Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W181–W186***
DOI: 10.1093/nar/gkh404

ASmodeler: gene modeling of alternative splicing from genomic alignment of mRNA, EST and protein sequences

Namshin Kim[1,2], Seokmin Shin[2] and Sanghyuk Lee[1,*]

[1]Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea and [2]School of Chemistry, Seoul National University, Seoul 151-747, Korea

Received February 14, 2004; Revised March 9, 2004; Accepted March 29, 2004

ABSTRACT

Alternative splicing is in important mechanism of modulating gene function and expression which greatly expands transcriptome diversity. ASmodeler is a novel web-based utility that finds gene models including alternative splicing events from genomic alignment of mRNA, EST and protein sequences. User-supplied sequences are aligned against the genome map using the BLAT and SIM4 programs. Resulting exon connectivity is analyzed by applying graph-theoretic methods to build all possible gene models including splice variants. The algorithm essentially combines the genome-based sequence clustering and transcript assembly procedures in a coherent fashion. In addition to the user-supplied sequences, UniGene clusters and many well-known gene predictions such as Genscan, Ensembl and Acembly may be included in gene modeling. The current implementation supports human, mouse and rat genomes. ASmodeler is available at http://genome.ewha.ac.kr/ECgene/ASmodeler/.

INTRODUCTION

Alternative splicing (AS) is an important mechanism in regulating gene expression and function in eukaryotes. Many splice variants with missing motifs or domains are directly related to disease and are therefore major targets for therapeutic drug development (1,2). The complete sequencing of the human genome greatly accelerated discovery of gene variants due to AS events. The last four years have seen numerous databases and methods for identifying alternative splicing (3–8). It is estimated that 40–60% of human genes show evidence of encoding transcripts that are alternatively spliced, and that alternative splicing became a major mechanism in expanding proteome diversity (9–11).

Identification of alternative splicing events is pursued in two directions. Early methods aligned Expressed Sequence Tag (EST) sequences against mRNA sequences to look for any insertion or deletion. Since the human genome map become available, genome-based methods seem to have been dominant since they give more extensive and reliable results. Genome-based methods align mRNA and EST sequences against the genome and examine exon–intron boundaries to find splice variants. Recently, we developed a novel algorithm, ECgene, which combines genome-based EST clustering and transcript assembly procedures in a coherent and consistent fashion (N. Kim, S. Shin and S. Lee, manuscript in preparation; http://genome.ewha.ac.kr/ECgene/). Genomically aligned mRNA and EST sequences are clustered according to shared exon–intron boundaries. The resultant exon connectivity is analyzed by applying graph-theoretic methods to build all possible gene models including splice variants.

ASmodeler builds gene models inferred from user-supplied sequences using the ECgene algorithm. Sequences compatible with each gene structure are grouped together as supporting evidence for the gene model. Gene structure and its supporting evidence are visualized using the University of California Santa Cruz (UCSC) Genome Browser (12) in order to take advantage of its ample annotations and special features.

Instead of being a simple web implementation of ECgene, ASmodeler has many additional features.

(i) Users may input protein sequences. This option can be utilized as a means for comparative gene modeling. For example, protein homologs from mouse may be supplied to the human genome to see if the same AS events are conserved between two species. This is particularly important since the resulting splice variations occur inside the CDS region.
(ii) We support analysis of NCBI's UniGene clusters. UniGene recently switched to the genome-based method (http://www.ncbi.nlm.nih.gov/UniGene/build2.html) for human, mouse and rat, which is quite similar to the

*To whom correspondence should be addressed. Tel: +82 232772888; Fax: +82 232772384; Email: sanghyuk@ewha.ac.kr

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

ECgene algorithm. However, it still does not provide the gene models for each cluster. ASmodeler may be used as a transcript assembler for the UniGene clusters.

(iii) Gene predictions in the UCSC genome database can be included in the transcript building procedure. When input transcripts cover only partial genes, *ab initio* gene-prediction programs are helpful in recovering the full length.

(iv) Public sequences in the GenBank database are pre-aligned against the genome and can be included in the clustering procedure using simple check-box options.

(v) Many parameters for alignment are under user control for more flexibility.

MATERIALS AND METHODS

Datasets

Genome-based sequence clustering requires the genome map and transcript sequences. We used the July 2003 human reference sequence (UCSC version hg16), which is based on NCBI Build 34. The genome sequence was downloaded from the UCSC Genome Site. Genome maps for mouse and rat are still at the draft stage. We used the most recent releases, which are the October 2003 assembly (UCSC version mm4) for mouse and the June 2003 assembly (UCSC version rn3) for rat.

To support UniGene querying, UniGene databases were obtained from NCBI UniGene No. 163 for human (September 22, 2003), No. 129 for mouse (October 3, 2003) and No. 124 for rat (October 18, 2003). ESTs and mRNA sequences were obtained from NCBI GenBank release 138 (October 15, 2003). GenBank flat files were parsed to obtain organism-specific sequences. ESTs in gbestNN.seq and mRNAs in gbhtcNN.seq and gbpriNN.seq were extracted. Sequences from models in UniGene were obtained from NCBI RefSeq release 2 (November 4, 2003, vertebrate mammalian).

Gene predictions and models supported in ASmodeler were obtained from the UCSC download site. UCSC known genes (based on SWISS-PROT, TrEMBL, mRNA and RefSeq), RefSeq genes (13), Vega genes (Vertebrate Genome Annotation), Ensembl genes (14), Fgenesh++ genes (15) and Genscan genes (16) are available at the time of writing. Acembly genes (D. Therry-Mieg, J. Thierry-Mieg, M. Potdevin and M. Sienkiewicz, unpublished; http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly) will be added when they are available from the download site.

Overview of the algorithm

We present a brief summary of ASmodeler algorithm, which is almost identical to the ECgene algorithm. Details of the algorithm will be described elsewhere.

(i) *Genomic alignment of mRNA, EST and protein sequences*. The goal of this first step is to align input sequences against the genome. We use the BLAT (17) C/S version for this purpose. Erroneous BLAT alignments—small gaps, suspicious terminal exons are corrected for valid splice sites. Suspicious alignments are corrected using the SIM4 (18) program, which is based on a dynamic programming algorithm for more sensitive alignment. The combination of BLAT and SIM4 makes the genomic alignment process reliable and superfast, so that PC-based Linux clusters can handle web-based requests. We keep only the best hits that satisfy the minimum percentage identity and alignment coverage criteria unless the user selects the option allowing multiple genomic alignments. Genomic alignments of transcripts from selected gene predictions were downloaded from the UCSC genome site and used without further modification.

(ii) *Primary EST clustering*. In an effort to resolve the problem of contamination by genomic DNA sequences, only the sequences with introns (i.e. spliced sequences) are used in this step since most contaminated ESTs are expected to be unspliced (19). Sequences that share a splice site are grouped together to produce what we call the 'primary' clusters. Considering the low sequence quality of ESTs, variations within ±6 bp are allowed. Primary clusters are equivalent to the version of UniGene built using a genome-based algorithm, except that unspliced sequences are not included.

(iii) *Transcript assembly procedure*. The connectivity of exons in each primary cluster is represented as a directed acyclic graph (DAG). All possible paths along exons are found using the depth-first search method. Each path represents a putative gene model, i.e. the splice variants. In an effort to filter out gene models without sufficient evidence, sequences in the primary cluster are mapped onto the putative gene models, which serve as clone evidence of the model. Exons without clone coverage are trimmed away from the gene model, and the result is compared with other gene models as a redundancy check. The presence of polyA tails, detected from careful analysis of genomic alignment of mRNA and EST sequences, is specifically used to determine the gene boundary and direction. The direction of each gene model is determined by examining the intron sequences of GT→AG consensus.

(iv) *Addition of unspliced sequences*. Unspliced sequences with correct orientation are added so as not to change the exon–intron boundaries of existing gene models. Extension of terminal exons is allowed in this step. The primary cluster obtained so far corresponds to multi-exon genes, whose subclusters represent splice variants.

(v) *Clustering for unspliced genes*. Remaining unspliced sequences are further clustered according to overlap in the genomic loci. The resulting clusters, representing single-exon genes, are added to the list of primary clusters.

It is important to note that some splice forms may not exist in nature. Graph-theoretic analysis usually overpredicts splice variants since it combines all alternative splicing types occurring in different exon regions. It is easy to obtain several hundreds or even thousands of variants for genes studied intensively. The only direct evidence for the validity of a gene model is to verify the full-length clone experimentally. As an aid to judging the reliability of a gene model, we provide the minimal set of clones ('Min. Clones') to cover all exons in each gene model. Having fewer clones in the minimal set implies a higher chance of being a real transcript in a cell. Recently, exon junction microarrays have been used to monitor the splicing pattern at every exon–exon junction (20).

Another source of wrong predictions is misaligned sequences. Even though alignments from the BLAT and SIM4 programs go through extensive correction and filtering steps (details will be described elsewhere), it is not an easy task

to align small exons accurately, especially if the introns are non-canonical. Alignments with small terminal exons or non-canonical introns should be examined critically.

RESULTS

Input options

Figure 1 shows the GUI for the ASmodeler input page. Most items are self-explanatory, but a user guide is available on the website for more details. Input sequences can be any combination of the UniGene cluster ID, mRNA, EST and protein sequences. Minimum percentage identity is under user control, and the defaults values are 96, 93 and 70% for mRNA, EST and protein sequences, respectively. The user can adjust the minimum alignment coverage, too. The default value of 50% means that the aligned part should be over half of the sequence length.

Genomic alignment is the most time-consuming part of the algorithm. In an effort to speed up the computation, we use the pre-calculated genomic alignments for sequences in the UniGene cluster ID (input type A), making querying by UniGene ID a superfast process suitable for a web application. Other input-type sequences (mRNA, EST, protein in box B, C, D) are aligned against the genome using BLAT. Therefore, providing input sequences will take extra execution time, but with more flexibility in the alignment process in terms of the percentage identity cutoff and alignment coverage. It takes about one minute to align and cluster 500 sequences on our Pentium4 2.8 GHz computer.

If the genomic region is specified, ASmodeler chooses the best hit within the specified genomic region even if it finds better hits outside the region. Therefore, specification does not speed up the calculation.

mRNA and EST sequences in GenBank whose pre-calculated genomic alignments overlap with the user-supplied sequences may be included in clustering procedure by selecting a check-box option. This is useful for researchers who are trying to build gene models by combining a small number of their own sequences with public resources in the GenBank. We download the GenBank sequences periodically and update the alignment for this purpose. By default, only the best alignment within the user-specified genomic region is kept. However, multiple hits above the minimum percentage

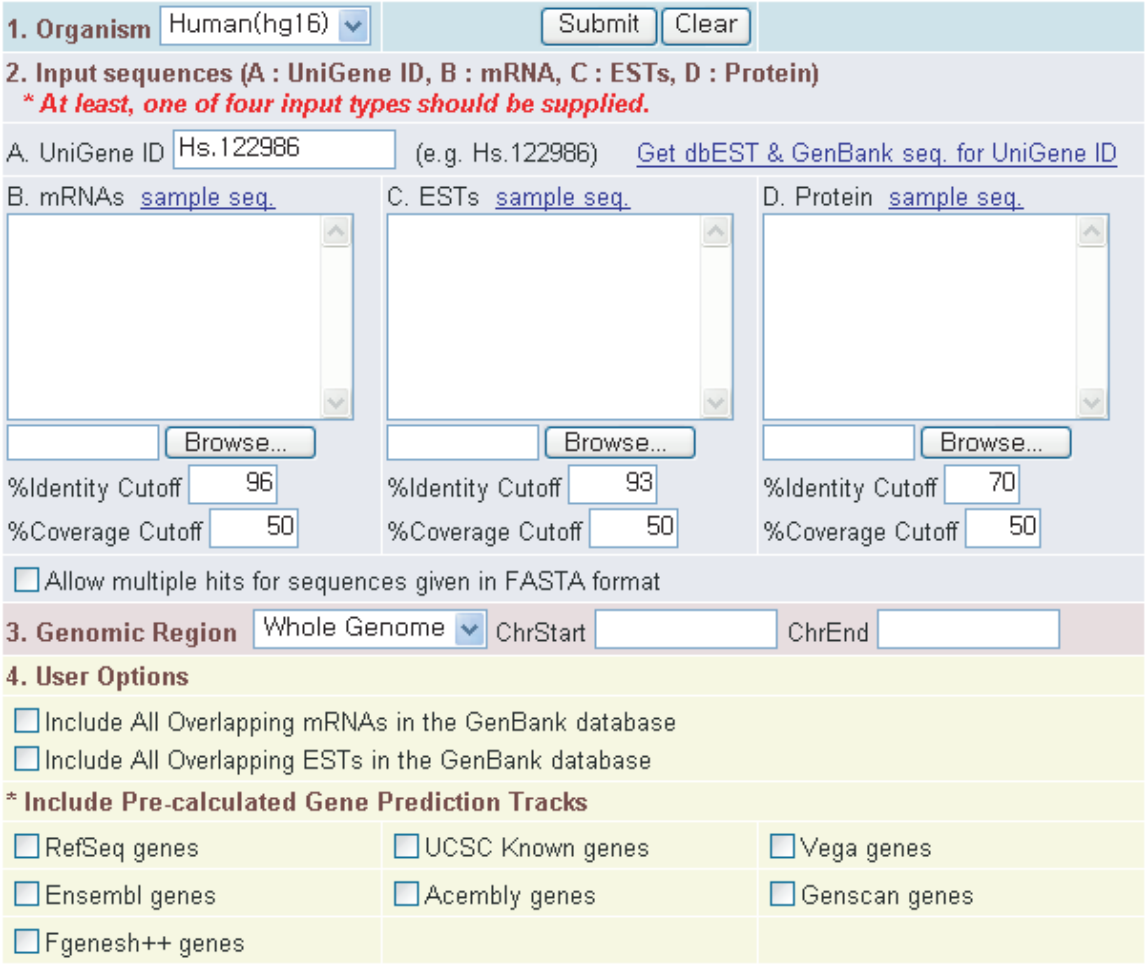

Figure 1. Input GUI design for ASmodeler. Any combination of input types is allowed. (**A**) UniGene cluster ID, (**B**) mRNA, (**C**) EST, (**D**) protein sequences in FASTA format. See text for more details on the check-box options. We provide a script to extract sequences from dbEST and GenBank for a given UniGene ID. Using downloaded sequences from the UniGene website is not recommended since they are the raw sequences that contain contaminants, vector sequences and low-quality parts. A user guide is available on the ASmodeler website.

identity may optionally be allowed. Finally, mRNAs from pre-calculated gene predictions can be added using simple check-box options.

Output features

We will describe the output of ASmodeler by taking the transcript assembly for a UniGene cluster as an example. The main output is the sequence clusters and gene models for each cluster. Figure 2 is part of the output from ASmodeler for the UniGene cluster Hs.122986, corresponding to the TPTE (transmembrane phosphatase with tensin homology) gene. It consists of 5 mRNA and 52 EST sequences. The upper table is the list of inferred clusters with brief summary information that includes the number of splice variants in each cluster and the number of mRNA, EST and spliced EST sequences. Clicking on the cluster ID shows the clustering result—i.e. sequence members for each cluster and splice variants. The lower table is the summary of transcript models in each cluster. In addition to the basic information (the number of mRNA, EST and spliced EST sequences), it shows the size of the resultant mRNA and protein sequences. Information on the coding and untranslated regions is also shown. Links to mRNA and protein sequences are provided for each splice model. The column 'Min. Clones' is the minimum number of clones necessary to reconstruct the gene structure, and (e) means that all representative clones are EST sequences, not mRNA. Having fewer clones in the minimal set implies a higher chance of being a real transcript in a cell.

Genomic positions in Figure 2 are linked to the UCSC Genome Browser to display the gene structure and clustering results graphically. ASmodeler's result is added to the UCSC Genome Browser as custom tracks, as shown in Figure 3. The initial layout shows all clusters within the genomic region in the dense-view mode. Cluster size and the presence of a polyA tail are indicated in the title line. The presence of a polyA tail strongly implies the end of transcription. Clicking on each gene model displays the sequence alignment of cluster members in the pack-view mode, as can be seen for the AS21C1.1 cluster in Figure 3. Sequences with polyA tail and mRNA sequences are clearly marked to aid interpretation.

According to our analysis, Hs.122986 is further classified into four clusters. Most members belong to the AS21C1 cluster, which has six variant forms. All sequences in other clusters are unspliced, which reduces the reliability of genomic alignment. Even though they have partial overlap with sequences in the AS21C1 cluster, AS21C3 and AS21C4 mostly align at the intronic region of the AS21C1 cluster. Two ESTs in the second cluster (AS21C2) have overlapping alignment of 2 bp within the intron region of the AS21C1 cluster. Since exon–intron mismatch is strictly forbidden in adding unspliced sequences to existing clusters, they are clustered as a separate gene even though it is quite possible to allow small mismatches. No sequence in these clusters shares any exon–intron boundary with the AS21C1 cluster. Therefore, we group these sequences into separate clusters.

Cluster AS21C1 consists of 4 mRNA and 43 EST sequences, which imply six variants. The first variant, AS21C1.1, has 37 sequence members, 34 of these being spliced—i.e. containing introns. Since RefSeq NM_013315 is a member of this variant (indicated as mRNA with polyA), this may be the representative gene model. Four exons show alternative exon-skipping events. Variant 5 (AS21C1.5) has identical gene structure to variant 4, except for the terminal exon (cf. gene direction on the negative

Gene Models

Cluster ID	#Var	#mRNA	#EST	#Spliced	genomic position	strand	polyA
AS21C1	6	4	43	43	chr21:9928622-10013226 (UCSC)	-	O
AS21C2	1	0	2	0	chr21:9929047-9929348 (UCSC)	+	X
AS21C3	1	1	6	0	chr21:9	-	X
AS21C4	1	0	1	0	chr21:9	-	X

Genome Browser

Transcript ID	mRNAs #	ESTs #	Spliced #	Exons #	polyA #	Min. Clones	mRNA (bp)	CDS (bp(aa))	5'UTR (bp)	3'UTR (bp)
AS21C1.1	2	35	34	24	2	1(m)	2163	1653(551)	344	165
AS21C1.2	0	12	9	4	0	1(e)	557	252(84)	142	162
AS21C1.3	0	1	1	8	0	1(e)	581	279(93)	297	0
AS21C1.4	1	36	33	23	0	1(m)	2558	1599(533)	368	590
AS21C1.5	1	21	21	23	1	1(m)	2128	1599(533)	368	160
AS21C1.6	0	2	2	9	0	1(e)	694	357(119)	333	0
AS21C2.1	0	2	0	1	0	1(e)			99	0
			0	1	0	1(m)			712	392
			0	1	0	1(e)	767	0(0)	0	766

Sequences

Cluster Members

Figure 2. Summary output table for gene models. The cluster's ID, AS21C2, implies that it is located on chromosome 21. A variant number is added as an extension to the cluster ID. Links in the table provide relevant information, as indicated in the box.

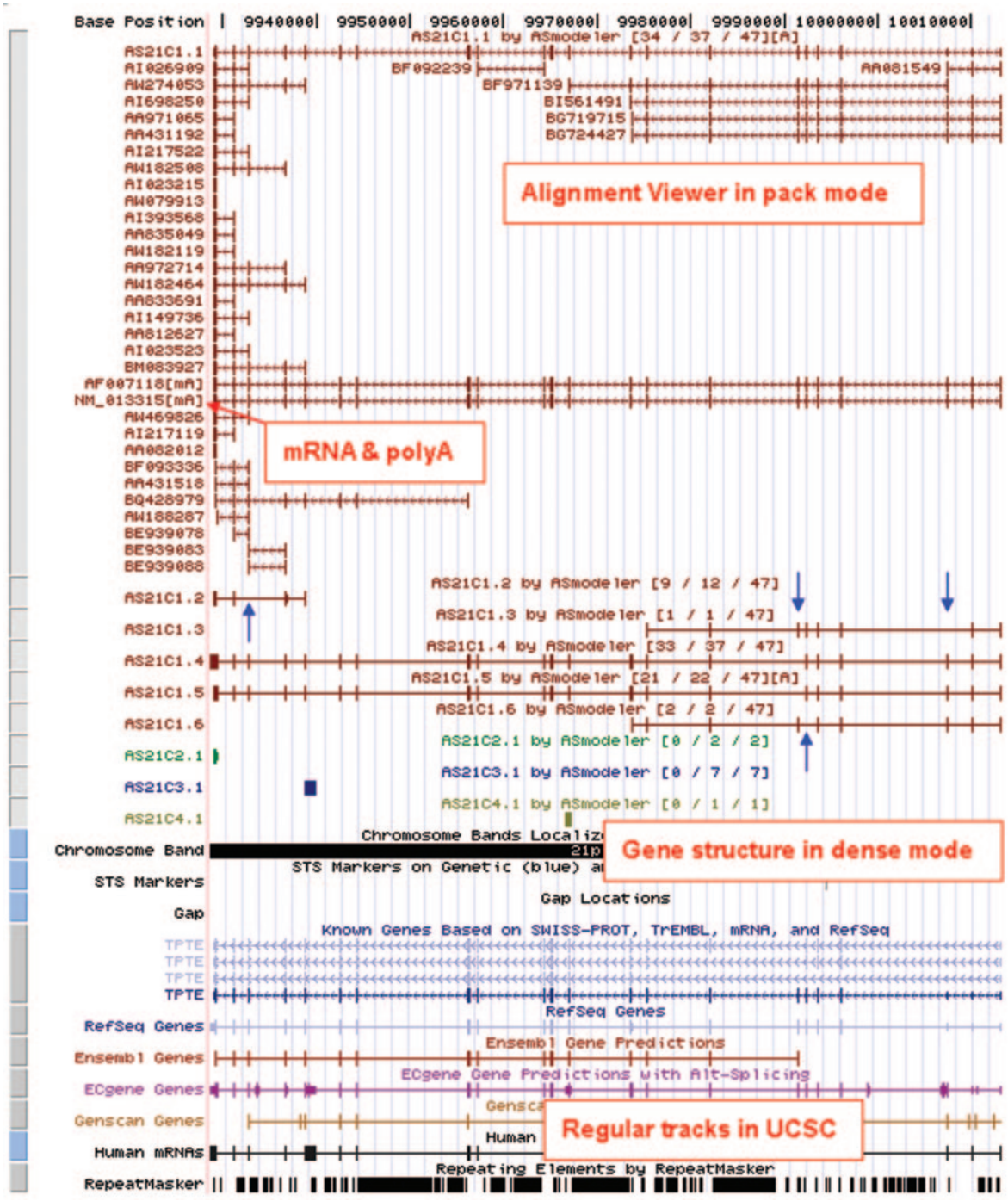

Figure 3. The UCSC genome browser for viewing gene structure and clustering. Each cluster is shown in a different color. Variant no. 1 (AS21C1.1) is expanded in the pack-view mode to show the sequence alignment. Other variants are in the dense-view mode. The title line 'AS21C1.5 by ASmodeler [21 / 22 / 47][A]' means that cluster AS21C1 has 47 sequences, 22 of those being consistent with the gene structure of the fifth variant. It includes 21 spliced sequences and 1 unspliced sequence. This transcript shows evidence of a polyA tail. Four exons involved in alternative exon-skipping events are marked with arrows.

strand). The expanded view shows that variant 5 has an mRNA clone (BC028719) with a valid polyA tail that indicates the end of transcription. On the other hand, variant 4 has another mRNA clone (AL833345) whose terminal exon is longer than BC028719. Therefore, it seems that all six variants show some sort of evidence of mRNA and/or EST. More recent data show that there exist additional isoforms supported by new mRNA clones (AY219887, AY219888).

ASmodeler can be applied for many purposes. The user guide on the website illustrates three typical applications: (i) genome-based sequence clustering and gene modeling for alternative splicing, (ii) transcript assembly for UniGene clusters as described in this paper and (iii) comparative gene modeling of alternative splicing using homologous protein sequences. Each example includes a detailed description and discussion.

DISCUSSION

ASmodeler is a versatile web application for general gene modeling. The fact that transcripts are assembled from user-supplied sequences gives tremendous flexibility. For example, one can remove sequences with suspicious alignment. The UCSC Genome Browser does not show arrows for non-canonical introns that do not satisfy the GT→AG consensus. One may remove those parts of the alignment from the input sequence. On the other hand, gene predictions or even suboptimal exons can be added to find any additional variants or to extend the existing models.

Even though NCBI's UniGene switched to the genome-based method for human, mouse and rat genomes, UniGene itself does not provide any consensus sequences or splice variant analysis. As illustrated in the examples in Figures 2 and 3, ASmodeler can be utilized as a gene-modeling program for the UniGene clusters that includes alternative splicing events. Furthermore, web implementation enables the user to add private sequences to find any additional variants. Sequence quality in the resulting gene models is high since exon sequences are extracted from the genome sequence, not from the overlapping EST sequences.

It should be noted that ASmodeler's role as a transcript assembler for UniGene clusters is quite similar to the SpliceNest program (21). However, it generates consensus sequences by analyzing sequence alignments between mRNA and EST sequences in a UniGene cluster without using the genome map. Resultant transcript-specific contigs are mapped onto the genome. Furthermore, real-time calculation in ASmodeler can accommodate user-supplied sequences with many other optional features, whereas SpliceNest shows pre-calculated results.

Presenting the clustering result using the UCSC Genome Browser allows users to take advantage of the tremendous resources at the UCSC Genome Center (12). Ample annotations, such as putative transcript start sites, transcription factor binding sites, neighboring genes, mRNA sequences from other organisms and conserved regions between species, will not only be of great help in testing the validity of the model, but also provide better insight into the biological function of the gene. Furthermore, microarray and SAGE tracks are available to explore the gene expression *in silico*. Utility functions at the Genome Center are also useful. For example, one can extract the DNA sequences with exons from each gene model colored differently by using the DNA link of the genome browser. The table browser link can be used to combine other annotation tracks with the ASmodeler result for additional information.

ACKNOWLEDGEMENTS

We are grateful to the UCSC Genome Center for making such wonderful resources available to the public. This work was supported by the Ministry of Science and Technology of Korea through the bioinformatics research program of MOST NRDP (M1-0217-00-0027) and the Korea Science and Engineering Foundation through the Center for Cell Signaling Research at Ewha Womans University.

REFERENCES

1. Caceres,J.F. and Kornblihtt,A.R. (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet*., **18**, 186–193.
2. Levanon,E.Y. and Sorek,R. (2003) The importance of alternative splicing in the drug discovery process. *Targets*, **2**, 109–114.
3. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res*., **32**, D64–D69.
4. Pospisil,H., Herrmann,A., Bortfeldt,R.H. and Reich,J.G. (2004) EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Res*., **32**, D70–D74.
5. Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res*., **12**, 1837–1845.
6. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res*., **31**, 101–105.
7. Zavolan,M., Kondo,S., Schonbach,C., Adachi,J., Hume,D.A., Hayashizaki,Y. and Gaasterland,T. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res*., **13**, 1290–1300.
8. Lee,B.T., Tan,T.W. and Ranganathan,S. (2003) MGAlignIt: a web service for the alignment of mRNA/EST and genomic sequences. *Nucleic Acids Res*., **31**, 3533–3536.
9. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Ann. Rev. Biochem*., **72**, 291–336.
10. Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
11. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*., **17**, 100–107.
12. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
13. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res*., **31**, 34–37.
14. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res*., **30**, 38–41.
15. Salamov,A.A. and Solovyev,V.V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res*., **10**, 516–522.
16. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78–94.
17. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
18. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*., **8**, 967–974.
19. Sorek,R. and Safer,H.M. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res*., **31**, 1067–1074.
20. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
21. Krause,A., Haas,S.A., Coward,E. and Vingron,M. (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res*., **30**, 299–300.