3554–3559 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg549

PromoSer: a large-scale mammalian promoter and transcription start site identification service

Anason S. Halees[1], Dmitriy Leyfer[1] and Zhiping Weng[1,2,*]

[1]Bioinformatics Program and [2]Biomedical Engineering Department, Boston University, 44 Cummington Street, Boston, MA 02215, USA

Received February 16, 2003; Revised March 12, 2003; Accepted March 31, 2003

ABSTRACT

Proximal promoters have a major impact on transcriptional regulation. Studies of the sequence-based nature of this regulation usually require collection of proximal promoter sequences for large sets of co-regulated genes. We report a newly implemented web service that facilitates extraction of user specified regions around the transcription start site of all annotated human, mouse or rat genes. The transcription start sites have been identified computationally by considering alignments of a large number of partial and full-length mRNA sequences to genomic DNA, with provision for alternative promoters. The service is publicly available at http://biowulf.bu.edu/zlab/PromoSer/.

INTRODUCTION

Recent advances in biotechnology have opened the door for large scale and high throughput analysis approaches to the genomes of many organisms. Using microarray technology for example, patterns of similar expression profiles (conditionally or temporally) have been linked to shared regulatory mechanisms (1–5). The methods used to analyze and elicit these control mechanisms vary. Common approaches include searching for novel *cis*-elements using expectation maximization (6) or Gibbs sampling (7) algorithms. Alternatively, known motifs are used to look for significant clusters (8) of these elements or for recurring patterns that seem to correlate well with certain clusters of genes (9,10).

The common prerequisite for all computational analysis methods is the availability of promoter sequence data. It is well known that enhancer and suppressor control elements can exist at sites tens of thousands of bases upstream or even downstream of the transcription start site (TSS) (11). In many cases however, the essential control elements are present within the proximal promoter a few hundred to a couple of thousand bases upstream of the TSS (12–14).

The reliable identification of the TSS and the extraction of the proximal promoter is not a trivial task (15) and seems to be the bottleneck for many large scale projects that attempt to analyze microarray data, because it has not yet been automated to a satisfactory level. For example, if a researcher identifies a few hundreds of genes with similar expression patterns, it would require tedious script parsing to obtain their promoter sequences. Since most sequences on a microarray are not full length, one at least has to search for the corresponding entry in RefSeq, align the RefSeq entry to the genome and extract the upstream sequence. Due to the wide applications of microarrays, there have been substantial redundant efforts in such promoter finding activities, and sometimes the biologist may simply take the upstream sequence corresponding to the microarray entry, which may in fact correspond to a coding region.

The bioinformatics community's response to this problem takes three forms: one is to collect databases of experimentally verified TSSs as in the Eukaryotic Promoter Database (16). The data set is of high quality but is also relatively small due to the experimental verification requirement. Another approach is to generate and sequence libraries of full-length mRNA molecules using a variety of biochemical methods such as cap-trapping and oligo-capping (17–19). Large-scale projects have yielded rich sets of such full-length mRNA sequences (19–21). The availability of the draft assemblies of an increasing number of mammalian genomes, currently human (22), mouse (23) and rat (Rat Genome Sequencing Consortium), has allowed for yet a third approach where large sets of truncated as well as full-length mRNA sequences are clustered by sequence similarity and aligned to the genome. The TSS is identified as the furthest 5′ aligned position of the cluster (19,24).

Until now, however, the available resources have not been directly helpful to researchers looking for a large set of promoters. Possible reasons include: (i) the systems only provide the full-length mRNA sequence without localizing it to the genome; (ii) the dataset is too small and many genes sought are not included; and (iii) the user interface simply does not allow for batch jobs of many queries. In this study, we aim to provide such a resource. The service is publicly accessible at http://biowulf.bu.edu/zlab/PromoSer/. A typical application of our resource is to couple with the analysis of microarray results.

WEB SERVICE DESCRIPTION AND CONSTRUCTION METHOD

PromoSer is a freely accessible web-based service to facilitate the extraction of a large number of proximal promoter

*To whom correspondence should be addressed at Biomedical Engineering Department, Boston University, 44 Cummington Street, Boston, MA 02215, USA. Tel: + 1 617 353 3509; Fax: +1 617 353 6766; Email: zhiping@bu.edu

Table 1. Distribution of successfully aligned regions[a]

	Human	Mouse	Rat
EST	1 304 059/4 952 197 (26%)	465 016/3 664 685 (13%)	57 202/281 826 (20%)
mRNA	79 217/111 658 (71%)	79 968/102 440 (78%)	6697/9487 (71%)
RefSeq	15 959/18 241 (87%)	10 590/12 903 (82%)	3143/4322 (73%)
EPD[b]	1778/1871 (95%)	110/194 (57%)	74/118 (63%)
RIKEN	—	52 206/60 770 (86%)	—
RefFull[c]	4056/4802 (84%)	—	—
FLmRNA[d]	10 215/12 340 (83%)	—	—
Total	1 415 284/5 101 109 (28%)	607 890/3 841 644 (16%)	67 116/295 755 (23%)

[a]The number of alignments that remained after the filtering process (nominator) and the number of sequences obtained for each genome (denominator). Mouse and rat results are strongly influenced by the draft quality of their genome assemblies.
[b]Eukaryotic Promoter Database.
[c]RefSeq sequences extended by DBTSS.
[d]Full-length human mRNA sequences from IMS, Tokyo, Japan.

sequences. The user supplies a list of mRNA accession numbers (without version numbers) and selects the required sequence range around the TSS. Upon request execution, the user will receive a FASTA formatted text file of the sequences. If the indicated range overlaps with the transcribed region of the immediate upstream gene, the user will be notified and given the choice of retrieving only intergenic sequences. A genome assembly can have gaps. Gaps with known lengths are treated as regular nucleotides (marked with 'N'). Gaps with unknown lengths are considered 'breaks' in the genome assembly; we will notify the user if such a gap occurs in the sequence range requested by the user and return the sequence only up to the gap. Currently, the user can query for any genomic (non-organelle) mRNA from the human (*Homo sapiens*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) genomes. The system utilizes the most recent genome assemblies of each organism and the mRNA set will be updated frequently to keep pace with an expanding GenBank mRNA collection. By 1 June 2003, we should have entire sets of 1, 2 and 5 kb upstream sequences for most commonly used Affymetrix chips available for download at the PromoSer website.

To allow for fast interactive response times, alignments have been pre-computed and stored into a database. To construct the database, we first downloaded the most recent assemblies of the human (22) genome (14 November 2002), mouse (23) genome (February 2002) and rat genome (Rat Genome Sequencing Consortium, November 2002) from the UCSC (25) genome browser (http://genome.ucsc.edu/). These sequences had already been masked with RepeatMasker and Tandem Repeat Finder. During the compute-intensive alignment phase, masked regions were excluded from consideration as likely TSSs.

We then downloaded all available mRNA sequences for the three genomes. These include all available TSS flanking sequences from the Eukaryotic Promoter Database (16), all EST and non-EST mRNA sequences from GenBank (the dbEST and nr databases) and from RefSeq (26). We also downloaded the publicly available set of full-length cDNA sequences from RIKEN (20). In addition we downloaded the available DBTSS (19) extensions to the RefSeq sequences. The human mRNA dataset contains a large subset of full-length cDNA sequences deposited by the Institute of Medical Science, University of Tokyo, Japan.

Using a powerful cluster of 128 dual-processor compute nodes and the efficient BLAT tool (27), each of these >9 000 000 mRNA or EST sequences were aligned to their corresponding genomes and localized to specific chromosomal regions. BLAT is a local alignment tool, which means it occasionally can produce spurious high scoring short alignments; therefore, the alignments were then scored and filtered according to the following criteria:

1. EPD sequences (which are genomic) had to match at ≥95% identity over the length of the query sequence.
2. All other sequences >250 bases had to have their full length aligned to the genome, minus ≤50 bp to allow for poly-A tail truncation. Sequences <250 bp had to align for ≥80% of their length.
3. In addition to length requirements, the alignments had to achieve a minimum match identity to the genomic region they aligned to, according to the sequence type; EST: >90%, 'regular' mRNA: >95%, full-length mRNA: >97%.
4. Only spliced EST sequences were retained to reduce the danger of a genomic contamination to the EST library from which the sequence was obtained.

Alignments that satisfied the filter criteria were scored based on match, mismatch and indel counts. Currently we only keep the best genomic alignment for each query mRNA or EST sequence. Table 1 shows the number of sequences considered and the number of alignments retained after the filtering process. The percentage of aligned human sequences from EPD is much higher than those for mouse and rat, possibly reflecting the quality of genome assembly. A sharp reduction in the number of ESTs can be observed due to the exclusion of non-spliced ESTs.

All the sequences that hit the same genomic region in the same orientation and overlapped fully or partially were grouped into one cluster extending from the 5′ most genomic position to the 3′ most position. Sequences that shared a minimum of 80 bases of *transcribed* region were linked together producing a graph. We resolve all disconnected components of the graph, which represent independent groups of transcripts within this cluster. This manipulation is necessary to untangle interleaved transcripts and recover genes that are embedded within the introns of larger genes. Table 2

Table 2. Number of clusters after combining overlapping alignments in the same orientation

Genome	All clusters[a]	Non-EST[b]	Confirmed[c]
Human	37 572	27 415	16 410
Mouse	39 470	34 028	29 922
Rat	15 915	4847	2738

[a]Counting all alignments as indicated in Table 1 (cluster quality level 1 and above).
[b]Excluding clusters that are entirely EST (cluster quality level 2 and above).
[c]Clusters with at least one known full-length mRNA which was not truncated during alignment (i.e. from Table 1, entries under EPD, RIKEN, RefFull or FLmRNA; cluster quality level 4).

shows the current number of clusters thus obtained. Clusters consisting purely of ESTs are considered of the lowest quality and assigned a quality level 1. Clusters that contain a single non-EST sequence are assigned quality level 2. Those that have >1 non-EST sequence but no full-length sequences are given a quality level 3; for this purpose, sequences presumed to be full length but had >10 bases truncated from their 5′ side were downgraded to 'ordinary' mRNA. All other clusters were assigned quality level 4. The TSS prediction is the 5′ most genomic position of each alignment within the cluster and upstream of the TSS of a full-length sequence if available. If multiple TSS positions >20 bp apart were found, they were reported as alternative promoters. Except in quality 4 clusters, individual ESTs are not considered for alternative promoters and only the 5′ most position from all the ESTs in the cluster is considered a potential TSS.

All that information was pre-computed and stored in a highly indexed MySQL database. A web-based user interface allows users to submit queries using almost all available GenBank accession IDs (for the supported organisms and referencing an mRNA or EST sequence) to extract promoters of the required genes. Users may request up to 2000 sequences per operation and may specify a large range for the promoter region (10 000 bases upstream of the TSS and 1000 bases downstream). In case of multiple promoters, the user has the choice of extracting all of them or only the one that corresponds to the 3′ most TSS or the 5′ most TSS (representing the most conservative and most aggressive degrees of extension, respectively). Alternatively, the user may choose to extract only the longest extension that is supported by the largest number of sequences in the cluster. If the requested region overlaps with another cluster on the same chromosome that is upstream of the cluster in consideration, the user may wish to ignore this fact or stop extraction at the boundary of the upstream cluster, which can be restricted to the same strand or be on either one.

There are a number of options in promoter extraction. We believe that the choice and information should be passed to the users so that they would have the freedom to decide on the course of action in case of ambiguity. This is in contrast with adopting certain solutions that would inevitably seem inappropriate for the purposes of one user or the other. In the result page, we display a summary table indicating various statistics of each extracted promoter region, e.g. the starting and ending coordinates of the predicted TSS, the quality and size of the cluster to which the promoter belongs, the number of supporting sequences and the extent of genomic extension (in base pairs).

RESULTS

PromoSer is an easy-to-use service for extracting high quality promoter sequences in a batch mode. Without PromoSer, the common approach has usually been to work with a small set of promoters and to extract them manually. RefSeq and LocusLink have often been the resource of choice for these situations, but it requires a fair bit of manual work to trace and extract the genomic region of the promoters. A drawback of this method has been well known but perhaps not fully appreciated. Many sequences in RefSeq are not full length; they are often truncated before the 5′ end of the actual transcript. It has previously been reported (19) that ∼34% of RefSeq mRNA sequences could be extended toward the 5′ direction by 87 bases on average, using alignments with sequences obtained from a full-length mRNA sequence library.

Compared to earlier efforts, PromoSer has compiled a larger and more varied data set. This revealed a surprising sequel to the previous reports. Using PromoSer, we found that at least 63% of human RefSeq entries (58% and 17% for mouse and rat, respectively) could be further extended towards the 5′ end. The TSS position was shifted, on average, by a surprising 16 981 bases upstream on the chromosome. The average number of sequences overlapping an extendable RefSeq sequence was 123. For the mouse genome, the shift averaged 6801 bases and the clusters contained on average 47 sequences. In the case of the rat genome, an average shift of 8880 bases upstream from that identified by aligning a RefSeq sequence alone was found, with on average 14 sequences overlapping an extendable RefSeq entry. Moreover, 17% of human RefSeq, 12% of mouse RefSeq and 4% of rat RefSeq sequences could be extended by at least 2000 bases on the chromosome. Figure 1 provides a histogram of the amount of genomic extension obtained by PromoSer for RefSeq sequences in each organism. It is worth noting, however, that the rat genome is still in an early draft stage and the figures reported throughout the paper are very likely to change as resources for the rat genome mature to the levels of the human and mouse genomes. Even without relying on any EST data, ∼40% of human RefSeq and 46% of mouse RefSeq sequences can be extended towards the 5′ end by utilizing recent full-length mRNA data only.

There are a number of interesting examples in which utilizing clusters of alignments produces an unexpected large change in the TSS location prediction. In one extreme case, MMP26 (matrix metalloproteinase 26, LocusID: 56547) is a well-characterized gene. Figure 2 contains snapshots from the UCSC genome browser that illustrates this gene. The RefSeq entry for MMP26 (NM_021801) is curated and confirmed to be complete only at the 3′ end. PromoSer localized NM_021801 to a cluster on chromosome 11 that also contained four mRNA sequences and several spliced ESTs (Fig. 2A). All these sequences aligned well with NM_021801 in its coding regions (Fig. 2B). Two ESTs, BG189720 and

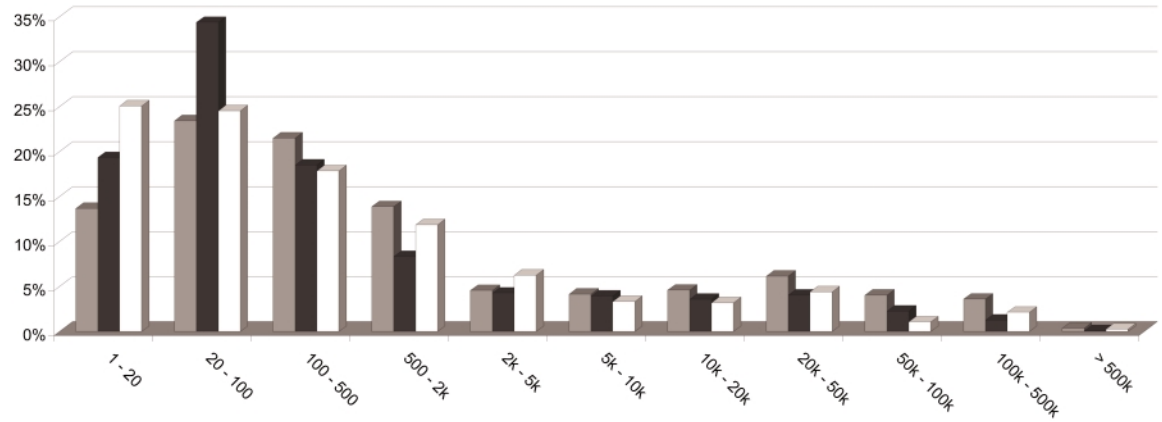

Figure 1. Length distribution of the amount of genomic extensions (in base pairs) obtained for RefSeq sequences. Grey columns are for human, black for mouse and white for rat. The column heights represent the percentage of all RefSeq sequences that could be extended to within the range given.

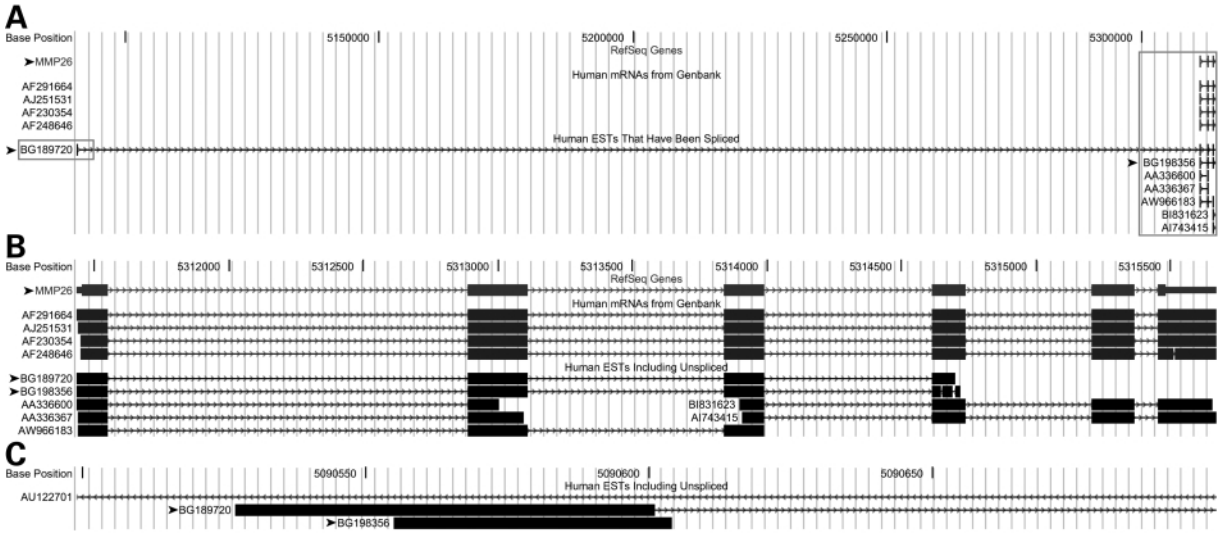

Figure 2. The genomic alignment of MMP26 and the other overlapping sequences. Thin line segments represent introns, with arrowheads indicating the 5′ to 3′ direction. (**A**) The entire alignment, indicating the extension resulting from ESTs BG189720 and BG198356. Two boxes circle the 5′ and 3′ ends of the alignment, with the 3′ end shown in greater detail in (**B**) and the 5′ end in (**C**). Note that the 5′ extension of BG198356 is only shown in (C), since BLAT aligned this EST to the genome in two fragments (see text for more details).

BG198356, extended NM_021801 in the 5′ direction. BG189720 was spliced into five regions, the four regions in the 3′ end aligned well with NM_021801, and the 5′ most region was about 80 bases long and aligned at a position more than 220 000 bases upstream of NM_021801. The other EST BG198356 is 97% identical to BE189720. Interestingly, BLAT aligned BG198356 to the genome in two pieces, even though it is 100% identical to BG189720 at the break point of these two pieces. Thus, the 5′ fragment is classified as an unspliced EST in the UCSC browser (a snapshot is shown in Fig. 2C). We note that BG189720 is a high quality sequence obtained using a random gene activation method (28) that can induce activation patterns not normally observed in the cell lines used. Therefore, it is likely that there is an alternative TSS more than 220 000 bases upstream of the TSS defined using RefSeq. For sequences in this cluster, PromoSer reports two alternative promoters, one according to the RefSeq sequence NM_021801 and the other one defined by the 5′ end of BG189720.

In another example, Figure 3 shows a case where a RefSeq sequence (NM_033543) was extended by more than 26 700 bases upstream using mRNA data, where the mRNA sequence AK023602 overlaps the RefSeq sequence completely and extends it in both the 5′ and 3′ directions. Several ESTs that overlap and cover the entire cluster also support this large 5′ extension.

DISCUSSION

Large-scale computational analysis of transcription regulation is a powerful and promising technology that should provide us

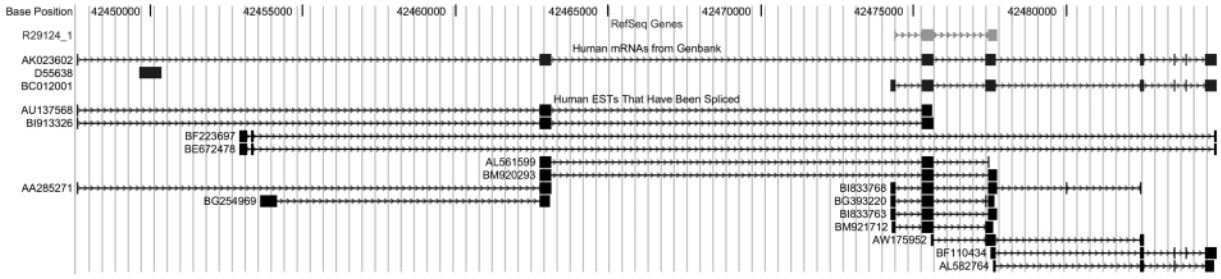

Figure 3. Another example of a long 5′ genomic extension relying only on mRNA data.

with a better understanding of the nature of the intricate network of regulatory proteins that provide living cells with their remarkable properties. Biological results clearly show that the regulation mechanism is a multilevel high complexity system that involves physical, informational, proximal, distal, upstream and downstream *cis*-regulatory elements on the DNA, as well as protein/protein, protein/DNA and protein/small molecule interacting factors. Hopeless as it may look, computational analysis has nevertheless focused on the most obvious and likely the most revealing component of this system, that is, the proximal promoter. Defined operationally as the region immediately upstream of the transcription start site and extending for a few hundred to a couple thousand bases, this region still remains an elusive target for the exact characterization of its structure.

Much has been learned by grouping sets of promoter regions and comparing their sequences to look for shared motifs of short (5–20 bp) *cis*-elements. Both grouping and analysis can each be done in a large number of ways. The data from which groups and clusters can be identified can also come from many sources, most notably and commonly from microarray experiments. Yet, a common element in all of these methods is the need for an accurate set of proximal promoter sequences to analyze. Until recently this task seemed like the bottleneck of the computational pipeline. To the best of our knowledge, resources for accurate large-scale promoter extraction did not exist publicly.

A program previously developed, PEG, was based on computational text analysis of GenBank mRNA records, tracing the annotations and iteratively stitching and composing the promoter using a complex algorithm (29). However, the program requires non-trivial installation on a local UNIX computer, which is not feasible for experimental biologists. A simpler approach (the EZ-Retrieve web server) is based on parsing NCBI UniGene, RefSeq and LocusLink entries for annotated TSSs (30). However, the TSSs for many genes are not annotated, and the mRNA TSS record annotations can be error prone and unreliable. It is known that mRNAs are commonly truncated before the 5′ end of the molecule. This truncation exists even in the highly curated RefSeq (19,31) set of reference sequence data. A java-based tool Toucan allows online genomic sequence retrieval from the Ensembl database (including orthologous sequences) and *cis*-element analysis (32). It assumes the most 5′ position of the annotated transcript to be the TSS and therefore is likely to suffer from similar shortcomings as the above two methods.

PromoSer has set out to fill this gap by providing a publicly accessible and easy to use service that relies on a large data set. By utilizing nearly all major public data sets of full-length cDNA sequence information, PromoSer achieves both coverage and accuracy. We have made a conscious decision not to include TSS annotations in existing databases or *de novo* promoter predictions due to their undetermined accuracy. PromoSer will continue to be updated frequently to utilize the most recent data sets. As more mammalian genomes come off the sequencing pipeline, they will also be incorporated. Future directions for PromoSer include the consideration of cross-species conservation and the ability to localize and identify the TSS for a user supplied set of sequences.

Although the backbone for PromoSer shares much with the UCSC Genome Browser, we decided against using the UCSC alignment information as provided for a number of reasons. Firstly, our dataset consists of a more heterogeneous collection than exists in the Genome Browser and is expanding. If we simply add on to the UCSC set, we risk conflicts due to non-uniform filtering criteria of the two methods that would be difficult to resolve. Secondly, we plan to keep the service up-to-date by frequent updating and incorporation of new datasets as they emerge. Thirdly, a separate server gives us finer control over the quality of alignments that are considered for inferring the TSS. Obviously, PromoSer and the Genome Browser serve different purposes and need not handle things exactly the same way.

The utility of accurately identified promoter sequences is illustrated by the recent work of Ohler *et al.* on the core promoters in the *Drosophila* genome. Several new regulatory motifs have been identified and the computational TSS prediction has also been improved with the more accurate training data (33). Likewise, PromoSer should prove useful in studying mammalian gene regulation, and aid the development of computational tools for promoter prediction (34–37).

ACKNOWLEDGEMENTS

We wish to thank Martin Frith, Peter Haverty and Joel Graber for thought provoking discussions and advice and the sharing of helpful resources. We also would like to thank the reviewers for valuable feedback. This work has been supported in part by NSF grants DBI-0078194 and MRI DBI-0116574 and NIH grant 1P20GM066401-01.

REFERENCES

1. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
2. Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Van de Rijn,M., Waltham,M. *et al.* (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227–235.
3. Chiang,D.Y., Brown,P.O. and Eisen,M.B. (2001). Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, **17**, S49–S55.
4. Ueda,H.R., Chen,W., Adachi,A., Wakamatsu,H., Hayashi,S., Takasugi,T., Nagano,M., Nakahama,K., Suzuki,Y., Sugano,S. *et al.* (2002). A transcription factor response element for gene expression during circadian night. *Nature*, **418**, 534–539.
5. Wu,Q., Kirschmeier,P., Hockenberry,T., Yang,T.Y., Brassard,D.L., Wang,L., McClanahan,T., Black,S., Rizzi,G., Musco,M.L. *et al.* (2002). Transcriptional regulation during p21WAF1/CIP1-induced apoptosis in human ovarian cancer cells. *J. Biol. Chem.*, **277**, 36329–36337.
6. Lawrence,C.E. and Reilly,A.A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
7. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
8. Frith,M.C., Hansen,U. and Weng,Z. (2001). Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.
9. Gasch,A.P. and Eisen,M.B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**, RESEARCH0059.
10. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**, 153–159.
11. Blackwood,E.M. and Kadonaga,J.T. (1998). Going the distance: a current view of enhancer action. *Science*, **281**, 61–63.
12. McKnight,S.L. and Kingsbury,R. (1982). Transcriptional control signals of a eukaryotic protein-coding gene. *Science*, **217**, 316–324.
13. Mitchell,P.J. and Tjian,R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **245**, 371–378.
14. Novina,C.D. and Roy,A.L. (1996). Core promoters and transcriptional control. *Trends Genet.*, **12**, 351–355.
15. Ohler,U. and Niemann,H. (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56–60.
16. Praz,V., Perier,R., Bonnard,C. and Bucher,P. (2002). The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.*, **30**, 322–324.
17. Carninci,P., Kvam,C., Kitamura,A., Ohsumi,T., Okazaki,Y., Itoh,M., Kamiya,M., Shibata,K., Sasaki,N., Izawa,M. *et al.* (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**, 327–336.
18. Das,M., Harvey,I., Chu,L.L., Sinha,M. and Pelletier,J. (2001). Full-length cDNAs: more than just reaching the ends. *Phys. Genom.*, **6**, 57–80.
19. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
20. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
21. Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F. *et al.* (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
22. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
23. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
24. Trinklein,N.D., Aldred,S.J., Saldanha,A.J. and Myers,R.M. (2003). Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.
25. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002). The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
26. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
27. Kent,W.J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
28. Harrington,J.J., Sherf,B., Rundlett,S., Jackson,P.D., Perry,R., Cain,S., Leventhal,C., Thornton,M., Ramachandran,R., Whittington,J. *et al.* (2001). Creation of genome-wide protein expression libraries using random activation of gene expression. *Nat. Biotechnol.*, **19**, 440–445.
29. Zhang,T. and Zhang,M. (2001). Promoter Extraction from GenBank (PEG): automatic extraction of eukaryotic promoter sequences in large sets of genes. *Bioinformatics*, **17**, 1232–1233.
30. Zhang,H., Ramanathan,Y., Soteropoulos,P., Recce,M.L. and Tolias,P.P. (2002). EZ-Retrieve: a web-server for batch retrieval of coordinate-specified human DNA sequences and underscoring putative transcription factor-binding sites. *Nucleic Acids Res.*, **30**, e121.
31. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003). NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
32. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and Moor,B.D. (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.
33. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002). Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, RESEARCH0087-0087.
34. 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.
35. Down,T.A. and Hubbard,T.J. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
36. Hannenhalli,S. and Levy,S. (2001). Promoter prediction in the human genome. *Bioinformatics*, **17** (Suppl. 1), S90–S96.
37. Scherf,M., Klingenhoff,A., Frech,K., Quandt,K., Schneider,R., Grote,K., Frisch,M., Gailus-Durner,V., Seidel,A., Brack-Werner,R. *et al.* (2001). First pass annotation of promoters on human chromosome 22. *Genome Res.*, **11**, 333–340.