

Gene content evolution by a gene gain-loss-duplication model

István Miklós

***MTA-ELTE Theoretical Biology and Ecology Group, Budapest, Hungary
Genome Analysis and Bioinformatics Group, University of Oxford, UK***

28-29 May, 2005, Stockholm, Sweden

Gene content evolution

Presence-absence models

- Ignores information on copy numbers

Finite state models

- Threshold on the copy number
- Gets obsolete if a new genome discovered with number of gene copies more than the threshold

Unlimited models

- Computational complexity???
- EASY!!!**

Gene gain-loss-duplication model

- Gain (horizontal gene transfer) with rate κ
- Duplication with rate λ , for each gene, independently
- Deletion with rate μ , for each gene, independently

Kolmogorov forward equation

$$\frac{dp_n(t)}{dt} = -(\kappa + n(\lambda + \mu))p_n(t) + (\kappa + (n-1)\lambda)p_{n-1}(t) + (n+1)\mu p_{n+1}(t)$$

Small problems I.

Solve the following infinite differential equation system to get transition probabilities

$$\frac{dp_0(t)}{dt} = -\kappa p_0(t) + \mu p_1(t)$$

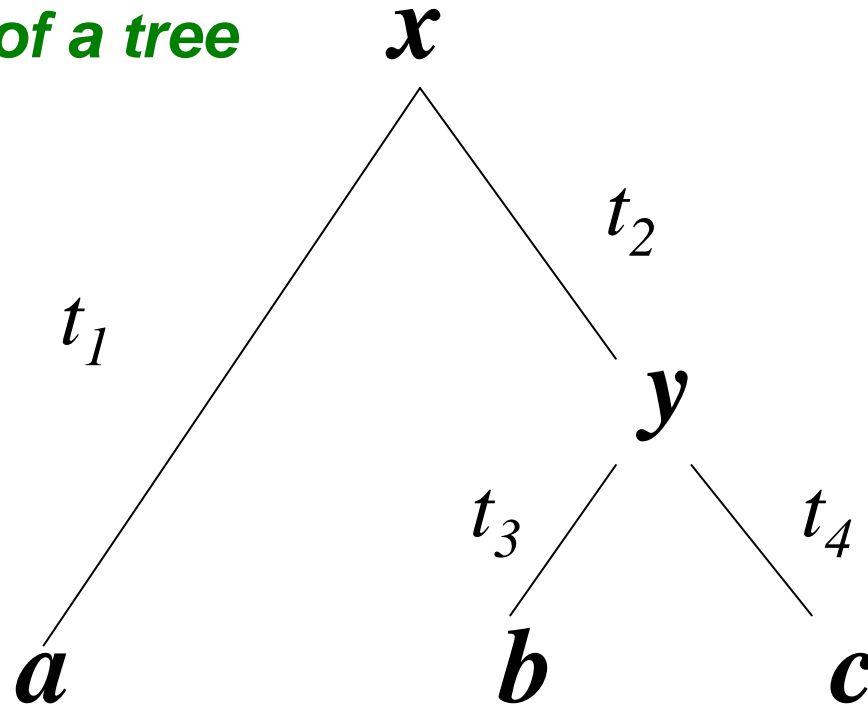
$$\frac{dp_1(t)}{dt} = -(\kappa + \lambda + \mu) p_1(t) + (\kappa + \lambda) p_0(t) + 2\mu p_2(t)$$

⋮

$$\frac{dp_n(t)}{dt} = -(\kappa + n(\lambda + \mu)) p_n(t) + (\kappa + (n-1)\lambda) p_{n-1}(t) + (n+1)\mu p_{n+1}(t)$$

Small problems II.

Calculate the likelihood of a tree



$$\sum_{x=0}^{\infty} \sum_{y=0}^{\infty} P_{t_1}(a | x) P_{t_2}(y | x) P_{t_3}(b | y) P_{t_4}(c | y)$$

Felsenstein's algorithm does not help!!!

Solutions I.

Solving an infinite differential equation system

$$\frac{dp_n(t)}{dt} = -(\kappa + n(\lambda + \mu))p_n(t) + (\kappa + (n-1)\lambda)p_{n-1}(t) + (n+1)\mu p_{n+1}(t)$$

Generating variable ξ , and generating function

$$G(\xi, t) = \sum_{n=0}^{\infty} p_n(t) \xi^n$$

Multiplying the n th equation with ξ^n , and summing them

$$\begin{aligned} \frac{\partial G(\xi, t)}{\partial t} = & -\kappa G(\xi, t) - (\lambda + \mu)\xi \frac{\partial G(\xi, t)}{\partial \xi} + \kappa\xi G(\xi, t) + \\ & + \lambda\xi^2 \frac{\partial G(\xi, t)}{\partial \xi} + \mu \frac{\partial G(\xi, t)}{\partial \xi} \end{aligned}$$

$$\frac{\partial G(\xi, t)}{\partial t} + (-\lambda \xi^2 + (\lambda + \mu)\xi - \mu) \frac{\partial G(\xi, t)}{\partial \xi} = \kappa(\xi - 1)G(\xi, t)$$

Solving with the method of Lagrange

$$\frac{dt}{1} = \frac{d\xi}{-\lambda \xi^2 + (\lambda + \mu)\xi - \mu} = \frac{dG}{\kappa(\xi - 1)G}$$

$$\int (\mu - \lambda) dt = \int \left(\frac{1}{\xi - 1} + \frac{\lambda}{\mu - \lambda \xi} \right) d\xi \quad \text{has solutions} \quad e^{-(\mu - \lambda)t} \frac{\xi - 1}{\mu - \lambda \xi} = C_1$$

$$\int \frac{d\xi}{\mu - \lambda \xi} = \int \frac{dG}{\kappa G} \quad \text{has solutions} \quad G(x, t) (\mu - \lambda \xi)^{\frac{\kappa}{\lambda}} = C_2$$

General solution

$$G(\xi, t)(\mu - \lambda\xi)^{\frac{\kappa}{\lambda}} = \Phi\left(e^{-(\mu-\lambda)t} \frac{\xi - 1}{\mu - \lambda\xi}\right)$$

We are interested in the particular solution $G(\xi, 0)=1$, which is satisfied for

$$\Phi(a) = \left(\frac{\mu - \lambda}{\lambda a + 1}\right)^{\frac{\kappa}{\lambda}}$$

This yields

$$G(\xi, t) = \left(\frac{\mu - \lambda}{\mu - \lambda e^{-(\mu-\lambda)t} - \lambda(1 - \lambda e^{-(\mu-\lambda)t})\xi}\right)^{\frac{\kappa}{\lambda}}$$

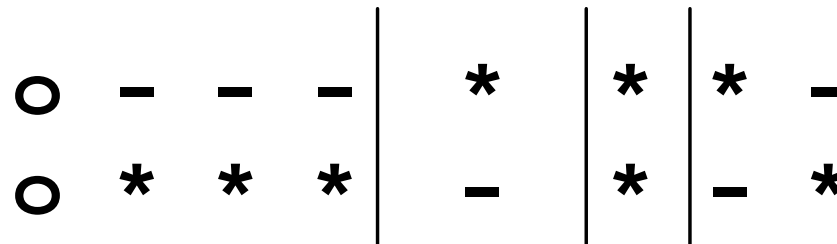
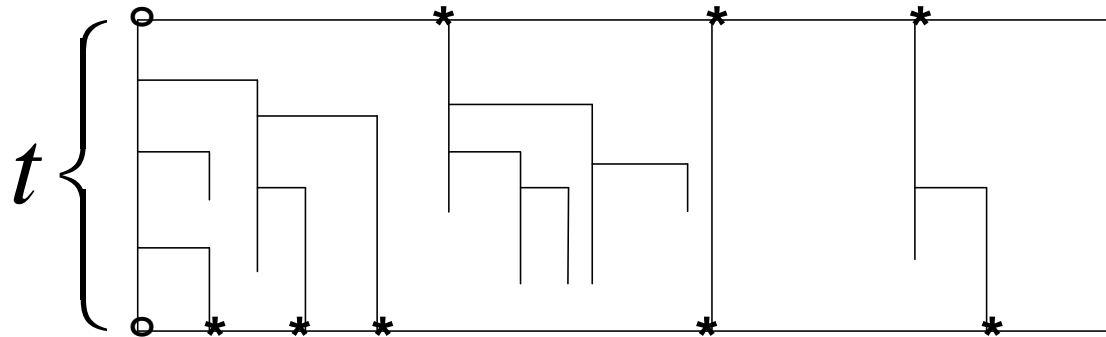
The Taylor series of $G(\xi,t)$ gives the solutions for $p_n(t)$

$$p_n(t) = \frac{\Gamma\left(\frac{\kappa}{\lambda} + n - 1\right)}{n!} (1 - \lambda\beta(t))^{\frac{\kappa}{\lambda}} [\lambda\beta(t)]^n$$

where Γ is the generalized factorial function and

$$\lambda\beta(t) = \frac{1 - e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}}$$

Galton-Watson forest and pseudo-alignment



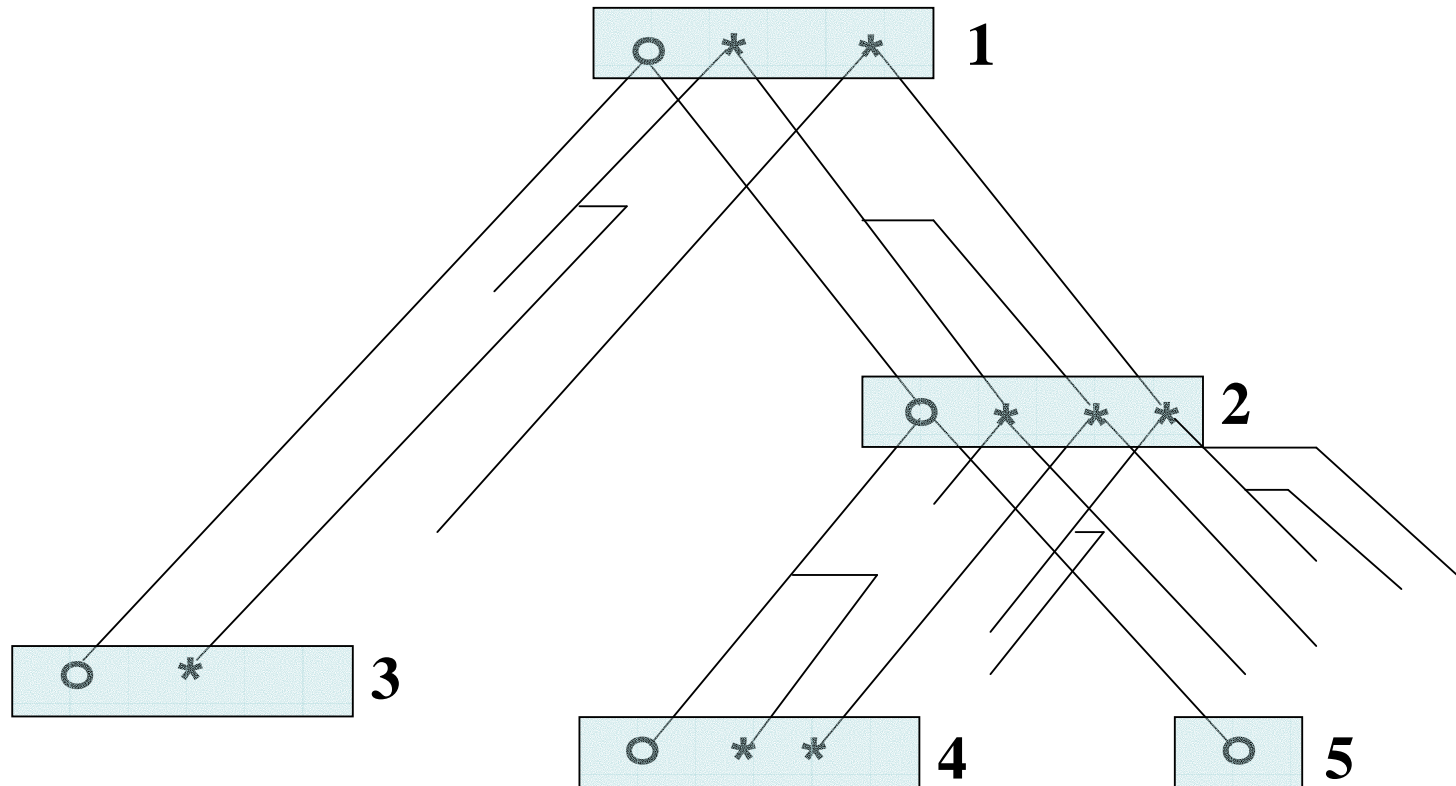
$$p_n(t)$$

$$p_0(t) = \mu\beta(t)$$

$$p_n(t) = (1 - \mu\beta(t))(1 - \lambda\beta(t))[\lambda\beta(t)]^{n-1}$$

Independent fate of Galton-Watson trees \rightarrow the probability of a pseudo-alignment is the product of probabilities of patterns

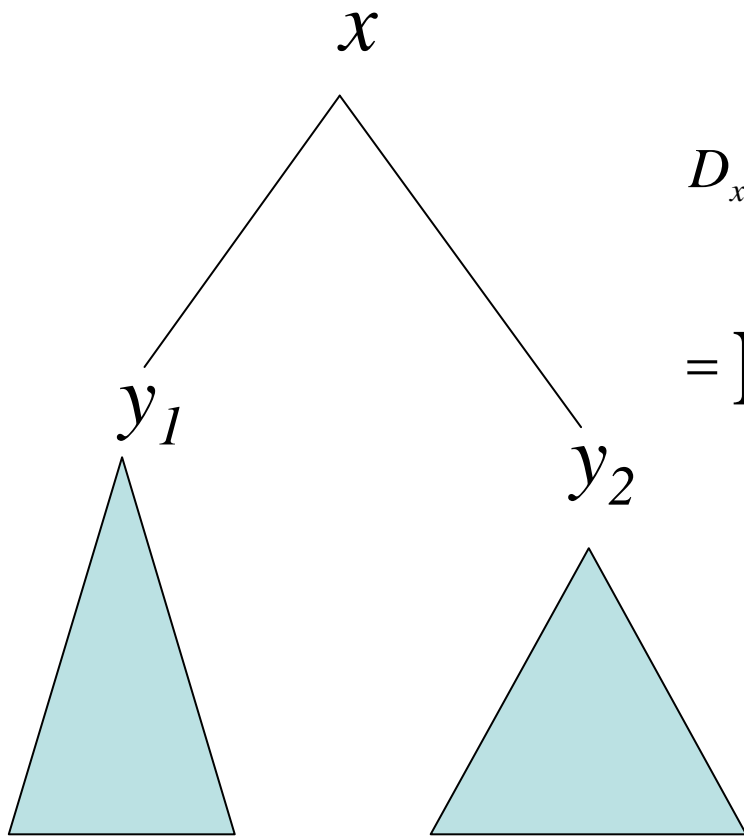
Galton-Watson superforest



Only finite many genes at internal nodes can have any descendants at any leaves.

Folding probabilities I.

Probability that a gene dies out on all leaves



$$D_x = \prod_{j=1}^2 \left(\mu(t) + \sum_{i=1}^{\infty} (1 - \mu\beta(t))(1 - \lambda\beta(t)) [\lambda\beta(t)]^{i-1} D_{y_i}^i \right) =$$
$$= \prod_{j=1}^2 \left(\mu(t) + \frac{(1 - \mu\beta(t))(1 - \lambda\beta(t)) D_{y_j}}{1 - \lambda\beta(t) D_{y_j}} \right)$$

Folding probabilities II.

Effective probability

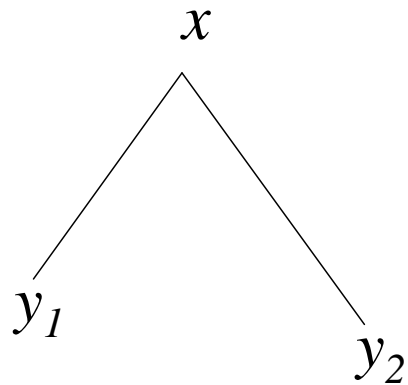
Having n descendants surviving + some other dieing out.

$$\begin{aligned} p_n^{eff}(t) &= \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{\kappa}{\lambda} + n + i - 1\right)}{(n+i)!} (1 - \lambda\beta(t))^{\frac{\kappa}{\lambda}} [\lambda\beta(t)]^{n+i} \binom{n+i}{i} D_y^i = \\ &= \frac{\Gamma\left(\frac{\kappa}{\lambda} + n - 1\right)}{n!} \frac{(1 - \lambda\beta(t))^{\frac{\kappa}{\lambda}} [\lambda\beta(t)]^n}{(1 - \lambda\beta(t)D_y)} \end{aligned}$$

Dynamic programming I.

Effective probability

There are m_1 surviving genes at node y_1 , m_2 surviving genes at node y_2 , given that there are n surviving genes at node x .



x	○	-	-	-	*	*	-	*	-
y_1	○	*	*	-	-	*	-	*	*
y_2	○	-	-	*	*	*	*	*	-

- Each gene at node x must have at least one gene in at least one child.
- Triplewise dynamic programming
- To avoid over-counting, genes first in y_1 , then in y_2 : two layer DP table.

Dynamic programming II.

Conditional likelihoods

$$L_x(n) = \sum_{m_1=0}^{M_1} \sum_{m_2=0}^{M_2} T_x(m_1, m_2, n) L_{y_1}(m_1) L_{y_2}(m_2)$$

where M_1 and M_2 are the sum of genes on the leaves of left and right sub-trees, and $T_x(m_1, m_2, n)$ is the transition probability introduced on the previous slide

Equilibrium distribution

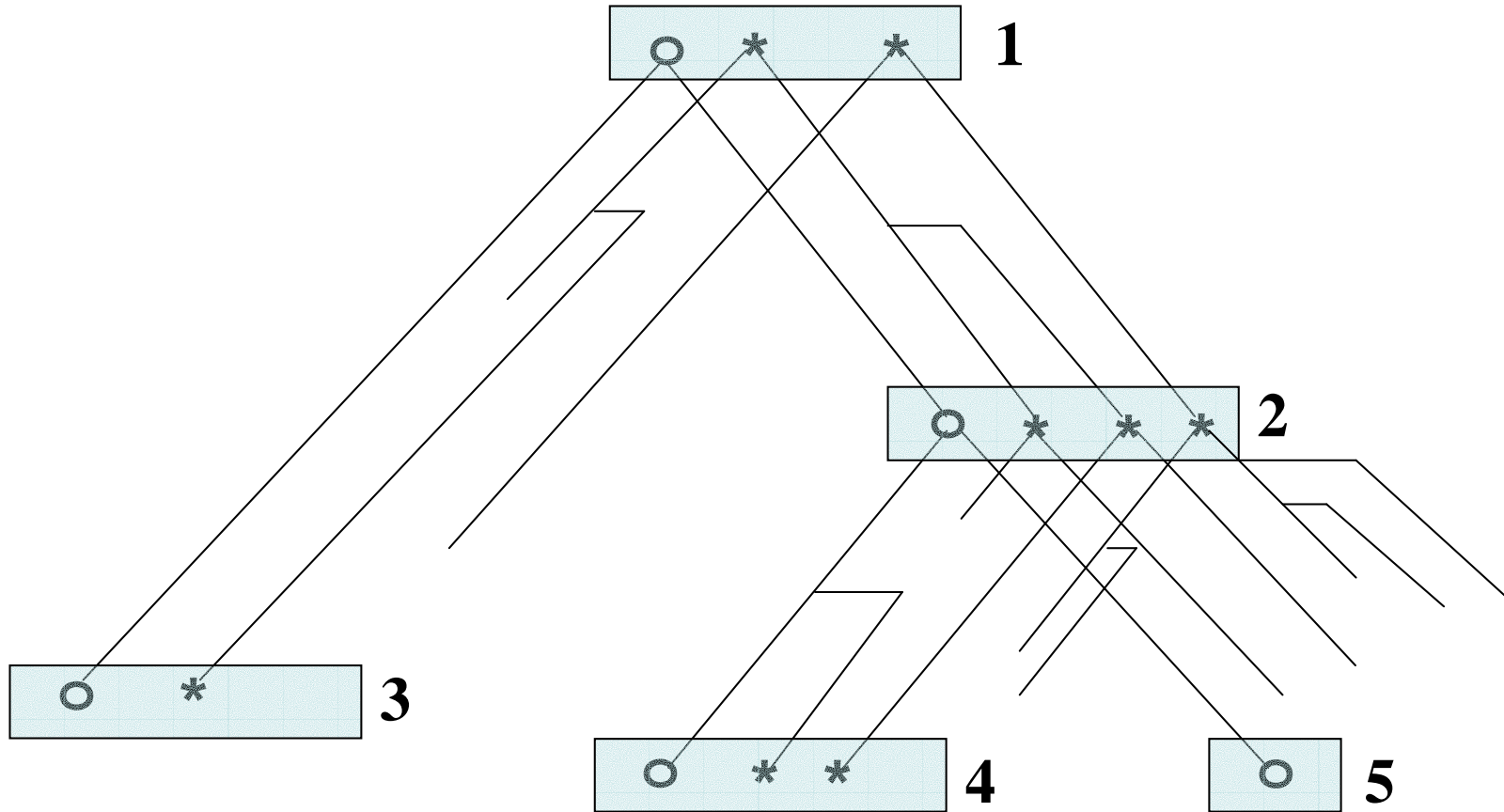
$$\gamma_n = \frac{\Gamma\left(\frac{\kappa}{\lambda} + n - 1\right)}{n!} \left(1 - \frac{\lambda}{\mu}\right)^{\frac{\kappa}{\lambda}} \left(\frac{\lambda}{\mu}\right)^n$$

Likelihood of the tree

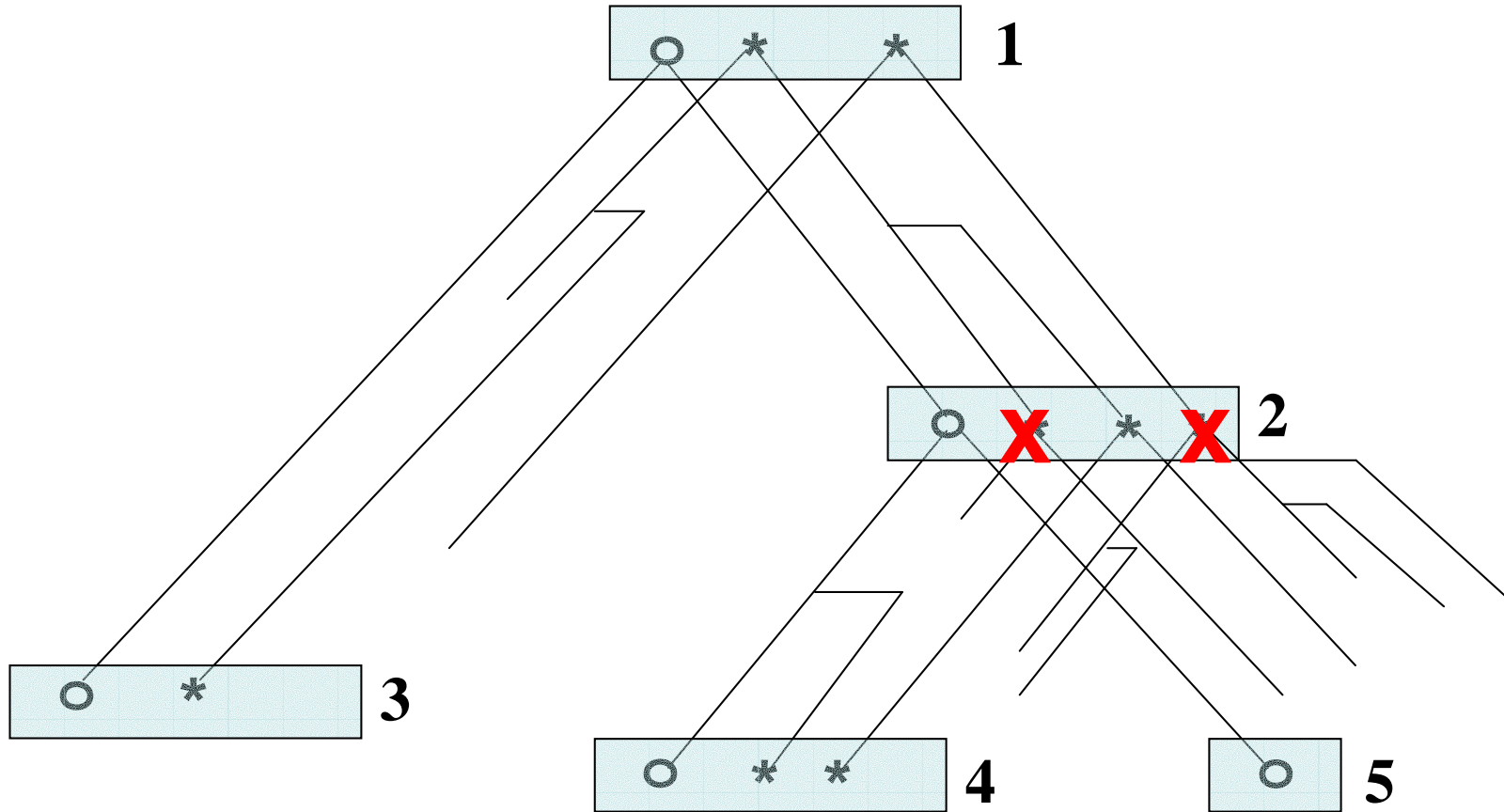
Folding the probabilities of dieing out genes at the root

$$L = \sum_{n=0}^{M_{root}} \sum_{i=0}^{\infty} L_{root}(n) \gamma_{n+i} \binom{n+i}{i} D_{root}^i =$$
$$\sum_{n=0}^{M_{root}} L_{root}(n) \frac{\Gamma\left(\frac{\kappa}{\lambda} + n - 1\right) \left(1 - \frac{\lambda}{\mu}\right)^{\frac{\kappa}{\lambda}} \left(\frac{\lambda}{\mu}\right)^n}{n! \left(1 - \frac{\lambda}{\mu} D_{root}\right)^{n + \frac{\kappa}{\lambda}}}$$

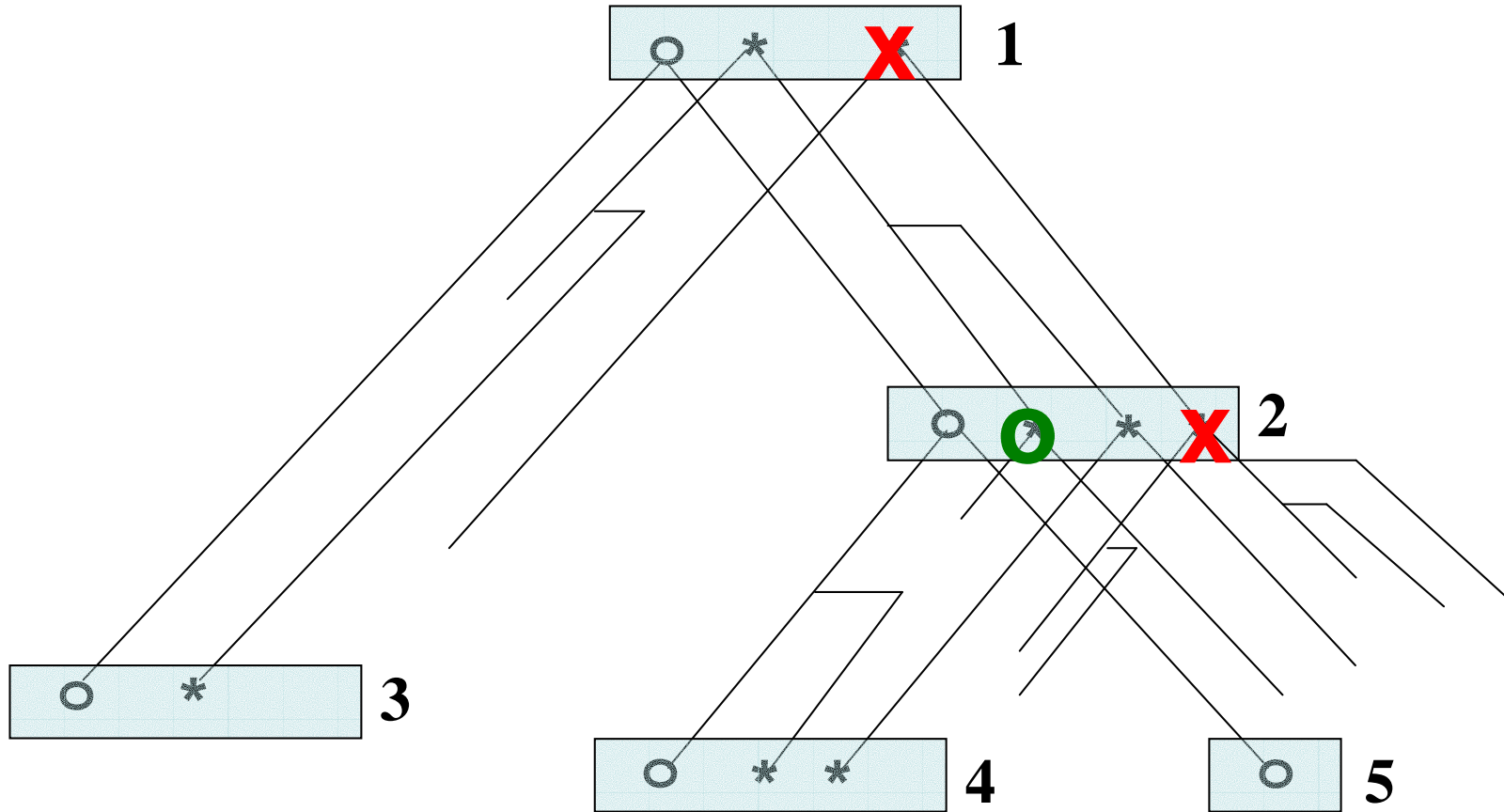
Example



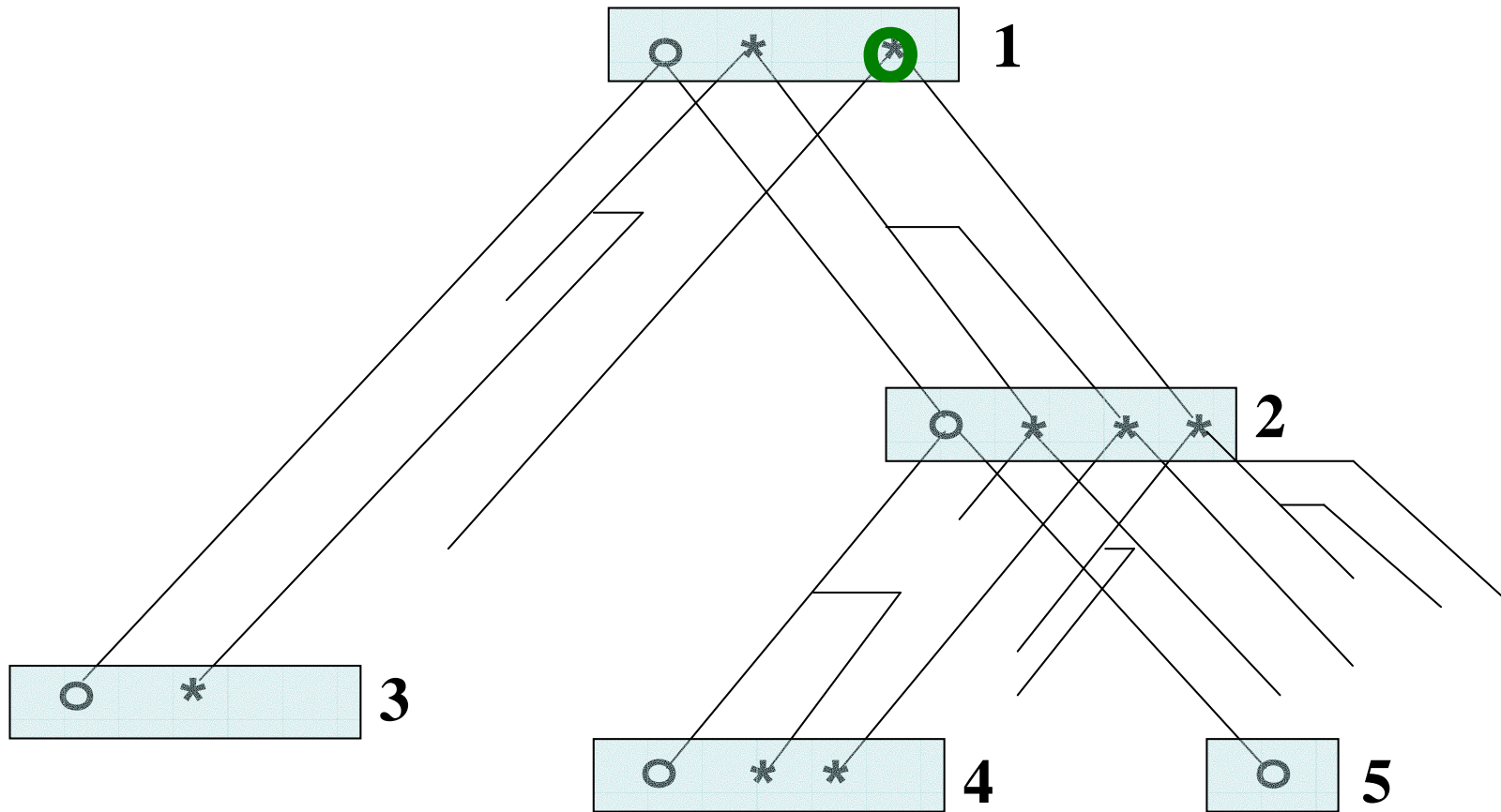
Example



Example



Example



Conclusions

- Gene content evolution by gene gain-loss-duplication
- Can handle arbitrary copy numbers
- Polynomial time algorithm, $O(NM^3)$ running time