BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1769–1770
doi:10.1093/bioinformatics/btu096

Structural bioinformatics Advance Access publication February 14, 2014

SKINK: a web server for string kernel based kink prediction in α-helices

Tim Seifert[1], Andreas Lund[2], Benny Kneissl[3,*], Sabine C. Mueller[4], Christofer S. Tautermann[5] and Andreas Hildebrandt[1]

[1]Department of Software Engineering and Bioinformatics, Johannes Gutenberg University of Mainz, 55128 Mainz, [2]Institute for Computer Science, Johann Wolfgang Goethe-University Frankfurt am Main, 60054 Frankfurt am Main, [3]Data Science, Pharma Research and Early Development Informatics (pREDi), Roche Diagnostics GmbH, 82377 Penzberg, [4]Department of Human Genetics, Saarland University Faculty of Medicine, 66421 Homburg and [5]Lead Identification and Optimization Support, Boehringer-Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The reasons for distortions from optimal α-helical geometry are widely unknown, but their influences on structural changes of proteins are significant. Hence, their prediction is a crucial problem in structural bioinformatics. Here, we present a new web server, called SKINK, for string kernel based kink prediction. Extending our previous study, we also annotate the most probable kink position in a given α-helix sequence.

Availability and implementation: The SKINK web server is freely accessible at http://biows-inf.zdv.uni-mainz.de/skink. Moreover, SKINK is a module of the BALL software, also freely available at www.ballview.org.

Contact: benny.kneissl@roche.com

Received on November 11, 2013; revised on January 31, 2014; accepted on February 7, 2014

1 INTRODUCTION

In our previous study (Kneissl *et al.*, 2011), we presented a new method for predicting kinks from amino acid sequences of α-helices implemented in the Biochemical Algorithm Library (BALL; Hildebrandt *et al.*, 2010). Using string kernel based support vector machines (SVM), we were able to predict kinks in ~80% of all cases correctly, exceeding recently reported accuracies of alternative approaches (Hall *et al.*, 2009; Langelaan *et al.*, 2010; Meruelo *et al.*, 2011; Rigoutsos *et al.*, 2003; Yohannan *et al.*, 2004). To make our approach publicly available by providing a user-friendly interface, we developed the web server SKINK. Moreover, we annotate the most probable kink position and calculate the corresponding probability value described shortly in the Section 2. In consequence, the region around the identified residue can be treated as flexible, e.g. in Molecular Dynamics (MD) simulations, while the remainder of the helix can be restrained.

2 METHODS

The prediction (classification) of an input sequence as kinked or non-kinked is described in our previous study (Kneissl *et al.*, 2011) in detail. In addition, the most probable position of a putative kink is computed. Therefore, we divided the original sequence into overlapping k-mers of all odd sizes between 7 and 15 amino acids. Let the kink be in the center position of the k-mer; this corresponds to one to two neighbored helical turns, which are taken into account. These fragments are then predicted with the corresponding support vector machine as kinked or non-kinked. Using a trapezoidal-shaped function, the probability p_s that the kink is located at position s in the currently considered subsequence is calculated by

$$p_s = \sum_{i=3}^{7} \sum_{j=-i}^{i} \frac{7-|j|}{7} \cdot \mathrm{k}_{s+j-i,\, s+j+i}, \qquad (1)$$

where k $_{a,b}$ is 1 if subsequence $[a\colon b]$ is predicted as kinked and otherwise 0. Hence, a position is more likely to be the kink if it appears near the center position of the k-mers that are labeled as kinked. The calculated probabilities are then normalized by the maximum possible value at each position such that each amino acid in the input sequence is assigned with a value between 0 and 1. Because we are not interested in kinks located at the end of a helix, the first and last two turns (seven positions) are ignored. Finally, the maximum values are chosen to determine the kink positions. If there are multiple maxima, two are merged if they are closer than four positions, else they are annotated as two kinks.

Applying this kink annotation approach on our previously used training dataset (downloadable from the Web site), we are able to identify 85.6% of 339 correctly predicted kinks (out of 366) within one helical turn.

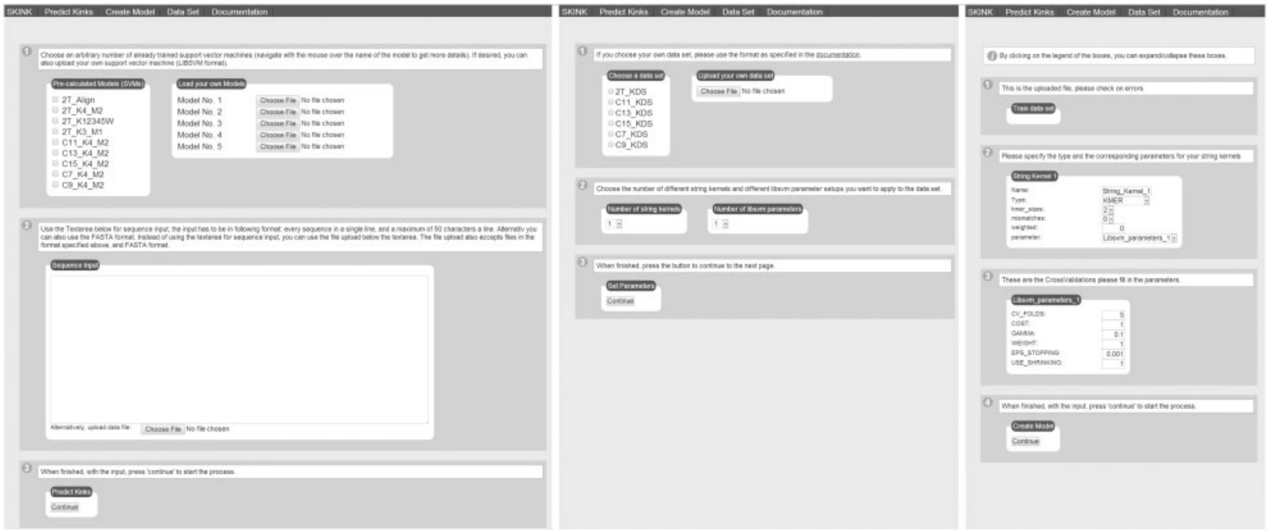
3 SKINK

The SKINK web server is kept deliberately simple but well-structured to maintain clarity and is geared to the abilities of the users (see Fig. 1). First, there is a page *Documentation*, where some information about the dataset format as well as the precomputed support vector machine models is given. A detailed description and some analysis results of the datasets can be found in the page *Data Sets*. Moreover, all datasets can be downloaded here.

The two main functionalities are provided in the pages *Predict Kinks* and *Create Model*.

Users can query the server via the input text field or upload a file containing all sequences, which shall be classified as kinked or non-kinked. We allow two different input formats: The sequence can be in plain text (one sequence per line) or in the FASTA format. In addition, they can upload a file containing

*To whom correspondence should be addressed.



1 Choose an arbitrary number of already trained support vector machines (navigate with the mouse over the name of the model to get more details). If desired, you can also upload your own support vector machine (LIBSVM format).

□ 2T_Align
□ 2T_K4_M2
□ 2T_K12345W
□ 2T_K3_M1
□ C11_K4_M2
□ C13_K4_M2
□ C15_K4_M2
□ C7_K4_M2
□ C9_K4_M2

Model No. 1 Choose File No file chosen
Model No. 2 Choose File No file chosen
Model No. 3 Choose File No file chosen
Model No. 4 Choose File No file chosen
Model No. 5 Choose File No file chosen

2 Use the Textarea below for sequence input, the input has to be in following format: every sequence in a single line, and a maximum of 50 characters a line. Alternativ you can also use the FASTA format. Instead of using the textarea for sequence input, you can use the file upload below the textarea. The file upload also accepts files in the format specified above, and FASTA format.

Alternatively, upload data file: Choose File No file chosen

3 When finished, with the input, press 'continue' to start the process.

1 If you choose your own data set, please use the format as specified in the documentation.

○ 2T_KDS
○ C11_KDS
○ C13_KDS
○ C15_KDS
○ C7_KDS
○ C9_KDS

Choose File No file chosen

2 Choose the number of different string kernels and different libsvm parameter setups you want to apply to the data set.

Number of string kernels: 1
Number of libsvm parameters: 1

3 When finished, press the button to continue to the next page.

By clicking on the legend of the boxes, you can expand/collapse these boxes.

1 This is the uploaded file, please check on errors

2 Please specify the type and the corresponding parameters for your string kernels

Name: String_Kernel_1
Type: KMER
kmer_sizes: 2
mismatches: 0
weighted: 0
parameter: Libsvm_parameters_1

3 These are the CrossValidations please fill in the parameters.

CV_FOLDS: 5
COST: 1
GAMMA: 0.1
WEIGHT: 1
EPS_STOPPING: 0.001
USE_SHRINKING: 1

4 When finished, with the input, press 'continue' to start the process.

Fig. 1. Snapshots of the web server input pages: input page for kink prediction (left), page for creating own models (center) and page for setting the corresponding LIBSVM parameters (right)

all sequences in one of these two formats. Before continuing, the user has to choose at least one of the precomputed models or upload his/her own one in the LIBSVM (Chang and Lin, 2011) format. All chosen support vector machines are then applied on each input sequence. The classification whether the sequence is kinked or not is quite fast (about 15 s), whereas the annotation of the kink position can be up to three times longer because of the computation of all feature vectors of the different subsequences. The results are presented in a new page, where all information (input sequences, chosen models, logfile) of the request is collected. For each model a subdirectory is created where the corresponding result file is stored. In this file, each input sequence is labeled as non-kinked (0) or kinked (1), whereas in the latter case the kink position and the probability value are given, too.

To create a new model, the user has to choose a training dataset, which can either be one of the provided ones on the Web site or an own dataset. In both cases, the BALL::ParameterFile format is required. Thereafter, he/she has to decide how many different string kernels and how many different LIBSVM parameter setups he/she wants to apply, which can be combined for each training run on the following setup page. Hence, for one dataset all generated support vector machines are located in a separate subdirectory. Clicking 'continue' shows the setup page. Here, the user can re-check the uploaded dataset and choose different parameters for the string kernel functions. Each setup can be named in a unique user-specified name. Finally, the LIBSVM parameters, e.g. for the cost function or the number of cross validation folds, can be set. Clicking 'continue' shows the result page, where for each fold a separate SVM is created. Moreover, it shows a file with all sequences and the corresponding predicted labels as well as a file containing a short analysis, e.g. the confusion matrix and prediction accuracies.

Funding: Boehringer Ingelheim Pharma GmbH & Co. KG.

Conflict of Interest: none declared.

REFERENCES

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

Hall,S.E. *et al.* (2009) Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. *J. Mol. Graph. Model.*, **27**, 944–950.

Hildebrandt,A. *et al.* (2010) BALL–Biochemical Algorithms Library 1.3. *BMC Bioinformatics*, **11**, 531.

Kneissl,B. *et al.* (2011) String kernels and high-quality data set for improved prediction of kinked helices in α-helical membrane proteins. *J. Chem. Inf. Model.*, **51**, 3017–3025.

Langelaan,D.N. *et al.* (2010) Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. *J. Chem. Inf. Model.*, **50**, 2213–2220.

Meruelo,A.D. *et al.* (2011) TMKink: a method to predict transmembrane helix kinks. *Protein Sci.*, **20**, 1256–1264.

Rigoutsos,I. *et al.* (2003) Structural details (kinks and non-alpha conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors. *Nucleic Acids Res.*, **31**, 4625–4631.

Yohannan,S. *et al.* (2004) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl Acad. Sci. USA*, **101**, 959–963.