# Evolving Visual Metaphors and Dynamic Tools for Bioinformatics Visualization

**Panel Organizer:**
**Theresa-Marie Rhyne, North Carolina State University**

**Panelists:**
**Thomas H. Dunning Jr., MCNC/North Carolina Supercomputing Center**
**Gus Calapristi, Pacific Northwest National Laboratory**
**Chris North, Virginia Polytechnic Institute and State University**
**Donna Gresh, IBM T.J. Watson Research Center**

## Introduction:

This panel examines issues relating to establishing visual metaphors and building effective software tools for bioinformatics visualization. Our discussion focuses on three major issues:

(1) The potential of new visual metaphors to bring insight to unfamiliar biological or genomic data sets.

(2) The capabilities of bioinformatics visualization to stimulate collaborative and dynamic inquiry.

(3) The danger that visual metaphors can restrict rather than expand horizons for bioinformatics or other similar scientific discovery.

Bioinformatics poses a challenging domain for computer generated visualization techniques. For starters, is this a domain of scientific or information visualization? Typical scientific data sets have inherent spatial metaphors such as fluid flow in the human heart or chemical bonding of molecules. In contrast, exploration of cyberspace or searching of document repositories are traditional information visualization problems requiring the creation of a visual metaphor. How do we handle data mining of genomic information where the visual metaphor might not be obvious but the results can push forward scientific discovery?

Genomic-sequencing is a scientific domain with the complex challenge of evaluating its existing and developing new visual metaphors. Other biological disciplines are experiencing similar challenges. Our panel highlights these concerns by comparing and contrasting work underway at various centers and universities examining bioinformatics visualization issues today.

## Position Statements:

**Theresa-Marie Rhyne:**

Bioinformatics includes the application of efficient algorithms for manipulating data stored and processed on high performance computers to genomic or protein sequence data. Often bioinformatics software is text-based where letters and numbers are displayed as a sequence. Basic information visualization methods that apply simple color coding techniques to these text-based bioinformatics tools and develop graphical interfaces have evolved. Recently, information visualization researchers have applied hyperbolic projection and information mural methods to genomic sequence data. Web-based systems that support simultaneous, linked access to disparate databases and dynamic visual discovery among geographically dispersed researchers are under development. Clearly, information and statistical visualization techniques will continue to play an important role in bioinformatics research.

Perhaps scientific visualization and other computer graphics techniques can also be of assistance. John Avise, in his October 5, 2001 article in Science magazine, noted that metaphors of the genome have and will continue to develop and change as scientific research progresses. These metaphors range from the beads on a string notion to a social collective where DNA sequences exhibit intricate divisions of labor. Scientific visualizations and computer graphics methods can assist in the visual display of these concepts. Virtual and augmented reality systems can visually depict and immerse genetic researchers into the complex metaphors they wish to consider.

There is always a danger with visual metaphors. We can tend to believe them too much and perhaps restrict our horizons to new and different ways of thinking. This is true of any field of science including genomic discovery. As information and scientific visualization practitioners we have a responsibility to be aware of the power of our visual imagery and to continue to work closely with genetic

researchers and bioinformatics scientists as they explore their wealth of data.

John C. Avise, "Evolving Genomic Metaphors: A New Look at the Language of DNA", Science, Vol. 294, No. 5540 (October 5, 2001), pp. 86 – 87.

**Thom H. Dunning Jr.:**

The genomic revolution is producing mass quantities of data critical to a broad range of biological fields from taxonomy through agriculture to human health. This data includes genomic sequence data and protein structural data as well as data on metabolic and regulatory pathways, *etc*. The analysis and interpretation of this data presents challenges unlike any encountered in biology heretofore.

Molecular biology, especially mammalian molecular biology, is far more complex than envisioned just a few short years ago. Instead of hundreds of thousands of genes, the human genome contains tens of thousands of genes plus many alternate splicings of those genes to produce alternate sets of proteins as well as regulatory regions that control the expression of these genes. Even the great "deserts" of the genome sequence were found to contain much of interest—immigrant DNA, pseudogenes, and more. Genomics and its follow-on science, proteomics, is now poised for the next phase—understanding the time evolution of gene expression, production of proteins, and activation of pathways in a cell. This information will provide detailed insights into how diseased cells differ from healthy cells, how old cells differ from young cells, *etc*. It also ensures that the data explosion will continue, unabated, in biology.

Although visualization techniques are currently being heavily used to analyze the data from genomic research, it is clear that new approaches will be needed as the quantity and variety of data escalates. In addition, it is not at all clear that the visualization techniques currently being used are the most effective possible. Developing a new generation of visualization techniques appropriate for the analysis of genome-scale data will require close collaboration between computer scientists and molecular biologists, particularly those biologists focused on bioinformatics.

**Gus Calapristi:**

Traditional techniques for analysis of biological data, such as pair-wise comparison algorithms that find similarities among proteins, are not sufficient for today's voluminous, data rich environment. Today the researcher must work with multiple data types but still expects the precision and predictability of the traditional tool set. New algorithms and visual metaphors offer the promise of handling larger, more diverse data sets, but may be completely invalid if subtle details or perspectives related to the data are missed. For example, the cluster analysis on a set of proteins may "look" good, but is that because the feature extraction algorithm has selected statically interesting, but biologically insignificant features of the protein?

If the algorithms are correct, are the visualizations providing the utility needed by the researcher? Principle components based proximity plots generate two or three dimensional approximations of the multi-dimensional space and do a good job of exposing generalized gene expression behaviors. They enable relationships among entire genome(s) to be displayed on a single computer screen, but they do little to identify specific findings of value. Parallel coordinate plots clearly show specific genetic behaviors but quickly become unusable when displaying results for thousands of genes simultaneously. Since no single analysis or visualization approach can fully address the needs of today's bioinformaticist, new systems must carefully integrate wide-ranging visualizations, support multiple methods for processing data, and do it in a near real-time, interactive environment. Achieving the appropriate blend of processing algorithms, visual metaphors, and integration cannot be done without the expert knowledge of the biologist applied in concert with the visualization expertise of the computer scientist.

**Chris North:**

Bioinformatics is a rapidly evolving field. The fast pace of new technologies, new experimental methods, and new fields of endeavor make it extremely difficult to build supportive visualization tools. The explosion in data collection is resulting in continuously changing data formats and dynamic schemas. Data is often stored in ad hoc formats, and new types of data are constantly being added to the mix. Emerging data standards leave open-ended hooks for new data, and data storage centers modify database schemas regularly. Furthermore, data integration is the often critical to scientific discovery in bioinformatics. Biologists must analyze the larger context of the data to construct models of organism functions.

The design and development of visualizations is very dependent on the data and user tasks. As a result, visualization developers in bioinformatics often find that their visualizations are obsolete by the time they are implemented. While some visualizations have been created for specific smaller problems [1], these tend to provide only minimal or temporary value, while broader needs go unmet. Software companies in this field spend much of their effort on customizing solutions, rather than shrink-wrapped products. Visualization developers simply cannot keep pace with the rate of change.

At Virginia Tech, our approach is to enable extreme flexibility in visualization to match the variability in data

and tasks. Our research on Snap-Together Visualization enables end-users to quickly construct or modify multiple-view visualizations by snapping together components on the fly without programming [2]. Much like graphical database schemas enable users to manipulate data storage, graphical visualization schemas can enable them to manipulate data display as well. As data schemas evolve, visualizations can equivalently evolve (perhaps automatically). Similarly, broader frameworks or problem solving environments can be constructed to link visualization to other tasks in bioinformatics research [3][4].

[1] Tian, Y., Clement, M., Ellis, M., Steele, J., North, C., "Gene Expression Mural: Visualizing Gene Expression Databases", *IEEE Visualization 2001 WiP*, 2001.
[2] C. North, B. Shneiderman. "Snap-Together Visualization: A User Interface for Coordinating Visualizations via Relational Schemata", *Proc. ACM Advanced Visual Interfaces 2000*, pg. 128-135, 2000.
[3] Philip Isenhour, James Begole, Winfield S. Heagy, and Clifford A. Shaffer, "Sieve: A Java-Based Collaborative Visualization Environment," *IEEE Visualization '97,* October 1997.
[4] N. Ramakrishnan, L.S. Heath, R.G. Alscher, and B.I. Chevone, "Expresso - A PSE for Bioinformatics: Finding Answers with Microarray Technology", in *Proceedings of the High Performance Computing Symposium*, Advanced Simulation Technologies Conference, pages 64-69, Apr 2001.

**Donna Gresh:**

One of the most interesting characteristics of bioinformatics data, from a visualization point of view, is that it often includes a variety of sorts of data, some of which may have a spatial context (molecular structure, radiology images, sequence information, biological structures) and much of which does not. Some of the information may be numerical (gene expression levels, blood chemistry measurements, patient age) while others are often categorical (patient diagnosis, patient sex, experimental protocol). A key characteristic of many bioinformatics data sets is that there are often a large number of interrelated but diverse variables whose relationships are not necessarily well understood ahead of time. Thus typical scientific visualization techniques, such as coloring a continuous variable using some sort of colormap to see the variation is often of less value than many information visualization techniques, which have been created to browse and explore often unfamiliar and complex data. Using the right tools for each type of data, with interplay between them, has been a focus of our work in the Visual Analysis group at T.J. Watson over the past several years.

In order to better help create better tools for problem solving about multi-type data sets, we have worked directly with life science researchers. One application we created was for visualizing the results of simulations of proteins. This application included both standard scientific visualization presentations of the protein structure with linked views of residue angles and distance correlation matrices. We have also collaborated with biomedical engineers at Johns Hopkins University to create an application explicitly combining information visualization techniques and sci-viz presentations to better understand a simulation of the mammalian heart. We are currently working with a medical research team at Hadassah Hospital to visualize clinical data concerning bone marrow transplant patients using information visualization techniques. In each case we have found value in presenting multiple linked views which serve to better understand the relationships between the variables in the data, and are using this insight as our guidepost in developing a library of linked info-viz and scientific visualization components that can be integrated into bioinformatics applications.

## Biographical Sketches for Panelists:

**Theresa-Marie Rhyne: (tmrhyne@ncsu.edu)**

Theresa-Marie Rhyne is a multimedia and visualization expert in Learning Technology Service at North Carolina State University. As of January 2002, she began contributing to the NC BioGrid effort (http://www.ncbiogrid.org) under development at the North Carolina Supercomputing Center/ a division of MCNC. From 1990 - 2000, she was a government contractor (initially for Unisys Corporation (1990 - 1992) and then for Lockheed Martin Technical Services (1993 - 2000)) at the US EPA Scientific Visualization Center. She was the founding visualization expert at the Center. She was the Lead Conference Co-Chair for IEEE Visualization 1998 and the Past Conference Co-Chair for IEEE Visualization 1999. She serves on the Editorial Board of IEEE Computer Graphics & Applications and is a senior member of IEEE.

**Thomas H. Dunning Jr.: (thom.dunning@ncsc.org)**

Thom H. Dunning, Jr. joined MCNC on March 1, 2001 as Vice President for High Performance Computing and Communications. In that position he oversees the North Carolina Supercomputing Center and the North Carolina Research and Education Network for the University of North Carolina System. Dr. Dunning is also a professor in the Department of Chemistry at the University of North Carolina at Chapel Hill. Before joining MCNC, Dr. Dunning was Assistant Director for Scientific Simulation in the Office of Science at the U.S. Department of Energy, on leave from Pacific Northwest National Laboratory. In that position, he was instrumental in securing funding for a new DOE scientific computing program, *Scientific Discovery through Advanced Computing*

Dr. Dunning is a theoretical and computational chemist who has authored more than 100 scientific publications on topics ranging from advanced techniques for molecular calculations to computational studies of the spectroscopy of high power lasers and the chemical reactions involved in combustion. Dr. Dunning is a member of the American Chemical Society, and a Fellow of the American Physical Society and of the American Association for the Advancement of Science. He has served on numerous national advisory committees, including NRC's AFOSR Chemistry Review Committee (1987-1990), NSF's Advisory Committee for Chemistry (1991-3), and DOE's Council on Chemical Sciences (1996-9). He was the founding vice-chair of NRC's Chemical Sciences Roundtable (1996-1999).

Dr. Dunning received his B.S. in Chemistry in 1965 from the University of Missouri-Rolla and his Ph.D. in Chemical Physics from the California Institute of Technology in 1970.

**Gus Calapristi: (Gus.Calapristi@pnl.gov)**

Gus Calapristi is a software project manager at Pacific Northwest National Laboratory (PNNL). He recently directed the development and deployment of visualization and analysis software for the pharmaceutical and life sciences industries. He is now engaged in projects dealing with text visualization and analysis for the federal government. Prior to joining PNNL in 1996, Mr. Calapristi managed software engineering groups at Boeing Computer Services, Richland and was responsible for developing enterprise level business systems, technical data repositories, and environmental compliance systems. After receiving his undergraduate degree in Biology from Florida State University in 1980 he worked as software developer and database designer for Rockwell Hanford Operations.

**Chris North (north@cs.vt.edu)**

Chris North is an assistant professor of computer science at Virginia Tech. He is a member of the Center for Human-Computer Interaction, heads the Lab for Information Visualization and Evaluation, member of the Bioinformatics in Computer Science research group, and collaborates with faculty at the Virginia Bioinformatics Institute. He spent a year at the U.S. Bureau of the Census developing visualization tools for GIS and health statistics. He received his Ph.D. from the University of Maryland, College Park, where he worked at the Human-Computer Interaction Lab.

**Donna L. Gresh: (gresh@us.ibm.com)**

Donna Gresh is a visualization scientist in the Visual Analysis group at IBM's T.J. Watson Research Laboratory, where she has worked since 1990. From 1990 to 1998 she was one of the developers of IBM Visualization Data Explorer, a scientific visualization application that is now available in the open source community. Since 1998 she has explored the use of both scientific and information visualization techniques in a variety of application areas, including finance and biology. She earned her Ph.D. in Electrical Engineering from Stanford University in 1990, where her dissertation concerned properties of the Uranian rings as inferred from Voyager radio occultation measurements.