PART II.
Prediction of functional regions within disordered proteins

Zsuzsanna Dosztányi
MTA-ELTE Momentum Bioinformatics Group
Department of Biochemistry Eotvos Lorand University, Budapest, Hungary
dosztanyi@caesar.elte.hu
Large-scale analysis of IDPs

Possible through prediction methods
- Functional properties
- Evolutionary properties

Disorder characterization
- Percentage of properties with long (>30 or >40 aa)
- Percentage of disordered residues
How common is protein disorder?

- Around 50% of human proteins have long disordered regions
- Around 30% of residues in the human proteome are predicted as disordered
- Disorder content increases with evolutionary complexity
Protein disorder is prevalent
Protein disorder complements the functional repertoire of globular proteins

Table 2. Correlation and anticorrelation of structural disorder with Swiss-Prot functional categories

<table>
<thead>
<tr>
<th>Top functions that correlate with long disorder&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Top functions that anticorrelate with long disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differentiation</td>
<td>GMP biosynthesis</td>
</tr>
<tr>
<td>Transcription</td>
<td>Amino acid biosynthesis</td>
</tr>
<tr>
<td>Transcription regulation</td>
<td>Transport</td>
</tr>
<tr>
<td>Spermatogenesis</td>
<td>Electron transport</td>
</tr>
<tr>
<td>DNA condensation</td>
<td>Lipid A biosynthesis</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>Aromatic hydrocarbons catabolism</td>
</tr>
<tr>
<td>mRNA processing</td>
<td>Glycolysis</td>
</tr>
<tr>
<td>mRNA splicing</td>
<td>Purine biosynthesis</td>
</tr>
<tr>
<td>Mitosis</td>
<td>Pyrimidine biosynthesis</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>Carbohydrate metabolism</td>
</tr>
<tr>
<td>Protein transport</td>
<td>Branched-chain amino acid biosynthesis</td>
</tr>
<tr>
<td>Meiosis</td>
<td>Lipopolysaccharide biosynthesis</td>
</tr>
</tbody>
</table>

Xie H et al. J Proteome Res. 2007, 6, 1882
Functions of intrinsically disordered proteins

I  Entropic chains
II Linkers
III Molecular recognition
IV Protein modifications (e.g. phosphorylation)
V Assembly of large multiprotein complexes
Interaction of IDPs

Coupled folding and binding
Coupled folding and binding

Functional advantages

- Weak transient, yet specific interactions
- Post-translational modifications
- Flexible binding regions that can overlap
- Evolutionary plasticity

Signaling Regulation
Various complexes of IDPs

- Can be grouped according the adopted secondary structure elements
  - alpha helical
  - beta strand
  - polyproline
  - irregular
Small interfaces

SH3 domain
Large interfaces

- Cyclin-dependent kinase (Cdk) inhibitor, p27Kip1 (p27)
- Binds to cdk-cyclin komplex and inhibits their activity
- Fully disordered protein
Conformational plasticity

C-terminal transactivation domain (CTAD) of the hypoxia inducible factor-1α

Berlow et al. FEBS Lett. 2015;589:2433
Fine tuning the entropic component

\[ \Delta G = \Delta H - T \Delta S \]

\[ \Delta G < 0: \Delta H < 0 \quad \text{Minimising entropy loss upon binding} \]
\[ \Delta G < 0: \Delta S \approx 0 \]

Maximising enthalpy of binding

Maximising enthalpy gain upon binding (\(\Delta H_{\text{binding}} \ll 0\))

\[ \uparrow \text{interface area} \]

Minimising entropy loss upon binding

Minimising change in conformational entropy (\(\Delta S_{\text{conform}}\))

\[ \uparrow S_{\text{conform}} \]

Providing additional configurational entropy (\(\Delta S_{\text{config}}\)) via repeated motifs in multivalent IDR

Extensive interface with secondary structure

Motif interface without secondary structure

Intrinsically disordered region-binding domain

Intrinsically disordered region

Flock et al Curr Opin Struct Biol. 2014; 26:62
Phase transitions

- Multivalency and weak interactions
- Regulated by phosphorylation
- Transition from small complexes and large, dynamic supramolecular polymers.

Disordered binding regions

- Complexes of IDPs in the PDB: ~ 200
- Known instances: ~ 2 000
- Estimated number of such interactions in the human proteome: ~ 1 000 000

- Experimental characterization is very difficult
- Computational methods
Disordered protein complexes

- Interaction sites are usually *linear* (consist of only 1 part)
- Enrichment of interaction prone amino acids

Complex between p53 and MDM2

Sequence

No need for structure, binding sites can be predicted from sequence alone

Binding sites
Prediction of disordered binding regions – ANCHOR

- What discriminates disordered binding regions?
  - A cannot form enough favorable interactions with their sequential environment
  - It is favorable for them to interact with a globular protein

- Based on simplified physical model
  - Based on an energy estimation method using statistical potentials
  - Captures sequential context
nucleoprotein from Nipah virus (DP00697)
Machine learning approaches

MORFchibi

- Uses two SVMs
  - One recognizes the different amino acid composition of flanking regions compared to the binding region
  - One recognizes the similarity to known binding regions
- trained on short chains in complex

Malhis N, Gsponer J. Bioinformatics. 2015; 31:1738
Machine learning approaches

DISOPRED3

- Uses three SVMs
  - Simple sequence profile
  - PSI-Blast profiles (very slow)
  - PSI-Blast profiles with global features
- trained on short chains in complex

Jones DT, Cozzetto D. Bioinformatics. 2015; 31:857
Amount of disordered binding regions

- What is the amount of disordered protein regions in the human proteome?
  - ANCHOR: 93429
  - MORFchibi: 275013
  - DISOPRED3: 63848

- We cannot tell what is the *false positive rate* of these methods.
The functionality of a protein segment is often approached by investigating the evolutionary history of its primary sequence.

Can this approach used for disordered proteins?

Sometimes …
Constrained and flexible disorder

- ‘constrained’,
  - if both features (amino acid sequence and the property of disorder) are conserved
- ‘flexible’,
  - if only disorder is conserved
- ‘non-conserved’ positions
  - where disorder is not conserved

**DISCONS** novel server (preliminary)

Varadi et al. BMC Bioinformatics. 2015;16:153
Conservation patterns of linear motifs

- No evolutionary constraints to keep the structure
- Strong constraints on functional site

SlimPrints

- Generates sequence alignments of orthologous sequences
- Relative conservation score per position
- Filters out less reliable regions
- Fails if sequences are too divergent, or too similar
The next challenge:

- Characterizing the ensemble of conformations of IDPs
- And their relationship with function
Ensemble characterization for IDPs

- Experimental methods cannot detect a single conformation, only time or ensemble averages

- Combination of methods are needed (NMR, SAXS)

- Methods are used to characterize
  - Radius of gyration
  - Transient secondary structure elements
  - Transient long range contacts
### PED database

**Rg distribution of Ensemble #1**

<table>
<thead>
<tr>
<th>Ensemble index</th>
<th># of conformers</th>
<th>Rg mean</th>
<th>Dmax mean</th>
<th>Show/Hide</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>130 conformers</td>
<td>20.81±0.94 Å</td>
<td>70.89±4.49 Å</td>
<td>Show List</td>
</tr>
</tbody>
</table>

**Select a conformer from the list below to display it**

<table>
<thead>
<tr>
<th>Conformer</th>
<th>Rg</th>
<th>Dmax</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conformation 1</td>
<td>Rg: 20.19</td>
<td>Dmax: 65.17</td>
<td>View conformer - Up</td>
</tr>
<tr>
<td>Conformation 2</td>
<td>Rg: 18.99</td>
<td>Dmax: 60.44</td>
<td>View conformer - Up</td>
</tr>
<tr>
<td>Conformation 3</td>
<td>Rg: 20.16</td>
<td>Dmax: 67.9</td>
<td>View conformer - Up</td>
</tr>
</tbody>
</table>