Protein Structure Analysis

Majid Masso

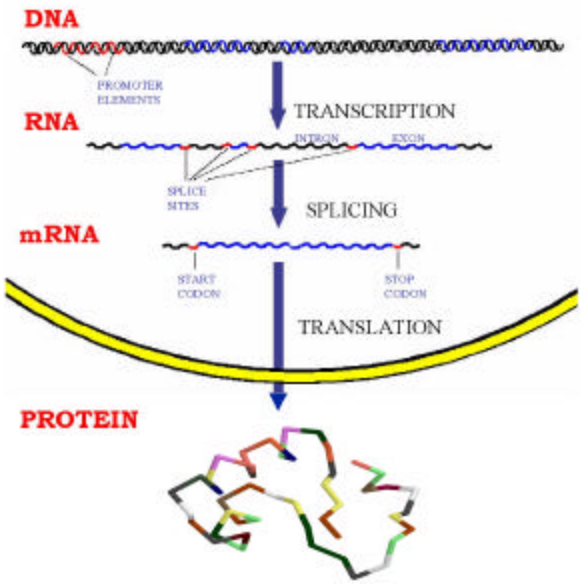
Secondary Structure: Computational Problems

Secondary structure characterization
Secondary structure assignment
Protein structure classification
Secondary structure prediction

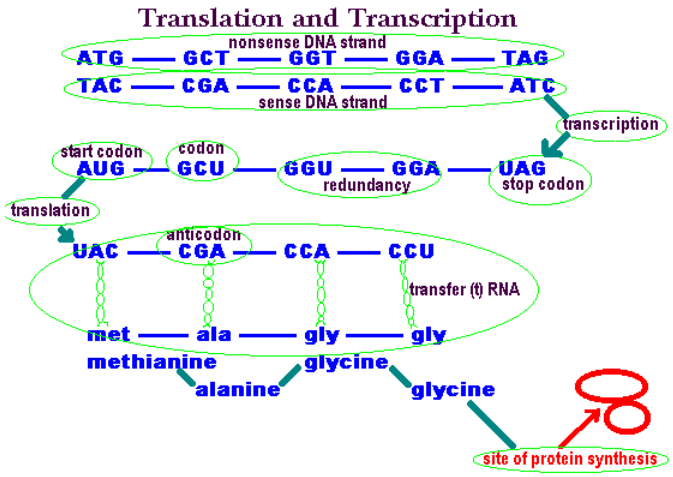
Protein Basic Structure

- A protein is made of a chain of amino acids.
- There are 20 amino acids found in nature
- Each amino acid is coded in the DNA by one or more codons, i.e. a three base sequence.

Cell Informatics



Transcription and Translation

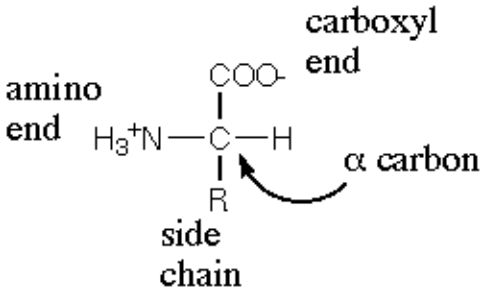


From http://www.agen.ufl.edu/~chyn/age2062/lect/lect_07/of7_1a.GIF

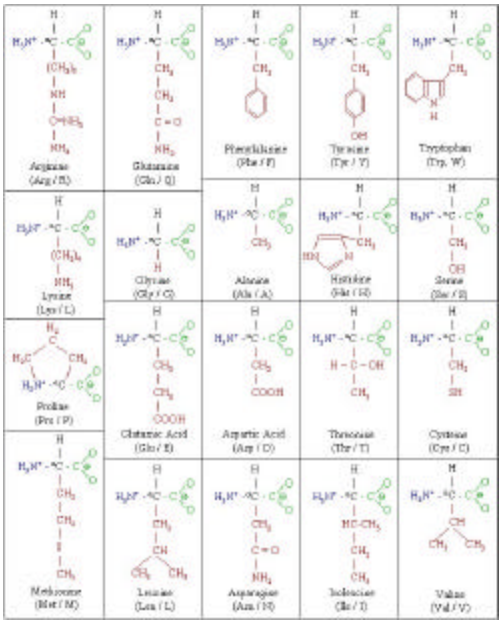
Finding the Protein Sequence

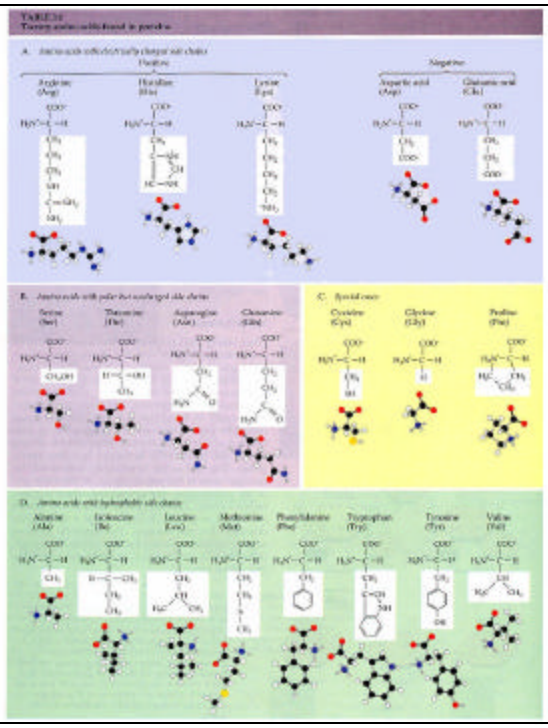
- From DNA sequence
- From protein sequencer
- From mRNA sequence

Amino Acid Residue

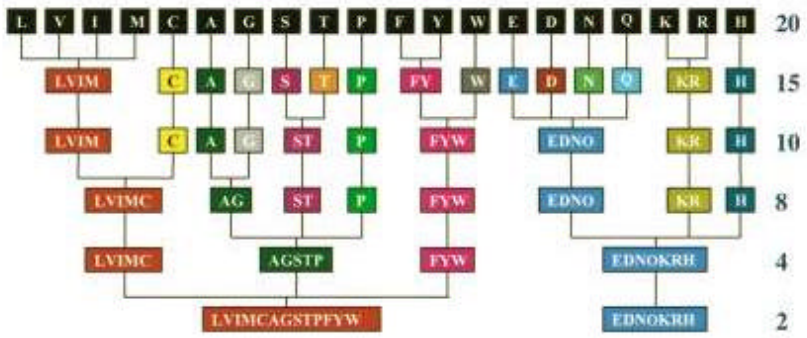


Amino Acids



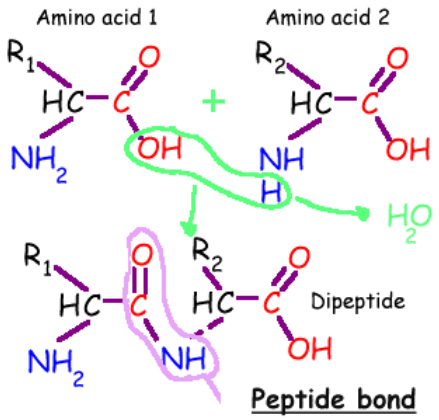


Amino Acid Residue Clustering

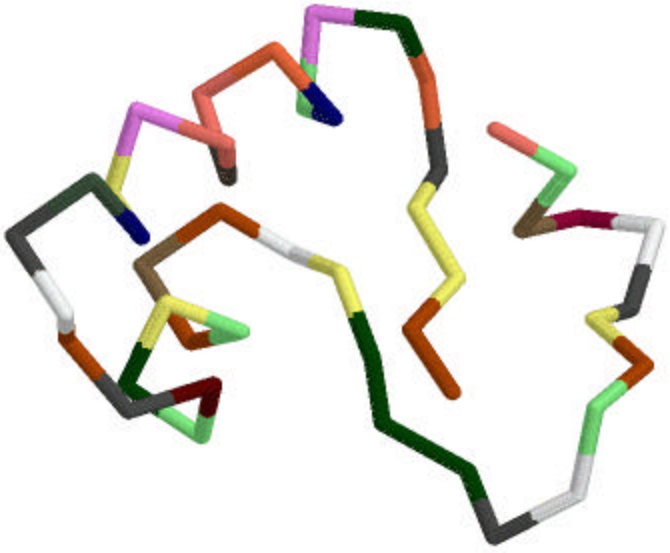


Adopted from: L.R.Murphy et al., 2000

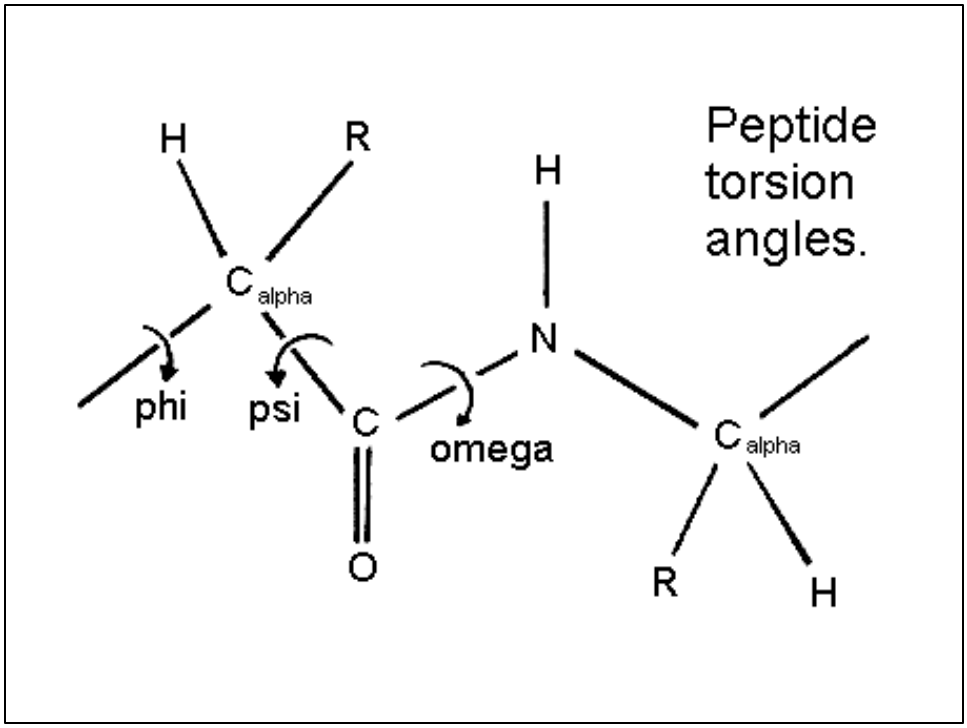
Peptide Bond

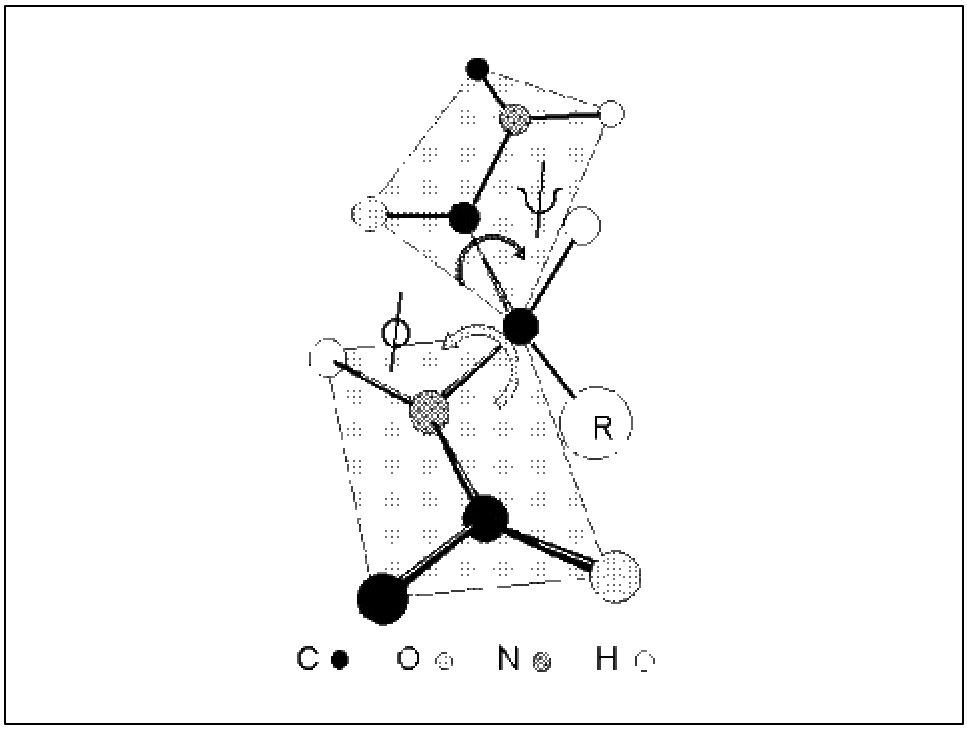


Protein backbone



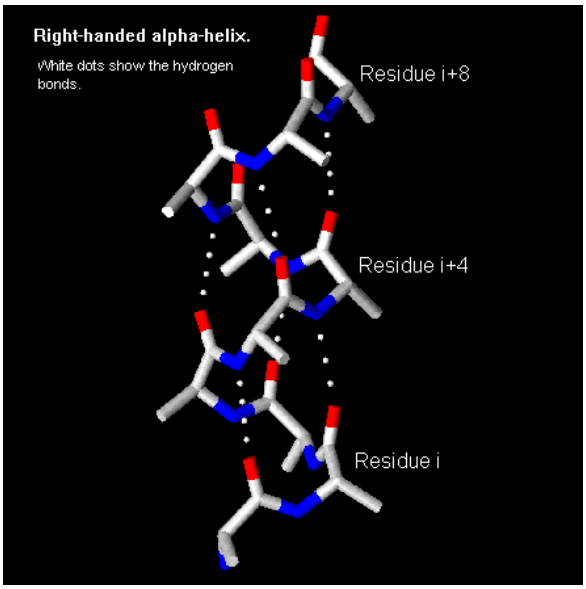
NWVLSTAADMQGVVTDGMASGLDKD


Peptide torsion angles.
H
R
C alpha
phi
psi
C
O
H
N
omega
C alpha
R
H


R
C
O
N
H

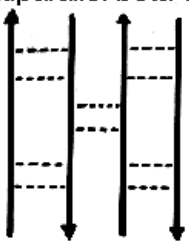
Secondary Structure (Helices)

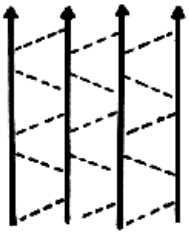
Helix



Secondary Structure (Beta-sheets)

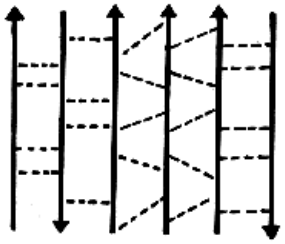
Antiparallel beta-sheet



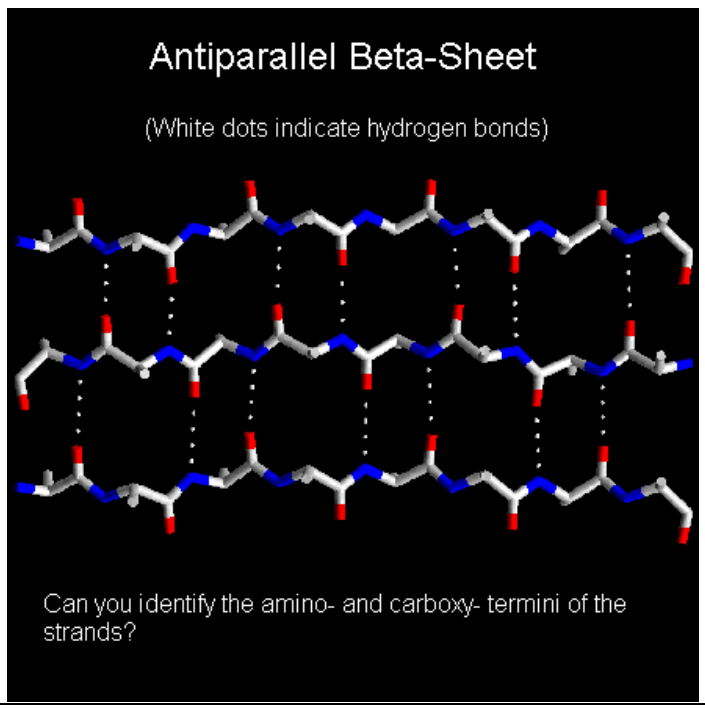


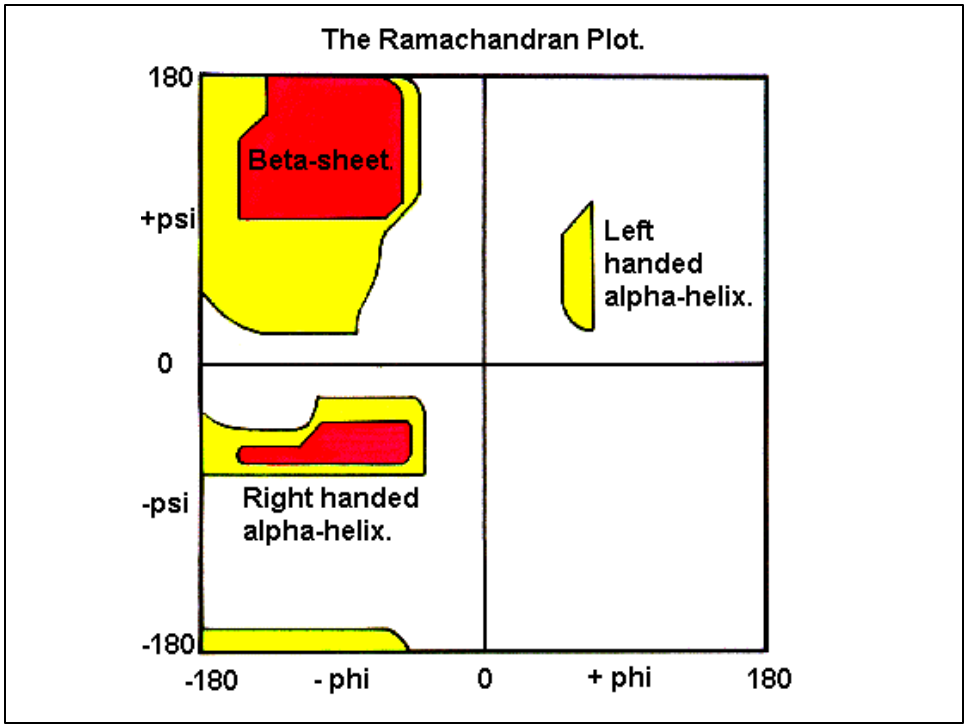
Parallel beta-sheet

The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.



Mixed beta-sheet

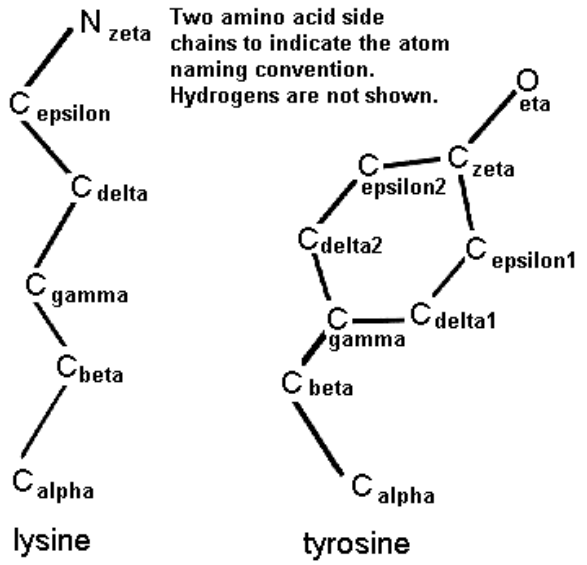




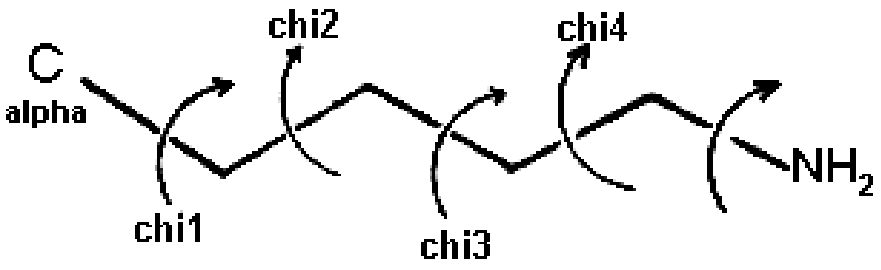
Secondary Structure Conformations

	ϕ	ψ
alpha helix	-57	-47
alpha-L	57	47
3-10 helix	-49	-26
π helix	-57	-80
type II helix	-79	150
β-sheet parallel	-119	113
β-sheet antiparallel	-139	135

Side-Chain Atom Nomenclature



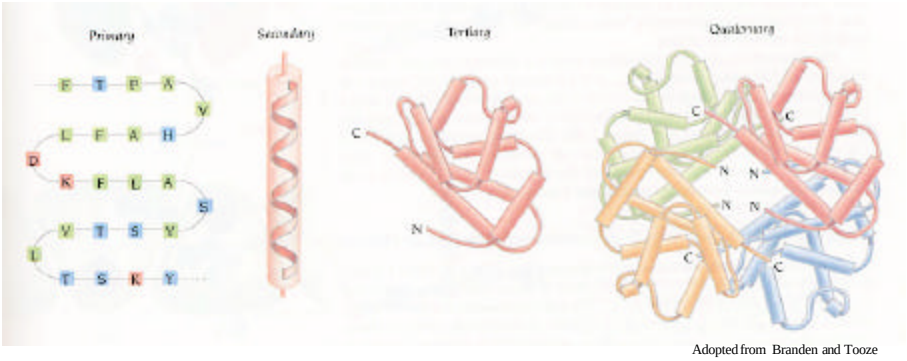
Side-Chain Torsional Angles



Four Levels of Protein Structure

- Primary Structure – Sequence of amino acids
- Secondary Structure – Local Structure such as α-helices and β-sheets.
- Tertiary Structure – Arrangement of the secondary structural elements to give 3-dimensional structure of a protein
- Quaternary Structure – Arrangement of the subunits to give a protein complex its 3-dimensional structure.

Protein Structure Hierarchy



Adopted from Branden and Tooze

- Primary - the sequence of amino acid residues
- Secondary - ordered regions of primary sequence (helices, beta-sheets, turns)
- Tertiary - the three-dimensional fold of a protein subunit
- Quaternary - the arrangement of subunits in oligomers.

Protein Structure Determination

X-ray crystallography

NMR spectroscopy

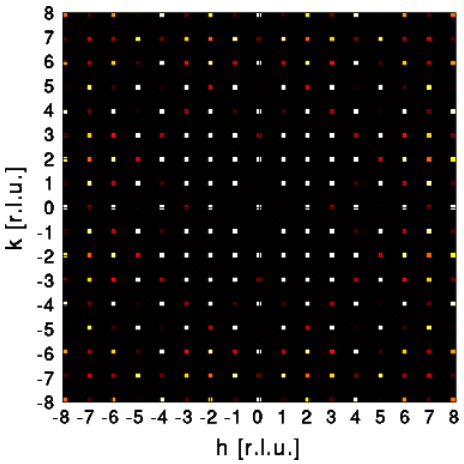
Neutron diffraction

Electron microscopy

Atomic force microscopy

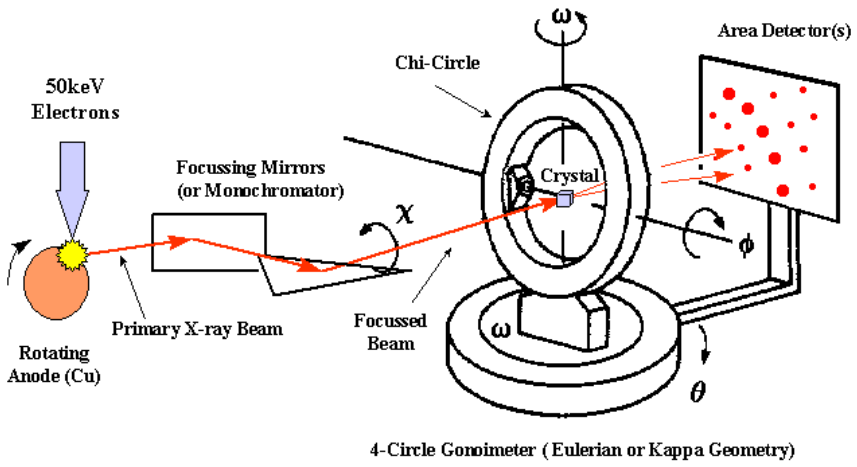
Measuring Protein Structure

- Determining protein structure directly is difficult
- X-ray diffraction studies – must first be able to crystallize the protein and then calculate its structure by the way it disperses X-rays.



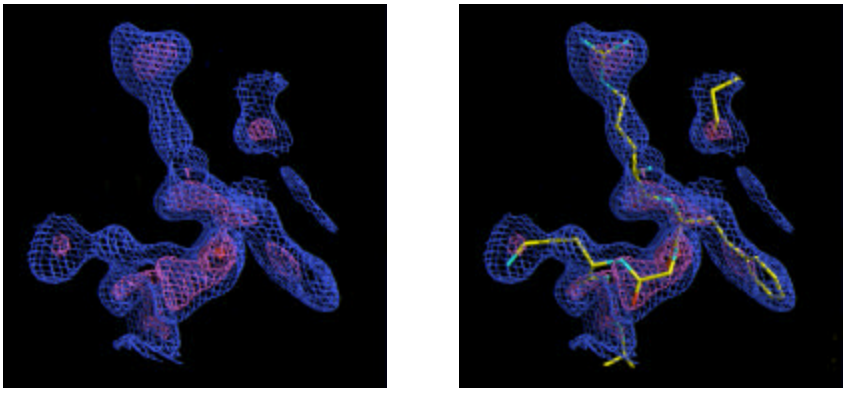
From http://www.uni-wuerzburg.de/mineralogie/crystal/teaching/inv_a.html

X-ray crystallography



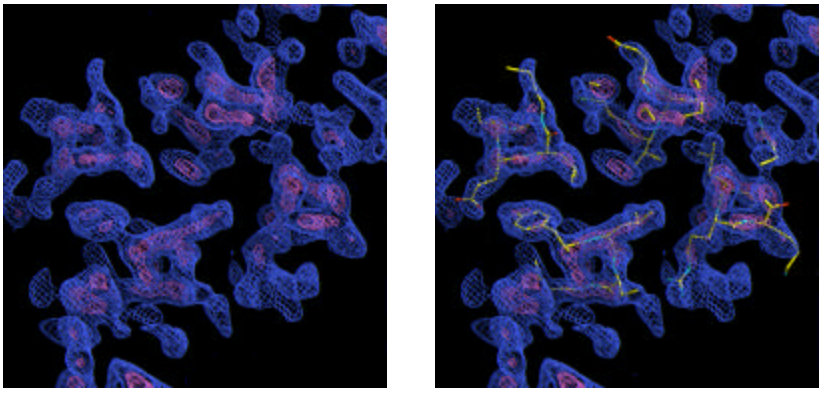
4-Circle Goniometer (Eulerian or Kappa Geometry)

X-ray crystallography



Electron density map created from multi-wavelength data (Arg)

X-ray crystallography



Experimental electron density map and model fitting
(apoE four helix bundle)

X-ray crystallography

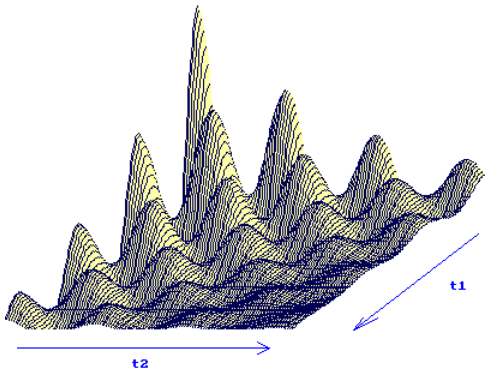
Confidence in structural features of proteins determined by X-ray crystallography

(These are *rough* estimates, and depend strongly on the quality of the data.)

Structural feature	Resolution				
	5 Å	3 Å	2.5 Å	2.0 Å	1.5 Å
Chain tracing	—	Fair	Good	Good	Good
Secondary structure	Helices fair	Fair	Good	Good	Good
Sidechain conformations	—	—	Fair	Good	Good
Orientation of peptide planes	—	—	Fair	Good	Good
Protein hydrogen atoms visible	—	—	—	—	Good

Measuring Protein Structure

- NMR – Use nuclear magnetic resonance to predict distances between different functional groups in a protein in solution. Calculate possible structures using these distances.



http://www.cis.rit.edu/htbooks/nmr/inside.htm

Why not stick to these methods?

- X-ray Diffraction –
  - Only a small number of proteins can be made to form crystals.
  - A crystal is not the protein's native environment.
  - Very time consuming.
- NMR Distance Measurement –
  - Not all proteins are found in solution.
  - This method generally looks at isolated proteins rather than protein complexes.
  - Very time consuming.

Structure verification and validation

Biotech Validation Suite:
http://biotech.ebi.ac.uk:8400/
ERRAT
Verify3D
Procheck

Procheck programs

CLEAN - cleaning PDF file

SECSTR - assigning secondary structure

NB - identifying non-bonded interactions

ANGLEN - calculating bond lengths and bond angles

TPLOT, PPLOT, BPLOT - graphical output

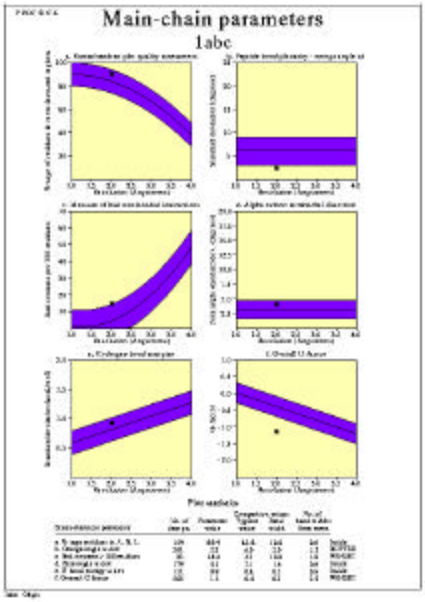
Bond lengths (Procheck)

Bond	labeling		Value	sigma
C-N	C-NH1	(except Pro)	1.329	0.014
	C-N	(Pro)	1.341	0.016
C-O	C-O		1.231	0.020
Calpha-C	CH1E-C	(except Gly)	1.525	0.021
	CH2G*-C	(Gly)	1.516	0.018
Calpha-Cbeta	CH1E-CH3E	(Ala)	1.521	0.033
	CH1E-CH1E	(Ile,Thr,Val)	1.540	0.027
	CH1E-CH2E	(the rest)	1.530	0.020
N-Calpha	NH1-CH1E	(except Gly,Pro)	1.458	0.019
	NH1-CH2G*	(Gly)	1.451	0.016
	N-CH1E	(Pro)	1.466	0.015

Bond angles (Procheck)

Angle	labeling		Value	sigma
C-N-Calpha	C-NH1-CH1E	(except Gly,Pro)	121.7	1.8
	C-NH1-CH2G*	(Gly)	120.6	1.7
	C-N-CH1E	(Pro)	122.6	5.0
Calpha-C-N	CH1E-C-NH1	(except Gly,Pro)	116.2	2.0
	CH2G*-C-NH1	(Gly)	116.4	2.1
	CH1E-C-N	(Pro)	116.9	1.5
Calpha-C-O	CH1E-C-O	(except Gly)	120.8	1.7
	CH2G*-C-O	(Gly)	120.8	2.1

Procheck output



a. Ramachandran plot quality - **percentage** of the protein's residues that are in the **core** regions of the Ramachandran plot.

b. Peptide bond planarity - **standard deviation** of the protein structure's **omega** torsion angles.

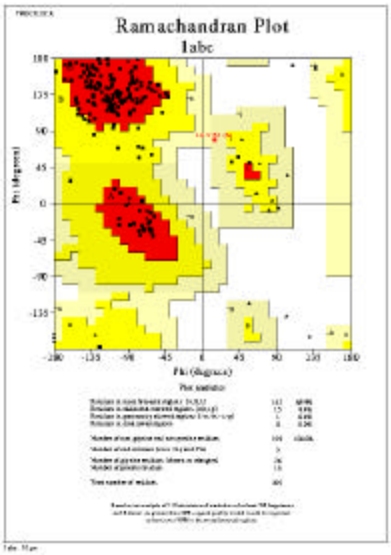
c. Bad non-bonded interactions - **number of bad contacts** per **100** residues.

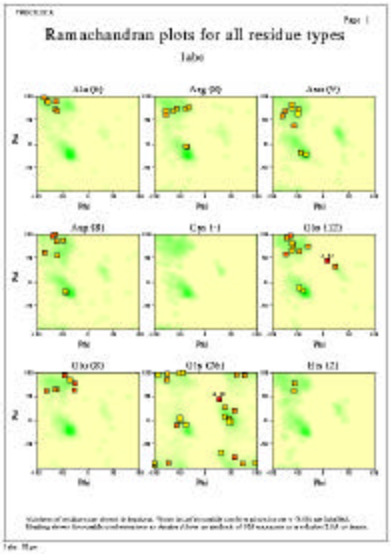
d. Ca tetrahedral distortion - **standard deviation** of the z **torsion angle** (**C**a , **N**, **C**, and **C**b**)**.

e. Main-chain hydrogen bond energy - **standard deviation** of the **hydrogen bond energies** for **main-chain hydrogen bonds**.

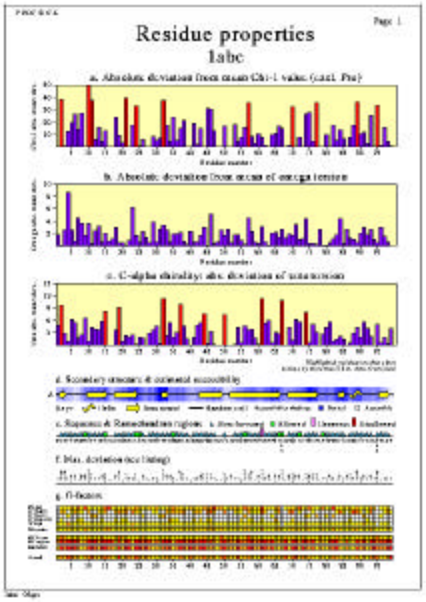
f. Overall G-factor - average of different G-factors for each residue in the structure.

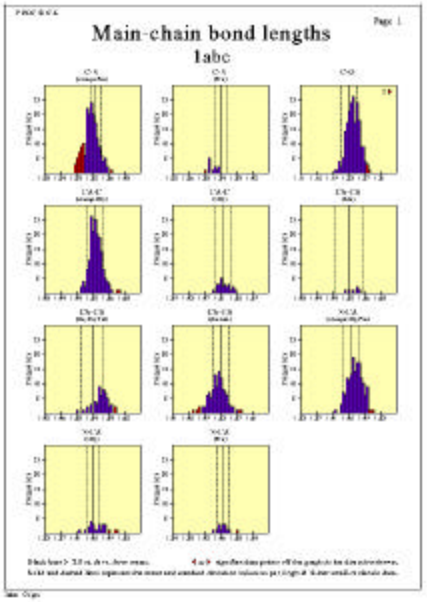
Procheck output



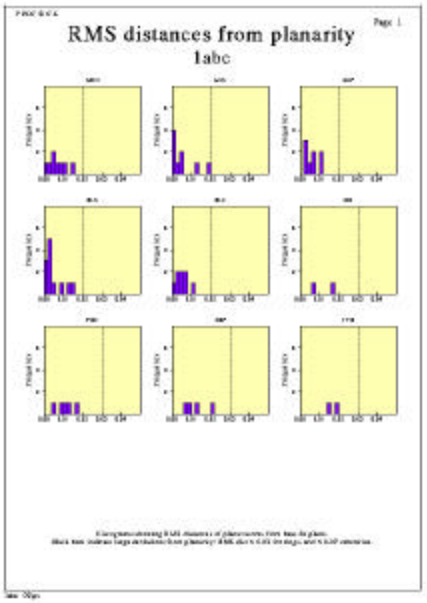


Procheck output

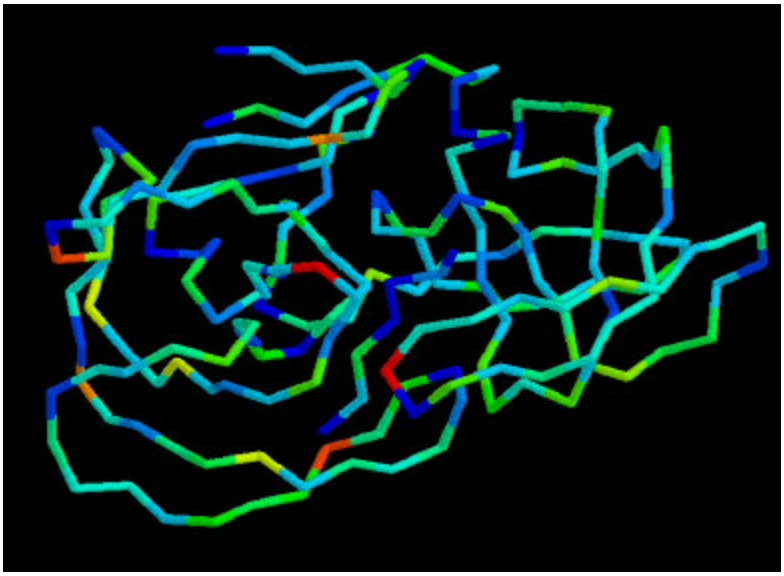




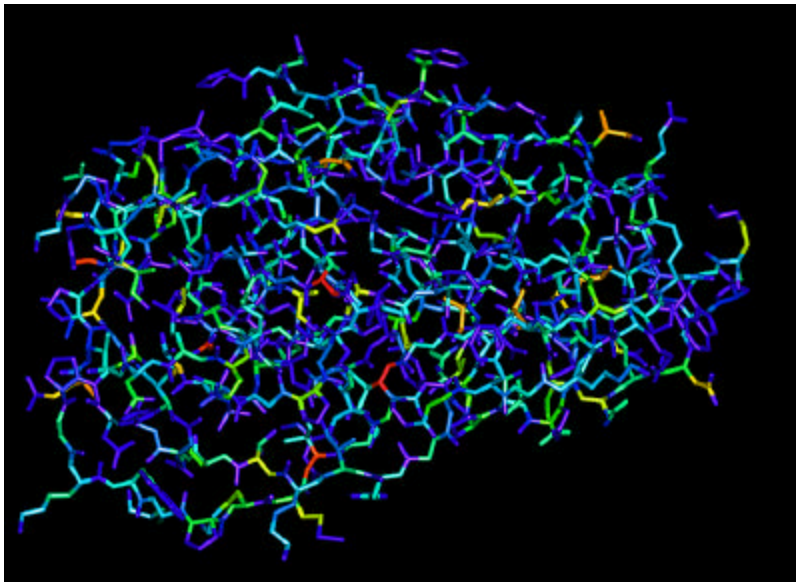
Procheck output



Procheck output - backbone G factors



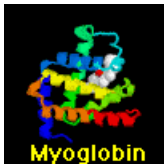
Procheck output - all atom G factors

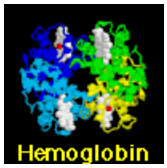


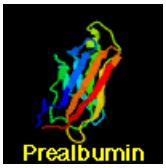
Secondary Structure Assignment

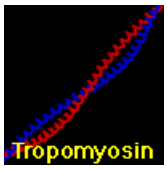
- DSSP
- Stride

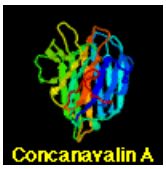
Structural classes of proteins











all α all β α/β

Protein Structure Classification

SCOP - Structural Classification of Proteins
http://scop.mrc-lmb.cam.ac.uk/scop/

FSSP - Fold classification based on Structure-Structure alignment of Proteins
http://www.ebi.ac.uk/dali/

CATH - Class, architecture, topology and homologous superfamily
http://www.cathdb.info/latest/index.html

SCOP: Structural Classification of Proteins

Essentially manual classification

Current release: 1.69
25973 PDB Entries (July 2005). 70859 Domains.

http://scop.mrc-lmb.cam.ac.uk/scop/

The **SCOP** database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy; the principal levels are family, superfamily and fold

Family: *Clear evolutionarily relationship*
Superfamily: *Probable common evolutionary origin*
Fold: *Major structural similarity*

SCOP: Structural Classification of Proteins

Family: *Clear evolutionarily relationship*

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absense of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.

SCOP: Structural Classification of Proteins

Superfamily: *Probable common evolutionary origin*

Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily.

SCOP: Structural Classification of Proteins

Fold: *Major structural similarity*

Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

SCOP Statistics

Class	**Folds**	**Super families**	**Families**
All alpha proteins	179	299	480
All beta proteins	126	248	462
Alpha and beta proteins (a/b)	121	199	542
Alpha and beta proteins (a+b)	234	349	567
Multi-domain proteins	38	38	53
Membrane and cell surface proteins	36	66	73
Small proteins	66	95	150
Total	**800**	**1294**	**2327**

FSSP Database

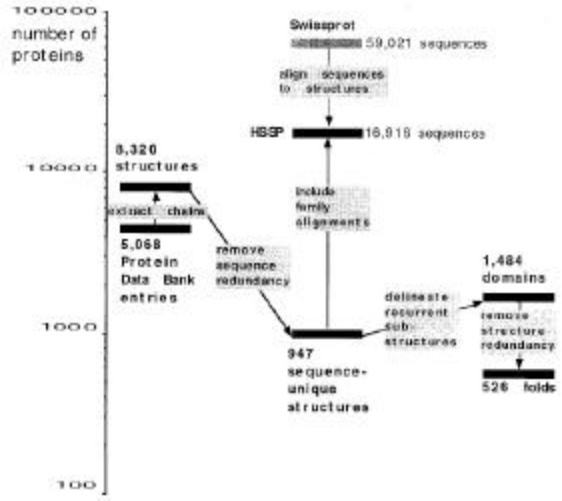
Essentially automated classification

Current release: September 2005

3724 sequence families representing 30624 protein structures

The FSSP database is based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB). The classification and alignments are automatically maintained and continuously updated using the Dali search engine.

Structure processing for Dali/FSSP



Adopted from Holm and Sander, 1998

Dali Domain Dictionary

http://www.ebi.ac.uk/dali/

Structural domains are delineated automatically using the criteria of recurrence and compactness. Each domain is assigned a Domain Classification number DC_l_m_n_p , where:

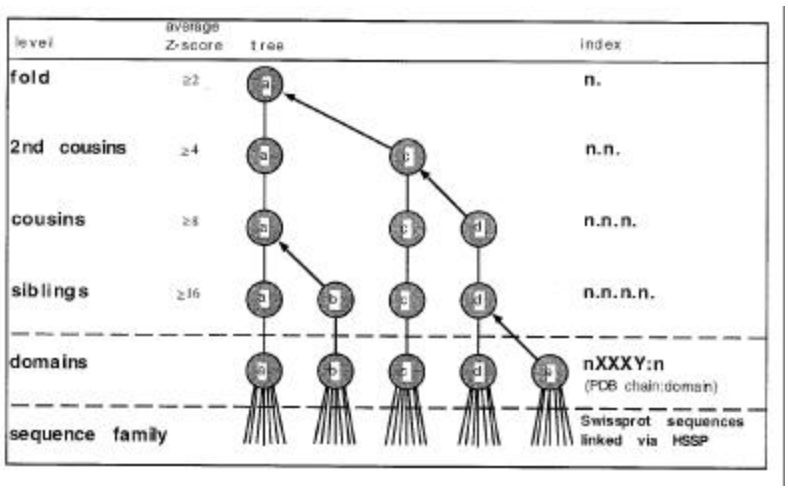
l - fold space attractor region

m - globular folding topology

n - functional family

p - sequence family

Hierarchical clustering of folds in Dali/FSSP

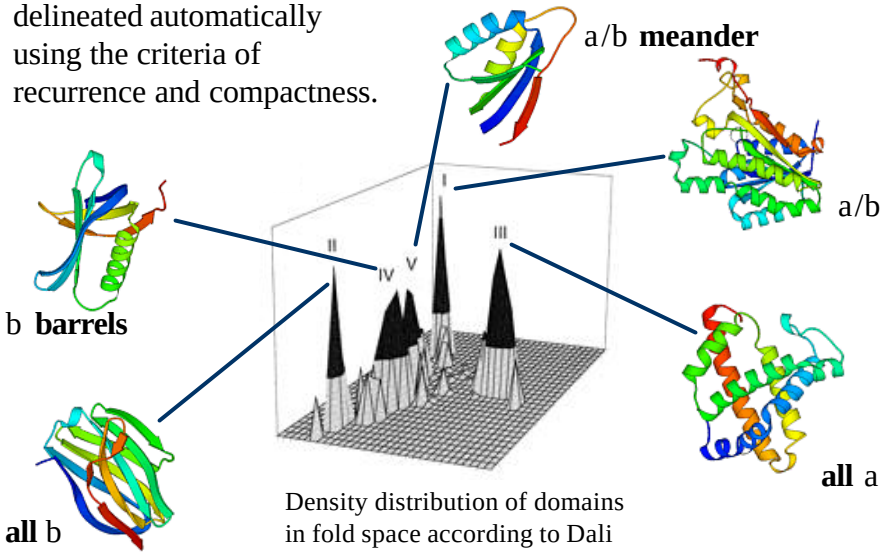


Adopted from Holm and Sander, 1998

Dali Domain Dictionary

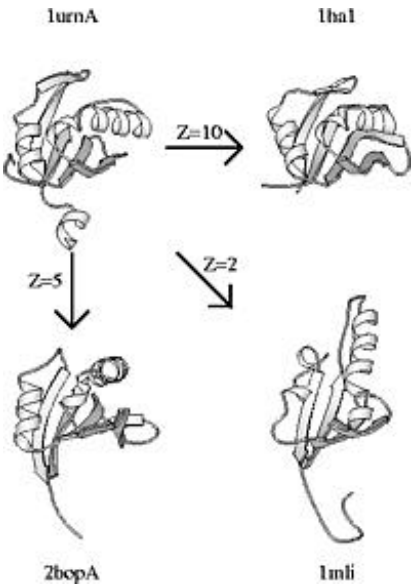
Structural domains are delineated automatically using the criteria of recurrence and compactness.

Fold space attractor regions



Density distribution of domains in fold space according to Dali

Dali Domain Dictionary



Fold types

Fold types are defined as clusters of structural neighbors in fold space with average pairwise Z-scores (by Dali) above 2.

Structural neighbours of 1urnA (top left). 1mli (bottom right) has the same topology even though there are shifts in the relative orientation of secondary structure elements

Dali Domain Dictionary

Functional families

The third level of the classification infers plausible evolutionary relationships from strong structural similarities which are accompanied by functional or sequence similarities. Functional families are branches of the fold dendrogram where all pairs have a high average neural network prediction for being homologous. The neural network weighs evidence coming from: overlapping sequence neighbours as detected by PSI-Blast, clusters of identically conserved functional residues, E.C. numbers, Swissprot keywords.

Dali Domain Dictionary

Sequence families

The fourth level of the classification is a representative subset of the Protein Data Bank extracted using a 25 % sequence identity threshold. All-against-all structure comparison was carried out within the set of representatives. Homologues are only shown aligned to their representative.

CATH - Protein Structure Classification

Combines manual and automated classification

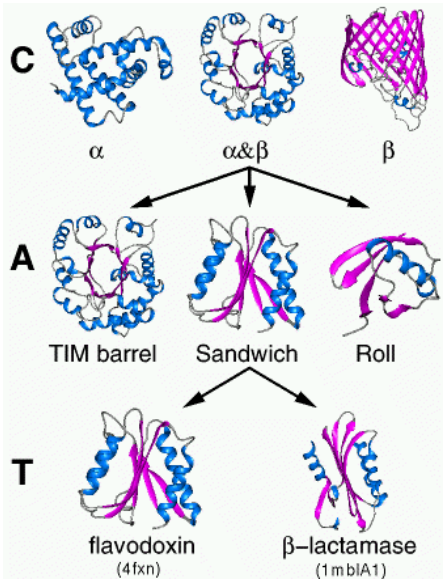
Current release: 2.6.0 (April 2005)

http://www.cathdb.info/latest/index.html

CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels:

Class
Architecture
Topology
Homologous superfamily

CATH - Protein Structure Classification



CATH - Protein Structure Classification

Class, C-level

Class is determined according to the secondary structure composition and packing within the structure. It can be assigned automatically (90% of the known structures) and manually.

Three major classes:

- mainly-alpha
- mainly-beta
- alpha-beta (alpha/beta and alpha+beta)

A fourth class is also identified which contains protein domains which have low secondary structure content.

CATH - Protein Structure Classification

Architecture, A-level

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures.

It is currently assigned manually using a simple description of the secondary structure arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g the beta-propellor or alpha four helix bundle).

Procedures are being developed for automating this step.

CATH - Protein Structure Classification

Topology (Fold family), T-level

Structures are grouped into fold families at this level depending on both the overall shape and connectivity of the secondary structures. This is done using the structure comparison algorithm SSAP.

Some fold families are very highly populated and are currently subdivided using a higher cutoff on the SSAP score.

CATH - Protein Structure Classification

Homologous Superfamily, H-level

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified first by sequence comparisons and subsequently by structure comparison using SSAP.

Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria:

- Sequence identity >= 35%, 60% of larger structure equivalent to smaller
- SSAP score >= 80.0 and sequence identity >= 20% 60% of larger structure equivalent to smaller
- SSAP score >= 80.0, 60% of larger structure equivalent to smaller, and domains which have related functions

CATH - Protein Structure Classification

Sequence families, S-level

Structures within each H-level are further clustered on sequence identity. Domains clustered in the same sequence families have sequence identities >35% (with at least 60% of the larger domain equivalent to the smaller), indicating highly similar structures and functions.

Predicting Protein Structure from the Amino Acid Sequence

- Goal: Predict the 3-dimensional (tertiary) structure of a protein from the sequence of amino acids (primary structure).
- Sequence similarity methods predict secondary and tertiary structure based on homology to know proteins.
- Secondary structure predictions methods include Chou-Fasman, GOR, neural network, and nearest neighbor methods.
- Tertiary structure prediction methods include energy minimization, molecular dynamics, and stochastic searches of conformational space.

Evolutionary Methods

Taking into account related sequences helps in identification of “structurally important” residues.

Algorithm:

- find similar sequences
- construct multiple alignment
- use alignment profile for secondary structure prediction

Additional information used for prediction

- mutation statistics
- residue position in sequence
- sequence length

Sequence similarity methods for structure prediction

- These methods can be very accurate if there is > 50% sequence similarity.
- They are rarely accurate if the sequence similarity < 30%.
- They use similar methods as used for sequence alignment such as the dynamic programming algorithm, hidden markov models, and clustering algorithms.

Secondary Structure Prediction Algorithms

- These methods are 70-75% accurate at predicting secondary structure.
- A few examples are
  - Chou Fasman Algorithm
  - Garnier-Osguthorpe-Robson (GOR) method
  - Neural network models
  - Nearest-neighbor method

Secondary Structure Prediction

Three-state model: helix, strand, coil

Given a protein sequence:

- `NWVLSTAADMQGVVTDGMASGLDKD...`

Predict a secondary structure sequence:

- `LLEEEELLLLHHHHHHHHHHLHHHL...`

Methods:

- statistical
- stereochemical

Accuracy: 50-85%

Statistical Methods

Residue conformational preferences:

Glu, Ala, Leu, Met, Gln, Lys, Arg -	helix
Val, Ile, Tyr, Cys, Trp, Phe, Thr -	strand
Gly, Asn, Pro, Ser, Asp -	turn

Chou-Fasman algorithm:

Identification of helix and sheet "nuclei"
Propagation until termination criteria met

Chou-Fasman Algorithm

- Analyzed the frequency of the 20 amino acids in α helices, β sheets and turns.
- Ala (A), Glu (E), Leu (L), and Met (M) are strong predictors of α helices.
- Pro (P) and Gly (G) break α helices.
- When 4 of 5 amino acids have a high probability of being in an α helix, it predicts a α helix.
- When 3 of 5 amino acids have a high probability of being in a β strand, it predicts a β strand.
- 4 amino acids are used to predict turns.

Garnier-Osguthorpe-Robson Method

- Chou-Fasman assumes that each individual amino acid influences secondary structure.
- GOR assumes the the amino acids flanking the central amino acid also influence the secondary structure.
- Hence, it uses a window of 17 amino acids (8 on each side of the central amino acid).
- Each amino acid in the window acts independently on influencing structure (to save computational time).
- Certain pair-wise combinations of amino acids in the window also contribute to influencing structure.

Garnier - Osguthorpe - Robson (GOR) Algorithm

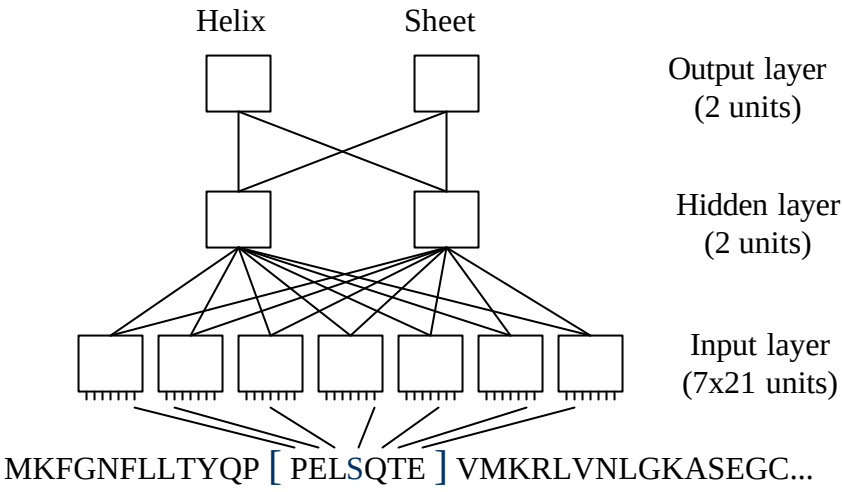
Likelihood of a secondary structure state depends on the neighboring residues:

$$L(S_j) = \Sigma\ (S_j;R_{j+m})$$

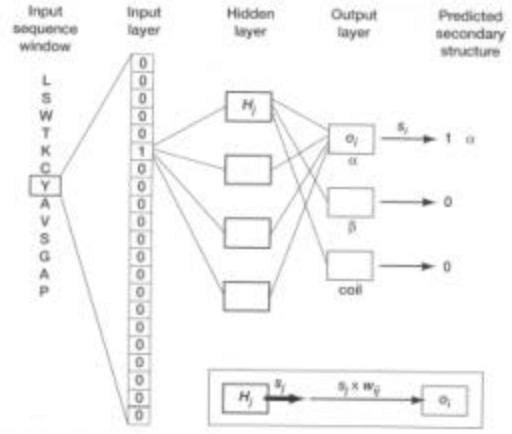
Window size - [j-8; j+8] residues

Accuracy for a single sequence - 60%
Accuracy for an alignment - 65%

Neural Networks Methods



Rost and Sander Neural Network Model



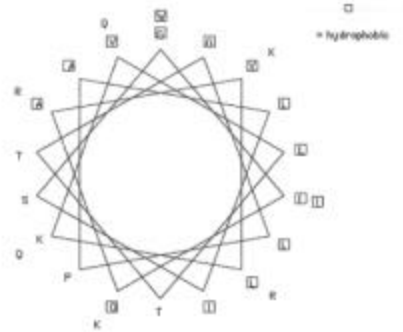
From *Bioinformatics: Sequence and Genome Analysis*
by David Mount

Nearest Neighbor Method

- Like neural networks, this is another machine learning approach to secondary structure prediction.
- A very large list of short sequence fragments is made by sliding a window (n=16) along a set of 100-400 training sequences of know structure but with minimal similarity.
- A same-size window is selected from the query sequence and the 50 best matching sequences are found.
- The frequencies of the of the secondary structure of the middle amino acid in each of the matching fragments is used to predict the secondary structure of the middle amino acid in the query window.
- Can be very accurate (up to 86%).

Hydrophobicity/Hydrophilicity Plots

- Charge amino acids are hydrophilic, i.e. Asp (D), Glu (E), Lys (K), Arg (R).
- Uncharged amino acids are hydrophobic, i.e. Ala (A), Leu (L) Ile (I), Val (V), Phe (F), Trp (W), Met (M), Pro (P).
- In an α helix, hydrophobic amino acids might line up on one side, which suggests that that side is on the interior of a protein or protein complex.



From *Bioinformatics: Sequence and Genome Analysis* by David Mount – Helicalwheel plot by GCG

Stereochemical Methods

Patterns of hydrophobic and hydrophilic residues in secondary structure elements:

- segregation of hydrophobic and hydrophilic residues
- hydrophobic residues in the positions 1-2-5 and 1-4-5
- oppositely charged polar residues in the positions 1-5 and 1-4 (e.g. Glu (i), Lys (i+4))

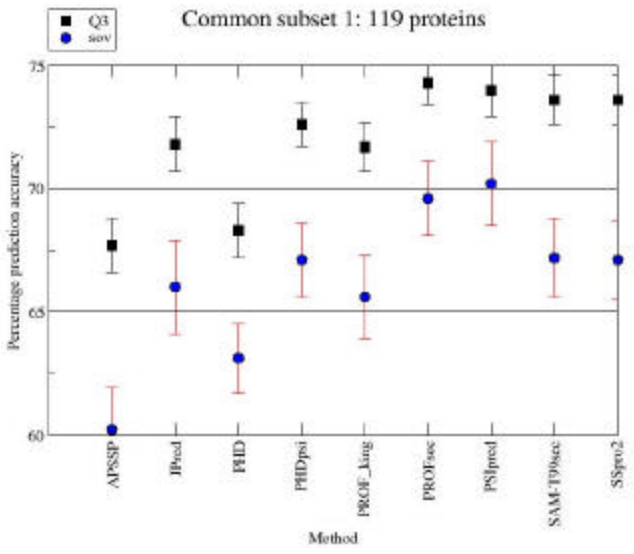
Definitions of hydrophobic and hydrophilic residues (hydrophobicity scales) are ambiguous

Stereochemical Methods

Hydropathic correlations in helices and sheets

		F-F	F-L	L-F	L-L
a	*i, i+2*	-	+	+	-
	i, i+3	+	-	-	+
	i, i+4	+	-	-	+
	i, i+5	-	+	+	-
b	*i, i+1*	-	+	+	-
	i, i+2	+	-	-	+
	i, i+3	-	+	+	-

Accuracy of prediction



EVA (http://cubic.bioc.columbia.edu/eva/)

Accuracy of Prediction

$$Q_3 = \frac{PH + PE + PC}{N}$$

$$W = \log \frac{TP \times TN}{FP \times FN}$$

Range: 50-85%

Energy Potential Functions

- Contains terms for electrostatic interatction, van der Wals forces, hydrogen bonding, bond angle and bond length energies.
- Common software packages have their own implementation: Charmm, ECEPP, Amber, Gromos, and CVF.
- Structural predictions only as good as the assumptions upon which it is based (mainly the energy potential function).

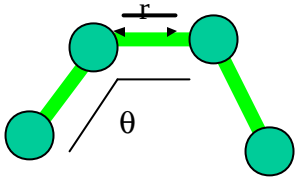
Bonded Terms

Bond Length

$$E_{\text{bond-length}} = \Sigma_{\text{bonds}}\, k_b(r - r_0)^2$$

Bond Angle

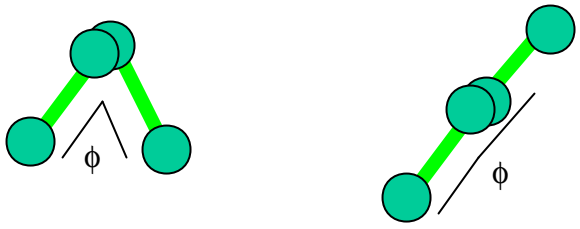
$$E_{\text{bond-angle}} = \Sigma_{\text{angle}}\, k_\theta (\theta - \theta_0)^2$$



Bonded Terms

Dihedral Angle

$$E_{\text{dihedral-angle}} = \Sigma_{\text{dihedrals}}\, K_{\phi}\,(1 + \cos\,[n\phi(R)-\gamma]$$



Non-Bonded Terms

Lennard-Jones potential (van der Waals force)

$$E_{vdW} = \Sigma_{i,j}\, \underset{\text{repulsive}}{A_{ij}/r_{ij}^{12}} - \underset{\text{dispersion}}{B_{ij}/r_{ij}^{6}}$$

Electrostatic interactions

$$E_{elec} = \Sigma^{r}_{i,j}\, q_i q_j/(4\pi\varepsilon_0\varepsilon_r r_{ij})$$

ε_0 = permittivity of free space
ε_r = dielectric constant of medium around charges

Non-Bonded Terms

Hydrogen Bonding – Some atoms (O, N, and to a lesser degree S) are electronegative, i.e. the attact electrons to fill their valence shells. Hydrogen tends to donate electrons to these atoms forming hydrogen bonds. This is common in water.

Salt Bridges – A positively charged lysine or arginine residue can form a strong interaction with a negatively charged aspartic acid or glutamic acid residue.

Energy Minimization

- Assumes that proteins are found at or near the lowest energy conformation.
- Uses a empirical function that describes the interaction of different parts of the protein with each other (energy potential function).
- Searches conformation space to find the global minimum using optimization techniques such as steepest descents and conjugate gradients.
- To avoid the multiple-minima problem, approaches such as dynamic programming, or simulated annealing have been used.

Molecular Dynamics

$F_i = m_i a_i$ Motion	force by Newton's Second Law of
$a_i = dv_i/dt$	acceleration
$v_i = dr_i/dt$	velocity
$-dE/dr_i = F_i$	Work = force x distance
$-dE/dr_i = m_i d^2r_i/dt^2$	put it all together

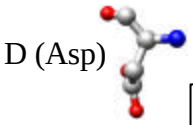
Molecular Dynamics

- Model System – Choose protein model, energy potential function, ensemble, and boundary conditions.
- Initial Conditions – Need initial positions of the atoms, an initial distribution of the velocities (assume no momentum i.e. $\Sigma_i m_i v_i = 0$), and the acceleration which is determined by the potential energy function.
- Boundary Conditions – If water molecules are not being explicitly included in the potential function, the solvent boundary conditions must be imposed. The water molecules must not diffuse away from the protein. Also, usually a limited number of solvent molecules are included.

Molecular Dynamics

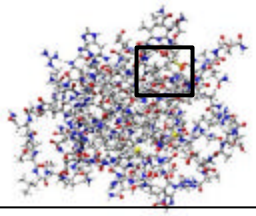
- Result
  - The result of the simulation is a time series of the trajectories (path) followed by the atoms governed by Newton's law of motion.
  - The time scales are usually very small (picoseconds).
  - The motion of the molecule can be seen.
  - The motion will move the atoms into the near-equilibrium conformation of the protein.

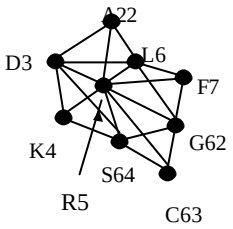
Delaunay Tessellation of Protein Structure



D (Asp) → C_a or center of mass

Abstract each amino acid to a point
~~Atomic coords – Protein Data Bank (PDB)~~





Delaunay tessellation: 3D "tiling" of space into non-overlapping, irregular tetrahedral simplices. Each simplex objectively defines a quadruplet of nearest-neighbor amino acids at its vertices.

Counting Amino Acid Quadruplets

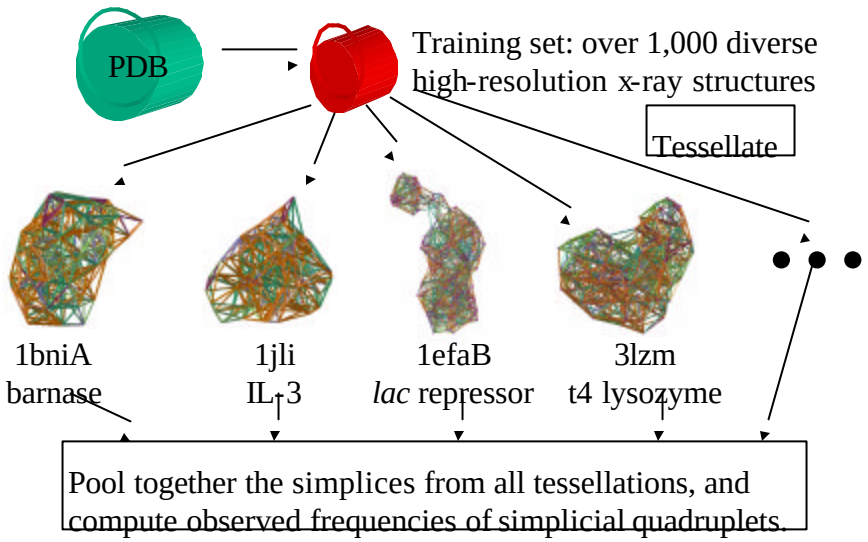
Ordered quadruplets: $20^4 = 160{,}000$ (too many)

Order-independent quadruplets (our approach):

C D E F	$\binom{20}{4}$
C C D E	$20 \cdot \binom{19}{2}$
C C D D	$\binom{20}{2}$
C C C D	$20 \cdot 19$
C C C C	20

Total: 8,855 distinct unordered quadruplets

Four-Body Statistical Potential



Four-Body Statistical Potential

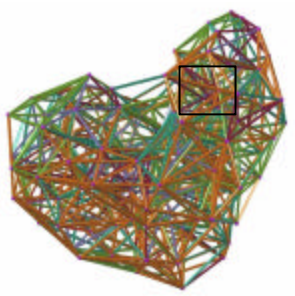
- Modeled after Boltzmann potential of mean force:

  $? E_i = -KT \ln(p_i / p_{ref})$

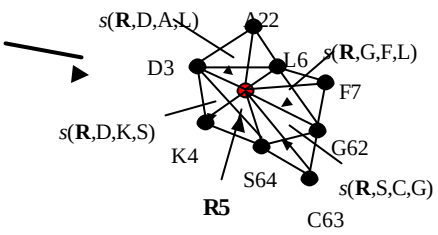
- For amino acid quadruplet (i,j,k,l), a log-likelihood score ("pseudo-energy") is given by $s(i,j,k,l) = \log(f_{ijkl} / p_{ijkl})$
- f_{ijkl} = proportion of training set simplices whose four vertex residues are i,j,k,l
- p_{ijkl} = rate expected by chance (multinomial distribution, based on training set proportions of residues i,j,k,l)
- Four-body statistical potential: the collection of 8855 quadruplet (or simplex) types and their respective log-likelihood scores

Application: Protein Topological Score

- Global measure of sequence-structure compatibility
- Obtained by summing log-likelihood scores of **all** simplicial quadruplets defined by the tessellation

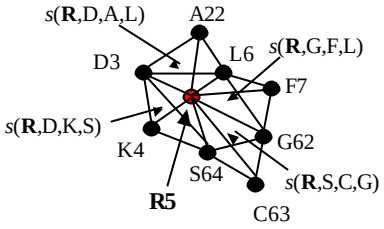


S = ? $_{\hat{\imath}}$ s(**î**), sum taken over **all** simplex quadruplets **î** in the entire tessellation.



Application: Residue Environment Scores

- For each amino acid position, locally sum the scores of only simplices that use the amino acid point as a vertex

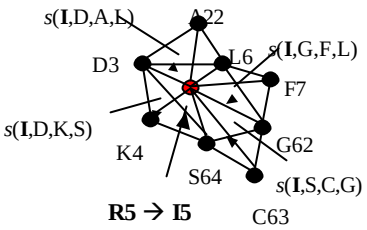


$q_5 = q(\mathrm{R5}) = ?_{\hat{\imath}}\, s(\hat{\imath})$, sum taken over only simplex quadruplets $\hat{\imath}$ that contain amino acid R5

- The scores of all the amino acid positions in the protein structure form a **Potential Profile** vector $Q = < q_1,\ldots,q_N >$ (N = length of primary sequence in the solved structure)

Computational Mutagenesis Methodology

- Observation: mutant and wild type (wt) protein structure tessellations are very similar or identical
- Approach: obtain mutant topological score and potential profile from wt structure tessellation, by changing residue labels at points and re-computing



- **Residual Score** = mutant – wt topological scores = $S_{mut} - S_{wt}$
- **Residual Profile** = mutant – wt potential profiles = $Q_{mut} - Q_{wt}$
- For single point mutants only, residual score ° residual profile component at the mutant position