Scalable Phylogenetics through Input Preprocessing

Roberto Blanco, Elvira Mayordomo,
Esther Montes, Rafael Mayo, and Angelines Alberto

Abstract. Phylogenetic reconstruction is one of the fundamental problems in computational biology. The combinatorial explosion of the state space and the complexity of mathematical models impose practical limits on workable problem sizes. In this article we explore the scalability of popular algorithms under real datasets as problem dimensions grow. We furthermore develop an efficient preclassification and partitioning strategy based on guide trees, which are used to intently define an evolutionary hierarchy of groups of related data, and to determine membership of individual data to their corresponding subproblems. Finally, we apply this method to efficiently calculate exhaustive phylogenies of human mitochondrial DNA according to phylogeographic criteria.

1 Motivation

The organization of living organisms (extant or not) into a "tree of life", as conceived by Darwin, is the purpose of the modern discipline of phylogenetics [4]. Continuous advances in sequencing technologies have supported an exponential growth in publicly available biological sequences over the last quarter century. This abundance offers extraordinary potential to shed light into the inner workings of evolution.

Phylogenetic techniques infer trees, or generalizations thereof, from multiple sequence alignments according to a certain optimality criterion. This mathematical

Roberto Blanco · Elvira Mayordomo
Departamento de Informática e Ingeniería de Sistemas (DIIS)/Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, María de Luna 1,
50018 Zaragoza, Spain
e-mail: {robertob,elvira}@unizar.es

Angelines Alberto · Rafael Mayo · Esther Montes
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT),
Avenida Complutense 22, 28040 Madrid, Spain
e-mail: {angelines.alberto,rafael.mayo,esther.montes}@ciemat.es

M.P. Rocha et al. (Eds.): IWPACBB 2010, AISC 74, pp. 123–130, 2010.
springerlink.com

scoring scheme acts as a computational surrogate of the true biological objective: to recover the true evolutionary relationships between living organisms, as represented by (usually) aligned, meaningful sequences.

Nevertheless, this is a very challenging problem. Not only are there no efficient methods to calculate an optimal phylogeny given an alignment and a tree scoring function, but the preservation of the optimality of a solution while adding new data to it has been proven to be NP-hard for even the elementary parsimony criterion [3]. Therefore, standard practice depends on heuristic methods, which still remain very costly.

From an information-rich perspective, the main defect of conventional methods is that they are completely blind, whereas it would be desirable to provide hints and facts that may constrain and help to simplify problem solution. In this paper, we resort to evolutionary hypotheses as classifiers and definers of smaller, simpler subproblems, and illustrate the biological and computational benefits of such a methodology.

2 Methods

In this section we introduce the working principles of our technique, providing the experimental grounds for it as well as its theoretical basis. The illustration of the practice and benefits of this framework is deferred to the next section.

2.1 Stand-Alone Algorithms

There exist many phylogenetic reconstruction methods, but their principal classification regards whether they treat statistical quality evaluation as an additional step or as part of the process itself [6]. This assessment is of the utmost importance because of the infeasibility of exact algorithms for even small datasets. Both approaches have strengths and weaknesses, and thus method selection is a very important choice, not least due to the very poor scaling of most techniques.

To assess the severity of these decisions, we have evaluated the relative performance and scalability of the main families of methods; our results are summarized in Fig. 1. Problem dimensions have been chosen to scale up to moderately sized, real datasets, with which we will deal in Sect. 3. PHYLIP and MrBayes [5, 8] have been used for traditional and Bayesian methods, respectively, due to their availability and generality.

All three traditional methods are seen to follow similar trends, though parsimony is more abrupt in its progression. Generally speaking, more thorough methods incur in significantly higher costs, quickly becoming impractical. It should be noted that these times must be multiplied by the desired number of bootstrap samples for the analysis, which can nevertheless be executed in parallel.

The raw computational needs of Bayesian methods are more difficult to estimate due to their simulational nature, as is their parallelization. Cost ultimately depends on both problem dimensions, number of iterations (and its relation to model

convergence and stop conditions), and possibly number of processors. For the sake of measurement we have fixed an appropriate, catch-all number of iterations and assume approximately linear speed-up for the execution setup [1]. Time growth is comparable, though steeper and orders of magnitude above likelihood search, which together with indivisibility make these methods impracticable for large datasets.

Generally speaking, it can be concluded that, since problem complexity is always superlinear, partitioning offers very significant improvements and allows agile production of cleaner results; therefore, it should be exploited whenever possible. These gains can be invested in the calculation of a higher degree of statistical support or the selection of more sophisticated methods and substitution models.

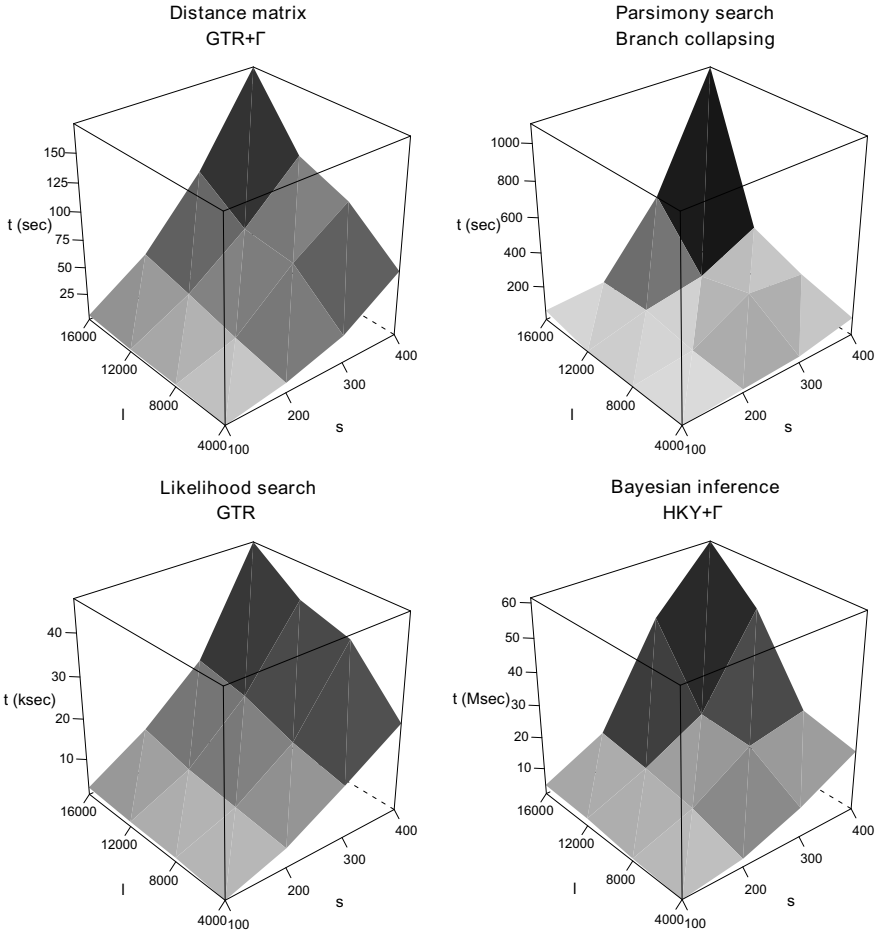

Fig. 1 Performance and scalability of phylogenetic reconstruction methods along both problem dimensions: number of sequences s and sequence length l

2.2 The Supertree Approach

Supertree methods have been conceived to combine, and possibly reconcile, a number of input subtrees according to the underlying relations between them that are manifested through a set of overlapping taxa, which are used to puzzle the trees together [10]. This approach is advantageous in that it is ideally cost-effective and allows comprehensive studies while preserving known results.

If inputs are reasonably structured and compatible, i.e., there are no unsolvable contradictions between the inferred sets of clades, it is possible to produce exact solutions in polynomial time [11]. What we develop here is a further simplification of the compositional burden based on prior structural knowledge about the data.

Our proposal is an extension of the hierarchical classification problem. If a decision tree can be provided that allows simple, recursive categorization of single sequences and that reflects well-supported phylogenetic knowledge, such a "skeleton tree" can be employed to obtain groups of related sequences, thus splitting the data into their corresponding subclades. Each of these can be subsequently computed independently and in parallel. In essence, we use supertrees to reduce problem dimensionality through a classic divide and conquer scheme.

Insofar as we know for certain, or are willing to assume as hypotheses, the relations between the clades defined by the preprocessing hierarchy, composition of partial results into the final tree is limited to substitution of symbolic "clade nodes" by their associated subtrees. The substitution skeleton must, however, accommodate all clades as leaves of the tree, transferring internal categories down dummy leaf branches if needed. The process is straightforward save for this provision.

3 Case Study: Human Mitochondrial DNA

Mitochondrial DNA (mtDNA) is one of the most important evolutionary markers for phylogenetics due to its remarkable features: absence of effective recombination, high mutation rate and ease of sequencing, among others. Its prime metabolic roles also grant it prominence in the study of rare genetic disease [13].

Consequently, comprehensive research on human mtDNA is of great interest, though the very own wealth of available information deters straightforward, manually supervised trees; we have previously addressed the question in [2] and subsequent work. At present we can effectively perform periodic updates to the human mitochondrial phylogeny, though as we endeavor to show there is plenty of room for improvement.

3.1 Structural Properties

For the proposed supertree methods to be applicable, firstly a suitable set of classifiers needs to be identified. In the case of mtDNA, its matrilineal inheritance along with the migrations that scattered the human species around the world gave rise to mitochondrial haplogroups: large population groups related by common descent,

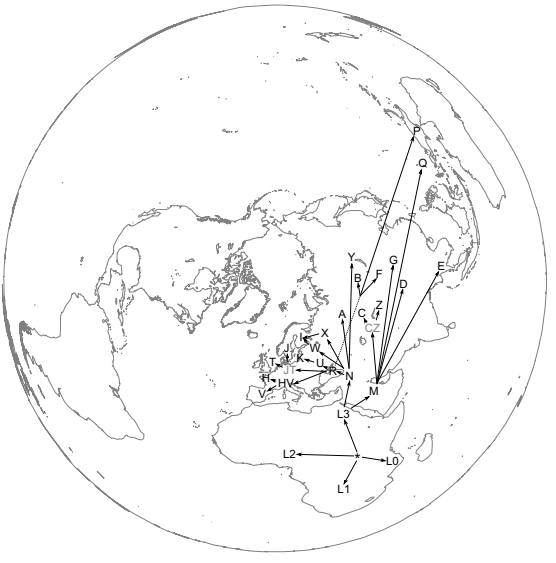

Fig. 2 Phylogeographical history of the major human mitochondrial haplogroups, fitted to MITOMAP's tree skeleton. The star marks the root of the tree. Internal clades that were excluded from the hierarchy are greyed out

to which membership can be ascribed simply by checking a handful of defining polymorphisms.

These groups spread in an arborescent fashion, not unlike a conventional species tree, as depicted in Fig. 2. The study of human haplogroups is a well-founded area of ongoing research with diverse applications [12] and a standard cladistic notation established in [7]. It is therefore perfectly adequate to our purposes as a recipient of phylogenetic knowledge.

3.2 Materials and Methods

For the conduction of our experiments we have prepared a curated database of completely sequenced, human mtDNA sequences, partly based on MITOMAP's query on GenBank. The aligned sequences ($s = 4895$, $l = 16707$) can be assumed to be homogeneous and show no obvious flaws.

Also from the MITOMAP phylogeny [9] we have selected its "Simplified mtDNA lineages" as an adequate guide tree that offers a good level of detail for a first classification. Minor corrections have been made regarding haplogroups B (allowing room for ambiguity in the definition of its distinctive indel) and I (avoiding a reticulation event for which no evidence has been found and promoting it as a direct descendant of N).

Recursive classification has been performed to assign each sequence to its related haplogroup; sequences that fail to match exactly one category have been excluded from the analysis (1.7% of the total, a very small fraction given the low complexity of the skeleton). Haplogroup subtrees have been computed using distance matrices with neighbor joining clustering and bootstrap sampling, and Bayesian inference; both under adequate substitution models. Executions have been distributed across high-performance clusters to exploit the potential parallelism unveiled by our method.

3.3 Results

As a result of the classification we obtain 28 clade subsets with sizes up to $s = 583$ owing to unequal sequencing of population groups around the globe. Due to the progression of computation times with s, a handful of prolific groups clearly dominates total execution times. However, the recursive nature of the tree makes it possible to easily refine copious nodes by simply identifying and adding children

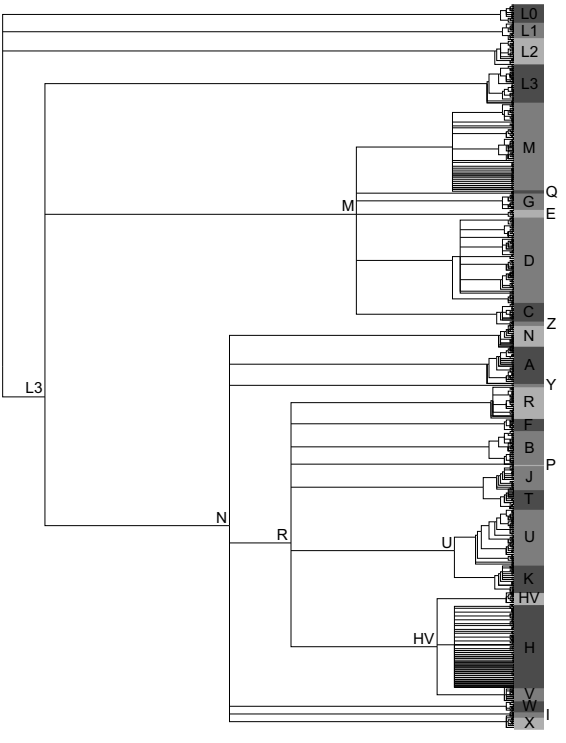

Fig. 3 Human mitochondrial haplogroup supertree. Parent haplogroups have been propagated to the top of their subtrees as leaf haplogroups; they are also labels of their associated clades

subtrees as required; this has indeed been the case of Bayesian inference of groups D, H, M and U, due to technical limitations. In fact, although traditional methods yield satisfactory throughput, Bayesian inference remains only marginally tractable and on the whole depends on more extensive partitioning.

The improvements derived from preprocessing are very remarkable. As a reference, the cost of computing a single distance matrix for the selected dataset with a suitable, reasonably complex evolutionary model is approximately 10^5 sec. By contrast, a single workstation requires the same time to produce reasonably supported (e.g., 100 bootstrap samples per haplogroup) combined phylogenies under the same algorithm and model, therefore amounting roughly to a 100-fold increase in performance; additional gains can be achieved by exploiting the larger number of simpler, less costly tasks.

The phylogenies we obtain are qualitatively better than standard, blind results in that results are clear, and noise and discordance are confined to individual subproblems, hence bounding their potential effect and improving overall quality and robustness. One such phylogeny can be seen in Fig. 3.

4 Discussion and Future Work

We have presented an efficient and effective supertree technique that effectively performs a reduction on s, the costliest dimension of the phylogeny reconstruction problem, through a divide and conquer approach where the penalties of the split and merge operations are negligible. As a result, computational load is substantially lowered, known properties are respected by construction, and the number of completely independent, distributable tasks is increased. These improvements afford more detailed treatment of individual problem instances and feasible resolution of continuously growing inputs.

We also deem it possible to develop reasonably accurate cost prediction, which, despite some measure of input dependence and irregularity, may be of great assistance in selecting methods (and models) in accord with available time and resources. Such estimates may also lead to better scheduling and load balancing in distributed environments.

It has been stressed that solutions are, as a matter of fact, qualitatively better. Moreover, we would like to know how prior knowledge may affect quantitative scores when pitted against atomic methods, as well as the effect and progression of individual algorithms and substitution models, and pinpoint the sources of error, including those that might cause incorrect group ascription.

Finally, considering that reliable phylogenetic knowledge is the source of all classification hierarchies, their appraisal merits further attention. The fit of an alignment and a guide tree, or lack thereof, may evidence a need for (possibly iterative) refinement of one or both parts. Likewise, the capacity to generate restricted datasets and treat them in finer detail could be used to support (sub)haplogroup identification while providing feedback and increasing guide tree resolution.

Acknowledgements. This work was supported in part by the Spanish Ministry of Science and Innovation (MICINN) under Action CAC-2007-52 and Project TIN2008-06582-C03-02. Roberto Blanco was supported by Grant AP2008-03447 from the Spanish Ministry of Education.

References

1. Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., Ronquist, F.: Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20, 407–415 (2004)
2. Blanco, R., Mayordomo, E.: ZARAMIT: a system for the evolutionary study of human mitochondrial DNA. In: Omatu, S., Rocha, M.P., Bravo, J., Fernández, F., Corchado, E., Bustillo, A., Corchado, J.M. (eds.) IWANN 2009. LNCS, vol. 5518, pp. 1139–1142. Springer, Heidelberg (2009)
3. Böckenhauer, H.J., Hromkovič, J., Královič, R., Mömkea, T., Rossmanith, P.: Reoptimization of Steiner trees: changing the terminal set. Theor. Comput. Sci. 410, 3428–3435 (2009)
4. Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P.: Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287 (2006)
5. Felsenstein, J.: PHYLIP – Phylogeny Inference Package (version 3.2). Cladistics 5, 164–166 (1989)
6. Holder, M., Lewis, P.O.: Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. 4, 275–284 (2003)
7. Richards, M.B., Macaulay, V.A., Bandelt, H.J., Sykes, B.C.: Phylogeography of mitochondrial DNA in western Europe. Ann. Hum. Genet. 62, 241–260 (1998)
8. Ronquist, F., Huelsenbeck, J.P.: MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574 (2003)
9. Ruiz-Pesini, E., Lott, M.T., Procaccio, V., Poole, J.C., Brandon, M.C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P., Wallace, D.C.: An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 35, D823–D828 (2007)
10. Sanderson, M.J., Purvis, A., Henze, C.: Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13, 105–109 (1998)
11. Steel, M., Dress, A.W.M., Böcker, S.: Simple but fundamental limitations on supertree and consensus tree methods. Syst. Biol. 49, 363–368 (2000)
12. Torroni, A., Achilli, A., Macaulay, V., Richards, M., Bandelt, H.J.: Harvesting the fruit of the human mtDNA tree. Trends Genet. 22, 339–345 (2006)
13. Wallace, D.C.: A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. Annu. Rev. Genet. 39, 359–407 (2005)