

Hetero: a program to simulate the evolution of DNA on a four-taxon tree

Lars S Jermiin,^{1,2} Simon YW Ho,¹ Faisal Ababneh,³ John Robinson,³ Anthony WD Larkum^{1,2}

¹School of Biological Sciences, University of Sydney, NSW, Australia; ²Sydney University Biological Informatics and Technology Centre, University of Sydney, NSW, Australia; ³School of Mathematics and Statistics, University of Sydney, NSW, Australia

Abstract: We present a computer program to simulate the evolution of a nucleotide sequence on a phylogenetic tree with four tips. The program, Hetero, allows users to assign lineage-specific differences in the rate matrices used to describe the evolutionary process. It has a simple user interface and output, making it equally useful in the teaching and research of phylogenetics.

Keywords: phylogenetics, Monte Carlo simulation, nucleotide sequence evolution, long-branch attraction effect, teaching phylogenetics, bioinformatics

Availability: Hetero is available for academic use at <http://www.bio.usyd.edu.au/~jermiin/programs.htm>

Contact: Lars Jermiin (lsj@bio.usyd.edu.au)

Introduction

Phylogenetic inference is rated among the more difficult challenges facing investigators of biological questions. Many scientists shy away from using a phylogenetic approach – even where it would be appropriate – because it is considered too difficult. However, it need not be that difficult; with practice, it is possible for workers in many scientific disciplines to infer a generalised phylogeny of any specific area because user-friendly phylogenetic programs are available for most computer systems. Instead, the challenge lies in: (1) choosing the appropriate phylogenetic data for the question in mind; (2) choosing the phylogenetic method that most appropriately applies to these data; and (3) determining whether the evolutionary events and processes of the past are indeed correctly inferred.

The performance of molecular phylogenetic methods is usually debated on the basis of well justified criteria (Penny et al 1992). It is known that many of the methods perform well under ideal conditions (ie those that are specified by the assumptions under which the phylogenetic methods operate), and often less so when: (1) the sequences are short (Hillis et al 1994); (2) multiple substitutions at the same positions have eroded the footprints of evolutionary events (Huelsenbeck and Hillis 1993); or (3) the assumptions of the phylogenetic methods are violated by rate heterogeneity among lineages (Felsenstein 1978) or among sites (Yang 1996), by compositional heterogeneity (Lockhart et al 1992), by invariant sites (Fitch 1986b; Lockhart et al 1996), or by covariotides and covarions (Fitch 1986a; Lockhart et al

1998). The message emerging from these and many other studies is the importance of choosing a phylogenetic method that applies most appropriately to the data. However, phylogenetic methods are still often chosen on the basis of prior experience with a phylogenetic approach and rarely after an objective assessment of the appropriateness of various methods in the context of the data to be analysed.

One way to appreciate the importance of choosing the appropriate phylogenetic method is to simulate the evolution of a nucleotide sequence on a known phylogenetic tree. Several programs (Table 1) offer a great deal of flexibility with respect to the substitution models and the tree topology. However, one shortcoming is that all the substitution models used are of the time-reversible form, implying that the sequences generated by these programs will have the same composition of nucleotides, codons or amino acids, which is not always so in real data (eg Jermiin et al 1994; Foster et al 1997).

Hetero is a computer program designed to simulate the evolution of a nucleotide sequence along a rooted four-taxon tree. The nucleotide substitution models are lineage-specific, allowing for different evolutionary processes along adjacent edges, both in terms of the average rate of change and in the equilibrium nucleotide content. Hetero is intended to be a tool for teaching and research; as such, it produces results in

Correspondence: Lars S Jermiin, School of Biological Sciences, Heydon-Laurence Building A08, University of Sydney, NSW 2006, Australia; tel +61 2 9351 3717; fax +61 2 9351 4119; email lsj@bio.usyd.edu.au

Table 1 Publicly available computer programs that can simulate the evolution of a sequence of nucleotides, codons or amino acids on a tree

Program	Data	URL
Seq-Gen	Nucleotides	http://evolve.zoo.ox.ac.uk/
Pseq-Gen	Amino acids	http://evolve.zoo.ox.ac.uk/
Treevolve	Nucleotides	http://evolve.zoo.ox.ac.uk/
PTreevolve	Amino acids	http://evolve.zoo.ox.ac.uk/
Rose	Nucleotides and amino acids	http://bibiserv.techfak.uni-bielefeld.de/download/
ProSeq	Nucleotides	http://helios.bto.ed.ac.uk/evolgen/filatov/proseq.html
evolver	Nucleotides, codons and amino acids	http://abacus.gene.ucl.ac.uk/software/paml.html
Pal	Nucleotides, codons and amino acids	http://www.stat.uni-muenchen.de/~strimmer/pal-project/
Vanilla	Nucleotides, codons and amino acids	http://www.stat.uni-muenchen.de/~strimmer/pal-project/vanilla/

NOTE: All websites accessed 5 Oct 2003.

a format that can be analysed using phylogenetic programs from PHYLIP (Felsenstein 2002).

The program and its features

Fundamental to our approach is the belief that pattern and process are two sides of the same coin: the pattern is the phylogeny that describes the time and order of divergence events, and the process is the mechanism by which mutations in nucleotide sequences accumulate over time in diverging lineages. It makes no sense to consider the pattern and process independently, even when only one of these two components is of interest, because any inference of evolutionary pattern is dependent on the evolutionary process and vice versa.

Hetero uses a rooted, four-leaved binary tree and user-specified edge lengths to allow the evolution of an ancestral nucleotide sequence to occur along the edges of the tree (Figure 1). Information must be entered in six blocks before the Monte Carlo simulation can proceed:

1. *Entry of edge lengths in the tree.* The edge lengths are given in terms of average number of nucleotide substitutions per site (K) or units of time (t). These measures of evolutionary time are proportional if the sequences evolve under the same time-reversible model of nucleotide substitution, because the average rate of change per site per time unit (k) is the same for all the edges. However, when different substitution models are assigned to different edges, the correct measure of evolutionary time is t . Hetero uses default values for the length of edges in the tree ($a=b=c=d=0.95$; $e=f=0.05$); these values can be overwritten if the user wants to consider another set of edge lengths (which may be necessary if the evolutionary time is given in average number of nucleotide substitutions per site).

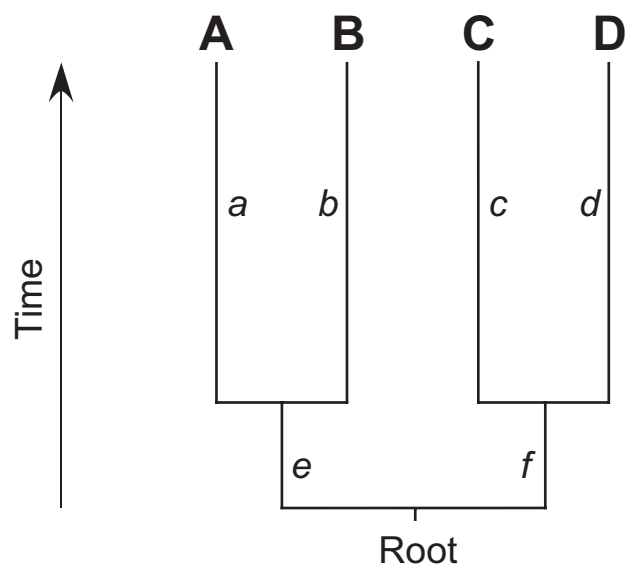


Figure 1 The phylogenetic tree used by Hetero. The root (Root), the edges ($a \dots f$) and the tips (A, B, C, D) are labelled, and the direction of time from the past to the present is indicated by an arrow. Each Monte Carlo simulation begins with a randomly generated ancestral nucleotide sequence at the root of the tree; this sequence is duplicated into two arrays. The sequences in these two arrays are allowed to evolve along the edges e and f . At a later stage during the simulation, each of the sequences in these two arrays is duplicated into another two arrays, and the final simulation occurs along the edges a , b , c and d . The simulations along the individual edges are independent and may proceed by using different substitution models.

2. *Entry of the ancestral nucleotide composition.* It is necessary to know the nucleotide content of the ancestral DNA at the root of the tree (Figure 1) in order to run the Monte Carlo simulation. By default, Hetero uses uniform nucleotide content (ie $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$), but the user may select another composition of nucleotides (eg $\pi_A = 0.3$, $\pi_C = 0.2$, $\pi_G = 0.2$, $\pi_T = 0.3$). If the user selects another composition, the program will determine whether the frequencies add up to 1.0, and an error message is issued if this is not so.

3. *Entry of substitution models for the six edges.* To run the Monte Carlo simulation, it is necessary to specify the rate of change per site per time unit (\mathbf{R}_i) for the i th edge in the tree (Figure 1):

$$\mathbf{R}_i = \begin{bmatrix} -\sum_{y \neq A} \alpha_{iAy} \pi_{iy} & \alpha_{iAC} \pi_{iC} & \alpha_{iAG} \pi_{iG} & \alpha_{iAT} \pi_{iT} \\ \alpha_{iCA} \pi_{iA} & -\sum_{y \neq C} \alpha_{iCy} \pi_{iy} & \alpha_{iCG} \pi_{iG} & \alpha_{iCT} \pi_{iT} \\ \alpha_{iGA} \pi_{iA} & \alpha_{iGC} \pi_{iC} & -\sum_{y \neq G} \alpha_{iGy} \pi_{iy} & \alpha_{iGT} \pi_{iT} \\ \alpha_{iTA} \pi_{iA} & \alpha_{iTC} \pi_{iC} & \alpha_{iTG} \pi_{iG} & -\sum_{y \neq T} \alpha_{iTy} \pi_{iy} \end{bmatrix} \quad (1)$$

where α_{ixy} is the conditional rate of change from nucleotide x to nucleotide y in \mathbf{R}_i , and π_{iy} is the frequency of nucleotide y in \mathbf{R}_i . By default, Hetero uses time-reversible \mathbf{R}_i matrices with predefined parameters (ie $\pi_{iy} = 0.25$; $\alpha_{ixy} = 0.2$), but they can be changed whenever this is needed; if so, Hetero will determine whether any of the diagonal elements are less than -1.0 , and an error message is issued if this is the case.

4. *Entry of the sequence length, number of simulations and seed for simulation.* To run the Monte Carlo simulation, it is necessary to specify the sequence length, the number of simulations and the seed number that is needed to prime the random number generator. Default values for these parameters are not provided; instead, the user is asked to specify the sequence length (it must be in the range of 10 to 10000 nucleotides), the number of simulations (it must be in the range of 1 to 10000 simulations) and a randomly chosen number to prime the ran2 random number generator, which has a periodicity of more than 2×10^{18} (Press et al 1992).
5. *Multiple substitutions at the same site.* There is widespread evidence to suggest that multiple substitutions may have occurred at many sites in coding and non-coding DNA throughout evolution. The result is the disappearance of ancestral states and erosion of the signal that is needed to correctly infer the evolutionary history. Under some conditions, it may be of interest to allow the sequences to evolve without the accumulation of multiple substitutions at the same site (eg to assess the performance of parsimony methods under conditions that differ in terms of multiple substitutions at the same site). Hetero can allow this to occur if the expected tree length is less than 1.0 substitution per site. If the evolutionary time is given as t , Hetero will estimate the length of the i th edge in terms of substitutions per site, K_i :

$$K_i = t_i (\pi_{iA} \sum_{y \neq A} \alpha_{iAy} \pi_{iy} + \pi_{iC} \sum_{y \neq C} \alpha_{iCy} \pi_{iy} + \pi_{iG} \sum_{y \neq G} \alpha_{iGy} \pi_{iy} + \pi_{iT} \sum_{y \neq T} \alpha_{iTy} \pi_{iy}) \quad (2)$$

and the sum of the K_i values will be compared to 1.0. Alternatively, the program will compare the sum of the user-specified K_i values to 1.0.

6. *Entry of names of files that will contain results of simulation.* Hetero produces output that appears on the monitor and in two text files. The user is asked to enter the names of these two files and, if the names are valid and non-identical, Hetero will perform the simulation while updating the user on its progress.

Hetero directs its output to two files and to the standard output. One file contains the alignments of simulated nucleotide sequences; they are stored in the sequential PHYLIP format, allowing them to be analysed directly using the phylogenetic programs in PHYLIP (Felsenstein 2002). The other file contains the details used in the simulation and information collected during the simulation, which may be used to obtain a better understanding of molecular evolution as a dynamic process and to understand why phylogenetic methods do not always perform as well as expected. In particular, the program produces a table with:

1. Pairwise differences in the GC content for all pairs of sequences (columns 1–6);
2. The number of constant sites (eg with A in all four sequences) (column 7);
3. The number of singleton sites (eg with A in one sequence and G in the other three sequences) (columns 8–11);
4. The number of parsimoniously informative sites (eg with A in two sequences and T in the other two sequences) (columns 12–14);
5. The number of hyper-variable sites (ie sites containing three or four different nucleotides) (column 15).

These observations are obtained for each simulation and the average is estimated across all the simulations. In addition, the program also produces a table listing the number of sites that have changed 0, 1, 2 ... n times; this table may be useful in developing an understanding of how multiple substitutions at the same site accumulate.

Binary and executable codes of Hetero are available for research and teaching from <http://www.bio.usyd.edu.au/~jermiin/programs.htm>. Hetero is very fast on standard platforms and, for obvious reasons, produces very large output files – the maximum size of the two output files is

~403 MB. We therefore recommend that users be considerate towards other users of the computer when they run Hetero. If the results are to be stored for a long period, we suggest that users save the first section of the file containing the details and seed that was used to run the simulation; based on this information, it is possible to recreate the data.

Using Hetero to teach phylogenetic inference

Hetero may be used to illustrate specific problems relating to phylogenetic analysis of sequence data. It has long been known (eg Felsenstein 1978; Hasegawa et al 1991; Steel et al 1993) that rate heterogeneity among lineages can lead to what is commonly called a 'long-branch attraction' effect, but how can this be shown effectively to students studying molecular systematics?

One approach is to use Hetero to design **R**-matrices that differ in their average rates of change. For example, let us use the default values for **R_e** and **R_f**, change the values of α_{bxy} and α_{cxy} from 0.20 to 0.38 and change those of α_{axy} and α_{dxy} from 0.20 to 0.02; this ensures that the average rates of change for **R_b** and **R_c** are 19 times faster than those for **R_a** and **R_d**, respectively. Using these matrices, a sequence length of 2000 nucleotides, and allowing multiple substitutions to occur at the same site, we generated 1000 datasets that were analysed using maximum parsimony, as implemented in DNAPARS (Felsenstein 2002).

The correct tree (AB|CD) was found in 3.3% of all cases, whereas incorrect trees were found in 96.7% (BC|AD) and 0.0% (AC|BD) of all cases. When repeated with $\alpha_{axy} = \alpha_{bxy} = \alpha_{cxy} = \alpha_{dxy} = 0.2$, the analysis gave a very different result: the correct tree was recovered in 98.9% of all cases. Increasing the sequence length from 2000 to 10 000 nucleotides worsened the recovery rate to 0.0% in the first analysis and improved it to 100% in the second analysis. As expected, using maximum likelihood with a transition/transversion ratio of 0.5 and a constant rate of change across all sites, as implemented by DNAML (Felsenstein 2002), the recovery rate of the first analysis increased to nearly 100%.

To appreciate why these results are so different, we need to examine the distributions of sites that have changed *X* times, and the number of parsimoniously informative sites that support the three splits (AB|CD), (AC|BD) and (AD|BC). It is clear (Table 2) that the number of sites that have changed more than once is similar (11.7%), so the difference between the two analyses cannot be explained by a change in the proportion of sites that have superimposed mutations. The average number of sites supporting different splits offers the

Table 2 The average number of sites that have changed *X* times^a

<i>X</i>	Uniform rates ^b	Non-uniform rates ^c
0	1114.94	1115.18
1	651.56	650.86
2	190.29	190.51
3	37.17	37.22
4	5.37	5.49
5	0.62	0.66
6	0.06	0.07
7	0.01	0.02
8	0.00	0.00

^a These values are based on a simulation with default values chosen for the ancestral nucleotide sequence; the sequence length equal to 2000 nucleotides; with 1000 simulations; and with multiple nucleotide substitutions allowed to occur at the same site. Uniform nucleotide content was used in the analyses.

^b Default values were used in all **R**-matrices.

^c The default values were used in **R_e** and **R_f**, whereas $\alpha_{bxy} = \alpha_{cxy} = 0.38$ and $\alpha_{axy} = \alpha_{dxy} = 0.02$ were used in **R_b** and **R_c**, and in **R_a** and **R_d**, respectively.

explanation needed (Table 3): when the rate of molecular evolution was homogeneous among lineages, the strongest support was for the correct split (ie AB|CD, being 33.78 versus 16.98 and 16.72 sites); with rate heterogeneity among lineages, the strongest support is for another split (ie AD|BC, being 33.06 versus 20.33 and 3.36 sites).

Based on these observations, we can infer that because the molecular evolution along *b* and *c* occurred at a faster rate than along *a* and *d* (Figure 1), B and C are often more similar to each other than they are to A and D, respectively. Convergent, parallel and back substitutions must be the primary reasons for this higher than expected similarity between B and C, and the maximum parsimony method is clearly unable to identify the correct tree using sequence data that have arisen through a more complex process of

Table 3 The average number of sites supporting different splits^a

Splits	Uniform rates ^b	Non-uniform rates ^c
(ABCD)	1130.63	1144.38
(A BCD)	170.70	18.27
(B ACD)	169.30	335.23
(C ABD)	169.61	336.47
(D ABC)	169.98	18.14
(AB CD)	33.78	20.33
(AC BD)	16.98	3.36
(AD BC)	16.72	33.06
Hyper-variable	122.30	90.77

^a These values are based on a simulation with default values chosen for the ancestral nucleotide sequence; the sequence length equal to 2000 nucleotides; with 1000 simulations; and with multiple nucleotide substitutions allowed to occur at the same site. Uniform nucleotide content was used in the analyses.

^b Default values were used in all **R**-matrices.

^c The default values were used in **R_e** and **R_f**, whereas $\alpha_{bxy} = \alpha_{cxy} = 0.38$ and $\alpha_{axy} = \alpha_{dxy} = 0.02$ were used in **R_b** and **R_c**, and in **R_a** and **R_d**, respectively.

molecular evolution. Interestingly, but perhaps not surprisingly (for example, see Hillis et al 1994), the maximum likelihood method performs better under the very same conditions.

Conclusion

We have developed a small but efficient computer program to simulate the evolution of a nucleotide sequence on a phylogenetic tree with four tips. The program allows users to assign lineage-specific differences in many of the relevant evolutionary parameters. It has a simple user interface and a simple output, making it equally useful in the teaching and research of phylogenetics.

Acknowledgements

We wish to thank Ross H Crozier, Andrew Grimm and an anonymous reviewer for their constructive comments on the manuscript. Simon YW Ho was supported by an AE and FAQ Stephens Scholarship, and Faisal Ababneh was supported by a postgraduate scholarship from Al-Hussein Bin Talal University, Jordan. This is research paper #001 from SUBIT.

References

- Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*, 27:401–10.
- Felsenstein J. 2002. PHYLIP (Phylogeny Inference Package). 3.6(alpha3). Seattle: Department of Genome Sciences, University of Washington.
- Fitch WM. 1986a. The estimate of total nucleotide substitutions from pairwise differences is biased. *Philos Trans R Soc Lond B Biol Sci*, 312:317–24.
- Fitch WM. 1986b. An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. *Prog Clin Biol Res*, 218: 149–59.
- Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol*, 44:282–8.
- Hasegawa M, Kishino H, Saitou N. 1991. On the maximum likelihood method in molecular phylogenetics. *J Mol Evol*, 32:443–5.
- Hillis DM, Huelsenbeck JP, Swofford DL. 1994. Hobgoblin of phylogenetics. *Nature*, 369:363–4.
- Huelsenbeck JP, Hillis DM. 1993. Success of phylogenetic methods in the four-taxon case. *Syst Biol*, 42:247–64.
- Jermiin LS, Graur D, Lowe RM, Crozier RH. 1994. Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. *J Mol Evol*, 39:160–73.
- Lockhart PJ, Larkum AWD, Steel MA, Waddell PJ, Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc Natl Acad Sci USA*, 93: 1930–4.
- Lockhart PJ, Penny D, Hendy MD, Howe CJ, Beanland TJ, Larkum AWD. 1992. Controversy on chloroplast origins. *FEBS Lett*, 301:127–31.
- Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol Biol Evol*, 15:1183–8.
- Penny D, Hendy MD, Steel MA. 1992. Progress with methods for constructing evolutionary trees. *Trends Ecol Evol*, 7:73–9.
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. Numerical recipes in C. New York: Cambridge Univ Pr.
- Steel MA, Hendy MD, Penny D. 1993. Parsimony can be consistent! *Syst Biol*, 42:581–7.
- Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol Evol*, 11:367–72.