A FRAMEWORK FOR LARGE-SCALE INTERACTIVE VISUALIZATION OF PHYLOGENETIC TREES

Joerg Meyer
University of California, Irvine
644E Engineering Tower, Irvine, CA 92697-2625, U.S.A.
jmeyer@uci.edu

ABSTRACT

The integration of rendering services for large-scale datasets is a challenging task when it comes to complex datastructures. While regular and irregular grids have been widely addressed before, we are going to discuss other complex datastructures, including trees and super-trees, and our new approaches to visualize those very large datasets hierarchically. The framework covers techniques suitable for both web-based and immersive presentation.

Keywords: Phylogenetic trees, interactive visualization, information visualization, large-scale visualization, web-based rendering, immersive rendering

1 INTRODUCTION

Interactive visualization tools for large-scale volumetric grids [30] and large unstructured meshes have been studied in great detail [15]. Applications in the fields of medical imaging, biomedical visualization, and bioinformatics have driven joint efforts to develop software to provide tools for physicists and biologists to organize their data, analyze it, and setup distributed databases and visualization systems. These systems turned out to be very useful to capture image data, to allow for navigating the data, and to make it accessible to the community to foster collaborative and cross-disciplinary research.

However, one common datastructure, which is used by a wide-spread group of biologists, trees and super-trees, was left out. We try to fill this gap by introducing new rendering techniques and hierarchical data transmission schemes for so-called phylogenetic trees, which are used to sort various kinds of species or their gene sequences.

Phylogenetic trees depict the hierarchical pattern of common ancestry of species, genes, sequences or other entities (taxa). Phylogenies have become a widely and routinely used tool in the biological and biomedical sciences, especially in the analysis of molecular sequence variation.

In order to study phylogenies and to explore their predictive properties, large-scale tree databases have been created using various formats. At the same time, work is advancing on the synthesis of extremely large phylogenetic trees from these large-scale databases of smaller trees. The quality of diagnostic predictions strongly depends on the completeness of a set of samples of a particular group of organisms or genes, and the correctness of the connectivity information between the taxa. Therefore it is essential to have precise navigation and visualization tools to interact with the tree structure and the structure of relationships between trees [16].

In order to visualize enormously large data sets now available, we need tools ranging from real-time immersive 3-D environments to web-based desktop PC applications. Hierarchical data storage and new query techniques are used to access the data. Tree data is transmitted over the internet using progressive schemes. The user is provided with several visualization, navigation, and query options, ranging from global views down to detailed views of single trees or taxa, or comparative side-by-side and superimposed imagery. Intuitive visualization paradigms, such as stick models, funnels, magnifying lenses, icons for clusters, etc., are used to present the user with a comprehensive user interface to navigate the trees.

2 DATASETS

Since the 1960's when phylogenetic techniques were first used to study molecular variation within and between human populations [7] and among vertebrate proteins [52][53], they have been applied in an increasingly diverse array of problems. These range from "conventional" phylogenetic reconstructions using homologous sequences in different species, such as the small-subunit rRNA trees spanning all life [36][37], to less conventional comparative studies of genes within single genomes, multigene families, or between hosts and pathogens. These latter studies often bear directly on basic understanding of molecular biology or human health issues. For example, based on phylogenies of R1 retrotransposons within a single species (*Drosophila*), inferences about mechanisms of recombination were possible [27]. Another example is the huge superfamily of G-protein-coupled receptors, which function in intercellular communication. Phylogenetic relationships

(e. g., [50]) have suggested useful pharmaceutical applications. In biomedical research, phylogenetic trees are being increasingly applied in investigating the biology of human pathogens, such as *Plasmodium* [38], fungi such as *Pneumocystis* and *Candida* [5], viruses such as *HIV* [10][25][32], and many others. Phylogenetic techniques have been used to trace contact histories for infectious diseases [35] and identify geographic origins of new outbreaks, as in the case of *West Nile Virus* [26], and even the timing of new introductions [29], suggesting their broad explanatory power in epidemiology [18]. Phylogenetic analyses are beginning to have direct clinical implications as in the study of Wade et al. [49], which used phylogenies of *HIV-1 gag* and *env* genes to infer mechanisms of transmission of multiple sequence variants of the virus from mother to infant.

Phylogenies are predictive: the diagnostic features of a clade can be used to predict features of an unstudied member known to belong to that clade. A more complete sample of taxa from any clade, and therefore a more comprehensive phylogeny can provide much more precision in these diagnostic predictions. Including more taxa may also lead to a more accurate reconstruction of their relationships [4][17].

Until recently most phylogenetic studies have been relatively small, including on the order of 100 taxa or less, but large-scale phylogenies are becoming more commonplace with the advent of high throughput sequencing. Phylogenies of several thousand homologous sequences have now been published, and there is good reason to expect that the trend toward larger trees will continue. The number of homologous sequences known for many genes has spiraled upward in recent years, which means that the size of aligned datasets suitable for phylogenetic analysis has increased dramatically. Even a cursory survey of the genome databases reveals many genes with 100+ homologous sequences. There were 384 aligned sequences of the *env* gene of *HIV-1* in the Los Alamos *HIV* database as of January 1999 [24]; but even this is a "modest" number in comparison to the nearly 10,000 small subunit rRNA sequences reported in the latest release of the RDP II database [39]. For pathogens with small genomes and few genes, such as *HIV* and other RNA viruses, the number of sequences of homologus regions across different samples will undoubtedly continue to skyrocket, prompting more and more efforts to estimate comprehensive phylogenetic trees. Over 50,000 sequences from *HIV-1* are presently in GenBank.

Yet there is a looming technical problem with reconstructing large phylogenetic trees. Large trees are difficult to visualize, which makes it difficult to evaluate results and come to synthetic conclusions. The structure embodied by a tree of 1000 (or even 100) sequences is extremely complex (Fig. 1).

There are currently almost no software tools that even permit the display of such large trees, much less any informative browsing of them, or data analysis. Crude standard tools for visualization of large trees, which merely display a tree in a rectangular window, offer the user a choice between pruning away most of the tree that will not fit in the window, or compressing it to the point where all resolution is lost. This leads to more than just esthetic problems. For example, an investigator may have intensively sampled one clade of closely related viral strains (say from one population or individual), producing one cluster with a very large number of branches. In attempting to understand the relatives of these strains, any view of the tree will be dominated by the overwhelming number of taxa in one group. Some alternative scheme is needed to visually "downweight" (based on various criteria) this part of the tree to permit its relationship to the remainder of the tree to become apparent.

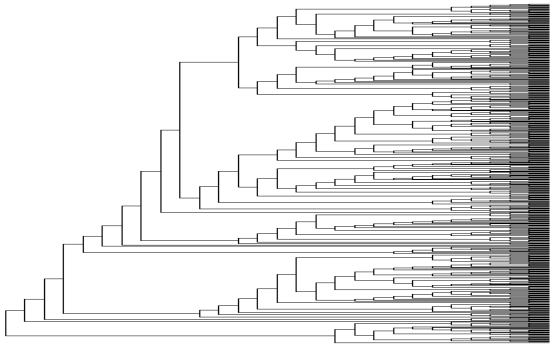

Fig. 1. Conventional representation of a sequence tree

Annotation of a large tree with other kinds of information is also problematic [12]. Investigators are often interested in reconstructing ancestral states of either the sequences themselves (e. g., to investigate adaptive amino acid replacements) or more derivative information, such as the geographic range of the strain, the host, or a classificatory term such as the subtype. Displaying these kinds of information intelligibly in a large tree poses a significant challenge. Not only is it not possible to understand the fine structure of the topology of the tree (the relationships), it is obviously impossible to display the names of the taxa and any associated information about these taxa. Unfortunately, the taxon names are the essential keys that connect the phylogeny to the investigator's understanding of the biology of the organisms. Clearly, this condensed representation as in Fig. 1 is entirely unsatisfactory.

3 DATABASE

We are using an object-relational type database to store our large-scale tree structures. Traditionally, databases supported very organized and structured data that are flat. Relational databases over the years proved to be very efficient, reliable and simple to use. One of the main reasons for the success of relational databases is the existence of a robust core query language that is equivalent to the first-or-

der logic that gives this language a sound mathematical basis. This core language is further grounded on a procedural language such as relational algebra, on which the popular and declarative query language SQL is based. SQL is relatively expressive and efficient, thanks to numerous optimization techniques that enrich this paradigm [19][20][28][43][48].

However, relational models are very poor in supporting complex data structures, hierarchical and taxonomical data, stored procedures or the so called methods, modularity or encapsulation, complex but useful relationships, recursion, deduction, and so on. These observations lead to the development of recursive relational [47], nested relational [9], object-oriented [3], deductive [48], and deductive object-oriented [22] data models. These models evolved into the now famous object-relational [2][21][44] data models. All these new models come with various strengths and weaknesses. Although these models extend the relational model in significant ways, they are still limited in supporting somewhat structured data. For each of these models, query languages exist, and one of them, SQL3, has been adopted in the meantime as an international standard for object-relational databases.

The World Wide Web is a good example. With the popularity of HTML and XML documents and their complicated structures, there arose the need for a more expressive query language. Although there is no generally accepted query language for web documents, several interesting ones have been proposed. Of them XML-QL [11], XSLT [8], XQL [40] for web documents, and Lorel [1], StruQL [14], and UnQL [6] for semi-structured data are good examples. These languages, however, are designed basically for navigation through the object structure and for selecting vertices for display. In particular they do not allow "tree" processing and do not handle tree type values. In other words, they essentially handle flat data in tree-like structures.

In phyloinformatics, tree-structured data that represent the phylogentic information are arbitrary in shape but have somewhat regular vertex types. Good examples of such data types are Tree of Life (`http://phylogeny.arizona.edu/tree/phylogeny.html`), TreeBASE (`http://www.mcb.harvard.edu/BioLinks/Evolution.html`) and Ribosomal Database (`http://www.cme.msu.edu/`). These databases currently do not support any querying as the majority of them are web enabled and support only hyperlink based traversals or prefabricated form based queries that heavily rely on extensive coding and system development. For example, it should be possible to ask if there exists a tree that is almost similar to another tree in the database either in structure or in information content, or if there is a subtree that when added as a child of a node in another tree becomes equivalent to a given tree, and so on. These queries cannot be asked in the current phylogenetic databases or data repositories.

Therefore we are developing a query language for tree-structured data which permits the formulation of queries with tree-valued data and produce tree valued results. In other words, trees are regarded as the unit of data. The results are visualized using one of the rendering paradigms given below.

4 VISUALIZATION PARADIGMS

Phylogenetic tree structures are usually too complex to be visualized in full detail. Therefore it is necessary to stagger the visualization hierarchically. Existing methods include hyperbolic trees [33] or fractal techniques [23]. Based on these ideas, we are taking a somewhat different approach. The main differences are in the graphical user interface (GUI) and in the underlying database. Multiple levels of abstraction are needed in the user interface and in the database. The GUI provides the biologist with the following options: (a) a general overview, (b) a simplification (abstraction), and (c) methods to refine the image by navigating the scene (database query).

The visualization of collections of trees and their relationships to each other is the goal of this framework. Different kinds of collections of trees can have uses ranging from browsing tree databases in an exploratory fashion, to answering specific biological questions. The software allows the biologist to observe and explore a phylogeny database. The biologist not only has some ideas about the tree he or she is looking for in the collection, but also wants to be guided by the system to locate unknown trees which are related to the specified ones. The trees found in the database can be *selected* for further detailed analysis.

Fig. 2 gives a possible graphical presentation of a tree database. The disk-shaped surface represents the database and the shaded regions on this surface sets of “similar” trees. Similarity between trees is represented by a measure chosen from a given selection by the biologist [13]. An example for a measure between two trees is the number of their common taxa (this measure was introduced by Sanderson et al. [41]). Dark shaded regions contain trees that are specified by the biologist to analyze them in the context of similar trees in the database [50].

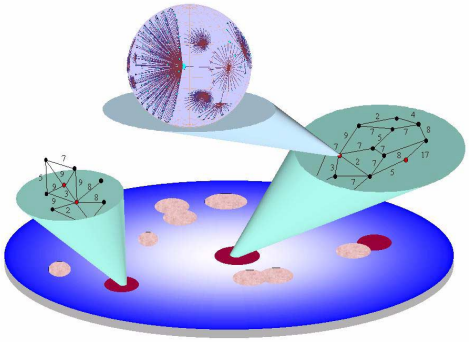

Fig. 2. Collection of trees (database)

Navigation among similar or "related" trees can be useful in various contexts. For example, an investigator studying the evolution of a gene or gene family involved in disease resistance to a pathogen may be interested in using its phylogeny to make inferences about the history of amino acid substitutions that might generate testable hypotheses about protein structure and function. To estimate the direction of evolutionary change it is necessary to root the tree correctly, which is often difficult because rooting is generally accomplished via selection of outgroups that may be obscure or unknown to the investigator. However, outgroups will often be found in other phylogenetic studies that are "nearby" in the sense described above – that is studies that share at least a few sequences in common with the study at hand. Rapid navigation among studies containing potential outgroups assists the experimental design of phylogenetically informed studies.

Another kind of similarity or relatedness between trees is the relationship imposed by specificity of host and pathogen. Given the increasing availability of phylogenies of both host and their pathogens it is feasible and useful to navigate from host to pathogen tree and back again, examine congruence between their histories, and make inferences about specificity and the likelihood of host shifts.

The dark shaded regions can be selected for analysis. We imagine for each selected region a funnel pops up, representing the trees and their similarity relation as a graph $G = (V, E)$ (second level). The node set V represents the trees, and the edge set E represents "similar" trees. Edges are weighted by the number of common taxa between edge related trees. Highlighted nodes represent the trees that were chosen by the biologist [33].

Basically the whole database is represented as one graph $G = (V, E)$, in which edges in E connect trees that have at least one common taxon. Thus, connected components of the graph represent all trees that have at least one common taxon. Fig. 2 represents these components by shaded regions. Depending on the selected measure, the connected components could also be put in relation to each other by visualizing the similarity as edges that connect regions (nodes) on the surface (Fig. 3).

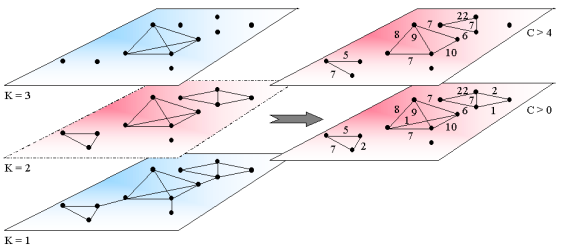

Fig. 3. Connected components and edge weight components

Often connected components of the graph are huge. Thus the biologist must be supported by tools that help him or her with further analysis. Subgraphs of a connected graph can be highlighted in two different biologically meaningful ways. Some trees in a connected component might be highly connected to each other and some trees might be more "similar" to each other than other trees. Thus, the biologist can select subgraphs by their connectivity as well as by their similarity.

Connectivity can be represented by the graph-theoretical construct of k-connected components of a graph. A component is k-connected if there is no way to disconnect it by removing k - 1 edges. As an example the left side of Fig. 3 shows 1-, 2- and 3-connected components of a tree graph selected from a tree database. The lowest layer, $k = 1$, shows a given connected component. On the next higher layers the 2- and 3-connected components of the given graph are shown. Highly similar graphs can be highlighted by selecting a cut-off value for the similarity. Every edge that is below the cut-off value is removed, which leaves only highly similar edges. An example is shown in Fig. 3. We assume that the biologist wants to analyze highly similar trees, given by the cut-off value 4, of the 2-connected components of the given graph. The right side shows in bottom-up order the 2-connected components which are chosen to be analyzed for similarity, and then the 2-connected graph that has only edges that connect trees with a similarity greater than 4.

As shown in this image, connected components can be selected by the user and then analyzed in detail. Within a connected component, some trees might be "similar" to a large number of other trees, while others are not. Highlighting k-connected components of the graph permits one to distinguish between highly connected and weakly connected trees. A three-dimensional scheme, as shown in Fig. 3, shows all related k-connected components in one image. The biologist may remove some of these levels to narrow the focus. In addition, edges can be colored or highlighted according to intervals or ranges of similarity. This way low similarity edges can be hidden to focus only on stronger similarities by giving them no color or a lower intensity. Extremely similar or almost identical structures, which might be already known or not interesting, might be excluded as well. We call this technique "windowing". The colors or intensities may refer to different levels in a three-dimensional projection of the graph as shown in Fig. 3.

Further analysis can be made of selected subsets of the trees in the database. The selected tree set allows to manipulate and display a set of similar trees together with a "consensus" tree, which displays aspects of commonality. A large variety of consensus tree methods have been described in the literature [46]. Most biologists use some variant of strict or majority-rule consensus methods, which convey those groups present on all or a majority of trees in a collection, respectively. Recently, largest common pruned trees have been used more widely, and are especially promising for studies in which trees only partially overlap in their taxa. Largest common pruned trees remove the

fewest number of taxa necessary to obtain agreement among the trees in a collection. All consensus methods keep some information about a collection of trees and lose other information, and therefore it is especially critical to convey the properties of these consensus trees and permit interactive exploration of alternative views.

The biologist is provided with a set of different methods to obtain a consensus tree [45], such as strict or local consensus and super-tree methods [42]. Our notion of visualization represents a set of trees as "satellites" of their consensus tree.

The consensus tree appears in the center, and the satellite trees in the peripheral region. For some consensus methods a measure between a satellite tree and the consensus tree can be calculated that describes how well the given tree is represented by the consensus tree. Our approach allows the user to graphically distinguish between trees that are well represented and trees that are less well represented by the consensus tree. Trees that are well represented are drawn on an "orbit" close to the consensus tree, while other trees are on a more distant "orbit". Structures in a satellite tree that are common in the consensus tree can be optionally highlighted using a different color.

5 GRAPH VISUALIZATION

For complex tree structures, it is necessary to display them in 3-D, either as a 2-D projection of a 3-D world on a desktop screen, or as a real 3-D scenery on a stereoscopic display. The first option allows to make the contents of the database, which are stored on our visualization server, available on the internet. A web-based interface enables interactive navigation, manipulation, and visualization on a desktop PC or workstation. The second option allows to immerse the user in a 3-D environment, so that he or she can interact with the tree structure in a natural way: by grabbing objects, picking and moving subtrees, and using various navigation tools. The web-based interface serves as a smaller version of the immersive 3-D interface, and makes the same tools and paradigms, which are available in the immersive environment, available to the user at home or at her or his office.

The tree database is stored on a visualization server. The database serves as a repository which can be accessed from a rendering client. This client can be either an immersive 3-D application or a web-based application for desktop displays. The web-based client uses Java3D and OpenGL to render the tree structures.

The performance of the rendering client strongly depends on the complexity of the scene. Due to the complexity of the datasets, it is not possible to render an entire dataset in full detail at interactive frame rates. Therefore we must find methods to reduce the complexity of the geometry, which is derived from the tree structure.

There are two different approaches: (i) reduce the complexity of the underlying tree structure by abstraction (subtrees are symbolized by icons), and (ii) reduce the complexity of the geometry that needs to be transmitted over the internet (mesh reduction). Most of the rendering is done on the server side, so that the client only needs to render the geometry information which is transmitted over the internet. This way we avoid performance problems on the client side.

Rendering a complex geometry scene is still a challenging task, and we need to develop asymmetric compression schemes which allow to compress the data on a high level on the server side and uncompress the data fast and efficiently on the client side. Here again most of the workload is on the server side, because the server is much more powerful than a rendering client, which is connected through the internet.

5.1 JAVA-BASED WEB INTERFACE

The server sends a Java applet to the rendering client [31]. The user at the client side can select a database, and the client sends a request to the server. The server responds by transmitting a general description of the database, which is presented in graphical form to the user. The graphical representation consists of the convex hull of smaller trees, or icons, depending on the level of detail. The user is then enabled to refine her or his request, and the server responds with updated geometry data for the subtrees. The image is updated continuously as the user changes the focus area or the perspective.

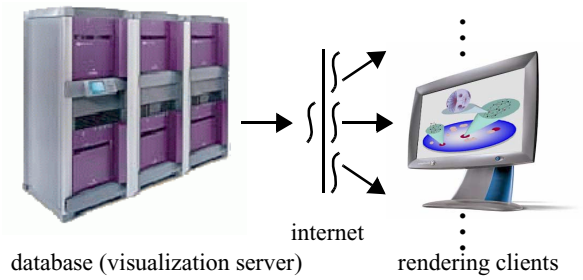

Fig. 4. Java-based web-interface

Standard web protocols (http) and ports are used to send and receive data. All geometric data is wavelet encoded before transmission. Vector quantization is used on the datastream to allow for progressive transmission [30]. The client stores the data locally. A caching scheme [34] is used to minimize transmission costs.

5.2 STEREOSCOPIC IMMERSIVE RENDERING

The NSF-Engineering Research Center for Computational Systems at Mississippi State University provides facilities to render complex objects in an immersive stereoscopic en-

vironment. This device is called CAVE (Cave Automatic Virtual Environment), a room-sized, multi-person, high-resolution, three-dimensional video and audio theater, which surrounds the user by a set of up to six (four in the current setup) projection walls, which fill the entire viewing area with a stereoscopic projection, driven by a powerful graphics engine (SGI Onyx 2 with 2 rendering pipelines, each one split into two channels). Iowa State University has a C6 with 6 rendering pipelines (one per wall).

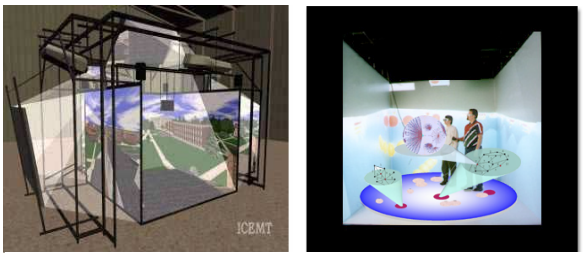

Fig. 5. CAVE (Cave automatic virtual environment)

The CAVE allows the user to navigate and walk the virtual tree, sit on its branches, pull oneself along the branches until she or he finds an interesting subtree. This kind of direct interaction with a complex data structure is only possible in an immersive environment. Subtrees can be selected, broken off like branches, and reattached at another part of the tree. Similar to the web-based interface, the server in the background provides the data, and as the user moves to a different focus area, the scene is continuously updated. The geometry is small enough to maintain interactive frame rates [30]. A map and a jump function allow the user to move quickly between different parts of the tree.

6 CONCLUSIONS

We have presented a framework for interactive rendering of large tree structures. The primary user group that we address is biologists who need to store their data in large tree structures and need navigation and visualization tools to analyze the data. Our framework provides flexible output presentation modes, ranging from a Java-based web interface to a virtual environment (CAVE). The tools are useful to improve predictive content-guided navigation of large-scale phylogenetic trees and collections of trees.

ACKNOWLEDGEMENTS

I would like to thank Oliver Eulenstein (Iowa State University), Hasan M. Jamil (Wayne State University), and Michael J. Sanderson (UC Davis) for their contributions and for some of the images.

REFERENCES

[1] Abiteboul, Serge; Quass, Dallan; McHugh, Jason; Widom, Jennifer; Wiener, Janet L.: The Lorel Query Language for Semistructured Data, *Int. J. on Digital Libraries*, Vol. 1, No. 1, pp. 68-88, 1997.

[2] Ananthanarayanan, R.; Gottemukkala, Vibby; Kafer, Wolfgang; Lehman, Tobin J.; Pirahesh, Hamid: Using the Co-existence Approach to Achieve Combined Functionality of Object-Oriented and Relational Systems, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., pp. 109/ 118, May 26-28, 1993.

[3] Banchilhon, F.; Delobel, C.; Kanellakis, P.: Building an Object-Oriented Database System: The story of O2, Morgan Kaufmann, 1992.

[4] Bininda-Emonds, Olaf R. P.; Bryant, Harold N.: Properties of matrix representation with parsimony analyses, *Systematic Biology*, Vol. 47, pp. 497-508, 1998.

[5] Bowman, B. H.; Taylor, J. W.; White, T. J.: Molecular evolution of the fungi: human pathogens, *Mol. Biol. Evol.*, Vol. 9, pp. 893-904, 1992.

[6] Buneman, Peter; Davidson, Susan B.; Hillebrand, Gerd G.; Suciu, Dan: A Query Language and Optimization Techniques for Unstructured Data, *SIGMOD Conference*, pp. 505-516, 1996.

[7] Cavalli-Sforza, L. L.; Edwards, A. W. F.: Analysis of human evolution. *Proc. 11th Intl. Congress Genet.*, pp. 923-933, 1964.

[8] Clark, J.: XSL transformations, Version 1.0, `http://www.w3.org/TR/WD-xslt`, August 1999.

[9] Colby, L. S.: A Recursive Algebra and Query Optimization for Nested Relations, *Proceedings of the ACM SIGMOD*, pp. 273-283, 1989.

[10] Crandall, K. (ed.): The evolution of HIV, Johns Hopkins, Baltimore, 1999.

[11] Deutsch, Alin; Fernandez, Mary F.; Florescu, Daniela; Levy, Alon Y.; Suciu, Dan: A Query Language for XML, *Proc. of WWW8 Conference*, 1999.

[12] Eulenstein, Oliver; Mirkin, Boris; Vingron, Martin: Comparison of Annotation Duplication, Tree Mapping, and Copying as Methods to Compare Gene Trees with Species Trees, *DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences*, AMS, Vol. 37, pp. 71-93, 1997.

[13] Eulenstein, Oliver; Mirkin, Boris; Vingron, Martin: Duplication-Based Measures of Difference Between Gene- and Species Trees, *Journal of Computational Biology*, Vol. 5, pp. 135-148, 1998.

[14] Fernandez, Mary F.; Florescu, Daniela; Kang, Jaewoo; Levy, Alon Y.; Suciu, Dan: Catching the Boat with Strudel: Experiences with a Web-Site Management System, *SIGMOD Conference*, pp. 414-425, 1998.

[15] Hamann, B.: A data reductions scheme for triangulated surfaces, *Computer Aided Geometric Design*, Vol. 11, No. 2, pp. 197-214, 1994.

[16] Herman, I; Melancon, G.; Marshall, M. S.: Graph Visualization and Navigation in Information Visualization: A Survey, *Transactions on Visualization and Computer Graphics*, Vol 6, No. 1, January-March 2000.

[17] Hillis, D. M: Taxonomic sampling, phylogenetic accuracy, and investigator bias, *Systematic Biololgy*, Vol. 47, pp. 3-8, 1998.

[18] Holmes, E. C.: Using phylogenetic trees to reconstruct the history of infectious disease epidemics, in Harvey, P. H. et al. (ed.): *New uses for new phylogenies*, Oxford University Press, Oxford, England, UK, New York, New York, USA, pp. 169-186, 1996.

[19] Jamil, H. M.; Lakshmanan, L. V. S.: A Declarative Semantics for Behavioral Inheritance and Conflict Resolution, in Lloyd, John (ed.): *Proceedings of the 12th International Logic Programming Symposium (ILPS)*, Portland, Oregon, MIT Press, pp. 130-144, December 4-7, 1995.

[20] Jamil, H. M.: GQL: A Reasonable Complex SQL for Genomic Databases, *IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2000)*, Washington D.C., USA, November 8-10, 2000.

[21] Keller, Arthur M.; Jensen, Richard; Agrawal, Shailesh: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., pp. 523-528, May 26-28, 1993.

[22] Kifer, M.; Lausen, G.; Wu, J.: Logical Foundations for Object-Oriented and Frame-Based Languages, *Journal of the Association of Computing Machinery*, Vol. 42, No. 3, pp. 741-843, 1995.

[23] Koike, H.; Yoshihara, H.: Fractal Approaches for Visualizing Huge Hierarchies, *Proceedings of the 1993 IEEE Symposium on Visual Languages*, pp. 55-60, IEEE/CS, 1993.

[24] Korber, B.; Hahn, B.; Foley, B.; Mellors, J. W.; Leitner, T.; Myers, G.; McCutchan, F.; Kuiken, C. L. (eds.): Human Retroviruses and AIDS: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 1997.

[25] Korber, B.; Muldoon, M.; Theiler, J. et al.: Timing the ancestor of HIV-1 pandemic strains, *Science*, Vol. 288, pp. 1789-1796, 2000.

[26] Lanciotti, R. S.; Roehrig, , J. T.; Deubel, V. et al.: Origin of the West Nile virus responsible for an outbreak of encephalitis in the Northeastern United States, *Science*, Vol. 286, pp. 2333-2337, 1999.

[27] Lathe, W. C. III; Burke, W. D.; Eickbush, D. G.; Eickbush, T. H.: Evolutionary stability of the R1 retrotransposable element in the genus Drosophila, *Molecular Biology and Evolution*, Vol. 12, pp. 1094-1105, 1995.

[28] Maier, D.: The Theory of Relational Databases, Computer Science Press, MD, 1993.

[29] McGuire, K.; Holmes, E. C.; Gao, G. F.; Reid, H. W.; Gould, E. A.: Tracing the origins of louping ill virus by molecular phylogenetic analysis, *Journal of General Virology*, Vol. 79, pp. 981-988, 1998.

[30] Meyer, Joerg: Interactive Rendering of Medical and Biological Data Sets; *Ph. D. thesis*; Department of Computer Science, University of Kaiserslautern, Germany, Shaker Verlag, September 29, 1999 (ISBN 3-8265-7009-X).

[31] Meyer, Joerg; Borg, Ragnar; Hamann, Bernd: VR-based Rendering Techniques for Time-Critical Applications; *Scientific Visualization 2000*, Schloss Dagstuhl, Germany, accepted for publication, May 22-26, 2000.

[32] Mindell, D. P.; Shultz, J. W.; Ewald, P. W.: The AIDS pandemic is new but is HIV new?, *Systematic Biology*, Vol. 44, pp. 77-92, 1995.

[33] Munzner, Tamara: Exploring Large Graphs in 3D Hyperbolic Space, *IEEE Computer Graphics and Applications*, Vol. 18, No. 4, pp. 18-23, July/August 1998.

[34] Nadeau, David R.; Baley, Michael J.: Visualizing Volume Data using Physical Models, *IEEE Visualization Conference*, pp. 49-50, 2000.

[35] Ou, C.-Y. et al.: Molecular epidemiology of HIV transmission in a dental practice, *Science*, Vol. 256, pp. 1165-1171, 1992.

[36] Van de Peer, Y.; Rensing, S. A.; Maier, U.-G.; de Wachter, R.: Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae, *Proc. Natl. Acad. Sci.*, Vol. 93, pp. 7732-7736, 1996.

[37] Van de Peer, Y.; de Wachter, R.: Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-ste rate variation in 18S rRNA, *J. Mol. Evol.*, Vol. 45, pp. 619-630, 1997.

[38] Rich, S. M.; Licht, M. C.; Hudson, R. R.; Ayala, F. J.: Malaria's eve: evidence for a recent population bottleneck throughout the world's populations of Plasmodium falciparum, *Proc. Natl. Acad. Sci.*, USA, Vol. 95, pp. 4425-4430, 1998.

[39] Rijk, P.; van de Peer, Y; de Wachter, R.: The rRNA WWW Server. Univ Of Antwerp, 1998.

[40] Robie, J.: The design of XQL, `http://www.texcel.no/whitepapers/xql-design.html`, 1999.

[41] Sanderson, Michael J.; Purvis, Andy; Henze, Chris: Phylogenetic supertrees: assembling the tree of life, *Trends in Ecology & Evolution*, Vol. 13, No. 3, pp. 105-109, March 1998.

[42] Semple, Charles; Steel, Mike: A supertree method for rooted trees, *Discrete Applied Mathematics*, Vol. 105, pp. 147-158, 2000.

[43] Silberschatz, A.; Korth, H. F.; Sudarshan, S.: Database System Concepts (Third Edition), McGraw-Hill, 1996.

[44] Snodgrass, Richard T.; Winslett, Winslett: UniSQL/X Unified Relational and Object-Oriented Database System, *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, Minneapolis, Minnesota, p. 481, May 24-27, 1994.

[45] Steel, Mike: The complexity of reconstructing trees from qualitative characters and subtrees, *Journal of Classification*, Vol. 9, pp. 91-116, 1992.

[46] Swofford, D. L.: When are phylogeny estimates from molecular and morphological data incongruent? in Miyamoto, M. M.; Cracraft, J. (eds.): *Phylogenetic analysis of DNA sequences*, Oxford University Press, New York, pp. 295-333, 1991.

[47] Teuhola, Jukka: An Efficient Relational Implementation of Recursive Relationships using Path Signatures, *Proceedings of the Tenth International Conference on Data Engineering*, Houston, Texas, pp. 348-355, February 14-18, 1994.

[48] Ullman, J. D.: Principles of Database and Knowledge-base Systems, Part I and II, Computer Science Press, 1988.

[49] Wade, C. M.; Lobidel, D.; Leigh Brown, A. J.: Analysis of human immunodeficiency virus type 1 env and gag sequence variants derived from a mother and two vertically infected children provides evidence for the transmission of multiple sequence variants, *J. Gen. Virol.*, Vol. 79, pp. 1055-1068, 1998.

[50] Wilkie, T. M.; Yokoyama, S.: Evolution of the G protein alpha subunit multigene family, in Fambrough, D. M. (ed.): *Molecular evolution of physiological processes*, Rockefeller Univ. Press, New York, pp. 249-270, 1994.

[51] Yuan, Yan P.; Eulenstein, Oliver; Vingron, Martin; Bork, Per: Towards detection of orthologues in sequence databases, *Bioinformatics*, Vol. 14, No. 3, pp. 285-289, 1998.

[52] Zuckerkandl, E.; Pauling, L.: Molecular disease, evolution, and genetic heterogeneity, in Kasha, M.; Pullman, B. (eds.): *Horizons in biochemistry*, Academic Press, New York, pp. 189-225, 1962.

[53] Zuckerkandl, E.; Pauling, L.: Evolutionary divergence and convergence, in Bryson, V.; Vogel, H. J. (eds.): *Evolving genes and proteins*, Academic Press, New York, pp. 97-166, 1965.