Computational approaches to the discovery of regulatory elements in noncoding DNA

Michael Koldobskiy

MB&B 452a
December 13, 2002

INTRODUCTION

Biological research in the post-genomic era has been charged with the formidable task of assigning cellular functions to thousands of gene products. Large-scale gene deletion *(1)* and protein structure determination *(2)* efforts are underway, and creative methods for global characterization of protein activities, such as protein microarray technology *(3)*, have been devised. Thus, much progress is being made in annotating the protein-coding regions of genomes. However, a more complete understanding of the genome also requires the annotation of noncoding DNA, and in particular the identification of regulatory elements responsible for transcriptional control of gene expression. Noncoding regulatory elements serve as binding sites for specific transcription factors, and thus dictate which genes are expressed, at which times, and in which cells. This mechanism is therefore responsible for cellular differentiation, function, and response to stimuli *(4)*. Discovery of regulatory motifs by traditional genetic and biochemical methods is quite laborious, requiring the construction of a series of deletions in the noncoding region upstream of a gene, and testing each one for effects on gene expression. Fortunately, bioinformatic analysis can greatly simplify this process by predicting likely regulatory elements.

STRATEGIES

1. Using microarray-based expression profiles

A. Overview

One strategy for identifying unknown regulatory elements in gene promoters benefits from microarray-based expression profiling *(5, 6)*. An expression profile is determined by quantifying relative levels of mRNA transcripts as time (or another experimental variable) changes *(7-9)*. Genes with similar expression profiles can then be clustered *(10 and references therein)*. Since mRNA quantitation directly reflects transcriptional activity, genes that are transcriptionally coregulated should cluster together. Similarities in the upstream sequences of clustered genes can then be identified as potential transcription factor binding sites.

Ohler and Niemann report that in practice identifying the conserved regulatory motif is not so simple: the motif is of unknown size, it might not be well conserved between the various promoters analyzed, the promoter sequence used for the analysis might not be complete, and the microarray clustering algorithm may produce results that are not representative of *in vivo* coregulation *(11)*. Nevertheless, the approach has been successful in identifying yeast promoters *(5)*. In higher eukaryotes, the situation is more complicated since regulatory elements may be dispersed over very large distances, and identification of the transcription start site is itself a challenge *(12)*.

B. Identification of conserved motifs

A local multiple alignment algorithm capable of detecting subtle similarities can be used to compare the promoters of putatively coregulated genes. A highly effective approach is Gibbs sampling *(13)*. Briefly, given N sequences, it seeks to find a pattern of a given width W within each sequence. One sequence, z, is chosen at random or in a specified order. For a segment of width W starting at a random position in all sequences excluding z, the pattern description (probabilistic model of residue frequencies at each position in the selected segments) and background frequencies (probabilistic model of residue frequencies at positions not in the selected segments) are calculated. Every possible segment of width W in z is scored according to the ratio of the probability that it was generated by the current pattern probabilities over the probability that it was generated by the background probabilities. The algorithm continues iteratively, such that once a correct segment is picked by chance in one sequence, the process will tend to recruit further correct segments. Notably, the Gibbs sampling algorithm is able to find an optimized local alignment model for N sequences in N-linear time.

Another alignment approach based on a statistical algorithm is Multiple Expectation-maximization for Motif Elicitation, or MEME *(14, 15)*. The expectation-maximization (EM) algorithm is used to fit a statistical model to each input sequence; for each motif, MEME maximizes a likelihood function that balances width of the motif, accuracy of the match, and the number of sequences in the data set that exhibit a match. The algorithm determines the optimal width of the motif, whether the motif occurs in all sequences or a subset, and whether it occurs multiple times or once per sequence. Having identified a motif, MEME can search for more motifs sequentially, excluding previously identified motifs from subsequent searches. MEME is therefore capable of identifying a motif containing a gap, but would consider it to be two separate motifs identified in different runs of the algorithm.

A modified Gibbs sampling algorithm is also able to determine the width of the motif, and the number of copies of a motif in a sequence (16). Additionally, recent work has shown that representing the background residue frequency by a higher-order Markov process improves detection of regulatory elements by the Gibbs sampling approach. The authors suggest using an independent data set of intergenic regions for each organism to establish a background model whose quality does not depend on the input sequences for a given analysis *(16, 17)*.

A different class of algorithms examines oligomers of a certain length and reports those that occur more often than the background promoter sequence composition *(18, 19)*. These methods yield a list of over-represented oligomers, rather than a weight matrix model of the motifs as do Gibbs sampling and MEME *(11)*.

2. Using cross-species comparisons

An alternative strategy, independent of microarray analysis, relies on comparisons of noncoding DNA between different species – for example, human and mouse (20, 21). The noncoding regions upstream of homologous genes in several species are compared, and used as a guide for discovering regulatory elements. Conserved noncoding stretches are likely to contain regulatory elements; comparisons of these sequences (using the methods described above, for instance) can identify these elements.

OUTLOOK

Bioinformatics can revolutionize the field of gene regulation by predicting regulatory elements *in silico*. Two independent approaches are available, one utilizing whole-genome expression profiles, and the other relying on conservation of important regulatory regions among homologues. The two methods may be used in a complementary fashion: cross-species comparisons can be used to demarcate possible regulatory motifs (or conserved regions likely to contain them) either before or after microarray clustering and promoter sequence comparison. Such an integrated approach would provide a check on the quality of results, as significant regulatory motifs should be detected by both methods. It would also simplify the task of promoter recognition in higher eukaryotes by distinguishing between junk and non-junk portions of noncoding regions.

In addition to contributing to the prediction of regulatory element location, microarray data can be used to predict promoter function. As a result of the microarray experiment, the putatively coregulated gene clusters are known to be affected by a given stimulus or set of experimental conditions. This functional information is only limited by creativity in the design of microarray experiments, and becomes more complex as advances in microarray clustering and analysis are made; notably, Qian *et al.* report an approach for clustering of time-shifted and inverted gene expression profiles *(22)*. An accompanying challenge to using microarray data for the functional annotation of promoters is the development of appropriate database methods.

Some have suggested that the detection of regulatory elements would benefit from a joint modeling of DNA physical structure along with sequence *(23)*. The rationale for this approach is that transcription factor proteins recognize regions of DNA with specific conformational, bendability, and protein-induced deformability properties (*24, 25*). Along the same lines, separate algorithms could be devised to find particular classes of transcription factors with specific binding properties. Thus, a synergistic relationship between structural biology and bioinformatics would be created: biochemical and biophysical characterization of transcription factors and their binding sites would lead to the creation of improved algorithms for detecting novel regulatory elements, which (upon further analysis) would in turn contribute to our biochemical and biophysical understanding. An important finding demonstrating the value of additional knowledge about noncoding DNA is that, in vertebrates, gene promoters are often found near CpG islands; this information has been successfully exploited for large-scale mapping of human promoters *(26)*.

Finally, computational prediction of regulatory elements must be verified in the laboratory. Reporter genes can be placed under control of the putative control element and used in various functional assays *in vitro* or *in vivo* in a transgenic system *(27)*. An important application of the discovery of novel regulatory elements would be the discovery of novel transcription factors. Armed with a knowledge of the transcription factor's target DNA sequence and a reporter gene construct for assaying activity, DNA affinity chromatography methods could potentially be employed to isolate the desired transcription factor.

Overall, progress in the field of gene regulation has been (and will continue to be) greatly accelerated via a partnership with bioinformatics.

REFERENCES

1. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387-91 (2002).
2. Brenner, S. E. A tour of structural genomics. *Nat. Rev. Genet.* **10**: 801-9 (2001).
3. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**: 2101-5 (2001).
4. Purves, W. K., Sadava, D., Orians, G. H., & Heller, H. C. *Life: The Science of Biology.* Sunderland, MA: Sinauer (2001).
5. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**: 939-46.
6. Bucher, P. Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* **9**: 400-7.
7. De Risi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680-6 (1997).
8. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., & Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**: 3273-97 (1998).
9. Soukas, A., Cohen, P., Socci, N. D. & Friedman, J. M. Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev.* **14**: 963-80 (2000).
10. Gerstein, M. & Jansen, R. The current excitement in bioinformatics – analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Opin. Struct. Biol.* **10**: 574-84 (2000).
11. Ohler, U. & Niemann, H. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet*. **17**: 56-60 (2001).
12. Down, T. A. & Hubbard, T. J. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**: 458-61 (2002).
13. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**: 208-14 (1993).
14. Bailey, T. L. & Elkan, C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21**: 51-83 (1995).
15. Bailey, T. L. & Elkan, C. An artificial intelligence approach to motif discovery in protein sequences: application to steroid dehydrogenases. *J. Steroid Biochem. Molec. Biol.* **62**: 29-44 (1997).
16. Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. & Moreau, Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.* **9**: 447-64 (2002).
17. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P. & Moreau, Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**: 1113-22 (2001).
18. Van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by cimputational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**: 827-42 (1998).
19. Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. Predicting gene regulatory elements *in silico* on a genomic scale. *Grenome Res.* **8**: 1202-15 (1998).
20. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet*. **16**: 369-72 (2000).
21. Wassermann, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26**: 225-8 (2000).
22. Qian J., Dolled-Filhart M., Lin J., Yu H., Gerstein M. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol.* **314**: 1053-66 (2001).
23. Ohler, U., Niemann, H., Liao, G. & Rubin, G. M. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* **17**: S199-S206 (2001).
24. Voet, D. & Voet, J. G. *Biochemistry*. New York: Wiley (1995).
25. Liao, G. C., Rehm, E. J. & Rubin, G. M. Insertion site preferences of the P transposable element in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **97**: 3347-51 (2000).
26. Ioshikhes, I. P. & Zhang, M. Q. Large-scale human promoter mapping using CpG islands. *Nat. Genet.* **26**: 61-3 (2000).
27. Sumiyama, K., Irvine, S. Q., Stock, D. W., Weiss, K. M., Kawasaki, K., Shimizu, N., Shashikant, C. S., Miller, W. & Ruddle, F. H. Genomic structure and functional control of the Dlx3-7 bigene cluster. *Proc. Natl. Acad. Sci. U. S. A.* **99**: 780-5 (2002).