Phylogenetic Trees: An Information Visualisation Perspective

Savrina F. Carrizo
School of Information Technologies
University of Sydney
New South Wales, Australia, 2006
scarrizo@it.usyd.edu.au

Abstract

The development of powerful visualisation tools is a major challenge in bioinformatics. Phylogenetics, a field with a growing impact on a variety of life science areas, is experiencing an increasing but poorly met requirement for software supporting the advanced visualisation of phylogenetic trees. Visualisation problems within the domain are commonly experienced by its researchers, but are poorly documented. Furthermore, the applications in the domain have not been reviewed from an information visualisation perspective. In this paper, the problems are defined and the methods employed by phylogenetic applications are reviewed with respect to related research within *information visualisation*. The results of a survey of the visualisation needs of phylogenetics researchers are also presented.

The full version of this report is available at http://www.it.usyd.edu.au/~scarrizo.

Keywords: Phylogenetic trees, Information visualisation, software.

1 Introduction

Phylogenetics is a field with a growing impact on a variety of life science areas and can benefit greatly from the use of *information visualisation* techniques.

Phylogenetic analyses rely heavily on visual inspection, structural comparison, manipulation and exploration of phylogenetic trees, and thus present a number of visualisation challenges. While phylogenetic inference methods are comparatively well developed, tools in the domain are characterised by a lack of effective visualisation techniques (Munzner *et al.*, 2003).

There is a scarcity of literature regarding the visualisation needs of the domain and no reviews detailing current approaches. This paper fills this gap by presenting a comprehensive review of phylogenetic tree visualisation problems and current approaches employed that aim to address them (section 3). The results of an international survey of phylogenetics researchers, regarding their visualisation problems and requirements, are also presented (section 4). Related research from *information visualisation* that may be used to meet some requirements is also identified.

Copyright © 2004, Australian Computer Society, Inc. This paper appeared at the *2nd Asia-Pacific Bioinformatics Conference (APBC2004)*, Dunedin, New Zealand. Conferences in Research and Practice in Information Technology, Vol. 29. Yi-Ping Phoebe Chen. Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included

2 Information Visualisation

Information visualisation, the application of computer methods for displaying information graphically, is an integral part of the field of bioinformatics. It has roles not only in analysis, but also in building more user-friendly interfaces, implementing methods to navigate large information spaces intuitively, and powerful techniques to browse and query data interactively via the visualisation (Robinson and Flores, 1997).

A good visual representation of the data under consideration will aid the cognitive processes involved in interpreting the data, while a poor visual representation will hinder or may even mislead the viewer.

3 Problem Taxonomy

There are five main categories of phylogenetic tree visualisation problems:

1. Layout
2. Labelling and annotation
3. Navigation
4. Tree Comparison
5. Manipulation and editing

Editing may be considered as not strictly an information visualisation issue. However, as indicated in section 4, researchers report that output formats of many programs are unsuitable for publication and often use Adobe Illustrator and CorelDraw to edit trees manually.

3.1 Layout

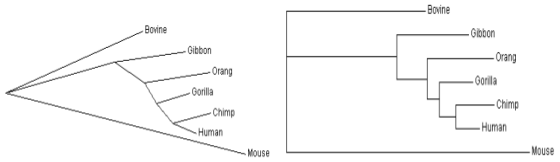
There are various tree topologies, of varying utility, used for displaying phylogenetic trees. The most commonly used are the phylogram, radial and slanted cladogram (Munzner, *et al.* 2003) (Figure 1).

Currently the choice of what phylogenetic tree styles to include in applications and to publish appears to be arbitrary or based on conventions in the field rather than consideration of the effects of layout on understanding. For example, most trees published in journals are cladograms or phylograms with equal edge lengths, despite branch lengths being an important indicator of evolutionary time.

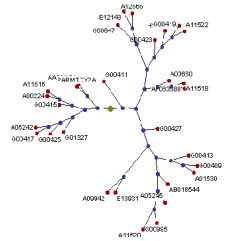
There is a need to assess the effects of phylogenetic tree styles on understanding as authors (Dengler and Cowan,

1998) report differences in the perception of a graph depending on layout. Furthermore, the complexity of designing algorithms to address various phylogenetic visualisation issues differs with different tree styles. For example, resolution of the static labelling problem described in section 3.2 is more complex for radial trees than for phylograms. Also, interactive navigation techniques built on top of a poor visual representation increases the cognitive load of the viewer (Ware, 2000).

Overall, studies on the effects of layout would assist in defining a priority for tree styles to work on when attempting to resolve visualisation issues.



Cladogram **Phylogram**



Radial

Figure 1: Common phylogenetic tree layouts

3.2 Labelling and Annotation

As per Fekete and Plaisant (1998) the labelling problem can be defined as follows; given a set of graphical objects, find a layout to position all the labels such that each label is:

1. readable
2. non-ambiguously related to its graphical object
3. does not obscure any pertinent information

"Completeness, the labelling of all objects, is often desired but not always possible" (Fekete and Plaisant, 1998).

Any labelling strategy must take into account the following labelling considerations: (1) A phylogenetic tree can have labels for any node or edge (2) Labels may be long (e.g. latin nomenclature) (3) There may be more than one label on any given node or edge (4) Edge lengths are biologically meaningful. Resolution of this issue varies in complexity for different tree topologies; for example, this problem is more challenging for a radial tree than for a phylogram.

3.2.1 Information Visualisation Approaches

Labelling is a challenging problem throughout information visualisation; both static and dynamic techniques have been employed to address the issue.

The goal of static labelling techniques is to visually associate a maximum number of labels with their associated graphic objects in the best way possible (Fekete and Plaisant, 1998). Static label placement has been shown to be an NP-complete problem (Neyer, 2001).

Beyond a certain number of nodes (approx. 30-40), dynamic labelling techniques need to be used to achieve that which static techniques cannot.

Fekete and Plaisant, (1998) present a taxonomy of both static and dynamic labelling techniques.

3.2.2 Phylogenetics Software Approaches

The label-all-leaves approach is typically used. This often results in overlapping labels and occlusion, ambiguous label placement with respect to the corresponding graphical object, and the need to zoom-in to see labels. Such problems result in a poor visual representation and cause difficulty in interpreting even relatively small trees (30-40 nodes).

In phylogenetics, only one paper mentions the labelling problem and attempts to address the issue, with limited results. Munzner *et al.* (2003) report translucent labels as difficult to read and in their application, TreeJuxtaposer, use a "contrasting border rather than the usual opaque background rectangle for the label". However, information can still be hidden. When trees have more than 40 nodes, users must occasionally turn off labelling briefly to locate areas of interest (Munzner *et al.*, 2003).

3.3 Navigation

The exploration of large phylogenetic trees is important, as biologists need to visualise trees with up to hundreds of thousands of nodes.

Displaying an entire tree may give an indication of the overall structure within it, but makes it difficult to comprehend, thus techniques with which to explore or navigate the tree are required.

3.3.1 Information Visualisation Approaches

Visual data exploration usually follows a three-step process: overview, zoom and filter, and details-on-demand (Keim, 2001). Thus a first step in viewing a large graph is usually to reduce the size of the graph to display by clustering. This is a common and intuitive technique, which can be used in phylogenetics since subtrees can be clustered with their common ancestor and thus collapsed and expanded.

Preservation of the mental map as one navigates through a large graph means that successive views of the graph should not be radically different. Smooth animation methods showing transition from one view to the next can be used to preserve the mental map, although no phylogenetic programs currently employ this method.

A typical method of navigation involves zooming, however this does not preserve the mental map as the viewer typically looses sight of the global context as they zoom deeper into the visualisation.

The focus+context technique is common throughout information visualisation and allows the user to focus on a subset of their data while retaining a sense of its context within the global scheme (Robinson and Flores, 1997).

3D and virtual reality environments have also been used to aid in the navigation of large data sets. It is argued that the third dimension "adds space" in which to pack extra data.

3.3.2 Phylogenetics Software Approaches

Despite the crucial importance of data exploration for large phylogenetic trees, few applications implement algorithms to aid in this process.

Zmasek and Eddy (2001) state that ATV can handle large trees, however, extensive scrolling is necessary, thus making ATV unsuitable for exploration of large trees.

TreeJuxtaposer (Munzner *et al.*, 2003) was designed by information visualisation researchers to support structural comparison of trees with several hundred thousand nodes (Figure 2).

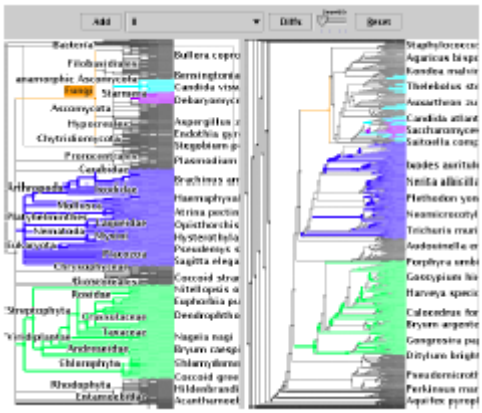


Figure 2. TreeJuxtaposer (Munzner *et al.*, 2003)

Unlike other programs for large trees, it uses a maximum visible detail principal rather than visual aggregation such as clustering (Munzner *et al.*, 2003). 'Guaranteed visibility' is the property that marked areas are always visible no matter what navigation is performed by the user. These areas serve as landmarks and thus help to maintain the user's mental map (Munzner *et al.*, 2003).

TreeJuxtaposer employs 'Accordion tree' navigation, a new rectilinear focus+context distortion technique of expanding and contracting rectangular areas. As if laid out on a stretchable rubber sheet, the tree can be deformed in both the vertical and horizontal directions. The effects of the distortion are global, thus giving the effect of growing some areas and shrinking others (Munzner *et al.*, 2003).

TreeWiz (Rost and Bornberg-Bauer, 2002) uses multiple windows as a means of navigating subtrees. Although reported to scale to 75,000 nodes, the use of multiple windows results in screen clutter and divided visual focus. While this application has attempted to employ intuitive approaches for navigating large trees, Munzner *et al*. (2003) report that navigation is awkward because each viewpoint change spawns a new window. Figure 3 shows an example of window generation during navigation.

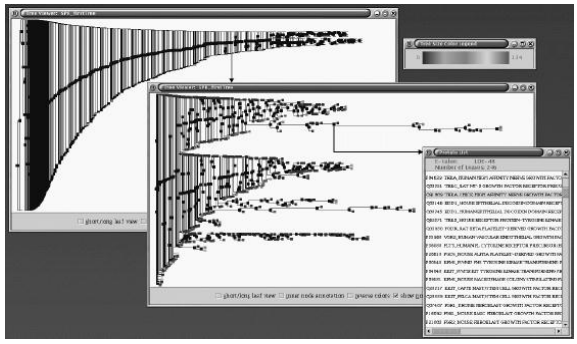


Figure 3. TreeWiz. The arrows have been added to the screen shot and points from the node, which spawned the new window, to the new window (Rost and Bornberg-Bauer, 2002).

An example of a 'focus+context' technique, mostly used with trees, is the hyperbolic projection view, that maps points into hyperbolic space and displays them as either a 2D or 3D representation (Robinson and Flores, 1997; Munzner, 1998). This distorted view makes it possible to interact with potentially large trees (Herman *et al*., 2000).

Figure 4 shows an example of an application written to query a taxonomic database maintained by the National Centre for Bioinformatics (NCBI). The application uses a hyperbolic projection to display the resulting phylogenetic tree. The user is able to drag the underlying tree around, select nodes to be the origin (i.e. the point of maximum magnification), and rotate the scene about the origin (Robinson , 1998).

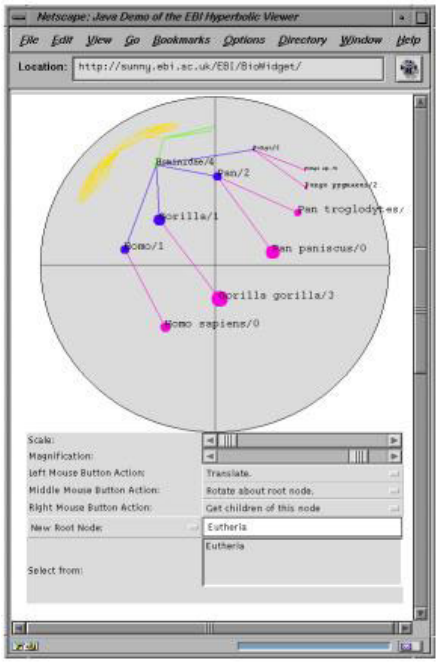


Figure 4. 2D Hyperbolic projection viewer used at the NCBI (Robinson and Flores, 1997).

A drawback of the hyperbolic projection approach is that the coordinate space of the points is distorted. A user might wish to examine the complete data set without focusing upon a particular portion in order to view the spatial relationship. Clustering or distance between points in a phylogenetic tree are biologically significant, thus, inspection of such features will be hampered by the distortion of the points in data space (Robinson and Flores, 1997).

Authors have stated that human perceptual capabilities can be better exploited with 3D visualisations (Can and Vogelmann, 1998; Herman *et al*., 2000). However, 3D visualisations also give rise to problems including occlusion and the difficulty of choosing the best "view" to initially present to the user (Herman *et al*., 2000). A recent application has applied 3D visualisation to phylogenetic trees.

Arbor 3D (Ruths *et al*., 2000) is a system to visualise and interrogate large phylogenetic trees in real-time, in a virtual semi immersive 3D environment. It allows for viewing of the entire data set on a single screen. Interrogation is conducted using an input device to select a node and view its associated data. Trees can be rotated about the root to aid conceptualisation of the relationships. Arbor 3D uses the third dimension as a data dimension to visualise extra information on the tree (Ruths *et al*., 2000). Figure 5 shows the Arbor 3D interface.

The scale and cost of the hardware required limits the utility of this system as biologists rarely have access to such equipment or the funds to acquire them. Phylogenetics thus may benefit from research in the use of 3D metaphors in a standard PC and Macintosh environment to overcome the issue of accessibility.

While the suitability of representing an inherently 2D structure in 3D can be questioned, the use of the third dimension as a dimension on which extra data can be mapped is a promising idea.

Figure 5. The Arbor 3D interface (Ruths *et al*., 2000).

3.4 Comparison of phylogenetic trees

Phylogenetic data is characteristically noisy and incomplete. As such phylogenetic inference is an optimisation problem, which may result in hundreds or thousands of near optimal or equally optimal trees (Amenta and Klinger, 2000). Phylogenetic tree comparison is thus a significant visualisation challenge and is of significance to other biological problems, which also require tree comparison, such as hierarchical clustering of microarray data (Munzner *et al.*, 2003).

A number of phylogenetic programs address tree comparison with varying methods, each with their own merits. The techniques include 3D visual comparison of a set of trees where the third dimension is the time axis (Stewart *et al.*, 2001), point set visualisation for a large set of trees (Amenta and Klinger, 2002), and structural comparison of two trees via colour coding (Munzner *et al.*, 2003).

FastDNAml (Stewart *et al.*, 2001) is a program for the maximum likelihood inference of phylogenetic trees and includes a 3D tree viewer that enables interactive comparison of many different trees.

Two applications have been built on this framework, both of which display trees along a time axis and have the facility for tracing the position of selected nodes or subtrees among the multiple trees (Stewart *et al.*, 2001) as seen in figure 6.

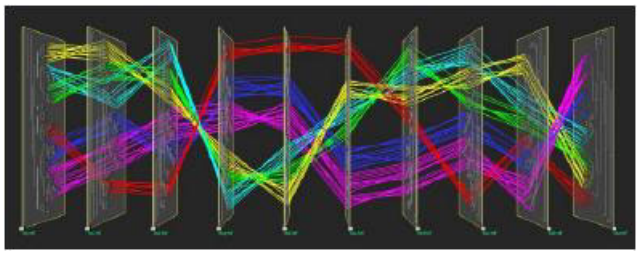


Figure 6. The fastDNAml 3D viewer (Stewart *et al.*, 2001).

One application allows for real-time monitoring and analysis of the computational runs of the fastDNAml algorithm (Stewart *et al.*, 2001). As the algorithm iterates each optimal tree is added to the appropriate location on the time axis. Growth and refinement of the tree can be studied as nodes are added and rearranged. The second application allows any number of trees to be loaded and arranged to allow detailed off-line analysis of trees.

Amenta and Klinger (2002) developed TreeSet, which integrates a point set visualisation of the distribution of a large set of trees, with detail views of an individual tree or of a consensus tree summarising a subset of trees. See Figure 7. The point set view uses a multi-dimensional scaling heuristic (MDS) to arrange points such that a set of points close together have a well resolved consensus tree i.e. the consensus tree is nearly binary, thus supporting a single hypothesis about the evolutionary relationship of the taxa (Amenta and Klinger, 2002). When a user selects a single point the corresponding tree is displayed. When a group of points is selected, their consensus tree is displayed. Selecting multiple points or groups allows for side-by side comparison of the trees (Amenta and Klinger, 2002). Once the search for optimal

trees has been completed, a visual representation of the search process, using colouring and animation features, is displayed. This application is particularly simple to use and effective.

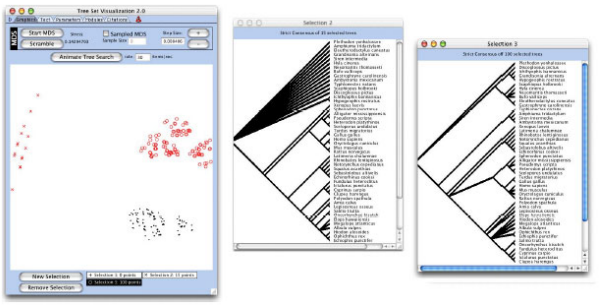


Figure 7. TreeSet Interface with 2 consensus trees of two subsets of points selected (Amenta and Klinger, 2002).

Munzner *et al*. (2003) state that single number summaries of the differences between trees, as used in TreeSet, is too course for biologists who need to understand the structural differences between two trees. TreeJuxtaposer (Munzner *et al.*, 2003) was thus designed to support structural comparison of trees with several hundred thousand nodes.

Structural comparison using TreeJuxtaposer is assisted by a search function on node names, highlighting of structurally different areas, linked views, and interactive colouring of areas in other trees that correspond to the area under the mouse in the active window (Munzner *et al.*, 2003). See Figure 2. Future work suggested by the authors of TreeJuxtaposer includes the consideration of edge lengths in the similarity measure, which are currently not considered.

These applications cater for tree comparison under different conditions; comparsion of a large set of trees and comparison of two very large trees. Thus each have useful applications. Consideration of the suitability of the similarity metrics used for comparison by the applications is necessary before use, as two different metrics may give different results for the same data set being compared.

4 Survey of Phylogenetics Researchers

Given the sparsity of literature on the topic of visualisation in phylogenetics, the problems and requirements of the domain are unclear. A survey was designed and sent via email to many of the domains foremost researchers, including attendees of the 2003 Annual Phylogenetics meeting held at Kaikoura in New Zealand. The aims were to catalogue their visualisation preferences, problems, needs and priorities for these needs. It is hoped that the requirements and problems defined via this survey will lead to a focused effort in improving phylogenetic visualisation tools in the near future.

4.1 Results

The twenty-one responses to the survey revealed many visualisation problems commonly experienced by phylogenetics researchers. One respondent provided a representative comment; “Visualisation tools for the domain are woefully inadequate for trees of non-trivial size i.e. 25+ leaves…I spend too much time wrestling the tree into a usable and publishable format”.

There is a consensus on the main problems experienced as defined in section 3, namely problems with the inability to handle large trees, lack of editing support and the lack of suitable styles for presentation purposes, lack of software to handle phylogenetic networks.

Respondents were asked to priorities their visualisation problems and needs on a scale of 1-5, where 1 was very high priority to be resolved, and 5 was very low priority to be resolved. The problems, requirements and priorities are summarised in Table 1.

Problem/Needs	Priority
Labelling	
Overlapping labels	3
Annotation should be possible on any part of the tree	1
Layout	
Default layouts are unsuitable for publishing, extensive manual re-drawing and editing required	2
Editing	
Full Word-like functionality required e.g. fonts, line widths, cut and paste	1
Ability to save to various file formats required e.g. jpeg, gif	1
Large Trees	
Can’t read and explore on screen, editing to reduce the size of tree sections required	1
Option for zooming in required	1
General	
Disparate programs required i.e. a lack of visualisation software integrated with analysis software	1
Buggy software	1
Lack of software for Macintosh and Unix	3
Extra Features	
Visualise trait evolution on the tree	2
Visualise which sites support or do not support particular branches	2
3D software for phylogenetic trees, networks and associated information	1

Table 1: Common problems and needs identified in the survey. Priorities shown are the averages of the responses.

5 Future Work

Despite some recent attention, further research is required to address the above visualisation issues and meet the needs of researchers.

Future research should include empirical evaluations to provide direction as to which methods should be followed up and developed further and which should be reconsidered and replaced by other methods.

6 Conclusion

All disciplines within biology have the complex challenge of developing and evaluating its visual metaphors. Since no single class of techniques can adequately address the visualisation needs in bioinformatics, successful applications will employ a range of techniques, fine-tuned to best present the data under consideration.

The development of powerful visualisation software can be achieved with biologists and visualisation experts working together; yet another challenge experienced at the junction of a number of disciplines. Biologists must teach visualisation researchers about the complexities of biological analyses and the visualisation researchers must inform biologists about the possibilities and the current limitations in information visualisation.

7 References

Alan Robinson. (2000) About Visualisation at the EBI. http://industry.ebi.ac.uk/~alan/VisSupp/AboutVisSupp.html

Amenta N. and Klinger J. (2002) Case Study: Visualizing Sets of Evolutionary Trees. IEEE Symposium on Information Visualization (InfoVis'02). Boston, Massachusetts, USA, p. 71.

Can K. and Vogelmann V. (1998) Effective Visualisation of Hierarchical Graphs with the Cityscape Metaphor

Card, S. K., Mackinlay, J. D. & Shneiderman, B. (1999): *Readings in Information Visualization.* Morgan Kaufmann.

Dengler E. and Cowan W. (1998): Human Perception of laid-out graphs. Proceedings of the symposium on Graph Drawing GD '98, Springer-Verlag, pp. 441-444.

Fekete, J., D., and Plaisant, C. (1998): Excentric Labeling: Dynamic Neighborhood Labeling for Data Visualization. *Proceedings of CHI'99, Pittsburgh, PA*, USA, May 15-20, 1999, ACM, New York, 512-519. HCIL-98-09.

Herman I. (2002) Graph Visualisation and Navigation in Information Visualisation: a survey. In: *IEEE Transactions on Visualization and Computer Graphics*, 6(1).

Keim D. A. (2001) Visual exploration of large data sets. Communications of the ACM august 2001 Vol 44, No 8

Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L. and Zhou, Y. (2003): TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility. In *Proc. SIGGRAPH 2003.*

Munzner T. (1998) Drawing Large Graphs with H3Viewer and Site Manager (System Demonstration). S. H. Whitesides (Ed.):GD'98, LNCS 1547, pp. 384-393, 1998.

Neyer G. (2001) Map Labelling with Application to Graph Drawing. In M. Kaufmann and D. Wagner (Eds.): Drawing Graphs, LNCS 2025, pp. 247 - 273, 2001. Springer-Verlag, Berlin Heidelberg 2001.

Pagel M. (1999): Inferring the historical patterns of biological evolution. *Nature* 401: 877-884

Rhyne T. M. (2002): Evolving Visual Metaphors and Dynamic Tools for Bioinformatics Visualisation. *IEEE Visualization Conference Proceedings*: 579 - 582

Robinson, A. J., and Flores, T. P., (1997): Novel Techniques for Visualising Biological Information. *Ismb* **5**:241-9.

Robinson A. J. EBI Hyperbolic Viewer, European Bioinformatics Institute, http://industry.ebi.ac.uk/~alan/components, 1998.

Rost U. and Bornberg-Bauer E. (2002) TreeWiz: interactive exploration of huge trees. *Bioinformatics* 18: 109-114

Ruths D. A., Chen E. S. and Ellis L. (2000) Arbor 3D: an interactive environment for examining phylogenetic and taxonomic trees in multiple dimensions. *Bioinformatics* 16: 1003-1009

Stewart C. A., Hart D., Berry D. K., Olsen G. J., Wernet E. A. and Fischer W. (2001) Parallel implementation and performance of fastDNAml – a program for the maximum likelihood phylogenetic inference. *Proceedings of SC2001*, Denver, CO, November 2001.

Ware C. (2000): *Information Visualisation. Perception for Design*. Academic Press, USA.

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