Biological Sequence Modeling with Convolutional Kernel Networks

Dexiong Chen [†] Laurent Jacob [‡] Julien Mairal [†]

[†]Inria - firstname.lastname@inria.fr [‡]CNRS - firstname.lastname@univ-lyon1.fr

Overview

Biological sequence modeling as a **supervised learning** problem

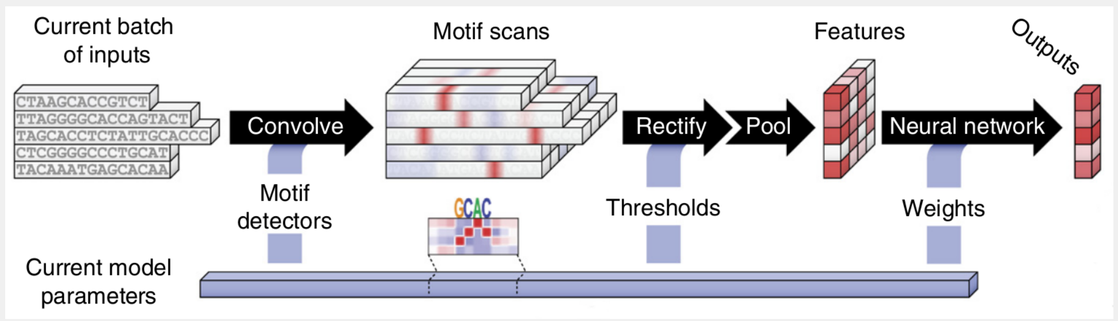
$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\Omega(f)$$

- $x_1, \dots x_n \in \mathcal{X}$ are biological sequences (DNA or proteins).
- Each sequence x_i is associated to some measurement $y_i \in \mathbb{R}$.

Goal: learning a predictive and interpretable function f.

CNNs and Kernel methods

CNNs [1]:



- $\mathcal{F} = \{f(x) = W_2 \max_i(\mathrm{ReLU}(W_1 x[i:i+k] + b_1)) + b_2 \,|\, W_1, W_2, b_1, b_2\}$.
- Yields a non-convex optimization problem in huge dimension.
- Provides good representations via backpropagation in practice.
- Open problems: Interpretation? Robustness?

Kernel Methods:

- $\mathcal{F}$ is a Hilbert space endowed with a Hibertian norm.
- Generic and flexible to type of data.
- Relatively easy to regularize by controlling $\|f\|^2_{\mathcal{F}}$.
- Lack of scalability.

Our approach [2]: mixing CNNs with kernel methods

- Build deep models (special case of CNNs) that are easy to regularize when few data are available.
- No tricks: no dropout, no batch normalization, parameter-free initialization.
- Two ways of learning:
  - **Simple unsupervised representation learning** algorithm without backpropagation (but high dimensions).
  - **Supervised learning** with back-propagation (low dimensions).
- Leverage interpretation from classical string kernels.

From mismatch kernel to convolutional kernel

String mismatch kernel:

$$K(x, x') = \frac{1}{|x||x'|}\sum_{i=1}^{|x|}\sum_{j=1}^{|x'|} K_0\left(x[i:i+k], x'[j:j+k]\right)$$

- K_0 equals to 1 if two k-subsequences are identical up to some mismatches otherwise 0.
- K_0 is fixed, thus not data or task-adaptive.
- Lacks scalability and interpretability in terms of motifs.

Convolutional kernel:

- K_0 becomes a Gaussian kernel over one-hot representations of k-subsequences.
- A natural feature map of x is $|x|^{-1}\sum_{j=1}^{|x|}\phi_0(x[j:j+k])$.
- K_0 is differentiable but still fixed, still not data or task-adaptive.
- Still lacks scalability and interpretability.

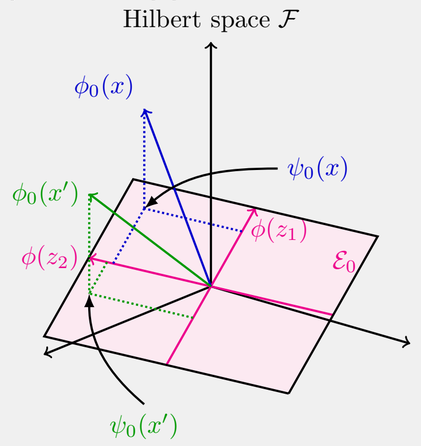
Software

Our Pytorch code is freely available at

`https://gitlab.inria.fr/dchen/CKN-seq`

Convolutional kernel network

Nyström approximation for CKN



- Regular Nyström approximation: randomly choose and fix p samples $Z := (z_1 \dots z_p)$, then

$$K_0(x, x') \simeq \langle \psi_0(x), \psi_0(x')\rangle,$$

and solve the linear problem.

- Convolutional kernel is differentiable, we can optimize over the filters Z

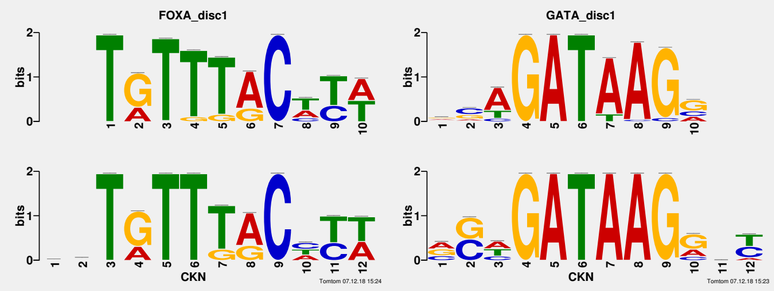
$$\min_{\beta\in\mathbb{R}^q,\, Z} \sum_{i=1}^{n} L(\beta^\top\psi(x_i), y_i) + \lambda\|\beta\|^2,$$

$$\psi_0(x) := K_{ZZ}^{-\frac{1}{2}} K_Z(x),$$

where $[K_{ZZ}]_{ij} = K_0(z_i, z_j)$, $[K_Z(x)]_i = K_0(z_i, x)$.

$$\psi(x) := |x|^{-1}\sum_{j=1}^{|x|}\psi_0(x[j:j+k]).$$

Visualization of learned filters



CKN variants

Unsupervised CKN (uCKN)

- Learns z_j with K-means over subsampled k-subsequences.
- Outperforms supervised CKN on some small-scale tasks.

Data augmented CKN (CKN+)

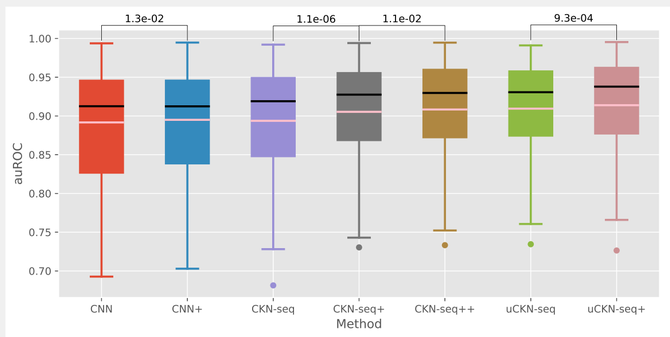
- Define a perturbation distribution Δ of sequences (e.g. adding SNPs), then augment each sequence x by $x + \delta$ with $\delta \sim \Delta$.

Hybrid CKN between uCKN and CKN+ (CKN++)

- Data augmented version, but we use the prediction of unsupervised CKN instead of y_i for $x_i + \delta, \delta \neq 0$.

Applications

DNA Transcription factor binding site prediction:



Protein homology detection on SCOP1.67

Method	auROC	auROC50
Mismatch	0.878	0.543
LA-kernel	0.919	0.686
LSTM	0.942	0.773
CNN (128 filters)	0.960	0.799
CKN-seq (128 filters)	0.965	0.819
CKN-seq (128 filters) + BLOSUM62	**0.973**	**0.835**
unsup CKN-seq (32768 filters)	0.958	0.806
Profile-based methods		
Mismatch-profile on SCOP 1.53	0.980	0.794
SW-PSSM on SCOP 1.53	0.982	0.904
CKN-seq (128 filters) + profile	**0.986**	**0.906**
unsup CKN-seq (4096 filters) + profile	0.968	0.863

Reference

[1] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 2015.

[2] D. Chen, L. Jacob, and J. Mairal. Biological sequence modeling with convolutional kernel networks. *Bioinformatics*, February 2019.