Diffusion Approximations for Demographic Inference: ∂a∂i

Ryan N. Gutenkunst[*], Ryan D. Hernandez[†], Scott H. Williamson[‡], and Carlos D. Bustamante[§]

Models of demographic history (population sizes, migration rates, and divergence times) inferred from genetic data complement archeology and serve as null models in genome scans for selection. Most current inference methods are computationally limited to considering simple models or non-recombining data. We introduce a method [1] based on a diffusion approximation to the joint frequency spectrum (FS) of genetic variation between populations. Our implementation, ∂a∂i, can model up to three interacting populations and scales well to genome-wide data. We have applied ∂a∂i to human data from Africa, Europe, and East Asia, building the most complex statistically well-characterized model of human migration out of Africa to date.

Given sequence from individuals in each of P populations, the resulting FS is a P-dimensional matrix. Each entry records the number of SNPs in which the derived allele was found in the corresponding number of samples from each population. For example, the [2,0] entry records the number of derived alleles seen twice in population 1 and seen zero times in population 2. In the absence of linkage, the FS is a complete summary of the data. Furthermore, it is known that linkage does not bias demographic inference.

Efficient solution of the relevant multi-dimensional diffusion equation is challenging. In particular, finite-difference grids must be coarse to be tractable, but this can cause artifacts in the calculated FS. We overcome this by solving at three progressively finer grids, extrapolating to an infinitely fine grid. This dramatically increases both computational speed and accuracy.

As an application, we consider human migration out of Africa, using 5 Mb of noncoding sequence (generated by the Environmental Genome Project [2]) from each of 12 Yoruba (YRI), 22 CEPH European (CEU) and 12 Han Chinese (CHB) individuals. Figure 1 shows the data, along with the maximum-likelihood parameters and FS for our demographic model. Importantly, ∂a∂i's speed enables extensive bootstrapping for statistically characterizing the model, including estimating parameter uncertainties and significance values for hypothesis and goodness-of-fit tests.

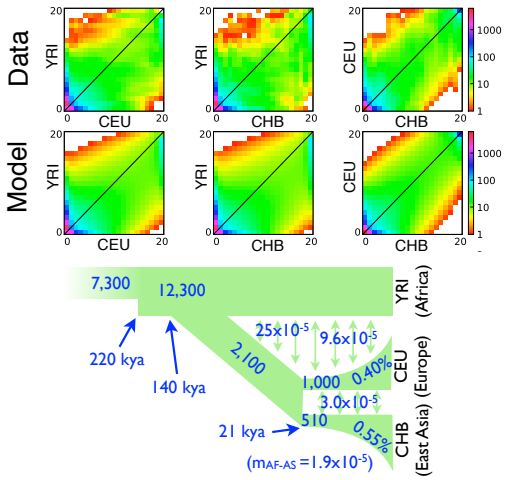

Figure 1: Genetic model for human expansion out of Africa [1]. Using ∂a∂i, the 14 free parameters were fit to 5 Mb of noncoding sequence. Parameter uncertainties are typically about 20%. (Parameters are drift effective sizes, divergence times in thousands of years ago, migration rates per chromosome per generation, and growth rates per generation.)

The method implemented in ∂a∂i is general and widely applicable. To our knowledge, ∂a∂i has been applied to humans [1, 3, 4], rice [5], orangutans [6], and cattle.

∂a∂i is implemented in Python and C, making extensive use of the NumPy and SciPy libraries. The code is publicly available at `http://dadi.googlecode.com`, under the New BSD License.

[1] Gutenkunst RN, et al. (2009) PLoS Genet 5:e1000695

[2] Livingston RJ, et al. (2004) Genome Res 14:1821

[3] Andrés AM, et al. (2009) Mol Biol Evol 26:2755

[4] Nielsen R, et al. (2009) Genome Res 19:838

[5] Xu X, et al. (submitted)

[6] Locke D, et al. (submitted)

[*]Theoretical Biology and Biophysics & Center for Nonlinear Studies, Los Alamos National Laboratory; `ryang@lanl.gov`

[†]Human Genetics, University of California, San Francisco

[‡]Biological Statistics and Computational Biology, Cornell University

[§]Genetics, Stanford University