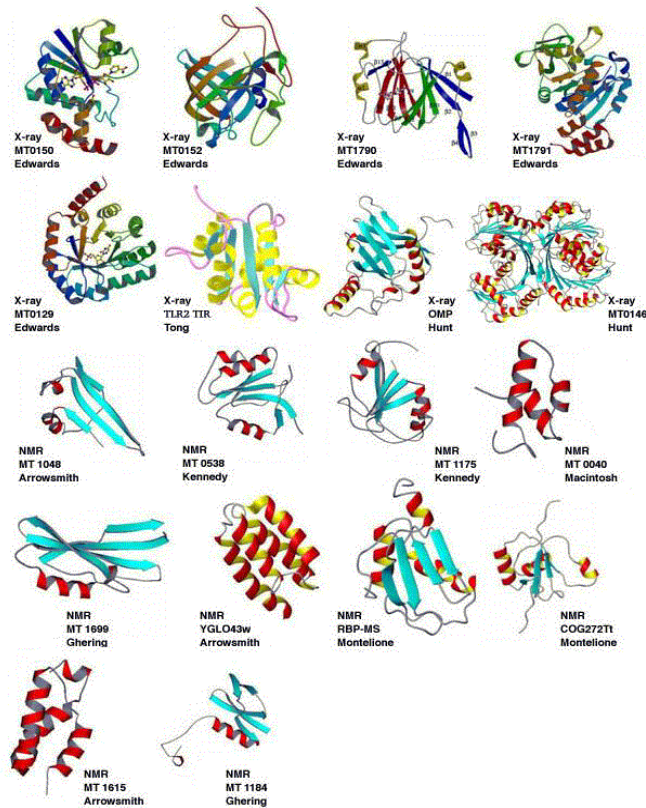


The Protein Space

From Sequence to Structure to Function



Michal Linial
Life Sciences Institute
The Hebrew University
Jerusalem, Israel



January, 2003



Protein Sequences

1,000,000 pr
(static)

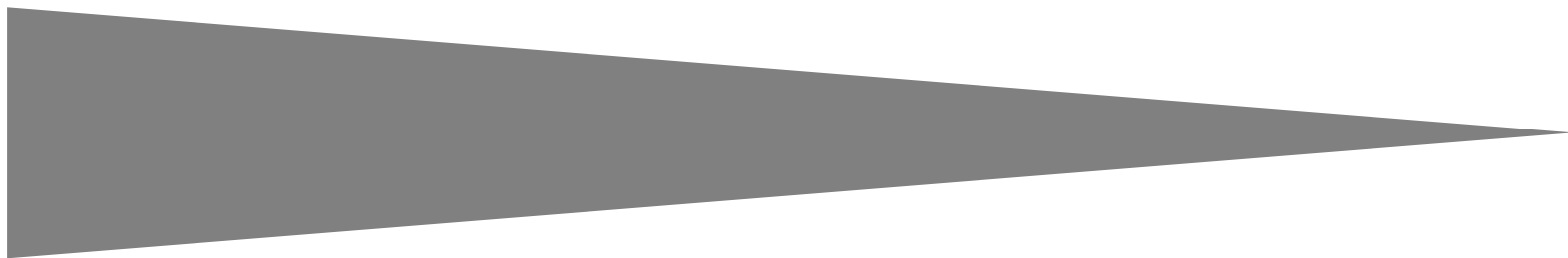
Protein Variants

10,000,000 pr
(dynamic)

Exon combinations, post-translation modification, p-p interaction...

Protein Function

?????



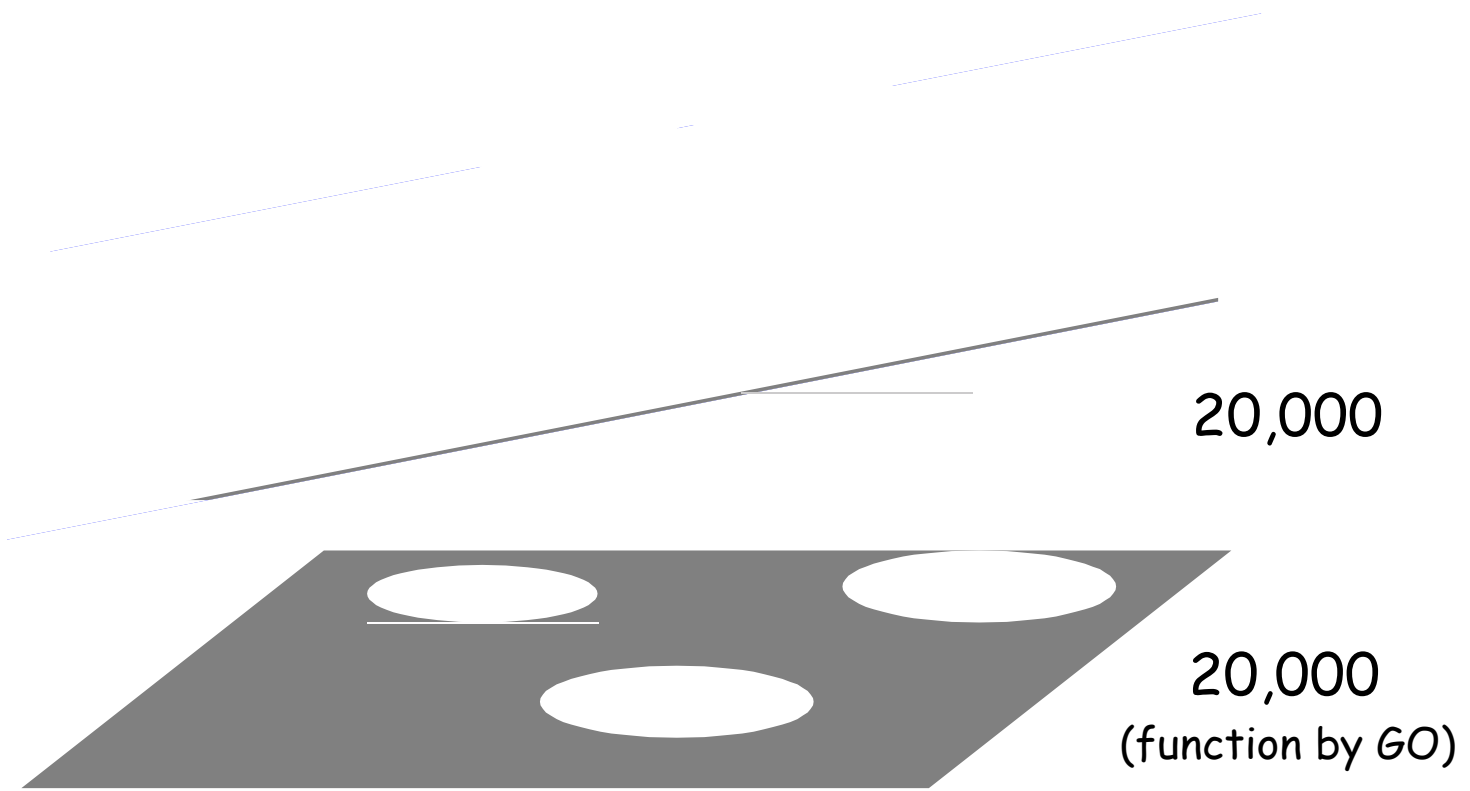
A link between sequence, structure and function

Protein **structures** are much more conserved than protein **sequences**

Proteins of identical (similar) **Structure** tend to have identical (similar) **function**

Extract **structural** information from **sequence** alone
(The Holy Grail)

Assign **function** based on **structural** characterization



January, 2003

The Scheme of the talk

THE TASK: Construct a map of the protein space

ProtoClass -rationale and concept

ProtoNet in brief

AND BEYOND: Functional roadmap in ProtoClass

Biological examples

THE PURPOSE: New superfamilies for SG

ProTarget - ranked list of proteins

THE APPLICATION: AraNet - SG for Arabidopsis

QUO VADIS: Quality assessment of clusters for FG

vis-à-vis InterPro, SCOP, FSSP etc

ProtoClass - Set of automatic classifications of all proteins

Seeking statistically significant regularities (clusters)
Reconstruct the 'geometry' of the sequence space



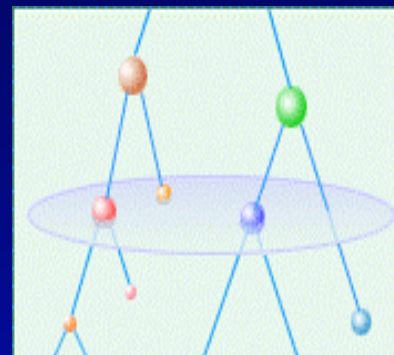
Guiding principle

Homologous proteins evolved from common ancestor protein

Homology is a transitive relation that can be deduced based on statistical similarities

ProtoClass - Set of classifications of all proteins

ProtoClass systems generate graphs and maps that yield views at any levels of granularity.



ProtoMap



release May 1997

ProtoNet - A (arithmetic)

release July 2002

ProtoNet - G (geometric)

release July 2002

ProtoNet - H (harmonic)

release July 2002

January, 2003



ProtoNet

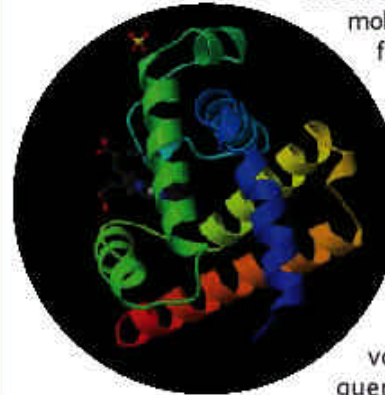
ProtoNet - Ver 2.1

Dec 2002

DATABASE

Parse Protein Pedigrees

A new Web site from Hebrew University in Jerusalem aims to simplify the analysis of protein structure and function. Along with the usual sequence information, ProtoNet automatically clusters proteins by similarity, creating a family tree that allows researchers to compare individual proteins or related groups. For more than 100,000 proteins,



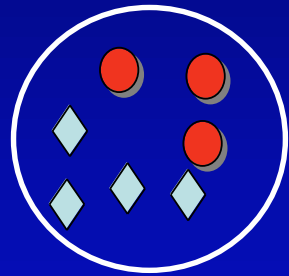
the site holds a data card that lists each molecule's amino acid sequence, identifies functional regions, and charts the taxonomy of the organism it comes from. You can compare each protein to other members of its immediate family or climb up the tree to contrast different groups, which might help deduce the function of mystery molecules or tease out evolutionary trends. If you don't find your favorite protein here, submit its sequence to find out how it fits into known clusters.

www.protonet.cs.huji.ac.il/protonet/index.php

Science 298 : p329 (11 October 2002)

www.protonet.cs.huji.ac.il

Protein Sequence-Structure Space



ProtoNet - in brief

ProtoNet.cs.huji.ac.il
automatic hierarchical classification of proteins

Version 1.4

main page
search
classify your protein
introduction
methods
guided tour

related links
ProtoNet team
help
feedback

search _____

Individual protein

Get protein card
Returns information about a protein.

Protein Sequence Alignment
Displays sequence alignment for two proteins.

Check protein in cluster
Testing whether a protein belongs to a given cluster.

Vertical perspective

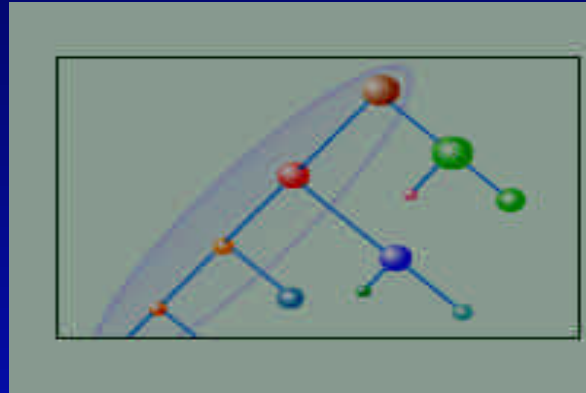
Get cluster card
Returns information about a ProtoNet cluster.

January, 2003

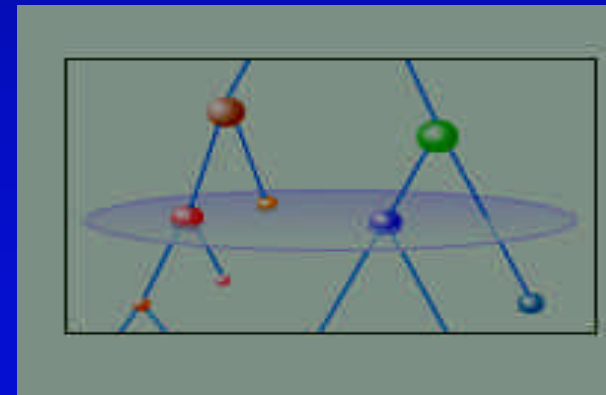
ProtoNet Perspectives



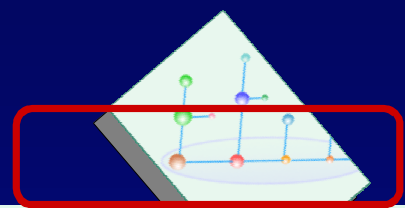
Clustering Chain (to the root - subfamilies)



Horizontal (to the correct level - maps)



Clustering Chain



ProtoNet.cs.huji.ac.il
automatic hierarchical classification of proteomes

Version 1.4 Type of classification "Geometric"

Class: 749494

main page search classify your proteome introduction methods guided tour related links ProtoNet team help feedback

73 176639 175662 9661 160338 9930 136448 0.0 148484 146012 0.0

size: 4 size: 70 size: 1 size: 4 size: 1

8: 76,848 41,926 8,187 2,495 0.0 0.0

9641 12126 18200 18400 18400 18400

136448 0.0 148484 146012 0.0

ProtoLevel

No. of clusters

Get cluster info Get neighbors

Number of proteins in cluster 148484:

Total proteins	Number of proteins of child 1	Number of proteins of child 2
54	50	4

θ - with PDB
θ - no Protein ID
θ - hypothetical proteins
? - are fragments

Get Keywords Appearances (select a type)

Sweepnet + ProtoNet keywords

Proteins of cluster 148484.
Sorted by "(Depth)"

Cluster G140494
Cluster G9029
Protein 9629 (CCAA_MOUSE)

Root

How Pure is your cluster?

Keywords of cluster 148484
Total number of proteins in this cluster: 54

Keywords of type "InterPro"						
Keyword description	Number of proteins in cluster with this keyword	Keyword Deviation from Expectation	Keyword frequency among proteins (%)			General keyword frequency (%)
			in this cluster	in child1 136448 (size: 50)	in child2 146012 (size: 4)	
Calcium and sodium channel pore region (S4-S6) IPR001682	54	275.44	100.0	100.0	100.0	0.07
Cation channels (non-ligand gated) IPR000636	54	174.9	100.0	100.0	100.0	0.17
Cation channels TM region (not potassium) IPR002111	51	260.13	94.44	96.0	75.0	0.07
Calcium channel IPR002077	38	257.37	70.37	68.0	100.0	0.04
Sodium channel IPR001696	15	161.68	27.77	30.0	0	0.01
IQ calmodulin-binding motif IPR000048	11	42.24	20.37	22.0	0	0.12
EF-hand family IPR002048	5	7.84	9.25	9.99	0	0.65
HMG-I and HMG-Y DNA-binding domain (A+T-hook) IPR000637	3	22.01	5.55	5.99	0	0.03
Voltage-dependent potassium channel IPR003091	1	6.29	1.85	1.99	0	0.04

Additional options
complex queries..
Horizontal view..

January, 2003

Microsoft Internet Explorer window: **List of proteins - Microsoft Internet Explorer**

Address bar: <http://www.proteinatlas.org/ptmsearch.html?cluster=148484&searchTerm=&proteinid=136448>

Get Protein Sequence Alignment

Proteins of cluster "148484"
[entry 1 - 136448, entry 2 - 146012]

Sorted by "Depth"

Choose two proteins from list.
(You can press CTRL+F to make search on this page)

No.	Protein ID	Swissprot ID	Belongs to child	Check protein
1	9627	CCAA_DROME	child 1	<input type="checkbox"/>
2	9634	CCAB_DISOM	child 1	<input type="checkbox"/>
3	9642	CCAC_HUMAN	child 1	<input type="checkbox"/>
4	9628	CCAA_HUMAN	child 1	<input type="checkbox"/>
5	9645	CCAC_RAT	child 1	<input type="checkbox"/>
6	9630	CCAA_RABIT	child 1	<input type="checkbox"/>
7	9631	CCAA_RAT	child 1	<input type="checkbox"/>
8	9655	CCAE_RABIT	child 1	<input type="checkbox"/>
9	9638	CCAB_RAT	child 1	<input type="checkbox"/>
10	9663	CCAM_MOUSE	child 1	<input type="checkbox"/>
11	9646	CCAD_CHICK	child 1	<input type="checkbox"/>
12	9657	CCAF_HUMAN	child 1	<input type="checkbox"/>
13	9667	CCAG_HUMAN	child 1	<input type="checkbox"/>
14	9656	CCAE_RAT	child 1	<input type="checkbox"/>
15	9635	CCAB_HUMAN	child 1	<input type="checkbox"/>
16	9637	CCAB_RABIT	child 1	<input type="checkbox"/>

Protein ID	Swissprot ID	Accession	Length	Weight	Isotype	Parent Level	No. of clusters
9653	CCAE_HUMAN	P33026, P33027, P33028	2312	17	13648		
9547	CCFL_RAT	P33026, P33027, P33028	2009	16	13648		
9545	CCFL_HUMAN	P33026, P33027, P33028	2009	16	13648		
9688	CCAS_RABIT	P33026, P33027, P33028	1873	16	13648		

Cluster: G50000
Cluster: G9629
Protein: 9520 [CCAA_MOUSE]

history

Internet

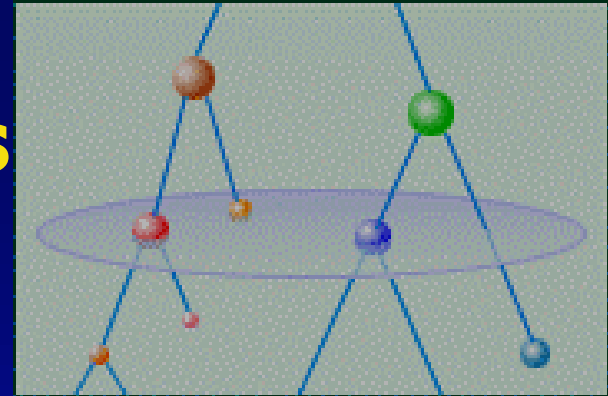
Taskbar: Start, Deleted 2..., myscripte..., answers..., PROTOONE..., Microsoft..., PROTOONE, http://www..., Protein, List of pr..., 14:18

ProtoNet top 20

<i>S/N</i>	<i>Cluster ID</i>	<i>Size</i>	<i>Family</i>
1	176127	995	GPCR
2	176194	790	Kinase
3	176876	725	Homoebox
4	176008	590	Cytochrome P450
5	176252	527	Cytochrome B/B6
6	174059	524	★ ABC transporter **
7	176475	512	★ Intermediate filament *
8	176004	467	Zinc finger, C2H2
9	176689	459	Ras GTPase
10	175479	434	GTP-binding elongation
11	176624	426	Rubisco
12	176400	416	EF-hand
13	176407	405	Tyrosine kinase
14	176639	380	Immunoglobulin C-type
15	176531	374	AAA ATPase
16	174961	354	NADH-Ubiquinone
17	176445	347	★ Short Dehydrogenase *
18	173944	343	ATP synthase
19	176025	342	Serine proteases
20	176946	336	★ G-protein beta **

20 largest clusters in the ProtoNet (Arithmetic) tree at pre-selected horizontal level: Clusters including substantial fraction of hypothetical proteins. 7-15% (*), 15-20% (**).

ProtoClass Road-Maps



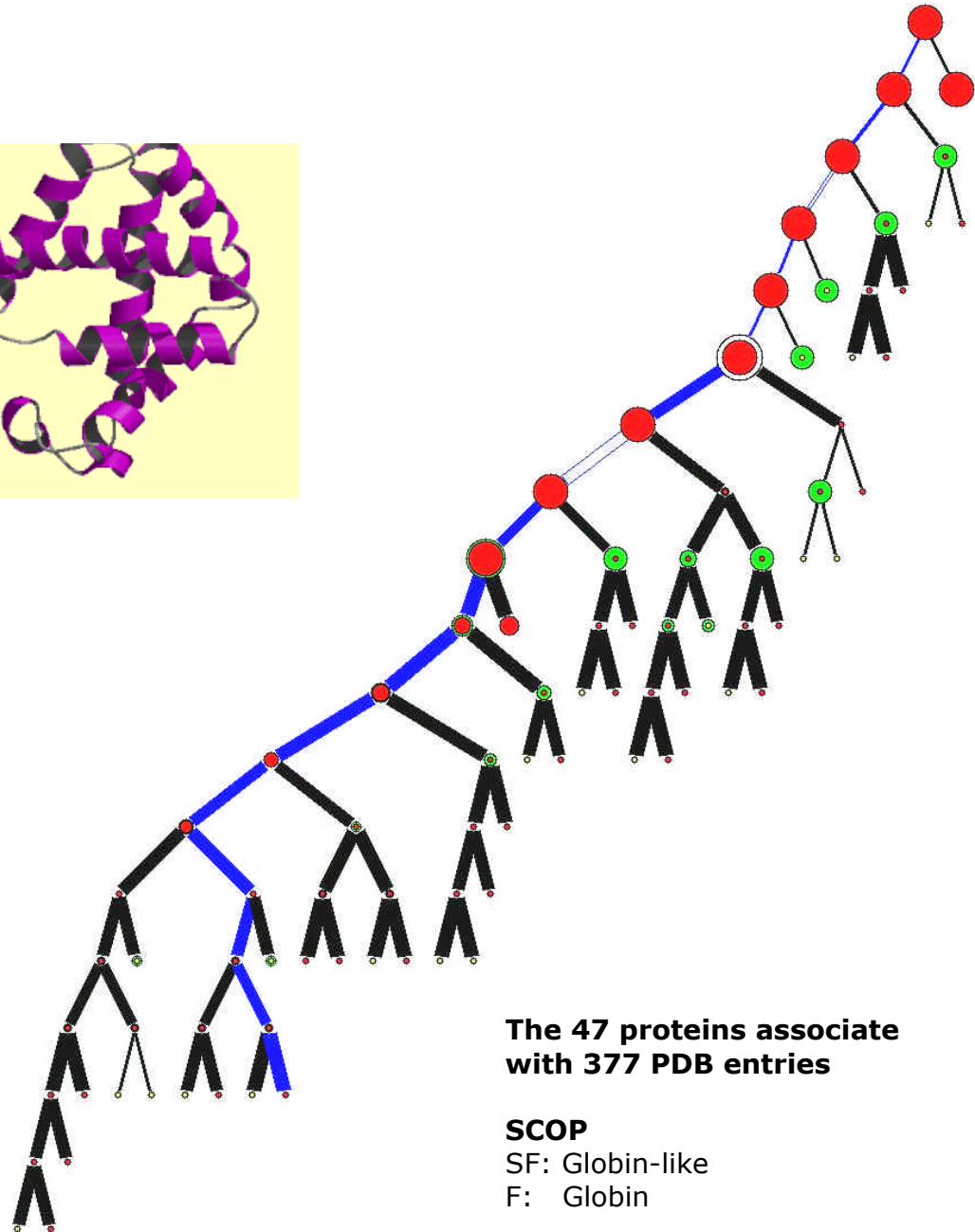
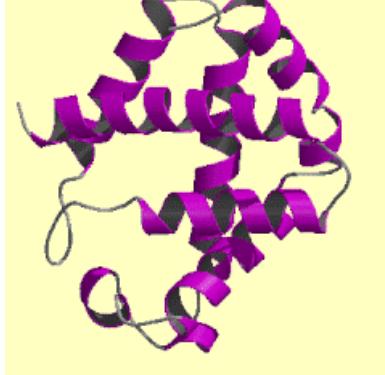
A horizontal view provides 'distances' between clusters. Those are the basis for creating **Road-Maps**.

We test the **biological content** of those road maps in term of biological information: I.e., domain, feature, **structure**, function, taxonomy etc.

January, 2003



**The
leaving
Tree :
the most
informative
subtree**



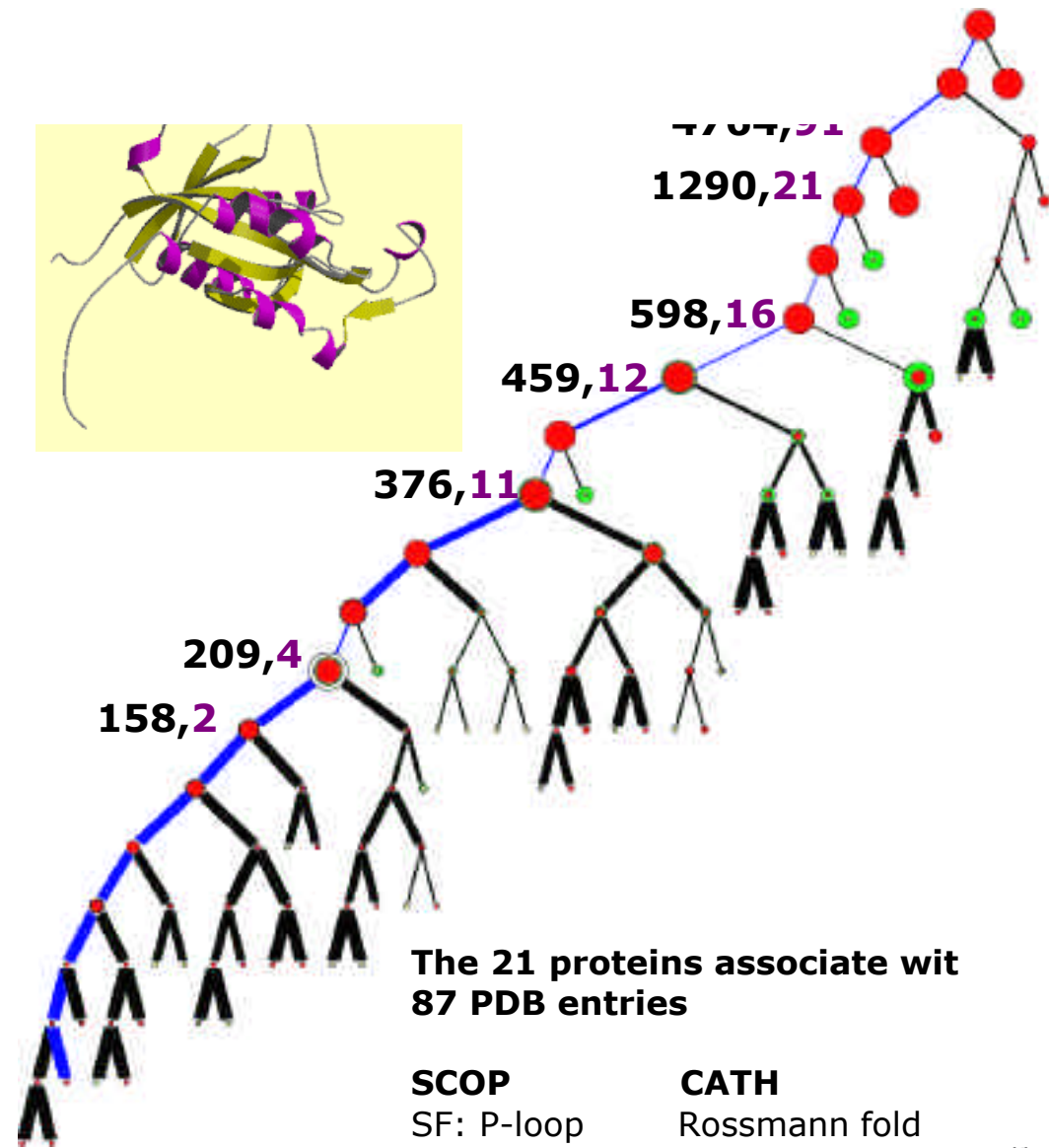
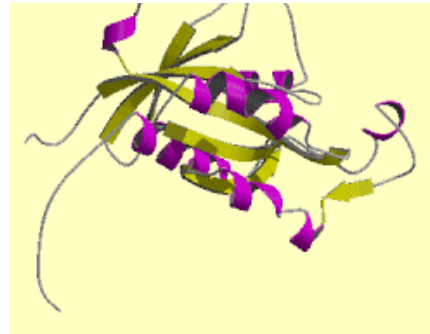
**The 47 proteins associate
with 377 PDB entries**

SCOP

SF: Globin-like

F: Globin

**The
leaving
Tree :
the most
informative
subtree**



**The 21 proteins associate with
87 PDB entries**

SCOP
SF: P-loop

CATH
Rossmann fold

The Twilight Zone

Proteins with **>30%** sequence identity can be reliably modeled and their fold may be predicted.

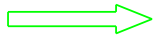
Remote homologues: Sequences with only **10-30%** identical amino-acids may belong to the same structural superfamily (or fold). **The Twilight Zone**

State of the art search engines (**PSI-BLAST, SAM-99**) fail to cross the Twilight Zone.

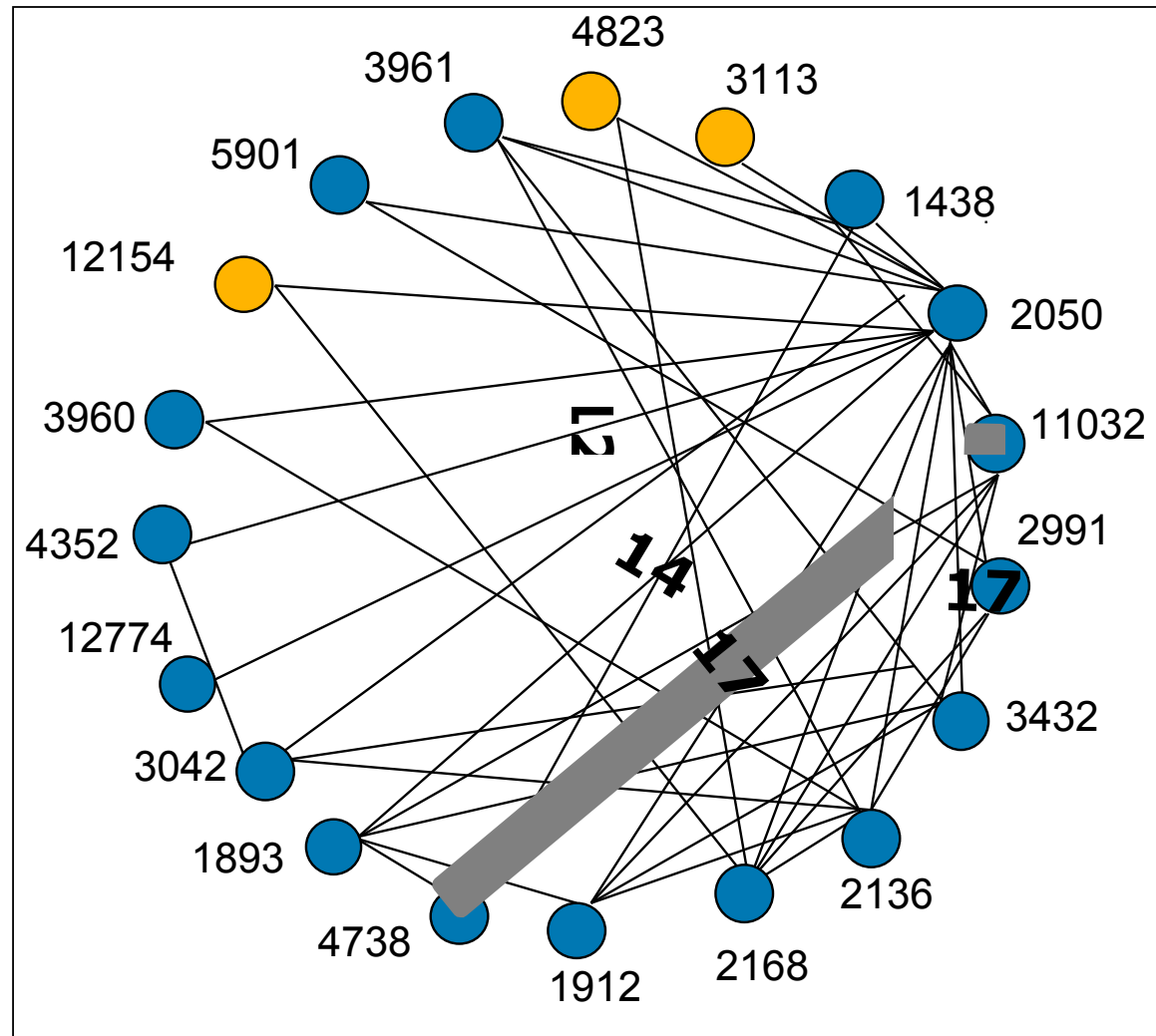
The sequence identity in the Twilight Zone provides no signal for structural similarity (Rost et al., 2002)

A roadmap with high complexity

multiple functions: antibiotic resistance, catalysis,
transcription regulation, histone acetylation..



The case of the GCN5 sequence group



GCN5- Related Superfamily (10 solved domains)

All proteins in the road-map
belong to the same superfamily
(SCOP - NAT)

The road-map connect proteins
that are considered in the
twilight zone



**HISTONE
ACETYLTRANSFERASE
(1BOB) Ec: 2.3.1.48**

R. N. Dutnall et al. Structure of the Histone Acetyltransferase Hat1:
A Paradigm for the Gcn5- Related N-Acetyltransferase Superfamily.
Cell 94 pp. 427 (1998)

