Exercise 1:
Pymol and structural classification databases
Administrative issues

- Course number 81813 – separate enrollment
- Course site: moodle2.cs.huji.ac.il (or http://www.cs.huji.ac.il/~fora/81855/)
- Submit a minimum of 8 exercises via email
  - Submission dates: 28/03 (ex1), 04/04 (ex2), 11/04 (ex3), 02/05 (ex4), 16/05 (ex5), 23/05 (ex6), 30/05 (ex7), 06/06 (ex8), 13/06 (ex9), 20/06 (ex10), 27/06 (ex11)
- Contact:
  - Email: alisa.khramushin@mail.huji.ac.il
  - Phone: 87063
Amino acids structures and nomenclature

Hydrogen

Amino

Carboxyl

R-group (variant)

peptide
Classification of protein “closeness”

- Describe structural and evolutionary relationships of proteins with determined three-dimensional structures
- Hierarchical classification
- Three commonly used databases:
  - SCOP
  - CATH
  - ECOD
Hierarchical division to classes, folds, superfamilies, families and domains
Mostly manual (last updated in 2009) – highly reliable
Classified unit is usually a protein domain.

same major secondary structures in same arrangement; same topological connections

low sequence identity; indication of common evolutionary origin (according to structure / function)

> 30% sequence identity / same function and structure

CATH

http://www.cathdb.info/

- Manual curation, aided by classification and prediction algorithms (last update correct for 05/2017)

- Hierarchical clustering into Class, Architecture, Topology (fold), and Homologous superfamily
  - **Class** - secondary structure content
  - **Architecture** - similar arrangement of secondary structures in space
  - **Topology** - fold groups; finer similarity in connectivity (no clear evolutionary relationship)
  - **Homologous superfamily** – high sequence similarity & some structural similarity and *vice versa* (common ancestor)

New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Sillitoe I *et. al.*, NAR (2013)

So what’s the difference?

- Assignment is of domains
  - 82% agreement on number of domains per entry
  - half of the 2-domain CATH entries considered a single domain in SCOP

<table>
<thead>
<tr>
<th>SCOP (number of domains)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4475</td>
<td></td>
<td>51</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>80</td>
<td>1007</td>
<td>51</td>
<td>104</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>7</td>
<td>159</td>
<td>10</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>38</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ECOD

http://prodata.swmed.edu/ecod/

- Groups primarily by evolutionary relationships rather than topology
- Based on an automated pipeline; updated weekly (contains almost all structures in the PDB)
- Clusters structures into slightly different levels:
  - Similar secondary structure compositions and geometric shapes
  - Some evidence of (broad) homology.
  - Common ancestry (sequence-structure similarity, functional similarity, shared unusual features, literature)
  - Similar topological connections
  - Significant sequence similarity