Mining the Structural Genomics Pipeline: Identification of Protein Properties that Affect High-throughput Experimental Analysis

by Chern-Sing Goh, Ning Lan et al.

Presented by Jingjun Sun
Outline

- Background
- Methods
- Results
- Conclusion
Structural genomics generates unique datasets in terms of protein physical and chemical properties

discover significant protein features that influence a protein being determined structurally

identify potential bottlenecks in various stages of the structural genomics pipeline
Methods

- decision trees
- conservation, charged residues, hydrophobic patches, binding partners, length
- cloning, expression, purification, structural determination
Results
Structure vs No Structure

TargetDB

COG < 1 [27267]

- proteins have COGs
  \[
  \frac{15952}{27267} = 0.59
  \]

- proteins structure COGs
  \[
  \frac{297}{350} = 0.85
  \]

DE < 9.7

- 2432 6

GAVLI < 31.7

- 2600 16

C < 1.8

- 8044 249

- 2579 26

15655

297
### Results con'd

<table>
<thead>
<tr>
<th>Feature</th>
<th>Total targets</th>
<th>Structure targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>COGs</td>
<td>34.6%</td>
<td>38.4%</td>
</tr>
<tr>
<td>GAVLI</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>Hp_aa</td>
<td>291</td>
<td>243</td>
</tr>
<tr>
<td>Length</td>
<td>12.5%</td>
<td>12.6%</td>
</tr>
<tr>
<td>KR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\frac{\text{proteins have COGs}}{\text{total target}} = \frac{15952}{27267} = 0.59
\]

\[
\frac{\text{proteins structure COGs}}{\text{total structure target}} = \frac{297}{350} = 0.85
\]

15952 proteins have COGs

350 proteins structure COGs
### Results con'd

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Total targets</th>
<th>SD targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAVLI</td>
<td>Average GAVLI composition(%)</td>
<td>34.6</td>
<td>38.4</td>
</tr>
<tr>
<td>hp_aa</td>
<td>average No. of hp residues</td>
<td>15</td>
<td>7</td>
</tr>
</tbody>
</table>

*within a hp stretch*

```
SEGAVLIWKAAAVLIRGAVLI...
NQSEGAVLIEKDWRLIRMS...
```

20
Expressed vs cloned & not expressed

COG < 1 [14385]

any_partners<1

length < 524

.44
1816

.13
416

3182

.76
208
659

pl < 5.9

2785

979

.73

2203

3043

.58

5827

proteins not expressed COGs = \frac{3182}{5622} = 0.57

proteins expressed COGs = \frac{5827}{8319} = 0.70
Results con'd

Purified vs expressed & not purified

\[
\frac{\text{proteins not purified COGs}}{\text{total not purified}} = 0.65
\]

\[
\frac{\text{proteins purified COGs}}{\text{total purified}} = 0.75
\]

Structure vs purified & not structure

\[
\frac{\text{proteins not structure COGs}}{\text{total not structure}} = 0.75
\]

\[
\frac{\text{proteins structure COGs}}{\text{total structure}} = 0.85
\]
Percents that belong to COGs

85% structures (370)
- 75% purified (4115)
- 70% expressed (8587)

63% cloned (14767)
59% total (27711)

75% not structures (3745)
- 65% not purified (4472)
- 53% not expressed (6180)
- 54% not cloned (12944)
<table>
<thead>
<tr>
<th>Structure</th>
<th>Numbers of Hydrophobic Resides</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total (27711)</td>
</tr>
<tr>
<td></td>
<td>Cloned (14767)</td>
</tr>
<tr>
<td></td>
<td>Expressed (8587)</td>
</tr>
<tr>
<td></td>
<td>Structures (370)</td>
</tr>
<tr>
<td></td>
<td>Not expressed (6180)</td>
</tr>
<tr>
<td></td>
<td>Not purified (4472)</td>
</tr>
<tr>
<td></td>
<td>Not structures (12944)</td>
</tr>
<tr>
<td></td>
<td>6.6 Structures (3745)</td>
</tr>
</tbody>
</table>

Results con'd

Numbers of Hydrophobic residues
Results con'd

- charged residue composition
- bottleneck at purification
- number of binding partners
- bottlenecks at purification and structure
- length
- bottleneck at expression
Conclusion

- COGs
- charged residue composition
- number of hydrophobic residues in hydrophobic stretches
- number of binding partners it has
- Length
- nuclear localization signals ...
Questions & Comments

Thanks!