Protein Structure Complexity (AlphaFold)

Proteins are large, complex molecules that play critical roles in nearly all biological processes. They are composed of chains of amino acids, which are the building blocks of proteins.

Composition of Proteins

Proteins are primarily made up of carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and sometimes sulfur (S). These elements combine to form amino acids, linked by **peptide bonds** to form polypeptide chains. A typical protein may consist of hundreds to thousands of amino acids.

Each amino acid has:

- An **amino group (-NH$_2$)**
- A **carboxyl group (-COOH)**
- A **side chain (R-group)** unique to each amino acid determines its properties (e.g., hydrophobic, hydrophilic, acidic, or basic).
- Polypeptide > 10 amino acids
- Protein = 300 - 1,000 amino acids

Protein structure:

1. Primary Structure: The primary structure is the linear sequence of amino acids in a polypeptide chain.

2. Secondary Structure: The secondary structure refers to local folding of the polypeptide chain into regular structures.
3. Tertiary Structure: The tertiary structure is the overall three-dimensional shape of a single polypeptide chain, formed by further folding of the secondary structures.
4. Quaternary Structure: The quaternary structure refers to the arrangement and interaction of multiple polypeptide chains (subunits) to form a functional protein complex.

Types of Proteins:

- **Enzymes** are proteins that act as biological catalysts, speeding up chemical reactions in cells without being consumed.

- **Transcription factors** are proteins that regulate gene expression by binding to specific DNA sequences. They control the transfer of genetic information from DNA to messenger RNA (mRNA), which is the first step in producing proteins from genes.

Here are some examples of enzymes and transcription factors that are widely used in research.

- Enzyme: **Hexokinase** is a key enzyme in the first step of glycolysis, where it phosphorylates glucose to form glucose-6-phosphate. This is a crucial step in cellular metabolism, enabling cells to harvest energy from glucose. particularly in studies related to metabolism, diabetes, and cancer.
- Transcription Factor: **p53** (Tumor Protein 53) is a transcription factor that regulates the expression of various genes involved in DNA repair, cell cycle arrest, apoptosis, and senescence. It acts as a tumor suppressor and is often called the "guardian of the genome." Its role in regulating cell fate makes it critical for understanding cellular responses to stress, particularly in cancer research.

Source of 3D Protein Structure:

AlphaFold is an AI system developed by Google DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiments.

For Our Project Search:

- **Hexokinase (Human) UniProt ID:** P52789 (Hexokinase 1, one of the common human isoforms, species: Homo sapiens)
- **p53 (Human) UniProt ID:** P04637

Fractal Dimension

The **fractal dimension** is a mathematical concept that quantifies the complexity of shapes or patterns in space. It extends the notion of dimensionality from whole numbers (like 1D, 2D, and 3D) to non-integer values, capturing how detail in a pattern changes with scale.

Interpretation:

- **D = 1**: represents a line; the structure is simple and linear.
- **D = 2**: represents a surface; the structure has more complexity.
- **D = 3**: Represents a volume; the structure is fully three-dimensional and may fill space.
- **1 < D < 2**: Indicates a complex, but not fully space-filling structure, common in many biological macromolecules like proteins.

A higher fractal dimension suggests a more complex arrangement of atoms, potentially indicating functional or structural significance.

Data Exploratory Report: Analyzing Protein Complexity through Fractal Dimensions

Introduction

The goal of this project is to explore the structural complexity of proteins by calculating their fractal dimensions. Proteins' complexity can be linked to their structure and function, and fractal dimension provides a mathematical way to quantify how detailed a structure is. We use tools like `requests` to fetch protein structures from **AlphaFold**, visualize them with **py3Dmol**, and calculate their fractal dimensions using a Python implementation.

Libraries Used

We used the following libraries in our project:

- `requests`: To fetch protein structures from the AlphaFold Protein Structure Database.
- `py3Dmol`: For 3D visualization of protein structures.
- `Bio.PDB`: To parse and analyze PDB files.
- `matplotlib`: For plotting and visualizing results.
- `numpy` and `scipy.stats`: For handling data and calculating fractal dimensions.

Data Collection and Visualization

1. The protein used in our research is the hexokinase enzyme uniprot_id = `P52789`

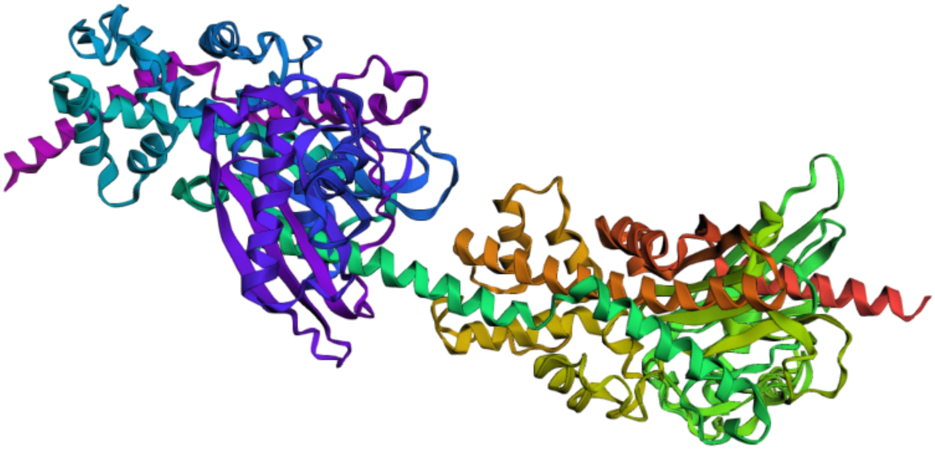

Fig. 1: Hexokinase Enzyme

2. The second protein used in our research is the p53 transcription factor.

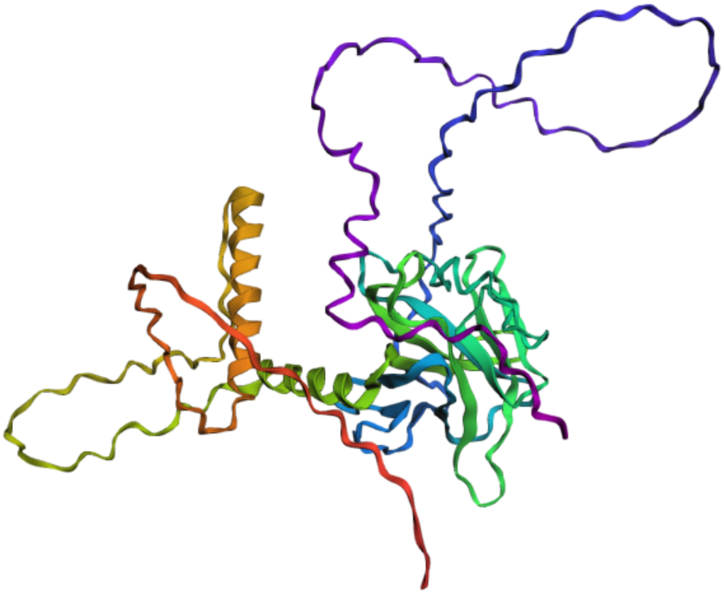

fig.2::p53 transcription factor

Feedback: As you can see from the p53 structure above from the alphaFold protein database, machine learning still requires a lot of training. That is why we used the Protein Data Bank Database experimentally measured structure to get a more accurate p53 structure.

Data Acquisition: Retrieving Protein Structures Using BioPython

BioPython allows you to fetch and parse structures from PDB. We will need to retrieve structures for both enzymes and transcription factors.

We only use this database for the **p53** transcription factor because the alphafold protein accurately predicted the **hexokinase enzyme** .

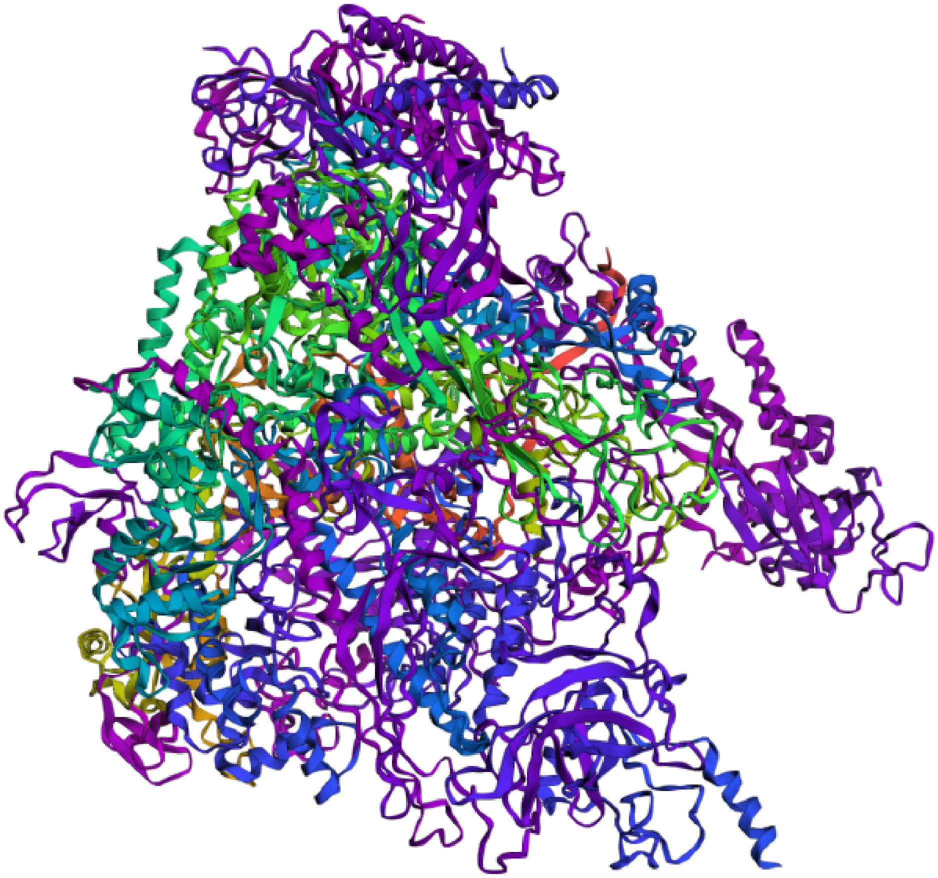

Fig.3: p53 Transcription factor from PDB

Fractal Dimension Calculation

Fractal dimensions give us a quantitative way to assess the complexity of a protein's structure. The higher the fractal dimension, the more complex the structure is, suggesting a more detailed or irregular structure. In our Python implementation of the **box-counting algorithm,** we used to calculate the fractal dimension.

Results:

Hexokinase enzyme fractal dimension

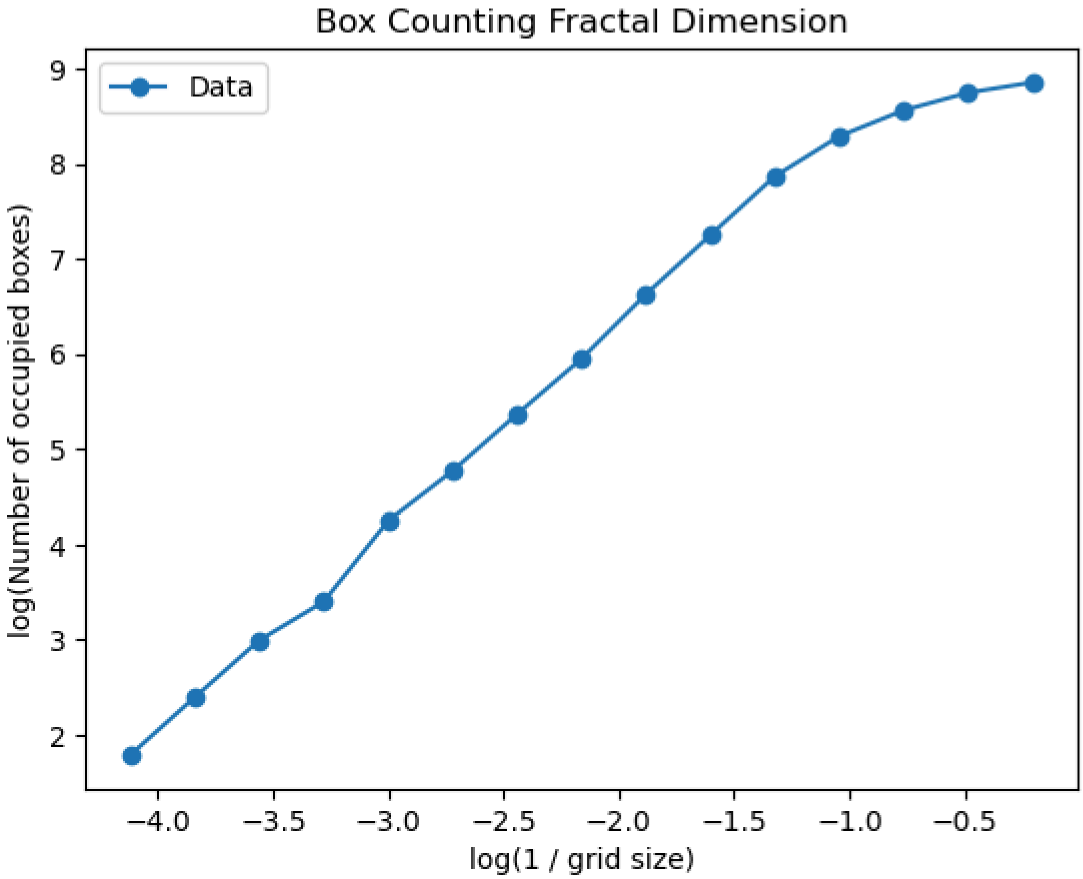

Fig. 4: Hexokinase enzyme fractal dimension

```
Estimated Fractal Dimension: 1.9426
Fractal Dimension of the protein in hexokinase.pdb: 1.9426
```

The **estimated fractal dimension of 1.9426** indicates that the 3D structure of hexokinase exhibits a moderately complex spatial arrangement. Since fractal dimensions range from 1 to 3, a value close to 2 suggests that the protein's structure is not too simple (like a straight line). nor maximally complex (like a space-filling object), but it shows a more intricate, branched, or folded pattern.

This type of fractal dimension is typical for biological macromolecules, reflecting the complex folding and interactions within the protein. If you want to compare it with other proteins or further explore how structural complexity relates to functionality, this method provides a robust way to analyze such properties.

`p53 transcription factor` fractal dimension

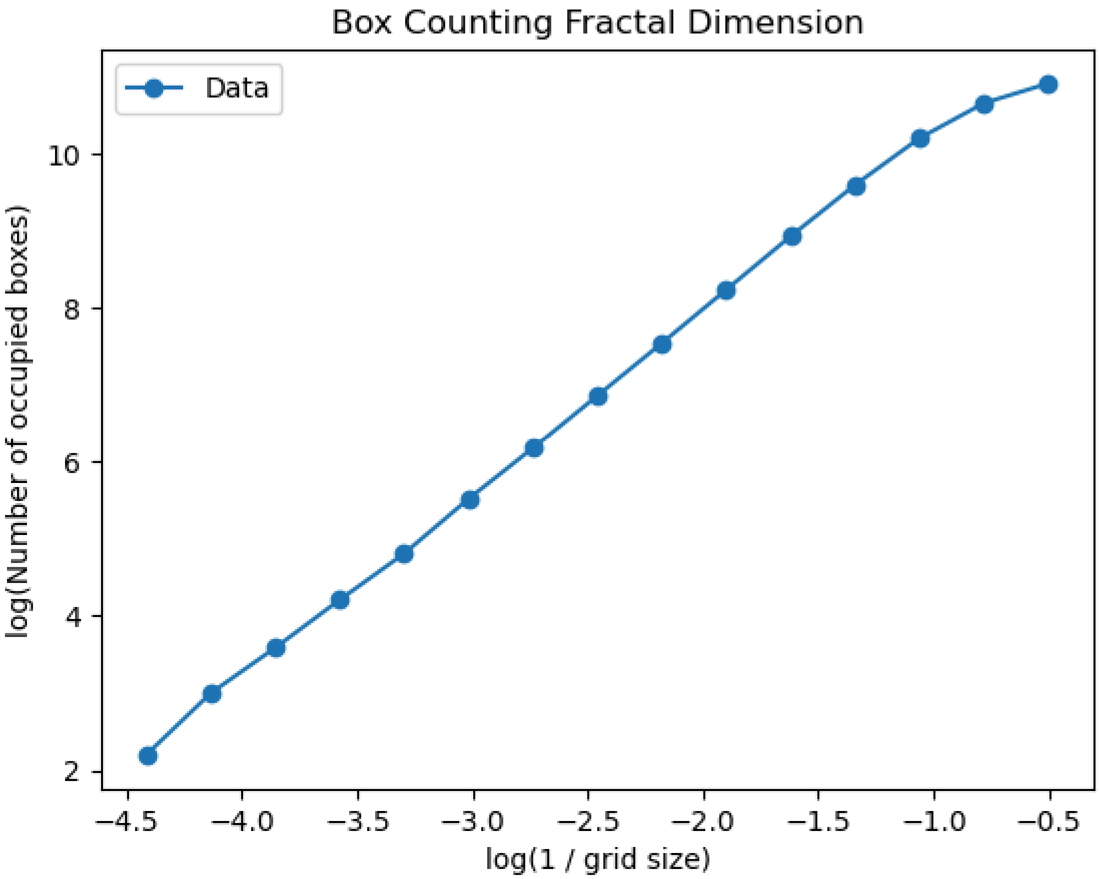

Fig 5: p53 TF Box Counting Fractal Dimension

Estimated Fractal Dimension: 2.3085
Fractal Dimension of the Protein in p53.pdb: 2.3085

The estimated fractal dimension of 2.3085 indicates that the p53 protein structure is more complex than a simple 2D surface but does not fully occupy 3D space uniformly. The protein likely has a rich, irregular structure with significant detail.

More Protein Examples:

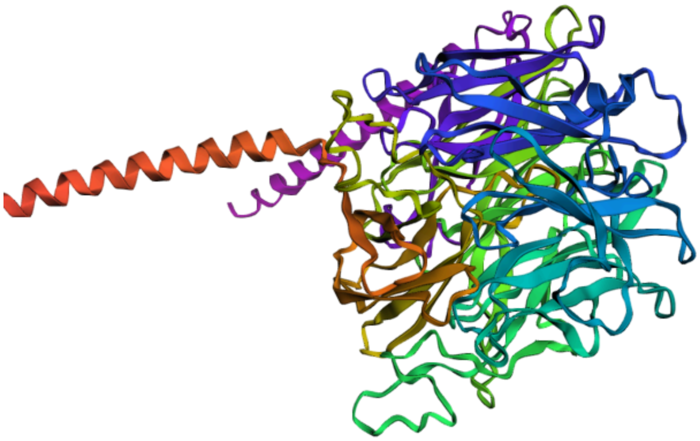

Fig. 6: T-cell immunomodulatory protein homolog Protein structure

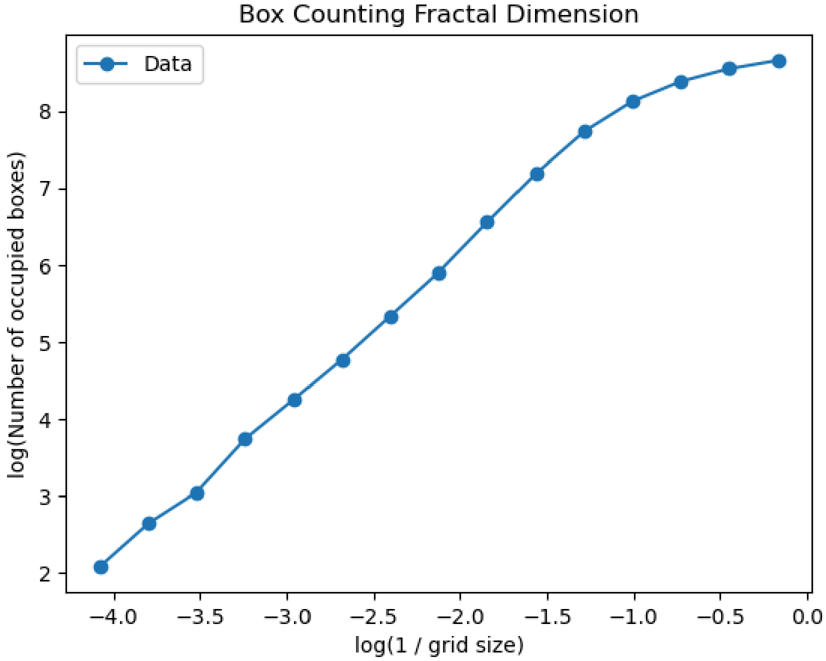

Fig. 7: T-cell immunomodulatory protein homolog fractal dimension

```
Estimated Fractal Dimension: 1.8196
```

Comparison with Other Proteins:

- If you compare this fractal dimension to other proteins, such as **hexokinase** (fractal dimension ~1.94), p53 has a **higher fractal dimension**, suggesting that it might be structurally more complex. which could be necessary for p53's multifunctional roles in DNA repair, cell cycle regulation, and apoptosis.
- **Functional Relevance:** Understanding the fractal dimension allows you to hypothesize that proteins with higher fractal dimensions are involved in more complex or variable biological processes.
- **Comparative Analysis:** By comparing the fractal dimensions of different proteins, you can identify patterns in which regulatory proteins (e.g., transcription factors like p53) have higher fractal dimensions than enzymes like hexokinase, which may have simpler, more defined structures.

The Complexityeqn.py script leverages existing Chou Fasman data to predict the amount of alpha helices, beta sheets and coils to give us a preliminary structure. Then, the structure is “smoothed” by taking the most repeated type of structure in a predetermined window for accuracy purposes and to avoid false positives. The Shannon entropy is calculated to give us a good idea of the complexity of the structure.
For higher accuracy, we devised an equation, which takes in to account

1) The Shannon Entropy
2) The Number of transitions between helices and sheets
3) The Standard Deviation in the lengths of the helices/sheets

All of these variables are weighted depending on how much they affect the structure complexity and the reasoning for including these parameters as well as their weights are listed below.

Since we only have the secondary structure, made of helices and sheets, estimated by the Chou Fasman method, the variables help to factor in the complexity associated with tertiary and quaternary structures.

The Shannon Entropy is self explanatory, as it by definition gives you the diversity of the different types of structures and the more diverse the amount of structures, the higher likelihood of the protein being complicated.
The Number of transitions indicate how many times in the sequence there was a change from a helix to a sheet or vice versa. Logically speaking, we can extrapolate that the more transitions there are, (i.e the more times the sequence changes from a sheet to a helix or vice versa) the higher likelihood of the protein folding on itself and becoming more complicated.
The Standard Deviation helps to calculate complexity because it measures how different the lengths of helices and sheets are. If a structure has a lot of helices or sheets with varying lengths, it's more likely to be complicated, as opposed to a structure that contains similar length helices/sheets, which indicates long chains, but no folding within each other.

The equation is also normalized by dividing by the total of the fractions of each type, so an edge case for example where a high amount of helices is present doesn’t appear as complicated.

The Shannon Entropy was deemed to affect the complexity the most, as it is directly structure related which is why it is weighted to 1.
The transitions were weighted less because although they are a good indicator, high transitions values could also be due to noise, hence the 0.3. (The number also tends to overpower the entropy value which ranges from 1 to 2)
In general, the variability of the length of structures doesn't contribute as much to the complexity.

Notes:

We can also include line dimension calculations here for an even more accurate model, but the line dimension algorithm we came up with isn't as accurate as we'd like it to be.

The complexity score ultimately obtained is in arbitrary units and can only serve to compare between different proteins.

In the case of analyzing proteins with more coils, (more structurally disorganized structures)/ IDPs (Intrinsically Disorganized Proteins) the number of coils and the variables associated with it can be calculated as well and added to the equation.