BMC Bioinformatics

BioMed Central

Software

Open Access

Recodon: Coalescent simulation of coding DNA sequences with recombination, migration and demography

Miguel Arenas* and David Posada

Address: Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, 36310 Vigo, Spain

Email: Miguel Arenas* - miguelab@uvigo.es; David Posada - dposada@uvigo.es

* Corresponding author

Published: 20 November 2007

BMC Bioinformatics 2007, **8**:458 doi:10.1186/1471-2105-8-458

Received: 2 August 2007

Accepted: 20 November 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/458

Abstract

Background: Coalescent simulations have proven very useful in many population genetics studies. In order to arrive to meaningful conclusions, it is important that these simulations resemble the process of molecular evolution as much as possible. To date, no single coalescent program is able to simulate codon sequences sampled from populations with recombination, migration and growth.

Results: We introduce a new coalescent program, called *Recodon*, which is able to simulate samples of coding DNA sequences under complex scenarios in which several evolutionary forces can interact simultaneously (namely, recombination, migration and demography). The basic codon model implemented is an extension to the general time-reversible model of nucleotide substitution with a proportion of invariable sites and among-site rate variation. In addition, the program implements non-reversible processes and mixtures of different codon models.

Conclusion: *Recodon* is a flexible tool for the simulation of coding DNA sequences under realistic evolutionary models. These simulations can be used to build parameter distributions for testing evolutionary hypotheses using experimental data. Recodon is written in C, can run in parallel, and is freely available from http://darwin.uvigo.es/.

Background

Coalescent theory [1] provides a very powerful framework for the simulation of samples of DNA sequences. Coalescent simulations can be very useful to understand the statistical properties of these samples under different evolutionary scenarios [2], to evaluate and compare different analytical methods [3], to estimate population parameters [4] and for hypothesis testing [5]. Not surprisingly, several simulation programs have recently been developed under this framework [6-12]. In order to obtain meaningful biological inferences from simulated data it is important that the generating models are as realistic as possible. However, increasing model complexity usually results in longer computing times, and most programs usually focus on a restricted set of biological scenarios. Currently, we lack a tool for the simulation of samples of coding sequences that have evolved in structured populations with recombination and fluctuating size, typical for example of fast evolving pathogens and MHC genes [13,14]. Here, we introduce a new simulation program, called *Recodon*, to fill this gap.

Implementation

The simulation of data in *Recodon* is accomplished in two main steps. First, the genealogy of the sample is simulated under the coalescent framework with recombination, migration and demographics. Second, codon sequences

are evolved along this genealogy according to a nucleotide or codon substitution model.

Simulation of genealogies

For each replicate, genealogies are simulated according to thecoalescent under a neutral Wright-Fisher model [15,16]. Waiting times to a coalescence, recombination or migration event are exponentially distributed, and depend on the number of lineages, effective population size (N), recombination, migration and growth rates. Time is scaled in units of $2N$ generations. Recombination occurs with the same probability between different sites (either nucleotides or codons). A finite island model [16,17] is assumed, where migration takes place at a constant rate between different demes. Multiple demographic periods can be specified, each one with its own initial and final effective population size, and length (number of generations). Positive or negative exponential growth is assumed.

Simulation of nucleotide and codon sequences

Recodon implements several nucleotide and codon models that include different parameters (Table 1). The most complex nucleotide model implemented is the general time non-reversible model (GTnR; extended from Tavaré [18]), while the most general codon model is GY94∞GTnR_3∞4, which is the Goldman and Yang codon model [19], crossed with GTnR, and with codon frequencies predicted from the nucleotide frequencies at each codon position. Usually, the sequence at the root (most recent common ancestor or MRCA) is built according to the equilibrium frequencies, but the user has the option of specifying its own sequence. Note that in the presence of recombination, such sequence is just a concatenation of the MRCA sequences for the different recombinant fragments.

Table 1: Key arguments for Recodon. The user can specify several parameters to implement different simulation scenarios. These arguments can be entered in the command line or read from a text file.

Parameter	Example value	Application
Number of replicates	1000	All
Sample size	12	All
Number of sites (bp or codons)	3000	All
Effective population size	1000	All
Exponential growth rate	2.1×10^{-5}	Demography
Demographic periods[1]	1000 5000 200	Demography
Recombination rate	5×10^{-6}	Recombination
Migration rate	1.2×10^{-4}	Migration
Number of demes	4	Migration
Mutation rate	5.1×10^{-4}	All
Nucleotide frequencies[2]	0.4 0.3 0.1 0.2	Nuc/codon models
Transition/transversion ratio	2.1	Nuc/codon models
Relative substitution rates	1.0 2.3 2.1 3.0 4.2 1.0	Nuc/codon models
Nonsynonymous/ synonymous rate ratio[3]	1.8	Codon models
Rate variation among sites[4]	0.5	Nuc/codon models
Proportion of invariable sites	0.2	Nuc/codon models

[1]from 1000 to 5000 effective size during 200 generations.
[2]can be specified for each codon position in codon models (3 × 4).
[3]*dN*/*dS*.
[4]shape of the gamma distribution.

Program input

The input of the program consists of a series of arguments that can be entered in the command line or, more conveniently, specified in a text file (Table 1). These arguments fully parameterize the simulations, and control the amount of information that is sent to the console or output files.

Program output

The principal output of the program is a set of sampled aligned nucleotide or codon sequences in sequential Phylip format. Additional information that can be saved to different files includes the genealogies, divergence times, breakpoint positions, or the ancestral sequences. Replicates can be filtered out depending on the number of recombination events, and an independent outgroup sequence can also be evolved. At the end of the simulations, a summary of the different events is printed to the console.

Results and Discussion

We have developed a new program, called *Recodon*, for the simulation of coding DNA sequences. The program can run in parallel over multiple processors using the MPI libraries. The models implemented imitate the simultaneous action of several evolutionary processes, like recombination, migration, non-constant population size or selection at the molecular level. Understanding the joint effects of these processes is important in order to obtain more realistic estimates of population genetic parameters from real data [3,20-22].

Program validation

Recodon has been validated in several ways. The output of the program was contrasted with the theoretical expectations for the mean and variances for different values, like the number of recombination and migration events, or the times to the most recent common ancestor [23]. In addition, results obtained with *Recodon* were in agreement with those obtained with other programs [10] under different evolutionary scenarios. Finally, substitution and codon model parameters were estimated from the simulated data using

HYPHY [24] and PAUP*[25]. The average parameter estimates from these programs agreed very well with the expected values from the simulations.

Application

Coalescent simulations like those implemented in *Recodon* can be used to generate numerical expectations for different parameters under complex evolutionary scenarios, in which different processes interact in a simultaneous fashion. This can be very important to understand the interaction of different parameters, which complicates enormously their estimation [3]. Indeed, realistic simulation models are essential to evaluate different methods and strategies for estimating parameters and testing hypotheses from real data.

One potential application of *Recodon* could be the study of fast-evolving pathogens like HIV-1, which show high recombination and adaptation rates for coding genes [26]. For example, we could use this program to understand whether intrapatient genetic diversity for the *env* gene should increase with decreasing migration rates. Then we could test whether the number and diversity of *env* haplotypes sampled from a patient, all other conditions equal, ressemble the simulated cases with (or without) compartmentalization. Simulated data can also be used to obtain numerical estimates of population genetic parameter using approximate Bayesian computation [4,27-30]. Estimation by simulation can be especially useful in situations where the likelihood for a model is not known, or is computationally prohibitive to evaluate, which is often the case under complex biological scenarios.

In addition, we carried out a very simple experiment to illustrate another possible use of *Recodon*. In particular, we studied the effect of population structure on the footprint of molecular adaptation. Results suggest that population subdivision tends to increase both *dN* and *dS* divergences, as a result of longer times to the most recent common ancestor (Figure 1). This increase is similar in magnitude, and the *dN*/*dS* ratio is not affected by different migration rates when the simulated value is below one or one, but there seems to be a slight increase when the simulated *dN*/*dS* is 10.

Future development

In the future we plan to relax some of the current assumptions, like an homogeneous recombination rate [31].

Conclusion

Recodon is a versatile program for the simulation of codon alignments under complex population models. This program fills a gap in the current array of coalescent programs for the simulation of DNA sequences, as no single program is able to simulate codon sequences sampled from populations with recombination, migration and growth. Data simulated with this program can be used to study both theoretical and empirical properties of DNA samples under biologically realistic scenarios.

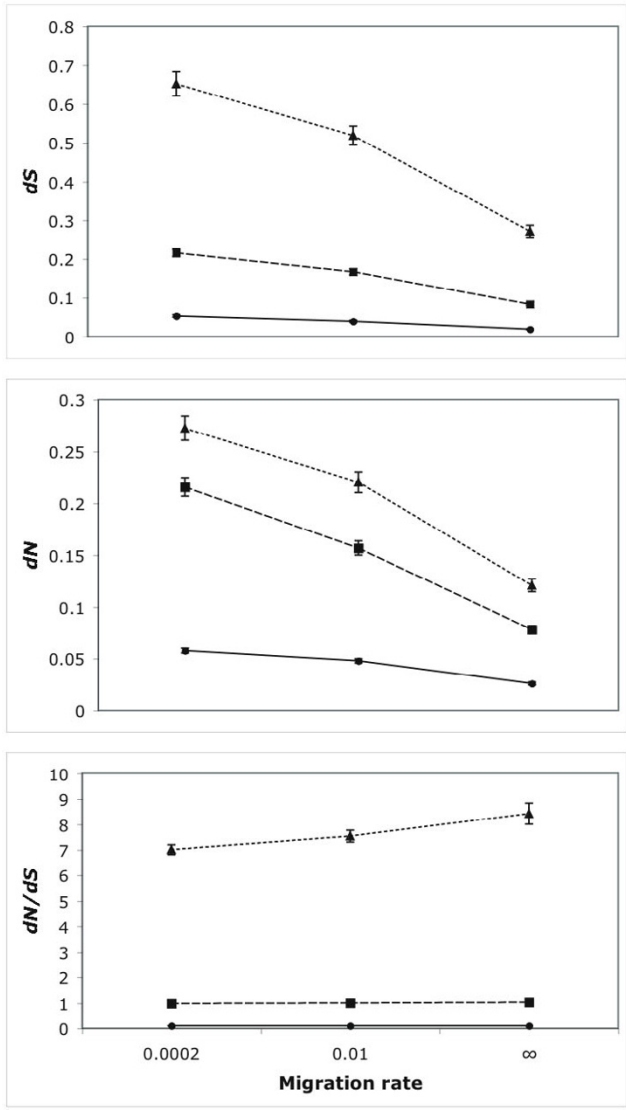

Figure 1

Effect of population structure on the estimation of synonymous and nonsynonymous divergence. Nine different scenarios were simulated, combining three migration rates (m = 0.0002, 0.01 and ∞ (= one deme)) and three dN/dS ratios (dashed line = 0.1, solid line = 1, dotted line = 10). For each scenario, 500 alignments with 10 sequences 333 codons long, were simulated. In all cases, the mutation rate was $5.4 \infty 10^{-5}$, the transition/transversion ratio was 1.0, and the effective population size was 1000. Mean synonymous divergence per synonymous site (dS), nonsynonymous divergence per nonsynonymous site (dN), and their ratio (dN/dS) were estimated according to Nei and Gojobori [32] with a modified version of SNAP [33]. Error bars indicated approximate 95% confidence intervals (± s.e. ∞ 1.96).

Availability and requirements

Recodon is written in ANSI C, and it has been compiled without problems in Mac OS X, Linux Debian and Windows. It can run in parallel using the MPI libraries in architectures with several processors. The program is freely available at http://darwin.uvigo.es/, including executables, source code and documentation. The program is distributed under the GNU GPL license.

Authors' contributions

Recodon is an extension of a coalescent program written by DP, who conceived the idea and supervised its development. MA wrote and validated the program. Both authors drafted the manuscript, and both read and approved its final version.

Acknowledgements

This work was partially supported by grant BFU2004-02700 (MCyT) to DP and by the FPI fellowship BES-2005-9151 (MEC) to MA from the Spanish government. Several functions were taken from code provided by R. Nielsen and Z. Yang. We want to thank J. Carlos Mouriño at the Supercomputing Center of Galicia (CESGA) for extensive help with code parallelization.

References

1. Kingman JFC: **The coalescent.** *Stochastic Processes and their Applications* 1982, **13:**235-248.
2. Innan H, Zhang K, Marjoram P, Tavare S, Rosenberg NA: **Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites.** *Genetics* 2005, **169(3):**1763-1777.
3. Carvajal-Rodriguez A, Crandall KA, Posada D: **Recombination Estimation Under Complex Evolutionary Models with the Coalescent Composite-Likelihood Method.** *Mol Biol Evol* 2006, **23(4):**817-827.
4. Beaumont MA, Zhang W, Balding DJ: **Approximate Bayesian computation in population genetics.** *Genetics* 2002, **162(4):**2025-2035.
5. DeChaine EG, Martin AP: **Using coalescent simulations to test the impact of quaternary climate cycles on divergence in an alpine plant-insect association.** *Evolution Int J Org Evolution* 2006, **60(5):**1004-1013.
6. Excoffier L, Novembre J, Schneider S: **SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography.** *J Hered* 2000, **91:**506-509.
7. Spencer CC, Coop G: **SelSim: a program to simulate population genetic data with natural selection and recombination.** *Bioinformatics* 2004, **20(18):**3673-3675.
8. Mailund T, Schierup MH, Pedersen CN, Mechlenborg PJ, Madsen JN, Schauser L: **CoaSim: a flexible environment for simulating genetic data under coalescent models.** *BMC Bioinformatics* 2005, **6:**252.
9. Marjoram P, Wall JD: **Fast "coalescent" simulation.** *BMC Genet* 2006, **7:**16.
10. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation.** *Bioinformatics* 2002, **18(2):**337-338.
11. Hellenthal G, Stephens M: **msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots.** *Bioinformatics* 2007, **23(4):**520-521.
12. Posada D, Wiuf C: **Simulating haplotype blocks in the human genome.** *Bioinformatics* 2003, **19(2):**289-290.
13. Edwards SV, Hedrick PW: **Evolution and ecology of MHC molecules: from genomics to sexual selection.** *Trends in Ecology and Evolution* 1998, **13(8):**305-311.
14. Awadalla P: **The evolutionary genomics of pathogen recombination.** *Nat Rev Genet* 2003, **4(1):**50-60.
15. Fisher RA: **The Genetical Theory of Natural Selection.** Oxford: Oxford University Press; 1930.
16. Wright S: **Evolution in Mendelian populations.** *Genetics* 1931, **16:**97-159.
17. Hudson RR: **Island models and the coalescent process.** *Mol Ecol* 1998, **7:**413-418.
18. Tavaré S: **Some probabilistic and statistical problems in the analysis of DNA sequences.** In *Some mathematical questions in biology – DNA sequence analysis Volume 17*. Edited by: Miura RM. Providence, RI: Amer. Math. Soc; 1986:57-86.
19. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994, **11(5):**725-736.
20. Anisimova M, Nielsen R, Yang Z: **Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites.** *Genetics* 2003, **164(3):**1229-1236.
21. Shriner D, Nickle DC, Jensen MA, Mullins JI: **Potential impact of recombination on sitewise approaches for detecting positive natural selection.** *Genet Res* 2003, **81:**115-121.
22. Posada D: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** *Mol Biol Evol* 2002, **19(5):**708-717.
23. Hudson RR: **Gene genealogies and the coalescent process.** *Oxf Surv Evol Biol* 1990, **7:**1-44.
24. Kosakovsky Pond SL, Frost SD, Muse SV: **HYPHY: Hypothesis testing using phylogenies.** *Bioinformatics* 2005, **21:**676-679.
25. Swofford DL: **PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods).** 4th edition. Sunderland, Massachusetts: Sinauer Associates; 2000.
26. Rambaut A, Posada D, Crandall KA, Holmes EC: **The causes and consequences of HIV evolution.** *Nature Review Genetics* 2004, **5:**52-61.
27. Excoffier L, Estoup A, Cornuet JM: **Bayesian analysis of an admixture model with mutations and arbitrarily linked markers.** *Genetics* 2005, **169(3):**1727-1738.
28. Tanaka MM, Francis AR, Luciani F, Sisson SA: **Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data.** *Genetics* 2006, **173(3):**1511-1520.
29. Tallmon DA, Luikart G, Beaumont MA: **Comparative evaluation of a new effective population size estimator based on approximate bayesian computation.** *Genetics* 2004, **167(2):**977-988.
30. Shriner D, Liu Y, Nickle DC, Mullins JI: **Evolution of intrahost HIV-1 genetic diversity during chronic infection.** *Evolution Int J Org Evolution* 2006, **60(6):**1165-1176.
31. Wiuf C, Posada D: **A coalescent model of recombination hotspots.** *Genetics* 2003, **164(1):**407-417.
32. Nei M, Gojobori T: **Simple method for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3(5):**418-426.
33. Korber B: **HIV Signature and Sequence Variation Analysis.** In *Computational Analysis of HIV Molecular Sequences* Edited by: Rodrigo AG, Learn GH. Dordrecht, Netherlands: Kluwer Academic Publishers; 2000:55-72.