mGene: A Novel Discriminative Gene Finding System

G. Schweikert,* A. Zien,* G. Zeller,* C.S. Ong, F. de Bona, S. Sonnenburg, P. Philips, and G. Rätsch

Contact: Gunnar.Raetsch@tuebingen.mpg.de

Motivation: As an increasingly large number of genomes is being sequenced and has to be annotated, the computational problem of gene finding has never been more important. We present a novel discriminative machine learning technique [1, 2] for predicting structured outputs used to build an accurate computational gene prediction system, called mGene.

Results: We decouple the gene prediction problem into two steps: First, we identify various signal sequences such as transcription and translation start or stop sites, as well as exon/intron boundaries (i.e. splice sites). All these classification problems are solved independently using Support Vector Machines (SVM) employing an appropriate combination of string kernels [2, 4]. We use our freely available Shogun toolbox [3] allowing us to train on millions of examples and to achieve higher recognition accuracies.

Second, we approach the gene structure prediction problem with a newly developed discriminative *label sequence learning* algorithm, which is related to generalized Markov models and similarly described in [1, 2]. As input it takes the predictions from the first step which indicate possible transitions between segments, i.e. a splice donor signal marks the transiton from exon to intron. Together with additional content information (e.g. coding potential or length distributions), all signal information is combined into globally optimal gene structures.

Due to its modular architecture, mGene can be readily extended from an *ab initio* gene finder to one which also incorporates extrinsic information, such as sequence conservation or known transcript sequences.

With mGene we participated in the nGASP genome annotation competition on nematode genomes. While the official evaluation of the predictions will be announced in the near future, we made a preliminary assessment comparing our predictions to those of our competitors and to experimentally annotated genes as a ground truth. Each of three submission categories (*ab initio*; using conservation; using known sequences) was evaluated on nucleotide, exon and transcript level and according to seven out of these nine criteria, mGene performed best among 15 other methods including Genscan, Fgenesh and Augustus.

[1] G. Rätsch and S. Sonnenburg. Large scale hidden semi-Markov SVMs. In *Proc. NIPS'06*. MIT Press, 2007. In press.

[2] G. Rätsch, 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.

[3] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, pages 1531–1565, 2006. Special Topic on "Machine Learning and Large Scale Optimization".

[4] S. Sonnenburg, A. Zien, and G. Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. pages e472–e480, July 2006.

*authors contributed equally