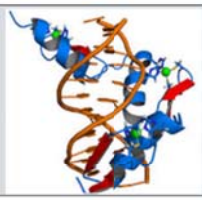

The 16th Israeli Bioinformatics Symposium – IBS

Bar-Ilan University, Ramat-Gan, June 24th 2014

Bar-Ilan University

Program

9:00-9:30 Gathering and Coffee

Opening session / Chairperson: Yael Mandel-Gutfreund

9:30-9:40 Welcome Remarks

9:40-10:30 Keynote Lecture 1: Lior Pachter (UC Berkeley):
What gene expression reveals about a person (and their biology)

10:30-11:00 Coffee

First Session: Protein interaction, recognition & engineering / Chairperson: Ron Unger

11:00-11:20 Miki Kosloff (Haifa):
Deciphering and re-designing interaction specificity between signaling proteins

11:20-11:40 Dana Reichman (HUJI): A role of protein disorder in cellular stress response

11:40-12:00 Avraham Samson (BIU):
Elastic network normal modes reveal the GPCR activation mechanism

12:00-14:00 lunch & posters

13:40-14:00 Business meeting

Second Session: Genomes and Evolution / Chairperson: Edward Trifonov

14:00-14:20 Gal Chechik (BIU): Spatio-temporal patterns of the mammalian brain transcriptome

14:20-14:40 Itay Mayrose (TAU): The evolutionary consequences of whole genome duplications

14:40-15:00 Idan Menashe (BGU): Co-expression profiling of autism genes in the mouse brain

15:00-15:50 Coffee & posters

Third Session: Applications / Chairperson: Ron Pinter

15:50-16:10 Ido Bachelet (BIU): Thought-controlled molecular robots in a living host

16:10-16:30 Avner Schlessinger (Mount Sinai School of Medicine):
Computer-aided discovery of nutrient uptake modulators in reprogrammed cancer pathways

16:30-16:40 Hagit Alon (Prize4Life): PRO-ACT: big data – the next step towards ALS cure

16:40-17:00 Coffee

Nobel Session/ Chairperson: Joel Sussman

17:00-17:15 Poster prize Ceremony

17:15-17:20 Introduction

17:20-18:10 Keynote Lecture 2: Michael Levitt (Stanford):
Birth and Future of Multi-Scale Modeling of Macromolecules

18:10-18:20 Concluding remarks

Sponsors:

DELL

i-CORE
Israeli Centers of Research Excellence

Edmond J. Safra Center for Bioinformatics Tel Aviv University

The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering
המרכז הבין תחומי למדעי החיים וההנדסה ע"ש לורי לוקי

TAU HUJI HH BIU SMC
Gene Regulation in Complex Human Disease

Chromatin & RNA Gene Regulation

Bar-Ilan University:
Office of the VP of research
The faculty of Exact Sciences
The faculty of Medicine
The Mina and Everard Goodman Faculty of Life Sciences

SPEAKER ABSTRACTS

What gene expression reveals about a person (and their biology)

Lior Pachter

Departments of Mathematics, Molecular & Cell Biology and Electrical Engineering & Computer Science, University of California at Berkeley

lpachter@math.berkeley.edu

Deciphering and re-designing interaction specificity between signaling proteins

Mickey Kosloff

Department of Human Biology, University of Haifa, Israel

kosloff@sci.haifa.ac.il

For cellular signaling cascades to function correctly, their protein components must recognize their appropriate partners accurately. This requirement presents a challenge for living cells, as *related* components are used repeatedly in both parallel and intersecting cascades within the same cell. Signaling therefore requires that the interactions of *particular* protein-family members be tailored to each signaling cascade via interaction specificity. Understanding the structural basis for such selectivity is a major goal in both experimental and computational biology. Yet, beyond single representative examples, little is known of how specificity is determined among members of large protein families, including those involved in signal transduction. We developed a “bottom-up” approach to decipher interaction specificity, integrating experimental and computational methods to map specificity determinants at the protein family level. The resulting “residue-level maps” are then used to redesign proteins with altered activities and specificities, offering new insights into protein-protein interactions and paving the way for the engineering of signaling networks at the cellular level.

[1]A role of protein disorder in cellular stress response

Dana Reichman

The Alexander Silberman Institute of Life Science, The Hebrew University of Jerusalem

danare@mail.huji.ac.il

Elastic network normal modes reveal the GPCR activation mechanism

Dikla Kolan, Gennadiy Fonar, Avraham O. Samson

Faculty of Medicine in the Galilee, Bar-Ilan University, Zfat

Avraham.Samson@biu.ac.il

G-protein-coupled receptors (GPCR) are a family of membrane-embedded metabotropic receptors which translate extracellular ligand binding into an intracellular response. Here, we calculate the motion of several GPCR family members such as the M2 and M3 muscarinic acetylcholine receptors, the A2A adenosine receptor, the β2-adrenergic receptor, and the CXCR4 chemokine receptor using elastic network normal modes. The normal modes reveal a dilation and a contraction of the GPCR vestibule associated with ligand passage, and activation, respectively. Contraction of the vestibule on the extracellular side is correlated with cavity formation of the G-protein binding pocket on the intracellular side, which initiates intracellular signaling. Interestingly, the normal modes of rhodopsin do not correlate well with the motion of other GPCR family members. Electrostatic potential calculation of the GPCRs reveal a negatively charged field around the ligand binding site acting as a siphon to draw-in positively charged ligands on the membrane surface. Altogether, these results expose the GPCR activation mechanism and show how conformational changes on the cell surface side of the receptor are allosterically translated into structural changes on the inside.

Spatio-temporal patterns of the mammalian brain transcriptome

Gal Chechik

The Gonda brain research center, Bar Ilan University

gal.chechik@biu.ac.il

The evolutionary consequences of whole genome duplications

Itay Mayrose

The Bioinformatics Unit, George S. Wise Faculty of Life Sciences, Tel-Aviv University,

itaymay@post.tau.ac.il

Polyploidy (or whole-genome duplication) is a widespread feature of extant organismal diversity, particularly in plants, yet its importance to evolution has long been debated. In this talk I will present a computational framework for the identification of polyploid lineages across a vast array of taxonomic groups. I will then exemplify how such a framework can be used to tackle long standing questions in polyploidy research.

Co-expression profiling of autism genes in the mouse brain.

Idan Menashe[1,2], Pascal Grange[3], Eric Larsen[2], Sharmila Banerjee-Basu[2], and Partha Mitra[3]

[1]Department of Public Health, Faculty of Health Sciences, Ben-Gurion University of the Negev, [2]MindSpec Inc., [3]Cold Spring Harbor Laboratory, Cold Spring Harbor

idanmen@bgu.ac.il

There is significant evidence that the onset and severity of autism spectrum disorder (ASD) is governed in part by complex genetic mechanisms affecting the normal development of the brain. To date, a number of genes have been associated with ASD, however, the temporal and spatial expression of these genes in the brain remains unclear. To address this issue, we examined the co-expression network of 26 autism genes from AutDB (http://mindspec.org/autdb.html), in the framework of 3,041 genes with the highest quality expression data from the Allen Mouse Brain Atlas database (http://mouse.brain-map.org). These data were derived from in situ hybridization experiments conducted on male, 56-day old C57BL/6J mice co-registered to the Allen Reference Atlas, and were used to generate a normalized co-expression matrix indicating the cosine similarity between expression vectors of genes in this database. The network formed by the autism-associated genes showed a higher degree of co-expression connectivity than seen for the other genes in this dataset (Kolmogorov–Smirnov P = 5x10-28) suggesting common neuroanatomical properties. Using Monte Carlo simulations, we identified two cliques of co-expression genes that are significantly enriched with autism genes (A Bonferroni corrected P< 0.05). Genes in these cliques were significantly over-expressed in the Cerebellar Cortex (P = 1 x 10-5) suggesting possible implication of this brain region in autism. In conclusion, our study, provides a detailed profiling of co-expression patterns of autism genes in the mouse brain, and suggests both specific brain regions and new candidate genes involved in autism etiology.

Thought-controlled molecular robots in a living host

Ido Bachelet

The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel

Ido.Bachelet@biu.ac.i

We report a new type of brain-machine interface enabling a human operator to control nanometer-size robots inside a living animal by thought. Recorded EEG patterns indicating cognitive load are recognized in real-time by an algorithm, which in turn controls the state of an electromagnetic field. The field induces the local heating of billions of DNA origami robots inside the animal, leading to their reversible activation and subsequent exposure of a therapeutic payload. This technology enables the online switching of a molecule on and off in response to a subject's cognitive state, with potential implications to therapeutic control in disorders such as schizophrenia, depression, and attention deficits, which are among the most challenging conditions to diagnose and treat.

Computer-aided discovery of nutrient uptake modulators in reprogrammed cancer pathways

Avner Schlessinger

Department of Pharmacology and Systems Therapeutics, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai

avner.schlessinger@mssm.edu

Alterations in cell metabolism support rapid growth and proliferation of cells - key hallmarks of cancer. Solute Carrier (SLC) transporters are membrane proteins that transport solutes such as metabolites and drugs across membranes, and play a major role in mediating nutrient delivery in reprogrammed cancer metabolism networks. Here, we describe a structure-based discovery approach to identify small molecule modulators for key nutrient SLC transporters. In particular, we use comparative modeling, virtual screening, and experimental testing to identify ligands, including drugs, metabolites, and lead-like molecules for the amino acid transporters LAT-1 and ASCT2, which function cooperatively in cancer metabolism and are highly expressed in the blood-brain-barrier (BBB). Our results may explain some of the pharmacological effects (i.e., efficacy and/or side effects) of known drugs via polypharmacology, and rationalize the enhanced brain permeability of two drug-like molecules. Finally, two of our hits inhibited proliferation of a cancer cell line by distinct molecular mechanisms, providing useful chemical tools to characterize reprogrammed metabolic networks, as well as a framework for developing efficacious drugs against these key targets and other SLC transporters.

PRO-ACT: big data – the next step towards ALS cure

Hagit Alon

Prize4Life

alonh@prize4life.org

Large datasets are critical for identifying and emphasizing biologically-relevant observations that are statistically significant, especially in rare diseases like ALS. The Pooled Resource Open-access ALS Clinical Trials (PRO-ACT) platform provides an unprecedented opportunity to increase our understanding of the heterogeneity of the ALS patient population and of the natural history of the disease. The PRO-ACT database covers the records of over 8,500 ALS patients who participated in 17-phase II and phase III clinical trials. Over 8 million longitudinally collected data-points include demographics, family history, vital signs, clinical assessments, lab data, medications, and survival information of the patients. Open-access to researchers worldwide has been made in December, 2012, and has attracted the attention of over 200 researchers from 23 different countries, which yielded at least 6 research papers in different publication stages. Given the nature of ALS as a rare disease with heterogeneous characteristics predicting disease progression is highly important. To that end, we undertook an innovative crowd-sourcing initiative to shed light on the difficult challenge of ALS prognosis. The DREAM- Phil Bowen ALS Prediction Prize4Life was launched to incentivize the development of novel methods to accurately predict future change in ALSFRS at the individual patient level. This program brought in over 1000 solvers from around the world and led to the development of several valuable algorithms to predict the progression of ALS, with potential to aid both clinicians and future ALS clinical trials. The challenge also led to the identification of new features predictive of ALSFRS progression that have been later verified on the full PRO-ACT database, and will be presented as well. Those algorithms have a remarkable added value: as they reduce the cost of clinical trials by 20%. Additional preliminary results also include insights in terms of correlation between ALSFRS slope and several baseline variables; the ability to establish initial relations between different ALSFRS questions and disease progression in patients; and identification of features predictive of ALS survival. Initial analysis of the data provided insights in terms of correlations between ALSFRS slope and several baseline variables. In addition, relations between different ALSFRS questions could help categorize disease progression in patients. These early results demonstrate the value of large datasets for developing a better understanding of ALS natural history, prognostic factors, and disease variables. More sophisticated and targeted analyses will continue to reveal even more about this disease, which has for so long defied our understanding.

Birth and Future of Multi-Scale Modeling of Macromolecules

Michael Levitt

Department of Structural Biology, Stanford University School of Medicine

michael.levitt@stanford.edu

POSTER ABSTRACTS

What can we learn from a system view on hypertension?

Alon Botzer, John Moult and Ron Unger
Bar-Ilan University
alon.botzer@biu.ac.il

Background: Essential hypertension is a condition affecting over 1 billion people worldwide, commonly associated with serious illnesses such as stroke, cardiovascular disease, diabetes mellitus and renal disease; hence it is considered \the silent killer\. Since hypertension is a complex trait disease its causes and underlying mechanisms remain poorly understood. The consequences of this condition and its growing prevalence are primary drivers for innovative methods required to further explore new directions that can promote effective treatment and reduce the likelihood of occurrence of this complex condition. Goal: To study essential hypertension on a system perspective making use of bioinformatics tools in order to gain new insights that are not pronounced when focusing on a details-based resolution. Methods: A] Construction of a protein- protein interaction network graph for hypertension associated genes in order to identify key elements that may play a significant role in this constellation based on graph centrality analysis. Enriched gene regulatory elements (transcription factors and microRNAs) were extracted by motif finding techniques and knowledge-based tools that could enhance the understanding of the nature of hypertension system regulation. B] Creation of an organism-to-gene orthology map based on pairwise ortholog scores relative to human. Implementing a two-way hierarchical clustering on map scores suggests the evolutionary order by which the circulatory system has evolved. C] Comparing gene targets of hypertension-indication drugs with targets of hypertension-induced side- effect drugs. This analysis clarifies the clinical pathways and mechanisms that influence hypertension. Results: This combined approach yielded a new insight regarding elements that regulate the hypertension gene-network on the translation and post-transcription levels - SP1 EZH2 miRNA27 and miRNA548C as well as highlighting the central role that Insulin may play in inducing hypertension. The evolutionary analysis confirms the developmental order of the blood circulatory mechanism in respect to affiliated genes. The pharmacological study confirms the role of dopaminergic and adrenergic receptors as major pathways affecting hypertension and strengths the understanding regarding the causes of hypertension as a prevalent consequential side-effect among many common drugs. Discussion: We view blood pressure regulation module as a system-of- systems comprising several contributing sub-systems and pathways rather than a single mechanism. Such a regulatory pattern is affected by a large number of distributed elements that could be considered for therapeutic purposes. To our surprise Insulin proved to be significantly central suggesting a primary role in hypertension. This finding also highlights the tight link between essential hypertension and the metabolic syndrome diseases.

Variability in genes and processes across human tissues

Shoval Tirman and Esti Yeger-Lotem
Ben-Gurion University of the Negev
shovalt@bgu.ac.il

Genetically identical cells exhibit remarkable phenotypic diversity. One of the sources for this diversity is variation in the expression of genes and the processes they constitute. Here we aim to characterize this variation across human tissues. Starting by analyzing the variability of genes across tissues. We find that ubiquitous genes are less variable than genes expressed in a restricted number of tissues, and that genes associated with diseases are more variable than non-disease genes. In addition, we find that genes that undergo transcriptional regulation are more variable than other genes, and this tendency is increased among genes that are regulated by a high-variability transcription factor. We then define process variability by the variability in expression of process genes, and apply this definition to biological pathways from the KEGG database. We find that the proteasome, an essential protein complex responsible for protein degradation, is of low variability, as opposed to the high variability PPAR signalling pathway that is composed of nuclear hormone receptors that are activated by fatty acids and their derivatives.Genetically identical cells exhibit remarkable phenotypic diversity. One of the sources for this diversity is variation in the expression of genes and the processes they constitute. Here we aim to characterize this variation across human tissues. Starting by analyzing the variability of genes across tissues. We find that ubiquitous genes are less variable than genes expressed in a restricted number of tissues, and that genes associated with diseases are more variable than non-disease genes. In addition, we find that genes that undergo transcriptional regulation are more variable than other genes, and this tendency is increased among genes that are regulated by a high-variability transcription factor. We then define process variability by the variability in expression of process genes, and apply this definition to biological pathways from the KEGG database. We find that the proteasome, an essential protein complex responsible for protein degradation, is of low variability, as opposed to the high variability PPAR signalling pathway that is composed of nuclear hormone receptors that are activated by fatty acids and their derivatives.

VarElect: phenotype-based variation prioritizer in GeneCards

Gil Stelzer, Anna Alkelai, Tsviya Olende, Shahar Zimmerman, Marilyn Safran and Doron Lancet

Weizmann Institute of Science

gil.stelzer@weizmann.ac.il

Next generation sequencing has provided the scientific community with a key technology for deciphering the genetic cause of human diseases. Typical whole exome sequences depict ~25,000 non-reference coding variants; only a few signify disease. As a rule, one relies on criteria such as rarity in the general population, predicted damage, and evolutionary conservation in the encoded protein and segregation in more than one affected individual, to shorten the potential gene list.. We have constructed the VarElect tool for phenotype-dependent variant prioritization, leveraging the rich information and scoring mechanisms of GeneCards and MalaCards, the human gene and disease compendiums, to enable the zooming in on a handful or even just one candidate gene. Its algorithm affords inferring direct as well as indirect links between genes and phenotypes, matching the provided keywords best describing the disease and its symptoms. An example of an indirect GeneCards-based inference of GeneA to Phenotype PhenX is when the disease is found to be linked to GeneB, which in turn shares a pathway with GeneA. Such gene-to-gene relations are also formed (among others) by interaction networks, paralogy relations, domain-sharing, and mutual publications. MalaCards, in turn, allows one to produce a comprehensive phenotype search expression by utilizing this database's built-in information about diseases, their relationships, and their underlying symptoms. Thus VarElect provides a robust algorithm for ranking genes within a short list, and pointing out their likelihood to be related to a disease, which will prove indispensable in future clinical applications.

Understanding the Connection between Transcriptome and Proteome in Resting Non- Proliferating Tissues through the Example of the Auditory System

Kobi Perl, Yoni Bhonker, Ofer Yizhar-Barnea, Shaked Shivatzki, Kathy Ushakov, Yair Pozniak, Orly Yaron, Noam Shomron, Tamar Geiger, Karen B. Avraham and Ron Shamir

Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine and Sagol School of Neuroscience, Blavatnik School of Computer Science, Tel-Aviv University

kobipe3@gmail.com

The correlation between expression levels of protein and mRNA in mammals is relatively low with a Pearson correlation coefficient of ~0.40. To explain the remaining variation, some combination of post transcriptional regulation and measurement noise are invoked. This lack of correlation makes it difficult to integrate protein and mRNA data. Tools for data integration are sparse. Initial findings from such tools suggest that the transcriptional and the translational regulation evolved independently except in the rare occasions where a strong selection factor in favor of correlation was present. However, such claims are based on data from perturbed systems, where the observed discordance between the proteome and the transcriptome is significantly affected by the temporal de-synchronization of the transcriptional and translational regulation levels. We decided to focus on the connection of mRNA and protein levels in non- proliferating tissues, through the example of the auditory system. By performing a joint-analysis of RNA-seq and protein mass spectrometry data from the cochlea and the vestibule, we hope to shed light on the roles of the two regulation systems, to identify genes that are mainly regulated on one level or the other, and to infer their general features. The two tissues investigated are quite similar in structure, but have unique roles in hearing and balance. This allows us to ask questions about the differentiation of cells in respect to those two levels of regulation, specifically at which regulation system the functionality in encoded. (Schwanhäusser et al., 2011) measured the reaction rates of mRNA transcription, mRNA degradation, protein translation, and protein degradation. We asked whether those rates, measured in a cell line of mouse fibroblasts, are common to non-proliferating tissues from non-fibroblast origin. We tested this claim on our auditory system data, and on previously obtained proteomic and transcriptomic data from different mouse tissues. The performed analysis suggests that those rates can improve our prediction of protein abundance from mRNA data in other systems as well. Moreover, it suggests that while the protein translation and degradation rates are similar between functionally related tissues, they are less similar between tissues with very distinct functions, in a way that preserves a similar protein profile across all tissues. This is quite surprising as the mRNA profile is quite different between functionally distinct tissues. We hypothesize that functionally distinct tissues possess different mRNA profiles but similar protein profiles, in rest, as part of a preparation for a stimulus.

The asymmetric organization of phosphorylation systems across eukaryotes

Ilan Smoly, Esti Yeger-Lotem and Michal Ziv-Ukelson
Ben-Gurion University of the Negev
smolyi@bgu.ac.il

Post-translational modifications of proteins are a common cellular mechanism for regulating protein functions and adapting to changing conditions. The protein modifications are typically reversible, and are carried by specialized enzymes with opposing functions. A prominent modification is protein phosphorylation, which controls cellular responses to signals in all living organisms. Protein phosphorylation is exerted by kinases, which covalently bind a phosphate group to the target protein, and by phosphatases, which detach the phosphate group. Here we use the well-studied protein phosphorylation system of budding yeast to unravel key organizational features of phosphorylation systems. We show that yeast kinases can be organized hierarchically into layers with distinct target-specificity, essentiality, and amenability to regulation by phosphorylation. Phosphatases, in contrast, form a small set of robust and relatively unmodified regulators. Their tendency to remain static is in agreement with the unique impact of phosphatases on response time to signals. Additionally, phosphatases compensate for their markedly reduced gene numbers by broad target specificity and high protein abundance. These asymmetric characteristics are shared by phosphorylation systems of evolutionary-distant organisms including plant, fly and human. A subset of these characteristics is common to the protein acetylation and protein ubiquitination systems of eukaryotes.

Solving the microtargetome as a stable marriage problem quantifies the systemic mode of action of miRNAs

Nathanael Weill, Véronique Lisi, Nicolas Scott, Paul Dallaire, Julie Pelloux and Francois Major

Université de Montréal

naweill@gmail.com

In eucaryotes, gene expression is regulated by microRNAs (miRNAs). MiRNAs bind to messenger RNAs (mRNAs) and repress their translation or promote their degradation. There is still no algorithm to predict the miRNA- induced translational repression on each gene. Here, we solve the whole set of miRNA::mRNA interactions at equilibrium (microtargetome) in a given cellular context, which improves the precision of target predictions by more than 20% over the commonly employed target prediction tools, and allows us to run virtual experiments, such as gene deletion or overexpression. The algorithm implements two fundamental properties of the microtargetome. First, miRNAs bind a target mRNA depending on its abundance, that of the target, and their affinity of binding. Second, the miRNA-induced translational repression applied to any target depends on the cellular context. Using simulations, we inferred systemic consequences of these properties on the microtargetome activity.

Short peptides that stall the ribosome are underrepresented in the proteome

Renana Sabi and Tamir Tuller

Tel-Aviv University

renanasabi@gmail.com

The ribosomal exit tunnel is the site in the ribosome through which nascent proteins leave the ribosome during gene translation. This tunnel may undergo strong biochemical interactions with some specific short amino acids sequences. The presence of such sequences in a protein may stall or even arrest the ribosome during gene translation; thus, these sequences are expected to undergo a process of elimination from the proteome during evolution to improve translation efficiency. Based on the output of a new experimental approach of ribosome profiling data, we find that indeed regions upstream of ribosomal pauses positions are enriched with short peptides that are underrepresented in the proteome. The computational approach described here can be used for detecting short peptides that can cause ribosomal pauses for various biotechnological and biomedical applications.

Pathways as robust biomarkers for cancer classification: the power of big expression data

David Amar, Tom Hait and Ron Shamir
Tel-Aviv University
ddam.am@gmail.com

Background: Gene expression signatures, serving as biomarkers, have been used successfully for prognosis, diagnosis and patient stratification in cancer. However, such signatures are often not sufficiently robust and sometimes perform poorly on new datasets. Moreover, standard case- control studies may yield a signature that is not specific to the tested disease. In addition, the biomedical interpretation of gene set signatures is often difficult. Methods: We collected data from 174 gene expression studies in GEO, covering 13,314 samples from 17 different array technologies covering 48 diseases. Each sample was manually annotated with Disease Ontology terms. We analyzed each sample separately by calculating two pathway features: the mean and standard deviation of the pathway genes' weighted rank. We used 1700 pathways from the Reactome, NCI, KEGG and Biocarta databases. We call the new database Peptalk (Pathway Expression, Phenotype and Tissue - A Learning Knowledgebase). Results: Using Peptalk, we constructed a classifier for each disease by comparing the pathway features of the disease samples to those of all other samples. We used 48 diseases that have at least 5 datasets and 100 samples each. Our analysis produced high performance classifiers for 16 disease terms, which include cancer, gastric cancers, breast cancer, and immune system cancers. For these diseases, classification was accurate even when validated on new datasets of different technologies. Our cancer classifier is based on a signature that of cell cycle and DNA replication pathways. Notably, the classifier for gastrointestinal cancer is based on several pathways not associated previously with the disease, including down-regulation of aquaporin-mediated transport. We summarize our results in a network in which nodes diseases and pathways. Edges connect a pathway to a disease if the pathway is markedly differential in the disease patients compared to both healthy controls from the same study and to samples of other diseases. Hence edges indicate a pathway-disease coupling that is specific to the tested disease. The network provides an overview of associations between pathways and diseases, and is valuable for further biological interpretation. Conclusions: Our analysis shows how large public databases can be used to infer disease specific diagnostic tools and can reveal pathways that have not been recognized before as pertinent to a specific type of cancer.

Nucleotide Composition in Retained Introns Improves Translation Costs in Fungi

Zafrir Zohar, Hadas Zur and Tamir Tuller

Tel-Aviv University

zoharzaf@post.tau.ac.il

RNA splicing is the central process of intron removal in eukaryotes, known to regulate, among others, growth, development, and response to external signals. Intron retention, in which an intron remains in the mature mRNA transcript, is widespread in various organisms. Accordingly, the roles of various intronic sequence features affecting intronic retention and translation regulation have yet to be thoroughly studied. Focusing on four fungi as model organisms, we performed a comprehensive large scale systems biology study to characterize for the first time how the interactions between gene splicing and translation, are encoded in transcripts. Analysis of the intronome of these fungi demonstrates that the beginning of introns is selected for various signals related to gene translation. Ribosomal profiling data analysis in S. cerevisiae supports the conjecture that in this organism intron retention frequently occurs; thus, introns are partially translated by the ribosome and their translation efficiency affects organismal fitness. These new discoveries are contributory steps towards a broader understanding of splicing regulation, intron evolution, and the crosstalk between splicing and translation; in addition, they can be used for engineering gene expression in various biotechnological and synthetic biology applications.

Network orientation via shortest paths

Dana Silverbush and Roded Sharan
Tel-Aviv University
dana.silverbush@gmail.com

The graph orientation problem calls for orienting the edges of a graph so as to maximize the number of pre-specified source-target vertex pairs that admit a directed path from the source to the target. Most algorithmic approaches to this problem share a common preprocessing step, in which the input graph is reduced to a tree by repeatedly contracting its cycles. While this reduction is valid from an algorithmic perspective, the assignment of directions to the edges of the contracted cycles becomes arbitrary, and the connecting source-target paths may be arbitrarily long. In the context of biological networks, the connection of vertex pairs via shortest paths is highly motivated, leading to the following problem variant: Given a graph and a collection of source-target vertex pairs, assign directions to the edges so as to maximize the number of pairs that are connected by a shortest (in the original graph) directed path. This problem is NPcomplete and hard to approximate to within sub-polynomial factors. Here we provide a first polynomial-size integer linear program formulation for this problem which allows its exact solution in seconds on current networks. We apply our algorithm to orient protein-protein interaction networks in yeast and compare it to two state-of-the-art algorithms. We find that our algorithm outperforms previous approaches and can orient considerable parts of the network thus revealing its structure and function.

MalaCards: an integrated compendium for diseases and their annotation

Noa Rappaport, Michal Twik, Noam Nativ, Gil Stelzer, Frida Belinky, Tsippi Iny-Stein, Iris Bahir, Marilyn Safran and Doron Lancet

Weizmann Institute of Science

noa.rappaport@weizmann.ac.il

Comprehensive disease classification, integration and annotation are crucial for biomedical discovery. At present, disease compilation is incomplete, heterogeneous and often lacking systematic inquiry mechanisms. We introduce MalaCards, an integrated database of human maladies and their annotations, leveraging the architecture and strategy of the GeneCards database of human genes. MalaCards mines and merges 63 data sources to generate a computerized card for each of 19,486 human diseases. Each MalaCard contains disease-specific prioritized annotations, as well as inter-disease connections, empowered by the GeneCards relational database, its searches, and GeneDecks set- analyses. First, we generate a disease list from 15 ranked sources, using disease-name unification heuristics. Next, we employ four schemes to populate MalaCards sections: 1) Directly interrogating disease resources, to establish integrated disease names, synonyms, summaries, drugs/therapeutics, clinical features, genetic tests, related publications, variations and anatomical context; 2) Searching GeneCards for associated genes with corresponding relevance scores; 3) Analyzing disease-associated gene-sets in GeneDecks to yield affiliated pathways, phenotypes, compounds, and GO terms, sorted by a composite relevance score and presented with GeneCards links; 4) Searching within MalaCards itself, e.g. for additional related diseases and anatomical context. The latter forms the basis for the construction of a disease network, based on shared MalaCards annotations, embodying associations based on etiology, clinical features and clinical conditions. This broadly disposed network has a power-law degree distribution, implying inherent properties of such networks. Diseases are also classified and grouped by various methods, striving to make MalaCards an effective tool for biomedical research.

Interactions between distant RNAs in regulatory networks

Mor Nitzan, Avital Steiman-Shimony, Yael Altuvia, Ofer Biham and Hanah Margalit

The Hebrew University of Jerusalem

mor.nitzan@mail.huji.ac.il

Competing endogenous RNAs (ceRNAs) were recently introduced as RNA transcripts that affect each other's expression level through competition for their microRNA (miRNA) co-regulators. This stems from the bi-directional effects between miRNAs and their target RNAs, where a change in the expression level of one target affects the level of the miRNA regulator, which, in turn, affects the level of other targets. By the same logic, miRNAs that share targets compete over binding to their common targets, and therefore also exhibit ceRNA- like behavior. Taken together, perturbation effects could propagate in the post-transcriptional regulatory network through a path of co-regulated targets and miRNAs that share targets, suggesting the existence of distant ceRNAs. Here we study the prevalence of distant ceRNAs and their effect in cellular networks. Analyzing the network of miRNA-target interactions deciphered experimentally in HEK293 cells, we show that it is a dense, intertwined network, suggesting that many nodes can act as distant ceRNAs of one another. Indeed, using gene expression data from a perturbation experiment we demonstrate small, yet statistically significant, changes in gene expression caused by distant ceRNAs in that network. We further characterize the magnitude of the propagated perturbation effect and the parameters affecting it by mathematical modeling and simulations. Our results show that the magnitude of the effect depends on the generation and degradation rates of involved miRNAs and targets, their interaction rates, as well as the distance between the ceRNAs and the topology of the network. While demonstrated for a miRNA- mRNA regulatory network, our results offer a new view on various post-transcriptional cellular networks, expanding the concept of ceRNAs, and implying possible distant cross-talk within the network with consequences for the interpretation of indirect effects of gene perturbation.

Genome-wide detection of regulated transcriptional termination in bacteria

Daniel Dar, Iris Karunker and Rotem Sorek

Weizmann Institute of Science

daniel.dar@weizmann.ac.il

Regulated transcriptional termination is a key gene regulatory process in bacteria. Implemented through cis regulatory elements, it controls the expression of genes involved in essential metabolic processes, virulence, antibiotic resistance, and phage host interactions. To uncover novel regulators in a genome-wide manner we developed a method for direct sequencing of RNA 3' ends in bacteria. We demonstrate that our method can identify the vast majority of termination events in the cell in a genome wide manner and to the single nucleotide resolution. In Bacillus subtilis our approach detects over 90% of the known regulators with high specificity and identifies over a dozen novel regulators implicated in swarming behavior, antibiotic resistance, rRNA metabolism and key metabolic reactions. To characterize novel and known regulators, we devised an RNA-Seq platform that allows unbiased, parallel in-vivo measurements of multiple regulator activities in different physiological conditions and enables high-throughput ligand screening for riboswitch candidates. Our results provide the first experimental genome-wide method for analyzing termination events in bacteria and can be readily applied in additional organisms.

Expression signatures of cancerous vs. non-cancerous cellular proliferation and their clinical relevance

Yedael Waldman, Tamar Geiger and Eytan Ruppin
Tel-Aviv University
yedaelwa@post.tau.ac.il

Understanding cell proliferation mechanisms has been a long-lasting goal of the scientific community and specifically of cancer researchers. Previous genome-scale studies of cancer proliferation determinants have mainly relied on knockdown screens aimed to gauge their effects on cancer growth. This powerful approach has several limitations such as off-target effects, partial knockdown, and masking effects due to functional backups. Here we employ a complementary approach and assign each gene a cancer Proliferation Index (cPI) that quantifies the association between its expression levels and growth rate measurements across 60 cancer cell lines. Reassuringly, genes found essential in cancer gene knockdown screens exhibit significant positive cPI values, while tumor suppressors exhibit significant negative cPI values. Cell cycle, DNA replication, splicing and protein production related processes are positively associated with cancer proliferation, while cellular migration is negatively associated with it - in accordance with the well known \go or grow\ dichotomy. A parallel analysis of non-cancerous proliferation indices (nPI) across 224 lymphoblastoid cell lines reveals surprisingly marked differences between cancerous and non-cancerous proliferation. These differences highlight genes in the translation and spliceosome machineries as selective cancer proliferation-associated proteins. A cross species comparison reveals that cancer proliferation resembles that of microorganisms while non-cancerous proliferation does not. Furthermore combining cancerous and non-cancerous proliferation signatures leads to enhanced prediction of patient outcome and gene essentiality in cancer. Overall these results point to an inherent difference between cancerous and non- cancerous proliferation determinants whose understanding may contribute to the future development of novel cancer- specific anti-proliferative drugs.

Exploring Coding Sequence Determinants of Expression in S. cerevisiae via Combined Synthetic-Computational Biology Approach

Shimshi Atar, Tuval Ben-Yehezkel, Ehud Shapiro and Tamir Tuller
Tel-Aviv University
shimshi@post.tau.ac.il

A major challenge in functional genomics is to understand the way different parts of the transcript affect aspects of its expression. Specifically, it is impossible to deduce causal relation between sequence features and transcript expression by analyzing endogenous genes. Here, to study the distinct and causal effect of different parts of the transcript, we generated a novel large YFP based library (463 variants); it was composed of randomizations of a different part of the viral HRSVgp04 gene's transcript while maintaining the protein it encodes, and fusing it to a reporter protein. Specifically, we randomized the end of the 5'UTR and the first 80 codons of the ORFs. These libraries enabled understanding the effect of different parts of the transcript on translation efficiency and protein levels, while controlling for possible effects of other transcript regions. Among others, we show that a few mutations in the non-coding region can have an effect of more than one order of magnitude on the measured protein levels. The reported results improve the understanding of gene translation regulation and transcript evolution; in addition they promote a better engineering of synthetic gene expression systems.

Comperhensive study of gene expression regulating information encoded in viral genomes

Eli Goz and Tamir Tuller

Department of Biomedical Engineering, Tel-Aviv University

guzeliez@gmail.com

Gene expression is a central biological mechanism in all domains of living organisms. It is a complex process of deciphering the information encoded in the mRNA molecule by the ribosome to produce a protein, regulated via complex interactions of various cis and trans acting factors, many of which are still unknown and/or poorly understood. The aim of this work is to use computational methods for a comprehensive analysis and modeling of gene expression in viruses - intracellular parasites consisting of encapsulated genome, which rely entirely on the translational machinery of the host cell for the synthesis of their proteins. Using HIV and Dengue as model organisms we show that these viruses undergo extensive evolutionary selection and some of the selection sites seem to be related directly to biophysical variables attributed to gene expression. We believe that our analysis will enable a better understanding of viral evolution and gene expression and intend to use our findings to computationally design general efficient methodologies for anti-viral vaccines engineering.

BindDB: an in-silico reverse-ChIP platform for epigenomic analysis in embryonic stem cells

Ilana Livyatan, Yair Aaronson, David Gokhman, Ran Ashken and Eran Meshorer
The Hebrew University of Jerusalem
iforst@yahoo.com

The exponential increase in epigenomic data from next generation sequencing techniques has greatly advanced the study of epigenetics in a variety of biological processes. Specifically, in embryonic stem cells (ESCs), chromatin accessibility regulated by histone modifications, and the resulting transcription factor binding, play a key role in regulating the properties of stem cells and the process of differentiation. Yet, the interplay among the modifications and factors and their influence on biological function is not always clear and requires a more general, integrative view of all of the data. In addition, the field has commonly viewed epigenetics from the perspective of the factor, trying to deduce its functional role from its binding profile. Here we developed a \reverse\ methodology whereby research stems from the perspective of one or more genomic regions or genes and we attempt to identify which factors bind these regions and/or what their chromatin status is.In line with this approach we compiled a database of hundreds of epigenetic datasets from human and mouse ESCs including epigenetic profiles for transcription factors chromatin modifiers and remodelers insulators structural proteins such as cohesin subunits as well as a wide battery of histone modifications. We developed BindDB (www.meshorerlab.ac.il/bindDB) a webtool to easily query the database to obtain a comprehensive and integrative epigenetic profile of all factors/modifications that show evidence of binding within the region(s) queried. For large queries the tool also runs downstream enrichment level analysis of each factor in the database to determine the specificity of factor binding to the regions of interest and clustering analysis of binding profiles which provides insights into the organization of binding factors into complexes and how they crosstalk with histone modifications. With the aid of the BindDB tool we were able to revisit and expand our analysis of histone gene regulation leading toa better understanding of how these clustered genes are epigenetically regulated. We could also better pinpoint and define the factors and modifications regulating bivalent genes reaching far beyond the standard \K4-K27\ definitions. Finally we found sound epigenetic signatures for the large group of the recently discovered lincRNAs confirming tight epigenetic regulation indicative of important functional roles for this group of non-coding genes. We expect the combination of the integrated data and the tool to provide a \breakthrough\ platform for the epigenetic aspects of stem cell research.

A computational study about the effect of oscillations in tRNA pool and translation elongation/initiation factors on translation rate

Lior Trieman and Tamir Tuller
Tel-Aviv University
lior.trieman@gmail.com

Gene translation is the process by which messenger RNA (mRNA) is decoded by a ribosome to produce a specific amino acid chain protein. This fundamental process is related to all biomedical disciplines. Specifically, it may be engineered for various biotechnological applications. Oscillations and clocks are important components of many hardware systems, suggesting that they are required also in synthetic biology. In the current study, we aimed at understanding the effect of oscillations in various factors related to the translation process on translation rate. To this end we analyzed computational models that consider the interactions between ribosomes and their size, concentration of tRNA molecules, the efficiency of their interaction with codons, and the stochastic nature of translation. Among others, we show that oscillating the different translational factors such that they all have a common period (T) results in a convergence to a periodic translation rate with period T. These findings may be utilized in future Synthetic biology.

A Code for Transcription Elongation Speed

Eyal Cohen and Tamir Tuller
Tel-Aviv University
coheneya@post.tau.ac.il

The two major steps of gene expression are transcription and translation. While hundreds of studies about the effect of sequence features on the translation elongation have been published, there are very few studies that connect sequence features to transcription elongation rate. Here we suggest that short sub-sequences of the transcript have a typical effect on relative RNA polymerase density; the density score of 5-mers learned from different parts of the gene and among different groups of genes are highly correlated. Thus, the density and elongation rate of the RNA polymerase are partially encoded in the underlying DNA sequence. In addition, we show that mRNA levels are partially related to the elongation rate of the RNA polymerase. Finally, we find a high correlation between the relative density of ribosomes and RNA polymerase on short sequences, suggesting that short DNA/RNA sequences have a similar effect on elongation speed of RNA-Polymerase/Ribosome respectively.

A mathematical model of mouse NK cell maturation population dynamics shows that CD27-Mac-1-NK cells are not the main source of later NK cell subsets

Michal Sternberg-Simon[*,1] *, Nissim Pinhas*[*,1]*, Laura Chiossone*[†]*, Gitit Shahaf*[*]*, Thierry Walzer*[‡]*, Eric Vivier*[§,¶,‖,#]*, Ramit Mehr**

*The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. [†] Department of Experimental Medicine (DIMES), University of Genoa, Italy. [‡]Centre International de Recherche en Infectiologie (CIRI), INSERM – CNRS, Lyon, France. [§]Centre d'Immunologie de Marseille-Luminy (CIML), Université de la Méditerranée, Marseille, France. [¶] Inserm, U1104, Marseille, France. [‖] CNRS, UMR7280, Marseille, France. [#] Service d'Immunologie, Assistance Publique – Hôpitaux de Marseille, Hôpital de la Conception, Marseille, France

[1]These authors contributed equally.

nissimp12@gmail.com

NK cells can be divided by the expression of two cell-surface markers, CD27 and Mac-1 (a.k.a. CD11b), into four separate subsets, suggesting a linear development model: CD27-Mac-1 -> CD27+Mac-1 -> CD27+Mac-1+ -> CD27-Mac-1+. We used a combination of BrdU labeling experiments and mathematical modeling in order to quantitatively test this model. In addition, we examined the possibility that NK cells may skip maturation steps, and transfer directly from CD27-Mac-1 to CD27-Mac-1+, since this transition requires changing the expression of only one marker (Mac-1). Fitting modeling results to the experimental data shows that the majority of NK cells already express CD27 upon entering the proposed pathway, questioning the role of the CD27-Mac-1 subset as a first developmental stage. Additionally, only a small fraction of NK cells exit the BM to other sites, suggesting that peripheral NK cell populations originate from site-specific immature NK cells more than from BM-derived mature NK cells.

Towards the Mapping of the "Prenylosome" A Structure based Prediction of Farnesylation and Geranylgeranylation Targets

Michal Sperber, Nir London, James Hougland, Corissa Lamphear, Elaina Zeverina, Elia Wright, Carol Fierke and Ora Schueler-Furman

Department of Microbiology and Molecular Genetics, Faculty of Medicine, Hadassah Medical School, The Hebrew University of Jerusalem

sperber1984@gmail.com

Prenylation- the attachment of a Farnesyl / Geranylgeranyl group to the carboxy terminus (C') of a protein is an important post-translational modification that targets the modified protein to the membrane. Protein farnesyltransferase (FTase) and Geranylgeranyltransferase (GGTase) are responsible for catalyzing this modification. Until recently it was believed that a C' CaaX box motif (Cys-aliphatic-aliphatic-any amino acid) is required for farnesylation and geranylgeranylation, but recent experiments have revealed a larger diversity of possible substrates. Here we propose a general structural modeling scheme to account for peptide specificity, and present its performance in recovering the experimentally derived selectivity profile of FTase. We obtain good discrimination between substrates and non-substrates for farnesylation in a recently published data set (Area under ROC curve; AUC = 0.88/0.91 on training/test sets accordingly), and identify 85% of reported proteins known to be farnesylated. Experimental validation of newly predicted peptides show that most (26/29) are indeed farnesylated in vitro, revealing a novel class of targets (CxxD/E). Encouraged by the good distinction of FTase substrates, we presumed that the same protocol could also provide selectivity for GGTase substrates. FTase and GGTase show significant similarity both in structure and sequence (same alpha subunit; very similar beta subunit). Our results here for GGTases show an AUC value of 0.83/0.86 on our training set (using minimization only/FlexPepDock accordingly). Using this protocol, we search for new and specific GGTase targets. This approach is easily adapted to additional systems: we have applied it with minor modifications to bcl and HDAC8 substrate prediction (see poster by Nawsad Alam). These protocol suggests a convenient avenue to advance the elucidation of the cellular network of peptide-mediated interactions.

The role of RNA conformations in RNA-protein recognition

Efrat Kligun and Yael Mandel-Gutfreund
Technion - Israel Institute of Technology
efratk@techunix.technion.ac.il

RNA-proteins interactions have a key role in many biological processes. Those interactions are mediated through a variety of protein domains. The most common RNA binding domain is the RNA recognition motif (RRM). Here we studied protein-RNA structures from 9 RNA binding protein families extracted from the protein data bank (PDB). Characterizing the structural properties of RNA within the interfaces revealed that in some family, specifically in the RRM family, there is enrichment of unique RNA conformations. Looking on RNA-proteins interactions showed that the RRM domains are also enriched by interactions with nucleotides having unique RNA conformations. Further, examination of the sequences binding the RRM domain showed mainly a preference of G nucleotides in syn conformation to precede U nucleotides (not in syn conformation) and the opposite. Those findings may imply a general code of RNA recognition by the RRM domains which are able to recognize a wide variety of different RNA sequences and shapes and until now their recognition code was still unknown. Overall, this study suggests an additional recognition code of RNA- proteins, conformation readout, used only in specific protein families.

Structure-based prediction of novel Histone Deacetylase 8 (HDAC8) non-Histone substrates: Insights into the functional context of HDAC and its substrates

Nawsad Alam, Lior Zimmerman, Noah Wolfson, Carol Fierke and Ora Schueler-Furman
The Hebrew University of Jerusalem
nawsad.alam2010@gmail.com

HDAC8, a member of the Histone deacetylases (HDAC) family, catalyzes the deacetylation of histones in the cell nucleus. Current evidence of interactions of HDAC8 with additional, non histone substrates such as p53 and SMC3 indicates a more complex functionality of this enzyme. This motivated us to develop a comprehensive, structure-based approach to detect more yet unknown HDAC8 substrates. We have previously developed the Rosetta FlexPepBind protocol that evaluates peptide binding ability to a receptor based on structural models of this interaction. Here we adapt and apply this protocol to the characterization of HDAC8 substrates using peptide fragments with known acetylated sites. Application of this protocol will hopefully identify new HDAC8 substrates, which can be used to investigate the interactome of this enzyme and its substrates.

Protein Complexes Assembly and Function Revealed by Structural Motion Prediction

Adva Yeheskel[1], Rony Seger[2] and Malka Cohen-Armon[3]

[1] The Bioinformatics Unit, George S. Wise Faculty of Life Sciences, Tel-Aviv University,

[2] Department of Physiology and Pharmacology and the Neufeld Cardiac Research Institute, Sackler Faculty of Medicine, Tel-Aviv University

[3] Department of Biological Regulation, The Weizmann Institute of Science

suezadva@tauex.tau.ac.il

PARP1 (polyADP-ribose polymerase-1) is a DNA binding protein which transfers ADP-ribose from nicotineamide dinucleotide to glutamic and aspartic residues. It contains 7 domains, 4 of them bind DNA and one is a catalytic domain. We explore the relationship between function and dynamics for PARP1 using structural motion prediction (normal mode analysis). A network of inter-domain contacts links nicked DNA binding domains to the catalytic domain in PARP1. Recently, the PARP1 was shown to be activated after binding to phosphorylated ERK2 in the absence of nicked DNA (Cohen-Armon et al.Molecular Cell 2007). The ability of PARP1 to bind nicked DNA is associated with DNA repair, whereas binding phosphorylated ERK2 mediates a variety of physiological functions associated with regulation of gene expression. This includes epigenetic mechanisms and mechanisms promoting proliferations and differentiation. In this work we predict the binding site of PARP1 to phosphorylated ERK2 and demonstrate the allosteric effects implicated in the catalytic activity of PARP1 using normal mode analysis. We show conformational changes in PARP1, rendering the protein accessible to target proteins binding, which may allow catalytic activity. These conformational changes are maintained either with nicked DNA or with ERK2 bound to PARP1. We suggest that the observed motion underlies PARP1 function, a hypothesis that should be further validated.

Predicting Protein-DNA Interactions Using Geometric Docking Algorithms – Can DNA specificity Improve Results?

Shula Shazman and Barry Honig
Columbia University NY
ss4179@c2b2.columbia.edu

Protein-DNA interactions are involved in many essential biological activities. Since there is no simple mapping code between DNA base pairs and protein amino acids, the prediction of Protein-DNA interactions is a challenging problem. Geometric Docking Algorithms are based on the assumption that there is a shape complementarity between the participating molecules. Existing docking algorithms between protein and DNA consider slightly the specificity between the protein sequence and the sequence of the DNA. Based on a systematic consideration of DNA specificity an improvement of the docking algorithms results can be observed. In order to test whether a systematic consideration of DNA specificity might improve the docking algorithms results by making them more accurate, sequence- dependent DNA shape models were used for docking instead of standard B-DNA. A non-redundant dataset of 169 protein-DNA complexes was selected from PDB. Three DNA shape models were docked to each of the 169 proteins in the dataset using the PatchDock algorithm1, 2. All three DNA structural models had the same DNA as the DNA structure in the PDB complex. The first DNA model was the DNA structure taken from PDB, the second DNA shape model was built using a coarse-grained model (to study sequence-dependent DNA shape) based on Molecular Dynamic simulations which was developed in Honig lab (paper in preparation) and the third DNA shape model was built as a standard B-DNA using 3DNA software. In order to find the best solution in each of the different docking cases CAPRI evaluation was applied. In 150 out of the 169 (89%) PDB complexes in dataset, docking a sequence dependent DNA model produced better results than the regular B-DNA. A statistical pair potential for Protein-DNA interactions (1) was used to further test the contribution of protein-DNA specificity in PatchDock (2,3). Incorporating this statistical pair potential produced an improved ranking of Patchdock solutions for 126 out of 169 (75%) complexes in the dataset. The results of docking a sequence-dependent DNA shape and using a statistical pair potential to re-rank PatchDock solutions both make it clear that the specificity of protein-DNA interactions could contribute to finding the protein-DNA binding site. The insights from such a study could help to characterize the unique properties of protein-DNA interfaces and identify new drug target sites. REFERENCES 1. Mu Gao and Jeffery Skolnick. Nucleic Acids Research 36, 3978-92, (2008). 2. Duhovny D., Nussinov R. and Wolfson HJ. Algorithms in Bioinformatics Lecture Notes in Computer Science Volume 2452, pp 185-200 (2002). 3. Duhovny D., Nussinov R. and Wolfson HJ. Nucleic Acids Res 33, W363-7 (2005).

NMR structure of the soluble amyloid beta (17-34) peptide

Gennadiy Fonar and Avraham O. Samson
Bar-Ilan University
fonar_g@yahoo.com

Alzheimers disease is the most common neurodegenerative disorder in the world. Its most significant symptoms are memory loss and decrease of cognition. Alzheimers disease is characterized by aggregation of two proteins in the brain namely amyloid β (Aβ) and tau. Recent evidence suggests that the interaction of soluble Aβ with nicotinic acetylcholine receptors (AChR) contributes to disease progression. In this thesis we determine the nuclear magnetic resonance (NMR) structure of an 17-34 peptide solubilized by the addition of two glutamic acids at each terminus. This study shows that the Aβ peptide adopts an -helical structure for residues 19-26 and 28-33. The -helical structure is broken around residues S26, N27, and K28, which form a kink in the helical conformation. This -helix was not described earlier in an aqueous solution without organic solvents, and at pH 7. These data are in agreement with Aβ adopting an -helical conformation in the membrane before polymerizing into Aβ-sheets and provide insight into the intermediate state of Aβ in Alzheimer's disease.

Large Scale Computational Scanning for Peptide-Mediated Protein-Protein Interactions

Yuval Sedan, Orly Marcu and Ora Schueler-Furman
The Hebrew University of Jerusalem
orly.marcu@mail.huji.ac.il

PeptiDerive is a protocol that systematically evaluates the free energy of each possible peptide-protein interaction derived from known protein- protein interfaces, in order to identify energetically significant linear segments in the interactions. In this work we have run this protocol on the PROTCID database, containing all dimeric interfaces observed in multiple crystal forms. Our goal is to characterize the derived peptides and to identify at a large scale interactions that can be inhibited using small molecule ligands or designed peptides.

How do different peptide-binding domains within a protein influence each other? A case study of interaction between SH3 and SH2 domains

Adily Basal and Ora Schueler-Furman
The Hebrew University of Jerusalem
adily777@gmail.com

How do peptide-binding domains communicate with each other to obtain regulatory responses? Many proteins contain several peptide binding domains that are connected by flexible linkers. These domains often interact mutually and regulate the response of the protein. According to PFAM-based annotation, there are above 180 annotations of human proteins that include a tandem arrangement of SH3 and SH2 domains. Most of these are Src-kinases, Tyrosine-kinases and adaptors. Recently, the structure of one of these adaptors, the Src-like adaptor protein 2 (slap2), has revealed a unique interaction between its SH3 and SH2 domains: they form one continuous beta-sheet that creates one globular domain. This special close association led us to look for the factors that determine the orientation of the two domains relative to each other, in slap2 as well as other SH3-SH2 domain containing proteins, and to investigate possible consequences on peptide-binding specificities of these domains. Using sequence analysis and structural modeling, we developed a predictor for genome-wide identification of SH3-SH2 domain combinations that can create one globular domain rather than the more generally observed loose interactions between SH3 and SH2 domains, and the characterization of peptide binding specificity of the combined domain. This tool will allow improved understanding of a particular example of how two domains interact to form a more complex regulation than each domain on its own.

Functional annotation of nucleic acid binding proteins using 3D surface historgrams

Noam Kremen and Yael Mandel-Gutfreund
Technion - Israel Institute of Technology
noam.kremen@gmail.com

The availability of 3D protein structures has increased dramatically over the past years. At the start of the millennium, The Protein Databank (PDB) held approximately 14,000 three dimensional (3D) structures of biological macromolecules. As of the time of writing, this database holds 95,000 structures, 25,000 of those added during the last four years. As the total number of structures increases, so does the number of those structures that do not easily lend themselves to analysis by established methods based on sequence or structural comparison approaches. This thesis presents a new approach to analyzing a 3D protein structure by looking at the local geometry of the protein molecular surface. Using a discrete three-dimensional point-mesh to approximate the vdW surface, it is possible to define a neighborhood of arbitrary size around each point on the mesh and calculate its local curvature. When applied to a dataset consisting of RNA and DNA binding proteins, this method reveals that the distribution of curvature values on the surface of a protein is not dictated by the protein fold and is not directly correlated to fold family membership. However, we found that proteins with similar surface-curvature distributions tend to exhibit common functional characteristics, despite the fact that they have evolved independently. Overall, our method provides an additional insight regarding the tight relationship between the protein structure and its function.

Elastic network modes refinement of the electron microscopy structure of the translocator protein

Amir Shechvitz and Avraham O. Samson
Faculty of Medicine in the Galilee, Bar-Ilan University, Zfat
amirshechvitz@hotmail.com

The translocator proteins (TSPOs) is a family of membrane receptors localized mainly on the mitochondria of higher species, and on the cellular membrane of bacteria. As implied by their name, the receptors are responsible for translocating hydrophobic molecules, such as cholesterol and porphyrin intermediates, across the membrane lipid bilayer. Attempts to crystallize the proteins have been unsuccessful, and even with the advent of the NMR structure of mouse TSPO, structure determination of the bacterial TSPO (TspO) remains incomplete. Recent attempts to solve one such structure using electron microscopy were also fruitless, and to the best of our knowledge, the translocator proteins structure is unsolved. In this study, we will determine the structure of bacterial TSPO (TspO) using elastic network mode refinement of a homology model constrained by an electron-microscopy density map. TspO is found to adopt a helical bundle fold composed of five -helices. This study sheds light on the TSPO structure involved in mitochondrial permeability and paves the ways for computer aided drug design of drugs involved in apoptosis, neurosteroidogenesis and many other processes.

Characterizing of functional human coding RNA editing from evolutionary, structural and dynamic perspectives

Oz Solomon, Lily Bazak, Erez Levanon, Ninette Amariglio, Ron Unger, Gideon Rechavi and Eran Eyal

Bar-Ilan University, Sheba Medical Center

oz.solomon@live.biu.ac.il

A-to-I RNA editing has been recently shown to be a widespread phenomenon with millions of sites spread in the human transcriptome. However, only few are known to be located in coding sequences and modify the amino acid sequence of the protein product. Here, we used high-throughput data, variant prediction tools and protein structural information in order to find structural and functional preferences for coding RNA editing. We show that RNA editing has a unique pattern of amino-acid changes characterized by enriched stop-to-tryptophan changes, positive-to-neutral and neutral-topositive charge changes. RNA editing tends to have stronger structural effect than equivalent A-to-G SNPs but weaker effect than random A-to-G mutagenesis events. Sites edited at low level tend to be located at conserved positions with stronger predicted deleterious effect on proteins comparing to sites edited at high frequencies. Lowly edited sites tend to destabilize the protein structure and affect amino acids with larger number of intra-molecular contacts. Still, some highly edited sites are predicted also to prominently affect structure and tend to be located at critical positions of the protein matrix and are likely to be functionally important. Using our pipeline, we identify and discuss several novel putative functional coding changing editing sites in the genes COPA (I164V), GIPC1 (T62A), ZN358 (K382R) and CCNI (R75G).

BEERT: A tool for the fast prediction of change in external ligand entropy

Ido Ben-Shalom[1], *Stefania Pfeiffer-Marek*[2] *and Holger Gohlke*[1]

[1] Heinrich-Heine-University of Düsseldorf

[2] *Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany.*

ido.ben.shalom@uni-duesseldorf.de

Entropy is an essential component of binding free energy (delta G): binding enthalpy (delta H) does not correlate with delta G due to entropy-enthalpy compensation. However, many methods to compute delta G neglect entropy, as it is difficult to estimate. Here we aim at estimating differences in the rotational and translational component of binding entropy delta S(R/T), which arise from the varying extent of restriction of the respective degrees of freedom compared the unbound state. Our Binding Entropy Estimation for Rotation and Translation (BEERT) method allows the fast prediction of delta S(R/T) in three steps: (i) Generation of a configurational ensemble by using AutoDock, representing the energy landscape; (ii) clustering of poses that represent alternative energy minima according to their intermolecular interaction pattern; (iii) estimation of the degree of rotational and translational restriction. In all, this approach mimics the reduction of accessible microstates for estimating delta S(R/T). For validation, we compared a linear combination of delta S(R/T) and MMPBSA effective energies to experimental affinities of inhibitors binding to HIV1 protease, FXa, and HSP90. The obtained correlations (R^2 = 0.55 - 0.60 ± 0.1 - 0.7, $p < 0.001$) are significant and improved against MMPBSA results alone (R^2 = 0.01 - 0.38). For comparison, an alternative method for entropy estimation widely used in scoring functions by counting the rotatable bonds of ligands resulted in significantly lower R^2 = 0.01 - 0.42. The robustness of our method was tested in a leave-one-out cross-validation, resulting in q^2 = 0.34 - 0.51 ± 0.3 - 0.6, $p < 0.05$ for all datasets. Spurious correlations were excluded by applying our approach to the datasets with scrambled experimental affinities. In summary, BEERT allows an efficient prediction of delta S(R/T) to enhance binding affinity predictions for lead optimization.

Approaching Reality: Guides toward full-resolution modeling of peptide-protein interactions

Emma Joy Dodson, Nawsad Alam, Dima Kozakov and Ora Schueler-Furman
The Hebrew University of Jerusalem
emma-joy.dodson@mail.huji.ac.il

FlexPepDock ab initio1 is a Rosetta protocol designed to predict the location and configuration of a peptide on the surface of its receptor protein, starting from an input structure positioned within relative proximity of the native complex. The approach is based on high resolution modeling of atomic details at the surface of the peptide-protein interface. Our protocol adjusts side chain orientations for both peptide and receptor, but rigid body orientation and backbone conformational changes are confined to the peptide. With FlexPepDock ab initio, the peptide undergoes alternating rounds of rigid body orientation and backbone conformation optimization, arriving at a final position characterized by the lowest binding energy in the energy landscape. Here we evaluate the success of both Brix2,3- and Peptimap4- based approaches in creating input structures best suited for coordination with FlexPepDock ab initio.

Widespread evidence to the role of flanking regions on transcription factor binding specificity

Iris Dror, Remo Rohs and Yael Mandel-Gutfreund
Technion - Israel Institute of Technology
dmotid@techunix.technion.ac.il

Identifying transcription factor (TF) binding sites (BSs) is a key step in understanding gene regulation. Genome wide technologies now enable the identification of sequence preferences (motifs) for hundreds of different TFs. However, recent in vivo studies suggest that while a short sequence motif can appear myriad of times in the eukaryotic genome, 99.8% of its appearances are unbound by TFs. This raises the question, what distinguishes bound from unbound motifs? While it is well established that TF binding depends on different mechanisms such as the DNA methylation level, chromatin structures and binding of cofactors, these have not been sufficient to explain the majority of TF binding preferences. Moreover, recent in-vitro binding assays, which do not consider the effect of cellular environment, demonstrate unexplained binding preference for many TFs. This suggests that the information determining TF specificity is also directly encoded within the DNA context surrounding a motif and can be intrinsically recognized by the protein. Here we aimed to investigate the direct contribution of the DNA sequences flanking the core motif on a TF's DNA binding specificity. To this end, we analyzed the DNA sequences flanking the BSs of 105 and 72 TFs, extracted from in-vitro binding assays and in-vivo ChIP-seq data, respectively. In addition we employed our high- throughput DNA shape predictor to analyze the shape preferences of these sequences. Selecting all bound sequences containing the known motifs and comparing their DNA sequence and structural properties to non-binding sequences revealed significant differences between functional and non- functional BSs at the regions flanking the core motifs. Notably, the BSs of TFs belonging to similar families exhibited common features, which expanded well beyond the core motif. Intriguingly, the characteristic features of the flanking regions extracted from in-vitro and in- vivo data for TFs belonging to the same families were similar, emphasizing that the features of the flanking regions represent the intrinsic preference of the proteins to bind within a specific environment, which is characterized by unique sequence and structural features. We propose that these unique features help to guide the TFs bind to their cognate binding sites.

Reconstructing Cancer Karyotypes

Rami Eitan and Ron Shamir
Tel-Aviv University
ramieita@post.tau.ac.il

Reconstructing Cancer Karyotypes Rami Eitan and Ron Shamir, Tel Aviv University. Cancer genomes change during the disease progression in a series of rearrangements. These rearrangements include, among others, segmental deletions, insertions, translocations and inversions. The result is a highly complex, patient- specific cancer karyotype. Using new technologies of deep sequencing and microarrays it is possible to interrogate a cancer genome and obtain lists of copy number variations and breakpoints (\jumps\) relative to the normal genome. This information is very detailed but local in nature and does not give the overall picture of the cancer genome. One of the basic challenges in cancer genome research is to use such information to infer the cancer karyotype. We suggest an algorithmic approach that receives segmental copy number and breakpoint data as input and produces a cancer karyotype that is concordant with them. Moreover by modifying and implementing an algorithm suggested by Greenman et al. (2010) we are able to derive a chronological series of rearrangements that produced the observed cancer genome.

RBPmap: a web server for mapping binding sites of RNA-binding proteins

Inbal Paz, Idit Kosti, Manuel Ares, Jr., Melissa Cline and Yael Mandel-Gutfreund

Technion - Israel Institute of Technology

inbalos@tx.technion.ac.il

Regulation of gene expression is executed in many cases by RNA-binding proteins (RBPs) that bind to mRNAs as well as to non-coding RNAs. RBPs recognize their RNA target via specific binding sites on the RNA. Predicting the binding sites of RBPs is known to be a major challenge. We present a new webserver, RBPmap, freely accessible through the website http://rbpmap.technion.ac.il/ for accurate prediction and mapping of RBP binding sites. RBPmap has been developed specifically for mapping RBPs in human, mouse and Drosophila melanogaster genomes, though it supports other organisms too. RBPmap enables the users to select motifs from a large database of experimentally defined motifs. In addition, users can provide any motif of interest, given as either a consensus or a PSSM. The algorithm for mapping the motifs is based on a Weighted-Rank (WR) approach, which considers the clustering propensity of the binding sites and the overall tendency of regulatory regions to be conserved. In addition, RBPmap incorporates a position-specific background model, designed uniquely for different genomic regions, such as splice sites, 5' and 3' UTRs, non-coding RNA and intergenic regions. RBPmap was tested on high-throughput RNA binding experiments and was proved to be highly accurate.

NeuroPID: Genomic Prediction of Neuropeptides

Dan Ofer and Michal Linial

The Hebrew University of Jerusalem

ddofer@gmail.com

Neuropeptides (NPs) are short secreted peptides produced in neurons. NPs act by activating signaling cascades governing broad functions such as metabolism, sensation and behavior throughout the animal kingdom. NPs are the products of multistep processing of longer proteins, the NP precursors (NPPs). We present NeuroPID (Neuropeptide Precursor Identifier), an online machine-learning tool that identifies metazoan NPPs. NeuroPID was trained on 1418 NPPs annotated as such by UniProtKB. A large number of sequence-based features were extracted for each sequence with the goal of capturing the biophysical and informational-statistical properties that distinguish NPPs from other proteins. Training several machine-learning models, including support vector machines and ensemble decision trees, led to high accuracy (89-94%) and precision (90-93%) in cross-validation tests. For inputs of thousands of unseen sequences, the tool provides a ranked list of high quality predictions based on the results of four machine-learning classifiers. The output reveals many uncharacterized NPPs and secreted cell modulators that are rich in potential cleavage sites. NeuroPID is a discovery and a prediction tool that can be used to identify NPPs from unannotated transcriptomes and mass spectrometry experiments. NeuroPID predicted sequences are attractive targets for investigating behavior, physiology and cell modulation. NeuroPID is available at http://www.neuropid.cs.huji.ac.il.

Lineage tree analysis of high throughput immunoglobulin sequencing clarifies the B cell maturation pathways

Helena Hazanov, Yu-Chang Wu**, Alex Ademokun**,*
*Deborah K. Dunn-Walters** and Ramit Mehr**
*The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. **Guy's Hospital, King's College London, UK.
Lena1907@gmail.com

Transitional (TR) B cells are immature B cells that have migrated from the bone marrow to peripheral lymphoid organs, but can still undergo selection against autoreactivity. TR cells that survive selection eventually develop into mature naïve B cells (CD27+ IgD, NA). Upon exposure to antigen, NA cells become IgM memory (CD27+IgD+, MM) or \classical\ class-switched memory cells (CD27+ IgD SM). MM immunoglobulin (Ig) genes are thought to undergo somatic hypermutation albeit with lower frequency than SM but not class-switch recombination. Hence it is assumed that MM B cells originate from T-independent immune responses while SM cells originate from T-dependent responses. Alternatively MM cells maybe early emigrants from T-dependent germinal centers. The consensus opinion today is that double negative B cells (CD27+ IgD DN) are exhausted memory cells. However it is unclear which of the memory populations MM SM or both gives rise to DN cells. In this study we used lineage tree analysis of Ig heavy chains gene sequences of five B cell subsets (TR NA MM SM and DN) from three individuals to study the relationships between the above B cell populations and garner insights regarding their role in immune responses. Our analyses confirmed that both MM and SM branches can include DN Ig sequences sometimes identical to SM Ig sequences. MM trees were significantly shorter than SM trees. Our finding of combined trees that included both MM and SM sequences suggests that at least some MM cells originate from the same clones as SM rather than develop separately.

Inferring binding site motifs from HT-SELEX data

Yaron Orenstein and Ron Shamir
Tel-Aviv University
yaronore@post.tau.ac.il

Transcription factors (TF) play a vital role in gene regulation. By binding the DNA at promoter regions they encourage or impede gene transcription. Thus, much research has been focused on learning TF binding preferences. Technological advancements in the last decade have made it possible to measure the binding of a single TF to thousands of DNA sequences. Protein binding microarrays (PBM) and high-throughput-SELEX (HT- SELEX) utilize the power of microarrays and high-throughput sequencing, respectively, to measure in vitro binding in a high-throughput manner. Recently, we conducted a large-scale comparison between models derived from these technologies. Through our analysis, we revealed several biases in HT-SELEX technology, such as sequence biases and over-enrichment. In this study, we suggest ways to overcome these biases and present a new algorithm to infer a binding site motif from HT-SELEX data. To evaluate the performance of our algorithm, we tested how well models inferred by our algorithm predict PBM and ChIP-seq binding. Our algorithm performs better in predicting both in vitro and in vivo binding compared to the extant algorithm due to J. Toivonen, and slightly worse compared to published models, which were generated with manual intervention. We hope that our algorithm and analysis will help to advance the research of the rich HT-SELEX data.

Identifying functional residues in proteins using a multilevel alphabet

Ronit Hod and Yael Mandel-Gutfreund
Technion - Israel Institute of Technology
hodronit@gmail.com

The protein sequence dictates its structure and the function. In the process of evolution proteins accumulate changes allowing the divergence of species and individuals. Within the protein sequence different positions evolve at varying rates. While some are highly conserved, others tend to diverge extensively. Thus the evolutionary rate of individual residues in proteins serves as the basis for prediction of protein functional sites, assuming that these tend to be highly conserved. While sequence conservation is the basis for many prediction methods, additional information on structural conservation has proven to contribute to function prediction. We present a novel Multi-Level Alphabet (MLA), which we employed to search for functional residues in proteins which are conserved both at the physicochemical level and the secondary structure. We demonstrated that MLA conserved residues are correlated with functional positions in proteins such as catalytic sites, hot spots, disease related mutations and cancer related mutations. While the correlation between functional sites and conserved residues at the MLA level was usually higher than the correlation with amino acid (AA) conserved residues, the union of conserved MLA residues and conserved AA residues significantly improved the detection of different types of functional sites in proteins. The low intersection between the conserved populations point out to the different evolutionary pressure on individual positions in proteins.

Gene and microRNA expression as tools to infer spectrums of Inflammatory Bowel Diseases

Metsada Pasmanik-Chor, S. Ben Shachar, H. Yanai, L. Baram, H. Elad, A. Ofer, E. Brazowski, N. Shomron, H. Tulchinsky and Iris Dotan

Tel-Aviv University

metsada@post.tau.ac.il

Pouchitis is inflammation of the previously normal small bowel reservoir (ileal pouch) which may develop in ulcerative colitis (UC) patients undergoing large bowel resection and pouch surgery. We aimed to characterize pouch disease behavior using a molecular approach. UC pouch patients were prospectively stratified according to disease behavior into normal pouch (NP), chronic pouchitis (CP), and Crohns-like disease of the pouch (CLDP) groups. These were compared to Crohns disease (CD). Gene expression analysis of intestinal mucosal biopsies was performed using Affymetrix microarrays, in sixty six subjects, validated by real-time PCR. MicroRNA expression was performed by Illumina miR-Seq. Gene ontology was studied using Bioinformatics tools. While in UC ileum there were no significant gene or microRNA expression alterations, NP patients had 168 differentially-expressed genes (fold change ≥|2|, corrected p value ≤0.05). In CP and CLDP 490 and 1152 gene expression alterations were detected, respectively. Gene expression and microRNA profiles reflected disease behavior. CD ileitis had 358 alterations, with a 96% overlap with the various pouch groups. Gene ontology analyses revealed multiple biological processes associated with pouch inflammation, including response to chemical stimulus, small molecule metabolic and immune system processes and specific infectious-related pathways such as staphylococcus aureus, leishmaniasis and tuberculosis. There were 190 genes with significant negative correlation with 63 microRNAs. Interestingly, only 6% of these genes were up regulated, while the majority were down- regulated. Gene and microRNA alterations in pouch inflammation and CD overlap, suggesting that IBD is a spectrum, rather than distinct diseases. Altogether, our work shows that, gene and microRNA expression patterns could be used to characterize IBD subgroups. Acknowlegment: The authors would like to thank Dr. Varda Oron-Karni for the most professional performance of microarrays. This study was partially supported by a generous grant from the Leona M. and Harry B. Helmsley Charitable Trust.

From gastritis to gastric lymphomas: Ig gene repertoire diversification and selection

*Miri Michaeli, HillaTabibian-Keissar[#], Ginette Schibi[#], Gitit Shahaf, Yishai Pickman, Helena Hazanov, Kinneret Rosenblatt[#], Deborah K. Dunn-Walters**, Iris Barshack[#] and Ramit Mehr*

*The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. **Guy's Hospital, King's College London, UK. [#]Dept. of Pathology, Sheba Medical Center, Tel-Hashomer, Israel.

miri.michaeli1@gmail.com

Chronic gastritis is characterized by gastric mucosal inflammation due to autoimmune responses or infection, frequently with Helicobacter pylori. Gastritis with H. pylori background can cause gastric mucosa-associated lymphoid tissue lymphoma (MALT-L), which sometimes further transforms into diffuse large B cell lymphoma (DLBCL). However, gastric DLBCL can also be initiated de novo. The mechanisms underlying transformation into DLBCL are not completely understood. We analyzed immunoglobulin repertoires and clonal trees to investigate whether and how immunoglobulin gene repertoires, clonal diversification and selection in gastritis, gastric MALT-L and DLBCL differ from each other and from normal responses. The two gastritis types (positive or negative for H. pylori) had similarly diverse repertoires. MALT-L dominant clones presented higher diversification and longer mutational histories compared with all other conditions. DLBCL dominant clones displayed lower clonal diversification, suggesting the transforming events are triggered by similar responses in different patients. These results are surprising, as we expected to find similarities between the dominant clones of gastritis and MALT-L and between those of MALT-L and DLBCL.

Filling the gaps in Non-coding DNA using comparative genomics-a review

S. Anandakumar and P. Shanmughavel

Bharathiar University

bioinfoanand@gmail.com

Abstract Genomes of all organisms are made of several highly conserved non-protein coding regions in addition to the coding segments. Until recent genome-wide analysis, the importance of these non-coding segments was not appreciated. Biochemical analyses by isolating these non-coding regions from cells, tissues or whole organism studies are powerful tools for their identification. In lieu of this, identifying and annotating these regions using comparative and functional genomics approaches is on a high radically. Understanding and identifying their location and what these segments constitute would pave way for functional annotation. Large scale functional genomics approaches help to identify novel genes and allow hypothesize its in vivo function systematically in turn aid in annotating the conserved regions obtained from comparative genomics at the sequence level. In this review, analysis of the non-coding regions in Plasmodium species was carried out to elucidate the concealed information contained in them. Considering the size of the gene numbers discovered in the genomes of the 3 Plasmodium species, it is reasonable to urge the PlasmoDB to re-annotate these genomes to incorporate the novel coding regions predicted in this study.

Exploiting Regulatory Hidden Information Interleaved in the Redundancy of the Genetic Code without Prior Knowledge

Hadas Zur and Tamir Tuller
Department of Biomedical Engineering, Tel-Aviv University
zurhadas@post.tau.ac.il

Dozens of studies in recent years have demonstrated that codon usage encodes various aspects related to all stages of gene expression regulation. Thus, information related to gene expression regulation is encoded in the ORF itself and affects its evolution. When relevant high quality large scale gene expression data is available it is possible to statistically infer and model these signals, enabling analysing and engineering gene expression. However, when these data are not available it is impossible to infer and validate such models. In the current study we suggest Chimera - an unsupervised computationally efficient approach for exploiting hidden high dimensional information related to the way gene expression is encoded in the ORF, based solely on the genome of the analysed organism. One version of the approach, named Chimera Average Repetitive Substring (Chimera ARS), estimates the adaptability of an ORF to the intracellular gene expression machinery (e.g. ribosomes, RNA polymerases, microRNAs, spliceosomes, etc) of a genome (host), by computing its tendency to include long sub-sequences that appear in its coding sequences; the second version, named ChimeraMap, engineers the codons of a protein such that it will include long sub-sequences of codons that appear in the host coding sequences, improving its adaptation to a new host's gene expression machinery. Both versions are based on the assumption that evolution shapes the ORFs of endogenous genes to improve their expression. We demonstrate the applicability of the new approach for analyzing and engineering heterologous genes and for analyzing endogenous genes. Specifically, focusing on E. coli, we show that it can exploit information that cannot be detected by conventional approaches (e.g. the CAI - Codon Adaptation Index), which only consider single codon distributions; for example, we report correlations of up to 0.67 for the ChimeraARS measure with heterologous gene expression, when the CAI yielded no correlation. Furthermore, our analysis of the ORFs of endogenous genes demonstrates that indeed high dimensional information is encoded in their codon usage bias and is correlated with their expression levels.

Editing Cancer

Nurit Paz Yaacov, Lily Bazak, Hagit Porath, Miri Danan, Ilana Sychikov, Binyamin Knisbacher, Eli Eisenberg and Erez Levanon

Bar-Ilan University

nuritpaz@gmail.com

RNA editing is a site-specific modification, altering the sequence of transcript from that encoded in the genome. The most dominant alteration is adenosine to Inosine conversion (A-to-I RNA editing), mediated by the ADAR family of enzymes. It is most frequent in primates, especially in Alu elements, estimated to contain over 100 million editing sites but contain also essential role in many recoding sites. The exact frequency and patterning of this modification was not yet estimated across cancer tissues. We systematically characterized A-to-I RNA editing modification along various cancer tissues using RNA-Seq and whole exome sequences from the TCGA collection in nine different tissues, that contained matched tumor and normal sequences from the same patient: Bladder Urothelial Carcinoma 13 patients, Breast invasive carcinoma 95 patients, Colon adenocarcinoma 18 patients, Head and Neck squamous cell carcinoma 29 patients, Kidney renal clear cell carcinoma 62 patients, Liver hepatocellular carcinoma 30 patients, Lung adenocarcinoma 36 patients, Prostate adenocarcinoma 31 patients and Thyroid carcinoma 42 patients. We show that A-to-I editing and ADARs are significantly altered in most cancer type tested. In most tumors editing level is elevated compared to their matched normal, especially in breast, head and neck, thyroid, bladder and lung. In the liver we see gender biased editing alteration, correlated to survival rate. Kidney demonstrates opposite trend and no major alteration identified in prostate. In coding sequence the picture is more complex revealing dominance of A-to-I mismatches and patient specific editing alterations in many nonsynonimus sites. Our results demonstrate that editing is frequent and distinctive in many tumor tissues. We suggest that editing can serve as additional layers of mutation, accompanying the traditional cancer genomic mutation. Similarly, we speculate that most epi-mutations will be \passengers\ and only few will serve as \drivers\ in each patient. Those can provide as a potential new layer of therapeutic targets. Thus we suggest that the classification (and treatment) of the patient mutation profile should be done on both DNA and RNA levels.

Construction of a Microsatellites-Based Linkage Map for the White Grouper (Epinephelus aeneus)

Lior Dor, Andrey Shirak, Sergei Gorshkov, Mark R. Band, Abraham Korol, Yefim Ronin, Arie Curzon, Gideon Hulata, Eyal Seroussi and Micha Ron

ARO

lior.dor@mail.huji.ac.il

The white grouper (Epinephelus aeneus) is a promising candidate for domestication and aquaculture due to its fast growth, excellent taste and high market price. A linkage map is an essential framework for mapping QTL for economic traits, and study of genome evolution. DNA of a single individual was deep-sequenced and microsatellite markers were identified in 177 of the largest scaffolds of the sequence assembly. Success rate of developing polymorphic homologous markers was 94.9 % as compared to 63.1 % of heterologous markers from other grouper species. Of the 12 adult mature fish present in the broodstock tank, two males and two females were identified as parents of the assigned offspring by parenthood analysis using 34 heterologous markers. A single full-sib family of 48 individuals was established for construction of first-generation linkage maps based on genotyping data of 222 microsatellites. The markers were assigned to 24 linkage groups in accordance to the 24 chromosomal pairs. The female and male maps consisting of 203 and 202 markers spanned 1,053 and 886 cM, with an average inter-marker distance of 5.8 and 5.0 cM, respectively. Mapping of markers to linkage groups ends was enriched by using markers originating from scaffolds harboring telomeric repeat-containing RNA. Comparative mapping showed high synteny relationships among the white grouper, kelp grouper (E. bruneus), orange-spotted grouper (E. coioides) and Nile tilapia (Oreochromis niloticus). Thus, it would be useful to integrate the markers that were developed for different groupers, depending on sharing of sequence data, into a comprehensive consensus map.

Conserved lincRNA expression patterns during vertebrate evolution

Hadas Hezroni and Igor Ulitsky
Weizmann Institute of Science
hadas.hezroni@weizmann.ac.il

Over the past decade, increasingly thorough examinations of the RNA species in mammalian cells have revealed the fascinating complexity of the transcriptome, in which genomic regions previously thought to be transcriptionally silent give rise to a range of processed and regulated transcripts that do not appear to code for functional proteins. These include long intervening noncoding RNAs (lincRNAs), which contain at least 200 bases, and similarly to mRNAs, begin with a 5 cap and end with a poly(A) tail. Thousands of lincRNAs have been described in human and mouse and an increasing number of them are now implicated as key regulators in a variety of processes, including establishment of cell identify, proliferation, apoptosis and response to stress. It is well appreciated that lincRNA genes evolve much faster than coding sequences or UTRs. In order to characterize lincRNA evolution we have collected RNA-seq and 3P-seq data from at least ten tissues or developmental stages in each of 12 vertebrate species and identified thousands of lincRNA genes in each species, which shared several conserved features such as a small number of exons, typical length of about 1kb, and low but tissue-specific expression levels. We find that lincRNA loci evolve rapidly - over 80% of the lincRNAs in each species are lineage-specific, less than 100 lincRNAs are conserved between mammals and fish, and in conserved lincRNAs, only a small portion of the sequence is under selective constraint. Nevertheless, hundreds of lincRNAs have been retained since the last common ancestor of all amniotes and thousands of lincRNAs are found in syntenic positions throughout vertebrates, forming a large cohort of candidates for functional roles. In order to evaluate the potential regulatory and functional conservation of sequence- or position-similar lincRNAs, we compared their gene expression patterns across species. We found a clear conservation of expression level of lincRNAs sharing sequence conservation, with correlation coefficients ranging from 0.4 to 0.6 in different species and tissues, slightly lower than those observed for conserved protein-coding genes. Furthermore, when we compared relative expression of sequence-conserved genes across different tissues, we found similar levels of correlation in lincRNAs and in coding genes. lincRNAs appearing in syntenic positions but sharing no detectable sequence conservation exhibit lower, but still significant expression pattern similarities, suggesting that many of such pairs are functionally orthologous.

ConCord - coordinates conversion system

Mark Sokol, Gil Stelzer and Mali Salmon-Divon
Ariel University
marksokol89@gmail.com

Due to advances in microarray, high throughput sequencing and high throughput proteomic technologies, biology has become an increasingly data- rich field. These technologies have opened new opportunities to study the entire genome sequence of organisms (genomics), the complete set of RNA transcripts produced by the genome (transcriptomics), and large scale investigation of the function and structure of proteins (proteomics). Analysis of the generated wealth of data involves integration of multiple layers of information. Such integration frequently requires conversion between different coordinate systems. Since a single gene is transcribed to several transcripts which in turn code for different proteins this transformation is not a trivial task, especially for the non-bioinformatician users. Here we introduce ConCord, a user friendly web platform which converts coordinates between genome transcriptome and proteome systems in all possible directions. This useful software, which is written in Java and relies on the Ensmebl database, has the flexibility to be used in a various genomic applications.

Computational prediction of splice-altering genetic variations

Noa Cohen, Amit Lavon and Liran Carmel
The Hebrew University of Jerusalem
noa.cohen@carmelab.com

Splicing is a crucial stage in gene expression of the majority of human genes. Mutations affecting splicing are known to cause a wide range of genetic diseases, and revealing them is essential for understanding better the underlying mechanism of different diseases and for suggesting targets for gene therapy. According to current estimates, about 30% of human genetic diseases are due to such alterations [1] and thus research of this field can have significant implications. We propose a computational method to estimate the effect of any mutation on the normal splicing pattern of nearby genes. Existing variety of computational tools attempt to solve this problem; however, they usually rely on the knowledge of specific splicing regulatory elements (SREs), a significant known drawback. Our technique uses k-mer distributions around splicing junctions and thus bypasses the pitfalls of the traditional methodologies. In order to estimate the tool's performance we use mutations whose effect on splicing is known. The mutations dataset includes mutations that do not affect splicing, as well as mutations that result is alternative splice site or exon skipping. Given a mutation, the tool outputs the predicted effect on splicing. So far we successfully predicted 63% and 43% of the mutations that are known to generate an alternative 3' splice site and an alternative 5' splice site accordingly. The tool can contribute to the research of Mendelian disorders with unknown genetic basis and to identify human disease susceptibility polymorphisms that affect splicing. Reference 1. Lim, K.H., et al., Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A, 2011. 108(27): p. 11093-8.

Computational Analysis of Pasilla Binding Motifs Location in Drosophila Genome Points to the Protein Role in Splicing Regulation

Maya Polishchuk, Ben Brown, Alexander Favorov, Peter Bick and Vladimir Tumanyan
Vavilov Institute of General Genetics of the Russian Academy of Sciences, Engelhardt Institute of Molecular Biology of the Russian Academy of Sciences
pl.maya@gmail.com

We performed a bioinformatics study of RNA-binding protein Pasilla function using an original algorithmic approach. Pasilla is the Drosophila melanogaster homologue of mammalian RNA- binding splicing regulatory factors NOVA1 and NOVA, a change in the expression of which is observed in POMA syndrome in humans. Binding sites of Pasilla and its homologues are known from experiments. We have demonstrated that Pasilla binding sites tend to occur relatively close on the RNA sequence (to form clusters). Clusters of binding sites indicate the location of regulatory elements. Whole genome analysis of the Drosophila genome revealed for the first time tens of thousands of statistically significant clusters of Pasilla binding motifs, including dense clusters of up to 100-200 bp. We performed the first of its kind whole- genome analysis of Pasilla binding sites clusters position relative to other known elements of D.melanogaster genome that are important for splicing regulation: location of the annotated today D.melanogaster exons (one hundred thousand exons) and location of experimentally confirmed alternatively spliced Pasilla-affected exons (several hundred exons). It has been shown that the boundaries of exons are significantly more often found within clusters than random genome positions. Thus, splice donor sites are found in clusters two times more frequently than acceptor sites. This effect has been shown separately for exons annotated as constitutive, exons annotated as alternative and exons for which Pasilla-affected alternative splicing was confirmed experimentally. The observed positioning of Pasilla binding motif clusters in D.melanogaster pre-mRNA matches the in vivo and in vitro derived sequence patterns in the genes of mice for the well-characterized splicing factor NOVA. The results indicate the splicing regulatory role for Pasilla protein. Our methodology and software can be used to predict regulatory proteins, their binding regions both in DNA and RNA, as well as identify potential locations of new exons.

Clonal lineage trees analysis reveals non-random multiplicity of identical heavy immunoglobulin chain sequences in gastric Diffuse Large B cell lymphoma.

Irina Iosselevitch, Hilla Tabiban-Keissar#, Helena Hazanov*, Miri Michaeli*, Iris Barshack# and Ramit Mehr**

*The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. #Dept. of Pathology, Sheba Medical Center, Tel-Hashomer, Israel.

ira.iosselevitch@gmail.com

Lymphoid malignancies can be tracked via the antigen receptor sequence of the malignant clone. The 454 system is a High Throughput Sequencing platform often used to characterize the BCR repertoires or track B cell lymphomas by sequencing of Immunoglobulin (Ig) genes. Sequencing, however, requires prior amplification by PCR, which may arbitrary amplify certain sequences over others, due to primer biases or random factors. On the other hand, multiplicity of sequences may be due to presence of B cells with identical Ig genes. In this study, heavy chain variable region genes were amplified, using nine gastric Diffuse Large B cell lymphoma samples from different patients. Lineage trees were created, and their structures were analyzed. Non-random location of multiplicative sequences was found, as the majority of multiplicities are located just one mutation away from a leaf. In addition, the maximum multiplicity in a tree increases with clone size. These results may imply dependence of malignant clone vitality on the affinity of its receptor to an antigen. In any case, data show that the sequencing results are not highly affected by random PCR biases.

Catching genes escaping X inactivation: Current knowledge and beyond

Kerem Wainer and Michal Linial
The Hebrew University of Jerusalem
keremwai@gmail.com

Compensating the gene dosage between females and males is needed in all instances where the number of sex chromosomes differs between genders. In mammals, dosage compensation is achieved by silencing one of the female's X chromosomes during early embryogenesis in a process named X inactivation. It has been shown that in humans, the silencing of the X chromosome is partial and about 10-15% of the genes escape the X inactivation. These genes are expressed from the inactivated X chromosome (Xi) and are called Escapers. For two decades, many studies focused on characterizing the Escaper genes have been limited by the fact that female tissues are mosaic (i.e., composed of a paternal and maternal Xi). The mosaicism prevented direct monitoring of tissue expression that may support the identity and properties of the Escapers. Indirect methods were developed to overcome this obstacle. In this study, we provide a systematic survey that aims to provide a measure of the consistency of identifying genes as Escapers. To this end, we compiled evidence of expression from the Xi chromosome in human through mining the literature. We designed a normalized score that takes into account the consistency in escaping X chromosome inactivation with respect to numerous complementary methodologies. In addition, we show that the current technology of RNA-Seq from single cells is valuable for a direct measurement of allelic specific expression. We took advantage of the published transcriptomic data of GM12878 lymphoid cell, in addition to the fully sequenced diploid genomes. The analysis is also adapted to determine the allelic variations from the reported exome sequences of about 30 female transformed cell-lines (of the NCI-60 collection). Our developed pipeline is designed to determine the validity of an Escaper from rich clonal RNA-seq data. We conclude that the technologies of deep sequencing, single cell or clonally and a detailed genomic information, can be combined to seek consistency for the Escapers within healthy and cancerous cells.

BARCODE: Fast lossless compression via cascading Bloom Filters

Roye Rozov, Eran Halperin and Ron Shamir
Tel-Aviv University
roye.rozov@gmail.com

Data from large Next Generation Sequencing experiments present challenges both in terms of costs associated with storage and in time required for file transfer. When a reference genome is available, effective compression can be achieved by first aligning the reads to the reference genome, and then encoding each read using the alignment position combined with the differences in the read relative to the reference. These reference-based methods have been shown to compress better than reference-free schemes, but the alignment step they require demands several hours of CPU time on a typical dataset, whereas reference-free methods can usually compress in minutes. We present a new approach that achieves highly efficient compression by using a reference genome, but completely circumvents the need for alignment, affording a great reduction in the time needed to compress. In contrast to reference-based methods that first align reads to the genome, we hash all reads into Bloom filters to encode, and decode by querying the same Bloom filters using read- length subsequences of the reference genome. Further compression is achieved by using a cascade of such filters. Our method, called BARCODE, runs an order of magnitude faster than reference-based methods, while compressing an order of magnitude better than reference-free methods, over a broad range of sequencing coverage. In high coverage (50-100 fold), compared to the best tested compressors, BARCODE saves 80-90% of the running time while only increasing space slightly.

Automatic Classification of Insects' Complete Proteomes

Nadav Rappoport and Michal Linial
The Hebrew University of Jerusalem
nadavrap@cs.huji.ac.il

Insects are the most diverse clade among animals. There are 1 million species of insects according to a conservative estimate. At present already 40 genomes of insects were completely sequenced; many belong to the Drosophilidae family. We collected 300,000 proteins from 17 fully sequenced representatives that capture the diversity of the clade. We included the Arthropod Daphnia pulex as outgroup for the analysis. A classification procedure was applied based on the similarity distance measurements among all proteins. Using the hierarchical method of ProtoNet, all the sequences were clustered to produce 20,000 protein families. We took advantage of the completeness of the 18 species to estimate the degree of protein gain and loss along speciation. We found that when comparing the Hymenoptera (i.e., ants, honeybee, wasp) to Diptera (i.e., fruitfly, mosquitoes), the former had a higher rate of proteins' gain and loss. In addition we identified 650 families that were significantly expanded or shrank for at least one of the species.For a better understanding of the functional relevance of our findings, we associate each sequence with the appropriate Pfam keywords. Each protein family was named according to the dominant annotation of its proteins. We found that families that were maximally changed during speciation are enriched with DNA binding domains and surface receptors. We conclude that innovation of transcriptional regulation and cell communication functions underlies the diversity of the insects' clade. We present ProtoBug as a resource and as an unbiased approach for classification and automatic annotations of insect proteomes.

An epigenetic signature at intragenic exons with implication for expression

Idit Kosti, Meromit Singer, Lior Pachter and Yael Mandel-Gutfreund

Technion - Israel Institute of Technology

iditk@tx.technion.ac.il

Background : DNA methylation is an important epigenetic marker associated with gene expression regulation in eukaryotes. While promoter methylation is relatively well characterized, the role of intragenic DNA methylation remains unclear. Here we investigated the relationship of DNA methylation at intragenic exons with exon expression and histone modifications generated from a human fibroblast cell-line and primary B-cells. Results: Consistent with previous work we found that intragenic methylation is positively correlated with gene expression and that exons are higher methylated than their neighboring intronic environment. Intriguingly, we show that the overall elevated DNA methylation at exons relative to their surrounding introns is primarily a characteristic of silenced genes. These results were independent of the inclusion rate of the exons, suggesting a novel role for exon methylation that does not directly relate to active transcription or splicing processes. Furthermore, we observed a negative correlation between gene-body exon methylation and the density of the majority of histone modifications. Specifically, we demonstrate that hypo- methylated exons have a characteristic histone code comprised of significantly high levels of histone markings, which are positively correlated with exon expression. Conclusions: Overall, our comprehensive analysis of the human exome supports the presence of a novel regulatory mechanism at hypo-methylated intragenic exons. In particular our results reveal a previously unrecognized diverse and complex role of the epigenetic landscape at exonic regions within the gene body.

A genome-wide map of hyper-edited RNA reveals numerous new sites

Hagit T. Porath[1]*, Shai Carmi*[2] *and Erez Y. Levanon*[1]

[1] The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University

[2] Department of Computer Science, Columbia University

hagit_br@hotmail.com

Adenosine-to-Inosine RNA editing is one of the most frequent post-transcriptional modifications and is manifested as A-to- G mismatches when comparing RNA sequences to their source DNA. Recently, a number of RNA-seq datasets have been screened for the presence of A-to-G editing, and hundreds of thousands of editing sites were found. Here, we show that existing screens miss the majority of sites by ignoring reads with excessive ("hyper") editing that do not easily align to the genome. We show that careful alignment and examination of the unmapped reads in RNA-seq studies reveal numerous new sites, usually many more than originally discovered, and in precisely those regions that are most heavily edited. Specifically, we discovered 327,096 new editing sites in the heavily studied Illumina Human BodyMap data and more than doubled the number of detected sites in several published screens. We also identified thousands of new sites in mouse, rat, opossum, and fly. We demonstrate that our detection method is highly specific and, as expected, that the human hyper-edited regions are predicted to form particularly stable double-stranded RNA structure (the target of the ADAR editing proteins). Virtually all hyper-edited regions overlap either with known repeats or with repetitive regions that are currently unannotated. Our results establish that hyper-editing events account for the majority of editing sites.

VennBlast - whole transcriptome comparison tool

Tamar Zahavi and Mali Salmon-Divon
Ariel University
tamar.harel1@mail.huji.ac.il

RNA-seq is the method of choice for getting a primary list of genes for non- model organisms. Once this is achieved, one would proceed to annotate the newly discovered genes and consequently strive to map the organism in an evolutionary context. These kinds of studies involving high-throughput sequencing generate large amounts of data, whose analysis might be time consuming for the non-specialist user and merits computational skills. Here we describe VennBlast, a set of high-performance utilities that combines fast parallelized Blast filtering with a visualization tool for whole- transcriptomic alignment comparison using Venn diagrams. The software accurately illustrates simple set relationships between numbers of matching sequences and identifies transcriptome conservation among different organisms. The intuitive Venn diagram visualization allows researchers to easily select a desired subset of genes for further inspection, using the DAVID functional annotation tools, for instance, which enables investigators to understand biological meaning behind large lists of genes. The program is configured to run as a desktop application with an interactive user interface.

Using anatomic enrichment analysis to characterize stress-Induced resting-state functional connectivity changes

Adi Maron-Katz, Sharon Vaisvaser, Tamar Lin, Ron Shamir and Talma Hendler
Tel-Aviv University, Tel Aviv University and Tel Aviv Sourasky Medical Center
adimaron@gmail.com

Despite the growing interest in the neural activity during rest, few functional magnetic resonance imaging (fMRI) studies have linked changes in resting state functional connectivity (rsFC) to a prior stressful experience. It is known that the return to homeostasis following such a challenge involves large-scale network reorganization. To date this issue has been examined using a hypothesis driven approach, which may reveal only a fraction of the actual phenomena. Here we present a robust data driven approach for detecting and interpreting large scale rsFC modulations induced following stress. We compared resting-state fMRI profiles recorded from 57 male subjects before and after performing a stress induction task. Using a predefined functional parcellation of the gray matter combined with a univariate statistical analysis, we identified 490 parcel pairs that showed significant rsFC change following stress. A set of 189 pairs demonstrated significant rsFC increase and the remaining 301 pairs showed significant decrease. Lobe pairs that were highly enriched with such modulations were detected using the Hyper-geometric test. This analysis revealed a pattern of cross hemispheral parietal-temporal connectivity within the set of weakened connections, and patterns of thalamo- cortical (frontal, parietal and temporal) connectivity within the set of strengthened connections. None of the identified modulations was found to be associated with subjective stress reports. Integrating information regarding subjective stress into the analysis, using a report-based group partition, yielded a single connection between the right amygdala and precuneus that was found to be inversely correlated with change in subjective stress rating across all subjects. Our findings indicate large scale network modulations induced by a social stress test, and demonstrate the value of enrichment analysis for interpreting such large scale modulations. At the same time these findings emphasize the need to incorporate information relevant to the investigated process into the analysis.

The Influence of Neuronal Morphology and branching points on Electrical Activity

Netanel Ofer and Orit Shefi

Faculty of Engineering and Bar Ilan Institute of Nanotechnologies and Advanced Materials, Bar Ilan University

netanel.ofer@biu.ac.il

The structure of the neuron has a great influence on its electrical activity. Studying this effect is important to the understanding of brain function and information coding. Such study can contribute to the understanding of various neuronal diseases and to potential treatments. Our research goal is to have a better understanding of the influence of cell morphology on the electrical activity and to explore the information coding possibilities derive from the ramify structure of the neuronal tree. We analyze, using numerical simulations and complex systems analysis tools, the influence of cell geometry on the electrical activity. First, we examine the effect of the axon structure on the electrical response, and then we extend the discussion also for branching points. We found several firing patterns consist of action potentials and failures, which enable controlling the neuronal information. Modifying the cell structure parameters, such as diameter of the axon, enables information coding. Branching points that lead to different response in the two daughter branches also may be a tool for controlling and filtering neuronal activity. We test potential treatments that may affect the neuronal structure virtually, thus repairing the electrical activity in modified cells. Our study leads to possible mechanisms for information coding that may promote the understanding of communication between neurons, and information transformation in the brain.

PRO-ACT: big data - the next step towards ALS cure

Hagit Alon and Neta Zach
Prize4Life
alonh@prize4life.org

Large datasets are critical for identifying and emphasizing biologically-relevant observations that are statistically significant, especially in rare diseases like ALS. The Pooled Resource Open- access ALS Clinical Trials (PRO-ACT) platform provides an unprecedented opportunity to increase our understanding of the heterogeneity of the ALS patient population and of the natural history of the disease. The PRO-ACT database covers the records of over 8,500 ALS patients who participated in 17-phase II and phase III clinical trials. Over 8 million longitudinally collected data-points include demographics, family history, vital signs, clinical assessments, lab data, medications, and survival information of the patients. Open-access to researchers worldwide has been made in December, 2012, and has attracted the attention of over 200 researchers from 23 different countries, which yielded at least 6 research papers in different publication stages. Given the nature of ALS as a rare disease with heterogeneous characteristics predicting disease progression is highly important. To that end, we undertook an innovative crowd- sourcing initiative to shed light on the difficult challenge of ALS prognosis. The DREAM- Phil Bowen ALS Prediction Prize4Life was launched to incentivize the development of novel methods to accurately predict future change in ALSFRS at the individual patient level. This program brought in over 1000 solvers from around the world and led to the development of several valuable algorithms to predict the progression of ALS, with potential to aid both clinicians and future ALS clinical trials. The challenge also led to the identification of new features predictive of ALSFRS progression that have been later verified on the full PRO-ACT database, and will be presented as well. Those algorithms have a remarkable added value: as they reduce the cost of clinical trials by 20%. Additional preliminary results also include insights in terms of correlation between ALSFRS slope and several baseline variables; the ability to establish initial relations between different ALSFRS questions and disease progression in patients; and identification of features predictive of ALS survival. Initial analysis of the data provided insights in terms of correlations between ALSFRS slope and several baseline variables. In addition, relations between different ALSFRS questions could help categorize disease progression in patients. These early results demonstrate the value of large datasets for developing a better understanding of ALS natural history, prognostic factors, and disease variables. More sophisticated and targeted analyses will continue to reveal even more about this disease, which has for so long defied our understanding.

Improving Genome-Wide Association Studies Via #Principal Components Projection

Omer Weissbrod and Dan Geiger

Technion - Israel Institute of Technology

omerw@cs.technion.ac.il

Two prevailing difficulties in genome- wide association studies (GWAS) are spurious results due to genetic confounding, and power loss due to using covariates in ascertained case- control studies. To solve the former difficulty, association studies typically use principal component (PC) covariates, linear mixed models (LMMs), or a combination thereof. Recently, LMMs have emerged as the method of choice for GWAS. However, in ascertained case-control studies, LMMs may suffer from power loss stemming from the second difficulty. This difficulty arises because of the relation between an LMM and Bayesian linear regression. Given an oracle identifying causal variants, the power loss difficulty can be solved by omitting causal variants from the LMM genetic similarity matrix (GRM). We propose a novel method that approximates an oracle LMM, by removing the effects of causal variants from the LMM GRM. The principal idea behind our method is that causal signals can be identified by identifying principal components that explain a large proportion of the phenotypic but a small proportion of the genotypic variance in the data. By projecting the genotypic data to the subspace that is not spanned by these PCs, strong causal signals are no longer represented in the GRM. We demonstrate that our method can lead to significant power gains, with over 10 percent in average test statistic in real disease data sets.

Identification of Breast Cancer Subtypes Using RNA-Seq Data

Dvir Netanely and Ron Shamir
Tel-Aviv University
dnetanely@tau.ac.il

Breast cancer is one of the most common types of cancer with more than 450,000 deaths each year worldwide. It is a very heterogeneous disease, making it very hard to treat effectively. Traditionally, breast cancer tumors were categorized into three therapeutic groups, each with its own clinical protocol. This classification was based on factors such as tumor size, histologic grade and hormone receptor status. In the past decade, several studies used gene expression microarrays to develop breast cancer subtype predictors. The PAM50 predictor, which uses a molecular signature of 50 genes, can now distinguish between four intrinsic subtypes of breast cancer: luminal A, luminal B, HER2- enriched and basal-like. In this study we have analyzed a large RNA-Seq dataset containing 612 primary breast cancer tumor samples. Our unsupervised analysis partitioned the tumors into several clusters exhibiting high concordance with the PAM50 classes. Notably, the PAM50 predictor was based on supervised analysis while we obtained the results without using any prior information. Whereas the basal-like and HER2- enriched subtypes were quite easily separable from the rest, the accurate identification of other subtypes remains a challenge. This is a first attempt to achieve breast cancer classification based on RNA-seq data. Due to the higher sensitivity of the RNA-Seq technology and the large number of samples in our analysis, we expect to reveal finer molecular patterns, which may differentiate better between known subtypes, and may also uncover new breast cancer subtypes.

EXPANDER: A Platfor for Dissecting Networks and Functions Using NGS and Microarray Expression Profiles

Tom Hait and Ron Shamir
Tel-Aviv University
sthait@gmail.com

A major challenge in the analysis of gene expression data from microarray or deep-sequencing (RNA-seq) is to extract meaningful biological knowledge out of the huge volume of raw data. The EXPANDER (EXPression ANalyzer and DisplayER) software package is an integrative platform for the analysis of gene expression data, designed as a one-stop shop tool that implements various data analysis algorithms. EXPANDER is available with pre-compiled up-to-date data supporting analysis of 17 species: human, mouse, rat, chicken, fly, zebrafish, C. elegans, yeast (s.cerevisiae and s. pombe), arabidopsis, tomato, rice, grape, listeria, leishmania, A. fumigatus, and E. coli. Typical analysis using EXPANDER starts from basic normalization and filtering of the gene expression data. Then, the data can be analyzed using diverse clustering and biclustering algorithms. The gene groups identified (as well as user-provided gene groups), can then be tested for enrichment of functionally related genes (based on Gene Ontology), co-regulated genes (using promoter sequences and microRNA target predictions) or co-localized genes. In addition, protein interaction or signaling networks can be used for detection and analysis of gene modules, by integrating capabilities of the MATISSE and SPIKE tools. For disease data, the DEGAS tool performs de novo discovery of dysregulated pathways. The integrated analysis capabilities provided by EXPANDER and its built-in support of multiple organisms make it unique among the many tools available for microarray data analysis. EXPANDER is continuously enhanced and updated. Among the more recent additions to the software are gene set enrichment analysis (GSEA), wiki-pathway based enrichment analysis, and de novo motif finding using the AMADEUS tools. Support for expression analysis based on deep sequencing (NGS) includes EdgeR as well as Wilcoxon rank-sum test for differential expression analysis.

Characterizing Heterogeneity in Single Cell Studies

Itamar Kanter[1], Francisco Herrera[2], Robert Tjian[2], Stephen Quake[2] and Tomer Kalisky[1]

[1]Bar-Ilan University

[2]Stanford University

itamarkanter@gmail.com

Transcriptional heterogeneity within cell populations can be studied using microfluidic single cell qPCR. In order to examine the differentiation of Embryonic Stem Cells (ESC) into Motor Neurons we performed time-course experiments (Days 0,5,7,9) and measured the expression levels of ~90 genes related to “stemness”, proliferation, and differentiation, as well as genes related to transcription initiation. We examined different computational methods to characterize single cell gene expression heterogeneity in differentiating cells. First we explored the data via PCA and identified a 3D tetrahedron in the first 3 principal components. Using the Pareto front concept we identified 4 modes, or in evolutionary jargon 4 \archetypes\ of cells: ESC Progenitors and two types of differentiated cells. K-means clustering identifies these four cell types however the discrimination between the different modes was somewhat difficult due to fact that the more differentiated cells occupy a continuous regime in state space and cannot be clearly divided into separate well defined modes. Finally we examined the pairwise correlation between levels of different genes throughout the differentiation process. We found that at low gene expression levels there is a significant contribution of measurement noise that affects the observed pairwise correlation.

Bacterial Tyrosine Kinases as Targets for Novel Antibacterial Drugs: Atomistic Characterization of Mechanistic Principles and Druggable Sites

Mickal Abd Alhadi, Orly Tabachnikov and Meytal Landau
Technion - Israel Institute of Technology
sm1ckal@tx.technion.ac.il

The frightening rapid increase in the occurrence of antibiotic resistance to many common bacterial pathogens calls for an immediate demand for new classes of antibacterial agents. Bacterial tyrosine kinases (BY-kinases) have recently emerged as potential attractive drug targets, being vital for core bacterial processes, while having no homology to mammalian proteins. Accordingly, developing inhibitors to BY-kinases might offer new safe and effective antibacterial drugs. Based on our preliminary computational analyses, we postulated that BY-kinase surfaces contain several putative druggable sites suited for the binding of small-molecule allosteric inhibitors. Accordingly, our working hypothesis is that structural investigation of BY-kinases from multiple bacterial species would provide insight into the function and allosteric machinery of BY-kinases as well as opportunities for inhibitor design and directed small- molecule screens. Our goals are to determine structures of BY-kinases from multiple organisms, investigate their mechanistic principles, and to identify druggable sites for the binding of small-molecule inhibitors. Specifically, our structural studies, via x- ray crystallography, will be complemented by computational analyses, including molecular dynamic simulations, sequence and phylogenetic analyses, and predictions of functional and allosteric sites, such as protein- protein interfaces with substrates, phosphatases, and within homo- oligomers. All-in-all, structural information and predictions will guide biochemical analyses including site-directed mutagenesis and kinase activity assays to assess protein-protein interaction sites and the role of putative specificity determinants. We plan to use the information about newly identified allosteric sites to develop strategies for structure-based drug design. Our goals are significant on two levels: within protein science, BY-kinases raise fundamental questions regarding activation and regulation mechanisms in bacteria-specific tyrosine kinases. Within translational biology and medical sciences, BY-kinases present ideal targets for the development of novel and safe antibacterial drugs.

Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis

Hagay Enav, Yael Mandel-Gutfreund and Oded Beja

Technion - Israel Institute of Technology

hagayenav@gmail.com

Viral genomes often contain metabolic genes that were acquired from host genomes (auxiliary genes). It is assumed that these genes are fixed in viral genomes as a result of a selective force, favoring viruses that acquire specific metabolic functions. While many individual auxiliary genes were observed in viral genomes and metagenomes, there is great importance in investigating the abundance of auxiliary genes and metabolic functions in the marine environment towards a better understanding of their role in promoting viral reproduction. In this study, we searched for enriched viral auxiliary genes and mapped them to metabolic pathways. To initially identify enriched auxiliary genes, we analyzed metagenomic microbial reads from the Global Ocean Survey (GOS) dataset that were characterized as viral, as well as marine virome and microbiome datasets from the Line Islands. Viral-enriched genes were mapped to a "global metabolism network" that comprises all KEGG metabolic pathways. Our analysis of the viral-enriched pathways revealed that purine and pyrimidine metabolism pathways are among the most enriched pathways. Moreover, many other viral-enriched metabolic pathways were found to be closely associated with the purine and pyrimidine metabolism pathways. Furthermore, we observed that sequential reactions are promoted in pathways having a high proportion of enriched genes. In addition, these enriched genes were found to be of modular nature, participating in several pathways. Our naïve metagenomic analyses strongly support the well-established notion that viral auxiliary genes promote viral replication via both degradation of host DNA and RNA as well as a shift of the host metabolism towards nucleotide biosynthesis, clearly indicating that comparative metagenomics can be used to understand different environments and systems without prior knowledge of pathways involved.

Poster list

Poster board number = Poster number in the list = Page number in the abstract book

1. **What can we learn from a system view on hypertension?** *Alon Botzer, John Moult and Ron Unger*
2. **Variability in genes and processes across human tissues** *Shoval Tirman and Esti Yeger-Lotem*
3. **VarElect: phenotype-based variation prioritizer in GeneCards** *Gil Stelzer, Anna Alkelai, Tsviya Olende, Shahar Zimmerman, Marilyn Safran and Doron Lancet*
4. **Understanding the Connection between Transcriptome and Proteome in Resting Non- Proliferating Tissues through the Example of the Auditory System** *Kobi Perl, Yoni Bhonker, Ofer Yizhar-Barnea, Shaked Shivatzki, Kathy Ushakov, Yair Pozniak, Orly Yaron, Noam Shomron, Tamar Geiger, Karen B. Avraham and Ron Shamir*
5. **The asymmetric organization of phosphorylation systems across eukaryotes** *Ilan Smoly, Esti Yeger-Lotem and Michal Ziv-Ukelson*
6. **Solving the microtargetome as a stable marriage problem quantifies the systemic mode of action of miRNAs** *Nathanael Weill, Véronique Lisi, Nicolas Scott, Paul Dallaire, Julie Pelloux and Francois Major*
7. **Short peptides that stall the ribosome are underrepresented in the proteome** *Renana Sabi and Tamir Tuller*
8. **Pathways as robust biomarkers for cancer classification: the power of big expression data** *David Amar, Tom Hait and Ron Shamir*
9. **Nucleotide Composition in Retained Introns Improves Translation Costs in Fungi** *Zafrir Zohar, Hadas Zur and Tamir Tuller*
10. **Network orientation via shortest paths** *Dana Silverbush and Roded Sharan*
11. **MalaCards: an integrated compendium for diseases and their annotation** *Noa Rappaport, Michal Twik, Noam Nativ, Gil Stelzer, Frida Belinky, Tsippi Iny-Stein, Iris Bahir, Marilyn Safran and Doron Lancet*
12. **Interactions between distant RNAs in regulatory networks** *Mor Nitzan, Avital Steiman-Shimony, Yael Altuvia, Ofer Biham and Hanna Margalit*
13. **Genome-wide detection of regulated transcriptional termination in bacteria** *Daniel Dar, Iris Karunker and Rotem Sorek*
14. **Expression signatures of cancerous vs. non-cancerous cellular proliferation and their clinical relevance** *Yedael Waldman, Tamar Geiger and Eytan Ruppin*
15. **Exploring Coding Sequence Determinants of Expression in S. cerevisiae via Combined Synthetic-Computational Biology Approach** *Shimshi Atar, Tuval Ben-Yehezkel, Ehud Shapiro and Tamir Tuller*
16. **Comperhensive study of gene expression regulating information encoded in viral genomes** *Eli Goz and Tamir Tuller*

17. **BindDB: an in-silico reverse-ChIP platform for epigenomic analysis in embryonic stem cells** *Ilana Livyatan, Yair Aaronson, David Gokhman, Ran Ashkenazi and Eran Meshorer*
18. **A computational study about the effect of oscillations in tRNA pool and translation elongation/initiation factors on translation rate** *Lior Trieman and Tamir Tuller*
19. **A Code for Transcription Elongation Speed** *Eyal Cohen and Tamir Tuller*
20. **A mathematical model of mouse NK cell maturation population dynamics shows that CD27-Mac-1-NK cells are not the main source of later NK cell subsets** *Michal Sternberg-Simon, Nissim Pinhas, Laura Chiossone, Gitit Shahaf, Thierry Walzer, Eric Vivier and Ramit Mehr*
21. **Towards the Mapping of the "Prenylosome" A Structure based Prediction of Farnesylation and Geranylgeranylation Targets** *Michal Sperber, Nir London, James Hougland, Corissa Lamphear, Elaina Zeverina, Elia Wright, Carol Fierke and Ora Schueler-Furman*
22. **The role of RNA conformations in RNA-protein recognition** *Efrat Kligun and Yael Mandel-Gutfreund*
23. **Structure-based prediction of novel Histone Deacetylase 8 (HDAC8) non-Histone substrates: Insights into the functional context of HDAC and its substrates** *Nawsad Alam, Lior Zimmerman, Noah Wolfson, Carol Fierke and Ora Schueler-Furman*
24. **Protein Complexes Assembly and Function Revealed by Structural Motion Prediction** *Adva Yeheskel , Rony Seger and Malka Cohen-Armon*
25. **Predicting Protein-DNA Interactions Using Geometric Docking Algorithms – Can DNA specificity Improve Results?** *Shula Shazman and Barry Honig*
26. **NMR structure of the soluble amyloid beta (17-34) peptide** *Gennadiy Fonar and Avraham O. Samson*
27. **Large Scale Computational Scanning for Peptide-Mediated Protein-Protein Interactions** *Yuval Sedan, Orly Marcu and Ora Schueler-Furman*
28. **How do different peptide-binding domains within a protein influence each other? A case study of interaction between SH3 and SH2 domains** *Adily Basal and Ora Schueler-Furman*
29. **Functional annotation of nucleic acid binding proteins using 3D surface historgrams** *Noam Kremen and Yael Mandel-Gutfreund*
30. **Elastic network modes refinement of the electron microscopy structure of the translocator protein** *Amir Shechvitz and Avraham O. Samson*
31. **Characterizing of functional human coding RNA editing from evolutionary, structural and dynamic perspectives** *Oz Solomon, Lily Bazak, Erez Levanon, Ninette Amariglio, Ron Unger, Gideon Rechavi and Eran Eyal*
32. **BEERT: A tool for the fast prediction of change in external ligand entropy** *Ido Ben-Shalom, Stefania Pfeiffer-Marek and Holger Gohlke*
33. **Approaching Reality: Guides toward full-resolution modeling of peptide-protein interactions** *Emma Joy Dodson, Nawsad Alam, Dima Kozakov and Ora Schueler-Furman*
34. **Widespread evidence to the role of flanking regions on transcription factor binding specificity** *Iris Dror, Remo Rohs and Yael Mandel-Gutfreund*
35. **Reconstructing Cancer Karyotypes** *Rami Eitan and Ron Shamir*

36. **RBPmap: a web server for mapping binding sites of RNA-binding proteins** *Inbal Paz, Idit Kosti, Manuel Jr. Ares, Melissa Cline and Yael Mandel-Gutfreund*
37. **NeuroPID: Genomic Prediction of Neuropeptides** *Dan Ofer and Michal Linial*
38. **Lineage tree analysis of high throughput immunoglobulin sequencing clarifies the B cell maturation pathways** *Helena Hazanov, Yu-Chang Wu, Alex Ademokun, Deborah K. Dunn-Walters and Ramit Mehr*
39. **Inferring binding site motifs from HT-SELEX data** *Yaron Orenstein and Ron Shamir*
40. **Identifying functional residues in proteins using a multilevel alphabet** *Ronit Hod and Yael Mandel-Gutfreund*
41. **Gene and microRNA expression as tools to infer spectrums of Inflammatory Bowel Diseases** *Metsada Pasmanik-Chor, S. Ben Shachar, H. Yanai, L. Baram, H. Elad, A. Ofer, E. Brazowski, N. Shomron, H. Tulchinsky and Iris Dotan*
42. **From gastritis to gastric lymphomas: Ig gene repertoire diversification and selection** *Miri Michaeli, Hilla Tabibian-Keissar, Ginette Schibi, Gitit Shahaf, Yishai Pickman, Helena Hazanov, Kinneret Rosenblatt, Deborah K. Dunn-Walters, Iris Barshack and Ramit Mehr*
43. **Filling the gaps in Non-coding DNA using comparative genomics-a review** *S. Anandakumar and P. Shanmughavel*
44. **Exploiting Regulatory Hidden Information Interleaved in the Redundancy of the Genetic Code without Prior Knowledge** *Hadas Zur and Tamir Tuller*
45. **Editing Cancer** *Nurit Paz Yaacov, Lily Bazak, Hagit Porath, Miri Danan, Ilana Sychikov, Binyamin Knisbacher, Eli Eisenberg and Erez Levanon*
46. **Construction of a Microsatellites-Based Linkage Map for the White Grouper (Epinephelus aeneus)** *Lior Dor, Andrey Shirak, Sergei Gorshkov, Mark R. Band, Abraham Korol, Yefim Ronin, Arie Curzon, Gideon Hulata, Eyal Seroussi and Micha Ron*
47. **Conserved lincRNA expression patterns during vertebrate evolution** *Hadas Hezroni and Igor Ulitsky*
48. **ConCord - coordinates conversion system** *Mark Sokol, Gil Stelzer and Mali Salmon-Divon*
49. **Computational prediction of splice-altering genetic variations** *Noa Cohen, Amit Lavon and Liran Carmel*
50. **Computational Analysis of Pasilla Binding Motifs Location in Drosophila Genome Points to the Protein Role in Splicing Regulation** *Maya Polishchuk, Ben Brown, Alexander Favorov, Peter Bick and Vladimir Tumanyan*
51. **Clonal lineage trees analysis reveals non-random multiplicity of identical heavy immunoglobulin chain sequences in gastric Diffuse Large B cell lymphoma.** *Irina Iosselevitch, Hilla Tabibian-Keissar, Helena Hazanov, Miri Michaeli, Iris Barshack and Ramit Mehr*
52. **Catching genes escaping X inactivation: Current knowledge and beyond** *Kerem Wainer and Michal Linial*
53. **BARCODE: Fast lossless compression via cascading Bloom Filters** *Roye Rozov, Eran Halperin and Ron Shamir*
54. **Automatic Classification of Insects Complete Proteomes** *Nadav Rappoport and Michal Linial*
55. **An epigenetic signature at intragenic exons with implication for expression** *Idit*

Kosti, Meromit Singer, Lior Pachter and Yael Mandel-Gutfreund

56. **A genome-wide map of hyper-edited RNA reveals numerous new sites** *Hagit T. Porath, Shai Carmi and Erez Y. Levanon*
57. **VennBlast - whole transcriptome comparison tool** *Tamar Zahavi and Mali Salmon-Divon*
58. **Using anatomic enrichment analysis to characterize stress-Induced resting-state functional connectivity changes** *Adi Maron-Katz, Sharon Vaisvaser, Tamar Lin, Ron Shamir and Talma Hendler*
59. **The Influence of Neuronal Morphology and branching points on Electrical Activity** *Netanel Ofer and Orit Shefi*
60. **PRO-ACT: big data - the next step towards ALS cure** Hagit Alon and Neta Zach
61. **Improving Genome-Wide Association Studies Via #Principal Components Projection** *Omer Weissbrod and Dan Geiger*
62. **Identification of Breast Cancer Subtypes Using RNA-Seq Data** *Dvir Netanely and Ron Shamir*
63. **EXPANDER: A Platfor for Dissecting Networks and Functions Using NGS and Microarray Expression Profiles** *Tom Hait and Ron Shamir*
64. **Characterizing Heterogeneity in Single Cell Studies** *Itamar Kanter, Francisco Herrera, Robert Tjian, Stephen Quake and Tomer Kalisky*
65. **Bacterial Tyrosine Kinases as Targets for Novel Antibacterial Drugs: Atomistic Characterization of Mechanistic Principles and Druggable Sites** *Mickal Abd Alhadi, Orly Tabachnikov and Meytal Landau*
66. **Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis** *Hagay Enav, Yael Mandel-Gutfreund and Oded Beja*
67. **Human areal expression of most genes is governed by regionalization** *Lior Kirsch, Gal Chechik*

Author index