Abstracts

Hour Talks

Chip Aquadro

Locating the target of selective sweeps: theory and practice

Efforts to detect targets of recent positive selection ("selective sweeps") in the genome frequently involve an initial screen of a region of the genome at a low density of markers (either SNPs or microsatellites), followed by a denser screen in the region flanking markers that show reduced or skewed variation relative to that predicted by an equilibrium neutral model or relative to "background" variation in the genome. Direct sequencing is often then used to attempt to localize the gene/variation that was the target of the selective sweep. I will discuss how non-equilibrium demography seriously complicates the effort to initially detect departures due to natural selection, as well as subsequent efforts to localize the selective target. I will discuss a Goodness of Fit test that, when combined with Kim and Stephan's composite likelihood method for detecting and localizing sweeps, does reasonably well at distinguishing positive selection from many demographic perturbations. I will also provide examples from our relatively dense microsatellite screen of 800 kb of the X-chromosome in *Drosophila melanogaster* for both African and non-African population samples with follow-up sequencing that reveal departures from an equilibrium neutral model due apparently to positive selection, to demography alone, and to demographic amplification of an ancestral sweep. Finally, I will discuss our recent analysis of experimental strategies for localizing the targets of selective sweeps, and show that partial sequencing can lead to biased maximum likelihood estimates of selection parameters and reduced rejection rates. For common sample sizes and sampling strategies, the estimate of the target of selection has a very large confidence interval, and the strength of selection is often severely underestimated. A sequencing approach that leads to more accurate estimates will be discussed.

Steve Evans

A superprocess model for damage accumulation in fissioning organisms

One of the great challenges in biology is to understand the forces shaping life histories. Of particular interest is the apparent inevitability of decay and decline late in the organism's life. New technology has made it possible to study protozoans with previously unthinkable precision. Where microbes before could only be considered en masse, recent work has followed individual E. coli over many generations, following the fates of the ``old pole'' cell (the one that has inherited an end that has not been regenerated) and the ``new pole'' cell. Unpublished work from the same lab goes even further, by tracking the movement of damaged proteins through the generations, and showing that the growth of the population is maintained by a subpopulation of relatively undamaged cells. We analyze this phenomenon using ideas from Dawson-Watanabe

superprocesses and the theory of quasistationary distributions of killed diffusions. This is joint work with David Steinsaltz.

Susan Holmes

Studying the Immune System and Cancer with Multivariate Statistics and Microarrays

We will show how multivariate statistical methods can help explore the relationships between the immune system and cancer. We have analyzed data from microarray analyses of FACS sortedT cells using hierarchical clustering trees and network inference methods to pinpoint important genes for cancer diagnosis and treatment.

Paul Joyce

Statistical Inference for Population Genetics Models.

A large body of mathematical population genetics was developed by the three main speakers in this symposium. As a tribute to the substantial contributions of Ewens, Griffiths and Tavaré I will present an over view of some of my work, which builds upon their ideas. The focus will be on issues in the realm of Mathematical Statistics. The likelihood functions are based on the stationary distributions, under both infinite and K-alleles models, involving mutation, selection and genetic drift. The theoretical portion of the talk will consider limiting results that determine under what conditions models can be distinguished based on allele frequency data at a single locus. The computational portion of the talk will focus on new computationally efficient approaches to analyzing data under these models.

Steve Krone

Stochastic Demography, Coalescents, and Effective Population Size

In population genetics, deviations from the "usual assumptions" (e.g., constant population size, panmixia, etc.) sometimes have no measurable effect on polymorphism data; in other cases, the effects can be substantial. The key to understanding the difference can be found in relative time scales in the setting of the coalescent. We discuss conditions under which the genealogy for a model with stochastic demography or population structure converges to Kingman's coalescent with a linear change in time scale, and argue that this is a necessary condition for the existence of a meaningful effective population size. When these conditions are not met, the appropriate coalescent is shown to be a stochastic nonlinear time change of the standard coalescent. We illustrate these ideas with simulations of Fu and Li's F and Tajima's D under models of stochastically fluctuating population size and geographic structure. (Joint work with Ingemar Kaj, Magnus Nordborg, Martin Lascoux, Per Sjodin)

Vlada Limic

NK model

In 1987, Stuart Kauffman and Simon Levin introduced the NK model motivated by the problem of the evolution of DNA sequences. To each sequence of 0-1 bits of length N, they assigned a fitness as a sum of (random) quantities that depend only on bits observed in a sliding window of length K+1. The random map obtained in this way is called the fitness landscape. When $0 < K < N-1$, the fitness landscape is quite complicated and has many local maxima. Its properties have been extensively investigated by simulation but little is known rigorously. In joint works with Rick Durrett, and with Robin Pemantle we studied some qualitative and quantitative properties of the number of local maxima, their heights, and the height of the global maximum, in a fairly general setting. I will provide an introduction to the model, describe some of the rigorous results, and finally explain an interesting open problem.

Jason Schweinsberg

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

When a beneficial mutation occurs in a population, the new, favored allele may spread to the entire population, an event known as a selective sweep. As a result, when we sample individuals from a population and trace their ancestral lines backwards in time, many lineages may coalesce almost instantaneously at the time of a selective sweep. We discuss two approximations for the effect of a single selective sweep. The first is simple but not very accurate: flip coins with the same probability of heads to determine which lineages are descended from the one that had the beneficial mutation. The second approximation, which is based on approximating the population by a branching process, leads to very accurate results. We then consider the case when selective sweeps occur repeatedly throughout the history of the population and show that in this case the genealogy can be described by a coalescent process called a coalescent with multiple collisions. Finally, we show how this coalescent approximation can be used to get insight into tests that have been used to detect departures from the usual Kingman's coalescent.

Half Hour talks

Tibor Antal (Boston University Physics Department)

Fixation of strategies for an evolutionary game in finite populations

A stochastic evolutionary dynamics of two strategies given by 2 x 2 matrix games is studied in finite populations. We focus on stochastic properties of fixation: how a strategy

represented by a single individual wins over the entire population. The process is discussed in the framework of a random walk with site dependent hopping rates. The time of fixation is found to be identical for both strategies in any particular game and for any system size. The games can be classified according to whether the fixation probability or the fixation time exceeds the corresponding value for neutral games. Several statements and conjectures can be made. The asymptotic behavior of the fixation time and the fixation probabilities in the large population size limit can also be obtained directly from the discrete solution, without the use of a diffusion approximation.
(T. Antal, I. Scheuring: Fixation of strategies for an evolutionary game in finite populations. arXiv:q-bio.PE/0509008; to appear in Bulletin of Mathematical Biology.)

Julien Berestycki (Marseille)

Asymptotic formulae for the frequency spectrum of populations with heavy tailed offspring law

Suppose we sample n individuals from a population at a certain time. Due to mutations, at a given locus not all individuals in this sample will have the same allele. Moreover mutations also affect different sites. We may ask several questions. In the sample of size n how many different alleles should we observe at a given locus (site)? On how many sites should we expect to see different alleles? With which frequency should each of the different alleles be represented?

A fundamental result in this domain is the celebrated Ewens sampling formula. This result gives an explicit formula for the distribution of the allelic partition, under some standard assumptions on the reproduction mechanism of the population. More precisely the Ewens sampling formula is robust if the reproduction law has a second moment.

When this hypothesis is relaxed, the genealogy of the sample cease to be given by Kingman's coalescent and is better represented through the use of "multiple collisions" coalescents, or Lambda coalescents. By embedding a special family of Lambda coalescents in stable continuous random trees we can present here an asymptotic formula for the frequency spectrum (both in the infinite alleles and the infinite sites models), as the sample size n tends to infinity.

Nathanael Berestycki (UBC)

Global divergence of spatial coalescents.

Recently, Limic and Sturm (2006) introduced a class of processes which they called spatial coalescents and which generalizes the notion of a coalescent with multiple collisions to a setting where particles can only coalesce if they are on the same vertex of a given graph G and particles can travel on the graph according to some given transition probability. This corresponds to incorporating the effect of spatial structure and migration in the study of the genealogy of a population. We obtain various asymptotic results for

these processes, such as this one. On Z^d, if particles perform simple random walk and Kingman's coalescent, starting with N particles, at any given time the number of particles is typically log*(N). In general, for any graph and any coalescence mechanism, if started with infinitely many particles at a single site, the total number of particles remains infinite. Joint work with Omer Angel, Alan Hammond and Vlada Limic.

Adam Boyko (Cornell)

Quantifying the distribution of selective effects among newly arising mutations in the human genome

Here we report a novel approach for quantifying the distribution of selective effects of newly arising mutations based on SNP frequencies and the number of invariant sites and fixed differences against an outgroup. This likelihood method employs population genetic models that jointly handle natural selection and demography. Using 30000 coding SNPs identified by direct resequencing of 19 African- and 20 European-Americans, we investigate several models for the distribution of selective effects (Normal, Gamma, mixture models), the relationship between amino acid properties and exchangeability, and the degree to which deleterious and advantageous mutations influence human polymorphism and divergence.

Jake Byrnes, Vanja Dukic (Ecology and Evolution, U. of Chicago)

Hidden Markov Models for Detection of Gene Conversion Regions

Briefly, we are developing a hidden markov model to detect gene conversions between duplicate genes. The algorithm identifies regions of low divergence (the signature of conversion) between the aligned coding sequences of duplicates. We use MCMC with Metropolis-Hasting to explore the likelihood surface and then perform posterior decoding to determine the most likely location of the conversion. We have recently added a non-duplicated outgroup sequence to the alignment which allows us to distinguish domains of selective constraint from gene conversion.

John Chen (Bowling Green State)

On Hoppe's Urn and Ewens sampling formula

Consider the battle between human being and leukemia (or any other gene-related disease), if efforts (such as public intervention/inference) are made for the mutation towards the removal of the diseased-coded alleles, the genealogical process of the whole population can be viewed as a backward operation of Hoppe's urn model, which leads to Ewens sampling formula. In this talk, we will review conditions on Hoppe's urn model

and discuss alternatives to evaluate the joint probability of the numbers of different alleles in Ewens sampling formula.

Zachariah Dietz (Tulane)

Occupation laws for nonhomogeneous MC's and reinforcement schemes

We show that the induced proportional occupation measure of a non-stationaray discrete time Markov chain from a certain class of chains converges to a distribution defined through its moments to a measure absolutely continuous with respect to Lebesgue measure. This limiting measure may be dirichlet, and may intuitively be understood through classical GEM results.

Bjarki Eldon (Harvard)

Linkage disequilibrium between two loci under skewed offspring distribution among individuals

Expected linkage disequilibrium (LD) between two diallelic loci, as measured by the r^2 statistic, can be approximated in terms of the covariance in coalescence times for the two loci (McVean 2002). Eldon and Wakeley (2006) generalize the Moran and Wright-Fisher models of reproduction which result in positive probabilities of multiple coalescent events. We use these generalized neutral sampling processes to investigate effects of highly skewed offspring distribution among individuals on estimates of linkage disequilibrium (as measured by r^2) between two diallelic loci. In short, we investigate the scaling relationship between recombination and reproduction in a simple model of a population. A number of different limit processes are identified, which can give the same or very different predictions of LD. (Eldon and Wakeley (2006) Coalescent processes when the distribution of offspring number among individuals is highly skewed Genetics, in press)

Simona Grusea (Marseille)

Searching for conserved syntenies: mathematical model and statistical test.

We want to tell if a conserved synteny between two genomes (i.e. two genomic regions having in commun a certain number of orthologous genes, not considering the order of the genes) is really significant or it could have appeared by chance. We are taking into account the existence of gene families, i.e. the fact that we can't find in general a single ortholog for a given gene. We test the significance of clusters found not knowing a priori their size and how many orthologs they will contain. In the calculations we are using

limit theorems and gaussian approximations. We will try to have also results on real data by the time of the summer school.

Jeff Jensen (Cornell)

Inferring selection in the real world: demography, missing data, and the problem of outliers

One of the central goals of evolutionary biology is to understand the genetic basis of adaptive evolution. The availability of nearly complete genome sequence in a variety of organisms has facilitated the collection of data on naturally-occurring genetic variation on the scale of hundreds of loci to whole genomes. Such data have changed the focus of molecular population genetics from making inferences about adaptive evolution at single loci to attempting to identify which loci, out of hundreds to thousands, has recently been affected by natural selection. There are three major challenges to this effort: how to account for both non-equilibrium effects as well as ascertainment bias in interpreting outliers identified in these sub-genomic scans, how to efficiently sample identified loci in order to maximize power while minimizing cost, and finally how to distinguish the effects of selection from those of the demographic history of populations when evaluating sequence data spanning the localized region. To address issues of interpreting outliers, we simulate genomic scans and propose a corrected null-distribution that drastically improves Type I error of a widely-used test of selection. With regards to sampling schemes, we evaluate a wide range of sample sizes for a variety of strategies, varying from sparse to complete sequencing across a putative region, and propose a sampling scheme that, while no more costly in terms of either time or money than common approaches, is much more likely to accurately localize the target and strength of selection. And finally, we have modified a recently proposed composite likelihood ratio test, creating a goodness-of-fit statistic that is shown to have high sensitivity for differentiating between directional selection and demography.

Hye Won Kang (U. Wisconsin, Madison)

Modeling an infection cycle of the virusphage Q beta

A virusphage Q beta has an infection cycle: translation-replication-packaging-lysis. The number of viral genomes and viral proteins can be modeled with Markov processes. Since Markov processes are characterized by a martingale problem with generator, stochastic differential equations are obtained from generator. With a certain initial

condition, a limit of the scaled Markov processes converges weakly to the deterministic solution of the differential equations.

Amir R. Kermany (Department of Mathematics and Statistics, Concordia University)

Does the Fisher-Wright Model Apply to Stochastic Changes in Frequency of Historical Recombinations?

Recent examination of nucleotide polymorphisms, suggests a block-like configuration of alleles along the chromosome. These blocks are flanked by hotspots of recombination: regions for which higher probability of recombination can be inferred in the history of the sample. A haplotype block corresponds to a region of chromosome in the current population for which the allele configuration has been preserved since the time of the most recent common ancestor of the population.
Recombination hotspots can be formed either due to a variation in recombination rate along the chromosome, or due to the stochastic drift caused by a finite population size, or a combination of both factors. In this study, we provide a model, which describes the formation of blocks identical by descent as regions with no historical recombination.
By a proper definition of a recombination operator, the model enables us to keep track of historical recombination events in the chromosome. A simulation study shows that we detect a block-like structure in the population, which is merely due to the stochastic drift. In addition, in most of the cases, the blocks defined in this model coincide with the regions with high LD value.
The formalism of our model, suggests that the Fisher-Wright model (with a constant mutation rate) can be applied to calculate stochastic changes in the frequency of historical recombinations between two loci. Joint work with Donald Hickey (Department of Biology)

Amaury Lambert (Ecole Normale Superieure)

Discrete logistic branching populations and the canonical diffusion of adaptive dynamics.

Abstract. We start with the study of two competing subpopulations (resident and mutant) and find explicit second-order formulae for the probability of fixation of the mutant, also interpreted as the mutant's fitness, in the vicinity of neutrality. In particular, the second-order term is a linear combination of products of functions of the initial mutant frequency times functions of the initial total population size, called invasibility coefficients (fertility, defense, aggressiveness, isolation, survival). Then we apply a limit of rare mutations to a population subject to mutation, birth and competition where the number of coexisting types may fluctuate, while keeping the population size finite. This

leads to a jump process, the so-called `trait substitution sequence', where evolution proceeds by successive invasions and fixations of mutant types. Finally, we apply a limit of weak selection (small mutation steps) to this jump process, that leads to a diffusion process of evolution, called `canonical diffusion of adaptive dynamics', in which genetic drift (diffusive term) is combined with directional selection driven by the fitness gradient. (joint work with with N. Champagnat, Berlin)

Fernando Mendez (Ecology and Evolutionary Biology, Arizona)

Linear tests based on the frequency spectrum: One size does not fit all

The site frequency spectrum has been used to develop tests for demography and for selection. These tests, whose properties have been studied under different scenarios, cannot, simulatneoulsy, be sensitive, specific and test for different hypotheses. The best test will be determined by the hypotheses being tested. Here, I develop a framework for the search of the best linear tests. I also present some results on how this approach allows for an improvement in power.

Paul Munday (Oxford)

Importance sampling for inference on infectious disease transmission.

Given a configuration of infected individuals at the current time, and the relationships between them, can we use importance sampling to reconstruct the past history of disease transmission?

Raazesh Sainudiin (Oxford)

Exactly Approximate Bayesian Computations

When evaluation of the full likelihood function of parameter p in biologically realistic population genetic models from n observed DNA sequences D_o is computationally prohibitive, one may infer a population's history by approximating the posterior distribution $f(p \mid D_o)$, on the basis of summary statistics b_o of the full data D_o that are not necessarily sufficient for p. Such methods are known as approximate Bayesian computations (ABC).

In ABC, one first simulates data under p according to a prior distribution g, and summarizes it to b. Finally, p is accepted if $\| b - b_o \| < \varepsilon$, the acceptance-radius. Considerable effort is expended in finding the right epsilon that gives the optimal trade-off between efficiency and accuracy.

By concentrating first on a subset b' of b that can be expressed as linear combinations of

the site frequency spectrum (SFS) x, which is itself an (n-1)-dimensional summary of the full data D, we propose a modification to the standard ABC. The modification exploits the Markov basis methods of Algebraic Statistics intorduced by Diaconis and Sturmfels (Ann. Stat., 1998), to obtain the non-negative conditional SFS lattice that exactly satisfies the observed b'_o. The Markov basis ensures that any Markov chain run on our conditional SFS lattice with the base chain's transitions from the basis, are necessarily irreducible. Using the theoretical niceties of the infinitely many sites model of DNA mutation, we show that one can indeed make the acceptance-radius of the standard ABC schemes to be exactly zero and thus produce the exact conditional probability of interest $Pr(p \mid b'_o)$.

Koffi Sampson (Florida State U.)

Structured coalescent with nonconservative migration.

I consider the ancestral process of a sample from a subdivided population with stochastically varying subpopulation sizes and show that the corresponding finite-dimensional distributions converge to the structured coalescent.

Vishal Sood (Physics Dept., Boston U.)

Evolutionary Dynamics on Degree-Heterogeneous Graphs.

The evolution of two species with different fitness is investigated on degree-heterogeneous graphs. The population evolves either by one individual dying and being replaced by the offspring of a random neighbor (voter model (VM) dynamics) or by an individual giving birth to an offspring that takes over a random neighbor node (invasion process (IP) dynamics). The fixation probability for one species to take over a population of N individuals depends crucially on the dynamics and on the local environment. Starting with a single fitter mutant at a node of degree k, the fixation probability is proportional to k for VM dynamics and to 1/k for IP dynamics.

Dario Spano (Oxford)

Age-ordered frequencies, record indices and Gibbs random partitions

Brief description: Griffiths and Lessard (2005) provide a combinatorial derivation of the distribution of age-ordered frequencies in a sample (and in a population) of genes under the infinitely-many allele model with (possibly) variable population size. We discuss an extension of some of Griffitths and Lessards' results to Gibbs' family of random partitions.

Anja Sturm (Delaware)

On spatial Lambda coalescents

Lambda coalescents or coalescents with multiple mergers model the genealogies of populations with potentially large individual family sizes. We study some properties of Lambda coalescents for populations with a spatial substructure.