TRANSLATION OF PROTEIN SEQUENCES TO SECONDARY OR TERTIARY STRUCTURES WITH NATURAL LANGUAGE PROCESSING METHODS

PROMOTORS

Prof. dr. Willem Waegeman

Dr. ir. Gerben Menschaert

TUTOR

Jim Clauwaert

MASTER

C&G, C&L, Bioinformatica

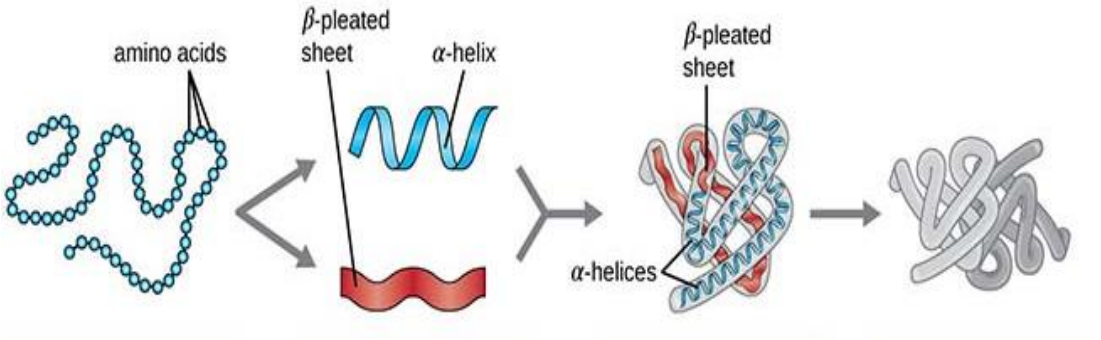
MORE INFO

willem.waegeman@ugent.be

jim.clauwaert@ugent.be

Introduction

The transformer unit is a novel technique that was first described for natural language processing (NLP). The technique adapts attention, an algorithm that focusses only on relevant parts of the input data, and has proven to heavily outperform previous techniques adapting recurrent units. The application of the transformer might prove to be successful in the field of proteomics, in which we will aim to predict a higher level structure from the amino acid sequence of the protein. This ambitious project is two-fold: (1) find or create (using auto-encoders) a vector representation which functions as a lower dimensional embedding for the secondary/tertiary structure of proteins and (2) train a transformer for the 'translation' of amino acid sequences into this embedding.



Aim of the thesis

With the help of an interested and persisting student, this ambitious project aims to gather enough information for publishing, and is therefore perfectly suited for those aiming to lay the groundworks for a future academic position. The thesis is going to utilize applied methods for data integration and machine learning (given as Predictive Modeling during the 1st semester), for which a strong engagement from the start is required.

[1] https://arxiv.org/pdf/1706.03762.pdf