Chapter 16

A Maximum Likelihood Method for Reconstruction of the Evolution of Eukaryotic Gene Structure

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

Abstract

Spliceosomal introns are one of the principal distinctive features of eukaryotes. Nevertheless, different large-scale studies disagree about even the most basic features of their evolution. In order to come up with a more reliable reconstruction of intron evolution, we developed a model that is far more comprehensive than previous ones. This model is rich in parameters, and estimating them accurately is infeasible by straightforward likelihood maximization. Thus, we have developed an expectation-maximization algorithm that allows for efficient maximization. Here, we outline the model and describe the expectation-maximization algorithm in detail. Since the method works with intron presence–absence maps, it is expected to be instrumental for the analysis of the evolution of other binary characters as well.

Key words: Maximum likelihood, expectation-maximization, intron evolution, ancestral reconstruction, eukaryotic gene structure.

1. Introduction

In eukaryotes, many protein-coding genes have their coding sequence broken into pieces – the exons – separated by the non-coding spliceosomal introns. These introns are removed from the nascent pre-mRNA and the exons are spliced together to form the intronless mRNA by the spliceosome, a large and elaborate macromolecular complex comprising several small RNA molecules and numerous proteins. No spliceosomal introns have ever been found in prokaryotes, and there are no eukaryotes with a completely sequenced genomes, not even the very basal ones, which would not possess introns *(1–3)* and the accompanying splicing machinery *(4)*.

Jason McDermott et al. (eds.), *Computational Systems Biology, vol. 541*

DOI 10.1007/978-1-59745-243-4_16

Despite the introns being such a remarkable idiosyncrasy of eukaryotic genomes, their origin and evolution are not thoroughly understood *(5, 6)*. It is generally accepted that introns can be regarded as units of evolution and that their presence/absence pattern is a result of stochastic processes of loss and gain. However, the nature of these processes is vigorously debated. Recent large-scale attempts to study these processes using extant eukaryotic genomes led to incongruent conclusions.

In a study on reconstruction of intron evolution, Rogozin et al. *(7)* analyzed ~700 sets of intron-bearing orthologous genes from eight eukaryotic species. The multiple alignment of the orthologs within each set was computed, and the intron positions were projected on the alignments to form presence/absence maps. Using Dollo parsimony to infer ancestral states, these authors observed a diverse repertoire of behaviors. Some lineages endured extensive losses, while others experienced mostly gain events. Early forbearers, such as the last common ancestor of multicellular life, were shown to be relatively intron-rich. This work suggested that both gain and loss of introns played significant roles in shaping the modern eukaryotic gene structure. However, as these inferences rely upon the Dollo parsimony reconstruction, the number of gains in terminal branches (leaves of the phylogenetic tree) is overestimated, resulting in underestimation (potentially, significant) of the number of introns in ancient lineages.

The same data set was analyzed by Roy and Gilbert *(8, 9)* using a different methodology. They adopted a simple evolutionary model, according to which different lineages are associated with different loss and gain probabilities. Using a variation on maximum likelihood estimation, they obtained considerably higher estimates for the number of introns in early eukaryotes and a correspondingly lower level of gains in all lineages, i.e., a clear dominance of loss events in the evolution of eukaryotic genes. Roy and Gilbert have substantially simplified the mathematics involved in the estimation procedure, at the expense of introducing into the computation considerations of parsimony, which yielded an inference technique that is a hybrid between parsimony and maximum likelihood. This hybrid, however, excludes from consideration different evolutionary scenarios, resulting in inflated estimates of the number of introns in early eukaryotes *(10)*.

The model of Roy and Gilbert is *branch-specific*, i.e., it assumes that the gain and loss rates depend only on the branch, thus tacitly presuming that all genes behave identically with respect to intron gain and loss. Exactly the inverse approach was adopted by Qiu et al. *(11)*. These authors developed a *gene-specific* model, whereby different gene families are characterized by different rates of intron gain and loss, but for a particular gene these rates are constant across the entire phylogenetic tree. They used a different data set combined with a Bayesian estimation technique and concluded

that almost all extant introns were gained during the eukaryotic evolution. This suggests evolution is dominated by intron gain events with few losses. However, the validity of a gene-specific model is disputable as it is hard to reconcile with the accumulating evidence on large differences between lineages *(12–15)*.

Recently, two maximum likelihood estimation techniques have been developed for essentially the same branch-specific evolutionary model as the one of Roy and Gilbert. Csuros *(10)* used a direct approach, while Nguyen et al. *(16)* developed an expectation-maximization algorithm. Both methods encountered the same problem of estimating the number of unobserved intronless sites. Each employed a technically different but conceptually similar method to evaluate this number. Both techniques were applied to the eight-species data of Rogozin et al. *(7)*, yielding very close estimates. As expected, these methods predict intron occupancy level of ancient lineages higher than those predicted by Dollo parsimony and lower than those predicted by the hybrid technique of Roy and Gilbert. Notably, these estimates are generally closer to those obtained using Dollo parsimony, and they imply an evolutionary landscape comprising both losses and gains, with some excess of gains.

While the Dollo parsimony *(7)* and the hybrid technique of Roy and Gilbert *(8, 9)* showed some methodological biases, the other analyses of intron evolution *(10, 11, 16)* used well-established estimation techniques. Nevertheless, these studies kept yielding widely diverging inferences. The reason seems to be the differences in the underlying evolutionary models, neither being sufficient to describe the complex reality of intron evolution. The branch-specific model fails to account for important differences between genes, whereas the gene-specific model ignores the sharp differences between lineages. Additionally, rate variability between sites, known to be an important factor in other fields of molecular evolution *(17, 18)*, should be taken into account also in the evolution of gene structure. This is particularly important for intron gain in light of the accumulating evidence in favor of the proto-splice model, according to which new introns are preferentially inserted inside certain sequence motifs *(19–21)*. This means that sites could dramatically differ in their gain rate depending on their position relative to a proto-splice site.

Here we describe a model of evolution that takes into consideration all of the above factors. In order to efficiently estimate the model parameters by maximum likelihood, we have developed an expectation-maximization algorithm. We also compiled a data set that is considerably larger than previously used ones, consisting of 400 sets of orthologous genes from 19 eukaryotic species. Applying our algorithm to this data set, we obtained high-precision estimates, revealing a fascinating evolutionary history of gene structure, where both losses and gains played significant roles

albeit the contribution of losses was somewhat greater. Moreover, we identified novel properties of intron evolution: (i) all eukaryotic lineages share a common, universal, mode of intron evolution, whereby the loss and gain processes are positively correlated. This suggests that the mechanisms of intron gain and loss share common mechanistic components. In some lineages, additional forces come into play, resulting either in elevated loss rate or in elevated gain rate. Lineages exhibiting an increased loss rate are dispersed throughout the entire phylogenetic tree. In contrast, lineages with excessive gains are much rarer, and all of them are ancient. (ii) Intron loss rates of individual genes show no correlation with any other genomic property. By contrast, intron gain rate of individual genes show several remarkable relationships, not always easily explained. In brief, intron gain rate is positively correlated with expression level, negatively correlated with sequence evolution rate, and negatively correlated with the gene length. Moreover, genes of apparent bacterial origin have significantly lower rates of intron gain than genes of archaeal origin. (iii) We showed that the remarkable conservation of intron positions is, mainly (~90%), due to shared ancestry, and only in a minority of the cases (~10%), due to parallel gain at the same location. (iv) We determined that the density of potential intron insertion sites is about 1 site per 7 nucleotides.

2. Materials

The algorithm learns the parameters of the model by comparing the structure of orthologous genes in extant species. To carry out this comparison, it requires two sets of input data, to be described in this section. The first is a phylogenetic tree, defining topological relationships between a set of eukaryotic species. The second is a collection of genes, for which one can identify orthologs in at least a subset of the species above.

2.1. Multiple Alignments

Suppose that we have G sets of aligned orthologous genes from S species. To represent the gene structure, we transform these alignments into intron presence–absence maps by substituting for each nucleotide (or amino acid) 0 or 1, depending on whether an intron is present or absent in the respective position. We allow for missing data by using a third symbol (*), and consequently a gene might be included in the input data even if it is missing in part of the species. Every site in an alignment, called *pattern*, is a vector of length S over the alphabet (0,1,*). Let Ω be the total number of unique patterns in the entire set of G alignments, denoted $\omega_1, \ldots, \omega_\Omega$, and let n_{gp} count the number of times pattern ω_p is found in the multiple alignment of gene g. Assuming that the sites evolve

independently, the set $M_g = (n_{g1}, \ldots, n_{g\Omega})$ fully characterizes the multiple alignment of the gth gene. Thus, all the relevant information about the multiple alignments is captured by the list of unique patterns $\omega_1, \ldots, \omega_\Omega$, and the list of vectors $M_1, \ldots, M_G$.

2.2. Phylogenetic Tree

Let T be a rooted bifurcating phylogenetic tree with S leaves (terminal nodes) corresponding to the S species above. The total number of nodes in T is $N = 2S - 1$, and we index them by $t = 0, 1, \ldots, N - 1$, with the convention that zero is the root node. The state of node t is described by the random variable q_t, which can take the values 0 and 1 (and * in leaves). We use V_t for the set of all leaves such that node t is among their ancestors. The entire collection of leaves is, obviously, V_0. The parent node of t is denoted $\mathrm{P}(t)$. We use the special notations q_t^P and V_t^P for $q_{\mathrm{P}(t)}$ and $V_{\mathrm{P}(t)}$, respectively. Analogously, the two direct descendents of node t are denoted $\mathrm{L}(i)$ and $\mathrm{R}(i)$, and we use the special notations q_t^L, q_t^R, V_q^L, and V_q^R for $q_{\mathrm{L}(t)}$, $q_{\mathrm{R}(t)}$, $V_{\mathrm{L}(t)}$, and $V_{\mathrm{R}(t)}$, respectively. We index the branches by the node into which they are leading, and use Δ_t to denote the length (in time units) of the tth branch. We assume that the tree topology, as well as all the branch lengths $\Delta_1, \ldots, \Delta_{N-1}$ are known.

3. Methods

3.1. The Probabilistic Model

A graphical model is a mathematical graph whose nodes symbolize random variables, and whose branches describe dependence relationships between them *(22)*. A bifurcating phylogenetic tree, when viewed as a graphical model, depicts the probabilistic model

$$\Pr(q_0) \prod_{t=1}^{N-1} \Pr(q_t | q_t^P). \qquad [1]$$

We use the notation $\pi_i = \Pr(q_0 = i)$ to describe the prior probability of the root, and $A_{ij}(g, t) = \Pr(q_t = j | q_t^P = i, g)$ to describe the transition probability for gene g along branch t. In our model, we assume that the transition probability depends on both the gene and the branch, and that it takes the explicit form

$$A(g, t) = \begin{pmatrix} 1 - \xi_t(1 - e^{-\eta_g \Delta_t}) & \xi_t(1 - e^{-\eta_g \Delta_t}) \\ 1 - (1 - \phi_t)e^{-\theta_g \Delta_t} & (1 - \phi_t)e^{-\theta_g \Delta_t} \end{pmatrix}. \qquad [2]$$

Here, η_g and θ_g are nonnegative parameters, determining the intron gain and loss rates, respectively, of gene g. Complementarily, ξ_t and ϕ_t determine the intron gain and loss coefficients of branch t, respectively, and are bound to the range $0 \leq \xi_t, \phi_t \leq 1$.

The probability of an intron present in gene g at the beginning of branch t to be retained along the branch is $(1-\phi_t)e^{-\theta_g \Delta_t}$, that is, it is retained only if the branch does not lose it (with probability $1-\phi_t$), and also the gene does not lose it (with probability $e^{-\theta_g \Delta_t}$). This comes to reflect a reality where strong forces to strip a gene off its introns will be practically unaffected by the particular lineage, and, oppositely, strong forces to strip a lineage off its introns will be practically unaffected by the particular gene. In the same spirit, the probability of an intron to be gained in gene g along branch t is $\xi_t(1-e^{-\eta_g \Delta_t})$, that is, it is gained only if both the branch "approves" it (with probability ξ_t) and the gene "approves" it (with probability $1-e^{-\eta_g \Delta_t}$).

In other fields of molecular evolution, it was long realized that analysis precision improves if one allows for rate variability across sites *(17, 18)*. Typically, such rate variability is modeled by introducing a *rate variable*, r, which scales, for each site, the time units of the phylogenetic tree, $\Delta_t \leftarrow r \cdot \Delta_t$. This rate variable is a random variable, distributed according to a distribution function with non-negative domain and unit mean, typically the unit-mean gamma distribution. The rate variability reflects the idea that sites differ in their rate of evolution. Specifically, there are fast-evolving sites ($r>>1$), as well as slow-evolving ones ($r<<1$). In our model of intron evolution we extend this idea by assuming that the gain and loss processes are subject to rate variability, independently of each other. Hence, a site can have any combination of gain and loss rates. To accommodate this idea, we use two independent rate variables, r^η and r^θ, that are used to scale, for each site, the gene-specific gain rate, $\eta_g \leftarrow r^\eta \cdot \eta_g$, and the gene-specific loss rate, $\theta_g \leftarrow r^\theta \cdot \theta_g$. We further assume that the distributions of these rate variables are independent of the genes, and are explicitly given by

$$\begin{aligned} r^\eta &\sim v\delta(\eta) + (1-v)\Gamma(\eta;\lambda_\eta) \\ r^\theta &\sim \Gamma(\theta;\lambda_\theta). \end{aligned} \qquad [3]$$

Here, $\Gamma(x;\lambda)$ is the unit-mean gamma distribution of variable x with shape parameter λ, $\delta(x)$ is the Dirac delta-function, and v is the fraction of sites that are assumed to have zero gain rate. These latter sites, denoted *invariant sites*, reflect these sites that are not a proto-splice site *(19–21)*. Intron loss does not have an invariant counterpart, as the assumption is that once an intron is gained, it can always be lost. Therefore, the loss rate variable is assumed to be distributed according to a gamma distribution, which is by far the most popular in describing rate variability *(17, 18, 23)*.

In practice, the rate distributions in **Eq. [3]** are rendered discrete *(24)*. We assume that the gain rate variable can take K_η discrete values $r_1^\eta = 0, r_2^\eta, \ldots, r_{K_\eta}^\eta$ with probabilities $f_1^\eta = v, f_2^\eta \ldots, f_{K_\eta}^\eta$ such

that $\sum_{k=1}^{K_\eta} f_k^\eta = 1$. Analogously, we assume that the loss rate variable can take K_θ discrete values $r_1^\theta, \ldots, r_{K_\theta}^\theta$ with probabilities $f_1^\theta, \ldots, f_{K_\theta}^\theta$ such that $\sum_{k=1}^{K_\theta} f_k^\theta = 1$. For a particular gain rate value r_k^η, we denote the actual gain rate $r_k^\eta \cdot \eta_g$ by η_{kg}. Similarly, for a particular loss rate value r_k^θ, we denote the actual loss rate $r_k^\theta \cdot \theta_g$ by θ_{kg}.

For notational clarity, we aggregate the model parameters into a small number of sets. To this end, let $\Xi_t = \{\xi_t, \phi_t\}$ be the set of parameters that are specific for branch t, and let $\Xi = (\Xi_1, \ldots, \Xi_{N-1})$ be the set of all branch-specific parameters. Similarly, let $\Psi_g = (\eta_g, \theta_g)$ be the set of parameters that are specific for gene g, and let $\Psi = (\Psi_1 \ldots, \Psi_G)$ be the set of all gene-specific parameters. Additionally, we denote by $\Lambda = (\nu, \lambda_\eta, \lambda_\theta)$ the parameters that determine the rate variability. When the distinction between the different sets of parameters is irrelevant, we shall use $\Theta = (\Xi, \Psi, \Lambda)$ as the set of all the model's parameters. We achieve further succinctness in notations by denoting the actual gene-specific rate values for particular values r_k^η and $r_{k'}^\theta$ of the rate variables as $\Psi_{kk'g} = (\eta_{kg}, \theta_{k'g})$.

3.2. The EM Algorithm

For each site, the S leaves form a set of observed random variables, their states being described by the corresponding pattern ω_p. The state of all the internal nodes, denoted σ, form a set of hidden random variables, that is, random variables whose state is not observed. In order to account for rate variability across sites, we associate with each pattern two hidden random variables, ρ_p^η and ρ_p^η, that determine the value of the rate variables in that site. To sum up, the observed random variables are ω_p, and the hidden random variables are $(\sigma, \rho_p^\eta, \rho_p^\theta)$.

We assume that sites within a gene, as well as the genes themselves, evolve independently. Therefore, the total likelihood can be decomposed as

$$L(M_1, \ldots, M_G|\Theta) = \prod_{g=1}^{G} L(M_g|\Xi, \Psi_g, \Lambda) = \prod_{g=1}^{G} \prod_{p=1}^{\Omega} L(\omega_p|\Xi, \Psi_g, \Lambda)^{n_{gp}}.$$

and so

$$\log L(M_1, \ldots, M_G|\Theta) = \sum_{g=1}^{G} \sum_{p=1}^{\Omega} n_{gp} \log L(\omega_p|\Xi, \Psi_g, \Lambda). \quad [4]$$

According to the well-known EM paradigm *(25)* $\log L(M_1, \ldots, M_G|\Theta)$ is guaranteed to increase as long as we maximize the auxiliary function

$$Q(\Theta, \Theta^0) = \sum_{g=1}^{G} \sum_{p=1}^{\Omega} n_{gp} Q_{gp}(\Xi, \Psi_g, \Lambda, \Xi^0, \Psi_g^0, \Lambda^0), \quad [5]$$

where

$$Q_{gp}(\Xi,\Psi_g,\Lambda,\Xi^0,\Psi_g^0,\Lambda^0)=\sum_{\sigma,\rho_p^\eta,\rho_p^\theta}\Pr(\sigma,\rho_p^\eta,\rho_p^\theta|\omega_p,\Xi^0,\Psi_g^0,\Lambda^0)\log\Pr(\omega_p,\sigma,\rho_p^\eta,\rho_p^\theta|\Xi,\Psi_g,\Lambda). \quad [6]$$

Using some manipulations (*see* **Note 1**), this can be written as

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

Denoting by $w_{gpkk'}$ and $Q_{gpkk'}$ the first and second square brackets, respectively, this expression becomes

$$Q_{gp}(\Xi,\Psi_g,\Lambda,\Xi^0,\Psi_g^0,\Lambda^0)=\sum_{k=1}^{K_\eta}\sum_{k'=1}^{K_\theta}w_{gpkk'}Q_{gpkk'}, \quad [7]$$

and consequently

$$Q(\Theta,\Theta^0)=\sum_{g=1}^{G}\sum_{p=1}^{\Omega}\sum_{k=1}^{K_\eta}\sum_{k'=1}^{K_\theta}n_{gp}w_{gpkk'}Q_{gpkk'}. \quad [8]$$

3.2.1. The E-Step

In this step we compute the function $Q(\Theta,\Theta^0)$, or, equivalently, the set of coefficients $w_{gpkk'}$ and $Q_{gpkk'}$. We accomplish this with the aid of an inward–outward recursion on the tree.

3.2.1.1. The Inward (γ) Recursion

Here we propose a variation on the well-known Felsenstein's pruning algorithm *(26)*. Let us associate with each node t (except for the root) a vector $\gamma_i^{gpkk'}(t)=\Pr(V_t|q_t^P=i,\Xi^0,\Psi_{gkk'}^0)$. In words, $\gamma_i^{gpkk'}(t)$ is the probability of observing the nodes V_t (which are a subset of the pattern ω_p) for a gene g, when the gain and loss rate variables are r_k^η and $r_{k'}^\theta$, respectively, and when the parent node of t is known to be in state i. By definition, this function is initialized at all leaves ($t\in V_0$) by

$$\gamma(t\in V_0)=\begin{cases}\begin{pmatrix}1-\xi_t(1-e^{-\eta_{gk}\Delta_t})\\1-(1-\phi_t)e^{-\theta_{gk'}\Delta_t}\end{pmatrix} & q_t=0\\ \begin{pmatrix}\xi_t(1-e^{-\eta_{gk}\Delta_t})\\(1-\phi_t)e^{-\theta_{gk'}\Delta_t}\end{pmatrix} & q_t=1.\end{cases} \quad [9]$$

Here, and in the derivations to follow, we omit the superscript from γ. For all internal nodes (except for the root), γ is computed using the recursion

$$\gamma_i(t)=\sum_{j=0}^{1}A_{ij}(g,t)\tilde{\gamma}_j(t), \quad [10]$$

where $\tilde{\gamma}_j(t)$ is defined as $\gamma_j[\mathrm{L}(t)]\gamma_j[\mathrm{R}(t)]$ (*see* **Note 2**).

The γ-recursion allows for computing the likelihood of any observed pattern ω_p, given the values of the rate variables:

$$\Pr(\omega_p|\Xi^0,\Psi^0_{\mathcal{g}kk'}) = \Pr(V_0|\Xi^0,\Psi^0_{\mathcal{g}kk'}) = \Pr(V_0^L, V_0^R|\Xi^0,\Psi^0_{\mathcal{g}kk'}) =$$

$$= \sum_{i=0}^{1} \Pr(V_0^L, V_0^R, q_0 = i|\Xi^0,\Psi^0_{\mathcal{g}kk'}) =$$

$$= \sum_{i=0}^{1} \Pr(q_0 = i|\Xi^0,\Psi^0_{\mathcal{g}kk'}) \cdot \Pr(V_0^L|q_0 = i,\Xi^0,\Psi^0_{\mathcal{g}kk'}) \cdot$$

$$\Pr(V_0^R|V_0^L, q_0 = i,\Xi^0,\Psi^0_{\mathcal{g}kk'}).$$

Given q_0, V_0^R is independent of V_0^L, and so

$$\Pr(V_0^R|V_0^L, q_0 = i,\Xi^0,\Psi^0_{\mathcal{g}kk'}) = \Pr(V_0^R|q_0 = i,\Xi^0,\Psi^0_{\mathcal{g}kk'}),$$

and

$$\Pr(\omega_p|\Xi^0,\Psi^0_{\mathcal{g}kk'}) = \sum_{i=0}^{1} \pi_i\tilde{\gamma}_i(0). \qquad [11]$$

This γ-recursion can be easily modified to incorporate missing data (*see* **Note 3**).

3.2.1.2. The Outward (α) Recursion

Once the γ-recursion is computed, we can use it to compute a second, complementary, recursion. To this end, let us associate with each node t (except for the root node) a matrix $\alpha_{ij}^{\mathcal{g}pkk'}(t) = \Pr(q_t = j, q_t^P = i|\omega_p,\Xi^0,\Psi^0_{\mathcal{g}kk'})$. It is beneficial to define for each node t (except for the root node) a vector $\beta_j^{\mathcal{g}pkk'}(t) = \sum_{i=0}^{1}\alpha_{ij}^{\mathcal{g}pkk'}(t) = \Pr(q_t = j|\omega_p,\Xi^0,\Psi^0_{\mathcal{g}kk'})$. Upon the computation of α, β is readily computed too. Again, omitting the superscripts, α can be initialized from its definition on the two direct descendants of the root,

$$\alpha(\mathrm{D}(0)) = \frac{1}{\Pr(\omega_p|\Xi^0,\Psi^0_{\mathcal{g}kk'})} \begin{cases} \begin{pmatrix} \pi_0\gamma_0(\bar{\mathrm{D}}(0))A_{00}(\mathcal{g},\mathrm{D}(0)) & 0 \\ \pi_1\gamma_1(\bar{\mathrm{D}}(0))A_{10}(\mathcal{g},\mathrm{D}(0)) & 0 \end{pmatrix} & \mathrm{D}(0) \in V_0, q_0^D = 0 \\ \begin{pmatrix} 0 & \pi_0\gamma_0(\bar{\mathrm{D}}(0))A_{01}(\mathcal{g},\mathrm{D}(0)) \\ 0 & \pi_1\gamma_1(\bar{\mathrm{D}}(0))A_{11}(\mathcal{g},\mathrm{D}(0)) \end{pmatrix} & \mathrm{D}(0) \in V_0, q_0^D = 1 \\ \begin{pmatrix} \pi_0\gamma_0(\bar{\mathrm{D}}(0))\tilde{\gamma}_0(\mathrm{D}(0))A_{00}(\mathcal{g},\mathrm{D}(0)) & \pi_0\gamma_0(\bar{\mathrm{D}}(0))\tilde{\gamma}_1(\mathrm{D}(0))A_{01}(\mathrm{D}(0)) \\ \pi_1\gamma_1(\bar{\mathrm{D}}(0))\tilde{\gamma}_0(\mathrm{D}(0))A_{10}(\mathrm{D}(0)) & \pi_1\gamma_1(\bar{\mathrm{D}}(0))\tilde{\gamma}_1(\mathrm{D}(0))A_{11}(\mathrm{D}(0)) \end{pmatrix} & \mathrm{D}(0) \notin V_0. \end{cases} \qquad [12]$$

Here, $\mathrm{D}(0)$ stands for any one of the direct descendants of the root, and $\mathrm{D}(0)$ is its sibling. For any other internal node, α is computed using the outward-recursion

$$\alpha(t) = \begin{pmatrix} \beta_0(\mathrm{P}(t))\tilde{\gamma}_0(t)A_{00}(\mathcal{g},t)/\gamma_0(t) & \beta_0(\mathrm{P}(t))\tilde{\gamma}_1(t)A_{01}(\mathcal{g},t)/\gamma_0(t) \\ \beta_1(\mathrm{P}(t))\tilde{\gamma}_0(t)A_{10}(\mathcal{g},t)/\gamma_1(t) & \beta_1(\mathrm{P}(t))\tilde{\gamma}_1(t)A_{11}(\mathcal{g},t)/\gamma_1(t) \end{pmatrix} \qquad [13]$$

(*see* **Note 4**).

Finally, for each leaf that is not a descendant of the root,

$$\alpha(t) = \begin{cases} \begin{pmatrix} \beta_0(\mathrm{P}(t)) & 0 \\ \beta_1(\mathrm{P}(t)) & 0 \end{pmatrix} & q_t = 0 \\ \begin{pmatrix} 0 & \beta_0(\mathrm{P}(t)) \\ 0 & \beta_1(\mathrm{P}(t)) \end{pmatrix} & q_t = 1. \end{cases} \qquad t \in V_0, \mathrm{P}(t) \neq 0 \qquad [14]$$

Again, this recursion can be straightforwardly modified when missing data are present (*see* **Note 5**).

These inward–outward recursions are the phylogenetic equivalent of the backward–forward recursions known from hidden Markov models, and other versions of it have already been developed *(27, 28)*. The version that we developed here can be shown to be the realization of the junction tree algorithm *(29)* on rooted bifurcating trees (*see* **Note 6**).

3.2.1.3. Computing the Coefficients $w_{gpkk'}$

Here we show that the γ-recursion is sufficient to compute the coefficients $w_{gpkk'}$. From the definition, $w_{gpkk'} = \Pr(\rho_p^\eta = k, \rho_p^\theta = k' | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0)$. Using the Bayes formula $\Pr(x, y|z) = \Pr(x, y, z)/\sum_{x,y} \Pr(x, y, z)$, we can rewrite it as

$$w_{gpkk'} = \frac{\Pr(\rho_p^\eta, = k, \rho_p^\theta = k', \omega_p | \Xi^0, \Psi_g^0, \Lambda^0)}{\sum_{h,h'} \Pr(\rho_p^\eta = h, \rho_p^\theta = h', \omega_p | \Xi^0, \Psi_g^0, \Lambda^0)} =$$

$$= \frac{\Pr(\rho_p^\eta = k | \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\rho_p^\theta = k' | \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\omega_p | \Xi^0, \Psi_{gkk'}^0)}{\sum_{h,h'} \Pr(\rho_p^\eta = h | \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\rho_p^\theta = h' | \Xi^0, \Psi_g^0, \Lambda^0) \cdot \Pr(\omega_p | \Xi^0, \Psi_{ghh'}^0)}.$$

But $\Pr(\rho_p^\eta = k | \Xi^0, \Psi_g^0, \Lambda^0)$ is just the current estimate of the probability of the gain rate variable to have the value r_k^η, namely $(f_k^\eta)^0$. Similarly, $\Pr(\rho_p^\theta = k' | \Xi^0, \Psi_g^0, \Lambda^0)$ is just $(f_{k'}^\theta)^0$. Therefore, the expression for the coefficients $w_{gpkk'}$ is reduced to

$$w_{gpkk'} = \frac{(f_k^\eta)^0 (f_{k'}^\theta)^0 \Pr(\omega_p | \Xi^0, \Psi_{gkk'}^0)}{\sum_{h,h'} (f_h^\eta)^0 (f_{h'}^\theta)^0 \Pr(\omega_p | \Xi^0, \Psi_{ghh'}^0)}. \qquad [15]$$

The function $\Pr(\omega_p | \Xi^0, \Psi_{gkk'}^0)$ is the likelihood of observing pattern ω_p for gain and loss rate variables r_k^η and $r_{k'}^\theta$, respectively. This is readily computed upon completion of the γ-recursion, using **Eq. [11]**.

3.2.1.4. Computing the Coefficients $Q_{gpkk'}$

Here we show that these coefficients require the α, β-recursion. By definition,

$$Q_{gpkk'} = \sum_\sigma \Pr(\sigma | \omega_p, \Xi^0, \Psi_{gkk'}^0) \cdot [\log f_k^\eta + \log f_{k'}^\theta + \log \Pr(\omega_p, \sigma | \Xi, \Psi_{gkk'})].$$

The probability $\Pr(\omega_p, \sigma|\Xi, \Psi_{\mathcal{g}kk'})$ is just the likelihood of a particular realization of the tree, thus from **Eq. [1]**

$$\log \Pr(\omega_p, \sigma|\Xi, \Psi_{\mathcal{g}kk'}) = \sum_{i=0}^{1} \delta(q_0, i) \cdot \log \pi_i + \sum_{i,j=0}^{1} \sum_{t=1}^{N-1} \delta(q_t, j)\delta(q_t^P, i) \cdot \log A_{ij}(\mathcal{g}, t). \quad [16]$$

Here, $\delta(a, b)$ is the Kronecker delta function, which is 1 for $a = b$ and 0 otherwise. Denote the expectation over $\Pr(\sigma|\omega_p, \Xi^0, \Psi^0_{\mathcal{g}kk'})$ by E_σ. Applying it to **Eq. [16]**, we get

$$E_\sigma[\log \Pr(\omega_p, \sigma|\Xi, \Psi_{\mathcal{g}kk'})] = \sum_{i=0}^{1} \log \pi_i \cdot E_\sigma[\delta(q_0, i)] + \sum_{i,j=0}^{1} \sum_{t=1}^{N-1} \log A_{ij}(\mathcal{g}, t) \cdot E_\sigma[\delta(q_t, j)\delta(q_t^P, i)].$$

But $E_\sigma[\delta(q_0, i)] = \Pr(q_0 = i|\omega_p, \Xi^0, \Psi^0_{\mathcal{g}kk'}) = \beta_i(0)$, and similarly $E_\sigma[\delta(q_t, j)\delta(q_t^P, i)] = \alpha_{ij}(t)$. Hence, $Q_{\mathcal{g}pkk'}$ is given by

$$Q_{\mathcal{g}pkk'} = \sum_\sigma \Pr(\sigma|\omega_p, \Xi^0, \Psi^0_{\mathcal{g}kk'})[\log f_k^\eta + \log f_{k'}^\theta + \log \Pr(\omega_p, \sigma|\Xi, \Psi_{\mathcal{g}kk'})] = \\ = \log f_k^\eta + \log f_{k'}^\theta + \sum_{i=0}^{1} \beta_i(0) \log \pi_i + \sum_{i,j=0}^{1} \sum_{t=1}^{N-1} \alpha_{ij}(t) \log A_{ij}(\mathcal{g}, t). \quad [17]$$

3.2.2. The M-Step

Substituting **Eq. [17]** in **Eq. [8]**, we obtain an explicit form of the function whose maximization guarantees stepping up-hill in the likelihood landscape,

$$\begin{aligned} Q = &\sum_{\mathcal{g}=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} n_{\mathcal{g}p} w_{\mathcal{g}pkk'} (\log f_k^\eta + \log f_{k'}^\theta) + \\ &+ \sum_{\mathcal{g}=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} n_{\mathcal{g}p} w_{\mathcal{g}pkk'} [\beta_0^{\mathcal{g}pkk'}(0) \log \pi_0 + \beta_1^{\mathcal{g}pkk'}(0) \log \pi_1] + \\ &+ \sum_{\mathcal{g}=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{\mathcal{g}p} w_{\mathcal{g}pkk'} \alpha_{00}^{\mathcal{g}pkk'}(t) \log[1 - \xi_t(1 - e^{-\eta_{\mathcal{g}k}\Delta_t})] + \\ &+ \sum_{\mathcal{g}=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{\mathcal{g}p} w_{\mathcal{g}pkk'} \alpha_{01}^{\mathcal{g}pkk'}(t) [\log \xi_t + \log(1 - e^{-\eta_{\mathcal{g}k}\Delta_t})] + \\ &+ \sum_{\mathcal{g}=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{\mathcal{g}p} w_{\mathcal{g}pkk'} \alpha_{10}^{\mathcal{g}pkk'}(t) \log[1 - (1 - \phi_t)e^{-\theta_{\mathcal{g}k'}\Delta_t}] + \\ &+ \sum_{\mathcal{g}=1}^{G} \sum_{p=1}^{\Omega} \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{t=1}^{N-1} n_{\mathcal{g}p} w_{\mathcal{g}pkk'} \alpha_{11}^{\mathcal{g}pkk'}(t) [\log(1 - \phi_t) - \theta_{\mathcal{g}k'}\Delta_t]. \end{aligned} \quad [18]$$

Actually, any increase in Q is sufficient to guarantee an increase in the likelihood, suggesting that a precise maximization of Q is not very important. Therefore, we speed computations by performing low-tolerance maximization with respect to each of the parameters individually. Except for the parameters λ_η and λ_θ, it is easy to differentiate Q twice with respect to any parameter. This lends itself into using simple zero-finding algorithms; we chose the Newton-Raphson algorithm *(30)*. Maximizing Q with respect to the shape parameters λ_η and λ_θ is more involved, as Q depends on these parameters only through the discrete approximation of the rate variability distributions, **Eq. [3]** (*see* **Note 7**).

4. Notes

1. If we replace the formal summing over all states of ρ_p^η and ρ_p^η in **Eq. [6]** by a direct sum, we get

$$\begin{aligned} Q_{gp}(\Xi, \Psi_g, \Lambda, \Xi^0, \Psi_g^0, \Lambda^0) = \sum_{k=1}^{K_\eta} \sum_{k'=1}^{K_\theta} \sum_{\sigma} \Pr(\sigma, \rho_p^\eta = k, \rho_p^\theta \\ = k' | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) \\ \log \Pr(\omega_p, \sigma, \rho_p^\eta = k, \rho_p^\theta = k' | \Xi, \Psi_g, \Lambda). \end{aligned} \quad [19]$$

Using our notational conventions, we can write the first term in **Eq. [19]** as

$$\begin{aligned} \Pr(\sigma, \rho_p^\eta = k, \rho_p^\theta = k' | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) \\ = \Pr(\rho_p^\eta = k, \rho_p^\theta = k' | \omega_p, \Xi^0, \Psi_g^0, \Lambda^0) \\ \cdot \Pr(\sigma | \omega_p, \Xi^0, \Psi_{gkk'}^0), \end{aligned} \quad [20]$$

and the second term as

$$\begin{aligned} \log \Pr(\omega_p, \sigma, \rho_p^\eta = k, \rho_p^\theta = k' | \Xi, \Psi_g, \Lambda) \\ = \log \Pr(\rho_p^\eta = k | \Xi, \Psi_g, \Lambda) + \\ + \log \Pr(\rho_p^\theta = k' | \Xi, \Psi_g, \Lambda) \\ + \log \Pr(\omega_p, \sigma | \Xi, \Psi_{gkk'}) \\ = \log f_k^\eta + \log f_{k'}^\theta + \log \Pr(\omega_p, \sigma | \Xi, \Psi_{gkk'}). \end{aligned} \quad [21]$$

Substituting **Eqs. [20]** and **[21]** back in **Eq. [19]** gives the desired result.

2. We expand

$$\begin{aligned}\gamma_i(t) &= \Pr(V_t|q_t^P = i) = \Pr(V_t^L, V_t^R|q_t^P = i)\\ &= \sum_{j=0}^{1} \Pr(V_t^L, V_t^R, q_t = j|q_t^P = i) =\\ &= \sum_{j=0}^{1} \Pr(q_t = j|q_t^P = i) \cdot \Pr(V_t^L|q_t = j, q_t^P = i)\\ &\quad \cdot \Pr(V_t^R|V_t^L, q_t = j, q_t^P = i).\end{aligned} \tag{22}$$

The first term is simply the definition of $A_{ij}(g, t)$. Given q_t, V_t^L is independent on q_t^P, thus the second term is just $\Pr(V_t^L|q_t = j) = \gamma_j(t^L)$. By similar arguments the third term is just $\Pr(V_t^R|q_t = j) = \gamma_j(t^R)$. By substituting those results in **Eq. [22]**, we recover the recursion formula, **Eq. [10]**.

3. One of the appealing features of this recursion is that it allows to treat missing data fairly easily. Only a single option has to be added to the initialization phase **Eq. [9]**,

$$\gamma(t \in V_0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad q_t = *.$$

4. To prove this recursion, let us start with the definition of α,

$$\begin{aligned}\alpha_{ij}(t) &= \Pr(q_t = j, q_t^P = i|\omega_p) = \Pr(q_t = j, q_t^P = i|V_0)\\ &= \Pr(q_t^P = i|V_0) \cdot \Pr(q_t = j|q_t^P = i, V_0)\\ &= \beta_i(\mathrm{P}(t)) \cdot \Pr(q_t = j|q_t^P = i, V_0).\end{aligned} \tag{23}$$

Let us make the decomposition $V_0 = V_t + \bar{V}_t$, with $\bar{V}_t$ being the set of all leaves such that node t is not among their ancestors. But, given q_t^P, the state of node t is independent on $\bar{V}_t$, and therefore **Eq. [23]** becomes

$$\alpha_{ij}(t) = \beta_i(\mathrm{P}(t)) \cdot \Pr(q_t = j|q_t^P = i, V_t). \tag{24}$$

From Bayes formula,

$$\begin{aligned}\Pr(q_t = j|q_t^P = i, V_t) &= \frac{\Pr(q_t = j, V_t|q_t^P = i)}{\Pr(V_t|q_t^P = i)}\\ &= \frac{\Pr(q_t = j|q_t^P = i) \cdot \Pr(V_t|q_t = j, q_t^P = i)}{\gamma_i(t)}\\ &= \frac{A_{ij}(g, t)}{\gamma_i(t)} \cdot \Pr(V_t|q_t = j, q_t^P = i).\end{aligned} \tag{25}$$

But given q_t, V_t is independent of $\mathrm{P}(t)$ and therefore

$$\Pr(V_t|q_t = j, q_t^P = i) = \Pr(V_t|q_t = j) = \tilde{\gamma}_j(t). \quad [26]$$

Combining **Eqs. [25]** and **[26]** in **Eq. [24]**, we get

$$\alpha_{ij}(t) = \frac{\tilde{\gamma}_j(t)\beta_i(\mathrm{P}(t))}{\gamma_i(t)} A_{ij}(g, p),$$

which is just another form of writing **Eq. [13]**.

5. When missing data are present, two simple modifications are required. First, we have to add to the initialization phase **Eq. [12]** an option

$$\alpha(\mathrm{D}(0)) = \frac{1}{\Pr(\omega_p|\Xi^0, \Psi^0_{gkk'})} \begin{Bmatrix} \pi_0\gamma_0[\bar{\mathrm{D}}(0)]A_{00}[g, \mathrm{D}(0)] & \pi_0\gamma_0\bar{\mathrm{D}}(0)]A_{01}[\mathrm{D}(0)] \\ \pi_1\gamma_1[\bar{\mathrm{D}}(0)]A_{10}[\mathrm{D}(0)] & \pi_1\gamma_1(\bar{\mathrm{D}}(0)]A_{11}[\mathrm{D}(0)] \end{Bmatrix} \quad \mathrm{D}(0) \in V_0, q_0^D = *$$

Second, we have to add to the finalization phase **Eq. [14]** an option

$$\alpha(t) = \begin{Bmatrix} \beta_0[\mathrm{P}(t)]A_{00}(g, t) & \beta_0[\mathrm{P}(t)]A_{01}(g, t) \\ \beta_1[\mathrm{P}(t)]A_{10}(g, t) & \beta_1[\mathrm{P}(t)]A_{11}(g, t) \end{Bmatrix} \qquad q_t = *.$$

6. The junction tree algorithm is a scheme to compute marginal probabilities of maximal cliques on graphs by means of belief propagation on a modified junction tree. Indeed, the matrix α computes marginal probabilities of pairs $(t, \mathrm{P}(t))$, but such pairs are nothing but maximal cliques on rooted bifurcating trees.
7. In our implementation, we used Yang's quantile method *(24)* to compute the discrete levels of the gamma distributions such that each level has equal probability. Formally, $f_1^\eta = v$, $f_k^\eta = (1 - v)/(K_\eta - 1)$ for $k = 2, \ldots, K_\eta$, and $f_k^\theta = 1/K_\theta$ for $k = 1, \ldots, K_\theta$. To perform the maximization in this case, we used Brent's maximization algorithm that does not require derivatives *(30)*.

References

1. Nixon JE, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J. A spliceosomal intron in Giardia lamblia. Proc Natl Acad Sci U S A 2002; 99:3359–3361.
2. Vanacova S, Yan W, Carlton JM, Johnson PJ. Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis. Proc Natl Acad Sci U S A 2005; 102:4430–4435.
3. Simpson AG, MacQuarrie EK, Roger AJ. Early origin of canonical introns. Nature 2002;419:270.
4. Collins L, Penny D. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol 2005;22:1053–1066.
5. Lynch M., Richardson AO. The evolution of spliceosomal introns. Curr Opin Genet Dev 2002;12:701–710.
6. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 2006;7:211–221.
7. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and

massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 2003;13: 1512–1517.
8. Roy SW, Gilbert W. Complex early genes. Proc Natl Acad Sci U S A 2005;102:1986–1991.
9. Roy SW, Gilbert W. Rates of intron loss and gain: implications for early eukaryotic evolution. Proc Natl Acad Sci U S A 2005;102:5773–5778.
10. Csuros M. Likely scenarios of intron evolution, Lecture Notes in Bioinformatics (McLysaght, A. and Huson, D., editors): Proc. RECOMB 2005 Comparative Genomics International Workshop (RCG 2005) 2005;3678:47–60.
11. Qiu WG, Schisler N, Stoltzfus A. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol Biol Evol 2004; 21:1252–1263.
12. Fedorov A, Roy SW, Fedorova L, Gilbert W. Mystery of intron gain. Genome Res 2003;13:2236–2241.
13. Cho S, Jin SW, Cohen A, Ellis RE. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. Genome Res 2004;14:1207–1220.
14. Roy SW, Hartl DL. Very little intron loss/gain in Plasmodium: intron loss/gain mutation rates and intron number. Genome Res 2006;16:750–756.
15. Jeffares DC, Mourier T, Penny D. The biology of intron gain and loss. Trends Genet 2006;22:16–22.
16. Nguyen HD, Yoshihama M, Kenmochi N. New maximum likelihood estimators for eukaryotic intron evolution. PLoS Comput Biol 2005;1:e79.
17. Nei M, Chakraborty R, Fuerst PA. Infinite allele model with varying mutation rate. Proc Natl Acad Sci U S A 1976;73: 4164–4168.
18. Uzzell T, Corbin KW. Fitting discrete probability distributions to evolutionary events. Science 1971;172:1089–1096.
19. Dibb NJ. Proto-splice site model of intron origin. J Theor Biol 1991;151:405–416.
20. Dibb NJ, Newman AJ. Evidence that introns arose at proto-splice sites. Embo J 1989;8:2015–2021.
21. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. Reconstruction of ancestral protosplice sites. Curr Biol 2004;14: 1505–1508.
22. Jordan IM (ed.). Learning in Graphical Models. Kluwer Academic Publishers, Boston, MA, 1998.
23. Jin L, Nei M. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 1990;7:82–102.
24. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 1994;39: 306–314.
25. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Statist Soc B 1977; 39:1–38.
26. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–376.
27. Friedman N, Ninio M, Pe'er I, Pupko T. A structural EM algorithm for phylogenetic inference. J Comput Biol 2002;9: 331–353.
28. Siepel A, Haussler D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 2004;21:468–488.
29. Castillo E, Gutierrez JM, Hadi AS. Expert systems and probabilistic network models (Monographs in Computer Science). Springer, New York, 1996.
30. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in C: The art of scientific computing. Cambridge University Press, New York, 2nd ed., 1992.