TAPASS : Tool for Annotation of Protein Amyloidogenicty in the context of other Structural States

Andrey KAJAVA: andrey.kajava@crbm.cnrs.fr
Théo FALGARONE: theo.falgarone@crbm.cnrs.fr

Centre de Recherche de Biologie Cellulaire de Montpellier
1919, route de Mende
34293 – Montpellier Cedex 05

1 Introduction

TAPASS (Tool for Annotation of Protein Amyloidogenicity in the context of other Structural States) provides consensual results on the occurrence and distribution of amyloid-forming regions in proteins assessed through the prism of the overall structural context. The pipeline allows the detection of Exposed Amyloidogenic Regions (EARs). It gather a total of 11 bioinformatic tools, each with a specific function.

2 TAPASS usage

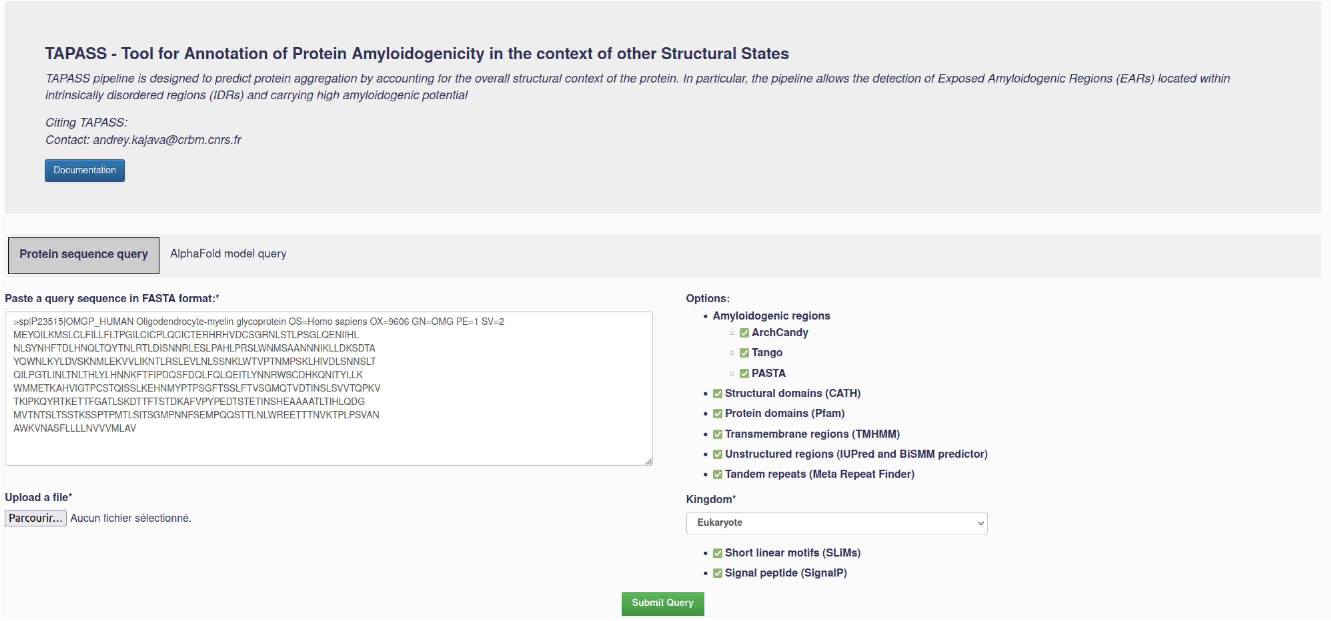
2.1 Protein sequence query

2.1.1 Sequence input

Input is a single protein sequence, either pasted or uploaded as a file in FASTA format. The sequence must contain a header starting by ">" and followed by the protein ID.

Example:

```
>sp|P23515|OMGP_HUMAN Oligodendrocyte-myelin glycoprotein OS=Homo sapiens OX=9606
GN=OMG PE=1 SV=2
MEYQILKMSLCLFILLFLTPGILCICPLQCICTERHRHVDCSGRNLSTLPSGLQENIIHL
NLSYNHFTDLHNQLTQYTNLRTLDISNNRLESLPAHLPRSLWNMSAANNNIKLLDKSDTA
YQWNLKYLDVSKNMLEKVVLIKNTLRSLEVLNLSSNKLWTVPTNMPSKLHIVDLSNNSLT
QILPGTLINLTNLTHLYLHNNKFTFIPDQSFDQLFQLQEITLYNNRWSCDHKQNITYLLK
WMMETKAHVIGTPCSTQISSLKEHNMYPTPSGFTSSLFTVSGMQTVDTINSLSVVTQPKV
TKIPKQYRTKETTFGATLSKDTTFTSTDKAFVPYPEDTSTETINSHEAAAATLTIHLQDG
MVTNTSLTSSTKSSPTPMTLSITSGMPNNFSEMPQQSTTLNLWREETTTNVKTPLPSVAN
AWKVNASFLLLLNVVVMLAV
```


TAPASS - Tool for Annotation of Protein Amyloidogenicity in the context of other Structural States

TAPASS pipeline is designed to predict protein aggregation by accounting for the overall structural context of the protein. In particular, the pipeline allows the detection of Exposed Amyloidogenic Regions (EARs) located within intrinsically disordered regions (IDRs) and carrying high amyloidogenic potential

Citing TAPASS:
Contact: andrey.kajava@crbm.cnrs.fr

Paste a query sequence in FASTA format:*

```
>sp|P23515|OMGP_HUMAN Oligodendrocyte-myelin glycoprotein OS=Homo sapiens OX=9606 GN=OMG PE=1 SV=2
MEYQILKMSLCLFILLFLTPGILCICPLQCICTERHRHVDCSGRNLSTLPSGLQENIIHL
NLSYNHFTDLHNQLTQYTNLRTLDISNNRLESLPAHLPRSLWNMSAANNNIKLLDKSDTA
YQWNLKYLDVSKNMLEKVVLIKNTLRSLEVLNLSSNKLWTVPTNMPSKLHIVDLSNNSLT
QILPGTLINLTNLTHLYLHNNKFTFIPDQSFDQLFQLQEITLYNNRWSCDHKQNITYLLK
WMMETKAHVIGTPCSTQISSLKEHNMYPTPSGFTSSLFTVSGMQTVDTINSLSVVTQPKV
TKIPKQYRTKETTFGATLSKDTTFTSTDKAFVPYPEDTSTETINSHEAAAATLTIHLQDG
MVTNTSLTSSTKSSPTPMTLSITSGMPNNFSEMPQQSTTLNLWREETTTNVKTPLPSVAN
AWKVNASFLLLLNVVVMLAV
```

Upload a file*

Parcourir... Aucun fichier sélectionné.

Options:

- **Amyloidogenic regions**
  - ☑ **ArchCandy**
  - ☑ **Tango**
  - ☑ **PASTA**
- ☑ **Structural domains (CATH)**
- ☑ **Protein domains (Pfam)**
- ☑ **Transmembrane regions (TMHMM)**
- ☑ **Unstructured regions (IUPred and BiSMM predictor)**
- ☑ **Tandem repeats (Meta Repeat Finder)**

Kingdom*

Eukaryote

- ☑ **Short linear motifs (SLiMs)**
- ☑ **Signal peptide (SignalP)**

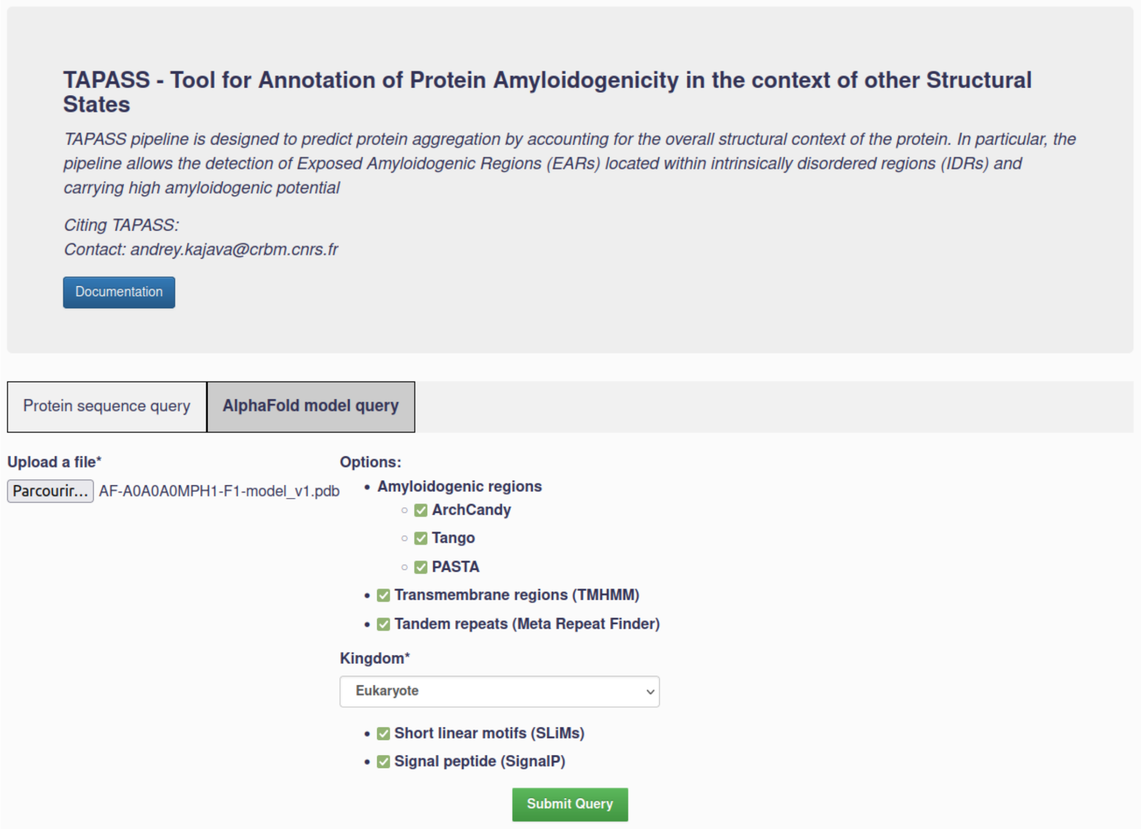
2.1.2 Option selection

TAPASS contains 11 distinctive bioinformatic tools (see more details in section 3). By default all tools are executed, but users can unselect them if they want to exclude some tools from the analysis. The kingdom parameter by default is *'Eukaryote'*, but it is possible to change it to *'gram-'* or *'gram+'*. This choice will affect the prediction of SignalP and SLiMs. In case the protein's origin is not known we recommend the selection of *'Eukaryote'*.

2.2 AlphaFold model query

2.2.1 File input

This mode allows to input an AlphaFold model recorded as a pdb file. The disordered regions are determined by a combination of the confidence score (pLDDT) given by AlphaFold and the relative accessible surface area (RASA) obtained by using DSSP. We consider a region as disordered if the pLDDT is lower than 70 and if in a window of 10 residues at least eight of them are exposed to the solvent (RASA > 0.15). Note that this mode does not have 'CATH', 'IUPred' and 'BiSMM' as they were meant to determine IDRs, which is now detected by using AlphaFold model.



TAPASS - Tool for Annotation of Protein Amyloidogenicity in the context of other Structural States

TAPASS pipeline is designed to predict protein aggregation by accounting for the overall structural context of the protein. In particular, the pipeline allows the detection of Exposed Amyloidogenic Regions (EARs) located within intrinsically disordered regions (IDRs) and carrying high amyloidogenic potential

Citing TAPASS:

Contact: andrey.kajava@crbm.cnrs.fr

Upload a file*

Parcourir... AF-A0A0A0MPH1-F1-model_v1.pdb

Options:

- **Amyloidogenic regions**
  - ☑ **ArchCandy**
  - ☑ **Tango**
  - ☑ **PASTA**
- ☑ **Transmembrane regions (TMHMM)**
- ☑ **Tandem repeats (Meta Repeat Finder)**

Kingdom*

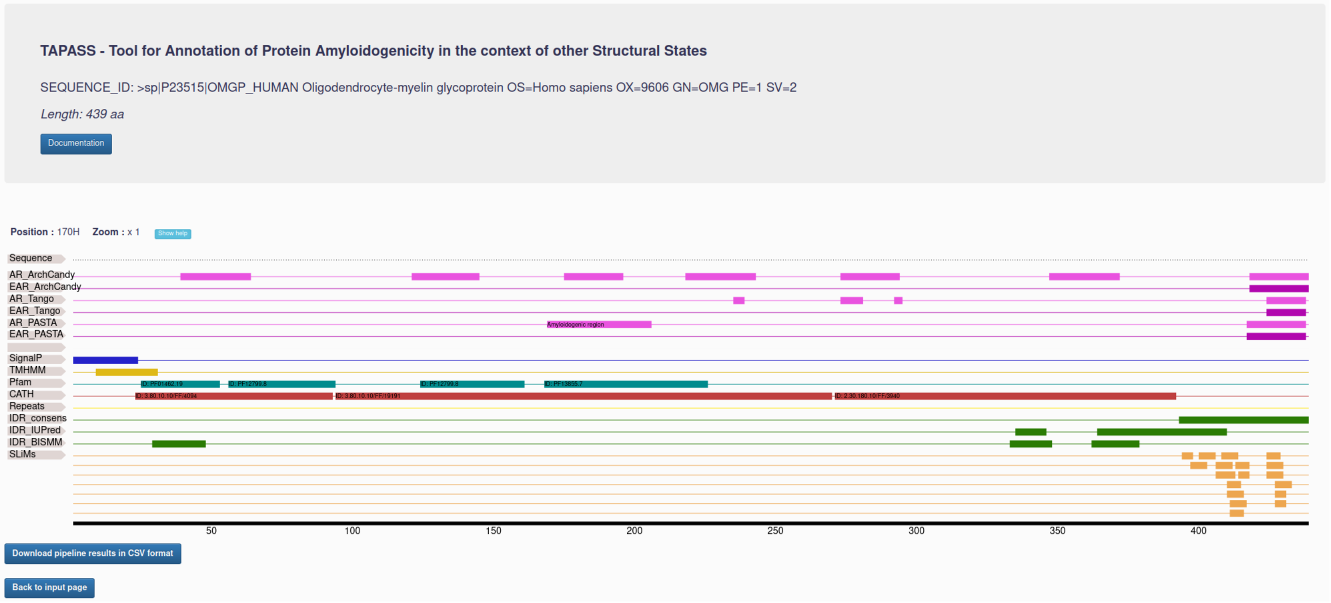
Eukaryote

- ☑ **Short linear motifs (SLiMs)**
- ☑ **Signal peptide (SignalP)**

2.3 Output

2.3.1 Graphical view

Predicted regions are represented by coloured boxes. ARs (light pink) and EARs (purple), which are the main focus of the pipeline, are grouped at the upper part of the output plot.



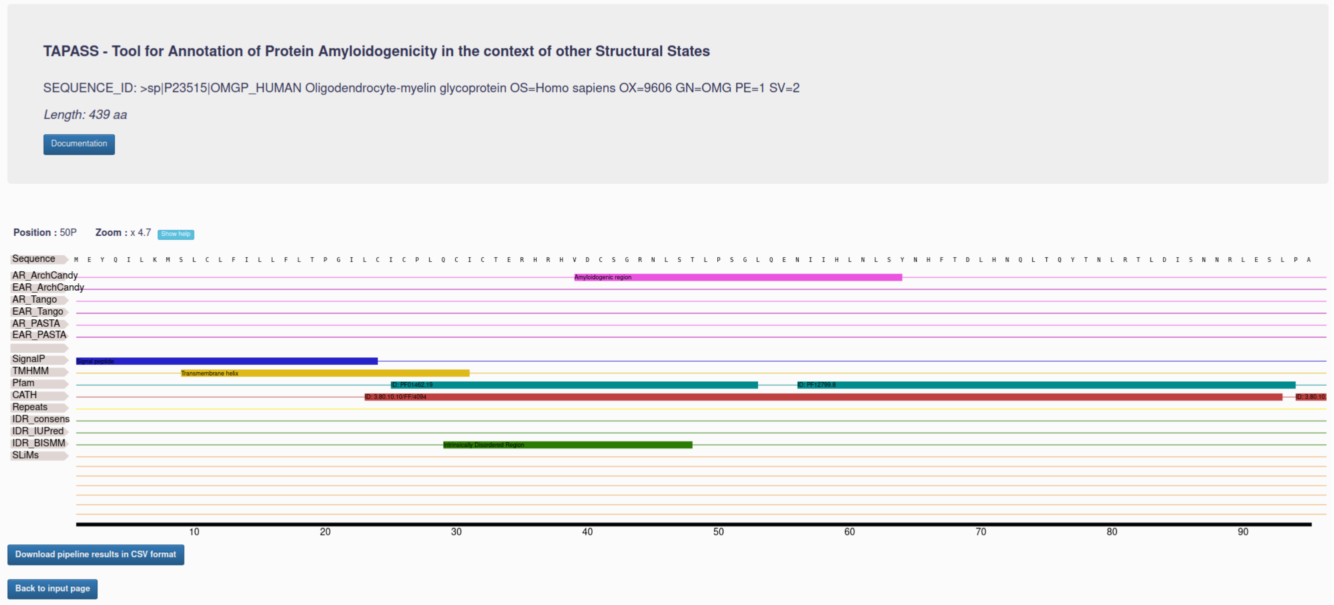
TAPASS - Tool for Annotation of Protein Amyloidogenicity in the context of other Structural States

SEQUENCE_ID: >sp|P23515|OMGP_HUMAN Oligodendrocyte-myelin glycoprotein OS=Homo sapiens OX=9606 GN=OMG PE=1 SV=2

Length: 439 aa

Position : 170H **Zoom :** x 1

It is possible for users to zoom in an area of interest by using the mouse left-click, the zoom out can be done by using the mouse right-click.



TAPASS - Tool for Annotation of Protein Amyloidogenicity in the context of other Structural States

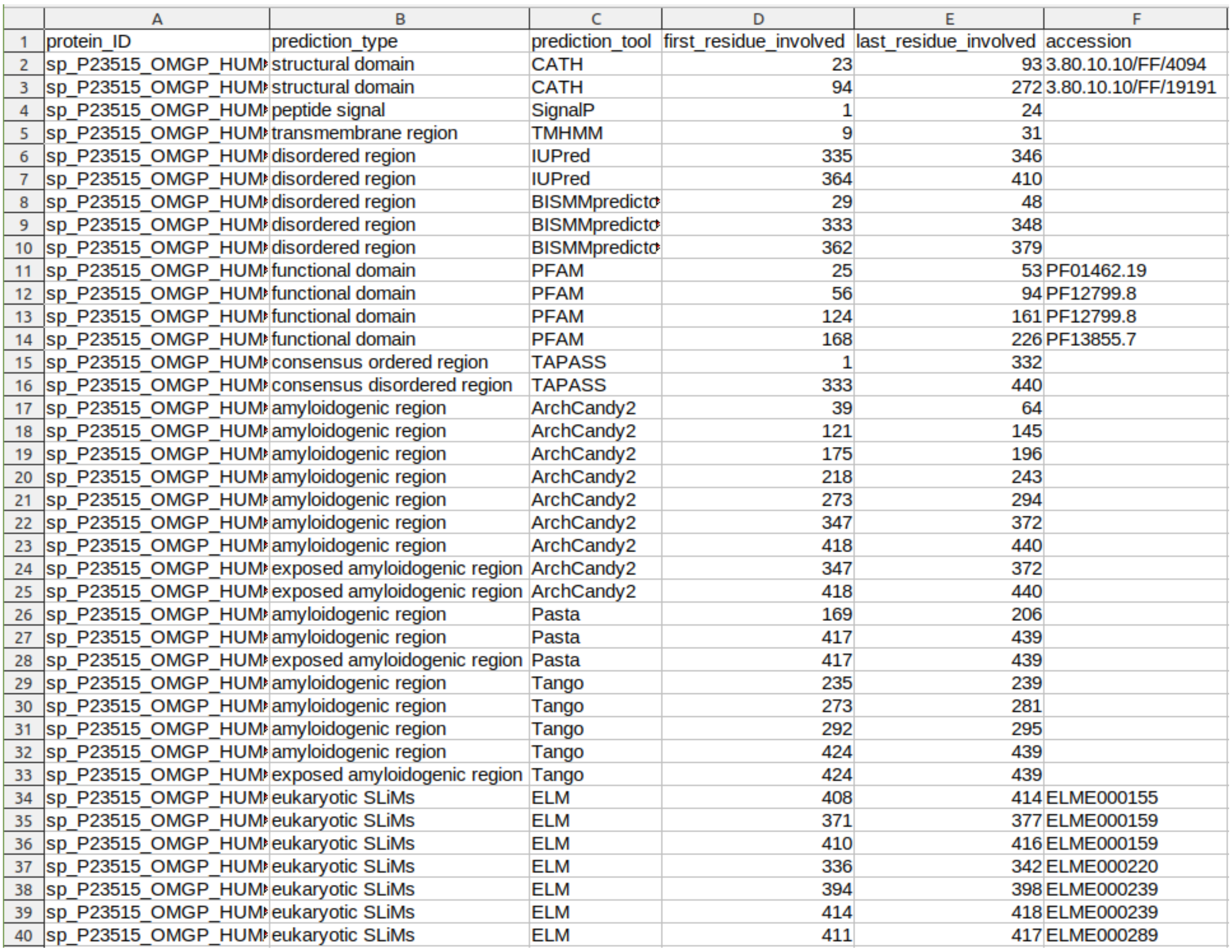
SEQUENCE_ID: >sp|P23515|OMGP_HUMAN Oligodendrocyte-myelin glycoprotein OS=Homo sapiens OX=9606 GN=OMG PE=1 SV=2

Length: 439 aa

2.3.2 CSV file

A CSV file summarising the prediction results can be download by clickling on the button at the bottom of the page *'Download pipeline results in CSV format'*. Each row represents a prediction and contain six columns :

- protein_ID : protein's identifier
- prediction_type : type of the predicted region (amyloidogenic region, transmembrane region, disordered region,...)
- prediction_tool : tool used in this prediction (ArchCandy2, Pasta, Tango, CATH,...)
- first_residu_involved : start position of the prediction
- last_residu_involved : end position of the prediction
- accession : accession identifier for CATH, PFAM and SLiMs predictions



	A	B	C	D	E	F
1	protein_ID	prediction_type	prediction_tool	first_residue_involved	last_residue_involved	accession
2	sp_P23515_OMGP_HUM	structural domain	CATH	23	93	3.80.10.10/FF/4094
3	sp_P23515_OMGP_HUM	structural domain	CATH	94	272	3.80.10.10/FF/19191
4	sp_P23515_OMGP_HUM	peptide signal	SignalP	1	24	
5	sp_P23515_OMGP_HUM	transmembrane region	TMHMM	9	31	
6	sp_P23515_OMGP_HUM	disordered region	IUPred	335	346	
7	sp_P23515_OMGP_HUM	disordered region	IUPred	364	410	
8	sp_P23515_OMGP_HUM	disordered region	BISMMpredicto	29	48	
9	sp_P23515_OMGP_HUM	disordered region	BISMMpredicto	333	348	
10	sp_P23515_OMGP_HUM	disordered region	BISMMpredicto	362	379	
11	sp_P23515_OMGP_HUM	functional domain	PFAM	25	53	PF01462.19
12	sp_P23515_OMGP_HUM	functional domain	PFAM	56	94	PF12799.8
13	sp_P23515_OMGP_HUM	functional domain	PFAM	124	161	PF12799.8
14	sp_P23515_OMGP_HUM	functional domain	PFAM	168	226	PF13855.7
15	sp_P23515_OMGP_HUM	consensus ordered region	TAPASS	1	332	
16	sp_P23515_OMGP_HUM	consensus disordered region	TAPASS	333	440	
17	sp_P23515_OMGP_HUM	amyloidogenic region	ArchCandy2	39	64	
18	sp_P23515_OMGP_HUM	amyloidogenic region	ArchCandy2	121	145	
19	sp_P23515_OMGP_HUM	amyloidogenic region	ArchCandy2	175	196	
20	sp_P23515_OMGP_HUM	amyloidogenic region	ArchCandy2	218	243	
21	sp_P23515_OMGP_HUM	amyloidogenic region	ArchCandy2	273	294	
22	sp_P23515_OMGP_HUM	amyloidogenic region	ArchCandy2	347	372	
23	sp_P23515_OMGP_HUM	amyloidogenic region	ArchCandy2	418	440	
24	sp_P23515_OMGP_HUM	exposed amyloidogenic region	ArchCandy2	347	372	
25	sp_P23515_OMGP_HUM	exposed amyloidogenic region	ArchCandy2	418	440	
26	sp_P23515_OMGP_HUM	amyloidogenic region	Pasta	169	206	
27	sp_P23515_OMGP_HUM	amyloidogenic region	Pasta	417	439	
28	sp_P23515_OMGP_HUM	exposed amyloidogenic region	Pasta	417	439	
29	sp_P23515_OMGP_HUM	amyloidogenic region	Tango	235	239	
30	sp_P23515_OMGP_HUM	amyloidogenic region	Tango	273	281	
31	sp_P23515_OMGP_HUM	amyloidogenic region	Tango	292	295	
32	sp_P23515_OMGP_HUM	amyloidogenic region	Tango	424	439	
33	sp_P23515_OMGP_HUM	exposed amyloidogenic region	Tango	424	439	
34	sp_P23515_OMGP_HUM	eukaryotic SLiMs	ELM	408	414	ELME000155
35	sp_P23515_OMGP_HUM	eukaryotic SLiMs	ELM	371	377	ELME000159
36	sp_P23515_OMGP_HUM	eukaryotic SLiMs	ELM	410	416	ELME000159
37	sp_P23515_OMGP_HUM	eukaryotic SLiMs	ELM	336	342	ELME000220
38	sp_P23515_OMGP_HUM	eukaryotic SLiMs	ELM	394	398	ELME000239
39	sp_P23515_OMGP_HUM	eukaryotic SLiMs	ELM	414	418	ELME000239
40	sp_P23515_OMGP_HUM	eukaryotic SLiMs	ELM	411	417	ELME000289

3 Tools available

- **ArchCandy2** : Amyloidogenic region predictor, an updated version of ArchCandy (Ahmed et al., 2015)

- **Pasta 2.0** : Amyloidogenic region predictor (Walsh et al., 2014)

- **Tango** : Amyloidogenic region predictor (Fernandez-Escamilla et al., 2004)

- **IUPred** : Intrinsically disorder predictor (Dosztányi et al., 2005)

- **BiSSM** : Intrinsically disorder predictor

- **SignalP** : Signal peptide predictor (Petersen et al., 2011)

- **TMHMM** : Transmembrane region predictor (Krogh et al., 2001)

- **CATH** : Structural domains predictor associated with HMMER (Dawson et al., 2017; Eddy, 2011)

- **PFAM** : Protein family predictor (El-Gebali et al., 2018)

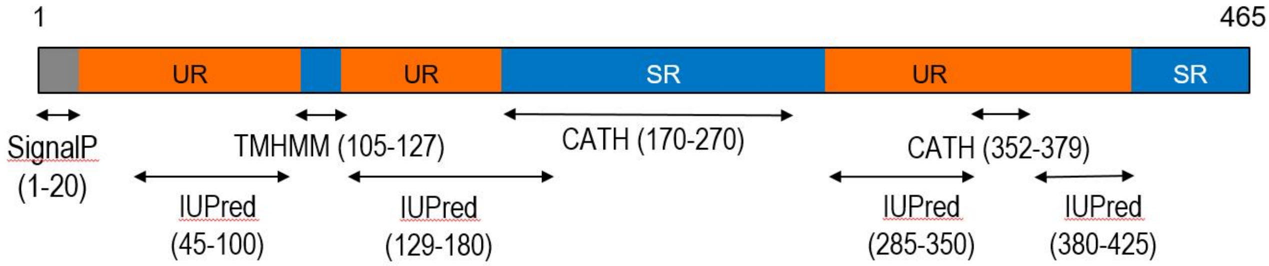
- **SLiMs** : Eukaryote short linear motifs predictor (Kumar et al., 2020; Ruhanen et al., 2014)

- **Meta Repeat Finder** : Predictor of repeats in sequences

4 Consensus IDRs, AR and EAR

4.1 Consensus IDRs

The pipeline assigns each residue of the analysed protein as belonging to a structured or an unstructured region. If both BISMM and/or IUPred predict a structured state at a given region, it is mapped as structured. If a structured region predicted by CATH or TMHMM overlaps with IDR prediction, this region is considered as structured. At the same time, structured regions of less than 30 residues are considered as unstructured. An exception is made for TMHMM prediction of transmembrane regions, which being shorter than 30 residues, are still considered structured. Consensus IDRs of less than 20 residues are considered as structured. N-terminal regions predicted as signal peptides are excluded from our analysis. Proteins shorter than 30 residues were predicted to be unstructured with exception of ones containing a transmembrane helix.



4.2 Amyloidogenic regions (ARs)

The results of the three amyloid predictors, ArchCandy2, TANGO and PASTA 2.0, were treated separately. Each predictor distinguishes between two types of regions : amyloidogenic regions (ARs) and non-amyloidogenic regions, with scores over and below the threshold, respectively. This binary outcome ignores both the exact values of the scores over the threshold and the number of the amyloidogenic hits at a given residue.

4.3 Exposed amyloidogenic regions (EARs)

EARs were defined in a similar way as with ARs, with the additional criteria that individual hits of amyloidogenic predictors should overlap with at least 80 % of an IDR.