Flash talks 2024

Highly accurate protein structure prediction with AlphaFold Jumper et al. *Nature* volume 596, pages 583–589 (2021)

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort1,2,3,4, the structures of around 100,000 unique proteins have been determined5, but this represents a small fraction of the billions of known protein sequences6,7. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'8—has been an important open research problem for more than 50 years9. Despite recent progress10,11,12,13,14, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)15, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

Link to PDF

Team: Jüli Kosar and Reem Majjani

Informed and automated k-mer size selection for genome assembly. Chikhi et al. Bioinformatics 2014, 30(1):31-7

Motivation: Genome assembly tools based on the de Bruijn graph framework rely on a parameter k, which represents a trade-off be- tween several competing effects that are difficult to quantify. There is currently a lack of tools that would automatically estimate the best k to use and/or quickly generate histograms of k-mer abundances that would allow the user to make an informed decision. Results: We develop a fast and accurate sampling method that con- structs approximate abundance histograms with several orders of magnitude performance improvement over traditional methods. We then present a fast heuristic that uses the generated abundance histo- grams for putative k values to estimate the best possible value of k. We test the effectiveness of our tool using diverse sequencing data- sets and find that its choice of k leads to some of the best assemblies. Availability: Our tool KMERGENIE is freely available at:

http://kmergenie.bx.psu.edu.

Link to PDF

Team: Tim Stadager and Lucie Marie Hasse