Richard Hudson and Norman Kaplan on the coalescent process

Many important questions in genetics involve looking back in time using data sampled in the present. The coalescent process describes the ancestry of a sample of genes: as lineages trace back, they coalesce at a rate inversely proportional to the effective population size. This remarkably simple approximation now dominates population genetics, both because it gives a direct intuition into the evolutionary process, and because it allows efficient simulation.

The coalescent predicts the genealogical relationships between sampled genes, and it depends only on the *effective population size*, regardless of the detailed life history. The coalescent was first described by Kingman (1982), but was developed independently by Hudson (1983) and Tajima (1983); its influence on population genetics came primarily through Hudson. The coalescent is rooted in older concepts: Malecot's (1948) idea of *identity by descent,* which is central to quantitative genetics (Kempthorne, 1954), and the diffusion approximation (Kimura, 1955), which depends on the same effective population size. The coalescent emerged not through radically new concepts, but rather, because it gives a natural way to analyse the samples of DNA sequences that were just becoming available. This paper, with its companion (Kaplan *et al*., 1988), extended the coalescent to include selection and used it to interpret data on sequence variation around the alcohol dehydrogenase (*Adh*) locus of *Drosophila melanogaster* (Kreitman 1983).

Under the coalescent, lineages coalesce at a rate equal to the inverse of the effective number of genes in the population, $1/2N_e$. Migration, recombination, and mutation can be included by allowing ancestral lineages to jump between locations, genetic backgrounds, or allelic states; this extension is known as the *structured coalescent* (Hudson 1983). Selection is much harder to incorporate, because the ancestry now depends on the allelic state. However, Kaplan *et al.* (1988) showed that if the selected backgrounds are taken as given, then the structured coalescent describes the genealogy of linked neutral alleles; random fluctuations in the frequency of the selected backgrounds can be treated by a diffusion that couples to the coalescent. A surprising conclusion from this method is that selection has to be extremely strong relative to random drift in order to distort neutral genealogies (Barton and Etheridge, 2004). This is a fundamental obstacle to detecting selection from sequence data.

The fast and slow (F/S) alleles of the *Adh* locus of *D. melanogaster* differ by a single amino acid. Kreitman (1983) sequenced 11 copies of the locus and found a sharp peak of polymorphism around the amino-acid difference. This is consistent with maintenance of these alleles by long-term balancing selection. Hudson and Kaplan (1988) showed that divergence between the F and S alleles is consistent with balancing selection, albeit with a lower than average recombination rate. However, there was also excess variation *amongst* the S alleles, which is not expected. Even though *Adh* in *Drosophila* is one of the most intensively studied polymorphisms, we still do not know how its sequence variation has been shaped by selection (Begun et al., 1999).

Population genetics is now focused on understanding the abundance of sequence data that has recently become available. Since the first work of Kreitman (1983), the goal has been to infer the nature and strength of selection across the genome directly from the DNA sequence; 'genome scans' of the kind introduced by Hudson and Kaplan (1988) are now being carried out for an enormous range of organisms. However, it is disconcerting that even the first and best-studied example is still unresolved.

Further Reading

Kaplan N. L., Darden T., Hudson R. R., 1988 The coalescent process in models with selection. Genetics **120**: 819–829.

Kingman J. F. C., 2000 Origins of the Coalescent: 1974–1982. Genetics **156**: 1–3.

Nagylaki T., 1989 Gustave Malecot and the transition from classical to modern population genetics. Genetics **122**: 253–268.

Literature cited

Barton N. H., Etheridge A. M., 2004 The effect of selection on genealogies. Genetics **166**: 1115–1131.

Begun D. J., Betancourt A. J., Langley C. H., Stephan W., 1999 Is the fast/slow allozyme variation at the Adh locus of *Drosophila melanogaste*r an ancient balanced polymorphism? Mol.Biol.Evol. **16**: 1816–1819.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. 23:183-201.

Hudson R. R., 1991 Gene genealogies and the coalescent process. Oxford Surveys in Evol. Biol. **7**: 1–44.

Kempthorne O., 1954 The correlation between relatives in a random mating population. Proc. R. Soc. Lond. Ser. B-Biol. Sci. **143**: 102–113.

Kimura M., 1955 Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. **20**: 33–55.

Kingman J. F. C., The coalescent. Stochastic Precesses and Application **13**: 235–248.

Kreitman M., 1983 Nucleotide polymorphisms at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304**: 412–417.

Laurie C. C., Bridgham J. T., Madhusudan C., 1991 Associations between DNA sequance variation and variation in expression of the Adh gene in natural populations of D melanogaster. Genetics **129**: 489–499.

Malécot G., 1948 *Les mathématiques de l'hérédité*. Masson et Cie. Paris.

Neuhauser C., Krone S. M., 1997 The genealogy of samples in models with selection. Genetics **145**: 519–534.

Tajima F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437–460.