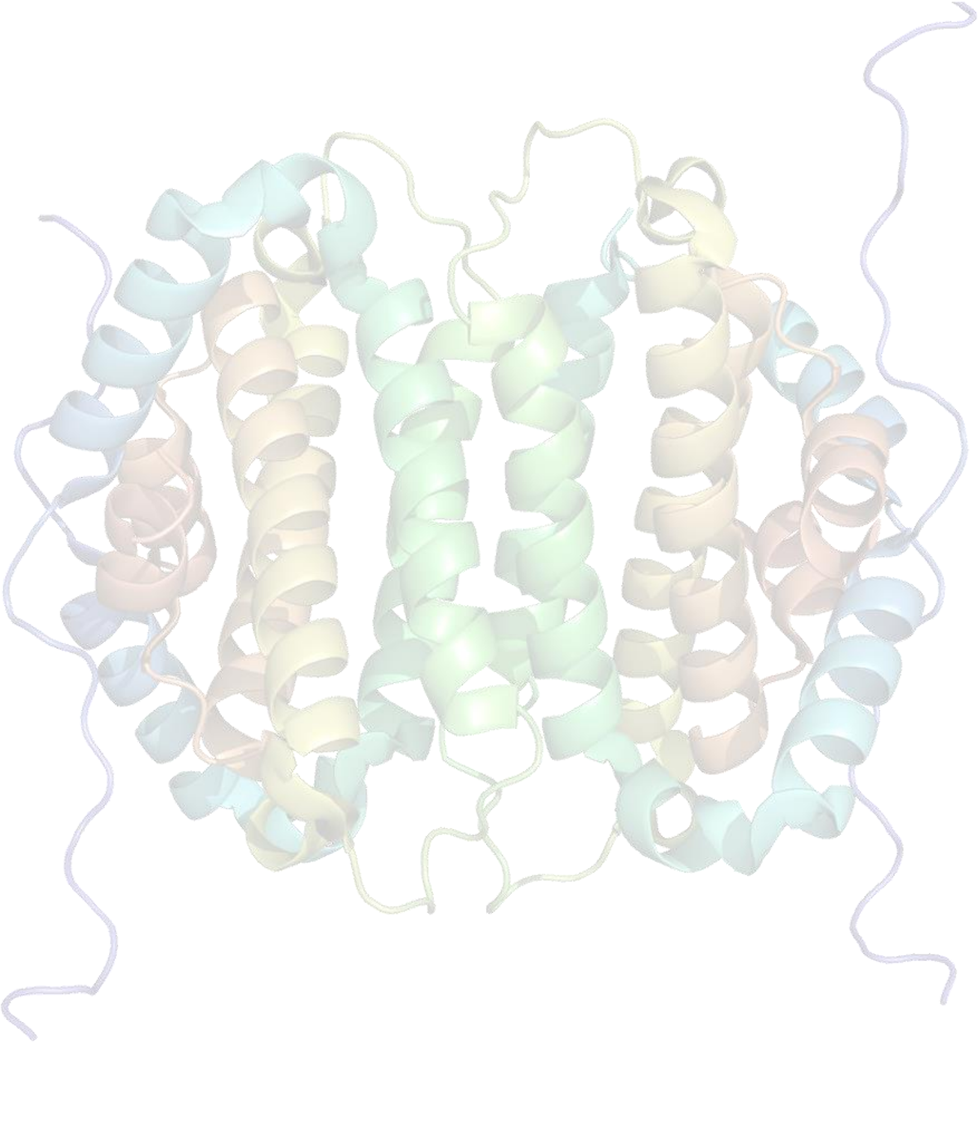
ANALYSIS OF PROTEIN FLEXIBILITY USING GNM TO PREDICT B-FACTOR-LIKE SCORES

Álvaro Ciudad Serrano, Clara Suárez Quintana, Júlia Vilalta Mor



SBI | PYTHON

MSc in Bioinformatics for Health Sciences

Apr, 2022

TABLE OF CONTENTS

INTRODUCTION

Proteins are dynamic elements

Proteins are macromolecules formed by amino acids connected through peptide bonds, where the carboxyl group of an amino acid binds to the amine of the alpha-carbon (αC) of another amino acid.

Proteins are the functional building blocks of biology, responsible for cell motility, structure building, signalling pathways, etc. In all these aspects, proteins have to interact with the environment, other proteins and elements such as membranes, nucleic acids. This is the reason why proteins cannot be static elements, as sometimes they are represented or thought of, but must be able to move in order to alternate between conformations that make interactions biologically significant.

The dynamism of proteins lies in their backbone, as well as in their side chains, mainly in the latter, which are in motion due to thermal and Brownian motion.

Protein flexibility

Protein structure can be obtained using different techniques, such as X-Ray crystallography or nuclear magnetic resonance (NMR). With NMR, the protein can be studied in an environment that represents more accurately the physiological state of the protein, but this technique is size-dependent and can only be used with small proteins.

Nevertheless, as described in previous studies, atomic details can also be retrieved from crystalline structures of high resolution, as some of the flexibility characteristics of proteins in solution are preserved in crystals. For instance, conformational disorder can be observed as the oscillation of atoms around their equilibrium positions.

In crystals obtained from X-rays, flexibility is represented by the atomic displacement parameters (ADP), which are also known as B-factors. These B-factors represent a decrease in diffraction that is due to two factors: 1) dynamic disorder caused by temperature-dependent vibration of the atoms, and 2) static disorder related to the orientation of the molecule.

Measuring flexibility: B-factors

B-factors are defined as the mean-square amplitude of displacements of the atoms around their equilibrium positions (u^2), averaged over the lattice:

$$B = 8\pi^2 \ \langle u^2 \rangle$$

Proteins with high B-factor-characterized regions have been found to have a higher flexibility than regions with lower B-factor, according to numerous studies. However, B-factors are not absolute measures and are affected by numerous parameters. Given two structures of the same protein, obtained with different resolution, B-factors of different values will be obtained. But not only is the resolution a factor to take into account. Elements affecting the values of the B-factors are listed below.

Resolution.

It has been shown that the higher the resolution at which crystals are obtained, the more reliable the B-factor that is generated. The explanation for this is based on the relationship between resolution and the ability to identify alternative atomic positions. If the resolution is high, alternative atomic positions can be identified individually and an appropriate refinement made. At low resolution, refinement is coarser and this leads to an increase in the values of the B-factors.

Refinement.

As indicated above, the refinement method affects the B-factor values obtained. When the diffraction data are sufficient, an anisotropic refinement can be carried out, which considers that the atomic displacements are not identical in the three directions of space. In the opposite case, which is most frequent, an isotropic refinement is performed, which tends to inflate the B-factors for technical reasons.

Occupancy.

The atomic occupancy reveals the presence of an atom in its mean position. It is shown as a value ranging from 0.0 to 1.0, taking the maximum value when the atom spends all the time in the same position during the production of the crystal, that is, when it is in the same conformation. On the other hand, values below 1 indicate that the atom spends only a fraction of the time in that position (indicating that there are different conformations that have been averaged into the final crystal).

Low occupancy values correlate with low values of the B-factors, because low occupancy means lower number of electrons and therefore lower displacement.

Limitations of B-factors

In the previous section we have mentioned some parameters that affect B-factors, which not only capture atomic displacement but also noise resulting from the determination of the crystals. Before attempting to draw conclusions about protein flexibility, some more limitations need to be considered in order not to make misleading extrapolations.

In recent years there has been an increasing trend in the average values of the B-factors obtained. The interpretation of electron density maps is relatively arbitrary and has gone through what can be considered "fashions". Nowadays, certain atoms are often included in the refinements which tend to increase the B-factors and which were not included in the past.

This can be very problematic and lead to incorrect interpretations, resulting in false structural or functional conclusions about proteins.

Many voices call for the establishment of a consensus threshold above which B-factors should be considered invalid, although this has not yet been achieved. In a recent publication a B_{max} has been defined and suggested as a threshold for omitting excessively large ADPs.

As mentioned above, B-factors are not absolute measures, so, tempting as it may be, they cannot be used to directly compare the flexibility of two structures. Before doing so, B-factors have to be standardised. Although different approaches are used (some specific for certain datasets, others more generic), the most widespread is the use of a standardised B-factor (B_n), using the following formula

$$B_n = (B - \langle B \rangle)/\sigma$$

where ⟨B⟩ is the average of the B-factor in the structure and σ, the standard deviation.

Predicting flexibility: other approaches

The conception of proteins as rigid structures is related to the determination of their structure by X-rays. However, as indicated above, proteins are far from being static elements. Their characterisation and study with other techniques, such as NMR, and using computational molecular dynamics (MD) methods helps to consider proteins as dynamic structures. And associated with dynamism is flexibility.

This is why the use of B-factors, obtained from crystallographic structures and which are not exclusive measures of flexibility, but also reflect intrinsic disorder and noise, is considered insufficient for assessing protein flexibility.

Other approaches, such as those indicated below, are increasingly used to try to obtain more reliable flexibility values.

RMSF.

The root mean-square deviation (RMSD) is a measure of the distance between two conformations calculated with their coordinates, for the same atoms in the same order in both. However, with RMSD, what is obtained is a global measure of the conformational stability of the structure during the simulation. To calculate flexibility, i.e. to identify the most mobile areas of the structure, the root mean-square fluctuation (RMSF) is used, which is calculated for each αC of the backbone according to the following formula

$$\rho_i = \sqrt{\langle (x_i - \langle x_i \rangle)^2 \rangle}$$

Where x_i represents the coordinates of atom i and ⟨x_i⟩ is the ensemble average position of i.

The higher the RMSF, the greater the divergence of that residual from the average, indicating high mobility and therefore, flexibility.

RMSF can be calculated from NMR chemical shift data, using the random coil index (RCI) to compute the RMSF accurately. It can also be computed from MD data, as well as from Elastic Network Models (ENM).

ENMs are models that describe protein dynamics. They represent the protein as a network of masses connected by springs (representing bonded and non-bonded interactions below a threshold), each mass (node) being a residue of the protein.

One of the simplest are the Gaussian network models (GNM), which consider that nodes are connected by identical harmonic springs. Those nodes separated by a distance above a certain cut-off distance are not connected. Despite the apparent simplicity of GNM, the results obtained with normal-mode analysis (NMA) have little to envy to other more complex approaches (such as energy-function potentials obtained from force fields).

A more detailed explanation about GNM can be found in the Methods section.

Structural Alphabets.

Historically, when analysing protein structures, a description of the protein based on three states has been used: alpha helices, beta sheets and coils. This description is rather coarse, but it can be improved by the use of structural alphabets, which provides a more detailed structure with information at the conformational level.

The flexibility information obtained by this approach lies in the fact that conformational variations of the backbone can be described as changes in the pattern of the structural alphabets.

A structural alphabet is a library of variable size *N* of structural prototypes, each prototype being representative of a local structure of the backbone of length *l*.

It is considered that any protein structure can be approximated in a very detailed way by combining prototypes.

One of the approaches to calculate flexibility using SAs has been recently described and employs one of the most widely used alphabets: the Protein Blocks. What they do is translate the three-dimensional structure of the protein into a sequence of PBs. To calculate flexibility, one must first obtain the multiple conformations by MD or Monte Carlo simulations. Then, PBs would be assigned to each of the conformations. Finally, the flexibility for each residue would be quantified with the following formula

$$N_{eq} = exp(-\sum_{x=1}^{16} f_x \ln(f_x))$$

Where *fx* is the frequency of PB *x* (*x* takes values from *a* to *p*). It computes the average of PBs in a given position in the conformers set.

According to this approach, regions that exhibit local conformational change will then be those that exhibit local flexibility.

Structural compliance.

We have seen that from NMR data, as well as from MD and ENM data, RMSF can be calculated and used to predict flexibility. However, a new approach using ENM simulations proposes a new measure of structural flexibility: Structural compliance.

According to the publication in which it is described, the structural compliance of a system is defined as the total resulting displacement in the direction of the force divided by the absolute value of the force.

The value of the structural compliance for each pair of residues is first calculated by applying equal and opposite forces to each of them. The pairwise structural compliance C of a pair of residues can be evaluated by calculating the total distance variation between the residues along the pulling direction, divided by the magnitude of the force acting on the nodes.

A compliance map of the entire protein can be generated by iterating the procedure over each pair of residues and it provides insight about the flexibility of the structure.

OBJECTIVE

In light of the above, we propose the prediction of flexibility of a protein from its sequence, using root mean-square fluctuations as B-factor-like scores, which can be compared with B-factors obtained from crystallographic structures after a normalisation step.

To obtain these scores, we use an approach based on GNMs, which allows us to simulate protein systems in a simpler and faster way than with MD and with good quality results.

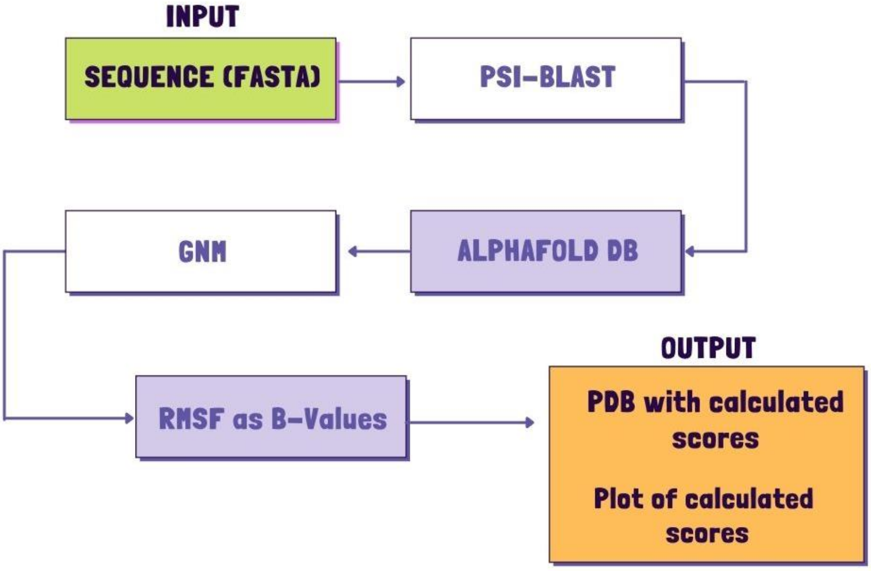
To do this, we integrate the following ideas into a program called **PyFlex**.

We use a sequence in FASTA format, which is used as input to carry out the calculations. With this sequence, a Psi-Blast is launched against the UniProt database to obtain homologues.

Once the homologues have been obtained, the best result is chosen based on the E-value and its structure is downloaded in PDB format from AlphaFold database.

Once the corresponding structure is obtained, we generate a GNM with our protein and obtain the mean square fluctuations for the alpha carbons.

The output is a PDB file with the simulated b-factors, as well as a plot of the B-factors-like scores versus protein length.



PYFLEX WORKFLOW

The specifics about PSI-BLAST and GNM are discussed in the Methods section.

METHODS

In this section we detail the mathematical basis behind the Gaussian Network Models, as well as the computations performed to calculate the RMSF.

PSI-BLAST

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) is a tool to compare a query sequence against a database of sequences with the aim of identifying local alignment regions. Alignments with homologous and distantly homologous sequences are given a score. It is an iterative process and with the results of the first round, an MSA with the highest scores is generated and a PSSM (position specific substitution matrix) is constructed. This matrix is used instead of BLOSUM62 in the following rounds of PSI-BLAST.

In our case, we use PSI-BLAST with the default iterations (5), against the UniProtKB/Swiss-Prot database, with BLOSUM62 as the initial matrix and an E-value cut-off of 10.0.

The best-scoring result from the PSI-BLAST is considered the best candidate and used to download its corresponding structure from the AlphaFold database.

Another PSI-BLAST is launched against the PDB database in order to obtain the homologous proteins and use the top ten results to obtain the average of them.

Gaussian Network Model (GNM)

The idea behind the use of GNM is to obtain a minimalist simplification of protein systems in order to study protein dynamics, keeping in mind that proteins in native state have access to a spectrum of motions (also referred to as “modes”). In GNM, proteins are represented by a network. The nodes are defined by the αC of the protein backbone, while the springs that connect them are representative of the bonded and non-bonded interactions between the nodes, taking into account a cut-off distance that is normally set at 7 Ä.

GNM assume that the fluctuations between the equilibrium positions of a node and a node position i are isotropic and Gaussian, and can be defined by the following vector

$$\Delta R_i = R_i - R_0$$

Where R_0 indicates the equilibrium position and *Ri*, an instantaneous position.

With this in mind, the fluctuations in the distance vector R_{ij} between two residues *i* and *j* are expressed as follows:

$$\Delta \boldsymbol{R}ij = \boldsymbol{R}ij - \boldsymbol{R}ij_0 = \Delta \boldsymbol{R}j - \Delta \boldsymbol{R}i$$

By considering the fluctuations to be isotropic (see *Predicting flexibility: B-factors: Refinement*), the potential of the network of N nodes (protein of N residues), V_{GNM} can be written in relation to the three spatial coordinates as follows:

$$V_{\mathrm{GNM}} = \frac{\gamma}{2}\left[\sum_{i,j}^{N} \Gamma_{ij}[(\Delta X_i - \Delta X_j)^2 + (\Delta Y_i - \Delta Y_j)^2 + (\Delta Z_i - \Delta Z_j)^2]\right]$$

$\boldsymbol{\Gamma}_{ij}$ represents the ijth element of the Kirchhoff matrix or connectivity matrix, which represents the contacts between residues as:

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } \mathrm{R}_{ij} \leq r_c \\ 0, & \text{if } i \neq j \text{ and } \mathrm{R}_{ij} > r_c \\ -\sum_{j,j\neq i} \Gamma_{ij}, & \text{if } i = j \end{cases}$$

In the Kirchhoff matrix, the diagonal values represent the coordination numbers of the nodes (residues). With the Kirchhoff matrix, all the necessary information about the network is obtained and the modes of motion of the structure can be evaluated.

To do this, Normal Mode Analysis (NMA) is done. It consists in a linear approximation of the motion followed by an eigenvalue decomposition of the Kirchhoff matrix to extract the global (low frequency, soft) and local modes (high frequency, stiff).

Applying calculations, beyond the scope of this analysis, the mean square fluctuations of residues and the correlation between residue fluctuations can be computed from the potential of the network as follows:

$$\langle \Delta \mathbf{R}_i^2 \rangle = \frac{3k_{\mathrm{B}}T}{\gamma}(\boldsymbol{\Gamma}^{-1})_{ii}$$

$$\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle = \frac{3k_{\mathrm{B}}T}{\gamma}(\boldsymbol{\Gamma}^{-1})_{ij}$$

Where k_B is the Boltzmann constant and T the absolute temperature. Γ^{-1} is the inverse of the Kirchhoff matrix and γ is the force constant taken to be uniform for all network springs.

In our case, when setting the GNM, we calculate all the modes from the αC from the backbone, using a γ value of 1.0, and a cut-off distance of 7.0.

TUTORIAL

A detailed walkthrough explaining the download, installation and functionalities of our implementation, PyFlex, is presented below.

PYFLEX

INSTALLATION

1. Download the archive file

Please, download the files following this LINK. In the PyFlex repository, click on the download button and save the .tar file onto your computer.

Please, notice that this software has been implemented to work with Linux OS.

2. Extract the files

In the terminal, type the following commands to extract the files and access the newly created directory.

```
tar -xvf PyFlex-0.4.tar.gz
cd PyFlex-0.4/
```

Inside, three folders are found: *PyFlex.egg-info*, *pyflex*, *blastscript*

Along with six files: *setup.py*, *setup.cfg*, *PKG-INFO*, *main.py*, *main_no_graph.py*, *Q12851.fasta*

3. Install all required modules

To install all the dependencies with the adequate versions, run the following code. It is recommended to use Python 3.9.

```
python3 –m venv ./pyflex_venv
source ./pyflex_venv/bin/activate
python3 setup.py install
```

It will load a new environment which will suffice for the program to run.

In case an error message is raised, you can install the following dependencies by hand.

```
sudo apt install libxkbcommon-x11-0
sudo apt install libxcb-xinerama0
```

When it finishes, three new folders will appear: *build*, *dist*, *outputs*

We need to give, in case they do not have them already, write permissions to the *outputs* folder for the outfiles of the analysis to be stored inside.

```
chmod a+wrx outputs/
```

USAGE

To run the program, make sure you are in the directory where the *main.py* file is located. The usage is displayed with the following command:

```
python3 main.py –h
```

In the terminal, the following will be shown:

```
usage: main.py [-h] [-I INFILE] [-o OUTFILE] [-g]

This program scores protein flexibility based on Elastic Networks. –I is used
for input and –o is used for the output name of the files.

optional arguments:
  -h, --help            show this help message and exit
  -i INFILE, --input INFILE
                        Input FASTA file/Sequence
  -o OUTFILE, --output OUTFILE
                        Output file
  -g, --graph           Use a graphical interface to run the app
```

In case the graphical interface raises some error due to interferences with system libraries, you can run the *main_no_graph.py*, which will only allow for the terminal version but with the same functionalities.

EXAMPLE

In the PyFlex-0.1 directory there is a FASTA file named *Q12851.fasta*. We will use it to show the functioning of the program, as well as the output that it generates.

Terminal

```
python3 main.py --input Q12851.fasta
```

Once the job is submitted, messages indicating the current step in the process will appear, as follows:

```
Psi-blast is running...
JobId: psiblast-R20220412-201133-0294-40437628-p1m
@> 6446 atoms and 1 coordinate set(s) were parsed in 0.14s.
@> Kirchhoff was built in 0.05s.
@> 819 modes were calculated in 0.17s.
Psi-blast is running...
JobId: psiblast-R20220412-201254-0668-7064250-p2m
```

```
@> 6446 atoms and 1 coordinate set(s) were parsed in 0.13s.
@> 6446 atoms and 1 coordinate set(s) were parsed in 0.19s.
Succesfully finished the job, output is in /outputs.
```

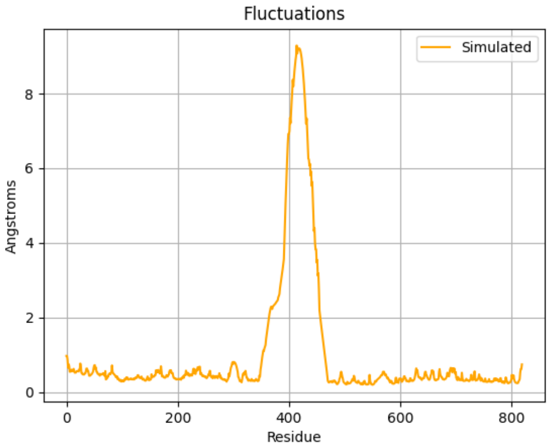
If we check in the *outputs/* directory, three files will have been created:

output_scaled.pdb, *output_scores_graph.png*, *output_simulated.pdb*

Since no name was indicated for the outfiles, they have been created using the default one.

The .pdb files, contain the typical coordinate's information, along with the B-factor-like scores that we predicted with our program, standing as the last column of the files.

The .png file is a plot of the B-factor-like scores by residue across the protein sequence.



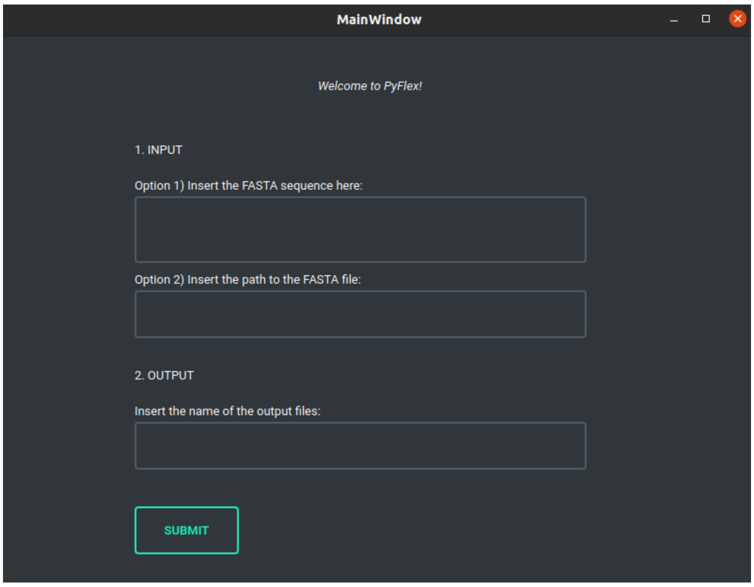
This three files are the outputs of the program.

Graphical interface

If you choose to run the graphical version, here is shown how and what to expect.

```
python3 main.py -g
```

A window like this will appear:



Welcome to PyFlex!

1. INPUT

Option 1) Insert the FASTA sequence here:

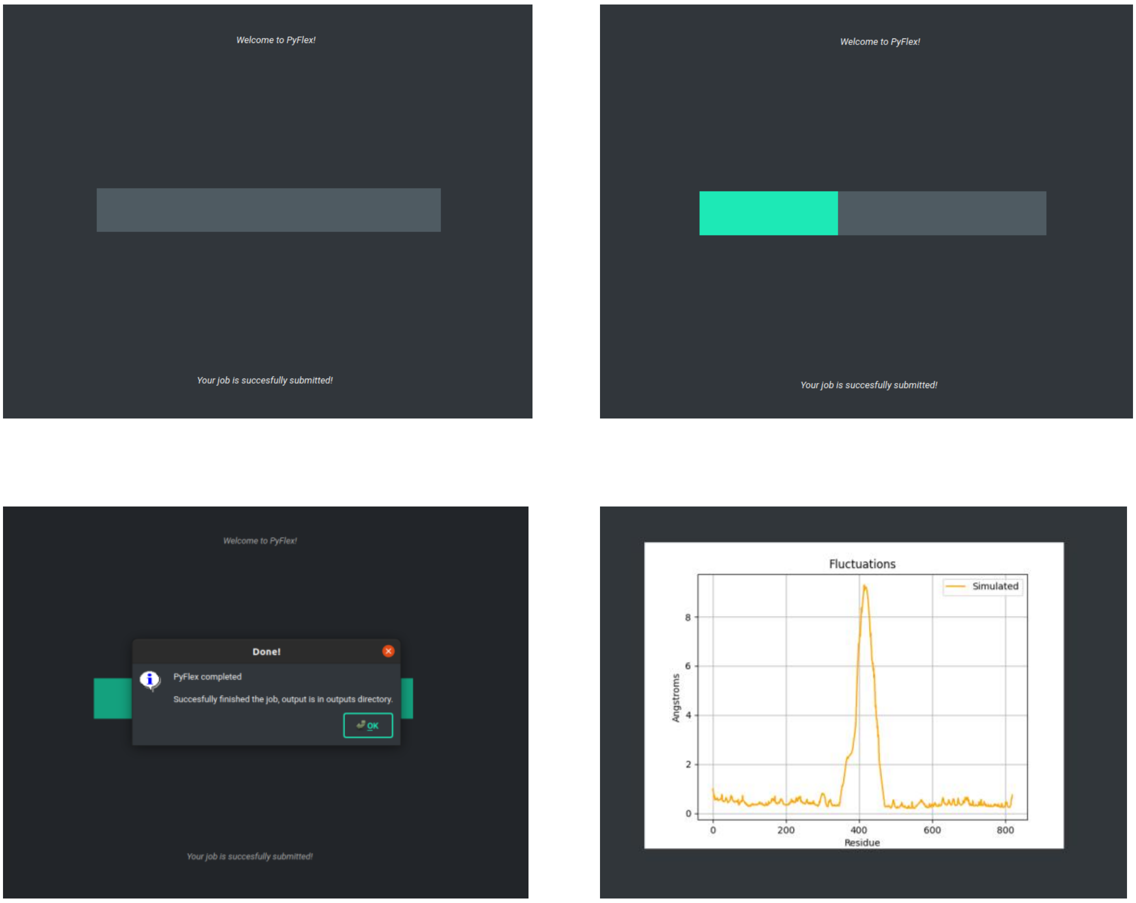
Option 2) Insert the path to the FASTA file:

2. OUTPUT

Insert the name of the output files:

You can enter a FASTA sequence on the first slot or provide the path to an existing file on your computer. The output name is not compulsory. If none provided, the default name will be used.

Once the job is submitted, you will get a progress bar which will be running until the job is finished. When finished, the result plot will appear in the same window.


Welcome to PyFlex!

Your job is succesfully submitted!

Welcome to PyFlex!

Your job is succesfully submitted!

Welcome to PyFlex!

Done!

PyFlex completed

Succesfully finished the job, output is in outputs directory.

Your job is succesfully submitted!

Figure 1. Graphical interface appearance through the job submission.

DISCUSSION

In order to check the accuracy of the predictions of our program, we perform and analysis and discussion of a set of proteins which is presented in this section. The flexibility of a protein is represented by these plots that contain on the x-axis all the residues of the protein along with the simulated flexibility (in Angstroms) on the y-axis.

P06401

This protein is the human progesterone receptor. As a receptor, it would be expected a high flexibility in regions involved in ligand recognition due to the necessary changes in conformation needed for both structures to interact properly.

The result returned by our program shows high flexibility indexes in the initial part of the protein (0-200 residues), as well as between 400-650.

When we get information on the regions and domains of the protein, we observe the following:

- The initial region of the protein 1-164 contains one activation function (AF3) involved in transcription regulation in progesterone receptors of the isoform B. Also, the region between residues 1-157 is recorded in UniProt as "disordered", which could be actually increasing the values of the indexes. In fact, if we check the structure, we observe that this region only contains one small helix, while the rest has no defined structure.
- The increased flexibility indexes between residues 400-650 could be explained as follows: that region corresponds also to one activation function (AF1), and two zinc fingers take part of the DNA binding region, where the linkers between structures and superstructures are usually flexible.

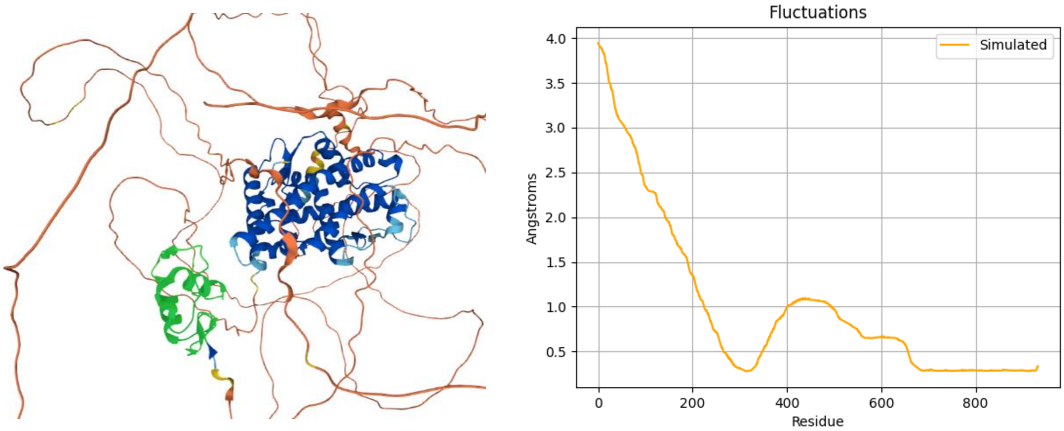


Figure 3. On the left, a capture of the structure of P06401 with the zinc-fingers coloured in green. On the right, the plot returned by PyFlex.

Q9P7Q4

This is a yeast protein involved in vesicle-mediated transport. According to the UniProt definition, it is necessary for the transport of proteins to the different compartments of the Golgi apparatus.

The plot obtained by our program shows very low indexes of flexibility, around 0 throughout the entire protein with the sole exception of the first 50-60 residues, which have very high indexes (above 15 Ä).

In UniProt this region is described as "disordered", and, when checking the structure using a viewer software, it can be seen that this region has no established structure, resembling a loop, while the rest of the protein is pretty structured, with helices and secondary structures, which are known to stabilize proteins and, in this case, cause the reduction of flexibility.

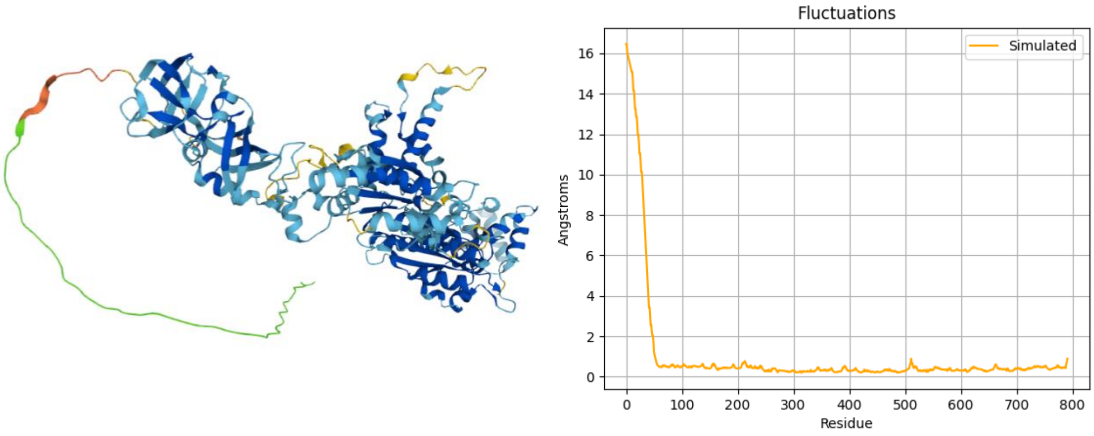


Figure 4. On the left, a capture of the structure of Q9P7Q4 with the initial disordered region coloured in green. On the right, the plot returned by PyFlex.

Q9VVG4

This is a protein from Drosophila melanogaster, its complete name being *Exocyst complex component 1*. It is a component of the exocyst, an octameric protein involved mainly in the tethering of vesicles from the Golgi apparatus to the cell membrane. It has a coiled coil comprising around 100 residues between residues 156-269.

The flexibility indexes observed in the plot returned by our program show higher values in the first 150 residues of the protein, while a steep decrease is observed between 150 and 200 that results in a stabilization of the indexes below 0.5 angstroms for the rest of the sequence. A couple of small peaks are observed around residues 480 and 550.

When we get information on the regions and domains of the protein, we observe the following:

- The initial region of the protein (until residues around position 150) is quite disordered, with loops surrounding a small helix. This could explain the high values of flexibility observed.

- Starting around 150 we observe a coiled-coil that afterwards is continued with a concatenation of different helices. This structured regions play different roles in different proteins, but in some structures, such as that of tropomyosin, coiled-coils have been pointed out as responsible for the stiffness of the structure. So we could think that they are actually playing a role increasing rigidity in this structure.
- The other two small peaks of flexibility remarked before (around 480 and 550) seem to be related to linkers or loops between helices in the structure, which are known for their flexibility.

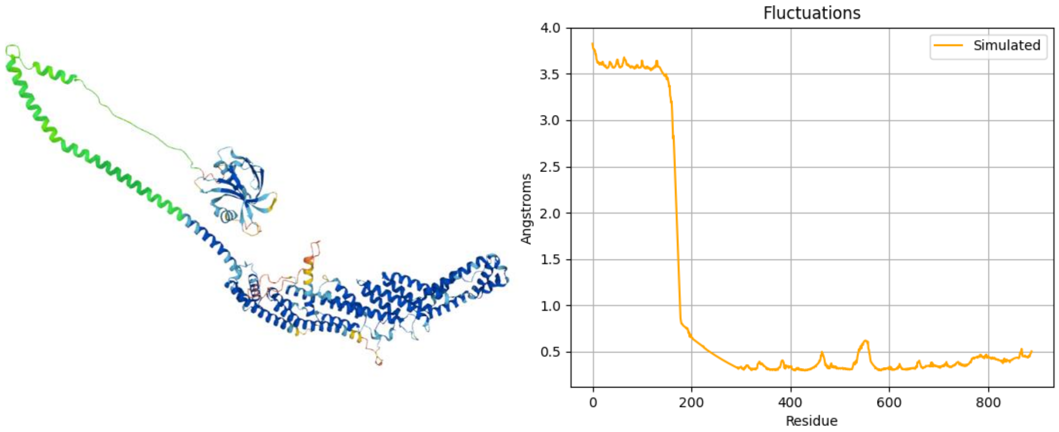


Figure 5. On the left, a capture of the structure of Q9VVG4 with the coiled-coil region coloured in green. On the right, the plot returned by PyFlex.

P38401

This protein is the *Guanine nucleotide-binding protein G(i) subunit alpha-1*, and functions as transducer downstream of G protein-coupled receptors (GPCRs) in numerous signaling cascades.

The high peak of flexibility that is observed in the plot can be due to the alpha-helix located at the start of the protein. Although it is a secondary structure and it is supposed to be more stable, the helix is not interacting with the rest of the protein, so it can have free different orientations that causes the flexibility of this part to be higher. However, the rest of the protein is interacting with each other forming a globular part that will be more stable due to the interactions between residues. In all this part we can observe different flexibility fluctuations around 0.5 and 1.5 angstroms approximately, the difference between this variety can be due to the parts containing secondary structure or the parts containing loops, that will show lower or high flexibility, respectively.

Furthermore, the parts with high flexibility can be related with the functionality of the protein, as it acts as a molecular switch inside cells, and is involved in transmitting signals from a variety of stimuli to the inside of the cell. It interacts with different nucleotides as well as binds GTP.

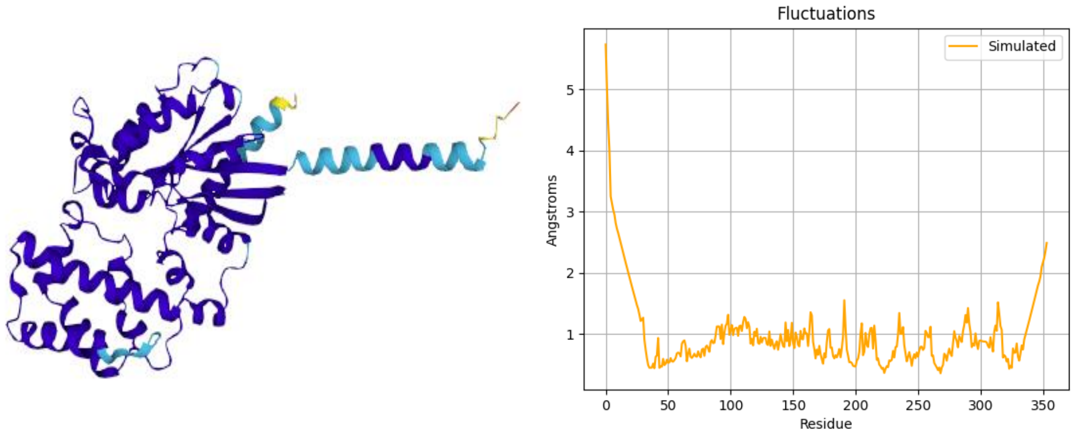


Figure 6. On the left, a capture of the structure of P38401. On the right, the plot returned by PyFlex.

P11433

This protein is the *Cell division control protein 24*. It promotes the exchange of CDC42-bound GDP by GTP, and also CDC24 may be involved in the initial selection and organization of the budding site. The functionality can also explain its structure.

If we look at the protein structure, we can see three large loops in the middle of the protein as well as the loops located at the starting and end residues, probably caused by some disordered regions as we saw in other proteins. Thus, this is a clear example that the structure is related to the flexibility. The parts that contain the loops show high flexibility and are represented by peaks in the fluctuation plot. Depending on the length of each loop, the flexibility in angstroms will be more or less proportional. On the other hand, the rest of the protein show low flexibility and it corresponds to the parts containing secondary structure, such as beta sheets or alpha helices.

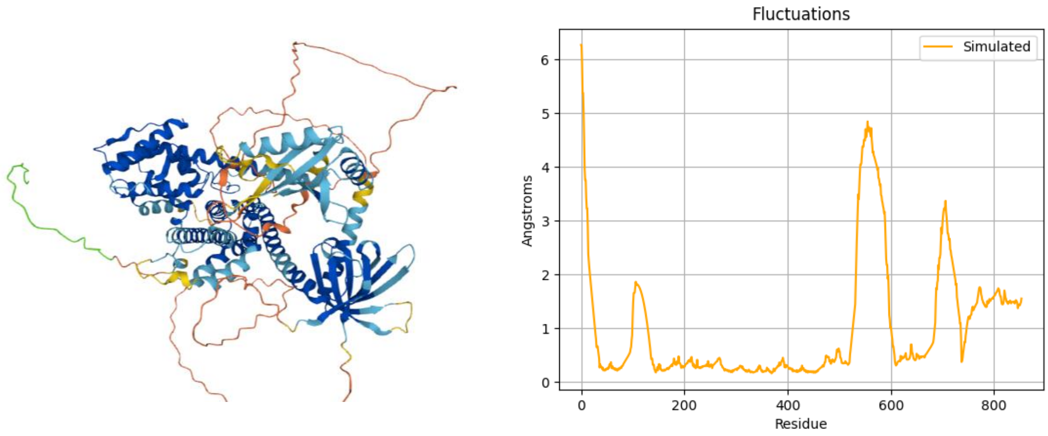


Figure 7. On the left, a capture of the structure of P11433, with the first disordered region coloured in green. On the right, the plot returned by PyFlex.

Q12851

This protein is a Mitogen-activated protein kinase 2, and we have used it as an example for PyFlex. The most interesting feature of this protein is that it works as a hinge. For this reason, there is a high peak of flexibility in the middle of the protein, where the rest of the residues maintain their flexibility score between 0.5 and 1 angstroms approximately, very different from 9 angstroms in the hinge.

It acts as an essential component of the MAP kinase signal transduction pathway. If we look at its structure, we can prove that the hinge with high flexibility belongs to a large loop, whereas the rest of the protein has almost a stable secondary structure. This loop is the one that allows the protein to act as a hinge. Thus, the structure makes sense with the simulated flexibility results.

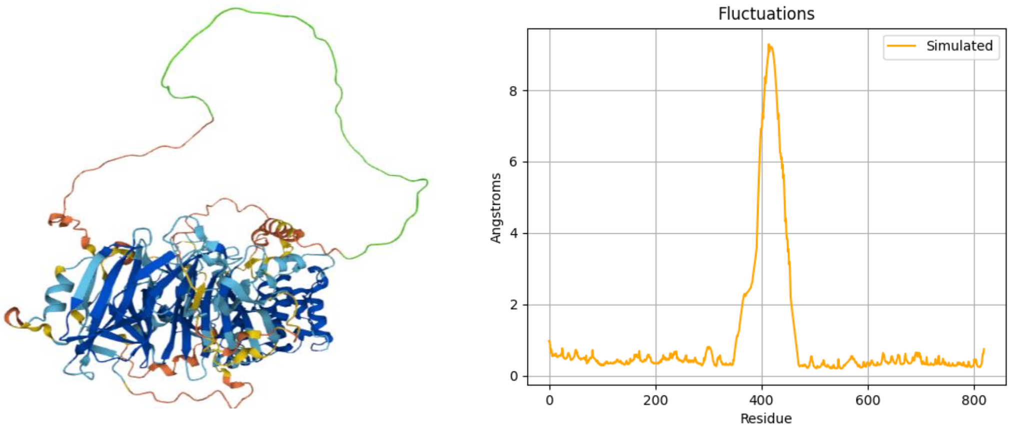


Figure 8. On the left, a capture of the structure of Q12851, with the hinge coloured in green. On the right, the plot returned by PyFlex.

P24062

Finally, we also computed our approach with a larger protein in order to check its performance, as all the proteins mentioned above contained between 350 and 800 residues approximately. The selected protein is the Insulin-like growth factor 1 receptor, consisting of almost 1400 residues. This protein binds IGF1 with high affinity and IGF2 and insulin (INS) with a lower affinity. Thus, it contains two different binding sites, one for nucleotide-binding and one for ATP binding. Both the structure of the protein and its functionality are important in order to understand their flexibility.

As we can see in the simulated flexibility plot, the flexibility varies among the protein. To begin with, both extremes showing high flexibility belong to two loops, as the lack of secondary structure increases the number of possible conformations, and thus, the flexibility. The same case happens with the peak represented in the middle of the plot, which can belong to a hinge-like functionality, as the previous case. Furthermore, the flexibility in the final residues are higher than the one in the first residues, on both sides of the centred peak. This is due to the secondary structures, as they are located differently along the protein.

Finally, we can prove that our approach can accurately predict the flexibility also in large proteins, but as a limitation, we can see that the plot could have more resolution if it would be wider.

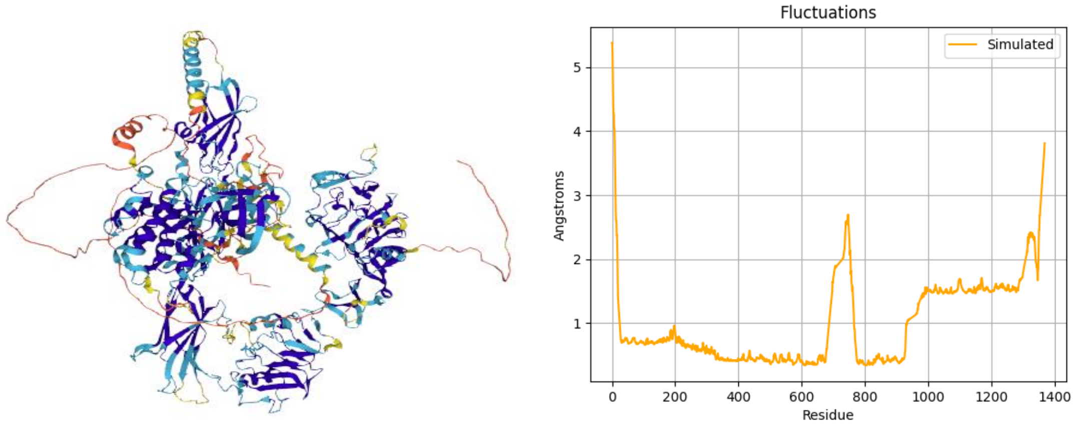


Figure 9. On the left, a capture of the structure of P24062. On the right, the plot returned by PyFlex.

MEDUSA

When comparing the PyFlex prediction results with the results predicted by Medusa web server, we can prove that our approach works well but it is difficult to compare with the other method because we have a large range of flexibility measured by angstroms, whereas Medusa uses a 3 or 5 classes of flexibility for each residue. However, this comparison is useful in order to get a clear idea of the accuracy of our method, and as a result we can see that in all proteins the results more or less are the same, although the comparison could not be precise. This is due to our method being highly sensitive to outliers, as low flexibility amino acids inside high flexibility regions are masked as also highly flexible.

Medusa results are attached in ***Supplementary Material*** for method performance evaluation if desired.

BIBLIOGRAPHY

Bhagwat M, Aravind L. PSI-BLAST Tutorial. In: Bergman NH, editor. Comparative Genomics: Volumes 1 and 2. Totowa (NJ): Humana Press; 2007. Chapter 10.

Bramer D, Wei GW. Blind prediction of protein B-factor and flexibility. J Chem Phys [Internet]. 2018;149(13):1–13. Available from: http://dx.doi.org/10.1063/1.5048469

Carugo O. Atomic displacement parameters in structural biology. Amino Acids. 2018;50(7):775–86.

Carugo O. How large B-factors can be in protein crystal structures. BMC Bioinformatics. 2018;19(1):1–9.

Craveur P, Joseph AP, Esque J, Narwani TJ, Noël F, Shinada N, et al. Protein flexibility in the light of structural alphabets. Front Mol Biosci. 2015;2(MAY):1–20.

De Brevern AG, Bornot A, Craveur P, Etchebest C, Gelly JC. PredyFlexy: Flexibility and local structure prediction from sequence. Nucleic Acids Res. 2012;40(W1):317–22.

Kmiecik S, Kouza M, Badaczewska-Dawid AE, Kloczkowski A, Kolinski A. Modeling of protein structural flexibility and large-scale dynamics: Coarse-grained simulations and elastic network models. Int J Mol Sci. 2018;19(11).

Li H, Chang YY, Yang LW, Bahar I. iGNM 2.0: The Gaussian network model database for biomolecular structural dynamics. Nucleic Acids Res. 2016;44(D1):D415–22.

Narwani TJ, Etchebest C, Craveur P, Léonard S, Rebehmed J, Srinivasan N, et al. In silico prediction of protein flexibility with local structure approach. Biochimie. 2019;165:150–5.

Rader AJ, Chennubhotla C, Yang LW, Bahar I. The Gaussian network model: Theory and applications. Norm Mode Anal Theory Appl to Biol Chem Syst. 2005;41–64.

Reinknecht C, Riga A, Rivera J, Snyder DA. Patterns in protein flexibility: A comparison of NMR “ensembles”, MD trajectories, and crystallographic B-factors. Molecules. 2021;26(5).

Scaramozzino D, Khade PM, Jernigan RL, Lacidogna G, Carpinteri A. Structural compliance: A new metric for protein flexibility. Proteins Struct Funct Bioinforma. 2020;88(11):1482–92.

Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins Struct Funct Genet. 2005;61(1):115–26.

Sonavane S, Jaybhaye AA, Jadhav AG. Prediction of temperature factors from protein sequence. Bioinformation. 2013;9(3):134–40.

Sun Z, Liu Q, Qu G, Feng Y, Reetz MT. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. Chem Rev. 2019;

Vander Meersche Y, Cretin G, de Brevern AG, Gelly JC, Galochkina T. MEDUSA: Prediction of Protein Flexibility from Sequence. J Mol Biol [Internet]. 2021;433(11):166882. Available from: https://doi.org/10.1016/j.jmb.2021.166882

Yang L, Song G, Jernigan RL. Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci U S A. 2009;106(30):12347–52.

Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins Struct Funct Genet. 2005;58(4):905–12.