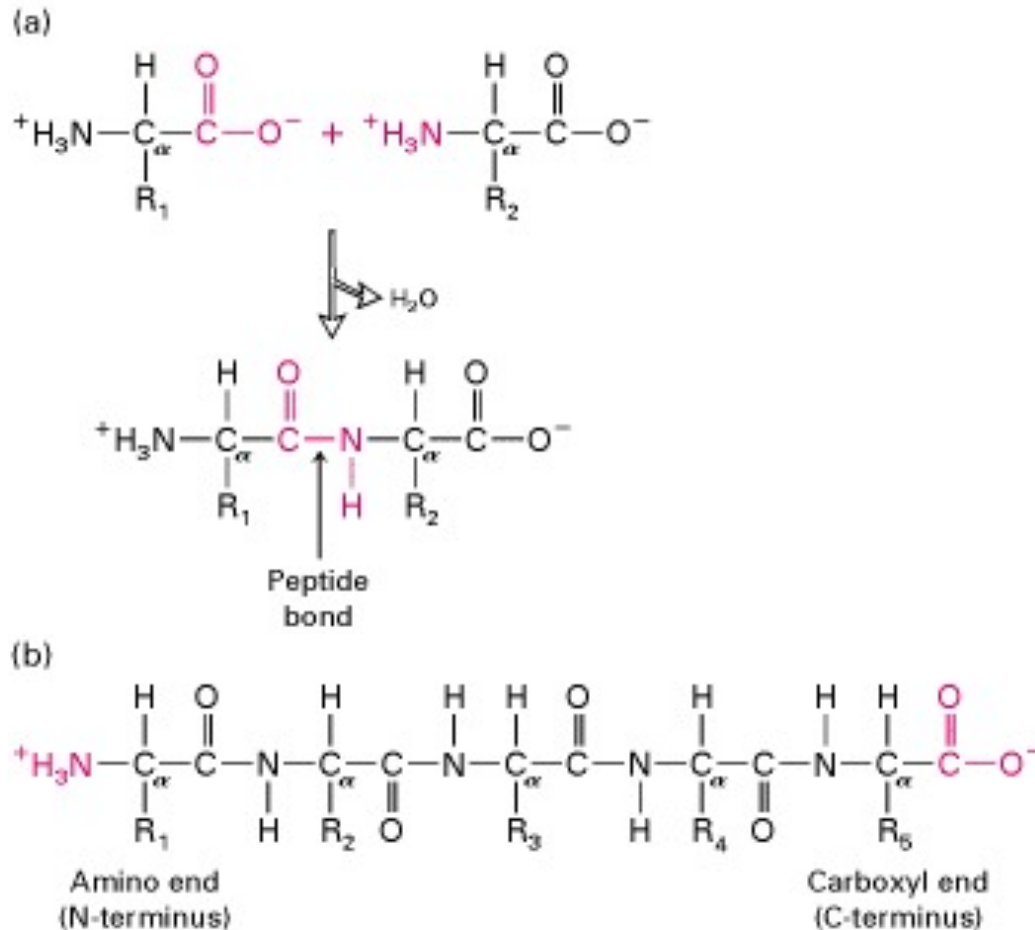


# Protein structure prediction

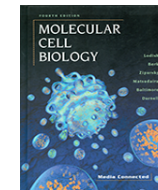
# Protein structure prediction

- Predictions of protein secondary structure
- Homology modeling
- Fold recognition
- *Ab initio* structure prediction (energy minimization)

# Protein polypeptide chain

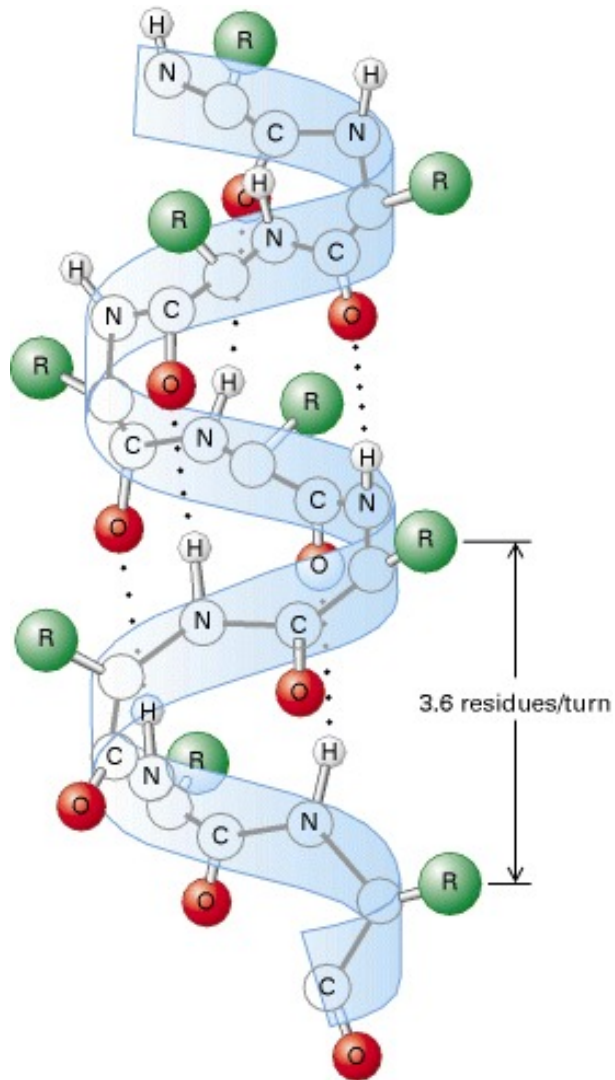


From: Section 3.1, Hierarchical Structure of Proteins

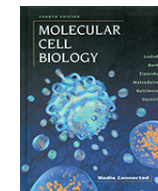


Molecular Cell Biology. 4th edition.  
Lodish H, Berk A, Zipursky SL, et al.  
New York: W. H. Freeman; 2000.

# Model of the $\alpha$ helix

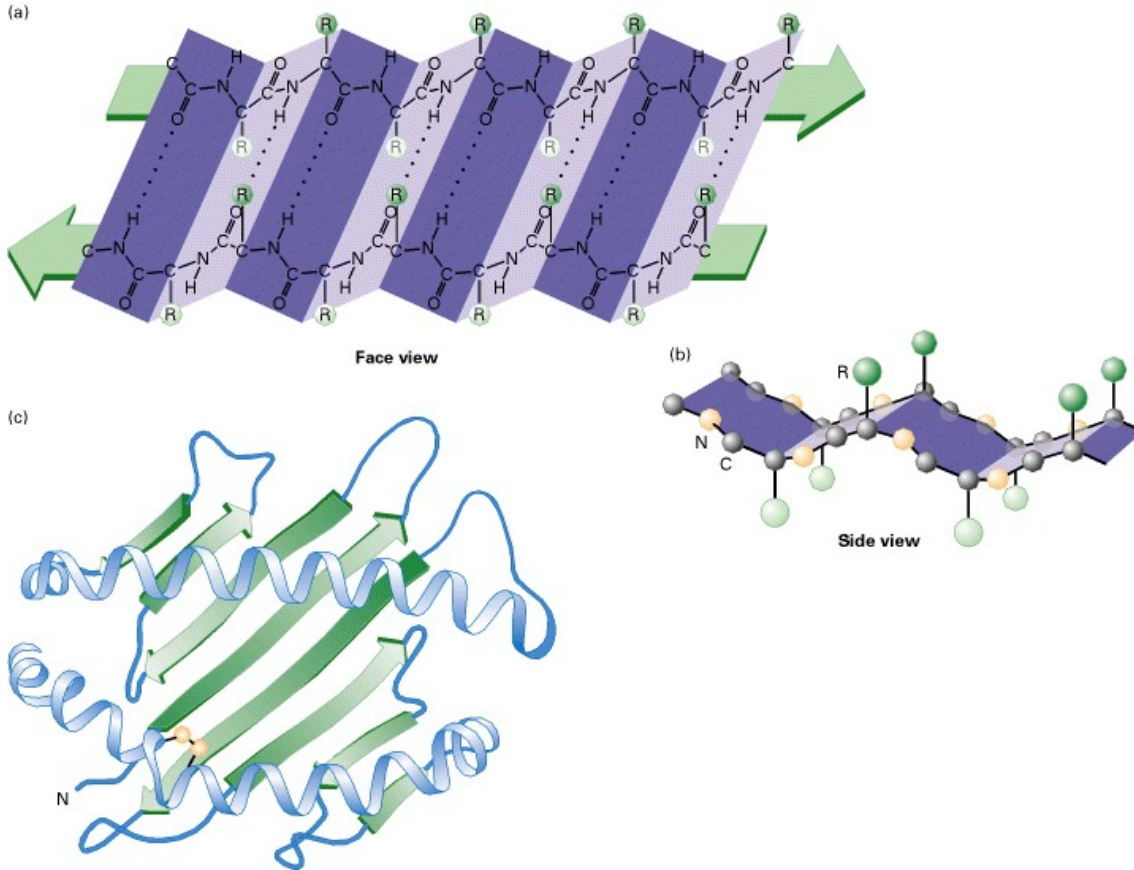


From: Section 3.1, Hierarchical Structure of Proteins

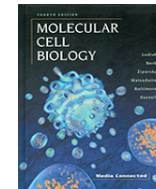


Molecular Cell Biology. 4th edition.  
Lodish H, Berk A, Zipursky SL, et al.  
New York: W. H. Freeman; 2000.

# $\beta$ sheets



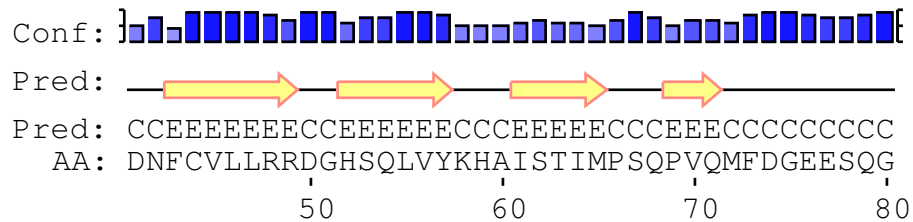
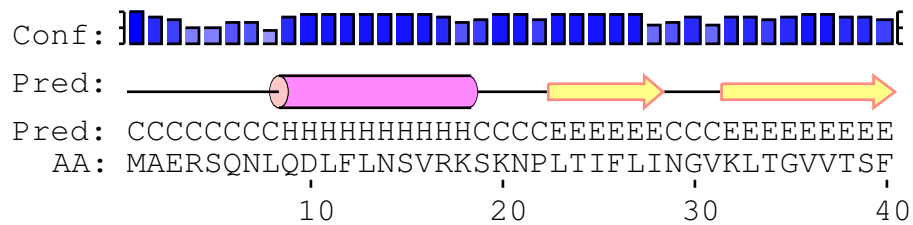
From: Section 3.1, Hierarchical Structure of Proteins


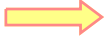



Molecular Cell Biology. 4th edition.  
Lodish H, Berk A, Zipursky SL, et al.  
New York: W. H. Freeman; 2000.

# Protein secondary structure prediction

- Protein secondary structure models consider amino acids of a polypeptide chain to be in one of the states typical for protein structures. Most simple models consider three states:  $\alpha$ -helix,  $\beta$ -strand and coil.
- Secondary structure prediction algorithms identify a state for every amino acid.
- For instance, below a result from the PSIPRED server is shown:




 = helix      Conf: ] [ = confidence of prediction  
 = strand      Pred: predicted secondary structure  
 = coil      AA: target sequence


# Protein secondary structure prediction

- Algorithms for secondary structure prediction mostly try to recognize similar patterns in local structural elements and extrapolate information from known structures to target sequences.
- The so-called first-generation algorithms were based on single amino acid propensities: for instance, Ala, Gln, Leu and Met are frequently found in helices, Gly, Tyr and Ser are not, Pro is a “helix-breaker”. A moving window with calculation of average propensity score along a sequence can indicate most likely states for sequence regions.

MAERSQNLQDLFLNSVRKSKNPLTIFLINGVKLTGVVTSF

[  $\dots S_{\text{avg}}(\alpha) \dots$  ] 

“ $\alpha$ -profile”

[  $\dots S_{\text{avg}}(\beta) \dots$  ] 

“ $\beta$ -profile”

# Protein secondary structure prediction

- Algorithms for secondary structure prediction mostly try to recognize similar patterns in local structural elements and extrapolate information from known structures to target sequences.
- The so-called first-generation algorithms were based on single amino acid propensities: for instance, Ala, Gln, Leu and Met are frequently found in helices, Gly, Tyr and Ser are not, Pro is a “helix-breaker”. A moving window with calculation of average propensity score along a sequence can indicate most likely states for sequence regions.

MAERSQNLQDLFLNSVRKSKNPLTIFLINGVKLTGVVTSF

[  $\dots S_{\text{avg}}(\alpha) \dots$  ]  $\longrightarrow$  “ $\alpha$ -profile”

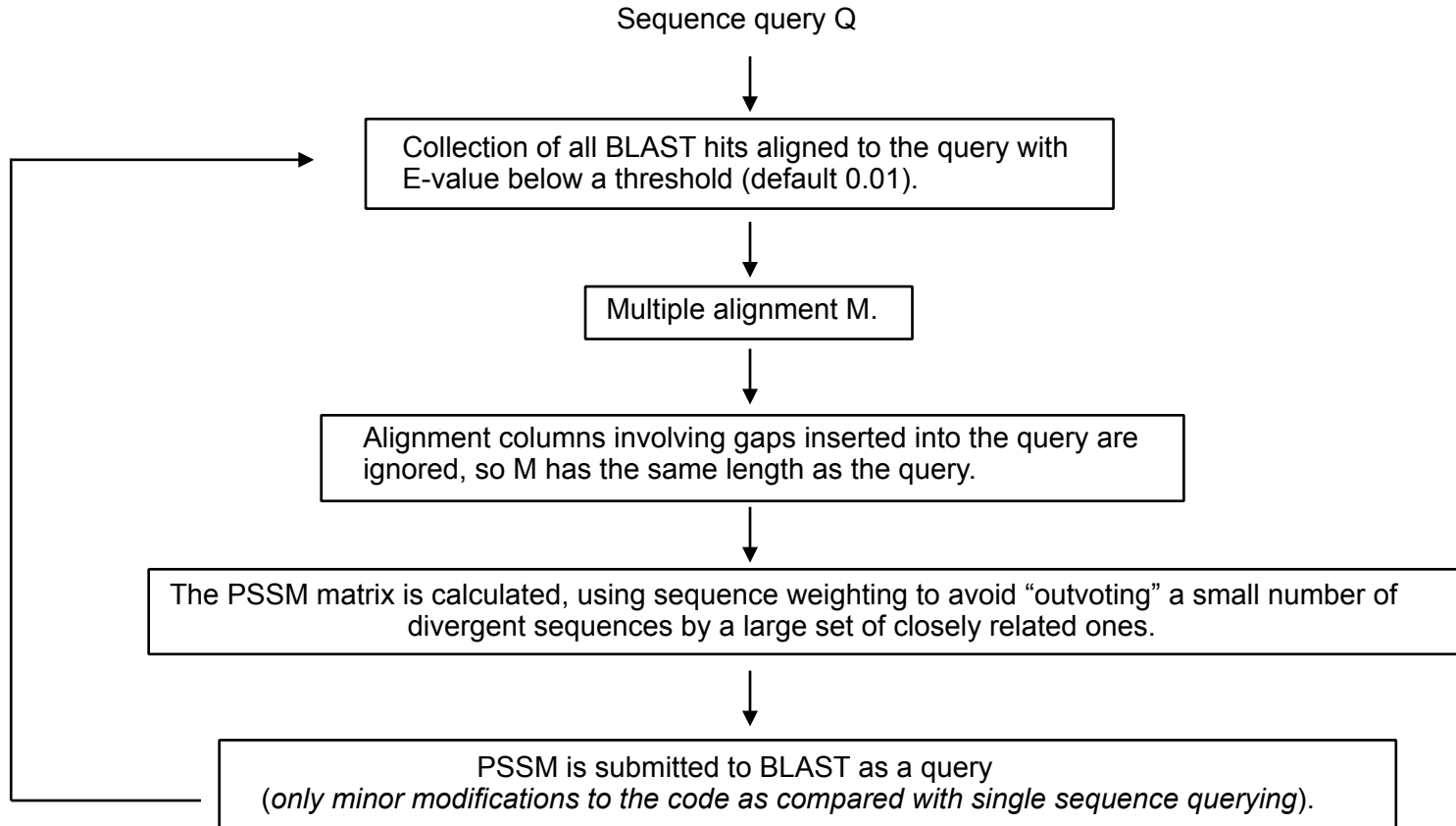
[  $\dots S_{\text{avg}}(\beta) \dots$  ]  $\longrightarrow$  “ $\beta$ -profile”

- Second-generation algorithms used propensities for segments of 3-51 amino acids.
- Currently used algorithms exploit the information from multiple alignments of related protein families, constructing profiles of patterns (PWM, Markov models) that identify most likely structural predictions.
- E.g. PSIPRED algorithm is based on profiles yielded by PSI-BLAST.



# PSI-BLAST (position-specific iterated BLAST)

(Altschul et al., 1997)

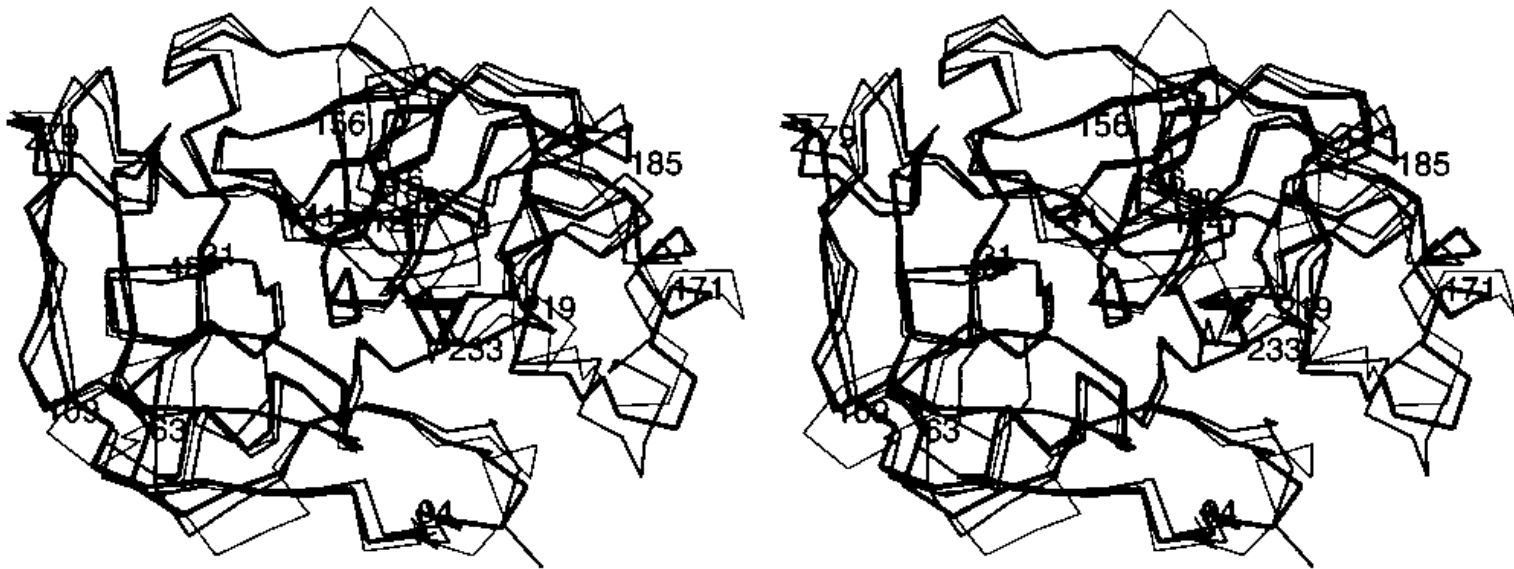


PSSM: position-specific score matrix (pronounced "possum") - a profile constructed using alignment of related sequences. PSSM dimensions are 4×N (nucleic acids) or 20×N (proteins), where N is the size of the aligned region (motif). PSSM columns correspond to motif positions, the matrix items reflect monomer frequencies at these positions.

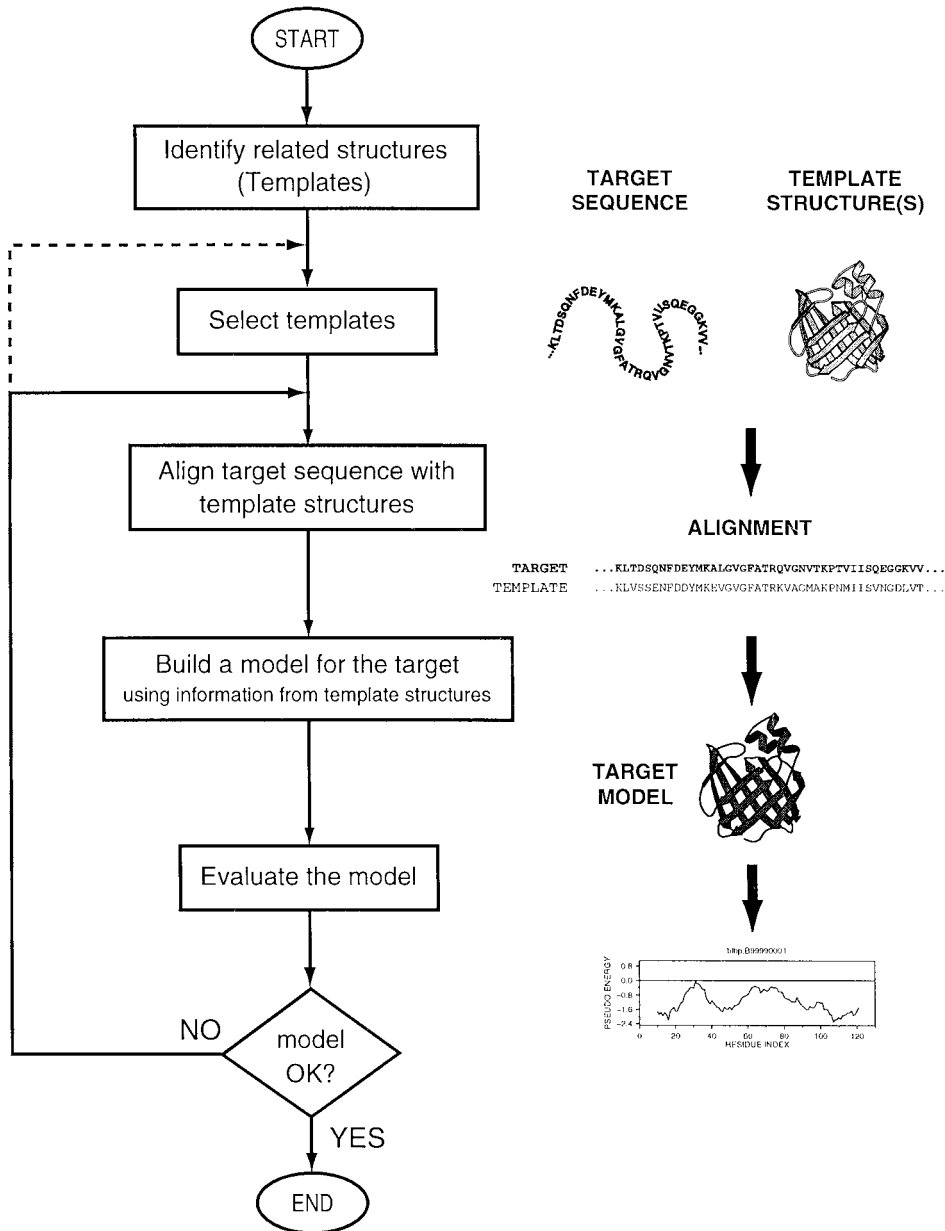
# Homology modeling

- Basic idea: “extrapolation” of the known structures to proteins with homologous sequences

Similar sequences may fold into very similar structures, e.g. below the superposition (stereoview) of the C $\alpha$  backbones of three proteins (elastase, tonin and trypsin) [Šali & Blundell, 1993].



# Main steps of homology modeling

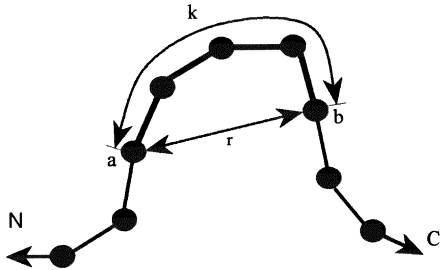


(Marti-Renom et al, 2000)

# Algorithms used in homology modeling

- **Search for template proteins of known structure.** Based on alignments of the target sequence against sequences stored in the database of structures (PDB). The best template(s) is (are) found using multiple alignments of sequences and profiles or profile hidden Markov models (HMMs).
- **Alignment of the target to template(s)** can be optimized after finding the best template(s).
- **Model building.** A number of algorithms exist. For instance, the core regions of the target can be modeled by averaging the positions of backbone atoms in the templates. Alternatively, the spatial restraints can be retrieved from multiple alignment and used to guide the modeling. The model is derived by minimizing the violations of restraints.
- **Loop modeling.** The regions with poor or no similarity to template sequences are modeled separately. E.g. using the conformations of similar fragments in the structure database or *ab initio* predictions.

# Calculation of knowledge-based (mean force) potentials using a database of protein structures



(M. Sippl, 1993)

A general idea to compute a pairwise interaction potential:

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

where frequency  $f^{ab_k}(r)$  is obtained from a database of known structures ( $a$  and  $b$ : some amino acid types),  
 $R$  - universal gas constant,  
 $T$  - temperature (K).

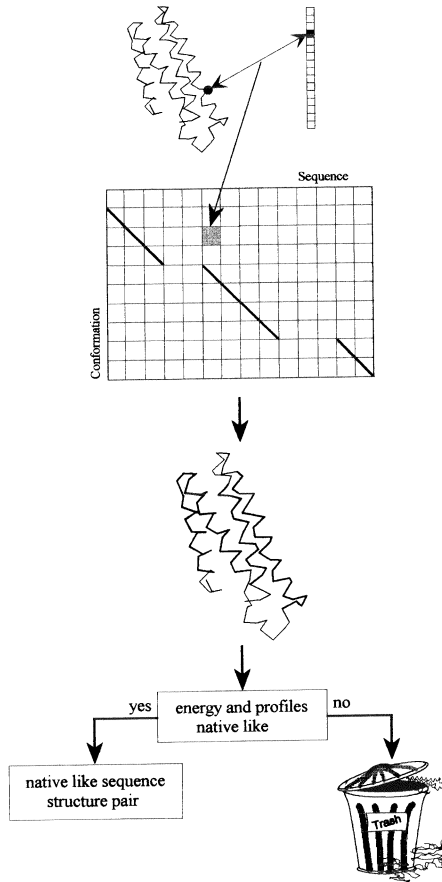
("Inverse" Boltzmann law)

Reference state can be defined as  $E_k(r) = - RT \ln [ f_k(r) ]$ ,  
where  $f_k(r)$  is an average value over all amino acid types.

Thus:

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

# Sequence/structure alignment (threading)



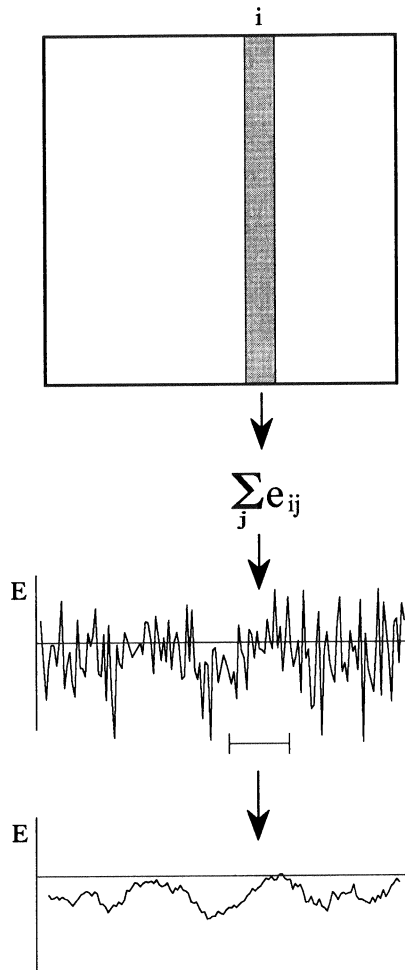
In sequence/structure threading, first the changes of total interaction energies of residues are calculated in assumption of the template structure interactions. E.g. amino acid at position  $i$  of the structure is replaced by amino acid  $j$  of the sequence, yielding the element  $[i, j]$  of the comparison matrix.

(frozen approximation)

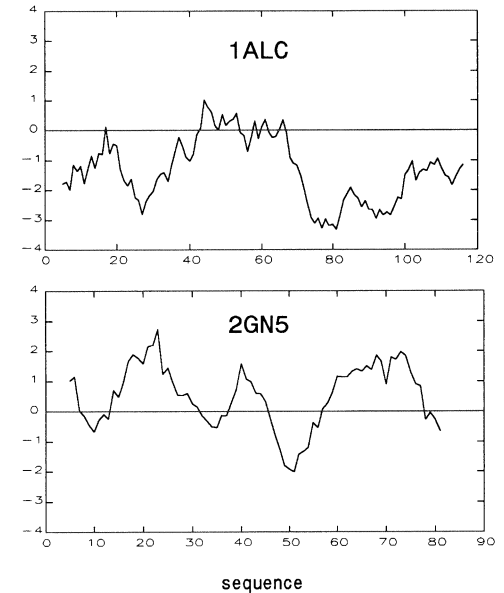
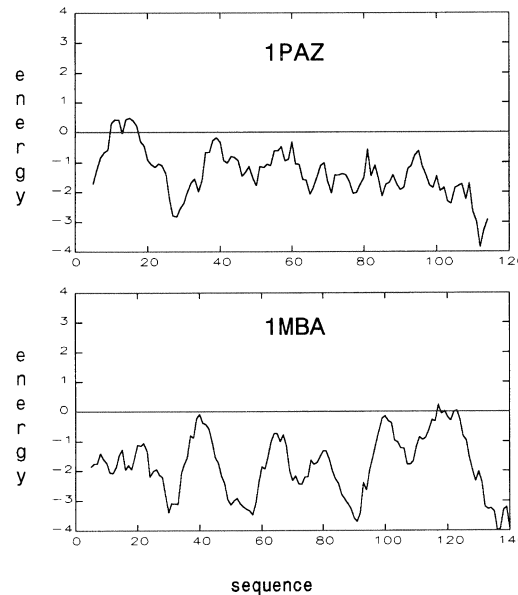
The alignment is computed by dynamic programming, yielding a structural model for the query sequence.

The quality of such models can be tested using calculations of total energies and energy profiles.

# Using energy profiles to evaluate structure models



(M. Sippl, 1993)



The profiles are smoothed using “gliding average” (e.g. over 10 residues).

Good models have relatively low energies along a sequence. The 2GN5 model does not seem to be good (positive peaks).

# *Ab initio (de novo)* protein structure prediction

## **- Lattice models**

*Rough approximation, with amino acid monomers occupying the discrete points in space determined by a lattice (usually cubic). Nowadays are mostly used for testing new ideas on folding algorithms rather than for real structure predictions.*

## **- Off-lattice models**

Low resolution models:

e.g. united residue model (UNRES), with side chain centroids.

High resolution models:

all-atom structures.

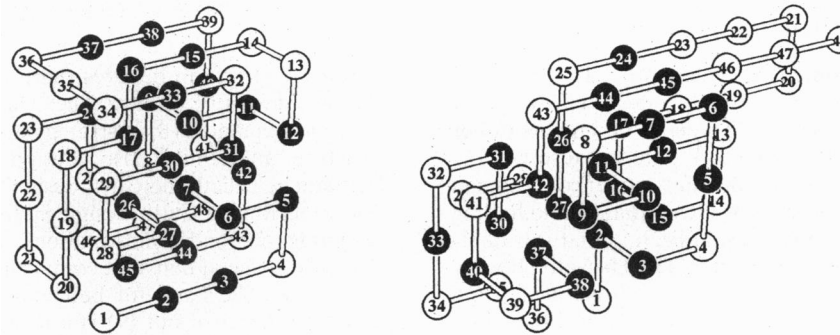


# Cubic lattice models

Mostly the HP-lattice model:

- Two monomer types: H (hydrophobic) and P (polar).
- Energy is proportional to the number of H-H contacts ( $h$ ) between closely located monomers that are not sequence neighbors:  $E = -\epsilon \times h$ .
- The energy minimum is usually searched by a Monte Carlo algorithm.

Below two alternative conformations for the same HP-sequence are shown:

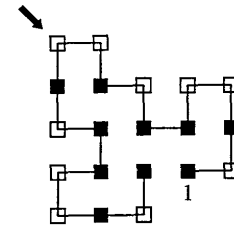
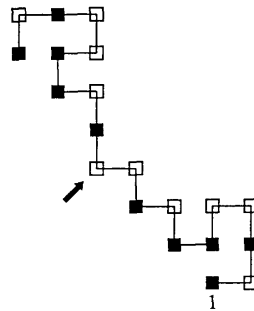


(Yue et al., 1995)

# Monte Carlo simulations of low free energy conformations

Monte Carlo simulations are based on generation of (quasi)random conformations. During the simulation, random changes are introduced. Lower free energies serve as a criterion to select structures for subsequent iterations.

The conformational transitions are usually rotations around some points (chosen randomly). In this example, the rotation makes the HP-structure more compact, changing the energy from - 4 to - 9:

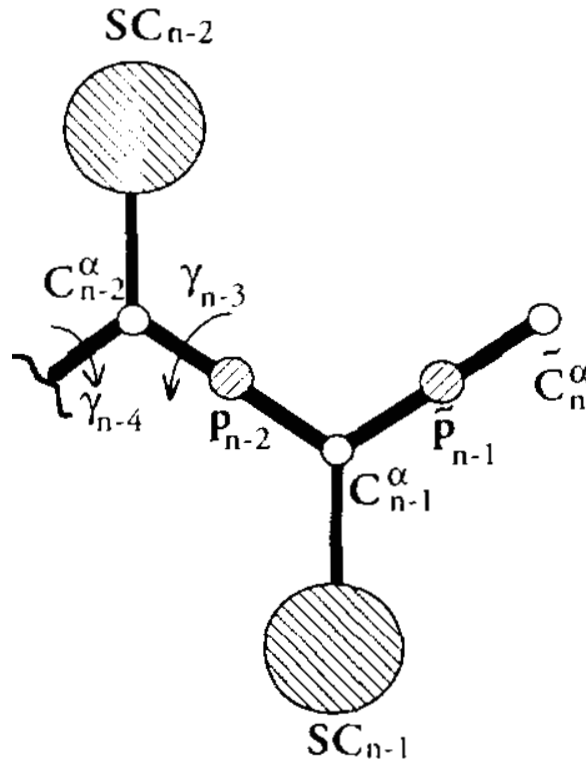


(Unger & Moult, 1993)

An example of simulation:

1. Start from a random coil conformation.
2. At every iteration: make a single change (rotation) from a conformation  $S_1$  with energy  $E_1$  to a conformation  $S_2$  with energy  $E_2$ .
3. If  $E_2 \leq E_1$ , accept the change to conformation  $S_2$ .  
If  $E_2 > E_1$ : accept with a probability criterion.  
E.g.  $p = \exp ( E_1 - E_2 / c )$ ,  
where  $c$  can be gradually decreased to “cool down” the simulation.

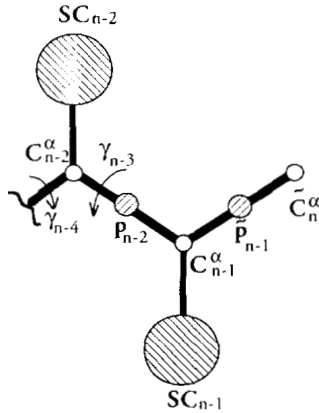
# An off-lattice model: virtual bond united-residue approximation (UNRES)



(A. Liwo et al., 1993)

- $C^{\alpha}$  - “virtual bonds” of 3.8 Å, with  $C^{\alpha}$  -  $C^{\alpha}$  -  $C^{\alpha}$  angles of  $90^{\circ}$ .
- Amino acids: approximated by “side-chain centroids” (SC).
- For each residue type, specific SC parameters (angles and centroid sizes).
- The only variables in this model are torsional angles  $\gamma$  of rotation around virtual bonds.

# Ab initio structure prediction using virtual bond united-residue approximation (UNRES)



Low energy conformations are usually searched by Monte Carlo (MC) algorithms in stepwise way, moving from the low resolution in UNRES to high-resolution all-atom structures. Energy potentials include various interactions (hydrophobic, hydrophilic, electrostatic) between atoms and/or molecular groups considered at a particular step.

(A. Liwo et al., 1993)

For instance:

- Begin with UNRES approximation with interactions between SC and peptide groups only: low energy structures can be found by MC simulation.

- The backbone atoms are introduced in these structures, and the folds are further optimized by MC.

*(An approximation with all-atom backbone and SC centroids is frequently called low-resolution refinement).*

- All atoms are introduced to the structures of the previous step, and MC simulation is performed on the all-atom model (high-resolution refinement).

# Variations in the algorithms for protein structure prediction

Different approaches can be combined in a single algorithm for structure prediction.

E.g. various combinations of conformational sampling (template-based or knowledge-based) with low/high resolution refinement.

One of the most successful applications is Rosetta methodology (D. Baker & coll.).

*E.g. according to **CASP**, Critical Assessment of Protein Structure Prediction, a biannual evaluation of prediction methods, carried out in a blind mode.*

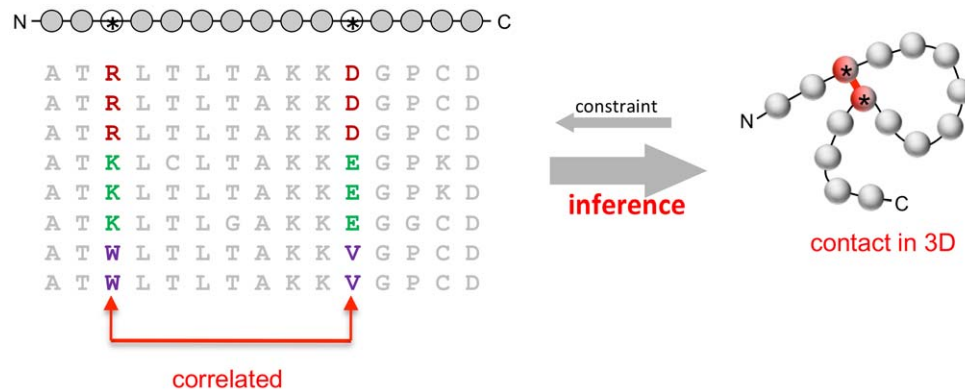
In Rosetta, protein folding is considered as an interplay of local interactions in the relatively small oligopeptide fragments and global assembly of these fragments by Monte Carlo energy minimization.

A single fragment is considered to fluctuate between several local structures. Such a fluctuation is modeled using the distribution of conformations observed in similar fragments of known crystal structures. In the first Rosetta stage, alternative minima of free energy can be identified using the coarse-grained low-resolution energy function.

The second stage starts from each of the low-resolution minima and returns back the atomic coordinates. The conformations are further optimized by a multistep Monte Carlo energy minimization procedure.

# Pairwise contacts predicted from amino acid covariations

Correlations between substitutions can be used for prediction of interactions:



(DS Marks et al., 2011)

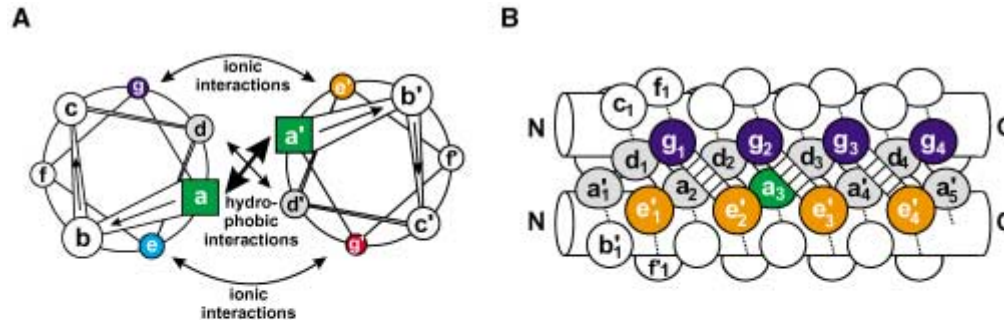
A straightforward application of Mutual Information (MI) values for detection of monomer contacts can yield a number of correlations that are not determined by interactions (*transitive indirect correlations*).

A (partial) solution for this problem can be provided by a model that is built for the whole alignment length in order to infer maximally informative correlations.

The growth of the sequence and structure databases may improve such approaches in future.

## A special case: coiled coil domains

Coiled coils are 2-5  $\alpha$ -helices wrapped around each other. They are stabilised by heptad repeats (usually denoted with a-b-c-d-e-f-g positions). The heptads can form a regular extended stable conformation because seven residues in a helix make a rotation close to two turns:  $7 \times 100^\circ = 720^\circ - 20^\circ$ . Stability is established via hydrophobic (a-d) and polar (e-g) interactions.

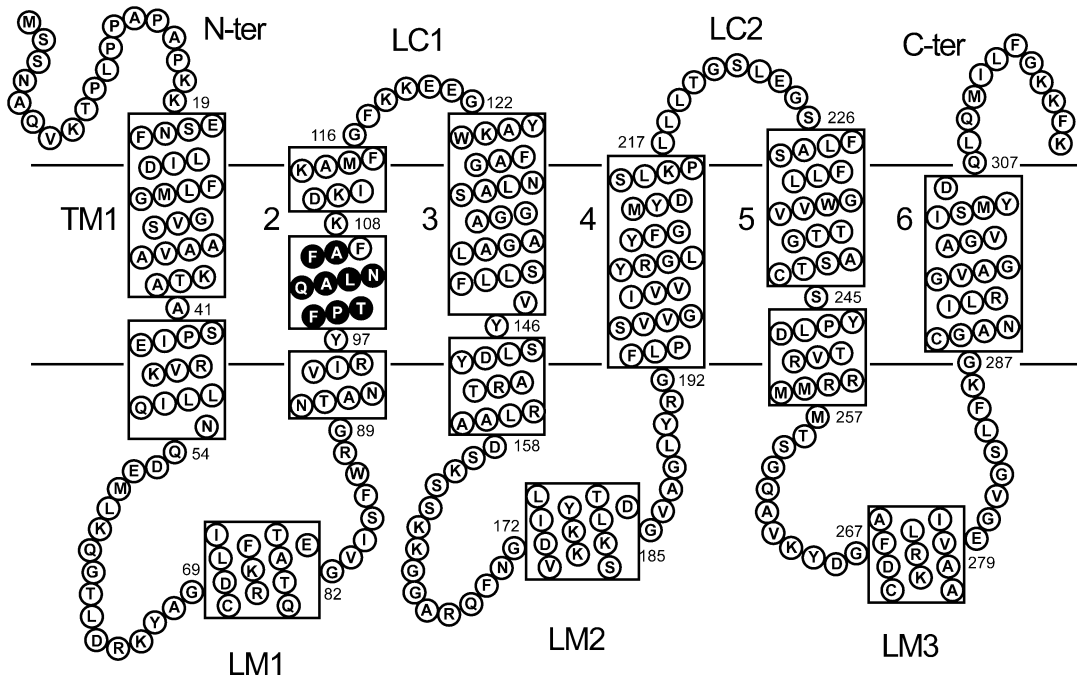


(Mason & Arndt, 2004)

Coiled coils are predicted by special algorithms, based on the estimates of probabilities of finding amino acid residues at specific heptad positions. For a given sequence, a total score can be calculated. Furthermore, additional side-chain interactions in the coiled coils can be taken into account.

# A special case: transmembrane proteins

A schematic representation of transmembrane (TM) protein with 6 TM segments:



(Kihira et al., 2004)

Transmembrane proteins have several transmembrane (TM)  $\alpha$ -helices. Predictions of TM topology require special algorithms, because the lipid environment differs from that of globular proteins. The algorithms usually calculate the most likely attributes for all residues, classifying them in the main structural elements such as (1) TM helix; (2) inside and (3) outside loops; (4) inside and (5) outside helix ends.