Assessment of Disorder Prediction

CASP7
Asilomar, November 27, 2006

Lorenza Bordoli
Florian Kiefer
Torsten Schwede
Prediction Format and Data

The symbols for the 2 state order/disorder prediction are 'O'=order, 'D'=disorder. [...] Last column should indicate a probability of a residue being in the disordered region. The value of this confidence level is in the range of 0.0 - 1.0.

- Disorder: D / O
- Disorder probability: 0.0 ≤ p ≤ 1.0
- A residue has been defined as disordered:
  - X-Ray: no coordinates present for crystallized residues (SEQRES)
  - NMR: Residues whose conformation is not sufficiently defined by NMR restraints (ensemble variation or definition by experimentalist in the REMARK section)
CASP7 Disorder Dataset

Target residue classification:

- 6% disordered residues (CASP6: 5.25%)
- 94% ordered residues (CASP6: 94.75%)

Prediction for 96 targets assessed

- On a per residue level
- Only segments > 3 residues were considered.
- 19,816 residues as prediction targets
## Submissions by 19 Predictor Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Targets predicted</th>
<th>Residues predicted</th>
<th>P-values</th>
<th>D(cutoff)</th>
<th>O(cutoff)</th>
</tr>
</thead>
<tbody>
<tr>
<td>132</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>140</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>153</td>
<td>95</td>
<td>19723</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>168</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.594843</td>
<td>&lt;=0.553115</td>
</tr>
<tr>
<td>188</td>
<td>10</td>
<td>2221</td>
<td>cont.</td>
<td>&lt;0.50</td>
<td>&gt;=0.50</td>
</tr>
<tr>
<td>253</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>271</td>
<td>94</td>
<td>19368</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>272</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>284</td>
<td>95</td>
<td>19494</td>
<td>0.0,1.0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>393</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.594843</td>
<td>&lt;=0.553115</td>
</tr>
<tr>
<td>443</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>470</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>538</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>572</td>
<td>95</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>590</td>
<td>95</td>
<td>19541</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>594</td>
<td>82</td>
<td>16687</td>
<td>cont.*</td>
<td>&gt;=0.85</td>
<td>&lt;=0.85</td>
</tr>
<tr>
<td>609</td>
<td>93</td>
<td>19231</td>
<td>cont.**</td>
<td>&gt;0.5</td>
<td>&lt;=0.50</td>
</tr>
<tr>
<td>681</td>
<td>75</td>
<td>15180</td>
<td>cont.</td>
<td>&gt;=0.50</td>
<td>&lt;=0.50</td>
</tr>
<tr>
<td>686</td>
<td>96</td>
<td>19816</td>
<td>cont.</td>
<td>&gt;0.5</td>
<td>&lt;=0.5</td>
</tr>
</tbody>
</table>

In blue: servers

**0, 0.25, 0.33, 0.5, 0.66, 0.75, 1**

*In steps of 0.01*
Naïve predictors

- Frequencies of disordered residues at the N- and C- termini is higher than the average
- Several naives predictors were included, predicting 2,3,4,5,6,7,8,9,10 disordered residues at the N- and C- terminus

<table>
<thead>
<tr>
<th>Group</th>
<th>Targets</th>
<th>Residues predicted</th>
<th>P-values predicted</th>
<th>D(cutoff)</th>
<th>O(cutoff)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naiv[2-10]</td>
<td>96</td>
<td>19694</td>
<td>0.2, 0.8</td>
<td>&gt;=0.50</td>
<td>&lt;0.50</td>
</tr>
</tbody>
</table>
Evaluation Criteria

- Continuous prediction (P-value):
  - ROC curves

- Binary Prediction (O / D)
  1. Sensitivity, Specificity, Product and Average
  2. Sw score (CASP6, Dunbrack [1])

ROC Curves

The area under the curve (AUC) is used as measure of prediction accuracy ($Sroc$)

- For each value of $P$:
  - $TP(P)$ rate: fraction of true positives in prediction with $p>P$
  - $FP(P)$ rate: fraction of false positives in prediction with $p>P$
Sensitivity and Specificity are commonly used measures to evaluate predictive accuracy:

\[
\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{TP}{N_{\text{disorder}}} = \frac{TP(P)}{N_{\text{disorder}}}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} = \frac{TN}{N_{\text{order}}} = \frac{TN(P)}{N_{\text{order}}}
\]
Binary Scores

- Combined scores of Sensitivity and Selectivity:
  - Good methods would ideally have both a high sensitivity and a high specificity
  - The impact of Specificity (Sp) can be seen on the rate of false positive predictions FP = (1 – Sp)
  - Combined to make a single measurement criteria:

\[
\sqrt{S_{\text{product}}} = \sqrt{S_{\text{sens}} \times S_{\text{spec}}} = \sqrt{\frac{TP \times TN}{N_{\text{disorder}} \times N_{\text{order}}}}
\]

\[
ACC = \frac{S_{\text{sens}} + S_{\text{spec}}}{2}
\]
Sw score (CASP6, Dunbrack)

Most residues (in CASP7: 94%) are ordered and a simple Q2 measure would not work well.

⇒ weighted score to reward for correctly predicting a disordered residue rather than an ordered one (Sw) introduced by Dunbrack:

\[
S_w = \frac{S}{S_{\text{max}}} = \frac{W_{\text{disorder}} TP - W_{\text{order}} FP + W_{\text{order}} TN - W_{\text{disorder}} FN}{W_{\text{disorder}} N_{\text{disorder}} + W_{\text{order}} N_{\text{order}}}
\]

⇒ Weights: \( W_{\text{disorder}} = 94, \ W_{\text{order}} = 6 \)

⇒ Sw varies between -1 and +1, with a random prediction resulting in \( Sw = 0 \)

We used a bootstrapping procedure to test the statistical significance of the assessment:

- 80% of randomly chosen target structures were assessed repeatedly 1000 times to derive standard error of each of the binary scores.
- 80% of randomly chosen target residues were assessed repeatedly 1000 times to derive standard error of each of the binary scores.

Standard error of ROC curves were estimated according to Hanley J.A. & McNeil B.J. (1983), Radiology, 148,839-843
ROC curves including naïve predictors

Relationship between AUC and Wilcoxon test statistics [1]:
- Performance of a method is different from random
- Statistical significance between two methods.


*Not assessed well by ROC curves*
Disorder Prediction ranked by ROC-AUC
Distribution of the length of disordered regions in CASP7
Short (<= 10) vs. Long (> 10) disordered regions

- Groups 253 and 470 are more accurate on long disordered regions
- Group 443 more accurate on short disordered regions
- Group 590 comparable accuracy on both long and short regions
Disorder Prediction ranked by ROC-AUC

10 residues N- C- terminal removed

- Internal disordered regions are more difficult to be detected than terminal disordered regions
Results

- In summary, groups 590, 253 followed by 443 & 470 give better performance than the other groups.
- Group 590 has highest sensitivity but a relatively high FP rate.
- At a lower FP rate groups 253 & 470 have comparable Sw and ACC scores to the rest of the top groups.
- Groups 253 and 470 more accurate on long disordered regions.
- Groups 443 more accurate on short disordered regions.
- Group 590 comparable accuracy on both long and short regions.
## Confusion matrix

### Gr.590

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Order</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td>15365</td>
<td>2991</td>
</tr>
<tr>
<td>Disorder</td>
<td>295</td>
<td>781</td>
</tr>
</tbody>
</table>

### Gr.253

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Order</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td>17990</td>
<td>637</td>
</tr>
<tr>
<td>Disorder</td>
<td>590</td>
<td>490</td>
</tr>
</tbody>
</table>

### Gr.443

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Order</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td>17218</td>
<td>1409</td>
</tr>
<tr>
<td>Disorder</td>
<td>478</td>
<td>602</td>
</tr>
</tbody>
</table>

### Gr.470

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Order</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td>17750</td>
<td>877</td>
</tr>
<tr>
<td>Disorder</td>
<td>621</td>
<td>459</td>
</tr>
</tbody>
</table>

### Gr.140

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Order</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td>15910</td>
<td>2717</td>
</tr>
<tr>
<td>Disorder</td>
<td>435</td>
<td>645</td>
</tr>
</tbody>
</table>

### Gr.609

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Order</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td>16492</td>
<td>1589</td>
</tr>
<tr>
<td>Disorder</td>
<td>491</td>
<td>551</td>
</tr>
</tbody>
</table>
Examples

- NMR-structure (T0357, 2HFQ):
Examples: T0357

- Nobody identified the disorder segment 24-30.
Examples identified correctly

- X-Ray structure (T0293, 2H00):
  - 32, 7 & 5 residue long disordered regions
  - Correctly identified by 590

590, 140, 609

590, 153

Majority of the groups
CASP6 vs. CASP7
CASP6 vs. CASP7

- Caveat: Comparison is based on *same scores* for a *different data set (CASP6)*.
- If we assume the comparison as valid, no improvement over CASP6 is observed.
Methods

➢ Group 590

➢ Group 253
CBRC-DR — K. Shimizu, S. Hirose, N. Inoue, S. Kanai and T. Noguchi (SVM, specialized predictors for long and short disordered regions.)

➢ Group 443
fais - T. Ishida, and K. Kinoshita

➢ Group 470
DISOPRED — K. Bryson and D. Jones (SVM trained on PSI-BLAST Profiles)
Acknowledgements

- Neil Clark
- Randy Read
- Michael Tress
- Alfonso Valencia
- Torsten Schwede
- Florian Kiefer
- Jürgen Kopp
- James Battey
- Roland Dunbrack
- Yumi Jin
- Anna Tramontano
- Krzysztof Fidelis
- Andriy Kryshtafovych
- John Moult
- Burkhard Rost
Discussion panel

- **Applications**: What are the main applications of disorder prediction? What are their requirements? How sensitive are they to over/under prediction?

- **Performance**: Are current disorder prediction methods performing sufficiently well for these practical applications? Do we deliver?

- **Definition of "Disorder"**: Is there consensus about what we define as a "disordered region"? Based on which experimental evidence?

- **Data**: What experimental data is required for method development?

- **Is the CASP7 data adequate for disorder prediction assessment?**

- **Human vs. servers**: Does human intervention improve disorder prediction?

- **Future?**