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 (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 (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
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)
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)
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.
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
A definition for NMR mobility disorder:

Mobile regions in an NMR experiment \textit{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
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.
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 accommodate new annotations
MobiDB's architecture

XML files
FASTA files
...

MongoDB

JSON

Node.js

JSON or XML (REST)

Angular.js + Bootstrap

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

### Disorder consensus

#### Disorder consensus [+]

---

### Detailed disorder annotations

#### DisProt [+]

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>DP00028</th>
<th>consensus [+]</th>
</tr>
</thead>
</table>

#### PDB-XRay [+]

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Resolution (Å)</th>
<th>Chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1wkx</td>
<td>2.1</td>
<td>B</td>
</tr>
<tr>
<td>2jgb</td>
<td>1.7</td>
<td>B</td>
</tr>
<tr>
<td>2gc</td>
<td>2.4</td>
<td>B</td>
</tr>
<tr>
<td>2v8w</td>
<td>2.3</td>
<td>B</td>
</tr>
<tr>
<td>2v8x</td>
<td>2.3</td>
<td>consensus [+]</td>
</tr>
<tr>
<td>2v8y</td>
<td>2.1</td>
<td>consensus [+]</td>
</tr>
<tr>
<td>3hkg</td>
<td>2.1</td>
<td>C</td>
</tr>
<tr>
<td>3hxi</td>
<td>1.8</td>
<td>C</td>
</tr>
<tr>
<td>3m93</td>
<td>2.9</td>
<td>C</td>
</tr>
<tr>
<td>3m94</td>
<td>2.05</td>
<td>C</td>
</tr>
<tr>
<td>3u7x</td>
<td>2.1</td>
<td>consensus [+]</td>
</tr>
</tbody>
</table>

#### Predictions [+]

dise-465
dise-HL
espritz-cl
espritz-n
espritz-x
iupred-f
iupred-s
# The MobiDB user interface

MobiDB

a database of protein disorder and mobility annotations

## Region details

<table>
<thead>
<tr>
<th>Location</th>
<th>Length</th>
<th>Annotation</th>
<th>Aitchley-plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-56</td>
<td>56</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>57-58</td>
<td>2</td>
<td>disordered</td>
<td></td>
</tr>
<tr>
<td>59-73</td>
<td>15</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>74-165</td>
<td>92</td>
<td>structured</td>
<td></td>
</tr>
<tr>
<td>166-167</td>
<td>2</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>186-182</td>
<td>15</td>
<td>structured</td>
<td></td>
</tr>
<tr>
<td>183-188</td>
<td>6</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>189-223</td>
<td>35</td>
<td>structured</td>
<td></td>
</tr>
<tr>
<td>224-227</td>
<td>4</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>228-290</td>
<td>63</td>
<td>structured</td>
<td></td>
</tr>
<tr>
<td>291-293</td>
<td>3</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>294-312</td>
<td>10</td>
<td>disordered</td>
<td></td>
</tr>
<tr>
<td>313-318</td>
<td>6</td>
<td>disordered</td>
<td></td>
</tr>
<tr>
<td>319-323</td>
<td>5</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>324-356</td>
<td>33</td>
<td>structured</td>
<td></td>
</tr>
<tr>
<td>357-360</td>
<td>4</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>361-390</td>
<td>30</td>
<td>structured</td>
<td></td>
</tr>
<tr>
<td>391-391</td>
<td>1</td>
<td>conflict</td>
<td></td>
</tr>
<tr>
<td>392-393</td>
<td>2</td>
<td>disordered</td>
<td></td>
</tr>
</tbody>
</table>

## Details for the selected region

**Location**

1-56

**Consensus disorder annotation**

conflict

**Average Aitchley values**

[0.094, 0.055, 1.113, 0.232, 1.095]

**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
- [19]: Phosphorylated on Thr-18 by VRK1
- [16]: Phosphorylated on Thr-18 by isoform 1 and isoform 2 of VRK2
- [18]: Phosphorylation on Thr-18 by isoform 1 of VRK2 results in a reduction in ubiquitination by MDM2 and in increased stability in vivo
- [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-mediated apoptosis
- [33]: Phosphorylated on Ser-33 by CDK7 in a CAK complex in response to DNA damage
- [46]: Phosphorylated on Ser-45 by HIPK2 upon UV irradiation
- [40]: Phosphorylation on Ser-46 is required for acetylation by CREBEP
- [46]: Phosphorylated by DYRK2 at Ser-45 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]: TADI

**Regions**

...
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

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

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**.

**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
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!

People

BioComputingUP lab, UniPD

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

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