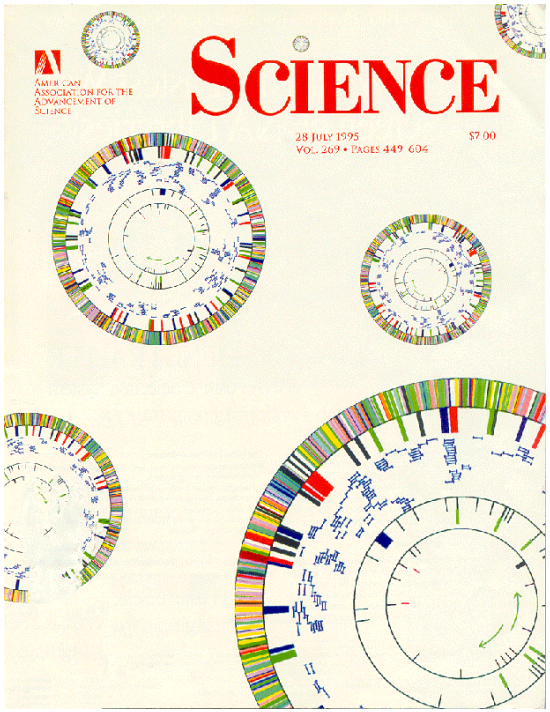
Comparative Genomics:

Surveys of a Finite Parts List

Mark Gerstein

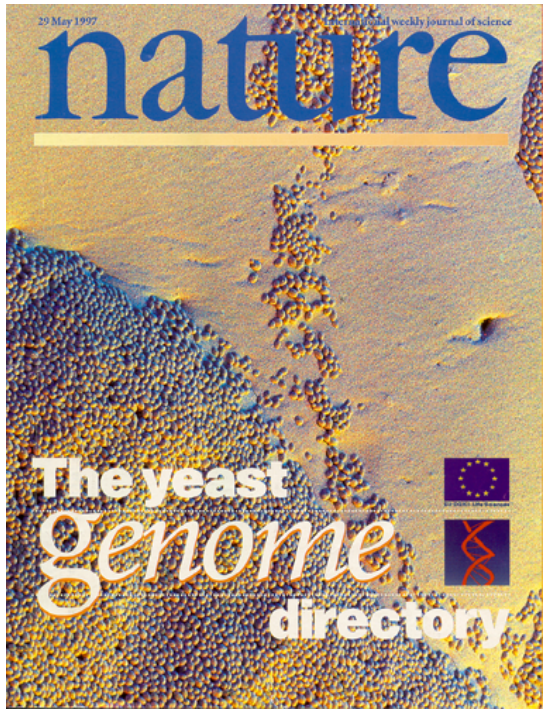
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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



1995

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

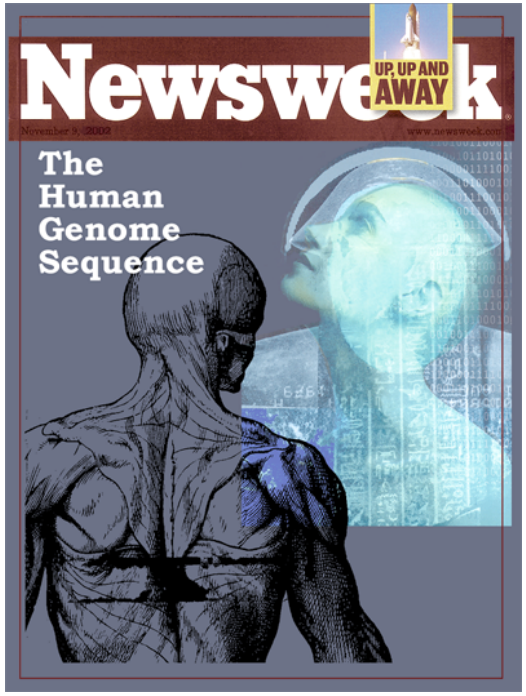


1997

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

1998

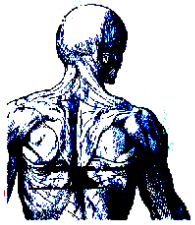
Animal, ~100 Mb, ~20K genes [*Science* **282**: 1945]



2000?

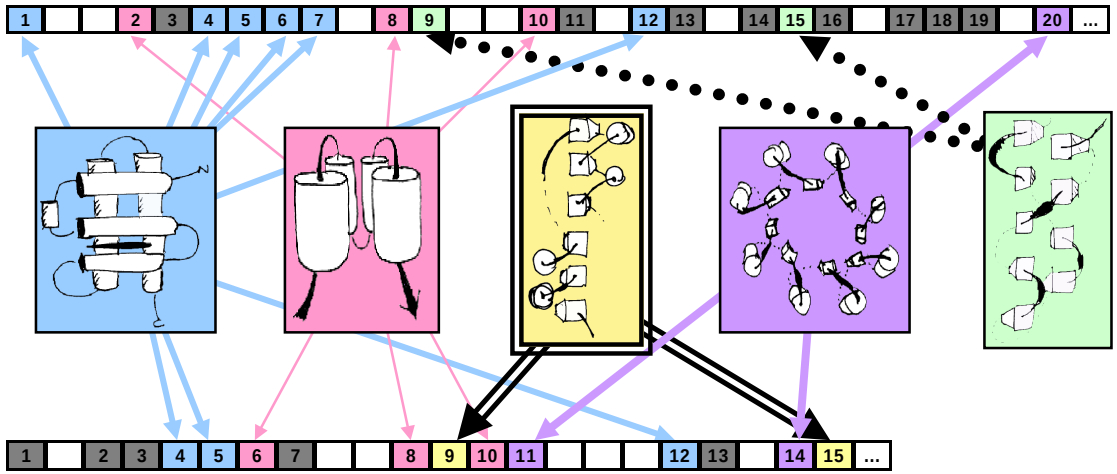
Human, ~3 Gb, ~100K genes [???]

World of Structures is even more Finite, providing a valuable simplification



(human)

(*T. pallidum*)



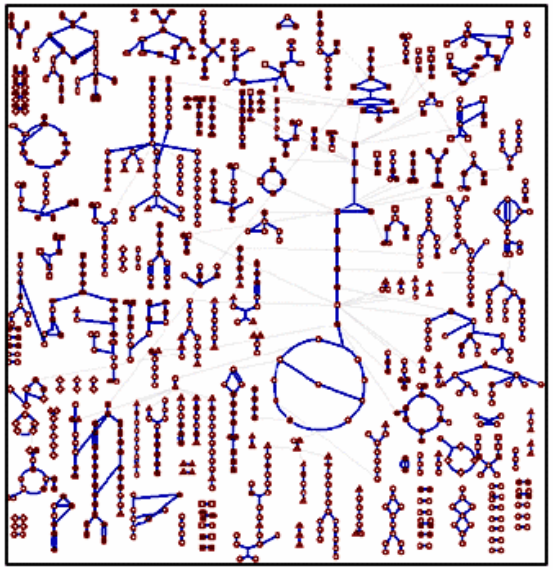
~100000 genes

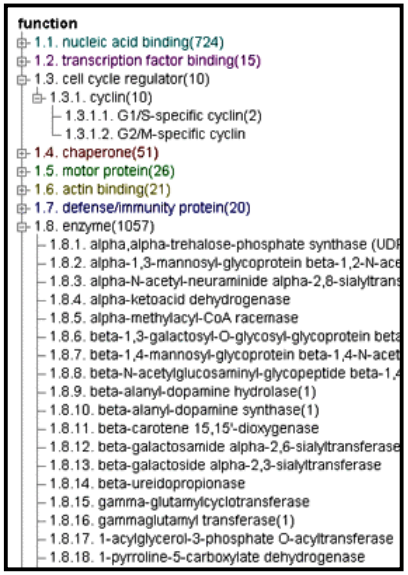
~1000 folds

~1000 genes

Likewise, for pathways, functions, sequence families, blocks, motifs....

Cross-referencing and interrelating among a Different Types of Parts

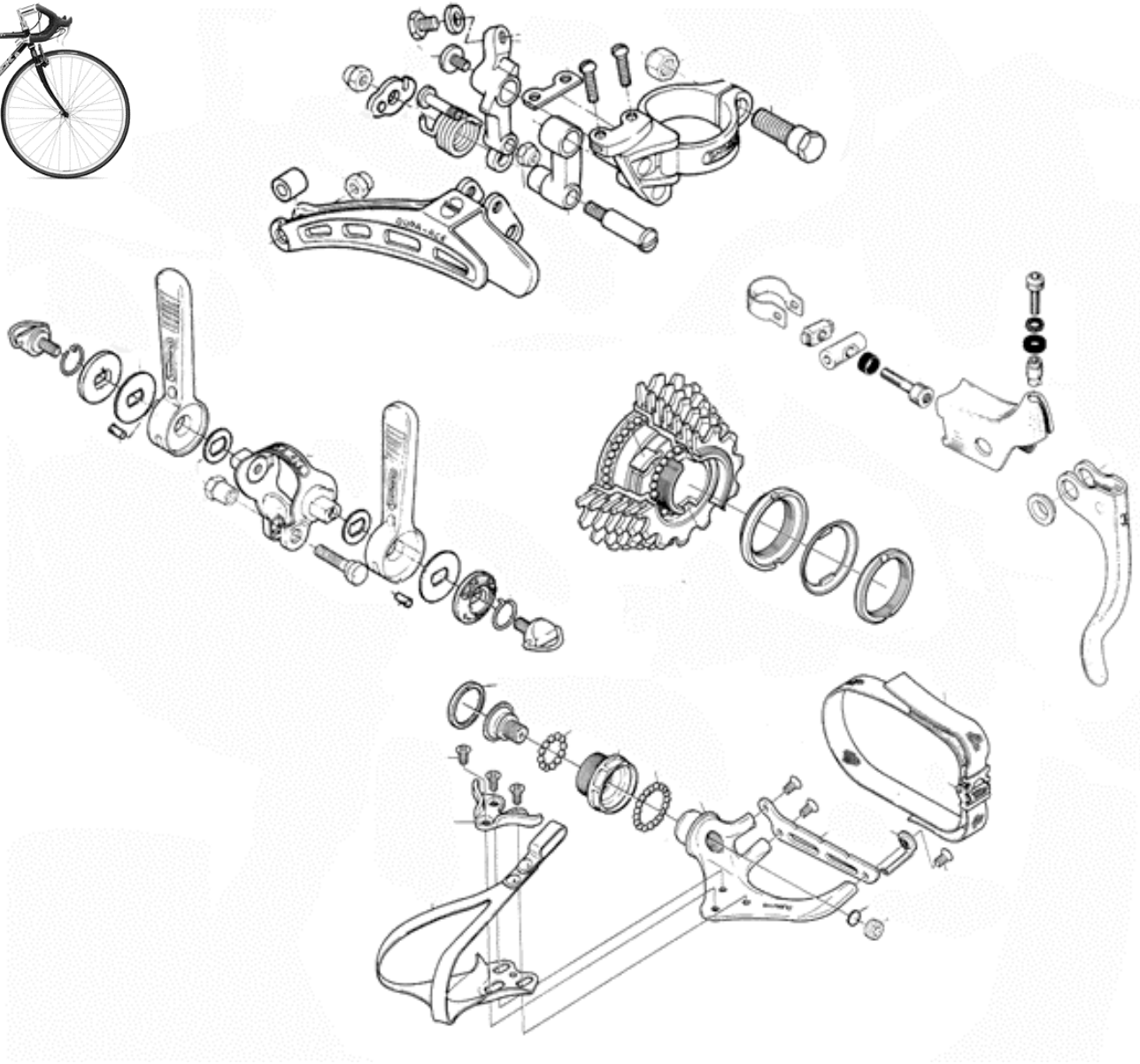




Functions picture from www.fruitfly.org/~suzi (Ashburner); Pathways picture from, ecocyc.pangeasystems.com/ecocyc (Karp, Riley). Related resources: COGS, ProDom, Pfam, Blocks, Domo, WIT, CATH, Scop....

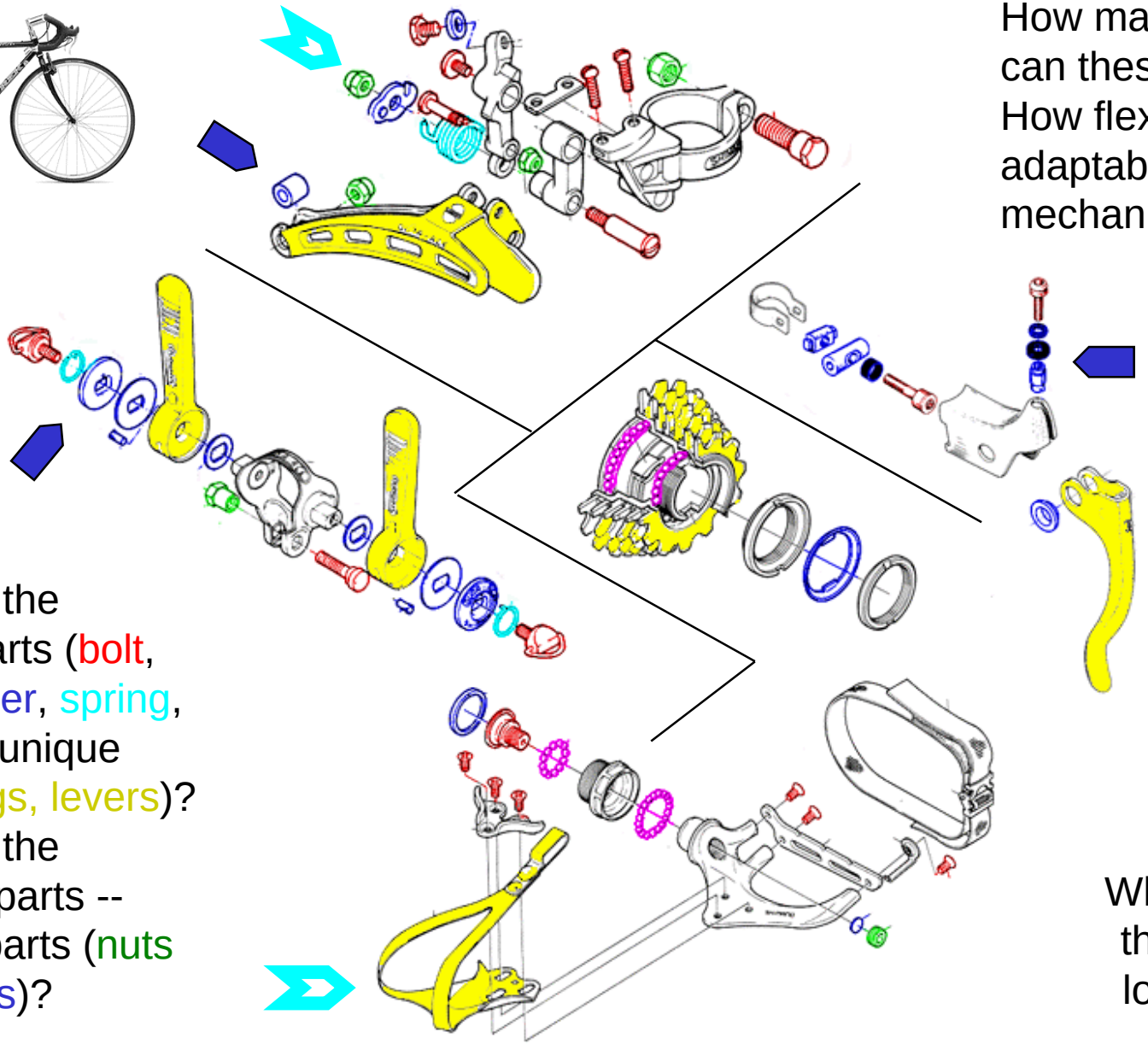
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

A Parts List Approach to Bike Maintenance



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

A Parts List Approach to Bike Maintenance



How many roles can these play? How flexible and adaptable are they mechanically?

What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

Where are the parts located?

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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

W Krebs, J Tsai, M Levitt, C Wilson, R Das, H Hegyi, J Lin, Y Kluger, C Arrowsmith, A Edwards, L Regan, S Balasubramanian, A Drawid, D Greenbaum, M Snyder, R Jansen

bioinfo.mbb.yale.edu

Structures ("Classic")

(now) Structural Genomics

(now) Func. Genomics

Arrays (future)

Structures ("Classic")

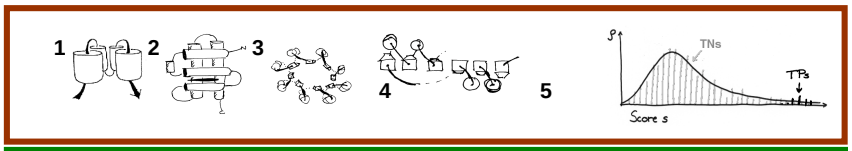
1 Fold Library (A parts list.) Structural Alignment, EVD P-value, Seq. Struc. diverg.

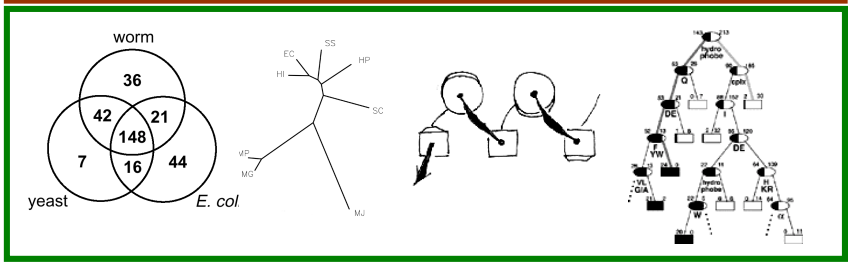
2 Folds in Genomes (Shared, common, and/or unique parts?) Known Folds. Fold Tree, Top-10. βαβ. Biases. MG fold assignment extent. MG Target Selection, MT retrospective decision tree.

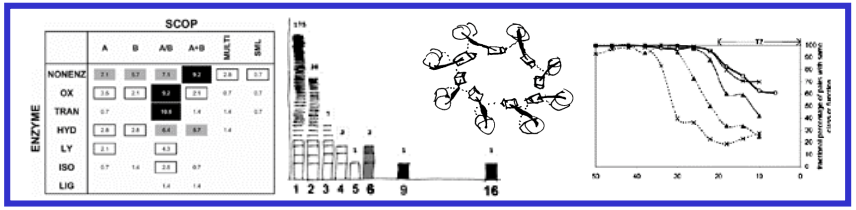
3 Folds & Functions (Roles/part?) How many folds /function? Mostly 1, but TIM versatile. Seq. diverg. vs. Func. diverg.

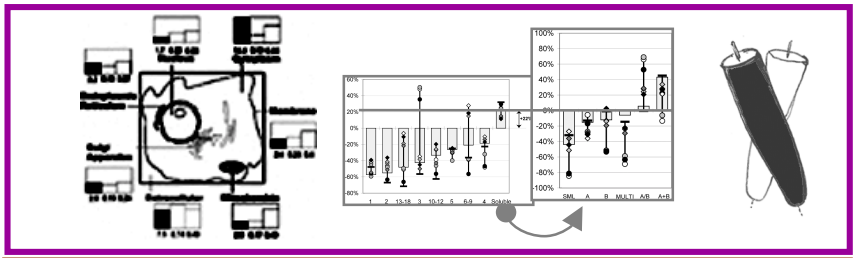
4 Folds in the Transcriptome (Common parts? Where are parts?) Enriched ↑ : VGA, TIM, αβ folds, energy, synthesis, cyt. Depleted ↓ : NS, long, TM folds, transport, transcription, Leu-zip, nuc. Bayesian Localizer, phenotypes clustering

5 Fold Flexibilty (How adaptable is a part?). Motions DB, morph server, interface packing, Voronoi Volumes



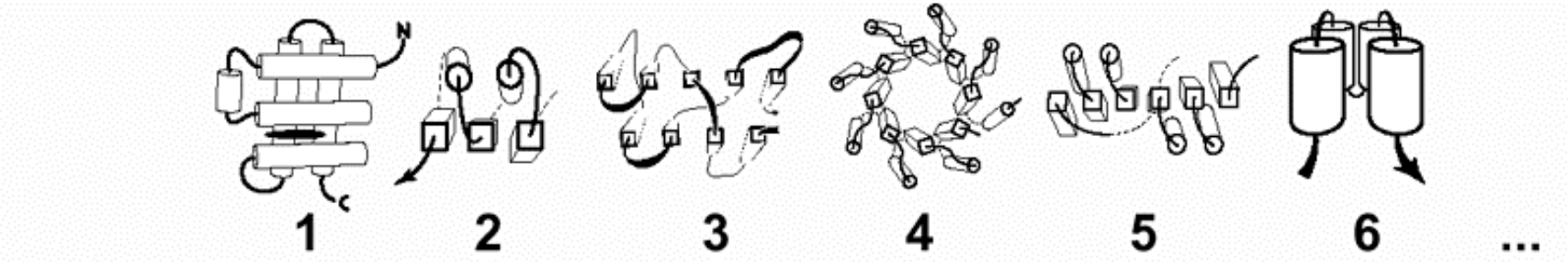






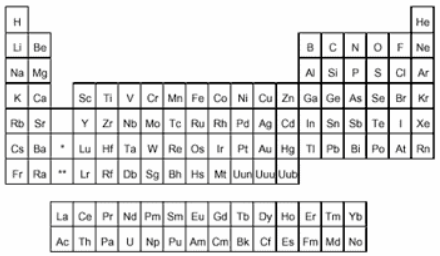
Fold Library vs. Other Fundamental Data structures

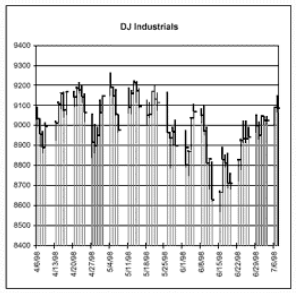
Parts List **Database**; **Statistical**, rather than mathematical relationships and conclusions



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}





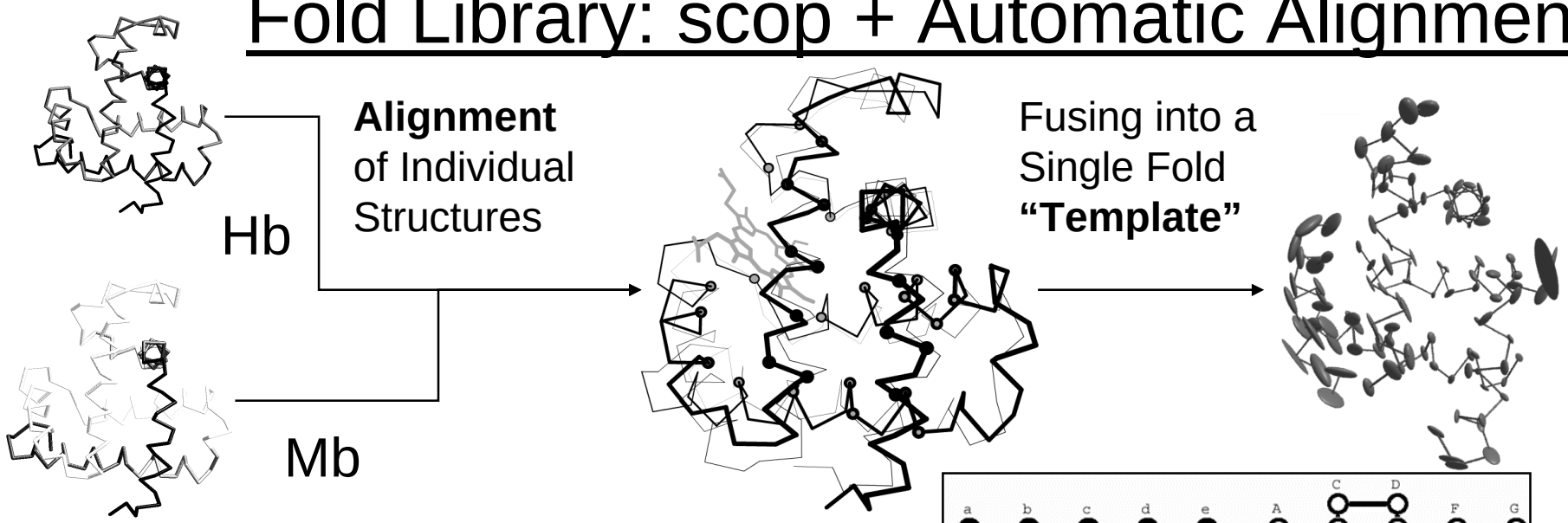
10 **100** **1000 -10000** **>1000000**

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.

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Fold Library: scop + Automatic Alignments

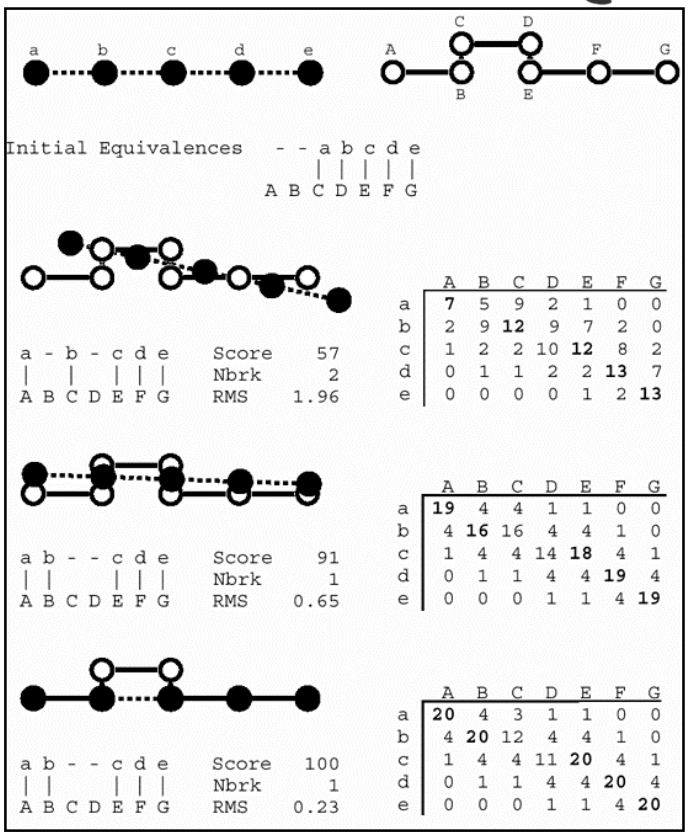


Iterative Dynamic Programming

Like repeated sequence alignment; 1 cycle doesn't converge, violates key assumption of D.P.

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

Derived from Program of G Cohen (Align, Satow et al. 1986); scop (Murzin et al., 1995). Much 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



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Some Similarities are Readily Apparent others are more Subtle

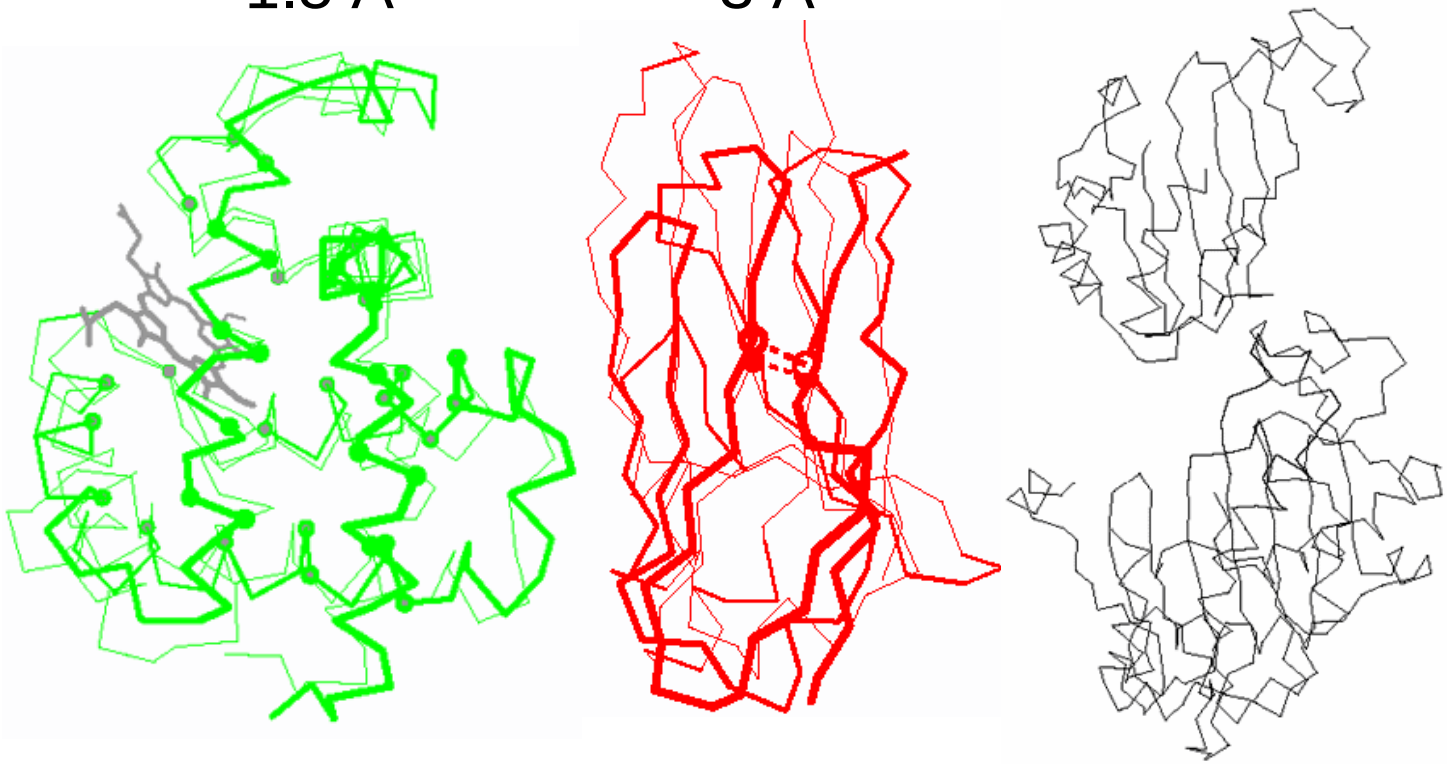
Easy:
Globins

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

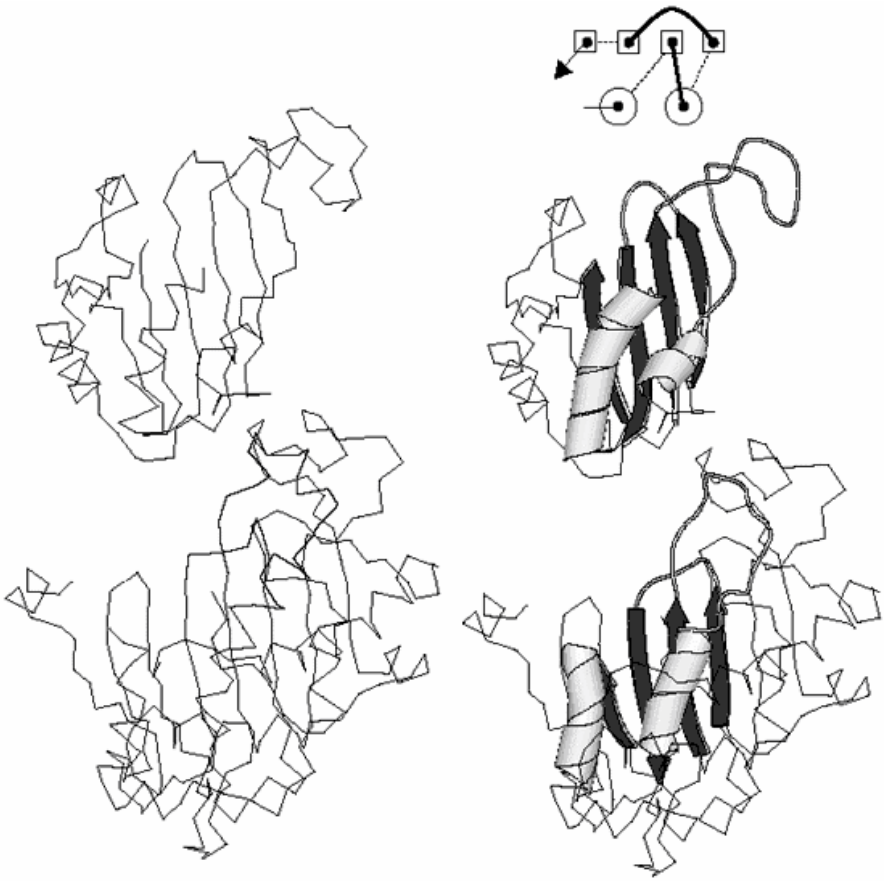
Easy:
Globins

125 res.,
~1.5 Å

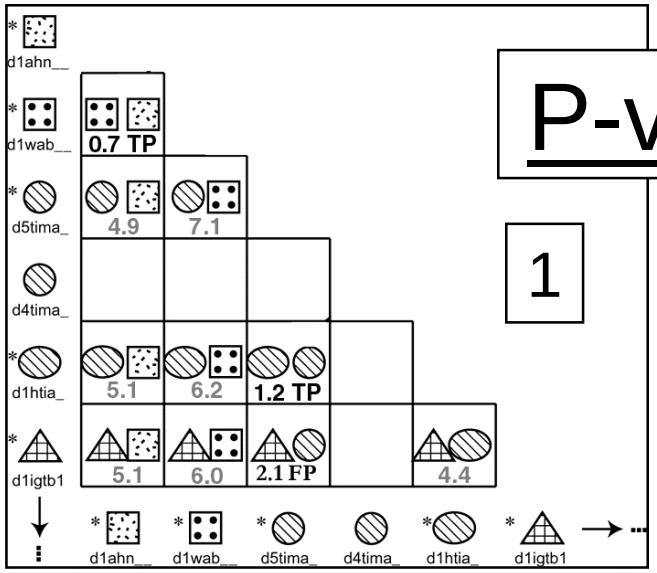
Tricky:
Ig C & V

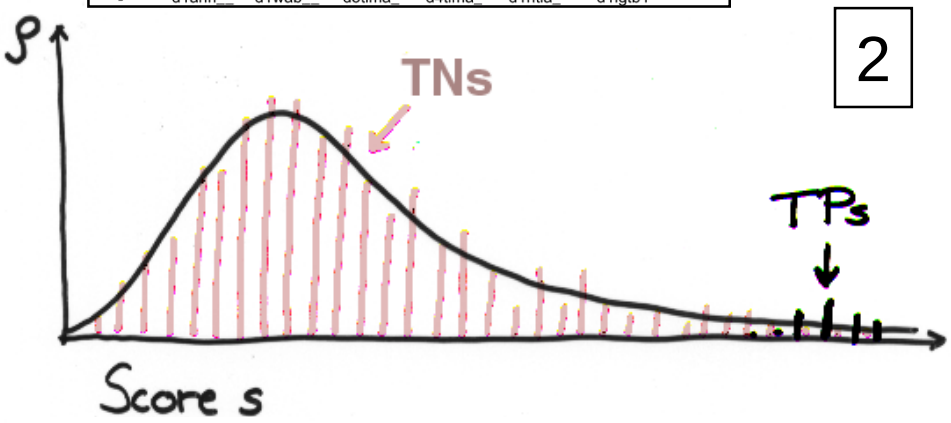
85 res.,
~3 Å

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

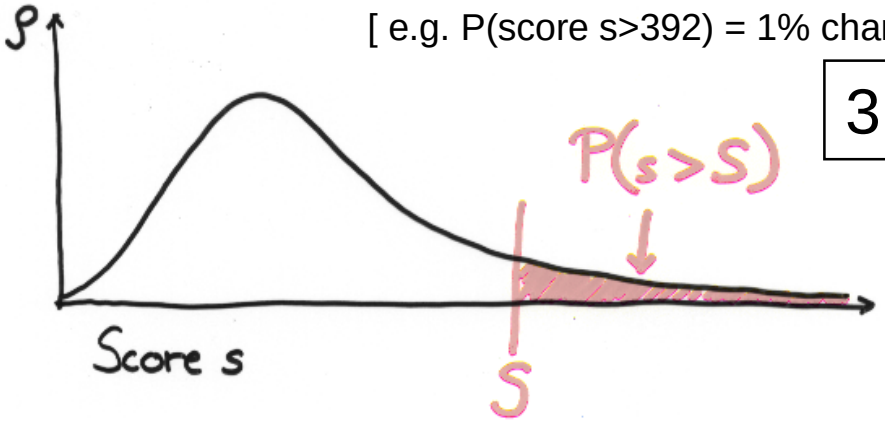


P-values

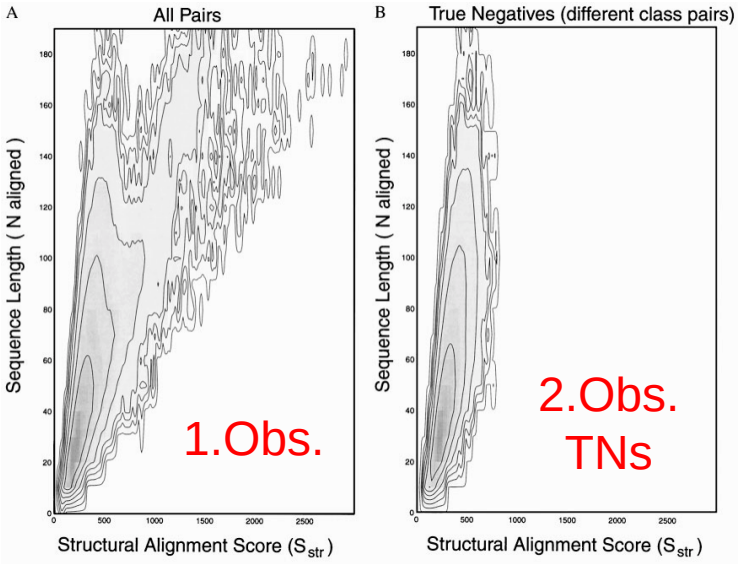


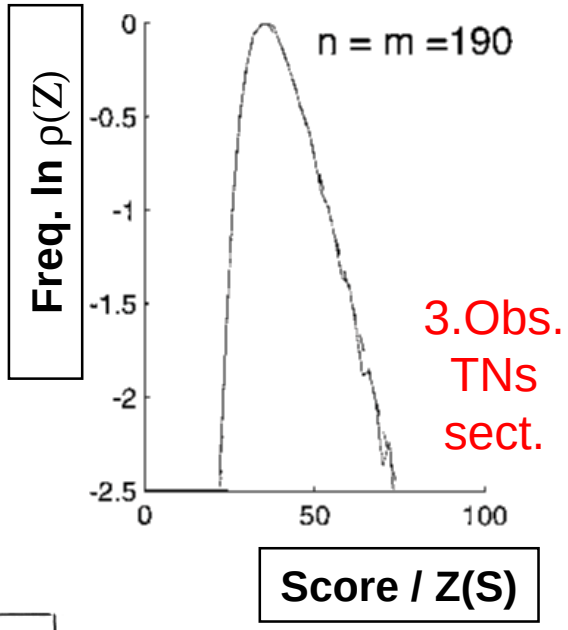


[e.g. P(score s>392) = 1% chance]



- 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 Empirical Distributions
  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),, chance of getting a score better than threshold S randomly
  4) **Same formalism for sequence & structure**





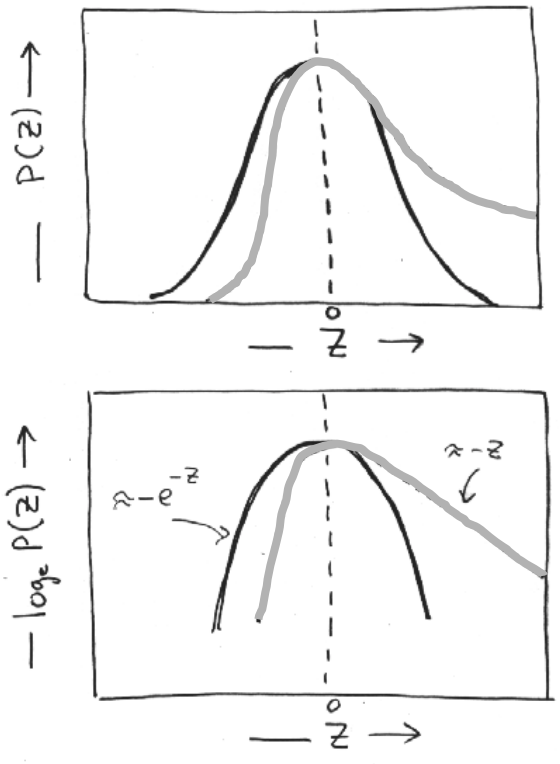
EVD Fits

3-parm fit works; Reasonable as Dynamic Programing maximizes over pseudo-random variables

5. Extreme **V**alue v. Normal Distributions:

EVD ~ **Max**(indep. random variables),

Normal ~ **Sum**(indep. random variables) [$\exp(-z^2)$]



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

4.

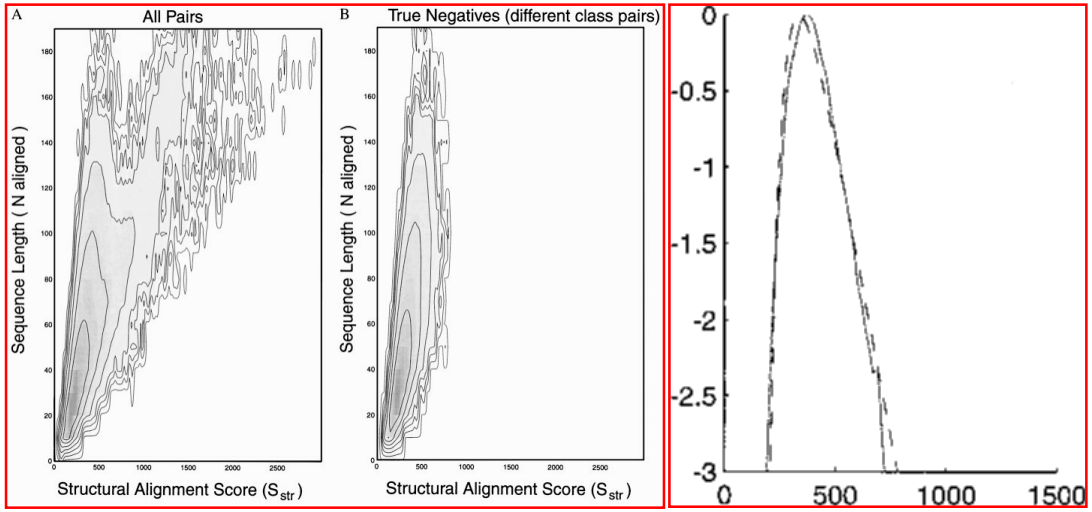
$$Z = \frac{S - (a \ln N + b)}{\sigma}$$

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

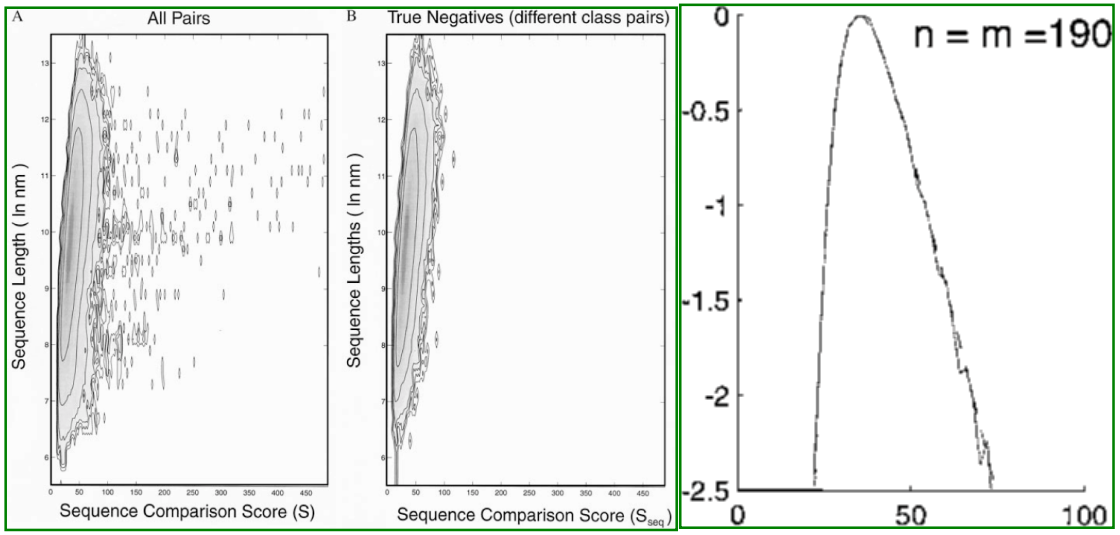
3 Free Parm. fit to EVD matches distributions bes
Fit involves: ***a, b,σ***. ***N*** = number of residues matched; ***G*** = total gap penalty; ***M***(i,j) = similarity matrix (Blossum for seq. or M_{str}(i,j), struc.)

Same EVD Results for Sequence & Struct.

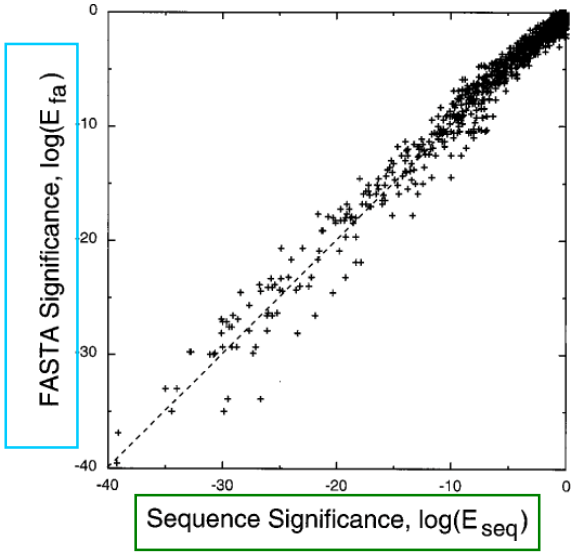
Structure



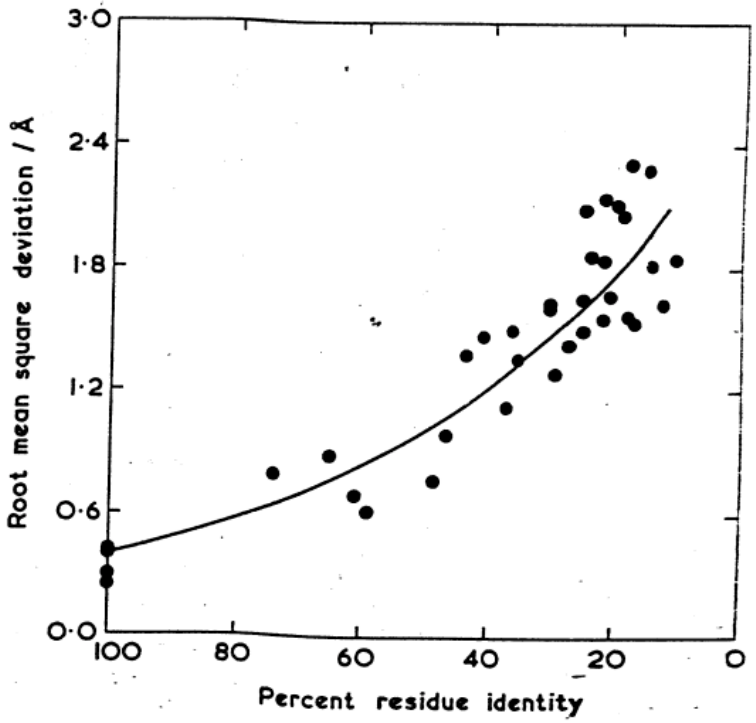
Sequence



- Use Sequence Scores to Validate
- Sequence P-value perfectly tracks FASTA e-value



Significance computation can be applied to **any** existing sequence or structure alignment; added Benefit: allows computation of an e-value without doing a db run



Chothia & Lesk, 1986 revisited, 32 →16K pts.

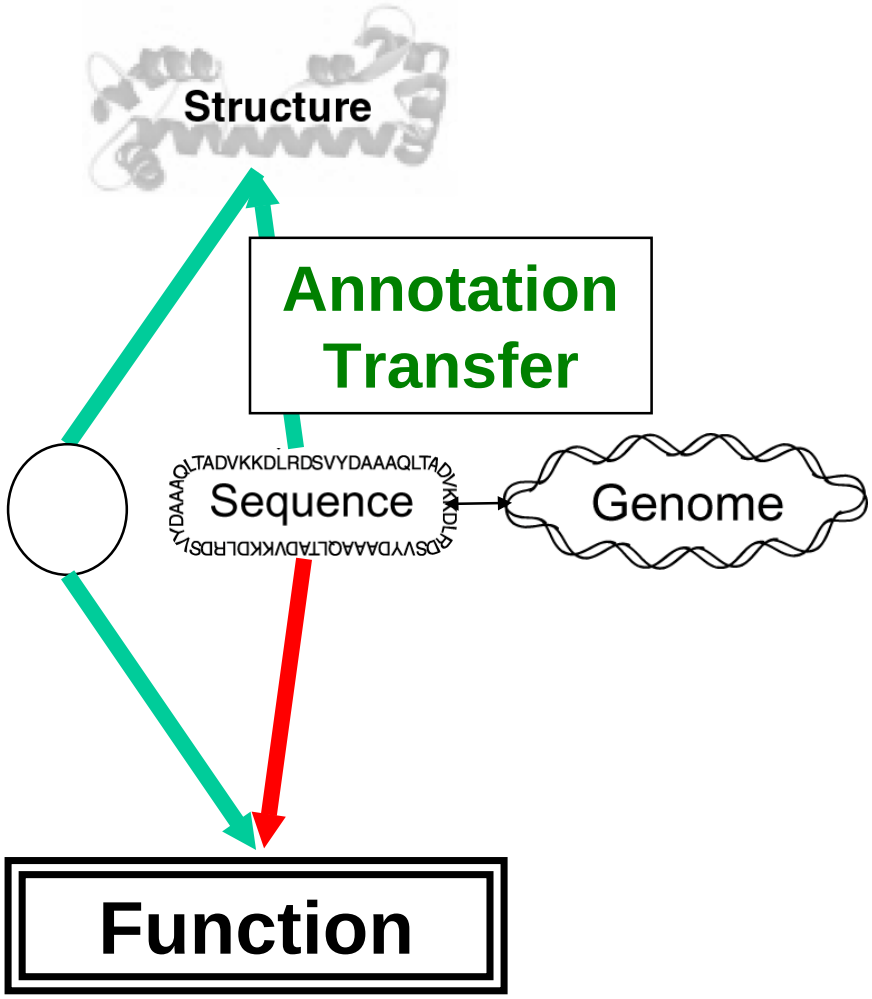
C&L Δ=.4 exp(1.9 H)

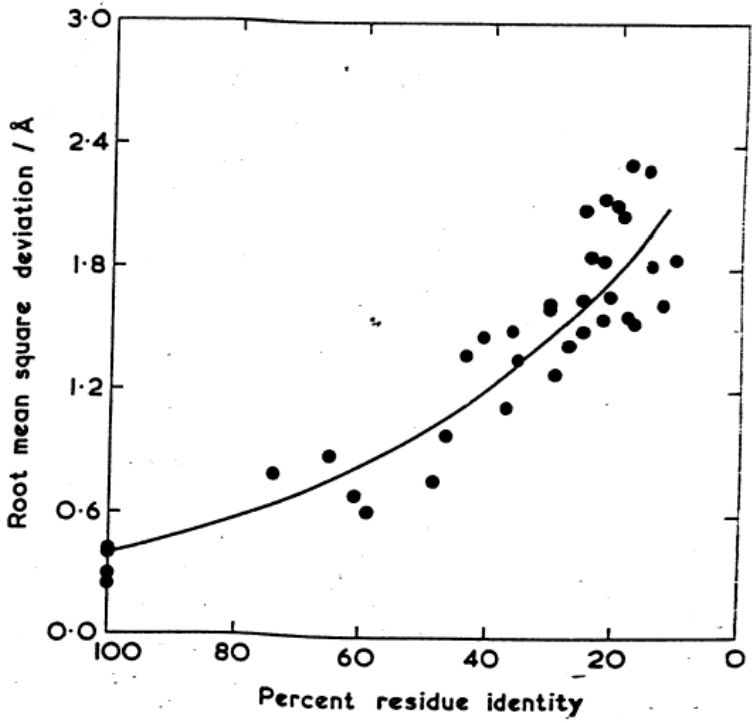
Multi: Δ= .2 exp(1.3 H)

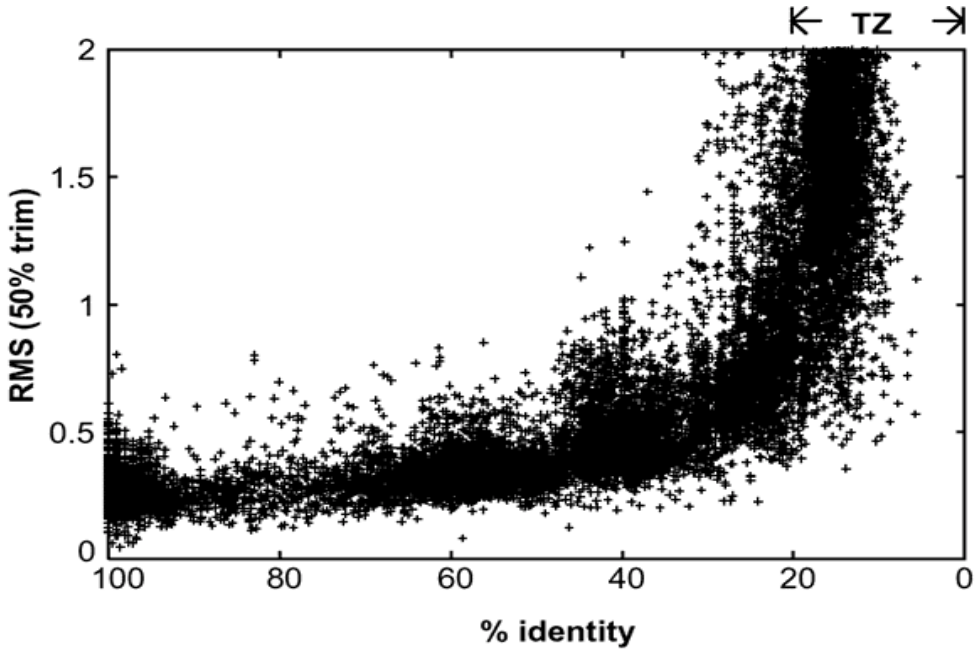
Single: Δ= .2 exp(1.9 H)

Related work done by Thornton, Pearson, Brenner

What's the use of sequence-structure scoring schemes? Precise Annotation Transfer for Genomics







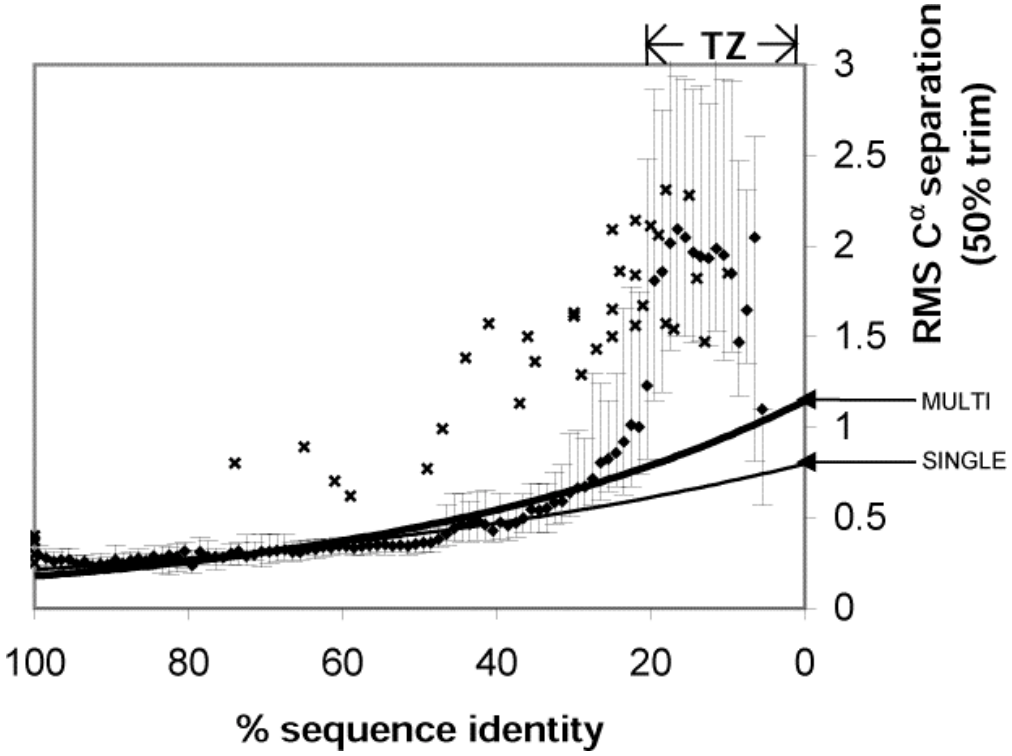
Chothia & Lesk, 1986 revisited, 32 →16K pts.

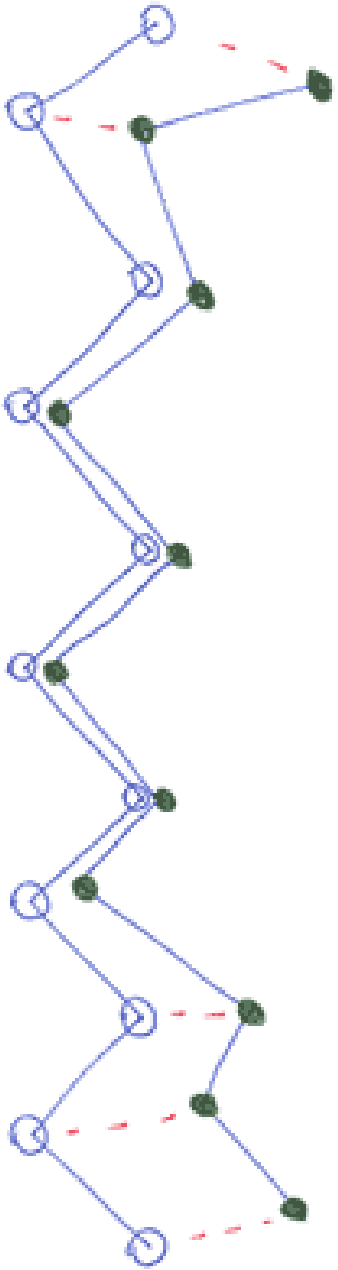
C&L Δ=.4 exp(1.9 H)

Multi: Δ= .2 exp(1.3 H)

Single: Δ= .2 exp(1.9 H)

Related work done by Thornton, Pearson, Brenner





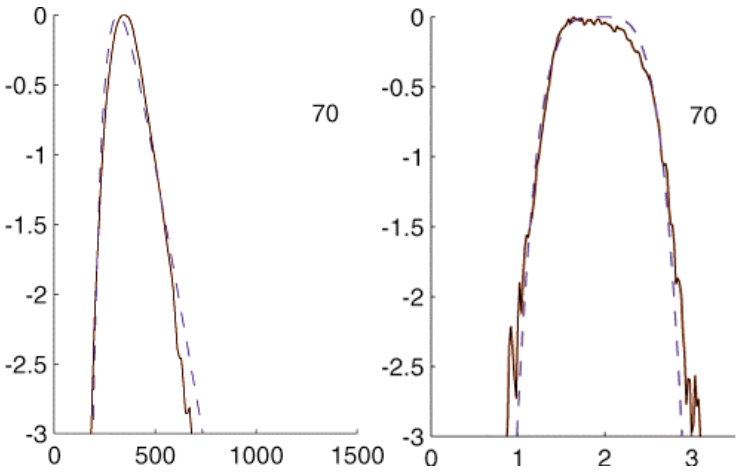
Problems with RMS and %ID

- Difference not similarity, NO EVD fit
- Dominated by worst-fitting atoms, easily skewed
- Trimming is arbitrary (50%)

S_{str} RMS

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

%ID problem: "Bunching up" between 20% and 0% identity



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

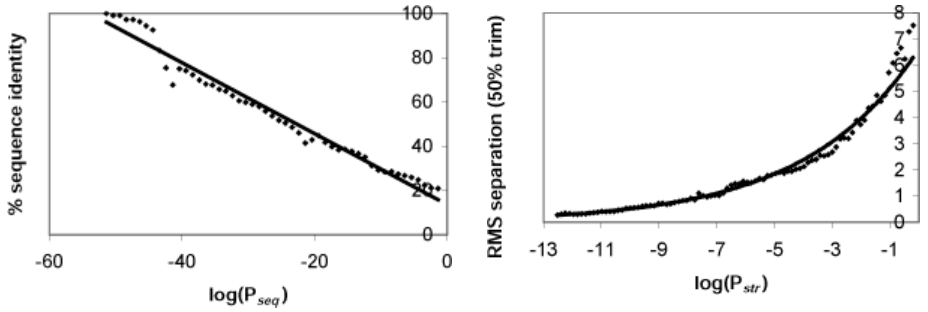
Modern statistical language

Not in TZ

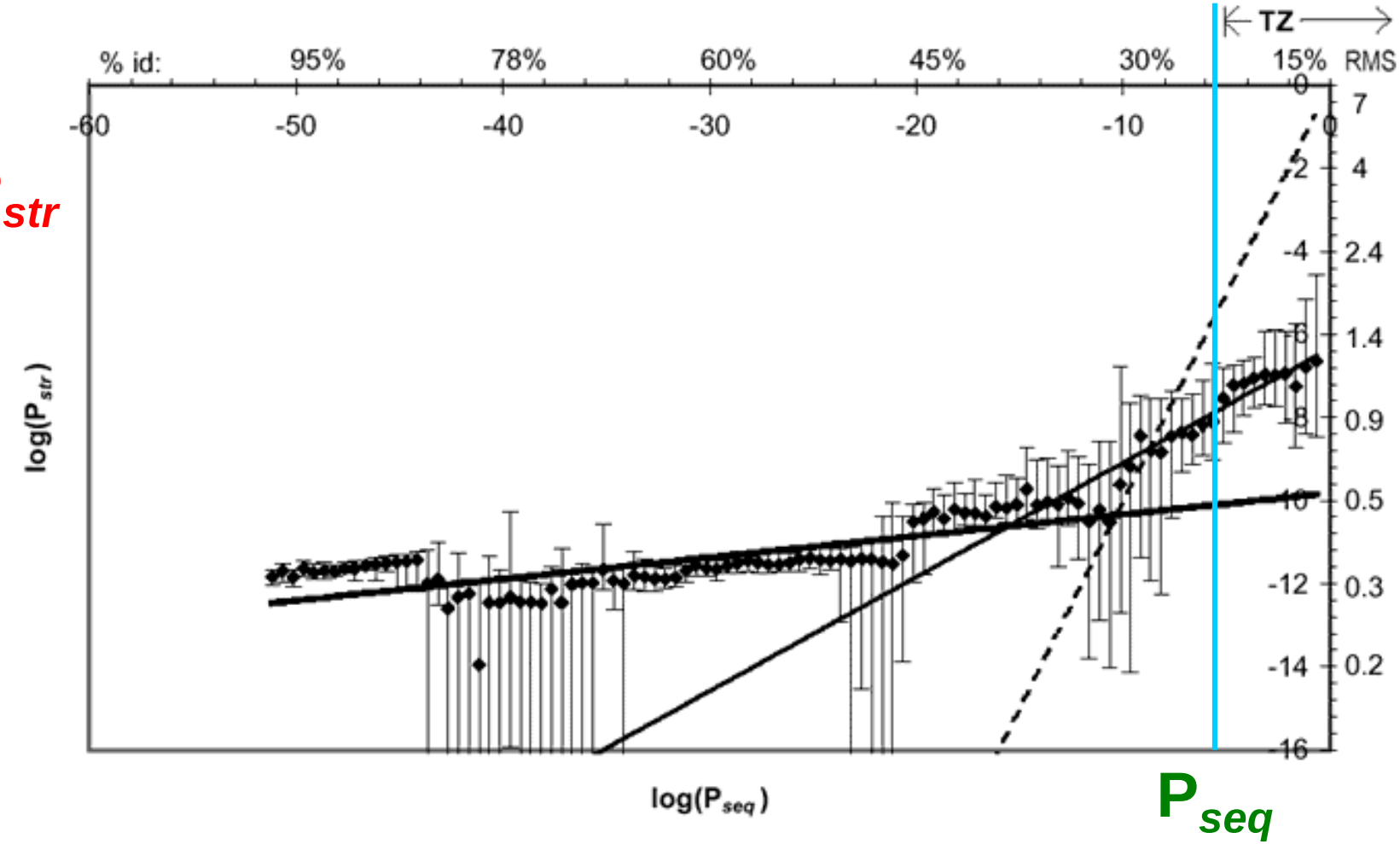
$P_{str}=10^{-10}P_{seq}^{.05}$

in TZ

$P_{str}=10^{-6}P_{seq}^{.274}$



overcomes length dependency



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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

W Krebs, J Tsai, M Levitt, C Wilson, R Das, H Hegyi, J Lin, Y Kluger, C Arrowsmith, A Edwards, L Regan, S Balasubramanian, A Drawid, D Greenbaum, M Snyder, R Jansen

bioinfo.mbb.yale.edu

Structures ("Classic")

(now) Structural Genomics

(now) Func. Genomics

Arrays (future)

Structures ("Classic")

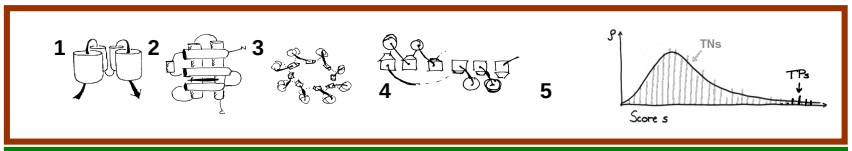
1 Fold Library (A parts list.) Structural Alignment, EVD P-value, Seq. Struc. diverg.

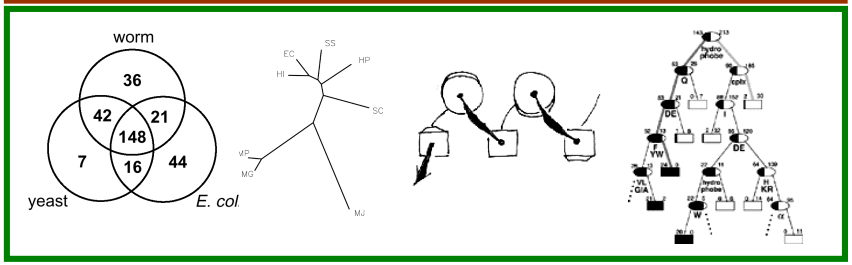
2 Folds in Genomes (Shared, common, and/or unique parts?) Known Folds. Fold Tree, Top-10. βαβ. Biases. MG fold assignment extent. MG Target Selection, MT retrospective decision tree.

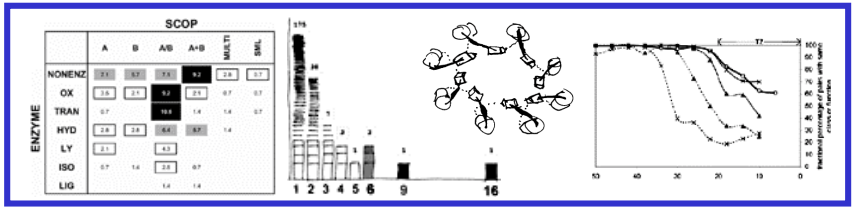
3 Folds & Functions (Roles/part?) How many folds /function? Mostly 1, but TIM versatile. Seq. diverg. vs. Func. diverg.

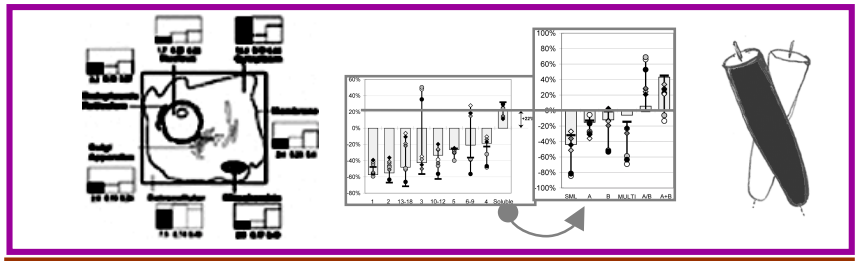
4 Folds in the Transcriptome (Common parts? Where are parts?) Enriched ↑ : VGA, TIM, αβ folds, energy, synthesis, cyt. Depleted ↓ : NS, long, TM folds, transport, transcription, Leu-zip, nuc. Bayesian Localizer, phenotypes clustering

5 Fold Flexibilty (How adaptable is a part?). Motions DB, morph server, interface packing, Voronoi Volumes



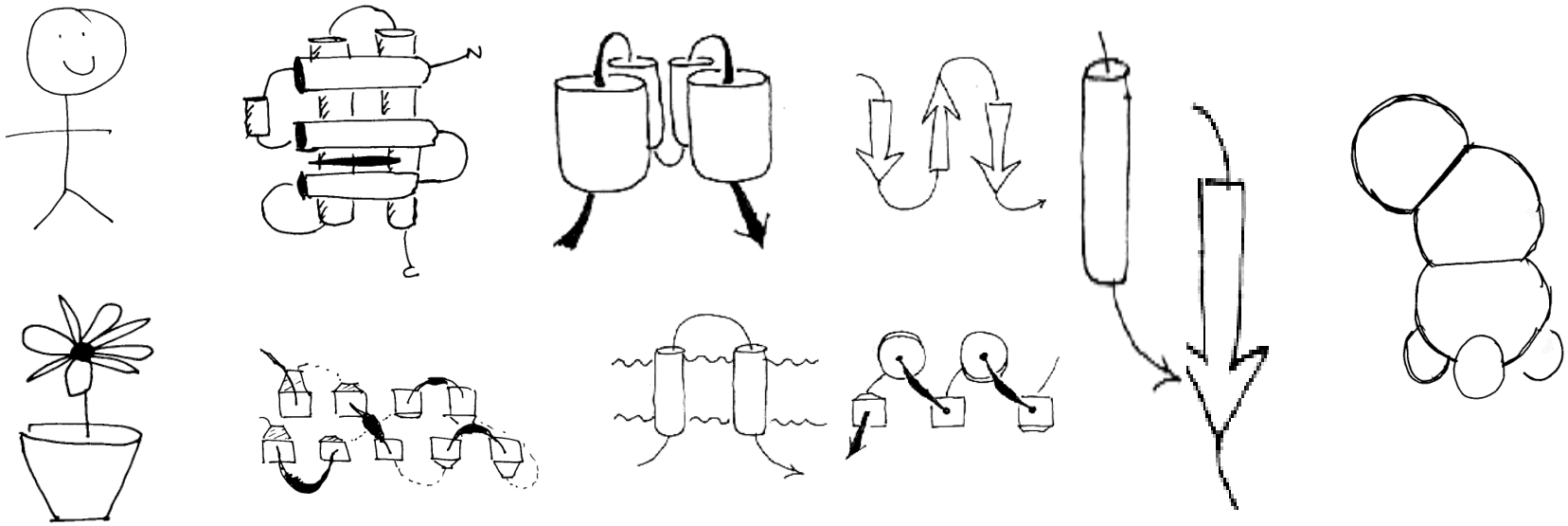






At What Structural Resolution Are Organisms Different?

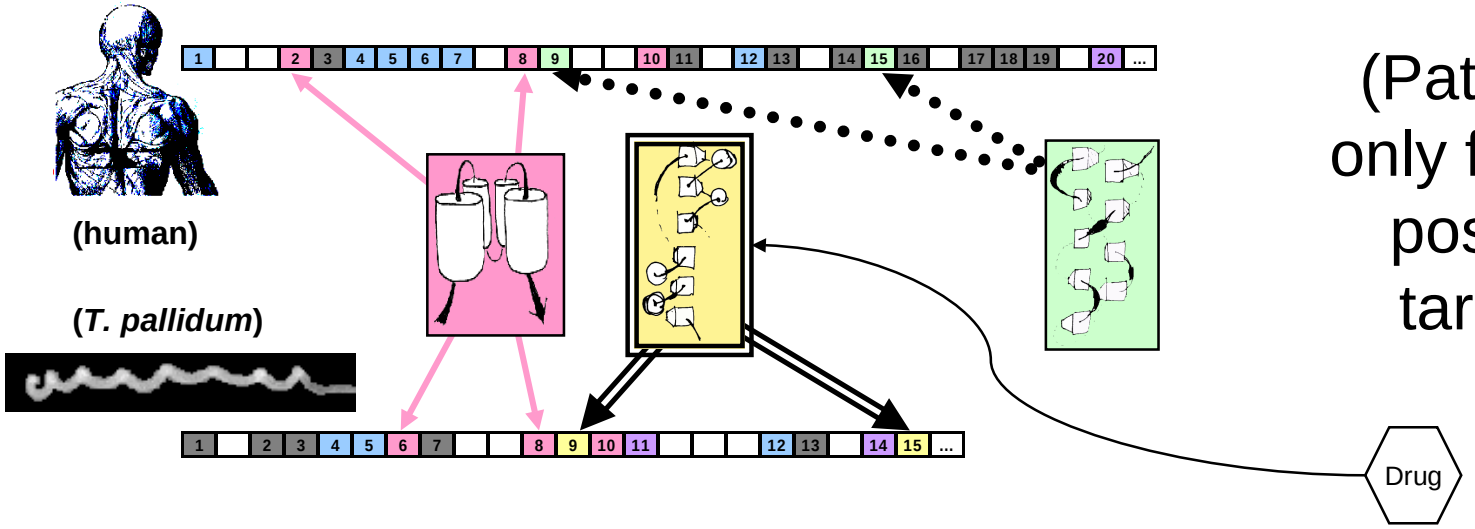
person plant	protein fold (lg)	super-secondary structure (ββ,TM–TM, αβαβ,ααα)	helix strand	individual atom (C,H,O...)



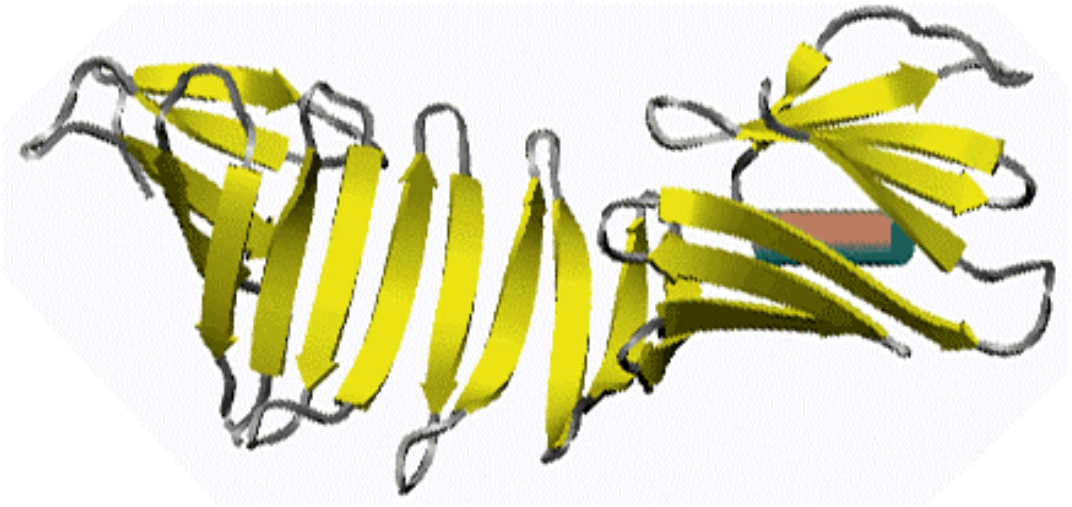
1m	100Å	10Å	1Å

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Practical Relevance of Structural Genomics



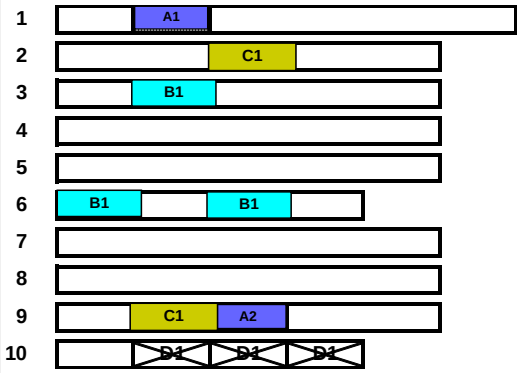
(Pathogen only folds as possible targets)



- OspA protein
  - ◊ in Lyme-disease spirochete *B. burgdorferi*
  - ◊ previously identified as the antigen for vaccine
  - ◊ has novel fold (C Lawson)

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Cross-Reference: Folds→Sequences → Organisms

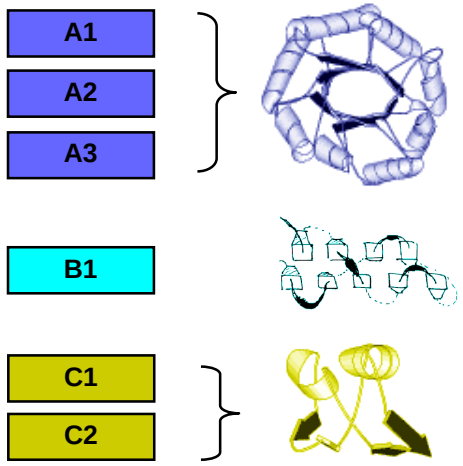


Match sequence representatives using fasta,blast, Ψ-blast, or other "fold assignment" methods

#2

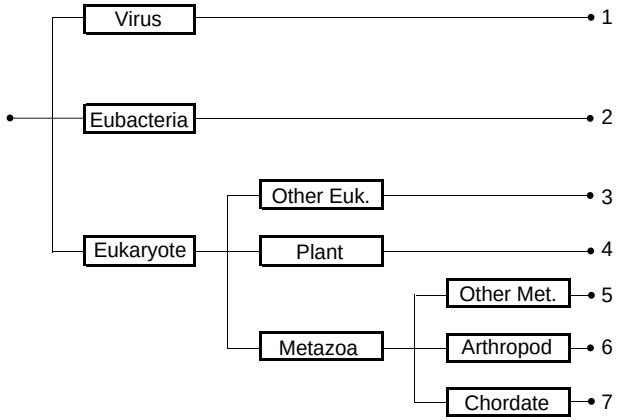
Structures in fold families with repre-sentive seq-uences (scop)

#1



Organize Sequences by Genome or Taxon

#3



3 + 5

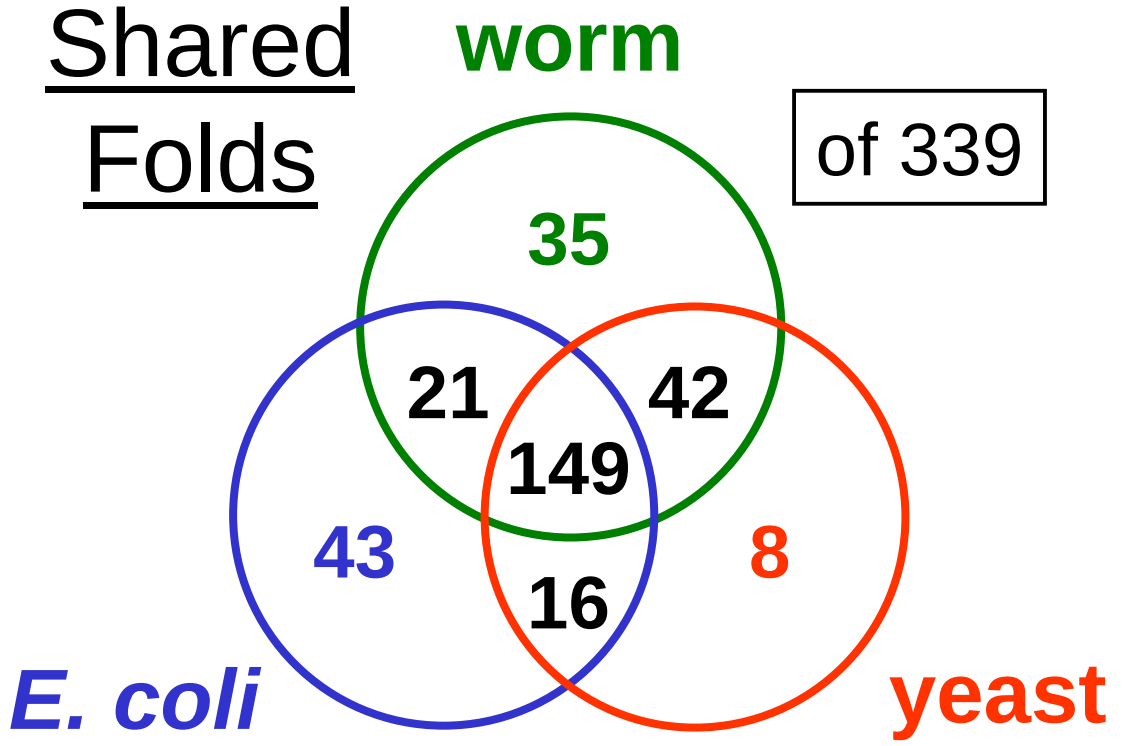
Abbrev.	Kingdom (subgroup)	Genome	Num. ORFs	Reference
EC	Bacteria (gram negative)	*Escherichia coli*	**4290**	Blattner et al.
HI	Bacteria (gram negative)	*Haemophilus influenzae*	**1680**	TIGR
HP	Bacteria (gram negative)	*Helicobacter pylori*	**1577**	TIGR
MG	Bacteria (gram positive)	*Mycoplasma genitalium*	**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.

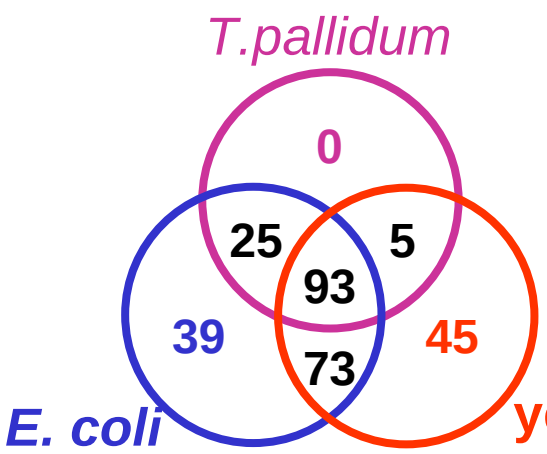
Tabulate Results in Database

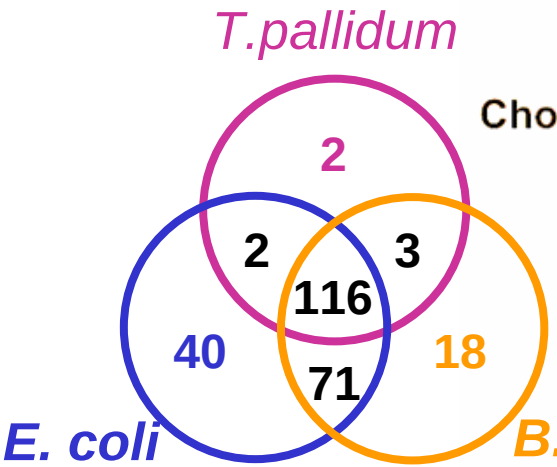
#4

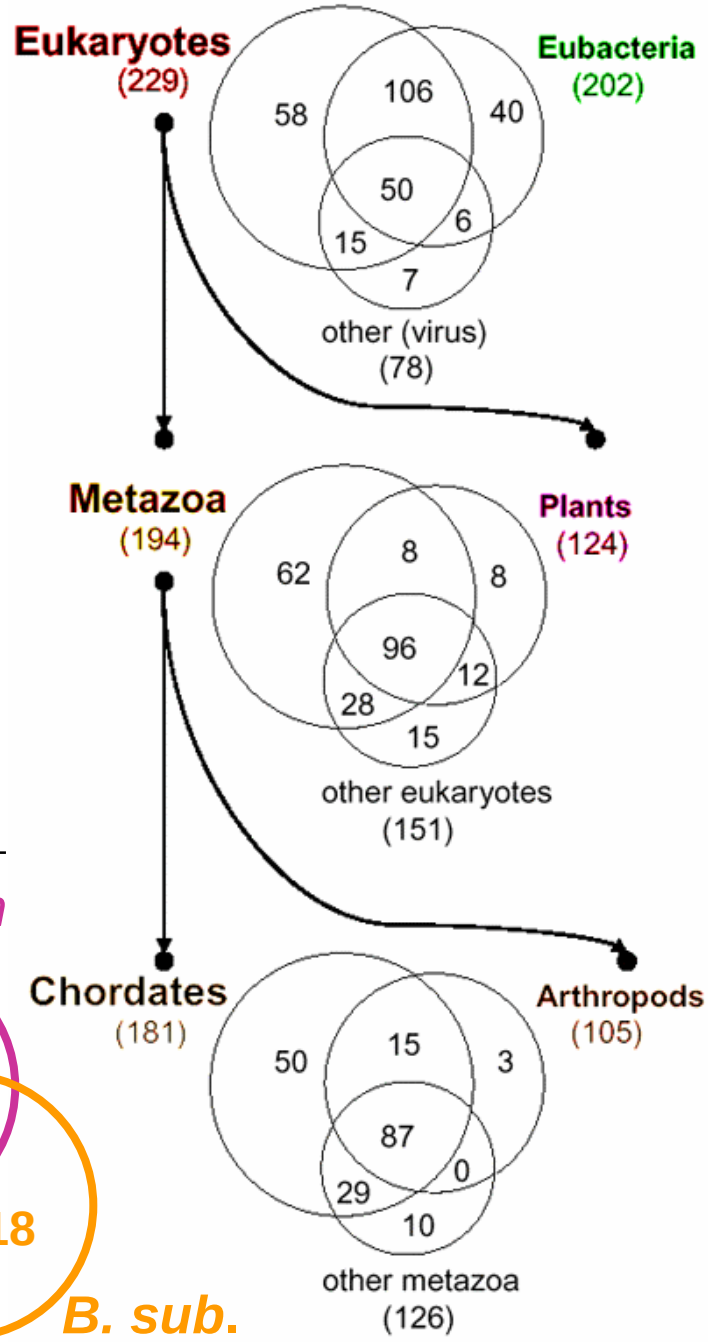
class	Fold#	EC	SC	HI	SS	HP	MJ	MP	MG	total	Fam.	PDB	Rep. Struc.	Name
α/β	18	60	46	23	40	19	7	4	3	202	16	183	1xel -	NAD(P)-binding Rossmann Fold
α/β	24	20	69	17	19	17	16	10	11	179	13	132	1gky -	P-loop Containing NTP Hydrolas
α+β	31	37	28	18	16	12	40	3	3	157	23	160	1fxd -	like Ferrodoxin
α/β	01	45	36	13	22	11	10	5	4	146	37	399	1byb -	TIM-barrel
α/β	23	18	17	7	9	4	8	2	2	67	5	36	1pyd a:2-181	Thiamin-binding
α/β	04	15	11	7	10	1	9	5	5	63	13	132	2tmd a:490-645	FAD/NAD(P)-binding
α+β	55	8	9	7	8	9	3	6	6	56	4	23	1sry a:111-421	Class-II-aaRS/Biotin Synthetase
β	27	7	10	8	8	4	4	3	3	47	5	19	1fnb 19-154	Reductase/Elongation Factor Do
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
β	55	9	10	5	5	2	2	2	2	37	7	19	1bdo -	Barrel-sandwich hybrid
α/β	15	5	5	4	4	5	6	3	3	35	3	22	2ts1 1-217	ATP pyrophoshatases

Shared Folds



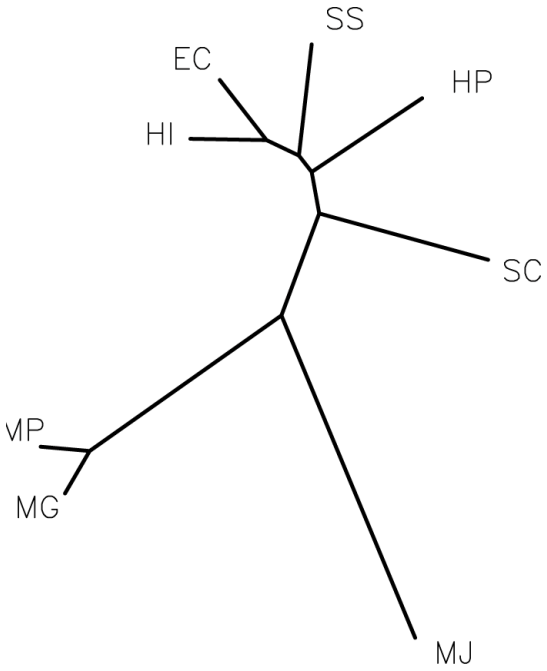




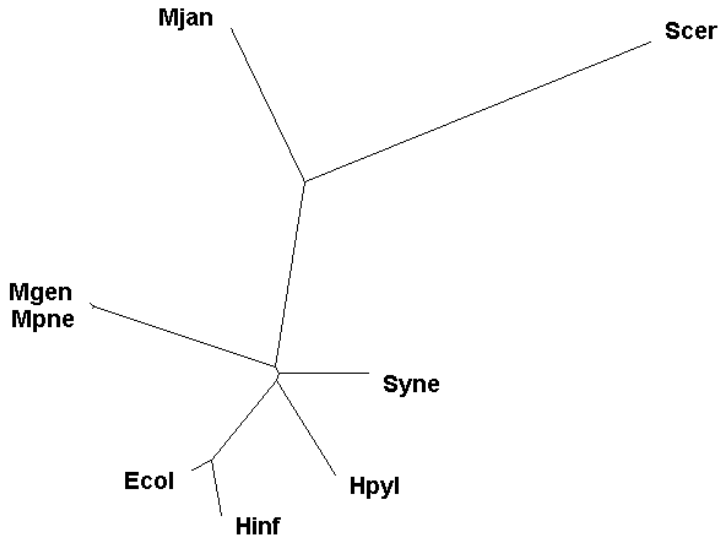


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

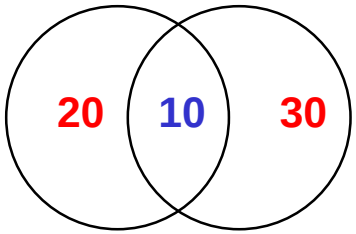
Cluster Trees Grouping Initial Genomes on Basis of Shared Folds



Fold Tree



"Classic" Tree

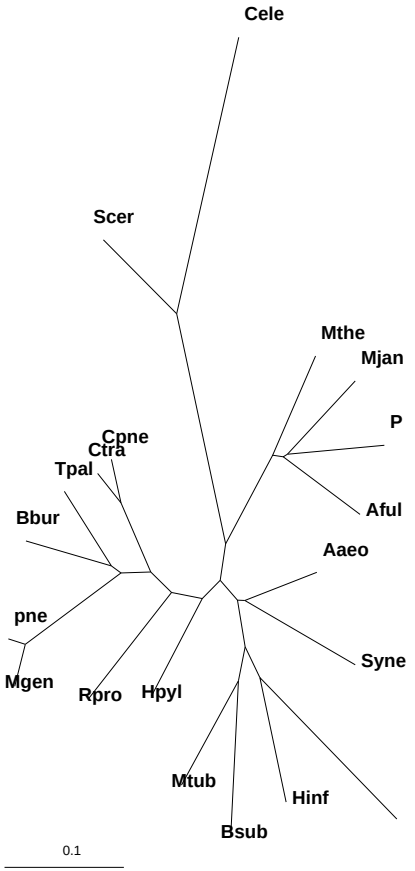


D=10/(20+10+30)

D=S/T **S** = # shared folds

D = shared fold dist. betw. 2 genomes

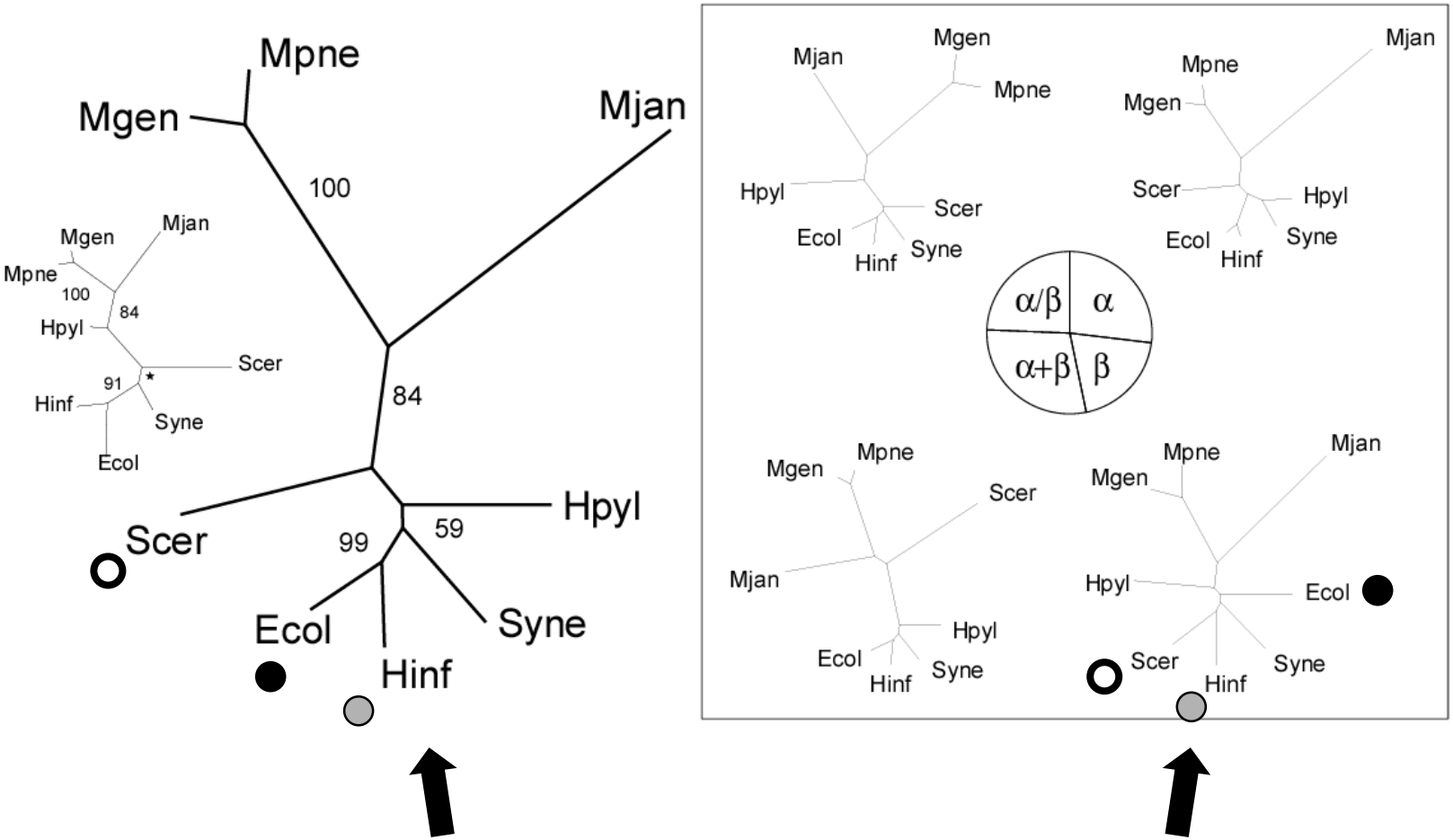
T= total # folds in both



20 Genomes

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Distribution of Folds in Various Classes

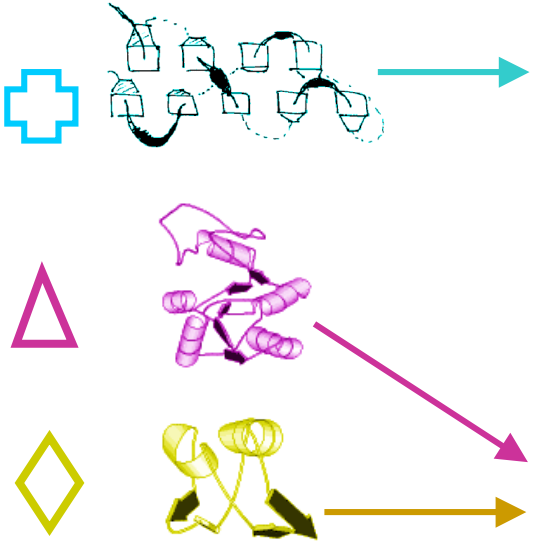


Unusual distribution of all-beta folds

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Common Folds in Genome, Varies Betw. Genomes

Depends on comparison method, DB, sfams v folds, &c (new top superfamilies via ψ-Blast, Intersection of top-10 to get shared and common)



Top-10 Worm Folds	class	num. matches in worm genome (N)	frac. all worm dom. (F)	in EC?	in SC?
Ig	B	830	1.7%		
Knottins	SML	565	1.1%		
Protein kinases (cat. core)	MULT	472	0.9%		
C-type lectin-like	A+B	322	0.6%		
corticoid recep. (DNA-bind dom.)	SML	276	0.5%		
Ligand-bind dom. nuc. receptor	A	257	0.5%		
alpha-alpha superhelix	A	247	0.5%		
C2H2 Zn finger	SML	239	0.5%		
P-loop NTP Hydrolase	A/B	235	0.5%		
Ferrodoxin	A+B	207	0.4%		

Rank	M.gen		B.sub		E.col	
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	M.the		A.ful	
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

Rank	S.cer	
1	P-loop hydrolyase	249
2	Protein kinase	123
3	Rossmann domain	90
4	RNA-binding domain	75
5	SAM methyl-transferase	63
6	Ribonuclease H-like	57
Total ORFs		6218
with Common Superfamilies		560 (9%)

Yeast

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Characteristics of Common, Shared Folds: βαβ structure

42 ARTICLES NATURE VOL. 336 3 NOVEMBER 1988

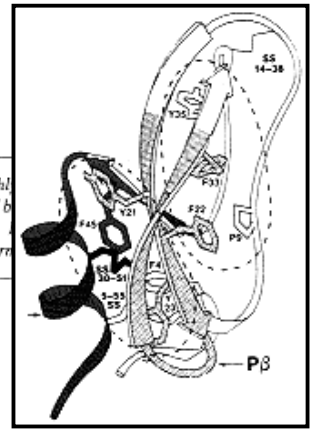
A peptide model of a protein folding intermediate

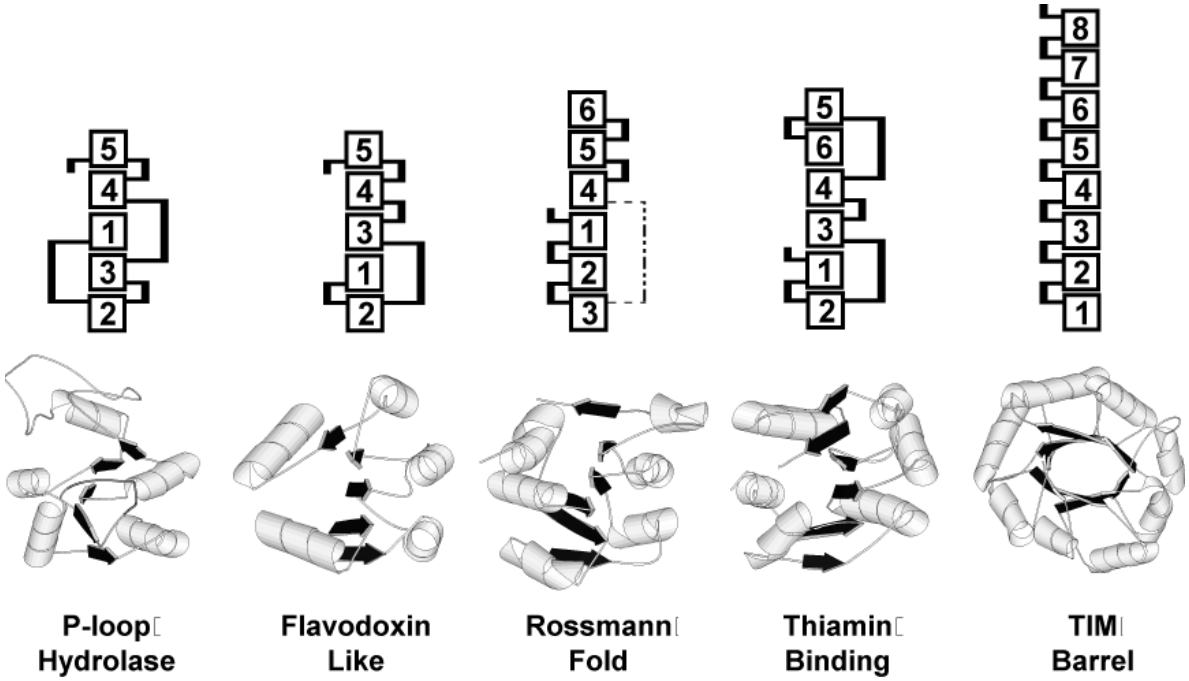
Terrence G. Oas & Peter S. Kim

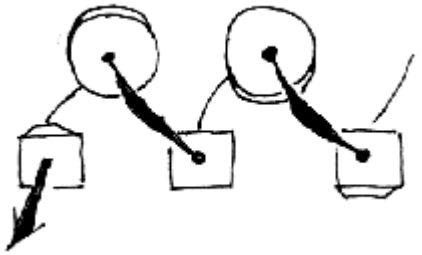
Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

It is difficult to determine the structures of protein folding intermediates because folding is a highl... disulphide-bonded peptide pair, designed to mimic the first crucial intermediate in the folding of b... inhibitor, contains secondary and tertiary structure similar to that found in the native protein. ... circumvent the problem of cooperativity and permit characterization of structures of folding interm...

336: 42







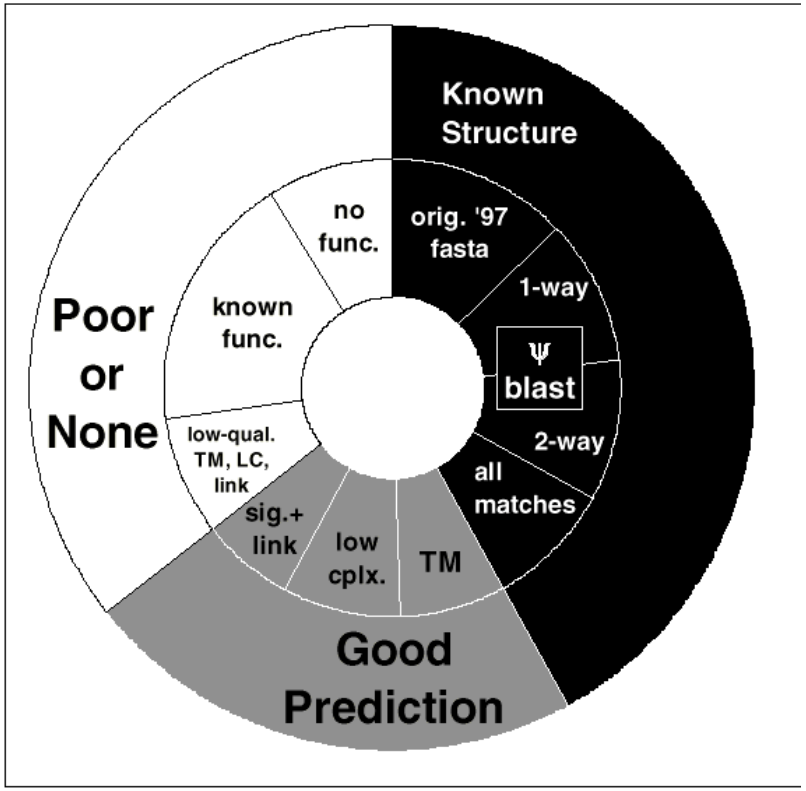
All share α/β structure with repeated R.H. βαβ units connecting adjacent strands or nearly so (18+4+2 of 24)

HI, MJ, SC vs scop 1.32

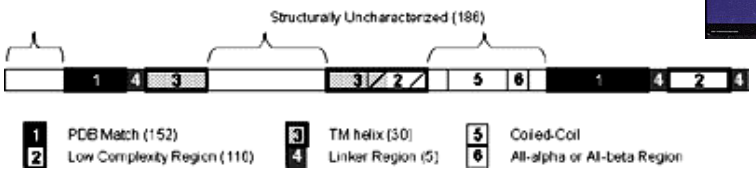
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

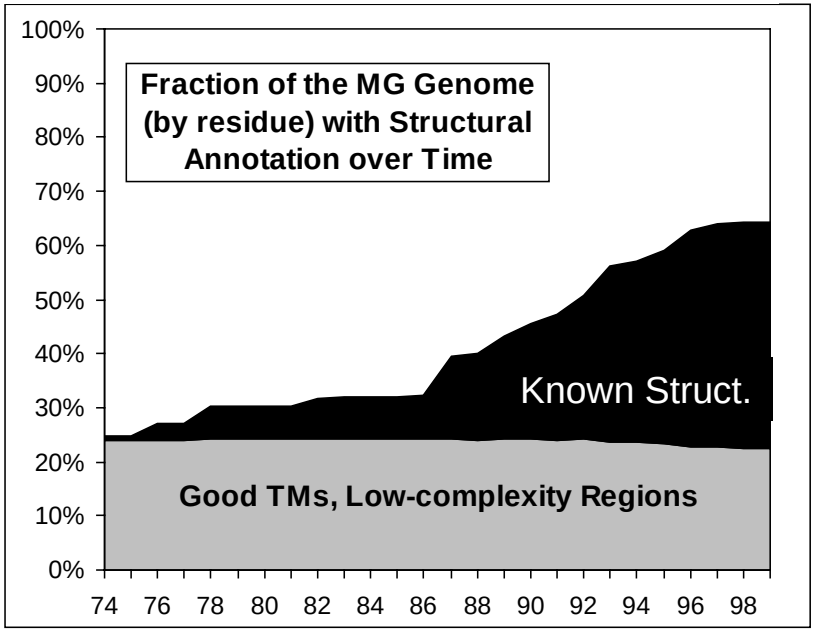
Bias Problem → Prediction, Expts.

- Known Structures are Incomplete, Biased Sample from Genome, so...
  - ◊ Resample
  - ◊ Expt. Structural Genomics...
  - ◊ Predict Structures...



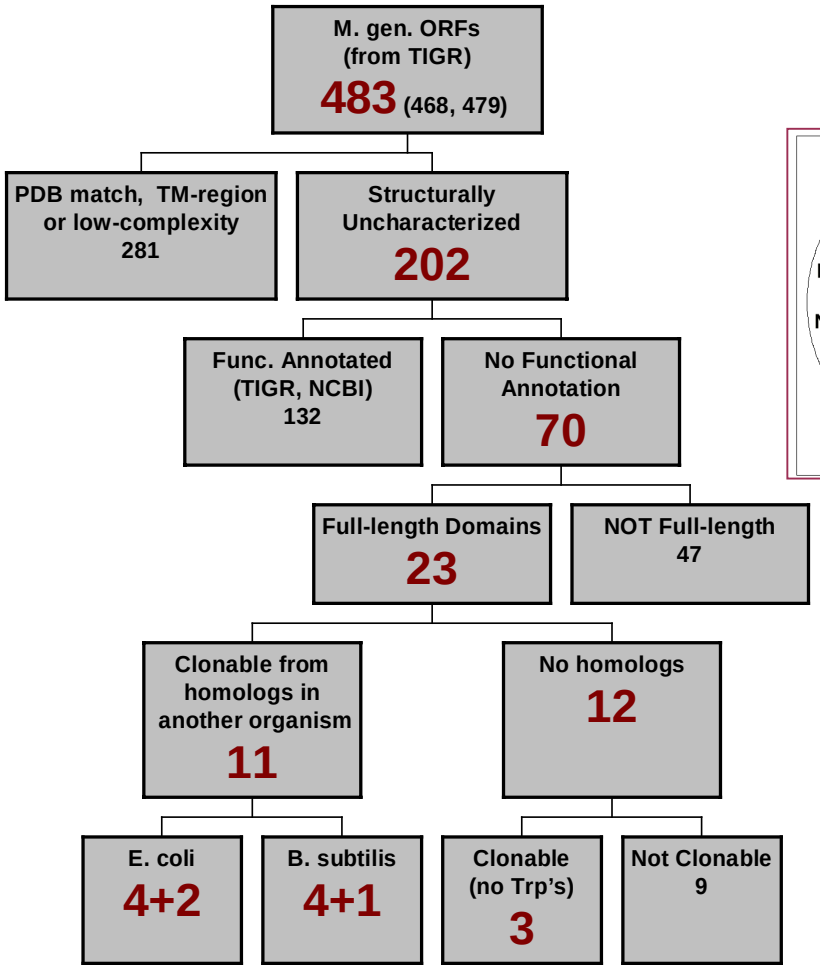
Same Sampling Issues with US Census!!

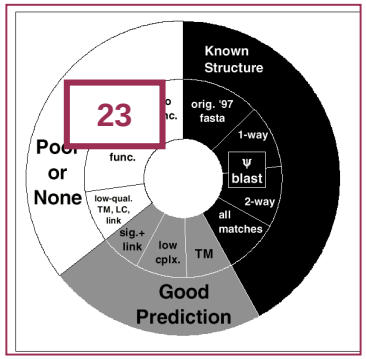


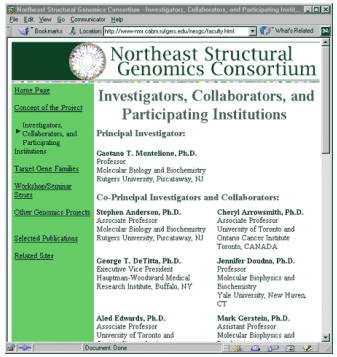


Finding Unusual Proteins for Expt. Structural Genomics

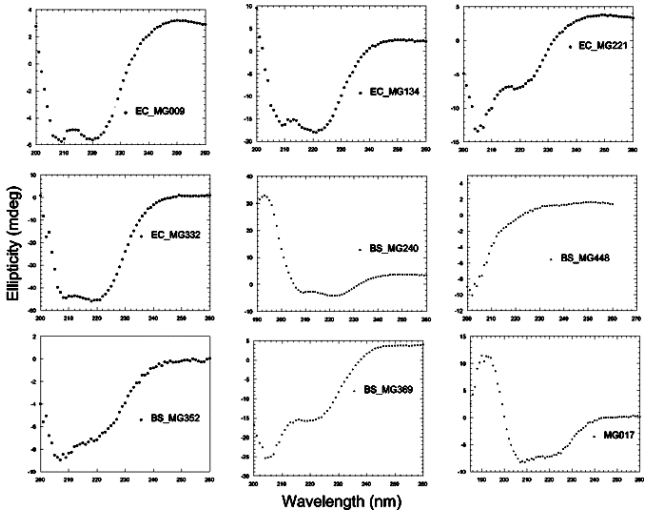
- **Prospective Target Selection**
- Identify Proteins in *M. genitalium* that are most atypical structurally (hardest)
- Characterize biophysically by CD (do they fold normally?)





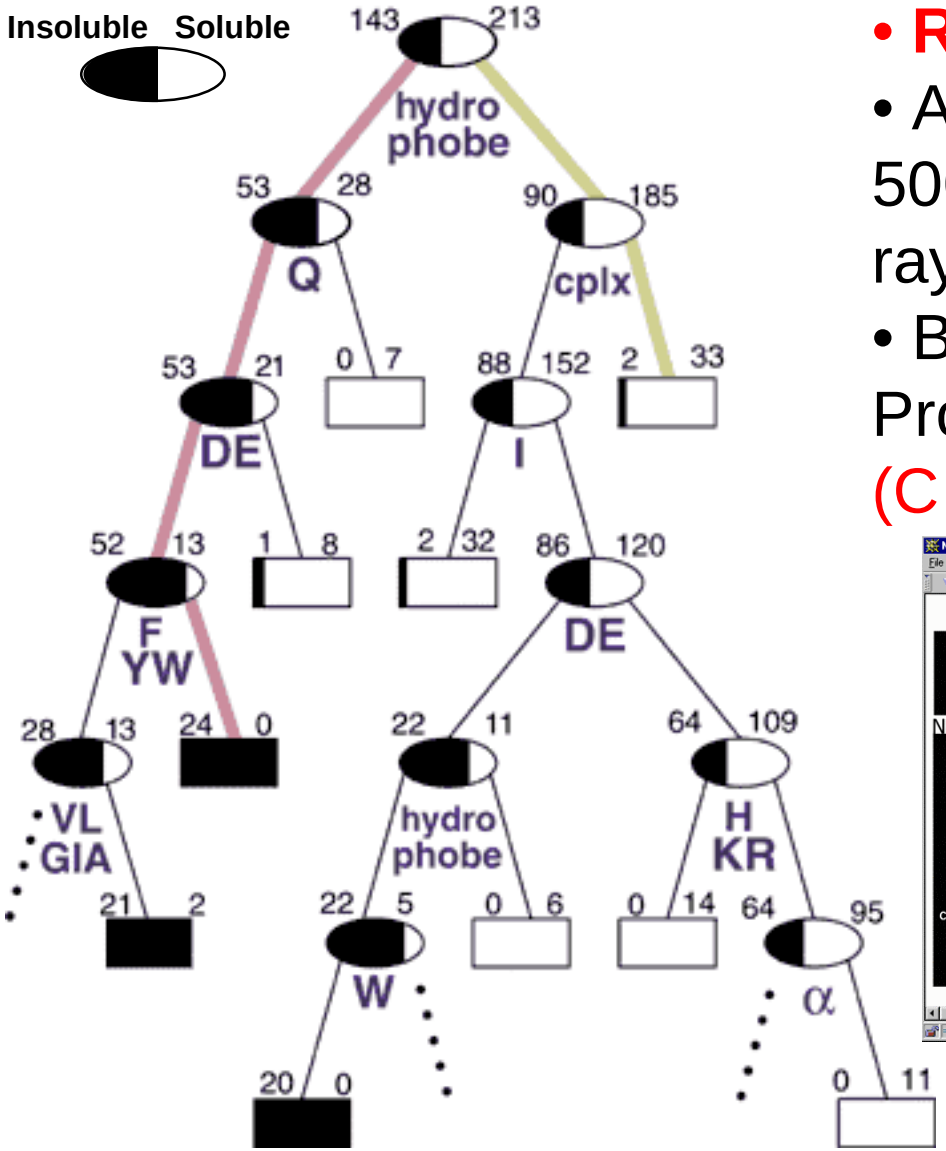


In collaboration with L Regan, S Balasubramanian

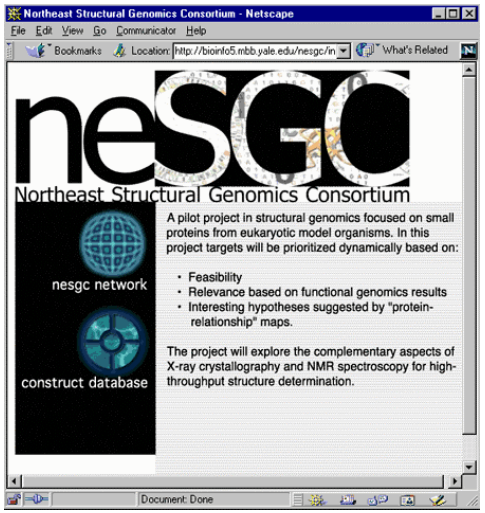


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Characterizing the Low-hanging Fruit for Experimental Structural Genomics



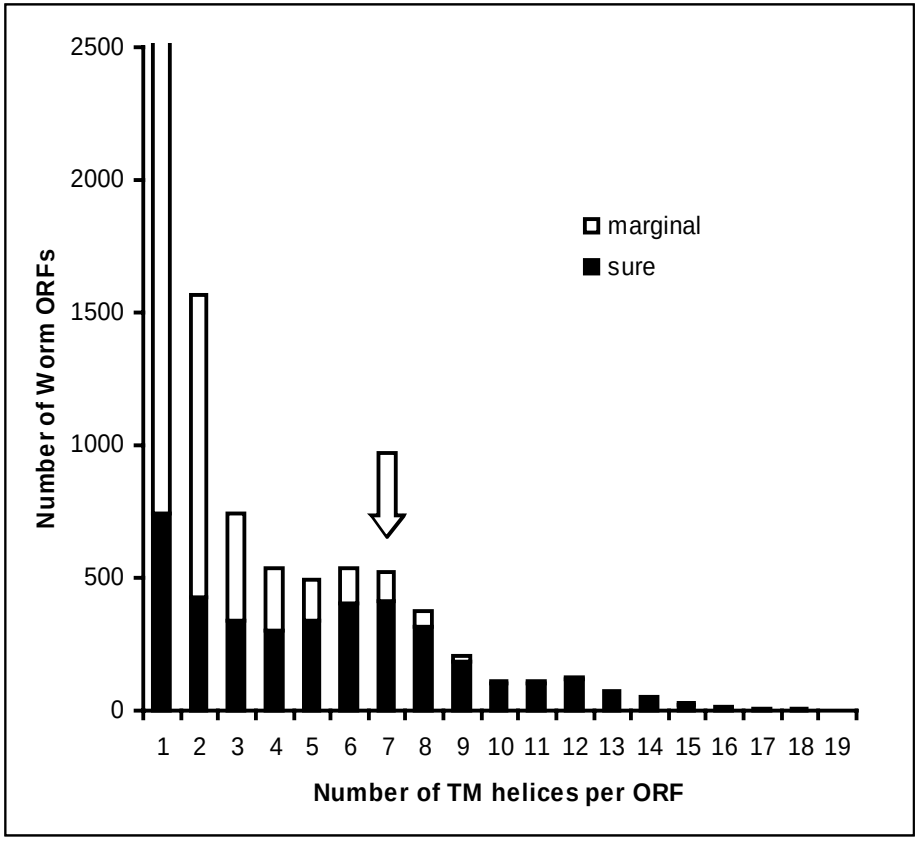
- **Retrospective Decision-Tree**
- Analysis of the Suitability of 500 M. thermo. proteins for X-ray/NMR work
- Based on results of Toronto Proteomics Group (C Arrowsmith, A Edwards)

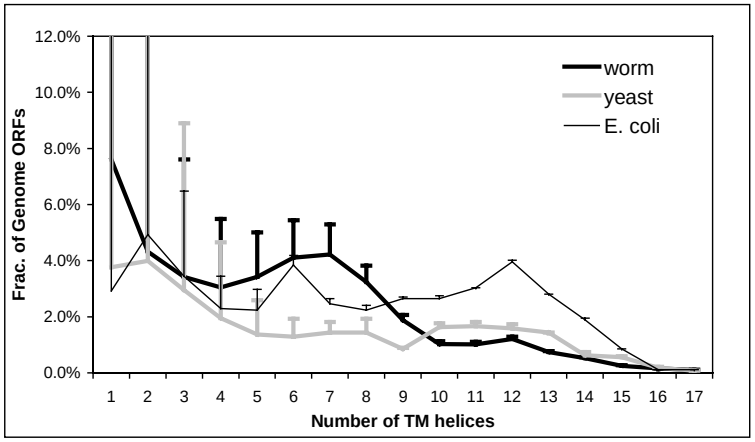


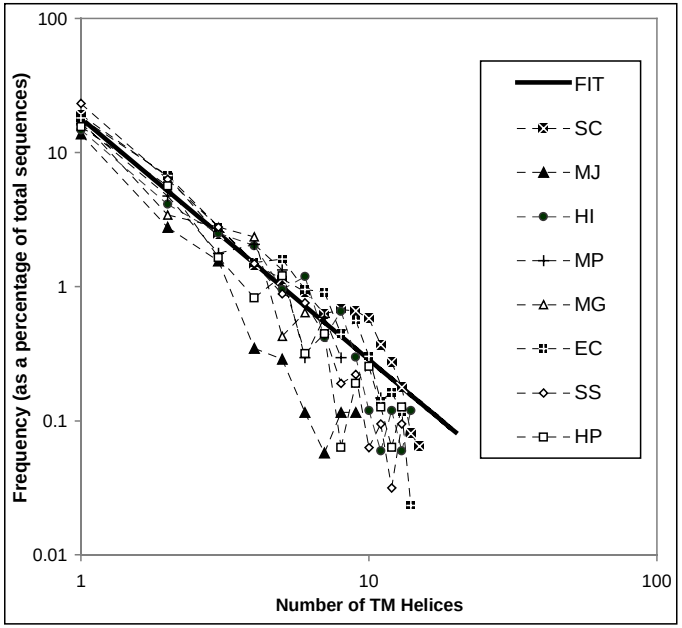
For example, proteins that fulfill the following sequence of four rules are likely to be insoluble: (1) have a hydrophobic stretch -- a long region (>20 residues) with average hydrophobicity less than -0.85 kcal/mole (on the GES scale); (2) Gln composition <4%; (3) Asp+Glu composition <17%; and (4) aromatic composition >7.5%. Conversely, proteins that do not have a hydrophobic stretch and have less than 27% of their residues in "low-complexity" regions are very likely to be soluble.

TM-helix "prediction"

- TM prediction (KD, GES). Count number with 2 peaks, 3 peaks, &c.
- Similar conclusions to others: von Heijne, Rost, Jones, &c.
- Divide Predictions into sure and marginal (Boyd & Beckwith's criteria)



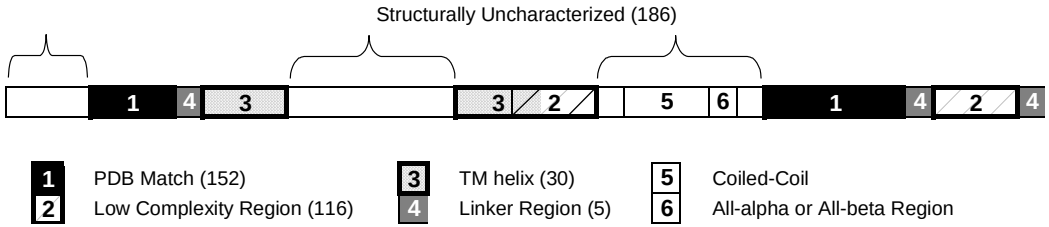




(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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……

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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			

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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

W Krebs, J Tsai, M Levitt, C Wilson, R Das, H Hegyi, J Lin, Y Kluger, C Arrowsmith, A Edwards, L Regan, S Balasubramanian, A Drawid, D Greenbaum, M Snyder, R Jansen

bioinfo.mbb.yale.edu

Structures ("Classic")

(now) Structural Genomics

(now) Func. Genomics

Arrays (future)

Structures ("Classic")

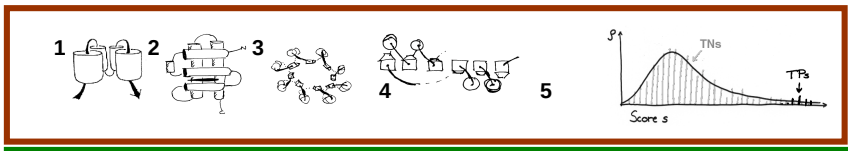
1 Fold Library (A parts list.) Structural Alignment, EVD P-value, Seq. Struc. diverg.

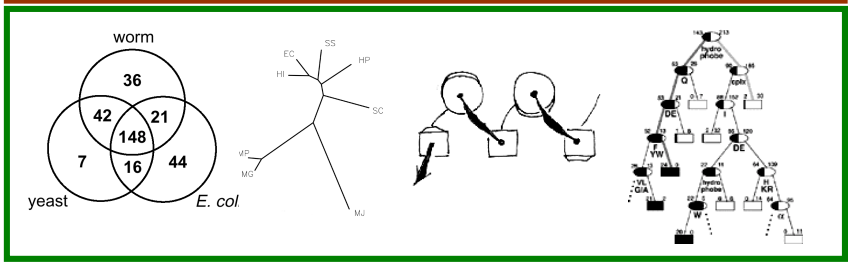
2 Folds in Genomes (Shared, common, and/or unique parts?) Known Folds. Fold Tree, Top-10. βαβ. Biases. MG fold assignment extent. MG Target Selection, MT retrospective decision tree.

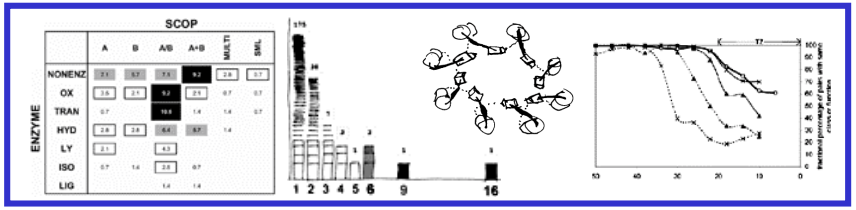
3 Folds & Functions (Roles/part?) How many folds /function? Mostly 1, but TIM versatile. Seq. diverg. vs. Func. diverg.

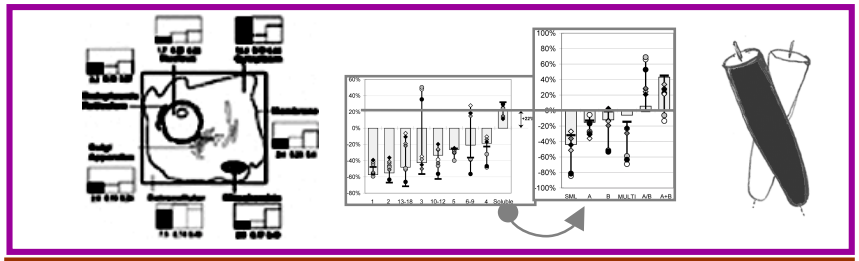
4 Folds in the Transcriptome (Common parts? Where are parts?) Enriched ↑ : VGA, TIM, αβ folds, energy, synthesis, cyt. Depleted ↓ : NS, long, TM folds, transport, transcription, Leu-zip, nuc. Bayesian Localizer, phenotypes clustering

5 Fold Flexibilty (How adaptable is a part?). Motions DB, morph server, interface packing, Voronoi Volumes

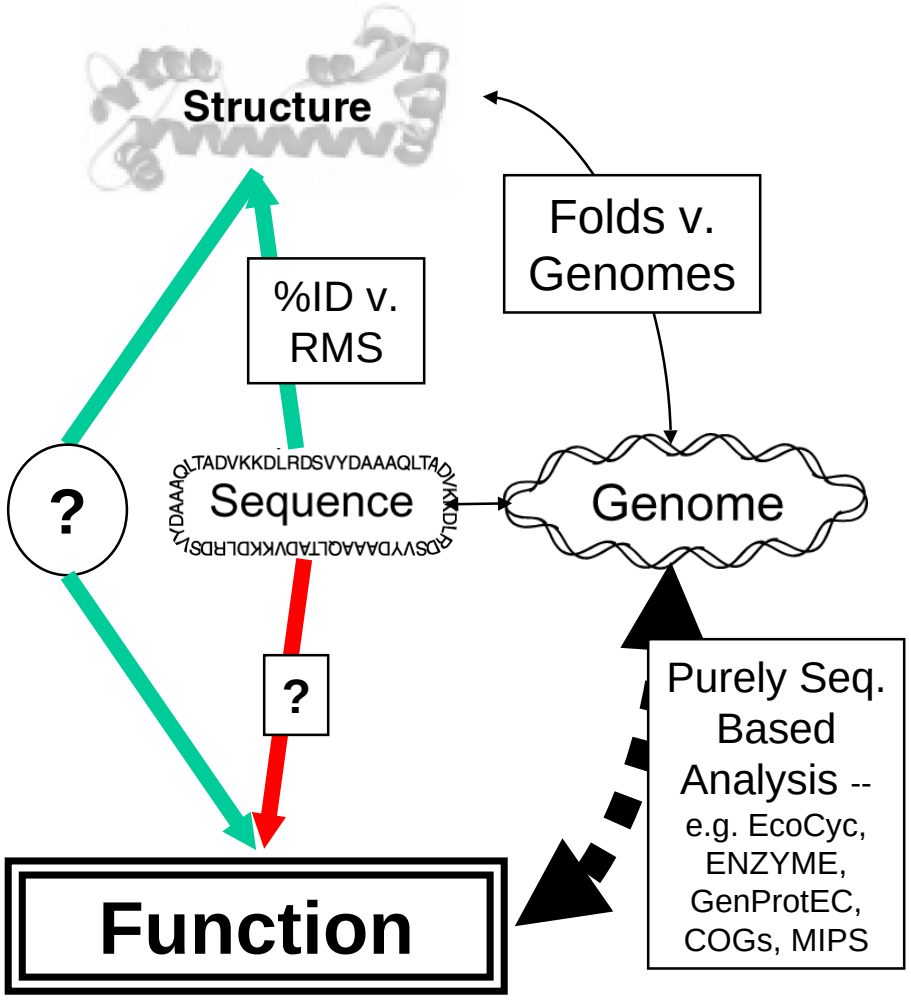






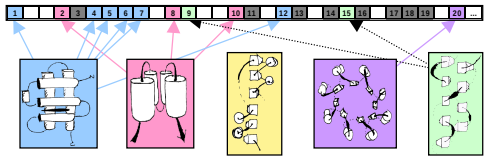


Adding Structure to Functional Genomics, Function to Structural Genomics

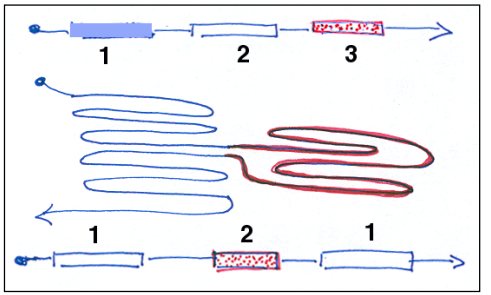


Why Structure? Do we really need it?

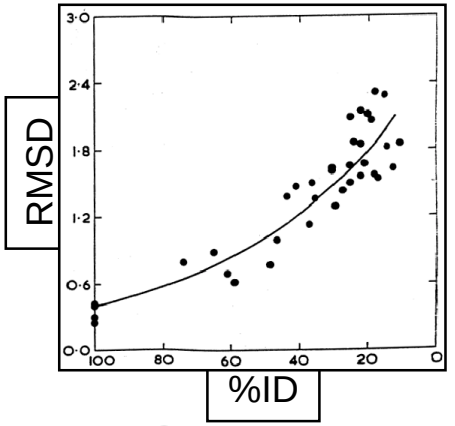
1 Most Highly Conserved



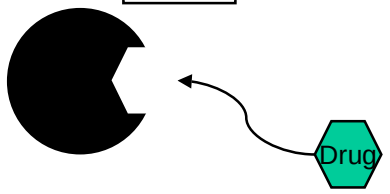
2 Precisely Defined Modules



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

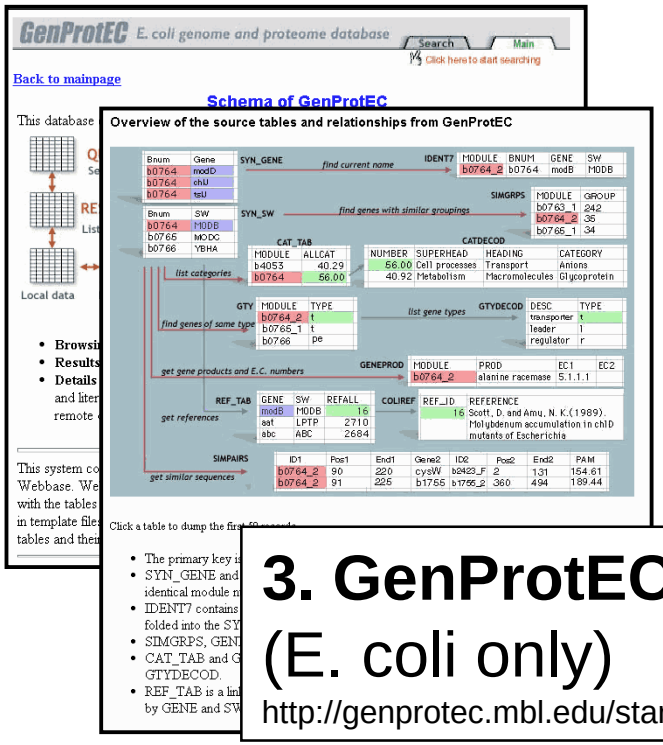


4 Link to Chemistry, Drugs

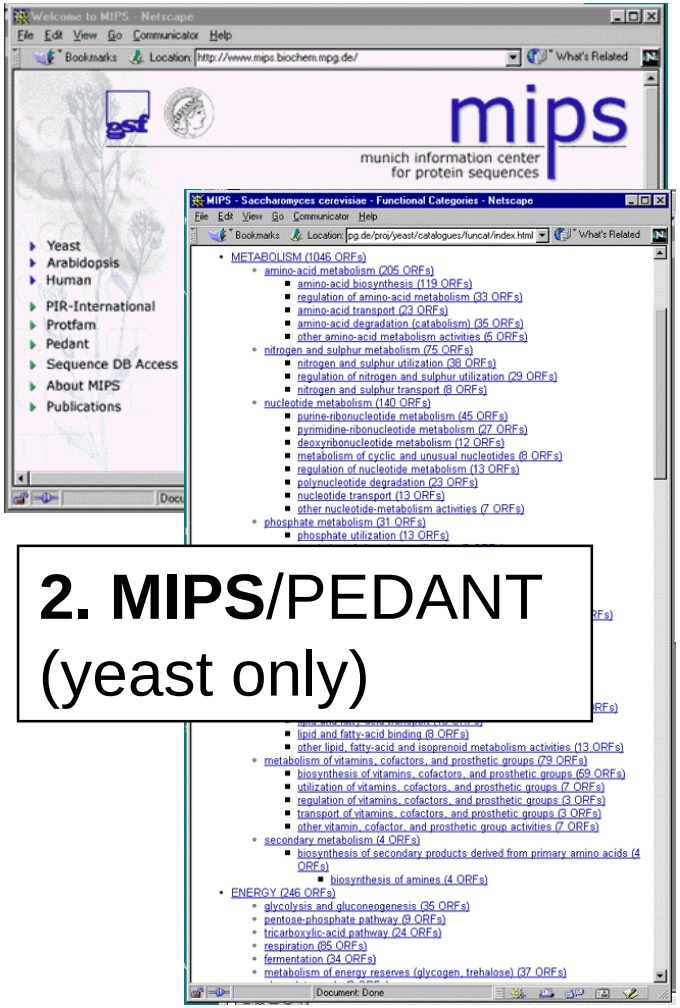


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Functional Classification



3. GenProtEC
(E. coli only)
http://genprotec.mbl.edu/start



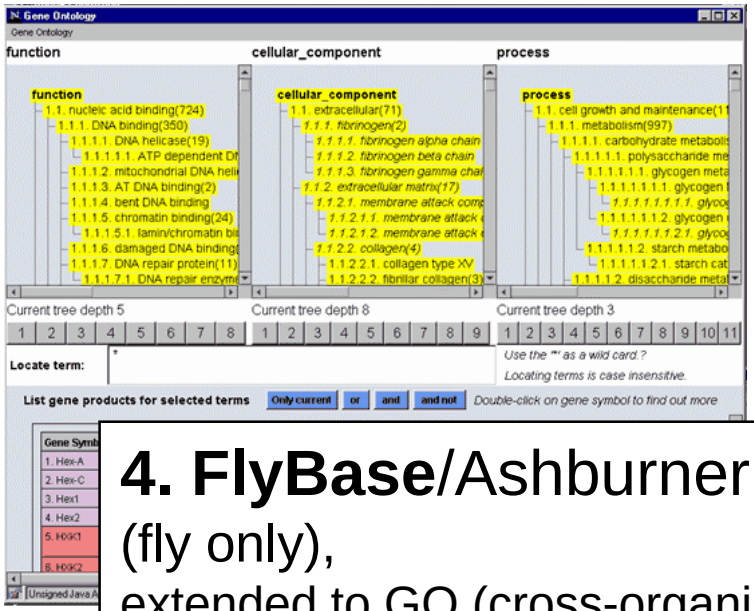
2. MIPS/PEDANT
(yeast only)

Also:

COGs
(cross
organism,
conserved)

WIT,
KEGG
(pathways)

TIGR,
EGAD

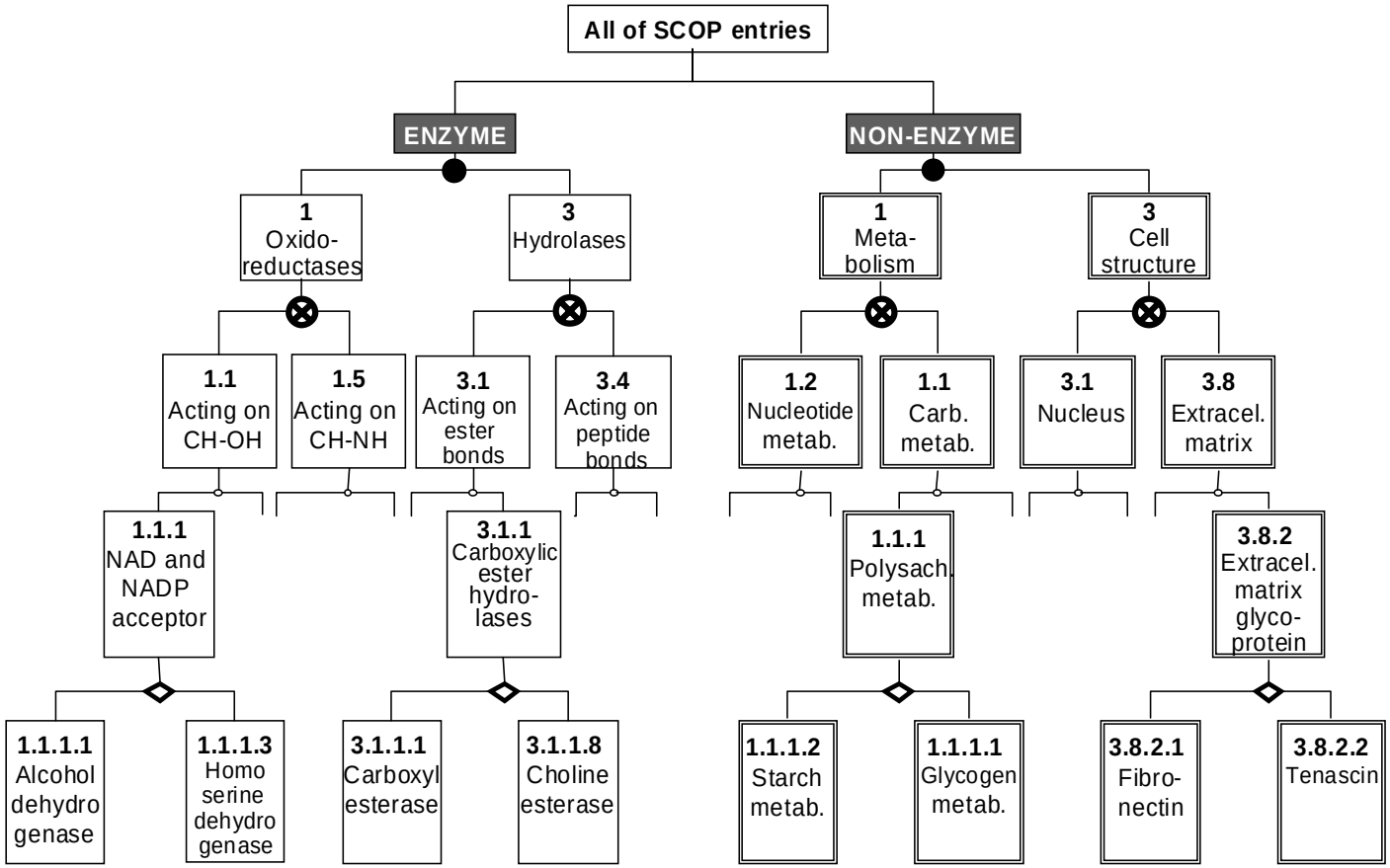


4. FlyBase/Ashburner
(fly only),
extended to GO (cross-organism)

1. ENZYME (cross-organism,
but just enzymes)

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

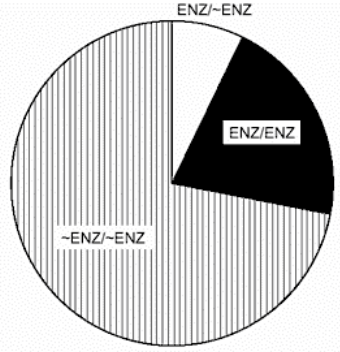
A Simple Scheme for Functionally Classifying Protein Structures



Focus on Pairs with Precise (1.1.1.*) and Broad (1.*) Similarity

A Combined Scheme, merging ENZYME + FLYBASE, with manual additions for proteins not in either (e.g. Ig's)

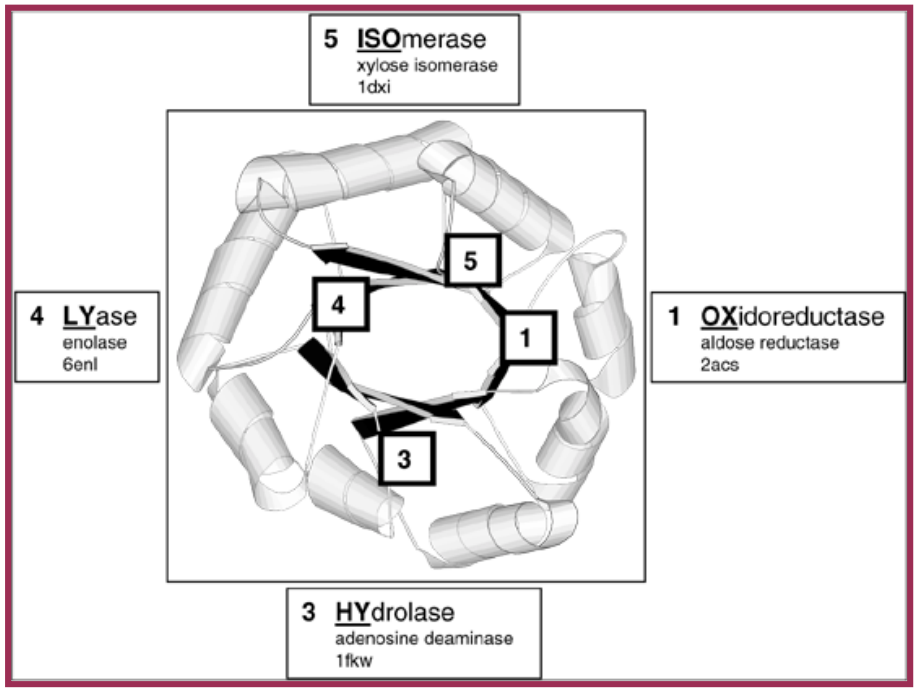
Also: MIPS, GenProtEC, GOGs



Fold-Function Combinations

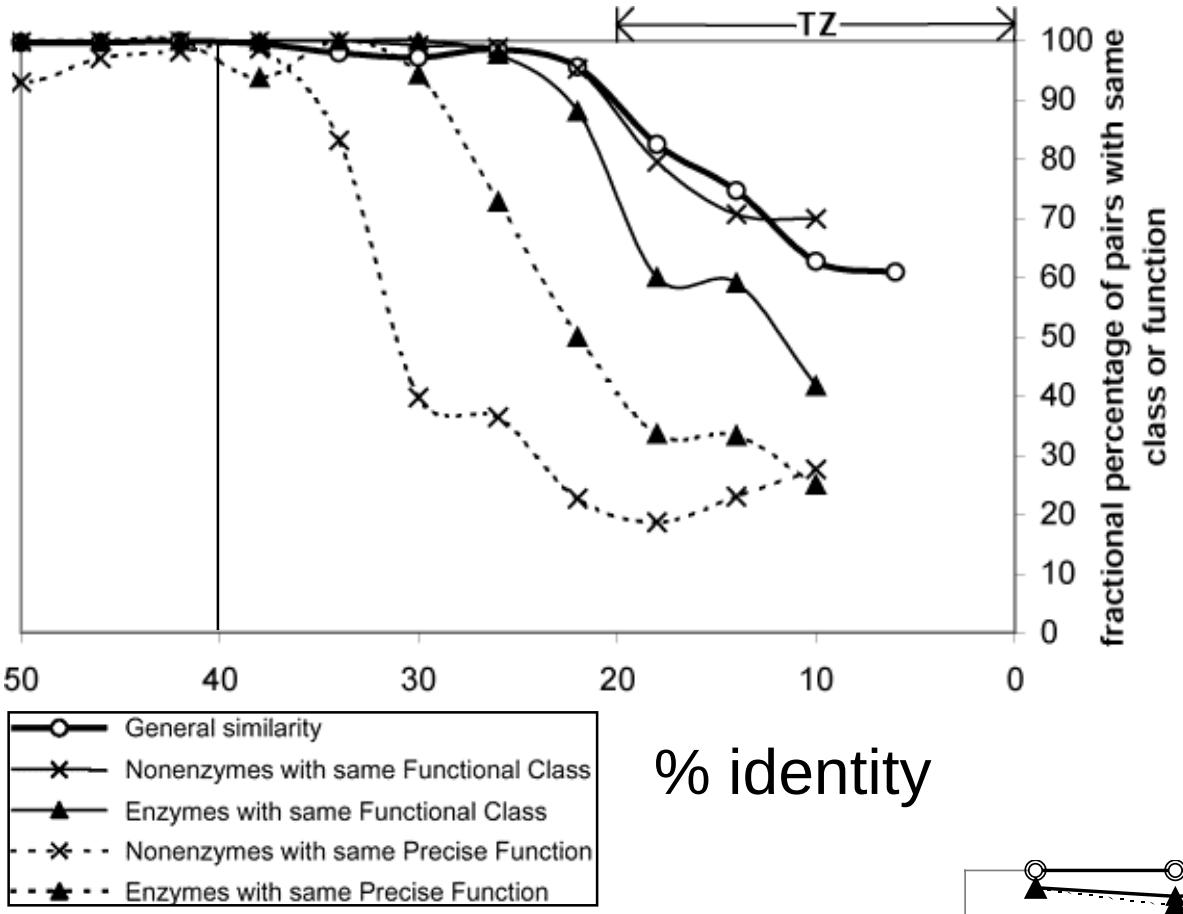
Many Functions on the Same Fold
-- e.g. the TIM-barrel

Have the sequences diverged beyond point of homology? Or does this occur at a certain sequence threshold?



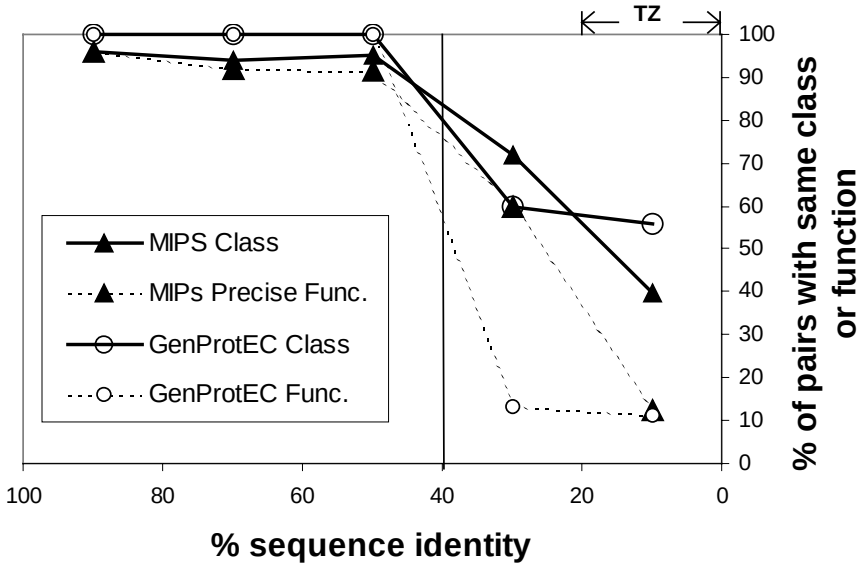
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Relationship of Similarity in Sequence & Structure to that in Function



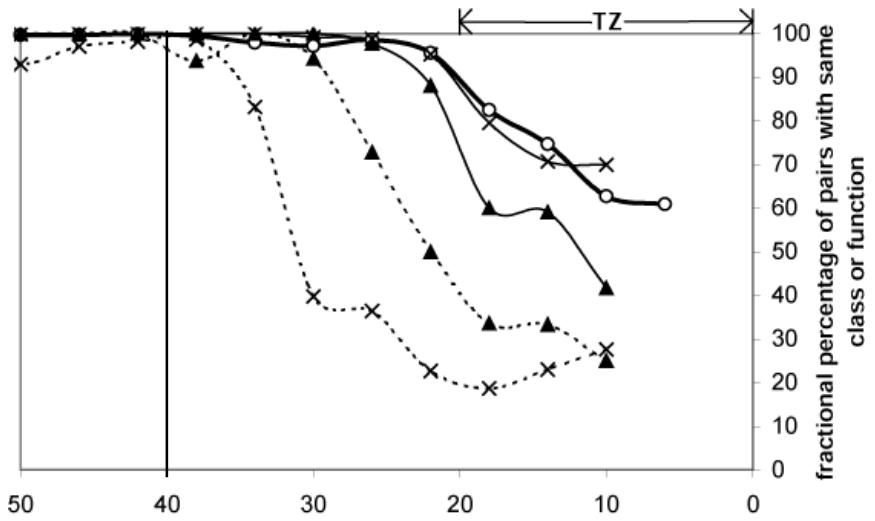
% identity

See at what %ID have diff. function (both broad & precise). Use 4 func. classifications -- ENZYME, FLYBASE (+extra), MIPS, GenProtEC



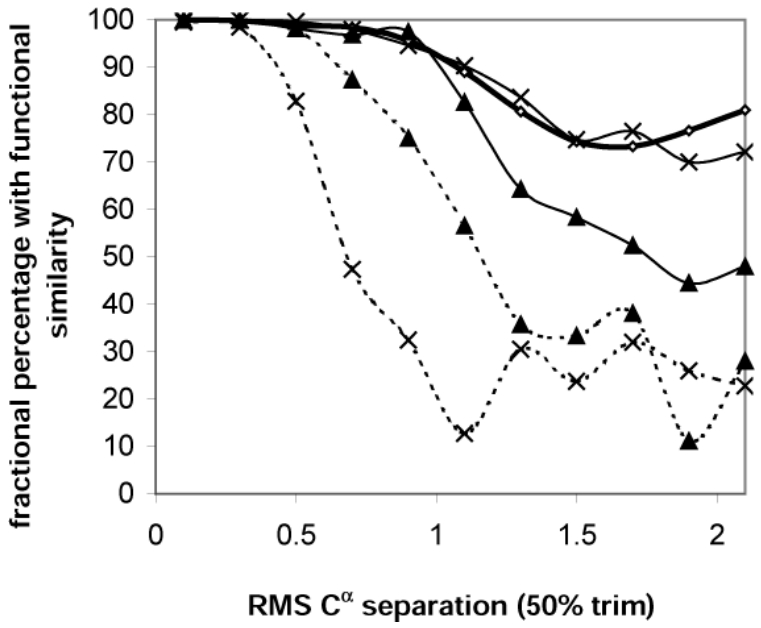
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

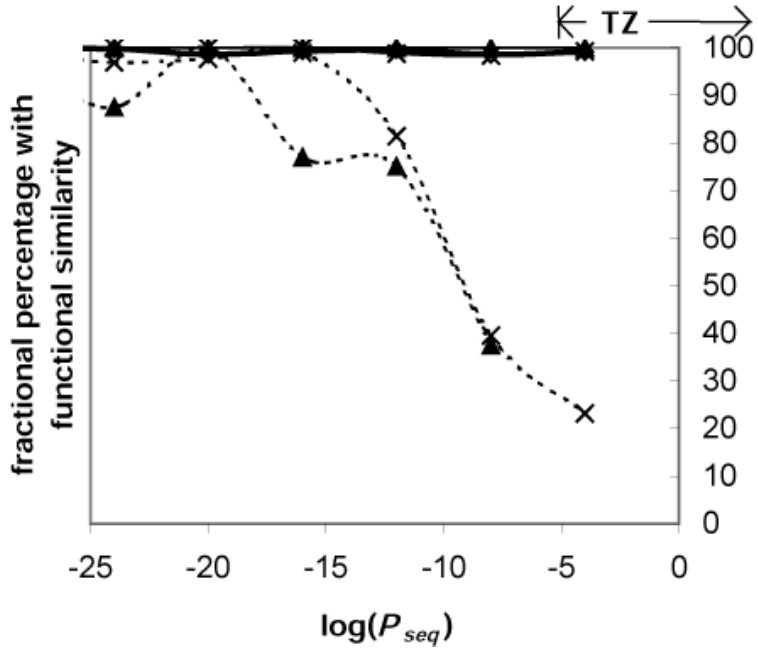
Relationship of Similarity in Sequence & Structure to that in Function II



% identity

Percent identity quite successful vs. structure sim. or statistical scores

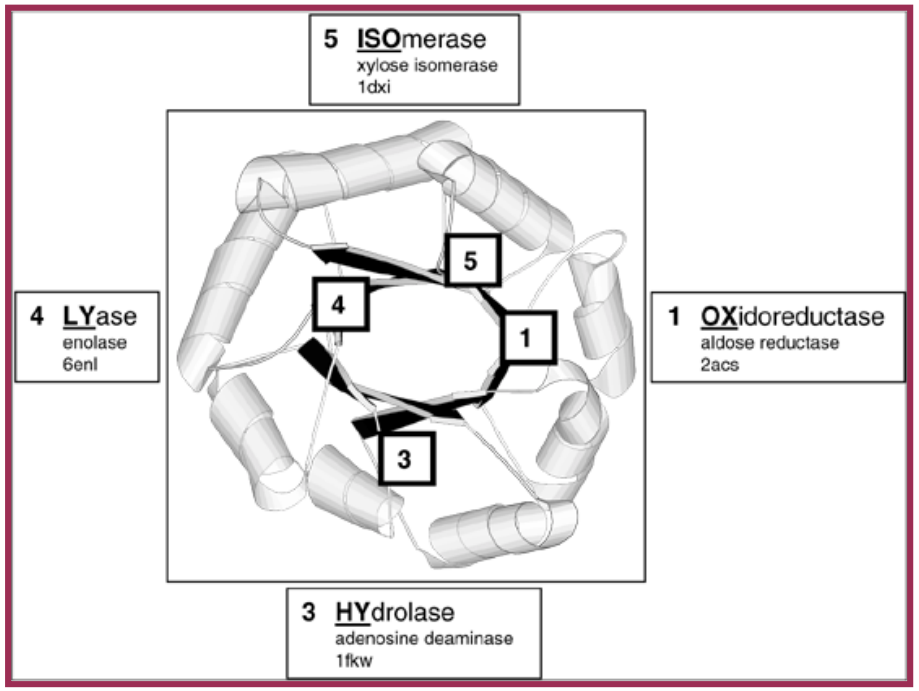




(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

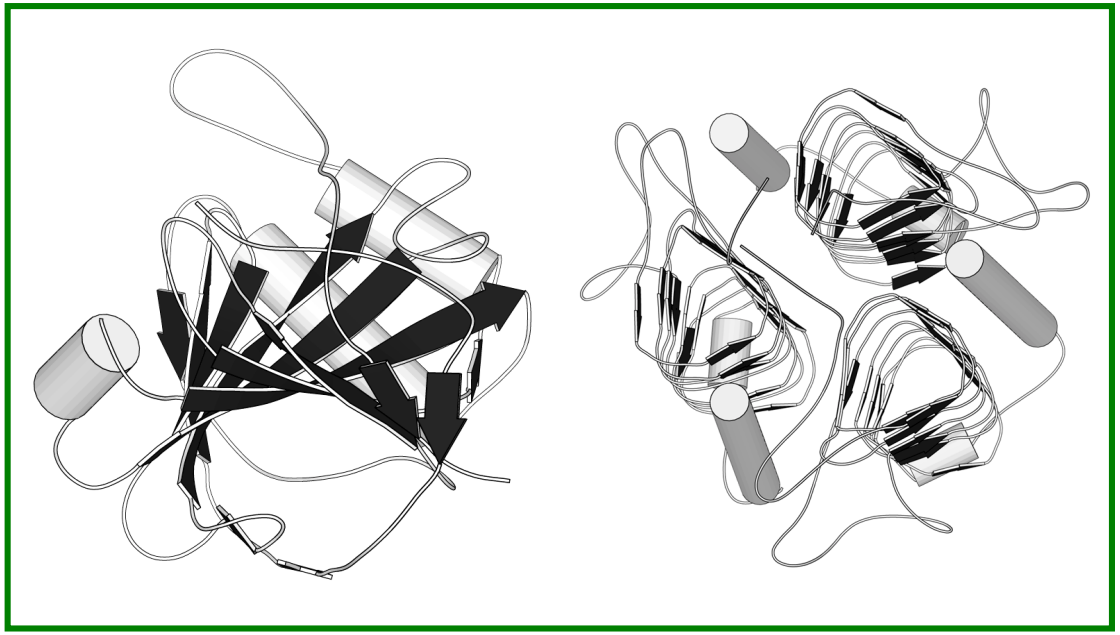
Fold-Function Combinations #2

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



Have the sequences diverged beyond point of homology? Or does this occur at a certain sequence threshold?

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



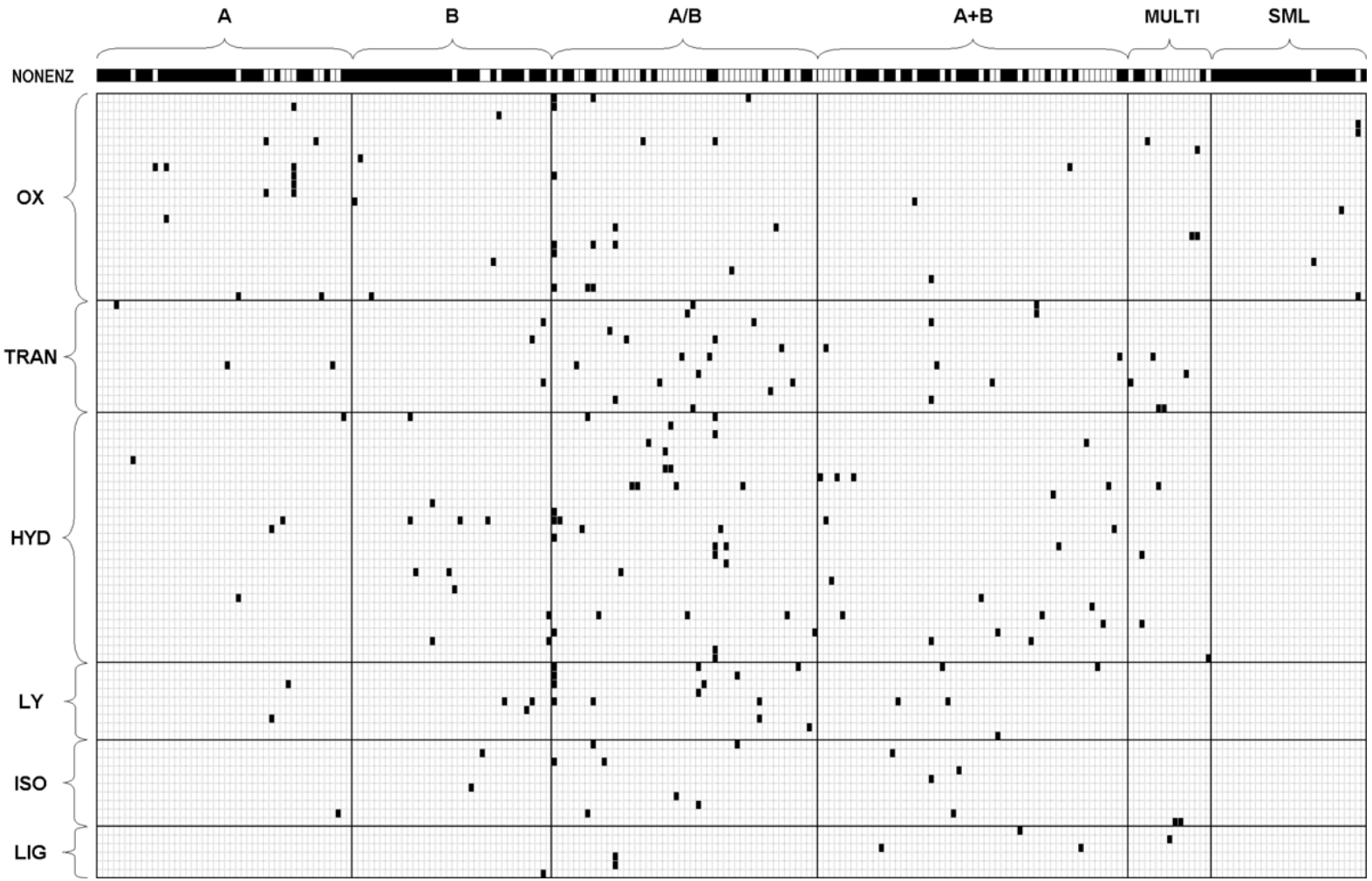
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Fold-Function Combinations

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

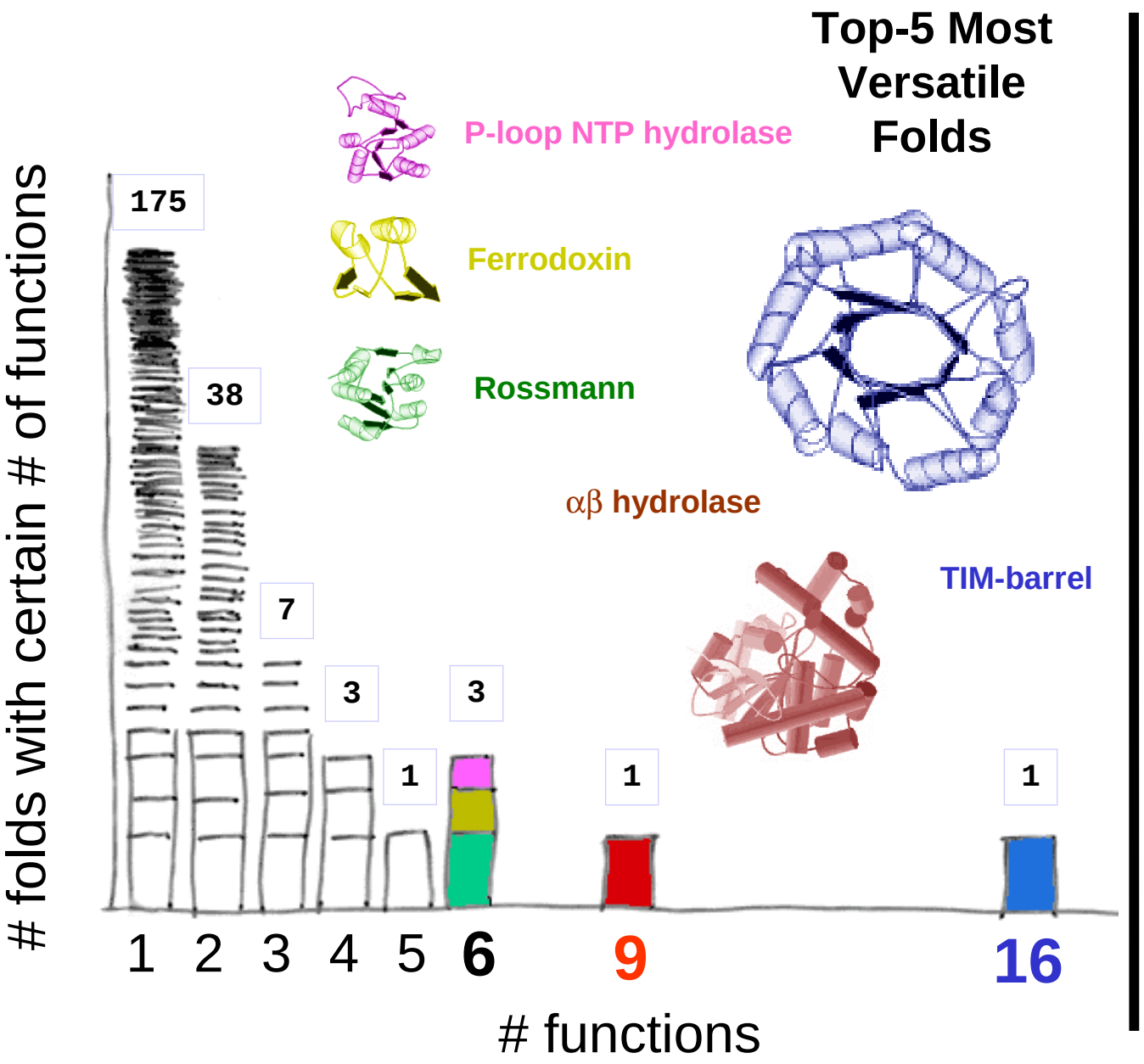
229 Folds

91 Enzymatic Functions
+ Non-Enzyme

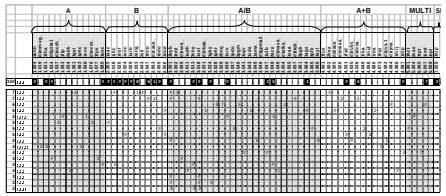


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Most Versatile Folds & Functions

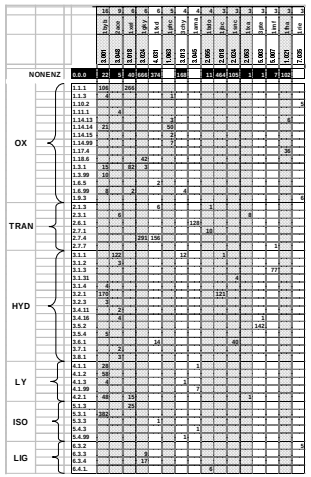


Top Multifold Functions →



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

Top Multifunctional Folds →

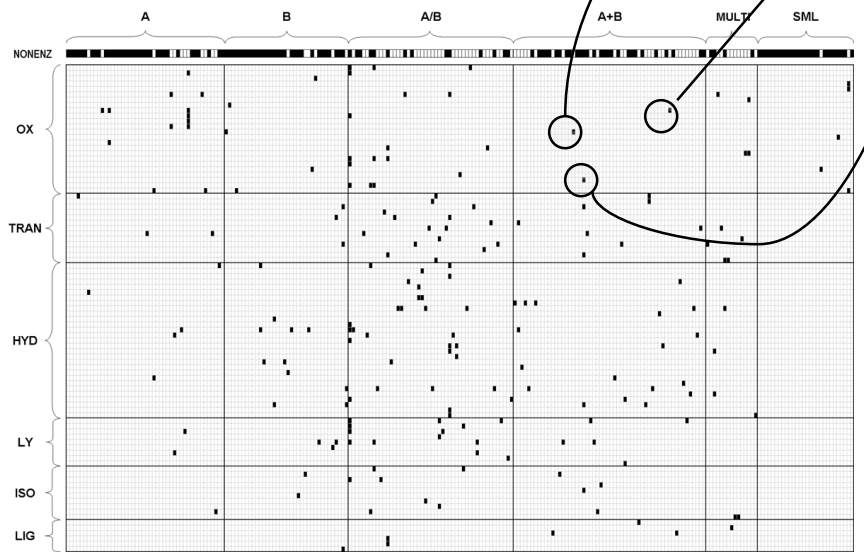


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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

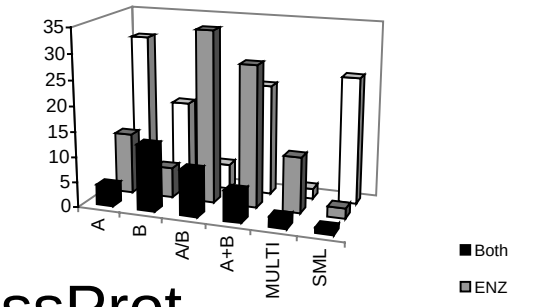
SCOP

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		

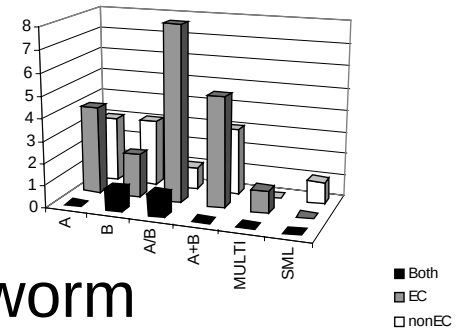
CATH (Thornton)

CATH

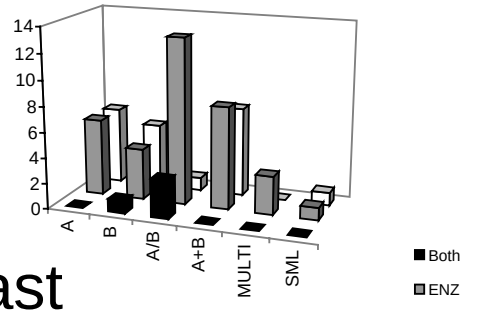
ENZYME	A	B	AB
NONENZ	10	9.0	15
OX	5.1	5.1	10
TRAN		1.3	13
HYD	2.6	1.3	14
LY		2.6	1.3
ISO	1.3	1.3	5.1
LIG			1.3



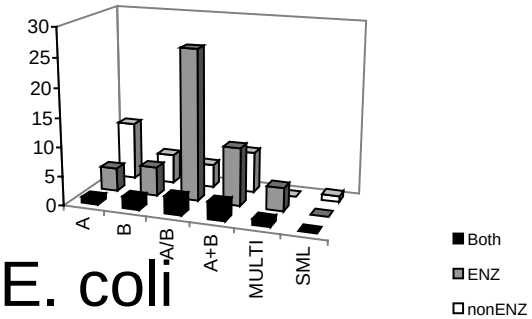
SwissProt



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	

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Different Structure Function Relationships for Most Ancient Proteins

All Yeast COGs

		SCOP A	SCOP B	SCOP A/B	SCOP A+B	SCOP MULTI	SCOP SML
Metabolism	C	2.2	2.6	4.8	3	0.4	
	E	2.2	1.1	7.4	2.6	0.7	
	F	1.1		3.7	1.8		
	G	0.4	0.4	3.3	0.7		
	H	1.1	0.7	4.8	3		
	I	0.7	0.7	2.2	0.4	0.4	
Information Storage & Processing	J	2.2	1.8	3	3	0.4	0.4
	K			1.1	0.4		
	L	1.1		1.5	1.1	1.1	
Cellular Processes	M		0.4	0.4	0.7		
	N	1.8	0.7	0.4	0.7		0.4
	O	1.5	1.1	3	2.2	0.4	0.4
	P		0.4	1.1	0.7	0.4	

Most Conserved COGs

		SCOP A	SCOP B	SCOP A/B	SCOP A+B	SCOP MULTI	SCOP SML
Metabolism	C			7.2	2.9		
	E	1.4		1.4	1.4		
	F			2.9			
	G			4.3	1.4		
	H	1.4	2.9		1.4		
	I						
Information Storage & Processing	J	8.7	7.2	7.2	10	1.4	1.4
	K						
	L					1.4	
Cellular Processes	M						
	N	1.4		1.4			
	O	2.9		7.2	2.9		
	P		1.4		2.9	1.4	

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

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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

W Krebs, J Tsai, M Levitt, C Wilson, R Das, H Hegyi, J Lin, Y Kluger, C Arrowsmith, A Edwards, L Regan, S Balasubramanian, A Drawid, D Greenbaum, M Snyder, R Jansen

bioinfo.mbb.yale.edu

Structures ("Classic")

(now) Structural Genomics

(now) Func. Genomics

Arrays (future)

Structures ("Classic")

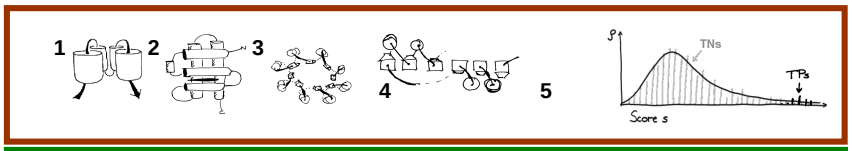
1 Fold Library (A parts list.) Structural Alignment, EVD P-value, Seq. Struc. diverg.

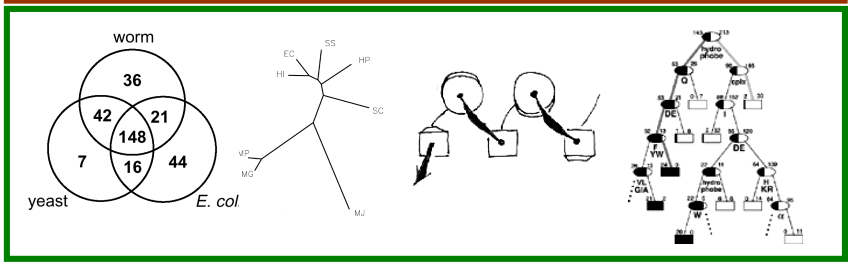
2 Folds in Genomes (Shared, common, and/or unique parts?) Known Folds. Fold Tree, Top-10. βαβ. Biases. MG fold assignment extent. MG Target Selection, MT retrospective decision tree.

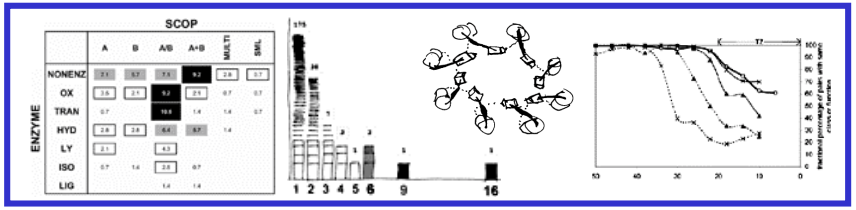
3 Folds & Functions (Roles/part?) How many folds /function? Mostly 1, but TIM versatile. Seq. diverg. vs. Func. diverg.

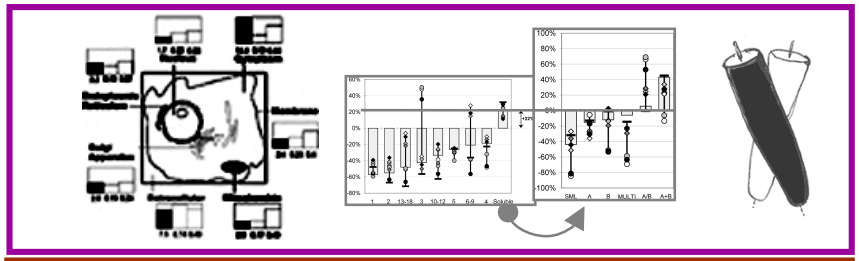
4 Folds in the Transcriptome (Common parts? Where are parts?) Enriched ↑ : VGA, TIM, αβ folds, energy, synthesis, cyt. Depleted ↓ : NS, long, TM folds, transport, transcription, Leu-zip, nuc. Bayesian Localizer, phenotypes clustering

5 Fold Flexibilty (How adaptable is a part?). Motions DB, morph server, interface packing, Voronoi Volumes





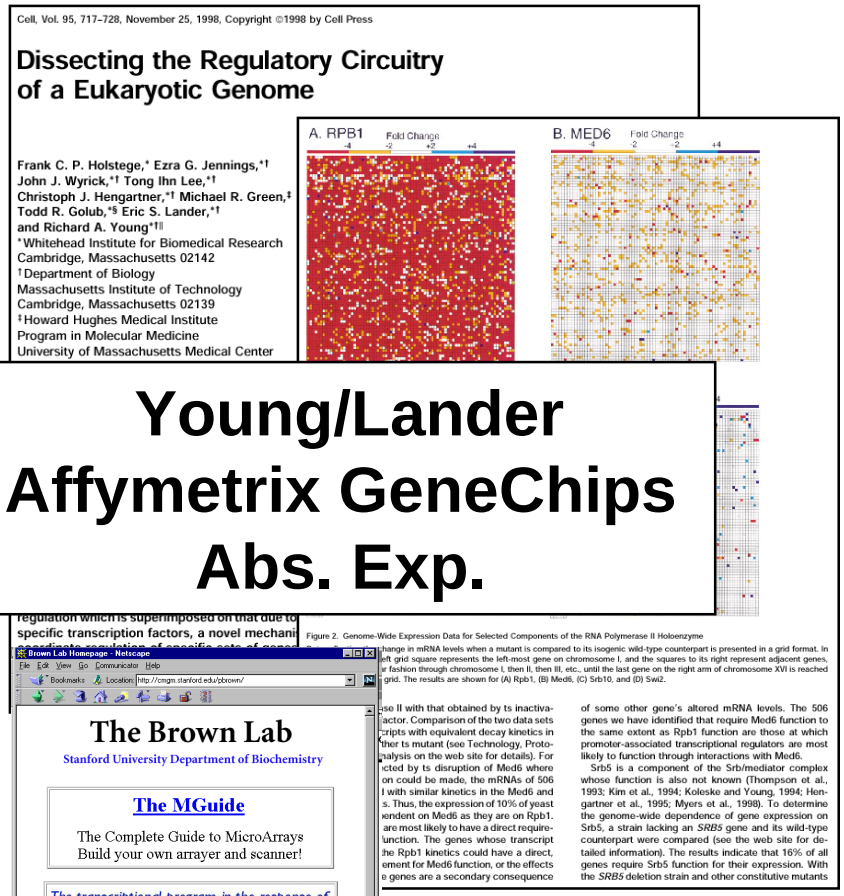




Gene Expression Datasets: the Transcriptome

Yeast Expression Data in Academia: levels for all 6000 genes!

X-ref. with other genome data: protein fold features common in Transcriptome....


Cell, Vol. 95, 717–728, November 25, 1998, Copyright ©1998 by Cell Press

Dissecting the Regulatory Circuitry of a Eukaryotic Genome

Frank C. P. Holstege,* Ezra G. Jennings,*† John J. Wyrick,*† Tong Ihn Lee,*† Christoph J. Hengartner,*† Michael R. Green,‡ Todd R. Golub,*§ Eric S. Lander,*† and Richard A. Young*†‖

*Whitehead Institute for Biomedical Research Cambridge, Massachusetts 02142
†Department of Biology Massachusetts Institute of Technology Cambridge, Massachusetts 02139
‡Howard Hughes Medical Institute Program in Molecular Medicine University of Massachusetts Medical Center

A. RPB1 Fold Change -4 -2 +2 +4

B. MED6 Fold Change -4 -2 -2 +4

Figure 2. Genome-Wide Expression Data for Selected Components of the RNA Polymerase II Holoenzyme

Young/Lander Affymetrix GeneChips Abs. Exp.

Brown Lab Homepage - Netscape

The Brown Lab

Stanford University Department of Biochemistry

The MGuide

The Complete Guide to MicroArrays
Build your own arrayer and scanner!

The transcriptional program in the response of human fibroblasts to serum

The web supplement to: Iyer V.R. et al., (1999) Science 283: 83-87

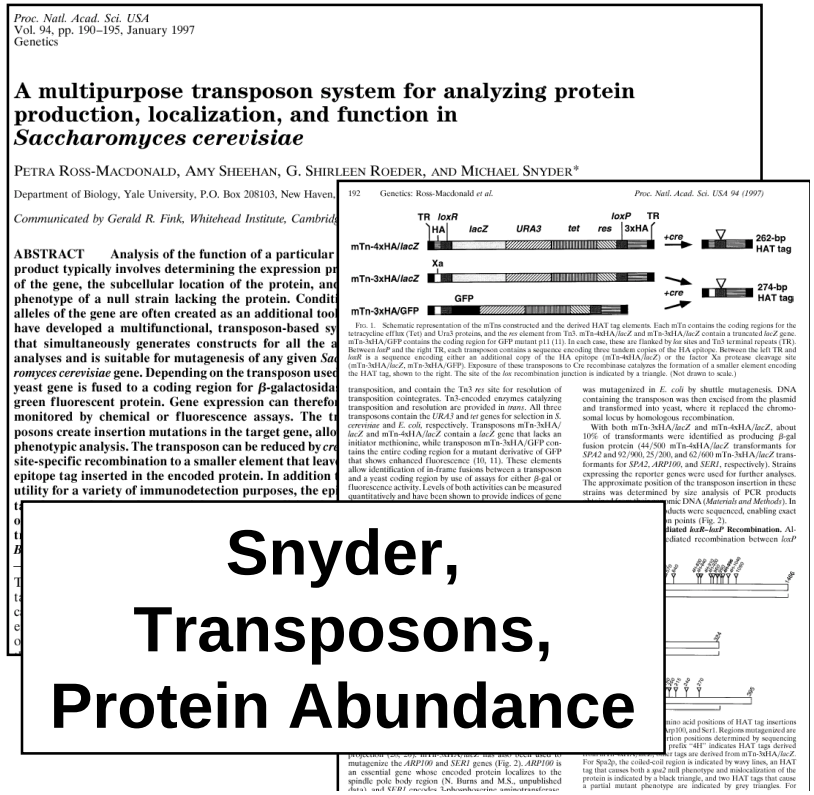
The Transcription of Sporulation in

The Web Companion

Science Magazine Re

Brown, μarrays, Rel. Exp. over Timecourse

Also: SAGE; Samson and Church, Chips; Aebersold, Protein Abundance


Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 190–195, January 1997
Genetics

A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*

PETRA ROSS-MACDONALD, AMY SHEEHAN, G. SHIRLEEN ROEDER, AND MICHAEL SNYDER*

Department of Biology, Yale University, P.O. Box 208103, New Haven,

Communicated by Gerald R. Fink, Whitehead Institute, Cambrid

ABSTRACT Analysis of the function of a particular product typically involves determining the expression pr of the gene, the subcellular location of the protein, an phenotype of a null strain lacking the protein. Condit alleles of the gene are often created as an additional too have developed a multifunctional, transposon-based sy that simultaneously generates constructs for all the a analyses and is suitable for mutagenesis of any given *Sa romyces cerevisiae* gene. Depending on the transposon usec yeast gene is fused to a coding region for β-galactosida green fluorescent protein. Gene expression can therefo monitored by chemical or fluorescence assays. The t posons create insertion mutations in the target gene, allo phenotypic analysis. The transposon can be reduced by cr site-specific recombination to a smaller element that leav epitope tag inserted in the encoded protein. In addition utility for a variety of immunodetection purposes, the ep

192 Genetics: Ross-Macdonald *et al.* *Proc. Natl. Acad. Sci. USA 94 (1997)*

TR loxR HA lacZ URA3 tet res loxP 3xHA TR

mTn-4xHA/*lacZ* +cre 262-bp HAT tag

Xa

mTn-3xHA/*lacZ* +cre 274-bp HAT tag

GFP

mTn-3xHA/GFP

FIG. 1. Schematic representation of the mTns constructed and the derived HAT tag elements. Each mTn contains the coding regions for the tetracycline efflux (Tet) and Ura3 proteins, and the *res* element from Tn3. mTn-4HA/*lacZ* and mTn-3HA/*lacZ* contain a truncated *lacZ* gene. mTn-3HA/GFP contains the coding region for GFP mutant p11 (11). In each case, these are flanked by *lox* sites and Tn3 terminal repeats (TR). Between *loxP* and the right TR, each transposon contains a sequence encoding three tandem copies of the HA epitope. Between the left TR and *loxR* is a sequence encoding either an additional copy of the HA epitope (mTn-4HA/*lacZ*) or the factor Xa protease cleavage site (mTn-3xHA/*lacZ*, mTn-3xHA/GFP). Exposure of these transposons to Cre recombinase catalyzes the formation of a smaller element encoding the HAT tag, shown to the right. The site of the *lox* recombination junction is indicated by a triangle. (Not drawn to scale.)

transposition, and contain the Tn3 *res* site for resolution of transposition cointegrates. Tn3-encoded enzymes catalyzing transposition and resolution are provided in *trans*. All three transposons contain the *URA3* and *tet* genes for selection in *S. cerevisiae* and *E. coli*, respectively. Transposons mTn-3xHA/*lacZ* and mTn-4xHA/*lacZ* contain a *lacZ* gene that lacks an initiator methionine, while transposon mTn-3xHA/GFP contains the entire coding region for a mutant derivative of GFP that shows enhanced fluorescence (10, 11). These elements allow identification of in-frame fusions between a transposon and a yeast coding region by use of assays for either β-gal or fluorescence activity. Levels of both activities can be measured quantitatively and have been shown to provide indices of gene

was mutagenized in *E. coli* by shuttle mutagenesis. DNA containing the transposon was then excised from the plasmid and transformed into yeast, where it replaced the chromosomal locus by homologous recombination.

With both mTn-3xHA/*lacZ* and mTn-4xHA/*lacZ*, about 10% of transformants were identified as producing β-gal fusion protein (44/800 mTn-4xHA/*lacZ* transformants for *SPA2* and 92/900, 25/200, and 62/600 mTn-3xHA/*lacZ* transformants for *SPA2*, *ARP100*, and *SER1*, respectively). Strains expressing the reporter genes were used for further analyses. The approximate position of the transposon insertion in these strains was determined by size analysis of PCR products

ducts were sequenced, enabling exact

n points (Fig. 2).

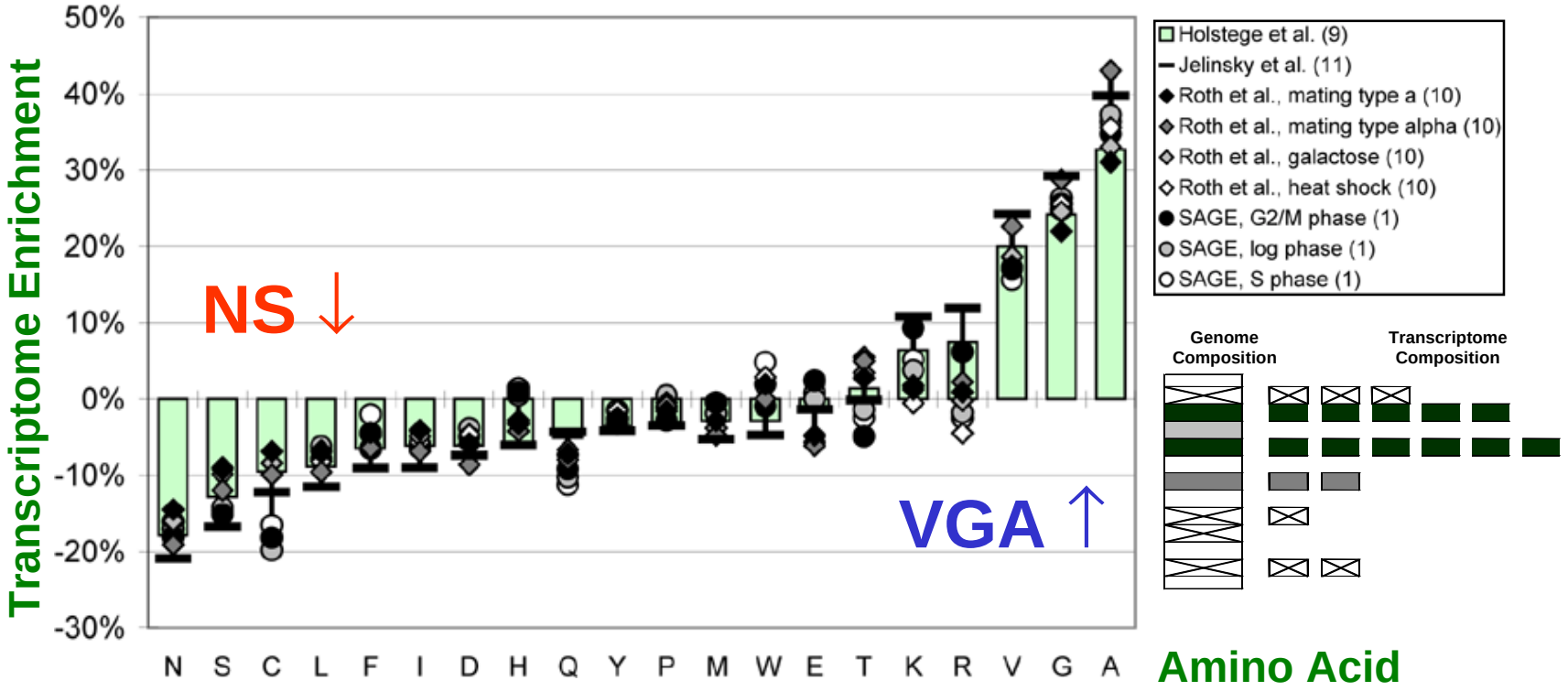
diated *lexR–lexP* Recombination. Al-

diated recombination between *lexP*

Snyder, Transposons, Protein Abundance

Composition of Genome vs. Transcriptome

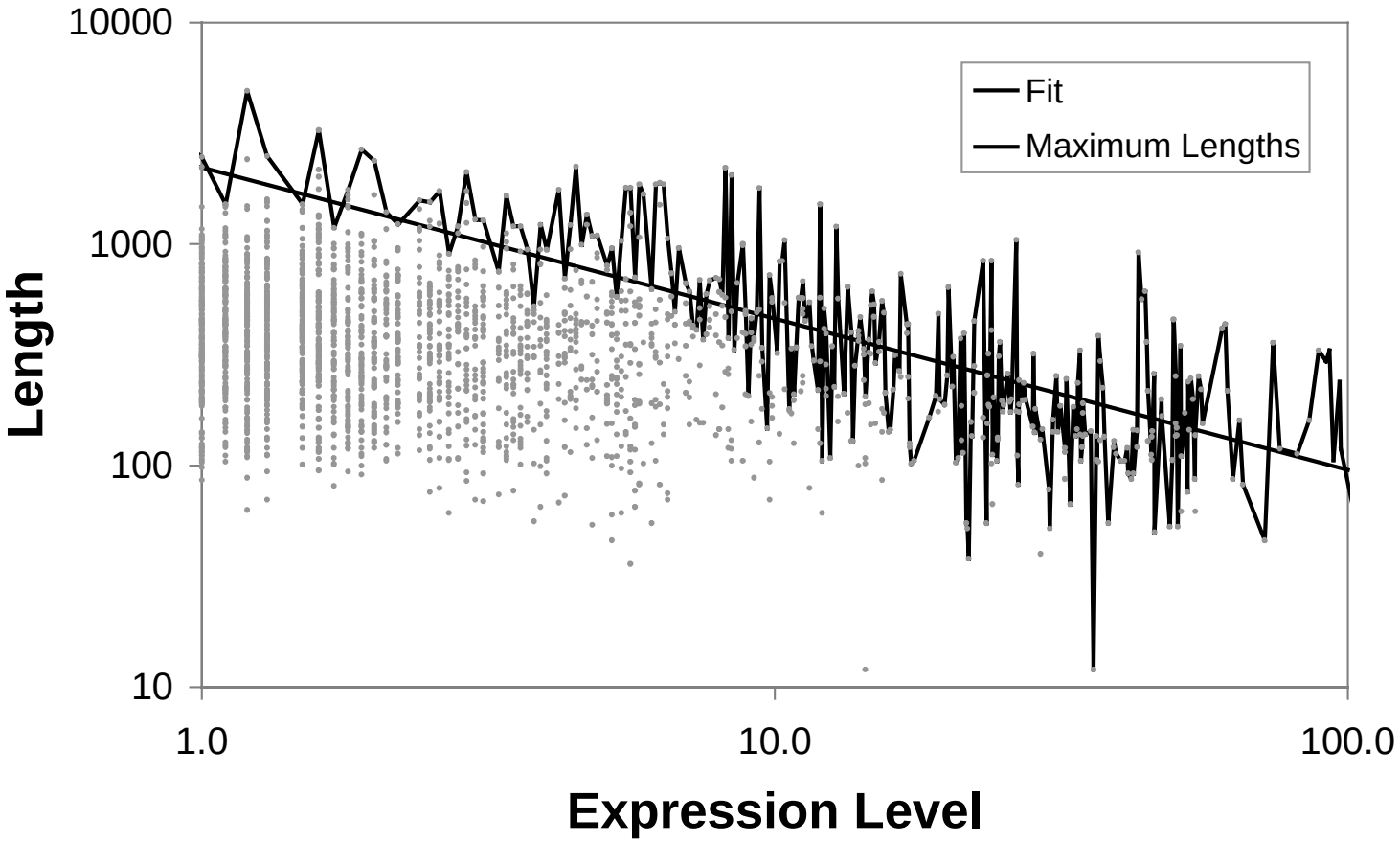
	$\sum_{\text{orf } i} n_i(F)$	$\sum_F \sum_{\text{orf } i} n_i(F)$	$G(F)$	$\sum_{\text{orf } i} e_i n_i(F)$	$\sum_F \sum_{\text{orf } i} e_i n_i(F)$	$T(F)$	$D(F)$
Feature *F* is Amino acids, in particular Ala	Number of Ala in yeast	Number of amino acids in yeast	Genome composition of Ala in yeast	Number of Ala weighted by expression	Number of amino acids weighted by expression	Transcriptome composition of Ala in yeast	Relative enrichment of Ala in transcriptome
Spec. Num.	**141890**	**2574876**	**5.5%**	**347807**	**4758441**	**7.3%**	**32.7%**
Feature *F* is Folds, in particular the TIM-barrel (3.1)	Number of TIM-barrel fold matches in yeast genome	Number of matches with all folds in yeast genome	Genome composition of TIM-barrel fold matches	Number of TIM-barrel fold matches weighted by expression	Number of matches with all folds weighted by expression	Transcriptome composition of TIM-barrel fold matches	Relative enrichment of TIM-barrel matches in transcriptome
Spec. Num.	**65**	**1560**	**4.2%**	**389**	**4709**	**8.3%**	**97.8%**



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

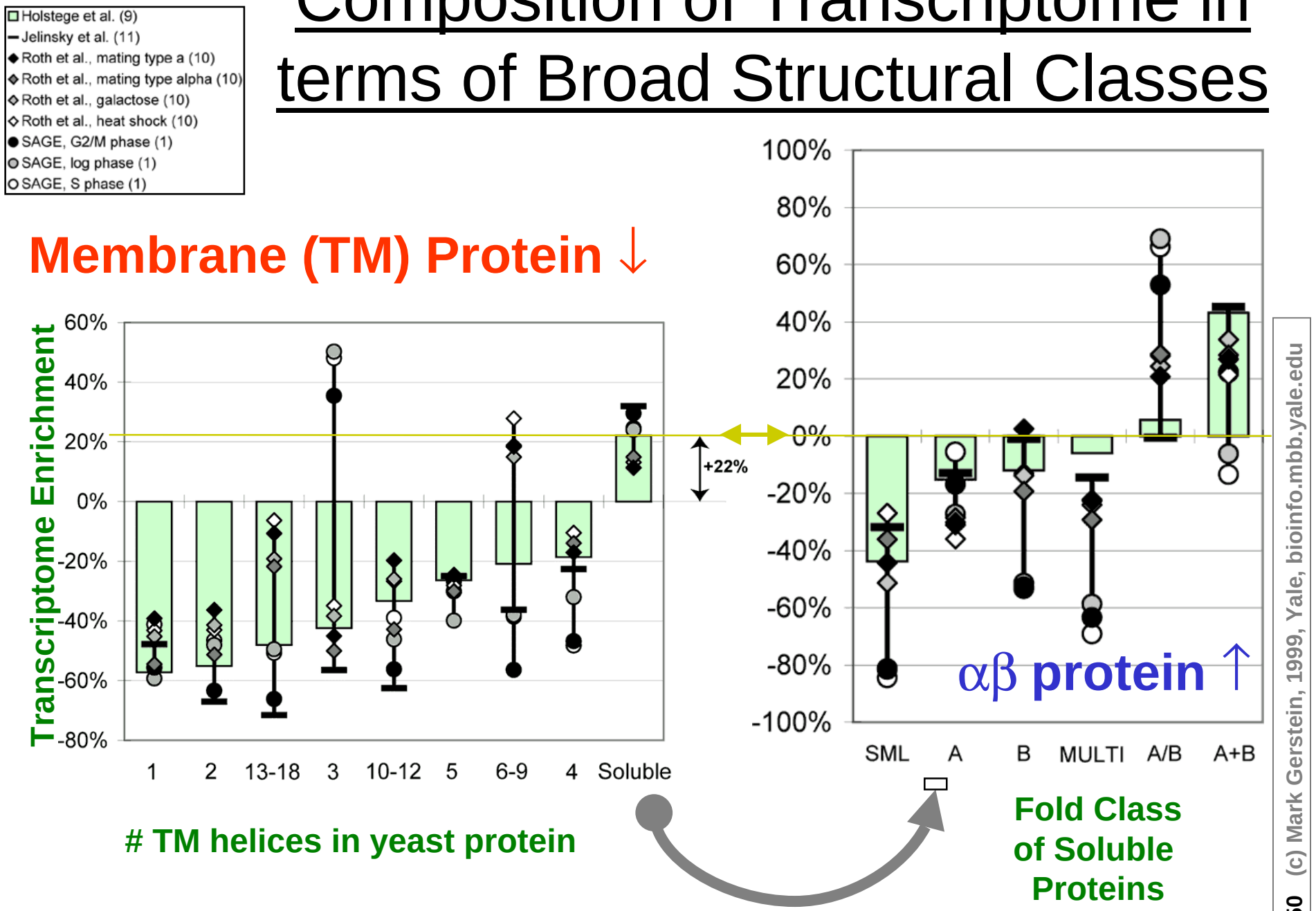
Relation between Length & Expression

Max Expression (e.g. transcripts/cell) ~ $(\text{Length})^{-2/3}$
Shorter proteins can be more highly expressed

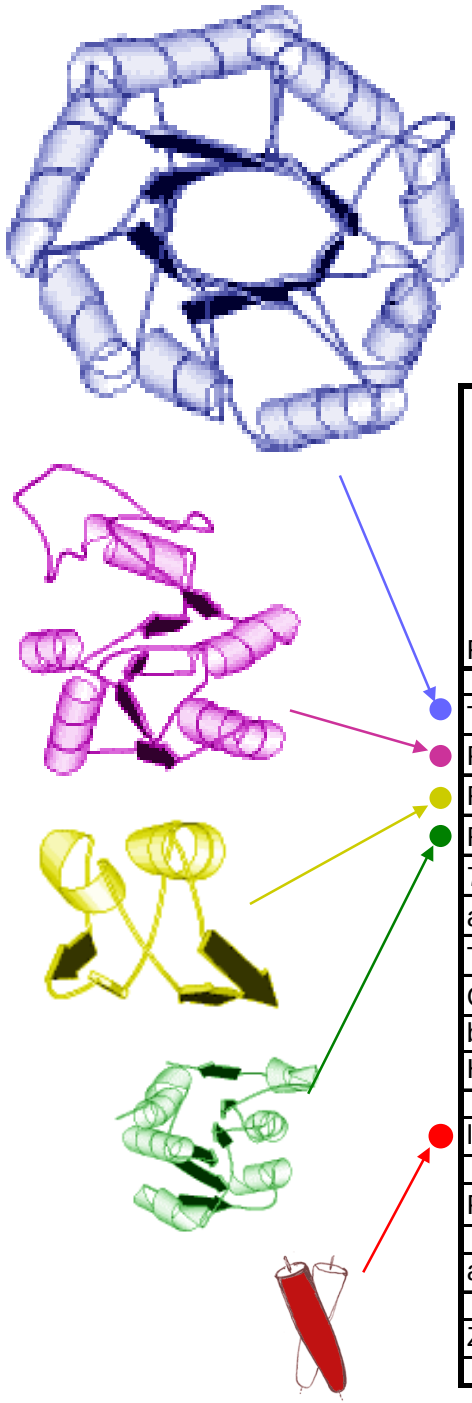


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Composition of Transcriptome in terms of Broad Structural Classes



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

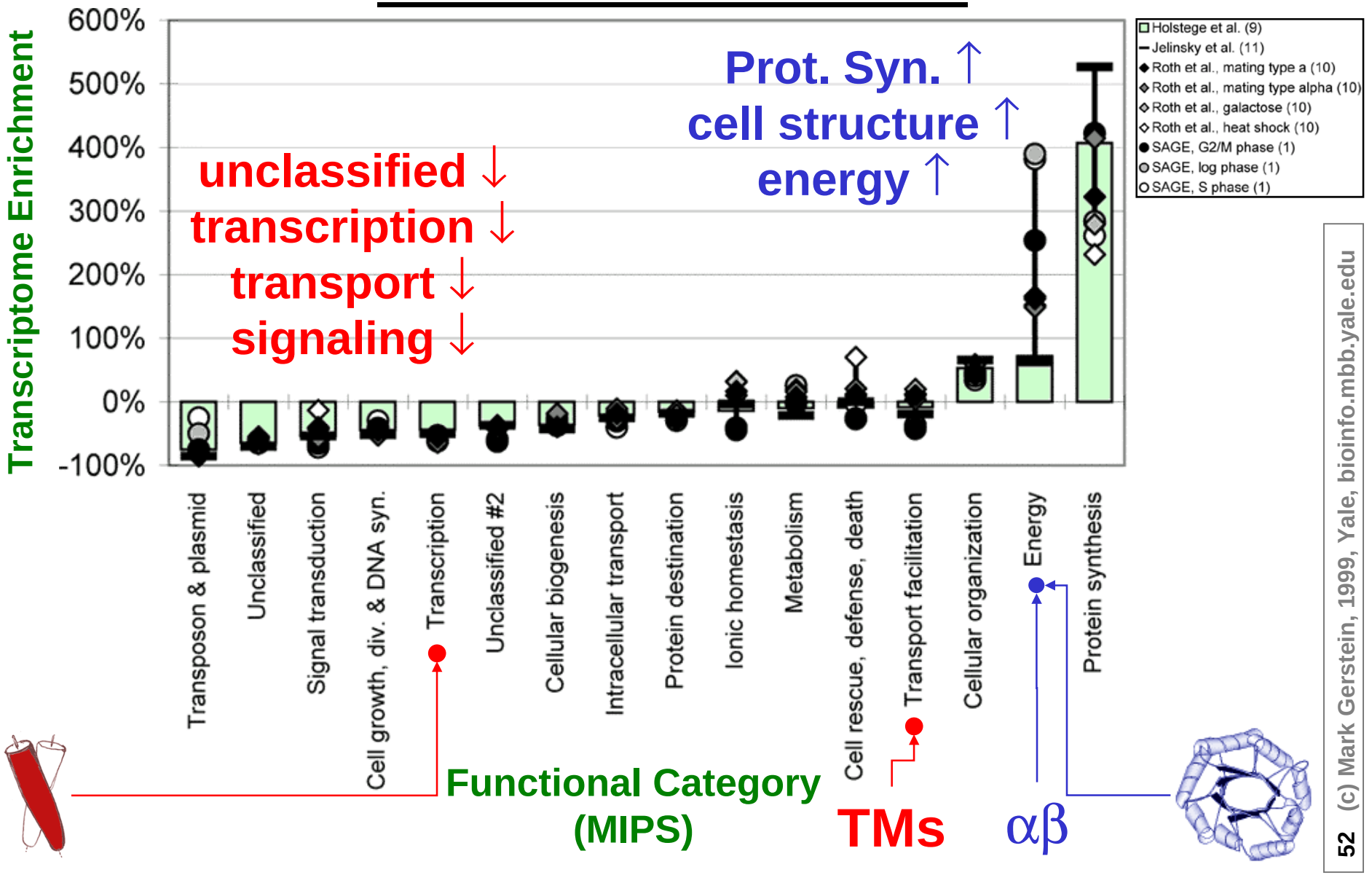


Which Protein Folds are Highly Expressed?

Top-10 folds in genome and transcriptome

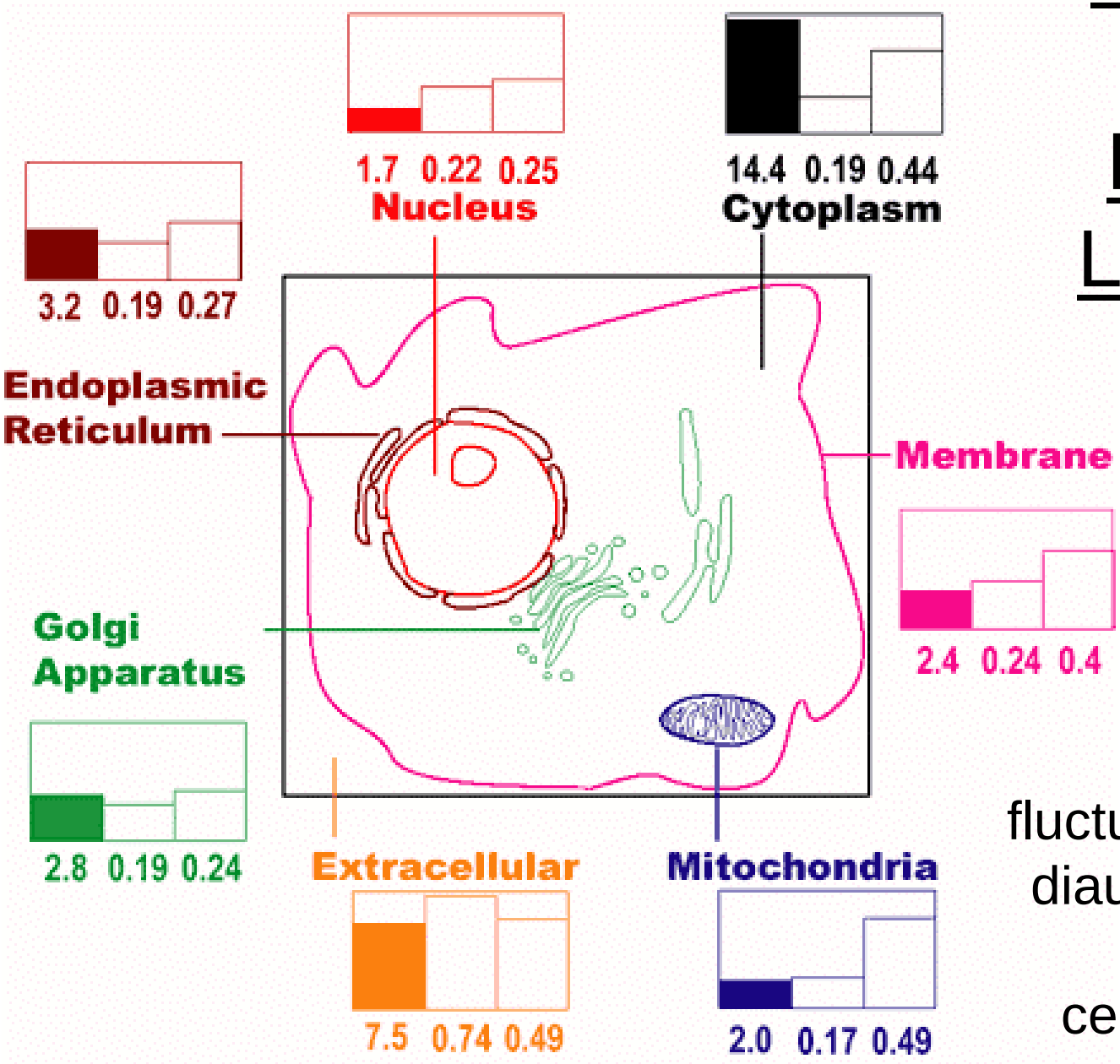
Fold	Fold Class	Rep. PDB	Composition: Genome [%]	Composition: Transcriptome [%]	Composition: Rel. Diff. [%]		Rank: Genome	Rank: Young	Rank: Samson	Rank: Church-a	Rank: Church-alpha	Rank: Church-gal	Rank: Church-heat	Rank: SAGE-GM	Rank: SAGE-L	Rank: SAGE-S
TIM barrel	α/β	1byb	4.2	8.3	+98	+	5	1	1	1	1	1	1	1	1	1
P-loop NTP hydrolases	α/β	1gky	5.8	5.2	-11	•	3	2	2	4	4	4	5	5	6	7
Ferredoxin like	α+β	1fxd	3.9	3.4	-14	•	6	3	7	11	9	8	10	4	10	11
Rossmann fold	α/β	1xel	3.3	3.3	0	•	8	4	3	3	3	2	2	19	15	9
7-bladed beta-propeller	β	1mda*	6.4	2.9	-55	-	2	5	4	5	6	6	7	9	9	16
aplha-alpha superhelix	α	2bct	4.4	2.7	-37	-	4	6	11	15	16	12	12	8	5	8
Thioredoxin fold	α/β	2trx	1.7	2.7	+63	+	14	7	6	8	2	5	4	11	10	6
G3P dehydrogenase-like	α+β	1drw†	0.2	2.7	+1316	+	81	8	12	2	5	3	3	35	19	30
beta grasp	α+β	1igd	0.6	2.6	+348	+	36	9	10	21	9	18	21	82	122	120
HSP70 C-term. fragment	multi	1dky	0.8	2.6	+231	+	31	10	16	17	11	16	12	48	25	56
long helices oligomers	α	1zta	3.8	2.1	-46	-	7	15	8	14	21	15	19	21	20	33
Protein kinases (cat. core)	multi	1hcl	6.8	1.6	-77	-	1	18	19	9	16	11	15	13	16	17
alpha/beta hydrolases	α/β	2ace	2.2	0.9	-62	-	10	32	31	25	26	21	23	26	26	26
Zn2/C6 DNA-bind. dom.	sml	1aw6	2.6	0.3	-89	-	9	75	94	27	50	32	40	48	39	50

Composition of Transcriptome in terms of Functional Classes

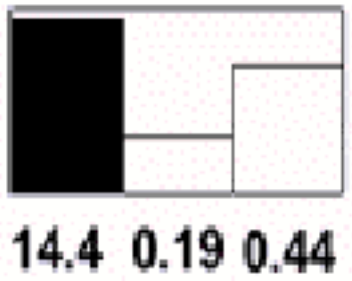


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Expression Level is Related to Localization



abs. exp. (GeneChip)

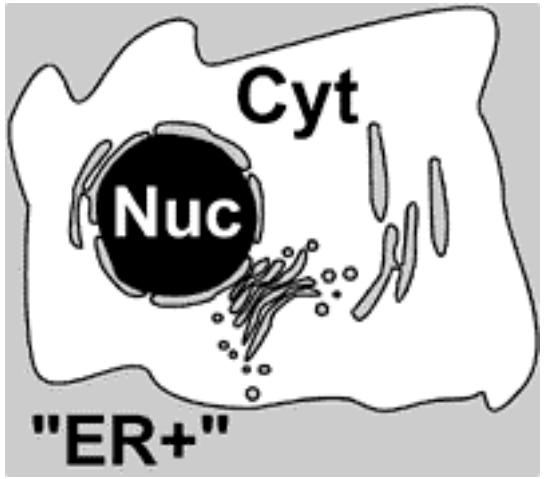


fluctuations in diauxic shift or cell-cycle

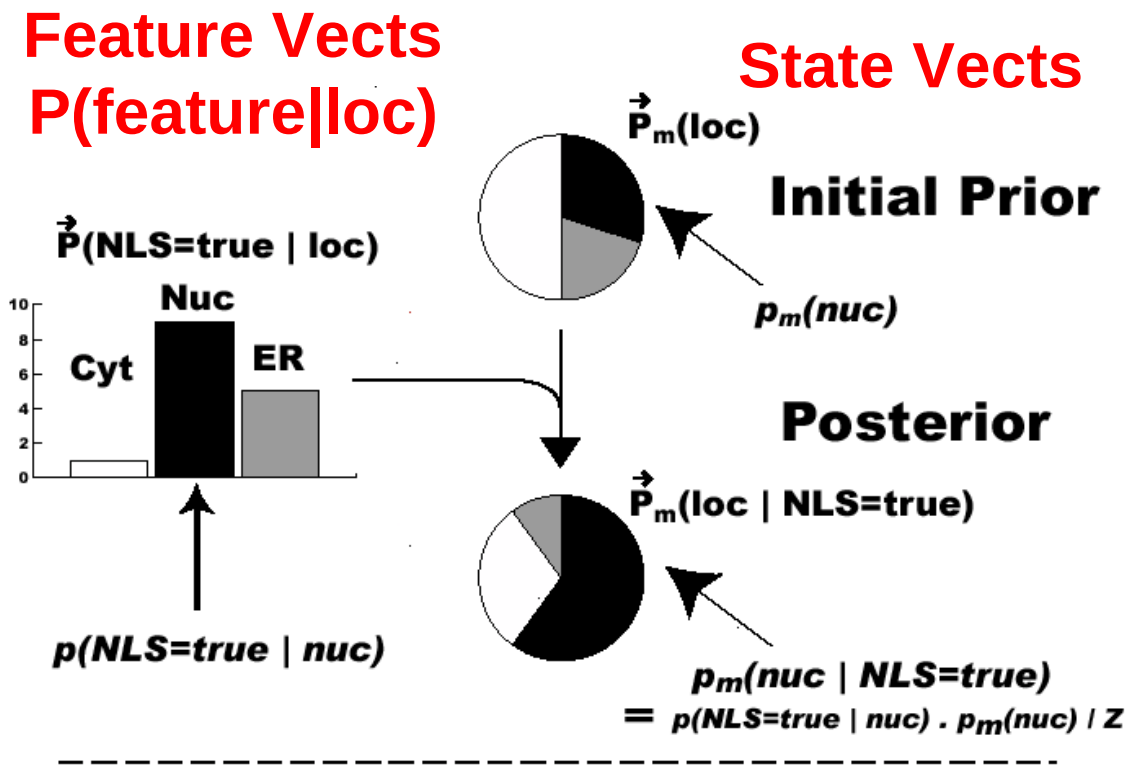
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Bayesian System for Localizing Proteins

loc=



Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.

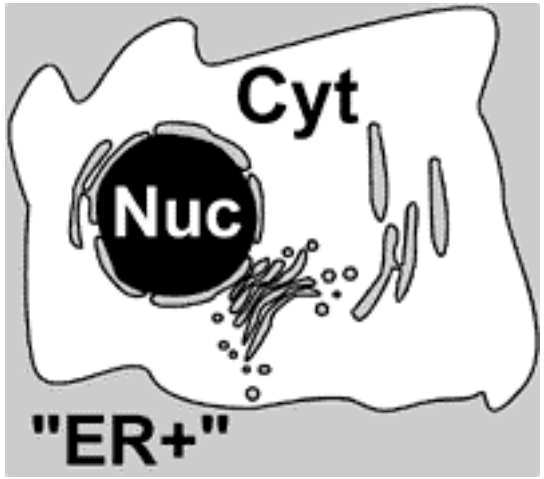


18 Features: Expression Level (absolute and fluctuations), signal seq., KDEL, NLS, Essential?, aa composition

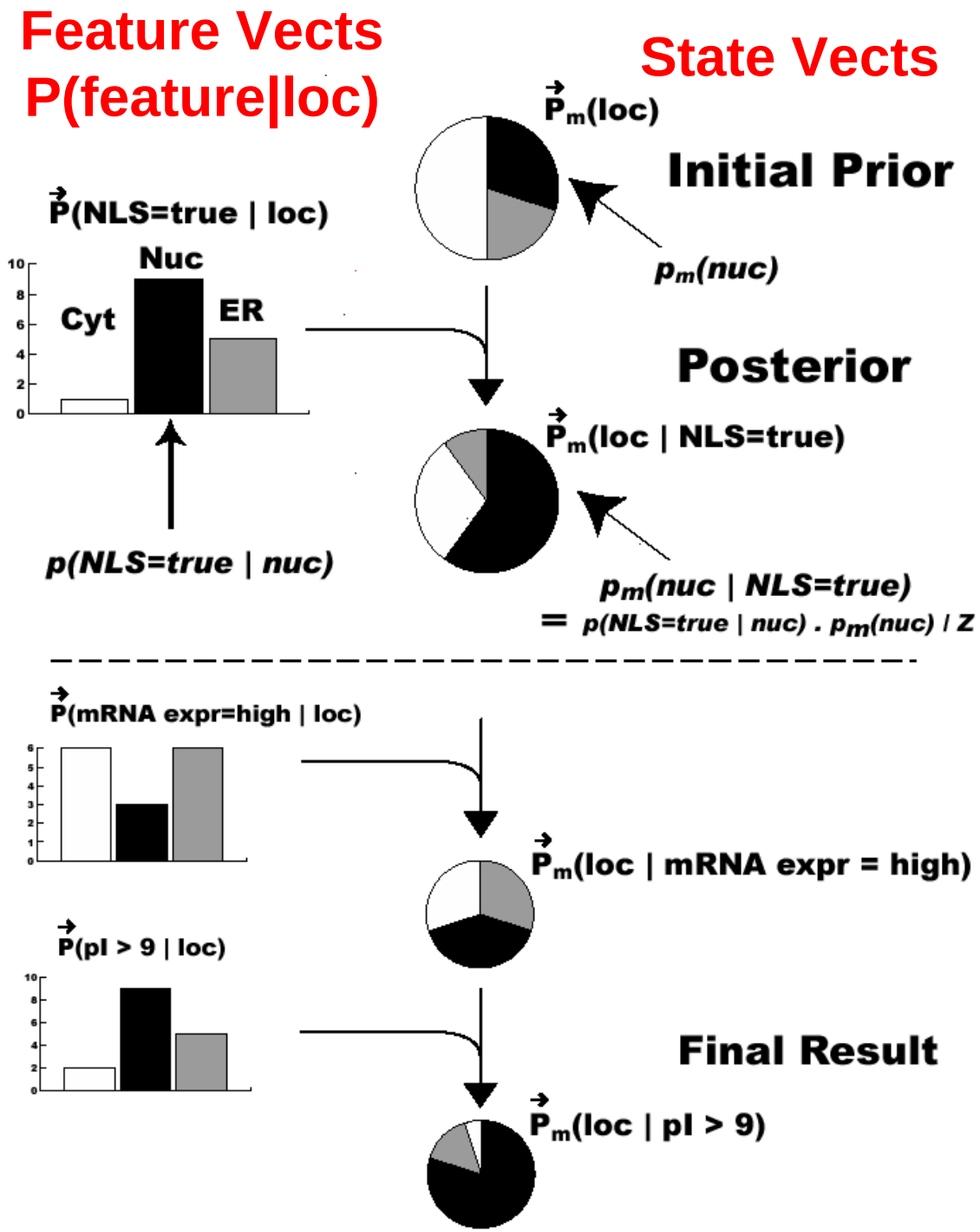
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Bayesian System for Localizing Proteins

loc=



Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.

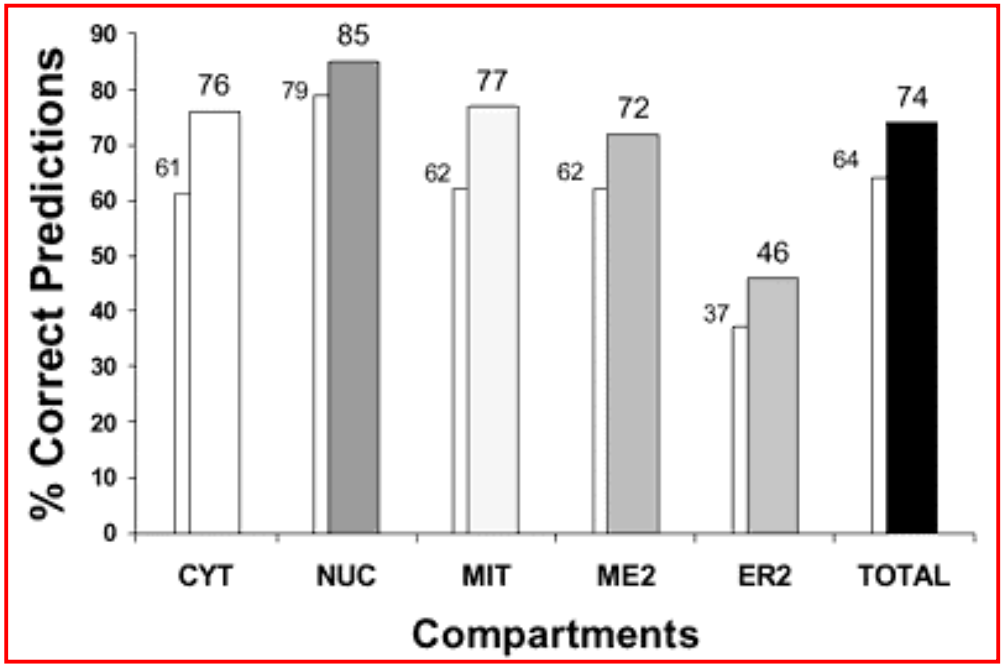


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Results on Testing Data

Individual proteins: 74% with cross-validation

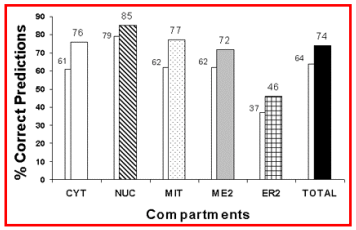
Testing, training data, Priors: ~2000 proteins from YPD, MIPS, SwissProt, Snyder Lab



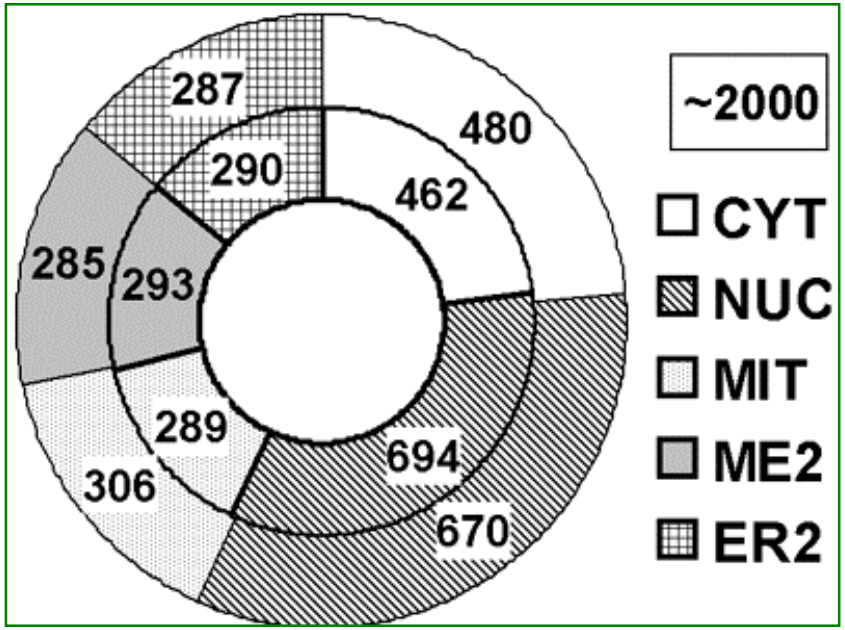
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Results on Testing Data #2

Individual proteins: 74% with cross-validation



Compartment Populations. directly sum state vectors to get population. Gives **96%** pop. similarity.



overall compartment population vector

state vector of protein 1

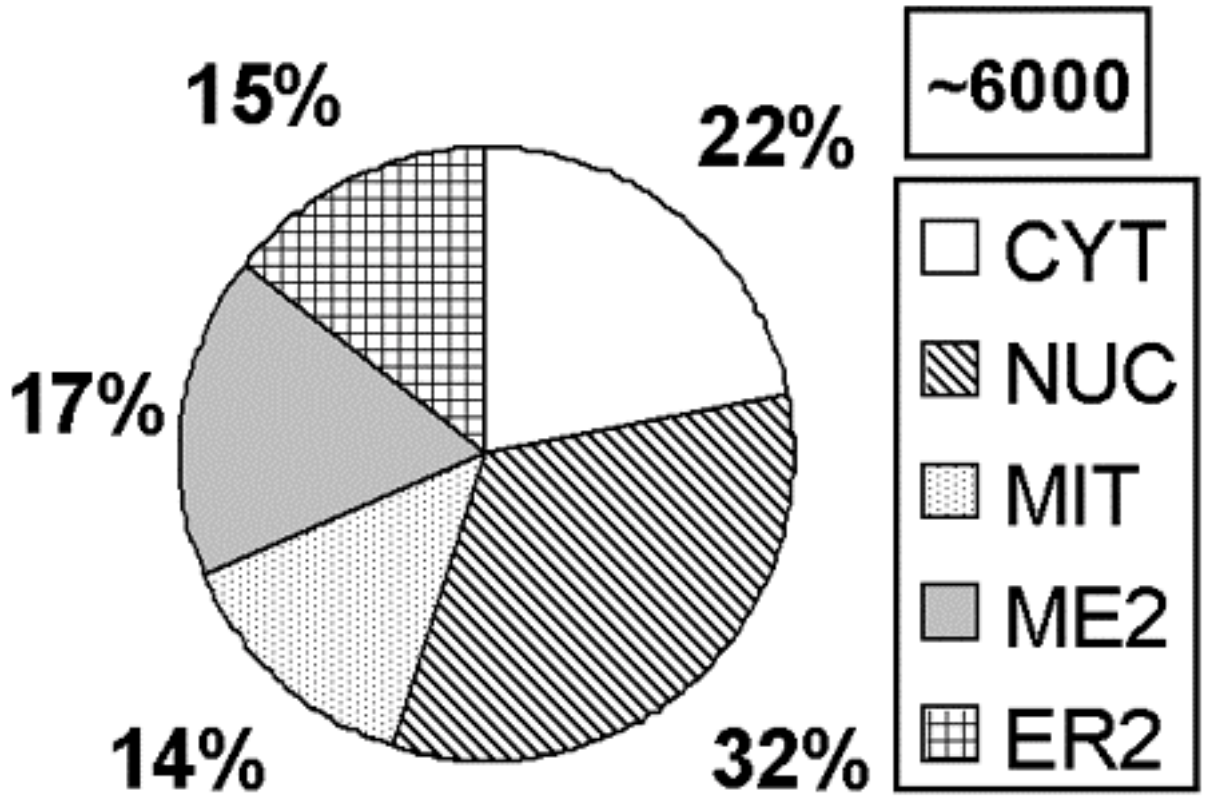
"Normal" sum of only well localized proteins

"QM-like" sum of all state vectors

$$\mathbf{N}(loc) = \mathbf{P}_1(loc) + \mathbf{P}_2 + \mathbf{P}_3 + \mathbf{P}_4 + \cdots + \mathbf{P}_m + \mathbf{P}_{6000}$$

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

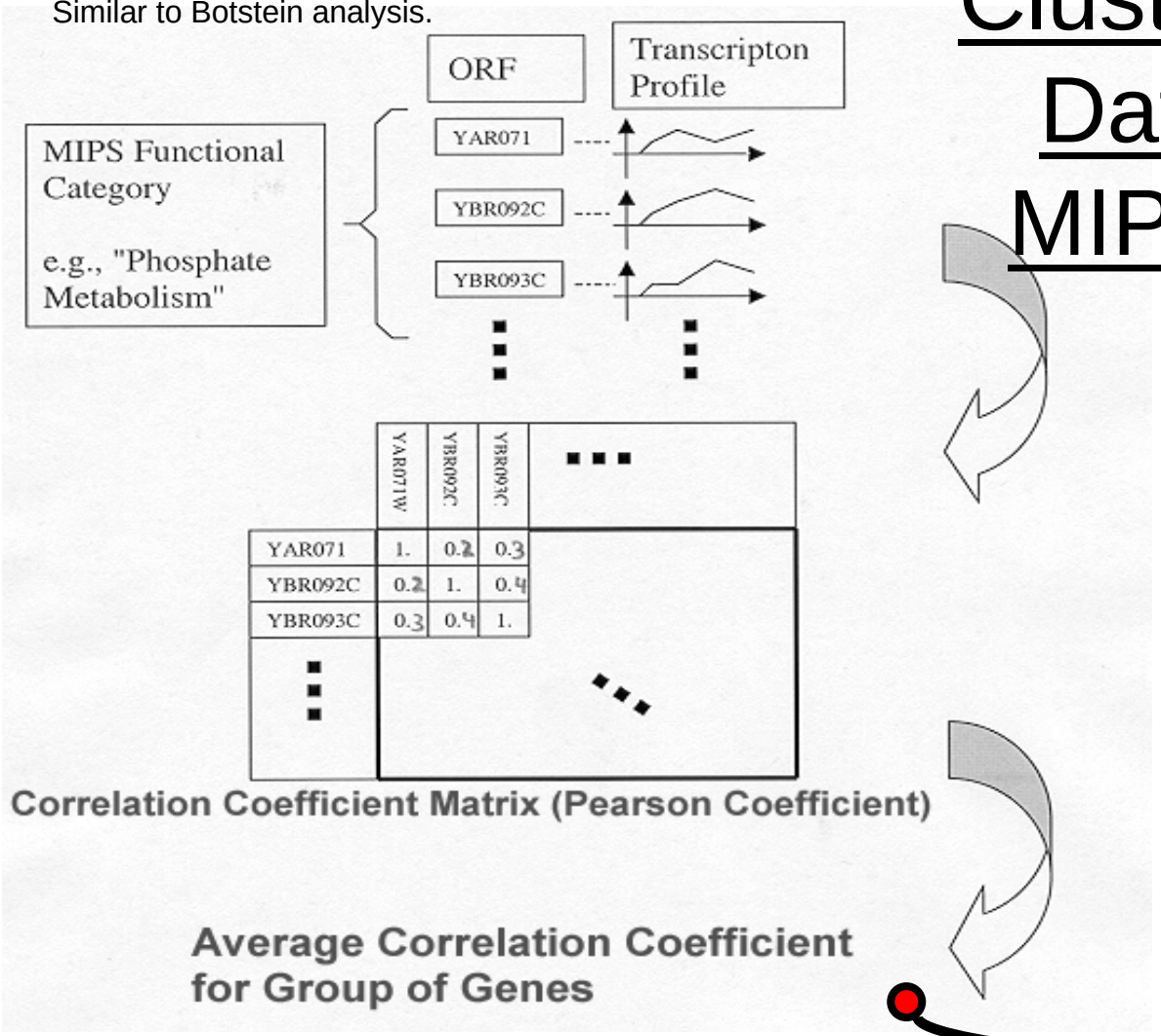
Extrapolation to Compartment Populations of Whole Yeast Genome: ~4000 predicted + ~2000 known



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Functional category number	Function	Average correlation	# ORFs
01	METABOLISM	0.1001	1005
01.01	amino-acid metabolism	0.1488	199
01.01.01	amino-acid biosynthesis	0.239	114
01.01.04	regulation of amino-acid metabolism	0.23	32

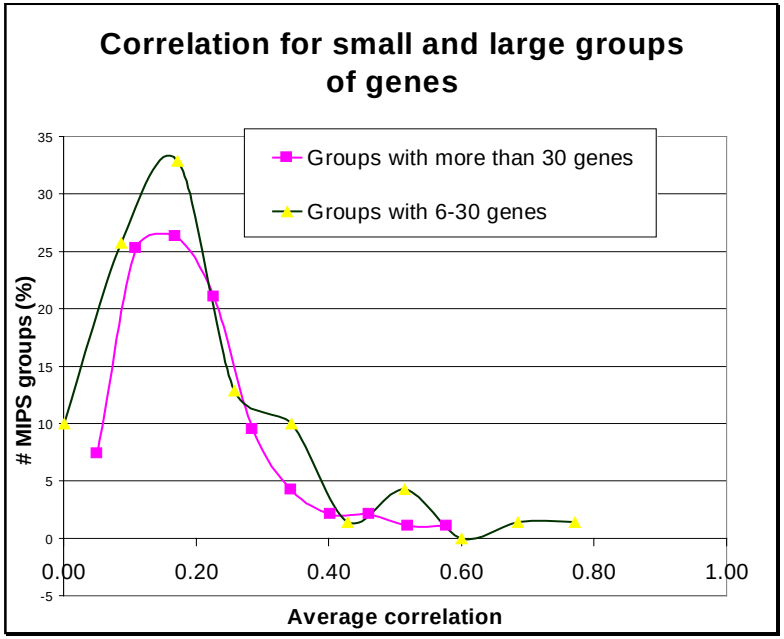
MIPS YFC: 66 bottom classes, 10 top classes
Average correlation of uncharacterized genes is 0.16
Similar to Botstein analysis.



Cluster Expression Data, Relate to MIPS Functional Category

Functional category number	Function	Average correlation	# ORFs
01	METABOLISM	0.1001	1005
01.01	amino-acid metabolism	0.1488	199
01.01.01	amino-acid biosynthesis	0.239	114
01.01.04	regulation of amino-acid metabolism	0.23	32
01.01.07	amino-acid transport	0.1198	23
01.01.10	amino-acid degradation	0.0524	30
01.01.99	other amino-acid metabolism activities	0.2205	1
01.02	nitrogen and sulphur metabolism	0.1869	73
01.02.01	nitrogen and sulphur utilization	0.0726	37
01.02.04	regulation of nitrogen and sulphur utilization	0.3715	28
01.02.07	nitrogen and sulphur transport	0.2829	8
01.03	nucleotide metabolism	0.1708	1
01.03.01	purine-ribonucleotide metabolism	0.3639	42
01.03.04	pyrimidine-ribonucleotide metabolism	0.176	28
01.03.07	deoxyribonucleotide metabolism	0.1095	12
01.03.10	metabolism of cyclic and unusual nucleotides	0.2848	8
01.03.13	regulation of nucleotide metabolism	0.2696	13
01.03.16	polynucleotide degradation	0.2461	20
01.03.19	nucleotide transport	0.1187	12
01.03.99	other nucleotide-metabolism activities	-0.032	7
01.04	phosphate metabolism	0.13	31
01.04.01	phosphate utilization	0.2	13
01.04.04	regulation of phosphate utilization	0.0599	8
01.04.07	phosphate transport	0.0724	10
01.05	carbohydrate metabolism	-0.0779	409
01.05.01	carbohydrate utilization	0.075	256
01.05.04	regulation of carbohydrate utilization	0.1174	120

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu



Correlation of Functional Class and Expression – not that strong

- Many groups of genes categorized by MIPS do not have higher correlation than random ORFs
- Smaller groups tend to have a slightly higher correlation

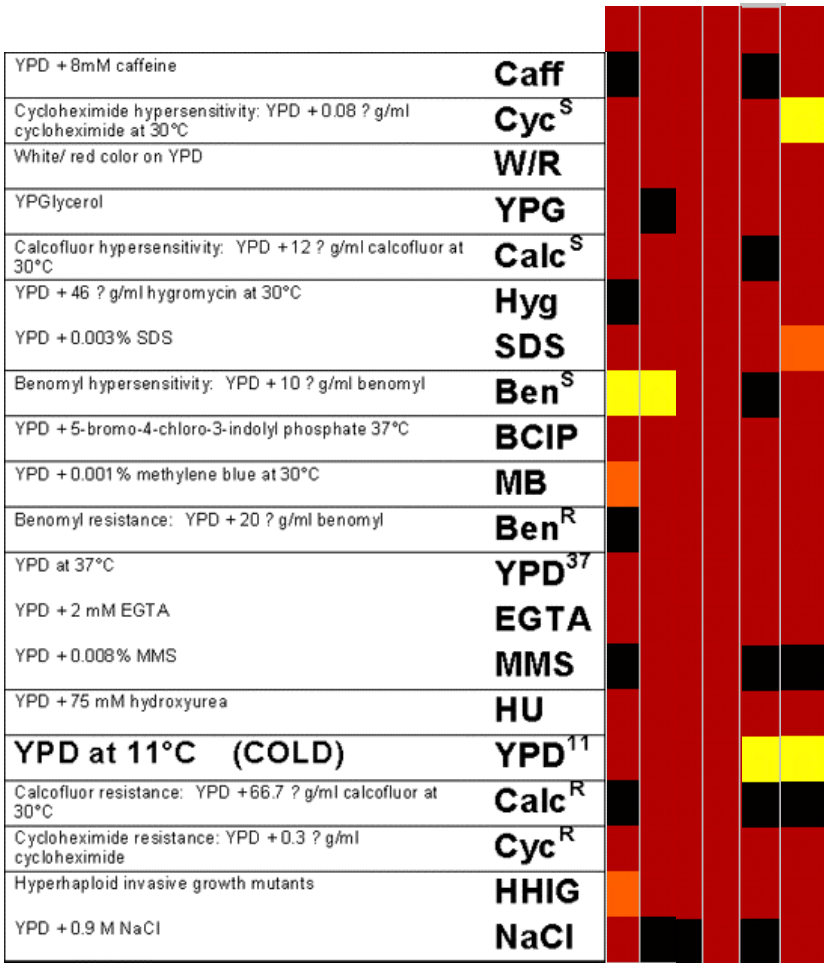
Highest Correlation

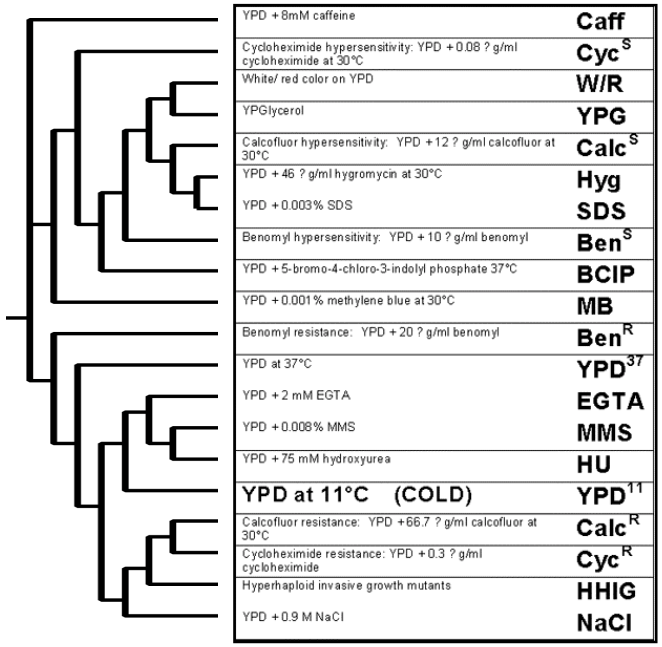
Functional category number	Function	Average correlation	# ORFs
10.04.11	key kinases	0.9403	2
10.04.13	key phosphatases	0.9283	2
11.11	ageing	0.8634	2
02.22	glyoxylate cycle	0.8136	6
10.02.07	G-proteins	0.8122	3
04.03.99	other tRNA-transcription activities	0.6932	4
09.08	biogenesis of Golgi	0.6647	2
09.19	peroxisomal biogenesis	0.6512	2
08.10	peroxisomal transport	0.646	12
04.01.04	rRNA processing	0.6074	53
01.20	secondary metabolism	0.5921	4
01.20.05	amines metabolism	0.5921	4
10.05.11	key kinases	0.5549	4
90	RETROTRANSPOSONS AND PLASMID PROTEINS	0.5299	7
02.10	tricarboxylic-acid pathway	0.5236	22
04.07	RNA transport	0.5111	27

.yale.edu

Whole Genome Phenotype Profiles

Transposon insertions into (almost) each yeast gene to see how yeast is affected in 20 conditions. Generates a phenotype pattern vector, which can be treated **similarly to expression data**

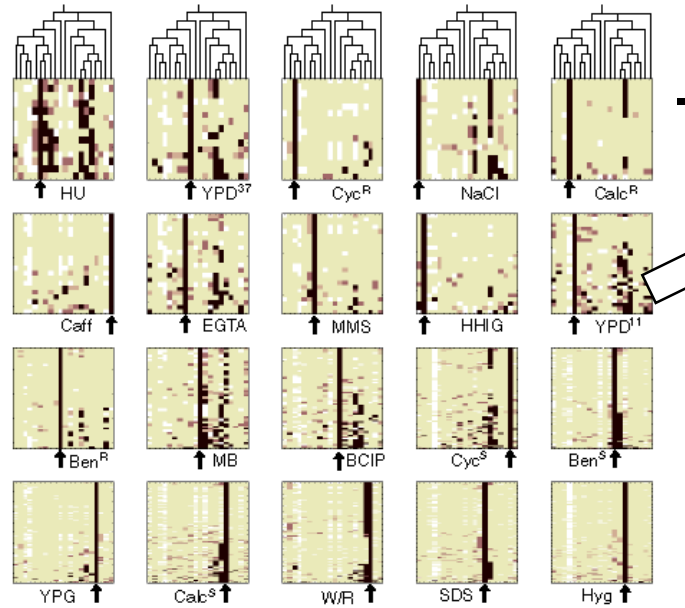


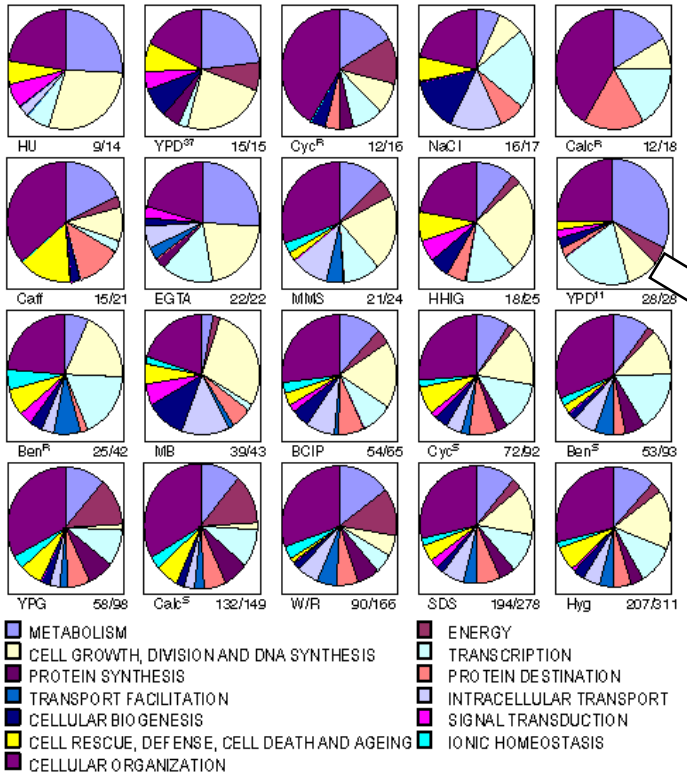


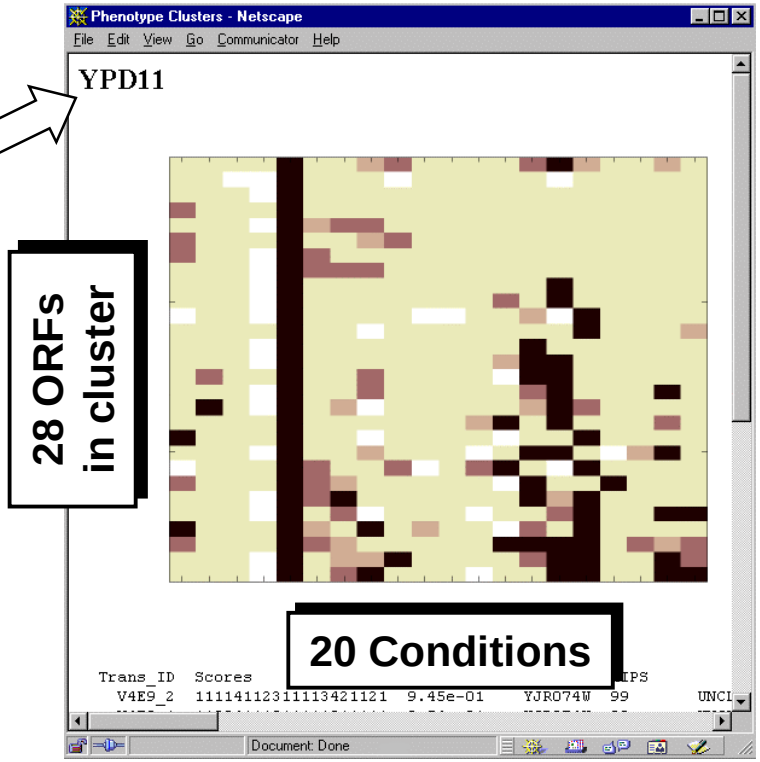
Clustering Conditions

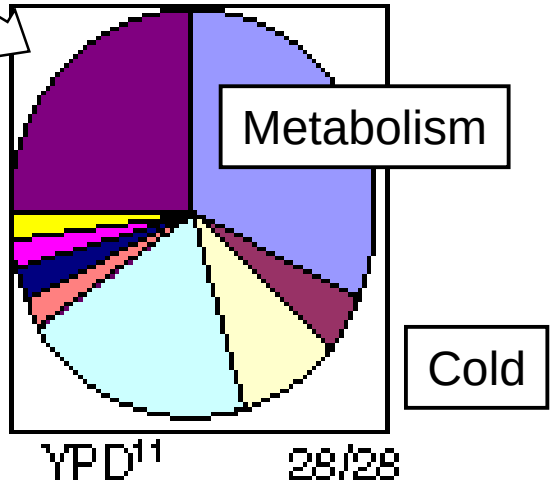
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Phenotype ORF Clustering







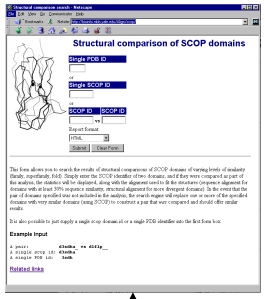


k-means clustering of ORFs based on "phenotype patterns," cross-ref. to MIPs Functional Classes

Cluster showing cold phenotype (containing genes most necessary in cold) is enriched in metabolic functions

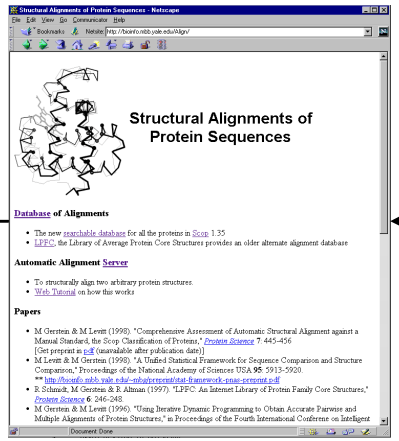
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

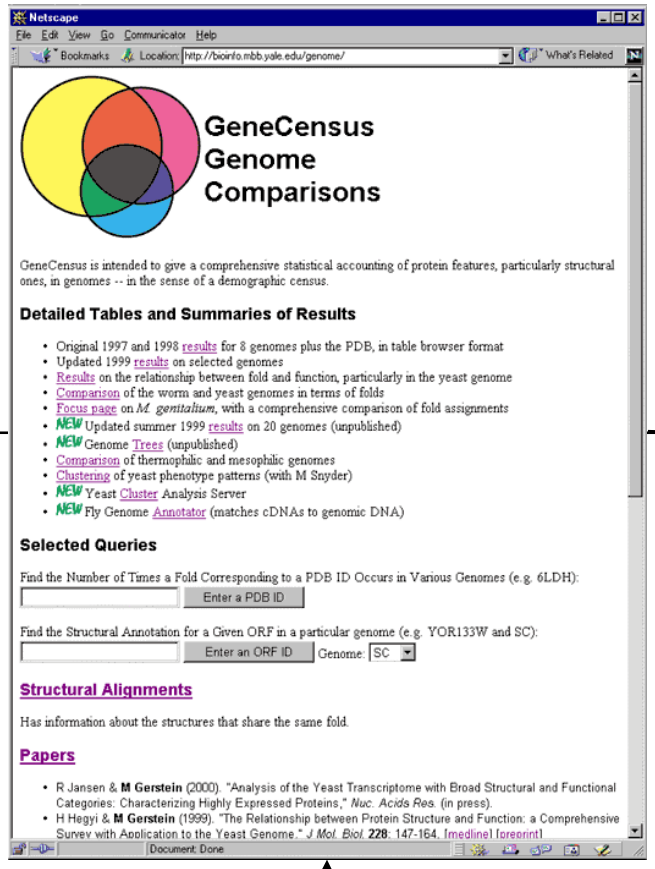
GeneCensus



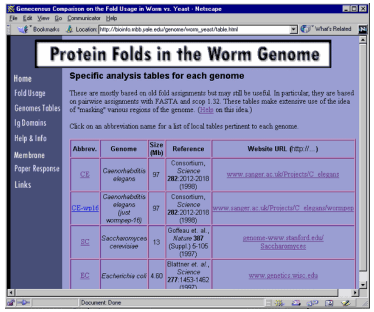
Alignment Database

Alignment Server

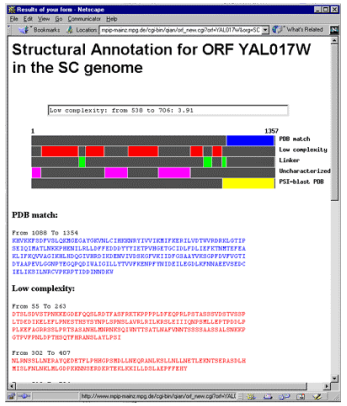




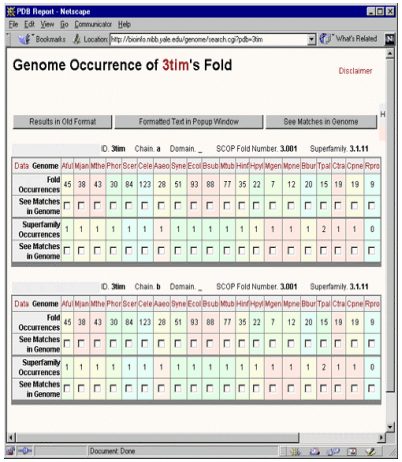
Detailed Tables



ORF Query



PDB Query



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

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

W Krebs, J Tsai, M Levitt, C Wilson, R Das, H Hegyi, J Lin, Y Kluger, C Arrowsmith, A Edwards, L Regan, S Balasubramanian, A Drawid, D Greenbaum, M Snyder, R Jansen

bioinfo.mbb.yale.edu

Structures ("Classic")

(now) Structural Genomics

(now) Func. Genomics

Arrays (future)

Structures ("Classic")

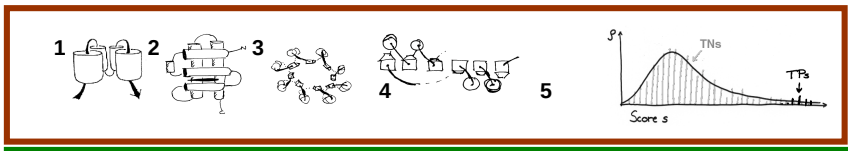
1 Fold Library (A parts list.) Structural Alignment, EVD P-value, Seq. Struc. diverg.

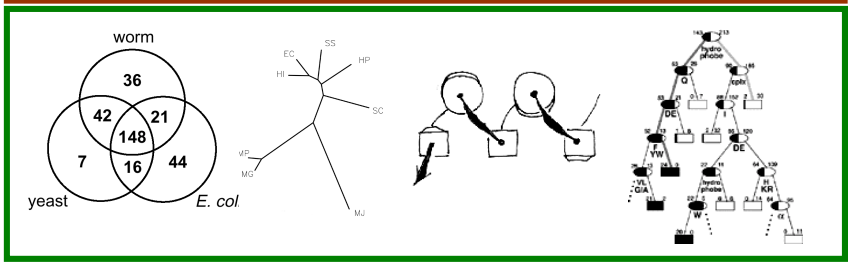
2 Folds in Genomes (Shared, common, and/or unique parts?) Known Folds. Fold Tree, Top-10. βαβ. Biases. MG fold assignment extent. MG Target Selection, MT retrospective decision tree.

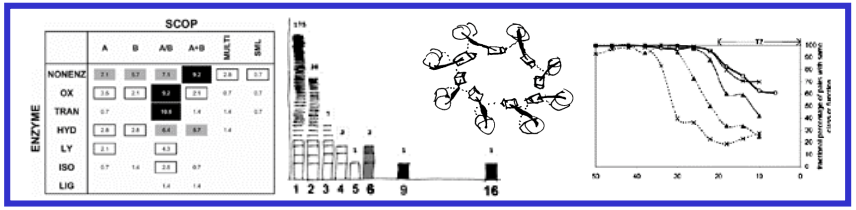
3 Folds & Functions (Roles/part?) How many folds /function? Mostly 1, but TIM versatile. Seq. diverg. vs. Func. diverg.

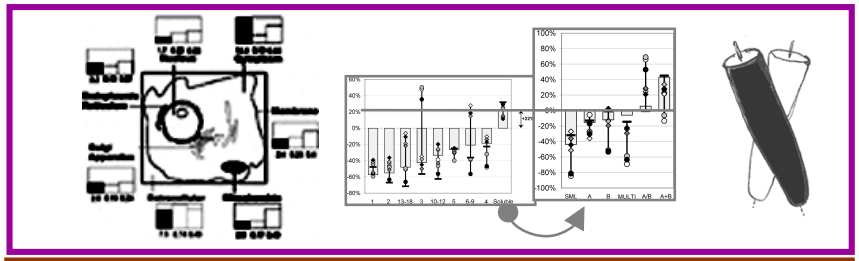
4 Folds in the Transcriptome (Common parts? Where are parts?) Enriched ↑ : VGA, TIM, αβ folds, energy, synthesis, cyt. Depleted ↓ : NS, long, TM folds, transport, transcription, Leu-zip, nuc. Bayesian Localizer, phenotypes clustering

5 Fold Flexibilty (How adaptable is a part?). Motions DB, morph server, interface packing, Voronoi Volumes



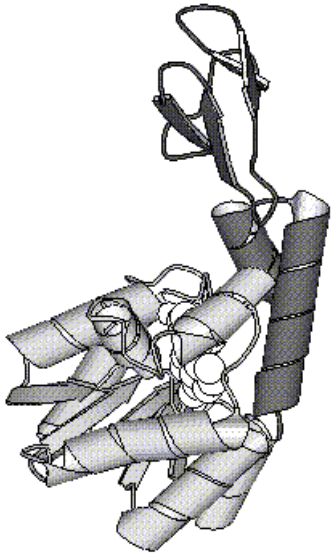


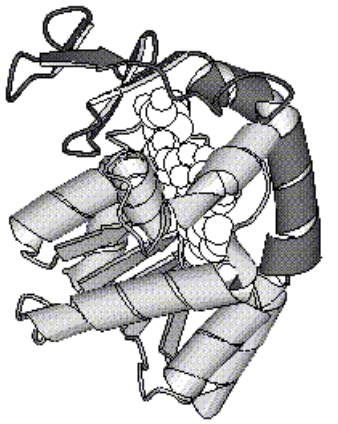


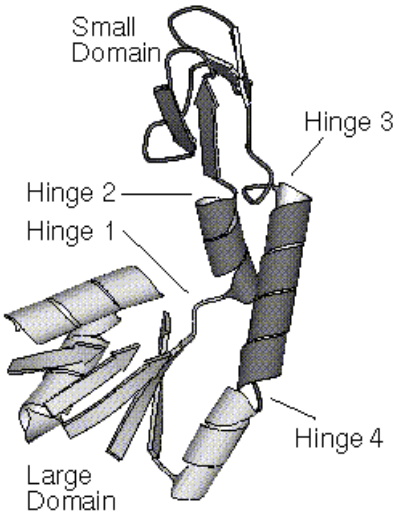


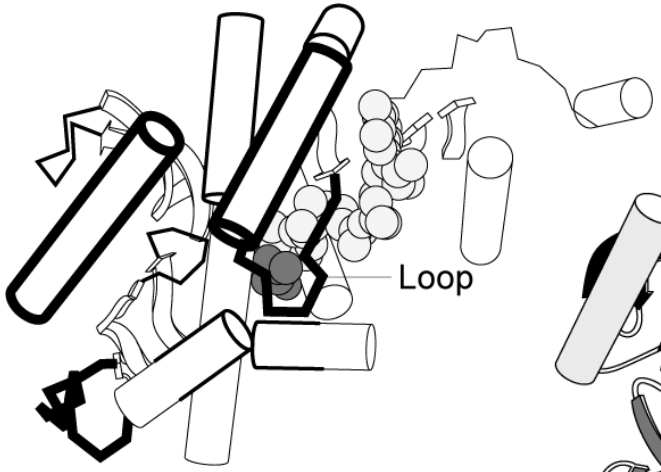
Macromolecular Motions

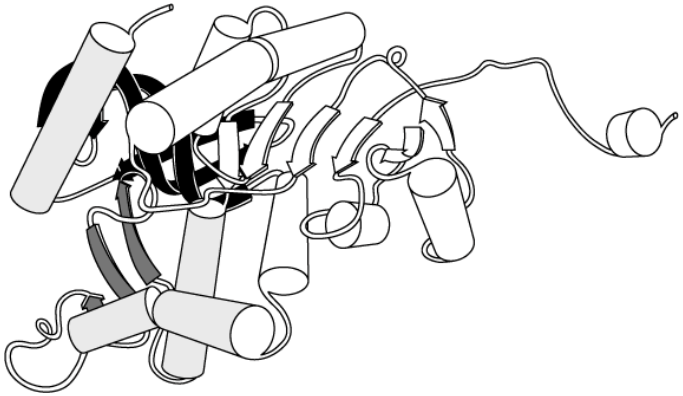
- What are they?
  - ◊ Proteins, Nucleic Acids (Hammerhead)
  - ◊ Sidechains (trivial), Loops (LDH), Domains (ADK), Subunits (Hb)
  - ◊ When a Ligand Binds: Open, Closed
- Essential link between structure and function
  - ◊ catalysis, regulation, transport, formation of assemblies, and cellular locomotion
- A complicated biological phenomena that can be studied in quantitative detail
  - ◊ changes in thousands of atomic coordinates









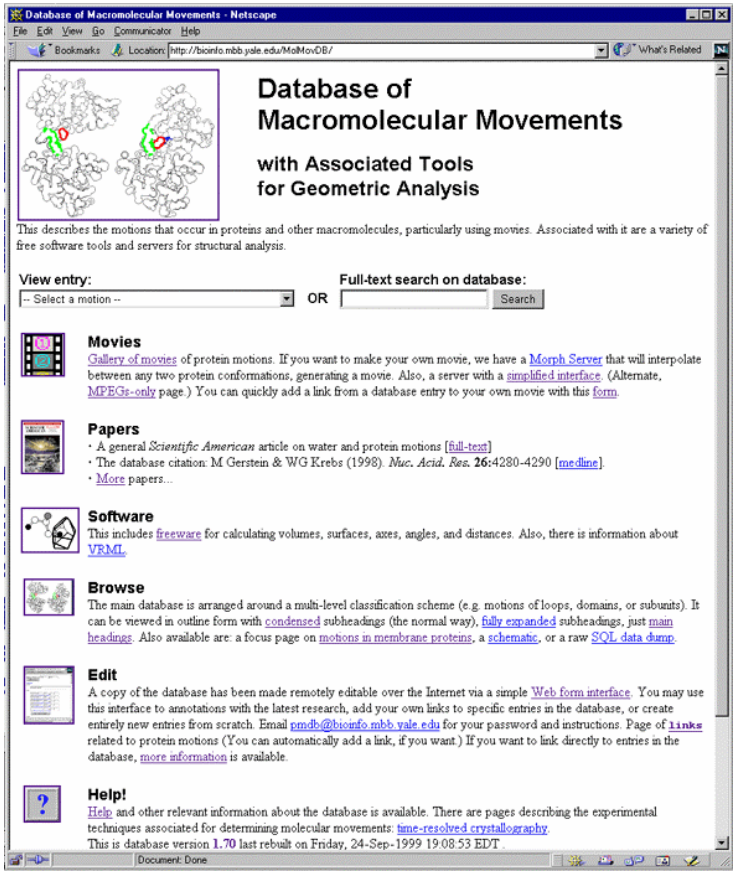


(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

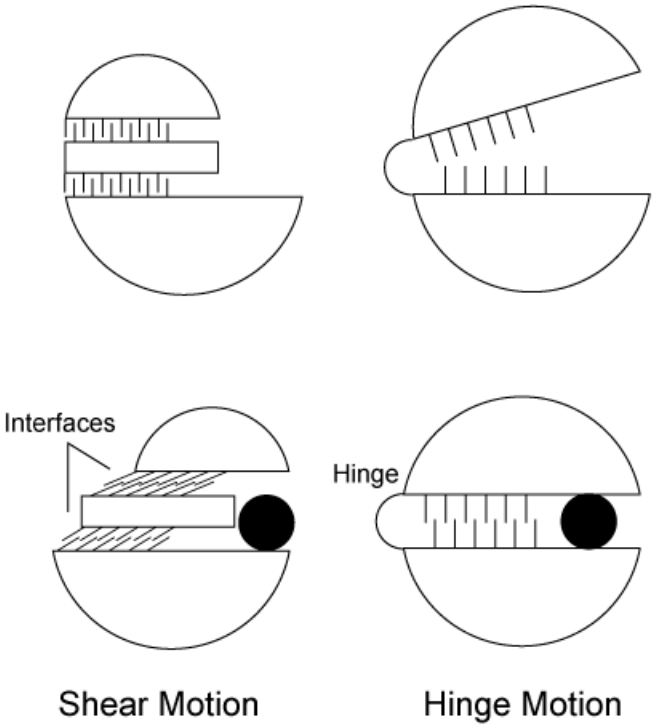
Motions DB: Information, Size, then Packing Based Classification

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

bioinfo.mbb.yale.edu/MolMovDB

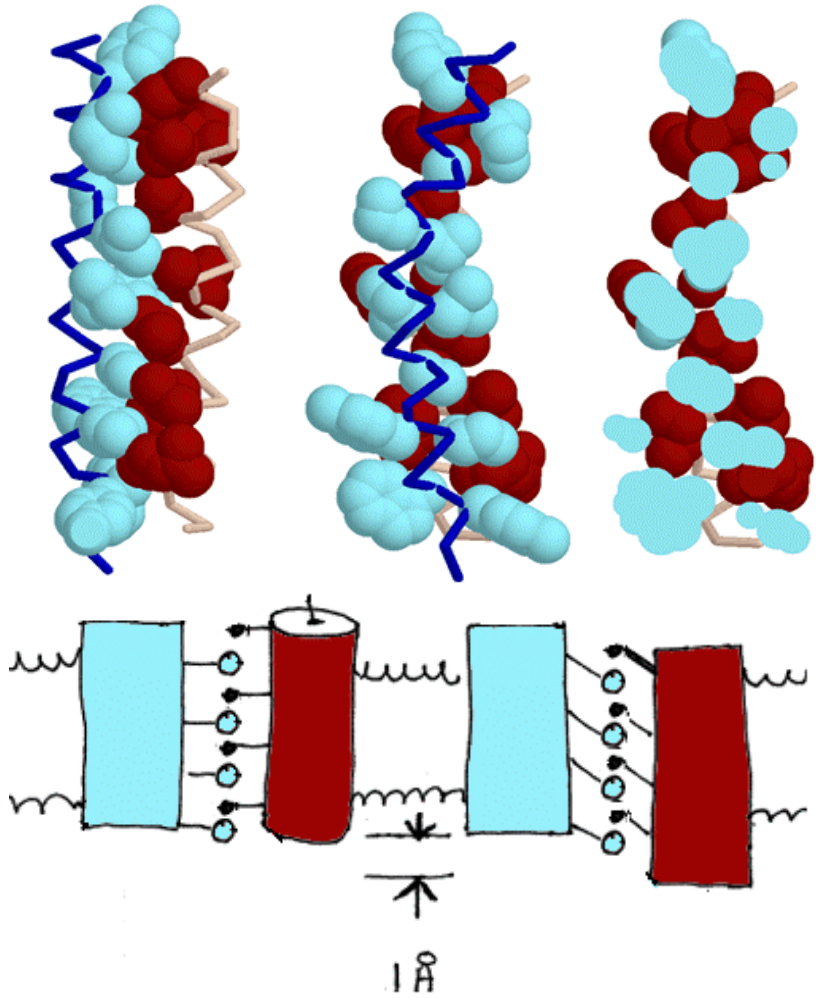


Interface Packing and Motions



Not applicable to allosteric motions, which are much slower (~1 ms) and do involve repacking interfaces

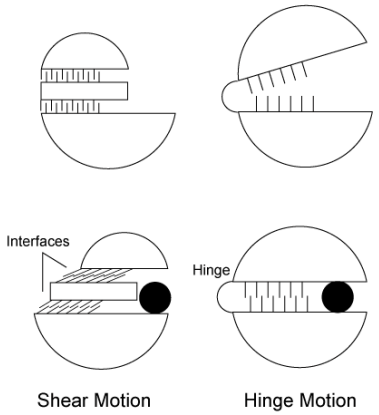
- Intercalating Interface, Knobs into Holes
- Packing is a strong constraint on motions
  - ◊ Domain or loop motions have to be fast (~10 ps – 100 ns)
  - ◊ Can't cross big energy barriers involved in repacking an interface

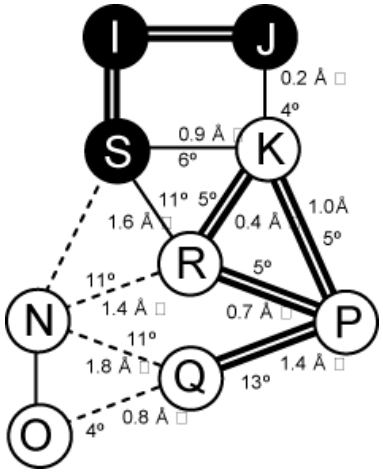


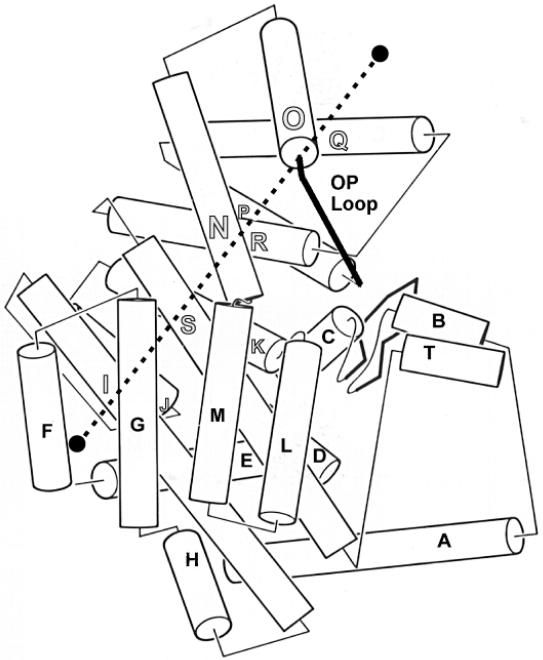
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

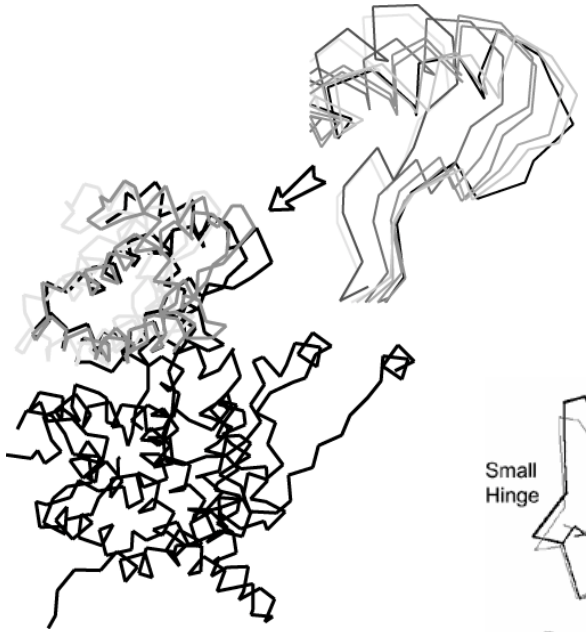
Packing Based Classification:

Hinge v ***Shear***

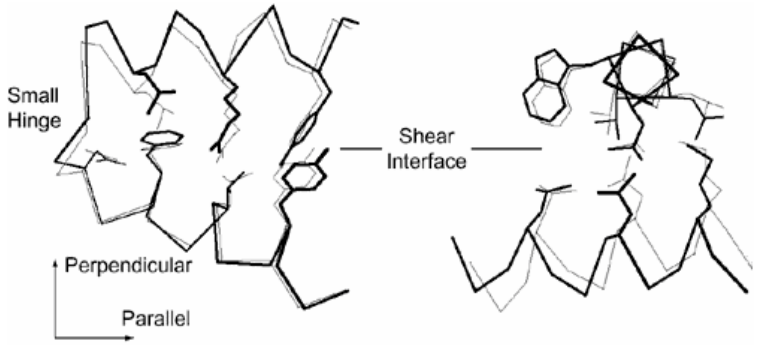






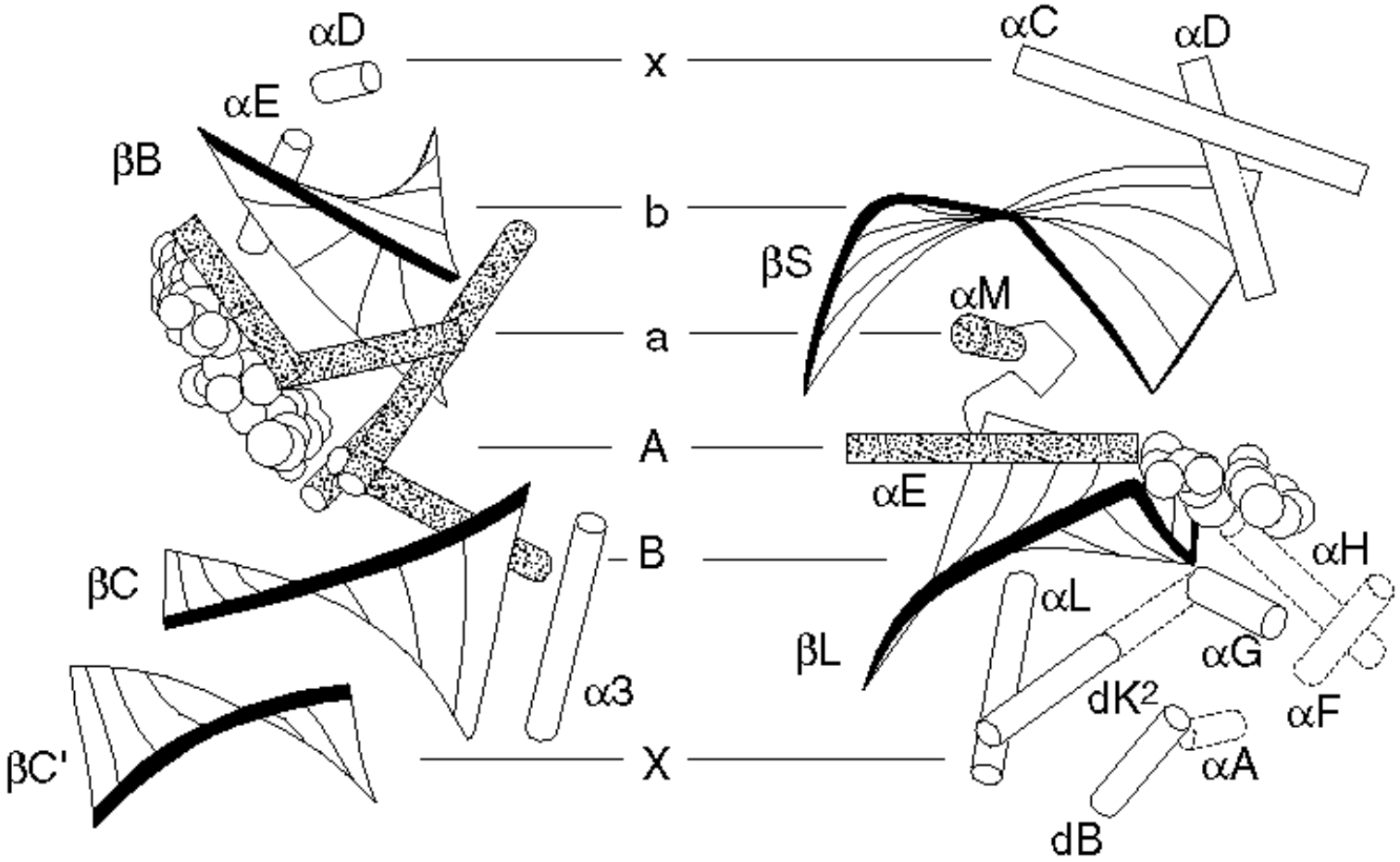


Shear Mechanism Involves Many Small Motions across a Continuously Maintained Interface



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Proteins With Shear Motions are Often Divided into Layers



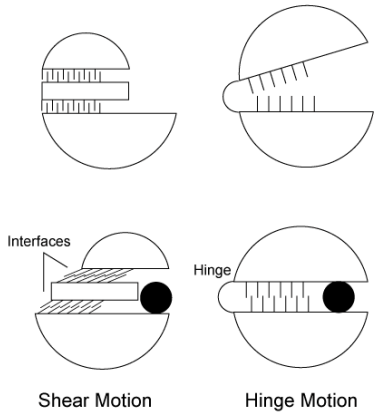
GAPDH

Hexokinase

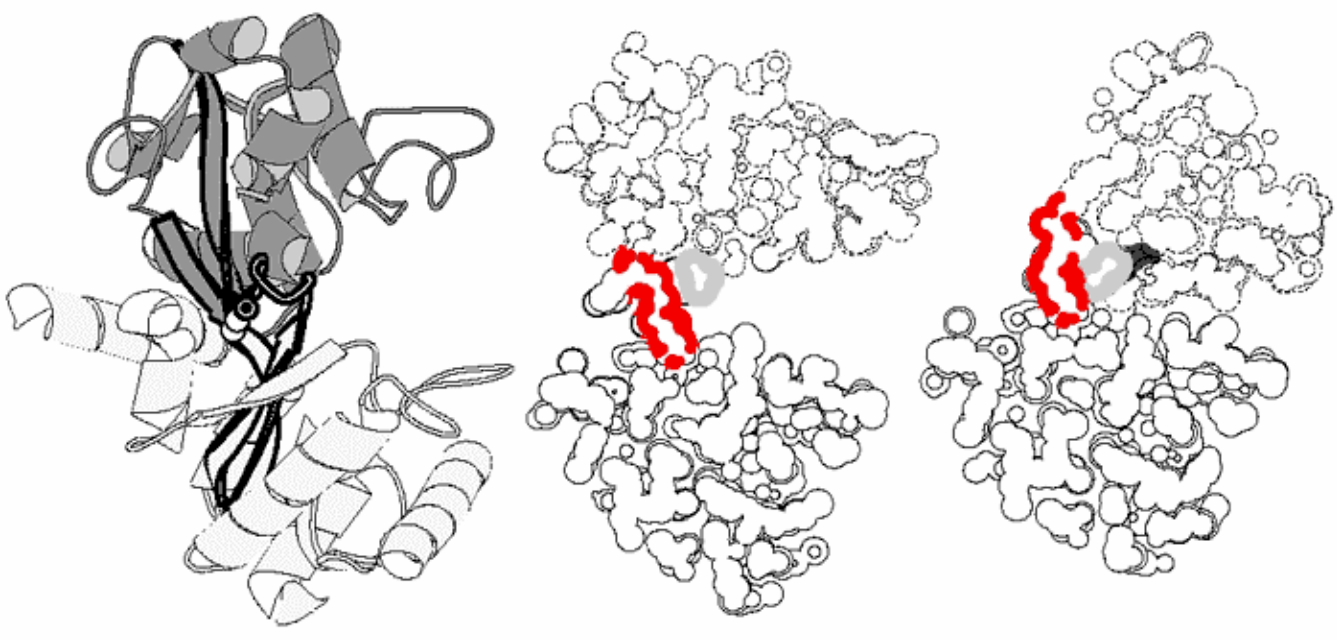
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

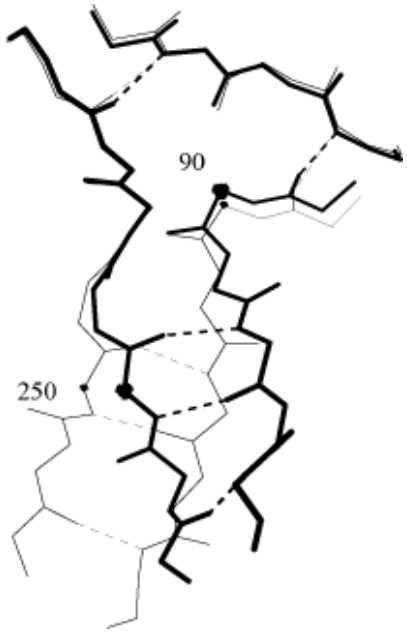
Packing Based Classification:

Hinge v Shear



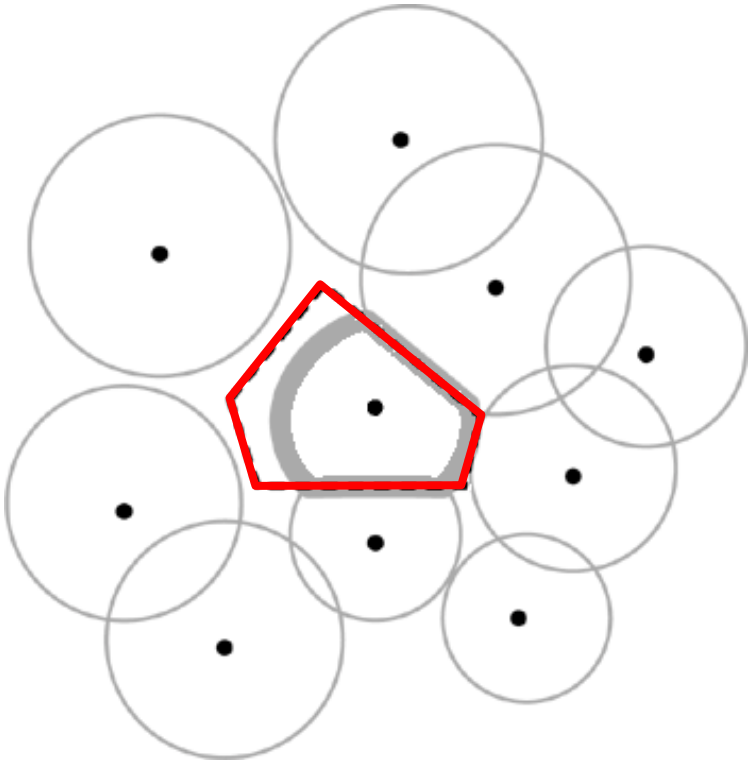
Hinge Mechanism involves absence of steric constraints (continuously maintained interface), esp. at hinge





(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Voronoi Volumes, the Natural Way to Measure Packing



- Each atom surrounded by a single convex polyhedron and allocated space within it
  - ◊ Allocation of all space (large V implies cavities)

Packing Efficiency

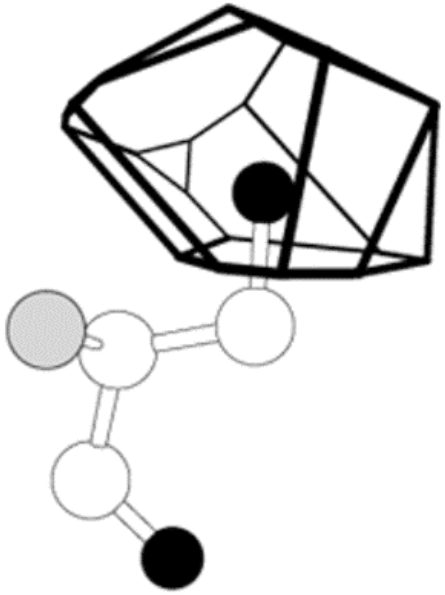
= Volume-of-Object

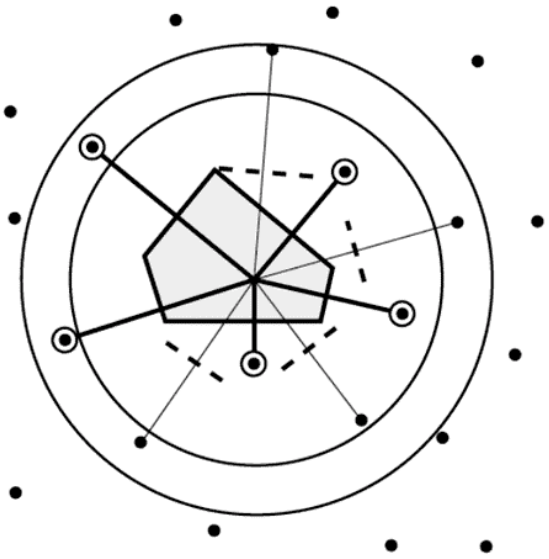
Space-it-occupies

= V(VDW) / V(Voronoi)

Absolute v relative eff.

V1 / V2





(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Unified Atoms		
atom	radii	volume
C3H0b	1.61	9.70
C3H0s	1.61	8.72
C3H1b	1.76	21.28
C3H1s	1.76	20.44
C4H1b	1.88	14.35
C4H1s	1.88	13.17
C4H2b	1.88	24.26
C4H2s	1.88	23.19
C4H3u	1.88	36.73
N3H0u	1.64	8.65
N3H1b	1.64	15.72
N3H1s	1.64	13.62
N3H2u	1.64	22.69
N4H3u	1.64	21.41
O1H0u	1.42	15.91
O2H1u	1.46	17.98
S2H0u	1.77	29.17
S2H1u	1.77	36.75

Residues	
aa	volume
Gly	63.8
Ala	89.3
Val	138.2
Leu	163.1
Ile	163.0
Met	165.8
Pro	121.6
His	157.5
Phe	190.8
Tyr	194.6
Trp	226.4
Cyh	112.8
Cys	102.5
Ser	94.2
Thr	119.6
Asn	112.4
Gln	146.9
Asp	114.4
Glu	138.8
Lys	165.1
Arg	190.3

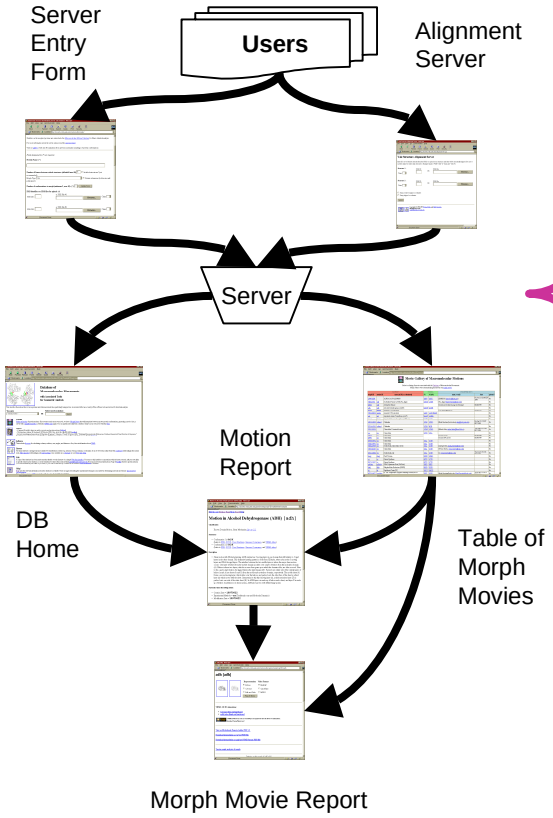
Parameters used in Protor Volume Derivation	
Typing Scheme	Hybrid chemical and numerical typing with 18 basic types
Radii Set	ProtOr Radii, Tsai et al. (1999)
Plane-Positioning Method	Ratio
Atom Selection Criteria	BL+
Structure Set	SCOP (87 structures)

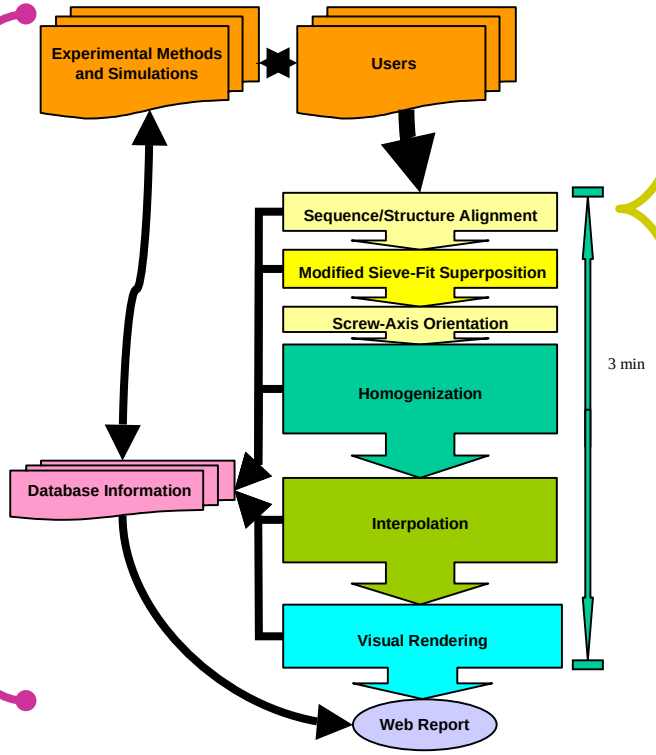
ProtOr Parameter Set: Standard Radii & Volumes

- Consistent Radii, Typing, and Volumes for Packing Calculations
- For comparison

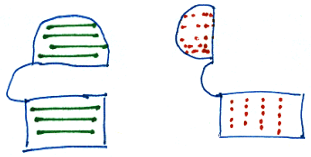
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Motion Analysis Server, Morph Movies

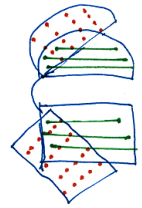




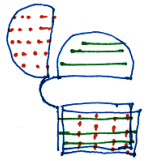
Struc-1 Struc-2



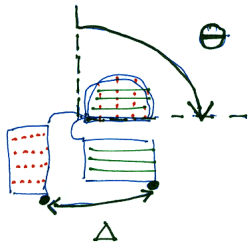
Overall Fit



Core-1 Fit

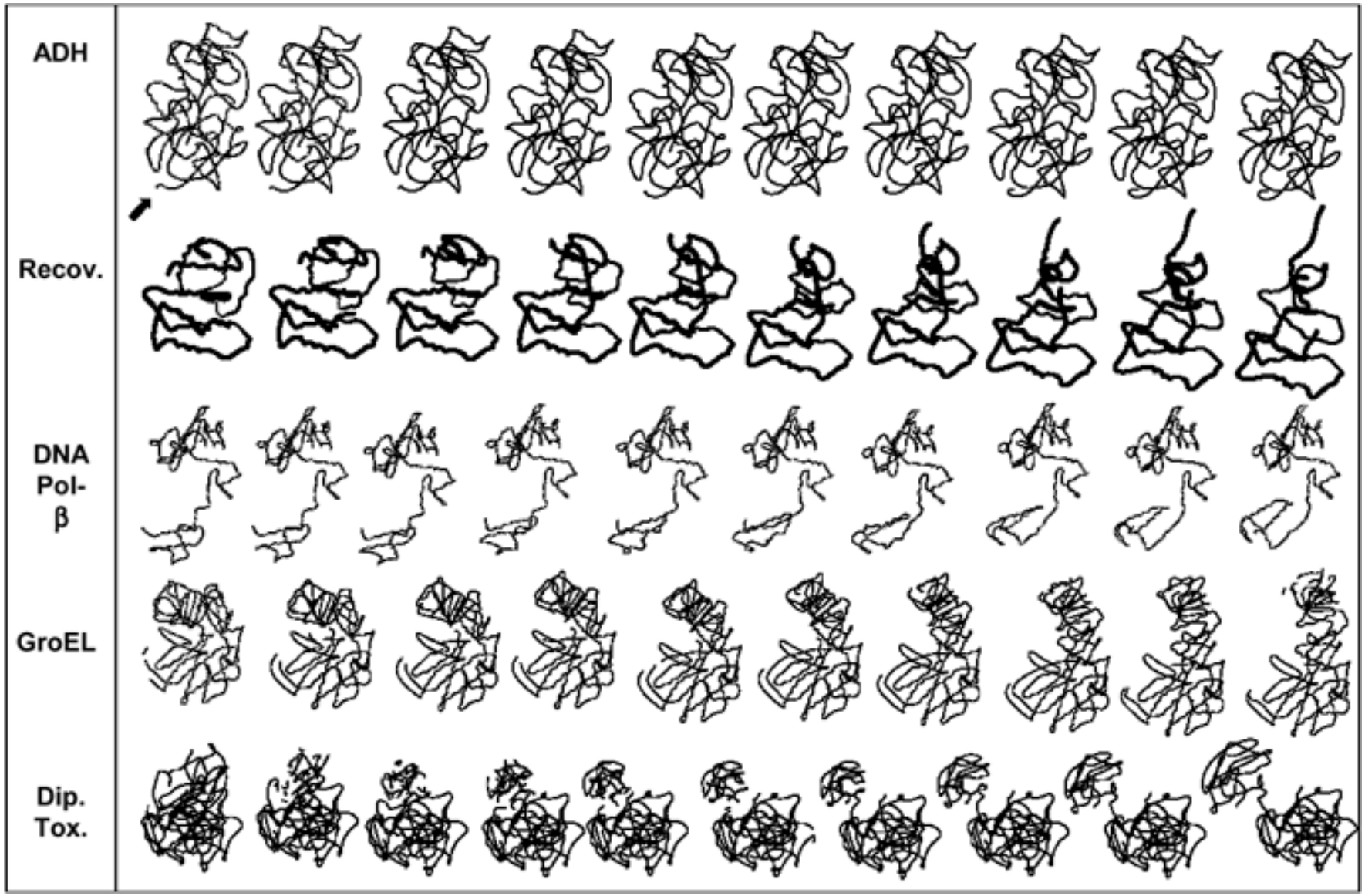


Core-2 Fit



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

5 Movies Generated by the Server



(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

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

W Krebs, J Tsai, M Levitt, C Wilson, Y Kluger, S Balasubramanin, L Regan, R Das, C Arrowsmith, A Edwards, H Hegyi, J Lin, A Drawid, D Greenbaum, M Snyder, R Jansen

bioinfo.mbb.yale.edu

Structures ("Classic")

(now) Structural Genomics

(now) Func. Genomics

Arrays (future)

Structures ("Classic")

1 Fold Library (A parts list.) Structural Alignment, EVD P-value, Seq. Struc. diverg.

2 Folds in Genomes (Shared, common, and/or unique parts?) Known Folds. Fold Tree, Top-10. βαβ. Biases. MG fold assignment extent. MG Target Selection, MT retrospective decision tree.

3 Folds & Functions (Roles/part?) How many folds /function? Mostly 1, but TIM versatile. Seq. diverg. vs. Func. diverg.

4 Folds in the Transcriptome (Common parts? Where are parts?) Enriched ↑ : VGA, TIM, αβ folds, energy, synthesis. Depleted ↓ : NS, very long, TM folds, transport, transcription, Leu-zip fold. Bayesian Localizer, phenotypes clustering

5 Fold Flexibilty (How adaptable is a part?). Motions DB, morph server, interface packing, Voronoi Volumes