Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

A comprehensive study of bidirectional promoters in the human genome

Mary Qu Yang
National Human Genome Research Institute
National Institutes of Health
yangma@mail.nih.gov

Laura L. Elnitski
Head, Genomic Functional Analysis
National Human Genome Research Institute
National Institutes of Health

***Abstract*—A *bidirectional promoter* is a region along a strand of DNA that regulates the expression of two genes flanking the region. Each of these genes is transcibed in a direction that points away from the other gene; two such genes are said to be in a *head-to-head* configuration. We search the UCSC List of Known Genes and GenBank Expressed Sequence Tag (EST) data for pairs of genes in such a configuration in order to identify new bidirectional promoters.**

The EST data constitutes a larger and more intricate dataset that the UCSC List of Known Genes. However, working with EST data presents a challenge, as the EST database may be highly redundant and may also contain overlapping ESTs. To deal with these problems, we have developed an algorithm to identify bidirectional promoters based on the above data sources; the algorithm is capable of handling redundant ESTs, and also ESTs that overlap or disagree in orientation.

This analysis resulted in the identification of thousands of new candidate head-to-head gene pairs, corroborated the 5' ends of many known human genes, revealed new 5' exons of previously characterized genes, and in some cases identified novel genes. Further analyses yielded evidence for coordinate expression of genes in a head-to-head configuration, and examined the prevalence of bidirectional promoters in different biological pathways.

I. Introduction

The mechanisms by which gene expression is regulated in the human genome are as yet not well-understood; it would greatly aid our understanding to be able to pin down prospective regulatory regions. It turns out that candidate regulatory regions can be identified by searching for genes arranged in a "head-to-head" configuration. Recall that a gene has a 5' end and a 3' end; in general, genes are transcribed in the $5' \rightarrow 3'$ direction ("downstream") by an RNA polymerase. The site where the RNA polymerase initially binds is a region of the DNA called a *promoter*; since transcription generally proceeds in the $5' \rightarrow 3'$ direction, the promoter must be located upstream of the 5' end of the gene. Two genes that have their 5' ends located fairly close together, say, within 1000 base pairs, and furthermore are transcribed in opposite directions are said to be in a *head-to-head*. configuration. The significance of this configuration is that it is likely that one or more regulatory regions will be located in the stretch between the 5' end of one gene and the 5' end of the other. This stretch is known as a *bidirectional promoter*[1], because it likely serves as the promoter for two genes that are transcribed in opposite directions.

Bidirectional promoters are abundant in the human genome [10], and help to regulate DNA repair, non-DNA housekeeping functions, and other processes. Most early instances of bidirectional promoters were discovered in the course of investigating individual genes [10], but recent computational searches by Adachi and Lieber [1] and Trinklein et al. [10] resulted in a substantial increase in the number of known bidirectional promoters. The data for these searches derives from several sources, including:

- the UCSC List of Known Genes [6]; a list of protein coding genes based on rigorous supporting evidence from UniProt and GenBank mRNA.
- Refseq mRNA data.
- Spliced Expressed Sequence Tags (ESTs) [3], which are short DNA sequences (usually 200-500 base pairs) obtained by sequencing one or both ends of a transcript of an expressed gene.

The EST data constitutes a larger and more intricate dataset than the UCSC List of Known Genes. However, working with EST data presents a challenge, as the EST database may be highly redundant and may also contain overlapping ESTs. To deal with these problems, we have developed an algorithm to identify bidirectional promoters based on the above data sources; the algorithm is capable of handling redundant ESTs, and also ESTs that overlap or disagree in orientation. The algorithm combines data from the three sources, so that if there is not sufficient evidence to conclude that a candidate region is in fact a bidirectional promoter based on EST data alone, it looks for supporting evidence by examining Known Gene and mRNA data.

This analysis resulted in the identification of thousands of new candidate bidirectional promoters, corroborated the 5' ends of many known human genes, revealed new 5' exons of

[1]Technically, the region that lies between two genes arranged in a head-to-head configuration should not be called a bidirectional promoter until it has been shown experimentally to regulate both genes, so we should really call it a *candidate bidirectional promoter*. However, to simplify the terminology, here we use the term *candidate bidirectional promoter* to denote a region between two genes in a head-to-head configuration, and the term *bidirectional promoter* to denote a region between two genes in a head-to-head configuration that also satisfies the various conditions imposed by the algorithm

1-4244-0710-9/07/$20.00 ©2007 IEEE

previously characterized genes, and in some cases identified novel genes. The fact that our algorithm extracts significantly more bidirectional promoters than were previously known raises the question as to whether these are in fact valid promoter regions; we provide supporting evidence to show that this is indeed the case. Further analyses yielded evidence for coordinate expression of genes in a head-to-head configuration, and examined the prevalence of bidirectional promoters in different biological pathways.

II. Data and Algorithm

The data that we use derives from 3 sources:

- The UCSC List of Known Genes [6].
- GenBank mRNA data [3].
- Spliced EST data from the GenBank dbEST database [3].

The algorithm for extracting bidirectional promoters is as follows:

I. **Known Gene Analysis:** Known Genes that overlap and have the same orientation are clustered; these clusters are defined by the furthest 3' and 5' ends of any gene in the cluster. The region between the 5' ends of two gene clusters is classified as a bidirectional promoter if the following conditions are satisfied:
   - The 5' ends of the two gene clusters are adjacent to one another, and the two arrows that define the $5' \rightarrow 3'$ direction for each gene cluster point away from each other.
   - The 5' ends of the two gene clusters are separated by no more than 1000 base pairs.
   - There are no other gene clusters between the 5' ends of the two gene clusters.

II. **EST Analysis:** ESTs were assessed for confidence in their orientation using the "ESTOrientInfo" table from the UCSC Genome Browser, which gives a measure of reliability of the orientation of the EST based on all overlapping transcripts from the region. Those with no score were excluded due to low confidence in their orientation. Once the orientation was confirmed, all ESTs were compared to the "intronEST" table to verify agreement; this table lists the intronic orientation for each intron of a spliced EST based on the presence of consensus splice sites.

   ESTs that overlap and have the same orientation are then clustered; these clusters are defined by the furthest 3' and 5' ends of any gene in the cluster. Candidate bidirectional promoter regions are formed by pairing an EST cluster with either another EST cluster or a Known Gene cluster, such that the two clusters are in a head-to-head configuration. The candidate bidirectional promoter is rejected if the two clusters overlap, or if the 5' ends of the two clusters are separated by more than 1000 base pairs.

   The candidate bidirectional promoters are then classified using a decision tree, as shown in Figure 3. The tree either rejects the candidate bidirectional promoter, or assigns it a class label "EST-Li", where i is an integer between 1 and 10. To streamline the notation, in the sequel we truncate the leading "EST-L" from the class label, so that the class label is just an integer between 1 and 10. The class label carries two pieces of information:
   - It gives a confidence level that the candidate is in fact a bidirectional promoter. The confidence level is an integer between 1 and 5, where 1 represents the lowest confidence level and 5 the highest. The confidence level can be obtained from the class label via:
   $$\text{confidence level} = 5 - \left\lfloor \frac{\text{class label} - 1}{2} \right\rfloor$$
   - It indicates whether the candidate bidirectional promoter is contained within a Known Gene or not. Odd-numbered class labels indicate that the candidate bidirectional promoter is contained within a Known Gene whereas even-numbered class labels indicate that it is not.

The classification proceeds as follows:

1. Candidate bidirectional promoters enter at the top of the tree in Figure 3. If the candidate bidirectional promoter is contained within a Known Gene, and there exists base pairs of the candidate bidirectional promoter that are more than 1000 base pairs away from the 5' end of the Known Gene in which the candidate bidirectional promoter is contained, then the candidate bidirectional promoter is rejected, otherwise we proceed Step 2 (the next level of the tree).
2. If the EST cluster(s) flanking the candidate bidirectional promoter satisfy the condition that the number of ESTs that overlap the cluster and disagree in orientation with the cluster is smaller than the number of ESTs comprising the cluster, then we say there is "majority agreement in orientation", and the candidate bidirectional promoter is classified to class 1 or class 2, depending on whether it is contained within a Known Gene. Otherwise we proceed Step 3 (the next level of the tree).
3. If the EST cluster(s) flanking the candidate bidirectional promoter satisfy the condition that, after disregarding ESTs that disagree in orientation with the cluster and overlap the 5' end of the cluster by no more than 1000 base pairs, the number of ESTs that overlap the cluster and disagree in orientation with the cluster is smaller than the number of ESTs comprising the cluster, then we say there is "majority agreement after excluding 5' overlap", and the candidate bidirectional promoter is classified to class 3 or class 4, depending on whether it is contained within a Known Gene. Otherwise we proceed Step 4 (the next level of the tree).
4. If the EST cluster(s) flanking the candidate bidirectional promoter agree in orientation with Known

Genes and mRNA transcripts, then the candidate bidirectional promoter is classified to class 5 or class 6, depending on whether it is contained within a Known Gene. Otherwise we proceed Step 5 (the next level of the tree).

5. If the EST cluster(s) flanking the candidate bidirectional promoter agree in orientation with Known Genes, then the candidate bidirectional promoter is classified to class 7 or class 8, depending on whether it is contained within a Known Gene. Otherwise we proceed Step 6 (the next level of the tree).
6. If the EST cluster(s) flanking the candidate bidirectional promoter exhibit majority agreement (as in Step 2) after excluding ESTs that overlap the 3' end of the cluster, then the candidate bidirectional promoter is classified to class 9 or class 10, depending on whether it is contained within a Known Gene. Otherwise the candidate bidirectional promoter is rejected.

The set of bidirectional promoters extracted by the algorithm consists of those extracted in Step I, which are precisely those flanked by Known Gene clusters, along with those extracted in Step II, which are precisely those flanked by either by two EST clusters, or by one EST cluster and one Known Gene cluster, and furthermore are not rejected by the decision tree.

Evidence that the extracted regions indeed serve as promoters can be obtained by looking for two features in the extracted regions that are associated with promoters:

- The presence of experimentally validated TAF250 (or TAF1) binding sites: a high percentage of TAF250 binding sites coincide with other markers of promoter regions [8].
- The presence of CpG islands.

For extracted regions that are flanked by two Known Gene clusters (those extracted in Step I), 74% overlapped a valid TAF250 binding site and 90% overlapped a CpG island, whereas for extracted regions that are flanked either by two EST clusters or by one EST cluster and one Known Gene cluster (those extracted in Step II), 52% overlapped a valid TAF250 binding site and 70% overlapped a CpG island. A summary of the percentages of extracted regions with valid TAF250 binding sites and/or CpG islands for each class is given in Table I.

Evidence that the extracted regions are in fact *bidirectional* promoters was obtained by dividing the extracted region in half and looking for experimentally validated TAF250 binding sites in each half. The last column of Table I gives the percentage of extracted regions with TAF250 binding sites in each half.

III. Results and Discussion

The algorithm identified 1006 bidirectional promoters flanked by two Known Genes, and 159 bidirectional promoters flanked by one Known Gene and one EST cluster (this situation is illustrated in Figure 4(c)). Of 5,575 candidate bidirectional promoters flanked by two EST clusters, 2,105 were rejected by the algorithm. Of the remaining 3,470 identified bidirectional promoters:

- 2,876 were supported by downstream sequences overlapping additional ESTs, mRNA or Known Gene data.
- 594 were located in Known Genes; these alternative promoters direct transcription of both a shorter form of the gene G in which they are embedded and a gene that has the opposite orientation to that of G (this situation is illustrated in Figure 4(b)).

A. Identification of Novel Genes and Exons

For each Known Gene G that is not in a head-to-head configuration with another Known Gene, let E be the closest EST to G that is in a head-to-head configuration with G. If the 5' end of E is no more than 1000 base pairs away from the 5' end of G, then:

- If E overlaps a downstream Known Gene G_2 having the same orientation as E, then E is considered to be an extension of the 5' end of G_2.
- If E overlaps a downstream gene G_2 having the opposite orientation to E, then E is considered to be a novel gene.
- If E does not overlap any Known Gene, but one or more downstream Known Genes have the same orientation as E, then E could either be a 5' extension or a novel gene (this situation is illustrated in Figure 4(f)). These ESTs require further investigation as well as experimental verification to determine if they represent 5' extensions or novel genes.

New functional elements identified in this analysis included novel 5' exons for characterized human genes (this situation is illustrated in Figure 4(d)). For instance, the EST AW169946 extended the 5' end of gene AK094318 by 144,000 base pairs to create a new transcription initiation site adjacent to the neighboring gene, AK125085.

In addition to extension of characterized genes, this analysis identified novel transcripts. These transcripts were absent from the List of Known Gene annotations and therefore were only detected by the EST analysis (this situation is illustrated in Figure 4(e)). These transcripts were spliced, however their protein-coding potential was not always obvious.

Of the 3,470 pairings of EST clusters in a head-to-head configuration, 40% represented extensions of the 5' ends of Known Genes and 43% represented novel transcripts. ESTs that confirmed the 5' ends of Known Genes were abundant (this situation is illustrated in Figure 4(g)).

B. Localization of regulatory intervals

The abundance of Known Genes whose 5' ends were extended by the EST analysis indicated that in many cases augmenting the Known Gene data with EST data resulted in narrower, more localized bidirectional promoter regions. To compare the widths of the bidirectional promoter regions extracted in the Known Gene analysis and in the EST analysis, the percentiles of the widths of the bidirectional promoter regions extracted in the Known Gene analysis and in the EST analysis are shown in Figure 2. The curve corresponding to

the EST analysis lies below that for the Known Gene analysis, indicating that the EST analysis resulted in narrower, more localized bidirectional promoter regions than the Known Gene analysis; 80% of the bidirectional promoters identified by the EST analysis were 300 base pairs or less, whereas 80% of the bidirectional promoters identified by the Known Gene analysis were 550 base pairs or less.

C. Coordinately-regulated expression groups

We looked for evidence of common regulatory patterns revealed by microarray expression profiles among 16,078 Known Genes. For each Known Gene, a cluster was formed consisting of that Known Gene, along with the 500 Known Genes with the most similar co-expression profiles according to the GNF expression data [9]. The association rate, defined as the proportion of genes in the same cluster that are regulated by bidirectional promoters, was then calculated for each cluster; it ranged from a low of .16 to a high of .56. A histogram of the association rates, shown in Figure 1, reveals a bimodal distribution. Genes with the highest rates clustered with other genes regulated by bidirectional promoters at a ratio of 2:1. The difference between the clusters obtained and those that would be expected by chance was statistically significant. Thus there was strong evidence of coordinated expression among subsets of genes in a head-to-head configuration.

D. Prevalence of bidirectional promoters in biological pathways

Bidirectional promoters are known to regulate a few categories of genes [1], [12]. Using the 26 biological pathway genes from the Reactome project [7] we examined additional biological categories for enrichment of bidirectional promoters. Compared to the human genome average in which 31% of genes contained bidirectional promoters, 13 Reactome pathways had a ratio of bidirectional promoters significantly larger than 31%, as shown in Figure 5. For example, the percentage of bidirectional promoters in the Influenza, HIV infection, and DNA repair pathways were respectively 48%, 42%, and 40%; these values yielded respective p-values of 0.04, 0.04, and 0.09 in a Chi-square test, indicating a statistically significant enrichment of bidirectional promoters as compared to the genome average. These results suggest that bidirectional promoters could provide potential therapeutic targets for disease intervention.

ACKNOWLEDGEMENTS

We gratefully acknowledge discussions with faculty of National Human Genome Research Institute for improvement of this manuscript. This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

REFERENCES

[1] N. Adachi and M. R. Lieber. Bidirectional gene organization: a common architectural feature of the human genome. *Cell*, 109(7):807–9, June 2002.

[2] Tim Beissbarth and Terence P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[3] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. GenBank: update. *Nucl. Acids Res.*, 32:D23–26, 2004.

[4] Chingfer Chen, Andrew J. Gentles, Jerzy Jurka, and Samuel Karlin. Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *PNAS*, 99(5):2930–2935, 2002.

[5] Sara J. Cooper, Nathan D. Trinklein, Elizabeth D. Anton, Loan Nguyen, and Richard M. Myers. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, 16(1):1–10, 2006.

[6] Fan Hsu, W. James Kent, Hiram Clawson, Robert M. Kuhn, Mark Diekhans, and David Haussler. The UCSC Known Genes. *Bioinformatics*, 22(9):1036–1046, 2006.

[7] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucl. Acids Res.*, 33:D428–432, 2005.

[8] Tae Hoon Kim, Leah O. Barrera, Ming Zheng, Chunxu Qu, Michael A. Singer, Todd A. Richmond, Yingnian Wu, Roland D. Green, and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436:876–880, August 2005.

[9] Andrew I. Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P. Cooke, John R. Walker, and John B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101(16):6062–6067, 2004.

[10] Nathan D. Trinklein, Shelley Force Aldred, Sara J. Hartman, Diane I. Schroeder, Robert P. Otillar, and Richard M. Myers. An Abundance of Bidirectional Promoters in the Human Genome. *Genome Res.*, 14(1):62–66, 2004.

[11] Nathan D. Trinklein, Shelley J. Force Aldred, Alok J. Saldanha, and Richard M. Myers. Identification and Functional Analysis of Human Transcriptional Promoters. *Genome Res.*, 13(2):308–312, 2003.

[12] Quan Zhao, Jianghui Wang, Ilya V. Levichkin, Stan Stasinopoulos, Michael T. Ryan, and Nicholas J. Hoogenraad. A mitochondrial specific stress response in mammalian cells. *The EMBO Journal*, 21:4411–19, 2002.

TABLE I
VERIFICATION OF REGULATORY REGIONS BY TAF250 AND CPG OVERLAP.

Leaf	Gene Pairs	Valid Taf250 (%)	CpG island (%)	Dual TAF250 (%)
K.G	1,006	74.55	90.15	71.37
EST-L1	2,083	53.77	72.11	50.17
EST-L2	240	50.83	80.41	49.58
EST-L3	225	50.22	73.78	47.11
EST-L4	173	61.85	79.19	58.96
EST-L5	184	37.50	47.80	35.32
EST-L6	103	61.17	67.96	57.28
EST-L7	21	42.86	66.67	33.33
EST-L8	24	29.16	58.33	0.25
EST-L9	363	66.92	83.27	60.84
EST-L10	54	53.70	74.07	48.15
Overall (EST)	3,470	52.30	70.40	48.85

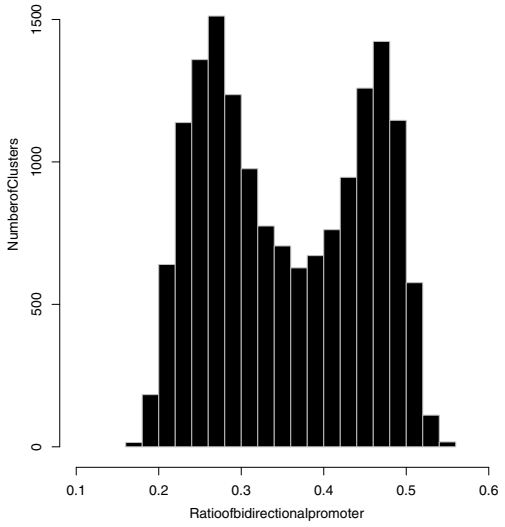

Fig. 1. Histogram of proportions of genes with bidirectional promoters after clustering genes with similar expression profiles.

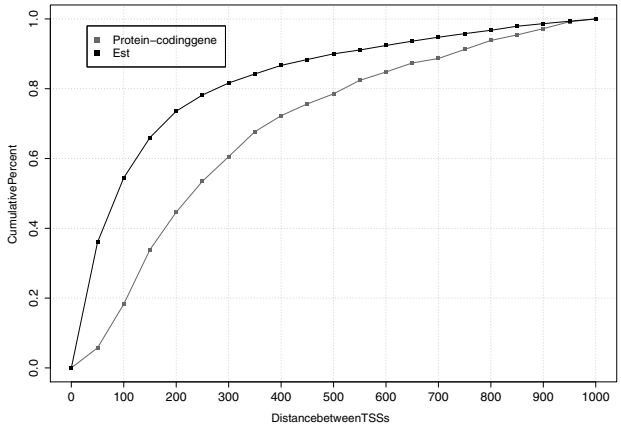

Fig. 2. Percentiles of the widths of the bidirectional promoter regions extracted in the Known Gene analysis and in the EST analysis

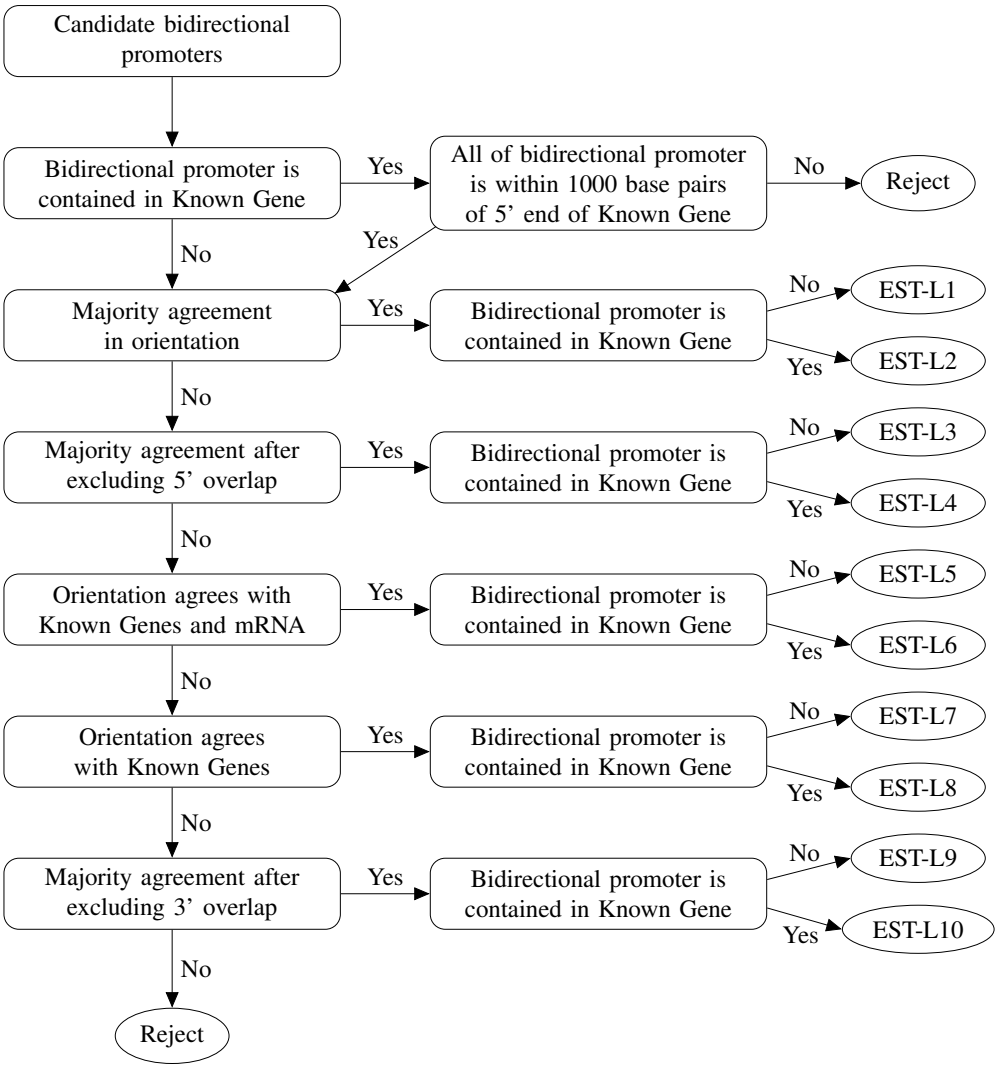

Fig. 3. Decision tree for classifying candidate bidirectional promoter regions flanked by at least one EST cluster.

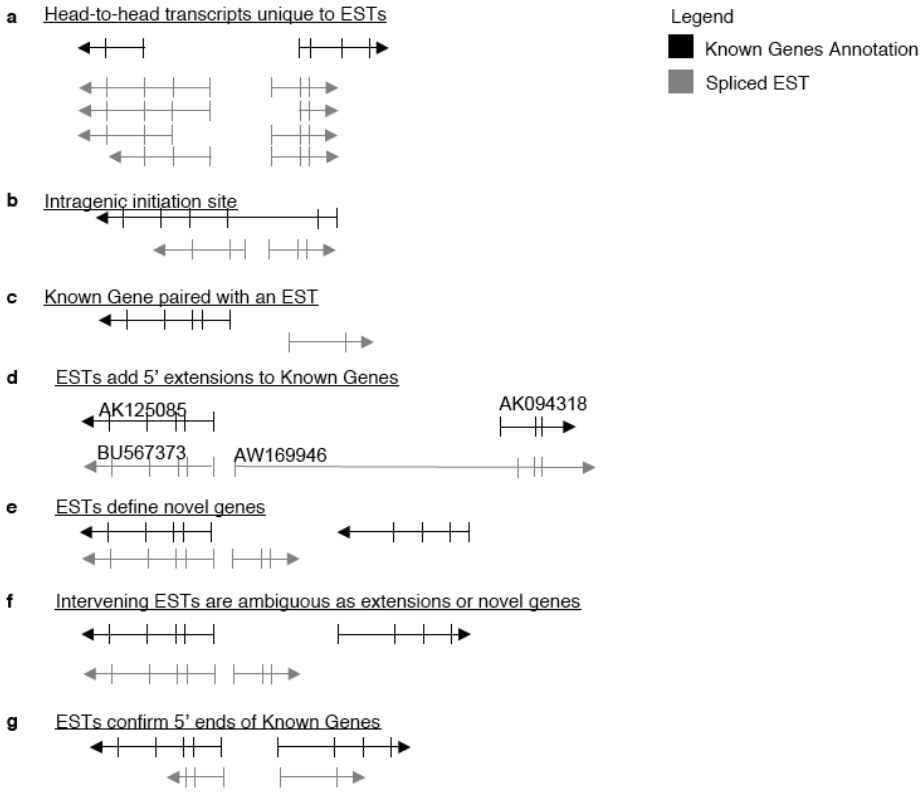

Fig. 4. Possible configurations of Known Genes and spliced ESTs.

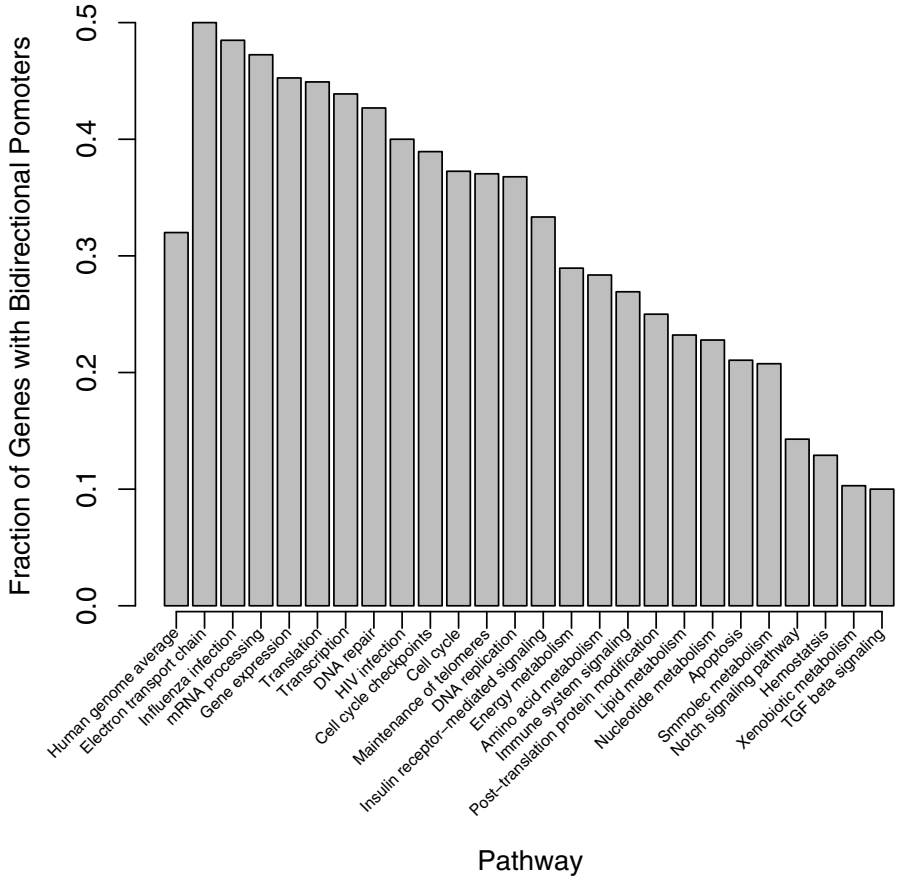

Fig. 5. Fraction of genes regulated by bidirectional promoters for different pathways. The first bar gives the average for the human genome, which is approximately .31.