OMICS A Journal of Integrative Biology
Volume 9, Number 3, 2005
© Mary Ann Liebert, Inc.

HumanUpstream and *MouseUpstream*: Databases of Promoter Sequences in the Human and Mouse Genomes

IGOR LEYKIN,[1,2] MING-CHIH J. KAO,[1] and WING HUNG WONG[1,2]

ABSTRACT

Large-scale genome annotations, based largely on gene prediction programs, may be inaccurate in their predictions of transcription start sites, so that the identification of promoter regions remains unreliable. Here we focus on the identification of reliable gene promoter regions, critical to the understanding of transcriptional regulation. We report the construction of databases of upstream sequences *HumanUpstream* and *MouseUpstream* based on information from both the human and mouse genomes and the database of expressed sequence tags (dbEST). Using the ENSEMBL generic genome annotation system, our approach allows more reliable identification of transcript start sites, and therefore extraction of more reliable promoters regions. The *HumanUpstream* and *MouseUpstream* databases are available free of charge.

INTRODUCTION

A COMMON AND FUNDAMENTAL PROBLEM in promoter analysis is the determination of transcription start sites. Despite the assembly of the human and mouse genomes, this remains an unsolved problem. Large-scale genome annotations are based largely on gene prediction programs, which may not predict full-length transcripts, so that the derived transcription start sites may be inaccurate (Liang et al., 2000; Quackenbush et al., 2000; Pertea et al., 2003; Zhu et al., 2003; Wasmuth and Blaxter, 2004; Foissac and Schiex, 2005; Hubbard et al., 2005; Lee et al., 2005). Since proximal promoter elements are of indispensable importance to transcription regulation, even small errors in the prediction of transcription start sites may eventually mislead promoter analyses. We focus on the identification of reliable gene promoter regions, a task critical to the understanding of transcriptional regulation. In our approach, we defined the transcription start sites of human and mouse genes based on the Ensembl EST GeneBuilder, which maps ESTs to the genome and then processes by merging the redundant ESTs and setting splice-sites to the most common ends (Clamp et al., 2003; Birney et al., 2004; Eyras et al., 2004; Hammond and Birney, 2004; Hubbard et al., 2005). ENSEMBLE EST Genes database contains resulting full-length transcripts for varies species. We then selected a subset of human and mouse 5′-reliable transcripts using stringent criteria for more accurate transcription start site predictions and then extracted upstream sequence for such transcripts. This approach allows us to extract more reliable gene promoter regions, compared to various genome annotation systems.

[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts.
[2]Department of Biostatistical Science, Dana-Farber Cancer Institute, Boston, Massachusetts.

We are reporting here the construction of *HumanUpstream* and *MouseUpstream*, databases of promoter sequences that should be useful for understanding the transcriptional regulation of genes. Whenever there is evidence for a set of genes to be co-regulated, one can attempt to analyze the corresponding promoter regions to identify binding sites (and their combinations) using the weight matrices of known transcription factor binding sites (Frech et al., 1993, 1998; Chen et al., 1995; Quandt et al., 1995; Prestridge, 1996; Werner, 1999, 2003; Scherf et al., 2000; Kielbasa et al., 2001; Aerts et al., 2003, 2004; Philippakis et al., 2005; Xing and van der Laan, 2005), or to discover over-represented sequence motifs as candidates of novel binding sites (Lawrence et al., 1993; Hughes et al., 2000; McGuire et al., 2000; Liu et al., 2001; Sinha and Tompa, 2003; Moses et al., 2004; Sinha et al., 2004; Thompson et al., 2004; Down and Hubbard, 2005; Nix and Eisen, 2005; Tan et al., 2005; Tompa et al., 2005).

MATERIALS AND METHODS

To extract ENSEMBL EST Genes and their upstream sequences, we used EnsMart–toolset for retrieving customized data sets from annotated genomes (Hammond and Birney, 2004; Kasprzyk et al., 2004; ⟨http://www.ensembl.org/Multi/martview⟩), which generates a number of different types of output, including sequence data. Various output formats, including HTML, text, and Microsoft Excel, are supported. We collected 59,820 human and 46,646 mouse EST Gene sequences and their upstream region 5 KB of length in FASTA format. To identify EST Genes that have reliable 5′ ends, we used BLAST to align them against the latest collection of human and mouse ESTs from NCBI's dbEST database. This database is updated on daily basis, at the time of analysis dbEST contained 4,806,831 human and 2,733,310 mouse ESTs. All EST sequences were used as a query, and 59,820 human and 46,646 mouse EST Gene sequences served as a database for BLAST alignment. All perfect matches at 5′ end of EST Genes were analyzed. Only EST Genes with at least one such perfect match were selected and further used.

To annotate selected EST Genes and, therefore, their promoter sequences using various resources we mapped EST Genes sequences to Affymetrix probe sequences (⟨http://www.affymetrix.com/support/technical/byproduct.affx?cat=arrays⟩) and created Affymetrix probe set ID—ENSEMBL EST Gene pairs for four commonly used Affymetrix GeneChip Arrays–human U95 and U133, and mouse U74 and MOE430. Mapping was done by direct matching of smaller sequence to larger one. Then we extended these links using the TIGR's RESOURCERER, which provides comprehensive annotation for all Affymetrix GeneChip Arrays. (Tsai et al., 2001; ⟨http://www.tigr.org/tigr-scripts/magic/r1.pl⟩). To each ENSEMBL EST Gene–Affymetrix probe set pair we added Genbank Acc, UniGene ID, RefSeq and LocusLink ID related to this Affymetrix probe set (Table 1).

Table 1. Annotation for Some Mouse EST Genes Created with Affymetrix NetAffx Data and RESOURCERER from the Institute for Genomic Research

Probe ID	*Genbank accession*	*UniGene ID*	*RefSeq Acc*	*LL ID*	*EST gene*
1451547_at	BC023358	Mm.24153	NM_027391	70337	>ENSMUSESTT00000000014
1453021_at	BM899291	Mm.286868	NM_030191	78808	>ENSMUSESTT00000000018
1423654_a_at	AF169300	Mm.21281	NM_011278	19822	>ENSMUSESTT00000000022
1422156_a_at	NM_008503	Mm.299566	NM_008503	16898	>ENSMUSESTT00000000024
1452072_at	W34301	Mm.33762	NM_026793	68632	>ENSMUSESTT00000000028
1428664_at	AK018599	Mm.98916	NM_011702	22353	>ENSMUSESTT00000000029
1417057_a_at	BC011499	Mm.295252	NM_026352	67738	>ENSMUSESTT00000000031
1438025_at	BE686616	Mm.55082	NM_175374	108853	>ENSMUSESTT00000000034
1421495_a_at	NM_024261	Mm.346733	NM_024261	73419	>ENSMUSESTT00000000035
1421495_a_at	NM_024261	Mm.346733	NM_024261	73419	>ENSMUSESTT00000000036

RESULTS

Extraction and verification of promoter sequences

ENSEMBL human and mouse EST Genes are assemblies of human and mouse ESTs (Clamp et al., 2003; Eyras et al., 2004; Hammond and Birney, 2004), and in many cases represent full-length transcripts. In some cases, EST genes also may contain full or partial cDNA sequences. 5′-ESTs are supposed to start directly on Transcription Start Site, so all transcripts with such a match or matches are very likely have correct 5′-end, what allow us to select and further use correct promoter sequences. We have used 5′-ESTs to verify quality of all human and mouse EST Genes by BLAST alignment. Only EST Genes showing perfect match of their 5′ end to at least one 5′ EST were further selected. From 59,820 human and 46,646 mouse EST genes, 34781 human and 31622 mouse EST Genes were found 5′-reliable, and their 5 KB upstream sequences were added to the set of promoter sequences organized as relational databases *HumanUpstream* and *MouseUpstream*. From them, 19097 (31.9%) human and 18862 (40.4%) mouse sequences have only one match to 5′ EST, 5742 (9.6%) human and 5563 (11.9%) mouse sequences have two matches, and 9942 (16.6%) human and 7191 (15.4%) mouse sequences have three or more such matches. Each promoter sequence in the database is marked with *, ** or *** in accordance with number of perfect matches to 5′ ESTs, as an indicator of the reliability of this promoter.

Annotation of promoter sequences

ENSEMBL EST Gene promoter sequences can be effectively utilized by various tools and methods, including computational prediction of organization of their promoter elements. However, ENSEMBL EST Genes database is not linked to any other resource, such as GenBank, LocusLink, UniGene, and RefSeq. We provided a comprehensive annotation for most of human and mouse EST Genes and their promoter sequences by linking them to individual probe sets from commonly used Affymetrix Gene Expression Analysis Arrays. We created Affymetrix probe set ID–ENSEMBL EST Gene pairs for 22,365 U74 probe sets, 20,248 MOE430 probe sets, 40,089 U95 probe sets, and 31,050 U133 probe sets. Such pairs then allowed us to also link most of EST Genes and their promoters to other identifiers, using publicly available the TIGR's RESOURCERER, which integrates information from various resources from the same species. Resulting tables contain Affymetrix probe set IDs, Genbank accession numbers, UniGene, RefSeq and LocusLink IDs, as well as corresponding ENSEMBL EST Gene IDs (Table 1). If necessary any other identifiers can be also linked to this data set.

CONCLUSION

Thousands of genes must be coherently differentially expressed in different cell types in response to extracellular and intracellular signals. Such diverse expression patterns can only be the result of combinatorial control of gene regulation since only a limited number of transcription factors are encoded by genome. Despite the assembly of several genomes, the identification of gene promoter regions remains unreliable. Large-scale genome annotations are based largely on gene prediction programs. A limitation of these annotations is that transcription start sites and, therefore, promoter sequences derived from them might be inaccurate. We have developed methods for an identification of gene promoter regions for various species using stringent criteria for more accurate transcription start site predictions. Compared to various genome annotation systems, our approach allows more reliable identification of transcription start sites, and therefore extraction of more reliable promoters regions. Using these methods we are able to derive high-quality promoter sequences for groups of co-regulated or functionally related genes. This allows us to define statistically significant combinations of transcription factors involved in their specific expression patterns. To determine their biological significances, these combinations should be verified experimentally.

DATA SETS

The *HumanUpstream* and *MouseUpstream* databases contain promoter sequences 5 KB of length in FASTA format for 34,781 human and 31,622 mouse EST Genes. They are available free of charge at ⟨http://www.biostat.harvard.edu/~ileykin/upstream.tar.gz⟩. The archive also contains README file and four annotation files for Affymetrix U74, MOE430, U95, and U133 probe sets (Table 1).

ACKNOWLEDGMENTS

W.H.W. was supported by NSF grants DMS-0090166 and DBI-9904701/0196176), and by The Dana Foundation. M.-C.J.K. was supported by the Howard Hughes pre-doctoral fellowship. I.L. was supported by NIH grant P20-CA 96970 and by NSF grant DMS-0090166.

REFERENCES

AERTS, S., THIJS, G., COESSENS, B., et al. (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res **31**, 1753–1764.

AERTS, S., VAN LOO, P., MOREAU, Y., et al. (2004). A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. Bioinformatics **20**, 1974–1976.

BIRNEY, E., ANDREWS, T.D., BEVAN, P., et al. (2004). An overview of Ensembl. Genome Res **14**, 925–928.

CHEN, Q.K., HERTZ, G.Z., and STORMO, G.D. (1995). MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput Appl Biosci **11**, 563–566.

CLAMP, M., ANDREWS, D., BARKER, D., et al. (2003). Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res **31**, 38–42.

DOWN, T.A., and HUBBARD, T.J. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res **33**, 1445–1453.

EYRAS, E., CACCAMO, M., CURWEN, V., et al. (2004). ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res **14**, 976–987.

FOISSAC, S., and SCHIEX, T. (2005). Integrating alternative splicing detection into gene prediction. BMC Bioinformatics **10**, 25.

FRECH, K., HERRMANN, G., and WERNER, T. (1993). Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. Nucleic Acids Res **21**, 1655–1664.

FRECH, K., QUANDT, K., and WERNER, T. (1998). Muscle actin genes: a first step towards computational classification of tissue specific promoters. In Silico Biol **1**, 29–38.

HAMMOND, M.P., and BIRNEY, E. (2004). Genome information resources—developments at Ensembl. Trends Genet **20**, 268–272.

HUBBARD, T., ANDREWS, D., CACCAMO, M., et al. (2005). Ensembl 2005. Nucleic Acids Res **33**, D447–D453.

HUGHES, J.D., ESTEP, P.W., TAVAZOIE, S., et al. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J Mol Biol **296**, 1205–1214.

KASPRZYK, A., KEEFE, D., SMEDLEY, D., et al. (2004). EnsMart: a generic system for fast and flexible access to biological data. Genome Res **14**, 160–169.

KIELBASA, S.M., KORBEL, J.O., BEULE, D., et al. (2001). Combining frequency and positional information to predict transcription factor binding sites. Bioinformatics **17**, 1019–1026.

LAWRENCE, C.E., ALTSCHUL, S.F., BOGUSKI, M.S., et al. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262**, 208–214.

LEE, Y., TSAI, J., SUNKARA, S., et al. (2005). The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res **33**, D717–D714.

LIANG, F., HOLT, I., PERTEA, G., et al. (2000). An optimized protocol for analysis of EST sequences. Nucleic Acids Res **28**, 3657–3665.

LIU, X., BRUTLAG, D.L., and LIU, J.S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput 127–138.

McGUIRE, A.M., HUGHES, J.D., and CHURCH, G.M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res **10**, 744–757.

MOSES, A.M., CHIANG, D.Y., POLLARD, D.A., et al. (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol **25**, R98.

NIX, D.A., and EISEN, M.B. (2005). GATA: a graphic alignment tool for comparative sequence analysis. BMC Bioinformatics **6**, 9.

PERTEA, G., HUANG, X., LIANG, F., et al. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19**, 651–652.

PHILIPPAKIS, A.A., HE, F.S., and BULYK, M.L. (2005). Modulefinder: a tool for computational discovery of cis regulatory modules. Pac Symp Biocomput 519–530.

PRESTRIDGE, D.S. (1996). SIGNAL SCAN 4.0: additional databases and sequence formats. Comput Appl Biosci **12**, 157–160.

QUACKENBUSH, J., LIANG, F., HOLT, I., et al. (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res **28**, 141–145.

QUANDT, K., FRECH, K., KARAS, H., et al. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res **23**, 4878–4884.

SCHERF, M., KLINGENHOFF, A., and WERNER, T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analyses approach. J Mol Biol **297**, 599–606.

SINHA, S., and TOMPA, M. (2003). YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res **31**, 3586–3588.

SINHA, S., BLANCHETTE, M. and TOMPA, M. (2004). PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics **5**, 170.

TAN, K., McCUE, L.A., and STORMO, G.D. (2005). Making connections between novel transcription factors and their DNA motifs. Genome Res **15**, 312–320.

THOMPSON, W., PALUMBO, M.J., WASSERMAN, W.W., et al. (2004). Decoding human regulatory circuits. Genome Res **14**, 1967–1974.

TOMPA, M., LI, N., BAILEY, T.L., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol **23**, 137–144.

TSAI, J., SULTANA, R., LEE, Y., et al. (2001). RESOURCERER: a database for annotating and linking microarray resources within and across species. Genome Biol **2,** software, 0002.1–0002.4.

WASMUTH, J.D., and BLAXTER, M.L. (2004). Prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics **5**, 187.

WERNER, T. (1999). Models for prediction and recognition of ukaryotic promoters. Mamm Genome **10**, 168–175.

WERNER, T., FESSELE, S., MAIER, H., et al. (2003). Computer modeling of promoter organization as a tool to study transcriptional coregulation. FASEB J **17**, 1228–1237.

XING, B., and VAN DER LAAN, M.J. (2005). A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data. J Comput Biol **12**, 229–246.

ZHU, Y., KING, B.L., PARVIZI, B., et al. (2003). Integrating computationally assembled mouse transcript sequences with the Mouse Genome Informatics (MGI) database. Genome Biol **4**, R16.

Address reprint requests to:
Dr. Igor Leykin
Joslin Diabetes Center
One Joslin Pl.
Boston, MA 02215

E-mail: Igor.Leykin@joslin.harvard.edu