Mathematisches Forschungsinstitut Oberwolfach

Report No. 40/2005

Mathematical Population Genetics

Organised by
Ellen Baake (Bielefeld)
Warren Ewens (Philadelphia)
Anton Wakolbinger (Frankfurt)

August 21st – August 27th, 2005

Abstract. The meeting was devoted to mathematical aspects of population genetics, a branch of theoretical biology that is concerned with the genetic structure of populations under the influence of various evolutionary processes such as genetic drift, mutation, selection, recombination, and migration. The main focus was on probabilistic aspects of the dynamics, with the Moran model and its relatives as a central theme, and the corresponding stochastic processes forward and backward in time as a unifying point of view.

Mathematics Subject Classification (2000): 92D15, 60xx.

Introduction by the Organisers

One of the aims of population genetics is to explain how the genetic variation within and between biological populations is generated and maintained. Population genetics theory describes the change in the genetic composition of populations under the influence of various evolutionary processes such as genetic drift, mutation, selection, recombination, and migration. Elements of randomness turn out to be essential in the modelling of these processes. Random genetic drift, for example, is a consequence of the fact that, even without fitness differences, some individuals may, by chance, have more offspring than others, so that the offspring of one genotype may displace another.

The basic processes of evolution are known in principle, along with fundamental equations which describe the effects of interactions between genes. Indeed, of the biological sciences, genetics – and population genetics in particular – is the one with the most clearly defined mathematical models, with a strong emphasis on probability. This fact has increasingly attracted probabilists who have given the theory a new impetus through powerful modern methods. Thus there is now an increasingly large community of biologists and mathematicians speaking a common language, learning from each other, and identifying problems that should be attacked. This spirit was manifest in our meeting, with its small group of senior mathematicians and biologists, and its larger fraction of young researchers.

The topics presented during the workshop crystallized into the following thematic groups:

> Genetic hitchhiking and selective sweeps (Stephan, Schweinsberg, Etheridge);
> Mutation-selection systems (Evans, Stannat);
> Speciation and adaptation (Bürger, Champagnat);
> Spatially structured systems (Birkner, Swart, Hutzenthaler);
> Stochastic processes associated with the coalescent (Krone, Pfaffelhuber);
> Poisson-Dirichlet distributions (Spano, Feng);
> Sampling formulae and coalescent-based inference (Ewens, Griffiths, Wakeley, Möhle, Hudson);
> Recombination, gene-rearrangement and sequence alignment (M. Baake, Beresticky, Metzler)

Stimulated by these lectures, there were many discussions and a rich scientific exchange during the workshop, between generations as well as between disciplines. Last not least, the meeting also profited greatly from the friendly and serene Oberwolfach atmosphere.

Workshop: Mathematical Population Genetics

Table of Contents

Abstracts

Deterministic recombination dynamics in continuous time

Michael Baake
(joint work with Ellen Baake)

The deterministic limit of the stochastic process of recombination in population genetics, in continuous time, leads to an interesting nonlinear ODE system that has been studied for a long time. The first major advances to understand the classical system are due to H. Geiringer [5] in the 1940s (though her formulation was slightly different). A full characterization of solutions (in algorithmic terms, but not in explicit form) was later obtained by Lyubich, compare [7] and references given there, and [4] for general background material.

Motivated by this problem in genetics, the relevant subclass that emerges from single crossover events was considered in [1], where an explicit solution to the resulting large system of nonlinear differential equations was constructed. Soon after, this was reformulated as a measure-valued differential equation and solved for a more general class of state spaces [3]. In this context, the main focus was on the exact solution and the compatibility with other genetic processes, in particular with mutation and (additive) selection.

On the other hand, nonlinear semigroups are rarely known explicitly, and if so, this usually rests upon the transformability of the system to a linear one. It is thus rather natural to start with the most elementary semigroup constituents. Among them, one can then identify mutually commuting ones, which may be used to reconstruct the solution to the full recombination equation of [3] in an alternative and perhaps more transparent way.

Given the solution, a mode decoupling on the basis of the combinatorial Möbius inversion is possible that explicitly shows how the nonlinear semigroup is related to a (larger) linear one and furthermore gives systematic access to the so-called linkage disequilibria, compare [4], of population genetics.

The key to the understanding of the nonlinear semigroup involved is the following result, whose proof is a simple direct verification (for details, see [2]).

Theorem 1. *Let K be a closed convex subset of a Banach space B, and assume that $\mathcal{R}\colon K \longrightarrow K$ is a (nonlinear) Lipschitz map which satisfies*

$$\mathcal{R}\big(ax + (1-a)\mathcal{R}(x)\big) \;=\; \mathcal{R}(x) \tag{1}$$

for all $a \in [0,1]$ and all $x \in K$. Let $\varrho \geq 0$ be arbitrary.

Then, the (nonlinear) Cauchy problem

$$\dot{x} \;=\; \varrho\big(\mathcal{R} - \mathbf{1}\big)(x)\;, \quad x(0) = x_0 \in K\,, \tag{2}$$

has the unique solution $x(t) = e^{-\varrho t}x_0 + (1 - e^{-\varrho t})\mathcal{R}(x_0)$ for $t \geq 0$, and the entire forward orbit remains in K. In particular, with $\varphi_t := e^{-\varrho t}\mathbf{1} + (1 - e^{-\varrho t})\mathcal{R}$, $\{\varphi_t \mid t \geq 0\}$ is a (nonlinear) semigroup that preserves K. □

Defining $\nu_1(t) := \mathcal{R}(x(t)) \equiv \mathcal{R}(x_0)$ and $\nu_2(t) := \mathcal{R}(x_0) - x(t)$, one has the decomposition $x(t) = \nu_1(t) - \nu_2(t)$ with $\dot{\nu}_1 \equiv 0$ and $\dot{\nu}_2 = -\varrho\nu_2$. This shows how the situation of Theorem 1 relates to a *linear* semigroup that acts on the larger space $B \otimes B$, bound to special initial conditions. Although the condition in the theorem looks a bit special, an example of such a nonlinear operator $\mathcal{R}$ is given by the so-called recombinator in population genetics.

Let $N := \{0, 1, \ldots, n\}$ denote the set of sites or *nodes* (of a genetic sequence, say), and $L := \{\frac{1}{2}, \frac{3}{2}, \ldots, \frac{n-1}{2}\}$ the corresponding set of *links*. Here, a half-integer α always denotes the link between nodes $\lfloor\alpha\rfloor = \alpha - \frac{1}{2}$ and $\lceil\alpha\rceil = \alpha + \frac{1}{2}$. Let X_i be a locally compact space, attached to node i. The total state space of sequences is $X = X_0 \times X_1 \times \ldots \times X_n$. If π_i denotes the canonical projection to X_i, one defines

$$\pi_{<\alpha} : \; X \longrightarrow X_0 \times \ldots \times X_{\lfloor\alpha\rfloor} \quad \text{and} \quad \pi_{>\alpha} : \; X \longrightarrow X_{\lceil\alpha\rceil} \times \ldots \times X_n$$

in the obvious way. Finally, if $\mathcal{M}(X)$ is the Banach space of finite measures on X (with total variation as norm), one defines the pullback $\pi_i : \; \mathcal{M}(X) \longrightarrow \mathcal{M}(X_i)$ via $\big(\pi_i.\omega\big)(E) := \omega\big(\pi_i^{-1}(E)\big)$ for all Borel sets E of X_i.

In this setting, the *elementary recombinator* $R_\alpha : \; \mathcal{M}(X) \longrightarrow \mathcal{M}(X)$, defined by $R_\alpha(0) = 0$ and

$$R_\alpha(\omega) = \frac{1}{\|\omega\|}\big((\pi_{<\alpha}.\omega) \otimes (\pi_{>\alpha}.\omega)\big) \tag{3}$$

for $\omega \neq 0$, is an example of the operator $\mathcal{R}$ of Theorem 1, with $B = \mathcal{M}(X)$ and K the closed cone of positive measures, see [3, 2] for details.

The deterministic recombination dynamics, in continuous time, on (pairs of) sequences with nodes according to N and with single crossover events at the links of L, is described by the measure-valued nonlinear ODE

$$\dot{\omega} \;=\; \sum_{\alpha \in L} \varrho_\alpha \big(R_\alpha - \mathbf{1}\big)(\omega)\,, \tag{4}$$

where ϱ_α is the individual recombination rate at link α. Each single term on the right hand side gives rise to a nonlinear semigroup $\{\varphi_t^{(\alpha)} \mid t \geq 0\}$ of the type discussed above. As these semigroups mutually commute (which is a stronger condition than commutativity of the recombinators, due to nonlinearity), one can now proceed via a multiple application of Theorem 1.

The corresponding Cauchy problem with a positive measure ω_0 as initial condition has the solution

$$\omega_t \;=\; \sum_{G \subset L} a_G(t)\, R_G(\omega_0) \tag{5}$$

for $t \geq 0$, where the coefficient functions are given by

$$a_G(t) \;=\; \prod_{\alpha \in \overline{G}} \exp(-\varrho_\alpha t) \prod_{\beta \in G} \big(1 - \exp(-\varrho_\beta t)\big). \tag{6}$$

These coefficients can be seen as the result of expanding the product (over $\alpha \in L$) of the commuting semigroups $\{\varphi_t^{(\alpha)} \mid t \geq 0\}$. They admit a nice probabilistic interpretation: $a_G(t)$ is nothing but the probability that the set of links hit by crossover events until time t is precisely G.

The next step is the decoupling into *modes* by means of the general inclusion-exclusion principle. One starts from the observation that

$$R_G = \sum_{H \supset G} T_H \qquad \Longleftrightarrow \qquad T_G = \sum_{H \supset G} (-1)^{|H-G|} R_H ,$$

which simultaneously defines the new operators T_G. If one further defines the signed measures $T_G(\omega_t)$, where ω_t is the solution from (5), they satisfy the *linear* ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t} T_G(\omega_t) \; = \; - \Big(\sum_{\alpha \in \overline{G}} \varrho_\alpha \Big) \cdot T_G(\omega_t), \tag{7}$$

see [3] for a proof and further details on this kind of Möbius linearization. This is the analogue of the decomposition encountered after Theorem 1.

For applications, it is important that the time evolution in forward direction converges exponentially fast to the total product measure, $T_L(\omega_0) = R_L(\omega_0)$. In other words, all deviations from the equilibrium of mutual independence of sites decay exponentially, at characteristic rates. With the help of suitably chosen correlation functions, one can now derive a complete set of so-called linkage disequilibria, compare [4] for the concept, that are the quantities used in practice to determine recombination patterns experimentally.

It is rather obvious that one can extend Eq. (4) to include site-wise mutation without disturbing complete solvability [1, 3]. Some further analysis reveals [3] that this remains true even for a model with additive selection. Further directions should include non-additive selection as well as multiple crossovers. Also, the corresponding models in discrete time need a better understanding. In all cases, the precise relation between the deterministic limit considered above and the full stochastic picture needs to be studied, though first results [6] indicate that expectation values are well described. Finally, recombination with sequences of different length seems a challenging extension, see [8] for first steps in that direction.

References

[1] E. Baake, *Mutation and recombination with tight linkage*, J. Math. Biol. **42** (2001) 455–488.

[2] M. Baake, *Recombination semigroups on measure spaces*, Monatsh. Math. **146** (2005) 267–278; `math.CA/0506099`.

[3] M. Baake and E. Baake, *An exactly solved model for mutation, recombination and selection*, Can. J. Math. **55** (2003) 3–41; `math.CA/0210422`.

[4] R. Bürger, *The Mathematical Theory of Selection, Recombination, and Mutation*, Wiley, Chichester (2000).

[5] H. Geiringer, *On the probability theory of linkage in Mendelian heredity*, Ann. Math. Statist. **15** (1944) 25–57.

[6] I. Hildebrandt and E. Baake, *Stochastic versus deterministic recombination dynamics with single crossovers*, in preparation.

[7] Yu. I. Lyubich, *Mathematical Structures in Population Genetics*, Springer, Berlin (1992).

[8] O. Redner and M. Baake, *Unequal crossover dynamics in discrete and continuous time*, J. Math. Biol. **49** (2004) 201–226; `math.DS/0402351`.

Of mice and men (and random walks)

Nathanaël Berestycki
(joint work with Rick Durrett)

Traditionally, biologists have studied rates of evolution induced by certain types of mutations with parsimony methods. Hannehalli and Pevzner [7] developed a polynomial algorithm to deal with this problem in the case of random reversals. Our work is motivated by Bourque and Pevzner's simulation study [4] of the effectiveness of this parsimony method in studying genome rearrangement. With the help of numerical simulations Bourque and Pevzner [4] concluded that the parsimony distance was an accurate estimate only as long as the distance was at most $0.4n$, where n is the size of the analyzed sample. To have a cleaner mathematical problem, we consider the analogous problem of random transpositions, and obtain a surprising result about the random transposition random walk $(\sigma_t, t \geq 0)$ on the symmetric group of order n. Let D_t be the minimum number of transpositions needed to go back to the identity from the location at time t. We show that D_t undergoes a phase transition around the critical time $n/2$.

Theorem 1. *Let $c > 0$.*

$$\frac{1}{n} D_{cn/2} \to_p u(c) = 1 - \sum_{k=1}^{\infty} \frac{1}{c} \frac{k^{k-2}}{k!} (ce^{-c})^k$$

Moreover $u(c) = c/2$ for $c \leq 1$, $u(c) < c/2$ for $c > 1$ and there is no second derivative at $c = 1$.

In other words, the distance to the identity is roughly linear during the subcritical phase, and after critical time $n/2$ it becomes sublinear. This result may be used to return to the original problem of random reversals, therefore providing a theoretical explanation for the observation of Bourque and Pevzner (2002).

In addition, in the random transposition case, we describe the fluctuations of $D_{cn/2}$ about its mean in each of the three regimes: subcritical, critical and supercritical. (The results can be found in [1] and [3]). The techniques used involve viewing the cycles in the random permutation as a coagulation-fragmentation process and relating the behavior to the Erdős-Renyi random graph model.

To say a few words about the proof of this result, the Erdős-Renyi random graph is obtained when all $\binom{n}{2}$ possible edges on a graph with n vertices are declared open independently with probability p. Here we may couple our random walk σ_t with an Erdős-Renyi random graph by drawing an edge between i and j

whenever a transposition (i, j) is performed on the random walk. Then it is easy to check that at time $cn/2$, the resulting graph is a realization of the Erdős-Renyi random graph with parameter $p \sim c/n$. Moreover in this coupling, the cycles of the permutation are subsets of the connected components of the random graph. Our phase transition may be understood as coming from the well-known double jump phenomenon for the size of clusters at $c = 1$.

While Theorem 1 is a nice theoretical result, some aspects are overly simplistic with respect to the original motivation. For instance it is naive to assume that all transpositions (i.e. all reversals) are equally likely. In fact, we expect that reversals can only involve markers that no further apart than L markers. Since for general L this a hard problem, we first investigate the simplest possible case where $L = 1$. This is the case of random adjacent transpositions, studied by computer scientists and biologists in this context for a long time ([5], [6]). To rephrase our problem, consider σ_t the composition of adjacent transpositions (i.e., those transpositions of the form $(i, i + 1)$) where we multiply σ_{t-} by a randomly chosen adjacent transposition at rate 1. The distance $d_{\text{adj}}(\sigma)$ between σ and the identity is the minimum number of adjacent transpositions that is needed to build σ starting from the identity. As mentioned this problem is not new but so far only formulae for the expectation of the distance the random walk were known. Moreover these were generally quite involved and not easy to analyze.

Using a simple excursion representation for this process we are able to prove:

Theorem 2. *Let $t > 0$. Then as $n \to \infty$, $n^{-1}Ed_{\text{adj}}(\sigma_{nt}) \to f(t)$ for an explicit smooth function $f(t)$. Moreover its behavior at ∞ is diffusive:*

$$\lim_{t\to\infty} \frac{f(t)}{\sqrt{t}} = \frac{1}{2}E[\max_{0\le s\le 1} B_{2s}] = \sqrt{2}/2$$

where B_t is a standard Brownian motion.

This should be regarded as a "diffusive behavior" type of result. When we decide to choose a time-scale of order n^3t, we obtain the following result.

Theorem 3. *Let $t > 0$.*

$$\frac{1}{n^2} d_{\text{adj}}(X_{n^3t}) \to_p \Pr[B_1(t) > B_2(t)]$$

where B_1 and B_2 are two reflecting Brownian motions started uniformly on $0 \le B_1(0) < B_2(0) \le 1$ evolving independently.

The last two results contrast sharply with the behavior of random transpositions. Although there are different regimes, we never observe a clear-cut phase transition. Moreover the distance is a strictly sublinear function of time in both regimes. This raises the question of what are the random walks where we observe some phase transition for the distance to the origin. The last part of our results deals with the study of examples where such a phase transition does or does not occur.

Consider first the composition of p-cycles. That is, suppose that the step distribution is now uniform on cycles of lengths p, where $p \geq 3$ (so $p = 2$ is exactly the case of random transpositions). To avoid complications we explain our results for $p = 3$. As the next result shows shows, this random walk displays a phase transition very similar to that of random transpositions. Let D_t be the distance to the identity of σ_t, where σ_t denotes the random walk which is a product of a Poisson number of 3-cycles.

Theorem 4. *For $c > 0$*

$$\frac{D_{cn}}{n} \to_p u(c) = 1 - \sum_{s=0}^{\infty} \frac{(2s+1)^{s-2}}{s!}(3c)^s e^{-6c(s+1/2)}$$

Of course the function has similar characteristics as in the case of random transpositions for $c < 1/6$, $u(c) = c$, it has no-second derivative at $c = 1/6$ and $u(c) < c$ for $c > 1/6$.

We also study a random walk on a random regular graph, i.e. a graph which is uniform on all graphs where each vertex has degree 3. The result says that the random walk also displays an interesting phase transition.

Theorem 5. *For fixed $t > 0$*

$$\frac{d(X_{\lfloor t \log_2 n \rfloor})}{\log_2 n} \to_p f(t) := \min\left(\frac{1}{3}t, 1\right)$$

We end by proposing a challenging open problem. Let L be a number that may even depend on n, and consider the random walk obtained by composition random L-reversals. For which values of L does this random walk have a phase transition? In particular, since $L = 1$ gives the diffusive example of adjacent transpositions and $L = n$ is the case of random transpositions, is there a critical value of L for the existence of a phase transition?

References

[1] Berestycki, N. and Durrett, R. (2005) A phase transition inthe random transposition random walk. *Probab. Theor. Rel. Fields*, to appear.

[2] Berestycki, N. and Durrett, R. (2005) Limiting behavior for the distance of a random walk. In preparation.

[3] Berestycki, N. (2005) The hyperbolic geometry of random transpositions. *Ann. Probab.*, to appear.

[4] Bourque, G. and Pevzner, P. A. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research.* 12, 26–36

[5] Eriksen, N. (2005) Expected number of inversions after a sequence of random adjacent transpositions - an exact expression. *Discrete Mathematics*

[6] Eriksson, H., Erikkson, K. and Sjöstrand, J. (2000) Expected number of inversions after k random adjacent transpositions. In D. Krob, A.A. Mikhalev, A.V. Mikhalev, eds. *Proceedings of Formal Power Series and Algebraic Combinatorics*, Springer-Verlag (2000) 677-685

[7] Hannehalli, S. and Pevzner, P.A. (1995) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the* 27^{th} *Annual Symposium on the Theory of Computing*, 178–189. Full version in the *Journal of the ACM.* 46, 1–27

Survival in the face of competition

Matthias Birkner
(joint work with Andrej Depperschmidt)

We study a microscopic stochastic model for the time evolution of a 'population', say of animals or plants, which live, move – in the case of plants, we think rather of the dispersal of seeds – and reproduce in space, subject to random fluctuations. Individuals live on the discrete space $\mathbb{Z}^d$ (we mostly think of $d = 2$) in non-overlapping generations. In the absence of competition, an individual would have on average $m > 1$ offspring. Due to competition for local resources, the average reproductive success of an individual at position x is reduced by an amount of $\lambda_{xy} \geq 0$ by each individual at position y. Here λ_{xy} is a positive finite range kernel on $\mathbb{Z}^d$ which depends only on the displacement $y-x$, and decays as $|y-x|$ increases. We assume that at least $\lambda_{xx} > 0$. Thus, an individual at x in generation n will have a random number of offspring with mean given by

$$\Big(m - \sum_y \lambda_{xy}(\xi_n(y) - \delta_{xy})\Big)^+,$$

where $\xi_n(y)$ denotes the number of particles at spatial position y in generation n. We take the positive part because it is impossible to have a negative number of children. For definiteness and simplicity, we assume that the actual number of offspring, given the present configuration, is Poisson-distributed with the above mean, and independent for different individuals. Once created, offspring take an independent random walk step with kernel p from the location of their mother, where p is a symmetric finite range kernel. In this way, our model incorporates individual-based random fluctuations in the number and spatial dispersal of offspring. We note that the regulation by competition is essential for the possibility of long-time stability of such systems: the situation $\lambda_{xy} \equiv 0$ corresponds to non-interacting branching random walks, and it is well known that then there exist no non-trivial equilibria in the biologically most interesting case $d = 2$.

Similar models have been studied in the ecology literature, using simulations and non-rigorous methods, see [2], [5]. Continuous mass versions (in the spirit of 'superprocesses') of these models have been investigated in [3] and [1], and the possibility of long-term survival in $d = 2$ for certain parameters has been proved there. Mathematical aspects of the spatially continuous model considered in [2], in particular a high-density rescaling which leads to a related deterministic integro-differential system, have been studied in [4].

We prove that for $m \in (1,4)$, if the coefficients of competition λ_{xy} are sufficiently small, the population, starting from any initial condition which has $\xi_1 \not\equiv 0$ with positive probability, will survive for all times with positive probability, locally as well as globally: it will spread out into the whole of $\mathbb{Z}^d$, and for any given site, the asymptotic fraction of times when one observes particles there is then strictly positive.

Our proof uses a suitable coarse graining and then comparison with finite-range oriented percolation. The restriction on m comes from the corresponding deterministic system with which we compare. We have at the moment no clear picture of the behaviour of our system when $m > 4$.

Furthermore we strongly suspect (and in fact outline a possible route to a proof via coupling) that when $m \in (1,3)$ and the λ_{xy} are sufficiently small, it is possible to couple versions of the system starting from initial conditions ξ_0, $\tilde{\xi}_0$ in such a way that on the event that both populations survive, we have

$$\cup_{m\in\mathbb{Z}_+} \cap_{n\geq m} \{\forall\, x \in B \,:\, \xi_n(x) = \tilde{\xi}_n(x)\}$$

almost surely for all finite $B \subset \mathbb{Z}^d$. This would imply in particular that in this parameter range, there is a unique non-trivial equilibrium μ for the process, and the distribution of ξ_n, given that $\xi_n \not\equiv 0$, converges to μ for any initial condition.

References

[1] J. Blath, J., A.M. Etheridge, M.E. Meredith, *Coexistence in locally regulated competing populations*, preprint (2005).

[2] B.M. Bolker, S.W. Pacala, *Using Moment Equations to Understand Stochastically Driven Spatial Pattern Formation in Ecological Systems*, Theoretical Population Biology **52**, no. 3 (1997).

[3] A.M. Etheridge, *Survival and extinction in a locally regulated population*, Ann. Appl. Probab. **14**, no. 1, (2004), 188–214.

[4] N. Fournier, S. Méléard, *A microscopic probabilistic description of a locally regulated population and macroscopic approximations*, Ann. Appl. Probab. **14**, no. 4, (2004), 1880–1919.

[5] R. Law, R., U. Dieckmann, *Moment Approximations of Individual-based models.* In *The Geometry of Ecological Interactions* (U. Dieckmann, R. Law and J.A.J. Metz, eds.) 252–270. Cambridge Univ. Press, 2000.

Intraspecific competition and sympatric speciation

Reinhard Bürger

(joint work with Kristan Schneider)

Ecologically driven sympatric speciation has received much attention recently. A multilocus model of a quantitative trait is treated, in which the trait is under frequency-dependent selection and acts as mating character for assortment. The purpose of the analysis is the identification of conditions that lead to competitive divergence and the establishment of reproductively isolated clusters in the population. This may be interpreted as evolutionary splitting or sympatric speciation. In our model, there are parameters that independently determine the strength of assortment, the costs for being choosy, and the strength of frequency-dependent natural selection. The latter results from intraspecific competition for a continuous resource spectrum. Sufficiently strong frequency dependence leads to disruptive selection on the phenotypes. Two modes of assortative mating are analyzed in detail, one without costs for being choosy, the other with high costs. The population is assumed to consist of sexual haploid or diploid individuals. The diallelic loci contribute additively to the trait. If frequency dependence is strong enough to induce disruptive selection and costs are absent or weak, the result of evolution depends in a distinctive nonlinear way on the strength of assortment: less genetic variation is maintained under moderately strong assortment than under weak or very strong assortment, sometimes none at all. Competitive divergence and evolutionary splitting can occur if frequency dependence and assortment are both strong enough. Even then, the evolutionary outcome depends on the genetics and the initial conditions; populations with little initial variation are likely to evolve to a monomorphic state (because mutation is ignored in the model). The roles of the number of loci, of linkage, and of asymmetric selection are explored. If assortative mating is very costly, competitive divergence never occurs. Instead, unless assortment is weak, populations convergence to one of the monomorphic states. In the absence of assortative mating, i.e., with random mating, the equilibrium and stability structure can be determined completely by assuming linkage equilibrium because a global Lyapunov function is found.

References

[1] R. Bürger, *A multilocus analysis of intraspecific competition and stabilizing selection on a quantitative trait*, J. Math. Biology **50** (2005), 355–396.

[2] R. Bürger and K. Schneider, *Intraspecific competitive divergence and convergence under assortative mating*, Amer. Natur., to appear.

A microscopic interpretation for a Markov jump model of evolution in adaptive dynamics

Nicolas Champagnat

(joint work with Régis Ferreière, Sylvie Méléard)

We study the links between two models of Darwinian evolution in an asexual population. The first one is a "microscopic" model, describing all individual's births and deaths in a finite population, natural in various biological settings, including plants populations dispersing in a spatial environment [1, 7] and asexual populations undergoing natural selection [9]. The second one, called "trait substitution sequence" (TSS), describes the dynamics of the population's dominant phenotype as a markov jump process in the phenotype space, and belongs to the recent biological theory of evolution called “adaptive dynamics” [9, 5], which has revealed an important tool to understand various biological phenomena, including evolutionary branching and speciation [4]. We propose to give a firm mathematical basis to this model by recovering it from the microscopic model under proper scaling.

1. The microscopic model

We consider a finite population in which each individual is characterized by a quantitative phenotypic trait (or simply *trait*) belonging to a closed subset $\mathcal{X}$ of $\mathbb{R}^d$. If at some time $t \geq 0$, the population is composed of $N(t)$ individuals with traits $x_1(t), \dots, x_{N(t)}(t)$ in $\mathcal{X}$, the state of the population is represented by the counting measure

$$\nu_t = \sum_{i=1}^{N(t)} \delta_{x_i(t)}.$$

In such a population, an individual with trait x may

- give birth to a new individual with rate $b(x)$,
- dye with rate $d(x) + \sum_{i=1}^{N(t)} \alpha(x, x_i(t)) = d(x) + \int_{\mathcal{X}} \alpha(x,y)\nu_t(dy)$;
- each birth event causes a mutation with probability $\mu(x)$, in which case the new individual has a mutant trait $x+z$, where z has law $M(x,z)dz$.

Hence the process $(\nu_t, t \geq 0)$ is a Markov process on the set $\mathcal{M}_F$ of finite measures on $\mathcal{X}$, with generator

$$\begin{aligned} L\phi(\nu) = &\int_{\mathcal{X}} [\phi(\nu+\delta_x) - \phi(\nu)](1-\mu(x))b(x)\nu(dx) \\ &+ \int_{\mathcal{X}} \int_{\mathbb{R}^d} [\phi(\nu+\delta_{x+z}) - \phi(\nu)]\mu(x)b(x)M(x,z)dz\,\nu(dx) \\ &+ \int_{\mathcal{X}} [\phi(\nu-\delta_x) - \phi(\nu)] \left(d(x) + \int_{\mathcal{X}} \alpha(x,y)\nu(dy) \right) \nu(dx). \end{aligned}$$

The function α governs the interaction between individuals in the population, which is the origin of selection. Note that there is no pre-defined fitness: a proper notion of fitness should (and will) be defined in terms of the individual parameters.

The following asumption ensures the existence of the process ν [7]:
(H) $0 \leq b(x) \leq \bar{d}$, $0 \leq d(x) \leq \bar{d}$, $b(x) - d(x) > 0$, $0 \leq \underline{\alpha} \leq \alpha(x,y) \leq \bar{\alpha}$ and $M(x,z) \leq C\bar{M}(z)$ for some constant C and probability density $\bar{M}$.

2. Large population scalings

We introduce a scaling parameter K (linked to the biological notion of “carrying capacity”), and we make all the parameters depend on K: $b_K, d_K, \alpha_K, \mu_K, M_K$ and we assume that $\alpha_K(x,y) = \alpha(x,y)/K$. We will see that K scales the population size: we introduce

$$X_t^K = \frac{1}{K} \sum_{i=1}^{N(t)} \delta_{x_i(t)} = \frac{1}{K}\nu_t^K.$$

The first large population limit corresponds to the simplest case where $b_K = b, d_K = d, \mu_K = \mu$ and $M_K = M$:

Theorem 1. *Assume (H), $X_0^K \Rightarrow \xi_0$ deterministic, $\sup_K \mathbf{E}[\langle X_0^K, \mathbf{1}\rangle^3] < +\infty$ and that b, d, α and μ are continuous. Then, on $\mathbb{D}(\mathbb{R}_+, \mathcal{M}_F)$, $X^K \Rightarrow \xi \in \mathcal{C}([0,T], \mathcal{M}_F(\mathcal{X}))$ determiniustic such that $\xi(0) = \xi_0$ and*

$$\begin{aligned}(1)\qquad \langle \xi_t, f\rangle = \langle \xi_0, f\rangle + \int_0^t \int_{\mathcal{X}} \Big\{ \big[(1-\mu(x))b(x) - d(x) - \int_{\mathcal{X}} \alpha(x,y)\xi_s(dy)\big] f(x) \\ + b(x)\mu(x) \int f(x+z)M(x,z)dz \Big\} \xi_s(dx)ds.\end{aligned}$$

This result re-establishes Kimura's equation [8]. The proofs of the results of this section [3] are based on tightness, martingale problems and weak convergence techniques.

The second large population limit corresponds to an acceleration of births and deaths where $b_K(x) = K^\eta r(x) + b(x), d_K(x) = K^\eta r(x) + d(x), \mu_K(x) = \mu(x)$ and $M_K(x,z)dz = \mathcal{N}(0, \sigma^2(x)\mathrm{Id}/K^\eta)$.

Theorem 2. *Under the same assumptions as in Theorem 1, in the case where $\eta = 1$, $X^K \Rightarrow Z \in \mathcal{C}([0,T], \mathcal{M}_F)$ where the stochastic process Z is defined by: $\sup_{t\le T} \mathbf{E}[\langle Z_t, \mathbf{1}\rangle^3] < \infty$,*

$$\begin{aligned}(2)\qquad \langle Z_t, f\rangle = \langle \xi_0, f\rangle + \int_0^t \int_{\mathbb{R}^d} \Big\{ (b(x) - d(x) - \int_{\mathcal{X}} \alpha(x,y) Z_s(dy)) f(x) \\ + \frac{1}{2} r(x)\mu(x)\sigma^2(x)\Delta f(x) \Big\} Z_s(dx)ds + M_t^f\end{aligned}$$

where M_t^f is a continuous martingale such that

$$\langle M^f\rangle_t = 2\int_0^t \int_{\mathbb{R}^d} r(x) f^2(x) Z_s(dx)ds.$$

In the case where $\eta = 1$, X^K converges to the deterministic process obtained by taking $M^f \equiv 0$ in (2).

The case $\eta < 1$ re-establishes Fisher's reaction-diffusion equation [8] and the *superprocess* limit of the case $\eta = 1$ generalizes Etheridge's [6] model for spatially structured populations. It is possible to obtain a similar result by taking $M_K = M$ and rescaling μ_K accordingly [3].

3. Convergence to the trait substitution sequence

The TSS model is based on a biological heuristic of *time scale separation* between the birth and death events and the mutation events (see [9]): the selection process has sufficient time between two mutations to eliminate the disadvantaged traits. Therefore, we scale the parameters of the microscopic model as for the large population limit of Theorem 1, except for the mutation probability: $\mu_K = u_k\mu$, where $u_K \to 0$ when $K \to +\infty$.

Observe that, in the case where $\mu \equiv 0$ and $X_0^K = n_0^K \delta_x$ (monomorphic population), (1) rewrites $\xi_t = n(t)\delta_x$ where $\dot{n} = (b(x) - d(x) - \alpha(x,x)n)n$. This well-known logistic equation has a unique stable equilibrium $\bar{n}_x = (b(x) - d(x))/\alpha(x,x)$. Similarly, in the dimorphic case, $\xi_t = n(t)\delta_x + m(t)\delta_y$, where

$$\begin{cases} \dot{n} = (b(x) - d(x) - \alpha(x,x)n - \alpha(x,y)m)n \\ \dot{m} = (b(y) - d(y) - \alpha(y,x)n - \alpha(y,y)m)m. \end{cases}$$

Let us assume that this system of equations has no equilibrium except the trivial ones $(0,0), (\bar{n}_x, 0)$ and $(0, \bar{n}_y)$, which writes:
(H') For any $x \neq y \in \mathcal{X}$, $f(x,y)f(y,x) < 0$, where $f(y,x) = b(y) - d(y) - \alpha(y,x)\bar{n}_x$.

Theorem 3. *Assume (H), (H'), $X_0^K = \gamma_K \delta_x$ with $\gamma_K \to \gamma > 0$ and*

$$\forall C > 0, \quad \log K \ll \frac{1}{K u_K} \ll \exp(CK), \tag{3}$$

then the process $(X^K_{t/Ku_K}, t \geq 0)$ converges to

$$Z_t = \begin{cases} \gamma\delta_x & si\ t = 0 \\ \bar{n}_{Y_t}\delta_{Y_t} & si\ t > 0 \end{cases}$$

for finite dimensional distributions, where the Markov jump process $(Y_t, t \geq 0)$ on $\mathcal{X}$ satisfies $Y_0 = x$ and has as infinitesimal generator:

$$A\varphi(x) = \int_{\mathbb{R}^d} (\varphi(x+z) - \varphi(x))\mu(x)b(x)\bar{n}_x \frac{[f(x+z,x)]_+}{b(x+z)} M(x,dz).$$

Condition (3) gives the proper scaling of the mutation probability ensuring the required time scale separation (t/Ku_K corresponds to the time scale of mutations). The function $f(y,x)$ corresponds to the notion of fitness of a mutant trait y in a monomorphic population with trait x. The proof of this result [2] makes use of the problem of exit from a domain (large deviations) and of comparisons with branching processes.

References

[1] B.M. Bolker, S.W. Pacala. *Using moment equations to understand stochastically driven spatial pattern formation in ecological systems.* Theor. Popul. Biol. **52** (1997), 179–197.

[2] N. Champagnat. *A microscopic interpretation for adaptive dynamics trait substitution sequence models.* Preprint MODALX 04/20, University of Paris X (2004).

[3] N. Champagnat, R. Ferrière, S. Méléard. *Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models via timescale separation.* To appear in Theor. Popul. Biol.

[4] U. Dieckmann, M. Doebeli. *On the origin of species by sympatric speciation.* Nature **400** (1999), 354–357.

[5] U. Dieckmann, R. Law. *The dynamical theory of coevolution: A derivation from stochastic ecological processes.* J. Math. Biol. **34** (1996), 579–612.

[6] A. Etheridge. *Survival and extinction in a locally regulated population.* Ann. Appl. Probab. **14** (2004), 188–214.

[7] N. Fournier, S. Méléard. *A microscopic probabilistic description of a locally regulated population and macroscopic approximations.* Ann. Appl. Probab. **14** (2004), 1880–1919.

[8] M. Kimura. *A stochastic model concerning the maintenance of genetic variability in quantitative characters.* Proc. Natl. Acad. Sci. USA **54** (1965), 731–736.

[9] J.A.J. Metz, S.A.H. Geritz, G. Meszéna, F.A.J. Jacobs, J.S. van Heerwaarden. *Adaptive Dynamics, a geometrical study of the consequences of nearly faithful reproduction.* In: van Strien, S.J., Verduyn Lunel, S.M. (Eds.), Stochastic and Spatial Structures of Dynamical Systems. North Holland, Amsterdam, (1996), 183–231.

Genetic Hitchhiking

Alison Etheridge
(joint work with Peter Pfaffelhuber, Anton Wakolbinger)

Suppose that a favourable mutation arises at a particular genetic locus and that the mutant allele rapidly sweeps to fixation (that is increases in frequency until everyone in the population carries it). Then the genetic variability at a linked neutral locus will be reduced during the sweep as the neutral allele that happened to be associated with the new favoured mutation will increase in frequency, a process known as hitchhiking. This suggests that one might be able to detect selection acting on a locus from its indirect effect on linked neutral loci. A difficulty with this approach is that we must distinguish the effects of selection from other possible causes of reduced diversity. The first step is to understand the nature of the effect of selection on a linked neutral locus and in particular in which ways it will be apparent from a sample.

Let us suppose then that a sweep originates at a time that we label zero and is completed at time T. We ignore the effect of mutation which would, in reality, 'blur' the signal. We are going to look for the signature of the sweep at a linked neutral locus precisely at time T. In this way we are maximising our chances of finding a pattern that characterises the sweep, since after time T the pattern of variation at the neutral locus would be broken down by (neutral) resampling. The variation in a sample will be described in terms of 'families' determined by common ancestry at the neutral locus at the time of the origin of the sweep (zero in our notation). Our aim is to approximate the family size distribution.

This problem was first considered by Maynard Smith and Haigh, [5], who also coined the term *hitchhiking.* In their approximation, the frequency at the selected locus increases deterministically and the sample has at most one non-singleton family, corresponding to individuals whose ancestor at the time of origin of the sweep was lucky enough to be on the same genome as the favourable mutation.

The rest of the sample form singletons, corresponding to individuals whose ancestral lineages experienced a recombination event during the time course of the sweep. However, work of Barton, [1], revealed that ignoring the stochastic changes in allele frequency close to the beginning of the sweep could lead to substantial errors and, in particular, as a result of recombination events occurring during this early stochastic pahse, there could be more than one non-singleton family. More recently, Durrett and Schweinsberg, [2], [6], examined the family size distribution arising in a Moran model of a population of N diploid individuals undergoing a selective sweep. They show that up to an error of $\mathcal{O}(1/(\log N)^2)$ the family size distribution can be approximated by sampling from a 'paintbox', obtained by a stick-breaking regime based on Beta random variables. This compares with an error of $\mathcal{O}(1/\log N)$ in the Maynard Smith and Haigh approximation.

The Moran model studied by Durrett and Schweinsberg is as follows. We model the population of N diploids as $2N$ haploids. Label the individuals that carry the favoured allele at the selective locus B and the rest of the population b. Each individual lives for an (independent) exponentially distributed amount of time, with parameter one, at the end of which it dies and is replaced by a copy of an individual chosen at random from the $2N$ members of the population (including the one that died). To incorporate selection, each substitution of a type b individual for a type B individual is rejected with probability s. This determines the dynamics at the selected locus. We now incorporate the neutral locus. Let us write r for the probability of a recombination event between the neutral and selected loci in each generation. When a new individual is born, with probability $1-r$ it inherits alleles at both the neutral locus and the selected locus from the same parent, but with probability r the new individual inherits the two alleles from two different parents with the second parent also chosen (independently) at random (with replacement) from the $2N$ individuals in the population. We suppose that there are $2N$ labels at the neutral locus, one for each individual alive at the beginning of the sweep, and these labels are passed on unchanged from parent to child. Individuals with the same label at time T are then in the same *family*.

In the work of Durrett and Schweinsberg, the selective advantage s is held fixed and large but finite populations are considered. In [4] we take a different approach. Rather than considering large N with s held fixed, we measure time in units of size $2N$, and let N tend to infinity with $\alpha = 2Ns$ and $\rho = 2Nr$ held fixed, in other words we take a diffusion approximation. We then let α tend to infinity. Now in the Moran model, the duration of the sweep is $\mathcal{O}(\log Ns)$ and so in order to see a non-trivial family distribution at the neutral locus we must take r to be $\mathcal{O}(1/\log N)$. Correspondingly, in the diffusion timescale (where time is measured in units of $2N$ generations) the sweep has duration $\mathcal{O}(\log \alpha/\alpha)$ and so we must take ρ to be $\mathcal{O}(\alpha/\log \alpha)$. In this diffusion setting we are able to produce an approximate sampling formula, accurate up to an error of order $\mathcal{O}(1/(\log \alpha)^2)$, which on setting $\alpha = 2Ns$ and $\rho = 2Nr$ also provides an approximate sampling formula for the Durrett-Schweinsberg paintbox.

At first sight this approach should not work. Diffusion approximations are appropriate if one is considering weak selection, but for strong selection details of the reproduction mechanism in the Moran model will persist in the diffusion limit. However, the sampling formula and the Durrett-Schweinsberg paintbox are based on two key approximations. To describe them it is convenient to think of lineages ancestral to a sample from the neutral locus *migrating* (as we trace backwards in time from T to 0) between the two genetic backgrounds determined by their type at the selected locus. Through recombination a neutral allele can find itself associated with a different type at the selected locus before and after a reproduction event. Notice that the migration rate will depend on the proportions of the population in each of the two backgrounds. For example, if the current frequency of favoured alleles is X, then a neutral lineage in the favoured background will migrate at rate $\rho(1 - X)$ (as a proportion X of the recombination events that it experiences are with parents of the same type at the selected locus and therefore do not result in a change in background). The two underlying approximations can then be described as follows. First, the probability that a lineage ancestral to our sample recombines out of the favoured background and then back in again (tracing backwards in time from T to 0) is negligible and second, the chance that a pair of ancestral lineages coalesces in the less favoured background b is negligible. This allows one to reduce the analysis of the family size distribution to the study of marked genealogical trees at the selected locus, with marks representing recombination events through which a neutral lineage migrates from the favoured to the less favoured background. Via a timechange, this reduces to the study of a sample from a marked Yule tree. The Yule tree is the same whether we consider the Moran model or the diffusion limit (although the way that marks appear is slightly different) and indeed is insensitive to changes in the Moran model provided that the growth rate in the diffusion limit is unchanged. As a result, the diffusion approach exhibits an unexpected robustness.

The beauty of our approach is that everything is absolutely explicit. The disadvantage is that depending on the population size N, there is a limit on the size of sample that we can consider. Our approximation allows for at most two non-singleton families in our sample, but simulations in [1] reveal several non-singleton families. The reason for this apparent inconsistency is that for a sample of size n the error in our approximation actually scales with $(n/\log\alpha)^2$ and so we need very large populations or small samples for this to be controlled. In more recent work (to be reported in [3]) we see a resolution of this difficulty which allows us to increase the sample size at the expense of not being able to approximate the full distribution of family sizes, but instead treating certain partitions of our sample into families as indistinguishable.

References

[1] N. Barton, *The effect of hitchhiking on neutral genealogies*, Gen. Res. **72** (1998), 123–133.

[2] R. Durrett & J. Schweinsberg, *Approximating selective sweeps*, Theor. Pop. Biol **66** (2) (2004), 129–138.

[3] A. Etheridge *Genetic Hitchhiking*, In preparation.

[4] A. Etheridge, P. Pfaffelhuber & A. Wakolbinger, *An approximate sampling formula under genetic hitchhiking*, Submitted to Ann. Appl. Probab.

[5] J. Maynard Smith & J. Haigh, *The hitch-hiking effect of a favourable gene*, Gen. Res. **23** (1974), 23–35.

[6] J. Schweinsberg & R. Durrett, *Random partitions approximating the coalescence of lineages during a selective sweep*, Ann. Appl. Probab. **15** (2005), 1591–1651.

Mutation-selection balance and models of aging

Steven N. Evans
(joint work with David Steinsaltz, Ken Wachter)

Brian Charlesworth issued the following challenge in 2001.

> Senescence of multicellular plants and animals is an almost universal phenomenon; it needs to be explained both in terms of cellular and physiological mechanisms, and of evolutionary forces.

The mortality rate of an organism at age t is $\mathbb{P}\{\text{die in } [t, t+dt] \,|\, \text{live to } t\}/dt$ Gompertz discovered in 1837 that mortality for humans is an exponential function of age. This observation has since been made for many multi-cellular organisms. Fisher (1930), Haldane (1941), Medawar (1946, 1952), Williams (1957), Hamilton (1966), and Charlesworth (1994, 2001) proposed models of aging and mortality involving:

- large numbers of *mildly deleterious* mutations that *meander towards extinction* in the population but are *constantly reintroduced*,
- effects of mutations are *age-specific* and may even be *positive at early ages* e.g. mutations for efficient fat metabolism and Alzheimer's disease.

The main idea is that natural selection won't oppose mutations with deleterious effects that are felt after the individual has been able to reproduce. As Charleworth said in 2001,

> From the evolutionary perspective $\cdots$ senescence is an evolved response to the greater selective impact of genes which affect survival or fecundity early in life, relative to genes with act later in life.

How do we turn this intuitively appealing idea into **MATHEMATICS?** In our work, we have a complete, separable *metric space* $\mathcal{M}$ of potential *mutations* (we can have a finite or infinite # of loci). The set of possible *genotypes* is the space $\mathcal{G}$ are integer–valued *measures* on $\mathcal{M}$ that assign finite mass to bounded sets (thus we have essentially countable (multi-) sets of mutant alleles). The *null genotype* has wild-type alleles at every locus and carries none of the mutant alleles and is the *zero measure*. The state of the population at time t is a *probability measure* P_t on $\mathcal{G}$, i.e. $P_t(dg) =$ "proportion of the population with genotype g".

First consider a model **without selection** in which mutations arise at rates described by a measure ν on $\mathcal{M}$ that assigns finite mass to bounded sets. Write P_tF for $\int F(g)\,P_t(dg)$. Then

$$\frac{d}{dt}P_tF = P_t\left(\int [F(\cdot+\delta_m)-F(\cdot)]\,\nu(dm)\right).$$

Let Π denote the *Poisson random measure* with *intensity measure* $\nu \otimes$ Lebesgue and define a $\mathcal{G}$-valued *Lévy process* $(X_t)_{t\geq 0}$ by

$$X_t := \int_{\mathcal{M}\times[0,t]} \delta_m\, d\Pi(m,u).$$

Then

$$P_tF = \mathbb{E}\left[F(W+X_t)\right],$$

where W is a random measure with distribution P_0, independent of Π.

Now introduce **selection costs** by supposing that each genotype g has a positive *selection cost* $S(g)$. The cost S vanishes on the null genotype, and vanishes for no other g. Costs will typically be decrements to the intrinsic rate of natural increase (so that we are essentially measuring fitness on a *logarithmic scale*). When $S(g+\delta_m)-S(g)$ is independent of g, the model is said to be *additive* or *non-epistatic.*

The appropriate non-linear evolution equation is now

$$\begin{aligned}\frac{d}{dt}P_tF = P_t&\left(\int [F(\cdot+\delta_m)-F(\cdot)]\;\nu(dm)\right)\\ &- P_t(F[S-P_tS]).\end{aligned}$$

Note that

$$P_tS = \textit{average selection cost} \text{ of population}$$

$$S(g)-P_tS = \textit{relative cost} \text{ of genotype } g$$

If we take $\mathcal{M}$ to be a *single point*, then mutations are identical and a genotype is specified by a natural number, the number of mutant alleles present in it. Our model then becomes

$$\begin{aligned}\frac{dP_t(n)}{dt} &= \nu P_t(n-1)-\nu P_t(n)\\ &\quad - P_t(n)\left(S(n)-\sum_m S(m)P_t(m)\right)\end{aligned}$$

This model was introduced by Kimura and Maruyama in 1966.

We can solve the model with selection as follows. Define a *linear operator*

$$AF := \int [F(\cdot+\delta_m)-F(\cdot)]\,\nu(dm) - S(\cdot)F(\cdot)$$

(= generator of Lévy process *killed* at rate $S(g)$ in state g). By the *Feynman-Kac* formula, $\frac{d}{dt}\Gamma_t F = \Gamma_t(AF)$, where

$$\Gamma_t F(g) = \mathbb{E}\left[\exp\left(-\int_0^t S(g+X_u)\,du\right)F(g+X_t)\right].$$

By calculus, a solution is therefore

$$P_t F = \frac{P_0 \Gamma_t F}{P_0 \Gamma_t 1}.$$

Using this representation, it is possible to give necessary and sufficient conditions for the probability measure P_t to converge in distribution as t goes to infinity

The *expectation measure* of P_t is the measure R_t on $\mathcal{M}$ given by $R_t(B) := \int_{\mathcal{G}} g(B) dP_t(g)$. If P_0 is the law of Poisson random measure with intensity ρ_0. Then (by a *Palm-Campbell* calculation)

$$R_t(dm) = \zeta_t(m)\nu(dm) + \eta_t(m)\rho_0(dm),$$

where

$$\zeta_t(m) := \mathbb{E}\Big[\exp\big(-\int_0^t S(\tilde{X}_u)du\big) \times \int_0^t \exp\big(-\int_\tau^t [S(\tilde{X}_u+\delta_m) - S(\tilde{X}_u)]du\big)d\tau\Big] \Big/ \mathbb{E}\left[\exp\big(-\int_0^t S(\tilde{X}_u)du\big)\right]$$

$$\eta_t(m) := \frac{\mathbb{E}\left[\exp\big(-\int_0^t S(\tilde{X}_u+\delta_m)du\big)\right]}{\mathbb{E}\left[\exp\big(-\int_0^t S(\tilde{X}_u)du\big)\right]}$$

with $\tilde{X}_t = W + X_t$.

When mutation rates are low relative to selective pressures, the burden of mutations can explode. More specifically, consider $B \subseteq \mathcal{M}$ with $\nu(B) < \infty$ and suppose

$$\sup\{S(g+g') - S(g) : g'(B) = g'(\mathcal{M})\} < \nu(B).$$

Then, for all n,

$$\lim_{t\to\infty} P_t\{g : g(B) \le n\} = 0.$$

Suppose now that S is *non-epistatic*. Set $S(m) := S(\delta_m)$, so $S(g) = \int S(m)\, g(dm)$. Let M_t be a Poisson random measure on $\mathcal{M}$ with intensity $(1/S(m))(1 - e^{-S(m)t})\nu(dm)$ and let N_t be an independent random measure on $\mathcal{M}$ with distribution

$$\mathbb{E}[F(N_t)] = \frac{\int e^{-S(g)t} F(g) P_0(dg)}{\int e^{-S(g)t} P_0(dg)}.$$

The distribution of P_t is that of $M_t + N_t$ (proof via Laplace functionals and PDE). and the random measure M_t converges to a Poisson random measure with intensity

$(1/S(m))\nu(dm)$. The random measure N_t becomes concentrated on the genotypes in the support of P_0 that have minimum selection cost (a consequence of *Varadhan's lemma*).

Charlesworth attempted to produce Gompertzian hazards using a special case of our model. We show that Charlesworth's results are an artifact of approximations that he makes. Moreover, the appearance of Gompertz rates in his work is tied to *specifics* of the model rather than the *general structure* – so his approach is *not robust*, even though **Gompertz mortality is ubiquitous**.

The model above is for *haploids* and does not incorporate *recombination* i.e. there is no *meiosis* (formation of *gametes* = eggs or sperm) to make mosaics of different genotypes from the population).

Suppose we assume

(1) Homozygotes for mutant alleles are negligible;
(2) Selection is weak relative to recombination;
(3) Recombination can split all parts of the genome.

Assumption 1 is equivalent to assuming that we have, not a diploid organism, but a haploid organism that goes through a phase of sexual reproduction and meiosis with recombination. Selection is active in the haploid phase.

Assumptions 1 and 2 are essentially those that underlie the *quasi-linkage equilibrium (QLE) approximation* of Barton and Turelli. Mathematically, we have a model arising as a *Trotter product*: i.e., a limit of *high frequency oscillations* between a *selective phase* and a *recombinant phase* on a *faster time scale*.

A natural sequence of discrete generation models satisfying the above assumptions converges to a model with the following description. For $\pi \in \mathcal{H} := \{$finite measures on $\mathcal{M}\}$, let X^π be a *Poisson* random measure with *intensity* measure π. Define a *non-linear operator* $D : \mathcal{H} \to \mathcal{H}$ by

$$(D\pi)(dm) := \mathbb{E}[S(X^\pi + \delta_m) - S(X^\pi)]\, \pi(dm).$$

Suppose that P_0 is the distribution of X^{ρ_0} for $\rho_0 \in \mathcal{H}$. Then P_t is the distribution of X^{ρ_t} where

$$\rho_t = \rho_0 - \int_0^t D\rho_s \, ds + t\nu.$$

This model is the subject of ongoing work in which are beginning to understand its long term equilibrium behavior for certain cost functions.

Two variance results in population genetics

Warrren J Ewens

In this talk two variance results in population genetics are discussed. The first relates to the optimal way of estimating the amount of genetic variation in a population, and the second to assessing the difference that two investigators, sampling from the same population at the same time, will have between their respective estimates of genetic variation.

This genetic variation is best described by the parameter θ, defined by $\theta = 4Nu$, where N is the (unknown) population size and u the (unknown) mutation rate to new alleles. Questions about genetic variation in the population are best approached as questions about the parameter θ, as discussed below.

The first problem can be described as follows. We do not of course have population information, and are given instead a sample of n DNA sequences of arbitrary length L, (possibly corresponding to the DNA for some gene), with no recombination between the sites in the sequence. Variation in this sample can be measured in two ways, first "horizontally" and second "vertically". In the horizontal case, there will be some number k of different sequences among the n sequences. It is known that k is a sufficient statistic for the parameter θ, so that any statement about θ as derived from the sample, and hence about genetic variation as measured horizontally, is best carried out by using k. The mean square error of the estimate of θ, using k, is approximately

$$\frac{\theta}{\sum_{j=1}^{n-1} \frac{j}{(j+\theta)^2}}.$$

The "vertical" estimate of θ is found by using the number of "polymorphic sites" in the sample of n sequences. A polymorphic site is one where there are two (or more) nucleotides represented in the sample. If the number of polymorphic sites is denoted by s, then θ is estimated by s/g_1, where $g_1 = \sum_{j=1}^{n-1} j^{-1}$, and the mean square error of this estimator is given by

$$\frac{\theta}{g_1} + \frac{\theta^2 g_2}{g_1^2},$$

where $g_2 = \sum_{j=1}^{n-1} j^{-2}$.

The comparison of the "horizontal" and the "vertical" estimates of genetic variation reduces to a comparison of this mean square error and the mean square error of the estimate of θ using k, as given above, the preferred estimating procedure attaching to that with the smaller mean square error.

It is straightforward to see that the ratio of the two mean square errors approaches 1 as θ approaches 0. This is as we expect, and forms a check on the results.

It is found that for some combinations of θ and n the vertical estimate is preferred and for other combinations the horizontal estimate is to be preferred. The vertical estimate is to be preferred at least whenever $\theta \leq 1$ and also whenever $n \leq 50$. When both $\theta > 1$ and also $n > 50$, however, the horizontal estimate can be preferred. For example, when $\theta = 3, n = 500$, the mean square error for the "horizontal" estimate is about 5% lower than that of the "vertical" estimate.

The reason for the fact that sometimes one estimate is preferred and sometimes the other has to do with the correlation between sites, due to the assumption that there is no recombination between them. Given the complete coalescent of sample, all information about where the mutations that caused the different sequences occurred would be available, and an optimal estimate of θ would then be available. Both the "horizontal" and the "vertical" data are incomplete in that they do not give these data, and in some cases the loss of information is greater in the "horizontal" case and in other cases it is greater in the "vertical" case.

The second problem has to do with the difference of the estimates of θ found by two investigators, each taking a sample of n sequences from the same population at the same time. The "narrow", or "sampling" variance of the difference $|k_1 - k_2|$ between the respective numbers k_1 and k_2 of sequences found by the two investigators is $\frac{1}{2}E(k_1 - k_2)^2$. This can be found by imagining a total sample of $2n$ sequences, of which we think of the first n belonging to the first investigator and the remaining n to the second.

For small values of n a direct calculation can be made by using the Ewens sampling formula. The simplest possible case is for $n = 2$. Here there are five possible "sequence configurations", namely {4}, {3,1},{2,2}, {2,1,1} and {1,1,1,1}. The configuration {2,1,1}, for example, means that of the sample of four sequences, two are identical and the other two are different from each other and from the two identical sequences. We can write this alternatively as the partition *AABC*.

For the configuration {4}, or equivalently the partition *AAAA*, both investigators see only one sequence, so that $k_1 - k_2 = 0$. At the other extreme, namely the configuration {1,1,1,1}, or equivalently the partition *ABCD*, both investigators see two sequences, so that again $k_1 - k_2 = 0$. The configuration {2,2} implies that both investigators see one allele (if investigator 1 sees *AA* and investigator 2 sees *BB*), or that both investigators see two alleles (both see *AB*). Again, in all cases, it is necessarily true that $k_1 - k_2 = 0$.

For the configuration {3,1}, it is necessarily the case that one investigator sees one sequence and the other sees two. Thus for this case, $|k_1 - k_2| = 1$. The Ewens sampling formula shows that the probability of this configuration is

$$\frac{8\theta}{(1+\theta)(2+\theta)(3+\theta)},$$

so that this configuration contributes half this amount to $\frac{1}{2}E(k_1 - k_2)^2$.

There are two possibilities for the configuration {2,1,1}. In the first of these, investigator 1 sees a configuration of the form *AA* and investigator 2 sees a configuration of the form *BC*. One third of {2,1,1} configurations are of this type, and for this type, $|k_1 - k_2| = 1$. In the second type, investigator 1 sees a configuration of the form *AB* and investigator 2 sees a configuration of the form *AC*. Thus for this type $|k_1 - k_2| = 0$. The probability of the configuration {2,1,1} is

$$\frac{6\theta^2}{(1+\theta)(2+\theta)(3+\theta)},$$

so that this configuration contributes one sixth of this amount to $\frac{1}{2}E(k_1 - k_2)^2$. All of this implies that the "narrow" variance is

$$\frac{4\theta+\theta^2}{(1+\theta)(2+\theta)(3+\theta)}.$$

The "total" variance of k is, for the case $n = 2, \theta/(1+\theta^2)$, and the ratio of the narrow variance to this is

$$\frac{(1+\theta)(4+\theta)}{(2+\theta)(3+\theta)},$$

which clearly approaches 2/3 as θ approaches 0, and approaches 1 as θ increases to large values.

A similar, but more complicated, calculation can be done for the case of a sample of three genes per investigator, and in principle for the case of an arbitrary number n of genes per investigator. Clearly, however, this approach becomes impractical for other than a small number of genes per investigator, and in general another approach is needed. This can be done by using univariate and bivariate frequency spectra and indicator random variables. It is found, using the frequency spectrum approach, that the narrow variance is asymptotically

$$\sum_{j=n}^{2n-1} \frac{\theta}{\theta + j} + O(n^{-1}),$$

and is thus approximately $\theta \log 2$ for large n. (It is perhaps unexpected that this variance is asymptotically independent of n, the sample size.) This conclusion implies that the narrow variance of the estimators of θ is $\theta \log 2/(\log n)^2$.

It is found that the narrow variance of s, the number of segregating sites seen by the two investigators, is asymptotically $\theta \log 2$. This implies that the narrow variance of the estimate of θ, using the number of segregating sites, is also asymptotically $\theta \log 2/(\log n)^2$. This result is perhaps surprising, since there is no a priori reason why these two narrow variances should be the same.

Behaviour of Poisson-Dirichlet distribution for large mutation rate

Shui Feng

(joint work with Donald A. Dawson)

The talk is based mainly on results in [1]. The large deviation results on age-class sizes are new.

Large Deviation for Poisson-Dirichlet Distribution. Let $U_1, U_2, ...$ be i.i.d. with common density function $f(u) = \theta(1-u)^{\theta-1}, 0 < u < 1$. Set

$$X_1 = U_1, X_n = U_n(1-U_1)\cdots(1-U_{n-1}), n \geq 2,$$

and $(P_1, P_2, ...)$ be the descending order of $(X_1, X_2, ...)$. Then the law Π_θ of $(P_1, P_2, ...)$ on space

$$\Delta = \{(p_1, p_2, ...) : p_1 \geq p_2 \geq \cdots \geq 0, \sum_{k=1}^{\infty} p_k \leq 1\}$$

is called the Poisson-Dirichlet distribution with parameter θ. Here Δ is equipped with the subspace topology of R^∞ and θ is proportional to the effective population size when individual mutation rate per generation is held constant.

Let $\mathbf{0} = (0, 0, ...)$. It is well known that Π_θ converges weakly to $\delta_{\mathbf{0}}$ when θ approaches infinity. For each fixed $m \geq 1$, let V_m be a continuous random variable with density function $\frac{1}{\Gamma(m)} \exp[-mv - e^{-v}], -\infty < v < \infty$. Then the following fluctuation theorem was obtained in [4].

Theorem 1. *For each $m \geq 1$, $\theta[P_m - \frac{\log\theta}{\theta} - \frac{\log\log\theta}{\theta}]$ converges weakly to V_m as θ converges to infinity.*

A family of probability measures $\{Q_\theta : \theta > 0\}$ on a topological space E is said to satisfy a large deviation principle (LDP) with speed $a(\theta)$ and rate function $I(\cdot)$ if for any closed set F and open set G

$$\limsup_{\theta\to\infty} \frac{1}{a(\theta)} \log Q_\theta\{F\} \leq -\inf_{x\in F} I(x),$$

$$\liminf_{\theta\to\infty} \frac{1}{a(\theta)} \log Q_\theta\{G\} \geq -\inf_{x\in G} I(x),$$

$$\text{for any } c > 0, \{x : I(x) \leq c\} \text{ is } \textit{compact}.$$

The following theorem is from [1].

Theorem 2. *The family of $\{\Pi_\theta : \theta > 0\}$ satisfies a LDP with speed $a(\theta) = \theta$ and rate function $I(\mathbf{p}) = \log \frac{1}{1-\sum_{k=1}^{\infty} p_k}$.*

Let Ψ be a bounded continuous function on Δ, $\sigma(\theta) > 0$. The Poisson-Dirichlet distribution with selection force Ψ and intensity $\sigma(\theta)$ is a probability measure on Δ defined as

$$\Pi^\Psi_{\sigma,\theta}(d\mathbf{p}) = \frac{\exp[\sigma(\theta)\Psi(\mathbf{p})]}{E^{\Pi_\theta}\{\exp[\sigma(\theta)\Psi(\mathbf{q})]\}} \Pi_\theta(d\mathbf{p}).$$

For any positive number c, set

$$I_c(\mathbf{p}) = sup_{\mathbf{q}\in\Delta}[c\Psi(\mathbf{q}) - I(\mathbf{q})] - [c\Psi(\mathbf{p}) - I(\mathbf{p})],$$

$$I_\infty(\mathbf{p}) = \begin{cases} 0, & \mathbf{p} = \mathbf{0} \\ \infty, & \text{else.} \end{cases}$$

Then the following result is also obtained in [1].

Theorem 3. *The family of $\{\Pi^\Psi_{\sigma,\theta} : \theta > 0\}$ satisfies a LDP with speed $a(\theta) = \theta$ and rate function*

$$I^\Psi_\sigma(\mathbf{p}) = \begin{cases} I(\mathbf{p}), & \textit{if } \lim_{\theta\to\infty} \frac{\sigma(\theta)}{\theta} = 0 \\ I_c(\mathbf{p}), & \textit{if } lim_{\theta\to\infty} \frac{\sigma(\theta)}{\theta} = c > 0 \end{cases}$$

If Ψ has a unique maximum and $lim_{\theta\to\infty} \frac{\sigma(\theta)}{\theta} = \infty$, then the family of $\{\Pi^\Psi_{\sigma,\theta} : \theta > 0\}$ satisfies a LDP with speed $a(\theta) = \theta$ and rate function $I_\infty(\mathbf{p})$.

Application. For each $m \geq 1$, let $H_m(\mathbf{p}) = \sum_{i=1}^{\infty} p_i^m$ be the mth order homozygosity. If the selection force $\Psi(\mathbf{p}) = -H_m(\mathbf{p})$, the heterozygote has advantage over the homozygote. In [3], simulation was done to study the role of population size in the infinie-alleles model with heterozygote advantage. It was observed and conjectured that if the selection intensity $\sigma(\theta)$ is scaled as the mutation rate θ, then the model looks like the neutral model. Let $\Phi_m(\mathbf{p}) = \frac{\exp[-\sigma(\theta)H_m(\mathbf{p})]}{E^{\Pi_\theta}\{\exp[-\sigma(\theta)H_m(\mathbf{p})]\}}$ denote the likelihood ratio. The following result was obtained in [5].

Theorem 4. *Under Π_θ, as θ goes th infinity,*

$$\Phi_m(\mathbf{p}) \Rightarrow \begin{cases} 1, & \textit{if } \lim_{\theta\to\infty}\frac{\sigma(\theta)}{\theta^{3/2}} = 0 \\ \exp[cZ_2 - c^2], & \textit{if } lim_{\theta\to\infty}\frac{\sigma(\theta)}{\theta^{3/2}} = c > 0 \\ 0, & \textit{if } \lim_{\theta\to\infty}\frac{\sigma(\theta)}{\theta^{3/2}} = \infty, \end{cases}$$

where Z_2 is a normal random variable with mean zero and variance 2. Thus in terms of the likelihood ratio, the model with heterozygote advantage behaves like the neutral model for large θ if $\sigma(\theta)$ grows slower than $\theta^{3/2}$, which confirms Gillespie's conjecture.

As a special case of Theorm 3, the LDP for the Poisson-Dirichlet distribution with heterozygote selection advantage holds as follow.

Theorem 5. *The family of $\{\Pi_{\sigma,\theta}^{-H_m} : \theta > 0\}$ satisfies a LDP with speed $a(\theta) = \theta$ and rate function*

$$I_\sigma^{-H_m}(\mathbf{p}) = \begin{cases} I(\mathbf{p}), & \textit{if } \lim_{\theta\to\infty}\frac{\sigma(\theta)}{\theta} = 0 \\ I(\mathbf{p}) + cH_m(\mathbf{p}), & \textit{if } lim_{\theta\to\infty}\frac{\sigma(\theta)}{\theta} = c > 0 \\ I_\infty(\mathbf{p}), & \textit{if } \lim_{\theta\to\infty}\frac{\sigma(\theta)}{\theta} = \infty. \end{cases}$$

From this theorem, we see that a phase transition occurs for the large deviation rate functions at the critical scale θ. Thus, in terms of large deviation rate functions, the model with heterozygote advantage behaves like the neutral model for large θ if $\sigma(\theta)$ grows slower than θ. At the critical scale θ, the selection can still be detected.

New Results. Consider a sample of size n from a Poisson-Dirichlet distribution with parameter θ. Let $X_{1,n}, ..., X_{n,n}$ be the age-class sizes in the sample. Then from [2], one has

$$P\{X_{1,n} = k\} = \frac{\theta}{n}\frac{\begin{pmatrix} n \\ k \end{pmatrix}}{\begin{pmatrix} \theta + n - 1 \\ k \end{pmatrix}} = \frac{\theta}{n}\frac{n!}{(n-k)!}\frac{(\theta+n-k-1)!}{(\theta+n-1)!},$$

and

$$P\{X_{1,n} = k_1, ..., X_{r,n} = k_r\} = \frac{(\theta/n)^r}{(1-k_1/n)\cdots(1-k_1/n-\cdots-k_{r-1}/n)} \times \frac{n!}{(n-k_1-\cdots-k_r)!}\frac{(\theta+n-k_1-\cdots-k_r-1)!}{(\theta+n-1)!}.$$

Then depending on the ratio between n and θ, the LDP results for each fixed $1 \leq r \leq n$ are summarized in the following table with

$$\begin{aligned} S(x_1,...,x_r) &= (c+1)\log(c+1) + (1-\sum_{i=1}^{r} x_i)\log(1-\sum_{i=1}^{r} x_i) \\ &- (c+1-\sum_{i=1}^{r} x_i)\log(c+1-\sum_{i=1}^{r} x_i). \end{aligned}$$

ratio	speed a_θ	rate function
n fixed, θ large	$\log\theta$	$(\sum_{i=1}^r k_i - r)$
$\lim_{\theta\to\infty} \frac{\theta}{n} = \infty$	$n\log\frac{\theta}{n}$	$\sum_{i=1}^r x_i$
$\lim_{\theta\to\infty} \frac{\theta}{n} = c > 0$	n	$S(x_1,...,x_r)$
$\lim_{\theta\to\infty} \frac{\theta}{n} = 0$	θ	$\log\frac{1}{1-\sum_{i=1}^r x_i}$

References

[1] D.A. Dawson and S. Feng, *Asymptotic behavior of Poisson-Dirichlet distribution for large mutation rate* , to appear Ann. Appl. Probab..

[2] P. Donnelly and S. Tavaré, *The ages of alleles and a coalescent*, Adv. Appl. Prob.,**12**, 1-19.

[3] J.H. Gillespie, *The role of population size in molecular evolution*, Theor. Pop. Biol. **55** (1999), 145–156.

[4] R.C. Griffiths, *On the distribution of allele frequencies in a diffusion model*, Theor. Pop. Biol. **15** (1979), 140–158.

[5] P. Joyce, S. M. Krone, and T.G. Kurtz, *When can one detect overdominant selection in the infinite-alleles model?*, Ann. Appl. Probab. **13**, No.1, (2003), 181–212.

Ewens' sampling formula

Bob Griffiths

(joint work with Sabin Lessard)

Ewens' (1972) sampling formula (ESF) is the probability distribution of a configuration of alleles in a sample of genes under the infinitely-many-alleles-model of mutation where every mutation is a new type. Warren Ewens is a pioneer in Mathematical Genetics so it is a pleasure to speak at this meeting on the history of the ESF and new results.

Ewens' sampling formula, the probability of a sample of n genes having k types with b_j types represented j times, $\sum jb_j = n$, and $\sum b_j = k$, is

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

The formula is derived by a short combinatorial argument. There are a number of different stochastic models giving rise to this important formula, particularly an urn model of Hoppé (1984), which is related to random partitions in Joyce and Tavaré (1987). In Hoppé's urn model initially there is one black ball of mass θ in the urn. Successively balls are drawn from the urn. If a ball is black it is returned with a ball of a new colour, if not black a ball of mass 1 of the same colour as the ball drawn is added to the urn. When there are n non-black balls with k different colours, the colours are randomly labelled $1, 2, \dots, k$. The distribution of the configuration of balls of different colours is ESF. To obtain a cycle representation in a random permutation in Hoppé's urn model, label the balls according to the order that they enter the urn. New colours start a new cycle, or if ball k's colour was determined by choosing ball j insert it in a cycle to the left of j. The cycle lengths are distributed as in the ESF. Arratia *et. al.* (2003) use the ESF and related limit Poisson Dirichlet distribution as $n \to \infty$ in studying cycle lengths in random permutations.

In a coalescent model of ancestry (Kingman, 1982) a binary tree is formed back in time beginning at the leaves of the tree, with coalescence at rate $\binom{k}{2}$ while k edges in the tree, and mutations occur at rate $\theta/2$ on the edges of the tree. Mutant families in the leaves of the tree have their type determined by the mutation subtending the family and mutations above these defining mutations have no effect on the sample configuration of types. A random forest which determines the sample configuration is defined by stopping lineages back in time on the tree when a mutation first occurs. Each tree in the forest then represents the ancestry of a mutant family. In this reduced process, lineages are lost back in time at rate $\binom{j}{2}$ by coalescence and $j\theta/2$ by mutation while j edges. The combinatorial derivation of the ESF is found by considering arrangements of the forest giving rise to families of sizes $n_1, n_2, \dots, n_k$. Let $T_n, T_{n-1}, \dots, T_1$ be times while lineages are lost back in time by coalescence or mutation. In a variable population size model, let $\lambda(t)$ be the relative population size at time t back to the present size. For example with exponential growth forward in time, $\lambda(t) = \exp(-\beta t), \beta > 0$.

The rate of coalescence at time t when j ancestor lines is $\binom{j}{2}\lambda(t)^{-1}$ and the rate of mutation is $\frac{j\theta}{2}$. An age-ordered sampling formula when the population size varies is

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

where the constants are

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

This reduces to the age-ordered ESF from Donnelly and Tavaré (1986) in the constant population size case

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

A connection is explored between the distribution of age-ordered frequencies in a sample and record values and record indices in a random permutation. The extension of this to the population as $n \to \infty$ is to consider record values and record indices in a sequence of independent uniform $[0,1]$ random variables. The connection is that age-ordered allele frequencies (from oldest to youngest) $n_1, \dots, n_k$, conditional on mutations arising by mutation from the top of the tree down while $i_1(=1), i_2, \dots i_k$ edges in the forest is identical to the conditional ditribution of the increments $s_1, s_2 - s_1, s_3 - s_2, \dots$ given record indices $i_1, i_2, \dots, i_k$. A coalescent tree and associated forest is well defined as $n \to \infty$ because of the quadratic coalescence rate. In this limit the partial sums of age-ordered allele frequencies $\big\{ \sum_{\nu=1}^m X_\nu, m \geq 1 \big\}$ given $i_1, i_2, \dots$ are distributed as record values in a sequence of independent uniform random variables $\{U_l; l \geq 1\}$ given they occur at record indices $i_1, i_2, \dots$. This leads to a random partition of the population

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

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

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

The indices $\{i_j; j \geq 1\}$, where $i_1 = 1$ form a Markov chain with transition probabilities

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

The distribution of the oldest allele frequency X_1 can be found from the record value representation. Conditional on $\{T_j,\ j \geq 2\}$

$$-\log(X_1) = \sum_{j=2}^{\infty} \gamma_j$$

where $\{\gamma_j,\ j > 1\}$ are independent random variables. γ_j has an atom at zero with probability $(1 + \rho_j)^{-1}$, and a continuous density of

$$\frac{\rho_j}{1 + \rho_j} \cdot (j-1) \cdot e^{-(j-1)\gamma},\ \gamma > 0$$

where $\rho_j = \theta\lambda(T_j)/(j-1)$. The mth oldest allele frequency has a representation

$$-\log(X_m) = \sum_{k=2}^{i_m} \delta_k + \sum_{j=i_m+1}^{\infty} \gamma_j$$

where $\{\delta_j, j > 1\}$ are independent exponential random variables such that δ_j has rate $j - 1$.

The research in this talk is published in Griffiths and Lessard (2005).

References

[1] Arratia, A., Barbour, A.D., Tavaré, S. *Logarithmic Combinatorial Structures: A Probabilistic Approach.* European Mathematical Society Publishing House, Switzerland. (2003).

[2] Donnelly, P., Tavaré, S. *The ages of alleles and a coalescent.* Advances in Applied Probability **18** (1986), 1–19.

[3] Ewens, W.J. *The sampling theory of selectively neutral alleles.* Theoretical Population Biology **3** (1972), 87–112.

[4] Ewens, W.J. *Population genetics theory – The past and the future.* In: Lessard, S. (Ed.), Mathematical and Statistical Developments of Evolutionary Theory. NATO ASI Series C: Mathematical and Physical Sciences, Vol. 299, pp. 177–227, Kluwer Academic Publishers, Dordrecht, The Netherlands. (1990)

[5] Griffiths, R. C. and Lessard, S. *Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles.* Theoretical Population Biology (2005) In Press.

[6] Hoppé, F.M. *Polya-like urns and the Ewens' sampling formula.* Journal of Mathematical Biology **20** (1984), 91–94.

[7] Joyce, P., Tavaré, S. *Cycles, permutations and the structures of the Yule process with immigration.* Stochastic Processes and Their Applications **25** (1987), 309–314.

[8] Karlin, S. McGregor, J.L. *Addendum to a paper of W. Ewens.* Theoretical Population Biology 3 (1972), 113–116.

[9] Kingman, J.F.C. *The coalescent.* Stochastic Processes and Their Applications **13** (1982), 235–248.

Estimating the time to the most recent common ancestor of a sample of sequences

Richard Hudson

The times to the most recent common ancestor of a sample of DNA sequences can be estimated by an variety of methods. The estimates are of great interest to biologists for assessing a range of hypotheses about population genetic processes and population history, in humans and other organisms. Some methods are based on specific population genetic models, others are "model-free" at least in terms of the population genetic ascpects. A new variance result is obtained for a simple estimate of Thomson et al. (2001):

$$\sigma_t^2 = t - \frac{n-1}{n}\frac{E_{ij}}{2}$$

Ergodic behaviour of locally regulated branching populations

Martin Hutzenthaler
(joint work with Anton Wakolbinger)

We consider a diffusion limit of a branching coalescing particle system on the lattice $\mathbb{Z}^d$ in which

- ➔ each particle migrates with rates $\alpha\, m^{\dagger}(i,j)$,
- ➔ each particle splits with rate $\beta + s$ into two new particles,
- ➔ each particle dies with rate β and
- ➔ each pair of particles on the same site coalesces with rate 2γ.

The model is given by the solution of

$$\begin{aligned} dX_t(i) &= \alpha \sum_{j\in\mathbb{Z}^d} m_{ij}\big(X_t(j) - X_t(i)\big)\,dt \\ &\quad + \gamma X_t(i)\big(K - X_t(i)\big)\,dt + \sqrt{2\beta X_t(i)}\,dB_t(i) \end{aligned} \tag{1}$$

where the $B(i)$ are independent Brownian motions and $\alpha, \beta, \gamma, K > 0$.

The first of two main results is concerned with a phase transition. Etheridge [1] proved that there is a critical capacity $K_c = K_c(\alpha, m, \gamma, \beta) \in [0, \infty)$ such that the process survives for $K > K_c$ and suffers local extinction for $K < K_c$. Survival means that $X_t(0)$ does not converge to zero in probability whenever started from a spatially homogeneous initial state. We complement this result with the following

Theorem 1. *Define $\overline{K} > 0$ as the unique solution of*

$$\int_0^\infty \exp\big(\overline{K}\gamma y - \frac{\gamma\beta}{2}y^2\big)\,\cdot\,\alpha\exp\big(-\alpha y\big)\,dy = 1. \tag{2}$$

Then $\overline{K} \le \inf_m K_c(m)$, i.e. for all $K \le \overline{K}$ the process dies out almost surely

$$\forall |x| < \infty:\ \mathbf{P}^x\big(\exists t \ge 0 : |X_t| = 0\big) = 1 \tag{3}$$

and suffers local extinction

$$\mathcal{L}(X_t) \Longrightarrow \delta_{\underline{0}} \qquad \text{as } t \to \infty. \tag{4}$$

whenever started from a spatially homogeneous initial state. Here, $\underline{0}$ denotes the zero configuration.

The main idea of the proof is a comparison with a *mean field model* given by the solution of

$$dV_t = \alpha(\mathbf{E}V_t - V_t)\,dt + \gamma V_t(K - V_t)\,dt + \sqrt{2\beta V_t}\,dB_t\,. \tag{5}$$

This process is mathematically better tractable. We show that $\mathcal{L}(X_t(i))$, $i \in \mathbb{Z}^d$ is *dominated by $\mathcal{L}(V_t)$ in the Laplace transform order.*

More precisely, we prove

Proposition 2. *Let X be a solution of* (1) *with associated initial distribution $\bar{\mu}$ satisfying $\mathcal{L}(X_0(i)) = \mathcal{L}(X_0(0))$ for all i and $\mathbf{E}X_0(0) < \infty$. Denote by V the solution of* (5) *with initial distribution $\mu := \mathcal{L}(X_0(0))$. Then*

$$\mathbf{E}^{\bar{\mu}} \exp\big(-\lambda X_t(i)\big) \geq \mathbf{E}^{\mu} \exp\big(-\lambda V_t\big), \qquad \lambda \geq 0. \tag{6}$$

The consequence of this proposition is that extinction of the mean field model implies local extinction of X.

The second main result considers the long term behaviour of the process. There is at most one non-trivial translation invariant equilibrium. The process converges to this invariant measure (or to $\delta_{\underline{0}}$) whenever started from a non-trivial spatially homogeneous initial state. On an intuitive level, the reason for this is as follows: There are two forces working against each other: super-critical branching and individual competition. Super-critical branching increases mass, whereas fighting amongst the individuals decreases it. If a (local) population size is large then competition takes more effect, whereas as long as a local population size is small the competition is negligible in comparison to the mass producing branching. Thus, there should be some attracting equilibrium state in which the two forces balance each other.

Theorem 3. *There is an invariant measure $\bar{\nu}$: If $\mathcal{L}(X_0)$ is translation invariant and does not charge the zero configuration $\mathbf{P}(X_0 = \underline{0}) = 0$, then*

$$\mathcal{L}(X_t) \implies \bar{\nu} \qquad \text{as } t \to \infty \tag{7}$$

The principal tool for proving ergodic behaviour is a duality.

Proposition 4. *Let $X^\dagger$ be a solution of* (1) *with transposed migration kernel $m^\dagger$ given by $m^\dagger(i,j) = m(j,i)$. Then we have the following self-duality:*

$$\mathbf{E}^{x} \exp\big(-\frac{\gamma}{\beta}\langle X_t, y\rangle\big) = \mathbf{E}^{y} \exp\big(-\frac{\gamma}{\beta}\langle x, X_t^\dagger\rangle\big) \tag{8}$$

for all suitable x, y.

The main advantage of this self-duality is that instead of starting in infinite total mass we can analyse the evolution of the process started with finite total mass. For example, choose $y = \lambda\delta_0$ and $x(i) \equiv$ const. Then the self-duality tells us that it makes no difference whether we study the law of $(X_t(0))_{t\geq 0}$ started in x or whether we study the total mass $|X_t^\dagger|$ with $X_t^\dagger$ started in $\lambda\delta_0$, $\lambda > 0$. For the total mass process one can apply martingale methods to study its long term behaviour.

References

[1] A. M. Etheridge, *Survival and extinction in a locally regulated population*, Ann. Appl. Probab. **14** (2004), 1, 188–214.

Stochastic Demography, Coalescents, and Effective Population Size

Stephen M. Krone
(joint work with Ingemar Kaj, Magnus Nordborg, Martin Lascoux, Per Sjödin)

The notion of "effective population size" has been a fixture in population genetics for a long time. It is, however, a concept that is too often misused and, at times, serves simply as an informal device for avoiding or ignoring the demographic complications that invariably arise in real populations.

The classical concepts of effective population size have typically been used when, in the calculation of a particular quantity (e.g., the probability of identity by descent), a given population model behaves in the same way as the standard neutral, panmictic Wright–Fisher model with constant population size. Some common examples are the "inbreeding effective population size," the "variance effective population size," and the "eigenvalue effective population size." Unfortunately, such quantities sometimes do not exist and, perhaps more importantly, even when they do, they need not be equal. In [2]–[4], we have proposed the notion of "coalescent effective population size" which avoids much of the ambiguity of earlier effective sizes and is much less susceptible to misuse.

By definition, the coalescent effective size exists when the suitably re-scaled ancestral process (with one unit of time corresponding to N generations) converges to a linear time change of Kingman's coalescent. When this is the case, all polymorphism data (from samples that are not of the same order as the population size) will be indistinguishable from those arising from a standard Wright–Fisher model. Thus one can use the Wright–Fisher model to calculate all quantities of interest. In Nordborg and Krone [3], it was shown that a coalescent effective size exists when demographic events occur on a fast time scale relative to coalescence events. This condition is more informative than the classical concepts of effective size, even in a theoretical setting, since the existence of, say, an inbreeding effective size does not imply that the population behaves like a Wright–Fisher model in any other respect.

In Kaj and Krone [2] (cf. also Donnelly and Kurtz [1]), it was shown that when a population model with stochastically fluctuating population size experiences large size changes on the same time scale as for coalescence events, then there is no coalescent effective size. In fact, under quite general conditions, the re-scaled ancestral process will converge weakly to a nonlinear stochastic time change of Kingman's coalescent. In particular, if the population size τ generations in past is a Markov chain given by $M_N(\tau) = NX_N(\tau)$, where the relative size process $X_N([Nt]) = N^{-1}M_N([Nt])$ converges weakly to a continuous-time Markov process $X(t)$ as $N \to \infty$ (e.g., a diffusion process or a continuous-time jump chain), and if the probability, $c_N\big(M_N(\tau-1), M_N(\tau)\big)$, that two lineages coalesce when going from generation $\tau - 1$ to generation τ (in past) satisfies

$$c_N(k,m) = \frac{1}{N}H_N(\frac{k}{N},\frac{m}{N}),$$

where $H_N(\frac{k}{N}, \frac{m}{N}) \to H(x, y)$ as $k/N \to x$ and $m/N \to y$, then the time change is given by

$$\int_0^t H(X_s, X_s) ds.$$

The limiting coalescent is then of form

$$A_N([Nt]) \Rightarrow A(\int_0^t H(X_s, X_s) ds),$$

where $A(t)$ denotes Kingman's coalescent. In such cases, the effects of size fluctuations will show up in polymorphism data. For example, in Cannings-type models with exchangeable reproduction, one has

$$\begin{aligned} H_N&(\frac{k}{N}, \frac{m}{N}) \\ &= \Big(\frac{k}{N}(\frac{k}{N} - \frac{1}{N})\Big)^{-1} \frac{md}{N} \to \frac{yd}{x^2} \equiv H(x, y). \end{aligned}$$

In Sjödin et al. [4], computer simulations of Fu and Li's F statistic were used to assess the effects on polymorphism data of deviations from the standard Wright–Fisher assumptions. In particular, two simple demographic models–one with randomly fluctuating population size, and the other with subdivided populations linked by migration–were simulated to uncover differences in cases for which there is a coalescent effective size and cases for which there is not. When size fluctuations have an effect on F, this effect was seen to be very much dependent on sample size and initial population size. For example, large negative values of F (which can be a signature of population expansion forward in time) were exclusively obtained when the population size at the time of the sample was the larger of two possible sizes.

In the case of subdivided populations, if migration between subpopulations is sufficiently fast compared to coalescence events, the effects of subdivision will be felt in the coalescent only in an average sense. Essentially, the migration process has time to reach equilibrium between coalescence events. In this case there will be a coalescent effective population size and the genealogy will be given by Kingman's coalescent with a linear time change. If, on the other hand, migration events are "intermediate" in the sense that they occur on the same time scale as coalescences, then the resulting genealogical process will be described by a structured coalescent. In this case, the genealogy cannot be thought of a standard coalescent and there is no coalescent effective population size. Simulations resulted in positive values of Fu and Li's F when migration rates were not sufficiently fast, as expected when the genealogy is described by a structured coalescent. An interesting feature of the simulations was that they pointed out how fast the migration had to be to result in the effects of subdivision being averaged out, and they showed the effects of subdivision on F when migration was not fast enough. The effects of subpopulation size were also important.

References

[1] P. Donnelly and T.G. Kurtz, *Particle representations for measure-valued population models*, Ann. Probab. **27** (1999), 166–205.

[2] I. Kaj and S.M. Krone, *The coalescent process in a population with stochastically varying size*, J. Appl. Probab. **40** (2003), 33–48.

[3] M. Nordborg and S.M. Krone, *Separation of time scales and convergence to the coalescent in structured populations*, In Modern Developments in Theoretical Population Genetics. M. Slatkin and M. Veuille (eds.) (2002)

[4] P. Sjödin, I. Kaj, S.M. Krone, M. Lascoux, M. Nordborg, *On the meaning and existence of an effective population size*, Genetics **169** (2005), 1061–1070.

A Poisson Model Heuristic for Judging the Significance of Gapped Local Alignments

Dirk Metzler

1. Introduction

When two DNA or protein sequences differ only by a few mutations, the following question of significance arises. Could this similarity between the sequences be due to pure chance or does it indicate a common ancestry or function? The mutations we consider here are nucleotide or amino acid substitutions in single positions as well as insertions and deletions of sequence fragments. Similarities and differences in sequences can be displayed in an *alignment*, as for example the following one:

```
AGTC___AGTTC__GTG
ACTCACTAG_TCAAGCG
 ^             ^
```

Two positions marked with ˆ in this alignment are *mismatches*, which indicate that substitutions have occurred. The three stretches of underscores, so-called *gaps*, correspond to fragments that have been inserted or deleted. The similarity of two sequences can be measured by a score function. For example, we could give each match a reward of $+1$, each mismatch a penalty of $-\mu$, each gap a penalty of $-\Delta$ and each position in a gap a penalty of $-\delta$, so that the score of the alignment above would be $9 - 2\mu - 3\Delta - 6\delta$. (In commonly used scoring schemes the match rewards and mismatch penalties may depend on the involved nucleotide or amino acid types.) For a given pair of sequences the algorithm of Smith and Waterman [9] finds the local alignment of highest score, *local* means that only parts of the sequences are related. The BLAST software [1, 2] is suitable for finding high-scored local alignments between given sequences and sequence databases.

The significance of a local alignment score s can be judged by its E-value, which is the expected number of non-overlapping local alignments of score $\geq s$ under a null hypothesis of unrelated sequences. Dembo et al. [4] consider the null hypothesis that all positions in both sequences are taken independently from a distribution on the base types or amino acid types. They showed that (under certain conditions on the scoring parameters) the E-value of score $\geq s$ for *gapless* local alignments is for large s asymptotically $nmke^{-\lambda s}$, where n and m are the sequence lengths and the constants k and λ can be computed numerically. Altschul et al. [3] conjectured that also in the case of *gapped* local alignments the E-value for score $\geq s$ is asymptotically of the form $nmk'e^{-\lambda' s}$ and estimate k' and λ' from simulation studies and data base comparisons. The results are used for E-value estimations in newer versions of the BLAST program. Siegmund and Yakir [7, 8] showed that in the asymptotic of large s and $\Delta \sim \log s$, such that $s \cdot e^{-\lambda\Delta} \to \beta > 0$, the E-value is $nmce^{-\lambda s}$ with the same constant λ in the gapped and in the ungapped case. The constant c can be computed numerically.

2. A Poisson Model for Gapped Local Alignments

In [5] we suggest a heuristic Poisson model for gapped local alignment. We replace the set of sequence pairs $\{1, \dots, n\} \times \{1, \dots, m\}$ by a continuous rectangle $R = [0, n] \times [0, m]$. Inspired by the results of Dembo et al. [4] we assume that high-scored gapless local alignments are thrown into R according to a Poisson point process of intensity k and that their scores are independently exponentially distributed with parameter λ. A high-scored gapped alignment in the Poisson model is a sequence of high-scored ungapped alignments following each other closely. Given an ungapped alignment at $x \in \mathbb{R}$ we assume that the range $U_x(g) \subset R$ of points that we could reach for a gap penalty of $\Delta + g$ scales in g like a 2-dimensional shape, i.e. there is a constant ω such that the area of $U_x(g)$ is $\omega g/2$.

In [6] a formula for computing E-values of high-scored local alignments from given parameters k, λ, and ω is derived. In [5] we show that the asymptotic E-value for high scores s and $s \cdot e^{-\lambda\Delta} \to \beta > 0$ is $k \cdot e^{-\lambda s} \cdot e^{\omega k \beta/\lambda}$, which coincides with Siegmund and Yakir's result if we set ω appropriately. In simulation studies in [5] it turned out that this value of ω fits well with the heuristic motivation of ω given above.

The Poisson model also gives a heuristic explanation why the asymptotic $s, \Delta \to \infty$ with $s \cdot e^{-\lambda\Delta} \to \beta > 0$ is tractable. Under this assumption, the number of gaps in the best local alignment in the Poisson model is asymptotically Poisson distributed with a finite, positive expectation value. We conjecture that the same is true in the sequence-based model in [7, 8].

3. Applications

Former approaches for assigning an E-value to local-alignment scores are based on the assumption that all positions of each DNA or protein sequence are i.i.d., cf. [4, 7, 8], or that all protein sequences are similarly composed, cf. [3]. Both assumptions are uncertain, and also not necessary when the E-values are based on the Poisson model. Given a pair of sequences, we first estimate k, λ and ω for this particular pair of sequences from their configuration of high-scored ungapped local alignments. Then we can use these values to compute the E-value of the best gapped local alignment. Simulation studies in [6] give evidence that this works well, especially when the asymptotic formula is used. In simulations with slightly untypical amino acids compositions or with correlations between neighboring sequence positions, the E-values based on the Poisson model were much more reliable than the E-values given by BLAST.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Basic local alignment search tool*, J. Mol. Biol. **215** (1990), 403–410

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucl. Acids. Res. **25** (1997), 3389-3402

[3] S. F. Altschul, R. Bundschuh, R. Olsen, T. Hwa, *The estimation of statistical parameters for local alignment score distributions*, Nucl. Acids. Res. **29** (2001), 351-361

[4] A. Dembo, S. Karlin and O. Zeitouni, *Limit distribution of maximal non-aligned two-sequence segmental score*, Ann. Probab. **22** (1994), 2022–2039

[5] D. Metzler, S. Grossmann, A. Wakolbinger, *A Poisson model for gapped local alignments*, Prob. Stat. Letters **60** (2002), 91–100

[6] D. Metzler, *Robust E-Values for Gapped Local Alignments*, Preprint (2005)

[7] D. Siegmund, B. Yakir, *Approximate p-values for Local Sequence Alignments*, Annals of Statistics **28** (2000), 657–680

[8] D. Siegmund, B. Yakir, *Correction: Approximate p-values for local sequence alignments*, Annals of Statistics **31** (2003), 1027–1031

[9] T. F. Smith, M. Waterman *Identification of common molecular subsequences*, J. Mol. Biol., **147** (1981), 195–197

Coalescent theory – simultaneous multiple collisions and sampling distributions

Martin Möhle

1. Coalescent processes with multiple collisions

Let Λ be a finite measure on $[0,1]$. The Λ-coalescent (Pitman [16], Sagitov [17]) is a time-continuous Markovian process $R = (R_t)_{t\geq 0}$ with state space $\mathcal{E}$, the set of all equivalence relations on $\mathbb{N} := \{1,2,\ldots\}$, and infinitesimal rates

$$q_{\xi\eta} \;=\; \begin{cases} \displaystyle\int_{[0,1]} \frac{1-(1-x)^{b-1}(1-x+bx)}{x^2}\,\Lambda(dx) & \text{if } \xi=\eta,\\ \displaystyle\int_{[0,1]} x^{b-a-1}(1-x)^{a-1}\,\Lambda(dx) & \text{if } \xi\prec\eta,\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

where $a := |\eta|$ and $b := |\xi|$ are the number of blocks (equivalence classes) of the equivalence relations η and ξ, and $\xi \prec \eta$ means (by definition) that exactly $b-a+1$ (≥ 2) blocks of ξ merge together to form one block of η, while all the other

$a-1$ blocks of ξ remain unchanged. If $\Lambda = \delta_0$ is the Dirac measure concentrated in 0, then the Λ-coalescent is Kingman's coalescent [10, 11] with binary mergers of ancestral lineages. If $\Lambda = U$ is uniformly distributed on $[0,1]$, then R is the Bolthausen-Sznitman coalescent [1]. If Λ is concentrated in 1, then the process R is called a star-shaped coalescent. From (1) it follows that the block counting process $D = (D_t)_{t\geq 0} := (|R_t|)_{t\geq 0}$ is a Markovian death process with rates

$$(2)\quad g_{nk} \;=\; \frac{n!}{(k-1)!(n-k+1)!}\int_{[0,1]} x^{n-k-1}(1-x)^{k-1}\,\Lambda(dx), \quad n,k\in\mathbb{N}, k<n$$

and total rates $g_n = \sum_{k=1}^{n-1} g_{nk} = \int_{[0,1]}(1-(1-x)^{n-1}(1-x+nx))x^{-2}\Lambda(dx)$, $n\in\mathbb{N}$. For example, for the Bolthausen-Sznitman coalescent we have $g_n = n-1$ and $g_{nk} = n/((n-k)(n-k+1))$.

2. Sampling distributions

In population genetics the ancestry of a sample of n genes is often modelled by the process $(\varrho_n R_t)_{t\geq 0}$, where $\varrho_n : \mathcal{E}\to\mathcal{E}_n$, $\varrho_n(\xi) := \{(i,j)\,|\,1\leq i,j\leq n, (i,j)\in\xi\}$, denotes the natural projection to the (finite) set $\mathcal{E}_n$ of all equivalence relations on $\{1,\ldots,n\}$. By definition, two individuals i and j of the sample have the same ancestor at time t in the past if and only if $(i,j)\in\varrho_n R_t$. Assume now that each individuals is of a certain type. Mutations are superimposed on the genealogical tree $(\varrho_n R_t)_{t\geq 0}$ as follows: Conditional on the tree, mutations occur independently of the tree at the points of a homogeneous Poisson process with rate $r>0$ on each branch of the tree. Usually, the infinitely-many-alleles model is assumed, i.e. each mutation leads to a new type (allele) never seen before.

Fix $n\in\mathbb{N}$ and sample n individuals from the population. For $i\in\mathbb{N}$, let a_i denote the number of types in the sample which appear i times. Note that $n=\sum_{i=1}^\infty ia_i$, i.e. $\mathbf{a} := (a_1,a_2,\ldots)\in\mathbb{N}_0^\infty := \{0,1,2,\ldots\}^\infty$ is a partition of n. In particular, $a_i = 0$ for $i>n$. Of fundamental interest in population genetics is the probability $q(\mathbf{a})$ that we have sampled a specific partition $\mathbf{a}$ of types. The sampling probabilities $q(\mathbf{a})$ satisfy the following recursion on n (Möhle [12]): $q(1,0,0,\ldots)=1$ and

$$(3)\qquad q(\mathbf{a}) \;=\; \frac{nr}{g_n+nr}q(\mathbf{a}-\mathbf{e}_1) + \sum_{i=1}^{n-1}\frac{g_{n,n-i}}{g_n+nr}\sum_{j=1}^{n-i}\frac{j(a_j+1)}{n-i}\,q(\mathbf{a}+\mathbf{e}_j-\mathbf{e}_{i+j})$$

for any partition $\mathbf{a}=(a_1,a_2,\ldots)\in\mathbb{N}_0^\infty$ with $n=\sum_i ia_i\geq 2$, where $\mathbf{e}_j$ denotes the jth unit vector in $\mathbb{R}^\infty$ and the convention $q(\mathbf{a}) := 0$ is used whenever some of the entries of $\mathbf{a}$ are negative. Define the probability generating function $f_n(s^{(n)}) := \sum_{\mathbf{a}} q(\mathbf{a})s_1^{a_1}\cdots s_n^{a_n}$, $s^{(n)} := (s_1,\ldots,s_n)\in\mathbb{R}^n$, where the sum extends over all partitions $\mathbf{a}$ of n. In terms of f_n, the recursion (3) is equivalent to

$$(4)\qquad (g_n+nr)f_n(s^{(n)}) \;=\; nrs_1f_{n-1}(s^{(n-1)}) + \sum_{i=1}^{n-1}\frac{g_{ni}}{i}\sum_{j=1}^{i} js_{n-i+j}\frac{\partial}{\partial s_j}f_i(s^{(i)}),$$

i.e. $f_1(s_1)=s_1$, $f_2(s_1,s_2) = (2rs_1^2+g_2s_2)/(g_2+2r)$ and so on. Solutions of the recursion (3), or equivalently (4), are only known for special cases. For the

Kingman coalescent ($\Lambda = \delta_0$), the solution of the recursion (3) is the celebrated Ewens sampling formula (Ewens and Tavaré [2])

$$q(\mathbf{a}) \;=\; \frac{n!}{[\theta]_n}\prod_{i=1}^{n}\left(\frac{\theta}{i}\right)^{a_i}\frac{1}{a_i!}, \tag{5}$$

where $[\theta]_n := \theta(\theta+1)\cdots(\theta+n-1)$. For the star-shaped coalescent ($\Lambda = \delta_1$), the solution q (Möhle [12, Section 4]) corresponds to a hook composition structure (Gnedin and Pitman [6, Section 6]). For general finite measure Λ, the recursion (3) seems to be difficult to solve for arbitrary partitions $\mathbf{a}$. However, for certain partitions $\mathbf{a}$ of n, solutions can be derived easily. For example, $q(n,0,0,\ldots) = \prod_{i=2}^{n}(ir/(g_i+ir))$.

Only a few explicit examples for sampling distributions are known from the literature (Gnedin [3, 4]). Pitman [15] studied a family of sampling distributions depending on two parameters $\alpha,\theta \in \mathbb{R}$ such that either $0 \le \alpha < 1$ and $\theta \ge -\alpha$ or $\alpha < 0$ and $\theta = -m\alpha$ for some fixed $m \in \mathbb{N}$. The case $\alpha < 0$ goes at least back to Keener et al. [9]. The corresponding sampling distributions have the form

$$q_{\alpha,\theta}(\mathbf{a}) \;=\; \frac{n!\,[\theta+\alpha]_{k-1,\alpha}}{[\theta+1]_{n-1}}\prod_{i=1}^{n}\left(\frac{[1-\alpha]_{i-1}}{i!}\right)^{a_i}\frac{1}{a_i!}, \tag{6}$$

where $n = \sum_i ia_i$ is the (given) sample size, $k := \sum_i a_i$ is the number of types and the notation $[\theta]_{0,\alpha} := 1$ and $[\theta]_{k,\alpha} := \theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)$, $k \in \mathbb{N}$, is used. For $\alpha = 0$, equation (6) reduces to the Ewens sampling formula (5). For the sub-range of parameters $0 \le \alpha < 1$ and $\theta \ge 0$, the composition structure which corresponds to Pitman's sampling distributions (6) is regenerative (Gnedin and Pitman [5, Section 8]). For more information on regenerative composition structures, in particular on their asymptotics for large n, we refer to Gnedin, Pitman and Yor [7, 8]. In contrast (Möhle [13]), the composition structure which corresponds to the sampling distributions (3) induced by a Λ-coalescent with mutation rate $r > 0$ is regenerative if and only if the measure Λ is either concentrated in 0 (Kingman case) or concentrated in 1 (star-shaped case). The Ewens sampling formula ($\alpha = 0$) is the only case in which Pitman's two parameter sampling distribution coincides with a sampling formula induced by a Λ-coalescent, namely the Kingman coalescent (Möhle [13]).

3. Extensions to processes with simultaneous multiple collisions

There exists a wider class of sampling distributions (Möhle [12, Section 5]), in which the underlying coalescent process allows for simultaneous multiple collisions of ancestral lineages. These coalescent processes are characterized in terms of a sequence of measures Λ_j on the simplex $\Delta_j := \{x = (x_1,\ldots,x_j) \in [0,1]^j \,|\, x_1+\cdots+x_j \le 1\}$, $j \in \mathbb{N}$ (Möhle and Sagitov [14], Schweinsberg [18]). We conjecture that all the corresponding composition structures are non-regenerative, except for the case when Λ_j is the zero-measure for all $j \ge 2$ and $\Lambda := \Lambda_1$ is either concentrated in 0 or concentrated in 1.

References

[1] E. Bolthausen and A.-S. Sznitman, *On Ruelle's probability cascades and an abstract cavity method*, Comm. Math. Phys. **197** (2) (1998), 247–276.

[2] W.J. Ewens and S. Tavaré *The Ewens sampling formula*, In *Multivariate Discrete Distributions* (N. S. Johnson, et al., eds), (1995) Wiley, New York.

[3] A. Gnedin, *Three sampling formulas*, Combin. Probab. Comput. **13** (2004), 185–193.

[4] A. Gnedin, *The Bernoulli sieve*, Bernoulli **10** (2004), 79–96.

[5] A. Gnedin and J. Pitman, *Regenerative composition structures*, Ann. Probab. **33** (2005), 445–479.

[6] A. Gnedin and J, Pitman, *Regenerative partition structures*, The Electronic Journal of Combinatorics **11** (2) #R12 (2004/2005), 1–21.

[7] A. Gnedin, J. Pitman and M. Yor, *Asymptotic laws for regenerative compositions: gamma subordinators and the like*, (2004) Preprint.

[8] A. Gnedin, J. Pitman and M. Yor, *Asymptotic laws for compositions derived from transformed subordinators*, (2005) Preprint.

[9] R. Keener, E. Rothman and N. Starr, *Distributions on partitions*, Ann. Stat. **15** (1987), 1466–1481.

[10] J.F.C. Kingman, *The coalescent*, Stoch. Process. Appl. **13** (1982), 235–248.

[11] J.F.C. Kingman, *Origins of the coalescent: 1974-1982*, Genetics **156** (2000), 1461–1463.

[12] M. Möhle, *On sampling distributions for coalesent processes with simultaneous multiple collisions*, Bernoulli **11** or **12** (2005 or 2006), in press.

[13] M. Möhle, *On a class of non-regenerative sampling distribution*, (2005) Preprint.

[14] M. Möhle and S. Sagitov, *A classification of coalescent processes for haploid exchangeable population models*, Ann. Probab. **29** (2001), 1547–1562.

[15] J. Pitman, *Exchangeable and partially exchangeable random partitions*, Probab. Theory Relat. Fields **102** (1995), 145–158.

[16] J. Pitman, *Coalescents with multiple collisions*, Ann. Probab. **27** (1999), 1870–1902.

[17] S. Sagitov, *The general coalescent with asynchronous mergers of ancestral lines*, J. Appl. Probab. **36** (1999), 1116–1125.

[18] J. Schweinsberg, *Coalescents with simultaneous multiple collisions*, Electron. J. Probab. **5** (2000), 1–50.

All about Eve. On the evolution of the time since the last MRCA

Peter Pfaffelhuber
(joint work with Anton Wakolbinger)

In a continuum population whose forward evolution follows a standard Wright-Fisher diffusion, the time span back from time 0 to the most recent common ancestor (MRCA) is distributed like S_1^∞, where $S_i^\infty = \sum_{j=i+1}^\infty T_j$, with T_j independent and $\exp\left(\binom{j}{2}\right)$-distributed; this is the time Kingman's coalescent needs to come down from infinitely many to j lines (see [Lit75], [Gri80]. [Kin82]). When the population evolves, the current MRCA (of all those that live at time 0) will remain in business up to a random time E when the next MRCA's offspring takes over; let us denote the time when the new MRCA lived by B. (B stands for *begin* and E for *end*, since during the time interval (B, E) the frequency of the new MRCA's offspring develops from 0 to 1.) We compute the joint distribution of the random pair (E, B) given that the current MRCA lived at time $-d < 0$, and we investigate the dynamics of the time stationary process $\mathcal{F} = \big((E_n, B_n)\big)_{n\in\mathbb{Z}}$. This is summarized in Figure 1.

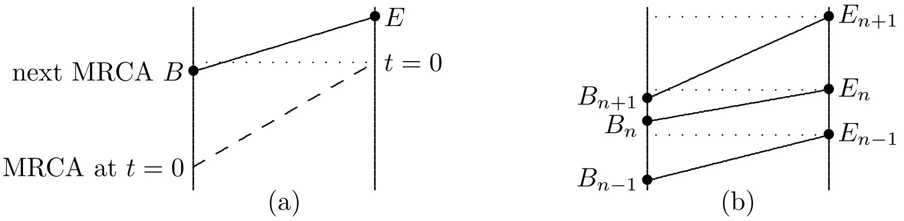

Figure 1. Times when the MRCA changes (E-axis) and when the new ancestor will have lived (B-axis) (a) as seen from $t = 0$, (b) as a stationary process

1. From today's MRCA to the next MRCA

In a time-stationary Wright-Fisher population, the time E has an exponential distribution with mean 1, see e.g. [Wat82]. A quick argument is that the split induced by the sizes of the two oldest families at time 0 is uniformly distributed, and remains also at time $t > 0$ given no one of the two families has fixated up to time t. Let us name the two oldest families at time 0 by the S- and the H-family.

To study the joint distribution of E and B we make use of two random variables L and I defined as follows. L is the number of individuals living at time 0 that still have offspring at the time E. In case $L > 1$ we let I denote the number of lines in the full coalescent back from time 0 at the time B, and in case $L = 1$ we put $I = \infty$.

To calculate the distribution of L we use deFinetti's Theorem to see that, given the split of the whole population into the H- and the S-family is $(p, 1-p)$, a randomly sampled individual is in the H-family with probability p and in the S-family with probability $1-p$, independent of all others. This also applies to that individual among all those living at time t whose offspring lives longest, second longest etc. We call these these individuals the *most persistent, second most persistent* individual etc.

We have $L = \ell$ exactly when the most persistent, ..., ℓth most persistent individual are altogether either in the H- or the S family and the $(\ell+1)$st most persistent individual is in the other family. This gives

$$\mathbf{P}[L = \ell] = 2\int p^{\ell}(1-p)dp = \frac{2}{(\ell+1)(\ell+2)}.$$

From this equation we can calculate the probability that the next MRCA will live after time 0, i.e. in the future of time 0. This is the case iff $L = 1$, and consequently

$$\mathbf{P}[\text{next MRCA lives after time } 0] = \mathbf{P}[B > 0] = \mathbf{P}[L = 1] = \tfrac{1}{3}. \tag{1}$$

The L individuals that still have offspring up to time E form a subsample of all those alive at time 0. Their ancestral lines can be traced back in the full coalescent and using Pólya urns one can prove that the joint distribution of the two is

$$\mathbf{P}[L = \ell, I = i] = \begin{cases} \dfrac{\ell!(\ell-1)}{3}\dfrac{1}{\binom{i+\ell}{i}}, & \ell \geq 2, i \geq 3 \\ \frac{1}{3}, & \ell = 1, i = \infty. \end{cases} \tag{2}$$

To find the joint distribution of (E, B) we have to find, given L and I, the distribution of the times when the $(L+1)$st most persistent line disappears and when there are I lineages in the full coalescent back from time 0, which represent the times E and B, respectively. The latter is given by a random split of the coalescence time S_1^{∞}, i.e. when the current time to the MRCA is $d > 0$ this is given by $R_{i,d}$ which is distributed as S_i^{∞} given $S_1^i + S_i^{\infty} = d$. The time when the $(L+1)$st most persistent lineage disappears can be studied using the (modified) look-down process which was introduced in [DK99]. In this process every individual has a label and every individual *looks down* to any other individual with a smaller label at rate 1. When a look-down event occurs from individual j to individual i the individual i has one offspring with label j. All individuals with label j or more increase their label by 1.

In this process individuals disappear as soon as their label has been increased infinitely often. Given $L = \ell$ the label of individual $\ell + 1$ is increased infinitely often after a time which is distributed as S_{ℓ}^{∞}. All this gives the following Theorem.

Theorem. Let (L, I) be as in (2). Given, the time of the MRCA at time 0 is time $-d$, the next MRCA will be at time E and it will have lived at time B, where the joint distribution of (E, B) is represented by

$$(E, B) = \begin{cases} (S_1^2 + S_2^\infty, S_1^2) & \text{if } L = 1, \\ \\ (S_L^\infty, -R_{I,d}) & \text{if } L > 1. \end{cases}$$

2. The process of MRCAs

Let us now turn to the point process $\mathcal{F} := \big((E_n, B_n)\big)_{n\in\mathbb{Z}}$ of consecutive pairs of times when a MRCA enters the population and when it is found. It turns out that the look-down process is very helpful in this respect.

Consider the look-down process at a time when individual 2 looks down to individual 1. This is at a time B_n since there will be a time when the first three, four, five,... individuals in the look-down will be descendants of these two at time B_n. Ultimately (that is at time E_n) all individuals will have the individuals 1 and 2 from time B_n as ancestors. Thus, it turns out that times B_n and E_n are linked by a line through the look-down process which we call a *fixation line.*

Closely attached to this fixation line is the coalescent of the whole population back from time E_n, a process which occurs also in [Taj90]. Unlike Kingman's coalescent, when there are n lines, the coalescence rate for this coalescent is $\binom{n+1}{2}$.

Using the look-down picture we can describe also the interaction of fixation lines, and obtain a complete description of the process $\mathcal{F}$. Indeed, the fixation lines interact since it is possible that a future MRCA is born in the population still before an older one has fixed.

The interaction of the fixation causes the process $\mathcal{F}$ to be non-Markov. Nevertheless a measure for the memory of the process $\mathcal{F}$ is the number Z of fixation lines to come that overlap with the one between B_0 and E_0. The distribution of Z can be given explicitly, some values are

$$\mathbf{P}[Z = 0] = \tfrac{1}{3}, \qquad \mathbf{P}[Z = 1] = \tfrac{11}{27}, \qquad \mathbf{E}[Z] = 1.$$

The $\frac{1}{3}$ already appeared in (1)., the connection being that $Z = 0$ corresponds to the case, seen from time E_n, that no fixation line overlaps with the nth, or, equivalently, all MRCAs to come will live in the future of E_n.

More details and proofs can be found in [PW05]. We are grateful to John Wakely and Dick Hudson who pointed out the connections to Watterson's and Tajima's work [Wat82] and [Taj90] during the meeting.

References

[DK99] P. Donnelly and T.G. Kurtz. Particle representations for measure-valued population models. *Annals of Probability*, 27(1):166–205, 1999.

[Gri80] R. C. Griffiths. Lines of descent in the diffusion approximation of neutral Fisher-Wright models. *Theor. Pop. Biol.*, 17:37–50, 1980.

[Kin82] J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982.

[Lit75] R.A. Littler. Loss of variability at one locus in a finite population. *Math. Biosci.*, 25:151–163, 1975.

[PW05] P. Pfaffelhuber and A. Wakolbinger. The MRCA process in an evolving coalescent. *submitted.*

[Taj90] F. Tajima. Relationship between DNA polymorphism and fixation time. *Genetics*, 125:447–454, 1990.

[Wat82] G.A. Watterson. Mutant substitutions at linked nucleotide sites. *Adv. Appl. Prob.*, 14:166–205, 1982.

A coalescent model for the effect of advantageous mutations on the genealogy of a population

Jason Schweinsberg
(joint work with Rick Durrett)

When an advantageous mutation occurs in a population, the favorable allele may spread to the entire population in a short time, an event known as a selective sweep. To model a selective sweep, we consider a population of size $2N$, which represents the $2N$ chromosomes of N diploid individuals. We consider two sites on the chromosomes. At one site, there are two possible alleles, denoted by B and b, and B is advantageous. At the other site, all possible alleles are neutral. We assume that at time zero, $2N-1$ chromosomes have the b allele, while one chromosome, which has just had a beneficial mutation, has the B allele. We assume that each individual independently lives for an exponential time with mean 1 and then is replaced by a new individual whose parent is chosen at random from the population, except we disregard disadvantageous replacements of a B chromosome by a b chromosome with probability s. We assume that the new individual inherits its alleles at both sites from the same chromosome with probability $1-r$. With probability r, because of recombination, the new individual inherits its alleles at the two sites from two ancestors chosen independently at random from the population. We then condition the number of B chromosomes to reach $2N$ before 0, which is the event that a selective sweep occurs.

A selective sweep affects the genealogy not only of the site at which the mutation occurred but also of the neutral site. When the recombination probability r is small, the allele at the neutral site on the chromosome that had the beneficial mutation will increase in frequency as a result of a selective sweep, a process known as "hitchhiking". At the neutral site, most likely many of the lineages will be traced back to the individual that had the beneficial mutation at the beginning of the selective sweep. However, others will "escape" the selective sweep because of recombination and be traced back to different ancestors. We focus on the case of strong selection, where the selective advantage s is $O(1)$ and the recombination probability r is $O(1/(\log N))$.

Define a random partition Θ of $\{1, \dots, n\}$, obtained by sampling n individuals from the population at the end of the selective sweep and declaring i and j to be in the same block of Θ if and only if the ith and jth individuals in the sample inherited their allele at the neutral site from the same ancestor at the beginning of the sweep. To get a simple approximation to the distribution of Θ, first approximate the probability that a lineage fails to escape the selective sweep by $p = (2N)^{-r/s}$. Then define a random partition Θ_p as follows. Flip n independent coins which come up heads with probability p. One block of Θ_p consists of all of the integers whose coins come up heads (corresponding to lineages that do not experience recombination), while all of the other integers are in blocks by themselves. We show [11] that there is a constant C such that for all partitions π of $\{1, \dots, n\}$,

$$|P(\Theta = \pi) - P(\Theta_p = \pi)| \leq \frac{C}{\log N}.$$

Thus, the distribution of Θ converges to that of Θ_p as $N \to \infty$. However, as noted by Barton, approximating Θ by Θ_p does not work well in practice. Typically not all lineages that escape the selective sweep get traced back to different ancestors, which means that Θ has more than one non-singleton block. This happens because, when we trace the lineages backwards in time, some groups of lineages will coalesce and then escape the sweep together, usually near the beginning of the sweep.

To obtain a more accurate approximation, we take advantage of the fact that near the beginning of the selective sweep, the number of individuals with the B allele can be approximated by a continuous-time branching process in which each individual gives birth at rate 1 and dies at rate $1 - s$. The individuals in such a branching process who have an infinite line of descent form another branching process in which there are no deaths and each individual gives birth at rate s. For the purposes of considering the genealogy of a sample taken a long time into the future, it is a good approximation to consider only individuals with an infinite line of descent.

Let τ_k be the first time at which there are k individuals with an infinite line of descent. The probability of a recombination along one of the k lineages with an infinite line of descent between times τ_k and τ_{k+1} is approximately r/s. If such a recombination occurs, then the fraction of the population, a long time into the future, descended from the lineage with the recombination has approximately the beta distribution with parameters 1 and $k - 1$. Therefore, we can define a random partition Π, which approximates the distribution of Θ, as follows. Let $M = \lfloor 2Ns \rfloor$. Let $(W_k)_{k=2}^M$ be independent random variables such that W_k has a Beta distribution with parameters 1 and $k - 1$.

Let $(\zeta_k)_{k=2}^M$ be a sequence of independent random variables, also independent of the W_k, such that $P(\zeta_k = 1) = r/s$ and $P(\zeta_k = 0) = 1 - r/s$ for all k. The event that $\zeta_k = 1$ corresponds to a recombination between times τ_k and τ_{k+1}. For $k = 2, 3, \dots, M$, let $V_k = \zeta_k W_k$, and let $Y_k = V_k \prod_{j=k+1}^L (1 - V_j)$, which approximates the fraction of lineages that escape between times τ_k and τ_{k+1}. Let $Y_1 = \prod_{j=2}^L (1 - V_j)$.

Finally, define random variables $Z_1, \dots, Z_n$ to be conditionally independent given $(Y_k)_{k=1}^M$ such that for $i = 1, \dots, n$ and $j = 1, \dots, M$, we have $P(Z_i = j | (Y_k)_{k=1}^M) = Y_j$. Here the integers i such that $Z_i = k$ correspond to lineages that recombine when there are k members of the B population with an infinite line of descent.

Then define Π such that i and j are in the same block if and only if $Z_i = Z_j$. It is shown in [11] that for all partitions π of $\{1, \dots, n\}$, we have

$$|P(\Theta = \pi) - P(\Pi = \pi)| \leq \frac{C}{(\log N)^2}.$$

Simulation results in [3] show that this approximation is very accurate. Recently, Etheridge, Pfaffelhuber, and Wakolbinger [5] have shown that many aspects of this analysis carry over to the case of weak selection, where s and r are both $O(1/N)$ and the process can be described by a diffusion limit.

We also consider in [4] how the genealogy is affected by recurrent selective sweeps. Because, under strong selection, the duration of a selective sweep is short, many ancestral lines can coalesce almost instantaneously at the time of the selective sweep. If selective sweeps happen at times of a Poisson process, as proposed by Gillespie [8], then as $N \to \infty$ the genealogy converges, under suitable conditions, to a coalescent process called a coalescent with multiple collisions. Coalescents with multiple collisions, introduced in [9, 10], are coalescent processes such that whenever there are b clusters, each possible merger of k clusters into one happens at rate $\lambda_{b,k}$, where $\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k} \, \Lambda(dx)$ for some finite measure Λ on $[0, 1]$. Kingman's coalescent, which describes the genealogy of the sample in the absence of selective sweeps, is the special case in which Λ is a unit mass at zero.

Using properties of coalescents with multiple collisions, we obtain approximations for how recurrent selective sweeps would affect test statistics that are used to detect departures from Kingman's coalescent. These statistics include Tajima's D-statistic [12], which compares the number of "segregating sites" at which not all n sequences in the sample agree to the average number of pairwise differences over the $\binom{n}{2}$ pairs of sequences in the sample, and Fu and Li's D-statistic [7], which considers the number of mutations that affect just a single lineage. One can show [4] that for large samples, Tajima's D-statistic should be negative when there are recurrent selective sweeps, as has been observed repeatedly in simulations (see e.g. [2]), and that Tajima's D-statistic should have more power to detect selective sweeps than Fu and Li's D-statistic. However, much work remains in this area. In particular, it is an open problem to get a handle on how multiple mergers of ancestral lines affect the full site frequency spectrum, that is, for each k, the number of mutations that affect exactly k lineages. Such information would be needed to analyze, for example, the H-statistic of Fay and Wu [6].

References

[1] N. H. Barton, *The effect of hitch-hiking on neutral genealogies*, Genet. Res. Camb. **72** (1998), 123–133.

[2] J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan, *The hitchhiking effect on the site frequency spectrum of DNA polymorphisms*, Genetics, **140** (1995), 783-796.

[3] R. Durrett and J. Schweinsberg, *Approximating selective sweeps*, Theor. Pop. Biol. **66** (2004), 129–138.

[4] R. Durrett and J. Schweinsberg, *A coalescent model for the effect of advantageous mutations on the genealogy of a population*, Ann. Appl. Probab. **15** (2005), 1591–1651.

[5] A. M. Etheridge, P. Pfaffelhuber, and A. Wakolbinger, *An approximate sampling formula under genetic hitchhiking*, Preprint, available at http://front.math.ucdavis.edu/ math.PR/0503485.

[6] J. C. Fay and C.-I Wu, *Hitchhiking under positive Darwinian selection*, Genetics **155** (2000), 1405-1413.

[7] Y. X. Fu and W. H. Li, *Statistical tests of neutrality of mutations*, Genetics **133** (1993), 693-709.

[8] J. H. Gillespie, *Genetic drift in an infinite population: the pseudohitchhiking model*, Genetics **123** (2000), 887–899.

[9] J. Pitman, *Coalescents with multiple collisions*, Ann. Probab. **27** (1999), 1870–1902.

[10] S. Sagitov, *The general coalescent with asynchronous mergers of ancestral lines*, J. Appl. Probab. **36** (1999), 1116–1125.

[11] J. Schweinsberg and R. Durrett, *Random partitions approximating the coalescence of lineages during a selective sweep*, Preprint, available http://front.math.ucdavis.edu/math.PR/ 0411069.

[12] F. Tajima, *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism*, Genetics, **123** (1989), 585-595.

On transition functions with Dirichlet and Poisson-Dirichlet stationary distributions

Dario Spanò

(joint work with Robert C. Griffiths)

The dynamic of the allele frequencies in a neutral, d-allele model with infinite population size and $2 \leq d \leq \infty$, is described by a Markov process in $(d-1)$-dimensional simplex with transition density given by

$$(1) \qquad p_t^{(\theta)}(x,dy) = d\pi_\theta(y)\{1 + \sum_{n=1}^{\infty} \rho_n^{(|\theta|)}(t) Q_n^{(\theta)}(x,y)\}, \qquad (t \geq 0)$$

where for $\theta \in \mathbb{R}_+^d$: (i) $d\pi_\theta(y)$ is the stationary measure of the process i.e. the Dirichlet distribution on the or its Poisson-Dirichlet limit if $d = \infty$; (ii) $Q_n^{(\theta)}(x,y) = \sum_{|\mathbf{n}|=n} P_{\mathbf{n}}^\theta(x) P_{\mathbf{n}}^\theta(y)$ is determined by any choice of (non-unique) multivariate polynomials $\{P_{\mathbf{n}}^\theta(x)\}$ orthonormal with respect to $d\pi_\theta$, and (iii)

$$(2) \qquad \rho_n^{|\theta|}(t) = e^{-\frac{1}{2}tn(n+|\theta|-1)}$$

is the only quantity depending on the time-parameter t (see Griffiths (1979)).

A natural question to ask is whether $\{\rho_n^{(|\theta|)}\}_{n\geq 1}(t)$ is the only possible choice of coefficients making the expansion (1) the transition function of a Markov process with Dirichlet or Poisson-Dirichlet stationary measure. A crucial, preliminary step in this direction is to characterize all possible nonzero sequences $a_n(\theta)$ which, if replaced to $\rho_n^{|\theta|}(t)$ in (1), make the function $p_t^{(\theta)}(x, dy)$ non-negative. This is the topic of the present discussion.

Bochner (1954) provides an answer to the problem for $d = 2$ and $\theta = (\alpha, \alpha)$ with $\alpha > 1/2$. In this case, $Q_n^\theta(x, y) = P_n^\alpha(x)P_n^\alpha(y)$, where $P_n^\alpha(x)$ are modified Gegenbauer Polynomials, orthonormal with respect to the beta(α, α) distribution. The function $p_t^{(\theta)}(x, dy)$ turns out to be non-negative if and only if

$$a_n(\theta) = \int_0^1 \frac{P_n^\alpha(x)}{P_n^\alpha(1)} dH(x) \tag{3}$$

for a positive, sigma-finite measure dH. Such result is extended by Gasper (1972) to all $\theta \in \Theta^* := \{(\alpha_1\alpha_2) : 1/2 < \alpha_1 \leq \alpha_2 \textit{ or } \alpha_1 + \alpha_2 > 2\}$: a necessary and sufficient condition is an analogue of (3) with $\{P_n^\alpha(x)\}$ now replaced by a system $\{P_n^{(\alpha_1,\alpha_2)}(x)\}$ of modified Jacobi polynomials, orthonormal with respect to the beta(α_1, α_2) distribution. The proof of Bochner-Gasper's result relies on the following key property of Jacobi Polynomials for $\theta = (\alpha_1, \alpha_2) \in \Theta^*$:

$$\frac{P_n^\theta(x)P_n^\theta(y)}{P_n^\theta(1)P_n^\theta(1)} = \int_0^1 \frac{P_n^\theta(z)}{P_n^\theta(1)} dm_{(\theta;x,y)}(z), \tag{4}$$

where

$$dm_{(\theta;x,y)}(z) = \left(\sum_{n\geq 0} [P_n^\theta(1)]^{-1} P_n^\theta(x)P_n^\theta(y)P_n^\theta(z) \right) d\pi_\theta(z) \tag{5}$$

is a probability measure. No results are known for more general $\theta \in \mathbb{R}^2$ as the property (4) holds if and only if $\theta \in \Theta^*$.

We provide an extension of Bochner-Gasper's characterization for $d \geq 2$. Let $\Theta_d^* = \{\theta = (\alpha_1 \leq \ldots \leq \alpha_d) \in \mathbb{R}^d : \alpha_1 \geq 1/2 \textit{ or } 2 \leq \alpha_j + \alpha_{j+1},\ j = 1, \ldots, d-1\}$.

Theorem. For $\theta \in \Theta_d^*$,

$$\{1 + \sum_{n=1}^\infty a_n(\theta) Q_n^{(\theta)}(x, y)\} \geq 0$$

if and only if

$$a_n(\theta) = \int \frac{P_n^{(\alpha_1,|\theta|-\alpha_1)}(z)}{P_n^{(\alpha_1,|\theta|-\alpha_1)}(1)} dH(z) \tag{6}$$

for some positive measure dH on $[0, 1]$.

We also show that condition (6) also implies

$$a_n(\theta) = \int \frac{P_n^{(\alpha_i+\nu,|\theta|-\alpha_i-\nu)}(z)}{P_n^{(\alpha_i+\nu,|\theta|-\alpha_i-\nu)}(1)} dH'(z)$$

for $i = 1, \ldots, d$ and $0 \leq \nu \leq |\theta| - \alpha_i$.

The key for the proof of the theorem is a surprising extension of the product formula (4) to the d-dimensional kernel $Q_n^\theta(x, y)$: for $\theta \in \Theta_d^*$,

$$Q_n^{(\theta)}(x,y) = \left[P_n^{(\alpha_1,|\theta|-\alpha_1)}\right]^2 \int_0^1 \frac{P_n^{(\alpha_1,|\theta|-\alpha_1)}(z)}{P_n^{(\alpha_1,|\theta|-\alpha_1)}(1)} dw_{(x,y;\theta)} \tag{7}$$

with

$$dw_{(x,y;\theta)} = \prod_{j=1}^{d-1} dm_{(\phi_i(x),\phi_i(y);(\alpha_i,\alpha_{i+1}))},$$

where $dm_{(z,w;(\alpha_i,\alpha_{i+1}))}$ is defined by (5) and the functions $\phi_i(x) \in [0,1]$ have an explicit, constructive representation.

Unfortunately the restriction to $\theta \in \Theta_d^*$ is not irrelevant, as it forces the total mass $|\theta|$ of the parameter of $d\pi_\theta$ to be proportional to the dimension d of the type space. Therefore the property (7) is not helpful to characterize "good" sequences $a_n(\theta)$ in the infinitely-many-types case (apart from those limit processes with Poisson-Dirichlet stationary measure with parameter $\theta = \infty$). It is still unclear whether the condition (6) is necessary and sufficient for the positivity of $p_t(x, y)$ even when $\theta \in \mathbb{R}^d \setminus \Theta_d^*$. We finally show that, besides $\rho_n^\theta(t)$ as in (2), a nontrivial example of positivity, holding for general θ, is given when

$$a_n(\theta) = r^n \qquad |r| < 1.$$

However this result cannot be applied to prove (6) for general θ, as long as one cannot prove that it is valid even for $r = a + ib$.

References

[1] Bochner, S., *Positive zonal functions on spheres*, Proc. Nat. Acad. Sci. U.S.A. **40**, (1954). 1141-1147

[2] Gasper, G., *Banach algebras for Jacobi series and positivity of a kernel*, Ann. of Math. (2) **95** (1972), 261-280.

[3] Griffiths, R.C., *A transiton density expansion for a multi-allele diffusion model* , Adv. Appl. Prob. **11**, (1979), 2, 310-325.

On stability of the optimal filter for nonergodic signals

Wilhelm Stannat

A simple nonlinear problem in stochastic filtering theory in continuous time can be formulated as follows. Consider a system of two stochastic differential equations

$$dX_t = B(X_t)dt + dW_t \tag{1}$$

$$dY_t = GX_t dt + d\tilde{W}_t, \quad Y_0 = 0\,. \tag{2}$$

Here, $(W_t)_{t\geq 0}$ and $(\tilde{W}_t)_{t\geq 0}$ are independent Brownian motions on $\mathbb{R}^d$ and on $\mathbb{R}^p$, $B : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field and G is a $p \times d-$matrix.

Equation (1) models a stochastic signal process describing the state of a system, for example the position of some airplane. The state of the system cannot be observed directly but only through some measurement procedure that adds an additional error to the final observation $Y_\cdot$. Equation (2) is a simple model for this measurement process in the case where the measurement error is independent of the signal process.

In stochastic filtering one is now interested in calculating the optimal filter, that is, the conditional distribution

$$\eta_t(A) := E\left[1_A(X_t)|\mathcal{Y}_t\right]\,, A \in \mathcal{B}(\mathbb{R}^d)\,,$$

of the signal X_t given the information

$$\mathcal{Y}_t := \sigma\left(Y_s|s \in [0,t]\right)$$

provided by the observation up to time t. Note that η_t depends on the initial distribution of the signal, that is, on the distribution of X_0 (in the following denoted by μ_0), which is unknown, since we do not observe the signal directly. We are therefore interested in the dependence of η_t w.r.t. μ_0 and we will use in the following the notation

$$\eta_{t,\mu_0} \text{ and } E_{\mu_0}\left[1_A(X_t)|\mathcal{Y}_t\right]$$

to indicate explicitely the dependence on μ_0.

It is widely believed that if the signal process is ergodic, so that X_t forgets μ_0 for large t, the same is true for η_{t,μ_0}. Corresponding results have been obtained, in cases where the state space of the signal process is compact, by Kunita, Stettner, Ocone, Pardoux, da Prato, Malliavin, Fuhrmann, Zeitouni, Atar, Del Moral and Miclo (see [4] for references). However, it is already known from the linear case (Kalman-Bucy Filter) that η_{t,μ_0} may become independent of μ_0 also for nonergodic signals.

In the papers [2], [3], [4] a variational approach is introduced to study the long-time behaviour of η_{t,μ_0} and a clear explanation is given, why η_{t,μ_0} may become independent of μ_0 also for nonergodic signals. Moreover, explicit lower bounds on the rate of stability of η_{t,μ_0} w.r.t. its initial condition μ_0 are obtained.

The main tool is the analysis of the associated pathwise filter equation

$$\dot{\mu}_t^y = \hat{A}_t^y \mu_t^y + \sigma_t^y \mu_t^y - \int \sigma_t^y d\mu_t^y \cdot \mu_t^y \,. \tag{3}$$

Here, $y \in C([0,\infty[;\mathbb{R}^p)$ is a parameter, $\hat{A}_t^y$ denotes the dual of the differential operator

$$A_t^y f = Af - G^T y_t \cdot \nabla f \,,$$

where

$$Af = \frac{1}{2}\Delta f + B \cdot \nabla f$$

is the generator of the signal process, and

$$\sigma_t^y(x) = -G^T y_t \cdot B(x) + \frac{1}{2}\|G^T y_t\|^2 - \frac{1}{2}\|Gx\|^2 \,.$$

For the notion of a classical solution of equation (3) we refer to [4].

The solution of the pathwise filter equation provides - up to some density - a regular conditional probability of X_t given $\mathcal{Y}_t$. More precisely, in good cases (see [2], [3], [4] for precise statements) it follows that

$$\eta_{t,\mu_0}(A) = \frac{\int 1_A(X_t) e^{G_t^T Y_t \cdot x}\, \mu_t^Y(dx)}{\int e^{G_t^T Y_t \cdot x}\, \mu_t^Y(dx)} \qquad \text{a.s.} \tag{4}$$

where μ_t^Y is the solution of (3) with initial condition $\mu_0^Y = \mu_0$.

From now on suppose that $B = \frac{\nabla\varphi}{\varphi}$ for some strictly positive differentiable function φ. To state our main results we make the following assumptions:

Assumption 1 The potential

$$W(x) := \|Gx\|^2 + \frac{\Delta\varphi}{\varphi}(x)$$

is in $C_p^2(\mathbb{R}^d)$ and uniformly strictly convex:

$$\exists\, \kappa_* > 0 \text{ such that } W_{xx} \geq \kappa_*^2 \cdot I \,.$$

Assumption 2 $\varphi \in C_p^2(\mathbb{R}^d)$, φ bounded, and there exists a log-concave function $\varphi_0 \in C^2(\mathbb{R}^d)$ and some finite constant M such that

$$M^{-1}\varphi_0 \leq \varphi \leq M\varphi_0 \,.$$

Fix a uniformly strictly log-concave function $m_0 \in C_p^2(\mathbb{R}^d)$ such that

$$-(\log m_0)_{xx} \geq \kappa_* \cdot I \text{ and } \int \frac{|\nabla m_0|^2}{m_0}\, dx < \infty$$

and define the probability measure

$$\nu(dx) := \left(\int m_0 \varphi \, dx\right)^{-1} m_0(x)\varphi(x)\, dx \,.$$

Theorem 1. *Let $y \in C([0,\infty[,\mathbb{R}^p)$ and suppose that Assumptions 1& 2 hold. Let $g \in C_b^2(\mathbb{R}^d)$ be log-concave and $\mu_0^i \ll \nu$, $i = 1,2$, with densities $h_i \in C_p^2(\mathbb{R}^d)$ bounded from below and from above satisfying $\int \frac{|\nabla h_i|^2}{h_i} m_0 \, dx < \infty$. Let $\delta > 0$ be such that $\delta < h_i < \delta^{-1}$. Then there exist unique classical solutions $\mu_t^{i,y}$, $t \geq 0$, of (3) with initial condition μ_0^i, and*

$$\left\| \frac{g \, d\mu_t^{1,y}}{\int g \, d\mu_t^{1,y}} - \frac{g \, d\mu_t^{2,y}}{\int g \, d\mu_t^{2,y}} \right\|_{var} \leq \frac{2}{\delta^2} e^{-\frac{\kappa_*}{2M^4} t} \tag{5}$$

for any $t > 0$. In particular,

$$\limsup_{t\to\infty} e^{\frac{\kappa_*}{2M^4} t} \, \|\mu_t^{1,y} - \mu_t^{2,y}\|_{var} < \infty \, .$$

The proof of the theorem, as well as comparison with existing results, can be found in [4]. Time-dependent generalizations of the result can be found in [2] and [3]. For a discussion of the measure-valued semilinear equation (3) in the context of genetic algorithms, including nonlinear generalizations, see [1].

Corollary 2. *Let μ_0^i and $\mu_t^{i,y}$, $i = 1,2$, be as in the theorem. Assume that*

$$\eta_{t,\mu_0^i}^y(A) := \frac{\int 1_A(X_t) e^{G_t^T y_t \cdot x} \, \mu_t^{i,y}(dx)}{\int e^{G_t^T y_t \cdot x} \, \mu_t^{i,y}(dx)} \, , \qquad y \in C([0,\infty[,\mathbb{R}^p) \, ,$$

is a regular conditional distribution of X_t given $\mathcal{Y}_t$ w.r.t. $P_{\mu_0^i}$. Then

$$\limsup_{t\to\infty} e^{\frac{\kappa_*}{2M^4} t} \left\| \eta_{t,\mu_0^1}^Y - \eta_{t,\mu_0^2}^Y \right\|_{var} < \infty \text{ almost surely.}$$

The corollary provides a simple sufficient criterion for stability of the optimal filter w.r.t. its initial condition. Moreover, a lower bound on the exponential rate of stability is determined that mainly depends on the lowest eigenvalue of the Hessian of the potential $W(x) = \|Gx\|^2 + \frac{\Delta\varphi}{\varphi}(x)$. The more convex W, the higher the rate of stability. Note that W consists of two parts: the second part $\frac{\Delta\varphi}{\varphi}$ depends on the signal, whereas the first part $\|Gx\|^2$ depends on our choice G how to measure the signal. The more precise our measurement, the more convex $\|Gx\|^2$. Conversely, our criterion provides a priori lower bounds on our choice G to reach a certain exponential rate κ_*. Also note that ergodic and non-ergodic directions of the signal process can be "separated" in the criterion.

References

[1] W. Stannat, *On the convergence of genetic algorithms - a variational approach*, Probab. Theory Relat. Fields **129** (2004), 113–132.

[2] W. Stannat, *Stability of the pathwise filter equation for a time-dependent signal on R^d*, Appl. Math. Optim. **52** (2005), 39–71.

[3] W. Stannat, *Stability of the optimal filter via pointwise gradient estimates*, In: Stochastic Partial Differential Equations and Applications II, Editors: G. da

Prato et al., Lecture Notes in Pure and Applied Mathematics, Marcel Dekker, New York, 2005.

[4] W. Stannat, *Stability of the pathwise filter equation on R^d*, Bielefeld 2004, submitted.

Genetic hitchhiking and linkage disequilibrium

Wolfgang Stephan

1. Introduction

During the past 15 years, studies of genetic variation in Drosophila and several other genetically well-characterized species, including humans and mice, have focused on the detection of natural selection at the DNA sequence level by analyzing the relationship between patterns of variation and recombination rates. Most of these studies have found a strong positive correlation between the local recombination rate experienced by a gene and its level of nucleotide polymorphism (Aguadé et al. 1989, Stephan and Langley 1989), whereas divergence (to closely related species) was not correlated with recombination (Begun and Aquadro 1992). This observation was not consistent with the standard neutral model (i.e. constant mutation rate, constant population size; Kimura 1983). It led to an intensive search for alternative models invoking natural selection and/or demography. In particular, the so-called hitchhiking model (proposed by Maynard-Smith and Haigh in 1974) re-surfaced and played an important role in the following years until today.

The hitchhiking model considers the effect of rare, strongly advantageous substitutions on linked, neutral (or weakly selected) polymorphisms. It predicts a reduction of genetic variation near the target site of selection, an excess of rare polymorphisms, and high-frequency derived alleles. These properties can be used to detect "footprints" of selection in scans of variation along the genome described next.

2. Results

Genome scan: We measured nucleotide sequence polymorphism along the X and third chromosomes in two *Drosophila melanogaster* populations, an ancestral population from Africa and a derived population from Europe (Glinka et al. 2003, Ometto et al. 2005). This comparison allowed us to test whether frequent selective events occurred during the colonization of Europe after the last ice age (about 10,000 years ago), as would be expected if adaptation of fruit flies to novel environments leaves footprints of selection in the genome. Indeed, about 10% of the genomic regions surveyed in the European population showed a severe reduction of variation, as predicted by the hitchhiking model, whereas little evidence for selection was found in the African sample.

Distinguishing selection and demography: Patterns of variation caused by genetic hitchhiking, such as the reduction of variation, look similar to those caused by demographic processes (particularly population size bottlenecks). We have therefore developed statistical techniques to distinguish between them, using the coalescent and maximum likelihood approaches (Kim and Stephan 2002, Li and Stephan 2005, Ometto et al. 2005). This work is based on the idea that demography (such as bottlenecks) affects the entire genome, while selection acts locally on individual genes (see above). Using this strategy, we were able to identify candidate regions in the genome where selection is likely to have occurred in the recent past. Furthermore, we have begun to study these regions in detail to localize the targets of selection and find the genes and associated phenotypes involved in adaptation (Beisswanger et al. 2005).

The effects of hitchhiking on linkage disequilibrium (LD): We have analyzed the effect of hitchhiking on LD (nonrandom association between polymorphisms) using a three-locus model with two neutral sites and one selected one (Stephan et al. 2005). LD is measured between the neutral sites. Employing similar analytical approximations as Maynard-Smith and Haigh (1974), we found that hitchhiking increases LD in the first half of the selected phase; i.e., shortly after the occurrence of the advantageous mutation. However, LD is destroyed quickly before the selected mutation reaches fixation. As a consequence, LD around the target sites of selection in the genome should be reduced (even though variation may not be very low). We expect that this property may also be a useful signature of selection in the genome, similar to the footprint of reduced variation described above.

3. Acknowledgements

I am grateful to many colleagues who have contributed to this project, in particular S. Beisswanger, D. De Lorenzo, S. Glinka, J. Hermisson, S. Hutter, Y. Kim, C. Langley, H. Li, S. Mousset, L. Ometto, P. Pennings, P. Pfaffelhuber, Y. Song, and N. Svetec. This research was funded by the VolkswagenStiftung and the DFG.

References

Aguadé, M., N. Miyashita, and C. H. Langley, 1989. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics 122: 355-362.

Begun, D. J. and C. F. Aquadro, 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356: 519-520.

Beisswanger, S., W. Stephan, and D. De Lorenzo, 2005. Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. In press.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo, 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multilocus approach. Genetics 165: 1269-1278.

Kim, Y. and W. Stephan, 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-777.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Li, H. and W. Stephan, 2005. Maximum likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics 171: 377-384.

Maynard-Smith, J., and J. Haigh, 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23-35.

Ometto, L., S. Glinka, D. De Lorenzo, and W. Stephan, 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol. Biol. Evol. 22: 2119-2130.

Stephan, W. and C. H. Langley 1989. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. Genetics 121: 89-99.

Stephan, W., Y. S. Song, and C. H. Langley, 2005. The hitchhiking effect on linkage disequilibrium between linked neutral loci. In press.

Branching-coalescing particle systems

Jan Swart
(joint work with Siva R. Athreya)

Let Λ be a finite or countable set and let G be a group of bijections $g : \Lambda \to \Lambda$ that is transitive, i.e., $\forall i, j \in \Lambda\ \exists g \in G$ s.t. $gi = j$. Examples are $\Lambda = \mathbb{Z}^d$ or T_d (a regular tree), and G the group of translations on $\mathbb{Z}^d$ or automorphisms of the tree.

Consider a system of particles on Λ evolving according to the following laws. 1° Particles jump from i to j with rate $a(i,j)$. 2° Each particle splits with rate $b > 0$ into two new particles at the same site. 3° Each pair of particles on the same site coalesces with rate $2c > 0$ to one particle. 4° Each particle dies with rate $d \geq 0$. We make the following assumptions.

(1) The transition rates $a(i,j)$ are irreducible.
(2) $\sum_j a(i,j) = \sum_j a(j,i) < \infty$.
(3) $a(i,j) = a(gi, gj)\ \forall g \in G$.

Let $X_t(i)$ be the number of particles present at site $i \in \Lambda$ and time $t \geq 0$. We call $X = (X_t)_{t\geq 0}$ the (a,b,c,d)-braco-process.

Consider moreover a process where at each site $i \in \Lambda$ there lives a large fixed number of organisms, which can be of two genetic types: healthy and defective. Assume that: 1° With rate $a(i,j)$, an organism at site i migrates to site j. 2° Healthy organisms with rate b choose another organism, living on the same site, and replace it by a healthy organism. 3° Each pair of organisms living at the same site is resampled with rate $2c$, i.e., a random member of the pair is replaced by an organism with the type of the other member. 4° With rate d, a healthy organism mutates into a defective one. Then, in the limit that the number of organisms at each site is large, the frequencies $\mathcal{X}_t(i)$ of healthy organisms at site i and time t are described by the following system of SDE's:

$$\begin{aligned} \mathrm{d}\mathcal{X}_t(i) = &\sum_j a(j,i)(\mathcal{X}_t(j) - \mathcal{X}_t(i))\,\mathrm{d}t + b\mathcal{X}_t(i)(1-\mathcal{X}_t(i))\,\mathrm{d}t - d\mathcal{X}_t(i)\,\mathrm{d}t \\ &+ \sqrt{2c\mathcal{X}_t(i)(1-\mathcal{X}_t(i))}\,\mathrm{d}B_t(i). \end{aligned} \tag{1}$$

We call $\mathcal{X} = (\mathcal{X}_t)_{t\geq 0}$ the (a,b,c,d)-resem-process.

For any $\phi \in [0,1]^\Lambda$ and $x \in \mathbb{N}^\Lambda$, let $\mathrm{Thin}_\phi(x)$ denote a thinning of x with ϕ, i.e., for each i, the $x(i)$ particles at site i are independently kept with probability $\phi(i)$ and thrown away otherwise. Let $\mathrm{Pois}(\rho)$ denote a Poisson point measure with local intensity $\rho \in [0,\infty)^\Lambda$. The next theorem describes some important relations between braco-processes and resem-processes. Here, $X^\dagger$ and $\mathcal{X}^\dagger$ denote the $(a^\dagger, b, c, d)$-braco-process and $(a^\dagger, b, c, d)$-resem-process, respectively, where $a^\dagger(i,j) := a(j,i)$ are the reversed jump rates.

Theorem 1. (Dualities and Poissonization)

(a) (Duality) $P^x[\mathrm{Thin}_\phi(X_t) = 0] = P^\phi[\mathrm{Thin}_{\mathcal{X}^\dagger_t}(x) = 0]$.

(b) (Self-duality) $P^\phi[\mathrm{Pois}(\frac{b}{c}\mathcal{X}_t\psi) = 0] = P^\psi[\mathrm{Pois}(\frac{b}{c}\phi\mathcal{X}^\dagger_t) = 0]$.

(c) (Poissonization) $\mathcal{L}(X_0) = \mathcal{L}(\mathrm{Pois}(\frac{b}{c}\mathcal{X}_0))$ *implies* $\mathcal{L}(X_t) = \mathcal{L}(\mathrm{Pois}(\frac{b}{c}\mathcal{X}_t))$.

The duality (a) goes back to [5]. It has an interpretation in terms of potential healthy ancestors which is due to [4].

It has been shown in [2] that branching-coalescing particle systems can be started in infinity. In fact:

Theorem 2. *Let* $X^{(n)}_0 = x^{(n)}$ *with* $x^{(n)}(i) \uparrow \infty\ \forall i \in \Lambda$. *Then*

(a) *The processes* $X^{(n)}$ *may be coupled such that* $X^{(n)}_t \uparrow X^{(\infty)}_t$ *for all* $t > 0$.

(b)

$$E[X^{(\infty)}_t(i)] \leq \begin{cases} \frac{r}{c(1-e^{-rt})} & \text{if } r \neq 0, \\ \frac{1}{ct} & \text{if } r = 0 \end{cases} \qquad \text{wherer } r := b - d + c.$$

(c) *The law of* $X^{(\infty)}$ *decreases stochastically to a limit*

$$\mathcal{L}(X^{(\infty)}_t) \downarrow \mathcal{L}(X^{(\infty)}_\infty) =: \overline{\nu} \quad \text{as } t \uparrow \infty,$$

called the upper invariant law.

(d) $\overline{\nu}$ *is uniquely characterised by*

$$P[\mathrm{Thin}_\phi(X^{(\infty)}_\infty) = 0] = P^\phi[\mathcal{X}^\dagger_t = 0 \text{ for some } t \geq 0].$$

The explicit bound in part (b) is new compared to [2]. Resampling-selection processes have an upper invariant law too:

Theorem 3. *If* $\mathcal{X}^1_0 = 1$, *then:*

(a) $\mathcal{X}^1_t$ *decreases stochastically to a limit*

$$\mathcal{L}(\mathcal{X}^1_t) \downarrow \mathcal{L}(\mathcal{X}^1_\infty) =: \overline{\mu} \quad \text{as } t \uparrow \infty \tag{2}$$

called the upper invariant law.

(b) $\overline{\mu}$ *is uniquely characterised by*

$$P[\mathrm{Thin}_{\mathcal{X}^1_\infty}(x) = 0] = P^x[X^\dagger_t = 0 \text{ for some } t \geq 0]. \tag{3}$$

(c) *If* Λ *is infinite, then* $\mathcal{L}(X^{(\infty)}_\infty) = \mathcal{L}(\mathrm{Pois}(\frac{b}{c}\mathcal{X}^1_\infty))$.

Theorem 3 (c) says that $\overline{\nu}$ is a Poissonization of $\overline{\mu}$. We mention the following open problem: is each invariant law of X the Poissonization of an invariant law of $\mathcal{X}$? On certain lattices, such as trees, there is probably a multitude of invariant laws, so this question is nontrivial.

Say that the (a, b, c, d)-braco-process *survives* if $P[X_t \neq 0 \ \forall t \geq 0] > 0$ for some finite nonzero initial state X_0. Survival of (a, b, c, d)-resem-processes is defined similarly. It has been shown in [5] that (a, b, c, d)-braco-processes on $\mathbb{Z}^d$ survive if b is sufficiently large. By comparison with critical branching, it is easy to see that (a, b, c, d)-braco-processes die out if $b \leq d$.

Say that a probability law is *homogeneous* if it is invariant under translations with the group G, and *nontrivial* if it gives zero probability to the zero configuration. Since the transition laws are invariant under G, the upper invariant measures are obviously homogeneous. By Theorem 2 (d) and Theorem 3 (b) and (c), it is easy to prove that the following statements are equivalent:

- The (a, b, c, d)-braco-process survives.
- The (a, b, c, d)-resem-process survives.
- The $(a^\dagger, b, c, d)$-braco-process has a nontrivial invariant law.
- The $(a^\dagger, b, c, d)$-resem-process has a nontrivial invariant law.

But here is an open question: does survival of the (a, b, c, d)-braco-process imply survival of the $(a^\dagger, b, c, d)$-braco-process? This is obvious if a and $a^\dagger$ are isomorphic, as is the case if the lattice is an abelian group such as $\mathbb{Z}^d$, but in general the question is nontrivial.

Using duality, one can prove the following result about convergence to the upper invariant law.

Theorem 4.

(a) *If $\mathcal{L}(X_0)$ is homogeneous and nontrivial, then $\mathcal{L}(X_t) \underset{t\to\infty}{\Longrightarrow} \overline{\nu}$.*

(b) *If $\mathcal{L}(\mathcal{X}_0)$ is homogeneous and nontrivial, then $\mathcal{L}(\mathcal{X}_t) \underset{t\to\infty}{\Longrightarrow} \overline{\mu}$.*

Part (b) has been proved before in [5]. Once can show that part (a) implies part (b) by Poissonization, but not vice versa. The analogue of Theorem 4 for the contact process has been proved long ago by Harris in [3]. All these proofs follow the same scheme. The main ingredient in the proof of Theorem 4 (a) is the next lemma, that says that the dual $(a^\dagger, b, c, d)$-resem-process exhibits 'extinction versus unbounded growth'. Here, for any $\phi \in [0, 1]^\Lambda$, we write $|\phi| := \sum_i \phi(i)$.

Lemma 5. *If $|\mathcal{X}_0^\dagger| < \infty$, then*

$$P\big[\mathcal{X}_t^\dagger = 0 \quad \textit{for some } t \geq 0 \quad \textit{or} \quad \lim_{t\to\infty} |\mathcal{X}_t^\dagger| = \infty\big] = 1.$$

In fact, it turns out that $e^{-\frac{b}{c}|\mathcal{X}_t|}$ is a submartingale, so $|\mathcal{X}_t|$ has a random limit by submartingale convergence. All the hard work is in showing that the limit is $\{0, \infty\}$-valued, and that if the limit is zero, then $|\mathcal{X}_t|$ becomes zero in finite time.

One moreover needs the following somewhat technical fact.

Lemma 6. *Let $\mathcal{L}(X_0)$ be homogeneous and nonrivial, and $t > 0$. Then*

$$\lim_{n\to\infty} P\big[\mathrm{Thin}_{\phi_n}(X_t) = 0\big] = 0$$

for all $\phi_n \in [0,1]^\Lambda$ satisfying $|\phi_n| \to \infty$.

Using these lemmas and Theorem 2 (d), Theorem 4 (a) can be proved in one line. One has

$$\begin{aligned}\lim_{t\to\infty} P[\mathrm{Thin}_\phi(X_t) = 0] &= \lim_{t\to\infty} P[\mathrm{Thin}_{\mathcal{X}^\dagger_{t-1}}(X_1) = 0]\\ &= P[\exists t \geq 0 \text{ such that } \mathcal{X}^\dagger_t = 0] = P[\mathrm{Thin}_\phi(X^{(\infty)}_\infty) = 0].\end{aligned} \tag{4}$$

Since this holds for all $\phi \in [0,1]^\Lambda$, it follows that $\mathcal{L}(X_t) \Rightarrow \overline{\nu}$.

References

[1] S.R. Athreya and J.M. Swart, *Branching-coalescing particle systems*, Probab. Theory Relat. Fields **131(3)** (2005) 376–414.

[2] W. Ding, R. Durrett, and T.M. Liggett, *Ergodicity of reversible reaction diffusion processes*, Probab. Theory Relat. Fields **85(1)** (1990) 13–26.

[3] T.E. Harris, *On a class of set-valued Markov processes*, Ann. Probab. **4** (1976) 175–194.

[4] S.M. Krone and C. Neuhauser, *Ancestral processes with selection*, Theor. Popul. Biol. **51(3)** (1997) 210–237.

[5] T. Shiga and K. Uchiyama, *Stationary states and their stability of the stepping stone model involving mutation and selection*, Probab. Theory Relat. Fields **73** (1986) 87–117.

Gene genealogy when the distribution of offspring number among individuals is highly skewed

John Wakeley
(joint work with Bjarki Eldon)

We consider a continous-time limit process for the ancestry of a sample of genetic data in the case where the variance of offspring number is very large. The limit process is a special case of the general coalescent with multiple mergers introduced by Sagitov and Pitman. The process differs from the usual coalescent of Kingman in which only binary mergers can occur. The appearance of multiple mergers dramatically changes the predictions of the model regarding patterns of genetic variation, relative to the predictions of Kingman's coalescent. For example, as multiple mergers become more frequent, a greater fraction of the genealogy is in the external branches of the tree. The coalescent with multiple mergers also occurs on a faster time scale than Kingman's coalescent, the result being, that the mutation parameter scales differently with the population size than in Kingman's coalescent, where it scales linearly with the population size. We use the fact that the model predicts an excess of singleton polymorphisms (again relative to Kingman's coalescent) to make inferences about the distribution of offspring number from genetic data from the Pacific oyster.

Reporter: Ellen Baake

Participants

Roland Alkemper
alkemper@mathematik.uni-mainz.de
Institut für Mathematik
Universität Mainz
Staudingerweg 9
55099 Mainz

Prof. Dr. Ellen Baake
ebaake@techfak.uni-bielefeld.de
Technische Fakultät
Universität Bielefeld
Postfach 100131
33501 Bielefeld

Prof. Dr. Michael Baake
mbaake@math.uni-bielefeld.de
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld

Dr. Nathanael Berestycki
berestyc@math.cornell.edu
Pacific Institute for the
Mathematical Sciences
University of British Columbia
1933 West Mall
Vancouver, BC V6T 1Z2
CANADA

Matthias Birkner
birkner@wias-berlin.de
WIAS
Mohrenstr. 39
10117 Berlin

Prof. Dr. Reinhard Bürger
reinhard.buerger@univie.ac.at
Fakultät für Mathematik
Universität Wien
Nordbergstr. 15
A-1090 Wien

Dr. Nicolas Champagnat
champagn@clipper.ens.fr
WIAS
Mohrenstr. 39
10117 Berlin

Charles Cuthbertson
cuthbert@stats.ox.ac.uk
Department of Statistics
University of Oxford
1 South Parks Road
GB-Oxford OX1 3TG

Andrej Depperschmidt
depperschmidt@math.tu-berlin.de
Fachbereich Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin

Prof. Dr. Alison M. Etheridge
etheridg@stats.ox.ac.uk
etheridge@stats.ox.ac.uk
Magdalen College
Oxford University
GB-Oxford OX1 4AU

Prof. Dr. Steven N. Evans
evans@stat.berkeley.edu
Department of Statistics
University of California
367 Evans Hall
Berkeley, CA 94720-3860
USA

Warren J. Ewens
wewens@sas.upenn.edu
Department of Biology
University of Pennsylvania
221 Leidy Laboratories
Philadelphia PA 19104
USA

Prof. Dr. Shui Feng
shuifeng@mcmaster.ca
Department of Mathematics and
Statistics
Mc Master University
1280 Main Street West
Hamilton, Ont. L8S 4K1
CANADA

Prof. Dr. Robert Charles Griffiths
griff@stats.ox.ac.uk
Department of Statistics
University of Oxford
1 South Parks Road
GB-Oxford OX1 3TG

Prof. Dr. Richard Hudson
rr-hudson@uchicago.edu
Department of Ecology & Evolution
University of Chicago
1101 E. 57th St.
Chicago, IL 60637
USA

Martin Hutzenthaler
hutzenth@math.uni-frankfurt.de
Institut für Stochastik und
Mathematische Informatik
Johann Wolfgang Goethe-Universität
Robert-Mayer-Str. 10
60325 Frankfurt

Prof. Dr. Steve Krone
krone@uidaho.edu
Department of Mathematics
University of Idaho
P.O.Box 441103
Moscow ID 83844-1103
USA

Dr. Dirk Metzler
metzler@informatik.uni-frankfurt.de
Institut für Stochastik und
Mathematische Informatik
Johann Wolfgang Goethe-Universität
Robert-Mayer-Str. 10
60325 Frankfurt

Dr. Martin Möhle
martin.moehle@uni-tuebingen.de
Mathematisches Institut
Universität Tübingen
Auf der Morgenstelle 10
72076 Tübingen

Paul Munday
paul.munday@lincoln.oxford.ac.uk
Department of Statistics
University of Oxford
1 South Parks Road
GB-Oxford OX1 3TG

Pleuni Pennings
pennings@zi.biologie.uni-muenchen.de
Universität München
Department Biologie II
Großhardener Str. 2
82152 Planegg-Martinsried

Dr. Peter Pfaffelhuber
pfaffelhuber@zi-biologie.uni-muenchen.de
Department Biologie II
Universität München
Großhadernerstr.2
82152 Planegg

Kristan Schneider
a9900273@unet.univie.at
kristan.schneider@univie.ac.at
Fakultät für Mathematik
Universität Wien
Nordbergstr. 15
A-1090 Wien

Prof. Dr. Jason Schweinsberg
jschwein@math.ucsd.edu
Dept. of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0112
USA

Dr. Dario Spano
spano@stats.ox.ac.uk
Department of Statistics
University of Oxford
1 South Parks Road
GB-Oxford OX1 3TG

Dr. Wilhelm Stannat
stannat@mathematik.uni-bielefeld.de
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld

Prof. Dr. Wolfgang Stephan
stephan@zi.biologie.uni-muenchen.de
Department Biologie II
Universität München
Großhadernerstr. 2
82152 Planegg

Dr. Jan M. Swart
swart.@utia.cas.cz
UTIA AV CR
Pod vodarenskou vezi 4
18208 Praha 8
Czech Republik

Prof. Dr. John Wakeley
wakeley@fas.harvard.edu
Dept. of Organismic & Evolutionary
Biology, Harvard University
2102 Biological Laboratories
16 Divinity Avenue
Cambridge MA 02138
USA

Prof. Dr. Anton Wakolbinger
wakolbin@math.uni-frankfurt.de
Institut für Stochastik und
Mathematische Informatik
Johann Wolfgang Goethe-Universität
Robert-Mayer-Str. 10
60325 Frankfurt