Molecular Structures
Molecular structures
Why is it important?

Answers to scientific questions such as:

What does the structure of protein ‘X’ look like?

Can we predict the binding of molecule ‘X’ to ‘Y’?

Does molecule ‘Z’ has the potential to become a good drug candidate?

Can we find a promising subset of drug candidates from a huge database containing millions of compounds?
Biological background
What is a gene?

A *gene* is a stretch of DNA which influences the organism by encoding a *protein* or *structural* or *functional* RNA. The human genome contains about 25,000 genes …

1953: The double-helix structure is discovered, A-T and C-G basepairs

1960: The genetic code is cracked (codons)

2001: The human genome is sequenced, total size is roughly $3 \times 10^9$ bp
Cells and chromosomes

The human body has about $10^{14}$ cells

Each cell has 46 chromosomes, DNA molecules, which store genetic information
**DNA - a chemical encoding**

**Nucleotides**
- Adenine (A)
- Cytosine (C)
- Guanine (G)
- Thymine (T)

**Cytosine**

\[
\begin{align*}
\text{C} & \quad \text{NH}_2 \\
\text{H} & \quad \text{C} \\
\text{C} & \quad \text{N} \\
\text{N} & \quad \text{C} \\
\text{H} & \quad \text{C} \\
\text{C} & \quad \text{O} \\
\end{align*}
\]
DNA, RNA, and Proteins

Before a gene comes into use, its coding DNA is transcribed to RNA that in turn is translated into a protein ...
# Amino acid encoding

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU Phe</td>
<td>UCU Ser</td>
<td>UAU Tyr</td>
<td>UGU Cys</td>
</tr>
<tr>
<td></td>
<td>UUC UCC UAC UGG</td>
<td>UCC UAC UAG Amb UGGA Trp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>CUU Leu</td>
<td>CCU Pro</td>
<td>CAU His</td>
<td>CGU Arg</td>
</tr>
<tr>
<td></td>
<td>CUC CCC CAC CGC</td>
<td>CCA CAG CGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>CUA</td>
<td>CCA CAA G1n CGA</td>
<td>CCG CAG CGG</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>AUG Met</td>
<td>ACA AAA Lys AGA Arg</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUG Met</td>
<td>ACA AAA Lys AGA Arg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>AUU Ile</td>
<td>AUU Thr</td>
<td>AAU Asn</td>
<td>AGU Ser</td>
</tr>
<tr>
<td></td>
<td>AUC ACC AAC AGC</td>
<td>ACA AAA Lys AGA Arg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>AUA ACA AAA Lys AGA Arg</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUG Met</td>
<td>ACG AAG AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>AUG Met</td>
<td>ACG AAG AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GGU Val</td>
<td>GCU Ala</td>
<td>GAA Glu</td>
<td>GGA</td>
</tr>
<tr>
<td></td>
<td>GUC GCC GAC GGC</td>
<td>GCA GAA Glu GGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GUA GCA GAA Glu GGA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUG GCG GAG GGG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

First base  

Third base
What are proteins?

Proteins are large molecules that have many important biological functions. They can act as both structural components or active agents.

Examples

**Structural components:** proteins that constitute hair, skin, and muscles

**Active agents:** enzymes that catalyze most intracellular reactions, and transport proteins such as hemoglobin that carries oxygen to our tissues.
Protein building blocks

Amino acids
Consists of a central carbon atom $C^\alpha$ which is bonded to an amino group and a carboxyl group and a side-chain
The backbone in proteins

The backbone is the sequence of amino groups, $C_\alpha$, and carboxyl groups.

Proteins differ in the number of amino acids linked together, and the sequential order in which these amino acids occur, and the manner in which they fold based on the local environmental condition.
So what do they look like?

ARNDAAQMFQPQSSTWYCTCPPFCATACGT
Level of representation

A gene reveals only the blueprint of a protein…

Knowing the structure of a protein is an important step towards understanding its functionality.

The structure of a biomolecule is classified in structural levels…

Primary

Secondary

Tertiary

AAUCUGC...

Met Asp Phe... ααααβββααβ...

The tertiary structure is believed to be encoded in the primary structure.
Representing biological structures

Simple folding model, hydrophobic/polar amino-acid types (H,P) or (0,1)

Protein backbone and side-chains, angles (phi, psi, CB, rotamers)

Full atom models, atomic coordinates (x,y,z)
Available structural data

Protein Data Bank (PDB) – a database of 43,238 protein structures (November 2007), http://www.pdb.org

Structures are obtained by time consuming experimental methods: X-Ray Crystallography (37,105), NMR (5,941)
PDB entry: Crambin

<table>
<thead>
<tr>
<th>HEADER</th>
<th>PLANT SEED PROTEIN</th>
<th>30-APR-81</th>
<th>1CRN</th>
<th>1CRND</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPN</td>
<td>CRAMBIN</td>
<td></td>
<td></td>
<td></td>
<td>1CRN</td>
</tr>
<tr>
<td>SOURCE</td>
<td>ABYSSINIAN CABBAGE (CRAMBE ABYSSINICA) SEED</td>
<td>1CRN</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUTHOR</td>
<td>W.A.HENDRICKSON,M.M.TEETER</td>
<td>1CRN</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| ATOM | 1  N   THR     1      17.047  14.099   3.625  1.00 13.79 | 1CRN | 70   |
| ATOM | 2  CA  THR     1      16.967  12.784   4.338  1.00 10.80 | 1CRN | 71   |
| ATOM | 3  C   THR     1      15.685  12.755   5.133  1.00  9.19 | 1CRN | 72   |
| ATOM | 4  O   THR     1      15.268  13.825   5.594  1.00  9.85 | 1CRN | 73   |
| ATOM | 5  CB  THR     1      18.170  12.703   5.337  1.00 13.02 | 1CRN | 74   |
| ATOM | 6  OG1 THR     1      19.334  12.829   4.463  1.00 15.06 | 1CRN | 75   |
| ATOM | 7  CG2 THR     1      18.150  11.546   6.304  1.00 14.23 | 1CRN | 76   |
| ATOM | 8  N   THR     2      15.115  11.555   5.265  1.00  7.81 | 1CRN | 77   |
| ATOM | 9  CA  THR     2      13.856  11.469   6.066  1.00  8.31 | 1CRN | 78   |
| ATOM | 10 C   THR     2      14.164  10.785   7.379  1.00  5.80 | 1CRN | 79   |
| ATOM | 11 O   THR     2      14.993  9.862    7.443  1.00  6.94 | 1CRN | 80   |
| ATOM | 12 CB  THR     2      12.732 10.711   5.261  1.00 10.32 | 1CRN | 81   |
| ATOM | 13 OG1 THR     2      13.308  9.439    4.926  1.00 12.81 | 1CRN | 82   |
| ATOM | 14 CG2 THR     2      12.484 11.442   3.895  1.00 11.90 | 1CRN | 83   |
| ATOM | 15 N   CYS     3      13.488 11.241   8.417  1.00  5.24 | 1CRN | 84   |
| ATOM | 16 CA  CYS     3      13.660 10.707   9.787  1.00  5.39 | 1CRN | 85   |
| ATOM | 17 C   CYS     3      12.269 10.431  10.323  1.00  4.45 | 1CRN | 86   |
| ATOM | 18 O   CYS     3      11.393 11.308  10.185  1.00  6.54 | 1CRN | 87   |
| ATOM | 19 CB  CYS     3      14.368 11.748  10.691  1.00  5.99 | 1CRN | 88   |
| ATOM | 20 SG  CYS     3      15.885 12.426  10.016  1.00  7.01 | 1CRN | 89   |
| ATOM | 21 N   CYS     4      12.019  9.272  10.928  1.00  3.90 | 1CRN | 90   |
| ATOM | 22 CA  CYS     4      10.646  8.991  11.408  1.00  4.24 | 1CRN | 91   |

...
Example: Crambin

TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN
Example of problems in structural bioinformatics
Protein 2nd structure prediction

Predict the secondary structure $\alpha, \beta, \text{coil}$ of proteins from the primary sequence…

Alpha helix (1PHB)  Beta sheet (1CDY)
Protein 2nd structure prediction

Predict the secondary structure $\alpha, \beta, \text{coil}$ of proteins from the primary sequence…

http://bioinf.cs.ucl.ac.uk/psipred/

Alpha helix (1PHB)  

Beta sheet (1CDY)

JPred  
http://www.compbio.dundee.ac.uk/~www-jpred/

PSIPred  
http://bioinf.cs.ucl.ac.uk/psipred/
Side-chain prediction

Predict the side-chain angles given the backbone of the protein…
Side-chain prediction

Predict the side-chain angles given the backbone of the protein...

O. Eriksson, Y. Zhou and A. Elofsson. 
*Side-Chain Positioning as an Integer Programming Problem.* 
In Proceeding of WABI'01, LNCS 2149, 128-141, 2001.

J. Desmat et al. 
*The dead end elimination theorem and its use in side-chain positioning.* 
Molecular docking
Virtual screening

Ligand database

Molecular docking

Target Protein

Ligand docked into protein’s active site
RNA 2nd structure prediction
RNA 2nd structure prediction

Mfold

http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi
Protein Folding
The protein folding problem

Determine the 3D structure of a protein based on its primary (amino acid) sequence

ARNDAAQMFQPQSSTW
YCTCPPFCATACGT...
Protein folding remains one of the great unresolved problems of molecular biology!

Folding process: Transformation from inactive to active (native) state…

The native state is assumed to be the conformation with lowest free energy

Early experiments by Christian B. Anfinsen (1950s) showed that a protein refolds to its native state from its unfolded state without any intervention…

Some proteins require so-called chaperones in order to fold properly…
Why predict protein structures?

- X-Ray crystallography is expensive and time-consuming
- Some important proteins (membrane proteins) are difficult (or impossible) to crystallize so we only know few 3D structures of membrane proteins...
- Simulations may provide us with new knowledge about the physics (e.g. what drives the natural folding process?)
- Understand mis-folding which causes diseases…
Protein structure prediction

Numerous methods are available...

**Homology**, use structural elements from related/similar sequences with known structure as a starting point...

**Threading**, *thread* the sequence on other known structures. Choose the one who seems to fit the most...

**Simulation** *(ab initio)* of folding dynamics …

**Minimize free energy**, model legal folds, assign an energy (score) to each legal fold, find the legal fold with minimum free energy …
Protein structure prediction

Folding@home
http://folding.stanford.edu

foldit
http://fold.it/portal
Lattice Models
HP lattice models

Introduced by Ken A. Dill in 1985...

Predicting protein structures by minimizing free energy

Assumption: Formation of a **hydrophobic core** is a principal force in protein structure formation...

Sequence is \( S \in \{H, P\}^* \) modeling **hydrophilic/polar** (water loving) and **hydrophobic** (water hating) amino acids

Folding is an embedding in a 2D lattice, i.e. a self-avoiding walk along the grid
**HP lattice models**

Hydrophobic / hydrophilic amino acids

http://en.wikipedia.org/wiki/Amino_acid#Hydrophilic_and_hydrophobic_amino_acids

http://en.wikipedia.org/wiki/Hydropathy_index


Sequence is $S \in \{H, P\}^*$ modeling hydrophilic/polar (water loving) and hydrophobic (water hating) amino acids

Folding is an embedding in a 2D lattice, i.e. a self-avoiding walk along the grid
Energy function (scoring)

Goal: maximise the number of non-bonded H-H contacts

Very simple energy function, -1 for each direct contact (occupying neighboring non-diagonal lattice points) of non-bonded H-H amino acids

Energy score = -4
Observations

Using the simple energy function, compact low-energy conformations are generated with a hydrophobic core, since $H-H$ interactions are rewarded.

Thus, hydrophobic residues tend to be on the inside, while the hydrophilic residues are forced to the ’surface’.
Pros and cons

- Residues are represented by a single atom, what about side-chains? Bonds are not mimicking ’reality’

- The energies are very short range and electrostatic interactions (repulsion/attraction) are not considered...

- Parity problem, two residues must be an odd distance (of at least three) apart to be in contact with each other

- Unable to reveal the structure of any particular protein (illustrate general principles governing protein folding?)

- All conformations can be found for short sequences

- Easy to understand for non-biologists, easy to implement
The number of possible valid (i.e., self-avoiding) folds for a sequence of length L on a two-dimensional square lattice approaches (Guttmann et al, 1996):

\[ 2.638^n \]

Thus, the number of possible solutions is exponential in the length of the sequence being folded.

So if n=50 and evaluating 1 million folds takes 1 second, it will take about 36,722,560 years to find the optimal (lowest energy) fold…
How can we obtain good folds?

A simple lattice model…

however finding an *optimal* solution is NP-complete!

Hart and Istrail presented (in 1995) a 1/4-approximation algorithm for the 2D HP lattice model, i.e., the algorithm finds a fold of a given sequence $S$ in time $O(|S|)$ with

$$\text{Score} \geq \frac{1}{4} \cdot \text{OPT}(S)$$
Improvements? (Storm, Lyngsø)

The **U-Fold** removes the parity restriction…
However, \( S = (10)^i 0(10)^i 00(10)^i (01)^i \) \( \Rightarrow U\text{-Fold} = \frac{1}{4} \cdot \text{OPT}(S) \)

The **S-Fold** allows multiple bends…
However, \( S = (10)^i (0^{2i+1})^4 (10)^i \) \( \Rightarrow S\text{-Fold} = \frac{1}{4} \cdot \text{OPT}(S) \)

The **C-Fold** allows two bends that fold the two ends towards each other…
However, \( C\text{-Fold}(S) \geq R \cdot \text{OPT}(S) \) where \( \frac{1}{4} \leq R \leq \frac{1}{3} \)
Further improvements

Alantha Newman presented in 2002 a linear-time 1/3 approximation algorithm.

Numerous heuristic algorithms have been applied to the 2D (and 3D) HP lattice model, such as genetic (evolutionary) algorithms, Monte Carlo (simulated annealing), tabu-search, ant-systems, etc.
Exercise
Exercise

Make a fold of the following sequences in the 2D HP-model with as many non-local H-H bonds as possible:

\( S_1 = \text{hphphhphpphphphpphph} \)

\( S_2 = \text{hhphpphphpphphpph} \)

Use ‘pen and paper’ or

http://www.cs.au.dk/~jn/HPCanvas/HPCanvas.html

Can you give a general upper bound on the optimal score (maximum number of non-local H-H bonds) for a HP-sequence?