

Using machine learning to accelerate empirical protein design

Enzymes form essential active ingredients for numerous industrial applications, including aiding the release of nutrients during animal feed production, and as degrading agents in machine dish and laundry products. Thanks to stability and activity optimization, these enzymes can drastically reduce material and energy waste.

In situations where the structures of the enzyme and substrate are known, sequence optimization can be tackled by so called quantitative structure activity relationship (QSAR) models, which typically focus on a small number of candidate sites determined by mechanistic understanding of the enzyme. However, in the setting where we know only relatively little about the enzyme and its exact mechanism, one frequently targets positions throughout the entire sequence for modification, resulting in a search-space too vast to explore systematically.

This minimum 6 month project aims to train neural networks on our unique large screening data, to recognize sequence-activity relationships without the use of pre-computed structure-based features or rules. We will attempt to improve performance using hyper-parameter sweeps and data augmentation and compare performance with currently applied methods.

We offer an opportunity to get experience in applied machine learning using unique, large high-quality datasets. You will work together with a team of enthusiastic specialists in data science and protein design and have access to a local compute cluster. Expect to get plenty of experience and make real contributions to an important field of research using data you won't find anywhere else.

Qualifications

We are looking for ambitious MSc students interested in developing their skills in machine learning, bioinformatics and statistics that

- have a background in computer science, machine learning, artificial intelligence, bioinformatics or statistics,
- are proficient in Python and/or R
- have an interest in life science applications such as structural biology or biochemistry
- are available for minimum 6 months for 40 hours per week

The internship is expected to start first half of 2018.

Location and Contact

Genencor International BV, Leiden. Send your CV and motivational letter and/or any questions you may have to Henning Redestig <henning.redestig@dupont.com>.