ZARAMIT: a system for the evolutionary study of human mitochondrial DNA *

Roberto Blanco and Elvira Mayordomo

Departamento de Informática e Ingeniería de Sistemas
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza. María de Luna 1, 50018 Zaragoza, Spain
{robertob, elvira} at unizar.es

Abstract. ZARAMIT is an information system capable of fully automated phylogeny reconstruction. Methods have been tailored to mitochondrial DNA sequences, with focus on subproblem partitioning. We have built exhaustive human mitochondrial phylogenies (approximately 5500 sequences) and detected problems in existing haplogroup hierarchies through data-driven classification.
Information on the project can be found on zaramit.org.

1 The case for mitochondrial DNA

Mitochondria, organelles present in most eukaryotic cells, are responsible for the generation of most of the cell's chemical energy. They are also remarkable for possessing their own, separate genome, which coexists with nuclear DNA and is inherited independently.

Further, mitochondrial DNA (mtDNA) has several features which make it an ideal candidate for conducting evolutionary studies. Firstly, it is small in mammals (15000 to 17000 base pairs) and encodes a homogeneous set of genes with little variation between species. Secondly, it exists within a very reactive environment where ROS are common: this provokes high mutation rates, approximately an order of magnitude above those of nuclear DNA. Thirdly, it displays matrilineal inheritance, which coupled with the virtual absence of recombination results in a pure evolutionary marker.

These properties make mtDNA suitable for studying evolutionary relations between closely related organisms due to its comparatively high resolution. Despite the high proportion of changes between individuals, their absolute number is small, owing to the short length of these sequences. This, in turn, permits a compact expression of mtDNA sequences as differences from a canonical reference sequence [1].

We are especially interested in the reconstruction of exhaustive human mitochondrial phylogenies which may let us spot potentially deleterious mutations. These are among the most common causes of rare genetic diseases, such as LHON

* This research was supported in part by Projects PM063/2007 of the Government of Aragon and TIN2008-06582-C03-02 of the Spanish Government's MICINN

and MELAS. Such harmful phenotypes are quickly discarded through natural selection; thus, they are only expected around terminal branches [2].

Historically, the number of publicly available sequences has increased exponentially, which allows comprehensive studies in population genetics but, at the same time, represents a difficult computational challenge.

2 Methods for phylogeny reconstruction

Computational phylogenetics [3,4] concerns itself with the algorithms and methods used to infer phylogenies: relational hierarchies between a set of data, usually biological sequences, which are most often expressed as phylogenetic trees.

Ideally, we desire to recover the true evolutionary relations between the inputs, referred to as taxa. However, evolution is largely an unknown process that must be approximated by macroscopic, statistical models. Furthermore, in order to assess robustness and quality of the solution, extensive data sampling is required, effectively resulting in a large number of more or less related problems.

Obtaining an optimal solution to this problem (a tree which maximizes an optimality criterion subject to a certain model of evolution) is generally known to be an NP-complete problem [5,6,7]; only for very particular, mostly unrealistic models, are polynomial, exact algorithms known. On the other hand, the optimal tree need neither be unique nor "real". In practice, we settle for suboptimal, biologically significant trees which are likely to reflect reality to a large extent.

Desirable properties of methods include statistical consistency and tolerance to mild violations of their evolutionary hypotheses or restrictions. Execution time is also a concern for large datasets, which favor fast yet consistent heuristics like neighbor joining instead of sophisticated likelihood models.

3 Exhaustive human mitochondrial phylogenies

We are aware of two different efforts to construct mitochondrial phylogenies comprising all available human mtDNA sequences. Firstly, there is MITOMAP [8], which can be considered the main reference database for the mitochondrial genome. Secondly, PhyloTree.org [9], a recently started project aiming to provide periodically updated phylogenies, which emphasizes haplogroup classification (the search for mutation patterns associated to genetic populations).

Both approaches suffer when confronted with the exponential growth experimented by sequence databases. Reconstruction methods are not designed for split or incremental execution, which, though conceivable, has yet to be explored. Separate haplogroup processing may decouple dependencies to some extent; however, the mitochondrial haplogroup hierarchy is subject to frequent changes, and in turn relies on phylogenies for its definition. Moreover, manually annotated phylogenies pose further maintenance challenges.

Instead, we favor an automated iterative approach, where phylogenies are built with minimal, initial intervention. The MITOMAP method of manual construction has reportedly proven unworkable and similar procedures suffer from

a serious lack of quality evaluation. Besides basic tree topology, we also wish to infer unknown ancestral sequences as labels of internal tree nodes, since these are needed to analyze purifying selection across phylogenies.

4 Current results

Over the past year, we have developed an information system, ZARAMIT [10], capable of fully automated phylogeny reconstruction. It manages every step of the process, from database retrieval and synchronization to multiple sequence alignment and construction of labeled, statistically supported trees.

Particular care has been put into subproblem definition and incremental construction, where applicable. Any work can be defined as a sequence of tasks: typically, sequence alignment and tree reconstruction methods. Many of these processes are relatively easy, but very time-consuming and often repetitive; therefore, computation stages are interleaved with storage stages, which can save much time in the long term.

Aligned databases have been created using Clustal incremental alignments. To accommodate the peculiarities of mitochondrial DNA and the huge size of experiments, we have combined sophisticated substitution models with neighbor joining: a very fast distance method with provable good performance. We have selected PHYLIP as the initial phylogenetic engine due to its renown, thorough implementation of most classical methods and source code availability. Quality of the solution has been assessed via bootstrap, resulting in 100–1000 fully independent trees built in parallel with the help of a Condor cluster.

To date, we have built such phylogenies for all currently available (in GenBank) human mtDNA sequences, together with chimpanzee outgroups: as of the time of this writing, close to 5500 unique sequences. Incremental tree construction has been first approached by means of automatic, data-driven haplogroup classification, which has allowed us to identify several problems with the simple MITOMAP haplogroup hierarchy, used as a basic tree skeleton.

Finally, we have considered alternative distance models based on compression and pure information theory concepts [11,12,13], as opposed to biologically-dependent models. Because sequences are very close to each other in absolute terms, it remains to be seen how differences can be emphasized to compensate raw similarity, a complication which may arise in all distance methods.

5 Work in progress

In order to further the goals of the system, effort must be made to extend process autonomy. Whereas some tasks (tree reconstruction) should be triggered by user request, most intermediate results can be updated bottom-up, without human intervention, given adequate configuration schemes. Automated search for potentially deleterious mutations from currently available, fully labeled trees, is another of our chief objectives, along with integration of alignment positions with gaps. The latter would allow the joint study of coding and control regions,

which is currently impossible, as there is a significant number of entries whose control regions have not been sequenced.

We also want to research incremental tree construction, considering both supertree methods guided by haplogroup hierarchies, and hybrid direct-incremental methods. Other related problems include model selection and evaluation of convergence and robustness applying not only bootstrap, but also Bayesian inference and phylogenetic networks.

Lastly, boosting efficiency is one of the main concerns when faced with such large datasets as ours. Parallelism and distribution are obvious choices, though much work remains to be done. Algorithm engineering [14] has achieved some outstanding results speeding up existing methods. Fixed-parameter complexity [15] tries to constrain parameter values to find subproblems where efficient solutions exist, and could be of great use to avoid the NP barrier. These are all areas of interest, together with the development of specialized methods for the recovery of extremely large phylogenies.

References

1. Anderson, S., et al.: Sequence and organization of the human mitochondrial genome. Nature **290** (1981) 457–465
2. Ruiz-Pesini, E., et al.: Effects of purifying and adaptive selection on regional variation in human mtDNA. Science **303** (2004) 223–226
3. Felsenstein, J.: Inferring Phylogenies. Sinauer (2003)
4. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press (2000)
5. Foulds, L.R., Graham, R.L.: The Steiner problem in phylogeny is NP-complete. Advances in Applied Mathematics **3** (1982) 43–49
6. Day, W.H.E.: Computational complexity of inferring phylogenies from dissimilarity matrices. Bulletin of Mathematical Biology **49** (1987) 461–467
7. Chor, B., Tuller, T.: Maximum likelihood of evolutionary trees: hardness and approximation. Bioinformatics **21** (2005) i97–i106
8. Ruiz-Pesini, E., et al.: An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Research **35** (2007) D823–D828
9. van Oven, M., Kayser, M.: Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation **29** (2008) E386–E394
10. Blanco, R.: Definición y prototipo de herramienta de análisis filogenético para el ADN mitocondrial humano. Master's thesis, Centro Politécnico Superior (Universidad de Zaragoza) (2008)
11. Li, M., et al.: An information based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics **17** (2001) 149–154
12. Li, M., et al.: The similarity metric. IEEE Transactions on Information Theory **50** (2004) 3250–3264
13. Urcola, P.: Algoritmos de compresión para secuencias biológicas y su aplicación en árboles filogénicos construidos a partir de ADN mitocondrial. Master's thesis, Centro Politécnico Superior (Universidad de Zaragoza) (2006)
14. Moret, B.M.E., et al.: High-performance algorithm engineering for computational phylogenetics. The Journal of Supercomputing **22** (2002) 99–111
15. Gramm, J., et al.: Fixed-parameter algorithms in phylogenetics. The Computer Journal **51** (2008) 79–101