Protein domain sequences analysis using LSTM-RNN

Marseille JOBIM 2018

Tristan Bitard-Feildel – Laboratory of Computational and Quantitative Biology

Introduction

An shallow walkthrough into deep learning territory

Background :

NVIDIA Titan Xp for molecular dynamics project (GPU acceleation of MDs).

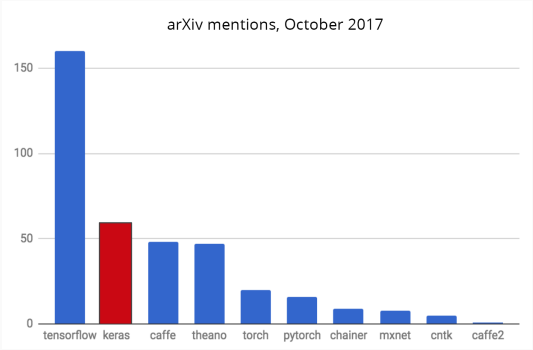
It's nice to learn new things, could we try to do some deep learning ?

Where to start ?

Found 26 different softwares/libraries. 20 with C/C++ support, 18 with Python support.

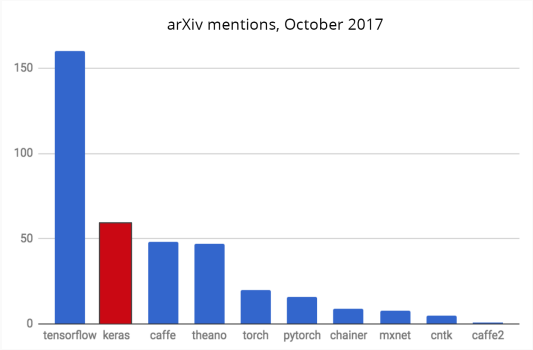
- Large and active community of users
- Prototyping is fast
- Python and R (Keras), C++/Python (Tensorflow)
- Different backends available on Keras
- TFLearn (Tensorflow) $\sim$ Scikit-learn (Python)

An shallow walkthrough into deep learning territory



- Large and active community of users
- Prototyping is fast
- Python and R (Keras), C++/Python (Tensorflow)
- Different backends available on Keras
- TFLearn (Tensorflow) $\sim$ Scikit-learn (Python)

An shallow walkthrough into deep learning territory



- Large and active community of users
- Prototyping is fast
- Python and R (Keras), C++/Python (Tensorflow)
- Different backends available on Keras
- TFLearn (Tensorflow) $\sim$ Scikit-learn (Python)

From molecular dynamics to protein sequences

An shallow walkthrough into deep learning territory

What is a biological protein sequence ?

A biological protein sequence

Definition

- A protein sequence : list of amino acids (set of more or less 20 different types)
- A biological protein : has a function associated to a context and has been explored by evolution

A biological protein sequence

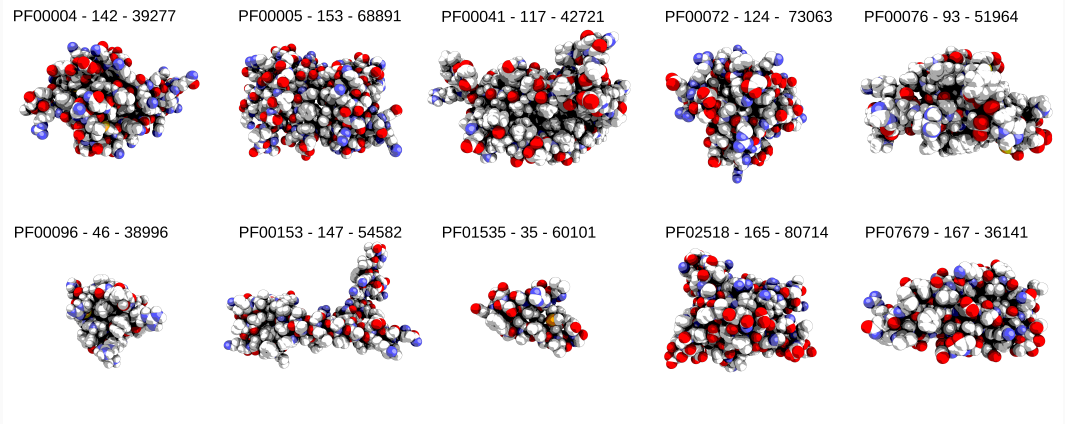
Definition

- A protein sequence : list of amino acids (set of more or less 20 different types)
- A biological protein : has a function associated to a context and has been explored by evolution

Which properties have been maintained and why ?
Can we explore new protein sequences ?

The context

Protein families represent the molecular diversity associated to a fold / function (broad meanning)



protein family – max size – number of sequences

Biological sequence generation

The infinite monkeys typewriter theorem

- random generation according to propensity
- Hidden Markov Models (order 1) :
  $P(X_k = x|X_{k-1} = x_{k-1}, ..., X_{k-n} = x_{k-n}) = P(X_k = x|X_{k-1} = x_{k-1})$
- ...

Long Short Term Memory Recurrent Neural Networks

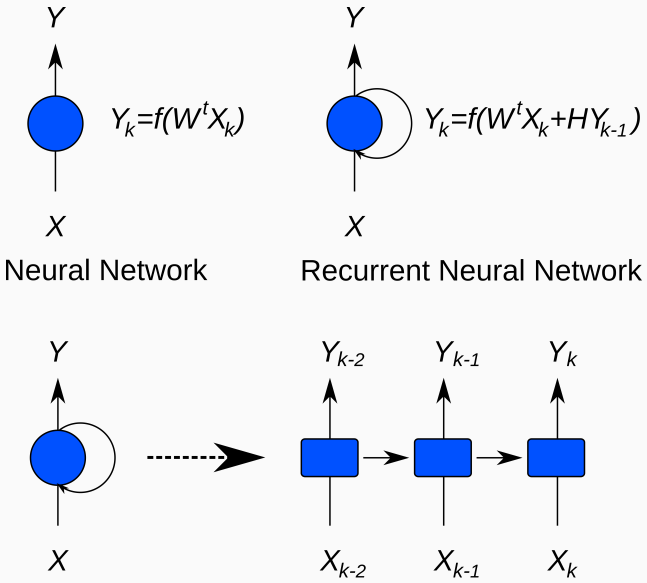
LONG SHORT-TERM MEMORY

Neural Computation 9(8):1735–1780, 1997

Sepp Hochreiter
Fakultät für Informatik
Technische Universität München
80290 München, Germany
hochreit@informatik.tu-muenchen.de
http://www7.informatik.tu-muenchen.de/~hochreit

Jürgen Schmidhuber
IDSIA
Corso Elvezia 36
6900 Lugano, Switzerland
juergen@idsia.ch
http://www.idsia.ch/~juergen

Long Short Term Memory Recurrent Neural Networks



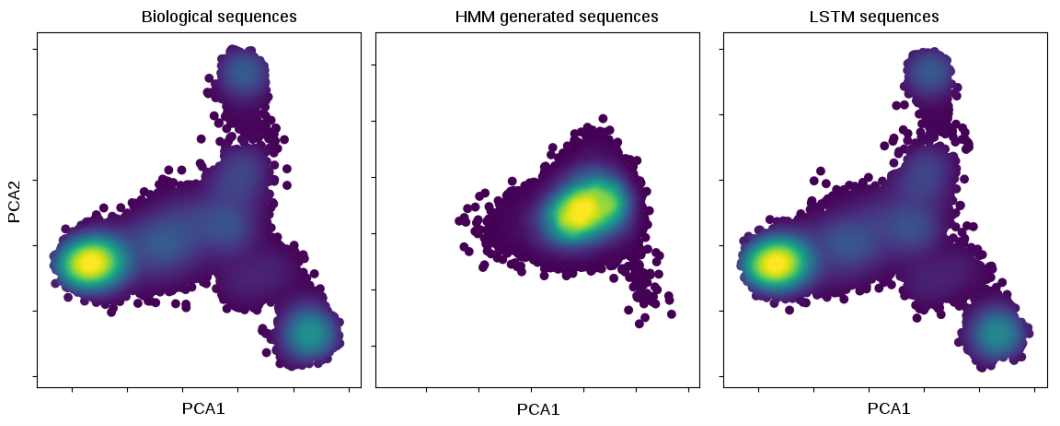
Long Short Term Memory Recurrent Neural Networks

Generate protein sequences using an amino acid based LSTM generator.

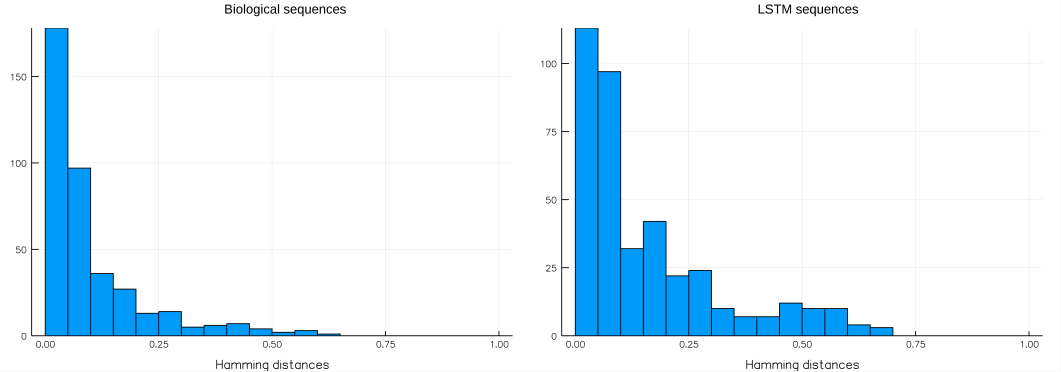
- Input : unaligned protein sequences of the family (with “start” and “stop” symbols)
- Network : three layers of LSTM cells (unit size 256), many-to-many network)
- Output : same number of sequenecs thant the original protein family

Evaluation

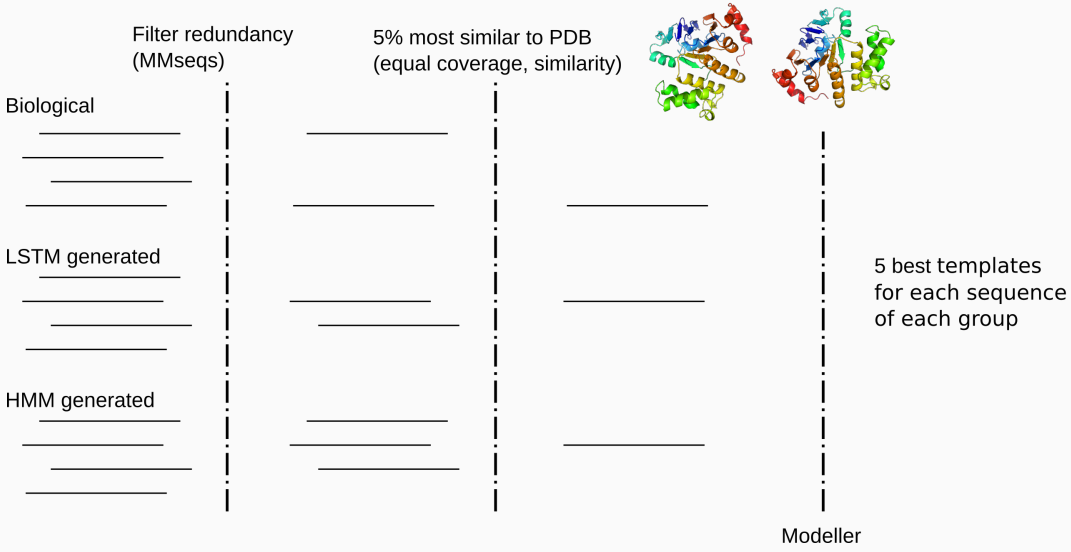
Sequence inspection



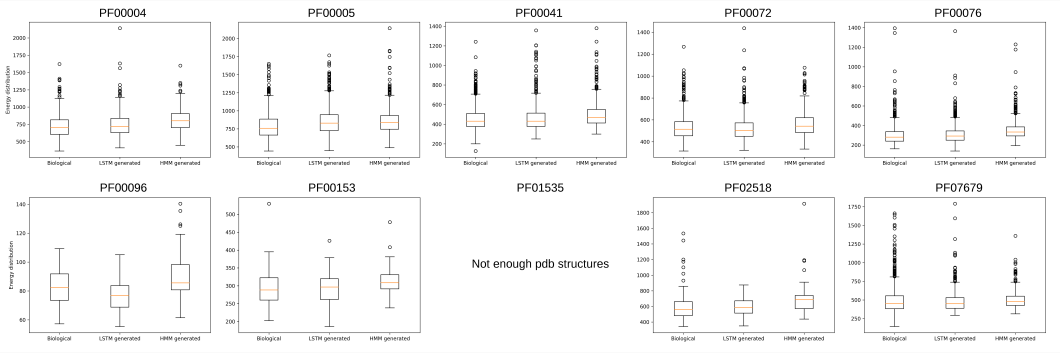
Sequence inspection



The best to be modelled



The best to be modelled



Classification

Quick classification using a small different LSTM (64 hidden units, many-to-one network, shown below the**validation set accuracy**) :

Biological vs	PF00004	PF00005	PF00041	PF00072	PF00076
HMM sequences	0.868	0.837	0.681	0.787	0.655
LSTM sequences	**0.496**	**0.502**	**0.502**	**0.565**	**0.504**
Biological vs	PF00096	PF00153	PF01535	PF02518	PF007679
HMM sequences	0.56	0.775	0.576	0.796	0.787
LSTM sequences	**0.494**	**0.499**	**0.518**	**0.496**	**0.501**

Better results with hyperparameters optimization [1] (but time consuming), exemple PF00076 :

- Biological versus HMM sequences : 0.655 $\rightarrow$ 0.954
- Biological versus LSTM sequences : 0.504 $\rightarrow$ 0.55

1. Bergstra, Bardenet, Bengio & Kégl, Algorithms for Hyper-Parameter Optimization, NIPS. 2011.

Classification

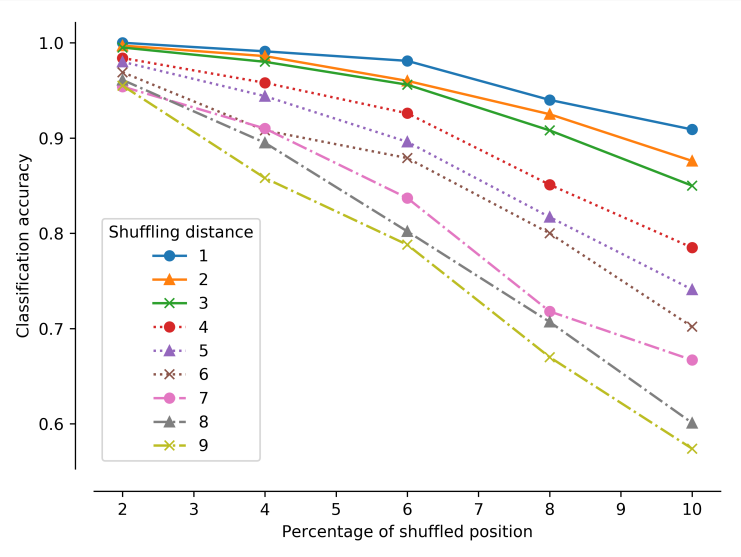
Summary

- Very hard to distinguish **Biological** from **LSTM generated** sequences
- Hyperparameter optimization can greatly improved performances

Inside the classification

- Take N sequences
- Take a random position fo each sequence and an other at distance d
- Randomize the position
- Repeat for $x\%$ of the positions

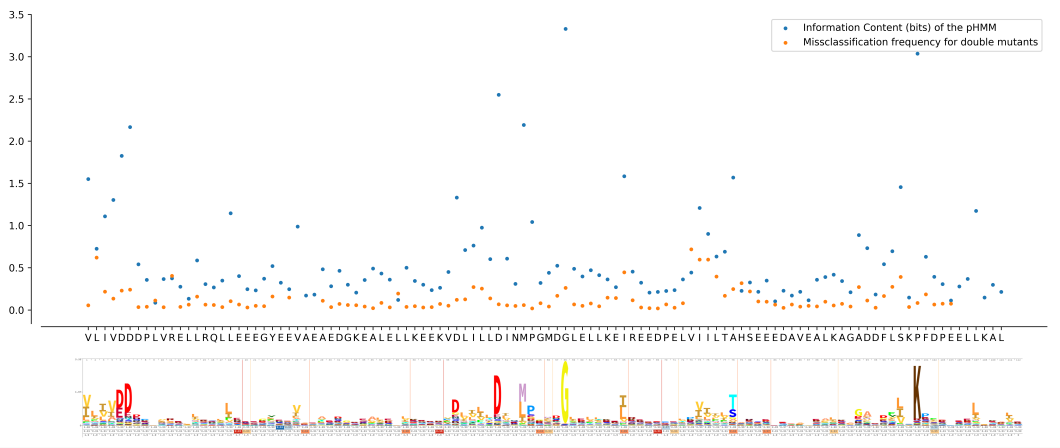
Inside the classification



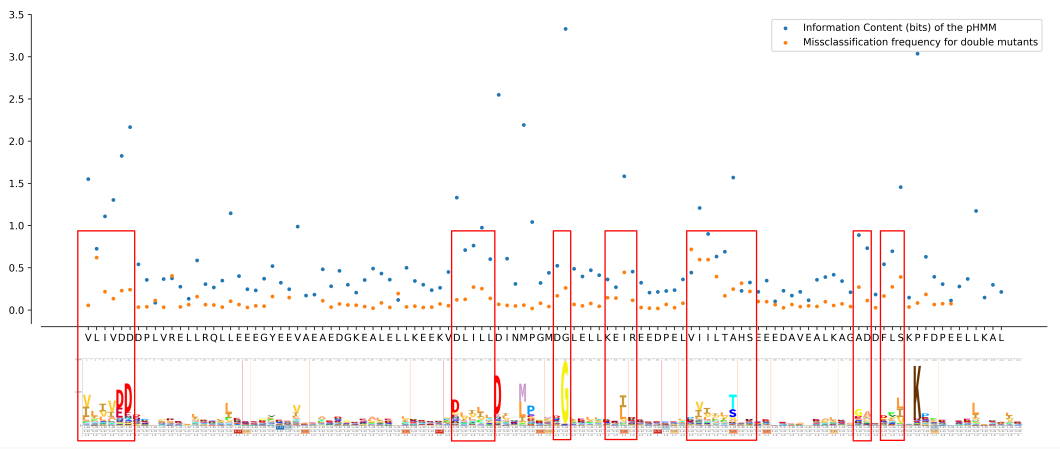
Inside the classification

- Take a sequence of a protein
- Make all possible double mutants and classify them
- For each single position compute the number of time it is missclassified (and normalize)

Inside the classification



Inside the classification



The direct coupling analysis (DCA) framework

Can we distinguish biological sequences from DCA generated sequences ?
(DCA generated sequences are known to be near native like).

Summary

LSTM-RNN are an interesting architecture to analyse protein sequences and can :

- **capture structured relationships between positions**
- **generate "biological" like sequences**
- **be used for classification (protein domain detection ?)**

Next :

- **using Variationnal AutoEncoder to modelize the sampling space**
- **adding structural information to the model**

Laboratory of Computational and Quantitative Biology

Analytical Genomics

- Alessandra Carbone

Statistical Genomics and Biological Physics

- Martin Weigt
- **Pierre Barrat-Charlaix**