Protein Structure Prediction

Notes from Chapter 5
of
Computational Biology an Application-Oriented View
by A.P. Gultyaev

1

Protein Structure Prediction

- Primary Structure
- Secondary Structure Prediction
- Tertiary Structure Prediction
- Prediction of Coiled Coil Domains
- Prediction of Transmembrane Segments

2

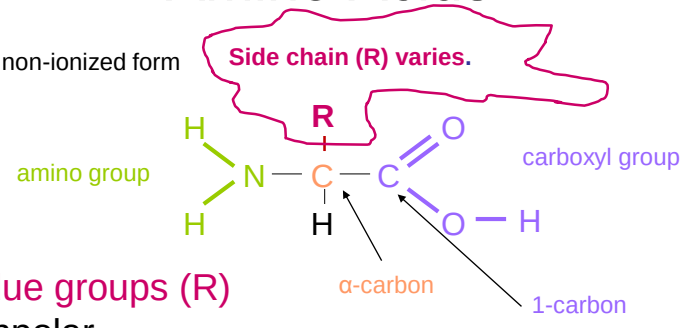
Why Structure Prediction?

Important Macromolecules in Living Organisms

- DNA
  - mostly long sequences that carry genetic information
- RNA:
  - mRNA carries the genetic information for protein synthesis
  - tRNA used to deliver amino acids to ribosomes
  - Keeper of genetic information of many viruses, etc.
  - Shorter sequences
  - **Mostly single stranded adopting 3d structures**
  - **The functional form of RNA sequences require a specific 3d structure**
- **Proteins**

Amino Acids

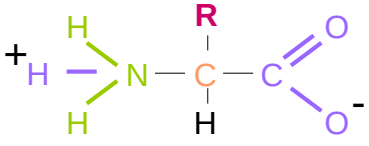
General non-ionized form



Residue groups (R)

- Nonpolar
  - hydrophobic
- Polar
  - Acidic
  - Uncharged
  - Basic

Ionized form



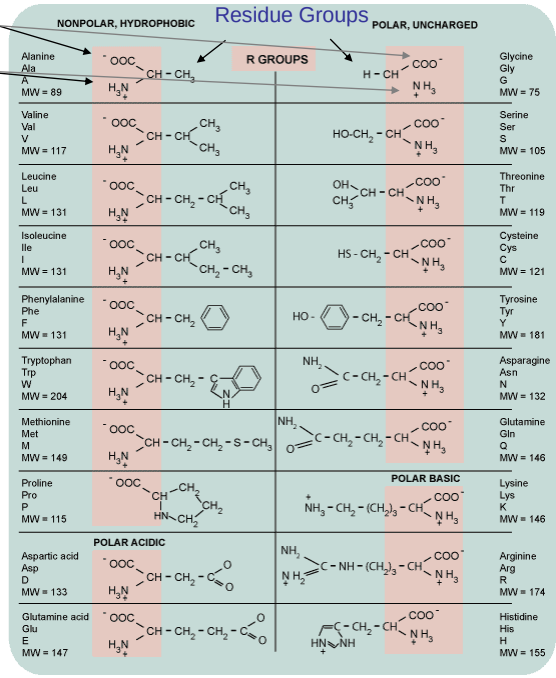
4

Amino acids

Carboxyl-group

Amino-group

Residue Groups

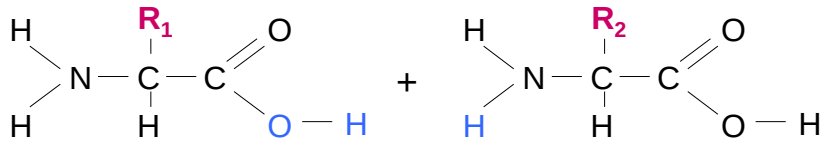


From: http://biotech.matcmadison.edu/resources/proteins/labManual/chapter_2.htm

Residue groups

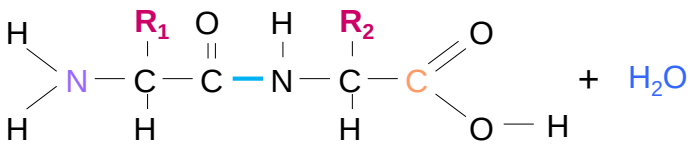
- Nonpolar
  - hydrophobic
- Polar
  - Acidic
  - Uncharged
  - Basic

Amino Acids: Peptide Bonds



Start:
N-terminus

End:
C-terminus



When the protein is translated from messenger RNA,
it is created from N-terminus to C-terminus.

6

Protein Primary Structure

- Chain of amino acids

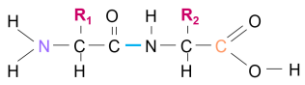
Example: Epidermal Growth Factor (EGF)

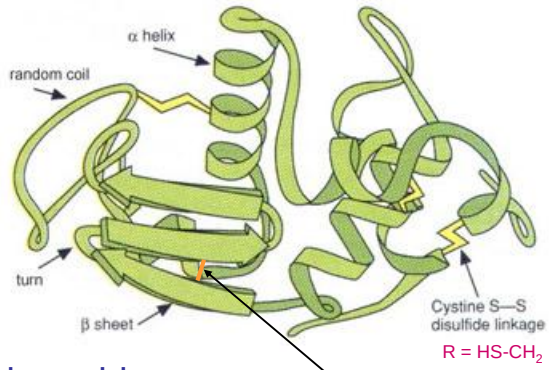
>1EGF:A|PDBID|CHAIN|SEQUENCE

NSYPGCPSSYDGYCLNGGVCMHIESLDSY
TCNCVIGYSGDRCQTRDLRWWELR

7

Protein Secondary Structure Prediction





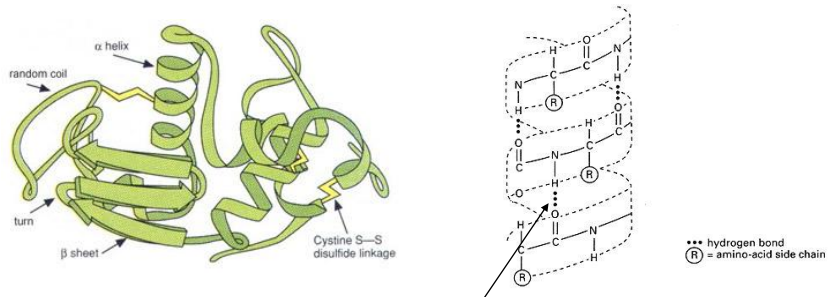
States of polypeptide residues

- Helix, strand, coil
- Helix, strand, coil, turn
- To as much as 8 states: helix, strand, coil, turn, …

H-bridges between NH-groups and nearby CO-groups in other direction

8

Protein Secondary Structure Prediction



Empirical (statistical) approaches extrapolating the statistics from known structures:

- Ala, Gln, Leu and Met are commonly found in α-helices
- Pro, Gly, Tyr and Ser usually **not** found in α-helices
- Pro is found to be a *'helix-breaker' (bulky ring prevents n/n+4 H-bonds formation)*

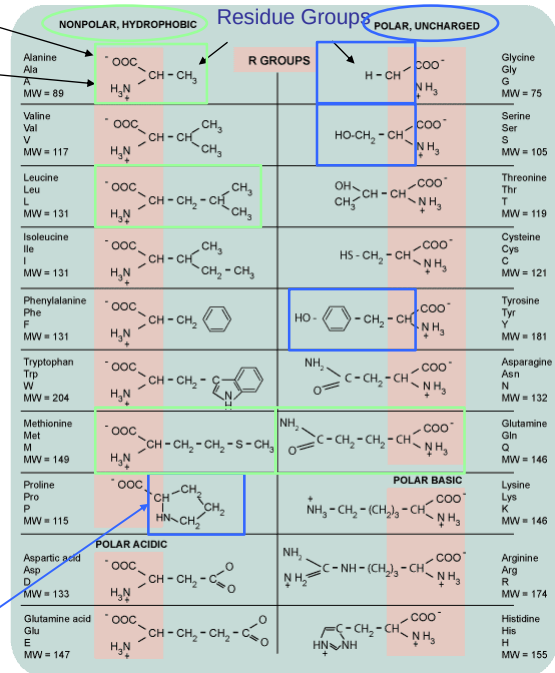
9

Amino acids

Carboxyl-group

Amino-group

Residue Groups



From: http://biotech.matcmadison.edu/resources/proteins/labManual/chapter_2.htm

Residue groups

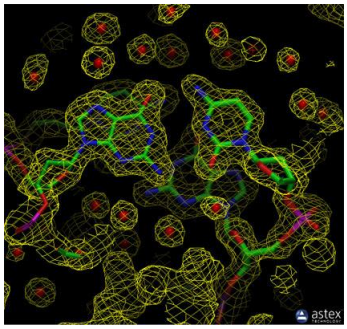
- Nonpolar
  - hydrophobic
- Polar
  - Acidic
  - Uncharged
  - Basic

Helix breaker

Protein Tertiary Structure Determination

X-ray crystallography

- Purified crystallized protein
- X-ray diffraction patterns and phase determination



⇒ electron densities

⇒ atom locations

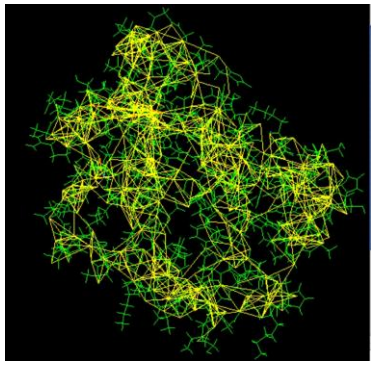
From: http://www.pdb.org

11

Protein Structure Determination

NMR spectroscopy

- Purified protein
- Strong magnetic field
- Analysis of resonances



⇒ List of nuclei that are close to each other

⇒ Characterize local conformation of atoms that are bonded together

⇒ List of restraints

⇒ Build model of the protein with the location of each atom

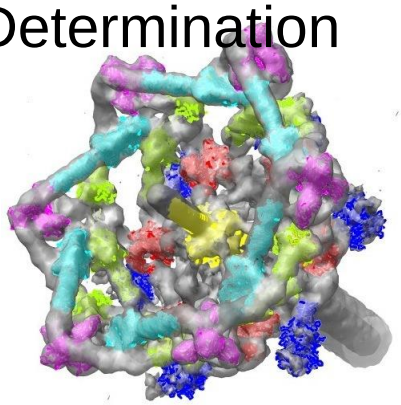
- Limited to small or medium sized proteins
- Produces information of the protein in a solution! As opposed to a crystallized form => study of flexible proteins possible

12

NMR = Nuclear Magnetic Resonance

From: http://www.pdb.org

Protein Structure Determination



Electron microscopy

- Beam of electrons
- Projections of the molecule
- Shape of the molecule
- 3D density maps
- Electron diffraction patterns
  - If proteins packed in membranes or small crystals
- 3D alignment and averaging for obtaining electron tomography (only if molecule is very symmetrical)

- In general not able to see each atom separate

13

From: http://www.pdb.org

Some Examples

- Tertiary structure and their functions

Epidermal Growth Factor

Keywords: cell signaling, cancer, ErbB, HER

Categories: cell signaling, cancer

Introduction

The cells in your body constantly communicate with each other, negotiating the transport and use of resources and deciding when to grow, when to rest, and when to die. Often, these messages are carried by small proteins, such as epidermal growth factor (EGF), shown here in red from PDB entry 1egf. **EGF is a message telling cells that they have permission to grow.** It is released by cells in areas of active growth, then is either picked up by the cell itself or by neighboring cells, stimulating their ability to divide. The message is received by a receptor on the cell surface, which binds to EGF and relays the message to signaling proteins inside the cell, ultimately mobilizing the processes needed for growth.

From:

June 2010 Molecule of the Month by David Goodsell Previous Features

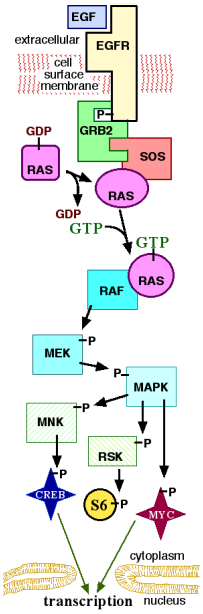
doi: 10.2210/rcsb_pdb/mom_2010_6

http://www.pdb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb126_1.html

Additional reading about EGF

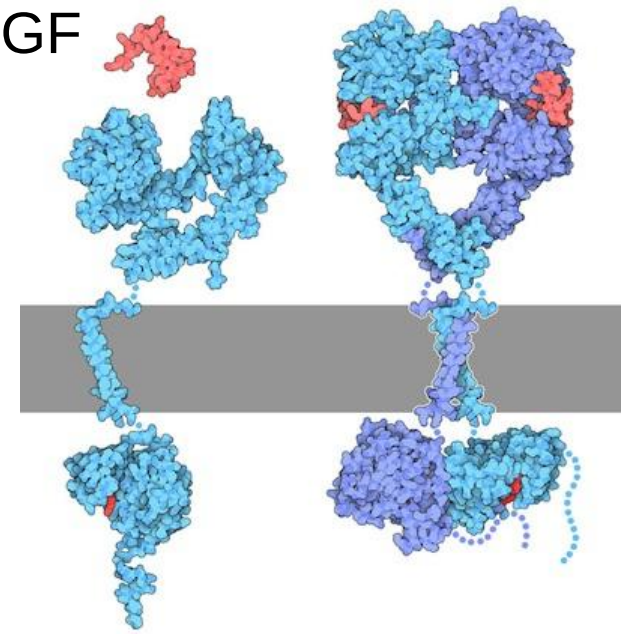
1. M. Lemmon (2009) Ligand-induced ErbB receptor dimerization. Experimental Cell Research 315, 638-648.
2. R. Bose and X. Zhang (2009) The ErbB kinase domain: structural perspectives into kinase activation and inhibition. Experimental Cell Research 315, 649-658.
3. K. M. Ferguson (2008) Structure-based view of epidermal growth factor receptor regulation. Annual Review of Biophysics 37, 353-373.

15



Picture from wikipedia

1EGF



Picture from www.pdb.org

Sodium-Potassium Pump

Introduction

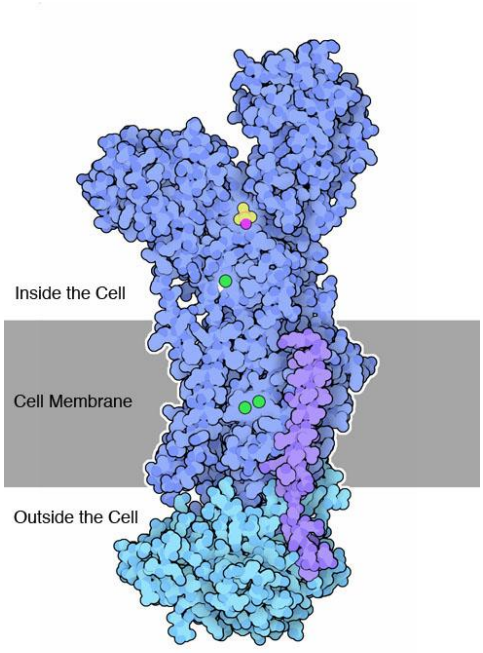
Our bodies use a lot of energy. ATP (adenosine triphosphate) is one of the major currencies of energy in our cells; it is continually used and rebuilt throughout the day. Amazingly, if you add up the amount of ATP that is built each day, it would roughly equal the weight of your entire body. This ATP is spent in many ways: **to power muscles, to make sure that enzymes perform the proper reactions, to heat your body.** The lion's share, however, goes to the protein pictured here: roughly a third of the ATP made by our cells is spent to power the sodium-potassium pump.

From:

- October 2009 Molecule of the Month by David Goodsell Previous Features doi: 10.2210/rcsb_pdb/mom_2009_10

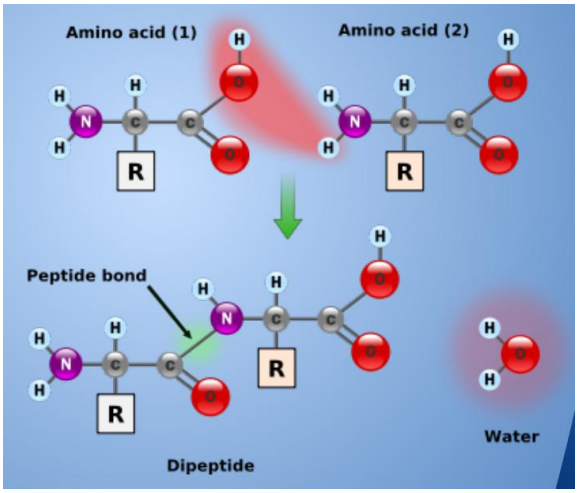
17

- 2ZXE Crystal structure of the sodium - potassium pump in the E2.2K+.Pi state
- 3B8E Crystal structure of the sodium-potassium pump

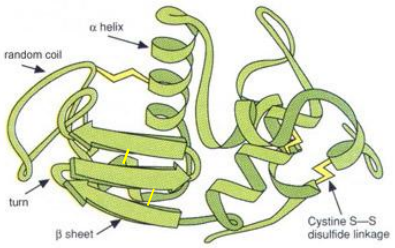


18

Secondary Structure Prediction



Protein Secondary Structure Prediction



States of polypeptide residues

- Helix, strand, coil
- Helix, strand, coil, turn
- To as much as 8 states: helix, strand, coil, turn, …

Empirical (statistical) approaches extrapolating the statistics from known structures:

- Ala, Gln, Leu and Met are commonly found in α-helices
- Pro, Gly, Tyr and Ser usually **not** found in α-helices
- Pro is found to be a *'helix-breaker' (bulky ring prevents n/n+4 H-bonds formation)*

20

Protein Secondary Structure Prediction Algorithms (1970 – 80)

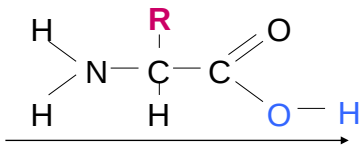
Chou and Fasman (1978)

- Moving average of values that indicate the probability of a residue type to adopt α-helix, β-sheet and turn
- Probabilities derived from frequency observation normalized by frequency expected by chance
- Together with heuristics for determining the ends of secondary structures

21

Protein Secondary Structure Prediction Algorithms (1970 – `80)

Garnier et al. (1978)



- Consider for each residue it's surrounding region: from 8 residues towards the N-terminal, and to 8 residues towards the C-terminal.
- Estimate the effect that these residues have on the possible structure (state) at their position.
- **Profile:** For each of the 20 amino acids, a profile (17 residues long) quantifies the contribution the respective residue makes towards the probabilities of other residues to be in one of four states:
  α-helix, β-sheet, turn and coil
- For each of the four states ***probability profiles*** are produced, and at any position the highest profile value predicts the structure.

22

Protein Secondary Structure Prediction Algorithms

Pattern Recognition approach:
hydrophobicity patterns.

- Hydrophobicity is an important driving effect for protein folding. => Segregation from water molecules.
- In Lim (1974) the "half-buried" α-helix has the following pattern:
  - the i^{th} residue pointing towards the core, should have hydrophobic residues at positions
    - *i*, *i*+3 and *i+4* or
    - *i, i-1* and *i*-4.
- Various algorithms use many different rules to recognize these kind of patterns.

23

Protein Secondary Structure Prediction Algorithms

1990s

- In the 1990s the so-called second-generation methods used different ways for calculating propensities in windows of 3 to 51 residues.
- However: prediction accuracy stalled at levels slightly above **60%** (Q3-score)

Q3 score: percentage of residues predicted correctly in one the three states: helix, strand, and other.

24

Protein Secondary Structure Prediction Algorithms

Next generation structure prediction based on multiple alignments:

- Homologous proteins should have similar structures.

For example:

Observation: from all naturally evolved proteins with more than 35% pair wise identical residues over more than 100 *aligned* residues have similar structures (Rost, 1999).

Also: improved structure prediction using neural networks:

input layer (<= sequence)

↓

hidden layers (learning) ← *training data (known structures)*

↓

output layer (=> structure prediction as probabilities)

25

Protein Secondary Structure Prediction Algorithms

PSIPRED (Jones, 1999):

a combination of multiple alignments with neural networks for secondary structure predictions

1. Apply PSI-BLAST to construct a profile (PSSM) corresponding to the query protein.
2. Use this profile as the input to a neural network.
3. The final output is a 3-state prediction (helix/strand/coil).

26

Protein Secondary Structure Prediction

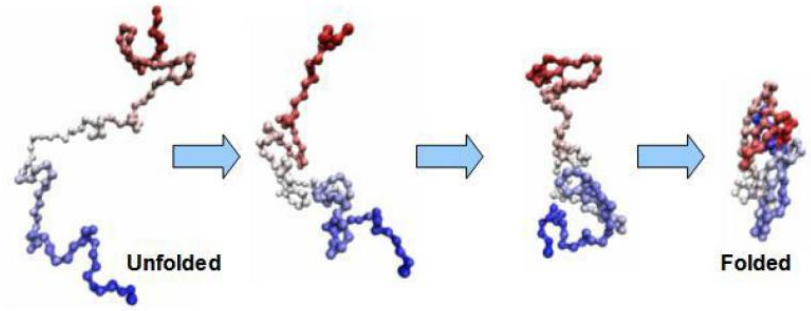
Other algorithms:

- LiveBench
- JPred
- BLAST-ERT-RICO
  - Q3 score on RS126 data set: 92.19%
  - uses multiple sequence alignment information [Lee et al, 2012]

Q3-score = percentage of correct 3-state predictions

JPred	84.5
PHD	75.5
PHDpsi	82.7
PROF_king	80.9
PROFsec	80.9
Prospect	70.9
PSIpred	70.9
SAM-T99sec	78.2
SSpro2	79.1

Tertiary Structure Prediction



Protein Tertiary Structure Prediction

Classified into three general strategies:

1. **Comparative (homology) modeling**.
   - Exploit the sequence homology between the target and known structures.
   - Obtain the most accurate structural model for the target, consistent with the known set.

2. **Fold recognition**: try to recognize a known fold in a domain within the target protein.
   - Use a so-called **threading** algorithm to detect common structures even in the absence of high sequence similarity.
   - The target protein sequence is threaded through templates from the structure database.
   - Alternative sequence structure alignments are scored using conformational energy calculations, based on statistics of known structures.
   - Highest scores are used to determine the locations of the folds.

3. ***Ab initio* methods**. Modeling of structures using potential energy calculations.

29

Protein Tertiary Structure Prediction: Comparative Homology Modeling

Homology modeling:

- sequence similarity between a target protein and at least one related protein with known structure (the templates) => 3D similarities
- extrapolate template structures to the target sequence.

(Fiser et al, 2001) Comparative Protein Structure Modeling:

Algorithm (outline):

1. Identify related structures
2. Select templates
3. Align target with templates
4. Build a model for the target (using information from template structures)
5. Evaluate the model
6. If model is not satisfactory, repeat the steps 2 to 5 or 3 to 5

30

Comparative Modeling (Step 1 and 2)

Identification of structures related to the target sequence

- usually done by searching the database of known protein structures (PDB) using the *target sequence* as the query.

Methods:

- Standard alignment procedures (e.g. FASTA or BLAST).
- Multiple sequence comparisons, for example PSI-BLAST (Position-Specific Iterated BLAST)
  - Improves the search sensitivity.
  - PSI-BLAST uses (Re)calculation of position-specific score matrices (PSSM)
- IMPALA (Integrating Matrix Profiles And Local Alignments)
  - uses a database of PSSMs and the *target sequence* as a query.

31

Template Search in PDB: PDB-BLAST (Step 1)

PDB-BLAST

- Builds a multiple sequence alignment using the *target sequence* as a query
- **And** constructs similar multiple alignments using all found potential templates (related sequences) as queries (each alignment is called a *sequence profile*).
- The final templates are found by
  - comparing the target sequence profile with each of the template sequence profiles
  - using a dynamic programming method and BLOSUM62

32

Selecting Templates (Step 2)

Template Quality:

- Increases with overall sequence similarity to the target.
- Decreases with the number and length of gaps in the alignment.

Furthermore:

- The quality of the experimentally determined structure is another important factor in template selection *(e.g. resolution of a crystallographic structure; the number of restraints per residue for an NMR structure; …).*

Note:

- Multiple templates may be aligned with different target domains.
- Often beneficial using different templates that are overall similar to the target sequence.

33

Protein Tertiary Structure Prediction: Comparative Homology Modeling

Homology modeling:

- sequence similarity between a target protein and at least one related protein with known structure (the templates) => 3D similarities
- extrapolate template structures to the target sequence.

(Fiser et al, 2001) Comparative Protein Structure Modeling:

Algorithm (outline):

1. Identify related structures
2. Select templates
3. Align target with templates
4. Build a model for the target (using information from template structures)
5. Evaluate the model
6. If model is not satisfactory, repeat the steps 2-5 or 3-5

34

Target Sequence Alignment with Template(-s) (Step 3)

A template search usually does not yield the optimal target-template alignment.

If sequence identity is high

- An accurate alignment can be calculated automatically using standard alignment algorithms.

If sequence identity is low

- The alignment may need manual intervention with inspection of gaps and misaligned residues.

Note:

The alignments can be also improved by including structural information from the template (automatically or manually).

35

Protein Tertiary Structure Prediction: Homology Modeling

Homology modeling:

- sequence similarity between a target protein and at least one related protein with known structure (the templates) => 3D similarities
- extrapolate template structures to the target sequence.

(Fiser et al, 2001) Comparative Protein Structure Modeling:

Algorithm (outline):

1. Identify related structures
2. Select templates
3. Align target with templates
4. Build a model for the target (using information from template structures)
5. Evaluate the model
6. If model is not satisfactory, repeat the steps 2-5 or 3-5

36

Model Building (Step 4):
(I) Modeling by Assembly of Rigid Bodies.

- *Rigid Bodies* are obtained from the aligned protein template structures.
- Dissect the protein structure into
  - conserved core regions
  - variable loops that connect them, and
  - side chains
- Build Model
  - Core regions of the target can be modeled using a superposition of templates.
  - Loops are generated by scanning a database of all protein structures to identify suitable structurally variable regions *(fit the core regions and have a compatible sequence).*

37

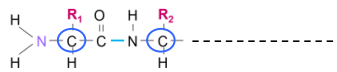
Model Building (Step 4):
(II) Modeling by segment matching or coordinate reconstruction.

Experimental finding:

Most hexa(6)-peptide segments of protein structures can be clustered into ~100 structurally different classes.

Comparative models can be constructed as follows:

- Use a subset of atomic positions from template structures as "guiding" positions.
- Identify and assemble short all-atom segments that fit these guiding positions.
- The guiding positions usually correspond to the C_α atoms of the segments conserved in the target-template alignment.



38

Model Building (Step 4):
(III) Modeling by Satisfaction of Spatial Restraints.

Algorithm (sketch):

- Use an alignment to templates to generate (many) constraints or restraints on the target structure.
  - Homology derived restraints by assuming that the *corresponding distances* between aligned residues in the template and in the target structures *are similar*.
  - Stereo-chemical restraints on *bond lengths*, *bond angles* and *dihedral angles (angles of bonds between N and C_α, and between C_α and C_1)*.
- Derive the final model by minimizing the violations of all the restraints.

39

Model Building (Step 4):
(IV) Loop Modeling.

Inserted or low similarity regions relative to the templates frequently correspond to surface loops.

Two main type of approaches for loop modeling:

1. Database Search Approaches:
   - Assume that similar loop conformations could be found among known structures.
2. Conformational Search Approaches:
   - Attempt to optimize the loop conformation.

40

Protein Tertiary Structure Prediction: Homology Modeling (Step 5 and 6).

Homology modeling:

- sequence similarity between a target protein and at least one related protein with known structure (the templates) => 3D similarities
- extrapolate template structures to the target sequence.

(Fiser et al, 2001) Comparative Protein Structure Modeling:

Algorithm (outline):

1. Identify related structures
2. Select templates
3. Align target with templates
4. Build a model for the target (using information from template structures)
5. Evaluate the model
6. If model is not satisfactory, repeat the steps 2-5 or 3-5

41

Protein Homology-Modeling Implementations

(Web-based) automated homology modeling systems for structure prediction for one or many sequences without human intervention.

SWISS-MODEL

- One of the first servers for protein structure predictions.
- SWISS-MODEL was initiated in 1993 and accessible via the ExPASy web server.
- See: *http://swissmodel.expasy.org*

42

Fold Recognition Methods

Fold Recognition Methods

If a target protein shows relatively low *sequence* similarity to known structures:

- recognize a known *fold* within the target by a search for an optimal sequence-to-structure compatibility.

Threading Algorithms:

- A target sequence is *threaded* through templates from the structure database.
- Alternative *sequence-structure alignments* are scored according to some measure of compatibility between the target sequence and the template structures.
- Scoring is done using *threading potentials*.

44

Fold Recognition: Threading Potentials (1/2)

Two major types of knowledge-based threading potentials:

1. Mean Force Potentials

- derived by applying the inverse Boltzmann equation to statistics of pairs of residues (a,b), located at various distances in the sequences and in the structures:

$$E - E^* = - RT \ln [f / f^*]$$, where

  - E is the statistical potential of the interaction at some state
  - f is the pairing frequency in this state
  - E* and f* are corresponding values for a reference state
  - R is the molal gas constant
  - T is the absolute temperature

45

Threading algorithms Remarks

Threading Algorithms

- More complex than sequence-sequence alignments.
- Scoring measure of compatibility between the target sequence and the template structures
- Approximations required.

For Example

Approximation 1: *Ungapped Threading*

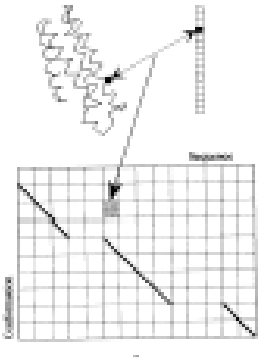
- The query sequence is mounted over an equally long part of a template fold.
- The total alignment score is computed as sum of pairwise potentials for all query residues.
- Mostly not used for real predictions, but rather for testing and adjusting the energy functions.

46

Threading Algorithms
Gaps in Sequence-Structure Alignments

Note: A score for a given residue in a query sequence, assumed to be aligned to a residue in a template structure, depends on:

1. The type of the two residues (as in sequence-sequence alignments).
2. The gaps that may be introduced at other alignment positions.

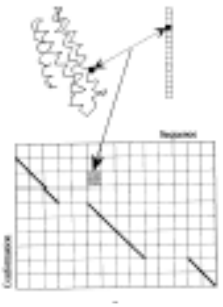


47

Threading Algorithms
Gaps in Sequence-Structure Alignments

Approximation 2: Frozen Approximation (Sippl, 1993)

- A comparison N×M matrix (N #residues in template and M #residues in the query) is calculated by replacing the amino acids in the template structure with amino acids from the target sequence one at a time.
- The rest of the structure is kept intact, and it is assumed that the field created by the native protein will also favor the correct replacement.
- Sequence-sequence alignment: the scores in the comparison matrix are used for calculating a dynamic programming matrix leading to the final alignment.
- Very crude approximation, that does not solve the full threading problem.
- But: efficient



48

Combined Scoring in Fold Recognition Methods

Often in fold recognition methods several scores are produced, related to different aspects of the sequence-structure alignment, some of the most important:

- initial sequence profile alignment score
- number of aligned residues
- length of target sequence
- length of template sequence
- pairwise energy sum
- solvation energy sum

Each of these scores separately may be not sufficient, therefore:

1. A neural network can be very effective to reduce a complexity to a single output value.
2. Performance of many programs may be improved by human intervention.
3. Modern fully automated methods are approaching and in some cases even challenging the accuracy of human-curated predictions.

49

Ab Initio Protein Structure Prediction

50

Ab Initio Protein Structure Prediction

Ab initio, or *de novo* approaches:

- Predict a protein structure and folding mechanism from knowledge of its amino acid sequence only
- Often used to denote a method or algorithm that is entirely based on physico-chemical interactions
- But the most successful *ab initio* methods utilize information from the sequence and structural databases
- Basic idea: search for the native state which is presumably in the minimum energy conformation

51

Protein Lattice Models

Lattice Models

- Represent proteins as simple monomer units connected by bonds (ignoring side chain dimensions).
- Simplified models are very useful for understanding protein folding process.

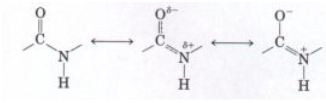
Mostly the HP Lattice Model is used:

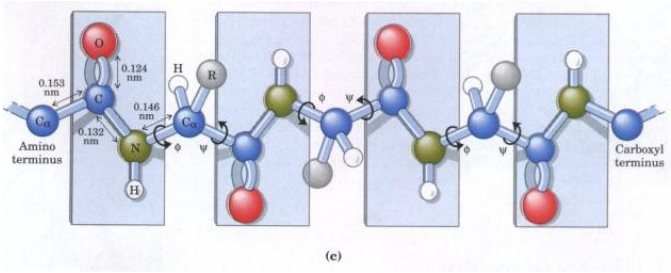
- Proteins are sequences of H (hydrophobic) and P (polar) monomers.
- Monomers are connected by bonds with bond angles taking a few discrete values (rather than a continuum).
- Different conformations conform to lattices in two or three dimensions.
  - In 3D, each configuration is a *self-avoiding walk* on the lattice.
- Contacts between H (hydrophobic) monomers are favorable:
  - The energy is determined by the number of H-H contacts h: $E = - \varepsilon h$, where ε is a positive constant.

52

Dihedral angles

- Bond length (C-N = 0.132 nm)
- Stretch angle (120˚)
- Torsion angle (between -180˚ and +180˚)
- Φ (phi), ψ (psi)
- ω (omega = 180˚) →



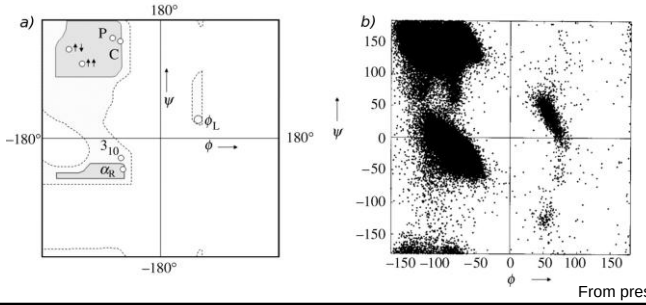


From presentation: J. Neuteboom

Dihedral angles (2)

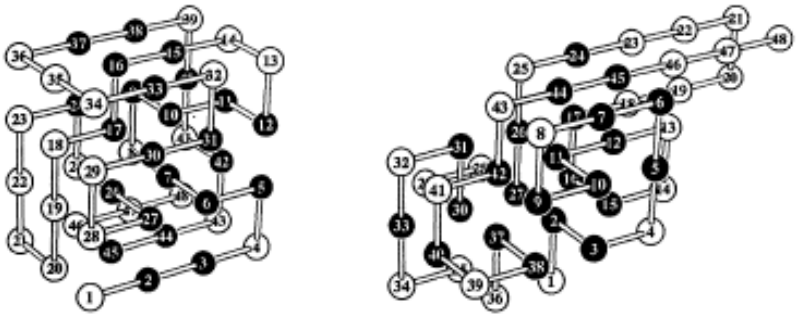
Ramachandran Graph:

- 1. For all the amino acids in a protein: calculate Φ (phi) and ψ (psi)
- 2. Plot the Φ-ψ couples in a 2D plot
- 3. The result would be:



From presentation: J. Neuteboom

Protein Lattice Models



Lattice Models (K.Yue et al., 1995)

- Simplified model, but the number of possible conformations may be rather large.
- Here two alternative lattice folds for the same 48-mer sequence is shown

55

Algorithms to find (Global) Minima in Lattice Models

Monte Carlo (MC) / Simulated Annealing

1. Start from a random coil conformation.
2. At every iteration, from a conformation S_1 with energy E_1 make a single change to a conformation S_2 and evaluate its energy E_2.
   - The single change can be a rotation around some monomer.
3. If $E_2 \leq E_1$ accept the change to conformation S_2,
4. If $E_2 > E_1$ decide non-deterministically, whether to accept the change, according to the energy increase.

Usually the acceptance criterion is Random < exp [E1 - E2 / C], where

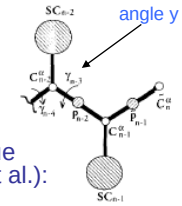
- Random is a random number in the interval [0,1] and
- C is gradually decreased ("cooled") during the simulation to achieve convergence (Simulated Annealing)

A related approach for finding low energy conformations in the HP lattice model:

- Use a genetic algorithm:
- A population of structures is generated that undergo a quasi-Darwinian evolution with the free energy as the fitness criterion.

56

A Virtual-Bond United-Residue Approximation (UNRES)



For side chains in *ab initio* predictions, a so-called United Residue Approximation (UNRES) has been suggested (H.Scheraga et al.):

- Side chains are represented by spheres ("side-chain centroids", SC).
- Each centroid represents all the atoms belonging to a real side chain.
- For every residue type (side-chain type) a van der Waals radius r°SC is introduced.

- A polypeptide chain is represented by a sequence of C_α atoms with attached united side chains (SC) and peptide group centers (p) centered between two consecutive C_α atoms.
- The distance between successive C_α atoms is set to 3.8 Å (a virtual-bond length, characteristic of a planar *trans* peptide group CO-NH).
- It is assumed that C_α - C_α - C_α virtual bond angles have a fixed value of 90° (close to what is observed in crystal structures).
- The united side chains have fixed geometry, with parameters being taken from crystal data: (see next slide).

- The only variables in this model of protein conformation are virtual-bond torsional angles γ.

57

Parameters taken from Crystal Data for Fixed Geometries of United Side Chains (SC) :

Residue	b_{sc} (Å)	θ_{sc} (deg)	ϕ_{sc} (deg)*	r^0_{sc} (Å)
Cys	1.38	120.7	-148.5	5.0
Met	2.34	120.5	-154.3	6.2
Gly	0.00	-	-	3.8
Trp	3.58	125.8	-154.2	7.2
...etc.				

* ϕ_{sc} - dihedral angle defined by SC_i - C^α_i - C^α_{i-1} - C^α_{i+1}

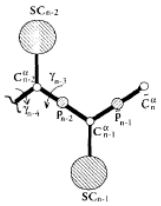
sum + 90 deg ~ 360 deg

58

Lattice Models Energy Functions (Liwo et al., 1993)

The energy function for the simplified chain can be represented as the sum of the following interactions between side chains and peptide groups:

- Hydrophobic interaction
- hydrophilic interaction
- Electrostatic interaction



Note:

- The potential functions depend on the nature of interactions, distances and dimensions of side chains.
- The parameters in the expressions for contact energies are estimated empirically from crystal structures and all-atom calculations.

59

Structure Prediction using UNRES (algorithm sketch)

1. Low-energy conformations in UNRES approximation are searched using Monte Carlo energy minimization.
   - A cluster analysis divides the set of low-energy conformations => lowest-energy representatives (structures).
   - Structures having energies ≤ the lowest energy structure + threshold are saved
2. The virtual-bond united-residue structures are converted to an **all-atom backbone** (preserving distances between C_α's).
   - The backbone is generated by Monte Carlo simulations using a "hybrid" representation of the polypeptide chain: all-atom backbone and united side chains.
   - Whereby the united side chains are still subject to the constraints following the UNRES simulations: some or even all the distances of the virtual-bond chain are substantially preserved.
3. **Full (all-atom) side chains** are introduced with accompanying minimization of steric overlaps, allowing both the backbone and side chains to move.
   - Monte Carlo simulations explore conformational space in the neighborhood of each of the low-energy structures.

60

Combinations of Approaches
Rosetta (D. Baker & coworkers)

Rosetta - combines both *ab initio* and fold recognition approaches.

Underlying idea:

In protein folding it is assumed that **local sequence fragments (3 - 9 residues)** rapidly alternate between different possible local structures.

- The distribution of conformations sampled by an isolated chain segment is approximated by
  - the distribution adopted by that sequence segment and **related sequence segments** in the protein structure database.
- Folding occurs when the conformations and relative orientations of these **local segments combine to form low energy global structures.**
- **Non-local interactions are optimized by a Monte Carlo search** through the set of conformations that can be built from the ensemble of local structure fragments.

61

Combinations of approaches
Rosetta (D. Baker & coworkers)

In the standard Rosetta protocol an approximated protein representation is used:

- Backbone atoms are explicitly included.
- **Low-resolution refinement:** side chains represented by centroids
- **High-resolution refinement:** all-atom protein representation.

Note:

- Similar stepwise refinement protocols can be used to improve predictions yielded by other methods, for instance, in loops (variable regions) of **homology-modeling structures**.

For a long period of time Rosetta turned out to be one of the most successful prediction methods in recent CASP experiments (Critical Assessment of Structure Prediction),

No optimal prediction approaches exists:

- Try to combine the best features of many different procedures
- Try to derive a consensus, meta-prediction:
  - The **3D-Jury system** generates meta-predictions using models produced by a set of servers. The algorithm scores various models according to their similarities to each other.

62

CASP: Critical Assesment of Techniques for Protein Structure Prediction

Goals

- Model similarity
- Mapping
- Structure identification
- Accuracy of comparative models
- Progress
- Most effective methods
- Focus foir future efforts

See: http://predictioncenter.org
For current lists and challenges.

#	GR name	SUM Z-score (GDT_TS)
1.	QUARK	115.788
2.	Zhang-Server	113.242
3.	RaptorX-MSA	103.270
4.	RaptorX	103.010
5.	RaptorX-Boost	99.845
6.	HHpredB	93.104
7.	HHpredA	93.104
8.	HHpredC	91.821
9.	Seok-server	89.542
10.	MULTICOM-CLUSTER	88.944
11.	BAKER-ROSETTASERVER	87.240

References

*Note: Free text availabiity, e.g. via PubMedCentral (**PMC**) is indicated. Other article are available via Leiden University Library Information Portal (U-LIP, http://metalib.leidenuniv.nl/) or (within University) at journal sites.*

Altschul S.F. et al. The statistics of sequence similarity scores. (BLAST tutorial). http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Brudno M. et al. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. **13**:721-731. (free full-text at www.genome.org)

Eddy S.R. (2004a). What is dynamic programming? *Nature Biotechnology* **22**:909-910

Eddy S.R. (2004b). Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology* **22**:1035-1036.

Gribskov M., McLachlan A.D. & Eisenberg D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84**:4355-4358. (PMC)

Pearson W.R. & Lipman D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**:2444-2448. (PMC)

64

References

Sippl M.J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *Journal of Computer-Aided Molecular Design* **7**:473-501. http://www.came.sbg.ac.at/Publications/JCAMD.7-473-501.pdf

Schwartz S. et al. (2003). Human-mouse alignments with BLASTZ. *Genome Res*. **13**:103-107. (free full-text at www.genome.org)

Thompson J.D., Higgins D.G. & Gibson T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**:4673-4680. (PMC)

Wheeler D.L. et al. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **36**:D13-D21. *(Reference source/summary of NCBI resources, in particular, Entrez retrieval system.)*

Zuker M. (2000). Calculating nucleic acid secondary structure. *Curr. Opinion Struct. Biol*. **10**:303-310.

65

WWW References

- **NCBI homepage:** http://www.ncbi.nlm.nih.gov/
- **European Bioinformatics Institute:** http://www.ebi.ac.uk
- **The ExPASy (Expert Protein Analysis System):** http://www.expasy.org
- **The PSIPRED protein structure prediction server:** http://www.psipred.net
- **MFOLD server for RNA structure prediction:**
- http://frontend.bioinfo.rpi.edu/applications/mfold/

66

Appendix

Note: no exam material

67

Protein Lattice Models

Algorithms for Lattice Models

- The Simple Cubic lattice (SC) model.
- The Body-Centered Cubic lattice (BCC).
- The Face-Centered Cubic lattice (FCC).

68

Fold Recognition: Threading Potentials (2/2)

Second type of knowledge-based threading potentials:

2. Optimized Potentials

- Optimization such that the native conformations of a set of proteins are forced to have significantly lower energies than alternative (**random**, or **decoy**) conformations.
- Optimization can be performed using the **random energy model (REM).**
  - REM assumes that decoy energies are independent random values.

69

Mean Force Potentials

Knowledge-based (database-derived) mean force potentials incorporate all forces (electrostatic, van der Waals, etc) acting between atoms as well as the influence of the environment (solvent).

For the interaction between two residues (a, b) with a sequence separation k and distance r between specified types of atoms (e.g. $C_\beta \rightarrow C_\beta$, $C_\beta \rightarrow N$ etc.) a general definition of the potential is:

$$E^{ab}_k(r) = -RT \ln [f^{ab}_k(r)]$$

where $f^{ab}_k(r)$ is obtained from a database of known structures.

The reference state is defined as

$$E_k(r) = -RT \ln [f_k(r)]$$

where $f_k(r)$ is an average value over all residue types.

=> Thus the potential for the specific interaction of residues is

$$\Delta E^{ab}_k(r) = E^{ab}_k(r) - E_k(r) = -RT \ln [f^{ab}_k(r) / f_k(r)]$$

70

Solvation Potential

The Solvation Potential for an amino acid residue a is defined as:

$$\Delta E^{a}_{solv}(r) = - RT \ln (f^{a}(r) / f(r)),$$

where:

- r is the degree of residue burial,
- $f^{a}(r)$ is the frequency of occurrence of residue a with burial r,
- f(r) is the frequency of occurrence of an arbitrary residue with burial r.

71

Optimization of Potentials

GOAL:

define a threading potential to maintain a close relationship with the underlying true potential.

Various approaches are developed to optimize the potential parameters so that:

the native conformations of proteins are discriminated from alternatives

Random energy model (REM) (Shakhnovich & coll.)

72

Optimization of Potentials

Random energy model (REM) (Shakhnovich & coll.):

- The energy of each threading alignment is defined as a sum of all pairwise contacts between particular atoms located at specified distance (with some cutoff, usually 7.5-9 Å between C_α or C_β atoms).
- The contact energies are specified in a 20×20 matrix U for all types of amino acids
- Summations are taken over all residues that separated >2 positions along the sequence.

- It is assumed that:
  - the set of alignments consists of the "native" alignment with energy EN and a set of decoys;
  - the energies of decoys take statistically independent random values.

73

Random Energy Model (REM)

- The average energy E_{av} and standard deviation σ of decoys can be estimated explicitly or derived from generated alignments.

- Given the matrix U and query sequence, the Z-score can be calculated for the native alignment: $Z_{REM} = (E_N - E_{av})/\sigma$

- Estimate pairwise potentials U by a Monte Carlo optimization procedure
  - simultaneously maximizing thermodynamic stability for all proteins in the training set database.
- Some other approaches for the optimization of contact potentials have been proposed as well.

74

CASP 11 (2014)

See: http://predictioncenter.org
For current lists and challenges.

#	GR code	GR name	Domains Count	SUM Z-score (>-2.0)
1	204	Zhang	78	76.4117
2	169	LEE	78	68.7497
3	290	MULTICOM	78	66.7849
4	044	LEER	78	66.5034
5	277	Zhang-Server	78	65.9858
6	425	Seok-refine	78	63.2947
7	499	QUARK	78	59.5585
8	065	Jones-UCL	75	58.0721
9	042	TASSER	78	56.5341
10	338	ProQ2	78	56.3264
11	132	ProQ2-refine	78	55.5291
12	333	Kiharalab	76	55.0840
13	347	Wallner	78	54.3184
14	358	Skwark	78	53.0744
15	067	CNIO	78	51.5664
16	282	PML	77	48.9282
17	144	Mufold	78	46.3855
18	438	QA-Recombinelt_H	74	43.2909
19	241	SHORTLE	75	42.5868
20	364	QA-Recombinelt_WFH	72	39.8400
21	064	BAKER	78	39.7843
22	162	McGuffin	78	35.8029
23	038	nns	78	35.5076
24	482	wfMix-KPa	72	35.1184
25	434	QA-Recombinelt_H2	72	31.8436
26	056	wfMix-KPb	72	29.9382
27	317	keasar	78	25.0417
28	310	MUFOLD-R	73	21.0179
29	368	Seder1	73	16.4363
30	008	MULTICOM-CONSTRUCT	78	16.0272
31	184	BAKER-ROSETTASERVER	78	15.7884

Protein Design

76

Protein Design: Inverse Folding

Inverse folding task:

the search for a sequence that will fold into a desired structure.

The inverse problem may have many solutions.

- There are examples of structurally similar proteins with very different amino acid sequences.

Problems to solve:

- find a sequence that folds into a given topology, and
- ensure that the sequence will not fold into any alternative structure (because of lower free energy).

Naïve: Lattice models for solving the Inverse Folding Problem:

- for a given structure, produce a design that maximizes the number of hydrophobic residues at possible contact locations.
- This does not guarantee that such a sequence would not form even more contacts in some different configuration.

77

Protein Design: Inverse Folding

Protein design methods:

- based on probabilistic approaches that gradually improve the quality of solution (sequence folded into unique structure).
- Until 2003 successful examples were restricted by relatively simple topologies such as coiled coils.
- A breakthrough in 2003:
  - a novel 93-residue fold designed using a computational strategy with multiple iterations between sequence design and structure refinement (Rosetta) in order to produce a desired topology.

78

Prediction of Coiled Coil Domains

Prediction of coiled coil domains:

- The coiled coil is a motif consisting of several α-helices wrapped around each other to form a super-coil.
- The coiled coils were first described in 1953 by Crick.

79

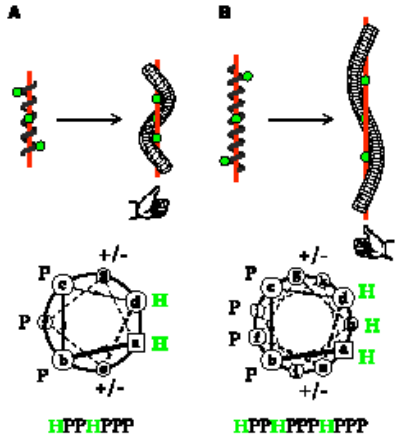
Prediction of Coiled Coil Domains

Main sequence requirements for coiled coils:

- For left-handed super-coils:
  - each of the helices contain repeats of seven residues (a-b-c-d-e-f-g)n and (a'-b'-c'-d'-e'-f'-g')n.
  - In normal α-helix, each residue would rotate about 100° around helix axis, thus 7 residues would rotate 700°.
  - The residues a and d are usually nonpolar (e.g. Leu, Val, Ile), yielding hydrophobic interactions with a' and d'
  - Residues e and g are usually charged (e.g. Glu, Lys), maintaining electrostatic interactions.
  - Positions b, c and f are typically hydrophilic.
  - Two slightly left-handed supercoiled helices (20° every 7 residues) can face each other at the axis of super-helical rotation with the same positions of the repeats.
- For right-handed super-coils:
  - Repeats of 11 residues (11 × 100° = 1100°).
  - Here in the repeat (a-b-c-d-e-f-g-h-i-j-k)n positions a, d and h are hydrophobic.

80

Prediction of Coiled Coil Domains



Two types of super-coils

- Super-coiling brings repeat units (heptads or undecatad) in identical positions relative to the superhelix axis, as seen in the helical wheel projections.

81

Prediction of Coiled Coil Domains

Straightforward approach to predict coiled coils (Lupas et al., 1991):

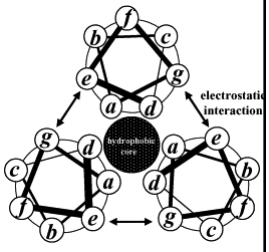
- Based on the frequencies of amino acids found in each of the seven positions in the heptad repeats contained in the database.
- Use these frequencies to score a given sequence and determine the probability for the formation of a left-handed coiled coil.

Improvement (e.g. program PairCoil, Berger et al., 1995):

- Include the frequencies of each pair of residues in each pair of heptad positions (repeat of 7).

Extension (MultiCoil, Wolf et al., 1997):

- For the identification of three-stranded coiled coils.



Prediction of Coiled Coil Domains

Coiled coils exist in

- two-, three-, four- and five-stranded conformations in both parallel and anti-parallel orientations.

An algorithm to identify coiled coil motifs in protein structures based on a search for a potential for the specific "knobs into-holes" packing noted by Crick (program SOCKET, Walshaw & Woolfson, 2001):

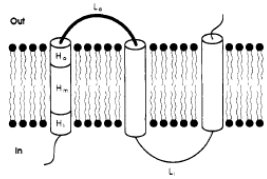
- A knob is a side-chain contacting four or more side-chains of another helix.
  - Typically residues *i*, *i*+3, *i*+4, *i*+7.
  - One of these four side-chains is in turn itself a knob in a hole formed by four side-chains of the first helix.
- Different cycles of arrangements of knobs and holes can be observed in higher-order coiled coils (other than two-stranded).

Note: The algorithm recognizes specifically cyclic knob arrangements.

83

Prediction of Trans-Membrane Segments

- Due to the specific environment in a membrane, the folding of trans-membrane proteins occurs differently as compared to globular proteins.
- Therefore, specific algorithms are needed to distinguish trans-membrane proteins and to predict their structures.

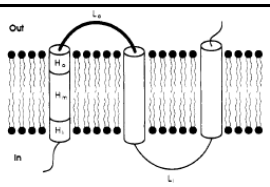


A first approximation,

- the sequences of trans-membrane proteins can be represented as helical segments of high hydrophobicity (buried in the membrane)
- alternating with the hydrophilic loops inside or outside the membrane.

84

Prediction of Trans-Membrane Segments



First approaches based on the analysis of hydrophobicity:

- A so-called hydropathy scale is introduced, wherein each amino residue has some score based on its physical characteristics (water solubility etc.).
- One of the most known is Kyte-Doolittle hydropathy index (Kyte & Doolittle, 1982).

Sliding Window Algorithm:

- Use a sliding window of say 7-10 residues and compute a moving average of the hydropathy value for the protein.
- Transmembrane segments are regions with relatively high computed hydropathy values

Note:

- the algorithm detects all hydrophobic regions, not only transmembrane helices.

85

Prediction of Trans-Membrane Segments

Statistics on known trans-membrane protein structures show:

- Clear biases in the frequencies of occurrencies of certain residues in different regions of transmembrane proteins:
  - membrane helix
  - inside loop
  - outside loop
  - inside helix
  - end (or tail)
  - outside helix end (or tail)

An expectation maximization method (Jones et al., 1994):

- Uses a set of statistical tables (computed as log likelihood ratios) to calculate the most likely topology.
- Based on a dynamic programming (similar to sequence alignment).

The modern generation of methods for trans-membrane protein structure prediction is based mostly on probabilistic methods such as hidden Markov models or Bayesian approaches.

86