Splice Form Prediction using Machine Learning

G. Rätsch[1], S. Sonnenburg[2,*] J. Srinivasan[3,*] H. Witte[4] K.-R. Müller[2] R. Sommer[4] B. Schölkopf[5]

[1] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany; [2] Fraunhofer FIRST, Berlin, Germany [3] CalTech, Division of Biol., Pasadena, USA; [4] MPI for Developmental Biology, Tübingen; [5] MPI for Biological Cybernetics, Tübingen; * contributed equally

For modern biology, precise genome annotations are of prime importance as they allow the accurate definition of genic regions. However, accurate *ab initio* gene finding is still a major challenge in computational biology. We employed state-of-the-art machine learning methods to assess and improve the accuracy of genome annotations. Our system is trained to recognize exons and introns on the unspliced mRNA. First, we have developed a novel Support Vector Machine (SVM) based method that very accurately predicts splice sites. Then, we adopted a so-called *label sequence learning* technique similar to Conditional Random Fields and Hidden Markov SVMs [3] to the problem of predicting the splice form of a gene. The parameters of mappings (shown as green arrows in Figure 1) determine the contribution of the detector outputs to the score. During training they are adjusted to *maximize the margin* between the true splicing and all other ones (one of them is shown in red). The prediction on new genes works by selecting the splicing with the best score via dynamic programming.

We applied our system, called *mSplicer*, to the genome of *C. elegans* in order to improve its annotation. In 87-95% of all tested genes, our method correctly identified all exons and introns. Notably, only 37-50% of the presently unconfirmed gene annotations agree with our predictions. We hypothesized that a sizable fraction are not correctly annotated. A retrospective evaluation of the WormBase WS120 annotation [1] revealed that splice form predictions on unconfirmed gene segments in WS120 are inaccurate in about 18% of the considered cases, while our predictions deviate in only 10-13%. We experimentally analyzed 20 controversial genes on which our system and the annotation disagree. While our method correctly predicted 75% of those cases, the standard annotation was never completely correct. We conclude that the genome annotation of *C. elegans* can be greatly enhanced using modern machine learning.

Our method is the first that learns to predict splice forms discriminatively. A benefit compared to generative probabilistic methods is that it can be extended to include additional features for instance related to alternative splicing. Also, the method can in principle be extended to the simultaneous prediction of several several alternative isoforms, which would allow us to learn to predict so called splice graphs of a gene. We are currently combining this idea with our previous work on alternative splicing (e.g. [2]) and intend to present preliminary results at the conference.

[1] T.W. Harris, N. Chen, F. Cunningham, et al. Wormbase: a multi-species resource for nematode biology and genomics. *Nucl. Acids Res.*, 32, 2004. D411-7.

[2] G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21(S1):i369–i377, 2005.

[3] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured output spaces. *J. Mach. Learn. Res*, 6:1453–1484, 2005.

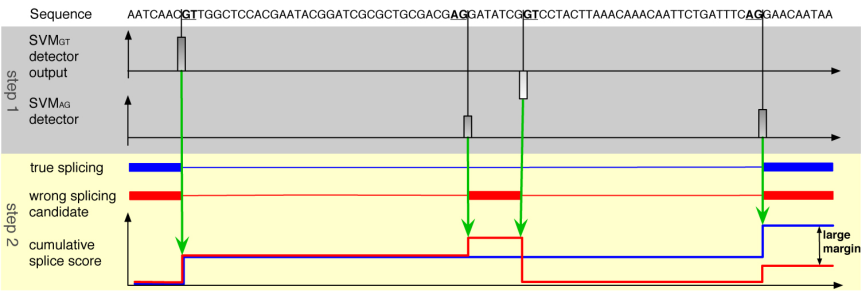

Figure 1: Given the start of the first and the end of the last exon, we first scan the DNA using SVM detectors trained to recognize intron starts ($\mathrm{SVM}_{\mathrm{GT}}$) and ends ($\mathrm{SVM}_{\mathrm{AG}}$): they assign an output to each candidate site. Each putative splicing gets a score to which the SVM splice site predictions and other information, such as the exon/intron lengths, contribute. Large Margin Learning determines the optimal contribution of the different factors to the splicing score.