CANCER GENOMICS & PROTEOMICS *2*: 301-306 (2005)

Review

Phylogenetics: Applications, Software and Challenges

ALEXANDROS STAMATAKIS

Foundation for Research and Technology-Hellas, Institute of Computer Science, Crete, Greece

\Abstract. *Inference of phylogenetic trees comprising hundreds of organisms based on elaborate statistical models of evolution is an intensive computational task. However, in recent years there has been an impressive improvement in search algorithms, which currently allow for inference of huge phylogenetic trees comprising more than 1,000 taxa within a couple of hours on a single PC. This paper provides an overview of applications of phylogenetic trees to various areas of biological and medical research and reviews some of the most efficient software available for phylogenetic inference. Finally, some of the new challenges that the field currently faces in the areas of high performance computing and information visualization are discussed.*

Phylogenetic trees are used to represent the evolutionary history of a set of n organisms which are often also called taxa within this context. A multiple alignment of a suitable, in a biological context, small region of their DNA or protein sequences can be used as input for the computation of phylogenetic trees.

In a computational context, phylogenetic trees are usually strictly bifurcating (binary) unrooted trees. The organisms of the alignment are located at the tips (leaves) of such a tree, whereas the inner nodes represent extinct common ancestors. The branches of the tree represent the time which was required for the mutation of one species into another, new, one. The most fundamental algorithmic problem computational phylogeny faces consists of the immense amount of potential alternative tree topologies. This number grows exponentially with the number of sequences n, *e.g.,* for $n=50$ organisms there already exist $2.84*10^{76}$ alternative topologies; a number almost as large as the number of atoms in the universe ($\approx 10^{80}$). Thus, given some, biologically meaningful, optimality criterion for evaluating all alternative configurations (topologies) in order to search for the best tree, one can quickly assume that the problem might be NP-hard. NP-hard problems in computer science are hard-to-solve optimization tasks. Typically, there is a great number of different configurations – tree topologies in the specific case – which have to be evaluated using some function $f()$, *e.g.,* the likelihood function, in order to find the configuration which minimizes or maximizes $f()$. Due to the sheer size of the search space for NP-hard problems, it is usually not possible to compute the optimal solution. Therefore, appropriate intelligent heuristics have to be deployed which only explore a small fraction of this *gigantesque* search space and typically only yield suboptimal solutions. For some heuristics and problems a guaranteed worst-case performance exists, *i.e.* a formula that states that the solution found by the algorithm will be, for example, at most 10% from optimal. In phylogenetics, however, no such guarantee exists, especially for very large trees computed with Maximum Likelihood (ML). The NP-hardness of ML has recently been demonstrated (9). Another important aspect for the design of heuristic tree searches consists of the very high degree of accuracy (difference to the score of the optimal or best-known solution) which is required to obtain reasonable biological and topologically closely-related results. While an accuracy of 90% is considered to be a "good" value for heuristics designed to solve "classic" NP-hard optimization problems, such as the traveling salesman problem, recent results suggest that phylogenetic analyses require an accuracy $\geq 99.99\%$, in particular for large trees (44). This observation yields the whole field more difficult and challenging. Regarding the various evolutionary models which have been proposed for phylogenetic inference, a *trade-off* exists between speed and quality. This means that a phylogenetic analysis conducted with an elaborate model such as ML requires significantly more time but yields trees with superior accuracy than, for example, Neighbor Joining (NJ) or Maximum Parsimony (MP). However, due to the higher

Correspondence to: Alexandros Stamatakis, Foundation for Research and Technology{Hellas, Institute of Computer Science, P.O. Box 1385, Heraklion, Crete, GR-71110 Greece. e-mail: stamatak@ics.forth.gr

Key Words: Phylogenetic interference, RAxML, phylogenetic software, review.

accuracy, it is desirable to infer large and complex trees with ML. Within this context, it is important to emphasize that the design of ML programs used to be mainly an *algorithmic discipline*. Therefore, progress in the field has been attained through algorithmic innovations rather than by brute force allocation of all available computational resources, *e.g.* large supercomputers or even grids of supercomputers. However, because of the major algorithmic advances over the last couple of years, technical implementation aspects and High Performance Computing (HPC) implementations are becoming increasingly important. Many state-of-the-art sequential search algorithms, which in principle are able to reconstruct very large trees of 5,000 taxa and more, face considerable technical problems with respect to memory and resource shortages.

Applications

The inference of phylogenies with computational methods has many important applications in medical and biological research, such as drug discovery and conservation biology. A result published by Korber *et al.* (19), that times the evolution of the HIV-1 virus, demonstrates that ML techniques can be effective in solving biological problems. Phylogenetic trees have already witnessed applications in numerous practical domains, such as in conservation biology (3) (illegal whale hunting), epidemiology (5) (predictive evolution), forensics (27) (dental practice HIV transmission), gene function prediction (7) and drug development (14). Other applications of phylogenies include multiple sequence alignment (11, 25), protein structure prediction (31), gene and protein function prediction (12, 22) and drug design (30). A paper by Bader *et al.* (2) addresses important industrial applications of phylogenetic trees, *e.g.* in the area of commercial drug discovery. Due to the rapid growth of available sequence data over recent years and the constant improvement of multiple alignment methods, it has now become feasible to compute very large trees which comprise more than 1,000 organisms. The computation of the tree-of-life containing representatives of all living beings on earth is considered to be one of the *grand challenges* in Bioinformatics. Some large multi-institutional/multi-disciplinary projects are underway which aim at building the tree of life: CIPRES (Cyber Infrastructure for Phylogenetic Research www.phylo.org) and ATOL (Assembling the Tree of Life project, tolweb.org).

Software for Phylogenetic Analysis

The review of current software for phylogenetic analysis is restrained to statistical phylogeny methods, since the general consensus is that they represent the most accurate methods currently available.

Performance studies. A thorough comparison of popular phylogeny programs using statistical approaches such as fastDNAml (26), MrBayes (15), PAUP* (paup.csit.fsu.edu) and TREE-PUZZLE (41) on small simulated datasets (up to 60 sequences) has been conducted by Williams *et al.* (43). The most important result of this paper is that MrBayes outperforms all other phylogeny programs in terms of speed and tree quality. However, the results of this survey do not necessarily apply to large real datasets, since simulated alignment data has different properties and a significantly stronger phylogenetic signal than real world data [see (37) for a discussion]. Therefore, much more computational effort is required to find a "good" phylogenetic tree for real-world data. Due to the significant differences between real and simulated datasets, comparative surveys should include collections of simulated and real datasets in order to yield a more complete image of program performance. In fact, some real datasets exist for which MrBayes fails to yield acceptable trees within reasonable time (36). Huelsenbeck *et al.* (16) provide an in-depth discussion of the potential pitfalls of Bayesian inference.

Sequential algorithms. In 2003, Guidon *et al.* published an interesting paper about their very fast new program PHYML (13). The performance of PHYML has been tested on medium-sized simulated datasets of 100 sequences and two well-studied real datasets containing 218 and 500 sequences. The main advantage of PHYML consists of its speed and in a very comprehensive implementation of nucleotide and amino acid substitution models. On real-world data, however, the current search algorithms implemented in RAxML clearly outperform PHYML, both in terms of execution time and final tree quality (34, 37). The requirement to improve accuracy on real data has been recognized by the authors of PHYML. In fact, they have recently integrated a very promising improvement of the *lazy subtree rearrangement* technique from RAxML (34) into PHYML (personal communication). Another main advantage of RAxML over most other programs consists of an extremely efficient technical implementation of the likelihood function, which consumes ≥90% of the total execution time in most ML implementations. A further advantage, which is becoming more important at present, is the significantly lower memory consumption of RAxML. Irrespective of these differences between RAxML and PHYML, the results in (13) show that well-established sequential programs like PAUP*, TREE-PUZZLE (41) and fastDNAml (26) are prohibitively slow on datasets containing more than 200 sequences, at least in sequential execution mode. More recently, Vinh *et al.* (42) published a program called IQPNNI – an improved version of TREE-PUZZLE – which yields better trees than PHYML on real-world data, but is significantly slower (37). Finally,

MetaPIGA (21), though also slower than RAxML and PHYML due to a different search technique (genetic algorithm), represents a very user-friendly program. A significantly more efficient version will soon be released (personal communication).

A slightly different class of search algorithms are the so-called divide- and- conquer approaches, that intelligently split the problem into smaller subproblems which can be solved more efficiently. Rec-I-DCM3 (29) is a recently introduced meta-method for divide-and-conquer tree searches. It has to be used in conjunction with a base method for ML (*e.g.* RAxML) which is applied to compute subtrees and further optimizes the comprehensive tree. Initial results indicate that it is able to significantly improve upon the quality of final trees compared to stand-alone RAxML on large alignments (10).

In the final analysis, it can be stated that PHYML, RAxML, IQPNNI and MetaPIGA are currently among the fastest freely available software packages for phylogenetic inference. MrBayes is also a very fine tool, but should be used in conjunction with one of the above programs to avoid the aforementioned potential pitfalls.

Parallel phylogeny programs. Most parallel implementations of ML programs are technically very solid in terms of performance and parallelization techniques. However, they drag behind algorithmic development, *i.e.* relatively old and slow search algorithms are parallelized. For example, the largest tree computed with parallel fastDNAml (39) (2001), which is based on the fastDNAml algorithm from 1994, contained 150 taxa. The same argument holds for a technically very interesting JAVA-based distributed implementation of fastDNAml: DPRml (18). In addition to using the same old search algorithm, significant performance penalties are caused by using JAVA both in terms of memory efficiency and speed of numerical calculations. Those technical limitations will become more intense when trees comprising over 417 taxa (currently largest tree with DPRml, personal communication) are computed. The authors of DPRml are working on the aforementioned issues (personal communication). The largest phylogenetic analysis conducted with the parallel version of TREE-PUZZLE contained 257 taxa due to limitations caused by the data structures (personal communication). IQPNNI has also recently been parallelized and exhibits very good parallel efficiency (23). Brauer *et al.* (4) are currently working on an improved parallel version of GAML which can compute trees of up to 3,000 organisms. Once again, the main limitation for the computation of larger trees is memory consumption (personal communication). Nonetheless, the new tree search algorithm of GAML appears to be at least as powerful as the RAxML algorithms, but requires higher inference times (personal communication). As already mentioned, a parallel version of Rec-I-DCM3 for maximum likelihood also exists which is based on RAxML (10). The current implementation faces some scalability limitations due to load imbalance, which in turn leads to a relatively "bad" parallel efficiency, caused by significant differences in the subproblem sizes. Finally, the previous parallel and distributed implementations of the RAxML algorithm exist (33, 35). Parallel RAxML has been used to compute the – to the best of the author's knowledge – largest ML-based phylogeny to date, containing 10,000 organisms on a medium-sized PC cluster using a total of approximately 3,200 accumulated CPU hours (35).

All parallel implementations that incorporate recent algorithmic advances like RAxML, Rec-I-DCM3(RAxML), IQPNNI and GAML represent a good choice for inference of huge phylogenetic trees, though the latter two can potentially encounter memory shortages.

Challenges

Memory efficiency. A very important challenge for the design and implementation of phylogeny programs is reduction of the memory consumption and improvement of the cache efficiency for two main reasons: firstly, the computation of very large trees has become feasible with the new generation of search algorithms and, due to the immense accumulation of data, alignments are also constantly growing in both dimensions: alignment length and number of taxa. As outlined in the previous section, the applicability to larger problems of most current programs is limited by memory consumption. Secondly, for several years now the speed of Central Processing Units (CPUs) has been growing at a higher rate than the memory access speed, such that the performance of a large class of scientific applications is nowadays limited by their memory access pattern rather than by the CPU speed. Currently, it is unlikely that this trend will be reversed. Some strategies to overcome this burden consist of optimizing programs to reduce memory consumption on a technical level, deployment of divide-and-conquer strategies to reduce the size of the problem and exploitation of powerful shared-memory processors (32). Finally, recent approaches also exploit the immense computational potential of peripheral hardware such as Graphics Processing Units (GPUs) (8).

Implementation and optimization of the likelihood function. Another important area of research focuses on the optimization of the likelihood function, which typically consumes over 90% of total execution time in programs such as RAxML or PHYML (32). Some approaches (20, 38) focus on detecting equal patterns in the alignments and re-using previously calculated values instead of re-computing

them each time. In addition, using a separate implementation of the compute-intensive functions for each individual model of nucleotide substitution, as well as applying low-level technical optimizations (*e.g.* manual loop unrolling) to the likelihood functions, will become essential for computation of very large trees.

Information visualization. Despite the algorithmic advances in the field, only a few adequate visualization tools are available for the analysis of very large trees. Thus, the design of novel tree-viewing tools is crucial in order to accelerate the analysis process, as well as to extract useful information from the data and expedite the cognitive process. Among the most popular representations are phylogram, radial and slanted cladogram drawings (24). Such representations are provided by common tree-viewing programs such as ATV (45). However, these layouts and programs are targeted at medium-sized trees comprising a maximum of 300-400 taxa. Thus, they are not well-suited to visualize large trees with thousands of taxa. Approaches for larger trees make use of two-dimensional and three-dimensional (17) hyperbolic space in order to simultaneously provide a detailed, as well as contextual, view of the tree. Other approaches such as SpaceTree (28) only display representative parts of very large trees. However, biologists usually prefer a simultaneous detailed display and contextual view of phylogenies. The use of treemaps to display phylogenetic trees has recently been proposed (1), but this concept is also limited to a maximum of 2,000-3,000 taxa. There are also approaches based on virtual reality (40) which are, however, not accessible to most researchers due to the sheer cost of the respective infrastructure. Carrizo (6) provides a readable and comprehensive review of efforts to appropriately display phylogenetic trees from an information visualization perspective. Nonetheless, since no really satisfying solution currently exists, the design of appropriate visualization tools is becoming an issue of increasing importance, since otherwise the information contained in large phylogenies will be useless.

Other issues. Other current issues concern, for example, the development of more complex and realistic statistical models of sequence evolution, the assessment of final tree quality as well as accuracy, new methods to infer phylogenetic networks and phylogenetic inference based on gene-order data. Moreover, the simultaneous computation of multiple alignments and phylogenetic trees is also receiving more attention.

Outlook

As outlined in this paper, there are already a large number of applications of phylogenetic trees to real-world biological and medical problems. Moreover, numerous efficient programs are freely available – mostly as open source code – for sequential and parallel phylogenetic inference, which incorporate a plethora of advanced search techniques. Nonetheless, high performance computing techniques will have to be increasingly deployed in order to cope with the pace of algorithmic development and data accumulation. Another serious problem is the absence of appropriate visualization tools for very large trees, which has a negative impact on the interpretability of the results. However, promising new algorithmic developments, which significantly reduce the amount of required evaluations of the likelihood function, coupled with an increasing awareness about the aforementioned high performance computing issues in the community, are likely to allow for parallel inference of trees containing 10,000-20,000 sequences on medium-sized PC clusters in the near future.

Acknowledgements

This work was funded by a Postdoc-fellowship granted by the German Academic Exchange Service (DAAD).

References

1 Arvelakis A, Reczko M, Stamatakis A, Symeonidis A and Tollis IG: Using treemaps to visualize phylogenetic trees. Proc of ISMBDA2005, in press.
2 Bader DA, Moret BME and Vawter L: Industrial applications of high-performance computing for phylogeny reconstruction. Proc of SPIE ITCom *4528*: 159-168, 2001.
3 Baker CS and Palumbi SR: Which whales are hunted? A molecular genetic approach to whaling. Science *265*: 1538-1539, 1994.
4 Brauer MJ, Holder MT, Dries LA, Zwickl DJ, Lewis PO and Hillis DM: Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. Mol Biol Evol *19*: 1717-1726, 2002.
5 Bush RM, Bender CA, Subbarao K, Cox NJ and Fitch WM: Predicting the evolution of human in uenza. Science *286(5446)*: 1921-1925, 1999.
6 Carrizo SF: Phylogenetic trees: an information visualisation perspective. Proc of APBC2004, 315-320, 2004.
7 Chang BS and Donoghue MJ: Recreating ancestral proteins. Trends Ecol Evol *15*: 109-114, 2000.
8 Charalambous M, Trancoso P and Stamatakis A: Initial experiences supporting a bioinformatics application to a graphics processor. Proc of PCI 2005, in press.
9 Chor B and Tuller T: Maximum likelihood of evolutionary trees is hard. Proc of RECOMB05, 2005.
10 Du Z, Stamatakis A, Lin F, Roshan U and Nakhleh L: Parallel divide-and-conquer phylogeny reconstruction by maximum likelihood. Proc of HPCC05, in press.
11 Edgar RC: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucl Acid Res *32(5)*: 1792-1797, 2004.
12 Eisen JA: Phylogenomics: intersection of evolution and genomics. Science *300*: 1706-1707, 2003.

13 Guindon S and Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol *52(5)*: 696-704, 2003.
14 Halbur P, Lum MA, Meng X, Morozow I and Paul PS: New porcine reproductive and respiratory syndrome virus DNA and proteins encoded by open-ended frames of an Iowa strain of the virus are used in vaccines against PRRSV in pigs. Patent-Filing WO9606619-A1, 1994.
15 Huelsenbeck JP and Ronquist F: MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17(8)*: 754-755, 2001.
16 Huelsenbeck JP, Larget B, Miller RE and Ronquist F: Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol *51(5)*: 673-688, 2002.
17 Hughes T, Hyun Y and Liberles DA: Visualizing very large phylogenetic trees in three dimensional hyperbolic space. BMC Bioinformatics *5(48)*, 2004.
18 Keane TM, Naughton TJ, Travers SAA, McInerney JO and McCormack GP: Dprml: distributed phylogeny reconstruction by maximum likelihood. Bioinformatics *21(7)*: 969-974, 2005.
19 Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BW, Wolinsky S and Bhattacharya T: Timing the ancestor of the HIV-1 pandemic strains. Science *288*: 1789-1796, 2000.
20 Kosakovsky-Pond SL and Muse SV: Column sorting: rapid calculation of the phylogenetic likelihood function. Syst Biol *53(5)*: 685-692, 2004.
21 Lemmon AR and Milinkovitch MC: The metapopulation genetic algorithm: an effcient solution for the problem of large phylogeny estimation. PNAS *99*: 10516-10521, 2002.
22 La D, Sutch B and Livesay DR: Predicting protein functional sites with phylogenetic motifs. Prot Struct Funct Bioinf *58(2)*: 309-320, 2005.
23 Minh BQ, Vinh LS, Haeseler A v and Schmidt HA: pIQPNNI – parallel reconstruction of large maximum likelihood phylogenies. Bioinformatics, advance on-line access, 2005.
24 Munzner T, Guimbretiere F, Tasiran S, Zhang L and Zhou Y: Treejuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. Proc of SIGGRAPH 2003, 2003.
25 Notredame C, Higgins D and Heringa J: T-coffee: A novel method for multiple sequence alignments. J Mol Biol *302*: 205-217, 2000.
26 Olsen G, Matsuda H, Hagstrom R and Overbeek R: fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput Appl Biosci *10*: 41-48, 1994.
27 Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT, Mullins JI, Schochetman G, Berkelman RL and Economou AN: Molecular epidemiology of HIV transmission in a dental practice. Science *256(5060)*: 1165-1171, 1992.
28 Plaisant C, Grosjean J and Bederson BB: Spacetree: supporting exploration in large node link tree, design evolution and empirical evaluation. Proc of InfoVis2004, 57-70, 2002.
29 Roshan U, Moret BME, Warnow T and Williams TL: Rec-I-DCM3: a fast algorithmic technique for reconstructing large phylogenetic trees. Proc CSB2004, Stanford, California, USA, 2004.
30 Searls DB: Pharmacophylogenomics: gene, evolution, and drug targets. Nat Rev Drug Disc *2*: 613-623, 2003.
31 Shindyalov IN, Kolchanov NA and Sander C: Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Prot Eng *7*: 349-358, 1994.
32 Stamatakis A, Ott M and Ludwig T: RAxML-OMP: An efficient program for phylogenetic inference on SMPs. Proc of PaCT05, in press.
33 Stamatakis A, Lindermeier M, Ott M, Ludwig T and Meier H: Draxml@home: a distributed program for computation of large phylogenetic trees. FGCS *51(5)*: 725-730, 2005.
34 Stamatakis A, Ludwig T and Meier H: RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics *21(4)*: 456-463, 2005.
35 Stamatakis A, Ludwig T and Meier H: Parallel inference of a 10,000-taxon phylogeny with maximum likelihood. Proc Euro-Par2004, 997-1004, 2004.
36 Stamatakis A, Ludwig T and Meier H: New fast and accurate heuristics for inference of large phylogenetic trees. Proc of IPDPS2004, Santa Fe, New Mexico, April 2004.
37 Stamatakis A: An efficient program for phylogenetic inference using simulated annealing. Proc of IPDPS2005, Denver, Colorado, April 2005.
38 Stamatakis A, Ludwig T, Meier H and Wolf MJ: AxML: A fast program for sequential and parallel phylogenetic tree calculations based on the maximum likelihood method. Proc of CSB2002, 21-28, Palo Alto, California, August 2002.
39 Stewart C, Hart D, Berry D, Olsen G, Wernert E and Fischer W: Parallel implementation and performance of fastDNAml – a program for maximum likelihood phylogenetic inference. Proc of SC2001, November 2001.
40 Stolk B, Abdoelrahman F, Koning A, Wielinga P, Neefs JM, Stubbs A, de Bondt A, Leemans P and van der Spek P: Mining the human genome using virtual reality. Proc of EGPGV02, 17-21, 2002.
41 Strimmer K and Haeseler AV: Quartet puzzling: a maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol *13*: 964-969, 1996.
42 Vinh LS and Haeseler AV: IQPNNI: Moving fast through tree space and stopping in time. Mol Biol Evol *21(8)*: 1565-1571, 2004.
43 Williams TL and Moret BME: An investigation of phylogenetic likelihood methods. Proc of BIBE'03, 2003.
44 Williams TL, Berger-Wolf BM, Roshan U and Warnow T: The relationship between maximum parsimony scores and phylogenetic tree topologies. Tech. Report, TR-CS-2004-04, Department of Computer Science, The University of New Mexico, 2004.
45 Zmasek CM and Eddy SR: Atv: Display and manipulation of annotated phylogenetic trees. Bioinformatics *17*: 383-384, 2001.

Received July 29, 2005
Accepted August 31, 2005