Newly Determined Structures - Number of New Folds
Newly Determined Structures - Fraction of New Folds

% of Total Folds which are New

Year

Classification of structures

SCOP:  http://scop.mrc-lmb.cam.ac.uk/scop/  
       (domains, good annotation)

CATH:  http://www.biochem.ucl.ac.uk/bsm/cath/

CE:     http://cl.sdsc.edu/ce.html

Dali Domain Dictionary:  
                        http://columba.ebi.ac.uk:8765/holm/ddd2.cgi

FSSP:    http://www2.ebi.ac.uk/dali/fssp/  
          (chains, updated weekly)

HOMSTRAD:  
            http://www-cryst.bioc.cam.ac.uk/~homstrad/

HSSP:    http://swift.embl-heidelberg.de/hssp/
SCOP Hierarchy of Structures

**Family:** evolutionarily related with a significant sequence identity - 2327 in SCOP

**Superfamily:** different families whose structural and functional features suggest common evolutionary origin - 1294 in SCOP

**Fold:** different superfamilies having same major secondary structures in same arrangement and with same topological connections - 800 in SCOP
Scop Classification Statistics

1.65 release 20619 PDB Entries (1 August 2003).
54745 Domains

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of folds</th>
<th>Number of superfamilies</th>
<th>Number of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha proteins</td>
<td>179</td>
<td>299</td>
<td>480</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>126</td>
<td>248</td>
<td>462</td>
</tr>
<tr>
<td>Alpha and beta proteins (a/b)</td>
<td>121</td>
<td>199</td>
<td>542</td>
</tr>
<tr>
<td>Alpha and beta proteins (a+b)</td>
<td>234</td>
<td>349</td>
<td>567</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>38</td>
<td>38</td>
<td>53</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>36</td>
<td>66</td>
<td>73</td>
</tr>
<tr>
<td>Small proteins</td>
<td>66</td>
<td>95</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>800</td>
<td>1294</td>
<td>2327</td>
</tr>
</tbody>
</table>
Scop Classification Statistics
1.69 release
25973 PDB Entries (1 Oct 2004). 70859 Domains
(excluding nucleic acids and theoretical models)

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of folds</th>
<th>Number of superfamilies</th>
<th>Number of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha proteins</td>
<td>218</td>
<td>376</td>
<td>608</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>144</td>
<td>290</td>
<td>560</td>
</tr>
<tr>
<td>Alpha &amp; beta proteins ((\alpha/\beta))</td>
<td>136</td>
<td>222</td>
<td>629</td>
</tr>
<tr>
<td>Alpha &amp; beta proteins ((\alpha+\beta))</td>
<td>279</td>
<td>409</td>
<td>717</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>46</td>
<td>46</td>
<td>61</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>47</td>
<td>88</td>
<td>99</td>
</tr>
<tr>
<td>Small proteins</td>
<td>75</td>
<td>108</td>
<td>171</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>945</strong></td>
<td><strong>1539</strong></td>
<td><strong>2845</strong></td>
</tr>
</tbody>
</table>
Domain may be more sensitive but depends on correct partition
Hundreds of thousands of gene sequences translated to proteins (SwissProt, PIR)

27,112 solved structures (PDB)

Goals:
- Predict structure from sequence
- Predict function based on sequence
- Predict function based on structure
## PDB Current Holdings Breakdown

**Tuesday May 30, 2006**

<table>
<thead>
<tr>
<th>Exp. Method</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein/NA Complexes</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray</td>
<td>28945</td>
<td>900</td>
<td>1350</td>
<td>28</td>
<td>31223</td>
</tr>
<tr>
<td>NMR</td>
<td>4583</td>
<td>699</td>
<td>121</td>
<td>6</td>
<td>5409</td>
</tr>
<tr>
<td>Electron Microscopy</td>
<td>88</td>
<td>9</td>
<td>28</td>
<td>0</td>
<td>125</td>
</tr>
<tr>
<td>Other</td>
<td>73</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>33689</strong></td>
<td><strong>1612</strong></td>
<td><strong>1502</strong></td>
<td><strong>34</strong></td>
<td><strong>36837</strong></td>
</tr>
</tbody>
</table>

- **Molecule Type**

**Tuesday May 30, 2006**
Many proteins with dissimilar sequences fold into similar structures

- Estimated number of folds: 600-5000

Protein Folds: sequential and spatial arrangement of secondary structures
SH3 domains are small protein modules. Five SH3 domains are shown.

Structural similarities are observed among these different protein families but no detectable sequence similarities are found.
Fraction of new folds (PDB new entries in 1998)

3358 new chains

95% sequence identity

664 unique chains

Sequence comparison with previously known proteins

490 similar

174 non-similar

Elimination of transmembrane and virus capsid proteins

147 globular proteins

Sequence comparison

196 domains

Structure comparison

147 known fold

49 novel fold

Relationship of Similarity in Sequence to that in Function

Wilson et al. JMB 297: 233
Relationship of Similarity in Sequence to that in Function

Can transfer both Fold & Functional Annotation

Wilson et al. JMB 297: 233
Relationship of Similarity in Sequence to that in Function

- Can transfer both Fold & Functional Annotation
- Can transfer Annotation related Fold but not Function
- Cannot transfer Fold or Functional Annotation ("Twilight Zone")

Wilson et al. JMB 297: 233
Can we predict function via fold similarity?

E.G. cspA OB fold suggests DNA binding [Montelione]

**Issue**: To what degree does fold determine function, globally?
Different Folds with Same Function (Carbonic Anhydrases, 4.2.1.1)
Many Functions on Same Scaffold (TIM-barrel)
To what degree is fold associated with function?
Folds with multiple functions

Number of functions associated with a fold

Frequency in database of 229 folds

Hegyi & Gerstein, JMB 288: 147

[Similar results by Thornton]