About the conference: The 2019 ISMB (Intelligent Systems in Molecular Biology) conference was held in Basel. ISMB is the annual meeting of the International Society of Computational Biology and the largest in the field. Once in three years it co-locates with the smaller European compbio conference ECCB. The five-day event had 8-10 parallel sessions and a record number of >2000 participants. The society has been struggling for years to make the conference more attractive and better organized, since with so many activities in parallel it is quite hard to find what one wants. The innovation this year was the organization of the program around Communities of Scientific Interest or COSIs – these are subgroups of scientists interested in one area (e.g. Medical informatics, Metagenomics, Computational Oncology, Machine Learning in Computational and Systems Biology, RNA, CAMDA, BOSC, EvolCompGen,... there are already more than 20 COSIs. I mention some more below. See the society’s website. Each COSI had a committee that chose talks for presentation. There was also a possibility of submitting new works for proceedings, and accepted papers will appear in the journal Bioinformatics. There were >1000 posters in three sessions.

The conference took place in the Basel Conference Center, an impressive and highly professional venue. The conference was extremely well organized and pleasant, in spite of simmering heat outside (38°C). Still, I found it hard to cherry pick the best talks and moving between sessions was not easy. I am also concerned about the COSIs forming "silos" in the society if they persist. I guess this is the inevitable price of the growing size.

Topic-wise, machine learning techniques and computational medicine were everyone, and were also overrepresented in the selected proceedings papers. I will review below some selected talks. All talks were videotaped and will be available on the ISCB website and the ISCBV u-tube channel. If a talk I sketch (or any) is of interest to you – watch it. Most talks except the keynotes were less than 20 minutes long.

I chose to sketch several lectures that are of broad interest to all, and others that may interest more some group members. There are many good topics that I missed completely due to overlaps, e.g. the sessions on Metagenomics, Computational Oncology and RNA.

I marked in boldface some valuable resources and tools that were new to me (or I thought may be new to some of you).

Nikolaus Rajewsky (MDC Berlin): Gene regulation in space and time by single cell analysis

Rajewsky is the head of a new large research center in Berlin, which combines advanced technologies with bioinformatics to study cellular processes on a very large scale. His focus is RNA. In this keynote, he described how they reconstructed the spatial organization cells based on measurements of scRNA in many thousands of single cells (RNA was measured using a variant of the Dropseq technique). In profiling the single cell NRA, the spatial information is lost. With Mor Nitzan (HUJI and now Harvard) and Nir Friedman (HUJI) they developed a computational method called NOVOSPARC that uses linear transport theory and linear programming to reconstruct the RNA profiles of the cells in space based on profile proximity
(paper in BioRxiv). He showed striking examples from drosophila, mouse gut and liver and even brain (!). The vision is to use this technique in order to analyze disease biopsies, in order to link tissue and genetic information.

He then moved on to describe a technique to organize scRNA profiles in time. The technique uses the ratio of spliced and unspliced transcripts in cells in order to compute "RNA velocity" (Plass et al., Science 18) and use this to order cells in time. In planaria, where regeneration is always ongoing, they used it to reconstruct the complete lineage tree (Wolf et al., GB 19). They also reconstructed cell cycle temporal order using RNA velocity. In a cancer model in mice they could identify the tumorigenesis trajectory.

Rajewsky’s vision is summarized in a huge project called LifeTime aiming to study cellular spatial and temporal patterns on a very large scale (eventually all human tissues), in order to identify cells in early disease stages that are undetectable today. The rationale is that treating cells that are only slightly deviating from the normal behavior – if we can identify them – would be much easier than doing that in late disease stages, where the RNA profile is dramatically and irreversibly changed. About a hundred institutes and companies are involved in LifeTime, which still seeks funding.

*Talk of interest to: All.*

**William Noble (U. Washington): Embedding genomic and proteomic data**

Noble, who won the mid-career award of the ISCB, gave a clear and impressive keynote about ML methods in his research. One problem he discussed was integration of single cell genomic data of different types (multi-omics) by embedding the different data types to a shared low dimensional manifold. Several techniques were previously described for that challenge (e.g. MAGAN, Amodio et al, ICML 19, GUMA, Cui et al NIPS 14). His method is based on MMD – a measure for distance between distributions. The algorithm called **MMD-MA** (developed with J.P. Vert) was used for joint embedding of scHIC and single-cell imaging data.

Noble also described a new algorithm called **AVOCADO** for imputation of ENCODE data. The matrix of tissues x assays in ENCODE is highly missing and not at random. Similar to the idea of the Netflix competition for predicting viewer preferences, they decompose the data into genomic positions x assay factors x cell type factors and show that they improve the TF binding predictions versus prior methods. The general vision is to use latent embedding in order to understand biology and not just as a computational black box.

Incidentally, he mentioned that the benchmark used by the authors of TargetFinder to identify promoter-enhancer interaction was criticized for bias, and that their analysis was performed on a corrected benchmark.

*Most relevant to Nimrod, Tom, Hagai.*

**Isaac Kohane (Harvard) – 5 ways compbio can accelerate medicine**

This was a keynote talk in the Systems Medicine COSI. Kohane gave a very high-level talk on the pitfalls and potential in combining computational and clinical work. As usual, his talk had
a lot of flare and was quite funny, but one should listen to him, as he is at the forefront of translating computational analysis to physicians. His main messages:

(1) "Quants" (computational scientists) should undergo short and intensive apprenticeships (e.g. postdocs) in medical departments. Otherwise, purely computational findings could be nonsense due to not understanding the health systems dynamics. (Example: taking blood sample at 3am is a great but useless predictor of death).

(2) Create collaborations between the meds and the quants. (Example: Undiagnosed Disease Network of 12 medical centers. All patients clinical and genotypic data is available identified in Amazon clouds. By joining forces, 30% of the cases reached new significant diagnosis that combines genomics and understanding of the disease).

(3) "Pitch in". In desperate cases, the computational methods can be the last resort. (Example: a kid with acute IBD, who deteriorated and was on the verge of surgery to remove his intestines. By combined analysis of the kid's RNA and the Connectivity Map, a simple food supplement saved him).

(4) Become a leader: get funding and build your own team. (Story: a postdoc of his found a way to identify cases of domestic abuse from emergency clinic data, and built a group, got grants and established a career around it).

(5) There are hundreds of clinical studies (failed and successful) of pharma companies where large-scale data is publicly available. This is a grossly underused resource.

Of broad interest, especially those who interface with clinical questions.

Janet Kelso (MPI Leipzig): Neanderthals and modern humans.

Kelso was the keynote speaker in the VarI COSI. She gave an engaging talk about the state of the art study of ancient DNA. Topics covered were very diverse, including admixture between Neanderthals, Denisovans, and humans; the methods used to identify admixture; and the specific functional reasons to depletion and preservation of Neanderthal alleles in modern humans. Two vignettes: PheWAS analysis and comparison to UKBiobank showed significant differences in skin color phenotypes and in mood-related features. (Apparently our Neanderthal cousins were more moody and less social than us humans). Fascinating, non-technical stuff!

Of broad interest.

Gunnar Raetsch (ETH Zurich) – Representation Learning of patient health data

Raetsch was on the faculty of Memorial Sloan Kettering Hospital and moved three years ago to ETH. He gave the keynote in the MLSCB COSI. His talk was quite rich and I will cover only a fraction of it. His (and the community's) long-term vision is to combine as many data types as possible in order to create an up-to-date picture of an individual's health status. (Think about a "Doctor App" that one can probe on his cellphone daily at home, and get clues how to change behavior/diet to put one's health back on track, and not wait a week or more to visit
the physician whereby a serious problem emerges.) I did not write down the citations but it seems most studies were already published.

**Prediction of serious health events in ICU.** This project used data of all parameters collected on 54K patients hospitalized in an intensive care unit in Bern Hospital over about a decade. Some parameters were measured and recorded every five minutes, while others were measured when prescribed by the physician. The goal was to predict events of circulatory dysfunction eight hours ahead of their occurrence.

Out of 700 parameters collected, ~200 were used in the analysis. Imputation of missing values was done based on previous measurements, in different window sizes for measurements of different resolution. A variety of ML prediction algorithms was tried, and an algorithm called **LightGBM** based on boosted trees worked best (Deep Learning did not work so well). The system called circEWS is now being deployed. The method was validated on the **MIMIC** dataset (ICU data from Boston hospital, publicly available). As expected, one had to retrain the method to fit the MIMIC data. There was deterioration in the results on MIMIC, and the authors established that the reason was the lower resolution of the MIMIC data (1 hour vs. 5 mins in Bern). They also observed that the size of the training data had substantial effect on the performance, which leads one to hope that as data accumulates prediction will improve. (Raetsch also noted that cases where the doctor intervened in the 8-hr prediction window and thus prevented the event are treated now as false positive.)

**Unsupervised learning of patient’s health state.** The goal of this study was to represent a patient’s state based on all historical data on him/her in a low dimensional space. Previous methods tried to do this without factoring in the time dimension (e.g. DeepPatients – see paper in Scientific Reports). Here they used auto-encoder and in the low dimensional space combined a representation of RNN generalized to LSTM. Another extension vs. prior method was trying to predict not only current but also future state of the patient. Their method, called **Seq2Seq-Forecaster**, outperformed prior art and simpler auto-encoders. Testing was done on **Philips EICU dataset** (which apparently is also publicly available).

To come up with discrete patient states that are more valuable to the clinicians, the authors developed a method called **SOM-VAE**, which combined auto-encoder and SOM in the lower dimension (with 64 clusters) plus a Markov model for the time dimension. Some insights were visually demonstrated using this approach.

*Of special interest to Dan, plus Nimrod, David and Hagai for the ML.*

**Laura Furlong (Barcelona) – Gene-disease associations**

Furlong was another keynote in the Varl COSI. She started by describing the **DisGenNet** resource – an extensive catalog and network of gene variants in diseases based on 15 different resources and the literature. It covers both Mendelian and complex diseases and spans over 120k variants and 380k associations. Apparently it is accompanied by a PPI network. The main topic of her talk was studying topological features of the network in order to understand the connection between a particular gene’s topology and function. In addition to the well-studied local gene properties (e.g. degree) and global ones (e.g. betweenness centrality), she defined interesting meso-scale properties that depend on the gene’s larger neighborhood. Genes
were also clustered using **InfoMap**. She then defined six types of genes based on the combination of within-module degree and participation coefficient (a meso-scale parameter).

Analysis of the disease genes based on these six types identified enrichment of disease genes in some types - but deeper analysis showed that all the enrichment is due to cancer genes. The same thing happened when looking at complex diseases only. One conclusion was that the reports in the literature that disease genes have high centrality were due to the cancer genes. Finally, Furlong described analysis of the effect of mutations in genes for diseases that are autosomal recessive or dominant based on the data of the **EXAC project**, which contains 60K exomes.

The talk and analysis was rather descriptive, but I thought the new meso-scale network properties are interesting.

*Of special interest to: Hagai.*

---

**Julia Zeitlinger** (**Stowers Institute**) **BPNet for TFBS organization**

Zeitlinger is an experimentalist who collaborates with Anshul Kundaje, a ML expert from Stanford CS. She spoke in the RegSys COSI about using her experimental data to understand TF organization. She developed a technique called **ChIP-nexus** that identifies TFBSs at extremely high resolution (the method is an extension of the ChIP-exo method). Using the experimental data in windows of 1000 bases, a deep learning algorithm was developed and trained. It predicted the signal in accuracy equal to that between repeated experiments (!). They also developed a method called **DeepLift** for extracting meaning from the DL model by changing one base at a time and measuring the effect on the model (Shrikumar et al. arxiv 18). They showed the method is more accurate in predicting peaks than PWM-based methods.

The study focused on eight pluripotency TFs. In terms of results, the study found numerous transposable elements that contained functional binding sites. It also identified TF pairs that had preferable distance between them. Interestingly, the Nanog TF was found to have periodicity of 10.9bp, which fits one full cycle of the DNA helix, suggesting that it actually binds to the nucleosome. Another method for identifying TFBSs called **ChExMix** was also mentioned.

The overall method, called **BPNet**, was shown to work also based on ChIP-seq data, albeit giving lower resolution results than with ChIP-nexus.

*Of special interest to Tom and those interested in ML.*

---

**Krister Swenson** (**CNRS**) – **Large scale mammalian genome rearrangement and 3D structure**

This talk in the EvolCompGen COSI described joint work with Mathiew Blanchette (McGill). The idea was to check on a large scale if the breakpoints involved in genome rearrangements between species tend to happen between proximal genomic areas as measured in Hi-C data. The same type of analysis was performed more than five years ago by Sagot’s group, but it was then limited to intra-chromosomal contacts and was done on a small scale. The methodology here was based on sampling and moving around segments of the HiC map (not very clearly described; see Proceedings for details). The analysis reaffirmed the past conclusion that genome rearrangements are indeed affected by 3D proximity. Oddly enough,
analysis was done for HiC data from multiple human tissues, though only germline is relevant to species evolution. The analysis also used only human HiC data, though to do the analysis more accurately the other species (e.g. mouse) HiC should also be factored in.

*Of interest primarily to Lianrong and Nimrod.*

**Guillaume Marcais (CMU) – LSH for edit distance**

Marcais described work from Kingsford's group on a novel method that adapts the locality sensitive hashing (LSH) technique to edit distance. The MinHash algorithm uses the following trick due to Broder: Given a sequence, create all the k-mers in it and use a permutation of the k-mers to identify the smallest. Now repeat the process with m random permutations, and create an m-long sketch, i.e. the m minimizers. As was proven by Broder, the overlap in the sketches between two sequences divided by their union is an approximation of the Jaccard similarity between the sequences. While this is an elegant way to quickly evaluate the Jaccard distance, it does not fit edit distance between sequences, since (1) it ignores the multiplicities of k-mers and (2) it ignores the relative order of the k-mers in each sequence.

The suggested method, called **OMH** (Order MinHash), takes care of both problems. First, it keeps the location of each minimizer in the sequences. Second, in windows of length l (a parameter) minimizers are reordered by their locations. The authors prove that OMH is a LSH for edit distance. The result opens the question if more realistic sequence similarity measures can be developed. Also, OMH was not proven at this point to be an approximation of the edit distance like MinHash is.

This is an elegant piece of work, which connects well to minimizers and universal hitting sets. However, its practical value is unclear.

*Relevant mainly to David, Yael, Lianrong.*