A Database Designed to Computationally Aid an Experimental Approach to Alternative Splicing

C.L. Zheng, T.M. Nair, M. Gribskov, Y.S. Kwon, H.R. Li, and X.-D. Fu

Pacific Symposium on Biocomputing 9:78-88(2004)

A DATABASE DESIGNED TO COMPUTATIONALLY AID AN EXPERIMENTAL APPROACH TO ALTERNATIVE SPLICING

C.L. ZHENG[1], T.M. NAIR[1], M. GRIBSKOV[1,2]
University of California, San Diego
[1]San Diego Supercomputer Center
[2]Department of Biology
9500 Gilman Dr.
La Jolla, CA 92093, USA
{czheng, nair, gribskov} @sdsc.edu

Y.S. KWON, H.R. LI, X.-D. FU
University of California, San Diego
Department of Cellular and Molecular Medicine
9500 Gilman Dr.
La Jolla, CA 92093, USA
{ykwon, hairili, xdfu} @ucsd.edu

A unique microarray approach has been developed to profile alternative splicing in the cell. To support the development of this approach, we have developed the Manually Annotated Alternatively Spliced Events (MAASE) database system, which is a unique alternative splicing information resource designed specifically with experimentalists in mind. MAASE is an online resource for the convenient access, identification, and annotation of alternative splicing events (ASEs). MAASE consists of two components: an annotation system and a curated database. The annotation system is a web-based workspace that combines manual and computational approaches to identifying and annotating ASEs, a combination that is vital if a comprehensive collection is to be obtained. The annotation system is publicly available and provides a scalable solution to acquiring as well as contributing to annotated ASEs. MAASE annotated ASEs are deposited into the database component, which can either be queried one entry at a time or multiple entries at a time with convenient access to alternatively spliced junctional and surrounding sequences to facilitate the design of microarray experiments.

1 Introduction

The frequency and importance of alternative splicing (AS) is evidenced by studies, indicating that up to 60% (1-4) of all human genes are alternatively spliced and that it may be one of the major mechanisms in expanding and regulating the composition of the proteome (5). With this in mind, substantial effort has been devoted to the identification, annotation and prediction of alternatively spliced genes and their alternatively spliced isoforms. A brief list of some database efforts focused on AS includes the Alternative Splicing Database (ASDB) (6); Alternative Splicing Database of Mammals (AsMamDB) (7); SpliceDB (8); Putative Alternative Splicing Database (PALS db) (9); Intron Information System (ISIS) (10); and Alternative

Splicing Annotation Project (ASAP) (11). Each of these database efforts has contributed to the further understanding of the field. For example ASDB has clustered and identified alternatively spliced variants based on analysis of Swiss-Prot and GenBank while AsMamDB contains information on alternatively spliced genes of human, mouse, and rat species. SpliceDB is a database of canonical (GT-AG) and non-canonical (GC-AG; AT-AC) splice sites inferred from EST sequences. ISIS is a database of intron sequences, extracted from GenBank, that are involved in AS. PALS db reveals putative alternative splicing events (ASEs) by visually aligning UniGene clusters to the longest cDNA sequence. The quality of each aligned portion is displayed to allow users to judge the veracity of putative alternatively spliced junctions. ASAP is currently the largest database of alternatively spliced genes in humans and provides information on gene structure, AS, tissue specificity, and protein isoforms.

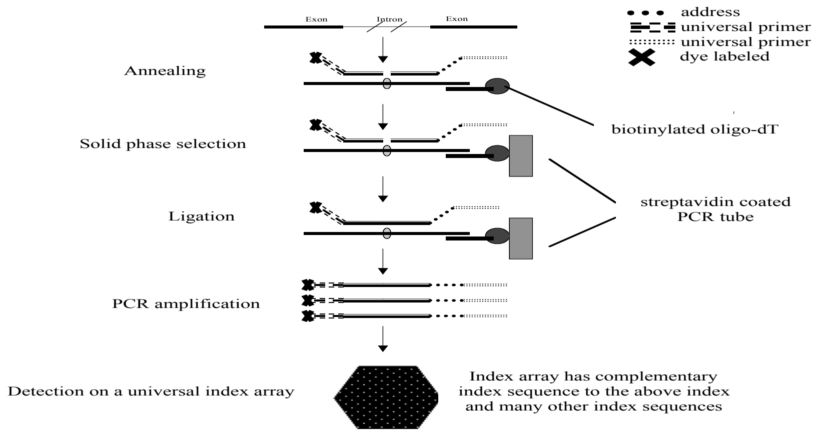
Informative as each AS database effort is, experimental labs have not been able to fully utilize them for several reasons. Firstly, strict heuristics, which are applied in selecting entries in many of the computationally derived databases to ensure a high degree of accuracy, have rendered them incomplete. Many are also missing information in which experimentalists are truly interested, such as the complex AS modes (e.g. mutually exclusive exons and even more complex modes such as those found in CD44 [reviewed (12, 13)]) and information that can be found only in literature. Finally, interest in looking at the global effects on/of AS have prompted some initial attempts to address alternative splicing using microarray platforms (14-17); however current AS databases are not structured for convenient access to the information needed by experimentalists to design and perform these experiments.

To meet the needs of experimentalists, the Manually Annotated Alternatively Spliced Events (MAASE) database system is designed to accurately annotate ASEs by a combination of manual and computational efforts and to allow for convenient access to its content. The inspiration for MAASE grew out of the parallel development of the RASL splicing array platform (17). This collaborative effort has resulted in a successful bridge linking AS databases and the needs of those who use them.

2 Results/Discussion

2.1 The RASL Approach to Profiling Alternative Splicing

In order to profile AS on a large scale by microarray approaches, an oligonucleotide ligation-dependent hybridization approach (Fig 1), RASL (RNA Annealing Selection and Ligation), has been developed (17). The first step in RASL is to synthesize oligonucleotides complementary to specific splice junction donor and acceptor sequences (the target oligonucleotides). To distinguish between different ASEs, oligonucleotides complementary to specific exonic splice junctions are linked to individual index (or address) sequences, a collection of computer-generated and experimentally verified sequences which are not found in the genomic sequence. Furthermore, each oligonucleotide is also linked to a universal primer landing site for PCR amplification.



address
universal primer
universal primer
dye labeled
Exon
Intron
Exon
Annealing
biotinylated oligo-dT
Solid phase selection
Ligation
streptavidin coated PCR tube
PCR amplification
Detection on a universal index array
Index array has complementary index sequence to the above index and many other index sequences

Fig 1. The RNA Annealing Selection and Ligation (RASL) Strategy

The RASL assay consists of five steps: (1) Annealing: Pooled oligonucleotides are mixed with isolated total cellular RNA along with biotinylated oligo-dT. (2) Solid phase selection: The mix is then transferred to a streptavidin coated PCR tube. Biotinylated oligo-dT is thus immobilized on the surface; mRNA is annealed to the oligo-dT; and the target oligonucleotides are annealed to specific splice junctions in the mRNA. After the selection, unhybridized oligonucleotides are washed away. (3) RNA-mediated oligonucleotide ligation: Target oligonucleotides corresponding to a particular splice junction are juxtaposed. The aligned oligonucleotides are then ligated by T4 ligase. This step is key for the specificity of the assay, as only oligonucleotides, which are annealed next to each other on a targeted RNA, will be ligated. Furthermore, because each target oligonucleotide carries only one primer site, only ligated oligonucleotides have primer sites on both ends, and thus can serve as a template for PCR amplification. (4) PCR amplification: The pair of universal primers, one of which is dye-labeled, is used to amplify the ligated products. This step is the basis for the high sensitivity of the assay. (5) Detection on a universal index array: The dye-labeled PCR products are hybridized to an array of index sequences to allow quantification of specific ASEs.

The RASL approach combines high specificity and sensitivity in profiling AS in the cell. The approach, however, requires prior knowledge of AS, and thus a high quality AS database is essential. We therefore decided to build such a database in order to efficiently facilitate this experimental approach.

2.2 MAASE Database System Overview

To facilitate the RASL approach, we set out to construct a comprehensive and user-friendly AS resource — the MAASE database system. The system comprises two components: an annotation component and a database component. The annotation component is an environment for manual annotation (with computational support) of ASEs. The combination of manual and computational annotation of ASEs addresses many of the shortcomings of purely computationally derived databases. The MAASE annotation system is publicly available and is intended to be a community-based effort. The high level of effort required for manual annotation is one of the driving forces for this project. A community-based effort greatly enhances the scalability of this data resource. With this database system, the AS community can contribute as well as obtain information.

2.3 Annotation of ASEs Using MAASE

Current AS databases are not comprehensive due to their lack of coverage of more complex AS splicing modes and the exclusion of AS information found only in the literature. In addition, those databases do not provide the information or the query capability needed for designing microarray experiments. Because of these limitations, we began to manually annotate ASEs for the RASL platform. Inspired by such manual annotation efforts, the MAASE annotation pipeline has been developed to incorporate computational efforts that enhance the speed and accuracy of manual annotation.

The manual annotation pipeline has two entry points: the Swiss-Prot database and NCBI's GenBank database. Swiss-Prot is the preferred point of entry due to its well-curated gene entries however not all gene loci have a corresponding Swiss-Prot entry at which time GenBank can be used. Next, related cDNA and EST sequences from a variety of databases are collected along with AS information from published literature. Each of the sequences is aligned to the genomic sequence to visualize the ASEs. The mode of splicing of each ASE is then determined. The data obtained from such a tedious annotation task proved to be well worth the time and effort, as it leads to a complete and detailed AS annotation. The MAASE annotation pipeline is a web-based tool that automates many of the manual annotation steps described above.

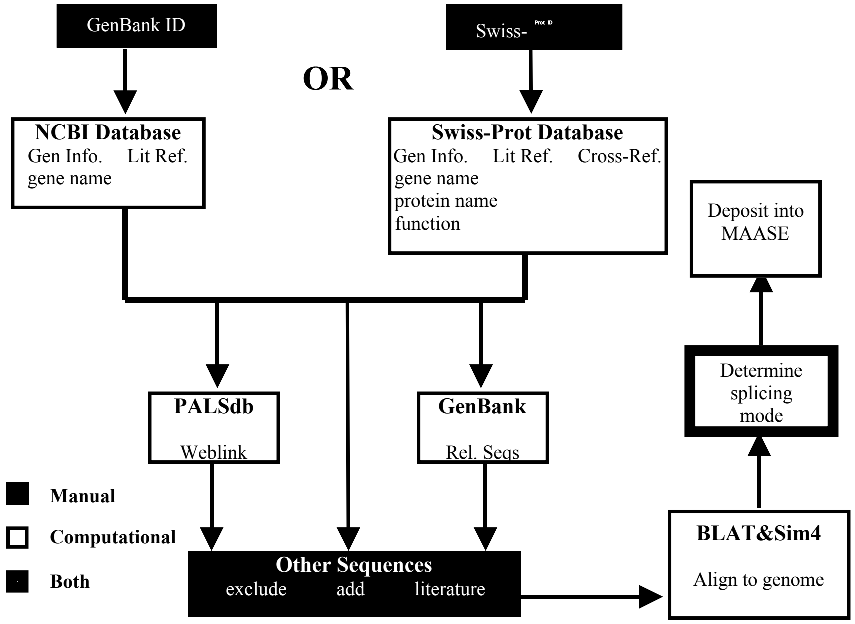


Fig 2. Flowchart of MAASE Annotation Tool

Beginning with either a Swiss-Prot or GenBank entry, MAASE automatically retrieves information such as literature references and GenBank cross-references; obtains a weblink to PALSdb; obtains related GenBank entries; allows the user to exlude/include sequences from other databases or published literature; aligns sequences to the genome to obtain a gene model; determines splicing mode; and deposits into MAASE.

The details of how the manual and computational efforts work in synergy are what make the MAASE annotation unique (Fig 2). The first step is for the user to enter the desired Swiss-Prot ID or GenBank ID. From this entry point, the MAASE annotation system automatically obtains useful information such as gene name, protein name, functional informational and a list of GenBank cross-references. MAASE then queries GenBank for other related cDNA sequences based on the cross-referenced sequences. A web-link to the PALS db is also retrieved for the specific Swiss-Prot entry to allow the user to judge and include EST sequences as they see fit. All of this information is then graphically presented to the user on the web. At this point, the user can include/exclude sequence based on existing evidence. Any AS information found in published literature can also be added at this point. Once all of this information is entered, MAASE takes over and aligns each of the sequences to the genomic sequence using BLAT (18) and Sim4 (19). BLAT is used to pinpoint the genomic location of the entered sequences; Sim4 is used to align each of the sequences to the genomic sequence. After alignment, MAASE indicates all internal ASEs. MAASE does this by first constructing a master sequence of all non-redundant sequence segments. These sequence segments consists of whole exon segments as well as subsequences of exons showing a splicing difference (Fig 3). Each entered sequence is compared to the master sequence in search of missing sequence regions, and then with all other sequences entered to determine the splicing mode. Once again all information is presented to the user. At this point the user does a final check of the sequences and splicing mode. Although MAASE is able to automatically assign the splicing mode, complicated splicing modes may not always be assigned correctly due to the use of certain heuristics, and therefore some manual intervention is needed. Once the user verifies the entry, it is deposited into the database. We believe that such an intricate tag-team system is required to achieve accurate annotation of ASEs.

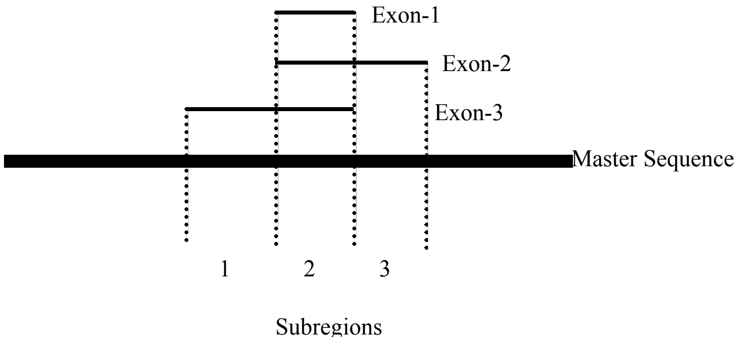


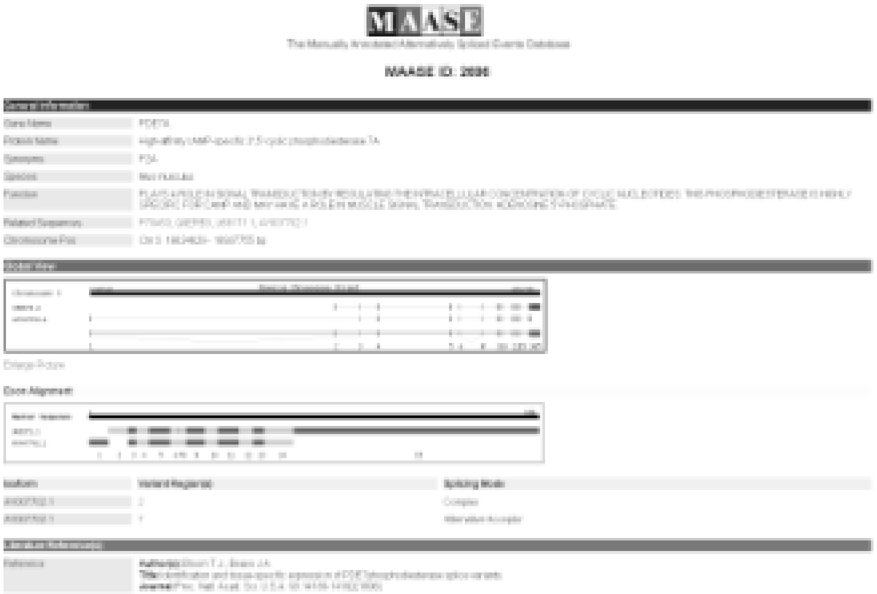
Fig 3. Construction of the Master Sequence

The master sequence is constructed with all exonic information from collected cDNA and ESTs. Here is an example of how overlapping exons would be separated into subregions in the master. Once the master sequence is constructed, each entered sequence is compared to the master sequence in search of missing regions. The splicing modes are assigned by comparing all entered sequences using the corresponding genomic region as reference. The accuracy of the assignment is ensured by manual inspection.

2.4 MAASE Database

The MAASE database is built for easy access to alternatively spliced junction sequences, either individually or collectively. These features are not addressed by other AS databases, but are essential for the design of microarray experiments. The MAASE database can be queried for a list of alternatively spliced junction sequences sorted by keyword or splicing mode. MAASE also contains a built-in program to pair index (or address) sequences to targeting oligonucleotides for the RASL assay. The most well suited pairs are chosen by pairing each oligonucleotide sequence with each possible address sequence and calculating potential RNA secondary structure, using RNAFold (20). In this way, the chosen index-targeting oligonucleotide pair has the least stable structure to minimize internal hybridization to allow for maximal hybridization potential on the index chip. MAASE can be queried individually by keyword, gene name, NCBI accession number, Swiss-Prot ID and splicing mode. Each database entry consists of the following sections (Fig 4):

- **General Information**: Gene Name, Protein Name, Synonyms, Species, Function, Related Sequences, Chromosome Position
- **Global View**: Graphical alignment of each isoform with the genome and with each other, assessment of individual exon and intron sequences, graphical representation of each alternatively spliced event
- **Exon Alignment**: Alignment of each isoform's exons to all other isoforms with alternating colors for visual clarity
- **Splicing Mode Information**: Name of the variant sequence, variant region, splicing mode
- **Literature References**: Standard literature citations



MAASE

The Manually Annotated Alternatively Spliced Events Database

MAASE ID: 2696

General Information

Gene Name	PDE7A
Protein Name	High-affinity cAMP-specific 3',5'-cyclic phosphodiesterase 7A
Synonyms	P3A
Species	Mus musculus
Function	PLAYS A ROLE IN SIGNAL TRANSDUCTION BY REGULATING THE INTRACELLULAR CONCENTRATION OF CYCLIC NUCLEOTIDES. THIS PHOSPHODIESTERASE IS HIGHLY SPECIFIC FOR CAMP AND MAY HAVE A ROLE IN MUSCLE SIGNAL TRANSDUCTION. [illegible]
Related Sequences	[illegible]
Chromosome Pos	Chr 3: [illegible] bp

Global View

Enlarge Picture

Exon Alignment

Isoform	Variant Region(s)	Splicing Mode
[illegible]	2	Complex
[illegible]	1	Alternative Acceptor

Literature Reference(s)

Reference	Author(s): Bloom T.J., Beavo J.A. Title: Identification and tissue-specific expression of PDE7 phosphodiesterase splice variants Journal: Proc. Natl. Acad. Sci. U.S.A. 93: 14188-14192 (1996)

Fig 4. MAASE Database Entry Information Page

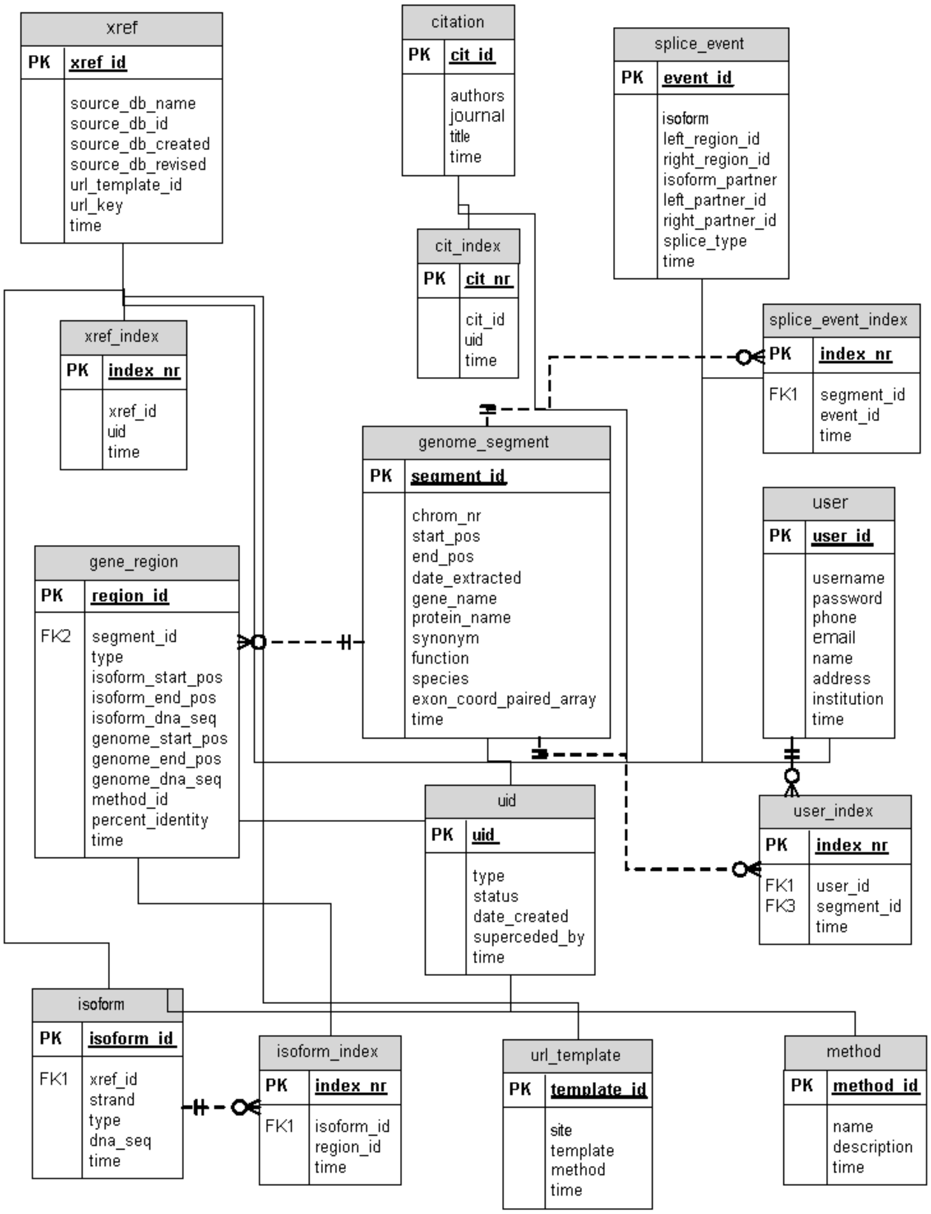


Fig 5. MAASE Database Schema

For simplicity, not all foreign key relationships involving the **UID** table are shown.

The MAASE database schema (Fig 5) was designed for simplicity both of data entry and of retrieval. The design is similar to a star schema in which most of the tables (shown in **bold**) can be thought of as children of the unique identifier (**uid**)

table. The **uid** table assigns unique identifiers (ids) to individual table objects and stores each of the table object ids, types and their current status in the database. The tables fall into several basic groups: core tables essential in a functional genomics database, tables for sequences, tables for annotation, and tables used to link internal and external information together. The core tables provide information on each gene locus (**genome_segment)**, information on related sequences (**xref**), information on database users (**user**) and information used by the database (**method**, **url_template**, and **uid**). The tables providing sequence information are **gene_region**, and **isoform**. Annotation is managed by the **splice_event** table. The tables that link information together (**splice_event_index**, **user_index**, **isoform_index**, **cit_index**, and **xref_index**) allow for quick and convenient updates to the database while relationships between tables allow for the handling of the special needs of an AS database. The **genome_segment** table is the primary table for each database entry with all other (e.g., sequence and annotation) tables relating to it. An example of this relationship can be seen in the construction of an isoform from individual gene region objects. Each gene region, identified in the **gene_region** table, is a defined subsequence of a contiguous genomic DNA segment specified in the **genome_segment** table, as well as a subsequence of an individual isoform sequence. The **isoform_index** table indicates how a specific **isoform** is to be constructed from the **gene_region** objects. In this manner, individual **gene_region** objects can be separate entities as well as part of its **isoform,** allowing for easy access to individual exon/intron sequences as well as whole isoforms.

MAASE is implemented using the MySQL (21) relational database

management system, and the core application programming interface (API) used by MAASE is written by Modulewriter (22), an object-relational mapping (ORM) tool. The MAASE database system can be accessed at http://splice.sdsc.edu.

Acknowledgements

This work was supported by the U.S. National Institutes of Health (CA-88351) and assisted by the facilities of the National Biomedical Computation Resource at SDSC (NIH RR08605).

References

1. I.H.G. Consortium, "Initial Sequencing and Analysis of the Human Genome" Nature 409, 860 (2001)
2. D. Brett, J. Hanke, G. Lehmann, S. Hasse, S. Delbruck, S. Krueger, J. Reich, and P. Bork, "EST Comparison Indicates 38% of Human mRNA Contain Possible Alternative Splice Forms" FEBS Letters 474, 83 (2000)
3. B. Modrek, A. Resch, C. Grasso, and C. Lee, "Genome-wide Detection of Alternative Splicing in Expressed Sequences of Human Genes" Nucleic Acids Research 29, 2850 (2001)
4. A.A. Mironov, J.W. Fickett, and M.S. Gelfand, "Frequent Alternative Splicing of Human Genes" Genome Research 9, 1288 (1999)
5. B.R. Graveley, "Alternative Splicing: Increasing Diversity in the Proteomic World" Trends in Genetics 17, 100 (2001)
6. I. Dralyuk, M. Brudno, M.S. Gelfand, M. Zorn, and I. Dubchak, "ASDB: Database of Alternatively Spliced Genes" Nucleic Acids Research 28, 296 (2000)
7. H. Ji, Q. Zhou, F. Wen, H. Xia, X. Lu, and Y. Li, "AsMamDB: An Alternative Splice Database of Mammals" Nucleic Acids Research 29, 260 (2001)
8. M. Burset, I.A. Seledtsov, and V.V. Solovyev, "SpliceDB: Database of Canonical and Non-canonical Mammalian Splice Sites" Nucleic Acids Research 29, 255 (2001)
9. Y.-H. Huang, J.-J. Chen, S.-T. Yang, and U.-C. Yang, "PALS db: Putative Alternative Splicing Database" Nucleic Acids Research 30, 186 (2002)
10. L. Croft, S. Schandorff, F. Clark, K. Burrage, P. Arctander, and J.S. Mattick, "ISIS, the Intron Information System, Reveals the High Frequency of Alternative Splicing in the Human Genome" Nature Genetics 24, 340 (2000)
11. C. Lee, L. Atanelov, B. Modrek, and Y. Xing, "ASAP: The Alternative Splicing Annotation Project" Nucleic Acids Research 31, 101 (2003)
12. D. Naor, S. Nedvetzki, I. Golan, L. Melnik, and Y. Faitelson, "CD44 in Cancer" Clinical Reviews in Clinical Laboratory Sciences 39, 527 (2002)
13. J. Lesley, and R. Hyman, "CD44: Structure and Function" Frontiers in Bioscience 3, d616 (1998)
14. D.D. Shoemaker, E.E. Schadt, C.D. Armour, Y.D. He, P. Garrett-Engele, P.D. McDonagh, P.M. Loerch, A. Leonardson, P.Y. Lum, G. Cavet, L.F. Wu, S.J. Altschuler, S. Edwards, J. King, J.S. Tsang, G. Schimmack, J.M. Schelter, J. Koch, M. Ziman, M.J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M.R. Meyer, M. Mao, J. Burchard, M.J. Kidd, H. Dai, J.W. Phillips, L.P. S., R. Stoughton, S. Scherer, and M.S. Boguski, "Experimental Annotation of the Human Genome Using Microarray Technology" Nature 409, 922 (2001)
15. H. Wang, E. Hubbell, J.-S. Hu, G. Mei, M. Cline, G. Lu, T. Clark, M.A. Siani-Rose, M. Ares, D.C. Kulp, and D. Haussler, "Gene Structure-Based

Splice Variant Deconvolution Using a Microarray Platform" Bioinformatics 19, 315 (2003)

16. G.K. Hu, S.J. Madore, B. Moldover, T. Jatkoe, D. Balaban, J. Thomas, and Y. Wang, "Predicting Splice Variant from DNA Chip Expression Data" Genome Research 11, 1237 (2001)
17. J.M. Yeakley, J.-B. Fan, D. Doucet, E. Wickham, Z. Ye, M.S. Chee, and X.-D. Fu, "Profiling Alternative Splicing on Fiber-Optic Arrays" Nature Biotechnology 20, 353 (2002)
18. W.J. Kent, "BLAT - The BLAST-Like Alignment Tool" Genome Research 12, 656 (2002)
19. L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller, "A Computer Program for Aligning a cDNA Sequence With a Genomic DNA Sequence" Genome Research 8, 967 (1998)
20. M. Zuker, "Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction" Nucleic Acids Research 31, 3406 (2003)
21. MySQL, Ed. P. Dubois (New Riders Publishing, Indianapolis, 2000).
22. C.L. Zheng, F. Fana, P.V. Udupi, and M. Gribskov, "Modulewriter: A Program For Automatic Generation of Database Interfaces" Computational Biology and Chemistry 27, 135 (2003)