Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos

Alexander V. Alekseyenko
alexander.alekseyenko@ucla.edu

June 27, 2008

Classes of Parsimony Models

Wagner

- Main principle: If you can get from state *a* to state *b*, you can return at some point in the future
- Example: A nucleotide or an amino acid site
- Parsimony class: Wagner
- Stochastic counterpart: Continuous-time Markov chains (CTMC)

Dollo

- Main principle: Some states are hard to reach, and/or some states have no return.
- Example: Genomic content traits, e.g. an intron or an exon, a whole gene
- Parsimony class: Dollo
- Stochastic counterpart: Birth/Immigration-death process, absorbing CTMCs

What If No Class Fits?

Allele gain-mutation-loss

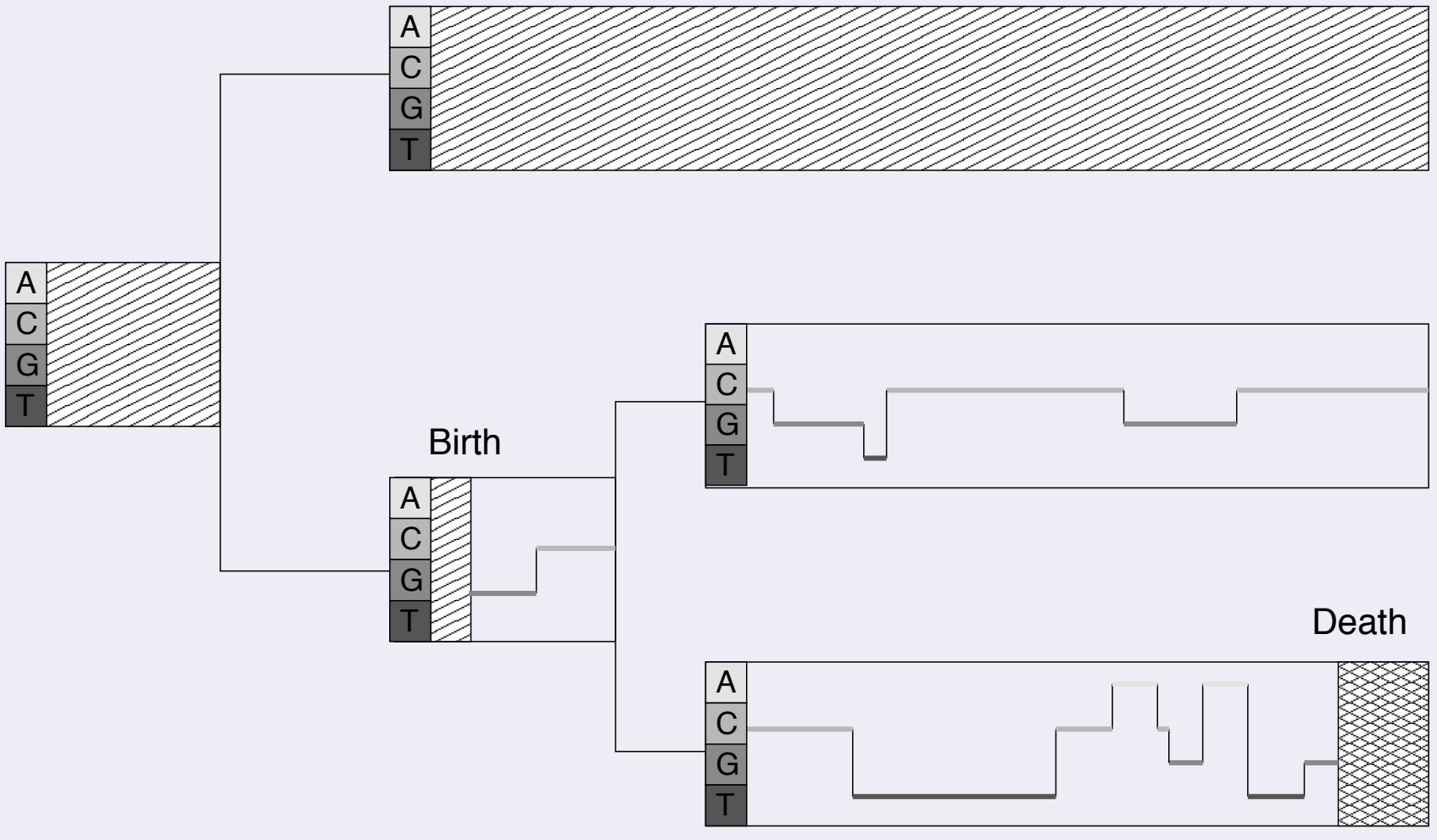
- Several alleles at a locus: A_1, A_2, A_3, A_*
- A_* denotes absence of the marker at the locus
- Gain and loss of marker operates on different scale from mutation
- Loss of a marker is irreversible

An exon sequence and splice-sites

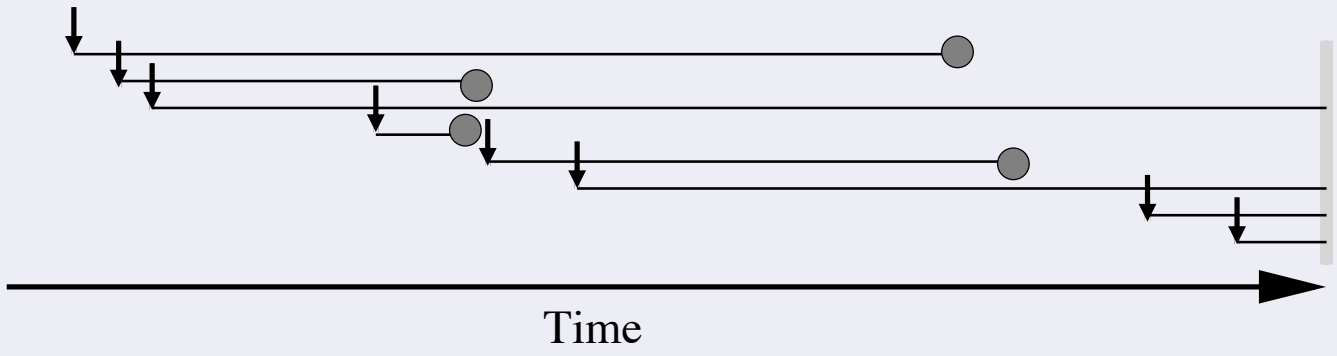
- Mutations in the splice sites are easy to observe in sequences (frequent)
- Gain and loss of homologous exon sequence is a more complex process

Multi-State Stochastic Dollo

Motivating Example



Gain-Loss Process



Parameters

- λ – rate of gain
- μ – per capita rate of loss
- t – time before present

Number of characters observed $N(t)$

- $N(t) \sim \text{Poisson}(\Omega_t)$
- Expected number of characters

$$\begin{aligned} \Omega_t &= \int_0^t \lambda e^{-(t-s)\mu} \mathrm{d}s \\ &= \tfrac{\lambda}{\mu}(1 - e^{-\mu t}). \end{aligned}$$

Tree-time Definition

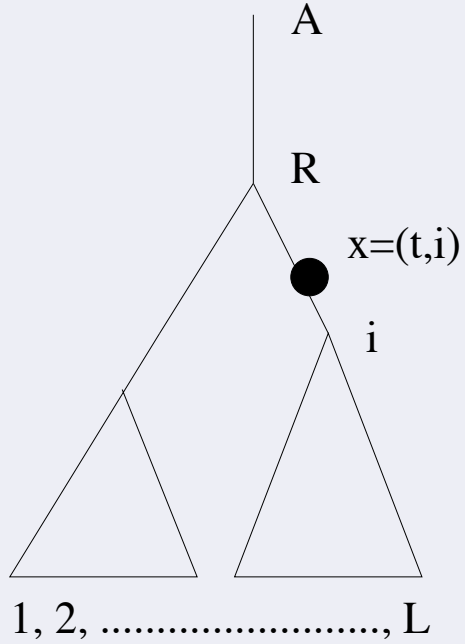
Tree

Let $\mathbf{g} = (\mathcal{V}, E)$ be a phylogenetic tree, directed acyclic

Nodes, $\mathcal{V}$

- L tip nodes;
- R root node;
- A "Adam" node, ancestral to the root;
- internal nodes, such that $\mathbf{g}$ is connected.
- $[\mathbf{g}]$ is collection of all time points on $\mathbf{g}$

$\mathbf{g}$ time

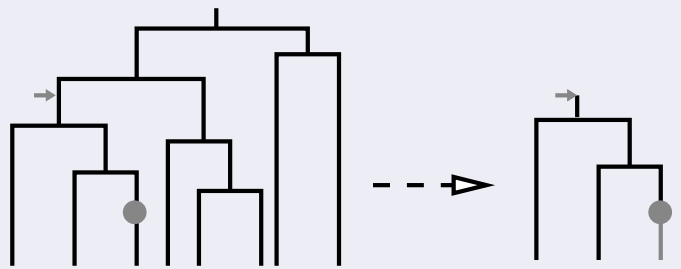


CTMC Likelihood

Data

- Observed data $D = \left\{C^{(j)}\right\}_{1 \leq j \leq N}$;
- Patterns $C = (c_1, \ldots, c_L)$ in L tip species.

Induced Subtree



Main idea

- Condition on gain times
- Integrate out the gain times

$$P(D|\mathbf{g}, \mu, \lambda) = \int_{\mathbf{X} \in [\mathbf{g}]^N} P(D|\mathbf{X}, \mathbf{g}, \mu, S(\mathbf{X})) f(\mathbf{X}|\mathbf{g}, \mu, \lambda) d\mathbf{X}.$$

More on Likelihood

Joint Density of Gains

Joint gain density is given by Poisson point process with intensity $\omega_{\mathbf{g}}(x) = \lambda\, \sigma_{\mathbf{g}}(x)$,

$$f(\mathbf{X}|\mathbf{g}, \mu, \lambda) = \frac{1}{N!} e^{-\Omega_{\mathbf{g}}} \prod_{c=1}^{N} \omega_{\mathbf{g}}(x_c) dx_c.$$

Likelihood Expression After Integration

Sum of partial likelihoods at each node weighted by the probability of gain above that node.

$$P(D|\cdot) \propto \left(\frac{\lambda}{\mu}\right)^N e^{-\Omega_{\mathbf{g}}} \prod_{c=1}^{N} \sum_{\langle i,j\rangle \in E} P(C^{(c)}|(t_i, i), \mathbf{g}, \mu)(1 - e^{-\mu(t_j - t_i)}).$$

Embedded process

Embedding the mutation process $\mathbf{Q}$

If $\mathbf{Q}$ is a standard mutation process (e.g. HKY), then to account for trait loss we extend it with an absorbing death state to arrive at $\mathbf{Q}'$.

$$\mathbf{Q}' = \begin{pmatrix} \mathbf{Q} - \mathbf{I}\mu & \mathbf{1}\mu \\ \mathbf{0}^T & 0 \end{pmatrix} \quad \mathbf{Q}'(t) = \begin{pmatrix} \mathbf{Q}(t) \times e^{-\mu t} & \mathbf{1}(1 - e^{-\mu t}) \\ \mathbf{0}^T & 1 \end{pmatrix}$$

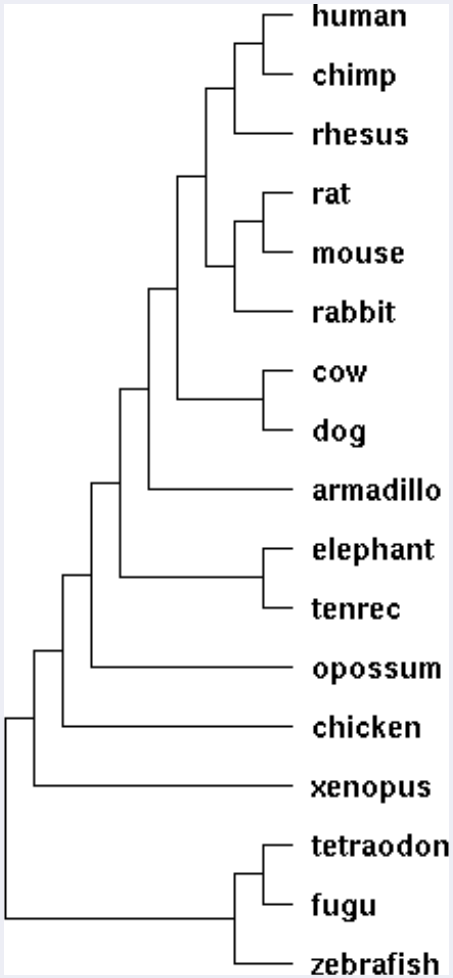
Integrating transitions of the mutation process

Using detailed balance identity we write the partial likelihoods at node i in terms of conditionals of the state in which the trait realizes.

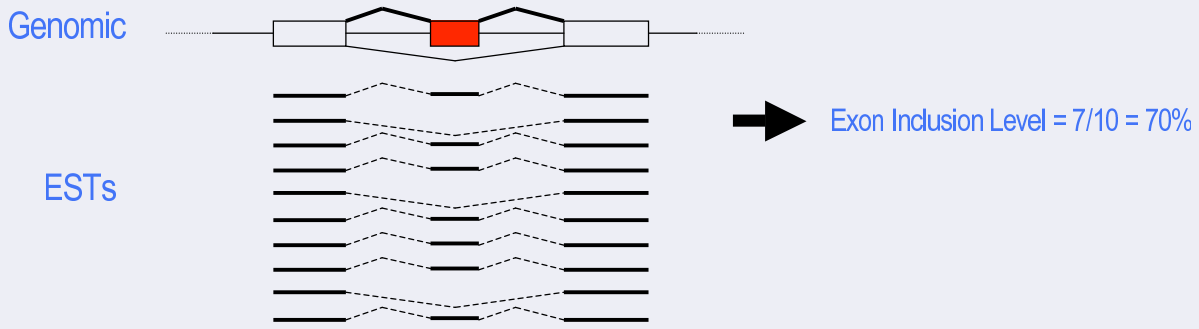
$$P(C^{(c)}|(\tau, i), \mathbf{g}, \mu) = \sum_l \pi_l P(C^{(c)}|(t_i, i), l, \mathbf{g}, \mu)$$

Vertebrate Exon Evolution

17-Species Phylogeny



Splicing and Inclusion Rates

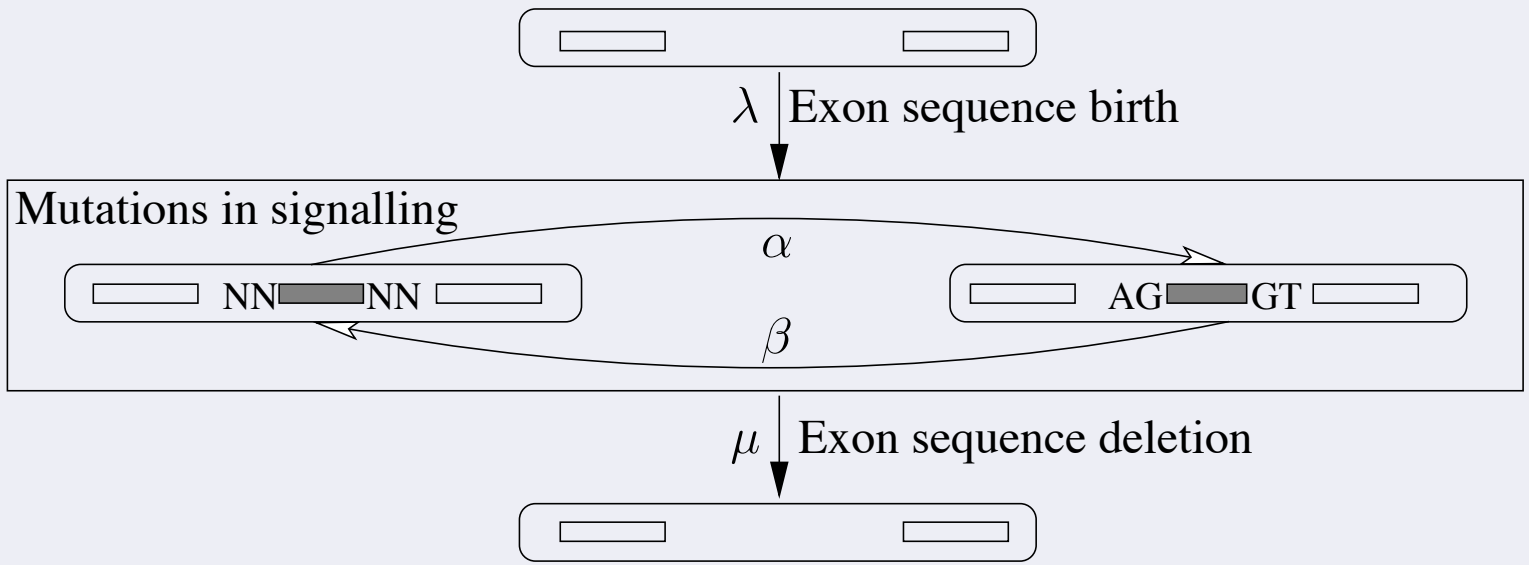


Basic Dataset Description

- Exons are ascertained in Human as Constitutive, Alternative (major, medium, minor inclusion)
- Genomic searches for exon sequence conservation in UCSC genomic alignments
- Splice site mutations are monitored

Alternative Splicing Model

The Model



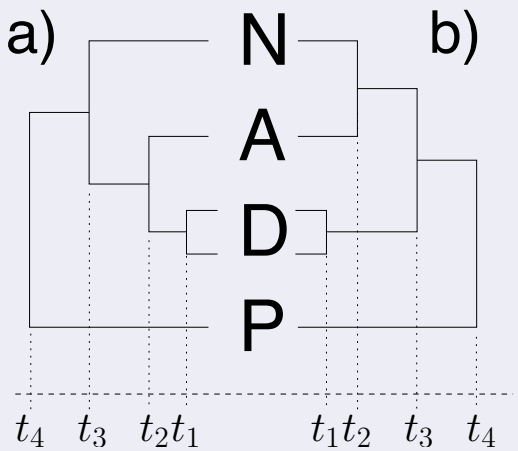
The Results

	(a) Constitutive	(b) Major	(c) Medium	(d) Minor
μ	0.458 (0.451 – 0.465)	0.568 (0.543 – 0.594)	1.63 (1.58 – 1.68)	6.42 (6.27 – 6.57)
α	17.7 (17.2 – 18.2)	17.2 (15.6 – 18.9)	4.20 (3.80 – 4.63)	4.87 (4.23 – 5.55)
β	2.79 (2.75 – 2.84)	2.69 (2.56 – 2.83)	6.95 (6.75 – 7.14)	24.9 (23.9 – 25.9)
π	0.864 (0.861 – 0.866)	0.865 (0.856 – 0.873)	0.377 (0.355 – 0.399)	0.164 (0.148 – 0.180)

Topology Inference Based on Intron Conservation

Which Phylogeny?

a) Coelomata
b) Ecdysozoa



A, arthropod;
D, deuterostome;
N, nematode;
P, plant (outgroup).

Bayes Factor Test

$$B_{ij} = \frac{P(\mathbf{g}_i|D)}{P(\mathbf{g}_i)} \Big/ \frac{P(\mathbf{g}_j|D)}{P(\mathbf{g}_j)} = \frac{P(D|\mathbf{g}_i)}{P(D|\mathbf{g}_j)}.$$

Marginal Likelihood

$$P(D|\mathbf{g}_i) = \int P_{\mathbf{g}_i}(D|\theta)\, f(\theta) d\theta.$$

Harmonic Mean Marginal Likelihood Estimator

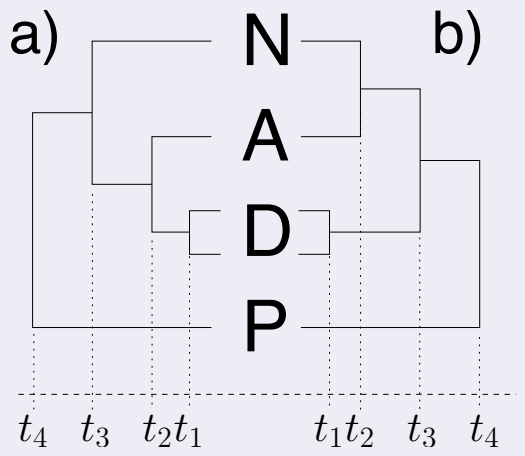
$$P(D|\mathbf{g}_i) \approx \left[\frac{1}{B} \sum_{b=1}^{B} \frac{1}{P(D|\mathbf{g}_i, \Theta^{(b)})} \right]^{-1}.$$

"Ecdysozoa" is Supported by the Data Under MSSD

Results

	(a) "ecdysozoa"	(b) "coelomata"	(c) "complement"
t_1	0.128 (0.117 – 0.140)	0.133 (0.121 – 0.145)	0.134 (0.122 – 0.146)
t_2	0.513 (0.480 – 0.548)	0.712 (0.683 – 0.740)	0.718 (0.690 – 0.747)
t_3	0.704 (0.678 – 0.730)	0.729 (0.701 – 0.757)	0.724 (0.695 – 0.753)
t_4	0.856 (0.828 – 0.886)	0.876 (0.848 – 0.905)	0.876 (0.847 – 0.905)
$\log_{10} P(D)$	-36.62	-65.47	-64.87

Topologies



Acknowledgments

Intellectual Contributions

- Geoff Nicholls;
- Marc Suchard;
- Igor Rogozin (intron conservation dataset);
- Christopher Lee (alternative splicing).

Material Contributions

- Microsoft Research Gift to Marc Suchard;
- NIH Systems and Integrative Biology training grant.