

Evolutionary-scale interpretation of protein functions in the human gut microbiome

The human gut is home to a vast array of microscopic organisms, which make up our gut microbiota. These bacteria, viruses, and other prokaryotic organisms play a vital role in regulating various physiological processes and influencing human health. Over the past decade, scientists have used large-scale metagenomic sequencing to study the diversity of these organisms across space and time, and how variations in their populations relate to human health and behaviour. This research has uncovered more than 600 million protein-coding genes from nearly 5,000 prokaryotic species. However, only a half of these genes can be understood based on their similarity to other well-studied, leaving many "dark proteins" with unknown functions.

Recently, deep learning has caused a revolution in the field of computational biology. By learning from the wealth of protein information deposited in various protein databases, it has allowed the development of multiple computational tools for the analysis of such proteins. By applying deep learning methods to predict the three-dimensional structures of proteins at unprecedented levels of accuracy, researchers can now study the functions and interactions of proteins that have no known homologs. In the protein sequence space, by treating proteins as strings of words, deep learning models can detect evolutionary relationships that were previously out of reach, allowing for the annotation of remote homologs and orphan genes. To better understand what proteins do functionally, combining sequence and protein 3D structure information within deep learning algorithms, unlocked access to vast functional repertoire encoded in proteins. This is thus the right time to carry out a large-scale analysis of these proteins, combining deep learning based methods for protein structure prediction, the detection of very distance evolutionary relationships and function predictions, beyond the ability of standard approaches.

In this project, we will create an atlas of human gut protein structures annotated with functions and a protein universe map to help us navigate this vast space. We will use deep learning and large-scale evolutionary modeling to better understand the functionally dark proteins of the human gut metagenome. By studying the proteins and their genomic context, we will seek to identify unique proteins and their potential roles in human health and behavior. Our findings will help to prioritize future research and provide a more comprehensive view of the proteins found in the human gut.