Modelling the genetics of spatially structured populations

Alison Etheridge, University of Oxford

The last century has seen remarkable developments both in the nature and scale of genetic data, and in the tools available with which to interrogate them. However, fundamental questions remain unresolved. The genetic composition of a population can be changed by natural selection, mutation, mating, and other genetic, ecological and evolutionary mechanisms. How do they interact, and what was their relative importance in shaping the patterns that we see today?

The genetic history of a species is encoded in the patterns of genetic variation determined by mutations laid down on the genealogical trees that describe the ancestral relationships between genes sampled from the population. In order to compare with data, models for the (forwards in time) dynamics of the population must come hand in hand with consistent (backwards in time) models for the genealogical trees. To see how this works, consider the simplest possible model of inheritance, in which a population of fixed size, N, evolves in discrete generations. Each individual is assumed to have exactly one parent, chosen uniformly at random from the previous generation. Offspring inherit the genetic type of their parent. This model is known as the Wright-Fisher model. If we accept this model, then we can easily describe the ancestral relationships between individuals in a sample from the population. For example, if we sample just two individuals, then the chance that they have a common parent in the previous generation is just $1/N$; if they do not, then the chance that they have a common grandparent is $1/N$; and so on. In other words, the number of generations until we reach a common ancestor has a geometric distribution with parameter $1/N$. Since this has mean N, we immediately see that the appropriate timescale on which to view evolution of the population is in units of N generations. If we assume that N is large, then, in these units, the time to the most recent common ancestor of two individuals is (approximately) distributed as an exponential random variable with parameter one. More generally, for larger samples, since the chance of three or more lineages merging in a single generation is order $1/N^2$, with high probability we will only see pairwise mergers, and the family tree relating individuals has a simple and elegant description in terms of independent exponentially distributed random variables. This is the process we call Kingman's coalescent.

No real population evolves according to the Wright-Fisher model. Nonetheless, the Kingman coalescent has found almost unreasonable success in modelling the genealogical relationships between individuals in a vast array of biological populations. The twist is that in order to relate the model to data, we must revert to time units of single generations and it turns out that to do this we must use not the census population size, but instead an 'effective' population size. This single parameter is somehow able to capture the effects of natural selection, spatial structure, fluctuating population size and so on, that are absent from the Wright-Fisher model. That this works (at least if we sample individuals from far enough apart) is all the more astonishing as the difference between census and effective population sizes can be many orders of magnitude. For example, the human census population size is 7 billion, but the effective population size is around ten thousand.

The Kingman coalescent is just one facet of the rich mathematical structure that emerges if one views biological populations on successively larger spatial scales. In order to investigate the patterns of variation at smaller scales, we need models that explicitly incorporate spatial structure. One successful approach assumes that the population is subdivided into colonies. However, many

populations are not subdivided, but rather evolve in a spatial continuum. In the 1940's Malécot and Wright considered populations distributed across $\mathbb{R}^1$ or $\mathbb{R}^2$. We follow Malécot. The population evolves in discrete generations. Each individual, independently, produces a random number of offspring with mean one. Their spatial locations are randomly distributed about the location of the parent (according to a Gaussian distribution). Malécot started the process from a spatially homogeneous random configuration, believing that the configuration at any later time would still be spatially homogeneous. This allowed him to find an expression for the probability that two individuals sampled at a given separation are of the same genetic type. It was not until 1975, that Felsenstein, in his famous paper *'a pain in the torus'*, pointed out that the assumption of spatial homogeneity is inconsistent; in $\mathbb{R}^1$ or $\mathbb{R}^2$, a population evolving according to Malécot's dynamics will not be homogeneous, but instead will either die out or develop clumps of arbitrary density and extent.

To prevent clumping, population density must be controlled through local rules. One can achieve this by basing reproduction events on regions of space, rather than on individuals. When a region is affected by an event, an individual (the parent) is chosen at random from those within the affected region; each individual within the region, independently, dies with a probability u; and finally a random number of offspring, with mean λu times the area of the region, is thrown down uniformly across the affected region. If λ is large enough, there is a nontrivial, spatially homogeneous, stationary distribution. The backwards in time process describing genealogies is unwieldy, but can be approximated by a much simpler process, obtained by taking $\lambda \to \infty$. The forwards in time process also converges as $\lambda \to \infty$ to a limit which is known as the spatial Lambda Fleming-Viot process (SLFV), [1]. The pain in the torus is overcome and we have a consistent description of the genealogical trees relating individuals in a sample from the population.

Like the Wright-Fisher model, the SLFV is really a framework for modelling. The mathematical structure is surprisingly rich. Within it we find as scaling limits the Brownian net, branching Brownian motion, superprocesses, various stochastic partial differential equations, the celebrated Fisher-KPP equation, and much more besides. (See [2, 3, 4] for some examples.) Crucially, because it provides a way to combine genetic drift with spatial structure, it provides a tractable setting in which to investigate the interplay between spatial structure and other forces of evolution.

References

[1] N H Barton, A M Etheridge, and A Véber. Modelling evolution in a spatial continuum. *J. Stat. Mech.*, PO1002, 2013.

[2] J A Chetwynd-Diggle and A M Etheridge. SuperBrownian motion and the spatial Lambda-Fleming-Viot process. *Electron. J. Probab.*, 23(71):1–36, 2018.

[3] A M Etheridge, N Freeman, and S Penington. Branching Brownian motion, mean curvature flow and the motion of hybrid zones. *Elect. J. Probab.*, 22(paper 103):40pp, 2017.

[4] A M Etheridge, N Freeman, and D Straulino. The Brownian net and selection in the Spatial Λ-Fleming-Viot process. *Elect. J. Probab.*, 22(paper 39):36pp, 2017.

[5] J Felsenstein. A pain in the torus: some difficulties with the model of isolation by distance. *Amer. Nat.*, 109:359–368, 1975.

[6] G Malécot. *Les Mathématiques de l'hérédité.* Masson et Cie, Paris, 1948.