CRG Centre for Genomic Regulation

X CRG Annual Symposium
Computational Biology of Molecular Sequences
10-11 November 2011

NOTES DAY 2

The second day of the X CRG Annual Symposium on Computational Biology started this morning with **Gary Stormo**, one of the pioneers in protein-DNA modelling. With around thirty years of experience in this field, Stormo's talk summarized the top-four features in the evolution of DNA-protein modelling: consensus sequences, position weight matrices, higher-order models and progress in recognition models.
In the era where the use of the –Omics technologies is producing an enormous amount of different types of data, Stormo remarked the importance of developing specific models depending of the specific type of high-throughput data.

Ron Shamir focused on gene regulation, protein interaction and disease during his talk. Shamir explained in detail the algorithms that have been developed by his research group in two general areas: for finding regulatory motifs in genes and to establish the regulatory networks using expression data.
They have developed AMADEUS (A Motif Algorithm for Detecting Enrichment in mUltiple Species), a motif search algorithm based on hyper-geometric score and binned enrichment scores. Shamir's team found that AMADEUS is the more accurate and faster algorithm for discovering regulatory motifs in Metazoan, even when applied to recovering motifs from protein binding microarray data.
ALLEGRO is an expression model they have used to for example, study the yeast osmotic shock pathway, in which around 6000 genes participating and 133 conditions. They discovered multiple motifs with diverse expression patterns, even if the response is a small fraction of the conditions. It has also been applied to UT3'R analysis in human stem cells, with 14000 genes and 124 conditions (Biases in length/GC content of 3' UTRs).
Other tools (http://www.cs.tau.ac.il/~rshamir/) developed by Shamir's group are Matisse, a Modular Analysis for Topology of Interactions and Similarity Sets, which has been successfully applied to establish the bases of pluripotency in 150 human stem cell lines. Shamir also discussed Graph connectivity as a helpful paradigm for modelling, and a NICK kernel based method to integrate data.

In the fourth session of the Symposium **Ana Tramontano** revealed more surprises from the RNA world; specifically in the long non-coding RNA (lncRNA) universe, where a novel gene-expression regulatory circuitry has emerged.
In the 55% of the cases, the structure of proteins will be radically different from what we expect from the sequence analyses. Using structural analysis Tramontano has shown that with more than half-unverified RNA transcripts the picture is the same. **The analysis of lncRNA has revealed a novel regulatory circuitry in the cell**. In specific cases, lncRNAs compete with the transcription factor transcripts for microRNAs, blocking somehow the effect of the microRNAs on the regulation of the expression of some transcription factors. "The million-dollar question now is whether or not this is a general mechanism for gene-expression regulation", concluded Ana Tramontano.

In line with Tramontano's talk, **Peter F. Stadler** discussed the structural evolution of lncRNAs. By definition, lncRNAs are mRNA-like spliced and often polyadenylated RNAs. This includes microRNA precursors, snoRNA precursors, piRNA precursos, lincRNAs associating with protein complexes, and ceRNAs. However, there are other types of lncRNAs, such as those which are totally or partially intronic transcripts, independent UTRs, long unspliced RNAs, or macroRNAs that are hundreds of kilobases long transcribed regions.
Functionally, lncRNAs are involved in regulatory pathways, such as cell cycle, oncogenic and tumour suppressor pathways. In fact, genome-tiling arrays have revealed that differentially expressed segments are highly pathway specific and sometimes dependent on the genomic context.

Trying to answer the question of how well conserved are lncRNAs, Stadler gave us 2 answers: there is a relatively low degree of sequence constraint (Marques and Ponting 2009), but however there are some very well conserved examples (Chodroff et al 2010).

Tim Hubbard started his talk remembering how the human consensus human race was won by a public project concerned with open access data for all. But now we have moved to a managed access model of sequencing access (must be *bona fide* researcher), although the access is currently very limited.

Drop in sequencing prices and data storage are the starting points of genomic medicine, specifically towards personal genome sequence, cancer genome sequence and pathogen genome sequences. However, the clinical services to analyse whole genomes require a database of variants validated for clinical effect and health economics value. And also a network of open, federated national databases that is appropriate (national specifity). **But what is the economic impact of the Human Genome Project? What is the health economic value of having so much sequence information?** These are definitely questions to reflect on, to which **Søren Brunak**‘s added some others at the end of his talk. Brunak understands that for a fine-grained characterisation of diseases the integration of sequencing data and phenotypic data is essential. In Denmark, a repository of all electronic hospital records contains medical data of all hospitalised Danish citizens during the last 15 years. These records provide a large source of phenotypic data that allow the clustering of patient regarding of their phenotypes rather than their genotypes.

After mentioning a *problem with a difficult solution*, that is the $1,000 genome-sequence, but the $100,000 cost of analysing the genome, Brunak finished his talk lodging a petition to the audience: “**Add fine-grained information to biological information and approach the network biology not only from genotypes *but also* from phenotypes**”.

Terry Speed uses statistics like a drunk uses a lamp-post to stand, or so he said in the beginning of his talk. Library preparation, fragmentation, end-repair, adaptor ligation, size selection (melting) and PRC. What is the common characteristic of all these molecular biology techniques? Bias.

Bias is a generalised problem in analytical science, and genome sequencing is not an exception.

Bredan Frey gave a very interactive and interesting talk about splicing code. Talking about the regulatory codes for genome expression Frey summarised in a single phrase the change of paradigm about DNA structure and regulation:

“Exons contain the message; “non-coding” DNA contain the regulatory code”. Creating predictive regulatory codes that allow us to predict which RNAs would be expressed in a specific tissue or in a specific cell is the next challenge to fully understand genome regulation, and here the splicing code will have a lot to say.

In the last talk of the Symposium, **Mark Gerstein** focused on the **importance of the *dark matter of the genome*, the non-coding regions**. And here the numbers talk by themselves: only a 1,2% of the human genome corresponds to exons.