*Bioinformatics*, 34(12), 2018, 2132–2133
doi: 10.1093/bioinformatics/bty056
Advance Access Publication Date: 7 February 2018
Applications Note

OXFORD

Sequence analysis

# Kpax3: Bayesian bi-clustering of large sequence datasets

**Alberto Pessia[1,*] and Jukka Corander[1,2,3]**

[1]Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland, [2]Department of Biostatistics, University of Oslo, 0317 Oslo, Norway and [3]Pathogen Genomics, Wellcome Trust Sanger Institute, CB10 1SA Hinxton, UK

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on October 1, 2017; revised on January 6, 2018; editorial decision on February 1, 2018; accepted on February 6, 2018

## Abstract

**Motivation:** Estimation of the hidden population structure is an important step in many genetic studies. Often the aim is also to identify which sequence locations are the most discriminative between groups of samples for a given data partition. Automated discovery of interesting patterns that are present in the data can help to generate new biological hypotheses.

**Results:** We introduce Kpax3, a Bayesian method for bi-clustering multiple sequence alignments. Influence of individual sites will be determined in a supervised manner by using informative prior distributions for the model parameters. Our inference method uses an implementation of both split-merge and Gibbs sampler type MCMC algorithms to traverse the joint posterior of partitions of samples and variables. We use a large Rotavirus sequence dataset to demonstrate the ability of Kpax3 to generate biologically important hypotheses about differential selective pressures across a virus protein.

**Availability and implementation:** Kpax3 is implemented as a Julia package and released under the MIT license. Source code and documentation are available at: https://github.com/albertopessia/Kpax3.jl.

**Contact:** alberto.pessia@helsinki.fi

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

## 1 Introduction

Identification of crucial positions in DNA and protein sequences is an important problem in biology and medicine. Types of target locations vary considerably depending on the scope of the analysis, from single sites to long contiguous segments. As an example, this kind of information is valuable for vaccine development (Kilbourne *et al.*, 2002) or increasing our understanding of antibiotic resistance (Chewapreecha *et al.*, 2014). However, interesting patterns hidden in the data are often masked by the unknown population structure, which typically must also be inferred from the same data. When approached more formally, this problem can be formulated as a two-way cluster analysis where both samples and variables are partitioned into disjoint subsets. Such an approach is generally referred to as bi-clustering, or co-clustering (Mirkin, 1996). Here, we introduce the next-generation version of the bi-clustering software Kpax (Marttinen *et al.*, 2006; Pessia *et al.*, 2015). Our new build significantly improves from the previous two by employing a new statistical formalization for an increased amount of detailed inference information. To highlight the possibilities offered by Kpax3 we demonstrate it using both synthetic and real datasets.

## 2 Implementation

Kpax3 employs a Bayesian bi-clustering model (Supplementary Material S1) that extends and improves the one originally introduced by Pessia *et al.* (2015). Datasets can be loaded either as *fasta* files or as more general *csv* files. Kpax3 output consists of several text files containing the clustering of both the rows (sequences) and columns (sites) of the input dataset. The software is entirely written in the Julia programming language (Bezanson *et al.*, 2017) and is freely

available through the Julia's built-in package manager. The package includes easy step-by-step tutorials to help the user to get started.

## 3 Results

Details of a simulation study using synthetic datasets can be found in Supplementary Material S2. Here, we will illustrate an application of Kpax3 to a real dataset of Rotavirus protein sequences. Data was retrieved from NCBI's Virus Variation Resource database (Brister *et al.*, 2013) (accessed 2016-08-15) and protein accession numbers are provided in Supplementary Material S3. The dataset consisted of 841 protein sequences whose length, after alignment, was 783 amino acids of which 683 were polymorphic. When converted into binary variables, the final dataset dimension was 841 rows and 2612 columns. Rotavirus is a double-stranded RNA virus belonging to the family *Reoviridae* and can cause gastroenteritis, affecting mostly children and infants. Although it is a common virus and nearly every child get infected by the age of five, the implied illness can develop into serious condition and is still a major cause of death in developing countries (Bernstein, 2009). We chose to focus only on species A of Rotavirus because it is the most common virus of this type in humans and we analyzed the VP4 structural protein because it is responsible of binding the virus to the target cell and under constant selection pressure from the immune system. VP4 is located on the surface of the virion, has a spike shape, and is also known to cluster into different serotypes (Hoshino and Kapikian, 2000). We expect to recover this structure through our unsupervised clustering. In total, $10^6$ MCMC samples were collected with default prior hyperparameters. Figure 1 shows the plots of the three posterior distributions estimated by Kpax3. The 95% credible interval for the total number of clusters is between 11 and 17 groups with a mode at 13 (Fig. 1A). Clusters of protein sequences are evident when examining the posterior similarity matrix (Fig. 1B), where each element is the posterior probability of the corresponding two sample units belonging to the same cluster. As expected, recovered groups mirror the known protein serotypes with group P[8] further split into four sub-clusters. The major contribution of our method is the ability to highlight regions of the gene where selective pressure might be happening. From the posterior distribution of column classification (Fig. 1C) we can observe regions at the extremities of the protein that are responsible for the discrimination of the clusters.

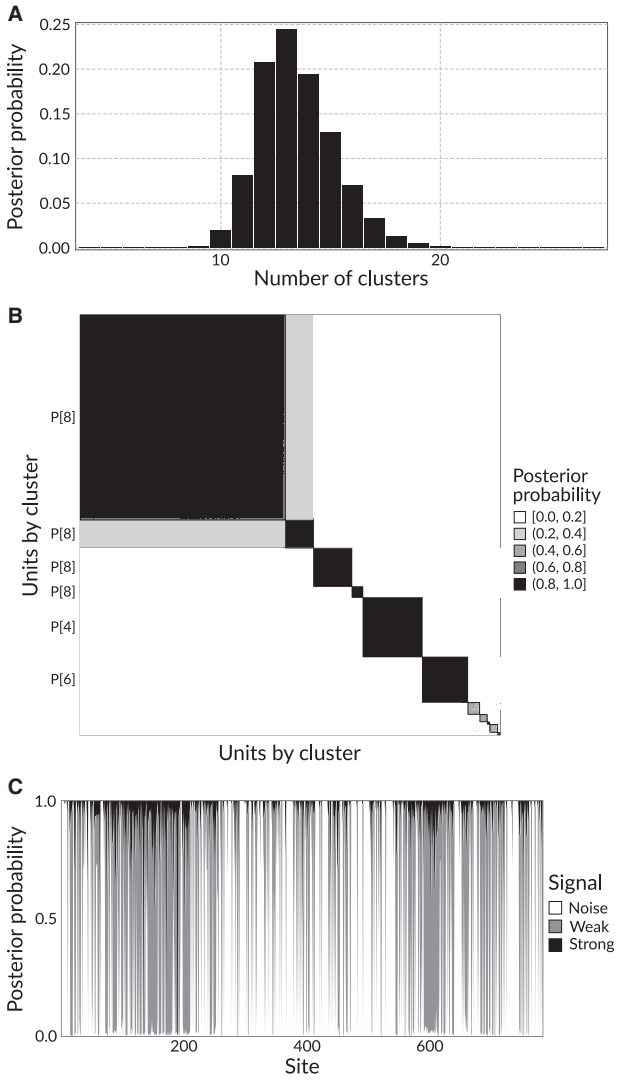


**Fig. 1.** Plots of the posterior distributions generated by Kpax3, representing uncertainty about the corresponding quantities. (**A**) Total number of Rotavirus VP4 protein groups. (**B**) Matrix of pairwise sequence proximities. Each pixel represents a pair of strains. Colours indicate the posterior probability of them belonging to the same group. Clusters of known serotypes, as recovered by Kpax3, have been annotated on the left side of the figure. (**C**) Classification of each protein site into the three levels of clustering discrimination power. Dark colours indicate that amino acid residues were found to be different, among clusters, at that particular position

## 4 Conclusion

We extended the model of Pessia *et al.* (2015) to a more general framework and derived extensive analytical formulas to enable simulation of samples from the posterior distribution using a hybrid MCMC approach. Kpax3 is particularly useful for clustering protein sequences, providing valuable information about differences in selection pressure and mutation rates across different genes. In a more general setting, any kind of matrix-variate categorical data can be bi-clustered by Kpax3.

## Acknowledgements

We acknowledge financial support from the COIN Centre of Excellence funded by the Academy of Finland. We would like to thank Elli-Noora Virkkunen for her valuable comments on the final draft of the manuscript.

*Conflict of Interest*: none declared.

## References

Bernstein,D.I. (2009) Rotavirus overview. *Pediatric Infect. Dis. J.*, **28**, S50–S53.

Bezanson,J. *et al.* (2017) Julia: a fresh approach to numerical computing. *SIAM Rev.*, **59**, 65–98.

Brister,J.R. *et al.* (2013) Virus variation resource – recent updates and future directions. *Nucleic Acids Res.*, **42**, D660–D665.

Chewapreecha,C. *et al.* (2014) Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.*, **46**, 305–309.

Hoshino,Y. and Kapikian,A.Z. (2000) Rotavirus serotypes: classification and importance in epidemiology, immunity, and vaccine development. *J. Health Popul. Nutrit.*, **18**, 5–14.

Kilbourne,E.D. *et al.* (2002) The total influenza vaccine failure of 1947 revisited: major intrasubtypic antigenic change can explain failure of vaccine in a post-World War II epidemic. *Proc. Natl. Acad. Sci. USA*, **99**, 10748–10752.

Marttinen,P. *et al.* (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, **22**, 2466–2474.

Mirkin,B. (1996) *Mathematical Classification and Clustering, Volume 11 of Nonconvex Optimization and Its Applications*. Springer, New York.

Pessia,A. *et al.* (2015) K-Pax2: Bayesian identification of cluster-defining amino acid positions in large sequence datasets. *Microb. Genomics*, **1**, 1–11.