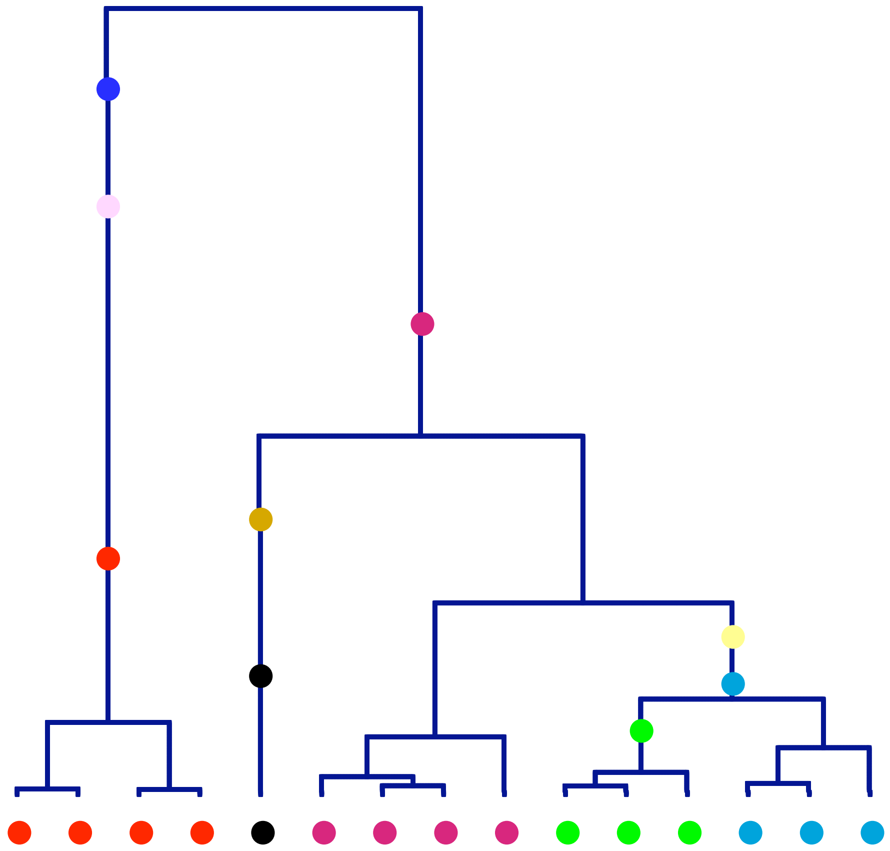
Ewens' sampling formula; a combinatorial derivation

Bob Griffiths, University of Oxford

Sabin Lessard, Université de Montréal

Infinitely-many-alleles-model: unique mutations



Sample configuration of alleles 4 A_1, 1 A_2, 4 A_3, 3 A_4, 3 A_5.

Ewens' sampling formula (1972)

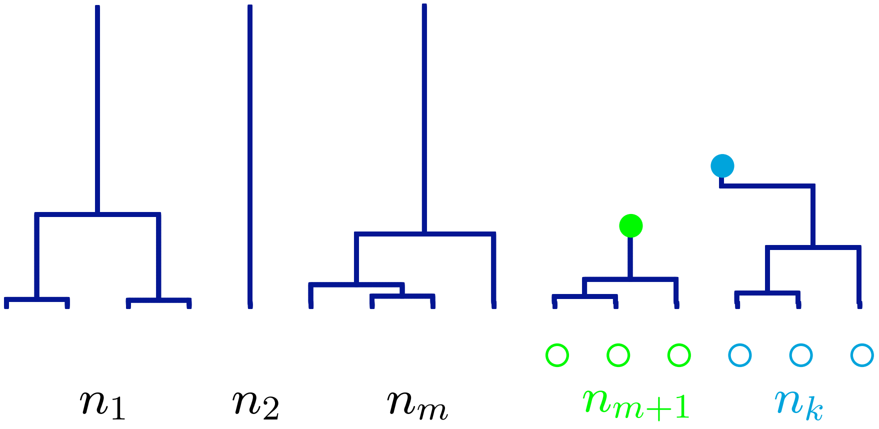
n sampled genes

Probability of a sample having k types with b_j types represented j times, $\sum jb_j = n$, and $\sum b_j = k$, is

$$\frac{n!}{1^{b_1}\cdots n^{b_n}} \cdot \frac{1}{b_1!\cdots b_n!} \cdot \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

Example: Sample 4 A_1, 1 A_2, 4 A_3, 3 A_4, 3 A_5.
$b_1 = 1$, $b_2 = 0$, $b_3 = 2$, $b_4 = 2$.

Old and New lineages



Old and new lineages, Watterson (1984)

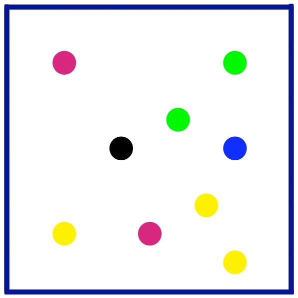
n sample genes traced to m ancestral genes.

Probability of having n_l genes of type l for $l = 1, \ldots, k$, types $1, \ldots, m$ ancestral and types $m + 1, \ldots, k$ mutant.

$$\frac{(n-m)!\theta^{k-m} \prod_{l=1}^{m} n_l! \prod_{l=m+1}^{k} (n_l - 1)!}{\prod_{i=m+1}^{n} i(\theta + i - 1)}$$

Kingman's (1982) partition formula when $\theta = 0$.

Hoppé’s (1987) urn model



1. Start with 1 black ball of mass θ in the urn.

2. Select a ball from the urn. If it is black return it with a ball of a new colour, if not add a ball of mass 1 of the same colour as the ball drawn.

3. Stop when n non-black balls and randomly label them $1, 2, \ldots, k$ if k different colours.

The Chinese restaurant process

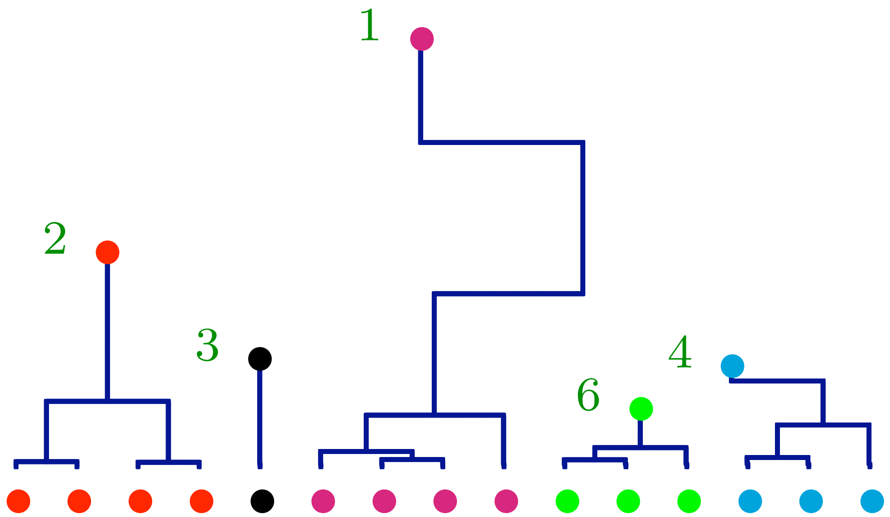
Imagine people $1, 2, \ldots, n$ arriving sequentially at an initially empty restaurant with a large number of tables.

Person j sits at the same table as person i (with probability $1/(j-1+\theta)$, for each $i < j$), or else sits at an empty table (with probability $\theta/(j-1+\theta)$). The distribution of the configuration of the number of people at the tables $n_1, n_2, \ldots$ is the Ewens' sampling formula

$$\frac{n!}{1^{b_1} \cdots n^{b_n}} \cdot \frac{1}{b_1! \cdots b_n!} \cdot \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

Random permutations, Joyce and Tavaré (1987)

In Hoppé's urn model label the balls according to the order that they enter the urn. If ball k's colour was determined by choosing ball j insert it in a cycle to the left of j.



(12 11 5 2)(3)(15 10 7 1)(14 9 6)(13 8 4)

Random permutations

If π is a permutation with k cycles

$$P_\theta(\pi_n = \pi) = \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

The number of permutations with b_1 cycles of length 1, b_2 cycles of length $2, \ldots, b_n$ cycles of length n is

$$\frac{n!}{\prod_{j=1}^n j^{b_j} b_j!}$$

Ewens sampling formula is

$$\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \cdot \frac{n!}{\prod_{j=1}^n j^{b_j} b_j!}$$

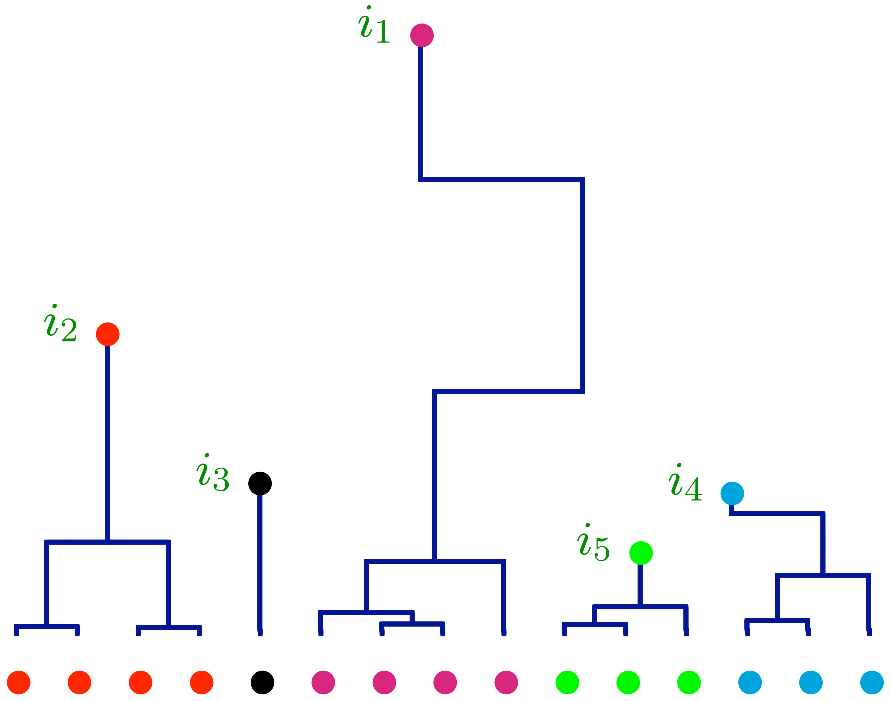
Birth Process with Immigration, Joyce and Tavaré (1987)

Immigrants enter the population according to a Poisson Process of rate θ, then reproduce according to a binary branching process.

If each immigrant is a new type and offspring are the same type as their parents, then the sequence of states the branching process with immigration moves through has the same distribution as those generated by Hoppé's urn.

End History

Forest of non-mutant ancestral lineages, to defining mutations



Ancestral lineages are lost back in time by coalescence or mutation at rates $\binom{i}{2}$ and $\frac{i\theta}{2}$ while i non-mutant lineages.

Ewens' sampling formula derivation: Griffiths and Lessard (2005)

$\frac{n!}{n_1!\cdots n_k!}$ assignments of k types

$\times\frac{1}{b_1!\cdots b_n!}$ if types are unlabelled

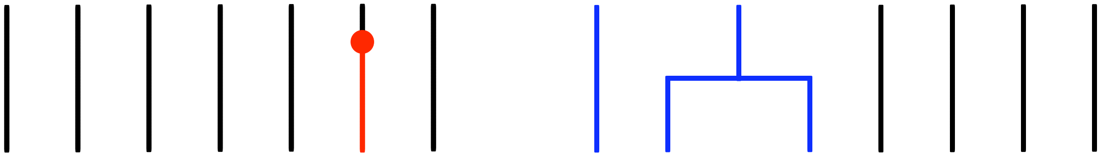
$\times n!$ arrangements of loss by mutation or coalescence

$\times\frac{\theta}{i(i+\theta-1)}$ if the i gene lost is the last of its type or $\frac{j-1}{i(i+\theta-1)}$ if it is the jth last of its type for $i = 1, \ldots, n$.

Probability of a sample having k types with b_j types represented j times is

$$\frac{n!}{1^{b_1}\cdots n^{b_n}}\cdot\frac{1}{b_1!\cdots b_n!}\cdot\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

Next event back in time

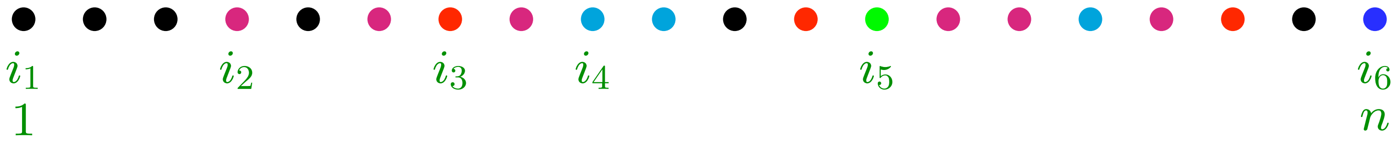


$\frac{1}{i} \cdot \frac{\theta}{\theta+i-1}$ $\frac{j-1}{i-1} \cdot \frac{i-1}{\theta+i-1}$

Probability of a mutation on a particular lineage when i ancestor lineages.

Probability of a coalescence in a group of j lineages when i ancestor lineages.

Combinatorial arrangement of age-ordered frequencies

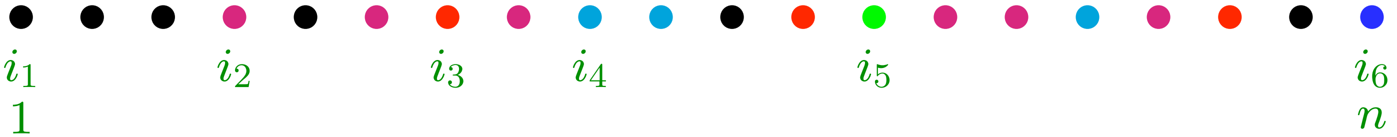


Allocate genes $n_1, \ldots, n_k$ to ordered events from 1 to n in the ancestral lines starting from the oldest type.

n_m genes are allocated in positions $\geq i_m$.

A particular labelling is possible if and only if for each $1 \leq m \leq n$ events in positions $i_1 = 1, \ldots, i_m - 1$ are labelled from $n_1, \ldots, n_{m-1}$.

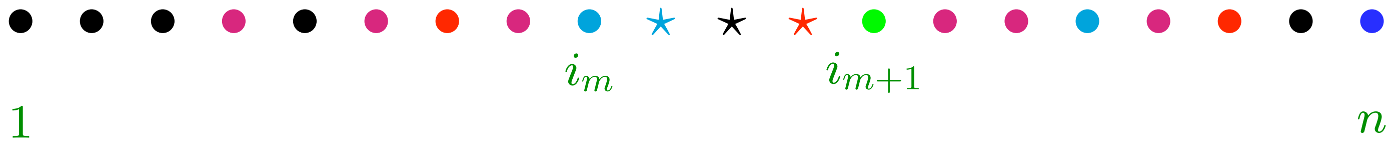
That is, if and only if $i_m - 1 \leq \sum_{\nu=1}^{m-1} n_\nu$.



$i_m - 1 \leq \sum_{\nu=1}^{m-1} n_\nu$ for $1 \leq m \leq n$

The probability of an arrangement is

$$a_{\mathbf{i}} = \frac{1}{n!} \cdot \prod_{m=1}^{k} n_m \cdot (\sum_{\nu=1}^{m} n_\nu - i_m)_{[i_{m+1}-i_m+1]}$$



Variable population size

$\lambda(t)$ is the relative population size at time t back to the present size.

Rate of coalescence at time t when i ancestor lines is $\binom{i}{2}\lambda(t)^{-1}$ and the rate of mutation is $\frac{i\theta}{2}$.

$T_n, T_{n-1}, \dots, T_1$ are times when ancestor lines are lost by mutation or coalescence.

Age-ordered Sampling formula

$$\frac{n! \cdot \theta^{k-1}}{\left(\prod_{l=1}^{k} n_l\right)} \sum_{\mathbf{i}} a_{\mathbf{i}} \mathbb{E}\left\{\frac{\prod_{l=2}^{k} \lambda(T_{i_l})}{\prod_{i=2}^{n}[\theta\lambda(T_i) + i - 1]}\right\}$$

Age-ordered sampling formula

$$\frac{n!\cdot\theta^{k-1}}{\left(\prod_{l=1}^{k} n_l\right)}\sum_{\mathbf{i}} a_{\mathbf{i}}\mathbb{E}\left\{\frac{\prod_{l=2}^{k}\lambda(T_{i_l})}{\prod_{i=2}^{n}[\theta\lambda(T_i)+i-1]}\right\}$$

Constant population size, Donnelly and Tavaré (1986), $\lambda(t) = 1$

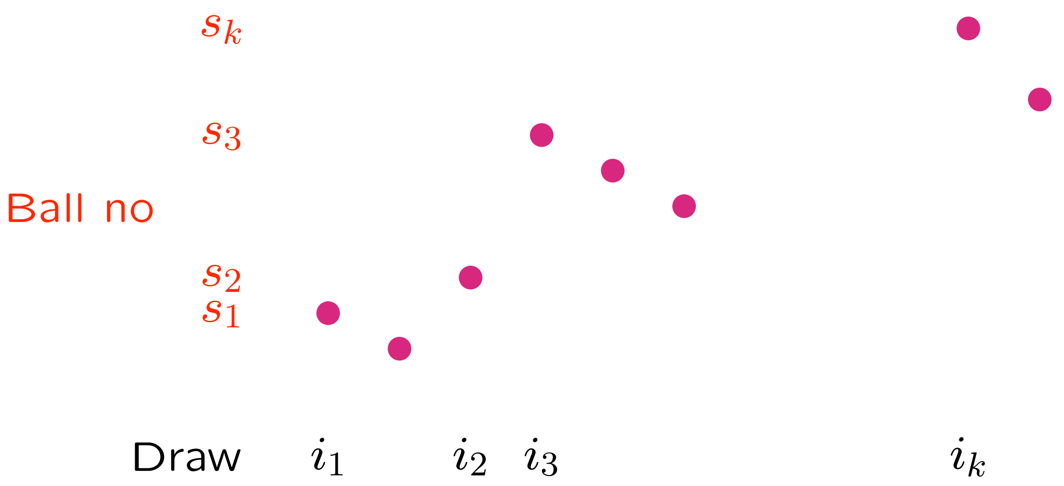
$$\frac{(n-1)!}{n_k\cdot(n_k+n_{k-1})\cdots(n_k+\cdots+n_2)}\cdot\frac{\theta^k}{\theta\cdots(\theta+n-1)}$$

$$\sum_{\mathbf{i}} a_{\mathbf{i}} = \frac{n_1}{n}\cdot\frac{n_2}{n_2+n_3+\cdots n_k}\cdots\frac{n_{k-1}}{n_{k-1}+n_k}$$

is the size-biassed probability of an age-ordering $1, 2, \ldots, n$

Record indices and record heights in an urn model

Balls labelled $1, 2, \ldots, n$ in an urn.



Age-ordered allele frequencies $n_1, n_2, \ldots, n_k$ given their ancestral lines are lost by mutation when $i_1, i_2, \ldots, i_k$ genes are distributed as the increments $s_1, s_2 - s_1, \ldots, s_n - s_{n-1}$ in the urn model given $i_1, i_2, \ldots, i_k$ (Griffiths and Lessard, 2004).

Age ordered sample frequencies

Random permutation (2,1,3,6,5,4,8,7).

Record Epoch	1	3	4	7
Record Value	2	3	6	8
Sample frequency	2	1	3	2

Age-ordered population frequencies $\{X_m; m \geq 1\}$

Partial sums $\{\sum_{\nu=1}^{m} X_\nu, m \geq 1\}$ given $i_1, i_2, \ldots$ are distributed as record values in a sequence of independent uniform random variables $\{U_l; l \geq 1\}$ given they occur at record epochs $i_1, i_2, \ldots$.

Random Partition

$$X_m = \xi_{m-1} \prod_{l=m}^{\infty} (1 - \xi_l),\ m \geq 1$$

where $\{\xi_l; l \geq 1\}$ are independent with $\xi_0 = 1$, and for $m \geq 1$, ξ_m has a density

$$(i_{m+1} - 1)(1 - z)^{i_{m+1}-2},\ 0 < z < 1$$

Random Partition $X_m = \xi_{m-1} \prod_{l=m}^{\infty}(1-\xi_l)$, $m \geq 1$ where $\{\xi_l; l \geq 1\}$ are independent with $\xi_0 = 1$, and for $m \geq 1$, ξ_m has a density

$$(i_{m+1}-1)(1-z)^{i_{m+1}-2},\ 0 < z < 1$$

Markov chain $\{i_j; j \geq 1\}$, where $i_1 = 1$ and

$$P(i_j = b \mid i_{j-1} = a) = \frac{a}{\theta + a} \cdots \frac{b-2}{\theta + b - 2} \cdot \frac{\theta}{\theta + b - 1},\ b > a.$$

GEM distribution Unconditional age-ordered distribution of population frequencies in a constant sized population model.

$$X_m = Z_m(1 - Z_{m-1}) \cdots (1 - Z_1),\ ,m \geq 1,$$

where $\{Z_j; j \geq 1\}$ are independent with density

$$\theta(1-z)^{\theta-1},\ 0 < z < 1$$

Pitman's two parameter Ewens sampling formula
The Chinese restaurant construction

Imagine people $1, 2, \ldots, n$ arriving sequentially at an initially empty restaurant with a large number of tables.

Before the $n + 1$th person arrives suppose there are k occupied tables.

Person $n + 1$ sits at the same table as person i with probability $(n_i - \alpha)/(n + \theta)$, for each $i < n + 1$, or else sits at an empty table with probability $(\theta + k\alpha)/(n + \theta)$.

The distribution of the configuration of the number of people at the tables $n_1, n_2, \ldots$ is the two-parameter Ewens' sampling formula

$$\frac{n!}{1^{b_1} \cdots n^{b_n}} \cdot \frac{1}{b_1! \cdots b_n!} \cdot \frac{(\theta + \alpha)_{k-1\uparrow\alpha} \prod_{i=1}^{k} (1-\alpha)_{n_i-1\uparrow 1}}{\theta(\theta+1)\cdots(\theta+n-1)}$$

where $(x)_{n\uparrow\alpha} = \prod_{i=0}^{n-1}(x + i\alpha)$. Usually $0 \leq \alpha \leq 1$.

Limit Frequencies in age-order

$X_1 = B_1, X_2 = (1-B_1)B_2, X_3 = (1-B_1)(1-B_2)B_3, \ldots$ where $\{B_i\}$ is an independent sequence and B_i has a Beta $(1-\alpha, \theta+i\alpha)$ distribution.

This Beta distribution form characterizes the product distribution form which is invariant under size-biassing.

Limit age-ordered frequencies given record indices

Let $\{i_j, j = 1, 2, \dots\}$ be the limit sequence of record indices and $X_1, X_2, \dots$ be the age-ordered limit frequencies. A representation is

$$X_j = \xi_{j-1} \prod_{m=j}^{\infty} (1 - \xi_m)$$

where $\{\xi_j\}$ are independent, $\xi_0 = 1$ and for $j > 1$, ξ_j is Beta $(1 - \alpha, i_{j+1} - j\alpha - 1)$.

$\{i_j, j = 1, 2, \dots\}$ is a Markov chain with

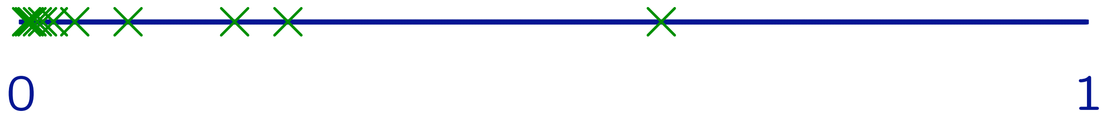
$$P_j(i_{j+1} \mid i_j) = (i_j - \alpha j)_{(i_{j+1} - i_j - 1)} \frac{\prod_{l=i_j+1}^{i_{j+1}} (\theta + \alpha(l-1))}{\theta + j}$$

Griffiths and Spanò (2007).

Poisson Dirichlet Process

$\{x_{(i)}\}$ is a point process on (0,1),

$$x_{(1)} > x_{(2)} > \cdots, \sum_{1}^{\infty} x_{(i)} = 1.$$



Relative frequencies in the Ewens sampling formula converge in distribution to the Poisson Dirichlet Process.

Kingman (1993), Poisson Processes; Arratia, Barbour and Tavaré (2003), Logarithmic combinatorial structures; Pitman (2006), Combinatorial stochastic processes

Definition Let $\{Y_i\}$ be a non-homogeneous Poisson process with mean measure density

$$\theta y^{-1} e^{-y},\ y > 0,\ (\theta > 0),$$

and $Y = \sum_{j=1}^{\infty} Y_j$. Then the Poisson Dirichlet point process is defined as

$$\left\{ X_{(i)} = \frac{Y_{(i)}}{Y} \right\}.$$

Y has a Gamma (θ) distribution and is independent of $\{X_{(i)}\}$.

Multidimensional frequency spectra h_k

$$P(\text{Points in } (x_1, x_1 + dx_1), \ldots, (x_k, x_k + dx_k))$$

$$= h_k(x_1, \ldots, x_k) dx_1 \cdots dx_k$$
$$= \theta^k (x_1 \ldots x_k)^{-1} (1 - \sum_1^k x_i)^{\theta - 1} dx_1 \cdots dx_k$$

for $x_1, \ldots, x_k > 0$, $\sum_1^k x_i < 1$.

Ewens sampling formula from the PD(θ) distribution

$$\int \frac{n!}{n_1! \cdots n_k!} x_1^{n_1} \ldots x_k^{n_k} h_k(x_1, \ldots, x_k) dx_1 \cdots dx_k$$
$$= \frac{n!}{n_1 \cdots n_k} \cdot \frac{\theta^k}{\theta(\theta+1) \cdots (\theta + n - 1)}$$

Size-Biassed Poisson Dirichlet

Let $\{Z_i\}$ be *iid* random variables with density

$$\theta(1-z)^{\theta-1},\ 0<z<1,$$

and

$$\begin{aligned} X_1 &= Z_1, \\ X_2 &= Z_2(1-Z_1), \\ X_3 &= Z_3(1-Z_2)(1-Z_1), \ldots \end{aligned}$$

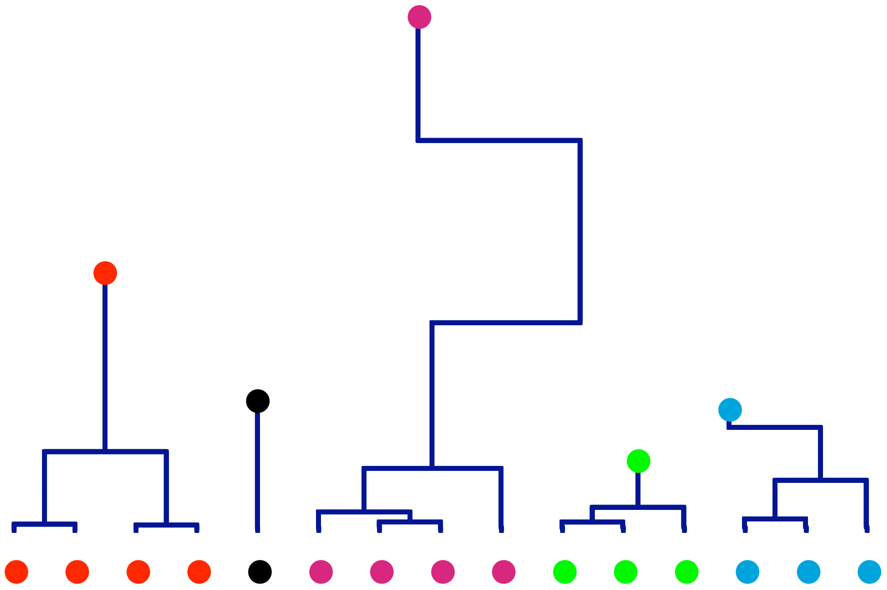
then $\{X_i\}$ is distributed as a Poisson Dirichlet process.

This is a size-biassed representation of the process.

An important result is that the size-biassed distribution is the age-ordered distribution of allele frequencies.

Coalescent lineage distributions

Forest of non-mutant ancestral lineages



$A_n^\theta(t)$ is the number of edges in the forest at time t back, with $A_n^\theta(0) = n$. It is possible that $n = \infty$.

Mutant family sizes in the Poisson Dirichlet process

In an infinite-leaf coalescent tree the joint distribution of family sizes from non-mutant lineages at time t back, given $A^{\theta}_{\infty}(t) = l$ is Dirichlet

$$\frac{\Gamma(\theta + l)}{\Gamma(\theta)}(1 - x_1 - \cdots - x_l)^{\theta-1},\ 0 < \sum_1^l x_j < 1$$

New mutant family sizes, scaled to have a total frequency 1, have an independent PD(θ) distribution. The total frequency of old mutations Y is Beta (l, θ), and total frequency of new mutations is $1 - Y$.

Poisson Dirichlet random measure

$$\mu = \sum_{i=1}^{\infty} x_i \delta_{\xi_i}$$

where $\{x_i\}$ is PD(θ) and independent of $\{\xi_j\}$ which are *i.i.d.* $\nu_0 \in \mathcal{P}(S)$, with S a compact metric space.

Stationary distribution of the random measure

$$\Pi_{\theta,\nu_0}(\cdot) = P(\mu \in \cdot)$$

Fleming-Viot process with type space S, and mutation operator

$$(Af)(x) = \frac{\theta}{2}\int_S (f(\xi) - f(x))\nu_0(d\xi)$$

Denote $\eta_n(y_1, \dots, y_n)$ as the empirical measure of points $y_1, \dots, y_n \in S$,

$$\eta_n(y_1, \dots, y_n) = n^{-1}(\delta_{y_1} + \cdots + \delta_{y_n})$$

The Fleming-Viot process with type space S and mutation operator A has transition function $P(t, \mu, d\nu)$ for given $\mu \in \mathcal{P}(S)$

$$\begin{aligned} P(t, \mu, .) &= q_0^\theta(t)\Pi_{\theta,\nu_0}(\cdot) \\ & \qquad + \sum_{n=1}^{\infty} q_n^\theta(t) \int_{S^n} \mu^n(dy_1 \times \cdots \times dy_n) \\ & \qquad\qquad \Pi_{n+\theta,(n+\theta)^{-1}\{n\eta_n(y_1,\dots,y_n)+\theta\nu_0\}}(\cdot) \end{aligned}$$

Ethier and Griffiths (1993). A review paper is Ethier and Kurtz (1993), Fleming-Viot Processes in Population Genetics.

Lineage distribution, sample of n genes

$$P(A_n^\theta(t) = j) = \sum_{k=j}^{n} \rho_k^\theta(t)(-1)^{k-j} \frac{(2k+\theta-1)(j+\theta)_{(k-1)} n_{[k]}}{j!(k-j)!(n+\theta)_{(k)}}$$

for $j = 0, 1, ..., n$, where $\rho_k^\theta(t) = e^{-k(k+\theta-1)t/2}$
and $a_{(j)} = a(a+1)\cdots(a+j-1)$, $b_{[j]} = b(b-1)\cdots(b-j+1)$

$\{A_n^\theta(t), \, t \geq 0\}$ is a death process with edges lost by coalescence or mutation at rate $\binom{j}{2} + j\frac{\theta}{2}$, $j = n, n-1, \ldots, 1$.

If $\theta = 0$ then $A_n^0(t)$ is the number of edges at time t back in the coalescent tree.

Lineage distribution, infinite-leaf coalescent tree

$$P(A_\infty^\theta(t) = j) = \sum_{k=j}^{\infty} \rho_k^\theta(t)(-1)^{k-j} \frac{(2k+\theta-1)(j+\theta)_{(k-1)}}{j!(k-j)!}$$

where

$$\rho_k^\theta(t) = e^{-k(k+\theta-1)t/2}$$

$\{A_\infty^\theta(t), t \geq 0\}$ is a death process with edges lost by coalescence or mutation at rate $j(j+\theta-1)/2$, $j = \ldots 5, 4, 3, 2, 1$.

Functional form of $\rho_k^\theta(t)$ suggests a connection with Brownian motion. Griffiths (2006).

Complex variable representations: X_t is $N(0,t)$ and $Z_t = e^{iX_t}$

The distribution of the number of non-mutant ancestor lineages in the population at time t back is

$$P(A_\infty^\theta(t) = j) = e^{\frac{1}{8}t}\frac{\Gamma(2j+\theta)}{\Gamma(j+\theta)j!}E\Big[\frac{(\rho Z_t)^j(1-\rho Z_t)}{\sqrt{Z_t}(1+\rho Z_t)^{2j+\theta}}\Big]$$

for $j = 0, 1, \dots$ where $Z_t = \exp(iX_t)$ and $\rho = e^{-\frac{1}{2}\theta t}$.

Time to the most recent common ancestor

The distribution of the time to the most recent common ancestor of the population T° is

$$P(T^\circ < t) = e^{\frac{1}{8}t} E\Big[\frac{(1-\beta Z_t)}{\sqrt{Z_t}(1+\beta Z_t)^2}\Big]$$

where $\beta = e^{-t}$.

The distribution of the time to the most recent common ancestor of a sample T_n° is

$$P(T_n^\circ < t) = e^{\frac{1}{8}t} E\Big[\sqrt{Z_t}(1-Z_t)(1-VZ_t)^{n-2}\Big]$$

where V is independent of Z_t with a Beta $(2, n-2)$ distribution.

Age of a mutation in the population

The distribution of the age of a mutation ξ_p, observed to be of frequency p in the current population is

$$P(\xi_p \leq t) = \frac{e^{\frac{t}{8}}}{2(1-p)} E\Big[\frac{(1 - Z_t^2)}{\sqrt{Z_t} R(Z_t, p)}\Big]$$

where

$$R(Z_t, p) = \big[(1 + Z_t)^2 - 4(1-p) Z_t\big]^{\frac{1}{2}}$$

References

Arratia, R., Barbour, A. D., Tavaré, S. (2003) Logarithmic combinatorial structures: a probabilistic approach. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich.

Donnelly, P., Kurtz, T. G. (1996) A countable representation of the Fleming-Viot measure-valued diffusion. *Ann. Appl. Prob.* 24, 698–742.

Donnelly, P.J., Kurtz, T.G. (1999) Particle representations for measure-valued population models. *Ann. Prob.* 24, (1999), 166–205.

Ethier, S.N., Griffiths, R.C. (1987) The infinitely many sites model as a measure valued diffusion. *Ann. Prob.* 15, 515-545.

Ethier, S.N., Griffiths, R. C. (1993) The transition function of a Fleming–Viot process. *Ann. Prob.* 21, 1571–1590.

Ethier, S.N., Kurtz, T.G. (1993) Fleming-Viot processes in population genetics. *SIAM J. Control Optimization* 31, 345–386.

Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Popul. Biol.* 3, 87–112.

Griffiths, R.C. (1980) Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoret. Popul. Biol.* 17, 37–50.

Griffiths, R.C. (1984) Asymptotic line-of-descent distributions. *J. Math. Biol.* 21, 67-75.

Griffiths, R.C. (2006) Coalescent lineage distributions. *Adv. Appl. Prob.* 38, 405–429.

Griffiths, R. C., Lessard, S. (2005) Ewens' sampling formula and related formulae: Combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theoret. Popul. Biol.* 68,167–177.

Griffiths, R.C., Spanò, D. (2007) Record indices and age-ordered frequencies in Exchangeable Gibbs Partitions. Electronic Journal of Probability, 1101-1130

Hoppé, F. (1984) Polya-like urns and the Ewens sampling formula. *J. Math. Biol.* 20, 91–99.

Joyce, P., Tavaré, S. (1987) Cycles, permutations and the structures of the Yule process with immigration, *Stochastic Proc. Appl.* 25, 309–314.

Kingman, J.F.C. (1982) The coalescent. *Stochastic Proc. Appl.* 13, 235–248.

Kingman, J.F.C. (1993) Poisson Processes. Oxford University Press.

Pitman, J. (2006) Combinatorial Stochastic Processes Lecture Notes in Mathematics 1875 Springer-Verlag Berlin Heidelberg

Tavaré, S. (1984) Line-of-descent and genealogical processes, and their application in population genetics models. *Theoret. Popul. Biol.* 26, 119–164.

Tavaré, S. Zeitouni. (2004) Lectures on Probability Statistics Lecture Notes in Mathematics 1837 Springer-Verlag Berlin Heidelberg