NOTES DAY 1

"*This century is the century of the human genome*". With this words **Luis Serrano**, director of the Centre for Genomic Regulation, welcomed the more than 250 assistants of the 10th CRG Annual Symposium on Computational Biology of Molecular Sequences. The organiser of the symposium and head of the Bioinformatics and Genomics programme, **Roderic Guigó**, presented the session about Protein Analysis after a brief introduction of the state-of-the-art of computational biology of molecular sequences.

Michele Vendruscolo talked about protein aggregation and folding in a very interesting talk entitled "Life on the edge". The amino acid sequence that encodes de free energy landscape of a protein also integrates information about the solubility of the protein. In the extremely dense cellular environment proteins are expressed at their critical concentrations. Therefore, small increases in protein expression might lead to aggregation and thus to disease. In fact, changes in the physico-chemical properties of proteins induced by changes in their amino acid sequence can decrease the solubility and induce aggregation of proteins that can lead to disease. **Eugene Koonin** reviewed the universals of genome evolution: ubiquitous powerlaws, scaling of functional classes of genes with genome size, the distribution of evolutionary rates across orthologous genes sets, among others. Koonin argued that simple models, such as modelling misfolding-drive protein evolution quantitatively reproduce universal distribution of evolution rates or the dependence of evolution rate on expression.

Nick Goldman highlighted in his talk "Adventures in Evolutionary Allignment" the problem of the progressive aligners that are generally used. Tradicional aligners make phylogenetically unrealistics gaps creating deletion hotspots and making insertions rare, implying sequences shrink over time. "This happens because they don't use phylogenetic info properly, but we can correct this to some extent", explained Goldman. Other problems associated to traditional aligners is related to the fact that insertions are not deletions: "insertions mistreated, denser sampling does not change it!", argued Goldman. Alignment is about sequence evolution: result should be phylogenetically meaningful and alignment affect downstream evolutionary distance, concluded Goldman. **Mathieu Blanchette** put on the table the ancestral genome reconstruccion problem, the algorithmic challenges in ancestral reconstruction and the applications of multiple alignment in annotating the human genome. In a very interesting talk Blanchette proposed a human computing framework for multiple alignment as the next challenge in computing for comparative genomics. Given that the human brain is probably the most powerful computing device ever invented, why don't we turn sequence alignment into a game? Phylo is an interactive whole-genome multiple alignment game that lets scientists and non-scientists internauts contribute to science.

In the first talk of the symposium, **Temple Smith** addressed the Homo sapiens Neanderthal mystery of Lgf5, a regulator of apoptosis found in both invertebrate and vertebrate organisms. The origin of the mammalian version of Lgf5 is not clear, as it seems evolutionary closer to an invertebrate family member that includes *Drosophila melanogaster*. "*Like you, I share a few genes with our cousin the Neanderthal, but there is an stop codon in the Lgf5 protein that is not in all of us*", said Smith. **Amos Bairoch** explored the modern human protein universe and presented the new resource for protein knowledge neXtProt. In the CALIPHO Computer Analysis and Laboratory Investigation of Proteins of Human Origine, Bairoch's research team focus on the about 5000 human proteins for which we lack functional knowledge: proteins similar to characterized proteins in distant organisms such as bacteria, plants or yeast but not validated in mammals, and 'Orphans', proteins with no similarity to any characterized protein, but that can be conserved across a more or less wide taxonomic space.

In the third session of the X Symposium on Computational Biology of Molecular Sequences, **Philipp Bucher** focused on the computational promoter analysis in the era of ultra-high-throughput sequencing. After introducing the ChIP-Seq technique and data analysis as well as other short read sequencing data call for a revision of the promoter concept, Buchers presented the new Eukaryotic Promoter Database (EPDnew), an attempt to improve promoter annotation using ChIP-Seq and related data. Launched in May 2011, EPDnew intends to give higher quality and coverage to the old annotated, non-redundant collection of eukaryotic Pol II promoters EPD database. EPDnew is based on mass genome annotation data: CAGE, ChIP-Seq. EPDnew is organism specific and is automatically compiled from archived data.

Alfonso Valencia, the last speaker of the day, brought to the conference room other questions in the protein universe. Trying to answer the questions 'what is in a cell? Proteins? Which proteins?' Valencia discussed the largely unknown process of alternative splicing at the protein level to finally conclude that nowadays "we don't know what is inside of a cell". How protein complexes are form, how proteins encounter each other and how cellular pathways form, are also obscure biological processes for which Science can still not provide answers.