Available online at www.sciencedirect.com
ScienceDirect
Mathematical Biosciences 209 (2007) 470–485
Mathematical Biosciences
www.elsevier.com/locate/mbs

Convergence rate estimation for the TKF91 model of biological sequence length evolution

Alexander Y. Mitrophanov [a,1], Mark Borodovsky [a,b,*]

[a] School of Biology, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA
[b] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332-0535, USA

Received 6 September 2006; received in revised form 17 February 2007; accepted 23 February 2007
Available online 13 March 2007

Abstract

The TKF91 model of biological sequence evolution describes changes in the sequence length via an infinite state-space birth–death process, which we term the TKF91-BD process. The TKF91 model assumes that, for any pair of modern sequences, the ancestral sequence has equilibrium length distribution, an assumption whose validity has not been rigorously investigated. We obtain explicit upper and lower bounds on the rate of convergence to equilibrium for the distribution of the TKF91-BD process. We show that the rate of convergence of the TKF91-BD process for protein sequences with parameter values inferred from sequence data on α and β globins is too low to guarantee convergence to equilibrium on a reasonable timescale. For the analyzed nucleotide sequences, the convergence is faster, but the equilibrium sequence length is unrealistically small. The Jukes–Cantor model of nucleotide substitutions can converge considerably faster than the length evolution model for both amino acid and nucleotide sequences, while the speed of convergence for the Kimura model is close to that for the TKF91-BD process describing nucleotide sequences.

Keywords: Evolution of biological sequences; Mutation; TKF91 model; Birth–death process; Exponential convergence; Convergence bound

* Corresponding author. Address: Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332-0535, USA. Tel.: +1 404 894 8432; fax: +1 404 894 0519.
E-mail address: mark.borodovsky@biology.gatech.edu (M. Borodovsky).
[1] Present address: Department of Molecular Microbiology, Washington University School of Medicine, Campus Box 8230, St. Louis, MO 63110-1093, USA.

0025-5564/$ - see front matter © 2007 Published by Elsevier Inc.
doi:10.1016/j.mbs.2007.02.011

1. Introduction

In recent years, stochastic modeling of evolutionary changes in the length of biomolecular sequences was given considerable attention [1–8]. The proposed models take into account randomly occurring insertion and deletion events; if combined with an appropriate nucleotide or amino acid substitution models, they provide complete probabilistic description of sequence evolution. Not only do such models enhance our understanding of the evolutionary process at a quantitative level, but they also serve practical purposes such as providing a solid evolutionary basis for pairwise and multiple sequence alignments [1,4,9–12], estimating rates of point indel (insertion/deletion) mutations [1,2,10], and reconstructing phylogenies [5]. Originally stochastic indel models were designed to describe the evolution of DNA [7,8] and protein sequences [1,13]; recently they were extended to describe the evolution of RNA structure [14].

The two main types of mathematical formalism used to model indels are continuous-time Markov processes and hidden Markov models (HMMs). The major modeling tool to represent sequences with indels has been a special case of Markov chain termed birth–death process. Markov chains of this type possess a number of useful properties, and are well studied. However, in the case of infinite state space certain important properties of birth–death processes are quite difficult to analyze. In this paper we investigate one of such properties, the rate of convergence to equilibrium (stationary distribution). Our primary goal is to obtain estimates on the rate of convergence to equilibrium for the well-known TKF91 model, which underlies numerous approaches to statistical sequence alignment and evolutionary modeling [1,2,7–9,11–16]. Originally the model was formulated for nucleotide sequences [7]; later it was used to describe the evolution of proteins as well [1].

The assumption that sequence divergence occurs within the stationary distribution is very common in quantitative modeling of molecular evolution [17]. It is an important postulate in modeling indels under the TKF91 model, stating that the length distribution of the sequences at the time of divergence is the stationary distribution of the appropriately constructed birth–death process [1,2,7,8]. Indeed, under the assumptions of the TKF91 model, the likelihood of a pair of modern sequences S_1 and S_2, originated from the ancestral sequence, S_a, t years ago, is written as

$$P(S_1, S_2, t) = \sum_{S_a} P(S_a \to S_1, t) P(S_a \to S_2, t) P(S_a, \infty).$$

Here $P(S_a \to S_i, t)$, $i = 1, 2$, is the probability of transition from S_a to S_i in time t, and $P(S_a, \infty)$ is the equilibrium probability of sequence S_a. $P(S_a, \infty)$ is equal to the product of the equilibrium probability that S_a has the given length and the equilibrium probability that a sequence of that length consists of the specific nucleotides which constitute S_a. As we see, the equilibrium assumption plays a central role in the formulation of the TKF91 model. While such an assumption greatly simplifies the analysis, its validity requires additional investigation. This question motivated the development of the convergence rate estimation method described in this paper.

We obtain rigorous upper and lower bounds on the rate of convergence of the birth–death process in the TKF91 model, and compare that rate to the rate of convergence for the Jukes–Cantor and Kimura substitution models, for which the equilibrium assumption is standard [17]. In addition to establishing rigorous mathematical results, we analyze recent estimates of indel rates and comment on the validity of the equilibrium hypothesis for the TKF91 in the light of the available

data. Our computations show that, when the TKF91 model is applied to protein sequences, convergence to equilibrium occurs on biologically unreasonable time scales. For nucleotide sequences, convergence is faster; however, the equilibrium sequence length turns out to be too small to be biologically realistic. For the substitution models, especially for the Jukes–Cantor model, convergence appears to be fast enough so that the equilibrium hypothesis can be used as a realistic assumption.

2. Birth–death processes and the TKF91 indel model

The TKF91 model, or the "links" model, introduced by Thorne, Kishino, and Felsenstein in 1991 [7], comprises a substitution model and an indel model; these models can be considered independently of each other. Here we concentrate on the indel part of the TKF91 model, which describes the evolution of the sequence length. The length evolution of the sequence (consisting of nucleotides or amino acids) is represented in the TKF91 model by an infinite state-space birth–death process (due to the special form of the expression for the birth rates, the process is called birth–death process with immigration in [1]). We refer to this process as the *TKF91-BD process*. To proceed, we need a few definitions and statements concerning birth–death processes in general.

Being a special case of continuous-time Markov chain, a birth–death process $X = \{X(t), t \geqslant 0\}$ on a state space $S_N = \{0, 1, \ldots, N\}$ ($N \leqslant \infty$) can be defined by the transition rate matrix $Q = (q_{ij})$, $i,j \in S_N$, and the initial probabilities $p_i(0) = P\{X(0) = i\}$, $i \in S_N$ (for a background on continuous-time Markov chains and birth–death processes, see [18]). The matrix Q of a birth–death process has a special structure: for every $i \in S_N$, $q_{i,i+1} \geqslant 0$, $q_{i,i-1} \geqslant 0$, and all the other q_{ij} with $i \neq j$ are equal to 0. Therefore, a transition from state i can occur only to the states $i+1$ or $i-1$, $i \in S_N$. For the diagonal entries, we have $q_{ii} = -(q_{i,i+1} + q_{i,i-1})$. Thus, a birth–death process is determined by the sets of up-transition (birth) rates $\{\lambda_i\}$, $\lambda_i = q_{i,i+1}$, and down-transition (death) rates $\{\mu_i\}$, $\mu_i = q_{i,i-1}$, $i \in S_N$. We assume that $\mu_0 = 0$, and $\lambda_N = 0$ if $N < \infty$. All the other birth and death rates, maybe except λ_1, are positive real numbers. There can be broader definitions of birth–death process, but here we restrict ourselves to the one given above.

For a given birth–death processes, one of the important questions is whether it is possible to uniquely define the transition probabilities $p_{ij}(t) = P\{X(t) = j | X(0) = i\}$, $i,j \in S_N$. A sufficient condition for the uniqueness of the transition probabilities is the bound

$$\sup_{i \in S_N} (\lambda_i + \mu_i) < \infty \tag{1}$$

(see Proposition 2.9 in [18], Chapter 2). Thus, all birth–death processes on a finite state space have uniquely defined transition probabilities.

Another important question concerning the behavior of birth–death processes is whether there exists a unique stationary probability distribution $\pi = \{\pi_i\}$, such that

$$\sum_{i \in S_N} \pi_i p_{ij}(t) = \pi_j$$

for all $t \geqslant 0$, $j \in S_N$. A stationary distribution always exists for birth–death processes on finite state spaces. If the distribution π exists, then

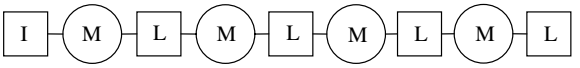

Fig. 1. The "links" model representation for a sequence consisting of 4 monomers (showed as circles). The links, including the leftmost "immortal" link, are shown by squares.

$$|p_{ij}(t) - \pi_j| \to 0 \quad \text{as } t \to \infty$$

for all $i,j \in S_N$ [18]. This is the type of convergence studied in this paper.

We now turn to the TKF91-BD model of sequence length. First, we informally describe the model in terms of the possible indel events in a sequence. A sequence can be represented as a set of monomers and connecting links; in Fig. 1, we show such a representation for a sequence of length 4. Monomers are shown as circles, while links are shown as squares. The leftmost link is the "immortal" link. Each monomer is associated with the link to the right, and an evolutionary event is either a deletion of such a monomer-link pair or an insertion of a similar pair to the right of an existing link. Note that, in such a setting, the immortal link can never be deleted. The intensity of the insertion process per link is λ, and the intensity of the deletion process per monomer is μ. Thus, for a sequence of length n, the overall insertion rate is $\lambda(n+1)$, and the overall deletion rate is μn.

The TKF91-BD process is a birth–death process on the infinite state space, with birth and death rates $\lambda_i = \lambda(i+1)$ and $\mu_i = \mu i$, $i \in S_\infty$. The states of the process correspond to different possible values of the sequence length. Applying Theorem 2.2 of ([18], Chapter 3), it is easy to prove that the transition probabilities for the TKF91 process are uniquely defined. As follows from Theorem 4.5 of ([18], Chapter 5), if $0 < \lambda < \mu$, a unique stationary distribution exists. It has the form (also given in [1])

$$\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i, \quad i \in S_\infty.$$

If $0 < \mu \leqslant \lambda$, there is no stationary distribution; the sequence length drifts to infinity as $t \to \infty$. Thus, we always assume $\lambda < \mu$. Another important characteristic of the TKF91-BD process is the existence of an explicit expression for the mathematical expectation [1]

$$E(i,t) := \sum_{j=1}^{\infty} j p_{ij}(t) = i + \left(\frac{\lambda}{\mu - \lambda} - i\right)(1 - \mathrm{e}^{-(\lambda-\mu)t}). \tag{2}$$

The expression (2) implies that, for all $i \in S_\infty$, the stationary value of the mathematical expectation is

$$E(\infty) = \frac{\lambda}{\mu - \lambda}. \tag{3}$$

We will use (2) and (3) in the following sections.

3. Convergence bounds for the TKF91-BD process

To estimate the distance between the probability distributions $p = (p_i)$ and $\tilde{p} = (\tilde{p}_i)$ on S_N, it will be convenient to use the metric $\|\cdot\|_{l_{1D}}$, which is frequently used in the studies of convergence

and stability properties of infinite state-space birth–death processes [19–22]. This metric is defined with respect to a sequence of positive real numbers $\{d_i\}$, $i \in S_N$, as follows:

$$\|p - \tilde{p}\|_{l_{1D}} = \sum_{i=1}^{N} g_i |p_i - \tilde{p}_i|,$$

where

$$g_i = \sum_{k=0}^{i-1} d_k.$$

In the general case, the l_{1D}-distance between some distributions on S_∞ may be infinite, but in all the situations that we consider the distance is finite. In what follows, we use exclusively the l_{1D} metric defined by $d_i = 1$, $i \in S_N$; we denote this metric by $\|\cdot\|$ (without a subscript). Our major goal is to obtain upper and lower bounds on the quantity $\|p(i,t) - \pi\|$, where $p(i,t)$ is the distribution of the process at time t if the process started in state i at time 0 ($i \in S_\infty$, $t \geqslant 0$).

In this work, we draw heavily upon the method of bounding the rate of convergence to stationarity for birth–death processes developed by Zeifman and coauthors [20–22]. Theorem 1 below is a direct consequence of Zeifman's results [22]. For X, assume that the condition (1) is satisfied. Let $p(t) = (p_i(t))$ and $\tilde{p}(t) = (\tilde{p}_i(t))$, $i \in S_N$, be the state probability vectors for X, corresponding to the initial distributions $p(0)$ and $\tilde{p}(0)$. Consider a sequence $\{\alpha_i\}$, $i \in S_{N-1}$, of real numbers defined by

$$\alpha_i = \lambda_i + \mu_{i+1} - \lambda_{i+1} - \mu_i, \quad i = 0, 1, \ldots, N-1, \tag{4}$$

and put

$$\alpha = \inf_i \alpha_i. \tag{5}$$

Theorem 1. *If $\alpha > 0$, then for any $p(0)$ and $\tilde{p}(0)$ satisfying $\|p(0) - \tilde{p}(0)\| < \infty$, the following inequality holds*:

$$\|p(t) - \tilde{p}(t)\| \leqslant \mathrm{e}^{-\alpha t} \|p(0) - \tilde{p}(0)\|, \quad t \geqslant 0. \tag{6}$$

Proof. Follows from Theorem 3.1 of [22] and the Remark following its proof. □

Notably, the condition (1), necessary to prove the bound (6), is not satisfied for certain practically important birth–death processes with infinite state space, including the TKF91-BD process; and Zeifman's theory cannot be directly applied. Here we show that in some cases (6) can be extended to processes for which (1) does not hold. We do so by proving a convergence bound for the TKF91-BD process.

Theorem 2. *Let X be the TKF91-BD process. For $i,j \in S_\infty$ ($i \neq j$)*

$$\|p(i,t) - p(j,t)\| \leqslant \|p(i,0) - p(j,0)\| \mathrm{e}^{-(\mu-\lambda)t} = (i+j)\mathrm{e}^{-(\mu-\lambda)t}, \quad t \geqslant 0. \tag{7}$$

Proof. Fix arbitrary $i_0, j_0 \in S_\infty$ ($i_0 \neq j_0$), and select a positive integer $K > \max(i_0, j_0)$. Consider the family $\{X^M\}$, $M = K, K+1, \ldots$, of birth–death processes on S_M, such that X^M has birth rates $\{\lambda_i^M\}$ and death rates $\{\mu_i^M\}$ defined as follows:

$$\lambda_i^M = \lambda_i, \quad \mu_i^M = \mu_i, \quad i \in S_{M-1}; \quad \lambda_M^M = 0, \quad \mu_M^M = \mu_M.$$

We first apply Zeifman's bound (6) to the processes X^M as follows. Denote $p_{ij}^M(t) = P\{X^M(t) = j | X^M(0) = i\}$, $i,j \in S_M$. For any process X^M, by (4),

$$\alpha_i = \mu - \lambda, \quad i = 0, 1, \ldots, M-2,$$

$$\alpha_{M-1} = \lambda M + \mu,$$

therefore, by (5), $\alpha = \mu - \lambda$. Also, notice that $\|p^M(i,0) - p^M(j,0)\| = i + j$, where $p^M(k,t)$ is defined for X^M in a similar way to $p(k,t)$. Thus, for each M, the inequality (6) gives

$$\sum_{i=1}^{K} i|p_{i_0 i}^M(t) - p_{j_0 i}^M(t)| \leqslant \sum_{i=1}^{M} i|p_{i_0 i}^M(t) - p_{j_0 i}^M(t)| \leqslant \mathrm{e}^{-(\mu-\lambda)t} \sum_{i=1}^{M} i|p_{i_0 i}(0) - p_{j_0 i}(0)| = (i_0 + j_0)\mathrm{e}^{-(\mu-\lambda)t};$$

for convenience, we rewrite this as follows:

$$\sum_{i=0}^{K} i|p_{i_0 i}^M(t) - p_{j_0 i}^M(t)| \leqslant (i_0 + j_0)\mathrm{e}^{-(\mu-\lambda)t}. \tag{8}$$

Consider now the family of processes $\{\hat{X}^M\}$, where $\hat{X}^M$ is a birth–death process on S_∞ whose birth and death rates coincide with those of X^M for the transitions inside S_M, and are equal to 0 for all other transitions. Thus, for $\hat{X}^M$, S_M is a closed irreducible set of states. Therefore, the transition probabilities corresponding to the transitions inside S_M coincide for X^M and $\hat{X}^M$; all other transition probabilities for $\hat{X}^M$ are 0 for all $t > 0$. We denote the birth and death rates for $\hat{X}^M$ by $\hat{\lambda}_i^M$ and $\hat{\mu}_i^M$, respectively. Obviously, $\hat{\lambda}_i^M \to \lambda_i$ and $\hat{\mu}_i^M \to \mu_i$ as $M \to \infty$. As the condition Eq. (1) holds for all $\hat{X}^M$, their transition probabilities are uniquely defined. Since all $\hat{X}^M$, as well as X, have uniquely defined transition probabilities, the sequence of processes $\{\hat{X}^M\}$ and the process X satisfy the conditions of Theorem 1 of Kurtz [23]. Hence $p_{ij}^M(t) \to p_{ij}(t)$ as $M \to \infty$ for all $i,j \in S_\infty$, $t > 0$. Passing to the limit as $M \to \infty$ in Eq. (8), we thus obtain

$$\sum_{i=1}^{K} i|p_{i_0 i}(t) - p_{j_0 i}(t)| \leqslant (i_0 + j_0)\mathrm{e}^{-(\mu-\lambda)t}. \tag{9}$$

Since (9) holds for every K such that $K > \max(i_0, j_0)$, passing to the limit as $K \to \infty$ in (9), we get

$$\|p(i_0,t) - p(j_0,t)\| \leqslant (i_0 + j_0)\mathrm{e}^{-(\mu-\lambda)t}.$$

As our choice of i_0 and j_0 was arbitrary, the theorem follows. □

Theorem 3. *If X is the TKF91-BD process, then, for all $i \in S_\infty$,*

$$\|p(i,t) - \pi\| \leqslant \left(i + \frac{\lambda}{\mu - \lambda}\right)\mathrm{e}^{-(\mu-\lambda)t}. \tag{10}$$

Proof. We have

$$\|p(i,t)-\pi\| = \sum_{j=1}^{\infty} j|p_{ij}(t)-\pi_j| = \sum_{j=1}^{\infty} j\left|\sum_{k=0}^{\infty}\pi_k p_{ij}(t) - \sum_{k=0}^{\infty}\pi_k p_{kj}(t)\right|$$
$$\leqslant \sum_{j=1}^{\infty} j \sum_{k=0}^{\infty}\pi_k|p_{ij}(t)-p_{kj}(t)|. \tag{11}$$

The series on the right-hand side of (11) converges, as can be seen from the following:

$$\sum_{j=1}^{\infty} j\sum_{k=0}^{\infty}\pi_k|p_{ij}(t)-p_{kj}(t)| \leqslant \sum_{j=1}^{\infty} j\sum_{k=0}^{\infty}\pi_k(p_{ij}(t)+p_{kj}(t)) = \sum_{j=1}^{\infty} j\sum_{k=0}^{\infty}\pi_k p_{ij}(t) + \sum_{j=1}^{\infty} j\sum_{k=0}^{\infty}\pi_k p_{kj}(t)$$
$$\leqslant \sum_{j=1}^{\infty} jp_{ij}(t) + \sum_{j=1}^{\infty} j\pi_j = E(i,t)+E(\infty) < \infty.$$

Next, using (7) and the definition of mathematical expectation, we have

$$\sum_{k=0}^{\infty}\pi_k\sum_{j=1}^{\infty} j|p_{ij}(t)-p_{kj}(t)| = \sum_{k=0}^{\infty}\pi_k\|p(i,t)-p(k,t)\| \leqslant \sum_{k=0}^{\infty}\pi_k(i+k)\mathrm{e}^{-(\mu-\lambda)t}$$
$$= (i+E(\infty))\mathrm{e}^{-(\mu-\lambda)t}. \tag{12}$$

Therefore, the series on the left-hand side of (12) is convergent. This, together with the convergence of the series on the right-hand side of (11), allows us to change the order of summation

$$\sum_{j=1}^{\infty} j\sum_{k=0}^{\infty}\pi_k|p_{ij}(t)-p_{kj}(t)| = \sum_{k=0}^{\infty}\pi_k\sum_{j=1}^{\infty} j|p_{ij}(t)-p_{kj}(t)|.$$

This expression, combined with (11) and (12), gives the bound

$$\|p(i,t)-\pi\| \leqslant (i+E(\infty))\mathrm{e}^{-(\mu-\lambda)t}. \tag{13}$$

Substituting (3) into (13), we prove the theorem. □

Theorem 4. *Let X be the TKF91-BD process. For $i \in S_\infty$*

$$\|p(i,t)-\pi\| \geqslant \left|i-\frac{\lambda}{\mu-\lambda}\right|\mathrm{e}^{-(\mu-\lambda)t}, \quad t \geqslant 0. \tag{14}$$

Proof. We have that

$$\|p(i,t)-\pi\| = \sum_{k=1}^{\infty} k|p_{ik}(t)-\pi_k| \geqslant \left|\sum_{k=1}^{\infty} kp_{ik}(t) - \sum_{k=1}^{\infty} k\pi_k\right| = |E(i,t)-E(\infty)|. \tag{15}$$

The expression (2) implies that

$$|E(i,t) - E(\infty)| = \left| i - \frac{\lambda}{\mu - \lambda} \right| e^{-(\mu-\lambda)t}.$$

Combining this with (3), we complete the proof. □

Theorems 3 and 4 allow us to obtain upper and lower convergence bounds that are uniform over i in a certain range. Such bounds are given by Proposition 1.

Proposition 1. *If* $i \leqslant \lambda/(\mu - \lambda)$, *then, for all* $j = 0, \ldots, i$,

$$\left| i - \frac{\lambda}{\mu - \lambda} \right| e^{-(\mu-\lambda)t} \leqslant \|p(j,t) - \pi\| \leqslant \left(i + \frac{\lambda}{\mu - \lambda} \right) e^{-(\mu-\lambda)t}, \quad t \geqslant 0.$$

Proof. Follows directly from Theorems 3 and 4. □

When studying convergence to stationarity, we are frequently interested in knowing how much time it would take the process to reach a certain vicinity of the equilibrium. The key value of Theorems 3 and 4 is that they allow us to obtain upper and lower estimates on this time. For given $i \in S_\infty$, define

$$\tau(i,\varepsilon) = \inf\{t > 0 : \|p(i,t) - \pi\| \leqslant \varepsilon\}.$$

Thus, $\tau(i,\varepsilon)$ is the time required for the distributions to get as close as ε in our metric $\|\cdot\|$. Theorems 3 and 4 provide the following bounds:

$$\frac{1}{\mu - \lambda} \log\left(\varepsilon^{-1} \left| i - \frac{\lambda}{\mu - \lambda} \right| \right) \leqslant \tau(i,\varepsilon) \leqslant \frac{1}{\mu - \lambda} \log\left(\varepsilon^{-1} \left(i + \frac{\lambda}{\mu - \lambda} \right) \right) \tag{16}$$

Since $\|p(i,0) - \pi\| = i + \lambda/(\mu - \lambda) - 2i\pi_i$, we consider only the case $\varepsilon \leqslant i + \lambda\ /(\mu - \lambda) - 2i\pi_i$. While the upper bound in (16) is always informative, the lower bound becomes less than 0 if $|i - \lambda\ /(\mu - \lambda)| \leqslant \varepsilon$ and, therefore, can be replaced by the trivial bound $0 \leqslant \tau(i,\varepsilon)$. In such cases, the upper bound (10), which implies the upper bound in (16), is all we have. Thus, it is necessary to investigate how tight the bound (10) is.

Having an exponential bound of the form

$$\|p(i,t) - \pi\| \leqslant C_i e^{-bt}, \tag{17}$$

we would like to investigate if the constant C_i in this bound could be made smaller, and the constant b could be made larger. If there is some b_0 such that, for any $b \geqslant b_0$, there is no such C_i that (17) holds, then b_0 can be regarded as the "true" rate of exponential convergence of $\|p(i,t) - \pi\ \|$ to 0. The ultimate goal of the convergence rate analysis is to determine such b_0. The constant b is far more important than C_i, since it determines the speed of exponential convergence to equilibrium. This becomes clear when we wish to estimate the time $\tau(i,\varepsilon)$. As follows from (17),

$$\tau(i,\varepsilon) \leqslant b^{-1} \log(C_i/\varepsilon). \tag{18}$$

Since the right-hand side of (18) depends linearly on b^{-1} and logarithmically on C_i, the influence of b is significantly greater than that of C_i. As a matter of fact, since

$$\log(2) \approx 0.69 \quad \text{and} \quad \log(10^{20}) \approx 46.05, \tag{19}$$

it is b^{-1} that determines the order of magnitude of $\tau(i,\varepsilon)$. Thus, if our estimate of b is near b_0, we are likely to be able to estimate the order of magnitude of $\tau(i,\varepsilon)$ accurately, despite the fact that the corresponding C_i (such that (17) holds for b, C_i) may be huge. For the lower bound on $\tau(i,\varepsilon)$, such as the one on the left-hand side of (16), the dependency on the pre-exponential constant (in our case, $|i-\lambda/(\mu-\lambda)|$) will also be logarithmic, so that the order of magnitude of $\tau(i,\varepsilon)$ will be determined by the exponential rate.

Returning to the TKF91-BD process, it is easy to see that the rate $\mu-\lambda$ of exponential convergence cannot be increased. If we could increase it, then for some $t'>0$ we would have

$$\|p(i,t') - \pi\| < |i-\lambda/(\mu-\lambda)|\mathrm{e}^{-(\mu-\lambda)t'},$$

which is impossible (by Theorem 4). Thus, for the TKF91 process, Eq. (10) is a sharp upper bound of the form (17). Since the rates of exponential convergence of the upper and lower bounds coincide, we come to the conclusion that the lower bound also has the exact rate of exponential convergence, and in this sense is tight.

It is also important to note that, for any fixed $t \geqslant 0$, the lower bound Eq. (14) approaches the upper bound (10) as $i \to 0$. Accordingly, our bounds are especially tight for small i, and become exact for $i=0$. This property is advantageous for the analysis of sequence data because there exists compelling evidence that the early, ancestral proteins tended to be quite short [24]. In fact, it is conceivable to model sequence length evolution setting the initial length equal to 0, in which case we account for the "birth" of the sequence.

4. Substitution models and their convergence

In the TKF91 model, substitutions are described by a continuous-time Markov chain on the state space S_N, with $N=4$ in the case of nucleotide sequences, and $N=20$ in the case of amino acid sequences. While the primary focus of this paper is to assess the speed of convergence for the TKF91-BD process, it is of interest, for the sake of comparison, to investigate the convergence of the substitution models as well. Here we discuss the convergence of the well-known Jukes–Cantor [25] and Kimura [26] models of nucleotide substitution (for a description of these models, see also [27]). Similar studies can be carried out for Markov chain models of amino acid substitution in protein sequences; a method of constructing the transition rate matrix for such models is described in [28] (see also Chapter 8 of [29]).

As above, our main goal is to estimate the distance between the distributions $p(i,t)$ and π, which are defined for finite Markov chains in a similar way to the case of birth–death processes. While the metric $\|\cdot\|$ utilized in the preceding sections is well-suited for birth–death processes on infinite state space, for finite probability vectors $p=(p_i)$ and $\tilde{p}=(p_i)$, $i \in S_N$, a more simple and natural distance is the l_1-distance defined by

$$\|p-\tilde{p}\|_{l_1} = \sum_{i\in S_N} |p_i - \tilde{p}_i|.$$

This metric is frequently applied in the studies of stability and convergence of finite Markov chains [20,22,30], and this is the metric we use for the substitution models.

The case of finite Markov chains is much simpler compared to infinite state-space birth–death models, because we can always compute the distributions using standard numerical procedures. The situation is even better for the Jukes–Cantor and Kimura models which are analytically tractable, meaning that it is possible to obtain explicit expressions for $\|p(i,t)-\pi\|_{l_1}$. The Jukes–Cantor model is defined by the single parameter ρ, the state-independent nucleotide substitution rate (the states correspond to the 4 nucleotides). All the stationary probabilities are equal to 1/4. Using the known expressions for the transition probabilities of the model [27], it is easy to show that, for $i=1,\ldots,4$,

$$\|p(i,t)-\pi\|_{l_1}=\frac{3}{2}\mathrm{e}^{-4\rho t}. \tag{20}$$

For the Kimura model, all the stationary probabilities are also equal to 1/4, and

$$\|p(i,t)-\pi\|_{l_1}=\frac{3}{4}\mathrm{e}^{-4\beta t}+\frac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t}+\left|\frac{1}{4}\mathrm{e}^{-4\beta t}-\frac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t}\right|,\quad i=1,\ldots,4, \tag{21}$$

where α and β are the rates of transitions (in the genetics sense) and transversions, respectively.

5. Results and discussion

In this section we apply our bounds to estimate the rate of convergence for the TKF91-BD process with parameters derived from biological sequence data. The methodology underlying our analysis is as follows. We calculate the upper and lower bounds for the distance $\|p(i,t)-\pi\|$, and use them to estimate the values $\tau(i,\varepsilon)$ for different ε. Since we know that the convergence bounds give the correct order of magnitude for $\|p(i,t)-\pi\|$, we will be able to judge whether or not convergence with accuracy ε occurs on realistic timescales. For ε, we use the following general definition

$$\varepsilon=\omega I, \tag{22}$$

where $I=i+\lambda/(\mu-\lambda)-2i\pi_i$ is the initial value for $\|p(i,t)-\pi\|$, and ω is a small positive number. Such a definition is justified by the idea that convergence can be interpreted in the relative sense: we can regard convergence accuracy as sufficient if the norm $\|p(i,t)-\pi\|$ is reasonably small compared to $\|p(i,0)-\pi\|$. Thus, we say that a $100\omega\%$ convergence was achieved at time $\leqslant t$ if

$$\|p(i,t)-\pi\|\leqslant\omega\|p(i,0)-\pi\|.$$

To estimate the parameters of the TKF91 model, Hein et al. [1] applied the TKF91 indel model to the human α and β globins, and obtained the estimates $\lambda=0.03718$, $\mu=0.03744$. The same estimates were obtained by Knudsen and Miyamoto [10]. The units for these estimates were indels per substitution. As can be inferred from Fig. 5 of [1], the authors assumed the absolute substitution rate to be (approximately) 1 substitution per site per 10^9 years (Byr). This number agrees with the known estimates for globins [31]. Thus, we have enough data to apply our results to estimate convergence. The rate of exponential convergence is $\mu-\lambda=0.00026$. (The equilibrium sequence length for the rates as above is $\lambda/(\mu-\lambda)\approx 143$ amino acids. This is very close to the observed lengths of 141 (α) and 146 (β) of actual proteins, as was noticed by Hein and coauthors

[1].) The dependency of the upper and lower bounds (10) and (14) on t for different i (initial sequence lengths) is shown in Fig. 2; the possibility to obtain bounds uniform over i is provided by Proposition 1. As was discussed above, in the case of exponential convergence, the convergence rate is mostly determined by the exponent, and the exponents in our bounds are exact. Therefore, we can expect the distribution of the process to be close to the stationary distribution when the upper and lower bounds are close to each other, compared to the distance between the bounds at time 0. If at time t the distance between the upper and lower bounds is almost as large as the corresponding distance at time 0, then we should expect that at time t the distribution of the process is likely to be far from equilibrium. It is obvious from Fig. 2 that the convergence of the upper and lower bounds can be noticed only after 100 Byr.

We now characterize convergence via the convergence times $\tau(i,\varepsilon)$. As follows from general considerations, if the distance $\|p(i,t)-\pi\|$ is not less than 25% of the initial distance, we can hardly speak of convergence to equilibrium. Fig. 3 shows the dependencies of the upper and lower bounds on $\tau(i,\varepsilon)$ in (18) on ω via (22) for different initial sequence lengths. We chose the initial length values assuming that the ancestral sequence is shorter than the actual globin. It is easy to see that for all initial sequence lengths the 25% convergence time is greater than 100 Byr. The relatively weak influence of the initial sequence length on the upper bound in (16) is illustrated by Fig. 4; a similar picture can be generated for the lower bound. Note that the dependence on ε is also relatively weak, since ε enters the upper bound in (16) under the logarithm.

One of the main difficulties with using the equilibrium hypothesis within the TKF91 framework stems from the relationship between the rate of exponential convergence and the equilibrium sequence length. Indeed, if we know that the equilibrium length of a sequence should be greater than a, then the expression (3) for the equilibrium length gives us the inequalities

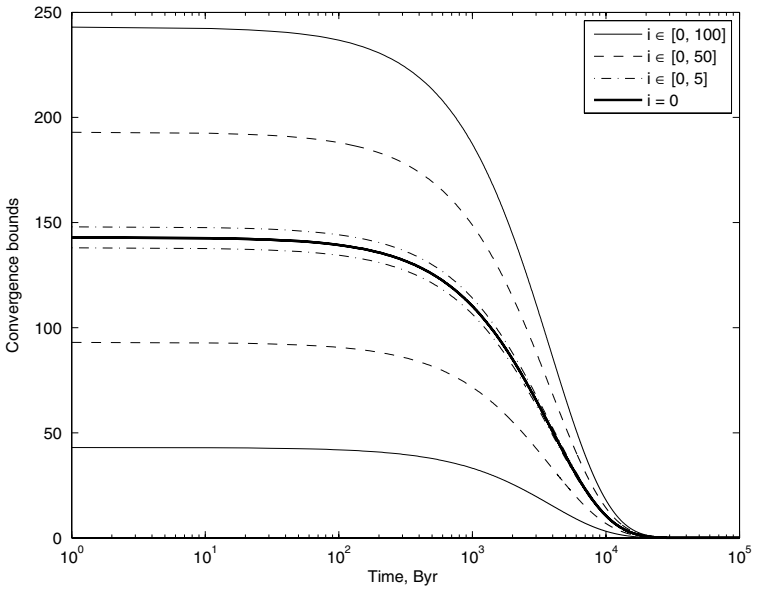

Fig. 2. Upper (10) and lower (14) convergence bounds for the TKF91-BD process; i is the initial sequence length. The bounds are uniform over i belonging to the intervals specified in the figure legend.

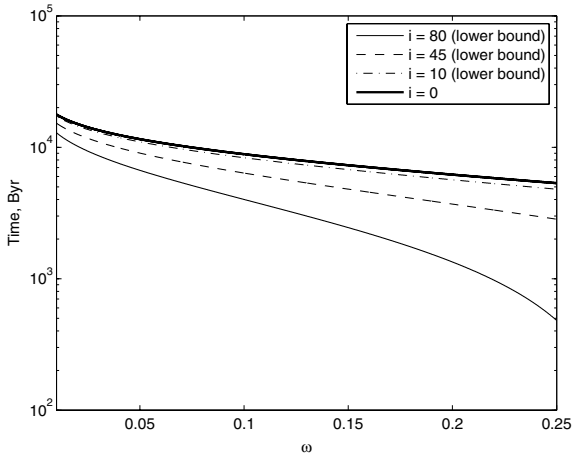

Fig. 3. Dependency of the upper and lower bounds in (16) on ω (via (22)) for protein sequences (see text for details); i is the initial sequence length. In this figure, the upper bounds are indistinguishable from the exact bound for $i = 0$.

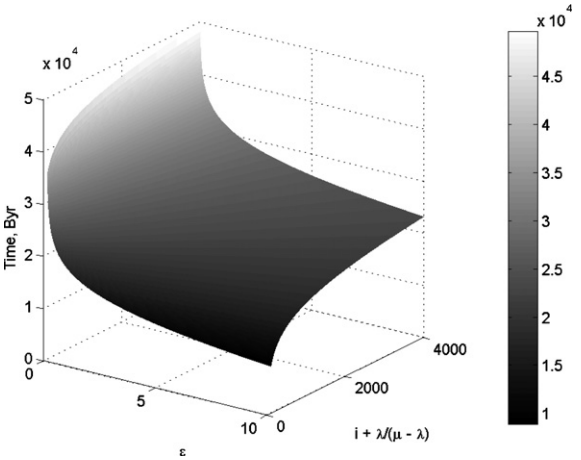

Fig. 4. Dependency of the upper bound in (16) on ε and $i + \lambda/(\mu - \lambda)$ for protein sequences.

$$\frac{a}{a+1} < \frac{\lambda}{\mu} < 1.$$

Since $a/(a+1)$ approaches 1 for large a, if a is large, λ/μ is very close to 1, in which case the convergence rate, $\mu - \lambda$, will necessarily be small. Conversely, if the rate of convergence is large, then λ/μ and, therefore, a, will be small. The conclusion is that it may be quite difficult to have realistic convergence rates and realistic equilibrium sequence lengths at the same time. An example of slow convergence for realistic equilibrium length was given above. Now we illustrate the complementary situation by an analysis of the human DNA evolution.

Indel mutation rates in human processed pseudogenes were estimated by Ophir and Graur [32]. They found that an insertion occurs once per every 100 substitutions, and a deletion occurs once

per every 40 substitutions. The average insertion size is approximately 5 nt, and the average deletion size is 8 nt. We shall treat all indel events as one-nucleotide indels; this is definitely a stretch, but it is necessary to make one in order to use the TKF91-BD process as a model for the length evolution of real sequences. We thus have about 5 insertions per 100 substitutions, and 8 deletions per every 40 substitutions. To infer the insertion and deletion rates, we need an estimate for substitution rates in pseudogenes. This rate has been estimated to be 3.9 substitutions per site per Byr [33], which implies that we can take $\lambda = 0.195$ and $\mu = 0.78$ for the TKF91-BD process as a model of pseudogene length evolution. The corresponding rate of exponential convergence is $\mu - \lambda = 0.585$, which is three orders of magnitude larger than the convergence rates for proteins discussed above. In a similar way to the protein sequences, it can be shown that 25% convergence for i on the order of 1000 can be achieved in approximately 2 Byr, which is biologically realistic. However, the equilibrium length of the sequence is $\lambda/(\mu - \lambda) \approx 0.33$ nt, which is obviously too small to reflect any reasonable pseudogene length (the average length of a human processed pseudogene is close to 1100 nt [34]).

Now we consider the convergence rates for substitution models. For the Jukes–Cantor model, we take the exponential convergence parameter $\rho = 3.9$, as above. The plot of the distance between the distributions, calculated according to formula (20), is given in Fig. 5. It is natural to say that the inequality $\|p(i,t) - \pi\|_{l_1} \leqslant 0.05$ means that the chain is sufficiently close to equilibrium: indeed, the mean absolute difference per state probability, $\|p(i,t) - \pi\|_{l_1}/4$, is on the order of 0.01. This degree of closeness to equilibrium is attained after about 230 Myr (millions of years) of evolution. After 100 Myr of evolution, we have $\|p(i,t) - \pi\|_{l_1}/4 \approx 0.08$, which is probably close enough to equilibrium. As an example of Kimura's model, we can consider the one with transition rate 1.71 and transversion rate 1.22 events per site per Byr, as was estimated by comparison of human and rodent protein-coding sequences [35]. The plot of the distance between the distributions for this model, calculated according to (21), is given in Fig. 5. As we can see, convergence occurs at a smaller rate than for the Jukes–Cantor model; $\|p(i,1) - \pi\|_{l_1} \approx 0.008$, which indicates sufficient closeness to equilibrium at approximately 1 Byr.

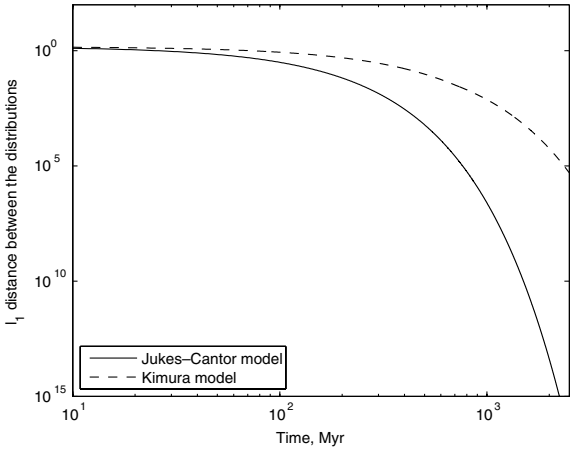

Fig. 5. Convergence to equilibrium for Markov chain substitution models: the l_1-distance for the Jukes–Cantor model (Eq. (20)) and for the Kimura model (Eq. (21)).

The TFK91 model is based on the following main assumptions: (a) the size of all insertions and deletions is 1 nucleotide (amino acid); (b) insertion and deletion rates do not change over time; (c) the ancestral sequences have equilibrium length. Our results show that, for the considered examples, these assumptions, applied simultaneously, contradict biological reality. The 25% convergence time of 100 Byr that we stated above is clearly beyond biological reality, since this time period exceeds the Earth's age. Of course, if the process has started from the equilibrium distribution at time 0, it would have been in equilibrium at all times. However, this would be possible only if the sequence evolution process has started before the time 0 from some specific state and has had enough time to converge to the equilibrium distribution; starting from equilibrium by pure chance appears to be highly improbable. By considering the sequence length distributions which are fully concentrated in a single state, we assumed that the time 0 is the very beginning of the sequence evolution process, so that there is no reason to expect any prior convergence.

We showed that if the convergence rate for the TKF91 model is high enough, the resulting equilibrium sequence length can become too small to describe any real sequence. It is plausible that taking into account larger-size insertions and deletions, as well as inhomogeneity in the insertion and deletion rates, would decrease the convergence time and keep the equilibrium length reasonable. But if the TKF91 model is used in its original form, the findings regarding sequence evolution obtained under the equilibrium hypothesis should be treated with care. It is noteworthy that applicability of the equilibrium hypothesis for the practical purposes of statistical sequence alignment depends on the sensitivity of the methods in question to deviations from this hypothesis. This issue requires further investigation.

Our analysis showed that, for the substitution models, the equilibrium assumption is considerably more realistic. Our estimate of convergence time for nucleotide sequences for the Kimura model was 1 Byr. If the ancestral sequence has existed for 1 Byr before the modern sequences have originated from it, then the equilibrium assumption can be regarded as justified. For DNA sequences of some ancient proteins, such as ribosomal proteins, this can be true, since it was estimated that life on the Earth began more than 3 Byr ago [36]. For the Jukes–Cantor model, with the convergence time of a few hundreds of Myr, the equilibrium assumption is applicable in a much wider variety of situations.

Our overall conclusion is that the TKF91 model in combination with equilibrium hypothesis fails to capture certain important characteristics of the process of length evolution of biological sequences. Further investigations are needed to develop more realistic evolution models that would match the tractability and practical convenience of the TFK91 model. Finally, we would like to point out that the methods of convergence rate estimation for birth–death process developed in this paper are quite general. The formalism of birth–death processes is an important tool of mathematical modeling in biology [37]. Our Theorems 2–4 can be generalized to other birth–death processes, thus providing means of convergence analysis and validity verification for a class of biologically important mathematical models.

Acknowledgments

The authors are grateful to Dr. Svetlana Ekisheva for valuable comments on the manuscript. This work was supported in part by the NIH Grant HG00783 to M.B.

References

[1] J. Hein, C. Wiuf, B. Knudsen, M.B. Moller, G. Wibling, Statistical alignment: computational properties, homology testing and goodness-of-fit, J. Mol. Biol. 302 (2000) 265.
[2] I. Holmes, Using evolutionary expectation maximization to estimate indel rates, Bioinformatics 21 (2005) 2294.
[3] I. Holmes, W.J. Bruno, Evolutionary HMMs: a Bayesian approach to multiple alignment, Bioinformatics 17 (2001) 803.
[4] D. Metzler, Statistical alignment based on fragment insertion and deletion models, Bioinformatics 19 (2003) 490.
[5] I. Miklós, G.A. Lunter, I. Holmes, A "long indel" model for evolutionary sequence alignment, Mol. Biol. Evol. 21 (2004) 529.
[6] E. Rivas, Evolutionary models for insertions and deletions in a probabilistic modeling framework, BMC Bioinformatics 6 (2005), Paper No: 63.
[7] J.L. Thorne, H. Kishino, J. Felsenstein, An evolutionary model for maximum-likelihood alignment of DNA sequences, J. Mol. Evol. 33 (1991) 114.
[8] J.L. Thorne, H. Kishino, J. Felsenstein, Inching toward reality—an improved likelihood model of sequence evolution, J. Mol. Evol. 34 (1992) 3.
[9] J. Hein, J.L. Jensen, C.N.S. Pedersen, Recursions for statistical multiple alignment, Proc. Natl. Acad. Sci. USA 100 (2003) 14960.
[10] B. Knudsen, M.M. Miyamoto, Sequence alignments and pair hidden Markov models using evolutionary history, J. Mol. Biol. 333 (2003) 453.
[11] G.A. Lunter, I. Miklós, Y.S. Song, J. Hein, An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees, J. Comp. Biol. 10 (2003) 869.
[12] I. Miklós, An improved algorithm for statistical alignment of sequences related by a star tree, Bull. Math. Biol. 64 (2002) 771.
[13] J.L. Thorne, G.A. Churchill, Estimation and reliability of molecular sequence alignments, Biometrics 51 (1995) 100.
[14] I. Holmes, A probabilistic model for the evolution of RNA structure, BMC Bioinformatics 5 (2004), Paper No: 166.
[15] D. Metzler, R. Fleissner, A. Wakolbinger, A. von Haeseler, Assessing variability by joint sampling of alignments and mutation rates, J. Mol. Evol. 53 (2001) 660.
[16] M. Steel, J. Hein, Applying the Thorne–Kishino–Felsenstein model to sequence evolution on a star-shaped tree, Appl. Math. Lett. 14 (2001) 679.
[17] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, Sunderland, MA, 2004.
[18] W.J. Anderson, Continuous-Time Markov Chains: An Applications-Oriented Approach, Springer, New York, 1991.
[19] B.L. Granovsky, A. Zeifman, Nonstationary queues: estimation of the rate of convergence, Queueing Syst. 46 (2004) 363.
[20] B.L. Granovsky, A.I. Zeifman, Nonstationary Markovian queues, J. Math. Sci. 99 (2000) 1415.
[21] A.I. Zeifman, Some estimates of the rate of convergence for birth and death processes, J. Appl. Probab. 28 (1991) 268.
[22] A.I. Zeifman, Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes, Stoch. Process. Appl. 59 (1995) 157.
[23] T.G. Kurtz, A note on sequences of continuous parameter Markov chains, Ann. Math. Stat. 40 (1969) 1078.
[24] C.P. Ponting, R.R. Russell, The natural history of protein domains, Ann. Rev. Biophys. Biomol. Struct. 31 (2002) 45.
[25] T.H. Jukes, C.R. Cantor, Evolution of protein molecules, in: H.N. Munro (Ed.), Mammalian Protein Metabolism, Academic Press, New York, 1969, p. 21.
[26] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, J. Mol. Evol. 16 (1980) 111.
[27] R. Durrett, Probability Models for DNA Sequence Evolution, Springer, New York, 2002.
[28] H. Kishino, T. Miyata, M. Hasegawa, Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, J. Mol. Evol. 31 (1990) 151.

[29] M. Borodovsky, S. Ekisheva, Problems and Solutions in Biological Sequence Analysis, Cambridge University, Cambridge, 2006.
[30] A.Y. Mitrophanov, Stability and exponential convergence of continuous-time Markov chains, J. Appl. Prob. 40 (2003) 970.
[31] M. Kimura, T. Ohta, On some principles governing molecular evolution, Proc. Natl. Acad. Sci. USA 71 (1974) 2848.
[32] R. Ophir, D. Graur, Patterns and rates of indel evolution in processed pseudogenes from humans and murids, Gene 205 (1997) 191.
[33] S.J. Fleishman, T. Dagan, D. Graur, pANT: A method for the pairwise assessment of nonfunctionalization times of processed pseudogenes, Mol. Biol. Evol. 20 (2003) 1876.
[34] C. Chen, A.J. Gentles, J. Jurka, S. Karlin, Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22, Proc. Natl. Acad. Sci. USA 99 (2002) 2930.
[35] D.W. Collins, T.H. Jukes, Rates of transition and transversion in coding sequences since the human-rodent divergence, Genomics 20 (1994) 386.
[36] W.M. Fitch, F.J. Ayala, Tempo and mode in evolution, Proc. Natl. Acad. Sci. USA 91 (1994) 6717.
[37] A.S. Novozhilov, G.P. Karev, E.V. Koonin, Biological applications of the theory of birth-and-death processes, Briefings Bioinform. 7 (2006) 70.