ProtFlex: Protein flexibility assessing program

ProtFlex: a python standalone program to predict and assess the flexibility of proteins. Artigues-Lleixà M[1] and Torrén P[1].

[1] MSc Bioinformatics for Health Sciences, Pompeu Fabra University, Barcelona, Spain. 2022.

1. What is ProtFlex?

ProtFlex is a new bioinformatics tool, developed in python, used to assess the flexibility of a given protein sequence and predicted structure. From an input FASTA formatted file, ProtFlex performs a blastp to identify the protein as the hit with lowest e-value and recovers its predicted structure from the AlphaFold database, using the UniprotID. Once the structure is recovered, ProtFlex constructs a Gaussian Network Model to obtain the Normalised Square Fluctuations, to establish the flexibility score of each aminoacid.

2. Biological framework

Proteins are dynamic structures where their inherent flexibility allows them to function correctly through molecular interactions. So, flexibility calculation is of great importance to understand the biological relevance of the proteins. [1]

First, to understand how the flexibility is calculated, we will infer into Normal mode analysis and its evolution. Normal mode analysis is a technique that provides information on the equilibrium modes accessible to a system, assuming that the system is stabilised by harmonic potentials.
A feature of the normal mode analysis is the robustness of the global modes (represent reconfigurations along directions that require the least energy ascent for a given size deformation) which are defined by the entire structure.This means that these global modes are insensitive to local interactions or specific energy functions and parameters that define the force field, which are result from their systemic nature.
So, taking inspiration with the insensitivity of global modes to structural and energetic details, elastic network models appeared. Elastic network models are entropic models that capture the overall internal motions and provide an exact solution for unique dynamics of each structure. Inside these elastic network models we encounter two different types of models: Anisotropic network model (ANM) and Gaussian network model (GNM). [2]

2.1 Elastic network models: ANM and GNM

The Anisotropic network model (ANM) is a coarse-grained normal mode analysis, subject to potential. The overall potential is the sum of harmonic potentials between the interacting nodes. The position of the nodes are identified by the coordinates of Cα atoms for amino acids. A major utility of the ANM is its ability to generate alternative conformations (substates or microstates) in the close neighbourhood of a given structure upon deforming the original structures along the dominant (lowest frequency) modes. [2][3]

Gaussian network model (GNM) is a minimalist, coarse-grained normal mode analysis and describes the vibrational dynamics of proteins and their complexes. In the model, proteins are also represented by nodes corresponding to the Cα atoms of the amino acids. This model is based on the assumption that all residue fluctuations (and inter-residue distances) are gaussianly distributed around their equilibrium coordinates. [2][4] It seems that GNM performed with Cα - atoms (as it can be also performed with other points of reference like Cβ atoms) performs the best comparing it against other GNM. [5]

Whereas the GNM is limited to the evaluation of the mean squared displacements and cross-correlations between fluctuations, the ANM approach permits the evaluation of directional preferences, providing a 3-D description of the 3N - 6 internal modes. In the GNM, the molecule is viewed as a collection of N sites, one for each residue, resulting in an ensemble of N - 1 independent modes, instead of the 3N - 6 modes obtained with the ANM.
Even with its simplicity, it has been observed that GNM fluctuation predictions agree better with experiments than those computed with ANM. In ANM, we need a higher cutoff distance for interactions than in GNM (usual cutoff distances are 13-15Å and 7Å respectively), as it has been observed that it gives rise to excessively high fluctuations compared to the GNM results or experimental data. With a higher cutoff distance, each residue is connected to more neighbours in a more constrained and consolidated network. [6]
The higher performance of the GNM can be attributed to two different features: 1) on a coarse-grained scale, molecular motions can be approximated by normal fluctuations with rescaled force constants, even if the motions of individual atoms depart from harmonicity; and 2) the GNM yields an analytical solution, without the sampling inaccuracies found in other types of simulations. [7]

In conclusion, GNM is more accurate and should be chosen when evaluating the deformation magnitudes or the distribution of motions of individual residues, and that is why ProtFlex uses GNM with Cα atoms as nodes. [5]

2.2 Flexibility parameters: B-factors and Square Fluctuations

While protein folds to a unique structure, it also fluctuates and makes thermodynamic movements. Protein molecules exhibit varying degrees of flexibility throughout their structure, with some segments showing little mobility while others may be disordered, unresolvable by techniques such as X-ray crystallography.

The B-factor, also called temperature factor, Debye-Waller factor, or atomic displacement parameter, is a term used in protein crystallography to describe the attenuation of X-ray or neutron scattering caused by thermal motion, therefore measuring the structural fluctuations of a protein caused by the temperature-dependent vibration of the atoms crystalized. This factor can be used to identify the flexibility of atoms, side chains or whole regions in a protein. [8]

Mean square fluctuation or mean squared displacement is a measure of the deviation of the position of a particle with respect to a reference position over time. It is the most common measure of the spatial extent of random motion.

We can compare the B-factors obtained in crystallography with residue mean-square fluctuations as B-factors are a consequence of the dynamic disorder in the crystal caused by the temperature-dependent vibration of the residues in the protein. However, protein crystals also have static disorder (molecules or parts of them in different unit cells do not occupy the same position or exactly the same orientation). What static disorder do is that B-factors may not always reflect the fluctuations correctly.

2.3 ProtFlex

The program that we created has the objective of developing a flexibility score for proteins derived from AlphaFold. Proteins predicted in AlphaFold do not have the experimental values as they are predicted and not crystalised. The fact that the proteins are not crystallised is important, as we can not use the B-factor as a parameter to study the flexibility of the input protein. [10] For this reason, we adopt as a flexibility score the normalised square fluctuations obtained from a GNM [5]. These normalised square fluctuations indicate to us the degree of movement for each residue.

With ProtFlex, we use AlphaFold as a database to obtain the modelled protein. We use this database as PDB only has experimentally obtained proteins and, often we can't find the entire structure of the protein, only certain domains. By using AlphaFold, we can access the whole protein structure, achieving a more complete look of the protein behaviour. Moreover, we can also calculate the flexibility of proteins whose structure can't be empirically resolved.

When working with AlphaFold structures we have to keep in mind that they are predictions and so, every residue has associated an estimate of its confidence.

ProtFlex is an innovative tool as up to this day, B-factors are the parameters that most programs use to obtain the flexibility of a protein, but to obtain these parameters, we have to obtain the structures by X-ray crystallography with a very good resolution for them to be reliable. Moreover, B-factors are not purely a flexibility score per se and can be very influenced; that is why we chose square mean fluctuations as our flexibility score.

3. How does ProtFlex work?

ProtFlex is composed of two main scripts, one with the core algorithm and the other with the functions developed to perform the flexibility assessment. In the tutorial section can be found a detailed description of every part of the algorithm.

4. Requirements

To work with this package a Python3 interpreter must be installed on the computer. This package has been developed using Python 3.8.10.

ProtFlex requires the library BioPython, can be installed using pip3:
$ pip3 install biopython

5. Installation

To download the package is needed to clone the git repository:

$ git clone https://github.com/mariaartlle/PYT-SBI-project-ProtFlex.git
$ cd PYT-SBI-project-ProtFlex

Once we are inside the cloned directory, the package can be installed manually using the setup.py file:

$ pip3 install .

6. Tutorial

Arguments description

- -i / --input

Required argument. The input argument can be a FASTA formatted file or, alternatively, a single UniprotID code. If the input is not in the correct format, the program will raise an exception.

- -o / --output

Optional argument. The name of the desired output file, if it already exists, the file will be overwritten, if not the program will create a new one with the name provided. If this argument is not defined the output file will be named after the UniprotID of the provided input file.

- -g / --graph

Optional argument. If this argument is defined, a graphical representation of the flexibility scores will be provided besides the parseable output text file.

- -pdb

Optional argument. Needs to be a PDB ID of a structure resolved by X-ray crystallography. When defined along the graph argument (-g), the output graph produced will include the

normalised experimental B-factors of the PDB provided. Useful to visualise the empirical parameters along the predicted ones.

Execution

To execute ProtFlex we need the fasta formatted file of the protein we want to calculate the flexibility scores or its Uniprot ID. Either way, the input argument is the same:

$ python3 protflex_core.py -i Q9Y223.fasta -g

Here we are using a fasta file (with only one protein) to execute ProtFlex and indicating that we want a graphical representation of the results.
$ python3 protflex_core.py -i Q9Y223 -g -pdb 2YHW

In this example, apart from calculating the flexibility scores of Q9Y223, we are also providing the program with an X-ray crystallography resolved PDB file. The graph that will be obtained will also have represented the normalised experimental B-factors along the normalised square fluctuations to facilitate the comparisons to the user.

$ python3 protflex_core.py -i Q9Y223 -o my_output.txt

In this example we are using the UniprotID and indicating the name we want for the output file. We won't retrieve any graphical representation of the results, only the default text file with the flexibility values.

ProtFlex algorithm's description

- **FASTA file**

When a FASTA formatted file is used as input the first step is performing a BLASTP search against the uniprot database to retrieve the best hit and its UniprotID, which will be used to retrieve its predicted AlphaFold structure (pdb file).

- **UniprotID**

When a UniprotID is given as input, the program retrieves its AlphaFold predicted structure (pdb file).

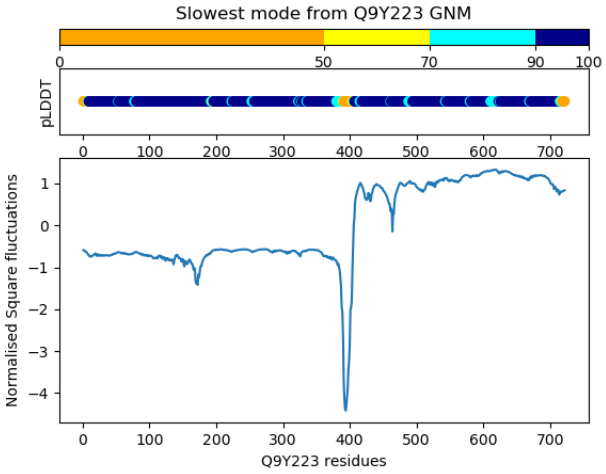
Once the structure has been retrieved and parsed, a Gaussian Network Model is constructed with its alpha carbons. A Kirchoff matrix is calculated with the alpha carbon coordinates set and then it is diagonalized to calculate the normal modes. To obtain our flexibility score the slowest mode is used to calculate the Normalised Square Fluctuation of each amino acid of the provided protein. These values represent the sum of square-fluctuations for a set of normal modes, when a single mode is provided the values are calculated by multiplying the square of the mode array with the variance along the mode. For our flexibility score we use these values normalised: the higher the values are, the more flexible are the aminoacids (as their fluctuation is bigger). ProtFlex's flexibility score units are arbitrary because of the way the GNM's Square Fluctuations are calculated.

As for the output results, the program creates a parseable output text file with the information of the alpha carbons used (residue name, number and chain), the confidence value of the prediction by AlphaFold (pLDDT) and the Normalised Square Fluctuation value of each aminoacid.

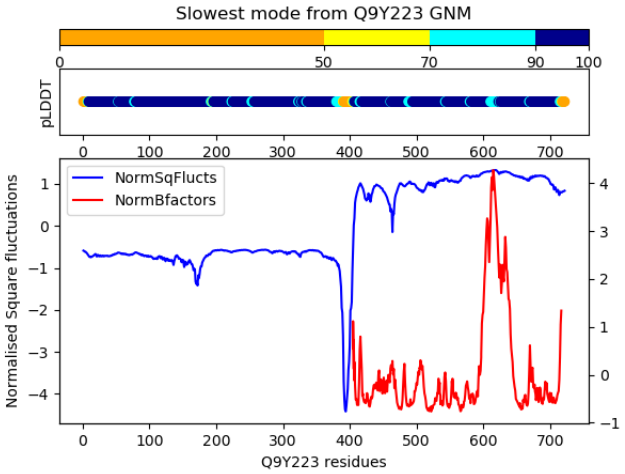
ProtFlex_Q9VVG4_out.txt

```
ProtFlex parseable output file
Analyzed protein: Q9VVG4
AT  AA  CH  N   pLDDT   NormSqFluct
CA  MET A   1   23.23   2.079
CA  LEU A   2   27.02   2.076
CA  SER A   3   30.05   2.078
CA  ILE A   4   30.77   2.078
CA  GLY A   5   31.58   2.067
CA  ALA A   6   38.08   2.062
CA  MET A   7   39.93   2.055
CA  ALA A   8   47.81   2.045
CA  ASN A   9   63.12   2.032
CA  ILE A   10  73.30   2.017
CA  LYS A   11  73.23   2.028
CA  HIS A   12  69.63   1.994
CA  THR A   13  81.74   1.999
```

If indicated in the execution of the program the user can also retrieve a graphical representation of the results.



If along the graph argument, the user provides an X-ray crystallography PDB structure, the graphical representation of the results will also have the experimental B-factors represented in the plot along the square fluctuations.



Throughout the program's execution the user can follow its progress through updates in the terminal. To ensure easy problem solving, a logging file is available to the user when the program is finished.

7. Limitations

The most important limitation of the program is that it can only assess the flexibility of proteins available in the Uniprot database whose predicted structure can be retrieved from the AlphaFold database.

Furthermore, as we are working with predicted protein structures, we have to bear in mind the confidence in their prediction. To address this conflict, AlphaFold produces a per-residue estimate of its confidence, called pLDDT, that corresponds to the model's predicted score on the IDDT-Cα metric. This estimate goes from 0 to a 100: the regions with pLDDT >90 are expected to be modelled to high accuracy, while the ones with pLDDT between 50 and 70 should be treated with caution due to their low confidence. This metric is included in ProtFlex results, as it is crucial to understand the reliability of the flexibility scores calculated.

Also, ProtFlex is designed to work only with a protein at a time, therefore, it cannot process files with multiple fasta sequences or multiple UniprotIDs.

8. Examples

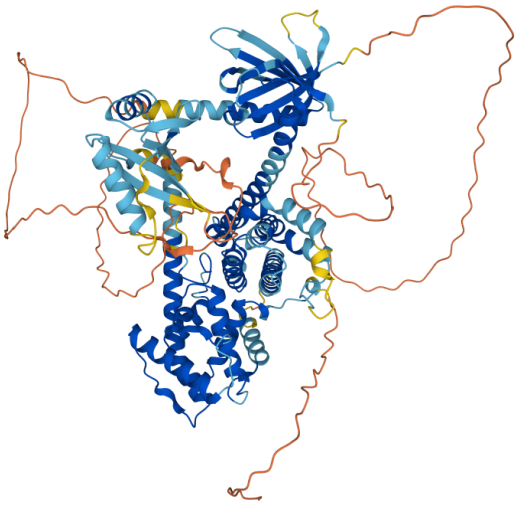
To enlighten the reader about ProtFlex's functionality, we have performed the analysis of 4 different proteins, availables in the github repository and the compressed package.

Example 1: P11433

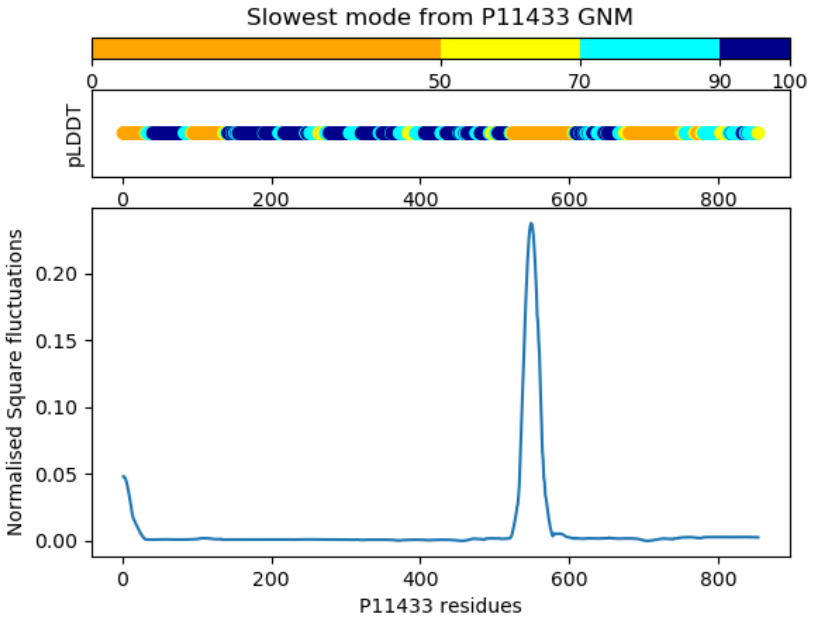
P11433 corresponds to a cell division control protein (CDC24) present in Saccharomyces cerevisiae. Its function relies on the promotion of the exchange of CDC42-bound GDP by GTP, a crucial step in the yeast mitotic cycle. It also plays a role in calcium regulation and the selection and organisation of the budding site. This protein can be found partially resolved by NMR in the PDB database. To stick to the purpose of the program we will analyse the results of the whole protein predicted by AlphaFold.

- **AlphaFold structure**

Taking a look into its structure we can rapidly identify different parts in the protein, we see a very organised core, mainly conformed by alpha helices and some other beta supersecondary structure. Enveloping this core there are two loops, both with very low confidence estimates.



- **ProtFlex results:**



The region with the higher values corresponds to the loop delimited by LYS525 and, roughly, SER600 (visible in the right upper quadrant of the structure photograph). In the Uniprot entry for this protein we can find that there's a domain delimited between residues 478-668 that is also found in other proteins involved also in GTP/GDP exchange.

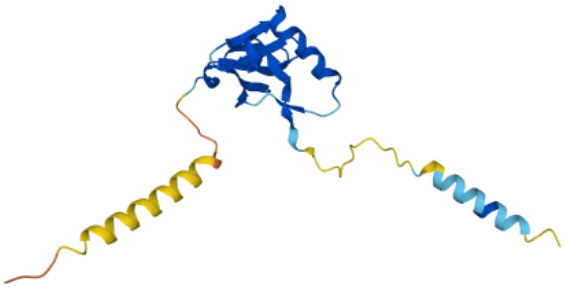
Due to the low values of the pLDDT estimate in this region we can't say with confidence that it is the most flexible region. Since the creation of AlphaFold, there has been numerous studies trying to gain new insights in this prediction technology and its capabilities. Recently it has been seen that there is a negative correlation between the pLDDT score from AlphaFold and the main chain flexibility of proteins, meaning that highly flexible regions tend to have a lower pLDDT score [11].

<u>Example 2: P16041</u>

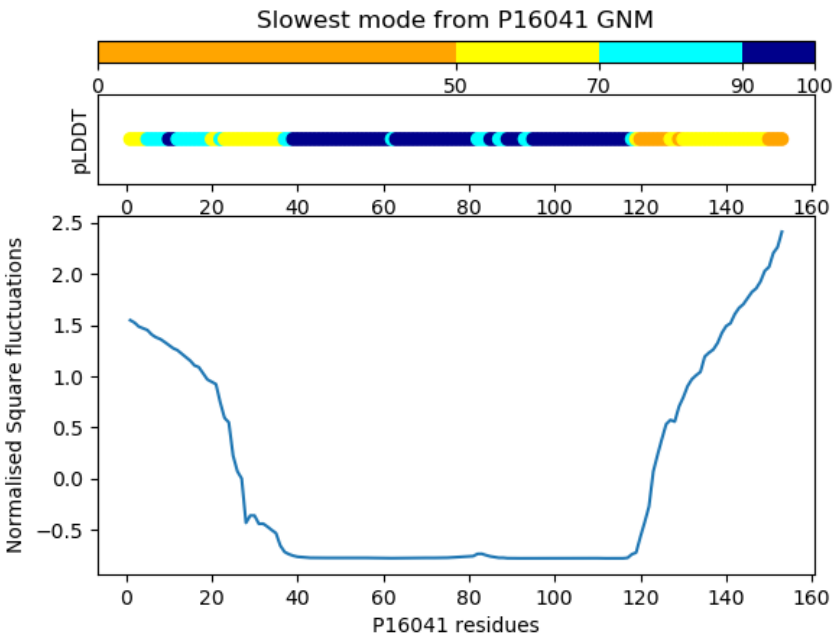
P16041 corresponds to the protein Vasotocin-neurophysin VT1, an antidiuretic hormone from a salmon species.

- **AlphaFold structure**

The structure is only available in the AlphaFold database. If we take a look at the predicted structure we can see that the core of the protein is composed mainly by beta strands, while both the extremes have a single alpha helix.



- **ProtFlex results**



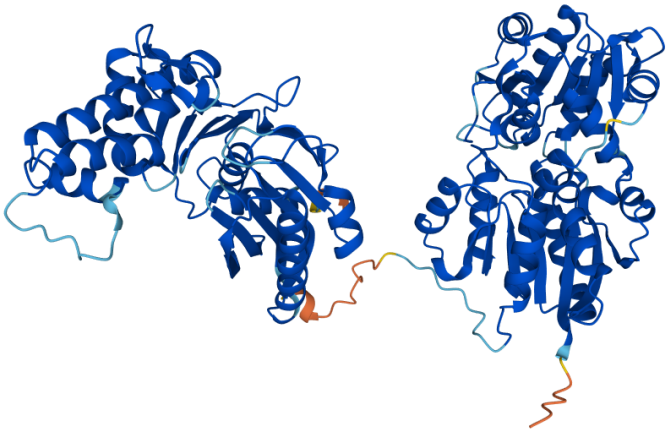
Regarding the ProtFlex's results we can easily identify the N-terminal and C-terminal as the most flexible regions of this protein (they have higher flexibility scores) . We have to take into account that these regions are also the ones with lowest confidence estimates for their prediction. These results are consistent with the protein structure, as the core conformed by beta strands is much more stable and rigid than the single helices that conform the beginning and end of the protein.

- Example 3: Q9Y223

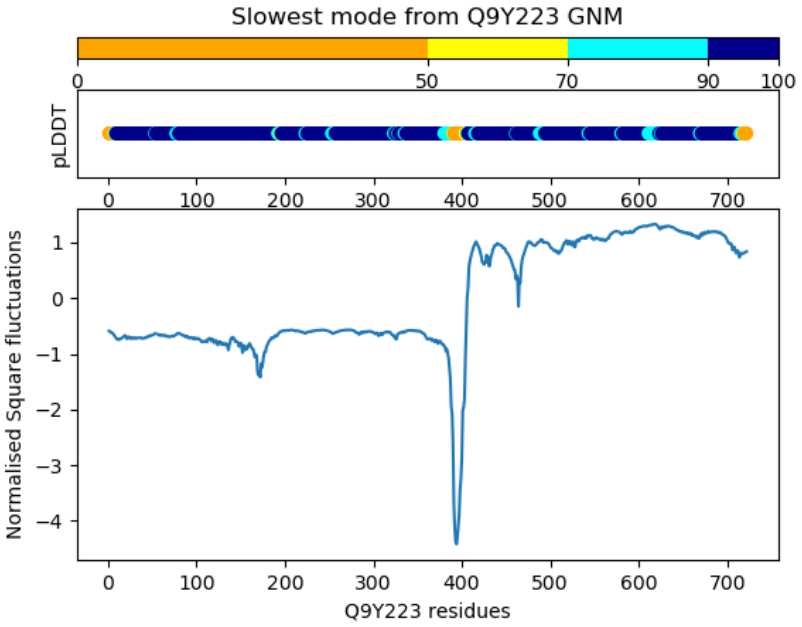
Q9Y223 corresponds to a human Bifunctional UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase. This protein is involved in the regulation of N-acetylneuraminic acid biosynthesis and plays an essential role in early development, as it is required to synthesise sialic acids.

- **Alphafold structure**

This protein has been partially resolved by X-ray crystallography experiments, to look at the whole structure we will analyse the AlphaFold predicted structure. The protein is formed by two differentiated domains, formed by both alpha and beta supersecondary structures. Another remarkable aspect of this structure is the binding of different substrates, zinc and ATP.

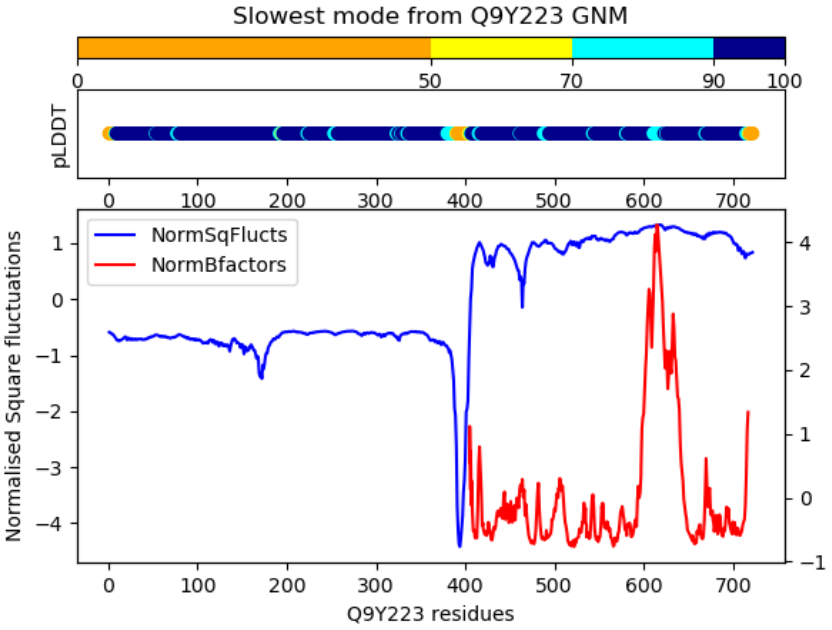


- **ProtFlex results**



In comparison to the other studied proteins, the flexibility values for this protein are very low, and all the residues have been predicted with high confidence estimates. Interestingly, the region with the lowest flexibility scores is the one that comprises a loop that joins the two domains of the protein and it is also the part of the protein with lowest confidence estimate. This region isn't well characterised, so we can not know for sure the nature of this relationship.

The most studied region of this protein is the one between positions 406 and 720 and has been experimentally resolved by X-ray crystallography. For this reason it may be interesting to take a look at the experimentally determined B-factors:

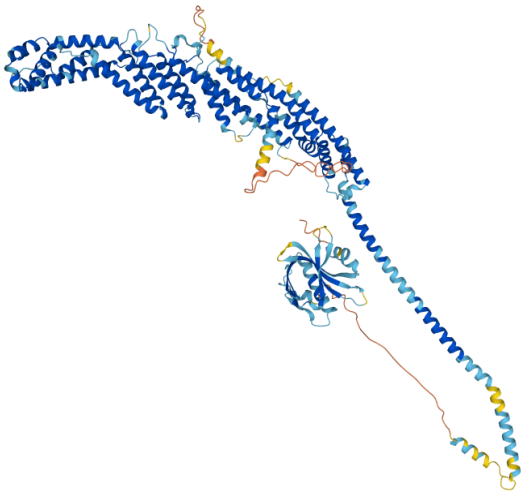


We don't have a reference point to compare the B-factors' values as the protein is only partially resolved, but we can see that both the square fluctuations and B-factors have high values in this region. Also it is worth mentioning that the residue with higher value in both parameters is the same. This region comprises all the known binding sites for zinc, ATP and multiple substrates: it is expected of this kind of region to have more flexible residues that can adopt more than one conformation to bind other agents.

- Example 4: Q9VVG4

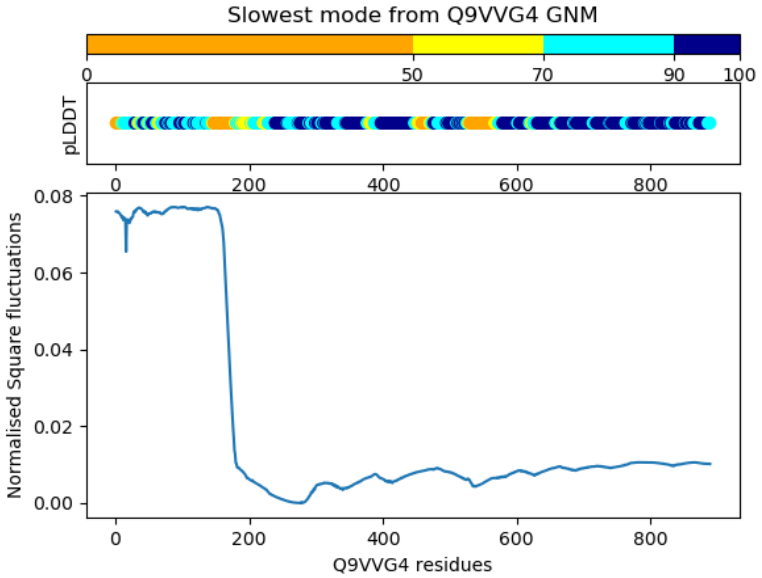
Q9VVG4 corresponds to the Exocyst complex component 1 of *Drosophila melanogaster*. This protein is involved in exocytosis processes, more specifically in the docking of exocytic vesicles onto the plasma membrane.

- **AlphaFold structure**



This protein has not been resolved by any empirical method so we only have the predicted structure available in the AlphaFold database. If we take a look at it we can see that it is mainly formed by alpha helices, although the N-terminal has a beta supersecondary structure connected to the main structure with a loop.

- **ProtFlex results**



The region with higher flexibility values is the N-terminal one, including both the beta structure and the loop. This region is also the one involved in binding PIP2 molecules according to the information available in Uniprot and PFAM databases. It would make sense that, since it is the domain in charge of binding other agents this region would be the one with highest flexibility scores. From residue 200 the flexibility scores barely change, taking into account that most of the residues are part form alpha helices superstructures, it wouldn't be expected much flexibility.

9. Bibliography

[1] Teilum, K., Olsen, J., & Kragelund, B. (2009). Functional aspects of protein flexibility. *Cellular And Molecular Life Sciences*, *66*(14), 2231-2247. doi: 10.1007/s00018-009-0014-6

[2] Bahar, I., Lezon, T., Bakan, A., & Shrivastava, I. (2009). Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chemical Reviews*, *110*(3), 1463-1497. doi: 10.1021/cr900095e

[3] Eyal, E., Yang, L., & Bahar, I. (2006). Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, *22*(21), 2619-2627. doi: 10.1093/bioinformatics/btl448

[4] Zhang, H., Jiang, T., Shan, G., Xu, S., & Song, Y. (2017). Gaussian network model can be enhanced by combining solvent accessibility in proteins. *Scientific Reports*, *7*(1). doi: 10.1038/s41598-017-07677-9

[5] Park, J., Jernigan, R., & Wu, Z. (2013). Coarse Grained Normal Mode Analysis vs. Refined Gaussian Network Model for Protein Residue-Level Structural Fluctuations. *Bulletin Of Mathematical Biology*, *75*(1), 124-160. doi: 10.1007/s11538-012-9797-y

[6] Rader, Andrew & Chennubhotla, Chakra & Yang, Lee-Wei & Bahar, Ivet. (2006). The Gaussian Network Model: Theory and Applications. Normal Mode Analysis. Theory and Applications to Biological and Chemical Systems.

[7] Atilgan, A., Durell, S., Jernigan, R., Demirel, M., Keskin, O., & Bahar, I. (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical Journal*, *80*(1), 505-515. doi: 10.1016/s0006-3495(01)76033-x

[8] Sun, Z., Liu, Q., Qu, G., Feng, Y., & Reetz, M. (2019). Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chemical Reviews*, *119*(3), 1626-1665. doi: 10.1021/acs.chemrev.8b00290

[9] Drenth, J., 2011. *Principles of protein X-ray crystallography*. New York: Springer.

[10] Bramer, D., & Wei, G. (2018). Blind prediction of protein B-factor and flexibility. *The Journal Of Chemical Physics*, *149*(13), 134107. doi: 10.1063/1.5048469

[11] Saldaño, T., Escobedo, N., Marchetti, J., Zea, D., Mac Donagh, J., & Velez Rueda, A. et al. (2021). Impact of protein conformational diversity on AlphaFold predictions. doi: 10.1101/2021.10.27.466189