Poster H-3

Splice Form Prediction using Machine Learning

Authors:
Gunnar.Raetsch@tuebingen.mpg.de (*Friedrich Miescher Laboratory of the Max Planck Society*)
Sören Sonnenburg (*Fraunhofer FIRST*)
Jagan Srinivasan (*Caltech*)
Klaus-Robert Müller (*Fraunhofer FIRST*)
Ralf Sommer (*MPI Developmental Biology*)
Bernhard Schölkopf (*MPI Biological Cybernetics*)

Short Abstract: Accurate ab initio gene finding is still a major challenge in computational biology. We employ cutting edge machine learning similar to Hidden-Markov-SVMs to assay and improve the accuracy of genome annotations. We applied our system on the C_elegans genome and were able to drastically improve its annotation.

Long Abstract:
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 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 the prediction of splice

graphs representing several alternative isoforms — an important problem in genome analysis. An easy way is to extend the state-space to the product space, leading to several alternative intron/exon segmentations. This allows us to learn to predict splice graphs of a fixed maximal width. We are currently combining this idea with our previous work on alternative splicing (e.g. [2]). Preliminary experiments indicate that using this technique one can accurately predict splice graphs for C. elegans.

[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.