Addressing the Intrinsic Disorder Bottleneck in Structural Proteomics

Sang-Gook Han
National Institute of General Medical Sciences

• Support biomedical research for disease diagnosis, treatment, and prevention with understanding life processes.

• Major Initiative
  1. Protein Structure Initiative
  2. Pharmacokinetic Research Network: Medicine and Genes
  3. Model of Infectious Disease Agent Study
  4. Collaborative Research/Glue Grants : Solve problems together
Protein Structure Initiative

Human genes: 20,134
Mouse: 20,063
Yeast: 6,680
from Ensembl
## Protein Structure Initiative

### Class

<table>
<thead>
<tr>
<th>Class</th>
<th>Folds</th>
<th>Superfamilies</th>
<th>Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>218</td>
<td>376</td>
<td>608</td>
</tr>
<tr>
<td>$\beta$</td>
<td>144</td>
<td>290</td>
<td>560</td>
</tr>
<tr>
<td>$\alpha / \beta$</td>
<td>136</td>
<td>222</td>
<td>629</td>
</tr>
<tr>
<td>$\alpha + \beta$</td>
<td>279</td>
<td>409</td>
<td>717</td>
</tr>
<tr>
<td>Multi-domain</td>
<td>46</td>
<td>46</td>
<td>61</td>
</tr>
<tr>
<td>Membrane and cell surface</td>
<td>47</td>
<td>88</td>
<td>99</td>
</tr>
<tr>
<td>Small</td>
<td>75</td>
<td>108</td>
<td>171</td>
</tr>
<tr>
<td>Total</td>
<td>945</td>
<td>1539</td>
<td>2845</td>
</tr>
</tbody>
</table>

**SCOP 1.69v**

- Genome Sequences
- Protein Sequences
- Homology: 30 ~ 35 % sequence Identity
- Protein Structures
- Representative structures for each homology
Protein Structure Initiative

1. Target Selection
2. PCR amplification of the selected gene
3. Cloning the gene
4. Express the protein
5. Check the cloned gene and the expressed protein
6. X-ray or NMR measurement
7. Seek appropriate crystallization or NMR solution conditions
8. Protein Purification
9. Calculate comparative protein structure models
10. Functional Inference

Structure-determination Pipeline
Protein Structure Initiative

Discover High-throughput methods for protein structures, such as Cloning, purification, crystallization, Target selection, Bottlenecks

Appropriate candidates for structural determination
Under standard set of conditions
Target Selection

scoring system: 10000000000000 ~ 99999999999999

(1): 1 is the highest score and 9 is the lowest score calculated by other digits (properties)
(2): homologues structures in PDB
(3): fragment score
(4): state of progress of any similar protein in TargetDB
(5): structure class
(6): number of predicted transmembrane helices
(7): presence of a precursor sequence
(8): number of cysteines present in the sequence
(9): presence of low-complexity regions
(10): predicted likelihood that the protein will have a new fold
(11): evidence from gene chip data
(12): predicted disorder
(13): predicted solubility
(14): blacklist
Problem with disordered proteins

Lack of secondary or tertiary structure
Folding is condition-dependent
Various conformations in solution
33%~66% in Eukaryotes

Problem with disordered proteins

Measurement:
$^1$H-$^{15}$N heteronuclear single-quantum correlation (HSQC) NMR spectroscopy

Chemical shift depends on chemical nature of neighbor groups

Partially fold, multiple conformations, aggregated proteins

Folded protein: HSQC+  Unfolded protein: HSQC-
Problem with disordered proteins

X-ray Crystallography

Unlike NMR, No Positive or negative outcome

Lack of suitable crystal contact and other factors on protein structures makes even well-ordered proteins difficult to be crystallized.

Crowded environment enables disordered proteins to be stable.

More complicated than NMR
Problem with disordered proteins

Disorder predictors filter out the disordered proteins before cloning using protein sequence. It downgrade priority for structure determination.

SEG
GlobPlot and GlobDom
DisEMBL
DISOPRED
Charge-hydropathy prediction
Example: SEG

>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR
MANLGCWMLVLVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNNRYKFQGGQGWQGF
HGGGWGGQPQHGGGWQPHQUGGQPPHYQGQGGQPHQUGGGWGQTHSQWNKPSKPKNMKHAMGAAAGAA
VVGGLGGYMLGSAMSPRIIHFSGDYEDRYYRENMRHYPNQVYYPRMDEYSNQNNFVHDCVINITIKQHTVT
TTTKGENFTETDVKMMERVVEQMCITQVERESQAYYYQRGSSMVLFSFBRV
ILLISFLIFLIVG

Low complexity = a subset of disordered proteins

>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR

1-49 MANLGCWMLVLVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNNRY
50-94 WNTTGGSRYPGQGSPGGNNRY

ppqggggwqpphggwggqphggwqphgg 50-94 gwgqphggwggqphgg

95-112 THSQWNKPSKPKNMKHAMGAAAGAA

agaaaaagavvglggylgsams 113-135
136-187 RPIIHFSGDYEDRYYRENMRHYPNQVYYPRMDEYSNQNNFVHDCVINITIKQH

202-236 DVKMMERVVEQMCITQVERESQAYYYQRGSSMVLFS

237-252 sppvillisflifliv

253-253 G

Low complexity sequence

High complexity sequence
Example: DisEMBL

- **Loop or coil**: regions of lack secondary structure
- **Hot loop**: regions of lack secondary structure + high B-factors
- **Remark 465**: regions without apparent electron density in crystal structure
Example: GlobPlot

GlobDom: regions of globular domains, ordered regions
Results

• Effectively filtering out disordered proteins:

The number of HSQC+ targets increases
The number of HSQC- targets decreases

\[ T = Op + Dp \quad P(Op) \quad P(Dp) \quad P(HSQC+|Op) \quad P(HSQC-|Dp) \quad P(HSQC+) \]

With threshold 0.3 for PONDR VL-XT (30% of the residues are disordered)
Results

ROC Curve with all false disorder prediction rate equal weight
## Results

After 10,000 resampling iterations

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Summary</th>
<th>Whole curve</th>
<th>Partial curve (0 to 0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VL-2S</td>
<td>FR</td>
<td>0.728</td>
<td>(0.600, 0.839)</td>
</tr>
<tr>
<td>VL-2C</td>
<td>NR</td>
<td>0.776</td>
<td>(0.660, 0.871)</td>
</tr>
<tr>
<td>VL-XT</td>
<td>LR</td>
<td>0.758</td>
<td>(0.637, 0.859)</td>
</tr>
<tr>
<td>VL-2</td>
<td>LR</td>
<td>0.709</td>
<td>(0.583, 0.820)</td>
</tr>
<tr>
<td>DISOPRED2</td>
<td>NR</td>
<td>0.750</td>
<td>(0.623, 0.853)</td>
</tr>
<tr>
<td>DisEMBLE remark 465</td>
<td>NR</td>
<td>0.678</td>
<td>(0.549, 0.795)</td>
</tr>
<tr>
<td>VL-XT CDF</td>
<td>NA</td>
<td>0.674</td>
<td>(0.560, 0.778)</td>
</tr>
<tr>
<td>GlobPlot</td>
<td>LR</td>
<td>0.607</td>
<td>(0.473, 0.734)</td>
</tr>
<tr>
<td>SEG</td>
<td>LR</td>
<td>0.667</td>
<td>(0.561, 0.763)</td>
</tr>
<tr>
<td>Charge-Hydropathy</td>
<td>NA</td>
<td>0.602&lt;sup&gt;b&lt;/sup&gt;</td>
<td>(0.466, 0.727)</td>
</tr>
<tr>
<td>DisEMBL hotloops</td>
<td>LR</td>
<td>0.576&lt;sup&gt;b&lt;/sup&gt;</td>
<td>(0.441, 0.705)</td>
</tr>
<tr>
<td><strong>VL-2V</strong></td>
<td>LR</td>
<td>0.550&lt;sup&gt;b&lt;/sup&gt;</td>
<td>(0.415, 0.681)</td>
</tr>
<tr>
<td><strong>DisEMBL coils</strong></td>
<td>FR</td>
<td>0.551&lt;sup&gt;b&lt;/sup&gt;</td>
<td>(0.415, 0.684)</td>
</tr>
<tr>
<td><strong>GlobDom</strong></td>
<td>NR</td>
<td>0.507&lt;sup&gt;b&lt;/sup&gt;</td>
<td>(0.380, 0.635)</td>
</tr>
</tbody>
</table>

**Desired range of FDP range**

- Below 0.5 for whole curve
- 0.02 for partial curve in CI

NR: Number of residues
FR: Fraction of residues
LR: Longest region
Results

A. Filtering disordered proteins
   Increase ordered proteins in protein population
   Increasing FDPR, Increasing DPs, Decreasing OPs

B. Actual fraction of DPs = 0.35
Fund related to PSI

- **Structural Biology of Membrane Protein (R01)**
- NIH Guide: PA-06-119
- Release Date: Jan/20/2006
- Expiration Date: May/2/2009