Meeting report

Understanding the language of gene regulation

Wynand Alkema* and Wyeth W Wasserman†

Addresses: *Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden. †Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, Canada.

Correspondence: Wyeth W Wasserman. E-mail: wyeth@cmmt.ubc.ca

Published: 18 June 2003

Genome ***Biology*** 2003, **4:**327

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/7/327

A report on the Cold Spring Harbor Laboratory meeting 'Systems Biology: genomic approaches to transcriptional regulation', Cold Spring Harbor, USA, 6-9 March 2003.

On the snow-covered coast of Long Island, the community of researchers dedicated to understanding how DNA sequences selectively activate gene transcription gathered to assess progress in the field, to celebrate recent successes and to plot future directions. The theoretical foundations of the field were established in the 1980s. In laboratory studies, the concept of regulatory modules was established, introducing the idea that transcription-factor binding sites are grouped in functional clusters in regulatory sequences. In bioinformatics, the first predictive models were introduced for the identification of potential binding sites for well-characterized transcription factors and 'phylogenetic footprinting' was formally introduced for the identification of regulatory sequences conserved between orthologous genes. The recent influx of researchers into the field is a testament to the opportunities presented by emerging genomic resources such as large-scale expression data, transcription-factor binding data and full genome sequences of multiple eukaryotic genomes. For researchers pursuing studies in diverse model organisms (such as yeast, worm, fly, vertebrates, and *Arabidopsis*), the clear message of the conference was that the early theoretical ideas are broadly applicable. The presentations centered on three main themes: quantitative descriptions of protein-DNA interactions, prediction and characterization of clusters of transcription-factor binding sites, and the analysis of co-regulated systems of genes.

Describing protein-DNA interactions

Efforts to study gene regulation are founded on the determination of how transcription factors and DNA interact. Modeling of the DNA-binding preferences of transcription factors has so far mainly used single-order positional weight matrices - that is, matrices of the bases preferred at each position of a binding site assuming that the nucleotide observed at one position is independent of the nucleotide found at any other position. Recent published reports in which large collections of transcription-factor binding sites were generated and analyzed indicate that this underlying assumption is false. In a retrospective analysis of the collections, Gary Stormo (Washington University, St. Louis, USA) explained that the underlying assumption is adequate in most cases, because inclusion of positional correlations gives only a marginal improvement in the specificity of binding-site predictions. Stormo gave an inspiring call for researchers to use the techniques for generating large sets of transcription-factor binding sites to define quantitatively the target-nucleotide preferences of amino acids that directly interact with DNA. Using *in vitro* binding data for zinc-finger transcription factors, Stormo generated a quantitative matrix profile for the prediction of amino-acid:base interactions. The current shift in focus from the analysis of target sequences to the protein-DNA interface was reflected in posters from Barry Honig's lab (Columbia University, New York, USA) presenting methods for predicting the binding properties of uncharacterized transcription factors using the refined structures of DNA-bound factors from the same structural class.

A particular constraint on the analysis of regulatory sequences is the sparse data available on the binding preferences of transcription factors. A practical approach towards the analysis of protein-DNA interaction was presented by Martha Bulyk (Harvard Medical School, Boston, USA). She used 'protein binding microarrays', in which phage-displayed transcription factors are bound directly to microarrays of double-stranded DNA. Quantitative data on the level of binding of each protein to each spotted sequence were used to determine the binding-site specificity for

zinc-finger transcription factors directly from differences in fluorescence intensity. The resulting matrix describing the observed binding preferences of each protein provides better specificity in predicting suitable transcription-factor binding sites than previous models. In one of the highlights of the conference, Rick Young (Whitehead Institute, Massachusetts Institute of Technology, Boston, USA) introduced results from a high-throughput screening procedure to identify binding sites for yeast transcription factors. The experiments used crosslinked chromatin immunoprecipitation of transcription factors followed by microarray analysis ('ChIP on chip'), and the results obtained define sets of genes containing promoter sequences that are bound by the transcription factors tested. The new approach to the genome-scale characterization of transcription-factor binding properties has influenced research in the field enormously.

Prediction of *cis*-regulatory elements

The study of composite response elements - or regulatory modules - dominated the presentations on regulatory regions in metazoan genomes. Because the ability of current approaches to predict isolated transcription-factor binding sites is poor (as was widely noted during the conference), various groups presented novel methods - and experimental assessments of published methods - for the detection of clusters of binding sites. Most algorithms involve counting the occurrences of predicted binding sites for user-defined sets of transcription factors; the algorithms differed in the procedures for counting and how the significance of predictions was assessed. Susan Celniker (Lawrence Berkeley National Laboratory, Berkeley, USA) and Marc Halfon (Howard Hughes Medical Institute and Brigham and Women's Hospital, Boston, USA) identified functional regulatory regions in *Drosophila melanogaster* by counting instances of binding sites for transcription factors involved in regulation of the *even-skipped* (*eve*) gene, which itself encodes a transcription factor important in developmental patterning. Both used a phylogenetic-footprinting step, incorporating comparison with the nearly complete genome of *Drosophila pseudoobscura*, to prioritize their predictions for experimental testing. Some of the regions identified by these methods showed regulatory activity, but further analysis of wider samples of predictions indicated that the overall performance of the methods was low. On this point, Halfon remarked that these clustering methods may be better suited to identifying true regulatory regions in the very early stages of embryonic development than clusters that are functional in later stages of development.

Many speakers touched on the idea of using phylogenetic footprinting to increase the specificity of algorithms that predict regulatory sequences. The broad applicability of phylogenetic footprinting was supported by promising results in species ranging from bacteria to multicellular eukaryotes. Dario Boffelli (Lawrence Berkeley National Laboratory, Berkeley, USA) introduced the concept of 'phylogenetic shadowing', in which multiple sequence comparisons are made between orthologous genes across short evolutionary distances, taking relationships into account. Applying this method across a set of closely related primates, he demonstrated that it can reveal functional regulatory sequences. In contrast to these closely related species, Elliott Margulies (National Human Genome Research Institute, National Institutes of Health, Bethesda, USA) analyzed a set of orthologous sequences from a spectrum of vertebrates, from human to zebrafish. He introduced a scoring function for the analysis of multi-species conserved sequences (MCS). In calculating the MCS score, the contribution of the sequence from each non-human species is weighted according to the percentage of identical nucleotides observed in alignments of syntenic, neutrally-evolving sequences from the species and humans. It appeared that the scores for coding MCSs were significantly higher than the scores for non-coding MCSs. Using a threshold heuristically determined to separate the two classes, he predicted potential regulatory regions in a portion of human chromosome 7.

Phylogenetic footprinting alone is not sufficient to define regulatory regions, however. Eric Siggia (Rockefeller University, New York, USA) demonstrated that when phylogenetic footprinting is used to locate known regulatory regions in *D. pseudoobscura* and *Drosophila virilis,* only 50% of the known regulatory regions fall within sequences that are conserved between these species. Furthermore, several groups doing human-rodent comparisons reported a failure to detect the anticipated functions in well-conserved regions. The combination of multi-species phylogenetic footprinting with robust models of composite response elements holds the most promise for unraveling the complex regulatory mechanisms governing transcription.

Modeling regulatory systems

Two distinct approaches were introduced for the study of gene interactions in transcriptional regulation. The first, the construction and study of simple artificial regulatory systems in *Escherichia coli*, was presented by Stanislas Leibler (Rockefeller University, New York, USA) and Kenzie MacIsaac (University of Toronto, Canada). Regulatory circuits were created by coupling well-characterized inducible promoters (such as those controlled by IPTG or arabinose) to repressor proteins. Leibler emphasized that there is rarely a one-to-one relationship between the observed phenotypic characteristics of a system and hypotheses about the underlying genetic regulatory circuit: in the absence of detailed measurements, even three-gene systems can lead to results open to multiple interpretations. In short, the analysis of large gene networks cannot be conclusive with current data.

At the other extreme of complexity, several groups presented a second approach: models for the entire regulatory network

of yeast. David Gifford (Massachusetts Institute of Technology, Boston, USA) described the modeling of functional gene modules (sets of co-regulated genes) using Young's ChIP-on-chip data supplemented by gene-expression data. The genes in the identified modules contained similar promoter sequences, and their expression profiles correlated significantly. Subsequent comparison of the expression of genes in the modules with the transcription factors regulating them meant that the factors could be classified into activators and repressors. The idea of using co-expression data as a tool to define regulatory regions was also the basis of the closing keynote talk by Stuart Kim (Stanford University Medical Center, USA). He has analyzed data from human, fly, worm and yeast microarrays to identify groups of genes that are co-expressed in more than one species.

As well as the use of phylogenetic footprinting in the detection of regulatory regions in genomic sequences, the same technique has been applied to detect evolutionarily conserved regulatory networks in yeast (Saeed Tavazoie, Princeton University, USA) and bacteria (poster presented by W.A.). Tavazoie demonstrated that the false-positive rate of binding-site predictions derived from yeast gene-expression data can be reduced if the data are filtered by analyzing the conservation of networks across related organisms.

Innovative genomic methods to probe transcriptional regulation have helped to fulfil the promise of the techniques established in the early years of bioinformatics - phylogenetic footprinting, transcription-factor binding-site profiling and the identification of regulatory modules of binding sites. Algorithms are now emerging that can reveal critical information about the regulatory mechanisms governing expression of sets of genes. It was apparent from many presentations, however, that one challenge for the short term is to produce reliable reference collections of transcription-factor binding sites that can be used for the training and benchmarking of methods for the analysis of regulatory sequences. The current lack of such reference data results in an over-reliance on anecdotal evidence to justify methods: a surprising proportion of the methods presented at the meeting were justified by observations that a selected portion of the results agree with information found in the biological literature.

Taken together, the impressive results shown at this meeting raise optimism for the future. Investigators may now wish to venture into new challenging areas: for instance, despite success in the analysis of yeast regulatory sequences, attempts to find the control sequences for co-expressed sets of human genes have rarely been fruitful. Alternatively, researchers may wish to respond to Stormo's challenge to decipher the amino-acid:nucleotide interaction code, or they may venture into the study of chromatin. The combination of new data resources and algorithmic advances is fueling real and meaningful progress in making sense of the mechanisms governing gene expression.