INTRODUCTION

Aminoacid flexibility in proteins have been studied using several scores and analysis. The most usual approach is crystallographic B-factors (1). Another aproaches are NMR spin relaxation data, where a parameter called S^2 is used (2), molecular dinamics using RMSF as a mesure of flexibility (3) and Conformational entropy (4).

In order to have a method to predict protein flexibility these parameters have been used in diferent ways even combinated to develop a program for this purpose. However each parameter has its own advantages and disadvantages. To use crystallographic b-factor as predictor for flexibility we need to have some considerations. First one, resolution of the X-ray structure. A resolution below 1,5 Amstrongs biases B-factors having overestimated values (5).

Using NMR to predict flexibility has some problems. First one, there are less protein structures obtained using NMR than X-ray. Morover dinamic information provided by this method is only available for a proteins smaller than 40 kDa (6). However using NMR for calculate flexibility in disordered proteins or regions is more acurate than other methods (7).

ESTENO 1.0

Esteno1.0 (Fl**e**xibility a**s**sesment wi**t**hout prot**e**in co**n**volutional netw**o**rk) is an approach for prediction of protein flexibility from sequence. The programme works without a deep learning approach, instead it uses a combination of different features that can be retrived directly from the sequence or from data derived from structure predictors (as the PAE matrix from Alpha Fold). Esteno's score has been refined in the way of minimizing the differences between its sister programme Medusa.

As mentioned, Esteno uses different features to build its score. The first feature used is the mean of the PAE values of the aminoacids centered in the diagonal of the PAE matrix generated by Alpha Fold. This first feature was our first score approximation(the method that computes it is diag_score), when we compared it with the 5 classes prediction of Medusa we observed that some regions of high or low flexibility where correlated with picks (maximums or minimus respectively) so we code some functions that detect those picks and assign them a value that varies depending on how many positions in a row the values of PAE are ascending or descending and it also gives a value to the surronding values of this picks that is weighted depending on the relative increase or decrease of the PAE value for that step (the function that returns this list is max_min_rm). Another problem we observed is that some proteins present a PAE matrix with low valued regions but Medusa classifies some of that regions as flexible, this is the case of P11344. By taking a closer look to that concrete case we observe that the problematic regions presented a domain with some loops enriched with aminoacids such as proline or glycine. This lead us to use another feature that takes into account the aminoacid composition, this took us to a paper (8) that gives a mean flexibility score for each of the 20 aminoacids. Using this we divided the aminoacids in two groups (flexible and non-flexible), the last feature scans the protein and gives a score of 1 for that positions where there are 3 or more flexible aminoacids counting from 2 aminoacids downstream to downaminoacids upsetream and a score of 0 to that positions where there are only 2 or less flexible aminoacids in that environment (the function that gives this is flexible_env).

Once we had the different features we comined them by summing them in a total score and standarized this score. Once standarized we established some cutting values to classify each position with a 0, 1 or 2 that indicates from less to more flexible. The standarization method that

better worked for our score was to move the distribution to a distribution situaded between 0 and 2 and using as cut-off values 0.66 and 1.33.

Before deciding ourselves with the method to classify each position we analyzed different scores and parameters by comparing them with the Medusa outputs, we did this with 4 different proteins:

SS1=Sum of diag_score and max_min_rm Standarized, Cut-values: 0, 1*
SS2=Sum of diag_score and max_min_rm Standarized, Cut-values: 0.66, 1.33
SM1=Sum of diag_score and max_min_rm standarized by moving to distribution situated between 0 and 2
XS1=diag_score Standarized Cut-values: 0,1
XS2=diag_score Standarized Cut-values: 0.66, 1.33
XM1=diag_score moved to 0-2
TS1=Sum of diag_score, max_min_rm and flexible_env Standarized, Cut-values: 0, 1
TS2=Sum of diag_score, max_min_rm and flexible_env Standarized, Cut-values: 0.66, 1.33
TM1=Sum of diag_score, max_min_rm and flexible_env moved to 0-2
RMSE=RMSE between the predicted class using the concrete score and the class given by medusa (3 classes prediction).

*Note: The cut-values mentioned are the ones used to classify in 3 classses. To compare with the two-classes results of Medusa (S and NS) the cut-off values were: 0 in the case of the normal standarization and 1 in the case of moving the distribution to 0-2.

P11344

	3	S	NS	RMSE
SS1	0.551595	0.690432	0.630394	0.968549
SS2	0.549719	0.690432	0.630394	0.960769
SM1	0.547842	0.731707	0.664165	0.970484
XS1	0.562852	0.701689	0.645403	0.982969
XS2	0.566604	0.701689	0.645403	0.992467
XM1	0.564728	0.707317	0.647280	0.982015
TS1	0.534709	0.681051	0.677298	0.851001
TS2	0.547842	0.681051	0.677298	0.982015
TM1	0.542214	0.703565	0.639775	0.849898

P06401

	3	S	NS	RMSE
SS1	0.444802	0.663451	0.679528	0.948288
SS2	0.430868	0.663451	0.679528	1.280575
SM1	0.454448	0.540193	0.511254	0.883334
XS1	0.473741	0.727760	0.748124	0.903133
XS2	0.409432	0.727760	0.748124	1.203781
XM1	0.551983	0.733119	0.757771	0.957847
TS1	0.455520	0.670954	0.661308	1.013309
TS2	0.467310	0.670954	0.661308	1.132204
TM1	0.491961	0.696677	0.687031	0.872959

Q9P7Q4

	3	S	NS	RMSE
SS1	0.547980	0.684343	0.664141	1.054093

SS2	0.598485	0.684343	0.664141	1.149769
SM1	0.547980	0.655303	0.625000	1.054093
XS1	0.584596	0.675505	0.655303	1.038404
XS2	0.627525	0.675505	0.655303	1.080708
XM1	0.588384	0.679293	0.646465	1.042045
TS1	0.529040	0.705808	0.685606	0.971825
TS2	0.587121	0.705808	0.685606	1.104399
TM1	0.539141	0.672980	0.655303	0.976362

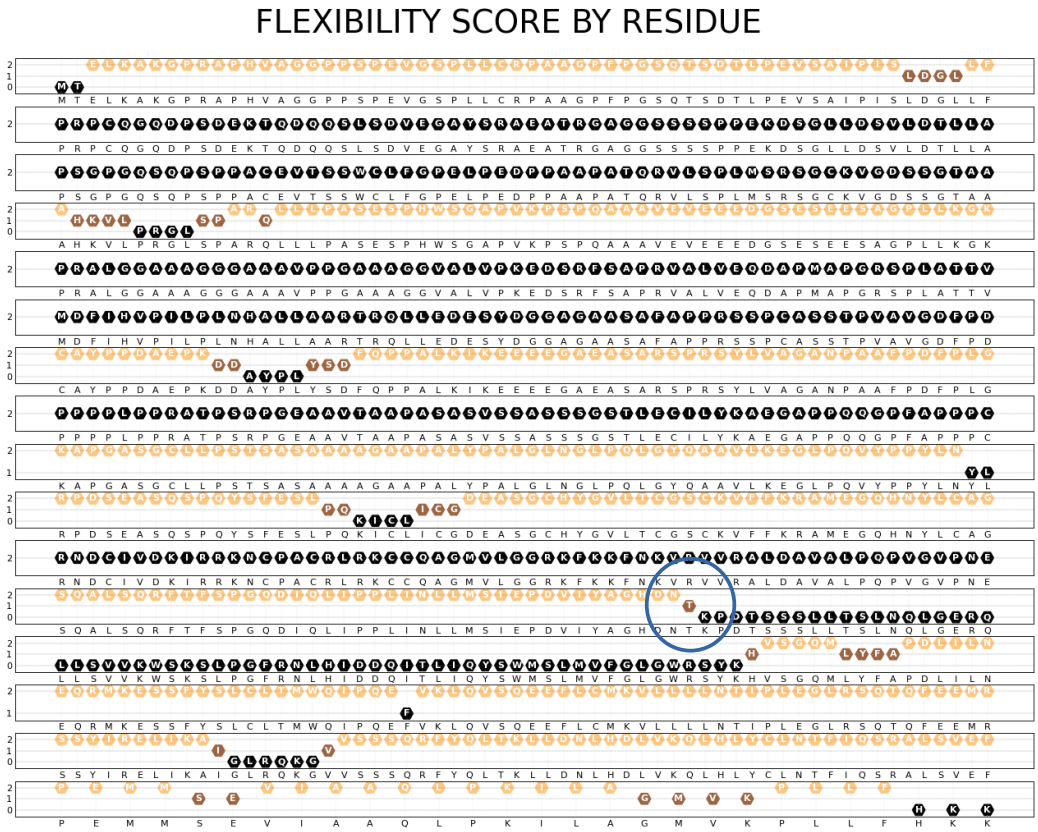
Q13888

	3	S	NS	RMSE
SS1	0.475949	0.724051	0.678481	0.938623
SS2	0.445570	0.724051	0.678481	1.056625
SM1	0.498734	0.731646	0.681013	0.905678
XS1	0.584810	0.779747	0.744304	0.811312
XS2	0.567089	0.779747	0.744304	0.880161
XM1	0.582278	0.797468	0.746835	0.826767
TS1	0.508861	0.693671	0.688608	0.900070
TS2	0.468354	0.693671	0.688608	0.997465
TM1	0.518987	0.716456	0.691139	0.881598

We can conclude by observing this data that the general success in classifiying well the aminoacid as flexible or rigid rounds the 70% (when classifying into 2 classes). By comparing the ratio of success when classifying in 3 classes and the RMSE (a mesure that is even more indicative of the accurateness of the method because it distinguish between a bad classification and an awful one) we selected as a score to implement into our programme TM1. This scoring approach has the lower RMSE on average and, as we commented, works notably better with special cases as the one of the protein P11344 because it takes into account a higher number of features so it is more adaptable when one of them does not perform as expected.

Now we are going to show out graphical representation of the six proteins proposed.

P06401

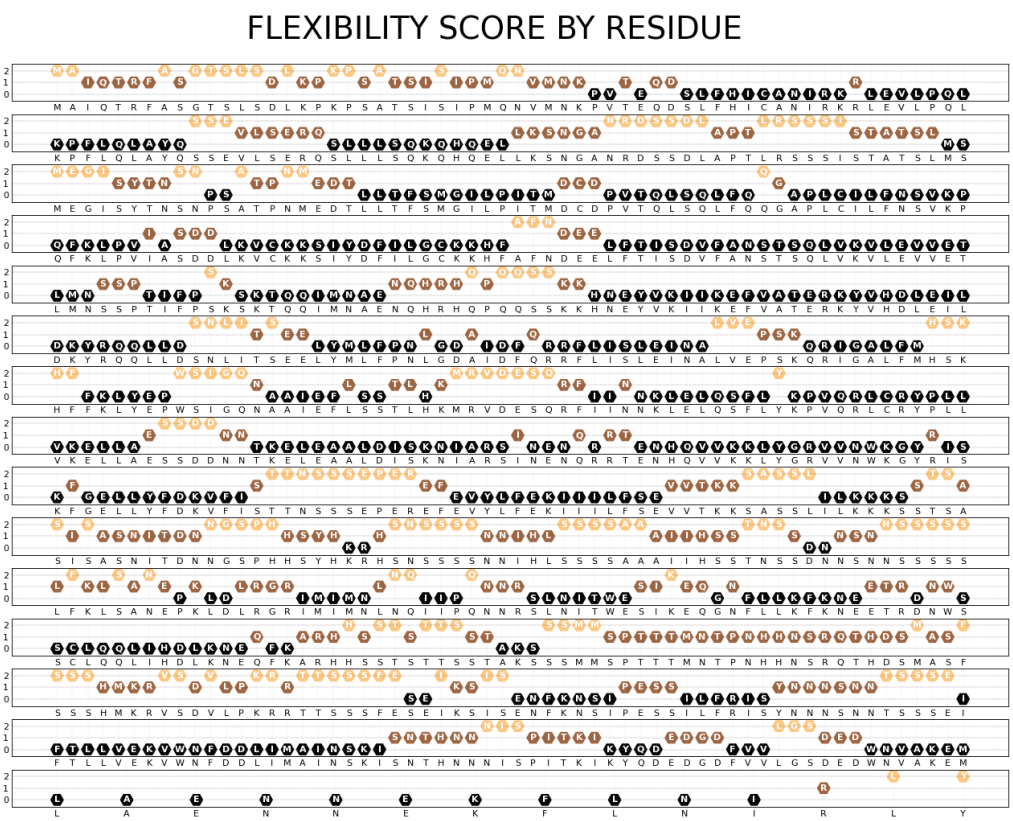


As we can see in the graphic, until aminoacid number 700 (blue circle), almost every aminoacid scores a flexibility 2. There are some regions which have a score of 0 (rigid) as well but they are to short and the may have an important function in the protein. After the aminoacid 705 we have a large region with all the aminoacids rigids. This region finishes around aminoacid 765.

If we take into acount the crystallographic X-ray structures on PDB, these structures are obtained after the aminoacid 680 more or less. That means all previous regions are dificult to crystallize. This can be because these regions are so flexible and are difficult to obtain a regular crystall which can be used in X-ray crytallography.

We are using PAE matrix from alphafold, however, aminoacids before 700 tend to have a pLDDT very low. That means alpha fold is not sure at all of the structure proposed for that aminoacid. These regions need to be traited as a intrinsecally disordered regions as some authors propose (9) (10)

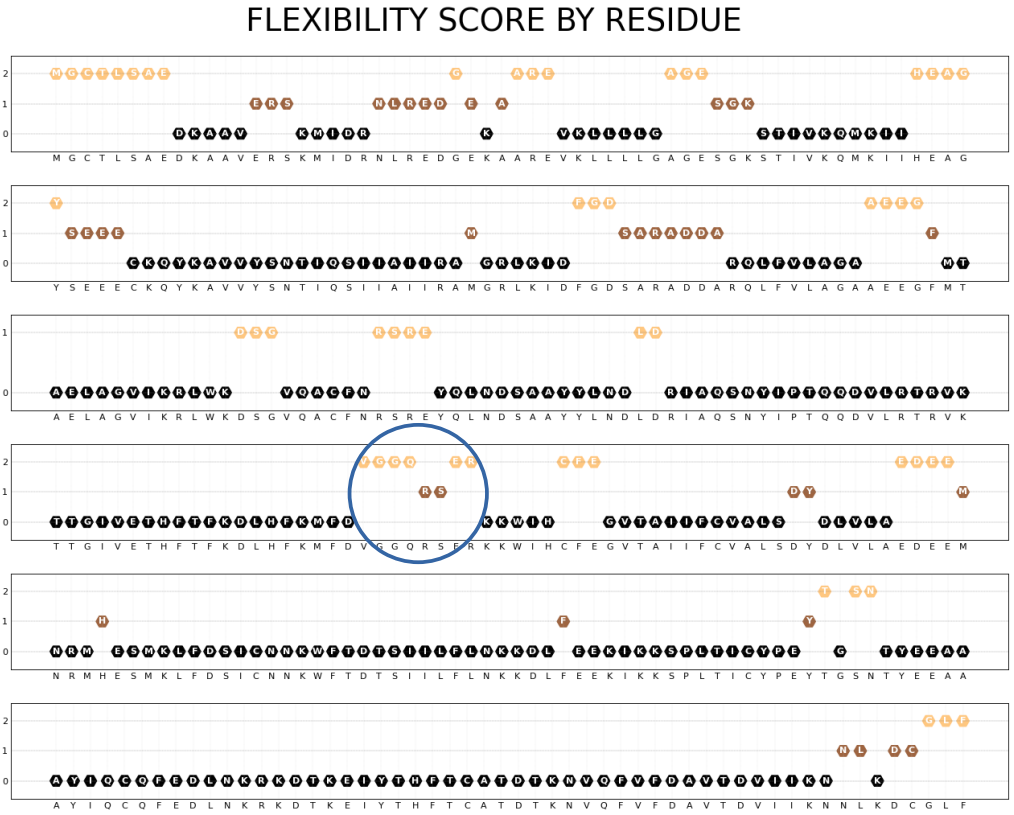
P11433



Here we have the representation of CDC24 of *S. cerevisiae* an important protein in the regulation of the cell cycle. PDB structures are only available for the C-terminal part of the protein using NMR. In this case, structure from AlphaFold is much better than the previous one, so here we have an example supporting our approach as a predictor of flexibility only using PAE matrix and some features always availables.

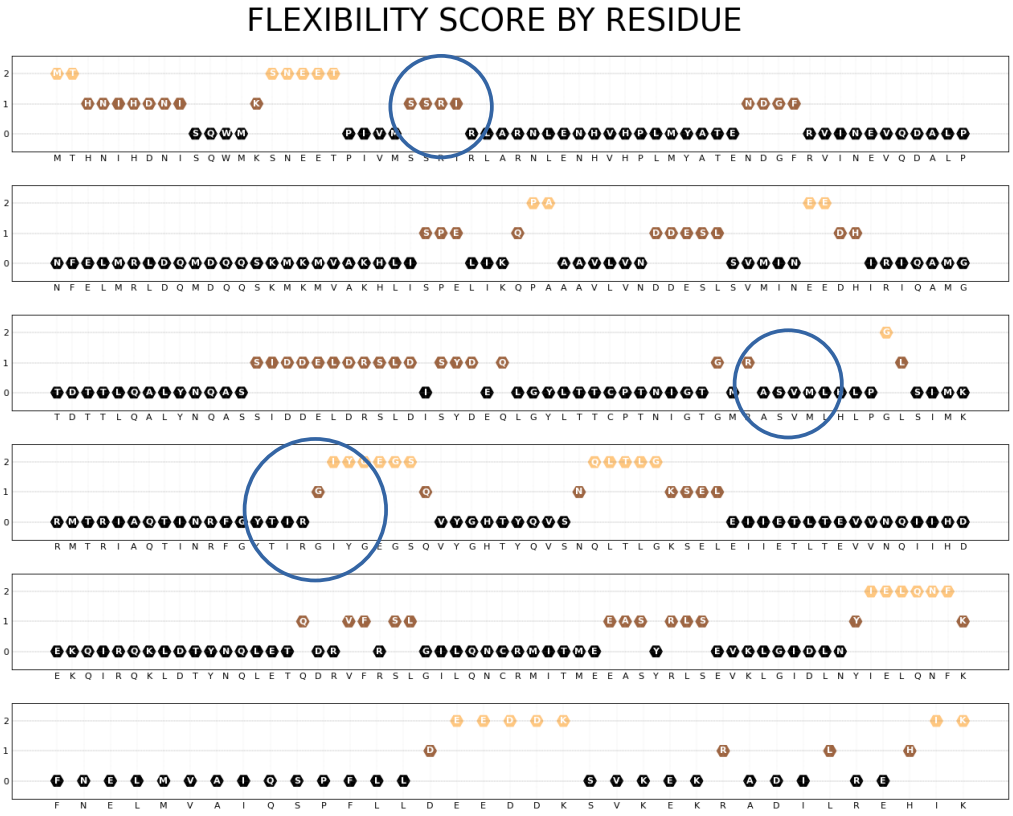
The first aminoacids have a score of 1 or 2 (flexibles), and comparing with the pLDDT this region migth be disordered. There are some regions which are disordered as well and in this graphic match with a score of 1 or 2.

P38401



This protein does not have any structure on PDB. So here we can not compare with any empiric data. Alpha fold structure defines with a high precision this protein and our analysis shows that rigid regions are predominants. Regions with flexible aminoacids match with loops in the structure. The blue circle corresponds to the loop between an alpha helix and a beta strand in the positions of the aminoacids 201-206. N-terminal and C-terminal aminoacids are flexibles too in out analysis because these regions tend to be less structured.

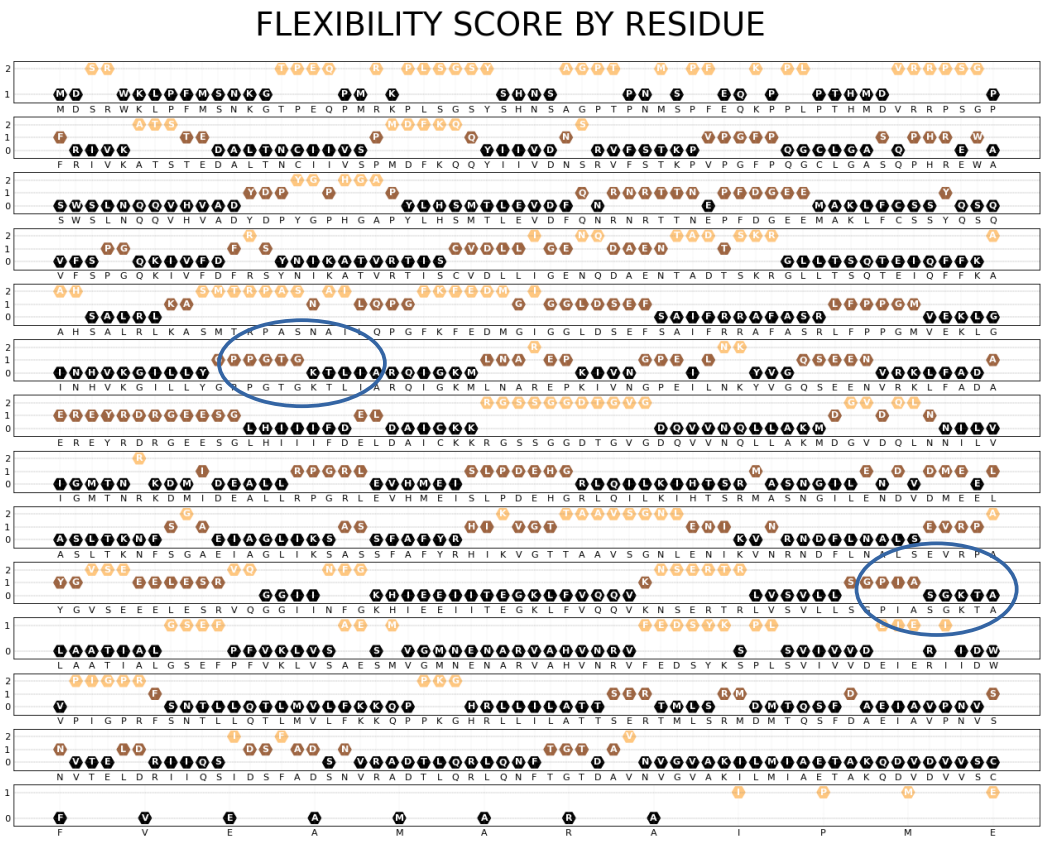
P65206



This protein is an arginine-kinase which uses ATP and no PDB structure is available only AlphaFold one. It has a ATP binding region composed by 3 regions. In this case we are not going to focus in the general aspects of this protein only see which fleixibility have the aminoacids implicated in the binding of ATP. Blue circles represent the binding regions implied in the binding of the ATP and as we can see, first one has an intermediate flexibility (score 1), second one is rigid but has an aminoacid with a score of 1, and the third region has an increasing flexibility.

Studying binding regions of molecules has a lot of importance in order to unfold the caracteristics and properties of these special regions. Here we observe that regions 1 and 3 are intermediate flexible and region 2 is preferently rigid. Some flexibility in these regions is requiered in order to allow the conformational change between the form with the ATP binded and with no ATP.

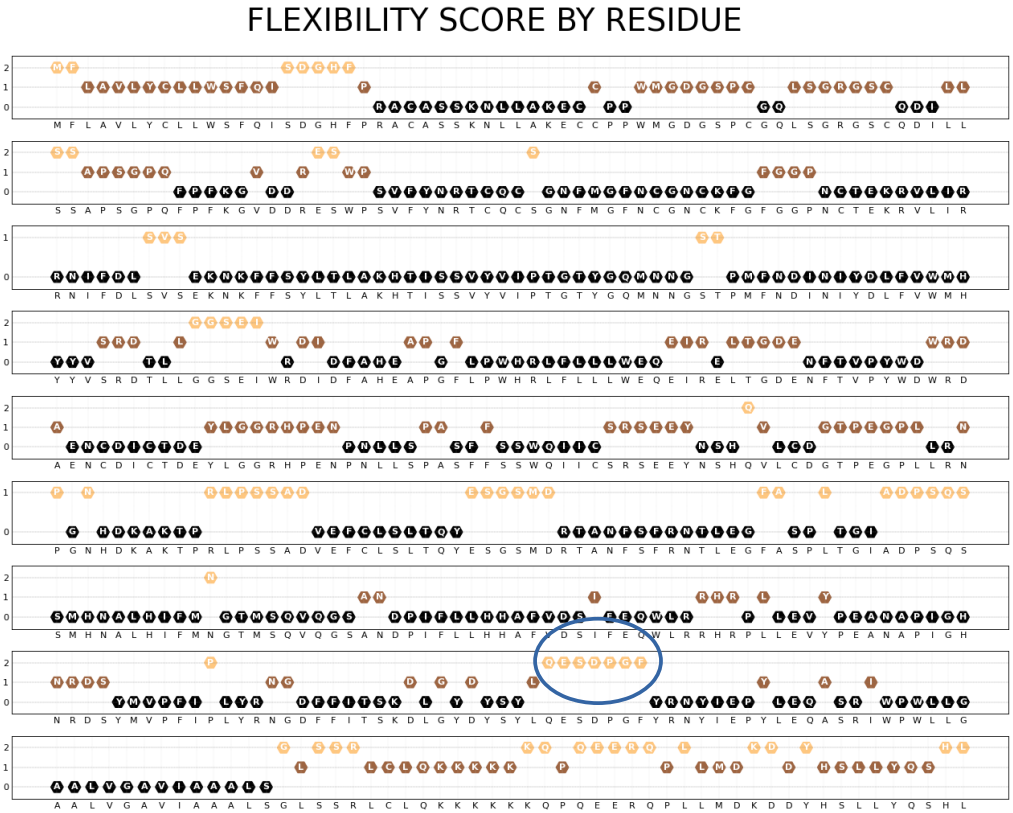
Q9P7Q4



This protein has also a binding site for the ATP, and as before, blue circles represent the aminoacids which are implied in the binding. We observe that these aminoacids (first aminoacids of both regions) have a score of 1, which has sense with we have said before.

Now we are going to analize two proteins which are not the proposed ones

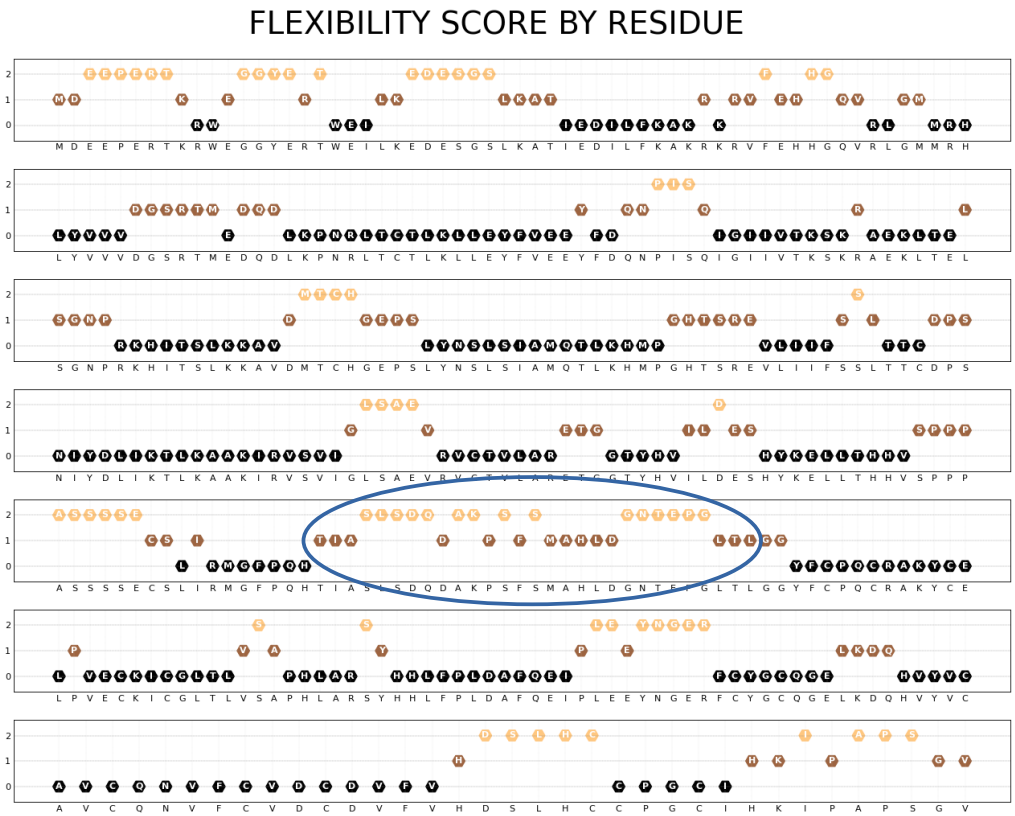
P11344



This protein has two well defined domains which are separated by a linker (blue circle). This linker is flexible (score 2) and allows the two domains to have independence. Usually protein domains tend to be connected by a flexible linker and here we have one example.

Other regions as N-terminal and C-terminal show high flexibility. In this case this regions do not have a good pLDDT and might be disordered as we have said before.

Q13888



Here the region with the blue circle has a lot of loops which conect Beta-strands. These beta-strands are short. This region is good to see the effect of the enviroment of residues. They are connected by a lot of loops and beta-strands are short that is why in this region are a lot of aminoacids with scores 1 or 2.

BIBLIOGRAPHY

(1) A. Schlessinger, B. Rost, Protein flexibility and rigidity predicted from sequence, Proteins 61 (2005) 115–126.

(2) G. Lipari, A. Szabo, Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules, J. Am. Chem. Soc. 104 (1982) 4546–4570.

(3) Narwani, T. J., Etchebest, C., Craveur, P., Léonard, S., Rebehmed, J., Srinivasan, N., ... & de Brevern, A. G. (2019). In silico prediction of protein flexibility with local structure approach. *Biochimie*, *165*, 150-155.

(4) Dong, Q., Wang, K., Liu, B., & Liu, X. (2016). Characterization and prediction of protein flexibility based on structural alphabets. *BioMed Research International*, *2016*.

(5) Na, H.; Song, G. The performance of fine-grained and coarse-grained elastic network models and its dependence on various factors. Proteins: Struct., Funct., Genet. 2015, 83, 1273−1283.

(6) O.F. Lange, P. Rossi, N.G. Sgourakis, Y. Song, H.W. Lee, J.M. Aramini, A. Ertekin, R. Xiao, T.B. Acton, G.T. Montelione, D. Baker, Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 10873e10878.

(7) Cook, E. C., Usher, G. A., & Showalter, S. A. (2018). The use of 13C direct-detect NMR to characterize flexible and disordered proteins. In *Methods in Enzymology* (Vol. 611, pp. 81-100). Academic Press.

(8) Bhaskaran, R. P. P. K., & Ponnuswamy, P. K. (1988). Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research*, *32*(4), 241-255.

(9) Tunyasuvunakool, K., Adler, J., Wu, Z. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596,** 590–596 (2021). https://doi.org/10.1038/s41586-021-03828-1

(10) Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, *433*(20), 167208.