Molecular Systematics and Evolution: Theory and Practice
ed. by R. DeSalle, G. Giribet and W. Wheeler

Theory and practice of parallel direct optimization

Daniel A. Janies and Ward C. Wheeler

Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA

Summary. Our ability to collect and distribute genomic and other biological data is growing at a staggering rate (Pagel, 1999). However, the synthesis of these data into knowledge of evolution is incomplete. Phylogenetic systematics provides a unifying intellectual approach to understanding evolution but presents formidable computational challenges. A fundamental goal of systematics, the generation of evolutionary trees, is typically approached as two distinct NP-complete problems: multiple sequence alignment and phylogenetic tree search. The number of cells in a multiple alignment matrix are exponentially related to sequence length. In addition, the number of evolutionary trees expands combinatorially with respect to the number of organisms or sequences to be examined. Biologically interesting datasets are currently comprised of hundreds of taxa and thousands of nucleotides and morphological characters. This standard will continue to grow with the advent of highly automated sequencing and development of character databases. Three areas of innovation are changing how evolutionary computation can be addressed: (1) novel concepts for determination of sequence homology, (2) heuristics and shortcuts in tree-search algorithms, and (3) parallel computing. In this paper and the online software documentation we describe the basic usage of parallel direct optimization as implemented in the software POY (ftp://ftp.amnh.org/pub/molecular/poy).

Introduction

The first step in phylogenetic analysis is to establish putative homology statements for characters observed among study species. When considering morphology, putative homology statements result from comparative analysis by a trained specialist. However, the establishment of homologies across many sequence positions and species are not easily or optimally conducted by eye (see Giribet et al., this volume). In the analysis of molecular sequence data, multiple alignment algorithms can assign provisional homologies among residues (e.g., nucleotides, amino acids). Putative statements of morphological and molecular homology are tested by phylogenetic analysis. Cladograms are constructed from those putative homologues that are shown to be shared derived features.

The problem

The number of cells in a multiple alignment matrix are exponentially related to the number of taxa and sequence length. An alignment of m sequences of length N nucleotide bases will require N^m elements of storage (Needleman and Wunsch, 1970). One commonly used heuristic approach is to provide an ini-

tial topology of relationships among the taxa (guide tree) for accreting sequences into a matrix of provisional homologies (Sankoff et al., 1973). In theory, an alignment procedure could be repeated for each possible set of relationships among taxa. However, the number of topologies is combinatorially dependent on the number of taxa. Multiple alignment of more than a few short sequences requires heuristics.

Furthermore, the results of heuristic multiple alignment are dependent on the order in which the sequences are accreted and the functions chosen for the relative costs of insertion-deletion and substitution events in sequences. For evolutionary studies, the objective of performing a multiple alignment is often to proceed to a phylogenetic analysis with a set of putative homologies unbiased by initial assumptions. Clearly, computationally efficient and assumption-minimizing alternatives to the existing paradigm of multiple alignment are essential for evolutionary and molecular biology.

A solution

Direct optimization is a novel method of comparing putatively homologous sequence residues during cladogram diagnosis, thus obviating multiple alignment (Wheeler, 1996). Alignment algorithms create correspondences between sequence strings of various lengths by inserting gaps. In multiple alignment the relative costs of insertion-deletion and substitution events determine the number and position of gap characters inserted in sequences. Direct optimization works by creating parsimonious hypothetical ancestral sequences at internal cladogram nodes. The key difference between direct optimization and multiple alignment is that evolutionary differences in sequence length are accommodated not by the use of gap characters but rather by allowing insertion-deletion events between ancestral and descendant sequences. Evolutionary base substitution and insertion–deletion events between ancestor and descendant sequences are treated with the same cost functions (e.g., Sankoff matrices) as in multiple alignment.

Theory

Determination of DNA sequence homology

The phylogenetic analysis of DNA sequences, like that of all other comparative data, is based on schemes of putative homology that are then tested via congruence to determine synapomorphy and cladistic relationships. Unlike some other data types, however, putative molecular homologies or characters are not directly observable. DNA sequences from various organisms are often unequal in length. Hence, the correspondences among sequence positions are not evident and some sort of procedure is required to determine which regions

are homologous. This procedure is typically multiple sequence alignment. Alignment inserts gaps to make the corresponding (putatively homologous) nucleotides line up into columns. These columns (characters) comprise the data used to reconstruct cladograms. Many investigators try to hand-align raw data or hand-edit algorithmic alignments to reduce "errors" and ambiguity, but this is certainly a subjective and unrepeatable process. Whether alignment is accomplished manually or algorithmically, the resultant characters are then submitted to phylogenetic analysis as column vectors in the same manner as other forms of data, such as morphological characters scored by an investigator. Whatever the analytical pathway, alignment is an artificial manipulation of DNA observations via the insertion of gap characters that are not data but rather just place-holders. The primary reason in phylogenetics to create an alignment has only an operational basis—to make it possible to submit these data to standard phylogeny programs that were designed to handle column vectors of morphological characters. This is not a reason to believe that construction of an alignment followed by a separate tree search procedure is the only or the best way to do phylogenetics.

Limitations of multiple alignment

Alignment-based homology schemes rest on a notion of base-to-base correspondence in which individual nucleotide bases transform among five states (A, C, G, T or U, and gap) within a single character. The use of a base-to-base framework to view DNA homology is in large part responsible for the the phenomenon of long-branch attraction because of the paucity of character states (A, C, G, T/U, or –) in a column. A method available in POY, fixed-state optimization, can be used to avoid this pitfall because the method views the whole sequence, not individual bases, as characters (Wheeler, 1998, 1999b). In a fixed-states approach the number of possible character states are related to the length (n) of the sequences (up to 4^n) thus reducing the chance of random non-historical similarity to a negligible probability.

Static versus *dynamic homology*

In standard phylogenetic analysis, once an alignment is created it is not revised during or as a result of subsequent phylogenetic analysis. In this sense the putative homologies defined in the alignment are static. Reexamination is often done by hand but users will likely fall prey to biases and rearrange bases in favor of preferred groups. However, as implemented in MALIGN (Wheeler and Gladstein, 2000), randomization of an alignment's guide tree can achieve reexamination of putative homology and the alignments can be judged by an optimality criterion applied to the trees produced from the alignment via phylogenetic analysis.

As pointed out by Phillips et al. (2000), Mindell (1991) advocated using "known" phylogenies to guide alignments but the required phylogenetic information is often unavailable. In most evolutionary studies, the object of performing a multiple alignment is to allow phylogenetic analysis with a set of putative homologies unbiased by initial assumptions of relationship. Topology-based alignment comes at the cost of results that are dependent on the addition order of sequences as determined by the guide tree (Fitch and Smith, 1983). Thus, preconceived notions of relationships will bias the analysis. Randomization of the alignment topology is the most objective course of action.

The most significant advantage of direct optimization is that homology assessment is dynamic. In direct optimization, nucleotide homologies are fluid in the sense that they change not only when different guide trees are used, but also when various data are combined. Statements of putative homology depend not only on the addition order of sequences during the initial build of a cladogram and base transformation costs (as with standard alignment) *but also* on congruence among characters. In direct optimization, many optimization schemes, each implying a distinct set of putative homologies, can be examined via variable sequence alignments that occur concurrently with initial cladogram building. The diagnosis of each cladogram involves finding the lowest-cost hypothetical ancestral sequences possible. Direct optimization is accomplished by examining all possible homologies between the nucleotide bases of two descendant vertices. Dynamic programming is used (in a step akin to pairwise sequence alignment) to optimize each hypothetical ancestral sequence for the minimum weighted number of insertion-deletion events and base substitutions. At each vertex in a cladogram, all possible hypothetical ancestral sequences are implicitly constructed and their costs determined. The minimum cost ancestral sequence is retained and used to optimize the next vertex down the cladogram. Wheeler formally describes the algorithm's downpass in this volume.

Dynamic homology and combined analysis

A logical assumption is that there is one phylogeny of a natural group of organisms because there is one evolutionary history. Comparative data of various sorts reflect the phylogeny of groups under study with different levels of support. No one type of data has been demonstrated to have a high fidelity record of evolutionary history across groups of very different ages. The basic strength of the combined analysis approach lies in the ability of synapomorphies from different types of data to provide additive support for related groups. Dynamic homology takes combined analysis one step further by allowing co-optimization of molecules and morphology. Putative sequence homologies are tested and revised via optimization of their congruence with morphological synapomorphies. This contrasts sharply with standard com-

bined analyses in which prealigned sequences are attached to morphological characters. Standard analysis is restricted by static alignment to seeking for a common signal at the level of the tree search. It has been demonstrated that, in terms of character congruence and topological congruence, combining prealigned datasets produces cladograms which are suboptimal to those produced when the same raw data are analyzed with direct optimization (Wheeler, 1998).

Computational complexity of phylogenetics and heuristic solutions

Alignment

As introduced earlier, the number of cells in a multiple alignment matrix are exponentially related to the number of taxa and sequence length. Furthermore, the number of multiple alignments becomes very large with a small number of short sequences (Slowinski, 1998). As a consequence, exact solutions are intractable and heuristics are required to produce multiple alignments. Heuristic alignment algorithms get the job done at the cost of alignment ambiguity. As discussed above, one common heuristic is the use of a guide tree to direct the addition order of sequences in multiple alignment (Sankoff et al., 1973). In theory, an alignment procedure could be repeated for each possible set of relationships among the taxa. However this is intractable because of the large number of evolutionary trees with just a few taxa (discussed below). Alignment heuristics are reviewed in detail in Phillips et al. (2000). In common practice, one topology is used (e.g., as implemented in CLUSTAL [Thompson et al., 1994] and in TREEALIGN [Hein, 1990]). As discussed above, topology-based alignment comes at the cost that results are dependent on the addition order of sequences as determined by the guide tree (Fitch and Smith, 1983). This bias can be addressed by increasing the number of random additions performed which increases runtime (e.g., as implemented in MALIGN, [Wheeler and Gladstein, 2000]). Furthermore, various parameter sets for base transformation costs in alignment may lead a limited set of groups or few groups in common. In many cases when results of many parameter sets are compared, phylogenies share few groups (e.g., W.C. Wheeler, 1995; O'Leary, 1999; Giribet, 1999; Giribet et al., 2000; Giribet and Ribera, 2000; Janies, 2001). However, some analyses have shown consistent results despite parameter variation (Edgecombe et al., 1999). The implementation of topology-based alignment can be improved by concurrent examination of many guide trees and can be explored in reasonable time with an inexpensive cluster of PCs using POY or MALIGN (discussed below). Parallellization of software implemented on inexpensive computing clusters and evermore popular multiprocessing PCs provide a very efficient (in terms of maximizing analytical rigor within available time and money) means to rationally address large phylogenetic datasets at the level of sequence alignment.

Topologies

The number of networks facing a topology-based alignment or a phylogenetic tree search is combinatorily dependent on the number of taxa. Thus the number of possible topologies becomes astronomical as taxa are added to the analysis.For example, the number of possible rooted topologies increases as a power series (let y = the number of rooted topologies, let i = the starting point of 3 taxa, let t = the total number of taxa) (Cavalli-Sforza and Edwards, 1967).

$$y = \prod_{i=3}^{t} (2i - 3)$$

The number of possible rooted topologies reaches 34,459,425 with only 10 taxa, 8.2×10^{21} with 20 taxa, and 2.75×10^{76} with 50 taxa.

Practice

The challenges presented by alignment of DNA and phylogenetic tree search have prompted research in heuristics and parallellization. There are several operational reasons to do parallel direct optimization as implemented in POY. Commonly used alignment algorithms produce one (or sometimes many) alignments based on a single parameter set and distance-based addition sequence. Then the investigator has to run a phylogenetic tree search algorithm. POY produces trees, is reasonably fast under a variety of platforms and runs very fast in parallel on inexpensive clusters of PCs (Gee, 2000; Sterling et al., 1999; Janies and Wheeler, 2001). This new paradigm offered by POY permits the investigator to examine many alignment topologies (up to millions of trees per second) and ratchet and swap replicates. Furthermore, the speedup that parallelism affords permits searching a wide parameter space in reasonable time. New, fast phylogenetic search algorithms will produce short trees from a single alignment at unprecedented speed (e.g., Goloboff, 1999; Nixon, 1999). However the speed and quality of the phylogeny produced by these algorithms is dependent on the speed and quality of the alignment(s). Multiple alignment can take weeks of processing time on desktop computers. POY challenges the existing paradigm of alignment followed by a separate tree search, by unifying these steps into a single algorithm that is efficiently scalable to the large datasets necessary to make sense of the large amounts of data being produced by high-throughput DNA sequencing and character coding.

Efficiency of parallel direct optimization

Four major algorithms of POY were tested for parallel efficiency: two types of initial cladogram building and two types of branch swapping. Random repli-

cates of initial cladogram builds were distributed to several processors via a one-processor-per-replicate strategy (via the commands -parallel -multibuild n). Alternatively, single cladogram builds were partitioned across many processors (via the command -parallel). Branch swapping jobs were partitioned across many processors (via the commands -parallel -tbr -spr). These algorithms were tested on several datasets comprised of DNA and morphology ranging from 40–500 taxa (Janies and Wheeler, 2001).

The results of these studies are straightforward and very informative on the scaling properties of POY on large and small clusters. The results on the large cluster (256 processor cluster comprised of Intel Pentium 500 MHz PIIIs networked via 100 Mbps switched Ethernet) contrast significantly (for some but not all algorithms) with those derived from similar studies on a small cluster (11 processor cluster comprised of Intel Pentium 200 Mhz PIIs networked via a 10 Mbps Ethernet hub) previously in service at the AMNH. Various algorithms in POY show fundamentally different properties within and between clusters.

The multibuild command exhibits excellent parallel efficiency in the large cluster (Fig. 1). Speedup (trees examined per second) is very close to linear with the addition of processors regardless of dataset or cluster size. In contrast, parallel building shows poor parallel efficiency in the large cluster with only slight speedup up to 128 slave processors (Fig. 2). This result is similar in large and small clusters. Branch swapping commands show excellent speedup for 10

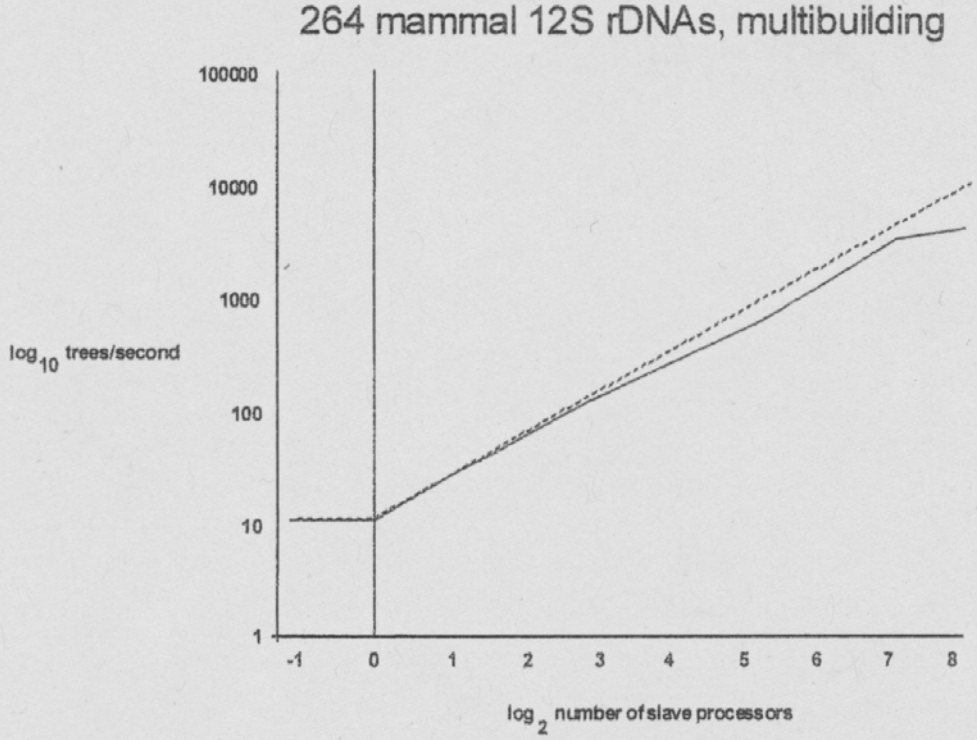

Figure 1. Parallel efficiency of the one-processor-per-replicate strategy using the POY commands (-parallel -multibuild n). The dotted line represents perfect parallel speedup. The solid line represents actual speedup. The multibuild command exhibits excellent parallel efficiency for 264 mammal 12S rDNA and results are similar for other large datasets.

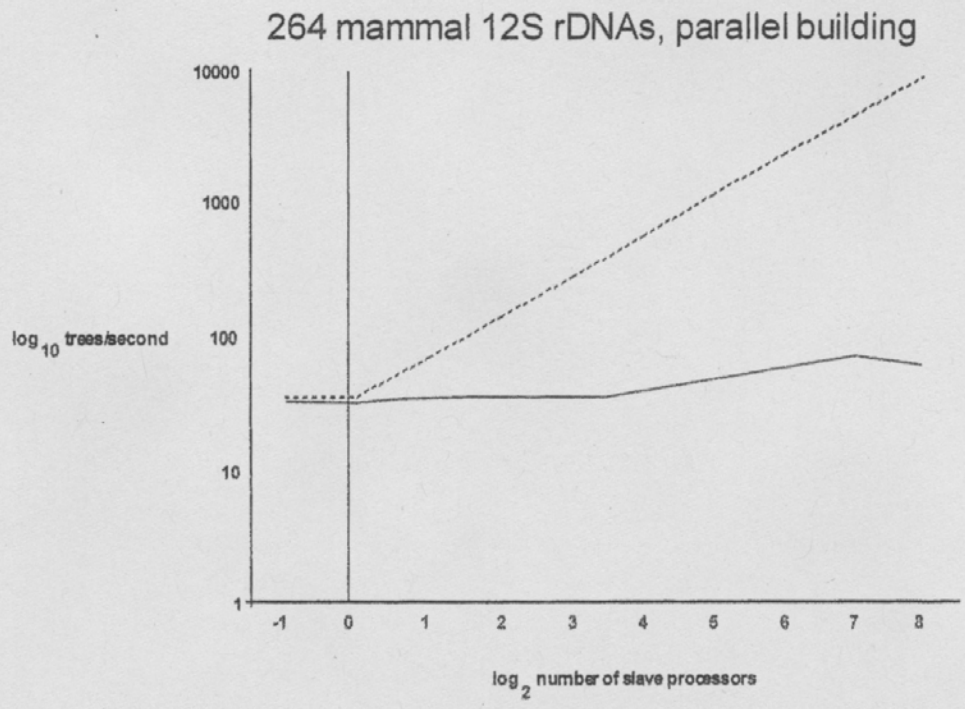

Figure 2. Parallel efficiency of a strategy in which work of each single cladogram build is partitioned across many processors using the POY command -parallel. The dotted line represents perfect parallel speedup. The solid line represents actual speedup. The parallel command exhibits poor parallel efficiency for 264 mammal 12S rDNA and results are similar for other large datasets.

slave processors on the small cluster and excellent speedup for 32 slave processors on the large cluster (Fig. 3.). However, there is no appreciable speedup with the addition of slave processors and this result is independent of dataset size.

These results are fundamental to improving the algorithms for hierarchical parallelism and multi-user load balancing to achieve maximum performance per unit investment. Furthermore, the excellent parallel efficiency of the multibuild command is very encouraging. This result demonstrates the viability of building clusters comprised of several hundred of processors without investing in expensive, non-standard, network hardware. Also, it will be important to invest resources in obtaining higher clockspeed processors to shorten per–node runtimes when using multibuild.

Progress in phylogenetic analysis of DNA sequence data is limited by computational capacity. Advances in DNA sequencing technology have permitted the accumulation of phylogenetic data sets with hundreds to thousands of taxa, each with thousands of nucleotides. Parallelism offers a tractable means to create the computational power required for aggressive heuristic searches. The ongoing development of parallel algorithms combined with the low cost and simplicity of off-the-shelf hardware make cluster computing a revolutionary technology for evolutionary biology.

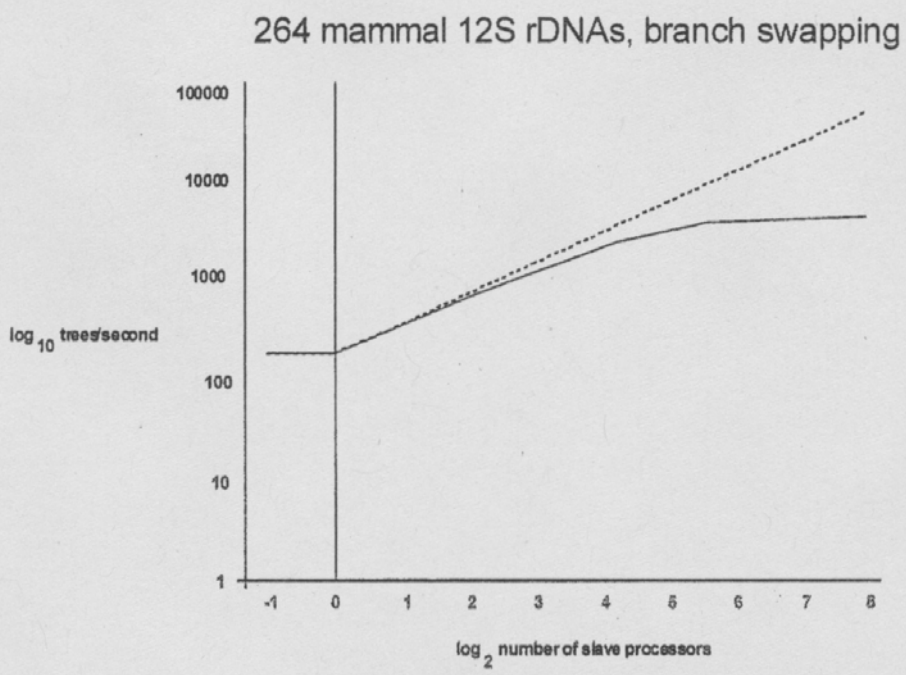

Figure 3. Parallel efficiency of branch swapping using the POY commands -parallel -tbr -spr. The dotted line represents perfect parallel speedup. The solid line represents actual speedup. Branch swapping on trees based on 264 mammal 12S rDNAs in parallel shows excellent speedup for 32 slave nodes but additional processors provide no appreciable speedup. This result is independent of dataset size.

Acknowledgements

The National Aeronautics and Space Administration, the American Museum of Natural History and the New York City Department of Cultural Affairs provided research funding. Lisa Gugenheim, Tim Mohrmann, Pete Makovicky, Diego Pol, Estelle Perrera, Al Phillips, Julian Faivovich and Rebecca Klasfeld of the AMNH were instrumental in the procurement and construction of the parallel cluster. Valuable comments on the manuscript were provided by Gonzalo Giribet, Susan Perkins and Marc Allard. Test dataset for mammals was provided by Ginny Emerson.