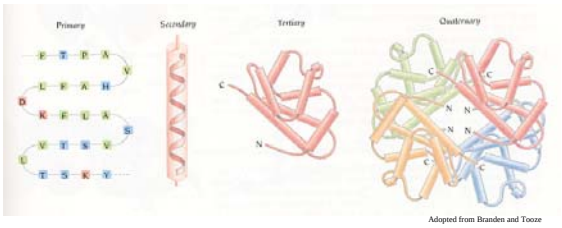
BINF 731

Protein Structure Analysis

Iosif Vaisman

2004

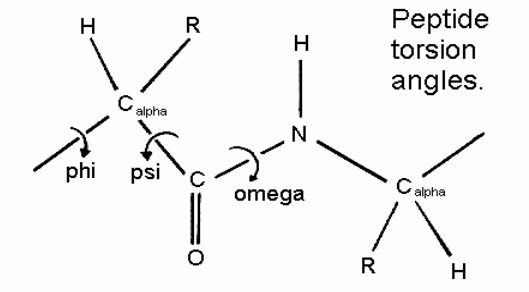
Protein Structure Hierarchy



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.

Secondary Structure Conformations



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

Secondary Structure Prediction

Three-state model: helix, strand, coil

Given a protein sequence:
- `NWVLSTAADMQGVVTDGMASGLDKD...`

Predict a secondary structure sequence:
- `LLEEEELLLLHHHHHHHHHHHLHHHL...`

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 Parameters

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

Chou-Fasman Algorithm

Identification of helix and sheet "nuclei"

- helix - 4 out of 6 residues with high helix propensity (P > 100)
- sheet - 3 out of 5 residues with high sheet propensity (P > 100)

Propagation until termination criteria met

Turn prediction

1) `p(t) > 0.000075`
2) `P(turn) > 1.00`
3) `P(a) < P(turn) > P(b)`

where `p(t) = f(j)f(j+1)f(j+2)f(j+3)`

P.Y. Chou, G.D. Fasman, *Biochemistry*, 1974, **13**, 211-222

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%

Evolutionary Methods

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

Algorithm:

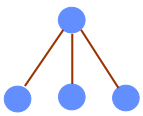
- find similar sequences
- construct multiple alignment
- use alignment profile for secondary structure prediction

Additional information used for prediction

- mutation statistics
- residue position in sequence
- sequence length

Neural Networks

Perceptron



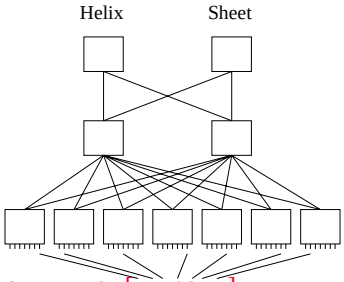
Output layer

Input layer

$$Y = \begin{cases} 1 \text{ if } \Sigma\ w_i i_i > \Theta \\ 0 \text{ otherwise} \end{cases}$$

Learning process: $\Delta w_i = (T_p - Y_p)i_{pi}$

Neural Networks Methods



Stereochemical Methods

Patterns of hydrophobic and hydrophilic residues in secondary structure elements:

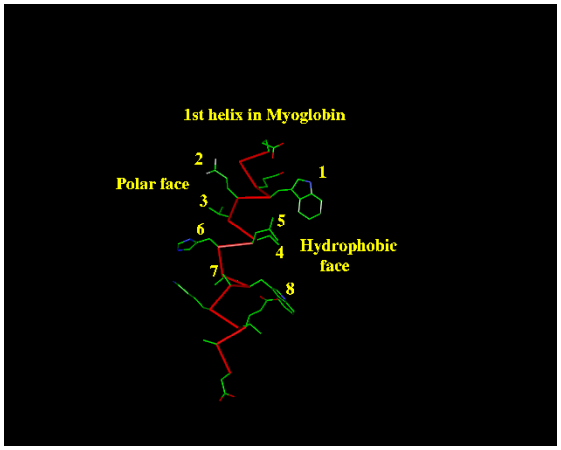
- segregation of hydrophobic and hydrophilic residues
- hydrophobic residues in the positions 1-2-5 and 1-4-5
- oppositely charged polar residues in the positions 1-5 and 1-4 (e.g. Glu (i), Lys (i+4))

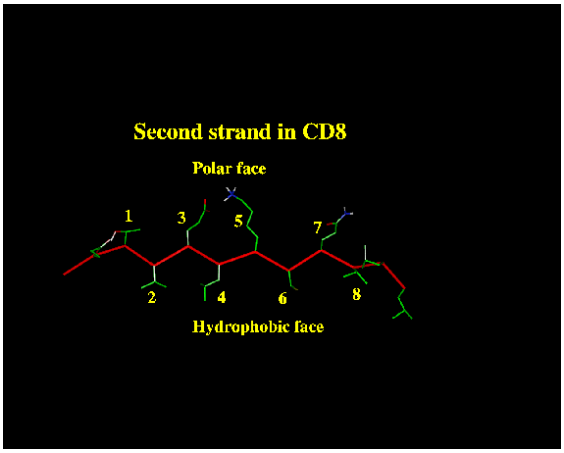
Definitions of hydrophobic and hydrophilic residues (hydrophobicity scales) are ambiguous

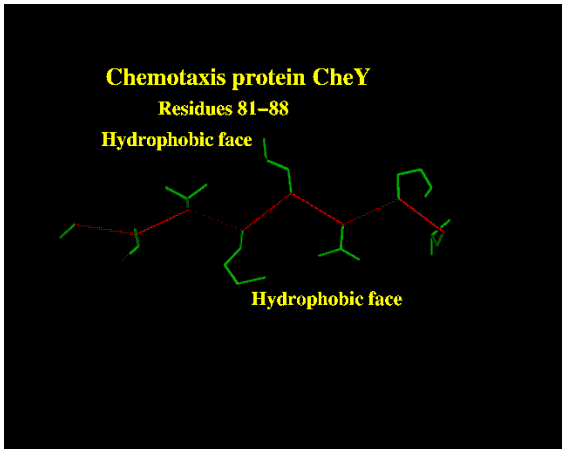
Stereochemical Methods

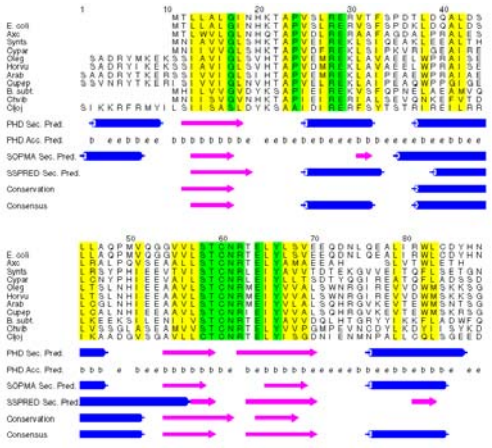
Hydropathic correlations in helices and sheets

		F-F	F-L	L-F	L-L
α	*i, i+2*	-	+	+	-
	i, i+3	+	-	-	+
	i, i+4	+	-	-	+
	i, i+5	-	+	+	-
β	*i, i+1*	-	+	+	-
	i, i+2	+	-	-	+
	i, i+3	-	+	+	-

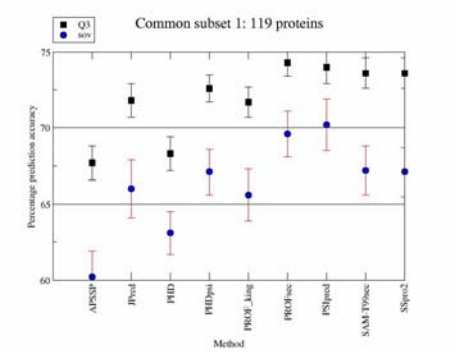








Accuracy of prediction



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

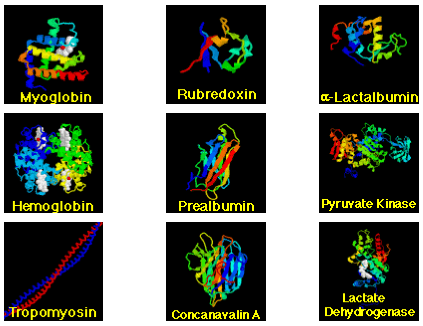
Accuracy of Prediction

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

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

Range: 50-85%

Structural classes of proteins



all α all β α/β

Protein Structure Classification

SCOP - Structural Classification of Proteins

FSSP - Fold classification based on Structure-Structure alignment of Proteins

CATH - Class, architecture, topology and homologous superfamily

SCOP: Structural Classification of Proteins

Current release: 1.65
20619 PDB Entries (1 Aug 2003). 54745 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

Current release: May 2003
3242 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.

Dali Domain Dictionary

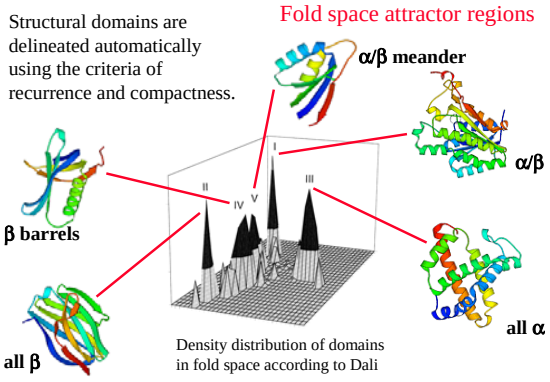
http://www2.ebi.ac.uk/dali/domain

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:

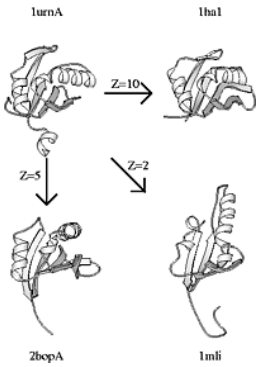
- l - fold space attractor region
- m - globular folding topology
- n - functional family
- p - sequence family

Dali Domain Dictionary

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



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

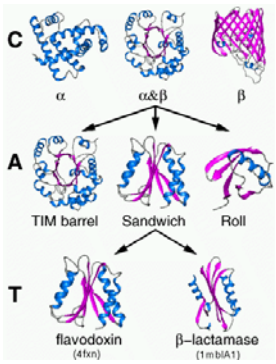
Current release: 2.5.1 (Jan 2004)

http://www.biochem.ucl.ac.uk/bsm/cath_new/

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.