Biological Sequence Modeling with Convolutional Kernel Networks

Dexiong Chen, Laurent Jacob, Julien Mairal

Biological Sequence Modeling with Convolutional Kernel Networks

Dexiong Chen[1], Laurent Jacob[2], and Julien Mairal[1]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
{fisrtname.lastname}@inria.fr
[2] Univ. Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, Lyon, France
laurent.jacob@univ-lyon1.fr

Understanding the relationship between biological sequences and the associated phenotypes is a fundamental problem in molecular biology. Accordingly, machine learning techniques have been developed to exploit the growing number of phenotypic sequences in automatic annotation tools. Typical applications include classifying protein domains into superfamilies [6, 9], predicting whether a DNA or RNA sequence binds to a protein [1], its splicing outcome [3], or its chromatin accessibility [4], predicting the resistance of a bacterial strain to a drug [2], or denoising a ChIP-seq signal [5]. Choosing how to represent biological sequences is a critical part of methods that predict phenotypes from genotypes. Kernel-based methods [6, 9, 8] have often been used for this task. They have been proven efficient to represent biological sequences in various tasks but only construct fixed representations and lack scalability to large amount of data. By contrast, convolutional neural networks (CNN) [1] have recently shown scalable and able to optimize data representations for specific tasks. However, they typically lack interpretability and require large amounts of annotated data, which motivates us to introduce more data-efficient approaches.

In this work we introduce CKN-seq, a strategy combining kernel methods and deep neural networks for sequence modeling, by adapting the convolutional kernel network (CKN) model originally developed for image data [7]. CKN-seq relies on a convolutional kernel, a continuous relaxation of the mismatch kernel [6], and the Nyström approximation. The relaxation makes it possible to learn the kernel from data, and we provide an unsupervised and a supervised algorithm to do so — the latter leading to a special case of CNNs.

On a transcription factor binding prediction task and a protein remote homology detection task, both approaches show better performance than DeepBind, another existing CNN [1], especially when the amount of training data is small. On the other hand, the supervised algorithm produces task-specific and small-dimensional sequence representations while the unsupervised version dominates all other methods on small-scale problems but leads to higher dimensional representations. Consequently, we introduce a hybrid approach which enjoys the benefits of both supervised and unsupervised variants, namely the ability of learning low-dimensional models with good prediction performance in all data size regimes. Finally, the kernel point of view of our method provides us simple ways to visualize and interpret our models, and obtain sequence logos. On some

simulated data, the logos given by CKN-seq are more informative and match better with the ground truth in terms of any probabilistic distance measures. We provide a free implementation of CKN-seq for learning from biological sequences, which can easily be adapted to other sequence prediction tasks and is available at https://gitlab.inria.fr/dchen/CKN-seq.

The fact that CKNs retain the ability of CNNs to learn feature spaces from large training sets of data while enjoying a reproducing kernel Hilbert space structure has other uncharted applications which we would like to explore in future work. First, it will allow us to leverage the existing literature on kernels for biological sequences to define the bottom kernel instead of the mismatch kernel, possibly capturing other aspects than sequence motifs. More generally, it provides a straightforward way to build models for non-vector objects such as graphs, taking as input molecules or protein structures. Finally, it paves the way for making deep networks amenable to statistical analysis, in particular to hypothesis testing. This important step would be complementary to the interpretability aspect, and necessary to make deep networks a powerful tool for molecular biology beyond prediction.

A full version of the paper is available at https://doi.org/10.1101/217257.

References

1. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology **33**(8), 831–838 (2015)
2. Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V.G., Bourgault, A.M., Laviolette, F., Corbeil, J.: Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics **17**(1), 754 (2016)
3. Jha, A., Gazzara, M.R., Barash, Y.: Integrative deep models for alternative splicing. Bioinformatics **33**(14), 274–282 (2017). https://doi.org/10.1093/bioinformatics/btx268
4. Kelley, D.R., Snoek, J., Rinn, J.L.: Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research **26**(7), 990–999 (2016)
5. Koh, P.W., Pierson, E., Kundaje, A.: Denoising genome-wide histone chip-seq with convolutional neural networks. Bioinformatics **33**(14), i225–i233 (2017)
6. Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S.: Mismatch string kernels for discriminative protein classification. Bioinformatics **20**(4), 467–476 (2004)
7. Mairal, J.: End-to-end kernel learning with supervised convolutional kernel networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 1399–1407 (2016)
8. Rangwala, H., Karypis, G.: Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics **21**(23), 4239–4247 (2005)
9. Saigo, H., Vert, J.P., Ueda, N., Akutsu, T.: Protein homology detection using string alignment kernels. Bioinformatics **20**(11), 1682–1689 (2004)