PhD Thesis

Molecular phylogenetic reconstructions with a discrete mathematical method, the Boolean analysis

Eszter Ari

Dr. Éena Jakó, senior research fellow
supervisor

Dr. Eörs Szathmáry, professor
head of doctoral program

Dr. Anna Erdei, professor
head of doctoral school

2012

Theoretical and Evolutionary Biology doctoral program,
Biology doctoral school,
Department of Genetics,
Department of Plant Taxonomy, Ecology and Theoretical Biology,
Eötvös Loránd University, Budapest
INTRODUCTION

With molecular phylogenetic methods the phylogeny of organisms can be reconstructed according to molecular data, mostly DNA, RNA and protein sequences. The widely used, so called standard phylogenetic methods are the distance based Neighbor-Joining (NJ), and the character based Maximum Parsimony (MP), Maximum Likelihood (ML) and Bayesian analysis. Although the standard methods use different probability and stochastic models for tree reconstruction, they all generally treat sequences as collections of “independently and identically” distributed sites. In order to avoid computational difficulties, the standard phylogenetic methods assumed that all nucleotide or amino acid positions and changes along the sequence are independent. But according to the biological and chemical adequacy the sites of a sequence cannot be treated independently if we accept the fact that the most important information about the structure and function of a protein, a cell or an organism are coded by ordered sets of different monomers. Nevertheless, the standard phylogenetic methods are able to reconstruct acceptable species trees, if the “phylogenetic signal” in the sequences is satisfactory. But there are cases, when one may get controversial results by using different methods and/or different sequences. In such cases, an algorithm based on an alternative, non-statistical approach, that considers the order of nucleotides or amino acids if sequences can be useful for classification. This gives us an opportunity to verify and compare the results (congruence analysis). The discrete mathematical method and software, presented in the dissertation, are based on the Iterative Canonical Form of Boolean functions (ICF; Jakó, 1983) called Boolean analysis or BOOL-AN (Jakó et al, 2009). By using the ICF algorithm we can evaluate certain structural invariants which characterize the given sequences in a canonical way. According to the ICF invariants the sequences can be classified in reasonable time and with reasonable memory consuming.

OBJECTIVES

The aim of my investigations was to test the performance of the Boolean analysis for making reliable phylogenetic trees. If we want to use the novel BOOL-AN method to investigate sequences of species with unknown or controversial phylogeny, first we have to prove its adequacy and reliability for calculating phylogenetic trees.
Correspondingly, the aim of my investigations was to answer the following questions:

- Do the Boolean analysis and the standard phylogenetic methods produce the same or different results?
- How reliable are the BOOL-AN trees?
- How do the properties of sequences and the different parameter settings of BOOL-AN influence the results?

To answer the above questions I tested the BOOL-AN method by using:

- artificially evolved or simulated sequences, and
- mt-tRNS sequences of great apes.
- The BOOL-AN trees were compared to trees produced by standard phylogenetic methods.
- The topology of resulting tree in each case was compared with the guidetree of simulations and with the known species tree of great apes.
- I tested also the BOOL-AN method by using different parameter settings and distance formulas.

**METHODS**

**The BOOL-AN method**

The BOOL-AN method encodes sequences into discrete mathematical language, by transforming them to a system of binary strings, called *molecular codes* (Jakó, 2007; Jakó et al, 2009; Fig. 1. b).

The ICF is a decomposition algorithm, that by means of a series of iterative steps produces the canonical normal form of a Boolean function (Jakó, 1983). The information content of any sequence can be described rigorously by this canonical form, used as a generalized molecular descriptor (Jakó, 2007). In order to calculate the ICF for any Boolean function, it previously has to be ordered in an *n* dimensional space. The ICF transforms then the sequence to a certain form that explores the functionally important key positions of it (Fig. 1. c). From these key positions the original sequence can be reconstructed without loss of information. The molecular descriptors can be represented in form of ICF graphs as well (Fig. 1. e). The topology of ICF graphs shows the
characteristic features of the primary structure and the consequences of structural changes (mutations), respectively. Therefore the topology of the ICF graphs for evolutionarily more closely related species is more similar than in case of relatively distantly related species. This fact can be shown by using distance calculations between the corresponding ICF graphs as well (Jakó et al, 2009; Ari et al, 2012).

Fig. 1. Main steps of the BOOL-AN and visualization of the results

a: sequence alignment; b: encoding of the sequence information; c: ICF computation; d: distance calculation. Visualization: e: ICF graphs; f: tree construction (NJ); g: metric multidimensional scaling (PCoA). The original source of revised figure: Jakó et al, 2009.

To classify the sets of sequences we should calculate distances between the ICF results (Fig. 1. d). From the distance matrix trees can be yield by Neighbor-Joining method (Fig. 1. e), or pointmaps can be calculated by multidimensional scaling (g). For calculateing distance matrices some well known formulas can be applied, like Euclidean and Jaccard distances. In the present work there was used also the special merged ICF graph distance (IGD), that was developed directly for quantifying distances between the ICF graphs (Jakó et al, 2009; Jakó et al, 2012).

Testing the BOOL-AN method for phylogenetic reconstructions

To test the reliability of BOOL-AN and standard phylogenetic methods I used two different approaches. In the first test trees were reconstructed from artificially evolved or simulated DNA sequences (Ari et al, 2012). Then the result trees were compared by the guide trees of simulations. Four different approaches were used for sequence simulations. A hundred sequences were simulated by each set of model parameters.
Simulation 1: The goal of this test was to investigate the impact of the number of taxa on phylogenetic reconstruction. The number of taxa in the guide trees was changed (5, 10, 20, 50, 100 and 200) while other parameters of simulation were fixed.

Simulation 2: Here I would wish to explore the effect of different sequence lengths. For testing this, the lengths of simulated sequences were altered (30, 50, 80, 100, 200, 500, 1000 and 5000 nucleotides).

Simulation 3: This test is used to reveal the influence of different guide tree branch lengths. Therefore six trees were used for guide trees with different branch lengths and same topology.

Simulation 4: This test examines the effects of sequence diversities, i.e. various combinations of different nucleotide substitution parameters. Eight set of sequences were simulated with different substitution model parameters.

In the second test natural sequences were investigated, the mitochondrial tRNA (mt-tRNA) genes of great apes. The phylogeny of great apes seems to be cleared nowadays, so according to their sequences phylogenetic methods can be tested and compared (Ari et al, 2008; Ari et al, 2012).

To measure the reliability of each phylogenetic method Robinson-Foulds distances (RF) were calculated between result trees and the guide tree of simulation or the known tree of great apes. The average of branch support percentages of consensus trees made from trees of each applied methods were also used as a reliability values (Ari et al, 2012).

RESULTS
According to the two applied reliability values (RF distances and branch support values) the reliability of phylogenetic methods and applied parameters were distinguishable. I tested the BOOL-AN method with different parameters and distance formulas also, and the best results were got by using the merge ICF graph distance.

The averages of reliability values for all simulation tests and approaches together are shown in Table 1. BOOL-AN (with IGD) trees had the best reliability values in most cases, except when branch lengths of the guide tree (Simulation 3) or the diversity of sequences (Simulation 4) was altered. In these two tests, Maximum Parsimony was as good or
sometimes slightly better than BOOL-AN. In most cases Maximum Likelihood proved to be the least adequate.

*Table 1. The averages of reliability values of simulation trees*

<table>
<thead>
<tr>
<th>Main factor</th>
<th>Reliability values</th>
<th>BOOL-AN</th>
<th>MP</th>
<th>NJ</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Number of taxa</td>
<td>Average RF distances divided by No. of taxa</td>
<td>1.2429</td>
<td>1.2687</td>
<td>1.2779</td>
<td>2.2087</td>
</tr>
<tr>
<td>(2) Sequence length</td>
<td>Average branch support values</td>
<td>78.2325</td>
<td>77.165</td>
<td>71.6963</td>
<td>62.465</td>
</tr>
<tr>
<td></td>
<td>Average RF distances</td>
<td>3.0475</td>
<td>3.4288</td>
<td>3.9625</td>
<td>5.255</td>
</tr>
<tr>
<td>(3) Branch length of guide tree</td>
<td>Average branch support values</td>
<td>47.881</td>
<td>49.846</td>
<td>45.7381</td>
<td>38.119</td>
</tr>
<tr>
<td></td>
<td>Average RF distances</td>
<td>6.5733</td>
<td>6.6885</td>
<td>7.3133</td>
<td>8.43</td>
</tr>
<tr>
<td>(4) Sequence diversity</td>
<td>Average branch support values</td>
<td>73.3393</td>
<td>73.4857</td>
<td>44.1964</td>
<td>53.9643</td>
</tr>
<tr>
<td></td>
<td>Average RF distances</td>
<td>3.6675</td>
<td>3.6519</td>
<td>7.71</td>
<td>6.37</td>
</tr>
</tbody>
</table>

*Average branch support values*: the method is more reliable while this number is great (the maximum is 100). *Average RF distances*: the method is more reliable while this number is small. The best values highlighted with red, the worst with green.

Although the topology of trees made from each mt-tRNA specificity were not all the same as the known phylogeny of great apes, the BOOL-AN and the Bayesian consensus trees have the same topology as the widely accepted tree of great apes. The topology of Maximum Parsimony, Maximum Likelihood and Neighbor-Joining consensus trees differs more or less from the established cladogram of great apes. In contrast, the BOOL-AN and the Bayesian consensus trees have the same topology as the widely accepted one (*Fig. 2*).
CONCLUSIONS

In contrast to standard molecular phylogenetic methods the BOOL-AN method does not require predefined evolutionary hypotheses or substitution models. However the Boolean analysis cares about the positions of nucleotide (or amino acids) differences within genes, genoms (or proteins) of the investigated species. Therefore the BOOL-AN can use this additional information for better classification while counting a phylogenetic tree. Thereby with the BOOL-AN it is possible to classify quite short and very similar sequences (like tRNAs, mt-tRNAs, 5sRNAs) that do not have satisfactory phylogenetic signal for standard phylogenetic procedures.

Fig. 2. The widely accepted cladogram of great apes and mt-tRNA trees made by different phylogenetic methods

a: The widely accepted tree of great apes. Consensus trees based on trees obtained for each molecular phylogenetic method from the 22 mt-tRNA sequences of apes: b: BOOL-AN and Bayesian tree; c: Maximum Parsimony tree; d: Maximum Likelihood tree; e: distance based Neighbor-Joining tree. The original source of revised figure: Ari et al, 2012.
It has been demonstrated that the new discrete mathematical method, which is entirely different in its basic principles from the existing phylogenetic approaches, produced interpretable results in case of both simulated and natural sequences. BOOL-AN method produced particularly good results while analyzing relatively short sequences, like the very similar mitochondrial tRNA genes of great apes. What is important, the BOOL-AN software is much faster than the standard character based molecular phylogenetic procedures. Therefore it can be applied in a reasonable time for making analyses with thousands of taxa, like a Tree of Life project.
REFERENCES


