Gain, loss, duplication model

István Miklós[1,2]

[1]Department of Statistics, University of Oxford

[2]Rényi Institute

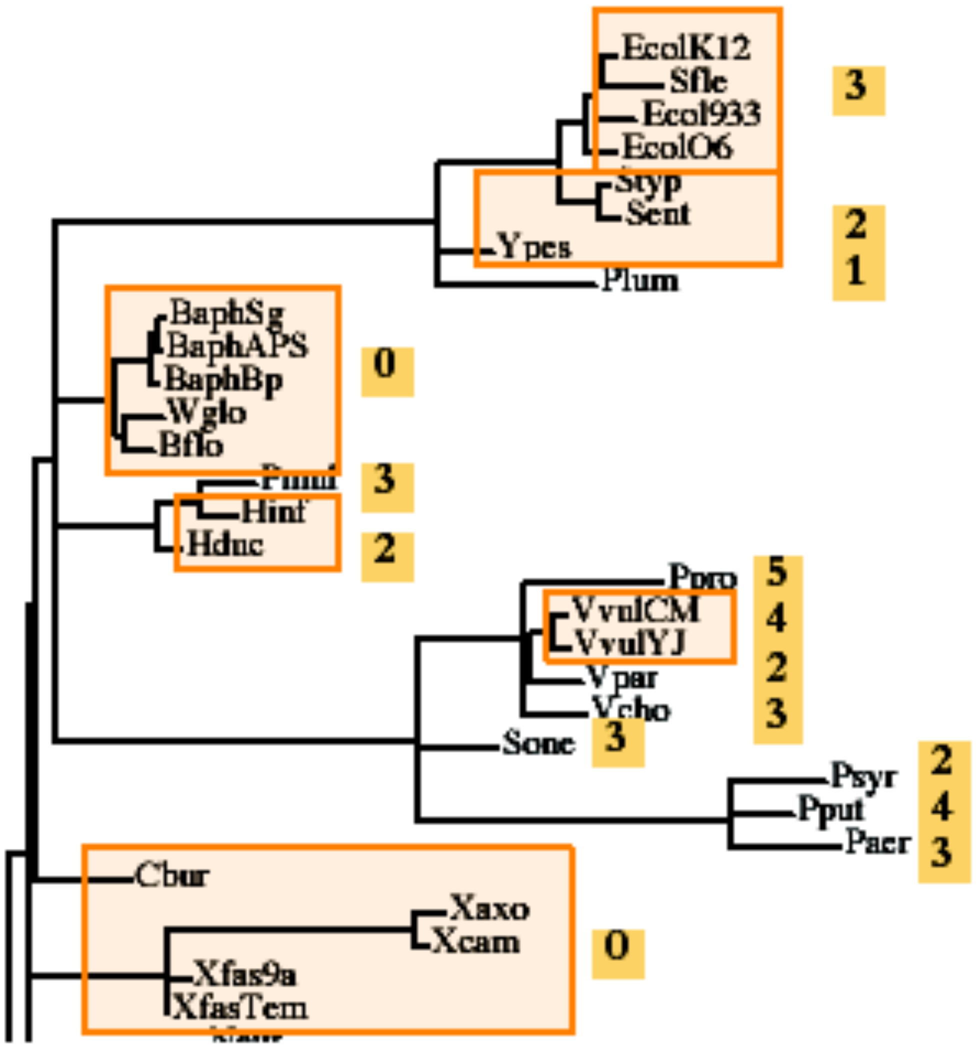
Bayesian Phylogeny Workshop
25-29 June, 2008, Budapest

Joint work with Miklós Csűrös

Why gene content evolution?

- compute **rates** of loss, duplication, and transfer
- complete **history** of a gene family
- ancestral gene content
- modes of adaptation (transfer or duplication+specialization)
- **phylogeny** reconstruction

Example: COG0247 (Kinesin like protein)

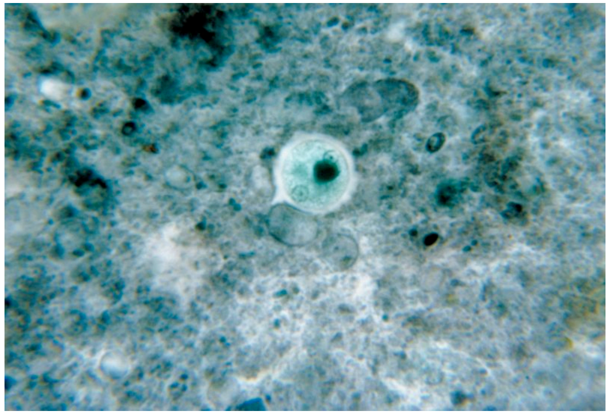


Mutations changing gene copy numbers

Deletions

“If you don’t need it – lose it!”

An extreme example: *Entamoeba histolytica*



- Loss of complete synthetic pathways, for example, purin - pyrimidine *de novo* synthesis
- No mitochondrion

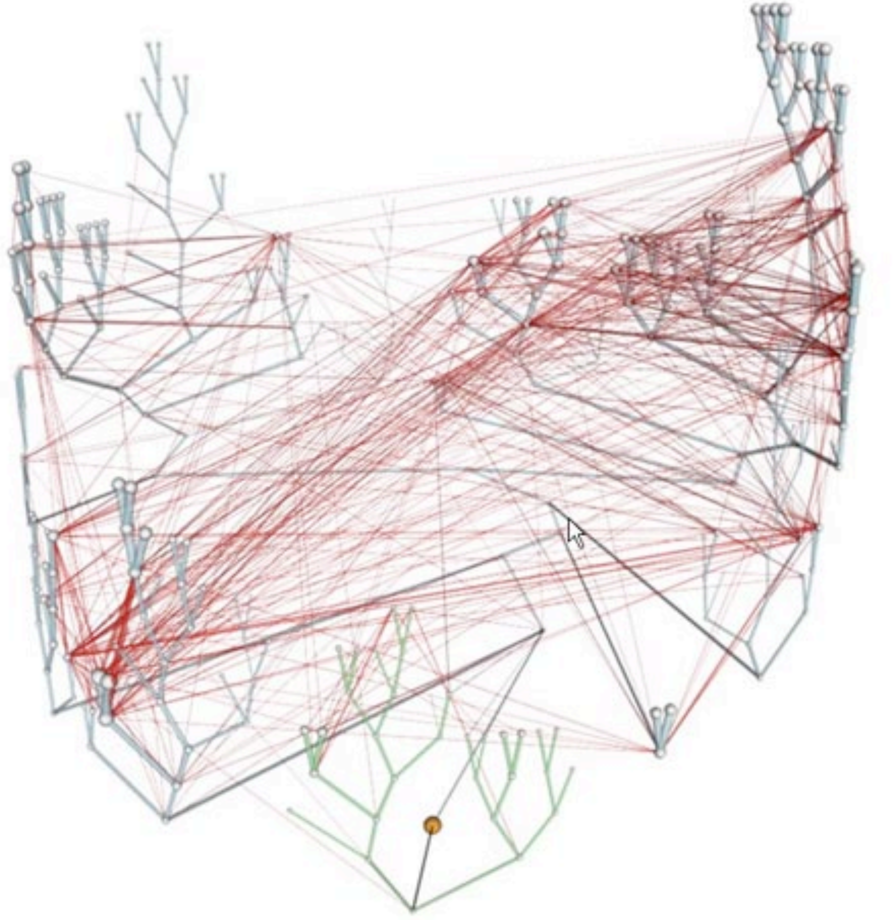
Mutations changing gene copy numbers

Horizontal gene transfer

"If you need it – borrow it!"

Horizontal gene transfer is common amongst Prokaryotes, evidences are accumulating that it also happen quite frequently amongst Eukaryotes.

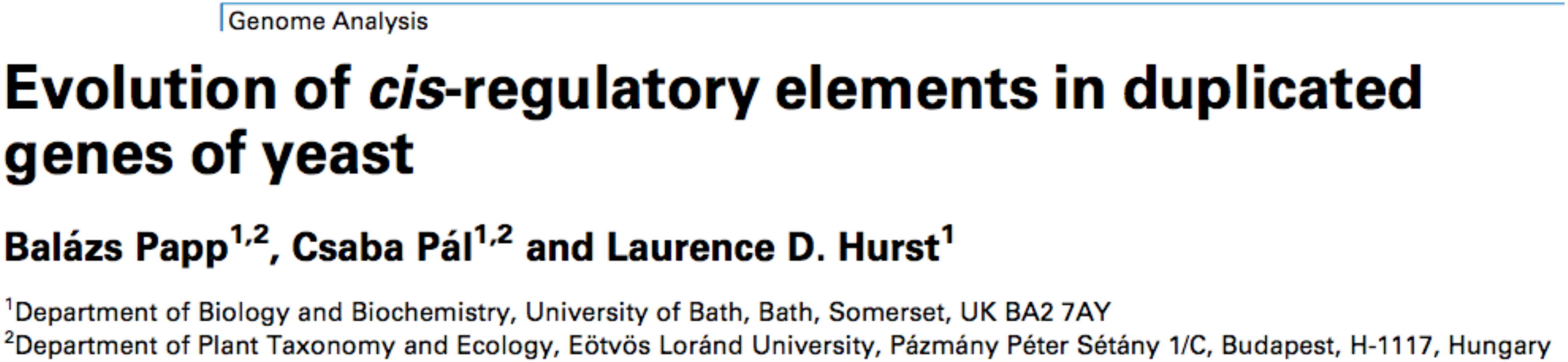
E. hystolitica also borrowed several genes from Human's Prokaryotic symbionts and pathogens.



A bird's-eye view of the tree of life, showing the vines in red and the tree's branches in grey [Bacteria] and green [Archaea]. The last universal common ancestor is shown as a yellow sphere.

Mutations changing gene copy numbers

Gene duplication



Genome Analysis

Evolution of *cis*-regulatory elements in duplicated genes of yeast

Balázs Papp[1,2], Csaba Pál[1,2] and Laurence D. Hurst[1]

[1]Department of Biology and Biochemistry, University of Bath, Bath, Somerset, UK BA2 7AY
[2]Department of Plant Taxonomy and Ecology, Eötvös Loránd University, Pázmány Péter Sétány 1/C, Budapest, H-1117, Hungary

Many duplicated genes in yeast could find a novel role by changing regulatory elements.

Available data and methods

Gene copy numbers

- Number of orthologous genes in an ortholog group.
- Do not care with sequence similarity
- The geneology of the genes is not predicted

Methods

- Maximum Likelihood tree
- Bayesian analysis

Conclusion: Fast likelihood calculation is needed.

Computational approaches

Presence-absence models (parsimony and ML)

- Ignores information on copy numbers

Finite state models (ML)

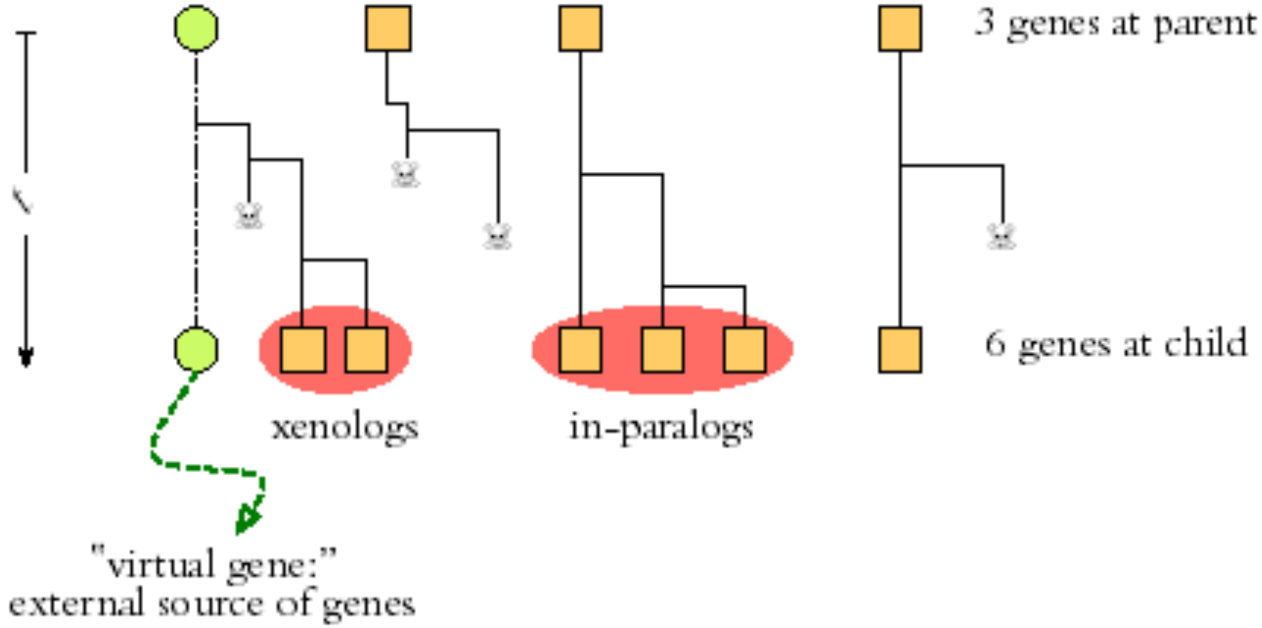
- Threshold on the copy number
- Gets obsolete if a new genome discovered with number of gene copies more than the threshold

Unlimited models

- Want: likelihood model for duplication, loss, and horizontal transfer, with exact and fast computations

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

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

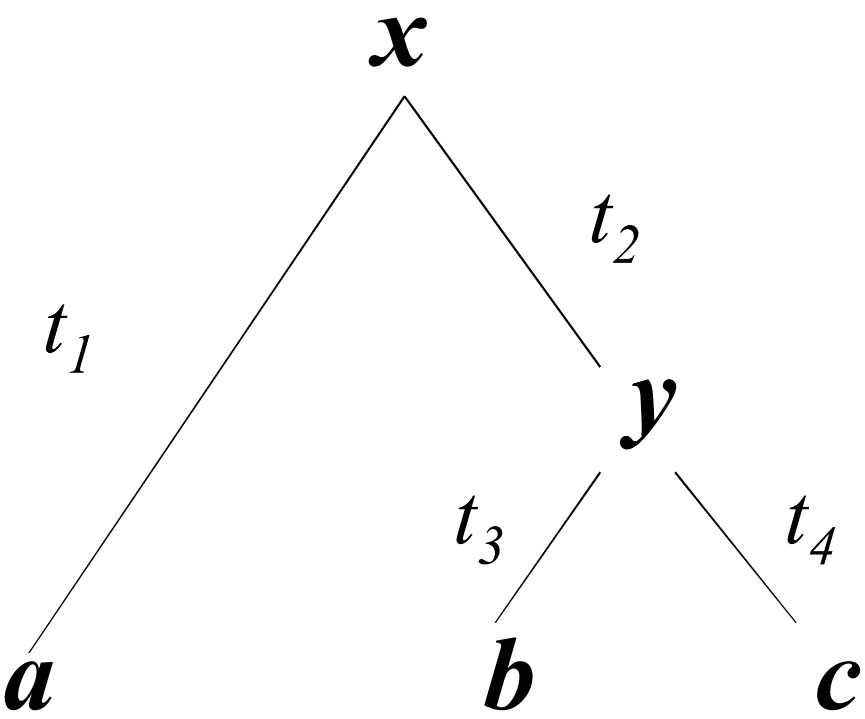
$$\frac{dp_1(t)}{dt} = -(\kappa + \lambda + \mu) p_1(t) + (\kappa + \lambda) p_0(t) + 2\mu p_2(t)$$

$$\vdots$$

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

Small problems II.

Calculate the likelihood of a tree



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

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 nth equation with ξn, and summing them

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

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

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

Solving with the method of Lagrange

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

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

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

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(ξ,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(ξ,t) gives the solutions for pn(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

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

Transition probabilities I.

$h_t(n)$: probability of observing n xenologs after evolutionary time t.

$$h_t(n) = \binom{\frac{\kappa}{\lambda} + n - 1}{n} \left(1 - \lambda\beta(t)\right)^{\frac{\kappa}{\lambda}} \left(\lambda\beta(t)\right)^{n-1}$$

where $\beta(t) = \dfrac{1 - e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}}$ and

$$\binom{\frac{\kappa}{\lambda} + n - 1}{n} = \begin{cases} 1 & \mathit{if}\ n = 0 \\ \dfrac{\left(\frac{\kappa}{\lambda}\right)\left(\frac{\kappa}{\lambda}+1\right)\ldots\left(\frac{\kappa}{\lambda}+n-1\right)}{n!} & \mathit{if}\ n > 0 \end{cases}$$

[This is a classical birth-death process with immigration (Karlin-McGregor, 1958)]

Transition probabilities II.

$g_t(n)$: probability of observing n in-paralogs after evolutionary time t (starting with one gene).

$$g_t(n) = \begin{cases} \mu\beta(t) & n = 0 \\ (1-\mu\beta(t))(1-\lambda\beta(t))\left(\lambda\beta(t)\right)^{n-1} & n > 0 \end{cases}$$

[Simple birth-death process, e.g. Feller, 1950]

Finite computation

Note: The Felsenstein's algorithm does not work for our model, since the number of characters (how many copies of a gene exist) is infinite.

Solution I: Approximate the infinite sum with a finite summation

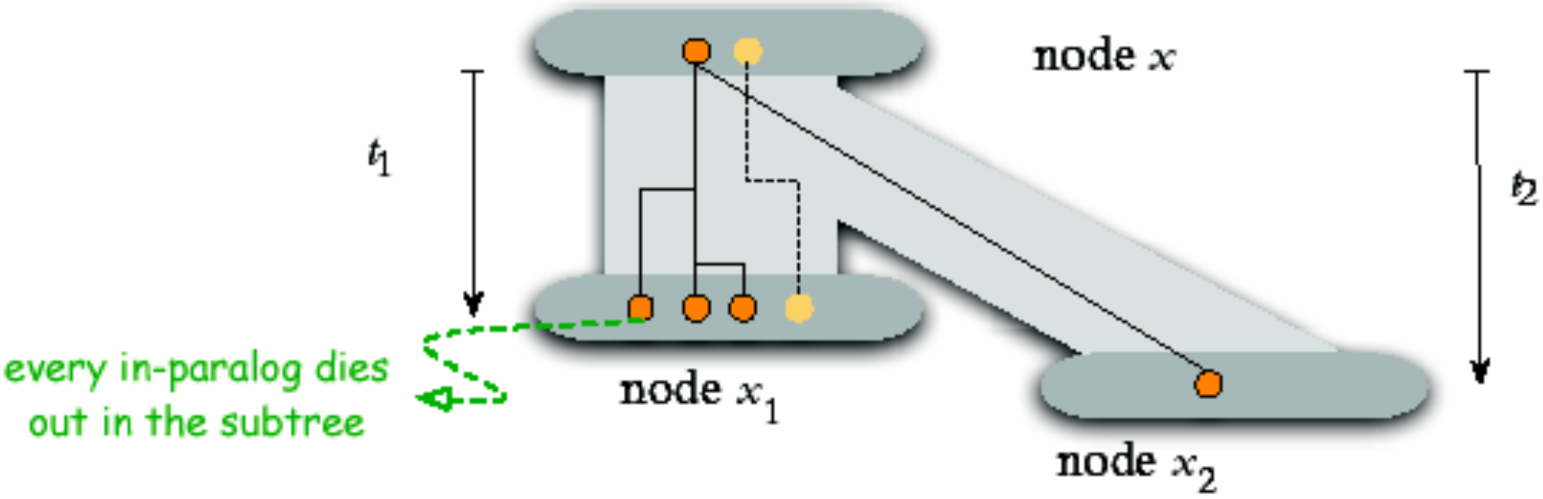
Solution II: Exact likelihood calculations.

Key idea: condition on the number of genes that have surviving modern offsprings.

Interestingly, it is not only an exact calculation, but in some cases is faster than the first solution...

Probability of extinction

What is the probability D_x that a particular gene at an intermediate node x has no modern descendant?



$$D_x = \prod_j \sum_{m=0}^{\infty} g_{t_j}(m) \left(D_{x_j}\right)^m$$

Now, plug in $g_t(m)$ and replace infinite sum by a closed formula.
→ dynamic programming to compute all D_x from leaves towards the root

Effective transition probabilities I.

Consider transition probabilities for in-paralog and xenolog blocks when only surviving genes are counted

Before: $h_t(n)/g_t(n)$: n xenologs/paralogs after time t
Now: $H_y(n)/G_y(n)$ xenologs/paralogs at node y which all have modern offsprings

By conditioning on the number of all xenologs/paralogs at node y (edge xy of length t):

$$H_y(n) = \sum_{i=0}^{\infty} \binom{n+i}{i} h_t(n+i) (D_y)^i (1-D_y)^n$$

$$G_y(n) = \sum_{i=0}^{\infty} \binom{n+i}{i} g_t(n+i) (D_y)^i (1-D_y)^n$$

Effective transition probabilities II.

Same trick: plug in $h_t(n)/g_t(n)$ and replace infinite sum by a closed formula:

$$H_y(n) = \binom{\frac{\kappa}{\lambda} + n - 1}{n} \left(\frac{1 - \lambda\beta(t)}{1 - D_y \lambda\beta(t)} \right)^{\frac{\kappa}{\lambda}} \left(\frac{(1 - D_y)\lambda\beta(t)}{1 - D_y \lambda\beta(t)} \right)^{n}$$

$$G_y(0) = 1 - \frac{(1 - \mu\beta(t))(1 - D_y)}{1 - D_y \lambda\beta(t)}$$

Geometric tail!!

$$G_y(n) = \frac{(1 - \mu\beta(t))(1 - \lambda\beta(t))}{\lambda\beta(t)\left(1 - D_y \lambda\beta(t)\right)} \left(\frac{(1 - D_y)\lambda\beta(t)}{1 - D_y \lambda\beta(t)} \right)^{n}$$

A complication...

We want to compute conditional likelihoods $L_x(n)$ probability of gene counts in the leaves rooted at node x, given that there are n genes at x that have modern offsprings

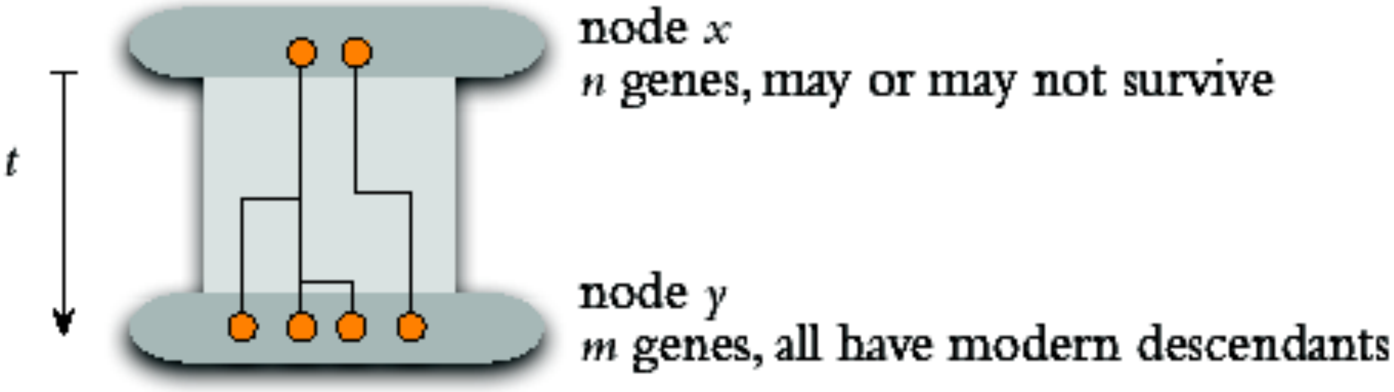
Unlike in the Felsenstein's algorithm, the fates of genes at different branches are not independent due to the condition!

It would lead to a cubic time algorithm (in the number of characters at leaves...), while the Felsenstein's algorithm's running time grows only quadratic with the alphabet size.

Can we do better? YES!!!

Survival probabilities

Probability $p_y(m|n)$


t
node x
n genes, may or may not survive
node y
m genes, all have modern descendants

$$p_y(m\,|\,0) = H_y(m) \qquad p_y(0\,|\,n) = H_y(0)G_y(0)^n$$

$$p_y(1\,|\,n) = G_y(0)\,p_y(1\,|\,n-1) + G_y(1)\,p_y(0\,|\,n-1)$$

$$\begin{aligned} p_y(m\,|\,n) = {} & G_y(0)\,p_y(m\,|\,n-1) + \\ & (G_y(1) - G_y(0))\,p_y(m-1\,|\,n-1) + \\ & \frac{\left(1 - D_y\right)\lambda\beta(t)}{1 - D_y\lambda\beta(t)}\,p_y(m-1\,|\,n) \end{aligned}$$

Conditional likelihoods I.

$L_x(n)$: likelihood of gene counts in the subtree rooted at x, given that there are n surviving genes

The easy case:

$$L_x(0) = \prod_j \sum_m^{M_j} p_{x_j}(m \mid 0) L_{x_j}(m)$$

M_j is the sum of gene copy numbers at the leaves of the subtree rooted at x_j.

Conditional likelihoods II.

Idea: consider $l_x(n)$, the likelihood of gene counts in the subtree rooted at x, given that there are n genes (may not survive)

Conditioning on the number of genes that survive at x:

$$l_x(n) = \sum_{i=0}^{n} \binom{n}{i} \left(D_y\right)^{n-i} \left(1 - D_y\right)^{i} L_x(i)$$

Conditioning on the number of surviving genes at the children x_j:

$$l_x(n) = \prod_j \sum_{m=0}^{M_j} p_{x_j}(m \mid n) L_{x_j}(m)$$

From the equality of the RHS and the base case of $L_x(0)$, we have the necessary recursions

Conditional likelihoods III.

From the equality of the RHS and the base case of $L_x(0)$, we have the necessary recursions

$$L_x(n) = \left(1 - D_y\right)^{-n} \left(\prod_j \sum_{m=0}^{M_j} p_{x_j}(m \mid n) L_{x_j}(m) - \sum_{i=0}^{n-1} \binom{n}{i} \left(D_y\right)^{n-i} \left(1 - D_y\right)^{i} L_x(i) \right)$$

For complete likelihood, combine $L_{root}(n)$ and equilibrium probabilities for surviving genes at root (or another distribution if it is more appropriate)

Algorithm

1. Compute sum of gene counts in each subtree
2. Compute extinction probability D_x for all nodes x
3. Compute $p_x(m|n)$ on all edges xy where $0 \leq m \leq M_x$ and $0 \leq n \leq M_y$
4. Compute $L_x(n)$ for all nodes x and $0 \leq n \leq M_x$
5. Compute weighted sum at root to get total likelihood

Running time: $O(M^2N)$ for a tree with N leaves, and sum of gene counts M

(In fact, it is $O(N+M^2h)$ where h is the height of the tree.)

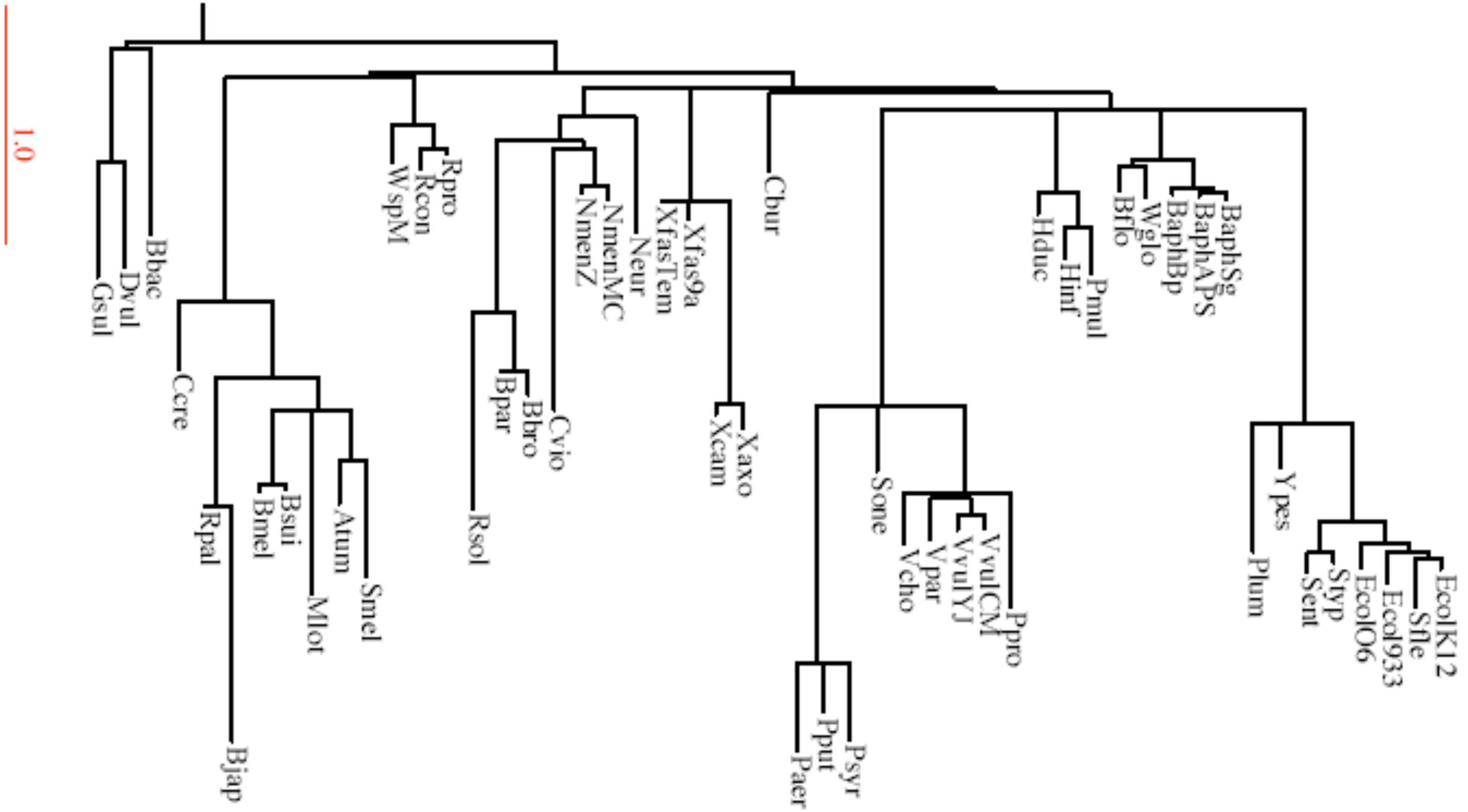
A remark...

The introduced algorithm is a dynamic programming algorithm with inclusion-exclusion

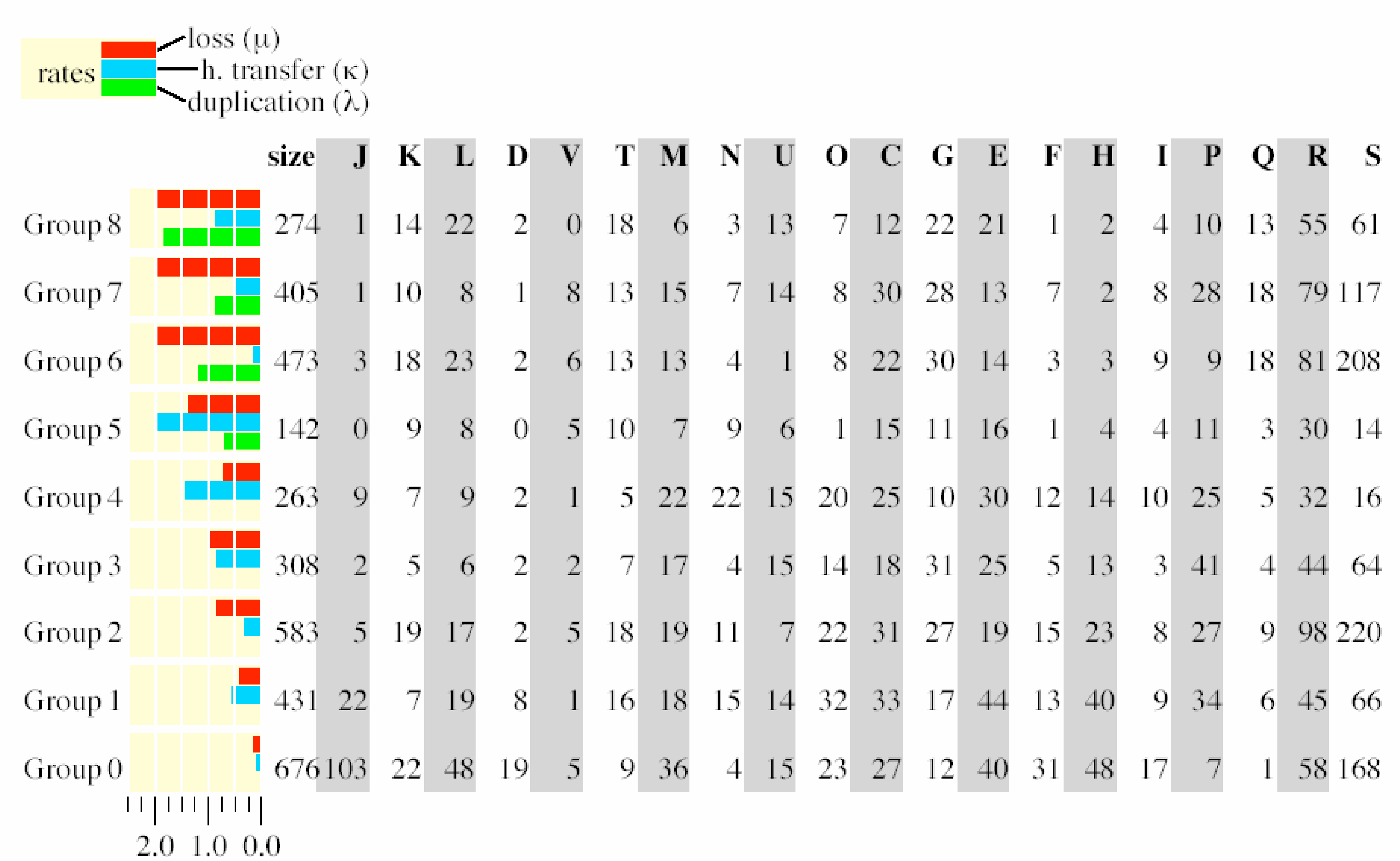
Another example is the one-state recursion by Lunter, Miklós, Song & Hein, an acceleration of the Forward algorithm when the HMM describes a birth-death model.

- Deeper understanding why it is possible
- Other examples?
- **Numerical instability?**

An example: proteobacteria + COGs



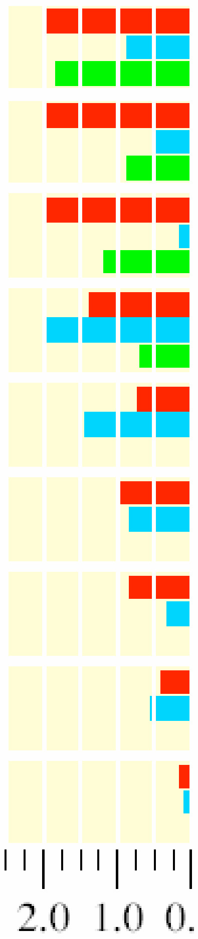
Clustering rates



	size	J	K	L	D	V	T	M	N	U	O	C	G	E	F	H	I	P	Q	R	S
Group 8	274	1	14	22	2	0	18	6	3	13	7	12	22	21	1	2	4	10	13	55	61
Group 7	405	1	10	8	1	8	13	15	7	14	8	30	28	13	7	2	8	28	18	79	117
Group 6	473	3	18	23	2	6	13	13	4	1	8	22	30	14	3	3	9	9	18	81	208
Group 5	142	0	9	8	0	5	10	7	9	6	1	15	11	16	1	4	4	11	3	30	14
Group 4	263	9	7	9	2	1	5	22	22	15	20	25	10	30	12	14	10	25	5	32	16
Group 3	308	2	5	6	2	2	7	17	4	15	14	18	31	25	5	13	3	41	4	44	64
Group 2	583	5	19	17	2	5	18	19	11	7	22	31	27	19	15	23	8	27	9	98	220
Group 1	431	22	7	19	8	1	16	18	15	14	32	33	17	44	13	40	9	34	6	45	66
Group 0	676	103	22	48	19	5	9	36	4	15	23	27	12	40	31	48	17	7	1	58	168

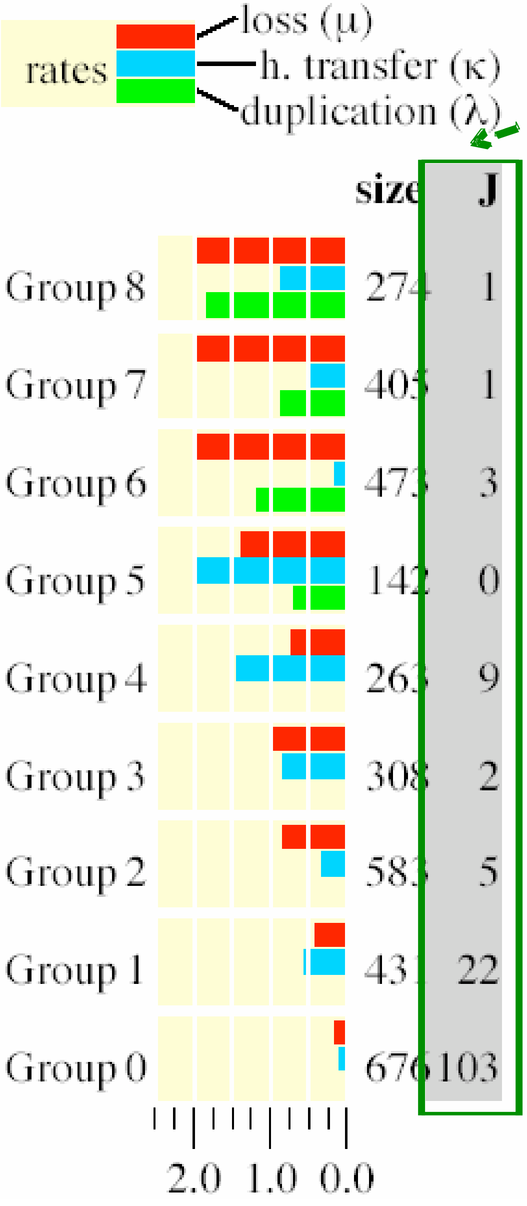
Metabolic functions and cell motility genes evolve by horizontal transfer

rates 🟥 loss (μ) 🟦 h. transfer (κ) 🟩 duplication (λ)



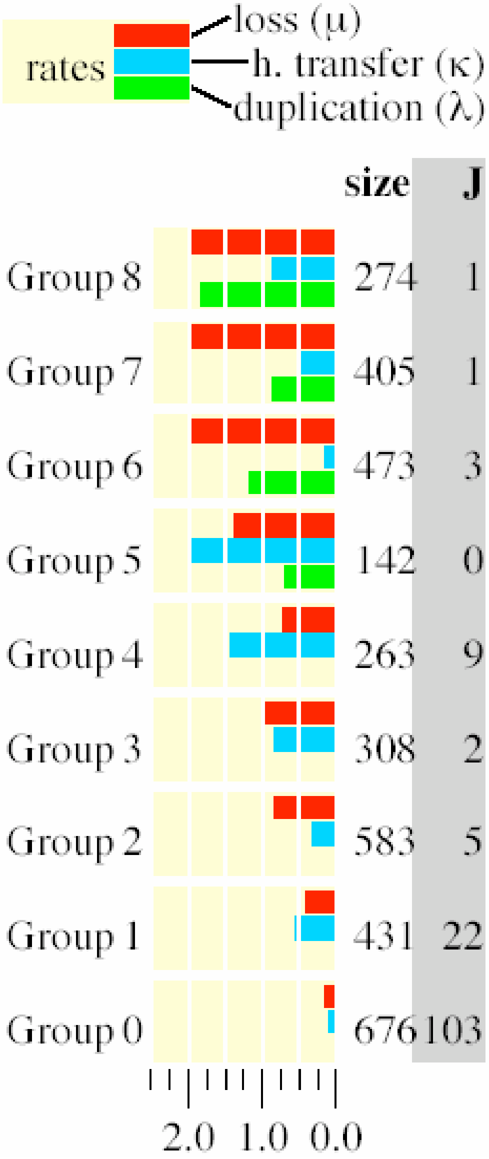
	size	J	K	L	D	V	T	M	N	U	O	C	G	E	F	H	I	P	Q	R	S
Group 8	274	1	14	22	2	0	18	6	3	13	7	12	22	21	1	2	4	10	13	55	61
Group 7	405	1	10	8	1	8	13	15	7	14	8	30	28	13	7	2	8	28	18	79	117
Group 6	473	3	18	23	2	6	13	13	4	1	8	22	30	14	3	3	9	9	18	81	208
Group 5	142	0	9	8	0	5	10	7	9	6	1	15	11	16	1	4	4	11	3	30	14
Group 4	263	9	7	9	2	1	5	22	22	15	20	25	10	30	12	14	10	25	5	32	16
Group 3	308	2	5	6	2	2	7	17	4	15	14	18	31	25	5	13	3	41	4	44	64
Group 2	583	5	19	17	2	5	18	19	11	7	22	31	27	19	15	23	8	27	9	98	220
Group 1	431	22	7	19	8	1	16	18	15	14	32	33	17	44	13	40	9	34	6	45	66
Group 0	676	103	22	48	19	5	9	36	4	15	23	27	12	40	31	48	17	7	1	58	168

Translation and cell cycle control are very stable



	size	J	K	L	D	V	T	M	N	U	O	C	G	E	F	H	I	P	Q	R	S
Group 8	274	1	14	22	2	0	18	6	3	13	7	12	22	21	1	2	4	10	13	55	61
Group 7	405	1	10	8	1	8	13	15	7	14	8	30	28	13	7	2	8	28	18	79	117
Group 6	473	3	18	23	2	6	13	13	4	1	8	22	30	14	3	3	9	9	18	81	208
Group 5	142	0	9	8	0	5	10	7	9	6	1	15	11	16	1	4	4	11	3	30	14
Group 4	263	9	7	9	2	1	5	22	22	15	20	25	10	30	12	14	10	25	5	32	16
Group 3	308	2	5	6	2	2	7	17	4	15	14	18	31	25	5	13	3	41	4	44	64
Group 2	583	5	19	17	2	5	18	19	11	7	22	31	27	19	15	23	8	27	9	98	220
Group 1	431	22	7	19	8	1	16	18	15	14	32	33	17	44	13	40	9	34	6	45	66
Group 0	676	103	22	48	19	5	9	36	4	15	23	27	12	40	31	48	17	7	1	58	168

Secondary metabolites biosynthe
evolve by gene duplication



	size	J	K	L	D	V	T	M	N	U	O	C	G	E	F	H	I	P	Q	R	S
Group 8	274	1	14	22	2	0	18	6	3	13	7	12	22	21	1	2	4	10	13	55	61
Group 7	405	1	10	8	1	8	13	15	7	14	8	30	28	13	7	2	8	28	18	79	117
Group 6	473	3	18	23	2	6	13	13	4	1	8	22	30	14	3	3	9	9	18	81	208
Group 5	142	0	9	8	0	5	10	7	9	6	1	15	11	16	1	4	4	11	3	30	14
Group 4	263	9	7	9	2	1	5	22	22	15	20	25	10	30	12	14	10	25	5	32	16
Group 3	308	2	5	6	2	2	7	17	4	15	14	18	31	25	5	13	3	41	4	44	64
Group 2	583	5	19	17	2	5	18	19	11	7	22	31	27	19	15	23	8	27	9	98	220
Group 1	431	22	7	19	8	1	16	18	15	14	32	33	17	44	13	40	9	34	6	45	66
Group 0	676	103	22	48	19	5	9	36	4	15	23	27	12	40	31	48	17	7	1	58	168

Conclusions

First exact likelihood calculation for the three parameter model

Future:

- Ancestral gene content
- Incorporate pathway information and sequence similarity

Thanks:
Hervé Philippe, Martin Lercher, Csaba Pál, Balázs Papp

Hungarian Ministry of Education (Békésy postdoctoral position)
Hungarian National Office for Research and Technology
(eScience Knowledge Centre)

Natural Sciences and Engineering Research Council of Canada
Fonds québecois de la recherche sur la nature et les technologies.