Stochastic Processes in Evolutionary and Disease Genetics

Ellen Baake (Universität Bielefeld),
Don Dawson (Carleton University, Ottawa),
Warren Ewens (University of Pennsylvania, Philadelphia),
Bruce Rannala (University of Alberta, Edmonton)

August 7–12, 2004

1 General overview of the field and its developments

The broad subject area of the meeting was that of *mathematical population genetics*. It is concerned with the analysis of the generation, nature, and maintenance of genetic variation within and between biological populations. In its evolutionary aspects it describes the change in the genetic composition of populations under the influence of various evolutionary forces, the most important of which are mutation, selection, recombination, migration and random genetic drift. The latter is a consequence of the fact that even without fitness differences, some individuals may, just by chance, have more offspring than others, so that the offspring of one genotype may displace another one in a finite population. Thus there is a significant element of randomness in genetic systems. From the point of view of disease genetics, many diseases are caused by deleterious mutant genes, and the analysis of the variation in a population for the disease and the normal gene is a significant component of this area of research.

These two components of the theory have hitherto been somewhat separate. However, recent trends in evolutionary genetics theory have brought them together, and one of the aims of this meeting was to further this fusion of two important areas of population genetics.

Three new developments are shaping the area at present: a change in biological thinking, the emergence of new data, and new mathematic(ian)s; these are, of course, all interrelated. Let us explain this in some more detail.

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 is the one with the most clearly defined mathematical models. The evolutionary behavior of a population may be described by a stochastic model of gene frequency change, which is similar to corresponding models in interacting particle systems. These models are well understood if mutation and drift are the only forces present, or if selection is also present but acts on the genes at one or a small number of gene loci. In particular, the pattern of genetic variation generated under these scenarios is quite well known.

But for several decades, the area suffered from a lack of data to support - or reject - the hypotheses about the evolutionary process and the genetic basis of various diseases. It was therefore often criticized as "l'art pour l'art". This situation has suddenly been reversed due to the wealth of molecular data flowing in during the past few years. The data come from the various genome projects, and from studies aimed at the genetic basis of human diseases. The most valuable data

derive from samples from a population (population sequence data), as opposed to single individuals. For the first time in the history of the field, the theory is now lagging behind the data, and the lack of analytical results translates into a lack of statistical methods for data analysis. The immediate need of data evaluation methods is often satisfied by heuristic techniques of a preliminary nature. But in the long run, there is a real need for methods which rest on a solid foundation with respect to the underlying genetic stochastic processes.

Evolutionary genetics theory has thus moved in large part to an analysis of the corresponding inverse problem, namely the reconstruction of evolutionary history from the observed patterns of genetic variation. A particularly challenging problem is the detection of selection at the molecular level. Selective forces are not easy to analyze since their effects must be discerned against a background of stochastic effects. Thus the analysis of the genetic data used to assess the effects of selection presents particularly difficult statistical problems, which have no entirely satisfactory answer even today.

The theoretical foundations of such analyses have been laid by the change in direction in population genetics from the classical prospective theory, considering the evolution of a population forwards in time, to the retrospective theory, which considers the past history of the currently-observed population. Mathematically, the backward view corresponds to the dual of the forward process. Coalescent theory, the most frequently used area of the retrospective theory, is concerned specifically with properties of the ancestry of a sample of genes as they trace back to a common ancestor. If, for example, a disease mutation occurs only once, two or more disease genes in a contemporary sample have an ancestry that traces back to a most recent common ancestor disease gene.

Problems beyond those listed above are far more complex, have not been solved, and will require significant mathematical analysis for their resolution. This is particularly so if processes like recombination are included, or if selection acts in a complicated way. These are exactly the problems encountered in disease genetics. Significant properties of the disease locus itself are in practice often unknown, including its location - indeed a major aim of the theory is to attempt to locate it. Inferences about its location are made by using genes at known marker loci. This leads to the problem that the coalescent process of the disease gene is different from that of the markers, because of recombination between disease and marker loci. The situation is further complicated by the fact that diseases are often polygenic (and possibly under selection, which may be of complicated type due to interaction between loci). Such diseases are called complex diseases, and their study forms the center of current genetic investigation.

Another important development is the increasing interest of the mathematical community in theoretical biology in general, and genetics in particular. Many professional probabilists have recently moved into the area, with powerful modern methods at their fingertips, which has helped to turn the mathematics of biological evolution into an active and growing field. This has resulted in a productive interplay in which the problems of population biology have stimulated new mathematics which in turn has provided powerful new analytical tools to address the emerging problems of biological evolution. In particular there have been important developments in the theory of Markov processes stimulated by population biology - these include the introduction of important families of interacting particle systems and the class of measure-valued Fleming-Viot processes including the infinitely many types and infinitely many sites processes that now play an important role in population genetics. A number of effective mathematical tools for the analysis of these systems have been developed. Other mathematical tools will be described below, where appropriate.

After this general overview, let us now give a more detailed description of the various matters described above, as discussed at the meeting. It had six main topics, each led by a key speaker: Particle systems (Rick Durrett), The Coalescent (John Wakeley), Evolutionary population genetics (John Gillespie), Branching processes (Peter Jagers), Human genetics (Robert Elston), and Haplotype blocks (Peter Donnelly). We will proceed from the more theoretical to the more applied, and put special emphasis on the connections between the topics.

2 Particle systems and coalescent process

A fundamental model class in population genetics is defined by the Moran model and its relatives (and the closely related Wright-Fisher model with its variants). This is best described as an interacting particle system with a fixed number of N individuals, each of which is assigned a type; individuals reproduce and mutate independently, in discrete or continuous time. Every time an individual reproduces, the offspring is assigned a type (according to a Markov chain that describes mutation), and replaces another individual that is randomly chosen to die, thus keeping population size constant. If N gets large, the system is described by a diffusion limit, known as the Fleming-Viot measure-valued process. A very general particle representation that also remains valid in this limit is provided by the so-called look-down process, which yields a joint representation of particles and their genealogies [4].

¿From an evolutionary perspective, it is of particular interest to consider these particle systems *backward* in time. Given a sample from the present population, one aims at finding its genealogy. Here, a reproduction event forward in time corresponds to merging of individual lineages to a common ancestor backward in time, that is, a coalescence event. Since its invention by Kingman [22], the coalescence process has revolutionized population genetic thinking and data analysis.

This coalescent process is tractable and has been much studied in the case of neutral evolution, that is, all types of individuals have the same reproduction rate (this is the 'vanilla-flavoured' coalescent). The emphasis of current research is on the extension of the underlying ideas and methods to more complex systems involving selection, recombination, migration, and variable population size.

The mathematical description of the process becomes a great challenge when types have different reproduction rates, that is, if selection is involved. This situation is particularly relevant for many questions in molecular evolution, in particular, when one wants to infer the (most likely) evolutionary history from a sample of individuals of a present-day population, and pinpoint selective events that have happened in the past.

One major step has been the construction of Neuhauser and Krone [24, 26], which uses, forward in time, two different reproduction events: definitive ones that will be used by every individual regardless of its type, and potential ones that may only be used by 'fit' individuals. Backward in time, this now induces a coalescing/branching structure, where the branching events correspond to unresolved birth events, meaning that the ancestry here may only be decided in a second step, when the types of the ancestors have been resolved. This process is rather complex, but some explicit results may be obtained, with considerable technical effort, about the time to the most recent common ancestor, for example.

The process becomes much simpler if, rather than full geneologies, only the ancestral lines of single individuals are considered. This seems to have been overlooked for quite some time; some explicit results (for two types) have recently appeared [13].

If more than two types are considered, explicit analytical results seem out of reach at present, and even simulation of the backward coalescent is a challenge. The 'first generation' simulation algorithms require sampling from the stationary distribution, which is, however, known only for the unrealistic case of parent-independent mutation. Recently, however, some progress could be made through *exact sampling* algorithms [10]. They do not require explicit knowledge of the stationary distribution and are, therefore, more generally applicable.

A second important direction concerns the inclusion of spatial structure into the Moran model and the resulting coalescent. A popular model here is the *stepping-stone model*, where individuals perform a random walk on a one- or two-dimensional lattice (or torus, in a mathematical idealization). Genetic structure may then be analyzed through the homozygosity as a function of the separation of the colonies, and a genetic distance known as FST (fixation index of subpopulations relative to the total population). Various limits, depending on the scaling of migration rate, subpopulation size and number of subpopulations, must be considered. Recent results include the logarithmic growth of FST with the number of colonies, the identification of parameter regimes where the stepping-stone model is effectively panmictic, the structure of genealogies, the effect of migration on the mutation patterns expected under the infinite-sites model, and the additional effect of selection [2, 5, 33, 32].

Another important line of development concerns the assumption of constant population size,

which is a severe limitation. Traditionally, variable population size has been treated with the help of the concept of *effective population size*; but a certain confusion has been associated with this notion. Recently, this has been analyzed within the framework of the coalescent [30]: Having identified conditions under which a model with stochastic demography converges to the coalescent with a linear change in time scale, Sjodin et al. [30] have argued that this is a necessary condition for the existence of a meaningful effective population size. Such a linear time scale change is obtained when demographic fluctuations and coalescence events occur on different time scales.

3 Branching processes

Branching processes have a long history in population genetics theory. They were first used by Wright to determine the fate (fixation or loss) of a rare mutant within a finite population (for review, see [9, p. 27ff]); for this purpose, a single-type Galton-Watson process is relevant. Recently, multi-type branching processes have been used in the context of mutation-selection models for large populations [18]; here, as in the coalescent process, the view back in time has become important, and earlier results by Jagers et al. [21] can now be used to investigate the relationship between the forward and the backward process, and the present and ancestral distribution of types, respectively.

But coalescent and branching processes have more in common than the backward view along single lines. Motivated by an earlier meeting on mathematical population genetics, Geiger [17] has recently started to investigate an analogue to the coalescent for branching processes. If k individuals are sampled uniformly at random from one generation of a large Galton-Watson population that has persisted for n generations, then the shape of the subtrees spanned by the sampled vertices and the root depends essentially on the tail of the offspring distribution: While in the finite variance case the subtrees are asymptotically binary (as $n \to \infty$), multiple branch-points do persist in the limit if the variance of the offspring distribution is infinite.

Apart from concrete questions like this one, branching processes and particle systems can also be subsumed under the general framework of particle systems and look-down processes mentioned earlier [4].

4 Evolutionary genetics

An important topic in modern evolutionary genetics is the identification of selective events in the history of a sample from the patterns of genetic variation observed in a present-day population. This is often done by means of the so-called *hitchhiking effect*, namely the fact that the fixation of a strongly selected beneficial mutation is accompanied by the increase of variants at other loci linked with the beneficial mutation. This effect leaves numerous signatures of diversity in DNA sequences, both within and between species, and affects the frequency spectrum of alleles, as well as linkage disequilibria and codon bias. Depending on whether there has been a single (recent) hitchhiking event or several repeated ones, the effects may be local or over a broader range. By comparing theoretical predictions with actual sequence data, one can infer the rate and strength of beneficial mutations in nature (among the many references available, see [23] for a recent example).

The hitchhiking effect has recently entered the level of large-scale analysis of SNP data. SNP's, '*single nucleotide polymorphisms*', are single nucleotide sites that are polymorphic in a population. Much effort is devoted to the problem of detecting selective sweeps using large SNP data sets from genomic scans. However, special care must be taken to overcome the ascertainment problem: Most population genetical methods do not correctly accommodate the special discovery process used to identify SNPs, which results in biased allele frequency distributions that must be corrected for [27].

Last but not least, our traditional understanding of the interplay of selection and genetic drift is challenged by the *pseudohithhiking model* proposed by Gillespie [19]. Strongly selected substitutions at one locus can induce stochastic dynamics that resemble genetic drift at a closely linked neutral locus. The pseudohitchhiking model is a one-locus model that approximates these effects and can be used to describe the major consequences of linked selection. The coalescent of the pseudohitchhiking model has a random number of branches at each node, which leads to a frequency spectrum that

is different from that of the equilibrium neutral model. If *genetic draft*, the name given to these induced stochastic effects, is a more important stochastic force than genetic drift, then a number of paradoxes that have plagued population genetics disappear – but, at the same time, the estimation procedures commonly employed in genetic analyses may be estimating parameters other than those that are assumed.

Apart from its impact on population genetics, this approach is also having a significant impact on mathematical research. Since the model relies on *strongly* selected mutants, the usual diffusion limit and associated coalescence theory is not applicable. Durrett and Schweinsberg [6] have approximated it with the help of random partitions created by a stick-breaking process, and Etheridge, Pfaffelhuber and Wakolbinger have modelled the ancestry at the neutral locus by means of a structured coalescent in a random background, and derived a corresponding sampling formula [8].

5 Recombination and haplotype blocks

Recombination is the formation of a chromosome passed on by a parent to an offspring by physical exchange between the two parental chromosomes, so that the transmitted chromosome consists of parts of each of the two parental chromosomes. There has been much recent speculation (based on patterns of genetic variation), and occasionally experimental confirmation (via sperm typing), that rates of recombination across the human genome vary on a fine scale. In particular, some regions of the genome appear to contain *recombination hotspots*, where recombination occurs at rates several times higher than the background average rate. Aside from inherent interest, an understanding of this local variation is essential for the appropriate design and analysis of many studies aimed at elucidating the genetic basis of common diseases or of human population histories. Standard pedigree-based approaches do not have the fine scale (< 0.1 cM) resolution that is needed to address this issue, because thousands of meioses are needed per recombination event. In contrast, samples of DNA sequences from unrelated chromosomes in the population carry relevant information, as there are a large number of meioses in the history of a sample of population data. But inference from such data is extremely challenging in several respects: the underlying stochastic model (the coalescent with recombination, a process that is practically intractable in the full-fledged version required here), the statistical analysis, and the computational requirements.

Although there has been much recent interest in the development of full likelihood inference methods for estimating local recombination rates from such data, they are not currently practicable for data sets of the size being generated by modern experimental techniques. Fearnhead and Donnelly [11, 12] introduced and studied two approximate likelihood methods. The first, a marginal likelihood, ignores some of the data. For larger sequences, they introduced a *composite likelihood*, which approximates the model of interest by ignoring certain long-range dependencies. With a combination of both methods, data from the lipoprotein lipase gene have been analyzed. A different approach was pursued by Li and Stephens [25], who have related the patterns of genetic variation to the underlying recombination process through their PAC model (product of approximate conditionals). This method has already been applied to two problems: determining whether a recombination hotspot identified in human males via sperm typing is also present in chimps; and quantifying the frequency of recombination hotspots in genes in the human genome.

Closely related to the local variation of recombination rates is the concept of haplotype blocks. A haplotype block is a region of a chromosome that tends to be passed on intact, without recombination, from parent to offspring. Partly as a result of this, the region of a chromosome corresponding to a block tends to exhibit only a few haplotypes in the entire population. Identification of haplotype blocks is a way of examining the extent of linkage disequilibrium in the genome, which generally provides useful information for the planning of association studies in human genetics (see the next Section). The aim is to identify a minimal subset of SNPs that can characterize the most common haplotypes. No uniform definition of a haplotype block has yet been agreed upon; however, various operational definitions are in use, see, e.g., Daly et al. [3]. The *Hap-Map project* (http://www.hapmap.org/index.html.en) describes haplotype blocks in the human genome. Particular interest is in the question whether there is similar haplotype block-structure between and within

populations (Nigeria/Yoruba, Asia, African Americans, Europeans), see, e.g., [16].

6 Human genetics

In human genetics, finding genes underlying heritable traits has been a long-standing problem. In recent years attention has shifted from 'Mendelian' disorders (that is, diseases caused by one defective gene, such as Huntington's disease or cystic fibrosis) to so-called complex traits, which are thought to be influenced by multiple genes possibly interacting with each other and with environmental risk factors. Many of these are common diseases, like diabetes.

As mentioned above, inference relies on the association with known marker genes, i.e., on linkage disequilibrium; this association is complete if there is no recombination between disease and marker locus, and decreases with distance (i.e. recombination rate) between them, thus giving a method of estimating this distance.

On a finer scale, the coalescent-based methods of the previous Section are the methods of choice; but for larger distances, pedigree-based methods are more appropriate. Here, one takes advantage of one basic difference between general population genetics and human genetics, which uses family (or pedigree) data rather than population data. Observations are made on a collection of markers (usually SNP's) transmitted from parents to affected (and sometimes unaffected) offspring, and as a result an assessment can be made about which SNP's the disease gene is close to. Since the locations of the SNP's are known, inferences can be made about the location of the disease gene. This area of research is known as *linkage analysis*, is based on probability models and parametric inference, and has a very long tradition; for review, see [28]. Over the years, each major development in parametric statistical inference has been adopted by the developers of linkage analysis methods, and questions of genetic analysis have prompted new statistical developments, from the work of Fisher [15] onwards. In many ways, statistical inference and genetic analysis have developed in parallel over the last 100 years.

Currently, the field is moving from a situation in which marker typing was hard and expensive to an era where this is relatively cheap, fast and easy, and the major cost of a study of a complex trait is now in the family collection and trait phenotyping. Recent progress in sequence analysis has made available the joint analysis of thousands and even hundreds of thousands of SNP's, thus making possible genome-wide screens for disease genes. Indeed, many researchers are already taking advantage of this fact. The statistical challenge is now how to deal with the vastly larger number of variables than observations: The enormous number of genotype configurations leads to a *curse of dimensionality*. This situation is analogous to that in microarray expression analysis, where expression levels of large numbers of genes are measured on a comparatively very small number of individuals. In both cases, false positives are the main problem. This is now an area of intensive statistical research; some recent approaches are discussed in [20].

References

[1] J.D. Balding, M. Bishop, C. Cannings (eds.), Handbook of Statistical Genetics, 2nd ed., Wiley, Chichester, 2003.

[2] T. Cox and R. Durrett, The Stepping Stone Model: New Formulas Expose Old Myths, *Ann. Appl. Prob.* **12** (2002), 1348-1377.

[3] M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander, High-resolution haplotype structure in the human genome, *Nature Genetics* **29** (2001), 229–232.

[4] P. Donnelly and T.G. Kurtz, Genealogical processes for Fleming-Viot models with selection and recombination, *Ann. Appl. Prob.* **9** (1999), 1091–1148.

[5] R. Durrett, Probability Models for DNA Sequence Evolution, Springer 2002.

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

[7] A. M. Etheridge, Introduction to Superprocesses, American Mathematical Society, Providence, RI, 2000.

[8] A.M. Etheridge, P. Pfaffelhuber, and A. Wakolbinger, An approximate sampling formula under genetic hitchhiking, submitted. math.PR/0503485

[9] W.J. Ewens, Mathematical Population Genetics, I. Theoretical Introduction, 2nd ed., Springer 2004.

[10] P. Fearnhead, Perfect simulation from population genetic models with selection, *Theor. Pop. Biol.* **59** (2001), 263–279.

[11] P. Fearnhead and P. Donnelly, Estimating recombination rates from population genetic data, Genetics **159** (2001), 1299-1318

[12] P. Fearnhead and P. Donnelly, Approximate likelihood methods for estimating local recombination rates, J. Roy. Stat. Soc. B. **64** (2002), 657-680.

[13] P. Fearnhead, The common ancestor at a non-neutral locus, *J. Appl. Prob.* **39** (2002), 38-54.

[14] P. Fearnhead, Ancestral processes for non-neutral models of complex diseases, *Theor. Pop. Biol.* **63** (2003), 115-130.

[15] R.A. Fisher, The systematic location of genes by means of crossover observations, *Amer. Nat.* **56** (1922), 406–411.

[16] S.B. Gabriel et al., The structure of haplotype blocks in the human genome. *Science* **296** (2002), 2225-2229.

[17] J. Geiger, The tree structure of a k-sample from a large Galton-Watson generation, in preparation.

[18] H.-O. Georgii and E. Baake, Supercritical multitype branching processes: the ancestral types of typical individuals, *Adv. Appl. Prob.* **35** (2003), 83–114; q-bio.PE/0311020.

[19] J.H. Gillespie, Genetic drift in an infinite population: The pseudohitchhiking model, *Genetics* **155** (2000), 909-919.

[20] J. Hoh and J. Ott, Mathematical multi-locus approaches to localizing complex human trait genes, *Nat. Rev. Genet.* **4** (2003), 701-709.

[21] P. Jagers, Stabilities and instabilities in population dynamics, *J. Appl. Prob.* **29** (1992), 770–780.

[22] J.F.C. Kingman, The coalescent, *Stoch. Proc. Appl.* **13** (1082), 235–248.

[23] Y. Kim and R. Nielsen, Linkage disequilibrium as a signature of selective sweeps, *Genetics* **167** (2004), 1513-1524.

[24] S.M. Krone and C. Neuhauser, Ancestral processes with selection, *Theor. Pop. Biol.* **51** (1997), 210-237.

[25] N. Li and M. Stephens, (2003) Modelling linkage disequilibrium and identifying recombination hotspots using single nucleotide polymorphisms data, Genetics **165** (2003), 2213-2233.

[26] C. Neuhauser and S.M. Krone, The genealogy of samples in models with selection, *Genetics* **145** (1997), 519–534.

[27] R. Nielsen et al., A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* (2005), in press.

[28] J. Ott, *Analysis of human genetic linkage*, 3rd ed., Johns Hopkins, 1999.

[29] J. Schweinsberg and R. Durrett, Random partitions approximating the coalescence of lineages during a selective sweep, submitted.

[30] P. Sjödin, I. Kai, S. Krone, M. Lascoux and M. Nordborg, On the meaning and existence of an effective population size, *Genetics* **165** (2005), 1061-1070.

[31] M. Stephens and P. Donnelly, Ancestral inference in population genetics models with selection, *Austral. New Zeal. J. Stat.* **51** (2003), 901–931,

[32] J. Wakeley and T. Takahashi, The many-demes limit for selection and drift in a subdivided population, *Theor. Pop. Biol.* **66** (2004), 83–95.

[33] I. Zähle, T. Cox and R. Durrett, The stepping stone model, II: Genealogies and the infinite sites model, submitted.