Nucleic Acids Research, 2005, Vol. 33, Database issue ***D103–D107***
doi:10.1093/nar/gki004

TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies

Fang Zhao, Zhenyu Xuan, Lihua Liu and Michael Q. Zhang*

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

Received August 12, 2004; Revised August 31, 2004; Accepted September 8, 2004

ABSTRACT

In order to understand gene regulation, accurate and comprehensive knowledge of transcriptional regulatory elements is essential. Here, we report our efforts in building a mammalian Transcriptional Regulatory Element Database (TRED) with associated data analysis functions. It collects *cis*- and *trans*-regulatory elements and is dedicated to easy data access and analysis for both single-gene-based and genome-scale studies. Distinguishing features of TRED include: (i) relatively complete genome-wide promoter annotation for human, mouse and rat; (ii) availability of gene transcriptional regulation information including transcription factor binding sites and experimental evidence; (iii) data accuracy is ensured by hand curation; (iv) efficient user interface for easy and flexible data retrieval; and (v) implementation of on-the-fly sequence analysis tools. TRED can provide good training datasets for further genome-wide *cis*-regulatory element prediction and annotation, assist detailed functional studies and facilitate the decipher of gene regulatory networks (http://rulai.cshl.edu/TRED).

INTRODUCTION

To understand gene regulatory mechanisms and networks requires accurate and comprehensive knowledge of transcriptional regulatory elements. They include *cis*-elements, such as promoters and *trans*-elements, such as transcription factors. A number of databases have been created to facilitate such studies. However, most of them are only dedicated to either promoter annotation or transcription factor binding and functional information, which make data access disconnected and correlation of different types of data difficult. Hence, we are motivated to build a unique resource for both *cis*- and *trans*-regulatory elements, and provide easy access of the correlation between promoter sequences and transcription factor binding information.

Although current promoter databases have provided much value for the development of promoter-finding programs and gene regulation studies (1,2), many have their own limitations. These include incomplete datasets, inadequate data accuracy, restricted accessibility of the data and lack of sequence analysis functionalities. On top of single-gene-based and whole genome experimental promoter identification, computational methods are greatly needed for efficient genome-wide promoter annotation. However, in higher eukaryotes, promoter finding *in silico* has turned out to be one of the most difficult problems in computational biology (3). Therefore, accurate promoter annotation for all the genes in higher eukaryotes is still an outstanding challenge.

Collecting comprehensive and precise transcription factor binding and regulation information currently known is a daunting task. It involves painstaking and time-consuming literature curation by transcription study experts. Although there are a limited number of databases (4,5) dedicated to this aspect of data collection, they often do not conveniently correlate functional information to the relevant promoter sequences and its genomic context that are required in most of the regulation studies. Furthermore, inevitably, data completeness is always an issue.

Here, we report our efforts in building a Transcriptional Regulatory Element Database (TRED) with associated data analysis functions. With the availability of complete genome sequences for human and draft sequences for mouse and rat, we have mapped out and documented in the database gene transcription start sites (TSSs) and core promoters for the whole genomes through both automated pipeline and hand curation. In addition, we have been carrying out continuous expert curation of transcription factor binding and regulation information on these promoters. Our short-term goal is to provide comprehensive and accurate *trans*-regulatory information for target genes of cancer-related transcription factors. We have so far included binding data for a few transcription factors, with emphasis on two major cell cycle regulators, E2F and Myc. For each, we have recorded thousands of target genes of different binding qualities as demonstrated by various experiments.

*To whom correspondence should be addressed. Tel: +1 516 367 8393; Fax: +1 516 367 8461; Email: mzhang@cshl.edu

A web-based user interface has been implemented for easy data visualization, retrieval and analysis for both single-gene-based studies and large-scale sequence manipulation and gene regulatory network studies. We intend to build TRED to contain information of both *cis*- and *trans*-regulatory elements for every annotated gene, and to serve as a one-stop data provider for researchers interested in gene regulation studies.

DATA SOURCES

Promoter annotation

Promoters in TRED came from two sources: automated genome-wide annotation and hand curation. They complement each other, and together realize the relative completeness and accuracy of the data. The automated annotation pipeline was built to extract and merge known promoters from databases such as EPD and DBTSS (1,6), employing promoter-finding programs such as FirstEF (7) combined with mRNA/EST information and cross-species comparisons to predict promoters, and associating them with known or predicted genes (Z. Xuan, F. Zhao, J. Wang, G. Chen and M. Q. Zhang, submitted for publication). Given the difficulty and complexity of promoter prediction in higher eukaryotes, accuracy of computational promoter annotation is limited. Therefore, hand curation was applied as a crucial part of our data collection to assess computational prediction and ensure data accuracy. After we pooled data from both sources, further data cleaning and integration were carried out. Based on the reliability of the supporting evidence for each promoter, a quality level was assigned.

Transcription factor binding curation

Curation was carried out for transcriptional regulation information on promoters. Exhaustive literature search for target genes of individual transcription factors was carried out, binding motifs and experimental evidence were recorded, and transcription factor binding motifs were mapped on promoters of the target genes. Binding quality levels were assigned based on definitiveness of the binding evidence, which was determined by the experimental approaches employed to demonstrate the binding and expert data interpretation. A standardized curation format has been developed for easy data entry and automated data loading into the database. To best preserve the curated association between motifs and promoters through changes such as genome assembly releases and genome annotations, we also record motif flanking sequences.

Curation is a time-consuming and laborious process, and we started out by focusing on target genes of cancer-related transcription factors. In compliance with the broad interest in cell cycle regulatory network studies, we have completed curation for transcription factor E2F and Myc target genes. They are involved in various biological pathways and have profound effects in cell proliferation (8–13). Many E2F and Myc target genes have been identified by traditional transcription studies as well as newly developed, large-scale functional genomics studies.

DATABASE CONSTRUCTION AND IMPLEMENTATION

A MySQL relational database was constructed for storage and query of the data. It includes three key entities: 'Promoter', 'Gene' and 'Factor'. 'Promoter' is a weak entity

Table 1. Promoter and gene numbers in TRED, with gene numbers in parentheses

Promoter quality	1	2	3	4	5 + 6	Sum
Human	1971 (1779)	13 120 (9769)	24 563 (14 363)	9217 (7214)	9358 (8877)	58 229 (30 981)
Mouse	214 (179)	8385 (6675)	20 318 (12 122)	13 252 (10 812)	8595 (8442)	50 764 (31 683)
Rat	91 (84)	808 (534)	7157 (3987)	819 (614)	21 511 (21 437)	30 386 (26 064)

Promoter qualities ranked from high to low: 1, known, curated promoters; 2, known, pipeline collected promoters; 3, predicted promoters with Refseq evidence and putative promoters taking 5′ ends of Refseq as TSSs; 4, predicted promoters with mRNA (other than Refseq and EST) evidence; 5, predicted promoters with EST evidence; 6, predicted promoters supported only by gene prediction.
Promoters included in a higher ranking are automatically excluded from the lower ranking categories.

Table 2. Numbers of curated E2F and Myc target promoters and genes, with gene numbers in parentheses

	E2F targets			Myc targets		
	Human	Mouse	Rat	Human	Mouse	Rat
Promoter quality[a]						
1	221 (164)	59 (47)	9 (9)	298 (263)	5 (5)	4 (3)
2	388 (355)	14 (14)	0 (0)	1125 (651)	43 (31)	26 (15)
3	496 (454)	29 (29)	2 (2)	1230 (730)	59 (34)	90 (54)
4	249 (229)	18 (18)	0 (0)	15 (12)	7 (5)	4 (4)
5 + 6	239 (231)	21 (21)	0 (0)	8 (7)	10 (5)	4 (3)
Sum	1593 (1329)	141 (127)	11 (11)	2676 (785)	124 (38)	128 (62)
Binding quality[b]						
Known	166 (127)	20 (13)	0 (0)	2667 (782)	70 (28)	108 (54)
Likely	10 (10)	0 (0)	0 (0)	4 (1)	10 (3)	9 (3)
Maybe	1217 (1048)	70 (69)	0 (0)	5 (3)	28 (7)	11 (5)
Computationally predicted	200 (175)	51 (48)	11 (11)	0 (0)	0 (0)	0 (0)
Sum	1593 (1329)	141 (127)	11 (11)	2676 (785)	108 (38)	128 (62)

[a]Number break-down by promoter quality (promoter quality definition is the same as that in Table 1).
[b]Number break-down by E2F and Myc binding quality.

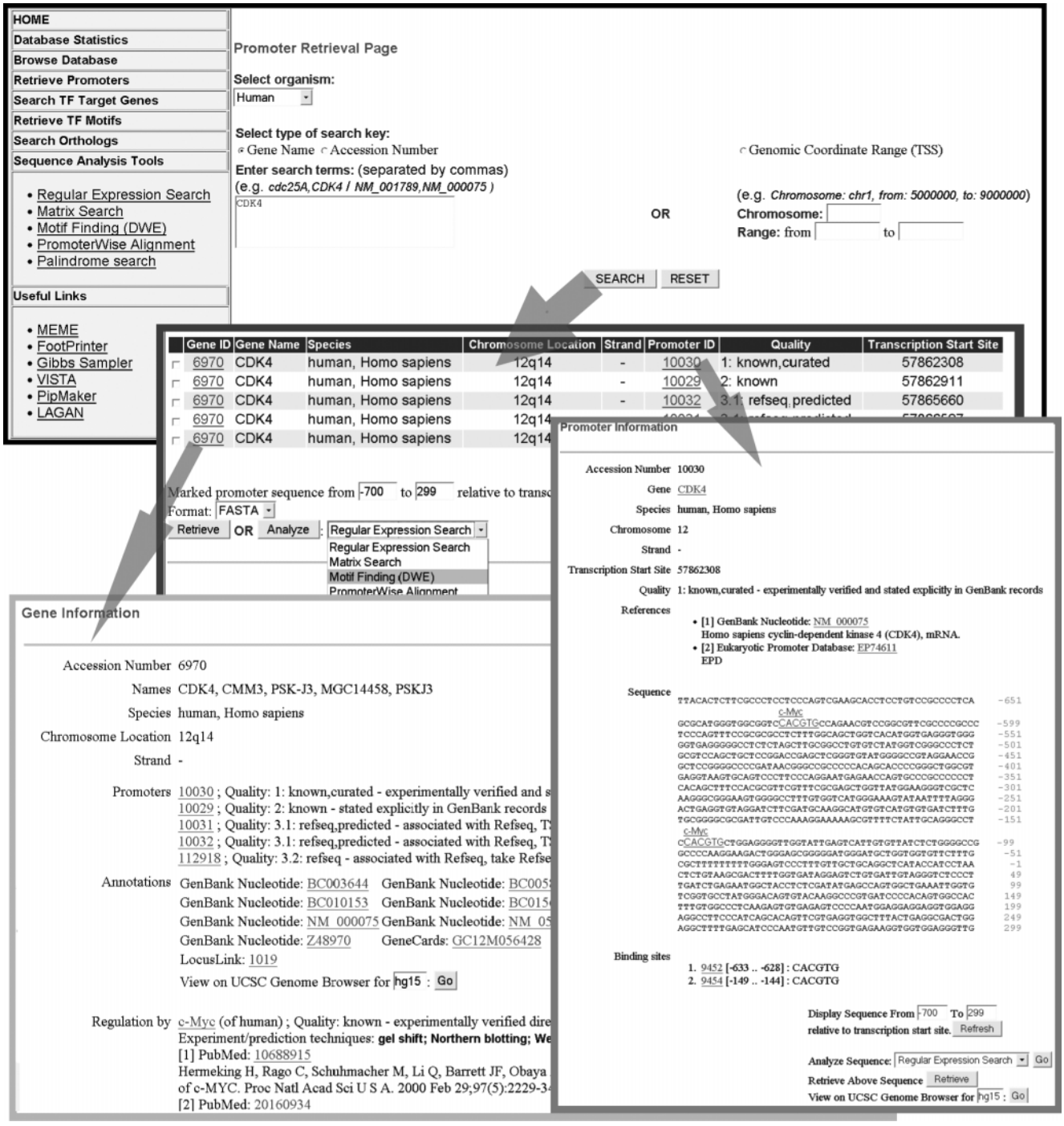

Figure 1. Sample pages showing TRED user-interface for gene promoter search, promoter search results, gene information and promoter information.

because our model would not allow a promoter to exist without the associated gene. There are two key relationships: (i) a promoter regulates a gene, which is a many-to-one relationship; and (ii) a factor binds a promoter, which is a many-to-many relationship. Other entities in the relational schema include promoter qualities, binding motifs, binding qualities and external data sources. Other relationships include gene annotation, promoter supporting evidence, factor annotation and binding supporting evidence. An automated data look-up, integration and loading pipeline has been developed for easy populating and updating the database.

DATABASE CONTENT

TRED contains whole genome promoter annotation for human, mouse and rat from both curation and computational prediction. The number of genes and promoters in various quality categories are listed in Table 1. From our extensive

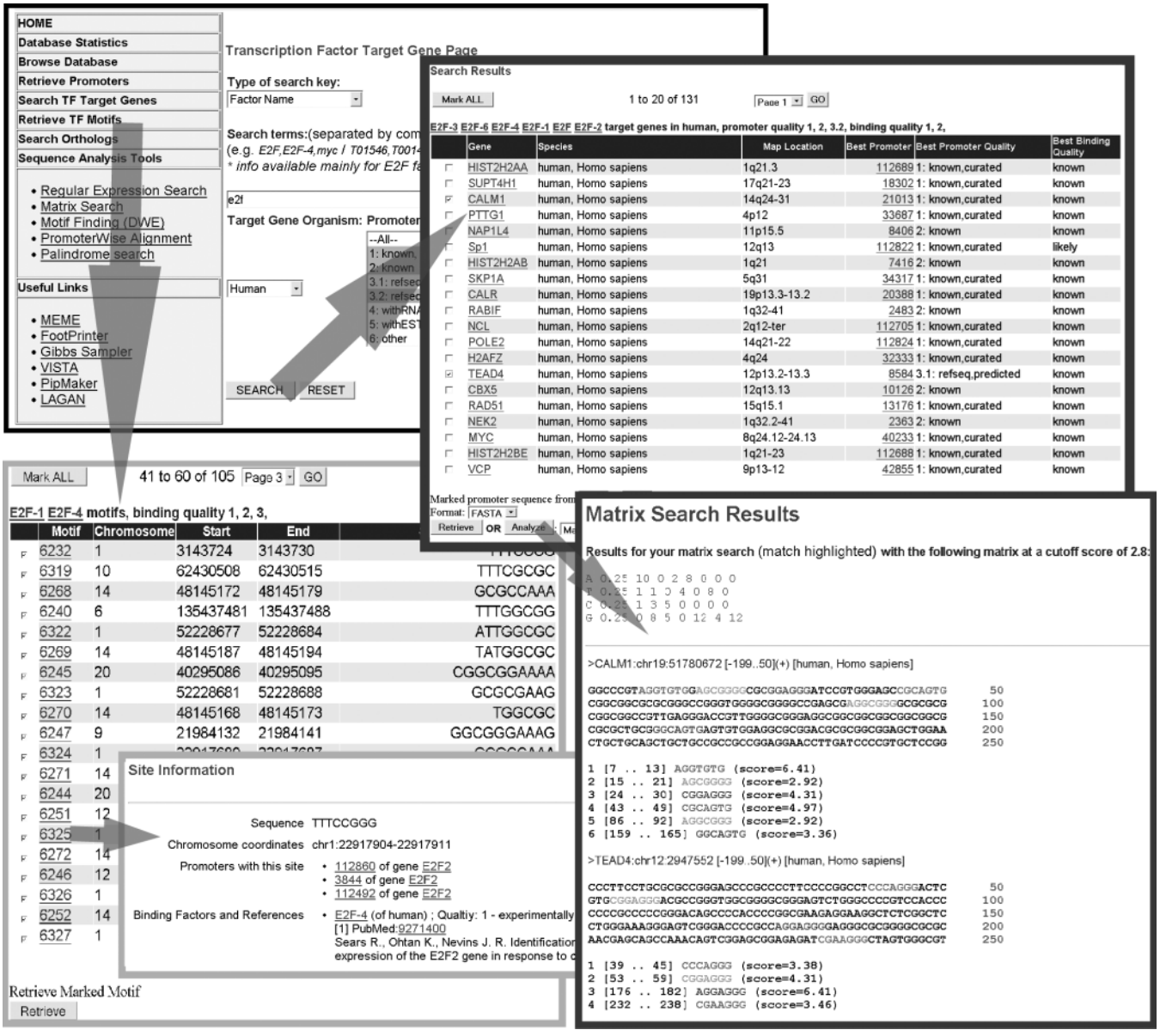

Figure 2. Sample pages showing TRED user-interface for promoter retrieval for target genes of a transcription factor, on-the-fly sequence analysis results, transcription factor binding motif retrieval and biding site information.

literature curation, TRED holds functional annotations of hundreds of direct target genes for E2F and Myc in human, mouse and rat with concrete binding evidence (high binding quality levels) (Table 2). Many of them have experimentally verified promoter sequences and known E2F or Myc binding motifs. It also has a collection of thousands of genes shown to be regulated by E2F and Myc of lower binding confidence (e.g. only demonstrated by expression experiments or computational prediction). This is a more comprehensive collection than that recorded for these two transcription factors in the Transfac database (4). Some target genes for a few other transcription factors are also included in the current release of TRED. To provide users with further information of the genes, cross-references to other well-known databases such as GenBank, PubMed, GeneCards (14) and Transfac were established.

WEB INTERFACE

Data access and retrieval

A CGI/Perl-based web interface was built to facilitate easy visualization and retrieval of both single-gene-based and batch data. It carries the following major functionalities.

(i) Search promoters for a gene or a list of genes by gene name, GenBank ID or chromosome location (Figure 1). The resulting page contains all annotated promoters for the gene, ranked from the highest quality to the lowest. Links for gene information and promoter information (including localization of transcription factor binding sites) are provided by the hotlinks in 'Gene ID' and 'Promoter ID' columns, respectively (see Figure 1). Sequence retrieval of desired promoters can be achieved

by checking the box on the left of each entry. Sequence length for retrieval can be decided by users, with the default being 1 kb (700 bp upstream and 299 bp downstream of TSS). Promoter sequences of interest can also be conveniently sent to 'on-the-fly analysis' page for further analysis (see below).

(ii) Gene information page displays the annotation and promoter links for a particular gene, as well as transcription factors that regulate the gene, experimental evidence and literature references. A link is provided to locate the gene on UCSC Genome Browser and access additional annotations (Figure 1).

(iii) Promoter information page includes genomic localization of the promoter, annotation references and the sequence, with transcription factor binding sites marked and hot linked to detailed binding information and literature references. A link is provided to locate the promoter on UCSC Genome Browser and access its genomic context (Figure 1).

(iv) Retrieve promoter sequences for all target genes of a transcription factor, with the option of filtering sequences for desired promoters and binding qualities (Figure 2). This will conveniently produce good datasets for computational studies on transcriptional regulons and networks, as well as for the development and training of computational tools such as motif-finding programs.

(v) Retrieve all binding motifs for a transcription factor (Figure 2). This can greatly facilitate the construction of transcription factor binding positional weight matrices (PWMs) for target gene identification and gene regulation studies.

(vi) Browse the genome for genes/promoters located in a particular chromosome.

(vii) Search for orthologous genes based on the annotation in Ensembl.

On-the-fly analysis tools

On-the-fly analysis tools were implemented for sequences retrieved from TRED or imported from other resources (Figure 2). They currently include simple sequence manipulation and analysis tools for users' convenience and motif-matching programs based on regular expression and PWM. A word counting-based motif searching tool DWE (15) and PromoterWise, a program specifically for pair-wise promoter local alignment (E. Birney, unpublished), are also implemented. Promoters on various TRED sub-pages can be directly sent to these analysis tools at a click of a button.

In addition to the on-the-fly tools, TRED also provides links to many other sequence analysis and motif-finding programs such as MEME (16) and Gibbs sampler (17).

FUTURE DEVELOPMENTS

Updating of genome-wide promoter annotation based on newer genome assembly releases can be automated and will be done for the next release. Promoter annotation for mammals other than human, mouse and rat will be carried out and included in TRED. For transcription factor binding and regulation information, literature curation has been a continuing effort. We hope to finish target genes of cancer-related transcription factors in the near future, and eventually expand to targets of other transcription factors.

ACKNOWLEDGEMENTS

We thank Ewan Birney for providing the PromoterWise program. This work is supported by NIH grants (HG01696, HG02600 and GM06513) to M.Q.Z.

REFERENCES

1. Perier,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res.*, **28**, 302–303.
2. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.
3. Zhang,M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nature Rev. Genet.*, **3**, 698–709.
4. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
5. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
6. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.*, **32**, D78–D81.
7. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet.*, **29**, 412–417.
8. Ren,B., Cam,H., Takahashi,Y., Volkert,T., Terragni,J., Young,R.A. and Dynlacht,B.D. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.*, **16**, 245–256.
9. Ishida,S., Huang,E., Zuzan,H., Spang,R., Leone,G., West,M. and Nevins,J.R. (2001) Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. *Mol. Cell. Biol.*, **21**, 4684–4699.
10. Farnham,P.J., Slansky,J.E. and Kollmar,R. (1993) The role of E2F in the mammalian cell cycle. *Biochim. Biophys. Acta*, **1155**, 125–131.
11. Muller,H. and Helin,K. (2000) The E2F transcription factors: key regulators of cell proliferation. *Biochim. Biophys. Acta*, **1470**, M1–M12.
12. Coller,H.A., Grandori,C., Tamayo,P., Colbert,T., Lander,E.S., Eisenman,R.N. and Golub,T.R. (2000) Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl Acad. Sci. USA*, **97**, 3260–3265.
13. Fernandez,P.C., Frank,S.R., Wang,L., Schroeder,M., Liu,S., Greene,J., Cocito,A. and Amati,B. (2003) Genomic targets of the human c-Myc protein. *Genes Dev.*, **17**, 1115–1129.
14. Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adato,A., Ben-Dor,U., Esterman,N., Rosen,N., Peter,I. *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene c ompendium. *Bioinformatics*, **18**, 1542–1543.
15. Sumazin,P., Chen,G., Hata,N., Smith,A.D., Zhang,T. and Zhang,M.Q. (2004) DWE: discriminating word enumerator. *Bioinformatics*, in press.
16. Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.
17. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.