Gene trees: enumeration and min-max theorems

Éva Czabarka[1], Péter L. Erdős[2], Virginia Johnson[1], Anne Kupczok[3], Vincent Moulton[4], László A. Székely[1], Todd J. Vision[5]

1. University of South Carolina, USA 2. Rényi Institute of Mathematics, Hungary 3. Center for Integrative Bioinformatics, Austria 4. University of East Anglia, UK 5. University of North Carolina, USA

Phylogenetic and gene trees represent evolutionary relationships between copies of a gene, species or taxonomic units. The root of such a tree is the common ancestor, internal nodes represent speciation events or (in case of gene trees) duplication events. Thus, the number of children for each internal node is at least two, and the tree ideally (but not necessarily) is rooted and binary, The leaves are labeled with the name of the corresponding species/taxonomical unit. For phylogenetic trees, labels are unique, for gene trees, labels may repeat

É. Czabarka, P.L. Erdős, V. Johnson, V. Moulton obtained generating function identities and recursions to count rooted and unrooted gene trees with a given number of leaves using a given label set (where labels may be omitted). The tools for these formulas are the fact that removal of a root and rooting the resulting forest at the neighbors of the old root establishes a bijection between rooted phylogenetic trees and certain forests of such rooted phyloegentic trees, and a generalization of Otter's formula [4] to multi-leaf-labeled trees. Otter's formula states that in an unrooted tree the number of different trees minus the number of different non-symmetry edges is 1; this can be used to connect counts of rooted trees to counts of unrooted trees. The resulting recursions can be easily coded, the tables below were obtained by a program available at

http://www.math.sc.edu/~czabarka/programfiles/treecode.html

Rooted binary gene trees with n leaves and k labels where all labels are used at least once.

$n\backslash k$	1	2	3	4	5
1	1	0	0	0	0
2	1	1	0	0	0
3	1	4	3	0	0
4	2	14	27	15	0
5	3	48	180	240	105
6	6	171	1,089	2,604	2,625
7	11	614	6,333	24,180	42,075
8	23	2,270	36,309	207,732	554,820
9	46	8,518	207,255	1,710,108	6,578,550
10	98	32,576	1,184,829	13,739,550	73,169,250

Unrooted binary gene trees with n leaves and k labels where all labels are used at least once.

$n\backslash k$	1	2	3	4	5
1	1	0	0	0	0
2	1	1	0	0	0
3	1	2	1	0	0
4	1	4	6	3	0
5	1	10	30	36	15
6	2	27	140	310	300
7	2	74	663	2376	3990
8	4	226	3,186	17,304	44,850
9	6	710	15,642	123,508	462,735
10	11	2,354	78,441	874,998	4,550,955

Rooted non-binary gene trees with n leaves and at most k labels.

$n\backslash k$	1	2	3	4	5
1	1	2	3	4	5
2	1	3	6	10	15
3	2	10	28	60	110
4	5	40	156	430	965
5	12	170	948	3,396	9,376
6	33	785	6,206	28,818	97,775
7	90	3,770	42,504	256,172	1,068,450
8	261	18,805	301,548	2,357,138	12,081,605
9	766	96,180	2,195,100	22,253,672	140,160,650
10	2,312	502,381	16,307,598	214,370,398	1,658,936,806

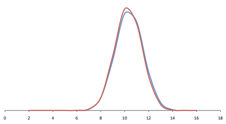
A combinatorial array $A(n,k)$ $(k=1,2,\ldots d_n)$ is an infinite sequence of finite sequences. Let $E_n=A(n,1)+\ldots+A(n,d_n)$, and define the random variable Z_n by $\mathrm{P}(Z_n=k)=A(n,k)/E_n$. The array $A(n,k)$ is asymptotically normal, it the cumulative density function of the standardized variable $(Z_n-\mathrm{E}(Z_n))/\sigma(Z_n)$ converges uniformly to the cumulative density function of the standard normal random variable (central limit condition). A stronger, local limit condition essentially states uniform convergence of the appropriate quantity to the probability density function of the standard normal random variable. Earlier results of P.L. Erdős and L.A. Székely [2] imply that $F^*(n,k)=S^*(n,\ n-k+1)$, where is the number of partitions of an n-element set into k classes, each of which has size two, and is the number of rooted phylogenetic trees with k leaves and n non-root vertices. Using this result, Harper's method for asymptotic normality[3], and Canfield's asymptotic results[1] on the Bell numbers, É. Czabarka, P.L. Erdős, V. Johnson, A. Kupczok, L.A. Székely showed that the array $F^*(n,k)$ is asymptotically normal and also satisfies the stronger local limit conditions. Also, they showed the same for the biologically more relevant distribution of $T_{n,k}$, the number of phylogenetic trees with n leaves and k internal vertices The graphs below illustrate the local limit conditions.

$S^*(n,k)$ expectation and variance

$$E\left(Z_n^*\right)=\frac{n}{r}-r-\frac{1}{2r}+\frac{1}{2r\left(r+1\right)^2}+O\left(\frac{1}{n}\right)$$

$$\sigma^2\left(Z_n^*\right)=\frac{n}{r\left(r-1\right)}-r+1+\frac{2}{r+1}-\frac{1}{2\left(r+1\right)^1}-\frac{1}{2\left(r+1\right)^3}+\frac{1}{\left(r+1\right)^4}+O\left(\frac{1}{n}\right)$$

$S^*(30,k)$ 31−k leaves, k internal vertices
Red curve the distribution from theorems
blue curve data normalized



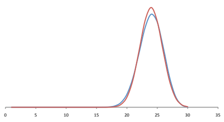
$T_{n,k}$ expectation and variance

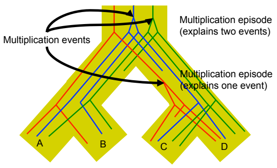
$$E\left(Z_n\right)=\frac{1-\rho}{2\rho}n+\frac{.75-\ln 2}{\rho}+O\left(\frac{1}{n}\right)$$

$$\sigma^2\left(Z_n\right)=\frac{n}{4}\left(\frac{1}{\rho^2}-\frac{2}{\rho}-1\right)+\frac{1-2\rho-2\rho^2}{8\rho^2}+O(1),$$

where $\rho=-1+2\ln(2)$

$T_{31,k}$: 31leaves, k internal vertices
Red curve distribution from theorems
blue curve data normalized




Multiplication events
Multiplication episode (explains two events)
Multiplication episode (explains one event)
A
B
C
D

The problem: given a species tree and gene trees that developed on it.
Biologists identified the nodes of the gene trees that are multiplication events, and.also identified intervals in which those events were likely to occur.
Place the minimum number of multiplication episodes on the species tree that explain these events (an episode can explain several events, but ancestral relations between events must be respected) and put them in the right interval.
Several versions of the problem exists with corresponding algorithms (Guigo et. al.[5], Page-Cotton[6], Burleigh-Bansal-Eulenstein-Vision[7]., etc)

Éva Czabarka, L.A. Székely, T.J. Vision developed a mathematical model for the problesm that allow for a generalization to more natural models (e.g. ones that give different ways to episodes of different types), a combinatorial notion of equivalent solutions hat potentially allows for counting the number of different solutions, and showed that the minimum number of episodes that explain all events is the maximum number of pairwise disjoint intervals in which we have to place the events (in the biologically relevant case, when intervals are open upwards). Similar minimax theorems exists for the case when the intervals are closed upwards.

1. Canfield*: Engel's inequality for Bell numbers*, JCTA **72** (1995) no 1. 184-187
2. P.L. Erdős and L.A. Székely: *Applications of antilexycographic order. I. An enumerative theory of trees*, Adv. in. Applied Math **10** (1989) no 4, 448-496
3. L.H. Harper: *Stirling behavior is asymptotically normal*, Ann. Math. Stat **38** (1967) 410-414
4. R. Otter: *The number of trees*, Ann. of Math. **49** (1948) 583-599
5. R. Guigo et. al: *Reconstruction of ancient molecular phylogeny*, Mol. Phylogenet. Evol**. 6** (1996) 189-203
6. R.D.M. Page, J.A. Cotton: *Vertebrea phylogenomics, reconciled trees and gene duplications*, Pacific symposium on Biocomputing (2002) 536-547
7. G. Burleigh, M.S. Bansal, O. Eulenstein, T.J. Vision: *Inferring species trees using genome duplication episodes*, Mol. Biol. Evol.