & * & u \\ \lambda_1 u & u & u & * \end{array} \right) \end{array} . \tag{17}
$$

Recall that all the mutation rates are equal to u, and that the λ factors arise from the population genetics, not directly from the mutations. This is another time reversible substitution rate matrix for which the relevant quantities can be calculated using similar methods as before. The stationary frequencies of the four bases under this model are $\pi_A = \lambda_1/(\lambda_1 + 3\lambda_2)$ and $\pi_i = \lambda_2/(\lambda_1 + 3\lambda_2)$ for $i \neq$ A. Due to selection, π_A is larger than the others. If this model is fitted to sequence data, the ratio λ_1/λ_2 can be obtained from the observed value of π_A:

$$
\lambda_1/\lambda_2 = 3\pi_A/(1-\pi_A). \tag{18}
$$

It is necessary to distinguish two types of substitutions. $\hat{D}_1$ is the fraction of substitutions in which one is A and one is anything other than A, and $\hat{D}_2$ is the fraction of substitutions in which the two bases are different but neither is A. Rate parameters can then be estimated:

$$
\widehat{\lambda_1 ut} = -\pi_A \ln\left(1 - \frac{\hat{D}_1}{2\pi_A(1-\pi_A)}\right), \tag{19}
$$

$$
\widehat{ut} = -\frac{1}{3}\ln\left(1 - \frac{3\hat{D}_2 + \hat{D}_1}{2(1-\pi_A)}\right) - \frac{\widehat{\lambda_1 ut}}{3}. \tag{20}
$$

We simulated a population evolving according to this population genetics model with parameters $2N = 500, u = 2.5 \times 10^{-5}$, and $s = 1 \times 10^{-3}$, and constructed time series for the base state as in section 2. For these parameters, $\lambda_1 = 1.58, \lambda_2 = 0.58, \lambda_1/\lambda_2 = 2.71$ and the expected frequency of A is $\pi_A = 0.475$. The sampling interval was 10^4 generations; hence $ut = 0.25$ and $\lambda_1 ut = 0.395$. The results are shown in Table 3. As the sampling interval is large, the random individual and consensus state methods are close to the result from counting fixations, and all three methods give a good estimate of the substitution process. Once again, as long as we are interested in intervals of time longer than the coalescence time, it is reasonable to approximate the population by a single sequence and to fit the data with a Markov model of substitutions. However, this example makes it clear that the substitution model describes the fixation process, not the mutation process.

Table 3. Estimating the substitution rate parameters from simulation data for the single site model with selection

method	π_A	$\hat{D}_1$	$\hat{D}_2$	λ_1/λ_2	$\lambda_1 ut$	ut
random individual	0.472	0.298	0.154	2.68	0.429	0.280
consensus state	0.477	0.288	0.149	2.73	0.411	0.266
counting fixations	0.482	0.281	0.143	2.80	0.398	0.251

To make this model for selection at a single site more practical for phylogenetics, it would be necessary to generalize it to deal with a whole sequence. There is no reason that the preferred base at every site should be A, so it would be necessary to allow different sites to have different preferred bases. Also, there is no reason why the same strength of selection should apply at every site, so it would be preferrable to include some kind of distribution of selection strengths across sites. These complications go beyond what is usually done in phylogenetics, where one substitution matrix is typically used to fit the average properties of a heterogenous bunch of sites. We will not pursue this here. Instead, we will consider two cases that are relevant for practical phylogenetic studies and that illustrate the importance of population genetics in the substitution process - pairs of sites in the stem regions of RNAs and triplets of sites in codon-based models for coding regions of DNA.

6. RNA pairs and codon triplets. RNA molecules such as ribosomal RNA and transfer RNA have secondary structures that are strongly conserved over evolutionary time. Each site in a stem region of an RNA structure is paired with another site on the opposite side of the stem. These pairs of sites evolve under the constraints of base-pair complementarity. In the simplest population genetics model for a pair of interacting sites, we suppose that the four combinations that form a Watson-Crick pair (AU, UA, GC, CG) are all equally fit ($w = 1$), and that all the other combinations are less fit ($w = 1 - s$) because the secondary structure is destabilized. In real RNA sequence alignments we frequently observe pairs of species that differ by pairs of mutations in the stem regions: for example, an AU pair in one species and a GC pair in another. This requires two mutations at separate sites. As mutation rates are very low, it is unlikely that both these mutations occurred simultaneously in the same gene copy. There must have been genes in the population where the sites did not match. In this example, the mismatch state would be either GU or AC. When the second mutation occurred in the mismatched sequence, the GC matching pair would be created. These pairs of mutations are called compensatory mutations, because the second mutation compensates for the deleterious effect of the first.

If the population is initially monomorphic for the AU pair, the rate of fixation of either GU or AC mismatch states is $\lambda_2 u$ (equation 16). If the mismatch state is fixed, the rate of fixation of the compensatory mutation GC (or of return to AU) should be $\lambda_1 u$ (equation 15). From this argument, the net rate of the complementary change would be slow, because it would be dominated by the slow rate of fixation of the initial deleterious mutation. However, this arguments fails to capture the nature of the complementary substitution process. If $Ns \gg 1$, and $u/s \ll 1$, it is very unlikely that a mismatch state will be fixed in the population, but there is nevertheless a continual rate of deleterious mutations to mismatch states all the time. For a large population, the frequency of a mismatch state in a population that is dominated by a matched state is u/s, determined by mutation-selection balance. Mismatch sequences will thus be present in the population at a low level, although we will typically not see them if we only sample one individual from the population. The rate of creation of GC sequences in an AU population is $2N \times 2u/s \times u$, where the second 2 comes from the fact that there are both GU and AC mismatches. Although the GC sequence is advantageous with respect to the mismatch, it is neutral with respect to the majority of the population, which is still AU. The probability of fixation of the neutral GC sequence is therefore $1/2N$. Thus, the net rate of the compensatory substitution is $2u^2/s$, which is much larger than $\lambda_2 u$ in the parameter range that we are considering. A more detailed calculation of the rate of compensatory subtitutions is given by Stephan [8]. The allele frequency distribution, analagous to Figure 1 can also be calculated for this situation (Higgs [9]).

Analysis of real RNA sequence alignments shows that pairs of compensatory subsitutions are frequent, and are often more frequent than single substitutions to mismatch states (Higgs [10]). The observation that double substitutions are more frequent than single substitutions is counter-intuitive initially, but the above argument shows how it can be explained in terms of population genetics.

An important aspect of real RNA structures is that GU and UG pairs occur quite frequently, and they are usually treated as weakly matching pairs rather than as mismatches. GU and UG pairs would therefore have a smaller selective disadvantage s than other mismatch pairs. As a result, changes that are double substitutions (AU $\leftrightarrow$ GC or UA $\leftrightarrow$ CG), which can go via GU or UG intermediates, occur more frequently than double transversions (such as AU $\leftrightarrow$ UA), which can only go via true mismatches. Also the severity of a mismatch mutation depends on the context of the structure in which it occurs. Free energies of helix formation depend on stacking interactions between neighbouring pairs, not just on individual pairs. A mismatch occurring in a very strongly bonded helix might be almost neutral, because the helix would still form easily when the mismatch is present, but the equivalent mismatch occurring in a weakly bonded helix might make the helix unstable, and therefore have a much more negative effect. At sites where the mismatch is nearly neutral, it would be relatively easy for the single mismatch to be fixed. We do in fact see a substantial number of sites with mismatches (especially GU and UG), as well as sites that appear to evolve almost exclusively via compensatory pairs.

Ribosomal RNA is one of the most frequently used genes in phylogenetic studies. Therefore it is of interest to develop realistic substitution rate models for the evolution of paired sites in RNA. When fitting substitution rate models to sequence data, we will be averaging over many sites with different selection regimes. A good model must therefore allow for all the types of substitutions that might occur - in particular, it must allow both single substitutions and compensatory double substitutions as possibilities. The relative rates of these processes can be estimated by fitting to real data. Among the first to use a model of this type were Tillier and Collins [11], although a variety of previous models had been proposed that only allowed single substitutions. Savill et al. [12] compared a large number of models for paired sites and used statistical tests to determine which gave better fits to sequence data. It was found that models that allowed double substitutions always fitted the data better than those that did not, and that the estimated rate of double substitutions was relatively large.

We have since gone on to implement RNA-specific models in a software package known as PHASE, that uses the Bayesian Markov Chain Monte Carlo technique for phylogenetic inference (Jow et al. [13], Hudelot et al. [14]). In phylogenetics, we are interested in determining the relative likelihood of alternative trees, given the sequence data. Standard single-site models assume that sites evolve independently and the full likelihood is a product of likelihoods of sites. RNA-pair models account for the strong correlation between the two sites in each pair. If this is ignored, relative likelihoods of alternative trees can be seriously in error. We therefore recommend using a base pair model for phylogenetics with RNA sequences, and we also recommend using a rate model that allows double substitutions, because this is an essential part of the way compensatory substitutions are fixed in populations. As we have emphasized in this article, the substitution rate model in phylogenetics needs to describe the fixation process, not the mutation process.

We now turn to models for amino acid sequences. A 20×20 substitution rate matrix can be used to describe the rate of substitution of any amino acid to any other. The earliest of these is the PAM model (Dayhoff et al. [15]), and more recent versions include those of Jones et al. [16], Adachi and Hasegawa [17], Müller and Vingron [18], and Whelan and Goldman [19]. These papers describe several different methods by which rate parameters can be estimated by fitting sequence data. As many parameters are being estimated, a very large set of sequences is required (alignments of many genes, each containing many sequences).

These models are called empirical models, because the rates are estimated from sequence data without any prior theory of which rates should be high or low. Nevertheless, the resulting matrices show some important trends that can be understood from first principles. The amino acid sequences evolve because mutations occur in the DNA sequences that code for them. Amino acid substituions are therefore influenced by the genetic code (i.e. the mapping from codons to amino acids). Amino acid changes that can be achieved by a DNA mutation at a single codon position occur more frequently than those that require two or three DNA mutations. Substitution rates are also influenced by selection at the protein level. Natural selection is, in general, conservative. Changes between amino acids that have very different physical properties tend to be disruptive of protein structure and function; therefore, they have a sizeable selective disadvantage and they tend not to be fixed in the population. Changes between amino acids with very similar properties are much more likely to be nearly neutral with respect to protein function; therefore they are more likely to be fixed in the population. A simple demonstration of this is given by Higgs and Attwood (Figure 4.6 of [6]) using the matrix of Jones et al. [16]. The set of amino acid pairs for which substitutions occur most frequently is a subset of the amino acid substitutions that can be achieved by a single mutation, i.e. all the rapid substitutions can be achieved by a single mutation, but not all the single mutation changes are rapid. Once again, this illustrates that the substitution process is describing fixation of new variants in the population, and is dependent on selection, not just mutation.

When a DNA sequence is translated to an amino acid sequence, information is lost, because more than one codon codes for each amino acid. Therefore, it is desireable to develop substitution rate models for coding sequences that work directly at the DNA level. Although 4-state models can describe single sites in DNA, in order to describe the effects of the genetic code and selection on proteins, we need 64-state models that treat triplets of sites in a codon as a single unit. Codon-based models are often used (Muse and Gaut [20], Yang and Nielsen [21], Yang et al. [22], Pond and Muse [23]) in which each codon can mutate to any other codon that differs at a single position. Differences in base frequences and differences between transitions and transversions can be accounted for in a similar way to the TN model (equation 9). An additional important parameter in these models is ω, the ratio of non-synonymous to synonymous substitution rates. Variants of these models allow ω to differ between genes (lower ω indicates stronger stabilizing selection), or between sites (high ω indicates unusual sites where beneficial mutations have occurred).

In these models, all non-synonymous substitutions are equivalent. The model of Goldman and Yang [24] is more interesting, in that it allows non-synonymous substitution rates to depend on the difference in the physical properties of the amino acids coded by the two codons. A physical property distance matrix due to Grantham [25] is used, and it is assumed that the substitution rate decreases exponentially with this distance. This builds in the fact that conservative amino acid substitutions occur more frequently than disruptive ones, as was observed with the empirical matrices. A similar idea is used by Higgs et al. [26] using a different distance matrix. In this case, the distance between amino acids is a weighted sum of the differences in several different physical properties. The weights are estimated from the data. This gives an indication of which physical properties have the strongest influence on substitution rates.

A further feature of the model used by Higgs et al. [26] is that it allows changes between codons that differ at two or three positions, not just those that differ at a single position (as in all the codon models mentioned previously). Parameters are included in the model that control the relative rates of double and triple changes to single changes. It is found that double and triple changes occur less frequently than single changes, but they do occur. Statistical tests show that if the rates of double and triple changes are fixed to be zero, the fit to the data is much worse than if these rates are allowed to be non-zero. We interpret this as an indication that double and triple changes are being fixed in the population at the same time, even if they did not occur simultaneously as mutations. The situation seems to be similar to that occurring in RNA pairs. There is one case in codon-based models where the analogy with compensatory substitutions in RNA is quite good. The two blocks of serine codons, UCN and AGY, are not accessible to one another by a single mutation. It is necessary to pass through an intermediate codon for cysteine or threonine. At a site where serine is the favoured amino acid, the two codon blocks can exchange via compensatory substitutions. However, in most double or triple changes, the final amino acid is different from the initial one, and the intermediate amino acid is different from either of these. The compensatory substitution model then seems less applicable, because it is not clear that the intermediate state is less fit than the initial and final states. Nevertheless, double and triple substitutions seem to be important when fitting the data. Similar conclusions have been reached by Whelan and Goldman [27] and Kosiol et al. [28]. The case of double substitutions between the two blocks of serine codons was also studied by Averof et al. [29], who argue that these occur as a result of mutational events that affect two neighbouring bases simultaneously, rather than because of compensatory substitutions. It should also be remembered that in the empirical amino acid substitution models discussed above, all amino acid substitutions are allowed. There is no restriction that substitution rates must be zero if the amino acids are not interchangeable via a single mutation. Thus, there seems to be no need to make this restriction in codon-based models either. The bottom line in phylogenetic studies is that we want to infer which is the best tree. We are most likely to get a reliable answer for the tree if we use a model of evolution that gives a good fit to the sequence data.

7. Towards more realistic models. Maximum likelihood methods in phylogenetics require the calculation of the likelihood that a known set of sequences evolved on a proposed tree. The maximum likelihood priciple is to select the tree for which this likelihood is the largest. The likelihood calculation is facilitated by the assumption that sites evolve independently; hence the likelihood for the whole sequence is the product of the likelihood of the sites. This assumption is made for convenience and practicality: it is clearly not true. The sites in RNA pairs or in codon triplets are obviously not independent, but it is relatively easy to incorporate this by using a model with a larger number of states that treats a pair or a triplet as a single unit, as in the previous section. In reality, correlations between different parts of a sequence occur in a way that cannot be parcelled up into neat pairs and triplets. For example, in RNA, stacking interactions between neighbouring pairs are important for the stability of the helix [10][30], so each pair is influenced by its neighbours, which are influenced by their neighbours, and so on. Similar correlations exist in protein structures too, and not just between neighbours along the chain. Residues in a protein can be in close proximity spatially even if they are not close along the backbone of the sequence.

In general, the fitness of a molecule depends on the whole molecule. Sites in a molecule are linked, so selection acting at one site can influence the fate of mutations occurring at other sites. Recent work has begun to take this into account. In both RNA and protein sequence evolution, the structure of the molecule tends to be more conserved than the sequence. The structure acts as a constraint on the evolution of the sequence. In order for a sequence to fold to a given structure, that structure must be thermodynamically favourable in comparison to other structures that the sequence could form. Computational methods are available that predict the energy of a sequence when folded to a specified structure. In the RNA case, thermodynamic parameters for formation of various kinds of base pairs and loop structures are known from experimental measurement (Freier et al. [30]). These are incorporated into routines that calculate the energy of a sequence when folded to a specified secondary structure. In the protein case, separation of secondary and tertiary structure is not so simple. An amino acid interacts with its close neighbours within the tertiary structure. There is no convenient, experimentally measured energy scale for these interactions, but pseudo-energies can be calculated statistically by analyzing the contacts that occur in many known examples of protein structures (Jones [31]). The more frequent the observed contacts between two amino acids (relative to the expectation in random sequences), the lower should be the pseudo-energy for interaction of these amino acids. It is then possible to calculate the pseudo-energy of any sequence when threaded onto the known structure of a real protein.

Given either a real energy or a pseudo energy function, the evolution of sequences on a fixed structure can be studied. This has been done for proteins (Robinson et al. [32], Rodrigue et al. [33]) and RNAs (Yu and Thorne [34], Thorne et al. [35]). It is assumed that fitness depends on energy: mutations that lower the energy are treated as advantageous and those that increase the energy are treated as deleterious. Substitution rates are calculated in these models as functions of population genetics parameters N, s and u, as in equations 15 and 16. This leads to the intriguing possibility that by fitting data in the form of single sequences per species, as in phylogenetics, it is possible to estimate parameters that are relevant to population genetics, i.e. one can do population genetics without data on within-species variation [35].

In all these studies, it is assumed that a sequence can mutate into any sequence that differs from it by one mutation. For a nucleic acid sequence of length L, there are 4^L possible sequences and $3L$ other sequences to which each sequence could mutate in a single step. Multiple substitutions could, in principle, be incorporated into this framework, but have not yet been included, for reasons of computational complexity. It can be seen that this type of model is of a different level of complexity to those considered in all the previous sections, because the number of states increases exponentially with the sequence length, whereas in the previous cases it is fixed independently of L (e.g. 4 for a single site model or 64 for a codon model). In the likelihood calculations for standard models, the quantity calculated is a sum over the likelihoods of all possible histories of sequence evolution that could have existed in the interior portions of a tree. This requires solution of equation 4, which involves determining the eigenvalues of the r_{ij} matrix. For the models discussed in this section, an alternative method is used in which possible histories of substitution along the tree are sampled using a Markov Chain Monte Carlo method. This allows model parameters to be estimated in a Bayesian framework.

8. Conclusions. Here we will summarize the central points considered in this article. Population genetics and phylogenetics are both fields in which stochastic models have been developed to explain biological data. The models used in the two fields have largely been independent of one another because they have focused on different kinds of data: within-species variation in the case of population genetics and between-species variation in the case of phylogenetics. As the volume of useful sequence data for phylogenetics has increased over the past decades, the potential for meaningful estimation of large numbers of parameters from the data has also increased. This has driven the development of ever-more-complex models in phylogenetics, supported by the observation that more complex models can often be shown to fit real data better than simpler ones.

Phylogenetic models have now reached the level of complexity where it is desirable to think about population genetics parameters in the model specification. However, this raises the main issue with which we began this paper. Phylogenetic models consider a substitution process as a Markov process going on in a single sequence. This ignores variability among members of a population, which is the whole raison d'etre of population genetics. Is this a valid thing to do? We concluded that, in many cases, the answer is 'yes'. If mutation rates are very small and population sizes are not too large, as is very often the case, most sites in a gene sequence will be monomorphic, and differences between species will usually be larger than differences between members of the same species. A randomly sampled member of the population will then be reasonably representative of the species as a whole. We have shown that Markov substituion rate models can be used to fit data consisting of randomly sampled sequences taken from the populations of separate species, or from the lineage of one species at broad intervals of time. The substitution rates estimated in this way correspond to the rates of fixation of new sequence variants in the population. In the case of neutral evolution, the fixation rate is equal to the mutation rate, so fitting the substitution model gives a direct estimate of the mutation rates. However, in any model applied to real data, this will not be the case. The substitution matrix will give an estimate of fixation rates, which depend on selection and random drift, in addition to mutation. A clear example of the qualitative difference between mutation and fixation rates is given by compensatory mutations, which are known to be frequent in RNA and which also appear to be occurring in protein coding sequences.

When specifying the functional form of the substitution models to be used in phylogenetics, it is necessary to allow for the types of complexity that we expect to see in the fixation rates. This complexity is a consequence of the fact that there really is a population genetics process generating the sequence data that we analyze in phylogenetics.

References

[1] S. Wright, *The Theory of Gene Frequencies*, Vol. II, University of Chicago Press, Chicago, 1969.

[2] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory*, Harper and Row, New York, 1970.

[3] R. Durrett, *Probability models for DNA sequence evolution*, Springer, New York, 2002.

[4] T. H. Jukes and C. R. Cantor, *Evolution of protein molecules*, in: Mammalian Protein Metabolism, H. N. Munro (ed.), Academic Press, New York, 1969, 21–123.

[5] W. H. Li, *Molecular Evolution*, Sinauer Associates, Sunderland Massachusetts, 1997.

[6] P. G. Higgs and T. K. Attwood, *Bioinformatics and Molecular Evolution*, Blackwell Publishing, Malden, Massachusetts, 2005.

[7] K. Tamura and M. Nei, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees*, Mol. Biol. Evol. 10 (1993), 512–526.

[8] W. Stephan, *The rate of compensatory evolution*, Genetics 144 (1996), 419–426.

[9] P. G. Higgs, *Compensatory neutral mutations and the evolution of RNA*, Genetica 102/103 (1998), 91–101.

[10] P. G. Higgs, *RNA secondary structure: Physical and computational aspects*, Quarterly Reviews in Biophysics 33 (2000), 199–253.

[11] E. R. M. Tillier and R. A. Collins, *Neighbor joining and maximum likelihood with RNA sequences: Addressing the interdependence of sites*, Mol. Biol. Evol. 12 (1995), 7–15.

[12] N. J. Savill, D. C. Hoyle and P. G. Higgs, *RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum likelihood methods*, Genetics 157 (2001), 399–411.

[13] H. Jow, C. Hudelot, M. Rattray and P. G. Higgs, *Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution*, Mol. Biol. Evol. 19 (2002), 1591–1601.

[14] C. Hudelot, V. Gowri-Shankar, H. Jow, M. Rattray and P. G. Higgs, *RNA-based phylogenetic methods: Application to mammalian mitochondrial RNA sequences*, Mol. Phyl. Evol. 28 (2003), 241–252.

[15] M. O. Dayhoff, R. M. Schwartz and B. C. Orcutt, *A model of evolutionary change in proteins*, Atlas of Protein Sequence and Structure 5 (1978), 343–352. National Biomedical Research Foundation, Washington, DC.

[16] D. T. Jones, W. R. Taylor and J. M. Thornton, *The rapid generation of mutation data matrices from protein sequences*, CABIOS 8 (1992), 272–282.

[17] J. Adachi and M. Hasegawa, *A model of amino acid substitution in proteins encoded by mitochondrial DNA*, J. Mol. Evol. 42 (1996), 459–468.

[18] T. Müller and M. Vingron, *Modelling amino acid replacement*, J. Comp. Biol. 7 (2000), 761–776.

[19] S. Whelan and N. Goldman, *A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach*, Mol. Biol. Evol. 18 (2001), 691–699.

[20] S. V. Muse and B. S. Gaut, *A likelihood approach for comparing synonymous and non-synonymous nucleotide substitution rates, with application to the chloroplast genome*, Mol. Biol. Evol. 11 (1994), 715–724.

[21] Z. Yang and R. Nielsen, *Synonymous and non-synonymous rate variation in nuclear genes of mammals*, J. Mol. Evol. 146 (1998), 409–418.

[22] Z. Yang, R. Nielsen, N. Goldman and A. M. Krabbe Pedersen, *Codon-substitution models for heterogeneous selection pressure at amino acid sites*, Genetics 155 (2000), 431–449.

[23] S. K. Pond and S. V. Muse, *Site to site variation of synonymous substitution rates*, Mol. Biol. Evol. 22 (2005), 2375–2385.

[24] N. Goldman and Z. Yang, *A codon-based model of nucleotide substitutions for protein-coding DNA sequences*, Mol. Biol. Evol. 11 (1994), 725–736.

[25] R. Grantham, *Amino acid difference formula to help explain protein evolution*, Science 185 (1974), 862–864.

[26] P. G. Higgs, W. Hao and G. B. Golding, *Identification of conflicting selective effects on highly expressed genes*, Evolutionary Bioinformatics 2 (2006), 1–13.

[27] S. Whelan and N. Goldman, *Estimating the frequency of events that cause multiple-nucleotide changes*, Genetics 167 (2004), 2027–2043.

[28] C. Kosiol, I. Holmes and N. Goldman, *An empirical codon model for protein sequence evolution*, Mol. Biol. Evol. 24 (2007), 1464–1479.

[29] M. Averof, A. Rokas, K. H. Wolfe and P. M. Sharp, *Evidence for a high frequency of simultaneous double-nucleotide substitutions*, Science 287 (2000), 1283–1286.

[30] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Nielson and D. H. Turner, *Improved free energy parameters for prediction of RNA duplex stability*, Proc. Nat. Acad. Sci. USA 83 (1986), 9373–9377.

[31] D.T. Jones, *GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences*, J. Mol. Biol. 287 (1999), 797–815.

[32] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman and J. L. Thorne, *Protein evolution with dependence among codons due to tertiary structure*, Mol. Biol. Evol. 20 (2003), 1692–1704.

[33] N. Rodrigue, N. Lartillot, D. Bryant and H. Philippe, *Site interdependence attributed to tertiary structure in amino acid sequence evolution*, Gene 347 (2005), 207–217.

[34] J. Yu and J. L. Thorne, *Dependence among sites in RNA evolution*, Mol. Biol. Evol. 23 (2006), 1527–1537.

[35] J. L. Thorne, S. C. Choi, J. Yu, P. G. Higgs and H. Kishino, *Population genetics without intraspecific data*, Mol. Biol. Evol. (in press).