Intrinsic disorder proteins

Peter Tompa
Institute of Enzymology
Hungarian Academy of Sciences
Budapest, Hungary

Keith Dunker
CCBB & IUSM
Indiana University
Indianapolis, USA
Why do we study disorder proteins?

1) Gap in knowledge (684 (803) vs. thousands of IUPs)
2) Structural genomics initiatives
3) Bioinformatics studies
4) Single protein studies
Analysis of Signaling Interactions

• Examined each interaction on Pawson’s website.

• Almost all of the interactions involved ordered regions binding to disordered partners.

• Conclusion: if Pawson’s examples are typical, then a very significant proportion of protein-protein signaling interactions use disordered regions.
Parallel Paradigms

**Catalysis**

AA seq → 3-D Structure → Function

**Signaling**

AA seq → Disordered Ensemble → Function
## Disorder and Function

<table>
<thead>
<tr>
<th>Category</th>
<th>Change</th>
<th>Example(s)</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Recognition</td>
<td>D → O</td>
<td>113</td>
<td>Inter- and Intra-protein, ssDNA, dsDNA, tRNA, rRNA, mRNA, nRNA, bilayers, ligands, cofactors, metals</td>
</tr>
<tr>
<td>Protein Modification</td>
<td>Variable</td>
<td>36</td>
<td>Acetylation, fatty acylation, glycosylation, methylation, phosphorylation, ADP-ribosylation, ubiquitination, proteolytic digestion</td>
</tr>
<tr>
<td>Entropic Chains</td>
<td>Variable</td>
<td>17</td>
<td>Linkers, spacers, bristles, clocks, springs, detergents, self-transport</td>
</tr>
</tbody>
</table>

Basic approaches to predict disorder

1) Machine learning
2) Structural approach
The unusual AA composition of IDPs

Dunker et al. (2001) J. Mol. Graph. Model. 19, 26
AA feature space: AAindex database
http://www.genome.jp/aaindex

A number is associated with every amino acid, which quantitatively describes how characteristic the given feature is to the AA (has 566 different scales at present)
Based on AA compositions, two things you might not want to do...

1) low complexity regions
2) regions w/o secondary structure
Low complexity regions

Sequence databases contain a lot of regions in which only a few amino acids occur (simple), or some dominate (biased), this can be described by an entropy function - Wootton (1993)

Shannon's entropy

\[
K_2 = - \sum_{i=1}^{N} \frac{n_i}{L} \left( \log_2 \frac{n_i}{L} \right)
\]

$L > 20$; window size,
$N$ alphabet size (20),
$n_i$: az i. aminosav száma
...may correspond to disorder

<table>
<thead>
<tr>
<th>SRY_MOUSE</th>
<th>TRANSCRIPTIONAL ACTIVATOR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1-143</strong></td>
<td>MEGHVKRPMNAFMVWSRGERHKLAQQNP</td>
</tr>
<tr>
<td></td>
<td>QNTekiSKQLGCWRKSLTEAEKRPFFQEAQR</td>
</tr>
<tr>
<td></td>
<td>LKILHREKYPNYKQPHRRAKVSPQRSGLQ</td>
</tr>
<tr>
<td></td>
<td>PAVASTKLYNLLQWDRNPHAIETYRDWSRA</td>
</tr>
<tr>
<td></td>
<td>AHLYSKNNQSFYWPVDIPTGHL</td>
</tr>
</tbody>
</table>

| **144-366** | qqqqqqqqqqqfhhhhqgqqqqgqfdhhqgqgqqfdhhqgqgqqfhhhhqgqgqqfhh | |
|            | dhhhhheeegqfhdhhqgqgqfhdhhqgqgqfhhhhqgqgqfhdhhqgqgqfhh | |
|            | qqqqqqqqqqqfhhhhqgqgqfhhddhhhhqgqgqfhhddhhhhqgqgqfhhdd | |
|            | hqqqqqqqqqqfhdhhqgqgqfhdhhqgqgqfhdhhqgqgqfhdhhqgqgqfhdh | |
|            | hqqqqqqqqqqfhdhhhqqqqgqfhdhhqgqgqfhdhhqgqgqfhdhhqgqgqfhdh | |
|            | fhdhhqgqgqfhdhhqgqgqfhdhhqgqgqfhdhhqgqgqfhdhhqgqgqfhdhh | |
|            | qqqqqqqqqqqfhhhhqgqgqfhh | |

| **367-395** | LTYLLTADITGEHPTYQEHLS | STALWLAQVS |
The relationship of low complexity and disorder
Based on AA compositions, two things you might not want to do...

1) low complexity regions
2) regions w/o secondary structure
...but there are globular proteins without Secondary structure and IDPs with secondary structure

Loopy Proteins Appear Conserved in Evolution

Jinfeng Liu\textsuperscript{1,2,3}, Hepar Tan\textsuperscript{2} and Burkhard Rost\textsuperscript{2,3,4*}

Radhakrishnan (1997) \textit{Cell} 91, 741

1) Machine learning
Artificial neural network (NN)
Artificial neural networks

Basic unit: a neuron

\[ \sigma \left( \sum_{j=1}^{n} w_j x_j + b_j \right) \]

Hidden layer

\[ \hat{y}(\Theta) = g(\Theta, x) = \sum_{i=1}^{n_b} \sigma \left( \sum_{j=1}^{n} w_{i,j}^1 x_j + b_{i,j}^1 + b_i^2 \right) + b^2 \]
Predictor of naturally disordered regions (PONDR®)

Input

- 18 amino acids
- Hydrophobicity
- Sequence complexity

VL2, VL3
VLXT
VSL2
Dunker, 1998

Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization

Ethan Garner 1        Paul Cannon 2        Pedro Romero 2

THOUSANDS OF PROTEINS LIKELY TO HAVE LONG DISORDERED REGIONS

PEDRO ROMERO, ZORAN OBRADOVIC
School of Electrical Engineering and Computer Science
Washington State University, Pullman, WA 99164

PROTEIN DISORDER AND THE EVOLUTION OF MOLECULAR RECOGNITION: THEORY, PREDICTIONS AND OBSERVATIONS

A. K. DUNKER, E. GARNER, S. GUILLIOT
dunker@:mail.wsu.edu
Department of Biochemistry & Biophysics,
Washington State University, Pullman, WA 99164-4660
Tumor suppressor p53

Uversky et al. (2005) *J. Mol. Recogn.* 18, 343
**Support vector machine (SVM)**

How would you classify this data? But which is best?

Support vectors against which the margin pushes up

This is the simplest SVM, the linear SVM (LSVM)

The maximum margin linear classifier is considered best
2) Structural approach (interaction potential)
The protein non-folding problem

• **Protein folding problem**
  How does amino acid sequence determine protein structure?

• **Protein non-folding problem**
  How does amino acid sequence determine the lack of protein structure?
• Globular proteins have special sequences that enable the formation of a large number of favorable interactions

• IDPs contain (disorder-promoting) amino acids, which tend to avoid interacting with each other.

• An IDP thus cannot fold into a low-energy conformation.
A simple implementation, FoldUnfold

Calculates the contact number of amino acids
Estimating the total pairwise interresidue interaction energy of a sequence: IUPred

1) Calculate interresidue interaction energies from structure

2) Try to estimate the energy without knowing the structure

3) Apply the estimation to sequences w/o structure (e.g. to IDPs, which have no structure)
Interresidue interaction energy calculated for known structures

How to estimate the interresidue interaction energy of a protein of unknown structure or w/o structure?

**Structure**

<table>
<thead>
<tr>
<th>MODEL</th>
<th>1</th>
<th>MET A 23</th>
<th>2.191</th>
<th>28.312</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>1</td>
<td>N</td>
<td>-4.381</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>2</td>
<td>CA</td>
<td>-3.305</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>3</td>
<td>C</td>
<td>-3.706</td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>4</td>
<td>O</td>
<td>-4.867</td>
<td></td>
</tr>
</tbody>
</table>

**Sequence**

MKVPPHSIEA EQSVLGLML
DNERWDDVAE RVVADDFYTR
PRHIITEMA RLQESGSPID
LITLAESLER QGQLDSVGGF
AYLAELSKNT PSAANISAYA
DIVRERA VVR EMIS

Calculated energy per residue

Estimated energy per residue
Estimation of pairwise interresidue interaction energy from sequence

The contribution of individual AAs depends on its potential partners, i.e. its neighborhood. A quadratic formula is needed to take this into consideration.

\[ E(estimated) / L = \begin{pmatrix} n_A & n_C & \cdots & n_Y \end{pmatrix} \begin{pmatrix} P_{AA} & P_{AC} & \cdots & P_{AY} \\ P_{CA} & P_{CC} & \cdots \\ \vdots & \ddots & \vdots \\ P_{YA} & \cdots & P_{YY} \end{pmatrix} \begin{pmatrix} n_A \\ n_C \\ \vdots \\ n_Y \end{pmatrix} \]

The relationship between AA composition and energy is given by an optimized 20x20 energy predictor matrix, \( P_{ij} \)

\[ P_{ij} : \sum_k \left( E_k(calc) - E_k(est) \right)^2 \rightarrow \text{min} \]
Correlation of calculated and estimated interaction energies

The estimated energy for globular proteins and IDPs
Making it position-specific: the IUPred algorithm
IUPred, p53

[Graph showing disorder tendency for p53 residues]
DisEMBL™ Intrinsic Protein Disorder Prediction
DISOPRED2 Disorder Prediction Server
DRIPRED Web based predictor for disordered regions in proteins
FoldIndex© Estimate the fold probability of a protein
GlobPlot 2 Intrinsic Protein Disorder, Domain & Globularity Prediction
IUPred Prediction of Intrinsically Unstructured Proteins
PONDR® Predictors of Natural Disordered Regions
PreLink Prediction of unfolded segments in a protein sequence based on amino acid composition
RONN Regional Order Neural Network
VL2 DisProt Predictor of Intrinsically Disordered Regions
VL3, VL3H, VL3E DisProt Predictor of Intrinsically Disordered Regions
Alternative Splicing and Intrinsic Disorder

• Find proteins with both ordered and disordered regions.

• Find mRNA alternative splicing information for these proteins and map to the ordered and disordered regions.

• For alternatively spliced regions of mRNA, do they code for ordered protein more often or do they code for disordered protein more often?
Alternative Splicing
Alternative Splicing

5’ UTR  Coding Sequence  3’ UTR

mRNA  Transcription

Protein sequence  Translation
Alternative Splicing

5' UTR  Coding Sequence  3' UTR

mRNA 1  Transcription  mRNA 2

Isoform 1  Translation  Isoform 2
Structural Studies of AS

- Disordered AS regions
  - Pyrophosphorylase
  - RAC1 Tumor necrosis factor
  - Sulphotransferase

- Structured AS regions
  - EDA-A1
  - EDA-A2
  - Glutathione S-transferase
Studying the Relationship ID\(\leftrightarrow\)AS

DisProt

Database of proteins with experimentally determined structure and disorder
www.disprot.org

ASG (AS Gallery)

SwissProt (VarSplic)

ASED dataset:
46 proteins
74 characterized AS regions
>19,000 characterized residues, 35% ID
Results on ASED

Distribution of structurally characterized AS regions

- Fully disordered: 57%
- Fully ordered: 19%
- Partial: 24%
Enlarging the Dataset

PONDR® VSL1 ID predictor (> 80% accuracy) → Validation → ASED dataset

Analysis → ASSP dataset
- 558 AS human proteins from SwissProt
- 1,266 AS regions
Global Results

AS regions disorder distributions in ASED and ASSP

- ASED experimental
- ASED predicted
- ASSP predicted

Disorder content (AS regions)

Relative frequency
Alternative Splicing and Disorder

- **Ordered Proteins**: active site residues non-local in sequence, become associated by protein folding

- **Disordered Proteins and regions**: functional residues localized in sequence

- Functional regions for signaling and regulation are located one after another

- Alternative splicing edits functional sets and thereby leads to regulatory and signaling diversity
Disorder and Cell Signaling

Disorder and Drug Discovery

- The p53-MDM2 interaction is blocked by several drugs; one is in clinical trials and shows promise as an anti-cancer drug.
- The drug molecules bind to the ordered partner, preventing the disordered partner from binding.
- Such interactions are typically weak per unit of surface area, and the interaction surfaces can be small, thus such interactions are ideal drug targets.
- Molecular Kinetics has strategy to find all druggable MoRF-based interactions; bioinformatics indicates that more than one hundred are in cancer-associated proteins.
- Is this approach a new drug discovery pathway?