



## Data, They Have Lots & Lots of Data

The Human Genome Project, one of the largest single investigative projects in modern science, lasted 13 years. The raw data it generated — approximately equal to the information contained in 102 Libraries of Congress — catapulted scientists into a “Post-Genomic” world that some expect to last several generations.

“Completion of the Human Genome Project in 2003 marked the beginning of a ‘post-genomic’ era and *in silico* (computing-based) biology,” explains Dr. Sumeet Dua, lead investigator of the LA EPSCoR CyberTools Information Services and Portals team and Louisiana Tech University Upchurch Endowed Professor of Computer Science.

“Coupled with the advances in automated data collection technologies, it has led to an unprecedented growth in the size, complexity, and quantity of collected data, a large proportion of which is currently inaccessible for analysis by computational scientists.

“Yet computational challenges of translating this knowledge base into meaningful diagnostics, innovative new therapies, and associate treatment methodologies have just begun.”

### Enter NSF EPSCoR

Taking on the challenges are Dr. Dua and his team of six researchers from Louisiana Tech’s Data Mining Research Laboratory and LSU Health Sciences Center in the New Orleans’ Eye Center. Together, they have developed a number of unique algorithms — sequences of logical instructions to achieve a specific computer programmatic goal — as well as four CyberTools with tutorials, and new international collaborations (See *About the NSF-Funded CyberTools Project*, page 2).

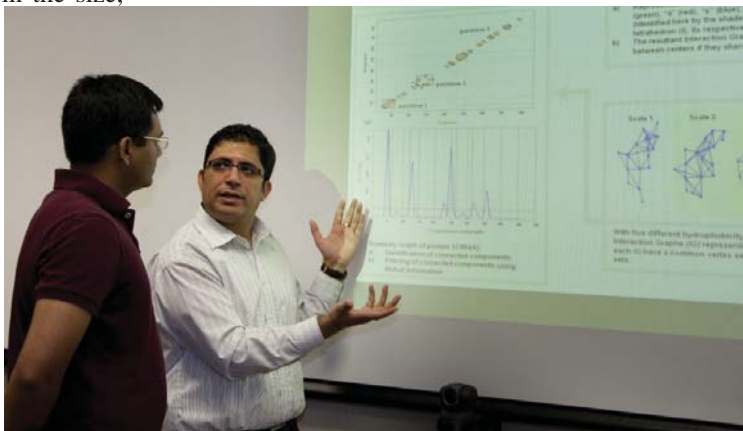
### What Challenges?

“Biology is on the verge of becoming one of the most data-rich disciplines in science, providing an enormous opportunity

for leading computer scientists in mining information ‘nuggets’ in the tsunami of genomic data that is being generated,” says Dr. Dua.

Discovering protein structures and assigning them functions is a computationally intensive task. Consider that... there are more than 500,000 known protein sequences, and the number is growing.

Defining data mining as a prospective science focusing on the delivery of previously unknown relationships among



Dr. Sumeet Dua, right, a CyberTools lead investigator, and Post Doctoral Fellow Dr. Pradeep Chowriappa discussing a Cyber Tool design.

existing data, he adds: “The rapid growth in technology and science has increased the complexity of the relationship between computer science and biological research, particularly in the field of genetics.”

The interdisciplinary field of Computational Biology and Bioinformatics is utilized to provide computer-based methods for coping with and interpreting the vast data that researchers encounter in the biological sciences.

Bioinformatics is the science of storing, extracting, analyzing, interpreting, and

utilizing information from biological databases, large unorganized bodies of life sciences data.

Identifying genes in DNA sequence is high on the list of computational biology challenges. The smallest sequence of molecules connected in a strand of DNA that stores information about heredity, genes carry the information that instructs other cells within the body on how to function properly. They are also responsible for coding the proteins that are essential to every aspect of our physiology.

### In Search of the Key

“The molecular sequence of the gene itself holds the key to this ‘code.’ If captured, it will help reveal the structure of the protein that it will build, a major factor in defining the protein’s function,” says Dr. Dua. “Ultimately, if the parent gene of each protein can be discerned, we may be able to treat diseases and improve human health on a sub-molecular level.”

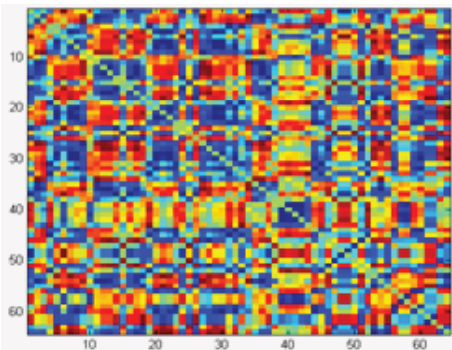
He notes, for example, that protein complexes have been implicated in such genetic disorders as Alzheimer’s and Cystic Fibrosis.

“Although the interpretation of genome data is still in its initial stages, genetic tests also indicate a predisposition to a variety of other illnesses, including breast cancer, bleeding disorders such as

hemophilia, and liver diseases. This has driven researchers to further investigate proteins associated with human disease.”

Discovering protein structures and assigning them functions is a computationally intensive task. Consider, for example, that more than 600 properties exist per amino acid (AA), the smallest unit of protein sequence...that proteins can be several thousand AA long...that there are more than 500,000 known protein sequences, and the number is growing.

*Data, Continued on page 2*



Map-color coded representation of complex residual interactions of a protein sequence.

## Data Continued

Dr. Dua and his research team have developed unique algorithms that will efficiently and accurately fuse these properties to generate new synthetic sets of properties with greater specificity and sensitivity in the characterization of protein structures and sequences.

In one algorithm, a unique approach was developed that provides insights into the protein's functional (operative) form. They have also designed a CyberTool to successfully predict conformational changes in proteins.

In another study, they created a unique Protein Map, a succinct visual and interactive 2-D representation of the complex intra-protein interactions for comparative mining and discovery.

Using the protein map, a CyberTool was developed and made available to the broader scientific community. A data mining tool that analyses the relationships between the physico-chemical properties of the protein and relates them to its structure has also been developed.

"While we are intuitively aware of the constantly changing nature of proteins, we

have little understanding as to how they drive structural flexibility. We have thus developed a tool that paves the way for us to identify local sequence modulations that impact protein function without changing the protein structure," adds Dr. Dua.

The CyberTool features data mining algorithms for classification and prediction and connects to known datasets and databases such as the worldwide Protein Data Bank (PDB) repository.

Additionally, a tool training session presented at an international workshop resulted in new international collaborations that are underway for the further development of the algorithmic tools.

## Leveraging Resources

The LA EPSCoR project leverages the supercomputing resources provided by the Louisiana Optical Network Initiative (LONI) to mine the vast, heterogeneous and rapidly growing PDB. A detailed tutorial that enables a novice to understand and execute the tools was also developed.

"This project is a good example of the CyberTools axiom that cyberinfrastructure development must be guided by scientific questions and, conversely, the scientific strategies must include advanced cyberinfrastructure," adds Dr. Michael Khonsari, LA EPSCoR Project Director and Board of Regents Associate Commissioner for Research and Development.

"In this particular case, a computational researcher is provided with an accurate understanding of a biological research project and the questions that it raises. This enables him or her to develop computational tools specifically designed for the biologist and the investigations being undertaken.

"It's a Post-Genomic collaboration at its best."

## About the NSF-Funded CyberTools Project

The National Science Foundation EPSCoR program awarded LA EPSCoR a \$9 million Research Infrastructure Improvement (RII) Grant in 2007. With matching funds of \$3 million from the Board of Regents Support Fund and \$3.2 million from the nine participating institutions, the three-year grant total is over \$15.2 million.

The focus of the LA EPSCoR CyberTools grant is on cyberinfrastructure — the technology and network systems that have infiltrated every aspect of today's modern world — specifically on the development of Cyber Tools to help investigators effectively utilize cyberinfrastructure at a level that would not otherwise be possible.

CyberTools don't fit neatly in a kit or a box. You can't even touch them. In the right hands, however, they create the infrastructure required to support current and future discoveries in science and engineering.

The innovative collection of cyber services and computational toolkits being developed by the NSF EPSCoR RII project are enabling Louisiana researchers in many disciplines to participate in worldwide grid computing — a network of many separate computers addressing large-scale computational problems.

Participating institutions are: Louisiana State University, LSU Health Sciences Center-New Orleans, Louisiana Tech University, Southern University-Baton Rouge, Tulane University, Tulane Health Sciences Center, University of Louisiana-Lafayette, University of New Orleans and Xavier University.