*Bioinformatics*, 32(7), 2016, 1091–1093
doi: 10.1093/bioinformatics/btv705
Advance Access Publication Date: 1 December 2015
Applications Note

OXFORD

Genome analysis

# LedPred: an R/bioconductor package to predict regulatory sequences using support vector machines

**Denis Seyres[1,2,†], Elodie Darbo[3,†], Laurent Perrin[1,2,4], Carl Herrmann[5] and Aitor González[1,2,*]**

[1]INSERM, UMR1090 TAGC, Marseille, F-13288 France, [2]Aix-Marseille Université, UMR1090 TAGC, Marseille, F-13288 France, [3]Cancer Research UK, London Research Institute, London WC2A 3LY, UK, [4]CNRS, Marseille, France and [5]IPMB, Universität Heidelberg and Department of Theoretical Bioinformatics, DKFZ, Heidelberg 69120, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

Received on 20 July 2015; revised on 23 November 2015; accepted on 26 November 2015

## Abstract

**Summary:** Supervised classification based on support vector machines (SVMs) has successfully been used for the prediction of *cis*-regulatory modules (CRMs). However, no integrated tool using such heterogeneous data as position-specific scoring matrices, ChIP-seq data or conservation scores is currently available. Here, we present LedPred, a flexible SVM workflow that predicts new regulatory sequences based on the annotation of known CRMs, which are associated to a large variety of feature types. LedPred is provided as an R/Bioconductor package connected to an online server to avoid installation of non-R software. Due to the heterogeneous CRM feature integration, LedPred excels at the prediction of regulatory sequences in Drosophila and mouse datasets compared with similar SVM-based software.

**Availability and implementation:** LedPred is available on GitHub: https://github.com/aitgon/LedPred and Bioconductor: http://bioconductor.org/packages/release/bioc/html/LedPred.html under the MIT license.

**Contact:** aitor.gonzalez@univ-amu.fr

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Transcriptional programs are tightly controlled by *cis*-regulatory modules (CRMs) that control gene expression. CRMs are stretches of DNA that contain specific combinations of transcription factor binding sites (TFBS) and are often associated with histone modification or nucleosome-free regions (Spitz and Furlong, 2012). A well-established strategy to predict CRMs is to analyze combinations of CRM features described above in a set of experimentally validated CRMs. Support vector machine (SVM) algorithms are appropriate for this task since they can learn sequence features that distinguish CRMs from background sequences. Existing tools (Supplementary Table 1) score sequences based on *k*-mer discovery (KIRMES, Schultheiss *et al.*, 2009; KeBABS, Palme *et al.*, 2015; *k*-mer-SVM Fletez-Brant *et al.*, 2013; gkm-SVM, Ghandi *et al.*, 2014), position-specific scoring matrices (PSSMs) (CLARE, Taher *et al.*, 2012) or ChIP-seq data (Tagliazucchi, 2014). The DEEP tool theoretically integrates a large palette of features but is specific to mammals, does not return a model to score new sequences and is currently unavailable (Kleftogiannis *et al.*, 2014). Here, we present an R/Bioconductor package based on the libsvm library (Chih-Chung and Chih-Jen 2011) to model CRMs in many species and score new

sequences based on PSSMs, NGS data and pre-computed custom values such as conservation.

## 2 Workflow

The workflow is composed of five main steps: (1) numerical mapping of features to sequences, (2) tuning the SVM parameters, (3) feature selection, (4) model creation and evaluation and (5) scoring of new sequences (Fig. 1). Steps 1–4 are related to the model training, and the model is applied in Step 5. In Step 1, the user submits a BED file with positive training sequences. The negative training sequences can be generated from a set of background sequences (Bedtools; Quinlan and Hall, 2010) or submitted directly in a BED file. Users can use PSSMs (Transfac format), ChIP-seq peaks (BED) or signals (WIG or BED format) or BED files with regions and custom values (e.g. conservation scores or RNA-seq levels). Each sequence is scanned with the PSSMs using RSAT matrix-scan (Medina-Rivera *et al.*, 2015) and scored according to binding probabilities (see Bioconductor vignette). In the case of BED peaks, positive binary values are set for an overlap of >20% with the query sequence, whereas for the WIG signals, the average overlap is calculated (Java Genomics Tool Kit, Palpant, 2012). Step 1 is calculated by an online server to avoid installation of third-party programs. The numerical matrix returned by Step 1 is used in Steps 2–4 to create a model. In Step 2, the optimal SVM parameter values are computed from a grid of values. In Step 3, a ranking of the features is created using a recursive feature elimination algorithm while the optimal number of features is found by calculating Cohen's kappa coefficient for different number of features. In Step 4, an SVM model and a plot of its performance on the training sequences are created. Finally, in Step 5, the optimal features are mapped to the query sequences, which are then scored with the SVM model. The mouse example below, with around 170 positive and 1700 negative training sequences and 600 PSSMs, took around 1-h user time and 1.3-GB maximum resident set size using four CPUs (Supplementary Fig. 3). Execution time scales exponentially while memory scales linearly (Supplementary Fig. 3). Sequence and annotation data, and the workflows below, are available as Supplementary data and in the Bioconductor vignette.

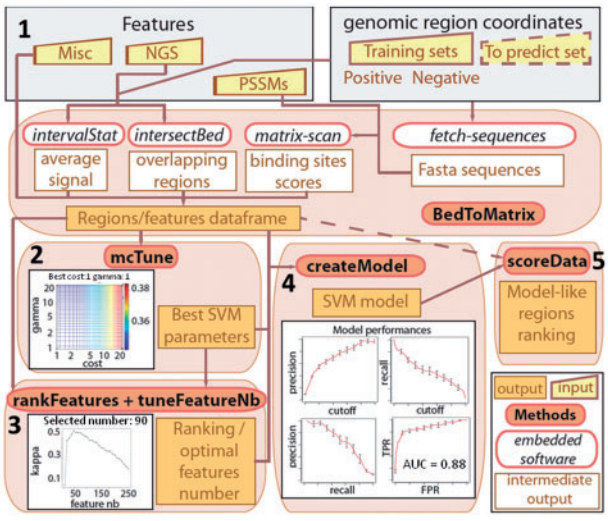


**Fig. 1.** Overview of the LedPred workflow. Genomic features are mapped to the training sequence sets. The resulting data frame is used to compute optimal SVM parameters, rank the features and find the best feature number. The optimal parameters and features are used to create a model that can be stored and used in the last part to score query sequences

## 3 Study cases

To illustrate our approach, we have analyzed *in vivo* validated *Drosophila melanogaster* enhancers from the CAD2 database studied in the context of mesodermal development (Bonn *et al.*, 2012). The positive training set is a subset (80%) of these enhancers (114/143) and 10 times more random sequences as negative training set (Supplementary file 1). PSSMs from different public databases (Supplementary data), mesoderm-specific ChIP-seq data for H3K4me1 (Bonn *et al.*, 2012) and ChIP-on-CHIP data on five mesoderm transcription factors (Tinman, Zinzen *et al.*, 2009; Dorsocross, Pangolin, phosphorylated MAD and Panier, Junion *et al.*, 2012) were used as descriptive features. Best performances were obtained when 90 out of 576 of the best ranked features were used. H3K4me1 ChIP-seq data and the Pnr and dTcf ChIP-on-CHIP data, known for their mesodermal roles, were classified among the best discriminative features. The model was used to scan the whole non-coding genome split into 1500 bp regions with 750 bp overlap (116 736 regions in total) (Fig. 1). Of the 5848 regions predicted with an FPR cut off of 5%, we retrieved 44% of the 29 remaining CAD2 mesodermal CRMs not included in the training set (Binomial test $P$-value $5.3 \times 10^{-09}$). Finally, we compared the performance of LedPred with other SVM-based CRM predictors using this Drosophila dataset and one from mouse in two different ways (Supplementary file 3). First, we compared the AUC returned by the software during the training (Supplementary Fig. 1). Second, we created a model with a training set composed of 80% sequences and drew ROC curves for the remaining 20% test sequences (Supplementary Fig. 2). In the Drosophila dataset, LedPred showed equivalent best performances as gkm-SVM, kebabs and *k*-mer-SVM (Supplementary Figs 1 and 2A). In the mouse, we used an enhanced dataset that has experimentally been characterized as high, medium and low level activity together with vertebrate PSSMs and NGS data related to this dataset (Supplementary File 2) (Vanhille *et al.*, 2015). We predicted enhancers using NGS data specific to this dataset, and PSSMs from the Jaspar (Mathelier *et al.*, 2014) and the Human Protein–DNA Interactome database (Xie *et al.*, 2010). In this dataset, LedPred achieved the best performances (AUC0.81 in Supplementary Fig. 1 and AUC0.75 in Supplementary Fig. 2) with a clear difference to the following classifiers (AUC0.66 in Supplementary Fig. 1 and AUC0.64 in Supplementary Fig. 2).

## 4 Concluding remarks

The LedPred package is a new, fast and accurate SVM workflow to predict CRMs based on PSSMs, NGS data and other features in different species. An online server facilitates the mapping of PSSMs, ChIP peaks and signals to the sequences, while the subsequent steps create an optimized SVM model and use it to score new sequences. The SVM specific steps after feature mapping can be applied to any numerical matrix, and therefore this package can be useful for other binary classification problems.

## Acknowledgements

We acknowledge Charles Chapple for critical reading of this manuscript, Jacques van Helden for insightful discussions, Jaime Castro for advice with RSAT matrix-clustering and for PSSMs and Salvatore Spicuglia for the mouse dataset. We acknowledge the French Institute of Bioinformatics (ANR-11-INBS-0013, www.france-bioinformatique.fr) for providing the bioinformatics

cloud infrastructure used for the software developments and the computing resources related to this study.

## Funding

ANR, partner of the ERASysBio+ supported under the UE ERA-NET Plus scheme in FP7 (C.H. and L.P.).

*Conflict of interest:* none declared.

## References

Bonn,S. *et al*. (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet*., **44**, 148–156.

Chih-Chung,C. and Chih-Jen,L. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol*., **2**, 1–27.

Fletez-Brant,C. *et al*. (2013) kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res*., **41**, 544–556.

Ghandi,M.D. *et al*. (2014) Enhanced regulatory sequence prediction using gapped K-Mer features. *PLoS Comput. Biol*.,**10**, e1003711.

Junion,G. *et al*. (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, **148**, 473–486.

Kleftogiannis,D. *et al*. (2014) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res*., **43**, e6.

Mathelier,A. *et al*. (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*., **42**(Database issue), D142–D147.

Medina-Rivera,A. *et al*. (2015) RSAT 2015: regulatory sequence analysis tools. *Nucleic Acids Res*., **43**, W50–W56.

Palme,J. *et al*. (2015) KeBABS: an R package for kernel-based analysis of biological sequences. *Bioinformatics*, **31**, 2574–2576.

Palpant,T. (2012) Java genomics tool kit, http://palpant.us/java-genomics-toolkit/.

Quinlan,A.R. and Hall,I.M.(2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Schultheiss,S.J. *et al*. (2009) KIRMES: kernel-based identification of regulatory modules in euchromatic sequences. *Bioinformatics*, **25**, 2126–2133.

Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet*., **13**, 613–626.

Tagliazucchi,G.M. (2014) SVM2CRM: support vector machine for *cis*-regulatory elements detections. *R package version 1.0.0*. https://bioconductor.org/packages/release/bioc/html/SVM2CRM.html.

Taher,L. *et al*. (2012) Clare: cracking the LAnguage of regulatory elements. *Bioinformatics*, **28**, 581–583.

Vanhille,L. *et al*. (2015) High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun*., **6**, 6905.

Xie,Z. *et al*. (2010) hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics*, **26**, 287–289.

Zinzen,R.P. *et al*. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.