Protein flexibility of steroid receptors and protein kinases analyzed by ProtFLEXpreD.

Laura Ciaran Alfano and Neus Pou Amengual

MSc in Bioinformatics for Health Sciences, Universitat Pompeu Fabra

Structural Bioinformatics and Introduction to Python project, 2022

INTRODUCTION

Proteins rely on flexibility to respond to environmental changes, ligand binding and chemical modifications. [1] Structural flexibility enables their conformational change, which is associated with numerous biological activities such as molecular recognition, allosteric regulation, catalytic activity, and protein stability. [2] Thus, information on protein flexibility is as important as tertiary structure to provide more insights into understanding protein function, and consequently will have significant impact on genomic study, disease research, and drug-design. [2]

However, discerning protein flexibility by either experiments or computational simulations is a difficult process. In particular, many computational approaches to understanding protein flexibility require an experimentally determined protein structure. [3] And, most of the flexibility prediction methods use the crystallographic data as the only indicator of protein's inner flexibility and predict them as rigid or flexible. [4]

The Debye-Weller factors, commonly referred to as B-factors or temperature factors, describe the attenuation of X-ray scattering or coherent neutron scattering caused by thermal motion. [5, 6, 7, 8] They indicate both the static mobility, related to the molecule orientation, and dynamic mobility, caused by the atom's vibration. [2] B-factors are the most common experimental descriptors of protein flexibility, reflecting both atom vibration and static structural disorder. [6, 8]

Although each atom in protein atomic resolution structures has its B-factor, the B-factor of the whole residue is generally represented by its Cα B-factor. [2, 7]

Protein residues are classified into two states as rigid or flexible on the basis of a B-value threshold. [6] The residues with low B-factor values are usually more stable in structure than the ones with large B-factor values. Indicating that residues with high B-factors have higher than average flexibility as opposed to residues with low B-factors are believed to occur at more rigid positions. [5, 6]

Comparing the B-factor values from highly similar pairs of crystallized chains has demonstrated that flexibility is encoded at the amino acid sequence level to a significant degree and therefore should be predictable, at some level, from the amino acid sequence. [5]

This is why B-factors were used as predictors of protein flexibility in this project. But, because they are the result of a number of factors, including degree of resolution, crystal contacts, and the particular refinement procedure. This prevents the direct comparison between non-normalized B-factors from different structures. Thus, they have to be normalized. The following normalization was applied:

$$Bfactor = \frac{observed\ Bfactor - \mu}{\sigma}$$

where μ is the mean of B-factors of a given structure, and σ is the standard deviation. [2, 5]

NUCLEAR RECEPTOR FAMILY

The nuclear receptor (NR) superfamily is composed of a family of transcription factors (TFs) that bind and respond to certain steroids and other signaling molecules, such as vitamin D3, thyroid hormone, and retinoids, playing an important role in a number of biological processes including metabolism, reproduction, and inflammation. [9, 10, 11] They modulate the expression of target genes by recruiting co-regulatory complexes to specific sites in the genome. [11, 12]

There are homologues in Drosophila for several of the receptors, indicating that they derived from a common ancestor gene that precedes the divergence of invertebrates and vertebrates. [13]

Conformational changes within the receptor occur when the ligand binds, which in turn binds specific DNA sequences throughout the genome. This causes the recruitment of co-regulator proteins, chromatin remodelers, and the general transcriptional machinery to the DNA for the activation or repression of the target gene expression. [9]

Regardless of the diversity in the size, shape, and charges of activating ligands, almost all members of the nuclear receptor superfamily share a common modular five domain structure. Each of these subdomains plays a specific role in receptor biology. [9]

The N-terminal domain is highly disordered, making it difficult to perform a structural analysis, in addition to having little conservation and a lot of difference in its size between the different NRs. It contains the Activator Function-1 region (AF-1), which interacts with a variety of coregulator proteins in a specific manner. [9, 10, 13]

The DNA binding domain is the most conserved domain and it confers specificity for binding to hormone response elements. It is approximately 70 amino acids long and has two subdomains that each contain four highly conserved cysteine residues that coordinate a zinc ion to create the canonical DNA-binding zinc finger motif. Each zinc finger is then followed by an amphipathic helix and a peptide loop. [13]

The zinc fingers of the nuclear receptors fold towards each other in order to form a more globular domain, containing two helices, which are located at the carboxy-terminal end of the zinc fingers. These helices are oriented perpendicular to each other and form the base of a hydrophobic core. [13]

The first subdomain contains the DNA reading helix, which interacts with the major groove to make base-specific interactions with the DNA. The second subdomain helix makes non-specific contacts with the DNA backbone, defining the optimal spacing and alignment of half-sites. The peptide loop, of five amino acids, in this subdomain, contains the distal box, or “D box,” that contains residues for receptor dimerization. [9, 11, 12, 13]

The ligand-binding domain (LBD) is a signalling domain that not only binds to ligands but also interacts directly with coregulator proteins located along the C-terminal 200– 300 residues. While the amino acid sequence conservation is low, it is a structurally conserved domain that commonly contains 11-12 α-helices and four β-strands that fold into three parallel layers to form an alpha-helical sandwich. This folding creates a hydrophobic ligand-binding pocket (LBP) at the base of the receptor. This pocket is composed of ~75% hydrophobic residues, but also contains critical polar residues that make key hydrogen bonding interactions with the ligand. These hydrogen bonds help position the ligand in the correct orientation. The variability across the nuclear receptors at the ligand-binding region allows them to recognize a diverse cadre of ligands. [9, 10, 11, 12]

This domain contains another activation function surface (AF-2), which is composed of helices 3, 4, and 12. Helix 12, or the activation function helix (AF-H) has been shown to be conformationally dynamic upon ligand binding, altering the orientation of AF-2 to facilitate interaction with different co-regulator proteins. Other α helices in the LBDs also shift in positions in subtle but still meaningful ways that can impact receptor activation. [9, 10, 11, 12]

Co-activator proteins interact with the AF-2 surface of the nuclear receptor via an alpha helix containing a specific motif. The ends of the helical peptide are generally held in place by a charge clamp formed by a lysine on the nuclear receptor H3 and a glutamate on H12 that cap the helix dipole. Co-repressors contain a longer conserved motif that interacts with the same hydrophobic surface but, because of their length, the canonical clamp formation is inhibited. The discrimination between either co-activator or co-repressor binding has been linked to the conformational flexibility of H12. The binding of a ligand stabilizes this helix into a more fixed conformation. [9]

The hinge region is a short, flexible linker between these two fundamental domains. This region has the least sequence and size conservation between nuclear receptors. It is also a site for regulatory post-translational modifications. The hinge can also contain a nuclear localization signal. [9, 10]

STEROID RECEPTORS SUBFAMILY

The nuclear receptor superfamily can be divided into seven subfamilies, and the two receptors that are explained in this report belong to the third subgroup which comprises the steroid receptors, which are key regulators of a host of metabolic, reproductive, and developmental processes. [9, 13]

When steroid receptors are unligated, they are bound in a complex with heat-shock proteins. Most of these complexes are located in the cytoplasm, with exception of the estrogen receptor (ER) complexes, which are found in the nucleus. The binding of the ligand dissolves the complexes and the ligated receptors form homodimers that bind to hormone response elements in the promoter region of target genes. [9, 13]

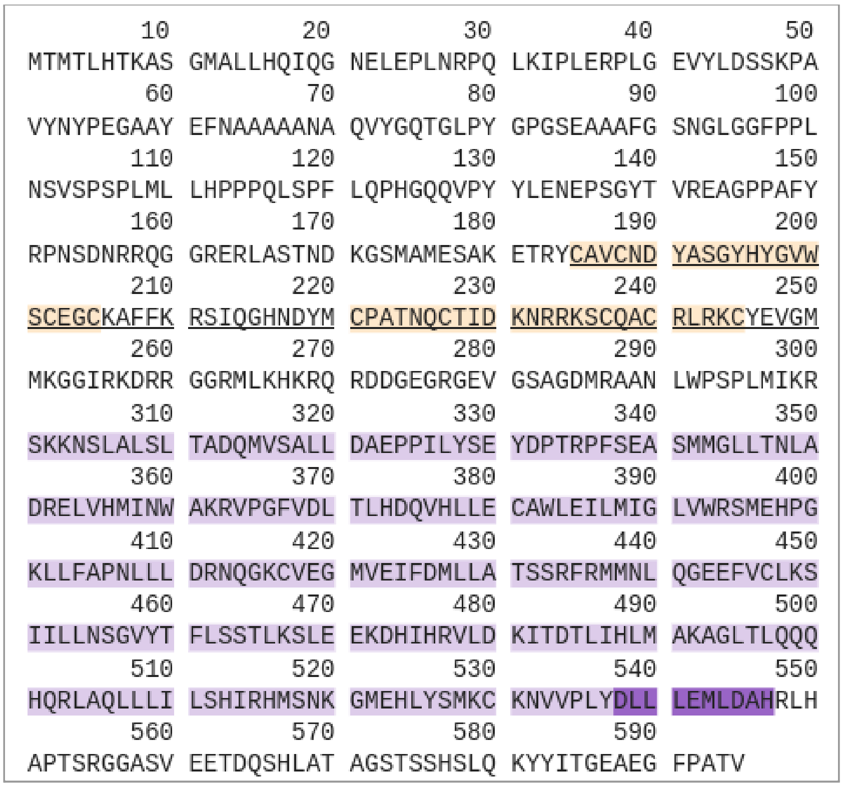
Another characteristic is that the carboxy-terminal extension seems to extend the protein–DNA interface by contacting the minor groove of the response element and also plays an important role in sequence-specificity. [11]

The steroid receptors bind palindromic repeats of the DNA response elements. These palindromes contain two AGGACA repeats that can be separated by a spacer region that varies in length. The length of this spacer has been shown to allosterically modulate SRs, resulting in varied transcriptional outputs. [9]

In both proteins we mostly have structural information about the ligand-binding domain, information about the DNA binding domain is limited. So the report will be focused on the ligand-binding domain.

ESTROGEN RECEPTOR (P03372)

In the protein sequence, with the information taken from Uniprot, we can see the different domains that are part of the nuclear receptor family (**Fig. 1**). The DNA binding domain is found between residues 185 and 250, with the two zinc fingers located in residues 185-205 and 221-245. Then we have the hinge region, located between residues 251 and 310. And directly after the ligand-binding domain, from residue 311 to 547.



```
        10         20         30         40         50
MTMTLHTKAS GMALLHQIQG NELEPLNRPQ LKIPLERPLG EVYLDSSKPA
        60         70         80         90        100
VYNYPEGAAY EFNAAAAANA QVYGQTGLPY GPGSEAAAFG SNGLGGFPPL
       110        120        130        140        150
NSVSPSPLML LHPPPQLSPF LQPHGQQVPY YLENEPSGYT VREAGPPAFY
       160        170        180        190        200
RPNSDNRRQG GRERLASTND KGSMAMESAK ETRYCAVCND YASGYHYGVW
       210        220        230        240        250
SCEGCKAFFK RSIQGHNDYM CPATNQCTID KNRRKSCQAC RLRKCYEVGM
       260        270        280        290        300
MKGGIRKDRR GGRMLKHKRQ RDDGEGRGEV GSAGDMRAAN LWPSPLMIKR
       310        320        330        340        350
SKKNSLALSL TADQMVSALL DAEPPILYSE YDPTRPFSEA SMMGLLTNLA
       360        370        380        390        400
DRELVHMINW AKRVPGFVDL TLHDQVHLLE CAWLEILMIG LVWRSMEHPG
       410        420        430        440        450
KLLFAPNLLL DRNQGKCVEG MVEIFDMLLA TSSRFRMMNL QGEEFVCLKS
       460        470        480        490        500
IILLNSGVYT FLSSTLKSLE EKDHIHRVLD KITDTLIHLM AKAGLTLQQQ
       510        520        530        540        550
HQRLAQLLLI LSHIRHMSNK GMEHLYSMKC KNVVPLYDLL LEMLDAHRLH
       560        570        580        590
APTSRGGASV EETDQSHLAT AGSTSSHSLQ KYYITGEAEG FPATV
```

Figure 1. *Estrogen receptor protein (P03372) sequence. Residues that belong to the DNA binding domain are underlined, and the residues of the zinc fingers are highlighted in yellow. The residues that comprise the ligand-binding domain are highlighted in light purple, and the helix 12 is highlighted in dark purple.*

If we look at the flexibility plots (**Fig. 2**) we can see different peaks of protein flexibility. In this figure we can not only observe the flexibility profile of the protein but also that the residues considered as flexible have higher values than the ones considered as rigid, agreeing with the literature. [8] In this part of the report, only the flexibility peaks longer than 4 residues are explained, they are indicated in **Fig. 3**.

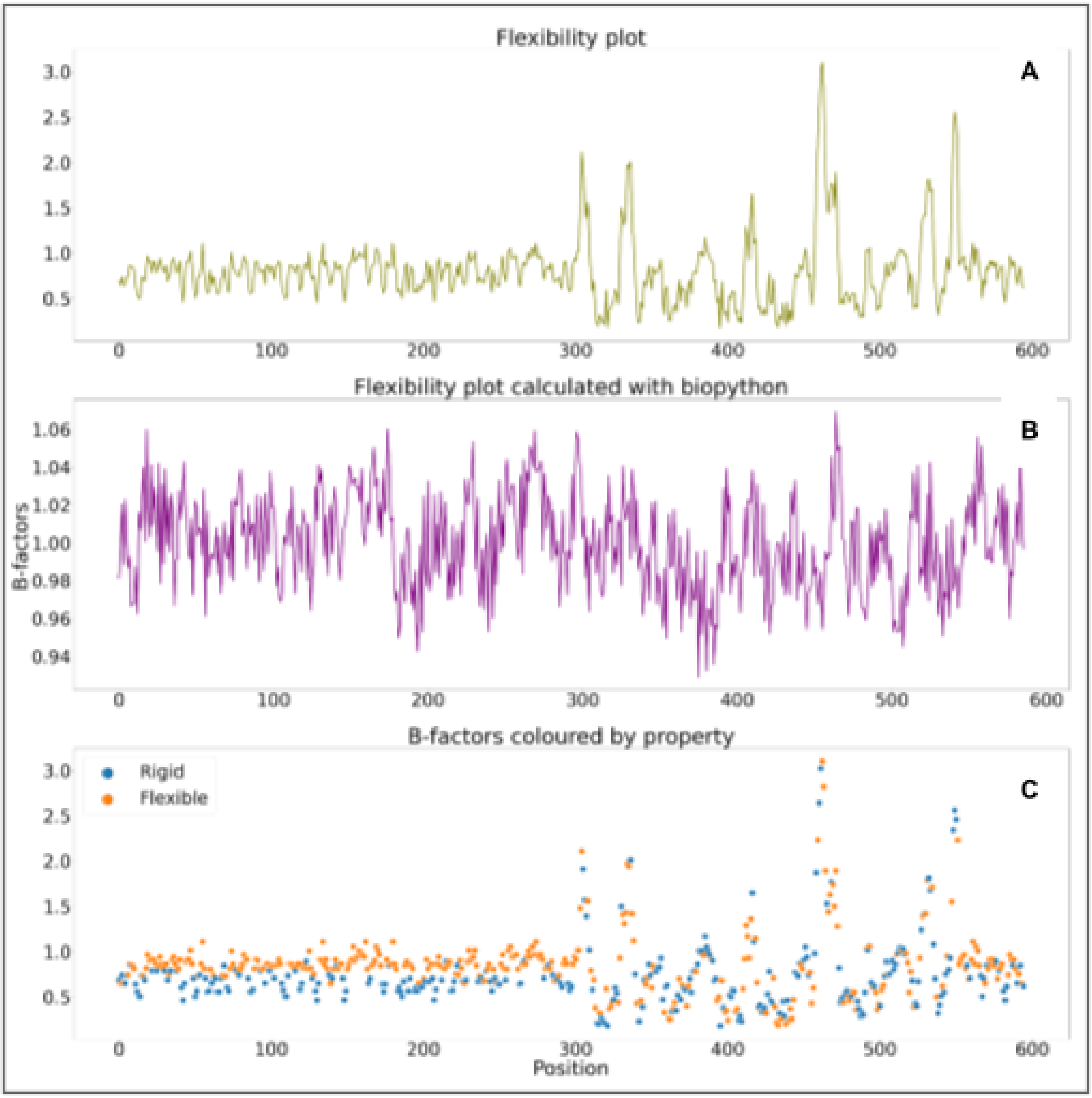
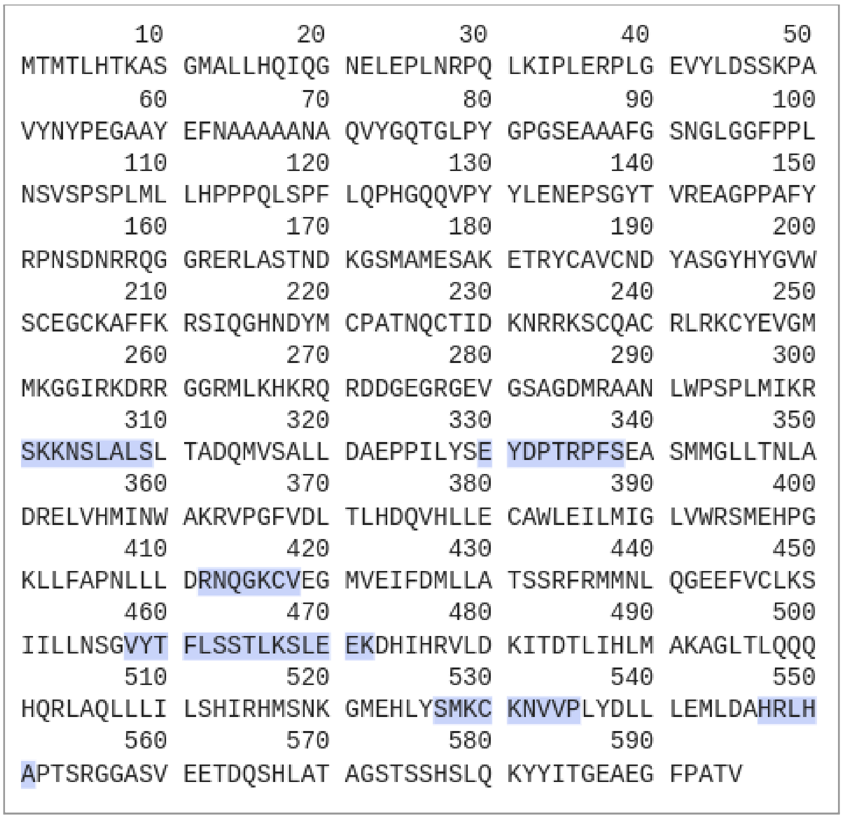


Figure 2. *ProtFLEXPred results of estrogen receptor protein from Homo Sapiens.* ***A****. Line plot that represents the flexibility profile predicted with the ProtFLEXpreD program (PFD profile). Residues 260 - 270 within the brown square.* ***B.*** *Line plot of the flexibility profile predicted by the biopython package (BP profile). [29].* ***C.*** *Scatter plot that represents the B-factors values coloured by flexibility or rigidity obtained by the ProtFLEXpreD program (PFD coloured). [8]*



```
        10         20         30         40         50
MTMTLHTKAS GMALLHQIQG NELEPLNRPQ LKIPLERPLG EVYLDSSKPA
        60         70         80         90        100
VYNYPEGAAY EFNAAAAANA QVYGQTGLPY GPGSEAAAFG SNGLGGFPPL
       110        120        130        140        150
NSVSPSPLML LHPPPQLSPF LQPHGQQVPY YLENEPSGYT VREAGPPAFY
       160        170        180        190        200
RPNSDNRRQG GRERLASTND KGSMAMESAK ETRYCAVCND YASGYHYGVW
       210        220        230        240        250
SCEGCKAFFK RSIQGHNDYM CPATNQCTID KNRRKSCQAC RLRKCYEVGM
       260        270        280        290        300
MKGGIRKDRR GGRMLKHKRQ RDDGEGRGEV GSAGDMRAAN LWPSPLMIKR
       310        320        330        340        350
SKKNSLALSL TADQMVSALL DAEPPILYSE YDPTRPFSEA SMMGLLTNLA
       360        370        380        390        400
DRELVHMINW AKRVPGFVDL TLHDQVHLLE CAWLEILMIG LVWRSMEHPG
       410        420        430        440        450
KLLFAPNLLL DRNQGKCVEG MVEIFDMLLA TSSRFRMMNL QGEEFVCLKS
       460        470        480        490        500
IILLNSGVYT FLSSTLKSLE EKDHIHRVLD KITDTLIHLM AKAGLTLQQQ
       510        520        530        540        550
HQRLAQLLLI LSHIRHMSNK GMEHLYSMKC KNVVPLYDLL LEMLDAHRLH
       560        570        580        590
APTSRGGASV EETDQSHLAT AGSTSSHSLQ KYYITGEAEG FPATV
```

Figure 3. *Estrogen receptor protein (P03372) sequence. Residues that belong to high B-factor peaks and that are explained in this report are highlighted in blue.*

The flexibility peaks that are easily identified are the ones found between residues 527-535 and 547-551. Those correspond to the flanking amino acids of the helix 12 and are loop regions.

The ligand-binding domain, as previously mentioned, is folded into a three-layered antiparallel a-helical sandwich comprising a central core layer of three helices (H5/6, H9 and H10) sandwiched between two additional layers of helices (H1–4 and H7, H8, H11). The remaining secondary structural elements, a small two-stranded antiparallel b-sheet and H12, are located at the ligand-binding portion at the narrower end of the molecule and flank the main three-layered motif. [14]

The repositioning of helix 12 is essential for the AF-2 activity, the binding of the ligand (**Fig. 4**). When the ligand is an agonist, the H12 sits over the ligand-binding cavity and is located closely against H3, H5/6 and H11. It forms a sort of "lid" of the binding cavity. This conformation seems to be a necessity for transcriptional activation because this "sealing" of the cavity generates a competent AF-2 that can interact with the co-activators. But, when the ligand is an antagonist, because of their size, the alignment of the AF-H over the cavity is blocked, and the helix lies in a groove formed by H5 and the carboxy-terminal end of H3. [14, 15]

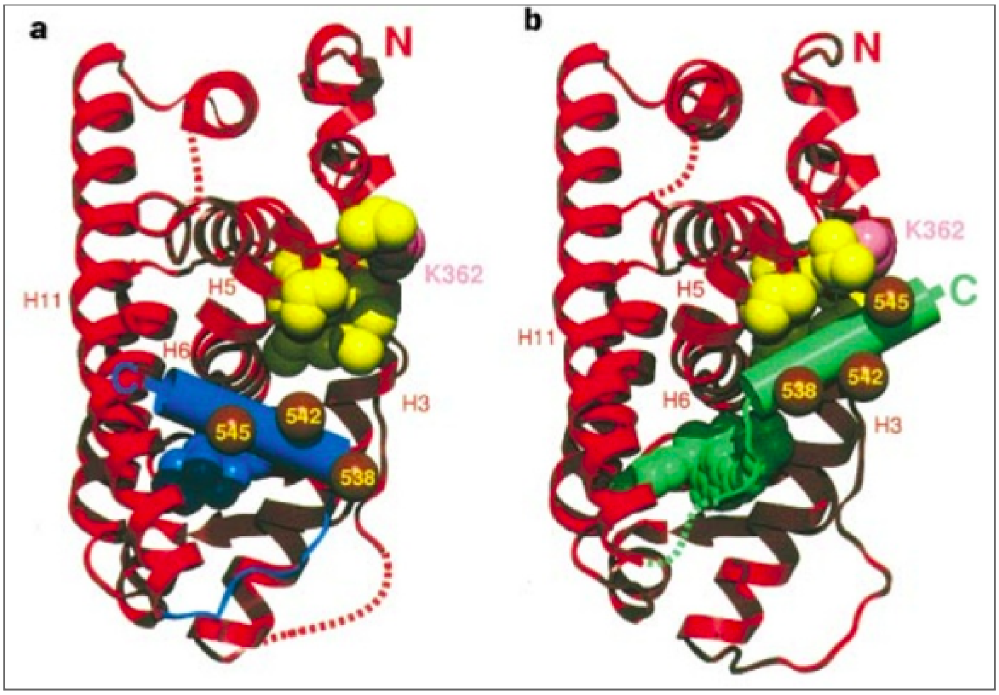


Figure 4. *Positioning of helix H12 in estrogen receptors.* ***A.*** *The LBD–agonist complex.* ***B.*** *The LBD–antagonist complex. H12 is drawn as a cylinder and coloured blue (agonist complex) or green (antagonist complex). The remainder of the LBD is shown in red. [14]*

We have another peak in the hinge region, between residues 301 and 309, this can be explained by the nature and function of this region and, as mentioned in the nuclear receptor family introduction, this is a highly flexible region, the secondary structure is not available. If we look at Alpha Fold, this region is predicted with very low confidence (**Fig. 5**). All of this agrees with the fact that high B-factors are found in this region.

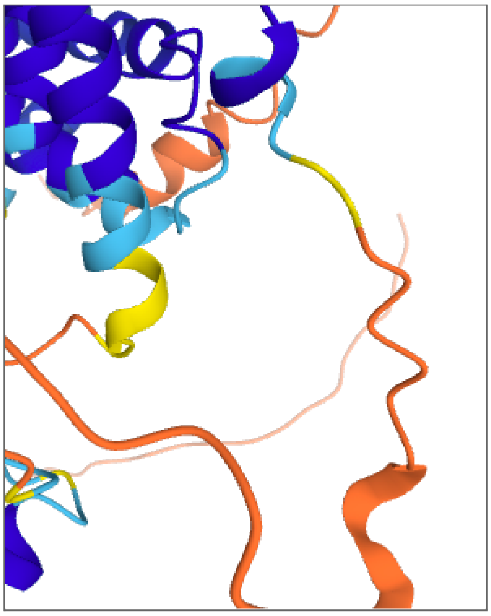


Figure 5. *Alpha Fold structure of estrogen receptor (P03372). Residues* 301 to 309 *correspond to the part of the structure on the right coloured in yellow and orange indicating low (50-70%) and very low (>50%) confidence respectively.*

The peaks between residues 330 to 338, 412 to 418 and 458 to 472, correspond to the n-terminal ends of the helices 3, 5 and 10. These helices are part of the AF-2 and as previously mentioned, are part of helices that shift position when the receptor is bound to the ligand. And, for example, if we look at Alpha Fold, we can see that the last high B-factor region (loop between H10 and H11) has been predicted with low and very low confidence (**Fig. 6**). This can be the cause of the presence of high B-factors in these regions.

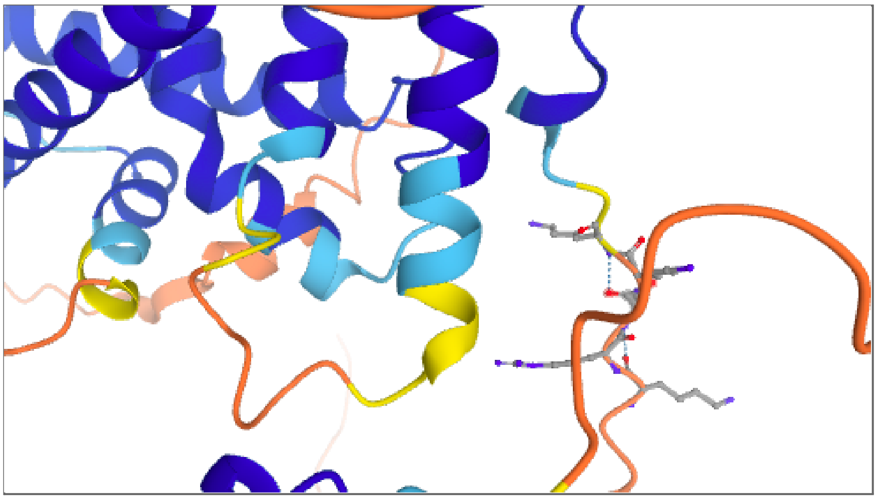


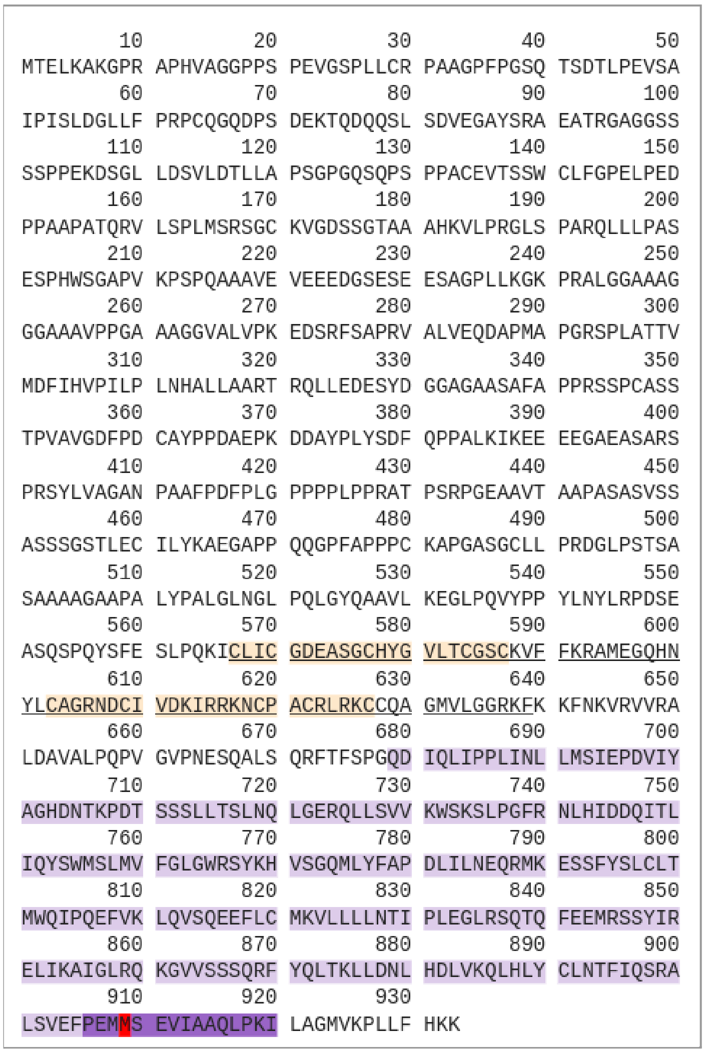
Figure 6. *Alpha Fold structure of estrogen receptor (P03372). Residues 458 to 472 correspond to the part of the structure coloured in yellow and orange indicating low (50-70%) and very low (>50%) confidence respectively.*

It is important to remember that the active site residues occur mainly in regions of low B-factors, while the residues lining the binding pocket tend to exist in higher B-factor regions. [5] Furthermore, it has been seen that residues in coil regions have higher B-values compared to the residues present in other regular secondary structure types (helix and strand). [2] So this could be what is influencing the obtained B-factors.

PROGESTERONE RECEPTOR (P06401)

The progesterone receptor, because it pertains to the same family as the estrogen receptor, has a similar structure and conformation. Using the information from Uniprot, the DNA binding domain is found between residues 567 and 639, with the two zinc fingers located in residues 567-587 and 603-627. Then we have the hinge region, located between residues 640 and 678. And directly after the ligand-binding domain, from residue 679 to 913 (**Fig. 7**).

Methionine 909, located in helix-12, has a crucial function in the agonism/antagonism balance. In the agonist conformation, it is oriented towards the ligand-binding protein mentioned in the estrogen receptor. It is typically the only residue directly in contact with the ligands, and its interactions are key for the function of the receptors. A shift of this residue leads to the destabilization of the entire helix, hindering the agonistic response. [16, 17]



```
        10         20         30         40         50
MTELKAKGPR APHVAGGPPS PEVGSPLLCR PAAGPFPGSQ TSDTLPEVSA
        60         70         80         90        100
IPISLDGLLF PRPCQGQDPS DEKTQDQQSL SDVEGAYSRA EATRGAGGSS
       110        120        130        140        150
SSPPEKDSGL LDSVLDTLLA PSGPGQSQPS PPACEVTSSW CLFGPELPED
       160        170        180        190        200
PPAAPATQRV LSPLMSRSGC KVGDSSGTAA AHKVLPRGLS PARQLLLPAS
       210        220        230        240        250
ESPHWSGAPV KPSPQAAAVE VEEEDGSESE ESAGPLLKGK PRALGGAAAG
       260        270        280        290        300
GGAAAVPPGA AAGGVALVPK EDSRFSAPRV ALVEQDAPMA PGRSPLATTV
       310        320        330        340        350
MDFIHVPILP LNHALLAART RQLLEDESYD GGAGAASAFA PPRSSPCASS
       360        370        380        390        400
TPVAVGDFPD CAYPPDAEPK DDAYPLYSDF QPPALKIKEE EEGAEASARS
       410        420        430        440        450
PRSYLVAGAN PAAFPDFPLG PPPPLPPRAT PSRPGEAAVT AAPASASVSS
       460        470        480        490        500
ASSSGSTLEC ILYKAEGAPP QQGPFAPPPC KAPGASGCLL PRDGLPSTSA
       510        520        530        540        550
SAAAAGAAPA LYPALGLNGL PQLGYQAAVL KEGLPQVYPP YLNYLRPDSE
       560        570        580        590        600
ASQSPQYSFE SLPQKICLIC GDEASGCHYG VLTCGSCKVF FKRAMEGQHN
       610        620        630        640        650
YLCAGRNDCI VDKIRRKNCP ACRLRKCCQA GMVLGGRKFK KFNKVRVVRA
       660        670        680        690        700
LDAVALPQPV GVPNESQALS QRFTFSPGQD IQLIPPLINL LMSIEPDVIY
       710        720        730        740        750
AGHDNTKPDT SSSLLTSLNQ LGERQLLSVV KWSKSLPGFR NLHIDDQITL
       760        770        780        790        800
IQYSWMSLMV FGLGWRSYKH VSGQMLYFAP DLILNEQRMK ESSFYSLCLT
       810        820        830        840        850
MWQIPQEFVK LQVSQEEFLC MKVLLLLNTI PLEGLRSQTQ FEEMRSSYIR
       860        870        880        890        900
ELIKAIGLRQ KGVVSSSQRF YQLTKLLDNL HDLVKQLHLY CLNTFIQSRA
       910        920        930
LSVEFPEMMS EVIAAQLPKI LAGMVKPLLF HKK
```

Figure 7. *Progesterone receptor protein (P06401) sequence. Residues that belong to the DNA binding domain are underlined, and the residues of the zinc fingers are highlighted in yellow. The residues that comprise the ligand binding domain are highlighted in light purple, and the helix 12 is highlighted in dark purple. Methionine 909, important for ligand binding, is highlighted in red.*

If we look at the flexibility plots (**Fig. 8**) we can see different peaks of protein flexibility. In this figure, once again it seems that the residues are well classified depending on if they are rigid or flexible. In this part of the report, only the flexibility peaks longer than 4 residues are explained, they are indicated in **Fig. 9**.

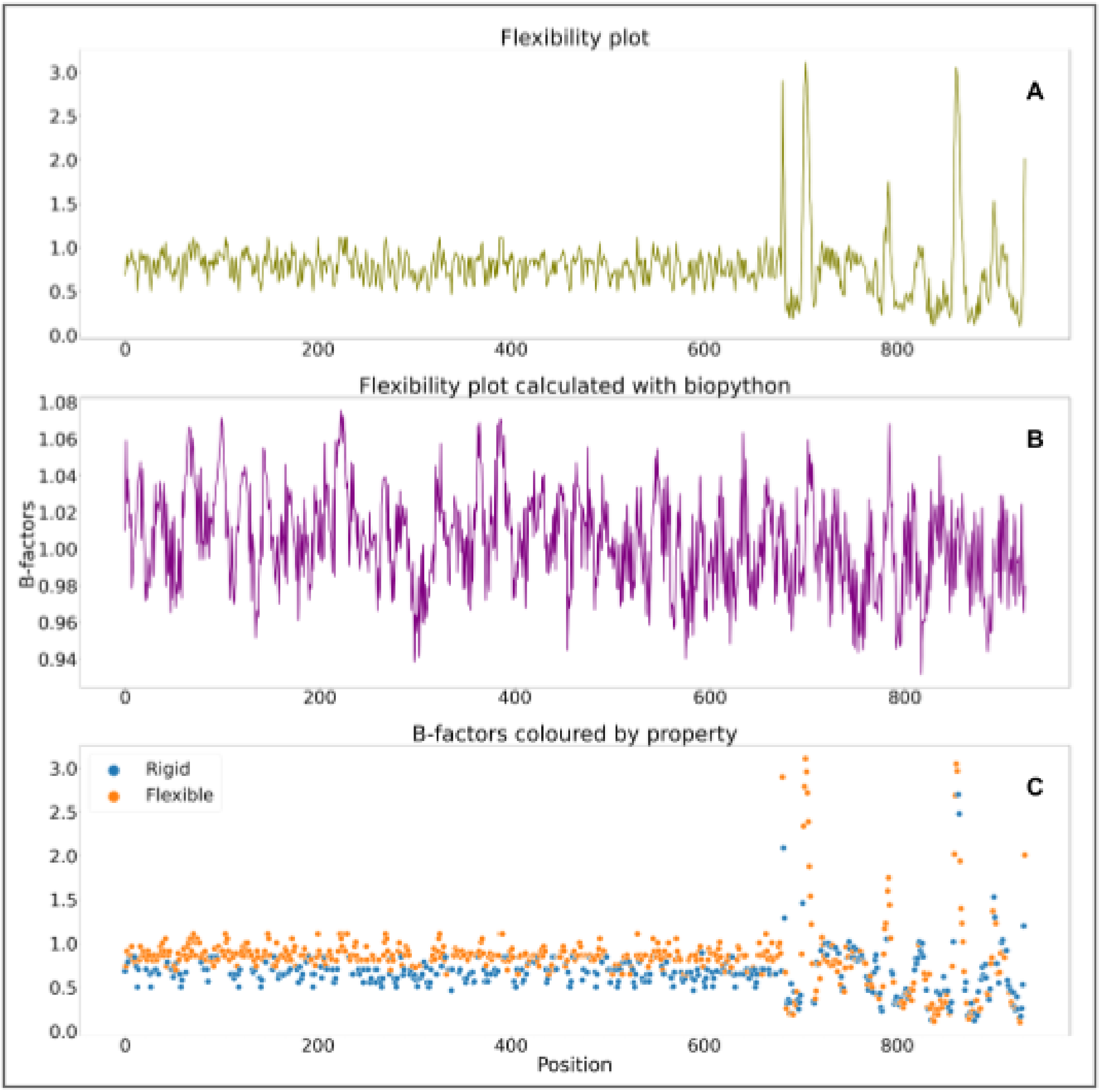
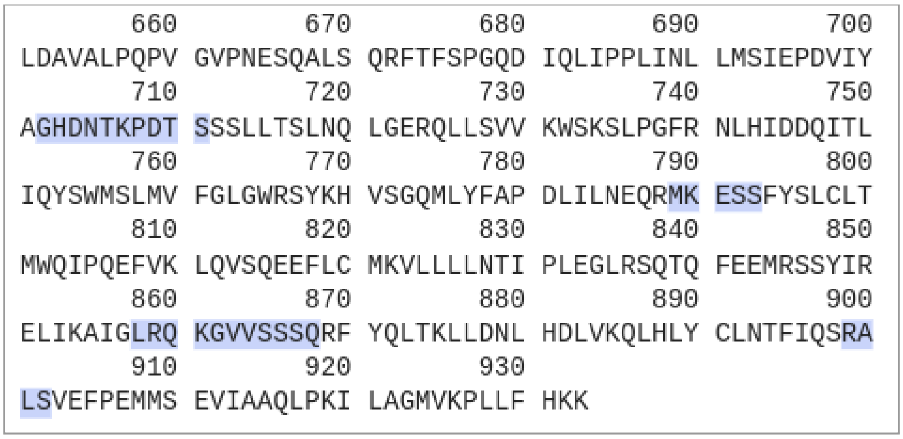


Figure 8. *ProtFLEXPred results of progesterone receptor protein from Homo Sapiens.* ***A****. Line plot that represents the flexibility profile predicted with the ProtFLEXpreD program (PFD profile). Residues 260 - 270 within the brown square.* ***B.*** *Line plot of the flexibility profile predicted by the biopython package (BP profile). [29].* ***C.*** *Scatter plot that represents the B-factors values coloured by flexibility or rigidity obtained by the ProtFLEXpreD program (PFD coloured). [8]*



```
       660        670        680        690        700
LDAVALPQPV GVPNESQALS QRFTFSPGQD IQLIPPLINL LMSIEPDVIY
       710        720        730        740        750
AGHDNTKPDT SSSLLTSLNQ LGERQLLSVV KWSKSLPGFR NLHIDDQITL
       760        770        780        790        800
IQYSWMSLMV FGLGWRSYKH VSGQMLYFAP DLILNEQRMK ESSFYSLCLT
       810        820        830        840        850
MWQIPQEFVK LQVSQEEFLC MKVLLLLNTI PLEGLRSQTQ FEEMRSSYIR
       860        870        880        890        900
ELIKAIGLRQ KGVVSSSQRF YQLTKLLDNL HDLVKQLHLY CLNTFIQSRA
       910        920        930
LSVEFPEMMS EVIAAQLPKI LAGMVKPLLF HKK
```

Figure 9. *Progesterone receptor protein (P06401) sequence from residue 651 to the end. Residues that belong to high B-factor peaks and that are explained in this report are highlighted in blue.*

Similar to the estrogen receptor, we find a peak of flexibility before the helix 12. When helix 12 cannot adopt its agonistic conformation, the correct formation of the AF-2 surface is prevented, causing the exclusion of the binding of co-activators. This helix can adopt two different positions depending on the biological activity of the ligand (**Fig. 10**), but it is worth mentioning that both of them differ from the unliganded conformation. [18]

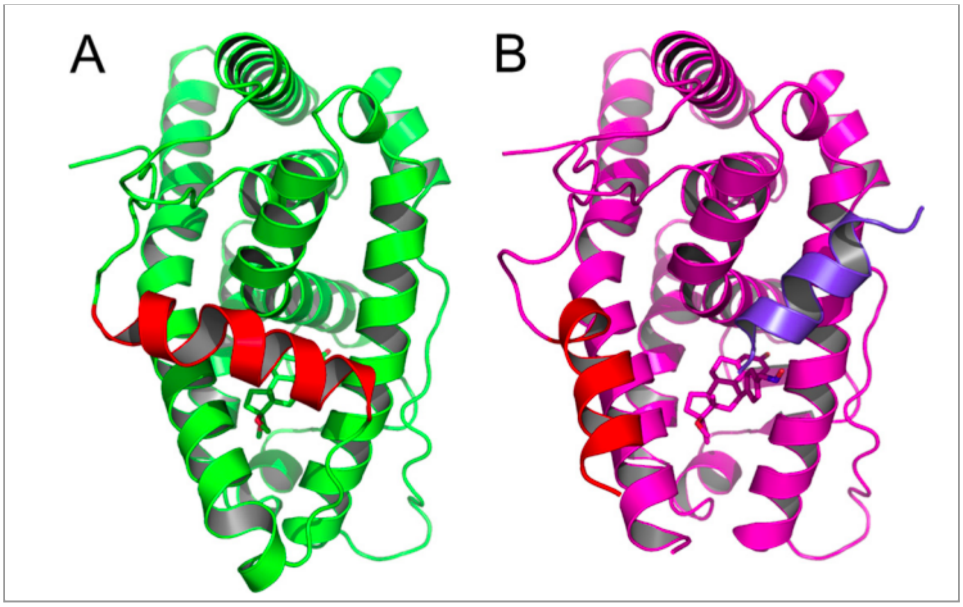


Figure 10. A. *Secondary structure of the PR-Agonist complex. Helix-12 is coloured red and is oriented in the classic agonist conformation.* ***B.*** *Secondary structure of a previously described PR-antagonist complex. Helix-12 is coloured red and is shifted from the agonist position to allow the binding of a co-repressor peptide, coloured blue. [18]*

The binding of a ligand has been shown to increase the flexibility of the helix 12 region, independently of if it is an agonist or antagonist. [17] This, together with the fact that the flanking regions of the helix are loops, are factors that can cause the high B-factors observed.

High B-factors are also found in residues 702-711, 787-793 and 858-868, all corresponding to loops at the N-terminal of helices 3, 6 and 10 respectively (**Fig. 11**). These loops have shown conformational flexibility. They have been demonstrated to cooperate in the helix 12 displacement when an antagonist binds to the receptor. In addition, the second flexible loop is involved in ligand entry, and the same happens in estrogen receptors. [17]

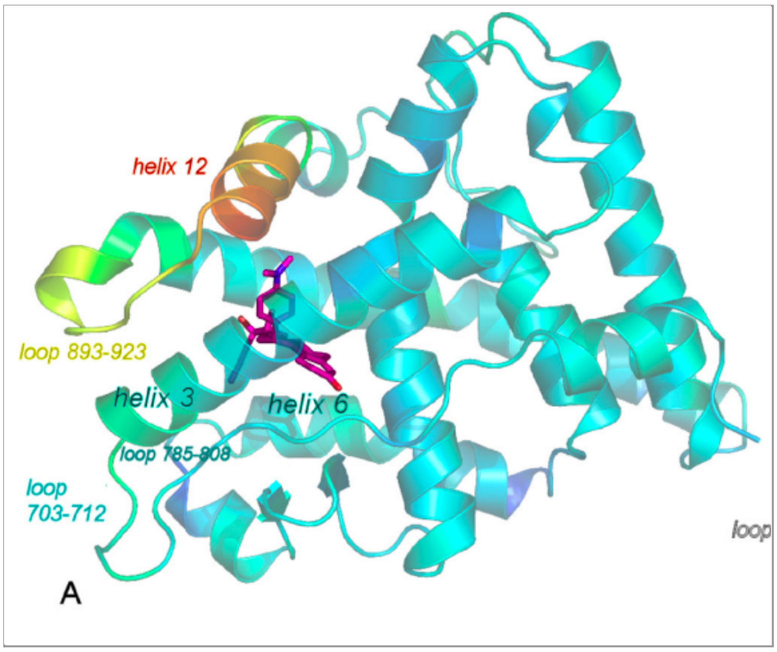


Figure 11. *Binding of a special antagonist (purple) in the progesterone receptor ligand-binding domain (blue/red). Ribbon view of the ligand-binding domain that binds to the antagonist. The coloring represents B-factor changes compared with a receptor-agonist complex. Changes range from -4 (blue) to +44 (red) Å^2 and are predominantly restricted to the loop 785– 808 and helix 12. The conformation of helix 12 is agonistic, closely packed against the LBD core. [17]*

PROTEIN KINASES FAMILY

Kinases are part of the larger family of phosphotransferases, which are enzymes that catalyze a process known as phosphorylation. It is characterized by the transfer of phosphate groups from high-energy molecules, such as ATP, to specific substrates. They modify lipids, carbohydrates or other molecules taking part in important biological functions, such as metabolism, cell signaling, protein regulation, cellular transport, secretory processes and other cellular pathways. [18, 19]

Protein kinases are enzymes that selectively modify proteins, in contrast to kinases, by covalently adding phosphates to them. Modifications promote changes in the functional activity of the target protein, such as enzyme activity, cellular location or association with other proteins. As kinases, they take part in important biological functions, especially, they are involved in signal transduction [19]. The human genome encodes over 500 protein kinases that correspond to nearly 2% of the entire genome[19, 20]. The most known types are Serine/Threonine kinases, which phosphorylate serines and threonines in their targets, and Tyrosine kinases, which phosphorylate tyrosines in their targets [19]. A less common type are protein-arginine kinases, they differ in mechanisms and structure from the previous protein-kinases explained [21].

Two protein-arginine kinases have been analyzed with the ProtFLEXpreD program. The results are explained below.

McsB PROTEIN (P65206)

McsB is a protein-arginine kinase that catalyzes the specific phosphorylation of arginine residues in proteins. It is codified by the mcsB gene in the Staphylococcus *aureus* species [23]. It regulates transcription factors and marks aberrant proteins for degradation taking part in the protein quality control process [22].

The protein is linearly organized as PD-DD-DD*-PD* (the asterisk denotes the partner protomer), where PD is the phosphotransferase domain and DD is the dimerization domain. The PD contains a nine-stranded β-sheet surrounded by seven α-helices, containing the active site at its center. [22] Specifically, there are two binding sites in positions 82 and 115 and three nucleotide bindings between positions 24-28, 166-170 and 197-202. [23] The DD domain is a four-helix bundle that is tightly associated with the PD domain. [22] It has been observed that the most flexible regions in a protein are the ones where the loops are [24], the most rigid regions are the ones where an active site is [25] and binding sites flexibility is variable. [24]

The predicted structure in Alpha Fold presents a high confidence score in almost all the protein, you can observe it in dark blue in **Fig. 12A**. Loops and N-terminal domain are the predicted structures that present low confident scores [26, 27]. We could expect high B-factors values in our flexibility analysis with the ProtFLEXpreD program in these regions.

There are no structures of McsB in the PDB, but there are from its homologous proteins. **Fig. 12B** shows the McsB structure of *Bacillus subtilis* species, obtained from the PDB [28]. Comparing both structures, it can be observed that they are very similar. In contrast to *Staphylococcus aureus* predicted structure, the N-terminal domain is well folded in the *Bacillus subtilis* structure. If the *Staphylococcus aureus* McsB has a similar structure to *Bacillus subtilis*McsB, we could expect a smooth profile in the N-terminal region.

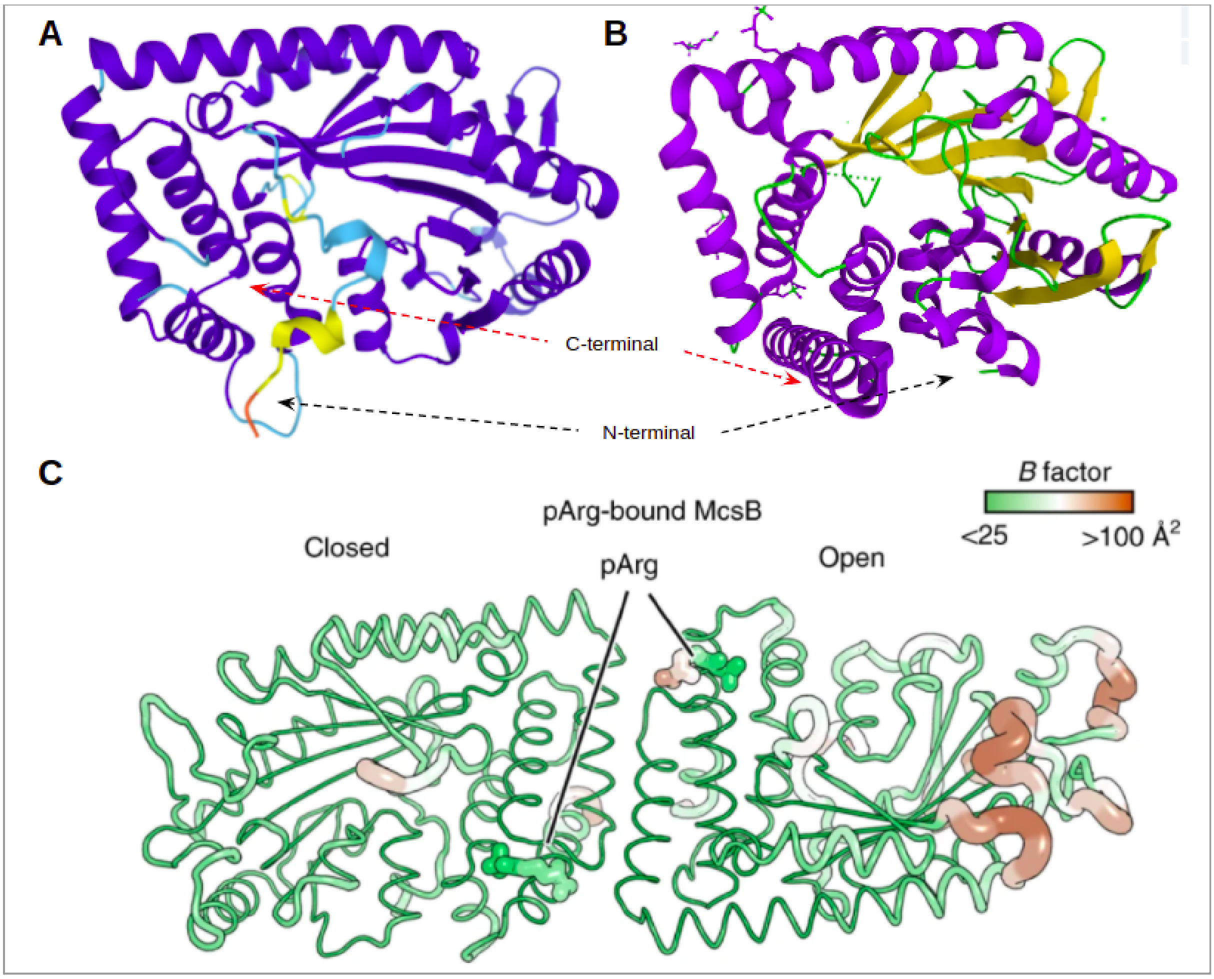


Figure 12. A. *Predicted structure of McsB in Staphylococcus aureus. High confidence score in dark blue followed, in descending order, by light blue, yellow and orange. [26, 27]* ***B.*** *Experimental structure of McsB in Bacilus subtilus. Helices are in purple, β-sheet strands in yellow and loops in green [28]* ***C.*** *Distribution of thermal motion factors of the pArg-bound structure. B factors are reflected in colour scheme and in cartoon putty radius. [22]*

The McsB protein results can be observed in **Fig. 13**. Comparing the PFD and the BP profiles we observed that they had a tendency to follow the same profile but the PFD profile was smoother than the BP profile. Moreover, there are three peaks of high B-factors values that are more visual in the PFD profile than in the BP profile. It could mean that our program better represents the flexibility profile of this protein because it is possible to better identify the most flexible regions in McsB protein.

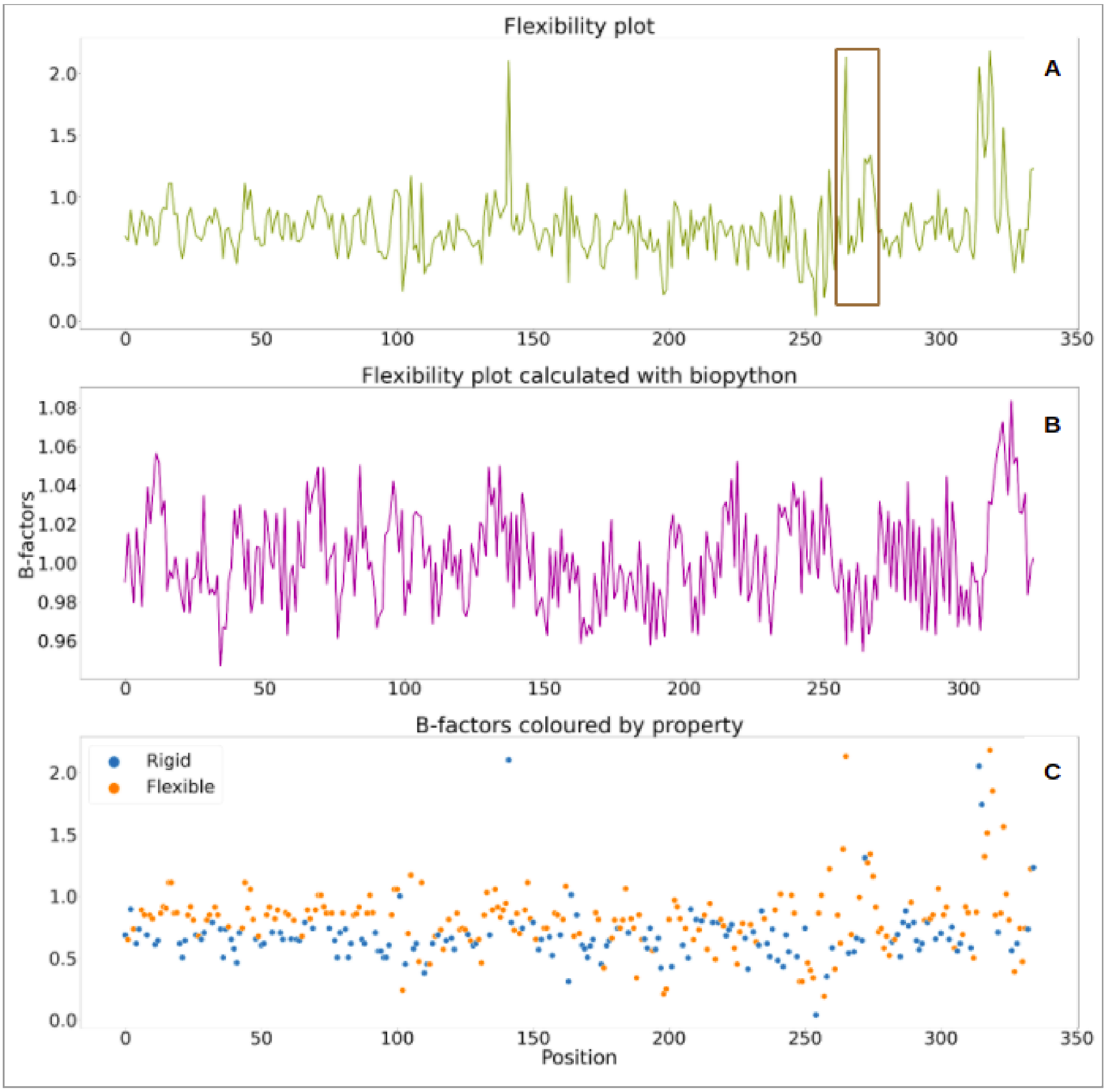


Figure 13. *ProtFLEXPred results of McsB protein of Staphylococcus aureus.* ***A****. Line plot that represents the flexibility profile predicted with the ProtFLEXpreD program (PFD profile). Residues 260 - 270 within the brown square.* ***B.*** *Line plot of the flexibility profile predicted by the biopython package (BP profile). [29].* ***C.*** *Scatter plot that represents the B-factors values coloured by flexibility or rigidity obtained by the ProtFLEXpreD program (PFD coloured). [8]*

As we explained above, we expected to see a smooth profile in the N-terminal domain and see the behavior in B-factors scores in binding sites and nucleotide sites, which are between residues 1 and 200 (**Fig. 14**). If we check the binding site positions, it seems that position 82 (**Fig. 14A**) and position 115 (**Fig. 14B**) have low B-factor scores. In the case of nucleotide bindings, we observed that positions 24-28 (**Fig. 14A**) and 197-202 (**Fig. 14B**) create a relative maximum of B-factors scores. In contrast, positions 166-170 (**Fig. 14B**) follow a low B-factors scores profile. Taking into account these observations, we could say that binding regions tend to have low B-factors scores in McsB protein, but not the highest ones. About the N-terminal domain, we observed that there is not a flexibility tendency, even with low confident scores in Alpha Fold, meaning that it could be folded as *Bacillus subtilis* McsB. In contrast, in the C-terminal region, we observed a peak of high B-factors scores when the predicted structure in Alpha Fold presents high confident scores in this region. In **Fig. 12B** we observed that N-terminal and C-terminal domains in the homologous protein are well folded, making it so that we do not expect high flexibility peaks in these regions. However, a high flexibility peak in C-terminal has been observed. Therefore, we could say that the confidence score in the predicted structure in Alpha Fold is not related to the flexibility profile as we expected and the folding of the homologous protein is not related to the flexibility profile as we expected. Another point of view of the high C-terminal flexibility could be its functional relationship. It is located in the second PD domain, characterized to have the promoter partner. A fact that could promote flexibility in this region. Analysis of the DD domain of the structure suggests that it may contain pArg-binding pockets, which would promote the formation of the pArg-bound McsB dimer (**Fig. 12C**). The McsB dimer shows an asymmetric conformation with two active sites, one closed and one opened, and an unequal distribution of B-factors [22]. This study [22] discovered that pArg-binding pocket is involved in the allosteric control of the enzymatic activity of the McsB kinase. These observations would be agreed with the high peak of flexibility above the residues 260 - 270 (**Fig. 12A**). And we could also think that this enzymatic activity could promote the high C-terminal flexibility that we observed in our results. It could be a hypothesis for future studies in this field.

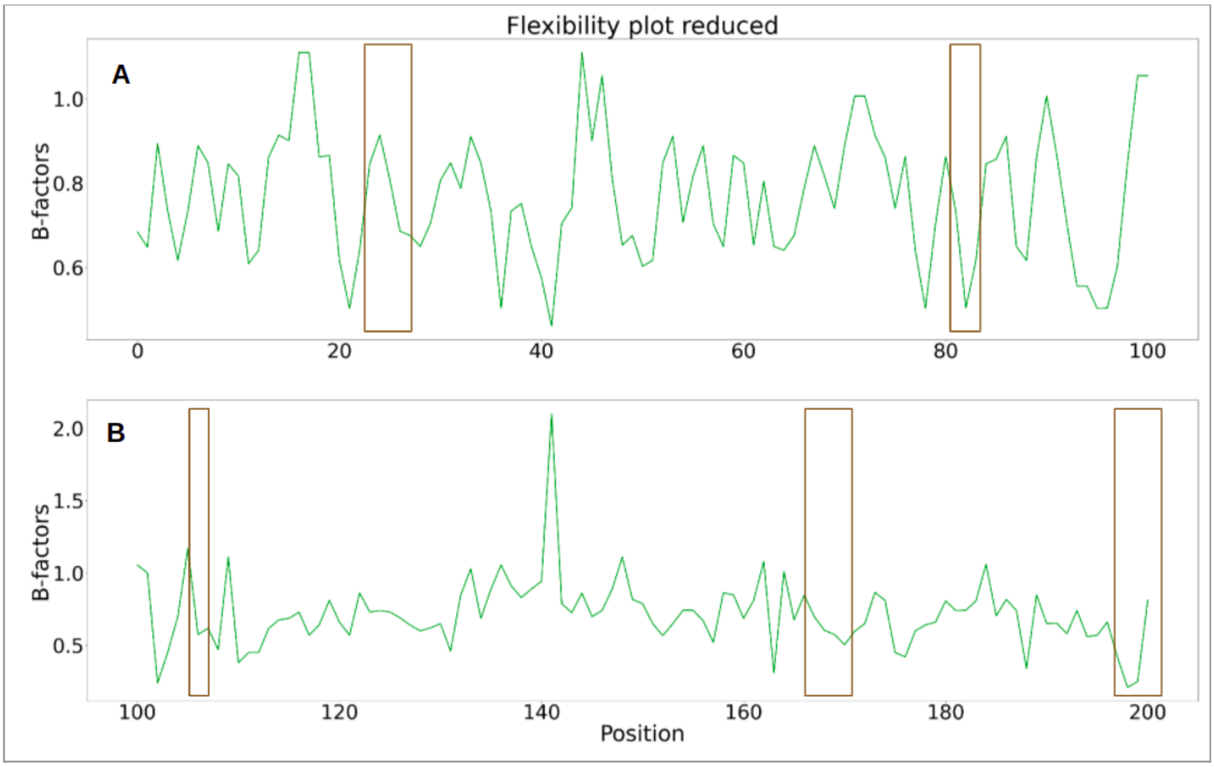


Figure 14. *ProtFLEXPred results of McsB protein of Staphylococcus aureus. A line plot that represents the flexibility profile predicted with the ProtFLEXpreD program (PFD profile).* ***A.*** *From residue 0 to 100. Residues 24 - 28 and 115 within the brown square.* ***B.*** *From residue 100 to 200. Residues 115, 166 - 170 and 197 - 202 within the brown square.*

Finally, the PFD coloured in **Fig. 13C** lets us observe that residues that we marked as flexible tend to have higher B-factors scores than residues marked as rigid. This observation could confirm our findings in the literature that some residues can be considered flexible and others rigid [8, 30]. Residues that do not follow this tendency could be an experimental error in writing the PDB B-factors or as important as to reject the hypothesis. It would be necessary to further study to get a consistent conclusion.

SRPK2 PROTEIN (P78362)

SRPK2 is a SRSF protein kinase 2 codified by the SRPK2 gene in *Homo sapiens*. [23] It belongs to the SR (serine/arginine rich proteins) family, whose members are very important in the regulation of mRNA metabolism. They are capable of interacting simultaneously with RNA and other protein components through RRM (RNA recognition motif) and RS domain (arginine and serine residues). [31] The genes that encode the SR family proteins are designated splicing factor (SF) [31] because they regulate the splicing [23], giving the name SRSF to these family proteins. They share their structure organization (**Fig. 15**), having one or two N-terminal RNA binding domains and a variable-length RS domain at C-terminal. The first one provides RNA-binding specificity and the second one allows protein-protein interaction. [31]

SRPK2 protein is located in the nucleus, cytoplasm, cytosol and chromosome. It phosphorylates serine residues located in RS rich regions of their substrates, promoting neuronal apoptosis by up-regulating cyclin-D1 (CCND1) expression, ACIN1 displacement to nucleoplasm to up-regulate cyclin A1, spliceosomal B complex formation and playing a negative role in the regulation of the HBV (hepatitis B virus) replication [23].

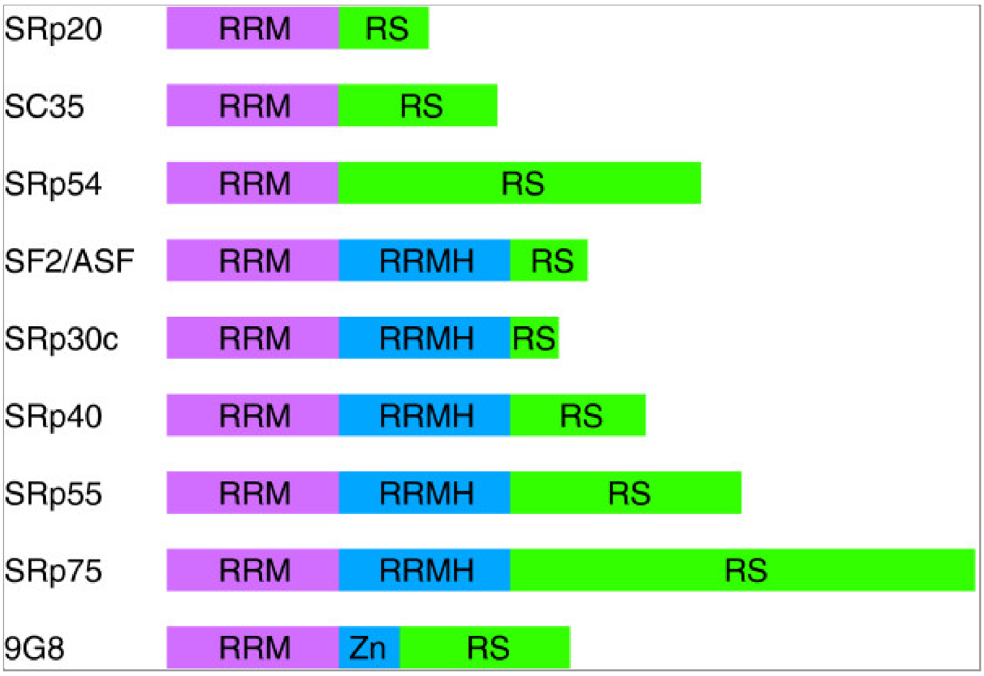


Figure 15. *Nine protein members of the human SR family. SC35 is a SRSF2 as SRPK2. RRM: RNA recognition motif; RRMH: RRM homology; RS: arginine/serine-rich domain; Zn: Zinc knuckle.*

The predicted structure in Alpha Fold is not very representative. It presents high confidence scores between residues 100 - 300 and 500 - 700, where some helices and a β-sheet strand can be observed (**Fig. 16A**). However, there are many regions that are not well folded making it a non-representative prediction [26, 27]. In the PDB there are two structures that represent SRPK2 (**Fig. 16C, 16D**), their IDs are 5MYV [32] and 2X7G [33]. Surprisingly, they are, respectively, in the second and third position in BLASTP results. The most homologous protein found is a SRPK1 (**Fig. 16B**), a SR family member, with 6FAD as PDB ID [34]. Checking the sequence length of these protein structures, we observed that the 5MYV and 2X7G sequence length is 389 residues [32, 33] and the 6FAD sequence length is 618 [34]. The query protein, SRPK2, has a sequence length of 688 residues [23], which lets us understand why the most homologous protein is 6FAD instead of 5MYV or 2X7G.

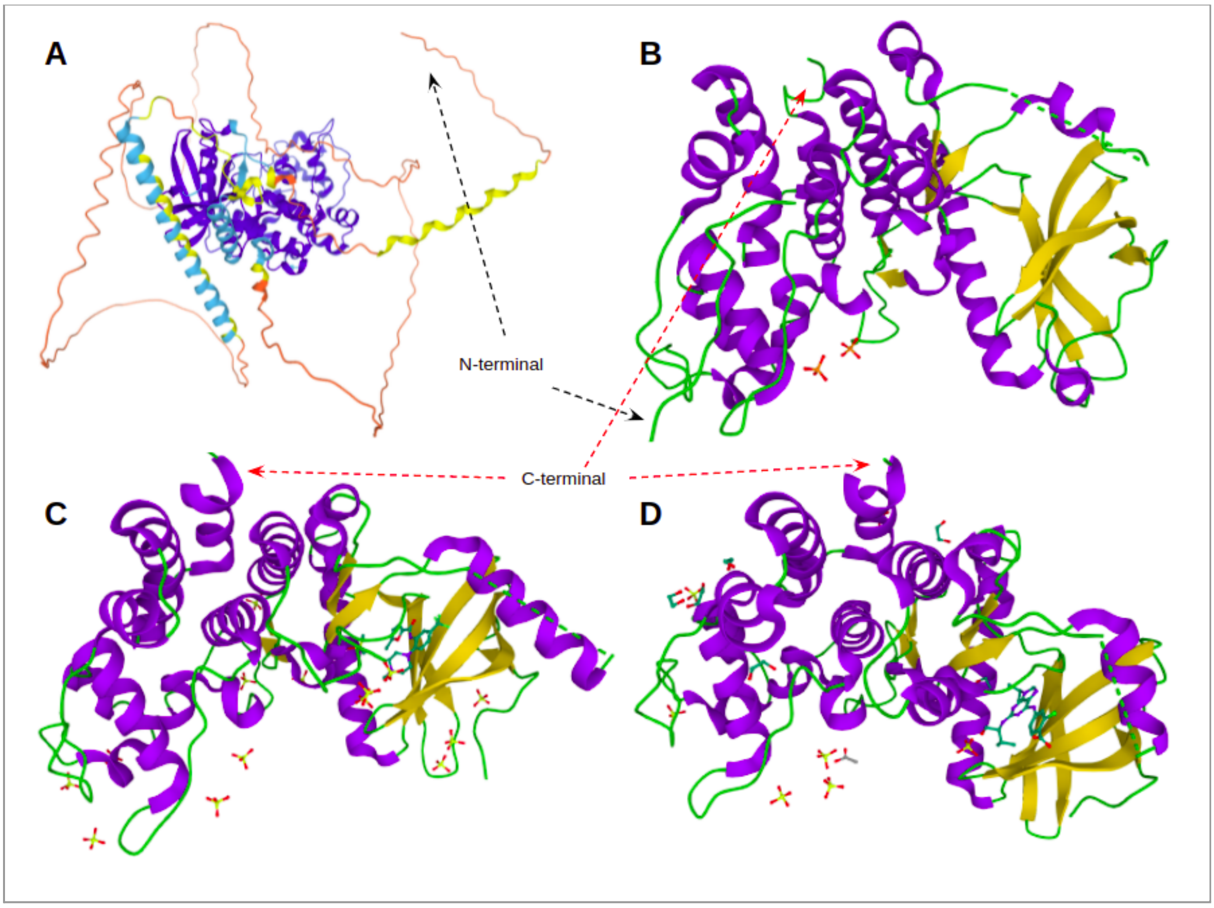


Figure 16. *Nine protein members of the human SR family. SC35 is a SRSF2 as SRPK2. RRM: RNA recognition motif; RRMH: RRM homology; RS: arginine/serine-rich domain; Zn: Zinc knuckle.*
A. *Predicted structure of SRPK2 in Homo sapiens. High confidence score in dark blue followed, in descending order, by light blue, yellow and orange. [26, 27]* ***B.*** *Experimental structure of SRPK1 in Homo sapiens with PDB ID 6FAD. [34]* ***C.*** *Experimental structure of SRPK2 in Homo sapiens with PDB ID 5MYV. [32]* ***D.*** *Experimental structure of SRPK1 in Homo sapiens with PDB ID 2X7G. [33] In* ***B***, ***C***, ***D***, *helices are in purple, β-sheet strands in yellow and loops in green.*

The structural characterisation of SR proteins is difficult due to the unknown phosphorylation state of serines within the RS domain and the free state's poor solubility [31], which may be another consideration to take into account analyzing the predicted and experimental structures in Alpha Fold and in PDB. Truncated versions of SRPK2 and SRPK1 allowed the discovery of their crystallographic structures. [35] However, as you can observe in **Fig. 16**, there is no information about the full-length structures. N-terminal region and SID, unstructured spacer insert domain between N-terminal and C-terminal regions, are unstructured [35]. Problem that also affects the predicted structure, which is really unclear

(**Fig. 16A**). [26, 27] SRPK2 is activated by phosphorylation on Ser-52 and Ser-588. It has an ATP binding site on residue 110, a proton acceptor active site in residue 214 and an ATP nucleotide binding in residues 87 - 95. [23] We expected to see low B-factors scores in active sites [25] and analyze the behavior of binding sites that is more variable. [24]

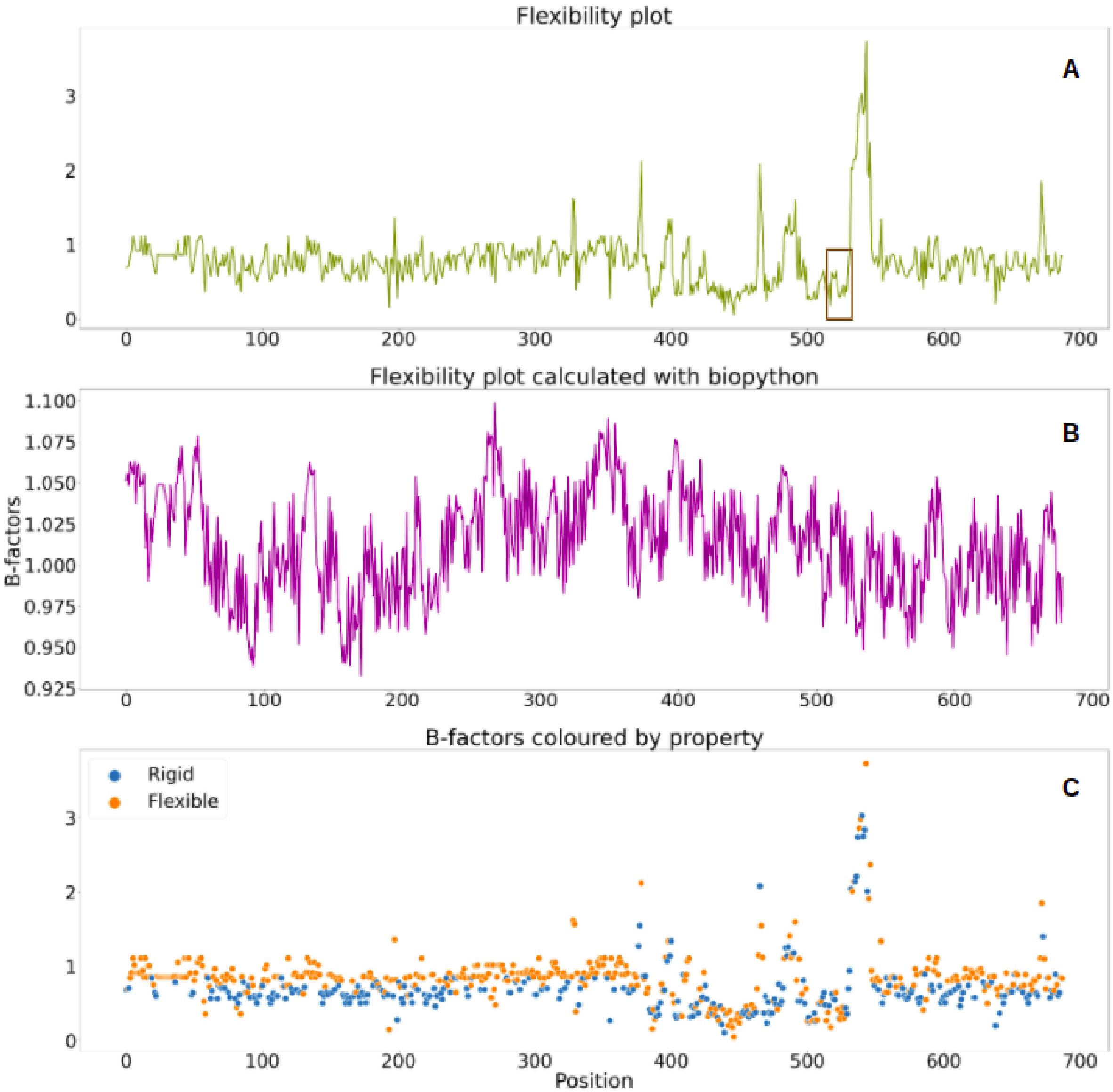


Figure 13. *ProtFLEXPred results of SRPK2 protein of Homo Sapiens.* ***A****. Line plot that represents the flexibility profile predicted with the ProtFLEXpreD program (PFD profile). Residues 260 - 270 within the brown square.* ***B.*** *Line plot of the flexibility profile predicted by the biopython package (BP profile). [29].* ***C.*** *Scatter plot that represents the B-factors values coloured by flexibility or rigidity obtained by the ProtFLEXpreD program (PFD coloured). [8, 30]*

SRPK2 protein results can be observed in **Fig. 13**. Comparing the PFD and the BP profiles we observed that they did not have a tendency to follow the same profile. Some maximum peaks that we observed in the PFD profile are not observed in the BP profile. It seems that our program better represents regions with high B-factors scores but the biopython function may better show its own flexibility profile because the difference between B-factors scores is smaller. It is an opposite observation to the one in McsB profiles, where we observed our program may better represent the flexibility profile of this protein. The flexibility profile from 1 to 250 residues is represented in **Fig. 14**. Between residues 87 - 95, nucleotide binding region, we can observe more high B-factor values than small, indicating that it may be a flexible region. A relative minimum in flexibility profile can be observed around residue 110, binding site, and residue 214, active site, confirming that these residues are located in rigid regions. These observations are agreed with ones observed by other studies. [24, 25] Analyzing the flexibility profile in Ser-52 (**Fig. 14A**) and Ser-588 (**Fig. 13A**), activity regulation residues, we observed high B-factor values, above 1, in both cases, indicating a flexible region. However, checking the results text file, we observed that there is not a serine in these positions. There is an asparagine in position 52, where the closest serine is shifted one position (residue 51) and there is a glycine in position 588, where the closest serine is shifted 9 (residue 607) and 10 (residue 568) positions. It could mean that these residues have been shifted due to mutations or fasta and pdb files in databases are not updated.

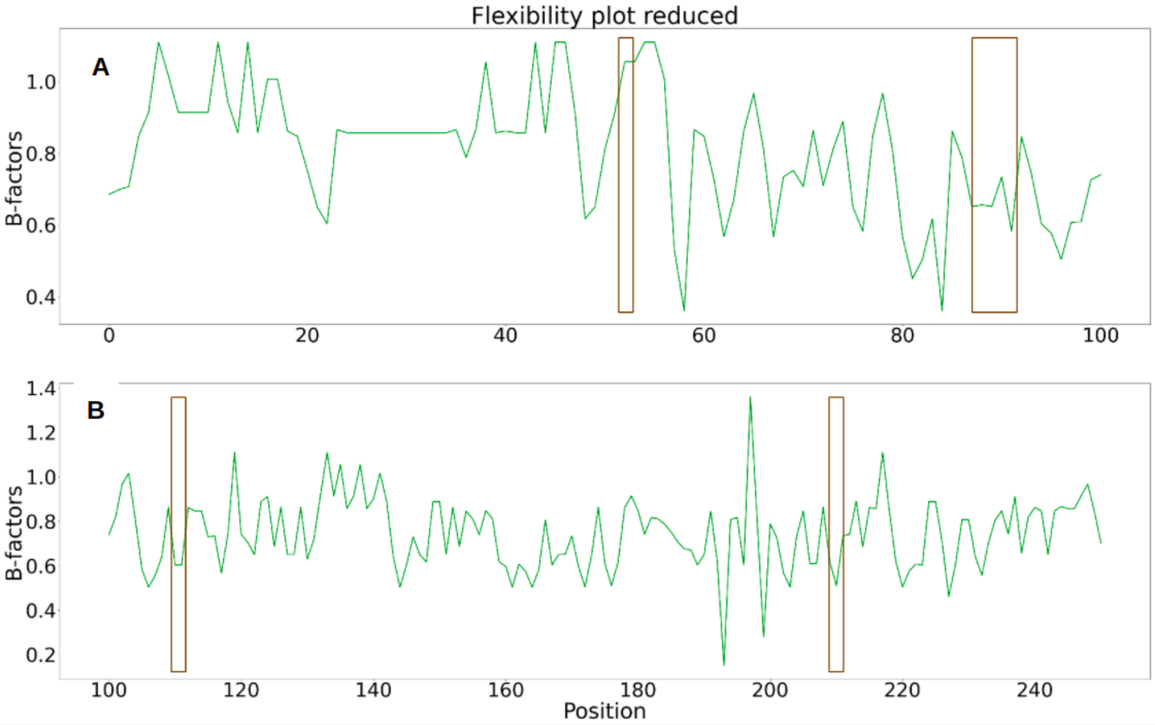


Figure 14. *ProtFLEXPred results of McsB protein of Staphylococcus aureus. Line plot that represents the flexibility profile predicted by the ProtFLEXpreD program (PFD profile).* ***A.*** *From residue 0 to 100. Residues 52 and 87 - 95 within brown square.* ***B.*** *From residue 100 to 250. Residues 110 and 214 within brown square.*

Finally, the PFD coloured in **Fig. 13C** lets us observe that residues that we marked as flexible tend to have higher B-factors scores than residues marked as rigid, as we observed in McsB protein, confirming again our findings in literature about the consideration of some residues as flexible and others as rigid [8, 30].

DISCUSSION

B-factors are experimentally values that can be affected by many factors such as overall resolution of the structure, crystal contacts, and, importantly, the particular refinement procedures. [4, 7] The degree of resolution achieved in the X-ray analysis can influence the resulting B-factors. Low resolution correlates with B-factors that are too high, the ones obtained from protein structures with a resolution ~1.5 Å are more likely to be more precise. But even with good resolution the obtained B-factors still have an uncertainty of around 15%. [4, 5, 7] The ProtFLEXpreD program uses B-factors from Protein Data Bank for the prediction of protein flexibility, which is an aspect to take into consideration, because the error that can be committed is proportional to the degree of resolution in the X-ray analysis.

Following this idea, we observed some incomplete results for proteins in PDB. Firstly, not the whole sequence is well structurally characterized. The number of residues in seqres lines is different from the number in atom lines and the residue location differs between the linear and folded sequence. Secondly, proteins that have the same B-factor value in each amino acid of the sequences. Thirdly, there are positions where a residue has been located but the specific amino acid is unknown. Finally, there are cases where an amino acid has been characterized twice and considered twice (we find A and B in the file) in atom lines when it appears only once in seqres lines. We tried to skip these inconveniences in our program, using the first amino acid characterized from the atom lines, considering unknown amino acids as X and calculating its B-factor according to its neighbors, and skipping homologous proteins that have the same B-factor all along the different residues. Taking into account that we have obtained reasonable and mostly justifiable results, we could say that we have managed these inconveniences well. However, problems that we did not take into account could appear and raise errors in our program in the future. Because the PDB is being replaced by 3DB, the Three-Dimensional Database of Biomolecular Structures [36], this new database could solve problems that we observed and the possible ones that can arise.

Studies that predicted protein flexibility using B-factors [8, 30] considered removing the first three amino acids in N-terminal and C-terminal domains because terminal residues are usually very flexible [30]. Others considered not removing them because they are not the most flexible regions and thus, only introduce noise. [8] Because in our proteins we did not encounter this problem, we decided to use the whole sequence without removing any residue, obtaining reasonable and mostly justifiable results. But this doesn’t exclude the possibility of this from happening. In addition, amino acids were divided into rigid (11 amino acids) or flexible (9 amino acids) in these studies. [8, 30] Taking this into account, we marked the residues of the sequence by this property. Results showed that flexible amino acids had higher B-factors values than rigid amino acids, which would confirm the observations of the previous studies. [8, 30] Therefore, this could be important to keep in mind for further studies.

Because our program heavily relies on homology and the PDB, this can introduce several limitations. The query protein may not have good enough homologues when performing the BLAST to extract the B-factors; we have tried to circumvent this problem by adding the possibility to do a computation based on B-factors found in the Smith et al. 2013 [8] publication. The hindrances caused by the use of PDB have already been mentioned.

All in all, we believe that we have created a good program to predict protein flexibility from the sequence using B-factors. Because, as we have seen, protein flexibility and its study are key to understanding how proteins respond to changes and modifications enabling their function, which can be used for genomic studies, disease research, and drug-design.

BIBLIOGRAPHY

[1] Teilum, K., Olsen, J. G., & Kragelund, B. B. (2011). Protein stability, flexibility and function. *Biochimica et biophysica acta*, *1814*(8), 969–976.

[2] Yaseen, A., Nijim, M., Williams, B. *et al.* (2016). FLEXc: protein flexibility prediction using context-based statistics, predicted structural features, and sequence information. *BMC Bioinformatics* 17, 281.

[3] Vera, R., Synsmir-Zizzamia, M., Ojinnaka, S., & Snyder, D. A. (2018). Prediction of protein flexibility using a conformationally restrained contact map. *Proteins*, *86*(10), 1111–1116.

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

[5] 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.

[6] Vander Meersche, Y., Cretin, G., de Brevern, A. G., Gelly, J. C., & Galochkina, T. (2021). MEDUSA: Prediction of Protein Flexibility from Sequence. *Journal of molecular biology*, *433*(11), 166882.

[7] Dong, Q., Wang, K., Liu, B., & Liu, X. (2016). Characterization and Prediction of Protein Flexibility Based on Structural Alphabets. *BioMed research international*, *2016*, 4628025.

[8] Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K., & Zhu, G. (2003). Improved amino acid flexibility parameters. *Protein science: a publication of the Protein Society*, *12*(5), 1060–1072.

[9] Weikum, E. R., Liu, X., & Ortlund, E. A. (2018). The nuclear receptor superfamily: A structural perspective. *Protein science: a publication of the Protein Society*, *27*(11), 1876–1892.

[10] Huang, P., Chandra, V., & Rastinejad, F. (2010). Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. *Annual review of physiology*, *72*, 247–272.

[11] Helsen, C., Kerkhofs, S., Clinckemalie, L., Spans, L., Laurent, M., Boonen, S., Vanderschueren, D., & Claessens, F. (2012). Structural basis for nuclear hormone receptor DNA binding. *Molecular and cellular endocrinology*, *348*(2), 411–417.

[12] Kumar, R., & Thompson, E. B. (1999). The structure of the nuclear hormone receptors. *Steroids*, *64*(5), 310–319.

[13] Ribeiro, R. C., Kushner, P. J., & Baxter, J. D. (1995). The nuclear hormone receptor gene superfamily. *Annual review of medicine*, *46*, 443–453.

[14] Brzozowski, A. M., Pike, A. C., Dauter, Z., Hubbard, R. E., Bonn, T., Engström, O., Ohman, L., Greene, G. L., Gustafsson, J. A., & Carlquist, M. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, *389*(6652), 753–758.

[15] Shiau, A. K., Barstad, D., Loria, P. M., Cheng, L., Kushner, P. J., Agard, D. A., & Greene, G. L. (1998). The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, *95*(7), 927–937.

[16] Lusher, S. J., Raaijmakers, H. C., Vu-Pham, D., Dechering, K., Lam, T. W., Brown, A. R., Hamilton, N. M., Nimz, O., Bosch, R., McGuire, R., Oubrie, A., & de Vlieg, J. (2011). Structural basis for agonism and antagonism for a set of chemically related progesterone receptor modulators. *The Journal of biological chemistry*, *286*(40), 35079–35086.

[17] Raaijmakers, H. C., Versteegh, J. E., & Uitdehaag, J. C. (2009). The X-ray structure of RU486 bound to the progesterone receptor in a destabilized agonistic conformation. *The Journal of biological chemistry*, *284*(29), 19572–19579.

[18] Lusher, S. J., Raaijmakers, H. C., Vu-Pham, D., Kazemier, B., Bosch, R., McGuire, R., Azevedo, R., Hamersma, H., Dechering, K., Oubrie, A., van Duin, M., & de Vlieg, J. (2012). X-ray structures of progesterone receptor ligand binding domain in its agonist state reveal differing mechanisms for mixed profiles of 11β-substituted steroids. *The Journal of biological chemistry*, *287*(24), 20333–20343.

[19] Cooper, J. A. (2018, July 23). kinase. Encyclopedia Britannica.

[20] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., & Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science (New York, N.Y.), 298(5600), 1912–1934.

[21] Taylor, S. S., & Kornev, A. P. (2011). Protein kinases: evolution of dynamic regulatory proteins. Trends in biochemical sciences, 36(2), 65–77.

[22] Suskiewicz, M. J., Hajdusits, B., Beveridge, R., Heuck, A., Vu, L. D., Kurzbauer, R., Hauer, K., Thoeny, V., Rumpel, K., Mechtler, K., Meinhart, A., & Clausen, T. (2019). Structure of McsB, a protein kinase for regulated arginine phosphorylation. Nature chemical biology, 15(5), 510–518.

[23] UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic acids research, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

[24] Gunasekaran, K., & Nussinov, R. (2007). How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. Journal of molecular biology, 365(1), 257–273.

[25] Yuan, Z., Zhao, J., & Wang, Z. X. (2003). Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein engineering, 16(2), 109–114. https://doi.org/10.1093/proeng/gzg014

[26] Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

[27] 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).

[28] Hajdusits, B., Suskiewicz, M. J., Hundt, N., Meinhart, A., Kurzbauer, R., Leodolter, J., Kukura, P., & Clausen, T. (2021). McsB forms a gated kinase chamber to mark aberrant bacterial proteins for degradation. eLife, 10, e63505.

[29] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England), 25(11), 1422–1423.

[30] Vihinen, M., Torkkila, E., & Riikonen, P. (1994). Accuracy of protein flexibility predictions. Proteins, 19(2), 141–149.

[31] Shepard, P. J., & Hertel, K. J. (2009). The SR protein family. Genome biology, 10(10),

[32] Batson, J., Toop, H. D., Redondo, C., Babaei-Jadidi, R., Chaikuad, A., Wearmouth, S. F., Gibbons, B., Allen, C., Tallant, C., Zhang, J., Du, C., Hancox, J. C., Hawtrey, T., Da Rocha, J., Griffith, R., Knapp, S., Bates, D. O., & Morris, J. C. (2017). Development of Potent, Selective SRPK1 Inhibitors as Potential Topical Therapeutics for Neovascular Eye Disease. ACS chemical biology, 12(3), 825–832.

[33] Pike, A.C.W., Savitsky, P., Fedorov, O., Krojer, T., Ugochukwu, E., von Delft, F., Gileadi, O., Edwards, A., Arrowsmith, C.H., Weigelt, J., Bountra, C., Knapp, S. (2010) Structure of Human Serine-Arginine-Rich Protein- Specific Kinase 2 (Srpk2) Bound to Purvalanol B.
[34] Tunnicliffe, R. B., Hu, W. K., Wu, M. Y., Levy, C., Mould, A. P., McKenzie, E. A., Sandri-Goldin, R. M., & Golovanov, A. P. (2019). Molecular Mechanism of SR Protein Kinase 1 Inhibition by the Herpes Virus Protein ICP27. mBio, 10(5), e02551-19.
[35] Barbosa, É., Seraphim, T. V., Gandin, C. A., Teixeira, L. F., da Silva, R., Righetto, G. L., Goncalves, K. A., Vasconcellos, R. S., Almeida, M. R., Silva Júnior, A., Fietto, J., Kobarg, J., Gileadi, C., Massirer, K. B., Borges, J. C., de Oliveira Neto, M., & Bressan, G. C. (2019). Insights into the full-length SRPK2 structure and its hydrodynamic behavior. International journal of biological macromolecules, 137, 205–214.
[36] Abola, E. E., Manning, N. O., Prilusky, J., Stampf, D. R., & Sussman, J. L. (1996). The Protein Data Bank: Current Status and Future Challenges. Journal of research of the National Institute of Standards and Technology, 101(3), 231–241.