# A GA for Maximum Likelihood Phylogenetic Inference using Neighbour-Joining as a Genotype to Phenotype Mapping

Leon Poladian
School of Mathematics and Statistics
University of Sydney
NSW 2006 Australia
L.Poladian@maths.usyd.edu.au

## ABSTRACT

Evolutionary relationships among species can be represented by a phylogenetic tree and inferred by optimising some measure of fitness, such as the statistical likelihood of the tree (given a model of the evolutionary process and a data set). The combinatorial complexity of inferring the topology of the best tree makes phylogenetic inference ideal for genetic algorithms. In this paper, two existing algorithms for phylogenetic inference (neighbour-joining and maximum likelihood) are co-utilised within a GA and enable the phenotype and genotype to be assigned quite different representations. The exploration vs. exploitation aspects of the algorithm are examined in some test cases. The GA is compared to the well known phylogenetic inference program PHYLIP.

## Categories and Subject Descriptors

J.3 [**Life and medical sciences**]: Biology and genetics; F.2.2 [**Nonnumerical algorithms and problems**]: Sorting and searching; I.5.3 [**Pattern recognition**]: Clustering

## General Terms

algorithms

## Keywords

phylogenetic inference, genetic algorithms, genotype to phenotype mapping, neighbour joining, maximum likelihood

## 1. INTRODUCTION

Phylogenetic inference is the construction of trees that represent the genealogical relationships between different species. It begins with a data set consisting of characters for each species; these characters might be nucleotide or amino acid sequences, protein shapes, anatomical characters, biosynthetic pathways or behavioural traits. In this paper we look at sequences of the nucleotides, $A$, $C$, $G$ and $T$. A typical set of sequences is shown in Table 1.

**Table 1: An example of aligned nucleotide sequences for 7 species. The degree of similarity of the sequences implies a particular genealogical relationship between the species.**

| | |
|---|---|
| species1 | CGTCGGATTGAGGTTT...GTACGACCATAATCTTAGA |
| species2 | AGTATGATGAAGCGAT...AATGCAAGAGCTACCATGA |
| species3 | AATCGGATAACAATGC...GATTGTTCTTCTTGATCGA |
| species4 | GCAGTAGTCCGTTAAG...TTCAGTGCAGTAGTTTCGG |
| species5 | ATAAGGCCCTGTCCTA...CTCTCCGCCGGACTATAGC |
| species6 | TCCGTGTGATTTTACA...GGGTAAGTATCAGTGGAAC |
| species7 | TATGCATTAGATTGGG...GGCGCACCACTTCAGCTCC |

Examples of trees can be seen in Fig. 1. The leaves of the tree represent the species in the data set; the internal nodes represent the inferred common ancestors; and the lengths of the edges or branches represent evolutionary time or an associated measure of distance between species. The trees shown are unrooted which is common when the model of the evolutionary process is time reversible and only a clustering of species into more or less closely related groups is required. The number of topologically distinct unrooted trees with $n$ leaves (species) is $(2n-5)!! = (2n-5)(2n-3)\dots 3.1$. Exhaustive search of all possible tree topologies becomes infeasible for even moderate numbers of species. The internal nodes on trees are usually unlabelled, but in studies of labelled histories [8] rooted trees are counted as distinct if the internal nodes (common ancestors) arise in a different order. The number of labelled histories is $2^{1-n}n!(n-1)!$. Labelling of internal nodes is also possible for unrooted trees, and can correspond to the order in which the internal nodes are inferred or emerge from an algorithm. The number of such distinct labelled trees is the same as for labelled histories.

For example, for 7 species there are 945 unlabelled unrooted trees and 56,700 labelled histories; for 10 species there are $2 \times 10^6$ unlabelled and $2.5 \times 10^9$ labelled trees; for 20 species the numbers grow to $2.2 \times 10^{20}$ and $5.6 \times 10^{29}$.

There are also $N = 2n-3$ internal branch lengths (or $2n-2$ for rooted trees) that need to be determined, so if


*GECCO'05*, June 25–29, 2005, Washington, DC, USA.

we label the relevant combinatorial/topological search space as $\mathbb{S}$ then the full search space is $\mathbb{S} \otimes (\mathbb{R}^+)^N$. This is the *phenotype space* in which the GA in this paper searches.

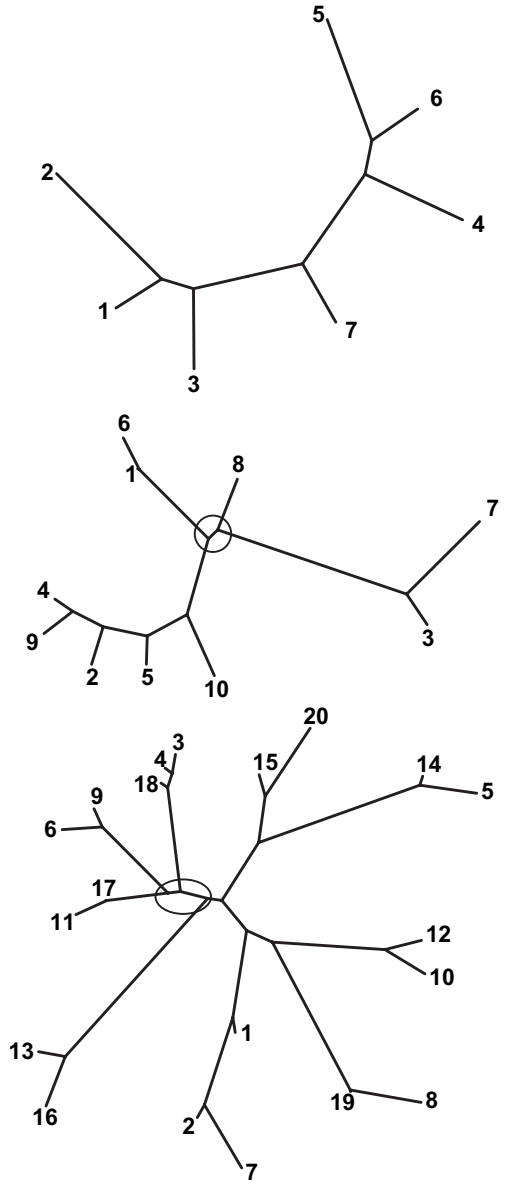



**Figure 1: The three tree topologies, with 7, 10 and 20 species respectively, used to test the genetic algorithm. The encircled regions are parts of the trees most difficult to reconstruct.**

For each pair of species, the total length of the edges which connect them can be calculated. This is a symmetric matrix containing a positive parameter for each pair of distinct species. The diagonal entries are zero. These $M = n(n-1)/2$ positive numbers will be used to represent the tree. The space of all matrices of this form (called distance matrices) is $(\mathbb{R}^+)^M$. This will become the *genotype space* in which the GA operates. An algorithm for constructing a tree from a distance matrix is discussed later.

The main idea of this paper is to have different representations for the genotype and phenotype: one is a matrix, the other is a tree. Separating the genotype from the phenotype necessitates the use of a good mapping from one space to the other, but also allows the use of a different set of genetic operators.

In Section 2 existing algorithms are briefly reviewed and compared to the new ideas proposed here. Details of the genotype to phenotype mapping and the fitness function are given in Section 3. The operation and parameters of the GA algorithm are given in Section 4. The performance of the algorithm is analysed in Section 5.

## 2. EXISTING ALGORITHMS FOR PHYLOGENETIC INFERENCE

Two recent books [29, 12] on phylogenetic methods give theoretical details and case studies for many current phylogenetic algorithms. Two specific approaches are of relevance here: maximum likelihood and distance based methods.

Maximum likelihood methods use a statistical model of the evolutionary process to compare candidate trees (hypotheses) and determine which is the most likely. Each competing hypothesis consists of three parts: the topology of the tree, the lengths of the edges, and the evolutionary model parameters (i.e. how nucleotide changes occur). Given a specific tree, a very efficient algorithm [12] exists for calculating its likelihood.

The definition of fitness (e.g. likelihood) and its calculation are independent of the method used to search for good candidate tree topologies or to optimise the edge lengths for a given tree shape. In many implementations, new candidate topologies are obtained by transformations applied directly to the tree. In order of increasing complexity (and disruption of the tree shape) these are: nearest-neighbour interchange (NNI) where one of the two subtrees on one end of an internal edge is swapped with one of the two subtrees at the other end; subtree pruning and regrafting (SPR) where a subtree is detached from one part of the tree and reattached elsewhere; tree bisection and reconnection (TBR) where an internal edge is deleted, the roots of the two subtrees are ignored, the trees are rotated and reattached arbitrarily. The space of all trees can be spanned more or less efficiently using any or all of these transformations. One of the best known phylogenetic inference packages, PHYLIP [11], has a maximum likelihood algorithm that constructs trees by gradually adding in each additional species and has options to perform various rearrangements of the trees as above.

The combinatorial part of the search is always regarded as the difficult aspect of the search, with the optimisation of the branch lengths considered unproblematic. Often hill climbing is used to optimise edge lengths for each given topology, but it has been shown [2] that even with as few as 4 species, there can be multiple maxima in the space $(\mathbb{R}^+)^N$ and hill climbing can fail to find the global optimum even if it is unique.

Distance-based methods are amongst the most rapid and simplest algorithms for phylogenetic inference and have two separate stages. The first stage determines an evolutionary distance between all pairs of species. This distance can be as simple as the fraction of nucleotide sites that differ between two gene sequences, or more elaborate by incorporating assumptions about different probabilities of different types of mutations. In the second independent stage, the matrix of pairwise distances is then used to construct a tree topol-

ogy. That is, a tree is sought such that the total length of the edges joining pairs of species is as similar as possible to the distance between species determined by the evolutionary model. One of the most efficient methods for this is a type of cluster analysis: the Neighbour-Joining (NJ) algorithm developed by Saitou and Nei, [28] and subsequently refined by Studier and Keppler [33]. The tree produced by this algorithm can only be regarded as a heuristic solution since it is not clear what it optimises; however, this tree is often very similar to the optimal trees found by other algorithms. Indeed the NJ tree can be used as a starting point for more sophisticated searches. The NJ algorithm has some interesting properties that suggest it would make a good genotype to phenotype mapping as discussed in the next section.

## 2.1 GA-based Searches

Thus, many aspects of the search problem make phylogenetic inference ideal for population based methods such as genetic algorithms. The first application of GA to phylogenetic inference appears to be by Matsuda in 1996 [21]. Since then several variations have been proposed by Lewis [20]; Moilanen [22]; Katoh, Kuma and Miyata [17]; Brauer *et al.* [1]; Lemmon and Milinkovich [19]; Congdon [3, 4, 5]; Poladian and Jermiin [24, 23]; and Shen and Heckendorn [31].These approaches all use some combination of mutation and crossover (recombination) operators based on NNI, SPR or TBR. They differ in whether the genetic operators are combined with exhaustive local searches, and other attributes of selection and diversity maintenance (eg. niching and crowding). What all these methods have in common is that the genetic operators are applied directly to the tree.

The advantages of directly manipulating trees is that genetic operators can be chosen that clearly exhibit properties such as locality (small mutations lead to similar looking trees) and heritability (offspring have features that resemble their parents). One disadvantage of working directly with the phenotype is that most common recombination operators need to be accompanied by a complicated repair mechanism to ensure that the offspring have meaningful phenotypes: mapping cleverly to another space may lead to representations and genetic operators that do not need repair or where repair is trivial. Other more theoretical advantages of using a alternate representation of trees are discussed in section 6.

# 3. REPRESENTATION AND FITNESS

There are a few studies that have explored the idea of representing the tree using an alternate structure. Reijmers *et al.* [26] compared a distance matrix representation with a Prüfer number [25] representation. The distance matrix representation performed poorly; however, only one set of mutation/recombiantion operators was investigated for each representation; thus, it the operators might be responsible for the poor performance rather than the representation itself. For example, Cotta and Moscato [6] compared two different types of recombination and mutation operators with the same Prüfer-like representation, and observed differences in performance. Gottlieb *et al.* [14] came to a definite conclusion that Prüfer numbers were a poor representation since for almost all common genetic operators only a negligible fraction of the genotype space provides high locality and heritability. These studies lead to the conclusion that representations cannot be considered in isolation from the genetic operators that will be applied to them, and more importantly both the representation and the genetic operators should exhibit locality and heritability.

## 3.1 Genotype and Phenotype

The genotype is a distance matrix: a symmetric $n$-by-$n$ matrix with strictly positive entries in all off-diagonal entries (the diagonal entries are zero and irrelevant). The rows and columns represent the species and the entries conceptually correspond to some evolutionary distance between pairs of species; but, most importantly they serve as parameters that undergo cross-over and mutation and map onto specific phenotypes.

The phenotype is a strictly bifurcating un-rooted tree with $n$ leaves (the number of species). The internal nodes are the inferred common ancestors and trees with differently labelled internal nodes are considered distinct; the lengths of the edges (branches) represent evolutionary time. The topology and edge lengths together are required to calculate the fitness function.

## 3.2 Genotype to Phenotype Mapping

The NJ algorithm is used as a mapping from the genotype space $(\mathbb{R}^+)^M$ to the phenotype space $\mathbb{S} \otimes (\mathbb{R}^+)^N$. The genotype to phenotype mapping is akin to an artificial embryogeny [32]. However, artificial embryogeny, especially in the context of ALife, is often about the unfolding of complexity (often with phenotypes of arbitrary complexity) from very compact genotypes with the mapping mimicking some developmental process. Whereas, here, it is more important that the genotype space have efficient and convenient crossover and mutation operators, and that the phenotype space have efficient fitness calculations.

The NJ Algorithm takes a matrix of pairwise distances, identifies the pair of species which are closest (neighbours) and joins both of them to their most recent common ancestor, determining the edge lengths and creating one internal node in the tree. The two species are then set aside and replaced by the single new node, and a new matrix of distances is calculated. This matrix has one less row and column, and the algorithm is applied recursively until all species and internal nodes are joined.

The algorithm will construct a tree from any matrix of distances. However, if the entries in the matrix do not satisfy what is known as a four-point metric inequality [30, 13], then some of the tree edges will be assigned negative lengths. When negative lengths occur they are simply replaced by their absolute value and the distance matrix is recomputed from this repaired tree. It is observed that repair becomes less and less frequent as the GA algorithm proceeds, unless the data corresponds to trees with genuinely small internal edges that may be obscured by statistical noise.

Some of the beneficial properties of NJ are discussed here. Small changes in the entries of the distance matrix usually result in small changes to the edge lengths of the corresponding tree (thus offspring usually resemble parents and the population evolves smoothly from generation to generation). However, sometimes small changes in the distance matrix can change the *order* in which nodes are joined; this can have a domino effect and produce moderate to large changes in the topology of the tree (thus sometimes offspring can be quite novel). However, unlike the tree transformations discussed above, this change in topology approximately con-

serves the inter-species distances and is therefore less likely to produce highly detrimental changes in the fitness of offspring. Furthermore, the distance between a specific pair of species depends on the edge lengths that join these two species and is insensitive to changes and rearrangements in the topology of other parts of the tree. Therefore, changes which only affect some entries in the distance matrix will tend to preserve aspects of the tree corresponding to the unchanged distances. Thus again, topological transformations can occur that preserve certain inter-species relationships.

### 3.3 Fitness Calculation

The Felsenstein pruning algorithm [10] is used to calculate the fitness of each candidate tree. The algorithm uses the original nucleotide data and the choice of evolutionary model is independent of all other aspects of the GA. The model requires the specification of the expected frequencies $p_x$ of the nucleotide bases and a transition probability matrix $P_{xy}(t)$ that describes the probability that a nucleotide $x$ will have changed into a nucleotide $y$ after an elapsed evolutionary time $t$. In this paper, the simplest model proposed by Jukes and Cantor [16] is used where all nucleotides occur with equal probability and mutate into each with equal probabilities at a constant rate. Likelihood values are usually extremely small probabilities and thus it is conventional to quote and discuss all results in terms of the logarithm of the likelihood, which will be a negative number.

## 4. THE GENETIC ALGORITHM

### 4.1 Initialisation

The starting population candidates were obtained from random distance matrices. Provided the initial population is not lacking in diversity, the outcome of a good genetic algorithm should be independent of the starting configuration. However, the time to convergence will depend on the starting population, and in a practical situation one might also seed the initial population with candidates obtained by other algorithms or certain prior information. To observe the performance of the algorithm under the harshest conditions, it was decided not to direct the GA towards specific regions of the phase space by seeding or biasing the initial population.

### 4.2 Selection Operators

A standard fitness based tournament selection is used for reproduction. Each individual is guaranteed to compete in at least one tournament, but also has a statistical expectation of competing in $m-1$ additional tournaments, where $m$ is the population size divided by the tournament size. The number of tournaments is equal to the population size and the winner of each tournament is selected into the next generation. Each pair of individuals is then replaced by two new individuals using the cross-over or recombination operator described below. After some trial and error, a population size of 500 and tournament size of 5 was used.

### 4.3 Recombination-Mutation Operators

Special genetic operators that act on matrices were invented. In earlier work, Reijmers *et al.* [26] adopted the obvious strategy of randomly selecting each entry of the distance matrix from either parent. Thus individual entries in the matrix play the role of independently assorted Mendelian genes. Their GA exhibited poor performance and this can be attributed to the lack of heritability in such a disruptive operation.

The crossover or recombination operation used here is uniform crossover with *entire* rows of the distance matrices playing the role of individual genes. Cross-over is achieved by swapping some subset of rows of the distance matrix between two individuals and then "symmetrising" the resulting distance matrices. Specifically:

1. A number $p$ is chosen randomly from the interval $[0, 1]$.
2. For each species, the corresponding rows of the distance matrix are interchanged with probability $p$. Thus, on average, each child inherits about $pn$ rows (species) from one parent and $(1-p)n$ rows (species) from the other parent.
3. If species $i$ and $j$ were both exchanged or both not exchanged then the entries in the distance matrix satisfy $d_{ij} = d_{ji}$. (The inter-species distance is thus also inherited unchanged from one parent.) No symmetrisation is required for such entries.
4. For those entries where $d_{ij} \neq d_{ji}$ a real-parameter stochastic average crossover is needed.

In the absence of selection pressure a good crossover operator should not reduce the diversity of the population (or change its average properties). Three well-known real parameter crossover operators that have the above properties are blend crossover BLX-$\alpha$ [9], Voigt's fuzzy recombination [34] and simulated binary crossover [7]. BLX was used since it is the easiest to implement, and after some trial and error a value of $\alpha = 1.0$ was chosen. The use of BLX also introduces new parameters into the population, therefore it is not immediately obvious that an independent mutation operator is also required. The large value of $\alpha$ is, in a sense, enhancing the role of mutation. In the interest of simplicity, no additional mutation was introduced, however the benefits of additional mutation or different values of $\alpha$ will be explored in future work.

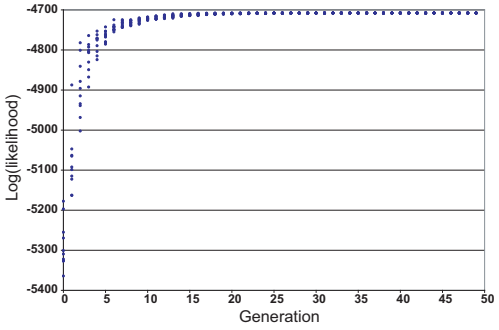
## 5. PERFORMANCE ON SIMULATED DATA

### 5.1 Simulation Method

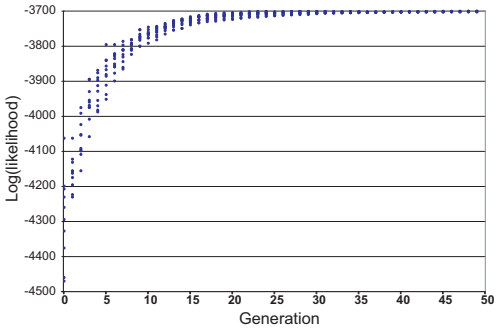
The benefits of using simulated data are that the evolutionary model, tree topology and edge lengths used to generate the nucleotide sequences are known and thus the results of the optimisation can be compared to these values. For each of the cases with 7, 10 and 20 species, a distance matrix was created with random entries in the range 0 to 1. The NJ algorithm was used to construct a tree. These three tree topologies are shown in Fig. 1. Nucleotide sequences of length 500 were then generated by simulating evolution on these trees. The simulation is essentially the inverse of the likelihood calculation. Starting from any one node, a random sequence of nucleotides is generated for this node. The probability matrix $P_{xy}(t)$ is then used to create a sequence of nucleotides at neighbouring nodes. A length of 500 nucleotides was used to keep the time taken to perform the fitness calculations down. In analyses with moderate numbers of species, it has been suggested that sequences of several thousand nucleotides are required for statistically

significant results. The analysis was also compared to the results from the widely used PHYLIP program [11] to compare the final trees and likelihood values. In every case, PHYLIP recovered the tree topology used to simulate the data, with edge lengths within a few percent of the original values. With only 500 nucleotides, small difference are expected due to statistical fluctuations during simulations, The edge-lengths shown in Fig. 1 are drawn proportional to the lengths found by the PHYLIP program.

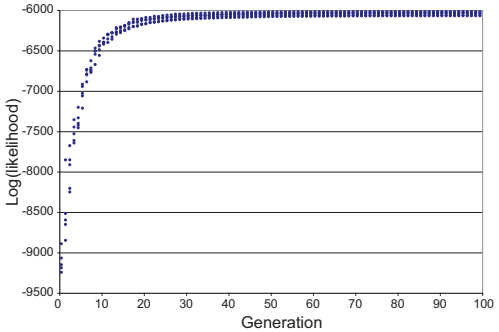
The GA was run 10 times for each tree using the simulated nucleotide sequences as input. Information was collected on the best likelihood value in each generation, the number of novel topologies emerging in each generation and the average and maximum number of generations that particular tree topologies persisted.




(a) GA-best -4707.64; PHYLIP -4707.64




(b) GA-best -3703.28; PHYLIP -3694.38




(c) GA-best -5999.89; PHYLIP -5909.47

**Figure 2: The log(likelihood) from each of the 10 runs as a function of generation for (a) 7 (b) 10 (c) 20 species. In each case the best value from the GA is compared to that from PHYLIP. N.B. The results for 20 species are shown over 100 generations.**

## 5.2 Convergence

In every run, the GA uncovered the tree topology used to simulate the data. However, as the number of species increased the best likelihood value attained by the GA fell slightly short of the maximum found by PHYLIP. The edge lengths of the trees in the final population also disagreed by a few to several percent. These results suggest that although this GA was good at searching the combinatorial space of topologies it could probably gain by being hybridised with a more conventional local search to optimise the edge lengths of the trees. The convergence of the log-likelihood with generation and the best values found are shown in Fig. 2.

In the 7 species case, the final populations contained only the correct tree topology, but with up to 4 different versions of labelling of the internal nodes. This occurs because the edge lengths of the optimal tree topology have not sufficiently converged and alternate orderings of the neighbour joining takes place. This is another indication that a hybrid approach is necessary.

In the 10 species case, the final populations contained two distinctly labelled versions of the same tree topology. The results are somewhat better than expected: the presence of a very short internal branch (as indicated by the circle in Fig. 1) combined with the lack of tight convergence of the edge lengths should lead to some alternate topologies persisting in the population.

In the 20 species case, three internal nodes of the tree were slightly problematic. These nodes are indicated on the tree in Fig. 1. The final populations contained the correct topology, but also contained different topologies corresponding to slightly different connections between these three internal nodes. In some cases as many as a dozen different labelled topologies and several distinct tree shapes still exist in the population even after 100 generations.

Table 2 shows the degree of edge length convergence for the distinct labelled topologies surviving in the final population of one of the 7 species runs. After only 50 generations, the edge lengths have converged to better than 3% for those connected to the species, and the internal edge lengths to 6%.

**Table 2: The four best (surviving) labelled topologies in the 7 species case. All have a log likelihood of -4707.9.**

| $i$ | $j, d_{ij}$ | | | | Phylip |
|---|---|---|---|---|---|
| 1 | 8 , 0.343 | 8,0.340 | 11, 0.346 | 12, 0.339 | 0.344 |
| 2 | 8 , 0.956 | 8,0.952 | 11, 0.954 | 12, 0.954 | 0.947 |
| 3 | 9 , 0.497 | 9,0.500 | 12, 0.506 | 11, 0.504 | 0.510 |
| 4 | 11, 0.675 | 12,0.668 | 9 , 0.666 | 9 , 0.661 | 0.685 |
| 5 | 12, 0.814 | 11,0.818 | 8 , 0.818 | 8 , 0.816 | 0.820 |
| 6 | 12, 0.362 | 11,0.358 | 8 , 0.359 | 8 , 0.358 | 0.354 |
| 7 | 10, 0.313 | 10,0.312 | 10, 0.314 | 10, 0.321 | 0.427 |
| 8 | 9 , 0.224 | 9 ,0.224 | 9 , 0.234 | 9 , 0.240 | 0.219 |
| 9 | 10, 0.845 | 10,0.846 | 10, 0.820 | 10, 0.812 | 0.695 |
| 10 | 11, 0.822 | 12,0.818 | 12, 0.835 | 11, 0.842 | 0.711 |
| 11 | 12, 0.226 | 12,0.234 | 12, 0.220 | 12, 0.226 | 0.211 |

## 5.3 Exploration vs. Exploitation

An interesting feature of any GA to try to understand is the balance between exploration and exploitation. In this GA, exploration is equivalent to topological innovation, that

is, the creation of labelled tree topologies that did not exist in earlier generations. It is important in early generations to search the space broadly and find many diverse candidates to avoid premature convergence to a local optimum. Later, in the exploitation phase, it is more important to refine the fitness of the best candidates and spend less time exploring since the chance of finding worthwhile novel trees is less.

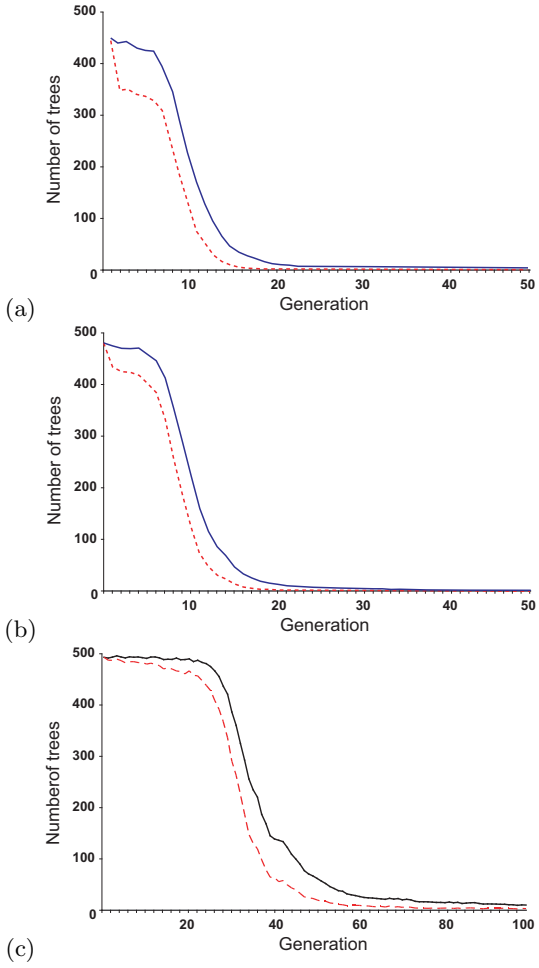



**Figure 3: The solid (upper) curve shows the number of distinct labelled trees in each generation. The dashed (lower) curve shows the number of novel trees in each generation (i.e. labelled trees that did not exist in the previous generation). Results are averages over 10 different runs for (a) 7 (b) 10 and (c) 20 species.**

The number of distinct trees and the emergence of novel trees per generation is shown in Fig. 3. Another indicator of progress is the persistence of good solutions; in this case: the number of generations that a labelled tree topology will exist in the population before disappearing. The average and maximum lifetime of the tree topologies as a function of the generation in which they first occurred is shown in Fig. 4.

For the 7 and 10 species cases, the exploration phase lasted about 5 to 7 generations; for the 20 species case, though, the exploration continued for more than 20 generations. In the exploration phase many new trees are discovered; on average these trees last less than two generations and even the most persistent ones survive less than 10 generations.

At some point, trees are discovered that essentially remain in the population for very long periods of time (sometimes forever). For 7 and 10 species this occurs around generation 10; for 20 species it is around generation 30 to 40. These are the trees that very strongly resemble the optimal tree topologies. The global peaks in Fig. 4 reveal the generation in which indefinitely persisting trees (i.e. the optimal topology) first appear.

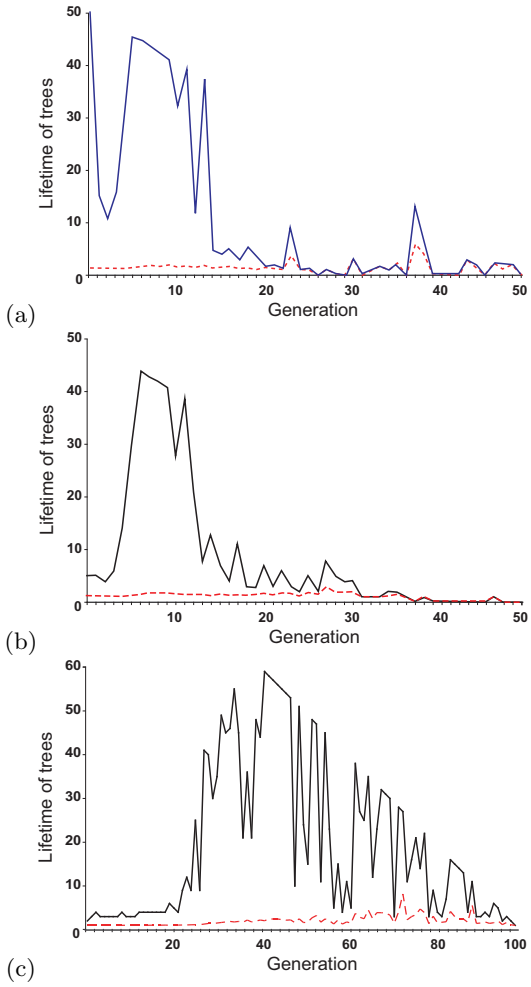



**Figure 4: The solid (upper) curve shows the maximum observed lifetime of trees born in that generation. The dashed (lower) curve shows the average lifetime of trees born in that generation. Results are averages over 10 different runs.**

After this discovery phase, new tree topologies are still discovered but most do not last very long, and most of the effort of the GA goes into refining the edge lengths of the trees already found. This is the consolidation or exploitation phase.

In the 7 species case, during the 10 runs, between 4060 and 5229 different labelled trees were examined. Depending on the run, the last novel tree appears sometime between generation 19 and 48. In most cases, there were two to four distinct labelled trees equally populating the last generation.

In the 10 species case, between 4892 and 5689 different labelled trees were examined. Depending on the run, the last novel tree appears sometime between generation 27 and 38. In every case, there were two distinct labelled trees equally populating the last generation.

In the 20 species case, between 17911 and 20332 trees were examined. New trees emerged at all stages up to 100 generations.

In the 7 species case, about 5% of the total number of possible trees are explored before the best topology is found. In the 10 species case, only slightly more trees are explored but this represents only $10^{-5}$% of the total space. Four times as many trees are explored in the 20 species case, although the combinatorial space is 20 orders of magnitude larger. This observation in itself suggests the GA has good scaling properties.

# 6. CONCLUSION

The preliminary analysis of the three cases studied here revealed both positive and negative attributes of the GA developed here. The behaviour of the GA can be considered in terms of three phases: an initial exploration phase, followed by discovery of good solutions (or partial solutions) and a final exploitation phase. The lengths of these phases varies with the number of species. One expects that the lengths of these phases should also depend on the parameters of the GA (such as population and tournament size) and this is a study that needs to be conducted next.

Much longer studies are needed to confirm typical convergence times for cases with 20 or more species, but the results here do indicate that the length of the exploration phase grows slowly with the number of species. More analysis is needed to determine the precise relationship (and hopefully confirm a low power law behaviour). Certainly, the search spaces varied by a very large number of orders of magnitude with only 3 to 4 times variation in the time taken to locate some good solutions. However, the above suggests that the algorithm has good initial behaviour and a reasonable balance of exploration and exploitation.

In all the cases tested, the GA found the original tree topology used to simulate the data, which agreed with the optimal solution found by PHYLIP. The algorithm did make a smooth transition to a more exploitative phase in later generations, but the performance of the algorithm in refining the values of the edge lengths was a bit slow. Slightly more worrying is that although the correct topology was present in the final population, several alternate topologies were also still present in the 20 species case after 100 generations. These alternate topologies sometimes appear to have higher likelihoods, but only because the edge lengths of these trees have yet to be accurately refined.

The persistence of these non-optimal topologies is hoped to be resolved by the same mechanism planned to resolve the slow convergence of the edge lengths: hybridisation with a specialised approach to refining the numerical values once or while the correct topology is being discovered.

A more detailed statistical analysis of the algorithm is worthwhile and underway as well as studying its performance on larger numbers of species and the incorporation of a special local optimiser.

In terms of a more theoretical analysis of the algorithm, a disadvantage of working directly with the trees has been the concept of the distance between two tree topologies in tree-space. There are a number of metrics that have been discussed in the literature. Some do not include any information on edge lengths, those that do include the Robinson-Foulds metric [27] and the branch score [18]. More recently, Holmes [15] has explored the need for good metrics in developing statistics for phylogenetic trees, and probability distributions in particular. Mapping from one space to another does allow a metric in one space to induce one in the other, and this may contribute to the discussion. However, it still remains a challenging question to quantitatively describe how various genetic operators reduce, preserve or increase diversity or flow through the phenotype space. By mapping to a simpler space, where well-defined distance metrics can be applied to matrices, some of these attributes may be more easily quantified. A better understanding of these concepts will also allow more sophisticated statistical approaches such as Bayesian inference to be applied.

# 7. ACKNOWLEDGMENTS

The author acknowledges the financial support of the Australian Research Council. I would also like to thank Lars Jermiin and Susan Holmes for useful discussions, Alan Huang for some programming, and the reviewers for their helpful comments and contributions.

# 8. REFERENCES

[1] M. J. Brauer, M. T. Holder, L. A. Dries, D. J. Zwickl, P. O. Lewis, and D. M. Hillis. Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. *Mol. Biol. Evol.*, 20:1717 – 1726, 2002.

[2] B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Mol. Biol. Evol.*, 17:1529–1541, 2000.

[3] C. B. Congdon. Gaphyl: A genetic algorithms approach to cladistics. In L. DeRaedt and A. Siebes, editors, *Principles of Data Mining and Knowledge Discovery*, pages 67–68. Lecture Notes in Computer Science, Vol 2168, Springer, Berlin, 2001.

[4] C. B. Congdon. Gaphyl: an evolutionary algorithms approach for the study of natural evolution. In *Genetic and Evolutionary Computation Conference*, pages 1057–1064, San Francisco, California, 2002.

[5] C. B. Congdon. Phylogenetic trees using evolutionary search: Initial progress in extending gaphyl to work with genetic data. In *Congress on Evolutionary Computation*, Canberra, Australia, 2003.

[6] C. Cotta and P. Moscato. Inferring phylogenetic trees using evolutionary algorithms. *Lecture Notes in Computer Science*, 2439:720–729, 2002.

[7] K. Deb and R. B. Agrawal. Simulated binary crossover for continuous search space. *Complex Systems*, 9:115–148, 1995.

[8] A. W. F. Edwards. Estimation of the branch points of a branching diffusion processes. *J. Royal Stat. Soc.*, 32:155–174, 1970.

[9] L. J. Eshelman and J. D. Schaffer. Real-coded genetic algorithms and interval-schemata. In L. D. Whitley, editor, *Foundations of Genetic Algorithms*, volume 2, pages 187–202. Morgan-Kauffman Publishers, San Mateo, CA, November 1993.

[10] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[11] J. Felsenstein. *PHYLIP (Phylogeny Inference Package) Version 3.5c.* Department of Genetics, University of Washington, Seattle, http://evolution.genetics.washington.edu/phylip.html, 1993.

[12] J. Felsenstein. *Inferring Phylogenies.* Sinauer Associates, Inc, Sunderland, Mass., 2004.

[13] W. M. Fitch. A non-sequential method for constructing trees and heirachical classifications. *J. Mol. Evol.*, 18:30–37, 1981.

[14] J. Gottlieb, B. A. Julstrom, G. R. Raidl, and F. Rothlauf. Prüfer numbers: A poor representation of spanning trees for evolutionary search. In L. Spector, E. Goodman, A. Wu, W. B. Langdon, H.-M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. Garzon, and E. Burke, editors, *Proceedings of the Genetic and Evolutionary Computation Conference 2001*, pages 343–350, San Francisco, California, 2001.

[15] S. Holmes. Statistics for phylogenetic trees. *Theor. Pop. Biol.*, 63:17–32, 2003.

[16] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In M. N. Munro, editor, *Mammalian Protein Metabolism, Vol III*, pages 21–132. Academic Press, New York, 1969.

[17] K. Katoh, K. Kuma, and T. Miyata. Genetic algorithm-based maximum-likelihood analysis for molecular phylogeny. *J. Mol. Evol.*, 53:477–484, 2001.

[18] M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11:459–468, 1994.

[19] A. R. Lemmon and M. C. Milinkovich. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *PNAS*, 99:10516–10521, 2002.

[20] P. O. Lewis. A genetic algorithm for maximum likelihood phylogeny inference using nucletide sequence data. *Mol. Biol. Evol.*, 15:277–283, 1998.

[21] H. Matsuda. Protein phylogenetic inference using maximum likelihood with a genetic algorithm. In L. Hunter and T. Klein, editors, *Pacific Symposium on Biocomputing '96*, pages 512–523, London, 1996. World Scientific.

[22] A. Moilanen. Searching for most parsimonious tree with simulated evolutionary optimisation. *Clad.*, 15:39–50, 1999.

[23] L. Poladian and L. S. Jermiin. Phylogenetic inference using evolutionary multi-objective optimisation. In *Genetic and Evolutionary Computation Conference*, Seattle, Washington, 2004.

[24] L. Poladian and L. S. Jermiin. What might evolutionary algorithms (EA) and multi-objective optimisation (MOO) contribute to phylogenetics and the total evidence debate. In *Genetic and Evolutionary Computation Conference*, Seattle, Washington, 2004.

[25] H. Prüfer. Neuer beweis eines satzes ueber permutationen. *Archiv für Mathematik und Physik*, 27:742744, 1918.

[26] T. H. Reijmers, R. Wehrens, and L. M. C. Buydens. Quality criteria of genetic algorithms for construction of phylogenetic trees. *J. Comp. Chem.*, 20:867–876, 1999.

[27] D. F. Robinson and L. R. Foulds. Comparison of weighted labelled trees. In A. F. Horadam and W. D. Wallis, editors, *Combinatorial Mathematics VI. Proceedings of the Sixth Australian Conference on Combinatorial Mathematics*, volume 53, pages 119–126, Berlin, 1979. Springer-Verlag.

[28] N. Saitou and M. Nei. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[29] M. Salemi and A.-M. VanDamme. *Handbook of phylogenetic methods.* Cambridge University Press, Cambridge, 2003.

[30] S. Sattah and A. Tsversky. Additive similarity trees. *Psychometrika*, 42:319–345, 1977.

[31] J. Shen and R. B. Heckendorn. Discrete branch length representations for genetic algorithms in phylogenetic search. *Lecture Notes in Computer Science*, 3005:94–103, 2004.

[32] K. O. Stanley and R. Miikkulainen. A taxonomy for artificial embryogeny. *Artificial Life*, 9:93–130, 2003.

[33] J. A. Studier and K. J. Keppler. A note on the neighbour-joining algorithm of saitou and nei. *Mol. Biol. Evol.*, 5:729–731, 1988.

[34] H. M. Voigt, H. Mühlenbein, and D. Cvetkovic. Fuzzy recombination for the breeder genetic algorithm. In L. Eshelman, editor, *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 104–111, San Mateo, CA, 1995. Morgan-Kauffman Publishers.