1670T

KeBABS: an R/Bioconductor package for kernel-based analysis of biological sequences. *J. Palme, S. Hochreiter, U. Bodenhofer.* Institute of Bioinformatics, Johannes Kepler University, Linz, Austria.

The computational analysis of biological sequences is a fundamental task. On the one hand, sequence analysis methods have supplied highly valuable insights into how patterns/motifs in amino acid sequences govern protein structure. On the other hand, a large proportion of our current knowledge about how DNA sequence patterns control transcription factor binding, nucleosome positioning and remodeling, alternative splicing, etc. , is the result of computational sequence analysis. In genetics, discriminative sequence analysis is becoming increasingly important to predict potential effects of single-nucleotide variations in the context of surrounding sequences.

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 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.

This contribution is devoted to introducing KeBABS, a powerful, flexible, and easy-to-use framework for kernel-based analysis of biological sequences based on the widely used scientific computing platform R. KeBABS is publicly and freely available via the Bioconductor project (for more information, see http://www. bioinf. jku. at/software/kebabs). 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. The features of the package will be described in detail along with illustrative biological examples.

1671F

HPMV: Human Protein Mutation Viewer. *W. A. Sherman[1], D. B. Kuchibhatla[1], V. Limviphuvadh[1], S. Maurer-Stroh[1,2], F. Eisenhaber[1,3,4], B. Eisenhaber[1].* 1) A*STAR Bioinformatics Institute (BII), Singapore, Singapore; 2) School of Biological Sciences (SBS), Nanyang Technological University (NTU), Singapore, Singapore; 3) Department of Biological Sciences (DBS), National University of Singapore (NUS), Singapore, Singapore; 4) School of Computer Engineering (SCE), Nanyang Technological University (NTU), Singapore, Singapore.

Next-generation sequencing advances are rapidly increasing thefraction of genetic disorders for which causative variants can beidentified. Our Human Protein Mutation Viewer (HPMV) can helpidentify causative variants in genetic disorders that are caused by asingle small change in a protein sequence (e. g. a non-synonymous pointmutation). HPMV fills a niche between initial variant filtering anddetailed analysis of individual mutations. It allows a researcher toquickly assess dozens of potential causative variants by presenting aninteractive cartoon that shows where each mutation of interest fallsalong the corresponding protein sequence - in relation to proteinfeatures that can help interpret the mutation. These protein featuresinclude post-translational modification sites, targeting signals,transmembrane helices, charge clusters, known conserved (Pfam)domains, and relevant 3D (PDB) structures, among others. As input, HPMV accepts protein mutations - as UniProt accessions withmutations (e. g. HGVS nomenclature), genome coordinates, or FASTAsequences. The main output of HPMV is its interactivecartoon. Clicking a sequence feature in the cartoon expands a treeview of additional information including multiple sequence alignmentsof conserved domains and a simple 3D viewer mapping the mutation torelevant PDB structures. A multiple sequence alignment of similarsequences from other organisms is provided directly below the cartoonand the cartoon itself includes a band showing the conservation ateach sequence position. In cases where a mutation is likely to havea straightforward interpretation (e. g. a point mutation in a wellunderstood targeting signal), this interpretation is suggested. The interactive cartoon is implemented as a web page using standardHTML, CSS, and JavaScript. It is also available as a standalone viewerimplemented in Java that can be downloaded in jar format, withembedded data, to be saved and viewed later with only a standard javaruntime environment. The HPMV website is: http://hpmv. bii. a-star. edu. sg/.

1672W

The ENCODE Analysis Pipelines: Tools for Repeatable, Standards-Based Analysis and Quality Control of Chromatin, Expression, and Methylation Experiments. *J. S. Strattan, T. Dreszer, B. C. Hitz, E. L. Hong, J. M. Cherry, ENCODE Data Analysis Center, ENCODE Data Coordinating Center.* Stanford University School of Medicine, Department of Genetics, Stanford, CA.

From Ammon's horn to zone of skin, members of the ENCODE Consortium have measured RNA quantity, RNA-protein interactions, DNA-protein interactions, DNA methylation, replication timing, chromatin structure, and histone modifications in over 4,000 experiments on more than 400 cell or tissue types. The ENCODE Data Analysis Center (DAC) have specified uniform processing pipelines for four ENCODE datatypes: ChIP-seq, RNA-seq, DNase-seq, and whole-genome bisulfite sequencing. The ENCODE Data Coordinating Center (DCC) have implemented these pipelines and deployed them to a cloud-based platform. Importantly, the pipelines can be run on the cloud via a web interface with no technical prerequisites other than input data. In this way researchers can perform the same analyses as ENCODE on their own data and repeat ENCODE analyses on ENCODE data. For ENCODE, the results of these analyses and metadata describing them are distributed through the ENCODE Portal, and illustrate general methods of accessing and interpreting ENCODE data. The ENCODE Portal is https://www. encodeproject. org/. The DCC codebase is freely available at https://github. com/ENCODE-DCC/.