CoalFace 0.2b Manual

Introduction

In population genetics, investigators are typically interested in the evolutionary history that can be inferred from population genetic data. However, one of the problems in this inference is the understanding of the potential data that could be generated given a particular demographic history. Since mating is essentially random, there are multiple possible genealogies that could result from the same demographic history. Furthermore, different genes may exhibit different gene genealogies dependent on their mutation rate, mode of inheritance and levels of recombination. It is this stochastity of the coalescent process that is imperative to understand when making inferences from population genetic data.

J.F.C. Kingman first introduced and assimilated coalescent theory (Kingman 1982), and since then several have contributed extensively to its development. This includes the work of R.R. Hudson, F. Tajima, S. Tavaré, R.C. Griffiths, P. Donnelly, J. Felsenstein and J. Wakeley. Since then several authors have developed population genetic analyses procedures that utilise the coalescent to draw inferences regarding demographic history from population genetic data. The software package presented here was initially developed as a teaching tool for students of population genetics. However, some basic analyses can also be conducted using the simulated data derived from this software program. I hope that you find this program both useful and fun in teaching coalescence to both undergraduate and postgraduate students.

Installation

Windows XP, 2000, NT

To install CoalFace on a windows based machine simply extract the .zip archive to somewhere sensible (eg: c:/Program Files/CoalFace) and then copy the file: qintf70.dll to your windows system directory (/windows/system32 or /WINNT/system32). The .dll file is the dynamic shared libraries used by the delphi script. It is necessary since this program has been coded as a cross-platform CLX application in Kylix, and makes use of the Kylix QT libraries instead of the native Windows GUI procedures. Then simply run the program by executing the CoalFace.exe file.

Linux

To install on a Linux based machine simply extract the tarball, using something like

$ tar -xvzf CoalFace.tar.gz

This should create a directory called CoalFace, where you will find some

subdirectories. The shell script CoalFace executes the program. Run this by simply typing:

$./CoalFace

Simple Genealogy simulation

In order to generate simple simulations of the coalescent process, under the *main* tab choose a sample size, k, and a population size, N. Remember, since you are simulating the genealogical history of k samples in a population size of N, you cannot have a sample size larger than k. Typically, a simulation would have a sample size of 50, with a population size of 10000-100000 (However, see note about theta). If you click *Run Simulation* you will be provided with a random simulation of the coalescent process. This coalescent tree or coalescent genealogy can be saved, as a bitmap file, by clicking the save button. The genealogy is also output to a tree file, defined under the output tab, in Newick tree format. This file can then be viewed with Rod Page's TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). If you choose *multiple simulations*, the coalescent simulation will be repeated and the Newick trees for each simulation will be written to the tree output file. CoalFace will draw the genealogy of the last simulation. However, by using the forward and back buttons one can flip through all the genealogies to get an idea of how much variation there is in the structure of a genealogy under the same population parameters.

Genealogy simulation with mutations

The default model is one where mutations are scattered on the genealogy according to a given mutation rate in mutations per site per generation. You can choose not to scatter mutations by deselecting *Export Segregating sites from infinite sites model* and *Scatter Mutations on Simulated DNA sequence*. The selection of the former will only scatter mutations on the genealogy, which in turn can be drawn on the genealogy in red by selecting *Draw mutations on genealogy.* Whereas, the selection of the latter will produce a set of simulated DNA sequences of the defined length, under the given substitution model. Allelic and nucleotide diversities for these sequences can be calculated by selecting *Calculate Diversity Indices* under the sequence tab. If you select *Generate sequence output file* these sequences will be output to a file defined in *Sequence Data output* under the *Output* tab in the required format. The substitution model default is a Jukes-Cantor, however F81, Kimura 2 parameter and HKY85 can also be implemented with or without gamma distrubuted rate variation among sites, and a proportion of invariable sites. The default setting is for the root sequence to be created at random under the base frequencies of the given model, but an input or hypothetical root sequence can be provided by checking *Input DNA sequence* and providing a text file. You may also simulate diploid data, for example nuclear gene sequences. In this case the number of samples will be twice that of the number of individuals output to the sequence output file. Remember to ensure your sample size is divisible by two, such that alleles can be assigned to individuals. All the alleles are first simulated in a single

genealogy, and alleles are then randomly assigned to individuals for output in the data file.

The simulation of DNA sequences is set as the default, yet microsatellite data can also be simulated by checking *Simulate Microsatellite data* under the *Microsatellite* tab. Again, since microsatellite data is diploid, you should ensure that your sample size, k, is divisible by two. The number of loci should be provided, and a microsatellite parameters file. This file defines the range of allele sizes, the type of repeat motif and the mutation rate for each microsatellite locus. Each locus is essentially an independent run of the simulation since complete linkage equilibrium is assumed. The format of the parameters file is as folllows:

[locus #] [Smallest allele] [Largest allele] [repeat motif] [mutation rate]

example:
1 122 148 2 0.001
2 225 269 4 0.0001
3 108 128 2 0.00001
4 105 150 5 0.00001
5 86 116 3 0.0001

The microsatellite mutation model is set as a default to an infinite alleles model, yet both a stepwise mutation and random allele model can be selected.

Simulation of demographic history

A very simple simulation of demographics is implemented in CoalFace. One can simulate variance in reproductive success, and an increasing or decreasing population size. *Variance in reproductive success* should result in quicker times to coalescence since fewer individuals in the population are contributing to the next generation. *Variance in reproductive success* is implemented by adjusting the population size according to Crow (1954). A period of population growth or decline (by setting the growth percentage as negative) can be implemented by setting the starting and ending generations of growth. Be careful not to make the % *population size change per generation* too large since the growth is exponential and can very quickly go beyond an integer that your computer can manage. A more advanced demographic model simulation is provided by Laurent Excoffier in the program SIMCOAL (http://cmpg.unibe.ch/software/simcoal/). This program allows for the simulation of data under custom defined demographic expansion models.

A note about Theta

Population geneticists are generally more interested in the value of theta, than in the population size or mutation rate. This is simply since both population size and mutation rate are difficult to estimate from genetic data. Therefore, a composite parameter, theta (2*N*mutation rate for haploid genes) is estimated. When running multiple simulations, or any simulations that generate data via

mutations, you should think about specifying a value for theta, rather than by specifying population size and mutation rate independently. For example a population size of 100 000 and a mutation rate of 10^{-6}, provides exactly the same value of theta as a population size of 10 000 with a mutation rate of 10^{-5}. If you are interested in obtaining estimates of molecular diversity then either combination would give similar results. However, due to the manner in which CoalFace is coded, it takes substantially longer for simulations of large population sizes and small mutation rates, than vice versa. Therefore, always try to implement theta by keeping population sizes low (1000-10000).

Program Outputs

The output files are provided under the *Files* tab. By selecting a working directory you can ensure all output files are directed to the given directory. This is useful on linux systems where the user may not have write access rights to the directory in which the program resides. The output file contains all the information aquired from multiple runs: Tmrca, number of segregating sites under infinte and finite models and diversity indices. This file can be imported into excel or matlab to draw distributions of these statistics under the simulated demographic parameters. Another option is to create Arelquin files for the simulated data. If you are running multiple simulations, this will create an Arlequin batch file, and the accompanying data files for the number of simulations. These can be analysed in Arlequin (http://lgb.unige.ch/arlequin/), and summaries of molecular diversity indices generated over multiple simulations. All the arlequin files will be saved to a created directory called Arlequin, in the CoalFace directory. Finally, a logfile is created that logs all the parameter settings chosen in the simulation run.

Coalescent Literature

Here are some good reviews of the coalescent, and its application to population genetic data.

Donnelly, P & Tavaré, S. 1995. Coalescents and genealogical structure under neutrality. Ann. Rev. Gen. 29: 401-421.

Hudson, RR. 1991. Gene genealogies and the coalescent process. In D. Futuyma & J. Antonovics (eds). Oxford Surveys in Evolutionary Biology, Vol 7: 1-44.

Fu, YX & Li, WH. 1999. Coalescing into the 21st Century: An overview and prospects of coalescent theory. Theoretical Population Biology 56: 1-10.

Nordborg, M. 2003. Coalescent theory. In. DJ. Balding, M. Bishop & C. Cannings (eds). Handbook of Statistical Genetics, 2nd edition. John Wiley & Sons, Ltd.

Rosenberg, NH. & Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics 3: 380-390.

Felsenstein, J. 2003. Coalescent Trees. In J. Felsenstein. Inferring Phylogenies. Sinauer Associates.

References

Crow, JF. 1954. Breeding structure of populations. II. Effective population number, pp 543-556. In O. Kempthorne, T. Bancroft, J. Gowen & J. Lush (eds). Statistics and Mathematics in Biology. Ames, Iowa State University Press.

Kingman, JFC. 1982. The Coalescent. Stoch. Proc. Appl. 13: 235-248.

Authors

Wayne Delport & Michael Cunningham
Molecular Ecology and Evolution Program
Department of Genetics
University of Pretoria
Pretoria
0002

bug report: wdelport@postino.up.ac.za