On the analysis of data derived from mitochondrial DNA distance matrices: Kolmogorov and a traveling salesman give their opinion.

Pablo Moscato, Luciana Buriol, and Carlos Cotta

Address for correspondence: ccottap@lcc.uma.es

July 10, 2002

Abstract

We aim at provoking new ideas for computational tools based on the powerful algorithms currently available for the TSP. They can help to order DNA sequences and can be viewed as complementary tools to the existing phylogenetic methods, uncovering interesting correlations not made explicit in most tree layouts. We study the problem of ordering the sequences such that they minimize the total intersequence distances. The distances are based on a measure that approximates Kolmogorov complexity. The method is fully automatizable, and the results are very encouraging.

1 Introduction

The task of dealing with the large-scale combinatorial problems arising in genomics is undoubtedly one of the greatest challenges to be addressed by scientists. Pairwise genome comparison and the construction of evolutionary trees are two such challenging tasks, due to the computational complexity and the sheer amount of data to process. While phylogenetic trees can undoubtedly help to order the data, the existence of "lateral gene transfer" might not be dismissed in future analysis and pose new challenges.

This paper is written with the aim of providing complementary computational tools. We mean they are complementary since, for instance, they can be used to display particular correlations in the data are seldom taken into account in the final layout of evolutionary trees. On the other hand, the methods can be used as a preprocessing step when dealing with large number of taxa. We want to attract the attention of the biological community, that there has been much progress in solving one particular combinatorial problem, the traveling salesman, and that the current public domain codes are extremely fast and can well handle large-scale ordering problems, providing useful heuristics to cluster data derived from distance measures.

2 A distance-measure based on mutual algorithmic information

We have decided to use a definition of distance between pairs of DNA sequences based on conditional Kolmogorov complexity. This is a novel approach that has recently been proposed in [2] and [7], following the success of the application of Kolmogorov complexity in other areas.

Given two sequences x and y, belonging to a generic alphabet Σ, the conditional Kolmogorov complexity, noted $K(x|y)$, is also known as the *algorithmic entropy* of x given y. $K(x|y)$ is defined to be the length of the shortest program causing a universal computer to output x given input y. We will note as $K(x)$, for any given sequence x, the conditional Kolmogorov complexity of x when we give as input an empty string ϵ, i.e., $K(x) = K(x|\epsilon)$. The concatenation of two strings x and y is denoted as xy. The value of $K(x) - K(x|y)$ is understood as the amount of information that the y sequence "knows" about sequence x, and $K(x|y)$ can be interpreted as the amount of randomness of sequence x given as input sequence y.

The expression $K(x) - K(x|y)$ is known as the *mutual algorithmic information*, and it is symmetrical within an additive logarithmic factor. Unfortunately, it does not satisfy the triangle inequality. For this reason, we will use the expression $d(x, y) = 1 - (K(x) - K(x|y))/K(xy)$ as distance measure following [3]. It can be shown that this distance measure satisfies the triangle inequality; furthermore, it has a good normalization factor $(K(xy))$ allowing to take into account the length of the sequences, as well as the situation in which a sequence is much shorter that the other.

3 An optimal traveling salesman tour in sequence space

Given a distance matrix $\{d_{ij} = d(x_i, x_j)\}$ between a set of sequences $S = \{x_1, \ldots, x_{|S|}\}$, our aim is to find an optimal path among all the sequences in the set, without repeating a visit to a sequence; thus, the objective is to find the minimum *Hamiltonian* path. This is an NP-hard problem, even when the starting and/or finishing sequence is selected *a priori*. This NP-hardness result notwithstanding, probably optimal solutions can be efficiently found using memetic algorithms [8].

We have tackled this *minimum Hamiltonian path problem* by reducing it to the TSP. We start by defining two auxiliary sequences x_0 and $x_{|S|+1}$ located "at infinity" in relation to the set S. This means that $d(x_0, x_0) = d(x_{|S|+1}, x_{|S|+1}) = d(x_0, x_{|S|+1}) = d(x_{|S|+1}, x_0) = 0$ and $d(x_0, x) = d(x, x_0) = d(x_{|S|+1}, x) = d(x, x_{|S|+1}) = M$, where M is an arbitrarily large number. We now run our MA for the TSP; any optimal tour T that has an edge connecting sequences x_0 and $x_{|S|+1}$ can be interpreted as a Hamiltonian path P that starts from the sequence which is consecutive of x_0 in T (which is not $x_{|S|+1}$) and ends in the sequence next to $x_{|S|+1}$ (which is not x_0). The length of T $(L(T))$ and the length of the hamiltonian path P $(L(P))$ are thus related by $L(T) - 2M = L(P)$. Since the value of $2M$ is a constant, minimizing $L(T)$ is equivalent to minimizing $L(P)$. Moreover, any tour that does not include the edge $(x_0, x_{|S|+1})$ includes four edges of weight M. Let

T_{4M} denote such a tour: if $M \geq |S| \max_{ij}\{d_{ij}\}$, then any such a tour is larger than any tour that contains the edge $(x_0, x_{|S|+1})$.

The advantage of using this transformation is that we can apply the well-developed traveling salesman codes to find such a path through sequence space. We have used $M = |S| \max_{ij}\{d_{ij}\}$ in our transformations to use our memetic algorithm for the TSP to search for optimal solutions. We remark that after we have found the solutions with our memetic algorithm code, we also send a request to Prof. William Cook from Princeton University, who has kindly run his TSP program (Concorde) on the same distance matrices we have generated. As a consequence we can safely affirm that the solutions that we have found are indeed optimal. Concorde is the fastest exact algorithm for the TSP and it is available on the public domain [1].

4 Application: whole mitochondrial genome phylogeny

To put our proposed methodology to a test, we decided to follow a study in [7] on the construction of whole phylogenies from complete DNA mitochondrial genomes. The experiments have been done using distance matrices computed from the coding regions of mitochondrial DNA genomes of several species of mammals. Due to space constraints we only report results for a 34-taxa instance. This instance contains the "phylogenetically controversial" Guinea pig, since its proper position is the subject of hot debates in systematic biology [6, 4, 1] [11, 10]. For this set, Li *et al.* produced a distance matrix, and analyzed it with the neighbor-joining and hypercleaning programs mentioned before. This gives as a result two different phylogenies, and they have chosen to display in Fig. 2 of [7] the consensus. It agrees, up to a certain extent, with the overall structure of the phylogeny presented in [9].

We have created the following heuristic to find the first (x_a) and final (x_b) pair of sequences. It adapts well to the Kolmogorov based definition of distance using here. We select as x_a the species that has the largest average distance to all other sequences of the set. Then we compute the set of maximally distant sequences from x_a. This set can have cardinality higher than one, so from this set we also select the one that has the largest average distance to all other sequences in the set.

Based on the result of this heuristic, the initial sequence (x_a) is the Platypus and the final sequence (x_b) is the Baboon. The result is shown in Fig. 1. The optimal sequence was found in just 0.5 seconds in a PIII 300MHz computer using Java (JDK 1.3). We note several things: the Guinea Pig is adjacent to the Orangutan; the Dog is the next one in the order, so the Guinea Pig groups rather far (regarding the linear relative order of the path) from both murid and nonmurid rodents, as it was also found in [7] from the phylogenetic trees created by the hypercleaning algorithm. The Dog, Harbor Seal, the Grey Seal, and the Cat appear in that order, being a carnivore outgroup of the ferungulates and giving additional support to the claims made in [5]. The Pig acts as a kind of "stepping stone" after the two rhinos and the two whales, but note that the phylogenetic tree methods of [7] have managed to group it with the perisodactyls rather than with the cetartiodactyls. Its

[1] Available at `http://www.math.princeton.edu/tsp/concorde.html`

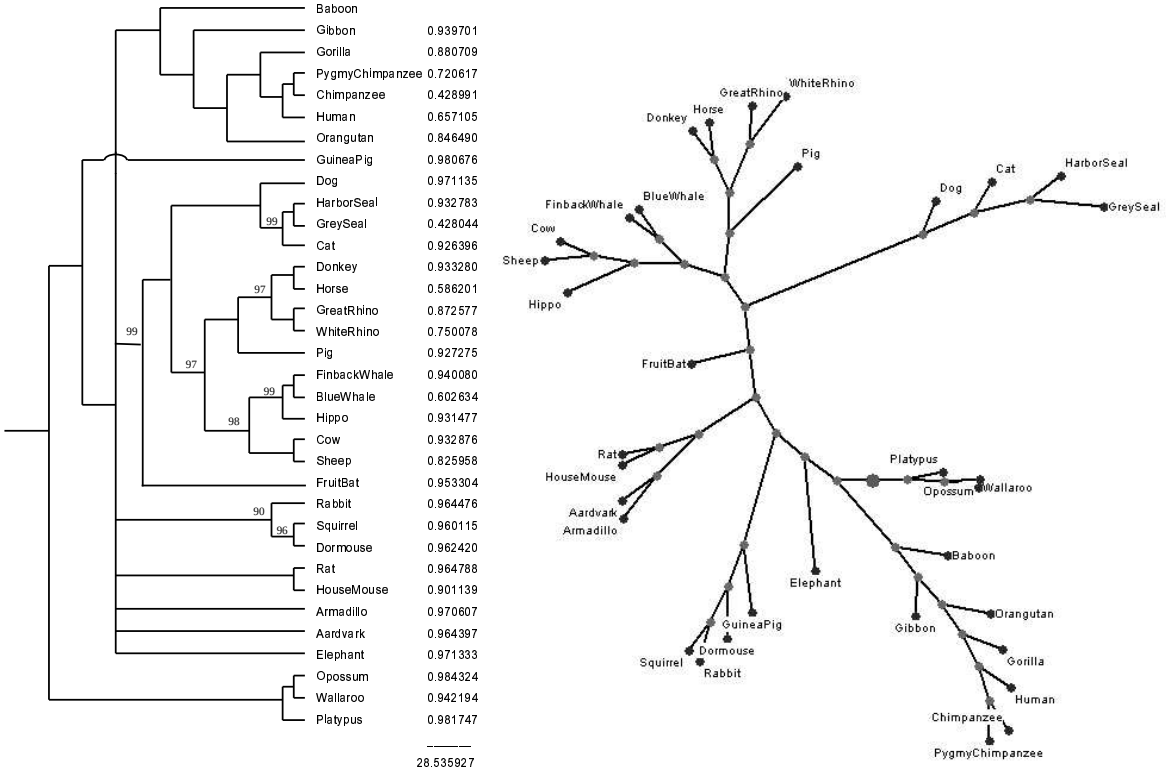

Figure 1: (Left) Evolutionary consensus tree of the neighbor-joining and hyperclearning algorithms for 34 mammals' species. The sequences are rearranged following an optimal Hamiltonian path starting from Platypus and ending with Baboon of total length 28.535927. (Right) Optimal ultrametric tree for this instance, computed by a Branch-and-Bound algorithm.

"central" position in our path order within the ferungulates appears to be more intuitive than the layout of Fig. 2 in [7] where it appears between the Donkey and the Fruit bat.

5 Conclusions

The main aim of the paper has been to show how currently available high-performance computer codes for a problem of high computational complexity, the traveling salesman problem, can be used to provide new methodologies for comparative genomic researchers. We stress the fact that the approach is complementary to other methods, and does not substitute phylogenetic studies.

An important aspect to highlight of the proposed methodology is is that it does not require gene identification and no additional domain knowledge. In addition to the fact that is based on a mathematically well-founded theory, the method allows it to be totally automatizable, running having as input a set of sequences. A current word of caution should be raised since a possible direction of improvement would be to develop a better conditional entropy estimator of DNA sequences (it is known that distances between highly divergent sequences tend to become similar to each other by the estimator proposed in [7]).

It would be also an interesting research direction to study if the optimal solution provided by the memetic algorithm can be used to "seed" a heuristic or exact algorithm for optimal phylogenetic tree construction. If so, this tandem of methods may allow the solution of larger instances of these problems.

References

[1] Y. Cao, N. Okada, and M. Hasegawa. Phylogenetic position of guinea pigs revisited. *Mol. Biol. Evol.*, 14:461–464, 1997.

[2] X. Chen, S. Kwong, and M. Li. A compression algorithm for DNA sequences and its applications in genome comparisons. *Genome Informatics*, 10:51–61, 1999.

[3] X. Chen, S. Kwong, and M. Li. A compression algorithm for DNA sequences based on approximate matching. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB), Tokyo, Japan*, page 107. Association for Computing Machinery, 2000.

[4] A.M. D'Erchia, C. Gissi, G. Pesole, C. Saccone, and U. Arnason. The guinea pig is not a rodent. *Nature*, 381:597–599, 1996.

[5] D. Graur, M. Gouy, and L. Duret. Evolutionary afinities of the order Perissodactyla and the phylogenetic status of the superordinal taxa Ungulata and Altungulata. *Mol. Phylogenet. Evol.*, 7:195–200, 1997.

[6] D. Graur, W.A. Hide, and W.-H. Li. Is the guinea pig a rodent ? *Nature*, 351:649–652, 1991.

[7] M. Li, J.H. Badger, C. Xin, S. Kwong, P. Kearney, and H. Zhang H. An information based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17:149–154, 2001.

[8] P. Moscato. Memetic algorithms: A short introduction. In D. Corne, M. Dorigo, and F. Glover, editors, *New Ideas in Optimization*, pages 219–234. McGraw-Hill, Maidenhead, Berkshire, England, UK, 1999.

[9] A. Reyes, C. Gissi, G. Pesole, F.M. Catze, and C. Saccone. Where do rodents fit ? evidence from the complete mitochondrial genome of *sciurus vulgaris*. *Mol. Biol. Evol.*, 17:979–983, 2000.

[10] A. Reyes, G. Pesole, and C. Saccone. Complete mitochondrial DNA sequence of the fat dormouse, Glis glis: further evidence of rodent paraphyly. *Mol. Biol. Evol.*, 15:499–505, 1998.

[11] J. Sullivan and D.L. Swofford. Are guinea pigs rodents ? the importance of adequate models in molecular phylogenetics. *J. Mammal. Evol.*, 4:77–86, 1997.