Computational Genomics Project Report
An Analysis of Alternative Splicing Behavior across Seven Organisms

Rose Hoberman

December 5, 2002

1 Motivation

One of the most surprising outcomes of the recent genome sequencing projects has been the discovery that humans have a comparatively small number of genes. For years, estimates have ranged from around 60,000 to 100,000 genes in the human genome. The recent sequencing of the human genome reveals only 32,000 genes compared to 14,000 in *Drosophila* and 19,000 in *C. elegans*. This mere doubling of the number of genes does not reflect the enormous difference in complexity between a human and a simple worm (which has only 959 cells). The question remains: what are the genetic properties that distinguish more complex organisms from simpler ones? One theory is that more complex organisms have higher numbers of *alternatively spliced* genes, resulting in many more proteins per gene than in simpler organisms.

In 1977 it was first discovered that different combinations of exons could be "alternatively spliced" to form multiple RNA molecules from a single gene [12]. Since then, molecular biologists have identified several hundred alternatively spliced genes. It is only recently, however, that the availability of large amounts of genome and expression data have allowed for large-scale computational studies of this phenomenon. While previously it was suggested that the percent of alternatively spliced genes in humans was only 5%, it has recently been estimated that this number is closer to 50% or perhaps even as high as 90%. For a more detailed history see [12, 14].

A study earlier this year [9] uses large-scale expressed sequence tag (EST) analysis to show that "the relatively low number of human genes identified is not accompanied by a uniquely high rate of alternative splicing." They claim that the amount of alternative splicing is essentially comparable among seven different eukaryotes. This claim, although quite possibly true, is *not* conclusively proven by their data. With small numbers of ESTs, they show that there is no statistically significant difference in the percentage of alternatively spliced genes. There are at least three possible problems with this analysis, which deserve further analysis. All three problems can be addressed by breaking their dataset into smaller subgroups.

The first possible problem is that aggregation of multiple ESTs from different sources is very likely to lead to a common statistical error known as Simpson's paradox. When aggregating proportions obtained from differently sized groups, one can easily draw incorrect conclusions [10]. For instance, suppose that genes with immune-related function are alternatively-spliced at a very high level, but genes which code for structural proteins are rarely alternative-spliced. If available cow ESTs are primarily derived from muscle (structural) and human ESTs are mostly from immune cells, then our averages will incorrectly report more alternative splicing in human than cow. This paradox can result in the opposite type of error as well. In order to discover if aggregation is a confounding issue in this analysis it will be necessary to break the data up into sub-classes.

A second issue is that the observed rate of alternative splicing increases with the number of ESTs sampled. In order to make the proportions comparable, the Brett et al study restricted the number of ESTs to 100,000 (the number available for *Arabidopsis*,the organism with the fewest number of ESTs sequenced). Although this may sound like a large amount of data, it's actually not even close to the number needed to saturate their model. First of all, many ESTs are derived from the same gene, meaning that with 100,000 ESTs it's extremely unlikely that they'll have even one EST from all 32,000 human genes. Second of all, because ESTs are very short sequences from the ends of mRNAs, it takes a huge number of ESTs to get even moderate coverage of the entire mRNA sequence. Third, without very large amounts of data, we are unlikely to be able to find alternative splice forms that are rare or occur only in specific tissues.

When Brett et al found no statistically significant differences at this small sampling size they concluded that there is no difference between the rates of alternative splicing in these different organisms. However, another very reasonable explanation is that their test simply does not have enough power. It is very likely that with more data the proportions will begin to diverge. For example, mice and humans seem to have a similar rate of alternative splicing when looking at 100,000 ESTs. With two million ESTs per organism, on the other hand, the authors find 25%- 35% more alternative splicing in humans than in mice. This is a substantial difference and we would like to analyze these 25% of RNAs to try to characterize where the differences are occurring.

Even if all these organisms do have comparable rates of alternative splicing, there is still an unanswered question regarding the relationship between complexity and alternative splicing. There is no clear consensus among biologists that there is a significant difference in complexity between human and mouse or cow (except in cognitive processes), and even between human and *Drosophila* (although from *Drosophila* to worm there is a huge gap, and an even larger gap between worm and yeast). A more interesting analysis would be to look for different rates of alternative splicing in homologous structures or systems that are *clearly* more developed in one organism than in another. Although there is no exact measure of biological complexity, there are many specific biological systems for which biologists generally agree upon a relative ordering among organisms. For example, the forebrain in human is more complex than in other mammals, and the immune system of vertebrates is more highly evolved than that of invertebrates. We would like to know if alternative splicing of specific *types* of mRNAs is associated with increased diversity.

Organism	**mRNA-EST AS Pairs**
Homo sapiens	13,527
Arabidopsis thaliana	2,151
C. elegans	1,395
Rattus norvegicus	986
Mus musculus	784
Drosophila melanogaster	307
Bos taurus	144
Total	**19,294**

Table 1: Number of alternatively spliced mRNA-EST pairs in the ASForms database.

2 Experiments

This section describes the collection and initial analysis of additional annotations for the alternative splicing dataset generated by Brett et. al [9] We consider two different ways of sub-classifying their data: 1) gene function and 2) EST tissue origin.

2.1 Determining Alternative Spliced Genes

We have downloaded the AsForms database [1], which provides alignments of mRNAs with alternatively spliced ESTs derived from seven different organisms. Table 1 lists the number of mRNA-EST pairs for which they found evidence of alternative splicing. Since there are multiple mRNA and ESTs which are derived from the same gene the total number of unique mRNAs and ESTs in their database are thus only 15,693 and 6,790 respectively.

2.2 Classifying Genes into Functional Classes

We have used the controlled vocabulary specified by the Gene Ontology (GO) [4] in order to classify each mRNA (or gene) into a functional class. The GO Ontology contains three main sub-ontologies: molecular function, biological process, and cellular component. Molecular function is defined as "the tasks performed by individual gene products," (examples include transcription factor and DNA helicase). Biological process is defined as "broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions." Lastly, cellular component refers to "subcellular structures, locations, and macromolecular complexes; examples include nucleus, telomere, and origin recognition complex." For the purposes of this study, biological process and molecular function will probably be more useful than cellular component.

The GO database specifies only definitions of and relations between GO terms; it doesn't directly provide gene annotations. The annotations were obtained from other participating databases. The particular databases we used were Flybase [3], the Rat Genome Database (RGD) [7], TIGR [5], and NCBI's Locuslink [6]. Flybase and RGD are organism-specific. Locuslink includes genes from four of the seven organisms in our database: human, mouse, rat, and fly. TIGR includes data for all seven organisms.

Organism	Bio	Molec	B or M	Cell	Total AS Genes
Arabidopsis	79	76	91	59	671
Cow	19	23	24	22	68
Worm	48	58	62	47	524
Fly	65	90	101	63	116
Mouse	221	219	240	180	265
Rat	251	249	289	196	478
Human	2852	2902	3324	2322	4668
All organisms	3535	3617	4131	2889	6790

Table 2: Number of genes in AsForms for which we obtained GO annotations. The columns specify which type of GO term was provided. "B or M" means either a biological process or molecular function or both were specified.

GO Term	GO Ancestor Term
cell-mediated immune response	response to biotic stimulus
perception of pathogenic bacteria	perception of biotic stimulus
immune cell chemotaxis	response to external stimulus
antibody	defense/immunity protein
MHC class I receptor	receptor
interleukin-6 receptor ligand	receptor binding

Table 3: Six non-intersecting GO categories related to the immune system.

Table 2 shows the coverage of the three types of GO annotations for the genes in the AsForms database. The coverage varies significantly for different organisms; about 90% of the alternatively spliced mouse genes are annotated, but only 14% of *Arabidopsis*. For about 40% of all the genes in the AsForms database we were able to find no GO annotations.

Unfortunately, three issues make it difficult to draw conclusions from this data. First of all, the number of alternatively spliced genes is very small, and the loss of the unannotated 40% just exacerbates this problem. As a result, there are very few terms in the GO ontology which map to more than a few different genes. The few terms that do map to larger numbers of genes are primarily cellular location specifiers. Another issue is that the groupings specified by the GO hierarchy do not necessarily group together genes at the level we were looking for. For instance, Table 3 shows six different GO terms related to the immune system. There is no intersection between any of the six terms until the top of the hierarchy is reached. Terms related to immune function are scattered about the GO hierarchy, and this is true (although to differing degrees) for other systems we are interested in as well.

2.3 Determining EST Tissue Source

NCBI's EST database (dbEST [2]) entries contains various fields which relate to tissue source. Unfortunately, however, the text in these fields is unconstrained, which leads to a number of difficulties:

- the vocabulary is inconsistent (T-cell vs. T lymphocyte)
- different types of information are combined in one field (infant brain, ovary tumor)
- extraneous information is given (schizophrenic brain, pregnant uterus)
- different levels of specificity are used (frontal lobe vs. brain, fovea vs. eye).
- many libraries are a pool of multiple samples from different tissues
- many libraries do not specify any clear source information

In an attempt to normalize away these differences, I wrote a script that parses the dbEST entries, mapping natural language phrases and keywords into a controlled vocabulary of different tissue types. In addition, with the consultation of a biologist (Javier Lopez), some tissues were combined into a single class. For instance, large and small intestine were combined into digestive tract, and all parts of the brain were grouped into a single class. Many EST libraries were derived from multiple organs; these were all grouped together into one class, called "pooled." In addition, some specific organs (such as eye, which is composed of connective tissue and neural tissue) were also considered pooled.

The tissue types for *Arabidopsis* are (clearly) not comparable to those for animals. In addition, fly and worm ESTs very rarely had tissue specified: the libraries were described in terms of developmental stage instead of tissue type. For this reason, only tissue information for the four mammals is reported.

The final tissue classification has 25 classes. Percentages for the top 20 classes are listed in Table 4 for each organism. The first column lists the percentage of ESTs originating in each of the 20 different tissue classes. The second column shows the percentage of *ASForms* (alternatively spliced) ESTs from each tissue class.

Table 4 clearly illustrates that the distribution of existing EST tissue sources for a particular organism is strongly biased toward particular tissue types. For instance, for cow there are many ESTs from mammary gland but almost none for neural tissue. For humans, on the other hand, the situation is reversed. Rat has an unusually large number of heart-derived ESTs. These types of differences could lead to significant bias when aggregating ESTs and testing for evidence of alternative splicing. Even if the total number of ESTs sampled is the same for each organism, to avoid Simpson's paradox the sample tissue proportions should be taken into account before drawing conclusions from summary statistics.

The table also shows that the primary factor effecting the percentage of alternatively spliced genes from a particular tissue is the total number of ESTs from that tissue. However, there are a few exceptions. For instance, almost 16% of alternatively spliced human ESTs come from liver, compared to a baseline of 2%. Both mouse and cow have triple the proportion of mammary gland ESTs in their alternatively spliced subset than overall. There could be many possible explanations for these outliers. For instance, the high rate of alternative splicing in human liver could be a result of a human liver EST library having a large number of ESTs which map to one (alternatively spliced) gene. In order to understand whether these outliers are biologically significant (or whether they are just a result of sampling or statistical error) it will be necessary to cluster the ESTs into gene clusters. This will

	Human		Mouse		Rat		Cow	
Class	% ESTs	% AS	% ESTs	% AS	% ESTs	% AS	% ESTs	% AS
unknown	4.10	2.65	17.73	0.55	7.46	1.59	4.27	0.71
pooled	12.77	3.22	13.62	16.64	40.48	48.67	57.79	30.50
neural	12.13	10.53	17.01	12.48	15.00	14.14	3.16	5.67
muscle	7.34	7.02	2.45	2.36	1.84	1.70	5.22	0.71
lung	6.81	5.27	4.07	4.72	1.22	1.06	1.74	1.42
colon	5.75	6.60	1.63	0.83	0.76	0.00	0.00	0.00
immune cell	5.30	7.24	6.06	4.85	0.00	0.00	0.00	0.00
placenta	4.44	1.65	0.72	1.94	3.14	2.23	5.85	14.18
mammary gland	4.03	2.60	**11.68**	**31.35**	0.04	0.00	**10.13**	**29.08**
germ cell	3.81	4.54	0.72	0.42	2.84	4.78	1.62	4.96
digestive tract	3.61	1.41	1.23	0.00	1.20	0.11	0.00	0.00
kidney	3.58	3.29	3.65	6.93	3.12	2.98	3.69	2.84
prostate	3.31	4.60	0.01	0.11	0.90	0.11	0.00	0.00
skin	3.17	0.65	2.28	1.39	0.00	0.00	0.00	0.00
male reproductive	3.17	3.81	3.99	0.83	1.20	1.06	0.00	0.00
pancreas	2.84	1.10	2.96	0.55	0.21	0.21	0.00	0.00
liver	**2.20**	**15.88**	3.20	6.38	0.52	0.64	0.15	0.00
heart	1.53	2.22	1.68	1.80	**16.06**	**18.07**	0.00	0.00
connective tissue	1.15	1.00	0.32	0.83	1.49	1.38	3.42	9.93
spleen	0.28	0.07	1.85	0.97	1.20	1.28	2.54	0.00
number ESTs	4,463,810	10,234	2,612,761	721	347,956	941	231,577	141

Table 4: Percentage of ESTs from 20 different tissues for four mammals. The left column shows the percentage for all ESTs in dbEST, and the second column shows the breakdown for only those ESTs in the ASForms database

also allow us to answer the question we asked in the introduction: “What is the proportion of alternatively spliced genes expressed in specific tissues and do these differ significantly across organisms?”

2.4 Conclusions

1. Getting data is hard: it doesn’t exist or is scattered about in multiple databases or uses non-standard formats and/or identifiers. It also requires getting biologists to make educated decisions about categories (e.g. what organs should be grouped together, how GO terms should be re-clustered).

2. There’s a lot of sampling bias in the data sets, and so breaking the data down into smaller sub-classes is very important.

3. The more sub-classes the data is broken into, the more data is required. (See point one.)

3 Future Work

3.1 Getting More Data

3.1.1 Getting EST-gene mappings

To carry out a complete statistical analysis it will be necessary to obtain EST-mRNA-gene mappings for all the data, not just the alternatively spliced genes. These mappings could possibly be obtained (at least for some of the organisms) from Unigene or TIGR. For the rest we will have to use Blast to align and cluster sequences.

3.1.2 Finding tissue information for remaining ESTs

The script I wrote to extract tissue information for dbEST entries is currently missing about 4-18% (depending on the organism) of ESTs. Many of these ESTs are probably from pooled or unspecified libraries, but it is also possible that library descriptions are present but use different terminology. This means that some of our tissue percentages could be underestimated. There are quite a few possibilities for classifying the tissue source of the remaining ESTs. These are listed in Table 5

3.1.3 Finding functional annotations for remaining genes

Currently I do not have functional annotations for 40% of the genes in the ASForms database. There are many resources for obtaining additional annotations. First, it is quite likely that many of the missing mRNAs might also be in LocusLink, just under a different Accession Number. LocusLink states that it “is not designed to report accession numbers for all published sequences from a given gene.... Other tools such as Entrez ’related sequences’ or BLAST, or BLink can be used to retrieve more records.”

Mouse and human library classification site	Classifies human and mouse libraries
NCBI's unigene library browser classification	Classifies libraries into 200 categories
ATCC's Cell Biology Collection	Contains source information for cell lines.
Anatomical ontologies TissueDB UPenn tissue hierarchy MESH Mouse Anatomical Dictionary Browser	Find tissue source for different tumor types. Find synonyms.
Someone must have done this already!	Find them.

Table 5: Additional resources for finding EST library tissue information.

The members of the GO consortium with the largest number of gene annotations are Compugen and EBI, both of which annotate proteins, not genes. Since GenBank lists the proteins corresponding to many mRNAs, these databases could also be used to annotate genes. Other members of the GO consortium (such as WormBase for *C. elegans* and TAIR for *Arabidopsis*) also contain additional annotations which I have not yet made use of. They use different sequence identifiers than GenBank and it is not always obvious how to find the corresponding sequences.

A third possibility is to use non-GO functional annotations, such as enzyme commission numbers. Genbank, LocusLink, and many other resources provide these types of information, and the GO consortium provides a mapping from the alternative terminologies to GO terms.

In addition, some work will probably need to be done to group GO terms into large categories (such as "immune"). Hopefully someone has done this already; if not, it may have to be done manually (with the help of a biologist).

3.1.4 Finding more alternatively spliced genes

Finally, another way of increasing the amount of data would be to use other existing alternative splicing databases which have been created for the same or different organisms. Alternatively we could create our own alternative splicing database directly from the available mRNA and EST libraries. In this case, the gene clusters built by Unigene and TIGR will most likely be useful. If data is available, it would be very interesting to compare the current organisms with a reptile, bird, or amphibian.

3.2 Other Directions

This project is just a small initial look into one of the many questions regarding alternative splicing. There are many directions in which it could be extended.

One obvious direction is to break ESTs down into non-tissue based classes. For instance, by histology (normal, pre-cancer, cancer, unknown, pooled), or developmental stage (embryo, fetus, larva, adult). Currently I'm combining tumor and normal samples from the same tissue, but cancerous cells might have different alternative splicing characteristics than normal cells. Differences in alternative splicing at different developmental stages would be particularly interesting for fly and worm, for which little tissue information was available.

An additional question that could be asked is whether increased *regulation* (e.g. developmental specificity) of alternative splicing has accompanied increased diversity. The fact that a gene is alternatively spliced does not imply that it is regulated: some alternative splicing appears to be purely stochastic. Nor does alternative splicing imply a regulatory effect. One way of addressing this question is to determine whether alternative splicing is occurring in coding vs. non-coding regions and classify the resulting effect on the protein product. First it will be necessary to try to identify an ORF (this will not always be possible for short ESTs). If we can find the ORF then we can determine if the effect of the splicing was on-off (e.g. ORF truncation) or modification (e.g. optional or variant domains). More specifically, we can identify internal deletions, insertions, substitutions, truncations of the protein product due to frame-shift or introduction of an early stop codon, and changes to the upstream or downstream untranslated regions.

Other possible directions include analyzing the available data on differential expression and regulation of isoforms as well as beginning to explore existing computational methods for identifying context effects on regulation of alternative splicing [11].

References

[1] ASForms database. http://rhodos.bioinf.mdc-berlin.de/splice/asforms/.

[2] dbEST. http://www.ncbi.nlm.nih.gov/dbEST/.

[3] FlyBase. http://flybase.bio.indiana.edu/.

[4] Gene Ontology (GO). http://www.geneontology.org/.

[5] The Institute for Genomic Research. http://www.tigr.org/tdb/tgi/.

[6] LocusLink. http://www.ncbi.nlm.nih.gov/LocusLink/.

[7] Rat genome database. http://rgd.mcw.edu/.

[8] D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich, and P. Bork. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, 474:83–86, 2000.

[9] David Brett, Heike Pospisil, Juan Valcrcel, Jens Reich, and Peer Bork. Alternative splicing and genome complexity. *Nature Genetics*, 30:29–30, January 2002.

[10] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Addison-Wesley, 2002.

[11] A. Javier Lopez. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, 32:279–305, 1998.

[12] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nucleic Acids Research*, 30(1):13–19, January 2002.

[13] Barmak Modrek, Alissa Resch, Catherine Grasso, and Christopher Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, 29(13):2850–2859, July 2001.

[14] Gavin C Roberts and Christopher WJ Smith. Alternative splicing: combinatorial output from the genome. *Curr Opin Chem Biol*, 6(3):375–383, June 2002.