

Supervisors: Rune Lyngsoe, Jotun Hein

Project: User Interface for Recombination Analysis

Usually when we think about the evolution of life and speciation, we have in mind a tree like structure. Two groups of a species eventually become so distant that they become different species, creating a new bifurcation in the history of evolution. However, this view completely ignores recombination where a new individual is created by combining genetic material from two other individuals. Including recombination requires evolution to be described by a type of directed acyclic graph, called an *ancestral recombination graph* (ARG), rather than a tree. The most prominent case of recombination is probably in sexual reproduction, where the progeny inherits one chromosome in each of its chromosome pairs from each of its parents. The chromosomes inherited, however, are not exact copies of one of the parental chromosomes. Rather it is a mosaic of the parent in question's pair of that chromosome, created by recombination or cross-over events switching the copying back and forth between the two chromosomes of the pair. In fact, mechanisms exist to discard chromosomes where no recombination has taken place. But meiotic recombination, as it is called, is not the only incidence of recombination. It is believed that all organisms to a smaller or larger extent exchange genetic material by recombination. This includes viruses and bacteria, and recombination is believed to be a major factor in e.g. the emergence of multiresistant pathogens like MRSA and new strains of influenza.

In our group, we have developed software currently the most powerful in existence to infer recombinations by parsimony under the infinite sites model of substitutions. A first challenge of this project would probably be to figure out what the previous sentence means. To give a brief explanation, assume that we have sequenced the same gene from a number of individuals. Though the sequences will be mostly identical, there will be some bases that differs between any two individuals. Positions where not all individuals have the same base are called *segregating sites*. The infinite sites model of substitutions essentially means that each segregating site can be traced back to one particular point in time. At this point, a mutation substituted the original base in the segregating site with a different base in some individual (making it segregating). This mutation was then passed on and multiplied by inheritance to eventually appear in some of the individuals we sequenced. It is easy to prove, that under this assumption not all data sets allow a simple tree like description of their evolutionary history. Our software finds the minimum number of recombinations that are required in any evolutionary history of the input data, and an accompanying evolutionary history with this number of recombinations. Inferring evolutionary histories by finding one requiring a minimum number of events is known as parsimony inference.

Our software only has a simple command line interface with no user interaction. At the end of the computation, the minimum recombination history can be output as a description of the corresponding ARG in formats used by several graph visualisation programs. It would be much more useful to biological researchers, if they in a graphical rendering of the ARG could e.g. look at the ancestors present at a particular time slice, explore alternative histories, move substitution events around, introduce a particular recombination etc. Our software has the essential functionality needed for the behind-the-scenes work, but to use it for the above tasks would mean laboriously tracking changes to sequences e.g. in a text editor. What we would like you to do, is to develop a graphical user interface that allows the user to make these explorations by point-and-click. Our software is implemented in C. It is your choice of language for the user interface, but you may be able to build on one of the two graph visualisation tools, implemented in C and java respectively, that we can output ARG descriptions for. Any improvements to the way we infer evolutionary histories would of course also be welcome.

A successful completion of this project, going beyond a prototype and actually implementing a fully fledged user interface, should be publishable as an application note in the bioinformatics literature. This project is thus perfect if you care to have a scientific publication to your name.

Suggested reading

Beagle homepage at www.stats.ox.ac.uk/~lyngsoe/beagle/

Gene Genealogies, Variation and Evolution by J. Hein, M.H. Schierup, and C. Wiuf, Oxford University Press (2005)

Minimum Recombination Histories by Branch and Bound by R.B. Lyngsø, Y.S. Song, and J. Hein, proceedings of the 5th Workshop on Algorithms in Bioinformatics (2005)