EVOLUTIONARY TREE RECONSTRUCTION AND TRAVELING SALESMAN PROBLEM: A POWERFUL ALGORITHM FOR SHAGGY TREES

***Korostishevsky M.*[1]*, *Burd A.*[2], *Mester D.*[2], *Bonne'-Tamir B.*[1], *Nevo E.*[2], *Korol A.*[2]**

[1] Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel 69978; [2] Institute of Evolution, University of Haifa, Haifa, Israel 31905
* Corresponding author: e-mail: korost@post.tau.ac.il

Keywords: *evolution, genetic trees, computer analysis*

Summary

Motivation: The correct Evolutionary Tree Reconstruction (ETR) based on the current genetic data is considered as an NP-complete problem. The number of "possible" trees rapidly grows with the number of "leaves". The full ETR model includes the number of ancestor vertices and the number of mutation events on the origin tree. Efficient algorithms are needed to meet the challenges of current molecular evolution studies in order to allow simultaneous treatment of hundreds and thousands of individual genotypes.

Results: We present a new ETR approach based both on the reduction the ETR problem to Traveling Salesman Problem (TSP) and on the minimization of the ETR model using Guided Evolution Strategy algorithm. The robustness of the model is defined by simulation experiments. The duration time on an ordinary computer, Pentium-4, is a few seconds for several hundreds of leaves.

Availability: http://study.haifa.ac.il/~aburd/genetic.html

Introduction

In any evolutionary process, speciation events cause a new species to split off from an existing one, thus creating the diversity of life forms we know today. A key issue in evolutionary biology is to reconstruct the history of these speciation events. An evolutionary tree is a rooted tree, where each internal vertex has at least two descendent vertices and the final vertices are labeled with distinct symbols representing recent species. The goal is: given the properties of the recent species, reconstruct what the tree is. Much of the current interest in evolutionary trees derives from the increasing availability of DNA sequence data. We consider here situations where the data represent non-recombining DNA sequences, such as DNA of Y-chromosome or mitochondrial DNA, where the origin for all sites of the given sequence is the same and the speciation events are defined only by the mutation process (Bonne-'Tamir *et al*., 2003).

The evolutionary tree is an oriented graph. A set of DNA sequences {s} defines the vertices of a tree. Length of an edge is defined by the number of generations between vertices. Equal number of generations is assumed from the root vertex to any final vertex. To a tree which contains n vertices, we shall number vertex with natural numbers 1, ..., n. To each vertex i of the tree, we define two corresponding numbers a_i and g_i, where a_i is the number of the proximate ancestral vertex ($a_i < i$) and g_i is the age of vertex i, i.e. number of generations from this vertex up to the contemporary generation ($g_i \geq 0$ and $g_i \leq g_j$, if $j = a_i$ and $i > 1$). For the root vertex $a_1 = 0$, for any final vertex (leaf) i, $g_i = 0$. We shall designate the sets of numbers a_i and g_i as {a} and {g}.

Let define μ(i/j) as the probability of a mutation from a nucleotide j to a nucleotide i per generation (Majewski, Ott, 2003). If μ(i/j) are given, a sequence of the root vertex uniquely determines the probability of sequences of all remaining vertices. The random mutation process causes the genetic distance between the leaves that is not certainly proportional to their distance in generations.

Because of that, a reconstruction of the origin tree based on similarity and difference between the leaves is a NP complete problem.

The full ETR model includes the number of ancestor vertices and the number of mutation events on the origin tree. The number of trees T_N rapidly increases with the number of leaves, N. Namely:

$$T_{N=}[2(N\text{-}1)]!/[2^N(N\text{-}1)!].$$

For $N = 50$ this number approaches $3*10^{76}$.

Method, algorithm, implementation, and scenarios

We present a new ETR approach based both on the reduction the ETR problem to the known Traveling Salesman Problem (TSP) and on the minimization of the ETR model. The TSP solution is achieved by rearranging the leaf order to minimize the cycle length over the leaves (Korostensky, Gonnet, 2000). A powerful algorithm of TSP solution, utilizing the strength of the Guided Evolution Strategies (GES) method (Mester, Bräysy, 2004), was developed and adapted to the ETR problem. The TSP solution reduces the set of acceptable binary trees to at least $N!/2$. For the above example with $N = 50$, the number of trees decreases more than 10^{12} times. A single tree of the reduced set is selected by using the Average Linkage Clustering (ALC) method. The final tree, hereafter referred to as MBK tree, is achieved by combining neighboring vertices of the binary tree to minimize the number of parameters defined the ETR model (Korostishevsky *et al*., 2001). The robustness and effectiveness of the model was established by simulation experiments. The resulted trees were compared with the trees obtained by the UPGMA algorithm (PHYLIP, 1995).

- The proposed algorithm includes the following functions:
- Simulating a tree of the given complexity.
- Simulating leaves of the given tree and the given root sequence.
- Evaluating ETR using the pairwise distance matrix between leaves.
- Estimating the quality and robustness of the ETR solution.

The software was built using C# language and Graphic User Interface (GUI) of Microsoft® Visual Studio.NET (http://msdn.microsoft.com/netframework/technologyinfo/howtoget/). The duration time on an ordinary computer, Pentium-4, is a few seconds for some hundreds of leaves. The simulation results on shaggy trees, $N > 100$, illustrate the effectiveness of the method.

Figure 1 is the program's screenshot, followed by the description of the input stages and the output forms. The user interface was designed as a one-window form application with the control buttons on the left and the resulted trees on the right. Most frequent ETR scenarios include:

- The tree for haplotype simulations defined by user.
- The haplotypes loaded from user's file.
- The haplotypes simulated on a random tree.

Simulation results and Discussion

Simulations were done on three different trees (with 19, 54 and 110 leaves).
Evaluation parameters: TH-tree height (1000, 10000), HL-haplotype length (50,100), MR-mutation rates (0.001, 0.0005). 1000 simulations were done for each of the 3*2*2*2=24 parameter combinations. The results are presented in Table and Figure 2.

Quality of the resulted trees was estimated by relative tree length, as a ratio reconstructed tree length / original tree length. The tree length is defined as one half of the minimal cyclic way through the leaves. This minimum way is achieved on original tree if the distances between leaves are proportional to age of MRCA. In our case, the distances between leaves are randomly deviating among 1000 simulations done for a given tree with fixed parameter values.

It can be easily seen that with the growing complexity of original tree our algorithm displays a growing advantage over the standard UPGMA algorithm.

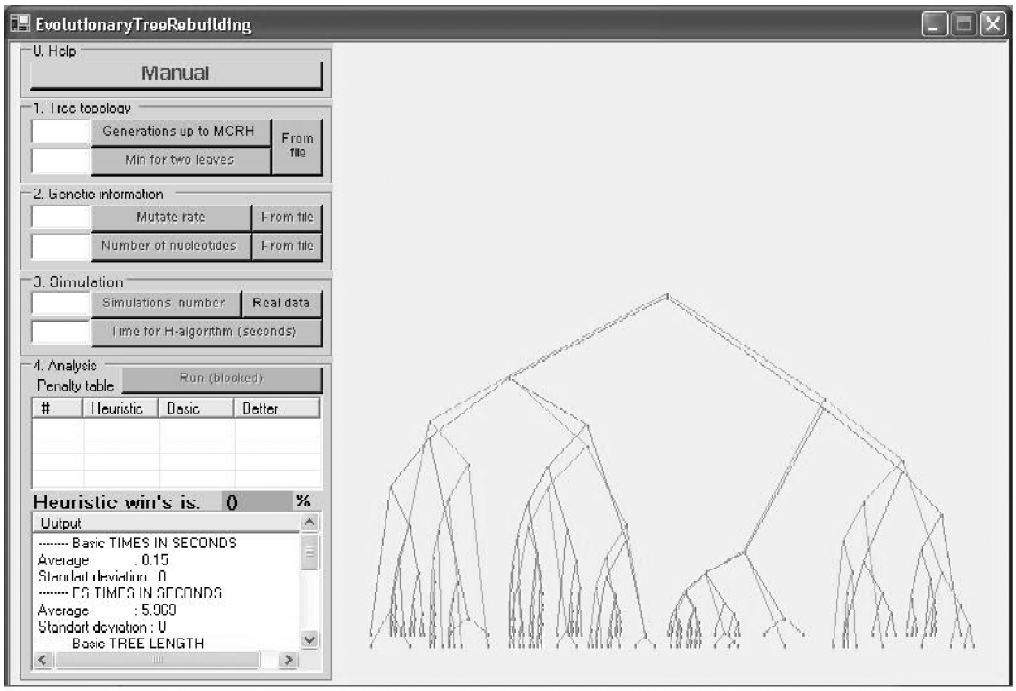

Fig. 1. User interface.

Table. Simulation results for a tree with 110 leaves

Simulation parameters				Simulated tree		UPGMA		MBK tree	
Leaf number	Tree height	Mutation rate	Haplotype length	Tree length	STDEV	Tree length	STDEV	Tree length	STDEV
110	10000	0.0005	100	5573	1.9229	5524	1.8243	5259	1.9189
110	10000	0.0005	50.0	2786	1.2790	2730	1.2265	2538	1.2802
110	10000	0.0010	100	6947	1.7577	6802	1.5834	6503	1.6927
110	10000	0.0010	50.0	3472	1.3079	3313	1.0249	3104	2.6594
110	1000	0.0005	100	1326	1.5285	1336	1.5410	1233	1.5369
110	1000	0.0005	50.0	689.2	0.9809	691.2	0.9954	616.7	0.9848
110	1000	0.0010	100	2244	1.7744	2247	1.7687	2100	1.8093
110	1000	0.0010	50.0	1125	1.2490	1123	1.2653	1015	1.2421

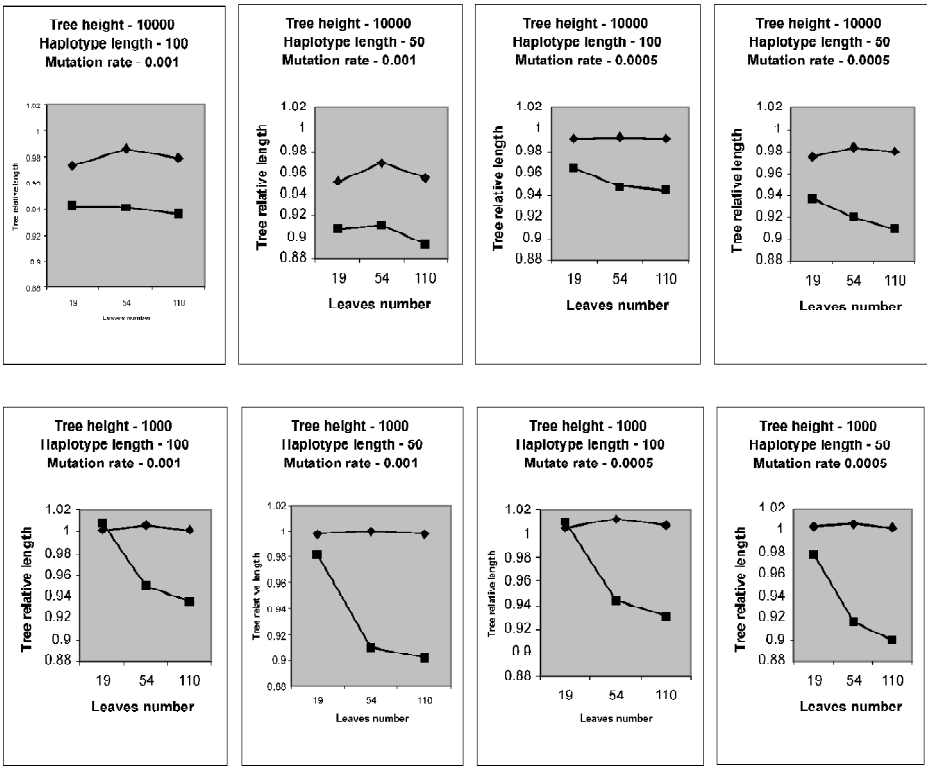

Fig. 2. Relative length of two reconstructed trees as a function of leaf number (squares for MBK and diamonds for UPGMA).

References

Bonné-Tamir B., Korostishevsky M., Redd A.J., Pel-Or Y., Kaplan M.E., Hammer M.F. Maternal and paternal lineages of the Samaritan isolate: mutation rates and time to most recent common male ancestor // Annals of Human Genetics. 2003. V. 67. P. 153–164.

Korostensky C., Gonnet G.H. Using traveling salesman problem algorithms for evolutionary tree construction // Bioinformatics. 2000. V. 16. P. 619–627.

Korostishevsky M., Ginzburg E., Bonné-Tamir B. Mutation origin reconstruction based on adjacent haplotypes // The First Workshop on Information Technologies Application to Problem of Biodiversity and Dynamics of Ecosystem in North Eurasia. Novosibirsk. Russia. 2001. Abstract P219.

Majewski J., Ott J. Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms // Gene. 2003. V. 305. P. 167–173.

Mester D., Bräysy O. Active guided evolution strategies for large-scale vehicle routing problems with time windows // Computers and Operations Res. 2004. (on line), 1–22.

PHYLIP. Phylogeny Inference Package, release 3.57c. University Washington, USA 1995. (http://evolution.genetics.washington.edu/phylip.html).