Burkhard Rost
TUM Computational Biology

Title:
Artificial Intelligence captures language of life written in proteins.

Abstract:
The objective of our group is to predict aspects of protein function and structure from sequence. The wealth of evolutionary information available through comparing the whole bio-diversity of species makes such an ambitious goal achievable. Our particular niche is the combination of evolutionary information (EI) with machine learning (ML) and artificial intelligence (AI). 30 years ago, the marriage of machine learning and evolutionary information (in the form of Multiple Sequence Alignments) allowed a breakthrough in secondary structure prediction. The same principle has been underlying all state-of-the-art predictions of protein structure and function and is also the root for the program that broke through in protein structure prediction, namely AlphaFold2.

Over the last two years, it has become possible to deep learn the language of life written in proteins through protein Language Models (pLMs). The information extracted is transfer learned to supervise learn protein prediction with annotations. I will present three particular new methods predicting protein structure (1D: secondary structure, membrane regions, & disorder, 2D: inter-residue distances/contacts, 3D: co-ordinates) and protein function (sub-cellular location, binding residues, GO terms), and the effects of sequence variation using pLMs. These embeddings allow for some applications to reach for others to surpass the state-of-the-art without using evolutionary information.

Crucial in all of this is the understanding of the AI and the control of database bias. For both computational biology could serve as a sandbox to prepare more sensitive applications of AI in society.