Exercise 2

Protein Prediction I for CS

Biological Databases
Structure Determination
Primary Structure:

- Linear sequence of amino acids
- Oriented from N- to C-terminus
- (Typically) starts with Methionine

Different Encodings:

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Protein Aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D-information: sequence of amino acids as string</td>
<td>Primary Structure: Amino Acid sequence</td>
</tr>
<tr>
<td>2D-information: 2D-array, contact map</td>
<td>Secondary structure elements like helices or sheets, ...</td>
</tr>
<tr>
<td>3D-information: coordinates or atom couplings</td>
<td>Tertiary structure: spatial arrangement of secondary structure elements (incl. amino acids, atoms, ...)</td>
</tr>
</tbody>
</table>
Sequence length distribution in UniProtKB/Swiss-Prot

X-ray Crystallography

- Crystallize
- Diffract
  - Measurement shows electron density
  - Not visible: H and flexible parts
  - Ambiguity (e.g. $NH_2/OH$)
- Build Model

©© Protein crystals grown in space by NASA
- Crystallize
- Diffract
  - Measurement shows electron density
  - Not visible: H and flexible parts
  - Ambiguity (e.g. $NH_2/OH$)
- Build Model

© Adapted from Example of electron density map by Bassophile
- Spinning nucleus absorbs radiation
- Exact frequency depends on environment → chemical shift
- J-Coupling → split
- Proteins too crowded for normal NMR

©️ Adapted from 1H NMR Ethanol Coupling shown by T.vanschaik
Protein NMR

- Sample preparation (<1 ml, <3 mmol/l) ($^{15}$N—, $^{13}$C—labeled samples help assign residues)
- Data collection (2D spectra)
- Resonance assignment
- Restraint generation
  - Distance (peak intensity)
  - Angle (coupling magnitude)
- Model building
  → ensemble of solutions
- No standard quality measures

Image from PDB101 determining structure, RCSB PDB
doi:10.1093/nar/28.1.235
Electron Microscopy

- Sample preparation (mostly CRYO)
- Measurement:
  - Electron diffraction
  - Electron tomography
- Model building:
  - Reconstructing electron density
  - Fitting atoms (e.g. from X-ray)
- No standard quality measures

Electron Interaction with Matter by Claudionico commonswiki
Rate of Protein Structure Determination (Log Scale)

Adapted from *Rate of Protein Structure Determination* by D Wells
BIOLOGICAL DATABASES
A study $S$ finds a possible link between protein $P$ and disease $D$. You are a researcher working on $D$ and after reading $S$ you would like to know more about $P$. Where do you look for further information?
Available Information:

- Sequence (only field which has to exist)
- Function
- Localization
- Taxonomy
- Interactions
- Expression (in which cell type or tissue does P appear)
- Pathology (relation between P and D which would you like to prove)
- Structure (links to PDB – if structure is available)
- Similar proteins
- Family/Domain

UniProtKB Components

- UniProtKB
  - UniProtKB/Swiss-Prot
  - UNiProtKB/TrEMBL
- UniPar: pure sequence archive, no annotations
- UniRef: consists of three databases of clustered sets of protein sequences (UniRef100, UniRef90, UniRef50) using the CD-HIT algorithm
- UniMes: data from metagenomic and environmental sample, not in UniProtKB
<table>
<thead>
<tr>
<th>Experimental Method</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein/NA Complex</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-Ray</td>
<td>127188</td>
<td>2015</td>
<td>6551</td>
<td>8</td>
<td>135762</td>
</tr>
<tr>
<td>NMR</td>
<td>11073</td>
<td>1279</td>
<td>259</td>
<td>8</td>
<td>12619</td>
</tr>
<tr>
<td>Electron Microscopy</td>
<td>2316</td>
<td>31</td>
<td>805</td>
<td>0</td>
<td>3152</td>
</tr>
<tr>
<td>Other</td>
<td>257</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>280</td>
</tr>
<tr>
<td>Multi Method</td>
<td>134</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>142</td>
</tr>
<tr>
<td>Total</td>
<td>140968</td>
<td>3334</td>
<td>7623</td>
<td>30</td>
<td>151955</td>
</tr>
</tbody>
</table>

**Comparison to UniProt**

**Sequence-Structure-Gap:**
- Sequencing proteins is easy, fast and cheap
- Determining the structure is hard and costly
  ➞ Many sequences with unknown structure
### PDB - Coordinates

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>1234567890123456789012345678901234567890123456789012345678901234567890</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>601</td>
<td>N</td>
<td>LEU A</td>
<td>75</td>
<td>-17.070</td>
<td>-16.002</td>
<td>2.409</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>602</td>
<td>CA</td>
<td>LEU A</td>
<td>75</td>
<td>-16.343</td>
<td>-16.746</td>
<td>3.444</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>603</td>
<td>C</td>
<td>LEU A</td>
<td>75</td>
<td>-16.499</td>
<td>-18.263</td>
<td>3.300</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>604</td>
<td>O</td>
<td>LEU A</td>
<td>75</td>
<td>-16.645</td>
<td>-18.789</td>
<td>2.195</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>605</td>
<td>CB</td>
<td>LEU A</td>
<td>75</td>
<td>-16.776</td>
<td>-16.283</td>
<td>4.844</td>
<td>1.00</td>
</tr>
<tr>
<td>TER</td>
<td>606</td>
<td></td>
<td>LEU A</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATOM</td>
<td>1185</td>
<td>O</td>
<td>LEU B</td>
<td>75</td>
<td>26.292</td>
<td>-4.310</td>
<td>16.940</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>1186</td>
<td>CB</td>
<td>LEU B</td>
<td>75</td>
<td>23.881</td>
<td>-1.551</td>
<td>16.797</td>
<td>1.00</td>
</tr>
<tr>
<td>TER</td>
<td>1187</td>
<td></td>
<td>LEU B</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HETATM</td>
<td>1188</td>
<td>H2</td>
<td>SRT A1076</td>
<td></td>
<td>-17.263</td>
<td>11.260</td>
<td>28.634</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1189</td>
<td>HA</td>
<td>SRT A1076</td>
<td></td>
<td>-19.347</td>
<td>11.519</td>
<td>28.341</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1190</td>
<td>H3</td>
<td>SRT A1076</td>
<td></td>
<td>-17.157</td>
<td>14.303</td>
<td>28.677</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1191</td>
<td>HB</td>
<td>SRT A1076</td>
<td></td>
<td>-15.110</td>
<td>13.610</td>
<td>28.816</td>
<td>1.00</td>
</tr>
<tr>
<td>HETATM</td>
<td>1192</td>
<td>O1</td>
<td>SRT A1076</td>
<td></td>
<td>-17.028</td>
<td>11.281</td>
<td>31.131</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>295</td>
<td>HB2</td>
<td>ALA A</td>
<td>18</td>
<td>4.601</td>
<td>-9.393</td>
<td>7.275</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>296</td>
<td>HB3</td>
<td>ALA A</td>
<td>18</td>
<td>3.340</td>
<td>-9.147</td>
<td>6.043</td>
<td>1.00</td>
</tr>
<tr>
<td>TER</td>
<td>297</td>
<td></td>
<td>ALA A</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ENDMDL
UniProtKB: General repository for sequence information

SwissProt: Manually curated sequence information [500K]

Many (nearly) identical sequences in UniProtKB -> Clustering using PIDE required
  - UniRef100 [189M], UniRef90 [95M], UniRef50 [35M]

PDB: General repository for structure information [151K]

PDB has less entries due to higher experimental complexity (Sequence-Structure-Gap)
EXERCISE
Bio.PDB – Resources:

- API: http://biopython.org/DIST/docs/api/Bio.PDB-module.html
- Tutorial (short): http://biopython.org/wiki/The_Biopython_Structural_Bioinformatics_FAQ
- Tutorial (long): http://biopython.org/DIST/docs/tutorial/Tutorial.html#htoc149

Structure → Model → Chain → Residue → Atom

Adapted from smcra.eps ©1999-2019, The Biopython Contributors All rights reserved.
Thank you!

QUESTIONS?