Comparing Genomes in terms of Protein Structure:

Surveys of a Finite Parts List

Mark Gerstein

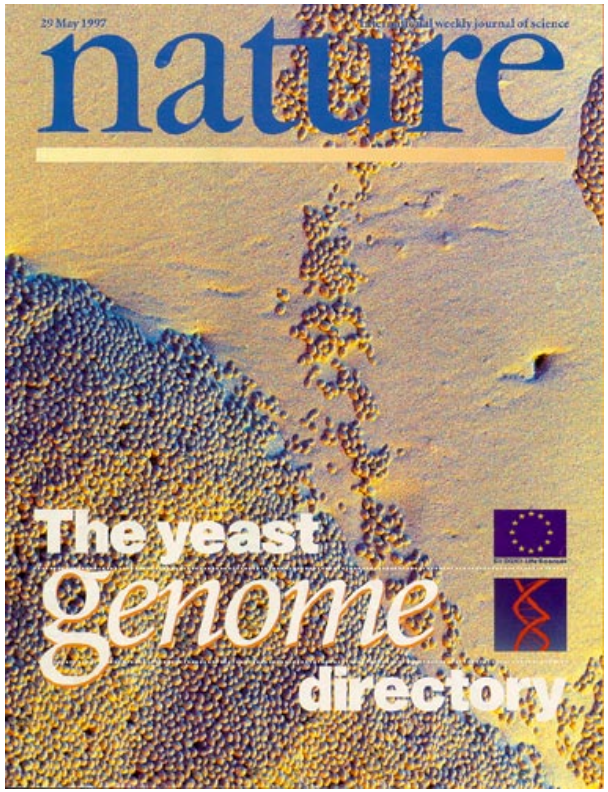
Genomes highlight the **Finiteness** of Biology

1995

Bacteria
1.6 Mb, ~1600 genes

[Fleischmann *et al.* (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd." *Science* **269**: 496-512.]

1997



Eukaryote
13 Mb, ~6000 genes

1998.....

Microbial Genomes
>15 completed,
~40 underway

The Worm:
75% of 100 Mb done, with ~13 K genes so far)

The Human:
3 Gb & 100 K genes, 2003?

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

1 Library of Known Folds

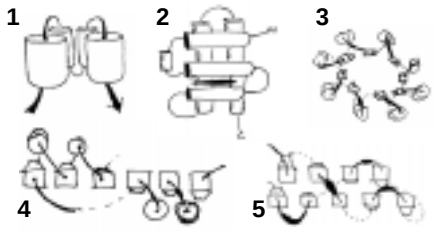
Importance of statistics. Scop auto-alignments. Significance follows EVD stats, same as sequences.

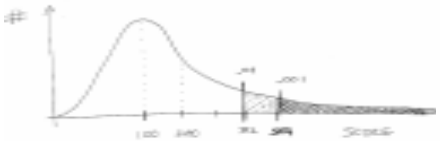
2 A Census of Known Folds

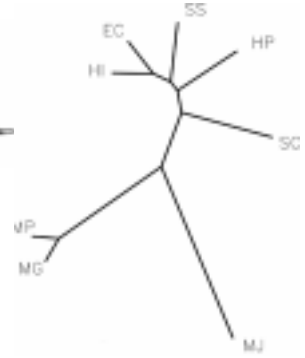
Which folds in which Organisms: Plants v. People, E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression, repeated βαβ supersecondary struc.

3 Prediction of Unknown Folds

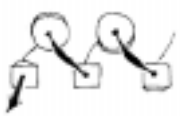
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Can extrapolate from known structures to genomes?

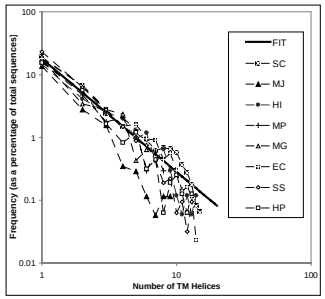


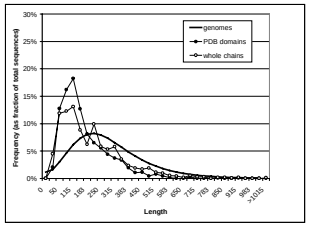




Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	1z5d	6	6	7
7-bladed beta-propeller	2bbk-N	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





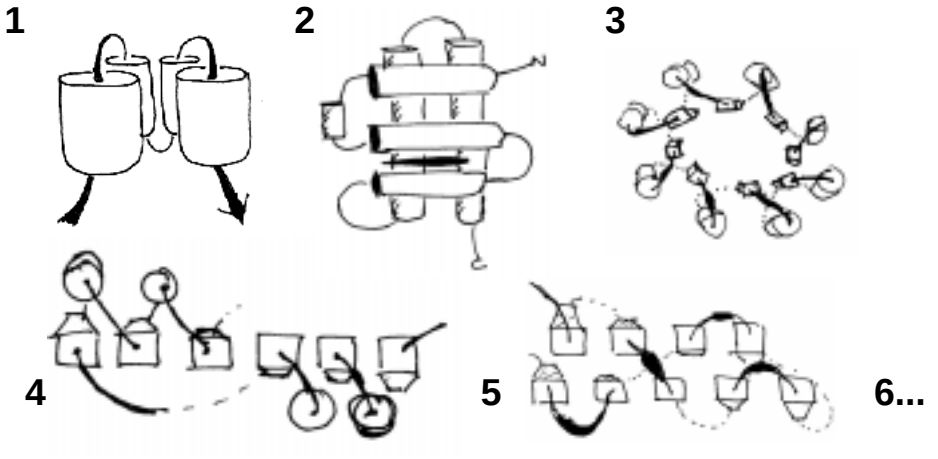


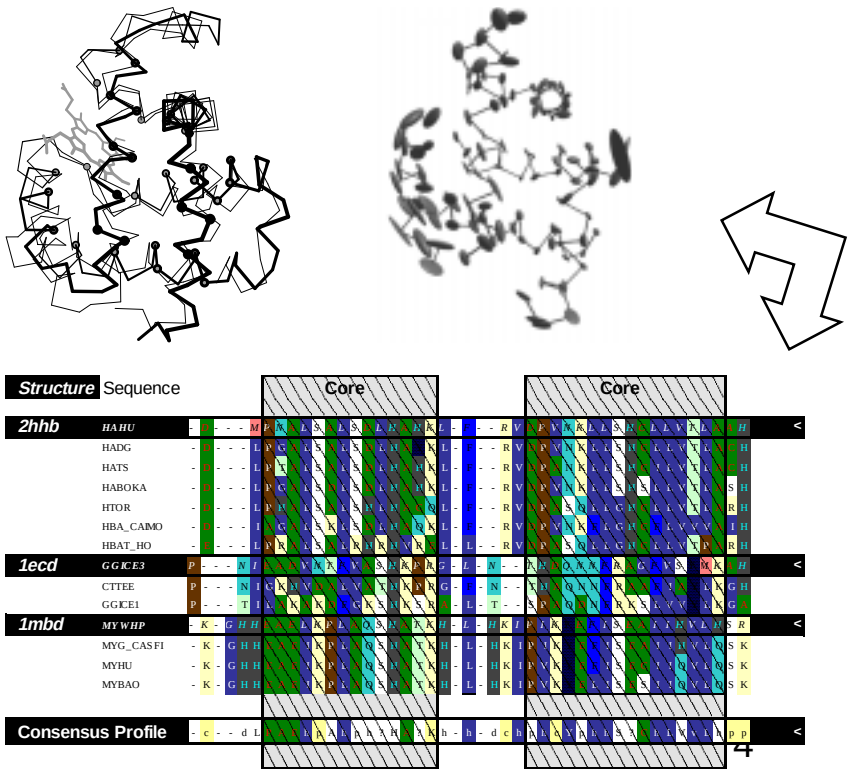
http://bioinfo.mbb.yale.edu/genome

Acknowledgements: M Levitt, scop (Murzin, Brenner, Ailey, Hubbard, Chothia)

Fold Library

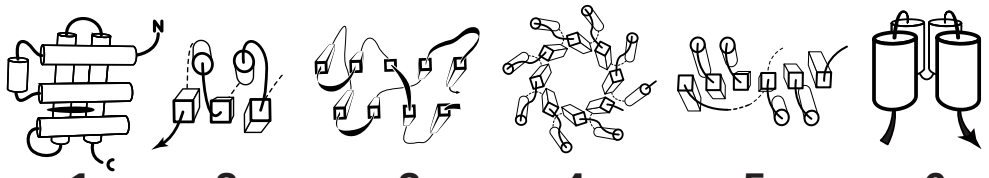
- Primary way to interpret Genome Sequences in terms of Structure
- Very Limited Number of Folds (~1K-10K, Chothia)
- Elements: Domain definitions; Aligned structures, collecting together Non-homologous Sequences; Core region annotation
- Many approaches to building Library
  - ◊ Automatic: FSSP-HSSP (Sander), Entrez-MMDB (Bryant)
  - ◊ Semi-automatic: CATH (Thornton), HOMALDB (Sali)
  - ◊ **Manual (scop, Murzin)**
  - ◊ Start with Sequences: Pfam (Durbin, Eddy), COGs (Koonin, Lipman), Blocks (Henikoff), ProSite (Bairoch)





Scale Fold Library vs. Other Fundamental Data structures

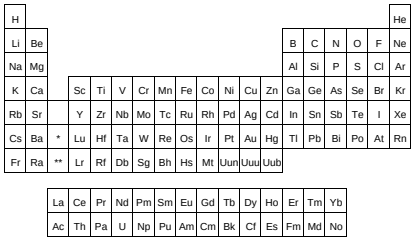
Statistical, rather than mathematical relationships and conclusions, Parts List Database

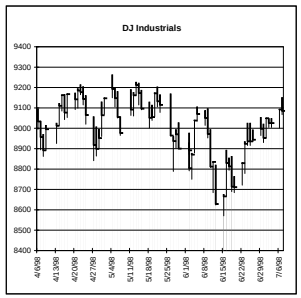


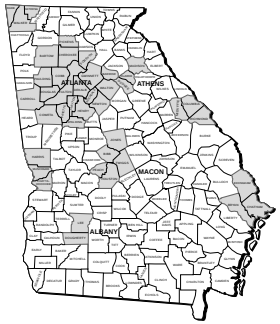
1 2 3 4 5 6 ...

Folds in Molecular Biology **1000-10000**

const.	mant.		exp.	unit
e	1.60	e	8	C
F	9.65	e	4	C/mol
ε_0	8.85	e	-12	F/m
μ_0	1.26	e	-6	H/m
h	6.63	e	-34	J •s
k	1.38	e	-23	J/K
m_e	9.11	e	-31	kg
m_p	1.67	e	-27	kg
m_n	1.68	e	-27	kg
a_0	5.29	e	-11	m
λ_C	2.43	e	-12	m
c	3.00	e	-19	m/s
G	6.67	e	-11	$m^3/kg•s^2$
N_A	6.02	e	23	mol^{-1}







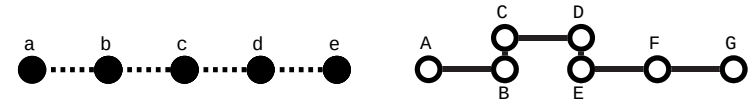
10 **100** **1000 -10000** **>1000000**

Physics Chemistry Finance Politics

(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.

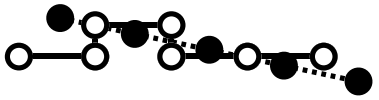
Automatic Alignments of Scop, Focussing on Statistics of Relationships

- Our Approach
  - ◊ Iterative Dynamic Programming, like repeated sequence alignment
  - ◊ Derived from Program of G Cohen (Align)
  - ◊ **Score** = $M_{str}(i,j) = \sum 100 / (5 + d^2)$
- Numerous other approaches to struc. alignment: SAP (Taylor), VAST (Bryant), Artymiuk, Sali, Sippl, Sander, Cohen, STAMP (Barton)



Initial Equivalences

```
- - a b c d e
    | | | | |
A B C D E F G
```



```
a - b - c d e    Score    57
|   |   | | |    Nbrk      2
A B C D E F G    RMS    1.96
```

	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    Score    91
| |     | | |    Nbrk      1
A B C D E F G    RMS    0.65
```

	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    Score   100
| |     | | |    Nbrk      1
A B C D E F G    RMS    0.23
```

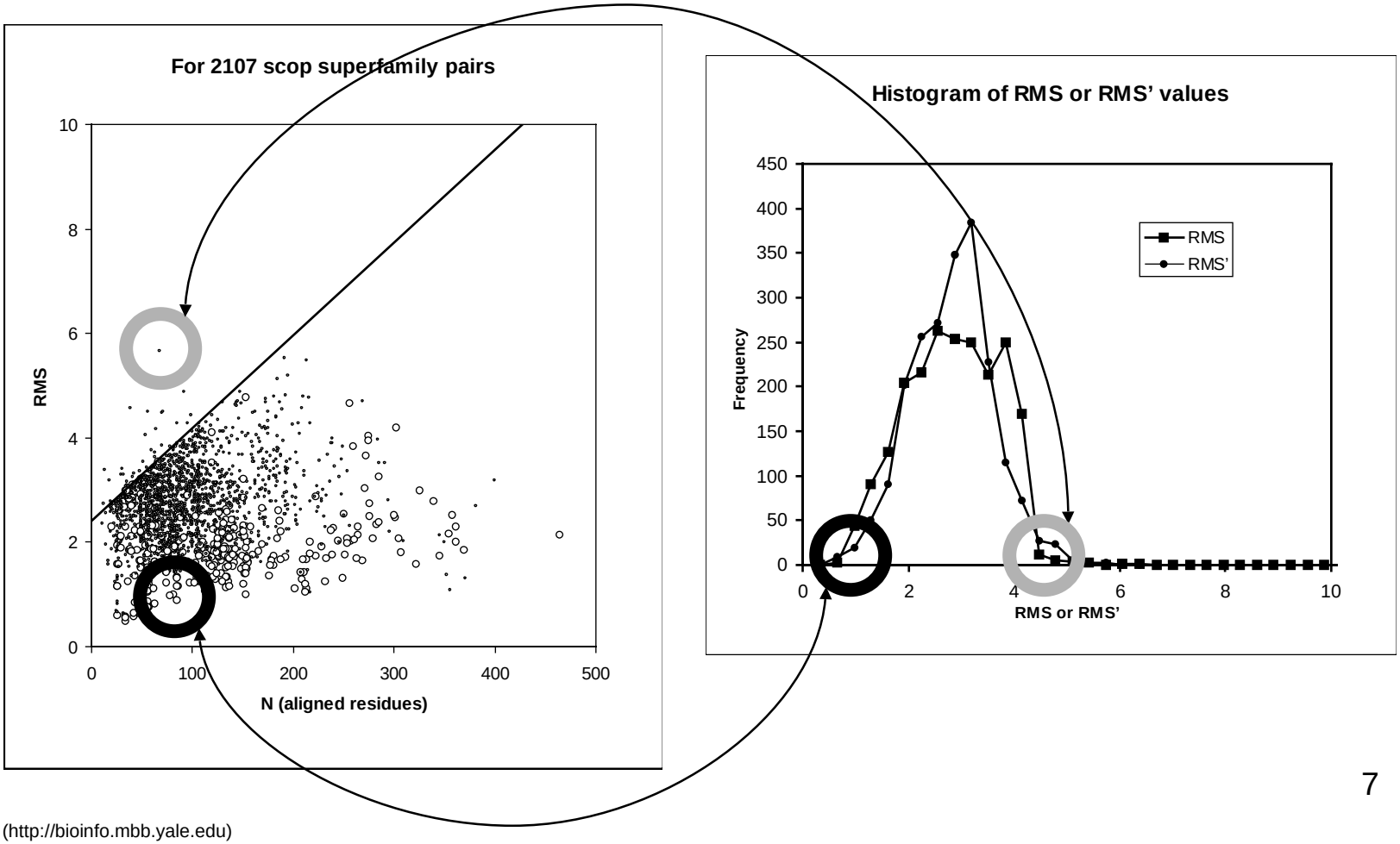
	A	B	C	D	E	F	G
a	**20**	4	3	1	1	0	0
b	4	**20**	12	4	4	1	0
c	1	4	4	11	**20**	4	1
d	0	1	1	4	4	**20**	4
e	0	0	0	1	1	4	**20**

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

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

Statistics on Range of Similarities

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

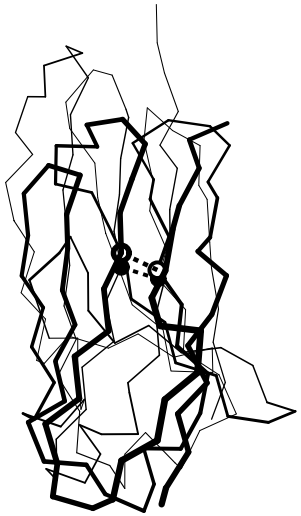


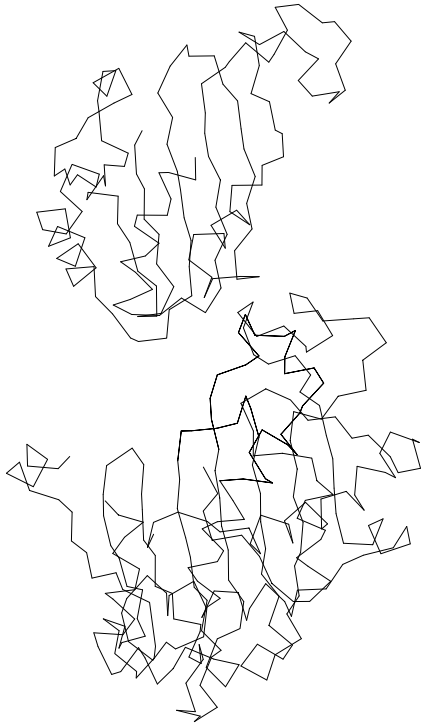
Some Similarities are Readily Apparent others are more Subtle

Easy: Globins

Tricky: Ig C, Ig V,

Very Subtle: G3P-dehydrogenase, C-term. domain



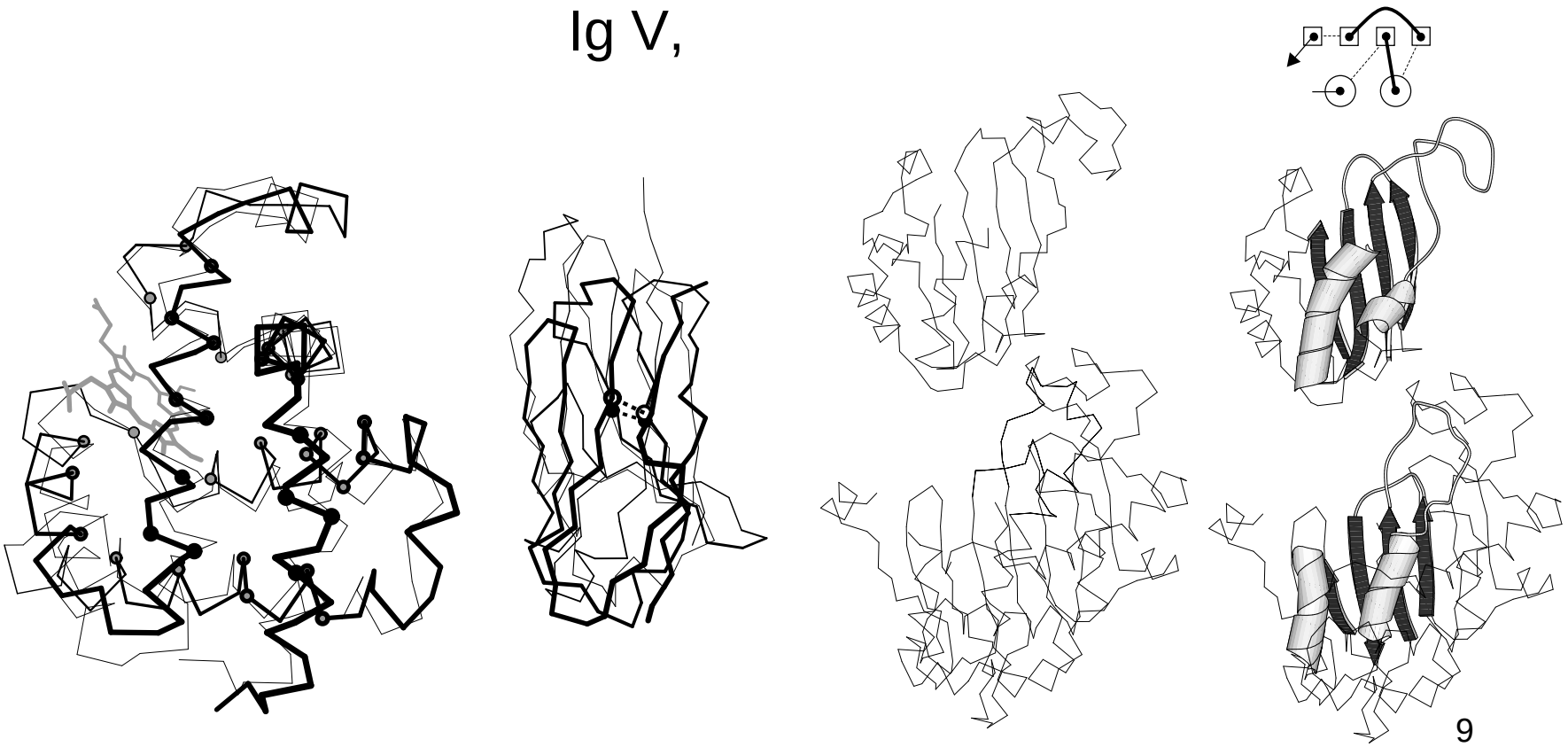


Some Similarities are Readily Apparent others are more Subtle

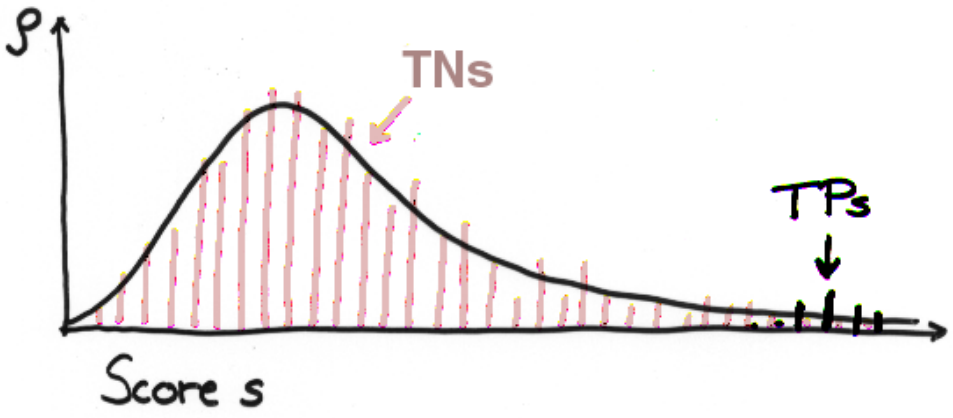
Easy: Globins

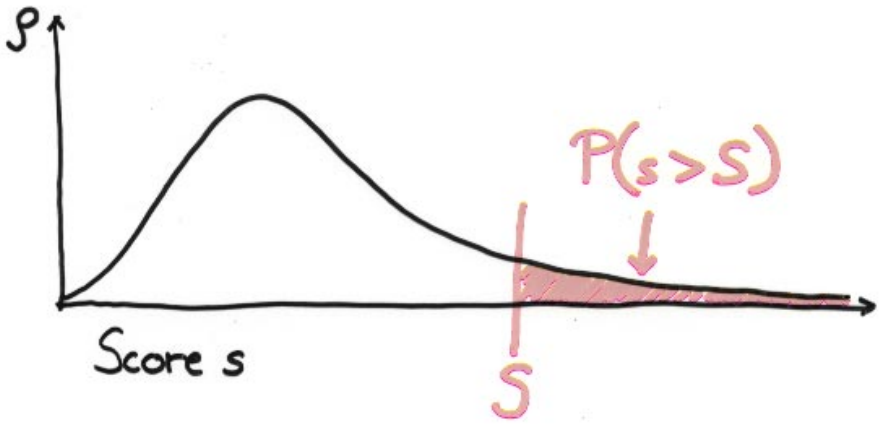
Tricky: Ig C, Ig V,

Very Subtle: G3P-dehydro-genase, C-term. domain



Strategy: same formalism for sequence & structure

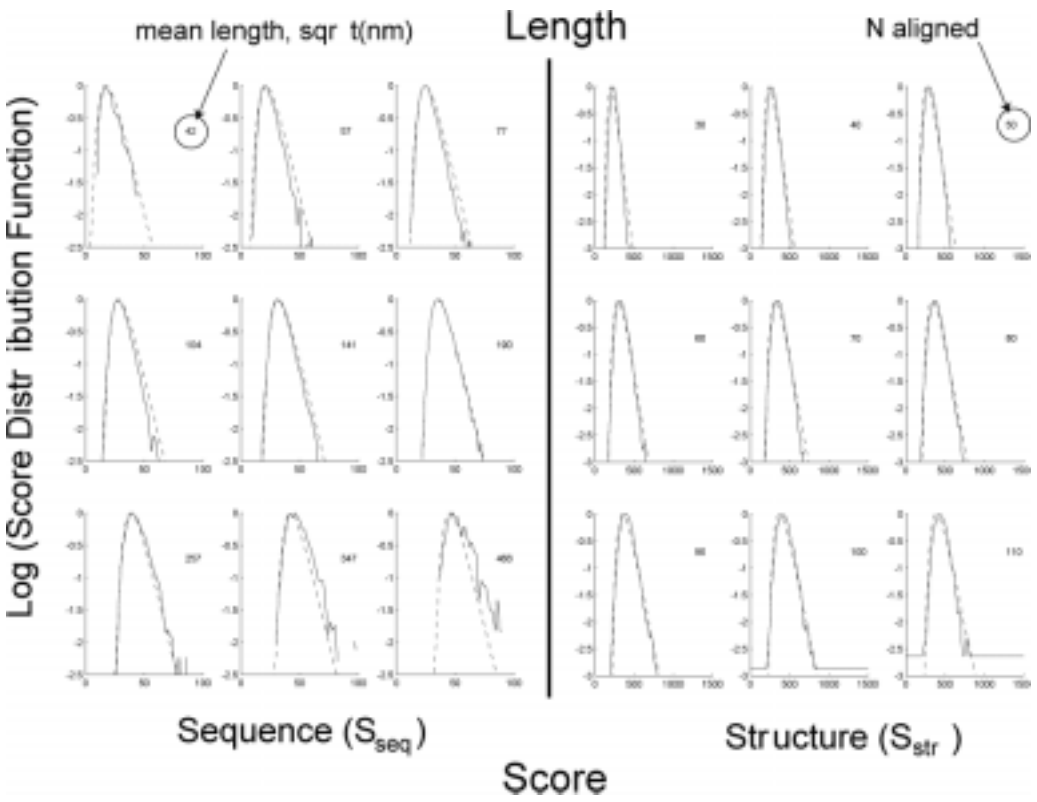


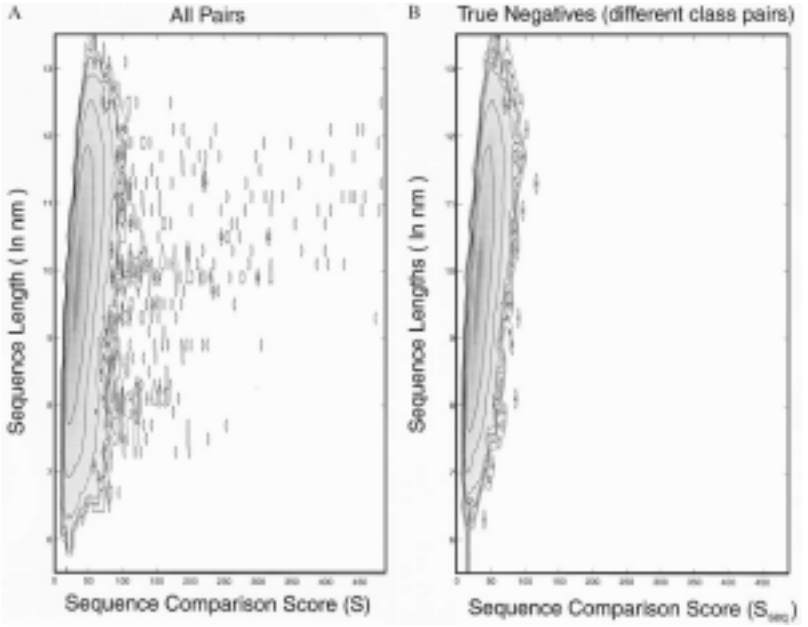


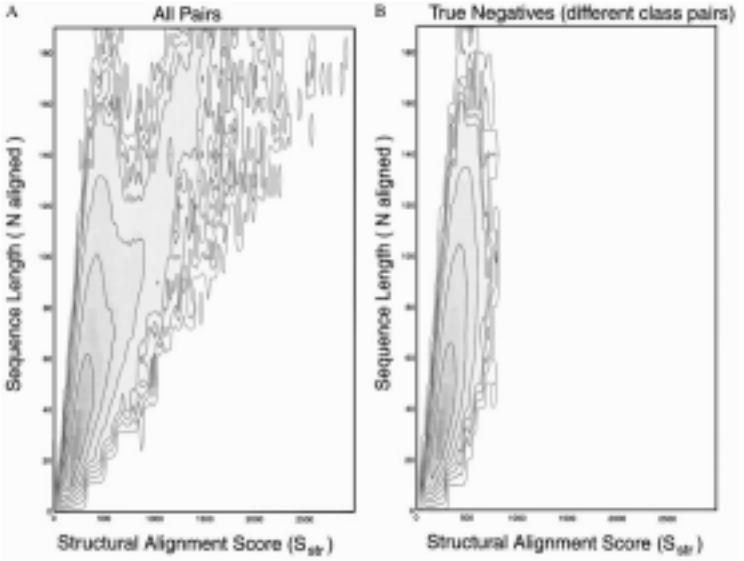
[e.g. P(score s>392) = 1% chance]

- Fit to Observed Distribution
  - ◊ All-vs-All comparison
  - ◊ Graph Distribution of Scores in 2D (N dependence)
  - ◊ 1K x 1K families -> ~1M scores; ~2K TP embedded in this
  - ◊ Fit a function ρ(S) to distribution of true negatives (as determined by scop)
- A P-value for Significance
  - ◊ Integration of ρ (ie the CDF) gives P(s>S), the chance of getting a score better than threshold S randomly
  - ◊ Extrapolated Percentile Rank: How does a Score Rank Relative to all Other Scores?
  - ◊ For sequences, originally used in Blast (Karlin-Altschul). Then in FASTA, &c.

Observed Distributions







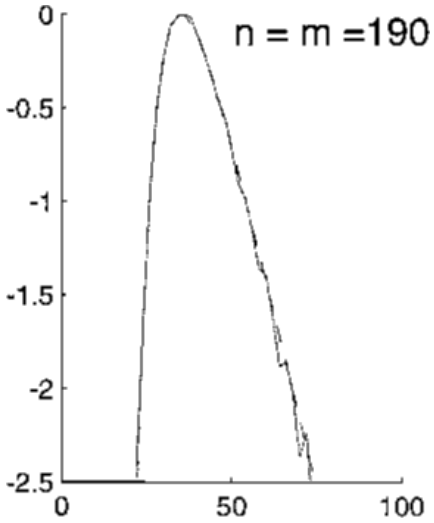
EVD Fits

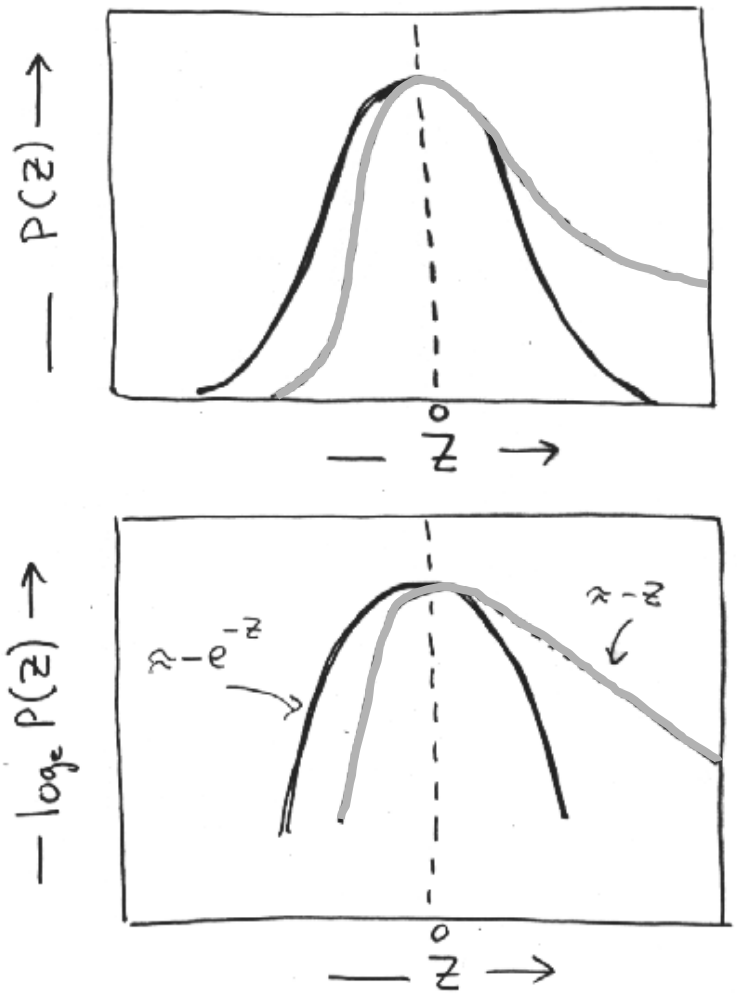
$$\rho(z) = \exp\left(-z - e^{-z}\right)$$

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

- 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:

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

Same Results for Sequence & Structure

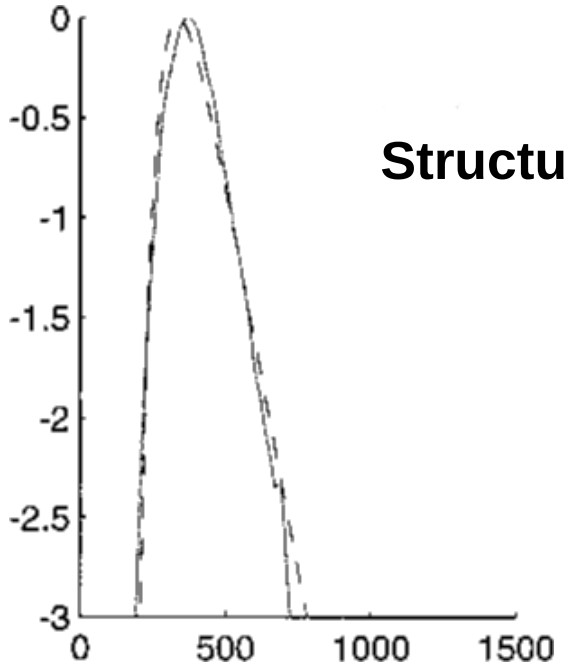
3 Free Parm. fit to EVD involving: ***a, b,*** σ.
These are the only difference betw. sequence and structure.

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

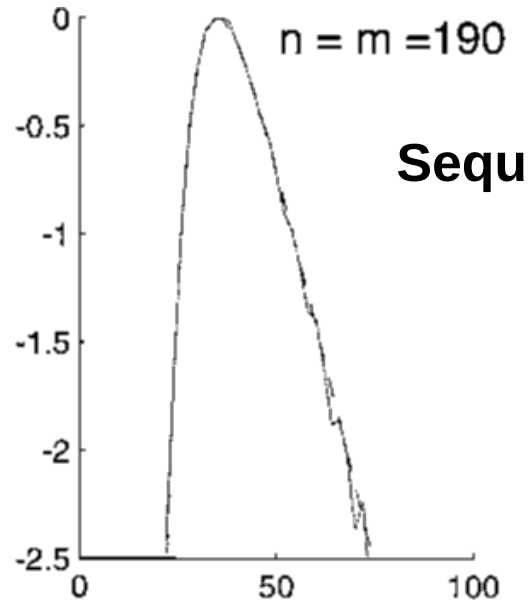
$$S = \sum_{i,j} M(i,j) - G$$

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

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.)



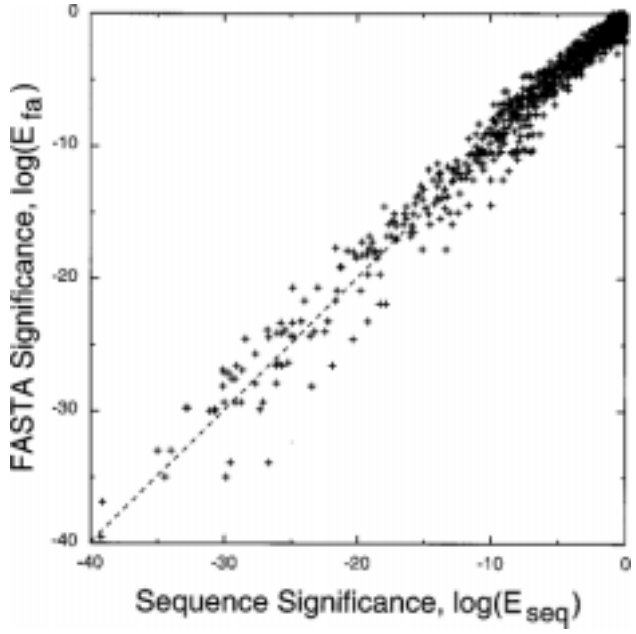
Structure



Sequence

Use Sequence Scores to Validate

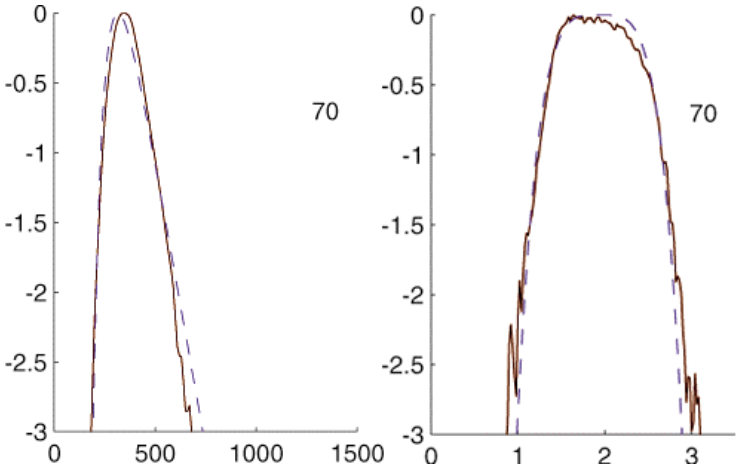
- Sequence P-value perfectly tracks FASTA e-value
  - ◊ Validates approach
  - ◊ Added Benefit: allows computation of an e-value without doing a db run
- Significance computation can be applied to **any** exisiting sequence or structure alignment



- 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

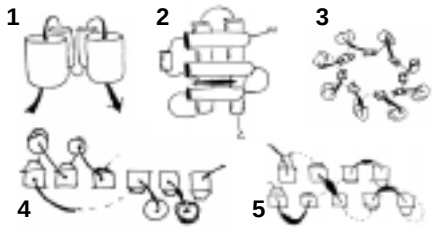
Importance of statistics. Scop auto-alignments. Significance follows EVD stats, same as sequences.

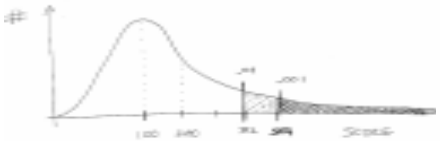
2 A Census of Known Folds

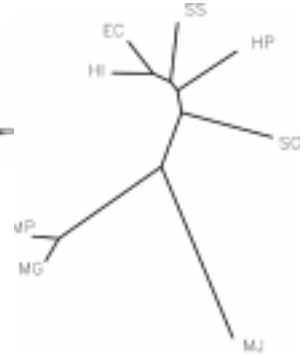
Which folds in which Organisms: Plants v. People, E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression, repeated βαβ supersecondary struc.

3 Prediction of Unknown Folds

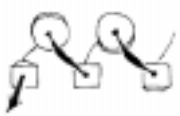
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Can extrapolate from known structures to genomes?

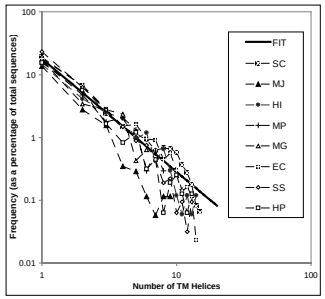


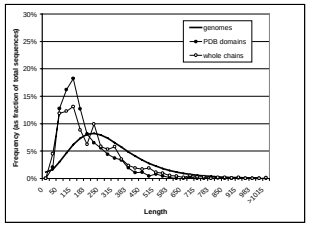




Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	1z5d	6	6	7
7-bladed beta-propeller	2bbk-N	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





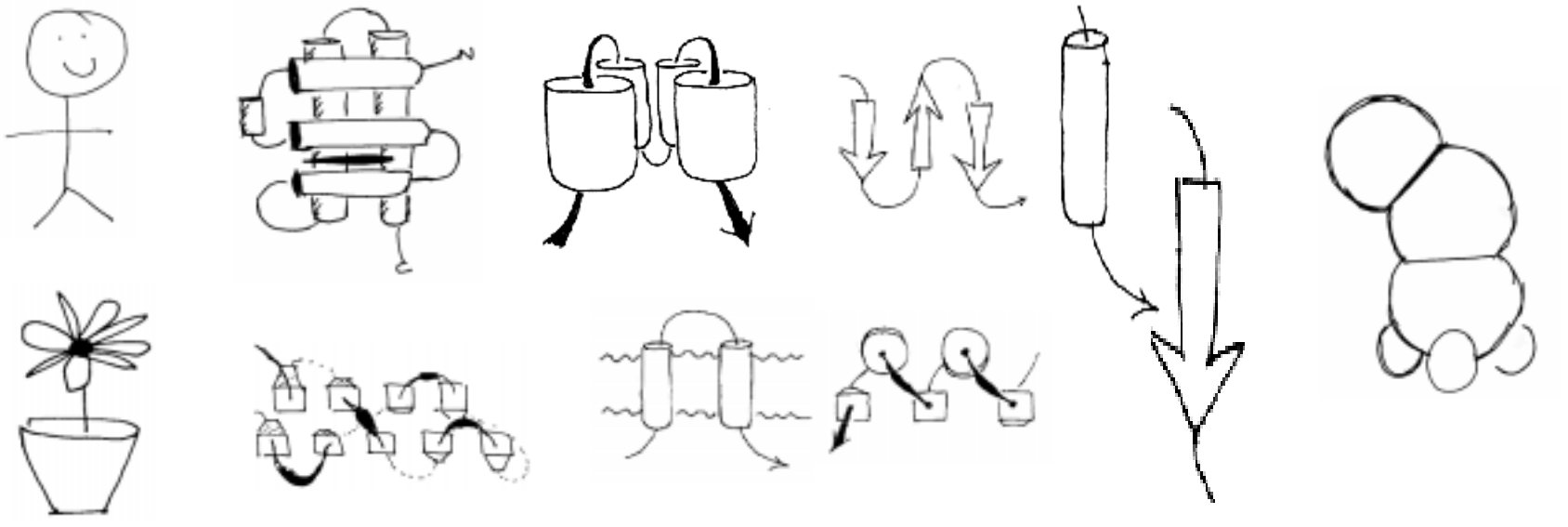


http://bioinfo.mbb.yale.edu/genome

Acknowledgements: M Levitt, scop (Murzin, Brenner, Ailey, Hubbard, Chothia)

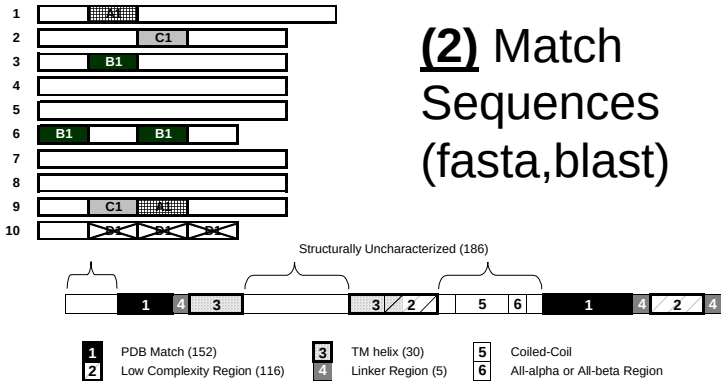
At What Structural Resolution Are Organisms Different?

person plant	protein fold (Ig)	super-secondary structure (ββ,TM–TM, αβαβ,ααα)	helix strand	individual atom (C,H,O...)



1m 100Å 10Å 1Å

Cross-Reference: Folds→Sequences → Organisms

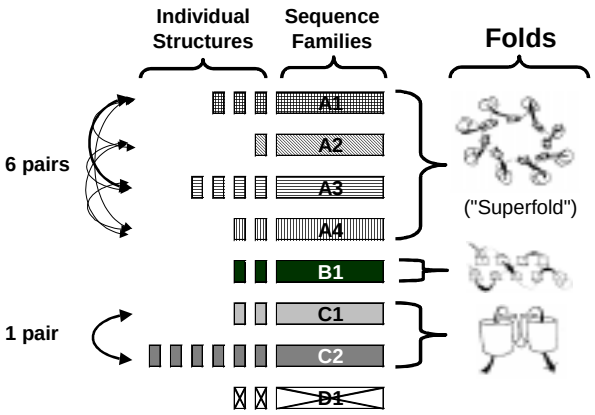


(2) Match Sequences (fasta,blast)

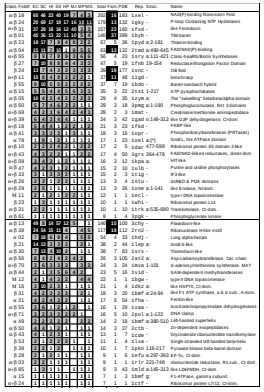
3 + 5

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.

(3) Organize Sequences by Genome or Taxon



(1) Structures in Folds (scop)



(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
α/β	24	20	69	17	19	17	16	10	11	179	13	132	1gky -	P-loop
α+β	31	37	28	18	16	12	40	3	3	157	23	160	1fxd -	like Fe
α/β	01	45	36	13	22	11	10	5	4	146	37	399	1byb -	TIM-ba
α/β	23	18	17	7	9	4	8	2	2	67	5	36	1pyd a:2-181	Thiami
α/β	04	15	11	7	10	1	9	5	5	63	13	132	2tmd a:490-645	FAD/N
α+β	55	8	9	7	8	9	3	6	6	56	4	23	1sry a:111-421	Class-I
β	27	7	10	8	8	4	4	3	3	47	5	19	1fnb 19-154	Reduct
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

Shared Folds in OWL and Initial Genomes

Venn Diagrams

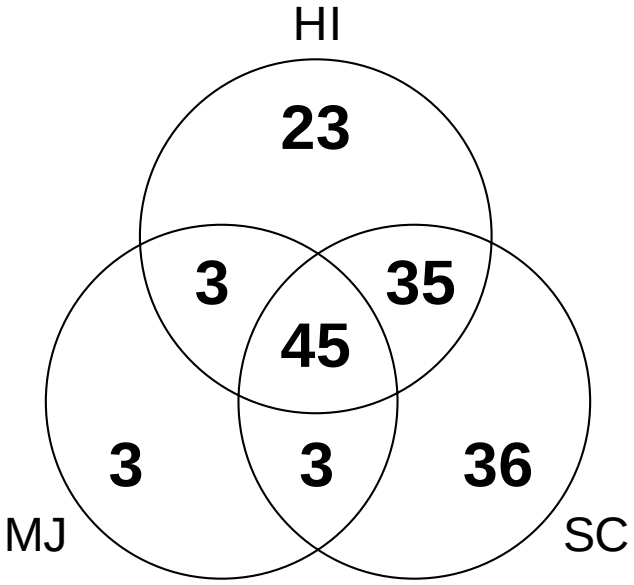
~300 folds (282 folds in scop 1.32 [‘96])

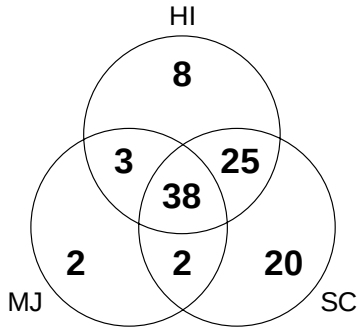
~120K sequences in OWL 27.1

7 groups of organisms

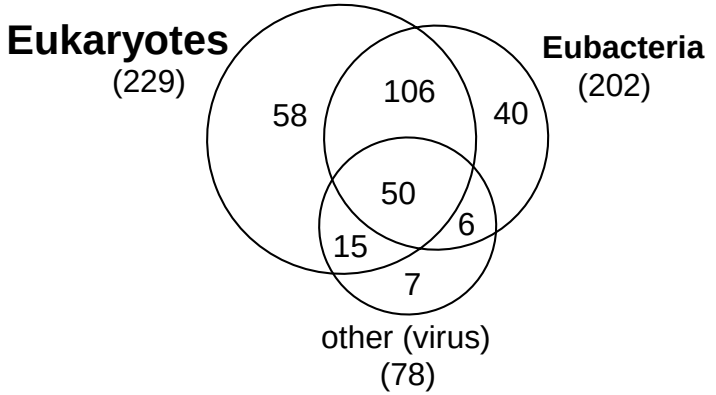
3 of the first genomes

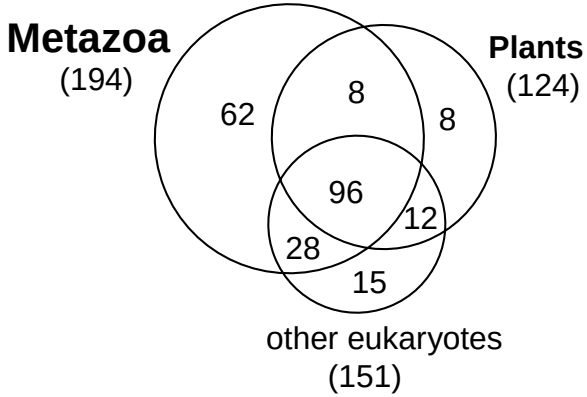
HI (bacteria), MJ (archeon), SC (eukaryote)

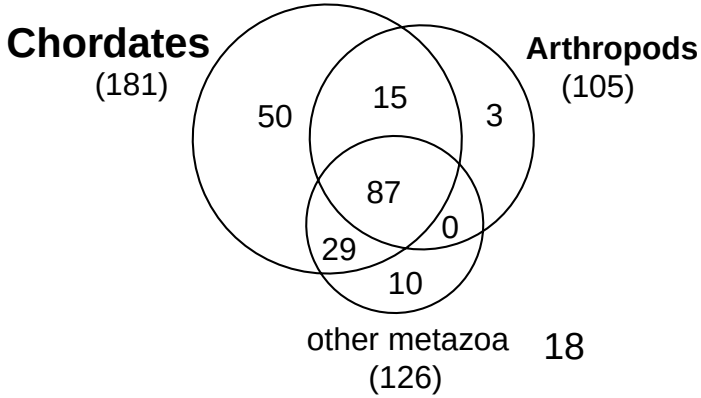




α/β







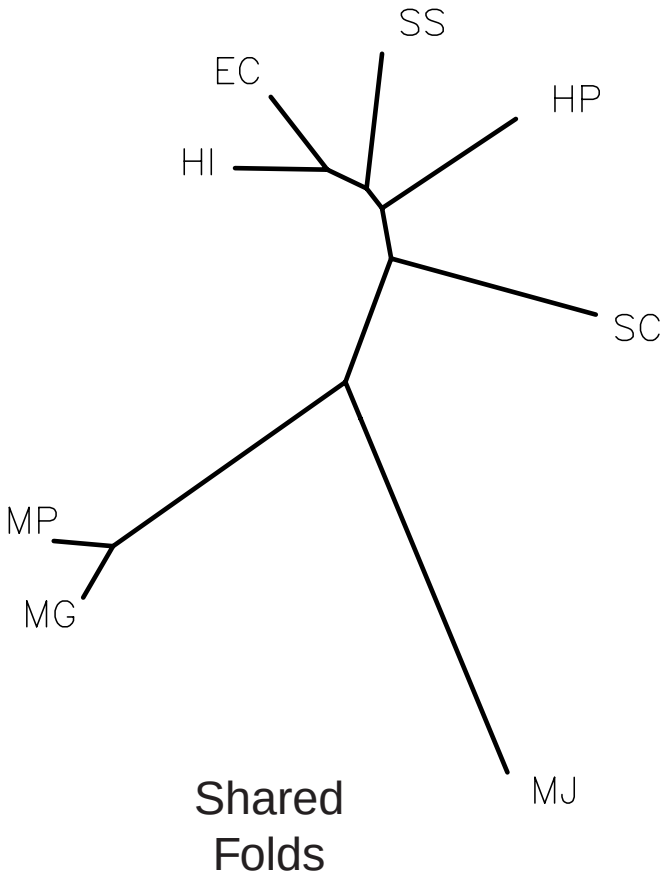
Cluster Trees Grouping 8 Initial Genomes on Basis of Shared Folds

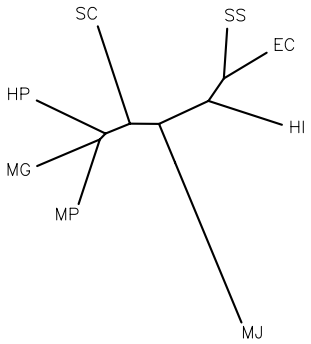
D=S/T

D = shared fold dist. betw. 2 genomes

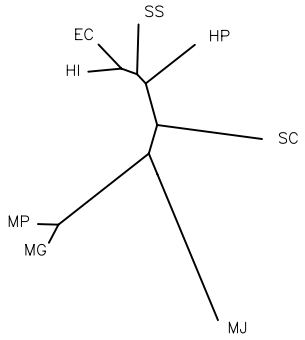
S = # shared folds

T= total # folds in both

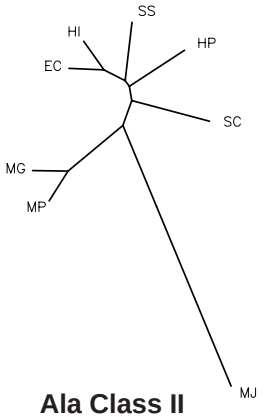




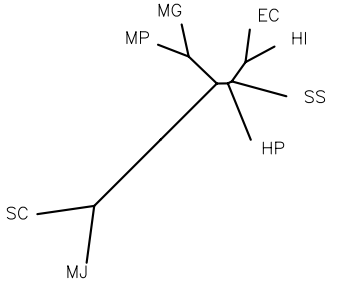
Overall Amino Acid Composition



Shared Sequence Families of Known Structure



Ala Class II synthetase similarity



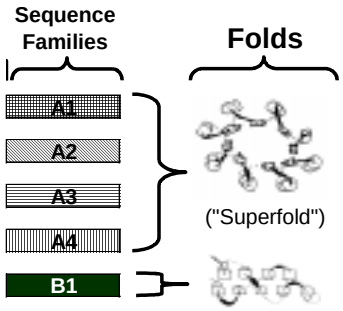
s17 ribosomal protein similarity (OB Fold)

Patterns of Folds Usage in 8 Genomes

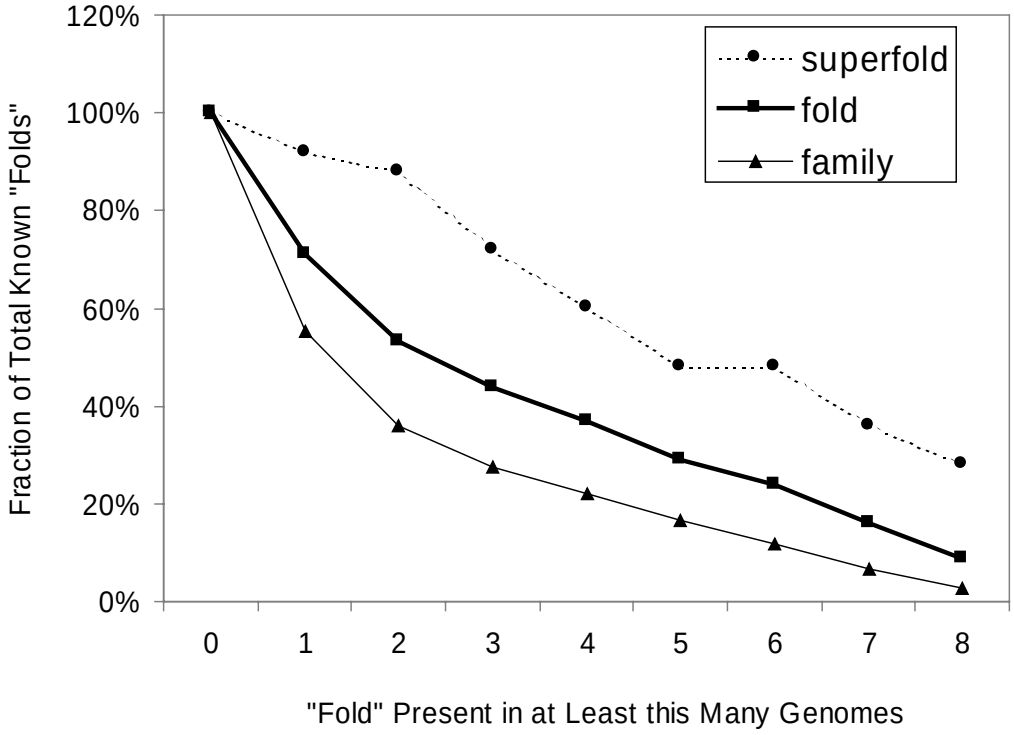
```
ESHSHMMM (##) ESHSHMMM (##) ESHSHMMM (##) ESHSHMMM (##) ESHSHMMM (##)
CCISPJPG      CCISPJPG      CCISPJPG      CCISPJPG      CCISPJPG

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)
```

	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/Orengo)



Top-10 Folds in 8 First Genomes

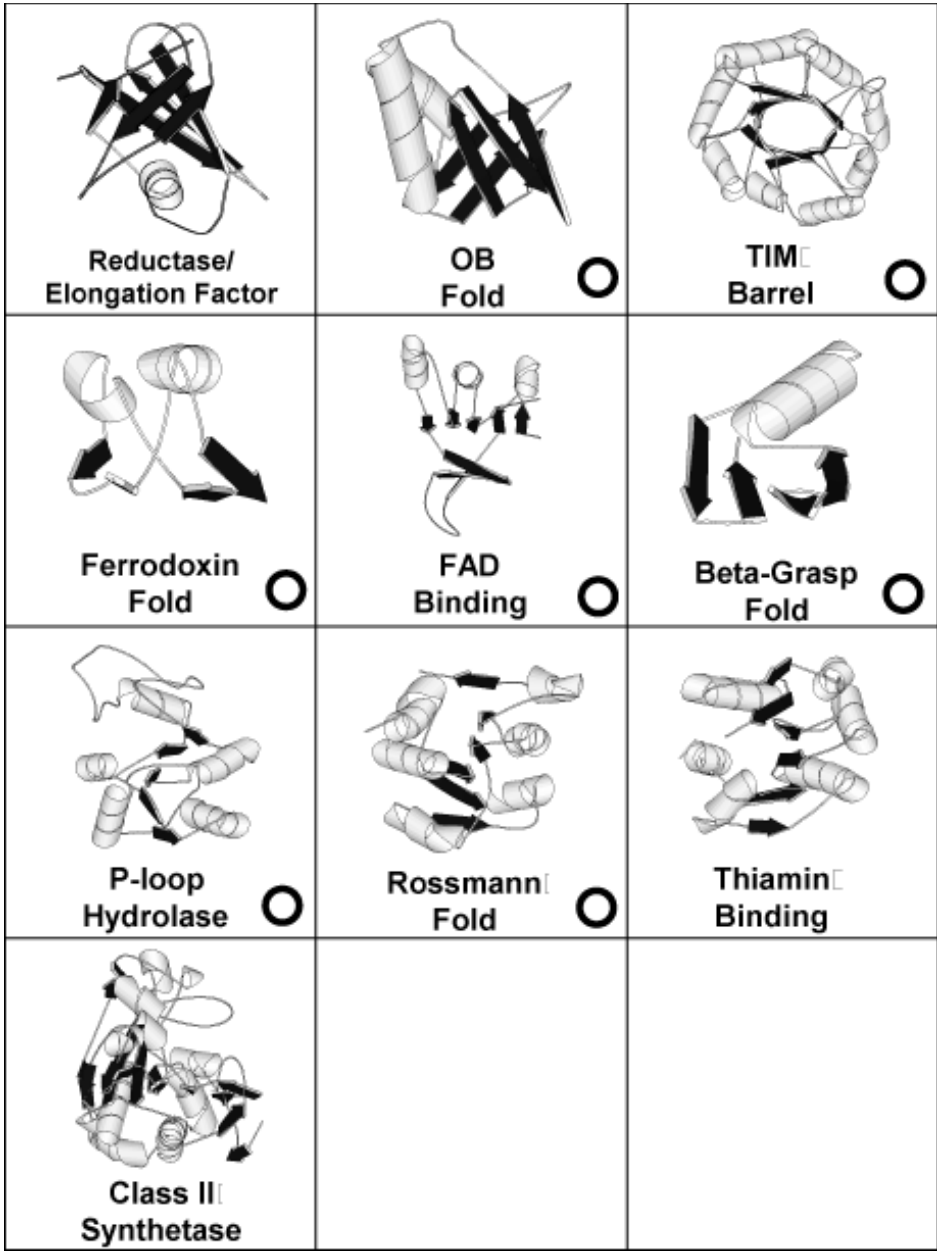
Num. in genome	Class	Fold Name	Representative Structure (PDB selection)
Top-10 in a eukaryotic genome (SC)			
84	α+β	Protein kinases (catalytic core)	1irk
49	α/β	**P-loop containing NTP hydrolases**	1gky
35	α/β	**Rossmann Fold**	2ohx A:175-324
31	α/β	**TIM barrel**	1tim A:
25	α/β	Ribonuclease H-like	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
Top-11 in a eubacterial genome (HI)			
18	α/β	**Rossmann Fold**	2ohx A:175-324
13	α/β	**P-loop containing NTP hydrolases**	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
Top-11 in an archaeal genome (MJ)			
19	α+β	Ferredoxin-like	1fxd
10	α/β	**P-loop containing NTP hydrolases**	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	Reductase/elongation factor common	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

What are the most common folds?

How many shared?

Many Super-folds

Mostly α/β

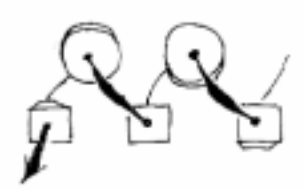


Top-5 Most Common Folds

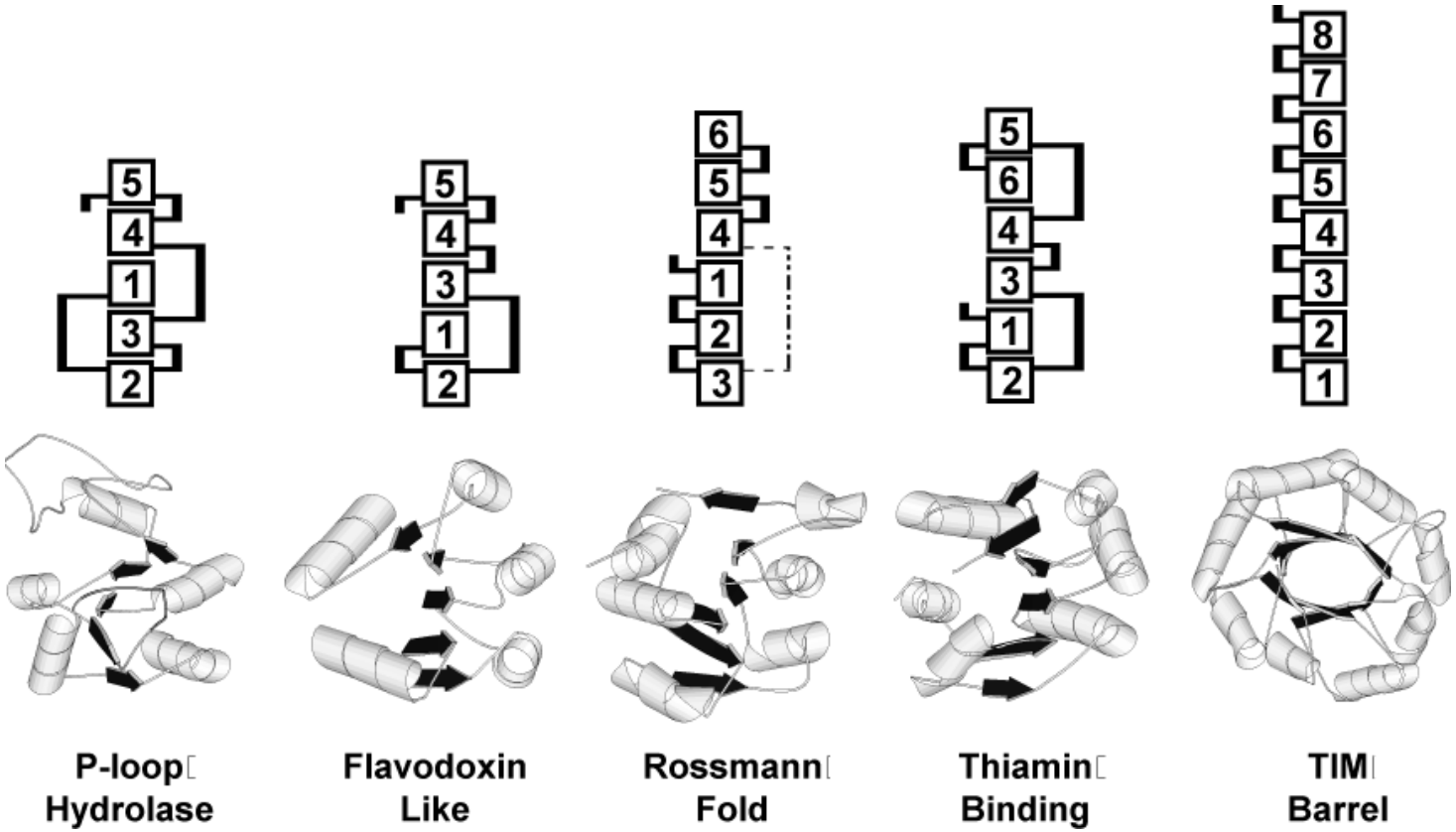
#	Class	SC Fold	Representative Structure (PDB selection)
84	α+β	Protein kinases (catalytic core)	1irk
49	α/β	NTP hydrolases with P-loop	1gky
35	α/β	Rossmann Fold	2ohx A:175-324
31	α/β	TIM barrel	1tim A:
25	α/β	Ribonuclease H-like	2rn2
18	S	Classic zinc finger	1zaa C:
14	α+β	Ubiquitin conjugating enzyme	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	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

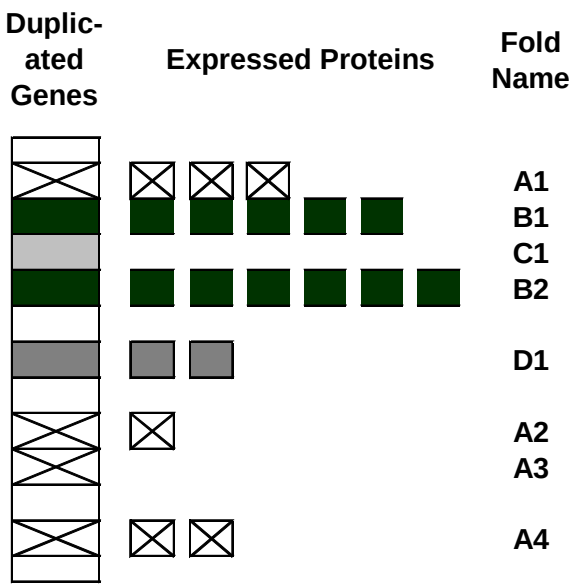
#	Class	MJ Fold	Representative Structure (PDB selection)
19	α+β	Ferredoxin-like	1fxd
10	α/β	NTP hydrolases with P-loop	1gky
7	α/β	TIM barrel	1tim A:
6	α/β	Rossmann Fold	2ohx A:175-324
5	α	Histone-fold	1ntx
4	α/β	Thiamin-binding Fold	1pvd A:2-181
4	α/β	Flavodoxin-like	3chy
4	β	Reduct./elongation fac. 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



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



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.

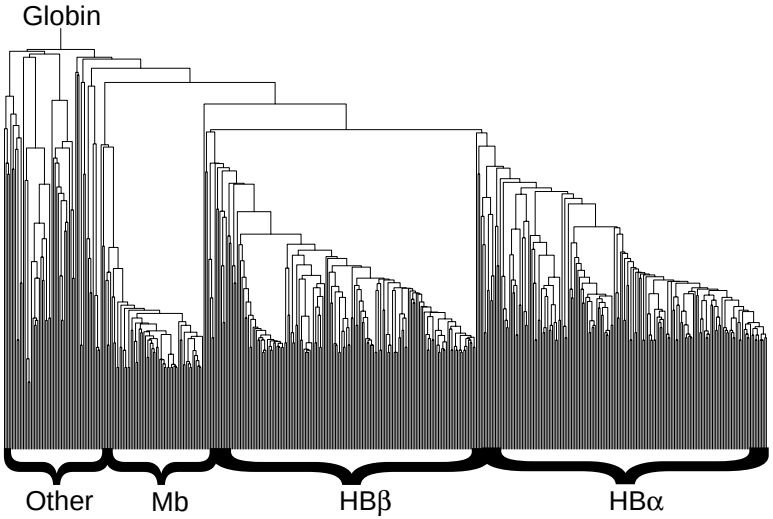
Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	**1**	**3**	**4**
NTP Hydrolases with P-loop	1gky	**2**	**1**	**2**
Classic Zn finger	1ard	**3**	9	**5**
Ribonuclease H-like motif	2rn2	**4**	**2**	**1**
Rossmann Fold	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	19319	1828
Overall Top-10				∇					
1REI-A	β	Immunoglobulin-like	32	13	◊	1	◊	25	◊
6TIM-B	α/β	TIM-barrel	29	6	◊	7	20	2	13
1ATP-E	O	Protein Kinases (catalytic core)	1	4	3	◊	3	6	6
1FXD	O	Ferredoxin-like	17	4	2	2	17	◊	8
1AKE-A	α/β	NTP Hydrolases containing P-loop	9	3	◊	5	3	2	7
1HDD-C	α	DNA-binding 3-helical bundle	13	3	◊	◊	2	5	◊
2HSD-A	α/β	Rossmann Fold (NAD binding)	11	3	◊	7	3	1	3
1MBD	α	Globin-like	3	2		1	◊	4	1
2RN2	α/β	like Ribonuclease H	15	2	5	1	2	1	5
1ZNF	S	Classic Zinc Finger	2	1	◊			3	1
Sequence Family Top-11				∇					
1REI-A	β	Immunoglobulin-like	32	13	◊	1	◊	25	◊
6TIM-B	α/β	TIM-barrel	29	6	◊	7	20	2	13
1FXD	O	Ferredoxin-like	17	4	2	2	17	◊	8
2RN2	α/β	like Ribonuclease H	15	2	5	1	2	1	5
1PYP	β	OB-fold	15	◊	◊	1	◊	◊	◊
1PTX	S	Small inhibitors, toxins, lectins	14	◊			3	◊	◊
2TBV-C	β	Viral coat and capsid proteins	14	1	12				
1HDD-C	α	DNA-binding 3-helical bundle	13	3	◊	◊	2	5	◊
2HSD-A	α/β	Rossmann Fold (NAD binding)	11	3	◊	7	3	1	3
1RCF	α/β	Flavodoxin-like	11	◊	◊	4	◊	◊	◊
1RCB	α	4-helical cytokines	11	◊	◊	◊		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

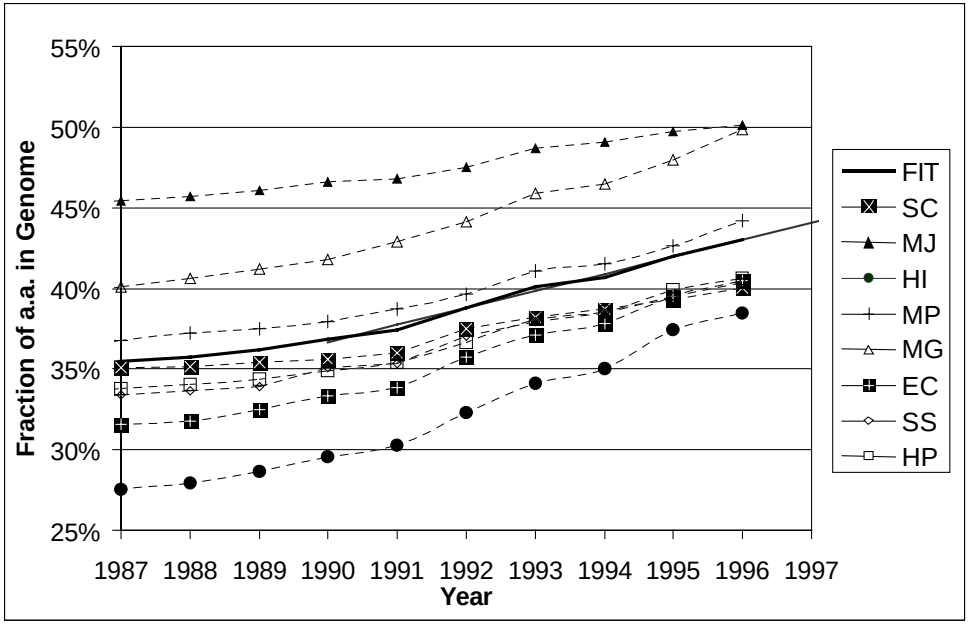
An Issue with Fold Counting: Biases in the Databanks



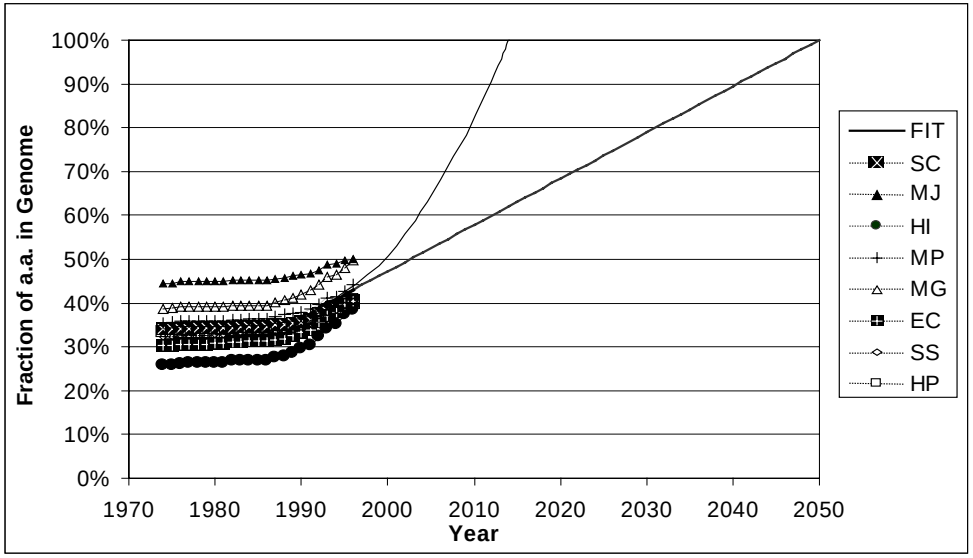
Example Structure (PDB)	Fold Name	Percentage of known folds in genome	Rank in eubacterial Top-10
Top-10 in a bacterial genome (H. influenzae)			
2HSD-A	Rossmann Fold (NAD binding)	9.6	**1**
1AKE-A	NTP Hydrolases containing P-loop	5.7	**3**
1RCF	Flavodoxin-like	5.1	**4**
6TIM-B	TIM-barrel	4.5	**2**
1FXD	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**

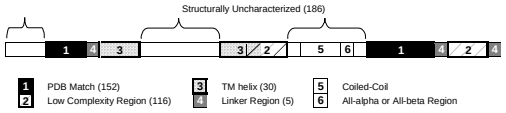
- Over-representation of certain species and functions in the databanks (e.g. human v. plant globins, Ig's)
  - Nevertheless HI top-10 like eubacterial top-10
- PDB is small, biased sample of genome (6-12%)
- Selection of structures in PDB is biased, anectdotal
- Different numbers with different comparison sensitivity: FASTA, HMM, &c
- Some Correction with Seq. Weighting, Diff. Sampling

The Problem: All Folds in Genome Not Known until 20?? → Prediction



- Separate TM, LC, linkers
- How many residues in genome matched by known folds, in 1975, '76, '77...'00...'50





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

1 Library of Known Folds

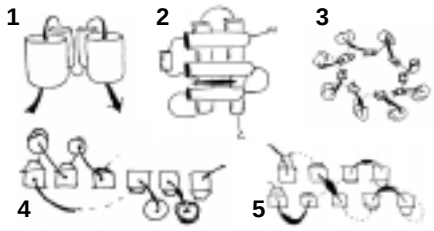
Importance of statistics. Scop auto-alignments. Significance follows EVD stats, same as sequences.

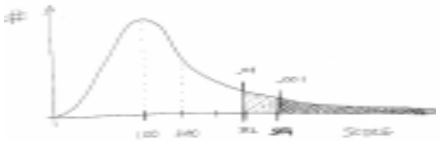
2 A Census of Known Folds

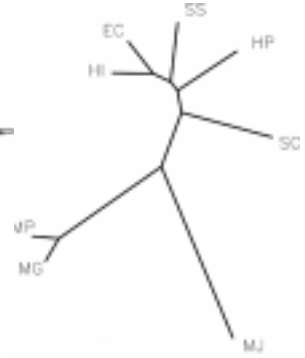
Which folds in which Organisms: Plants v. People, E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression, repeated βαβ supersecondary struc.

3 Prediction of Unknown Folds

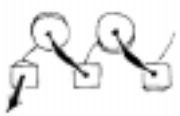
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Can extrapolate from known structures to genomes?

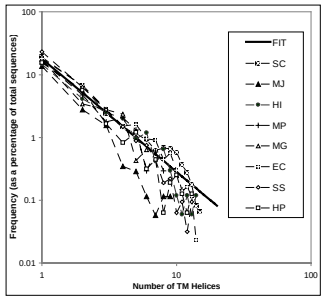


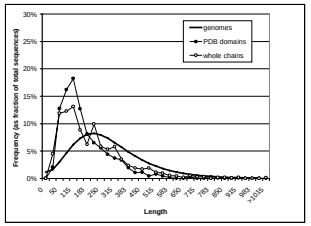




Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	1z5d	6	6	7
7-bladed beta-propeller	2bbk-N	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







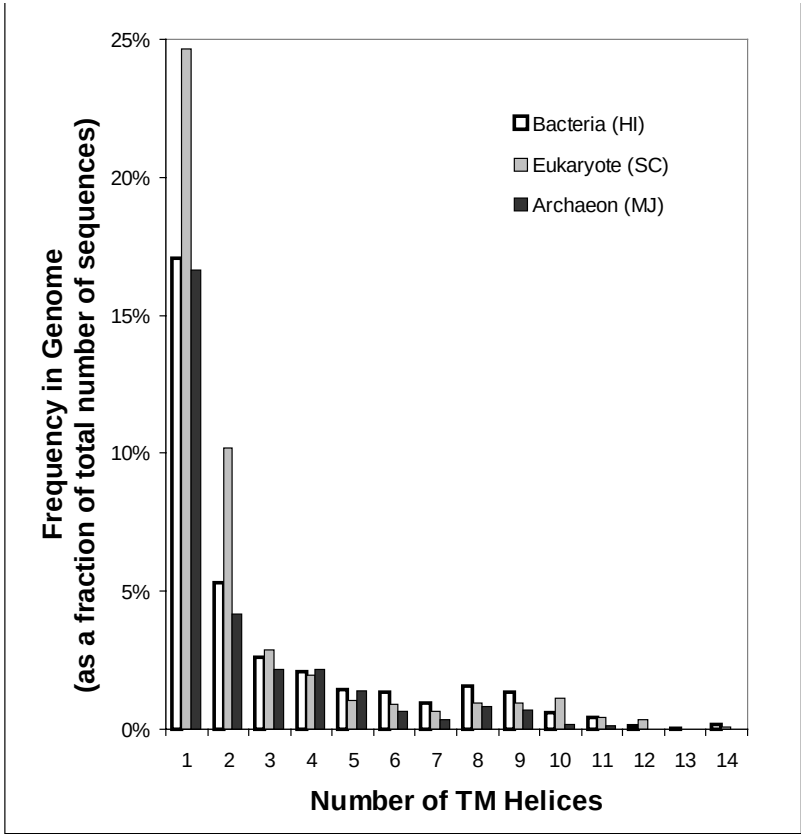
http://bioinfo.mbb.yale.edu/genome

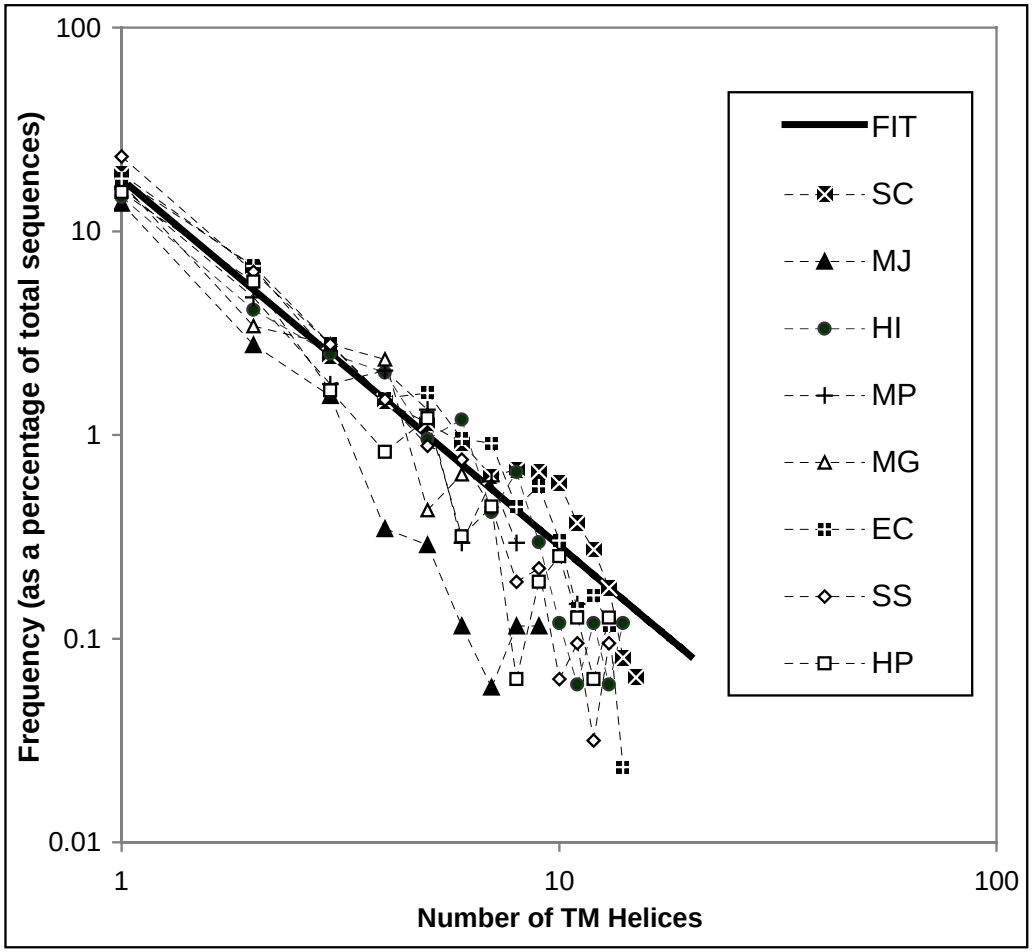
Acknowledgements: M Levitt, scop (Murzin, Brenner, Ailey, Hubbard, Chothia)

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.

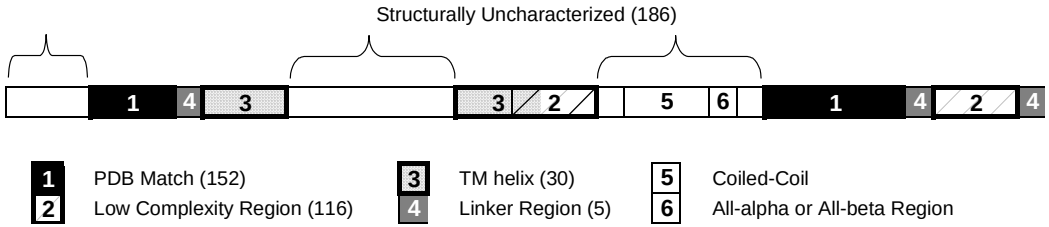
- No preference for particular supersecondary structures: 7-TM's
- Freq. of Number of TM helixes follows a Zipf-like law: **$F=1/[5n^2]$**





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

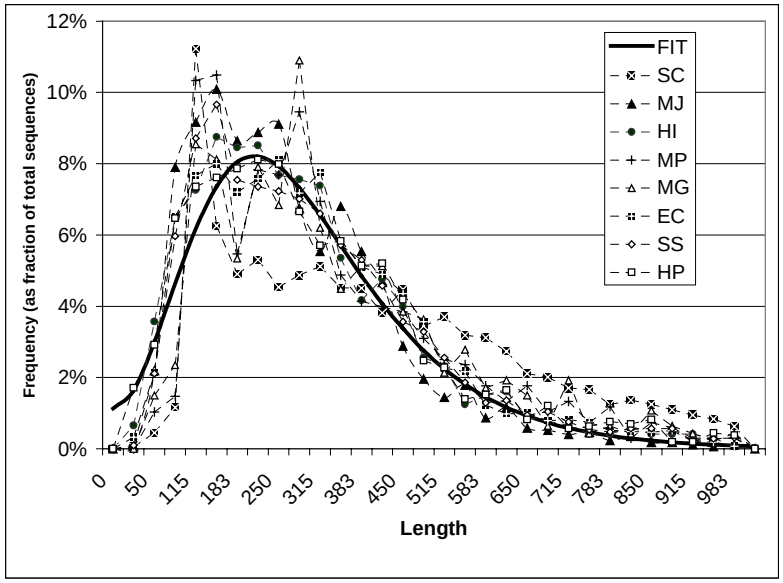
HI more αα, ααα, αααα ...

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
βββ	41%	0.96	1.15	**1.46**	1.62
ααα	19%	**1.01**	0.83	0.84	0.92
αβα	18%	1.04	1.03	0.87	1.16
ααβ	15%	1.03	0.97	0.89	0.70
βαβ	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
αααα	54%	1.03	1.35	**1.78**	2.28
αααα	29%	**1.10**	0.82	0.89	1.18
βββα	25%	0.85	0.94	1.10	0.98
βαβα	23%	1.11	**1.18**	0.94	1.48
αβαβ	21%	1.21	**1.23**	0.99	1.39
αβαα	21%	1.00	0.95	0.81	1.00
…	…	…	…	…	…

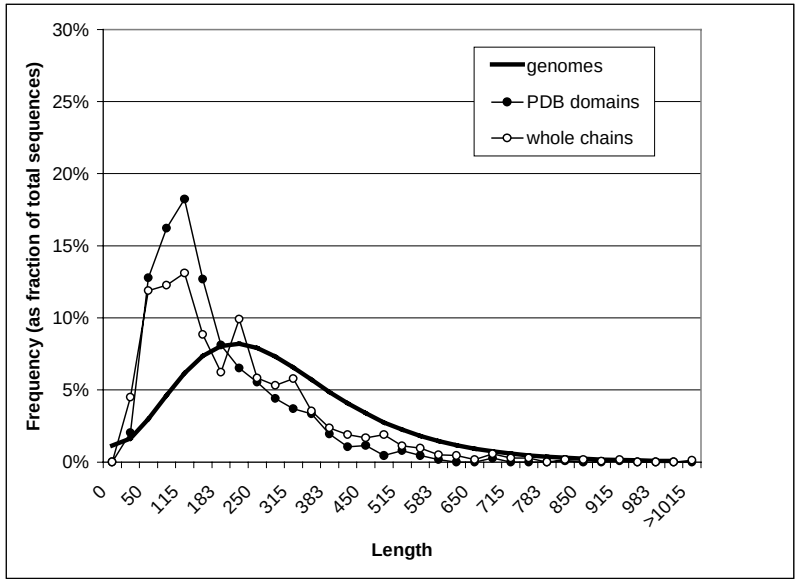
Genomes Sequences are longer than those in Known Structures



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

How Representative are the Known Structures of the Proteins in Complete Genome? Is prediction (extrapolation) based on known structures justified?

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



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

Amino Acid Composition

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

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

1 Library of Known Folds

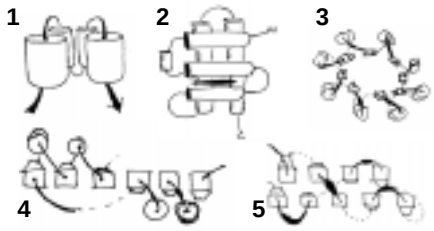
Importance of statistics. Scop auto-alignments. Significance follows EVD stats, same as sequences.

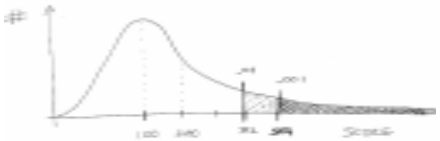
2 A Census of Known Folds

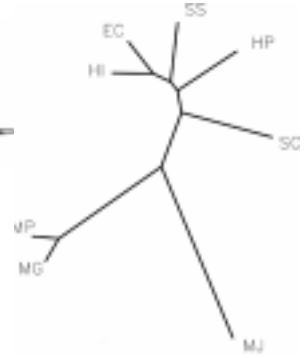
Which folds in which Organisms: Plants v. People, E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression, repeated βαβ supersecondary struc.

3 Prediction of Unknown Folds

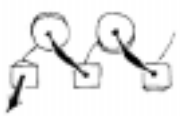
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Can extrapolate from known structures to genomes?

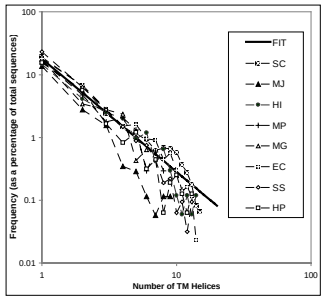


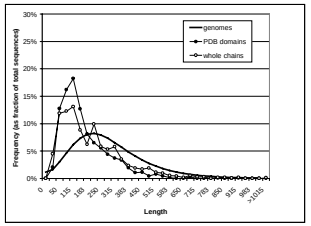




Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	1z5d	6	6	7
7-bladed beta-propeller	2bbk-N	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







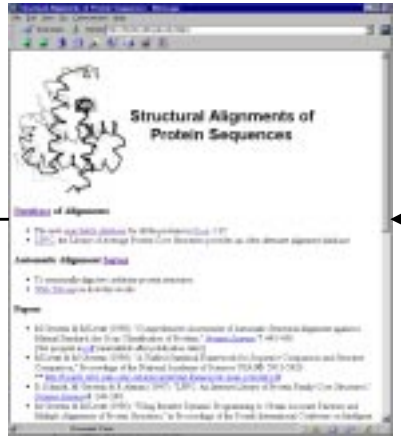
http://bioinfo.mbb.yale.edu/genome

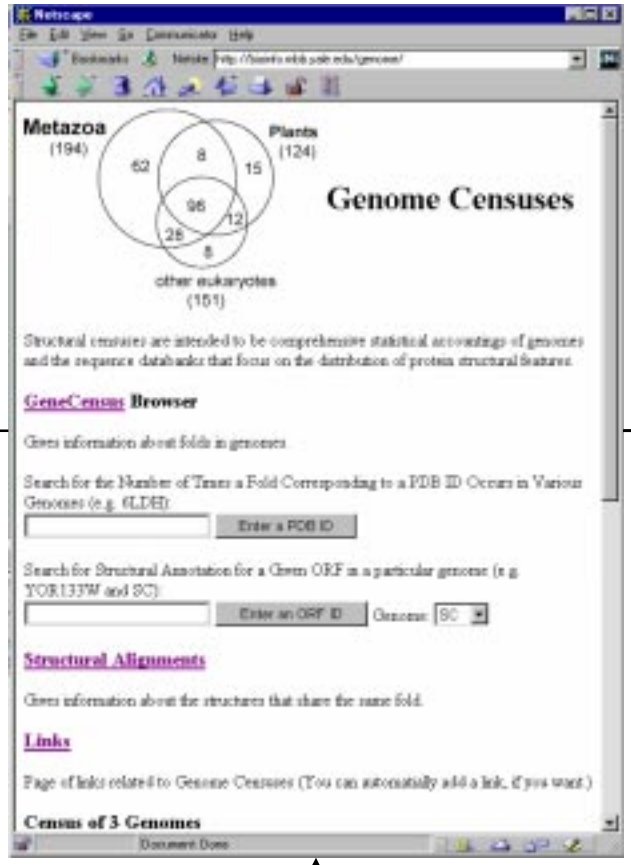
Acknowledgements: M Levitt, scop (Murzin, Brenner, Ailey, Hubbard, Chothia)

GeneCensus

Alignment Database

Alignment Server





Detailed Tables

PDB Query

ORF Query

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

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

1 Library of Known Folds

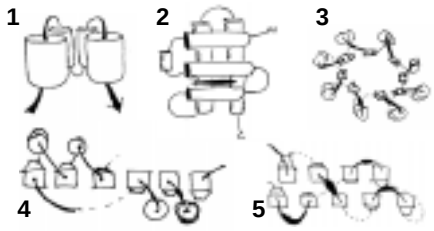
Importance of statistics. Scop auto-alignments. Significance follows EVD stats, same as sequences.

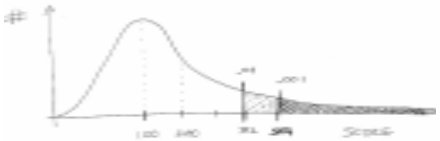
2 A Census of Known Folds

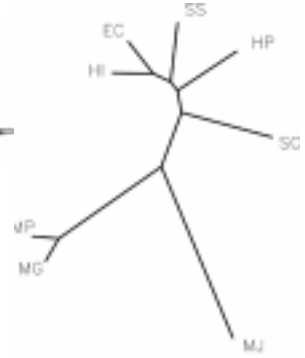
Which folds in which Organisms: Plants v. People, E coli v. yeast? Shared Fold Tree. Top-10 by duplication/expression, repeated βαβ supersecondary struc.

3 Prediction of Unknown Folds

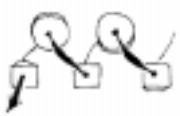
Zipf law for TM's but no 7-TM's. Same 2º comp. but different a.a. comp. Can extrapolate from known structures to genomes?

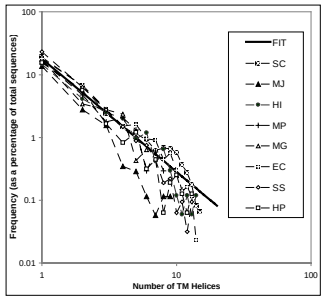


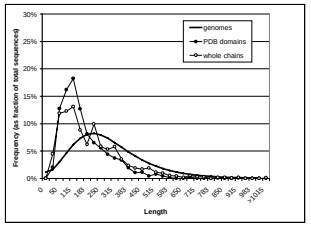




Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	1z5d	6	6	7
7-bladed beta-propeller	2bbk-N	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







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

Acknowledgements: M Levitt, scop (Murzin, Brenner, Ailey, Hubbard, Chothia)

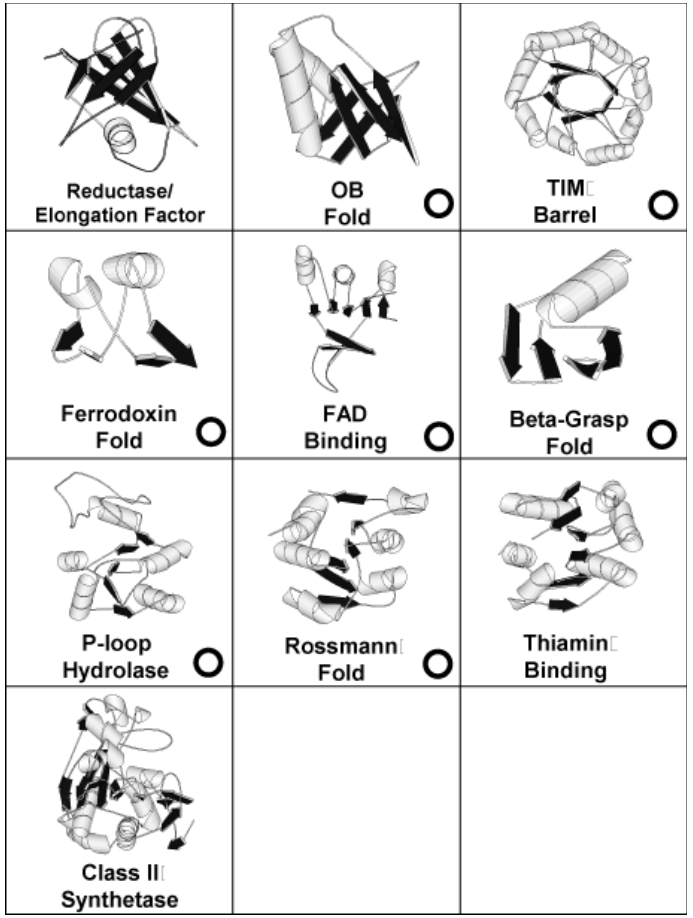
Issues with Top-10

Depend on:
Database (version) used, Biases
Which Genomes
Comparison Prog. (FASTA/BLAST)
Threshold (1e-3, 1e-4)

Want a Fair Census with Uniform & Consistent Sampling

Common Yeast Folds	Rep. Struct.	GeneCensus		Sacch3D, SGD		diff.
		count	rank	count	rank	
Protein kinases (cat. core)	1hcl	110	1	109	1	
NTP Hydrolases with P-loop	1gky	69	2	52	2	
Classic Zn finger	1ard	55	3	34	7	.
Ribonuclease H-like motif	2rn2	54	4	30	8	.
Rossmann Fold	1xel	46	5	41	5	
Zn2/Cys6 DNA-binding dom.	125d	46	6	30	9	.
7-bladed beta-propeller	2bbk-H	46	7	0	-	<
TIM-barrel	1byb	36	8	39	6	.
Ferrodoxin-like	1fxd	28	9	43	4	.
DNA-binding 3-helix bundle	1enh	22	10	22	10	
Long Helix Oligomers (coils)	1zta	1	-	47	3	<

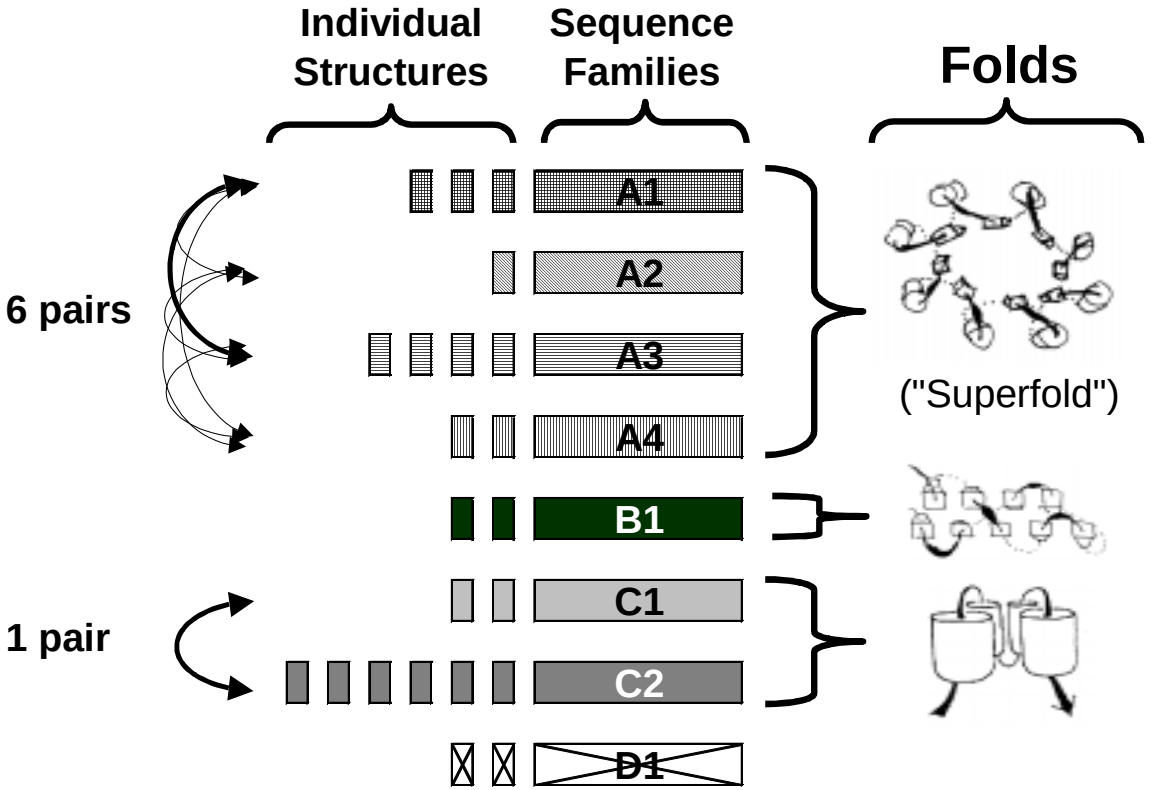
Top-10 in 8 Genomes



Sacc3D
(Chervitz, Cherry)
vs GeneCensus

Clustering the PDB, assess comparison methods

- Structure of “scop”
  - ◊ ~10K structure **domains**
  - ◊ ~1K sequence **families**
  - ◊ ~350 **folds**
  - ◊ 25 **superfolds** (>10 fam)
  - ◊ ~2K + ~2K true positive pairs (1.32)
- Different Sequence Thresholds for Clustering
  - ◊ high threshold, struc. sim.
  - ◊ Low threshold, seq. sim.
  - ◊ .01 e-value for all comparisons



- Brenner et al. used scop pairs to calibrate the statistics of comparison methods
  - ◊ How many of the TP pairs can be found with FASTA?
  - ◊ e-value ~ epq
- Credits: scop, Murzin, Brenner, Ailey, Hubbard, Chothia; Thornton, superfolds; Pearson, fasta