Meeting report

Genome regulation in the (so-called) post-genomic era

Mikhail S Gelfand

Address: State Scientific Center GosNIIGenetika, Moscow, 113545 Russia. E-mail: misha@imb.imb.ac.ru

Published: 10 November 2000

*Genome **Biology*** 2000, **1(5)**:reports4023.1–4023.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/5/reports/4023

A report on the Second International Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, 7-11 August, 2000.

The Second International Conference BGRS 2000 was mainly concerned with databases and recognition algorithms for transcription regulation sites, prediction of RNA secondary structure and its regulatory role, analysis of the three-dimensional structure of transcription factors and protein-DNA interactions.

The opening talk from Jim Fickett (SmithKline Beecham Pharmaceuticals, Philadelphia, USA) outlined the use of techniques such as phylogenetic footprinting and analysis of composite regulatory elements for finding regulatory regions involved in muscle-specific transcription. More than 95% of known muscle-specific regulatory sites reside within regions conserved between human and mouse. Analysis of phylogenetic footprinting patterns followed by a search for sequence similarity signals in clusters of co-expressed genes identified in expression array studies decreases the statistical noise and thus improves the performance of local multiple alignment algorithms. Also, because transcriptional regulation in eukaryotes is performed by cooperatively acting factors, simultaneous searching for multiple sites sharply decreases the number of false positives. The conclusion, however, was that although the methods such as these are already useful for providing clues for experimental studies, they are clearly insufficient for reliable annotation.

The modularity of regulatory elements was discussed in detail by Alexander Kel (Novosibirsk Institute of Cytology and Genetics, Novosibirsk, Russia), who presented the ‘fuzzy puzzle’ model of eukaryotic transcriptional regulation. Given that it is well known that a single site can be bound by different factors, that the binding of many factors is cooperative, and that transcription factors become fully structured only upon binding to DNA, Kel and colleagues argue that the seeming lack of strong sequence constraints in many eukaryotic transcription regulation sites is a natural consequence of the flexibility of the regulation machinery. Kel’s group have used this model to define binding sites for two transcription factors: the nuclear factor of activated T cells (NFAT) in the promoters of T-cell-specific genes, and elongation factor E2F in promoters of cell-cycle specific genes. As in Fickett’s study, it turned out that searching for composite elements instead of single sites greatly improved the specificity of predictions.

A very elegant study merging such seemingly disparate areas of computational biology as the analysis of single-nucleotide polymorphisms (SNPs) and the prediction of transcription regulation sites was presented by Tatiana Merkulova (Novosibirsk Institute of Cytology and Genetics, Novosibirsk, Russia). To explain the psychiatric disorder phenotypes caused by single-nucleotide mutations in intron 6 of the human tryptophan oxygenase gene (*TDO2*), Merkulova’s group identified candidate transcription-factor binding sites overlapping the mutated positions. The only transcription factor whose predicted affinities to the mutated sites explained the outcome of a mobility-shift gel-electrophoresis assay was YY-1, a multifunctional nuclear-matrix-associated protein that represses or stimulates gene expression, depending on the context. The prediction was confirmed by a gel-shift assay in the presence of anti-YY-1. The authors plan to apply this approach to other non-coding single-nucleotide polymorphisms with distinct phenotypes.

Although the majority of talks were concerned with eukaryotic genomes, considerable attention was also directed towards prokaryotes. My own report described a comparative approach to the recognition of transcription-factor binding sites. It is based on the following assumption: sets of co-regulated genes (regulons) are conserved in different genomes that contain orthologous transcription factors. Thus, when looking for candidate binding sites for a particular transcription factor, the presence of the same site occurring upstream of several orthologous genes is an indication that it is a true binding site, whereas false positives are scattered at random in the genome. This consistency check sharply increases the specificity of predictions, although it may lose species-specific members of regulons. This technique not only allows the transfer of data on regulatory interactions from well studied genomes to newly sequenced ones, but also makes it possible to find new members of old regulons and even describe regulons *de novo*. Julio Collado-Vides (Universidad Nacional Autonoma de Mexico, Cuernavaca, Mexico) described results of systematic study of transcription factors in *Escherichia coli* and other bacteria. The helix-turn-helix (HTH) factors were grouped into 20 families, and it was suggested that the repressors with the HTH motif at their amino termini share a common origin, whereas the LysR-related proteins that have dual action (repression of their own gene and activation of other genes) are not related to this superfamily, although they also possess an amino-terminal HTH motif.

These two studies potentially lead to an interesting situation: annotation of a prokaryotic genome can identify a number of regulons and, independently, a number of transcription factor genes. Thus, there emerges the problem of matching the transcription factors to the sets of binding sites. One way is positional analysis: transcription factor genes are often located within the same operons as the metabolic genes that they regulate, and thus regulate themselves. If this situation is found in at least one of a group of related genomes, the orthologous transcription factors in all other genomes are thus matched to a metabolic pathway and hence a regulon. On the other hand, as noted by Monica Riley (Marine Biological Laboratory, Woods Hole, USA), this raises a question about where transcription factors (and other proteins involved in information flow rather than metabolic pathways - for example, protein kinases) and transporters should be placed in general proteome classification schemes, as functionally they belong to multiple categories: for example, the lactose repressor belongs both to the sugar metabolism and transcriptional regulation categories.

Another way to match regulators and DNA sites is to study the general features of the protein-DNA interactions that, according to the report by Akinori Sarai (RIKEN Tsukuba Institute, Tsukuba, Japan), can be used to predict DNA targets of transcription factors. Conformations of Cα atoms around DNA bases were studied using a sample of known protein-DNA complex structures. This led to derivation of an empirical free energy potential for the interactions between bases and amino acids. It turned out that this procedure is sufficiently sensitive to allow for discrimination between cognate sites and random DNA sequences. In a similar vein, merging of sequence and structural features for recognition of *E. coli* promoters was discussed by Olga Ozoline (Institute of Cell Biophysics, Pushchino, Russia). She described sequence periodicities in promoter regions leading to possible structural deformations of DNA upon formation of the transcription initiation complex, and presented experimental results proving that these deformations modulate the activity of the RNA polymerase in the promoter T7D.

The comparative approach can be used not only for the recognition of regulatory sites, but also for gene recognition. Two talks described different approaches to this problem. The SGP program [http://www.soft.ice.mpg.de/sgp-1] presented by Thomas Wiehe (Max Planck Institute of Chemical Ecology, Jena, Germany) starts with alignment of nucleotide sequences and then constructs chains of conserved exons. It is fast and can be applied to long sequence fragments, but nucleotide alignment and the assumption of conservation of the exon-intron structure make it, like other similar programs such as Rosetta [http://plover.lcs.mit.edu], inapplicable outside vertebrates. On the other hand, the program Pro-Gen described by Andrey Mironov (State Scientific Center GosNIIGenetika, Moscow, Russia) performs alignment on the amino-acid level and does not require conservation of the exon junctions, and thus can be used for more distant comparisons, but is slower. Many sequencing projects do not involve genomic DNA, but are done on the expressed sequence tag (EST) level. A poster by Oleg Vishnevsky (Novosibirsk Institute of Cytology and Genetics, Novosibirsk, Russia) presented the program ORFScan, which finds the optimal reading frame in an alignment of EST sequences using two criteria: low number of gaps and high coding potential.

A large fraction of talks, a majority of posters, and a round-table discussion were dedicated to databases of eukaryotic transcription sites, regulatory regions, and transcription factors (Table 1; note the resources listed in Table 1 are underlined below and the corresponding URLs are linked online). Edgar Wingender (Biobase Biological Databases GmbH, Germany) described the database TRANSFAC, which stores information about eukaryotic transcription factors, their binding sites, and recognition rules in the form of sequence patterns and positional weight matrices. A similar database, TRRD, has been developed by the team of Nikolay Kolchanov (Novosibirsk Institute of Cytology and Genetics, Novosibirsk, Russia). Several other databases that are descendants of these two were described in other talks and many posters. These include COMPEL (Olga Kel-Margoulis, Novosibirsk Institute of Cytology and Genetics, Novosibirsk, Russia), PathoDB (Manuela Pruess, Biobase Biological Databases GmbH, Germany), and numerous TRRD spinoffs dedicated to specific regulatory systems (see Table 1).

An important issue that emerged was how diverse regulatory interactions taking place in a eukaryotic cell can be formalized, for example in databases, in descriptions of regulatory networks using discrete models and differential equations, and so on. The problem of formalization was also addressed

Table 1

Databases of transcriptional regulatory regions derived from TRRD

Database	URL
TRANSFAC	http://transfac.gbf.de
PathoDB	No web site yet available
TRRD	http://www.bionet.nsc.ru/trrd/
Databases derived from TRRD	
COMPEL	http://compel.bionet.nsc.ru/
Cell-cycle-specific genes	http://wwwmgs.bionet.nsc.ru/mgs/papers/kel_ov/celcyc/
Erythroid-specific genes	http://wwwmgs.bionet.nsc.ru/mgs/papers/podkolodnaya/esg-trrd/
Steroidogenesis-controlling genes	http://www.bionet.nsc.ru/trrd/papers/ignatiera/es-trrd/
Plant genes	http://wwwmgs.bionet.nsc.ru/mgs/papers/goryachkovsky/plant-trrd/
SELEX data	http://wwwmgs.bionet.nsc.ru/mgs/systems/Selex/
Data from large-scale experiments on site activity	http://wwwmgs.bionet.nsc.ru/mgs/systems/activity/

in a number of talks and posters dedicated to the analysis of regulatory networks, for example by John Reinitz (Mount Sinai School of Medicine, New York, USA). He also presented a theoretical model describing interactions of *Drosophila* homeobox genes, which adequately describes phenotypes of at least some null mutations. At present, however, these studies cover only transcriptional regulation and not other kinds of regulatory interactions. Blastoderm gene expression in *Drosophila* is thus an ideal system for testing the approach, as in this case neither protein degradation, differential intercellular transport, nor other developmental processes influence the concentrations of proteins. Reinitz also described implementation of wavelet image processing used for registration of experimental data, which allows one to eliminate systematic and random distortions of images.

The emerging trends in functional mapping of DNA sequences thus seem to be: using comparative genome analysis in various flavors; taking the diversity of the underlying phenomena into account, which leads to the simultaneous use of diverse techniques for each particular problem; and maintaining close links to experimental results at all stages of research.