**Nobel Prize in Chemistry**

**Design and prediction of protein structures**

**TV Venkateswaran, IISER, Mohali**

This year's Nobel Prize in Chemistry has been awarded for 'computational protein design' and 'protein structure prediction'. It has been shared by three brilliant minds: **David Baker** of the University of Washington, Seattle, USA, along with **Demis Hassabis and John M. Jumper**, both of Google DeepMind, London, UK. Their groundbreaking work, which involves the use of Artificial Intelligence (AI), has significantly advanced our ability to use protein structures for various purposes. For more on AI and the associated Nobel Prize, which was also awarded this year, see the article in this issue of ***JM***.

The key phrase is ‘protein structure’. It turns out that the structure of proteins is a very important aspect for their biochemical function. That is, the structure of the proteins, from their primary sequence to their complex three-dimensional forms, is the key to their diverse functions in our bodies.

**Proteins - a quick recap**

In our daily lives, 'protein' is often viewed as an essential nutrient found in pulses, eggs, and meat. However, the truth is that our bodies are a complex network of proteins. **Enzymes** like *pepsin* and **hormones** like *insulin* and *haemoglobin* in the blood are all proteins. They are the unsung heroes in our cells, carrying out most of the work and playing a crucial role in building, functioning, and regulating the body's tissues and organs.

Just as a wall is constructed by stacking bricks and binding it with a mortar, proteins are made up of small molecules known as **amino acids**. With 26 letters of the English alphabet, by various combinations, we are able to make the 2,17,000 distinct words found in the Oxford English Dictionary. Similarly by numerous permutations and combinations, many distinct proteins are formed out of just 20 different amino acids.

The instructions for making a protein chain or a protein sequence are coded in the genome in the form of DNA or RNA. Like one prepares the garland by joining several flowers in a specific order, the chain of amino acids in a different arrangement produces diverse proteins. Proteins are often called *macromolecules*, which distinguishes them from smaller molecules, such as glucose, water, etc.

Some proteins are made of a low number of amino acids; for example, *insulin* contains just 51 amino acids in its chain. In contrast, *Titin1*, also known as connectin, the biggest known protein, contains as many as 27,000 to 35,000 amino acids.

**Structure of proteins**

Just as you join the pearls in a string, some proteins are made up of only one chain of amino acids and are called *monomers*. Others may have more than one amino acid chain , that is, several monomers and they are called *oligomers*. *Myoglobin*, with 153 amino acids, is a typical example of a monomer with a single amino acid chain. It is used to carry store and transport oxygen in muscles. At the same time, haemoglobin is a globular protein composed of four chains of amino acids that come together to form a complex. It is used to carry oxygen from the lungs to the blood and tissues. The four chains are called as 'sub-units' because the molecule of haemoglobin is the functional unit. **BOX** **Myoglobin**

Myoglobin is an iron- and oxygen-binding protein found in the cardiac and skeletal muscle tissue of vertebrates in almost all mammals. Myoglobin is distantly related to hemoglobin. This protein was the first to have its structure solved by X-ray crystallography by Cambridge researchers **John Kendrew** and **Max Perutz**, for which they received the Nobel Prize in Chemistry in 1962. Since then, researchers have meticulously determined the structure of approximately 200,000 proteins.The picture shows a representation of the 3D structure of the protein myoglobin with turquoise coloured α-helices. Toward the right-center among the coils, is a prosthetic (non-amino acid) group called a *heme* group (shown in gray) with a bound oxygen molecule (red).

**END OF BOX**

**Protein Folding**

If you leave a piece of paper in the open for a few days, it will naturally curl. Similarly, an amino acid chain's sequence *folds* into a repeating pattern. *Alpha helices* and *beta sheets* are the **secondary** **structures** that are most prevalent stable folding patterns in proteins. Needless to say, the primary structure is the amino acid chain(s). Alpha helices resemble coiled crepe paper streamers (the turquoise coloured parts in the picture of myoglobin), while beta sheets resemble saree pleats. One section of the amino acid sequence folds spontaneously into an alpha helix, while another forms a pleated sheet. Thus, a protein may contain many helices, sheets, and other patterns.

After the amino acid chain folds into a repeating pattern, it further folds into a compact three-dimensional form known as the **tertiary structure** of the protein. This structure is crucial for the protein's function. For instance, myoglobin (see Box) first folds into eight alpha helices linked by loops. This unique fold creates a placeholder for the iron that binds the oxygen, allowing myoglobin to perform its oxygen storage function effectively.

For monomers, the tertiary structure is the final form. However, if there is more than one monomer in the protein, then there is a **quaternary structure**. An example of haemoglobin is shown in the picture. Another example is insulin, which comprises two subunits: A, which has 21 amino acids and two tiny alpha helices, and B, which contains 30 amino acids and one alpha helix. The two monomers have to combine; only then will insulin function. The entire structure is a tiny spherical shape like an oil drop in water.

**Structure and function**

Let us start with simple water: one oxygen and two hydrogen atoms - H2O. Hydrogen atoms share electrons with an oxygen atom in a **covalent** bond, but the sharing is not equal. The oxygen atom is more *electronegative* than the hydrogen atoms, meaning it attracts electrons more strongly. This results in the oxygen atom having a partial negative charge and the hydrogen atoms having a partial positive charge. The shared and unshared electron pairs repel each other, causing them to be 104.5° apart and giving the structure a peculiar Mickey Mouse form. Now, slightly negative oxygen may attract hydrogen from another water molecule. In contrast, positive hydrogen can attract oxygen from two separate water molecules. Thus, water molecules in a container create a weak bond with surrounding molecules, giving water its characteristic features.

The protein's three-dimensional tertiary structure also gives particular features. For example, specific folds form a placeholder for the iron that binds the oxygen in the middle of globular *myoglobin*, allowing it to store and release oxygen. In contrast, *collagen* is a lengthy protein comprising many amino acid chains twisted together like a rope or cable. The rope-like stiff characteristic of collagen strengthens the tendons and ligaments that link bones and muscles. *Porin* protein, conversely, is mostly beta pleats and is like a cylindrical tube with open ends, allowing it to function as a channel for a tiny chemical to diffuse across a cell membrane.

**Knowing the shape**

Knowing whether the end of the pencil is sharp or blunt allows us to predict its behaviour. Similarly, the structure and shape of the protein fold reveal information about how it will function. If the structure of the protein is altered it can misfold or misfunction and lead to disease. For eaxample, changing the sixth amino acid in the beta chain of the beta-hemoglobin gene from normal *glutamic acid* to *valine* causes significant alterations in the protein's behaviour. Healthy red blood cells are typically round (laddu)-shaped, but aberrant ones formed by mutant proteins are *sickle-shaped.*

Scientists have been keen to know how proteins fold and form specific shapes. Various experimental approaches, including X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, and cryo-electron microscopy, are employed to identify the structure. Yet, they take anywhere from a day to a year and are expensive.

**AI and protein structure**

By the 1990s, artificial intelligence (AI) had succeeded in image identification and facial recognition. For more details, see the article on Nobel Prize in Physics in this issue. Researchers attempted to apply AI methods to identify patterns in protein folding and underlying amino acid sequences. Enter our Nobel prize winners!

**Demis Hassabis** and **John Jumper** used these AI techniques to build a neural network-based AI model called **AlphaFold**. AlphaFold is a multicomponent artificial intelligence (AI) system that uses machine learning to predict a protein's 3D structure based on its primary amino acids. Initially, their prediction accuracy was not that good; however, the much-upgraded version **AlphaFold2**, published in 2020, was a massive success for *monomeric proteins*. AlphaFold2's accuracy was virtually identical to X-ray crystallography structures.

When Demis Hassabis and John Jumper had confirmed that AlphaFold2 really worked, they calculated the structure of all human proteins. Then they predicted the structure of virtually all the 200 million proteins that researchers have so far discovered when mapping Earth’s organisms.

**Google DeepMind**, where the two scientists work, has also made the code for AlphaFold2 publicly available, and anyone can access it. Recently, the even newer version, AlphaFold3, is also available freely. The AI model has become a gold mine for researchers. By October 2024, AlphaFold2 had been used by more than two million people from 190 countries. Previously, it often took years to obtain a protein structure, if at all. Now it can be done in a few minutes. The AI model is not perfect, but it estimates the correctness of the structure it has produced, so researchers know how reliable the prediction is.

**David Baker** also began experimenting with using AI to predict 3D structures. He developed an AI model named *Rosetta*. However, he thought, instead of trying to find the protein structure from the amino acid sequence, why not find the amino acid sequence for a given structure? He and his team modified Rosetta.

When a structure was provided, Rosetta first recognised its constituent parts before searching a database of all known proteins for a small piece that matched the detected components. The program then determined the amino acid sequences for that component. As a result, the amino acid sequences for the appropriate protein structure were found in stages.

Once the amino acid sequences were found for the designer 3D structure, his team synthesised the DNA sequence for the required amino acid sequence. They put this DNA into bacteria using genetic engineering and were able to produce the resulting protein! Thus, Baker demonstrated that functional designer proteins, or proteins with desired features, may be created.

In 2003, David Baker succeeded in using amino acids to design a new protein that was unlike any other protein. His team utilised the age-old technique of X-ray crystallography to determine if the protein structure they had produced was identical to the predicted one. They found that the structure determined was nearly similar to the protein they called *Top7* which they initially designed. Since then, his research group has produced one imaginative protein creation after another, including proteins that can be used as pharmaceuticals, vaccines, nanomaterials and tiny sensors.These designer proteins are known as '*de novo proteins*' as they do not exist in nature and are created using the structure as a guide.

**The road ahead**

Proteins’ amazing versatility as chemical tools is reflected in the vast diversity of life. That we can now so easily visualise the structure of these small molecular machines is mind boggling; it allows us to better understand how life functions, including why some diseases develop, how antibiotic resistance occurs or why some microbes can decompose plastic.

The ability to create proteins that are loaded with new functions is just as astounding. This can lead to new nanomaterials, targeted pharmaceuticals, more rapid development of vaccines, minimal sensors and a greener chemical industry – to name just a few applications that are for the greatest benefit of humankind.

***Sources: Several, including www.nobelprize.org***