



Bioinformatics Seminar

Speaker: Gabor T. Marth, D.Sc., Boston College Dept. of Biology

Title: Genome variation informatics: SNP discovery, demographic inference and human haplotype structure

Date: Monday, 17 November 2003

Time & Location

Refreshments: 11 am in the Applied Mathematics Common Room at MIT's Building 2, Room 349

Talk: 11:30 am to 1 pm in the Applied Mathematics Conference Room Building 2, Room 338

URL: <http://www-math.mit.edu/compbiosem/>

Abstract:

Allelic difference causes phenotypic diversity, may lead to disease, and can be used as a tool to track long-term demographic events in our evolutionary past. The vast majority of sequence variations are single nucleotide polymorphisms (SNPs). To discover SNPs in extant sequence resources, colleagues at the Washington University Medical School and I have developed PolyBayes, a statistically rigorous, Bayesian polymorphism detection algorithm. This tool was used for the discovery of hundreds of thousands of SNPs contributing significantly to a comprehensive polymorphic marker map of the human genome. At present, high genotyping cost, and statistical difficulties make it impractical to use all these markers simultaneously in genome-wide association studies. To reduce the number of markers needed, an ambitious project (the HapMap initiative) is under way to resolve human haplotype structure, and to tabulate local regions termed haplotype blocks where haplotype diversity is limited to a few common haplotypes indexed by a small but sufficient subset of SNPs. With such a resource in place the analysis of complex traits will be framed in terms of haplotype structure across phenotypic classes. To ensure that this expensive resource is truly useful, many questions involving marker density, marker placement relative to functional units of DNA, allele frequency, and the question of generality across world populations need to be investigated.

Our initial assessment of genome-scale variation data shows differential demographic structure within large world populations. Both marker density and allele frequency data from European samples speak of a bottleneck shaped history characterized by a severe collapse of effective size followed by a phase of recovery. In contrast, African data are best explained by a history of slow but uninterrupted population expansion. Simulation studies that take into account these individual histories can predict important quantities such as the extent of linkage disequilibrium or the distribution of haplotype block size characteristic for each population. For example, African haplotype blocks are predicted to be roughly half the size of European blocks. But are the African blocks fragmented versions of the European blocks? Although many SNPs are present in all world populations, we know that many markers are specific to a single population, and are monomorphic or rare in others. In light of this, how strongly does overall haplotype structure depend on the selection of markers used in its definition? Are SNPs that were discovered in one population suitable for defining haplotypes in another, or must one use "population-matched" SNPs? Using demographic models distilled from extant variation data, and by placing the variation structure of different world populations into the unified frame of reference of common human genealogy, we evaluate competing marker selection strategies for capturing haplotype structure in all subpopulations involved. These considerations can aid marker selection for the public haplotype project and satisfy the need for generality within the resource.

Massachusetts Institute
of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

For General Questions, please contact kvdickey@mit.edu