D68–D73 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj075

cisRED: a database system for genome-scale computational discovery of regulatory elements

G. Robertson*, M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O. L. Griffith, X. Zhang, Y. Pan, M. Hassel, M. C. Sleumer, W. Pan, E. D. Pleasance, M. Chuang, H. Hao, Y. Y. Li, N. Robertson, C. Fjell, B. Li, S. B. Montgomery, T. Astakhova, J. Zhou[1], J. Sander[1], A. S. Siddiqui and S. J. M. Jones

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada and [1]Department of Computing Science, University of Alberta, Edmonton, AB, Canada

Received August 15, 2005; Revised and Accepted October 8, 2005

ABSTRACT

We describe cisRED, a database for conserved regulatory elements that are identified and ranked by a genome-scale computational system (www.cisred.org). The database and high-throughput predictive pipeline are designed to address diverse target genomes in the context of rapidly evolving data resources and tools. Motifs are predicted in promoter regions using multiple discovery methods applied to sequence sets that include corresponding sequence regions from vertebrates. We estimate motif significance by applying discovery and post-processing methods to randomized sequence sets that are adaptively derived from target sequence sets, retain motifs with *p*-values below a threshold and identify groups of similar motifs and co-occurring motif patterns. The database offers information on atomic motifs, motif groups and patterns. It is web-accessible, and can be queried directly, downloaded or installed locally.

INTRODUCTION

Approaches for identifying transcriptional regulatory motifs computationally have been reviewed previously (1–3). Recently, progress has been made towards identifying a 'comprehensive catalog' of mammalian elements (4,5). As genome resources, data types and tools evolve, predictive approaches developed for such work can also be directed at a number of closely connected issues. Some of these include taking advantage in motif discovery of increasing numbers of genomes, including low-coverage sequences; quantifying the contributions to motif discovery of different genomes or sets of genomes; improving the predictive reliability of motifs by genome-scale clustering and co-occurrence; determining a best minimal set of motif discovery methods, probably in a discovery approach that uses multiple methods (6); and using coexpression and other functional data types.

In this report, we describe a new cisRED database that contains predictions for whole-genome discovery of regulatory elements in mammals and other eukaryotes. We also describe the predictive system behind the database, which uses genome-scale approaches to predict deeply conserved *ab initio* motifs and identifies groups of similar motifs and co-occurring patterns of motifs. Results are available in a web-accessible and downloadable MySQL database. The system is designed to be readily maintained and extended in the context of rapidly evolving resources, data types and tools.

DATA SOURCES AND PROCESSING

The system is outlined in Figure 1, and is described in detail at www.cisred.org/content/databases_methods. The upstream section of the pipeline loads the database with significant discovered atomic motifs, and the section downstream of the database identifies groups of similar atomic motifs and co-occurring patterns of motifs. An atomic motif consists of a set of sequences, typically with a common length between 6 and 12 bp, members of which are present in a sequence region on the target species and in corresponding regions on other genomes.

The system uses genome resources that are a combination of directly downloaded and processed sequences, annotations and relationships (e.g. orthology, coexpression and interactions). These are stored in an automatically updated local resource that holds a wide range of public and commercial databases.

*To whom correspondence should be addressed. Tel: +1 604 675 8170; Fax: +1 604 876 3561; Email: grobertson@bcgsc.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

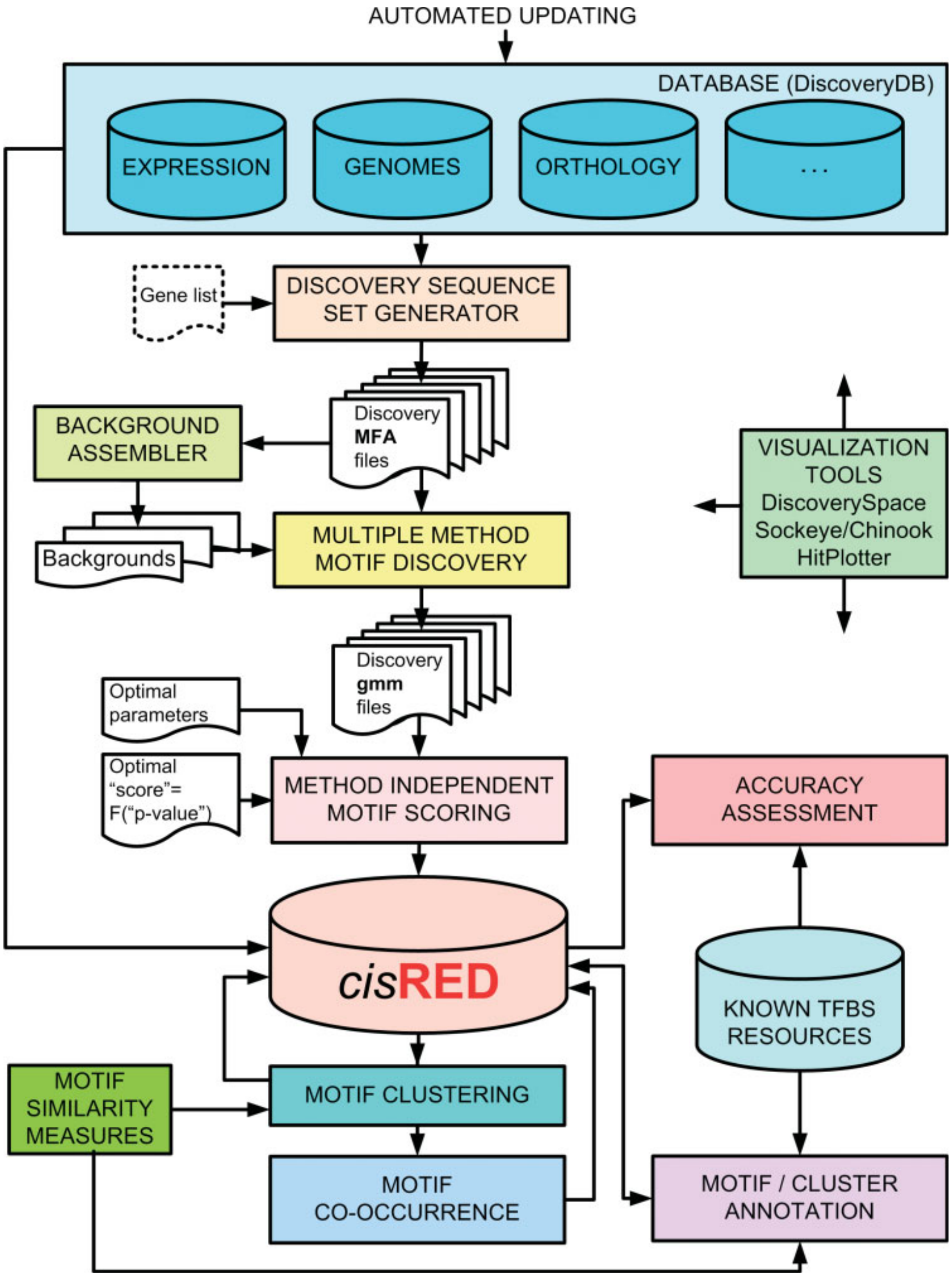

Figure 1. Data processing system for high-throughput motif discovery, clustering, co-occurrence, annotation and performance assessment.

Motif discovery was carried out in a search region based on a single major transcript for each gene. An input target sequence set consisted of a sequence from the target species, and corresponding sequences from homologous vertebrate genes. The following rules were used to assemble target sequence sets for cisRED 1.2e. We identified homologous genes by combining data from Compara (7), HomoloGene (8), Inparanoid (9) and KEGG (10). For each target human gene and each of its orthologues, we took the major Ensembl (7) transcript if its protein sequence was N-terminal complete. We further required that the human gene had an annotated Ensembl 5′-untranslated region (5′-UTR) that was at least 10 bp long. For each target gene's orthologue set, we required at least one of Dog, Mouse and Rat, and at least one of Chicken, Frog, Fugu, Tetraodon and Zebrafish. If the human 5′-UTR was <500 bp, we applied no UTR requirements to orthologues; however, if the human 5′-UTR was >500 bp, we required that all orthologues had annotated Ensembl 5′-UTRs that were at least 10 bp long. After this filtering, for any orthologous region that was missing from an input sequence set, we added a sequence region from the current UCSC (11,12) multiple sequence alignment. For a small subset of the target genes, corresponding regions from ENCODE-specific species were then added from the current ENCODE multiple sequence alignment (13). Finally, regions were added from an internally processed version of the unannotated

Macaca mullata genome from http://www.hgsc.bcm.tmc.edu/projects/rmacaque/. These rules delivered ~7500 human target genes with sequence sets that contained a minimum of three and an average of six orthologous sequences from other genomes. Sequence regions extended 1.5 kb upstream and 100 bp downstream of transcription start sites (TSS), net of all types of repeats except LTR/ERV1, LTR/ERVL, LTR/MaLR and of coding sequences, which were masked.

In addition to a target sequence set, we supplied each discovery program with a genomic 'background' input file. The background input for each species consisted of 1000 concatenated search regions that were randomly selected from the genome's entire set of search regions.

We used multiple discovery methods in parallel, running each method with a range of parameter settings; typically target motif width and motif occurrence model were varied. We used a compact but diverse base set of discovery methods that consisted of CONSENSUS (14), MEME (15) and MotifSampler (16). Raw discovered motifs were post-processed, for example, to remove identical motifs reported by the same algorithm, and to merge strongly overlapped motifs.

We assigned *p*-values to motifs discovered by multiple methods across large sets of target genes whose sequence sets varied in species composition, as follows. First, we identified a representative subset of target sequence sets that sampled the range of species compositions of the target sets. Then, for each representative target sequence set, we created random sequence sets by retaining the original sequence from the target species and replacing each orthologous sequence with a synthetic sequence. The random synthetic sequence was generated from the target sequence by a tool we developed to simulate neutral evolution using published substitution rates and indel rates and lengths. Each random sequence set was generated 30 times in order to avoid statistical bias. We then submitted each target sequence set and all random sequence sets to identical motif discovery and post-processing procedures. We assigned method-independent (MI) scores to all motifs discovered in target and random sequence sets, using a trainable function that contained four non-negative parameters:

$$\text{score}(\alpha) = \frac{(1+\alpha_1 D)(1+\alpha_2 C_{\text{remote}})(1+\alpha_4 W)}{(1+\alpha_3(1-C_{\text{close}}))}B,$$

where coefficients took the following possible values: D, $W \in R$, $B \in \{0, 1\}$ and $C \in [0, 1]$. In the above equation, D characterized the number of site sequences in a motif; W and B characterized the shape of a motif's information content profile; and C characterized motif sequence conservation. The score increased when the motif was conserved for species that are evolutionarily remote from the target, and decreased when the motif was not conserved for species that are close to the target (primates, in the case of human). We used the distribution of MI scores for motifs from a target gene's random sequences to transform the MI scores for the target gene's motifs into *p*-values. Finally, we loaded the database with motifs whose *p*-values were below a threshold (which, for cisRED 1.2e, was 0.05).

A library of known transcription factor binding sites, split into mutually exclusive training and testing fractions, was used to optimize the scoring function and to characterize the performance of the system. The library contained ~1000 sites for ~300 human genes from TRANSFAC v9.1 (17), and ~250 binding sites that we curated from the literature. We optimized the scoring function by simulated annealing, using a training fraction of known sites from randomly selected genes and two objective functions: the area under a receiver operating characteristic (ROC) curve and the number of experimentally known motifs that were not predicted. We assessed the system's predictive performance with a test fraction of known sites, using observables like sensitivity, specificity and positive predictive value (PPV) (e.g. www.cisred.org/content/databases_methods/human_1_2e/performance). Although comparisons of method performance are constrained by many factors, current system performance compared favourably to results in a recent study (6).

To identify groups of similar motifs, we defined two pairwise motif similarity metrics. For the first metric, we used a version of the Levenstein edit distance between two sequences that was modified to permit no internal gaps (18). For each motif, M, and its reverse complement, RC(M), we scanned motif pairs relative to each other, and reported the overall minimum average mutual edit distance to motif K, i.e. $\min(d(M,K), d(\text{RC}(M),K))$. The second metric was based on the maximum information content shared between position frequency matrices derived for each motif, and also treated a motif and its reverse complement as equivalent. We hierarchically clustered pairwise dissimilarity matrices with the local density-based OPTICS algorithm (19). We extracted clusters from OPTICS' reachability output by applying an automatic cluster recognition method that identifies cluster boundaries as inflection points in the reachability plot (20), then traversing these hierarchical segmentation results with an algorithm that traced a deepest available path, constrained by a maximum preset depth.

The large size of mammalian genomes makes it challenging to organize the computational hardware and software infrastructure required to address such issues routinely. We did the large-scale discovery, similarity and co-occurrence calculations on a Beowulf-style, ~400 CPU (Pentium III, Xeon, Opteron) OSCAR cluster (http://oscar.openclustergroup.org) running Red Hat Linux 9; and remotely on the Beowulf ~1700 CPU 'glacier' cluster at WestGrid (www.westgrid.ca). We clustered motifs on a 12 dual-core CPU (UltraSPARC IV) SMP server with 96 GB of RAM running Solaris 9.

CISRED DATABASE CONTENTS, STRUCTURE AND ACCESS

Figure 2 shows a schematic diagram of the database design from a user perspective. The database infrastructure is designed to evolve to hold results from a range of mammals, as well as results from model organisms. Because promoter regions are enriched sources of regulatory elements, motif discovery and the cisRED design were both based on regions around TSSs. cisRED human v1.2 contains motifs from promoter regions of ~7500 human genes using Ensembl v30 (NCBI 35) data, as well as a pilot result set of ~250 mouse genes.

The current database design makes three types or levels of information available for regulatory elements: (i) atomic motifs, which are discovered independently in each target

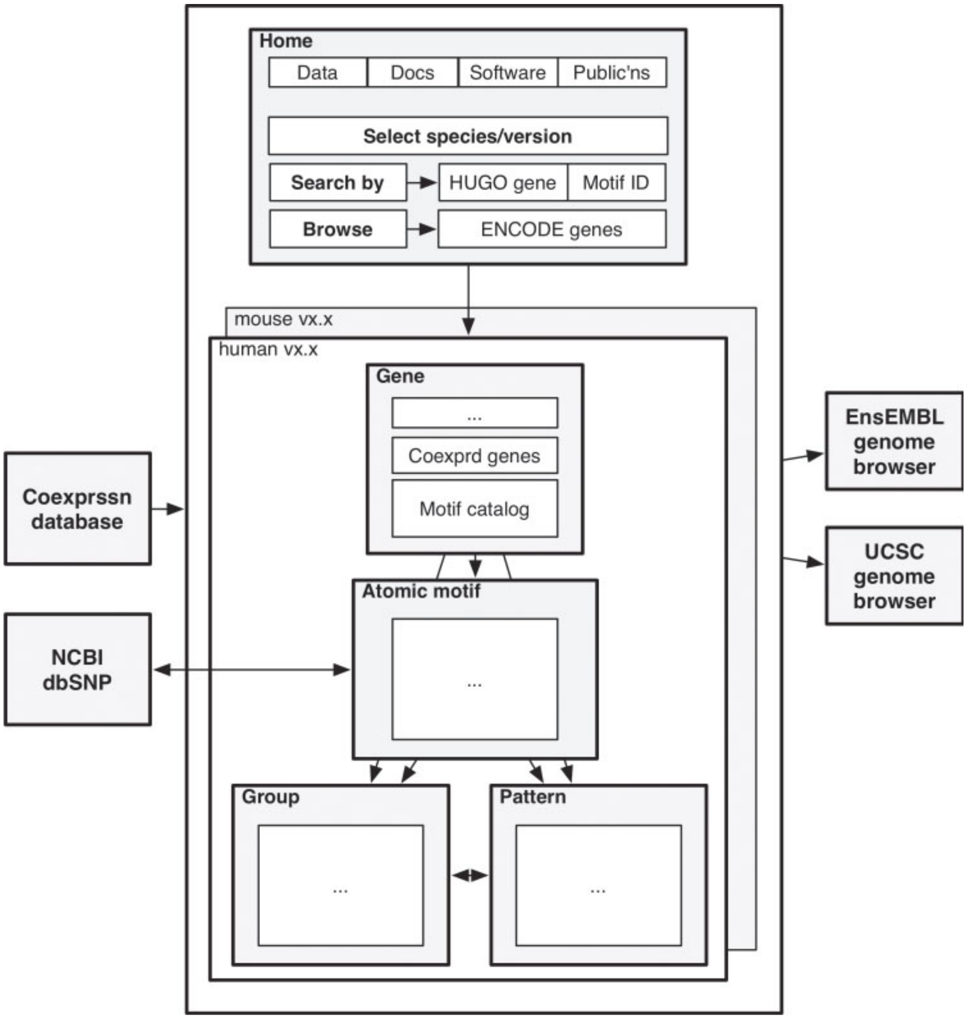

Figure 2. Database contents and high-level links, from a web user perspective, as of cisRED human v1.2. 'Atomic motifs' are motifs discovered in target sequence sets. 'Groups' are clusters of similar motifs that are identified by large-scale OPTICS (19) clustering. 'Patterns' are co-occurring sets of group-labelled motifs.

region sequence set; (ii) groups of similar motifs, each of which is a putative model for the binding site of a single transcription factor; and (iii) co-occurring patterns of models, which are putative regulatory modules.

Predicted regulatory elements can be viewed directly in cisRED's web user interface. From this interface, motifs can be viewed in the UCSC genome browser, in the Ensembl genome browser via a DAS server, or in the Sockeye comparative genomics workspace (21). A user can filter the displayed motifs by criteria like the *p*-value threshold, the orthologous species present in a motif and the discovery method.

The database contains a table of high-confidence globally coexpressed genes (22); genes coexpressed with each cisRED target gene are listed on that gene's page. Coexpression resources are available for download at http://www.bcgsc.ca/gc/bomge/coexpression. The database also contains a table of single nucleotide polymorphisms (SNPs) from dbSNP (8) that occur in predicted motifs. When a predicted human motif contains a SNP, the cisRED atomic motif page highlights this variation on the target sequence of a motif site sequence set, and the highlighted base hyperlinks to the SNP's primary source information in dbSNP.

The database can be accessed in several ways. A web user interface is available at www.cisred.org. A current schema diagram is available from the 'Databases & Methods' page, and direct SQL queries can be run on the MySQL databases at db.cisred.org. A user can download the data, with SQL files, as well as a compressed file that contains all input FASTA sequence sets. As genome resources and our data processing evolve, we update database contents. Because older versions of the database may not be compatible with the current user interface, only recent versions for each species can be accessed through the web or via a MySQL client. However, historical releases of cisRED databases are archived and can be downloaded.

Certain parts of the system's software are available from a tab on the cisRED home page. Sockeye (21) permits, for example, a user to assess details of conserved regions relative to genomic annotations in multiple sequence alignments. The HitPlotter visualizer displays large sets of discovered motifs from multiple-method discovery runs, and is available on request as a beta release. The database infrastructure is designed to facilitate installing cisRED locally.

FUTURE WORK

Database contents will be extended to include large-scale results for human, mouse, rat, *Caenorhabditis elegans* and *Drosophila melanogaster*. A new schema design supports these species. For example, cisRED human v2 will contain ∼18 000 human genes. The input sequence sets for this database were based on human TSSs that were identified by considering Ensembl (7) and RefSeq (23) annotations. To address the limitations of gene and transcript annotations for non-human species, corresponding vertebrate search regions were taken from UCSC (11,12) and ENCODE multiple sequence alignments (13). We are extending our ability to take advantage of unannotated and low-coverage genome sequence data.

We are continuing to improve motif post-processing and scoring. Optimizing the scoring function depends on having a large library of known motif sites; we anticipate that a newly created web database for submitting and curating binding sites from the literature may help us to enlarge and improve this resource (www.oreganno.org; S. B. Montgomery and O. L. Griffith, manuscript in preparation).

Given a scalable clustering method, we continue to assess motif similarity metrics and how best to use the hierarchical information output from OPTICS. Given groups of similar motifs, we are applying group labels to atomic motifs, then identifying overrepresented co-occurring motif patterns using hypergeometric statistics, imposing separation constraints on neighbouring motifs, and searching in two stages for patterns larger than pairs (24,25). We are implementing methods for annotating discovered motifs as known or novel against known site resources. Given annotated motifs, we will annotate overrepresented co-occurring patterns of human motif pairs as known versus novel using the TRANSCompel resource (26). We are applying genome-scale motif clustering and co-occurrence as filters for predicted motifs that may improve the predictive reliability and the resulting catalogue of conserved regulatory elements.

We have assembled a large multi-species coexpression resource that contains public microarray and SAGE data from diverse sources (22) (Table 1). We have shown that combining global coexpression data from multiple platforms improves confidence in coexpression predictions when assessed against the Gene Ontology (GO) (27). From this, we established GO-based Pearson correlation thresholds that identified high-confidence globally coexpressed gene pairs. The coexpression database makes results available from this global analysis and from two other recent analyses (28,29). Although coexpressed genes can have similar regulatory elements, the system's predictive performance improved only marginally when inputs included coexpressed genes in addition to orthologous genes (data not shown). Given these results, currently we are assessing an approach that includes no coexpressed genes in motif discovery inputs, but uses coexpression information to assess groups and co-occurring patterns identified in genome-scale sets of atomic motifs.

Table 1. Contents of coexpression database (22)

Species	Platform	Experiments	Unique genes
H.sapiens	SAGE (short)	272	20 312
	Oligo microarray	1640	12 452
	cDNA microarray	2852	13 111
M.musculus	SAGE (short)	85	12 715
	Oligo microarray	1802	8 164
	cDNA microarray	366	8 102
C.elegans	SAGE (long/short)	26	15 685
	cDNA microarray	1059	15 956
Total		8102	54 434

We will extend the database user interface to offer more complex user filtering, as well as motif searches based on consensus strings or matrices. For work with classes of regulatory elements that are defined by wet lab data types based on, for example, ChIP or DNase I hypersensitivity (e.g. see ENCODE tracks within the UCSC genome browser) (11,13), we have designed a new schema that is based on search regions rather than on genes, and will extend the user interface to support this.

We will continue to assess the contributions to regulatory element predictions of different genomes and sets of genomes. We will integrate and assess new motif discovery methods, and will identify a best minimal set of methods on an ongoing basis.

ACKNOWLEDGEMENTS

We gratefully acknowledge funding from Genome Canada, Genome British Columbia and the BC Cancer Foundation. S.J.M.J., M.C.S., S.B.M. and O.L.G. were supported by the Michael Smith Foundation for Health Research (MSFHR). E.D.P. was supported by the Canadian Institutes of Health Research (CIHR). S.B.M. and O.L.G. were also supported by the Natural Sciences and Engineering Research Council. H.H. was supported by the CIHR/MSFHR Strategic Training Program in Bioinformatics. This research has been enabled by the use of WestGrid computing resources, which are funded in part by the Canada Foundation for Innovation, Alberta Innovation and Science, BC Advanced Education, and the participating research institutions. Funding to pay the Open Access publication charges for this article was provided by Genome British Columbia.

Conflict of interest statement. None declared.

REFERENCES

1. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.*, **5**, 276–87.
2. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.
3. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
4. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
5. Elemento,O. and Tavazoie,S. (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.*, **6**, R18.
6. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

7. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
9. O'Brien,K.P., Remm,M. and Sonnhammer,E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
10. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
11. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., Weber,R.J., Haussler,D. and Kent,W.J. (2003) The UCSC Genome Browser database. *Nucleic Acids Res.*, **31**, 51–54.
12. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
13. Feingold,E.A., Good,P.J., Guyer,M.S., Kamholz,S., Liefer,L., Wetterstrand,K., Collins,F.S., Gingeras,T.R., Kampa,D., Sekinger,E.A. *et al.* (2004) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, **306**, 636–640.
14. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
15. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
16. Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2002) A Gibbs Sampling method to detect over-represented motifs in upstream regions of coexpressed genes. *J. Comp. Biol.*, **9**, 447–464.
17. Wingender,E. (2004) TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol.*, **4**, 55–61.
18. Levenstein,V.I. (1966) Binary codes capable of correcting insertions and reversals. *Sov. Phys. Docl.*, **10**, 707–710.
19. Ankerst,M., Breunig,M.M., Kriegel,H.-P. and Sander,J. (1999) OPTICS: Ordering Points To Identify the Clustering Structure. In *Proceedings of the Int. Conf. on Management of Data (ACM SIGMOD'99),* Philadelphia, PA.
20. Brecheisen,S., Kriegel,H.-P., Kröger,P. and Pfeifle,P.M. (2004) Visual mining through cluster hierarchies, Proc. SIAM Int. Conf. on Data Mining (SDM'04), Lake Buena Vista, FL, pp. 400–412.
21. Montgomery,S.B., Astakhova,T., Bilenky,M., Birney,E., Fu,T., Hassel,M., Melsopp,C., Rak,M., Robertson,A.G., Sleumer,M.C. *et al.* (2004) Sockeye: a 3D environment for comparative genomics. *Genome Res.*, **14**, 956–962.
22. Griffith,O.L., Pleasance,E.D., Fulton,D.L., Oveisi,M., Ester,M., Siddiqui,A.S. and Jones,S.J.M. (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics*, **86**, 476–488.
23. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
24. Zhu,Z., Shendure,J. and Church,G.M. (2005) Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, **15**, 848–855.
25. Kreiman,G. (2004) Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res.*, **32**, 2889–2900.
26. Kel-Margoulis,O.V., Kel,A.E., Reuter,I., Deinenko,I.V. and Wingender,E. (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.
27. Gene Ontology Consortium (2004), The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
28. Jensen,L.J., Lagarde,J., von Mering,C. and Bork,P. (2004) ArrayProspector: a web resource of functional associations inferred from microarray expression data. *Nucleic Acids Res.*, **32**, W445–W448.
29. Lee,H.K., Hsu,A.K., Sajdak,J., Qin,J. and Pavlidis,P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.