Protein Structure Prediction
- Genome sequence - sequence of bases - DNA
  - well annotated functional parts of the genome - cDNAs
  - cDNAs could translate into one or more proteins - part of post-translational modification
  - proteins carry out the functions in the cell
  - function of a protein - depends critically on its 3-dimensional structure (tertiary or native structure)
- Proteomics - derive structure of every protein coded by the genome, understand their co-operative properties and function
- Proteins - globular proteins or soluble proteins
  - membrane bound proteins

Proteomics - Membrane and Globular
- Globular proteins - can be crystallized and structure studied by X-ray crystallography, NMR techniques, mass spectrometry, electron microscopy
- Membrane proteins - are in the native form only when bound to membranes - lipids
  - difficult to crystallize in active form
  - crystal structures solved are few - normally the soluble loops
- 1/3rd of the genome are membrane bound proteins
- range from 1 to 12 membrane bound domains

Different types of database
- Protein Data Bank (PDB)
- SCOP - Structural Classification of Proteins
  - http://scop.mrc-lmb.cam.ac.uk/scop/
- FSSP - Fold classification based on Structure-Structure alignment of Proteins
  - http://jura.emb.ac.uk:8765/~holm
- CATH - Class, Architecture, Topology and Homologous superfamily
  - http://www.biochem.ucl.ac.uk/bsm/cath

Structural classification of Proteins
- Family: Proteins clustered together into families are evolutionarily related with pairwise residue identities between the proteins are 30% and greater. Some cases with low identity could have similar folds and evolution
- Superfamily: Probable common evolutionary origin - Proteins that have low sequence identities, but with structural and functional features suggest that a common evolutionary origin.
- Fold: Major structural similarity. Similar in sec structure elements with same topological connections. Proteins with same fold may not have common evolutionary origin
SCOP classification
structural and evolutionary similarities between proteins where structures is known

<table>
<thead>
<tr>
<th>class</th>
<th>No. of folds</th>
<th>No. superfamilies</th>
<th>No. families</th>
</tr>
</thead>
<tbody>
<tr>
<td>α proteins</td>
<td>151</td>
<td>257</td>
<td>409</td>
</tr>
<tr>
<td>β proteins</td>
<td>111</td>
<td>213</td>
<td>362</td>
</tr>
<tr>
<td>αβ proteins</td>
<td>212</td>
<td>308</td>
<td>488</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>39</td>
<td>39</td>
<td>52</td>
</tr>
<tr>
<td>Membrane and cell surface</td>
<td>12</td>
<td>19</td>
<td>34</td>
</tr>
<tr>
<td>Small proteins</td>
<td>59</td>
<td>84</td>
<td>128</td>
</tr>
<tr>
<td>Total</td>
<td>584</td>
<td>920</td>
<td>1473</td>
</tr>
</tbody>
</table>

Structure Prediction for Globular Proteins
Two types of methods

Homology Model-comparative model based on known structures and sequences in various Database

Ab initio Model-based on physical and chemical principles

Homology Model
• 500,000 sequences are known - homology modeling - structures of at least 150,000 proteins can be derived
• Widely used method
  – Comparative method
  – can reliably predict structures of proteins with high similarity
  – due to availability of structures (~18,000) in Protein Data Bank and various DB
• 3-D structures of proteins in a family are more conserved than their sequences.
• similarity at sequence level shows a definite similarity in structure.
• 1/3rd of all sequences are related to known protein structures
Step 1 in Homology Modeling - Fold Identification

Aim: To find a template or templates structures from protein database

- Improved Multiple sequence alignment methods
  - PSIBLAST, CLUSTAL

Sequence-Sequence Alignment Methods

- Pairwise Sequence alignment methods
  - BLAST, FASTA, WU-BLAST, SSEARCH - available on www
  - compares target sequence with sequences in DB pairwise alignment
  - scans the sequences for words (three character in length)
  - any statistically significant alignment would have a high scoring pair of words - hit
  - counts how many such hits are present
  - finds good homologs (for >30% sequence identities)

Multiple Sequence Alignments

- multiple sequence alignment methods - CLUSTALW
  - dynamic programming procedure
  - alignment between two sequences assign a score based on the number of similarities and gaps in sequences
  - position specific scoring matrix called a profile is constructed
  - profiles for query sequence and its homologues are compared to profiles of structures in PDB.
  - differs in scoring methods for gap penalties
  - identifies distant homologues - ideal for template generation

ACKNOWLEDGEMENT

The material presented here has been taken from the following source

Prediction of Protein Structure
Lecture 2

http://www.wag.caltech.edu/home/ch121/
by Vaidehi Nagarajan