Likely scenarios of intron evolution

Miklós Csűrös

Department of Computer Science and Operations Research
Université de Montréal
C.P. 6128, succ. Centre-Ville, Montréal, Qué., Canada, H3C 3J7
csuros@iro.umontreal.ca

Abstract. Whether common ancestors of eukaryotes and prokaryotes had introns is one of the oldest unanswered questions in molecular evolution. Recently completed genome sequences have been used for comprehensive analyses of exon-intron organization in orthologous genes of diverse organisms, leading to more refined work on intron evolution. Large sets of intron presence-absence data require rigorous theoretical frameworks in which different hypotheses can be compared and validated. We describe a probabilistic model for intron gains and losses along an evolutionary tree. The model parameters are estimated using maximum likelihood. We propose a method for estimating the number of introns lost or unobserved in all extant organisms in a study, and show how to calculate counts of intron gains and losses along the branches by using posterior probabilities. The methods are used to analyze the most comprehensive intron data set available presently, consisting of 7236 intron sites from eight eukaryotic organisms. The analysis shows a dynamic history with frequent intron losses and gains, and fairly — albeit not as greatly as previously postulated — intron-rich ancestral organisms.

1 Introduction

A major difference between eukaryotic and prokaryotic gene organization is that many eukaryotic genes have a mosaic structure: coding sequences are separated by intervening non-coding sequences, known as introns. Francis Crick's 1979 comment [1] on the evolutionary origins of spliceosomal introns — "I have noticed that this question has an extraordinary fascination for almost everybody concerned with the problem" — could have been said yesterday. The problem is still not completely resolved [2]. The question of whether or not the most recent common ancestor of eukaryotes and prokaryotes had introns, known as the "introns early/late" debate [3], is one of the oldest unanswered questions in molecular evolution. Recent advances [4–8] rely on whole-genome sequences for diverse organisms. It has become clear that introns have been gained and lost in different lineages at various rates. In this context it is of particular interest to estimate the intron densities in early eukaryotic organisms, as well as rates and patterns of intron loss and gain along different evolutionary lineages. The aim of this article is to describe a probabilistic model which allows for a maximum likelihood (ML) analysis of rates and scenarios. We describe some methods to

this end and apply them to a data set of 7236 introns from eight fully sequenced eukaryotic organisms.

2 A probabilistic model for intron evolution

In order to model the evolution of introns along an evolutionary tree, we use a Markov model that permits varying rates along different branches, described as follows. Let T be a phylogenetic tree T over a set of species X: T is a rooted tree in which the leaves are bijectively labeled by the elements of X. Let $E(T)$ denote the set of edges (directed away from the root), and let $V(T)$ denote the node set of the tree. Throughout the paper, intron presence is encoded by the value `1`, and intron absence is encoded by the value `0`. Along each edge $e \in E(T)$, introns are generated by a two-state continuous-time Markov process with *gain* and *loss rates* $\lambda_e, \mu_e \geq 0$, respectively. The length of an edge e is denoted by t_e. In addition, the root is associated with the *root probabilities* $\pi_{\mathtt{0}}$, $\pi_{\mathtt{1}}$ with $\pi_{\mathtt{0}} + \pi_{\mathtt{1}} = 1$. The tree T with its parameters defines a stochastic evolution model for the *state* $\tilde{\chi}(u)$ of an intron site at every tree node $u \in V(T)$ in the following manner. The intron is present at the root with probability $\pi_{\mathtt{1}}$. The intron state evolves along the tree edges from the root towards the leaves, and changes on each edge according to the transition probabilities. For every child node v and its parent u, $\mathbb{P}\left\{ \tilde{\chi}(v) = j \;\middle|\; \tilde{\chi}(u) = i \right\} = p_{i \to j}(uv)$, where $p_{i \to j}$ are determined by the edge parameters, which we discuss shortly. The values at the leaves form the *character* $\chi = (\tilde{\chi}(u) \colon u \in X)$. The input data set (or *sample*) consists of independent and identically distributed (iid) characters: $D = (\chi_i \colon i = 1, \ldots, n)$.

Using standard results [9], the transition probabilities along the edge e with rates $\lambda_e = \lambda, \mu_e = \mu$ and length t can be written as

$$p_{\mathtt{0}\to\mathtt{0}}(e) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-t(\lambda+\mu)} \qquad p_{\mathtt{0}\to\mathtt{1}}(e) = \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu} e^{-t(\lambda+\mu)}$$
$$p_{\mathtt{1}\to\mathtt{0}}(e) = \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu} e^{-t(\lambda+\mu)} \qquad p_{\mathtt{1}\to\mathtt{1}}(e) = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} e^{-t(\lambda+\mu)}.$$

In the absence of independent edge length estimates, we fix the scaling for the edge lengths in such a way that $\lambda_e + \mu_e = 1$.

A somewhat more complicated model of intron evolution was used by Rzhetsky *et al.* [10], who also accounted for possible intron sliding [11], whereby orthologous intron sites may differ by a few positions with respect to the underlying coding sequence in different organisms. In our case, the orthology criterion incorporates intron sliding a priori. Some other authors (e.g., [5]) imposed a reversible Markov model with identical rates across different branches, which is not entirely realistic for intron evolution, but nevertheless can result in important insights already.

3 ML estimation of parameters and scenarios

3.1 Unobserved intron sites

Our goal is to design a maximum likelihood approach to estimate the model parameters on a given tree T, and to calculate likely scenarios of intron gains and losses along the edges. The described probabilistic model is fairly simple, and the parameters can be estimated from a data set by usual optimization techniques [12]. There is, however, an inherent difficulty in analyzing an intron absence/presence data set: there is no obvious evidence of introns lost in all extant organisms in the study. Consequently, one has access only to a sample of iid characters from which the all-0 characters ("unobserved introns") have been removed. Maximizing the likelihood without the all-0 characters introduces a bias. At the same time, it is not possible to estimate the number of missing all-0 characters by maximizing either the likelihood (every added all-0 character decreases it), or the average likelihood (an unbounded number of all-0 characters can be added if their likelihood is large enough). It is therefore necessary to separate the estimation of unobserved sites from likelihood maximization.

The problem of augmenting the data set with a certain number of all-0 characters has a particular relevance for the complexity of ML estimation of phylogenies. Tuffley and Steel [13] showed that ML and maximum parsimony (MP) yield the same optimal tree topology when enough all-0 characters are added to the data set in a symmetric binary model. Their result was employed very recently [14, 15] to demonstrate the NP-hardness of ML optimization for phylogenies. The theoretical connection between ML and MP established by the addition of all-0 characters has direct practical consequences in the case of intron data sets. For instance, the analyses of the same sample carried out by two groups of researchers [4, 16–18], using ML and MP, arrived at different conclusions concerning intron gain/loss rates and ancient intron density. Some of the disagreements can be attributed to different assumptions about unobserved sites, instead of methodological issues.

For a formal discussion, define the following notions. An *extension* $\tilde{\chi}$ of a character χ is an assignment of states to every tree node that agrees with χ at the leaves. Let $H(\chi)$ denote the set of all extensions of χ. The likelihood of a character χ is the probability

$$f_\chi = \sum_{\tilde{\chi} \in H(\chi)} \pi_{\tilde{\chi}(\mathsf{root})} \prod_{uv \in E(T)} p_{\tilde{\chi}(u) \to \tilde{\chi}(v)}(uv).$$

The likelihood of a complete data set $D = (\chi_i \colon i = 1, \ldots, n)$ is simply $L(D) = \prod_{i=1}^{n} f_{\chi_i}$. Let f_0 denote the likelihood of the all-0 character $0^{|X|}$. The expected number of all-0 characters in a data set of size n is nf_0. Accordingly, the expected number of unobserved sites given that there are $\bar{n}$ observed ones (non-all-0 characters in the data set), is

$$\hat{n}_0 = \bar{n} \frac{f_0}{1 - f_0}. \tag{1}$$

(The distribution of the number of unobserved sites is a negative binomial distribution with parameters $\bar{n}$ and $(1 - f_0)$.)

Let $\bar{D} = (\chi_i \colon i = 1, \ldots, \bar{n})$ denote the observed sample, without the all-0 characters, and $n_0 = n - \bar{n}$ denote the true number of unobserved sites. Figure 1 sketches the algorithm Guess-the-sample for ML estimation of model parameters using a guess for n_0. The guess is used to optimize the model parameters and then to compute the expected number of unobserved sites using Eq. (1). Line G4 compares the latter with the original guess and if they differ too much, it rejects the optimized parameters. The exact definition of "too much" can rely on the concentration properties of n_0: for a given sample size n, it is binomially distributed with parameters n and f_0 with a variance of $nf_0(1 - f_0)$. For example, the guess Z can be rejected if

$$|\hat{n}_0 - Z| > c\sqrt{(\bar{n} + Z)f_0(1 - f_0)},$$

where c is a constant determining the desired confidence level. Figure 3 shows the behavior of this difference for a data set analyzed in Section 4. Notice that the plot suggests that n_0 could be estimated by an iterative technique, in which two steps are alternating: (1) estimation of the number of intron sites, based on model parameters, and observed introns, and (2) maximization of the likelihood given the estimated number of intron sites. In other words, $\hat{n}_0$ can be fed back to the algorithm in Line G4 in lieu of rejection, until convergence is reached. Based on the plot of Fig. 3, however, the convergence is very slow, and there is nothing gained over trying basically all possible values for n_0. (There is an upper bound given by the length of sequences from which $\bar{D}$ was obtained.)

Algorithm Guess-the-sample
Input A guess Z for n_0, observed sample $\bar{D} = (\chi_1 \colon i = 1, \ldots, \bar{n})$
G1 Set $D' = (\chi'_i \colon i = 1, \ldots, \bar{n} + Z)$ with $\chi'_i = \chi_i$ for $i \leq \bar{n}$ and $\chi'_i = \mathtt{0}^{|X|}$ for $i > \bar{n}$.
G2 Optimize the model parameters on the augmented sample D'.
G3 Calculate $\hat{n}_0$ by using the optimized model parameters in Eq. (1).
G4 Reject if $\hat{n}_0$ differs from Z by too much.

Fig. 1. ML parameter estimation with unknown number n_0 of unobserved sites.

3.2 Patterns of intron gain and loss along tree edges

Once the number of unobserved intron sites is estimated and the model parameters are optimized, the model can be used to infer likely scenarios of intron evolution. In particular, exact posterior probabilities for intron presence can be calculated at each node, or for intron loss and gain on each branch. Define the

lower conditional likelihood for every node u, site i, and state $x \in \{0, 1\}$ by:

$$\begin{aligned} L_i^{(x)}(u) &= I\{x = \chi_i(u)\} && \text{when } u \text{ is a leaf,} \\ L_i^{(x)}(u) &= \prod_{v \in \mathsf{children}(u)} \Big(\sum_{y \in \{0,1\}} p_{x \to y}(uv) L_i^{(y)}(v) \Big) && \text{when } u \text{ is not a leaf,} \end{aligned}$$

where $I\{A\}$ is the indicator function: $I\{A\} = 1$ if A is true, otherwise $I\{A\} = 0$. The value $L_i^{(x)}(u)$ is the probability of observing the states from character χ_i at the leaves of the subtree T_u rooted at u, given that u is in state x.

We also need the *upper conditional likelihood* $U_i^{(x)}(u)$, which is the probability of observing the states from character χ_i at leaves that are not in the subtree T_u, given that u is in state x. The upper conditional likelihoods can be computed by dynamic programming, using the following recursions in a breadth-first traversal.

$$\begin{aligned} U_i^{(x)}(\mathsf{root}) &= 1 \\ U_i^{(x)}(u) &= \sum_{y \in \{0,1\}} p_{y \to x}(vu) U_i^{(y)}(v) \prod_{w \in \mathsf{siblings}(v)} \Big(\sum_{z \in \{0,1\}} p_{y \to z}(vw) L_i^{(z)}(w) \Big), \end{aligned}$$

where v is the parent of u.

The posterior probability that node u is in state x at site i equals

$$q_i^{(x)}(u) \propto U_i^{(x)}(u) L_i^{(x)}(u).$$

Usual posterior calculations of ancestral states described in, e.g., [19, 12] apply to reversible mutation models, when the tree can be rerooted at u and then $L^{(x)}(u)$ can be used directly. Here we need the additional technicality of computing upper conditional likelihoods. One can also compute the posterior probability of site i undergoing a $x \to y$ transition on the edge leading to the node v from its parent u as

$$q_i^{(x \to y)}(uv) \propto U_i^{(x)}(u) p_{x \to y}(uv) L_i^{(y)}(v).$$

Working with posterior probabilities instead of the single most likely extension has the advantage that posterior probabilities can be summed to obtain expected counts for intron gains and losses. The *posterior mean counts* of states at a node u, or state transitions $(x \to y)$ on an edge uv are computed as

$$\begin{aligned} n^{(x)}(u) &= \sum_{i=1}^{n} q_i^{(x)}(u), \\ n^{(x \to y)}(uv) &= \sum_{i=1}^{n} q_i^{(x \to y)}(uv), \end{aligned} \tag{2}$$

respectively. (Notice that the sums include the unobserved intron sites.) In particular, $n^{(1)}(u)$ is the expected number of introns present at node u, given the model parameters and the observed data. Similarly, $n^{(0 \to 1)}(uv)$ is the expected number of introns gained, and $n^{(1 \to 0)}(uv)$ is the expected number of introns lost along the edge uv.

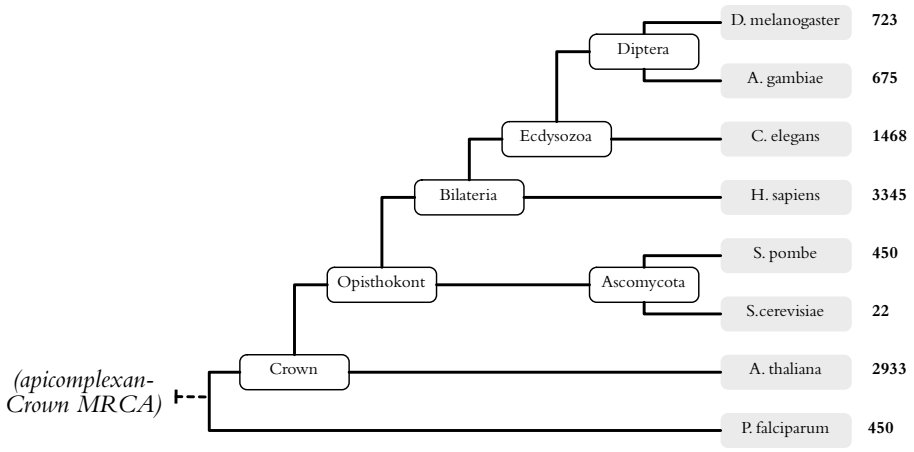

Fig. 2. Phylogenetic tree for the data set in Section 4, showing taxon names and intron counts. *P. falciparum* serves as an outgroup. Only the solid edges were used in the computations. The edge that connects *P. falciparum* to the tree accounts for changes between the Opisthokont node, and the most recent common ancestor (MRCA) of plants, animals, fungi, and apicomplexans, as well as for those leading from that MRCA to *P. falciparum*.

4 Intron evolution in eukaryotes

Rogozin *et al.* [4] compiled a data set based on orthologous protein groups in eukaryotic organisms. They aligned protein sequences with the genome sequences of eight fully sequenced organisms, and defined orthologous intron positions based on conserved regions in the alignments. The data set (downloaded from `ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/intron_evolution`) consists of 7236 orthologous intron positions, from 684 protein groups. Figure 2 shows the organisms involved in the study, as well as the number of introns for each organism.

We note in passing that there is some ongoing debate [20–23] as to whether the phylogenetic tree of Fig. 2 is correct, namely, whether Ecdysozoa are monophyletic. Philippe *et al.* [22] argue that they are, and that support for other hypotheses are due to long branch attraction phenomena. Roy and Gilbert [21] also argue for an ecdysozoan clade, based on the intron data set of [4]. We consider only one phylogenetic tree, and leave further analysis to a more complete version of this abstract.

We implemented a Java package for the analysis of intron data sets, which performs parameter optimization and posterior calculations. As we indicated in §3.1, it is necessary to estimate the number of unobserved intron sites before proceeding to likelihood maximization. Figure 3 shows the estimation procedure applied to the data at hand. The estimation reaches a fix point at around 35 thousand unobserved characters, i.e., likelihood optimization with that many

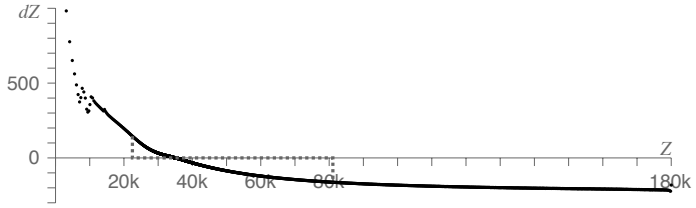

Fig. 3. Estimation of unobserved intron sites. The X axis shows the guess Z with which algorithm GUESS-THE-SAMPLE is invoked, and the Y axis shows the difference $\hat{n}_0 - Z$ calculated after parameter optimization. The dotted lines delineate the region in which the difference is below twice the standard deviation.

all-0 characters gives an equal expectation (within integer rounding) for the number of unobserved characters. Allowing for some statistical error, about 20–80 thousand unobserved characters give an expectation that is within twice the standard error after parameter optimization.

Using 35000 unobserved characters, we proceeded to parameter optimization, and then to the estimation of intron loss and gain patterns. Rogozin *et al.* [4] computed losses and gains using Dollo parsimony [24, 25], assuming that every intron arose only once along the tree.

Roy and Gilbert [16, 17] estimated transition probabilities and intron counts using "local" optimization, independently for each edge. (A similar method was used in [6].) Their principal technique is a tree contraction, in which a whole subtree is replaced by a single branch, and the corresponding characters are derived by computing a logical OR over the intron states at the subtree leaves. They provide separate sets of formulas for analyzing exterior and interior branches. In the case of exterior branches, three-leaf star trees are formed, in which the original edge is preserved, a second edge is contracted from the sibling subtree, and the third edge is contracted from the rest of the tree. In the case of internal branches, they contract the subtrees for the four neighbors of the edge endpoints to form a quartet. (The method applies only to binary trees.) The methods of [16, 17] estimate a larger number of parameters than our likelihood optimization: in addition to the probabilities of intron inheritance, various intron loss and gain counts are independently estimated on each branch. It is plausible that by not enforcing consistency between different estimates that depend on the same parameter (for instance, the same edge transition probabilities should appear in many different contractions), the results may get distorted. In addition, the Roy-Gilbert formulas do not account for the possibility of introns arising more than once.

Multiple origins of introns in an orthologous position are explicitly forbidden by Dollo parsimony. Parallel gains are allowed in our probabilistic model, and may in truth account for a number of shared introns between eukaryotic

kingdoms [5, 18]. Even if one disregards for a moment the question of parallel gains, Dollo parsimony still has its own shortcomings when used for reconstructing plausible histories. If intron gains are much less probable than intron losses, Dollo parsimony retrieves the most likely extension for every single character. It is not suitable, however, for determining cumulative values such as ancestral intron counts, since then the contribution of second, third, etc. most probable histories cannot be neglected. In particular, there is a chance that an intron is lost in such a pattern that its origin will be placed at a more recent inner node in the tree. For example, if an intron first appears in the MRCA for Ecdysozoa (similar example can be constructed for any phylogeny), it is possible that it is lost in *D. melanogaster* and *A. gambiae* and is only present in *C. elegans*. Then Dollo parsimony puts the origin of that intron onto the edge leading to *C. elegans*. Conversely, if the intron is lost in *C. elegans*, then Dollo parsimony places its origin at the node for Diptera. All methods agree (cf. Table 1) that such events cannot be too rare because many introns are lost on the branches leading to the insects and the worm. Another case in point are the 197 introns that are unique to *S. pombe* (44% of its introns). Dollo parsimony concludes that they were gained on that branch, which is doubtful.

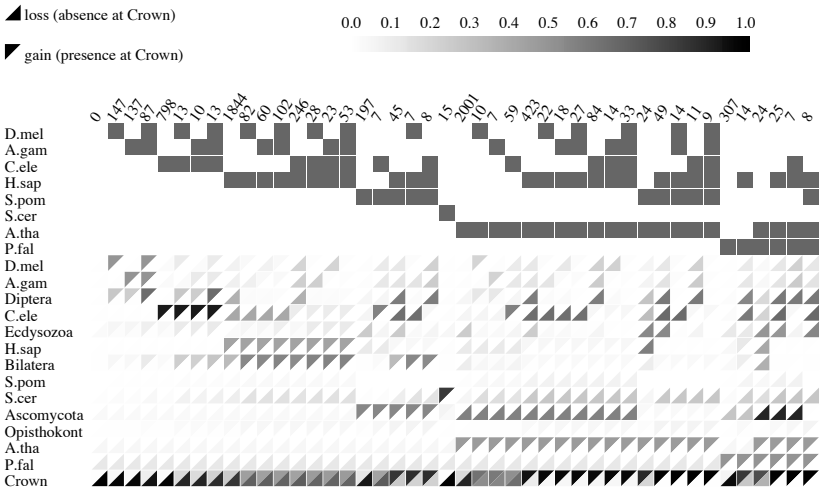

Fig. 4. Likelihoods for gains, losses, and presence at Crown for different characters. Columns correspond to characters: only those that occur at least seven times are shown. Character frequencies are displayed on top of the columns. Rectangles show the intron presence (shaded) or absence (empty) for each character. Shaded triangles show gain and loss posterior probabilities for each edge, and the posterior probabilities of intron presence/absence at the Crown taxon.

For characters that appear frequently in the data, Fig. 4 depicts probabilities for different scenarios. In some cases, the history is clear: if an intron is shared between *D. melanogaster* and *A. gambiae*, then there is a high probability of gain on the branch leading to Diptera, somewhat smaller one on the exterior branches leading to the two species, and some very small probabilities for gaining it earlier. In some cases, the posteriors show a mixture of possible histories: if an intron is present in *D. melanogaster* and *A. thaliana* (there are ten such cases), then it may have been gained more than once, or lost on several branches — which is not a surprising conclusion, but it illustrates the difficulty of choosing between such possibilities based on the intron presence/absence data alone. Notice also that the all-0 characters have no exciting history: most probably, they never had an intron present. Nevertheless, the small probabilities of gain and loss events associated with them add up to visible effects in the mean counts.

(a) Intron counts at interior nodes

Method	Diptera	Ecdysozoa	Bilateria	Ascomycota	Opisthokont	Crown
Dollo parsimony (DP)	732	1081	1613	254	1046	978
Local likelihood (LL)	968	2305	3321	667	1903	1967
Posteriors (P)	895	1762	2380	554	1239	1064
P: 95% confidence	824–962	1484–1972	2055–2669	108–880	965–1450	692–1333

(b) Intron gains and losses on external branches

	D.mel.		A.gam.		C.ele.		H.sap.		S.pom.		S.cer.		A.tha.	
Method	gain	loss	gain	loss	gain	loss	gain	loss	gain	loss	gain	loss	gain	loss
DP	147	156	137	194	**798**	411	**1844**	112	**197**	1	15	**247**	**2001**	46
LL	90	**335**	91	**384**	719	**1555**	849	825	0	**167**	14	**656**	**1726**	760
P	116	**288**	111	**329**	855	1150	**1163**	200	0	**104**	15	**546**	**2157**	286
conf.	±27	±54	±24	±57	±46	±235	±239	±153	0	0–226	±3	102–871	±169	42–487

(c) Intron gains and losses on internal branches

	Diptera		Ecdysozoa		Bilateria		Ascomycota		Opisthokont	
Method	gain	loss	gain	loss	gain	loss	gain	loss	gain	loss
DP	87	**436**	36	**568**	**594**	27	3	**795**	**92**	24
LL	134	**1470**	0	**1005**	**1452**	35	308	**1536**	169	232
P	159	**1024**	141	**752**	**1216**	73	274	**953**	**207**	32
conf.	±60	187–618	0–256	±307	±286	0–151	0–553	±297	0–413	0–72

Table 1. Intron evolution according to different methods. Values in the first row of each table are computed by Dollo parsimony, those in the second are computed by the formulas of Roy and Gilbert. The third row gives the posterior mean counts, computed via Eq. (2) assuming 35000 unobserved intron sites. The fourth row gives 95% confidence intervals for the posterior counts computed in a Monte Carlo procedure (see main text). Tree edges are identified by the nodes they lead to. Edges with a pronounced imbalance (at least 50%) towards gain or loss are emphasized in boldface.

Table 1 compares three optimization criteria. Our estimates for intron counts, gains, and losses are mostly between the two previous estimates. Our likelihood-

based approach gives only slightly more introns at the Crown than parsimony. The branch leading to *C. elegans* has more balanced gains and losses, which result in a net loss that is more modest than in [17]. The *H. sapiens* branch has almost six times as many gains than losses, as opposed to the likelihood calculations of [17] showing a balance. The branch leading to *A. thaliana* has a net gain predicted by all three methods. While we predict more gains on that branch than any of the other methods, the net change is close to what is computed by parsimony, due to more losses. Among the interior branches, we predict a significant net gain over the branch leading to the Opisthokont node, in agreement with parsimony, whereas [17] posit a modest net loss. As for pronounced biases towards gain or loss, our numbers agree with [17] concerning a tendency towards mass losses on a number of edges. At the same time, the mean counts tend to agree with parsimony regarding mass gains. In summary, our mean counts show more changes along the branches than parsimony, but are generally less extreme, and picture less intron-rich ancestral species than [17].

In order to assess the accuracy of the predictions in Table 1, we simulated intron evolution by the Markov model using the parameters optimized on the original data set. The methods were applied to the simulated data sets to estimate intron counts, gains and losses, which could be compared to the exact values observed in the simulation. We generated 1000 synthetic data sets with the same number of observed intron sites in order to assess the estimation error of different methods in our probabilistic model. Figure 5 plots the results of these experiments for some nodes. (For economy, only 100 experiments are shown: 1000 points would require a separate graph for every method at every node.) Our posterior counts generally perform better than the other two methods, which is not surprising in the case of Dollo parsimony (since its assumptions are decidedly different from those of our Markov model), but is more so for the formulas of Roy and Gilbert [16, 17]. These latter usually underestimate intron gains and systematically overestimate the number of ancestral introns. It is also noteworthy that the formulas may sometime result in negative values, which need to be corrected to 0 manually. Dollo parsimony also tends to be biased against gains on internal edges but may overestimate them on external edges (Bilateria-*H. sapiens* edge in particular). It usually underestimates the number of ancestral introns. Aside from their bias, parsimony-based estimates have remarkably low variance. (In the simulations, the vector of ancestral intron counts is distributed multinomially with parameters depending on the likelihood of different characters. The same holds for the vector of intron gains or the vector of intron losses. The estimates of the other two methods have more complex distributions.) Our posterior counts do not seem to have any bias. For ancestral intron counts, the estimates deviate by at most a few hundred from their real values. Specifically, the number of ancestral introns at the common ancestor of animals, plants, and fungi is estimated with an error between (-372) and (+269) in 95% of the cases and a median error of (+11), whereas Dollo parsimony underestimates it by 85 on average (42–134 in 95% of the cases), and the formulas of [17] overestimate

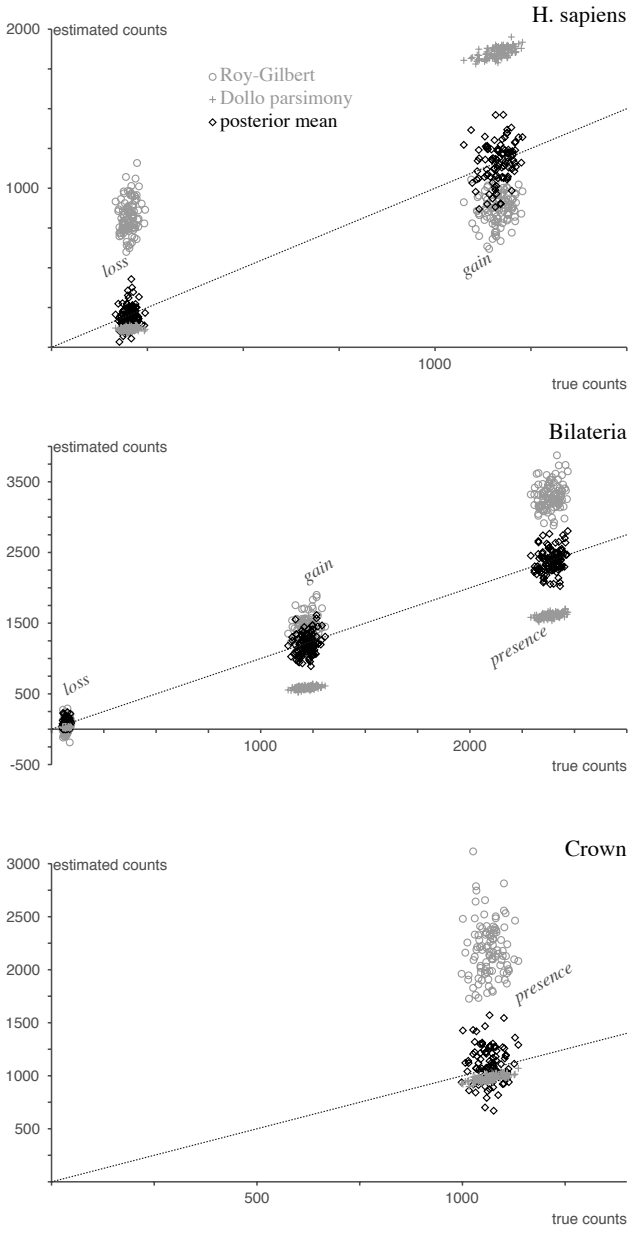

Fig. 5. Estimation error in 100 simulated data sets.

it by 1100 on average (710–1670 in 95% of the cases). The differences observed in the simulations are in fact very similar to those in Table 1.

	D.mel.		C.ele.		H.sap.		A.tha.	
Method	gain	loss	gain	loss	gain	loss	gain	loss
RG	0.7–0.9	1.4–2.0	3.4–4.8	1.6–2.2	2.4–3.3	0.4–0.5	2.2–2.9	0.2–0.3
This paper	17–53	3.3–4.0	33–80	1.4–2.0	140–840	0.6–1.5	44–200	0.3–0.7
(scaled)	(1.4–4.5)		(2.8–6.9)		(11.8–72.8)		(3.8–17)	

Table 2. Rates of intron gain/loss on some external branches estimated by two methods: Roy and Gilbert [17] (RG) and our optimization. Rates for gains are given in units of 10^{-12}, while losses are in units of 10^{-9} per year. The branch lengths (same as in [17] to permit comparisons) are as follows: *D. melanogaster* 250–300 million years (MY), *C. elegans* 500–700 MY, *H. sapiens* 600–800 MY, *A. thaliana* 1500–2000 MY. The gain rates in the RG row are based on an assumption of as many intron sites as nucleotide positions: around 480 thousand, while our calculations are based on the assumption of 35000 unobserved intron sites. This difference amounts to a factor of about 12 between the two gain rate estimates: in parentheses we give numbers scaled to 480 thousand intron sites to permit direct comparison.

Table 2 shows actual intron gain/loss rates calculated by optimization. Using the same actual time estimates for branch lengths as in [17], we computed the gain and loss rates in units of year^{-1}. Our ranges combine the uncertainty of branch lengths in years with 95% confidence intervals, calculated using the parametrized bootstrap procedure mentioned above, involving 1000 simulated data sets. Loss rates are comparable between previous and current estimates, but gain rates tend to be higher in our model. Most notably, gain rate on the branch leading from the MRCA of Bilateria to humans is by at least one magnitude higher than what was estimated in [17].

The Markov model enables predictions about intron dynamics in the future. Figure 6 compares current intron counts to the stationary probabilities for the appropriate branches: the Markov process on edge e converges to a ratio of $\mu_e : \lambda_e$ of intron absence:presence. *D. melanogaster*, *A. gambiae*, and *S. cerevisiae* are very close to equilibrium, but other organisms are farther from it. *C. elegans* is still within 50% of its stationary distribution, but *S. pombe* is losing introns, while humans and thale cress are heading toward much higher intron densities (six and four times as many introns as now, respectively).

5 Conclusion

We described probabilistic techniques for analyzing intron evolution, and applied them to a large data set. The probabilistic analysis assumes a Markov model of intron evolution, in which every intron site evolves independently, obeying the same rates, but the rates may be different on different branches. It is essential to allow for varying rates on branches because the mechanisms underlying intron

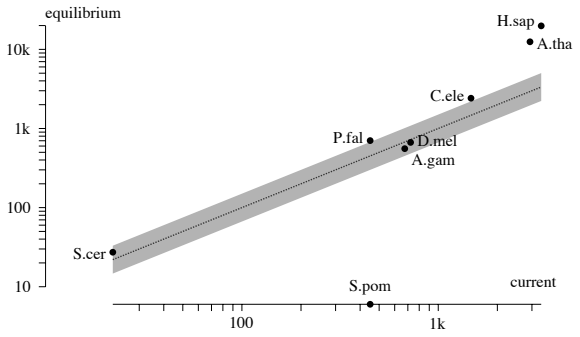

Fig. 6. Current and equilibrium intron counts. The latter are calculated from the stationary probabilities for the branch's Markov process. The dotted line shows identity, with a shaded band of ±50% around it. Species above the identity line are gaining introns, and species below it are losing introns.

gain and loss are fundamentally different, and their intensities vary between different organisms. We demonstrated that the model parameters can be estimated well from observing introns that evolved according to the model, and that the parameters provide sound estimates of ancestral intron counts. We described how posterior estimates can be computed exactly for ancestral intron counts and for gain and loss events. In contrast, Qiu *et al.* [5], relied on a reversible Markov model in which intron gain and loss rates are constant (for a particular gene family) across all branches of the tree. They further employed Markov chain Monte Carlo techniques to estimate posterior distributions.

Our analysis shows a dynamic history of introns, with frequent losses and gains in the course of eukaryotic evolution. We proposed a procedure for estimating unobserved intron sites. This procedure yields a more sound likelihood framework than what was used previously. Applied to the data set, which has 7236 orthologous intron sites, an additional 35000 unobserved intron sites were postulated to explain gains and losses. This equates to an intron site density of about one in every 12 nucleotides, which may characterize preferential intron insertion sites (such as exonic sequence motifs [5] enclosing the intron). All but 28 of 1064 introns present at the eukaryotic Crown node survived in at least one extant species, which means that about one seventh all introns predate the MRCA of animals, plants, and fungi, and the rest were gained more recently. Our counts show that about one third of human introns were gained after the split with Ecdysozoa, another third between that split and the split with fungi, and the rest mostly predate the MRCA of plants and animals.

It is conceivable that our model's assumptions of identical distribution and independence should be replaced by more realistic ones. We plan on exploring richer models in the future by enabling dependence between intron sites in the same gene, and by permitting varying rates among sites. Furthermore, by

combining data analyzed here with new sequences, especially in light of recent analyses of introns in fungi [6] and nematoda [7], one can produce more nuanced results concerning intron evolution by better sampling the phylogenetic tree.

Acknowledgments. This research was supported by NSERC and FQRNT grants. I am grateful to Hervé Philippe, Scott Roy, and Igor Rogozin for valuable comments on this manuscript and on intron evolution in general. I would also like to thank the anonymous referees for their careful reading, and for suggesting the simulated experiments to assess the estimation error.

Supplemental information. The mentioned Java package and various additional analyses can be found at `http://www.iro.umontreal.ca/~csuros/introns/`.

References

1. Crick, W.: Split genes and RNA splicing. Science **204** (1979) 264–271
2. Lynch, M., Richardson, A.O.: The evolution of spliceosomal introns. Current Opinion in Genetics and Development **12** (2002) 701–710
3. de Souza, S.J.: The emergence of a synthetic theory of intron evolution. Genetica **118** (2003) 117–121
4. 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. Current Biology **13** (2003) 1512–1517
5. Qiu, W.G., Schisler, N., Stoltzfus, A.: The evolutionary gain of spliceosomal introns: Sequence and phase preferences. Molecular Biology and Evolution **21** (2004) 1252–1263
6. Nielsen, C.B., Friendman, B., Birren, B., Burge, C.B., Galagan, J.E.: Patterns of intron gain and loss in fungi. PLoS Biology **2** (2004) e422
7. Coghlan, A., Wolfe, K.H.: Origins of recently gained introns in Caenorhabditis. Proceedings of the National Academy of Sciences of the USA **101** (2004) 11362–11367
8. Vaňáčová, Š., Yan, W., Carlton, J.M., Johnson, P.J.: Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis. Proceedings of the National Academy of Sciences of the USA **102** (2005) 4430–4435
9. Ross, S.M.: Stochastic Processes. Second edn. Wiley & Sons (1996)
10. Rzhetsky, A., Ayala, F.J., Hsu, L.C., Chang, C., Yoshida, A.: Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory. Proceedings of the National Academy of Sciences of the USA **94** (1997) 6820–6825
11. Rogozin, I.B., Lyons-Weiler, J., Koonin, E.W.: Intron sliding in conserved gene families. Trends in Genetics **16** (2000) 430–432
12. Felsenstein, J.: Inferring Pylogenies. Sinauer Associates, Sunderland, Mass. (2004)
13. Tuffley, C., Steel, M.: Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of Mathematical Biology **59** (1997) 581–607
14. Roch, S.: A short proof that phylogenetic reconstruction by maximum likelihood is hard. Technical report (2005) `math.PR/0504378` at `arXiv.org`.
15. Chor, B., Tuller, T.: Maximum likelihood of evolutionary trees is hard. In: Proc. Ninth Annual International Conference on Research in Computational Biology (RECOMB). (2005) In press.

16. Roy, S.W., Gilbert, W.: Complex early genes. Proceedings of the National Academy of Sciences of the USA **102** (2005) 1986–1991
17. Roy, S.W., Gilbert, W.: Rates of intron loss and gain: Implications for early eukaryotic evolution. Proceedings of the National Academy of Sciences of the USA **102** (2005) 5773–5778
18. Sverdlov, A.V., Rogozin, I.B., Babenko, V.N., Koonin, E.V.: Conservation versus parallel gains in intron evolution. Nucleic Acids Research **33** (2005) 1741–1748
19. Koshi, J.M., Goldstein, R.A.: Probabilistic reconstruction of ancestral protein sequences. Journal of Molecular Evolution **42** (1996) 313–320
20. Wolf, Y.I., Rogozin, I.B., Koonin, E.V.: Coelomata and not Ecdysozoa: Evidence from genome-wide phylogenetic analysis. Genome Research **14** (2004) 29–36
21. Roy, S.W., Gilbert, W.: Resolution of a deep animal divergence by the pattern of intron conservation. Proceedings of the National Academy of Sciences of the USA **102** (2005) 4403–4408
22. Philippe, H., Lartillot, N., Brinkmann, H.: Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Molecular Biology and Evolution **22** (2005) 1246–1253
23. Philip, G.K., Creevey, C.J., McInerney, J.O.: The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Molecular Biology and Evolution **22** (2005) 1175–1184
24. Le Quesne, W.J.: The uniquely evolved character concept and its cladistic application. Systematic Zoology **23** (1974) 513–517
25. Farris, J.S.: Phylogenetic analysis under Dollo's law. Systematic Zoology **26** (1977) 77–88