A Genetic Algorithm for Maximum-Likelihood Phylogeny Inference Using Nucleotide Sequence Data

Paul O. Lewis
Department of Biology, University of New Mexico

Phylogeny reconstruction is a difficult computational problem, because the number of possible solutions increases with the number of included taxa. For example, for only 14 taxa, there are more than seven trillion possible unrooted phylogenetic trees. For this reason, phylogenetic inference methods commonly use clustering algorithms (e.g., the neighbor-joining method) or heuristic search strategies to minimize the amount of time spent evaluating nonoptimal trees. Even heuristic searches can be painfully slow, especially when computationally intensive optimality criteria such as maximum likelihood are used. I describe here a different approach to heuristic searching (using a genetic algorithm) that can tremendously reduce the time required for maximum-likelihood phylogenetic inference, especially for data sets involving large numbers of taxa. Genetic algorithms are simulations of natural selection in which individuals are encoded solutions to the problem of interest. Here, labeled phylogenetic trees are the individuals, and differential reproduction is effected by allowing the number of offsprings produced by each individual to be proportional to that individual's rank likelihood score. Natural selection increases the average likelihood in the evolving population of phylogenetic trees, and the genetic algorithm is allowed to proceed until the likelihood of the best individual ceases to improve over time. An example is presented involving \textit{rbcL} sequence data for 55 taxa of green plants. The genetic algorithm described here required only 6% of the computational effort required by a conventional heuristic search using tree bisection/reconnection (TBR) branch swapping to obtain the same maximum-likelihood topology.

Introduction

Conventional wisdom holds that accurate solutions to large-scale phylogenetic problems require information from unattainable amounts of nucleotide sequence data. This conclusion is in large part due to simulation studies of four-taxa trees (Hillis, Huelsenbeck, and Swofford 1994). This view has recently been challenged by work suggesting that highly accurate phylogenetic inference can be accomplished with reasonable amounts of sequence data as long as taxon sampling is extensive (Hillis 1996). Unfortunately, increased taxon sampling does not come without a cost in terms of the computation time required by the analysis. Concomitantly, models being developed are increasingly more realistic from a biological perspective. These models recognize such important features of nucleotide sequences as nonindependence among the sites within codons (Goldman and Yang 1994; Muse and Gaut 1994), evolutionary correlation between adjacent nucleotides (Felsenstein and Churchill 1995), insertion-deletion processes (Thorne, Kishino, and Felsenstein 1991), and nonindependence of associated stem sites in ribosomal RNA (von Haeseler and Schoniger 1995; Muse 1995; Tillier and Collins 1995). These more complex models for phylogenetic inference require the use of the computationally expensive maximum-likelihood criterion. Thus, this ideal combination of more realistic evolutionary models and data sets with many taxa can substantially increase the computational cost of phylogenetic inference.

The phylogeny reconstruction problem is a classic NP-complete problem (Day 1987), with the number of possible solutions (unrooted, bifurcating tree topologies) being $$(2n-5)!/[(n-3)!(2^{n-3})]$$ (Edwards and Cavalli-Sforza 1964) for \(n\) included terminal taxa. There are no known efficient solutions to this class of computational problems, which means that, even when the optimality criterion used is relatively fast (e.g., maximum parsimony), it will never be possible to examine all trees when hundreds of taxa are included in the analysis. Fortunately, heuristic search strategies (Swofford and Begle 1993) provide a very good alternative to exhaustive enumeration, but even these can take several weeks or months to complete, often with the result that the search is stopped prematurely because of time constraints (Chase et al. 1993; Rice, Donoghue, and Olmstead 1997; Soltis et al. 1997). Model-based optimality criteria such as maximum likelihood require orders of magnitude more computational effort than does parsimony for each tree examined. I report here a different approach (a genetic algorithm) that substantially increases the efficiency of heuristic phylogeny searches involving large numbers of taxa while retaining the analytical power of model-based optimality criteria.

Genetic algorithms (GAs) make use of the power of natural selection to solve real-world problems (Forest 1993; Mitchell 1996). GAs have been applied to a diversity of complex optimization problems in engineering for many years, although, ironically, their use in problems involving biological data is only just being explored (for examples, see May and Johnson 1995; Parsons, Forrest, and Burks 1995; Vanbatenburg, Goldyaev, and Pleij 1995). The literature of GAs is littered with biological metaphors. GAs begin with a “population” of encoded random solutions to the problem of interest. Such encoded solutions are usually termed “chromosomes,” and a measure of their “fitness.” These “individuals” are subjected to simu-
lared natural selection, with those specifying fitter solutions leaving, on average, more "offspring" to the next generation than individuals of lower fitness. Each "generation," individuals are subjected to "mutation" and "recombination" events, with mutation and recombination operators defined according to the nature of the problem and the method used to encode solutions to the problem. Over time, the population increases in average fitness due to the action of natural selection on variation introduced by mutation and recombination.

The ability of GAs to find near-optimal solutions quickly in the face of complex data makes them ideal candidates for the problem of phylogenetic inference, especially when many taxa are included or complicated evolutionary models (necessitating the use of computer-intensive inference methods such as maximum likelihood) are applied. In the case of phylogeny reconstruction, the single chromosome of each individual can be designed to encode a single phylogenetic tree, along with its branch lengths and the values of other parameters comprising the substitution model used. Mutation and recombination operators can be defined for phylogenetic trees, and the fitness of an individual may be equated to its natural log likelihood (lnL) score. Trees with higher values of lnL thus tend to leave more offspring to the next generation, and natural selection increases the average lnL of the individuals in the simulated population. The tree with the highest lnL after the population fitness ceases to improve is taken to be the best estimate of the maximum-likelihood tree.

While not the first application of a GA to the phylogeny inference problem (see Matsuda 1996), the phylogenetic GA described here implements significant innovations designed to improve the ability of the algorithm to infer phylogeny for large numbers of taxa and parameter-rich substitution models. Genetic algorithms (and especially parallel implementations of GAs) offer the potential for inferring maximum-likelihood trees for large data sets involving hundreds of sequences.

Materials and Methods

The program written for this study (GAML: genetic algorithm for maximum likelihood phylogeny inference) implements the HKY (Hasegawa, Kishino, and Yano 1985) nucleotide substitution model and uses the maximum-likelihood criterion for ranking trees. It begins with a single population of $n$ individuals, each of which consists of a tree description (Newick standard format described in Swofford and Begle 1993, pp. 143–146; Felsenstein 1995, main.doc) and a value for the transversion/transversion rate ratio, $\kappa$, a parameter of the HKY model. Other parameters of the HKY model include branch lengths and equilibrium nucleotide frequencies. The branch lengths (measured as expected number of nucleotide substitutions per site) are specified as part of the tree description (fig. 1). While the equilibrium nucleotide frequencies could be incorporated into the individuals as well, they were simply equated to the empirical nucleotide frequencies for the purposes of this study to make the results directly comparable to the program PAUP* (Swofford 1998). The fully specified tree description (tree topology plus branch lengths), the value of the $\kappa$ parameter, and the base frequencies together allow the likelihood of the tree to be computed.

At the start of a GA run, each individual in the population is initialized with a random tree topology in which each branch length is set to an arbitrary value $x$ (for example, 0.05). Before the run begins, each branch length is changed slightly using the same method de-
scribed below for introducing branch length mutations. The $\kappa$ parameter for each individual is initialized to 4.0.

The following sequence of events occurs during each generation. Firstly, the fitness of each individual in the population is computed. This involves computing the lnL score for the tree specified as part of each individual’s “genotype.” This computation assumes the branch lengths specified in the tree topology; no optimization of branch lengths is performed in obtaining the lnL of the tree. In this aspect, the algorithm presented here differs from all existing maximum-likelihood phylogeny search algorithms, which optimize branch lengths for every tree considered during a search. It is at this step that considerable time is saved over other methods, including the GA described by Matsuda (1996), because it is the optimization of branch lengths that is the most computationally intensive portion of the traditional approach to maximum-likelihood phylogeny reconstruction.

Second, the individuals in the population are ranked on the basis of their fitness (i.e., lnL score). Because the branch lengths have not been optimized, two individuals specifying exactly the same tree topology can have different lnL values, and hence a potentially different ranking, at this stage. The probability of leaving an offspring to the next generation is defined to be $p(n-i+1)$, where $i$ is the position of the individual in the ranked list ($i = 1$ being the position of the individual having the highest value of lnL and $i = n$ being the lowest ranking individual) and $p$ is chosen such that the sum of such probabilities over all potential parents is 1.0 (i.e., $p = 2/[n(n+1)]$). This is, in GA parlance, a form of rank selection (Mitchell 1996, pp. 169–170) and is useful in both preserving variation early in the GA run and keeping the selection pressure high late in a run, when the differences in lnL between trees becomes small.

Third, the individual having the highest lnL is automatically allowed to leave $k$ offspring in the next generation. The remaining $n-k$ individuals are created by choosing a parental individual at random (based on the probabilities computed in the second step) and copying this parent’s genotype to the next generation. At this point, there are two populations of individuals, one representing the parental generation and the other representing the offspring generation. Unless specific reference is made to the parental population, all further discussion applies to the offspring population. All individuals except the first are subjected to branch length mutations and may also undergo topological mutation and/or recombination (described below) at this point. The first individual is protected from mutation and recombination to ensure that the genotype of the best individual found thus far will always be present. The following discussion of mutation and recombination thus refers only to the “mutable” individuals and not to the one being protected.

While all mutable individuals are subjected to branch length mutation, not all branch lengths are changed for any given individual. A random proportion $\lambda$ of the branches of any given individual are changed. For those branches selected to be mutated, a multiplicative factor is drawn from a gamma distribution having shape parameter $\alpha$ and mean 1.0. The new branch length is simply the old branch length times the gamma-distributed factor. The reason that a gamma distribution was chosen is that gamma random deviates are guaranteed to be in the range 0 to $\infty$, which means that the mutated branch lengths are guaranteed to remain in their valid range (from 0 to $\infty$). The variance of the gamma-distributed branch length mutation factors is inversely proportional to the shape of the gamma distribution. Thus, high values of the shape parameter (i.e., $\alpha = 500$) were used so that branch length mutations were of low effect. The probability that a branch is lengthened (versus shortened) is also related to the magnitude of $\alpha$. In all cases, the probability that a branch is shortened by mutation is greater than the probability that it is lengthened when using gamma-distributed factors to modify branch lengths. At the high value of $\alpha$ used in this study, however, the difference between these two probabilities is quite small (<0.02).

Topological mutations are incurred with probability $\mu$. Thus, on average, the topology of $(n-1)\mu$ individuals will be changed each generation. A topological mutation involves removing a randomly chosen subtree and reattaching it at a randomly chosen site on the remaining tree. Thus, topological mutations correspond exactly to the SPR (subtree pruning/regrafting) branch-swapping strategy (Swofford and Begle 1993, pp. 36–37), the only difference being that in SPR branch swapping, these mutations are applied in a very systematic fashion. That is, all possible subtrees are removed, and each is attached in turn to all possible remaining sites.

The $\kappa$ parameter for each individual is changed with probability $\pi$. This means that a proportion $\pi$ of individuals receives a new $\kappa$ value equal to the previous value of $\kappa$ multiplied by a gamma-distributed factor. Thus, the manner in which the $\kappa$ parameter is mutated is identical to mutation of branch lengths. It is because in general there is an excess of transitions compared to transversions. The $\kappa$ parameter as defined here is the ratio of the instantaneous rate of transitions divided by the instantaneous rate of transversions. A value of $\kappa$ less than 1.0 therefore implies that the instantaneous rate of transversions exceeds the instantaneous rate of transitions. The parameter $\kappa$ (the transition/transversion ratio) is often confused with the transition/transversion ratio, which is the probability (over one unit of time) of any transition divided by the probability of any transversion. Since there are twice as many ways of getting a transversion (A$\leftrightarrow$C, A$\leftrightarrow$T, G$\leftrightarrow$C, G$\leftrightarrow$T) as there are of getting a transition (A$\leftrightarrow$G, C$\leftrightarrow$T), the transition/transversion ratio equals 0.5 when $\kappa = 1.0$, assuming all four bases have equal frequencies.

Finally, recombination is performed with probability $r$. After a first parent has been chosen and given rise to an offspring individual in the next generation, a decision is made whether or not to allow a second parent to recombine with this new offspring individual. At this point, the offspring individual may have already expe-
The number of possible unrooted trees is too large to permit exhaustive enumeration, and heuristic search strategies must be employed. In a commonly used approach, a starting tree is obtained by random stepwise addition, and branch swapping is initiated on this stepwise addition tree. This tree search strategy is used in both DNAML (Felsenstein 1995) and fastDNAML (Olsen et al. 1994), as well as PAUP (Swofford and Begle 1993). The random-addition tree is created by starting with a tree composed of three randomly chosen taxa. A fourth randomly chosen taxon is attached to all three branches, in turn, and the placement resulting in the four-taxa tree having the highest likelihood score is used as the starting tree for adding the fifth randomly chosen taxon. This process continues until all $n$ taxa are added to the tree.

Branch swapping strategies include nearest neighbor interchange (NNI), subtree pruning/regrafting (SPR), and tree bisection/reconnection (TBR). These three strategies are explained fully by Swofford and Begle (1993). All involve making systematic changes to a tree in an attempt to find a topologically distinct tree having a higher likelihood score. For example, in SPR swapping, every possible subtree is removed (pruned) from the starting tree and reattached, in turn, to every remaining branch. Once a rearrangement with a higher likelihood score is found, the branch-swapping process begins again using that rearrangement as the starting tree.

Because of the systematic nature of branch swapping and the computational burden imposed by use of the maximum-likelihood optimality criterion, an analysis using a combination of these two approaches can be quite costly in terms of time (Kuhner and Felsenstein 1994). As a result, researchers attempt to speed up the process by (1) using a criterion other than maximum likelihood, (2) using a clustering approach rather than performing an heuristic search, or (3) including fewer taxa in the analysis. All of these alternatives have drawbacks.

The maximum-likelihood criterion is advantageous for a number of reasons, including the ability to model a variety of factors thought to affect nucleotide sequence evolution, robustness to violations of its model assumptions (Huelsenbeck 1995), and resistance to long-branch attraction (Gaut and Lewis 1995). Maximum likelihood also makes use of more information in the data than do other optimality criteria. For instance, some information is lost in converting a discrete data matrix (composed of observations of particular characters or nucleotide sites on each taxon) into a matrix of pairwise estimates of evolutionary distance (Penny, Hendy, and Steel 1992). Thus, criteria based on pairwise distances must necessarily work with less information than discrete character methods, such as parsimony or likelihood, that operate directly on the original observations. Parsimony, however, discards potentially relevant information as well. While constant and autapomorphic characters do not affect parsimony tree lengths, such characters are quite useful in identifying long branches and thus avoiding long-branch attraction. Maximum likelihood, by operating on the original observations and es-

Fig. 2.—Recombination involves selecting a second parent and using it to modify an offspring individual that has been copied, perhaps with modification, from the first parent. Specifically, the topology of the offspring individual is cut at a random point, producing a subtree (dotted lines in leftmost tree) and a remainder tree (solid lines in leftmost tree). Tip nodes represented in the subtree are then pruned from the second parent (dotted lines in middle tree). Finally, the subtree (dotted lines in rightmost tree) is added to what is left of the second parent after pruning (solid lines in rightmost tree).
Genetic Algorithm for Phylogeny Reconstruction 281

Fig. 3.—Progress of GA in the last of the three runs described in the text. The horizontal axis is time as measured in number of generations. The vertical axis is the natural logarithm of the likelihood (lnL). The solid curved line is the score of the best individual in the population for each generation. The arrow indicates the point (at generation 7278) at which the best tree was found; the GA was allowed to run for an additional 1,822 generations to see if a better tree could be found.

Estimating branch lengths from all characters, is the only criterion in current use that makes full use of the information present in the data.

Another common way of avoiding a time-consuming heuristic search is through the use of clustering algorithms that use a stepwise procedure to produce a single tree, which is considered to be the final estimate of phylogeny. Clustering approaches such as neighbor-joining are quite fast, but often do not result in trees that are optimal. Strimmer and von Haeseler (1996) found that the ability of neighbor-joining to reconstruct the true tree in computer simulations decreased exponentially with the number of taxa included in the study. Even in the four-taxon case, Huelsenbeck (1995b) found that a search using maximum likelihood was generally superior to neighbor-joining when the two were compared on an equal basis.

Finally, including fewer taxa in an analysis may actually make the phylogeny problem more difficult. There is some evidence from simulation studies (Hillis 1996) that including more taxa in an analysis improves the accuracy of the inferred tree. Genetic algorithms such as the one described here offer the possibility of avoiding all of the above-mentioned drawbacks by allowing heuristic searches for data sets with many taxa and using the maximum-likelihood criterion with sophisticated evolutionary models.

An Example Using rbcL Sequences from Green Plants

The above-described GA was applied to a 55-taxon problem involving sequences of the chloroplast gene rbcL from a diversity of green plants. The complete alignment used for this study in the form of a data file in nexus format is available over the Internet at the URL http://biology.unm.edu/~lewisp/gaml.html. The GA settings were as follows: number of individuals (n) was 25, automatic number of offspring for best individual each generation (k) was 5, branch length mutation probability (λ) was 0.05, topological mutation probability (μ) was 0.2, κ mutation probability (π) was 0.1, recombination probability (r) was 0.2, and the gamma shape parameter (α) used to modify branch lengths and the κ parameter was 500. Three separate GA runs were performed using three distinct random number seeds. All three runs were terminated when no improvement in the lnL of the best tree was observed in 2,000 generations. All three runs were performed on a Silicon Graphics Origin200 (180 MHz R10000 processor) running IRIX 6.4. The first run required 7,970 generations and 16.3 h of CPU time to obtain the final tree, which had a lnL score of 22,464,649.87077. When the branch lengths of this tree were optimized using PAUP* 4.0 (d57), the lnL improved only slightly, to 22,464,659.1 indicating that the GA performs reasonably well at fine-tuning branch lengths. The second GA run required 5,568 generations and 11.3 h of CPU time to arrive at a tree having lnL = 22,459,697.58 (which optimized to 22,459,707.58 using PAUP*). Finally, the third run required 7,278 generations and 14.8 CPU h and found a tree better than either of the first two runs with lnL = 22,458,328,507 (figs. 3 and 4). Optimization of branch lengths with PAUP* resulted in only a slightly increased lnL (22,458,308,27). In total, the three GA runs required 42.4 h of CPU time.
Figure 4.—Tree inferred by the GA on the last of three runs. Branch lengths are those inferred by the GA. The scale bar for branch lengths represents 0.1 expected substitutions per nucleotide site. The tree inferred by PAUP* is identical in topology to this tree, although the branch lengths were slightly different because PAUP* fully optimizes branch lengths for each tree examined.

It should be emphasized that the GA converged on a different tree topology in each of the three runs. Given that the best tree found by the GA upon termination differed between runs, can we have any confidence in the tree returned by the third run, which is the best of the three in terms of the maximum-likelihood criterion? Prior to the GA runs, a heuristic search was performed using PAUP* 4.0 (d54) on the same data set and using the same computer and model of nucleotide substitution. Like the GA, PAUP* was instructed to optimize not only branch lengths but also the $\kappa$ parameter of the HKY model. A single random-addition starting tree was swapped to completion using the TBR branch-swapping strategy. This process resulted in exactly the same tree topology (fig. 3) produced by GA run number three, but required 783.2 h to initially discover the tree that was ultimately chosen as best. A total of 859.9 h of CPU time was required by PAUP* to determine that no further swapping could improve the lnL of this tree. Thus, conservatively, PAUP* required more than 18 times as much computing effort as the GA (783.2 h for PAUP* vs. 42.4 h for GAML).

Conclusions

Genetic algorithms hold much promise for the future of molecular systematics, which is now experienc-
about phylogeny reconstruction. Thanks to Louise Lewis for much encouragement, and for critically reading many preliminary versions of this paper. Special thanks go to Derek Smith for much help in crossing the initial hurdles, and also to Stephanie Forrest for allowing me to participate in her course in genetic algorithms at the University of New Mexico. I also thank James Brown, Gretchen Hofmann, Tim Lowrey, Rob Miller, Jeff Thorne, and Terry Yates for helpful comments and suggestions for improvement.

APPENDIX

The computer program GAML used in this study is available for the Silicon Graphics (IRIX 5.3 and higher), Windows (Windows 95 or Windows NT), and Power Macintosh platforms. Nucleotide sequence data are read by GAML in PHYLIP (Felsenstein 1995) format, and GAML accommodates the standard ambiguity codes (although all ambiguities are treated as missing data). To obtain the program, please see the URL http://biology.unm.edu/~lewisp/gaml.html.

LITERATURE CITED


MARCY K. UYENOYAMA, reviewing editor

Accepted November 20, 1997