Bioinformatics Resources
- SQL -

Lecture & Exercises
Prof. B. Rost, Dr. L. Richter, J. Reeb
Institut für Informatik I12
Orga - Exam Date/Change?

- Exam scheduled for Friday, July 31\textsuperscript{st}
- Room: MW 0250 (Mechanical Engineering Building)
- Change requested: Wednesday, Jul 22\textsuperscript{nd}, in the range of 14-17 o’clock
- Voting
Classification of Structures: CATH/SCOP

- came up in the middle of the 1990s
- both are quite similar
- aim: organize the protein structures available in PDB, based on single domains
- hierarchical system (roughly):
  - secondary structure content
  - fold
  - super families
  - families
SCOP: a Structural Classification of Proteins

- fully manually curated, driven by expert analysis
- associated with the ASTRAL compendium
- latest news: SCOPe (UC Berkeley), SCOP2 (MRC Lab Mol Biol, Cambridge, UK)
CATH - Faces

taken from http://www.tgac.ac.uk/scientific-advisory-board/

taken from http://www.ebi.ac.uk/about/people/janet-thornton

BioinfRes SS 15
CATH

- semi-automatic procedure for deriving a novel hierarchical classification of protein domain structures

- four main levels:
  - C: protein class, mainly secondary structure composition of each domain
  - A: architecture, summarizes shapes based on orientation of secondary structure elements
  - T: topology, sequential connectivity is considered
  - H: homologous superfamily, high similarity with similar functions, evolutionary relationship assumed
Annual increase in the numbers of protein domain structures in the PDB (top plot, [11,12]). The lower lines show the numbers of identical families (H-level, 100% sequence identity between structures within the family and 100% overlap), non-identical families (N-level, > 95% sequence identity, 85% overlap), sequence families (S-level, > 35% sequence identity, 60% overlap), homologous superfamilies (H-level, > 25% sequence identity, SSAP >80 and 60% overlap), and topological or fold families (T-level, SSAP >70), where SSAP is a structural comparison score.
Numbering Scheme

- C: 1, 2, 3 (alpha, beta, alpha/beta) +1
- A: same architecture, different topology (31)+10
- T: Topology (connection of secondary structure elements) +10 (505)
- H: Homology (families) +10 (645)
The numbers of fold families (T-level), homologous superfamilies (H-level) and domain structures in different architectures are shown for the mainly $\alpha$, mainly $\beta$ and $\alpha$-$\beta$ classes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Architecture</th>
<th>Number of T-levels</th>
<th>Percentage of all T-levels*</th>
<th>Number of H-levels</th>
<th>Percentage of all H-levels*</th>
<th>Number of domains</th>
<th>Percentage of all domains*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainly $\alpha$</td>
<td>Non-bundle</td>
<td>86</td>
<td>17.03</td>
<td>93</td>
<td>14.42</td>
<td>1455</td>
<td>18.01</td>
</tr>
<tr>
<td></td>
<td>Bundle</td>
<td>34</td>
<td>6.73</td>
<td>39</td>
<td>6.05</td>
<td>226</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td>Few SS</td>
<td>25</td>
<td>4.95</td>
<td>25</td>
<td>3.88</td>
<td>112</td>
<td>1.39</td>
</tr>
<tr>
<td>Mainly $\beta$</td>
<td>Ribbon</td>
<td>17</td>
<td>3.37</td>
<td>17</td>
<td>2.64</td>
<td>114</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>Single sheet</td>
<td>5</td>
<td>0.99</td>
<td>6</td>
<td>0.93</td>
<td>56</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Roll</td>
<td>6</td>
<td>1.19</td>
<td>6</td>
<td>0.93</td>
<td>55</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Barrel</td>
<td>22</td>
<td>4.36</td>
<td>29</td>
<td>4.50</td>
<td>861</td>
<td>10.66</td>
</tr>
<tr>
<td></td>
<td>Clam</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Sandwich</td>
<td>21</td>
<td>4.16</td>
<td>43</td>
<td>6.67</td>
<td>1236</td>
<td>15.30</td>
</tr>
<tr>
<td></td>
<td>Distorted sandwich</td>
<td>14</td>
<td>2.77</td>
<td>14</td>
<td>2.17</td>
<td>83</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>Trefoil</td>
<td>1</td>
<td>0.20</td>
<td>4</td>
<td>0.62</td>
<td>49</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Orthogonal prism</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>4</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Aligned prism</td>
<td>1</td>
<td>0.20</td>
<td>2</td>
<td>0.31</td>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Four-propeller</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Six-propeller</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>37</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>Seven-propeller</td>
<td>2</td>
<td>0.40</td>
<td>2</td>
<td>0.31</td>
<td>11</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Eight-propeller</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Two-solenoid</td>
<td>2</td>
<td>0.40</td>
<td>3</td>
<td>0.47</td>
<td>5</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Three-solenoid</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>5</td>
<td>0.99</td>
<td>5</td>
<td>0.78</td>
<td>104</td>
<td>1.29</td>
</tr>
<tr>
<td>$\alpha$-$\beta$</td>
<td>Roll</td>
<td>24</td>
<td>4.75</td>
<td>33</td>
<td>5.12</td>
<td>469</td>
<td>5.81</td>
</tr>
<tr>
<td></td>
<td>Barrel</td>
<td>8</td>
<td>1.58</td>
<td>20</td>
<td>3.10</td>
<td>365</td>
<td>4.52</td>
</tr>
<tr>
<td></td>
<td>Two-layer sandwich</td>
<td>77</td>
<td>15.25</td>
<td>112</td>
<td>17.36</td>
<td>957</td>
<td>11.85</td>
</tr>
<tr>
<td></td>
<td>Three-layer ($\alpha\beta\alpha$) sandwich</td>
<td>78</td>
<td>15.45</td>
<td>115</td>
<td>17.83</td>
<td>1396</td>
<td>17.28</td>
</tr>
<tr>
<td></td>
<td>Three-layer ($\beta\alpha\beta$) sandwich</td>
<td>3</td>
<td>0.59</td>
<td>3</td>
<td>0.47</td>
<td>11</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Four-layer sandwich</td>
<td>4</td>
<td>0.79</td>
<td>4</td>
<td>0.62</td>
<td>12</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Box</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Horseshoe</td>
<td>1</td>
<td>0.20</td>
<td>1</td>
<td>0.16</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>34</td>
<td>6.73</td>
<td>34</td>
<td>5.27</td>
<td>253</td>
<td>3.13</td>
</tr>
<tr>
<td>Few SS</td>
<td>Irregular</td>
<td>14</td>
<td>2.77</td>
<td>14</td>
<td>2.17</td>
<td>98</td>
<td>1.19</td>
</tr>
</tbody>
</table>
Pfam

- current version is 27.0, March 2013, 14831 families
- hosted by the EBI
- Citation: “Pfam: the protein families database” Nucl. Acids Res. (1 January 2014) 42 (D1): D222-D230. doi: 10.1093/nar/gkt1223
Pfam

- Pfam-A: curated seed alignment derived from Pfamseq (UniProtKB based), profile HMMs for the seed alignment, full alignment with all HMM detected sequences
- Pfam-B: un-annotated, automatically generated from non-redundant cluster from ADDA
Terms

- **Family**: collection of related protein regions
- **Domain**: structural unit
- **Repeat**: shot unit which is unstable in isolation but forms a stable structure when found in multiple copies
- **Motif**: short unit found outside globular domains
- **Clans**: related group of Pfam entries based on similarity in sequence, structure of profile-HMM
Site organisation

- Clan
- Proteome
- Structure
- Sequence

Family
Pfam Numbers

- 14381 Pfam-A families
- 4563 are classified into 515 clans
- the Pfam-A release matches 79.9% of the 23.2 Mio sequences in the corresponding UniProt db
- coverage of 90.5% of SwissProt human
- use of jackhmmer (from HMMER3 package)
- consider CATH and PDB
Table 1. The reduction in size of RP versus full alignments

<table>
<thead>
<tr>
<th>Family identifier (accession)</th>
<th>Seed</th>
<th>Full</th>
<th>RP75</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC_trans (PF00005)</td>
<td>55</td>
<td>363 409</td>
<td>26% (93 265)</td>
</tr>
<tr>
<td>COX1 (PF00115)</td>
<td>94</td>
<td>254 351</td>
<td>1% (2006)</td>
</tr>
<tr>
<td>zf-H2C2_2 (PF13465)</td>
<td>163</td>
<td>227 898</td>
<td>61% (138 033)</td>
</tr>
<tr>
<td>WD40 (PF00400)</td>
<td>1804</td>
<td>193 252</td>
<td>65% (125 805)</td>
</tr>
<tr>
<td>MFS_1 (PF07690)</td>
<td>195</td>
<td>181 668</td>
<td>30% (55 719)</td>
</tr>
<tr>
<td>RVT_1 (PF00078)</td>
<td>152</td>
<td>172 360</td>
<td>5% (8257)</td>
</tr>
<tr>
<td>BPD_transp_1 (PF00528)</td>
<td>81</td>
<td>156 339</td>
<td>23% (36 523)</td>
</tr>
<tr>
<td>Response_reg (PF00072)</td>
<td>57</td>
<td>151 337</td>
<td>29% (44 329)</td>
</tr>
<tr>
<td>GP120 (PF00516)</td>
<td>24</td>
<td>146 453</td>
<td>N/A</td>
</tr>
<tr>
<td>HATPase_c (PF02518)</td>
<td>659</td>
<td>129 386</td>
<td>28% (36 085)</td>
</tr>
</tbody>
</table>

The reduction in the size of the full alignments varies from family to family, reflecting in part the bias in the diversity within a family. To provide more useable samples of the sequence diversity within the family, we now calculate model-matches for four additional sequence sets, based on RPs.
Databases - SQL

- Overlap with database lecture
- “SQL crash course”
- no design theory
- no normalization
- standard books like:
More Books

Selected SQL Topics

- Table modifications
  - insert, update, create, alter
- Data retrieval and reporting/aggregation
  - select, average, sum
- Combination and Performance
  - join
- Access control and permissions
  - grant
- Backup and Restore / Input-output
Reasons for DBMS

- redundancy, consistency
- limited access
- difficult multi-user access
- loss of information
- loss of integrity
- security issues
- expensive application development
Abstraction layers

View 1

View 2

Logical Layer

Physical Layer
Various Data Models

- Network model
- Hierarchical model
- **Relational Model**
- XML schema
- Object-oriented model
- Deductive model
Relational Model

<table>
<thead>
<tr>
<th>Students</th>
<th>Attends</th>
<th>Lectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matric</td>
<td>Name</td>
<td>Matric</td>
</tr>
<tr>
<td>123455</td>
<td>Mayer</td>
<td>123455</td>
</tr>
<tr>
<td>233457</td>
<td>Huber</td>
<td>233457</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Select** Name

**From** Students, Attends, Lectures

**Where** Students.Matric = Attends.Matric and Attends.LectureNo = Lectures.LectureNo and Lectures.Title = ‘Genomics’;

**Update** Lectures

**Set** Title = ‘Genomics of Mammalian’

**Where** LectureNo = 5;
Entity Relationship Model

- Graphical Notation
- Models real world “entities” and “relation”
- allows for “attributes”
- allows for functionalities (1:1, 1:n, n:m)
- allows to define keys
- key: a set for attributes which values combination allow unambiguous instance identification
Notation

- **Student**: (strong) Entity
- **Name**: Attribute, key: underlined
- **Attends**: Relation
- **weak Entity (depend on others)**
Functionality

- Student
- Lecture
- Grade

Attends

N

M

BioinfRes SS 15
taken from Prof. Kempers database lecture WS 13/14
BioinfRes SS 15
Prüfungen als schwacher Entitytyp

- Studenten ablegen Prüfungen
- MatrNr
- VorlNr
- Vorlesungen umfassen
- Professoren abhalten
  - Note
  - PrüfTeil
  - PersNr

- Mehrere Prüfer in einer Prüfung
- Mehrere Vorlesungen werden in einer Prüfung abgefragt

taken from Prof. Kempers database lecture WS 13/14
SQL

- implemented by most available dbms manufacturer
- but: not always all specified features implemented
- not everything is specified!
- especially admin/server maintenance is often vendor specific
SQL Data Types

- char
- varchar
- binary and varbinary
- blob and text
- numeric, decimal, integer (exact)
- approximate: float, double
SQL Data Types

- various formats for time and date
- enum: one out of a defined set
- set: zero or more items out of a predefined list

For more information see the live tour through