Applications Note

TreeQ-VISTA: An Interactive Tree Visualization Tool with Functional Annotation Query Capabilities

Shengyin Gu[1], Iain Anderson[2], Victor Kunin[2], Michael Cipriano[3], Simon Minovitsky[3], Gunther Weber[1], Nina Amenta[1], Bernd Hamann[1], Inna Dubchak[*,2,3]

[1] Institute for Data Analysis and Visualization (IDAV), Department of Computer Science, University of California, Davis, One Shields Ave., Davis, CA 95616, USA

[2] DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

[3] Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA, 94720, USA

Associate Editor: Chris Stoeckert

ABSTRACT

Summary: We describe a general multiplatform exploratory tool called TreeQ-Vista, designed for presenting functional annotations in a phylogenetic context. Traits, such as phenotypic and genomic properties, are interactively queried from a user-provided relational database with a user-friendly interface which provides a set of tools for users with or without SQL knowledge. The query results are projected onto a phylogenetic tree and can be displayed in multiple color groups. A rich set of browsing, grouping and query tools are provided to facilitate trait exploration, comparison and analysis.

Availability: The program, detailed tutorial and examples are available online at http://genome-test.lbl.gov/vista/TreeQVista.

Contact: vista@lbl.gov

1 BACKGROUND

In the genomic age, invaluable insights can be derived from comparisons between properties of different organisms. Such contextual information, including both phenotypic and genomic data, often results in important biological findings, such as the characterization of novel proteins or even whole genomic machineries. An example of these is a recent elucidation of the CRISPR/CAS system, a potential RNAi-based prokaryotic analog of an immune system (Haft, et al., 2005;Jansen, et al., 2002; Mojica, et al., 2005).

One of the commonly used comparative genomic methods is the analysis of co-occurrences of genes across organisms, termed phylogenetic profiles. There are multiple tools available for the analysis of phylogenetic profiles (Tatusov, et al., 1997; von Mering, et al., 2005). However, phylogenetic profiles of a gene consistently present in a small clade and another gene sporadically dispersed across distantly related organisms (i.e. direct inheritance vs. lateral gene transfer) result from very different evolutionary phenomena, and require very different interpretation. Currently, such analysis relies on detailed expert knowledge of phylogenetic relationships between the involved organisms. To facilitate a rapid interpretation of the profiles, there is a clear need to display multiple traits on a phylogenetic tree.

For visualization of phylogenetic trees, several efficient programs are available (Page, 1996)(Zmasek and Eddy, 2001) (Kumar, et al., 2005). For an updated list see [http://evolution.genetics.washington.edu/phylip/software.html]. These programs provide tools for visualizing, rendering and manipulating of trees, and in some cases they are attached to the tools for phylogenetic inference (Kumar, et al., 2005). However, these tools are not designed to query and display traits of organisms represented on a tree. There is no commonly used visualization software which reflects the phylogenetic paradigm for database access.

2 APPLICATION

TreeQ-Vista is an easy-to-use multiplatform interactive tool designed for querying functional annotations from a database and displaying them in context with a phylogeny. The tree is precomputed and provided by a user, and can reflect the phylogeny of a particular gene, a group of organisms or any alternative phylogeny a user prefers. Examples of traits that can be displayed by TreeQ-Vista include a phenotype (such as aerobic vs. anaerobic), a gene's properties (such as activity), and a genomic presence/absence profile of a gene family, domain, or pathway. Traits are displayed on a tree using a pre-defined color code, and multiple queries can be displayed at one time by color-coding different queries using separate colors. While our tool does not allow for the elucidation of links between microbial phenotypes and genotypes, which is a subject of other detailed studies (Goh, et al., 2006; Jim, et al., 2004), it gives a user the ability to query them simultaneously. For example, one could use a query for the phenotype "anaerobic" and the presence of pyrophosphate-dependent phosphofructokinase to test the theory that this enzyme is used mainly by anaerobic organisms to generate a higher ATP yield from glycolysis (Mertens, 1991).

While the TreeQ-Vista can display traits by parsing text-based files, its particular strength is in providing a powerful and user-friendly SQL query interface that extracts data from existing databases and projects the occurrence of traits onto a tree. In the last ten years , databases of various types, such as protein and RNA structures, expression data, and morphological data, have become ubiquitous. We are now accumulating data at an unprecedented, ever increasing speed and database technology has become crucial due to its advanced capability of constructing complex SQL queries that flat file storage approaches do not provide. Research in comparative genomics is often based on relational databases,

---

[*]To whom correspondence should be addressed.
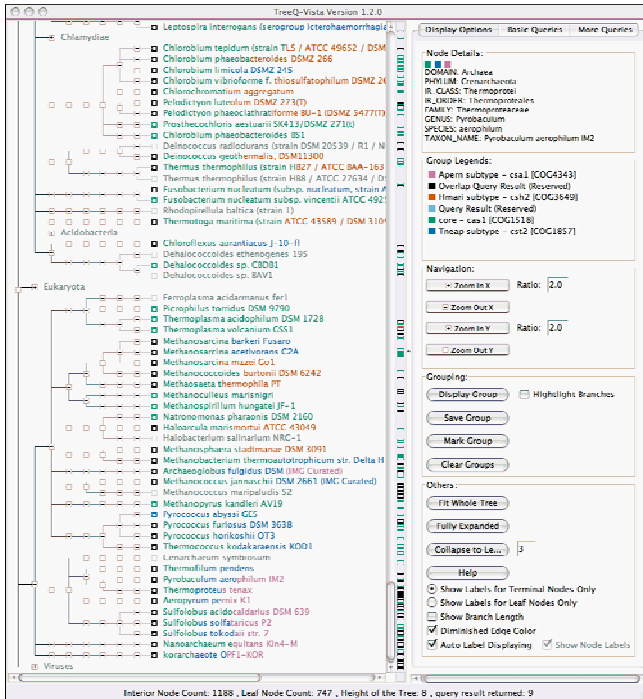
which store multiple properties of genomes as well as their pheno-types (Goldovsky, et al., 2005; Markowitz, et al., 2006; Peterson, et al., 2001; von Mering, et al., 2005). Therefore, complex queries can be addressed to a database of a user's choice and configuration, then processed and immediately displayed in a color-coded form on a tree. TreeQ-Vista can be used with databases of any sizes when a set of input files describing the database, tree and properties are provided to TreeQ-Vista. The ability of TreeQ-Vista to combine SQL queries against a database with the display of the query result onto a phylogenetic tree is unique to our knowledge. TreeQ-Vista greatly facilitates a much more efficient scientific discovery process.

Figure 1. A screenshot of a TreeQ-Vista window on a MacOS X computer showing a fragment of the taxonomic tree. Colors denote the presence of each of the four COG groups in the genomes resulting from an SQL query to the IMG database (Markowitz, et al., 2006). A major CAS core protein COG1518 (*cas1*) is depicted in green along with three subtype-defining genes: COG4343 (*csa1*) - Apern subtype in pink, COG1857 (cst2) - Tneap subtype in blue and COG3649 (*csh2*) - Hmari subtype in orange. Subtypes are defined according to (Haft, et al., 2005).

3   DISCUSSION

In the described example we aimed to investigate the distribution of CRISPR-associated sequences (CAS). CRISPRs (Clustered Regularly Interspaced Palindromic Repeats) are a novel class of repeats, separated by unique spacer sequences of similar length that are present in multiple prokaryotic genomes (Jansen, et al., 2002; Mojica, et al., 1995). The CAS genes appear in conjunction with CRISPRs and are thought to be involved in the propagation and functioning of these repeats. It has been recently proposed that the CRISPR/CAS system samples, maintains a record of, and inactivates invasive DNA that the cell has encountered, and therefore constitutes a exceptional prokaryotic analog of an immune system (Mojica, et al., 2005). Most CAS gene cassettes contain a set of

core proteins, present in most CRISPR-containing genomes, and a set of specific proteins define the particular subtype of CRISPR/CAS subsystem (Haft, et al., 2005).

A regular search for multiple marker genes in such case would consist of querying a database for each of the CAS subtype markers individually, then manually coloring branches on the tree using graphic editing software. TreeQ-Vista allows us to perform this whole exploratory routine, combined with a visual presentation in an efficient user-friendly manner in the matter of seconds. The graphical representation of the results of a database query for each of the marker COG groups is shown on Fig 1. The wide distribution of the core CAS in Archaea comes in stark contrast to the patchy distribution of the CAS subtypes. An extensive horizontal transfer as well as presence of multiple subtypes within some genomes can be readily seen. The TreeQ-Vista makes such associations rapid to discover from SQL queries, and the visualization makes it clearly visible by eye and can greatly speed up the discovery process.

4   IMPLEMENTATION AND AVAILABILITY

TreeQ-Vista was implemented in Java v1.5. TreeQ-Vista utilizes the VectorGraphics package to export the tree display along with traits to a variety of graphics formats [http://java.freehep.org/vectorgraphics/index.html]. The program accepts MySQL, Oracle and Derby database formats. User-specified database connection, tree description and property key description are read by the program, thus providing flexibility to handle different databases and properties to be queried. TreeQ-Vista also reads from a newick file format and a file containing grouping information to display traits from a flat file input. TreeQ-Vista is a general tool that can be utilized in many other areas of science.

TreeQ-Vista is available for download at http://genome-test.lbl.gov/vista/TreeQVista. More examples of sample queries and a step-by-step tutorial are available at the same website.

REFERENCES

Goh, C. et al. (2006) Integration of curated databases to identify genotype-phenotype associations, BMC Genomics, 7:257.

Goldovsky, L. et al. (2005) CoGenT++: an extensive and extensible data environment for computational genomics, *Bioinformatics*, **21**, 3806-3810.

Haft, D.H. et al. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes, *PLoS Comput Biol*, **1**, e60.

Jansen, R. et al. (2002) Identification of genes that are associated with DNA repeats in prokaryotes, *Mol Microbiol*, **43**, 1565-1575.

Jim, K. et al. (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes, Genome Res, 14(1):109-115.

Kumar, Y. et al. (2005) Graphical representation of ribosomal RNA probe accessibility data using ARB software package, *BMC Bioinformatics*, **6**, 61.

Markowitz, V.M. et al. (2006) The integrated microbial genomes (IMG) system, *Nucleic Acids Res*, **34**, D344-348.

Mertens, E. (1991) Pyrophosphate-dependent phosphofructokinase, an anaerobic glycolytic enzyme?, FEBS Lett. 285, 1-5.

Mojica, F.J. et al. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements, *J Mol Evol*, **60**, 174-182.

Mojica, F.J. et al. (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning, *Mol Microbiol*, **17**, 85-93.

Page, R.D. et al. (1996) TreeView: an application to display phylogenetic trees on personal computers, *Comput Appl Biosci*, **12**, 357-358.

Peterson, J.D. et al. (2001) The Comprehensive Microbial Resource, *Nucleic Acids Res*, **29**, 123-125.

Tatusov, R.L. et al. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.

von Mering, C. et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucleic Acids Res*, **33**, D433-437.

Zmasek, C.M. and Eddy, S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees, *Bioinformatics*, **17**, 383-384.