Protein structure prediction with machine learning

Shuichiro Makigaki and Dr Takashi Ishida, from the Department of Computer Science at Tokyo Institute of Technology, are developing a new sequence alignment generation model that employs machine learning and dynamic programming to predict protein structures. This novel methodology can also be applied to homology detection which is fundamental to bioinformatics.

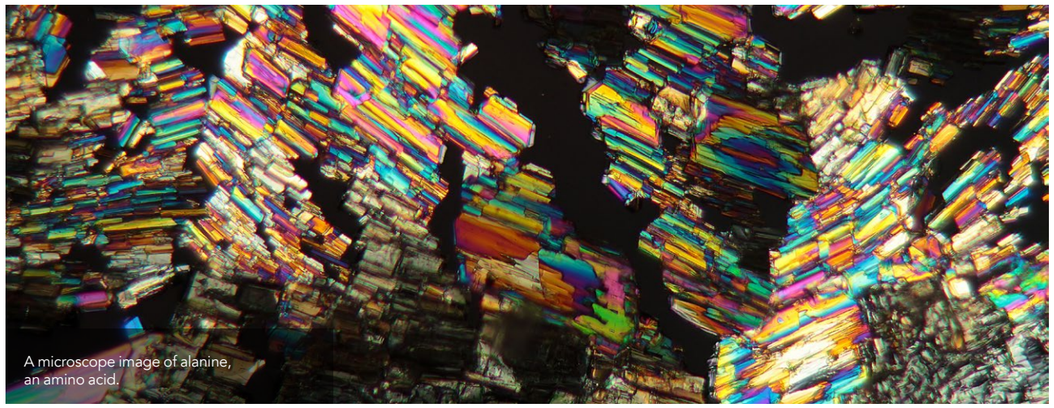
A protein's function is dictated by its three-dimensional structure. If the structure is known, then the protein's function can be predicted based on its structural similarity to other known proteins. It is also possible to predict which molecules or drugs can bind to the protein and how they will bind. The ability to make informed predictions of this kind is of particular importance in the design of new drugs.

Determining a protein's structure by experimentation, however, is an expensive, time consuming and difficult process. Computational modelling of protein structures is therefore often employed to ease the prediction process. Shuichiro Makigaki and Dr Takashi Ishida, from the Department of Computer Science at Tokyo Institute of Technology, are developing a new sequence alignment generation method that employs both machine learning and dynamic programming to accurately predict protein structures.

GETTING TO KNOW PROTEINS

Proteins are essential components of biology, biochemistry and pharmaceutical sciences. They take part in essentially every cellular process and are vital for the correct functioning of living organisms. After water, proteins make up the second-largest component in human tissues.

Some proteins have specific mechanical or structural functions. Enzymes, the catalysts for biochemical reactions, are proteins that are crucial to our metabolism. Proteins are also involved in the communication process governing basic cell activities and directing multiple-cell actions known as cell signalling. They are also fundamental to our immune response in the form of antibodies, large Y-shaped protein molecules employed to neutralise pathogens including viruses and pathogenic bacteria. Proteins are also important to both cell adhesion, where cells interact and attach themselves to neighbouring cells, and the cell cycle, the process that enables the duplication of a cell's DNA and the production of two daughter cells.



A microscope image of alanine, an amino acid.

PROTEIN STRUCTURES

Protein molecules are made from long chains of amino acids. These sequences of amino acids make up a polymer – the polypeptide – a long, continuous chain of amino acids linked by peptide bonds. But the structure of a protein is a unique three-dimensional shape, more complex than the polypeptide.

Protein structure can be described using four distinct structural features:

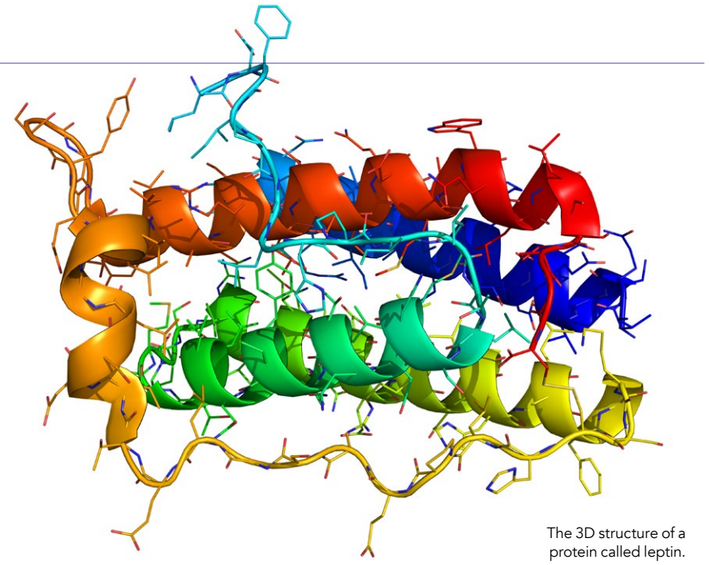
- A protein's primary structure denotes the sequence of amino acids making up the polypeptide chain.
- The secondary structure describes how small sections of the polypeptide chain form into highly regular shapes. There are two main types of secondary structure: the α-helix (a bit like a coiled spring) and the β-strand or β-sheets (more like a concertina or a pleat).
- The tertiary structure refers to the overall shape of an individual protein molecule. The polypeptide chain folds up into a compact globular structure.
- The quaternary structure is formed when several protein molecules form a protein subunit that can function as a single protein complex. Think of several 3D blocks fitting together to make a larger shape.

Experimental techniques such as X-ray crystallography, NMR spectroscopy, and dual polarisation interferometry can be used to determine the structure of proteins. The resulting protein structures are usually submitted to the online Protein Data Bank and are freely accessible on the internet. This means researchers like Makigaki and Ishida can reference these known structures for their own work.

PROTEIN STRUCTURE PREDICTION

Identifying the sequence of amino acids in the chain (the primary structure) is not enough to distinguish the protein's overall structure. However, given a sequence of amino acids, mathematical modelling techniques can be used to generate a model of the protein's tertiary structure to provide a comprehensive protein structure prediction.

Several of these protein structure prediction techniques have been developed to date. Some of them can be categorised as physicochemical (or de novo) simulations which are template-free modelling methods. These use just the sequence of amino acids to generate predictions and require vast amounts of computational power.



The 3D structure of a protein called leptin.

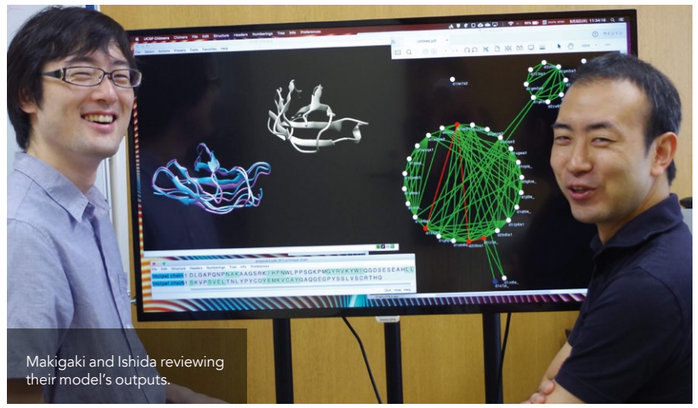
The protein's function can be predicted based on its structural similarity to other known proteins.

Other techniques reduce the computational power requirement by drawing on the biological fact that when two sequences are highly similar, their structures are also similar. Proteins with similar structures are known as 'homologs' and they are often evolutionarily related to one another. Techniques which use the homologs can be classified as template-based or homology modelling.

Template-based modelling starts with the amino acid sequence of the protein undergoing investigation and then identifies proteins with similar sequences whose structures are known – the homologs. The identified homologs are used as structural templates to build three-dimensional models that form predictions of the target protein's structure. Current research shows that template-based modelling methods produce more accurate results than de novo simulations provided suitable templates and protein sequence alignments are available.

A NEW TEMPLATE-BASED MODEL

Makigaki and Ishida are developing a new template-based model using



Makigaki and Ishida reviewing their model's outputs.

a machine learning model. Machine learning is an area of artificial intelligence established on the concept that systems can learn from data, identifying patterns and making decisions with little or no human intervention. This data analysis technique enables the automation of analytical model building.

SEQUENCE ALIGNMENT

The researchers' review of previous methods revealed that sequence profiles, using multiple sequence alignments, could accurately detect homology. They therefore chose to use a pairwise sequence alignment generation method. Sequence alignment involves arranging a protein's chain or sequence of amino acids as rows in a matrix with gaps inserted to allow similar or identical components to be aligned in successive columns. In essence, they break the amino acid sequence up into 'bitesize chunks' for the model. This facilitates the identification of similar or identical sections which can be the result of structural or evolutional relationships between the sequences.

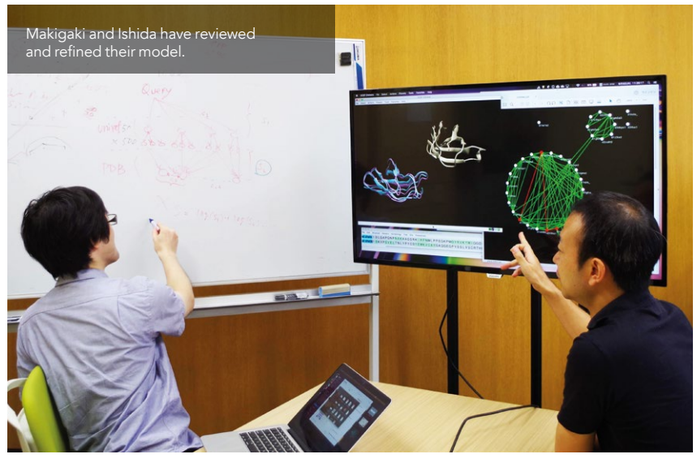
The team from Tokyo Institute of Technology then deploy machine learning to recognise the structural alignments of known homologs in their model. Structural alignment is a form of sequence alignment that attempts to ascertain equivalences between two or more polymer structures by comparing their shape and three-dimensional conformation. They also use dynamic programming, a problem-solving technique that breaks down a large problem into smaller and smaller sub-problems. The small problems are solved and an optimal solution to the large problem is discovered using the optimal solutions to the smaller problems.

FOCUSING ON ALIGNMENT GENERATION

The model needs information about known structurally similar proteins in order to create structural alignments. It is trained by carrying out pairwise structural alignments of pairs of proteins that are known to be structurally similar using a training dataset taken from the Structural Classification of Proteins.

Unlike other methods that integrate sequence alignment generation with homology detection, the researchers' methodology focuses only on alignment generation. Their input comprises a query and a subject. The query is the target amino acid sequence, and the subject is a known amino acid sequence, chosen as the template using homology detection. The model then aligns their sequences using the Smith–Waterman algorithm. This carries out local sequence alignment. Rather than comparing the entire sequence at once, the Smith–Waterman algorithm determines similar regions between two strings of protein sequences by comparing all possible segments of varying lengths in order to optimise the substitution score. The substitution score is a similarity measure that gives higher scores to similar characters and low or negative scores to dissimilar characters. The model outputs an alignment score and a sequence alignment that is now suitable for homology modelling. This technique is known as re-alignment.

The model also uses the k-Nearest Neighbour classification model to dynamically predict a substitution score. The k-Nearest Neighbour algorithm is a non-parametric method used in pattern recognition that is both simple and powerful. Having evaluated this method using a carefully split training and test dataset, the researchers were able to favourably compare the accuracy of their predicted structure models with those generated by state-of-the-art methods in order to measure their sequence alignment quality.



Makigaki and Ishida have reviewed and refined their model.

When the researchers evaluated the accuracy of this new alignment generation method, they found that it outperformed the state-of-the-art methods.

HIGH PERFORMANCE

Homology detection plays a fundamental role in bioinformatics and the study of protein structures in particular. The researchers' novel methodology can also be applied to homology detection by sorting the alignment scores incorporated in its result. The model's ability to detect remote homologs (i.e., pairs of proteins with similar structures and functions but whose sequence similarity is difficult to detect) has been prioritised because models cannot be generated without a template. Encouragingly, when the researchers evaluated the accuracy of this new alignment generation method, they found that it outperformed existing state-of-the-art methods.

Makigaki and Ishida have observed that their model's execution time is longer than desired. This is due to the k-Nearest Neighbour algorithm and the dataset size. They mention that a natural extension of this research is to employ faster k-Nearest Neighbour algorithms. They are also considering the use of higher-performance techniques such as convolutional neural networks.

Behind the Research

Shuichiro Makigaki

Takashi Ishida

E: ishida@c.titech.ac.jp **E:** makigaki@cb.cs.titech.ac.jp **W:** http://www.cb.cs.titech.ac.jp/

Research Objectives

Shuichiro Makigaki and Dr Takashi Ishida have developed a new model for protein structure prediction.

Detail

Shuichiro Makigaki and Dr Takashi Ishida
Department of Computer Science
School of Computing
Tokyo Institute of Technology
Meguro-ku
Tokyo 152-8550
Japan

Bio

Shuichiro Makigaki

After studying computer science and bioinformatics at Kyoto University, Makigaki entered the Tokyo Institute of Technology as a doctoral course student. Alongside research activity in the university, he works as a server-side architect and a researcher in some technology companies in Japan. Makigaki is interested in many topics over a wide range of computer science, but keeping writing codes is the motto.

Takashi Ishida

Ishida is Associate Professor in the Department of Computer Science at Tokyo Institute of Technology. He received a PhD from the University of Tokyo in 2006 and joined Tokyo Institute of Technology in 2009. His current research projects include the development of algorithms for protein tertiary structure prediction and computational drug design.

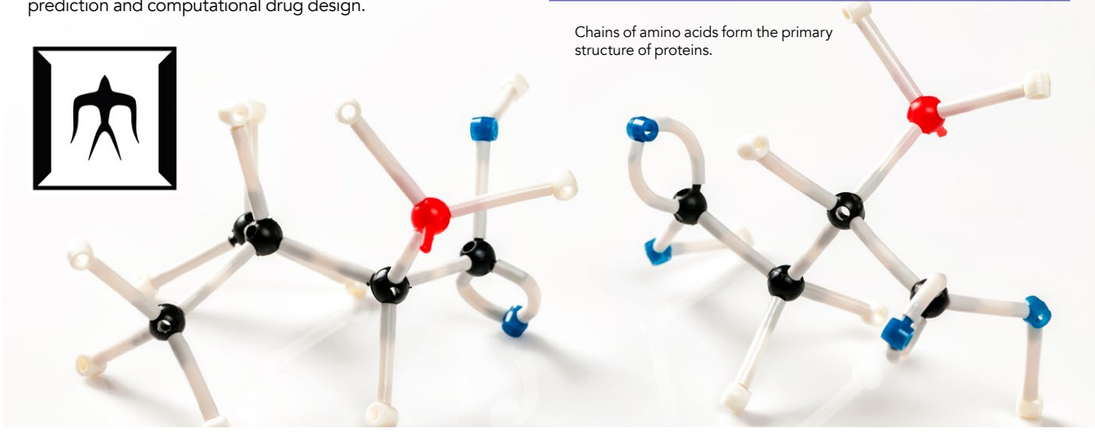
References

Makigaki, S. & Ishida, T. (2019). Sequence alignment using machine learning for accurate template-based protein structure prediction. *Bioinformatics*, [online] Available at: https://doi.org/10.1093/bioinformatics/btz483 [Accessed 30th September 2019].

Personal Response

What are your plans for the next phase in developing your protein structure prediction model?

"The proposed method often takes much execution time to generate accurate results. Instead of implementing faster k-Nearest Neighbour calculations, we plan to employ fast and accurate score prediction methods, such as a neural network. Also, we are developing a highly sensitive homology detection method because template-based modelling requires homologous proteins used as templates. We will develop a comprehensive modelling system by combining the detection tool with the alignment method presented here."



Chains of amino acids form the primary structure of proteins.