Evolution of eukaryotic exon-intron structure

- A Few Facts about Introns
- Sublinear-Time Evaluation of the Likelihood in Markov Models of Sequence Evolution
- MALIN Software Package
- Inferred Characteristics of Intron Evolution

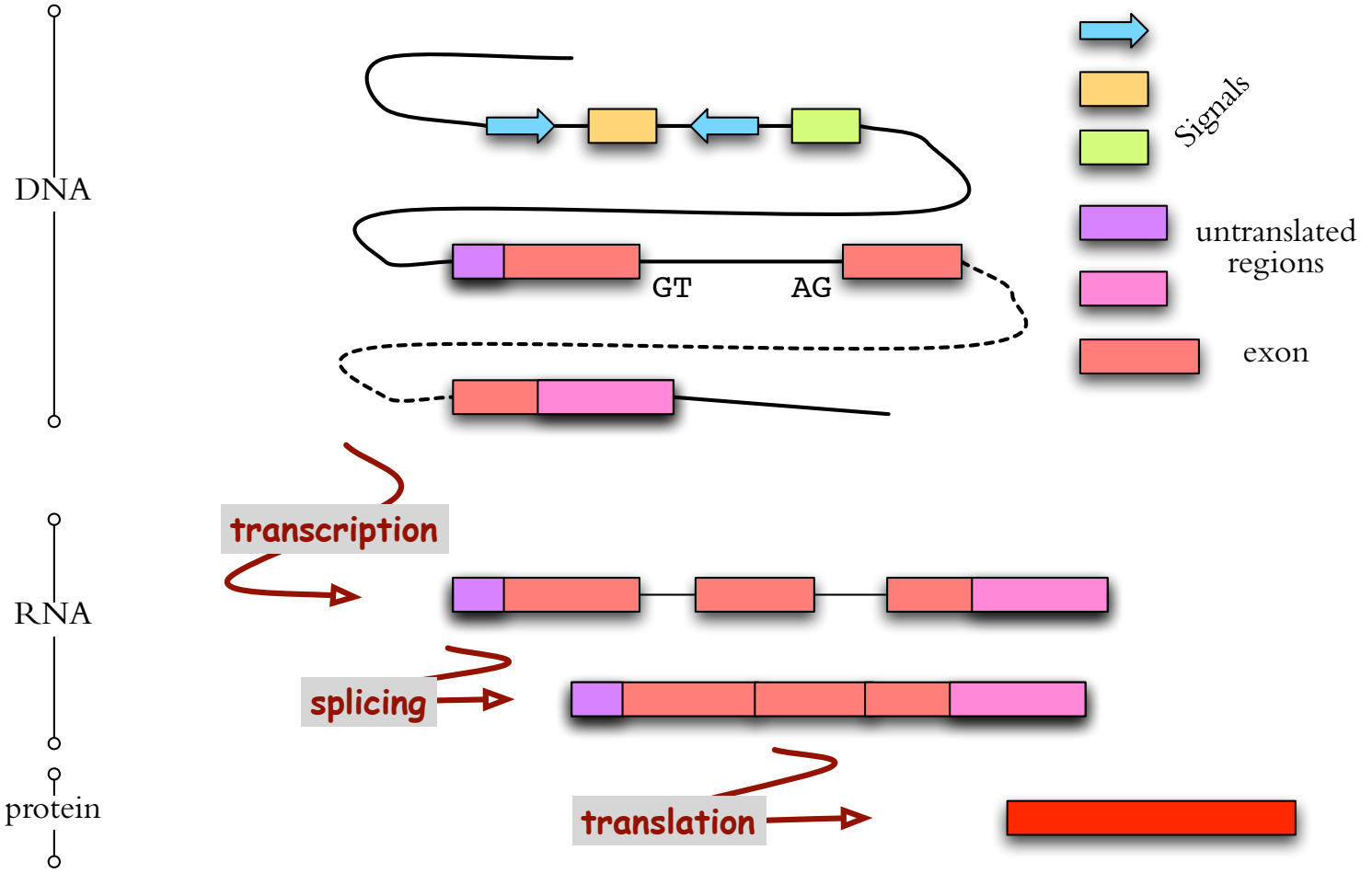
Miklós Csűrös

Department of Computer Science and Operations Research
Université de Montréal

Sabbatical affiliations:
Rényi Institute of Mathematics
Collegium Budapest Institute for Advanced Study

I. A Few Facts

Genes in pieces



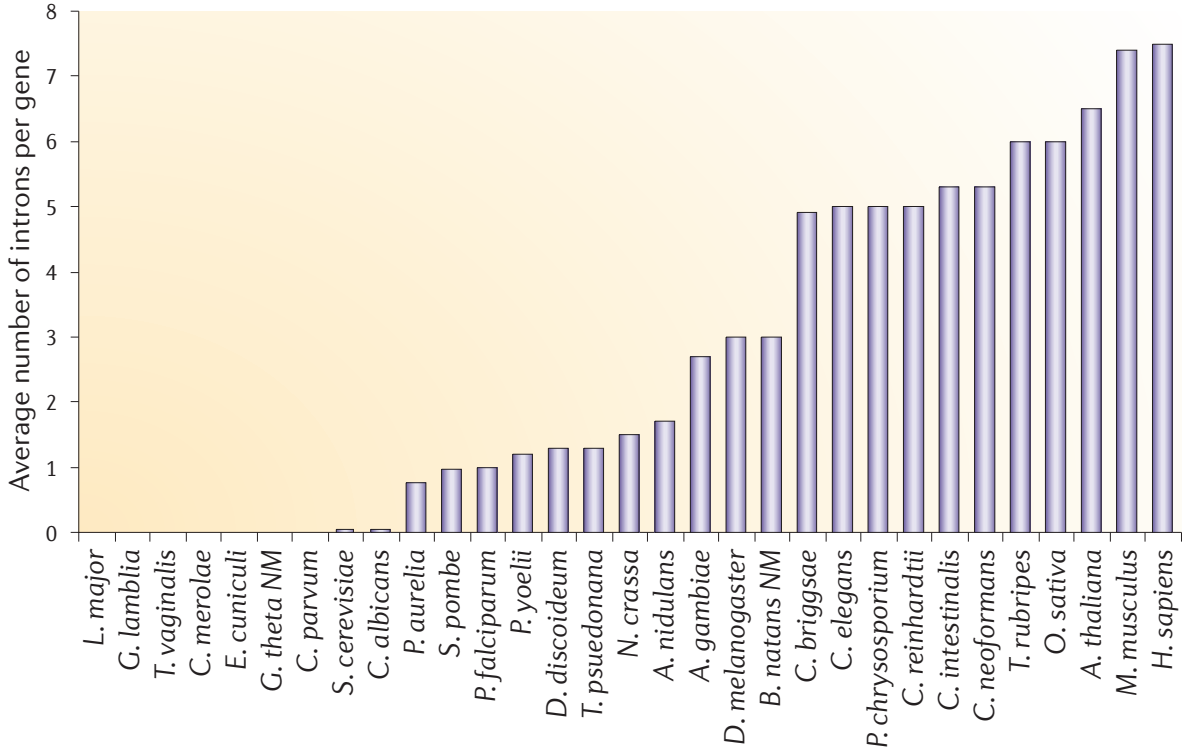
some eukaryotic genes are stitched together from ‘exons’ (as in *expressed*), intervening sequences, called introns, are removed from the messenger RNA

Prevalence of introns

Introns constitute a hallmark of eukaryotic gene organization

Prokaryotes have no spliceosomal introns, and no detectable traces of a spliceosome

Some eukaryotes have very few introns (yeasts), and some have very many (humans)



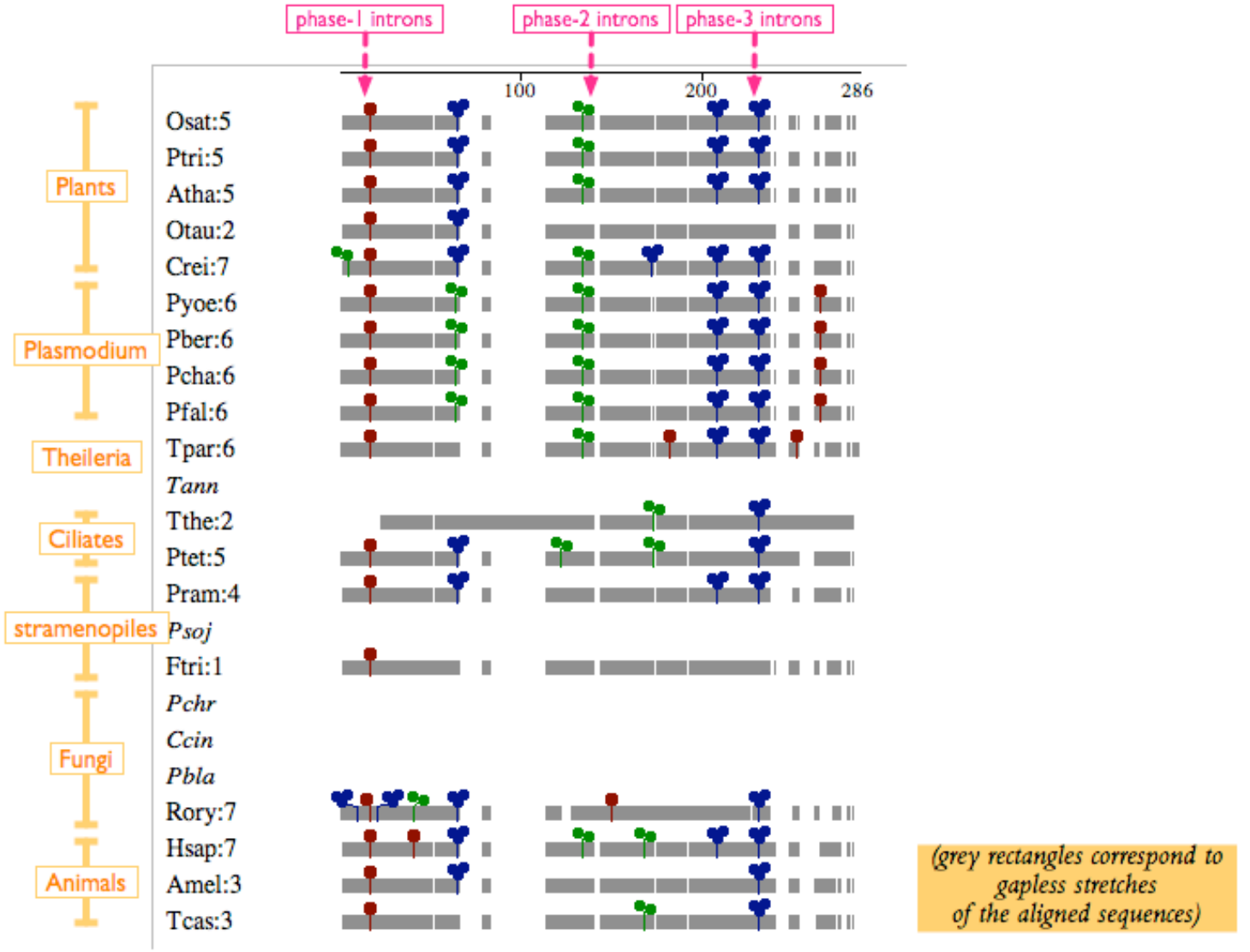
Roy & Gilbert *Nat Rev Genet* 7:211, 2006

Gene structure evolution stretches through eukaryotic evolution

Splicing is very old: [Collins & Penny *Mol Biol Evol* 22:1053, 2005]
Spliceosome: five snRNPs (a small RNA bound by several proteins), plus > 150 associated proteins
$\Rightarrow$ Already present in the last common ancestor of extant eukaryotes (**LECA**)

Gene structure is conserved: [Rogozin, Wolf, Sorokin, Mirkin, Koonin *Curr Biol* 13:1512, 2003]
intron sites are preserved across large evolutionary distances
(e.g., 1/3 of human-Arabidopsis introns coincide)

Structural similarity between orthologs



Some interesting questions

- origin of spliceosomal introns: how and when (with respect to the earliest eukaryotes) did they appear?

- dynamics of intron evolution: mechanisms, quantities and selection of intron loss and gain in lineages

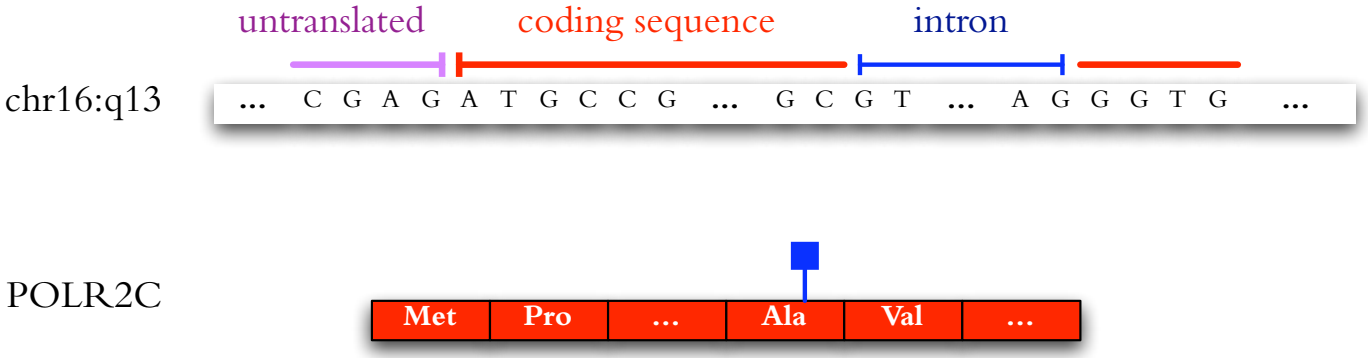
Intron site homology

Intron conservation ⇒ evolution can be traced back until the earliest eukaryotes
. . . if only one could establish homology between introns

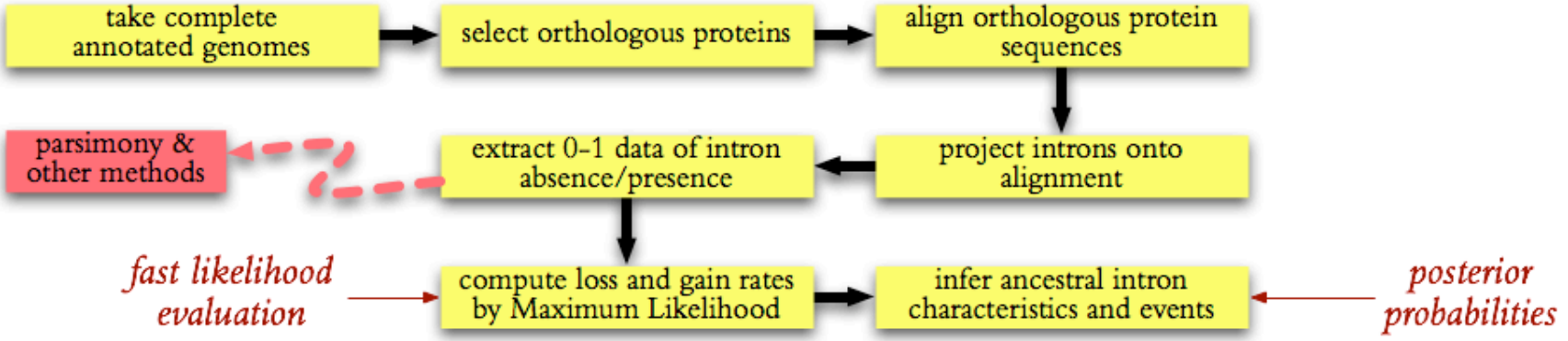
Intron gains and losses are rare (rates of $10^{-12}..10^{-9}$ per year)
but intronic sequences evolve largely neutrally

⇒ not much happens between organisms with alignable non-coding sequences
e.g., 122 intron losses in 17000 genes across human-mouse-rat-dog alignments
(Coulombe-Huntington & Majewski, *Genome Res* 17:23, 2007)

Project intron positions onto protein sequence to establish distant site homology



Analysis of intron data



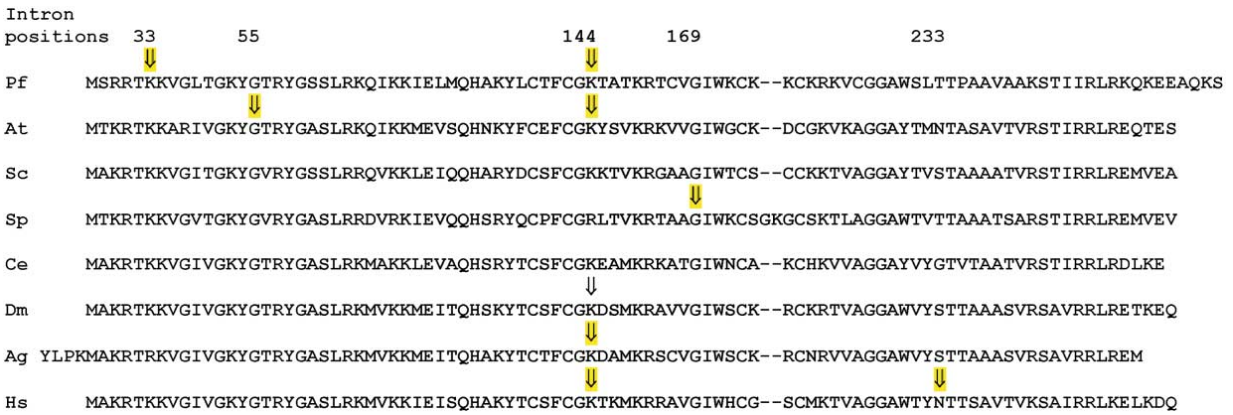

```
Intron
positions  33       55                              144      169
Pf      MSRRTKKVGLTGKYGTRYGSSLRKQIKKIELMQHAKYLCTFCGKTATKRTCVGIWKCK--KCK
At      MTKRTKKARIVGKYGTRYGASLRKQIKKMEVSQHNKYFCEFCGKYSVKRKVVGIWGCK--DCG
Sc      MAKRTKKVGITGKYGVRYGSSLRRQVKKLEIQQHARYDCSFCGKKTVKRGAAGIWTCS--CCK
Sp      MTKRTKKVGVTGKYGVRYGASLRRDVRKIEVQQHSRYQCPFCGRLTVKRTAAGIWKCSGKGCS
Ce      MAKRTKKVGIVGKYGTRYGASLRKMAKKLEVAQHSRYTCSFCGKEAMKRKATGIWNCA--KCH
Dm      MAKRTKKVGIVGKYGTRYGASLRKMVKKMEITQHSKYTCSFCGKDSMKRAVVGIWSCK--RCK
Ag YLPKMAKRTRKVGIVGKYGTRYGASLRKMVKKMEITQHAKYTCTFCGKDAMKRSCVGIWSCK--RCN
Hs      MAKRTKKVGIVGKYGTRYGASLRKMVKKIEISQHAKYTCSFCGKTKMKRRAVGIWHCG--SCM
```

	33	55	144	169	233
Pf	1	0	1	0	0
At	0	1	1	0	0
Sc	0	0	0	0	0
Sp	0	0	0	1	0
Ce	0	0	0	0	0
Dm	0	0	1	0	0
Ag	0	0	1	0	0
Hs	0	0	1	0	1

Rogozin, Wolf, Sorokin, Mirkin, Koonin *Current Biology* 13:1512 (2003)

Intron data

Once you determined which introns are in orthologous positions (from intron-annotated multiple alignments), you have a 0-1 data set (0= absent, 1= present)



	33	55	144	169	233
Pf	1	0	1	0	0
At	0	1	1	0	0
Sc	0	0	0	0	0
Sp	0	0	0	1	0
Ce	0	0	0	0	0
Dm	0	0	1	0	0
Ag	0	0	1	0	0
Hs	0	0	1	0	1

Rogozin, Wolf, Sorokin, Mirkin, Koonin *Current Biology* 13:1512 (2003)

II. Markov Models

Intron evolution - abstraction

intron presence/absence in a homologous position is encoded by 1/0

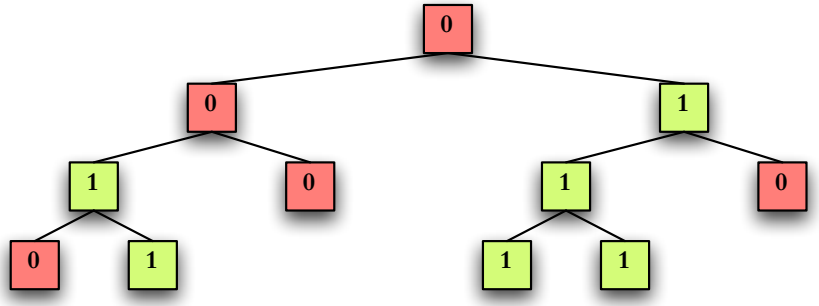
ℓ homologous sites (concatenate data for different genes)

data for an organism is a 0-1 sequence of length ℓ

T evolutionary tree over n organisms (rooted binary tree with labeled leaves)

intron states (0 or 1) evolve along the tree from the root towards the leaves

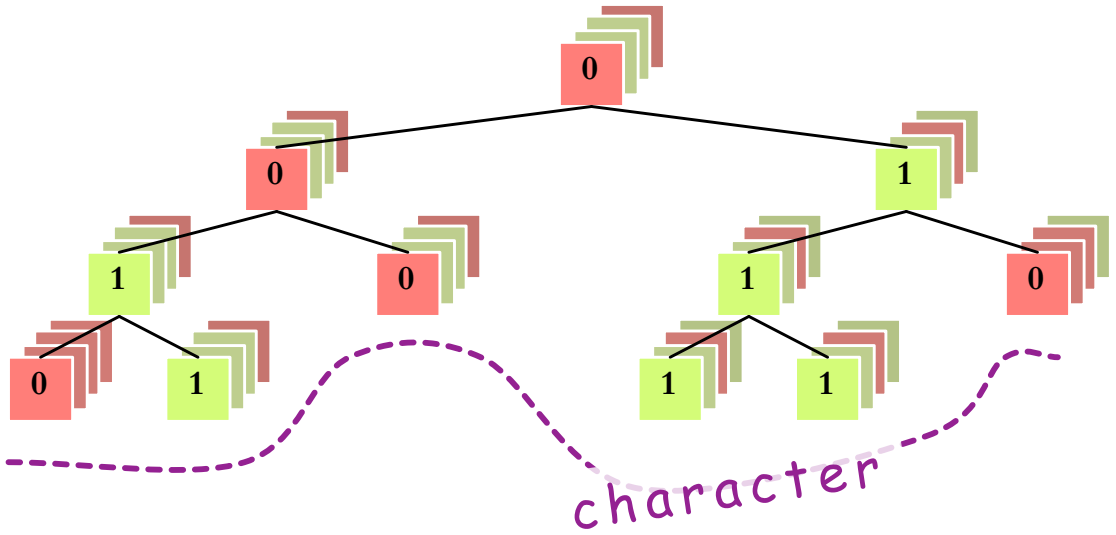
state may change along every branch; model defines the joint distribution of random intron states $\xi(u) \in \{0, 1\}$ for all nodes u



Probabilistic model: assumptions

- parallel intron gains on different tree edges are allowed
- introns evolve independently

$\Rightarrow$ every intron can be analyzed independently



character : vector of leaf states $\xi = \big(\xi(u) : u \in \{\text{leaves of } T\}\big)$ is observable

Probabilistic model: parameters

root state is 1 with probability π_1, or 0 with probability $\pi_0 = 1 - \pi_1$

state transition on edge e with probabilities $p_{0\to1}(e), p_{0\to0}(e), p_{1\to0}(e), p_{1\to1}(e)$

writing with branch length (t), gain (λ) and loss rates (μ):

$$p_{0\to1} = \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu}e^{-t(\lambda+\mu)}$$

- rates (λ, μ) may vary across branches
- can incorporate additional rate variation across sites

(Markov model for binary character)

Likelihood

Notation: $x(u) \in \{0, 1\}$ intron state observed at leaf u

Need to consider all possible states at ancestral nodes for the likelihood

Likelihood for observed states x of an intron site

$$f_x = \mathbb{P}\{\xi = x\} = \sum_{\tilde{x}\,:\,\text{possible states at nodes}} \pi_{\tilde{x}(\mathsf{root})} \prod_{\text{tree edges } uv} p_{\tilde{x}(u)\to\tilde{x}(v)}(uv).$$

Likelihood for the whole data

$$L(x_1, \ldots, x_\ell) = \prod_{i=1}^{\ell} f_{x_i}$$

Unobserved intron sites

Problem: there are no unobserved intron sites ($x = 0^n$) in the data but have non-zero probability in the model

$\Rightarrow$ simply using the presence/absence data at the leaves without all-0 columns introduces a bias in the likelihood optimization (underestimates intron loss)

Solution: compute the likelihood $\mathbb{P}\left\{\text{data} \;\middle|\; \textit{no all-absent sites}\right\}$

Mathematically:

$$L(x_1, \ldots, x_\ell) = \prod_{i=1}^{\ell} \frac{f_{x_i}}{\mathbb{P}\{\xi \neq 0^n\}} = (1 - f_{0^n})^{-\ell} \prod_{i=1}^{\ell} f_{x_i}$$

. . . just like Felsenstein's correction for restriction site data [Felsenstein, *Evolution* 46:159, 1992]

Computing the likelihood: peeling

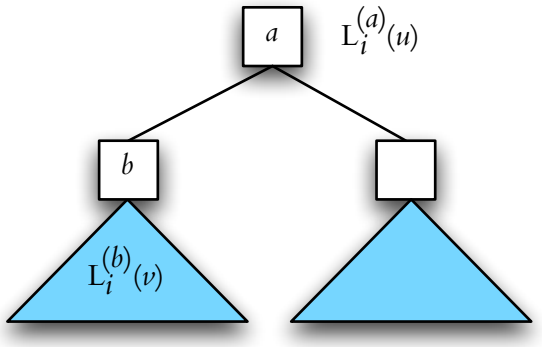
Classic dynamic programming method [Felsenstein 1983] for state set $\mathcal{A} = \{0, 1\}$

Principal tool: conditional likelihoods $L_i^{(a)}(u)$ — probability for leaf states at site i in the subtree of node u, when u is in state a

Recurrence for proceeding from leaves toward the root

$$L_i^{(a)}(u) = \mathbb{I}\{x_i(u) = a\} \qquad \text{when } u \text{ is a leaf,}$$

$$L_i^{(a)}(u) = \prod_{v \in \text{children}(u)} \left(\sum_{b \in \mathcal{A}} p_{a \to b}(uv) L_i^{(b)}(v) \right) \qquad \text{when } u \text{ is not a leaf,}$$



Fast evaluation of the likelihood

Parameters of the model (gain and loss probabilities on edges) are computed by maximizing the likelihood

$\Rightarrow$ likelihood needs to be evaluated many times, taking $\Theta(\ell n)$ time by the peeling algorithm

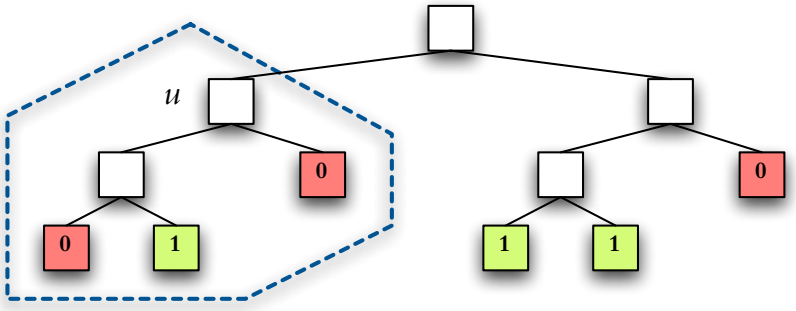
Theorem. The likelihood function can be computed in $O(n\ell/\log \ell)$ time on almost all phylogenies, after a one-time preprocessing step that takes $O(n\ell)$ time.

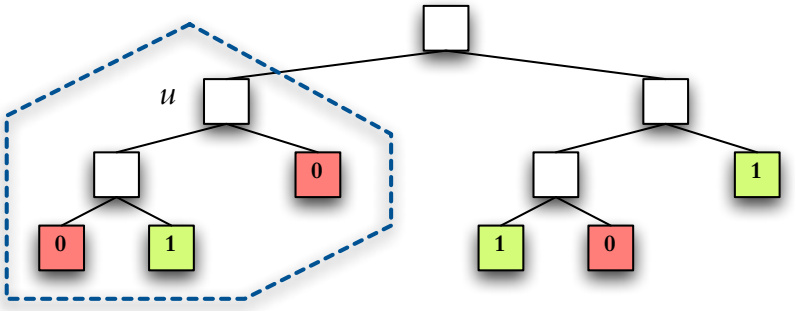
Practice: on intron data, 50–500 times faster than naïve implementation

Details: Csűrös, Holey, Rogozin, ISMB 2007

Fast evaluation of the likelihood

Observation: $L_i^{(a)}(u) = L_j^{(a)}(u)$ if different sample columns x_i and x_j assign the same labels to leaves in the subtree of u





χ_i=010110

χ_j=010101

Idea: first identify different subtree labelings, and then compute the conditional likelihoods $L_x^{(a)}(u)$ where x takes the values of the subtree labelings in the input data

Preprocessing

Preprocessing (compression): given the data $(x_1, \ldots, x_\ell)$,
determine the multiset of observed labelings (i.e., with multiplicities) within each subtree

Thm. The multiset of observed labelings can be computed for all nodes u (simultaneously) in $\Theta(\ell n)$ time.

(Difficulty: one needs to avoid the comparison of length-$O(n)$ vectors at each node, otherwise n^2 factor in the time complexity)

Previous applications:
- Larget and Simon [1998]: $O(n^2)$ label comparisons
- Kosakovsky Pond and Muse [2004]: heuristic ordering of pruning tasks
- Stamatakis et al. [2002]: only identity labelings

Algorithm — evaluation

After preprocessing, evaluating the conditional likelihoods at node u takes $\Theta(r|\mathcal{S}_u|)$ time where $r = |\mathcal{A}|$ is the alphabet size, and $\mathcal{S}_u$ is the set of observed labelings

$\Rightarrow$ computing the likelihood takes $O(rs)$ time where

$$s = \sum_{u \in \{\text{nodes of } T\}} \mathcal{S}_u.$$

Thm.

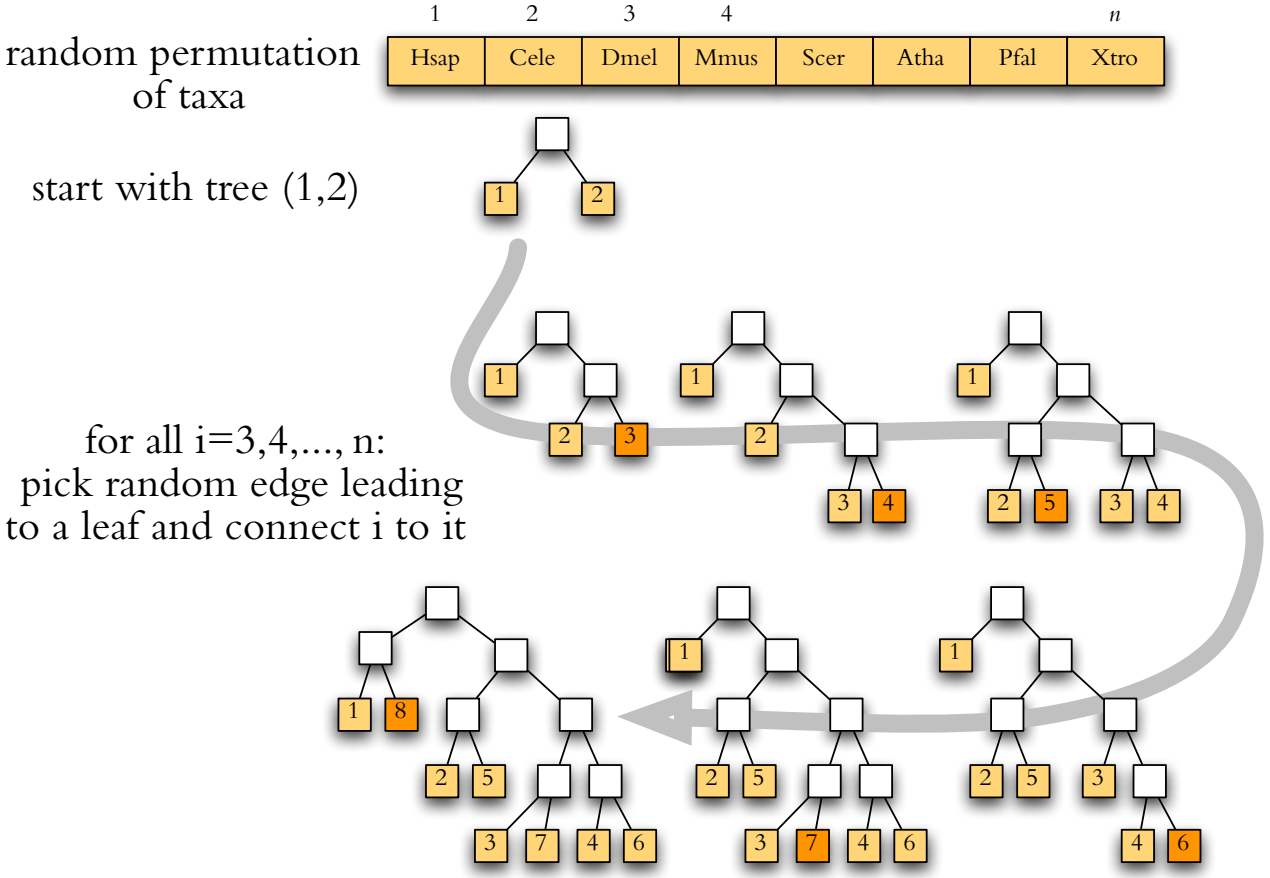
- For random* trees, $s = O(\ell n / \log_r \ell)$ on average
- For random* trees, $s \leq \frac{5\ell n}{1+\log_r \ell}$ with probability $1 - o(n/\log^4 \ell)$

* Yule-Harding model (coalescent)

$\Rightarrow$ after $O(\ell n)$ preprocessing, the likelihood can be evaluated in sublinear $O(\ell n / \log_r \ell)$ time for almost all phylogenies

Yule-Harding model

random tree generation procedure for Yule-Harding distribution of phylogenies



Note: tree shape is determined by the series of leaf selection (here 232431)

Proof

Recall: $s = \sum_{u \in \{\text{nodes of } T\}} \mathcal{S}_u$ where $\mathcal{S}_u$ is the number of different labelings of u's subtree seen in the sample

Now, if there are t leaves in the subtree, then $|\mathcal{S}_u| \leq \min\{\ell, r^t\}$ where r is the alphabet size ($r = 2$ for introns; $r = 3$ with ambiguous characters)

Therefore,

$$s \leq \sum_{k=1}^{n} C_k \min\{\ell, r^k\}$$

where C_k is the number of subtrees with k leaves

The proof for the bound on s follows from the fact that $C_k \approx 2n/k^2$ for all $k < n$.

Number of subtrees

C_k: number of subtrees with k leaves

Thm. For all $1 \leq k < n$ in Yule-Harding model,

- expected value $\mathbb{E}C_k = \frac{2n}{k(k+1)}$, i.e., $\frac{2}{k+1}$ fraction of leaves in size-k subtrees
- C_k is concentrated around its expected value:
  $\mathbb{P}\{|C_k - \mathbb{E}C_k| \geq \epsilon\} \leq 2e^{-2\epsilon^2/2n}$

Proof. Expectation: Devroye [*Random Structures and Algorithms*, 2:303, 1991], revisited by McKenzie & Steel ($k = 2$) and Rosenberg ($k > 2$) in 2000 and 2006, respectively

Concentration: C_k changes by at most 2 if you change one attachment in the leaf selection sequence (subtree prune and regraft) + McDiarmid's inequality for martingales with bounded differences

Even better in practice

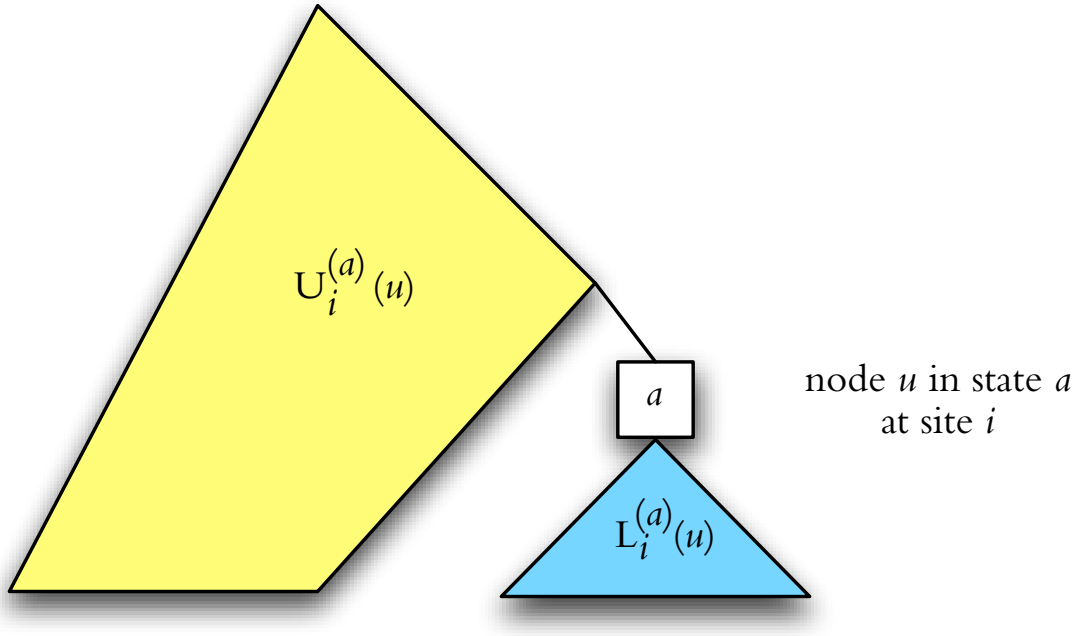
The theorem hold for any data set, but there is even less variation in true data

n	ℓ	r	$n\ell$	$n\lvert\mathcal{S}_{\mathsf{root}}\rvert$	bound	s
8	7236	2	101304	1386	368	183
18	8044	2	273496	19142	16764	1196
47	5216	3	479872	309120	65743	10305
23	10000	4	440000		46124	
47	10000	4	920000		148460	

Fourth column ($n\ell$): original peeling algorithm
Fifth column ($n|\mathcal{S}_{\mathsf{root}}|$): compression at root only
Sixth column: theorem's bound on s
Seventh column (s): exact value

Computing ancestral states and events

Method: (1) compute posterior probabilities for intron presence at nodes, as well as intron state transitions on edges
(2) sum posterior probabilities to obtain expected values for intron density at nodes, as well as losses and gains on edges + correction for absent sites



upper likelihood $U(u)$ computed via similar recursions as $L(u)$
(preorder traversal, use parent's U and siblings' L)

III. A Software Package

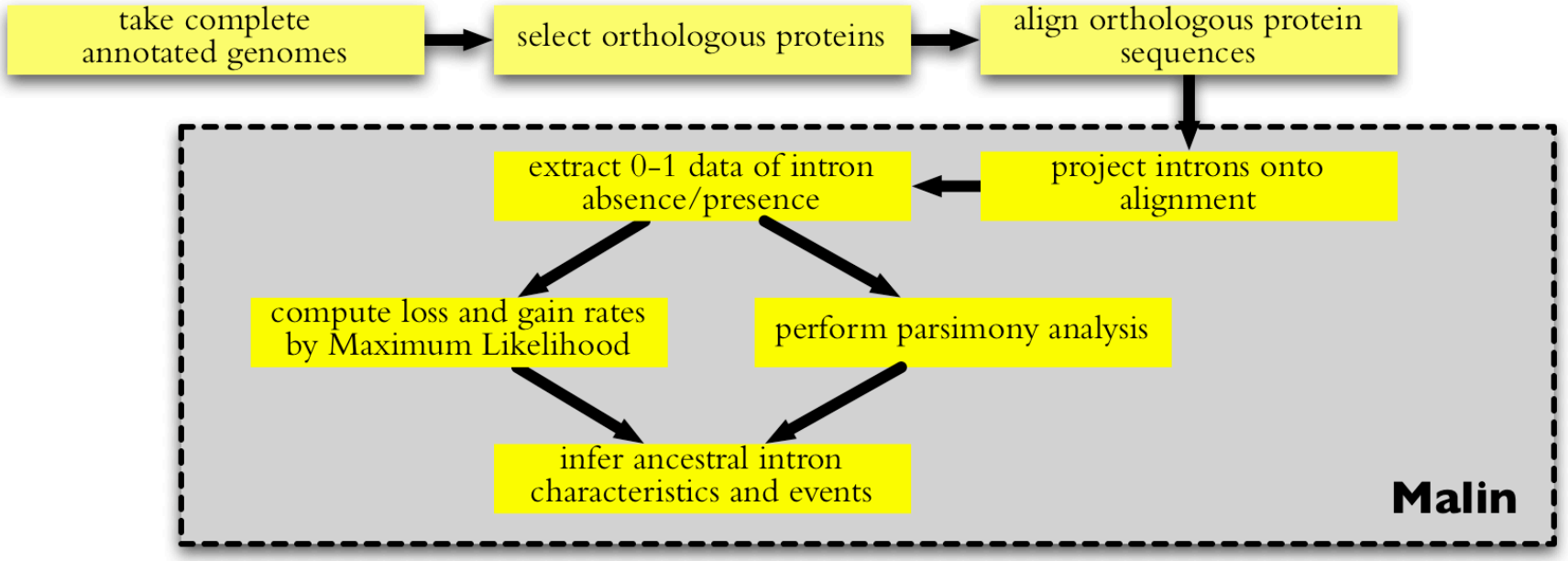
Advertisement: MALIN

MALIN — a software package for the evolutionary analysis of eukaryotic gene structure

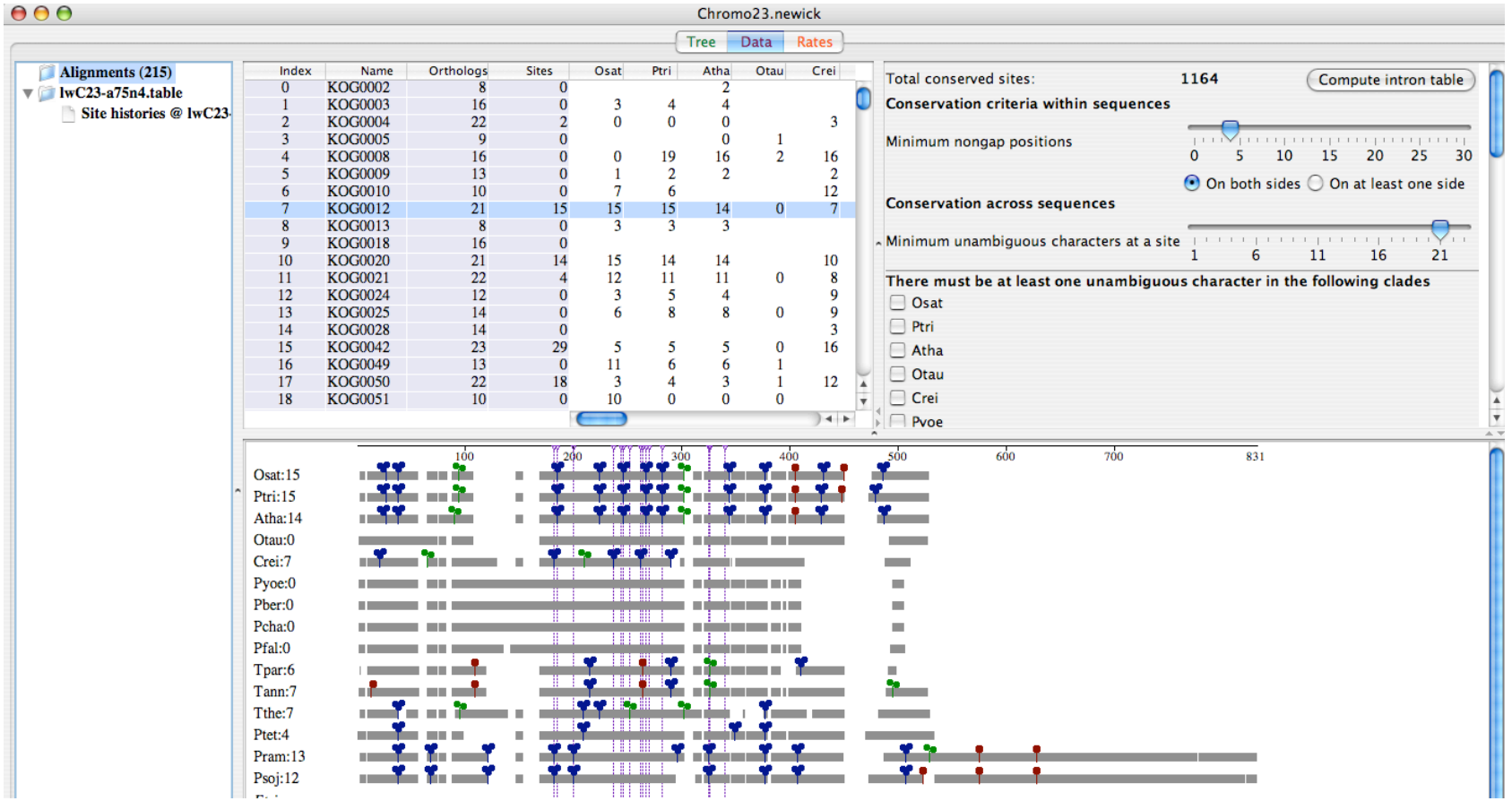
http://www.iro.umontreal.ca/~csuros/introns/malin/

Free Software!

Graphical User Interface!

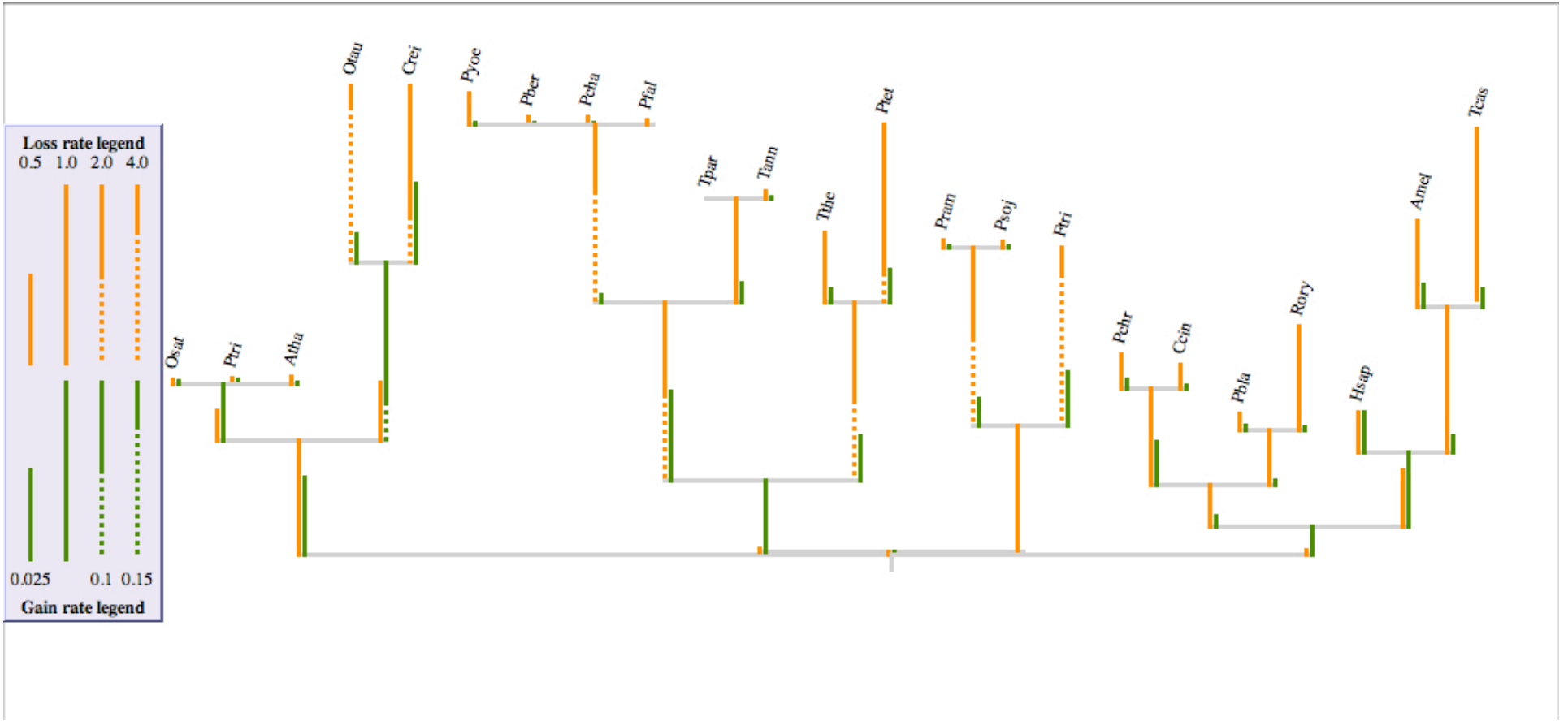


MALIN: alignments



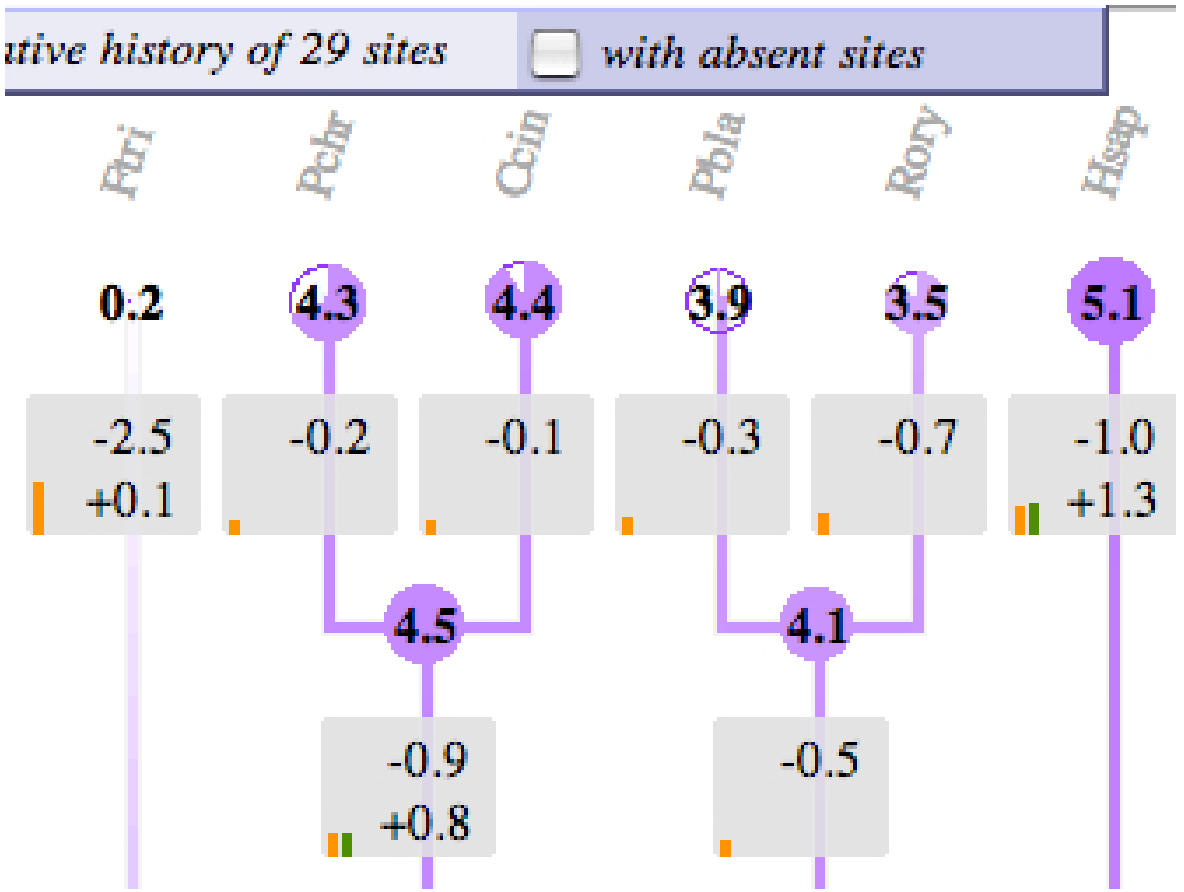
user specifies conservation criteria for intron site homology

MALIN: rates



displaying loss and gain rates

MALIN: histories

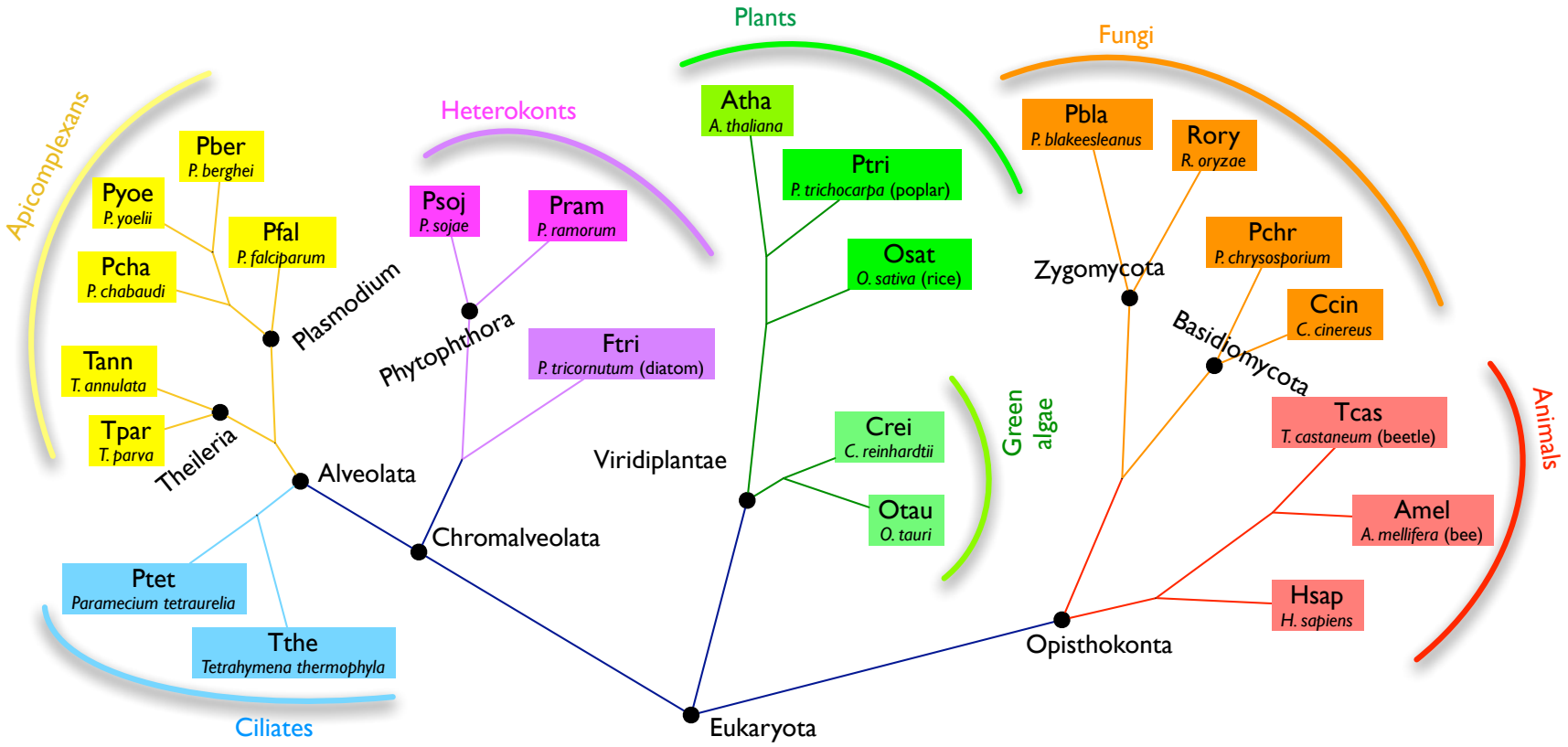


displaying aggregate history for multiple intron sites

IV. Inferred Characteristics

Exon-intron structure in Chromalveolates

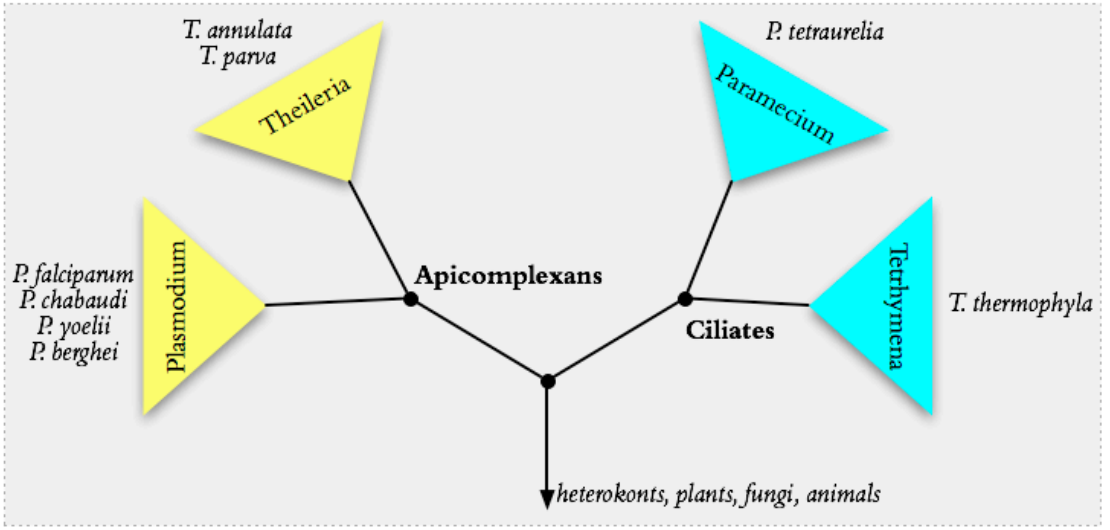
Data: 23 organisms, 392 orthologous gene families, 7030 intron-bearing sites

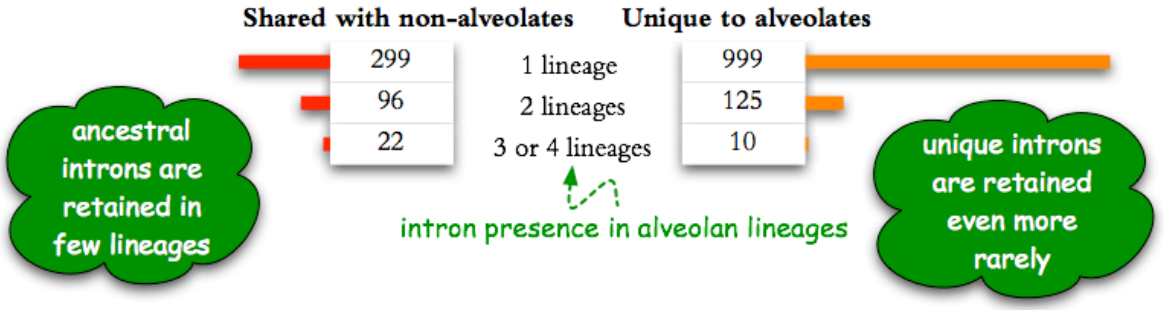


(sources: RefSeq, JGI, MIT/Broad)

Introns are retained with varying intensity

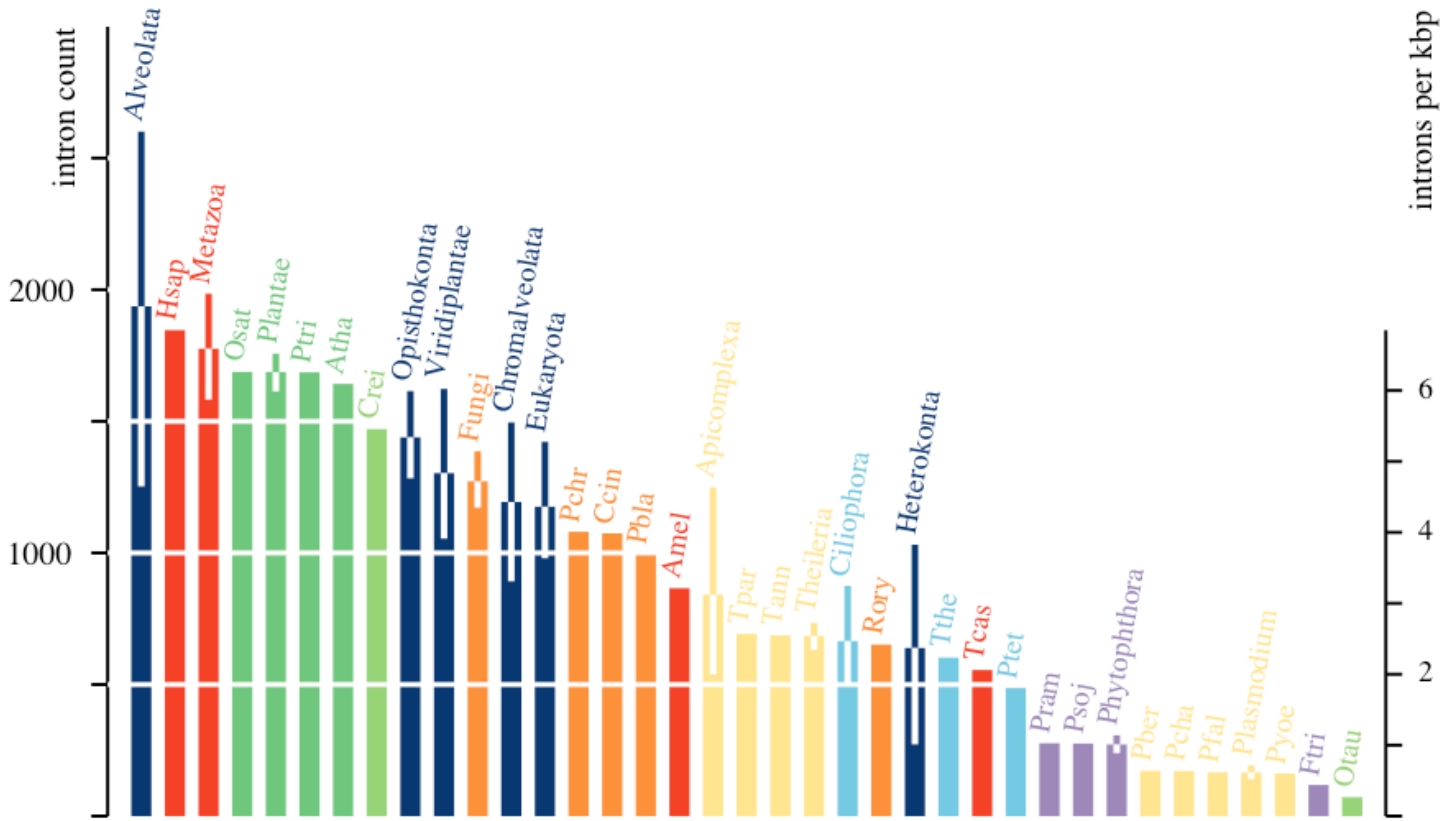
Check the distribution of intron sharing in four alveolate lineages





⇒ high incidence of intron loss, with varying rates at different sites

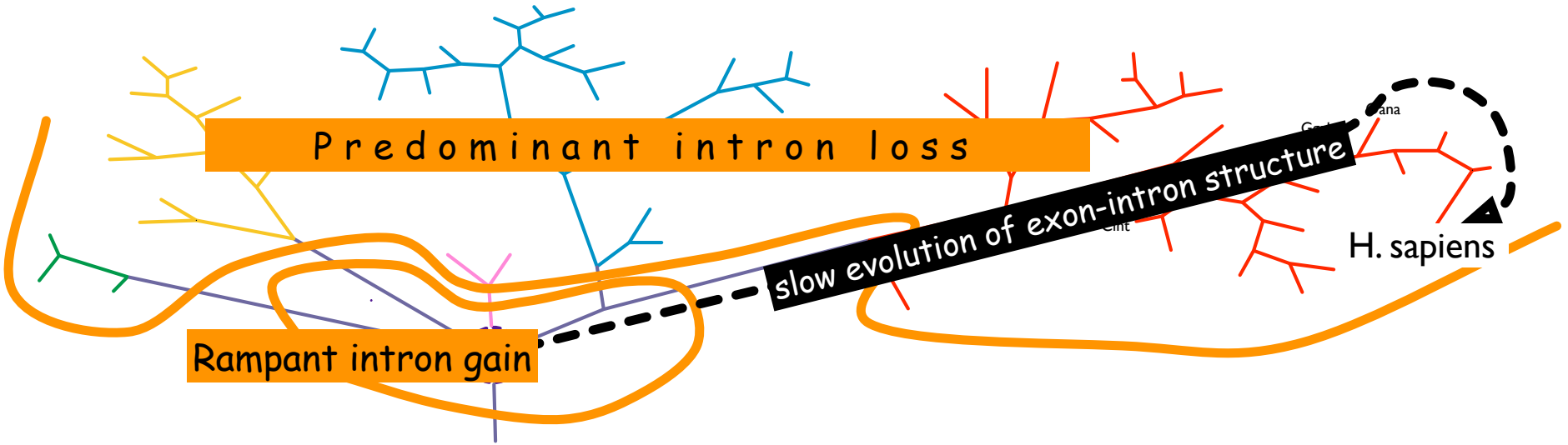
High ancestral intron density



⇒ ancestral alveolate intron density was comparable to humans

methods: continuous-time Markov process for intron gain and loss at a site, branch-specific rates, two loss rate categories, likelihood optimization, posterior probabilities for predicting presence at ancestors, error bars from bootstrap

Conclusion



Collaboration: Igor Rogozin and Eugene Koonin (NCBI), Andrew Holey (College of St. Benedict), István Miklós (Rényi Inst)
Help & advice: Hervé Philippe (U. Montréal), Scott Roy (Harvard-Massey-NCBI) Jacek Majewski (McGill), Liran Carmel (NCBI).

Funding: National Sciences and Engineering Research Council of Canada, NLM/NIH/DHHS Intramural Research program.

http://www.iro.umontreal.ca/~csuros/