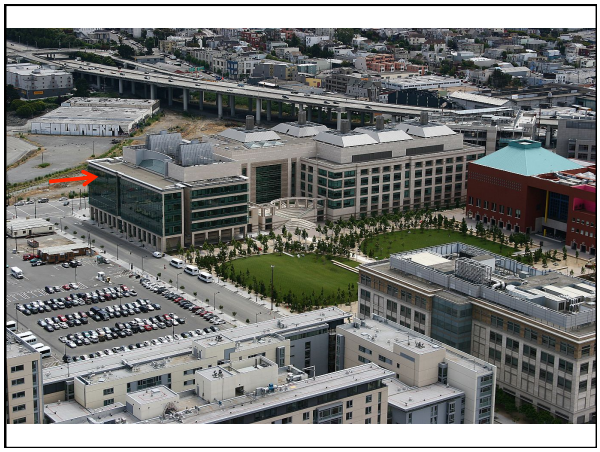
Structural Bioinformatics

Natalia Khuri
School of Pharmacy
University of California at San Francisco
San Francisco, CA
natalia.khuri@ucsf.edu
http://salilab.org/~nkhuri/Rabat2013

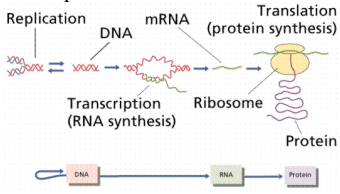


Topics to Cover

1) Introduction to Protein Structure
2) Classes of Protein Structure Prediction Methods
3) Quantitative Measures of Structural Differences
4) Critical Assessment of Protein Structure Prediction (CASP)
5) Template-Based modeling
   - Homology modeling
   - Threading
6) Free modeling
   - *De Novo*
   - *Ab Initio*

Central Dogma of Biology Revisited

The information for making proteins is stored in DNA. There is a process (transcription and translation) by which DNA is converted to protein. By understanding this process and how it is regulated we can make predictions and models of cells.



Proteins are the workhorses of the cell

- The function of a protein is affected by its structure and by its structural rearrangements during:
  - diffusion in the cell
  - ligand entry
  - ligand binding
  - protein–protein interactions, etc.

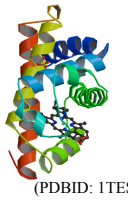
Proteins are responsible for many different functions in the living cell

- Enzymes
- Hormones
- Transport proteins
- Immunoglobulin or antibodies
- Structural proteins
- Motor proteins
- Receptors
- Signaling proteins
- Storage proteins

Tackling biology's big question

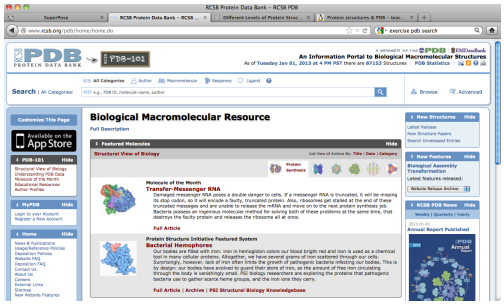
- How is all the necessary information specifying native protein structure contained in its primary amino acid sequence?

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIR
LFKSHPETLEKFDRFKHLKTEAEMKAEDLKKHG
VTVLTALGAILKKKGHHEAELKPLAQSHATKHK
IPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNK
ALELFRKDIAAKYKELGYQG



(PDBID: 1TES)

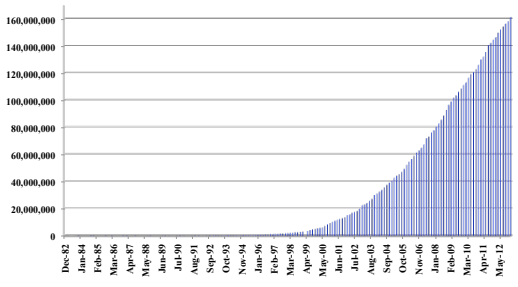
All published protein structures are deposited in the Protein Data Bank



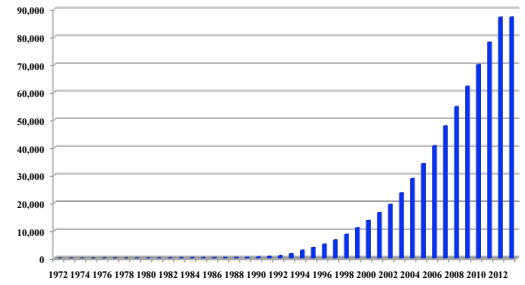
Protein Structure: Open questions

- Protein structure comparison
  - Assessing the degree of similarity between two protein structures
  - Given an unknown protein structure, can you identify the protein with most similar structure?
- www.procksi.org is a decision support system for protein structure comparison.

Sequence depositions into Genbank are growing exponentially



The rate of growth of solved protein structures is much slower



Protein Structure: Open questions

- Protein function prediction
  - The exact function of very few proteins is known
  - About 40% of proteins in the UniProt database are annotated as "hypothetical"
- Can we infer the function of an unknown protein
  - based on sequence alignment?
  - based on structure comparison?
- Can we predict for a protein, from its sequence or structure
  - its functionally important residues?
  - binding sites?

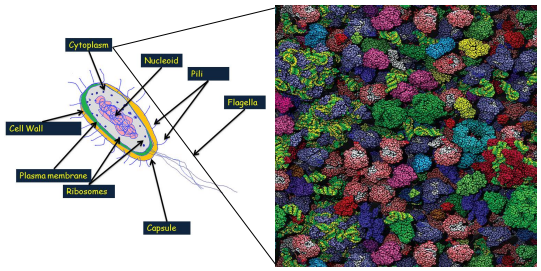
Protein Structure: Open questions

- Protein design
  - When modifying known proteins, or designing proteins de novo
    - How can we know that a mutation will not affect the structure/function?
    - How can we make sure that the protein will have the structure that we are interested in generating?

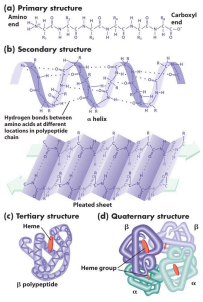
Protein Structure: Open Questions

- Prediction of interactions
  - How do viruses attach to cells?
  - How do proteins interact and mediate infection?
  - How do molecular machines organize themselves in healthy cells?
  - How do they differ in diseased cells?
- Can we predict from the primary sequences which proteins will interact with other proteins, DNA, or RNA?
- Can we predict how they will interact?

Protein Structure: Open Questions

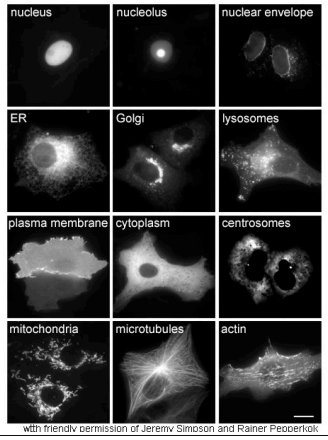


The structural organization of protein can be divided into four different levels

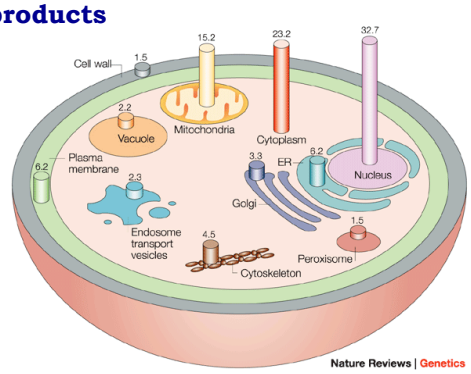


We can predict protein's location in the cell from primary sequence

- Membrane or trans-membrane proteins are located within cell membrane.
- Intracellular proteins are located within living cell and all functions are related with intercellular needs.
- External or secreted proteins function outside the cell they produced.
- Virus proteins are present only in viral organisms.



Relative frequencies of subcellular localization labels for yeast gene products



Case study: Where in the cell is my protein?

>unknown_cds Homo sapiens (human)
ATGCTGCGGAATCTGCTGGCTCTTCGTCAGATTGGGCA
GAGGACGATAAGCACTGCTTCCCGCAGGCATTTTAAAA
ATAAAGTTCCGGAGAAGCAAAAACTGTTCCAGGAGGAT
GATGAAATTCCACTGTATCTAAAGGGTGGGGGTAGCTGA
TGCCCTCCTGTATAGAGCCACCATGATTCTTACAGTTGG
TGGAACAGCATATGCCATATATGAGCTGGCTGTGGCTC
ATTTCCCAAGAAGCAGGAGTGA

Case Study Pipeline

- Get the sequence
- Translate
- Predict subcellular localization of the protein using WoLFPSORT (http://wolfpsort.org/)

WoLFPSORT predicts mitochondrial protein

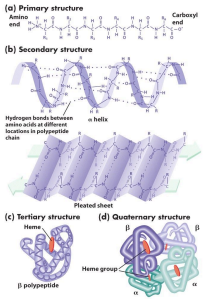
queryProtein WoLFPSORT prediction mito: 30.0

PSORT features and traditional PSORTII prediction

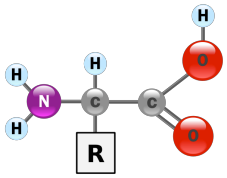
32 Nearest Neighbors

id	site	distance	identity	comments
COXJ_HUMAN	mito	0.0	100%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane.
COXJ_BOVIN	mito	26.0	84%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane.
COXJ_MOUSE	mito	48.7	81%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane.
COXJ_RAT	mito	49.7	81%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane.
GTA1_RABIT	cyto	52.1	16%	[Uniprot] SWISS-PROT45:Cytoplasmic.
NB4M_HUMAN	mito	56.1	15%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane; matrix side.
UCR6_PASHE	mito	56.6	12%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane.
SODM_HUMAN	mito	56.8	11%	[Uniprot] SWISS-PROT45:Mitochondrial matrix.
SODM_HORSE	mito	61.8	11%	[Uniprot] SWISS-PROT45:Mitochondrial matrix.
NIKM_HUMAN	mito	62.5	17%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane; matrix side.
RM12_MOUSE	mito	64.2	13%	[Uniprot] SWISS-PROT45:Mitochondrial.
SSB_DROME	mito	65.8	16%	[Uniprot] SWISS-PROT45:Mitochondrial. GO:0005739; C:mitochondrion; Evidence:IDA.
UCR6_ECHMU	mito	66.1	14%	[Uniprot] SWISS-PROT45:Mitochondrial inner membrane.
RM19_DROME	mito	68.4	10%	[Uniprot] SWISS-PROT45:Mitochondrial.
CCA1_HUMAN	mito	68.9	8%	[Uniprot] SWISS-PROT45:Mitochondrial. GO:0005739; C:mitochondrion; Evidence:IDA.
KAD3_BOVIN	mito	69.3	13%	[Uniprot] SWISS-PROT45:Mitochondrial matrix. GO:0005739; C:mitochondrion; Evidence:ISS.
SODM_BOVIN	mito	69.4	11%	[Uniprot] SWISS-PROT45:Mitochondrial matrix.
RM19_MOUSE	mito	69.6	8%	[Uniprot] SWISS-PROT45:Mitochondrial.

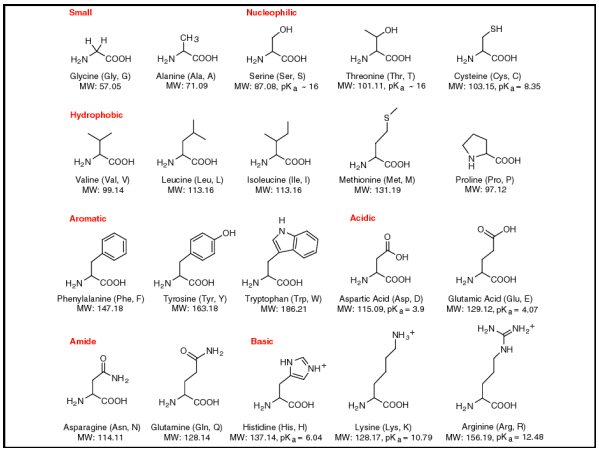
The structural organization of protein can be divided into four different levels



Proteins are large biological molecules consisting of one or more chains of amino acids

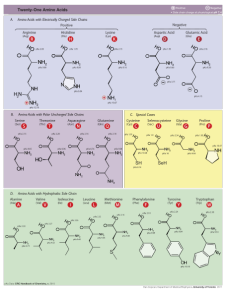


- All amino acids have a common part: the backbone
- Each amino acid type has a different side chain
- The Cα atom connects the backbone and the side chain
- The first carbon atom in the side chain is called Cβ (except for Gly)

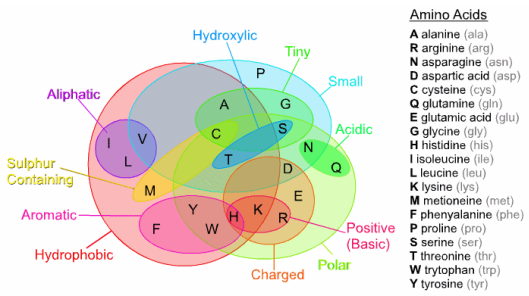


Classifications of Amino Acids

- Different amino acids have different properties
- These properties will affect the protein structure and function
- Hydrophobicity, for instance, is the main driving force (but not the only one) of the folding process



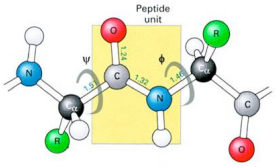
Overlapping Classes of Amino Acids



Amino Acids
A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

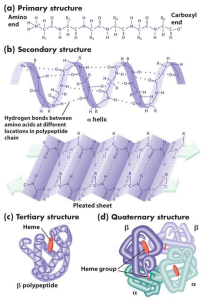
The polypeptide chain of proteins in joined together in a specific way

- Two dihedral angles (phi and psi) define the torsion of each amino acid in the chain.
- Phi is the angle of the Cα –N bond and psi is the angle of the Cα-C bond.



Hands-on Exercise 13 Questions 1-5

The structural organization of protein can be divided into four different levels

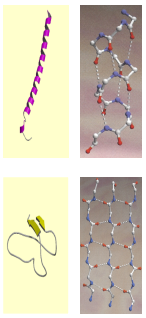


Proteins tend to fold into the lowest free energy conformation

- Proteins begin to fold while the peptide is still being translated.
  - Molecular chaperones work with other proteins to help fold newly synthesized proteins.
- Proteins bury most of its hydrophobic residues in an interior core.
- Folding begins with the formation of the secondary structures: α helices and β sheets.
- Much of the protein folding and modifications occurs in the endoplasmic reticulum and mitochondria.

Secondary Protein Structure

- There are two main kinds of secondary structure motifs:
  - α helices
  - β sheets
- Residues that do not fail in these two categories are said to be in coil state



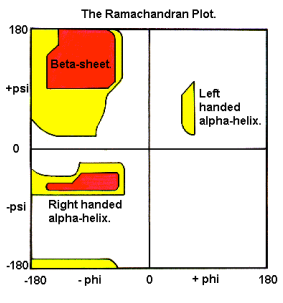
Residues form a loop of 3.6 residues/turn and 5.4Å wide

Residues lay flat in parallell strands. Called parallell sheets if all strands have the same N-to-C orientation, and antiparallell if adjacent strands have opposed orientations

Hands-on Exercise 13 Questions 6-9

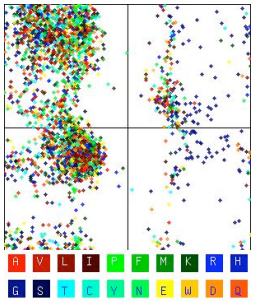
Ramachandran plots

- We saw that the backbone of a residue is characterised by two angles: psi and phi.
- Can they take any value?
- Fortunately not
- This effect was studied long ago by GN Ramachandran
- He proposed a diagram to visualize these angles (phi in the X axis, psi in the Y axi) of amino acid residues
- Different types of secondary structure are clustered in different regions of the diagram



Ramachandran plots

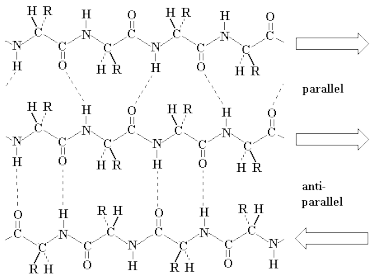
- You can create the Ramachandran plot for any protein in PDB at http://www.fos.su.se/~pdbdna/input_Raman.html
- At the right there is the plot for a set of 80 proteins



Amino Acid	3-Letter	1-Letter	Helical Propensity[24]
Alanine	Ala	A	0.0
Arginine	Arg	R	0.21
Asparagine	Asn	N	0.65
Aspartic acid	Asp	D	0.69
Cysteine	Cys	C	0.68
Glutamic acid	Glu	E	0.40
Glutamine	Gln	Q	0.39
Glycine	Gly	G	1
Histidine	His	H	0.61
Isoleucine	Ile	I	0.41
Leucine	Leu	L	0.21
Lysine	Lys	K	0.26
Methionine	Met	M	0.24
Phenylalanine	Phe	F	0.54
Proline	Pro	P	3.16
Serine	Ser	S	0.5
Threonine	Thr	T	0.66
Tryptophan	Trp	W	0.49
Tyrosine	Tyr	Y	0.53
Valine	Val	V	0.61

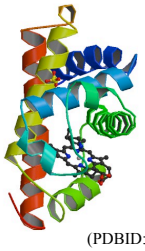
Higher value means lower propensity

Hydrogen Bonding in Beta Sheets

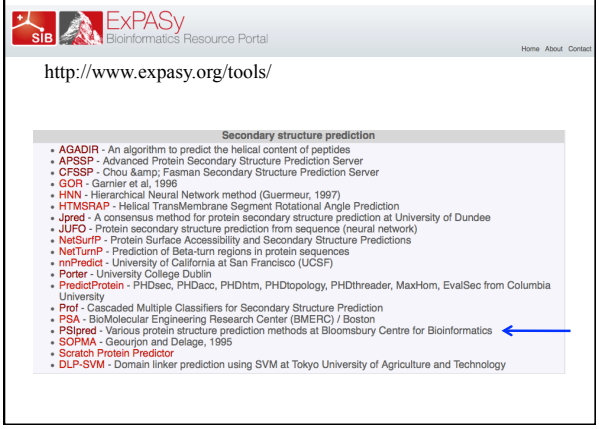


Case Study: predict secondary structure of the oxygen binding muscle protein

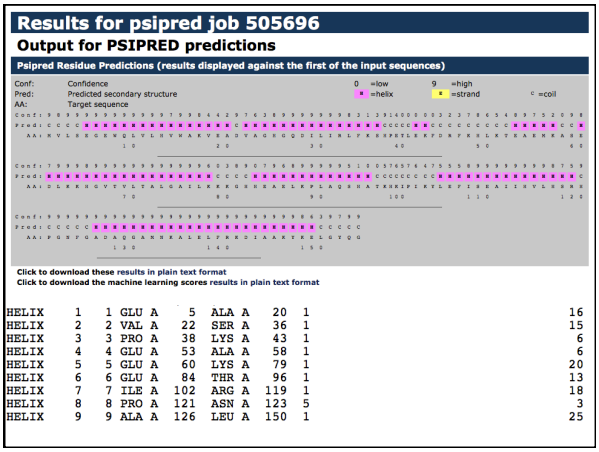
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKAE
DLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPG
NFGADAQGAMNKALELFRKDIAAKYKELGYQG

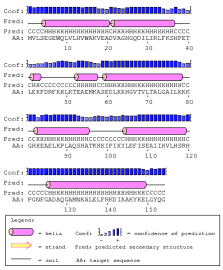


(PDBID: 1TES)



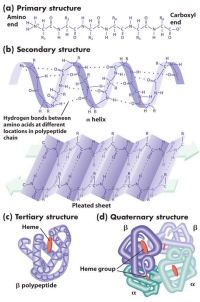
http://www.expasy.org/tools/





PSIPRED Results

The structural organization of protein can be divided into four different levels



Classifications of protein structure

- Several tertiary structure classification method exists, for instance, SCOP, CATH, and FSSP/ DDD.
- SCOP = Structural Classification Of Proteins http://scop.mrc-lmb.cam.ac.uk/scop/
  - uses a hierarchical system to catalog the proteins, according to evolutionary origin and structural similarity
  - the levels of the hierarchy are: class, fold, superfamily, family, protein and species

Main classes of SCOP (first level of hierarchy)

- All α proteins – proteins that have (almost) only α helices
- All β proteins – proteins that have (almost) only β sheets
- α+β proteins – proteins that have both α helices and (mostly) antiparallell strands, but segregated in different parts of the protein
- α/β proteins – proteins that have both α helices and (mostly) parallell strands, typically forming β+α+β units
- Multidomains proteins – proteins having two or more domains belonging to different classes
- Membrane and cell surface proteins
- Small proteins (metal ligans, heme and proteins with disulfide bridges
- Coiled coils proteins
- Low resolution protein structure
- Peptides
- Designed proteins

Case Study: Classify the following protein

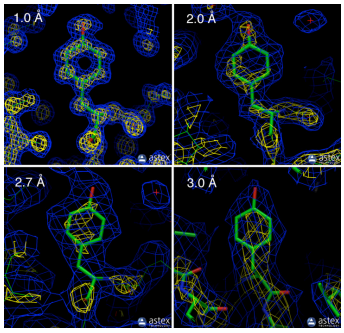
PDB ID: 5ULL

Case Study: Classify the following Protein

- SCOP classification of Flavodoxin from Clostridium beijerinckii
  - Class: α/β
  - Fold: Flavodoxin-like: 3 layers, α/β/α; parallel β-sheet of 5 strands
  - Superfamily: Flavoproteins
  - Family: Flavodoxin-related binds FMN
  - Protein: Flavodoxin
  - Species: Clostridium beijerinckii

Hands-on Exercise 14

Looking at Structures: Resolution



Hands-on Exercise 15 Questions 1-6

- Please, also complete Question 7 (register for Swiss-Model portal)

3D Structure Predictions

- The knowledge of protein structure and the dynamic behavior of the structure are critical for understanding how the protein performs its function.
- To fill the widening gap between the abundance of sequence availability and scarcity of experimentally resolved structure, computationally driven 3D structure prediction methodologies can be used to model structures of proteins, where no experimentally derived structure exists.

In Silico Methods for Determined Structures

- Even if we have an experimentally determined structure, *in silico* methods can be used to:
  - Model the effects of mutations
  - Predict the location of binding surfaces for other macromolecules and small-molecule effectors
  - Estimate binding energies
  - Predict local and non-local movements required for events such as binding, signaling and catalysis to occur.

Protein structure prediction methods are broadly divided into two groups

- Template-based methods
  - Homology (or comparative) modeling
  - Threading
- "Free-modeling" methods:
  - De Novo
  - Ab Initio

In template-based approaches a target protein structure is modeled using an experimentally solved template structure

- Homology (or comparative) modeling is based on the observation that evolutionarily related proteins (i.e., proteins that are related to one another in terms of amino-acid sequence) tend to have similar structures.
- Threading methods can be used to generate structures even if the target and template sequences are not evolutionarily related.

Free-modeling methods build 3D models using scoring (energy) functions

- ***De novo*** methods use knowledge-based scoring functions which can rely on experimentally derived understanding of protein folding, together with information on experimentally determined structures deposited in databases.
- ***Ab initio*** methods use scoring functions (force fields) based on first principles, without reference to solved structures.

CASP is a biennial competition in which protein structure prediction techniques are evaluated

Quantitative measures of structural differences

- The root mean square deviation (RMSD)
- The Global Distance Test Total Score (GDT-TS)
- The Z-score

The root mean square deviation

- Computes he average distance between the atoms of two structures when they are superimposed
- For reference:
  - 0.5 Å RMSD of alpha carbons occurs in independent determinations of the same protein
  - crystallographic models of proteins with about 50% sequence identity differ by about 1Å RMSD

The Global Distance Test Total Score

- Computes the largest set of amino acid residues' alpha carbon atoms in the model structure falling within a defined distance cutoff (1, 2, 3 and 8 Å) of their position in the experimental structure
- The GDT score has a value of 0-100
- A random superposition between unrelated structures will have a score of approximately 10-20

$$GDT = 100 \times \frac{(C1 + C2 + C3 + C4)}{4N}$$

- C1 = Count of number of residues superposed below threshold/4
- C2 = Count of number of residues superposed below threshold/2
- C3 = Count of number of residues superposed below a threshold
- C4 = Count of number of residues superposed below 2*threshold
- N = Total number of residues

The performance of computational modeling methods can be compared by using multiple targets

- The **z-score** is the distance, in standard deviations, between the observed alignment RMSD and the mean RMSD for random pairs of the same length, with the same or fewer gaps.
- **Z-scores** less than 2 are considered to lack statistical significance.

Threading

- In general, the structure of a protein is more conserved than its sequence.
- Thus, proteins can adopt the same fold even if there is no obvious sequence relationship between them.
- The best known threading programs are THREADER and RAPTOR

Threading: Sequence-to-Structure Alignment

- Threading methods assign (map/associate) the target sequence to templates with known folds, where each type of fold represents structures sharing closely similar architecture regardless of sequence.
- For every trial template, the optimal sequence-to-structure alignment is evaluated by a set of scoring functions based on physico-chemical parameters.

Limitations of Threading

- By their nature, **threading methods** are limited to a search of known folds and are unable to correctly predict the structure of the target if, in reality, it adopts a novel fold.
- Estimations of protein folds for water-soluble proteins: 400 to 10,000.
- Thus, 3D structures for many sequences cannot be predicted reliably by homology modeling or threading.

Template-Free Methods to the Rescue

- The 3D structures for many sequences cannot yet be predicted reliably by homology modeling or threading simply because suitable template structures do not presently exist.
- Because of this problem, and the additional difficulty that many modeled structures end up looking more like the template structure than they should, free modeling methods were developed.

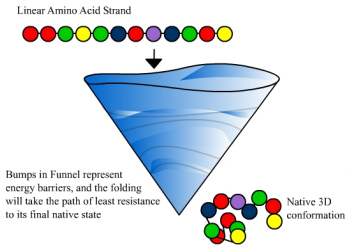
Thermodynamic Hypothesis of Anfinsen

- The template-free approach to modeling is guided by **Anfinsen's thermodynamic hypothesis**, which states that a protein's structure in a given environment is based on the sequence and corresponds to the global minimum of the potential energy of the system.

Levinthal's Paradox

- **Levinthal** formulated the paradox that the folding process cannot follow a random path to find the native conformation because it would take longer than the age of the universe.
- The concept of **folding funnels** was then developed in which protein folding follows an energy landscape, moving downhill to the global minimum

The Folding Funnel Concept



Hills and bumps represent the protein taking the path of least resistance when moving down the energy funnel. Funnels with the fewest bumps fold into their native forms faster since fewer energy boundaries exist.

Although these models are simplified attempts and do not account for misfoldings, they nonetheless prove accurate in the case of many proteins.

en.wikibooks.org/wiki/Structural_Biochemistry/Proteins/Protein_Folding

Free-Modeling Assumption

- Free-modeling methods draw on the following: 3D structures of target sequences are built using iterative processes in which the conformation of the folding structure is changed until a conformation with the lowest potential energy is found.
- Techniques used to search the energy landscape are combined with a scoring (energy) function used to estimate the value of potential energy.

Monte Carlo Simulation and Molecular Dynamics

- The parameters in the energy function may not be intended to faithfully reproduce energies, but rather promote computational tractability.
- The energy landscape is normally searched using Monte Carlo simulation or molecular dynamics approaches.

De Novo Methods

- Free-template methods can be divided into two groups, *ab initio* and *de novo*, based on their energy functions.
- ***De novo*** methods combine quantitative understanding of the physics of folding with knowledge about previously solved protein structures.
- Commonly used de novo methods include ROSETTA and I-TASSER.

Ab Initio Methods

- ***Ab initio*** methods use energy functions based on first principles of energy and atomic motion.
- The algorithms generally consist of a series of relatively simple terms to calculate the energies of structures.
- The computational demands are considerable.

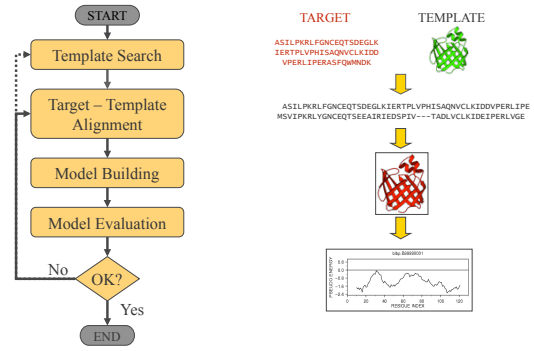
UNRES and ASTRO-FOLD

- Widely used methods include UNRES and ASTRO-FOLD .
- Despite attempts to reduce the computational costs, ***ab inito*** methods generally are limited to small molecules, including peptides, where they can be used to model the structures of fragments of sequence up to about 100 residues in length.

Homology Modeling

- **Homology modeling** is the most commonly used approach for modeling the 3D structures of proteins for which structures are not solved experimentally.
- The 3 steps of **homology modeling** are:
  1) Model building
  2) Refinement
  3) Evaluation.

Steps in Comparative Protein Structure Modeling



Model Building in Homology Modeling (I)

- The **model-building** step involves identifying the best template by aligning the sequence of the target with template sequences of proteins with known structures.
- A single template or multiple templates are chosen with the major (but not only) consideration being the extent of sequence identity/similarity between the target and the template.

Model Building in Homology Modeling (II)

- The chosen template acts as a "**pattern**" for the 3D coordinates of the target protein based on the conserved positions.
- In general, the sequences have to be at least 25% identical to be successfully employed for **homology modeling**.
- Many exceptions exist.
  - Example: rhodopsin was generated using the solved structure of bacteriorhodopsin as the template even though they do not share sequence similarity.

Model Building in Homology Modeling (III)

- Example of **pattern**: the hydrophobic periodicity of a helical transmembrane region from the solved structure of a membrane protein might be used as the basis of a "pattern" when aligning the target sequence.
- Stretches of amino acids in the target which do not fit the pattern of the template are often loop regions and are usually modeled using a database of fragment structures or by *ab initio* approaches.

Homology Modeling of Water-Soluble Proteins

- Most of the methods and assumptions used in homology modeling for water-soluble proteins are derived from the physico-chemical properties of water-soluble proteins.
- Properties of water-soluble proteins include experimentally determined low-resolution structures, as well as high-resolution structures.

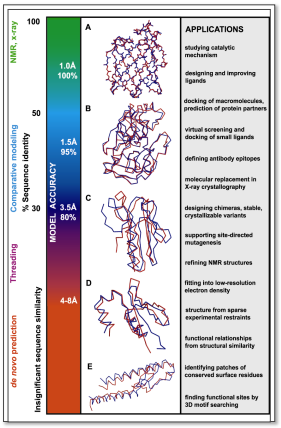
Homology Modeling of Transmembrane Proteins

- Homology model building of transmembrane proteins follow similar rules to water-soluble proteins.
- However, to increase the accuracy of the models, we might want to include information specific to membrane proteins such as:
  - the location of hydrophobic transmembrane regions
  - the incorporation of a lipid environment.

Hands-on Exercise 15 Questions 7-9

How do we choose templates?

- Homology modeling is based on choosing appropriate template models for the target under consideration.
- In the next slides, we study guidelines for choosing appropriate targets from databases.



D. Baker & A. Sali. *Science* 294, 93, 2001.

Target–Template Alignment

- Homology modeling programs use the target–template alignment as input but the alignments produced by the above search methods are usually sub-optimal and specialized alignment tools are often used to create a better alignment.
- This is crucial since the alignment of sequences is the most important step in the homology modeling procedure.

Modeler Expertise (I)

- Additional information is also often used to improve alignments, including:
  - The placement of hydrophobic regions
  - Secondary structure elements
  - Disulphide bonds.
- As is generally the case, predictions using intervention based on human expertise are mostly better than predictions from fully automated servers.

Modeler Expertise (II)

- Thus, the expertise of the modeler, drawing on biochemical information, such as:
  - function
  - family characteristics
  - mutagenesis observations
  - other information that may require manual intervention

  is frequently used to refine automated alignments.

Model Building

- There are different approaches for building the model:
  - Rigid-body assembly methods
  - Segment matching methods
  - Spatial restraint methods
  - Artificial evolution methods.
- As is often the case in bioinformatics, various studies have shown that no single model-building program is universally superior.

Refinement Phase in Homology Modeling

- In the **refinement phase**, the structures of loops and side chains are usually refined by **molecular dynamics** or energy minimization procedures.
- **Molecular dynamics** (**MD**) is a computer simulation of physical movements of atoms and molecules.
- The atoms and molecules are allowed to interact for a period of time, giving a view of the motion of the atoms. [wikipedia]

Model Refinement

- For homology modeling, refinement tends to focus on the correct orientation (rotamer position) of the side chains and the structure of the loops.
- Physical parameters and knowledge-based input are used to refine homology structures away from template structures.

Side-Chain Modeling

- **Side-chain modeling** is normally done by the homology modeling program but this is not always optimal.
- Therefore, models are often refined by standalone programs that use rotamer libraries derived from known structures or else molecular dynamics simulations are run for the entire model.

Handling Loops (I)

- The regions of the target sequence which do not have a corresponding homologous region in the template are often loops.
- Loops can play important structural roles, form ligand binding sites, etc.
- Loops are modeled using a database search or *de novo* conformational-search approach.
- In the database search approach, a database of loops, derived from known structures, is interrogated.

Handling Loops (II)

- Currently, loop searches are only carried out for loops of length up to 10 residues because the number of possible conformations for longer loops becomes very large.
- Conformations for shorter loops are well represented in databases such as wwPDB but recognizing them through appropriate scoring remains problematic.

Model Evaluation in Homology Modeling (I)

- In **model evaluation**, the refined models are evaluated for their agreement with information gathered from a number of sources, including generally known structural features and other experimental results.
- Each of these can take a variety of forms.
  - For example, structural features of the model can be evaluated by generating a Ramachandran plot and by calculating clash scores for steric overlap.

Model Evaluation in Homology Modeling (II)

- Experimental results can include:
  - independent measures of the location of disulphide bonds
  - secondary structure content as measured by circular dichroism or infrared spectroscopy
  - low-resolution structural data
  - information about conformation and function derived from mutagenesis studies.

Model Quality Assessment (I)

- After the model has been built, its stereochemistry can be checked using programs such as PROCHECK, WHATCHECK or MolProbity.
- These programs are not optimal because they check the capability of the homology modeling algorithm to build the structure rather than verifying the actual quality of the model.

Model Quality Assessment (II)

- Even though not optimal, these programs are still useful to detect errors in the modeling process and the models themselves (such as bad phi/psi angles or clashes).
- They can then use this information to choose the best model out of a set of predictions.

Model Quality Assessment (III)

- Another approach is to calculate a pseudo-energy profile of the model using programs as PROSA or Verify3D.
- These programs assign an energy value to each amino acid in the sequence derived from atomistic coordinates of correctly folded 3D structures.
  - Peaks in the profile indicate an unfavourable contribution to the potential energy of the structure and point to errors.

Topics Covered Today

1) Introduction to Protein Structure
2) Classes of Protein Structure Prediction Methods
3) Quantitative Measures of Structural Differences
4) Critical Assessment of Protein Structure Prediction (CASP)
5) Template-Based modeling
   - Homology modeling
   - Threading
6) Free modeling
   - *De Novo*
   - *Ab Initio*