research reports

no. 422 november 2001

jens ledet jensen

poisson approximation to the distribution of rare events in DNA evolution

department of theoretical statistics

university of aarhus

Poisson approximation to the distribution of rare events in DNA evolution

Jens Ledet Jensen
University of Aarhus
Aarhus, Denmark

Abstract

We consider a class of models for the evolution of a DNA sequence that allows for interaction among neighbouring elements. When the evolutionary distance between two sequences is small the number of changes along the sequence will be small. Categorizing the changes we prove a multivariate Poisson approximation to the distribution of the number of changes.

Key words:

1 Introduction

A DNA string consists of a long sequence of nucleotides of which there are four: A, G, C, and T. In a coding region these go together three by three (a reading frame) forming the codons that translate into the aminoacids. When comparing two aligned sequences it is customary to model the substitution part of the evolution of a DNA string as a continuous time Markov process. In the most simple models it is assumed that the different nucleotides evolve independently and the analysis is reduced to that of a Markov process with four states (Felsenstein, 1981). Similarly, for the slightly more general model where it is assumed that the codons evolve independently, the analysis is reduced to that of a Markov process with 64 states (Li et al., 1985). However, in some cases these models are not realistic. An instance of this is when there are overlapping reading frames. In such a case an aminoacid changing substitution

in reading frame I may or may not change the aminoacid coded for in reading frame II. This introduces interaction among neighbouring codons.

A model that takes the interaction among neighbouring codons into account has recently been proposed in Jensen and Pedersen (2000). However the analysis of this model is very cumbersome as the likelihood function can only be found via an MCMC simulation. In Pedersen and Jensen (2001) it is suggested to use a Poisson approximation to the observed number of changes. The Poisson approximation can be used both as a way of estimating the parameters and as a way of checking the model. The theoretical mean of the Poisson distribution must be found by simulation, but this simulation is straightforward and quick. In this note we show that the Poisson approximation can be theoretically justified.

We start in Section 2 by stating a Poisson approximation theorem of Arratia, Goldstein and Gordon (1989) for sums of dependent Bernoulli variables. In order to apply this theorem an estimate of the mixing properties of the variables is needed. A formal setup for an underlying process is introduced for which we can establish a coupling inequality and therby estimate the mixing properties. In Section 3 we consider the DNA evolutionary model of Jensen and Pedersen (2000) and Pedersen and Jensen (2001) and show how this model fits into the framework of Section 2. We also establish a central limit result to cover the cases where the Poisson approximation is no longer applicable. In the concluding Section 4 a small numerical example is given.

2 Preliminaries

2.1 Multivariate Poisson approximation

We will be using the multivariate Poisson approximation given in Arratia, Goldstein and Gordon (1989). Let Ω be an arbitrary index set and for $\alpha \in \Omega$ let U_α be a

Bernoulli random variable with $p_\alpha = P(U_\alpha = 1) = 1 - P(U_\alpha = 0) > 0$. Let Ω be partioned into disjoint nonempty subsets $\Omega(1), \ldots, \Omega(d)$, and define

$$S_j = \sum_{\alpha \in \Omega(j)} U_\alpha, \qquad \lambda_j = \sum_{\alpha \in \Omega(j)} p_\alpha.$$

For each $\alpha \in \Omega$ let M_α be a subset of Ω with $\alpha \in M_\alpha$. Let $Z_1, \ldots, Z_d$ be independent Poisson variables with $EZ_j = \lambda_j$.

Theorem 2.1 *(Arratia, Goldstein and Gordon, 1989)*
The total variation distance between the joint distribution of $(S_1, \ldots, S_d)$ *and* $(Z_1, \ldots, Z_d)$ *is bounded by*

$$2 \min \left\{ 1, \frac{1.4}{(\min \lambda_i)^{1/2}} \right\} (2b_1 + 2b_2 + b_3), \tag{2.1}$$

with

$$b_1 = \sum_{\alpha \in \Omega} \sum_{\beta \in M_\alpha} p_\alpha p_\beta, \tag{2.2}$$

$$b_2 = \sum_{\alpha \in \Omega} \sum_{\beta \in M_\alpha \setminus \{\alpha\}} E(U_\alpha U_\beta), \tag{2.3}$$

$$b_3 = \sum_{\alpha \in \Omega} E|E\{U_\alpha - p_\alpha | (U_\beta, \beta \in \Omega \setminus M_\alpha)\}|. \tag{2.4}$$

In our use of this theorem we have an underlying process $X_1, \ldots, X_n$ and

$$U_{il} = 1((X_{i-1}, X_i, X_{i+1}) \in B_{il}), \quad p_{il} = EU_{il}, \quad i = 1, \ldots, n, \quad l = 1, \ldots, d, \tag{2.5}$$

where $B_{i,1}, \ldots, B_{i,d}$ are disjoint subsets. We therefore take

$$\begin{aligned} \Omega &= \{1, 2, \ldots, n\} \times \{1, \ldots, d\}, \\ \Omega(l) &= \{(i, l) | i = 1, \ldots, n\}, \quad l = 1, \ldots, d, \\ M_{il} &= \{(j, m) | \, |j - i| < k(n)\}, \end{aligned} \tag{2.6}$$

for some number $k(n)$. Since the sets B_{il}, $l = 1, \ldots, d$, are disjoint we have that $U_{il} = 1$ implies that $U_{im} = 0$ for $m \neq l$, or

$$E(U_{il} U_{im}) = 0, \quad m \neq l. \tag{2.7}$$

The error terms b_1 and b_2 will be bounded by finding upper bounds on EU_{il} and $E(U_{il}U_{jm})$, respectively. Finding a bound on the error term b_3 is more complicated, and for that we need the coupling inequality of the next subsection.

2.2 A coupling inequality

Let X_i take values in a measure space $\mathcal{X}$ equipped with a measure μ. Assume that there exist functions

$$\phi_i(\cdot|\cdot,\cdot) : \mathcal{X}^3 \to [0,\infty), \quad i = 1,\ldots,n,$$

such that the joint density of $X_1,\ldots,X_n$ with respect to μ^n is

$$\frac{1}{C_n}\prod_{i=1}^{n}\phi_i(x_i|x_{i-1},x_{i+1}), \tag{2.8}$$

where C_n is a norming constant and x_0 and x_{n+1} are fixed points. This Gibbs form of the density introduces a second order Markov structure on the process and it will be convenient to have a notation for a consequtive pair of variables. Thus we let $x_{[i]} = (x_i, x_{i+1})$. The structure of the density (2.8) implies that the conditional density of $(X_{r+1},\ldots,X_{s-1})$ given the remaining variables depends on $x_{[r-1]}$ and $x_{[s]}$ only, and is given by

$$\frac{1}{Z(x_{[r-1]},x_{[s]})}\prod_{i=r}^{s}\phi_m(x_i|x_{i-1},x_{i+1}), \tag{2.9}$$

where $Z(x_{[r-1]},x_{[s]})$ is a norming constant. We say that sets $\mathcal{A}_k \subseteq \mathcal{X}^2$, $k = 1,\ldots,n$, are atoms for the process if the conditional density (2.9) is the same for all

$$(x_{[r-1]},x_{[s]}) \in \mathcal{A}_{r-1}\times\mathcal{A}_s.$$

In order to introduce a coupling we consider another process $\{Y_i\}$ independent of the $\{X_i\}$ process. We fix $I < J$ and consider the Y_i process for $i = I-1,\ldots,J+1$. We require that the conditional distribution of $(Y_{I+1},\ldots,Y_{J-1})$ given $(Y_{[I-1]},Y_{[J]}) = (y_{[I-1]},y_{[J]})$ has density (2.9) with $(x_{[I-1]},x_{[J]})$ replaced by $(y_{[I-1]},y_{[J]})$. The marginal

distribution of $(Y_{[I-1]}, Y_{[J]})$ is denoted Q_{IJ} and that of $(X_{[I-1]}, X_{[J]})$ is denoted P_{IJ}. Note that for variables with index between $I+1$ and $J-1$ the Y-process has the same Gibbs structure as the X-process.

To shorten the formulae below we use for $v \geq 1$ the notation

$$\begin{aligned} X(v) &= (X_{[I+v]}, X_{[J-v-1]}), \\ Y(v) &= (Y_{[I+v]}, Y_{[J-v-1]}), \\ \mathcal{A}(v) &= \mathcal{A}_{I+v} \times \mathcal{A}_{J-v-1}. \end{aligned}$$

For r with $2r+2 \leq J-I-1$ we introduce the coupling by the event

$$E_r = \left\{ (X(v), Y(v)) \notin \mathcal{A}(v)^2,\ v = 1, \ldots, r-1,\ (X(r), Y(r)) \in \mathcal{A}(r)^2 \right\}.$$

Furthermore we define

$$\begin{aligned} F_r &= (E_1 \cup \cdots \cup E_r)^{\mathrm{C}} \\ &= \left\{ (X(v), Y(v)) \notin \mathcal{A}(v)^2,\ v = 1, \ldots, r \right\}. \end{aligned}$$

Because of the structure (2.5) of the Bernoulli variables we formulate the following coupling inequality.

Lemma 2.2 *For any set B and for $r = \min\{k-I, J-k\} - 3$ we have*

$$|P((X_{k-1}, X_k, X_{k+1}) \in B) - P((Y_{k-1}, Y_k, Y_{k+1}) \in B)| < 2P(F_r).$$

Proof. Let $B_k(X)$ be the event that $(X_{k-1}, X_k, X_{k+1}) \in B$ and similarly for $B_k(Y)$. We write, with $a(v)$ an arbitrary point in the set $\mathcal{A}(v)$,

$$\begin{aligned} P(B_k(X)) &= \sum_{v=1}^{r} P(B_k(X), E_v) + P(B_k(X), F_r) \\ &= \sum_{v=1}^{r} P(E_v) P(B_k(X) | X(v) \in \mathcal{A}(v)) + P(B_k(X), F_r) \\ &= \sum_{v=1}^{r} P(E_v) P(B_k(X) | X(v) = a(v)) + P(B_k(X), F_r) \\ &= \sum_{v=1}^{r} P(E_v) P(B_k(Y) | Y(v) = a(v)) + P(B_k(X), F_r) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{v=1}^{r} P(E_v) P(B_k(Y) | Y(v) \in \mathcal{A}(v)) + P(B_k(X), F_r) \\
&= \sum_{v=1}^{r} P(B_k(Y), E_v) + P(B_k(X), F_r) \\
&= P(B_k(Y)) - P(B_k(Y), F_r) + P(B_k(X), F_r), \qquad (2.10)
\end{aligned}
$$

where in the second equality we use the Gibbs structure and the definition of the atoms $\mathcal{A}_i$ to obtain

$$
\begin{aligned}
P(B_k(X)|E_v) &= \int_{\mathcal{A}(v)} P(B_k(X)|X(v) = a(v)) P_{X(v)}(da(v)|E_v) \\
&= P(B_k(X)|X(v) = a(v)).
\end{aligned}
$$

The result of the lemma follows directly from (2.10). ■

When $I = 0$ we use a onesided version of the lemma, where the Y process has Y_0 fixed at x_0. Thus we use

$$
\bar{E}_r = \{(X_{[J-v-1]}, Y_{[J-v-1]}) \notin \mathcal{A}^2_{J-v-1},\ v = 1, \ldots, r-1,\ (X_{[J-r-1]}, Y_{[J-r-1]}) \in \mathcal{A}^2_{J-r-1}\},
$$

and

$$
\bar{F}_r = (\bar{E}_1 \cup \cdots \cup \bar{E}_r)^{\mathrm{C}} = \{(X_{[J-v-1]}, Y_{[J-v-1]}) \notin \mathcal{A}^2_{J-v-1},\ v = 1, \ldots, r\}.
$$

The result of the lemma is then true with $r = J - k - 3$. Similar modifications are made when $J = n + 1$.

In order to bound the probability of the event $F(r)$ we make the following assumption.

A1: There exist $\delta > 0$ and $\tau \geq 1$ such that for all i and all $z_1, z_2 \in \mathcal{X}^2$

$$
P(X_{[i]} \in \mathcal{A}_i | X_{[i-\tau-1]} = z_1, X_{[i+\tau+1]} = z_2) \geq \delta. \qquad (2.11)
$$

From this assumption we get the bound

$$
P\left((X(v), Y(v)) \notin \mathcal{A}(v)^2 \,\middle|\, \begin{array}{l} X(v-\tau-1) = x_1, X(v+\tau+1) = x_2 \\ Y(v-\tau-1) = y_1, Y(v+\tau+1) = y_2 \end{array} \right) \leq 1 - \delta^4, \qquad (2.12)
$$

where x_1, x_2, y_1, y_2 are arbitrary points in $\mathcal{X}^4$. To bound the probability of the event $F(r)$ we condition on the variables

$$\{X(-1+j(2\tau+1)), Y(-1+j(2\tau+1))\}, \quad j=0,1,\ldots,m(r),\ m(r)=\left[\frac{r+3}{2\tau+1}\right],$$

so that we can use the bound (2.12) $m(r)$ times. We thus get the upper bound

$$(1-\delta^4)^{m(r)} \tag{2.13}$$

for the probability of the event $F(r)$.

We now consider the Bernoulli variables defined by (2.5) and obtain a bound on the term b_3 from (2.4).

Proposition 2.3 *Under the assumption (2.11) we have the bound*

$$b_3 \le 2nd(1-\delta^4)^{(k(n)+1)/(2\tau+1)-1},$$

where $k(n)$ appears in the definition (2.6) of M_{il}.

Proof. We write b_3 as

$$\begin{aligned}
b_3 &= \sum_{i-1}^{n}\sum_{l=1}^{d} E|P(U_{il}=1|U_{jm},(j,m)\notin M_{il}) - P(U_{il}=1)| \\
&= \sum_{i-1}^{n}\sum_{l=1}^{d} E|E\{[P(U_{i,l}=1|U_{j,m},(j,m)\notin M_{i,l}, X_j, |j-i|\ge k(n)-1) \\
&\qquad\qquad -P(U_{i,l}=1)]|U_{j,m},(j,m)\notin M_{i,l}\}| \\
&\le \sum_{i-1}^{n}\sum_{l=1}^{d} E\int \Big|P(U_{i,l}=1|X_{[I-1]}=x_{[I-1]}, X_{[J]}=x_{[J]}) - P(U_{i,l}=1)\Big| \\
&\qquad\qquad P_{X_{[I-1]},X_{[J]}}(d(x_{[I-1]},x_{[J]})|U_{j,m},(j,m)\notin M_{i,l}),
\end{aligned} \tag{2.14}$$

where $I = i-k(n)+1$ and $J = i+k(n)-1$.

Let now the Y-process be defined as above with the marginal distribution of $(Y_{[I-1]}, Y_{[J]})$ concentrated at the point $(x_{[I-1]}, x_{[J]})$. Then

$$|P(U_{i,l}=1|X_{[I-1]}=x_{[I-1]}, X_{[J]}=x_{[J]}) - P(U_{i,l}=1)|$$

$$
\begin{aligned}
&= |P((Y_{i-1}, Y_i, Y_{i+1}) \in B_{i,l}) - P((X_{i-1}, X_i, X_{i+1}) \in B_{i,l})| \\
&\leq 2P(F_{k(n)-2} | Y_{[I-1]} = x_{[I-1]}, Y_{[J]} = x_{[J]}) \\
&\leq 2(1-\delta^4)^{(k(n)+1)/(2\tau+1)-1}, \qquad (2.15)
\end{aligned}
$$

where we have used (2.13). Inserting (2.15) into (2.14) we obtain the bound

$$
b_3 \leq 2nd(1-\delta^4)^{(k(n)+1)/(2\tau+1)-1},
$$

which is the result of the proposition. ■

3 Applications to DNA evolutionary models

In this section we will consider the class of models studied in Jensen and Pedersen (2000) and Pedersen and Jensen (2001). In these models the evolution of a string of nucleotides is described by a continuous time Markov process with a jump consisting in the substitution of one nucleotide.

We describe the DNA string in terms of the codons within a reading frame, which we denote reading frame I. Let $(D_1(t), \ldots, D_n(t))$ be the codons at time t. We imagine that there are also codons D_0 and D_{n+1} that are kept fixed during the evolution. A codon has three nucleotides and we use the notation $D_i = (d_i^1, d_i^2, d_i^3)$. As mentioned in the introduction we can have more than one reading frame. The i'th codon in reading frame II is $(d_i^2, d_i^3, d_{i+1}^1)$ and the i'th codon in reading frame III is $(d_i^3, d_{i+1}^1, d_{i+1}^2)$. There are three *stop codons*, namely (TAA), (TAG), and (TGA). If a mutation gives rise to a stop codon the protein being coded for is destroyed and this will most likely imply that the new individual does not survive. In the model we do therefore not allow the appearence of a stop codon in a reading frame. In the model the intensity of a change of a nucleotide in a particular position depends on the neighbouring nucleotides that together form a codon in a reading frames. We

generally write

$$\lambda(r, z|D_{i-1}, D_i, D_{i+1}), \quad r = 1, 2, 3, \quad z \in \{A, G, C, T\}\backslash\{d_i^r\}, \tag{3.1}$$

for the rate of a substitution of d_i^r by z. The rate is zero if the substitution of a nucleotide by z produces a stop codon. Thus, if we are in the situation with three reading frames and we have a substitutiton at position $r = 1$ the rate is zero if either (z, d_i^2, d_i^3), $(d_{i-1}^2, d_{i-1}^3, z)$ of (d_{i-1}^3, z, d_i^2) is a stop codon. If there is only one reading frame it is (z, d_i^2, d_i^3) alone that cannot be a stop codon. Excluding the cases where a stop codon is produced all other transitions are allowed. There will therefore exist constants $\kappa_1 > 0$ and κ_2 so that the nonzero rates are bounded below by κ_1 and above by κ_2,

$$\kappa_1 \leq \lambda(r, z|D_{i-1}, D_i, D_{i+1}) \leq \kappa_2. \tag{3.2}$$

3.1 Conditioning on the initial sequence

We imagine that we have observed two sequences $(D_1(0), \ldots, D_n(0))$ and $(D_1(t^0), \ldots, D_n(t^0))$, where the former has evolved into the latter. The probabilities in this subsection will be for the conditional process given the initial sequence $(D_1(0), \ldots, D_n(0))$.

At each codon position $i = 1, \ldots, n$ we can ask if there has been a change from $D_i(0)$ to $D_i(t^0)$. If so we can categorise the change into, say, d classes, where the category may depend on what happens at codon positions $i - 1$ and $i + 1$. Formally, we define a Bernoulli random variable U_{il} for each class $l = 1, \ldots, d$ by

$$U_{il} = 1\left((D_{i-1}(0), D_i(0), D_{i+1}(0), D_{i-1}(t^0), D_i(t^0), D_{i+1}(t^0)) \in B_{i,l}\right) \tag{3.3}$$

where the sets B_{il} are such that

$$U_{il} \leq 1(D_i(0) \neq D_i(t^0)).$$

Also the sets B_{il}, $l = 1, \ldots, d$ are disjoint so that $U_{il} = 1$ implies that $U_{im} = 0$ for $m \neq l$.

We start by deriving an upper bound for the success probability $P(U_{il} = 1)$ and an upper bound for the joint success probability $P(U_{il} = 1, U_{jm} = 1)$, $j \neq i$.

Lemma 3.1 *We have the bounds*

$$p_{il} = EU_{il} \leq p\lambda \quad \textit{and} \quad E(U_{il}U_{jm}) \leq \frac{(2p\lambda)^2}{2}\mathrm{e}^{2p\lambda}, \quad j \neq i,$$

with

$$p = \frac{3\kappa_2}{n\kappa_1} \quad \textit{and} \quad \lambda = 9n\kappa_2 t^0.$$

Proof. If there were no restrictions due to stop codons any codon would at a given instant in time have the possibility of 9 substitutions. However, considering all the possible restrictions one can see that there always is at least 3 possible substitutions. From (3.2) the total intensity for a change in the sequence is therefore bounded between $3\kappa_1 n$ and $9\kappa_2 n$ and the expected number of substitutions for the sequence is bounded between $3\kappa_1 n t^0$ and $9\kappa_2 n t^0$. Conditionally on the past the probability that a given substitution takes place in codon j is bounded between

$$\frac{3\kappa_1}{9\kappa_2 n} \quad \text{and} \quad p = \frac{9\kappa_2}{3\kappa_1 n}.$$

Let K be the total number of substitutions in the sequence. Then K is stochastically smaller that a Poisson random variable with mean $\lambda = 9\kappa_2 n t^0$. We then find

$$EU_{il} \leq E1(D_i(0) \neq D_i(t^0)) \leq 1 - E1(D_i(0) = D_i(t^0)), \tag{3.4}$$

and

$$\begin{aligned} E1(D_i(0) = D_i(t^0)) &\geq P(\text{no substitutions in codon } i) \\ &\geq E(1-p)^K \\ &\geq \sum_{k=0}^{\infty}(1-p)^k \frac{\lambda^k}{k!}e^{-\lambda} \\ &= \exp(-\lambda + \lambda(1-p)). \end{aligned} \tag{3.5}$$

Combining (3.4) and (3.5) we obtain

$$EU_{il} \leq 1 - \exp(-\lambda p) \leq \lambda p. \tag{3.6}$$

This gives the first statement of the lemma.

For the second statement we argue in the same way. We use

$$\begin{aligned} E(U_{il}U_{jm}) &\leq P(D_j(0) \neq D_j(t^0), D_i(0) \neq D_i(t^0)) \\ &\leq P(\text{at least two substitution in codon } j \text{ and codon } i \text{ together}) \\ &\leq E\left\{\sum_{l=2}^{K} \left(\begin{array}{c} K \\ l \end{array} \right) p_{max}^l (1-p_{min})^{K-l}\right\} \\ &= E\left\{(1-p_{min})^K \left[(1+\frac{p_{max}}{1-p_{min}})^K - 1 - K\frac{p_{max}}{1-p_{min}}\right]\right\}, \end{aligned} \tag{3.7}$$

where

$$p_{max} = 2p \ \text{ and } \ p_{min} = cp_{max}, \ c = \frac{\kappa_1^2}{9\kappa_2^2}.$$

Since the integrand in (3.7) is increasing in K we can replace the distribution of K by a Poisson distribution with mean λ. This gives the upper bound

$$\begin{aligned} & \mathrm{e}^{-\lambda}\left\{\mathrm{e}^{\lambda(1-p_{min}+p_{max})} - 1 - \lambda((1-p_{min}+p_{max})\right\} - \mathrm{e}^{-\lambda}\left\{\mathrm{e}^{\lambda(1-p_{min})} - 1 - \lambda((1-p_{min})\right\} \\ & \quad -p_{max}\lambda\left\{\mathrm{e}^{-\lambda}\{\mathrm{e}^{\lambda(1-p_{min})} - 1\right\} \\ = \ & \mathrm{e}^{-\lambda p_{min}}\left[\mathrm{e}^{\lambda p_{max}} - 1 - \lambda p_{max}\right] \\ \leq \ & \frac{(\lambda p_{max})^2}{2}\mathrm{e}^{\lambda p_{max}} \\ = \ & \frac{(\lambda 2p)^2}{2}\mathrm{e}^{\lambda 2p}, \end{aligned}$$

and the second statement of the lemma has been proved. ■

We next turn to the mixing properties of the evolution process and prove the assumption (2.11). The variable X_i of Section 2 is now the path of codon i over the time interval $[0, t^0]$, for which we here use the notation $D_i(\cdot)$. That we have the

Gibbs form (2.8) for the density is seen from (3.8) below. As the atom $\mathcal{A}_i$ we take $\mathcal{A}_i = A_i \cap A_{i+1}$ where A_i is the event that there has been no substitutions in codon i,

$$A_i = \{D_i(t) \equiv D_i(0),\, 0 < t \leq t^0\}.$$

The development of a codon over the time interval $[0, t^0]$ can be described by giving the times at which changes occur and by specifying the position within the codon of the change and specifying the new nucleotide. This corresponds to a point in the space

$$H = \cup_{k=0}^{\infty} [0, t^0]^k \times M^k,$$

where M is the mark space containing the position within the codon and the new nucleotide, and where $k = 0$ corresponds to no changes of the codon. The state space for the development of the whole sequence of codons is then the n-fold product of the space H. As dominating measure on H we use Lebesgue measure on $[0, t^0]$ and counting measure on M.

The rate for a substitution of the r'th nucleotide in codon i is $\lambda(r, z|D_{i-1}, D_i, D_{i+1})$ from (3.1), where (D_{i-1}, D_i, D_{i+1}) is the value of codons $i-1$, i, and $i+1$ prior to the substitution. The total rate for a substitution in codon i is

$$\lambda(D_{i-1}, D_i, D_{i+1}) = \sum_{r,z} \lambda(r, z|D_{i-1}, D_i, D_{i+1}).$$

Let n_i be the total number of substitutions in codons $i-1$, i, and $i+1$ and $t(1,i), t(2,i), \ldots, t(n_i, i)$ the times at which the substitutions occur. Furthermore, we let

$$(D_{i-1}[j], D_i[j], D_{i+1}[j])$$

be the three codons after the j'th substitution. Let the number of substitutions in codon i be $\tilde{n}_i \leq n_i$, let these be numbered by $k(1,i), \ldots, k(\tilde{n}_i, i)$ among the n_i substitutions and let the positions of the $\tilde{n}_i$ substitutions within the codon be

$r(1,i),\ldots,r(\tilde{n}_i,i)$. The time points for these substitutions are then $t(k(1,i),i),\ldots,$ $t(k(\tilde{n}_i,i),i)$. Define

$$\begin{aligned}
\lambda_i(j) &= \lambda(D_{i-1}[j], D_i[j], D_{i+1}[j]),\\
\tilde{\lambda}_i(j) &= \lambda(r(j,i), d_i^{r(j,i)}[k(j,i)] | D_{i-1}[k(j,i)-1], D_i[k(j,i)-1], D_{i+1}[k(j,i)-1]),\\
V_i &= \prod_{j=1}^{\tilde{n}_i} \tilde{\lambda}_i(j),\\
W_i &= \sum_{j=1}^{n_i+1} \lambda_i(j)[t(j,i) - t(j-1,i)], \quad t(0,i)=0,\ t(n_i+1,i)=t^0.
\end{aligned}$$

The density for the development of the codons over $[0,t^0]$ can now be written as

$$\prod_{i=1}^{n} V_i \exp(-W_i). \tag{3.8}$$

If we consider $(D_{I+1}(\cdot), D_{I+2}(\cdot),\ldots,D_{J-1}(\cdot))$ conditional on $(D_1(\cdot),\ldots,D_I(\cdot),\ D_J(\cdot),$ $\ldots,D_n(\cdot))$ this will depend on $(D_{I-1}(\cdot),D_I(\cdot),D_J(\cdot),D_{J+1}(\cdot))$ only, and the conditional density is

$$\frac{1}{Z}\exp\left(-\sum_{i=I}^{J} W_i\right)\left(\prod_{i=I+1}^{J-1} V_i\right) 1(V_I>0)1(V_J>0), \tag{3.9}$$

where $Z = Z(D_{I-1}(\cdot),D_I(\cdot),D_J(\cdot),D_{J+1}(\cdot))$ is a norming constant.

Lemma 3.2 *There exists $\delta>0$ such that for all i and all paths $(D_{i-3}(\cdot),D_{i-2}(\cdot),$ $D_{i+3}(\cdot),D_{i+4}(\cdot))$ we have*

$$P(A_i \cap A_{i+1} | D_{i-2}(\cdot),D_{i-1}(\cdot),D_{i+2}(\cdot),D_{i+3}(\cdot)) \geq \delta.$$

Proof. We first note the following simple bounds:

$$\begin{aligned}
&\kappa_1 \leq \tilde{\lambda}_i(j)1(\tilde{\lambda}_i(j)>0) \leq \kappa_2,\\
&\kappa_1^{\tilde{n}_i} \leq V_i 1(V_i>0) \leq \kappa_2^{\tilde{n}_i},\\
&3\kappa_1 \leq \lambda_i(j) \leq 9\kappa_2,\\
&3\kappa_1 t^0 \leq W_i \leq 9\kappa_2 t^0,
\end{aligned}$$

where the first bound is simply (3.2).

To understand intutitively the idea in the lower bound of the lemma we will imagine that there are no restrictions due to stop codons, i.e. all rates are positive. Then, on taking $I = i - 1$ and $J = i + 2$ in (3.9), we get the bound

$$P(A_i \cap A_{i+1} | D_{i-2}(\cdot), D_{i-1}(\cdot), D_{i+2}(\cdot), D_{i+3}(\cdot)) \geq \frac{\exp(-4 \cdot 9\kappa_2 t^0)}{Z}. \tag{3.10}$$

To get an upper bound on Z from (3.9) we use the lower bound $3\kappa_1 t^0(J - I + 1)$ for the sum of the W_i's and use the upper bound $(\kappa_2)^{\tilde{n}_{I+1}+\cdots+\tilde{n}_{J-1}}$ for the product of the V_i's. Summing over the possible substitutions for a fixed number of substitutions κ_2 is replaced by $9\kappa_2$ in the upper bound for the V_i's. Next, integrating over the substitution times we get the upper bound

$$(9\kappa_2)^{\tilde{n}_{I+1}+\cdots+\tilde{n}_{J-1}} \exp\{-3\kappa_1 t^0(J - I + 1)\} \prod_{j=I+1}^{J-1} \frac{(t^0)^{\tilde{n}_j}}{\tilde{n}_j!}. \tag{3.11}$$

Finally, summing over the number of substitutions $\tilde{n}_j$, $j = I + 1, \ldots, J - 1$ we obtain the bound

$$Z \leq \exp\{-3\kappa_1 t^0(J - I + 1) + 9\kappa_2 t^0(J - I - 1)\}. \tag{3.12}$$

Inserting this bound into (3.10) we get a lower bound for the conditional probability of the event $A_i \cap A_{i+1}$.

When there are restrictions due to stop codons a change in the path $D_I(\cdot)$ may necessitate a change in the path of codon $I + 1$, otherwise V_I becomes zero. We therefore take $I = i - 2$ and $J = i + 3$ and use the bound

$$\begin{aligned} &P(A_i \cap A_{i+1} | D_{i-3}(\cdot), D_{i-2}(\cdot), D_{i+3}(\cdot), D_{i+4}(\cdot)) \\ &\quad \geq \quad P(A_i \cap A_{i+1} \cap C \cap \tilde{C} | D_{i-3}(\cdot), D_{i-2}(\cdot), D_{i+3}(\cdot), D_{i+4}(\cdot)). \end{aligned} \tag{3.13}$$

Here C is the event that there are no substitutions in codon $i - 1$ if this event does not make $V_{i-2} = 0$, and there is exactly one substitution changing d^1_{i-1} to a C before time t^*_{i-2} if this is needed in order that $V_{i-2} > 0$. Similarly, $\tilde{C}$ is the event that there

are no substitutions in codon $i+2$ if this event does not make $V_{i+3}=0$, and there is exactly one substitution changing d^3_{i+2} to a C before time t^*_{i+3} if this is needed in order that $V_{i+3}>0$. Let I_1 and I_2 be the indicator functions for the necessity of a change in codon $i-1$ or codon $i+2$, respectively. We bound (3.13) from below by

$$\frac{1}{Z}\exp(-6\cdot 9\kappa_2 t^0)(\kappa_1 t^*_{i-2})^{I_1}(\kappa_1 t^*_{i+3})^{I_2}. \tag{3.14}$$

Instead of (3.11) we get the upper bound

$$\begin{aligned}\exp\{-6\cdot 3\kappa_1 t^0\}&\frac{(9\kappa_2 t^0)^{\tilde n_i}}{\tilde n_i!}\frac{(9\kappa_2 t^0)^{\tilde n_{i+1}}}{\tilde n_{i+1}!}\\ &\times\frac{(9\kappa_2)^{\tilde n_{i-1}}((t^0)^{\tilde n_{i-1}}-I_1(t^0-t^*_{i-2})^{\tilde n_{i-1}})}{\tilde n_{i-1}!}\\ &\times\frac{(9\kappa_2)^{\tilde n_{i+2}}((t^0)^{\tilde n_{i+2}}-I_2(t^0-t^*_{i+3})^{\tilde n_{i+2}})}{\tilde n_{i+2}!}.\end{aligned}$$

Summing over the possible values of $\tilde n_j$ ($\tilde n_{i-1}\geq 1$ if $I_1=1$) we obtain the bound

$$Z\leq \exp\left\{-6\cdot 3\kappa_1 t^0+4\cdot 9\kappa_2 t^0\right\}\left(1-\exp(-9\kappa_2 t^*_{i-2})\right)^{I_1}\left(1-\exp(-9\kappa_2 t^*_{i+3})\right)^{I_2}. \tag{3.15}$$

Dividing (3.14) by (3.15) we get the lower bound

$$\begin{aligned}&P(A_i\cap A_{i+1}|D_{i-3}(\cdot),D_{i-2}(\cdot),D_{i+3}(\cdot),D_{i+4}(\cdot))\\ &\quad\geq\ \exp(-(90\kappa_2-18\kappa_1)t^0)\left(\frac{\kappa_1 t^*_{i-2}}{1-\exp(-9\kappa_2 t^*_{i-2})}\right)^{I_1}\left(\frac{\kappa_1 t^*_{i+3}}{1-\exp(-9\kappa_2 t^*_{i+3})}\right)^{I_2}\\ &\quad\geq\ \exp(-(90\kappa_2-18\kappa_1)t^0)\left(\frac{\kappa_1 t^0}{1-\exp(-9\kappa_2 t^0)}\right)^{2}.\end{aligned}$$

This proves the result of the lemma. ■

We are now ready to bound $2b_1+2b_2+b_3$ of Theorem 2.1. From Lemma 3.1 and Proposition 2.3 together with Lemma 3.2 we find

$$2b_1+2b_2+b_3\leq nd\left\{4k(n)d\left((p\lambda)^2+\frac{(2p\lambda)^2}{2}\mathrm{e}^{2p\lambda}\right)+2(1-\delta^4)^{(k(n)+1)/(2\tau+1)-1}\right\}$$

with

$$p\lambda=\frac{27\kappa_2^2}{\kappa_1}t^0.$$

If we take $k(n) = 2(2\tau+1)\log(n)/\log(1-\delta^4)$ and assume that $t^0 \leq \kappa_1/(27\kappa_2^2)$ we end up with

$$2b_1 + 2b_2 + b_3 \leq c_1(\delta, \kappa_1, \kappa_2)\log(n)n(t^0)^2 \tag{3.16}$$

for some constant $c_1(\delta, \kappa_1, \kappa_2)$.

When using Theorem 2.1 we also want to take into account the term with $(\min \lambda_i)^{1/2}$. For this we need a lower bound on the success probability $P(U_{il} = 1)$. Here, however, we have the problem that we have not specified completely the sets B_{il} defining the Bernoulli variables U_{il}, and there may be combinations of $D_i(0)$ and B_{il} that makes U_{il} identically zero. Instead we make the following assumption.

A2: The starting sequence $D(0)$ and the sets B_{il} are such that for each $l = 1, \ldots, d$ the number of codons for which U_{il} can become 1 by one substitution only in $D_i(\cdot)$ and no substitutions in $D_{i-1}(\cdot)$ and $D_{i+1}(\cdot)$ is bigger than $c_2 n$ for some constant c_2.

Taking a codon i for which U_{il} can become 1 we can make a lower bound on

$$P(U_{il} = 1 | D_{i-2}(\cdot), D_{i-1}(\cdot) \equiv D_{i-1}(0) D_{i+1}(\cdot) \equiv D_{i+1}(0), D_{i+2}(\cdot))$$

as in (3.10) and (3.12). Instead of (3.10) the lower bound is

$$\frac{1}{Z}\exp(-3 \cdot 9\kappa_2 t^0)(\kappa_1 t^0), \tag{3.17}$$

and the estimation of Z in (3.12) is replaced by

$$Z \leq \exp\{-3 \cdot 3\kappa_1 t^0 + 9\kappa_2 t^0\}. \tag{3.18}$$

Next we use the bound

$$P(D_{i-1}(\cdot) \equiv D_{i-1}(0), D_{i+1}(\cdot) \equiv D_{i+1}(0)) \geq 1 - 2\lambda p, \tag{3.19}$$

obtained from (3.6). We then find

$$EU_{il} \geq (1 - 2\lambda p)\exp(-(36\kappa_2 + 9\kappa_1)t^0)\kappa_1 t^0.$$

For $t^0 \leq \kappa_1/(27\kappa_2^2)$ we can write the lower bound as

$$\lambda_i \geq c_2 n c_3(\kappa_1, \kappa_2) t^0, \tag{3.20}$$

for some constant $c_3(\kappa_1, \kappa_2)$

Combining (3.16) and (3.20) we get the following theorem from Theorem 2.1.

Theorem 3.3 *Let $S_j = \sum_{i=1}^n U_{ij}$, $j = 1, \ldots, d$, and let $t^0 \leq \kappa_1/(27\kappa_2^2)$. Under the assumption A2 the total variation distance between the joint distribution of $(S_1, \ldots, S_d)$ and independent Poisson variables with the same means is bounded by $c(\delta, \kappa_1, \kappa_2) \log(n)(n(t^0)^3)^{1/2}$ for some constant $c(\delta, \kappa_1, \kappa_2)$.*

Theorem 3.3 shows that the multivariate Poisson approximation is valid when $n(t^0)^3$ is small.

3.2 Unconditional result

In this section we let U_{jl} be defined as in (3.3) and as before let $X_i = D_i(\cdot)$. Instead of fixing $D_i(0)$, $i = 1, \ldots, n$ we let the initial sequence have the stationary distribution correponding to the Markov process with intensities given by (3.1). Since the estimates in Lemma 3.1 and in Lemma 3.2 are true for all values of the initial sequence $D_i(0)$, $i = 1, \ldots, n$, we can use the same estimates when we no longer condition on the initial sequence.

For the class of model studied in Jensen and Pedersen (2000) and Pedersen and Jensen (2001) the stationary distribution of a codon sequence can be viewed as a Markov chain along the sequence. We can then use a large deviation result to show that the number of three consequtive codons $(D_{i-1}(0), D_i(0), D_{i+1}(0))$ of a particular type is larger than cn except on a set having an exponentially small probability in n. We therefore obtain a result similar to Theorem 3.3 in the unconditional case.

3.3 A central limit theorem

When t^0 is so large that Theorem 3.3 is no longer applicable we can instead use a normal approximation to the distribution. Establishing a central limit theorem is, however, slightly complicated because we still want to include a limiting situation with $t^0 \to 0$ while $n \to \infty$. The main point seems to be a lower bound on the variance of the sum as $t^0 \to 0$. In order to establish a central limit theorem we will use the proof in Bolthausen (1982).

Let e be a d-dimensional unit vector and let $V_i = \sum_{l=1}^{d} e_l(U_{il} - EU_{il})$. Let m_n be a number to be chosen below and define

$$a_n = \sum_i E\left\{V_i\left(\sum_{j:|j-i|\leq m_n} V_j\right)\right\}.$$

According to Bolthausen (1982) we must prove the following statements,

$$a_n = \text{var}\left(\sum_{i=1}^{n} V_i\right)(1+o(1)), \tag{3.21}$$

$$a_n^{-2} \sum_{i_1,i_2,j_1,j_2:|i_1-j_1|\leq m_n,|i_2-j_2|\leq m_n} \text{cov}\left(V_{i_1}V_{j_1}, V_{i_2}V_{j_2}\right) \to 0, \tag{3.22}$$

$$a_n^{-1/2} \sup_i \sum_{j_1,j_2:|i-j_1|\leq m_n,|i-j_2|\leq m_n} E\left(V_{j_1}V_{j_2}\right) \to 0, \tag{3.23}$$

and

$$a_n^{-1/2} \sum_i E\left|E\left(V_i|V_j, |j-i| > m_n\right)\right| \to 0. \tag{3.24}$$

Lemma 3.4 *Under the assumption A2 and for t^0 bounded there exists a constant c such that*

$$var\left(\sum_{i=1}^{n} V_i\right) \geq cnt^0.$$

Proof. Since e is a unit vector let $|e_\alpha| \geq 1/\sqrt{d}$. Let $\{i_1, \ldots, i_K\}$ be the codons satisfying the condition in assumption A2 for the variables $U_{i\alpha}$. Starting from below ($r = 0$, $i_0 = 0$) we condition on the codon paths of codons $i_r+1, \ldots, i_s-1$, where $s > r$

is the smallest value such that we condition on at least two codon paths in between i_r and i_s. In this way we end up with a reduced set of codons $\{j_1, \ldots, j_L\} \subseteq \{i_1, \ldots, i_K\}$, where $L \geq c_2 n/3$, and where we have conditioned on at least two codon paths at each side of j_l. A lower bound on the variance is obtained as the mean of the conditional variance. This gives the lower bound

$$\sum_{l=1}^{L} E\left\{\text{var}\left(V_{j_l-1} + V_{j_l} + V_{j_l+1} | D_{j_l-2}(\cdot), D_{j_l-1}(\cdot), D_{j_l+1}(\cdot), D_{j_l+2}(\cdot)\right)\right\}. \tag{3.25}$$

To bound the individual terms in the sum we further condition on $D_{j_l-1}(\cdot) \equiv D_{j_l-1}(0)$ and $D_{j_l+1}(\cdot) \equiv D_{j_l+1}(0)$. Then $V_{j_l-1} = 0$, $V_{j_l+1} = 0$, and the possible values of V_{j_l} are $\{0, e_1(1-p_{j_l,1}), \ldots, e_d(1-p_{j_l,d})\}$. The conditional probability that V_{j_l} equals $e_\alpha(1-p_{j_l,\alpha})$ is bounded from below by $\exp(-(36\kappa_2 - 9\kappa_1)t^0)\kappa_1 t^0$ from (3.17) and (3.18). Arguing as in (3.17) and (3.18), but making an upper bound instead of a lower bound, the conditional probability that V_{j_l} equals $e_j(1-p_{j_l,j})$ is bounded from above by $\exp((27\kappa_2 - 6\kappa_1)t^0)\kappa_2 t^0$. Putting these bound together we see that

$$\text{var}\left(V_{j_l} | D_{j_l-2}(\cdot), D_{j_l-1}(\cdot) \equiv D_{j_l-1}(0), D_{j_l+1}(\cdot) \equiv D_{j_l+1}(0), D_{j_l+2}(\cdot)\right) \geq \omega_1 t^0 \tag{3.26}$$

for some constant ω_1. Combining (3.26) with the bound (3.19) each term in (3.25) is bounded from below by $\omega_1 t^0(1-2\lambda p)$ We therefore end up with

$$\text{var}\left(\sum_{i=1}^{n} V_i\right) \geq L\omega_1 t^0(1-2\lambda p) \geq \omega_2 n t^0,$$

where ω_2 is a constant. This then gives the result of the lemma. ■

Theorem 3.5 *If there exist constants c_3, c_4 and $0 \leq \gamma < 1/2$ such that $c_3 n^{-\gamma} \leq t^0 \leq c_4$, and if assumption A2 holds then*

$$(var(S))^{-1/2}(S - ES) \stackrel{\sim}{\to} N_d(0, I),$$

where $S = (S_1, \ldots, S_d)$ is the vector of counts.

Proof. We prove (3.21) to (3.24) by taking $m_n = n^\xi$ with $\gamma+\xi < 1/2$ and $\gamma+4\xi < 1$.

Let δ be given by Lemma 3.2 and define

$$\rho = (1-\delta^4)^{1/(2t+1)},$$

and let ω_i, $i = 1, \ldots,$ be constants. Using the line of argument as in (2.14) and (2.15) we get the bounds

$$|E\left(V_i | V_j, |j-i| > m_n\right)| \leq \omega_1 \rho^{m_n}, \tag{3.27}$$

$$|EV_iV_j| \leq \omega_2 \rho^{|i-j|}, \tag{3.28}$$

and for $|i_1 - j_1| \leq m_n$, $|i_2 - j_2| \leq m_n$, $|i_1 - i_2| > 2m_n$,

$$|\mathrm{cov}\left(V_{i_1}V_{j_1}, V_{i_2}V_{j_2}\right)| \leq \omega_3 \rho^{|i_1-i_2|-2m_n}, \tag{3.29}$$

and for arbritrary i_1, i_2

$$\begin{aligned} |\mathrm{cov}\left(V_{i_1}V_{j_1}, V_{i_2}V_{j_2}\right)| &\leq |E\left(V_{i_1}V_{j_1}V_{i_2}V_{j_2}\right)| + |E(V_{i_1}V_{j_1})E(V_{i_2}V_{j_2})| \\ &\leq \omega_4 \rho^{\min\{|i_1-j_1|, |i_1-i_2|, |i_1-j_2|\}}. \end{aligned} \tag{3.30}$$

To prove (3.21) we use (3.28) to write

$$\left|\mathrm{var}\left(\sum_{i-1}^{n} V_i\right) - a_n\right| \leq 2\omega_2 n \sum_{j=m_n}^{\infty} \rho^j = nt^0 \frac{2\omega_2}{1-\rho}\frac{1}{t^0}\rho^{m_n} = nt^0 o(1),$$

which from Lemma 3.4 proves (3.21). In particular, we then have

$$a_n \geq \omega_5 n t^0. \tag{3.31}$$

Turning to (3.22) we use (3.29) and (3.30) to get the bound

$$\begin{aligned} & a_n^{-2}\omega_6 n \left\{(2m_n+1)^2 \sum_{k=3m_n}^{\infty} \rho^{k-2m_n} + (6m_n)(2m_n+1)\sum_{j=0}^{3m_n}\rho^j\right\} \\ = \;& a_n^{-2}\frac{\omega_6 n}{1-\rho}\left\{(2m_n+1)^2\rho^{m_n} + 6m_n(2m_n+1)\right\} \to 0, \end{aligned}$$

where the convergence follows from $n^{2\gamma+2\xi-1} \to 0$. For (3.23) we have trivially the bound

$$\omega_7 a_n^{-1/2}(2m_n+1)^2 \to 0,$$

where the convergence follows from $n^{\gamma/2+2\xi-1/2} \to 0$. Finally, for (3.24) we get directly from (3.27) the bound

$$\omega_1 a_n^{-1/2} n\rho^{m_n} \to 0.$$

■

We note that the two theorems 3.3 and 3.5 overlap in the sence that the former can be used with $t^0 = n^{-\gamma}$ for $\gamma > 1/3$ and the latter can be used when $\gamma < 1/2$.

4 An example

To illustrate the result of Theorem 3.3 we simulate a model of the form described in Section 3. We consider a situation where reading frames I and II are used. A codon $D = (d^1, d^2, d^3)$ translates into an aminoacid $am(D)$. There are 64 codons, but only 20 different aminoacids. When we substitute z for d^r we can either have that the aminoacid $am(D)$ is unchanged, to be denoted by S (the substitution is *synonymous*), or the aminoacid $am(D)$ is changed, to be denoted by N (the substitution is *nonsynonymous*). When we have two reading frames we write $S(I)$ or $N(I)$ for what happens in reading frame I and $S(II)$ or $N(II)$ for what happens in reading frame II. Thus we let $1_{N(I),S(II)}$ be one if the substitution is synonymous in reading frame II and nonsynonymous in reading frame I. If both of z and d^r belong to $\{A, G\}$ or $\{C, T\}$ we speak of a transition and let 1_{ts} be one. Finally, we let $1_{nostop(I)}$ be one if the substitution of d^r by z does not produce a stop codon in reading frame I and similarly with $1_{nostop(II)}$. We will consider rates (3.1) of the form

$$\lambda(r, z|D_{i-1}, D_i, D_{i+1}) = \pi(z)\theta^{1_{ts}} f_I^{1_{N(I),S(II)}} f_{II}^{1_{S(I),N(II)}} f_{I/II}^{1_{N(I),N(II)}} \gamma^K 1_{nostop(I)} 1_{nostop(II)}, \tag{4.1}$$

where K is the number of times G follows after a C in the sequence $(d_{i-1}^3, d_i^1, d_i^2, d_i^3, d_{i+1}^1)$ minus the same number for the sequence with d_i^r replaced by z. Here $\theta \geq 1$, $f_I, f_{II}, f_{I/II}$ are all less than or equal to one, and $\pi(A) + \pi(G) + \pi(C) + \pi(T) = 1$.

With these rates there are several terms contributing to the interaction among the codons. The terms involving $f_I, f_{II}, f_{I/II}$ together with the term $1_{nostop(II)}$ all involve two codons. The term γ^K is introduced to model cases where there is a suppression of CG-dinucleotide pairs. Typically, this suppression will also be present across codon boundaries again introducing interaction.

The stationary distribution for a sequence evolving according to the rates in (4.1) can be found in Pedersen and Jensen (2001). The stationary distribution can be viewed as a Markov chain along the sequence and it is therefore easy to simulate a realization from the stationary distribution. For a fixed realization from the stationary distribution, $(D_1(0), \ldots, D_n(0))$, we have simulated the evolution of the DNA sequence 500000 times. As an illustration we have for each simulated value $(D_1(t), \ldots, D_n(t))$ counted the number of codons i for which $D_i(0)$ has been changed in one position, either position 1 or 3, the substitution is a transition, and the substitution is synonymous in both reading frames (the count being N_1) or the substitution is synonymous in one reading frame and nonsynonymous in the other reading frame (the count being N_2). We have performed two simulations that differ in the value t^0 of the evolutionary time:

$$t^0 = 0.1 \quad \text{and} \quad t^0 = 0.5.$$

The remaining parameters have been fixed at

$$n = 400, \quad f_I = f_{II} = 0.2, \quad f_{I/II} = f_I f_{II} = 0.04, \quad \theta = 2, \quad \gamma = \sqrt{0.2}, \quad \pi(\cdot) = 0.25.$$

Let us first consider the marginal distribution of N_1, the number of changes being synonymous in both reading frames. In Table 1 is a comparison of the simulated distribution with a Poisson distribution with the same mean as the simulated dis-

	$t^0 = 0.1$		$t^0 = 0.5$	
n	$P(N_1 = n)$	Pois(n)	$P(N_1 = n)$	Pois(n)
0	41.07 ± 0.14	41.91	3.38 ± 0.06	4.52
1	37.39 ± 0.14	36.45	12.56 ± 0.10	13.99
2	16.18 ± 0.10	15.85	22.10 ± 0.12	21.67
3	4.37 ± 0.06	4.60	24.34 ± 0.12	22.37
4	0.85 ± 0.03	1.00	18.99 ± 0.11	17.32
5	0.12 ± 0.012	0.17	11.15 ± 0.09	10.73
6	0.014 ± 0.005	0.025	5.01 ± 0.06	5.54
7			1.80 ± 0.04	2.45
8			0.52 ± 0.03	0.95
9			0.13 ± 0.02	0.33
10			0.023 ± 0.009	0.101

Table 1: Comparison of simulated distribution of N_1 with a Poisson distribution with the same mean. All entries are percentage values.

tribution. As can be seen for $t^0 = 0.1$ the true distribution is very close to the Poisson approximation whereas for $t^0 = 0.5$ the discrepancies are bigger with the Poisson approximation giving bigger probabilities for small values of N_1 and smaller probabilities for large values of N_1.

We next consider the conditional distribution of N_2, the number of changes being synonymous in one reading frame only, given the value of N_1. These are given in Figure 1. For $t^0 = 0.1$ the Poisson approximation is very accurate and N_1 and N_2 are close to being independent. Contrary to this we can for $t^0 = 0.5$ see a deviation from independence as well as a deviation from the Poisson approximation.

Let us conclude by illustrating the use of the normal approximation from Theorem 3.5. In fact we will go one step further and include the first correction term of an Edgeworth expansion. Thus from the simulations we estimate the mean $\mu(k)$, the variance $\sigma^2(k)$, and the third standardized moment $\kappa(k)$ of the distribution of N_2

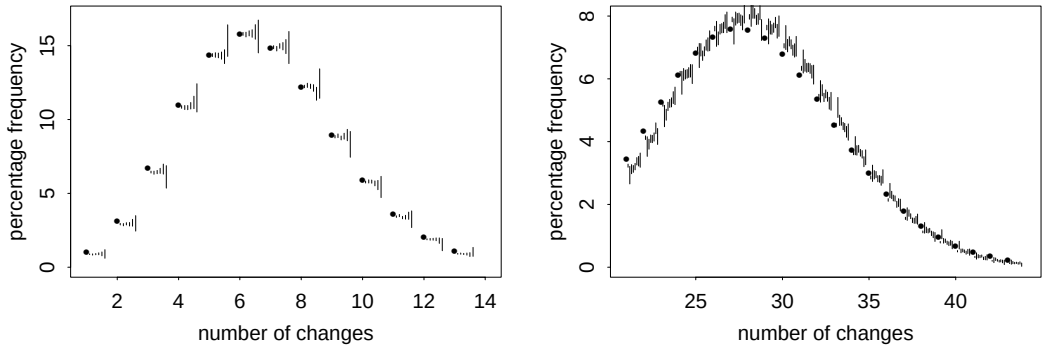

Figure 1: Comparison of conditional distribution of N_2 given N_1 with a Poisson approximation. The left figure is $t^0 = 0.1$ and the right figure is $t^0 = 0.5$. The $\bullet$ is the Poisson approximation and the vertical bars are the simulated probabilities plus minus two standard deviations. The leftmost bar is the marginal distribution of N_2 and moving right follows the conditional distributions of N_2 given $N_1 = n$ with $n = 0, 1, 2, 3, 4$ for $t^0 = 0.1$ and $n = 0, 1, \ldots, 6$ for $t^0 = 0.5$.

given that $N_1 = k$. We then approximate the conditional distribution by

$$P(N_2 = m | N_1 = k) \approx \frac{1}{\sigma(k)}\{1 - \frac{\kappa}{6}(3x(m) - x(m)^3)\}\frac{\mathrm{e}^{-x(m)^2/2}}{\sqrt{2\pi\sigma(k)^2}}, \qquad (4.2)$$

$$x(m) = \frac{m - \mu(k)}{\sigma(k)}.$$

For the case considered above with $t^0 = 0.5$ this gives a very good approximation. In Figure 2 is a typical example where we have taken $k = 0$.

5 References

Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.* **17**, 9-25.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376.

Jensen, J.L. and Pedersen, A-M.K. (2000). Probabilistic models of DNA sequence

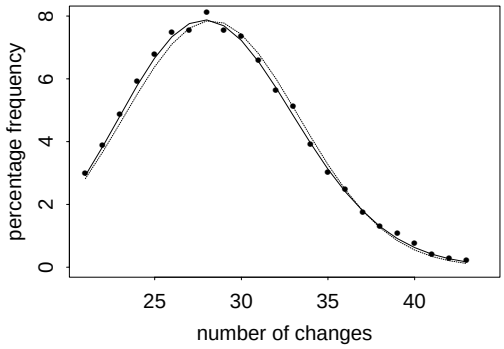

Figure 2: Comparison of the conditional distribution of N_2 given $N_1 = 0$ with a normal distribution for the case $t^0 = 0.5$. The dots are the simulated probabilities, the full drawn line is the Edgeworth approximation (4.2), and the dotted line is the normal approximation.

evolution with context dependent rates of substitution. *Adv. Appl. Probab.* **32**, 499-517.

Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150-174.

Pedersen, A-M.K. and Jensen, J.L. (2001). A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* **18**, 763-776.