
Probabilistic models for biological sequences: selection and Maximum Likelihood estimation

Svetlana Ekisheva

School of Biology, Georgia Institute of Technology,
Atlanta, GA 30332-0230, USA
E-mail: sveta.ekisheva@bme.gatech.edu

Mark Borodovsky*

Department of Biomedical Engineering, School of Biology,
Georgia Institute of Technology,
Atlanta, GA 30332-0230, USA
E-mail: mark.borodovsky@biology.gatech.edu
*Corresponding author

Abstract: Probabilistic models for biological sequences (DNA and proteins) are frequently used in bioinformatics. We describe statistical tests designed to detect the order of dependency among elements of the sequence and to select the most appropriate probabilistic model for an experimental biological sequence. For a model of given type, the independence model, the first-order Markov chain and the hidden Markov model (HMM), we derive the uniform lower bound for the rate of decay for the errors of the maximum likelihood (ML) estimates of the model parameters and, subsequently, the uniform confidence intervals for the parameters.

Keywords: statistical models for biological sequences; parameter estimation; Maximum Likelihood (ML) estimates; asymptotic properties of Maximum Likelihood (ML) estimates.

Reference to this paper should be made as follows: Ekisheva, S. and Borodovsky, M. (2006) 'Probabilistic models for biological sequences: selection and Maximum Likelihood estimation', *Int. J. Bioinformatics Research and Applications*, Vol. 2, No. 3, pp.305–324.

Biographical notes: Svetlana Ekisheva is a Research Scientist at the School of Biology at Georgia Tech, Atlanta, GA. She received a PhD Degree in Physics and Mathematics from the Saint-Petersburg State University, Russia. Her research interests are in bioinformatics, applied statistics and stochastic processes. Her expertise includes teaching of probability theory and statistics at Universities in Russia and the USA.

Mark Borodovsky is the Regents' Professor at the School of Biology, Georgia Tech and the Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University. He is also a Director of the Center for Bioinformatics and Computational Biology at Georgia Tech. He received a PhD in Physics and Mathematics from the Institute of Physics and Technology in Moscow, Russia. He was a recipient of the NIH Shannon Award in 1992. His research interests include development of probabilistic methods for computational genomics and systems biology.

1 Introduction

Many bioinformatics studies use probabilistic models that generate sequences with statistical properties close to those observed in empirical biological sequences (e.g., Gatlin, 1972; Almagor, 1983; Fitch, 1983; Borodovsky et al., 1986a, 1986b; Churchill, 1989; Tavare and Song, 1989; Cowan, 1991; Karlin and Macken, 1991; Karlin et al., 1992; Durbin et al., 1998; Reinert et al., 2000). The initial step of model selection is typically followed by the parameter estimation step based on the classical statistical theory.

Several alternative probabilistic models can be used to describe a sequence $x = (x_1, \dots, x_N)$ with symbols from a finite alphabet \mathcal{A} ; for biological sequences $|\mathcal{A}| = 4$ (DNA) or $|\mathcal{A}| = 20$ (proteins).

The simplest model, a homogeneous independence model M , assumes that the occurrences of symbols at different sequence positions are independent and at any position the occurrence of symbol α , $\alpha \in \mathcal{A}$, has probability

$$p_\alpha, \sum_{\alpha \in \mathcal{A}} p_\alpha = 1.$$

Assuming that an observed sequence x was generated by model M , the ML estimates of unknown parameters p_α are given by ratios

$$\hat{p}_\alpha = \frac{N(\alpha)}{N}, \quad (1)$$

where $N(\alpha)$ is the number of symbols α observed in sequence x .

A more general model of homogeneous type is the stationary ergodic m -order Markov chain $M(m)$ with transition probabilities

$$p_{\alpha_1, \dots, \alpha_m, \alpha_{m+1}} = P(x_i = \alpha_{m+1} \mid x_{i-1} = \alpha_m, \dots, x_{i-m} = \alpha_1),$$

$\alpha_k \in \mathcal{A}$, $i = m+1, \dots, N$, and a stationary distribution π . Obviously, the independence model is the Markov chain of order $m = 0$, $M = M(0)$. For a sequence x generated by the first order Markov model $M(1)$ the ML estimates of the transition probabilities $p_{\alpha\beta}$, $\alpha, \beta \in \mathcal{A}$, are delivered by formulas

$$\hat{p}_{\alpha\beta} = \frac{N(\alpha\beta)}{N(\alpha\bullet)}. \quad (2)$$

Here $N(\alpha\beta)$ is the number of occurrences of symbols (α, β) in adjacent positions of sequence x and $N(\alpha\bullet) = \sum_{\gamma \in \mathcal{A}} N(\alpha\gamma)$. Note that $N(\alpha\bullet) = N(\alpha)$ if $x_N \neq \alpha$ and $N(\alpha\bullet) = N(\alpha) - 1$ otherwise. Similarly, for a sequence x generated by the model $M(m)$ the ML estimates of the transition probabilities $p_{\alpha_1, \dots, \alpha_m, \alpha_{m+1}}$ are defined by formulas

$$\hat{p}_{\alpha_1, \dots, \alpha_m, \alpha_{m+1}} = \frac{N(\alpha_1 \dots \alpha_m \alpha_{m+1})}{N(\alpha_1 \dots \alpha_m \bullet)}.$$

Here $N(\alpha_1 \dots \alpha_k)$ designates the number of occurrences of k -letter word $(\alpha_1 \dots \alpha_k)$ in the sequence x and $N(\alpha_1 \dots \alpha_m \bullet) = \sum_{\gamma \in \mathcal{A}} N(\alpha_1 \dots \alpha_m \gamma)$.

2 Model selection

To briefly describe approaches to the model selection, we will follow arguments and notations used by Reinert et al. (2000). To identify the model type that would be the most appropriate for an anonymous sequence x , the χ^2 -test could be used. To determine whether the model M fits the sequence x , the null hypothesis of independence

$$H_0 : P(x_i = \alpha, x_{i+1} = \beta) = p_\alpha p_\beta, \quad i = 1, \dots, N-1, \quad \alpha, \beta \in \mathcal{A},$$

must be tested vs. the alternative hypothesis H_a which states that the probabilities $P(x_i = \alpha, x_{i+1} = \beta)$ do not depend on i , the only restriction. Under H_0 the ML estimate of $P(x_i = \alpha, x_{i+1} = \beta)$, $i = 1, \dots, N-1$, is

$$\hat{P}_{H_0}(\alpha, \beta) = \frac{N(\alpha \bullet) N(\bullet \beta)}{N-1 \quad N-1}.$$

Here $N(\bullet \beta) = \sum_{\gamma \in \mathcal{A}} N(\gamma \beta)$. Note that $(N(\alpha \bullet)/(N-1))(N(\bullet \beta)/(N-1)) = \hat{p}_\alpha \hat{p}_\beta$, where \hat{p}_α is the ML estimate of p_α based on sequence x_1, \dots, x_{N-1} , and \hat{p}_β is the ML estimate of p_β based on sequence x_2, \dots, x_N . Under the alternative hypothesis, the ML estimate of $P(x_i = \alpha, x_{i+1} = \beta)$, $i = 1, \dots, N-1$, is

$$\hat{P}_{H_a}(\alpha, \beta) = \frac{N(\alpha \beta)}{N-1}.$$

Then formula

$$\begin{aligned} \chi^2 &= \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \frac{((N-1)\hat{P}_{H_a}(\alpha, \beta) - (N-1)\hat{P}_{H_0}(\alpha, \beta))^2}{(N-1)\hat{P}_{H_0}(\alpha, \beta)} \\ &= \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \frac{((N-1)N(\alpha \beta) - N(\alpha \bullet)N(\bullet \beta))^2}{(N-1)N(\alpha \bullet)N(\bullet \beta)} \end{aligned}$$

defines the Pearson χ^2 -statistic which under H_0 follows asymptotically a χ^2 -distribution with $(|\mathcal{A}| - 1)^2$ degrees of freedom ($|\mathcal{A}|$ stands for the size of the alphabet). Therefore, H_0 is rejected if the sample value of χ^2 -statistic is larger than a critical value of the χ^2 -distribution corresponding to a specified significance level (false negative error rate). In application of this test to a DNA sequence, we have to consider the χ^2 -distribution with nine degrees of freedom.

If H_0 cannot be rejected at a predefined significance level (say, 5%), then the independence test allows to conclude that the model M fits the observed sequence x . Otherwise, if the null hypothesis does get rejected, a higher-order dependence should be tested. At the next step, for the first-order Markov chain $M(1)$ the null hypothesis can be formally stated as follows:

$$\begin{aligned}
H_0 : P(x_i = \alpha, x_{i+1} = \beta, x_{i+2} = \gamma) \\
&= P(x_i = \alpha)P(x_{i+1} = \beta | x_i = \alpha)P(x_{i+2} = \gamma | x_i = \alpha, x_{i+1} = \beta) \\
&= P(x_i = \alpha)p_{\alpha, \beta}p_{\beta, \gamma} = \sum_{x_1, x_2, \dots, x_{i-1}} \pi(x_1)p_{x_1, x_2} \times \dots \times p_{x_{i-1}, \alpha}p_{\alpha, \beta}p_{\beta, \gamma} \\
&= \pi(\alpha)p_{\alpha, \beta}p_{\beta, \gamma},
\end{aligned}$$

for any $i = 1, \dots, N-2, \alpha, \beta, \gamma \in \mathcal{A}$. Here π is a stationary distribution of the Markov chain $M(1)$. This null hypothesis should be tested against the alternative hypothesis H_a which assumes that probabilities $P(x_i = \alpha, x_{i+1} = \beta, x_{i+2} = \gamma)$ do not depend on i , the only restriction.

Under H_0 the ML estimate of $P(x_i = \alpha, x_{i+1} = \beta, x_{i+2} = \gamma), i = 1, \dots, N-2$, is

$$\hat{P}_{H_0}(\alpha, \beta, \gamma) = \hat{\pi}(\alpha)\hat{p}_{\alpha, \beta}\hat{p}_{\beta, \gamma} = \frac{N(\alpha \bullet \bullet)}{N-2} \frac{N(\alpha \beta \bullet)}{N(\alpha \bullet \bullet)} \frac{N(\bullet \beta \gamma)}{N(\bullet \beta \bullet)} = \frac{N(\alpha \beta \bullet)}{N-2} \frac{N(\bullet \beta \gamma)}{N(\bullet \beta \bullet)},$$

where $N(\alpha \beta \bullet)$ is the number of occurrences of three-symbol words starting with (α, β) ;

$$N(\alpha \bullet \bullet) = \sum_{\beta \in \mathcal{A}} N(\alpha \beta \bullet);$$

$N(\bullet \beta \gamma)$ is the number of occurrences of three-symbol words ending with symbols (β, γ) ;

$$N(\bullet \beta \bullet) = \sum_{\gamma \in \mathcal{A}} N(\bullet \beta \gamma).$$

Note that $\hat{\pi}(\alpha) = (N(\alpha \bullet \bullet)/(N-2))$ is the ML estimate of the stationary probability $\pi(\alpha)$ based on sequence x_1, \dots, x_{N-2} , $\hat{p}_{\alpha, \beta} = (N(\alpha \beta \bullet))/(N(\alpha \bullet \bullet))$ is the ML estimate of the transition probability $p_{\alpha, \beta}$ from sequence x_1, \dots, x_{N-1} , $\hat{p}_{\beta, \gamma} = (N(\bullet \beta \gamma))/(N(\bullet \beta \bullet))$ is the ML estimate from sequence x_2, \dots, x_N .

Under H_a the ML estimate of $P(x_i = \alpha, x_{i+1} = \beta, x_{i+2} = \gamma), i = 1, \dots, N-2$, is

$$\hat{P}_{H_a}(\alpha, \beta, \gamma) = \frac{N(\alpha \beta \gamma)}{N-2}.$$

Then under H_0 the Pearson χ^2 -statistic defined by formula

$$\begin{aligned}
\chi^2 &= \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \sum_{\gamma \in \mathcal{A}} \frac{((N-2)\hat{P}_{H_a}(\alpha, \beta, \gamma) - (N-2)\hat{P}_{H_0}(\alpha, \beta, \gamma))^2}{(N-2)\hat{P}_{H_0}(\alpha, \beta, \gamma)} \\
&= \sum_{\alpha \in \mathcal{A}} \sum_{\beta \in \mathcal{A}} \sum_{\gamma \in \mathcal{A}} \frac{(N(\alpha \beta \gamma)N(\bullet \beta \bullet) - N(\alpha \beta \bullet)N(\bullet \beta \gamma))^2}{N(\alpha \beta \bullet)N(\bullet \beta \gamma)N(\bullet \beta \bullet)}
\end{aligned}$$

asymptotically has a χ^2 -distribution with $(|\mathcal{A}|^2 - 1)(|\mathcal{A}| - 1) - l$ degrees of freedom. Here l is the number of triplets $(\alpha, \beta, \gamma), \alpha, \beta, \gamma \in \mathcal{A}$, such that the number of expected occurrences of triplet (α, β, γ) in the sequence is equal to zero (thus the corresponding l terms

$$\frac{((N-2)\hat{P}_{H_a}(\alpha, \beta, \gamma) - (N-2)\hat{P}_{H_0}(\alpha, \beta, \gamma))^2}{(N-2)\hat{P}_{H_0}(\alpha, \beta, \gamma)}$$

cannot be properly defined and they will be absent from the sum for χ^2 -statistic). Upon applying this test for a DNA sequence one accepts H_0 at a specified significance level if the observed value of the χ^2 -statistic is smaller than the critical value of the χ^2 -distribution with $45 - l$ degrees of freedom. If H_0 is rejected, the test for a higher-order Markov chain should be carried out in analogous way.

3 Uniform accuracy of the ML estimates and the rate of convergence in probability of the estimates to the model parameters

3.1 Independence model

Independence model assumes that occurrences of symbols at different sequence positions of an empirical sequence $x = (x_1, \dots, x_N)$ are independent of each other. Formally, the model is defined by a sequence $X = (X_1, \dots, X_N)$ of independent identically distributed (i.i.d.) random variables $X_i, i = 1, \dots, N$, such that $P(X_i = \alpha) = p_\alpha, \alpha \in \mathcal{A}, \sum_{\alpha \in \mathcal{A}} p_\alpha = 1$.

Therefore, the empirical sequence x becomes a realisation of random vector X (equivalently, we say that x is generated by the independence model M). Note that the vector of numbers of occurrences of symbols of different types in sequence x has the multinomial distribution.

Estimation of the unknown parameters $p_\alpha, \alpha \in \mathcal{A}$, by the ML approach raises a natural question: How close are these estimates to the unknown true values of parameters? Obviously, the ML estimates \hat{p}_α are unbiased:

$$\mathbf{E} \hat{p}_\alpha = \mathbf{E} \frac{N(\alpha)}{N} = \frac{Np_\alpha}{N} = p_\alpha.$$

The consistency ($\hat{p}_\alpha \xrightarrow{P} p_\alpha$) and the strong consistency ($\hat{p}_\alpha \rightarrow p_\alpha$ with probability 1) of the estimates \hat{p}_α follow from the law of large numbers and the strong law of large numbers, respectively. These properties indicate that ‘accuracy’ of the estimate $\hat{p}_\alpha, \alpha \in \mathcal{A}$, increases with the sample size N . Further in the text we assume that $\mathcal{A} = \{\alpha_1, \dots, \alpha_k\}, p_i = p_{\alpha_i}, i = 1, \dots, k$.

In practical applications probabilistic models are frequently used to compute probability of ‘data given model’ (the likelihood of the model). In the independence case, the probability of sequence fragment $x = (x_1, \dots, x_L)$ of the fixed length L is calculated as the product of probabilities of $x_i, i = 1, \dots, L$. The parameter estimates have to closely and in a uniform fashion approximate the true model parameters to provide accurate values for these products. To explore the uniform error rate of the ML estimates, we will study the asymptotic behaviour of $P(\max_{i=1, \dots, k} |\hat{p}_i - p_i| \leq \varepsilon)$ as the length of sequence x increases. The uniform bounds for the convergence rate of the ML estimates derived below could be translated to the bounds for the errors of computations of the likelihoods.

More formally, if for computations of the probability of sequence fragment (x_1, \dots, x_L) we use the ML estimates of parameters of the independence model instead of the unknown true parameter values, the approximation error would be:

$$\begin{aligned} \Delta &= \left| P_{(x_1, \dots, x_L)} - \hat{P}(x_1, \dots, x_L) \right| = \left| p_{x_1} \times \dots \times p_{x_L} - \hat{p}_{x_1} \times \dots \times \hat{p}_{x_L} \right| \\ &= \left| p_{x_1} \times \dots \times p_{x_L} - (p_{x_1} + (\hat{p}_{x_1} - p_{x_1})) \times \dots \times (p_{x_L} + (\hat{p}_{x_L} - p_{x_L})) \right| \\ &\leq \sum_{i=1}^L \binom{L}{i} \varepsilon^{i-1} (\max p_j)^{L-i} = \varepsilon L \sum_{l=0}^{L-1} \binom{L-1}{l} \frac{\varepsilon^l (\max p_j)^{(L-1)-l}}{l+1} \\ &\leq \varepsilon L (\max p_j + \varepsilon)^{L-1}. \end{aligned} \quad (3)$$

Therefore, we can achieve desirably small approximation error Δ by choosing sufficiently small ε and selecting the set of empirical realisations (the set of training sequences) where the condition $\max_{i=1, \dots, k} |\hat{p}_i - p_i| \leq \varepsilon$ is satisfied. The same arguments remain true for the Markov model case with obvious substitution of parameters p_i by transition probabilities p_{ij} when calculating the ‘data given model’ probabilities and assuming that $\max_{i,j=1, \dots, k} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon$.

To get an explicit result on the uniform closeness of \hat{p}_α to p_α , we will use the asymptotic normality property of the multidimensional ML estimator $\hat{\Theta} = (\hat{\theta}_1, \dots, \hat{\theta}_l)$ of a vector of parameters $\Theta = (\theta_1, \dots, \theta_l)$ (Cox and Hinkley, 1974; Section 9.2):

$$(\hat{\Theta} - \Theta) \sqrt{N} \Rightarrow \Phi_{0, I^{-1}(\Theta)}. \quad (4)$$

Here sign \Rightarrow designates a weak convergence, $\Phi_{0, I^{-1}(\Theta)}$ is a centered Gaussian vector with a covariance matrix $I^{-1}(\Theta)$, and $I(\Theta)$ is the Fisher information matrix of order $l \times l$ with elements

$$I_{i,j}(\Theta) = \mathbf{E}_\Theta \left(- \frac{\partial^2 \ln f_{X_1}(x_1, \Theta)}{\partial \theta_i \partial \theta_j} \right),$$

where $f_{X_1}(x_1, \Theta)$ stands for the likelihood function of the random variable X_1 (a single component of vector X). The asymptotic normality (4) holds if the likelihood function $f_X(x, \Theta)$ of random vector $X = (X_1, \dots, X_N)$ satisfies the regularity conditions (Cox and Hinkley, 1974; Section 9.1, conditions (a)–(d)).

It turns out that for the vector of parameters $\mathbf{P}' = (p_1, \dots, p_k)$ the likelihood function $f_X(x, \mathbf{P}')$ does not satisfy the regularity conditions, since

$$\mathbf{E}_{\mathbf{P}'} \frac{\partial \ln f_{X_1}(x_1, \mathbf{P}')}{\partial p_i} \frac{\partial \ln f_{X_1}(x_1, \mathbf{P}')}{\partial p_j} \neq \mathbf{E}_{\mathbf{P}'} \left(- \frac{\partial \ln f_{X_1}(x_1, \mathbf{P}')}{\partial p_i \partial p_j} \right)$$

for all $i, j = 1, \dots, k$. On the other hand, for the vector of parameters $\mathbf{P} = (p_1, \dots, p_{k-1})$ the likelihood function $f_X(x, \mathbf{P})$ does satisfy the regularity conditions; therefore, the asymptotical normality (4) holds for the ML estimator $\hat{\mathbf{P}} = (\hat{p}_1, \dots, \hat{p}_{k-1})$ of the unknown

vector of parameters $\mathbf{P} = (p_1, \dots, p_{k-1})$. The asymptotic normality of $\hat{\mathbf{P}} = (\hat{p}_1, \dots, \hat{p}_{k-1})$ helps derive the lower bound for the probability $P(\max_{i,j=1,\dots,k} |\hat{p}_i - p_i| \leq \varepsilon)$, since

$$\hat{p}_k = 1 - \sum_{i=1}^{k-1} \hat{p}_i.$$

For the log-likelihood function of a single component X_1 of the random vector X we have:

$$\ln f_{X_1}(x_1, \mathbf{P}) = \sum_{i=1}^{k-1} v_i \ln p_i + v_k \ln \left(1 - \sum_{i=1}^{k-1} p_i \right),$$

where v_i is the indicator of occurrence of symbol α_i in the first position of sequence X . Now we calculate the elements of the Fisher information matrix assuming that all probabilities p_i , $i = 1, \dots, k$, are positive:

$$I_{ii}(\mathbf{P}) = \mathbf{E}_{\mathbf{P}} \left(-\frac{v_i}{p_i^2} - \frac{v_k}{\left(1 - \sum_{i=1}^{k-1} p_i\right)^2} \right) = \frac{p_i}{p_i^2} + \frac{p_k}{\left(1 - \sum_{i=1}^{k-1} p_i\right)^2} = \frac{1}{p_i} + \frac{1}{p_k};$$

$$I_{ij}(\mathbf{P}) = \mathbf{E}_{\mathbf{P}} \left(-\frac{v_k}{\left(1 - \sum_{i=1}^{k-1} p_i\right)^2} \right) = \frac{p_k}{p_k^2} = \frac{1}{p_k},$$

$i \neq j$. The property (4) of asymptotical normality implies that a random vector $Y = (\hat{\mathbf{P}} - \mathbf{P})\sqrt{N}$ with components $Y_i = (\hat{p}_i - p_i)\sqrt{N}$, $i = 1, \dots, k-1$, is asymptotically the centered Gaussian vector with the covariance matrix $R = (r_{ij})$, $i, j = 1, \dots, k-1$, obtained by inversion from the information matrix I :

$$r_{ii} = p_i - p_i^2, \quad r_{ij} = -p_i p_j, \quad i \neq j.$$

The same arguments can be applied to random vector Y' with components $Y_i = (\hat{p}_i - p_i)\sqrt{N}$, $i = 2, \dots, k$. Therefore, random variables

$$Z_1 = \frac{Y_1}{\sqrt{\text{Var}(Y_1)}}, \dots, Z_k = \frac{Y_k}{\sqrt{\text{Var}(Y_k)}}$$

are asymptotically standard normal variables with covariances

$$\begin{aligned} \text{cov}(Z_i, Z_j) &= \frac{r_{i,j}}{\sqrt{\text{Var}(Y_i) \text{Var}(Y_j)}} = -\frac{p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} \\ &= -\frac{\sqrt{p_i p_j}}{\sqrt{(1-p_i)(1-p_j)}}, \end{aligned} \tag{5}$$

$i, j = 1, \dots, k$, $i \neq j$. Next we apply the following inequality by Li and Shao (2002), Corollary 2.1:

$$\begin{aligned} & \left| P(Z_1 \leq u_1, \dots, Z_k \leq u_k) - \prod_{i=1}^k P(Z_i \leq u_i) \right| \\ & \leq \frac{1}{4} \sum_{1 \leq i < j \leq k} |\text{cov}(Z_i, Z_j)| \exp\left(-\frac{u_i^2 + u_j^2}{2(1 + |\text{cov}(Z_i, Z_j)|)}\right) \end{aligned} \quad (6)$$

valid for any real numbers u_1, \dots, u_k .

Coming back to our main goal, the estimation of the probability \tilde{P} , for $\varepsilon > 0$, we have

$$\begin{aligned} \tilde{P} &= P(\max_{i=1, \dots, k} |\hat{p}_i - p_i| \leq \varepsilon) = P(|\hat{p}_1 - p_1| \leq \varepsilon, |\hat{p}_2 - p_2| \leq \varepsilon, \dots, |\hat{p}_k - p_k| \leq \varepsilon) \\ &= P\left(|Z_1| \leq \frac{\varepsilon\sqrt{N}}{\sqrt{p_1(1-p_1)}}, \dots, |Z_k| \leq \frac{\varepsilon\sqrt{N}}{\sqrt{p_k(1-p_k)}}\right) \\ &\geq P(\max_i |Z_i| \leq 2\varepsilon\sqrt{N}). \end{aligned} \quad (7)$$

The last probability in equation (7) can be expressed in terms of the joint cumulative distribution function of Z_1, \dots, Z_k (for proof see Appendix 1):

$$\begin{aligned} & P(|Z_1| \leq 2\varepsilon\sqrt{N}, \dots, |Z_k| \leq 2\varepsilon\sqrt{N}) \\ &= \sum_{l=0}^k (-1)^l \sum_{\substack{\{i_1, \dots, i_l\} \subset \{1, \dots, k\}, \\ \{i_1, \dots, i_l\} \cup \{i_{l+1}, \dots, i_k\} = \{1, \dots, k\}}} P(Z_{i_1} \leq -2\varepsilon\sqrt{N}, \dots, \\ & \quad Z_{i_l} \leq -2\varepsilon\sqrt{N}, Z_{i_{l+1}} \leq 2\varepsilon\sqrt{N}, \dots, Z_{i_k} \leq 2\varepsilon\sqrt{N}). \end{aligned} \quad (8)$$

Then we subtract and add to the right side of equation (8) the term

$$\begin{aligned} \prod_{i=1}^k (2\Phi(2\varepsilon\sqrt{N}) - 1) &= \sum_{l=0}^k (-1)^l \sum_{\substack{\{i_1, \dots, i_l\} \subset \{1, \dots, k\}, \\ \{i_1, \dots, i_l\} \cup \{i_{l+1}, \dots, i_k\} = \{1, \dots, k\}}} \\ & \prod_{j=1}^l P(Z_{i_j} \leq -2\varepsilon\sqrt{N}) \prod_{j=l+1}^k P(Z_{i_j} \leq -2\varepsilon\sqrt{N}), \end{aligned} \quad (9)$$

and apply inequality (6) to each difference

$$\begin{aligned} \delta_{i_1, \dots, i_l} &= P(Z_{i_1} \leq -2\varepsilon\sqrt{N}, \dots, Z_{i_l} \leq -2\varepsilon\sqrt{N}, Z_{i_{l+1}} \leq -2\varepsilon\sqrt{N}, \dots, Z_{i_k} \leq -2\varepsilon\sqrt{N}) \\ & \quad - \prod_{j=1}^l P(Z_{i_j} \leq -2\varepsilon\sqrt{N}) \prod_{j=l+1}^k P(Z_{i_j} \leq -2\varepsilon\sqrt{N}). \end{aligned}$$

For any choice of indices i_1, \dots, i_l and sufficiently large N we have

$$|\delta_{i_1, \dots, i_l}| \leq \frac{1}{4} \sum_{1 \leq i < j \leq k} \frac{\sqrt{p_i p_j}}{\sqrt{(1-p_i)(1-p_j)}} \exp\left(\frac{4\varepsilon^2 N}{2\left(1 + \sqrt{p_i p_j} / \sqrt{(1-p_i)(1-p_j)}\right)}\right) \quad (10)$$

$$\leq \frac{k(k-1)}{4} \exp(-2\varepsilon^2 N),$$

since $\sqrt{p_i p_j} / \sqrt{(1-p_i)(1-p_j)} \leq 1$. Then equations (7)–(10) yield

$$\tilde{P} \geq \prod_{i=1}^k (2\Phi(2\varepsilon\sqrt{N}) - 1) - \sum_{l=0}^k \binom{k}{l} \frac{k(k-1)}{4} \exp(-2\varepsilon^2 N) \quad (11)$$

$$\geq (2\Phi(2\varepsilon\sqrt{N}) - 1)^k - 2^{k-2} k(k-1) \exp(-2\varepsilon^2 N).$$

Since the right side expression in inequality (11) converges to 1 as the sample size N increases to infinity, this inequality provides the uniform lower bound for rate of convergence in probability for the ML estimates \hat{p}_i to true values of parameters p_i , $i = 1, \dots, k$ (or, equivalently, the lower bound for the rate of decay of the maximum error of the ML estimation).

Note that for $k = 2$ (X is the sequence of Bernoulli trials with probabilities p for ‘success’ and q for ‘failure’) the rate of convergence can be estimated directly from the Moivre-Laplace theorem on the normal approximation for binomial distribution with parameters p and N :

$$P(|\hat{p} - p| \leq \varepsilon, |\hat{q} - q| \leq \varepsilon) = P(|\hat{p} - p| \leq \varepsilon, |(1-\hat{p}) - (1-p)| \leq \varepsilon)$$

$$= P(|\hat{p} - p| \leq \varepsilon) = P\left(\left|\frac{N(\text{success}) - Np}{\sqrt{Npq}}\right| \leq \left(\varepsilon\sqrt{N} / \sqrt{pq}\right)\right) \quad (12)$$

$$\approx 2\Phi\left(\varepsilon\sqrt{N} / \sqrt{pq}\right) - 1 \geq 2\Phi(2\varepsilon\sqrt{N}) - 1.$$

Obviously, the estimate (12) is better than estimate (11) for $k = 2$. It can be explained by the fact that for the Bernoulli case (and only for this case) the distribution of the asymptotically Gaussian vector $((\hat{p} - p)\sqrt{N}, (\hat{q} - q)\sqrt{N})$ is uniquely defined by the distribution of its either component.

For nucleotide sequences ($k = 4$) inequality (11) becomes

$$\tilde{P} = P(\max_{i=1, \dots, 4} |\hat{p}_i - p_i| \leq \varepsilon) \geq (2\Phi(2\varepsilon\sqrt{N}) - 1)^4 - 48 \exp(-2\varepsilon^2 N). \quad (13)$$

Inequality (11) can also be used to derive the uniform confidence intervals for p_1, \dots, p_k with specified confidence level ε_1 as follows. For any $\varepsilon, \varepsilon_1: \varepsilon > 0, 0 < \varepsilon_1 < 1$, there exists $n(\varepsilon, \varepsilon_1)$ such that for any $N \geq n(\varepsilon, \varepsilon_1)$ the ML estimates $\hat{p}_i, i = 1, \dots, k$, satisfy the following inequality

$$P(\max_i |\hat{p}_i - p_i| \leq \varepsilon) \geq \varepsilon_1.$$

In the other words, if the length of sequence X is greater than $n(\varepsilon, \varepsilon_1)$, then for all $i = 1, \dots, k$, an interval $[\hat{p}_i - \varepsilon, \hat{p}_i + \varepsilon]$ is a confidence interval for p_i with confidence level greater than ε_1 . For instance, for a nucleotide sequence generated by the

independence model M we obtain from equation (13) that if the sequence length is at least 1,240 nt, then

$$P(\max_i |\hat{p}_i - p_i| \leq 0.05) \geq 0.9.$$

To achieve an uniform accuracy of estimates equal to $\varepsilon = 0.01$ with the same confidence level 0.9, the length of the nucleotide sequences should be at least 30,980 nt. For a protein sequence generated by an independence model inequality $P(\max_{1 \leq i \leq 20} |\hat{p}_i - p_i| \leq 0.01) \geq 0.09$ holds when the sequence is at least 103,556 amino acids long.

The other way to estimate \tilde{P} is to use the χ^2 approximation for the multinomial parameters. It is known from the classical χ^2 -theory (Kendall, 1945; Section 12.2) that under assumption that all probabilities p_i are positive, the statistic

$$\chi_v^2 = \sum_{i=1}^k \frac{(N_i - Np_i)^2}{Np_i}$$

is asymptotically χ^2 -distributed with $v = k - 1$ degrees of freedom. Then for sufficiently large N and $\varepsilon > 0$ the following inequalities hold:

$$\begin{aligned} \tilde{P} &= P(\max_i |\hat{p}_i - p_i| \leq \varepsilon) = P\left(\left|\frac{N_1}{N} - p_1\right| \leq \varepsilon, \dots, \left|\frac{N_k}{N} - p_k\right| \leq \varepsilon\right) \\ &= P\left(\frac{(N_1 - Np_1)^2}{Np_1} \leq \frac{\varepsilon^2 N}{p_1}, \dots, \frac{(N_k - Np_k)^2}{Np_k} \leq \frac{\varepsilon^2 N}{p_k}\right) \\ &\geq P\left(\sum_{i=1}^k \frac{(N_i - Np_i)^2}{Np_i} \leq \frac{\varepsilon^2 N}{\max p_i}\right) = P\left(\chi_{k-1}^2 \leq \frac{\varepsilon^2 N}{\max p_i}\right) \\ &\geq P(\chi_{k-1}^2 \leq \varepsilon^2 N) \approx \frac{1}{2^{(k-1)/2} \Gamma((k-1)/2)} \int_0^{\varepsilon^2 N} e^{-(x/2)} x^{(k-3)/2} dx. \end{aligned} \tag{14}$$

Here $\Gamma(y)$ is a gamma function. For nucleotide sequences ($k = 4$) estimate (14) yields:

$$\tilde{P} \geq 2\Phi(\varepsilon\sqrt{N}) - 1 - \frac{\sqrt{2N}\varepsilon}{\sqrt{\pi}} \exp\left(-\frac{\varepsilon^2 N}{2}\right). \tag{15}$$

Now inequality $P = (\max_i |\hat{p}_i - p_i| \leq 0.05) \geq 0.9$ follows from statement (15) for a nucleotide sequence with length at least 2,500 nt.

The comparison of the lower bounds L_1 and L_2 defined by right side expression in equations (11) and (14), respectively, shows that for any number k of parameters estimate (14) is better than estimate (11) ($L_1 < L_2$) for short sequences (but still long enough to use the normal and the χ^2 -approximations). However, starting from some $N = N(\varepsilon, k)$, the estimate (11) becomes better than estimate (14) ($L_1 > L_2$). For instance, for nucleotide sequences $L_1 < L_2$ if $\varepsilon\sqrt{N} < 1.51$; otherwise, $L_1 > L_2$. Therefore, by combining estimates (11) and (14) we obtain the following inequality for the ML estimates of the parameters of the independence model:

$$P(\max_i |\hat{p}_i - p_i| \leq \varepsilon) \geq \max \begin{cases} (2\Phi(2\varepsilon\sqrt{N}) - 1)^k - 2^{k-2} k(k-1)e^{-2\varepsilon^2 N} \\ P(\chi_{k-1}^2 \leq \varepsilon^2 N). \end{cases} \quad (16)$$

Formula (13) was used to compute the length of a DNA sequence sufficient for estimation of parameters of the independence model with desirable accuracy ε and the standard statistical confidence levels ε_1 equal to 0.90, 0.95 and 0.99 (Table 1).

Table 1 The minimal length of a nucleotide sequence generated by the independence model with parameters p_i , $1 \leq i \leq 4$, for which inequality $P(\max_{1 \leq i \leq 4} |\hat{p}_i - p_i| \leq \varepsilon) \geq \varepsilon_1$ holds according to lower bound equation (13)

| | $\varepsilon = 0.1$ | $\varepsilon = 0.05$ | $\varepsilon = 0.01$ |
|------------------------|---------------------|----------------------|----------------------|
| $\varepsilon_1 = 0.90$ | 310 | 1,240 | 30,980 |
| $\varepsilon_1 = 0.95$ | 345 | 1,377 | 34,410 |
| $\varepsilon_1 = 0.99$ | 425 | 1,698 | 42,436 |

3.2 Markov chain model

We consider the first-order ergodic stationary Markov chain $X = (X_1, \dots, X_N)$ with unknown transition probabilities estimated by equation (2) and stationary distribution π estimated by $\hat{\pi}(\alpha) = N(\alpha) / N$. It is easy to show that the ML estimates $\hat{\pi}(\alpha)$ are unbiased:

$$\begin{aligned} \mathbf{E}\hat{\pi}(\alpha) &= \frac{1}{N} \mathbf{E}N(\alpha) = \frac{1}{N} \sum_{i=1}^N P(x_i = \alpha) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{x_1, x_2, \dots, x_{i-1}} \pi(x_1) p_{x_1, x_2} \times \dots \times p_{x_{i-1}, \alpha} = \frac{N}{N} \pi(\alpha) = \pi(\alpha). \end{aligned}$$

From the theory of Markov processes it is known (Billingsley, 1961; Lemma 3.2, Theorem 3.3) that estimates $\hat{\pi}(\alpha)$ are consistent, and vector ξ with components $\xi_i = (\hat{\pi}(\alpha_i) - \pi(\alpha_i))\sqrt{N}$, $i = 1, \dots, k$, is asymptotically a centered Gaussian vector. Therefore, similarly to the case of the independence model, one could derive the analog of formula (11) that would define confidence intervals for stationary probabilities. In practical situations, however, the properties of the ML estimates \hat{p}_{ij} , $i, j = 1, \dots, k$, of transition probabilities are of greater interest. These estimates are consistent: $\hat{p}_{ij} \xrightarrow{P} p_{ij}$, $i, j = 1, \dots, k$, (Billingsley, 1961; Theorem 4.1). To determine a rate of the convergence in probability, we will use the asymptotic normality of a k^2 -dimensional vector $\eta = (\eta_{11}, \dots, \eta_{1k}, \dots, \eta_{k1}, \dots, \eta_{kk})$ with components

$$\eta_{ij} = \left(\frac{N(\alpha_i \alpha_j)}{N(\alpha_i)} - p_{ij} \right) \sqrt{N\pi_i} = (\hat{p}_{ij} - p_{ij}) \sqrt{N\pi_i},$$

$\pi_i = \pi(\alpha_i)$, $i, j = 1, \dots, k$ (Billingsley, 1961; Proof of Theorem 3.1). This vector is an asymptotically centered normal vector with covariance matrix $R = (r_{ij,kl}) = (\mathbf{E}\eta_{ij}\eta_{sl})$, $i, j, s, l = 1, \dots, k$: $r_{ij,sl} = 0$ for $i \neq s, j, l = 1, \dots, k$; $r_{ij,ij} = p_{ij} - p_{ij}^2$ for $i, j = 1, \dots, k$; $r_{ij,il} = -p_{ij}p_{il}$ for $i, j, l = 1, \dots, k, j \neq l$. Then, under assumption that all transition probabilities are positive, k^2 random variables $Z_{11}, \dots, Z_{1k}, \dots, Z_{k1}, \dots, Z_{kk}$, defined as

$$Z_{ij} = \frac{\eta_{ij}}{\sqrt{\text{Var}(\eta_{ij})}} = \frac{\eta_{ij}}{\sqrt{p_{ij}(1-p_{ij})}},$$

are asymptotically standard normal random variables with covariances

$$\begin{aligned} \text{cov}(Z_{ij}, Z_{il}) &= \frac{r_{ij,il}}{\sqrt{\text{Var}(\eta_{ij})\text{Var}(\eta_{il})}} = -\frac{\sqrt{p_{ij}p_{il}}}{\sqrt{(1-p_{ij})(1-p_{il})}}, \quad j \neq l; \\ \text{cov}(Z_{ij}, Z_{sl}) &= 0, \quad i \neq s. \end{aligned} \quad (17)$$

Note that covariances (17) are non-zeros only for pairs (Z_{ij}, Z_{il}) corresponding to components of one and the same vector of probabilities of transition from state i , $i = 1, \dots, k$, and a structure of such covariances is identical to covariances (5) of Z_i, Z_j introduced for the independence model. This observation helps apply inequality (6) to random variables Z_{ij} , $i, j = 1, \dots, k$, and to proceed similarly to the case of the independence model. For $\varepsilon > 0$ we have

$$\begin{aligned} \tilde{P} &= P(\max_{i,j} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon) = P(|\eta_{ij}| \leq \varepsilon\sqrt{N\pi_i}, \quad i, j = 1, \dots, k) \\ &= P\left(|Z_{ij}| \leq \frac{\varepsilon\sqrt{N\pi_i}}{\sqrt{p_{ij}(1-p_{ij})}}, \quad i, j = 1, \dots, k\right) \\ &\geq P(|Z_{ij}| \leq 2\varepsilon\sqrt{N\pi_i}, \quad i, j = 1, \dots, k) \\ &\geq P(|Z_{ij}| \leq 2\varepsilon\sqrt{N \min_{i,j} p_{ij}}, \quad i, j = 1, \dots, k), \end{aligned} \quad (18)$$

since $\pi_i \geq \min_j p_{ji}$, $i = 1, \dots, k$ (for proof see Appendix 2). Computations similar to ones performed at the steps (8)–(10) transform inequality (18) into

$$\tilde{P} \geq (2\Phi(2\varepsilon\sqrt{Np'}) - 1)^{k^2} - 2^{k^2-2} k^2 (k-1) \exp(-2\varepsilon^2 Np'), \quad (19)$$

where p' stands for $\min_{i,j} p_{ij}$. Inequality (19) yields for the estimates \hat{p}_{ij} the uniform lower bound of the rate of convergence in probability to true values of transition probabilities p_{ij} , $i, j = 1, \dots, k$, and can be used for finding the confidence intervals for p_{ij} (as it was done for the independence model). Note that the lower bound equation (19), being uniform with respect to an accuracy of ML estimates of the transition probabilities, does depend on the values of probability parameters of the Markov chain and requires the information about the positive lower bound of the $\min_{i,j} p_{ij}$.

For nucleotide sequences ($k = 4$) we have

$$\tilde{P} \geq (2\Phi(2\varepsilon\sqrt{Np'}) - 1)^{16} - 786,432 \exp(-2\varepsilon^2 Np'). \quad (20)$$

For example, for a nucleotide sequence generated by the Markov chain model with $p' \geq 0.15$, we obtain from equation (20) that inequality

$$P\left(\max_{i,j} |\hat{p}_{ij} - p_{ij}| \leq 0.05\right) \geq 0.9$$

holds if the sequence length is at least 21,140 nt.

Next we find a lower bound for \tilde{P} using χ^2 -approximation. From the theory of the Markov processes it is known (Billingsley, 1961; Theorem 3.1) that statistic $\chi_v^2 = \sum_{i,j=1}^k (\eta_{ij}^2 / p_{ij})$ is asymptotically χ^2 -distributed with $v = d - k$ degrees of freedom, where d is the number of positive entries of the transition probability matrix (p_{ij}) . Assuming that all transition probabilities are positive, we have $v = k^2 - k$.

Then for sufficiently large N and $\varepsilon > 0$ the following inequalities hold:

$$\begin{aligned} P(\max_{i,j} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon) &\geq P\left(\max_{i,j} |\eta_{ij}| \leq \varepsilon \sqrt{N \min_i \pi(i)}\right) \\ &= P(\max_{i,j} \eta_{ij}^2 \leq \varepsilon^2 N \min_i \pi(i)) \geq P\left(\sum_{i,j} \eta_{ij}^2 \leq \varepsilon^2 N \min_i \pi(i)\right) \\ &\geq P\left(\sum_{i,j} \frac{\eta_{ij}^2}{p_{ij}} \leq \varepsilon^2 N p'\right) \approx P(\chi_{k^2-k}^2 \leq \varepsilon^2 N p'). \end{aligned} \quad (21)$$

Therefore, for DNA sequences ($k^2 - k = 12$) inequality (20) yields

$$\begin{aligned} P(\max_{i,j} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon) &\geq P(\chi_{12}^2 \leq \varepsilon^2 N p') \\ &= \frac{1}{2^6 \Gamma(6)} \int_0^{\varepsilon^2 N p'} e^{-(x/2)} x^5 dx \\ &= 1 - \exp\left(-\frac{\varepsilon^2 N p'}{2}\right) \sum_{m=0}^5 \frac{1}{(2m)!!} (\varepsilon^2 N p')^m. \end{aligned} \quad (22)$$

Since the right side of equation (22) tends to 1 as the length N of sequence X grows to infinity, inequality (22) gives the uniform lower bound of the rate of convergence in probability of \hat{p}_{ij} to true values of transition probabilities p_{ij} , $i, j = 1, \dots, k$. A comparison of the lower bounds from inequalities (20) and (22) derived for the ML estimates of the transition probabilities of the Markov chain with $k = 4$ states reveals, similarly to the independence case, that estimate (22) based on χ^2 -approximation is better for smaller values of N (when $\varepsilon \sqrt{N p'} < 2.619$); otherwise, estimate (20) derived from the normal approximation is better. For instance, under assumption that $p' \geq 0.15$, it follows from equation (22) that $P(\max_{1 \leq i, j \leq 4} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon) \geq \varepsilon_1$ holds true for transition probabilities p_{ij} , $i, j = 1, \dots, 4$, if the ‘training’ DNA sequence is longer than 49,462 nt (21,140 nt from estimate (20)).

Once again, combining estimates (19) and (21) to cover all values of N where the normal approximation and the χ^2 -approximation are applicable, we derive the general

formula for the lower bound of the rate of convergence in probability of the ML estimates of transition probabilities of the Markov chain with k states:

$$P\left(\max_{i,j} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon\right) \geq \max \begin{cases} P(\chi_{k^2-k}^2 \leq \varepsilon^2 Np') \\ (2\Phi(2\varepsilon\sqrt{Np'}) - 1)^{k^2} - 2^{k^2-2} k^2 (k-1) e^{-2\varepsilon^2 Np'} \end{cases} \quad (23)$$

Formula (20) was used to compute the length of a DNA sequence sufficient for estimation of parameters of the Markov model with desirable accuracy ε and the standard statistical confidence level ε_1 equal to 0.90, 0.95 and 0.99 (Tables 2 and 3).

Table 2 The minimal length of a training nucleotide sequence sufficient for estimation of parameters p_{ij} , $1 \leq i, j \leq 4$, $\min_{i,j} p_{ij} \geq 0.1$, of the stationary ergodic Markov chain with accuracy characterised by inequality $P(\max_{1 \leq i, j \leq 4} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon) \geq \varepsilon_1$

| | $\varepsilon_1 = 0.1$ | $\varepsilon_1 = 0.05$ | $\varepsilon_1 = 0.01$ |
|------------------------|-----------------------|------------------------|------------------------|
| $\varepsilon_1 = 0.90$ | 7,930 | 31,720 | 792,990 |
| $\varepsilon_1 = 0.95$ | 8,272 | 33,086 | 827,138 |
| $\varepsilon_1 = 0.99$ | 9,060 | 36,240 | 906,010 |

Table 3 The same as in Table 2 with $\min_{i,j} p_{ij} \geq 0.15$

| | $\varepsilon_1 = 0.1$ | $\varepsilon_1 = 0.05$ | $\varepsilon_1 = 0.01$ |
|------------------------|-----------------------|------------------------|------------------------|
| $\varepsilon_1 = 0.90$ | 5,290 | 21,140 | 528,660 |
| $\varepsilon_1 = 0.95$ | 5,515 | 22,057 | 551,425 |
| $\varepsilon_1 = 0.99$ | 6,040 | 24,160 | 604,010 |

3.3 Hidden Markov Model (HMM)

Inequalities (16) and (23) can be also used to determine bounds for the rate of convergence in probability of the ML estimates of the HMM parameters. This application is relevant for the case when both a sequence of symbols and the corresponding sequence of hidden states are experimentally determined and available for parameter estimation (model training). For instance, for the HMM based algorithm of prediction of the protein secondary structure the training set is comprised from proteins sequences with known secondary structures. The estimation of parameters of the HMM describing gene organisation in genomic DNA from the training set of annotated genomic sequences (Durbin et al., 1998, p.62) could be another example.

We consider an HMM with k_1 hidden states $1, 2, \dots, k_1$, emitting k_2 distinct symbols x_1, \dots, x_{k_2} . Parameters of the HMM are transition probabilities p_{ij} , $i, j = 1, \dots, k_1$, and emission probabilities $e_i(x_j)$, $i = 1, \dots, k_1$, $j = 1, \dots, k_2$. We assume that all emission and transition probabilities have non-zero values and that the Markov chain of hidden states is a stationary ergodic Markov chain. Then the ML estimates \hat{p}_{ij} of transition probabilities

are defined by equation (2) the same as for a regular (non-hidden) Markov chain, and the ML estimates $e_i(x_j)$ of emission probabilities $e_i(x_j)$, $i = 1, \dots, k_1$, $j = 1, \dots, k_2$, are defined by equation

$$\hat{e}_i(x_j) = \frac{E_i(x_j)}{\sum_{l=1}^{k_2} E_i(x_l)},$$

where $E_i(x_j)$ designates the number of times that symbol x_j was emitted from hidden state i in training sequence (Durbin et al., 1998, p.62). A subsequence S^i of the training sequence consisting only of symbols emitted from a given hidden state i is generated by the independence model M with parameters $e_i(x_j)$, $j = 1, \dots, k_2$. Assuming that sequence S^i has length N_i , from inequality (16) we have for any $i = 1, \dots, k_1$:

$$\begin{aligned} P\left(\max_{i \leq j \leq k_2} |\hat{e}_i(x_j) - e_i(x_j)| \leq \varepsilon\right) &\geq \max\{P(\chi_{k_2-1}^2 \leq \varepsilon^2 N_i); \\ &(2\Phi(2\varepsilon\sqrt{N_i}) - 1)^{k_2} - 2^{k_2-2} k_2(k_2 - 1) \exp(-2\varepsilon^2 N_i)\}. \end{aligned} \quad (24)$$

Independence of sequences S^i , $i = 1, \dots, k_1$, allows to obtain the following inequality for the ML estimates of emission probabilities uniformly over all hidden states:

$$\begin{aligned} P\left(\max_{i \leq i \leq k_1} \max_{i \leq j \leq k_2} |\hat{e}_i(x_j) - e_i(x_j)| \leq \varepsilon\right) &= \prod_{i=1}^{k_1} P(\max_{i \leq j \leq k_2} |\hat{e}_i(x_j) - e_i(x_j)| \leq \varepsilon) \\ &\geq \max\left\{\prod_{i=1}^{k_1} ((2\Phi(2\varepsilon\sqrt{N_i}) - 1)^{k_2} - 2^{k_2-2} k_2(k_2 - 1) \exp(-2\varepsilon^2 N_i)); \right. \\ &\quad \left. \prod_{i=1}^{k_1} P(\chi_{k_2-1}^2 \leq \varepsilon^2 N_i)\right\}. \end{aligned}$$

To obtain similar result for transition probabilities, we have to turn to the underlying sequence of hidden states. This sequence is described by the first-order stationary ergodic Markov chain with k_1 states and transition probabilities p_{ij} , $i, j = 1, \dots, k_1$; therefore, inequality (23) holds for these ML estimates:

$$\begin{aligned} P\left(\max_{i \leq i, j \leq k_1} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon\right) & \\ &\geq \max\left\{P(\chi_{k_1^2 - k_1}^2 \leq \varepsilon^2 N p'); \right. \\ &\quad \left.(2\Phi(2\varepsilon\sqrt{N p'}) - 1)^{k_1^2} - 2^{k_1^2 - 2} k_1^2 (k_1 - 1) e^{-2\varepsilon^2 N p'}\right\}. \end{aligned} \quad (25)$$

Here N is the length of the training sequence (the same as the length of the sequence of underlying hidden states) and $p' = \min_{i,j} p_{ij}$. Note that the bound of convergence rate (24) depends on both k_1 and k_2 while the bound equation (25) depends on k_1 only. The explanation is obvious: an emission probability defined for a particular symbol emitted from a particular hidden state depends on both entities, while the Markov chain of hidden states is completely independent of emissions and does not depend on the size k_2 of the alphabet of emitted symbols. As described above, formulas (24) and (25) can yield confidence intervals for true values of parameters $e_i(x_j)$, $l = 1, \dots, k_2$, $i = 1, \dots, k_1$, and p_{ij} , $i, j = 1, \dots, k_1$, respectively, at a specified confidence level ε_1 .

As an example we consider an HMM for prediction of protein secondary structure with three hidden states (α -helix, β -strand and loop) and 20 observed amino acids ($k_1 = 3, k_2 = 20$). Application of formula (24) shows that the inequality

$$P\left(\max_{1 \leq i \leq 3} \max_{1 \leq j \leq 20} |\hat{e}_i(x_j) - e_i(x_j)| \leq 0.05\right) \geq 0.9$$

holds if the minimal among $N_i, i = 1, 2, 3$, is at least 4,360. Assuming that $\min_{i,j} p_{ij} > 0.15$, we have from equation (25) that the uniform accuracy of approximation for the transition probabilities is characterised by inequality

$$P\left(\max_{1 \leq i, j \leq 3} |\hat{p}_{ij} - p_{ij}| \leq 0.05\right) \geq 0.9$$

which holds if the length N of a training sequence is at least 13,393 amino acids.

Formulas (24) and (25) were used for computing the size of the training sets (total length of protein sequences with known secondary structure) sufficient for estimating the HMM parameters with desirable accuracy ε and the standard statistical confidence levels ε_1 equal to 0.90, 0.95 and 0.99 (Tables 4–6).

Table 4 The minimal number of amino acids situated in each structural confirmation (α -helix, β -strand and loop) of a training sequence which is sufficient for estimation of the HMM emission probabilities $e_i(x_j)$ with accuracy characterised by inequality $P(\max_{1 \leq i \leq 3} \max_{1 \leq j \leq 20} |\hat{e}_i(x_j) - e_i(x_j)| \leq \varepsilon) \geq \varepsilon_1$

| | $\varepsilon = 0.1$ | $\varepsilon = 0.05$ | $\varepsilon = 0.01$ |
|------------------------|---------------------|----------------------|----------------------|
| $\varepsilon_1 = 0.90$ | 1,090 | 4,360 | 108,966 |
| $\varepsilon_1 = 0.95$ | 1,125 | 4,500 | 112,494 |
| $\varepsilon_1 = 0.99$ | 1,207 | 4,825 | 120,618 |

Table 5 The minimal length of a training protein sequence sufficient for estimation of the HMM transition probabilities ($\min_{i,j} p_{ij} \geq 0.1$) with accuracy characterised by inequality $P(\max_{1 \leq i, j \leq 3} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon) \geq \varepsilon_1$

| | $\varepsilon = 0.1$ | $\varepsilon = 0.05$ | $\varepsilon = 0.01$ |
|------------------------|---------------------|----------------------|----------------------|
| $\varepsilon_1 = 0.90$ | 5,022 | 20,088 | 502,208 |
| $\varepsilon_1 = 0.95$ | 5,368 | 21,474 | 536,849 |
| $\varepsilon_1 = 0.99$ | 6,175 | 24,700 | 617,523 |

Table 6 The same as in Table 5 with $\min_{i,j} p_{ij} \geq 0.15$

| | $\varepsilon = 0.1$ | $\varepsilon = 0.05$ | $\varepsilon = 0.01$ |
|------------------------|---------------------|----------------------|----------------------|
| $\varepsilon_1 = 0.90$ | 3,348 | 13,393 | 334,806 |
| $\varepsilon_1 = 0.95$ | 3,579 | 14,316 | 357,900 |
| $\varepsilon_1 = 0.99$ | 4,117 | 16,468 | 411,682 |

4 Conclusion

In the present work, we have derived the lower bounds for the rate of convergence in probability of the maximum error for the ML estimates of parameters of the statistical models frequently used in bioinformatics: the independence model, the first-order Markov chain, and the HMM. For these models the inequalities (16), (23)–(25) that provide the lower bounds could be also used for finding the confidence intervals for unknown values of parameters at a specified confidence level. Each lower bound L is the maximum of two lower bounds, L_1 and L_2 , derived from the normal approximation and the χ^2 -approximation, respectively, for the vector of ML estimates. The bound L_2 is maximal for smaller sequence lengths, while L_1 is maximal for large sequence lengths. Note that inequalities (16) and (24) hold for all values of parameters, while the lower bounds given by equations (23) and (25) require additional information about the range of parameter values (transition probabilities).

We list the lengths of nucleotide or protein sequence sufficient to achieve a desirable accuracy ε of the ML estimates at a specified confidence level for the parameters of a given model (Tables 1–6). Formula (3) allows to translate the estimation error ε in equations (16), (23)–(25) to the error Δ of computations of the likelihood of a model for fragments of empirical sequences used in many bioinformatics algorithms (e.g., Lawrence et al., 1993; Borodovsky and McIninch, 1993; Burge and Karlin, 1997; Durbin et al., 1998).

The inequalities (16), (23)–(25) for the lower bounds of the convergence rates for different models present a new theoretical result articulating yet another property of the ML estimates of parameters.

Acknowledgements

The authors thank Dr. Alexander Mitrophanov for critical comments on the initial draft of the paper and further helpful discussions. We are grateful to Jialiang Wu for help with computations. This work was supported in part by the USA National Institutes of Health grant awarded to MB.

References

- Almagor, H. (1983) 'A Markov analysis of DNA sequences', *Journal of Theoretical Biology*, Vol. 104, pp.633–645.
- Billingsley, P. (1961) 'Statistical methods in Markov chains', *The Annals of Mathematical Statistics*, Vol. 32, pp.12–40.
- Borodovsky, M. and McIninch, J. (1993) 'GeneMark: parallel gene recognition for both DNA strands', *Computers and Chemistry*, Vol. 17, pp.123–133.
- Borodovsky, M.Y., Sprizhitsky, Y.A., Golovanov, E.I. and Alexandrov, A.A. (1986a) 'Statistical patterns in the primary structure of the functional regions of the Escherichia coli genome. I. Frequency characteristics', *Molecular Biology*, Vol. 20, pp.826–833 (English translation).
- Borodovsky, M.Y., Sprizhitsky, Y.A., Golovanov, E.I. and Alexandrov, A.A. (1986b) 'Statistical patterns in the primary structure of the functional regions of the Escherichia coli genome. II. Nonuniform Markov models', *Molecular Biology*, Vol. 20, pp.833–840 (English translation).

- Burge, C. and Karlin, S. (1997) 'Prediction of complete gene structures in human genomic DNA', *Journal of Molecular Biology*, Vol. 268, pp.78–94.
- Churchill, G.A. (1989) 'Stochastic models for heterogeneous DNA sequences', *Bulletin of Mathematical Biology*, Vol. 51, pp.79–94.
- Cowan, R. (1991) 'Expected frequencies of DNA patterns using Whittle's formula', *Journal of Applied Probability*, Vol. 28, pp.886–892.
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall, London.
- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge, UK, Cambridge University Press, New York.
- Fitch, W.M. (1983) 'Calculating the expected frequencies of potential secondary structure in nucleic acids as a function of stem length, loop size, base composition and nearest-neighbor frequencies', *Nucleic Acids Research*, Vol. 11, pp.4655–4663.
- Gatlin, L.L. (1972) *Information Theory and the Living System*, Columbia University Press, New York.
- Karlin, S. and Macken, C. (1991) 'Assessment of inhomogeneities in an *E. Coli* physical map', *Nucleic Acids Research*, Vol. 19, pp.4241–4246.
- Karlin, S., Burge, C. and Campbell, A.M. (1992) 'Statistical analyses of counts and distributions of restriction sites in DNA sequences', *Nucleic Acids Research*, Vol. 20, pp.1363–1370.
- Kendall, M.G. (1945) *The Advanced Theory of Statistics*, 2nd ed., C. Griffin, London.
- Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) 'Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment', *Science*, Vol. 262, pp.208–214.
- Li, W.V. and Shao, Q-M. (2002) 'A normal comparison inequality and its applications', *Probability Theory and Related Fields*, Vol. 122, pp.494–508.
- Reinert, G., Schbath, S. and Waterman, M.S. (2000) 'Probabilistic and statistical properties of words: an overview', *Journal of Computational Biology*, Vol. 7, pp.1–46.
- Tavare, S. and Song, B. (1989) 'Codon preference and primary sequence structure in protein coding regions', *Bulletin of Mathematical Biology*, Vol. 51, pp.95–115.

Appendix 1

Here we prove the statement: for any random variables X_1, \dots, X_k and positive numbers x_1, \dots, x_k the following equality holds:

$$\begin{aligned}
 & P(|X_1| \leq x_1, \dots, |X_k| \leq x_k) \\
 &= \sum_{l=0}^k (-1)^l \sum_{\substack{\{i_1, \dots, i_l\} \subseteq \{1, \dots, k\}, \\ \{i_1, \dots, i_l\} \cup \{i_{i_1+1}, \dots, i_k\} = \{1, \dots, k\}}} P(X_{i_1} \leq -x_{i_1}, \dots, \\
 & \quad X_{i_l} \leq -x_{i_l}, X_{i_{i_1+1}} \leq x_{i_{i_1+1}}, \dots, X_{i_k} \leq x_{i_k}).
 \end{aligned} \tag{26}$$

The inner sum in equation (26) is taken over all possible choices of indices i_1, \dots, i_l out of $1, \dots, k$.

We use the mathematical induction. For $k = 2$ we have

$$\begin{aligned}
 & P(|X_1| \leq x_1, |X_2| \leq x_2) = P(X_1 \leq x_1, |X_2| \leq x_2) - P(X_1 \leq -x_1, |X_2| \leq x_2) \\
 &= P(X_1 \leq x_1, X_2 \leq x_2) - P(X_1 \leq x_1, X_2 \leq -x_2) \\
 & \quad - P(X_1 \leq -x_1, X_2 \leq -x_2) + P(X_1 \leq -x_1, X_2 \leq x_2).
 \end{aligned}$$

Now, we assume that statement holds for $k = n$ and consider $k = n + 1$:

$$\begin{aligned}
 P(|X_1| \leq x_1, \dots, |X_{n+1}| \leq x_{n+1}) &= P(|X_1| \leq x_1, \dots, |X_n| \leq x_n, X_{n+1} \leq x_{n+1}) \\
 &\quad - P(|X_1| \leq x_1, \dots, |X_n| \leq x_n, X_{n+1} \leq -x_{n+1}) \\
 &= \sum_{l=0}^n (-1)^l \sum_{\substack{\{i_1, \dots, i_l\} \subset \{1, \dots, n\}, \\ \{i_1, \dots, i_l\} \cup \{i_{l+1}, \dots, i_n\} = \{1, \dots, n\}}} P(X_{i_1} \leq -x_{i_1}, \dots, X_{i_l} \leq -x_{i_l}, \\
 &\quad X_{i_{l+1}} \leq x_{i_{l+1}}, \dots, X_{i_n} \leq x_{i_n}, X_{n+1} \leq x_{n+1}) \\
 &\quad - \sum_{l=0}^n (-1)^l \sum_{\substack{\{i_1, \dots, i_l\} \subset \{1, \dots, n\}, \\ \{i_1, \dots, i_l\} \cup \{i_{l+1}, \dots, i_n\} = \{1, \dots, n\}}} P(X_{i_1} \leq -x_{i_1}, \dots, X_{i_l} \leq -x_{i_l}, \\
 &\quad X_{i_{l+1}} \leq x_{i_{l+1}}, \dots, X_{i_n} \leq x_{i_n}, X_{n+1} \leq -x_{n+1}) \\
 &= \sum_{l=0}^n (-1)^l \sum_{\substack{\{i_1, \dots, i_l\} \subset \{1, \dots, n\}, \\ \{i_1, \dots, i_l\} \cup \{i_{l+1}, \dots, i_n\} = \{1, \dots, n\}}} P(X_{i_1} \leq -x_{i_1}, \dots, X_{i_l} \leq -x_{i_l}, \\
 &\quad X_{i_{l+1}} \leq x_{i_{l+1}}, \dots, X_{i_n} \leq x_{i_n}, X_{n+1} \leq x_{n+1}) \\
 &\quad + \sum_{l=0}^n (-1)^{l+1} \sum_{\substack{\{i_1, \dots, i_l\} \subset \{1, \dots, n\}, \\ \{i_1, \dots, i_l\} \cup \{i_{l+1}, \dots, i_n\} = \{1, \dots, n\}}} P(X_{i_1} \leq -x_{i_1}, \dots, X_{i_l} \leq -x_{i_l}, \\
 &\quad X_{i_{l+1}} \leq x_{i_{l+1}}, \dots, X_{i_n} \leq x_{i_n}, X_{n+1} \leq -x_{n+1}) \\
 &= \sum_{l=0}^{n+1} (-1)^l \sum_{\substack{\{i_1, \dots, i_l\} \subset \{1, \dots, n+1\}, \\ \{i_1, \dots, i_l\} \cup \{i_{l+1}, \dots, i_{n+1}\} = \{1, \dots, n+1\}}} P(X_{i_1} \leq -x_{i_1}, \dots, X_{i_l} \leq -x_{i_l}, \\
 &\quad X_{i_{l+1}} \leq x_{i_{l+1}}, \dots, X_{i_{n+1}} \leq x_{i_{n+1}}).
 \end{aligned}$$

This completes the proof.

Appendix 2

Let X be the stationary ergodic first-order Markov chain with a finite set of states $1, 2, \dots, k$; transition probabilities p_{ij} , $i, j = 1, \dots, k$; and stationary distribution $\pi = (\pi_1, \dots, \pi_k)$. Then for any j the following inequalities hold:

$$\min_i p_{ij} \leq \pi_j \leq \max_i p_{ij}. \quad (27)$$

First we prove the left inequality in equation (27). Suppose that there exists j , $1 \leq j \leq k$ such that $\min_i p_{ij} > \pi_j$. Then for any $i = 1, \dots, k$, $\pi_j < p_{ij}$. We denote $\delta = \min_i p_{ij} - \pi_j$, $\delta > 0$.

Let $P = (p_{ij})$ be the matrix of transition probabilities of X . Since the limit of P^n when n grows to infinity must be equal to the matrix with vector π in each row, we have for the element (j, j) of P^n

$$b_n = \sum_{i_1, i_2, \dots, i_{n-1}} p_{j i_1} p_{i_1 i_2} \times \dots \times p_{i_{n-1} j} \rightarrow \pi_j$$

as $n \rightarrow \infty$. If in the sum for b_n we substitute $p_{i_{n-1} j}$ by π_j , we obtain a new sequence

$$a_n = \pi_j \sum_{l_1, l_2, \dots, l_{n-1}} p_{jl_1} p_{l_1 l_2} \times \dots \times p_{l_{n-2} l_{n-1}}$$

with properties:

$$a_n = \pi_j \sum_{l_{n-1}=1}^k \left(\sum_{l_1, \dots, l_{n-2}} p_{jl_1} p_{l_1 l_2} \times \dots \times p_{l_{n-2} l_{n-1}} \right) \rightarrow \pi_j \sum_{l_{n-1}=1}^k \pi_{l_{n-1}} = \pi_j,$$

$a_n \leq b_n - \delta$. Then the monotonicity property of the limit implies that inequality

$$\lim_{n \rightarrow \infty} a_n \leq \lim_{n \rightarrow \infty} b_n - \delta$$

must hold. We come to a contradiction: $\pi_j \leq \pi_j - \delta$ (while $\delta > 0$), and conclude that the assumption $\min_i p_{ij} > \pi_j$ was wrong; hence, $\pi_j \geq \min_i p_{ij}$. The right side inequality $\pi_j \leq \max_i p_{ij}$ can be proved similarly.