The History of Life Through Evolutionary Computation

Gary B. Fogel

Prior to our understanding of DNA and molecular sequence information, paleontologists inferred the history of life on Earth by comparing morphological characteristics (preserved phenotypic traits) found in fossils. Accurate estimation of this history of life or "phylogeny" was generally only possible for those organisms that had "hard parts" capable of preservation, died in a "fossilization-friendly" context, and were unearthed over time only to be discovered by a very fortunate paleontologist. According to Newton and Laporte, paleontologists are the first to recognize the incompleteness of the fossil record and the resulting difficulty of formulating evolutionary relationships between extant organisms. It is a shame that many of Earth's organisms have very likely left no trace of their existence.

Our ability to infer evolutionary histories changed significantly following the discovery of the genetic code. Fitch and Margoliash were the first to offer a computer process for the construction of phylogenetic trees from protein sequence information, specifically for 20 cytochrome C protein sequences (the most for any protein to that date) that had been elucidated from humans to yeast. This publication provided a tremendous opportunity for the use of biological sequence information as "molecular fossils" of information that could be compared between extant organisms to determine their evolution. Seemingly all that was required was additional sequence information and better computers and algorithms for their interpretation. Assessing the reliability of phylogenies involves difficult statistical and computational problems, including the NP-complete problems of sequence alignment and discovering the best phylogenetic tree that fits the data.

Given modern databases filled with sequence information, interest has turned from one of generating sequence to rapid interpretation and discovery of "true" phylogenies for their application in not only the resolution of the history of life but also for epidemiology as it relates to human disease. Three developments have been essential in this progression: 1) the development of criteria and algorithms for discriminating among potential phylogenies, 2) increased computational power over time, and 3) the rapid increase in sequence data availability. An assortment of algorithms has been offered to solve the phylogenetic reconstruction problem, some using evolutionary algorithms.

Challenges in phylogenetics

The phylogenetics problem can be loosely defined as the search for a treelike structure that defines ancestral relationships between related objects over time. The related objects can be anything from biological sequence information to morphological characters. The divergence over evolutionary time represented is captured in a treelike structure termed a "phylogeny" (Fig. 1). The number of unrooted, birfurcating tree topologies T for n taxa is given by

$$T = \frac{(2n-5)!}{\left((n-3)!2^{n-3}\right)}. \qquad (1)$$

In general, three possible methods are used to search for the best topology from this set of possible trees: 1) exhaustive methods, 2) branch and bound methods, and 3) heuristic methods. In analogy to the traveling salesman problem (TSP), exhaustive methods are useful when the number of

0278-6648/05/$20.00 © 2005 IEEE

possible tree topologies (TSP routes) is low. This is possible when the number of n taxa being compared is low (on the order of $\leq$12) but rapidly becomes infeasible with larger numbers of taxa. Branch and bound methods exclude trees that do not meet specific criteria, reducing the search space to a more reasonable size and increasing the probability of an exact solution of merit. However, there are still limitations on the upper number of taxa that can be used with this approach.

Heuristic searches commonly are used to build tree topologies either by changing the order in which the trees are built or via branch swapping with some metric of scoring (such as a parsimony) being used to determine which changes are more useful than others. It is easy to envision how evolutionary computation can be applied in this regard as a method of global optimization where the best resulting tree topology is estimated after searching only a small fraction of the search space. An unfortunate consequence of this approximation is that the resulting tree topology cannot be guaranteed to be optimal; however, it does allow the researcher to efficiently search large numbers of character states and infer "reasonable" historical relationships between organisms. Additional confidence in a proposed phylogeny results from overlap of "best" trees generated from different sequence or character sets.

It should be noted that one key discovery of 20th century evolutionary biology was the determination of a "tree of life" based purely on molecular sequences for a set of genes (ribosomal RNA genes) that are common to essentially all forms of extant life. According to Pace, the tree of life has reshaped our thinking of the evolution of life on Earth and helped us identify the three major kingdoms of life (eukarya, bacteria, and archaea). Phylogenetic methods are now integral parts of almost every major area of evolutionary biology and several parts of ecology. Those interested in learning more about recent discoveries in this area are referred to *Assembling the Tree of Life* and the Tree of Life web project <http://tolweb.org/tree/phylogeny.html>.

Phylogenetic methods

Correct alignment of the sequence information is generally a critical first component of phylogenetic analysis and can be difficult when there are large numbers of sequences to be compared, when the sequences are long, and when the sequences have limited evolutionary conservation.

Parsimony is a common method used to infer phylogenetic trees. This method is based on the hypothesis that the "best" tree in the space of all possible trees is the one that explains the relationship between the extant taxa with the fewest number of evolutionary changes throughout the tree. The principle of inheritance implies that when a characteristic state is modified in a species, the descendents of that species will have a high probability of sharing that characteristic. Thus the most parsimonious tree is the one that groups taxa together with similar characteristics and minimizes the number of observed overall changes in the tree. There are several statistical inconsistencies with this approach, most notably long-branch attraction and unequal rates of evolution.

Distance methods are a second common method to infer phylogeny. These methods use pairwise distance calculations for characters (e.g., a pairwise calculation of mismatches measured across all positions of a nucleotide sequence alignment). A model of sequence evolution is then applied to the distance matrix to correct for unobserved changes and chance similarities. If the model of sequence evolution is accurate, then the correct tree can be recovered. These models may be specific to the characters that are being investigated. It is clear that models of sequence evolution do not work equally well over all character sets. Thus, this approach works maximally as well as the model of sequence evolution.

Maximum likelihood methods represent a third common method of phylogenetic inference. This approach is generally similar to maximum parsimony. However, maximum likelihood methods use a specified model of sequence evolution to assign a value of confidence to ancestral states. Maximum parsimony methods assume that a shared character between two extant taxa must have also existed in the ancestor of those two taxa. Maximum likelihood methods assign a confidence to the existence of that ancestral character based on the distance (or length of time) that exists between the two extant taxa. As a result, maximum likelihood methods tend to be more computationally intensive than parsimony methods. They also share the same requirements as distance methods for a model of sequence evolution specific to the character set under investigation. However, maximum likelihood approaches can improve tree inference when the sequences are not closely related. All three of the above approaches are the subject of considerable investigation in the literature and are the subject of optimization with simulated evolution.

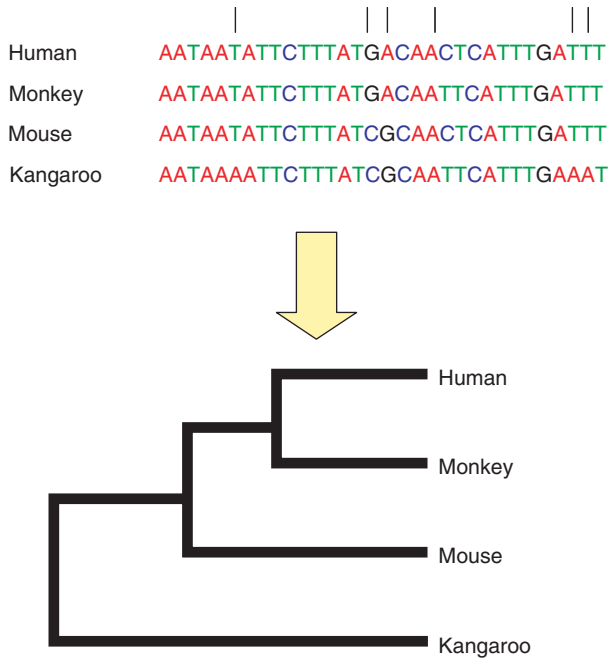

Fig. 1 Phylogenetic reconstruction. The nucleotide sequences for four organisms are aligned and in this simple example are highly similar except for six positions noted as ("|"). At each of these informative positions (characters), change has occurred over evolutionary time. All other nucleotide positions are invariant and not required for proper phylogenetic reconstruction. Combining key characters and resolving the true phylogeny is challenging when the number and type of characters/taxa increases.

Applications of evolutionary computation to phylogenetics

Evolutionary computation has been applied to phylogenetic reconstruction for 10 years. Matsuda was the first to use evolutionary algorithms for phylogenetic reconstruction, doing so with protein sequences. Since that time,

these strategies have been extended for use with different character sets, including Lewis who used evolutionary algorithms for phylogenetic reconstruction with nucleotide sequence information. For this purpose, individuals in the population were representations of phylogenetic tree topologies along with all branch lengths and values of other parameters regarding the model of sequence evolution used. Mutation and recombination operators were defined to generate new offspring solutions and fitness was scored with respect to the natural log likelihood (lnL) score. Higher lnL scores were favored over time (using ranked selection) until convergence on a solution was observed. The best resulting individual from this evolutionary search was believed to represent the best phylogenetic topology. This approach was applied to a phylogenetic reconstruction problem with 55 taxa involving sequences of the chloroplast *rbcL* gene in plants. In each of three runs, the evolutionary algorithm converged on a different tree topology (most likely as a result of premature convergence to local optima and/or insufficient parameter tuning). A comparison of the approach to a more standard heuristic method (PAUP v.4.0) on the same data set using the same computer and model of sequence evolution was made. The best resulting tree from PAUP was the same as one of the trees discovered with the evolutionary algorithm but required 783 hours to do this calculation. In comparison, the evolutionary approach only required 42.4 hours. Similar results and comparisons to standard methods were made by Reijmers et al. who coupled an evolutionary algorithm to a neighbor-joining approach to reduce the number of trees in the search space with successful evaluation relative to the FITCH algorithm.

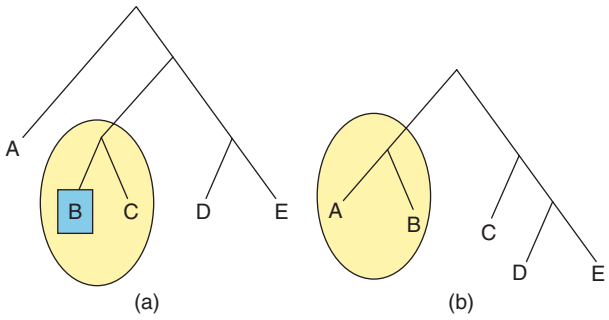

Fig. 2 Example of two parent phylogenies with five species each. (a) A subtree with species B is identified in the first parent solution. (b) A subtree that contains species B. The two circled branches would then be the subject of crossover (adapted from Congdon, 2002).

Congdon and Greenfest developed a program called "Gaphyl" for phylogenetic reconstruction with evolutionary algorithms. This was first used with binary character states using parsimony as a guide for optimal tree topology discovery. Gaphyl makes use of several intuitive variation operators for successful search. First, a crossover strategy is used as follows:

1. A species is selected at random from a parent phylogeny.

2. A subtree is selected at random that includes the species from 1 above (excluding the subtree that represents the entire tree itself and the subtree that represents only the direct lineage to the species in 1 above).

3. In a second parent phylogeny, the smallest subtree containing all the species from the subtree obtained in 2 above is found.

4. Two offspring trees are formed by swapping these two subtrees.

5. Any duplicate species found in the resulting offspring solutions are pruned.

Second, a mutation operator is used:

1. Select two species at random from within one parent and swap their positions in the tree.

2. Select a subtree at random within a parent phylogeny. Rearrange the evolutionary history of the species within that subtree.

The approach was extended in Congdon and Septor for use with DNA sequence information, and the early indications are that the Gaphyl program can outperform a common program called Phylip on datasets with l63 species with 1,588 nucleotides each in terms of CPU time, but both tools are able to find four equally parsimonies phylogenies for this data. A common theme in this work was that as the problem sets became more complex, Gaphyl was able to find more complete sets of phylogenies than Phylip, even when making use of the fitness function from Phylip. Gaphyl commonly demonstrated a wider variety of resulting "best" phylogenies than Phylip, suggesting that Gaphyl could search the space of possible solutions more effectively. Congdon presents a nice survey of the importance of variation operators in this problem area and suggested that as the number of species and attributes increases, the effective of Gaphyl over Phylip appeared to increase. Shen and Heckendorn have continued experimentation in this area.

Computational Intelligence in Bioinformatics

Advances in computational intelligence provide us with the opportunity to explore ways in which these methods can be applied to problems in bioinformatics, such as the phylogenetic reconstruction problem. However, this interdisciplinary research requires understanding the biological data, methods for optimal search, and where and how these can be applied. To bring these researchers from computer science and biology closer together, a number of workshops and special sessions have been organized at IEEE Computational Intelligence Society sponsored events and activities including:

- the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (IEEE CIBCB), sponsored by the IEEE Computational Intelligence Society <http://www.cibcb.org>
- a variety of special sessions on bioinformatics applications at IEEE sponsored conferences such as the Congress on Evolutionary Computation (CEC) <http://www.wcci2006.org>, International Joint Conference on Neural Networks (IJCNN) www.ijcnn.org, and FUZZ-IEEE <http://www.fuzzieee2005.org>
- *IEEE/ACM Transactions on Computational Biology and Bioinformatics* is a relatively new IEEE publication that serves this community.

In particular the IEEE CIS Bioinformatics and Bioengineering Technical Committee (IEEE BBTC) <http://ieee-cis.org/page/?sec=5&sub=6> promotes the research, development, education, and understanding of computational intelligence methods in computational biology.

Katoh et al., Lemmon and Milinkovitch, and Brauer et al. have all recently published in mainstream biology journals, using evolutionary algorithms for phylogenetic reconstruction. In particular, Brauer et al. focused on a parallelization of the evolutionary search, where a master process created a population of individuals and sent each of them to another single processor to be scored. Thus the population size of the evolutionary algorithm was equal to the number of slave nodes plus the master node. This method combined mutation of the branch lengths and tree topology, recombination, migration, and selection using a maximum-likelihood approach. Search-time improvement was roughly linear with respect to the number of processors that were used. Also of note was the use of both real biological data (288 plant taxa each with 4,822 nucleotides of sequence information) in addition to simulated data (a Monte Carlo simulation of 5,000 nucleotide characters across a model tree of the same 228 taxa, derived from the most parsimonious previously known answer for this data set. Use of simulated data provided a means of interrogating all of the "correct" ancestral states that are commonly missing in real data and proved to be very valuable in assaying the overall worth of the approach.

Future applications

Sequence data availability is no longer an issue for phylogenetics studies. Rather than simply compare genes from different organisms, the problem has transformed into complete genome comparison. Computational power continues to increase over time but the problem sizes of interest continue to expand as well. Development of better algorithms and methods for discriminating among potential phylogenies will continue to play a central role in this area.

Testing algorithm performance requires knowledge of a "ground truth" set of data from a known evolutionary history. This is difficult to obtain from the fossil record but can be generated artificially in the computer, in the laboratory with bacterial lineages or nucleic acids, or from particularly fast-evolving forms such as viruses. From each of these processes, ancestral states can be identified and stored as the evolutionary divergence unfolds. These data sets make particularly interesting examples for the evaluation of phylogenetic methods and are likely to be used more often in that manner in the future.

Pond and Frost utilized a genetic algorithm approach to assign lineages in a phylogeny to different rates of nonsynonymous and synonymous substitution (mutation rate) at the protein level. Very commonly, researchers assume a particular substitution/mutation rate exists over an entire phylogeny, when it is rather clear that the rate of mutation may vary significantly by lineage. The Pond and Frost study was the first to search for models of substitution in a phylogenetic context.

Read more about it

• 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.* vol. 19, no. 10, pp. 1717–1726, 2002.

• C.B. Congdon, "Gaphyl: A genetic algorithms approach to cladistics" in *Principles of Data Mining and Knowledge Discovery* (Lecture Notes in Computer Science, vol. 2168), L. DeRaedt and A. Siebes, Eds. Berlin: Springer-Verlag, 2002, pp 67–78.

• C.B. Congdon and E.F. Greenfest, "Gaphyl: A genetic algorithm approach to cladistics," in *Data Mining with Evolutionary Algorithms*, A.A. Freitas, W. Hart, N. Krasnogor, and J. Smith, Eds. 2000, pp. 85–88.

• C.B. Congdon, "Gaphyl: An evolutionary algorithms approach for the study of natural evolution," in *Proc. Genetic and Evolutionary Computation Conference (GECCO 2002)*, San Francisco, 2002, pp. 1057–1064.

• C.B. Congdon and K.J. Septor, "Phylogenetic trees using evolutionary search: Initial progress in extending gaphyl to work with genetic data," in *Proc. 2003 Congress on Evolutionary Computation (CEC 2003)*, 2003, pp. 320–326.

• W.M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, vol. 155, pp. 279–284, 1967.

• K. Katoh, K. Kuma, and T. Miyata, "Genetic algorithm-based maximum-likelihood analysis for molecular phylogeny," *J. Mol. Evol.*, vol. 53, no. 4–5, pp. 477–484, 2001.

• A.R. Lemmon and M.C. Milinkovitch, "The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation," *Proc. Nat. Acad. Sci. USA,* vol. 99, no. 16, pp. 10516–10521, 2002.

• P.O. Lewis, "A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data," *Mol. Biol. Evol.*, vol. 15, no. 3, pp. 277–283, 1998.

• H. Matsuda, "Construction of phylogenetic trees from amino acid sequences using a genetic algorithm," in *Proc. Genome Informatics Workshop,* 1995, no. 6, pp.19–28.

• H. Matsuda, "Protein phylogenetic inference using maximum likelihood with a genetic algorithm," in *Proc. Pac. Symp. Biocomput.*, 1996, pp. 512–523.

• C.R. Newton and L.F. Laporte, *Ancient Environments*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1989.

• S.L. Pond and S.D. Frost, "A genetic algorithm approach to detecting lineage-specific variation in selection pressure," *Mol. Biol. Evol.*, vol. 22, pp. 478–485, 2005.

• N.R. Pace, "A molecular view of microbial diversity and the biosphere," *Science*, vol. 276, pp. 734–740, 1997.

• T.H. Reijmers, R.L. Wehrens, and M.C. Buydens, "Quality criteria of genetic algorithms for construction of phylogenetic trees," *J. Comput. Chem.*, vol. 20, pp. 867–876, 1999.

• T.H. Reijmers, R. Wehrens, F.D. Daeyaert, P.J. Lewi, and L.M.C. Buydens, "Using genetic algorithms for the construction of phylogenetic trees: Application to G-protein coupled receptor sequences," *BioSystems*, vol. 49, pp. 31–43, 1999.

• J. Shen and R.B. Heckendorn, "Discrete branch length representation for genetic algorithms in phylogenetic search," in *Lecture Notes in Computer Science "Applications of Evolutionary Computing: EvoWorkshops 2004: EvoBIO, EvoCOMNET, EvoHOT, EvoISAP, EvoMUSART, and EvoSTOC, Coimbra, Portugal, April 5-7, 2004*," G.R. Raidl, S. Cagnoni, and J. Branke, Eds. Berlin: Springer, 2004, pp. 94–103.

About the author

Gary B. Fogel received his B.A. in biology from the University of California, Santa Cruz and a Ph.D. in biology from the University of California, Los Angeles (UCLA). While at UCLA, he was a fellow of the Center for the Study of Evolution and the Origin of Life. As vice president at Natural Selection, Inc., his current research interests focus on the application of computational intelligence methods to problems in the biomedical sciences. He is chair of the IEEE CIS Bioinformatics and Bioengineering Technical Committee and general chair for the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.