VBI e_Connections

feature technology in the news iGEM

Vol. 2 No. 3 Fall 2007 Newsletter of the Virginia Bioinformatics Institute at Virginia Tech

In search of exponential knowledge

The Virginia Bioinformatics Institute (VBI) hosted its second annual research symposium on September 6-7 at the Mountain Lake Hotel, Pembroke, VA. The event, which focused on research arising from VBI's team-oriented approach to science, included a host of talks as well as poster presentations describing VBI research projects.

In This Issue

Olga Troyanskaya, Assistant Professor at Princeton University's Department of Computer Science and Lewis-Sigler Institute for Integrative Genomics, NJ, was the keynote speaker at the symposium. During her presentation, Troyanskaya talked extensively about addressing the data-knowledge disconnect in bioinformatics as well as three data integration projects underway in her laboratory – SPELL, bioPIXIE and MEFIT.

"We're all aware of the exponential growth in data from large-scale biology. For example, microarray experiments have provided a deluge of information over the past ten years or so. However, we have not witnessed an exponential growth in knowledge to match this output. We need to address this data-knowledge disconnect through bioinformatics," Troyanskaya said.

Troyanskaya's view of the state-of-play of research is a four-way cycle of experiments, data, biological analysis and hypotheses. However she believes this cycle is currently blocked by a lack of large-scale feedback from computational methods back to experimental biology. She commented: "We need to integrate the analysis of diverse data sets, computational analysis of results as well as improve the accessibility and usefulness of this information to the scientific community. I believe computation can make a big difference and do something very real for biology."

Troyanskaya's laboratory is developing SPELL, bioPIXIE and MEFIT as tools to bridge the gap. SPELL is a "Google for microarray data" that allows the user to identify other genes related to a particular search. bioPIXIE looks beyond gene expression data and provides useful information on potential interaction networks and pathways for a particular gene or set of genes. MEFIT takes microarray data and predicts the likelihood of functional relationships between a gene pair in the context of a specific biological function.

Olga Troyanskaya

Recently Troyanskaya has been combining the three systems – SPELL, bioPIXIE and MEFIT – to make and experimentally verify novel predictions. So far, this approach has been more successful in finding moderate phenotypes or features. She said: "Many biologists may worry that we have discovered a tool that recapitulates the gene ontology or the functional assignment of a particular gene. However I think the results so far show that computational approaches are not just confirmatory – they can make a big difference and provide real biological insight. We're finding novel genes and we're also discovering under-annotations in the gene ontology. I would really like to see the gene ontology community adopt our tools to help facilitate their work."

Troyanskaya is beginning the second iteration of her work, which should provide an even more rapid, accurate and comprehensive characterization of particular biological processes. This could well be a step in the right direction in the quest for exponential knowledge.

VBI's 2nd annual research symposium

- Took place on September 6-7, Mountain Lake Hotel, Pembroke, VA
- 25 oral presentations from students, post-graduates and other scientists working at VBI
- Over 50 poster presentations on topics spanning infectious diseases, systems biology, bioinformatics, genomics, metabolomics, and proteomics
- Keynote presentation from Olga Troyanskaya, Assistant Professor at Princeton University's Department of Computer Science

Bioinformatics and the semantic web: distilling knowledge by mining heterogeneous data

In this article, Konstantinos Krampis, a graduate research assistant in Dr. Brett Tyler's laboratory at VBI, discusses some of the latest developments in applications for semantic web technologies. The semantic web is an evolving extension of the World Wide Web in which web content can be expressed not only in natural language, but also in a format that can be read and used by software agents, thus permitting them to find, share and integrate information more easily.

Within the field of bioinformatics, technologies have evolved to the point where storing and querying large amounts of biological data no longer pose a problem. The main conundrum instead, is related to integrating different types of data for efficient reasoning and being able to distill knowledge by looking at the whole picture of the scientific problem.

An open world

The power of the semantic web comes from the simple and universal data model that lies at its core – the Resource Description Framework (RDF). For bioinformatics, resources like genes, proteins, metabolites, diseases and experiments, can all be described with RDF as data objects with property values. A protein and a metabolic compound, measured in different experiments and their data stored in RDF with completely different properties, can still be causatively associated with a single property from each showing participation in or alteration of the same physiological process. Lacking the inflexibility of relational database schemata, the open-world data model of the semantic web can easily bring together biological data generated under different settings. A causative association revealed from within the data as described in the above example might point to a potential drug target for a metabolic compound. Besides this type of simple inference, more sophisticated machine-performed reasoning is possible in the semantic web, by using the Ontology Web Language (OWL) to design ontologies for classifying knowledge from a domain. OWL is built on top of RDF, allows for logical union, intersection and disjointness between data objects and also includes transitive, symmetric or inverse properties for describing them. This, in combination with automated reasoning, reveals more complex than binary associations between genes, proteins or metabolites, and exposes interrelations over a broader range in the dataset.

BioPAX: semantic web for pathway data

One of the early adopters of RDF/OWL technologies in the life sciences domain was the Biological PAthway eXchange (BioPAX, www.biopax.org) consortium, which agreed upon an ontology-standard for knowledge sharing. The BioPAX ontology is designed to not only include data objects for the metabolic reactions assembling the pathways. By drawing on the power of RDF, it weaves together in one interconnected graph data for all types of cellular apparatus making the pathways functional, such as catalyzing enzymes, genes that encode them, and even all their associated references in external databases. For the biologist this translates to being able to examine simultaneously all the biological entities participating in the pathway of interest, by looking at data from disparate experiments integrated within the same instance of the ontology. New data can be superimposed on the pathways, by getting for example a list of up-regulated genes from a microarray experiment and overlaying them to the BioPAX instance to see which pathways were activated during the experiment.

In another scenario, a user of BioPAX might choose to compare between species, by intersecting their pathways based on identity of only genes, proteins, metabolic reactions, or all of the above. Since the BioPAX data are encoded in an RDF graph, for the pathway intersection (or for the microarray data overlay described in the previous paragraph) the RDF Query Language (RQL) can be used, which is as efficient as its relational counterparts. One of the strengths of this approach comes from the fact that the topology of the pathways is natively represented in the structure of the RDF graph. Therefore, RQL queries can even reveal structures such as loops or cross-links, which may be shared at key points in the metabolic networks of the species under comparison. Last but not least, the user of BioPAX can take advantage of the machine reasoning capabilities embedded in OWL, by defining new classes and axioms within the ontology in order to reveal data objects that accord to certain patterns.

Momentum

The RDF/OWL technologies have gained wide acceptance, and many consortia choose to publish their data in such formats. Examples are the Gene Ontology, the National Cancer Institute, Microarray Gene Expression Data (MGED), and many more as the Open Biomedical Ontologies (OBO, http://obofoundry.org/) website demonstrates. RQL databases are easy enough to use, and their functionality to query heterogeneous data from different types of biological entities is unparalleled in any other querying and data storing technologies.

VBI e_Connections

VBI e_Connections is a quarterly publication of the Virginia Bioinformatics Institute produced by the Public Relations team. The newsletter includes feature articles, technology updates as well as interviews that may be of interest to VBI's audiences. Contributions are welcomed.
Please direct submissions to newsletter-editor@vbi.vt.edu
Newsletter team: Susan Bland: Editor; Barry Whyte: Editor; June Mullins: Graphic design.
For further information, please contact:
Barry Whyte at Tel: 540-231-1767, email: whyte@vbi.vt.edu
Website: www.vbi.vt.edu

Largest Ever Affymetrix Genechip® Plant Microarray Experiment

PR Newswire
United Business Media

BLACKSBURG, Va., May 25 /PRNewswire-USNewswire/ - The Virginia Bioinformatics Institute (VBI) at Virginia Tech today announced that it has completed the largest ever Affymetrix GeneChip® microarray study for a plant experimental system in an academic research setting. The 2600-chip experiment explores the counter-play of plant and pathogen genes during infection of soybean with the root-rot pathogen *Phytophthora sojae*, with a focus on mechanisms of long-lasting disease resistance. *P. sojae* causes severe damage in soybean crops and results in $100-200 million annually in losses for commercial farmers in the United States alone.

Much information has been obtained by careful study of single resistance genes in plants. However, this "low hanging fruit" approach has not resulted in significant long-lasting resistance of crops. Plants in which a single gene has been modified are quickly overcome by new strains of pathogens. The GeneChip experiment is part of a project aimed at understanding and improving a more long-lasting form of disease resistance called quantitative or multigenic resistance.

Source: PRNewswire May 25, 2007

Biochemical Profiling Research Group Implements New Software for Metabolomics

Blacksburg, USA - Researchers at the Virginia Bioinformatics Institute (VBI) at Virginia Tech are examining ways in which metabolomics can be applied to the study of systems biology. Recently, they invested in new technology, ACD/IntelliXtract, to further enhance their ability to handle the data obtained by liquid chromatography/mass spectrometry (LC/MS).

Metabolomics involves the systematic study of the metabolic processes of living cells and requires the high-throughput analysis of a large number of small-molecule cellular metabolites. While current instrument technology enables samples to be analyzed more and more quickly, the resulting avalanche of data must then be managed - often an overwhelming task.

Source: SpectroscopyNow.com August 30, 2007

Calendar of Events

VBI's calendar of events is available at: www.vbi.vt.edu

VBI Scientific Publications

Metabolic footprinting: a new approach to identify physiological changes in complex microbial communities upon exposure to toxic chemicals

Henriques ID, Aga DS, Mendes P, O'Connor SK, Love NG.

Environmental Science & Technology 2007: **41**: 3945-3951.

Researchers at the Virginia Bioinformatics Institute and the Department of Civil and Environmental Engineering at Virginia Tech have used metabolic footprinting coupled with statistical analysis to look at multiple, chemically stressed activated sludge cultures. The aim was to identify probable biomarkers that indicate community stress. The impact of cadmium (Cd), 2,4-dinitrophenol (DNP), and *N*-ethylmaleimide (NEM) shock loads on the composition of the soluble fraction of activated sludge cultures was analyzed by gross biomolecular analyses and liquid chromatography-mass spectrometry (LC-MS). The footprints contain information about specific biomolecular differences between the stressed samples. Since the experiments were conducted with mixed liquor from four distinct wastewater treatment plants, the discriminant *m/z* ratios may potentially be used as universal stress biomarkers in activated sludge systems.

Improvement of water use efficiency in rice by expression of *HARDY* an *Arabidopsis* drought and salt tolerance gene

Karaba A, Dixit S, Greco R, Aharoni A, Trijatmiko KR, Marsch-Martinez N, Krishnan A, Nataraja KN, Udayakumar M, Pereira A.

Proceedings of the National Academy of Sciences, 2007, available in advance on-line at
http://www.pnas.org/papbyrecent.shtml

An international team of scientists has produced a new type of rice that grows better and uses water more efficiently than other rice crops. Professor Andy Pereira at the Virginia Bioinformatics Institute has been working with colleagues in India, Indonesia, Israel, Italy, Mexico and The Netherlands to identify, characterize and make use of a gene known as *HARDY* that improves key features of this important grain crop. The research shows that *HARDY* contributes to more efficient water use in rice, a primary source of food for more than half of the world's population.

iGEM teams meet at VBI

Virginia Tech's iGEM team welcomed two other regional teams to the Virginia Bioinformatics Institute (VBI) on July 26 to share experiences and ongoing work for the upcoming national iGEM competition.

iGEM is the prestigious International Genetically Engineered Machines competition organized by the Massachusetts Institute of Technology (MIT). The objective of the competition is to design and build an engineered biological system using standard DNA parts.

Teams of undergraduates from the University of Virginia in Charlottesville, Va. and Davidson College in Davidson, NC traveled to Blacksburg to meet with Virginia Tech's team. Group members gave overviews of their current work and project goals and were given opportunities to ask questions, as well as spend time with other teams to discuss their experiences. One of the objectives of the meeting was to acclimatize the students to the competitive nature of the iGEM project and make newcomers aware that their work will be judged versus other team efforts.

While the Davidson College iGEM team has participated in the iGEM competition before, this is the first year that Virginia Tech and the University of Virginia have had iGEM teams. In fact, after learning that Virginia Tech had formed a team, an engineering student at the University of Virginia spearheaded the effort to create the school's own team.

The purpose of the competition is to help test the idea that biological engineering can be performed more reproducibly through the use of standardized parts, which, in the world of iGEM, are called BioBricks. These parts are provided by the iGEM competition to students, who use them to design and build genetic machines. Team members can also make their own parts and even combine their creations with BioBricks already available to form an endless number of complex systems.

Each team that visited VBI for the regional meeting presented their ongoing iGEM work, discussing their successes and challenges faced. They spoke about specific issues they have addressed throughout the process, including how they decided on a particular project, the different approaches that have been considered, and what the end results of the project could be.

Virginia Tech team

The Virginia Tech iGEM team is focusing on engineering and disease epidemics for their project, examining the development of an epidemic within and between populations. Their main goal is to create a population interaction model that predicts the spread of infection between groups of people. They are using *Escherichia coli* and bacteriophage lambda as a model population to understand what happens from the very beginning of a disease epidemic. The team will then design a network for the spread of infection and create models and verify them experimentally. According to team members, *E. coli* is being used because the effect of low-level parameters can be analyzed and experiments can be repeated.

University of Virginia team

The University of Virginia team is interested in the use of photobiological interfaces for the input and output of engineered biosystems. The team describes a cell in computing terms, with the genome representing its "Operating System" and BioBricks serving as individual software applications. Their goal is to find monitoring and input technologies for the "computer system," and they are looking specifically at electromagnetic radiation as a monitoring tool and input device, comparable to a computer's keyboard or mouse.

Although the team has done some preliminary work on several possible iGEM projects, they decided to concentrate their efforts on a project involving butanol biosynthesis. This project involves biofuel manufacturing to provide for the real-world need of alternative fuels. The team's goal is to create butanol in *E. coli* using agricultural waste as a starting point.

Davidson and Missouri Western team

Davidson College students are teaming with students from Missouri Western State University to create a collaborative iGEM team. Students from Davidson College visited VBI to present an overview of its project, which involves building an *E. coli* computer capable of solving the Hamiltonian path problem. The team hopes to manipulate *E. coli* into a mathematics problem solver, providing the large amount of processing power needed. The problem is the result of the discovery of a unique Hamiltonian path *in vitro* for a particular directed graph on seven nodes. The team hopes to make progress towards the solution of the problem *in vivo*. They are using a Hin/hixC DNA recombination mechanism to randomly generate possible paths through the graph. Gene expression and fragment length are then used to screen for a Hamiltonian path.

Until we meet again

The next time the three iGEM teams meet, it will be in November when they present their completed projects at the 2007 iGEM Jamboree in Boston. Each team hopes to donate any new BioBricks they create to the MIT Registry of Standardized Biological Parts. In addition, the Virginia Tech iGEM team members are considering future publication possibilities based on their work. The University of Virginia team is also developing material for a new course—"Biological Systems Design Seminar"—based on the work from its current project. Team members hope this course will help generate new project ideas, giving a substantial headstart for future iGEM teams.