Bioinformatics, 31(15), 2015, 2574–2576
doi: 10.1093/bioinformatics/btv176
Advance Access Publication Date: 25 March 2015
Applications Note

OXFORD

Sequence analysis

KeBABS: an R package for kernel-based analysis of biological sequences

Johannes Palme, Sepp Hochreiter and Ulrich Bodenhofer*

Institute of Bioinformatics, Johannes Kepler University, Linz, Austria

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on February 2, 2015; revised on March 4, 2015; accepted on March 20, 2015

Abstract

Summary: KeBABS provides a powerful, flexible and easy to use framework for **ke**rnel-**b**ased **a**nalysis of **b**iological **s**equences in R. It includes efficient implementations of the most important sequence kernels, also including variants that allow for taking sequence annotations and positional information into account. KeBABS seamlessly integrates three common support vector machine (SVM) implementations with a unified interface. It allows for hyperparameter selection by cross validation, nested cross validation and also features grouped cross validation. The biological interpretation of SVM models is supported by (1) the computation of weights of sequence patterns and (2) prediction profiles that highlight the contributions of individual sequence positions or sections.
Availability and implementation: The R package kebabs is available via the Bioconductor project: http://bioconductor.org/packages/release/bioc/html/kebabs.html. Further information and the R code of the example in this paper are available at http://www.bioinf.jku.at/software/kebabs/.
Contact: kebabs@bioinf.jku.at or bodenhofer@bioinf.jku.at

1 Introduction

The analysis of biological sequences is a fundamental task in bioinformatics. In the last two decades, kernel methods have been established as an important class of sequence analysis methods. For the classification of sequences, in particular, support vector machines (SVMs) have emerged as a sort of best practice. To apply SVMs for sequence analysis, it is necessary to either use a vectorial representation of the sequence data or to use *sequence kernels*, that is, positive semi-definite similarity measures for sequences. The use of sequence kernels, however, is not limited to sequence classification. For example, they can also be used for regression tasks and similarity-based clustering.

On the scientific computing platform R which is widely used in bioinformatics, only the kernlab package (Karatzoglou *et al.*, 2004) provides a limited selection of sequence kernels. This article presents KeBABS, an R/Bioconductor package for kernel-based sequence analysis that is primarily focused on biological applications. Compared with the SHOGUN Toolbox (Sonnenburg *et al.*, 2010), KeBABS provides a wider selection of up-to-date sequence kernels and facilitates seamless interplay with R and Bioconductor's sequence data packages.

2 Package description

Sequence kernels are *the* core functionality of KeBABS. Four commonly used kernels are provided: spectrum kernel (Leslie *et al.*, 2002), mismatch kernel (Leslie *et al.*, 2003), gappy pair kernel (a subset of spatial sample kernels according to Kuksa *et al.*, 2008) and motif kernel (Ben-Hur and Brutlag, 2003). These kernels consider occurrences of patterns regardless of their positions.

KeBABS also supports *position-dependent* variants for all its kernels except for the mismatch kernel: (i) position-specific variants only count occurrences of patterns if they appear at exactly the same position. (ii) Distance-weighted variants count occurrences of patterns if they appear at similar positions (Bodenhofer *et al.*, 2009), where the positional similarity is determined by a distance weighting function. The package provides three built-in distance weighting functions. Gaussian distance weights together with the spectrum kernel closely corresponds to the oligo kernel (Meinicke *et al.*, 2004). The distance weighting used by the shifted weighted degree kernel (Rätsch *et al.*, 2005) is available too and users can also supply custom distance weights. Therefore, the package also includes the weighted degree kernel and the shifted weighted degree kernel (Rätsch *et al.*, 2005), but without position weights.

As a unique new feature, KeBABS provides *annotation-specific* variants for all kernels except the mismatch kernel: each sequence can be accompanied by an aligned annotation sequence from a user-defined alphabet—to enable the kernel to distinguish patterns if they appear in different annotation contexts. For example, gene sequences can be annotated with an 'e' for all exon positions and 'i" for all intron positions. Then the kernel automatically distinguishes between exonic and intronic patterns. As another example, a coiled-coil sequence can be annotated with the heptade register. If used in conjunction with the gappy pair kernel, this corresponds to the coiled-coil kernel (Mahrenholz *et al.*, 2011).

KeBABS can compute kernel matrices for all kernels both in dense and sparse formats. For position-independent kernels, KeBABS also facilitates *explicit feature representations* that can be stored to dense or sparse matrices. These sparse feature representations in conjunction with LiblineaR allow for analyzing up to hundreds of thousands of sequences.

SVM framework: KeBABS provides a unified interface to three SVM implementations: kernlab (Karatzoglou *et al.*, 2004), LIBSVM (Chang and Lin, 2011; via the e1071 package) and LiblineaR (Fan *et al.*, 2008). The SVM framework in KeBABS can be used for classification (binary and multi-class) and regression tasks. Cross validation and hyperparameter selection are supported with all interfaced SVMs. For hyperparameter selection, accuracy, balanced accuracy and the Matthews correlation coefficient can be selected as performance objectives (the area under the ROC curve is also available for Version 1.2.0 or newer).

Grouped cross validation: Apart from the standard k-fold cross validation, KeBABS also supports *grouped cross validation*, i.e. cross validation that keeps pre-specified groups together in the same folds. As an example, to group sequences by their sequence identity is a common approach in protein structure prediction to assess whether a predictor is able to make use of sequence features beyond making a simple sequence identity-based prediction.

3 Example: Epitope-to-MHC binding

We analyzed the binding of protein fragments to the human MHC (major histocompatibility complex) class I molecule HLA-A*0201 based on epitope data from Roomp *et al.* (2010). The binding of protein fragments to the MHC molecule is an important step of the immune system to recognize proteins of questionable origin and trigger the immune system's reaction. Figure 1a illustrates the binding groove of the MHC molecule with the bound peptide fragment LLFGYPVYV (Madden *et al.*, 1993).

For our analysis, we used the strong binder/clear non-binder subset with 549 strong binders and 503 clear non-binders. The analysis was performed with the C-SVC from package kernlab. Upon hyperparameter selection on 40% of the data, the position-specific spectrum kernel with $k = 1$ turned out to be the best choice, which indicates that individual positions are highly relevant for the binding behavior. This setting resulted in a cross validation accuracy of 94.3% (average of 10 runs, with $\sigma = 0.453\%$ and an average area under the ROC curve of 0.983).

Figure 1b shows the feature weights computed from the SVM model and the relevance of each amino acid at each position. The prediction profiles of two sequences in Figure 1c show the contribution of each sequence position to the prediction. The upper sequence (peptide from Fig. 1a) is a strong binder with high positive contributions for Leu at pos. 2 and Val at pos. 9. The lower sequence is a clear non-binder with negative contributions from all positions except the last one. The heatmap of the prediction profiles for all sequences in Figure 1d reveals that the second and the last position have high relevance for all binders. Non-binders in the lower half of the image show higher negative contributions from the first and the third position. Positions 4, 5 and 8 generally have little importance.

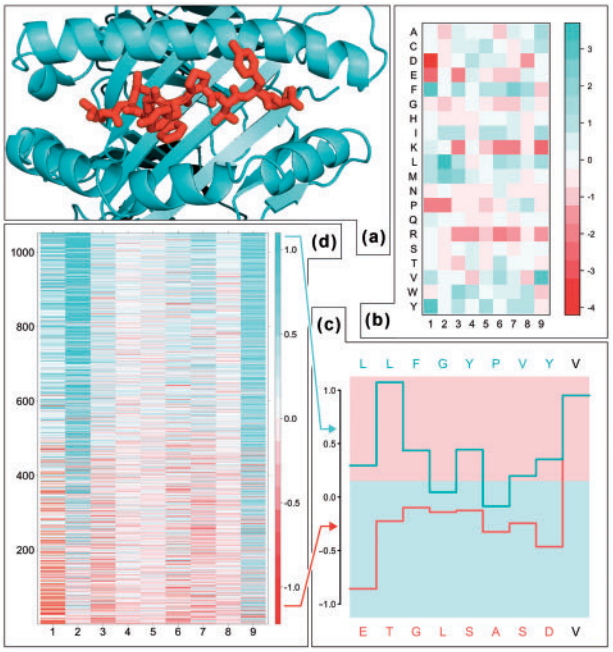

Fig. 1. (**a**) Structure of the binding groove of human class I MHC HLA-A*0201 with bound peptide LLFGYPVYV in red (PDB ID: 1hhk); (**b**) position-specific feature weights; (**c**) prediction profiles for the peptide from (a) and a non-binding peptide; (**d**) prediction profiles of all samples sorted by increasing decision value. Positive contributions to the decision value are shown in blue and negative ones in red. All results are based on a normalized position-specific spectrum kernel with $k = 1$

The importance of the anchor positions 2 and 9 for binding is well known (Madden *et al.*, 1993). The irrelevance of positions 4, 5 and 8 corresponds to the Janus face characteristics of the peptide, with some of the positions facing toward the MHC molecule and some toward a possibly binding T-cell receptor. KeBABS, via the computation of feature weights and prediction profiles, allows for mining such biological knowledge from classification models based on SVMs and sequence kernels.

Conflict of Interest: none declared.

References

Ben-Hur,A. and Brutlag,D.L. (2003) Remote homology detection: a motif based approach. *Bioinformatics*, **19**, 26–33.

Bodenhofer,U. *et al.* (2009) Modeling position specificity in sequence kernels by fuzzy equivalence relations. In: Carvalho,J.P. *et al.* (eds), *Proceedings of the Joint 13th IFSA World Congress and 6th EUSFLAT Conference*, Lisbon, Portugal, pp. 1376–1381.

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1–27:27.

Fan,R.-E. *et al.* (2008) LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.

Karatzoglou,A. *et al.* (2004) Kernlab—an S4 package for kernel methods in R. *J. Stat. Softw.*, **11**, 1–20.

Kuksa,P. *et al.* (2008) A fast, large-scale learning method for protein sequence classification. In: *8th International Workshop on Data Mining in Bioinformatics*, Las Vegas, NV, USA, Chapman & Hall/CRC Press, pp. 29–37.

Leslie,C.S. *et al.* (2002) The spectrum kernel: a string kernel for SVM protein classification. In: Altman,R.B. *et al.* (eds). *Pacific Symposium on Biocomputing*, World Scientific, Lihue, HI, USA, pp. 564–575.

Leslie,C.S. *et al.* (2003) Mismatch string kernels for discriminative protein classification. *Bioinformatics,* **1**, 1–10.

Madden,D.R. *et al.* (1993) The antigenic identity of peptide-MHC complexes: a comparison of five viral peptides presented by HLA-A2. *Cell,* **75**, 693–708.

Mahrenholz,C.C. *et al.* (2011) Complex networks govern coiled-coil oligomerizations—predicting and profiling by means of a machine learning approach. *Mol. Cell Proteomics,* **10**, M110.004994.

Meinicke,P. *et al.* (2004) Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, **5**, 169.

Rätsch,G. *et al.* (2005) RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics,* **21**(Suppl. 1), i369–i377.

Roomp,K. *et al.* (2010) Predicting MHC class I epitopes in large datasets. *BMC Bioinformatics,* **11**, 90.

Sonnenburg,S. *et al.* (2010) The SHOGUN machine learning toolbox. *J. Mach. Learn. Res.,* **11**, 1799–1802.