Phyl-Ariane

Phylogenomics: Integrated algorithms and visualizations for analyzing the evolution of life

Domaines Emergents - 2008

AGENCE NATIONALE DE LA RECHERCHE ANR

WWW.LIRMM.FR/PHYLARIANE

10/12/09

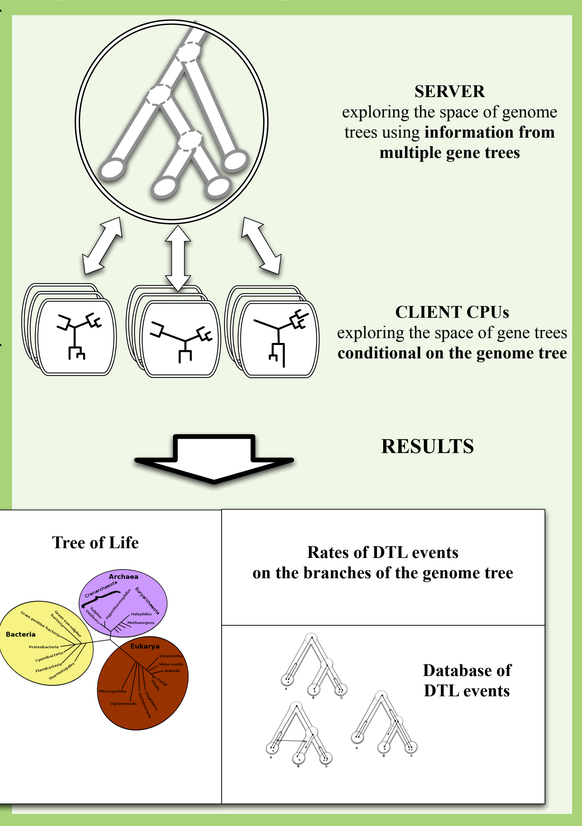
Objectives

The relentless march towards faster and cheaper sequencing technologies generates a wealth of genome sequence data that has a deep impact on many fields of biology and in particular on our picture of life evolution and organization. Despite this progress, many questions remain on the processes and patterns of genome evolution: phylogenies at many levels of the tree of life disagree, and there has been argument whether early evolution can be represented as a tree, or whether rampant horizontal gene transfers (HGT) have rendered vertical inheritance secondary. While comparative studies have established that gene duplication, horizontal transfer and loss have made a major contribution to the evolution of the genomes, a quantitative understanding of the extent and distribution of these processes is lacking.

The **Phyl-Ariane** project is aimed at developing methods able to use genome scale sequence data across a broad phylogenetic range to extract information on the historic pattern of genome evolution and the processes that have shaped it. This requires developing models of genome evolution as well as efficient algorithms to allow the reconstruction of the complex history of genomes. This information will be made publicly available in the form of a reconstructed tree of life and a database of DTL (*Duplication, Transfer and Loss*) events. Both of these resources have the potential to benefit a wide range of biological research by providing timely and relevant information on the pattern and process of genome evolution.

Simultaneous inference of the pattern and process of genome evolution

The problems of achieving a quantitative understanding of the processes that shape genomes and reconstructing the historic pattern of genomic descent are intertwined. The phylogenetic histories of individual genes in a given genome are not identical, but reflect individual genic histories of duplications, transfers and losses (*DTL*). These processes imply that, over sufficient phylogenetic distance, the history of genomic descent will be incongruent with individual genic histories for the majority of genes. To extract as much evolutionary information as possible from genomic sequences, it is necessary to reconcile the phylogenetic histories of individual genes (the forest of gene trees) with the phylogenetic history of the genomes in which they have resided (the tree of genomes).



We believe that in order to decipher the evolutionary information contained in the genomes of Bacteria, Archaea and Eukaryotes a statistical model of genome evolution is needed that, besides explicitly considering DTL processes, relates the prevailing trend of genomic descent to the histories of individual genes composing the genomes. A model is necessary that can infer from a set of homologous genes, represented as a forest of gene trees, a tree of genomes together with rates of DTL. We are developing such a statistical model, the outcome of which will be a statistically well defined reconstruction of the tree of genomes across the widest phylogenetic possible range, together with a database of DTL events. This computationally intensive approach benefits from progress in computer science, e.g. by relying on fixed-parameterized complexity.

Phylogenomic databases

Two databases conceived within the framework of the project contain genes from completely sequenced organisms. This feature is mostly important in the perspective of estimating DTL processes.

Orthomam is a database focusing on Mammals. This database will be helpful to study the rates of duplications and the evolution of these rates in organism. Here the deciphering of duplication patterns is made complex due to the mix of ancestral whole-genome duplications and duplications arising in specific lineages. **Link**: http://www.orthomam.univ-montp2.fr/orthomam/html

Hogenom is a database that mainly focuses on Bacteria, where transfers are observed in relatively large proportions. The database also contains a small number of Eukaryotes and Archaea in order to also study the relationships among early organisms. **Link**: http://pbil.univ-lyon1.fr/databases/hogenom

New methods

Extracting the speciation signal from multigene trees Due to duplication and transfer events, species genomes often have several copies of a same gene. Most gene trees are thus multi-labeled, *i.e.* some species names label more than one leaf. Since no supertree method exists to combine multi-labeled trees, until now they were simply discarded. This has been strongly criticized in a famous paper pointing out that the resulting supertree is just a «tree of 1%» of the available signal. We propose a series of fast algorithms to extract a maximum of speciation signals present within multi-labeled trees and put it under the form of single-labeled trees which can then be handled by supertree methods. When applied on HOGENOM tree collection, this leads to a particularly striking gain in the number of triplets, the basic building stones of elementary speciation information): from around 68k triplets (1% of all possible triplets) up to 23 millions (89% of those). Moreover, the whole process of extracting the speciation signal takes only minutes. This allowed us to infer a well resolved supertree containing the hundreds of taxa present in HOGENOM.

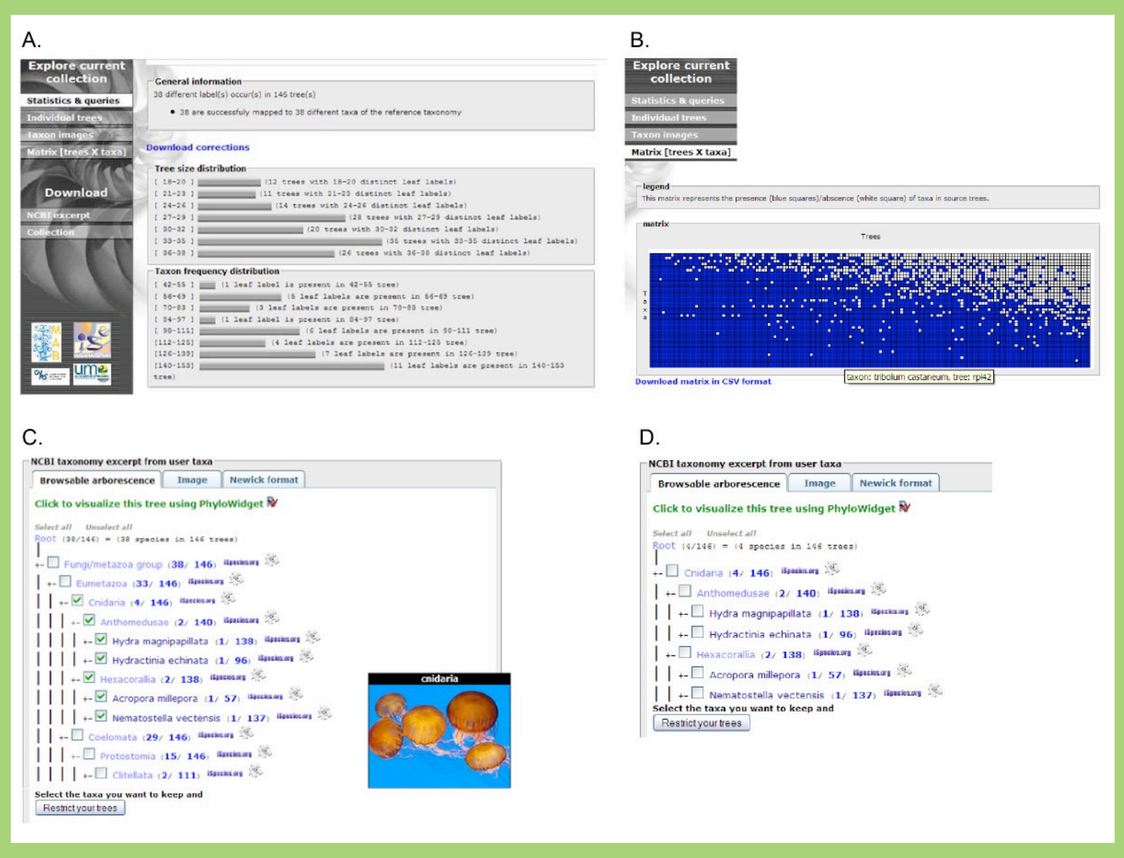
A fast and easy-to-use method to reconstruct galled networks This synthetic representation of a set of rooted gene trees helps to detect and visualize past hybridization or transfer events. Such a network contains all clusters of the input trees (or only those present in a chosen proportion of trees) and is built in two steps: first, a large subset of taxa leading to a largest set of compatible clusters is found, then the missing taxa are connected to the tree using as few reticulations as possible. Though this involves solving two NP-hard problems in sequence, two efficient FPT and branch-and-bound algorithms allow to handle hundreds of trees on a hundred taxa in a handful of seconds. The method is implemented in the freely available Dendroscope software. **Link**: http://www.dendroscope.org

New tools

We developed two operational tools to assist the studies conducted in this project. They have been advertised to the community at large, as their scope exceeds the scope of the project in several aspects.:

ScripTree: is a scripting language to perform graphical analyzes on collections of trees. It automates the graphic processing through a series of tree rendering operations, some of them involving additional annotations. This tool answers some of the challenges in visual exploration of tree collections that arise in phylogenomic analyses. **Link**: http://www.scriptree.org

PhyloExplorer facilitates assessment and management of phylogenetic tree collections. Given an input collection of rooted trees, it provides statistics describing collections (A,B), facilities for correcting invalid taxon names, extracting taxonomically relevant parts of collections (C,D), and identifying related trees in TreeBASE. **Link**: http://www.ncbi.orthomam.univ-montp2.fr/phyloexplorer



Ongoing work

We are developing a probabilistic model to explore the combined space of DTL rate parameters, genomes and gene trees. The model is implemented in a straightforward parallel architecture wherein *(i)* a server node will be searching for a genome tree, with the probability of a given genome tree enumerated based on multiple gene trees; *(ii)* client nodes for a given genome tree, search gene trees with the probability of a given gene tree topology being determined from both the multiple sequence alignment and an optimized DTL scenario with respect to the genome tree. The outcome of the model will be a posterior set of alternative genome trees reflecting scenarios of the prevailing trend of genomic descent together with posterior distribution of DTL rates along its branches.

First experiments on the probabilistic landscape of the space of reconciliations show that its probability mass is mostly located into a small and connected subspace of parsimonious reconciliations. Because this subspace can be explored in linear time, we hope that this will improve the approximation (according to a Bayesian framework) of the posterior probability of a given gene tree according both to its precision and the CPU time.

Selected publications

Journals

Genomes as documents of evolutionary history. B. Boussau, V. Daubin. *Tree*, 1192, 2009.

PhyloExplorer: a web server to validate, explore and query phylogenetic trees. V. Ranwez, F. Delsuc, N. Clairon, S. Pourali, N. Auberval, S. Diser, V. Berry. *BMC Evolutionary Biology*, 9:108, 2009.

Computing Galled Networks from Real Data. D. Huson, R. Rupp, V. Berry, P. Gambette & C. Paul. *Bioinformatics*, (ISMB spec. issue), 25(12), 2009.

Space of Gene/Species Trees Reconciliations and Parsimonious Models. J-P. Doyon, C. Chauve, and S. Hamel. *Journal of Comp. Biology*, 2009.

Databases of homologous gene families for comparative genomics. Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L, Gouy M, Perriere G. *BMC Bioinformatics*, 10, 2009.

International conferences

From gene trees to species trees through a supertree approach. C. Scornavacca, V. Berry, V. Ranwez. 3rd Int. Conf. on Lang. and Automata Theor. and Applications (LATA), LNCS, Springer, vol. 5457, 702-714, 2009.

Méthodes et Algorithmes pour la Bio-informatique LIRMM
Laboratoire d'Informatique de Microélectronique et Robotique de Montpellier

Institut des Sciences de l'Évolution de Montpellier

LBBE
Laboratoire de Biométrie et Biologie Évolutive