MobiDB: a comprehensive database of intrinsic protein disorder annotations

Tomás Di Domenico
University of Padua, Italy

@

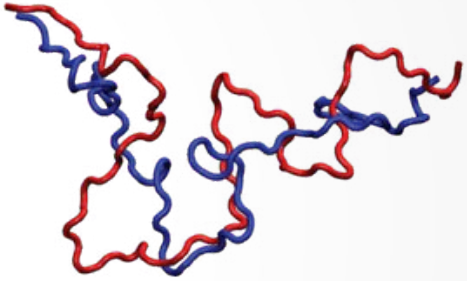
ISMB/ECCB 2013, Berlin
21/07/2013

Non-globular proteins

Non-globular proteins (NGPs) do not have the typical globular shape. They tend to be elongated.

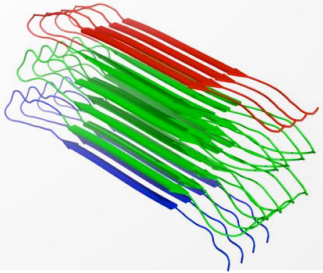
NGPs can be classified into three types:

I - Disordered proteins



II - Repeat proteins

III - Aggregating proteins



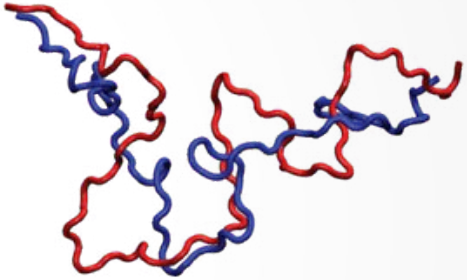
Non-globular proteins

Non-globular proteins (NGPs) do not have the typical globular shape. They tend to be elongated.

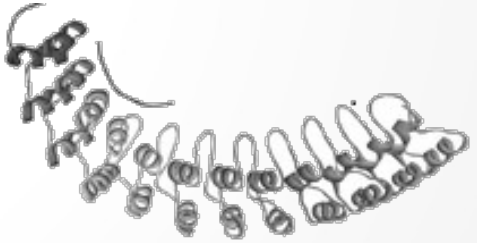
NGPs can be classified into three types:

We'll talk about these guys

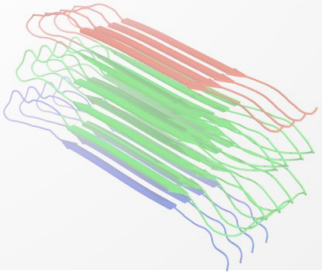
I - Disordered proteins



II - Repeat proteins



III - Aggregating proteins



The many faces of IDPs

So, let's talk about intrinsically disordered proteins.

Or should we rather say...

- natively denatured? (Schweers et al., 1994)
- natively unfolded? (Weinreb et al., 1996)
- intrinsically unfolded? (Baskakov et al., 1999)
- intrinsically unstructured? (Wright and Dyson, 1999)
- intrinsically disordered? (Dunker et al., 2000)
- **exceptionally flexible?!** (Ahmed et al., 2007)
- natively unstructured? (Schlessinger et al., 2007)
- naturally flexible? (Uversky et al., 2009)

The many faces of IDPs

So, let's talk about intrinsically disordered proteins.

Or should we rather say...

- natively denatured? (Schweers et al., 1994)
- natively unfolded? (Weinreb et al., 1996)
- intrinsically unfolded? (Baskakov et al., 1999)
- intrinsically unstructured? (Wright and Dyson, 1999)
- intrinsically disordered? (Dunker et al., 2000)
- exceptionally flexible? (Ahmed et al., 2007)
- natively unstructured? (Schlessinger et al., 2007)
- naturally flexible? (Uversky et al., 2009)

The many faces of IDPs

Independently of the chosen name, we'll define **IDPs** as those proteins that **do *not* fold into a fixed three-dimensional structure** under **physiological conditions**.

Source of IDP annotations

We can divide the sources of readily available IDP annotations in three groups:

- Manually curated
- Indirect
- Predicted

Manually curated: DisProt

- Entries manually extracted from literature
- 694 annotated proteins (v6.02)

- Slow updates
  - Few structures on each release
  - Hard to map to a UniProt entry
- 'Unstructured' complementary annotations

Indirect: PDB structures (X-Ray)

Arguably the most commonly used definition of disorder:

Residues not observed in an X-Ray experiment *may* be an indication of a disordered region.

Indirect: PDB structures (X-Ray)

Some issues

- A 'structured' annotation is much more authoritative than a 'disordered' annotation
- IDPs are famous for their ability to change, and we're looking at a still picture
- Only short disordered regions will be obtained from a crystallizable structure

Indirect: PDB structures (NMR)

A definition for NMR ~~mobility~~ disorder:

Mobile regions in an NMR experiment

may be an indication of a disordered region.

Indirect: PDB structures (NMR)

Some issues:

- NMR is limited to rather small molecules
- The models in an NMR ensemble are not necessarily different conformations of the protein

Predicted

A full rainbow of disorder predictors have been developed in the last few years.

They usually require only a sequence as input, which conveniently allows us to obtain disorder annotations for any protein we like.

Predicted

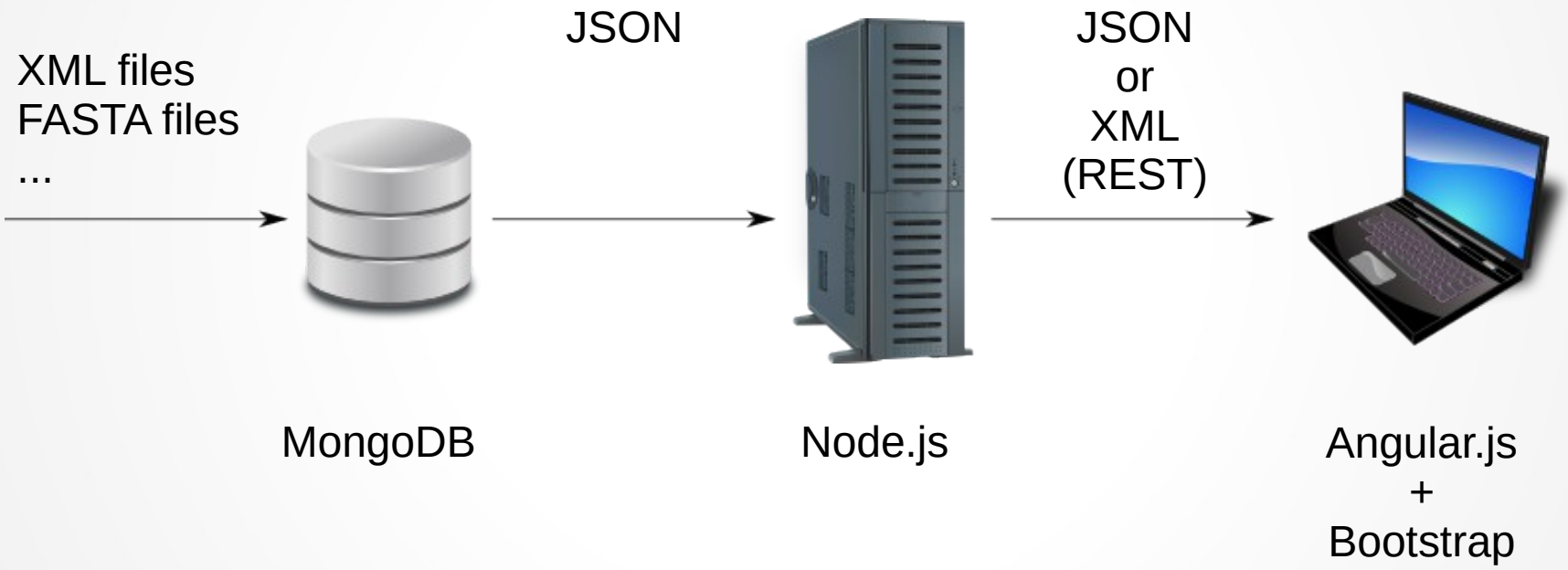
Some issues:

- 'All of the above'
- Again, the 'positive' cases are not exactly the opposite of the 'negative' cases.

MobiDB's motivation

- Provide extensive disorder annotations for all UniProt entries (~30 million)
- Take advantage of UniProt annotations that may be relevant to disorder
- Provide end users with a good UI and UX
- Give advanced users solid programmatic access through web services
- Easily accomodate new annotations

MobiDB's architecture



Quick development. Easily maintainable.

MobiDB's disorder data sources

- DisProt: XML distribution
- PDB: SIFTS database XML files
- Predictions: all the fast ones we can get
  - ESpritz x3
  - DisEMBL x2
  - IUPred x2
  - (GlobPlot, RONN, VSL2b, ...)

Map them all to a UniProt entry.

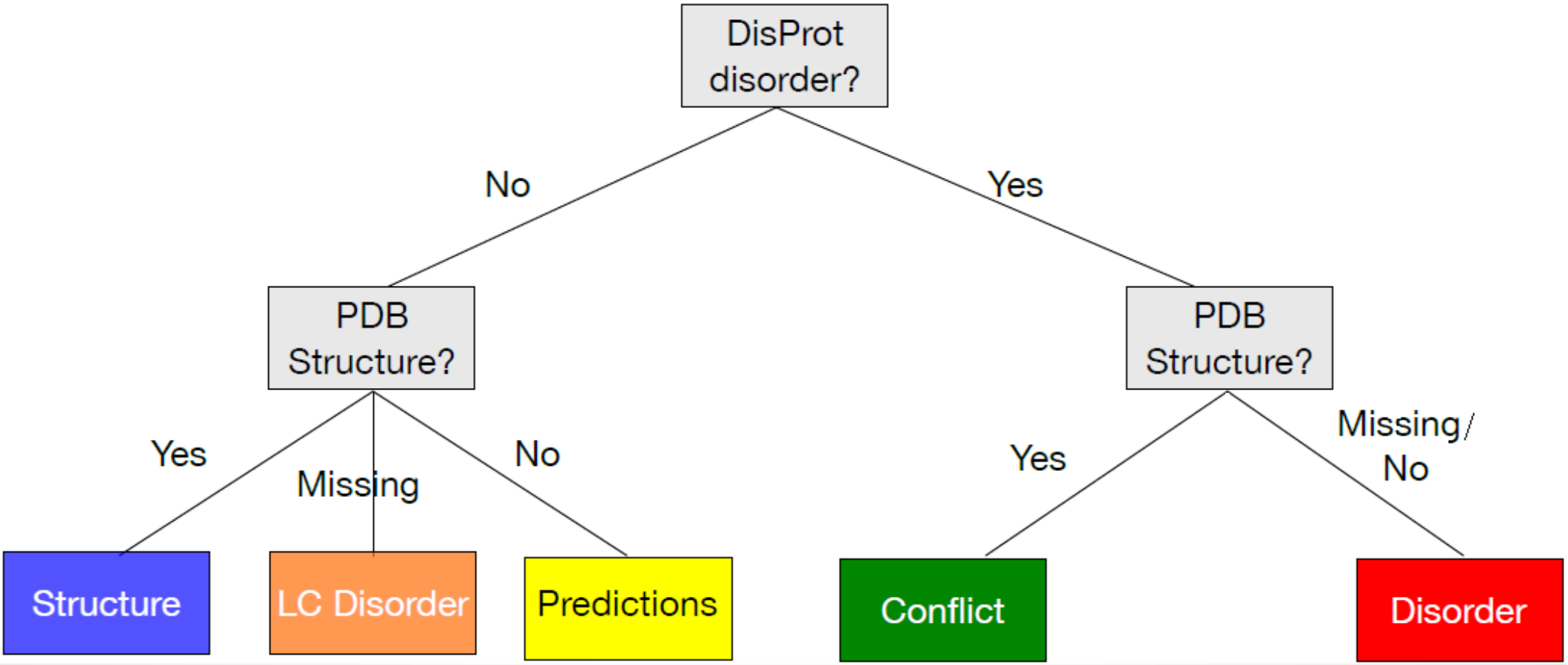
MobiDB's consensus

MobiDB provides an overview of disorder annotations for a protein by calculating a series of consensus:

- DisProt consensus (simple 'flattening')
- PDB consensus (structure wins)
- Predictors consensus (majority vote)
- **Full consensus**

MobiDB's consensus

MobiDB combines all available annotations into a single consensus annotation.



The MobiDB user interface

MobiDB

a database of protein disorder and mobility annotations

The MobiDB user interface

MobiDB

a database of protein disorder and mobility annotations

Region details [-]

Location	Length	Annotation	Atchley-plot
1-56	56	conflict	
57-58	2	disordered	
59-73	15	conflict	
74-165	92	structured	
166-167	2	conflict	
168-182	15	structured	
183-188	6	conflict	
189-223	35	structured	
224-227	4	conflict	
228-290	63	structured	
291-293	3	conflict	
294-312	19	disordered	
313-318	6	disordered	
319-323	5	conflict	
324-356	33	structured	
357-360	4	conflict	
361-390	30	structured	
391-391	1	conflict	
392-393	2	disordered	

Details for the selected region

Location

1-56

Consensus disorder annotation

conflict

Average Atchley values

[0.094,0.055,-1.113,0.232,-1.095]

Atchley values plot

PTMs

[9]: Phosphorylation at Ser-9 by HIPK4 increases repression activity on BIRC5 promoter
[15]: Phosphorylated on Ser-15 upon ultraviolet irradiation; which is enhanced by interaction with BANP
[15]: It is unclear whether AMP directly mediates phosphorylation at Ser-15
[18]: Phosphorylated on Thr-18 by VRK1
[18]: Phosphorylated on Thr-18 by isoform 1 and isoform 2 of VRK2
[18]: Phosphorylation on Thr-18 by isoform 2 of VRK2 results in a reduction in ubiquitination by MDM2 and an in
[20]: Phosphorylated on Ser-20 by CHEK2 in response to DNA damage, which prevents ubiquitination by MDM2
[20]: Phosphorylated on Ser-20 by PLK3 in response to reactive oxygen species (ROS), promoting p53/TP53-m
[33]: Phosphorylated on Ser-33 by CDK7 in a CAK complex in response to DNA damage
[46]: Phosphorylated on Ser-46 by HIPK2 upon UV irradiation
[46]: Phosphorylation on Ser-46 is required for acetylation by CREBBP
[46]: Phosphorylated by DYRK2 at Ser-46 in response to genotoxic stress
[55]: Phosphorylated on Thr-55 by TAF1, which promotes MDM2-mediated degradation
[55]: Dephosphorylated by PP2A-PPP2R5C holoenzyme at Thr-55

Motifs

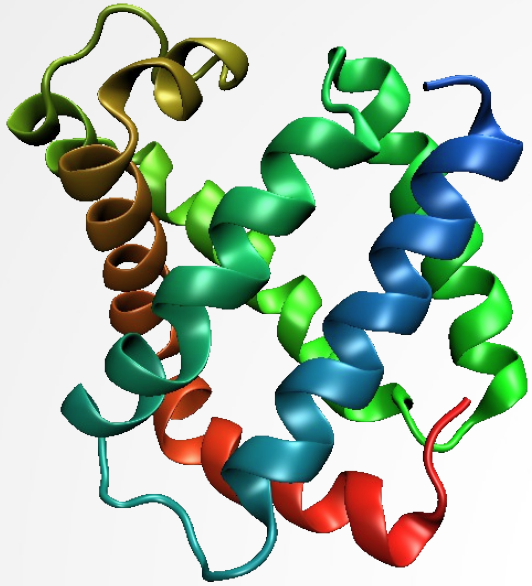
[17-25]: TADI
[48-56]: TADII

Regions

A few examples

Let's take a look at what a random user could find when looking for disorder annotations for his/her protein of interest

A few examples: Myoglobin

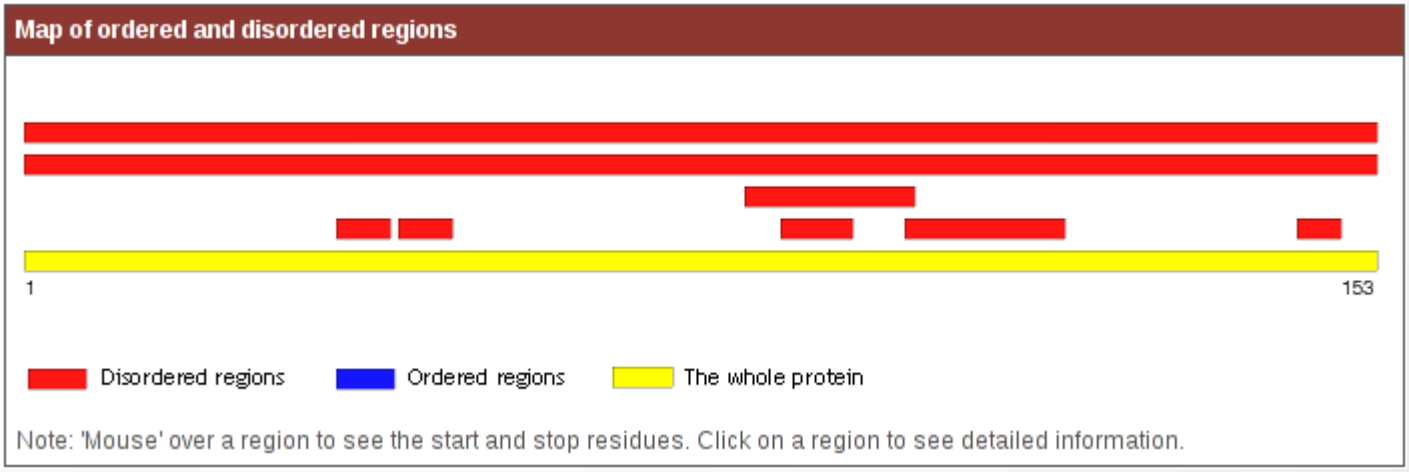


https://en.wikipedia.org/wiki/File:Myoglobin-1mba.png

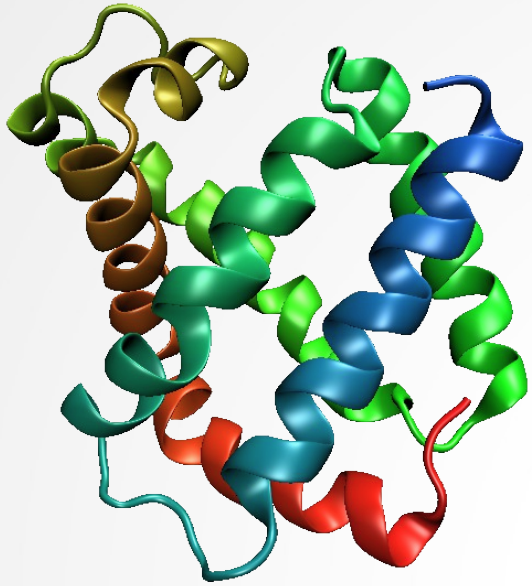
The first structure to be crystalized, it has more than a hundred entries on the PDB

...but...

...a quick search on the **DisProt** database tells us it's fully disordered.

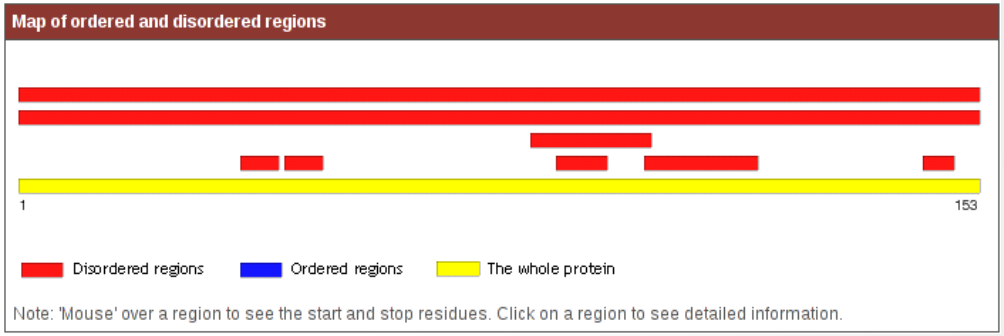


A few examples: Myoglobin



https://en.wikipedia.org/wiki/File:Myoglobin-1mba.png

vs.



So we've found that in the case of Myoglobin, there is a clear **conflict between** the information we find on the **PDB**, and the information we find on **DisProt**.

MobiDB:

Conflict reason: high pressure experiment concludes that Myoglobin has an unstructured state (like all proteins??)

A few examples: 4E-BP1

Eukaryotic translation initiation factor 4E-binding protein 1

Fletcher and Wagner, 98

'...4E-BPI has no regions of local order in the absence of eIF4E. […] appears to be an induced fit to a completely disordered protein molecule.'

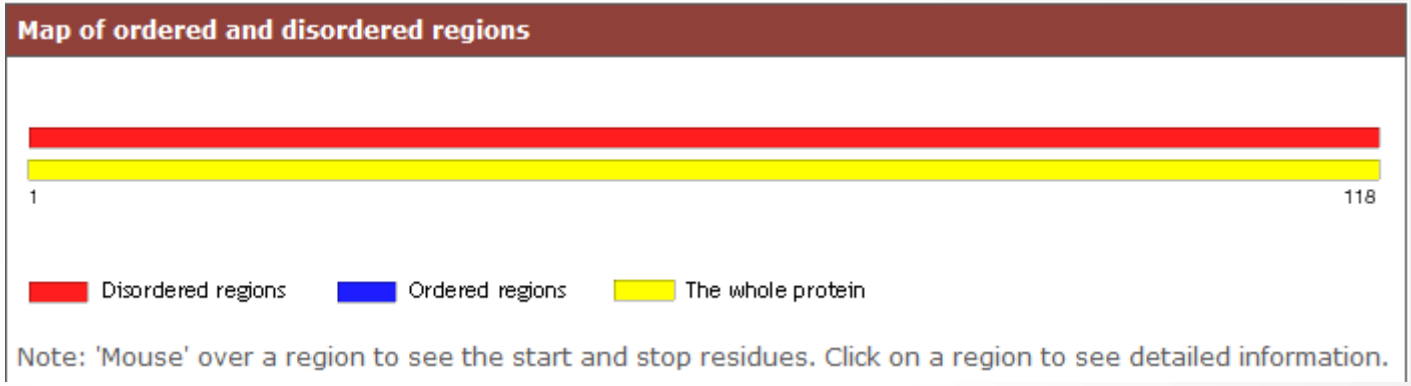
'NMR studies of 4E-BPI [...] have shown that these proteins have little or no folded structure under physiological conditions...'

'...appears to be mediated by a short central region of the 4E-BPS (within residues 49-68 of 4E-BP1) with the rest of the protein remaining unfolded in the bound state..'

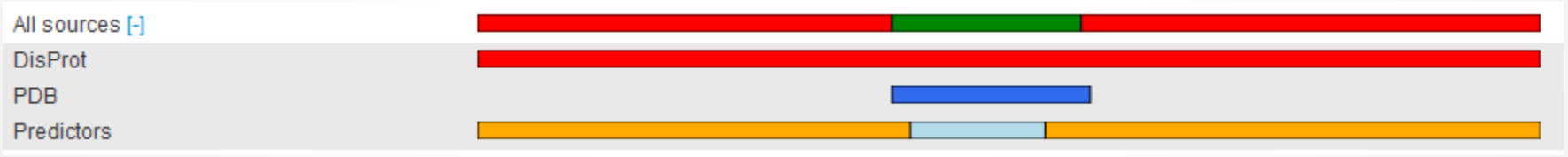
A few examples: 4E-BP1

Eukaryotic translation initiation factor 4E-binding protein 1

DisProt:



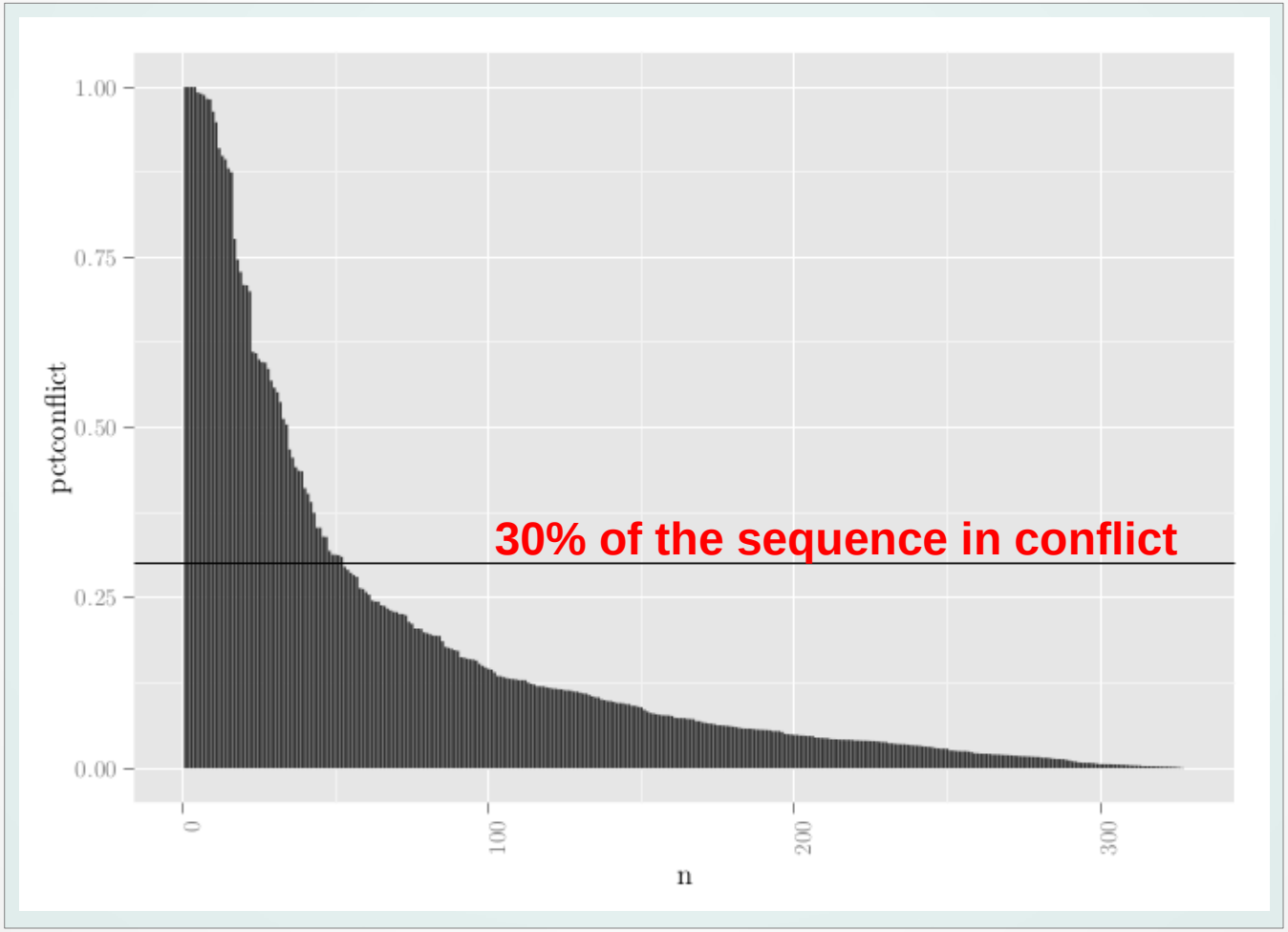
MobiDB:



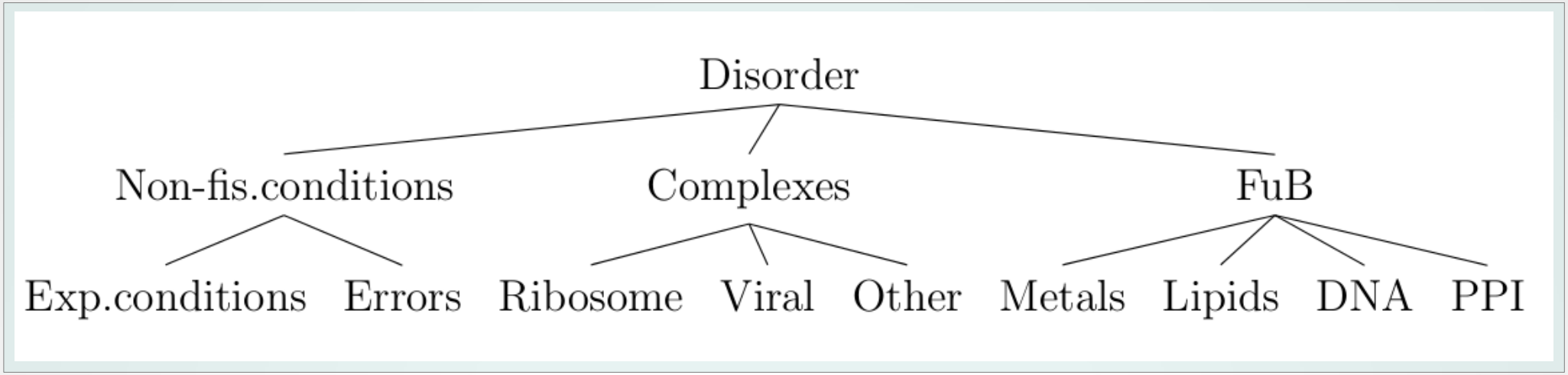
By adding other sources, we obtain a better representation of what was originally described on the publication

Getting results: conflicting regions

- 300 entries where at least five consecutive residues have conflicting disorder annotations.
- 52 entries have 30% or more of their sequence in conflict



Getting results: conflicting regions



- We were able to obtain this classification by manually reviewing the 52 proteins with more than 30% conflict.
- We are currently **improving and expanding this classification**.
- Ultimately, we hope to **better understand** what we believe is a group of phenomena, currently collectively identified as **Intrinsic Protein disorder**.
- We hope to expand the availability and relevance of this resource by integrating it into the italian node of the **ELIXIR project**

Conclusions

- By aggregating different sources of information, MobiDB provides a rich context of disorder annotations
- This context better represents the complexity of intrinsic protein disorder
- New annotations can be integrated seamlessly

Thank you!

MobiDB
a database of protein disorder and mobility annotations

http://beta.mobidb.bio.unipd.it

Visit me at poster **L070**!

Tomás Di Domenico <tdido@bio.unipd.it>

People

BioComputingUP lab, UniPD

- Prof. Silvio Tosatto
- Dr. Ian Walsh
- Dr. Giovanni Minervini
- Dr. Awais Ihsan