

# Protein Structure Prediction

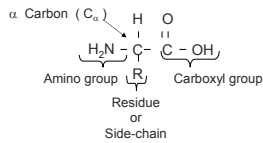
## Lecture 20

### Reading:

- Ref[JXZ, 2002]: chapters 16-18.
- D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*. (2001) 294(5):93-96.
- S. F. Altschul *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. (1997) 25:3389-3402.
- Y. Xu *et al.* An efficient computational method for globally optimal threading. *Journal of Computational Biology*. (1998) 5:597-614.
- Modeller-Program for comparative protein structure modelling. <http://salilab.org/modeller/modeller.html>

### Backgrounds

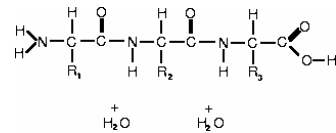
- **Protein** is a linear chain of amino acids (from a few hundreds to a few thousands), called Polypeptides.



General form of an amino acid

### A polypeptide backbone

- The carboxyl and amino groups of a pair of amino acids react through hydrolysis (removal of HO) to link and form a peptide bond.



### 20 amino acids

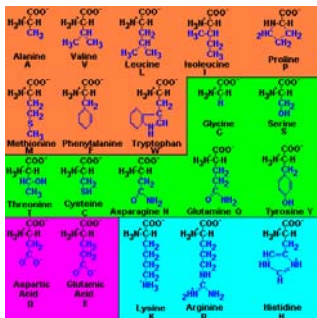


Photo credit: Robert J. Huskey

### Protein structure

- **Primary structure** is the sequence of amino acids that constitute the polypeptide chain.
- It folds to an intermediate structure known as the **secondary structure**, which consists of  $\alpha$  helix (spiral) or  $\beta$  sheets shapes.
- The **tertiary structure** is the 3 D structure of the protein

## A protein 3D structure



See [www.cmbi.kun.nl/gvteach/alg/infopages/proteins.shtml](http://www.cmbi.kun.nl/gvteach/alg/infopages/proteins.shtml)  
detailed atomic decomposition

## Structure prediction

- Big reasons for prediction:
  - Structure determines function.
  - Structure is better conserved than sequence.
  - Drug design.
  - A huge number of proteins (e.g. 300,000 human proteins).
  - A small number of structures been experimentally determined accurate.
  - Experimental structure determination is very expensive and slow.

## Structure prediction

- Roughly, 3 different approaches for prediction:
  - NMR spectroscopy and X ray crystallography.
  - Comparative modeling.
  - De novo structure prediction.

## NMR (Nuclear Magnetic Resonance) spectroscopy and X ray crystallography

- They are both experimental methods using knowledge from biology, physics, chemistry, biochemistry, and mathematics.
- NMR spectroscopy
  - a solution of protein is placed in a magnetic field and the effects of different radio frequencies on the resonance of different atoms in a protein are measured.
- X-ray crystallography
  - in this technique beams of x-rays are passed through a crystal of protein. Atoms in the protein crystal scatter the x-rays, which produce a diffraction pattern on a photographic film.

## Pros & Cons

- NMR
  - Size of protein limited (about 120 residues.)
  - Protein must be soluble
  - can provide the structure in near physiological condition
- X ray crystallography
  - Must be able to crystallize protein
  - Accurate
- Non of them is a high throughput technology

## Comparative modeling

- 4 steps:
  - find related known structures – templates
  - align the target sequence to the structure templates
  - build the backbone from the alignment
  - place the side-chains
- Templates can be found by
  - sequence comparison methods (PSI-Blast)
  - structure comparison methods
- Compute a sequence-structure alignment
  - using sequence alignment (homology search)
  - threading (next lecture)
- Produce all-atom model
  - cores, loops, side-chains
  - approximate positions of the conserved atoms

## De novo prediction

- No templates, predict from scratch
- Assumption: native structure is at the global free energy minimum
- Therefore, search limited to the conformational space with low global free energy.

## Accuracy comparison

Approach	req. seq. identity	accuracy	
NMR, X-ray	-	1.0 Å	
Comparative	sequence	> 50%	1.5 Å
	threading	> 30%	3.5 Å
	threading	< 30%	high error
De novo	insignificant	4-8 Å	

## Sequence homology search tool — PSI-BLAST

- Position-Specific-Iterated BLAST
  - encode the information about a whole protein family for the target sequence in a model to increase the chance of detecting remote homologies
- <http://www.ncbi.nlm.nih.gov/BLAST> (show example)
- One protein sequence  
>COG1024, CaiD, Enoyl-CoA hydratase/carnithine racemase  
mrcgdqhrf alwpgdkg lipatispyv varigeanal rfttsarlf aeeqrrigl  
hdvveaerld aaveaeikpy fstapaavaa skrlvhaiga sideavidmt ltrladtwet  
peaaegiaaf lnehqhgre gtkslnqkaa grgengdfrv persvlnvhe hrnaekrhq  
tgmaafgtds lltggadfrc scgylalaie srkcliits

## Protein function prediction

- Through structure prediction
- Indirectly through structure prediction
  - protein backbone structure prediction, via NMR or X-ray
  - threading to predict function
  - sequence similarity – homology search
  - key functional region identification
    - function motif recognition
    - structure motif recognition

## Threading or fold recognition

- When sequence comparison base methods fail.
- Fold recognition can be used to assign the correct fold to the target sequence.
- Threading or fold recognition is based on the theory that there may be only a limited number of distinct protein folds.