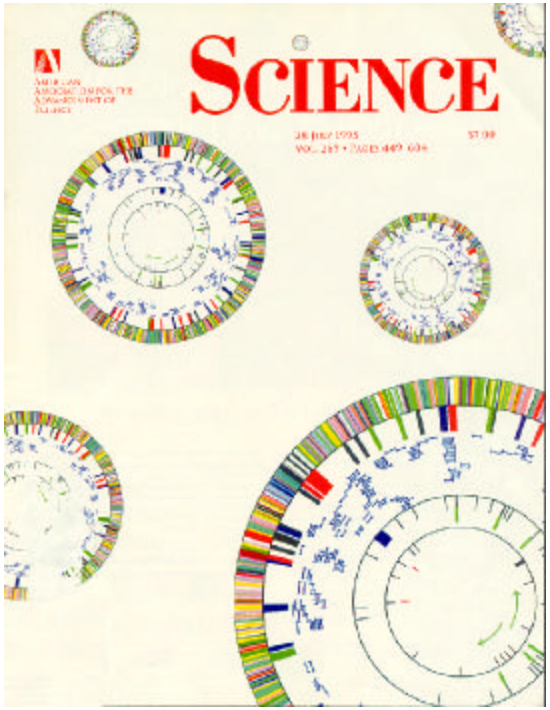
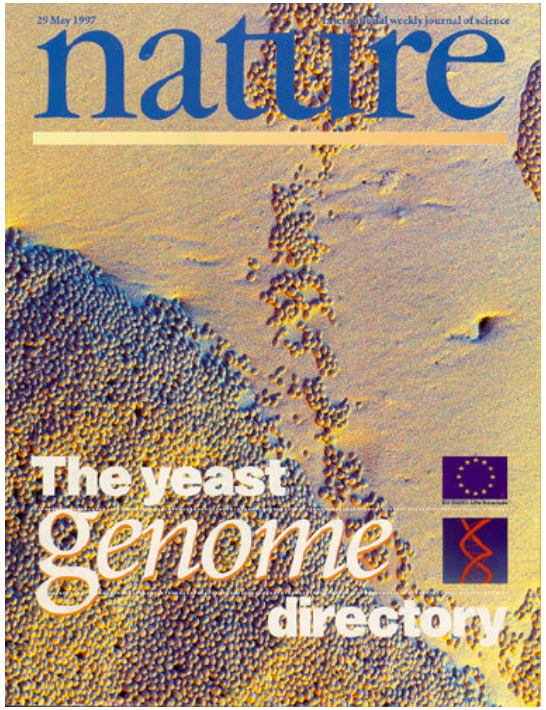
Comparing Genomes in terms of Protein Structure:

Surveys of a Finite Parts List

Mark Gerstein

Genomes highlight the **Finiteness** of the World of Sequences



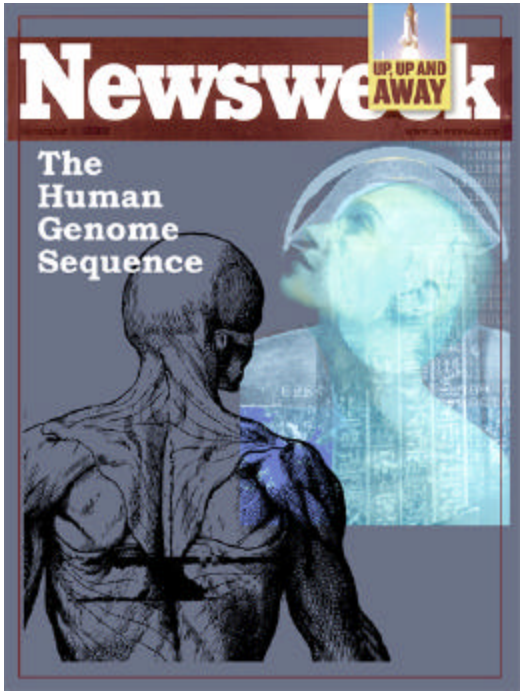


1995

Bacteria, 1.6 Mb, ~1600 genes [*Science* **269**: 496]

1997

Eukaryote, 13 Mb, ~6K genes [Nature **387**: 1]



1998

Animal, ~100 Mb, ~20K genes [*Science* **282**: 1945]

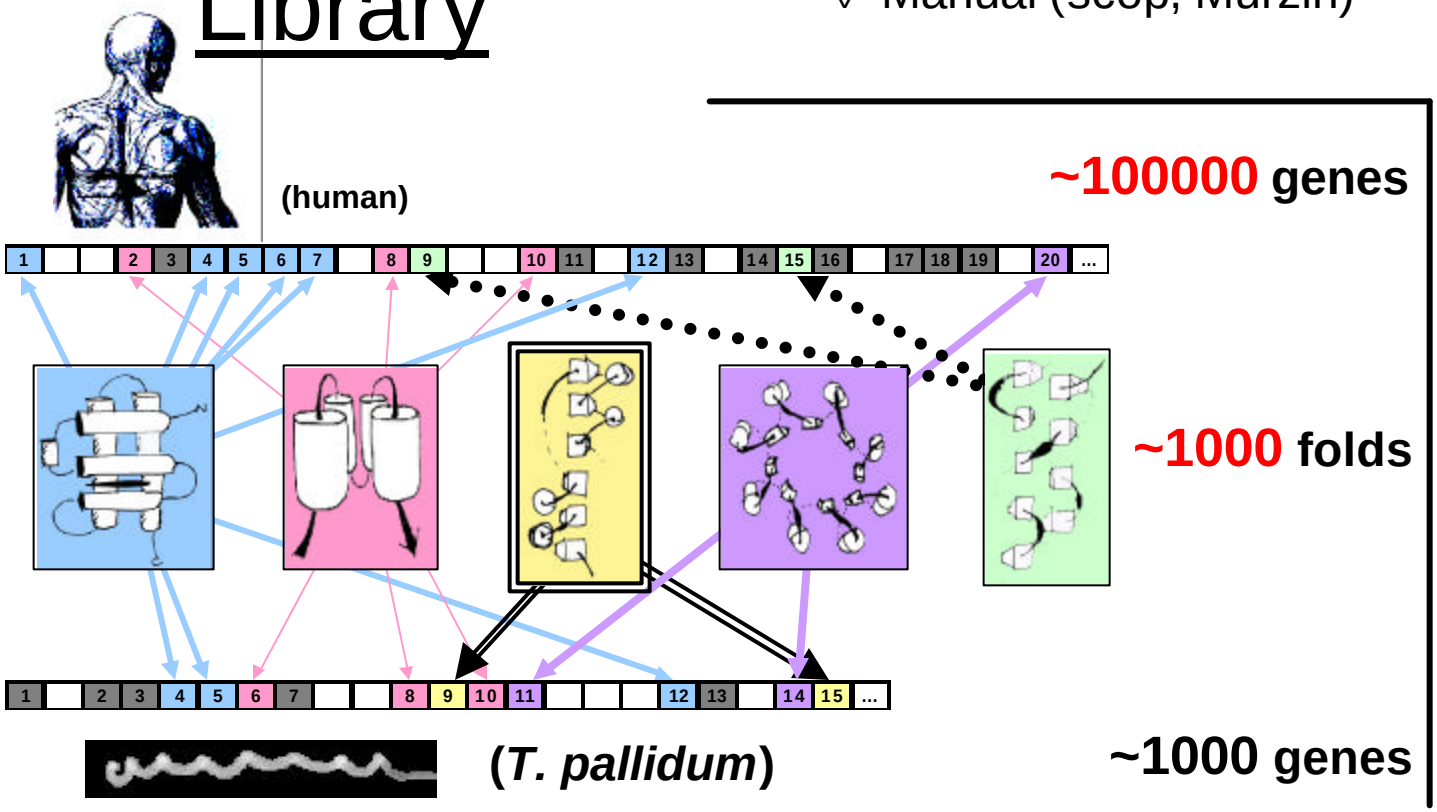
2000?

Human, ~3 Gb, ~100K genes [???]

The World of Structures is also Finite: A Fold Library

- Structure helps to understand genomes in simplest terms -- fewest parts & most duplication
- Structural domain more precisely defined than sequence module
- Sequence Similarity more reliably related to Structure than Function
- Many approaches to building Library
  - ◊ Manual (scop, Murzin)

Automatic:
FSSP-HSSP (Holm/Sander), Entrez-MMDB (Bryant)
Semi-automatic:
CATH (Thornton), HOMALDB (Sali)
Sequences 1st:
Pfam (Durbin/Eddy), COGs (Koonin/Lipman), Blocks (Henikoff), ProSite (Bairoch)



Comparing Genomes in terms of Protein Structure: Surveys of a **Finite** Parts List

1 Library of Known Folds

Importance of Statistics. Scop auto-alignments. P-values from EVD, same as sequences.

2 Census of Known Folds

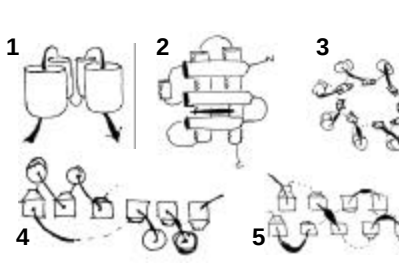
Which folds in which organisms: E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression. Repeated βαβ. Biases. Extent of MG fold assignment (65%)

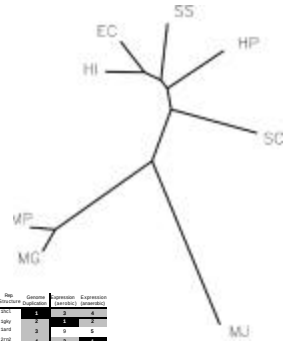
3 Prediction of Unknown Folds

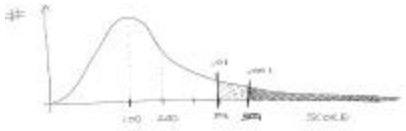
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Biases: Can extrapolate from known structures to genomes?

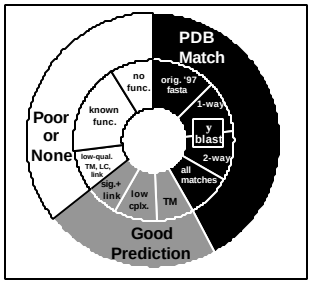
4 Fold-Function Relationships

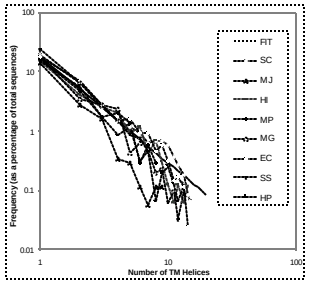
How many folds per function? Func. per fold? 331 of ~20K combinations. TIM most versatile scaffold.







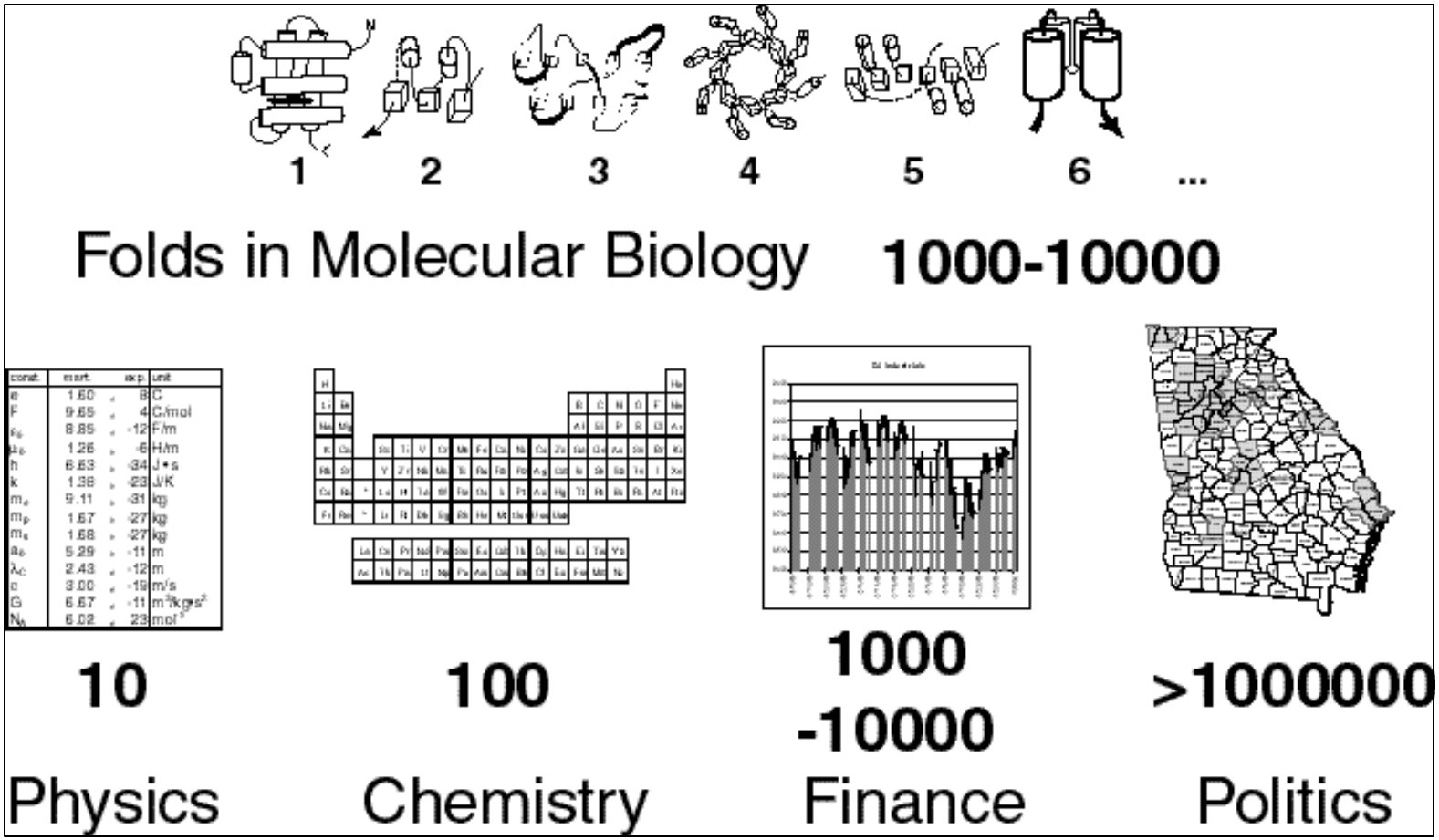




ENZYME	SCOP A	SCOP B	SCOP A/B	SCOP A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

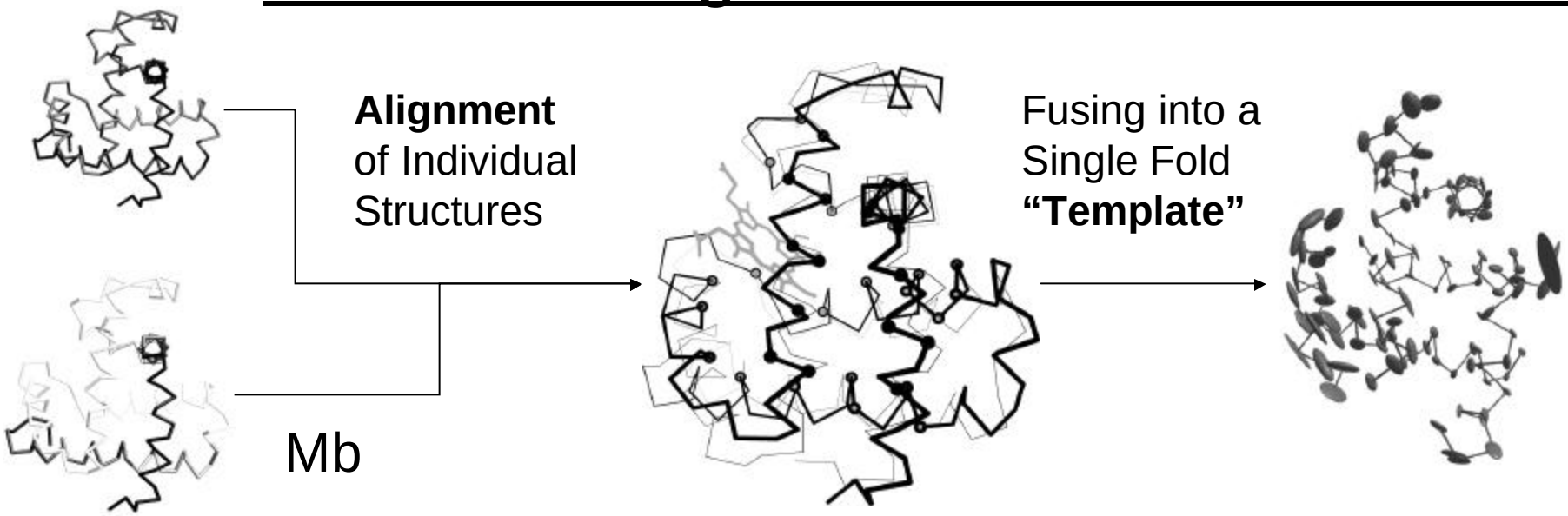
Fold Library vs. Other Fundamental Data structures

Parts List **Database**; **Statistical**, rather than mathematical relationships and conclusions



(Large than physics and chemistry, Similar to Finance (Exact Finite Number of Objects (3,056 on NYSE by 1/98), descrip. by Standardized Statistics (even abbrevs, INTC) and groups (sectors)) Smaller than Social Surveys, Indefinite Number of People, Not Well Defined Vocabulary and statistics.

Automatic Alignment to Build Fold Library

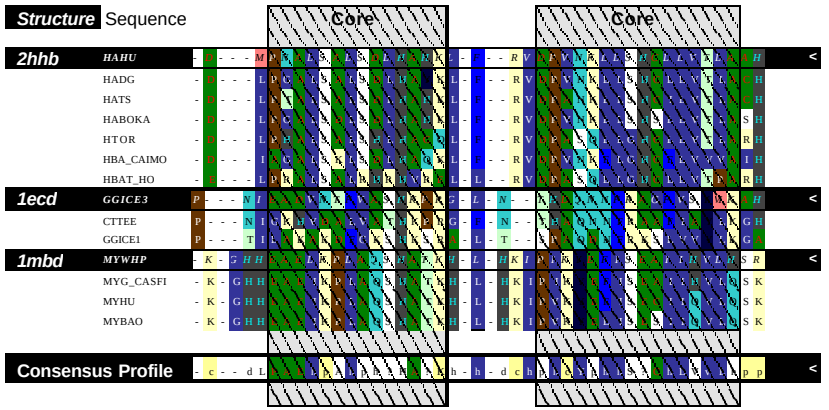



```
Hb VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQVKGHGKKVADALTNAV
     |||..   |   |.|| |  .  |.  | | | |   |  |  |       .|  .| ||  |  ||   .
Mb VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAIL

Hb AHVD-DMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
             |  | . || | .. .    .|  ..  |    |..|    . . | |      .  ||.
Mb KK-KGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
```

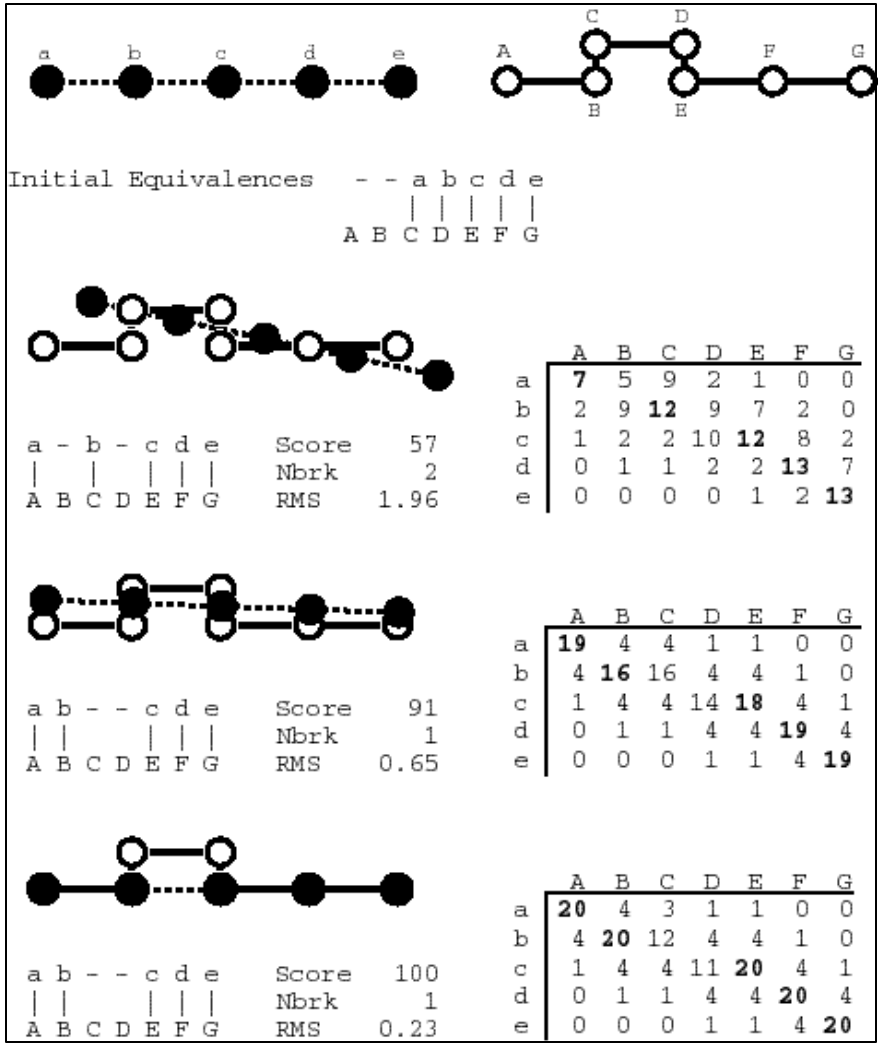
Elements: Domain definitions; Aligned structures, collecting together Non-homologous Sequences; Core annotation

Previous work: Remington, Matthews '80; **Taylor, Orengo '89**, '94; Artymiuk, Rice, Willett '89; Sali, Blundell, '90; Vriend, Sander '91; Russell, Barton '92; **Holm, Sander '93**; Godzik, Skolnick '94; Gibrat, Madej, Bryant '96; Falicov, F Cohen, '96; Feng, Sippl '96; G Cohen '97; Singh & Brutlag, '98



Automatic Alignments of Scop, Focussing on Statistics of Relationships

- Iterative Dynamic Programming, like repeated sequence alignment
  - ◊ Single cycle doesn't converge since violates key assum. D.P.
- Derived from Program of G Cohen (Align, Satow et al. 1986)



	A	B	C	D	E	F	G
a	**7**	5	9	2	1	0	0
b	2	9	**12**	9	7	2	0
c	1	2	2	10	**12**	8	2
d	0	1	1	2	2	**13**	7
e	0	0	0	0	1	2	**13**

	A	B	C	D	E	F	G
a	**19**	4	4	1	1	0	0
b	4	**16**	16	4	4	1	0
c	1	4	4	14	**18**	4	1
d	0	1	1	4	4	**19**	4
e	0	0	0	1	1	4	**19**

	A	B	C	D	E	F	G
a	**20**	4	3	1	1	0	0
b	4	**20**	12	4	4	1	0
c	1	4	4	11	**20**	4	1
d	0	1	1	4	4	**20**	4
e	0	0	0	1	1	4	**20**

```
ACSQRP--LRV-SH  -R  SENCV
A-SNKPQLVKLMTH  VK  DFCV-
```

Similarity Scores

$$S = \sum_{i,j} M(i,j) - nG$$

S = Total Score, where the sum is carried out over all aligned residues i and j

M(i,j) = Similarity matrix score for aligning i and j

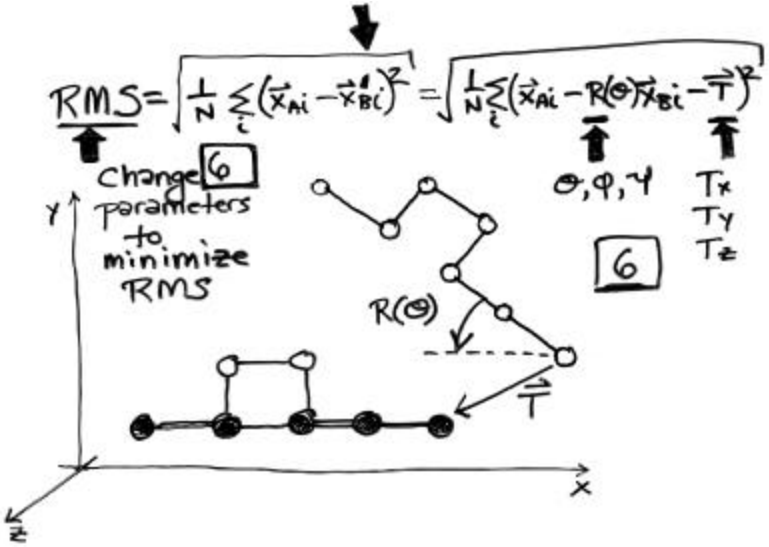
n, G = number of gaps (assuming no gap ext. penalty) and gap penalty

$M_{str}(i,j)$ = $100 / (5 + d(i,j)^2)$
(for Structural Similarity -> S_{str})

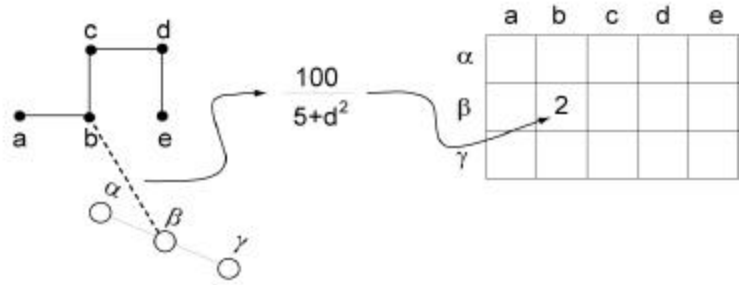
$M_{seq}(i,j)$ = Blossum matrix
(for Seq. Similarity -> S_{seq} = SWS)

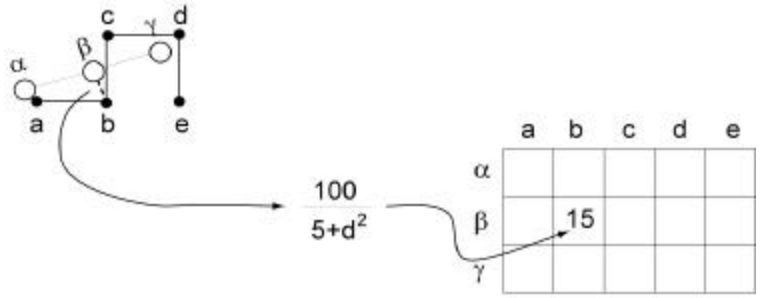
RMS =sqrt(sum($d(i,j)^2$)) [a difference]

%ID = Percentage identity



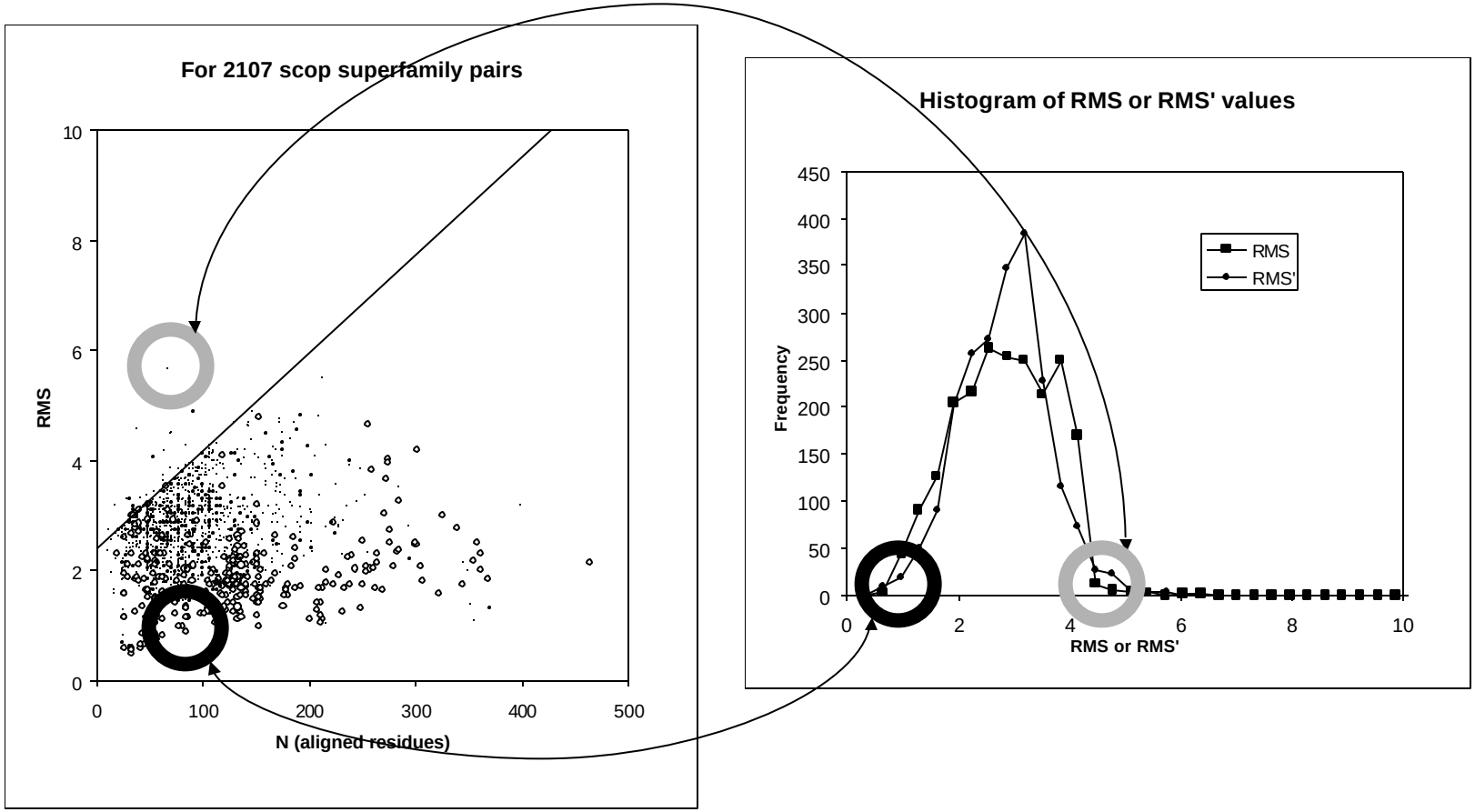
$$d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$





Statistics on Range of Similarities

For 2107 pairs, only 2% Outliers (with subtle similarity)



Num. Aligned

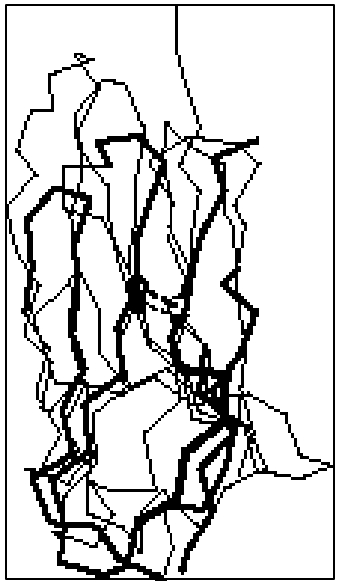
Some Similarities are Readily Apparent others are more Subtle

Easy:
Globins

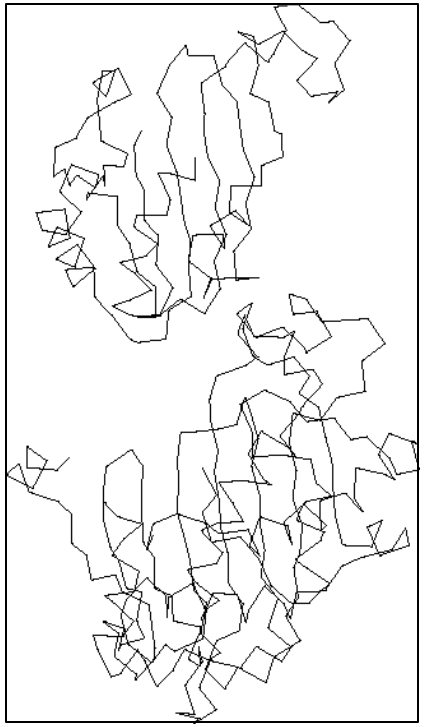
125 res.,
~1.5 Å

Tricky:
Ig C & V

85 res.,
~3 Å



Very Subtle: G3P-dehydro-genase, C-term. Domain
>5 Å



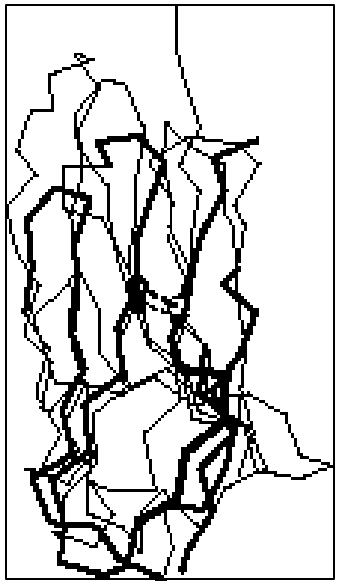
Some Similarities are Readily Apparent others are more Subtle

Easy:
Globins

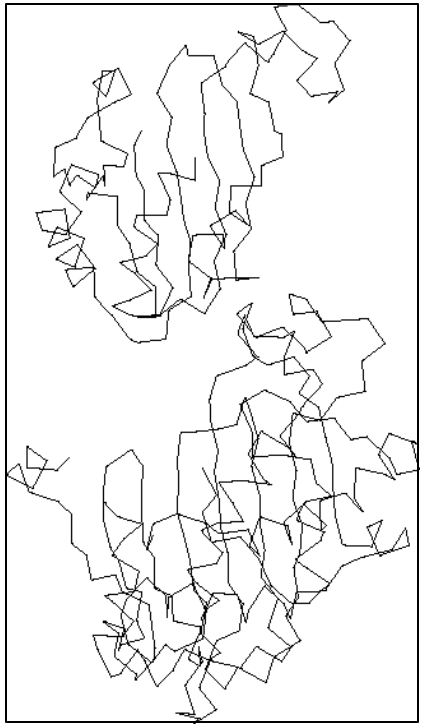
125 res.,
~1.5 Å

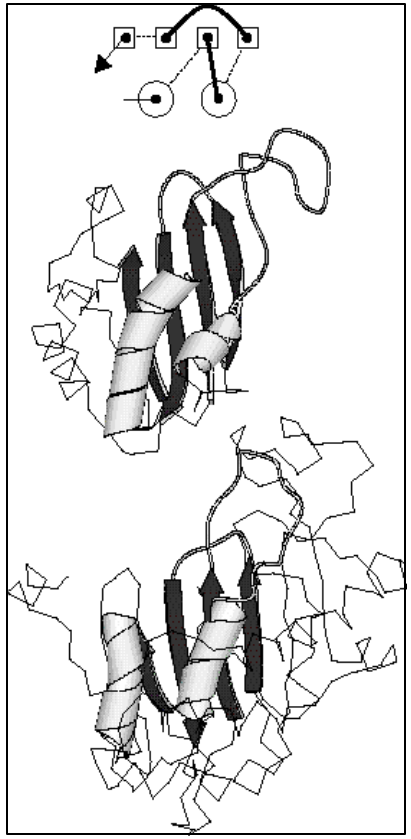
Tricky:
Ig C & V

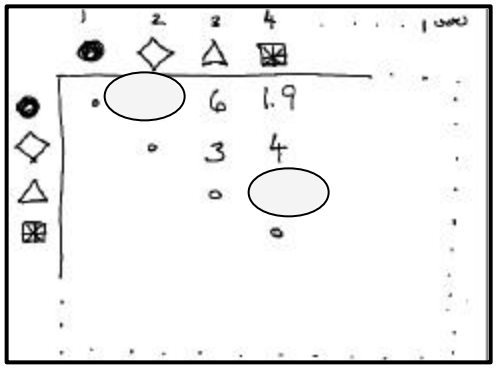
85 res.,
~3 Å



Very Subtle: G3P-dehydro-genase, C-term. Domain
>5 Å

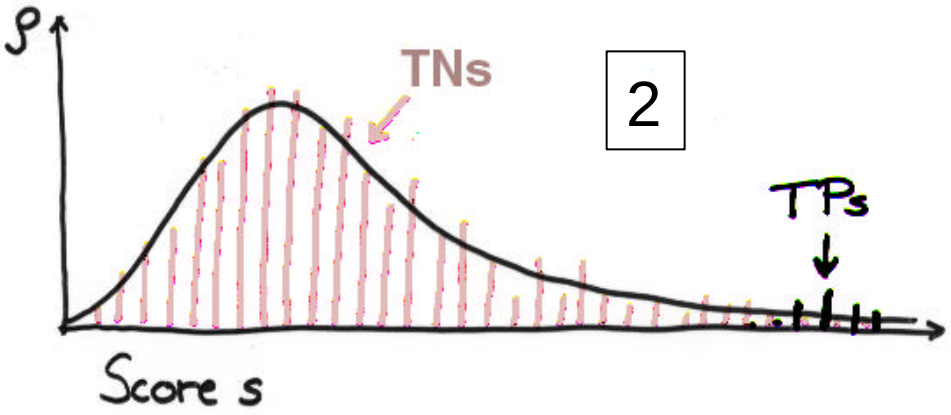




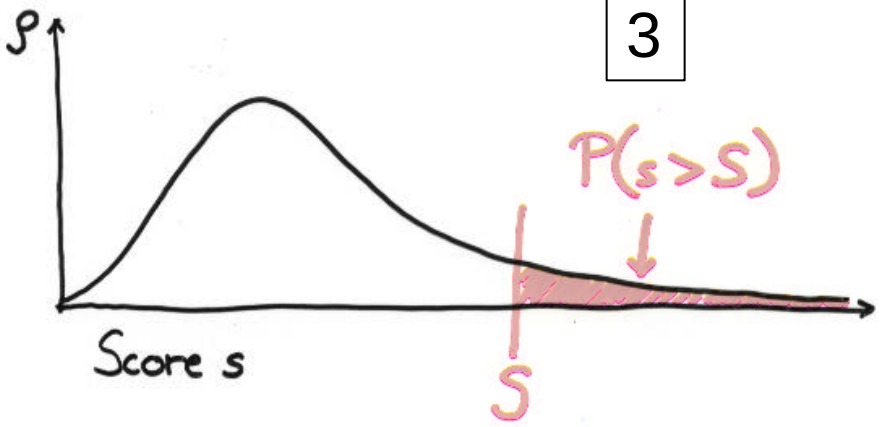


P-values

1



2

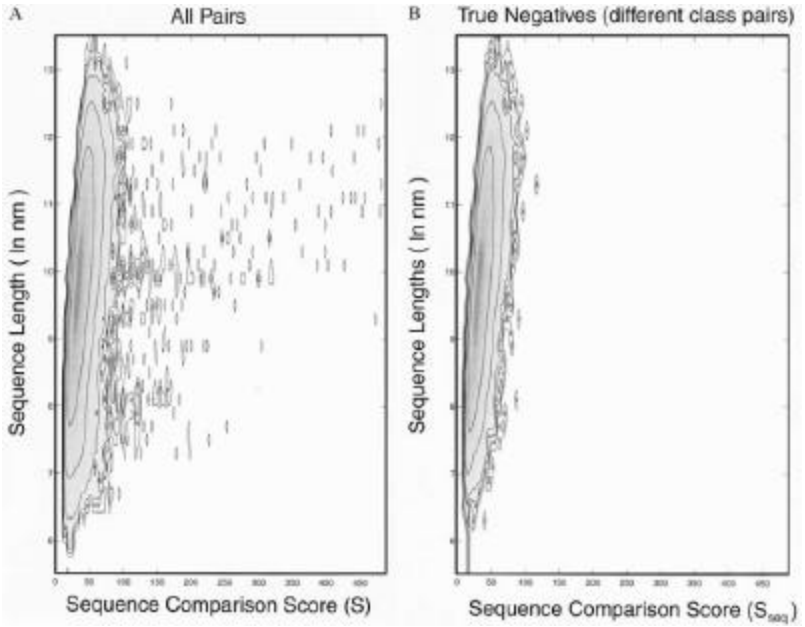


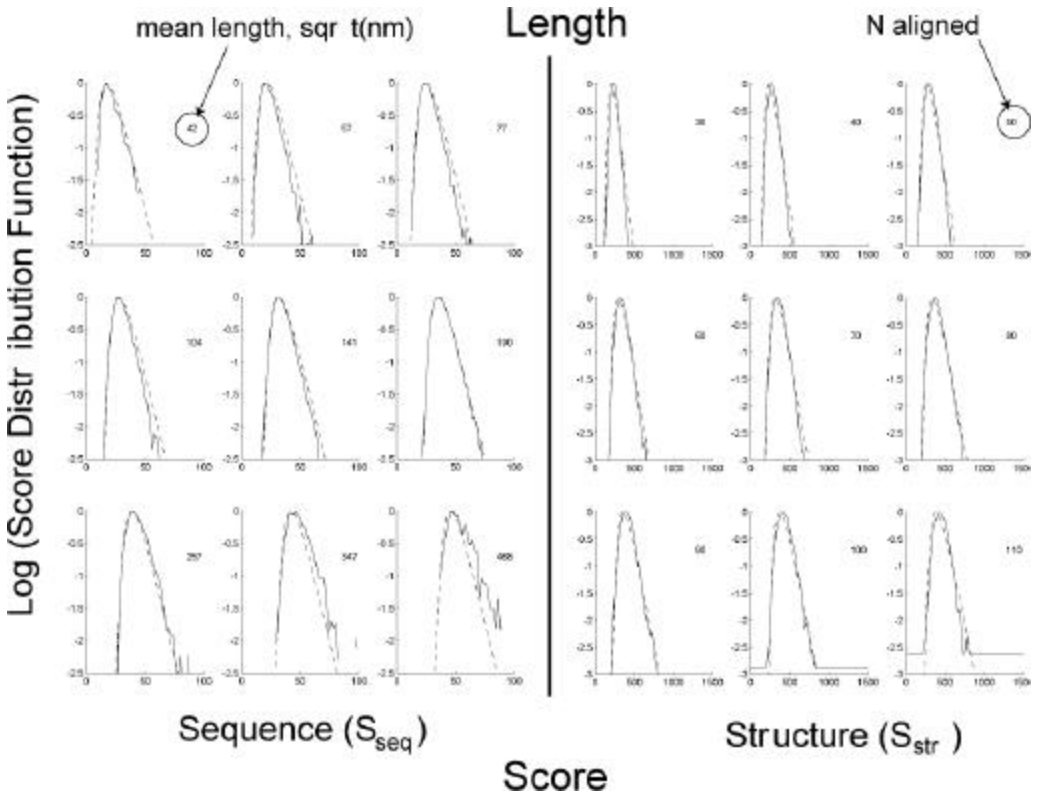
3

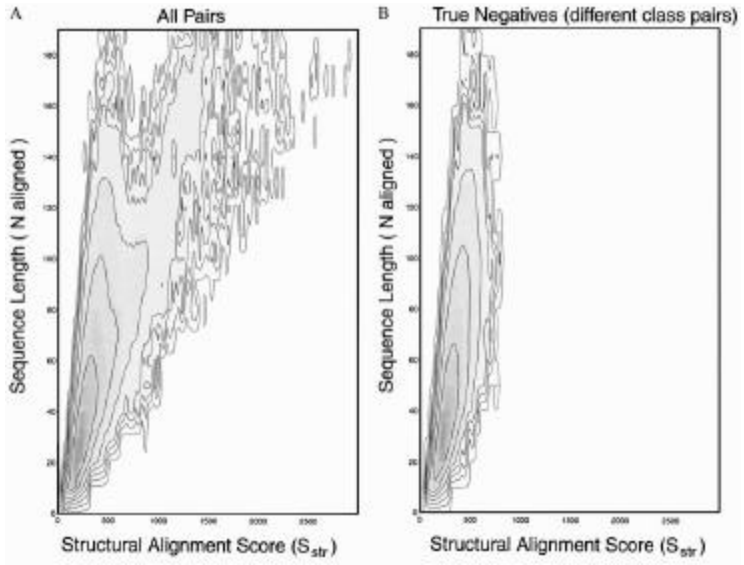
- Significance Statistics
  - ◊ For sequences, originally used in Blast (Karlin-Altschul). Then in FASTA, &c.
  - ◊ Extrapolated Percentile Rank: How does a Score Rank Relative to all Other Scores?
- Our Strategy: Fit to Observed Distribution
  - **1)** All-vs-All comparison
  - **2)** Graph Distribution of Scores in 2D (N dependence); 1K x 1K families -> ~1M scores; ~2K included TPs
  - **3)** Fit a function ρ(S) to TN distribution (TNs from scop); Integrating ρ gives P(s>S), the CDF, chance of getting a score better than threshold S randomly
  - 4) Use same formalism for sequence & structure

[e.g. P(score s>392) = 1% chance]

Observed Distributions







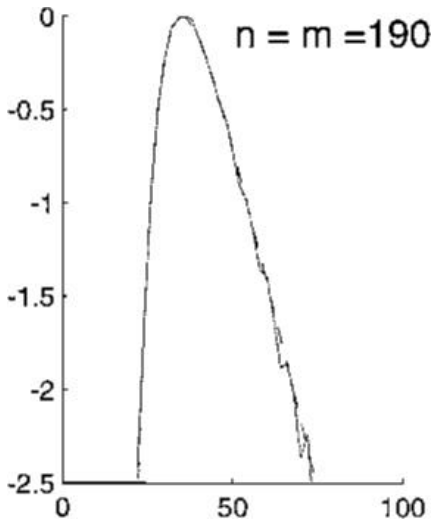
EVD Fits

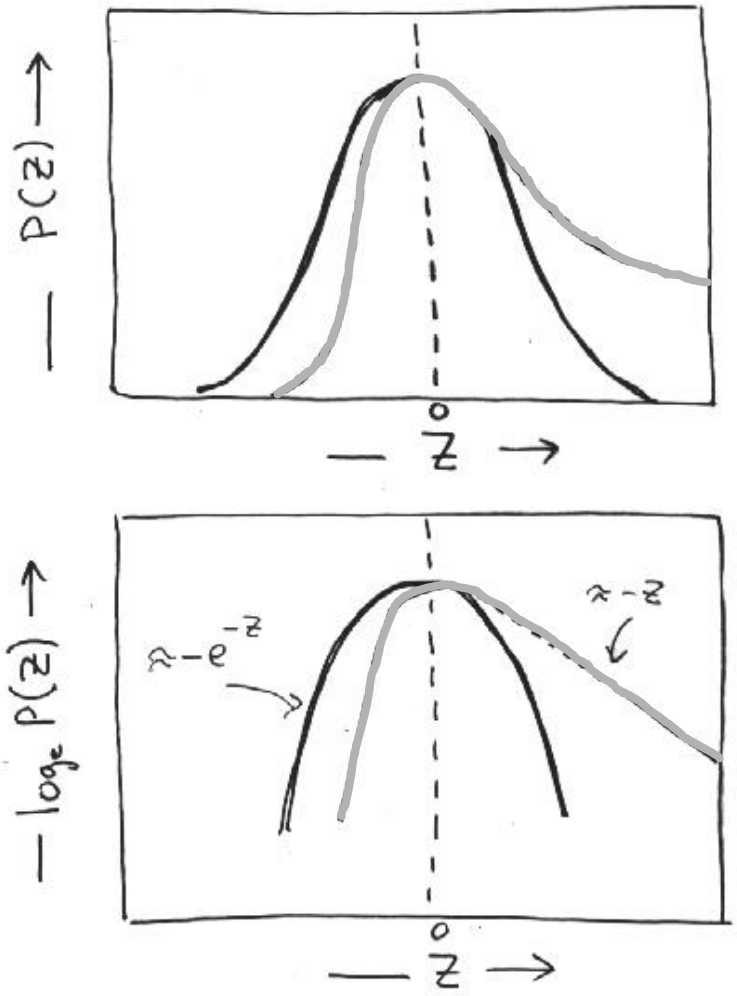
$$r(z) = \exp\left(-z - e^{-z}\right)$$

$$\left(\ln r(z) = -z - e^{-z}\right)$$

- Reasonable as Dyn. Prog. maximizes over pseudo-random variables
- EVD is **Max**(indep. random variables);
- Normal is **Sum**(indep. random variables)
  - ◊ $\rho(z) = \exp(-z^2)$, ln $\rho(z) = -z^2$

Observed





Extreme Value Distribution (EVD, long-tailed) fits the observed distributions best. The corresponding formula for the P-value:

$$P(z > Z) = \int r(z)dz = 1 - \exp(-e^{-Z})$$

Same Results for Sequence & Structure

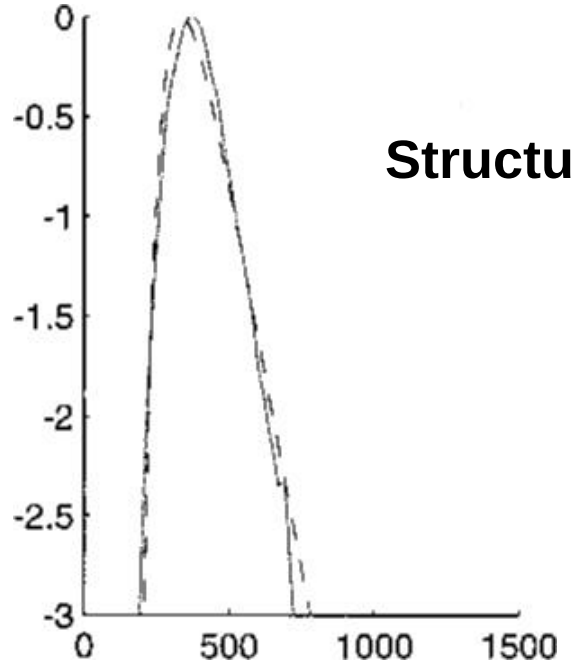
3 Free Parm. fit to EVD involving: ***a, b,s***.
These are the only difference betw. sequence and structure.

$$Z = \frac{S - (a \ln N + b)}{s}$$

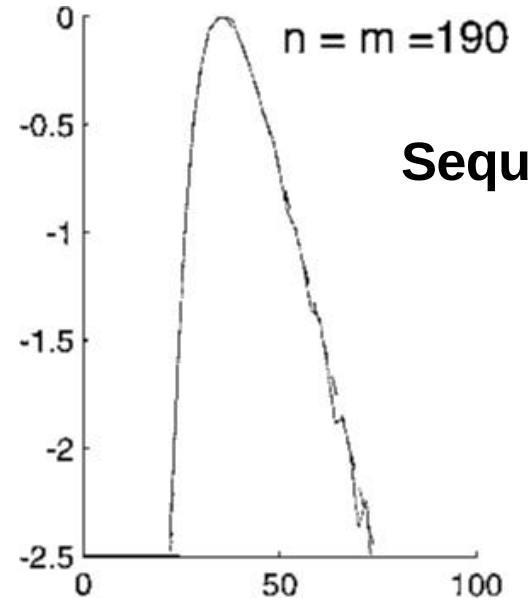
$$S = \sum_{i,j} M(i,j) - G$$

$$\boxed{r(z) = \exp\left(-z - e^{-z}\right)}$$

N, G, M also defined differently for sequence and structure.
N = number of residues matched.
G = total gap penalty.
M(i,j) = similarity matrix
(Blossum for seq. or $M_{str}(i,j)$, struc.)



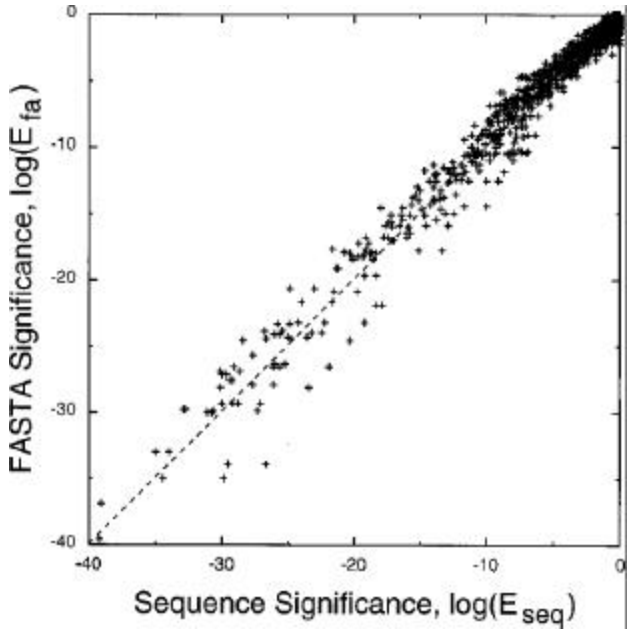
Structure



Sequence

Use Sequence Scores to Validate

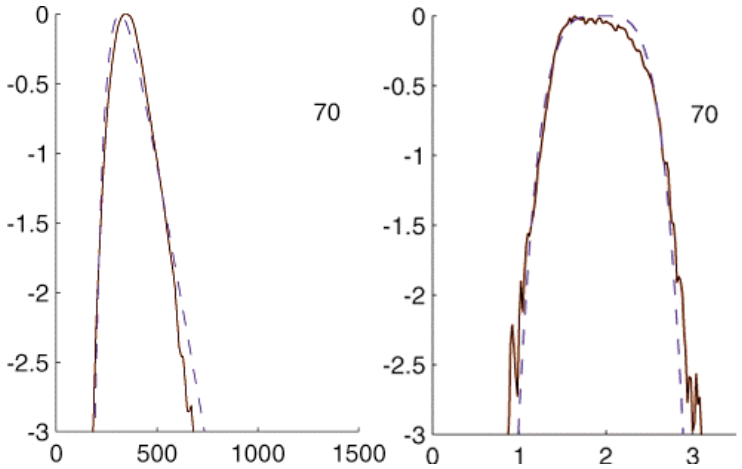
- Sequence P-value perfectly tracks FASTA e-value
  - ◊ Validates approach
  - ◊ Added Benefit: allows computation of an e-value without doing a db run
- Significance computation can be applied to **any** exisiting sequence or structure alignment



- Also, RMS doesn't work instead of structural alignment (no EVD fit)
  - ◊ RMS penalizes worst fitting atoms, easily skewed

S_{str} RMS

$$\sum \frac{100}{5+\mathbf{d}_i^2} \; vs \quad \sqrt{\sum \mathbf{d}_i^2}$$



Comparing Genomes in terms of Protein Structure: Surveys of a **Finite** Parts List

1 Library of Known Folds

Importance of Statistics. Scop auto-alignments. P-values from EVD, same as sequences.

2 Census of Known Folds

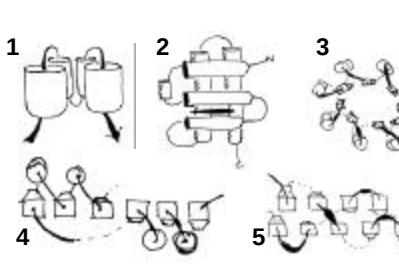
Which folds in which organisms: E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression. Repeated βαβ. Biases. Extent of MG fold assignment (65%)

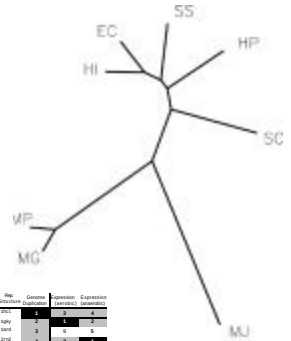
3 Prediction of Unknown Folds

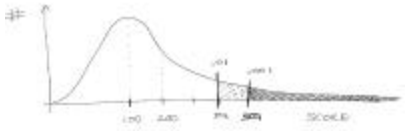
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Biases: Can extrapolate from known structures to genomes?

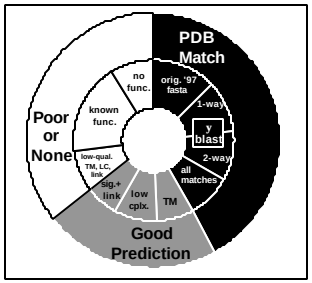
4 Fold-Function Relationships

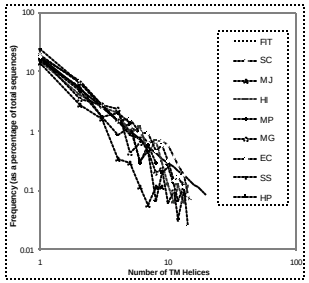
How many folds per function? Func. per fold? 331 of ~20K combinations. TIM most versatile scaffold.





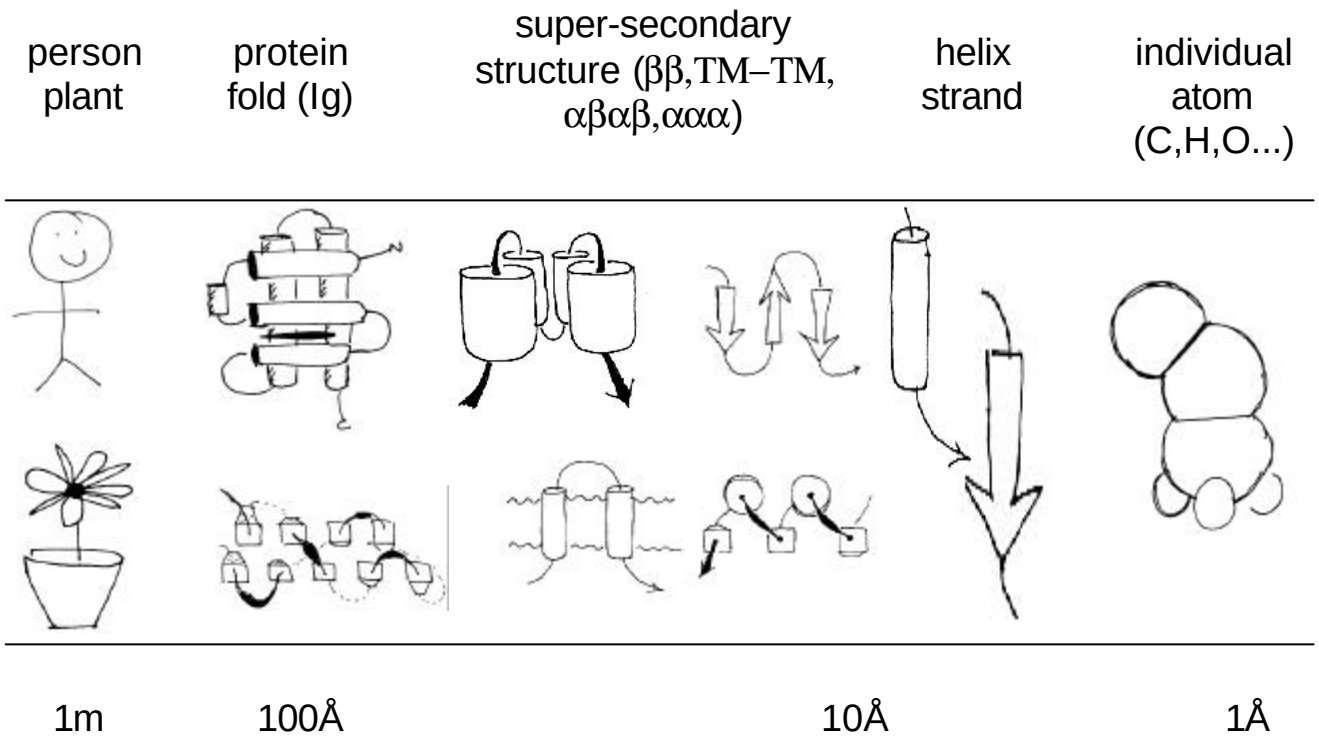


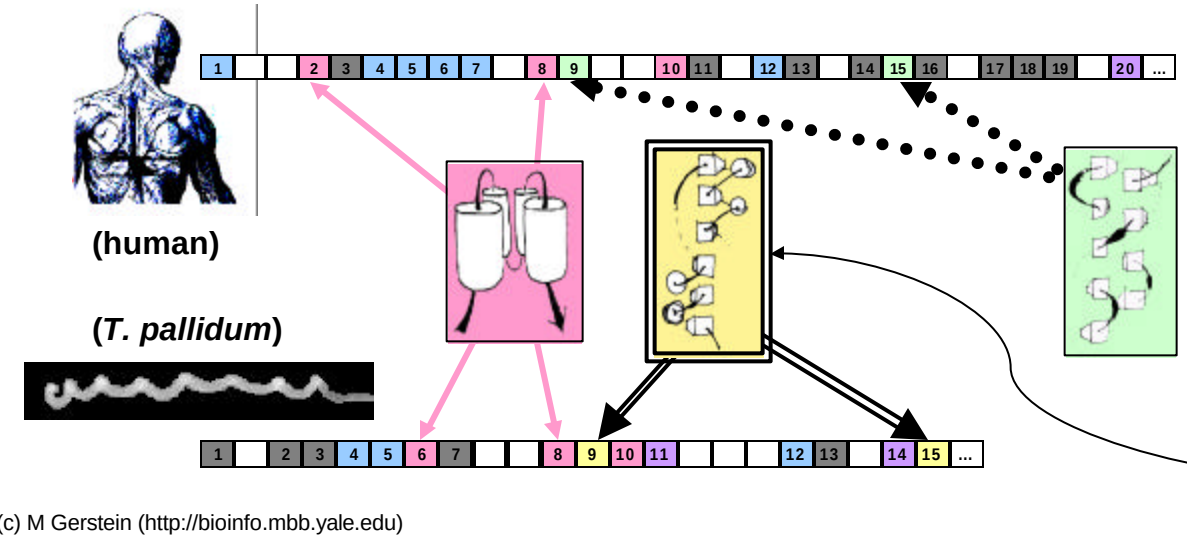




ENZYME	SCOP A	SCOP B	SCOP A/B	SCOP A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

At What Structural Resolution Are Organisms Different?

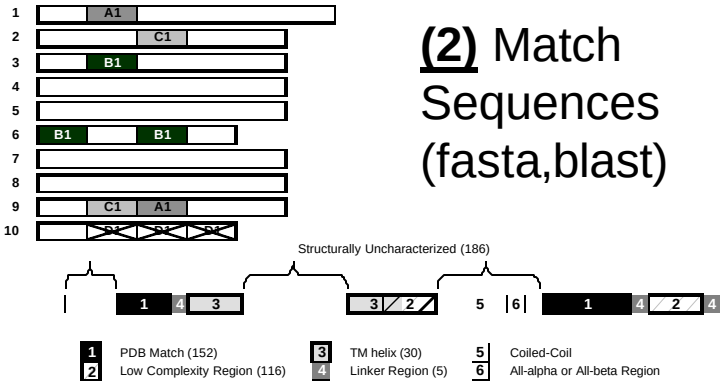




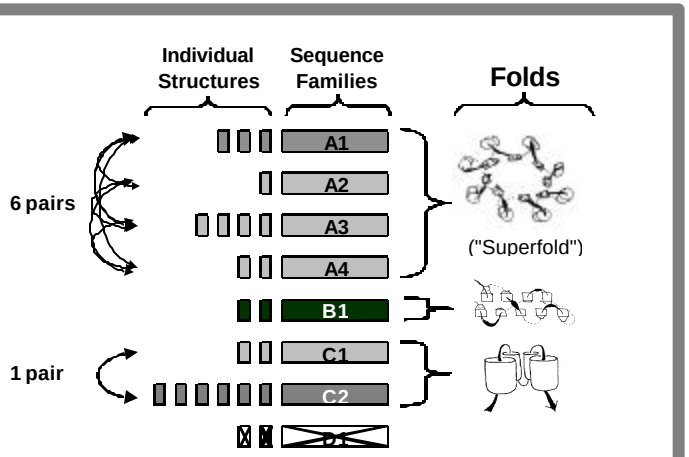
Practical Relevance

(Pathogen only folds as possible targets)

Cross-Reference: Folds→Sequences → Organisms

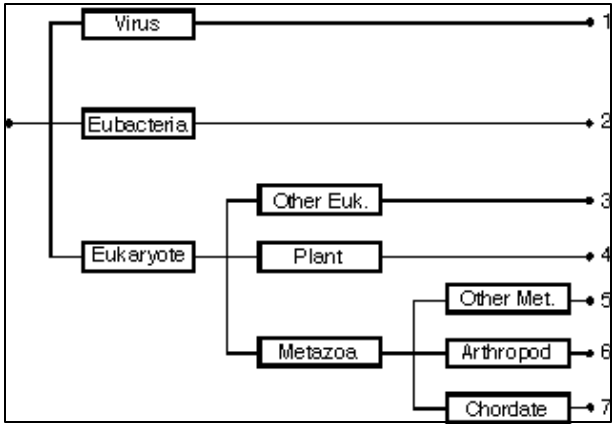


(2) Match Sequences (fasta,blast)



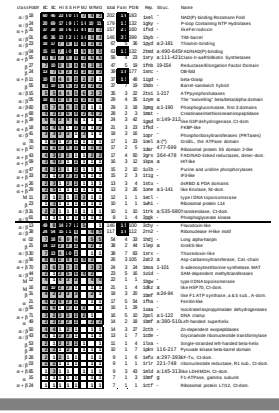
(1) Structures in Folds (scop)

(3) Organize Sequences by Genome or Taxon



3 + 5

Abbrev.	Kingdom (subgroup)	Genome	Num. ORFs	Reference
EC	Bacteria (gram negative)	*Escherichia coli*	**4290**	Blattner et al.
HI	Bacteria (gram negative)	*Haemophilus influenzae*	**1680**	TIGR
HP	Bacteria (gram negative)	*Helicobacter pylori*	**1577**	TIGR
MG	Bacteria (gram positive)	*Mycoplasma genitalium*	**468**	TIGR
MJ	Archaea (Euryarchaeota)	*Methanococcus jannaschii*	**1735**	TIGR
MP	Bacteria (gram positive)	*Mycoplasma pneumoniae*	**677**	Himmelreich et al.
SC	Eukarya (fungi)	*Saccharomyces cerevisiae*	**6218**	Goffeau et al.
SS	Bacteria (Cyanobacteria)	*Synechocystis* sp.	**3168**	Kaneko et al.



(4) Results in "Fold Table"

class	Fold#	EC	SC	HI	SS	HP	MJ	MP	MG	total	Fam.	PDB	Rep.	Struc.	Name
α/β	18	60	46	23	40	19	7	4	3	202	16	183	1xel	-	NAD(P)-bindi
α/β	24	20	69	17	19	17	16	10	11	179	13	132	1gky	-	P-loop Contai
α+β	31	37	28	18	16	12	40	3	3	157	23	160	1fxd	-	like Fe odox
α/β	01	45	36	13	22	11	10	5	4	146	37	399	1byb	-	TIM-ba el
α/β	23	18	17	7	9	4	8	2	2	67	5	36	1pyd	a:2-181	Thiami bindi
α/β	04	15	11	7	10	1	9	5	5	63	13	132	2tmd	a:490-645	FAD/N D(P)-
α+β	55	8	9	7	8	9	3	6	6	56	4	23	1sry	a:111-421	Class- aaR
β	27	7	10	8	8	4	4	3	3	47	5	19	1fnb	19-154	Reduct se/E
β	24	13	7	4	3	3	3	3	3	39	18	177	1snc	-	OB-fold
α+β	11	10	8	4	8	2	2	2	1	37	11	48	1igd	-	beta-Grasp

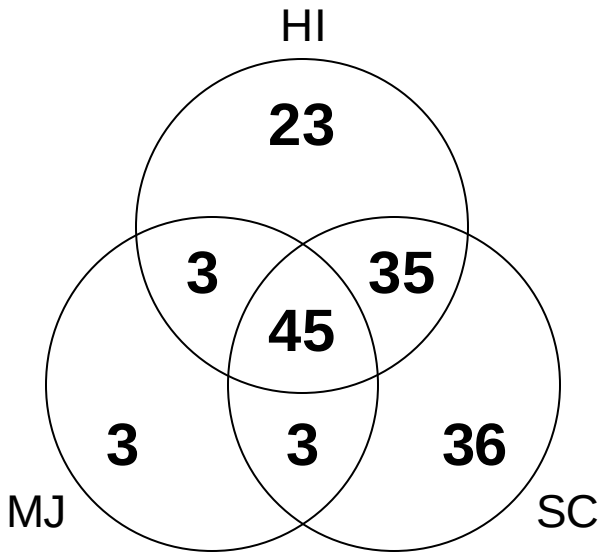
Venn Diagrams for Shared Folds in OWL and Initial Genomes

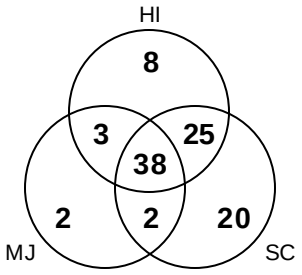
~300-350 folds (282 folds in scop 1.32 [‘96])

~120K sequences in OWL 27.1

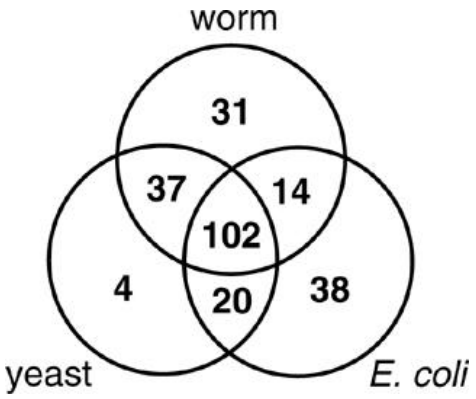
7 phylogenetic groups of organisms

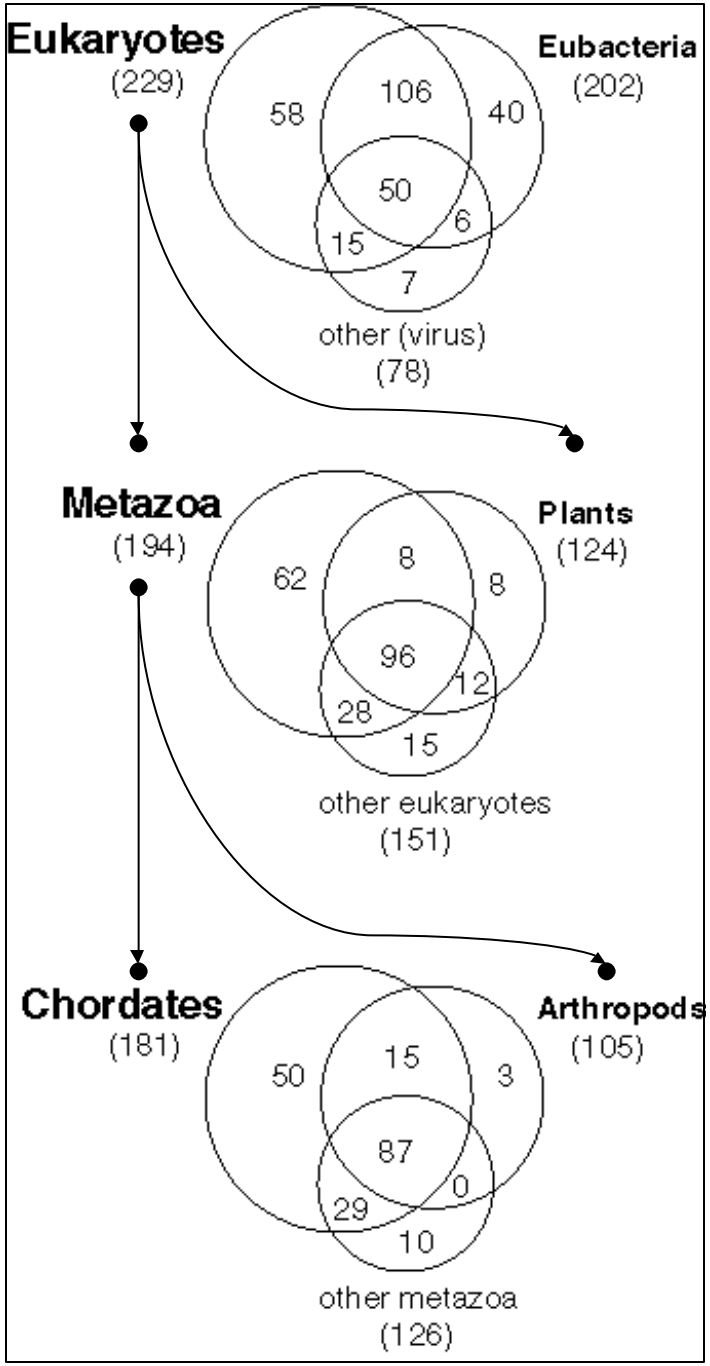
5 genomes --
HI, EC (bacteria),
MJ (archeon),
SC (eukaryote),
CE (worm, animal)





α/β





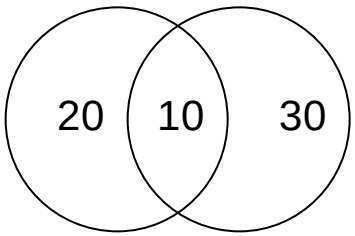
Cluster Trees Grouping 8 Initial Genomes on Basis of Shared Folds

D=S/T

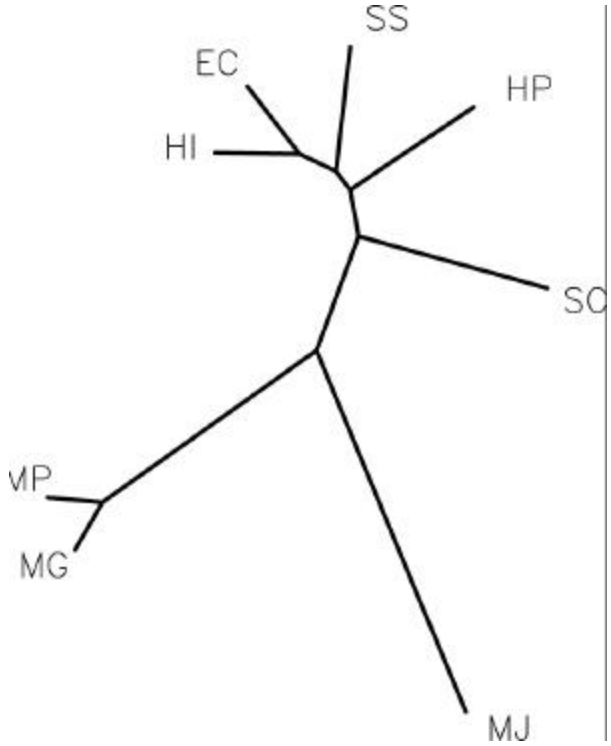
D = shared fold dist. betw. 2 genomes

S = # shared folds

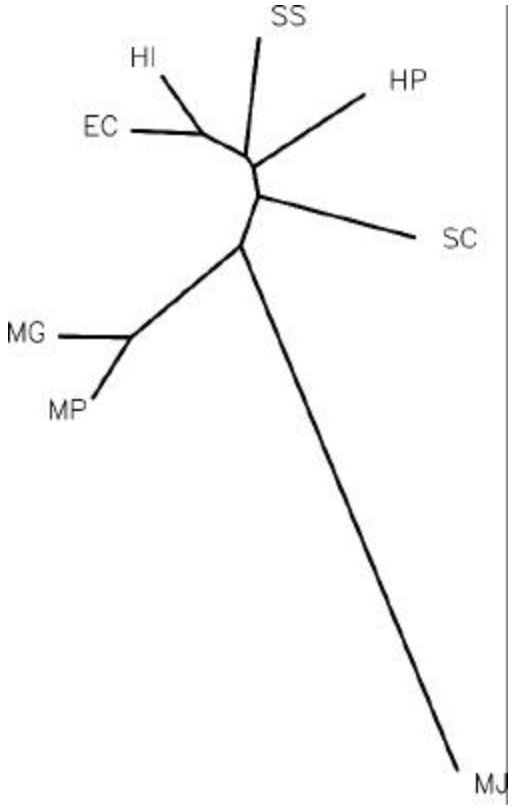
T= total # folds in both



D=10/(20+10+30)



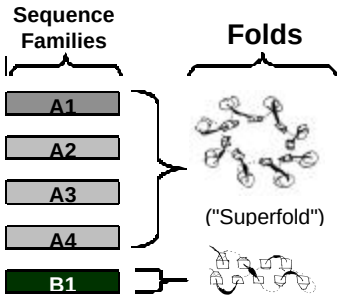
Fold Tree



"Classic" Tree

Patterns of Folds Usage in 8 Genomes

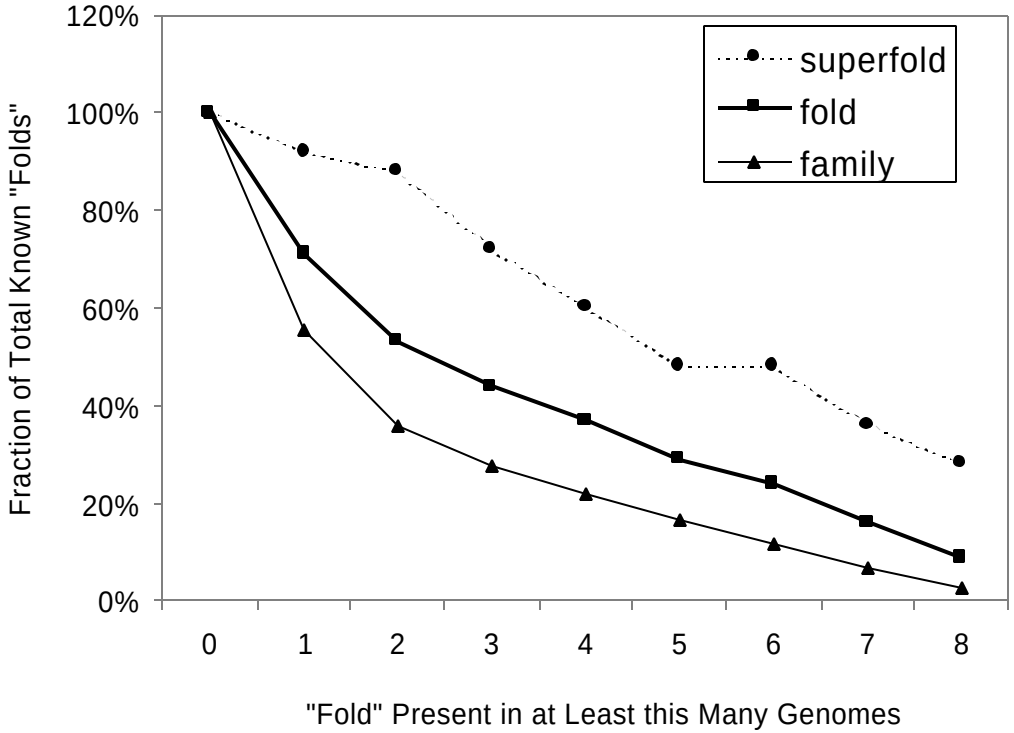
	fold	fam.	super fold
total in PDB	338	990	25
in at least one of 8 genomes	240	547	23
present in this many genomes			
1	60	192	1
2	32	82	4
3	23	54	3
4	27	53	3
5	17	50	0
6	27	49	3
7	24	41	2
8	30	26	7



Superfold = fold that allows many non-homologous seq. (Thornton)

```
ESHSHMMM (##) ESHSHMMM (##) ESHSHMMM (##) ESHSHMMM (##) ESHSHMMM (##)
CCISPJPG      CCISPJPG      CCISPJPG      CCISPJPG      CCISPJPG

11111111 (30) .1..... (23) 1..... (19) 11111.11 (16) 111111.. (16)
1111.... (09) 11111... (08) 1.1.... (08) 1.111.11 (06) 11..... (06)
...1.... (06) 1.11.... (05) .1.1.... (05) 1.111... (04) 11.1.... (04)
.1...1.. (04) ..1..... (04) 111111.1 (03) 1111111. (03) 1111..11 (03)
1111.1.. (03) .....1.. (03) 1111.111 (02) 111...11 (02) 111.11.. (02)
1.11.1.. (02) ..111... (02) .1.11... (02) 1..1.1.. (02) 1.1..1.. (02)
111..... (02) .11..... (02) ......1. (02) ....1... (02) 111..111 (01)
111.1.11 (01) 1.111..1 (01) 1.1111.. (01) .1.1..11 (01) .1.11.1. (01)
.11.1..1 (01) 1....111 (01) 1..111.. (01) 1.1...11 (01) 1.1..11. (01)
11....11 (01) 11.1.1.. (01) 11.11... (01) 111..1.. (01) 111.1... (01)
.11...1. (01) 1.....11 (01) 1...11.. (01) 1.1.1... (01) ......11 (01)
....1..1 (01) ...1.1.. (01) ...11... (01) ..1.1... (01) .1....1. (01)
1....1.. (01) .......1 (01)
```



Top-10 Folds in a Genome

Rank	*M. genitalium* Superfamily	#	*B. subtilis* Superfamily	#	*E. coli* Superfamily	#
1	P-loop hydrolase	60	P-loop hydrolyase	173	P-loop hydrolase	191
2	SAM methyl-transferase	16	Rossmann domain	165	Rossmann domain	158
3	Rossmann domain	13	Phosphate-binding barrel	79	Phosphate-binding barrel	64
4	Class I synthetase	12	PLP-transferase	44	PLP-transferase	38
5	Class II synthetase	11	CheY-like domain	36	CheY-like domain	36
6	Nucleic acid binding dom.	11	SAM methyl-transferase	30	Ferredoxins	35
Total ORFs		479		4268		4268
with Common Superfamilies		105 (22%)		465 (11%)		458 (11%)

Eubacteria

Rank	*S. cerevisiae* Superfamily	#	*C. elegans* Superfamily	#
1	P-loop hydrolyase	249	Protein kinase	429
2	Protein kinase	123	P-loop hydrolyase	411
3	Rossmann domain	90	Ligand-binding NR dom.	254
4	RNA-binding domain	75	C-type lectin	253
5	SAM methyl-transferase	63	alpha/beta hydrolase	180
6	Ribonuclease H-like	57	Ig superfamily	149
Total ORFs		6218		19,099
with Common Superfamilies		560 (9%)		1676 (8%)

Eukaryote

Rank	*M. thermo-autotrophicum* Superfamily	#	*A. fulgidus* Superfamily	#
1	P-loop hydrolyase	93	P-loop hydrolyase	118
2	Phosphate-binding barrel	54	Rossmann domain	104
3	Rossmann domains	53	Phosphate-binding barrel	56
4	Ferredoxins	48	Ferredoxins	49
5	SAM methyl-tranferase	17	SAM methyl-tranferase	24
6	PLP-transferases	15	PLP-transferases	18
Total ORFs		1869		2409
with Common Superfamilies		252 (14%)		309 (13%)

Archaea

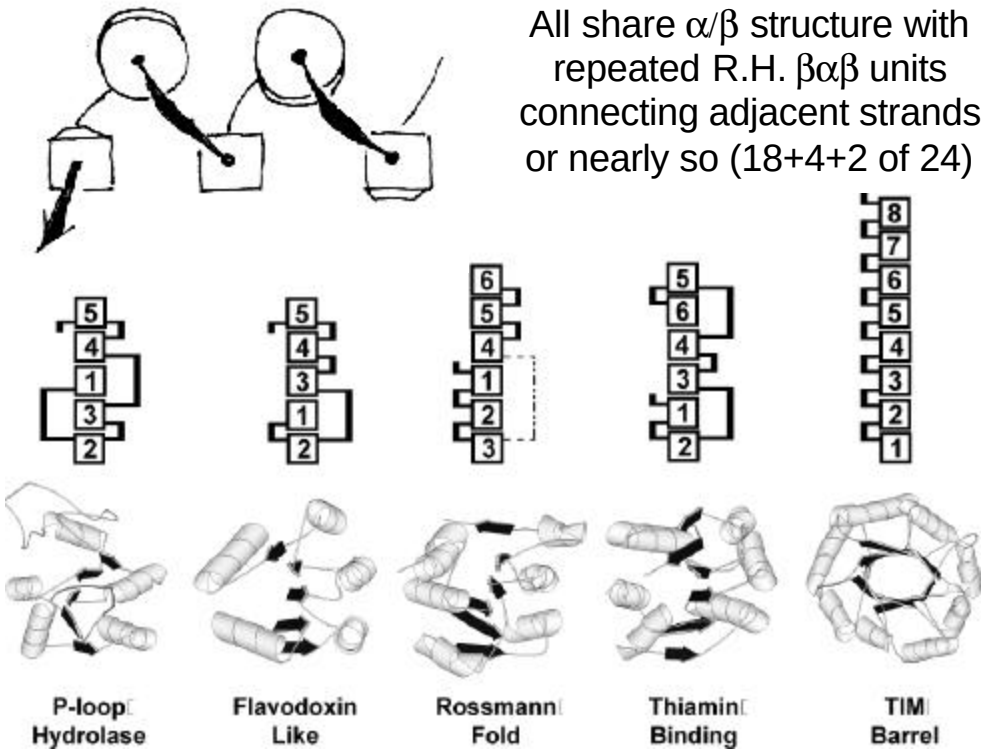
Depends on comparison method, DB, &c
(new top superfamilies via ψ-Blast left, old pairwise top folds via FASTA right)
Intersection of top-10 to get shared and common

#	Class	HI Fold	Representative Structure (PDB selection)
18	α/β	Rossmann Fold	2ohx A:175-324
13	α/β	NTP hydrolases with P-loop	1gky
12	α/β	Flavodoxin-like	3chy
10	α/β	TIM barrel	1tim A:
10	α+β	Ferredoxin-like	1fxd
10	α/β	Ribonuclease H-like	2rn2
6	α/β	Periplasmic binding protein-like II	1sbp
5	α/β	Periplasmic binding protein-like I	2dri
5	α+β	Like Class II aaRS synthetases	1sry A:111-421
4	β	OB-fold	1pyp
4	α/β	Thiamin-binding Fold	1pvd A:2-181

#	Class	SC Fold	Representative Structure (PDB selection)
84	α+β	Protein kinases (catalytic core)	1irk
49	α/β	NTP hydrolases with P-loop	1gky
35	α/β	Rossmann Fold	2ohx A:175-324
31	α/β	TIM barrel	1tim A:
25	α/β	Ribonuclease H-like	2rn2
18	S	Classic zinc finger	1zaa C:
14	α+β	Ubiquitin conjugating enzyme	1aak
12	β	GroES-like	1acy L:109-211
10	α/β	Thioredoxin-like	1trx
9	α/β	Thiamin-binding Fold	1pvd A:2-181
5 x 8	...	...	...
7	α/β	Flavodoxin-like	3chy

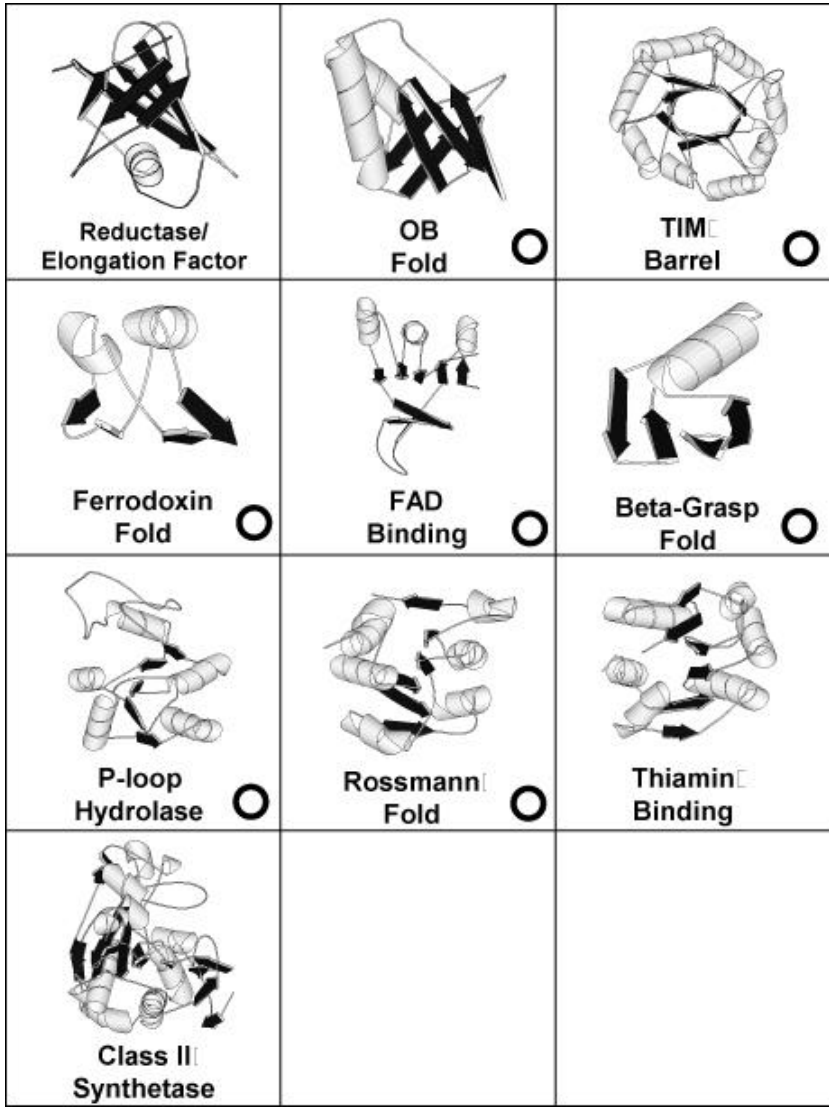
#	Class	MJ Fold	Representative Structure (PDB selection)
19	α+β	Ferredoxin-like	1fxd
10	α/β	NTP hydrolases with P-loop	1gky
7	α/β	TIM barrel	1tim A:
6	α/β	Rossmann Fold	2ohx A:175-324
5	α	Histone-fold	1ntx
4	α/β	Thiamin-binding Fold	1pvd A:2-181
4	α/β	Flavodoxin-like	3chy
4	β	Reduct./elongation fac. dom.	1efg A:283-403
3	α+β	ATP-grasp	1bnc A:115-330
3	α/β	PLP-dependent transferases	1dka
3	α/β	ATP pyrophoshatases	1gpm A:208-404

Characteristics of Common, Shared Folds: βαβ structure, superfold



All share α/β structure with repeated R.H. βαβ units connecting adjacent strands or nearly so (18+4+2 of 24)

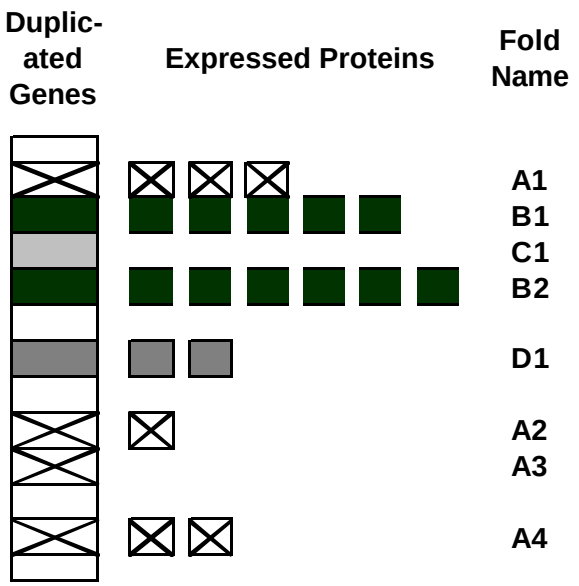
HI, MJ, SC vs scop 1.32



O =superfold (Thornton)

HI, MJ, MP, MG, SC, HP, SS, EC vs scop 1.35

Top-10 Folds according to Expression



Tot.	Dup.	Exp.
A	4	6
B	2	11
C	1	0
D	1	2

- Previous top-10 measures duplication
- Now weight by expression using data from Brown et al.

Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	**1**	**3**	**4**
NTP Hydrolases with P-loop	1gky	**2**	**1**	**2**
Classic Zn finger	1ard	**3**	9	**5**
Ribonuclease H-like motif	2rn2	**4**	**2**	**1**
Rossmann Fold	1xel	**5**	**4**	**3**
Zn2/Cys6 DNA-binding dom.	125d	**6**	**6**	**7**
7-bladed beta-propeller	2bbk-H	**7**	8	16
TIM-barrel	1byb	8	**5**	**6**
like Ferrodoxin	1fxd	9	**7**	10
DNA-binding 3-helix bundle	1enh	10	30	36
…	…			
GroES-like	1lep-A	17	10	9
…	…			
like HSP70, Ct-dom.	1dkz-A	22	11	8

What are the most common folds: Overall? In plants? In animals?

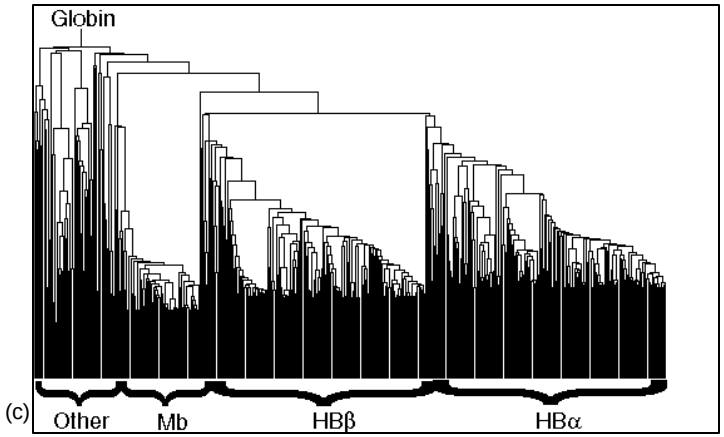
Example Structure (PDB)	Class	Fold Name	Families Num. Seq.	Num. of Sequences Total	Virus	Eubacteria	Eukaryote Plant	Eukaryote Metazoan	Eukaryote Other
Totals			719	37706	3139	7032	4960	13919	1928
Overall Top-10									
1REI-A	β	Immunoglobulin-like	32	13	0	1	0	25	0
6TIM-B	αβ	TIM-barrel	29	6	0	7	20	2	13
1ATP-E	◇	Protein Kinases (catalytic core)	1	4	3	0	3	6	6
1FXD	◇	Ferredoxin-like	17	4	2	2	17	0	8
1AKE-A	αβ	NTP Hydrolases containing P-loop	9	3	0	5	3	2	7
1HDD-C	α	DNA-binding 3-helical bundle	13	3	0	0	2	5	0
2KSD-A	αβ	Rossmann Fold (NAD binding)	11	3	0	7	3	1	3
1MBD	α	Globin-like	3	2		1	0	4	1
2RN2	αβ	like Ribonuclease H	15	2	5	1	2	1	5
1ZNF	S	Classic Zinc Finger	2	1	0			3	1
Sequence Family Top-11									
1REI-A	β	Immunoglobulin-like	32	13	0	1	0	25	0
6TIM-B	αβ	TIM-barrel	29	6	0	7	20	2	13
1FXD	◇	Ferredoxin-like	17	4	2	2	17	0	8
2RN2	αβ	like Ribonuclease H	15	2	5	1	2	1	5
1FXP	β	◇ β-fold	15	0	0	1	0	0	0
1FTX	S	Small inhibitors, toxins, lectins	14	0			3	0	0
2TBV-C	β	Viral coat and capsid proteins	14	1	12				
1HDD-C	α	DNA-binding 3-helical bundle	13	3	0	0	2	5	0
2KSD-A	αβ	Rossmann Fold (NAD binding)	11	3	0	7	3	1	3
1RCF	αβ	Flavodoxin-like	11	0	0	4	0	0	0
1RCB	α	4-helical cytokines	11	0	0	0		2	

	Fold Name	Number	Percent of Sequences Virus	Eubacteria	Eukaryote Plant	Eukaryote Metazoan	Eukaryote Other
Plant Top-10							
α/β	TIM-barrel	29	6	3	7	20	2	13
O	like Ferredoxin	17	4	2	2	17	3	8
α/β	NTP Hydrolases containing P-loop	9	3	3	5	3	2	7
O	Protein Kinases (catalytic core)	1	4	3	3	3	6	6
S	Small inhibitors, toxins, lectins	14	3			3	3	3
α/β	Rossmann Fold (NAD binding)	11	3	3	7	3	1	3
O	RuBisCO (small subunit)	1	3		3	2		3
β	like Concanavalin A	6	3	3	3	2	3	2
α	like Hydrophobic Seed Protein	2	3			2		
α/β	like Ribonuclease H	15	2	5	1	2	1	5
Metazoan Top-10								
β	like Immunoglobulin	32	13	3	1	3	25	3
O	Protein Kinases (catalytic core)	1	4	3	3	3	6	6
α	DNA-binding 3-helical bundle	13	3	3	3	2	5	3
α	like Globin	3	2		1	3	4	1
S	Classic Zinc Finger	2	1	3			3	1
α/β	NTP Hydrolases containing P-loop	9	3	3	5	3	2	7
β	Trypsin-like serine proteases	4	1	1	3		2	3
α	Cytochrome P450	1	1		3	3	2	1
S	like Glucocort. receptor (DNA-binding)	4	1	3			2	3
α	EF-hand	3	1		3	1	2	1

An Issue with Fold Counting: Biases in the Databanks

- Over-representation of certain species and functions in the databanks (e.g. human v. plant globins, Ig's)
  - Nevertheless HI top-10 like eubacterial top-10
- PDB small, biased sample of genome (6-12%)
- Diff. numbers with diff. comparison sensitivity
  - FASTA, HMM, &c
  - Some Correction with Seq. Weighting, Diff. Sampling
  - Uniform sampling is better than high sensitivity for some and low for others (ψ-blast problem)
  - Best to avoid FPs than FNs for Venn

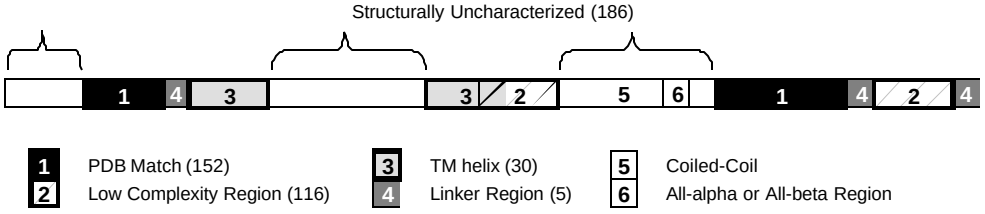
Example Structure (PDB)	Fold Name	Percentage of known folds in genome	Rank in eubacterial Top-10
Top-10 in a bacterial genome (H. influenzae)			
2HSD-A	Rossmann Fold (NAD binding)	9.6	**1**
1AKE-A	NTP Hydrolases containing P-loop	5.7	**3**
1RCF	Flavodoxin-like	5.1	**4**
6TIM-B	TIM-barrel	4.5	**2**
1FXD	Ferredoxin-like	4.2	**5**
2RN2	like Ribonuclease H	3.0	16
1SBP	like Periplasmic binding protein (class II)	3.0	**11**
2DRI	like Periplasmic binding protein (class I)	3.0	19
1SRY-*	Class II aaRS and biotin synthetases	2.7	50
1PYP	OB-fold	2.7	**9**

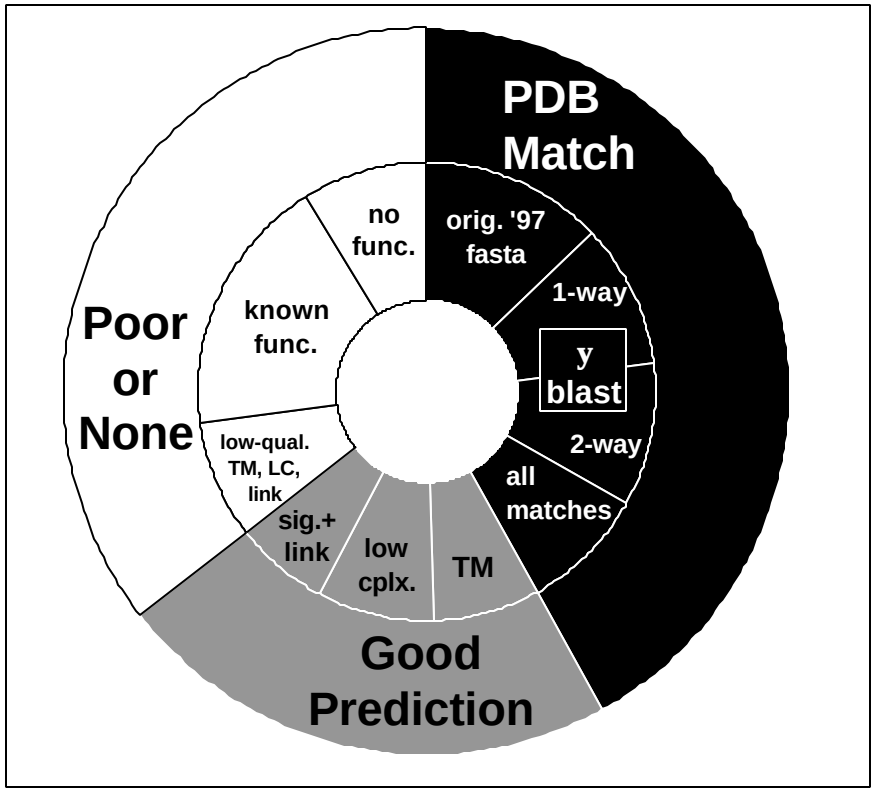


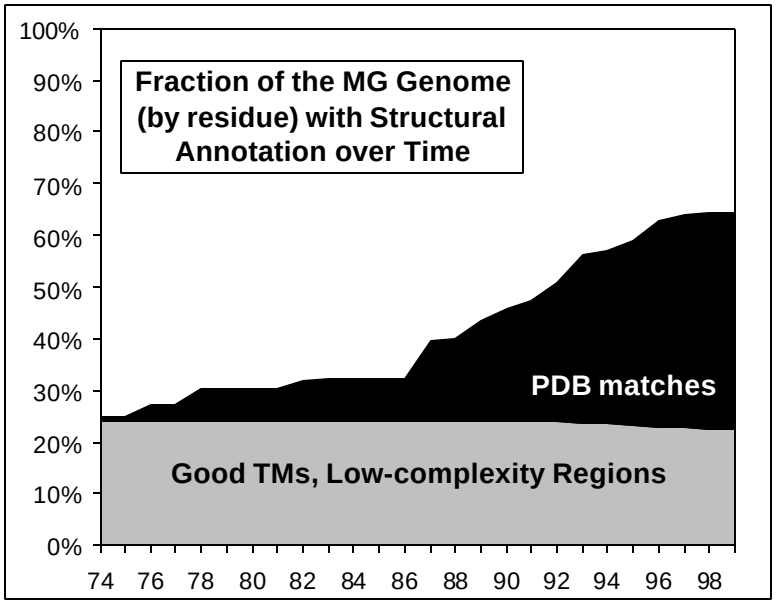
Same Issues with Real US Census!! Sampling

Know All Folds in a Genome: How are we doing on MG?

- MG smallest genome with 479 ORFs
- Separate PDB Match, TMs, LC (SEG), linkers
- How many residues in genome matched by known folds, in 1975, ‘76, ‘77...’00...’50
- The impact of PSI-blast in comparison to pairwise methods
  - ◊ Two way PSI-blast gives an improvement (genome vs PDB, PDB vs. genome)
- Union of many sets of PDB matches finds >40% of a.a. and more than half the ORFs (242/479)
  - ◊ (Eisenberg, Godzik, Bork, Koonin, Frishman)
- ~65% structurally characterized

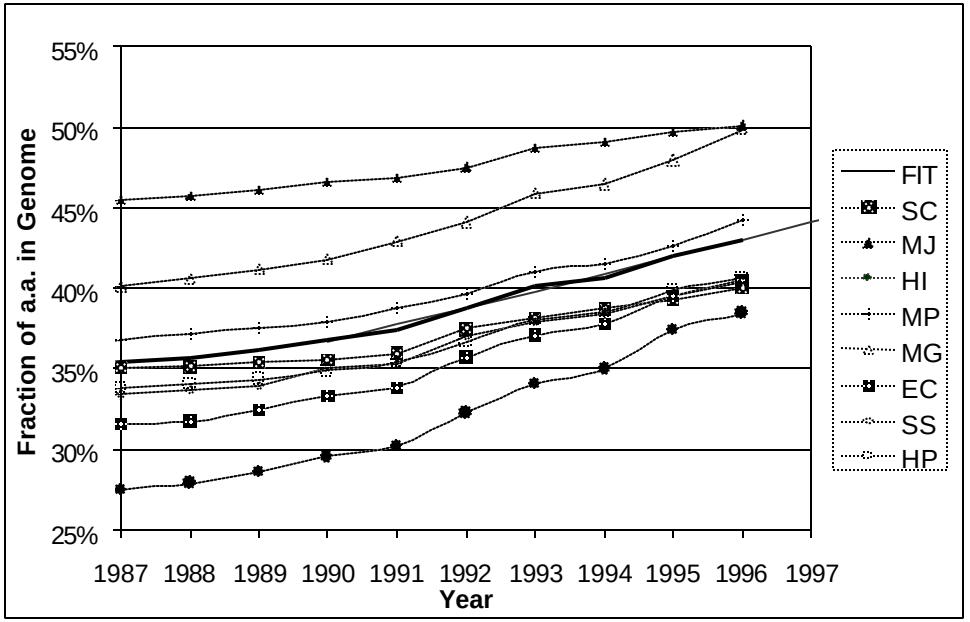


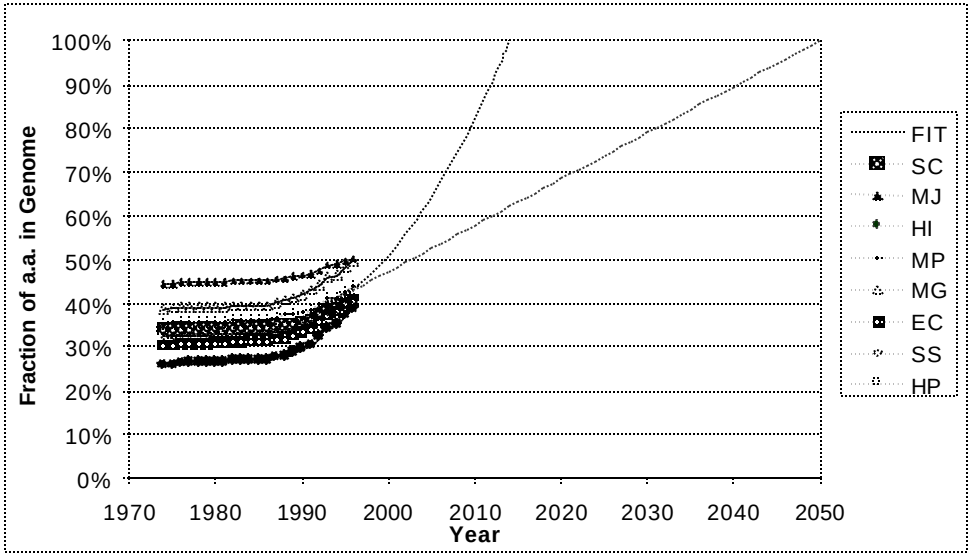


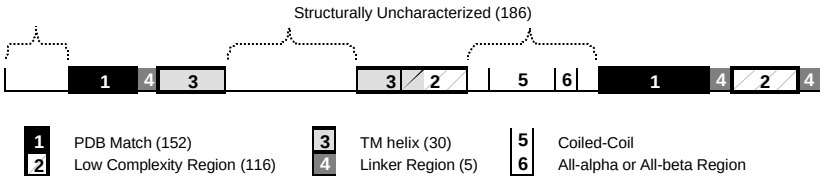


Know All Folds in Genome: MG Optimistic → Prediction

- Just use one pairwise method for matching
- Multiple, big genomes (e.g. SC)







Comparing Genomes in terms of Protein Structure: Surveys of a **Finite** Parts List

1 Library of Known Folds

Importance of Statistics. Scop auto-alignments. P-values from EVD, same as sequences.

2 Census of Known Folds

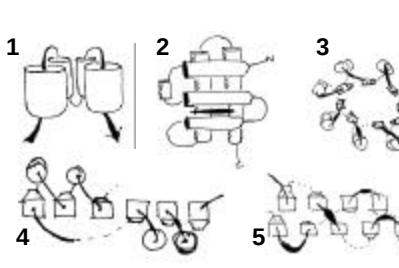
Which folds in which organisms: E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression. Repeated βαβ. Biases. Extent of MG fold assignment (65%)

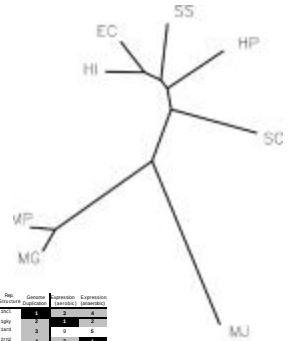
3 Prediction of Unknown Folds

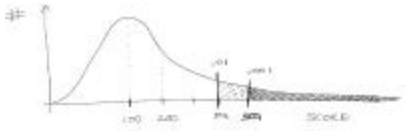
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Biases: Can extrapolate from known structures to genomes?

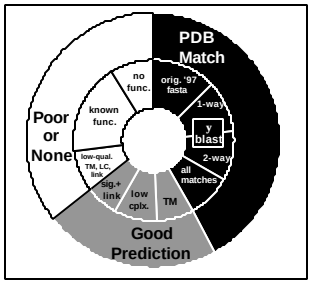
4 Fold-Function Relationships

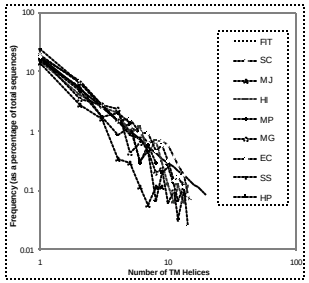
How many folds per function? Func. per fold? 331 of ~20K combinations. TIM most versatile scaffold.









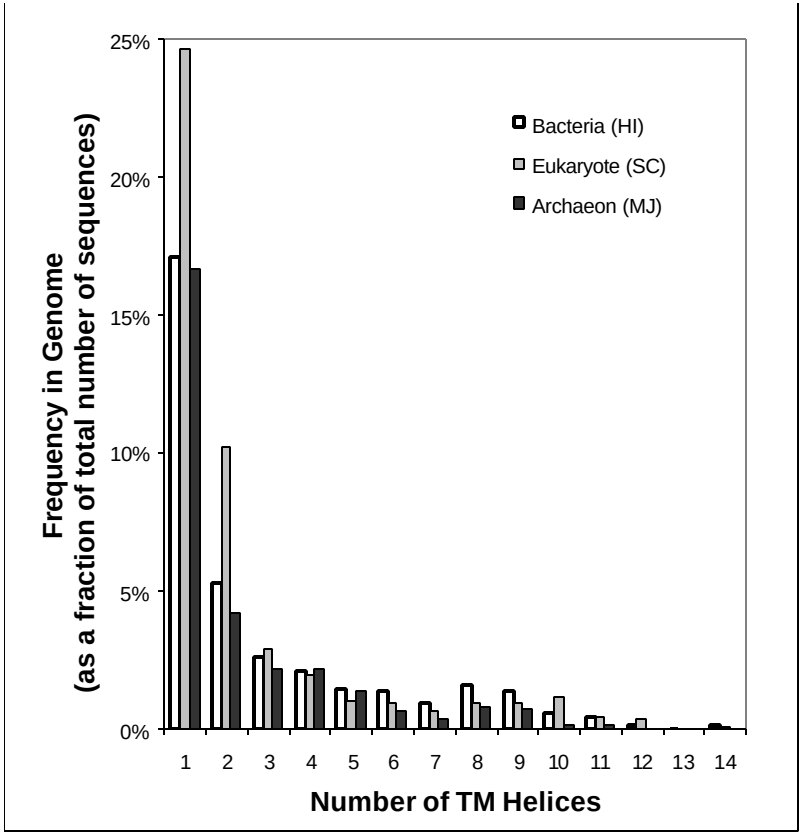


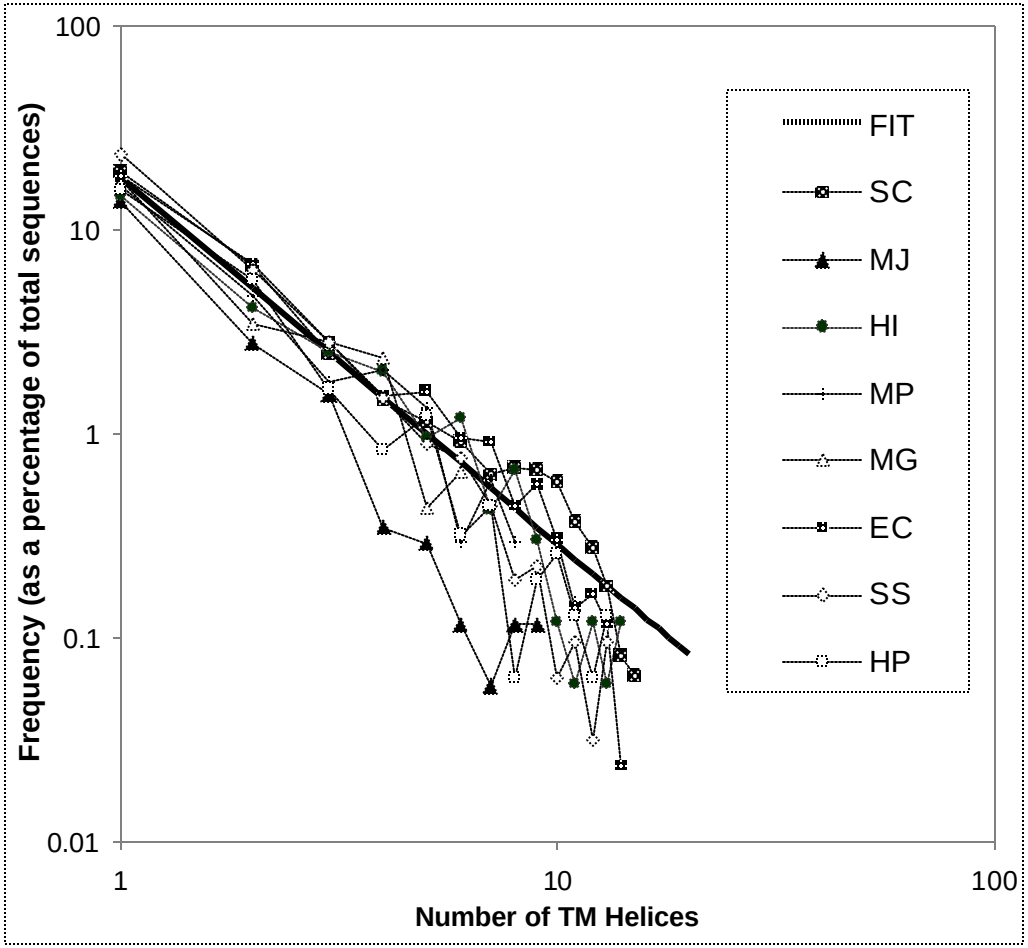
ENZYME	SCOP A	SCOP B	SCOP A/B	SCOP A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

TM-helix "prediction"

- TM prediction (KD, GES). Count number with 2 peaks, 3 peaks, &c.
- Yeast has more mem. prots., esp. 2-TMs
- Similar conclusions to others: von Heijne, Rost, Jones, &c.

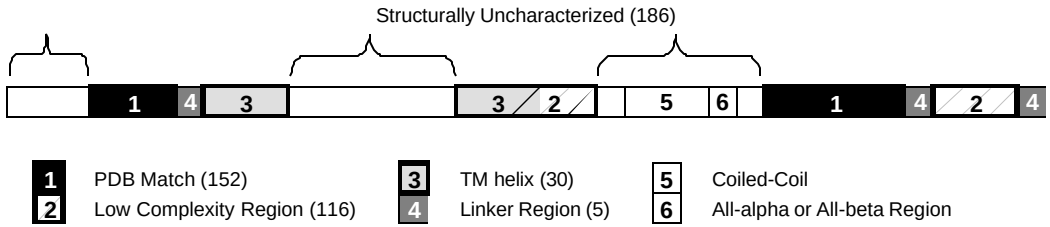
- No preference for particular supersecondary structures: 7-TM's
- Freq. of Number of TM helixes follows a Zipf-like law: $\mathbf{F=1/[5n^2]}$





2º Structure Prediction

- Bulk prediction of 2º struc. in genomes
- Same fraction of α and β (by element, half each)
- Both overall and only for unknown soluble proteins.



Fraction of residues Predicted to be in...	**strand**	**helix**
Avg	**17%**	**39%**
SD	**1%**	**2%**
EC	17%	39%
HI	16%	41%
HP	15%	42%
MG	17%	39%
MJ	19%	37%
MP	17%	39%
SC	17%	34%
SS	16%	38%

- Diff From PDB:
  31% helical and 21% strand.
- Related results: Frishman

Not expected since……

Different Amino Acid Composition Should Give Different 2º Structure

Each a.a. has different propensity for local structure
->
Different Compositions (K from 4.4 in EC to 10.4 in MJ, Q too)
->
Different Local Structure (but compensation?)

Propensities from Regan (beta) and Baldwin (alpha)

	Amino Acid Composition								Propensity (kcal/mole)		
	EC	HI	SS	SC	HP	MP	MG	MJ	TM-hlx	helix	strand
K	4.4	6.3	4.2	7.3	8.9	8.6	9.5	10.4	8.8	-1.5	-0.4
C	1.2	1.0	1.0	1.3	1.1	.8	.8	1.3	-2	-1.1	-0.8
R	5.5	4.5	5.1	4.5	3.5	3.5	3.1	3.8	12.3	-1.9	-0.4
N	4.0	4.9	4.0	6.1	5.9	6.2	7.5	5.3	4.8	-1	-0.5
Q	4.4	4.6	5.6	3.9	3.7	5.4	4.7	1.5	4.1	-1.3	-0.4
A	9.5	8.2	8.5	5.5	6.8	6.7	5.6	5.5	-1.6	-1.9	0
I	6.0	7.1	6.3	6.6	7.2	6.6	8.2	10.5	-3.1	-1.2	-1.3
H	2.3	2.1	1.9	2.2	2.1	1.8	1.6	1.4	3	-1.1	-0.4
S	5.8	5.8	5.8	9.0	6.8	6.5	6.6	4.5	-0.6	-1.1	-0.9
M	2.8	2.4	2.0	2.1	2.2	1.6	1.5	2.2	-3.4	-1.4	-0.9
P	4.4	3.7	5.1	4.3	3.3	3.5	3.0	3.4	0.2	3	>3.0
G	7.4	6.6	7.4	5.0	5.8	5.5	4.6	6.3	-1	0	1.2
F	3.9	4.5	4.0	4.5	5.4	5.6	6.1	4.2	-3.7	-1	-1.1
E	5.7	6.5	6.0	6.5	6.9	5.7	5.7	8.7	8.2	-1.2	-0.2
Y	2.9	3.1	2.9	3.4	3.7	3.2	3.2	4.4	0.7	-1.2	-1.6
V	7.1	6.7	6.7	5.6	5.6	6.5	6.1	6.9	-2.6	-0.8	-0.9
T	5.4	5.2	5.5	5.9	4.4	6.0	5.4	4.0	-1.2	-0.6	-1.4
D	5.1	5.0	5.0	5.8	4.8	5.0	4.9	5.5	9.2	-1	0.9
L	10.6	10.5	11.4	9.6	11.2	10.3	10.7	9.5	-2.8	-1.6	-0.5
W	1.5	1.1	1.6	1.0	.7	1.2	1.0	.7	-1.9	-1.1	-1
total propensity											
α	-1.00	-1.02	-0.96	-1.00	-1.05	-1.03	-1.05	-1.01			
β	-0.27	-0.33	-0.26	-0.36	-0.37	-0.38	-0.42	-0.36			

Supersecondary structure words

- Look at super-secondary patterns ("words" such as αα or βαβ) in predictions
- Compare observed freq. with expected freq.
  odds = f(αβ)/f(α)f(β)
  (Freq. Words, Karlin)
- Do have differences between genomes (and PDB) here

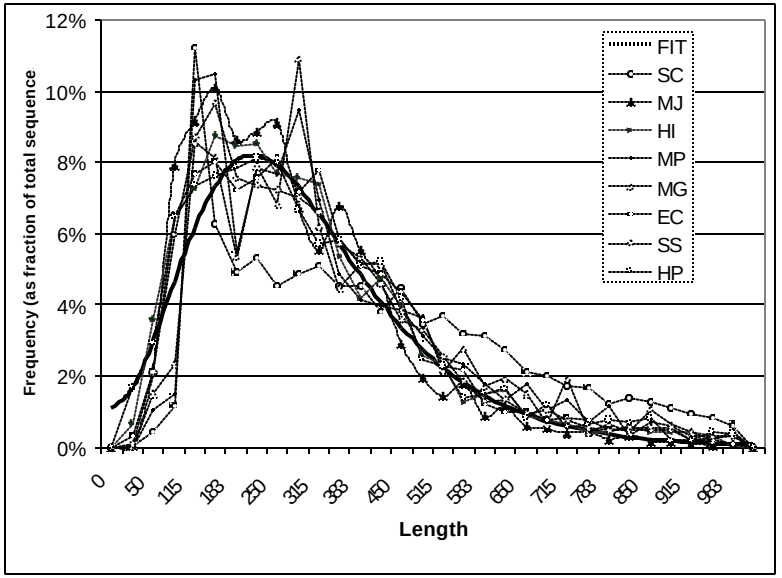
HI more αα, ααα, αααα ...

SC more ββ, βββ, βββββ...

MJ more αβαβ, βαβα …

Super-Secondary Structure "Word"	Maximum Difference between 3 Genomes	Relative Abundance (Odds Ratio)			
		HI	MJ	SC	PDB
ββ	26%	0.96	1.06	1.24	1.22
αα	15%	0.97	0.85	0.83	0.85
αβ	10%	1.09	1.09	0.99	0.95
βα	7%	0.98	1.00	0.93	0.99
bbb	41%	0.96	1.15	**1.46**	1.62
aaa	19%	**1.01**	0.83	0.84	0.92
aba	18%	1.04	1.03	0.87	1.16
ααβ	15%	1.03	0.97	0.89	0.70
bab	12%	1.15	**1.24**	1.10	1.19
βαα	11%	0.93	0.87	0.83	0.78
ββα	9%	0.90	0.94	0.99	0.82
αββ	6%	0.97	0.98	1.03	0.80
bbbb	54%	1.03	1.35	**1.78**	2.28
aaaa	29%	**1.10**	0.82	0.89	1.18
βββα	25%	0.85	0.94	1.10	0.98
baba	23%	1.11	**1.18**	0.94	1.48
abab	21%	1.21	**1.23**	0.99	1.39
αβαα	21%	1.00	0.95	0.81	1.00
…	…	…	…	…	…

How Representative are the Known Structures of the Proteins in a Complete Genome? The issue of Bias

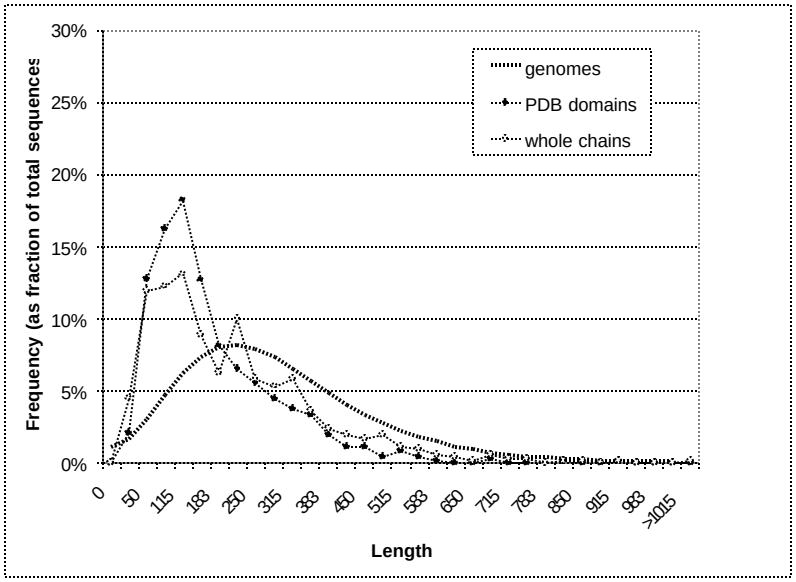


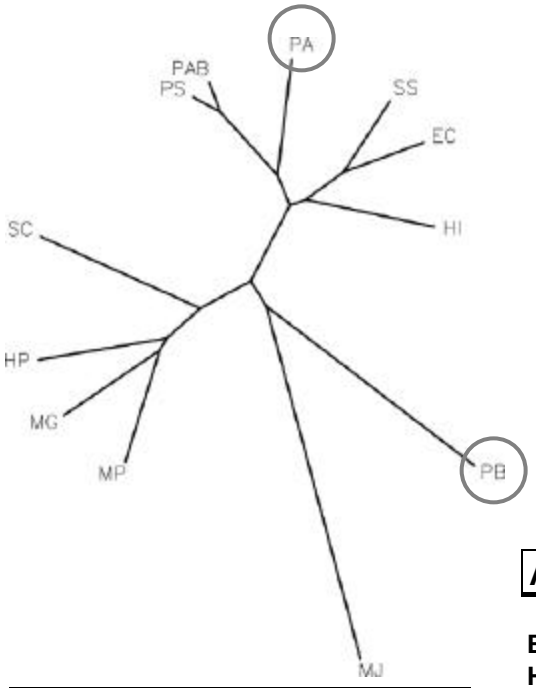
Assess 2º,TM predictions
- (+) comprehensive, statistical
- (-) predictions inaccurate (~65%)
- (-) extrapolate from PDB (esp. TM), domain problem

Is prediction (extrapolation) based on known structures justified?

Length: Genomes Sequences are longer than those in Known Structures

340 aa for avg. genome seq.
(470 aa for yeast)
205 aa for PDB chain
~160 aa for PDB domain





Amino Acid Composition

How Representative are the Known Structures of the Proteins in Complete Genome?

Name	Soluble PDB		= all-β	+ all-α
A	8.40%		6.8%	9.2%
C	1.72%		1.6%	1.4%
D	5.91%		5.9%	5.8%
E	6.29%		5.2%	7.3%
F	3.94%		4.2%	4.2%
G	7.79%		8.4%	6.4%
H	2.19%		2.1%	2.2%
I	5.54%		5.4%	5.1%
K	6.02%		5.6%	6.5%
L	8.37%		7.3%	9.6%
M	2.15%		1.7%	2.4%
N	4.57%		5.3%	4.4%
P	4.70%		5.1%	4.4%
Q	3.73%		3.5%	4.2%
R	4.78%		4.2%	5.4%
S	5.97%		7.2%	5.7%
T	5.87%		7.2%	5.2%
V	6.96%		7.6%	5.7%
W	1.46%		1.7%	1.5%
Y	3.64%		3.8%	3.5%

ABS.

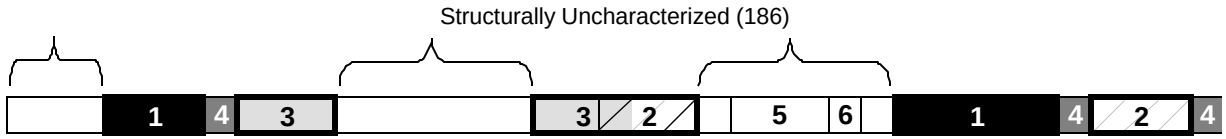
	rms	K	I	C	Q	W	N	F	L	G	A	P	S	R	H	M	E	D	T	Y	V
EC		4.4	6.0	1.2	4.4	1.5	4.0	3.9	10.6	7.4	9.5	4.4	5.8	5.5	2.3	2.8	5.7	5.1	5.4	2.9	7.1
HI		6.3	7.1	1.0	4.6	1.1	4.9	4.5	10.5	6.6	8.2	3.7	5.8	4.5	2.1	2.4	6.5	5.0	5.2	3.1	6.7
SS		4.2	6.3	1.0	5.6	1.6	4.0	4.0	11.4	7.4	8.5	5.1	5.8	5.1	1.9	2.0	6.0	5.0	5.5	2.9	6.7
SC		7.3	6.6	1.3	3.9	1.0	6.1	4.5	9.6	5.0	5.5	4.3	9.0	4.5	2.2	2.1	6.5	5.8	5.9	3.4	5.6
HP		8.9	7.2	1.1	3.7	.7	5.9	5.4	11.2	5.8	6.8	3.3	6.8	3.5	2.1	2.2	6.9	4.8	4.4	3.7	5.6
MP		8.6	6.6	.8	5.4	1.2	6.2	5.6	10.3	5.5	6.7	3.5	6.5	3.5	1.8	1.6	5.7	5.0	6.0	3.2	6.5
MG		9.5	8.2	.8	4.7	1.0	7.5	6.1	10.7	4.6	5.6	3.0	6.6	3.1	1.6	1.5	5.7	4.9	5.4	3.2	6.1
MJ		10.4	10.5	1.3	1.5	.7	5.3	4.2	9.5	6.3	5.5	3.4	4.5	3.8	1.4	2.2	8.7	5.5	4.0	4.4	6.9
AVG		7.5	7.3	1.1	4.2	1.1	5.5	4.8	10.5	6.1	7.0	3.8	6.4	4.2	1.9	2.1	6.5	5.1	5.2	3.3	6.4
SD		2.3	1.4	.2	1.3	.3	1.2	.8	.7	1.0	1.5	.7	1.3	.9	.3	.4	1.0	.3	.7	.5	.6

Diff.

EC	16	-25	8	-29	19	7	-15	-2	28	-6	13	-5	-3	16	3	28	-7	-14	-7	-22	1
HI	17	8	27	-38	24	-21	6	12	26	-15	-2	-20	-2	-6	-7	10	5	-17	-11	-14	-4
SS	20	-29	13	-39	49	9	-13	1	37	-6	1	11	-3	6	-15	-8	-2	-16	-6	-20	-4
SC	21	24	18	-21	5	-27	31	14	15	-36	-34	-7	51	-7	-2	-4	5	-4	0	-8	-20
HP	27	52	29	-34	0	-51	27	36	34	-26	-18	-29	14	-28	-4	2	11	-20	-25	1	-20
MP	28	45	18	-55	44	-17	35	41	24	-29	-20	-25	8	-27	-18	-28	-8	-17	2	-11	-7
MG	36	61	48	-50	27	-32	62	53	28	-41	-33	-36	11	-35	-28	-30	-8	-18	-8	-11	-12
MJ	38	77	88	-23	-61	-49	14	6	14	-19	-35	-28	-25	-20	-35	1	40	-8	-31	20	-2
AVG		26	31	-36	13	-23	19	20	26	-22	-16	-17	6	-13	-13	-4	4	-14	-11	-8	-9
RMS		45	39	38	35	31	30	28	27	25	24	23	21	21	18	18	16	15	15	15	11

Composition of Different Regions of Genomes

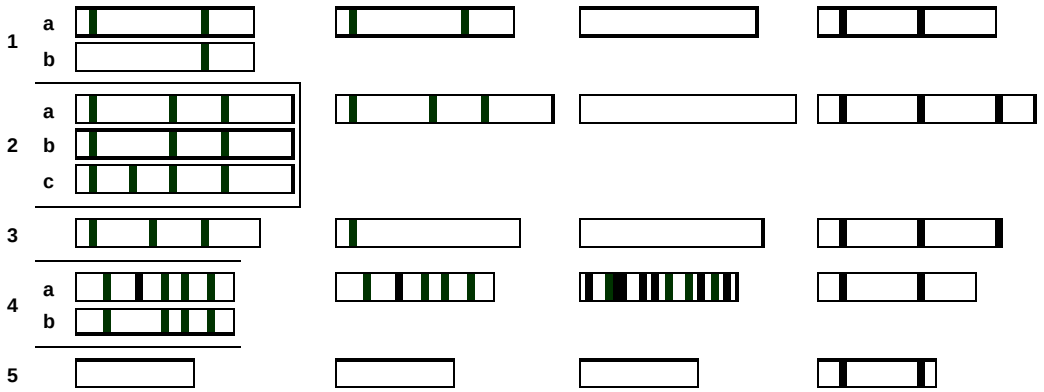
- Are composition differences uniform?
- Resampling
- Non-globular regions differ most in occurrence and composition
- Remove Repetitive Regions (SEG)



1 PDB Match (152)
2 Low Complexity Region (116)
3 TM helix (30)
4 Linker Region (5)
5 Coiled-Coil
6 All-alpha or All-beta Region

	AVG	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Statistics for Amino Acids										
Total Number	775998		1358465	505279	500616	170400	497968	237905	2900670	1033450
Fraction Masked by...										
PDB Match	8.7%	3.7%	11.1%	13.7%	8.8%	12.9%	7.1%	9.7%	6.2%	9.0%
Non-globular Region	21.7%	**6.9%**	16.7%	13.9%	22.2%	28.2%	35.1%	24.7%	23.9%	20.5%
TM-helix	4.9%	1.4%	7.3%	6.1%	4.8%	3.8%	2.9%	4.5%	5.2%	5.9%
Linker Region	5.1%	0.4%	5.3%	4.8%	4.8%	5.0%	5.0%	5.2%	4.6%	5.1%
Fraction Remaining										
Uncharacterized	59.7%	8.9%	59.6%	61.5%	59.4%	50.2%	49.9%	55.8%	60.0%	59.6%

	AVG	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Overall	23%	10%	16%	17%	27%	36%	38%	28%	21%	20%
PDB Match	18%	9%	12%	14%	24%	27%	34%	20%	12%	15%
Non Globular Region	**36%**	13%	32%	33%	39%	50%	52%	40%	42%	35%
TM-helix	**49%**	15%	55%	53%	55%	57%	55%	56%	56%	51%
Linker Region	27%	10%	22%	24%	29%	39%	33%	35%	21%	25%
Uncharacterized Region	23%	6%	15%	17%	26%	34%	32%	27%	20%	19%



PDB	Select	length	class	name
1sty	-	137	β	Staph nuclease
1cgp	a:9-137	129	β	CAP
1bgh	-	85	β	Gene V protein
1pht	-	83	β	SH3 domain
1tpf	a:	250	α/β	TIM
1wsy	a:	248	α/β	Trp Synthase
8dfr	-	186	α/β	DHFR
2rn2	-	155	α/β	Ribonuclease H
1brs	d:	87	α/β	Barstar
1gbs	-	185	α+β	Hen Lyzozyme
119l	-	162	α+β	T4 lysozyme
193l	-	129	α+β	alpha-Lactabumin
7rsa	-	124	α+β	RNAse A
1brn	l:	108	α+β	Barnase
1fkd	-	107	α+β	FK506
9rnt	-	104	α+β	RNAse T1
1sha	a:	103	α+β	SH2 domain
1ubi	-	76	α+β	Ubiquitin
1cse	i:	63	α+β	CI-2 inhibitor
1igd	-	61	α+β	B1 domain
1mbd	-	153	α	Globin
1hrc	-	105	α	Cytochrome c
2wrp	r:	104	α	Trp Repressor
1lli	a:	89	α	Cro Repressor
1cop	d:	66	α	Lambda Repressor
1rpo	-	61	α	ROP
1myk	a:	47	α	Arc Repressor
2zta	a:	31	α	GCN4 zipper
1btl	-	263	M	beta-Lactamase
1bpi	-	58	S	BPTI
AVG		116		

Name	Hydroph. Polar	Soluble PDB PS	biophys. proteins BP	Rel. Diff. BP/PS -1
P	**H**	4.7%	3.7%	-21%
F	**H**	4.0%	3.2%	-19%
M	**H**	2.1%	1.8%	-16%
D	*P*	6.0%	5.1%	-16%
V	**H**	7.0%	6.2%	-12%
C	**H**	1.7%	1.5%	-9%
S	*P*	6.0%	5.7%	-5%
G	.	7.8%	7.7%	-1%
I	**H**	5.6%	5.5%	-1%
N	*P*	4.6%	4.6%	0%
W	**H**	1.4%	1.5%	1%
T	*P*	5.8%	6.0%	2%
L	**H**	8.4%	8.7%	5%
A	.	8.4%	8.8%	6%
Y	.	3.7%	3.9%	6%
H	*P*	2.2%	2.4%	6%
Q	*P*	3.7%	4.0%	6%
R	*P*	4.8%	5.2%	9%
E	*P*	6.2%	7.0%	13%
K	*P*	5.9%	7.7%	30%

Biophysical Proteins

Proteins that inform our view of the folding process -- as compared to the PDB.

Shorter (116 v 161)

Fewer hydrophobes

Comparing Genomes in terms of Protein Structure: Surveys of a **Finite** Parts List

1 Library of Known Folds

Importance of Statistics. Scop auto-alignments. P-values from EVD, same as sequences.

2 Census of Known Folds

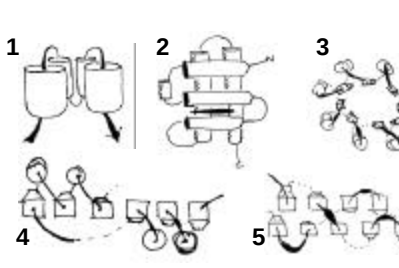
Which folds in which organisms: E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression. Repeated βαβ. Biases. Extent of MG fold assignment (65%)

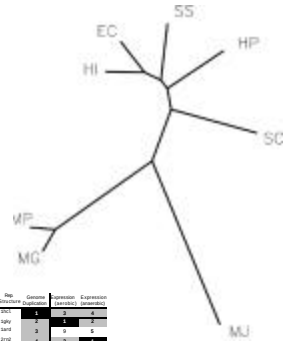
3 Prediction of Unknown Folds

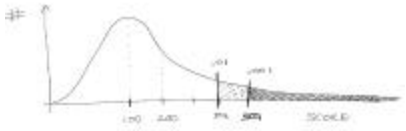
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Biases: Can extrapolate from known structures to genomes?

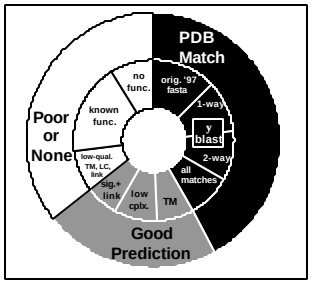
4 Fold-Function Relationships

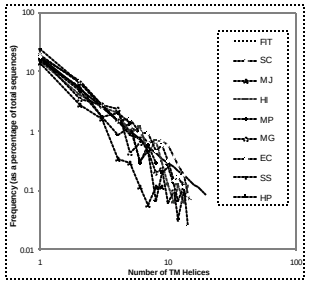
How many folds per function? Func. per fold? 331 of ~20K combinations. TIM most versatile scaffold.





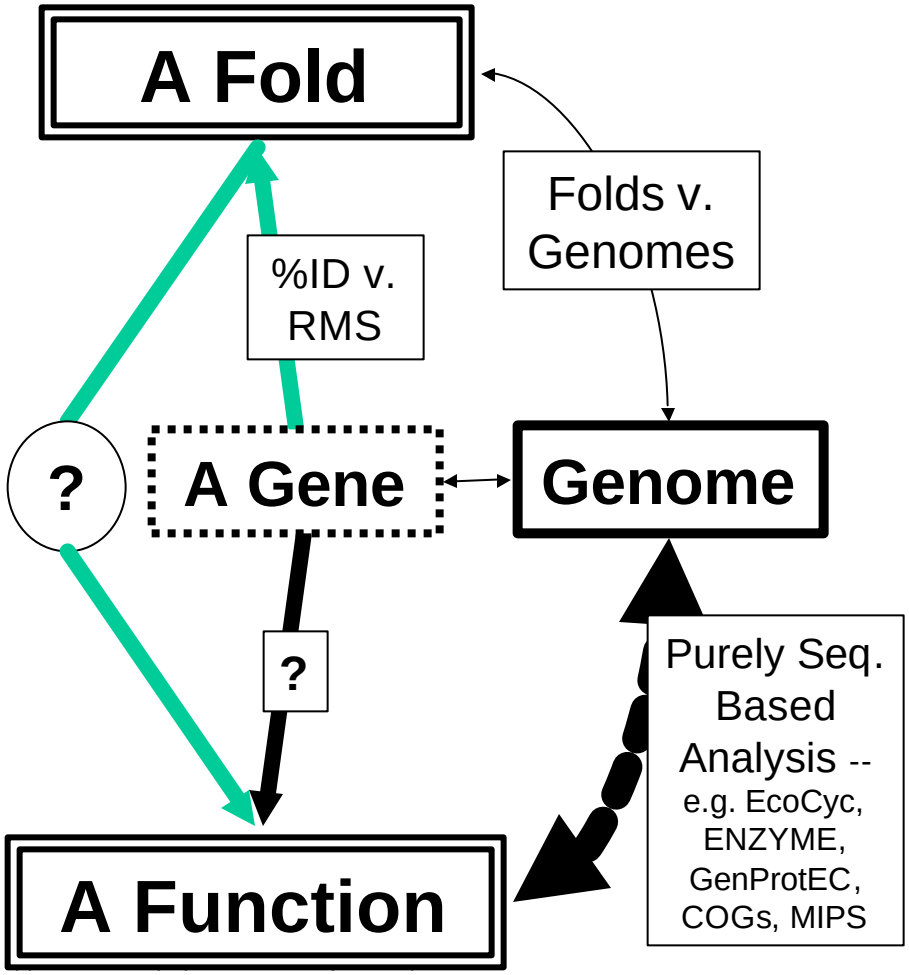






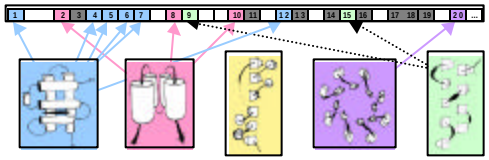
ENZYME	SCOP A	SCOP B	SCOP A/B	SCOP A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

Adding Structure to Functional Genomics, Function to Structural Genomics

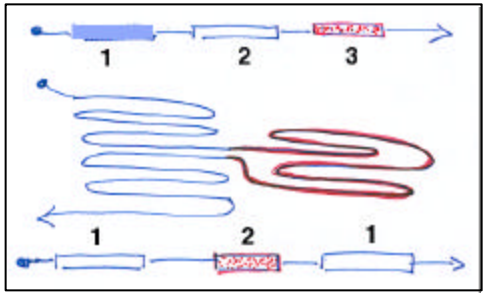


Why Structure? Do we really need it?

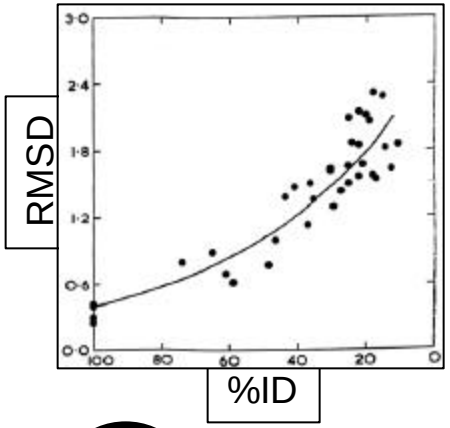
1 Most Highly Conserved



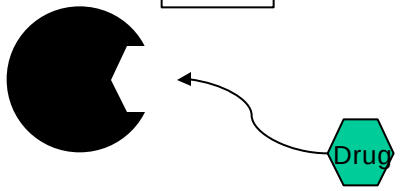
2 Precisely Defined Modules



3 Seq. ⇔ Struc. Clearer than Seq. ⇔Func.



4 Link to Chemistry, Drugs



Chothia & Lesk, 1986 -- 32 points

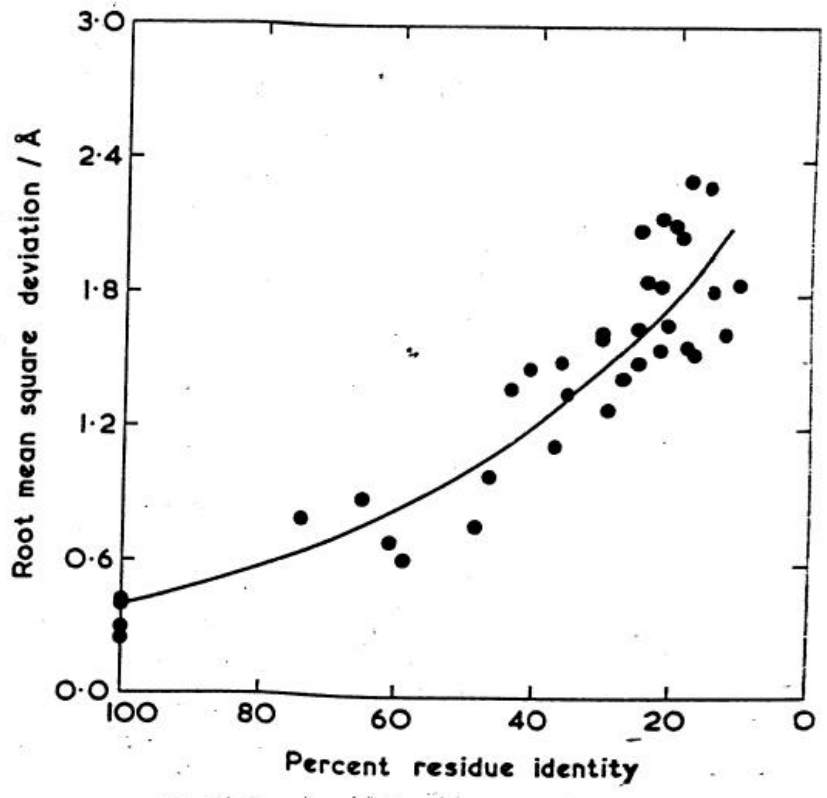


Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

EMBO J **4**: 823 (1986)

"The relation between the divergence of sequence and structure in proteins"

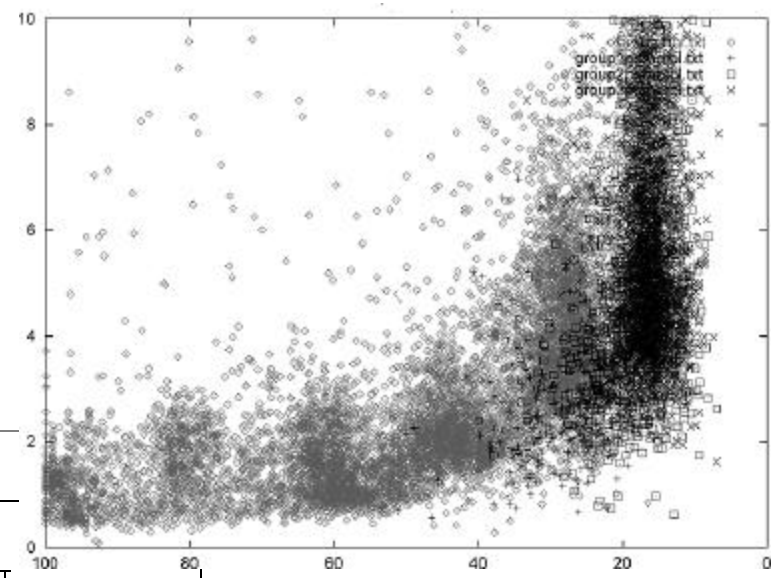
32 pairs of homologous proteins

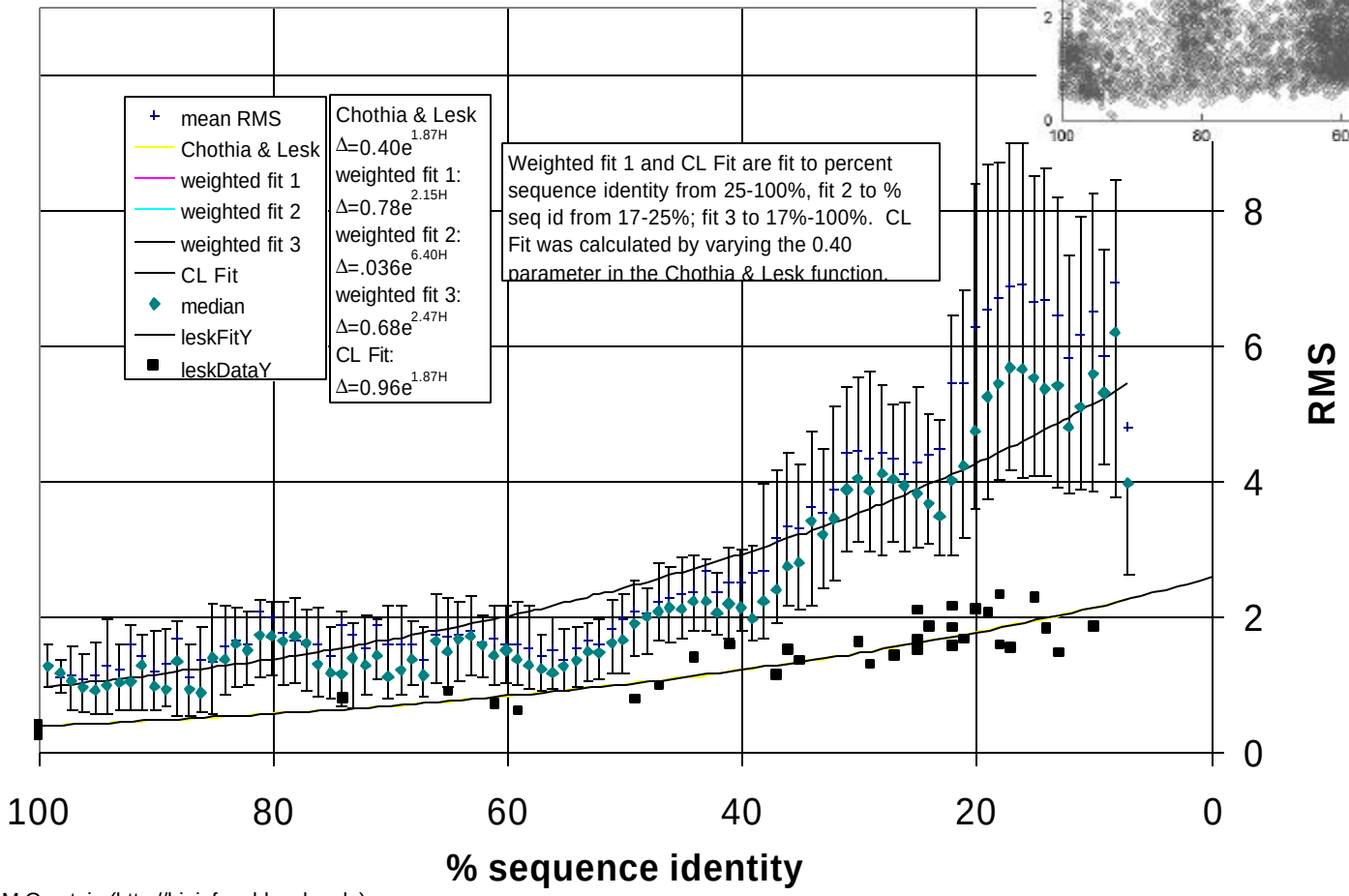
RMS, percent identity

$\Delta = 0.40\ e^{1.87H}$

Now redo with >16,000 pairs in scop + auto-alignments (pdb95d)....

Chothia and Lesk, revisited 16K points





C&L ‘86:

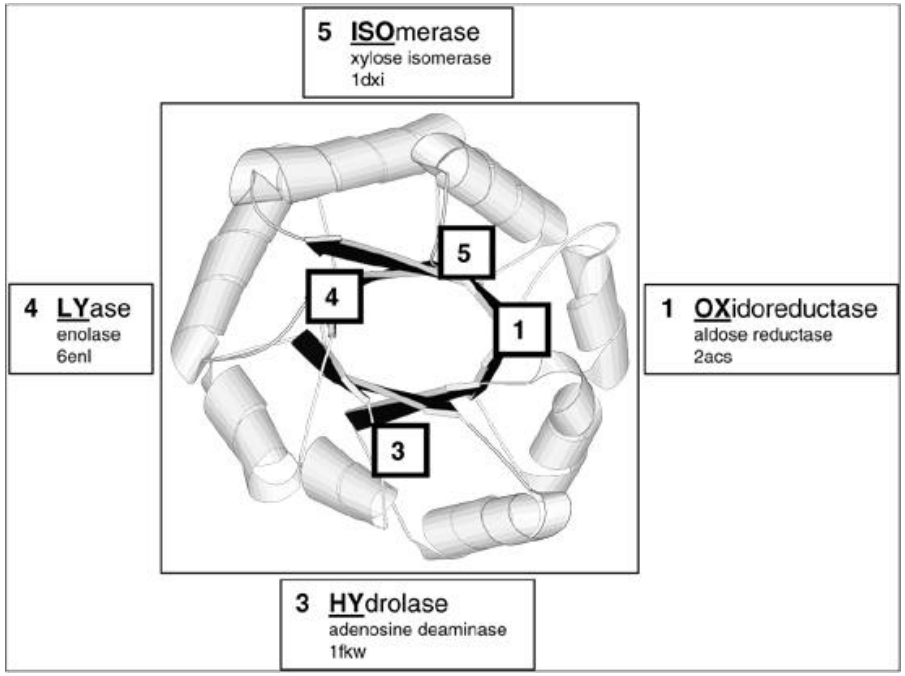
$\Delta = .4 \exp(1.9\ H)$

Here:

$\Delta = .8 \exp(2.2\ H)$

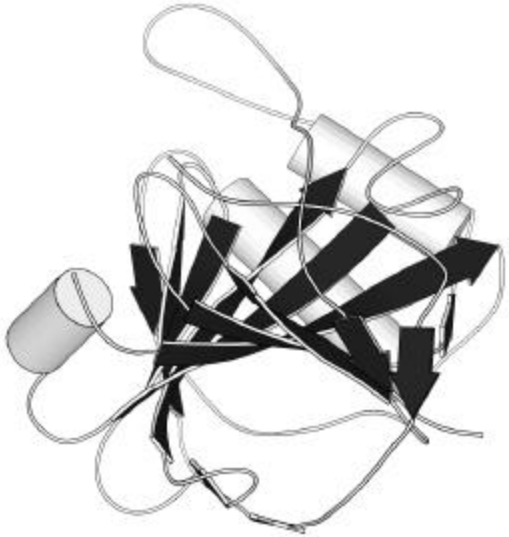
$\Delta = \exp(1.9\ H)$

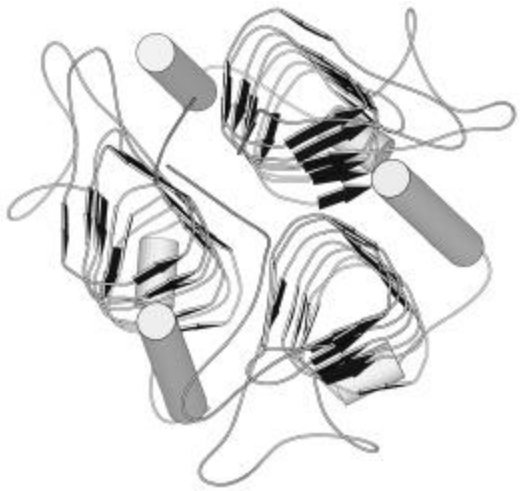
Fold-Function Combinations



Many Functions on the Same Fold
-- e.g. the TIM-barrel

- Two Different Folds Catalyze the Same Reaction -- e.g. Carbonic Anhydrases (4.2.1.1)



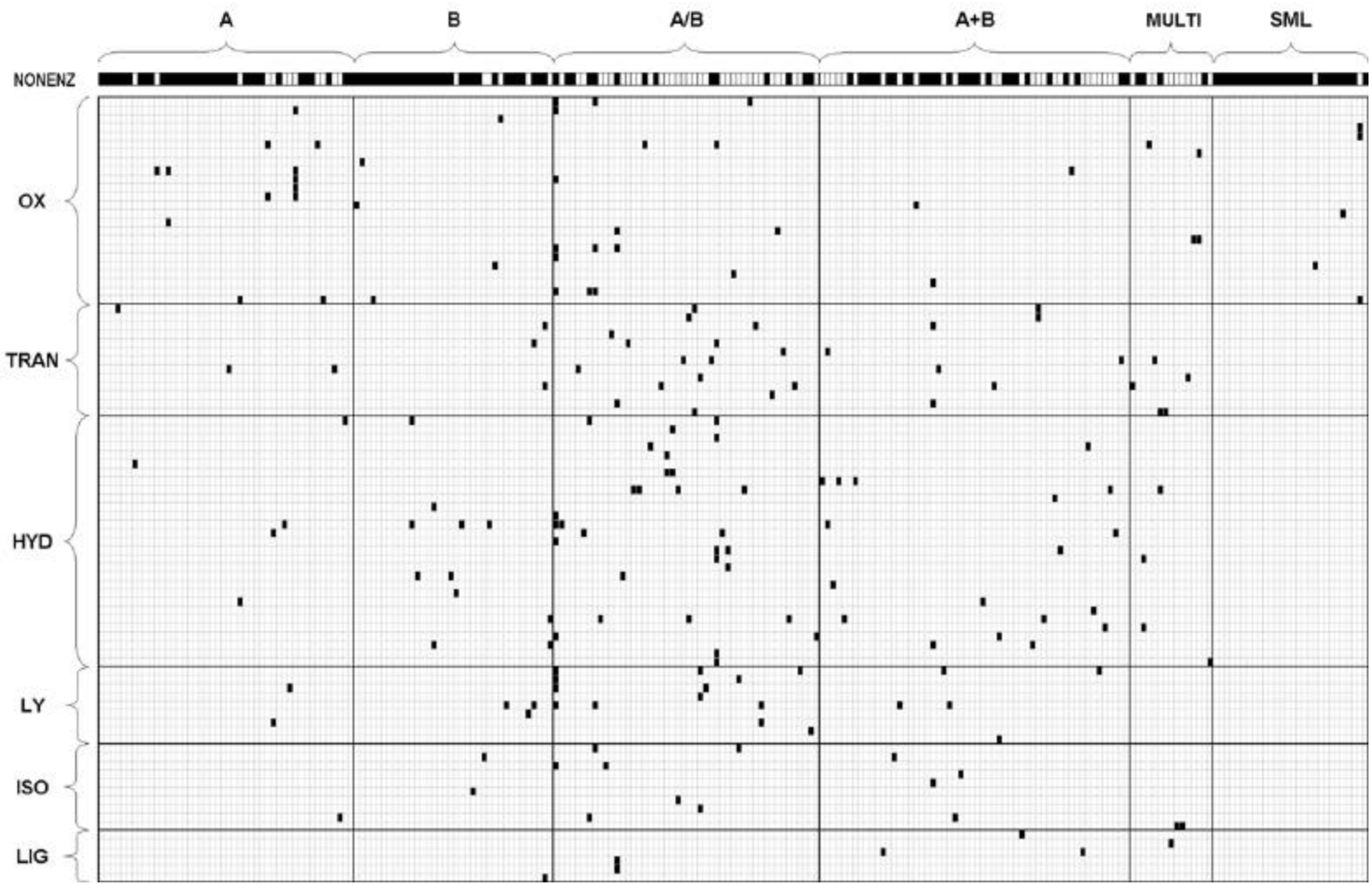


Fold-Function Combinations

~20K (=92x229) Possible,
331 Observed

229 Folds

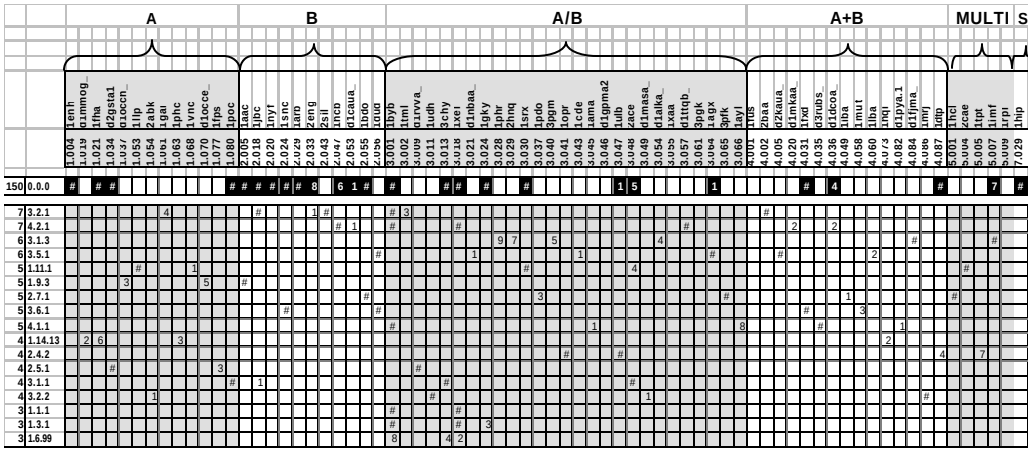
91 Enzymatic Functions
+ Non-Enzyme



The Most Versatile Folds, Versatile Functions

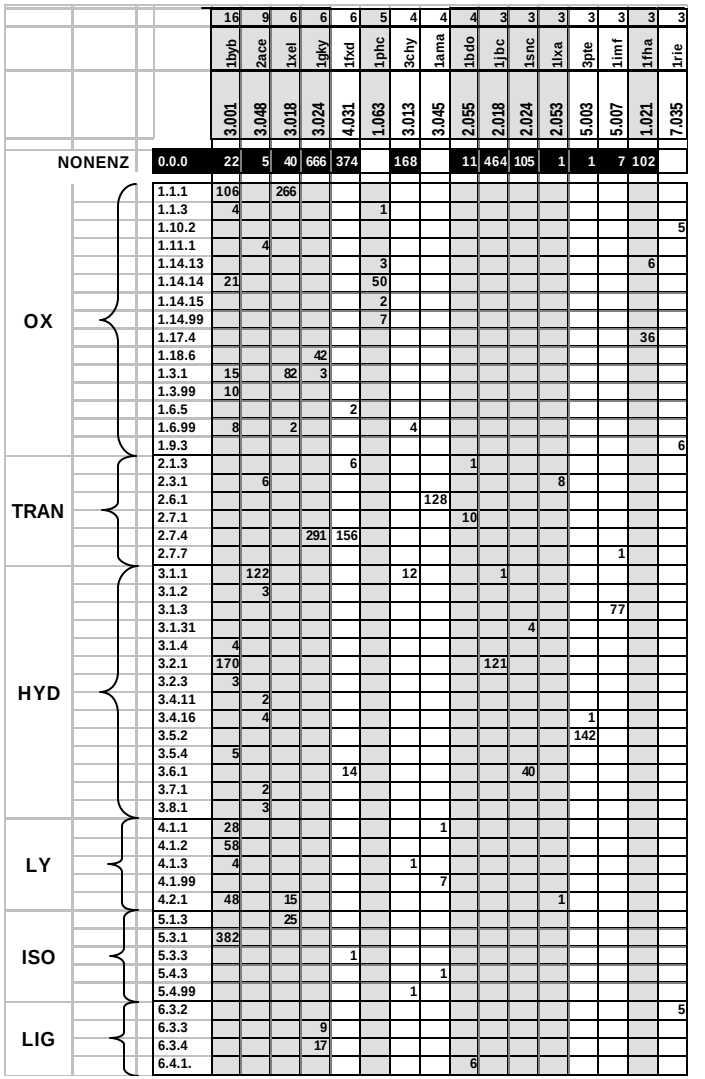
Top-4 Functions:
Glycosidases, carboxy-lyases, phosphoric monoester hydrolases, linear monoester hydrolases (3.2.1, 4.2.1 3.1.3, 3.5.1)

Top-5 Folds:
TIM-barrel (16),
alpha-beta hydrolase fold (9),
Rossmann fold (6), P-loop NTP hydrolase fold (6),
Ferrodoxin fold (6)



Top Multifold Functions ®

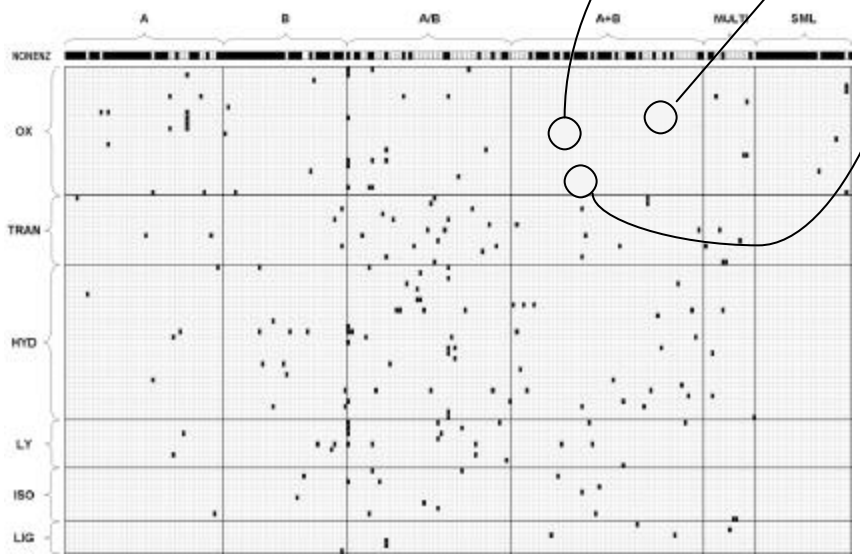
Top Multifunctional Folds ®



Fold-Function Combinations Cross-Tabulation Summary Diagram

	A	B	A/B	A+B	MULTI	SML	sum
NONENZ	34	30	14	28	4	26	136
OX	13	5	17	3	4	5	47
TRAN	3	3	16	8	5		35
HYD	4	11	30	18	4		67
LY	2	3	13	5			23
ISO	1	2	7	4	2		16
LIG		1	2	3	1		7
sum	57	55	99	69	20	31	331

3



SCOP

ENZYME

	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

[Similar analysis in Martin et al. (1998), *Structure* 6: 875]

Compare Classifications and Genomes

Compare 1 Structure-Function Cross-Tab for Different Genomes and Different Functional & Structural Classifications for the Yeast Genome

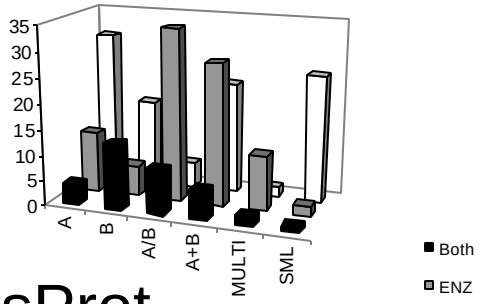
SCOP

ENZYME	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

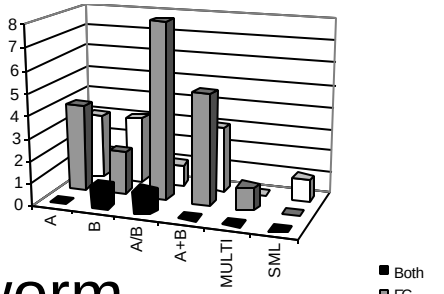
CATH (Thornton)

CATH

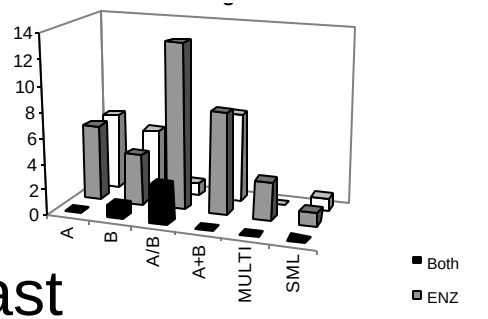
ENZYME	A	B	AB
NONENZ	10	9.0	15
OX	5.1	5.1	10
TRAN		1.3	13
HYD	2.6	1.3	14
LY		2.6	1.3
ISO	1.3	1.3	5.1
LIG			1.3



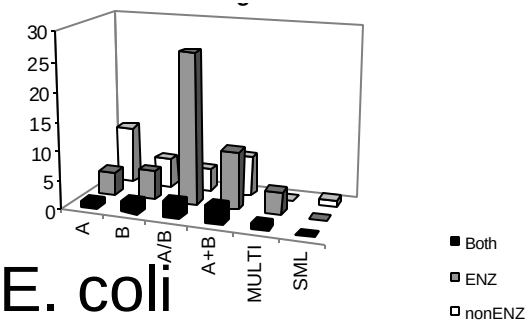
SwissProt



worm



Yeast



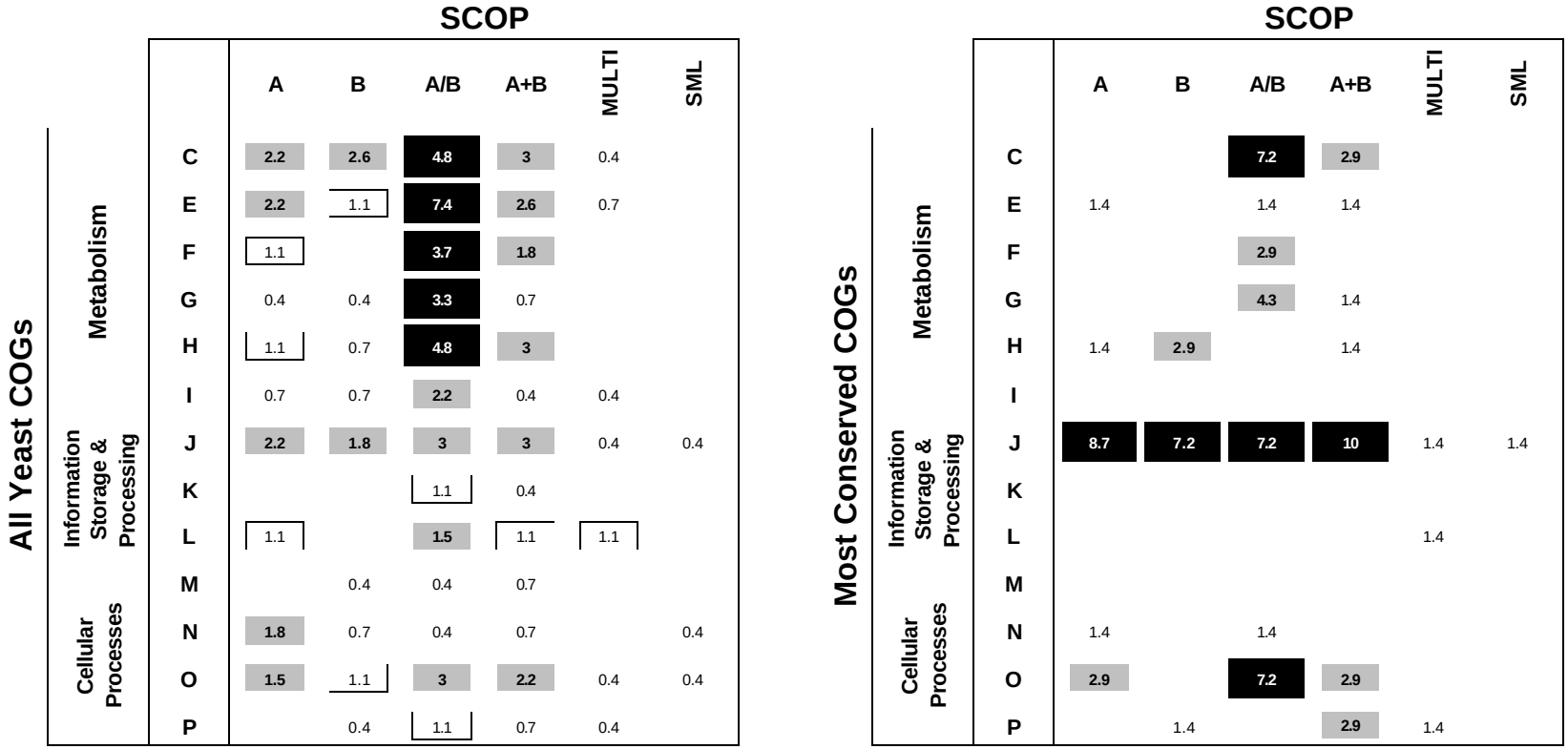
E. coli

MIPS YFC (Mewes)

SCOP

MIPS Functional Cat.		A	B	A/B	A+B	MULTI	SML
metabolism	1	3.5	2.3	10	4.5	1.3	0.8
energy	2	1.1	1.2	5	1.5	0.3	0.2
growth, div., DNA syn.	3	4.9	3.6	4	4.5	1.8	1.2
transcription	4	1.5	1.3	2.2	1.5	0.5	0.8
protein synthesis	5	1	0.9	0.7	1.3	0.3	0.2
protein targetting	6	1.2	1.7	2	1.6	0.5	0.3
transport facilitation	7	0.9	0.5	0.7	0.6	0.4	
intracellular transport	8	1.8	2.1	1.6	0.6	1	
cellular biogenesis	9	0.9	0.7	1.2	0.3	0.3	0.1
signal transduction	10	1	1	1.1	0.3	0.7	0.3
cell rescue, defense...	11	1.5	1	2.6	1.9	0.7	0.5
ionic homeostatis	13	0.5	0.3	0.4	0.4	0.2	

COGs vs SCOP: Different Structure Function Relationships for Most Conserved Proteins



(Scop, Murzin, Ailey, Brenner, Hubbard, Chothia; COGs, Tatusov, Koonin, Lipman)

Comparing Genomes in terms of Protein Structure: Surveys of a **Finite** Parts List

1 Library of Known Folds

Importance of Statistics. Scop auto-alignments. P-values from EVD, same as sequences.

2 Census of Known Folds

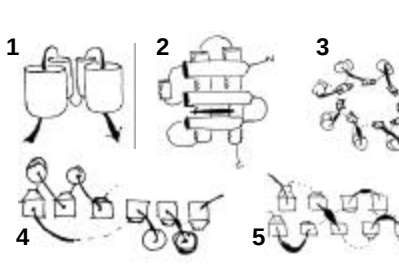
Which folds in which organisms: E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression. Repeated βαβ. Biases. Extent of MG fold assignment (65%)

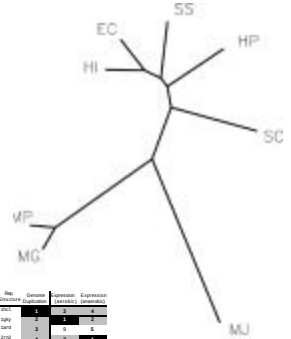
3 Prediction of Unknown Folds

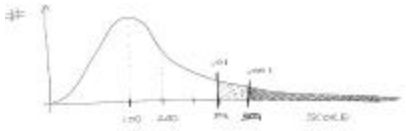
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Biases: Can extrapolate from known structures to genomes?

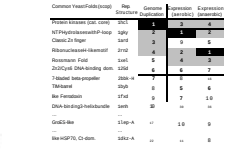
4 Fold-Function Relationships

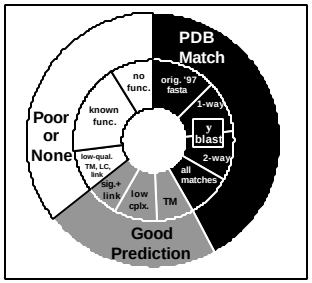
How many folds per function? Func. per fold? 331 of ~20K combinations. TIM most versatile scaffold.

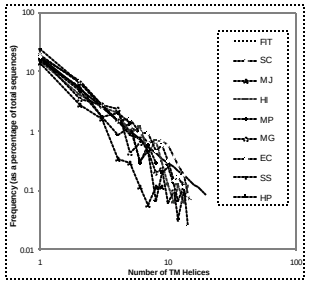












SCOP

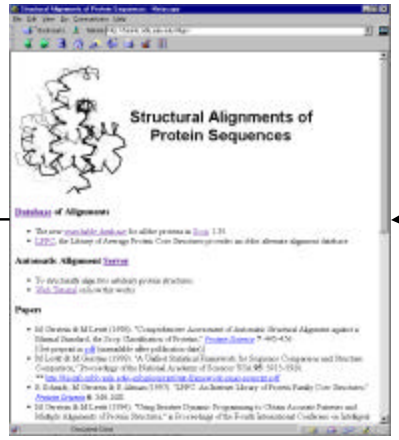
ENZYME	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

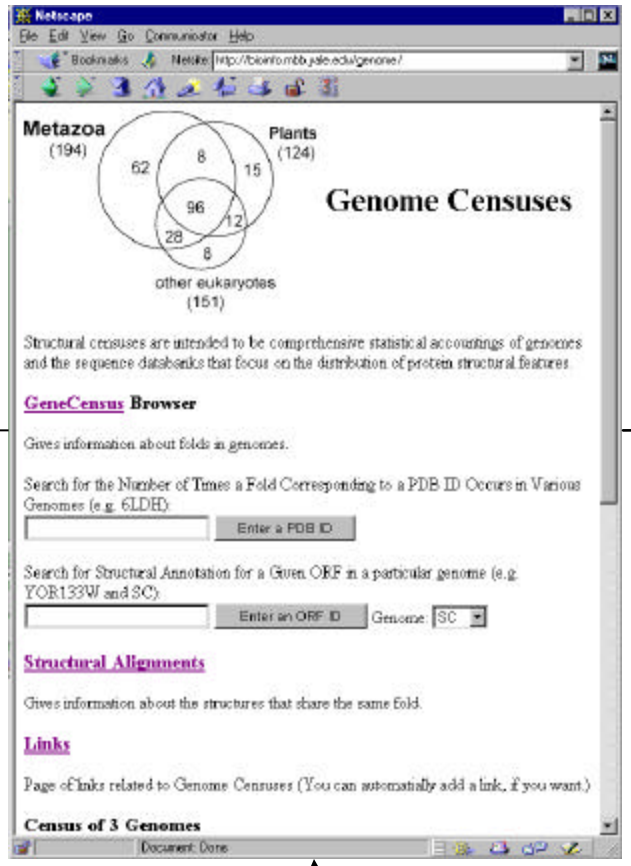
http://bioinfo.mbb.yale.edu/genome

GeneCensus

Alignment Database

Alignment Server

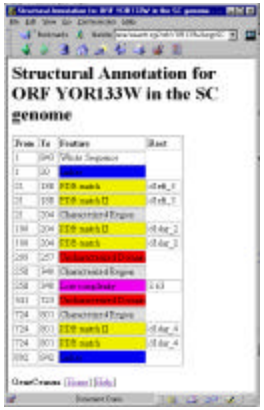




Detailed Tables

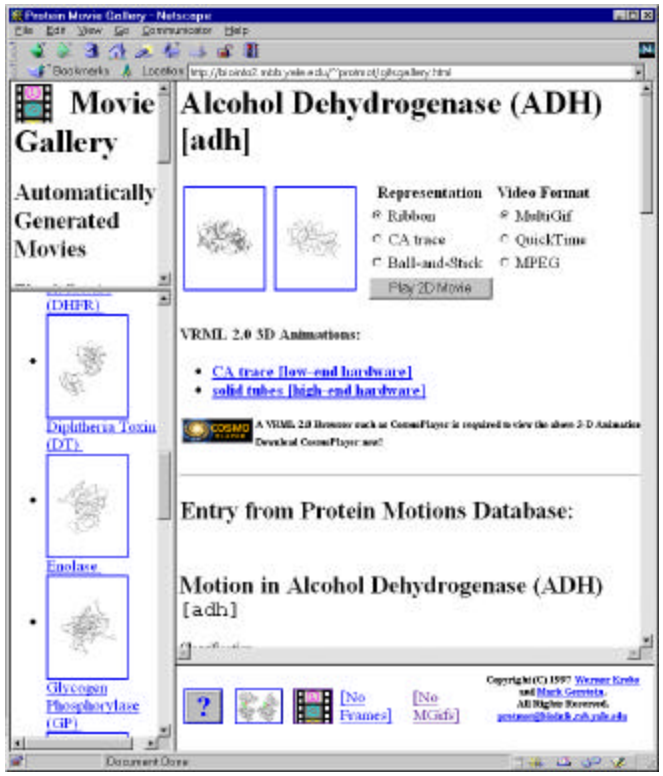
PDB Query

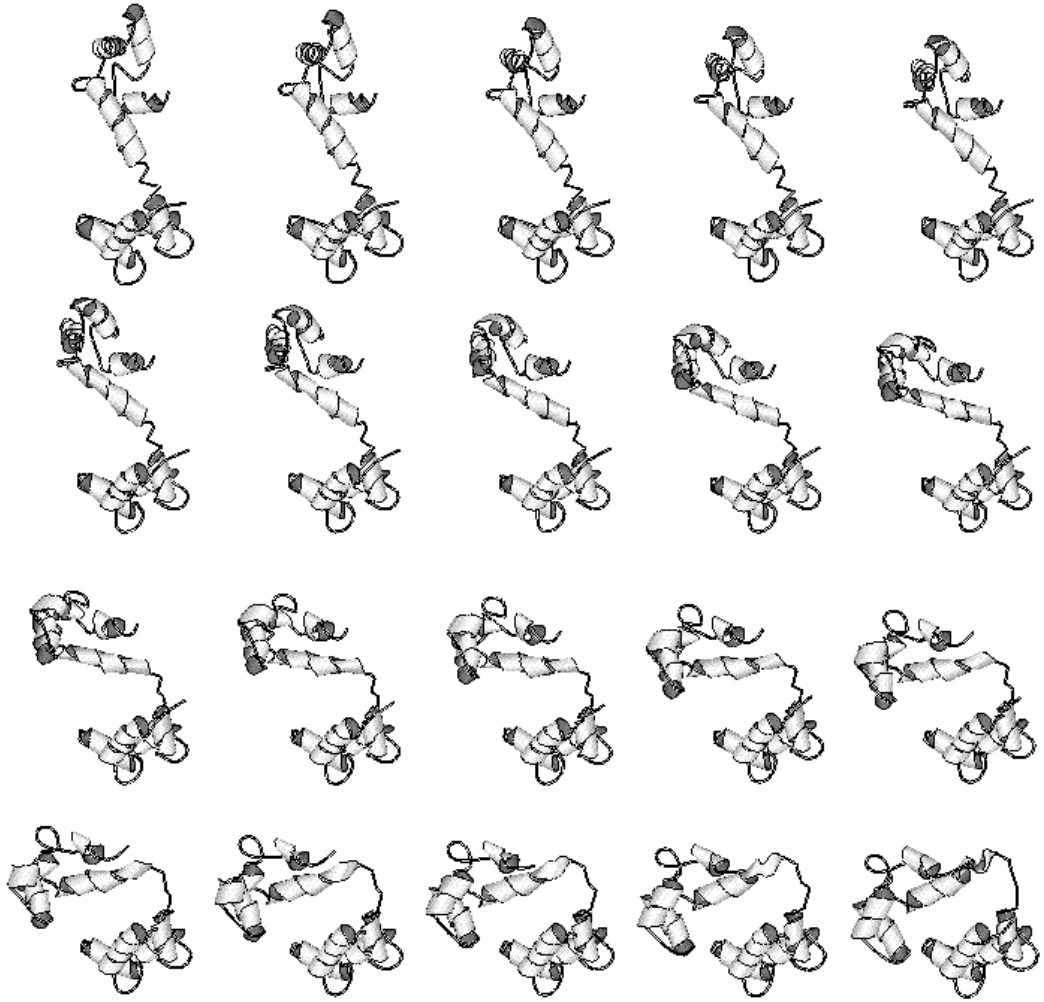
ORF Query



http://bioinfo.mbb.yale.edu/genome

“Morph” Movies of Protein Motions

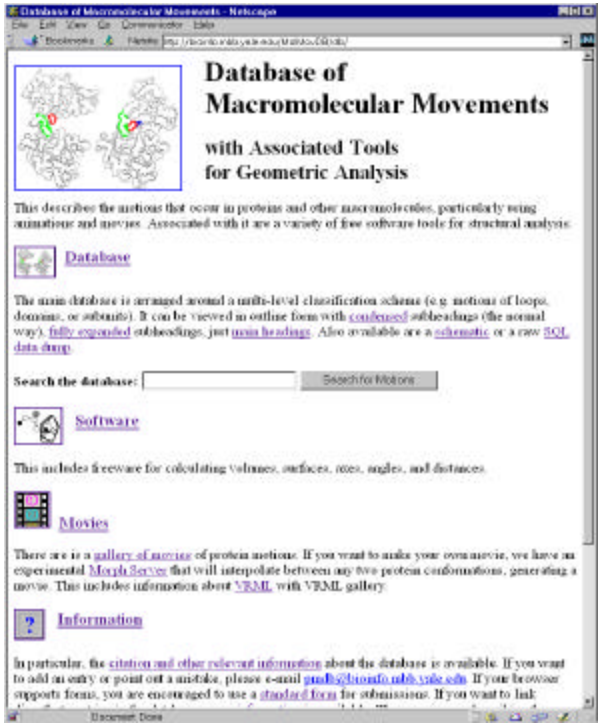


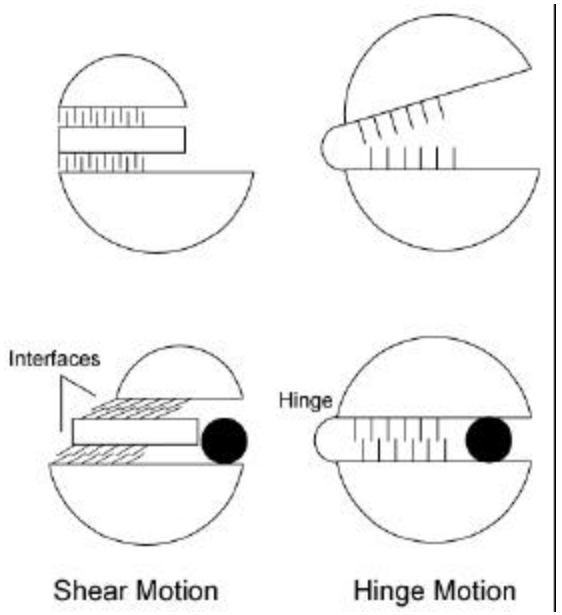


http://bioinfo.mbb.yale.edu/MolMovDB

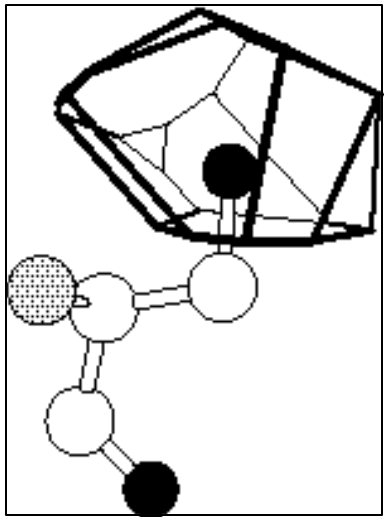
Server Produces Semi-realistic Minimized Interpolation (as MPEG, VRML, &c) between Any 2 Aligned Conformations, Analyzes the Motion

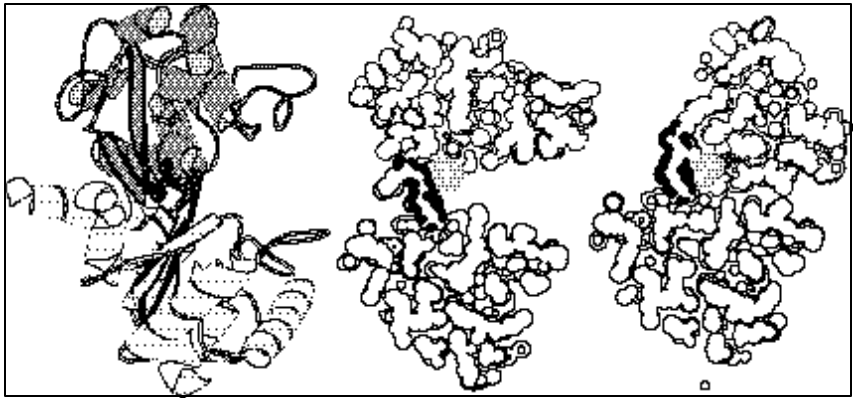
Server Helps Classify Motion in Database Based on Packing





Number Known Forms	Size of Motion	Mechanism of Motion	Examples	#
Motion				
2 forms	Fragment	**Hinge**	TIM, LDH, TGL	14
		Shear	Insulin	3
		Unclassifiable	MS2 Coat	3
	Domain	**Hinge**	LF, ADK, CM	16
		Shear	CS, TrpR, AAT	8
		Refold	Serpin, RT	3
		Special	Ig elbow	1
		Unclassifiable	TBP, EF-tu	3
	Subunit	Allosteric	PFK, Hb, GP	4
		Non-allosteric	Ig VL-VH	2
		Unclassifiable		
1 form	Fragment	**Hinge**		
		Shear		
		Unclassifiable	bR	1
	Domain	Refold		
		Hinge	LF~TF,SBP	10
		Shear	HK~PGK,HSP	4
		Special		
		Unclassifiable	Myosin	4
	Subunit	Allosteric		
		Non-allosteric		
		Unclassifiable	PCNA, GroEL	3





Comparing Genomes in terms of Protein Structure: Surveys of a **Finite** Parts List

1 Library of Known Folds

Importance of Statistics. Scop auto-alignments. P-values from EVD, same as sequences.

2 Census of Known Folds

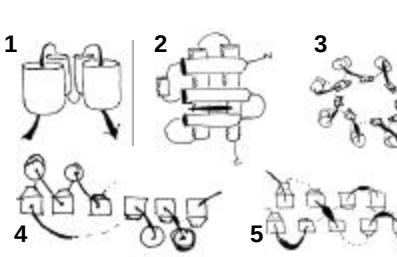
Which folds in which organisms: E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression. Repeated βαβ. Biases. Extent of MG fold assignment (65%)

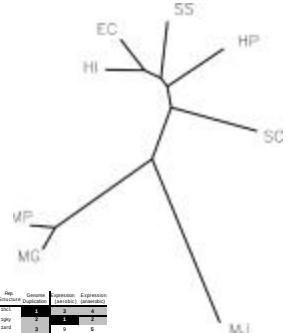
3 Prediction of Unknown Folds

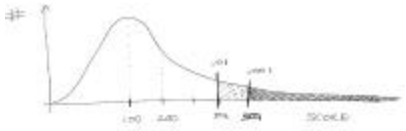
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Biases: Can extrapolate from known structures to genomes?

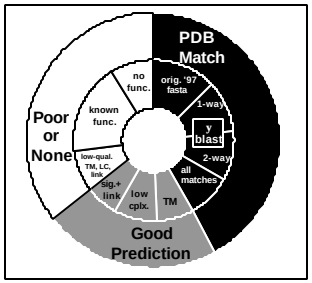
4 Fold-Function Relationships

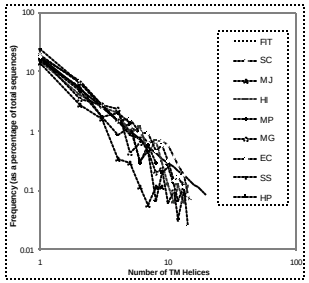
How many folds per function? Func. per fold? 331 of ~20K combinations. TIM most versatile scaffold.











http://bioinfo.mbb.yale.edu/genome

Acknowledgements: **M Levitt, H Hegyi, J Lin, C Chothia, S Teichmann, scop (Murzin et al.), W Krebs, C Wilson**

SCOP

ENZYME	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		