Levers and fulcrums: progress in *cis*-regulatory motif models

Ewan Birney

A new set of computational methods provide association of transcription factor binding motifs to gene function.

In contrast to our models for the function of protein-coding sequence, our understanding of noncoding DNA function is very limited. In an article in this issue of *Nature Methods*, Warner *et al.* present a new computational route to identify noncoding DNAs that exert regulatory function[1].

With protein-coding genes we have a good biochemical model of how the DNA sequence is transcribed, spliced and translated to protein, which allows us to generate a reasonable computational model of this process. Collectively, we have extensive experience on how to compare such genes and assign preliminary functions. The result has been a rich set of genomic annotation of protein-coding genes, in which new experimental evidence of protein existence or function can be readily transferred to other genomes.

In contrast, for noncoding DNA, despite many useful models to explain molecular biology experiments, such as the concepts of 'promoter', 'enhancer', 'silencer' and others, these experimental models have not lead to robust computational models of noncoding DNA function. Without robust discovery methods for these elements, much of the downstream work—such as comparing regions or assigning more detailed function—is impossible. In short, for protein-coding DNA we are some way into the mountain range of understanding, but for noncoding DNA we are still in the foothills of the problem and have not even established a sensible route into the task.

Warner *et al.* developed two *in silico* methods, Lever and PhylCRM (pronounced 'fulcrum' to resonate with 'lever'), that make a new set of assumptions for noncoding DNA[1]. The main assumptions—which are different from those in many other programs—are that there are, or shortly will be, reasonably complete 'dictionaries' of motifs (that is, the biochemical binding sites of transcription factors), and that the binding sites in such dictionaries are accurate enough to not require refinement.

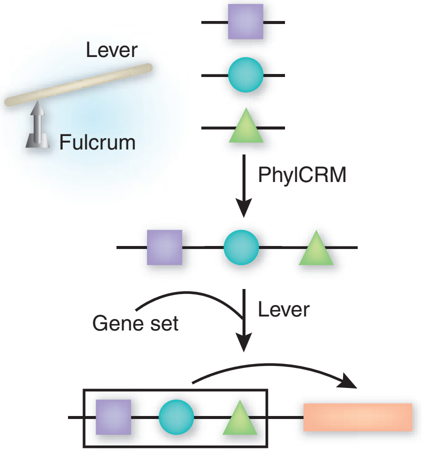

Figure 1 | Working together like a fulcrum and a lever, two new algorithms allow the identification of overrepresented *cis*-regulatory modules and their assignment to gene sets.

Previously much of the computational work has focused on generating such dictionaries, but there has both been steady progress on assembling such dictionaries computationally[2–4], and more importantly, direct high-throughput biochemical methods are coming on line, both *in vivo*[5] and *in vitro*—for example, high-throughput DNA protein arrays[6]. These *in vitro, in vivo,* and *in silico* methods are being developed to complement each other, and it is no surprise that these new computational methods come from one of the groups that is leading efforts to discover transcription factor binding *in vitro*.

PhylCRM is a *cis*-regulatory module (CRM) assessment process on an aligned region of genome; it relies on the MONKEY program[7] to place the motifs, but then assesses different combinations of motifs (factor x AND factor y) by generating a single scoring scheme that represents the presence of a proposed *cis*-regulatory module—that is, a collection of motifs. An empirical background distribution of all PhylCRM results is used to assess whether a particular *cis*-regulatory module is significantly high scoring. Lever then takes these results and applies them to a relatively standard enrichment analysis, in which a category of genes is assessed for overrepresentation of a particular *cis*-regulatory module (**Fig. 1**).

In this work the researchers used genes that are up- or downregulated in muscle development and are present in specific Gene Ontology (GO) categories. One can debate whether each of these steps is optimal: one could imagine different motif placement methods, perhaps more like the branch length score method used in the *Drosophila melanogaster* analysis[2], which does not make the assumption of a completely correct alignment in assessing whether the placement of the motif is conserved or not. Alternatively, different-scoring schemes for motif combinations could be applied, but this is the first end-to-end pipeline that works on the assumption of a (relatively) complete dictionary and outputs associations of *cis*-regulatory modules rather than just individual motifs to functions.

Formally the association is not to genes, but to a classification of gene sets, although the association does provide an immediate set of gene-specific hypotheses for experimental follow-up. Perhaps equally important is that this assignment process can work with anonymous transcription factor binding profiles and assign a first putative function to such factors.

Warner *et al.*[1] then provide a series of case studies in muscle development, which they follow up experimentally. They take six cases in which they applied a variety of experimental methods: quantitative PCR, chromatin immunopreciptation, luciferase assays and short hairpin RNA knockdowns of the putative binding factors. For all cases there was at least one experimental method showing a positive result.

Ewan Birney is at The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.
e-mail: birney@ebi.ac.uk

Although six case studies is not a formal statistical assessment of the method, it is clear that the method predicts results worth investing experimental efforts in.

Warner *et al.*[1] have provided a new set of computational methods in *cis* regulation, but perhaps more importantly, they have started attacking the problem with a different set of assumptions than is normally used. I expect that many other groups will be moving toward making these assumptions of a relatively complete and accurate dictionary of motifs and combine these genome-wide statistical models with more biophysically oriented models, such as those from the Segal lab[8]. These biophysical motivated models provide a more complete description of the binding of transcription factors, including weak sites, which are important in binding, in particular when cooperative effects are considered. When coupled with genome-wide *in vivo* models, such as those coming from the ENCODE consortium[5], and model organisms, such as *Drosophila*[9,10], there are grounds for optimism in making serious progress in our understanding of *cis*-regulatory DNA.

1. Warner, J.B. *et al. Nat. Methods* **5**, 347–353 (2008).
2. Stark, A. *et al. Nature* **450**, 219–232 (2007).
3. Ettwiller, L. *et al. Genome Biol.* **6**, R104 (2005).
4. Down, T.A., Bergman, C.M., Su, J. & Hubbard, T.J. *PLoS Comput. Biol.* **3**, e7 (2007).
5. ENCODE Project Consortium. *Nature* **447**, 799–816 (2007).
6. Bulyk, M.L., Huang, X., Choo, Y. & Church, G.M. *Proc. Natl. Acad. Sci. USA* **98**, 7158–7163 (2001).
7. Moses, A.M., Chiang, D.Y., Pollard, D.A., Iyer, V.N. & Eisen, M.B. *Genome Biol.* **5**, R98 (2004).
8. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. *Nature* **451**, 535–540 (2008).
9. Li, X.Y. *et al. PLoS Biol.* **6**, e27 (2008).
10. Jakobsen, J.S. *et al. Genes Dev.* **21**, 2448–2460 (2007).