Unravelling a flexibility score for proteins: Fancy FlexScore

Final Project Structural Bioinformatics and Python
2021/2022

LEIDY ALEJANDRA GONZÁLEZ MOLANO
IRIA POSE LAGOA
GERARD ROMERO SOLA

Contents

1 Introduction

Flexibility is an inherent property of proteins. Despite the fact of being crucial for the protein function, stability and interactions, no consensus for flexibility assessment has been established. Nowadays, a wide range of approaches are considered to estimate this parameter, ranging from experimental data, such as the temperature factor, to a combination of evolutionary and physico - chemical information found in the MEDUSA webserver [1]. The present project provides the flexibility score of a protein given its protein sequence. It is based on a combination of evolutionary, experimental, and physico-chemical information to predict the dynamic nature of proteins.

1.1 Background

The flexibility of a protein could be defined as the rate of protein conformational change. This flexible capability allows proteins to change and adapt to different biological and biochemical process that are essential for its function and interaction with other molecules and the environment. However, an accurate prediction of this behavior turned out to be a real challenge. [2]

Temperature factors from X-ray crystallography, commonly known as B-factors, give an experimentally determined indication of the degree of mobility in a protein structure. They reflect the degree of thermal motion and static disorder of an atom in a protein crystal structure. B-factors are also used to study many other biological properties such as the classification between biological interfaces and crystal packing contacts [3] or finding breakpoints in helices [4]. Some studies [5, 6], made use of this experimental value in order to asses a flexibility score of a protein, or to be more accurate, a flexibility value for each amino acid. Nevertheless, while in many cases B-factors can be easily gathered from PDB files, this measure is not always available. For this reason, the conducted approach consists in retrieving an approximate B-factor from homologous sequences after the evaluation of a multiple sequence alignment (MSA). In this context, this project aims at estimating a Flexibility score for each amino acid of a given protein for which no B-factor values are available.

1.2 Objectives

The main purpose of this project is to provide an assessment of the flexibility of a protein given its sequence. In detail, the function outputs a text file containing the flexibility scores per amino acid in addition to a graphical representation of these scores along the sequence. Furthermore, both outputs include complementary information regarding the hydrophobicity and secondary structures of the protein, to provide a comparison with parameters theoretically related to flexibility.

2 Materials and Methods

2.1 Evolutionary perspective

The first step of the function is obtaining the homolog candidates for the query protein. The evolutionary information of the protein family has been considered by conducting a BLASTP with a restrictive e-value threshold (1^{-20}) in order to obtain close homologs that share a high (%) of identity with the target. The homologs are obtained from a PDB [7] Database.

As an alternative to the BLASTP, if no homologs are found due to the restrictive threshold, two PSI-BLAST are conducted, with a more permissive threshold to obtain remote homologs. First, a PSSM (Position Specific Scoring Matrix) is obtained by using the UniProt database in order to avoid a non-evolutionary bias. Then, the PSSM is used to conduct a second PSI-BLAST, this time against the PBD database to find homologs with known structure.

The homolog candidates are then filtered in terms of quality, removing those with a resolution lower than 2Å [8]. Eventually, after applying the filter, only three candidates are kept in order to perform the MSA.

2.2 B-factors

Our following step, was to obtain the B-factor, as they will be the values to use when computing the flexibility score. To achieve this objective, first of all, we obtain and the normalize the b-factors of each homolog. B-factor values depend on several factors, including degree of resolution, crystal contacts, and refinement procedure [8]. Therefore, a normalization must be taken in order to compare different structures. The formula use to normalize the B-factors was the following:

$$B_i = \frac{B_i - \bar{B}}{\sigma}, \tag{1}$$

where $\bar{B}$ is the B-factors mean of the protein and σ the standard deviation.

Afterwards, we performed a Multiple Sequence Alignment (MSA) using T-Coffee [9] in order to get the similarity regions of our protein with the query. We decide to use t-coffee because it corrects misalignment errors in the MSA.

Then, to obtain the b-factors of our target protein, we scan the alignment and follow these rules:

- We remove the possible "X" amino acids, that is, the missing residues

- If there were coincidences between target and the three homologs we save the α-C b-factors b-factors and we compute the mean of them.
- If two homologs share the same amino acid with the target, we did the same as in the previous.
- If only one homolog shares the amino acid, we took its α-C b-factors.
- When there wasn't no homolog amino acid aligned with our target amino acid, we use the improved amino acid flexibility parameter computed by [5].

2.3 Flexscore

Once we have saved the b-values (or the improved amino acid flexibility parameters computed by [5] when b-values are not available), the flexibility score can be computed. The flexibility of a residue in an amino acid chain is dependent on whether the neighbor(s) of the residue is rigid or flexible, because a rigid neighbor decreases the residue flexibility and a flexible one does the opposite. In this way, we define the flexibility index as a weighted average of amino acid flexibility over whole residue chain of the protein. The neighbourhoods effect is considered using a sliding hat-shaped window. Therefore, the following formula was used to calculate the flexibility score of each amino acid of our protein problem [10],:

$$F_j = \lambda_j + \sum_{i=1}^{s-1} \frac{i}{s}(\lambda_{j-1} + \lambda_{j+1}), \tag{2}$$

where $s = \frac{ws+1}{2}$ is a "starting index" given by the window size (ws), L is the length of the protein and λ is the mean of the each amino acid b-factors given by the homologs protein we are using. When gaps etc

The final amino acid score was also scaled in a range of values between 0 and 1, in order to simplify the analysis.

Remark 1. It is important to note that with this approach we are losing the information of the first and last $s-1$ amino acids (according to the window size that is used).

The total flexibility score of the protein is given by:

$$F = \frac{1}{s(L-(ws-1))} \sum_{j=s-1}^{L-(s-1)} F_j \tag{3}$$

2.4 Physico-chemical restraints

In our analysis, we also take into account structural and hydrophobicity restrains to further validate our flexibility score. In this way, we made use of DSSP program [11] to save the corresponding

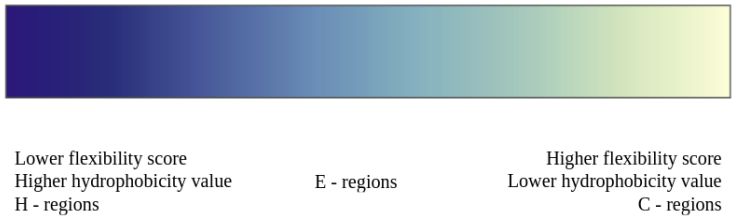
secondary structure, C - coil, E - extended (beta-strand) and H - helix, of our protein sequence retrieved from the AlphaFold repository [12, 13]. Moreover, we calculate the hydrophobicity score for each amino acid and for the global protein. To calculate the hydrophobicity scores by amino acid and the global hydrophobicity (GRAVY) we used the ProtParam module from Biopython.

From the literature [2, 14], we found a theoretical relationship between hydrophobicity and the type of secondary structure with the flexibility of a protein. On the one hand, hydrophobic regions tend to be in the core of the protein, resulting, in general, in a more rigid region. On the other hand, from the 3-state secondary structures (Loop, Helix and beta-strand), from the literature we extracted that beta-strands are, in general, more flexible than helix, for that reason, we associated a higher flexible score to beta-strands, being loops the structure with the highest flexibility score.

2.5 Visualization

Once we have the flexibility scores and physico-chemical restrains computed, they are plotted in a heat-map. The values of each feature amino acid are colored as a gradient, so that darker colors represent lower flexibility scores, higher hydrophobicity values and H regions, giving the user the idea that this region should be more rigid. In an opposite way, lighter colors represent higher flexibility scores, lower hydrophobicity values and C regions, representing the more flexible regions of the protein.

In detail, the secondary structures are just colored with a three-color palette representing each structure according to the theoretical flexibility. The following image shows the used palette with the labels corresponding to each score of the three parameters:



Moreover, we include a basic plot with the distribution of the flexibility amino acid scores along the protein sequence. This provides a quick look to those flexible regions and the global distribution of flexibility of the protein.

3 The Program

3.1 Prerequisites

Python Packages:

- `biopython` $\geq$ v.1.79
- `matplotlib` $\geq$ v.3.3.4
- `requests` $\geq$ v.2.25.1
- `pandas` $\geq$ v.1.4.1
- `seaborn` $\geq$ v.0.11.2
- `numpy` $\geq$ v.1.20

The above packages can be easily installed by executing the `requirements.txt` file:

```
pip install -r requirements.txt
```

External Programs. The following programs must be on the `$PATH`:

- NCBI programs (Blastp, Psiblast and makeblastdb) $\geq$ v.2.10.1+: Pairwise alignment and database creation.
- T-Coffee $\geq$ v.13.45.0: Multiple sequence alignment
- DSSP = v.2.3.0 (mkdssp): Secondary Structure assignment

The installation of the above programs can be automatically handle by executing the `requirements.py` file (linux users):

```
python3 requirements.py
```

3.2 Structure of the package: fancyflexscore-main

Description of the different modules and functions included into them.

- `__main__.py`: main execution of the program.
- `functions.py`: created functions.
- `requirements.py`: automatic installation of external packages (linux users).
- `requirements.txt`: automatic installation of python packages.
- `README.md`
- `LICENSE`
- `.gitignore`

3.3 Running the program

The user can obtain information about the program and the arguments that should write by:

```
python3 fancyflexscore-main/ -h
```

The arguments are:

- `-i INFILE`: Input FASTA file with one unique record containing the identifier and sequence of the protein of interest. The identifier must be a uniprot ID. Chain must be specified by: ¿uniprotID:chain. This argument is required.
- `-o OUTFILE`: Output filename without extensions. It generates two files: filename_results.csv and filename_visualization.pdf (default: results). Two more files are also generated: a PDB for homologs (located at structures/directory) and a MSA aln of target protein and PDB homologs ((located at structures/directory)). THis argument is optional.
- `-v verbose`: To have a follow up of the program (default: False). This argument is optional.
- `-ws window_size`: To change the window size when computing the flexibility score. The range of accepted values: 1,3,5,7,9,11,13. This argument is optional (default: 7).

Therefore, to run the program, the user should execute:

```
python3 fancyflexscore-main/ -i input_file -o output_file_name -ws window_size -v
```

4 Working Examples

4.1 Q9Y223

The Uniprot ID Q9Y223 corresponds to the *Homo sapiens* protein encoded by the gene GNE, a bifunctional UDP-N-acetylglucosamine 2-epimerase and N-acetylmannosamine kinase. It regulates and initiates the biosynthesis of N-acetylneuraminic acid (NeuAc), a precursor of sialic acids. It is also required for normal sialylation in hematopoietic cells.

Analysis of flexibility:

From the flexibility distribution plot we observe a main region of high flexibility from residues 600 to 630 (1). This region is found within the N-acetylmannosamine kinase domain, which comprises residues 406-722 of the protein. We could expect that this region with kinase activity has a higher flexibility than other regions since it is involved in the interaction with other proteins.

By conducting a deeper analysis of the results comparing against the PDB Flex server, we observe that the region of high flexibility is also found in the server as the region with the highest

Figure 1: Distribution of flexscores for Q9Y223

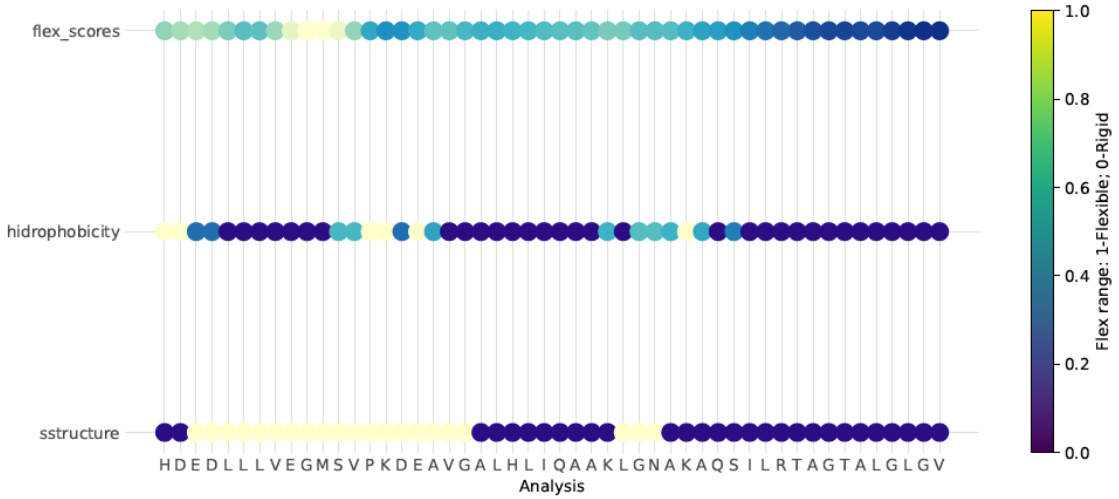


Figure 2: Flexscore peak heatmap region for Q9Y223

average RMSD as shown in the selected region. Furthermore, we see coincidence with the secondary structures of the PDB Flex.

4.2 P65206

The Uniprot ID P65206 corresponds to a protein-arginine kinase of *Staphylococcus aureus (strain N315)*. It catalyzes the specific phosphorylation of arginine residues in proteins, by using a molecule of ATP.

Analysis of the flexibility:

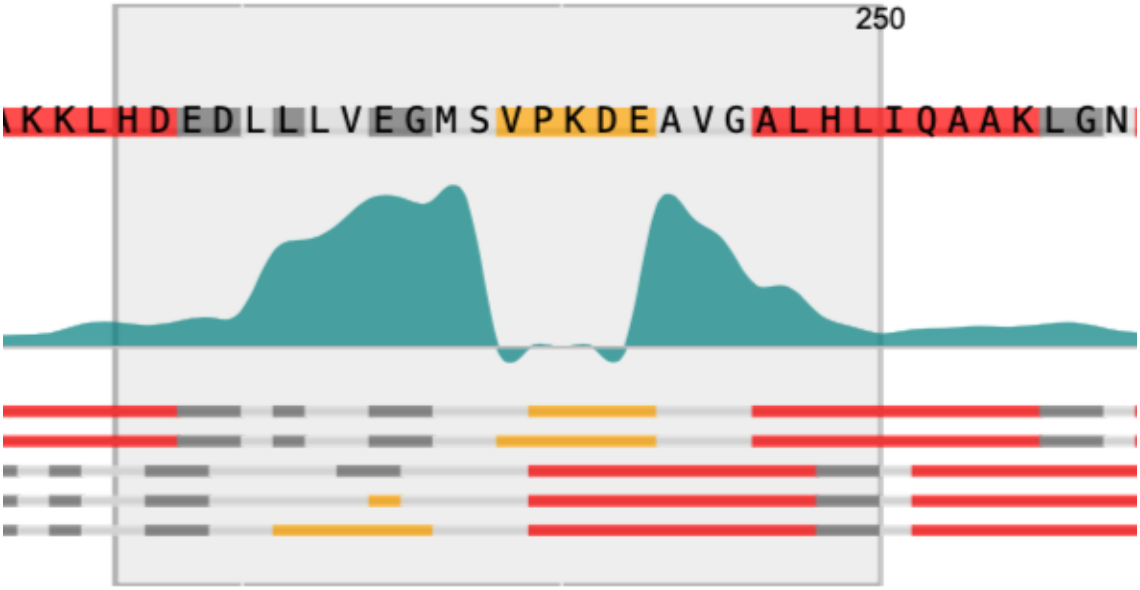


Figure 3: Highest average RMSD region for Q9Y223

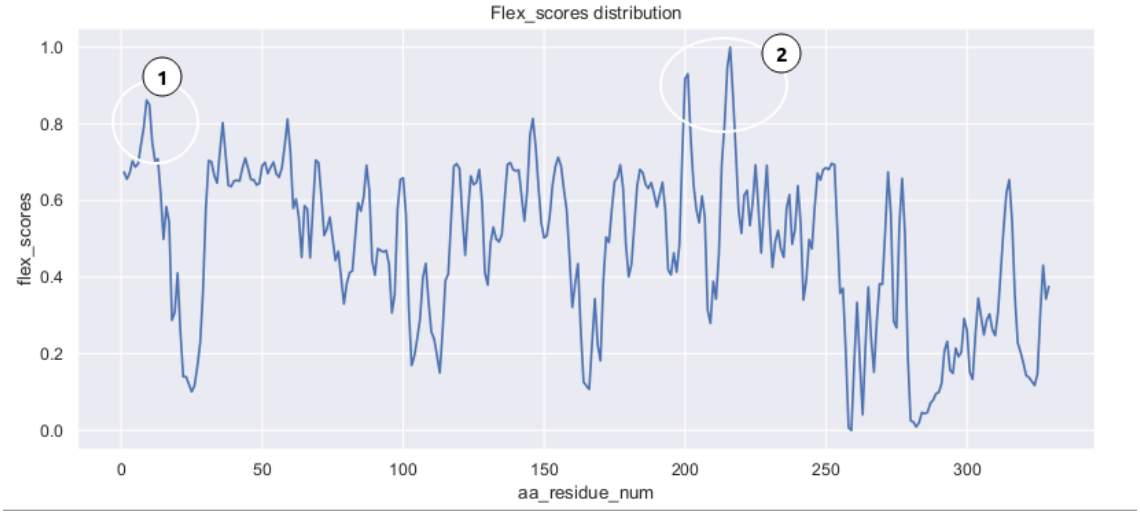


Figure 4: Distribution of flexscores for P65206. Two main regions of flexibility are highlighted

In the image, the two main regions of flexibility are highlighted. We could expect that the regions involved in the interaction with arginine residues and the ATP were highly flexible. According to Uniprot information, the nucleotide biding sites of the ATP are the followings:

- 24-28
- 166-170
- 197-202

Now, we will observe the flexscores for this regions:

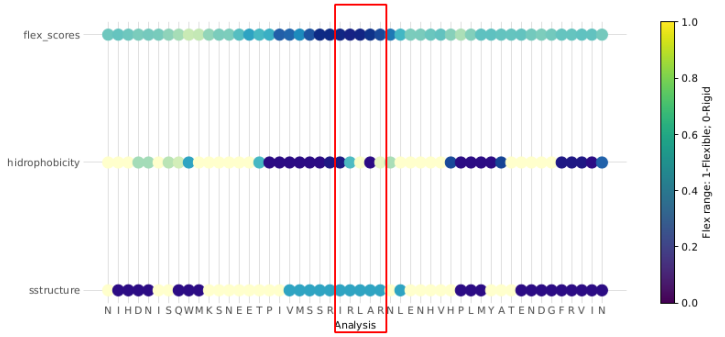


Figure 5: Flexscore peak 1 heatmap for P65206

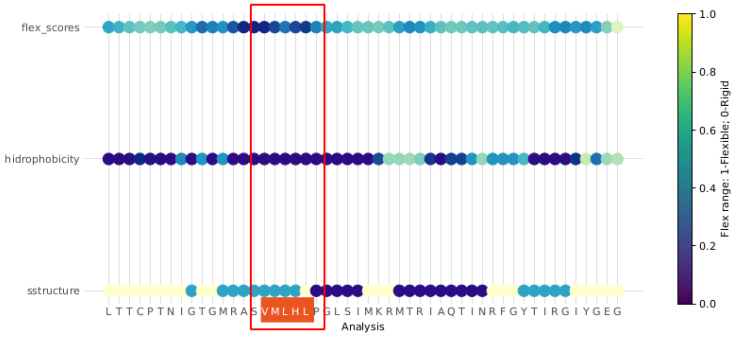


Figure 6: Flexscore peak 2 heatmap for P65206

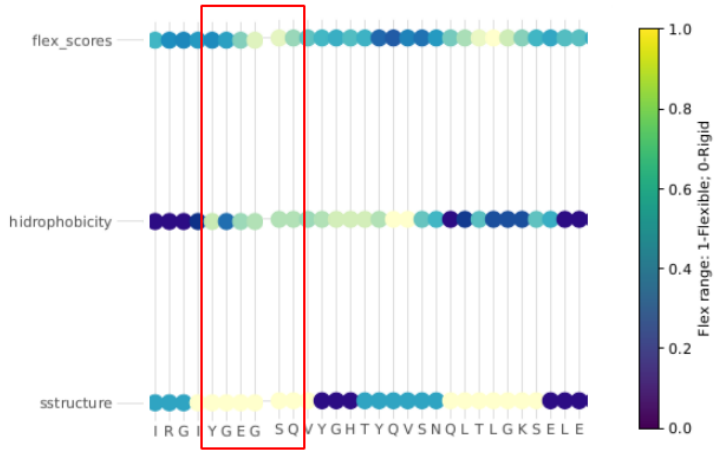


Figure 7: Flexscore peak 2 heatmap for P65206

And the RMSD distribution from the PDB Flex server:

With these results we see that the protein is quite flexible along the whole sequence. The regions found to be highly flexible in our distribution are also found in the PDB Flex server. Furthermore, we find concordance in the secondary structures obtained with the ones from the PDB.

From the previously mentioned ATP interaction, we obtain that some of the peaks of our

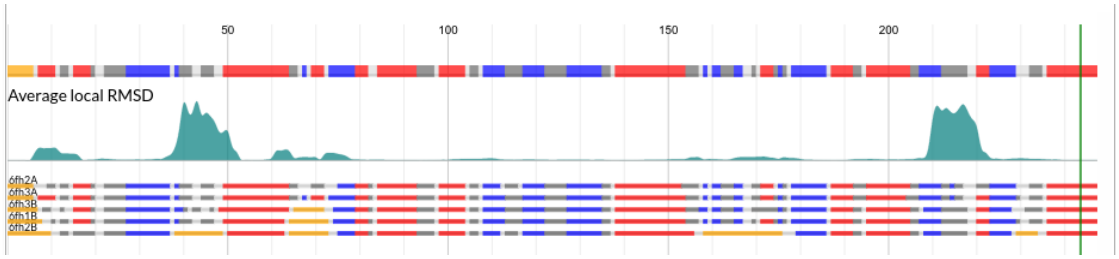


Figure 8: Average RMSD distribution PDB Flex for P65206

distribution correspond to those regions, meaning they could be expected to be flexible due to its interaction with ATP.

4.3 P04439

The Uniprot ID P04439 corresponds to the *Homo sapiens* protein encoded by the gene HLA-A, a HLA class I histocompatibility antigen, A alpha chain. It is an antigen-presenting major histocompatibility complex class I (MHCI) molecule. Its main function is the antigen presentation for the recognition by alpha-beta T cell receptor (TCR) of T cells, involved in the specific immune response.

Analysis of flexibility:

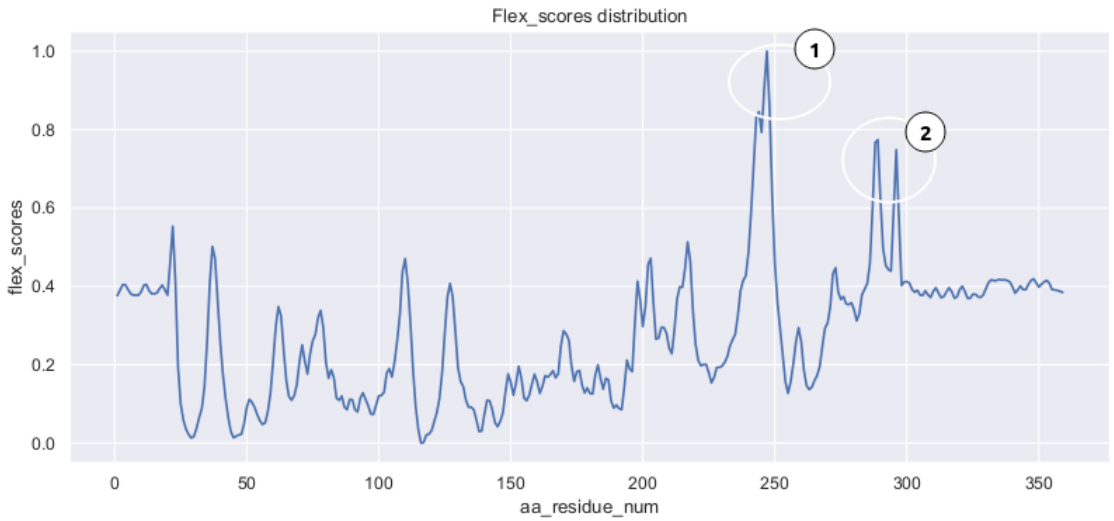


Figure 9: Distribution of flexscores for P04439. Two main regions of flexibility are highlighted

From the flexibility distribution plot we observe a main region of high flexibility from residues 230 to 250 (1). This region is found within the Ig-like C1-type domain, which comprises residues 209–295 of the protein. In detail, this region interacts directly with the CD8 coreceptor, which could explain the reason why this region is highly flexible. In addition, we find a highly flexible region from residues 280-297 (2), which is contained at the end of the previous domain and could

also be involved in the interaction.

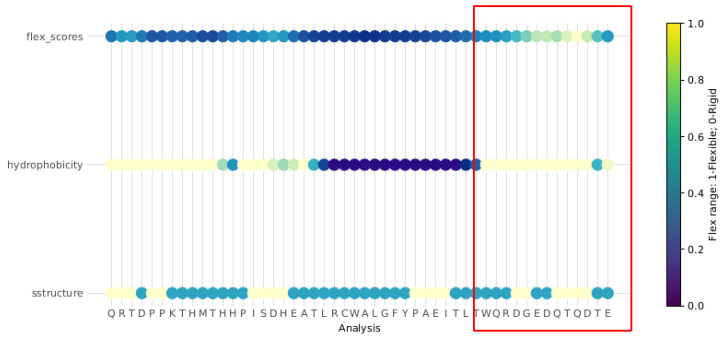


Figure 10: First region of flexibility for P04439

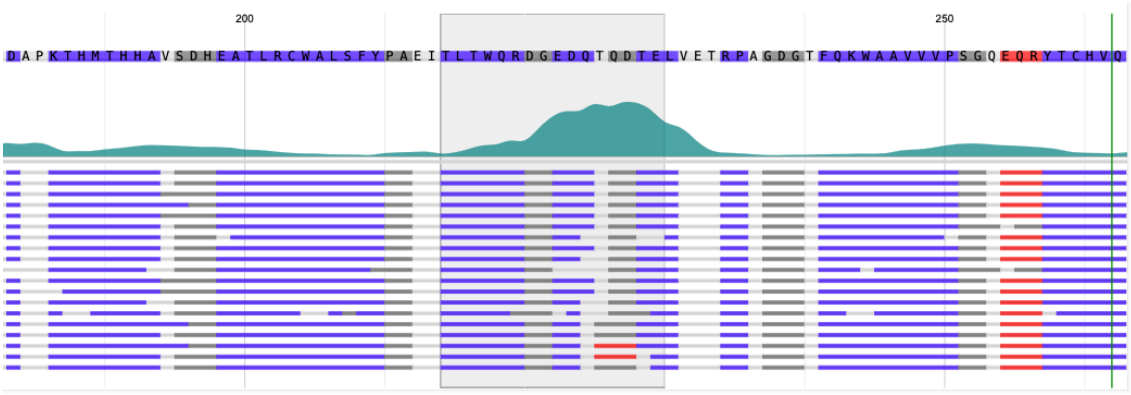


Figure 11: Highest average RMSD region for P04439

By conducting a deeper analysis of the results comparing against the PDB Flex server, we observe that the region of high flexibility is also found in the server as the region with the highest average RMSD as shown in the selected region. Furthermore, we also see coincidence with the secondary structures of the PDB Flex. Furthermore, we see that this region is more hydrophilic, which may translate in a more flexible region, as it has been observed.

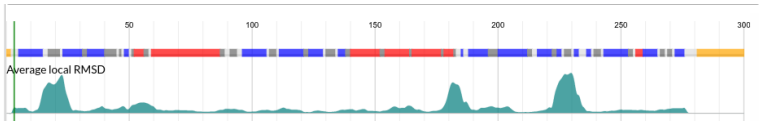


Figure 12: Average RMSD distribution PDB Flex for P04439

However, from the PDB Flex we see two other concrete highly flexible regions which are not spotlighted by our function.

4.4 Q6PML9

The Uniprot ID Q6PML9 corresponds to the *Homo sapiens* protein encoded by the gene SLC30A9, a Zinc transporter. It acts as a zinc transporter involved in intracellular zinc homeostasis. In addition to the transporter function, it plays a role in the transcriptional activation of Wnt-responsive genes. Furthermore, it functions as a secondary coactivator for nuclear receptors by cooperating with p160 coactivators subtypes

Analysis of flexibility:

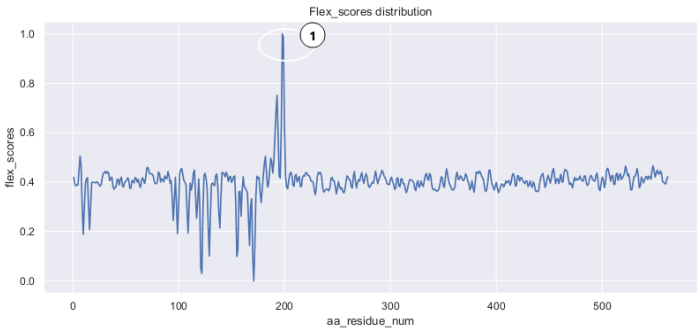


Figure 13: Distribution of flexscores for Q6PML9. The main region of flexibility is highlighted.

In this example, we are analysing a protein with no PDB structure and a high rate of loops. We wanted to demonstrate the results obtained by a query with less information. In this case, performing the alternative PSI-BLAST has been necessary.

We see an homogeneous distribution of the flexibility scores around 0.4 and a peak around the residue 200 (1). We could assume that this distribution is not very reliable, since it has been obtained from remote homologs. We found no remarkable information from the Uniprot record, since not special domains have been described. Taking into account the nature of the protein, since it is a transporter it makes sense that in general we find that the protein is homogeneously flexible. Furthermore, since it acts as an activator, it is supposed to interact with other proteins and maybe that region with high flexibility is important for this interaction.

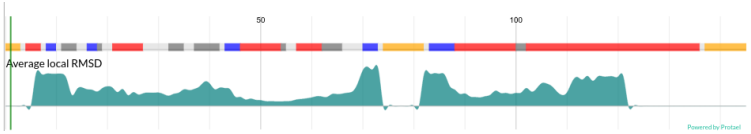


Figure 14: Average RMSD distribution PDB Flex for Q6PML9

By conducting a deeper analysis of the results comparing against the PDB Flex server, we observe that the results for this protein come from a cluster of 60(%), which means that they are also not very reliable. We also obtain an homogeneous flexibility along the whole protein shown. In this case, we will not compare the regions since the obtained protein is a remote homolog and no sequence coincidence is apparently found.

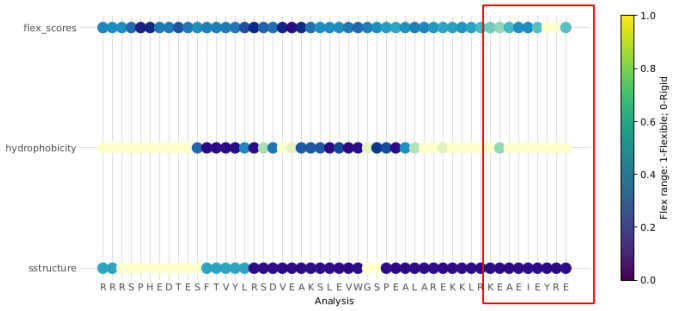


Figure 15: Flexscore peak heatmap region for Q6PML9

5 Limitations

- Many proteins have segments of unusually high mobility, which must be accounted for before normalization can be performed. Quality control studies was used to identify outliers.
- Using a window size of N to compute the weighted average of amino acid flexibility, implies that (N-1)/2 residues in each end of the protein are left without calculations. For instance, we have defined a window size of 7, giving no flexscore for the first and last 3 residues of the protein. From our point of view, there was no easy way to calculate the flexscore of those residues using the formula (2), as it will be difficult to manage different neighbour sizes for each residue. Future implementation must consider this drawback and provide a solution.
- Our approach is only based on b-factor values. We consider that this values should be corrected according the secondary structure and hydrofobicity scores.
- The retrieval of b-factor values comes from a MSA, meaning we obtained the b-factor where sequence coincidences were found. However, a proper approach would be account for sequence similarity, which would involve conducting a structural alignment for the retrieval of b-factors, since it would improve the accuracy.
- In MSA, regions of the target without homolog coverage are saved by retrieving b-factor values from the theory, but ideally, we should recover some homologs (distant perhaps) that do cover this section to obtain the b-factor.
- Our function is not suggested to work with protein fragments, but the whole protein. The main reason for this are the results obtained in the MSA, which in case of using a fragment, many gaps will be found and most part of the protein would be predicted from theoretical scores. It is also not suggested for whole protein complexes, since the considered only the most identical chains from the homologs are used.
- Our approach relies on PDB homologue with high resolution. Thus, in case of using a protein with few literature and poorly studied, it is less probable that close homologs are found and the function will mainly depend on theoretical b-factors, meaning the reliability of the scores would not be high.

6 Availability

The code of the program is available on: `https://github.com/G-Molano-LA/fancyflexscore`

References

[1] Vander Meersche, Yann, Gabriel Cretin, Alexandre G. de Brevern, Jean-Christophe Gelly, and Tatiana Galochkina. MEDUSA: Prediction of protein flexibility from sequence. *Journal of Molecular Biology.* 2021; https://doi.org/10.1016/j.jmb.2021.166882

[2] Teilum, K., Olsen, J.G. & Kragelund, B.B. Functional aspects of protein flexibility. *Cell. Mol. Life Sci.* 2009; **66**: 2231. https://doi.org/10.1007/s00018-009-0014-6

[3] Liu, Q., Li, Z. & Li, J. Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics 15.* 2014; **S3**. https://doi.org/10.1186/1471-2105-15-S16-S3

[4] Carugo, O. Detection of breaking points in helices linking separate domains. *Proteins: Structure, Function, and Bioinformatics.* 2001; **42(3)**:390-398.

[5] Smith, D.K., Radivojac, P., Obradovic,Z., Dunker, A.K. & Zhu, G. Improved amino acid flexibility parameters, *Protein Sci. .* 2003; **12**:1060-1072

[6] Vihinen, M., Torkkila, E., and Riikonen, P. Accuracy of protein flexibility predictions. *Proteins.* 1994; **19**: 141–149.

[7] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne, The Protein Data Bank, *Nucleic Acids Research.* 2000; **28, Issue 1,**:235–242, https://doi.org/10.1093/nar/28.1.235

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

[9] Notredame, C., Higgins, D. G., & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology.* 2000; **302(1)**: 205-217.

[10] Li Liu, Ying Fang, Jianhua Wu, Flexibility is a mechanical determinant of antimicrobial activity for amphipathic cationic α-helical antimicrobial peptides, *Biochimica et Biophysica Acta (BBA) - Biomembranes.* 2013; **1828, Issue 11**:2479-2486.

[11] Wouter G. Touw, Coos Baakman, Jon Black, Tim A. H. te Beek, E. Krieger, Robbie P. Joosten and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucl. Acids Res.* 2015; **43**: D364-D368.

[12] Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021.

[13] Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* 2021.

[14] Kopeć, K., Pedziwiatr, M., Gront, D., Sztatelman, O., Sławski, J., & Łazicka, M. et al. Comparison of α-Helix and β-Sheet Structure Adaptation to a Quantum Dot Geometry: Toward the Identification of an Optimal Motif for a Protein Nanoparticle Cover. *ACS Omega* 2019; **4(8)**: 13086-13099.