Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 5th, 2006

CONSENSUS OF PHYLOGENETIC TREES USING GENETIC ALGORITHMS

Michael L. Gargano, mgargano@pace.edu

Pace University Seidenberg School of Computer Science and Information Systems, Computer Science Dept., NYC, NY 10038

ABSTRACT

The problem of combining phylogenetic trees (also called evolutionary trees or dendrograms) based on the findings of various experts (or algorithms) in order to select a representative tree that is a consensus of the experts is considered. This research proposes using a genetic algorithm (GA) employing a novel phylogenetic tree code to solve this problem. This encoding scheme insures feasibility after performing the operations of crossover and mutation and also insures the feasibility of the initial randomly generated population (i.e., generation 0).

Keywords: consensus, phylogenetic tree, dendrogram, genetic algorithm

Introduction to the Problem

The evolution of flora, fauna, DNA structures, Roman spearheads, etc. into a given set of taxa can be intuitively and/or objectively reconstructed by experts to form a phylogenetric tree representing the splitting at each bifurcation of an initial population into the final set of taxonomical units. This problem is concerned with finding a consensus tree that "faithfully" represents and reflects the combined trees of many different experts and/or algorithms. This will provide a greater robustness overcoming an individual expert's methodologies, biases or other inadequacies.

The problem of combining the mathematical taxonomic findings (about possible patterns of evolution resulting in the given set of taxa) (i.e., phylogenetic, evolutionary, dendrogramic trees) of various experts (in

archaeology, zoology, bioinformatics, etc.) in order to select a representative tree that is a consensus of these is considered.

Given a set of expert trees, we wish to find a consensus that minimizes the sum of the distances between the consensus tree and all of the expert trees. This is a standard method used by mathematicians to quantify the difference between the two trees.

Here is an example of one expert's phylogenetic tree:

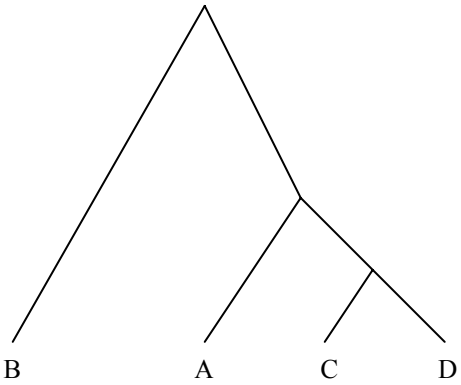


Figure 1. A phylogenetic tree with 4 taxa (i.e., A, B, C, D) and inter-leaf distance matrix D:

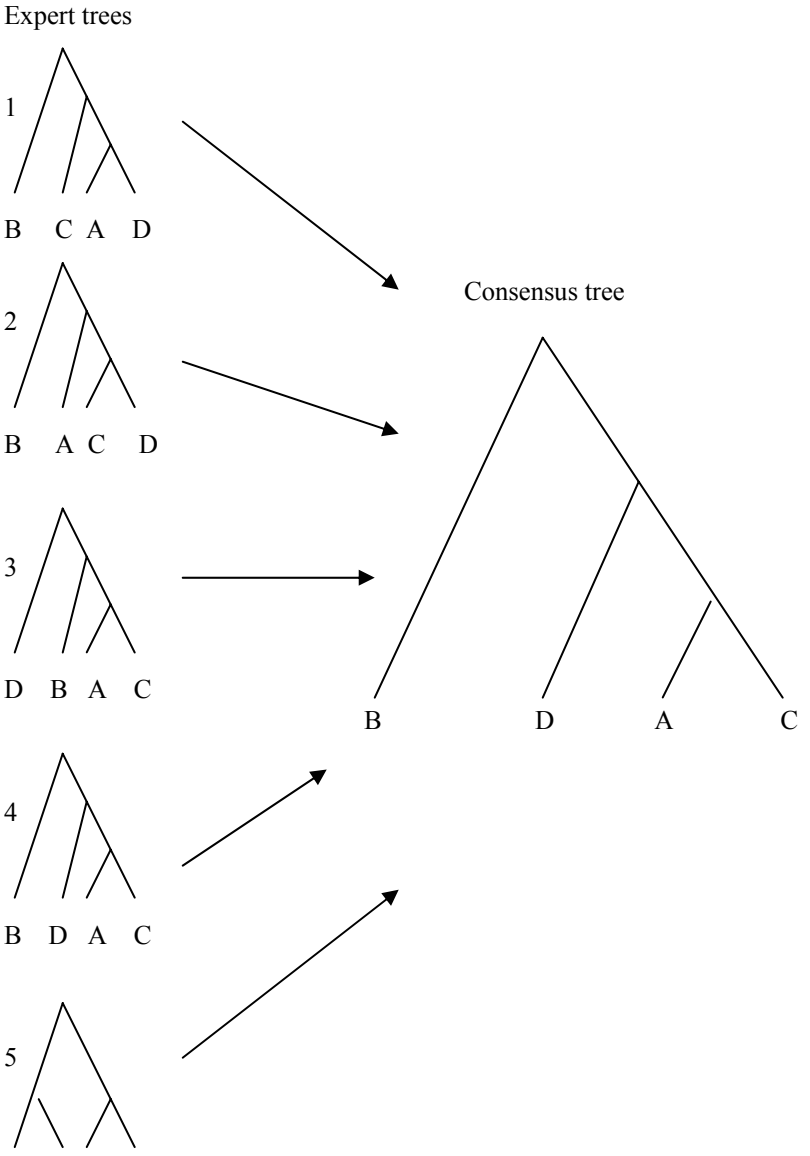
0	3	3	3
3	0	4	4
3	4	0	2
3	4	2	0

The genome (i.e., code) for this phenome (i.e., phylogenetic tree) is 3 1 1.

Mathematical Model

A phylogenetic tree is a rooted tree whose leaves are taxonomical units. The internal branches represent a splitting of the original population or subpopulation into two new subpopulations. Each tree is associated with an inter-leaf distance matrix that holds the distance from any leaf in the tree to any other leaf. Each expert given a set of known taxanomical units reconstructs what they believe to be the true the evolutionary process resulting in the known taxa. The problem is to find a tree (i.e., a consensus tree) whose inter-leaf distance matrix is closest to all the expert inter-leaf distance matrices. Thus, a consensus tree "faithfully" represents and reflects the combined trees of many different experts and/or algorithms. This will provide greater robustness

overcoming an individual expert's methodologies, biases or other inadequacies. Since finding a consensus tree is NP hard, we will use a genetic algorithm.
Figure 2. Consider a simple example of the consensus problem. Suppose we receive the resulting trees (with t = 4 taxa) from five different experts (M = 5) and wish to form a consensus tree using the distance between trees described above.



A C B D

Genetic Algorithm Methodology

A **genetic algorithm (GA)** is a biologically inspired, highly robust heuristic search procedure that can be used to find optimal (or near optimal) solutions to NP hard problems. The GA paradigm uses an adaptive methodology based on the ideas of Darwinian natural selection and genetic inheritance on a population of potential solutions. It employs the techniques of crossover (or mating), mutation, and survival of the fittest to generate new, typically fitter members of a population over a number of generations [1, 2, 3].

We propose GAs for solving this problem using a novel encoding scheme. Our GAs create and evolve an encoded population of potential solutions so as to facilitate the creation of new *feasible* members by standard mating and mutation operations. (A feasible search space contains only members which satisfy the problem constraints, that is, a dendrogram [4, 5, 6, 7, 13,14].) When feasibility is not guaranteed, numerous methods for maintaining a feasible search space have been addressed in [11], but most are elaborate and complex. They include the use of problem-dependent genetic operators and specialized data structures, repairing or penalizing infeasible solutions, and the use of heuristics.) By making use of problem-specific encodings, our problem insures a *feasible* search space during the classical operations of crossover and mutation and, in addition, eliminates the need to screen during the generation of the initial population.

We adapted many of the standard GA techniques found in [1, 2, 3] to this problem. A brief description of these techniques follows. Selection of parents for mating involves randomly choosing one very fit member of the population (i.e., one with a small total distance) and the other member randomly. The reproductive process is a simple crossover operation whereby two randomly selected parents are cut into sections at some randomly chosen positions and then have the parts of their encodings swapped to create new offspring (children). In our application the crossover operation produces an encoding for the offspring that have element values that always satisfy the position bounds (i.e., range constraints). Mutation is performed by randomly choosing a member of the population, cloning it, and then changing values in its encoding at randomly chosen positions subject to the range constraints for that position. A grim reaper mechanism replaces low scoring members in the population with newly created more fit offspring and mutants. Our fitness measure will be the smallest total distance between trees. The GA is terminated when, for example, either no improvement in the best fitness value is observed for a number of generations, a certain number of generations have been examined, and/or a

satisficing solution is attained (i.e., the result is not necessarily optimum, but is satisfactory).

The Generic Genetic Algorithm

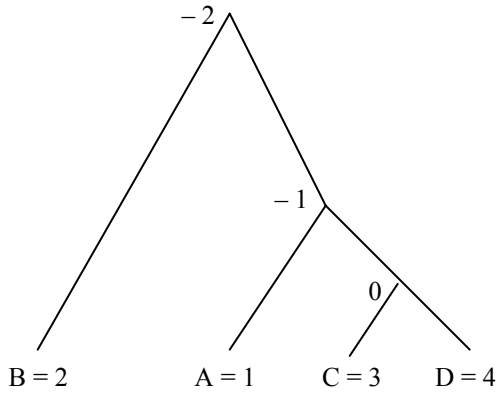
We can now state the generic genetic algorithm we used for each application:

1) Randomly initialize a population of encoded potential solutions.
2) Map each population member to its equivalent phenome.
3) Calculate the fitness of any population member not yet evaluated.
4) Sort the members of the population in order of fitness.
5) Randomly select parents for mating and generate and evaluate offspring using crossover.
6) Randomly select and clone members of the population to generate and evaluate mutants.
7) Sort all the members of the expanded population in order of fitness.
8) Use the grim reaper to eliminate the population members with poor fitness.
9) If (termination criteria is met) then return best population member(s)
   else go to step 5.

Encodings

In this problem the population of potential solutions with t taxa consists of possible consensus trees. Each tree has internal nodes with nonpositive values $-(t-2), -(t-1), \ldots, 0$ and leaves with positive values $1, 2, \ldots, t$ (we can think of A = 1, B = 2, etc.).

Here is the example of an expert's phylogenetic tree from Figure 1 with t = 4.



The value – (t – 2) is always considered the root. When the root bifurcates, there may be either two internal nodes or one internal node and one leaf. At any bifurcation of on internal node there may be either two internal nodes (the lowest valued and any other one) or one internal node (the lowest valued) and one leaf when there are still internal nodes, else there will be two leaves. In the above example – 2 bifurcates into – 1 next lowest valued internal node and the leaf 2. Then – 1 bifurcates into internal node 0 (the next lowest) and the leaf 1. Finally, since there are no internal nodes left 0 bifurcates into 3 (lowest) and 4.

Beginning with the sequence – 2 – 1 0 1 2 3 4
– 2 is the root and the next lowest value is – 1 so five positions remain
0 1 2 3 4
since B = 2 is in the third position, 3 goes into the first position of the phylogenetic code (i.e., the genome). The next lowest value is 0, now three positions remain 1 3 4
and since the leaf A = 1 is in the first position, 1 goes into the second position of the phylogenetic code. Finally, the lowest is 3, and one value is left
4
since 4 is in the first position, 1 goes into the last position of the code.

Thus, the genome (i.e., code) for this phenome (i.e., tree) is 3 1 1.

This process is easily reversed so that one may retrieve the tree from the code. Notice that there were 5 choices for the first position , 3 for the second, and 1 for the last. Therefore there are $5 \cdot 3 \cdot 1 = 15$ possible phylogenetic trees with t = 4 leaves. Here is a theorem for the number of rooted phylogenetic trees with t leaves.

. **Theorem**: The number of rooted phylogenetic trees with t taxonomical units for leaves is $((2t-3)!) / ((t-2)!\ 2^{t-2})$

Here's another example with the code 4 1 3 1. First notice that there are t = 5 = 4 + 1 leaves so that we consider the sequence
– 3 – 2 – 1 0 1 2 3 4 5
– 3 is the root that splits into – 2 (the smallest) and 2 the 4th position of
– 1 0 1 2 3 4 5 7 values

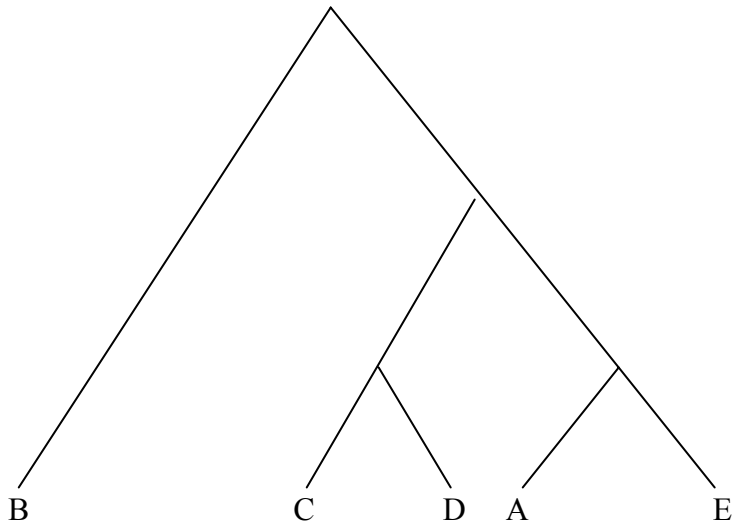
– 2 splits into – 1 (the smallest) and 0 the 1st position of
0 1 3 4 5 5 values

– 1 splits into 1 (the smallest) and 5 the 3rd position of
3 4 5 3 values

0 splits into 3 (the smallest) and 4 the 1 st position of
4 1 value

Notice that there are $7 \cdot 5 \cdot 3 \cdot 1 = 105$ possible phylogenetic trees with $t = 5$ leaves.

The phylogenetic tree with the code 4 1 3 1.



Fitness

Each tree is associated with an inter-leaf distance matrix that holds the distance between any two leaves in the graph (tree). For example the matrix of expert 1 in figure two is given by

$$D_1 = \begin{bmatrix} 0 & 4 & 3 & 2 \\ 4 & 0 & 3 & 4 \\ 3 & 3 & 0 & 3 \\ 2 & 4 & 3 & 0 \end{bmatrix}$$

While the matrix for the consensus tree is

$$C = \begin{bmatrix} 0 & 4 & 2 & 3 \\ 4 & 0 & 4 & 3 \\ 2 & 4 & 0 & 3 \\ 3 & 3 & 3 & 0 \end{bmatrix}$$

The distance between these two trees is defined as

$D(C, D_1) = \sum (d_1(i, j) - c (i, j))^2$ summed over all pairs of leaves and is equal to 8. The fitness is

$$\text{Fitness } (C) = \sum_e D(C, D_e)$$ summed over all the expert trees and is equal to 52.

Conclusions

The problem of combining phylogenetic trees (also called evolutionary trees or dendrograms) based on the findings of various experts (or algorithms) in order to select a representative tree that is a consensus of the expert trees was considered. Since finding such a consensus list is an NP hard problem, it was solved using a genetic algorithmic heuristic employing a novel phylogenetic tree code.

Ackowledgements

I wish to thank Pace University's Seidenberg School of Computer Science and Information Systems for partially supporting this research. I also wish to thank DIMACS at Rutgers University for a workshop that introduced me to this interesting and important problem.

References

[1] M. Mitchell, An Introduction to Genetic Algorithms, MIT Press, (2001).

[2] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, (1989).

[3] L. Davis, Handbook of Genetic Algorithms, Van Nostrand Reinhold, (1991).

[4] M. L. Gargano and W. Edelson, A Genetic Algorithm Approach to Solving the Archaeology Seriation Problem, Congressus Numerantium 119, (1996) pp. 193-203.

[5] W. Edelson and M. L. Gargano, Minimal Edge-Ordered Spanning Trees Solved By a Genetic Algorithm with Feasible Search Space, Congressus Numerantium 135, (1998) pp. 37-45.

[6] F. S. Roberts, Discrete Mathematical Models, Prentice-Hall Inc., (1970).

[7] F. S. Hillier and G. J. Lieberman, Introduction to Operations Research, Holden-Day Inc. (1968).

[8] K. H. Rosen, Discrete Mathematics and Its Applications, Fourth Edition, Random House (1998).

[9] M.L. Gargano, W. Edelson, Optimally Sequenced Matroid Bases Solved By A Genetic Algorithm with Feasible Search Space Including a Variety of Applications, Congressus Numerantium 150, (2001) pp. 5-14.

[10] M.L. Gargano, Maheswara Prasad Kasinadhuni , Self-adaption in Genetic Algorithms using Multiple Genomic Redundant Representations, Congressus Numerantium 167, (2004) pp. 183-192.

[11] G. Dunn and B.S. Everitt, An Introduction To Mathematical Taxonomy, Dover Publications (1980).

[12] C. Orton, Mathematics in Archaeology, Cambridge University Press, (1980).

[13] A.R. Lemmon, M.C. Milinkovitch, The Metapopulation Genetic Algorithm: an Efficient Solution to the Problem of Large Phylogeny Estimation, PNASvol. 99, no.16, pp. 10516-10521.