Ron’s summary of ISMB meeting - Berlin 21-23 Sept 2013

(Followed by Didi and Yaron's additions)

General:

ISMB is the largest bioinformatics conference, with some 1500 participants, 8-10 parallel sessions, hundreds of posters, and a bit of a carnival atmosphere. It is way too large to my taste, but is a good opportunity to meet everyone in the community. Since the meeting is so large and distributed you will miss many unless you schedule a meeting with them. There is huge variety in quality of talks. Usually the better ones are found in the proceedings track (refereed for the conference) and the highlight tracks (recently published papers from good journals) but there are exceptions in each directions.

The Berlin meeting took place in a large convention center, ICC, in the west end of Berlin. It was very well organized though a bit stingy on the coffee and refreshments.

Notable talks:

In the supplement I list the info from the program about the talks I attended. Here I summarize the most important points and action items.

**Hector Corrada Bravo (Maryland) Gene expression anti-profiles as a basis for accurate universal cancer signatures**

Highlight track paper - BMC Bioinfo 12

A new, very interesting approach to finding cancer signatures: Everybody is looking for genes that are differentially expressed between cases and controls, i.e. consistently high in one set and consistently low in the other. Instead, they look for genes that manifest consistent (typically low) expression in the controls, but hyper-variability (typically high, but with a large variance) in the cases. They demonstrate this on particular cancer types and give some evidence for a universal cancer anti-profile that accurately distinguishes cancer from normal regardless of tissue type. Importantly, this method uses single-chip normalization and quality assessment methods so no further retraining of signatures would be required before their application in clinical settings. R code is available.

Read: (1) Hansen et al Nat Gen 11 epigenetic experiments: loss of methylation in genomic blocks in colon cancer causes loss of silencing of certain genes (2) their paper BMC Bioinfo 12

**DIDI:** look into their single chip normalization methods. They seem to work for them and avoid the problems that we encounter with batches. Three methods were cited for QC and normalization (fRMA, GNUSE, Barcode).

**DVIR:** please read these papers carefully! This new approach may have something in it and the results look very promising. Also, talk to **ELLA** about this! It is very much in the spirit of the discussion we had last week, on using expression and methylation in normal tissues to understand breast cancer.

**Peter Glaus (Manchester) Identifying differentially expressed transcripts from RNA-seq data with biological variation**


Developed a new method (BitSeq) to compute DE in RNA-seq data. It uses a Bayesian approach with MCMC (and faster but less accurate collapsed Gibbs sampling). Shows improvement over cufflinks and others.

**DVIR:** please read. It is about time we start using read-based RNAseq values.)
Yves Lussier (Illinois)  Interpreting Personal Transcriptomes: Personalized Mechanism-Scale Profiling Predicts Survival in Oral, Prostate, Lung and Gastric Cancers

Highlight – PLOS CB 12 paper (and 3 followup papers)
Method to construct a signature based on pathway scores. Claims to outperform GSEA. The speaker (a clinician) was ultra-verbose so it was hard to assess the work based on the talk.

DIDI and DVIR – worth reading. This problem is in our ballpark.

Michael Baym (MIT, B. Berger's group) Compressive Genomics
HT – Nat Biotech 12
Another presentation of the compression method based on near identical sequence removal. Focus on CaBLAST and CaBLAT. Had another talk on protein BLAST in the conference. The work is oversold but the idea is important and there is tons of rooms for improvement.

ROYE: worth following.

Pedro Beltrao (EBI, work with N Krogan UCSF) Differential genetic interactions of S. cerevisiae stress response pathways
Report on differences in GIs between six stress conditions. Not impressive talk but the data is.

DIDI: take a look at the data. Can we apply the MMAP idea to these data? Also take a look again at the Bandyopadhay paper Science 10 (Ideker's lab)

Nils Gehlenborg (Harvard) Visual Exploration for Cancer Subtype Analysis
Presented and demonstrated StratomeX, a visualization tool for comparing by eye clustering solutions on patients along with additional TCGA data. No analysis but visualizations, some of them useful for human exploration. The analysis is done by FIREHOSE offline and imported regularly. Tool comes with all TCGA data (?!).

DVIR – worth playing a bit with this tool. May be a simpler way to import the TCGA data?

Nikolaus Shultz (MSKCC) The cBio Portal for Cancer Genomics
Presented a tool/portal for visualization of cancer data, with emphasis on mutations in individual genes. Includes nearly 10000 patient samples.

DVIR/RAMI – הלשון הכלית

Hi-Son Le (CMU, Ziv Bar Joseph group) Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation
Describes a method to model mRNA levels as a function of mirna levels, mirna-target predictions, PPIs. Statistical method based on forming modules, solved using local optimization and sparseness constrains. Role of modules is unclear but the method is intriguing.

DIDI – worth reading

Johannes Soedig (Ludwig Maximilian Univ Munich) XXMotif: A powerful algorithm and web server for de novo motif discovery in nucleotide sequences
Demo track – tool reported in Genome Res 2012
Described and demoed their motif finding tool which was reported to outperform all else in extensive tests. Stole lots of approaches and visualizations from our Amadeus paper. Claims that the use of biophysical model of binding is key to improvement.
**Yaron:** It is about time we learn and understand the biophysical models of Stormo and of these guys.

**Supplement**

**PP14 (HT) - Identifying differentially expressed transcripts from RNA-seq data with biological variation**

**Date:** Sunday, July 21  **Room:** Hall 14.2  
**Presenting author:** Peter Glaus, University of Manchester, United Kingdom  
**Additional authors:** Antti Honkela, University of Helsinki, Finland  
Magnus Rattray, University of Manchester, United Kingdom

**Presentation Overview:**

Analysing RNA-seq data poses multiple challenges due to base mismatches, non-uniform read distribution, reads shared by multiple splice variants and other factors which make the expression analysis especially difficult. The BitSeq method uses a Bayesian approach to model the read generation and sequencing processes and infers expression estimates of individual transcripts. Transcript expression levels can be used to obtain more accurate gene expression estimates, in comparison to popular count based methods, or for identifying differentially expressed transcripts or genes. Our differential expression model combines the uncertainty of the expression estimates with variances estimated from biologically replicated experiments to identify significantly differentially expressed transcripts with improved precision. We present advantages of using BitSeq in RNA-seq datasets dealing with multi-mapping reads and non-uniform read distribution. Experiments with real and synthetic datasets show that BitSeq produces state-of-the-art results in both expression estimation and differential expression analysis.

**PP16 (HT) - Gene expression anti-profiles as a basis for accurate universal cancer signatures**

**Date:** Sunday, July 21  **Room:** Hall 7  
**Presenting author:** Hector Corrada Bravo, University of Maryland, United States

**Presentation Overview:**

Gene expression anti-profiles are a new computational approach for developing cancer genomic signatures that specifically take advantage of gene expression heterogeneity. This presentation will describe the biological basis for this method derived from experimental findings suggesting that stochastic across-sample hyper-variability in the expression of specific genes is a stable and general property of cancer. Application of this methodology in screening patients for colon cancer based on expression measurements obtained from peripheral blood samples will be presented. We will also present results from development of a universal cancer anti-profile that accurately distinguishes cancer from normal regardless of tissue type. This method uses single-chip normalization and quality assessment methods so no further retraining of signatures would be required before their application in clinical settings. These results suggest that anti-profiles may be used to develop inexpensive and non-invasive universal cancer screening tests.

**Keyword:** Applied Bioinformatics, Gene Regulation & Transcriptomics

**PP20 (PT) - Poly(A) motif prediction using spectral latent features from human DNA sequences**

**Date:** Sunday, July 21  **Room:** Hall 14.2  
**Presenting author:** Bo Xie, Georgia Institute of Technology, United States
**Presentation Overview:**

Motivation: Polyadenylation is the addition of a poly(A) tail to an RNA molecule. Identifying DNA sequence motifs that signal the addition of poly(A) tails is essential to improved genome annotation and better understanding of the regulatory mechanisms and stability of mRNA. Existing poly(A) motif predictors demonstrate that information extracted from the surrounding nucleotide sequences of candidate poly(A) motifs can differentiate true motifs from the false ones to a great extent. A variety of sophisticated features has been explored, including sequential, structural, statistical, thermodynamic and evolutionary properties. However, most of these methods involve extensive manual feature engineering, which can be time-consuming and can require in-depth domain knowledge.

Results: We propose a novel machine learning method for poly(A) motif prediction by marrying generative learning (hidden Markov models) and discriminative learning (support vector machines). Generative learning provides a rich palette on which the uncertainty and diversity of sequence information can be handled, while discriminative learning allows the performance of the classification task to be directly optimized. Here, we employed hidden Markov models for fitting the DNA sequence dynamics, and developed an efficient spectral algorithm for extracting latent variable information from these models. These spectral latent features were then fed into support vector machines to fine-tune the classification performance. We evaluated our proposed method on a comprehensive human poly(A) dataset that consists of 14,740 samples from 12 of the most abundant variants of human poly(A) motifs. Compared with one of previous state-of-art methods in the literature (the random forest model with expert-crafted features), our method reduces the average error rate, false negative rate and false positive rate by 26%, 15% and 35%, respectively. Meanwhile, our method made about 30% fewer error predictions relative to the other string kernels. Furthermore, our method can be used to visualize the importance of oligomers and positions in predicting poly(A) motifs, from which we can observe a number of characteristics in the surrounding regions of true and false motifs that have not been reported before. Availability: website: http://sfb.kaust.edu.sa/Pages/Software.aspx

**Keyword:** Poly(A) motif, classification, generative learning, discriminative

Keynote (Overton Prize lecture): Insights from sequencing thousands of human genomes, Golncalo Abecasis, U. Michigan

**PP26 (HT) - Interpreting Personal Transcriptomes: Personalized Mechanism-Scale Profiling Predicts Survival in Oral, Prostate, Lung and Gastric Cancers**

**Date:** Monday, July 22  
**Room:** Hall 14.2  
**Presenting author:** Yves Lussier, The University of Illinois, United States

**Additional authors:**
- Xinan Yang, The University of Chicago, United States
- Kelly Regan, Ohio State University, United States
- Yong Huang, The University of Chicago, United States
- Jianrong Li, The University of Illinois at Chicago, United States
- Ezra Cohen, The University of Chicago, United States
- Tanguy Zeiwert, The University of Chicago, United States

**Presentation Overview:**

Gene expression signatures that are predictive of therapeutic response or prognosis are increasingly useful in clinical care; however, mechanistic interpretation of expression
arrays remains an unmet challenge. We developed a novel approach to generate "personal mechanism signatures" of molecular pathways and functions from gene expression arrays. FAIME, the Functional-Analysis-of-Individual-Microarray-Expression, computes mechanism scores using rank-weighted gene expression of an individual sample. In oral squamous cell carcinoma samples, the overlap of "Oncogenic Mechanisms of OSCC" (deregulated FAIME-derived scores of pathways and biological functions) accurately discriminate clinical samples in two additional datasets (n=35;91, F-accuracy=100%;97%, p<0.001), and predicts patients' survival in two studies (p=0.0018;p=0.032). Previous approaches depending on group assignment of individual samples before selecting features or learning a classifier are limited by design to discrete-class prediction. FAIME is more amenable for clinical deployment since it translates the gene-level measurements of each given sample into pathways and molecular function profiles that can be applied to analyze continuous phenotypes (e.g. survival-time).

**Keyword:** Applied Bioinformatics, Databases & Ontologies

**PP30 (HT) - Compressive Genomics**
**Date:** Monday, July 22
**Room:** Hall 14.2
**Presenting author:** Michael Baym, Harvard Medical School, United States

**Additional authors:**
Po-Ru Loh, MIT, United States
Bonnie Berger, MIT, United States

**Presentation Overview:**

The past two decades have seen an exponential increase in sequencing capabilities, outstripping advances in computing power. Extracting new insights from the datasets currently being generated will require not only faster computers; it will require smarter algorithms. However, most genomes currently sequenced are highly similar to ones already collected; thus the amount of novel sequence information is growing much more slowly. We show that this redundancy can be exploited by compressing the data so as to allow direct computation on the compressed data. This approach reduces the computational task of operating on many similar genomes to slightly more than that of operating on just one. Moreover, its relative advantage over existing algorithms grows with the accumulation of future genomic data. We demonstrate this compressive architecture by implementing versions of both BLAST and BLAT, and emphasize how compressive genomics, more generally, will enable biologists to keep pace with current data.

**Keyword:** Sequence Analysis, Databases & Ontologies

**LBR12: – Network based stratification of tumor mutants**
**Presenting author:** Matan Hofree (UCSD, Ideker's lab)

**Presentation Overview:**

**SS04 part A – Differential genetic interactions of S. cerevisiae stress response pathways**
**Presenting author:** Pedro Beltrao

**Presentation Overview:**

**WK04 part C – bioinformatics for the clinical audience**
**Presenting author:** Donna Slonim, Tufts University

**Presentation Overview:**
SS04 part LBR14 The Yule-Simpson effect cases doubt on DNA methylation differences at functional boundaries

Presenting author: Meromit Singer (Berkeley, Lior Pachter's group)

Presentation Overview:

Keynote: Gary Stormo: searching for signals in sequences

Presentation Overview:

Keyword: Sequence Analysis, Databases & Ontologies

P57 (HT) - Visual Exploration for Cancer Subtype Analysis

Date: Tuesday, July 23 Room: Hall 7

Presenting author: Nils Gehlenborg, Harvard Medical School, United States

Additional authors:
Alexander Lex, Harvard University, United States
Marc Streit, Johannes Kepler University Linz, Austria
Hans-Joerg Schulz, University of Rostock, Germany
Christian Partl, Graz University of Technology, Austria
Dieter Schmalstieg, Graz University of Technology, Austria
Peter Park, Harvard Medical School, United States

Presentation Overview:

This talk will introduce the promises and challenges of identifying and characterizing tumor subtypes in cancer genomics data sets from patient cohorts with hundreds of patients and how our visual exploration system Caleydo StratomeX (http://stratomex.caleydo.org) supports these processes. Heterogeneous data sets including multiple genomic (mRNA, miRNA, RPPA, copy number, gene mutations) and clinical data types can be loaded into the software to efficiently generate and confirm hypotheses about tumor subtypes and their functional and clinical effects.

In order to help analysts to identify promising candidate subtypes, StratomeX has been extended with computational methods to rank stratifications and identify stratifications that provide corroborating evidence for candidate subtypes. This previously unpublished feature as well as a new interactive website with large heterogeneous data sets from The Cancer Genome Atlas (TCGA) will be presented, too.

The talk will demonstrate the utility of StratomeX through a comprehensive case study from TCGA.

Keyword: Bioimaging & Data Visualization, Applied Bioinformatics

PP61 (HT) - The cBio Portal for Cancer Genomics

Date: Tuesday, July 23 Room: Hall 7

Presenting author: Nikolaus Schultz, Memorial Sloan-Kettering Cancer Center, United States

Additional authors:
Jianjiong Gao, Memorial Sloan-Kettering Cancer Center, United States
B. Arman Aksoy, Memorial Sloan-Kettering Cancer Center, United States
Benjamin Gross, Memorial Sloan-Kettering Cancer Center, United States
Gideon Dresdner, Memorial Sloan-Kettering Cancer Center, United States
S. Onur Sumer, Memorial Sloan-Kettering Cancer Center, United States
Ethan Cerami, Memorial Sloan-Kettering Cancer Center, United States
Anders Jacobsen, Memorial Sloan-Kettering Cancer Center, United States
Ugur Dogrusoz, Bilkent University, Turkey
Erik Larsson, University of Gothenburg, Sweden
Chris Sander, Memorial Sloan-Kettering Cancer Center, United States

Presentation Overview:

The cBio Portal for Cancer Genomics (cbioportal.org) provides an integrated and easy to use web resource for exploring, visualizing and analyzing multidimensional cancer genomics data. The portal reduces massive molecular profiling data from cancer tissues and cell lines to a readily understandable form as genetic, epigenetic, gene expression and proteomic events. The combination of a convenient query interface and customized data storage enables researchers to interactively explore genetic alterations across samples, genes and pathways and to link these to clinical outcomes, when available. The portal provides graphical summaries of gene-level data from multiple platforms, network visualization and analysis, survival analysis, and patient-centric queries. With its simple, yet powerful and flexible, interface and software programmatic access, the portal makes complex cancer genomics profiles accessible to researchers and clinicians without requiring bioinformatics expertise, thus facilitating biological discoveries.

Keyword: Bioimaging & Data Visualization, Databases & Ontologies

PP70 (PT) - Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model

Date: Tuesday, July 23  
Room: Hall 14.2

Presenting author: Nicola Bonzanni , VU University Amsterdam, Netherlands

Additional authors:  
Abhishek Garg, Swiss Institute of Bioinformatics, Switzerland  
K. Anton Feenstra, VU University Amsterdam, Netherlands  
Judith Schütte, University of Cambridge, United Kingdom  
Sarah Kinston, University of Cambridge  
Diego Miranda-Saavedra, University of Cambridge  
Jaap Heringa, VU University Amsterdam / Netherlands Bioinformatics Centre  
Ioannis Xenarios, Swiss Institute of Bioinformatics  
Berthold Göttgens, University of Cambridge, United Kingdom

Presentation Overview:

Motivation: Combinatorial interactions of transcription factors with cis-regulatory elements control the dynamic progression through successive cellular states and thus underpin all metazoan development. The construction of network models of cis-regulatory elements therefore has the potential to generate fundamental insights into cellular fate and differentiation. Haematopoiesis has long served as a model system to study mammalian differentiation, yet modelling based on experimentally informed cis-regulatory interactions has so far been restricted to pairs of interacting factors. Here we have generated a Boolean network model based on detailed cis-regulatory functional data connecting 11 haematopoietic stem/progenitor cell (HSPC) regulator genes. Results: Despite its apparent simplicity, the model exhibits surprisingly complex behaviour that we charted using strongly connected components and shortest-path analysis in its Boolean state space. This analysis of our model predicts that HSPCs display heterogeneous expression patterns and possess many intermediate states that can act as 'stepping stones' for the HSPC to achieve a final differentiated state. Importantly, an external perturbation or 'trigger' is required to exit the stem cell state, with distinct triggers characterising maturation into the various different lineages. By focussing on intermediate states occurring during erythrocyte differentiation, from our model we predicted a novel negative regulation of Fli1 by Gata1 which we confirmed experimentally thus validating our model. In conclusion, we demonstrate that an advanced mammalian regulatory network model based on experimentally validated cis-regulatory interactions has allowed us to make novel,
experimentally testable hypotheses about transcriptional mechanisms that control
differentiation of mammalian stem cells.

**Keyword:** Haematopoietic stem cell, cis-regulatory elements, transcription factor netw

---

**PP74 (HT) - Interplay of microRNAs, transcription factors and target genes:**
linking dynamic expression changes to function

**Date:** Tuesday, July 23  **Room:** Hall 14.2

**Presenting author:** Petr Nazarov, Centre de Recherche Public de la Sante, Luxembourg

**Additional authors:**
- Susanne Reinsbach, University of Luxembourg, Luxembourg
- Arnaud Muller, Centre de Recherche Public de la Sante, Luxembourg
- Nathalie Nicot, Centre de Recherche Public de la Sante, Luxembourg
- Demetra Philippidou, University of Luxembourg, Luxembourg
- Laurent Vallar, Centre de Recherche Public de la Sante, Luxembourg
- Stephanie Kreis, University of Luxembourg, Luxembourg

**Presentation Overview:**

MicroRNAs (miRNAs), small non-coding RNAs that negatively regulate gene expression at
the post-transcriptional level, are involved in fine-tuning fundamental cellular processes
and are believed to confer robustness to biological responses. Using microarray data we
investigated simultaneously the transcriptional changes of miRNA and mRNA expression
levels over time after activation of the Jak/STAT pathway by IFN-? stimulation of
melanoma cells. We observed delayed responses of miRNAs (after 24-48 h) with respect to
mRNAs (12-24 h) and identified biological functions involved at each step of the cellular
response. Inference of the upstream regulators allowed for identification of transcriptional
regulators involved in cellular reactions to IFN-? stimulation. Linking expression profiles of
transcriptional regulators and miRNAs with their annotated functions, we demonstrate the
dynamic interplay of miRNAs and upstream regulators with biological functions. Finally, our
data revealed network motifs in the form of feed-forward loops involving transcriptional
regulators, mRNAs and miRNAs.

**Keyword:** Gene Regulation & Transcriptomics

---

**PP78 (PT) - Integrating sequence, expression and interaction data to determine
condition-specific miRNA regulation**

**Date:** Tuesday, July 23  **Room:** Hall 14.2

**Presenting author:** Hai-Son Le, Carnegie Mellon, United States

**Additional authors:**

**Presentation Overview:**

Motivation: MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression
post-translationally. MiRNAs were shown to play an important role in development and
disease, and accurately determining the networks regulated by these miRNAs in a specific
condition is of great interest. Early work on miRNA target prediction has focused on
utilizing static sequence information. More recently, researchers have combined sequence
and expression data to identify such targets in various conditions. Results: Here we
propose a regression-based probabilistic method that integrates sequence, expression and
interaction data to identify modules of mRNAs controlled by small sets of miRNAs. We
formulate an optimization problem and develop a learning framework to determine the
module regulation and membership. Applying our method to cancer data we show that by
adding protein interaction data and modeling combinatorial regulation our method can
accurately identify both miRNA and their targets improving upon prior methods. We next
used our method to jointly analyze a number of different types of cancers and identified
both common and cancer type specific miRNA regulators.
**Keyword:** microRNA, gene regulation, transcriptomics, regulatory network

**TT35: XXMotif – A powerful algorithm and web server for de novo motif discovery in nucleotide sequences**

**Presenting author:** Johannes Soedig, Munich

**Presentation Overview:**

**Keyword:** Sequence Analysis, Databases & Ontologies

**ISCB Senior scientist accomplishment award keynote: David Eisenberg UCLA**

**Presentation Overview:**

---

**From:** ddam.am [mailto:ddam.am@gmail.com]
**Sent:** Sunday, July 28, 2013 6:37 PM
**To:** Ron Shamir
**Cc:** shamirgroup
**Subject:** Re: Summary of ISMB 2013 with action items

Hi all,

My additional two cents.

1. **Re - Hector Corrada Bravo (Maryland) Gene expression anti-profiles as a basis for accurate universal cancer signatures**

   I also like the idea of looking at changes in variation. However, I am not sure their single cell analysis will be helpful in cases similar to genepark. Actually, fRMA is one of the first things we tried, it did not help with our biases. Similar observation was made by Leek et al. (2012, these are the authors of fSVA) which pointed out that using fRMA did not help in leave-batch-out tests. This can be explained partially due to the fact that fRMA is good for removing statistical noise and making the training set "frozen" for downstream analyses. When the noise in the data is "biological" (i.e., mRNA of gene A appears in one batch and not in the other) this step will not help. I will try to look at their additional methods.

2. **Re - Yves Lussier (Illinois) Interpreting Personal Transcriptomes: Personalized Mechanism-Scale Profiling Predicts Survival in Oral, Prostate, Lung and Gastric Cancers**
I attended his talk. I got the feeling he spoke on many things but failed to explain the method clearly. Anyway, I see more and more patient-specific pathway analyses such as presented in this paper. Other example is GSVA (http://www.ncbi.nlm.nih.gov/pubmed/23323831) a recent method perform similar analysis.

One thing that is missing in these methods is using the actual pathway interactions. Basically, we can think about extending these ideas, the same as SPIA (http://www.ncbi.nlm.nih.gov/pubmed/18990722) extended GSEA based approaches.


This analysis is on taking a network (such as PPI or genetic networks) and automatically constructing a gene ontology. The method is based on running hierarchical clustering on the network and then post-processing the tree to understand how the ontology DAG is expected to look. Seems simple and elegant analysis that rediscovered most of the GO structure (in yeast) and suggested some corrections that were actually accepted by the GO consortium (!).


This talk was in the special session on dynamic networks and contained results from several works. Basically they show how to detect tissue-specific PPIs. The method is based on finding PPIs for which both proteins (i.e., their genes) are much more expressed in a tissue of interest. He said their method is three-fold better than just looking at if the gene pair is expressed or not (which is what is done by the lab of Esti Yeger-Lotem) - in terms of predictability across different datasets of the same tissue. Also, he showed that drug effects are biased towards increasing genes of tissue-specific PPIs.

5. Bioinformatics workshop on integrative analysis

This was almost a waste of time (although the name seemed promising). The session was way too technical without giving good explanation of the methods that were used. The session was given by guys that provide computational support to research groups (similar to the bioinfo support group in the LS department). However, they did mention an R package called party (http://cran.r-project.org/web/packages/party/party.pdf) for integrative analysis of survival data and expression data. They did not explain how and what this method does but mentioned that they are using it for subclass discovery - so this can be relevant to Dvir's project.

Used data from GEO of Affymetrix HGU Plus 2 data - ~14,000 samples.

The goal was to classify samples to tissues - organized using the MeSH tree of tissues.

The method was based on using SVM at each node of the tree and then applying Bayesian correction for the prediction (need to look into that in more detail). The method is called URSA.

Validation was performed by looking on small group of RNA-seq datasets and other Affy platform. I was hard to understand from the talk how good were the scores of the validation.

From: Yaron Orenstein [mailto:yaron.orenstein@gmail.com]
Sent: Monday, July 29, 2013 3:41 PM
To: ddam am
Cc: Ron Shamir; shamirgroup
Subject: Re: Summary of ISMB 2013 with action items

I will add my insights on the motif finding talks.

First, re the organization. Saturday (Special Interest Groups meetings) was very good - including cafe, cakes, lunch, soft drinks, and alcohol at the end of the day. I could not attend all talks due to volunteering assignments, but I got to fully attend the most relevant day (SIG on regulatory genomics). Keynotes were too general, in my opinion, and had little relevancy to students. You would get more from this broad viewpoint if you had a few years of experience in the field.

Two talks from regulatory genomics SIG:

1. Ivan Kulakovskiy (Russian Academy of Sciences, Moscow) - From binding motifs in chip-seq data to improved models of transcription factor binding sites.

Meeting abstract: http://light.ece.ohio.edu/~reggen/2013/Ivan_Kulakovskiy.pdf

These are the guys of ChIPMunk (motif finding in ChIP-seq data). They improved their models by using the di-nucleotide model, instead of the regular PWM. The only change is using an alphabet of 16 letter (di-nucleotide pair) instead of 4 at each position. They show improvement on five examples compared to the normal PWM model. They even compared to two HT-SELEX models. I discussed the SELEX data with them, and learned that they are also looking at it, and observed some of the phenomenon we found. They are not comparing it to PBM data.

2. Johannes Soding (Ludwig Maximilian University, Germany) - Drosophila Pol II core promoters cluster into four classes characterized by distinct sets of motifs, regulatory properties and nucleosome patterning

These are the guys of the new XXmotif motif finding tool (http://xxmotif.genzentrum.lmu.de/), published in NAR (http://nar.oxfordjournals.org/content/40/W1/W104.long) and Genome Research (http://genome.cshlp.org/content/early/2012/09/18/gr.139881.112.abstract). They have given a short promotion during the talk, but the main one was in the technology track, which I missed.

They claim that a significant contribution comes from the PWM refinement, which is more related to the energetic contribution of each position, as represented in the PWM model.

As for talks in the conference days:

3. Fantine Mordelet (Duke University, North Carolina) - Stability selection for regression-based models of transcription factor-DNA binding specificity

http://bioinformatics.oxfordjournals.org/content/29/13/i117.abstract

In this work, they investigated the use of more complex models for TF-DNA binding. They incorporated additional features, such as 2-mers and 3-mers. Their data were context-genomic PBM experiments, i.e. each oligo contains the CACGTG and different flanking sequences. They developed a machine learning approach that robustly selects a small set of parameters. They show (on the same data) that they capture the specificity of TFs with similar consensus sequence better than in the PW model. I plan to talk about it in the group meeting.

4. Limor Leibovich (Zohar Yakhini, Technion)- DRIMust: a web server for Discovering Rank Imbalanced Motifs Using Suffix Trees
Another motif finding platform. Limor gave a promotion for it in the technology track. It was published in NAR (http://nar.oxfordjournals.org/content/41/W1/W174.long and http://nar.oxfordjournals.org/content/early/2012/03/12/nar.gks206).

Their main advantage is using the minimum-hyper-geometric score, which exempts the user from defining a set of 'positives' (i.e., the target set) and by this utilizing the ranked list of genes.

Using suffix trees allows efficient search of k-mers and gapped k-mers. The webserver has a user-friendly interface. Talking to the guy of XXmotif, he complained to me that Limor didn't use his software correctly in the comparison; they gave the full list of genes to XXmotif instead of defining a positive set.

Sounds like a good solution for ranked gene lists.