Meeting report

Bioinformatics with a French accent

Laurence D Hurst* and Laurent Duret†

Addresses: *Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK. †Pole BioInformatique Lyonnais Laboratoire BBE - UMR CNRS 5558, Université Claude Bernard - Lyon 1, F-69622 Villeurbanne Cedex, France.

Correspondence: Laurence D Hurst. E-mail: l.d.hurst@bath.ac.uk. Laurent Duret. E-mail: duret@biomserv.univ-lyon1.fr

Published: 28 September 2005

Genome ***Biology*** 2005, **6:**349 (doi:10.1186/gb-2005-6-10-349)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/10/349

A report on the bioinformatics conference 'JOBIM: Journées ouvertes biologie informatiques mathématiques', Lyon, France, 6-8 July 2005.

Those with a fondness for Brazilian music will know of Antônio Carlos Jobim, best known to the rest of us as the composer of "*The Girl From Ipanema*". JOBIM is also the annual gathering of the French bioinformatics community. The title is no accident, the acronym being proposed by a French-Brazilian bioinformatician with a taste (we presume) for the bossa nova. This year JOBIM was held in Lyon and provided the opportunity to catch a cross-section of the challenges and opportunities opening up in the data-rich world of genomics and post-genomics.

Handling chips

For expression-related datasets the emphasis was very much on the challenges they pose. Although microarray data seem, in principle, a remarkable means to capture vast amounts of expression information, the quality of the data was of concern to many. Julie Aubert (L'Institut National Agronomique Paris-Grignon (INAPG), Paris, France) described the systematic and gene-specific dye-bias effects that have been observed in dual-color microarray experiments. The systematic effects may be normalized away but the gene-specific effects are more troublesome. She proposed a method (label bias index; LBI) for measuring such an effect and reported that after normalization this was the major source of experimental variability between replicates. The good news was that the artifact is not an inevitability, although why it is present only sometimes is not well understood. For the less computer-minded, the plethora of biases and the statistical techniques for handling them may seem daunting. In this regard, the development of Goulphar, reported by Sophie Lemoine (Ecole Normale Supérieure, Paris, France), may be just what is needed. This is a relatively user-friendly and versatile tool for the diagnosis and normalization of microarray data (available at the Goulphar website [http://www.transcriptome.ens.fr/goulphar]). Its inputs are a GenePix output file and a text file with the normalization parameters; the text file can be made at the Goulphar website using a simple web-based form.

First- and second-tier annotation

For those workers whose interests center on the sequence rather than on its expression, the current major challenges might be what one could describe as second-tier annotation, where the first tier is annotation directed at identifying protein-coding genes. Methods for first-tier annotation were not ignored, however. Sarah Djebali (Ecole Normale Supérieure, Paris, France) presented a new method of annotation that attempts to automate processes, such as the prediction of complete genes, that are usually left to manual curators. She compared the new method, known as Exogean, with eight other annotation methods, using manually annotated human genes as a benchmark. Exogean gave a significant improvement in the quality of prediction of complete genes. It also has the advantage of reporting all the details of the decision process, thus providing full traceability.

Some protein-coding transcripts will almost certainly be missed by this and most other methods. Michaël Bekaert (Université Paris Sud, Paris, France) considered one such class, namely proteins resulting from -1 frameshifting, a recoding process seen in some genomes. In such a situation, an mRNA has two potential translations, one of them starting one base pair out of the frame relative to the other. Annotation software will typically consider the frameshifted reading frame to be evidence either of a pseudogene or a sequencing error. Most such recoding events have been described in viruses, although a very small number of eukaryotic nuclear genes are known to function in such a manner. Might we have overlooked the majority of them, however, as we were not too sure what to look for? To this end Bekaeart described a method for scanning the yeast genome for potential candidates, in a manner that makes no assumptions about the mechanism of frameshifting. From a total of 189 candidate regions, 58 were further scrutinized. Of these, 28 expressed a full-length mRNA covering both predicted open reading frames (ORFs). This not only indicates a much higher prevalence of -1 frameshifting than was previously imagined to occur in eukaryotes, but also helps to explain why they have been missed. The previous methods for detecting them assumed a particular mechanism that appears not to apply.

On a broader scale, Marie Touchon (CNRS, Gif-sur-Yvette, France) examined well-described origins of replication in the human genome and showed that around these there are striking skews in the relative usage of G compared with C and A compared with T, much as there are associated with transcription. This permitted her to subsequently identify 1,000 replication origins. The position of replication origin sites appears to be a conserved feature in mammals. Possibly more contentious was an attempt to classify the human genome into isochores, broad-scale regions of the genome of more than 300 kb of approximately homogeneous GC content. Christelle Melo de Lima (Université Claude Bernard, Lyon, France) presented a new approach in which, instead of just relying on the genomic G and C content, the properties of genes are also taken into account. Genes in GC-rich regions tend, for example, to have small introns and small intergene spacers. Applying such a method under the assumption that only three classes of isochore exist (high, medium, and low GC content), she concluded that human chromosomes fit fairly well with an isochore organization for all chromosomes. It is unclear, however, how the method treats regions of the genome with gradually increasing GC content, such as those found as one goes towards the telomeres.

The issue attracting the most attention was the problem of identifying regulatory motifs. As discussed by Jacques van Helden (Université Libre de Bruxelles, Brussels, Belgium) in an invited lecture, pattern-discovery methods may have some utility in yeast and bacteria, but their application in human and mouse seems to be very poor, although one problem is the lack of a good benchmarking dataset. Along with pattern searching, conservation of residues can also provide good clues and underpins the phylogenetic footprinting method for finding regulatory sites. Fabrice Touzain (Lorraine Laboratory of Computer Science Research and Applications (LORIA), Villers-Lès-Nancy, France) showed, for example, how such a method could identify binding sites for sigma factor in bacterial DNA. One issue, however, is deciding which two or more species to consider. If they are too closely related there may be a weak signal; if they are too distant then some signal may be lost, possibly owing to changes in the transcription factors themselves. Rekin's Janky (Université Libre de Bruxelles, Belgium) described a new method for circumventing this problem. The strategy is to follow a phylogenetic tree and at each branch to apply a pattern-discovery method - dyad analysis. This looks for over-represented dyads by comparison with a background model. Applying this method to well-described control elements associated with the SOS-responding gene *lex4*, he showed that the method recovers known structures, but also that the motif diverges along parts of the phylogeny (notably the Firmicutes and Actinobacteria).

A bottom-up approach to the same problem of motif recognition and prediction was discussed in the invited lecture given by Richard Lavery (Institut de Biologie Physico-Chimique (IBPC), Paris, France). He considered the physics of proteins as they bind to DNA and asked where the relevant interaction forces come from. The only experimental input required for his calculations was an atomic structure of the protein specifically bound to a fragment of DNA. He reviewed detailed studies of hydrogen bonding that have revealed that, while interactions between specific side chains and bases certainly exist, the majority of hydrogen bonds involve the DNA backbones. Consequently, overall DNA conformation within a complex is also important, and sequence-dependent changes in DNA structure, or in its mechanical or dynamic properties, can also play a role in recognition. His energy-based method was able to predict consensus sequences, when they exist, with impressive ability.

Making and using trees

The method used by Janky presupposes that we have reliable phylogenetic trees. Intriguingly, there was little debate about methods for the analysis of alignments to construct such trees, although Alexis Criscuolo (L'Institut Supérieur de l'Entreprise de Montpellier (ISEM), Montpellier, France) proposed a new distance method for the construction of supertrees - trees in which not all taxa are represented by all genes in the alignment. Instead, there was a strong focus on steps in phylogeny construction before tree estimation from a given alignment. For example, Jean-François Dufayard (Université Claude Bernard, Lyon, France) detailed a new multiple alignment method that uses the phylogeny to aid the alignment. The important role of taxonomic and gene sampling was also highlighted. Simonetta Gribaldo (Université Paris 6, Paris, France), for example, examined the problem of archaeal phylogeny. She argued that genes of the transcriptional and translational systems are good choices for phylogenetic analysis as they tend not to be horizontally transferred and hence provide a good signal. Analyzing the proteins of these two systems separately, she showed that the two largely agreed, and support the dichotomy of Crenarchaeota and Euryarchaeota. An important point was that wider taxonomic sampling increases the convergence of the two phylogenies.

For some speakers, however, phylogenetics was not a means to decide which species are derived from which others, but rather to ask about the history of genes. Alexandra Calteau (Université Claude Bernard, Lyon, France) showed, for example, that two hyperthermophilic bacteria (*Aquifex aeolicus* and *Thermotoga maritima*) obtained many of their genes by horizontal transfer from archaea. Most notably, hyperthermophiles are unique in possessing reverse gyrase. The copy of this gene in the hyperthermophilic bacteria appears to have been derived from archaea, suggesting only one evolutionary origin for this unusual gene.

A comparable methodology was employed by invited speaker Martin Huynen (University of Nijmegen, The Netherlands) to investigate the early evolution of the mitochondrion. Assuming mitochondria were once α-proteobacteria, he asked whether one might identify genes in the eukaryotic nuclear genome that were likely to be derived from the primitive mitochondrion. From the set of proteins so identified he could identify aspects of the metabolism of the first mitochondria. Apart from the fact that the citric acid cycle was probably incomplete, the most striking conclusions concern what the mitochondria/α-proteobacteria were doing for their host. Nowadays mitochondria export ATP, but this would be strange behavior for any bacterium and was not supported by analysis of the ancestral gene set. Others have suggested that hydrogen was the main export, but no evidence for a hydrogen export metabolism was found. Indeed, the only export channel seen is one for iron-sulfur clusters. Although quite why an early eukaryote would need iron-sulphur clusters remains a matter for speculation, the results spoke highly of the role of bioinformatics in generating and testing hypotheses as much as in providing tools for analysis.

The concentration at this year's meeting on tools and methods over analysis reflects more the nature of the submission process (full papers had to be provided before the conference for selection) than the overall strengths of French bioinformatics. Although familiar in the computer sciences, such a procedure does not suit those from, for example, an evolutionary background. For next year's meeting in Bordeaux, the submission process will change and the flavor of the science will no doubt change with it. One idiosyncrasy will remain, namely that presentations will for the most part be given in French. *Plus ça change, plus c'est la même chose*.