



Contribution ID: 55

Type: **Student Micro-talk**

1-Dimensional Convolutional Neural Network for the Classification of Protein Sequences

Machine Learning (ML) algorithms have been used extensively in the classification and characterization of protein sequences and proteome data. However, the functional annotation of newly identified proteins relies heavily on sequence similarity analyses, i.e. functional annotations of characterized proteins are transferred to novel protein sequences based on amino acid sequence conservation. These methods, such as Hidden Markov models, Blast, C-Blast etc., hold several pitfalls: protein sequences have more and less conserved sub-sequences, and similarities in the conserved sub-sequences have higher relevance than similarities in the less conserved sub-sequences. This information is not taken into consideration in the computational tools mentioned above. Moreover, primary protein structures (the amino acid sequence) are three to ten times less conserved than tertiary protein structures (3D shape of a protein), leading to reduced signal strength for annotation purposes.

Deep Neural Networks (DNNs), are able to circumvent these constraints: they can be trained to identify patterns and classify input data without depending on manually defined features (e.g. information from sequence alignments). These algorithms can capture non-linear dependencies and interaction effects, spanning the wider sequence context. This makes DNNs attractive for the development of robust analyses tools for interrogation of biological sequences. Here, we introduce a 1-dimensional Convolutional Neural Network (1D-CNN) architecture for protein classification and discuss the advantages of using CNNs over other types of neural networks.

Student?

Yes

Supervisor name

Dr Uljana Hesse

Supervisor email

uhesse@uwc.ac.za

Primary author: ESTERHUYSEN, Fanechka

Presenter: ESTERHUYSEN, Fanechka

Session Classification: Micro-talks

Track Classification: Bioinformatics and Biological Sciences