Network Biology SIG 2011
July 15th, Vienna, Austria

Abstracts

Anand, Praveen – *"Proteome-wide binding site network analysis of* M. tuberculosis*" (Panel)*

Nagasuma Chandra and Praveen Anand. Indian Institute of Science, India.

Mycobacterium tuberculosis, causative agent of tuberculosis claims 4700 lives everyday. Special traits of the bacteria such as dormancy, ability to quickly take-over immuno-compromised hosts and resistance adds to the complexity of the problem. In the year 2010 highest rate of Multi-Drug Resistance strains (MDR-TB) were observed and in some of the places an alarming 28% of reported cases were found to be of MDR-TB, thus rendering most of the widely used drugs ineffective. This had led to serious concerns, warranting exploration of newer strategies and not merely new drugs. Polypharmacology, referring to targeting multiple proteins with a single drug, could be one such strategy1. Even if one of the targets were to acquire resistance, the drug would still be active through its other targets. A novel computational approach for polypharmacology is proposed here.

Availability of protein structures at genome scales, result of structural genomics consortium, and well established protocols to obtain high-confidence protein structures computationally has helped us to put forward the entire structural proteome of MTb. The obtained protein structures are then subjected to series of analysis involving binding site detection and ligand associations through consensus of well-validated tools including in-house algorithms that use different approaches. All together 9029 putative binding sites were predicted. All-vs all comparison of 9029 predicted binding sites was carried out using PocketMatch2 and a binding site network was constructed wherein each node represents a binding site and the edge represents relation of similarity amongst them determined through PocketMatch score. Network clustering algorithm MCODE was used to obtain the cluster of most similar binding sites, which can then be targeted using a single drug. Validation of clustering algorithm was carried out using MOAD dataset that contains 9836 protein ligand complexes. All the residues with 4.0Å of the ligand were considered to be binding sites resulting in 28195 pockets and all vs all comparison of binding sites were carried out using PocketMatch. The network of binding sites obtained after applying a cut-off of 80% similarity using PocketMatch was later clustered using MCODE consisted of 11005 binding sites and it was observed that all binding sites of the same ligands clustered separately into subnetworks.The same procedure was followed for the binding-sites of proteome, which yielded network 403 binding-sites with 23 clusters.

A detailed and systematic analysis has been carried out to find out the similarities among various proteins at the level of binding sites so as to adopt the polypharmacology approach. The binding-site similarity network constructed at the proteome level will give us more insights into the functioning of the organism and develop better strategies to efficiently control the spread of tuberculosis.

Bader, Gary – *"Network and pathway information: collecting, visualizing and using for gene function prediction"* (Invited keynote)

Barber, Alan – *"Pythoscape: A software framework for generation of large protein similarity networks" (Panel and Poster)*

Alan E. Barber and Patricia Babbitt. Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, USA.

Due to the rapidly increasing size of biological data available to computational scientists, new methods are necessary that can accommodate and organize ever larger data sets. Protein similarity networks leverage biological databases to enable hypothesis creation about sequence-structure-function relationships using datasets that are too large and unwieldy for many other methods. We present Pythoscape, an interface and set of plug-ins implemented in Python that serves as a framework for fast and efficient generation of large protein similarity networks that can be outputted and visualized in other software packages (e.g. Cytoscape). As an example case study, we have generated sequence similarity networks of the glutathione transferase (GST) superfamily, an enzyme superfamily whose members function in cellular chemical detoxification pathways. These networks demonstrate the utility of Pythoscape to manage and visualize large datasets and highlight technical challenges for correct classification of the enzymatic functions of GST superfamily members.

Birol, Inanc – *"Establishing correlation networks between gene and miRNA expression" (Presentation)*

Inanc Birol, Gordon Robertson, Andy Chu, Peter Westervelt, Richard K Wilson, Timothy J Ley, Marco A Marra and Steven Jm Jones. BC Genome Sciences Centre, Canada; Washington University Medical School, USA.

High throughput sequencing of mRNA and miRNA repertoire of cells for large patient cohorts is now feasible. With large studies underway, it is desirable to establish correlations between gene and miRNA expression profiles by characterizing changes between samples. Still, a data-driven exploration of these relations for transcriptomes representing a specific biological condition is a significant undertaking. There are three major challenges in establishing an informative picture of these relations: Interactions between genes and miRNAs are highly coupled, highly nonlinear, and require a large number of observations to support robust conclusions. We developed a method to address these challenges by considering differences of log-transformed expression levels between samples measured by high throughput sequencing experiments.

In this study, we report on relations between about 36,000 genes and 700 miRNA expression levels, for N patients, with N(N-1)/2 data points. All two-dimensional projections of this space for gene i and miRNA j are statistically symmetric around the origin. The expected direction of correlation between a gene and a miRNAs is positive when the miRNA is expressed along with its host gene. The direction is expected to be negative when a gene is targeted by a miRNA. To help with prioritization of the estimated correlations, we also report the goodness of our line fits, defined by the R2 metric, which in this formulation correspond to the square of the Pearson's correlation coefficients. With this approach we analyzed the gene and miRNA sequencing data from a cohort of 170 patients in a cancer study. We observed that 55% of the correlations between gene and miRNA expressions were consistent with zero (defined by the 99.9% confidence interval) and the remaining ones were equally distributed between positive and negative correlations. Fig. a shows the top 250 correlations, defined by the magnitude of the regression line slopes and by R2, as a network of relations. miRNAs are represented by the yellow nodes, and genes are represented by the pink nodes, with some genes of interest highlighted in green. Positive and negative correlations are indicated by red and blue edges, respectively, with the thickness of the line indicating goodness of fit.

Three correlations of the human miRNA mir-126 are highlighted in Fig. b – d. EGFL7 is a host gene for this intronic miRNA, and we see the expected strong positive correlation of expression (Fig. b). We observe that mir-126 expression level is also strongly positively correlated with the expression levels of a number of genes, including SHANK3 (Fig. c), and strongly negatively correlated with the expression levels of other genes, including DLX4 (Fig. d). In our presentation, we will further discuss quantification of similarities and differences between sample cohorts, and the implications of such similarities and differences for personalized genomics.

Donaldson, Ian – *"iRefScape: A Cytoscape plugin for visualization and data mining of protein interaction data from iRefIndex" (Presentation)*

Ian Donaldson. University of Oslo, Norway.

The iRefIndex [1] consolidates protein interaction data from ten databases in a rigorous manner using sequence-based hash keys (http://irefindex.uio.no). Working with consolidated interaction data comes with distinct challenges: data are redundant, overlapping, highly interconnected and may be collected and represented using different curation practices[2, 3].

The iRefScape plugin for the Cytoscape graphical viewer has a number of features that address these challenges. We show how these factors impact on data-mining tasks and how our solutions address them in a simple and efficient manner. A uniform accession space is used to limit redundancy and support search expansion and searching on multiple accession types. Multiple node and edge features support data filtering and mining. Node colours and features supply information about search result provenance. Overlapping evidence is presented using a multi-graph and a bi-partite representation is used to distinguish binary from n-ary source data. A rough path finding tool allows fast detection of potential paths. And a synchronized adjacency-matrix view supports mining relationships between sets of disease proteins or other user defined groups.

The iRefScape plugin will be of interest to advanced users of interaction data. The plugin provides access to a consolidated data set in a uniform accession space while remaining faithful to the underlying source data. Tools are provided to facilitate a range of tasks from a simple search to knowledge discovery. The plugin also represents a number of strategies that will be of interest to other plugin developers.

The plugin can be installed from the Cytoscape plugin manager and additional documentation is available at http://irefindex.uio.no (see iRefScape: Data availability via Cytoscape).

1. Razick S, Magklaras G, Donaldson IM: iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics 2008, 9(1):405.

2. Turinsky AL, Razick S, Turner B, Donaldson IM, Wodak SJ: Literature curation of protein interactions: measuring agreement across major public databases. Database (Oxford), 2010:baq026.

3. Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ: iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford), 2010:baq023.

Doncheva, Nadezhda – *"Novel approach to the interactive visual analysis of protein structure and function" (Panel)*

Nadezhda T. Doncheva, Karsten Klein, Francisco S. Domingues and Mario Albrecht. Max Planck Institute for Informatics and TU Dortmund University, Germany.

The study of individual amino acid residues and their molecular interactions in protein structures is crucial for understanding structure-function relationships. In recent years, many exciting applications of residue interaction networks (RINs) derived from 3D protein structures were found to provide additional insights into the structural and functional roles of interacting residues (Csermely, TiBS, 33(12):569-576, 2008, doi:10.1016/j.tibs.2008.09.006). Therefore, we developed a novel approach to the interactive visual analysis of RINs (Doncheva et al., TiBS, 36(4):179-182, 2011, doi:10.1016/j.tibs.2011.01.002). In particular, our new and freely available Cytoscape plugin RINalyzer (http://www.rinalyzer.de) supports the simultaneous viewing and exploration of a residue network in 2D and the corresponding 3D protein structure in UCSF Chimera (see figure below). RINalyzer also computes numerous topological centrality measures to identify critical residues in protein structures. Additionally, the 2D network representation is very useful for the comparison of protein structures and the investigation of long-range residue interaction paths, which are difficult to follow in 3D structure viewers.

Furthermore, RINalyzer is complemented by the command-line tool RINerator, which generates user-defined RINs and distinguishes different types of non-covalent residue interactions. In addition, the database RINdata (http://rinalyzer.de/rindata.php) contains over 50,000 pre-computed RINs for protein structures deposited in the PDB.

In conclusion, our software tools have the great potential to become standard applications for structural biologists, complementing other existing approaches to structure visualization and analysis. Our new approach can be particularly valuable in a number of biological and medical application scenarios that utilize protein structures. The possible applications include the characterization of the molecular effects of protein residue mutations and the study of residue interactions in protein binding interfaces, for example, of disease-associated proteins with residue substitutions.

Eichinger, Felix – *(see Mirel, Barbara)*

Gillis, Jesse – *"Making the best use of coexpression in network analysis with applications to human disease" (Presentation)*

Jesse Gillis, Meeta Mistry, Vaneet Lotay and Paul Pavlidis. University of British Columbia, Canada.

RNA coexpression data is commonly used to construct gene networks, but is often considered to be more difficult to interpret than protein interactions. This is in part due to the lack of consensus on methods for constructing networks from expression profiles, and the relatively poor performance of coexpression for function prediction. On the other hand, coexpression affords a major advantage over current large-scale protein interaction databases: it can be used to create "condition-specific" networks. Here we show that by appropriate consideration of data pre-treatment, aggregation and network construction, coexpression networks become a powerful tool for gene function analysis, on par or better than protein interaction networks in terms of many key properties while providing condition specificity, and discuss recent applications of these ideas to the study of human disease. Coexpression Network Analysis

In a recent paper [1], we laid out grounds for treating previous gene network analyses related to function with scepticism. We showed that gene networks (protein interactions, genetic interactions and coexpression) tend to encode very generic information about gene function in node degree, leading to highly multifunctional genes (which tend to have high node degree) dominating analyses to the point that details of network structure have a surprisingly small impact. We suggested that this property plays a dominant role in most previously reported network analyses. Here we consider approaches for addressing this problem for the case of coexpression networks. Multifunctionality bias can creep in to coexpression analyses in subtle ways (e.g., gene representation across microarray platforms). We find that while individual microarray studies are very noisy, careful aggregation yields a high quality network (as judged by gene function prediction performance using various algorithms, as well as semantic similarity) with much lower multifunctionality bias than comparable protein interaction networks. We will present our findings as to "best practises" surveyed across a large collection of public microarray data sets. Furthermore, because coexpression networks built in this way are not so generically swamped by enrichment of multifunctional/prevalent/promiscuous/hub genes, we show that they exhibit strong specificity to the data from which they were constructed (e.g., brain expression data, age specific data, disease specific).

We will present results applying these ideas to several application areas in neurogenomics. Our most detailed discussion will be of an analysis of difference in gene networks in schizophrenia as assessed by gene expression in human brain. We identified differences in functions enriched by node degree, as well as the modularity of candidate genes identified by differential expression analysis. We will also discuss analyses of human and mouse data for the purpose of prioritizing candidate genes in the genetics of other human neurological phenotypes and disorders including Huntington's disease.

1. Gillis, J. and P. Pavlidis, The impact of multifunctional genes on "guilt by association" analysis. PLoS One, 2011. 6(2): p. e17258.

Hermjakob, Henning – *"Reactome, IntAct, PSICQUIC: Bringing pathways and interactions together"* (Invited keynote)

Ideker, Trey – *"Rewiring of genetic networks by DNA damage"* (Invited keynote)

Iotem, Esti Yeger – *"The ResponseNet web server: Revealing signaling and regulatory networks linking genetic and transcriptomic screening data" (Panel)*

Alex Lan, Ilan Smoly, Guy Rapaport, Esti Yeger-Lotem. Department of Computer Science, Department of Software Engineering, and Department of Clinical Biochemistry and National Center for Biotechnology in the Negev, Ben-Gurion University, Israel.

Cellular response to stimuli is typically complex and involves both regulatory and metabolic processes. Large-scale experimental efforts to identify components of these processes often comprise of genetic screening and transcriptomic profiling assays. We previously established that in yeast genetic screens tend to identify response regulators, whereas transcriptomic profiling assays tend to identify components of metabolic processes (1).

ResponseNet is a network-optimization approach that integrates the results from these assays with data of known molecular interactions (1). Specifically, ResponseNet identifies a high-probability sub-network, composed of signaling and regulatory molecular interaction paths, through which putative response regulators may lead to the measured transcriptomic changes. Computationally, this is achieved by formulating a minimum-cost flow optimization problem and solving it efficiently using linear programming tools. When applied to screening data of a yeast model for alpha-synuclein toxicity, ResponseNet successfully mapped previously unknown as well as recognized pathways responding to alpha-synuclein toxicity. In particular, four de-novo predictions suggested by ResponseNet analysis were experimentally validated, including the presence of nitrosative stress, the involvement of the TOR pathway, the disturbance of the sterol biosynthesis pathway, and the mode-of-action of the genetic suppressor Gip2 in the response to alpha-synuclein toxicity (1).

The ResponseNet web-server offers a simple interface for applying ResponseNet. Users can upload weighted lists of proteins and genes and obtain a sparse, weighted, molecular interaction sub-network connecting their data. The predicted sub-network and its gene ontology enrichment analysis are presented graphically or as text. Consequently, the ResponseNet web-server enables researchers that were previously limited to separate analysis of their distinct, large-scale experiments, to meaningfully integrate their data and substantially expand their understanding of the underlying cellular response.

ResponseNet is available at http://bioinfo.bgu.ac.il/respnet

1. Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R.. Lindquist S. and Fraenkel E. (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet*, 41, 316-323.

Kelder, Thomas – *"Pathway interactions in insulin resistant mouse liver" (Presentation)*

Thomas Kelder, Lars Eijssen, Robert Kleemann, Marjan Van Erk, Teake Kooistra and Chris Evelo. Maastricht University, Department of Bioinformatics and Quality of Life, Vascular and Metabolic Disease, TNO, NL.

Complex phenotypes such as insulin resistance involve different biological pathways, which may interact and influence each other. To investigate relevant pathway interactions in the context of insulin resistant mouse liver, we developed an analysis approach that integrates gene/protein interaction networks, biological pathway information and experimental data to identify regulated paths between known pathways. This analysis was applied to the NuGO PPS2 transcriptomics dataset [1], which contains gene expression measurements in liver before and at different time points after a glucose challenge on normal and obesity-induced insulin resistant mice.

First, a protein interaction network was created by combining different resources, such as transcription factor targets (PAZAR), protein interactions (STRING) and curated reactions and interactions from pathway databases (KEGG and WikiPathways). Secondly, based on the experimental data, weights were assigned to the edges in this network. Thirdly, for each pair of pathways, a set of non-redundant shortest paths between their proteins was found where the length of each path is calculated by the sum of the weights of its participating edges. By assigning a lower weight for edges between nodes for which the corresponding genes are more differentially expressed in the dataset, paths that include regulated genes will get a shorter length. A pathway interaction network was generated based on the assumption that the more and shorter paths found between two pathways, the more likely it is that they interact.

We identified regulated pathway interactions for different comparisons between the diet groups and time points. The resulting networks provided new insights in which pathway interactions might be affected by insulin resistance and how these interactions change in response to a glucose challenge. We also zoomed in to the underlying protein interactions to study possible mechanisms and identify key proteins involved in pathway cross-talk. This helped us to generate hypotheses about underlying mechanisms and possible downstream effects of certain groups of differentially regulated genes, thereby providing starting points for more focused follow-up studies or experiments. Studying pathway interactions provides a new, systems level perspective on the data that complements established pathway analysis methods such as enrichment analysis. Including protein interaction networks increased the coverage of the analysis and allowed us to look beyond well established knowledge typically annotated to pathways. The analysis approach described here can be applied to different types of high-throughput data and could therefore be useful for analysis of other complex datasets as well.

1. Baccini M, Bachmaier E-M, Biggeri A, et al. The NuGO proof of principle study package: a

collaborative research effort of the European Nutrigenomics Organisation. Genes & nutrition. 2008;3(3-4):147-151.

Li, Jing – *"Disease gene prioritization by integrating multiple networks" (Panel)*

Yixuan Chen, Wenhui Wang and Jing Li. Case Western Reserve University, USA.

Identifying disease genes is crucial to the understanding of disease pathogenesis, and to the improvement of disease diagnosis and treatment. Traditional linkage analysis or association studies may return many candidate genes that show moderate to high signals. In recent years, many researchers have proposed approaches to prioritize these candidate genes by considering the relationship of candidate genes and existing known disease genes reflected in other data sources, such as human protein-protein interaction networks. In this paper, we propose an expandable framework for gene prioritization that can integrate multiple heterogeneous data sources by taking advantage of a unified graphic representation. Gene-gene relationships and gene-disease relationships are then defined based on the overall topology of each network using a diffusion kernel measure. These relationship measures are in turn normalized to derive an overall measure across all networks, which is utilized to rank all candidate genes. Based on the informativeness of available data sources with respect to each specific disease, we also propose an adaptive threshold score to select a small subset of candidate genes for further validation studies. We performed large scale cross-validation analysis on 110 disease families from the Online Mendelian Inheritance in Man (OMIM) database using three data sources based on protein interactions, gene expressions and pathway information. Results have shown that our approach consistently outperforms other two state of the art programs. Taking Parkinson disease (PD) as a case study, we tested our approach by considering all 3,243 genes that are shared by all three data sources. We identified four candidate genes (UBB, SEPT5, GPR37 and TH) that ranked higher than our adaptive threshold, all of which are involved in the PD pathway. In particular, a very recent study has observed a deletion of TH in a patient with PD, which supports the importance of the TH gene in PD pathogenesis. A web tool has been implemented to assist scientists in their genetic studies.

Mirel, Barbara – *"Gaining knowledge and coherence from complex networks and interactive activity trails" (Presentation)*

Barbara Mirel, Felix Eichinger, Juliana Freire, Mike Smoot and Terry Weymouth. University of Michigan, University of Utah, and University of California San Diego, USA.

For researchers in translational medicine who want to generate hypotheses about molecular mechanisms of a disease, biological networks integrate and display massive amounts of diverse measurement data, such as gene expression data. Through structure and visual codings, the networks characterize the data and bring in additional annotations. Annotations come from such sources as: the Gene Ontology (GO); pathway, disease/phenotype, and protein interaction databases; and data derived from Medline-mined natural language processing, term enrichment, and graph theoretic algorithms. However, cognitive load is high for exploring such visually and information-rich networks, especially when seeking to achieve and maintain a coherent flow, which is a prime success factor in this class of analysis (Neressian, 2008). As our prior user modeling research shows, coherence requires being able to monitor progress; recall reasoning processes; remember, repeat or adapt prior moves later in the workflow; and critically assess the implications of earlier choices on next steps and evolving meaning.

Part of the complexity of exploratory network analysis and its coherence-seeking tasks is translational researchers' conventional divide-and-conquer strategy. Researchers divide unwieldy, high dimensional hairballs into smaller subnetworks (subproblems) and explore iterative, nonlinear, often opportunistic ways. Then they have to reconstruct sub-solutions into an overall solution. Yet many of the subproblems have side conditions or bias, resulting from the artificial isolation process. Researchers need to factor this into the reconstruction as well. Thus within and across subproblems, coherence is challenging. The research we propose to discuss focuses on alleviating the cognitive burden of these metacognitive tasks for coherence. In collaboration with Cytoscape developers and the VisTrails team (an open source activity/provenance tracking tool from the University of Utah) we are researching, developing, and user-testing a system that will automatically record translational researchers' actions, present an interactive representation of their history at semantic/pragmatic levels and feed that information back into the network analysis program for visualization and meta-analysis.

In the workshop we will describe the realistic storyboard that guides our prototyping efforts, its component tasks for capture, and its representation as an activity trail. The prototyping is in-progress and will be completed by the workshop. We will explain the technologies facilitating activity capture and presentation, and highlight intrinsic challenges. We will discuss strategies for dealing with challenges, seeking insights from workshop participants. Some examples of challenges include the following:

1) Reconciling temporal capture of actions with spatial presentations of activity in trees or networks
2) Representing/grouping actions at levels of granularity meaningful to translational researchers
3) Switching between levels of abstraction
4) Using visualization and cognitive support to identify important steps and repetitive actions
5) Using task factories and/or macros to formalize and streamline task groups
6) Condense graphic events (e.g., manual move of nodes) into meaningful actions and capture them in context.

As background, our system aims to unite VisTrails and Cytoscape and give scientists a way to interactively track, replay, critically review and share their activities. It brings together our interdisciplinary expertise in Cytoscape 3.0; provenance tracking, prior integration of VisTrails into applications, integration of plug-ins into Cytoscape, cognitive task analysis, user modeling, and user testing; and biomedical research in renal disease. Task patterns, taxonomies, and representations from our earlier research and user models are the linchpin for turning automatically captured low-level actions into interactive presentations at the semantic and pragmatic levels at which researchers conceptualize their visual analytics.

Nitsch, Daniela – *“PINTA: A web server for network-based gene prioritization from expression data” (Presentation and Panel)*

Daniela Nitsch, Leon-Charles Tranchevent and Yves Moreau. KU Leuven, ESAT-SCD, Belgium.

A major challenge in human genetics is to identify novel disease genes. Genetic studies identify chromosomal regions involved in a disease or phenotype of interest, but often result in large lists of candidate genes of which only few can be followed up for further investigation. Identifying among such a list the most promising candidate genes for a disease of interest has been defined as the gene prioritization problem. Most of the already existing gene prioritization tools are using a guilt-by-association concept, and are therefore not applicable when little is known about the phenotype or when no confirmed disease genes are available beforehand.

In a recent paper [1], we have proposed a method that overcomes this limitation by replacing prior knowledge about the biological process by experimental data on differential gene expression between affected and healthy individuals. At the core of the method are a protein-protein interaction network and disease-specific expression data. The program propagates the expression data over the network using a Random Walk approach. Candidate genes are ranked based on the differential expression of their neighborhood. Our method relies on the assumption that strong candidate genes tend to be surrounded by many differentially expressed neighboring genes in a genomewide protein-protein interaction network. This allows the detection of a strong signal for a candidate even if its own differential expression value is too small to be detected by a standard analysis, as long as its interacting partners are highly differentially expressed.

We have implemented our method as a user-friendly and easy accessible web tool that we have termed PINTA [2], designed for the prioritization of candidate genes based on the differential expression of their neighborhood in a genome-wide protein-protein interaction network. PINTA is dedicated to the study of genetic disorders for which only little is known beforehand or when no background knowledge is assumed. PINTA relies on the presence of disease specific expression data, which makes it particularly attractive to study genetic conditions for which such expression data can easily be collected. PINTA propagates the expression data over the network using several random walk strategies. This allows the detection of a strong signal for a candidate gene even if its own differential expression value is small. PINTA is available for some prominent model organisms besides human (mouse, rat, worm, and yeast) and various array platforms are supported. PINTA provides for the top ranked genes a graphical view of the strongest contributing interacting genes and their expression signal in the network that leads to the candidate's strong scoring signal. By doing this, the user can easily assess the importance of a top-ranked gene in its network neighborhood influenced by its neighboring genes' expression levels. A benchmarked on 40 mouse knockout experiments has shown that PINTA outperforms traditional approaches [1].

1.Nitsch D et al. (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC Bioinformatics, 11, 460.

2.Nitsch D et al. (2011) PINTA - a web server for network-based gene prioritization from expression data. Accepted for publication in Nucleic Acids Research.

Pinney, John – *"Disease gene identification by data fusion over multiple networks" (Panel)*

Yoli Shavit, Nathan Harmston, Michael Stumpf and John Pinney. Imperial College London, UK.

The inference of novel causal and associated genes, given a disease of interest, is an important task at the crossroads of medicine and the biological sciences. Genes relate to each other in many ways, each type of information revealing another aspect of a complex picture. For example, data on human gene-gene relationships is now available from analyses of multiple transcriptomics and proteomics experiments, from genome annotations via the concept of semantic similarity, and more generally through text mining resources. Although several tools are available to produce lists of candidate genes to be tested for roles in a particular disease, existing methods are often based on a restricted number of input data types, and as such cannot integrate all of the available evidence. Tools that do exploit multiple data sources do so in a uniform manner, so that each data source is considered to have the same importance as the others, across all diseases. In this work, we develop an integrated, network-based approach, in which the evidence from each data source is combined with that of the others in proportion to its predictive value for a particular set of input genes, leading to improved results and supporting a greater understanding of the different causal pathways of genetic diseases.

A web-based application has been developed to allow the user to compute results for diseases and phenotypes taken from the OMIM database or to submit a custom list of genes. Results are presented as a ranked list of genes, with their overall relevance scores and genomic locations. In addition, the reliability weight of each network for the seed gene set can also be viewed, providing extra information about the types of evidence that were found to be informative for that particular disease. The JNets network visualization and investigation tool was chosen for the visualization of the top scoring candidate genes in the context of the various data networks. The top 50 candidates, seeds and induced networks are presented in an interactive network layout (see Figure). Various filter menus allow the user to view evidence associated with each network separately.

Pržulj, Natasa – *"Integrative network alignment and analysis: MI-GRAAL and GraphCrunch" (Presentation)*

Natasa Pržulj. Department of Computing, Imperial College London, UK.

We demonstrate that detailed topological comparisons between biological networks of different species are capable of producing new biological knowledge and corroborating existing sequence-based knowledge, including independently reproducing phylogenetic relationships.

Sequence-based computational approaches have revolutionized biological understanding. However, they can fail to explain some biological phenomena. Since proteins aggregate to perform a function instead of acting in isolation, the connectivity of protein-protein interaction (PPI) and other molecular networks will provide additional insights into the inner workings of the cell, over and above sequences of individual proteins. We argue that both network topology and sequence give insights into complementary slices of biological information, which sometimes corroborate each other, but sometimes do not. Hence, advances will depend on the development of sophisticated graph-theoretic methods for extracting biological knowledge purely from network topology before being integrated with other types of biological data (e.g., sequence). However, dealing with large networks is non-trivial, since many graph-theoretic problems are computationally intractable, requiring the development of heuristic algorithms.

Analogous to sequence alignments, alignments of biological networks will likely impact biomedical understanding. Network alignment (NA) allows detailed comparison between protein-protein interaction networks. Previous network alignment methods used information external to networks, e.g., sequence, because no good algorithm for purely topological alignment existed. Since it is important to understand how much information can be extracted from each source of biological data individually, e.g. sequence versus network topology, we introduce a topology-based NA algorithm, GRAphALigner (GRAAL), that produces by far the most complete alignments of biological networks to date that are obtained from network topology alone: our alignment of yeast and human PPI networks indicates that even distant species share a surprising amount of topology [1]. Next, we introduce a network alignment algorithm, called Matching-based Integrative GRAph ALigner (MI-GRAAL), which can integrate *any number and type* of similarity measures between network nodes including, but not limited to, any topological network similarity measure, sequence similarity, functional similarity, and structural similarity [2]. MI-GRAAL exposes the largest functional, connected regions of PPI network similarity to date: surprisingly, it reveals that 77.7% of proteins in the baker's yeast high-confidence PPI network participate in such a subnetwork that is fully contained in the human high-confidence PPI network. This is the first demonstration that species as diverse as yeast and human contain large, continuous regions of *global* network similarity. We apply MI-GRAAL's alignments to predict functions of un-annotated proteins in yeast, human and bacteria. Furthermore, using network alignment scores for PPI networks of different herpes viruses, we reconstruct their phylogenetic relationship. This is the first time that phylogeny is exactly reconstructed from purely topological alignments of PPI networks.

High-throughput methods for detecting molecular interactions have produced large biological network data sets with much more yet to come. Hence, the problems of biological network modeling, comparison, alignment and clustering are becoming important. We introduce GraphCrunch 2, a software tool that implements the latest research on biological network analysis. It is the only software tool that simultaneously implements methods to address all of the above mentioned problems based solely on network topology [3]. It parallelizes computationally intensive tasks to fully utilize the potential of modern multi-core CPUs. It is open-source and freely available for research use. It runs under the Windows and Linux platforms.

1. O. Kuchaiev, T. Milenkovic, V. Memisevic, W. Hayes, and N. Pržulj, Topological network alignment uncovers biological function and phylogeny, *Journal of the Royal Society Interface,* 7:1341-1354, 2010.

2. O. Kuchaiev and N. Pržulj, Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human, *Bioinformatics,* March 16, 2011.

3. O. Kuchaiev, A. Stefanovic, W. Hayes, and N. Pržulj, GraphCrunch 2: Software tool for network modeling, alignment and clustering, *BMC Bioinformatics*, 12(24):1-13, 2011. http://bio-nets.doc.ic.ac.uk/graphcrunch2/

Schelhorn, Sven-Eric – *"Computational identification of physical protein contacts from large-scale purifications of protein complexes" (Presentation and Panel)*

Sven-Eric Schelhorn, Julian Mestre, Mario Albrecht and Elena Zotenko. Max Planck Institute for Informatics, Germany; University of Sydney, Australia.

Background: High-throughput data sets of protein complex purifications have allowed new insights into the organization of cellular protein complexes. Several computational approaches have been developed to extract functionally relevant protein complexes from purification data. These methods, however, do not distinguish between direct physical protein contacts and indirect, bridging protein interactions. Consequently, the network of physical protein contacts within purified protein complexes is not well characterized, although this information is crucial for understanding protein complex structure and organization.

Methods: We have developed ISA [1], a novel computational approach for interpreting large-scale protein complex purifications. In contrast to other schemes for scoring protein purifications, ISA is specifically designed for inferring direct physical protein contacts from purification data and uses an advanced statistical null model to consider experimental replicates in a statistically correct fashion.

Results: We assessed ISA and four existing purification scoring schemes using several experimental reference sets. In contrast to prior studies that reported low enrichments of true physical protein contacts in protein complex purifications, our results show that purification data can be used to infer high quality physical contacts once a proper scoring method such as ISA is employed. Additionally, we demonstrate that ISA particularly excels at discovering physical contacts involving proteins that have been screened multiple times in purification experiment. We highlight this advantage of ISA by performing case studies on two biologically highly relevant purification data sets containing protein kinases and molecular chaperones, respectively.

Conclusion: Our findings indicate that protein complex purifications can be exploited to infer physical contacts at a quality level comparable to experimental binary assays such as yeast two-hybrid. By using a scoring scheme that is devised to consider experimental replicates, physical contacts can be reliably identified even in challenging purification data sets that are dominated by unspecific protein interactions. We therefore propose that physical contacts derived from protein complex purifications could be used to complement large-scale binary protein interaction assays in genome-wide interactome screens.

1. Sven-Eric Schelhorn, Julian Mestre, Mario Albrecht, and Elena Zotenko. Inferring physical protein contacts from large-scale purification data of protein complexes. Mol Cell Proteomics, 2011 (advance online publication), doi:10.1074/mcp.M110.004929

Turinsky, Andrei – *“Chromatin modification networks and disease annotations” (Presentation)*

Andrei L. Turinsky, Brian Turner, James A. Gleeson, Shuye Pu, Thomas Switzer, Sabry Razick, Ian M. Donaldson and Shoshana J. Wodak. Hospital for Sick Children, Canada; University of Oslo, Norway.

Chromatin modification (CM) is a crucial epigenetic mechanism that affects DNA replication, transcription and repair, and its disruption is implicated in the development of many diseases. CM is carried out by groups of physically interacting proteins. Yet there remains a dearth of public resources and analysis tools that explore the relationship between chromatin machinery and human diseases, especially in the context of protein-interaction networks.

To fill this gap, we integrated and analyzed data on CM genes and proteins, their interaction patterns, and their relationship to human diseases. This analysis was enabled by two bioinformatics resources recently developed in our group. One is the Disease-Annotated Chromatin Epigenetics Resource (DAnCER, http://wodaklab.org/dancer/), which contains a large collection of CM-related genes and proteins, their known disease annotations from OMIM and other supporting evidence. The other resource is iRefWeb (http://wodaklab.org/iRefWeb), a web interface to a broad landscape of protein-protein interaction data. The data were consolidated from ten major public databases using the iRefIndex procedure. DAnCER and iRefWeb are seamlessly integrated with each other, and are freely available to the community.

Initial examination revealed a significant enrichment of cancer annotations among all CM-related disease annotations. To further characterize the distribution of disease patterns across CM genes and their interacting partners, we combined the existing data from DAnCER and iRefWeb with new types of disease annotations and supporting evidence from five additional disease sources. This allowed us to quantify various network properties of both disease-related and CM-related proteins, including their propensity to interact with other disease- and/or CM-related proteins.

Furthermore, we characterized cases where disease annotations for a gene (protein) of interest may be reliably inferred from its interaction patterns with other disease-associated genes, especially when such associations are confirmed by evidence from multiple sources. We then systematically applied predictive methods to quantify the likelihood of such disease associations, based on the analysis of the interaction-network topology, the reliability of individual protein-protein interactions, and the similarity between different disease types. We also measured the effect of the choice of the initial disease-data sources, as well as the quality of the protein interactions, on the resulting network-based predictions. Our approach demonstrates a useful and practical way to explore disease-related data in the context of protein-interaction networks, using complementary tools and public resources. It also allows us to formulate new hypotheses on the function and disease associations of genes involved in important cellular processes, such as chromatin modification.

Valencia, Alfonso – *"Using and completing cancer networks"* (Invited keynote)

Wu, Guanming – "*Reactome Functional Interaction (FI) Cytoscape plugin: A network module-based tool for cancer data analysis" (Presentation)*

Guanming Wu, Irina Kalatskaya, Christina Yung, Robin Haw and Lincoln Stein. OICR, Canada.

Cancer re-sequencing projects and other high-throughput experiments on diseases usually generate many disease candidate genes. Some of them are true driver genes, while others are passengers. We believe that network-based approaches, supported by human-curated reliable pathway data, can help discover true disease driver genes, understand their mechanisms of action, and assist in the design of effective therapeutic agents.

By combining hand curated pathways in Reactome (http://www.reactome.org) and other pathway databases with functional interactions predicted using a machine learning technique, we have built a functional interaction (FI) network covering close to half of human proteins in SwissProt (Wu G, Feng X, and Stein L. Genome Biol 2010; 11(5):R53). To leverage this highly reliable FI network, we have built a Cytoscape plug-in called the Reactome FI Cytoscape plug-in (screenshot below). The FI plug-in can construct a FI sub-network based on a gene or protein list, cluster the network with a very fast modularity-drive spectral partition clustering algorithm we developed in house (Newman, ME. PNAS 2006;103(23): 8577-82), and then annotate those clusters. For cancer data analysis, the plug-in integrates the caBIG cancer gene index annotation data set (cabig.nci.nih.gov/inventory/data-resources/cancer-gene-index) to show the cancer disease ontology, highlight annotated genes for selected cancer terms, and display annotations for genes in the network. The plug-in has integrated connectivity to Reactome database, allowing it to retrieve and display human drawn pathway diagrams. We have also built tools into the plug-in to allow it to perform clinical survival analysis and correlate these results with network clustering results. The FI Cytoscape plug-in can be downloaded from wiki.reactome.org/index.php/Reactome_FI_Cytoscape_Plug-in, and chianti.ucsd.edu/cyto_web/plug-ins/displayplug-ininfo.php?name=Reactome%20FIs.

Posters

Canevet, Catherine – "*Using Ondex to predict a protein-protein interaction network for the cereal infecting fungal pathogen* Fusarium graminearum*"*

Duerr, Oliver – "*Using community structure for complex network layout"*

Ironi, Liliana – *"A computational tool for qualitative simulation of gene-regulatory network dynamics"*

Kuchinsky, Allan – *"An integrated network biology approach to elucidating anti-microbial mechanisms and counteracting drug resistant* Mycobacterium tuberculosis*"*

Kutmon, Martina – *"The importance of modularity in bioinformatics tools"*

Milenkovic, Tijana – "*Network analysis uncovers key biological processes and pathways in molecular networks"*

Rogojin, Vladimir – *"Predicting essential nodes for robustness in cancers via feedback loops in cellular networks"*

Scardoni, Giovanni – *“Network centralities interference and robustness”*

Valentini, Giorgio – *“A cost-sensitive neural algorithm to predict gene functions using large biological networks”*

Yu, Yi-Kuo – *“ppiTrim: constructing non-redundant and up-to-date interactomes”*

Yu, Yi-Kuo – “*Network analysis tools from the QMBP group”*