Statement of Research Interests

Raazesh Sainudiin

February 28, 2006

My general research interests lie in the statistical inference of stochastic processes superimposed on stochastically evolving networks. Stochastic processes are broadly defined to include Poisson processes, Markov chains, and non-linear stochastic differential equations. Stochastically evolving networks are meant to provide the biologically realistic contexts within which our stochastic processes arise. For example, such contexts are provided by phylogenetic X-trees in molecular evolution, structured ancestral recombination graphs in population genetics, contact networks of susceptible and infected nodes in epidemiology, or gene interaction networks in cell biology. I am interested in statistically sound, biologically realistic, computationally efficient, and numerically rigorous approaches to statistical inference in the biological sciences. During the next decade, I plan to conduct research and teach courses at the integrative interface of biological, computational, mathematical, and statistical sciences. My recent and current research can be broadly classified into two areas; (1) statistical genetics and (2) computational statistics.

Statistical Genetics

Microsatellites: Microsatellites are simple sequence repeats in DNA, for example the motif AT repeated twenty-five times in a row. Microsatellites mutate by changing the number of their repeats. They are used in the construction of genome-wide maps, in the search for disease-causing genes, and in identification for both forensics applications and paternity tests. For a recent mathematical treatment of microsatellites see Calabrese and Sainudiin (2004). My M.S. thesis was an effort to place several continuous time Markov chain models of microsatellite evolution in the literature into a cohesive framework that allows for statistical inference (Sainudiin et al., 2004). Biologically realistic models of complicated random walks on finitely many states in continuous time were shown to provide better descriptions of microsatellite evolution through their likelihood ratios using real data. Subtle differences in the nature of microsatellite evolution between humans and chimpanzees were captured in that study through the maximum penalized likelihoods of time non-homogeneous Markov chain models superimposed on species trees. Robustness in population genetic inference with microsatellites to model mis-specification, raised in Sainudiin et al. (2004), is currently being studied via simulations. More realistic molecular sequence data contains both fast-mutating repetitive DNA (microsatellite) and slow-mutating non-repetitive DNA. The first phase of my post-doctoral research at Oxford's Mathematical Genetics group is to develop novel sequential importance samplers that can obtain the likelihood under coalescent models from such DNA molecules which simultaneously evolve over two scales of time that can differ by up to five orders of magnitude.

Comparative Phylogenetics: We provided a simple generalization of a fairly popular codon substitution model used to detect positive Darwinian selection in rapidly evolving proteins (Nielsen and Yang, 1998), such as the human leukocyte antigen (HLA) of the immune system. Multiple independent projections of the stochastic dynamics of such molecules, modeled as simple mixtures of Markov chains, as they evolve down a phylogenetic tree, can shed light on positive physicochemical selection at the individual amino-acid level in the HLA system as well as the plant self-incompatibility (SI) system of crucifers (Sainudiin et al., 2005). First order empirical mutagenetic studies in the SI system have recently confirmed our predictions (Nasrallah, personal communication). Finer site-specific mutagenetic studies on the SI system are expected to provide empirical feed-backs that can improve the models employed in that study. We have extended these models to test the extent and nature of site-specific physicochemical pressures acting on the evolving proteins through likelihood ratio tests between more general mixtures of Markov chain models of codon substitution superimposed on phylogenetic trees (Wong et al., 2005).

Exact Bayesian Inference from Summaries: When full likelihood is computationally infeasible for realistic models of sequence evolution in a demographically complex population with rich sub-population structure, one may infer a population's history based on summary statistics of the data in a simulation-based framework. Such methods are known as approximate Bayesian computations (ABC) in population genetics (Marjoram et al., 2003).

In ABC methods, one first simulates data under a model parameter (drawn according to a prior) and summarizes it to obtain the summary statistics of interest. Finally, the simulated parameter value is accepted if it's summary statistics lie within a ball of an acceptance-radius ϵ centered at the observed summary statistics b_{Obs}. Several summary statistics used in ABC methods are linear combinations of the $n-1$ dimensional site frequency spectrum (SFS), which is a better summary of the the full data (often an $n \times m$ nucleotide matrix of homologous DNA sequences of length m from n individuals in the population of interest). Using algebraic statistical techniques introduced by Diaconis and Sturmfels (1998), we obtain a Markov basis of the conditional site-frequency spectra. Any finite set of summaries $b \in \mathbb{R}^k$, each of which is a linear combination of the SFS (e.g. number of segregating sites S and heterozygosity π), has linear weights encoded by a $k \times (n-1)$ matrix A. The matrix A can be used to condition the non-negative $n-1$ dimensional lattice $\mathbb{Z}_+^{n-1}$, the space of all possible spectra. A Markov basis not only guarantees the irreducibility of the conditional lattice $\Gamma_A^b \triangleq \{x \in \mathbb{Z}_+^{n-1} : Ax = b\}$ for any b, including b_{Obs}, but also gives the local moves for a Markov chain that can be made to sample from any stationary distribution on the state space Γ_A^b. Using the theoretical niceties of the infinitely many sites model of DNA mutation and a Markov basis for Γ_A^b, we are able to integrate the likelihood of *the exactly* observed summaries over other nuisance parameters in the space of structured ancestral recombination graphs. This novel integration over exact summaries allows us to make the acceptance-radius $\epsilon = 0$ and thus entirely avoid a major approximation in ABC. This project with Jim Booth, Mike Stillman and Kevin Thornton is in progress.

Computational Statistics

Rigorous Numerical Statistics: When statistical inference is conducted in a maximum likelihood (ML) framework, one is interested in the global maximum of the likelihood function over the parameter space. Explicit analytical solutions for the maximum likelihood estimates (MLEs) are only available for the simplest models. In practice one settles for a local optimization algorithm to numerically approximate the global solution. However, computers can only support a finite set of binary floating-point numbers called a number screen (P754, 1985). Conventional floating-point arithmetic with a number screen is not exact. Statistical decision procedures that rely on having found some global optimum through any numerical approach may suffer from the following sources of errors: *roundoff error* – the difference between computed and exact result, *truncation error* – from having to truncate an infinite sequence of operations, *conversion error* – inability to machine-represent decimals with infinite binary expansion, *ill-posed experiment* – due to possible presence of unknown nonidentifiable subspaces, and *measurement error* – usually induced by limits on empirical resolution. Furthermore, traditional nonlinear programming techniques that use local information start from some approximate trial point(s) and iterate by sampling only finitely many points. Thus, they can neither validate that the likelihood function has not soared between the sampled points, nor guarantee escape from a set of local maxima. Therefore, methods that use local information at finitely many points without accounting for all five major sources of errors, cannot be expected to yield anything more than an approximate decision.

The rigorous global optimization algorithm based on Hansen's method (Hansen, 1992) with Ratz's modifications (Ratz, 1992) in conjunction with automatic differentiation (Rall, 1981) can rigorously enclose the global maximum of the likelihood function and the MLEs through interval analysis (Moore, 1967). Such interval methods evaluate the likelihood function over a continuum of points including those that are not machine-representable and account for the five sources of errors described earlier. These enclosure techniques allow for *rigorous numerical statistics*, i.e. *computer-assisted proofs* of the statistical decisions, the subject of my Ph.D. thesis titled *machine interval experiments*. Two applications of such experiments are as follow:

Application 1 *One of the challenges in phylogenetics posed by Semple and Steel (2003) is to provide a general procedure that guarantees to solve for the MLEs of the branch lengths of a fixed phylogenetic X-tree $\mathcal{T}$. Assuming the simplest model of DNA evolution (Jukes and Cantor, 1969), Sainudiin (2004) rigorously enclosed the most likely trees, with two, three, and four leaves, based on primate mitochondrial DNA sequences (Brown et al., 1982) by operating in interval-extended Hessian differentiation arithmetic.*

Application 2 *I have used interval methods to enclose the ML and MLEs of univariate mixture models, with Gaussian and exponential densities. Using a notion of 'envelopes of empirical CDFs' (EeCDFs), one can obtain finite sample bootstraps of the rigorous enclosures of analytically intractable likelihood ratio test statistics. This approach can resolve problems that are induced by quasi-local searches with uncontrolled floating-point arithmetic, as mentioned in sections 6.5 and 6.6 of McLachlan and Peel (2000).*

Randomized Enclosure Algorithms: In Bayesian statistical inference, one is interested in obtaining samples from a target density $\pi(\theta)$, a high-dimensional, possibly multi-modal density, using only the shape of the target

π^* or only ratios of the form $\pi(\theta)/\pi(\vartheta)$, where $\theta, \vartheta \in \mathbf{\Theta} \subset \mathbb{R}^n$. The challenge is to obtain samples from $\pi \triangleq \pi^*/N_\pi$ without any knowledge of the normalizing constant $N_\pi \triangleq \int_{\mathbf{\Theta}} \pi^*(\theta)$. Several approaches to this problem rely on Monte Carlo methods such as RS, the rejection sampler of Neumann (1963), or MHS, the Metropolis-Hastings sampler (Metropolis et al., 1953; Hastings, 1970). Using an adaptive partitioning of the domain $\mathbf{\Theta}$ and the natural interval extension of π^*, we have been able to draw samples using an interval-extended RS (Sainudiin and York, 2005). This sampler *guarantees* independent samples from π by ensuring that the envelope function for the RS necessarily dominates π^* everywhere on $\mathbf{\Theta}$, a prerequisite for RS. Most conventional RSs cannot make this guarantee for nontrivial densities as they rely on non-rigorous numerical procedures to obtain the envelope function. Using our RS, we have been able to sample from various densities that have been previously considered "pathological," including the witch's hat density in up to 11 dimensions. We have also applied the sampler to non-Euclidean domains (piece-wise Euclidean tree spaces of negative curvature), complicated multivariate Rosenbrock densities (sharp banana-shaped ridge in up to 10 dimensions) and temperature-sensitive bumpy Levy densities (700 peaks of vastly different relative heights).

Chapter 2 of my Ph.D. thesis was partly concerned with the behavior of MHSs in $\mathbb{I}\mathbf{\Theta} \subset \mathbb{IR}^n$, the space of compact interval vectors or *boxes*. The goal was to find a MHS on $\mathbf{\Theta}$ that can quickly sample from π, even when it is highly multi-modal, using *enclosures* of ratios of the form $\pi^*(\Theta)/\pi^*(\mathit{\Theta})$, where $\Theta, \mathit{\Theta} \in \mathbb{I}\mathbf{\Theta}$. Given a base chain that can propose a move to a new state θ from the current state ϑ, the MHS is a Markov chain on $\mathbf{\Theta}$ with stationary distribution π. If the proposal distribution of a base chain is independent of the current state ϑ, then the base chain is said to be independent. We have shown that for the two parameter family of equi-weighted, symmetric mixture of two Beta densities given by, $\pi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{2\,\Gamma(\alpha)\Gamma(\beta)}\left(\theta^{\alpha-1}(1-\theta)^{\beta-1} + \theta^{\beta-1}(1-\theta)^{\alpha-1}\right)$, where $\theta \in [0,1]$, the interval-based MHS with independent base chains informed by interval arithmetic is almost as efficient as the optimal MHS with independent base chain given by $\pi_{\alpha,\beta}$ itself. The performance of the MHS with 'simple random walk' base chains or uniform independent base chains, in terms of the total variation distance to the target $\pi_{\alpha,\beta}$, is significantly worse. Using the spectral decompositions available for MHSs with independent base chains (Liu, 1995), we showed that there is a phase transition in the ratio of the second largest Eigen values of the interval-based MHS to the uniform independent MHS, as a function of $\gamma = \alpha = \beta$ in $\pi_{\alpha,\beta}$ and the partition size of $\mathbf{\Theta}$. The phase transition shows that even for extremely spiky densities the interval-based MHSs (induced by a coarse partition of $\mathbf{\Theta}$) have superior rates of convergence to π than the uniform independent MHS.

Most statistical inference today is done on computing machines through numerical methods that do not rigorously account for the physical realities of such machines with finite memory. We have extended the index set ($\mathbf{\Theta} \subset \mathbb{R}^n$) of the usual statistical experiment to $\mathbb{I}\mathbf{\Theta} \triangleq \{\, \Theta \triangleq [\underline{\theta}, \overline{\theta}] \subset \mathbf{\Theta},\ \underline{\theta} \leq \overline{\theta} \,\}$, a complete metric space under the Hausdorff metric, and shown that the extended experiment allows for rigorous enclosures of the globally optimal decision(s) with computers (Sainudiin, 2005). Possibly sub-optimal decisions may suffice for several decision problems. However, for others, such as parameter estimation for nonlinear stochastic differential equations, or training neural networks to aid in decisions concerning national security, or anti-missile computations in military decisions (Information management and technology division, 1992), or engineering design decisions in ocean (Selby et al., 1997) or space (Lions, 1996) applications, one may want or need to guarantee the globally optimal decision. The enclosure methods provide statisticians with powerful tools for rigorous numerical inference.

In the near future, I intend to efficiently apply such enclosure techniques to intermediate dimensional problems ($\mathbf{\Theta} \subset \mathbb{R}^{50}$). One approach will first employ algebraic statistical techniques (Pachter and Sturmfels, 2005) to find sufficient and computationally efficient model formulations and then take advantage of interval methods to enclose the optimal decisions, as already done for a simpler case in Sainudiin and Yoshida (2005). Another approach will exploit all available random access memory to store the pre-enclosed likelihood function for nearly constant time hash access through hull operations. This will produce tighter enclosures of the range and reduce the computational cost associated with interval evaluations on processors that are optimized for floating-point arithmetic. On the statistical genetics front, I intend to develop computationally efficient sequential particle filters that can evaluate the likelihood from homologous intra-species and inter-species DNA sequences within a coalescent framework that incorporates biologically realistic mutational models. Such mutational models will be motivated by the literature on the biochemistry and molecular biology of DNA damage and repair, which has been largely overlooked by most statistical geneticists working with population genetic data.

Acknowledgments

I was funded by the NSF grant DGE-9870631 to Strogatz and a joint NSF/NIGMS grant DMS 0201037 to Durrett, Aquadro, and Nielsen during my Ph. D. years at Cornell University. I am currently a research fellow of the Royal Commission for the Exhibition of 1851.

References

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson (1982). Mitochondrial DNA sequences of primates, tempo and mode of evolution. *Jnl. Mol. Evol.* *18*, 225–239.

Calabrese, C. and R. Sainudiin (2004). Models of microsatellite evolution. In R. Nielsen (Ed.), *Statistical Methods in Molecular Evolution*, Series: Statistics for Biology and Health, pp. 289–306. New York: Springer.

Diaconis, P. and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics 26*, 363–397.

Hansen, E. (1992). *Global optimization using interval analysis*. New York: Marcel Dekker.

Hastings, W. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika 57*, 97–109.

Information management and technology division (1992). Patriot Missile Defense: Software problem led to system failure at Dhahran, Saudi Arabia. Report GAO/IMTEC-92-26, United States, General accounting office.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian protein metabolism*, Volume 3, pp. 21–123. New York: Academic Press.

Lions, J. L. (1996). Chairman of the inquiry board, Ariane 5, Flight 501 failure: Report by the inquiry board. Inquiry, European Space Agency.

Liu, J. (1995). Metropolised independent sampling and comparisons to rejection sampling and importance sampling. *Statist. and Comput.* *6*, 113–119.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA 100*, 15324–15328.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Jnl. Chem. Phys.* *21*, 1087–1092.

Moore, R. E. (1967). *Interval analysis*. Englewood Cliffs, New Jersey: Prentice-Hall.

Neumann, J. (1963). Various techniques used in connection with random digits. In *John Von Neumann, Collected Works*, Volume V. Oxford, UK: Oxford University Press.

Nielsen, R. and Z. Yang (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics 148*, 929–936.

P754, I. T. (1985). *ANSI/IEEE 754-1985, Standard for Binary Floating-Point Arithmetic*. IEEE, New York.

Pachter, L. and B. Sturmfels (Eds.) (2005). *Algebraic statistics for computational biology*. Cambridge University Press.

Rall, L. B. (1981). *Automatic differentiation, techniques and applications*, Volume 120 of *Springer lecture notes in computer science*. New York: Springer-Verlag.

Ratz, D. (1992). *Automatische ergebnisverifikation bei globalen optimierungsproblemen*. pHd dissertation, Universitat Karlsruhe, Karlsruhe.

Sainudiin, R. (2004). *Enclosing the maximum likelihood of the simplest DNA model evolving on fixed topologies: towards a rigorous framework for phylogenetic inference*. BSCB Dept. technical report, BU-1653-M, Cornell University, Ithaca, New York.

Sainudiin, R. (2005). *Machine Interval Experiments*. pHd dissertation, Cornell University, Ithaca, New York.

Sainudiin, R., R. Durrett, C. Aquadro, and R. Nielsen (2004). Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. *Genetics 168*, 383–395.

Sainudiin, R., W. S. W. Wong, K. Yogeeswaran, J. Nasrallah, Z. Yang, and R. Nielsen (2005). Detecting site-specific physicochemical selective pressures: applications to the class-I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *Journal of Molecular Evolution 60*, 315–326.

Sainudiin, R. and T. York (2005). *An auto-validating rejection sampler*. BSCB Dept. technical report, BU-1661-M, Cornell University, Ithaca, New York.

Sainudiin, R. and R. Yoshida (2005). Applications of interval methods to phylogenetic trees. In L. Pachter and B. Sturmfels (Eds.), *Algebraic statistics for computational biology*. Cambridge, UK: Cambridge University Press.

Selby, R. G., F. J. Vecchio, and M. P. Collins (1997). The failure of an offshore platform. *Concrete International 19:(8)*.

Semple, C. and M. Steel (2003). *Phylogenetics*. Oxford, UK: Oxford University Press.

Wong, W. S. W., R. Sainudiin, and R. Nielsen (2005). Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *(submitted)*.