Perspectives

Anecdotal, Historical and Critical Commentaries on Genetics

Edited by James F. Crow and William F. Dove

Origins of the Coalescent: 1974–1982

J. F. C. Kingman

University of Bristol, Bristol BS8 1TH, United Kingdom

THE circle of ideas that has come to be known as the coalescent has proved to be a useful tool in a range of genetical problems, both in modeling biological phenomena and in making statistical sense of the rich data now available. In the 20 years since the concept crystallized, it has been extended in a number of directions. It is not the purpose of this note to document recent developments or to record the way in which others later arrived at similar conclusions by different routes; for that, consult, for instance, Hudson (1990), Donnelly and Tavaré (1995), and Stephens and Donnelly (2000).

The technique has been widely applied in recent studies of evolution, thanks to advances in molecular biology and computer technology. By being sample based, it has provided rigorous statistical analyses of population data and has provided a rationale for designing simulations. It has led to two different estimators of the key parameter, $\theta = 4N\mu$, where N is the effective population number and μ is the mutation rate, and, therefore, to a test of neutrality. It has provided estimates of the time to a common ancestor, and, in particular, a very long time estimate provided strong evidence for balancing selection in the ancestry of the HLA and other loci. This technique has also provided estimates of recombination and rate of selfing. It has been helpful in assessing migration patterns in human ancestry, in particular, sex differences as revealed by comparison of within- and between-group variability of Y chromosome and mitochondrial DNA. For a recent review, see Fu and Li (1999).

I shall not discuss these applications either. Mine is the much simpler aim of describing the way in which the ideas first came together, in the period leading to my 1982 articles. This is inevitably a personal account, but one that I hope is accurate, being based on records from these years. I have had the benefit of comments from Warren Ewens and Peter Donnelly, for which I am most grateful, but the interpretations and emphases are mine.

Three insights, in combination, comprise the essential basis of the coalescent. The first is the idea of tracing the ancestry of a gene backward in time and building up the family tree of the genes (at a particular locus) in a population sample back to the point at which they have a single common ancestor. This is just a generalization of Malécot's "identity by descent" (Nagylaki 1989) to more than two genes. It becomes powerful because of the second insight, that for a large class of demographic models, characterized by selective neutrality and constrained population size, the stochastic structure of the genealogy does not depend on the detail of the reproductive mechanism. Finally, for such models the effect of mutation is statistically independent of the genealogy.

What is surprising is that these rather simple ideas took so long to emerge. For me the story begins in 1974, when I was traveling in Australia meeting mathematicians who shared my interests in random processes and their applications. I had not worked on genetics since, as a Cambridge undergraduate, I had published juvenilia on polymorphisms maintained by single locus selection. But in Melbourne I encountered Warren Ewens, who was exploring some ideas of Ohta and Kimura (1973) on neutral evolution in finite populations (the "charge-state" model). Moving on to Canberra, I found that Pat Moran was working on a similar problem (Moran 1975), and the enthusiasm of these two Australians inspired their English visitor.

They considered a locus at which the different alleles can be labeled by a single numerical quantity and in which mutation causes a random addition or subtraction. Thus, a single line of descent will have genes that perform a random walk on the line. There was little biological credibility in such a description, but it accorded with the experimental techniques of gel electrophoresis, which were then the best way of distinguishing alleles (*e.g.*, Singh *et al.* 1976). A haploid population of fixed size n was supposed to evolve in discrete generations, the numbers of children of the members of one

Address for correspondence: University of Bristol, Senate House, Tyndall Ave., Bristol, BS8 1TH, United Kingdom.

generation having a symmetric multinomial joint distribution (the Wright-Fisher model).

Thus, the genes of one generation are represented by N points on the line. As we observe from generation to generation, the N points perform a "coherent random walk." The group strays farther and farther from its starting point, but the extent of the group remains relatively limited, and the distribution of the *relative* positions of the points converges to a proper limit. In my 1976 article I quoted Ewens as explaining this phenomenon by noting that "the probability that two points of G_t (the tth generation) have a common ancestor in G_s is $1 - (1 - N^{-1})^{t-s}$, which is near unity when $(t - s)$ is large compared with N. Thus, the whole of G_t is descended from a common ancestor in $G_{t-\Delta}$, where the random integer $\Delta = \Delta(t)$ remains stochastically bounded as $t \to \infty$. The relative distances are the result, therefore, only of displacements in these Δ generations."

This is tantalizingly close to the idea of deducing the genetic structure of the population from the genealogy back to the common ancestor, but the article then goes off into complicated mathematical analysis, which adds little to our understanding of the model. There are, however, two pointers to later work. First, the algebra is such that it forces consideration of samples of size n from the population and produces a recursion between n and $n + 1$. This led me to an interest in the Ewens sampling formula (Ewens 1972), which had already begun to take a central place on the population genetics stage.

The second pointer was the use of Fourier transforms, which made easy a generalization from a gene as a single number to one described by a family of d numbers. This led Kingman (1977a) to a theory of coherent random walks in space of d dimensions, at the price of no extra algebra. I think that I felt that escaping from the restriction to one dimension could lead to more realistic models, and this was confirmed when it became clear that all the formulae were simplified when d was allowed to tend to infinity. This corresponds to an assumption that a mutation always produces a new allele, the "infinite alleles" hypothesis (see Crow 1989).

It would not have been difficult to use the machinery of this article (Kingman 1977a) to derive the Ewens sampling formula, but only a special case was carried through. In Kingman (1977b), a different approach to that formula was introduced. Watterson (1974, 1976) had derived the sampling formula from the assumption that the population gene frequencies had a Dirichlet joint distribution, an assumption derived from the diffusion theory approach of the Kimura school, as well as earlier results of Wright. These frequencies, when arranged in descending order, have a limiting joint distribution known as the Poisson-Dirichlet limit, which I happen to have come across in a quite different connection (Kingman 1975). Watterson shows precisely that a population whose gene frequencies have this joint distribution will satisfy the Ewens sampling formula. The converse is proved in Kingman (1977b), and this is linked in Kingman (1978) to the consistency of the Ewens formula between different sample sizes.

Thus, by the end of 1978, the nature of the Ewens sampling formula and its links with, on the one hand, nonrecurrent mutation and, on the other, the classical Kimura diffusion approach to neutral evolution were well understood. Moreover, I had noted in Kingman (1977a) that the results were robust, in that they held when the Wright-Fisher multinomial model was replaced by other symmetric reproductive processes. What was still missing was the crucial connection with the genealogy.

As a result not only of this work but also of research into deterministic selective models, I was invited to be the speaker at a conference held at Iowa State University in June 1979 under the auspices of the Conference Board of the Mathematical Sciences and the National Science Foundation. The proceedings of that conference were published as Kingman (1980), and only one of its four chapters is devoted to neutral evolution. This adds little to the earlier articles, and for present purposes the most interesting feature is an annex, Appendix II, entitled "The genealogy of the Wright-Fisher model." This takes a rather subtle inequality, used in Kingman (1976) to prove a convergence result, and gives it a probabilistic meaning in terms of the random variable called $\Delta(t)$ above. It shows, in fact, that the probability that $\Delta(t)$ is greater than any integer r is at most $3(1 - N^{-1})^r$, the constant 3 being the best possible. But despite the title, there is no exploration of the family tree beyond the number of generations back to the common ancestor.

Our host at Iowa State, Oscar Kempthorne, had gathered an impressive group of participants, both mathematicians and biologists, and we discussed the problems of population genetics far into the night. It does not appear that the structure of the family tree entered into these debates, but it must have been there that the crucial idea was conceived, because the first account of the coalescent appears in Kingman (1982a), submitted for publication less than a year after the conference, in May 1980.

A footnote on the first page of that article observes that "genealogy means the whole family tree structure," so the cat is out of the bag. The argument starts from the observation that the Wright-Fisher multinomial model is equivalent to the rule that each member of a generation chooses its mother at random from the previous generation, the choices of different members being independent. This means that two members of the same generation have a probability $(1 - N^{-1})^r$ of having different ancestors r generations back. If time is measured in units of N generations and $N \to \infty$, the time to a common ancestor for the two has a negative exponential

distribution with probability density e^{-t}. Now consider n members of a particular generation, and trace their family tree backward through time. For some time there will be n ancestors, but at some instant two of the lines come together. The probability density of this coalescence time (in the limit as $N \to \infty$) is ke^{-kt}, where $k = n(n-1)/2$ is the number of pairs that might coalesce. Now trace back the $n-1$ lines until they coalesce; the argument is the same with n replaced by $n-1$ and so on, until the number of lines is reduced to 1. The article sets this up formally, by means of a Markov chain whose states are equivalence relations on $\{1, 2, \ldots, n\}$, and relates it to a representation of the Ewens sampling formula in terms of a certain "random paintbox." Thus the circle of ideas is complete.

Kingman (1982b) covers much the same ground in a more mathematical way, but a more important article is Kingman (1982c). This proves strong robustness results and goes a long way to explaining why the coalescent is useful for a wide range of neutral models. It also shows directly how, by allowing mutation to act on the branches of the family tree, the Ewens formula follows.

There is a moral to this tale. The first articles on coherent random walks (Ewens 1974; Moran 1975; Kingman 1976) were bogged down in complex algebra. If we had asked what the equations meant in probabilistic terms, we could not have missed the significance of the family tree or the simplification that comes if mutation is nonrecurrent (or if the mutant is independent of the parent). Those who analyze stochastic models should always lift their eyes from their equations to ask what they actually mean.

LITERATURE CITED

Crow, J. F., 1989 Twenty five years ago in genetics: the infinite allele model. Genetics **121:** 93–96.

Donnelly, P. J., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

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

Ewens, W. J., 1974 A note on the sampling theory for infinite alleles and infinite sites models. Theor. Popul. Biol. **6:** 143–148.

Fu, Y.-X., and W.-H. Li, 1999 Coalescing into the 21st century: an overview and prospects of coalescent theory. Theor. Popul. Biol. **56:** 1–10.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Kingman, J. F. C., 1975 Random discrete distributions. J. R. Stat. Soc. B **37:** 1–22.

Kingman, J. F. C., 1976 Coherent random walks arising in some genetical models. Proc. R. Soc. Lond. Ser. A **351:** 19–31.

Kingman, J. F. C., 1977a A note on multi-dimensional models of neutral mutation. Theor. Popul. Biol. **11:** 285–290.

Kingman, J. F. C., 1977b The population structure associated with the Ewens sampling formula. Theor. Popul. Biol. **11:** 274–283.

Kingman, J. F. C., 1978 Random partitions in population genetics. Proc. R. Soc. Lond. Ser. A **361:** 1–20.

Kingman, J. F. C., 1980 *Mathematics of Genetic Diversity*. SIAM, Philadelphia.

Kingman, J. F. C., 1982a On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kingman, J. F. C., 1982b The coalescent. Stochastic Process. Appl. **13:** 235–248.

Kingman, J. F. C., 1982c Exchangeability and the evolution of large populations pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino. North-Holland, Amsterdam.

Moran, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

Nagylaki, T., 1989 Gustave Malécot and the transition from classical to modern population genetics. Genetics **121:** 103–118.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Singh, R. S., R. C. Lewontin and A. A. Felton, 1976 Genetic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in *Drosophila pseudoobscura*. Genetics **84:** 609–629.

Stephens, M., and P. J. Donnelly, 2000 Inference in molecular population genetics. J. R. Stat. Soc. B **62:** 1–31.

Watterson, G. A., 1974 The sampling theory of selectively neutral alleles. Adv. Appl. Probab. **6:** 463–488.

Watterson, G. A., 1976 The stationary distribution of the infinitely-many neutral alleles diffusion model. J. Appl. Probab. **13:** 639–651.