

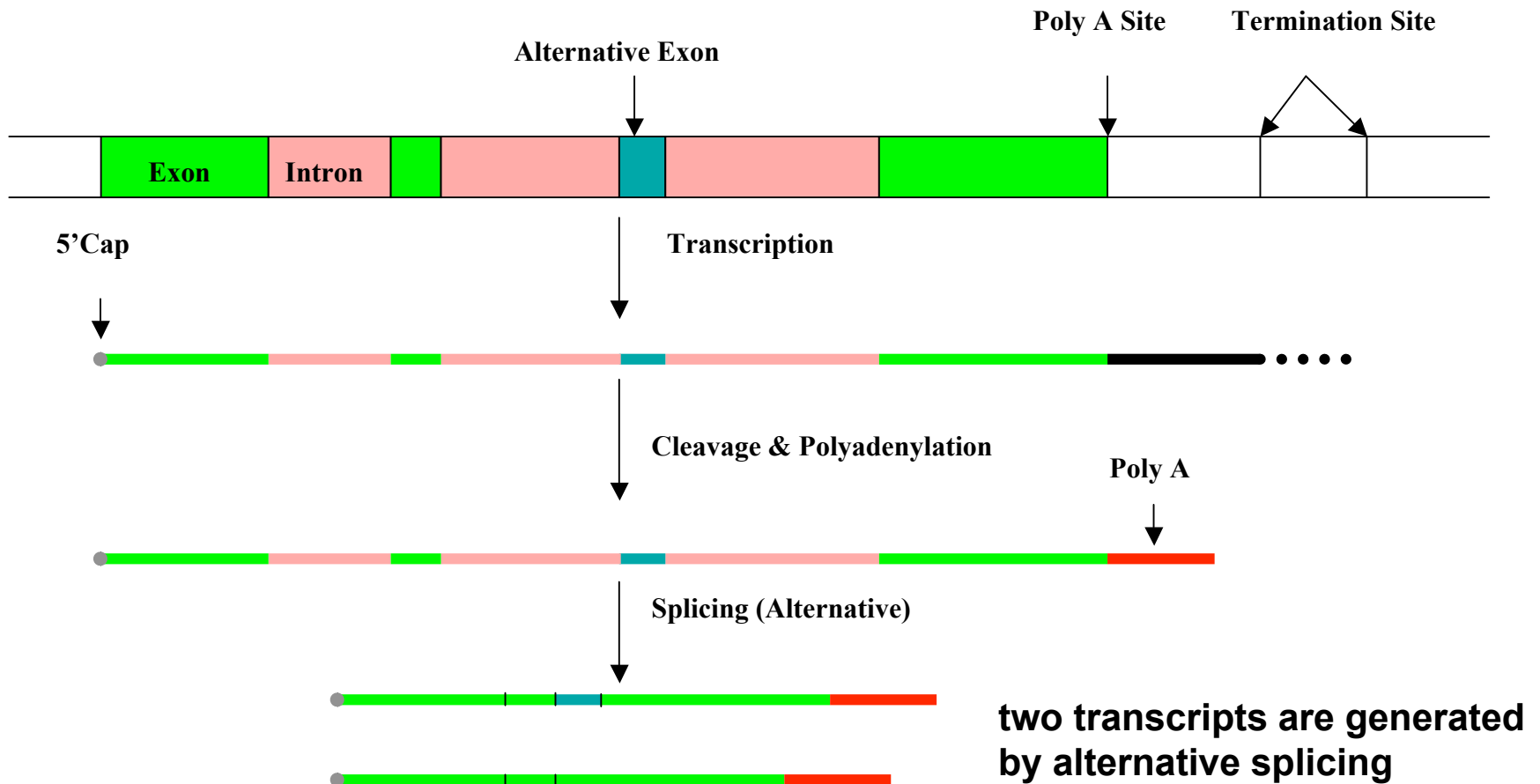
Inferring full-length transcript and protein isoforms from fragmentary sequence data

Yi Xing, Alissa Resch, and Christopher Lee
UCLA-DOE Center for Genomics and Proteomics
Molecular Biology Institute and
Department of Chemistry & Biochemistry
University of California, Los Angeles

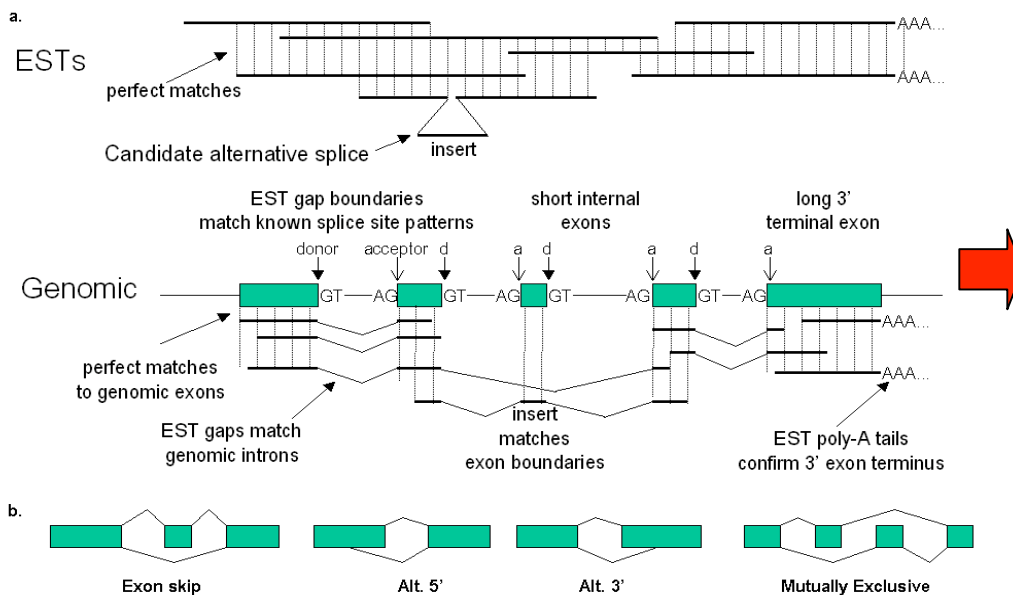
Abstract

Alternative splicing is a frequently observed event in mammalian genomes, and is a major force in increasing proteome diversity. Most high-throughput methods of novel splice variant detection (such as using ESTs, microarray & mass-spectrometry) only generate “local” information about fragments of the full length transcript or protein isoforms. The full length products of those novel splice variants are still left undetermined. We analyze this multi-assembly problem -- reconstructing the most likely set of full length isoform sequences from a mixture of fragmentary data, and provide a graph-based algorithm for solving it. In a variety of tests we demonstrate that this algorithm can appropriately deal with different types of transcript variation, increased fragmentation and removal of full length sequences from input data. Using this algorithm we constructed an Alternatively Spliced Protein database (ASP) by analyzing human genomic and expressed sequences, consisting of 13384 protein isoforms of 4422 human genes. ASP offers a useful resource for experimental discovery and characterization of novel full length isoforms, as well as large scale computational analyses of their proteomic impact.

RNA Processing & Alternative Splicing



Genome-wide detection of alternative splicing showed 30-60% of human genes are alternatively spliced



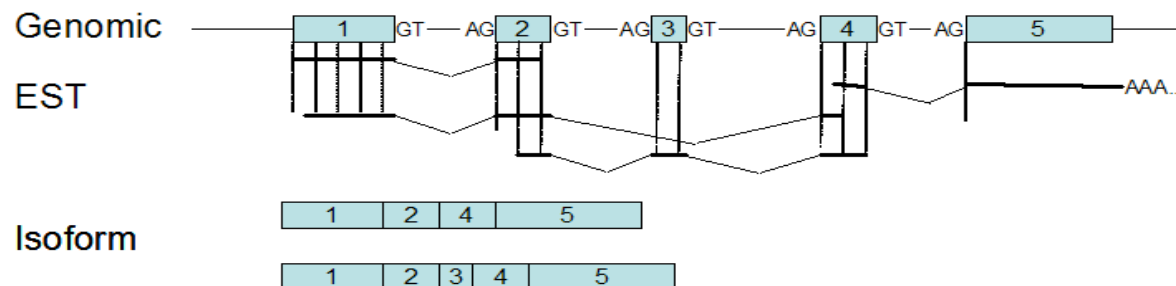
	All Genes		Genes w/ mRNA			
	splices	clusters	splices	clusters		
Total Clusters		96109		20817		
Mapped to Genome		68032	71%	17522	84%	
Detected Splices	133369	18173	27%	121172	12537	72%
Alternative Splice Relationships	30793	7991	44%	28947	7143	57%
(with multiple evidence)	14656	5205		14141	4924	
(novel)	26504	7393		24658	6545	
(novel, with multiple evidence)	10367	4094		9852	3813	

Modrek, B., et al. (2001) *Nucleic Acids Research* 29: 2850-2859
 Lee C. et. al. (2003) *Nucleic Acids Research* 31: 101-105

Assessing functional impact of alternative splicing in human proteome requires a complete catalog of transcript isoforms

	Dec-00	Jan-02
# of Splices	39862	133369
# of Splices in mRNA	24934 (62.6%)	114708 (86.0%)
# of Alternative Splice Relationships	6201	30793
# of Alternative Splice Relationships detected in mRNAs	815 (13.1%)	4289 (13.9%)

over 80% of the alternative splices we detected are *novel*

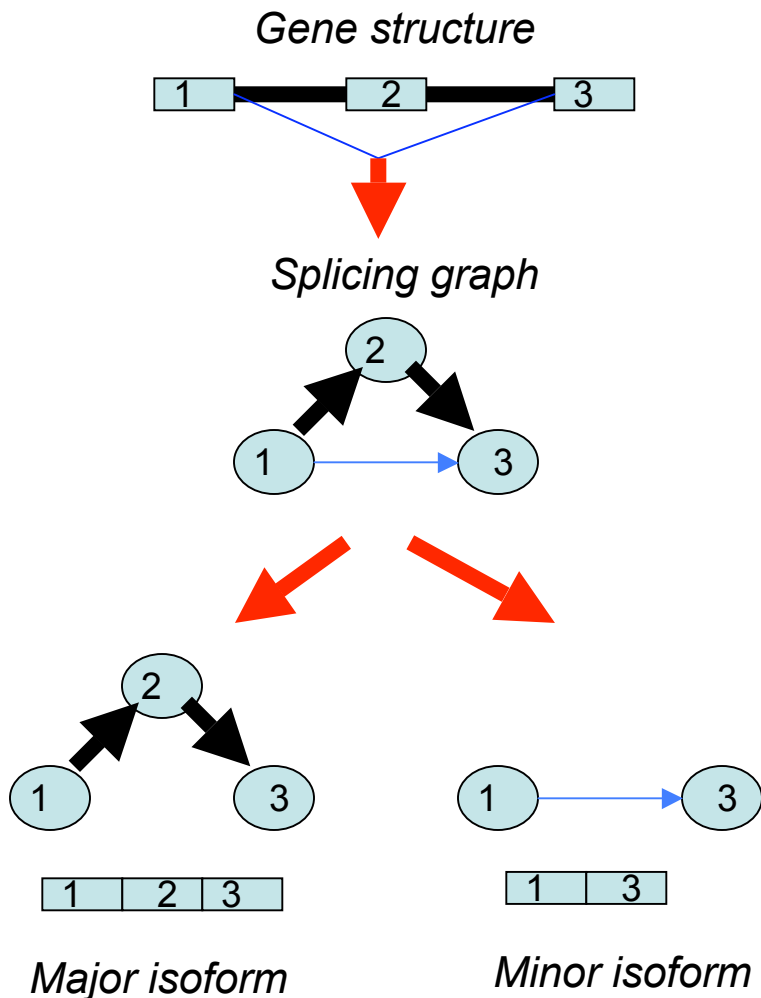


need to *infer* full length isoforms from novel splices detected in EST fragments

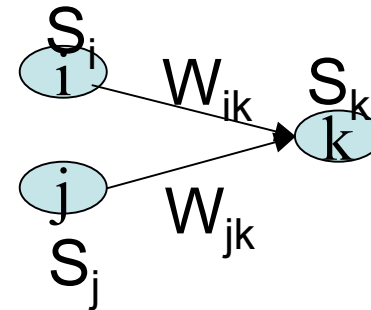
Isoform Problem Requires a Change of Mind in EST Assembly

- Classical EST assembly assumes the existence of a **single** consensus sequence for a set of EST sequences
- The high frequency of alternative splicing in human genes breaks the assumption of a single consensus sequence
- We call this problem a multiassembly problem – assembling **multiple** consensus sequences (isoforms) from a set of fragmentary EST data

Splicing Graph and Heaviest Bundling Algorithm



heaviest bundling (HB) algorithm



If $W_{ik} > W_{jk}$, $S_k = S_i + W_{ik}$

If $W_{ik} < W_{jk}$, $S_k = S_j + W_{jk}$

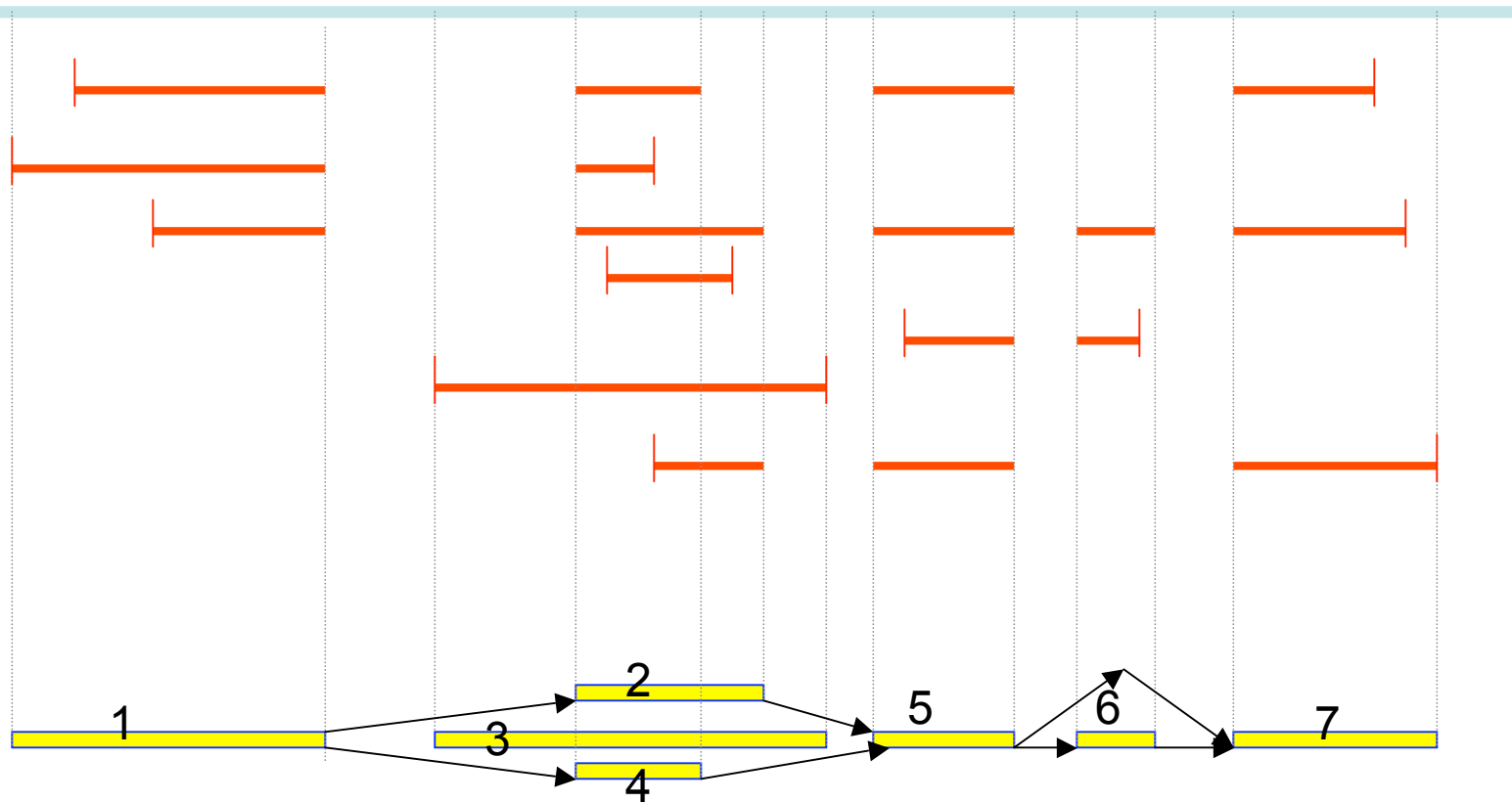
If $W_{ik} = W_{jk}$, $S_k = W_{ik} + \max(S_i, S_j)$

Finding traversals in the splicing graph maximizing EST observations

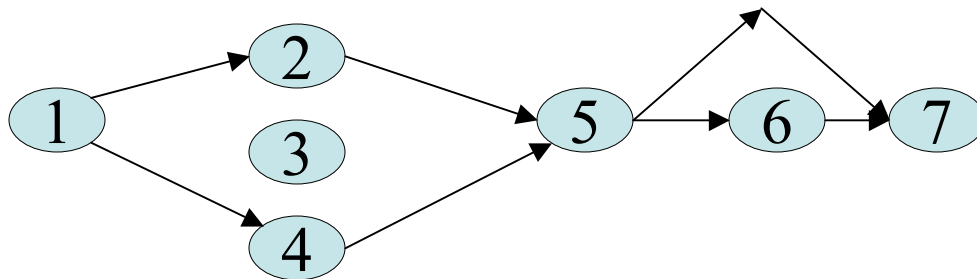
Constructing Splicing Graphs from EST-Genomic Alignments

GENOMIC

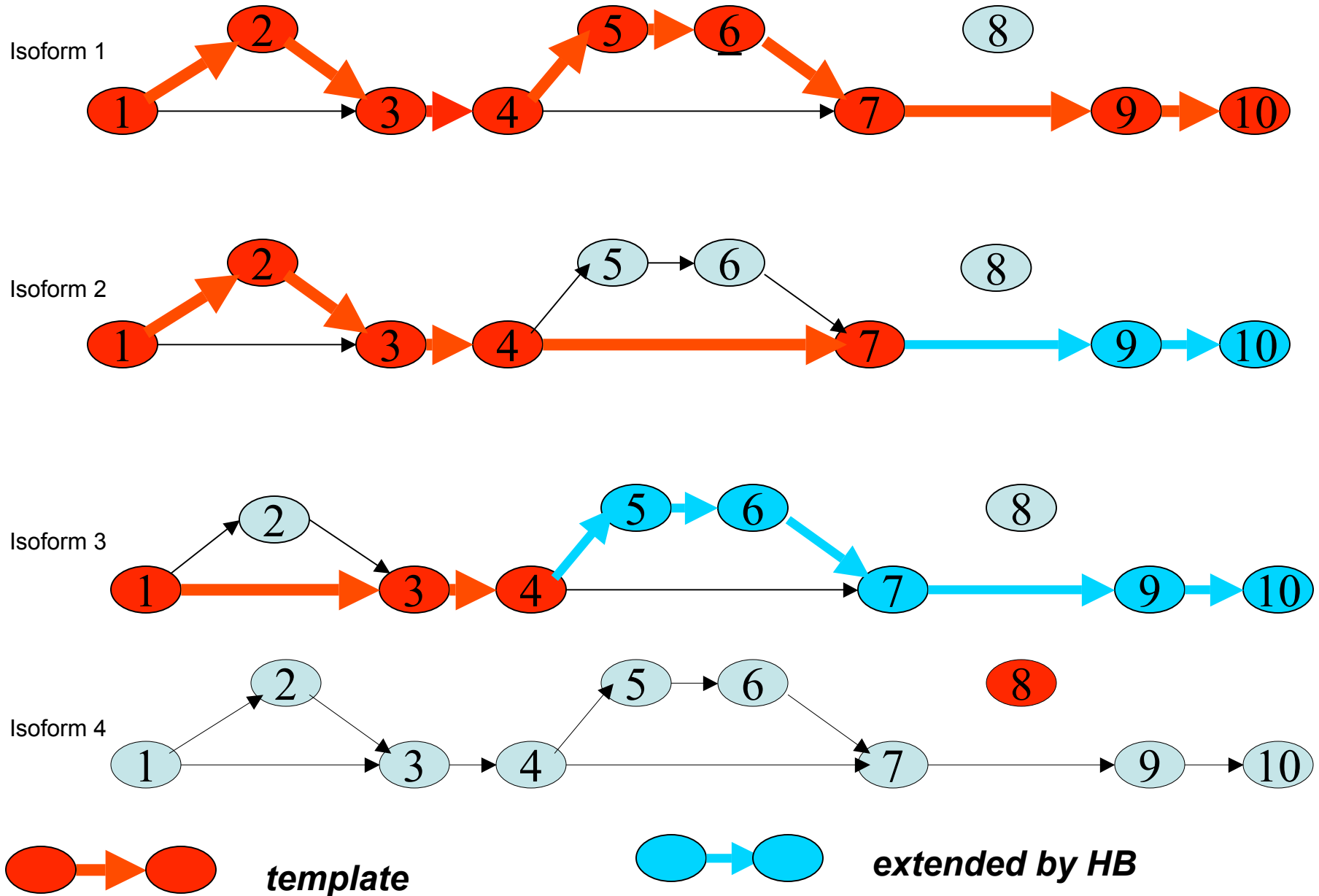
ESTs

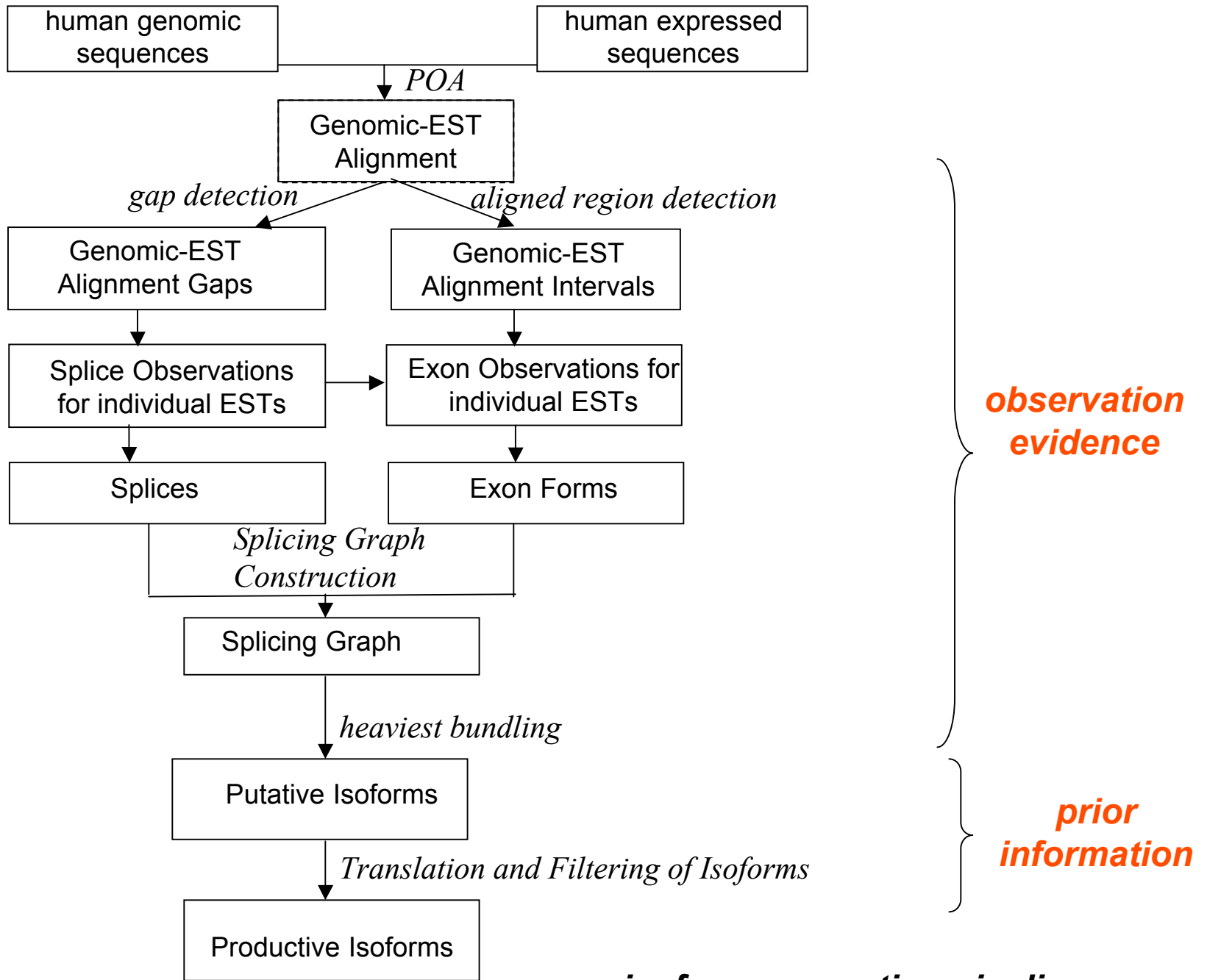


Splicing Graph



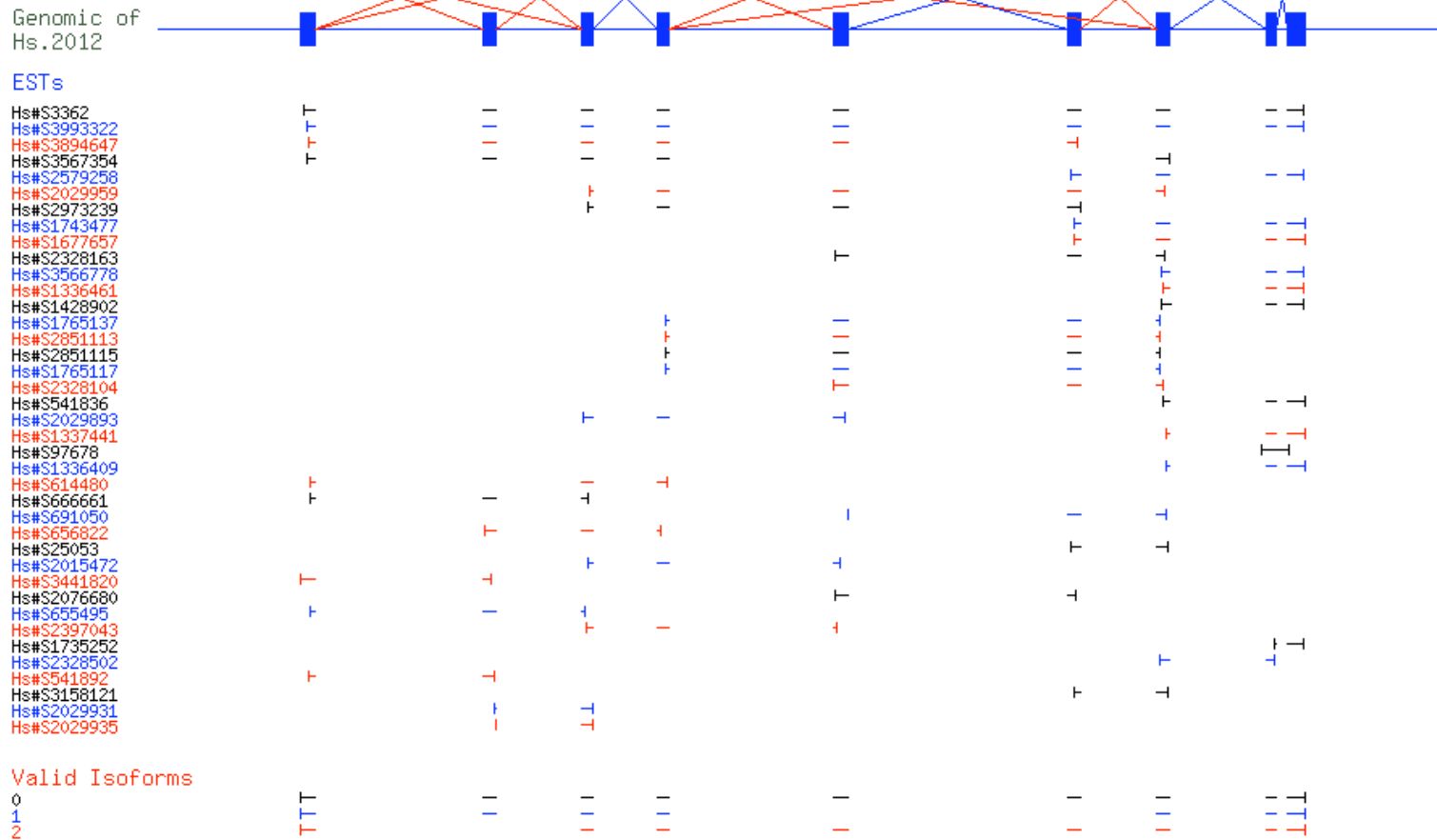
Splicing Graph and Heaviest Bundling of Hs.2012



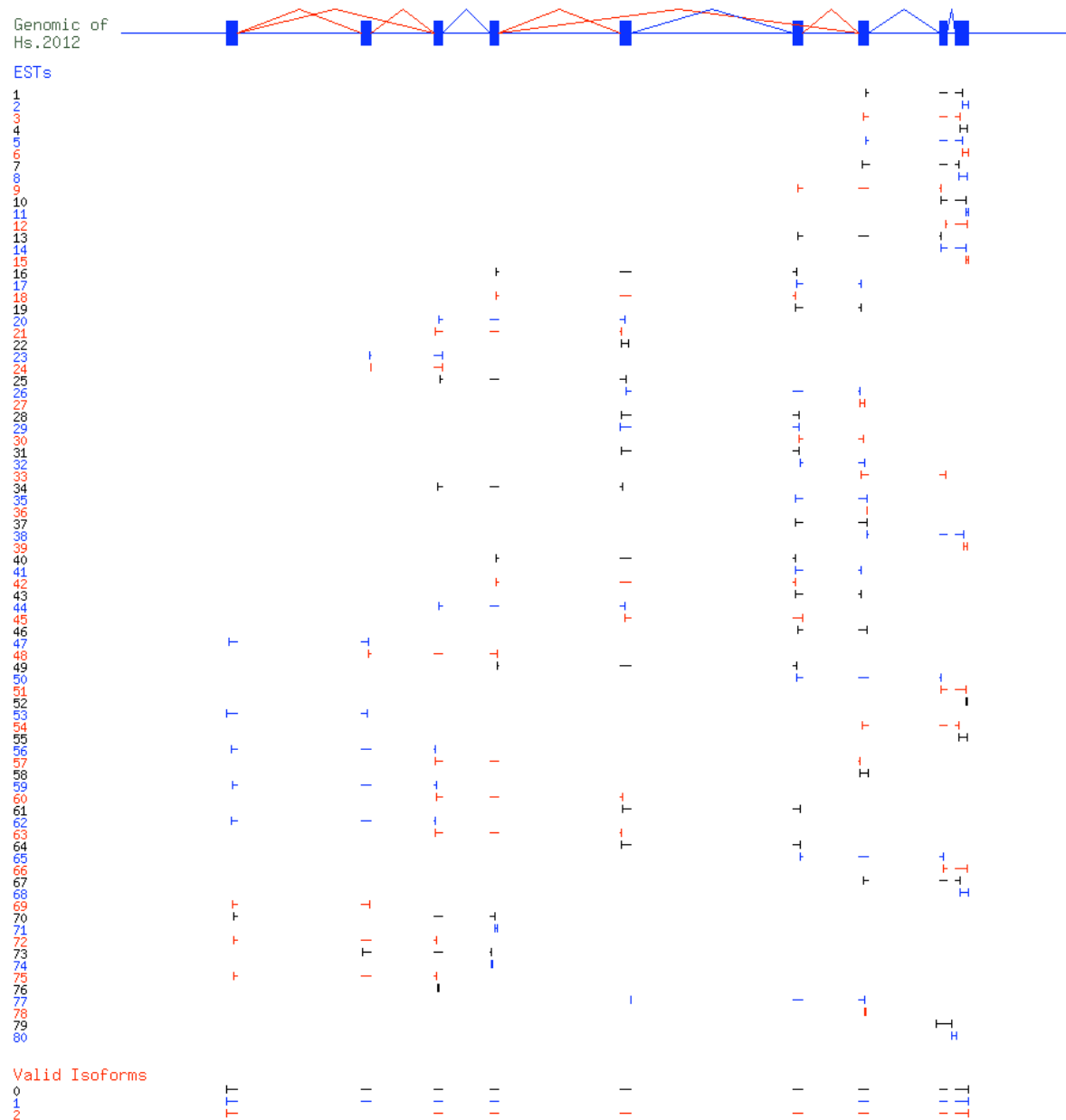


isoform generation pipeline

Alternative Splicing and Isoform Generation of Hs.2012 (TCN1)



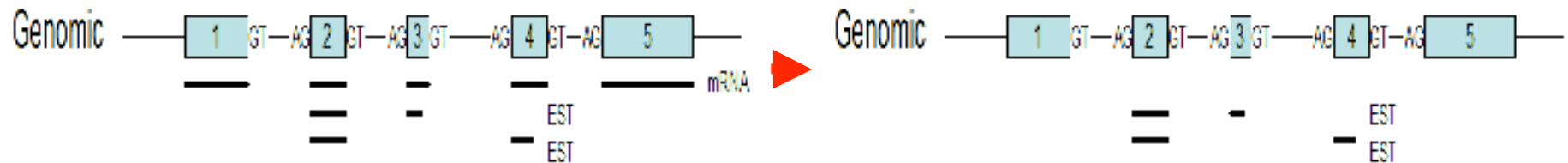
Isoform generation is robust to increased fragmentation of input EST data



chop ESTs into very short fragments, still generate isoforms correctly

Isoform generation doesn't rely on mRNA sequences as reference sequences

	With mRNA	mRNA removed from input sequences	
Genes	12	12	
Isoforms	40	Same	32
		Data Coverage Error	8
		Algorithmic Error	0



data coverage error: removing mRNA here leads to inadequate EST coverage on the gene

Database of Alternatively Spliced Proteins (ASP) in human

	UniGene Cluster		Isoforms	
Total Clusters	96109			
Mapped to Genome	68032	71%		
Detected Consensus Splices	18173	27%		
Produced Putative Isoforms	17581	97%	61508	
Productive mRNA Isoforms	13608	77%	29204	47%
Multiple Productive Isoforms Per Gene	6636	49%	22232	76%
Distinct Protein Isoforms			17742	80%
ASP (Alternatively Spliced Proteins) Database	4422		13384	

Discoveries using ASP (<http://www.bioinformatics.ucla.edu/ASP/>)

- [Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., Lee, C. \(2003\) Assessing the Impact of Alternative Splicing on Domain Interactions in the Human Proteome. J. Proteome Res.](#)
- [Xing, Y., Xu, Q., Lee, C. \(2003\) Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. FEBS Lett. 555: 572-8](#)

References

1. Xing, Y., Resch, A., Lee, C. (2004) The Multiassembly Problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Research* in press
2. Modrek, B., Resch, A., Grasso, C., Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research* 29: 2850-2859
3. Lee, C (2003) Generating Consensus Sequences from Partial Order Multiple Sequence Alignment Graphs. *Bioinformatics* 19: 999-1008
4. Lee, C., Atanelov, L., Modrek, B., Xing, Y. (2003) ASAP: The Alternative Splicing Annotation Project. *Nucleic Acids Res.* 31: 101-5.
5. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA. (2002) Splicing graphs and EST assembly problem. *Bioinformatics*. Suppl 1:S181-8
6. Kan Z, Rouchka EC, Gish WR, States DJ. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research*. 11(5):889-900