PUBLISHED IN ASSOCIATION WITH
COLD SPRING HARBOR LABORATORY PRESS

Computational identification of regulatory DNAs underlying animal development

Dmitri Papatsenko & Michael Levine

Department of Molecular and Cellular Biology, Division of Genetics & Development, University of California, Berkeley, California 94720, USA. Correspondence should be addressed to D.P. (dxp@berkeley.edu) or M.L. (mlevine@berkeley.edu).

Whole-genome sequence assemblies provide a rich resource for the *in silico* identification and characterization of regulatory DNAs, particularly enhancers, in different animal groups, including *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*. There are two major methods for the recognition of regulatory DNAs within complex genome assemblies[1]: (i) clustering of combinations of sequence motifs that correspond to known binding sites for defined transcription factors and (ii) phylogenetic analyses that identify sequence conservation in noncoding regions among two or more related species. We describe here the first method—clusters of binding sites for multiple transcription factors; the second method is described in the accompanying protocol[2]. Clustering methods require extensive prior knowledge of the binding preferences of known transcription factors. In ideal cases, the process under study relies on the activities of two or more well-defined transcription factors. Even in such optimal cases, however, there is a high incidence of false positives. A 'hit rate' of 30–50% is the limit of precision that can be obtained with these methods[3]. Nonetheless, they are considerably more efficient than the identification of enhancers via 'blind' functional assays, whereby random genomic DNA fragments near or within a given gene are analyzed for regulatory activities. Clustering methods have been used to identify many new enhancers engaged in common developmental processes, permitting the construction of genomic regulatory networks[4–7]. Clustering methods were first developed nearly a decade ago[8–11]; however, there is still no 'best' technique or 'universal' software (comparable to BLAST, a fast alignment search tool). Instead, new techniques are constantly being developed, and some even merge clustering methods with phylogenetic analysis. Despite the flourishing diversity of methods, most use common strategies that we describe here in a sequence of steps, providing a format for the identification of functionally related enhancers and coregulated genes in animal genomes.

MATERIALS

Any internet browser, running on a computer with access to the internet
Online tools for identifying enhancer sequences (**Table 1**)
Online tools for genomic analysis (see **Table 2** for a listing of programs referred to throughout the protocol)

PROCEDURE

Preparation of the sequence data and the motifs

1 | Prepare the sequence data. You can use coregulated enhancers or coregulated genes to identify potential DNA-binding sites as shared sequence motifs. If you have a set of enhancer or promoter sequences (training set), but no known binding motifs, you can proceed directly to Step 2 (motif extraction). If you launch your analysis with a set of coregulated genes and have no information about binding sites or enhancers, extract up to 500 bases upstream of the transcription start site (for example, using the Database of Transcriptional Start Sites (DBTSS)) or the first exon and then proceed to Step 2.

It is now possible to search entire genomes online. Most eukaryotic promoters can be retrieved automatically from dedicated databases, such as Eukaryotic Promoter Database (EPD), Human Promoter Database or Database of Orthologous Promoters (DoOP).

2| Prepare the motifs. If you start from a set of coregulated genes (from a microarray experiment), or with a set of enhancer or promoter sequences, you will need to extract putative binding motifs. Some of the best available resources are motif sampler MEME or Gibbs Motif Sampler run with default parameters. If you start from a set of separate binding sites, align them using MEME and save the alignment (motif) for your search.

➨ **TROUBLESHOOTING**

3| Select a program to scan your test sequences (**Table 1**). Your choice will depend on your goals: whether you plan to scan an entire genome online or just a set of sequences, and whether you plan to use advanced grammatical models (site *A* + site *B* separated by *N* bases) or site densities (see Step 4). We focused our current survey on practical aspects of online applications, ignoring their theoretical backgrounds, algorithm specifics and benchmarking tests.

Derivation and assessment of the recognition model

4| Construct your recognition models. A recognition model may include a single type of binding motif or a combination of different binding motifs. Derive a model from your training sequences, such as shared motifs among a set of coregulated enhancers. Depending on the program selected, you will have to specify either a motif combination (*A* + *B* + *C*) and a distance range (presumed cluster size, ~500 bases) or include advanced grammar. For advanced grammatical models, specify the distances between individual binding sites or define a model using site density, that is, the minimal number of sites in a window of a fixed width (for example, 10 sites/kilobase (kb)). To determine the distances between binding sites or distance ranges, identify binding site matches in your training sequences using motif-finding programs such as Possum.

Table 1 | Practical features of major enhancer-finding online tools.

Program	Recognition models	Genome-wide online scan	Custom input sequences	Phylogenetic filtering	Annotated output	URL
Cis-Analyst	Site density	Dm	N.A.	Dm-Dp	Plot, alignment	http://rana.lbl.gov/cis-analyst/cgi/viewer.php
Enhancer	Advanced grammar	Ag, Am, At, Ce, Ci, Cs, Dm, Fr	N.A.	N.A.	Plot, table	http://opengenomics.org/ http://cigbrowser.berkeley.edu/superenhancer.html
Cluster-buster	Site density	N.A.	<0.1 Mb	N.A.	Plot, table	http://zlab.bu.edu/cluster-buster/cbust.html
Target Explorer	Grammar	Ag, Dm	<1 Mb	Dm-Dp	Advanced options	http://trantor.bioc.columbia.edu/Target_Explorer/
AHAB	Site density	N.A.	<1 Mb	Dm-Dp	Plot, Table	http://gaspard.bio.nyu.edu/Ahab.html
rVISTA	Grammar	N.A.	<0.02 Mb	Advanced options	Plot, alignment	http://genome.lbl.gov/vista/rvista/submit.shtml
ConSite	Grammar	N.A.	<0.1 Mb	Hs-Mm	Plot, alignment	http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/
TFBScluster	Grammar	Hs, Mm, Rn	<10 Mb	Hs-Mm-Rn	Table, annotations	http://hscl.cimr.cam.ac.uk/genomic_tools.html

Genomes available for online scans abbreviated as follows: Dm, *D. melanogaster*; Dp, *Drosophila pseudoobscura*; Ag, *Anopheles gambiae*; Am, *Apis mellifera*; Ce, *C. elegans*; Ci, *Ciona intestinalis*; Cs, *Ciona savignyi*; Fr, *Fugu rubripes*; Hs, *Homo sapiens*; Mm, *M. musculus*; Rn, *Rattus norvegicus*; At, *Arabidopsis thaliana*. N.A., not available.

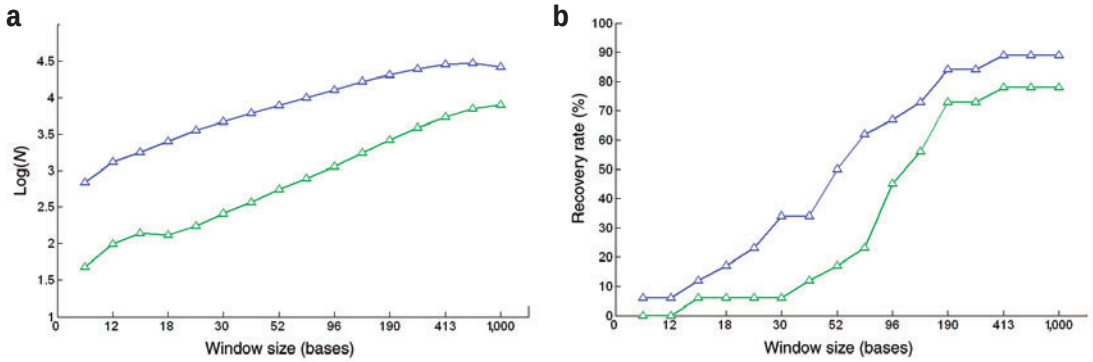

Figure 1 | Effect of phylogenetic footprinting on model sensitivity. (**a**) Dependence of logarithm of the number of hits in genome, log(*N*), on window size before and after phylogenetic filtering. Search model included two binding sites for Dorsal separated by 10–1,000 bases. Genome-wide scans were performed using *D. melanogaster* genome alone (blue) or a combination of *D. melanogaster* and *D. pseudoobscura* genomes (green). (**b**) Recovery rates (sensitivity) for the same genome-wide searches, calculated using control set of 18 enhancer sequences, containing functional Dorsal binding sites[15]. Despite considerable gain in selectivity (three- to sixfold), some functional sequences were lost after phylogenetic filtering.

Unfortunately, at present there are no convenient online tools to generate the recognition models automatically from a set of binding motifs and a set of training sequences. Interested users may try to download and run command-line applications such as Cluster-Trainer, Mtf-dist or motif sampler Co-Bind.
➨ **TROUBLESHOOTING**

Performing scans and annotating results

5| Estimate the selectivity and sensitivity of your models. If you base your recognition models on site distributions observed in your training set, you must assess model sensitivity (or rate of false negatives) by determining how many authentic enhancers are lost in the search. The simplest way to estimate the false-negative rate is to search your training sequences with the recognition models. If you plan to scan long sequences (1 Mb or more) or entire genomes, you may obtain a very large number of hits unless you control model selectivity (or rate of false positives). It is essential to maximize the stringency of your models, being careful not to eliminate known enhancers. Adding more types of binding sites, decreasing window size and increasing cutoff values will increase stringency of your models. The changes in the estimated false-positive and false-negative rates with progressively increasing model stringency (decreasing window size) are plotted in **Figure 1**.
You may choose different strategies by shifting the focus of your models toward reducing false positives (by increasing model stringency) or reducing false negatives (by relaxing your models). If the genome under study is well-annotated (for example, the D. melanogaster genome), and there are independent, efficient ways of filtering false positives (see Step 8), you may choose to focus your computational search on reducing false-negative rates and filter false positives using genome annotations afterwards.

6| Before running the programs to identify new enhancers, you will need to select parameter values and options. The input parameters may vary from one program to another. For instance, cutoff values for individual site matches might be expressed using match probability ($P = 0.0003$; as in the program Cis-Analyst) or informational score ($I = 5.5$; as in the program Cluster-buster). The dependence between *P* and *I* might be different for different binding motifs (see **Fig. 2**). Lower *P* values or higher *I* values will increase your search stringency. In the case of a string search, you can control the search stringency by adding ambiguous positions into your motif.
If you have already built and trained your recognition models (Steps 4 and 5), the choice of parameters should not present problems. The default values that are suggested by the selected programs might not apply if you are using custom motifs.

7| Annotate the results. New site clusters or motif combinations identified in whole-genome surveys (putative *cis*-regulatory modules; pCRMs) must be placed in the context of the existing genome annotations. Obviously, there is greater confidence in pCRMs that are located near genes engaged in the process under study. Most of the known regulatory modules are located within 10–20 kb from transcription start sites (either 5' or 3' of the start site). But for small genomes with 'dense' gene distribution (nematodes and insects) the expectation can be reduced to just several kilobases. In larger 'sparse' genomes (vertebrates) regulatory regions can be located much farther away from transcription start sites, as far as 10^5–10^6 bases[12]. Regardless of whether you scan an entire genome or just a genomic fragment, your annotation must include (i) sequence of the identified hit, (ii) position of the hit in the genome, (iii) names and accession numbers of genes within the annotated range, (iv) relative position of the hit within gene structure (3', 5', intron, exon) and (v) distances between the hit and transcription start sites of genes within the annotated range (see the program Target Explorer).

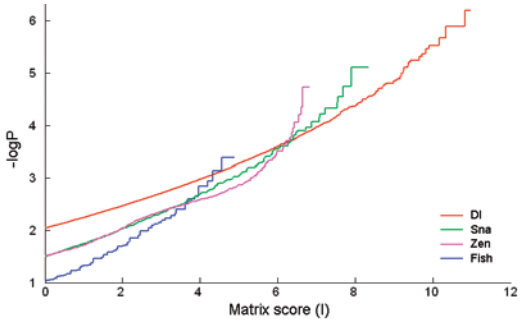

Figure 2 | Dependence of cumulative match *P* value from informational score (or weighted matrix score) (*I*). The same value of *I* may result in different match *P* values for different motifs. Among the reasons affecting the dependence are binding motif length, quality of the motif alignment and AT/GC content of the motif. The longest Dorsal motif (D1; consensus KGGAWTTYCC) produces matches with probability up to 10^{-6}; the shortest D/Fish motif (Fish; consensus TWYAAW) matches with probability up to ~4×10^{-4}.

8| Filter false positives and reveal most prominent candidates. Any genome-wide search is likely to produce a long list of pCRMs. Precision of recognition based on the 'search with a signal' methods (as well as any other prediction technique) gives no guarantee that any of the candidate sequences are functional. Use available annotation resources or databases to perform independent cross-validation of your candidate sequences. First, attention must be paid to gene density in the genomic region. It has been shown, for instance, that developmental genes of *D. melanogaster* and their vast *cis*-regulatory modules are frequently found within 'sparse' regions[13]. Within dense regions (>1 gene/2–3 kb) it is difficult to select the candidate target gene for the identified pCRM. Conversely, hits that are found 'too far' from any gene (>20–50 kb) should be given secondary consideration.

Unless you have already performed phylogenetic filtering described elsewhere[2,14], explore the conservation level of your candidate sequences using the UCSC or VISTA browsers. Typically, enhancer sequences are more conserved over short evolutionary distances (up to 50–100 million years) than other noncoding DNAs[6]. Retrieve functional information for all genes within your annotation range (~10 kb). In the case of the D. melanogaster genome there are several databases containing gene expression profiles (see The Drosophila Developmental Gene Expression Timecourse), expression patterns (see Patterns of gene expression in Drosophila embryogenesis) or gene interactions (see The GRID). These resources, listed in ***Table 2****, facilitate the identification of authentic enhancers and associated target genes.*

Table 2 | Related online tools for genome analysis.

Program	Authors	URL
Major databases for transcription regulatory regions		
Database of Transcriptional Start Sites (DBTSS)	Suzuki, Y., Yamashita, R., Sugano, S. & Nakai, K.	http://dbtss.hgc.jp/index.html
The Eukaryotic Promoter Database (EPD)	Schmid, C.D., Praz, V., Delorenzi, M., Perier, R. & Bucher, P.	http://www.epd.isb-sib.ch/index.html
Human Promoter Database (HPD)	Davuluri, R.V., Suzuki, Y., Sugano, S. & Zhang, M.Q.	http://zlab.bu.edu/~mfrith/HPD.html
Databases of Orthologous Promoters (DoOP).	Barta, E. *et al.*	http://doop.abc.hu/
***De novo* motif discovery and motif search tools**		
Multiple Em for Motif Elicitation (MEME)	Bailey, T., Elkan, C., Grundy, B. & Gribskov, M.	http://meme.sdsc.edu/meme/website/meme.html
The Gibbs Motif Sampler	Thompson, W., Rouchka, E.C. & Lawrence, C.E.	http://bayesweb.wadsworth.org/gibbs/gibbs.html
Possum	Fu, Y., Frith, M.C., Haverty, P.M. & Weng, Z.	http://zlab.bu.edu/~mfrith/possum/
Major phylogenetic analysis tools		
UCSC Browser	Karolchik, D. *et al.*	http://genome.ucsc.edu/cgi-bin/hgGateway
VISTA Browser	Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. & Dubchak, I.	http://genome.lbl.gov/vista/index.shtml
Other related databases and resources		
The GRID	Tyers, M., Breitkreutz, B., Jorgensen, P. & Breitkreutz, A.	http://biodata.mshri.on.ca/grid/servlet/Index
Drosophila Developmental Gene Expression Timecourse	Stolc, V. *et al.*	http://genome.med.yale.edu/Lifecycle/
Patterns of gene expression in *Drosophila* embryogenesis	Tomancak, P. *et al.*	http://www.fruitfly.org/cgi-bin/ex/insitu.pl
Toucan 2	Aerts, S. *et al.*	http://www.esat.kuleuven.ac.be/~saerts/software/toucan.php#tut
CREME	Sharan, R., Ben-Hur, A., Loots, G.G. & Ovcharenko, I.	http://creme.dcode.org/
Cluster-Trainer	Frith, M. & Weng, Z.	http://zlab.bu.edu/cluster-buster/ctrain.html
Co-Bind	GuhaThakurta, D. & Stormo, G.D.	http://ural.wustl.edu/~dg/co-bind.html
Mtf-dist	Papatsenko, D.	http://webdisk.berkeley.edu/~dap5/data_01/tools.html
ModelInspector (Genomatix)	Frech, K., Danescu-Mayer, J. & Werner, T.	http://www.genomatix.de/products/index.html

TROUBLESHOOTING TABLE

PROBLEM	SOLUTION
Step 2 Extracted putative motifs may be imprecise and even different from those actually present in your training set.	Motif extraction will work reliably only if you have one or more instances of a binding site in a majority of your sequences. Shorten your training sequences. Including long sequences (more than 500–1,000 bases) will increase noise level and will decrease the quality of the extracted motifs. Consider only the few best-scoring putative motifs. Perform motif extraction using orthologous sequences and compare the extracted motifs.
Step 4 Not every application will accept your custom recognition model.	Some programs have advanced grammar options; others, instead, focus on site density or motif combination (see **Table 1**). We advise you to explore specifics of the selected program first, and then proceed to the model construction. Steps 3–6 may require iterative model adjustment.

COMMENTS

We describe here only open-source online software that is easy to handle and requires no special skills to carry out, even with genome-wide computational screens. Programs performing similar analyses are also available as Java-based desktop applications (Toucan) or as commercial software suites. Commercial programs typically have interfaces that are more convenient and offer a better choice of options; however, they are quite expensive even with a limited (single-site, 1-year) subscription option. In addition to the identification of pCRMs using known or putative binding motifs, it is also possible to begin exploration from gene expression profiles . The CREME program uses this simple strategy, requiring as input a set of potentially coregulated human genes (after clustering genes by their expression profiles) and a database of binding motifs for human transcription factors, and identifies *cis*-regulatory modules in the promoter regions of genes in the set. Further development of enhancer recognition methods will likely lead to further integration of data types and analysis tools. That software will store and automatically process different kinds of data: motif combinations (clusters) in genome and functional data, including gene expression profiles and phylogenetic data. But custom computational surveys based on specific biological information will represent the dominant demands for the near future.

SOURCE

This protocol was contributed directly by the authors listed on the first page. For additional information describing the use of databases and search programs, please see Mount, D.W. *Bioinformatics: Sequence and Genome Analysis,* 2nd edn. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2004).

1. Ohler, U. & Niemann, H. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.* **17**, 56–60 (2001).
2. Bejerano, G., Siepel, A.C., Kent, W.J. & Haussler, D. Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nat. Methods* **2**, 535–545 (2005).
3. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).
4. Markstein, M., Markstein, P., Markstein, V. & Levine, M.S. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* **99**, 763–768 (2002).
5. Lifanov, A.P., Makeev, V.J., Nazina, A.G. & Papatsenko, D.A. Homotypic regulatory clusters in *Drosophila*. *Genome Res.* **13**, 579–588 (2003).
6. Schroeder, M.D. et al. Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* **2**, E271 (2004).
7. Berman, B.P. et al. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* **5**, R61 (2004).
8. Kondrakhin, Y.V., Kel, A.E., Kolchanov, N.A., Romashchenko, A.G. & Milanesi, L. Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Appl. Biosci.* **11**, 477–488 (1995).
9. Crowley, E.M., Roeder, K. & Bina, M. A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.* **268**, 8–14 (1997).
10. Wagner, A. A computational genomics approach to the identification of gene networks. *Nucleic Acids Res.* **25**, 3594–3604 (1997).
11. Pickert, L., Reuter, I., Klawonn, F. & Wingender, E. Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics* **14**, 244–251 (1998).
12. Ovcharenko, I. et al. Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137–145 (2005).
13. Nelson, C.E., Hersh, B.M. & Carroll, S.B. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5**, R25 (2004).
14. Kent WJ, Hsu F, Karolchik D, Kuhn RM, Clawson H, Trumbower H, & Haussler D. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* **15**, 737–41 (2005).
15. Stathopoulos, A. & Levine, M. Whole-genome analysis of *Drosophila* gastrulation. *Curr. Opin. Genet. Dev.* **14**, 477–84 (2004).