NIH National Center for Integrative Biomedical Informatics (NCIBI)

James Cavalcoli, Ph.D., Hosagrahar V. Jagadish, Ph.D., David J. States, M.D., Ph.D., Gilbert S. Omenn, M.D., Ph.D., Barbara Mirel, Ph.D., and Brian D. Athey, Ph.D.
University of Michigan, Ann Arbor, Michigan

NCIBI is one of seven National Centers for Biomedical Computing established under the NIH Roadmap Initiative

NCIBI Vision, Structure, and Team

Brian D. Athey, Ph.D., PI and Senior Scientific Director

The National Center for Integrative Biomedical Informatics (NCIBI) is developing the framework of conceptual models, computational infrastructure and an integrated knowledge repository, which modern scientists need to make effective use of the wealth of data flowing from molecular biology and translational research. The NCIBI provides researchers with web-accessible knowledge analysis, collaborative work environments to create and utilize computationally-enabled models, and workflows to better understand complex biomedical processes.

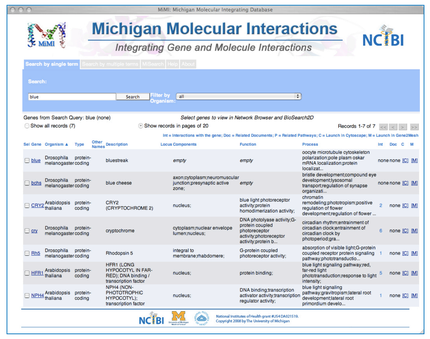
The NCIBI is currently:

- Developing a computing architecture that permits harnessing of diverse data sources, including information from the literature, into a single cohesive framework.
- Building an integrated information store containing experimental data, and models of biological processes.
- Implementing software tools for exploratory bioinformatics analysis, model construction, and model evaluation software.
- Accelerating the discovery of clinically relevant information in the biological domains of Prostate Cancer, Type 1 Diabetes, Type 2 Diabetes, and Bipolar Disorder.
- Conducting innovative education and training programs
- Developing problem solving techniques in the collaborative study of complex biological systems.

Core 1: Computational Technology

H.V. Jagadish, Ph.D. Senior Scientific Director

Core 1 designs and prototypes the database architectures, indexing and query tools, data analysis tools, and information science tools for complex concept display and analysis. It develops the computational methodologies to build databases containing the large variety of biological data and models which biomedical research scientists derive information from and create hypotheses with.



Core 2: Bioinformatics Technology

David J. States, M.D., Ph.D., Senior Scientific Director

Core 2 develops computational technology to organize biological data into structures that can be efficiently queried and linked using pattern inference and recognition, deep natural language parsing and intelligent information retrieval. Machine assisted and fully automated natural language processing is used to extract information from the literature and from the databases to inform molecular biological models. Advanced visualization and interface tools are created to convey information rapidly and effectively to biological scientists both locally in the Driving Biological Problems (Core 3) and nationally using web services.

Examples of Software Resources Available in NCIBI

- MiMIWeb, MiMI Plugin for Cytoscape
- SAGA approximate subgraph alignment and search
- SNP Function Portal: A web service for functional annotation of genetic polymorphisms
- GenePattern pipelines
- LexRank/Clairlib information retrieval and summarization tools
- Bayesian network analysis tools (BUBBLE, MarkIt and miniTuba)
- tfblast: efficient whole genome transcription factor binding site identification and clustering
- Shannon NLP pipeline
- Protege: ontology editor and knowledge-based framework
- OMSSA/X!Tandem/TransProteomicPipeline/CPAS - proteomics/mass spectrometry data analysis
- Literature-based automated annotation tools using MeSH terms (Gene2MeSH, Metab2MeSH)

Examples of Data Resources Available in NCIBI

- MiMI: Michigan Molecular Interactions Database
- MBI: Molecular Biology Integration Databases
  - Genes, taxonomoy, orthology relationships, sequences and functional annotation
- Shannon NLP resource
  - Relational implmentations of PubMed, PubMedCentral and task specific full text corpora
  - Document structure and deep NLP parsing and tagging
- Cell line ontology
- HUPO Plasma Proteome mass spectrometry and proteomics data sets
- Custom gene expression probe to genes mappings (CDF files)

Core 3: Driving Biological Problems (DBPs)

Gilbert S. Omenn, M.D., Ph.D., Senior Scientific Director

Criteria: Common diseases with complex and heterogeneous etiology—many unknowns; extensive complex datasets available

3.1: Prostate Cancer Progression – Arul Chinnaiyan, M.D., Ph.D., Lead
- Develop methods for identifying gene fusions in cancers
- Integrate gene expression, proteomics, protein-protein interaction data, starting with Oncomine 3.0, into a clinically-heuristic systems model.

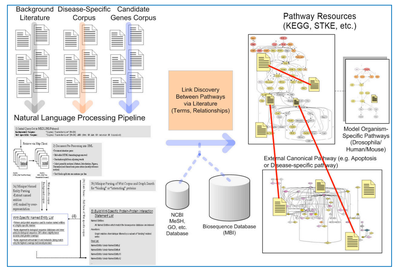
3.2: Type 1 Diabetes Neuropathy and Nephropathy: Mechanisms and Models – Eva Feldman, M.D., Lead
- Compare Nrf2 neuroprotection signaling to other pathways via SAGA tool and database workflow (Patel lab)
- Identify antioxidant response elements and tie to signaling pathways and metabolites
- Model T1DM-associated pathways via Bayesian Network analysis (Woolf lab)
- Collaborators will test therapies against reactive oxygen species in animal model (resveratrol trial)

3.3: Type 2 Diabetes: Genetic and Phenotypic Heterogeneity – Michael Boehnke, Ph.D., Lead
- Exploit ongoing FUSION study datasets
- Implement enhanced SNP analysis workflow
- Devise algorithms with much higher throughput
- Organize parallel analyses

3.4: Bipolar Disorder – Melvin McInnis, M.D., Lead
- Exploit ongoing NIMH studies of genetic, psychological, and psychiatric heterogeneity
- Import data from brain imaging
- Generate faster algorithms for analysis of pathways for candidate genes
- Use flexible tools to search pathways for similarities
- Model interactions of variants at two different gene loci (two different chromosomes)

NCIBI NLP Processing Overview: Linking Biomedical Literature to DBPs

Abstracts and complete text, where possible are parsed and tagged using a series of parsers and statistical and dictionary based tagging tools.Biological molecules are tagged and then linked to MiMI and the other sequence annotation databases as part of the data infrastructure.



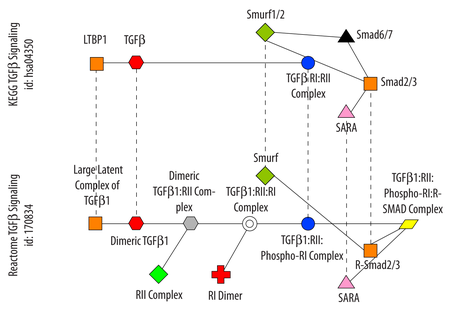
SAGA: Effective and Efficient Graph Matching

Motivation:
- Large amounts of biological graph data: e.g. MiMI, KEGG, Reactome, bioNLP, etc. Graph database sizes are large and increasing in size.
- Graph querying is a common requirement for many DBPs.

Challenge:
- Datasets are noisy/incomplete: so exact matching is not very useful.
- Subgraph Matching is computationally very expensive.

The SAGA approach: A database-driven approach that uses an index on the database graphs that are "digested" into smaller graph fragments.

Result: SAGA is more effective and efficient compared to existing methods.



In the above example SAGA is used to compare pathways in KEGG and Reactome.

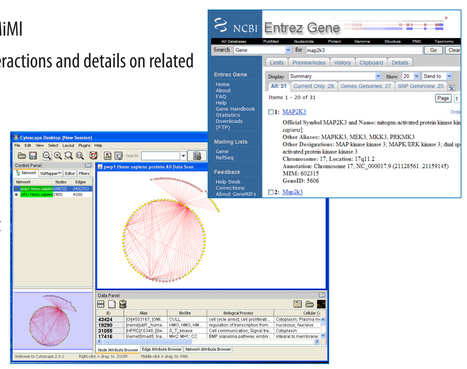
SAGA is able to find matches for similar information in these two different databases even though they organize pathways in different ways.

Note comparing pathways from different databases can be precursor to pathway data integration.

MiMI Integrated with Cytoscape

MiMI Plugin for Cytoscape is an open source interactive visualization tool that you can use for analyzing protein interactions and their biological effects. The MiMI Plugin for Cytoscape couples Cytoscape, a widely used software tool for analyzing bimolecular networks, with the MiMI database.

- Access the integrated molecular data assembled in MiMI
- Retrieve interactive graphics that display protein interactions and details on related attributes and biological concepts.
- Interacting with displays, you can expand networks to the next nearest neighbors and zoom and pan to relationships of interest.
- Literature information via nature language processing database (BioNLP) and a multi-document summarization system MEAD
- Match these sub-network to biological pathways using SAGA, a graph matching tool (SAGA).



Core 4: Computing and Data Infrastructure

James D. Cavalcoli, Ph.D., Core Interim Director

- Provide real-time computational, data, communication, networking, and collaborative support
- Implement NCIBI software tools
- Establish and maintain production databases and libraries
- Interface with UM Center for Advanced Computing (CAC) and DBP research sites to insure end-to-end capability
- Oversee links to national computing and data resources

Core 5: Education and Training

Barbara Mirel, Ph.D., Core Director

- Develop training guides for applications
- Create Demonstrations and Tutorials to promote utilization of tools
- Plan and Implement outreach to broader scientific community (with Core 6)

and

Core 6: Outreach and Dissemination

Brian D. Athey, Ph.D., Core Director

- Develop interdisciplinary training for NCIBI personnel
  - Lead the NCIBI "Tools and Technology" Outreach Series
- Establish education and training programs
- Train NIH researchers and user community to use NCIBI resources
- Evaluate NCIBI tools and technology effectiveness
- Leverage the environment
  - New UM Center for Computational Medicine and Biology (CCMB)
  - Bioinformatics and Computational Biology Graduate Training Program
- Enhanced training and research opportunities for T32 trainees
- Synergistic infrastructure
  - Courses, seminars, journal clubs, facilities
- Manage Collaborative R01 and R21 program
- Manage National Supercomputing Center Strategic Partnerships

Core 7: Administration, Collaboration and Interaction

Brian D. Athey, Ph.D., PI and Senior Scientific Director

NCIBI Executive Committee:
Brian D. Athey, (Chair), Gilbert S. Omenn, David J. States, Hosagrahar V. Jagadish, James Cavalcoli (Project Manager) Katrina Ward Hope, (UM Center for Computational Medicine & Biology Director of Administration)

NIH Officials:
NCIBI Program Officer (PO): Dr. Karen Skinner, NIDA NCIBI Lead Science Officer (LSO): Dr. Jane Ye, NLM, Science Officer (SO) German Cavelier, NIMH/NIH

NCIBI External Board of Advisors:
- Eric Jakobsson, Ph.D. Director of the National Center for Design of Biomimetic Nanoconductors
- Joel Saltz, M.D., Ph.D. Professor and Chair, Department of Biomedical Informatics The Ohio State University College of Medicine and Public Health
- Raymond Rudden, M.D., Ph.D. Professor Emeritus, University of Michigan School of Medicine
- John Quackenbush, Ph.D. Professor of Computational Biology and Bioinformatics, Harvard School of Public Health
- Kirstie Bellman, Ph.D. Principal Director Aerospace Integration Science Center.
- Mark H. Ellisman, Ph.D., Professor of Neurosciences & Bioengineering, University of California, San Diego

Applications-Oriented Interdisciplinary Research (IDR) and Implementation Team: NCIBI Subcontracting Partners

Institute for Systems Biology (ISB)
Prostate Cancer Informatics Partner
Leroy Hood, PI

Stanford University
Ontology Integration Partner
Mark Musen, PI

University of Southern California
FUSION Diabetes Type II Partner
Richard Watanabe, PI

The Broad Institute (MIT)
Gene Pattern and Workflows Partner
Jill Mesirov, PI

Carnegie Mellon University
Image-Based Information Retrieval Partner
Robert Murphy, PI

Other NIH Roadmap NCBCs NCIBI is Working With

Center for Computational Biology (CCB)
Arthur Toga, PI

Physics-Based Simulation of Biological Structures (SIMBIOS)
Russ Altman, PI

Informatics for Integrating Biology and the Bedside (i2b2) Isaac Kohane, PI

The National Center For Biomedical Ontology (NCBO)
Mark Musen, PI

National Alliance for Medical Imaging Computing (NA-MIC)
Ron Kikinis, PI

Multiscale Analysis of Genomic and Cellular Networks (MAGNet)
Andrea Califano, PI

Acknowledgements

This work has been done with the support from National Institutes of Health under grant #U54 DA021519 for the National Center for Integrative Biomedical Informatics.