CONF - 9104300--Absts.
CONF - 9308268--Absts.
CONF - 9308107--Exc.

DE-FG03-91ER61165

Workshop in Computational Molecular Biology

4/15/91 - 4/14/94

Funds from this award were used to support two workshops and a special session at a conference:

- *Workshop in Computational Molecular Biology*, '91 Symposium on the Interface: Computing Science and Statistics. Seattle, Washington. April 21, 1991. 9104300--Absts
- *Workshop in Statistical Issues in Molecular Biology*, Stanford, California. August 8, 1993. 9308268--Absts.
- *Session on Population Genetics*, 56th Annual Meeting, Institute of Mathematical Statistics. San Francisco, California. August 9, 1993. 9308107--Exc.

The invited speakers, together with the titles of their talks, are listed below. Abstracts of each talk are also included.

Simon Tavaré

April 12, 1995

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Workshop in Computational Molecular Biology

'91 Symposium on the Interface: Computing Science and Statistics

Seattle, Washington. April 21, 1991.

9:00-9:05	S. Tavaré, University of Southern California. *Introduction*
9:05-9:20	D. Galas, Department of Energy. *Overview*
9:20-10:00	F. Cohen, UC San Francisco. *Computational aspects of the protein folding problem*
10:00-10:40	T. Schlick, NYU. *New computational techniques for computing biomolecular structures and their dynamics*
10:40-11:10	Coffee Break
11:10-11:50	E. Branscomb, Lawrence Livermore National Labs. *Building physical genome maps by random clone overlap; a progress assessment of work on human chromosome 19*
11:50-12:30	E.A. Thompson, University of Washington. *Monte Carlo methods for linkage analysis and complex models*
12:30-1:50	Lunch
1:50-2:30	E.S. Lander, Whitehead Institute. *Dissecting complex inheritance: statistical and computational issues*
2:30-3:10	E. Myers, University of Arizona. *Practical and theoretical advances in sequence comparison*
3:10-3:40	Coffee Break
3:40-4:20	R.J. Roberts, Cold Spring Harbor Labs. *Error detection in DNA sequences*
4:20-5:00	M.S. Waterman, University of Southern California. *Computer methods for locating kinetoplastid cryptogenes*

Computational aspects of the protein folding problem.

F. Cohen, Dept. of Pharmaceutical Chemistry, UCSF.

The amino acid sequence of a protein uniquely determines its tertiary structure. Deciphering this relationship, the protein folding problem, has become increasingly important to molecular biologists. DNA sequencing has become routine. Sequencing the human genome will produce a flood of protein primary structure data. Methods for determining the three dimensional structure of a protein cannot keep pace with genomic data. Computational procedures for relating sequence to structure are required. Two semi-empirical approaches to the folding problem have evolved: detailed energy calculations and the hierarchic condensation model. Energy calculations which attempt to describe the thermodynamic transition between unfolded and folded states have become hopelessly lost in multiple minima along the way. The hierarchic condensation model postulates that secondary structure is a useful computational intermediate in protein folding. I will discuss the hierarchic condensation model and the value of simplified representations of protein chains.

Secondary structure prediction. Pauling anticipated the locally regular organization of polypeptides into α-helices and β-sheets. X-ray crystallographic structure determination has confirmed that these periodic elements dominate protein architecture. Early attempts at relating amino acid sequence to secondary structure achieved 65% accuracy relying on exclusively local effects. Refinements of these methods have failed to improve their accuracy. Presumably, this is because a significant fraction of the local peptide conformations are influenced by non-local effects. We have exploited the non-local aspects of globular structure to improve secondary structure prediction. Initial work focused on loops, irregular segments of the chain which join secondary structure elements. Consistent with their location on protein surfaces, loops (turns) are dominated by hydrophilic residues. This tendency can be exploited to locate 'strong' loops. Sequence pattern descriptors are written which embody this concept. 'Weak' or context dependent loops can be located based on hydrophilicity and their sequential juxtaposition to 'strong' loops. By placing an upper limit on the length of chain between successive loops, globularity is ensured and predictive accuracy improves to about 90%. This improvement is predicated on an ability to recognize whether the protein is all α-helical (α/α), all β-sheet (β/β) or alternating helix and sheet (α/β). We are extending this work to locate secondary structure between loops. Preliminary work suggests that in α/α proteins, it is possible to recognize the core of helices 89% of the time in a development set of 8 proteins and 84% of the time in a control set of 9 proteins. Patterns to recognize the N and C terminal caps of α-helices are currently under development.

Tertiary structure prediction. Our past computational experiments indicate that secondary structure is a useful intermediate for predicting tertiary structure. Myoglobin is an instructive example. Six of the helices in myoglobin contain clusters of hydrophilic residues along their surface. These docking sites can be located algorithmically. Crick and later Chothia and co-workers recognized that helices pack against other helices with significant geometric restrictions owing to the 'knobs and holes' created by side chains on the helix surface. A combinatorial problem is apparent, how many structures can be generated for myoglobin by docking hydophobic sites on helices using the idealized geometry of helix packing. While 3.4 x 10^8 structures are possible, only 20 respect the connectivity of the chain and steric constraints on amino acid packing. The best of these structures is 4.5Å r.m.s. from the crystal structure and contains a suitable heme binding site between the proximal and distal histidine. Similar calculations on more than 30 $\alpha/\alpha, \alpha/\beta, \beta/\beta$ proteins suggest that while 10^4 - 10^{15} structures 'are possible', only 1 - 10^3 are topologically and sterically reasonable. Unfortunately, the multiple minima problem is replaced by a multiplicity of alternative structures which are indistinguishable at the level of these crude models.

Our recent work suggests that these alternative structures are heterogeneous and may be separable using more detailed models. However, explicit energy calculations which follow the gradient of a complex potential function remain bound by the limitations of the multiple minima problem. We have examined a 'one sphere per residue' model of proteins. Aromatic residues are modelled by two or three spheres. At this level, 'correct' structures adopt a packing which is homogeneous and the sensibility of pairwise residue interactions can be quantified. I will discuss developments of a potential function consistent with this simplified representation. Molecular dynamics simulations, the time dependent integration of Newton's equations of motion, will be used to refine the combinatorially generated structures. Preliminary work suggests that these simplified energy calculations will be significantly less time intensive than traditional energy calculations using detailed molecular representations.

New computational techniques for computing biomolecular structures and their dynamics

T. Schlick, Courant Institute and Dept. of Chemistry, NYU.

The complex three-dimensional structures of biomolecules are key to our understanding of important biochemical processes. Outstanding problems include the complex folding of a protein into its characteristic three-dimensional shape and the subtle sequence and environment-dependent conformations of DNA double helices and a myriad of other DNA forms. Experimental techniques are one way to obtain structural information, but the technical difficulties are still formidable. Another approach is computational and includes potential energy minimization and molecular dynamics simulations (MD).

In the first approach, a function which approximates the potential energy of a molecular system is minimized to yield favorable regions in conformation space. In the second, the Newtonian equations of motion are solved repeatedly by some numerical scheme, and a space-time trajectory of the molecular system is obtained. While in theory, the dynamic picture generated by MD should reveal many local energy minima, serious computational obstacles make minimization and dynamics quite distinct approaches in practice. Minimization can provide important structural information and can generally be performed much faster than MD. However, the major difficulty is the multiple-minimum problem. MD bypasses this difficulty by relying instead on a numerical integration scheme to generate molecular conformations, but in order to guarantee numerical stability in explicit integration schemes, the time step Δt must be very small. This limits the total length of the trajectory that can be generated and, consequently, the scope of molecular motions that can be captured.

Our work has focused on improving computational methodologies in view of these fundamental difficulties. A new MD scheme has been developed to permit larger time steps: The implicit Euler integration scheme is combined with the Langevin dynamics formulation, which adds frictional and random forces to the systematic force; by choosing the frictional damping constant appropriately, we have shown that the high-frequency vibrational modes can be effectively damped. Furthermore, this damping can be set to mimic quantum-mechanical effects. Implementation of the scheme requires that a nonlinear system be solved at every time step. A truncated Newton minimization algorithm for large-scale potential energy functions has thus been developed, and the nonlinear system is formulated as a minimization problem related to (but easier than) potential energy minimization. Applications to several molecular systems have demonstrated the feasibility of larger time steps with the Langevin/Implicit-Euler scheme as well as its quantum-mimicking effects. Current work focuses on applications to the folding patterns of supercoiled DNA. These computational methodologies and applications will be presented at the workshop.

Building physical genome maps by random clone overlap; a progress assessment of work on human chromosome 19.
E. Branscomb, Lawrence Livermore National Labs.

At Livermore, a physical map of human chromosome 19 (ca 60Mbp) is being constructed. At the highest resolution, this map consists of an ordered asssembly of cloned DNA fragments of approximately 40Kbp size. To build this component of the map, a collection of over 8000 of these fragments (sufficient to form a 5-fold redundant coverage of the chromosome) is selected so that the fragments have approximately random locations along the chromosome. The task is then to rediscover the native order these cloned fragments had in the chromosome and to assemble them, so far as possible, into a continuous covering 'image' of the chromosome. The assembly depends in part on detecting fragments that were neighbors in the sense that they have a DNA segment in common (i.e. they 'overlap'). Our approach to this task involves obtained so-called 'restriction fingerprints' of each clone and comparing fingerprints. Evidence for overlap in such data is statistical and error-plagued. At present, approximately 9K cloned elements have been analyzed with about 7K yielding useful fingerprints. These have been assembled into about 700 islands.

I will first discuss briefly how the statistical analysis of overlap is done and how the cloned fragments are assembled. I will then assess in what way the observed results depart from expectations based on analytic models and Monte-Carlo simulations, what these differences reveal about the perverse 'biology' of human DNA, and how all this relates to locating and characterizing genes of interest.

Monte Carlo likelihoods for linkage analysis and complex models.
E. A. Thompson, Dept. of Statistics, University of Washington.

The computation of likelihoods for genetic models on the basis of data observed on related members of a pedigree has helped to resolve the inheritance of many genetic diseases, and to map the genes responsible. However, many traits with complex patterns of inheritance remain unresolved. Progress requires the use of more complex genetic models, in conjunction with linkage analysis, and the genetic homogeneity provided by large pedigrees. Methods of computation developed over the last 20 years have greatly extended the scope of pedigree analysis, but there remain strict limits on computational feasibility. Combination of these methods with Monte Carlo evaluation methods, including use of the Gibbs' sampler, provides the potential for substantial further progress.

Dissecting complex inheritance: statistical and computational issues.
E.S. Lander, Whitehead Institute for Biomedical Research and MIT.

Because of advances in molecular genetics, it has recently become possible (and almost straightforward) to map the genes that cause simple Mendelian traits and diseases in humans. The vast majority of inherited traits are not caused by single genes, but by the combined effects of multiple genes. We will discuss how to dissect such cases of polygenic inheritance by combining the tools of molecular genetics with statistical analysis and computer algorithms. On the theoretical side, we will show how key genetic questions hinge on the extreme value distributions of multi-dimensional random fields. On the experimental side, we will show how these results apply to the genetic basis of such diverse subjects as the study of hypertension and the production of tomato sauce.

Practical and theoretical advances in sequence comparison
G. Myers, Dept. of Computer Science, University of Arizona.

A new approach for rapid similarity searches, BLASTA, has been developed which directly approximates alignments that optimize a measure of local similarity, the Maximal Segment Pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of the alignments it generates.

The basic algorithm is simple, robust, and an order of magnitude faster than existing sequence comparison tools of comparable sensitivity such as FASTA.

The central algorithmic idea for BLASTA is a simplification of part of a more theoretical but interesting recent result. Given an inverted index for a database, we have devised an approximate match search algorithm whose expected performance is sublinear in the length of the database. This result represents a new advance in the arena of provably efficient algorithms for sequence comparison. Its practical utility for databases of current megabase size is still unknown. However, we will ultimately need sublinear methods for gigabase databases, as the time taken by any linear algorithm, such as FASTA or BLASTA, grows directly with the size of the database and so will eventually be too slow. Moreover, sublinear results for approximate match suggest further improvements for the general problem are possible.

Error detection in DNA sequences.
R.J. Roberts, Cold Spring Harbor Laboratory.

We have developed an algorithm that is able to detect certain errors within the coding regions of DNA sequences. The algorithm compares an input experimental sequence, translated in all 6 reading frames, with protein sequences present in a database such as SwissProt or PIR. Whenever a local similarity is found between a segment of the experimental sequence and a database sequence, the algorithm explores the continuation of that similarity in all three continuing reading frames. If a substantial similarity is detected across reading frames the region is recorded as one in which a DNA sequence error may have occurred. These potential sequence errors are recorded schematically and also printed as a more detailed version indicating precisely the region where errors may have occurred. The program is useful to check a newly-determined sequence for obvious errors and thus will help the investigator ensure the accuracy of the final product. We have conducted extensive tests of the algorithm using, as experimental sequences, many of the unidentified open reading frames flanking identified genes in the GENBANK database. Many clear examples of sequence errors can be found and in some cases, following correction of the errors, functions can be assigned to the products of these unidentified reading frames.

Computer methods for locating kinetoplastid cryptogenes.
M.S. Waterman, Depts. of Mathematics and Biological Sciences, USC.

RNA editing in the mitochondria of kinetoplastid protoza involves the insertion and/or deletion of precise numbers of uridine residues at precise locations in the transcribed RNA of certain genes. These genes are known as cryptogenes. In this paper we study computational algorithms to search for unknown cryptogenes and for the associated templates for insertion of uridines, gRNA sequences. The pairwise similarity search algorithm of Smith and Waterman is modified to study this problem. The algorithm searches for unknown gRNAs given the cryptogene sequence. The method is tested on 4 known cryptogenes which are known to have 7 associated gRNAs. The statistical distribution of the longest gRNA when comparing random sequences is derived. Finally we develop an algorithm to search for cryptogenes using amino acid sequences from related proteins.

Workshop in Statistical Issues in Molecular Biology

Stanford, California. August 8, 1993.

9:00-9:45	David Botstein, Stanford *Molecular biology for statisticians*
9:45-10:30	Samuel Karlin, Stanford *Assessing inhomogeneities in long DNA sequences*
10:30-11:00	Refreshments
11:00-11:45	Bruce Weir, North Carolina State *Statistical issues in the forensic use of DNA*
11:45-1:15	Lunch
1:15-2:00	Jurg Ott, Columbia *Statistical issues in human linkage analysis*
2:00-2:45	E.A. Thompson, Washington *Monte Carlo likelihoods in genetic analysis*
2:45-3:15	Refreshments
3:15-4:00	Terry Speed, UC Berkeley *Statistical aspects of genetic recombination*
4:00-5:00	Discussion

Assessing inhomogeneities in genomic sequences
Samuel Karlin, Stanford University

The unprecedented accumulation of DNA and protein sequence data (including the complete physical map of *E. coli* and the first complete nucleotide sequence of a eukaryotic chromosome (yeast chromosome III)) and more than 15 contigs exceeding 100 kb length poses challenges and opportunities in terms of genomic organization and analysis. My presentation will focus on statistical, computational, and informatics problems in this context. We describe methods and concepts for assessing and interpreting inhomogeneities in long DNA sequences. In particular, we (1) analyze patterns and anomalies in the occurrences of short oligonucleotides; (2) characterize the nature and locations of significant direct and inverted repeats; (3) delimit regions unusually rich in particular base types (e.g., G+C, purines); (4) analyze the distributions of markers of interest, e.g., delta elements, and ARS (autonomous replicating sequences) in yeast, Chi, Dam and Dcm sites in E. coli or other special oligonucleotides; (5) characterize rare and frequent oligonucleotides; and (6) analyze position-dependent fluctuations in sequence compositions. Marker arrays and compositional heterogeneity will be analyzed by employing three principal methods: (a) r-scan statistics to characterize extremes in marker spacings; (b) plots of position-dependent marker frequencies over a sliding window; and (c) quantile tables to contrast different count distributions.

Statistical issues in the forensic uses of DNA
Bruce Weir, North Carolina State University

The controversy surrounding the forensic use of DNA profiles continues, to the point that one judge can rule statistical calculations inadmissable because the prosecution did not use the 1992 NRC report, and another judge can rule the entire evidence inadmissable because the NRC report was used. Two issues will be discussed. An examination of a series of tests (conditional and unconditional exact, chi-square goodness of fit, and likelihood ratio) for independence of allele frequencies at one or two loci has been made. Different tests have different properties, but there is broad agreement between them over the status of the FBI databases. Secondly, the issue of population substructure will be examined with a series of measures of genetic relatedness. Four-gene measures are needed for the joint frequencies of crime scene and suspect genotypes at one locus, although it may be possible to use two-gene approximations.

Statistical issues in human linkage analysis
Jurg Ott, Columbia University

In linkage analysis, the random occurrence of recombination along chromosomes is exploited to define genetic distances between genetic loci. I will thus start out by briefly introducing recombination and map functions as far as this is required for an appreciation of genetic distances. Then, I will present a few examples of the many interesting statistical problems occurring in human linkage analysis.

Two loci with a probability for recombination (the recombination fraction, theta) of less than 0.5 are said to be genetically linked. The two main aspects of linkage analysis are estimation of theta and testing of the null hypothesis, theta=0.5. These seemingly ordinary tasks are complicated because the observations consist of nonindependent individuals. In fact, individuals MUST be related for the estimation of the recombination fraction. Most analyses are currently carried out by maximum likelihood. Test results are usually presented in terms of the likelihood ratio in favor of the alternative hypothesis rather than, for example, as empirical significance levels.

Various interesting statistical issue arise in gene mapping, where one aims to order a number of genetic marker loci. One order is considered to be "significantly" better than another when it is supported by a likelihood at least 1000 times larger than that under the alternative order. How often this leads to a wrong order is difficult to judge. For a given data set, computer simulation can give an approximate answer but is rarely carried out for this purpose.

Errors in determining a marker's phenotype have grave consequences for gene mapping. They currently occur with a frequency of 1-2search for unusually frequent occurrences of rare events, particularly multiple crossovers within a short distance. Of course, this strategy selectively eliminates multiple crossovers, which tends to show up as interference. I will discuss a few statistical methods of error detection.

Marker heterozygosity is a statistically interesting quantity. It is the proportion of heterozygous individuals in the population subject to the restriction that genotypes occur in Hardy-Weinberg equilibrium (alleles pair up at random to form genotypes). The MLE of heterozygosity is slightly biased; its variance is hard to calculate. In practice, biased estimates without standard errors are reported.

For mapping a new gene (eg. a disease gene), a high heterozygosity and a dense map of markers are important. One can show statistically that for localizing a new gene, one of these quantities can be substituted for the other. However, the precision of the estimated location depends almost entirely on the density of the marker map and not on the heterozygosity of the markers.

The mapping of disease genes also poses interesting problems. Recently, two-stage procedures were introduced, that is, one first tries to find those areas in the genome that exhibit mild evidence for harboring a disease gene. In a second stage, these areas are scrutinized. Various parameters involved in this strategy are then optimized with the aim to minimize the cost of finding a disease gene.

Monte Carlo likelihood in genetic analysis
Elizabeth Thompson, University of Washington

Markov chain Monte Carlo (MCMC) provides methods for obtaining realisations from probability distributions known only up to a normalising factor. Such distributions arise in "missing data" problems, as well as in many other areas. In such problems, the realisations obtained from the Markov chain can be used to obtain Monte Carlo approximants to a likelihood function, and thus extend the scope of likelihood analysis to situations where an exact likelihood cannot be computed.

Genetics provides many examples of complex, dependent, highly structured, data. The fitting of genetic models on the basis of observed characteristics of related individuals is a prime example of a "missing data" problem, since the underlying genes cannot be observed. The Monte Carlo estimation of likelihoods for genetic models will be presented, the possible increase in efficiency through partial exact computation will be considered, and a solution to the problem of estimating log-likelihood differences between widely disparate hypotheses will be given.

The example also shows that the Gibbs sampler is a rather poor MCMC method. Although the Gibbs sampler is adequate for the example presented, many problems will require more efficient samplers in the Metropolis-Hastings class of algorithms for MCMC.

Statistical aspects of genetic recombination
Terry Speed, University of California at Berkeley

"Genetic recombination is ubiquitous; all forms of life engage in either occasional or periodic reshuffling of the cards of their genetic decks. ... We must study genetic recombination because it is important. ... [But...] Recombination in our service is useful."
From the preface of *Genetic Recombination. Thinking about it in Phage and Fungi* by F.W. Stahl, San Francisco: W.H. Freeman & Co., 1979.

"The term re-combination seems to have been introduced into genetical literature by Bateson (1909), and was used initially for pairs of character differences showing independent inheritance."
From the introduction (p.1) of *Genetic Recombination. Understanding the Mechanisms* by H.L.K. Whitehouse. Chichester: John Wiley & Sons, 1982.

By the time you hear this talk at the Workshop, you will have met recombination in two, possibly three of the previous lectures: as an important genetic phenomenon, as the basis for constructing genetic maps, and in connection with establishing linkage between disease genes and mapped genetic markers. The aims of this talk are three fold: to spend some time examining recombination in its own right, from the perspective of a statistician; to exhibit the beautiful interplay of mathematics, statistics and biology brought to light by the study of recombination, with the aim of luring some of you into the intersection of statistics and genetics; and to indicate (briefly) some of the topics warranting further study. My discussion will be restricted to so-called eukaryotes, organisms such as fungi, animals and plants, which consist of cells with nuclei which undergo meiosis. (Bacteria and viruses have their own forms of recombination, and there are statistical aspects there as well, but there is only so much I can do in 3/4 hour.)

I will begin by showing you some genetic data exhibiting recombination. We can find many interesting features of such data, but the phenomenon of interference is the most interesting for us, not the least because it involves the central statistical notion of independence. The biological interpretation of such data requires a model of the process of recombination, and - here is where statistics gets involved - all such models have a stochastic component. Indeed, multilocus genetic mapping requires a full probability model to relate what are observed - recombination fractions - to genetic distances, which are the parameters of an underlying, unobserved crossover process. Statisticians will naturally be interested in the nature of the models used, in the evidence that they are biologically plausible, in the extent to which they fit the data, in the robustness of inferences to features of the model, in short, in all the technical problems posed by their use in addressing questions of biological interest.

Some of the topics I will discuss include the two-strand model, map functions, four-strand models, the distinction between crossover and chromatid interference, the Poisson, chi-squared and some other crossover models, ordering genetic loci, and the robustness and efficiency of the Poisson model.

Session on Population Genetics

56th Annual Meeting, Institute of Mathematical Statistics

San Francisco, California. August 9, 1993.

Joanna Mountain, Stanford
Frequencies of ancestral alleles: interpreting data from human populations on the basis of theoretical distributions and simulations

Ronald Lundstrom, Collaborative Research, Inc.
Estimating evolutionary parameters: population dynamics versus molecular dynamics

Sally Otto, UC Berkeley
Evolutionary consequences of recombination

Mike Steel, University of Canterbury, New Zealand
Doing the splits – a direct way to reconstruct phylogenetic trees based on Kimura's 3-parameter model

Frequencies of ancestral alleles: interpreting data from human populations on the basis of theoretical distributions and simulations.
Joanna L. Mountain, Stanford University

Comparison of allele frequencies in populations around the world provides insight into the evolutionary history of these populations. Although identification of the ancestral state for such polymorphisms provides further insight, we can rarely identify this state. Recently we examined 79 nuclear DNA polymorphisms in eight human populations from four continents. By examining the 79 loci in non-human primate species (chimpanzees, gorillas, and orangutans) we have now been able to identify the ancestral allele for a number of polymorphisms. For 61 out of 79 loci examined, chimpanzees share exactly one allele with humans; we consider these alleles to be ancestral at their respective loci. Such identification of ancestral alleles enables us to compare distributions of the frequencies of these alleles across human populations. Distributions vary dramatically from population to population. Wright suggests that the ancestral allele should more commonly have higher frequencies, leading to a distribution skewed to the right. Ancestral allele frequencies for several of the human populations appear to have such a distribution. Others have flatter or more U-shaped distributions. Although some differences between populations may be due to the fact that polymorphisms were originally detected in Europeans, changes in population size or natural selection are more likely to have given rise to these differences. The question addressed here, through simulation, is whether differences between distributions might be due to differences in population size changes.

Estimating evolutionary parameters: population dynamics versus molecular dynamics
Ronald S. Lundstrom, Collaborative Research, Inc.

Historical fluctuations in the size of human populations have influenced the genetic structure of modern day populations. Population surveys of DNA sequence data have used this fact to estimate dynamics of human population growth since the origin of the species. By comparing DNA sequences from diverse ethnic groups, advocates of DNA sequence data claim the ability to estimate the amount of time since a common ancestor of all humans, or at least an ancestor of the gene represented by the DNA sequences. Other studies have inferred population dynamics from the distribution of pairwise sequence divergence in a random sample from the population.

The *coalescent process* describes the genealogy of a random sample from a population. The population size as a function of time appears as a parameter of the process. We show that the number of ancestors of the sample at various times in the past is a non-homogenous pure death process, and an explicit formula for the death rates is given in terms of the population size function. Mutations occur according to independent homogeneous Poisson processes along each of the ancestral lineages determined by the pure death process. The distribution of the time to a common ancestor of the sample is obtained from the death process.

Since the population size function is a parameter in the coalescent process, inferences about population history can be tested by simulating the coalescent. We have used simulation to assess the resolving power of population DNA samples to estimate historical population sizes.

Evolutionary consequences of recombination
Sarah P. Otto, University of California, Berkeley.

Recombination shuffles together the genetic material carried by different members of a sexual species. This genetic mixing uncouples the evolutionary fates of neighboring genes, hence reducing the constraints which limit the capacity of genetic loci to respond to selection. In the process, however, recombination also separates advantageous gene combinations, the very gene combinations that enabled the parents to survive and reproduce. Whether or not the adaptation of a population to an environment is more rapid in the presence of recombination, that is, whether or not recombination speeds up the evolutionary process, depends critically on the ways in which this process is modelled. Using various stochastic and deterministic models, we have found that the effects of recombination depend on the population size, the initial population composition, and the selection regime under consideration.

Doing the splits – a direct way to reconstruct phylogenetic trees based on Kimura's 3-parameter model
Michael A. Steel, University of Canterbury, New Zealand.

We describe a new method for reconstructing evolutionary trees directly from DNA and RNA sequence data. The method has four key features: 1) It is statistically consistent under Kimura's 3-parameter substitution model, with no assumption required concerning the distribution of nucleotides for the ancestral taxon; 2) The method extends easily, for example, to allow for varying mutation rates at different sites in the sequences; 3) The method is computationally simpler than maximum likelihood as it does not require a "double optimization" (i.e. first finding a best fit for each tree, and then finding a best tree); and 4) The method generates explicitly a family of phylogenetic invariants for Kimura's 3-parameter model which distinguishes all trees for any number of taxa.

The method is an extension from 2-state to 4-state character sequences of an earlier method developed by M.Hendy and D. Penny, and based on Hadamard matrices. The extension relies on group-theoretic properties of Kimura's 3-parameter model, described by S. Evans and T. Speed. Preliminary trials suggest the new method works well.