An Expectation-Maximization Algorithm for Analysis of Evolution of Exon-Intron Structure of Eukaryotic Genes

Liran Carmel, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin

National Center for Biotechnology Information,
National Library of Medicine,
National Institutes of Health,
Bethesda, Maryland 20894, USA
{carmel, rogozin, wolf, koonin}@ncbi.nlm.nih.gov

Abstract. We propose a detailed model of evolution of exon-intron structure of eukaryotic genes that takes into account gene-specific intron gain and loss rates, branch-specific gain and loss coefficients, invariant sites incapable of intron gain, and rate variability of both gain and loss which is gamma-distributed across sites. We develop an expectation-maximization algorithm to estimate the parameters of this model, and study its performance using simulated data.

1 Introduction

Spliceosomal introns are one of the most prominent idiosyncrasies of eukaryotic genomes. They are scattered all over the eukaryota superkingdom, including, notably, species that are considered basal eukaryotes, such as *Giardia lamblia* [1]. This suggests that evolution of introns is intimately entangled with eukaryotic evolution; thus, the study of evolution of exon-intron structure of eukaryotic genes, apart from being interesting in its own right, might shed some light on the still enigmatic rise of eukaryotes. For example, one of the notorious, long-lasting unresolved issues in evolution of eukaryotic genomes is the intron-early versus intron-late debate. Proponents of the intron-early hypothesis posit that introns were prevalent at the earliest stages of cellular evolution and played a crucial role in the formation of complex genes via the mechanism of exon shuffling [2]. These introns were inherited by early eukaryotes but have been eliminated from prokaryotic genomes as a result of selective pressure for genome streamlining. By contrast, proponents of the intron-late hypothesis hold the view that introns had emerged, de novo, in early eukaryotes, and subsequent evolution of eukaryotes involved extensive insertion of new introns (see, e.g., [3,4]).

Various anecdotal studies have demonstrated certain features of intron evolution. But it was not until the accumulation of genomic information in the recent years that large-scale analyses became feasible. Such analyses yielded at least three different models of intron evolution. One model assumes parsimonious evolution [5]; another assumes a simple gene-specific gain/loss model and

A. McLysaght et al. (Eds.): RECOMB 2005 Ws on Comparative Genomics, LNBI 3678, pp. 35–46, 2005.

analyzes it using Bayesian learning [6]; and yet another one assumes a simple branch-specific gain/loss model on three-species phylogenetic topology and analyzes it using direct maximum likelihood [7]. It seems that none of these models is sufficiently general, and each neglects different aspects of this complex evolutionary process. This is reflected in the major contradictions between the predictions laid out by the three models. For example, the gene-specific model [6] predicts an intron-poor eukaryotic ancestor and a dominating intron gain process; the branch-specific model [7] predicts an intron-rich eukaryotic ancestor and a dominating loss process; while the parsimonious model [5] is somewhat in between, predicting intermediate densities of introns in early eukaryotes, and a gain-dominated kaleidoscope of gain and loss events.

Here, we introduce a model of evolution of exon-intron structure, which is considerably more realistic than previously proposed models. The model accounts for gene-specific intron gain/loss mechanisms, branch-specific gain/loss mechanisms, invariant sites (a fraction of sites that are incapable of intron gain), and rate distribution across sites of both intron-gain and intron-loss. Using data from extant species, we follow the popular approach of estimating the model parameters by way of maximum likelihood. Direct maximization of the likelihood is, however, intractable in this case due to a large number of hidden random variables in the model. These are exactly the circumstances under which the expectation-maximization (EM) algorithm for maximizing the likelihood might prove itself useful. None of the software packages that we are aware of, either using direct maximization or EM, can deal with our proposed model. Hence, we devised an EM algorithm tailored to our particular model. As this model is rather detailed, a variety of biologically-reasonable models can be derived as special cases. For this reason, we anticipate a broad range of applicability to our algorithm, beyond its original use. In the following we describe our model of exon-intron structure evolution and an EM algorithm for learning its parameters.

2 The Evolutionary Model

Suppose that we have multiple alignments of G different genes from S eukaryotic species, and let our observed data be the projection, upon the above alignments, of a presence-absence intron map. That is, at every site in each species we can observe either zero (absence of an intron), one (presence of an intron), or $\star$ (missing value, indicating lack of knowledge about intron's presence or absence). Let us define a *pattern* as any column in an alignment, and let $\Omega \leq 3^S$ be the total number of unique observed patterns, indexed as $\omega_1, \ldots, \omega_\Omega$. We shall use n_{gp} to denote the number of patterns ω_p that are observed in gene g.

Let the rooted phylogeny of the above S species be given by an N-node binary tree, where $S = (N+1)/2$. Let $q_0, \ldots, q_{N-1}$ be the nodes of this tree, with the convention that q_0 is the root node. We use the notations q^L, q^R and q^P to describe the left-descendant, right-descendant and parent, respectively, of node q (left and right are set arbitrarily). Also, let $\mathcal{L}(q)$ stand for the set of terminal nodes (leaves) that are descendants of q. We index the branches of the tree by the node into which they lead, and use $\varDelta_q$ for the length of the branch (in time units) leading into node q. Hereinafter, we assume that the tree topology, as well as the branch lengths $\varDelta_1, \ldots, \varDelta_{N-1}$, are known.

Assume that the root node has a prior probability π_i of being at state i ($i = 0, 1$), and that the transition matrix for gene g along branch t, $A^g_{ij}(q_t) = P(q_t = j | q_t^P = i)$, is described by

$$A^g(q_t) = \begin{pmatrix} 1 - \xi_t(1 - e^{-\eta_g \varDelta_t}) & \xi_t(1 - e^{-\eta_g \varDelta_t}) \\ 1 - (1 - \phi_t)e^{-\theta_g \varDelta_t} & (1 - \phi_t)e^{-\theta_g \varDelta_t} \end{pmatrix}, \tag{1}$$

where η_g and θ_g are gene-specific gain and loss rates, respectively, and ξ_t and ϕ_t are branch-specific gain and loss coefficients, respectively.

The common practice in evolutionary studies is to incorporate rate distribution across sites by associating each site with a *rate coefficient*, r, which scales the branch lengths of the corresponding phylogenetic tree, $\varDelta_t \leftarrow r \cdot \varDelta_t$. This rate coefficient is drawn from a probability distribution with non-negative domain and unit mean, typically the unit-mean gamma distribution. Such an approach is compatible with the notion that each site has a characteristic evolutionary rate. This, however, should be modified for intron evolution, where the gain and loss processes do not seem to be correlated. That is, sites that are fast to gain introns are not necessarily fast to lose them, and vice versa. Therefore, we model rate variation using two independent rate coefficients, r^η and r^θ, such that $\eta_g \leftarrow r^\eta \cdot \eta_g$ and $\theta_g \leftarrow r^\theta \cdot \theta_g$. These rates are independently drawn from the two distributions

$$\begin{aligned} r^\eta &\sim \nu\delta(\eta) + (1 - \nu)\varGamma(\eta; \lambda) \\ r^\theta &\sim \varGamma(\theta; \lambda). \end{aligned} \tag{2}$$

Here, $\varGamma(x; \lambda)$ is the unit-mean gamma distribution of variable x with shape parameter λ, $\delta(x)$ is the Dirac delta-function, and ν is the fraction of sites that are invariant to gain (i.e., sites that are incapable of gaining introns). Two comments are in order with respect to these rate distributions. First, a site can be invariant only with respect to gain, in accord with the proto-splice site hypothesis that presumes preferential gain of introns at distinct sites [8]. In contrast, once an intron is gained, it can always be lost. Second, we assumed the same shape parameter for the gamma distributions of both gain and loss. This is done solely to simplify the already complex model. At a later stage, we may consider extending the model to include different shape parameters.

3 The EM Algorithm

Phylogenetic trees can be interpreted as Bayesian networks that depict an underlying evolutionary probabilistic model. Accordingly, the terminal nodes form the observed random variables of the model, and the internal nodes form the hidden random variables. Under this view, estimating the model parameters using EM is natural. Indeed, different EM algorithms have been applied to phylogenetic trees with various purposes [9–11]. The algorithm that resembles the one described here most closely was developed by Siepel & Haussler [12] and used for branch length optimization and parameter estimation of time-continuous Markovian processes. However, our model does not fit into any of the existing schemes as it includes several unique properties, such as the branch-specific coefficients, the gain-invariant sites, and the different treatment of rate variability across sites. In the rest of this section, we develop the algorithm in the context of the proposed model; we attempt to do so using notations that are as general as possible, in order to allow the use of this algorithm with other models as well.

Denote by $\mathcal{N}_g = (n_{1g}, \ldots, n_{\Omega g})$ the counts of all observed patterns in the gth alignment, and by Θ the set of model parameters. We will use, whenever necessary, the decomposition $\Theta = (\Xi, \Psi, \Lambda)$ where $\Xi = (\Xi_1, \ldots, \Xi_{N-1})$ is the set of branch-specific parameters, $\Xi_t = (\xi_t, \phi_t)$ in our case, characterized by not being affected by the rate variability; $\Psi = (\Psi_1, \ldots, \Psi_G)$ is the set of gene-specific variables, $\Psi_g = (\eta_g, \theta_g)$ in our case, characterized by being subject to rate variability, and $\Lambda = (\nu, \lambda)$ is the set of rate variables. We assume independence between genes and between sites, hence the likelihood function is

$$L(\mathcal{N}_1, \ldots, \mathcal{N}_G|\Theta) = \prod_{g=1}^{G} L(\mathcal{N}_g|\Xi, \Psi_g, \Lambda) = \prod_{g=1}^{G}\prod_{p=1}^{\Omega} L(\omega_p|\Xi, \Psi_g, \Lambda)^{n_{gp}}, \quad (3)$$

and the log-likelihood is just

$$\log L(\mathcal{N}_1, \ldots, \mathcal{N}_G|\Theta) = \sum_{g=1}^{G}\sum_{p=1}^{\Omega} n_{gp} \log L(\omega_p|\Xi, \Psi_g, \Lambda). \quad (4)$$

To make the rate distributions (2) amenable to *in silico* manipulations, we rendered them discrete as was done previously by Yang [13], using K categories for the gamma distribution, and an additional category for the invariant sites. For the time being, we will keep our notations general and will not specify the rendering technique, and in particular, will not assume equi-probable categories. Accordingly, r^θ can take the values $(r_1^\theta, \ldots, r_K^\theta)$ with probabilities $(f_1^\theta, \ldots, f_K^\theta)$, and r^η can take the values $(r_1^\eta = 0, r_2^\eta, \ldots, r_{K+1}^\eta)$ with probabilities $(f_1^\eta = \nu, f_2^\eta, \ldots, f_{K+1}^\eta)$. Introducing rate variability across sites is equivalent to transforming the model into a mixture model, with the rates determining the mixture coefficients. Consequently, we will associate with each site two discrete random variables, ρ_p^η and ρ_p^θ, indicating the rate category of η and θ, respectively. According to the EM paradigm, we are guaranteed to climb up-hill in $\log L(\omega_p|\Xi, \Psi_g, \Lambda)$, if we maximize the auxiliary function

$$\begin{aligned} &Q_{gp}(\Xi, \Psi_g, \Lambda, \Xi^0, \Psi_g^0, \Lambda^0) = \quad (5)\\ &\quad = \sum_{\sigma, \rho_p^\eta, \rho_p^\theta} P(\sigma, \rho_p^\eta, \rho_p^\theta|\omega_p, \Xi^0, \Psi_g^0, \Lambda^0) \log P(\omega_p, \sigma, \rho_p^\eta, \rho_p^\theta|\Xi, \Psi_g, \Lambda) = \end{aligned}$$

$$= \sum_{\sigma,\rho_p^\eta,\rho_p^\theta} P(\sigma, \rho_p^\eta, \rho_p^\theta | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) \cdot$$
$$\cdot \sum_{k=1}^{K+1} \sum_{k'=1}^{K} 1_{\{\rho_p^\eta = k\}} 1_{\{\rho_p^\theta = k'\}} \left[\log f_k^\eta + \log f_{k'}^\theta + \log P(\omega_p, \sigma | \Xi, \Psi_{gkk'})\right] .$$

Here, σ is any realization of the internal nodes of the tree, $1_{\{\rho=k\}}$ is a function that takes the value 1 when $\rho = k$ and takes the value zero otherwise, and $\Psi_{gkk'}$ is the set of effective gene-specific rates which, in our model, is $\Psi_{gkk'} = (\eta_{gk}, \theta_{gk'})$, where we have introduced the notations $\eta_{gk} = r_k^\eta \cdot \eta_g$ and $\theta_{gk'} = r_{k'}^\theta \cdot \theta_g$. If we now use

$$P(\sigma, \rho_p^\eta = k, \rho_p^\theta = k' | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) = \qquad (6)$$
$$= P(\rho_p^\eta = k, \rho_p^\theta = k' | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) \cdot P(\sigma | \omega_p, \Xi^0, \Psi_{gkk'}^0)$$

in (5), we get

$$Q_{gp}(\Xi, \Psi_g, \Lambda, \Xi^0, \Psi_g^0, \Lambda^0) = \qquad (7)$$
$$\sum_{k=1}^{K+1} \sum_{k'=1}^{K} \left[\sum_{\rho_p^\eta,\rho_p^\theta} P(\rho_p^\eta, \rho_p^\theta | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) \cdot 1_{\{\rho_p^\eta = k\}} 1_{\{\rho_p^\theta = k'\}} \right] \cdot$$
$$\cdot \left[\sum_{\sigma} P(\sigma | \omega_p, \Xi^0, \Psi_{gkk'}^0) \left[\log f_k^\eta + \log f_{k'}^\theta + \log P(\omega_p, \sigma | \Xi, \Psi_{gkk'})\right] \right] .$$

Denoting by $w_{gpkk'}$ and $Q_{gpkk'}$ the first and second square brackets, respectively, the auxiliary function maximization of which assures increasing the likelihood is

$$Q = \sum_{g=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K+1} \sum_{k'=1}^{K} n_{gp} w_{gpkk'} Q_{gpkk'} . \qquad (8)$$

3.1 The E-Step

Here is how we compute $w_{gpkk'}$ and $Q_{gpkk'}$ for the current estimate Θ^0 of the model parameters.

$$w_{gpkk'} = \sum_{\rho_p^\eta,\rho_p^\theta} P(\rho_p^\eta, \rho_p^\theta | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) 1_{\{\rho_p^\eta = k\}} 1_{\{\rho_p^\theta = k'\}} = \qquad (9)$$
$$= P(\rho_p^\eta = k, \rho_p^\theta = k' | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) =$$
$$= \frac{P(\rho_p^\eta = k | \Xi^0, \Psi_g^0, \Lambda^0) \cdot P(\rho_p^\theta = k' | \Xi^0, \Psi_g^0, \Lambda^0) \cdot P(\omega_p | \Xi^0, \Psi_{gkk'}^0)}{\sum_{h,h'} P(\rho_p^\eta = h | \Xi^0, \Psi_g^0, \Lambda^0) \cdot P(\rho_p^\theta = h' | \Xi^0, \Psi_g^0, \Lambda^0) \cdot P(\omega_p | \Xi^0, \Psi_{ghh'}^0)} =$$
$$= \frac{(f_k^\eta)^0 (f_{k'}^\theta)^0 P(\omega_p | \Xi^0, \Psi_{gkk'}^0)}{\sum_{h,h'} (f_h^\eta)^0 (f_{h'}^\theta)^0 P(\omega_p | \Xi^0, \Psi_{ghh'}^0)} .$$

The function $P(\omega_p|\varXi^0, \varPsi^0_{gkk'})$ is the likelihood of the tree that we rapidly compute using a variant of Felsenstein's pruning algorithm [14]. To this end, let us define $\gamma^{gpkk'}(q) = P(\mathcal{L}(q)|q^P, \varXi^0, \varPsi^0_{gkk'})$, which is the probability of observing those terminal nodes that are descendants of q, for a given state of the parent of q. Omitting the superscripts for clarity, this function is initialized at all terminal nodes $q_t \in \mathcal{L}(q_0)$ by

$$\gamma(q_t) = \begin{cases} \begin{pmatrix} 1 - \xi_t(1 - e^{-\eta_{gk}\varDelta_t}) \\ 1 - (1 - \phi_t)e^{-\theta_{gk'}\varDelta_t} \end{pmatrix} & s_t = 0 \\ \begin{pmatrix} \xi_t(1 - e^{-\eta_{gk}\varDelta_t}) \\ (1 - \phi_t)e^{-\theta_{gk'}\varDelta_t} \end{pmatrix} & s_t = 1, \end{cases} \tag{10}$$

where s_t is the value observed at q_t. Then, γ is computed at all internal nodes (except for the root) using the inward-recursion

$$\gamma_i(q_t) = \sum_{j=0}^{1} A^g_{ij}(q_t)\tilde{\gamma}_j(q_t), \tag{11}$$

where $\tilde{\gamma}_j(q)$ is an abbreviation for $\gamma_j(q^L)\gamma_j(q^R)$. The likelihood of the tree is then

$$P(\omega_p|\varXi^0, \varPsi^0_{gkk'}) = \sum_{i=0}^{1} \pi_i\tilde{\gamma}_i(q_0). \tag{12}$$

Using this in (9) allows us to compute the coefficients $w_{gpkk'}$. In order to compute the coefficients $Q_{gpkk'}$ we need a complementary recursion to the above γ-recursion. To this end, let us define $\alpha^{gpkk'}(q, q^P) = P(q, q^P|\omega_p, \varXi^0, \varPsi^0_{gkk'})$. Again, omitting the superscripts, this function can be initialized on the two descendants of the root by

$$\alpha(q, q_0) = \frac{1}{P(\omega_p|\varXi^0, \varPsi^0_{gkk'})} \begin{cases} \begin{pmatrix} \pi_0\gamma_0(q^S)A^g_{00}(q) & 0 \\ \pi_1\gamma_1(q^S)A^g_{10}(q) & 0 \end{pmatrix} & q \in \mathcal{L}(q_0), \quad s = 0 \\ \begin{pmatrix} 0 & \pi_0\gamma_0(q^S)A^g_{01}(q) \\ 0 & \pi_1\gamma_1(q^S)A^g_{11}(q) \end{pmatrix} & q \in \mathcal{L}(q_0), \quad s = 1 \\ \begin{pmatrix} \pi_0\gamma_0(q^S)\tilde{\gamma}_0(q)A^g_{00}(q) & \pi_0\gamma_0(q^S)\tilde{\gamma}_1(q)A^g_{01}(q) \\ \pi_1\gamma_1(q^S)\tilde{\gamma}_0(q)A^g_{10}(q) & \pi_1\gamma_1(q^S)\tilde{\gamma}_1(q)A^g_{11}(q) \end{pmatrix} \\ & q \notin \mathcal{L}(q_0). \end{cases} \tag{13}$$

Here, q is a descendent of the root (either q_0^R or q_0^L), and q^S is its sibling. For any other internal node, α is computed using the outward-recursion

$$\alpha(q, q^P) = \begin{pmatrix} \frac{\tilde{\gamma}_0(q)}{\gamma_0(q)}\beta_0(q^P)A^g_{00}(q) & \frac{\tilde{\gamma}_1(q)}{\gamma_0(q)}\beta_0(q^P)A^g_{01}(q) \\ \frac{\tilde{\gamma}_0(q)}{\gamma_1(q)}\beta_1(q^P)A^g_{10}(q) & \frac{\tilde{\gamma}_1(q)}{\gamma_1(q)}\beta_1(q^P)A^g_{11}(q) \end{pmatrix}, \tag{14}$$

where $\beta(q) = P(q|\omega_p, \varXi^0, \varPsi^0_{gkk'}) = \sum_{q^P} \alpha(q, q^P)$ is computed for each node subsequently to the computation of α. Finally, for each terminal node that is not a descendant of the root,

$$\alpha(q,q^P)=\begin{cases}\begin{pmatrix}\beta_0(q^P)\ 0\\ \beta_1(q^P)\ 0\end{pmatrix} & s=0\\ \begin{pmatrix}0\ \beta_0(q^P)\\ 0\ \beta_1(q^P)\end{pmatrix} & s=1.\end{cases} \tag{15}$$

This inward-outward recursion is the phylogenetic equivalent of the backward-forward recursion known from hidden Markov models, and other versions of it have already been developed, see, e.g., [9,12]. We shall now see how the α's and β's allow us to compute the coefficients $Q_{gpkk'}$. Notice that, if we use the state variables as indices, we can replace the function $\log P(\omega_p,\sigma|\Xi,\Psi_{gkk'})$ in (7) by

$$\log P(\omega_p,\sigma|\Xi,\Psi_{gkk'})=\sum_{i=0}^{1}(q_0)_i\log\pi_i+\sum_{i,j=0}^{1}\sum_{t=1}^{N-1}(q_t)_j(q_t^P)_i\log A_{ij}^g(q_t). \tag{16}$$

Denote the expectation over $P(\sigma|\omega_p,\Xi^0,\Psi^0_{gkk'})$ by E_σ. Applying it to (16) we get

$$\begin{aligned}&E_\sigma\left[\log P(\omega_p,\sigma|\Xi,\Psi_{gkk'})\right]= \\ &=\sum_{i=0}^{1}\log\pi_i E_\sigma[(q_0)_i]+\sum_{i,j=0}^{1}\sum_{t=1}^{N-1}\log A_{ij}^g(q_t)E_\sigma[(q_t)_j(q_t^P)_i].\end{aligned} \tag{17}$$

But, $E_\sigma[(q_0)_i]=P(q_0=i|\omega_p,\Xi^0,\Psi^0_{gkk'})=\beta_i(q_0)$, and similarly $E_\sigma[(q_t)_j(q_t^P)_i]=\alpha_{ij}(q_t,q_t^P)$, so that $Q_{gpkk'}$ can be finally written as

$$\begin{aligned}Q_{gpkk'}&=\sum_\sigma P(\sigma|\omega_p,\Xi^0,\Psi^0_{gkk'})\cdot \\ &\cdot\left[\log f_k^\eta+\log f_{k'}^\theta+\log P(\omega_p,\sigma|\Xi,\Psi_{gkk'})\right]= \\ &=\log f_k^\eta+\log f_{k'}^\theta+\sum_{i=0}^{1}\beta_i(q_0)\log\pi_i+\sum_{i,j=0}^{1}\sum_{t=1}^{N-1}\alpha_{ij}(q_t,q_t^P)\log A_{ij}^g(q_t).\end{aligned} \tag{18}$$

One of the appealing features of EM is that is allows, in many cases, to treat missing data fairly easily. In our case, two simple modifications are required for this. Firstly, we have to add to the γ-recursion initialization (10) an option

$$\gamma(q_t)=\begin{pmatrix}1\\1\end{pmatrix}\qquad s_t=\star. \tag{19}$$

Secondly, we have to add to the α-recursion finalization (15) an option

$$\alpha(q_t)=\begin{pmatrix}\beta_0(q_t^P)A_{00}^g(q_t)\ \beta_0(q_t^P)A_{01}^g(q_t)\\ \beta_1(q_t^P)A_{10}^g(q_t)\ \beta_1(q_t^P)A_{11}^g(q_t)\end{pmatrix}\qquad s_t=\star. \tag{20}$$

3.2 The M-Step

Substituting the expressions for $w_{gpkk'}$ and $Q_{gpkk'}$ in (8), we obtain the final form of the function to be maximized at each iteration. Explicitly, this is

$$
\begin{aligned}
Q = & \sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K+1}\sum_{k'=1}^{K} n_{gp}w_{gpkk'}(\log f_k^{\eta} + \log f_{k'}^{\theta}) + \\
& + \sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K+1}\sum_{k'=1}^{K} n_{gp}w_{gpkk'}\left[\beta_0^{gpkk'}(q_0)\log\pi_0 + \beta_1^{gpkk'}(q_0)\log\pi_1\right] + \\
& + \sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K+1}\sum_{k'=1}^{K}\sum_{t=1}^{N-1} n_{gp}w_{gpkk'}\alpha_{00}^{gpkk'}(q_t)\log\left[1-\xi_t(1-e^{-\eta_{gk}\Delta_t})\right] + \\
& + \sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K+1}\sum_{k'=1}^{K}\sum_{t=1}^{N-1} n_{gp}w_{gpkk'}\alpha_{01}^{gpkk'}(q_t)\left[\log\xi_t + \log(1-e^{-\eta_{gk}\Delta_t})\right] + \\
& + \sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K+1}\sum_{k'=1}^{K}\sum_{t=1}^{N-1} n_{gp}w_{gpkk'}\alpha_{10}^{gpkk'}(q_t)\log\left[1-(1-\phi_t)e^{-\theta_{gk'}\Delta_t}\right] + \\
& + \sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K+1}\sum_{k'=1}^{K}\sum_{t=1}^{N-1} n_{gp}w_{gpkk'}\alpha_{11}^{gpkk'}(q_t)\left[\log(1-\phi_t) - \theta_{gk'}\Delta_t\right].
\end{aligned}
\tag{21}
$$

It is well-known that any increase in Q suffices to climb up-hill in the likelihood, and therefore it is not of utmost importance to maximize it precisely. Hence, we do not invest too much in finding precise maximum, but rather use low-tolerance maximization with respect to each of the parameters individually. Since it is easy to differentiate Q twice with respect to all the parameters (except for λ), we use the Newton-Raphson zero-finding algorithm for the maximization. Due to space limitations and because the derivation is, essentially, trivial, we do not present them here.

We must, however, devote a few words to the maximization of Q with respect to λ. In (21) we kept the rate distributions general, but (2) imposes the constraints $r_k^{\theta} = r_{k+1}^{\eta}$. Furthermore, in rendering the gamma distribution discrete, we assume equi-probable categories, thus

$$
f_{k+1}^{\eta} = (\nu - 1)f_k^{\theta} = \frac{\nu - 1}{K} \qquad k = 1, \ldots, K. \tag{22}
$$

Therefore, Q depends on λ through r^{η} and r^{θ}, making analytic differentiation impossible. Thus, in this case, we used Brent's maximization algorithm that does not require derivatives.

4 Validation

We intend to apply the algorithm to real data, namely, an amended version of the data set from [5], which consists of multiple alignments of over 700 orthologous genes from 8 eukaryotic species. However, prior to its application to real data, the algorithm must be carefully validated against simulated data. Thus, we have written simulation software that performs three tasks. Firstly, given the number of extant species, it builds a random phylogeny. Secondly, it assigns random lengths to the branches based on the exponential distribution (keeping the tree balanced). Thirdly, it draws the model parameters subject to some biologically plausible constraints. Given the phylogenetic tree and the model parameters, we then simulate any desired number of evolutionary scenarios, collecting the observations on the terminal nodes.

While EM algorithms always converge to a maximum of the likelihood, they are not guaranteed to find the global maximum. In practice, however, we have strong indications that our EM algorithm is highly effective in finding the global maximum. We cannot provide a proof for this, but at least it is clear that it always estimates model parameters that give a higher likelihood than the true model parameters, see Figure 1.

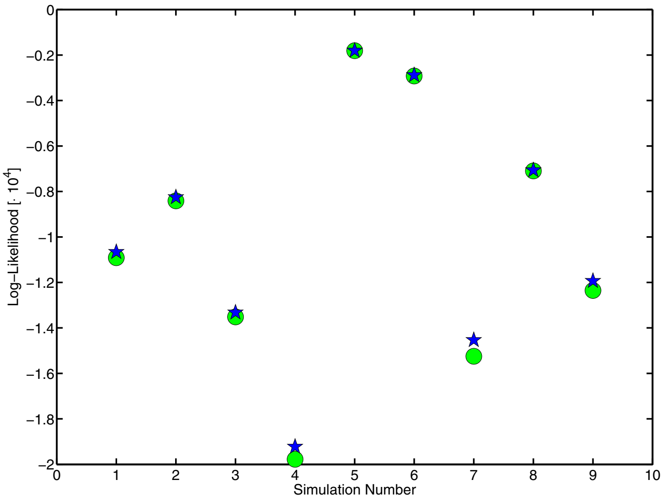

Fig. 1. Summary of 9 independent simulations. For each simulation, a 4 species random phylogeny spanning 400 million years and a set of model parameters were drawn randomly. Intron evolution was simulated for four multigenes of mean length of 5000 AA, with no rate variation. Parameters were estimated using tolerance of 10^{-2}. The dots indicate log-likelihood values computed for the true model parameters, and the pentagons indicate log-likelihood values computed for the estimated parameters. Note that the log-likelihood of the estimated parameters is always greater than that of the true parameters.

A well known property of maximum likelihood estimators is that they are not guaranteed to be unbiased for any finite sample size. In our model, and

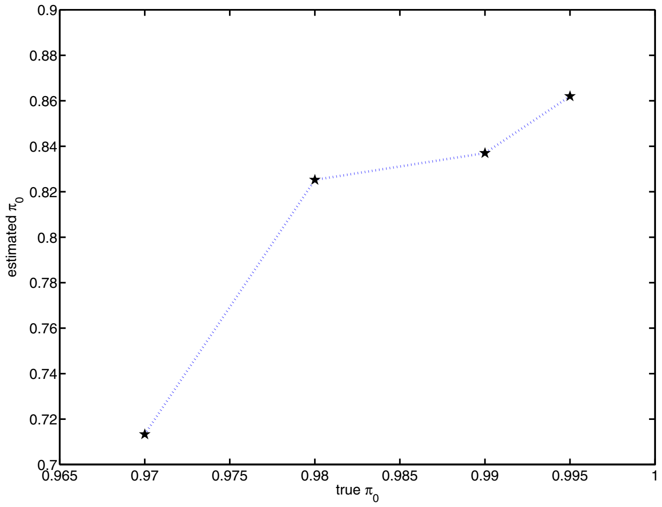

Fig. 2. Estimated π_0 versus true π_0.Each dot is the mean of three simulations of six-species random phylogeny spanning 300 million years. In each simulation, we assumed four multigenes of mean length of 50,000 AA, with no rate variation.

probably in other phylogenetic models, the bias might be significant, mainly due to the small number of species and to the paucity of informative patterns. An example is shown in Figure 2, where the probability π_0 of the root node is estimated. This problem can be less severe when a monotonic relation between the true parameter and the estimated one holds (Figure 2). We are currently investigating different approaches to map this bias more accurately.

5 Discussion

We describe here an algorithm that allows for parameter estimation of an evolutionary model for exon-intron structure of eukaryotic genes. Once estimated, these parameters could help resolving the current debate regarding evolution of introns, in particular, with regard to the relative contributions of intron loss and gain in different eukaryotic lineages.

Some of the assumptions of our model are worth discussion. Specifically, in Equations (3) and (4), we assumed that different sites evolve (i.e., gain and lose introns) independently. However, several observations show that such independence is only an approximation. First, introns in intron-poor species tend to cluster near the 5' end of the gene [15,16]. Second, adjacent introns tend to be lost in concert [16,17]. Nevertheless, it seems that such site-dependence of gain and loss is a secondary factor in intron evolution. First, non-homogeneous spatial distribution of introns along the gene is pronounced only in species with a low number of introns. Second, some anecdotal studies could not find any preference of adjacent introns to be lost together (e.g., [18].) Should subsequent studies indicate that the dependence between sites is more important than we currently envisage, our model probably can be extended using the context-dependent ideas developed in [12].

Similarly, in Equations (3) and (4), we assumed that different genes gain and lose introns independently. Currently, we are unaware of any strong evidence for such dependence, but if it is discovered, it can be easily accounted for in our model by concatenating genes with similar evolutionary trends and treating them as a single multigene.

Additionally, we assumed the same shape parameter for the gamma distribution of intron gain and loss rates. As mentioned above, this assumption was taken out of convenience, and due to the general impression that the exact shape of the gamma distribution is not a primary factor. However, our model can be rather easily extended to incorporate different shape parameters for gain and loss.

The computational complexity of the algorithm is, in the worst case, $O(G \cdot S \cdot K^2 \cdot 3^S)$. The exponential dependency arises because the number of unique patterns, Ω, is exponential with the number of species. However, if W is the total number of sites in all the alignments, it bounds Ω by $\Omega \leq \min(W, 3^S)$, thus keeping us, in practice, far away off the worst case.

The current Matlab® code is too slow to handle efficiently the real data (over two million sites) and the massive simulations. Therefore, we are in the process of writing the code in C++, allowing for its application to large data sets. The C++ software will be made available as soon as it is ready.

Acknowledgements

The authors would like to thank Jacob Goldberger for useful discussions regarding aspects of the EM algorithm.

References

1. Nixon, J. E., Wang, A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J., Samuelson, J.: A Spliceosomal Intron in Giardia Lamblia. Proc. Natl. Acad. Sci. USA **99** (2002) 3701–3705.
2. Gilbert, W.: The Exon Theory of Genes. Cold Spring Harb. Symp. Quant. Biol. **52** (1987) 901–905.
3. Cho, G., Doolittle, R.F.: Intron Distribution in Ancient Paralogs Supports Random Insertions and Not Random Loss. J. Mol. Evol. **44** (1997) 573–584.
4. Lynch, M.: Intron Evolution as a Population-genetic Process. Proc. Natl. Acad. Sci. USA **99** (2002) 6118–6123.
5. Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., Koonin, E.V.: Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution. Curr. Biol. **13** (2003) 1512–1517.
6. Qui, W.-G., Schisler, N., Stoltzfus, A.: The Evolutionary Gain of Spliceosomal Introns: Sequence and Phase Preferences. Mol. Biol. Evol. **21** (2004) 1252–1263.

7. Roy, S.W., Gilbert, W.: Complex Early Genes. Proc. Natl. Acad. Sci. USA **102** (2005) 1986–1991.
8. Dibb, N.J.: Proto-Splice Site Model of Intron Origin. J. Theor. Biol. **151** (1991) 405–416.
9. Friedman, N., Ninio, M., Pe'er I., Pupko, T.: A Structural EM Algorithm for Phylogenetic Inference. J. Comput. Biol. **9** (2002) 331–353.
10. Holmes, I.: Using Evolutionary Expectation Maximisation to Estimate Indel Rates. Bioinformatics **21** (2005) 2294–2300.
11. Brooks, D. J., Fresco, J. R., Singh, M.: A Novel Method for Estimating Ancestral Amino Acid Composition and Its Application to Proteins of the Last Universal Ancestor. Bioinformatics **20** (2004) 2251–2257.
12. Siepel, A., Haussler, D.: Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. Mol. Biol. Evol. **21** (2004) 468–488.
13. Yang, Z.: Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. J. Mol. Evol. **39** (1994) 306–314.
14. Felsenstein, J.: Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. J. Mol. Evol. **17** (1981) 368–376.
15. Mourier, T, Jeffares, D.C.: Eukaryotic Intron Loss. Science **300** (2003) 1393.
16. Sverdlov, A.V., Babenko, V.N., Rogozin, I.B., Koonin, E.V.: Preferential Loss and Gain of Introns in 3' Portions of Genes Suggests a Reverse-Transcription Mechanism of Intron Insertion. Gene **338** (2004) 85–91.
17. Roy, S.W., Gilbert, W.: The Pattern of Intron Loss. Proc. Natl. Acad. Sci. USA **102** (2005) 713–718.
18. Cho, S., Jin, S.-W., Cohen, A., Ellis, R.E.: A Phylogeny of Caenorhabditis Reveals Frequent Loss of Introns During Nematode Evolution. Genome Res. **14** (2004) 1207–1220.