DIMACS Center
Rutgers University

Workshops on Information Processing in the Biological Organism

Annual Report

July 2005

Participants who spent 160 hours or more

PI: Fred Roberts, DIMACS

Other Participants

Ron Levy, BioMaPS, Rutgers University

Wilma Olson, Center for Molecular Biophysics and Biophysical Chemistry, Rutgers University

Eduardo Sontag, BioMaPS, DIMACS, Rutgers University

Workshop on Information Processing in the Biological Organism (A Systems Biology Approach)
November 4-5, 2003

Organizers:
Fred S. Roberts, Chair, DIMACS / Rutgers University
Eduardo Sontag, Co-chair, Rutgers University

Workshop on Biomolecular Networks: Topological Properties and Evolution
May 11 - 13, 2005

Organizers:
Cenk Sahinalp, Case Western
Petra Berenbrink, Simon Fraser

Workshop on Information Processing by Protein Structures in Molecular Recognition
June 13 - 14, 2005

Organizers:
Bhaskar DasGupta, University of Illinois at Chicago
Jie Liang, University of Illinois at Chicago

Workshop on Detecting and Processing Regularities in High Throughput Biological Data
June 20 - 22, 2005

Organizer:
Laxmi Parida, IBM T J Watson Research

Other Collaborators

Petra Berenbrink, Simon Fraser, Co-Organizer, Workshop on Biomolecular Networks: Topological Properties and Evolution

Bhaskar DasGupta, University of Illinois at Chicago, Co-Organizer, Workshop on Information Processing by Protein Structures in Molecular Recognition

Rebecka Jornsten, Rutgers University, Co-Organizer, Workshop on DNA Barcode of Life

Jie Liang, University of Illinois at Chicago, Co-Organizer, Workshop on Information Processing by Protein Structures in Molecular Recognition

David Madigan, Rutgers University, Co-Organizer, Workshops on DNA Barcode of Life

Laxmi Parida, IBM T J Watson Research, Co-Organizer, Workshops on Detecting and Processing Regularities in High Throughput Biological Data

Cenk Sahinalp, Case Western, Co-Organizer, Workshop on Biomolecular Networks: Topological Properties and Evolution

Eduardo Sontag, Rutgers University, Co-Organizer, Workshop on Information Processing in the Biological Organism

Partner Organizations

Telcordia Technologies: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

AT&T Labs - Research: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

NEC Laboratories America: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Lucent Technologies, Bell Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Princeton University: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Avaya Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

HP Labs: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

IBM Research: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

Microsoft Research: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

Stevens Institute of Technology: Collaborative Research; Personnel Exchanges

Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Activities and Findings

This grant is for a workshop in Bethesda on Information Processing in the Biological Organism and subsequent preparation of recommendations for future initiatives in this area. The subsequent discussions among moderators led us directly to the idea of running a special focus at DIMACS on Information Processing in Biology, organized around a series of workshops. That special focus is partially funded by a grant on Information Processing in Biology (04-32013) and is a natural follow-up of the Bethesda meeting. Since these workshops were only partially funded, we used some of the leftover funds in this grant to increase the support for some of the workshops in this special focus. This also gave us an opportunity to have some of the Bethesda moderators join us for continuing discussions of the recommendations coming out of the Bethesda meeting. Moreover, in the present year, discussions with the moderators and others have led us to another topic, DNA barcoding, a rapid, cost-effective way to identify plant and animal species using a very short (~650 base pair) gene sequence from a standard region of the genome. The "Consortium for the Barcode of Life," based at the Smithsonian, is establishing a database inside GenBank at NIH, where hundreds of thousands of sequences will be deposited. Each sequence will be linked to the museum voucher specimen from which the sequence was obtained, and the species name attached to that specimen by a taxonomist. Unknown specimens can then be queried against these authority records. This system has tremendous potential to address applied problems in agriculture, pest control, infectious diseases, food quality, border inspections, etc. We have gotten involved with this project through David Schindel and in particular in its Working Group on Data Analysis, which is exploring the alternative ways the database could be used and should not be used. We have already identified some very exciting research issues in this area. We will be to using some of the remaining funds to continue interactions with this working group and to hold a series of workshops on the topic here at DIMACS. Plans are already underway to do so.

Workshop on Information Processing in the Biological Organism (A Systems Biology Approach)
Dates: November 4 - 5, 2003
Location: Four Points Sheraton, Bethesda, Maryland
Organizers, Fred S. Roberts, Chair, DIMACS / Rutgers University and Eduardo Sontag, Co-chair, Rutgers University
Attendance: 86

The workshop investigated information processing in biological organisms from the general point of view of systems biology. Traditional biological research has aimed to understand isolated parts of a cell or organism. It has achieved dramatic technological breakthroughs in understanding genes and proteins. The potential for dramatic new biological knowledge arises from investigating the complex interactions of many different levels of biological information. This is the heart of the systems approach to biology. It is aimed at studying genomic DNA, mRNA, proteins, and informational pathways and informational networks in conjunction, looking for "system-level" understanding. Systems biology requires an understanding of the basic system structures (the networks of gene interactions and biochemical pathways) and it requires an understanding of the dynamics of systems - how they change over time through metabolic changes, modifications in biochemical makeup, etc. Understanding the biological systems from this point of view can be greatly aided by the use of powerful mathematical and computer models. In turn, the systems understanding of biological systems can provide insights that might be useful for computer and information science.

Information processing is a key aspect of biological systems and the workshop concentrated on this aspect of systems biology. In general terms, the workshop sought to answer the following questions:

(1). What model systems concerning information processing in the biological organism are of broad utility to systems biology?
(2). What are the mathematical foundations relevant to the systems biology approach to the study of information processing in biological organisms, and specifically what algorithms are of broad and central utility for this topic?
(3). What is the next generation of reference resources needed for research in this field?
(4). How can the study of selected information processing processes within biological systems inform other disciplines, including computer science?

The workshop was organized around four main topics:
(a). Genetics to gene-product information flows, including temporal and spatial aspects.
(b). Signal fusion within the cell.
(c). Cell-to-cell communication.
(d). Information flow at the system level, including environmental interaction.

The following workshops were also supported by the DIMACS/BioMaPS/MB Center Special Focus on Information Processing in Biology:

Workshop on Biomolecular Networks: Topological Properties and Evolution
Dates: May 11 - 13, 2005
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Cenk Sahinalp, Case Western and Petra Berenbrink, Simon Fraser
Attendance: 102

The functioning of a biological system largely depends on the mutual interactions among its constituent components such as proteins. It is a common practice to represent such a system by a network, within which objects are represented as nodes and relations are represented as edges linking related pairs of nodes. A biological network is broadly defined as any network (graph) where the nodes are identified with some biologically relevant entities and edges define a relation over these entities. For example, in a protein-protein interaction (PPI) network nodes correspond to proteins and edges to interactions between them; in a metabolic network nodes usually correspond to metabolites and edges to reactions; yet another biological network may be used to describe co-occurrence of protein domains within proteins.

The structure of these biological networks resembles that of many other natural networks such as the world wide web (WWW) graph, where each vertex is a web page, and each edge is a hyperlink from one web page to another. The PPI and the WWW networks both exhibit a power law degree distribution and a small diameter. This is very different from networks generated by standard random graph models which are static and do not have power law degree distribution. Furthermore, the growth of both networks can be attributed to mechanisms of node duplication. Thus recent work on structural properties and evolution of the PPI network in conjunction with that of the WWW network has attracted considerable attention.

This workshop brought together researchers from diverse backgrounds who work on evolution and the structural properties of biological networks and how these properties relate to those observed in other natural networks. The workshop included talks on state of the art of and open questions in the following aspects of biological networks:

- Computational and experimental methods for discovering biological networks
- Evolutionary models for biological networks and their relationship to the models for WWW graph and other natural networks
- Combinatorial and statistical properties of biological network and their implications

- The structure of motifs and motif finding in biological networks

Leading specialists in the field gave invited presentations. There was a poster session and we invited poster contributions.

Workshop on Information Processing by Protein Structures in Molecular Recognition

Dates: June 13 - 14, 2005
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Bhaskar DasGupta, University of Illinois at Chicago, Jie Liang, University of Illinois at Chicago
Attendance: 32

Biological processes in cells are based on specific molecular recognitions, which triggers cascade of biological responses. The physical basis of complex network interaction is the three-dimensional structure of proteins and their functional regions. Understanding how information encoded in these biomolecules is recognized and processed by the interacting partners is a fundamental problem of biology.

In this workshop we discussed the development of algorithms for discovery of spatial patterns important for recognition, for uncovering deep evolutionary relationship of proteins, for predicting binding partners, and for simulating the protein-protein and protein-DNA recognition process. Specific topics of interest included protein-ligand and protein-protein binding site prediction, functional prediction of proteins with known structures but unknown functions, protein-protein interactions and docking, prediction of immune epitope, design of peptide modulators of protein-protein interactons, protein substructure matching, and evolution of structural biopattern. We hope further development in these areas will formulate new research problems and motivate new algorithms in combinatorics, optimization, discrete mathematics, mathematical programming, and additional areas.

Workshop on Detecting and Processing Regularities in High Throughput Biological Data

Dates: June 20 - 22, 2005
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Laxmi Parida, IBM T J Watson Research
Attendance: 66

The biological community is being inundated with a large amount of data and understanding this data is lagging behind the process of acquiring it. It is believed nature has left vital clues hidden in this data and there is a need for techniques and methodologies to work effectively in detecting these. Biological information processing exploits these regularities to gain understanding of the underlying model or phenomenon. For example, in its simplest form regularity could be repetition of functional or structural domains in a protein sequences or co-expression of genes in microarrays. When the data is in terms of networks, either representing protein-protein interactions or metabolic pathways, topological motifs tell a tale that will be fundamental in understanding the working of a biological system. The workshop contributed significantly to the research effort by bringing together researchers from the many different groups engaged in biological projects having the study of regularities in the data as an underlying theme.

Proposed Series of Workshops on DNA Barcoding

DNA barcoding has been proposed as a tool for differentiating species. Barcoding is based on the assumption that certain short gene regions evolve at a rate that produces clear interspecific sequence divergence while retaining low intraspecific sequence variability. The cytochrome c oxidase subunit 1 mitochondrial region ("COI") has emerged as the most likely barcode region for most eukaryotic animals. As the data bases with specimens that have been partially sequenced (e.g. COI region) grows, there is a

great need to develop more sophisticated analysis tools to understand how these data can be used to aid in species classification and species discovery. The Consortium for the Barcode of Life (CBOL; see www.barcoding.si.edu) is an international consortium of about 70 Member Organizations from six continents and more than 35 nations. These include natural history museums, herbaria, biodiversity and conservation organizations, university departments and other research organizations, government agencies and private sector companies. CBOL is devoted to exploring and developing the potential of DNA barcoding to become a tool for taxonomic research and for applications of species-level data to applied problems such as conservation, crop protection and sustainable development. To stimulate interaction between researchers in various fields already involved in DNA barcoding, and to bring attention to this emerging field of research outside the barcoding community, DIMACS will host a series of workshops for the CBOL Data Analysis Working Group (DAWG), initiate a Statitics/Mathematics Component (SMC) to DAWG, and develop a research agenda for SMC. Included in SMC will be researchers in machine learning, computer science, genomic research and statistics. The aim is to bring together experts in the above fields to identify the types of analytical, interpretive and display tools needed for the optimal treatment of DNA barcode data. In Spring of 2005 DIMACS members met with the CBOL steering committee members on several occasions, and Rebecka Jornsten (a DIMACS member) participated in the first international conference for the Barcode of Life and is now an active member of the DAWG. The first workshop for the DAWG and local experts in machine learning, statistics and genomic research will be scheduled in Fall 2005.

Findings

Screening for ortholog clusters using multipartite graph clustering

Akshay Vashist, and Casimir Kulikowski and Ilya Muchnik

Genes related through evolution are called homologous genes. They provide a good basis for extrapolating our knowledge from well-studied organisms to new ones, as in functional annotation of their genes. An important class of homologous sequences is that of orthologous genes, or gene sequences present in different genomes that have arisen through vertical descent from a single ancestral gene in the last common ancestor. Such genes usually perform the same function(s) in different organisms but the degree of sequence similarity across the organisms varies, and usually depends on the time elapsed since their divergence. Ortholog detection is a fundamental problem in estimating traces of the vertical evolution of genes; its practical uses include gene function annotation and finding targets for experimental studies. While many ortholog detection procedures have been proposed, they suffer from limitations that present real challenges. They may be limited to a pair of genomes, require phylogenetic information, which may not be reliable, or can handle only small sized data. The most widely used and trusted databases of orthologous groups require manual curation step(s) by experts and this is a rate limiting factor in addressing the current demand. Automatic procedures of ortholog screening, grounded on methodologies that can tackle the problem as completely as possible, while ensuring the sensitivity of the orthologous groups produced, could therefore be valuable adjuncts to speedup the process. Akshay Vashist, Rutgers University graduate student, Casimir Kulikowski, Department of Computer Science, Rutgers University, and Ilya Muchnik, DIMACS, developed a model for automatically extracting candidate ortholog clusters in a large set of genomes using a new clustering method for multipartite graphs. They designed a new kind of similarity function (linkage function) to capture the relationship between a gene and a subset of genes. The similarity relationships among genes from multiple genomes are represented as a multipartite graph, where nodes in a partite set correspond to genes in a genome. To this they apply a new clustering method for multipartite graphs. The method is fast and enables them to extract ortholog clusters from a large set of genomes. The key to the efficiency of the procedure is a particular property of the objective function, which is based on the linkage function. They evaluated the performance of their method by applying it to screening for orthologous genes in a large number of

prokaryote genomes. The analyses of the results shows that orthologous clusters obtained using their approach show a high degree of correlation with the manually curated ortholog clusters in one of the most trusted databases of ortholog clusters, COG. The multipartite graph clustering method produces smaller but conserved orthologous clusters. This feature ensures that clusters contain sequences that share an orthologous relationship, and is critical to balance for the manual curation of orthologous clusters. On the other hand, related conserved clusters can be merged, using a variant of the proposed method, to obtain a desired level of aggregation. This could be useful to biologists who search genomic data to discover relationships between genes in related organisms.

Outreach Activities

This project is closely intertwined with DIMACS efforts to link mathematics and computer science with biology in the high schools. The project organizers were involved in planning a DIMACS conference on this subject in April 2005 (see http://dimacs.rutgers.edu/Workshops/Biomath/). Also, the project organizers are working closely with the Summer 2005 DIMACS Bio-math Connect Institute (BMCI), which is aimed at introducing high school math/CS and Bio teachers to topics at the interface. This project is informing the BMCI effort and specific topics from the project are being adapted for use in BMCI.

Papers

Akshay Vashist, Casimir Kulikowski, Ilya Muchnik, "Screening for ortholog clusters using multipartite graph clustering by quasi-concave set function optimization," to appear in *Proceedings of The Tenth International Conference on Rough Sets, Fuzzy Sets,Data Mining, and Granular Computing* (RSFDGrC 2005).

Akshay Vashist, Casimir Kulikowski, Ilya Muchnik, "Ortholog groups as clusters on a multipartite graph," submitted to *Workshop on Algorithms in Bioinformatics* (WABI 05).

Akshay Vashist, Casimir Kulikowski, Ilya Muchnik, "Automating protein function annotation through candidate ortholog clusters from incomplete genomes," in preparation.

Talks

Akshay Vashist, Casimir Kulikowski, Ilya Muchnik, "Screening for ortholog clusters using multipartite graph clustering by quasi-concave set function optimization," *Tenth International Conference on Rough Sets, Fuzzy Sets,Data Mining, and Granular Computing* (RSFDGrC 2005).

Main website

http://dimacs.rutgers.edu/Workshops/InfoProcess/

Other Specific Products

Web pages

Workshop on Biomolecular Networks: Topological Properties and Evolution
http://dimacs.rutgers.edu/Workshops/Biomolecular/

Workshop on Information Processing by Protein Structures in Molecular Recognition
http://dimacs.rutgers.edu/Workshops/InformationProcessing/

Workshop on Detecting and Processing Regularities in High Throughput Biological Data
http://dimacs.rutgers.edu/Workshops/Detecting/

Contributions

Contributions within Discipline

This special focus is of course by nature multi-disciplinary. A major contribution is the impact on the research programs and careers of the participants. Here is a single comment from a participant, describing this.

"I just came back from a trip abroad that started with the DIMACS workshop on Biomolecular Networks, May 11-13 (followed by a couple of other conferences). It is still premature to tell what kind of collaborations or research results it has generated, but for now I can safely say that it was one of the best scientific meetings I have been to in the last 5 years. The setting (long talks, ample time for interaction, etc.) and selection of topics and participants was outstanding and I can already say that there are many issues that came up that I will definitely pursue in my research with my students here at the Technion and possibly with additional people. Thanks again for this great experience." Ron Y. Pinter, Dept. of Computer Science, Technion, Israel)

Contributions to Other Disciplines

Since the "discipline" is inherently multidisciplinary, there is no separate entry in this section.

Contributions Beyond Science and Engineering

We expect that many of the workshops will have an impact on the medical and public health fields.

Contributions to Human Resources Development

Many graduate students, undergraduates, and several postdocs are participating in the program. Local graduate students and many non-local students are involved as visitors and workshop attendees. More senior people are also heavily influenced by the project, being exposed to new directions of research. The impact on the careers of the students and faculty is illustrated by the following:

"I benefited a lot from the recent workshop on "Biological networks - topology and evolution…" It was good exposure to this developing field and an opportunity to discuss with other participants." Akshay Vashist, Rutgers University, a graduate student whose research was supported by this special focus.