BMC Bioinformatics

Oral presentation

KIRMES: kernel-based identification of regulatory modules in euchromatic sequences

Sebastian J Schultheiss[1,2], Wolfgang Busch[2,3], Jan Lohmann[2,4], Oliver Kohlbacher[5] and Gunnar Rätsch[1]

Address: [1]Machine Learning in Biology Research Group, Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tuebingen, Germany, [2]Max Planck Institute for Developmental Biology, 72076 Tuebingen, Germany, [3]Biology Department, Duke University, Durham, NC 27710, USA, [4]Department of Stem Cell Research, University of Heidelberg, 69120 Heidelberg, Germany and [5]Simulation of Biological Systems, Wilhelm Schickard Institute for Computer Science, University of Tuebingen, 72076 Tuebingen, Germany

from Fifth International Society for Computational Biology (ISCB) Student Council Symposium
Stockholm, Sweden 27 June 2009

Published: 19 October 2009

BMC Bioinformatics 2009, **10**(Suppl 13):O1 doi: 10.1186/1471-2105-10-S13-O1

This article is available from: http://www.biomedcentral.com/1471-2105/10/S13/O1

Background

We predict transcription factor (TF) target genes based on their regulatory sequence. A TF binding site is a short segment (~10 bp) near a gene's regulatory region that is recognized by respective TFs. Overrepresented motifs can be identified in regulatory sequences of a set of genes that is enriched with targets for a specific TF. Gibbs-sampling methods that try to identify position weight matrices to characterize binding sites have been successful for small genomes, but are problematic in higher eukaryotes, where motifs are degenerate and form *cis*-regulatory modules [1].

Methods

Our method classifies genes as TF targets. We use *de novo* motif finding and subsequently apply a Support Vector Machine employing a kernel that captures information about the motifs, their relative location, and sequence conservation (see Figure 1). The weighted degree kernel

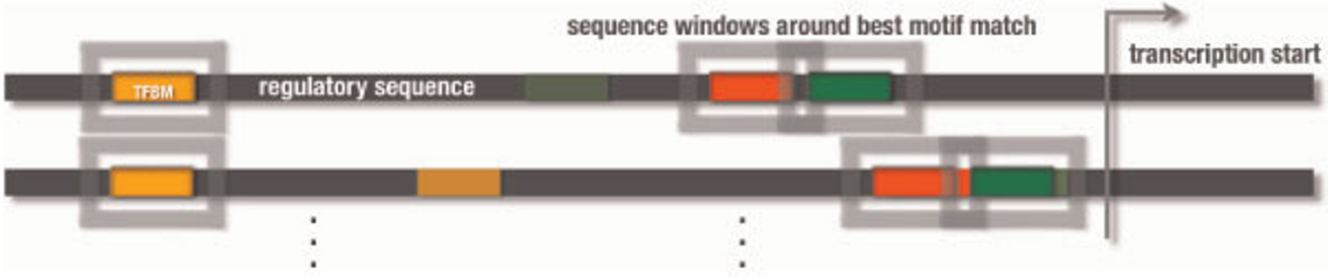

Figure 1
The idea behind the Regulatory Modules kernel: A motif finder is applied to regulatory sequences (long, gray bars) and identifies overrepresented motifs (colored segments). Around the best-matching motifs (boxed) in every sequence we excise 20 base pairs around the center. Conservation information and the pairwise distances of motifs to each other and to the end of the sequence are added to form the Regulatory Modules kernel, concatenating feature spaces.

with shifts (WDS) computes the similarity of fixed-length sequences. We extend this kernel with conservation information and information about motif co-occurrence to the Regulatory Modules kernel [2]. KIRMES is available on our Galaxy server http://galaxy.tuebingen.mpg.de. Using positional oligomer importance matrices [3], we are able to make the output of the kernel interpretable by displaying a sequence logo of the oligomers that contributed most to the correct classification.

Results

We compared our method to a state-of-the-art Gibbs sampler, PRIORITY [4], on its own dataset with the published settings with respect to successful classification. We achieve correct predictions on 74% of their sets *vs.* 63% for PRIORITY. We let KIRMES classify gene sets obtained from microarrays of *Arabidopsis thaliana*. Using conservation as weighting for the WDS kernel improves performance. These results illustrate the power of our approach in exploiting the relationship between motifs as well as conservation to improve the recognition of TF targets. Interpretable results and an easy-to-use web service make this a valuable tool for any researcher interested in gene regulation.

References

1. Gupta M and Liu J: **De novo cis-regulatory module elicitation for eukaryotic genomes.** *Proc Natl Acad Sci USA* 2005, **102(20):** 7079–7084.
2. Schultheiss SJ, Busch W, Lohmann JU, Kohlbacher O and Rätsch G: **KIRMES: Kernel-based identification of regulatory modules in euchromatic sequences.** *Bioinformatics* 2009, epub: 23 April 2009.
3. Sonnenburg S, Zien A, Philips P and Rätsch G: **POIMs: Positional Oligomer Importance Matrices – understanding support vector machine-based signal detectors.** *Bioinformatics* 2008, **24 (13):**i6–14.
4. Gordan R, Narlikar L and Hartemink A: **A fast, alignment-free, conservation-based method for transcription factor binding site discovery.** *Lecture Notes in Computer Science: RECOMB 2008* Springer, Heidelberg, Germany; **4955:**98–111.