EvolveAGene 3: A DNA coding sequence evolution simulation program

Barry G. Hall
Bellingham Research Institute
Bellingham, WA

drbh@mail.rochester.edu

Availability: Mac OS X and Windows executables and Perl source code plus a detailed manual from http://homepage.mac.com/barryghall/Software.html

Abstract
EvolveAGene 3 is a realistic coding sequence simulation program that separates mutation from selection and allows the user to set selection conditions, including variable regions of selection intensity within the sequence and variation in intensity of selection over branches. Variation includes base substitutions, insertions and deletions. Output includes a log file, the true tree and both unaligned coding sequence and protein sequences and the true DNA and protein alignments.

1. Introduction

A variety of methods and computer programs to implement those methods exist for the purpose of reconstructing evolutionary histories from molecular sequence data. Those methods are designed to reconstruct phylogenetic trees, ancestral sequences of common ancestors at the nodes of those trees, and to tease out aspects of adaptive evolution -both positive and purifying selection - along the branches of those trees. All of those methods depend upon data in the form of multiple alignments of the molecular sequences, and a variety of programs exist to estimate those multiple alignments. The problems of multiple alignment and phylogenetic reconstruction are very intertwined; phylogenetic reconstruction depends on correct multiple alignments and multiple alignment reconstruction depends upon correct phylogenetic reconstruction. There are several sets of proteins that have been aligned on the basis of tertiary structures of the proteins. Those protein alignments [1-3] and their corresponding DNA coding sequence alignments [4], while they are not "true" alignments, now serve as gold standards for assessing the accuracies of multiple sequence alignments. There is little in the way of known "true" phylogenies or multiple alignments that can be used to measure and compare the accuracies of phylogenetic methods and programs. There are a few experimental evolution studies that provide known phylogenies [5], but the small scale of the data sets limits their usefulness as test beds for comparing methods and programs. In the absence of known phylogenies, ancestral sequences, etc. simulations must be used to test the accuracies of the methods and programs.

Simulation programs generate true trees that can be compared with estimated trees, and can generate true alignments that can be compared with estimated alignments. The value of simulated data

sets is directly proportional to their biological realism. Unfortunately, most simulation programs are not very realistic. Genes evolve by nucleotide substitutions, insertions and deletions. Most programs, such as the popular Seq-gen [6] and Evolver [7], do not include insertions and deletions that result in gaps in alignments. When simulated data does not include insertions and deletions there is no need to align the sequences, eliminating an important step that contributes significantly to topological errors in phylogenetic trees. ROSE [8] and DAWG [9] are exceptions, and both employ mathematical models of sequence evolution (HKY, GTR, etc.). Since the evolutionary models are pre-defined, the methods and models that return the best trees, those most similar to the true tree, are the methods and models that best match the assumptions of the models used in the simulation.

In reality sequence evolution is a two step process in which spontaneous mutations occur and are then fixed into populations by selection and drift. Mutations occur as the results of replication and DNA repair errors and the relative proportions of the various kinds of mutations are called the *mutation spectrum*. Among those mutations that are base substitutions there are different proportions of each kind of substitution; e.g. AT to TA, AT to CG, etc. Within insertions and deletions there are different proportions of the various lengths. Selection acts on those mutations and greatly changes the proportions of the various mutations that are fixed into populations and thus appear in DNA sequences; i.e. the observed proportion of transitions relative to transversions is usually very different from their proportions in the mutational spectrum [10, 11].

EvolveAGene 3 is a sequence evolution simulation program that closely mimics real evolution by separating mutation from selection, by including insertions, deletions and base substitutions based upon a known mutational spectrum and by allowing the user to realistically specify the way selection operates on those mutations. In addition, EvolveAGene allows the user to specify that some region of a gene will be strongly conserved or that a region will be subject to positive (diversifying) selection. Similarly, the program can simulate differing adaptive constraints during the history of the sequences by allowing the user to specify that some branches will be subject to more intense purifying selection or to positive selection.

2. Program Overview

The user specifies the number of taxa (sequences) to be evolved. To create the true tree EvolveAGene takes a user-specified topology then assigns to each branch a random branch length that is between zero and twice the user-defined mean branch length. The topology must be strictly bifurcating, which is biologically realistic because we believe that speciation is a bifurcating process even though we may sometimes be unable to resolve the branching order. Because zero-length branches are permitted trifurcations do occur, although they are rare. The topology may be balanced or random; random in the sense that each branch has an equal probability of leading to an external (terminal or leaf) node or to an internal node. Alternatively, the topology may be supplied as an input tree, allowing the user to specify any desired topology.

Once the tree is determined EvolveAGene starts with a user specified sequence, usually an actual coding sequence, and moves outward from that root sequence along each branch for the required number of steps. At each step a random site in the sequence is chosen and EvolveAGene proposes a mutation according to the spontaneous mutational spectrum of *Escherichia coli*, which is better understood than that of any other organism [12, 13]. The probabilities of proposing each kind of mutation are discussed in [10]. The selection portion of the simulation consists of determining whether the proposed mutation will be accepted.

If the mutation is an insertion or deletion (an indel) it is rejected if the proposed indel length is not a multiple of three because, in reality, such frameshift mutations almost always result in loss of function and are not fixed into populations. Non-frameshift indels are accepted with probabilities that are specified by the user. The user can thus generate a data set that is as "gappy" as is desired.

If the mutation is a base substitution that does not result in an amino acid replacement it is accepted with a probability of that is set by the user, usually 1.0. If a base substitution mutation results in

a termination codon it is rejected because nonsense mutations almost universally result in loss of function. Otherwise it is accepted with a probability that is specified by the user. The user specifies the probability of accepting a replacement *relative to the probability of accepting a silent mutation.* The user specified value turns out to be very close to the resulting dN/dS ratio of the data set. The user can thus specify a realistic probability of accepting replacements based on actual dN/dS ratios from the literature.

A survey of 113 coding-sequence alignments [4] derived from the BaliBase set of tertiary-structure based alignments [3, 14, 15] in BaliBase sets 11, 12, 20 and 30 using the program codeml, part of the Paml suite [7] was used to evaluate typical dN/dS ratios. Figure 1 shows a histogram of dN/dS ratios. One data set, not shown on the histogram, was an extreme outlier with dN/dS = 1.41; i.e. those genes had been under positive selection. The remaining 112 data sets had a mean dN/dS ratio of 0.057±0.007, and a median of 0.016. 90% were < 0.191 and 90% were > 0.0075. The median was chosen as the default setting, and values between 0.0075 and 0.191 can certainly be considered realistic.

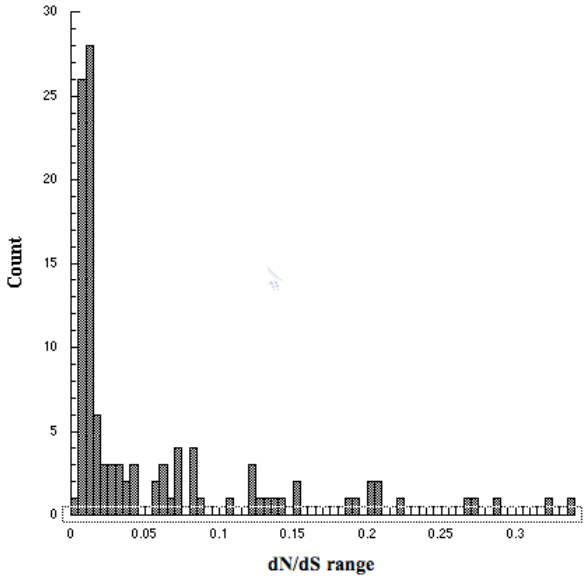

Figure 1

By default EvolveAGene assumes that amino acid replacement mutations are subject to a constant level of selection across the entire gene. In reality we know that some regions of a gene can be highly conserved, e.g. near an active site of an enzyme, while others can be quite tolerant of amino acid substitutions. Likewise, by default, EvolveAGene assumes that selection on amino acid replacements is constant over all of the branches; i.e. over time. In reality we know that there can be periods when a gene is subject to more intense negative (purifying) selection than at other times, and there can be times when it is subject to positive (diversifying) selection that favors amino acid replacements.

The user can over-ride those simplifying assumptions by choosing variable selection. If variable selection is chosen the user specifies the magnitude of either positive selection or more intense purifying selection and a random 10% segment of the gene is subjected to that selection.

Similarly, the user may specify that selection conditions are constant over branches or that they vary. If they vary a random 10% of the branches will be subjected to either positive or more intense purifying selection in a manner similar to that described above.

Proposed amino acid replacement mutations are thus accepted or rejected according to the probability that is set and, if variable selection is chosen, according to whether the mutation is on a branch or in a region that is subject to variable selection. If a mutation is rejected the process begins again by proposing a new mutation at a new site. When the required number of mutations have been accepted, according to the length of that branch, the resulting sequence is recorded either as an internal node or as

an external (leaf) node. As indels are accepted all of the sequences are updated so that the program maintains the true alignment of all sequences.

3. Input, output and specification of simulation conditions

The input is a text file that contains a simple coding sequence.

Ten output files are written. These include (1) a log file that records the simulation conditions, details of the mutations along each branch, specific regions and/or branches that have been subjected to positive or more intense purifying selection, and the total number insertions, deletions, silent and replacement base substitutions; (2) & (3) true trees, in Newick format, as rooted and unrooted trees with branch lengths and with and without interior node numbers as node labels; (4) & (5) DNA and corresponding protein sequences of the external (leaf) nodes; (6) & (7) DNA and protein sequences of the internal nodes; (8) & (9) the true alignments of the external and internal DNA node sequences; and (10) the true alignment of the external node protein sequences. The user may choose to have alignment in Fasta, Phylip (relaxed) or Nexus formats. True trees can be compared with estimated trees to evaluate phylogenetic methods and models, and true alignments can be compared with estimated alignments to evaluate alignment programs. Sequences of internal nodes are useful for evaluating ancestral sequence reconstruction programs.

The simulation conditions may be set from the command line or may be set interactively via a Phylip-like menu.

4. How realistic are the data sets simulated by EvolveAGene

Earlier versions of EvolveAGene have been used to compare the accuracies of some phylogenetic methods [10] and to assess a method for estimating ancestral sequences in deep phylogenies [16]. The underlying approach of EvolveAGene is very different from that of other sequence evolution simulation programs that depend upon mathematical models of nucleotide and amino acid substitutions. It is hoped that EvolveAGene's mutation and selection approach results in more realistic results, but there is no assurance that this is the case. Data sets generated by Rose, DAWG and EvolveAGene all look quite realistic at first glance. A study is underway to evaluate the realism of simulated data sets by comparison with real "gold standard" data sets in the BaliBase collection.

Literature Cited

1. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-7 (2004).
2. Raghava, G. P., Searle, S. M., Audley, P. C., Barber, J. D. & Barton, G. J. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics 4, 47 (2003).
3. Thompson, J. D., Koehl, P., Ripp, R. & Poch, O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61, 127-36 (2005).
4. Carroll, H. et al. DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins. Bioinformatics (2007).
5. Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. & Molineux, I. J. Experimental phylogenetics: generation of a known phylogeny. Science 255, 589-92 (1992).
6. Rambaut, A. & Grassly, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13, 235-8 (1997).
7. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13, 555-6 (1997).
8. Stoye, J., Evers, D. & Meyer, F. Rose: generating sequence families. Bioinformatics 14, 157-63 (1998).

9. Cartwright, R. A. DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics 21 Suppl 3, iii31-8 (2005).
10. Hall, B. G. Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences. Mol.Biol.Evol. 22, 792-802 (2005).
11. Li, W.-H. Molecular Evolution (Sinauer Assoc., Sunderland, MA USA, 1997).
12. Glickman, B. W., Burns, P. A. & Fix, D. F. in Antimutagenesis and Anticarcinogenesis Mechanisms (eds. Shankel, D. M., Hartman, P. E., Kada, T. & Hollender, A.) 259-281 (Plenum Press, New York, 1986).
13. Hall, B. G. The spectra of spontaneous growth-dependent and adaptive mutations in *ebgR*. J. Bacteriol. 181, 1149-1155 (1999).
14. Bahr, A., Thompson, J. D., Thierry, J. C. & Poch, O. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res 29, 323-6 (2001).
15. Thompson, J. D., Plewniak, F. & Poch, O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15, 87-8 (1999).
16. Hall, B. G. Simple and accurate estimation of ancestral protein sequences. Proc Natl Acad Sci U S A 103, 5431-6 (2006).