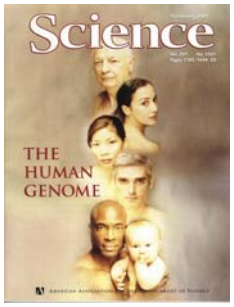
Genome Informatics conference 2008

Extracting biological knowledge from SVM models capable of
distinguishing alternative and constitutive splicing

T. Murlidharan Nair and Michael Gribskov

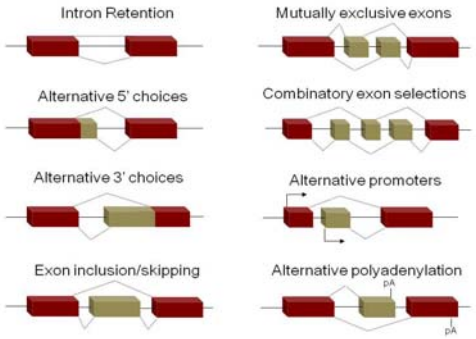
Alternative and Constitutive splicing

- Sequencing of the human genomes revealed that it is merely the skeleton of the underlying code of life
- Many layers of annotation will be required before it can be completely deciphered
- An important observation that came to light was the small number of genes identified from the genome



Alternative and Constitutive splicing

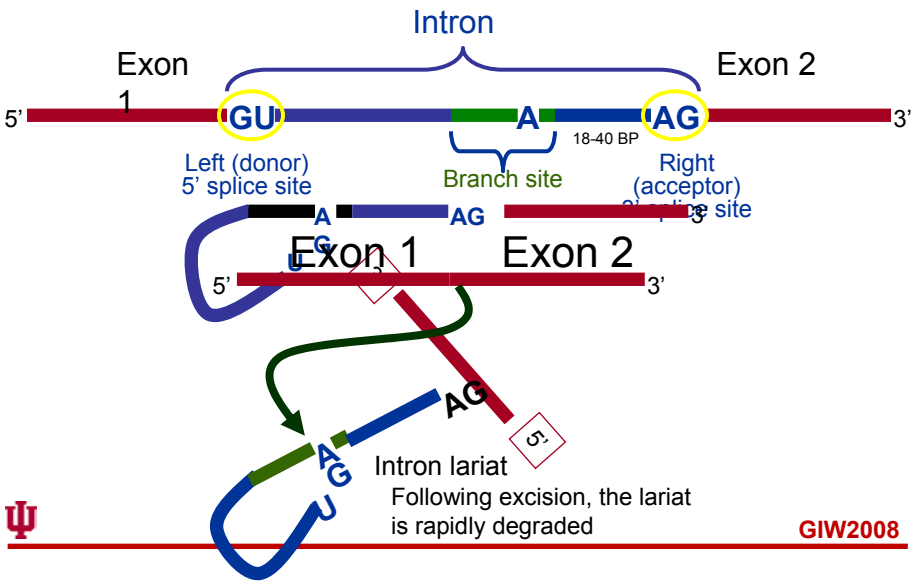
- Alternative splicing of pre-mRNA is a major contributor to proteomic diversity as well as control of gene expression levels.
- Splicing is regulated developmentally as well as tissue specifically and aberrant splicing can lead to wide range of human diseases



Alternative and Constitutive splicing

- Alternative splicing of pre-mRNA is a major contributor to proteomic diversity as well as control of gene expression levels.
- The long term goal is to understand the "splicing code" that will enable the prediction of splicing patterns from primary mRNA transcript.
- Core splicing signals are present in every intron
  - 5'splice site
  - 3' splice site
  - Branch point

Alternative and Constitutive splicing



Alternative and Constitutive splicing

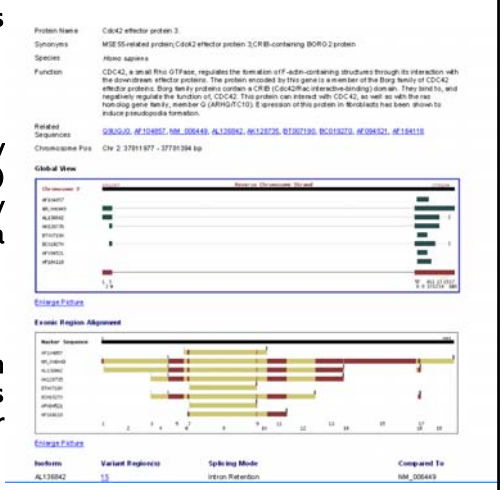
- 5'ss and 3'ss can be mapped precisely by aligning cDNA/ESTs to the genome and there is a large corpus of data available that can be used to build models
- Branch point sequence data sets are not as large
- Exon definition is an important step in mammalian splicing regulation
- Splicing process occurs with high fidelity despite the presence of pseudo-exons

Alternative and Constitutive splicing

- Splicing regulatory elements
  - ESEs and ESSs promote or inhibit inclusion of exons that they are part of
  - ISEs and ISSs promote or inhibit usage of adjacent splice sites or exons. They are present in introns.
- Can we identify distinguishing characteristics present in alternatively and constitutively spliced junctions using learning methods?
- Compare these with know regulatory elements and identify common subsets.

Alternative and Constitutive splicing

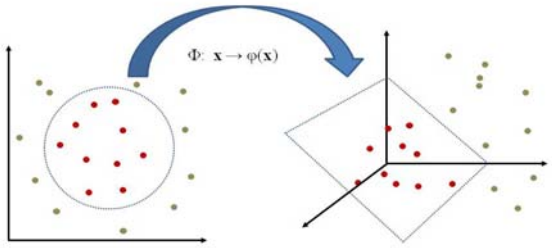
- For building learning methods requires good quality data
- Manually annotated alternatively spliced events database (MAASE) (Zhang et. al., 2005) contains highly curated data http://splice.bioinformatics.iusb.edu
- The database contains information on alternatively splice events and was developed to design primers for splicing arrays.



Alternative and Constitutive splicing

- Support vector machines are kernel based classifiers used for classification and regression.

- They classify input data by transforming them into an n-dimensional space and constructs a maximum margin hyperplane in that space separating the two classes.



Alternative and Constitutive splicing

- One of the problems with using SVMs as compared to other probabilistic methods, like weight matrices and Markov model, is that the resulting decision function is not easily interpretable. Thus no relevant biological knowledge can be extracted from them.

- Is it possible to understand the internal representation of an SVM model and use that information to extract biologically important features?

Alternative and Constitutive splicing

Algorithm Interpreting the SVM classifier for biological relevance

Inputs : *S* **the set of test examples not used for training the SVM; f(i), set of SVM models capable of optimally classifying** *S*

Output: *R*, **enumerated output reflecting the performance of the SVM classifier based on the region where the noise was introduced.**

W =Size of the window to introduce the noise

Begin

i =1

for m= 1 to 100 (the number of svm models)

S_{rand} **=Randomized subset set of test examples**

while j ≤ number of $S_{randomized}$ **do**

S_{rand} **[1 to W] = random sequence with the same probability distribution**

end while

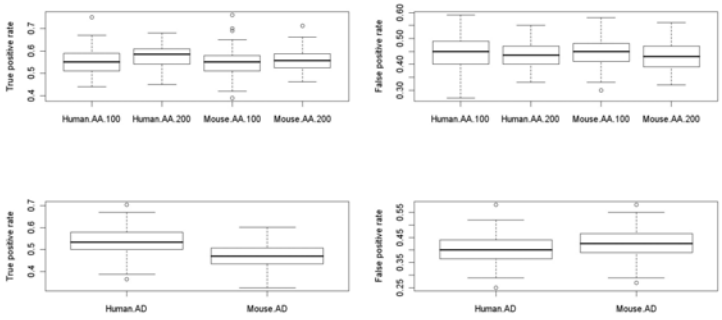
R[m]=f($S_{randomized}$ **)**

end for

end

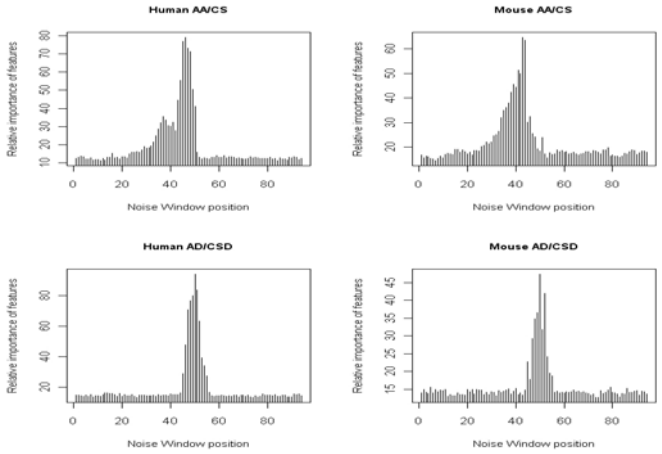
Alternative and Constitutive splicing

- **Separate Support vector machine (SVM) models were built to distinguish between**
  - **Constitutively spliced acceptor and alternatively spliced acceptor junctions**
  - **Constitutively spliced donor and alternatively splice donor junctions.**



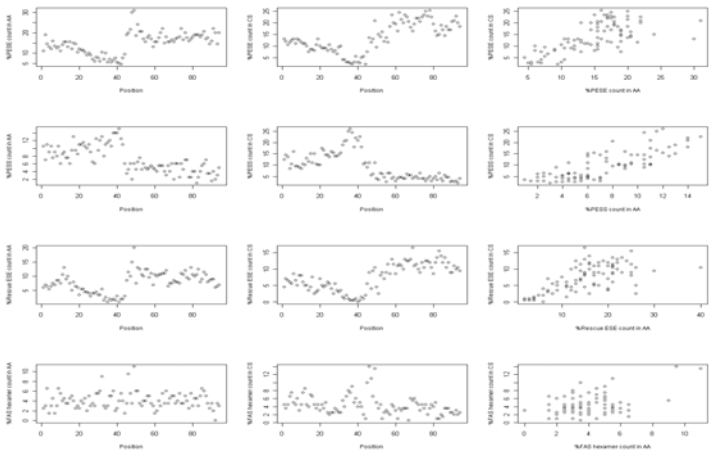
Alternative and Constitutive splicing

- **Feature extraction algorithm was then used to identify learned information.**



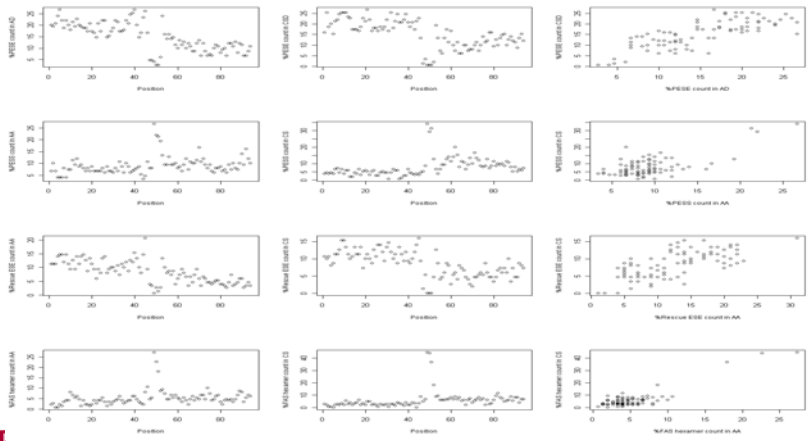
Alternative and Constitutive splicing

Percentage position occurrence of hexamers (silences/enhancers) around human alternatively spliced acceptor and constitutively spliced acceptor junctions



Alternative and Constitutive splicing

Percentage position occurrence of hexamers (silences/enhancers) around human Alternatively spliced donor and constitutively spliced donor junctions.



Alternative and Constitutive splicing

- **Preliminary results point to the fact that features present are complex to capture**
- **The learning space for SVM needs to be increased which will improve the performance of the model.**
- **With the SVM model we hope to contribute to the parts list towards derivation of the "splicing code"**
- **Eventual goal is to develop a splicing simulation algorithm by integrating the information derived with SVM models with those available in literature by developing a probabilistic model.**

Alternative and Constitutive splicing

Acknowledgements

- Prof. Xiang-Dong Fu UCSD
- Dr. Christina Zhang
- NSF for funding