Lecture Notes: Markov models of sequence evolution

Dannie Durand

Markov models of sequence evolution

Now that we have the Markov chain machinery under our belts, let's return to the question of modeling sequence evolution. The process of substitution at a single site in a nucleotide sequence can be modeled as a Markov chain, where each state represents a single nucleotide and the transition probability, P_{jk}, is the probability that nucleotide j will be replaced by nucleotide k in one time stop. Similarly, Markov chains can be constructed to model the evolution of amino acid sequences. Although in principle Markov models of sequence evolution are general and can be applied to nucleotide sequences and to amino acid sequences in exactly the same way, in practice working with a twenty-letter alphabet poses challenges that do not arise with a four-letter alphabet. In addition, the biophysical properties of the amino acids are more varied than those of the nucleotides. For these reasons, the Markov chain framework is applied somewhat differently in amino acid sequence models. For the moment, we will focus on nucleotide models and postpone amino acid models until later in the course.

Markov models of sequence substitution are used to answer a wide range of questions that arise in molecular evolution, including correcting for multiple substitutions at the same site, simulating sequence evolution, estimating rates of evolution, deriving substitution scoring matrices, and estimating the likelihood of observing a pair of aligned nucleotides, given a phylogenetic model.

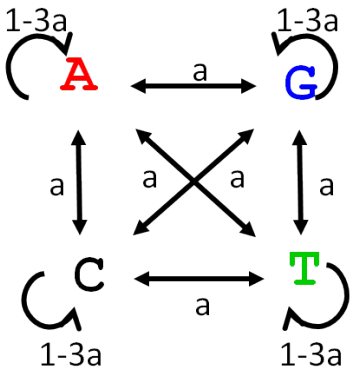
The simplest Markov model of sequence evolution for DNA is the *Jukes-Cantor model*[1], which assumes that all substitutions ($A \to C, A \to G, A \to T, C \to A...$) are equally probable and occur at a rate, α. The consequence of this assumption is that the overall rate of substitution is $\lambda = 3\alpha$. That is, λ is the probability that a given nucleotide will be replaced by *some* other nucleotide in one time step. The probability that the nucleotide remains unchanged is $1 - 3\alpha$.

The transition probability matrix for this Markov model is:

$$\begin{bmatrix} & A & G & C & T \\ A & 1-3\alpha & \alpha & \alpha & \alpha \\ G & \alpha & 1-3\alpha & \alpha & \alpha \\ C & \alpha & \alpha & 1-3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & 1-3\alpha \end{bmatrix}$$

[1] Jukes and Cantor, Evolution of protein molecules. In H. N. Munro, (ed.) *Mammalian Protein Metabolism*, 21-123, Adademic Press, NY, 1969.

This model can be represented graphically like this:



The stationary distribution of this Markov chain is $\varphi^* = (0.25, 0.25, 0.25, 0.25)$. (Verify that this is so by checking that $\varphi^* = \varphi^* P$).

The rate, α, of each possible substitution is an explicit parameter of the Jukes-Cantor model. In addition, the frequencies of A's, G's, C's and T's are implicitly specified by the model, since this is determined by the stationary distribution. Nucleotide substitution models can be made more realistic in two directions. First, the assumption that all substitutions occur at the same rate can be relaxed. Second, the specification of the rates can be adjusted to yield a non-uniform stationary distribution, since the assumption that all four bases have the same frequency ($\varphi_A = \varphi_C = \varphi_G = \varphi_T$) is unlikely to hold in most data sets.

Non-uniform transition probabilities: *The Kimura 2 Parameter (K2P) model* assumes that transitions and transversions occur at different rates. A *transition* is the substitution of a purine for another purine or a pyrimidine for another pyrimidine. A *transversion* is the substitution of a purine for a pyrimidine or a pyrimidine for a purine. Recall that the pyrimidines, including cytosine and thymine, are nucleotides with a ring with six elements. The purines, including adenine and guanine, have a pyrimidine ring fused to a five-sided imidazole ring. It make sense that transversions would proceed at a different rate than transitions, since substituting a purine with a pyrimidine, or vice versa, involves a greater change in size and shape than a substitution of two nucleotides from the same class. The transition matrix for the K2P model is

$$\begin{bmatrix} & A & G & C & T \\ A & 1-\alpha-2\beta & \alpha & \beta & \beta \\ G & \alpha & 1-\alpha-2\beta & \beta & \beta \\ C & \beta & \beta & 1-\alpha-2\beta & \alpha \\ T & \beta & \beta & \alpha & 1-\alpha-2\beta \end{bmatrix},$$

where α is the rate of transitions and β is the rate of transversions.

Non-uniform stationary distributions: Like the Juke Cantor model, the K2P model has a uniform stationary distribution, $\varphi^* = (0.25, 0.25, 0.25, 0.25)$. (Work through the algebra to convince yourself that this is true.) However, this is not a realistic model for the many genomes in which the G+C content deviates from 50%. The Felsenstein (1981) model, like the Jukes-Cantor model, assumes that all substitutions are equally likely, but can allow for an arbitrary stationary distribution, $\varphi = (\varphi_A, \varphi_C, \varphi_G, \varphi_T)$, where $\varphi_A \neq \varphi_C \neq \varphi_G \neq \varphi_T$. The *Felsenstein model* transition matrix is

$$\begin{bmatrix} & A & G & C & T \\ A & 1-\alpha\,(\varphi_C+\varphi_G+\varphi_T) & \varphi_G\,\alpha & \varphi_C\,\alpha & \varphi_T\,\alpha \\ G & \varphi_A\,\alpha & 1-\alpha\,(\varphi_A+\varphi_C+\varphi_T) & \varphi_C\,\alpha & \varphi_T\,\alpha \\ C & \varphi_A\,\alpha & \varphi_G\,\alpha & 1-\alpha\,(\varphi_A+\varphi_G+\varphi_T) & \varphi_T\,\alpha \\ T & \varphi_A\,\alpha & \varphi_G\,\alpha & \varphi_C\,\alpha & 1-\alpha\,(\varphi_A+\varphi_C+\varphi_G) \end{bmatrix}$$

The *Hasegawa, Kishino, Yano (HKY) model*, which combines both innovations, allows different rates for transitions and transversions and an arbitrary stationary distribution, $\varphi^* = (\varphi_A, \varphi_C, \varphi_G, \varphi_T)$. The HKY transition matrix is:

$$\begin{bmatrix} & A & G & C & T \\ A & 1-\alpha\,\varphi_G-\beta\,(\varphi_C+\varphi_T) & \varphi_G\,\alpha & \varphi_C\,\beta & \varphi_T\,\beta \\ G & \varphi_A\,\alpha & 1-\alpha\,\varphi_A-\beta\,(\varphi_C+\varphi_T) & \varphi_C\,\beta & \varphi_T\,\beta \\ C & \varphi_A\,\beta & \varphi_G\,\beta & 1-\alpha\,\varphi_T-\beta\,(\varphi_A+\varphi_G) & \varphi_T\,\alpha \\ T & \varphi_A\,\beta & \varphi_G\,\beta & \varphi_C\,\alpha & 1-\alpha\,\varphi_C-\beta\,(\varphi_A+\varphi_G) \end{bmatrix}$$

The *General Time Reversible (GTR) model* is an even more general model that allows a different rate for each of the six possible substitutions and an arbitrary stationary distribution. These models are discussed in greater detail in various molecular evolution textbooks; see, for example, Li's *Molecular Evolution*, (Sinauer Associates, 1997).

In deciding which model to use for a particular data set, we face the usual tradeoff: more general models with more parameters provide a more accurate representation of the underlying evolutionary process. However, with more complex models, more data is required to infer the parameter values and the danger of overfitting the parameters is greater.

Analyses of alignments of present-day sequences suggest that, in many sequence families, the rate of change varies from site to site. This is typically addressed by assuming that sequence substitution in a given family can be captured by a single model with a small number of rate categories. For example, one might model substitution in a given family using the Jukes Cantor model with four rates, $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. For each site, i, maximum likelihood estimation is used to estimate probabilities $(p_1(i), p_2(i), p_3(i), p_4(i))$, where $p_r(i)$ is the probability that site i is evolving at rate α_r.

Ziheng Yang discusses this approach in his textbook *Computational Molecular Evolution* (Oxford University Press, 2006).

These models do not allow for changes in rate or in GC-content over time. Developing models to account for temporal changes in rate or nucleotide composition is currently an active area of research.

Two applications of DNA substitution models

Markov models of sequence substitution are used to answer a wide range of questions that arise in molecular evolution, including simulating sequence evolution, estimating rates of evolution, and deriving substitution scoring matrices. Here we demonstrate how DNA substitution models can be used to estimate the likelihood of observing a pair of aligned nucleotides, given a phylogenetic model, and to correct for multiple substitutions. In future lectures, we will use an amino acid substitution model to derive a scoring matrix.

The likelihood of a pair of aligned sequences First, let's consider the problem of estimating the likelihood of a pair of aligned sequences. This problem arises in maximum likelihood approaches to estimating a phylogenetic tree. Maximum likelihood estimation (MLE) is a general method for estimating parameters of a model. It is based on the assumption that the observed data is best explained by the model that maximizes its likelihood; that is, the model for which the probability of the data is highest. Given a parameterized model, the parameter values are estimated by determining the values that maximize the probability of the data.

In the context of phylogeny estimation, the observed data is a set of k aligned sequences. The model has two components: a Markov model of sequence substitution and a rooted, binary tree with k leaves. The likelihood is the probability of observing the multiple alignment, under the assumption that the sequences evolved along the branches of the tree, sustaining mutations according to the rates specified by the substitution model. For a fixed tree topology, the branch lengths and the substitution rates are estimated by maximizing the probability of the multiple sequence alignment.

We demonstrate this calculation for the case where $k = 2$. Suppose we have two residues, x and y, that are the descendants of an ancestral nucleotide, z, and that time t has elapsed since their divergence (Fig. 1). The probability of observing x aligned with y is the product of $p_{zx}(t)$, the probability of observing an x after time t, and $p_{zy}(t)$, the probability of observing a y at time t, given that the ancestral residue was z. Since the base in the ancestral sequence is unknown, we estimate the probability by the weighted sum over all possible values of z:

$$\sum_{z \in \{A,C,G,T\}} p_z \; p_{zx}(t) \; p_{zy}(t), \tag{1}$$

where p_z is an estimation of the frequency of z in the ancestral sequence.

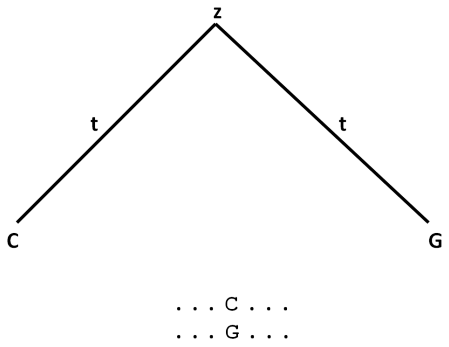


Figure 1: Hypothetical evolutionary scenario: C and G in present-day nucleotide sequences are descended from an unknown ancestral nucleotide, z.

For example, the likelihood of observing a cytosine in one sequence and a guanine in the other is

$$\mathcal{L}(GC|t) = p_A\, p_{AC}(t)\, p_{AG}(t) + p_C\, p_{CC}(t)\, p_{CG}(t) + p_G\, p_{GC}(t)\, p_{GG}(t) + p_T\, p_{TC}(t)\, p_{TG}(t) \qquad (2)$$

In order to estimate the probability of observing x aligned with y, given that a time interval t has elapsed since they diverged from their common ancestor, we need a way to estimate $p_{zx}(t)$ and $p_{zy}(t)$. Here, we derive expressions for the probability $p_{zx}(t)$ under the assumption that the sequence is evolving according to the Jukes Cantor model. The Jukes Cantor transition probability matrix,

$$\begin{bmatrix} 1-3\,\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\,\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\,\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\,\alpha \end{bmatrix}$$

is defined in terms of an instantaneous substitution rate, α. If the duration of a single time step in this Markov chain is Δt, $\Delta t \ll \frac{1}{\alpha}$, then the probability of a substitution between a given pair of nucleic acids in a single time step is $\alpha\,\Delta t$. We can define a transition probability matrix for the Jukes Cantor Markov model with this time interval as follows:

$$\begin{bmatrix} 1-3\,\alpha\,\Delta t & \alpha\,\Delta t & \alpha\,\Delta t & \alpha\,\Delta t \\ \alpha\,\Delta t & 1-3\,\alpha\,\Delta t & \alpha\,\Delta t & \alpha\,\Delta t \\ \alpha\,\Delta t & \alpha\,\Delta t & 1-3\,\alpha\,\Delta t & \alpha\,\Delta t \\ \alpha\,\Delta t & \alpha\,\Delta t & \alpha\,\Delta t & 1-3\,\alpha\,\Delta t \end{bmatrix}.$$

We use this transition matrix to derive an expression describing how changes accumulate at site i over a period of time t. First, we consider the event of observing, for example, an A at site i after a single time step; i.e., at time $t + \Delta t$. This event can occur in two ways: either site i contained an A at time t and no substitution occurred during the most recent time step or site i contained some

other nucleotide at time t and a substitution resulted in an A one time step later. Accounting for both of these scenarios, the probability of observing A at time $t + \Delta t$ is

$$\varphi_A(t + \Delta t) = (1-3\,\alpha\,\Delta t)\varphi_A(t) + \alpha\,\Delta t\,\varphi_C(t) + \alpha\,\Delta t\,\varphi_G(t) + \alpha\,\Delta t\,\varphi_T(t)$$

where $\varphi_x(t)$ is the probability of observing nucleotide x (i.e. of being in state E_x) at time t. Here, the first term gives the probability that the residue at site i at time t was an A and no substitution occurred. The second term is the probability that the residue at time t was not an A and a substitution did occur, replacing that residue with A. This reduces to

$$\varphi_A(t + \Delta t) = (1-3\,\alpha\,\Delta t)\,\varphi_A(t) + \alpha\,\Delta t\,\left[1-\varphi_A(t)\right].$$

Since the model is symmetric, this equation applies equally well to C, G or T. We can therefore rewrite the equation using the parameter x, where $x \in \{A, C, G, T\}$, and combine terms to obtain

$$\varphi_x(t + \Delta t) = (1-4\,\alpha\,\Delta t)\varphi_x(t) + \alpha\,\Delta t\,. \tag{3}$$

Having obtained an expression for the probability of observing a given nucleotide (x) after one time step, next, we need to derive an expression for the probability of observing x after a longer time interval. Subtracting $\varphi_x(t)$ from both sides of Equation 3 and some algebraic manipulation yields

$$\frac{\varphi_x(t + \Delta t) - \varphi_x(t)}{\Delta t} = \alpha\,(1-4\varphi_x(t))\,.$$

Taking the limit as $\Delta t \to 0$, we obtain a differential equation

$$\frac{d\varphi_x(t)}{dt} = \alpha\,(1-4\varphi_x(t))$$

that we can use to obtain an expression for the probability of observing nucleotide x at site i after an arbitrary time interval t. This differential equation has a standard form ($f'(t) = a - bf(t)$) with a known solution, from which we obtain

$$\varphi_x(t) = \frac{1}{4} + \left(\varphi_x(0) - \frac{1}{4}\right) e^{-4\alpha t}.$$

We now have an expression that depends on the initial state probability $\varphi_x(0)$. We have two cases. The probability that the present-day residue is the same as the ancestral nucleotide ($\varphi_x(0) = 1$) after time t is

$$p_{xx}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}. \tag{4}$$

The probability that the present-day nucleotide differs from the ancestral residue ($\varphi_x(0) = 0$) is

$$p_{zx}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}. \tag{5}$$

Equations 4 and 5 correspond to the terms in the expression for the likelihood in Equation 1. Substituting the right hand sides of Equations 4 and 5 into Equation 2, we obtain the likelihood for observing G aligned with C

$$\mathcal{L}(G, C|t, \alpha) = \frac{1}{2}\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^2 + \frac{1}{2}\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right) \cdot \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right),$$

assuming that $p_z = \frac{1}{4}$ for all z. We now have an expression for the probability of observing C aligned with G that depends on two parameters: the branch length, t, and the substitution rate, α. These parameter values are then estimated by finding the values of t and α that maximize $\mathcal{L}(G, C|t, \alpha)$.

This approach can be expanded to values of k greater than two by nesting multiple expressions with the same form as the right hand side of Equation 1. Under the assumption of positional independence, the likelihood for multiple sites is simply the product of the likelihoods for each site, individually.

Correcting for multiple substitutions. Another task that arises in molecular evolution is estimating the amount of sequence divergence between a pair of sequences. For example, the progressive alignment heuristic for multiple sequence alignment requires a matrix of the pairwise distances between all pairs of sequences, s_a and s_b.

A simple approach to estimating the distance between s_a and s_b would be to count the number of positions that are not identical in the pairwise alignment of s_a and s_b. If only a few changes have occurred, then the observed number of mismatches may, in fact, be the actual number of substitutions. However, as the divergence increases, so does the probability of two or more substitutions at the same site. In this case, the number of observed changes will underestimate the actual divergence as shown in Fig. 2.

Recall that the Jukes-Cantor model assumes that all substitutions ($A \to C, A \to G, A \to T, C \to A$...) are equally likely and occur at a rate α. The consequence of this assumption that the overall rate of substitution is $\lambda = 3\alpha$. Suppose that we have an ungapped[2] pairwise alignment of length n of two nucleotide sequences, σ and τ, that disagree at m positions. We wish to estimate the number of substitutions that actually occurred over t, the time interval that elapsed since they diverged from a common ancestor.

Here, we use the Jukes-Cantor model to derive a more accurate estimate of the number of substitutions. If we assume a constant rate of substitution, λ, in both lineages then the expected number of substitutions per site is

$$\begin{aligned} P_s &= 2\lambda t \\ &= 6\alpha t. \end{aligned}$$

[2]None of the substitution models we have discussed account for insertions and deletions.

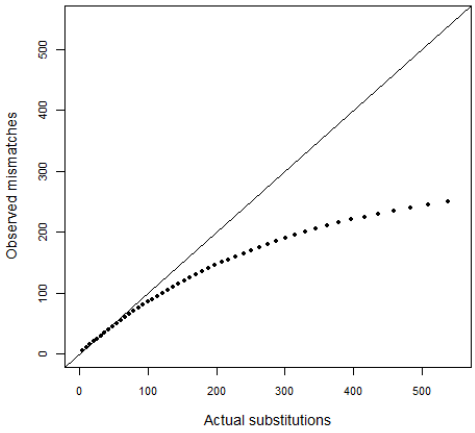


Figure 2: Observed versus actual number of substitutions.

Since both α and t are unknown, we estimate the expected number of substitutions from the frequency of mismatches in the current alignment. Given a Markov model of sequence substitution, we can use the observed frequency of mismatches to estimate $2\lambda t$ using the following strategy:

First, using the expressions for $p_{xx}(t)$ and $p_{xy}(t)$ that we derived in the previous section, we estimate the frequency of mismatches as a function of αt,

$$P_m = f(\alpha t).$$

We do this by estimating P_M, the frequency of matches, and subtracting to obtain $P_m = 1 - P_M$. Next, we invert this function to obtain an expression for the expected number of substitutions at a single site in terms of the number of mismatches.

$$\alpha t = f^{-1}(P_m).$$

The frequency of mismatches can be approximated by $\frac{m}{n}$, yielding an equation of the form

$$\alpha t \approx f^{-1}(\tfrac{m}{n}).$$

From this, we obtain an approximation for the frequency of substitutions:

$$\begin{aligned} P_s &= 6\alpha t \\ &\approx 6f^{-1}(\tfrac{m}{n}). \end{aligned}$$

Now we apply this strategy to obtain an estimate of the number of substitutions that occurred assuming that sequences are evolving according to the Jukes Cantor model. First, we derive an

expression for the probability of observing a match; for example, for observing two adenines aligned at site i. Given two sequences evolving independently from a common ancestral sequence, the probability that both sequences will have an A at site i is

$$P_M = [p_{AA}(t)]^2 + [p_{TA}(t)]^2 + [p_{CA}(t)]^2 + [p_{GA}(t)]^2 ,$$

where t is the elapsed time since their divergence. Replacing the first term with Equation 4 and the remaining terms with Equation 5, this reduces to

$$P_M = \left[\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right]^2 + 3\left[\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right]^2 .$$

The first term gives the probability of observing A's in both sequences if the ancestral nucleotide was also A. The second term represents the case where the ancestral nucleotide was not an A. By expanding the squared quantities and combining terms, we obtain

$$P_M = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}. \tag{6}$$

Note that since the Jukes Cantor model is symmetric, Equation 6 in fact gives the probability of observing the same nucleotide x in both sequences, where x may be any nucleotide. P_m, the probability of observing a mismatch at site i, is simply $1-P_M$ or

$$P_m = \frac{3}{4}(1 - e^{-8\alpha t}). \tag{7}$$

We solve the above equation to obtain an expression for αt in terms of P_m:

$$\alpha t = -\frac{1}{8}\ \ln\left(1 - \frac{4}{3}P_m\right).$$

Multiplying both sides of the equation by 6 yields the expected frequency of substitutions per site in terms of the probability of observing a mismatch,

$$P_s = -\frac{3}{4}\ln\left(1 - \frac{4}{3}P_m\right).$$

P_m can be estimated by the observed frequency of mismatches, allowing us to obtain an estimate in terms of the fraction of sites with an observable difference:

$$P_s \approx -\frac{3}{4}\ln\left(1 - \frac{4}{3}\ \frac{m}{n}\right).$$

Multiplying by n yields an estimate of the expected number of substitutions that actually occurred:

$$-\frac{3}{4}\ln\left(1 - \frac{4}{3}\ \frac{m}{n}\right) \cdot n. \tag{8}$$

So, for example, if we observe mismatches at 100 sites in a nucleotide sequence of length 1,000, then the Jukes-Cantor model predicts that the actual number of substitutions per site is 0.107 or 107 substitutions.

Applications with the K2P model Equations that are analogous to Equations 4, 5, and 8 can be similarly derived for the K2P model, which assumes that transitions and transversions occur at different rates. We state these results here without deriving them. The probability of observing the same nucleotide after elapsed time t in the K2P model is

$$p_{xx}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}, \tag{9}$$

where α is the rate of transitions and β is the rate of transversions. This equation is analogous to Equation 4. The probability of observing a different nucleotide is

$$p^s_{xy}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} \tag{10}$$

if $x \to y$ is a transition and

$$p^v_{xy} = \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \tag{11}$$

if $x \to y$ is a transversion. Note, this is the probability of the occurence of a *specific* transversion (e.g., `A` $\to$ `C`). The probability of the occurrence of *any* transversion (e.g., `A` $\to$ `C` or `A` $\to$ `T`) is

$$\frac{1}{2} - \frac{1}{2}e^{-4\beta t}. \tag{12}$$

Given an alignment of two sequences that diverged from a common ancestor that lived t years in the past and have since been evolving according to the K2P model, the probability of observing a transition (i.e., `AG`, `GA`, `CT`, `TC`) at a given site i is

$$P^s_m = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}). \tag{13}$$

The probability of observing a transversion at i is given by

$$P^v_m = \frac{1}{2}(1 - e^{-8\beta t}). \tag{14}$$

Given an alignment of length n, with m_s transitions and m_v transversions, the expected number of transitions and transversions that actually occured are

$$\left[-\frac{1}{2}\ln\left(1-\frac{2m_s}{n} - \frac{m_v}{n}\right) + \frac{1}{4}\ln\left(1-\frac{2m_v}{n}\right)\right] \cdot n \tag{15}$$

and

$$-\frac{1}{2}\ln\left(1-\frac{2m_v}{n}\right) \cdot n, \tag{16}$$

respectively. Summing these two quantities, we obtain the expected number of substitutions of all types:

$$\left[-\frac{1}{2}\ln\left(1-\frac{2m_s}{n} - \frac{m_v}{n}\right) - \frac{1}{4}\ln\left(1-\frac{2m_v}{n}\right)\right] \cdot n. \tag{17}$$