Bioinformatics: living on the edge

Evangelos Bellos[1], Lachlan J M Coin[2,3] and Myrsini Kaforou[4]

Author Affiliations

1. **Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London W2 1PG, United Kingdom**
2. **Department of Genomics of Common Disease, School of Public Health, Imperial College London, London W12 0NN, United Kingdom**
3. **BGI-Shenzhen, Shenzhen, 518083, China**
4. **Section of Paediatrics and Wellcome Trust Centre for Clinical Tropical Medicine, Division of Infectious Diseases, Department of Medicine, Imperial College London, United Kingdom**

Can music be “decomposed” to singular notes on a piece of paper? Can life be fully understood by the function of genes in isolation? These are some of the questions that Denis Noble posed in one of the more philosophical keynote addresses of this year’s ECCB. The famed critic of genetic reductionism, in concert with renowned guitarist Christoph Denoth, set out to explore how synergy begets complexity both in art and in science. And in doing so Prof Noble captured a central theme of the conference: holism.

From gene prediction to comparative genomics to protein structure prediction, the early days of post-genomic bioinformatics research were arguably characterized by a reductionist approach to understanding life. Systems biology, on the other hand, started on the opposite end of the spectrum, attempting to directly model complex behaviour arising from biomolecular interplay, largely abstracted from the individual components. In this post-ENCODE era in which tremendous progress has been made on understanding the biomolecules individually, the focus now seems to have shifted towards a middle ground: leveraging the treasure trove of information about the nodes of the network to elucidate the intricate web of interactions represented by the network edges.

Perhaps surprisingly, next generation sequencing talks were conspicuously scarce given the sustained development of these technologies and their essential role in the ascendancy of genomics. However, in reality, genomics were indirectly pervasive throughout the conference, from keynotes to posters. Sequencing technologies appear to be maturing into stable tools that are ubiquitously used to obtain deep knowledge of the biomolecules themselves as well as the relationships between them. Like silent partners, genomics are providing both the starting point and the means for investigating networks, without stealing the spotlight.

Genetic interactions: within genomes

Systematic identification and characterization of genome-wide genetic interactions is one area that has already been transformed by the availability of sequence data for multiple species. Colm Ryan (University of California) compared experimentally derived epistasis maps of fission yeast *S.pombe* and budding yeast *S.cerevisiae* to show that the overall level of genetic cross-talk

between different biological processes tends to be more conserved than the underlying interactions. He further demonstrated how epistasis maps can be used to annotate uncharacterised genes using the observation that genes with similar genetic interaction profiles are often involved in the same functional pathway. Approaches like this may be used as a model for genetic interaction analysis in more complex organisms, in an effort to understand not only biological functions but also mechanisms that are believed to play major roles in disease. Philipp Bucher (École Polytechnique Fédérale de Lausanne) exploited lineage specific whole genome duplication events in teleost fish (including zebrafish and pufferfish lineages) to investigate the post-duplication fate of ultraconserved non-coding elements (UCNEs). These elements tend to be organized in large clusters around developmental genes and were found to be retained together in one of the two genome copies. This so-called “winner takes all” scenario suggests that UCNEs operate as part of a dense cooperativity network that could be responsible for their high levels of conservation.

Genetic interactions: between genomes

In a tour de force keynote focusing on the social organization of fire ants *Solenopsis invicta,* Laurent Keller (University of Lausanne) demonstrated that ant behaviour is determined not only by their own genome but by other genomes in the population. By way of example, Keller described a “selfish” allele in a 13.9Mb-long non-recombining region containing the odorant protein Gp-9, that regulates its own frequency by mediating altruistic and aggressive instincts. Michal Linial (The Hebrew University of Jerusalem) explored interactions between virus and host genomes from an evolutionary perspective. Linial demonstrated how both amino acid and codon usage of viral proteomes are adapted for maximizing translational efficiency in specific host-environments. Linial also showed that virus proteins acquired from the host tend to have fewer protein domains and shorter linker regions than their host counterparts. Understanding these evolutionary adaptations can prove essential for managing viral infections effectively.

Protein-RNA interactions

The recently published results of the ENCODE project established that three quarters of the human genome is capable of being transcribed. One approach to understanding the functional role of this pervasive transcription may be the identification of new RNA binding proteins, as presented by Cristoph Dieterich (Max Delbrück Center for Molecular Medicine). By purifying proteins which have been UV cross-linked to polyadenylated mRNA, followed by quantitative mass spectrometry, Dieterich identified 797 mRNA-bound proteins, one third of which were previously unknown and likely participate in post-transcriptional gene-regulation networks. Moreover, Dieterich used PAR-CLIP to sequence the mRNA binding sites of these proteins. Mihaela Zavolan (Biozentrum) also reported a novel application of PAR-CLIP to the Argonaute component of the RNA-induced silencing complex that was used to discover non-canonical miRNA-target interactions. Intriguingly, Zavolan showed how a single miRNA molecule (miR-294) can target components of the chromatin remodelling complex and prevent differentiation in embryonic stem cells, thus concluding that a handful of miRNAs may be sufficient to determine cell fate.

Protein-protein interactions

While protein-protein interaction (PPI) networks have been extremely valuable as a starting point for modelling cellular processes, it is well known that PPI databases have a high false-positive rate while omitting many genuine interactions. The gold-standard for the field, as described by Barry Honig (Columbia University), is a three-dimensional structure of a complex that incorporates the interacting proteins, and is available for less than 0.5% of all recorded PPIs. Honig presented a computationally efficient method for structurally superimposing homologous proteins onto known structural interactions in order to generate interaction models which are subsequently combined with co-expression and functional similarity information. This approach makes clever use of homology to known structures, however, as Chris Sander (Memorial Sloan Kettering) noted, only half of the well-characterized PFAM families have a known 3D structure for any of their members. To address this, Sander identified co-evolving residues from multiple sequence alignments in order to predict residue-residue proximity in the folded protein structures. Philip Kim (University of Toronto) illustrated the tissue-specific nature of PPI networks by monitoring how they are affected by alternative splicing events. Kim presented evidence that as much as a third of all tissue-specific alternatively spliced exons lead to differential PPI via both creation of new and destruction of existing interactions.

Protein-drug interactions

Chemogenomics is an emerging research area that attempts to identify novel drug-protein interactions by screening libraries of small molecules against the proteome using computational analyses of high-throughput data. Yoshihiro Yamanishi (Kyushu University) presented a method based on sparsity induced binary classifiers that incorporates data on the chemical structures of drugs, genomic information about target proteins and all known protein-ligand interactions across different protein families into a predictive model. The model can predict underlying interactions between drug chemical substructures and protein functional sites, which are involved in drug-target interaction networks. With a view to understanding and minimising drug side effects in patients, Sayaka Mizutani (Kyoto University) presented an approach for integrating drug side effect data with drug-protein binding information, using sparse canonical correlation. This technique was used to derive highly correlated sets of drug-targeted proteins and side effects, along with the drugs which drive their correlation. Such methods can play an important role in predicting potential side effects and their mediating pathways even at the drug design phase.

Conclusions

The opening keynote lecture of ECCB was delivered by Nobel laureate Aaron Ciechanover (Technion - Israel Institute of Technology) who described his magnum opus: unravelling the previously unknown cellular mechanism of protein degradation. Ciechanover outlined many aspects of ubiquitin-mediated protein degradation including the role of E3 ubiquitin ligases in ensuring high target specificity. He further highlighted the link between dysregulation of the ubiquitin-proteasome pathway and disease. Various levels of this pathway have been identified as drug targets to treat or suspend the progress of diseases like

myeloma, leukaemia and viral encephalitis. While this talk was inspiring, the fact that these discoveries were made without the assistance of high-throughput technologies or bioinformatics was sobering, and raises the question whether computational analyses of high-throughput data can lead to the elucidation of still unknown, fundamental cellular mechanisms. In other words, when will bioinformatics have its Nobel Prize moment? Talks presented at this conference suggest that by transforming information on complex biomolecular interactions at high cellular and temporal resolution into sequence data, bioinformaticians may be well on the way to achieving such a feat.