ECCB 2014 Workshop W07 on:
Integrative Dynamic Analyses of Large Biomedical Network Data

List of abstracts

Igor Jurisica
University of Toronto

"Avoiding fusion of illusion and confusion – reducing bias in network-based analyses"

Protein-protein interactions (PPIs) are essential for understanding signaling cascades, associating proteins with disease or predicting their function, and fathoming drug mechanism of action. However, with estimated only ~10% of the human PPIs currently known, and about one-third of human proteins without any known interactions, human interactome is highly incomplete and potentially biased. We show that these "orphans" are not a random subset of the proteome, which poses a significant challenge since the current understanding of the interactome may be biased and unreliable.

Precision medicine needs to be data driven and corresponding analyses comprehensive and systematic. We will not find new treatments if only testing known targets and studying characterized pathways. Thousands of potentially important proteins remain pathway or interactome "orphans". Computational biology methods can help fill this gap with accurate predictions, but the biological validation and further experiments are essential. Intertwining computational prediction and modeling with biological experiments will lead to more useful findings faster, while optimizing resources.

We introduce a data mining-based method to predict physical PPIs proteome-wide. Overall, the in silico predicted interactome comprises 250,542 high confidence physical PPIs among 10,529 human proteins, including 1,089 PPI orphans. These computational predictions improved human interactome coverage relevant to both basic and cancer biology, and importantly, helped us to identify, validate and characterize prognostic signatures. Combined, these results may lead to unraveling mechanism of action for therapeutics, re-positioning existing drugs for novel use and prioritizing multiple candidates based on predicted toxicity, identifying groups of patients that may benefit from treatment and those where a given drug would be ineffective. Compared to previous PPI prediction methods, FpClass achieves better agreement with experimentally detected interactions. The high validation rate suggests that FpClass could help guide high-throughput screening, in a combined computational-experimental approach to interactome mapping.

Roded Sharan,
Tel Aviv University

"Protein networks: from topology to logic"

Protein networks have become the workhorse of biological research in recent years, providing mechanistic explanations for basic cellular processes in health and disease. However, these explanations remain topological in nature as the underlying logic of these networks is to the most part unknown. In this talk I will describe the work in my group toward the automated learning of the Boolean rules that govern network behavior under varying conditions. I will highlight the algorithmic problems involved and demonstrate how they can be tackled using integer linear programming techniques.

Alfonso Valencia
Spanish National Cancer Research Centre
"Building an Epigenetic Network with Co-Evolutionary Methods"
valencia@cnio.es, www.cnio.es

My group is interested in the reconstruction of protein networks using computational methods designed to disentangle direct and indirect interactions and in particular those classified under the general category of co-evolution based approaches (*Juan et al., 2013*).

In this case we have collected the heterogeneous high-throughput epigenomic datasets available in public repositories for mouse embryonic stem cells (mESCs) and processed the information to establish a comprehensive network of chromatin components and DNA methylation states.

The resulting network involves different types of cytosine modifications, epigenetic marks and chromatin remodelers, as well as their predicted functional relations. The network was analyzed to determine potential direct interactions (see *Lasserre et al., 2013*) and possible coevolutionary relations (see *Juan et al, 2008)*. In the workshop I will present the initial network model together with the new methodology developed for this study.

This work corresponds to the paper in preparation by *Carrillo de Santa Pau, Perner, Juan et al., (2014) and it was developed in collaboration with Martin Vingron's lab (MPIMG, Berlin) in the context Blueprint EU consortium (*www.blueprintepigenome.eu*).*

References

Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc. Natl. Acad. Sci. U.S.A. *105*, 934–939.
Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. Nat. Rev. Genet. *14*, 249–261.
Lasserre, J., Chung, H.-R., and Vingron, M. (2013). Finding associations among histone modifications using sparse partial correlation networks. PLoS Comp. Biol. *9*, e1003168.

Jan Baumbach
University of Southern Denmark
"Integrating multiple omics data types with biological networks"

Systems biology has emerged over the last decade. Driven by the advances in sophisticated omics technology the research community generated huge molecular biology data sets. This comprises rather static data on the interplay of biological entities, for instance protein-protein interaction network data, as well as quite dynamic data collected for studying the behavior of individual cells or tissues in accordance to changing environmental conditions, such as DNA microarrays, RNA sequencing or genome-wide methylation assays. Computational network enrichment brings the two different data types together for unraveling the molecular basis of complex diseases. It has become popular in systems biology to elucidate aberrant network modules. Traditionally, these approaches focus on combining gene expression data with protein-protein interaction (PPI) networks. Modern omics technologies, however, allow for inclusion of many more data sets, e.g. protein phosphorylation or epigenetic modifications. This creates a need for analysis methods that can combine these various sources of data to obtain a systems-level view on the dynamics of biological networks. In the talk, I will introduce our recent work on KeyPathwayMiner, a network enrichment software suite that is not limited to analyses of single omics data sets, e.g. gene expression, but is able to directly combine several different omics data types with biological networks. It is available via http://keypathwayminer.mpi-inf.mpg.de or by using the Cytoscape's app manager.

Citation: Alcaraz N, Friedrich T, Koetzing T, Krohmer A, Mueller J, Pauling J, Baumbach J (2012) Efficient key pathway mining - Combining networks and OMICS data. Integr Biol., 2012, 4(7), 756-764.

Michael Kramer
University of California San Diego
"Inferring gene ontologies from molecular networks"

Ontologies have proven very useful for capturing knowledge as a hierarchy of terms and their interrelationships. In biology a major challenge has been to construct ontologies of gene function given incomplete biological knowledge and inconsistencies in how this knowledge is manually curated. Here I will discuss our result that large networks of gene and protein interactions in *Saccharomyces cerevisiae* can be used to infer an ontology whose coverage and power are equivalent to those of the manually curated Gene Ontology (GO). Our first attempt, the Network eXtracted Ontology (NeXO, www.nexontology.org) contains 4,123 biological terms and 5,766 term-term relations, capturing 58% of known cellular components. New terms in NeXO were supported with quantitative genetic interaction profiling and chemogenomics and several were added as updates to GO. I will also discuss our recent work which has advanced our ability to infer ontologies from molecular networks by factoring the challenge into two steps: 1) combining experimental molecular networks into an integrated, quantitative similarity score network and 2) inferring an ontology in the form of a directed acyclic graph (DAG) from this integrated network. I will discuss the Clique Extracted Ontologies (CliXO) algorithm, which solves step 2 by directly inferring a DAG from a pairwise similarity network. CliXO, unlike all other algorithms tested, shows that nearly 100% of the syntactic information in a complex ontology such as GO can be encoded in a semantic similarity network (CliXO >99% precision, recall. Others <20% precision, recall). Furthermore, CliXO outperforms other algorithms at ontology inference when applied to experimental 'omics data. This work enables a shift from using ontologies to evaluate data to using data to construct and evaluate ontologies.

Co-authors: Michael Kramer[1], Janusz Dutkowski[1], Michael Yu[1], Michal A Surma[2,3], Rama Balakrishnan[4], J Michael Cherry[4], Nevan Krogan[2,5], Vineet Bafna[6], and Trey Ideker[1]
1. Department of Medicine, University of California San Diego, La Jolla, California, USA
2. Deparment of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, California, USA
3. Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
4. Department of Genetics, Stanford University, Stanford, California, USA
5. J. David Gladstone Institutes, San Francisco, California, USA
6. Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA

Teresa Przytycka
US National Institutes of Health
"Towards systems level analysis of tumor heterogeneity"

Uncovering and interpreting genotype-phenotype relationships are among the most challenging open questions in disease studies. In cancer, uncovering these relationships is complicated even further due to the heterogeneous nature of the disease. Over the years, we have developed several algorithms that help to analyze heterogeneous cancer data in the context of uncovering genotype-phenotype relations, identification of dysregulated pathways, and cancer classification. These approaches span a large spectrum of algorithmic techniques including optimization-based techniques and mixture models. Taken together, these approaches help to leverage datasets collected through TCGA and other initiatives for better understanding of cancer and cancer diversity.

Ben Raphael
Brown University
Associate Professor, Department of Computer Science
Director, Center for Computational Molecular Biology

"Algorithms for Analyzing Mutated Networks and Pathways in Cancer"

The rapidly declining costs of DNA sequencing have enabled large-scale measurement of somatic mutations in many cancer samples. However, distinguishing the subset of mutations that drive cancer development from random passenger mutations is a notoriously difficult problem. Individuals with the same cancer type typically exhibit different combinations of *driver* mutations, reducing the power of single mutation/gene tests of recurrence. We describe two approaches to predict *driver pathways*, groups of genes containing driver mutations, in a large cohort of cancer samples. In one approach, we identify subnetworks of genome-scale interaction network using our *HotNet* algorithm, which models mutations on a network as a heat diffusion process. In a second approach, we identify multiple sets of mutually exclusive mutations using our *Dendrix* algorithm. We describe new extensions of these algorithms and their application to genome/exome sequencing and array copy number data from several cancer types in The Cancer Genome Atlas (TCGA).

Topology-Function Conservation in Protein-Protein Interaction Networks

Darren Davis[1], Ömer Nebil Yaveroğlu[1,2], Noël Malod-Dognin[2], Aleksandar Stojmirovic[3,4] and Nataša Pržulj[2]

[1] Calit2, University of California, Irvine, USA
[2] Department of Computing, Imperial College London. UK
[3] National Center for Biotechnology Information (NCBI). USA
[4] Janssen Research and Development, LLC. USA

Proteins underlay the functioning of a cell, and the relationships between the functions of proteins and their wiring (i.e., topology) in protein-protein interaction network (PIN) have been extensively studied in systems biology. It has been shown that proteins with similar functions and cellular locations tend to cluster together [1], and 70-80% of interacting protein pairs share at least one function in yeast PIN [2]. Based on these observations, the functions of proteins are predicted from their topological characteristics such as shared neighbourhoods or network flow [3]. It has also been shown that the functional similarities between proteins depend on the similarities of their wiring patterns, independent of their location in the PIN [4]. Network alignment algorithms have also been used for finding similarly wired proteins between PINs of different species and transferring functional annotations across aligned proteins [5]. However, all of these studies assume that, for each biological function, the wiring patterns of the proteins are similar. This might be an incorrect assumption since some essential functions might be better preserved than some species-specific functions during evolution. Moreover, some functions might be carried out by proteins with different wiring patterns in the PIN, hence it may not be possible to link these functions with specific network topologies.

To improve our understanding of the topology-function relationships in PINs, we develop a statistical framework that: (1) characterises the statistically significant PIN topology-function relationships in a given species, and (2) uncovers the PIN topology-function relationships that are conserved between baker's yeast and human. Our framework captures the wiring patterns of proteins in a PIN by using their graphlet degrees: the *graphlet degree* of a node is the number of graphlets (i.e., small, connected, non-isomorphic, induced subgraphs) that the node touches at a topologically distinct place (i.e., at a particular graphlet orbit) [4]. The functions of proteins are given by their Gene Ontology (GO) annotations. We apply our framework to PINs of yeast and human and characterise the significant topology-function relationships for each PIN separately. Among many species-specific GO terms that are identified to have statistically significant correlation with the proteins being wired via a particular graphlet orbit, we identify 7 biological process and 2 cellular component GO terms that have evolutionary conserved wiring in the PINs of yeast and human (Figure 1). We validate the obtained topology-function relationships for three of these GO terms – namely transcription initiation, localisation and protein complexes – by finding studies that link the biological mechanisms involved in these functions with the identified wiring patterns in PINs.

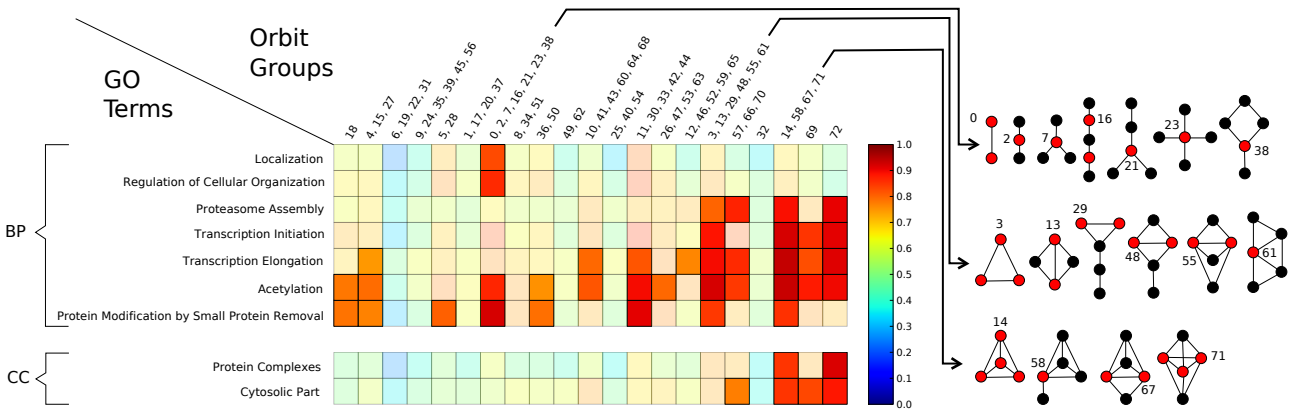

Figure 1: The wiring patterns of GO terms that have significantly conserved topology-function relationships for yeast and human. Each heatmap row represents the strength of the associations between a GO term and different wiring characteristics that are represented by graphlet orbits. Heatmap cells with solid colors represent the significantly strong associations between the GO terms and corresponding graphlet orbits. Similar wiring patterns (orbits) are grouped to simplify the illustration. Some of the important wiring patterns are illustrated on the right with the red nodes of the corresponding graphlets. For example, "Localization" is linked with mediator roles in sparse graphlets as illustrated by the red nodes of the graphlets at the top row.

The identified topology-function relationships improve our understanding on the evolution of PINs. The species-consistent topology-function relationships correspond to essential functions that are carried out similarly in different species. On the other hand, investigation of the species-specific topology-function relationships also raises interesting questions, such as why they are not preserved and how they differ across species. Furthermore, by utilising our framework as a filter for topology-based function prediction algorithms, we can further improve the accuracies of such algorithms. Finally, our framework is generic and can be applied to uncover consistent wiring patterns in different phenomena, including those of disease and KEGG pathway annotations of proteins.

References

[1] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, 2006.

[2] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nature Biotechnology*, vol. 21, no. 6, pp. 697–700, 2003.

[3] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, no. 1, 2007.

[4] T. Milenković and N. Pržulj, "Uncovering biological network function via graphlet degree signatures.," *Cancer Informatics*, no. 6, 2008.

[5] C. Clark and J. Kalita, "A comparison of algorithms for the pairwise alignment of biological networks," *Bioinformatics*, p. btu307, 2014.

Integrative, dynamic, and comparative biological network research of aging

Fazle E. Faisal,[1,2,3] Yuriy Hulovatyy[1,2,3] Han Zhao,[1,4] Vikram Saraph,[1,5] and Tijana Milenković[1,2,3,*]
[1]Department of Computer Science and Engineering, University of Notre Dame
[2]Interdisciplinary Center for Network Science and Applications, University of Notre Dame
[3]ECK Institute for Global Health, University of Notre Dame
[4]Department of Computer Science and Technology, Tsinghua University
[5]Department of Computer Science, Brown University
*Corresponding author (e-mail: tmilenko@nd.edu)

Since the US population is on average growing older because of ~78 million baby boomers who have begun turning 65 in 2011, and since susceptibility to diseases increases with age, studying molecular causes of aging gains importance. However, human aging is hard to study experimentally due to long lifespan and ethical constraints. Therefore, human aging-related knowledge needs to be inferred computationally. Analyses of gene expression or sequence data have been indispensable for investigating human aging computationally. However, these analyses are limited to studying genes (or their protein products) in isolation, ignoring their connectivities. But proteins do not function in isolation; instead, they carry out cellular processes by interacting with other proteins. And, this is exactly what biological networks, such as protein-protein interaction (PPI) networks, model; in PPI networks, nodes represent proteins and edges represent physical interactions between the proteins. Thus, analyzing topologies of proteins in PPI networks could contribute to the understanding of the processes of aging.

The majority of the current methods for analyzing *systems-level* PPI networks deal with their static representations, due to limitations of biotechnologies for PPI collection, even though cells are dynamic. For this reason, and because different data types can give complementary biological insights, we integrate current static PPI network data with age-specific gene expression data to computationally infer dynamic, age-specific PPI networks. Then, we apply a series of sensitive measures of topology to the dynamic PPIs to study cellular changes with age. We find that while global PPI network topologies do not significantly change with age, local topologies of a number of genes do. We predict such genes to be aging-related. We demonstrate the credibility of our predictions by: 1) observing significant overlap between our predicted aging-related genes and "ground truth" aging-related genes; 2) observing significant overlap between functions and diseases that are enriched in our aging-related predictions and those that are enriched in "ground truth" aging-related data; 3) providing evidence that diseases which are enriched in our aging-related predictions are linked to human aging; and 4) validating our high-scoring novel predictions in the literature. To validate the robustness of our approach, we base our predictions on two different gene expression data, where one is obtained by microarray technology and another one is obtained by RNA-seq technology. Indeed, the choice of gene expression data does not make significant effect on our aging-related predictions.

In our subsequent study, we show that network de-noising via state-of-the-art link prediction methods (including ours), followed by repeating the above analyses on the de-noised network data, improves the quality of the aging-related predictions.

In addition, the knowledge about human aging has typically been transferred from highly annotated model species (such as, yeast, fly, and worm) to poorly annotated human via genomic sequence comparison. However, PPI network topology data and genomic sequence data can give complementary biological insights, so non-sequence data can elucidate aging-related knowledge missed by current sequence-based methods. Also, since not all genes implicated in aging in model species have sequence orthologs in human, restricting comparison to sequence data may limit the knowledge transfer. We hypothesize that network alignment, which aims to find regions of topological and functional similarities between PPI networks of different species, can help with the transfer of aging-related knowledge between conserved network regions of different species. By using state-of-the-art network alignment algorithms, including those developed by us, we show that network alignment can indeed successfully predict new aging-related knowledge. We validate our predictions in independent aging-related data sets. To our knowledge, we are the first to use network alignment in the context of aging.

References:

1. Fazle E. Faisal and Tijana Milenković (2014), **Dynamic networks reveal key players in aging**, *Bioinformatics*, 30(12): 1721-1729.
2. Yuriy Hulovatyy, Ryan W. Solava, and Tijana Milenković (2014), **Revealing missing parts of the interactome via link prediction,** *PLOS ONE*, 9(3): e90073.
3. Boyoung Yoo, Huili Chen, Fazle E. Faisal, and Tijana Milenković (2014), **Improving identification of key players in aging via network de-noising**, *In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics* (ACM BCB), Newport Beach, CA, USA, September 20-23, 2014.
4. Fazle E. Faisal, Han Zhao, and Tijana Milenković (2014), **Global Network Alignment In The Context Of Aging**, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP(99):1, DOI: 10.1109/TCBB.2014.2326862.
5. Tijana Milenković, Han Zhao, and Fazle E. Faisal (2013), **Global Network Alignment In The Context Of Aging**, *In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics* (ACM BCB), Washington DC, USA, September 22-25, 2013 (acceptance rate: 28%).
6. Vikram Saraph and Tijana Milenković (2014), **MAGNA: Maximizing Accuracy in Global Network Alignment**, *Bioinformatics*, DOI: 10.1093/bioinformatics/btu409.

Global Analysis of coRegulation for the identification of functional modules

Rim Zaag[1], Etienne Delannoy[1], Marie-Laure Martin Magniette[1,2]

Plant Genomics Research, Evry, France[1], AgroParisTech, Paris, France[2]

One of the challenges faced by genomics currently is the understanding of gene function. Genome wide analysis of gene function mostly relies on guilt by association approaches through coexpression analysis taking advantage from the availability of transcriptome data. Indeed, cluster analysis of gene-expression profiles can be used to propose functions based on the assumption that coexpressed genes have likely related biological functions (Eisen et *al.*, 1998). Generally co-expression is performed by analyzing correlations between all the gene pairs from multiple microarray experiments collected from international repositories. Such approach has two drawbacks: First it leads to a local point of view about functional modules and second the dataset is composed of heterogeneous transcriptome results.

In constrast, we performed a global analysis of highly homogeneous transcriptome data extracted from CATdb (Gagnot et *al*., 2008). The whole dataset is composed of more than 18 000 genes described by 424 expression differences dealing with stress conditions. The coexpression analysis is performed through a model-based clustering method which allows the modelisation of the whole dataset by a mixture of distributions. A study of the Bayesian Information Criterion (Schwarz, 1978) as a function of the component number allows the evaluation of the fit between the data and the mixture. Once the assessment is done, the selected mixture is the one with the highest value of BIC.

Without a priori knowledge, the model has guided us to divide the whole dataset in twenty types of stresses leading to the identification of gene clusters having the same pattern of response under a single stress type. However coexpressed genes are not necessarily coregulated and then are less likely to be functional partners. To find groups of coregulated genes, we integrated these coexpression studies by calculating the occurrence number in a same cluster for each gene pair. Some pairs have a coordinated transcriptional response in up to 15 different types of stress and a resampling procedure showed that a gene pair observed in the same cluster in more than 4 stresses is significant. This approach allows us to focus our study on the potential key players of stress responses. Furthermore, the resulting coregulated gene network reveals an interesting topology of highly connected substructures. Preliminary analyses of these components containing orphan genes showed that they are more homogeneous than coexpression clusters highlighting probable functional modules.