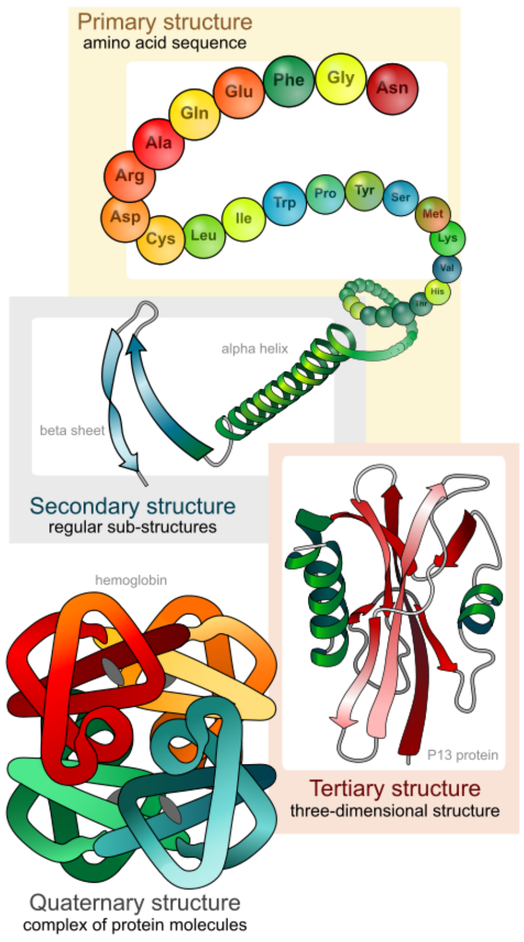
Deep Learning in Bioinformatics

Guillermo Romero Moreno, MSc in Artificial Intelligence, supervised by Mahesan Niranjan

Bioinformatics

Due to the technological advances in imaging and genomics, the field of biology has seen an explosion in the amount of data available. This bulky data, usually high-dimensional, requires a machine learning approach since more simple techniques cannot provide enough insights on the processes [1].

Of special relevance are the processes at the cellular level, since the molecular scale does not allow for direct observation and measurement, while it is possible to obtain some static information about DNA sequences, messenger RNA present at different times, and protein concentration. These three elements form an incredibly complex network (the gene regulatory network). DNA is transcripted into RNA, RNA is translated into proteins and proteins affect the two previous processes at which, leading to varying sorts of loops.



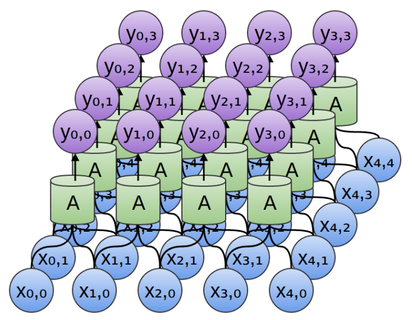
All DNA, RNA, and proteins are sequence sorts of data, so all their properties could be potentially derived exclusively from the sequence information.

Good problems in which to apply them are structural tagging on genomic data (promoter regions, non-regulatory fragments) or property predictions from protein sequences (secondary structure, binding properties) [2].

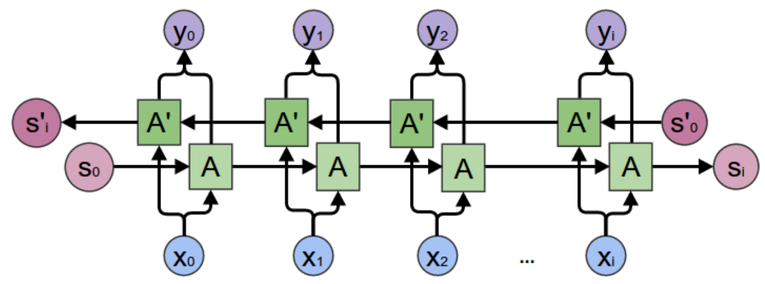
Deep Learning

Shallow machine learning techniques often depend heavily on pre-processing and right feature selection. This sort of information is usually not abundant enough, and requires knowledge from high experts since the processes being studied are very complex. Deep learning allows for an automatic feature creation purely based on the original data, and therefore can bring new insights on the biological processes.

CNNs have proven to be good at detecting features independently from where they are located and allow deeper structures due to the small amounts of weights per layer. These advantages can also be applied to 1D sequence data, as they are able to recognize motifs in any location of the Sequence.



LSTMs are naturally fitted for ordered sequence data, and their long-term memory trait allows them to capture non-local relations in a sequence that would be hard to detect otherwise.



Aims

The main aim of this project is building a machine learning algorithm for cellular sequence data, specifically by using deep learning architectures such as convolutional neural networks (CNNs) and recurrent neural network (LSTMs). Ideally, it should be able to improve a state-of-the-art algorithm by applying any of these techniques.

Steps

1) Review state-of-the-art algorithm and re-implement

2) Run it and analyze results. Spot weaknesses.

3) Think, design, and implement modifications that may overcome the flaws

4) Repeat 2) and 3) while time permits.

Methods

In machine learning applications it is really important to keep the test set aside from the process and only use it at the end for assessing a truthful generalization error. Cross-validation can also be used on the training set (nested cross-validation), for making various model decisions, such as tuning hyperparameters or early stopping. The different models trained in the nested splits can then form an ensemble and achieve higher generalization power.

Many researchers in the biological field also opt using each train and test datasets coming from different a source (after eliminating any existing overlap), thus ensuring the test data is absolutely new. The biological research field provides open access to rich databases [3].

Clean code standards [4] will be followed, in order to keep a self-explicative, easy-to-maintain code. A version control system (*git*) will be used as well.

References

[1] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. Nat Rev Genet, 16(6):321–332, 2015.
[2] W. Jones, K. Alasoo, D. Fishman, L. Parts, Computational biology: deep learning. Emerging Topics in Life Sciences. 1, 257–274 (2017).
[3] Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov. Applications of Deep Learning in Biomedicine, 2016.
[4] R. C. Martin. *Clean code*, Prentice Hall; 1 edition (August 11, 2008). **ISBN-10**: 0132350882
Protein secondary structure image: Jeremy Conn (Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License)
http://www.clearbiology.com/about/
CNN and LSTM images: Cristopher Ollah, https://colah.github.io/posts/2015-09-NN-Types-FP/

Author contact details
Email: grm1g17@soton.ac.uk
ECS department, University of Southampton