

Estimation of gene conversion rates in small gene families

Proposer: Jay Taylor

Gene families are sets of genes, belonging to the same genome, which encode related proteins. Examples from the human genome include the MHC genes (involved in adaptive immunity), the opsins (involved in color vision), and the olfactory receptor genes. Because the de novo origin of proteins from non-coding sequence probably occurs exceptionally rarely, the gain and loss and diversification of members of gene families is thought to be an important source of evolutionary innovation.

Evolution of gene families involves a number of processes that are usually ignored when analyzing non-duplicated DNA sequences. One of these is ectopic gene conversion, which occurs when a stretch of DNA sequence located within one member of a gene family is copied onto the corresponding sites in another member. Ectopic gene conversion plays an important role in the evolution of some gene families, both because it can prevent the diversification of different members of a family (concerted evolution) and because it can assemble mosaic genes from bits and pieces of other gene family members.

Existing methods for estimating ectopic gene conversion rates are fairly crude, e.g., method of moments-based estimators. The aim of this project will be to develop a more efficient method by generalizing the two-locus composite likelihood (CL) approach to small gene families. CL-based methods work by approximating the full likelihood of a set of DNA sequences under an evolutionary model by the product of the two-locus likelihoods calculated using all pairs of variable sites. The rationale for this approach is that although it is practically impossible to calculate the full likelihood directly (the state space is too large), it is fairly easy to calculate two-locus likelihoods using coalescent simulations or importance sampling. To extend this method to population data for gene families consisting of two homologous genes, we will consider pairs of pairs of segregating sites, i.e., sites (i_1, j_1) and (i_2, j_2) from genes 1 and 2, respectively, where i_1 and i_2 or j_1 and j_2 are corresponding sites. Ectopic gene conversion will have two effects on this kind of data: in addition to decoupling the genealogies at different sites, it will also homogenize corresponding sites on different genes.

This project would best suit someone interested in evolutionary biology who either knows or is willing to learn to program in C/C++ or some comparable language.

References:

- J. H. Gillespie (2004) Population Genetics. A Concise Guide, Johns Hopkins University Press. (Chapters 1-4).
- R. R. Hudson (2001) Two-Locus Sampling Distributions and Their Application. *Genetics* 159: 1805-1817.
- H. Innan (2002) A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. *Genetics* 161: 865-872.
- H. Innan (2004) A two-locus gene conversion model with selection and its application to the

human RHCE and RHD genes. PNAS 100: 8793-8798.

G. McVean, P. Awadalla and P. Fearnhead (2002) A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. Genetics 160: 1231-1241.

K. R. Thornton (2007) The neutral coalescent process for recent gene duplications and copy-number variants. Genetics 177: 987-1000.