

mGene: Accurate Gene-finding with Discriminative Methods

Gabriele Schweikert,^{1,2,3} Georg Zeller,^{1,3} Alexander Zien,^{1,4} Jonas Behr,¹ Christoph Dieterich,³ Cheng Soon Ong,^{1,2} Petra Philips,¹ Anja Bohlen,¹ Lisa Hartmann,¹ Nina Krüger,¹ Sören Sonnenburg,⁴ and Gunnar Rätsch

1 Introduction

Over the last few years the advent of next generation technologies has rendered genome-wide sequencing a routine process, which results in a quickly increasing number of completed genome projects. Harvesting the concealed biological meaning from this sea of data is now the true challenge. A first step in such an annotation process requires the detection of genes and the prediction of their precise structure, including untranslated regions (UTRs), exons and introns. To this end we have developed an accurate computational gene finding system, that is based on state-of-the-art discriminative learning techniques [1]. The long-term goal of this project is to provide a precise, ready-to-use gene finding system, that can be employed by biological practitioners to achieve a precise annotation for their genome of interest.

Our approach, mGene, combines discriminative machine learning techniques to solve a so-called structured output learning problem (see Figure 1). In a first step, we identify various signal sequences such as transcription and translation start or stop as well as splice sites [2].

These classification problems are solved independently using Support Vector Machines with specifically designed string kernels. In a second step, we approach the gene structure prediction problem with a novel label sequence learning algorithm related to generalized hidden Markov models [3, 4]. As input it takes the predictions from the first step, indicating possible transitions between segments (i.e. exon, intron, UTR, intergenic). Together with additional content information (e.g. coding potential), all information is combined into globally optimal gene structures. Using discriminative learning techniques throughout, we enforce a large margin between the score of the true gene structure and all other wrong structures.

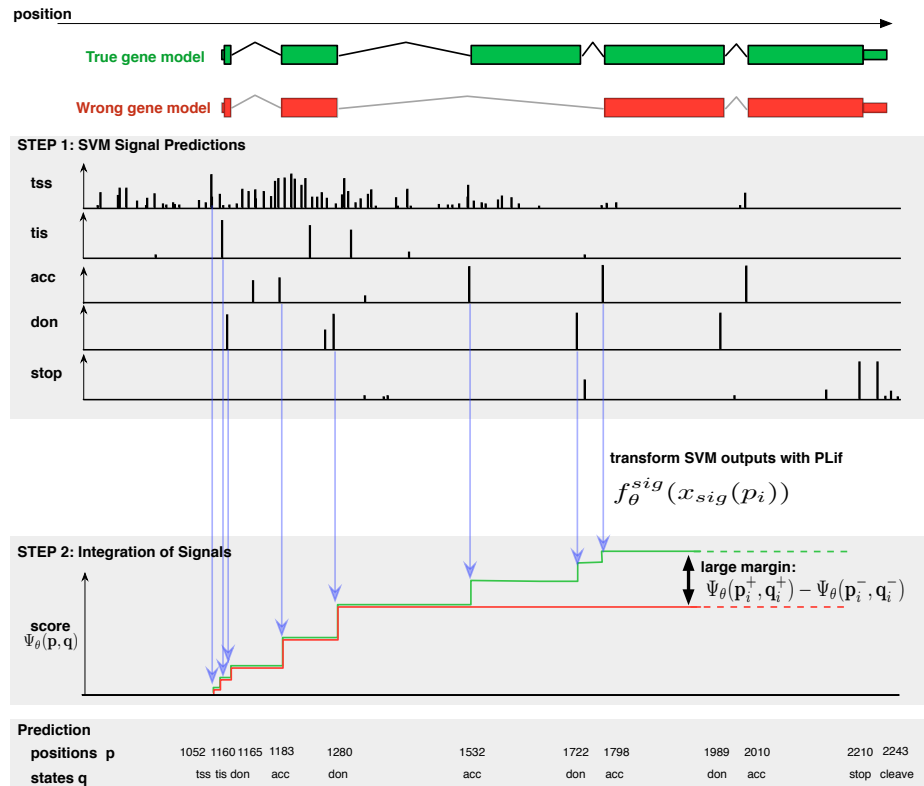


Figure 1: Two-step architecture employed by mGene: First individual signals are predicted independently with SVMs. These predictions are subsequently integrated with HSM-SVMs to form a valid gene structure.

¹Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany

²Max Planck Institute for Biological Cybernetics, Tübingen, Germany

³Max Planck Institute for Developmental Biology, Tübingen, Germany

⁴Fraunhofer Institute FIRST.IDA, Berlin, Germany

2 Experiments

Recently an international competition for automated annotation of nematode genomes, the nGASP project, showed the excellent performance of mGene on the model organism *C. elegans* when compared to 47 submitted predictions from seventeen research groups (see Table 1 and [5]). Testing a subset of 57 of 2200 newly predicted genes by PCR and sequencing we could further show the high accuracy of mGene. These findings suggest that even the gene catalog of a well-studied organism such as *C. elegans* can be substantially improved by mGene predictions. We have therefore created gene predictions for four newly sequenced nematode genomes, which have been included into the official wormbase genome browser and will constitute a valuable resource for the community.

	Base	Exon	Gene
Craig	93.23	79.16	39.58
Eugene	91.72	76.64	44.16
Fgenesh	92.65	79.96	45.90
Augustus	93.01	79.34	49.42
mGene	93.83	82.64	51.49

Table 1: Comparison of *ab initio* gene predictors evaluated according to the rules of the nGASP challenge [5, 1]; the average of sensitivity and specificity for single base, exon and whole gene evaluation is reported.

3 Discussion and Outlook

mGene's high accuracy is based on state-of-the-art predictions of functional elements using Support Vector Machines [3,4]. Equally important is a discriminative learning technique to appropriately combine information on parts of genes in order to predict gene structures [3]. In contrast to many HMM-based gene finders, mGene has the further advantage of being very flexible in terms of incorporating additional input data, such as conservation information (e.g. [1]). We have recently extended mGene such that transcriptome measurements are taken into account as complementary features. In this context we examined the benefit of Illumina-based mRNA-seq and Affymetrix-based transcriptome tiling array measurements with the expectation that the latter are advantageous for the lowly expressed transcripts and mRNA-seq is better suited to define exon/intron boundaries (preliminary results are given on the poster).

We have recently developed a web-service that provides all necessary tools to train mGene and to perform predictions using mGene. It is freely available at: <http://galaxy.tuebingen.mpg.de>.

References

- [1] Schweikert, G., G. Zeller, A. Zien, J. Behr, C. Dieterich, C.S. Ong, P. Philips, A. Bohlen, L. Hartmann, N. Krüger, S. Sonnenburg and G. Rätsch: mGene: Accurate Computational Gene Finding with Application to Nematode Genomes Under review for *Genome Research* (2009).
- [2] Sonnenburg, S., G. Schweikert, P. Philips, J. Behr and G. Rätsch: Accurate Splice Site Prediction using Support Vector Machines. *BMC Bioinformatics*, (Suppl 10):S7 (2007).
- [3] Rätsch, G., S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. Müller, R. Sommer and B. Schölkopf: Improving the *C. elegans* genome annotation using machine learning. *PLoS Computational Biology*, 3(2):e20 (2007).
- [4] Rätsch, G and S. Sonnenburg: Large Scale Hidden Semi-Markov SVMs. *Advances in Neural Information Processing Systems*, vol. 19, pp. 1161-1168, Cambridge, MA, MIT Press (2007).
- [5] Coghlan, A., T.J. Fiedler, S.J. McKay, P. Flicek, T. Harris, D. Blasiar, the nGASP Consortium, and L.D. Stein: nGASP - the nematode genome annotation assessment project. *BMC Bioinformatics*, 9:549 (2008).