Posters for NCBC Showcase November 8-9
Natcher Conference Center, NIH

1. **Preinitiation complex dynamics regulates RNA synthesis precision**

Kunal Poorey
University of Virginia

TATA-binding protein (TBP) nucleates the assembly of the transcription preinitiation complex (PIC), and although TBP binds promoters with high stability in vitro, recent results establish that virtually the entire TBP population is highly dynamic in yeast nuclei in vivo. This dynamic behavior is surprising in light of models which posit that a stable TBP-containing scaffold facilitates transcription reinitiation at active promoters. The dynamic behavior of TBP is a consequence of the enzymatic activity of the essential Snf2/Swi2 ATPase Mot1, suggesting that ensuring a highly mobile TBP population is critical for transcriptional regulation on a global scale. Here high-resolution tiling arrays were used to define the locations of TBP, to identify TBP binding sites susceptible to Mot1 action. Combined with computational analysis we define how perturbed TBP dynamics impact the precision of RNA synthesis in *Saccharomyces cerevisiae*. We find that Mot1 plays a broad role in establishing the precision and efficiency of RNA synthesis: in *mot1-42* cells, RNA length changes were observed for 713 genes, about twice the number observed in *set2*Δ cells, which display a previously reported propensity for spurious initiation within open reading frames. Loss of Mot1 led to both aberrant transcription initiation and termination, with prematurely terminated transcripts representing the largest class of events. Genetic and genomic analyses support the conclusion that these effects on RNA length are mechanistically tied to dynamic TBP occupancies at certain types of promoters. These results suggest a new model whereby dynamic disassembly of the PIC can influence productive RNA synthesis.

2. **Identifying genomic features by BLASTing through chromatin**

Michael Buck
SUNY at Buffalo

Sequencing the human genome was a monumental accomplishment that provided us with the roughly 3 billion base pairs, which encode the raw instructions for synthesizing every human cell, tissue, and organ. These instructions are encoded into an elegant but cryptic code which we have only recently begun to decipher. A key to understanding this code has been the identification of common genomic features, ie. promoters and enhancers. However, finding these locations within the genome can be a laborious and expensive undertaking requiring site specific assays. Even more difficult is identifying entirely new classes of genomic features. To facilitate identification and characterization of new classes of genomic features we have developed **Chrom**atin **A**rchitecture **B**asic **L**ocal **A**lignment **S**earch **T**ool (ChromABLAST). The ChromABLAST algorithm uses the chromatin architecture at a user defined region or regions of interest and identifies all similar regions within the genome as defined by their shared chromatin architecture. ChromABLAST differs from other methods in that it uses both the amplitude and spatial arrangement of the chromatin modifications to score similarity. Importantly, ChromABLAST allows for the identification of subtypes of known genomic features and can accurately predict previously uncharacterized locations. ChromABLAST uses an innovative weighted profile generated from only the most informative chromatin datasets and then scores the entire genome. We have validated the accuracy of our approach with multiple well characterized genomic features from yeast and

humans. We demonstrate how ChromABLAST identifies novel transcription start sites and show experimental evidence for their existence. In addition, we apply ChromABLAST across cell types and show how it can use the chromatin architecture at a transcription factors binding site in one cell type to predict its targets in another cell type.

Public Health Relevance

The proposed research is relevant to public health because identification and characterization of functionally important elements within the human genome is essential for our understanding of how genomic information is processed and disrupted during disease progression.

3. **Next Generation Molecular Surface Representation for Rapid Global/Local Protein Shape Comparison, Docking and Low-Resolution Data**

Daisuke Kihara
Purdue University

Protein structural bioinformatics is facing new challenges raised from needs of analyzing an increasing number of solved protein structures, low-resolution structural data from electron microscopy or tomography, predicted structures, as well as structures of unknown function. To enable efficient global and local protein structure analysis, our group has been using a novel protein surface representation with the 3D Zernike descriptor (3DZD). The 3DZD is a series expansion of a three dimensional mathematical function that can represent a 3D object in a rotation and translational invariant fashion. We show several successful applications of the protein surface representation using 3DZD in our group. 3DZD is very suitable for rapid real-time searching of global and local protein structures. On our webserver, 3D-Surfer, structure comparison against the entire Protein Data Bank can be performed within a few seconds. For local surfaces, particularly, we can perform efficient ligand binding pocket comparison as well as cross-comparison against ligand (drug) molecules. 3DZD can also identify complementary shapes in two molecules, which is very convenient for protein- protein docking prediction. Our new docking algorithm, LZerD (local 3D Zernike descriptor-based Docking algorithm), uses the 3DZD for capturing local shape complementarity of interacting protein surfaces. LZerD has been further extended recently for multiple protein docking (Multi-LZerD) and also for fitting multiple protein structures into a density map of electron microscopy (EM-LZerD). This work is supported by NIH (R01GM075004, R01GM097528).

Public Health Relevance

The presented computational methods allow fast identification of protein-ligand (or drug) interactions and protein-protein interactions by recognizing complementary surface shape of molecules. The methods provide crucial structural information for understanding protein-protein interactions that are involved in disease and help generating hypothesis in the early stages of drug discovery.

4. **From Connectivity Models to Region Labels: Identifying Foci of a Neurological Disorder**

Archana Venkataraman
MIT

We propose a novel approach to identify the foci of a neurological disorder based on anatomical and functional connectivity information. Specifically, we formulate a generative

model that characterizes the network of abnormal functional connectivity emanating from the affected foci. This allows us to aggregate pairwise connectivity changes into a region-based representation of the disease. We employ the variational EM algorithm to fit the model and subsequently to identify both the afflicted regions and the differences in connectivity induced by the disorder. We demonstrate our method on a population study of schizophrenia.

Public Health Relevance

Our work aggregates population differences in brain connectivity to localize regions associated with a neurological disorder.

5. **Structure-based studies of protein-protein interactions on a genome-wide scale: implications for cancer biology**

Antonina Silkov, Qiangfeng Cliff Zhang, Donald Petrey, Jose Ignacio Garzon, Lei Deng, Andrea Califano and Barry Honig
Columbia University

Experimentally determined protein complexes can be used to predict interactions between the sequence and structural homologs of the proteins involved. We describe a novel computational method based on geometric relationships between protein structures, which can be used to accurately predict protein-protein interactions on a genome-wide scale. The comparative study shows that the coverage and the accuracy compare favorably with methods derived from sequence and functional clues, e.g. sequence similarity, co-expression and phylogenetic similarity. A Bayesian evidence learning model, PrePPI, that combines structural information with non-structural evidence yields predictions of comparable quality to high-throughput experiments. We show how the novel protein-protein interactions predicted by PrePPI provide clues about the molecular function of poorly characterized potential oncogenes.

Public Health Relevance

PrePPI represents a unique open resource for predicting novel protein-protein interactions and visualizing their structural details.
These predictions can be used by researchers in many different areas of Biology and Public Health for mechanistic elucidation of disease-associated signaling cascades.

6. **Towards Collaborative Visualization of Biomedical Ontologies and Mappings on the Web**

Bo Fu, Lars Grammel, Eric Verbeek, Margaret-Anne Storey
University of Victoria, Canada

Improving cognitive support for ontology understanding is particularly important in the biomedical domain, where ontologies can be large and are typically created in a collaborative environment. Despite a large amount of research effort in developing ontology visualization techniques, there has been relatively little attention placed on supporting collaborative visualization for ontologies. This poster presents BioMixer (bio-mixer.appspot.com), a web-based visualization tool developed by the CHISEL (Computer Human Interaction and Software Engineering Lab) Group at the University of Victoria in collaboration with Stanford University as part of the National Center for Biomedical Ontology (NCBO). In particular, this poster demonstrates the collaborative features in BioMixer and showcases an application in the BioPortal library for biomedical ontology and mapping visualization. The research presented in

this poster is supported by the NCBO, the NHGRI, the NHLBI, the NIH Common Fund under Grant #U54-HG004028.

Public Health Relevance

The NCBO supports a burgeoning user community that is using ontologies to enhance biomedical research and to improve patient care. It supports bench scientists, clinician researchers, and workers in informatics in data annotation, data integration, information retrieval, natural-language processing, electronic patient record systems, and decision-support systems. It is a primary source of semantic-technology infrastructure and expertise for biomedical research and the development of advanced clinical information systems.

7. **Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing: A Novel Informatics Approach using i2b2**

Ashwin N Ananthakrishnan
Massachusetts General Hospital

Prior studies identifying patients with inflammatory bowel disease (IBD) utilizing administrative codes have yielded inconsistent results. Our objective was to develop a robust electronic medical record (EMR) based model for classification of IBD leveraging the combination of codified data and information from clinical text notes using natural language processing (NLP). Using the EMR of 2 large academic centers, we created data marts for Crohn's disease (CD) and ulcerative colitis (UC) comprising patients with $\geq$ 1 ICD-9 code for each disease. We utilized codified (i.e. ICD9 codes, electronic prescriptions) and narrative data from clinical notes to develop our classification model. Model development and validation was performed in a training set of 600 randomly selected patients for each disease with medical record review as the gold standard. Logistic regression with the adaptive LASSO penalty was used to select informative variables. We confirmed 399 (67%) CD cases in the CD training set and 378 (63%) UC cases in the UC training set. For both, a combined model including narrative and codified data had better accuracy (area under the curve (AUC) for CD 0.95; UC 0.94) than models utilizing only disease ICD-9 codes (AUC 0.89 for CD; 0.86 for UC). Addition of NLP narrative terms to our final model resulted in classification of 6-12% more subjects with the same accuracy.

Public Health Relevance

Crohn's disease and ulcerative colitis affect over 1.4 million Americans and are associated with considerable morbidity and healthcare costs. The novel approach used in this study will allow for efficient accrual of a large cohort, accurate definition of outcomes, and link to biospecimens that will be used to develop tools to predict disease related complications, response to treatments, as well as prevent unnecessary exposure to potentially toxic therapies with low likelihood of benefit.

8. **PROMISE : PRivacy-Oriented Medical Information Secure BiomEtric system**

Shuang Wang, Xiaoqian Jiang, Lijuan Cui, Samuel Cheng, Lucila Ohno-Machado
University of California and University of Oklahoma

Privacy is an important concern when biometrics are used in authentication systems for accessing Electronic Health Records (EHR) or other biomedical research data repositories involving human subjects. Biometrics of individuals deserve careful protection because they contain sensitive information closely related to personal privacy (e.g., personal health, ethnic group, etc.) and the

leakage of such information can be used to re-identify individuals. More importantly, biometrics are unique and they are not easily revocable. Existing secure biometric systems prevent attackers from collecting unprotected biometrics in databases, however, they cannot guarantee confidentiality in probing and transmitting biometrics. We propose a practical privacy-preserving biometric system for secure fingerprint authentication based on distributed source coding principles, for which biometric features from both probe devices and databases are protected (i.e., secure storage and transmission). The proposed framework is suitable for the scenario involving distributed and mobile devices. Besides biometric transmission and authentication in a privacy-preserving manner, experimental results show that the proposed system also achieves a reasonably high authentication accuracy (i.e., 96). The proposed framework can be integrated into the two-factor authentication system within iDASH for securely privacy-preserving access control.

Public Health Relevance

The proposed PROMISE framework tackles the privacy issue of integrating biometrics in authentication systems for accessing Electronic Health Records (EHR) or other biomedical research data repositories involving human subjects.

9. Tools for Integrating Metabolomics

Alla Karnovsky
University of Michigan

This poster describes 2 different NCIBI Tools (Metscape and Metab2MeSH) for identifying, and integrating Metabolite / metabolomics data along with Genes, Proteins, related pathways and reactions. These tools were developed as part of the NCIBI program and have been incorporated into a recently funded U24 Metabolomics Center grant.

Public Health Relevance

Parallel study of multiple types of molecular data can provide biomedical researchers with a comprehensive view of normal and disease physiology. Our tools help interpret complex genomic and metabolomics data. The new analysis methods and complex visualizations that we developed allow researchers to better learn from their data.

10. Chemical and Mechanical Regulation of Bacterial Cell Wall Expansion

Enrique Rojas
Stanford University

The bacterial cell wall is the target of many of our best antibiotics, and it is only by interfering with the growth of this macromolecule that these compounds are effective. Therefore, to develop novel antibacterial compounds, there is a need to deeply understand the molecular mechanisms that determine growth of the cell wall as well as the mechanisms by which cell wall-targeting compounds act. We are addressing these broad problems with a joint experimental and computational effort. Experimentally, we probe the cell wall both mechanically (by controlling osmotic pressure within the cell) and chemically (by applying antibiotics that target the cell wall and its precursors), and study the effects of these perturbations on single cell growth rate. Computational modeling allows us to relate observations at the cell level to phenomena at the molecular scale, and to thus test hypotheses with respect to

the molecular mechanisms of cell wall expansion by comparing our models to experiment. For example, our data concerning the Gram-negative enteric bacterium *Escherichia coli* is consistent with a model in which hydrolysis of cell wall material is highly coupled to insertion of new wall material. On the other hand, for the Gram-positive soil bacterium, *Bacillus subtilis*, it appears that these processes are not as tightly coupled, resulting in a drastic response to osmotic shock, as compared with *E. coli*. Finally, the development of computational models that describe these observations have allowed us to uncover not only the mechanisms of cell growth, but also the time scales associated with metabolism of cell wall precursors and those of antibiotic action.

Public Health Relevance

I am studying the biosynthetic pathway of the cell wall of bacteria, which is an important target for many of our best antibiotics. This research may lead to a better understanding of bacterial resistance to antibiotics and may also elucidate new pathways to drug development.

11. Querying the genome with GQL

Christos Kozanitis, Vineet Bafna
CSE/UCSD

With DNA sequencing costs dropping below $1, 000 for human genomes, the hope is that medicine will be personalized, based on the individual's genetic makeup. As hardware costs fall, software costs are a major obstacle to both discovery and personalized medicine. We propose attacking this obstacle using a hierarchical set of abstractions (akin to Internet layers) for genome processing. While interface formats exist for mapped files (BAM) and variants (VCF), we propose a new separation of variant calling into a deterministic evidence layer queried by a statistical inference layer via a proposed Genome Query Language (GQL). GQL allows inference engines to efficiently sift only the relevant evidence (e.g., the READs that flank a SNP site) from raw genomic data. Using a set of exemplar queries, we argue that the standard implementation of GQL using relational database tables with billions of rows for each genomic location will not scale; instead, we introduce efficient indices based on compressed strength vectors. We outline several benefits to separating evidence from inference including the ability to reuse genomic data, and enabling a shift from deep analysis of individual genomes to group inference on large populations.

Public Health Relevance

The vision of our project is twofold. The initial goal is to facilitate the discovery of genetic anomalies that result in diseases such as cancer. The ultimal goal is to make each individual's genome easily accessible to medical doctors for a more personalized diagnosis and treatment advice.

12. informed CONsent for clinical record and Sample use in research (iCONS)

Adela Grando, Elizabeth Johnstone
University of California San Diego, University of Minnesota, and University of California

Most research enterprises rely on paper-based informed consent documents. In contrast, we are proposing an *Electronic Informed Consent Management System*. Electronic informed consents offer the potential to embed multi-media resources to further educate patients on the research study they have been invited to participate and the consequences of participating, so they can

make a truly informed decision. Furthermore, a *Permission Ontology* can be used to capture, in a machine-interpretable standard formalism, the permissions that patients give through informed consent to share their clinical data and biosamples for future research. After patient permissions have been expressed in the *Permission Ontology*, they can be saved in a *Permission Repository*. Finally, a *Resource Allocator* can grant researchers' resources requests by accessing the *Permission Repository* and, using reasoning, determine compliance with subjects' permissions.

13. Protein Engineering through Computational Design: From Proof of Principle to Real Life Applications

Gert Kiss
Stanford University

In recent years, computational methods have matured to a point at which it has become possible to equip protein structures with new functions. Computational 'de-novo' design has taken the approach of engineering entire active sites around a QM transition state and grafting these into stable protein scaffolds. While this has led to the production of new biocatalysts, we are far from achieving the level of specificity, turnover rates, and chemical sophistication that is evident from nature's enzymes. We are working on significantly improving the current design approach by accounting for dynamic motions. Over the past year, the focus of my research has been on methodological developments and on their applications towards biochemically relevant systems. More specifically, I have been working a) on a protocol for the automated and reliable setup of molecular dynamics (MD) simulations, b) on rapid ways of analyzing the vast amount of data that is generated in the course of such simulations, c) on applying both (a) and (b) towards the analysis and the redesign of prenyltransferase NphB (in collaboration with Prof. Chaitan Khosla at Stanford), and d) on the high-throughput screening and evaluation of computationally designed enzymes, binders, and structural proteins (in collaboration with Prof. David Baker at UW). Here, I will give an overview of the techniques that have been developed and applied towards establishing a dynamics-based 'de-novo' design approach and will then introduce the two projects c) and d) in an attempt to bridge the gap between proof of principle and application.

Public Health Relevance

The methodological developments in computational protein engineering that I will describe are still in development and are currently in a transitory phase between proof of principle and their applicability to relevant problems. Two examples of such problems are highlighted in my poster and describe a) an approach towards the synthesis of meroterpenoid natural products with anti-cancer efficacy, and b) the engineering of flu-binder proteins.

14. Whole blood miRNA profiles in acute and convalescent Kawasaki disease using a pipeline tool for smallRNA-Seq

Jihoon Kim, Hubert D Lau, Chisato Shimizu, Jane C Burns, Petra Stepanowsky, Lucila Ohno-Machado
University of California

Kawasaki disease is an acute, self-limited vasculitis of childhood that can result in structural damage to the coronary arteries. Previous studies have implicated the TGF-β pathway in disease pathogenesis and generation of myofibroblasts in the arterial wall. microRNAs are small non-coding RNAs that modulate gene expression at the post-transcriptional level and can be transported between cells in extracellular vesicles. To understand the role that microRNAs play in

modifying gene expression in Kawasaki Disease, we studied microRNAs from whole blood during the acute and convalescent stages of the illness using an in-house developed pipeline analysis tool.

Methodology/Principal Findings: Sequencing was performed on size-selected RNA species (<30 nt) isolated from paired whole blood samples collected during the acute and convalescent phase from patients with Kawasaki disease (n=12) and during the acute phase from children infected with adenovirus (n=6). A new pipeline tool, MAGI (MicroRNA- seq Analysis in Gpu Infrastructure), was used to detect microRNA sequences and was used to analyze differential expression of microRNAs in acute and convalescent samples. miRs-143, -199b-5p, -618, -223, -145 and -145* (complementary strand) were confirmed by qRT-PCR to be differentially expressed in acute versus convalescent Kawasaki disease subjects (n=16). miR-145, which plays a critical role in the differentiation of neutrophils and vascular smooth muscle cells, was expressed at high levels in blood samples from acute Kawasaki disease but not adenovirus-infected patients. Transcript levels of miR-143 correlated strongly with peripheral blood polymorphonuclear leukocyte count (r2=0.53) suggesting that expression levels reflected circulating cell types. miR-145 was also detected in small extracellular vesicles isolated from acute Kawasaki disease plasma samples. Through MAGI, pathway enrichment analysis of the predicted targets of the 6 differentially expressed microRNAs identified the TGF-β pathway as the top pathway regulated by microRNAs in Kawasaki disease.

Conclusion: Sequencing of small RNA species allowed discovery of microRNAs that may participate in KD pathogenesis. miR-145 is transported in the peripheral blood in extracellular vesicles during acute Kawasaki disease and may participate, along with other differentially expressed microRNAs, in regulating expression of genes in the TGF-β pathway during the acute illness. Our pipeline tool enables a comprehensive and accurate discovery in a fast way.

Public Health Relevance

We developed a pipeline software for microRNA-Seq data analysis and applied this to the samples from Kawasaki Disease (KD) patients. We discovered the KD-specific microRNAs, which are validated and pending for patent.

15. Utilizing the electronic medical records for RA translational research with i2b2

Katherine P. Liao
Brigham and Women's Hospital & Harvard Medical School

I will demonstrate how we utilized the electronic medical records for rheumatoid arthritis translational research as part of the Informatics for Integrating Biology and the Bedside (i2b2) project. Specifically, I will provide an overview of (1) the methodology we developed to create an RA cohort within the EMR using an algorithm based approach and (2) logistics of how we linked clinical data with discarded blood samples. Finally, I will provide a brief example of a genetic association study we conducted to determine the association between RA risk alleles and low density lipoprotein levels (LDL) in the RA cohort.

Public Health Relevance

The development of an electronic medical record (EMR) platform for research harnesses a wealth of clinical data not typically available in most clinical studies, i.e. traditional cohort studies,

administrative databases. Research using EMR data can include a broad range of patients often not included in clinical trials which have strict selection criteria, as well as clinical outcomes and adverse events that occur as part of real-life practice.

16. A Study of the Impact of NCBO Technology on the Biomedical Community

Bo Fu, Cassandra Petrachenko, Margaret-Anne Storey
University of Victoria, Canada

Research conducted by the National Center for Biomedical Ontology (NCBO) has resulted in several tools and services aimed at addressing challenges arising from the biomedical domain. Notable examples of such tools and services, known as NCBO Technology, include the BioPortal library, the NCBO Annotator, the NCBO Resource Index and the NCBO Web services. In order to better support the needs of the growing biomedical community using NCBO Technology, it is essential to understand the impact of these technologies on the community. This poster presents key findings from a user study conducted by the CHISEL (Computer Human Interaction and Software Engineering Lab) Group at the University of Victoria in collaboration with Stanford University. The findings have shown strong evidence indicating that NCBO technologies are successfully supporting biomedical researchers during various stages of their work. The results have also highlighted the important role that the center plays in providing knowledge, support and standards to the biomedical community. The research presented in this poster is supported by the NCBO, the NHGRI, the NHLBI, the NIH Common Fund under Grant #U54-HG004028.

Public Health Relevance

The NCBO supports a burgeoning user community that is using ontologies to enhance biomedical research and to improve patient care. It supports bench scientists, clinician researchers, and workers in informatics in data annotation, data integration, information retrieval, natural-language processing, electronic patient record systems, and decision-support systems. It is a primary source of semantic-technology infrastructure and expertise for biomedical research and the development of advanced clinical information systems.

17. NIH National Center for Integrative Biomedical informatics (NCIBI)

James Cavalcoli
University of Michigan

This poster is an overview of National Center for Integrative Biomedical Informatics (NCIBI) and describes the different cores, collaborations and software tool development.

Public Health Relevance

Our poster is a descriptive overview of the. Our center develops software tools and datasets to address driving biological problems. We focus on areas of Cancer, Diabetes, Psychiatric disorders and we have developed software for integrating protein and metabolite data with literature annotation. The tools and data are available for use at www.ncibi.org.

18. NA-MIC Community Building

Tina Kapur
Harvard Medical School

Since the start of the NCBC program, NA-MIC outreach efforts have been organized around the concept of a shared software infrastructure as a way to build a strong community that spans basic algorithmic research through applied biomedical science. In this poster we summarize these activities in terms of their impact on both developers and end-users of our NA-MIC Kit of software. Training events, hands-on working meetings, and electronic communications have all been essential components our efforts to build an international network of like-minded researchers sharing code and accelerating the process of scientific discovery and clinical translation.

19. geWorkbench – Integrated genomics data analysis with MAGNet tools

Aris Floratos
Columbia University

geWorkbench (genomics Workbench, http://www.geworkbench.org/), the bioinformatics platform of the MAGNet Center, is an open source Java desktop application that provides access to an integrated suite of tools for the analysis and visualization of data from a wide range of genomic domains (gene expression, sequence, protein structure and systems biology). More than 70 distinct plug-in modules are currently available. These include state of the art algorithms for the reverse engineering of regulatory networks and for protein structure prediction developed by investigators in the MAGNet Center, as well as classical methods commonly used for the analysis of high-throughput genomic data (several variants of clustering, classification, homology detection, etc.). geWorkbench leverages standards-based grid technologies to provide seamless access to remote data, annotation and computational servers, thus enabling researchers with limited local resources to benefit from available public infrastructure. All geWorkbench modules are integrated through a user-friendly, unified user interface. For biomedical researchers with little or no computational training, this approach facilitates adoption by eliminating many steps that require programming skills (such as staging of databases/programs and transformations from one file format to another). For software developers, geWorkbench provides an open source, component-based architecture that enables the addition of new functionality in the form of plug-in modules. Extensive documentation is available (manuals, online help and tutorials) to guide users in the proper use of the application. An innovative logging framework collects and mines data about how various modules are being utilized, offering the possibility for novice users to learn from their more advanced peers.

Public Health Relevance

geWorkbench integrates many computational resources and makes them available through a unified user interface. For biomedical researchers with little or no computational training, this approach facilitates adoption by eliminating many steps that require programming skills. Additionally, by leveraging standards-based middleware technologies to provide seamless access to remote data, annotation and computational servers, geWorkbench enables researchers with limited local resources to benefit from available public infrastructure.

20. Mapping the Conformational Landscape of G-Protein Coupled Receptors Using Novel Computational Paradigms

Diwakar Shukla
Stanford University

G protein coupled receptors (GPCRs) are membrane proteins that modulate many important biological signal cascades, and are the targets of approximately 40% of all commercially available drugs. Besides binding G proteins, activated β2AR can be phosphorylated and bind arrestin, which redirects signaling to other pathways. The crystal structures of GPCR's reveal the structural differences between the active and inactive states but do not give insights into the activation mechanism or pathway relating the active and inactive states. An understanding of GPCR activation mechanism may facilitate the development of more selective drugs with improved therapeutic activity and minimal undesirable side effects. Recent simulation studies (Dror et. al., PNAS, 108, 46, 18684, 2011) have reported a deactivation pathway for β2AR. However, a more detailed and complete picture of the GPCR conformational landscape is desired. Furthermore, the conformational landscape of the apo receptor changes due to binding of ligands (confirmed by the NMR spectroscopy results (Liu et. al., Science, 335, 1055, 2012)). Here, we report results from extensive molecular dynamics (MD) simulations of the apo and ligand bound *β2* adrenergic receptor (β2AR) to elucidate the activation and deactivation mechanism of GPCRs and to identify the ligand induced conformational changes in the landscape. The simulations are performed using novel Markov state model based adaptive sampling algorithms on the Google exacycle distributed computing project.

Public Health Relevance

In this poster, we present results from a computational study of G-Protein coupled receptors (GPCRs), which are a family receptors on the cell surface responsible for sensing the presence of drugs, hormones, neurotransmitters etc. Considering that approximately 30% of all drugs target GPCRs, a more detailed mechanistic understanding of the GPCRs would help in designing drugs with less side effects.

21. Development of a Multi-Center Data Sharing Network for Post-market Surveillance

Susan Robbins, Michael E. Matheny, Fern Fitzhenry, Frederic S. Resnic
Vanderbilt University

iDASH supports automated medication safety surveillance utilizing existing informatics infrastructures within participating centers and leverages high performance, scalable capabilities of data analysis. The process for site participation, developed and tested at Brigham and Women's Hospital (Partners Healthcare), TVHS and UCSD, includes:

1. Data acquisition - Local retrospective EHR extraction
2. Case Identification - Selection of unique records based on high level condition and event filters
3. Common Data Model - Conversion/transformation of local EHR data to CDM
4. Statistical Toolset - Convert CDM cases to analytic table, summary stats and comparative analysis using open source statistical tools developed for longitudinal analysis e.g. risk-adjusted propensity date-range matched controlled cases
5. Analytical Methodology - Perform surveillance analysis at each site and communicate approved results and datasets to iDASH network

Preliminary results from local surveillance of hematologic medications Dabigatran vs Warfarin have been generated as table data and display graphics at the site level within the Analytic User Interface.

Public Health Relevance

Development and implementation of the automated safety surveillance module of iDASH, a national center for biomedical computing, provides an infrastructure for data analysis as a byproduct of usual care, and provides results to a more diverse group of medical practitioners and researchers, allowing for potential interventions to improve public health and safety.

22. Three-Dimensional Calculation and Quantification of Morphometric and Connectomic Changes in Traumatic Brain Injury (TBI)

John D Van Horn
UCLA

Each year, an estimated 1.7 million traumatic brain injury (TBI) cases occur in the United States alone, with an estimated 1.2 million emergency room visits and over 50,000 deaths. In spite of recent progress in the use of neuroimaging in TBI, the computational assessment of TBI images remains particularly challenging for multimodally quantifying TBI-related brain insults, changes in remaining healthy tissue, and connectivity over time. Such assessments are critically important for improving clinical outcome. In this poster, we illustrate 3D morphometric and connectomic change analysis featuring the 3D Slicer framework.

Neuroimaging data of severe TBI were obtained from patients admitted to the UCLA Medical Center and Brain Injury Research Center (BIRC). Acute (2 days post TBI) and chronic (~6 months post TBI) scans were obtained, including MP-RAGE T1, FSE T2, FLAIR, GRE T2, DWI, and SWI. Freely available NA-MIC Kit and 3D Slicer tools were used to obtain metrics of gray matter (GM) and white matter (WM) pathology and change after therapy/recovery. Automatic segmentation of lesions, hemorrhage and edema was performed using Atlas Based Classification (ABC), which includes multimodal image registration, model-based bias field correction and tissue classification. Clinical atrophy measures were computed: maximum width of anterior horns of lateral ventricles (HLV), minimum width of lateral ventricles (MLV), bifrontal index (BFI), bicaudate index (BCI), ventricular index (VI), Evan's index (EI) and Huckman's index (HI). Connectivity analysis was performed according to Irimia et al. Mean percentage changes in each metric were computed and visualized using 3D Slicer.

This is the first study to report calculation of TBI atrophy measures from automatically generated 3D models rather than from 2D MR images, which is customary in radiological practice though not as accurate. 3D Slicer was found to be robust for TBI analysis of longitudinal changes. For each metric, the mean over subjects was found to change by more than 5% at 6-month follow-up compared to baseline, (HLV: 6.7%, MLV: 19.3%, BFI: 6.2%, BCI: 25.4%, VI: 12.0%, EI: 6.2%, HI: 9.3%). The standard deviations over subjects of the change in each atrophy measure were found to be: 5.0% (HLV), 18.7% (MLV), 0.4% (BFI), 28.5% (BCI), 17.2% (VI), 5.0% (EI), 5.6% (HI). The BFI was found to exhibit the most consistent values over the TBI population, whereas the MLV and the BCI registered the largest changes from baseline to follow-up. Connectomic analyses indicated relative changes of fiber pathways connecting affected regions suggestive of diffuse axonal injury and connection-specific WM atrophy.

These atrophy and connectomic measures computed from automatically-generated segmentations provided high precision and accuracy for structural and connectomic analysis in TBI. With the goal of efficient multivariate analysis and visualization of morphometric and connectomic change with respect to neuropsychological metrics, through our NA-MIC collaboration we seek to gain additional insights on the relationships between TBI neuroanatomical pathology and clinical outcome variables.

Public Health Relevance

Millions of Americans suffer traumatic brain injury (TBI) annually. Computational assessment of TBI neuroimaging remains particularly challenging for multimodally quantifying TBI-related brain insults, changes in remaining healthy tissue, and connectivity over time. This work illustrates that connectomic analyses reveal diffuse axonal injury and connection-specific WM atrophy in TBI patients.

23. NCIBI transition to TranSMART

Brian Athey / Kevin Smith
University of Michigan

NCIBI leadership and staff are leading community building activities that bring together and align European Union (funded via ETRIKS) and US-based (current and proposed) resources to achieve the tranSMART vision, and by supporting and enhancing the informatics and data sharing platform for clinical and translational research. NCIBI and the Pistoia Alliance are working with industry, academic, nonprofit, government and value-add service provider organizations to establish a public private partnership that will 1) set scientific, data, analytics, and platform priorities; 2) coordinate major global initiatives through lightweight, transparent governance that includes promotion and outreach; and 3) secure long-term sustainable funding. Other NCBCs involved with tranSMART are i2b2 and NCBO. NCIBI is integrating its data, services and tools with tranSMART; prototype integration of NCIBI tools Metscape, ConceptGen and Metab2MeSH will be completed in September 2012 with beta release anticipated in Q1/Q2 2013.

Public Health Relevance

Disease agnostic, tranSMART integrates and allows for the analysis of high-throughput "'Omics" data such as next-generation sequencing (NGS), transcriptomics, proteomics, metabolomics, and phenotypic elements from electronic health records and clinical research management systems. The tranSMART value proposition relies on the global community of users, developers, and stakeholders who are the best source of innovation for applications and useful data. Continued development and use of tranSMART will create a means to enable "pre-competitive" data sharing broadly, saving money and, potentially, accelerating research translation to cures.

24. Simple and systematic parameterization of a polarizable water model

Lee-Ping Wang
Stanford University (SimBios)

In molecular mechanics simulations of biomolecules, the physical interactions are described using an empirical potential energy function (force field). Simple yet accurate force

fields are highly useful but challenging to develop. Force field development procedures usually involve fitting the empirical parameters to reference data from high-level theoretical calculations or experimental measurements. Generally speaking, a rich and diverse data set can lead to a highly accurate model but also a difficult optimization problem. We developed an open-source optimization program called ForceBalance to meet challenges in force field development. ForceBalance constructs force fields from accurate theoretical data combined with experimental data using systematic optimization methods and strict regularization schemes. We applied ForceBalance to optimize an inexpensive polarizable water model, largely based on the well-known AMOEBA model but with a substantially reduced computational cost. Our new model surpasses AMOEBA in accuracy for several properties of water.

Public Health Relevance

The present study represents a major improvement in the molecular modeling of water, the universal medium for all biomolecular dynamics and interactions. Our work provides a solid foundation for the study of processes such as protein aggregation at the molecular level, leading to an improved understanding of Alzheimer's disease and other neurodegenerative disorders."

25. Adoption of Ontology Design Patterns in Biomedical Ontologies

Jonathan Mortensen
Stanford University

Ontology Design Patterns (ODPs) provide a means to capture best practice, to prevent modeling errors, and to encode formally common modeling situations for use during ontology development. Given their benefits, we asked: Are ODPs used in biomedical ontologies? Which patterns do the ontology developers use? In which ontologies? How frequently are patterns used? We found that 33% of the BioPortal ontologies contained at least one pattern. While ontology design patterns provide benefits especially relevant for large scale biomedical ontologies, we show that today their use in BioPortal ontologies is only modest.

Public Health Relevance

Ontologies are utilized throughout the clinical and biomedical domains. By improving ontologies, one improves healthcare and therefore public health. In this work, we evaluate how ontology design patterns can be used to improve ontologies.

26. iDASH Biomedical Cyberinfrastructure Architecture

Claudiu Farcas
UCSD

The pace of biomedical scientific discovery is rapidly accelerating with new health information technologies that enable multi-disciplinary, multi-investigator collaborations and data sharing. We present the architectural underpinnings of the NIH funded “integrating Data for Analysis, Anonymization, and Sharing” (iDASH) Project’s Biomedical CyberInfrastructure (BCI), focusing on the implications of privacy, anonymization, large and complex data, and security on its design at multiple levels of abstraction, including the Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). We rely on an architectural blueprint that promotes encapsulation, separation of concerns, reusability, and service-orientation while enabling a direct and easy deployment leveraging the appropriate technologies at each architectural level. At the SaaS level, BCI has five subsystems for data acquisition, integration,

analysis, management, and presentation. PaaS level contains development support, collaboration, and sharing of knowledge artifacts. Finally, at the underlying infrastructure level, the iDASH BCI relies heavily on resource virtualization to maximize the utilization potential of the available physical resources and accommodate a variety of research needs. Given the uncertainty of privacy, security, and data retention policies in public clouds, iDASH employs a mix of modern cloud technologies to provide its own private cloud based on "cell architecture" to sandbox each subproject and all code intended for production. Consequently, each scientific subproject has specific allocations of storage and computing resources based on initial needs and planned growth patterns. This approach provides a solid ground for implementing fault-tolerance, increase system robustness, and support increasingly large computations.

Public Health Relevance

The iDASH Biomedical Cyberinfrastructure aims to enable unprecedented access to bio-medical data and support for sharing, privacy and anonymization. It provides a HIPPAA-compliant environment with scalable storage, computational resources, and specialized tools to help researchers collaborate in novel ways, while accounting for the inherent security and privacy risks.

27. Post-transcriptional regulators of microRNA biogenesis regulate pathogenesis of cancer

Pavel Sumazin
Columbia University

MicroRNAs (miRs) have emerged as key regulators of both normal and pathologic phenotypes, including cancer. Fine grained regulation of their biogenesis, however, is still poorly understood and only a few of their key regulators have been characterized. In order to understand the extent and specificity of miR-biogenesis control and its role in tumorigenesis and cancer progression, we set out to identify these regulators and profile their targets. We developed and experimentally validated Poseidon, an algorithm for genome-wide inference of miR-biogenesis regulators. Poseidon identifies biogenesis-regulator candidates by assessing the mutual information between mature miR expression and the expression of its transcriptional unit. We used Poseidon to identify candidate miR-biogenesis regulators in glioma and ovarian cancer using matching miR and gene expression profiles from TCGA. In total, Poseidon identified hundreds of regulators in these tumors, including known biogenesis regulators DGCR8, HNRNPA1, DDX5 and LIN28. We validated selected predictions that target miRs in glioblastoma tumors and are differentially expressed across glioma prognosis. Differential expression of (1) mature miRs, (2) their host genes, (3) co-transcribed miRs, and (4) miR precursors were measured upon ectopic expression or RNAi-mediated silencing of the predicted regulators in glioma cell lines. For example, DDX10 was predicted to up-regulate miR-218 biogenesis, and we showed that its ectopic expression up-regulates miR-218 at the pri-miRNA stage. We showed that RBBP4 inhibits miR-23b after pri-miRNA processing by DROSHA, and that WDR8 inhibits miR-18b, and PRIC285 inhibits miR-196a and miR-25. In total, we validated over a dozen novel miR-biogenesis regulators. Taken together, these results suggest that miR biogenesis is a complex, context specific, and finely-regulated process, and that miR biogenesis regulators may influence tumor initiation and progression by altering the expression of specific tumor-suppressors and oncomirs or by modifying large miR programs.

Public Health Relevance

MicroRNAs are key regulators of both normal and pathologic phenotypes and their aberrant regulation is a hallmark of cancer, however, only a few of the genes that regulate their maturation have been identified. We describe computational methodology for identifying post-transcriptional drivers of microRNA aberrant regulation, showing that alterations to these genes may lead to large-scale miRNome dysregulation and may play a role in cancer initiation and progression.

28. Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID) Network i2b clinical Research Chart Integration

Adam Davis, Chad Evans, William Seffens, Tahsin Kurc, Andrew Post, Joel Saltz, Emily Turner, Joshua Smith, Debbie Nickerson, Gary Gibbons, Rakale Quarells
Morehouse School of Medicine

The Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID) Network is the largest study of African American (AA) resistant hypertensive cases and controls in the US. MH-GRID Network has a centralized bio-repository of well annotated data elements from cryopreserved bio-specimens, laboratory data, whole exome sequencing, and EHR data from four main sites; Grady Hospital, Kaiser Permanente Georgia, Jackson Heart Study, and Jackson Hinds Clinic. A comprehensive repository has been established with information on anthropometric, social-environmental, and demographic data. The combined data repositories are used for in-depth, phenotypic characterization and longitudinal biosurveillance data combined with sequence data. To capture the multidimensional EHR, clinical and genomic data acquisition, we used a combination of Clinical Research Assistant (CR-Assist), RedCap and SQL databases. The MH-GRID Centralized Bio-repository (MGCB) was populated with the extracted, transformed, and de-identified data from the aforementioned databases. The Analytic Information Warehouse (AIW) is used to establish key common data elements, derived clinical phenotype definitions, and clinical data marts from EHR data in the MGCB. The AIW tool computed the clinical phenotypes based on categories, coded values, and temporal patterns creating a Clinical Data Warehouse (CDW) environment as clinical data marts. This provides an additional layer of security separation between backend EHR systems and the MH-GRID environment. i2b2 was populated with the MGCB clinical data and derived variables computed within the AIW. In conclusion the proposed infrastructure provides a federated computing environment for healthcare analytics and research data analysis to the client yet does not impede locally centralized databases. MH-GRID provides a platform for 'virtual' disease registries, catalyzes comparative effectiveness research in high-health disparity settings, and accelerates the translation of personalized medicine into minority communities.

29. Common Terminology Services 2 (CTS2)

Harold Solbrig
Mayo Clinic

Common Terminology Services Revision 2 (CTS2) is a joint HL7/Object Management Group Standard for the query, interchange and update of terminological resources. The standard specifies a formal model and a corresponding collection of XML Schemas and APIs that allows disparate terminological resource including resource catalogs, value set repositories, ontologies, coding and classification systems, etc. to be federated into a single, virtual ecosystem. The standard has been implemented in a collation of RESTful interfaces on BioPortal and projects are

currently underway to provide CTS2 portals to the UMLS, PhinVads, the ONC Quality Measures value sets and the NCI LexEVS terminology services.

Public Health Relevance

The Common Terminology Services 2 (CTS2) standard provides a critical component for the dissemination and use of shared terminology in healthcare reporting. The CTS2 standard provides the equivalent of an "HTML" for terminologies which, in turn, enables generic "terminology browsers" that can access value sets, definitions and other resources wherever they are created and published.

30. Torsion Angle Normal Mode Analysis

Jenelle Bray
Stanford University

Torsion angle normal mode analysis is a powerful tool for modeling protein conformational change. Optimized torsion-angle normal modes, in which the normal modes move along curved paths in Cartesian space, reproduce protein conformational changes more accurately than Cartesian normal modes. Furthermore, orthogonalizing the displacement vectors from torsion-angle normal-mode analysis and projecting them as straight lines in Cartesian space does not lead to better performance than Cartesian normal modes. Clearly, protein motion is more naturally described by curved paths in Cartesian space. Consequently, torsion angle normal modes can be used to describe conformational changes of proteins in very few dimensions, and thus to exhaustively map out and explore their conformational landscapes.

Public Health Relevance

Normal mode analysis efficiently explores the conformational landscape of a protein, which could aid rational drug design.

31. Coupled segmentation and registration for the prostate cancer and atrial fibrillation image analysis

Yi Gao
Harvard Medical School

Registration could aid segmentation: Extracting the prostate from magnetic resonance ~(MR) imagery is a challenging and important task for medical image analysis and surgical planning. We present a unified shape-based framework to extract the prostate from MR prostate imagery. Specifically, we first explicitly address the registration problem by representing the shapes of a training set as point clouds. In doing so, we are able to exploit the more global aspects of registration via a certain particle filtering based scheme. In addition, once the shapes have been registered, a cost functional is designed to incorporate both the local image statistics as well as the learnt shape prior. Experiments, which include several challenging clinical data sets, are performed to highlight the algorithm's capability of robustly handling prostate registration and the overall segmentation task.

Segmentation could aid registration. For the atrial fibrillation treatment, the longitudinal registration between the pre-ablation and post-ablation MR images is performed to assess the ablation success and the recurrence rate. For that purpose, high registration accuracy

has to be achieved for the left atrium region. To solve this problem, we combine the segmentation of the left atrium into the cost functional of the registration and provide a coupled deformable registration scheme for longitudinal data sets. Experiments are conducted and accurate matching is observed for the data sets in particular for the left atrium region.

Public Health Relevance

We design the medical image computing techniques to assist the prostate cancer treatment planning, neurosurgery, and atrial fibrillation ablation by synergizing image segmentation and registration.

32. Multi-scale modeling to evaluate the mechanisms underlying patellofemoral pain syndrome

Saikat Pal
Stanford University

We have developed a multi-scale modeling framework to evaluate the mechanisms underlying patellofemoral (PF) pain syndrome. We created whole-body computational models to predict patient-specific muscle forces during activities of daily living. We created finite element models of the knee joint from patient-specific geometry and predicted PF joint stress. We found that females have greater PF joint stress compared to males; this finding may explain the greater prevalence of PF pain in females. The multi-scale modeling framework provided new insights into the causes of PF pain syndrome that are difficult to acquire from clinical evaluation or experimental analysis.

Public Health Relevance

Anterior knee pain syndrome is common, affecting millions of individuals nationwide and costing billions in health care spending. Current treatment methods are unpredictable and often unsatisfactory as this syndrome has many possible causes that are difficult to diagnose. The goal of this study is to determine if patients with anterior knee pain have greater bone and cartilage mechanical stress than pain-free control subjects.

33. Virtual Proteomics: Inferring global protein activity profiles by network based analysis of gene expression signatures

Mariano J. Alvarez
Columbia University

Despite recent advances in molecular profiling, assessing protein activity on a proteome wide basis remains an elusive target in molecular biology. Lack of methodologies to accurately and systematically measure protein activity constitutes a major impediment to the elucidation of key drivers of pathophysiological phenotypes as well as of the activity of targeted drugs. Here we introduce a “virtual proteomics” approach that leverages the increasingly accurate and context specific knowledge of regulatory networks, to infer the differential activity of proteins on an individual sample basis, in proteome-wide fashion. Specifically, we propose that the activity of a protein’s transcriptional targets, either direct ones for a transcription factor or indirect ones for a signaling protein, provides the most accurate assessment of its activity. We show that this approach effectively capture the changes in transcription factor (TF) protein activity after RNAi-mediated knock-down of the TF itself or after perturbation with small molecule antagonists. We extended our approach beyond the transcription factor space showing that our method is also suitable to estimate the activity of proteins involved in upstream signaling events. Virtual

proteomics analysis of BCR and CD40 stimulated B-cells showed a significant change in receptor CD79A, CD79B, and CD40 protein activity, as well as on 17 of the 28 BCR and 9 of the 15 CD40 pathway proteins.

Public Health Relevance

The current lack of methodologies to accurately and systematically measure protein activity constitutes a major impediment to the elucidation of key drivers of pathophysiological phenotypes. We describe and validate a method that, by leveraging cell context-specific regulatory networks, infers protein activity in a proteome-wide fashion from the increasingly accessible genome-wide expression profile data.

34. Modeling and Simulation in Biomechanics: An Investigator's Perspective on the Role of Simbios

Ahmet Erdemir
Cleveland Clinic

Simbios, a National Center for Biomedical Computing at Stanford, has been an integral component of research activities in the Erdemir Laboratory at the Cleveland Clinic. Interactions of the research team with Simbios started in 2007 through a collaboration with the Center for coupling musculoskeletal movement simulations with tissue deformations (see https://simtk.org/home/multidomain). Simbios and Simtk.org have been a significant resource for a wide range of biomechanical data and they provided the infrastructure for collaboration, dissemination, and open development of models. By reusing musculoskeletal models, e.g. Torso + Lower Extremity Model, with state-of-the-art simulation software, e.g. OpenSim, reinvention of the wheel was prevented and our team had a chance to focus on scientific explorations. By significantly decreasing the burden of maintaining an infrastructure, Simbios allowed hosting of our other federally funded projects, e.g. those with multiple national and international teams (see https://simtk.org/home/j2c), where we have established platforms to quantify cell deformations from body level joint loads. The dissemination platform enabled the research team to reach out and deliver models, data, and results for others to reuse (also see https://simtk.org/home/cobi & https://simtk.org/home/meshcomparison). Last but not least, Simtk.org infrastructure also enabled testing of open science philosophy (see https://simtk.org/home/openknee). A project for the open development of a knee joint (not just free and open access to source code and release packages but encouragement of contributions from others) was able to recruit a volunteer developer from Canada, who streamlined some our model development scripts. This activity encouraged submission of an application for a collaborating R01 grant with Simbios to openly develop knee joint models of different populations and extend Simtk.org infrastructure towards cloud computing.

Public Health Relevance

Computational models provide the footing to understand normal and pathological mechanics of the musculoskeletal system and to guide simulation-based intervention design. If founded on comprehensive experimentation, dependable models utilized in a predictive simulation framework can increase effectiveness of clinical decision-making. Free and timely dissemination, supported by open development and accessible cloud-based simulation infrastructures, can provide the opportunity to realize the scientific and translational value of such models by national

and international research groups, clinical teams, and orthopedic industry.

35. SBGrid.org: Lowering Barriers to Structural Biology Computation

Andrew Morin, Piotr Sliz
Harvard Medical School

The SBGrid Consortium is a computing collaboration of over 220 structural biology laboratories at more than 70 academic institutions and 3 pharmaceutical companies in 15 countries. SBGrid disseminates, administers and supports over 260 different software titles used in structural biology research with an emphasis on X-ray crystallography, SAXS, SANS, NMR, electron microscopy, structure visualization and bioinformatics. SBGrid provides enabling research computing support to our member laboratories through curation and maintenance of a comprehensive software library. Each of the more than 260 programs the Consortium offers in the SBGrid Software Library are automatically configured, tested and deployed to computer workstations in Consortium member laboratories without need for user intervention, and support a variety of OS and hardware platforms. Complementing our research computing support mission, the SBGrid Science Portal is web-based set of applications for seamlessly submitting massively parallel computations to the US cyber-infrastructure utilizing the Open Science Grid. Portal applications provide ease of use and lower the barriers to entry for structural biology grid computing. As active intermediary between scientist software developers and users, SBGrid expands access, facilitates communication and lowers the costs of computational structural biology research. The Consortium is operated as a member supported, NIH-compliant non-profit service center of Harvard Medical School.

Public Health Relevance

Our work most closely fits with the aims of the Center for Computational Biology. However, because we provide computing support to any research group engaged in structural biology research, we may thus fall under the purview of multiple different centers. As for public health relevance - SBGrid provides enabling support for the computational tools and techniques that underlie structural biology research, from the basic to the translational. By lowering barriers and expanding access to structural biology computing resources, SBGrid helps scientists make fundamental advances in public health related research.

36. Structure and function of Cholesteryl ester transfer protein in cholesterol transferring by electron microscopy and molecular dynamic simulation

Lei Zhang and Gang Ren
Lawrence Berkeley National Laboratory

Human cholesteryl ester mass can be transferred from atheroprotective high-density lipoproteins (HDL) to atherogenic low-density lipoproteins (LDL) by cholesteryl ester transfer protein (CETP), resulting higher probability of cardiovascular diseases. Finding out the mechanism of CETP in CE transferring would be an important basis and key step toward the rational design of new CETP inhibitors for treating cardiovascular diseases. Using electron microscopy, computer image processing and molecular dynamics simulation, we discovered that CETP connects HDL and LDL or VLDL by penetrating into HDL with its N-terminal domain and LDL or VLDL with its C-terminal domain. By structural analyses, after special conformation change, the internal pores of CETP can connect to a hydrophobic central cavity, thereby forming a tunnel for transferring the cholesteryl ester from donor to acceptor lipoproteins. These new insights provide a key basis for revealing the mechanism of CETP and designing new CETP inhibitor drugs.

Public Health Relevance

Our research is to discover the insights into the molecular mechanism of how “good” cholesterol was converted to “bad” cholesterol by combination of electron microscope imaging and molecular dynamic simulation. These new insights provide a key basis for revealing the mechanism of molecule and designing new inhibitor drugs to treat cardiovascular diseases.

37. From Single-SNP to Wide-Locus GWAS: Identifying Functionally Related Genes and Intragenic Regions in Small Sample Studies

Knut M. Wittkowski
The Rockefeller University

Genome Wide Association Studies (GWAS) have had limited success when applied to complex diseases. Analyzing SNPs individually requires several large studies to integrate the often divergent results. In the presence of epistasis between SNPs, intragenic regions, or genes, multivariate approaches based on the linear model (including stepwise logistic regression) often have low sensitivity and generate an abundance of artifacts.

Methods: Recent advances in distributed and parallel processing spurred methodological advances in non-parametric statistics. U-statistics for multivariate data (µGWAS) are not confounded by unrealistic assumptions (linearity, independence) and were recently extended to incorporate information about hierarchical data structures. For µGWAS, a particular hierarchical structure reflects the sequence of neighboring SNPs and recombination hotspots.

Results: This computational biostatistics approach increases power and guards against artifacts. In particular, it can identify clusters of genes around biologically relevant pathways and pinpoint functionally relevant intragenic regions.

Application: A study of only 185 children with a narrowly defined neurodevelopmental diagnosis and publicly available controls sufficed to integrate previous findings into biologically plausible hypotheses about the interplay of genetic risk factors. Most available drugs target regulatory processes at the level of the nucleus or cell membrane. By moving from single-SNP to wide-locus GWAS, µGWAS was able to include a cluster of genes controlling functional processes in the cytoplasm, adding evidence for *PI3K*, currently investigated in the treatment of inflammatory diseases, as a potential additional drug target in epilepsy.

Public Health Relevance

A novel computational biostatistics approach reduces the sample size requirements for genome-wide association studies from (tens of) thousands of subjects to a few hundred, so that comparative effectiveness research and personalized diagnostics/treatment can provide decisions fine-tuned to particular populations or individuals, respectively. Moreover, sub-group analyses of phase III trials can now suggest risk factors for adverse events and novel directions for drug development.

The work was done within the CTSA component of the roadmap, yet to make the resources more widely available, we are seeking a closer collaboration with the NCBC.