Metagenomic Classification: a Deep Learning Approach

Daniele Bellani

01 • 02 • 2018

Some facts...

- Great excitement about **deep learning** models, due to successful applications in computer vision, speech recognition, natural language processing...
- In particular, CNN and RNN (in various flavours) obtained great results when treating **sequential data**
- **DNA** and **RNA** streams are a kind of sequential data
- First attempts on biomedical tasks, mainly involving **genomic data**
- Their application on **metagenomic data** is yet to come...

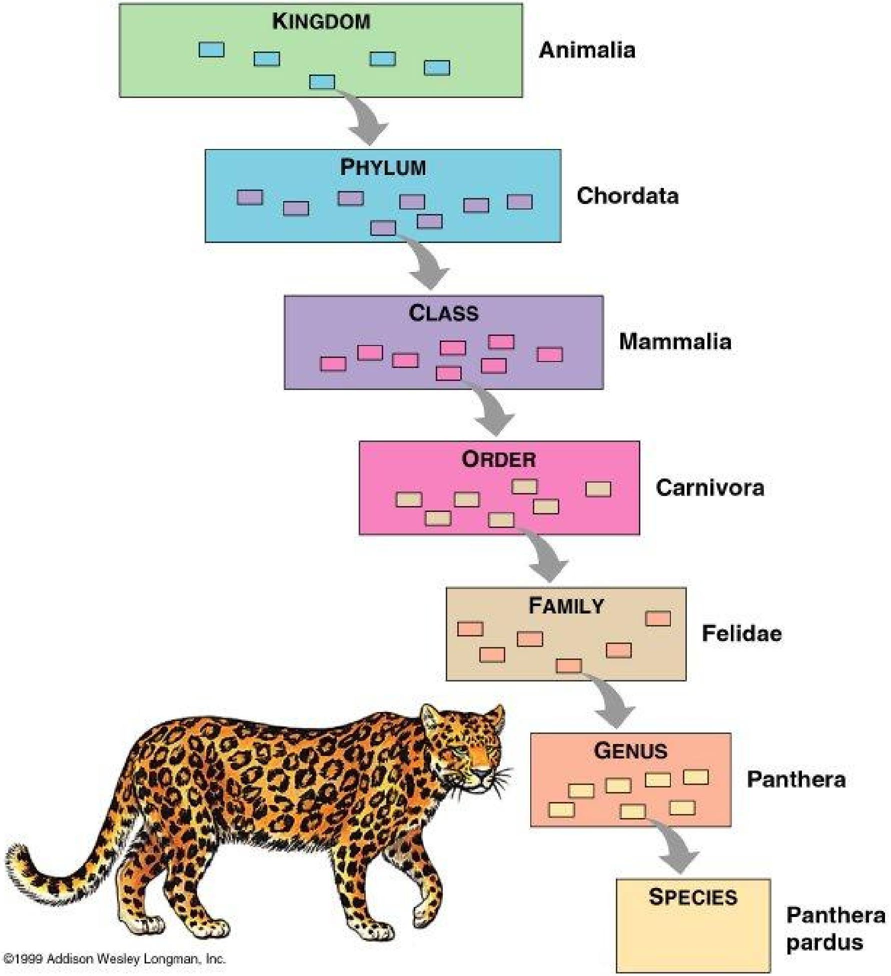
Biological background

Some new words...

- **Metagenomics** - study of genetic material recovered directly from **environmental** samples.
- **Reads** - fragments of genetic material.
- **Sequencing** - the process of extracting **reads** from biological samples.
- **K-mer** - DNA/RNA string of length **k**.
- **Prokaryotic cell** - type of cell which differs from the **eukaryotic cell** for not having a nucleus, having simpler internal structure and for **not** assembly in multicellular organisms (prokaryotes are **unicellular organisms**).

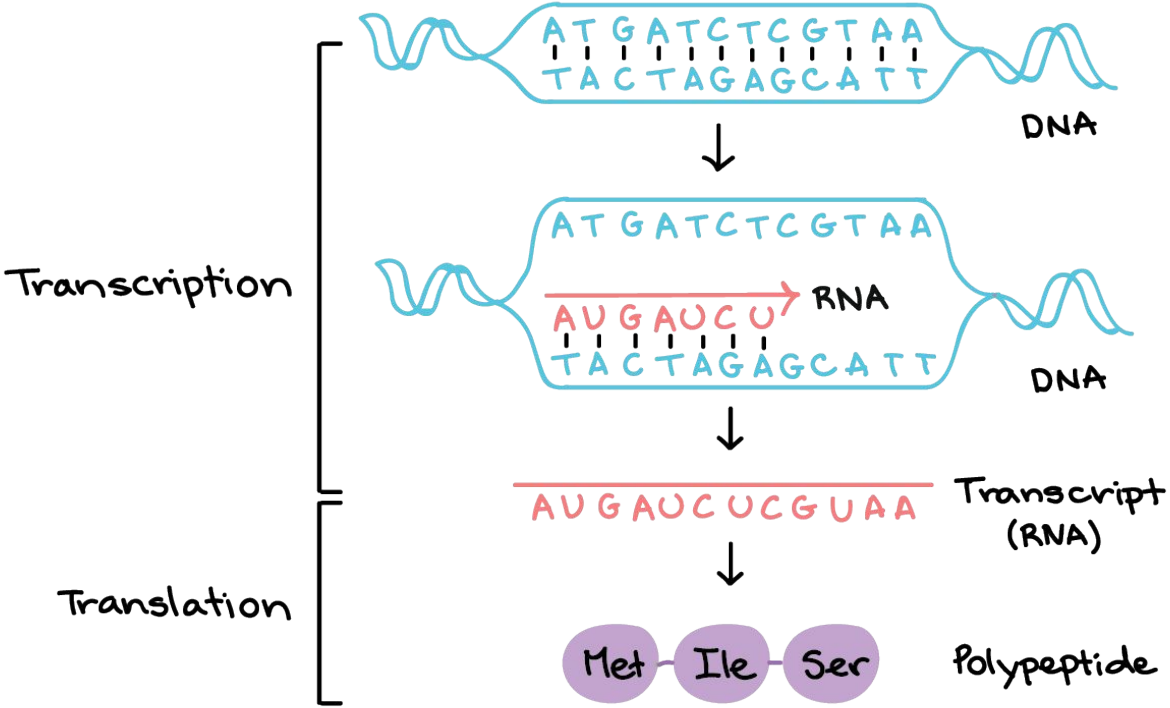
Some new words...

- **Virus** - protein shell containing genetic material.
- **Microbiome** - could indicate either a population of microorganisms or the collection of their genetic material (**genomes**).
- **Taxon** - a population, or group of populations of organisms which are usually inferred to be **phylogenetically** related and share characteristics which differentiate them from other groups. Taxons are organized in a **taxonomical ranking**.



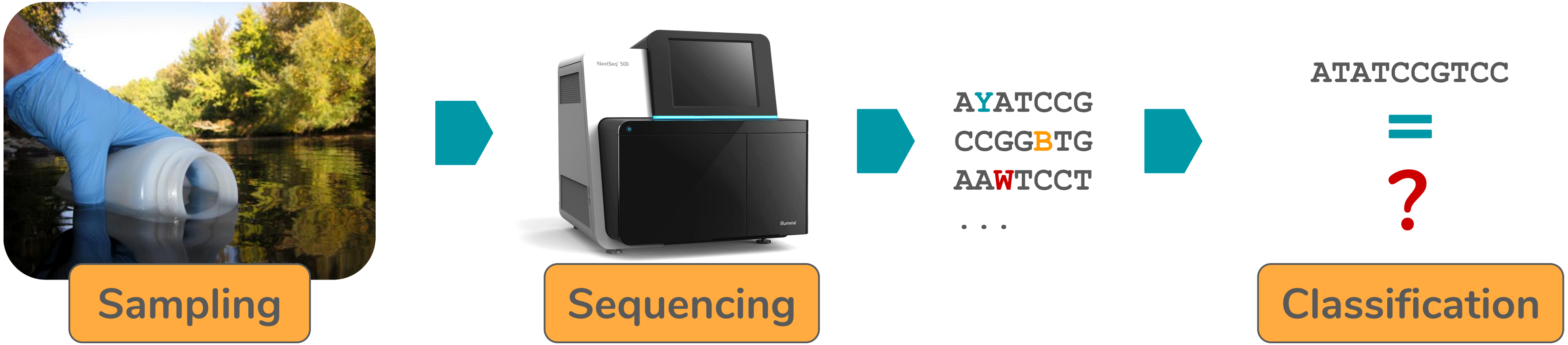
©1999 Addison Wesley Longman, Inc.

The central(ish) dogma of genomics



Metagenomic Classification

An ordinary metagenomic experiment



Sampling

Sequencing

AYATCCG
CCGGBTG
AAWTCCT
. . .

ATATCCGTCC
=
?

Classification

Metagenomic classification

Assign a **taxon** to a **read:**

ATCCACATATTCTTTCTAATCTCATTTTTATCTACATAAAGTAAAAGTTATTCACAAAAACGTAGCTTTA

Kingdom: *Bacteria*

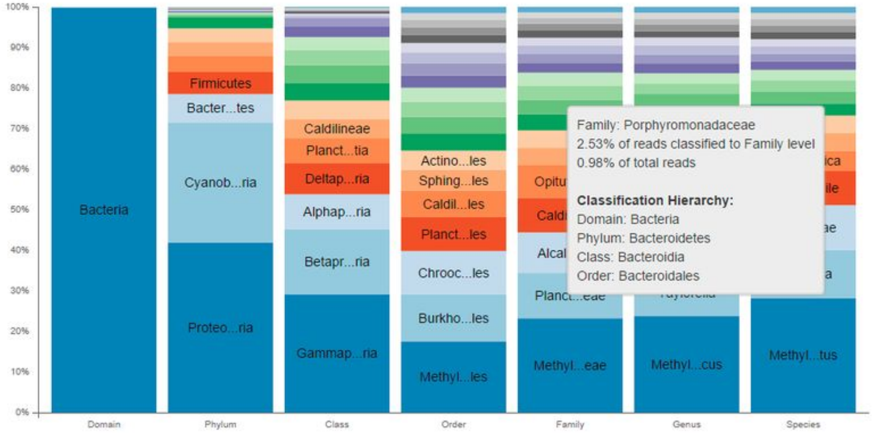
. . .

Genus: *Pelosinus*

Specie: *Pelosinus Fermentans*

Why is it important? - Data Analysis

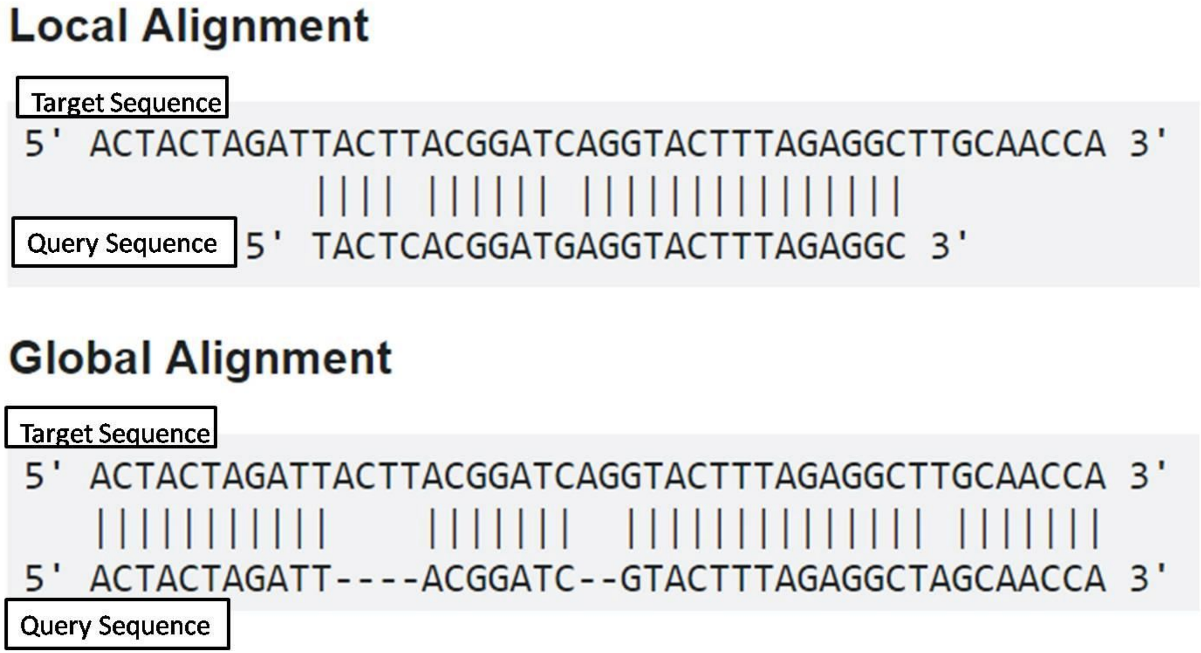
- Classification of reads is important to infer the composition of **microbial communities** (**microbiomes**) of the sample, and thus of the environment it comes from.
- Typical analyses relying on such operation include **pollution** analysis, **pathogens** detection, **air/water quality** analysis, etc.
- Samples can come even from the human body (we host a number of microorganism 3 times larger than the number of human cells!). These are used in **medicine** and **nutrition**.



https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/kraken-metagenomics.html

How to classify - Alignment

- Most of state-of-the-art procedures are based on **sequence alignment**



Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
          5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Query Sequence

Global Alignment

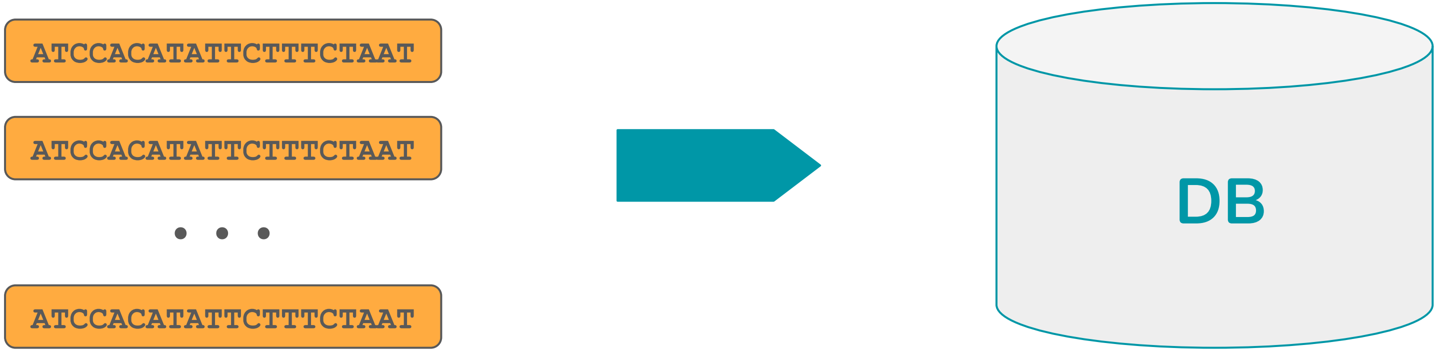
Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

How to classify - Alignment

- Each sequence produced by the sequencer serves as **query sequence**, while the **target sequence** is every sequence stored in the **genomes database**.
- The comparison can be performed on **nucleotide level** or **protein level**.



How to classify - Alignment

- Original dynamic programming algorithms could perform **global alignment** and **local alignment** with complexity **O(*mn*)**
- Alignment algorithms have been improved in the last 30 years, achieving a linear complexity - **O(*n*)**
- But alignment still remains a **time-consuming** operation:
  - Latest **Next Generation Sequencing** processes (**NGS**) produce billion of sequences
  - Databases contains thousands of fully-decoded genomes

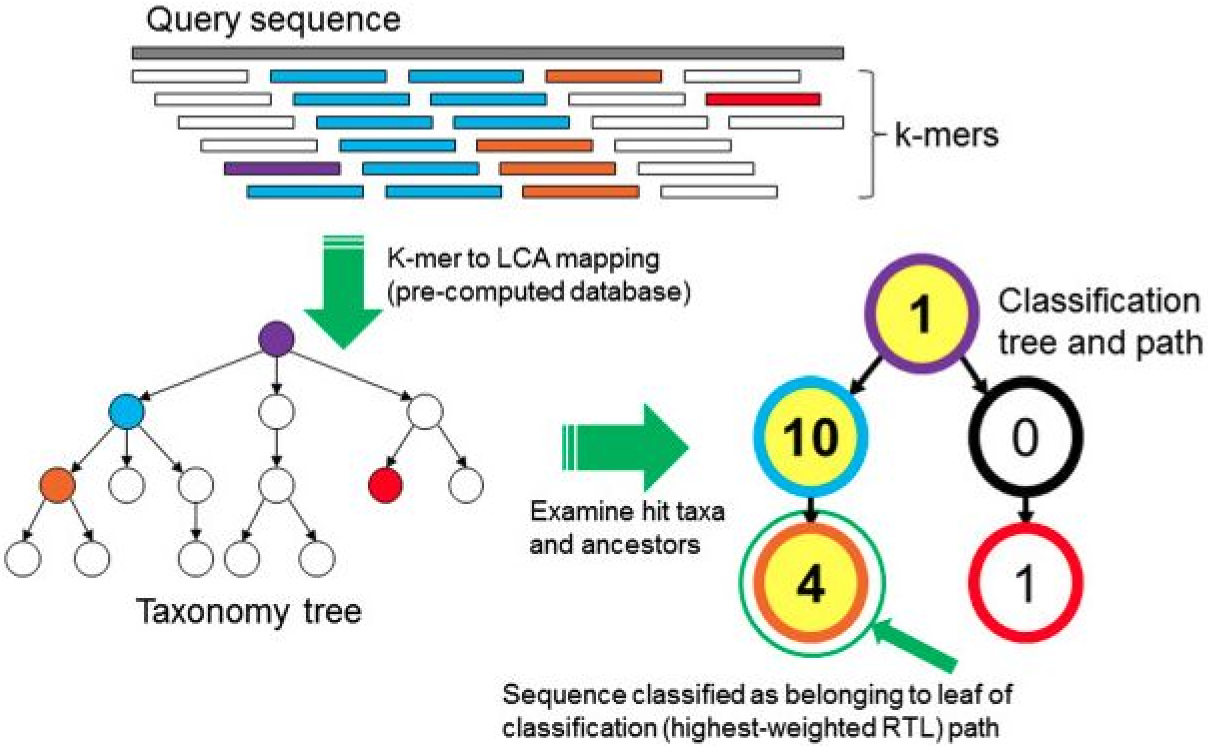
illumina®

NCBI National Center for Biotechnology Information

How to classify - Alignment-free methods

- In recent years, a lot of approaches not relying on alignment have been published
- These new methods are roughly distinguished in two classes:
  - **Marker gene approaches**
  - **K-mer based approaches**
- These methods have demonstrated good performances and high speed, but are weak on **real data** (due to variations) and suffer from **sampling bias.**
- The methods in these categories still rely on pre-constructed genomic databases or self-constructed mappings (these can be hundreds of GB in size!).

How to classify - Alignment-free methods



Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3), R46.

How to classify - Machine learning methods

- Some machine learning attempts to classification, mainly with:
  - Naive Bayes
  - SVM
- Minor attempts with:
  - Nearest Neighbor
  - Random Forest
- Machine Learning models were celebrated by biologists for their high **sensitivity (recall)**.
- Nevertheless, they have never catch the heart of biologists, due to small improvement on runtime.

Kaiju

NBC

MG-RAST

MetaPhlAn

Kraken

LMAT

MEGAN

DectICO

BLAST

Centrifuge

mOTU

CLARK

CARMA3

RAPSearch

GOTTCHA

k-SLAM

DIAMOND

Phymm

PhymmBL

How to classify - Validation

- Benchmarks can be performed on both **real** or **simulated** data.
- The principal metrics for the validation of a metagenomic classifier include:
  - Accuracy
  - Precision
  - Recall (**sensitivity**)
  - Measures combining Precision and Recall (**F1**, **ROC**)
  - Speed (rpm)
  - Predicted vs Real microbial distributions correlation
  - Fraction of unclassified sequences (main issue on real data)
- It is important to assess the robustness of the method by executing it on different-sized reads.

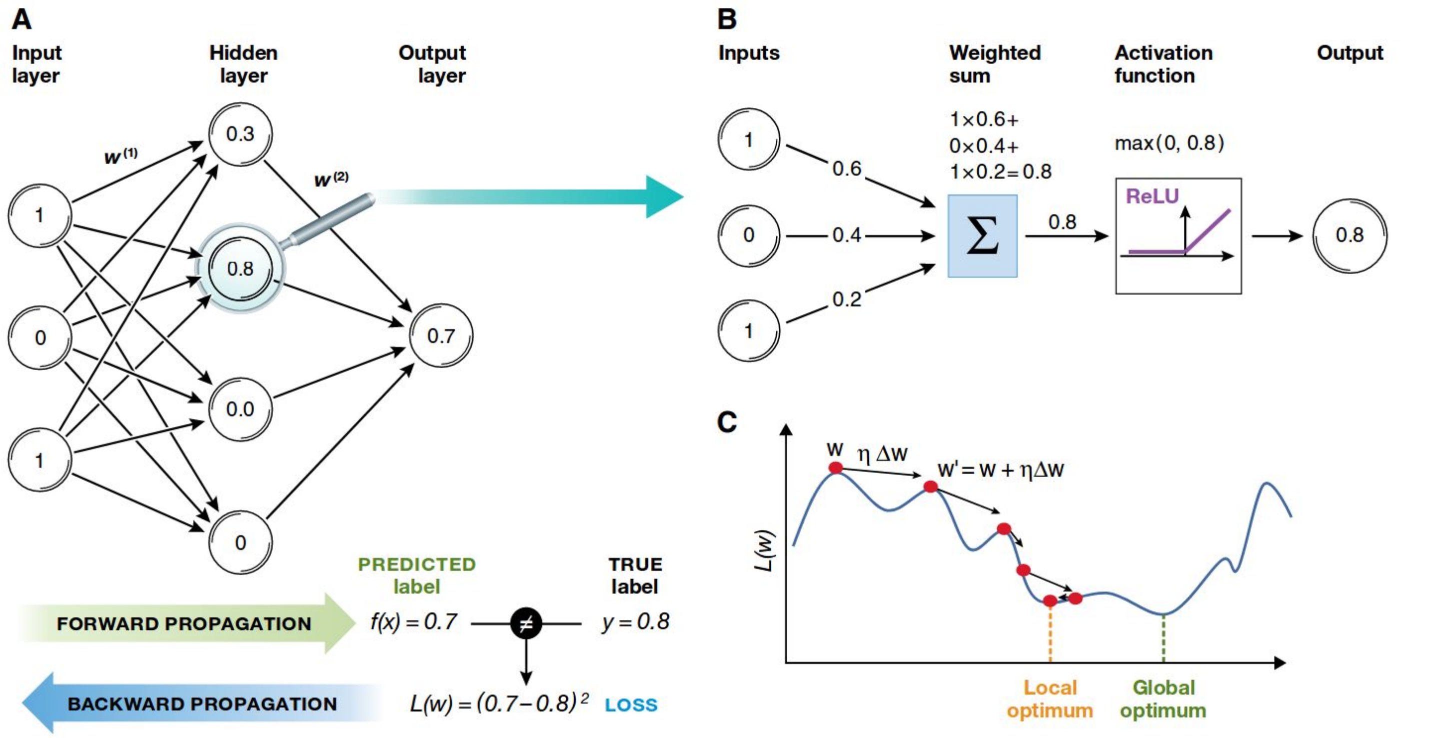
Metagenomic classification - Recap

- Sequence classification is a fundamental part of the metagenomic pipeline
- Alignment algorithms are still the most used approach but:
  - Time-consuming
  - Dependence on databases (which may not be complete)
- Alignment-free algorithms are faster and obtain even better results but:
  - Require massive memory and disk space
  - Still dependant on genomic databases (**sampling bias**)
  - Weak on real data

The goal is to find a method to accurately classify the highest percentage of different-sized reads from real and simulated datasets in the least time-consuming and memory-consuming way.

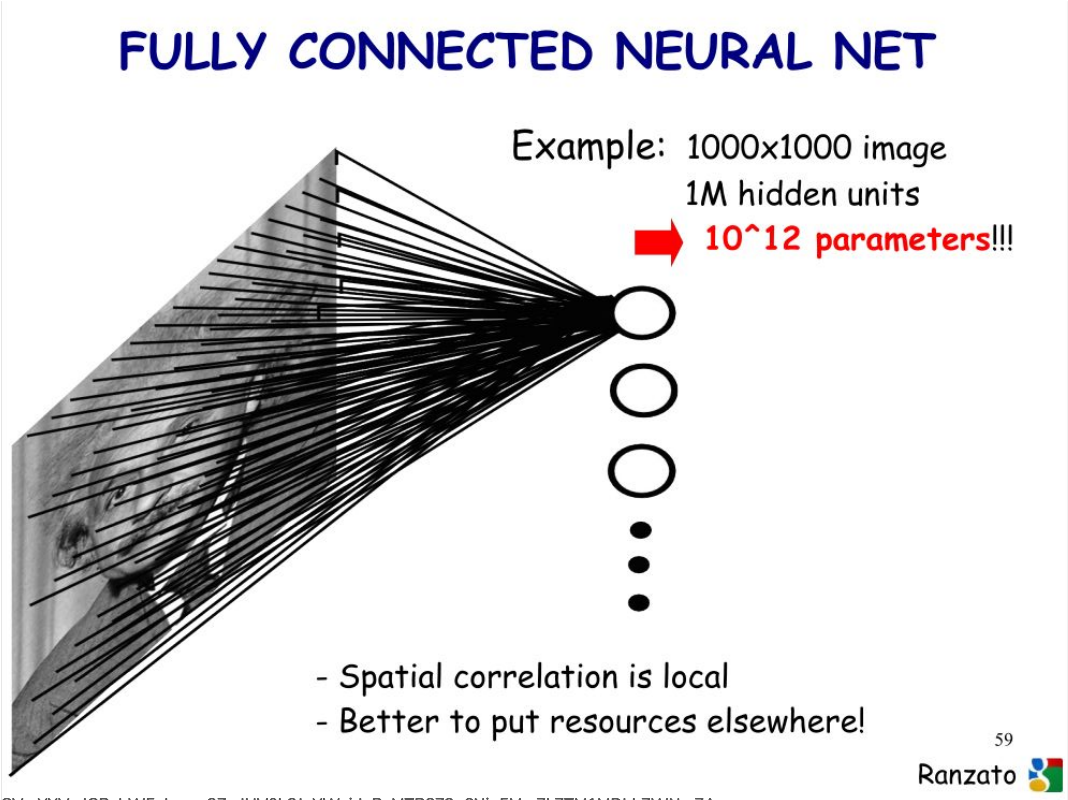
Deep Learning for Sequential Data

Perceptron and Multi-Layer Perceptron



Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular systems biology, 12(7), 878.

Convolutional Neural Networks



FULLY CONNECTED NEURAL NET

Example: 1000x1000 image

1M hidden units

10^12 parameters!!!

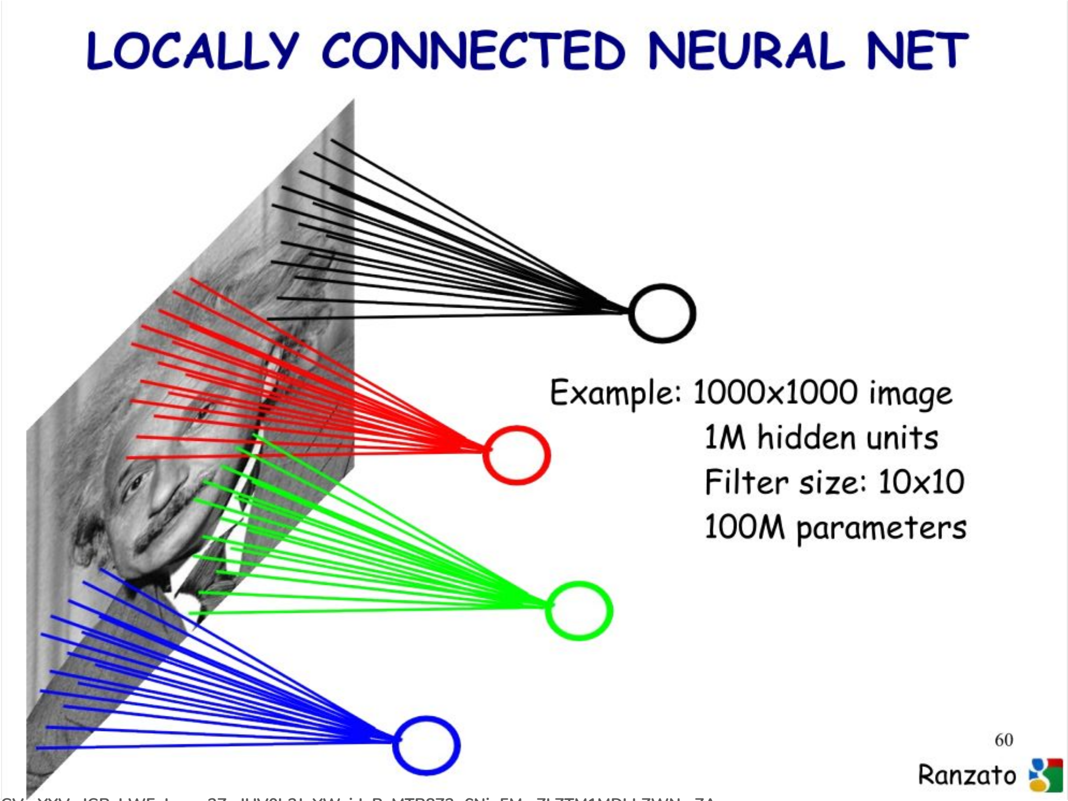
- Spatial correlation is local
- Better to put resources elsewhere!

59

Ranzato

https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxsc3ZydHV0b3JpYWxjdnByMTR8Z3g6Njg5MmZkZTM1MDhhZWNmZA

Convolutional Neural Networks



LOCALLY CONNECTED NEURAL NET

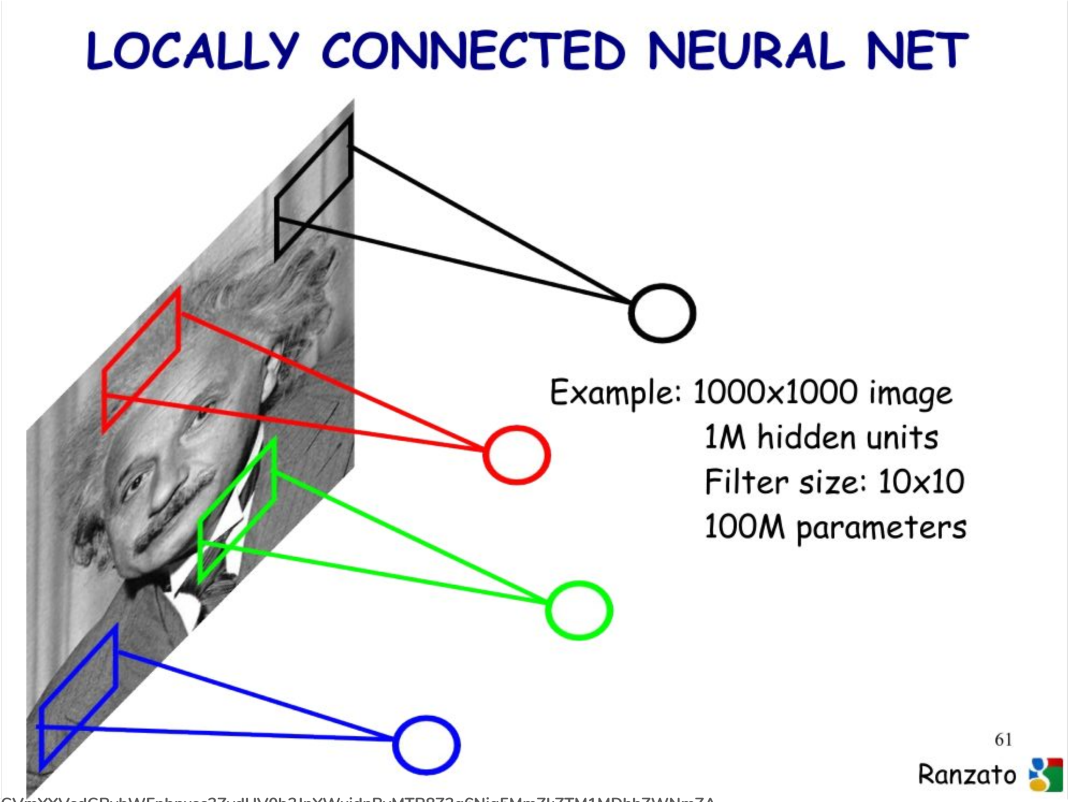
Example: 1000x1000 image
1M hidden units
Filter size: 10x10
100M parameters

60

Ranzato

https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxsc3ZydHV0b3JpYWxjdnByMTR8Z3g6Njg5MmZkZTM1MDhhZWNmZA

Convolutional Neural Networks



LOCALLY CONNECTED NEURAL NET

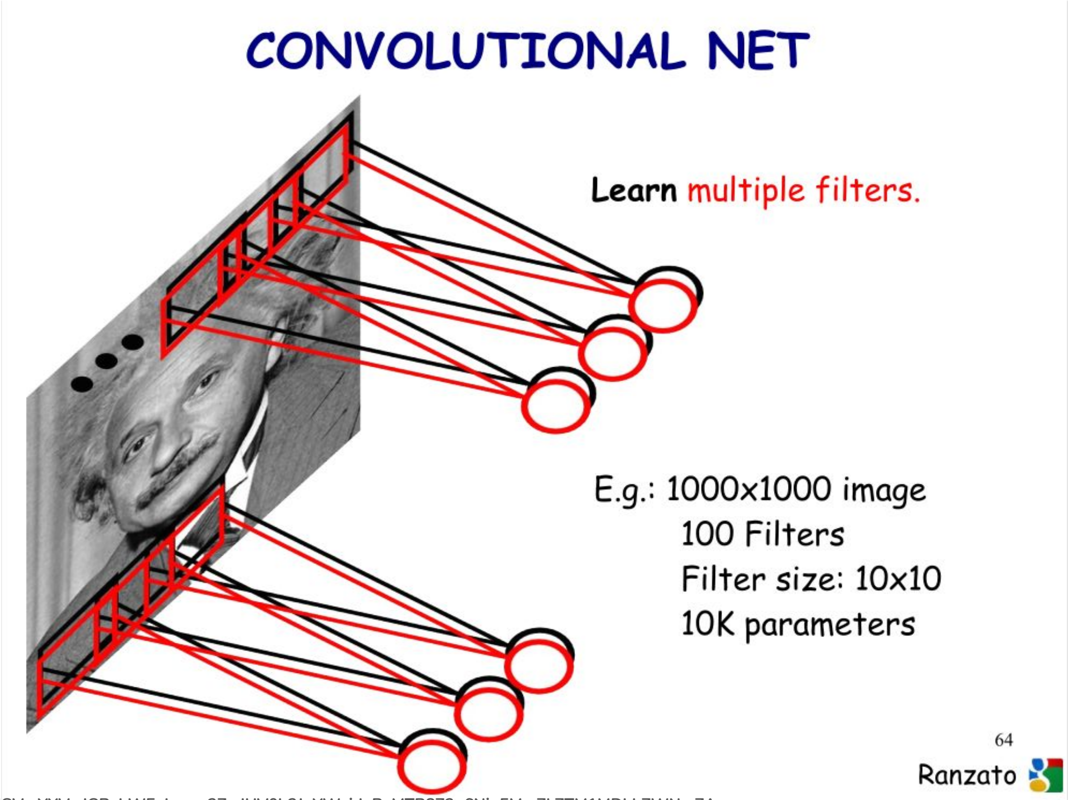
Example: 1000x1000 image
1M hidden units
Filter size: 10x10
100M parameters

61

Ranzato

https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxsc3ZydHV0b3JpYWxjdnByMTR8Z3g6Njg5MmZkZTM1MDhhZWNmZA

Convolutional Neural Networks



CONVOLUTIONAL NET

Learn multiple filters.

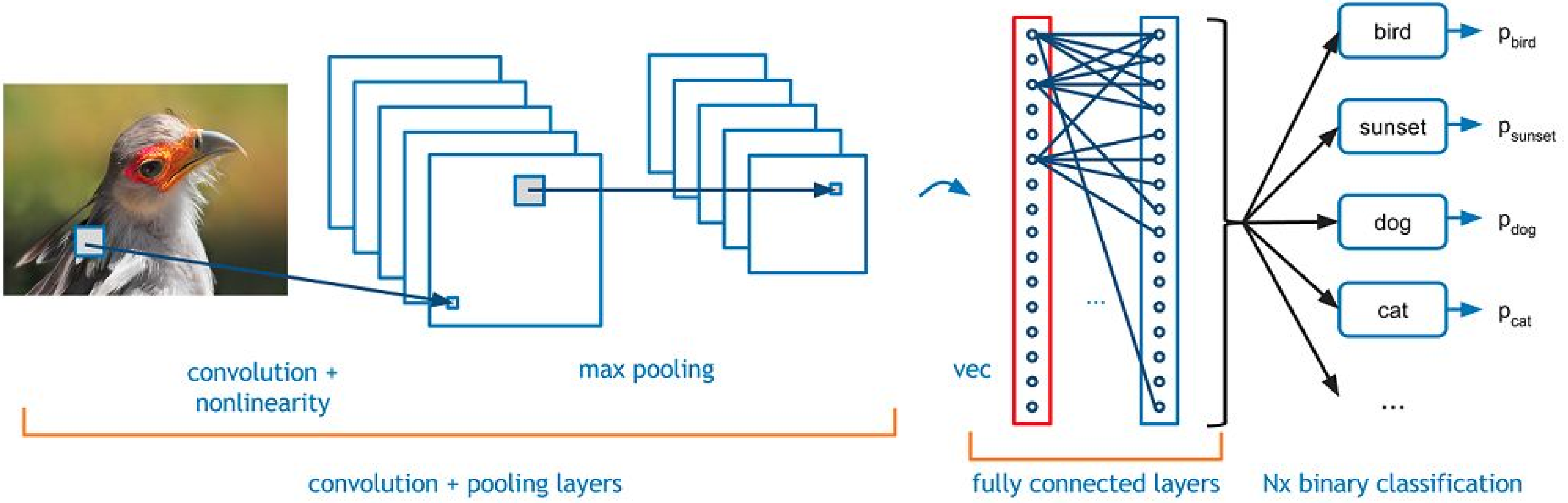
E.g.: 1000x1000 image
100 Filters
Filter size: 10x10
10K parameters

64

Ranzato

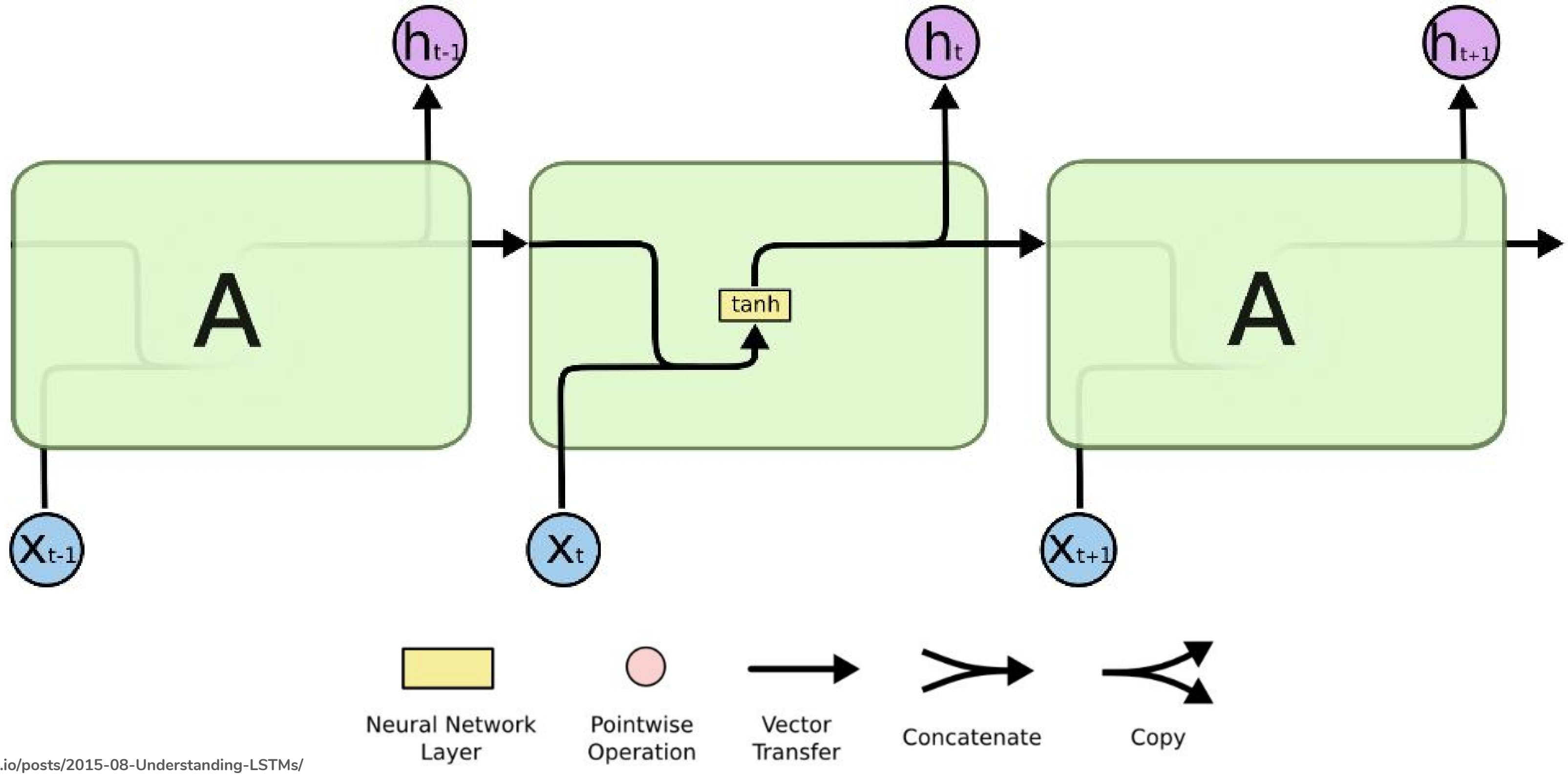
https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxsc3ZydHV0b3JpYWxjdnByMTR8Z3g6Njg5MmZkZTM1MDhhZWNmZA

Convolutional Neural Networks



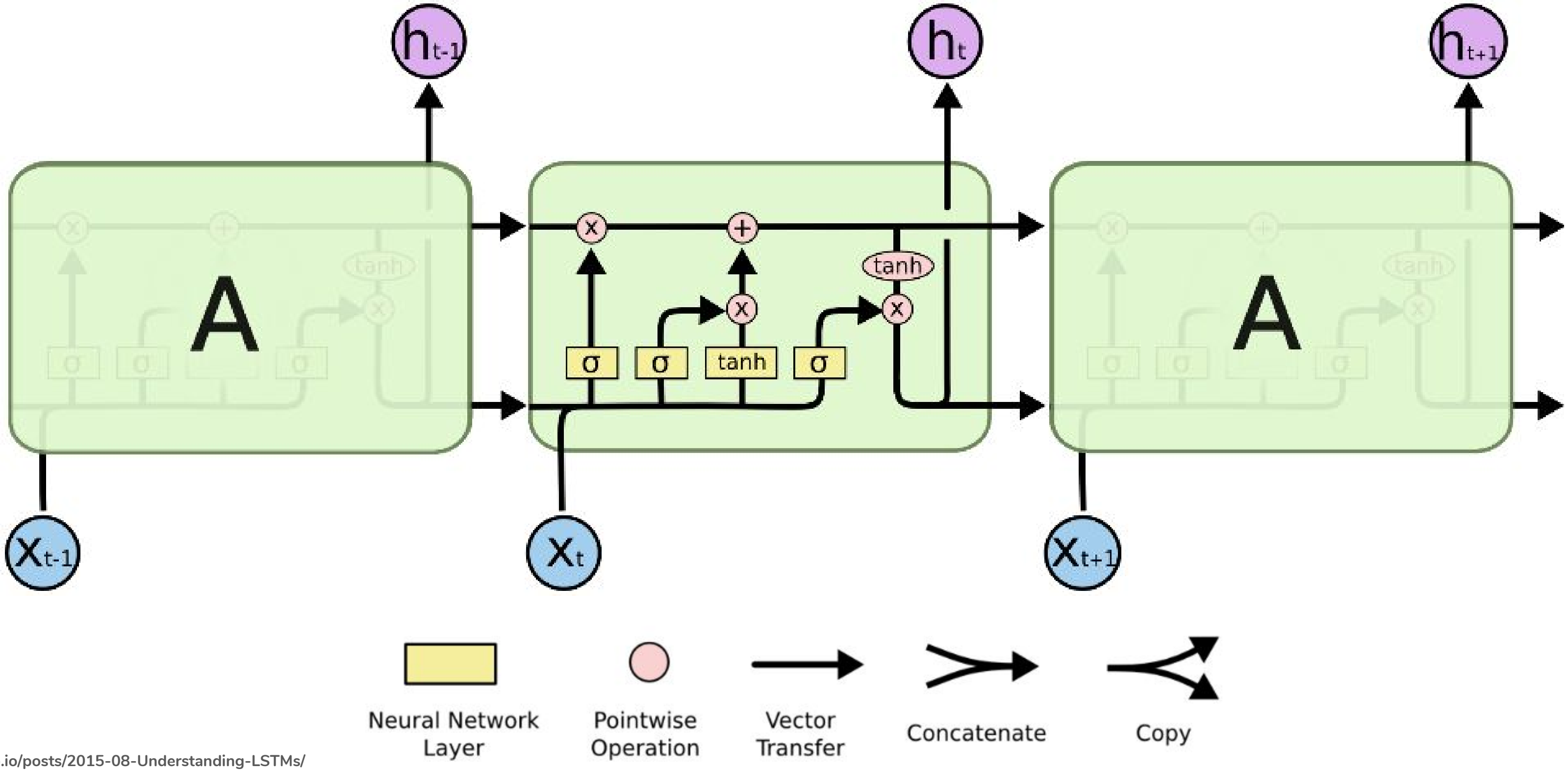
https://adeshpande3.github.io/adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/

Recurrent Neural Networks



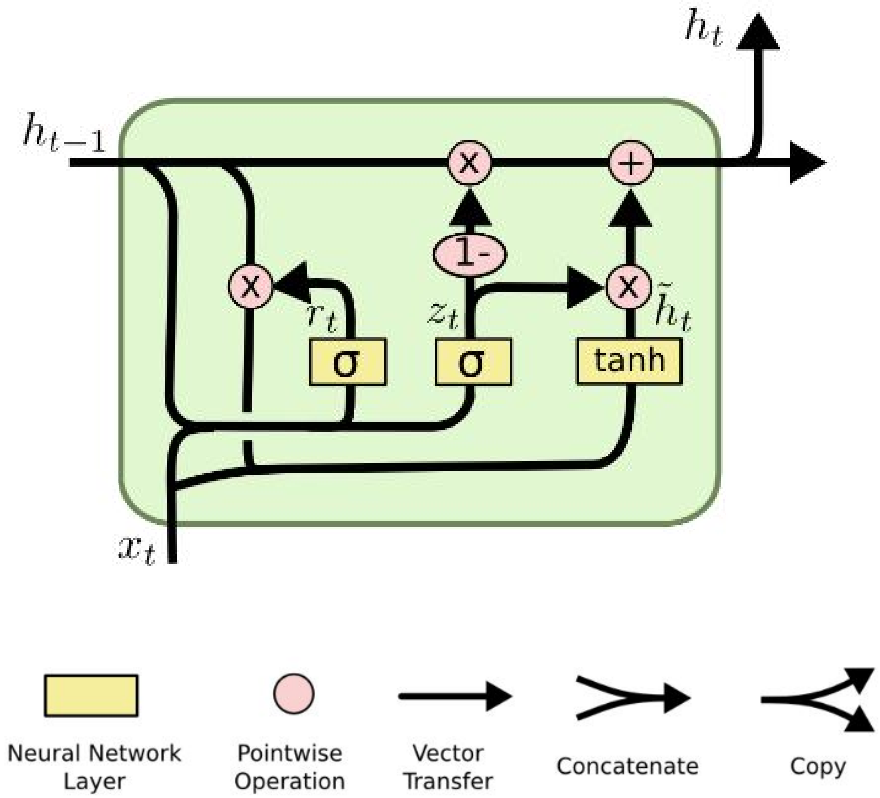
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Recurrent Neural Networks - LSTM



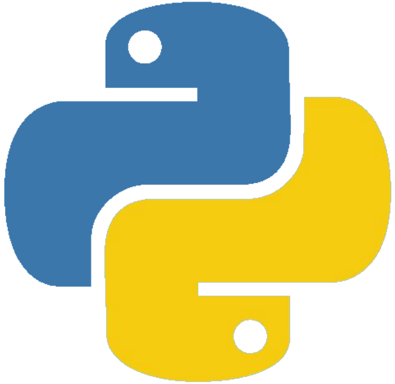
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Recurrent Neural Networks - GRU



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

DNN tools



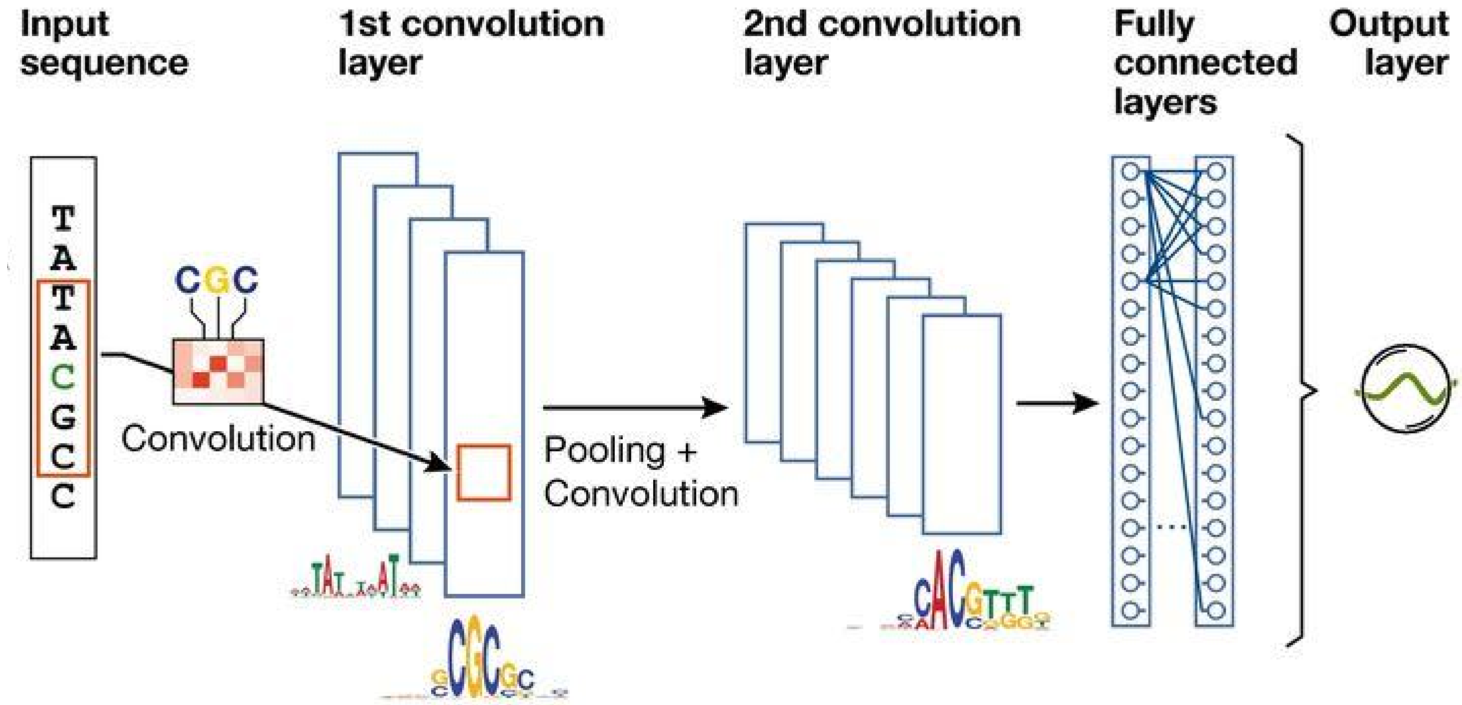
TensorFlow

DNN for Metagenomic Classification

Overview

- Deep learning models have shown good performances on **image classification** and **speech recognition**
- They manage to memorize patterns hidden in input data, which can be quite complex
- Unlike k-mer based approaches, neural networks (stateful RNN above all) can treat the input sequence as a whole stream of data, making them more robust to local variations
- Unlike alignment based approaches, no comparison between sequence is needed

Overview



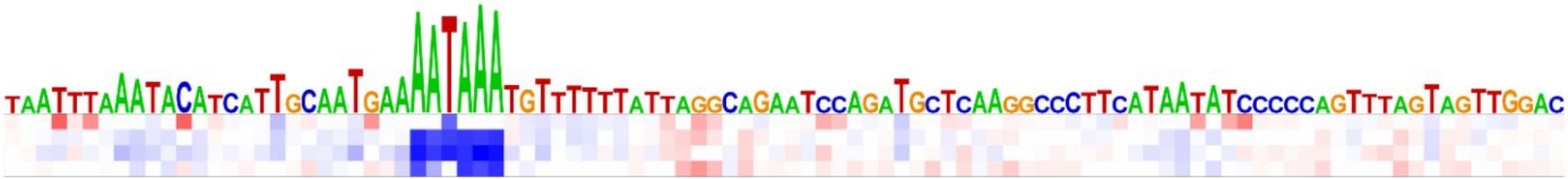
Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular systems biology, 12(7), 878.

Motivations

- **Speed**
    - Heavy training phase, but prediction phase grows linearly with the input
    - No comparison between sequences
    - GPU speed-up
- **No feature extraction**
    - Both CNN and RNN take raw (yet vectorized) sequences as input
- **Genome DB independence**
    - No need for complete reconstructed genome for each taxon, but a labeled set of sequences.
- **Local variations-tolerant**
    - All the sequence points are examined and weighed, no perfect match is seeked

Motivations

- **2-level support**
  - Could work with either protein-level or nucleotide-level sequences
- **Motif discovery**
  - Possibility to explain predictions by highlighting the most significative portions of input data



Leung, M. K. K., Delong, A., & Frey, B. J. (2017). Inference Of The Human Polyadenylation Code. *bioRxiv*, 130591.

Issues

- **Class number**
  - Performance of ML algorithms decreases with the increasing number of classes
- **Unknown species**
  - Metagenomic samples are quite often full of **uncategorized** or **unexpected** species
- **Motif discovery**
  - Novel approach, still a few works, could be the toughest part
- **Parameter tuning, structure choice and learning**
  - NN are full of parameters to be optimized

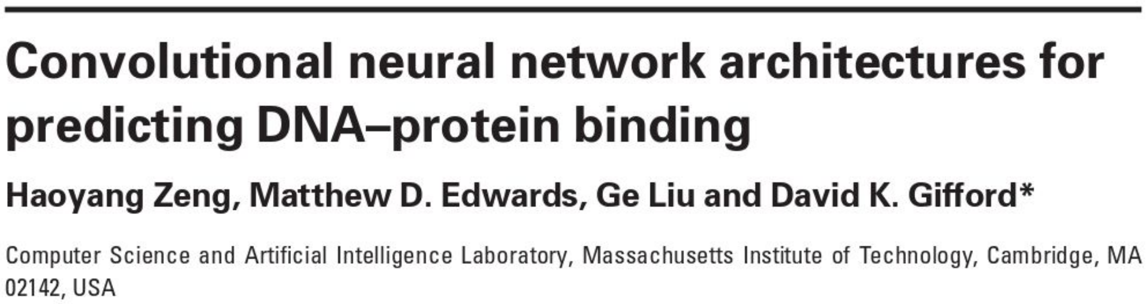
Looking around...

Predicting effects of noncoding variants with deep learning–based sequence model

Jian Zhou[1,2] and **Olga G Troyanskaya**[1,3,4]

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

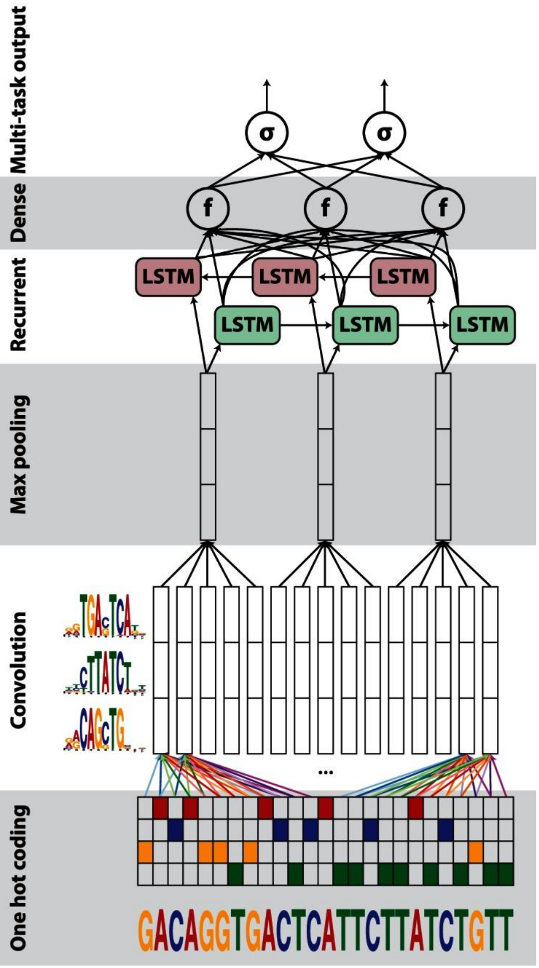
Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]



Convolutional neural network architectures for predicting DNA–protein binding

Haoyang Zeng, Matthew D. Edwards, Ge Liu and David K. Gifford*

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA



Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic acids research, 44(11), e107-e107.

SCELSE

Singapore Centre for Environmental Life Sciences Engineering

MAD

NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

DISCo

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

UNIVERSITA' DEGLI STUDI DI MILANO

BICOCCA

Bibliography

- Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics.
- Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications, 7.
- Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. Scientific reports, 6.
- Břinda, K., Sykulski, M., & Kucherov, G. (2015). Spaced seeds improve k-mer-based metagenomic classification. Bioinformatics, 31(22), 3584-3592.
- Ding, X., Cheng, F., Cao, C., & Sun, X. (2015). DectICO: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. BMC bioinformatics, 16(1), 323.
- Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular systems biology, 12(7), 878.

Bibliography

- Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC genomics, 16(1), 236.
- Marc'Aurelio Ranzato (2014), Deep Learning Tutorial https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxsc3ZydHV0b3JpYWxjdnByMTR8Z3g6Njg5MmZkZTM1MDhhZWNmZA
- Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3), R46.
- Mande, S. S., Mohammed, M. H., & Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. Briefings in bioinformatics, 13(6), 669-681.
- Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., & Sokhansanj, B. (2008). Metagenome Fragment Classification Using N-Mer Frequency Profiles. Advances in bioinformatics, 2008.
- Soueidan, H., & Nikolski, M. (2017). Machine learning for metagenomics: methods and tools. Metagenomics, 1(1).

Bibliography

- Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic acids research, 44(11), e107-e107.
- Lanchantin, J., Singh, R., Lin, Z., & Qi, Y. (2016). Deep motif: Visualizing genomic sequence classifications. arXiv preprint arXiv:1605.01133.
- Leung, M. K. K., Delong, A., & Frey, B. J. (2017). Inference Of The Human Polyadenylation Code. *bioRxiv*, 130591.
- http://colah.github.io/posts/2015-08-Understanding-LSTMs/

THANKS!