Poster C-42

Anâtaxis, a Simple Deterministic Phylogenetic Reconstruction Algorithm

Authors:
Bernhard Sonderegger (*Département d'informatique, Université de Genève, Swiss Institute of Bioinformatics*)
Gabriel Bittar (*Swiss Institute of Bioinformatics*)
Bastien Chopard (*Département d'informatique, Université de Genève, Swiss Institute of Bioinformatics*)

Short Abstract: We propose Anâtaxis, a simple phylogenetic tree reconstruction algorithm capable of analysing very large data-sets. Anâtaxis uses an intuitive two-step procedure to resolve the problems of variable rates of evolution and homoplasy. We show a biologically meaningful result and furthermore demonstrate the efficiency of each step using numerical methods.

Long Abstract:

Introduction

Phylogenetics is becoming an essential element in molecular and computational biology. Current trends in phylogenetics are leading towards ever more complex methods. Many of these methods are very computationally intensive and thus make use of heuristic search algorithms. A consequence of this is that the final result cannot be guaranteed to be the best tree. In this work, we propose a simple, deterministic algorithm called Anâtaxis for the rapid reconstruction of rooted phylogenetic tree topologies. Our method has the major advantage of allowing for heterogeneous rates of evolution and homoplasy, while still being fast enough to generate trees from data-sets of several hundred taxons in a few seconds on a standard desktop computer. Our aim is not to surpass model-based methods in theoretical quality, but to make the generation of very large trees feasible even in the presence of data with a high degree of homoplasy and highly heterogeneous rates of evolution.

Algorithm

The input for the Anâtaxis algorithm is an outgroup and a matrix of pairwise dissimilarities between taxons. An outgroup is a taxon or group of taxons known to be phyletically outside the ingroup of taxons being analysed. The algorithm consists of two major steps which divide the ingroup recursively to form a tree structure. These steps are the evolutionary normalisation of the dissimilarity matrix, and ingroup division.

The normalisation step directly addresses the problem of variable rates of evolution. In this step, information on the rates of evolution within the ingroup is obtained by using the outgroup as an external reference. The comparison of dissimilarities between the outgroup and the ingroup members gives an estimation of the relative speeds at which each ingroup member evolved from the closest common ancestor of both ingroup and outgroup. This information is used by the algorithm to correct the intra-ingroup dissimilarities. Since evolutionary rates from the common ancestor are the basis for this step, no knowledge of the internal tree structure of the ingroup is required. Initially, normalisation is performed on the

complete ingroup using an outgroup provided by the user. However, as the algorithm progresses and attempts to resolve lower branches of the tree, closer and more pertinent sub-outgroups are determined by the algorithm and used to calculate more finely corrected normalised matrices. No errors are accumulated at this level since each normalisation is performed on the original dissimilarities.

The second major step, ingroup division, deals with homoplasy-related errors. Homoplasy between two taxons will make their pairwise dissimilarity appear smaller. In this ingroup division step, all possible groups of three taxons from the ingroup are analysed separately. This means that if there is homoplasy between ingroup taxons i and j, these two will be analysed separately in the context of every other ingroup taxon. In this way, the algorithm takes advantage of the statistical effect: the likelihood that the error in the normalised dissimilarity D'i,j will cause a false interpretation in a large number of groups of three is small.

Validation

The effectiveness of each of these two steps in the Anâtaxis algorithm was demonstrated using numerical methods. To validate the normalisation procedure, sequence sets corresponding to ingroups with four taxons and single taxon outgroups were generated using a simple simulation of evolution. Simulation runs were performed for both symmetrical and asymmetrical topologies at various evolutionary distances. The ability to determine the correct tree topology from the normalised ingroup dissimilarities was compared to that of the well-known Neighbor-Joining algorithm.

The aim of the second set of numerical tests was to validate the ingroup division step. In order to avoid the need for accurate simulation of homoplasy and to isolate the division step from the normalisation step, a different approach was taken. Instead of simulating evolution directly, noise corresponding the type of error expected to arise from homoplasy was generated. The division step was then tested for its robustness to such noise.

The numerical tests showed that the normalisation step was able to resolve the correct tree structures across the full range of evolutionary distances. NJ on the other hand, was generally not very good at resolving asymmetrical trees and had some trouble resolving symmetrical trees with very short or very long evolutionary distances. For the validation of the ingroup division method, it was found that 100% of the ingroups were properly divided in virtually all conditions. Only the highest noise levels combined with the smallest ingroup sizes were able to reduce precision to 70%.

Biological example and comparison with other methods

Anâtaxis was compared to Neighbor-Joining and Tree-Puzzle (a maximum-likelihood method), using a large biological data-set comprising aligned chloroplast rps4 genes (chloroplastic 30~S ribosomal protein S4) from 163 species. It was found that dendrograms produced by Anâtaxis and Neighbor-Joining showed the same major branch divisions at the base of the tree, while the result from Tree-Puzzle had a large percentage of unresolved polychotomies. The main difference between the results of Neighbor-Joining and Anâtaxis was in the placement of the dicots and the monocots, with Anâtaxis giving a scenario closer to the view of modern plant systematics. Anâtaxis was found to run at speeds comparable to Neighbor-Joining when reconstructing trees with hundreds of taxons.

In conclusion, the Anâtaxis algorithm was described for reconstructing biologically

meaningful phylogenies. At the heart of the algorithm is a simple intuitive mechanism, which directly targets the problems of homoplasy and heterogeneous rates of evolution. With two sets of numerical tests we show that it holds promise as a simple phylogenetic tool capable of processing large data-sets or as a tool for seeding heuristic search algorithms of more time-consuming methods. The analysis of a biological data-set produced a meaningful tree, suggesting that the algorithm could potentially be used as a stand-alone tool.

Poster C-42

Phylogenetic inferences of existing Stramontita sp. (Mollusca) in Brazilian Northeastern Coast reveal unforeseen features of this subfamily

Authors:
Fátima de Cássia Evangelista de Oliveira (*NUGEN-UECE*)
Samara Cardoso da Silva (*NUGEN-UECE*)
Michel Toth Kamimura (*NUGEN-UECE*)
Daniel Pascoalino Pinheiro (*NUGEN-UECE*)
Sarah Ramos Medeiros (*NUGEN-UECE*)
Raimundo Bezerra da Costa (*NUGEN-UECE*)
Diana Magalhães de Oliveira (*NUGEN-UECE*)
Rodrigo Maggioni (*NUGEN-UECE*)

Short Abstract: Applying standard bioinformatics tools after directly sequencing DNA samples from Stramonitas spp from Brazilian Northeast Coast, we were able to build a phylogenetic tree that gives the first evidence about distance and similarities among these ancient animals in our region, as compared to reports in other parts of the world.

Long Abstract:
The molluscs are the most diverse phylum of animals next to arthropods. They have become adapted to living in almost all available habitats, being found in freshwater, in the deepest ocean trenches to the intertidal zone or on land, where they occupy a wide range of habitats. They are normally divided into 8 classes. The most important class of living molluscs is the Gastropoda, comprising more than 80% of all living mollusc species. Despite being such a large and conspicuous group, gastropods are surprisingly under-studied in evolutionary investigations. Many studies report the differences among species of gastropods based mainly on morphological characters. However, these studies were not capable to clarify many of the relationships in several groups, like Stramonita subfamily, one of the major great quantitative members in the Western Atlantic, in Ceara coast. In this study, we try to elucidate these relationships using molecular sequence data, which opened up a new way to classify organisms based on sequences rather than external features, supported by bioinformatics tools. This tools can provides us new opportunities to understand mollusc phylogenetic relationships and complement the extensive data obtained by functional, developmental, structural, and paleontological analyses. The samples were collected from Caponga/Ceará, Brazil, located in the Western Atlantic. All the samples were collected at Pirangi Estuary and individually stored. Genomic DNA was isolated from ventral foot using

the phenol/chloroform extraction method standart. PCR amplification was carried out and the products were sequeced to genomic analysis. Automated DNA sequencer used was ABI Prism 3100 (Applied Biossytems). It performs two kinds of analisis: sequencing analysis and fragment analisis. The analysis of the data was carried out with DNA Sequencing Analysis Software which separated a mixture of DNA fragments according to their lengths; it provided a profile of the separation and determined the order of the four deoxyribonucleotide bases. Sequencing analysis modules, created with DNA Sequencing Analysis software, provided the Auto Extractor with the parameters needed to analyze sequencing data. The Auto Extractor was used by the Data Colletion software to automatically extract and analyse the data after each run. Each peak in the electropherogram was represented to a single fragment. The position and shapes of the electropherogram peaks were used to determine one out f four possible bases. The analyzed data were stored as sample files on the hard drive. It was viewed with DNA Sequencing Analysis Software. Using these tools, it was possible to get the results of DNA sequencing to phylogenetic studies of the samples analysed. The phylogenetic analyses were carried out using standard bioinformatics tools (BLAST, Clustal and JPLOT). The sequences obtained for DNA sequencing were searched for homology against NCBI databases (through BLAST and its variants). Knowledge of molecular and evolutionary biology helped us to interpret the similarities and differences among the samples. Biological sequences showed complex patterns of similarity to one another, while comparisons between each of two species of gastropoda of the stramonita from northeast Brazilian were the main target in this survey. The Clustal series of programs are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequences and for preparing phylogenetic trees. The popularity of the programs depends on a number of factors, including not only the accuracy of the results, but also how robust, portable and user-friend they are. Clustal X provides an indication of the quality of an alignment by plotting a 'conservation score' for each column of the alignment. A high score indicates a well-conserved column; a low score indicates low conservation. The quality curve is drawn below the sequences / profiles, while NJplot is a tree drawing program able to draw any phylogenetic tree expressed in the Newick phylogenetic tree format (e.g., the format used by the PHYLIP package). NJplot is especially convenient for rooting the unrooted trees obtained from parsimony, distance or maximum likelihood tree-building methods. Phylogenetic trees constructed from the 16S rRNA sequences were used in this study. The sequences were aligned in Clustal X and the tree was constructed in molecular evolutionary genetics analysis (MEGA) using the neighbour joining method. Using these two tools after directly sequencing samples from the existing specimens of Stramonitas spp in the Brazilian Northeast Coast, we were able to build a tree of the remaining Stramonitas species. This phylogenetic tree gives the first evidence about distance and similarities among these ancient animals in our region, as compared to reports in other parts of the world.