DOI: http://dx.doi.org/10.7551/978-0-262-31050-5-ch015

Comparing Distance-Based Phylogenetic Tree Construction Methods Using An Individual-Based Ecosystem Simulation, EcoSim

Ryan Scott[1], Robin Gras[1]

[1]University of Windsor
scotto@uwindsor.ca

Abstract

Phylogenetic trees are constructed frequently in biological research to provide an understanding of the evolutionary history of the organisms being studied. Often, the actual phylogenetic tree is unknown and the phylogenetic tree constructed is an estimate. There are many methods of phylogenetic tree construction which fall into two main categories: distance-based methods and character-based methods. To test the accuracy of these methods, it is necessary that the system being studied is one for which the actual phylogenetic tree is known. EcoSim is an ecosystem simulation in which predator and prey agents possessing a complex behavioral model can interact, evolve and speciate. In this experiment, we used EcoSim to test the accuracy of the three main distance-based phylogenetic tree construction methods, when constructing a single tree and when performing phylogenetic bootstrapping. Since EcoSim provides data regarding speciation events, we were able to construct the actual phylogenetic trees from this data. We then performed the UPGMA, Neighbor-Joining, and Fitch-Margoliash methods at various time-steps and used symmetric distance as a metric to compare the topologies of the actual and estimated trees. On average, trees contained nearly 30 taxa. We found that the Fitch-Margoliash method with bootstrapping performed slightly better than the other methods, however no method constructed trees in which more than 50% of the partitions were correct.

Keywords: evolution, ecosystem, individual-based model, distance-based, phylogeny, consensus, speciation, phylogenetic bootstrapping.

Introduction

An interesting topic in biology is the construction of phylogenetic trees. Phylogenetic trees are constructed in an attempt to reconstruct the evolutionary past; to develop an understanding of when and which speciation events may have occurred to give rise to the organisms exhibited today. A phylogenetic tree consists of edges, internal nodes, and external nodes (leaves). Leaves represent operational taxonomic units (OTUs) which are the actual species from which data was gathered to construct the tree. The internal nodes are hypothetical taxonomic units (HTUs). They represent the hypothetical last common ancestors to all other species arising from them. The edges often represent the relatedness or genetic distance between two nodes, where a shorter edge length means species are more closely related. In some trees, edges may be considered an estimation of the time taken between speciation events. In the study of real organisms, constructed phylogenetic trees are often an estimate of the real phylogenetic tree, since the actual phylogenetic tree is usually unknown. Given different data types, there are many different methods that researchers can employ to estimate phylogenetic trees. There are two main groups of phylogenetic tree reconstruction methods: distance-based methods and character-based methods (consisting of subgroups parsimony, compatibility, and maximum likelihood methods) (Felsenstein, 1988).

Distance-based methods could rely on many different types of data to perform analysis including genetic distance from sequences, distances from immunological studies, and Euclidean distance applied in various ways (Wiley and Lieberman, 2011). In terms of distance-based phylogenetic tree construction methods, there are three methods that are more common: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sneath and Sokal, 1973), Neighbor-Joining (NJ) (Saitou and Nei, 1987), and Fitch-Margoliash (Fitch and Margoliash, 1967). Each algorithm has some known properties or cases in which the tree should be very similar to the actual tree. The UPGMA algorithm should produce a correct tree if the distance data is ultrametric, which also means that the evolutionary rates among taxa are constant. This is rarely the case in nature. The Neighbor-Joining algorithm and Fitch-Margoliash method perform well when the distance data is additive. Again, this is usually not the case either. These methods generate a single tree for any given distance matrix. Of the three methods, UPGMA is the most computationally efficient; the algorithm for UPGMA is of complexity $O(n^2)$ (Murtagh, 1984). The Neighbor-Joining algorithm is of complexity $O(n^3)$ (Mailund et al, 2006), and the least efficient of the three, the Fitch-Margoliash method, runs in complexity of $O(n^4)$ (Lespinats et al, 2011). Since distance matrices can be generated from pairwise Euclidean distance data, distance matrices usable in phylogenetic tree construction could be generated using Euclidean distances between points in n-dimensional space. Character-based methods can rely on a variety of phylogenetic characters such as genetic, morphological, behavioral, and molecular attributes to construct phylogenetic trees. Provided that there

Downloaded from http://direct.mit.edu/isal/proceedings-pdf/alife2012/24/105/1901132/978-0-262-31050-5-ch015.pdf by guest on 19 April 2024

© 2012 Massachusetts Institute of Technology

is variation among taxa in the attribute and that the attribute is heritable, it could potentially be used as a phylogenetic character (Grandcolas et Al, 2001). The characters, if necessary, may be discretized to allow for discrete character states to be generated (Wiley and Lieberman, 2011). The algorithms used to create phylogenetic trees using phylogenetic characters are generally more complex than distance-based methods (Felsenstein, 1988). Generally, these algorithms are based on an optimization criterion such as parsimony, maximum likelihood, or compatibility (Felsenstein, 1988). Character-based methods are quite commonly used in studies of nature, because it is said that data is lost when converting data for use with distance-based methods (Felsenstein, 1988). In this experiment, we focus solely on distance-based methods because we are not dealing with data from a real biological system, we are instead dealing with data that does not contain character-based attributes but instead contains numerical attributes for which it is more appropriate to use distance-based methods. Furthermore, character-based methods tend to be far more computationally complex.

A common practice in phylogenetic tree construction is bootstrapping, in order to test the repeatability of the results (Felsenstein, 1983). Bootstrapping is a resampling method in which the original data is resampled with replacement of characters (Felsenstein, 1983). Bootstrapping allows one to observe in what proportion of trees a particular partition of the tree is represented when data is resampled without removing data. Commonly, a large number (100-1000) of such resamplings are carried out. From these 100-1000 trees generated from bootstrapping, a single tree is generated that contains only the most represented partitions. The generated tree is known as a consensus tree (Felsenstein, 1988). There are several types of consensus tree construction methods, among them the "strict" consensus, "majority rule" consensus, and "majority rule extended" consensus (Felsenstein, 2004). Strict consensus creates a tree consisting only of partitions that were represented in all of the trees (Felsenstein, 2004). Majority rule consensus creates a tree consisting of partitions that occurred more than 50% of the time, but leaves all other partitions unresolved (Felsenstein, 2004). Lastly, majority rule extended creates a tree consisting of partitions that occurred more than 50% of the time, but then it resolves the rest of the tree by using the most represented partitions (Felsenstein, 2004). It is possible to calculate distances between trees, though there are many methods of doing so which are not verified in terms of accuracy. Furthermore, of those that have been verified, many are situational. The "symmetric distance" is a metric useful for determining distances between trees pertaining to topology, without considering branch lengths (Felsenstein, 2004). It is also a quite simple algorithm. If given two trees, you can simply count the number of partitions which do not exist in the other tree. This metric is useful because there is a maximum distance between two trees. Between two trees containing n taxa, the maximum distance is 2n-6. Therefore, these symmetric distance values are subject to normalization by dividing all values by 2n-6. Tree with a normalized symmetric distance of 1 are trees that share no partitions, and trees with a normalized symmetric distance of 0 are identical.

Researchers regularly attempt to create new methods or improve old ones, but little is known about what factors may determine which method is the best. In order to determine which factors favor which method, a study using a simulation would be intriguing, because a large amount of data could be generated very quickly, and the actual phylogenetic trees would be known. Thus, comparisons between the actual trees and the estimated trees could be made. The purpose of our experiment is to determine the accuracy of various distance based tree construction methods with and without bootstrapping. As we are most interested in tree topology and would like the ability to normalize tree distance values to allow comparison of results between different generations, symmetric distance is our distance of choice. This experiment requires a system from which a large amount of meaningful data can be efficiently acquired, and most importantly, for which the actual phylogenetic tree is known. Further, the conclusion of an experiment conducted by Hang et al (Hang et al, 2007) and Hagstrom et al (Hagstrom, et al, 2004) is that computer simulations often underestimate the accuracy of phylogenetic methods due to the non-existence of natural selection. Therefore, a system in which natural selection exists would be most valuable. For this experiment, our system of choice is EcoSim because like Avida, it exhibits natural selection, efficiently produces meaningful data, and tracks phylogenetic records.

The Ecosystem Simulation, EcoSim

EcoSim is an individual-based predator-prey ecosystem simulation in which agents can evolve (Gras et al, 2009). The agents have a behavior model which allows the evolutionary process to modify the behaviors of the predators and prey. Furthermore, there is a speciation mechanism which allows researchers to study global patterns as well as species-specific patterns. To our knowledge, EcoSim is the only simulation in which agent behaviors affect evolution and speciation. In EcoSim, an individual's genomic data codes for its behavioral model and is represented by a fuzzy cognitive map (FCM) (Kosko, 1986). The FCM contains sensory concepts such as foodClose or predatorClose, internal states such as fear or hunger, and motor concepts such as escape or reproduce. The FCM is represented as a 390-element array consisting of positive or negative floating-point values which represent the extent to which one concept influences another. For example, it would be expected that the sensory concept predatorClose would positively affect the internal concept fear, which would then positively affect the escape motor concept. Likewise, sensing that a predator is close should negatively affect hunger, which should result in a prey agent choosing not to eat when a predator is too close. Of course, these relationships among concepts evolve over time, sometimes giving a new meaning to a concept. This representation of the FCM allows for reasonable computational complexity while still allowing for a complex system with meaningful genomic information. Furthermore, the FCM is heritable, meaning that a new agent

Downloaded from http://direct.mit.edu/isal/proceedings-pdf/alife2012/24/105/1901132/978-0-262-31050-5-ch015.pdf by guest on 19 April 2024

is given an FCM which is a combination of that of its parents with possible mutations. The FCM is largely responsible for the evolution, speciation, and behavior model which makes EcoSim so unique. EcoSim subscribes to the "genotypic cluster" definition of a species, which states that "species are clusters of genotypes circumscribed by gaps in the range of possible multilocus genotypes between them" (Mallet, 1995). What this means, in EcoSim, is that if the difference between FCMs of the two most dissimilar conspecific individuals is greater than a set threshold, the species will then split and the new species will be reproductively isolated from the parent species (Aspinall and Gras, 2010). Each species of EcoSim is assigned a species ID, which is simply a count of how many species have existed in that run (starting at species 1). Thus, species 1 is the common ancestor of all other species in a run. All trees produced in this experiment (both actual and estimates) refer to species by their species ID. Since EcoSim has the capacity to allow speciation events to occur, it is possible to track speciation events throughout a run of the simulation and construct the actual phylogenetic tree. This is important because it offers us the opportunity to perform various tree reconstruction methods and compare the results with the actual tree, which is generally not possible with real data from biological systems. Since EcoSim uses an array of 390 floating-point values to represent an agent's genome, we can obtain the average FCM of any species at any time step in any particular simulation run. From this data, we are able to construct a pairwise distance matrix of all species alive any particular time step. Thus, we are able to perform and test distance-based phylogenetic tree construction methods on data generated by EcoSim. There have been several other studies conducted using EcoSim. EcoSim has been shown to have realistic species abundance patterns (Devaurs et al, 2010) and chaotic behavior with multi-fractal properties which has been observed in biological systems (Golestani and Gras, 2010). Another study observed disease diffusion patterns and disease control regimes in EcoSim (Farahani et al).

Data Preparation and Phylogenetic Methods

Five EcoSim runs of lengths 5658, 7098, 10000, 15500, and 19500 generations were carried out. The lengths of these runs are arbitrary and do not affect the results. These runs exhibited various run-specific characteristics. Respectively, the aforementioned EcoSim runs had an average global population of about 288740, 216320, 163675, 128530, and 149177 agents, and an average species count of 28.4, 16.3, 36.2, 30, and 29.4 species over the generations which we analyzed. Their average normalized symmetric distances (considering all phylogenetic construction methods) were 0.46, 0.48, 0.66, 0.59, and 0.54, respectively. The species population sizes ranged from 1 to 73242 over all of the runs. On average, there were 29.52 taxa per generation, ranging from 7 taxa to 47 taxa. Thus, the largest distance matrix from which a tree was constructed was 47x47. In this case, to calculate a single tree using the UPGMA or Neighbor-Joining method required less than one second, whereas when using the Fitch-Margoliash method it required nearly ten seconds. Even if the system has to handle hundreds of thousands of "intelligent" agents simultaneously, the overall complexity of the algorithm is linear and therefore it allows us to compute a very high number of time steps giving us the possibility to observe evolutionary phenomena. For reference, a run of 25000 generations of EcoSim takes approximately 40 days, but this depends on the number of predator and prey individuals produced.

A program was created to automatically generate phylogenetic trees in NEWICK format (Felsenstein, 2004) by extracting data regarding species splitting events from the simulation. The branch lengths of the trees generated by this program were exactly the number of generations passing between speciation events. Another program was implemented to edit the full phylogenetic trees, removing all species that did not exist at a given generation. The purpose of this was to generate actual trees that were comparable with results from the distance-based tree construction methods. Another program was then created to extract species-specific average FCMs at a given generation, and with that data construct distance matrices. This program used pairwise Euclidean distance between average FCMs to generate distance matrices. When analyzing biological systems, one would first have to convert the data (genetic or molecular sequences, enzyme binding data, or immunological data for example) into distance matrices. In the case of molecular or genetic sequences, one would first have to align the sequences and then calculate the genetic or molecular distance between them. Once this is completed, the distance-based phylogenetic tree construction methods can be applied.

Once these distance matrices were generated, the program "Neighbor" of PHYLIP (the PHYLogeny Inference Package) (Felsenstein, 1989) was used to perform Neighbor-Joining and UPGMA methods on the distance matrices. To perform the Fitch-Margoliash method, "Fitch" of PHYLIP was used. For a run of 10000 generations (for which 19000 trees are generated when performing phylogenetic bootstrapping), to compute all of the bootstrap Neighbor-Joining and UPGMA trees it only took about two hours, whereas to compute the bootstrap Fitch-Margoliash trees it took roughly ten hours. The trees generated from these algorithms were compared with the actual trees using symmetric distance. This was done using "TreeDist" of PHYLIP. In order to perform bootstrap analysis, another program was created to resample the FCM and generate distance matrices from these resampled FCMs. This was performed by choosing a replacement probability and then possibly replacing an FCM element with another for all species before calculating distances between species. The assigned replacement probability was 0.5, and 1000 such resamplings were performed. Then, "Consense" of PHYLIP was used to perform majority rule extended consensus. Majority rule extended was used as the consensus method because it generates fully resolved binary trees to allow for comparison with the actual phylogenetic trees. The consensus trees were then compared with the actual trees (again, using "TreeDist"). Tree construction (both the actual trees and distance-based estimates), consensus, and comparisons were

Downloaded from http://direct.mit.edu/isal/proceedings-pdf/alife2012/24/105/1901132/978-0-262-31050-5-ch015.pdf by guest on 19 April 2024

performed every 500 generations until the end of an EcoSim run, starting from a point in the run at which there were enough species in existence for it to be reasonable to test. This resulted in 100 analyzed time-steps, with 3006 tree estimates constructed per time-step.

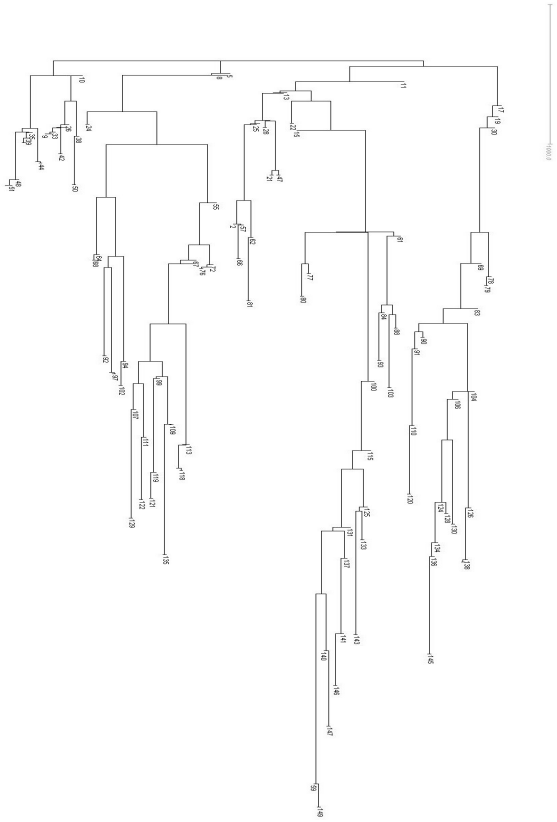


Figure 1: The actual phylogenetic tree for EcoSim run #2 of 5658 generations. Over the 5658 generations, 149 taxa were generated. The leaves of the tree represent the species indicated at the time of the last splitting event in which they were involved. The internal nodes of the tree are the species with the lowest species ID in the partition to the right of that node (since species are given an ID in the order in which they are generated), and represent that particular species at the time of that splitting event.

Results

Actual phylogenetic trees consisting of all species in a run were constructed for all five EcoSim runs, an example of which is shown in Figure 1. Edited trees, consisting of only species existing in a particular generation, were also created. Neighbor-Joining, UPGMA, and Fitch-Margoliash methods were used, and consensus trees using these methods were generated as well. Examples of each tree are shown in Figure 2. The UPGMA method is the only method of the three which creates a binary rooted tree when not performing consensus analysis. While not performing consensus analysis, Neighbor-Joining and Fitch-Margoliash methods create unrooted trees. All consensus trees are binary rooted trees. The trees produced by performing consensus analysis have branch lengths which are meaningless in terms of evolutionary distance between species. The branches of the consensus trees are actually the bootstrap value; this is number of trees in which the partition to the right of that branch was represented out of the 1000 resamplings performed. Thus, the longer the branch, the more represented that partition was. The tree distance metric we used, as previously mentioned, only deals with topology, so the branch lengths (in terms of comparison) are not necessarily important.

Symmetric distances between the edited actual trees and the estimated trees were calculated (Table 1). Ranked from most effective to least effective, the phylogenetic tree construction methods are as follows: 1) Fitch-Margoliash Consensus, 2) Fitch-Margoliash, 3) Neighbor-Joining Consensus, 4) UPGMA Consensus, 5) UPGMA, and 6) Neighbor-Joining. Note that although it was the most accurate, the Fitch-Margoliash method only classified, on average, 46% of the partitions.

Method	Avg. SD	SD Std. Dev.	Avg. Norm. SD	Norm. SD Std. Dev.
F-M (C)	28.98	11.41	0.54	0.15
F-M	29.24	11.5	0.55	0.15
N-J (C)	29.57	11.52	0.55	0.15
UPGMA (C)	29.86	12.43	0.56	0.16
UPGMA	31.23	13.02	0.59	0.18
N-J	32.44	11.92	0.6	0.14

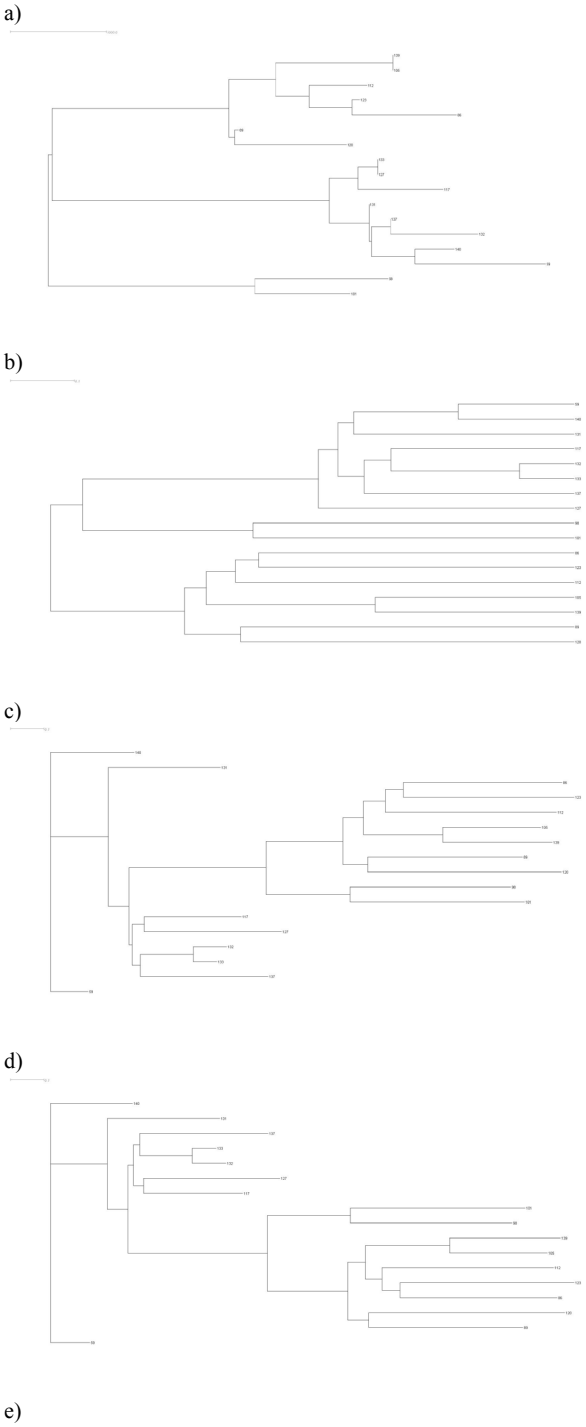
Table 1: The average and standard deviation of the symmetric distance (SD) and the normalized symmetric distance of all five EcoSim runs. The Fitch-Margoliash method generated the most accurate trees, with an average of 54% of partitions incorrectly reconstructed. The least accurate was the Neighbor-Joining method, with an average of 60% of partitions incorrectly reconstructed. The UPGMA method produced the most varying results, and the Neighbor-Joining method was the most consistent.

Conclusions

In our experiments based on data generated by our evolving ecosystem simulation, none of the distance-based methods performed well. None of the methods, on average, estimated over 50% of the partitions of the trees correctly. Though it is possible that these methods are just not as accurate as previously perceived, there could be several reasons why they performed poorly. It is possible that there are factors (such as mutation rates, small population sizes for some species, rate of evolution, probability of back-mutation, or large number of species) that make it difficult for distance-based phylogenetic tree construction methods to properly recreate the trees. It is also possible that Euclidean distance (employed in this manner) is just a poor metric for use with distance-based phylogenetic tree construction methods. Another possibility is that the distance matrices produced were not additive (and

Downloaded from http://direct.mit.edu/isal/proceedings-pdf/alife2012/24/105/1901132/978-0-262-31050-5-ch015.pdf by guest on 19 April 2024

thus not ultrametric either), but this is often the case in nature as well (Felsenstein, 2004). Lastly, rather than using the entire FCM, it may be better to choose specific FCM values to create phylogenies from, despite research in phylogenomics that



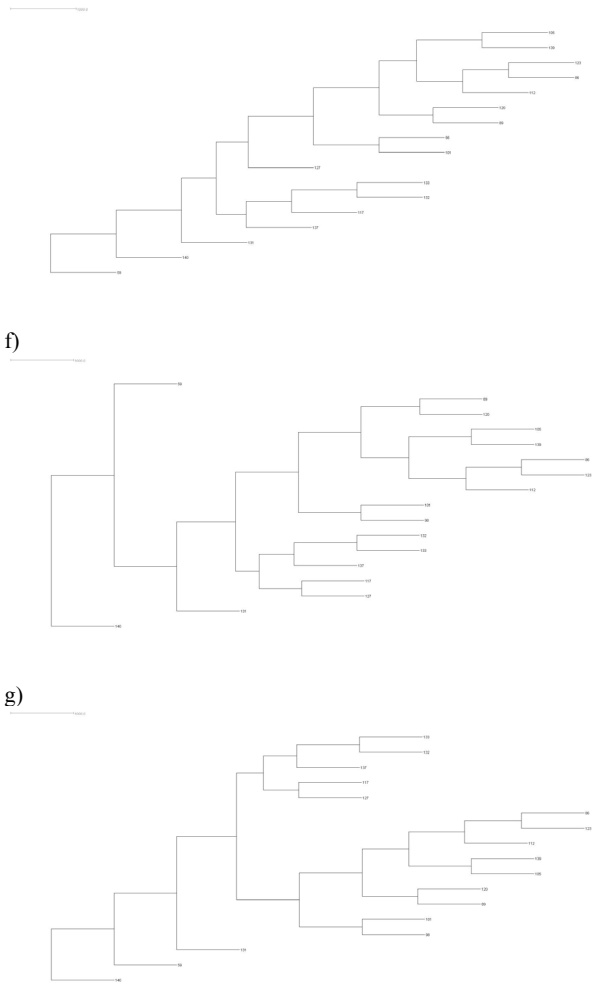


Figure 2: The actual (a) and estimated (b-g) trees for EcoSim run #2, generation #4158. Consensus trees (UPGMA (e), Neighbor-Joining (f), Fitch-Margoliash (g)) and UPGMA (b) trees are all binary rooted trees, while Neighbor-Joining (c) and Fitch-Margoliash (d) trees are unrooted. This example shows the similarities and differences between relatively small (17 taxa) trees generated by the various methods.

suggests using entire genomes (rather than a small number of genes) increases the phylogenetic signal-to-noise ratio (Phylippe et al, 2005; Snel et al, 2005). This is because our FCM may actually be noisy in terms of the phylogenetic data it generates, so determining and focusing on values with high phylogenetic signal-to-noise ratio may increase the accuracy. The Fitch-Margoliash method with consensus analysis performed slightly better than the other methods. It was expected that in all cases, performing phylogenetic bootstrapping and building consensus trees increased the accuracy of the methods.

Our results contrast from those of Hagstrom et al (Hagstrom et al, 2004), as in their experiments they have found that these methods are quite accurate (in many cases,

Downloaded from http://direct.mit.edu/isal/proceedings-pdf/alife2012/24/105/1901132/978-0-262-31050-5-ch015.pdf by guest on 19 April 2024

reproducing the exact phylogenetic tree) provided that there is an element of natural selection in the employed system. EcoSim is such a system, yet our results are quite different. One important difference between these experiments is that in our experiment, we attempted recreating phylogenies consisting of many (on average 29.52) taxa whereas in that of Hagstrom et al, phylogenies of only four taxa were reconstructed.

A study by Leitner et al (Leitner et al, 1996), in which researchers performed various phylogenetic tree construction methods on HIV-1 molecular data, also found that the Fitch-Margoliash method was most accurate, and it also found that Neighbor-Joining consensus was more accurate than UPGMA (though they considered branch lengths in their tree comparison, which may have increased the inaccuracy of UPGMA). They found that in some cases the true phylogeny was successfully reconstructed, whereas in all of our cases this did not occur. It is interesting to note, however, that they only had 9 taxa to analyze. Our best scenario was one in which we had only 7 taxa to analyze, which gave us 25% dissimilarity using Fitch-Margoliash and Neighbor-Joining , and 75% dissimilarity using UPGMA. On average, 29.52 taxa per generation were analyzed in our experiment. It is also interesting to note that choice of gene, in the case of HIV-1, accounted for an average symmetric distance difference of about 25%. This also leads us to believe that perhaps we should focus on specific FCM values (such as those that rapidly evolve or those that are most selected upon) rather than on the entire FCM. When considering the efficiency of the algorithms, the UPGMA and Neighbor-Joining methods are much more efficient than the Fitch-Margoliash method, so it may still be more appropriate to use Neighbor-Joining or UPGMA instead of Fitch-Margoliash in some cases (for example, those that require the computation of many trees).

In the future, we will attempt to determine which characteristics (for example relatedness of different species, speciation threshold, or rates of evolution) may allow each method to produce the most accurate tree. Furthermore, it would be intriguing to determine if these factors lead to better trees overall. We would also like to discover if selecting only certain FCM values produces better trees. It also may be interesting to discretize the FCM values and perform a similar analysis of the more popular character-based methods.

Acknowledgements

This work is supported by the NSERC grant ORGPIN 341854, the CRC grant 950-2-3617 and the CFI grant 203617 and is made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET, www.sharcnet.ca).

References

Aspinall, A., and Gras, R. (2010). K-means clustering as a speciation mechanism within an individual-based evolving predator-prey ecosystem simulation. *Active Media Technology*, pages 318–329, Toronto, Canada.

Devaurs, D., and Gras, R. (2010). Species abundance patterns in an ecosystem simulation studied through Fisher's logseries. *Simulation Modelling Practice and Theory*, *18*(1), 100-123.

Farahani, Y. M., Khater, M., and Gras, R. (in press). Modeling Epidemic Spread in a Predator-Prey Evolutionary Ecosystem Simulation. To appear in the *Journal of Artificial Life.*

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–791.

Felsenstein, J. (1988). Phylogenies From Molecular Sequences: Inference and Reliability. *Annual Review of Genetics*, 22:521-565.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5: 164-166.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760):279-284.

Golestani, A., and Gras, R. (2010). Regularity analysis of an individual-based ecosystem simulation. *Chaos (Woodbury, N.Y.)*, *20*(4), 043120.

Grandcolas, P., Deleporte, P., Desutter-Grandcolas, L., and Daugeron, C. (2001). Phylogenetics and Ecology: As Many Characters as Possible Should Be Included in the Cladistic Analysis. *Cladistics*, 17:104-110.

Gras, R., Devaurs, D., Wozniak, A., and Aspinall, A. (2009). An individual-based evolving predator-prey ecosystem simulation using a fuzzy cognitive map as the behavior model. *Artificial Life*, *15*(4), 423-63.

Hagstrom, G. I., Hang, D. H., Ofria, C., and Torng, E. (2004). Using Avida to Test the Effects of Natural Selection on Phylogenetic Reconstruction Methods. *Artificial Life 10,* pages 157-166. MIT Press, Cambridge, MA.

Hang, D., Torng, E., Ofria, C., and Schmidt, T. M. (2007). The effect of natural selection on the performance of maximum parsimony. *BMC Evolutionary Biology,* 7:94.

Kosko, B. (1986). Fuzzy cognitive maps. *International journal of man-machine studies*, *24*(1), 65-75.

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., and Albert, J. (1996). *Proceedings of the National Academy of Sciences of the United States of America*, 93:10864-10869.

Lespinats, S., Grando, D., Marechal, E., Hakimi, M., Tenaillon, O., and Bastien, O. (2011). How Fitch-Margoliash Algorithm can Benefit from Multi Dimensional Scaling. *Evolutionary Bioinformatics*, 7:61-85.

Mailund, T., Brodal, G. S., Fagerberg, R., Pedersen, C. N. S, and Phillips, D. (2006). Recrafting the neighbor-joining method. *BMC Bioinformatics*, 7:29.

Mallet, J. (1995). A species definition for the modern synthesis. *Trends in Ecology & Evolution*, 10:294–299.

Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: state of the art. *Computational Statistic Quarterly*, 1(2):101-113.

Phylippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, 36:541-562

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406-425.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*. Freeman, San Fransisco, CA.

Snel, B., Huynen, M., Dutilh, B. E. (2005). Genome Trees and the Nature of Genome Evolution. *Annual Review of Microbiology*, 59:191-209.

Wiley, E. O. and Lieberman, B. S. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. John Wiley & Sons, Hoboken, NJ.

Downloaded from http://direct.mit.edu/isal/proceedings-pdf/alife2012/24/105/1901132/978-0-262-31050-5-ch015.pdf by guest on 19 April 2024