Geometric analysis of protein structures

Mustafa Omar

12/21

Contents

1 Abstract

With the goal to create a database of representations of proteins, and a further aim to help solve the protein function problem, this investigation focuses on obtaining primary protein representations derived from the protein's 3D dimensional structure. Those representations are the 3D structure (as a base), Distogram and the Convexhull of the protein. The python classes and functions created can then be extended to include more complex protein representations, that may be used in machine learning systems, which aim at predicting a protein's function. Currently only the Distogram can be used as input for a successful machine learning model (convolutional neural networks), other representations such as the convex hull may need more transformative steps before they can be used as inputs. Plots produced by the code are displayed inside of figures, this is to showcase the output of the code accompanying this article. Finally, no significant data generating processes have been carried out in this investigation,this is due to lack of relevance at this step and storage reasons, it is however possible to start but it wont be of any immediate significance.

2 Introduction

Currently there are many known and important problems in science and technology, for which, we have no solutions. As we discover and answer questions in those fields, it is apparent that more abstract and difficult questions start to manifest themselves. We then start reaching the limit of what we as humans are able to interpret and solve without the aid of machines/computers. However, with the advancement in hardware and AI methodology (data driven science), it is now possible for us to answer questions previously unanswered, but in order to do so, there is a requirement...data. Without relevant and useful data, our methods cannot produces insights into anything, they (ML models) simply cannot work. Therefore it is crucial that a unique set of data is derived in order to make progress for a particular problem.

In biology proteins are the fundamental machines behind the operation of life. we have already build a substantial data set of protein sequences [Wheeler et al., 2006] and structures [team EMBL-EBI,], the former helped to build the latter and the latter will be used in this article to derive protein representations and later a data set that could prove useful in solving the protein function problem (Discussed in 2.3).

2.1 The protein

Proteins are the fundamental building block of life, each protein has a particular function to carry out throughout its lifetime. The classes of functions that proteins can carry out are as follows: Structural, functional and more (outside of scope). Structural proteins are self explanatory, they serve to provide structural support or give unique structural properties e.g elasticity, on the other hand, functional proteins, provide a unique interaction with other molecules present in the environment e.g CO2,O2, H2O etc, this interaction may alter the chemistry or simply server as a means of translocation of said molecules. We are particularly interested in understanding/predicting how the latter works (Functional proteins).

2.2 The protein structure problem

The fundamental structure of the protein is the primary structure, this is simply the order of the amino acids that define the protein. Through complex hydrophobic/philic interactions and electrostatic forces, the protein slowly folds into a 3D structure, this structure then determines the function of the protein. The problem that we had but is now solved [Senior, 2020], was the determination of the protein's 3D structure from its amino acid sequence. Having solved this problem a more abstract problem occurs, that being, how do we determine the function of a protein given its 3D structure.

2.3 The protein function problem

In theory we could simply run a simulation to determine how a protein will interact with (x), where (x) is a set of molecules we would like to test against the protein. This approach to determining the function of a protein is inefficient, this is because, we are limited by a fundamental theory of physics, quantum physics. To determine the real and exact position and function of a set of atoms arranged and linked by molecular bonds in 3D space, a substantial amount of classical computational power is required to simulate the quantum interactions, even for the most trivial of proteins, running an accurate quantum

simulation, although possible, is very time and resource consuming. Furthermore with the promise of quantum computation, running such simulations may become possible in the distant future. The question then becomes, what can we do right now to aid with solving the protein function problem ?

2.4 Geometric analysis of a protein

Different protein geometries produce different functions, therefore, we could potentially gain insight into the function of a protein by analysing the 3D geometry using representations that hold some insight about the chemistry of a protein. Those representation could be derived through a verity of methods and maybe of varying complexity.

In this article, a very simple set of representations are derived from the 3D structure. Those representation do not rely on physical calculations to obtain insight, they simply rely on distances between the atoms in the proteins to obtain potentially useful/significant insights. Though it is definitely going to be a requirement to include physics/chemistry informed representations of proteins in future projects.

3 Procedure

3.1 MMCIF file-type, the Parser class and Data

As a convention, the data of a particulate protein are stored in a particular type of file named MMCIF (The macro-molecular Crystallographic Information File), this file stores everything about the protein, from its authors, species, gene in which it is present, order of amino acid residues, XYZ coordinates of each atom and much more relevant information. For our purposes, we primarily used the the residues and the XYZ coordinates to derive our representations. To extract this information the package [Gemmiteam,] was used, then for usage, the class (Parser) was imported into relevant modules, this allows for code re-usability.

Finally the mmCIF files used in this project are obtained from Alphafold's protein data bank [team EMBL-EBI,].

3.2 3D geometry/structure

The structure of the protein, is simply a representation of how each atom is arranged in 3D space, this is directly obtained from the mmCIF file, and plotted using Matplotlib [Hunter, 2007], this will not be of any significance as direct input into a machine learning system, this is because our current architectures aren't optimised to train and make predictions from a complex 3D geometry such as this. Instead, the 3D geometry is used as a base from which other representation are derived, hence its inclusion in the code.

3.3 Distogram

The Distrogram is a representation of the geometry of the protein, it's short for distance histogram. For each amino acid residue we obtain a value for the distance between that amino acid and every other amino acid in the protein, therefore mathematically, the Distogram can be represented by a 2D matrix and thus as a 2D image. The implementation of the Distogram was simply done using Numpy arrays [Harris et al., 2020].

3.3.1 Significance of the Distogram

The Distogram was the primary target of prediction of Deepmind's alphafold [Senior, 2020], their algorithm used convolutional neural networks to find the optimum Distogram through the use of differential models for which the minima was found using gradient decent methodologies.

Furthermore, knowing the accurate Distogram of the protein and coupling this representation with other representations such as the average electric charge density or else, one could in theory derive a representation that has significant power in predicting the likelihood that a protein may interact with a specific molecule in its environment.

3.4 Convex-hull

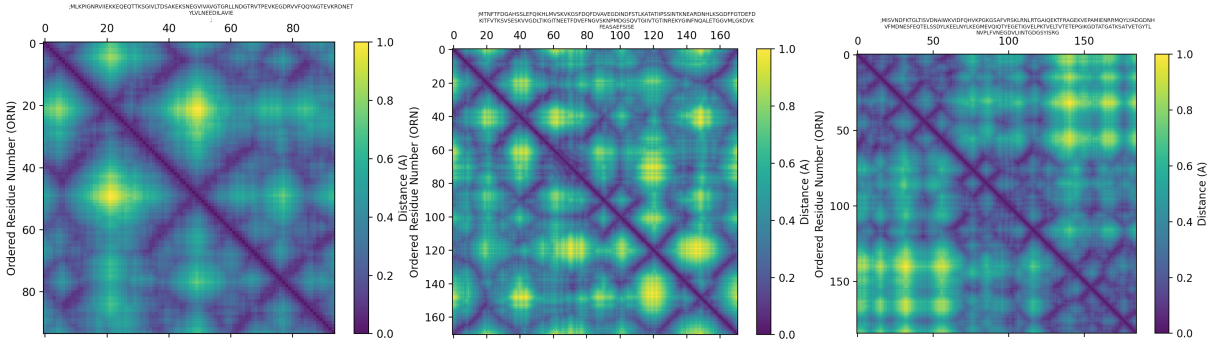
The Convex-hull is the subset of points that represent the edges/vertices of a particular set of points in 2D+.This representation simply tells us about the bounding limits of a protein. The convex hull is calculated using Scipy [Virtanen et al., 2020], which in turn applied some variation of the Convex-hull algorithm outline in [T.M, 1996].

3.4.1 Significance of the Convex-hull

Some functional proteins operate in different parts of the environment, for example some proteins operate inside of the cell (intra-cellular), others outside of the cell (extra-cellular) and others operate on cell surface membranes. Therefore, in theory, knowing the bulk structure of the protein as represented by the convex hull, could prove significant in predicting where a protein is likely to operate, this is because the size and shape of a protein may prevent it from operating in certain environments, for example round and irregularly shaped protein may have a low probability of being found to operate on cell surface membranes etc.

4 Results

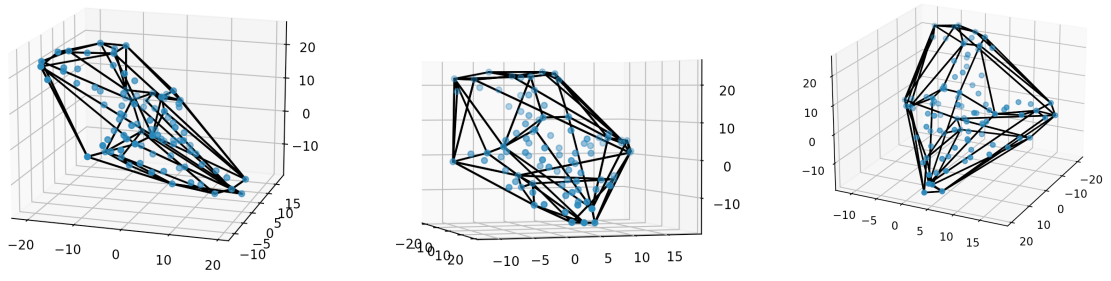
4.1 Distogram



(a) (b) (c)

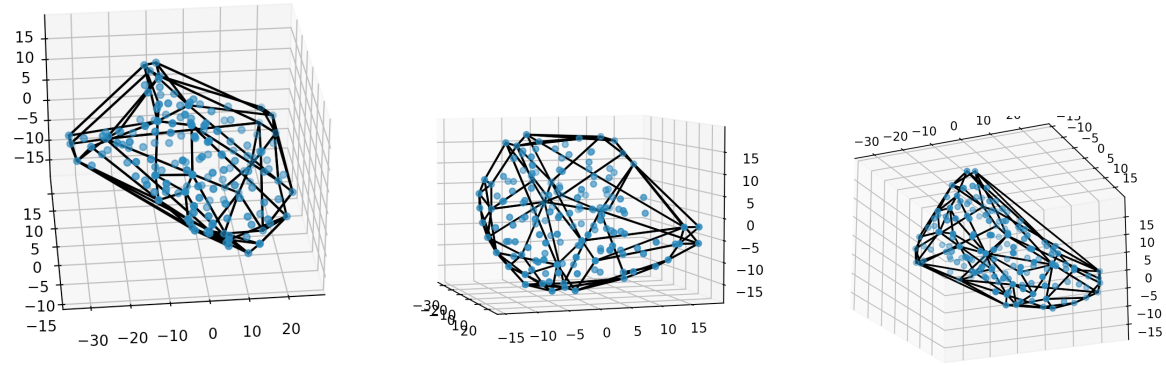
Figure 1: Showing the result of plotting a Distogram using the Distogram class, subfigures (a),(b) and (c) showing Distograms of arbitrary proteins obtained from Alphafold's protein databank[team EMBL-EBI,]

4.2 Convex hull



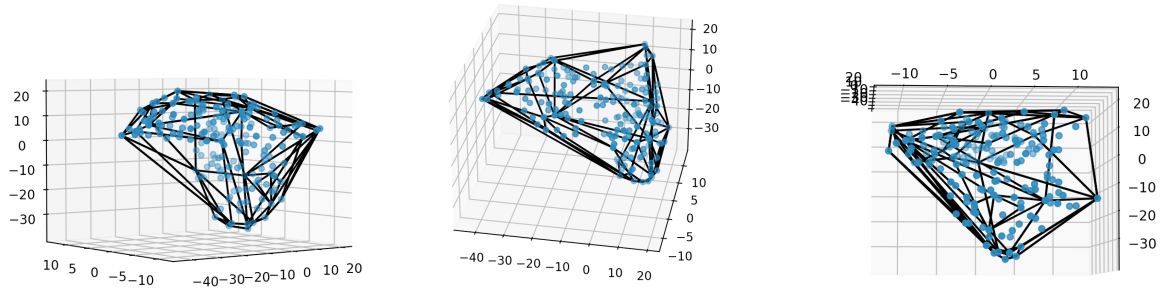
(a) (b) (c)

Figure 2: Showing the convex hull (From 3 different angles) of the protein in sub figure (a) in figure 1



(a) (b) (c)

Figure 3: Showing the convex hull (From 3 different angles) of the protein in sub figure (b) in figure 1



(a) (b) (c)

Figure 4: Showing the convex hull (From 3 different angles) of the protein in sub figure (c) in figure 1

5 Discussion

In sections 4.1 and 4.2 plots were obtained directly from code. Plots in section 4.1 show the Distograms of arbitrary proteins obtained from Alphafold's protein data bank, those plots/images, can be Directly used as input from a convolutional neural network that is part of a larger machine learning system. Furthermore the plots in section 4.2, show the Convex hull of the proteins in figure 1, where each of the 3 plots represents each of the subplots respectively. This representation is not very useful for a machine learning system, as our current architectures has no means to use as input the 3D plot of the convex hull, thus, more prepossessing/reformatting is required to obtain a representation more suitable for our current architectures/future architectures.

Furthermore as a point of improvement, more representations can be derived and added. Some will require physical calculations such as the electric charge around a protein or the Ramachandran plot (This plot shows how far atoms are allowed to rotate before steric collision occur). This is a developing area of research, and thus, there's yet more to calculate before the data generation step.

6 Conclusion

To conclude, The code produced in this project has allowed us to produce a set of plots that maybe useful in predicting the probability that a given protein will function in a particular way. It may also be the case that this investigation will serve as the seed from which more complex representations of proteins may be derived, that may have a higher significance when attempting to predict a protein's function.

Furthermore data generation can be done for the current representations we have (Distogram and convex hull), there are currently over 800,000 proteins structures available in Alphafold's protein databank. This

will serve as the seed data from which our representations will be made. For obvious storage and relevance reasons, no such data mining processes have been attempted yet.

References

[Gemmiteam,] Gemmiteam. Gemmi pakage .

[Harris et al., 2020] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585:357–362.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95.

[Senior, 2020] Senior, A.W., E. R. J. J. e. a. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577:706–710.

[team EMBL-EBI,] team EMBL-EBI, D. AlphaFold Protein Structure Databse .

[T.M, 1996] T.M, C. (1996). Optimal output-sensitive convex hull algorithms in two and three dimensions. *Discrete Comput Geom*, 16(10):361–368.

[Virtanen et al., 2020] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

[Wheeler et al., 2006] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2006). Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl_1):D5–D12.