Deep and cognitive learning applied to Precision Medicine: the initial experiments linking (epi)genome to phenotypes-disease characteristics.

D. Cirillo[#1], A. Valencia[#2]

[#]*Barcelona Supercomputing Center (BSC), c/Jordi Girona, 29, 08034 Barcelona (Spain)*

Life Sciences Department, Computational Biology Group

[1]davide.cirillo@bsc.es,
[2]alfonso.valencia@bsc.es

Keywords— **Deep learning, Network analysis, Personalized medicine**

EXTENDED ABSTRACT

Introduction

Present-day era of Big Data provides the unique opportunity to develop innovative approaches for data analysis to find new insights into specialized fields of biomedical research such as **Precision Medicine** [1]. Precision Medicine is defined as the integration of molecular research with clinical data in order to deliver better diagnoses and treatments tailored to the individual characteristics of each patient. Advanced analysis of health related data that is specific to a given individual must focus on both clinical information (e.g. clinical reports, medical images, patient histories) and biological data (e.g. gene and protein sequences, functions and pathways). This wealth of information has the potential to inspire systematic ways of making sense from the massive and heterogeneous stream of data and providing a unified view. In the regards, **Deep Learning** (DL) [2] and **Cognitive Computing** (CC) [3] are two branches of Artificial Intelligence (AI) representing convenient choices to tackle the problem of Big Data integration for Precision Medicine. DL comprises several machine learning techniques modeling multiple representations of data through many layers of nonlinear processing units. CC is a cross-disciplinary technology for adaptive and contextual knowledge representation and reasoning through sophisticated analytics aiming to mimic human learning mechanisms.

Objectives

1 - Ontologies of molecular and clinical annotations

A large fraction of biological knowledge is organized in the form of **ontologies**, i.e. sets of domain-specific categories (terms) with relations operating among them. The Gene Ontology (GO) [3] covers three domains: cellular components, biological processes and molecular functions. The Human Phenotype Ontology (HPO) [4] covers the domain of clinical signs or phenotypic anomalies in human diseases. Inspired by automatic language translation, specific Deep Learning approaches like "encoder-decoder" Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) architecture and Attention Mechanism (AM) (Fig. 1) can be used to model the "translation" of one ontology to the other and build a common function-disease reference framework. In the context of Personalized Medicine, this approach can be exerted to understand the relationships between the genetic variants found in a given patient and her/his specific set of disease-associated phenotypes and altered gene functions and pathways.

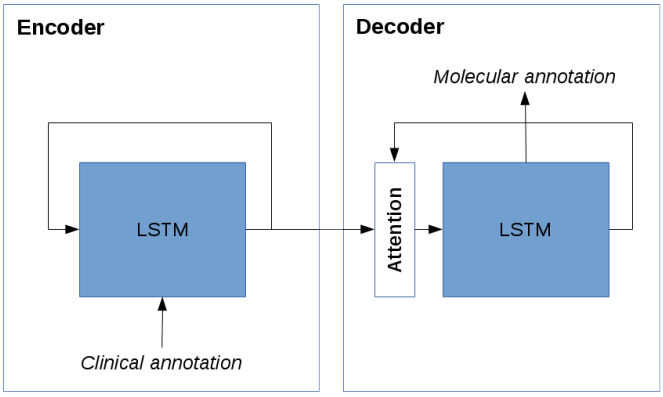

Fig. 1: Schematic view of Enconder-Decoder RNN that takes a Clinical annotation as an input and produces the correponding Molecular annotation as an output.

2 – Genotype-phenotype relationships

A compelling application of RNNs techniques is to uncover of the determinants of **time-dependent biological processes** such as cellular differentiation. Recently, Carrillo de Santa Pau et al. [5] proposed a model to link epigenetic signatures to cell fate during hematopoiesis, i.e. the process of formation of blood cellular components starting from stem cells in the red bone marrow (Fig. 2). Along with changes in the epigenome, additional 'omics' data (e.g. chromatin conformations, expression levels, protein abundances) can be taken into account to model complex time series by means of deep RNNs. In the context of Personalized Medicine, this approach can be applied to the detection of patient specific disease related (epi)genomics modifications affecting cellular

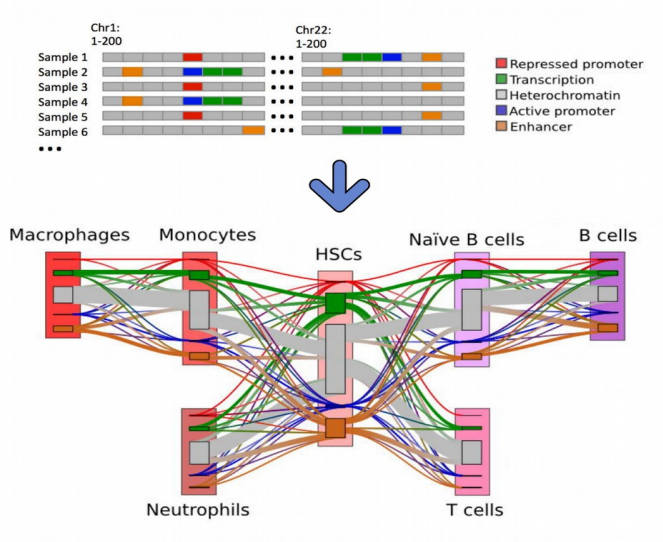

Fig. 2: Genome segmentation in 5 chromatin states across samples (upper panel). Changes of chromatine states between differentiation stages (lower panel; line thickness is proportinal to the number of regions) (adapted from Carillo de Santa Pau 2016 [5]).

differentiation processes and will be potentially applicable to tumor evolution processes.

3 – Cognitive computing with IBM Watson

IBM Watson (http://www.research.ibm.com/cognitive-computing/) [6] is one of the most advanced AI-platform for CC. This technology differs from current computing applications in that it moves beyond pre-configured rules as it reasons based on broad objectives. IBM Watson system is adaptive, interactive, stateful and contextual, meaning that it learns as information changes by asking questions and finding additional inputs based on user's needs at that point in time and from different sources of information (syntax, domain regulations, etc.). As a result, IBM Watson is able to addresses complex situations that are characterized by ambiguity and uncertainty such as question answering (QA) tasks based on large unstructured collections of natural language documents. In collaboration with IBM we are applying Cognitive Computing approaches to the above mentioned scenario relating (epi)genome-phenome/disease information sources.

Discussion

In recent years, many AI techniques emerged that can effectively learn from for very large and complex data sets achieving human-level performances in image and speech recognition as well as natural language processing. Incorporating biomedical knowledge into such a new generation of learning algorithms is the current paradigm in biomedical research. Indeed, the development of algorithms for a comprehensive analysis of health-related Big Data represents one of the main challenges for the Computational Biology community and also a rational aid to experimental scientists and physicians. In particular, DL such as deep RNNs, and CC such as IBM Watson system, offer the possibility to address Personalized Medicine problems prompting functional hypotheses and guiding better diagnoses and treatments. Multiple data sources, namely large-scale clinical and molecular data, combined into integrative models designed to reveal unexpected relationships among biological entities will help to unveil the broader context in which physiological and pathological events are occurring.

References

[1] Ashley EA. *Towards precision medicine*. Nat Rev Genet. 2016 Aug 16;17(9):507-22.
[2] LeCun et al. *Deep learning.* Nature. 2015 May 28;521(7553):436-44.
[3] Ashburner et al. *Gene ontology: tool for the unification of biology* (2000) Nat Genet 25(1):25-9.
[4] Köhler et al. *The Human Phenotype Ontology in 2017* (2017) Nucleic Acids Res 45 (D1): D865-D876.
[5] Carillo de Santa Pau et al.*Searching for the chromatin determinants of human hematopoiesis*. 206. bioRxiv doi: 10.1101/082917.
[6] Chen et al. *IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research.* (2015) Clin Ther. 2016 Apr;38(4):688-701.

Author biography

Davide Cirillo was born in Rome, Italy, in 1985. He received the M.D. degree in Pharamaceutical Biotechnology from University of Rome 'La Sapienza', Italy, in 2011, and the Ph.D. degree in Biomedicine from Universitat Pompeu Fabra (UPF) and Center for Genomic Regulation (CRG), of Barcelona, Spain, in 2016. Since March 2017, he is a Recognized Researcher at computational Biology Group within the Life Sciences Department of Barcelona Supercomputing Center (BSC), Spain. His current research interests include Deep Learning, Network Analysis, Artificial Intelligence, Precision Medicine.