Pacific Symposium on Biocomputing 5:200-202 (2000)

MOLECULES TO MAPS: TOOLS FOR VISUALIZATION AND INTERACTION IN COMPUTATIONAL BIOLOGY

EILEEN T. KRAEMER
Computer Science Department
The University of Georgia
Athens, GA 30602

THOMAS E. FERRIN
Computre Graphics Laboratory
University of California
San Francisco, CA 94143-0446

Tools for visualization and interaction in computational biology attempt to assist scientists in understanding and correlating the often vast quantities of data produced both by direct observation of biological samples and as a result of further analysis or correlation of such data. Other tools and techniques are being developed to assist scientists in presenting their results to others in richer, more meaningful formats. This conference track explores novel tools and techniques in visualization and interaction to support scientists in exploring, navigating, and evaluating this data, and tools that assist scientists in presenting their results in a format that takes advantage of the latest advances in publishing technology, exploiting characteristics of human perception to effectively convey information to viewers.

The Chimera molecular modeling system, first presented at PSB '96, is an extensible molecular modeling application constructed using standard components. Two papers in this track present extensions to Chimera. Huang, et al describe three new computational methods, each with an associated visual interface, designed to assist scientists in the integration of sequence and structure. MinRMS generates a family of structural alignments based on optimal RMSD (root mean square distance) superpositions of alpha-carbon atoms in the structures being compared. AlignPlot provides a graphical representation of the RMSD values for each alignment in a set, allowing users to quickly identify areas of structures that are most similar and serving as a navigation tool between alignments. MSFViewer displays multiple alignments of related sequences and supports interactive highlighting of alignments. Overall, the goal of this work is to provide a comprehensive set of tools for generating and testing hypotheses about sequence, structure, and functional relationships among proteins whose sequence similarities often fall below the level of statistical significance but whose relationships are nonetheless biologically significant.

In the second Chimera-related paper, Konerding presents the Ensemble/ Legacy Chimera Extension. The Ensemble component provides an object-oriented interface for accessing and manipulating collections of molecular conformations that share a common chemical topology. The Legacy component provides an object-oriented interface for incorporating and driving existing programs dealing with these same molecular conformations.

Shah and Hunter present a tool designed to visually manage a large-scale data-mining project. Rather than visualizing the biological data, the goal of this work is to provide a unified framework for visualizing the preparations for, and results of, hundreds of machine learning experiments. The goal of the experiments is to improve the accuracy of enzyme functional predictions from sequence. Components of the system include a graphical user interface to define and explore training set data and alternative representations, and to inspect hypotheses and results. The machine learning approaches discussed include information-theoretic decision tree induction and naive Bayesian learning on local sequence domain representations of problematic enzyme function classes.

MappetShow, presented by Guyon, Vaysseix, and Barillot, works to give users a clear view of very dense genome maps and permit efficient comparison of several of these maps. The authors point out that genome maps have changed recently with the advent of maps with hundreds of markers per chromosome, the creation of single nucleotide polymorphism maps that contain dozens to thousands of polymorphic sites, and the enrichment of maps from map-poor genomes with data from map-rich genomes. The problem presented by these new maps is that it is difficult to visualize maps with hundreds or thousand of markers. The objects representing the markers tend to overlap in the visualization. Further, it is problematic to define a method of navigating and comparing such maps. The authors propose as a solution the non-linear view, a form of distorted view in which detail is provided at one area, the area in focus, and contextual information is provided for surrounding data, but in much less detail. In addition, the authors describe different types of distorted views and mapping functions from data to placement of objects in visualizations.

Automatic tools have been created to solve a number of problems in computational biology. However, we often discover that human interaction can improve the performance of some components of these tools. Such is the case with the construction of evolutionary trees. Kearney, Secord, and Zhang profess that many methods for constructing evolutionary trees reduce to the edge selection problem and then go on to show that non-interactive methods of edge selection often perform poorly and can conceal alternative solutions. They provide an interactive tool that allows the user to explore and interact with an

ordered list of edges, and show that such a tool can be useful in constructing evolutionary trees.

Pulavarthi, Chiang, and Altman address the arena of electronic publishing by increasing the range of formats available for conveying scientific data. In this paper, the authors introduce GEML, Graphical Explanation Markup Language. GEML is implemented as an XML format and builds on the standard PDB format. They provide a library of commonly used actions and also support user-defined extensions. Their goal is to provide a format for defining molecular documentaries that can take advantage of the interactive capabilities of electronic publishing. In addition, they present a generator tool, GEIS (Generator of Exploratory Interactive Systems), which takes a GEML file as input and produces all the files necessary for an interactive, web-based molecular documentary.

Finally, Hansen, Lodha, and Pang present a cutting-edge, but now affordable, approach to understanding protein structure-structure alignments through haptic perceptualization. In this approach, haptic (force feedback) devices are employed to add to the range of channels through which information may be conveyed to the user. The authors point out that technology has reached a point where such devices are affordable and software and GUIs for creating such force-feedback applications have been developed to substantially reduce the effort required by the programmer.

Together, these papers address a wide range of topics related to tools for computational biology, from finding visualizations that help correlate protein sequence data (Huang) and map data (Guyon), through tools for management of computational experiments (Shah) and interaction with algorithms (Kearney) to new data presentation formats (Pulavarthi) and new hardware (Hansen).