Structural Genomics:

Surveys of a Finite Parts List

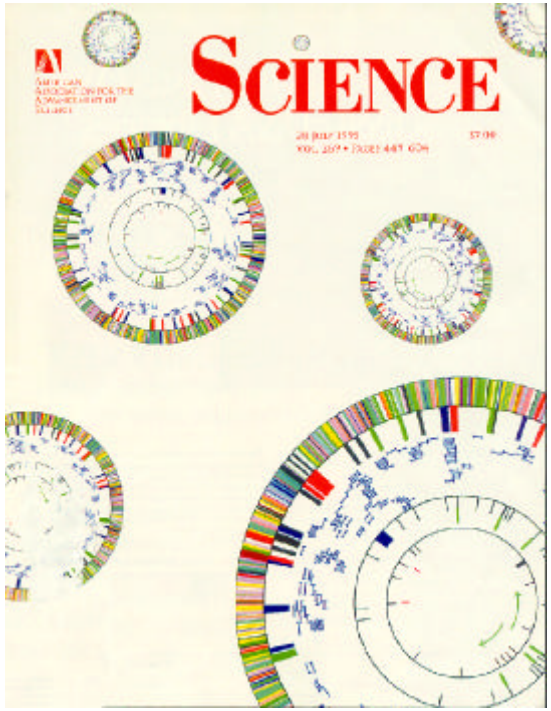
Mark Gerstein

H Hegyi, J Lin, B Stenger,
N Echols, P Bertone, J Qian,
L Regan, S Balasubramanian,
V Alexandrov, G Montelione,
A Edwards, C Wilson, Y Kluger,
C Arrowsmith, A Drawid,
R Jansen, D Greenbaum,
S Teichmann, P Harrison

Talk at DIMACS, 15 Dec. 2000

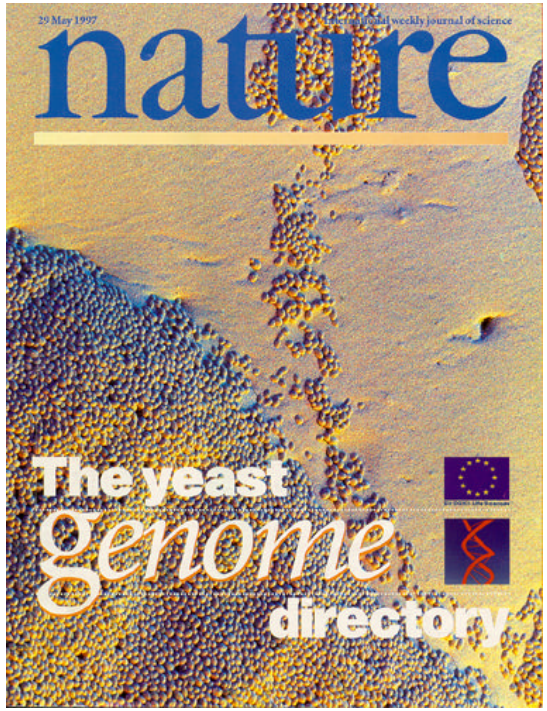
1995

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



1997

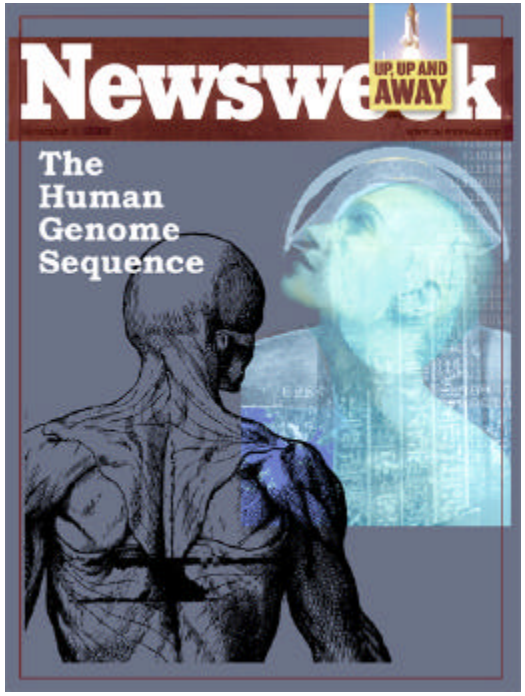
Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]



Genomes highlight the **Finiteness** of the "Parts" in Biology

1998

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



↔

2000?

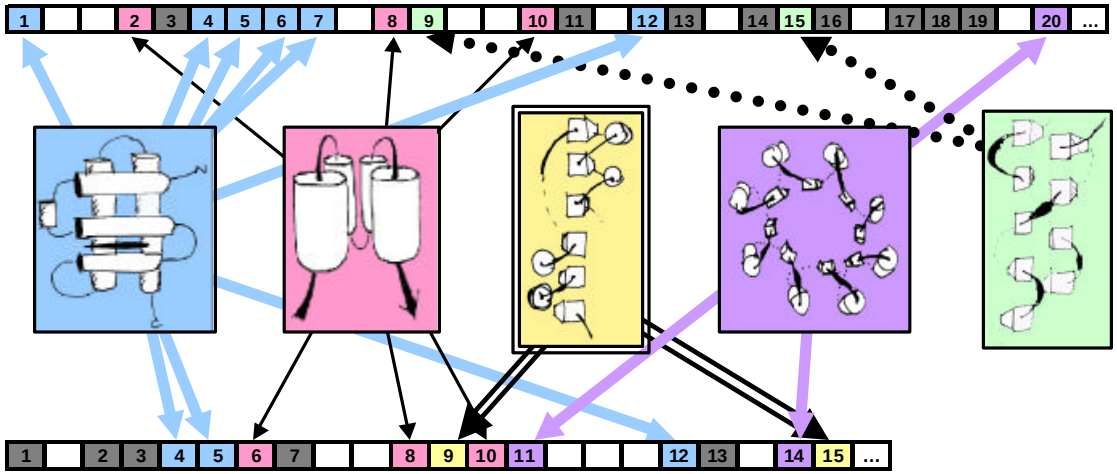
Human,
~3 Gb,
~100K genes [???]

'**98** spoof real thing, Apr '**00**

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

(human)

(*T. pallidum*)



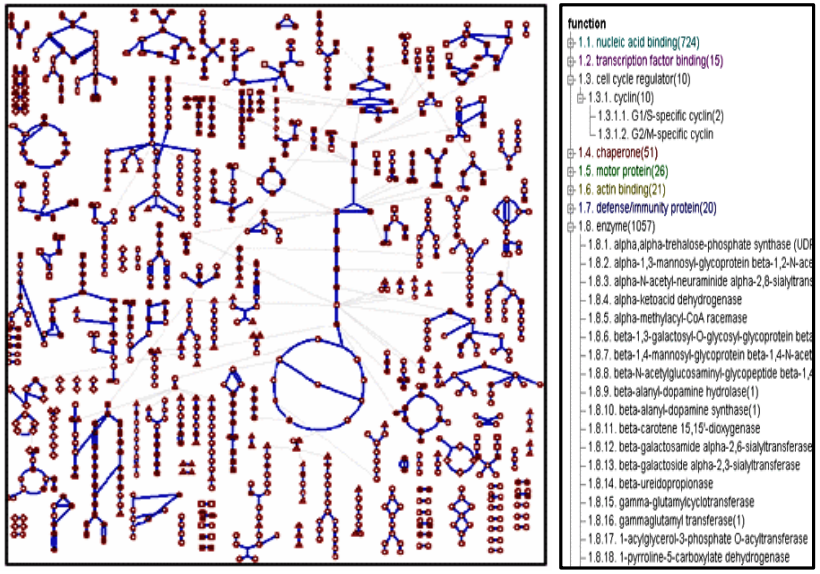
~100000 genes

~1000 folds

~1000 genes

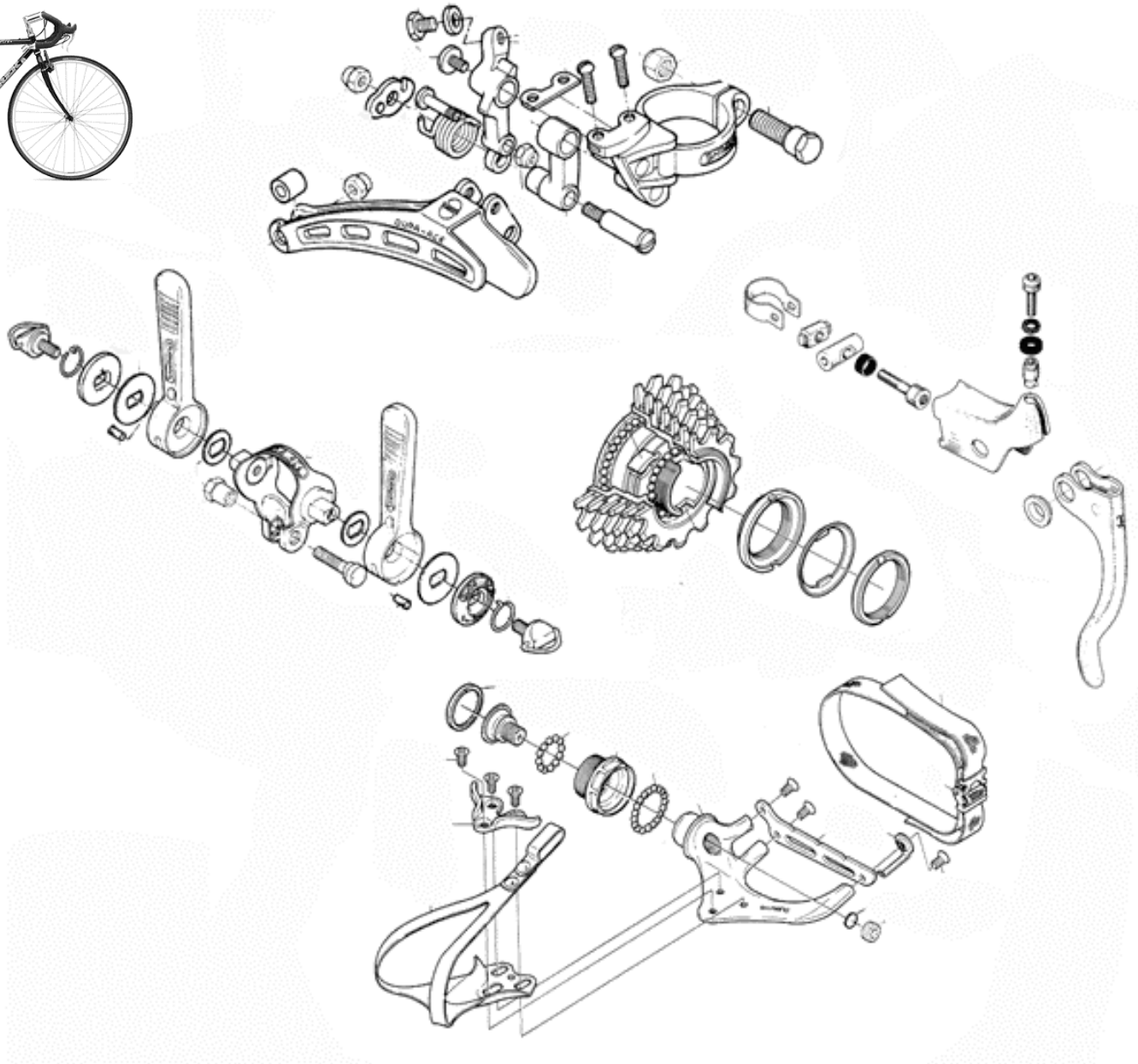
Same logic for pathways, functions, sequence families, blocks, motifs....

Global Surveys of a Finite Set of Parts from Many Perspectives

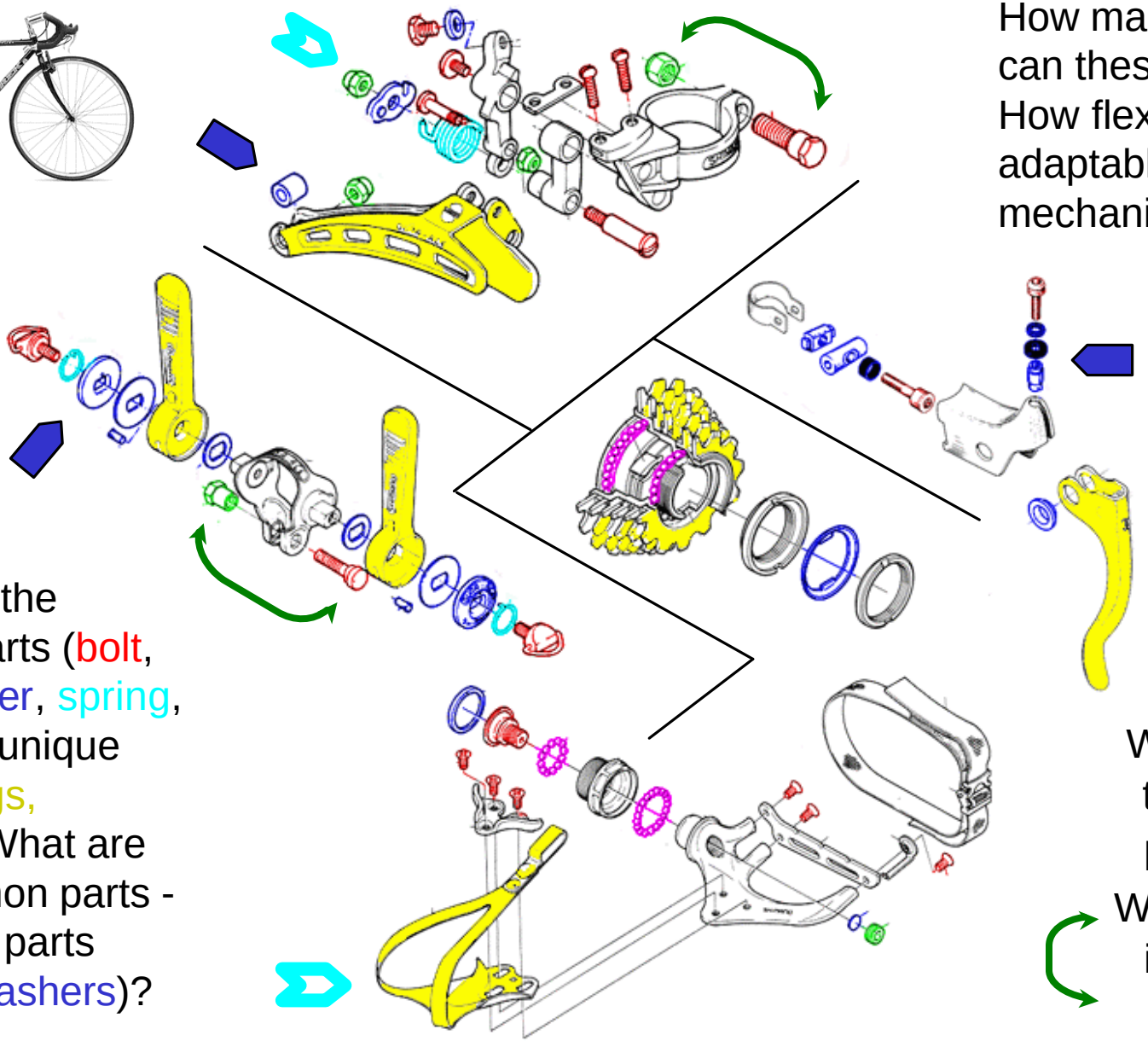


Functions picture from www.fruitfly.org/~suzi (Ashburner); Pathways picture from, ecocyc.pangeasystems.com/ecocyc (Karp, Riley). Related resources: COGS, ProDom, Pfam, Blocks, Domo, WIT, CATH, Scop....

A Parts List Approach to Bike Maintenance



A Parts List Approach to Bike Maintenance

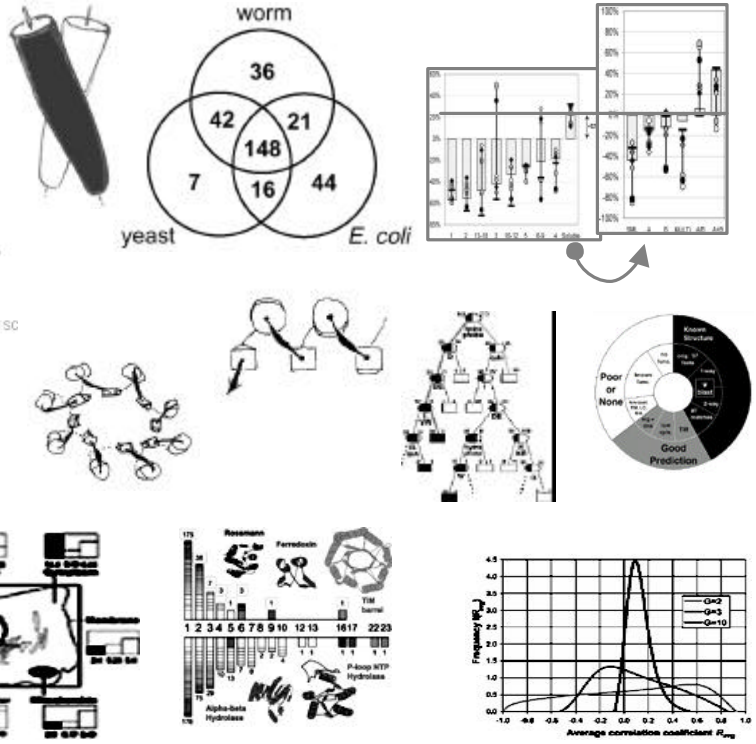


How many roles can these play? How flexible and adaptable are they mechanically?

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? Which parts interact?

Comparative Genomics: Surveys of a Finite Parts List



1 Using Folds to Interpret Genomes. Fold Library background. Shared and/or unique parts. Venn Diagram, Fold tree with all-β diff. Ortholog tree. Horizontal Transfer. Commor Parts: Top-10 folds with βαβ. Common Ψ-fold.

2 Tricky Issues →Expt. SG. Extent of fold assignment (MG, 65%, 98). Predictions. Biases. NESG.org Construct DB. Datamining this w/ decision trees. Selecting weird MG CD targets.

3 Folds & Functions. Roles/part? How many folds/func? 331 of ~20K combinations. Mostly 1 func/fold, but some versatile scaffolds -- TIM most versatile. Similar for interactions. Func. Divergence vs. Seq. & Struc. Diverg.

4 Using Folds to Interpret Expression Data. Top-10 parts in other terms. Enriched in transcriptome: VGA, αβ folds, energy, synthesis, Cytoplasmic, TIM fold. Depleted: NS, long, TMs, transport, transcription, Nuclear, Leu-zip fold. Bayesian localizer.

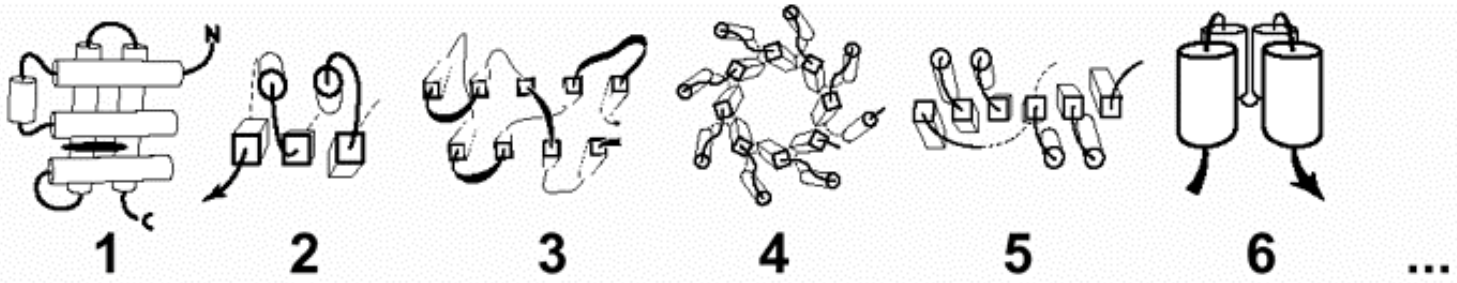
5 Tricky Issues: Relating Expression to Function. Expression relates to structure & localization but to function, globally? Weak relation to protein-protein interactions.

H Hegyi, J Lin, B Stenger, N Echols, P Bertone, J Qian, L Regan, S Balasubramanian, V Alexandrov, G Montelione, A Edwards, C Wilson, Y Kluger, C Arrowsmith, A Drawid, R Jansen, D Greenbaum, S Teichmann, P Harrison

bioinfo.mbb.yale.edu

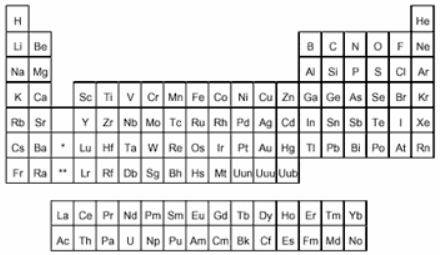
Fold Library vs. Other Fundamental Data structures

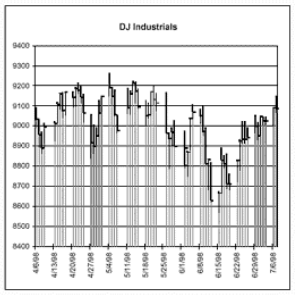
Parts List **Database**; **Statistical**, rather than mathematical relationships and conclusions



Folds in Molecular Biology **1000-10000**

const.	mant.	exp.	unit
e	1.60	8	C
F	9.65	4	C/mol
ε_0	8.85	-12	F/m
μ_0	1.26	-6	H/m
h	6.63	-34	J •s
k	1.38	-23	J/K
m_e	9.11	-31	kg
m_p	1.67	-27	kg
m_n	1.68	-27	kg
a_0	5.29	-11	m
λ_C	2.43	-12	m
c	3.00	9	m/s
G	6.67	-11	$m^3/kg•s^2$
N_A	6.02	23	mol^{-1}





10 Physics

100 Chemistry

1000 -10000 Finance

>1000000 Politics

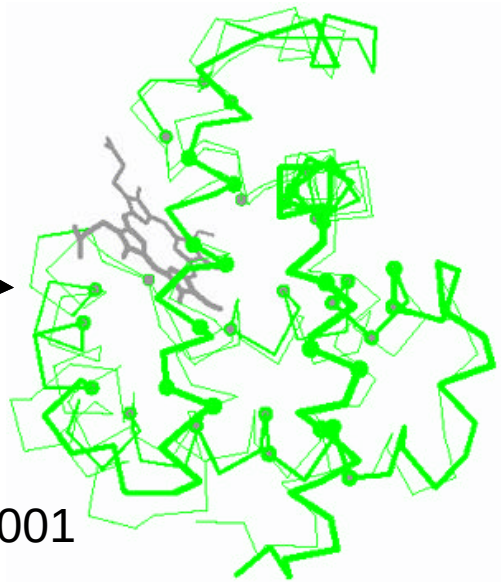
(Large than physics and chemistry, Similar to Finance (Exact Finite Number of Objects (3,056 on NYSE by 1/98), descrip. by Standardized Statistics (even abbrevs, INTC) and groups (sectors)) Smaller than Social Surveys, Indefinite Number of People, Not Well Defined Vocabulary and statistics.

The Parts List: A Library of Known Folds

Hb

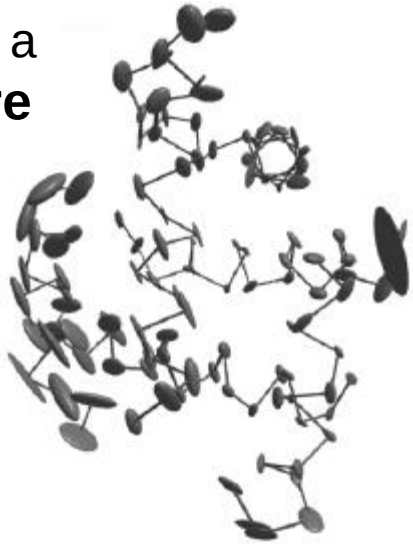
Mb

Alignment of Individual Structures

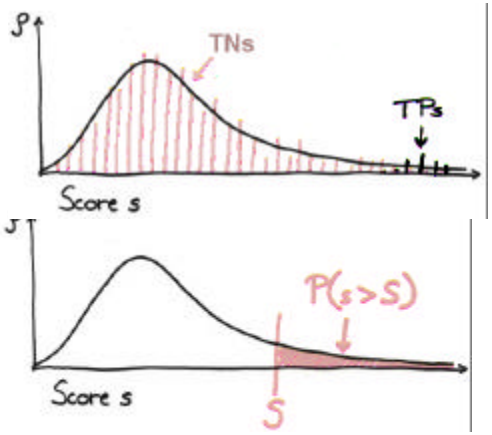


P<.000001

Fusing into a Single **Core** Structure Template

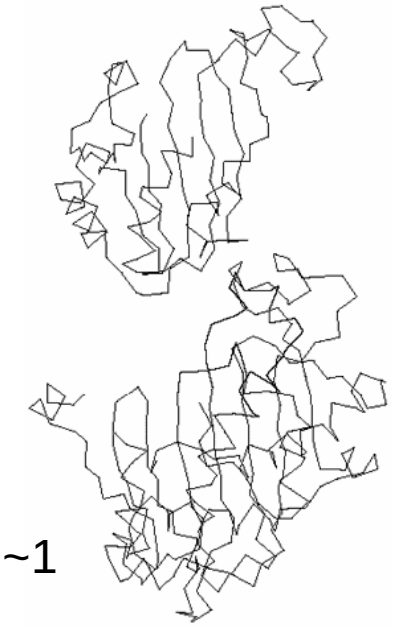


Statistics to Establish Relation-ships (P-values)



P<.001

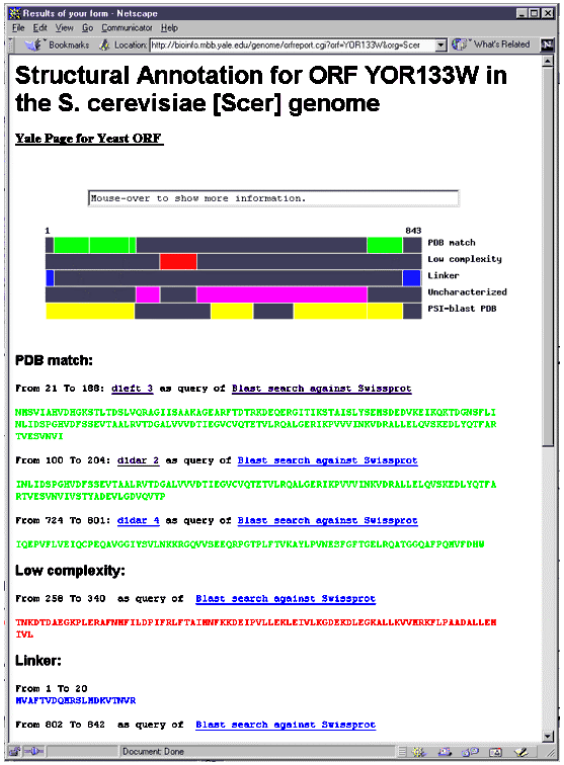
P~1



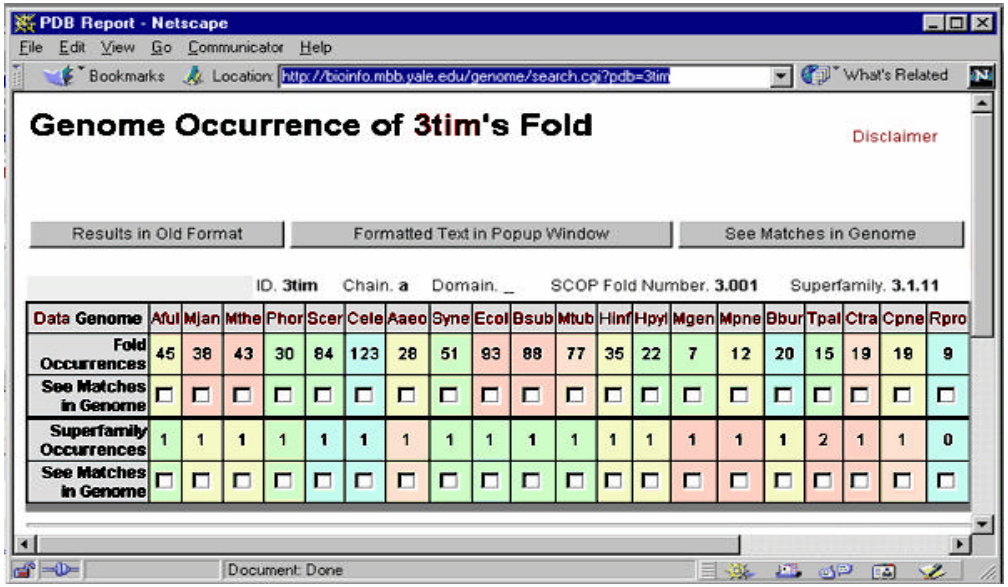
Integrated Analysis System: X-ref Parts with Genomes

Folds: scop+automatic
Orthologs: COGs
"Families": homebrew, ProtoMap

finding parts in genome sequences **blast**, ψ-blast, **fasta**, TM, low-complexity, &c (Altschul, Pearson, Wooton)



part occurrence profiles



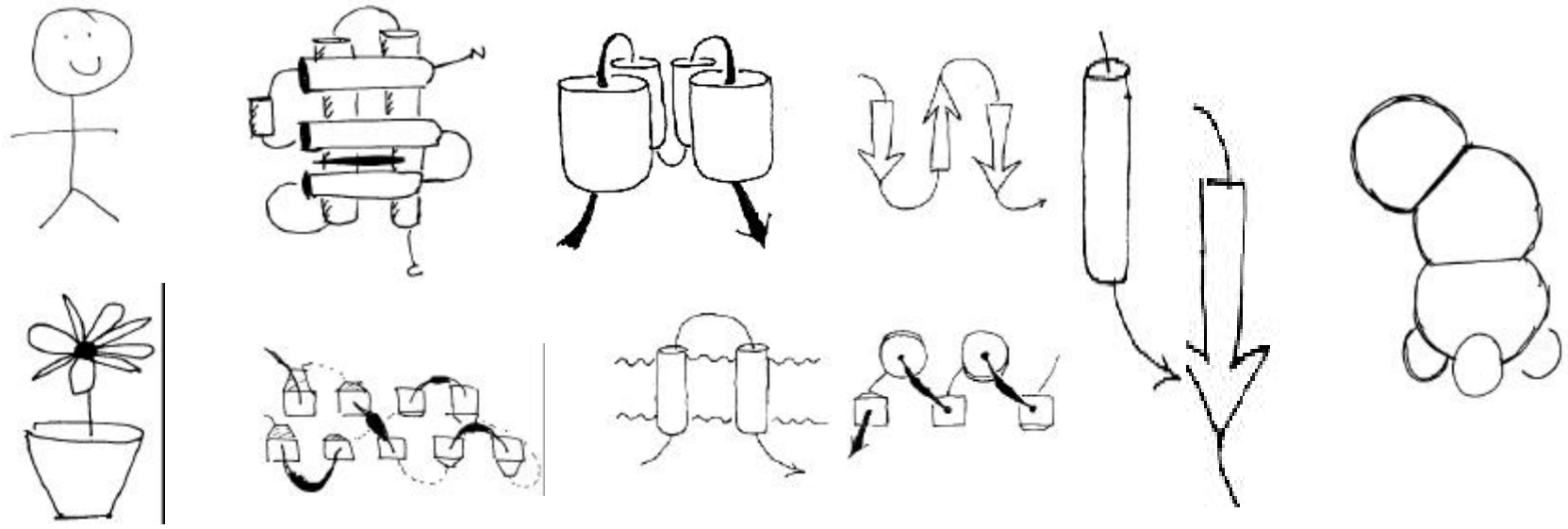
Data Genome	Aful	Mjan	Mthe	Phor	Scer	Cele	Aaeo	Syne	Ecol	Bsub	Mtub	Hinf	Hpyl	Mgen	Mpne	Bbur	Tpal	Ctra	Cpne	Rpro
Fold Occurrences	45	38	43	30	84	123	28	51	93	88	77	35	22	7	12	20	15	19	19	9
See Matches in Genome	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Superfamily Occurrences	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	0
See Matches in Genome	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

One approach of many...
Much previous work on Sequence & Structure Clustering
CATH, Blocks, FSSP, Interpro, eMotif, Prosite, CDD, Pfam, Prints, VAST, TOGA…

Remington, Matthews '80; **Taylor, Orengo '89**, '94; Thornton, CATH; Artymiuk, Rice, Willett '89; Sali, Blundell, '90; Vriend, Sander '91; Russell, Barton '92; **Holm, Sander '93+ (FSSP)**; Godzik, Skolnick '94; Gibrat, Bryant '96 (VAST); F Cohen, '96; Feng, Sippl '96; G Cohen '97; Singh & Brutlag, '98

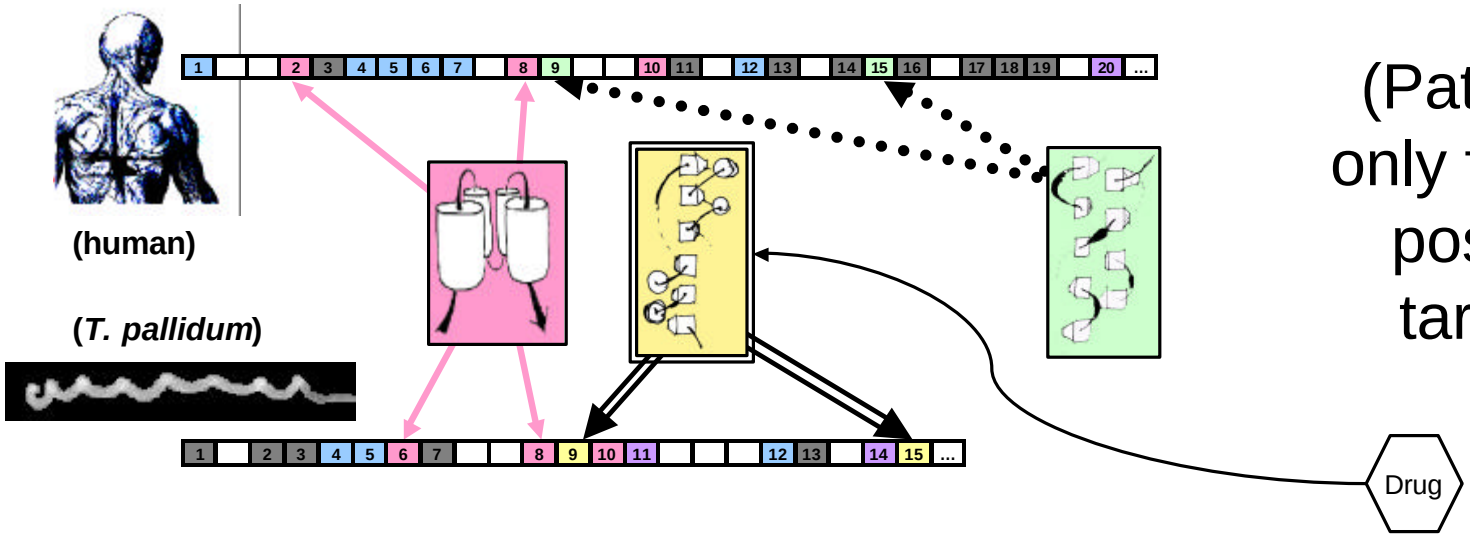
At What Structural Resolution Are Organisms Different?

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

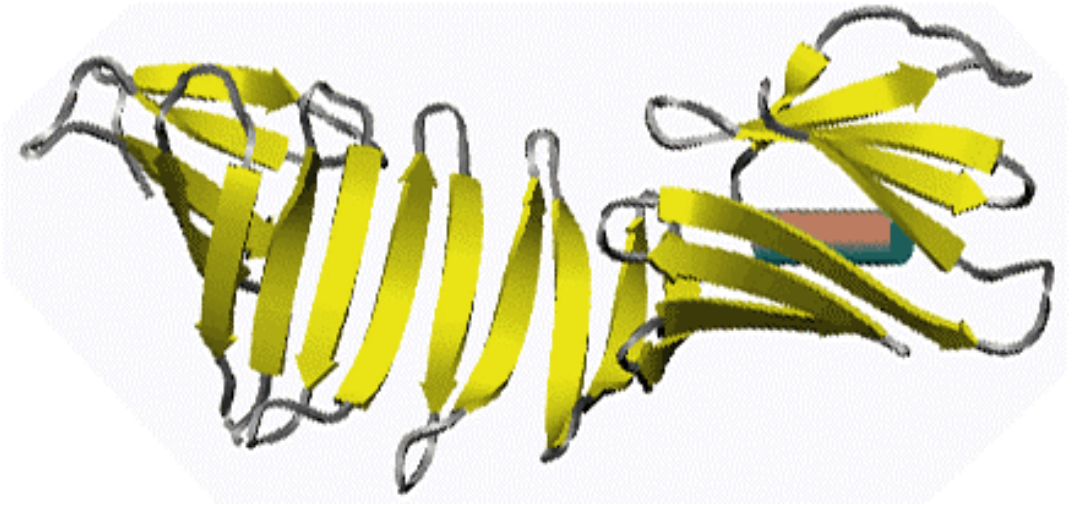


1m 100Å 10Å 1Å

Practical Relevance of Structural Genomics

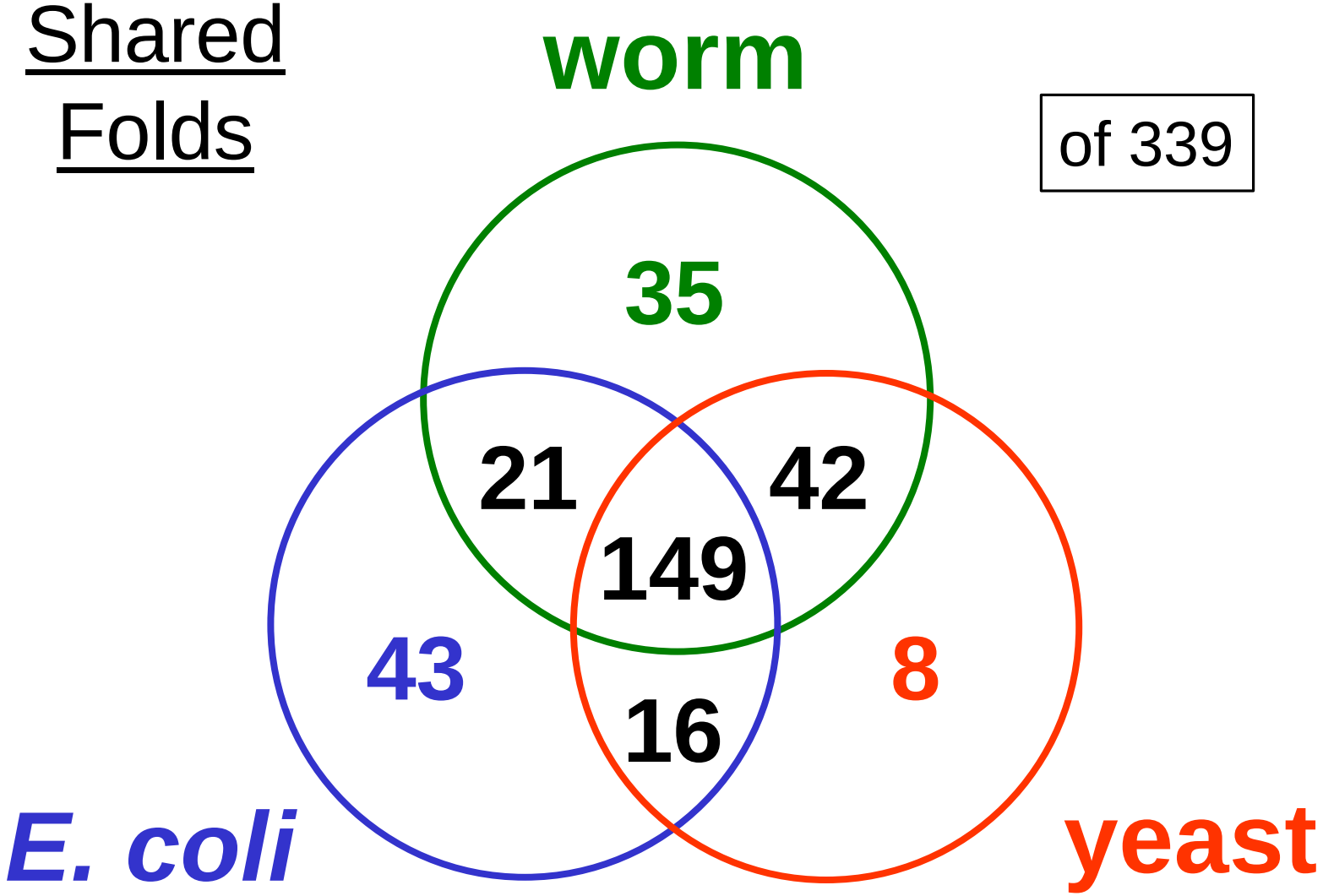


(Pathogen only folds as possible targets)

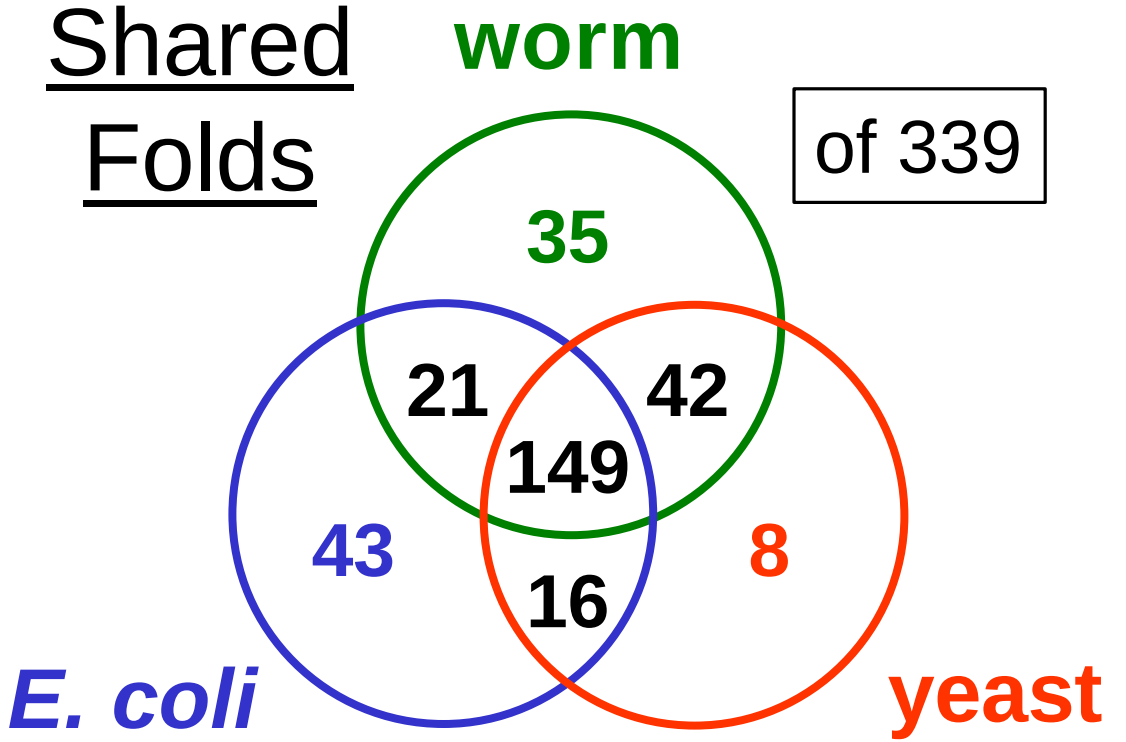


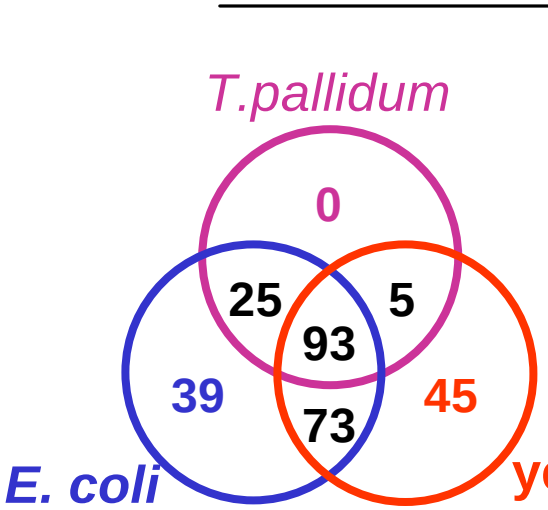
- OspA protein
  - ◊ in Lyme-disease spirochete *B. burgdorferi*
  - ◊ previously identified as the antigen for vaccine
  - ◊ has novel fold (C Lawson)

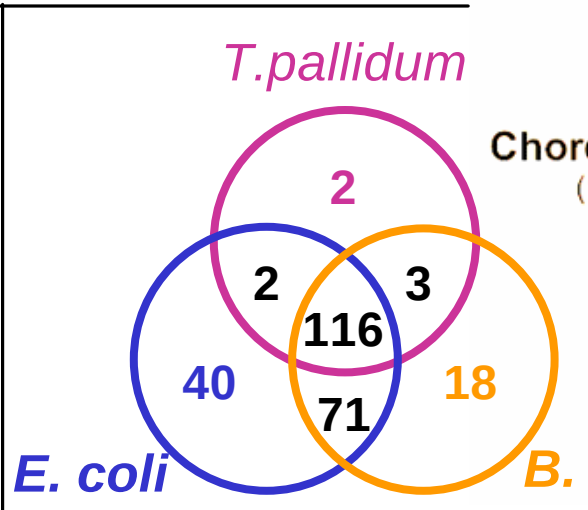
Shared Folds

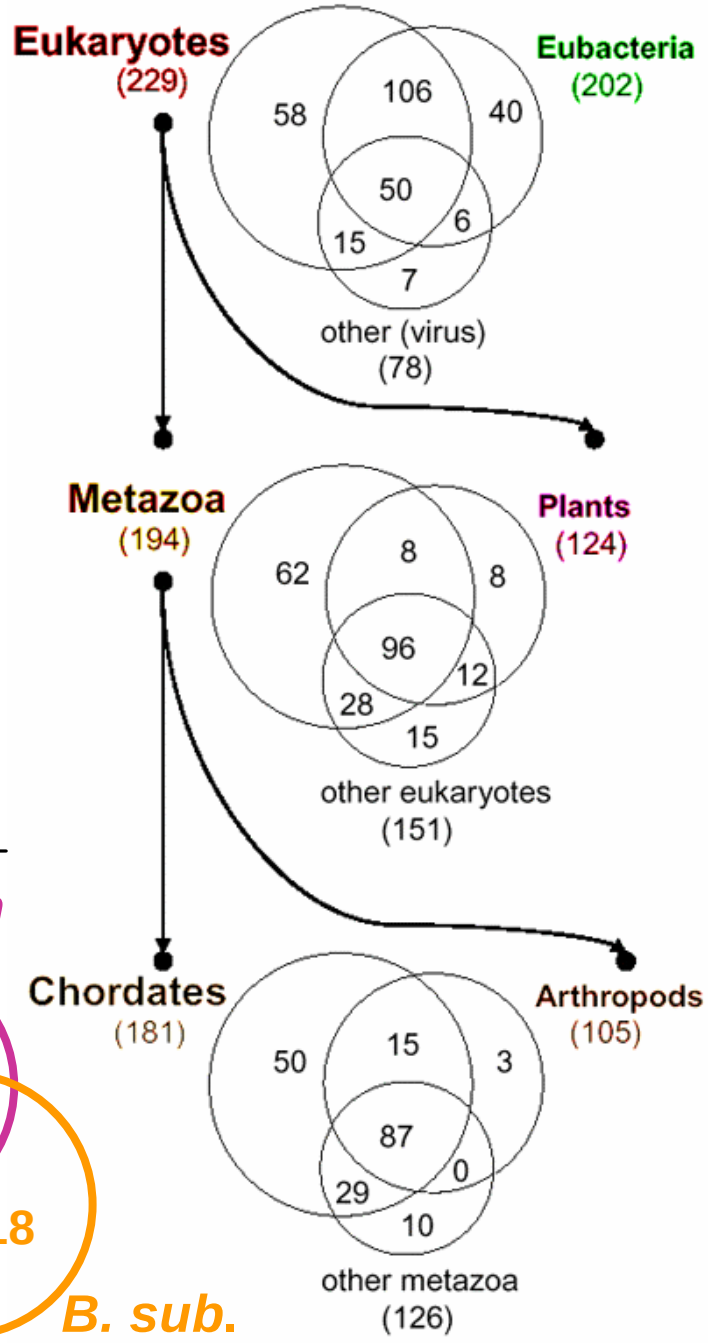


Shared Folds

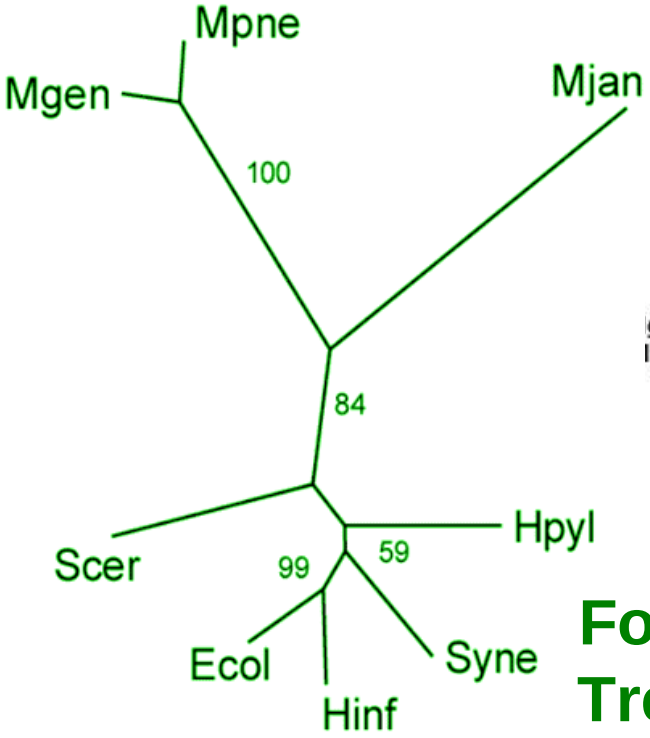




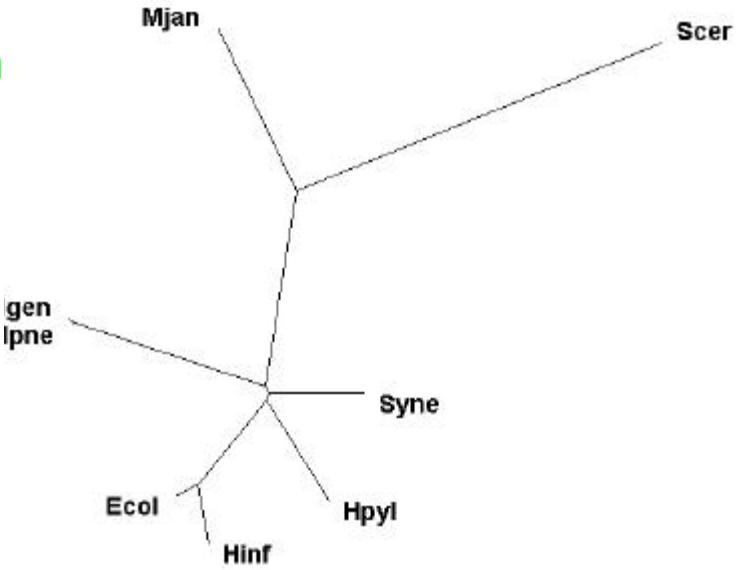




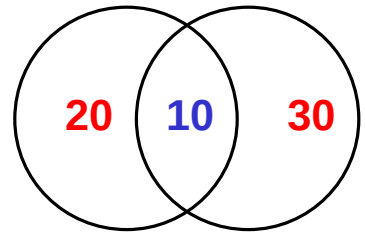
Cluster Trees Grouping Initial Genomes on Basis of Shared Folds



Fold Tree



"Classic" Tree

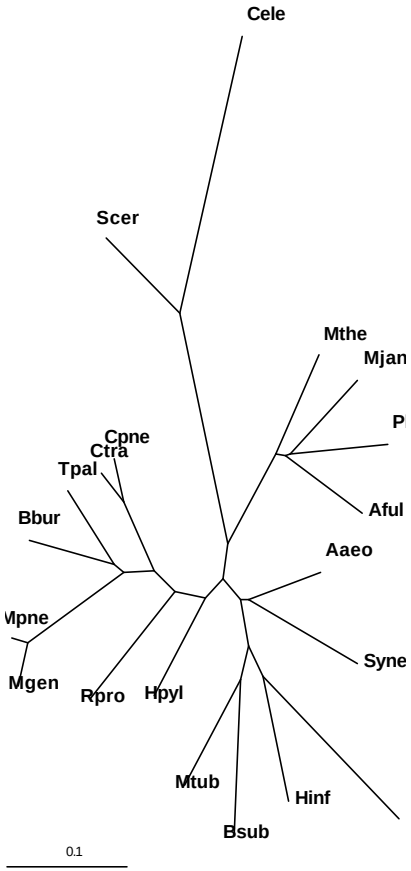


D=10/(20+10+30)

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

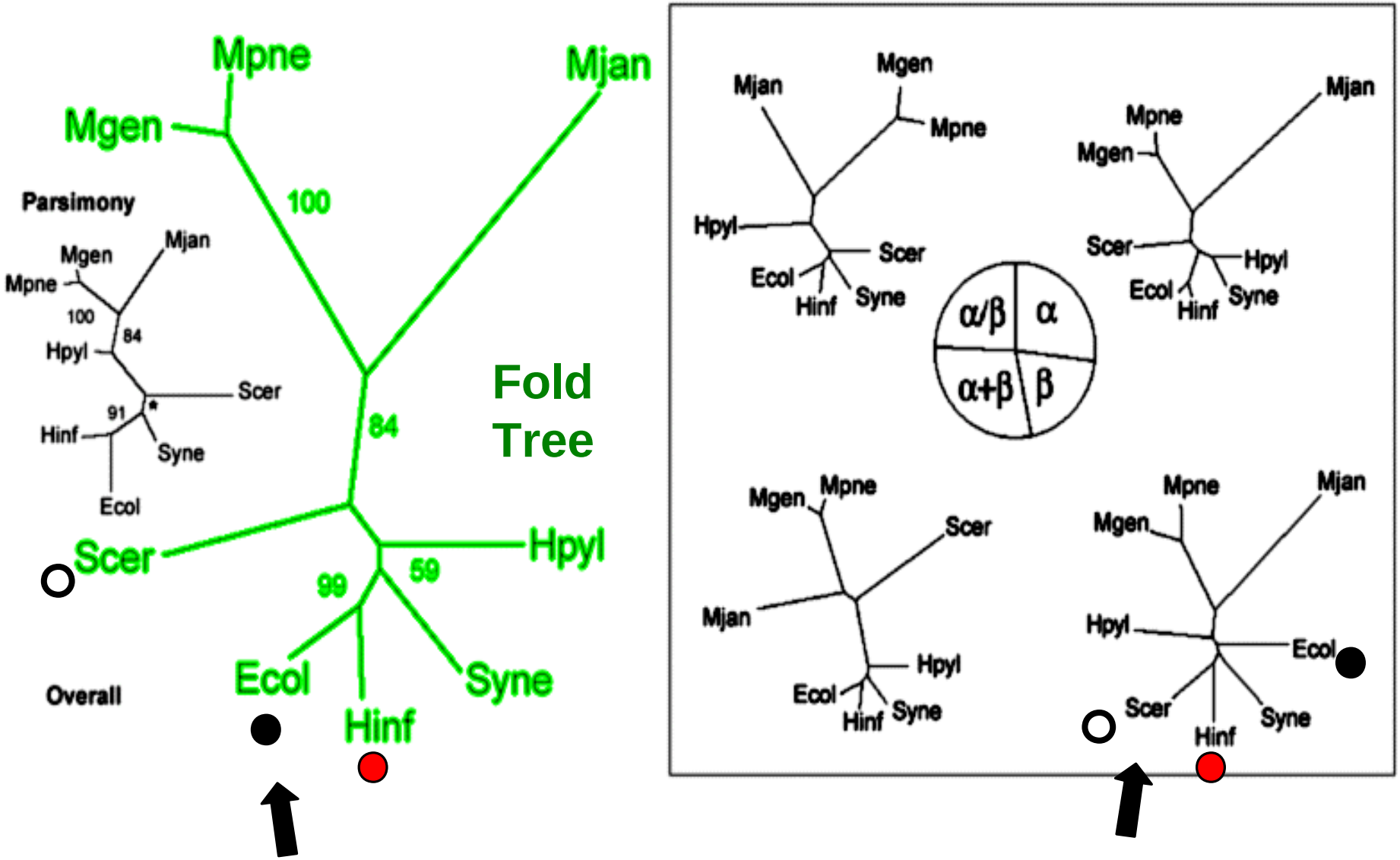
D = shared fold dist. betw. 2 genomes

T= total # folds in both



20 Genomes

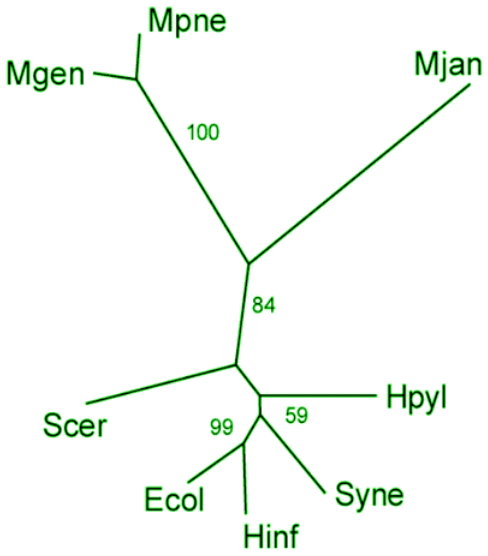
Distribution of Folds in Various Classes



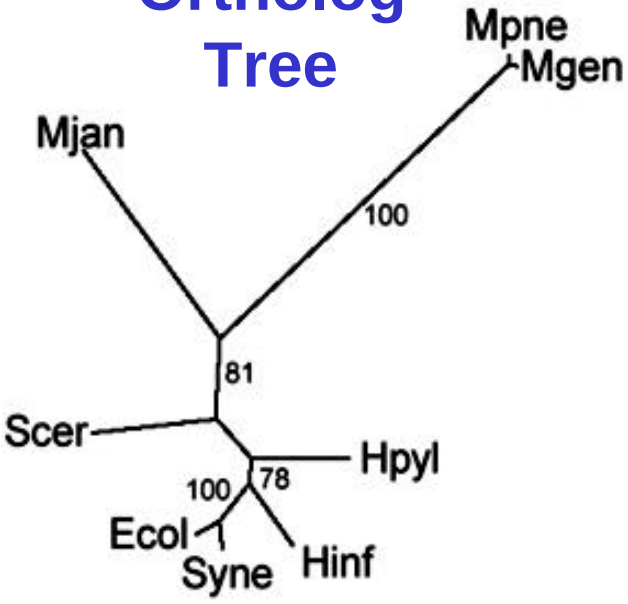
Unusual distribution of all-beta folds

Compare with Ortholog Occurrence Trees, another "partial- proteome" tree

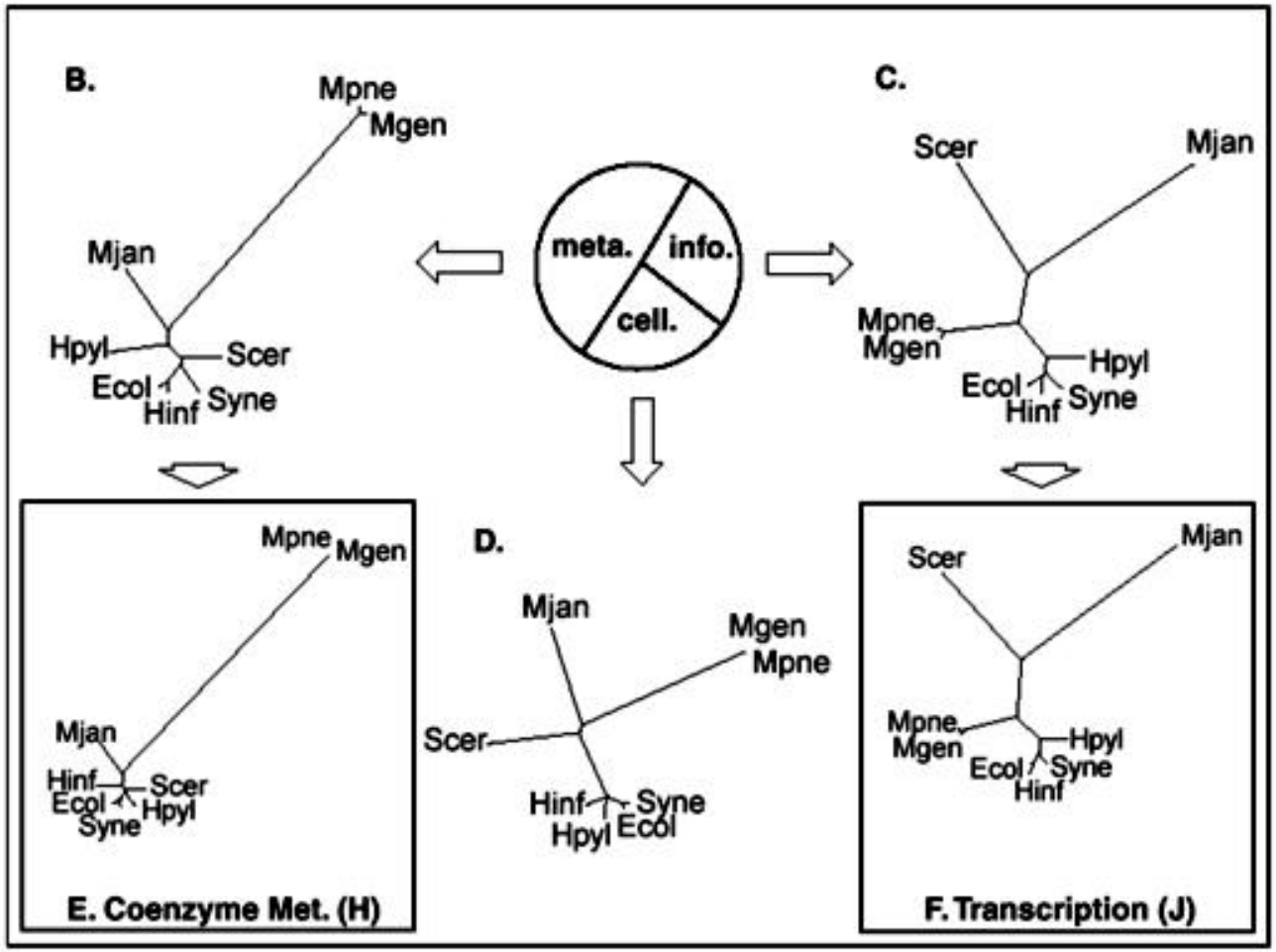
Fold Tree



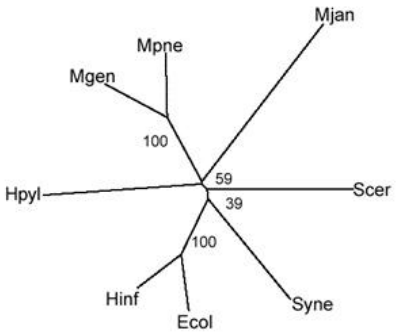
Ortholog Tree



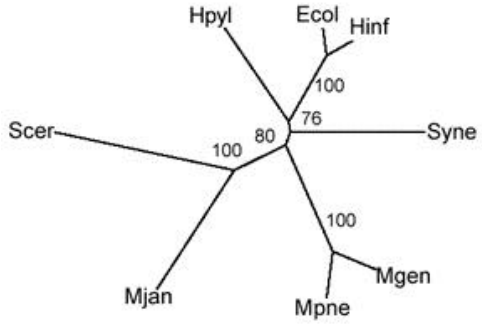
(based on COGs scheme of Koonin & Lipman, similar approaches by Dujon, Bork, &c.)



Compare with trees on spectrum of "levels": single-gene trees, whole-genome composition trees

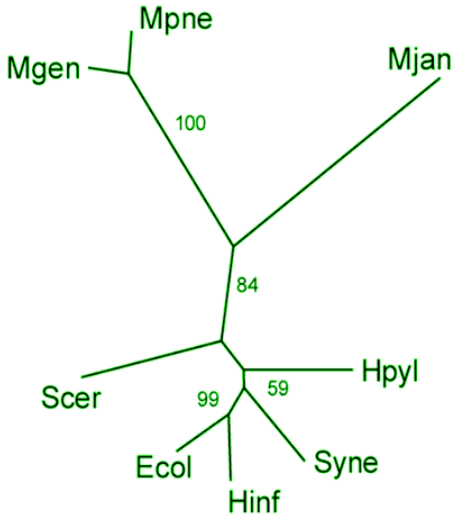


TIM

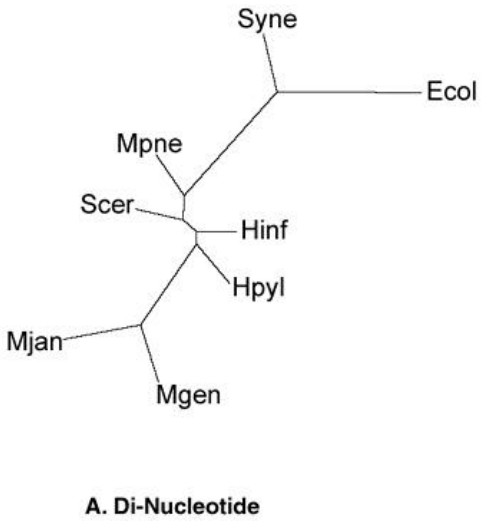


ribosomal protein

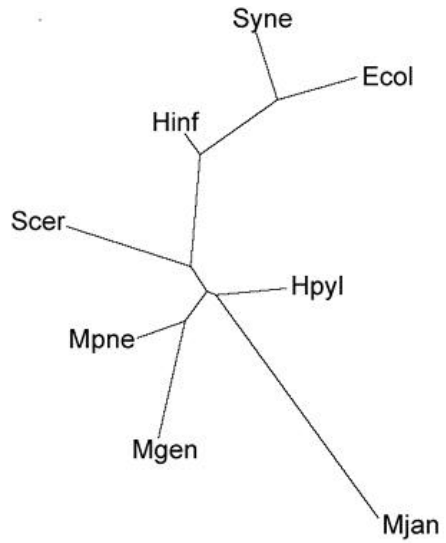
Single-gene Trees



Fold Tree



A. Di-Nucleotide



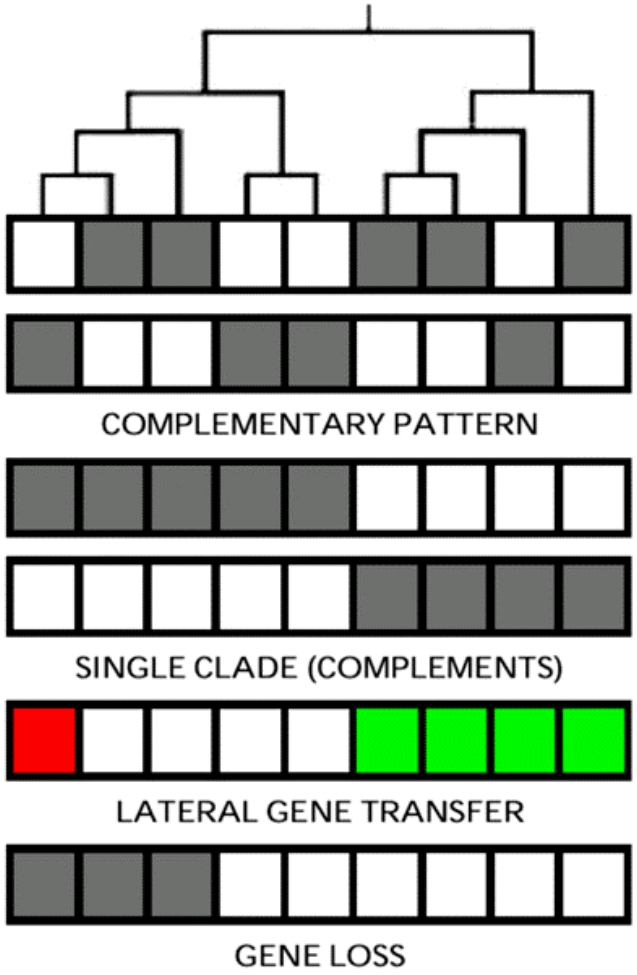
B. Amino Acid

Ortholog Tree

"Classic" Tree

AA & di-NT Composition Trees

(S Karlin)

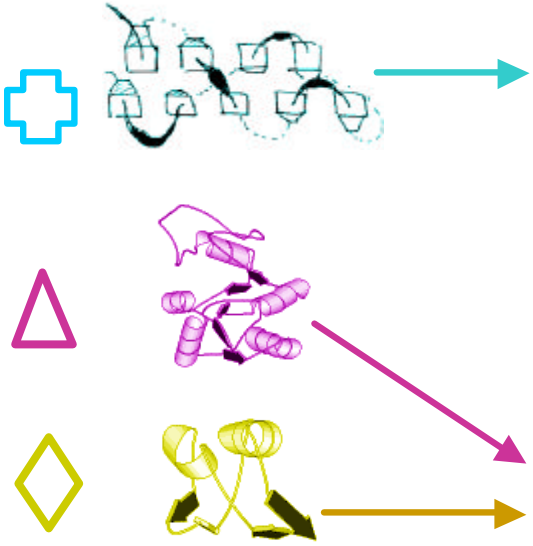


Specific Patterns of Fold Usage: Horizontal Transfer

aful	mjan	mthe	phor	scer	cele	aaeo	syne	ecol	bsub	mtub	hinf	hpyl	mgen	mpne	bbur	tpal	ctra	cpne	rpro	Sfam	domain	SCOP Function	Swissprot	Swissprot Function
•	•		•					•												1.86.1	**d1aora1**	Aldehyde FerOR C' domain	**AOR_PYRFU**	ALDEHYDE:FERREDOXIN OXIDOREDUCTASE
•	•		•					•												4.94.1	**d1aora2**	Aldehyde FerOR N' domain	**AOR_PYRFU**	ALDEHYDE:FERREDOXIN OXIDOREDUCTASE
•	•		•				•		•											3.1.10	**d5ruba1**	RuBisCo, C' domain	**RBL_NITVU**	RUBISCO LARGE SUBUNIT
				•							•									1.101.1	**d5csma_**	Chorismate mutase II	**CHMU_ARATH**	CHORISMATE MUTASE (EC 5.4.99.5)
				•			•					•								1.37.1	**d1rec__**	EF-hand	**TPC2_DROME**	TROPONIN C
					•		•													2.1.5	**d1suh__**	Cadherin	**CAD5_HUMAN**	VASCULAR ENDOTHELIAL-CADHERIN
					•			•					•							2.45.1	**d1eal__**	Lipocalins	**PGHD_HUMAN**	PROSTAGLANDIN-H2
				•	•			•												3.7.1	**d2bnh__**	Leucine-rich repeats	**RINI_PIG**	RIBONUCLEASE INHIBITOR
				•						•										4.70.1	**d1axx__**	Cytochrome b5	**NIA1_MAIZE**	NITRATE REDUCTASE (EC 1.6.6.1)
					•													•		4.112.1	**d1toh__**	Tyrosine hydroxylase	**TY3H_HUMAN**	TYROSINE 3-HYDROXYLASE (EC 1.14.16.2)

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. genitalium* Superfamily	#	*B. subtilis* Superfamily	#	*E. coli* Superfamily	#
1	P-loop hydrolase	60	P-loop hydrolase	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

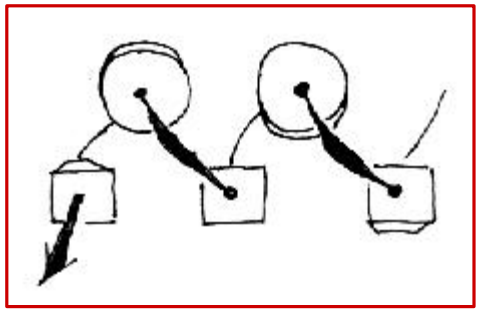
Rank	*M. thermo-autotrophicum* Superfamily	#	*A. fulgidus* Superfamily	#
1	P-loop hydrolyase	93	P-loop hydrolyase	118
2	Phosphate-binding barrel	54	Rossmann domain	104
3	Rossmann domains	53	Phosphate-binding barrel	56
4	Ferredoxins	48	Ferredoxins	49
5	SAM methyl-tranferase	17	SAM methyl-tranferase	24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

Archaea

Rank	*S. cerevisiae* Superfamily	#
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

Common, Shared Folds: βαβ structure



42 ARTICLES NATURE VOL. 336 3 NOVEMBER 1988

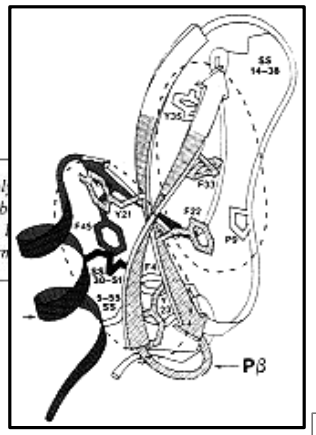
A peptide model of a protein folding intermediate

Terrence G. Oas & Peter S. Kim

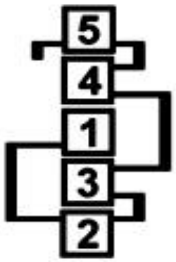
Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

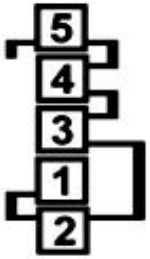
It is difficult to determine the structures of protein folding intermediates because folding is a highl disulphide-bonded peptide pair, designed to mimic the first crucial intermediate in the folding of b inhibitor, contains secondary and tertiary structure similar to that found in the native protein.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

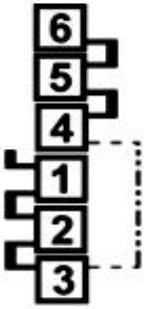
336: 42

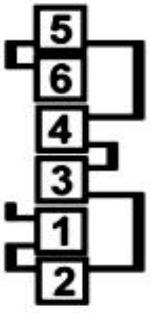


All share α/β structure with repeated R.H. βαβ units connecting adjacent strands or nearly so (18+4+2 of 24)

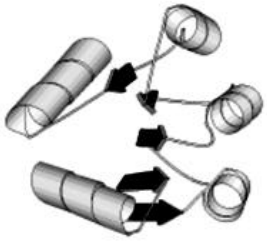


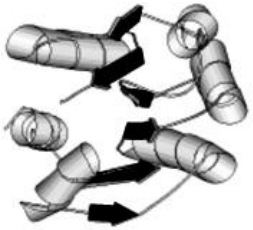


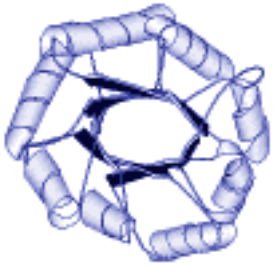




HI, MJ, SC vs scop 1.32

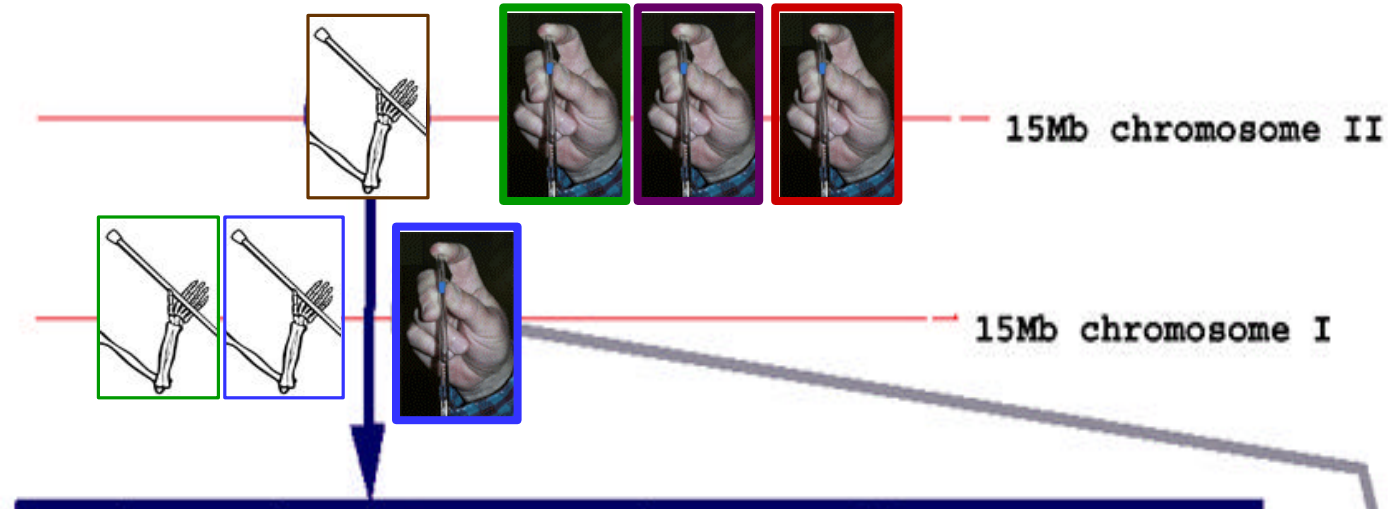




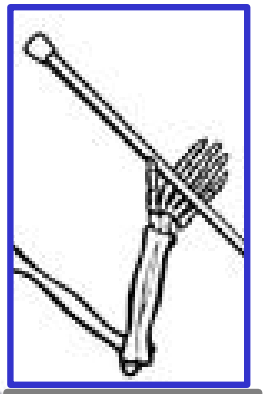


P-loop hydrolase | **Flavodoxin like** | **Rossmann Fold** | **Thiamin Binding** | **TIM-barrel**

Pseudogenomics: Surveying "Dead" Parts



Example of a potential ΨG with frameshift in mid-domain

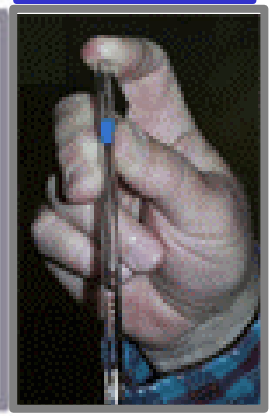


pseudogene fragment on worm chromosome II

TKRTSNGFGQDVVVDLFSILDSGLVARAHXVLQDIFEFFAS
KKMVTIFS#APHSPHSAPHYCAQFDNSAATVKV

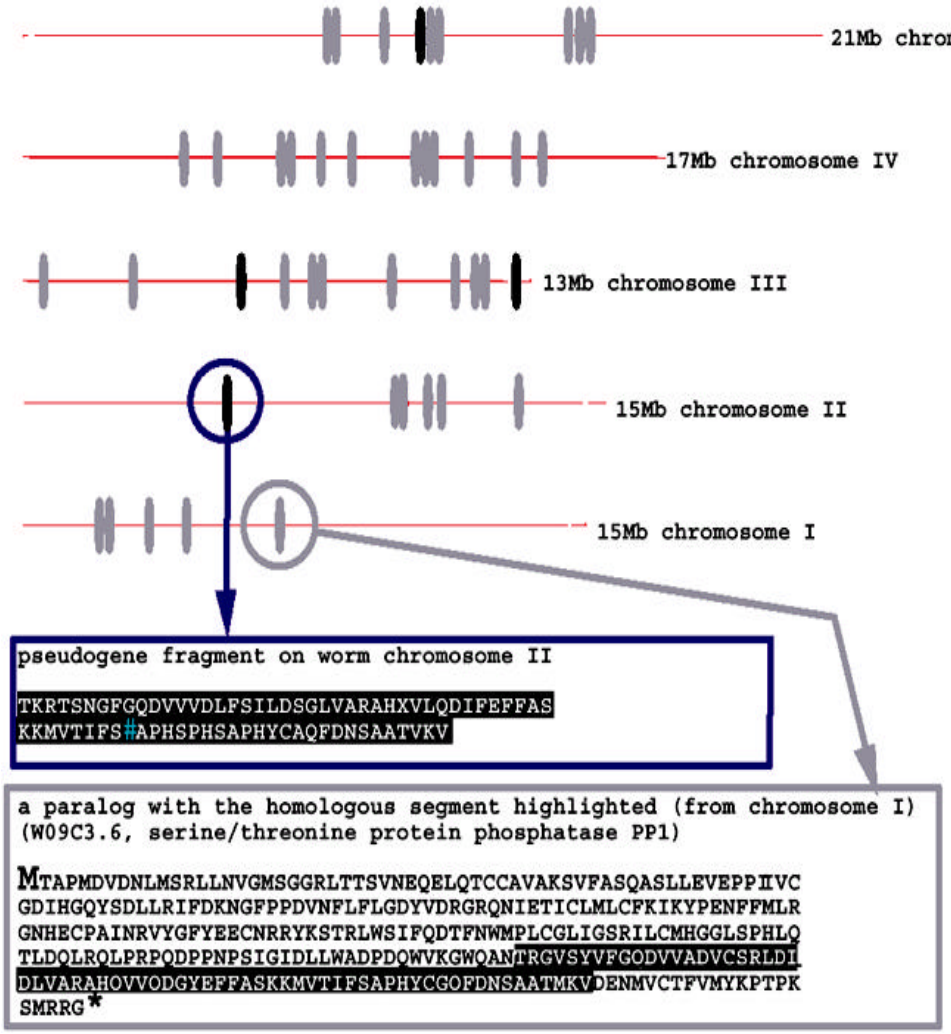
a paralog with the homologous segment highlighted (from chromosome I)
(W09C3.6, serine/threonine protein phosphatase PP1)

MTAPMDVDNLMSRLLNVGMSGGRLTTSVNEQELQTCCAVAKSVFASQASLLEVEPPIVC
GDIHGQYSDLLRIFDKNGFPPDVNFLFLGDYVDRGRQNIETICLMLCFKIKYPENFFMLR
GNHECPAINRVYGFYEECNRRYKSTRLWSIFQDTFNWMPLCGLIGSRILCMHGGLSPHLQ
TLDQLRQLPRPQDPPNPSIGIDLLWADPDQWVKGWQANTRGVSYVFGQDVVADVCSRLDI
DLVARAHQVVQDGYEFFASKKMVTIFSAPHYCGQFDNSAATMKVDENMVCTFVMYKPTPK
SMRRG*



(Our def'n: ΨG = obvious homolog to known protein with frameshift or stop in mid-domain)

Folds in Pseudogenes



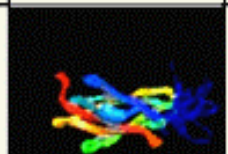
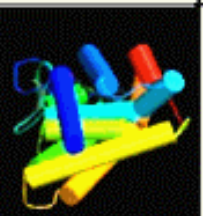
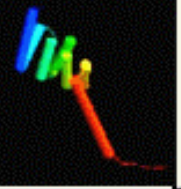
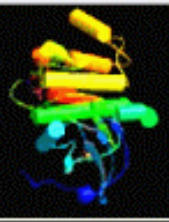
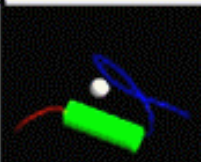
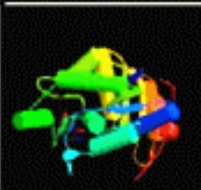
Example of a potential ΨG with frameshift in mid-domain

ΥG identification pipeline to Summary of Pseudogenes in worm

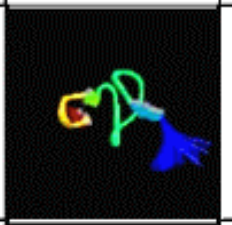
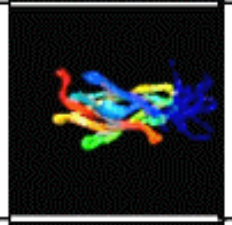
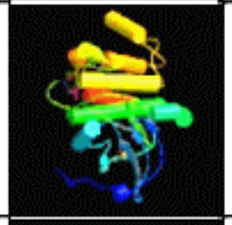
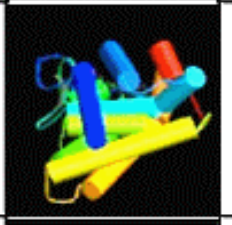
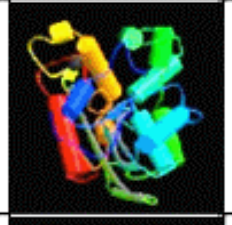
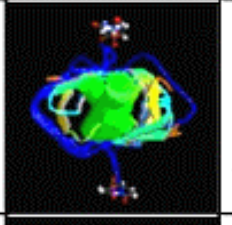
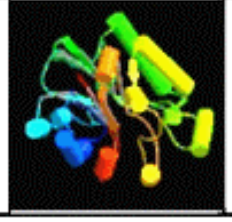
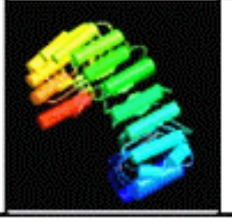
G=19K **G_E=8K** **ΥG=4K (2K)**

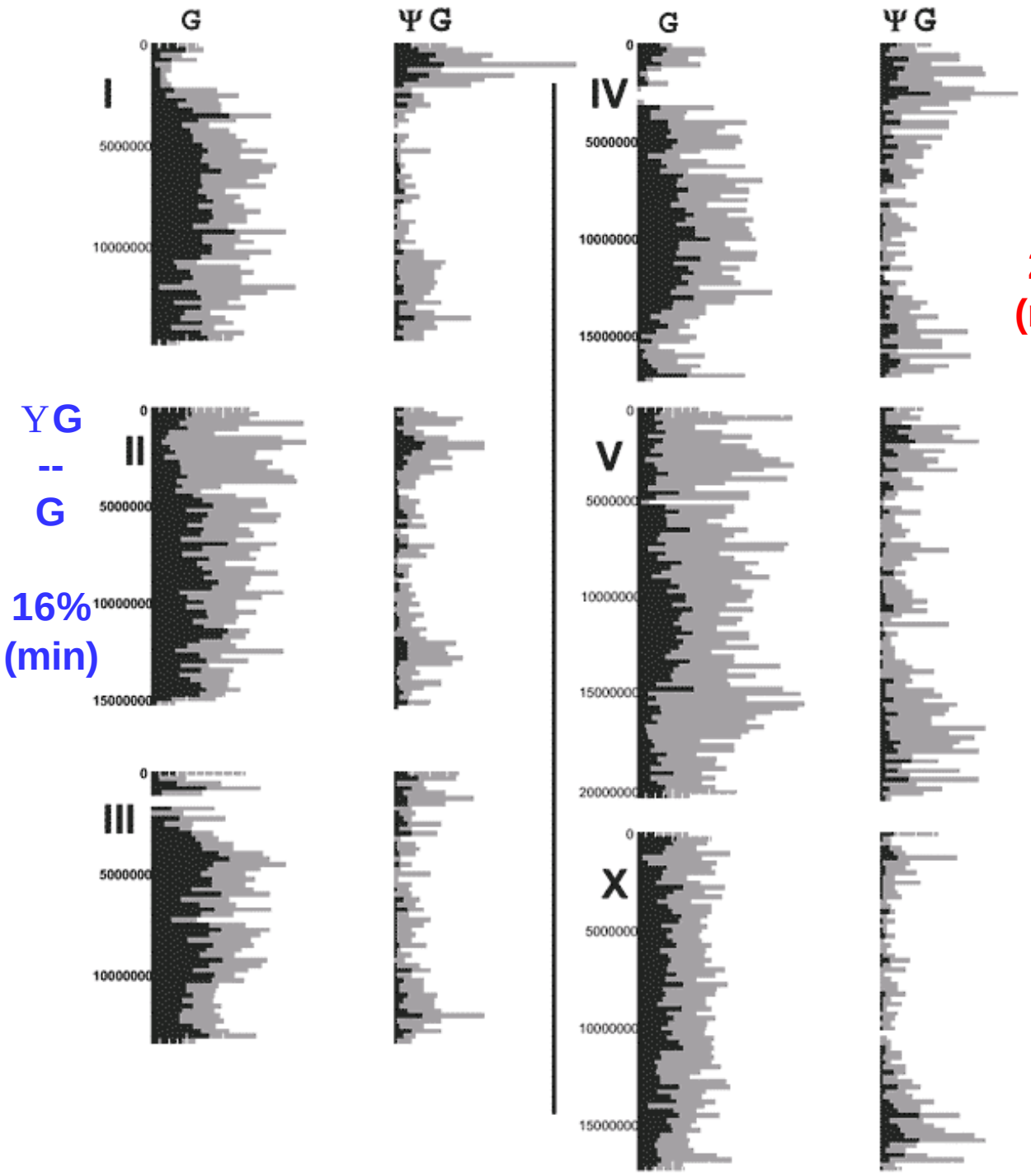
	Category	Total number	Number for genes with EST match	Genes with EST match as percentage of *Category*	Number for genes in paralog families with EST match	Genes in paralog families with EST match as percentage of *Category*
Genes	Total	18,576 (G)	7,829 (G_E)	42%	13,417 (G_P)	72%
	Singletons	5,913	2,788	47%	---	---
Pseudogenes and pseudogene fragments	Total	3,814 (ΨG)	997 (ΨG_E)	26%	2,729 (ΨG_P)	72%
	Singletons	637 (17% of ΨG)	233	36%	---	---
	Intronic pseudogenes *	1,155 (30% of ΨG)	351	30%	704	61%

Most Common Worm "Pseudofolds" #1

G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
1 (769)	2		d1ajw__ 2.1 Immuno-globulin	6 (246)	8		d2lbd__ 1.95 Nuc. receptor ligand-binding domain
2 (555)	6		d1dec__ 7.3 Knottin	7 (243)	34		d1a17__ 1.91 Alpha/alpha superhelix
3 (434)	3		d3lck__ 5.1 Protein kinase	8 (227)	17		d1sp2__ 7.31 Classic zinc finger
4 (302)	1		d1tsg__ 4.105 C-type lectin	9 (215)	20		d1dai__ 3.29 P-loop NTP hydrolase
5 (274)	7		d1zfo__ 7.33 Glucocorticoid receptor DNA-binding dom.	10 (197)	13		d2aw0__ 4.34 Ferredoxin

Most Common Worm "Pseudofolds" #2

ΨG Rank (Number matches)	G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	ΨG Rank (Number matches)	G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
1 (39)	4		d1tsg__ 4.105 C-type lectin	6 (18)	2		d1dec__ 7.3 Knottin
2 (32)	1		d1ajw__ 2.1 Immuno-globulin	7 (17)	5		d1zfo__ 7.33 Glucocorticoid receptor DNA-binding dom.
3 (27)	3		d3lck__ 5.1 Protein kinase	8 (15)	6		d2lbd__ 1.95 Nuc. receptor ligand-binding domain
4 (25)	11		d1cvl__ 3.56 Alpha/beta-hydrolase	9 (13)	58		d1bus__ 7.14 Ovomucoid PCI inhibitor fold
5 (23)	63		d1ako__ 4.93 DNAse-I fold	9 (13)	19		d2bnh__ 3.7 Leu-rich, right-handed β/α superhelix



ΥG -- G

16% (min)

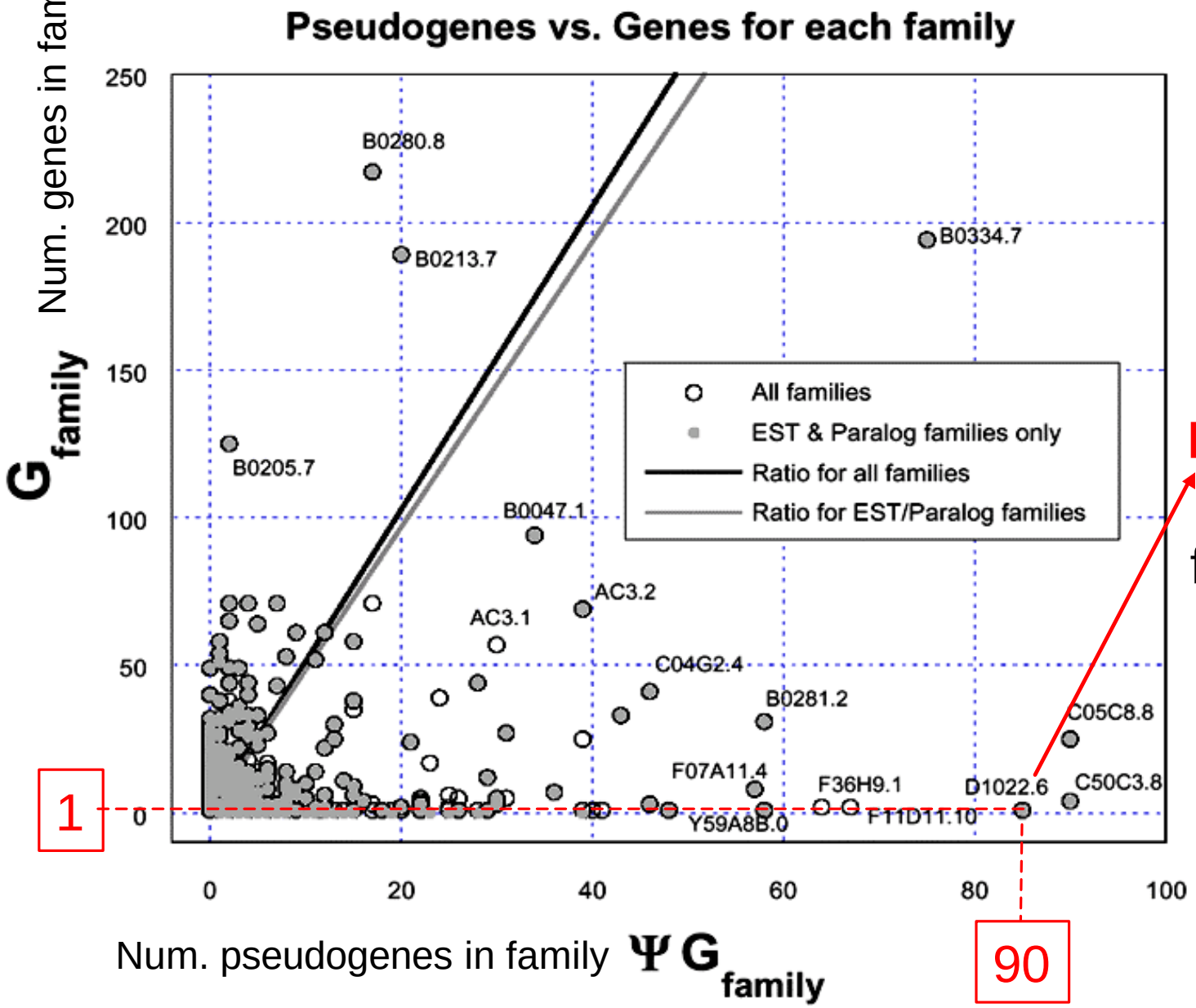
ΥG -- G

29% (max)

Pseudogene Distribution on Chomo-somes

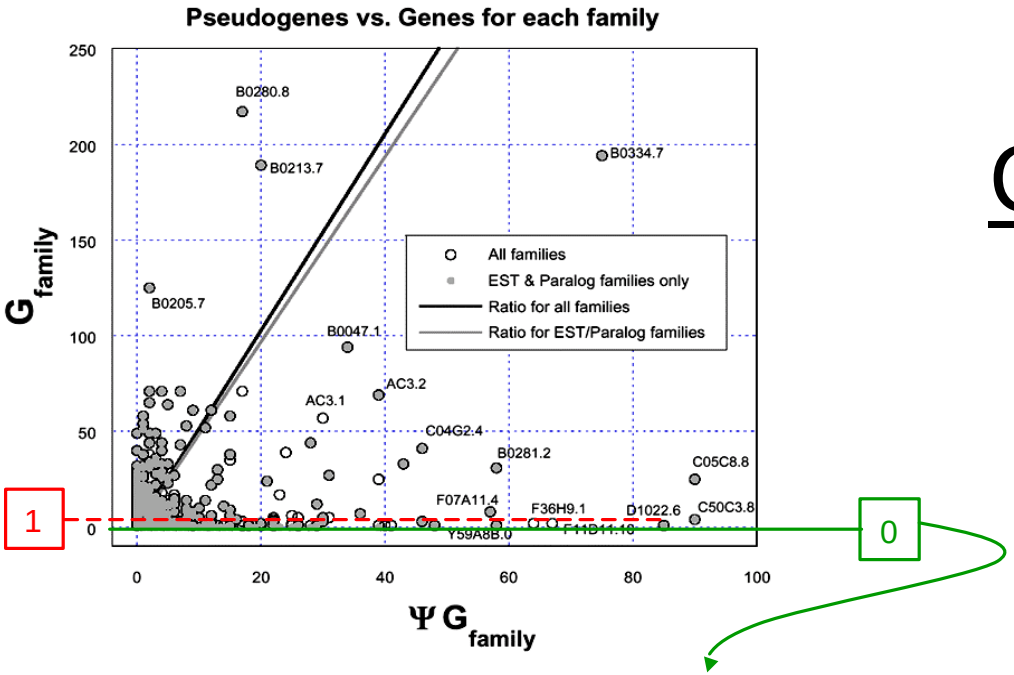
~50% ΨG in terminal 3Mb vs ~30% G

Decayed Lines of Genes?



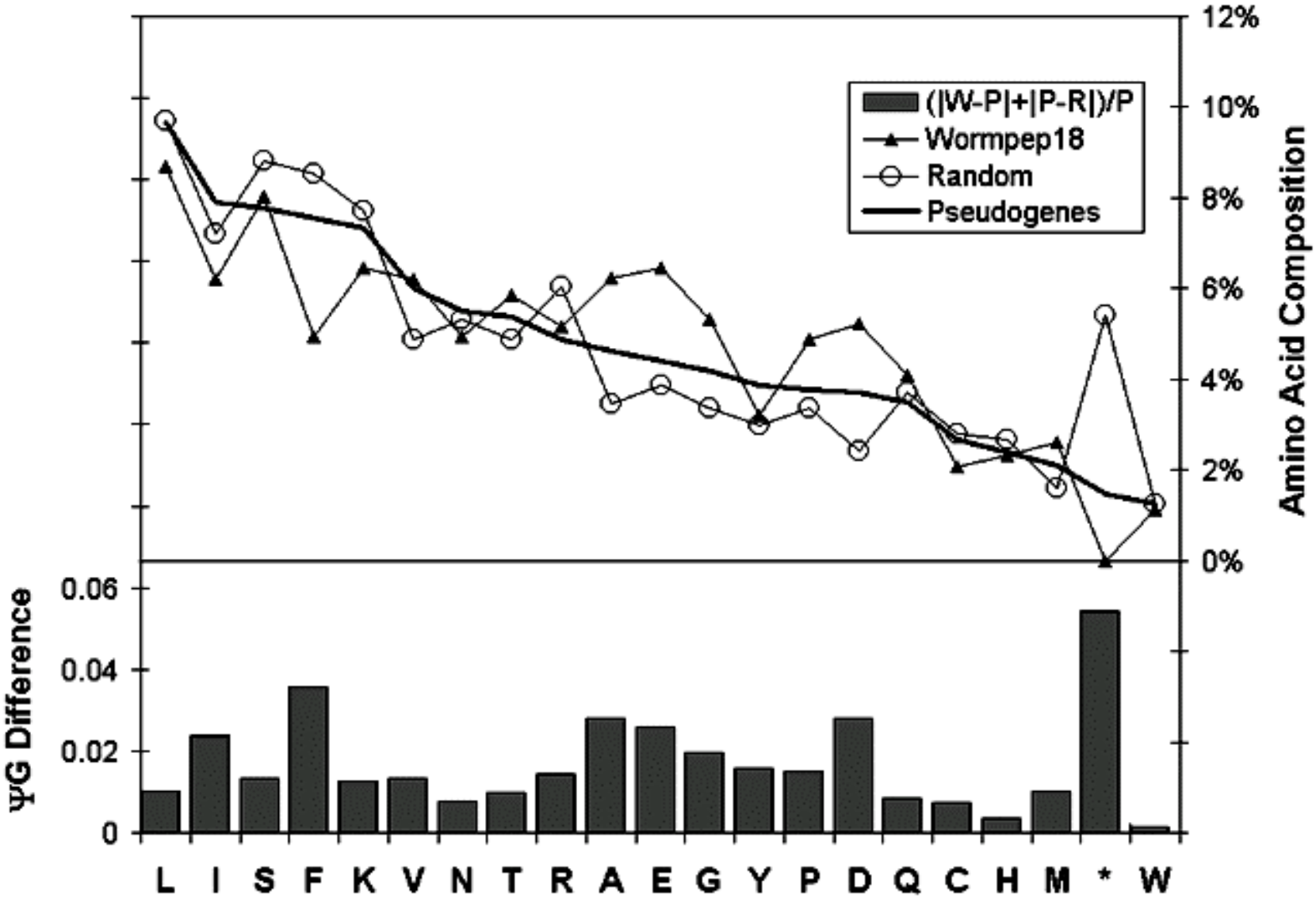
D1022.6 has 90 dead fragments of itself – a disused line of chemo-receptors?

Completely Dead Families

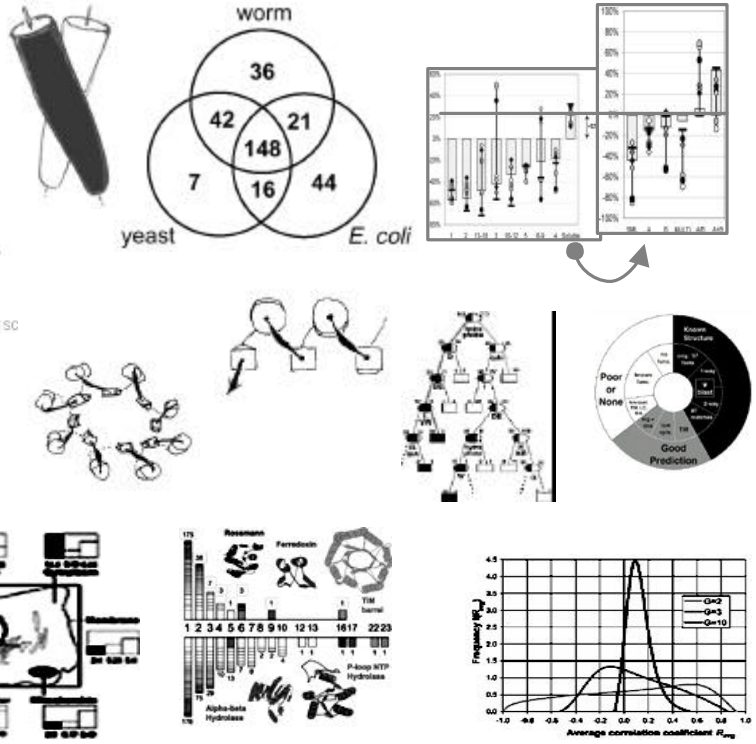


Rank	Number matches	Organism of closest match*	PROTOMAP family representative	Notes on representative
#1	7 *******	Yeast	YJA7_YEAST	Hypothetical protein in yeast
#2 =	5 *****	Human	XPD_MOUSE	Xeroderma pigmentosum group D complementing protein
#2 =	5 *****	Cow	CPSA_BOVIN	Cleavage and polyadenylation specificity factor
#4 =	4 ****	Frog	THB_RANCA	Thyroid hormone receptor beta
#4 =	4 ****	Human	SEX_HUMAN	SEX gene
#4 =	4 ****	Fly	MDR1_RAT	Multidrug resistance protein 1
#7 =	3 ***	Vaccinia virus	YVFB_VACCC	Hypothetical vaccinia virus protein
#7 =	3 ***	Fly	VHRP_VACCC	Host range protein from vaccinia
#7 =	3 ***	Human	IF4V_TOBAC	Eukaryotic initiation factor 4A
#7 =	3 ***	*E. coli*	ACRR_ECOLI	Acrab operon repressor

Amino Acid Composition of Pseudogenes is Midway between Proteins and Random



Comparative Genomics: Surveys of a Finite Parts List



1 Using Folds to Interpret Genomes.

Fold Library background. Shared and/or unique parts. Venn Diagram, Fold tree with all-β diff. Ortholog tree. Horizontal Transfer. Commor Parts: Top-10 folds with βαβ. Common Ψ-fold.

- **Tricky Issues →Expt. SG.** Extent of fold assignment (MG, 65%, 98). Predictions. Biases. NESG.org Construct DB. Datamining this w/ decision trees. Selecting weird MG CD targets.
- **Folds & Functions.** Roles/part? How many folds/func? 331 of ~20K combinations. Mostly 1 func/fold, but some versatile scaffolds -- TIM most versatile. Similar for interactions. Func. Divergence vs. Seq. & Struc. Diverg.

4 Using Folds to Interpret Expression Data.

Top-10 parts in other terms. Enriched in transcriptome: VGA, αβ folds, energy, synthesis, Cytoplasmic, TIM fold. Depleted: NS, long, TMs, transport, transcription, Nuclear, Leu-zip fold. Bayesian localizer.

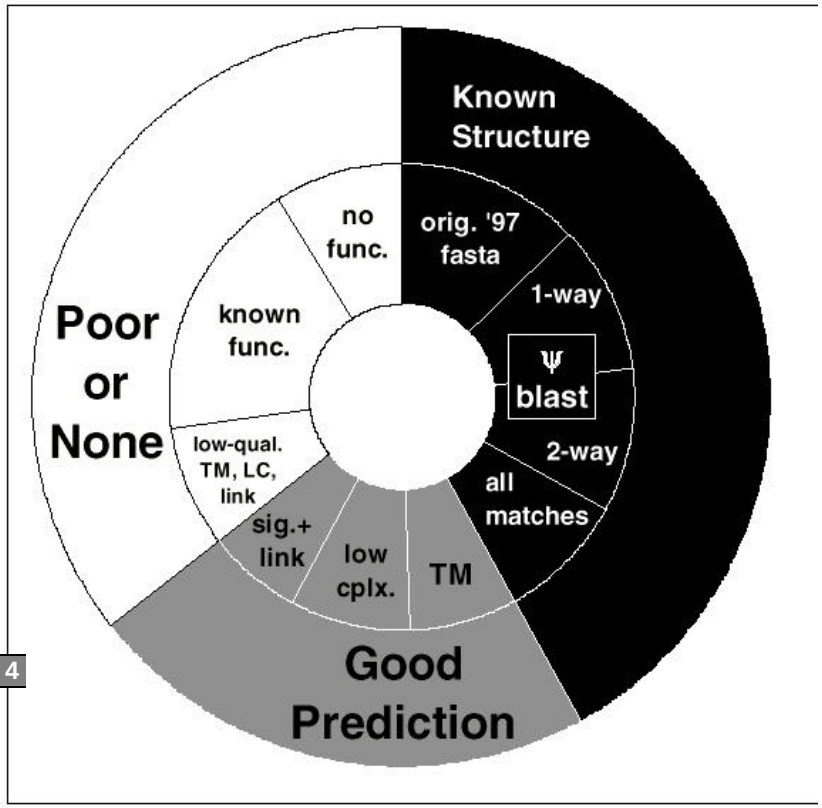
5 Tricky Issues: Relating Expression to Function.

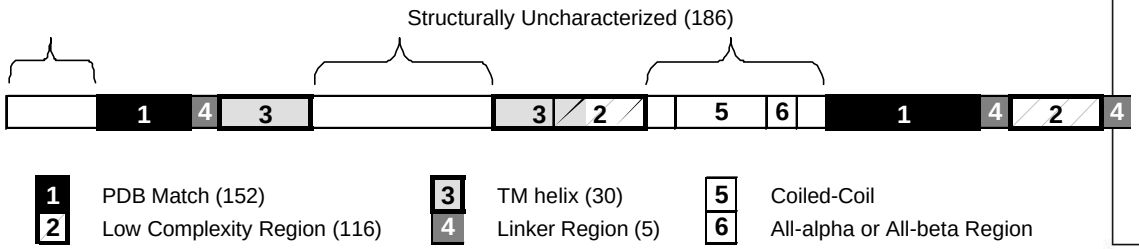
Expression relates to structure & localization but to function, globally? Weak relation to protein-protein interactions.

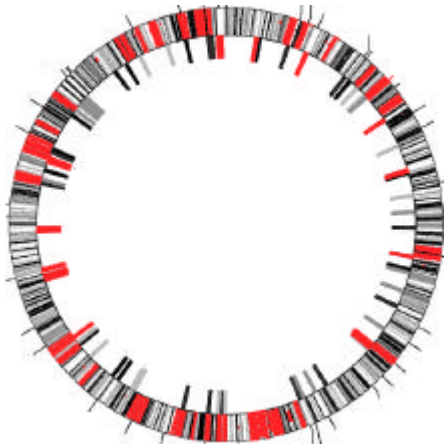
H Hegyi, J Lin, B Stenger, N Echols, P Bertone, J Qian, L Regan, S Balasubramanian, V Alexandrov, G Montelione, A Edwards, C Wilson, Y Kluger, C Arrowsmith, A Drawid, R Jansen, D Greenbaum, S Teichmann, P Harrison

bioinfo.mbb.yale.edu

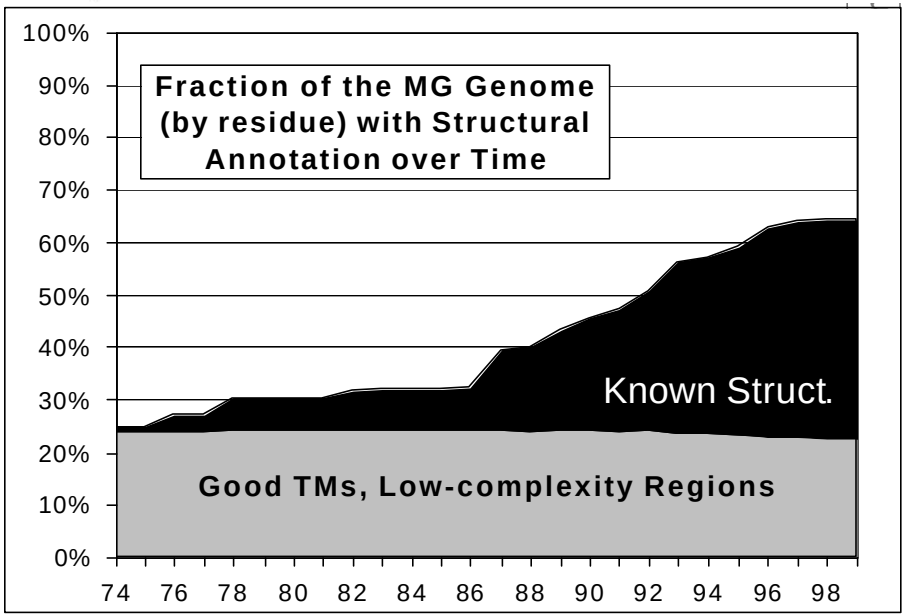
The Obvious Problem: Incomplete Coverage of the Genome by Known Folds







Status of M. gen. at end end of '98



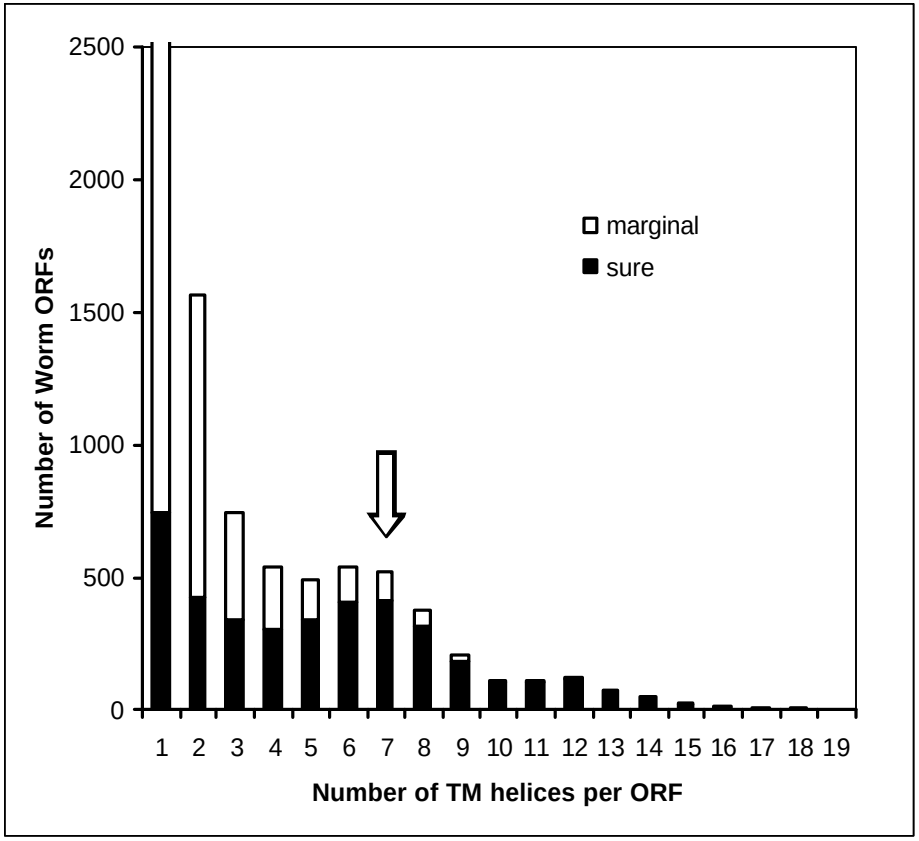
Bias Problem → Prediction, Experimental Structural Genomics

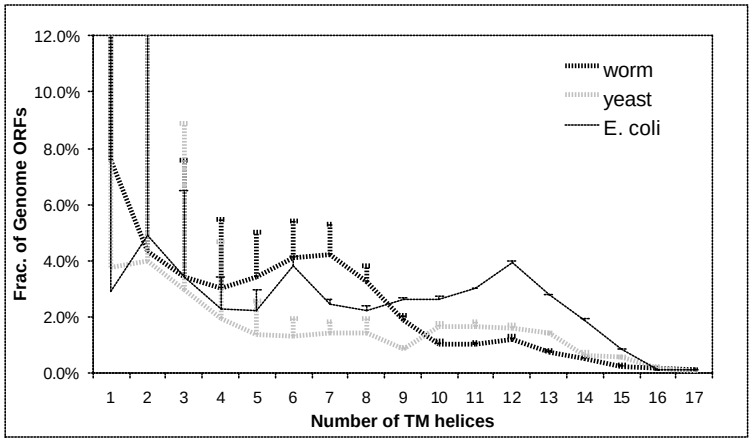
- Known Structures are Incomplete, Biased Sample from Genome, so...
  - ◊ Resample
  - ◊ Solve Structures
  - ◊ Predict Structures

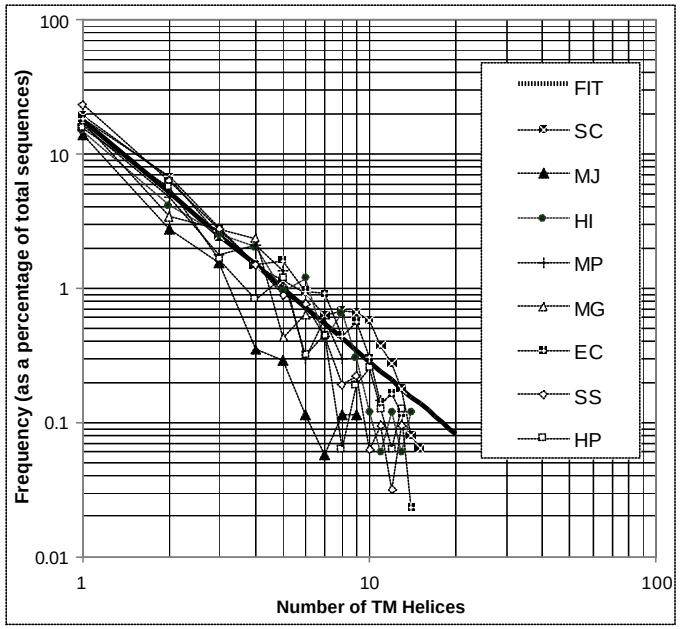
Same Sampling Issues with US Census!!

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)

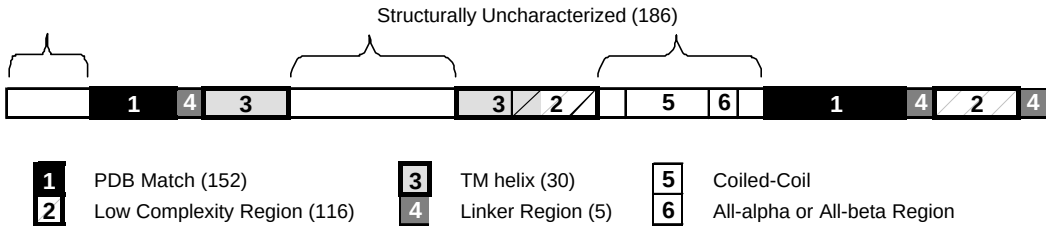






2º Structure Prediction

- 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.



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

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%

Not expected since……

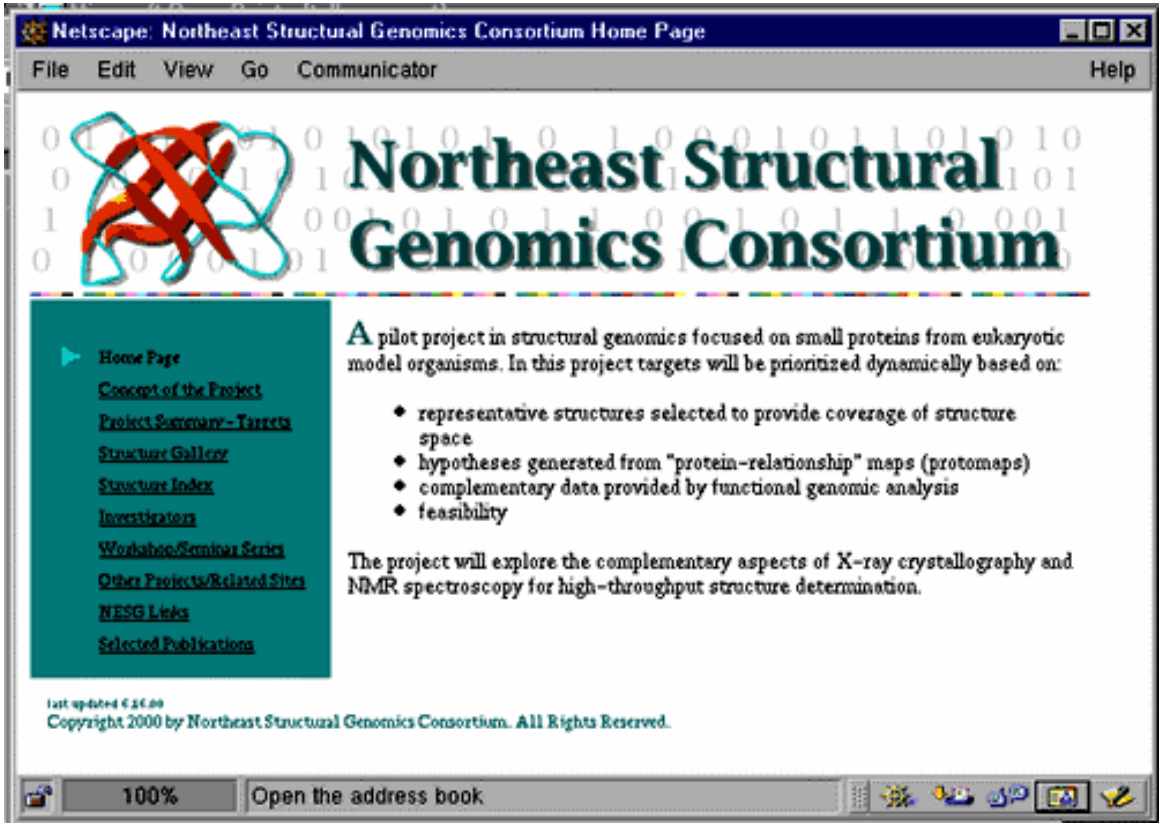
Different Amino Acid Composition Should Give Different 2º Structure

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			

nesg.org



G Montelione

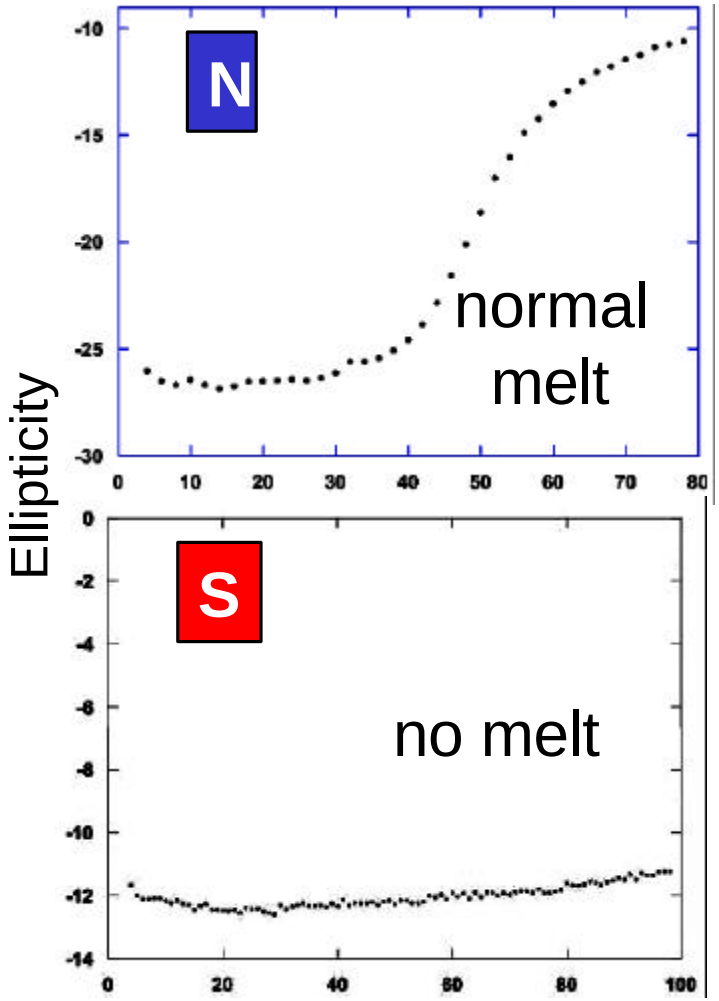
Columbia, Cornell, Yale, Rutgers, Toronto….

Strange S

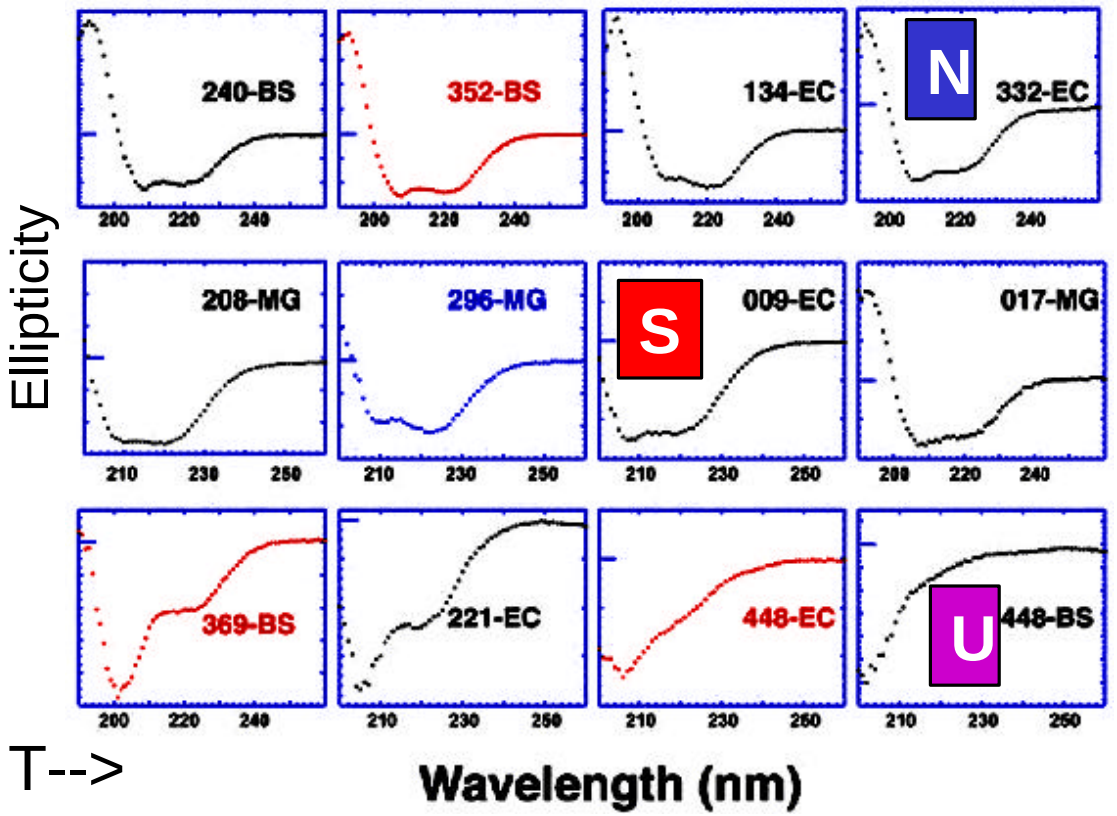
Normal N

Unstructured U

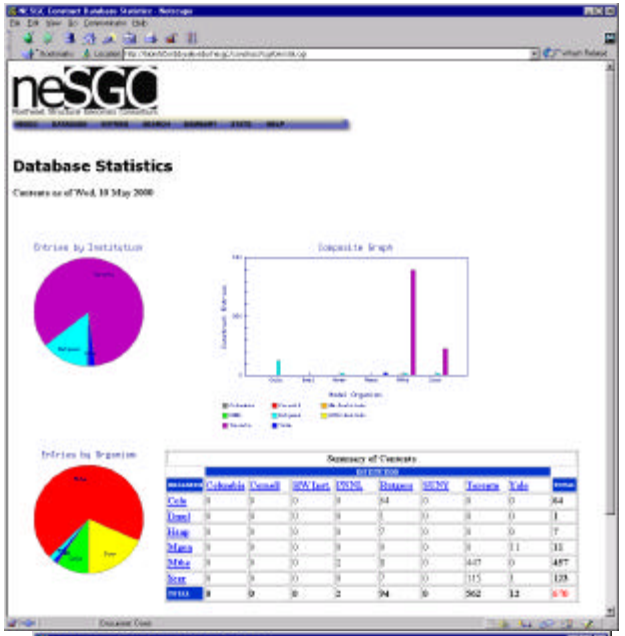
ORFID	Length	Source	Cloned	Expression	Fraction at 37°C	Purification	2° Structure	Thermal melts	Minimal	Essential
9	262	EC	Yes	++	IB	Sol (0.1mM IPTG)	H	No	Yes	No
17	176	MG	Yes	++	IB	IB	H	No	No	No
56	277	EC	Yes	-	-	-	-	-	Yes	Yes
134	100	EC	Yes	+	Sol	Sol	H	Yes	No	Yes
208	196	MG	Yes	+	IB+Sol	Sol	H	Yes	No	No
221	154	EC	Yes	++	IB+Sol	Sol [1]	H+C	No[2]	Yes	Yes
240	292	BS	Yes	++	IB+Sol	Sol	H	Yes	No	Yes
296	129	MG	Yes	++	IB	Sol (25°C)	H	Yes	No	No
332	239	EC	Yes	++	IB+Sol	Sol	H	Yes	Yes	Yes
352	166	BS	Yes	++	IB	Sol (30°C)	H	Yes	No	NO
369	557	BS	Yes	++	IB+Sol	Sol	H+C	Yes	No	Yes
448	150	BS	Yes	++	IB+Sol	Sol	C	-	Yes	Yes
448	150	EC	Yes	++	IB+Sol	Sol	C	-	Yes	Yes
461	425	BS	No	-	-	-	-	-	No	Yes



M.gen. CD results



Progress Database



NESGC Project Summary - Netscape

File Edit View Go Communicator Help

Bookmarks Location: http://bioinfo5.mbb.yale.edu/nesgc/construct/cgi-bin/summary.cgi What's Related

neSGC
Northeast Structural Genomics Consortium

NESGC DATABASE ENTRIES SEARCH SUMMARY STATS HELP

Project Progress Summary

Select the database entries to display: Advanced Form

Target Organism: Any Organism Attribute: Tertiary Structure

Homolog Organism: Any Organism

Laboratory: Any Laboratory

Institution: Any Institution

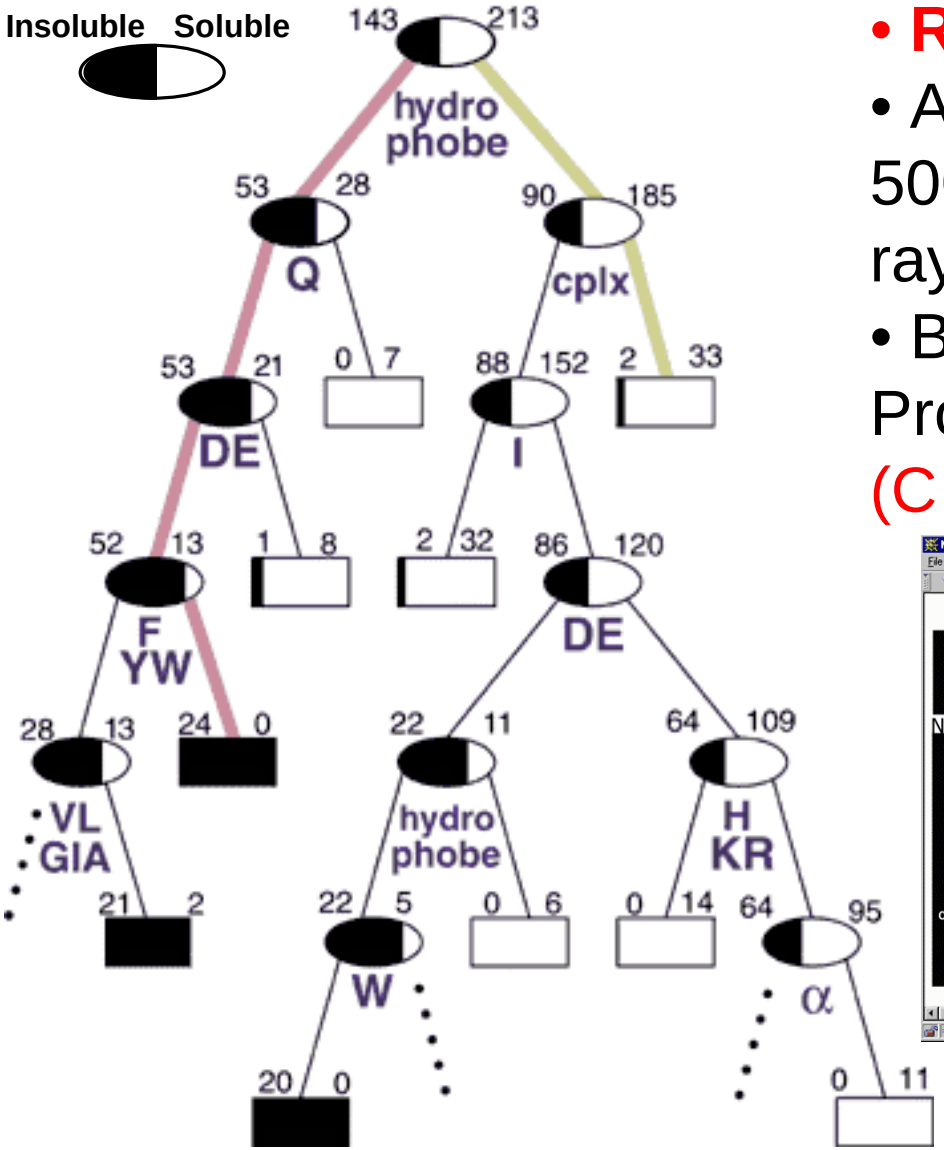
Create Table Display: Reduced

There are 14 entries which have either X-ray or NMR tertiary structure, excluding restricted access and test entries.

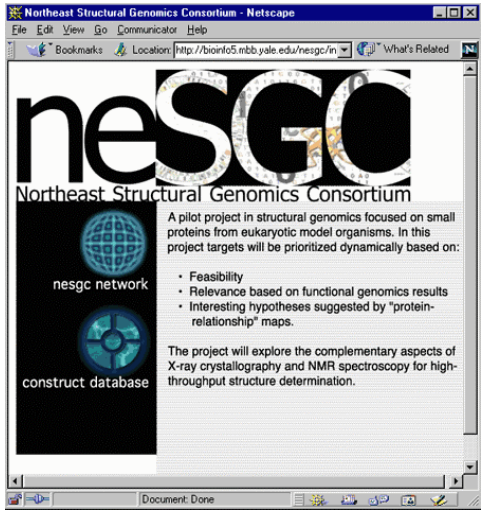
Download Sequences

NESGC ID	PDB ID	Organism	Target ORF	Cloned	Expressed	Purified	HSQC Spectrum	CD Assignments	Crystal	X-ray Data	X-ray Structure	NMR Assignments	NMR Structure	Biochemical Function
WR64Tt	-	Cele	-	*	*	*	*	-	-	-	-	*	*	*
TT1	-	Mthe	MT 0152	*	*	*	-	-	*	*	*	-	-	*
TT2	-	Mthe	MT0129	*	*	*	-	-	*	*	*	-	-	*
TT3	-	Mthe	MT1790	*	*	*	-	-	*	*	*	-	-	*
TT4	-	Mthe	MT0150	*	*	*	-	-	*	*	*	-	-	*
TT5	-	Mthe	MT1791	*	*	*	-	-	*	*	*	-	-	*
TT9	1eik	Mthe	mt1048	*	*	*	*	-	-	-	-	*	*	*
TT10	1eij	Mthe	mt1615	*	*	*	*	-	-	-	-	*	*	*
TT11	-	Mthe	mth0040	*	*	*	*	-	-	-	-	*	*	*
TT12	-	Mthe	mt1699	*	*	*	*	-	-	-	-	*	*	*
TT13	-	Mthe	mt1184	*	*	*	*	-	-	-	-	*	*	*

Characterizing the Low-hanging Fruit for Experimental Structural Genomics

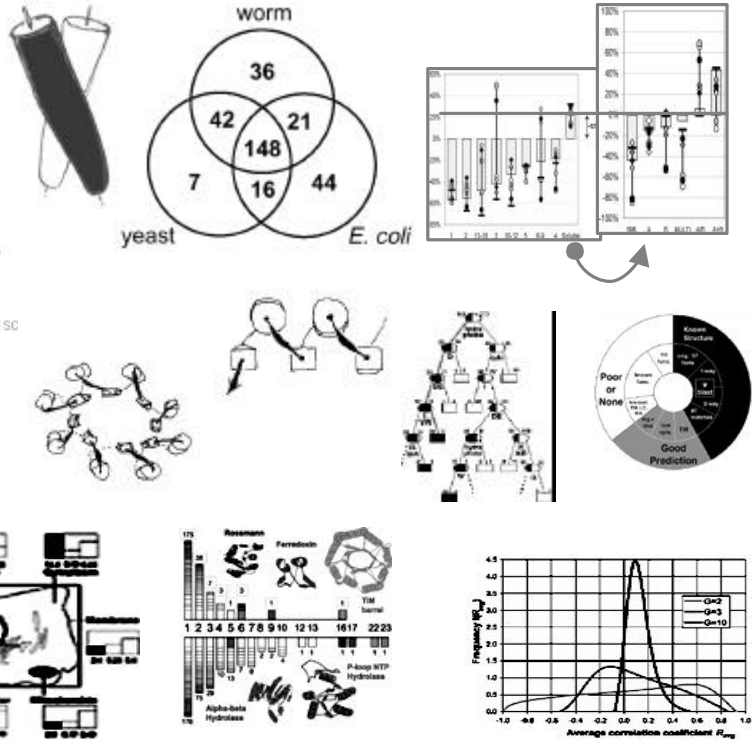


- **Retrospective Decision-Tree**
- Analysis of the Suitability of 500 M. thermo. proteins for X-ray/NMR work
- Based on results of Toronto Proteomics Group (C Arrowsmith, A Edwards)



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.

Comparative Genomics: Surveys of a Finite Parts List



1 Using Folds to Interpret Genomes.

Fold Library background. Shared and/or unique parts. Venn Diagram, Fold tree with all-β diff. Ortholog tree. Horizontal Transfer. Commor Parts: Top-10 folds with βαβ. Common Ψ-fold.

2 Tricky Issues →Expt. SG.

Extent of fold assignment (MG, 65%, 98). Predictions. Biases. NESG.org Construct DB. Datamining this w/ decision trees. Selecting weird MG CD targets.

3 Folds & Functions.

Roles/part? How many folds/func? 331 of ~20K combinations. Mostly 1 func/fold, but some versatile scaffolds -- TIM most versatile. Similar for interactions. Func. Divergence vs. Seq. & Struc. Diverg.

4 Using Folds to Interpret Expression Data.

Top-10 parts in other terms. Enriched in transcriptome: VGA, αβ folds, energy, synthesis, Cytoplasmic, TIM fold. Depleted: NS, long, TMs, transport, transcription, Nuclear, Leu-zip fold. Bayesian localizer.

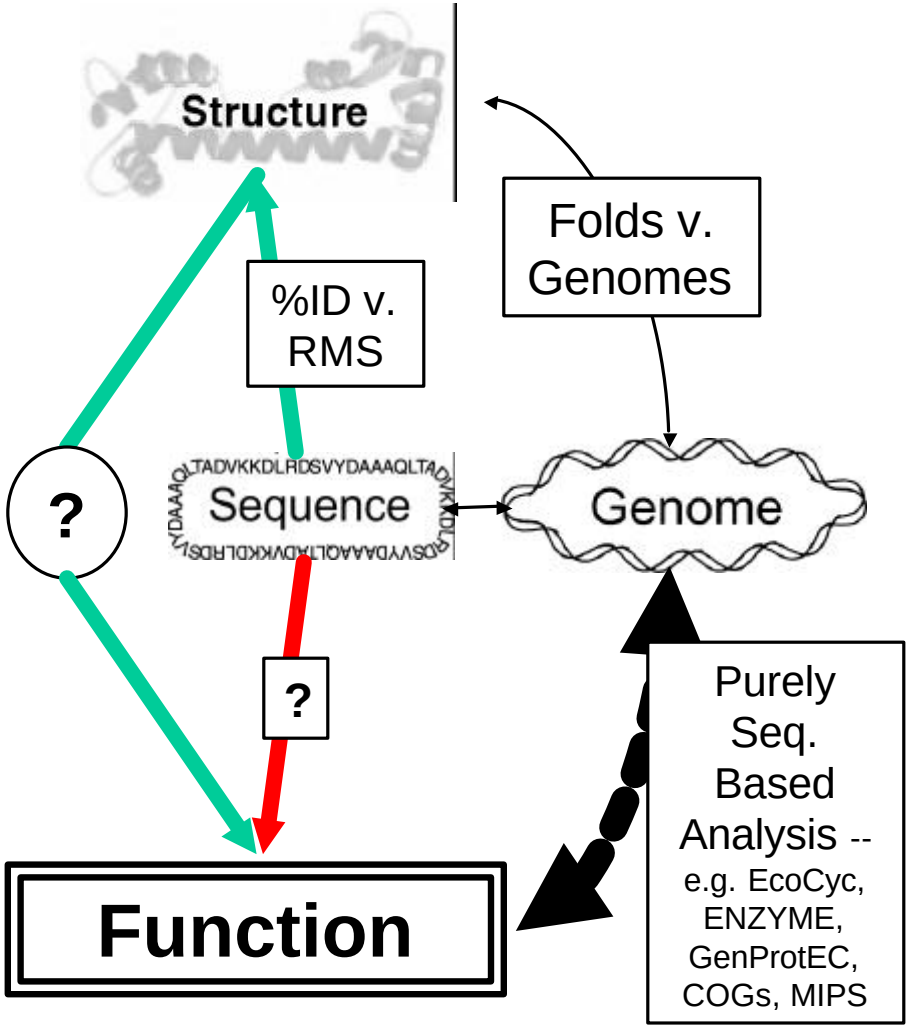
5 Tricky Issues: Relating Expression to Function.

Expression relates to structure & localization but to function, globally? Weak relation to protein-protein interactions.

H Hegyi, J Lin, B Stenger, N Echols, P Bertone, J Qian, L Regan, S Balasubramanian, S Teichmann, G Montelione, A Edwards, C Wilson, Y Kluger, C Arrowsmith, A Drawid, R Jansen, D Greenbaum, S Teichmann, P Harrison

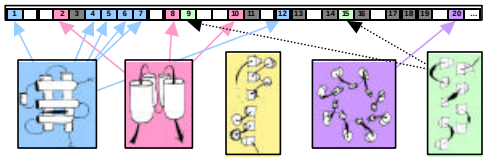
bioinfo.mbb.yale.edu

Adding Structure to Functional Genomics, Function to Structural Genomics

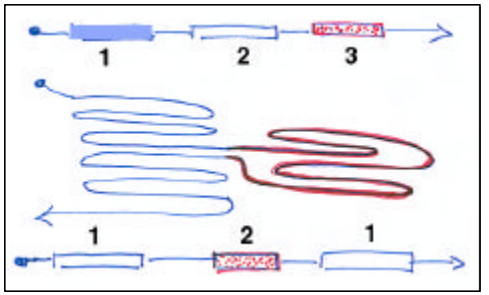


Why Structure? Do we really need it?

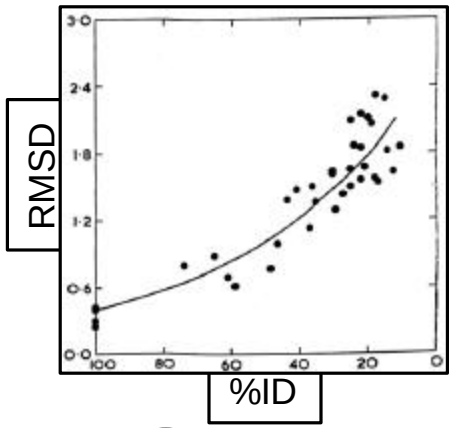
1 Most Highly Conserved



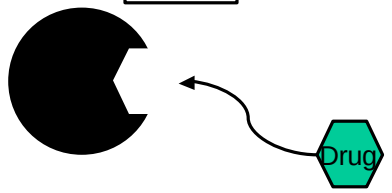
2 Precisely Defined Modules



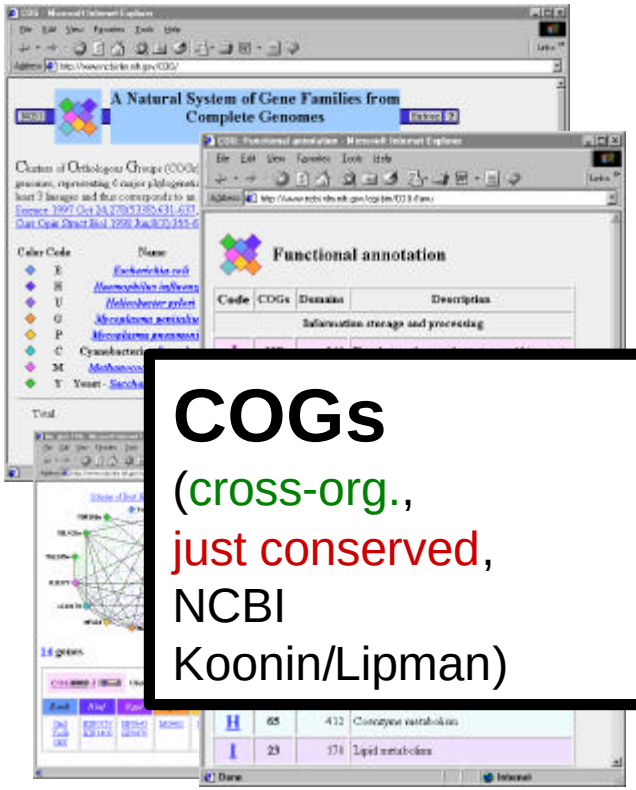
3 Seq. ⇔ Struc. Clearer than Seq. ⇔Func.



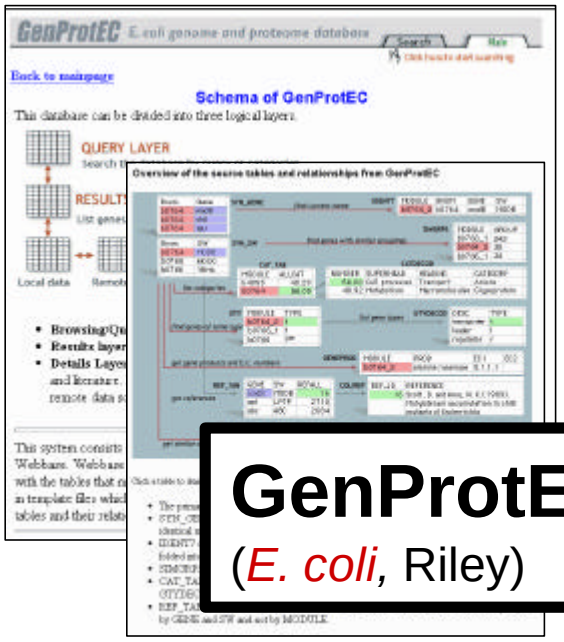
4 Link to Chemistry, Drugs



Functional Classification

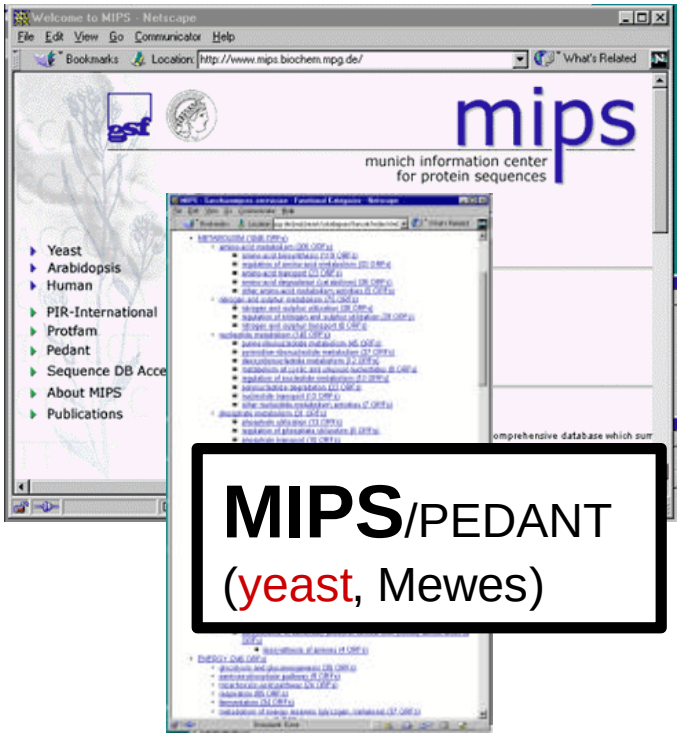


COGs
(cross-org.,
just conserved,
NCBI
Koonin/Lipman)

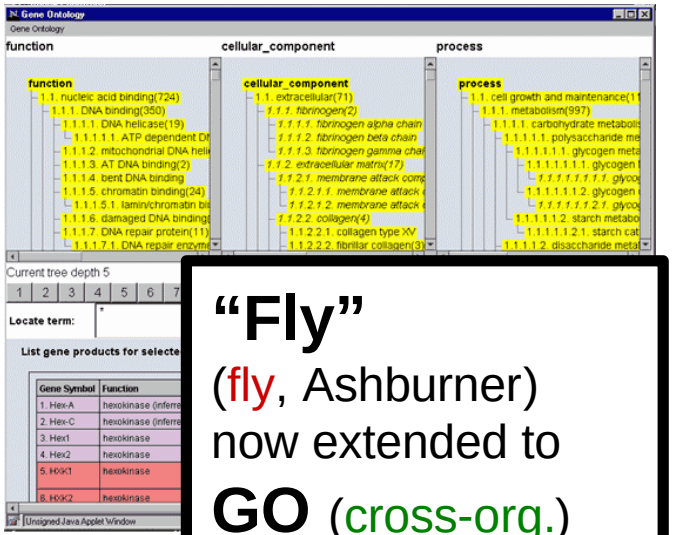


GenProtEC
(*E. coli*, Riley)

ENZYME
(SwissProt
Bairoch/
Apweiler,
just enzymes,
cross-org.)



MIPS/PEDANT
(yeast, Mewes)



"Fly"
(fly, Ashburner)
now extended to
GO (cross-org.)

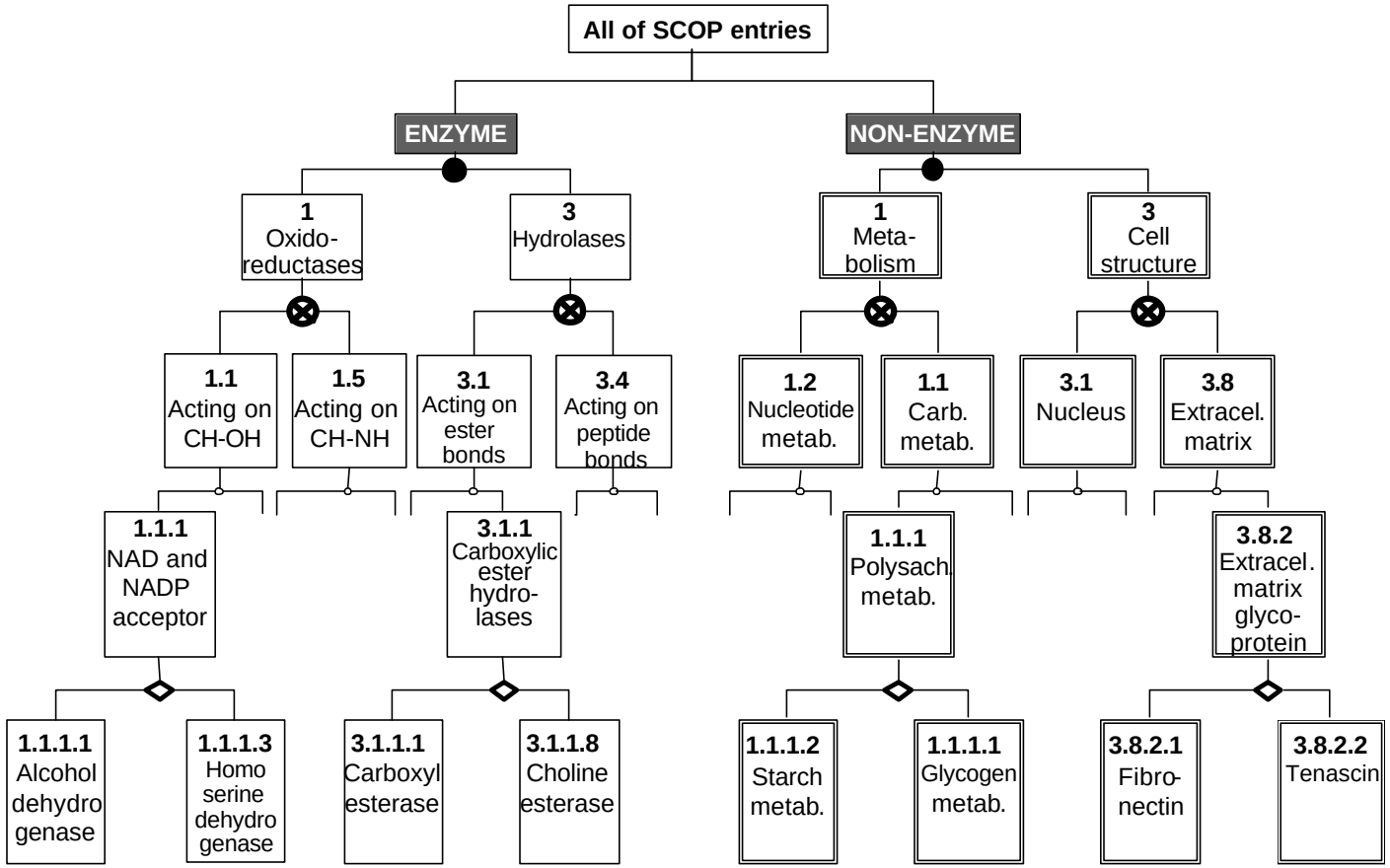
Also:

Other
SwissProt
Annotation

WIT, KEGG
(just pathways)

TIGR EGAD
(human ESTs)

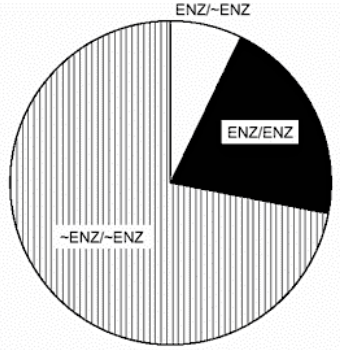
A Simple Scheme for Functionally Classifying Protein Structures



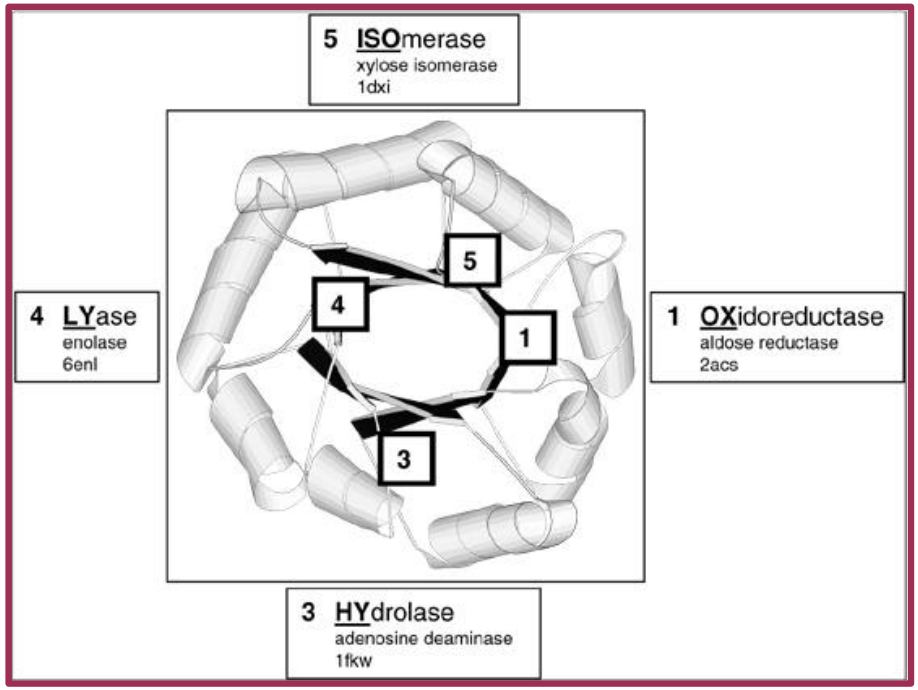
Focus on Pairs with Precise (1.1.1.*) and Broad (1.*) Similarity

A Combined Scheme, merging ENZYME + FLY, with some annotation from MIPS, GenProtEC, & manual additions for proteins not in either (e.g. Ig's)

Also: COGs



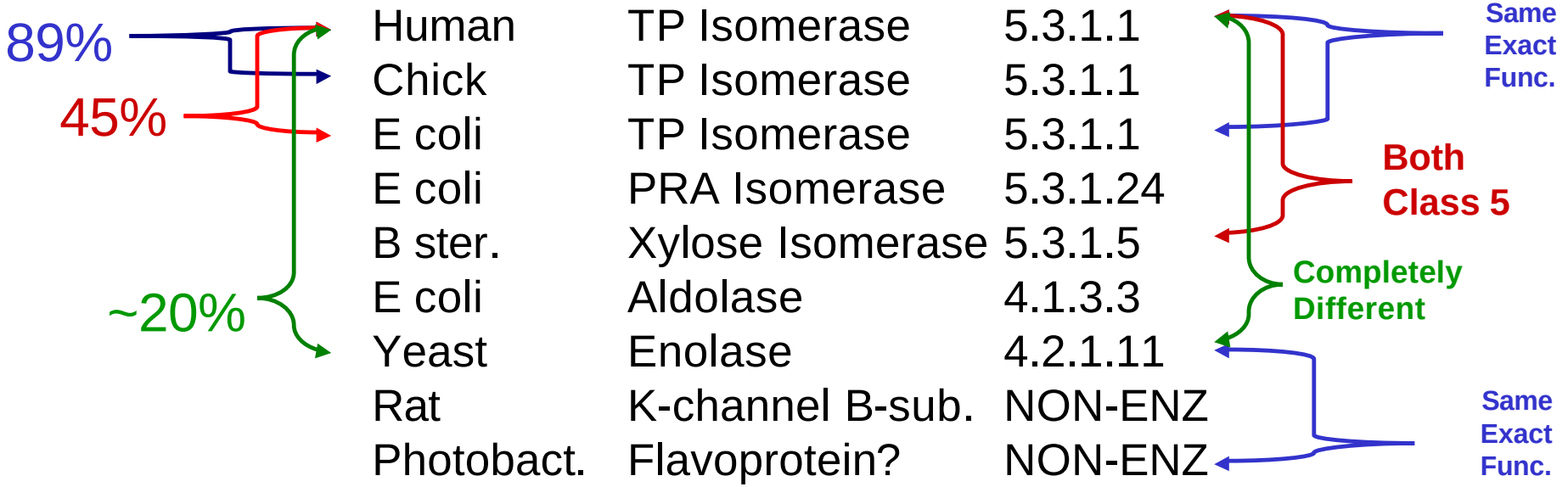
Fold-Function Combinations #1



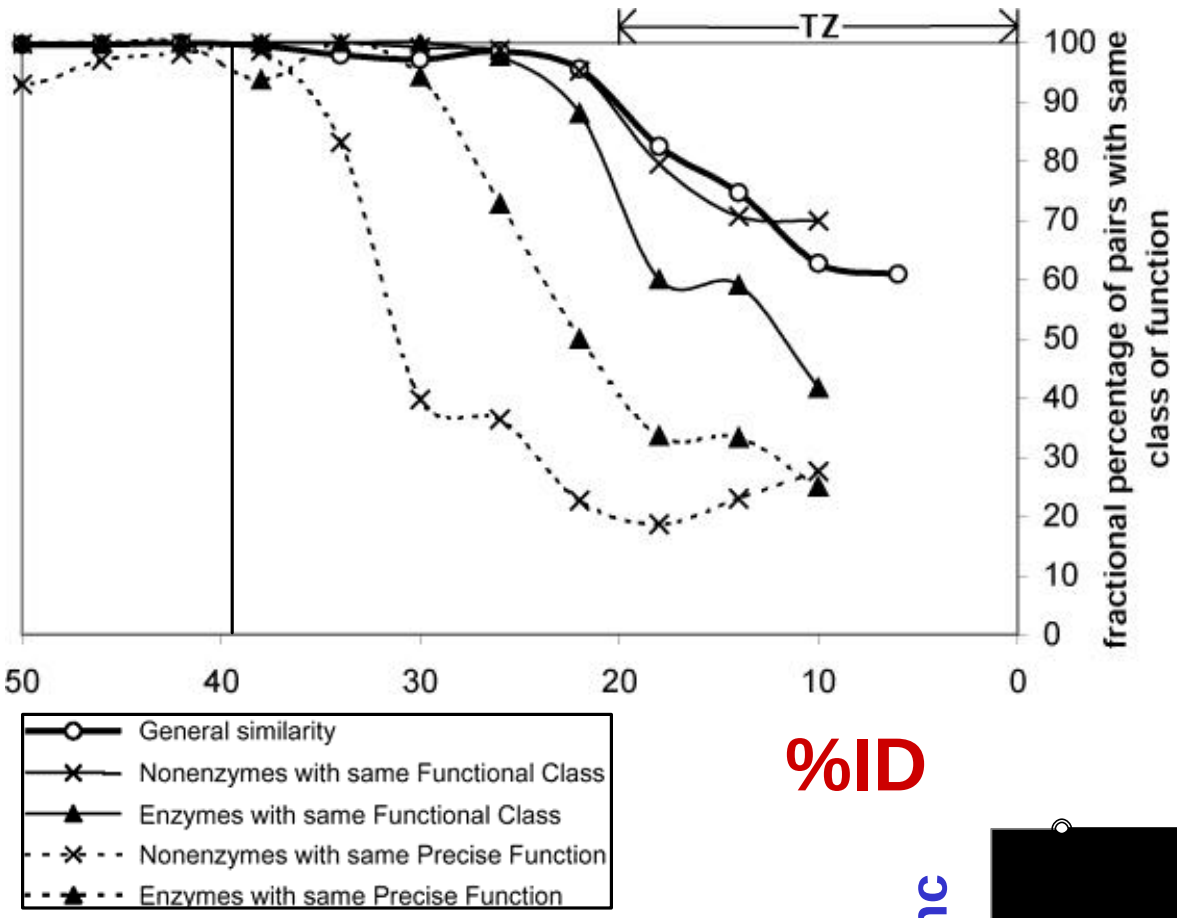
Many Functions on the Same Fold
-- e.g. the TIM-barrel

at what degree of divergence?
Sequence Diverg. (**%ID, P_{seq}**)
Structural Diverg. (**RMS, P_{str}**)
Functional Diverg. (**%SameFunc**)

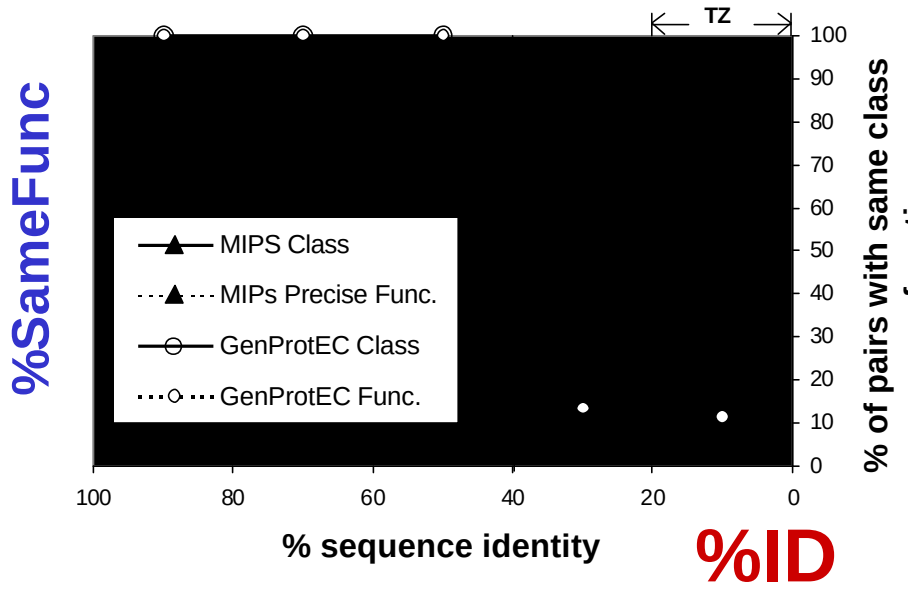
Compare large number of pairs of sequences that have same fold but different functions.



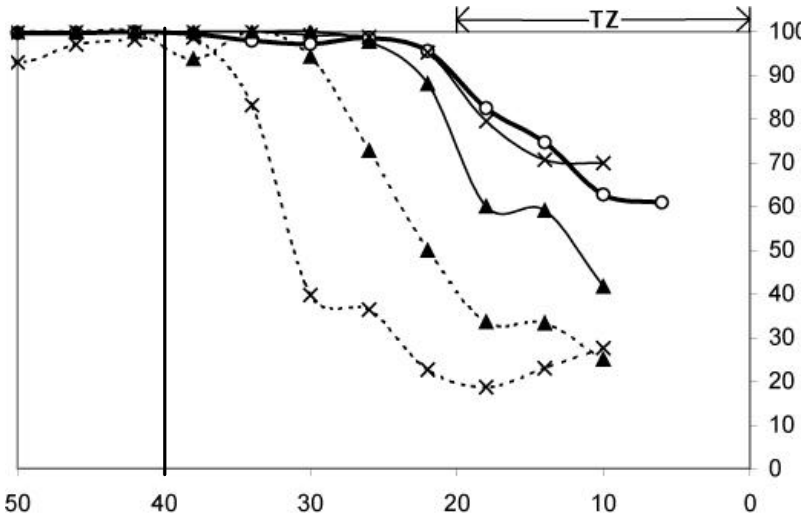
Relationship of Similarity in Sequence & Structure to that in Function



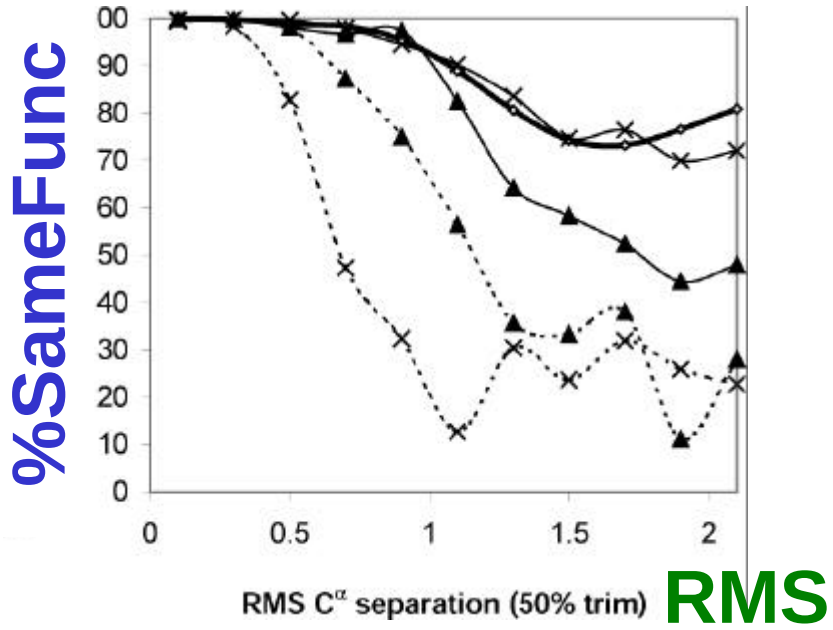
See at what %ID have diff. function (both broad & precise). Use 4 func. classifications -- ENZYME, FLY (+extra), MIPS, GenProtEC

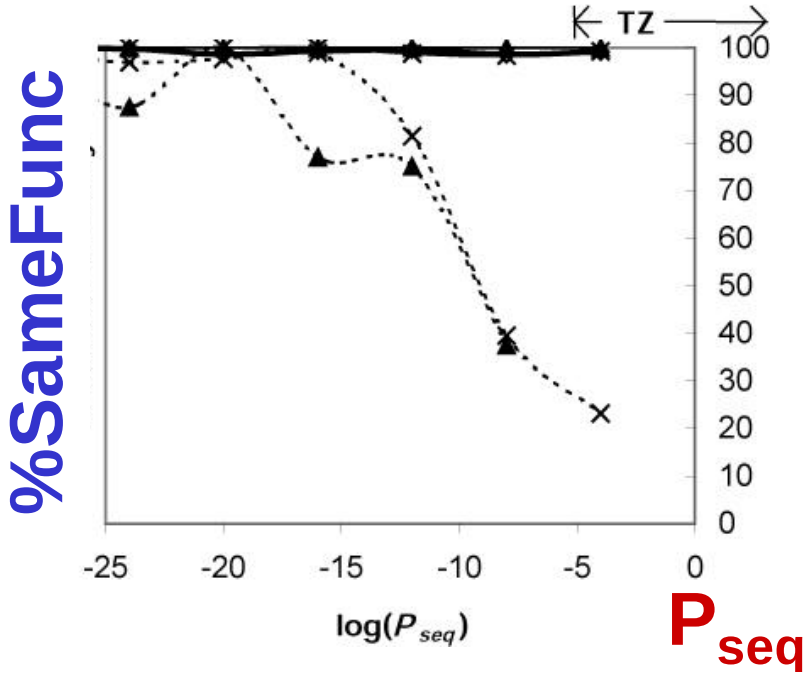


Relationship of Similarity in Sequence & Structure to that in Function II



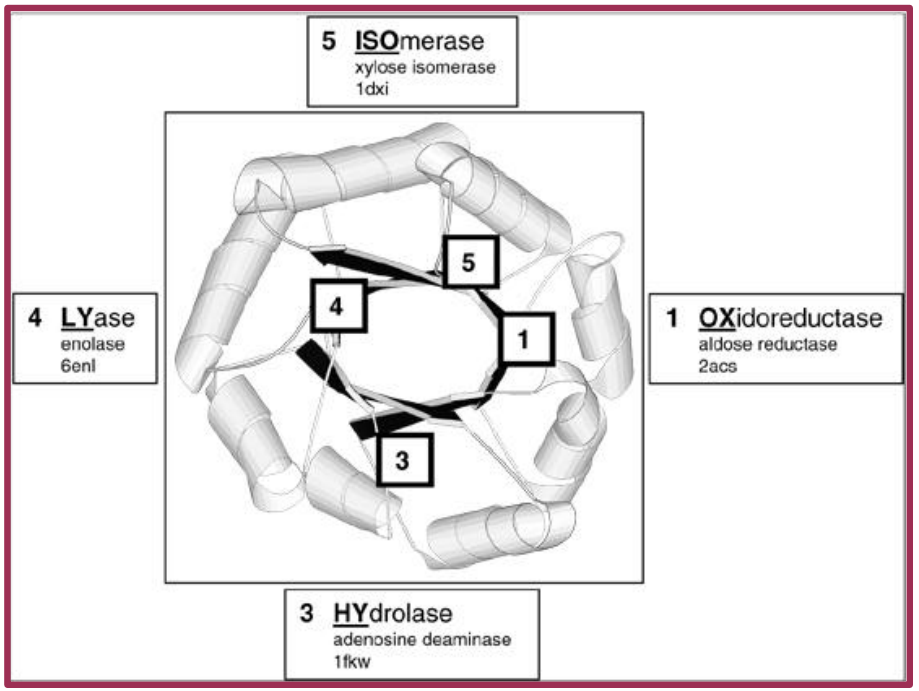
Percent identity quite successful vs. structure sim. or statistical scores



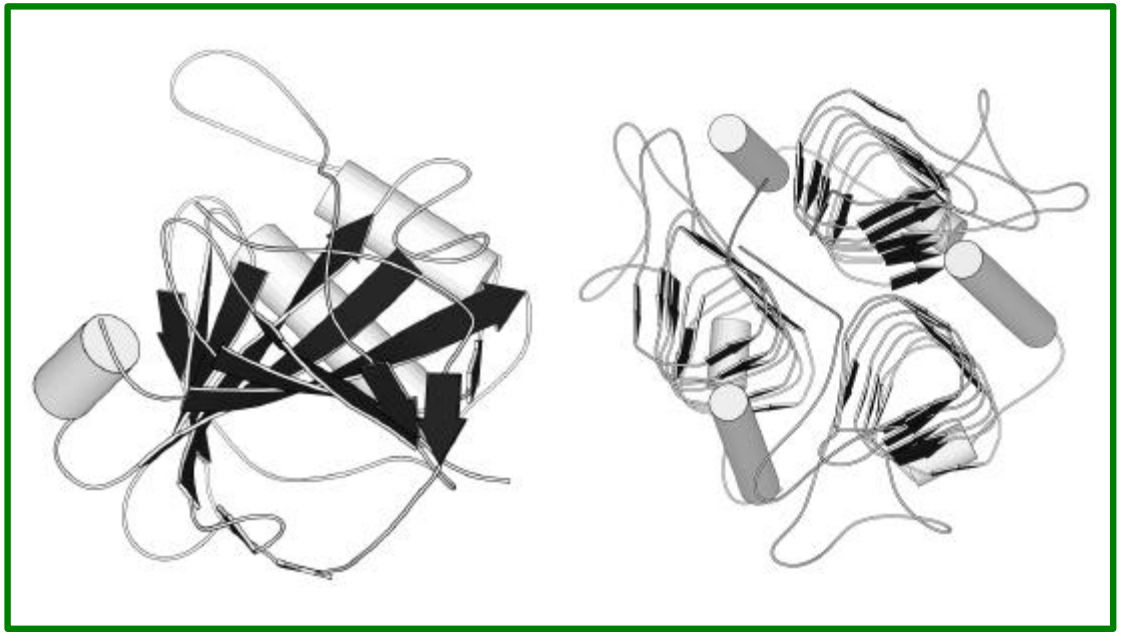


Fold-Function Combinations #2

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



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

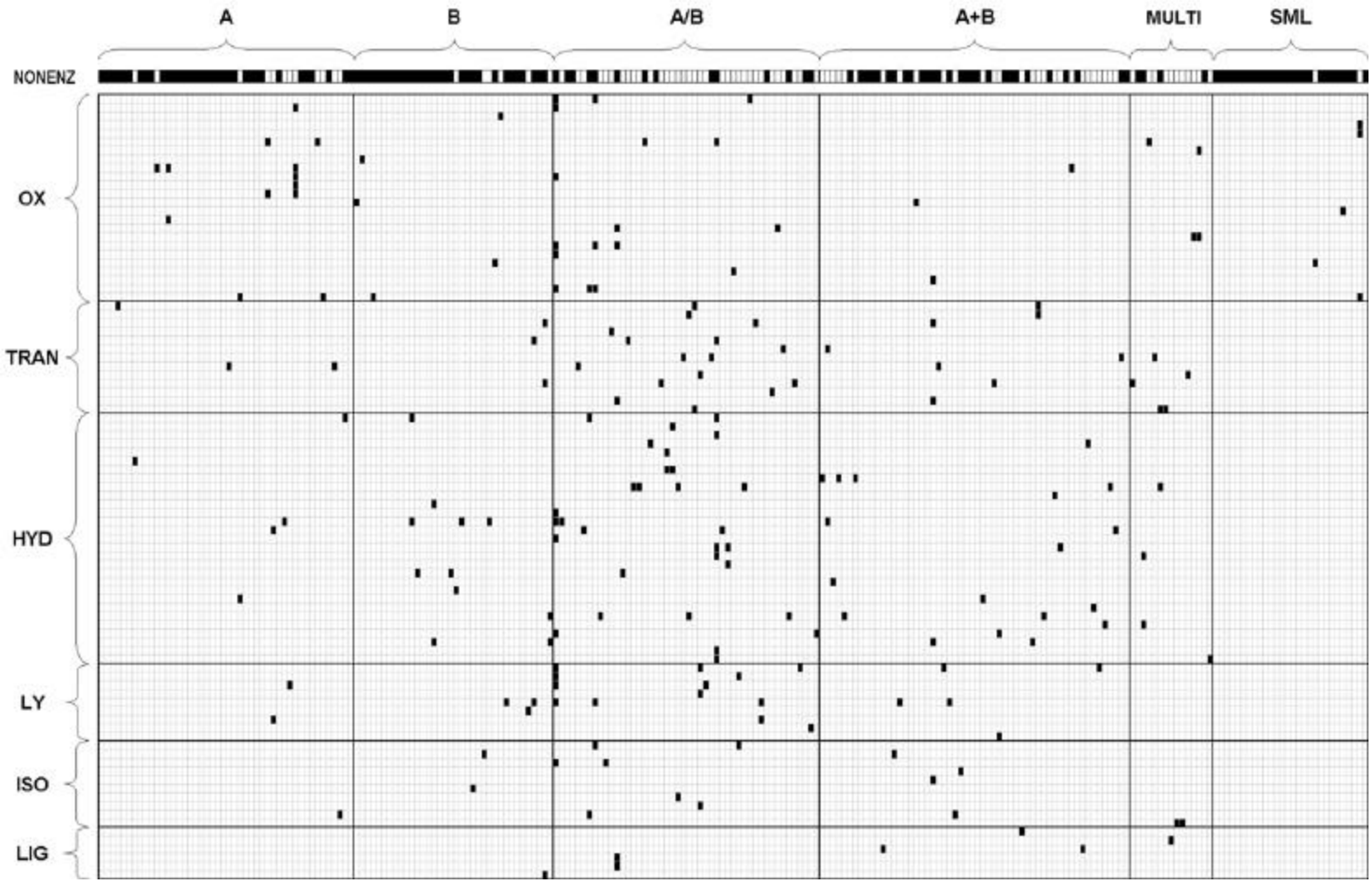


Fold-Function Combinations

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

229 Folds

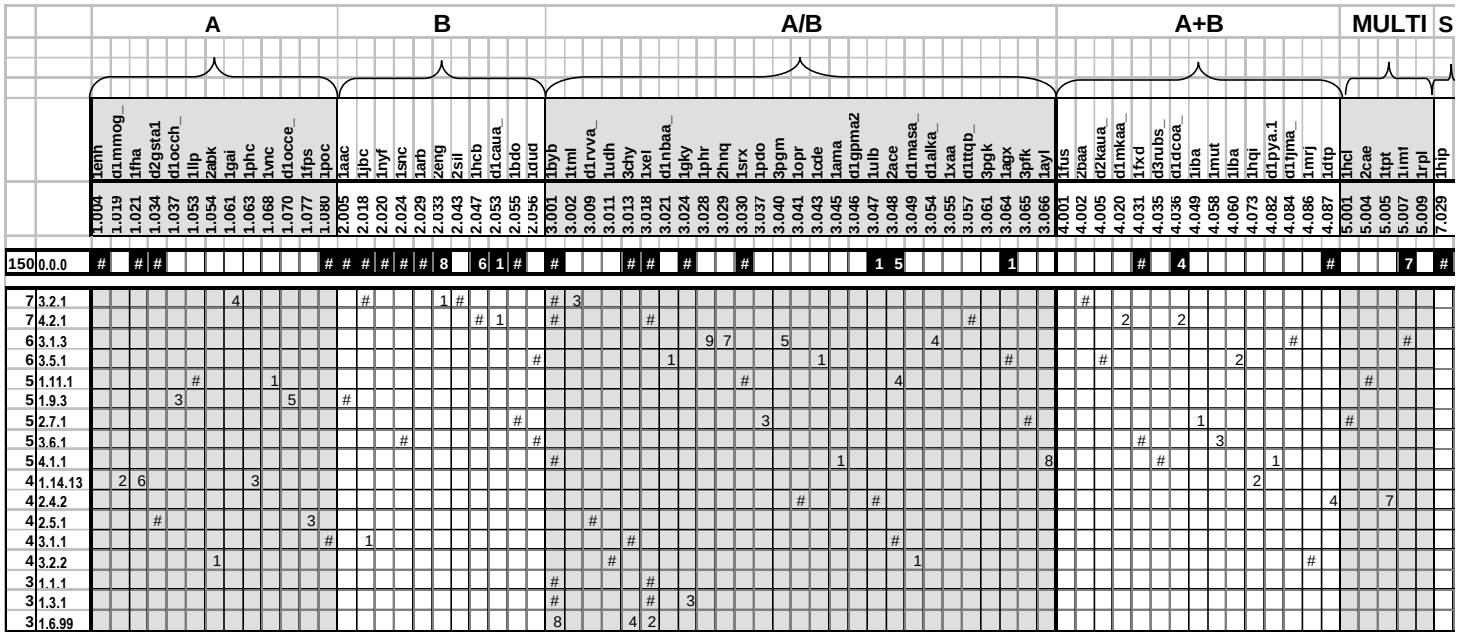
91 Enzymatic Functions
+ Non-Enzyme



Most Versatile Functions

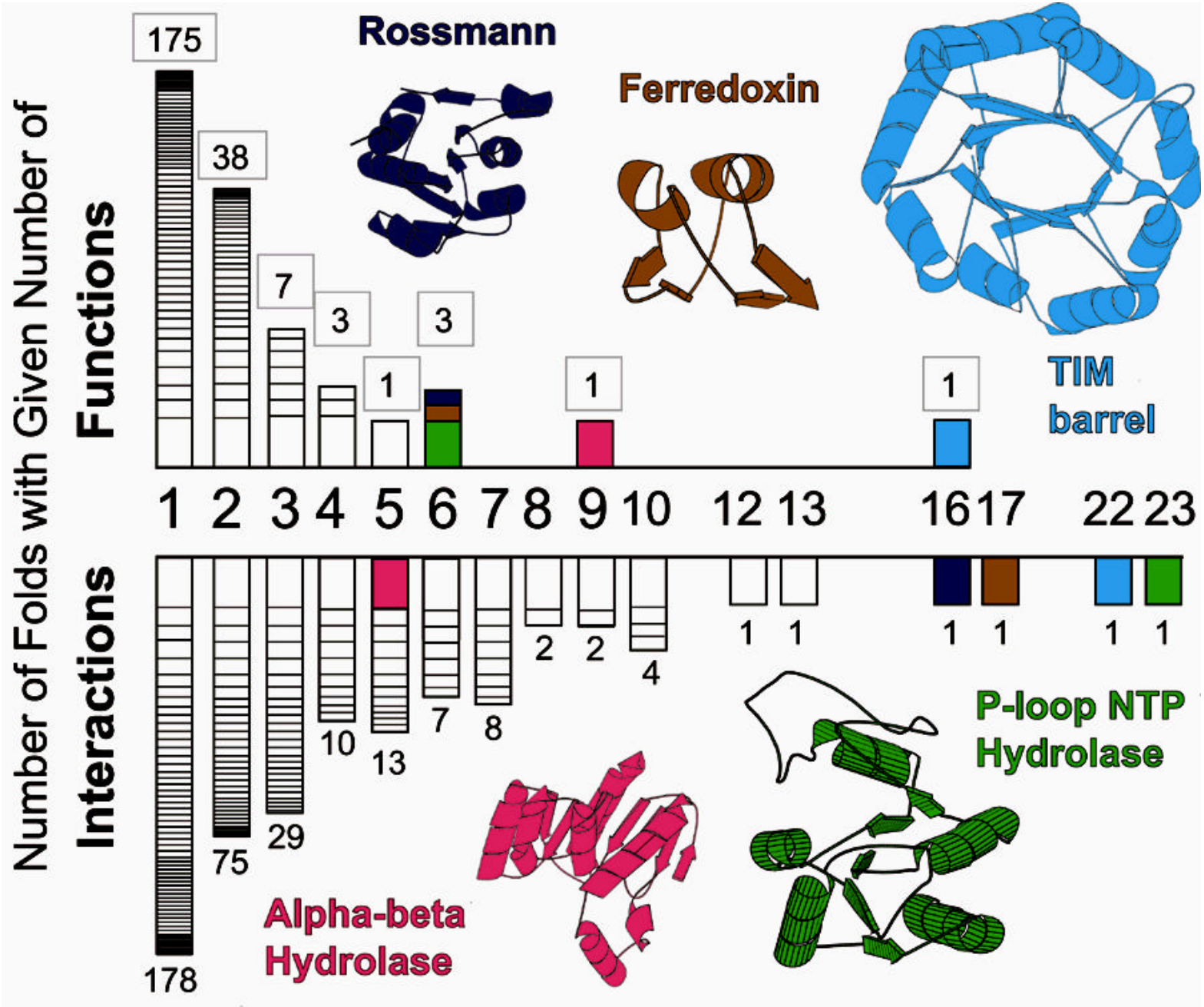
Sort table rows….

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)

Most Versatile Folds – Relation to Interactions



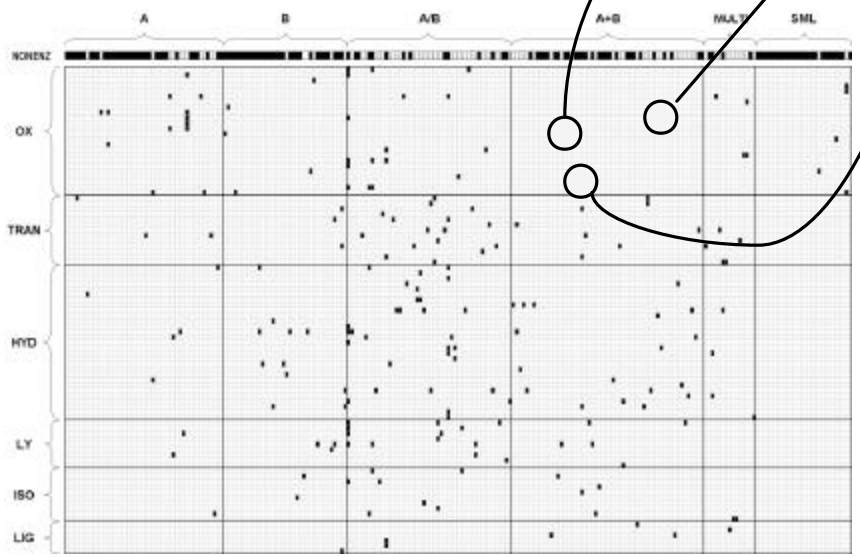
Similar results Martin et al. (1998)

The number of interactions for each fold = the number of other folds it is found to contact in the PDB

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]

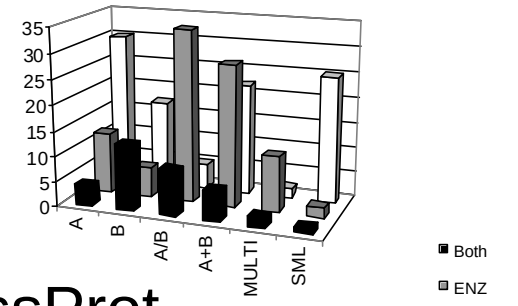
Compare Classifications and Genomes

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

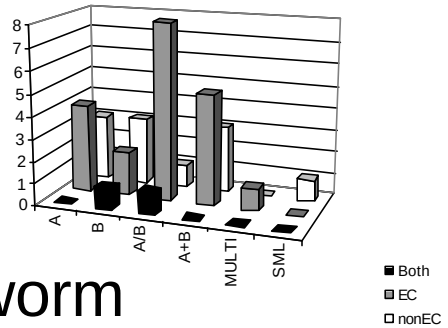
ENZYME \ SCOP	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

CATH (Thornton)

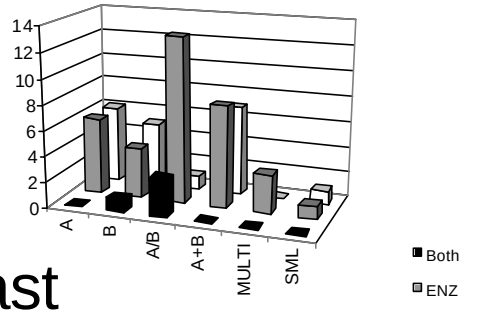
ENZYME \ CATH	A	B	AB
NONENZ	10	9.0	15
OX	5.1	5.1	10
TRAN		1.3	13
HYD	2.6	1.3	14
LY		2.6	1.3
ISO	1.3	1.3	5.1
LIG			1.3



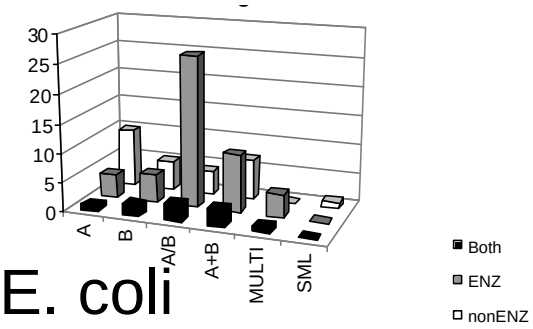
SwissProt



worm



Yeast

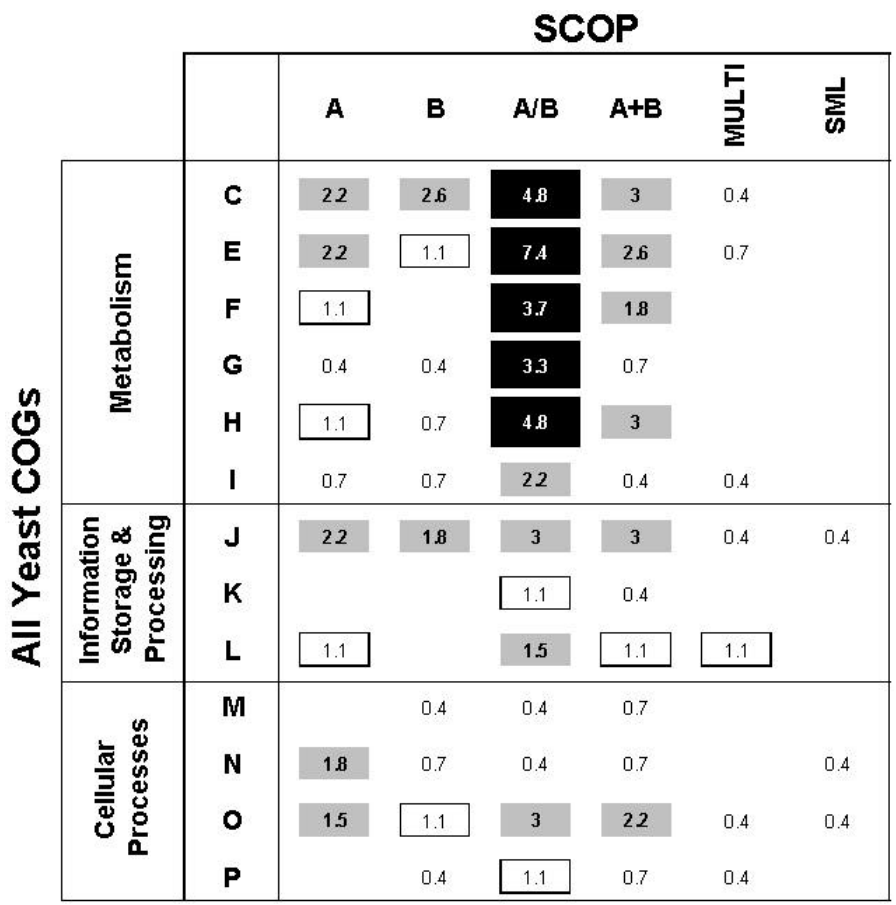


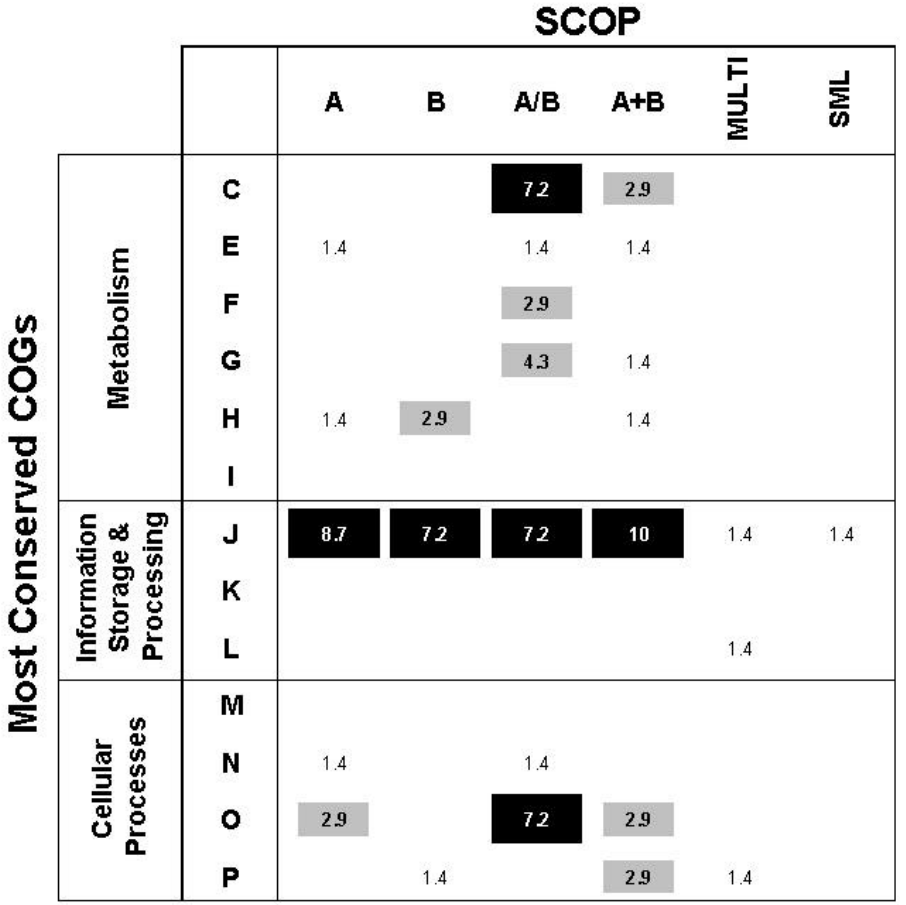
E. coli

MIPS YFC (Mewes)

MIPS Functional Cat.		SCOP 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	4.0	3.6	4	4.3	1.6	1.2
transcription	4	1.5	1.9	3.3	1.5	0.5	0.8
protein syn thesis	5	1	0.9	0.7	1.3	0.3	0.2
protein targeting	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	3.6	1.8	0.7	0.5
ionic homeostasis	13	0.5	0.9	0.4	0.4	0.2	

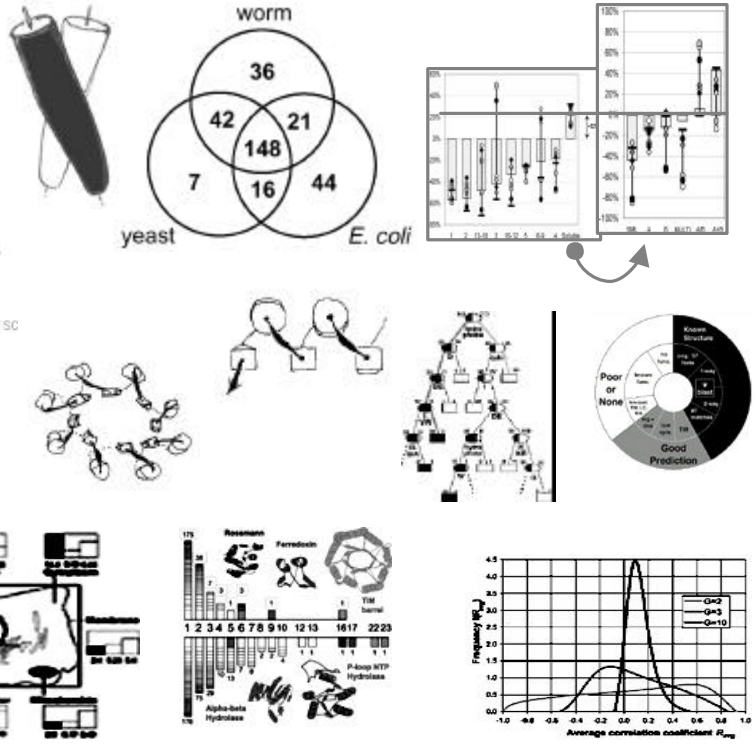
Different Structure Function Relationships for Most Ancient Proteins





(Scop, Murzin/Chothia; COGs, Koonin/Lipman)

Comparative Genomics: Surveys of a Finite Parts List



1 Using Folds to Interpret Genomes.

Fold Library background. Shared and/or unique parts. Venn Diagram, Fold tree with all-β diff. Ortholog tree. Horizontal Transfer. Commor Parts: Top-10 folds with βαβ. Common Ψ-fold.

- **Tricky Issues →Expt. SG.** Extent of fold assignment (MG, 65%, 98). Predictions. Biases. NESG.org Construct DB. Datamining this w/ decision trees. Selecting weird MG CD targets.
- **Folds & Functions.** Roles/part? How many folds/func? 331 of ~20K combinations. Mostly 1 func/fold, but some versatile scaffolds -- TIM most versatile. Similar for interactions. Func. Divergence vs. Seq. & Struc. Diverg.

4 Using Folds to Interpret Expression Data.

Top-10 parts in other terms. Enriched in transcriptome: VGA, αβ folds, energy, synthesis, Cytoplasmic, TIM fold. Depleted: NS, long, TMs, transport, transcription, Nuclear, Leu-zip fold. Bayesian localizer.

5 Tricky Issues: Relating Expression to Function.

Expression relates to structure & localization but to function, globally? Weak relation to protein-protein interactions.

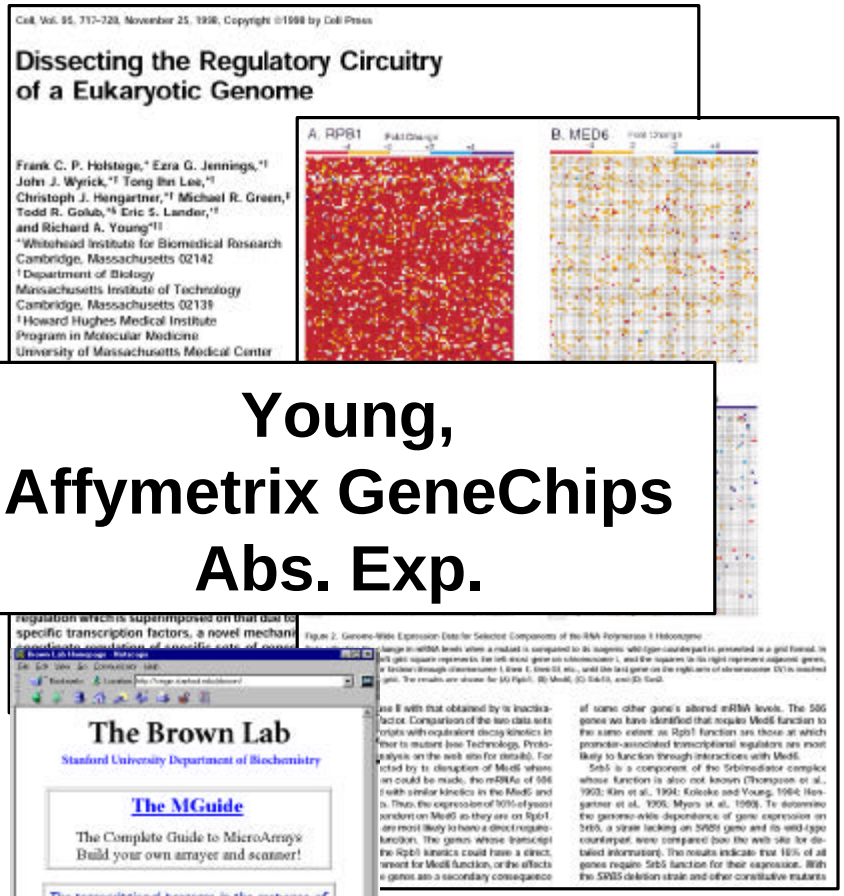
H Hegyi, J Lin, B Stenger, N Echols, P Bertone, J Qian, L Regan, S Balasubramanian, V Alexandrov, G Montelione, A Edwards, C Wilson, Y Kluger, C Arrowsmith, A Drawid, R Jansen, D Greenbaum, S Teichmann, P Harrison

bioinfo.mbb.yale.edu

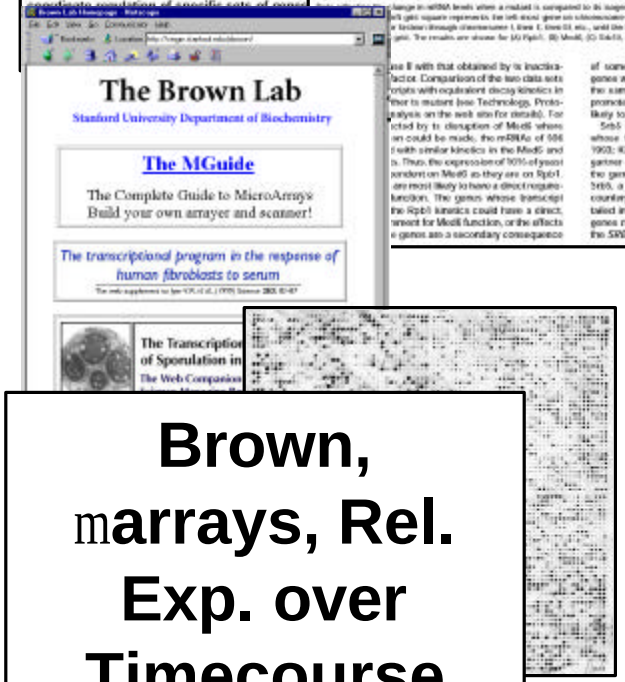
Gene Expression Datasets: the Transcriptome

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

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

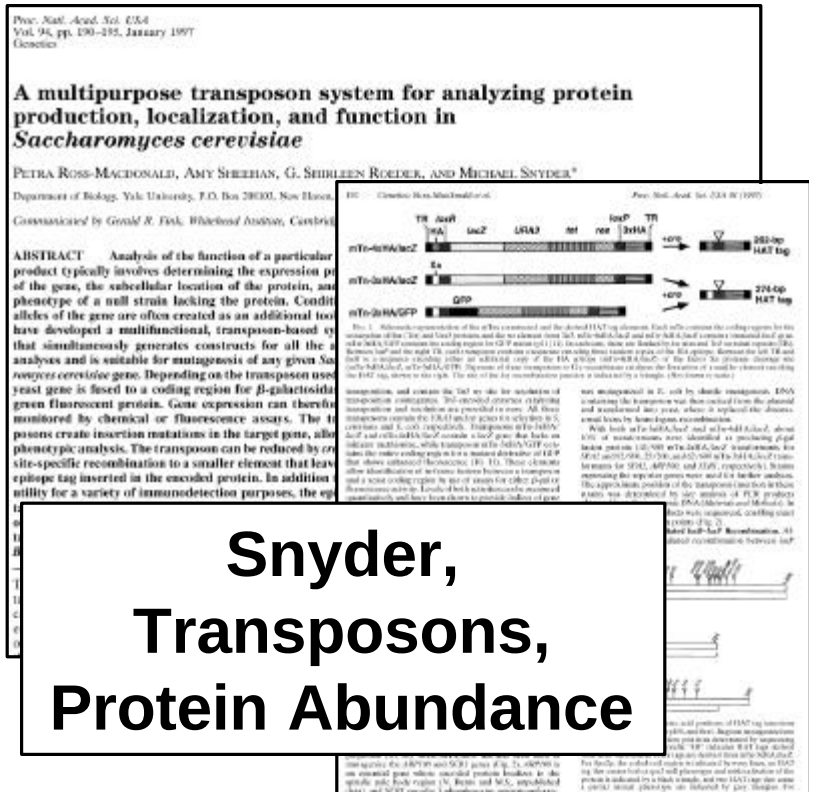


Young, Affymetrix GeneChips Abs. Exp.



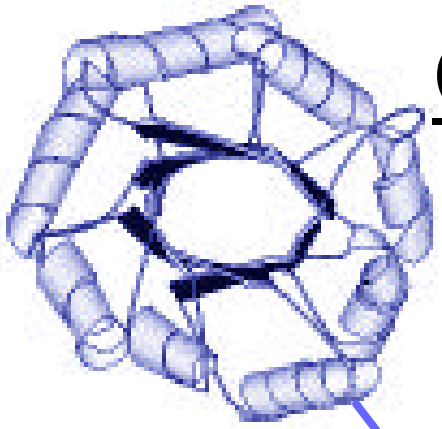
Brown, marrays, Rel. Exp. over Timecourse

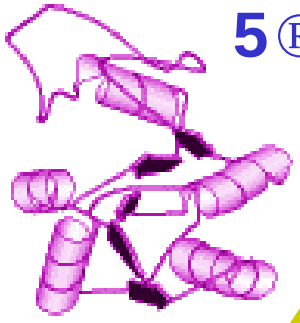
Also: SAGE (mRNA); 2D gels for Protein Abundance (Aebersold, Futcher)



Snyder, Transposons, Protein Abundance

Common Parts: the Transcriptome





5 ® 1

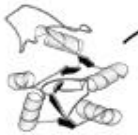
1 ® 18

7 ® 15

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
Leu-zipper	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

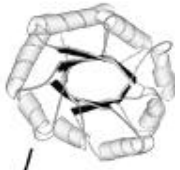
Feature *F* is Folds, in particular the TIM-barrel (3.1)	Number of TIM-barrel fold matches in yeast genome	Number of matches with all folds in yeast genome	Genome composition of TIM-barrel fold matches	Number of TIM-barrel fold matches weighted by expression	Number of matches with all folds weighted 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%**







Fold of	Freq. Genome	Freq. Transcriptome	Change CDC28	Change CDC15	Change Diauxic Shift	Change Sporulation	Change *E. coli* heat shock	Rep. PDB
Protein kinases (cat. core)	1	18	94	98	139	60	100	1p38
β-propeller	2	5	160	108	109	82	-	1mda
P-loop NTP hydrolases	3	2	100	88	91	57	39	1gky
α–α superhelix	4	6	136	90	110	44	55	2bct
TIM-barrel	5	1	58	57	39	24	91	1byb
Ferredoxin-like	6	3	135	61	63	70	144	1fxd
Rossmann fold	8	4	55	99	43	56	92	1xel
Ribonucleotide reductase (R1)	100	143	1	-	-	-	35	1rlr
ATPase dom. of HSP90	100	91	2	4	72	73	2	1ah6
Homing endonuclease-like	130	164	3	136	85	175	41	1af5
Aminoacid dehydrogenases; dim. dom.	-	-	4	169	121	3	51	1hup
DNA topo I (N-term)	-	-	175	1	148	126	-	1ois
DNA clamp	130	115	8	2	87	11	60	2pol
Metallothionein	100	14	89	3	33	12	-	1mhu
Phosphoenolpyruvate carboxykinase	130	190	51	26	1	96	169	1ayl
Citrate synthase	81	120	14	8	2	28	51	1csh
N-carbamoylsarcosine amidohydrolase	130	112	9	-	3	138	118	1nba
TBP-like	81	91	46	38	4	75	100	1bv1
5'-3' exonuclease	67	150	32	125	162	1	157	1tfr
α/α toroid	62	132	169	145	114	2	100	1gai
Cyclin-like	20	61	20	15	129	4	-	1vin
ATPase domain of GroEL	36	34	183	143	61	151	1	1aon
Head domain of GrpE	130	135	196	31	165	165	3	1dkg
HSP70 (C-term)	31	10	16	11	56	117	4	1dkz





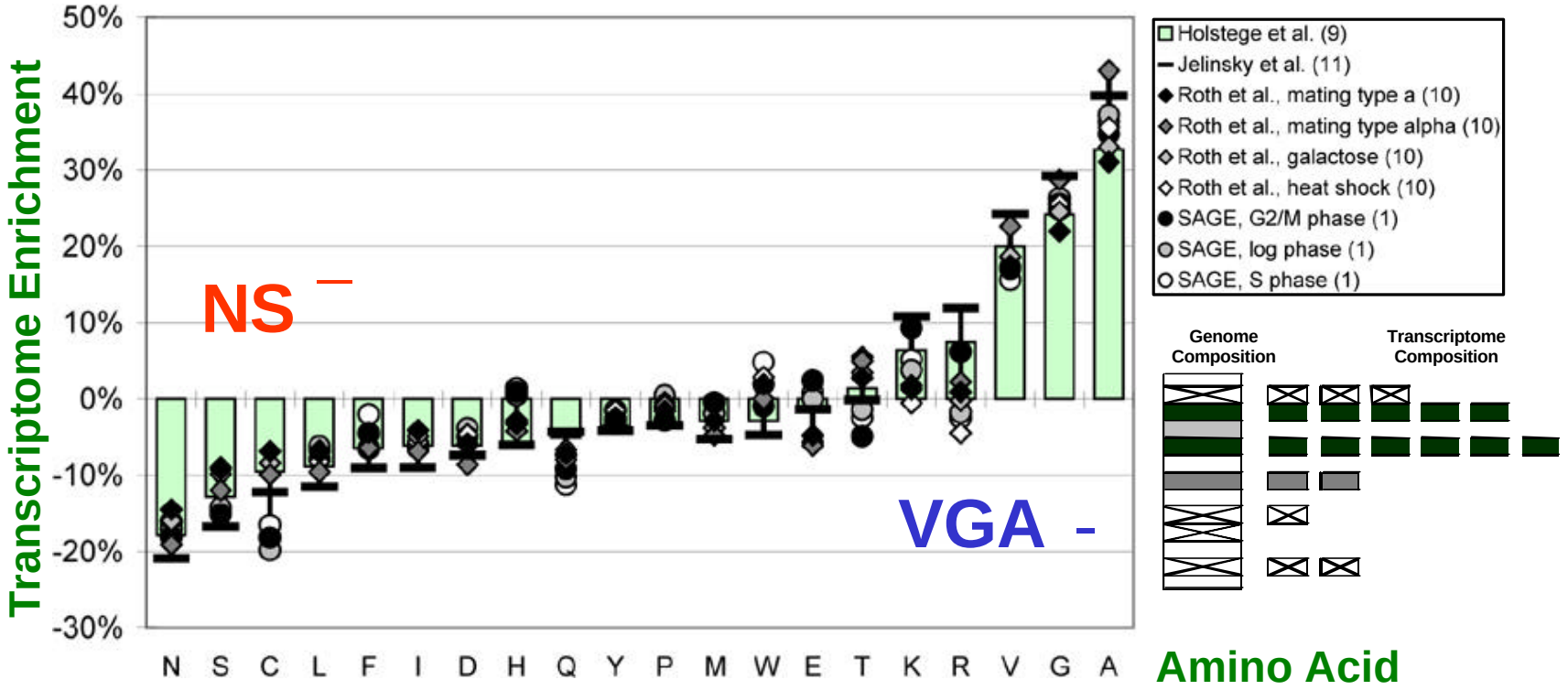
Common Folds

Changing Folds

Folds that change a lot in frequency are not common

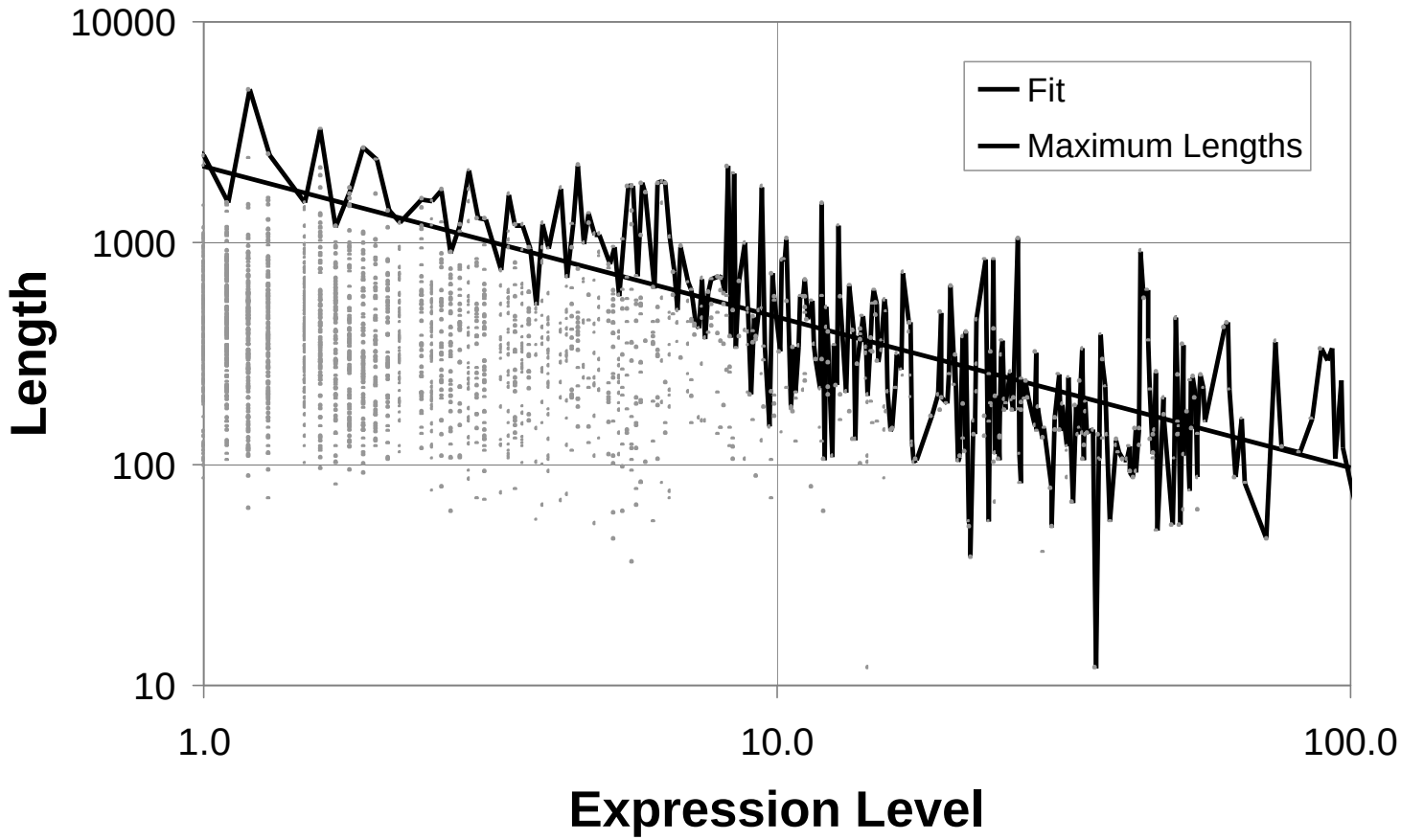
Composition of Genome vs. Transcriptome

	$\sum_{\text{orf } i} n_i(F)$	$\sum_{F} \sum_{\text{orf } i} n_i(F)$	$G(F)$	$\sum_{\text{orf } i} e_i n_i(F)$	$\sum_{F} \sum_{\text{orf } i} e_i n_i(F)$	$T(F)$	$D(F)$
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%**

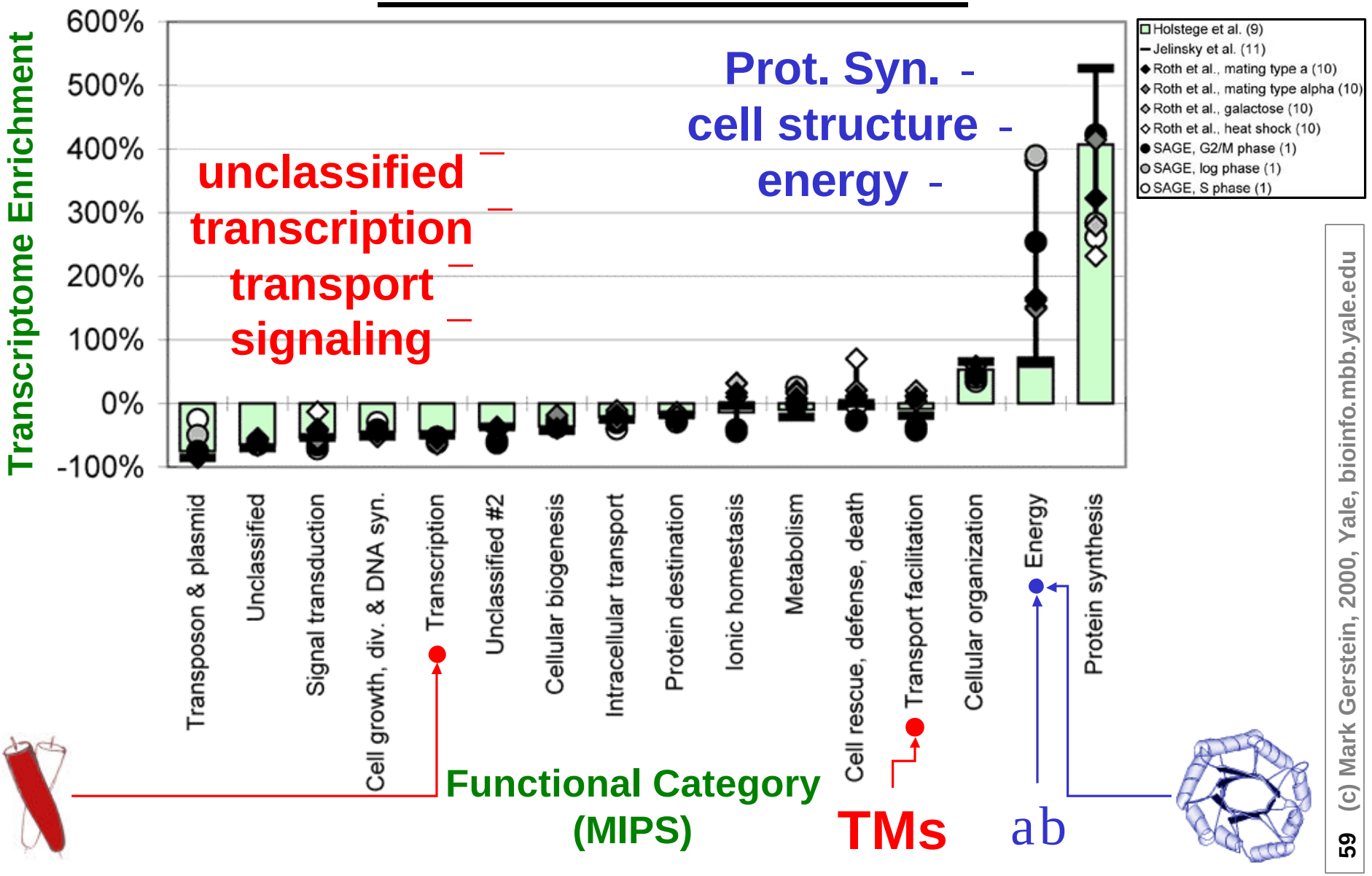


Relation between Length & Expression

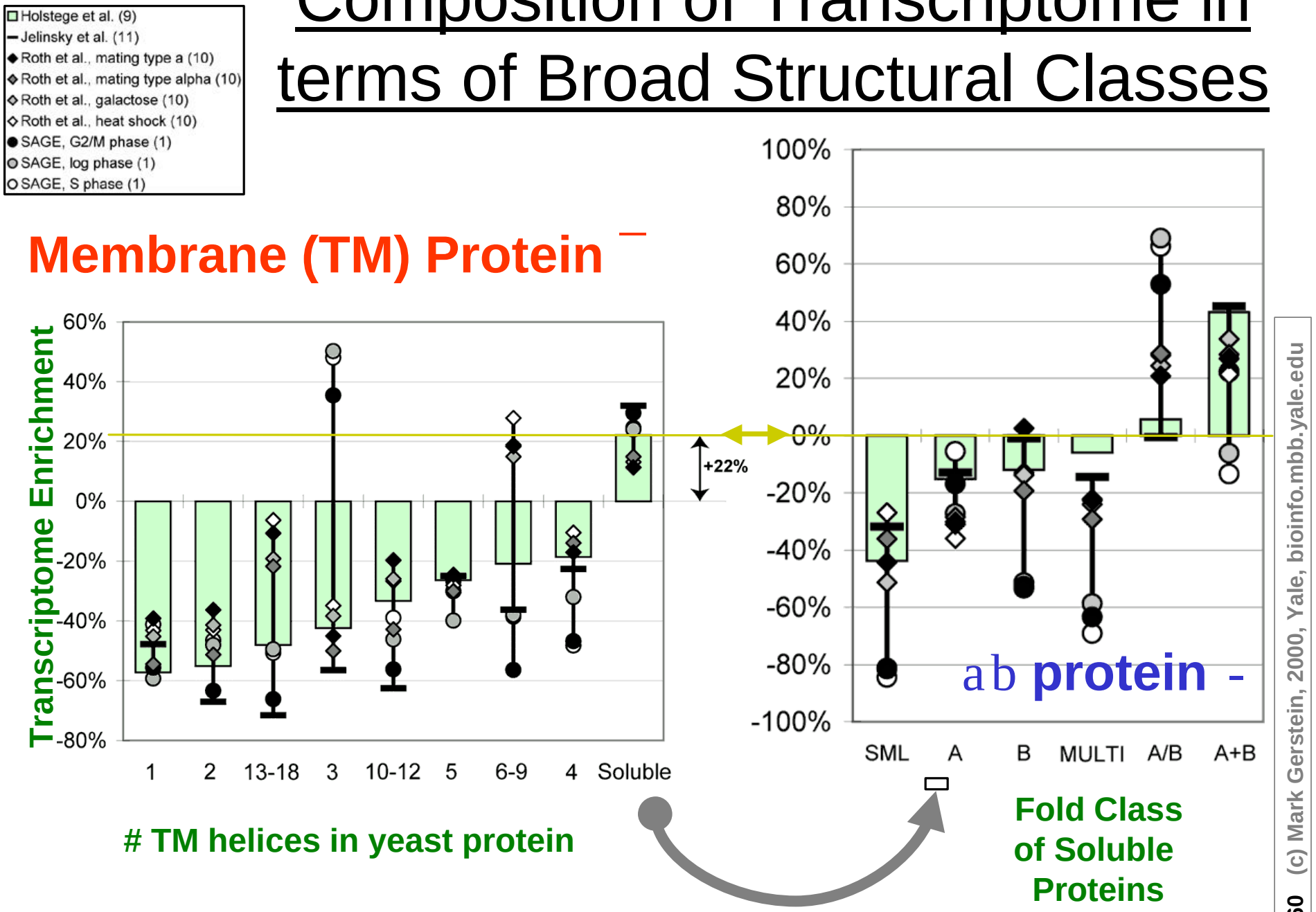
Max Expression (e.g. transcripts/cell) ~ (Length)$^{-2/3}$
Shorter proteins can be more highly expressed



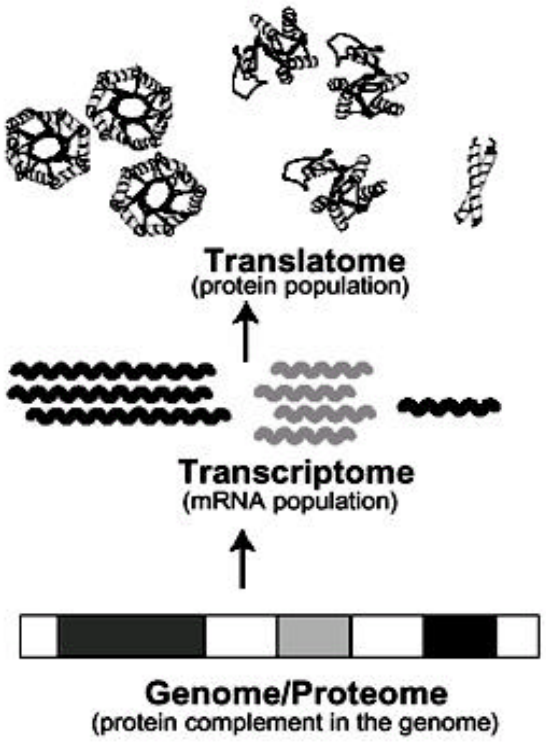
Composition of Transcriptome in terms of Functional Classes

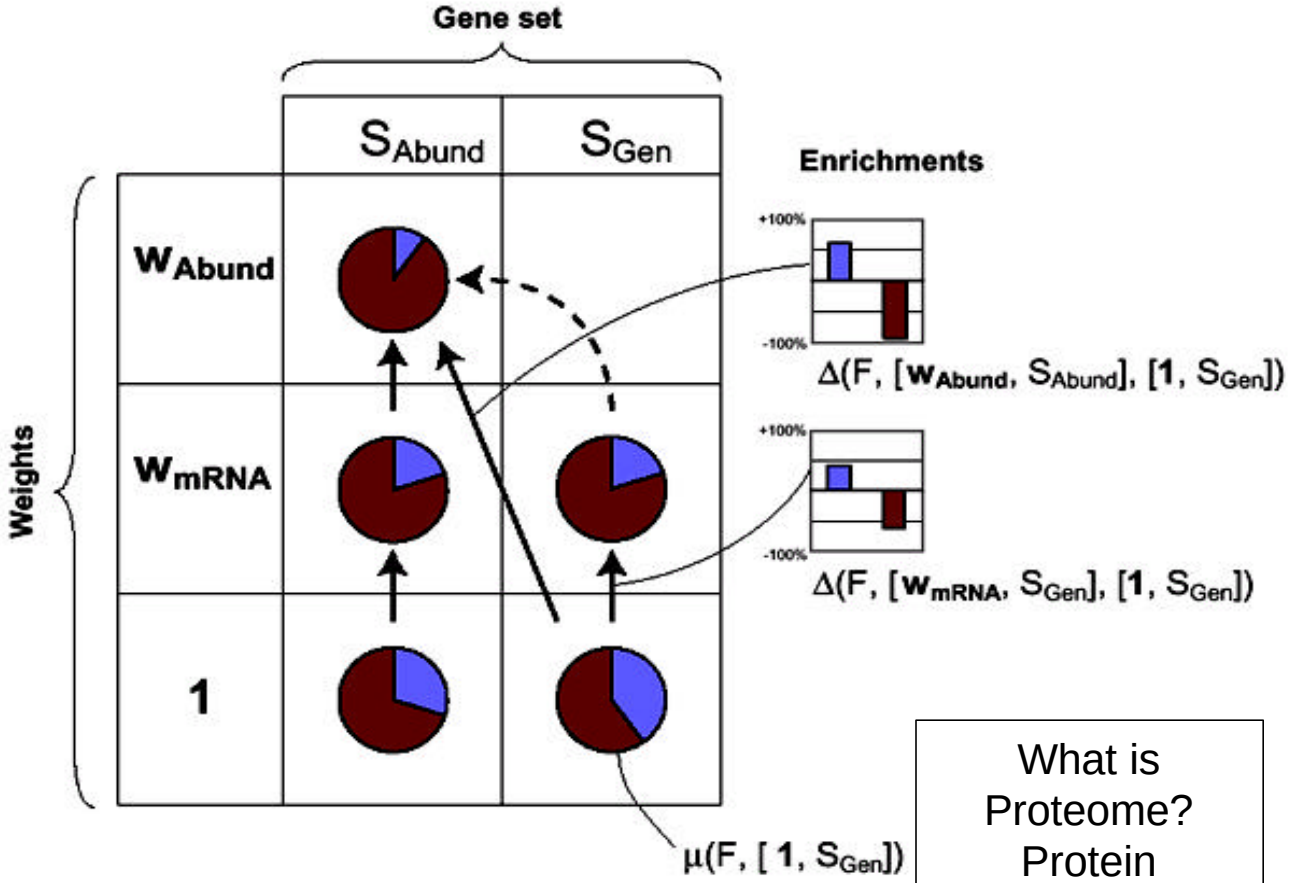


Composition of Transcriptome in terms of Broad Structural Classes



Relating the Transcriptome to Cellular Protein Abundance (Translatome)



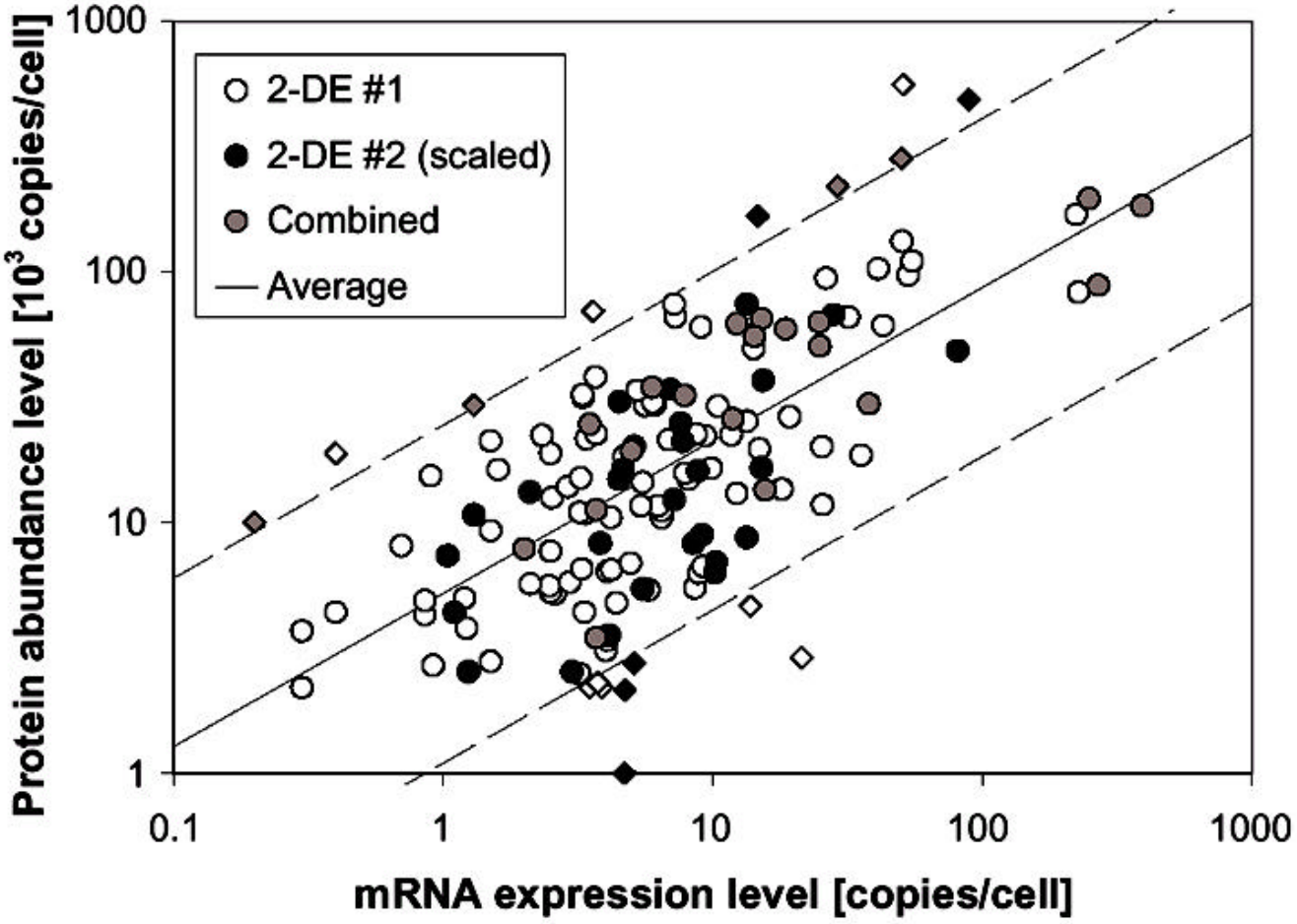


2D-gel electrophoresis Data sets: Futcher (71), Aebersold (156), scaled set with **171** proteins
New effect is dealing with gene selection bias

What is Proteome? Protein complement in genome or cellular protein population?

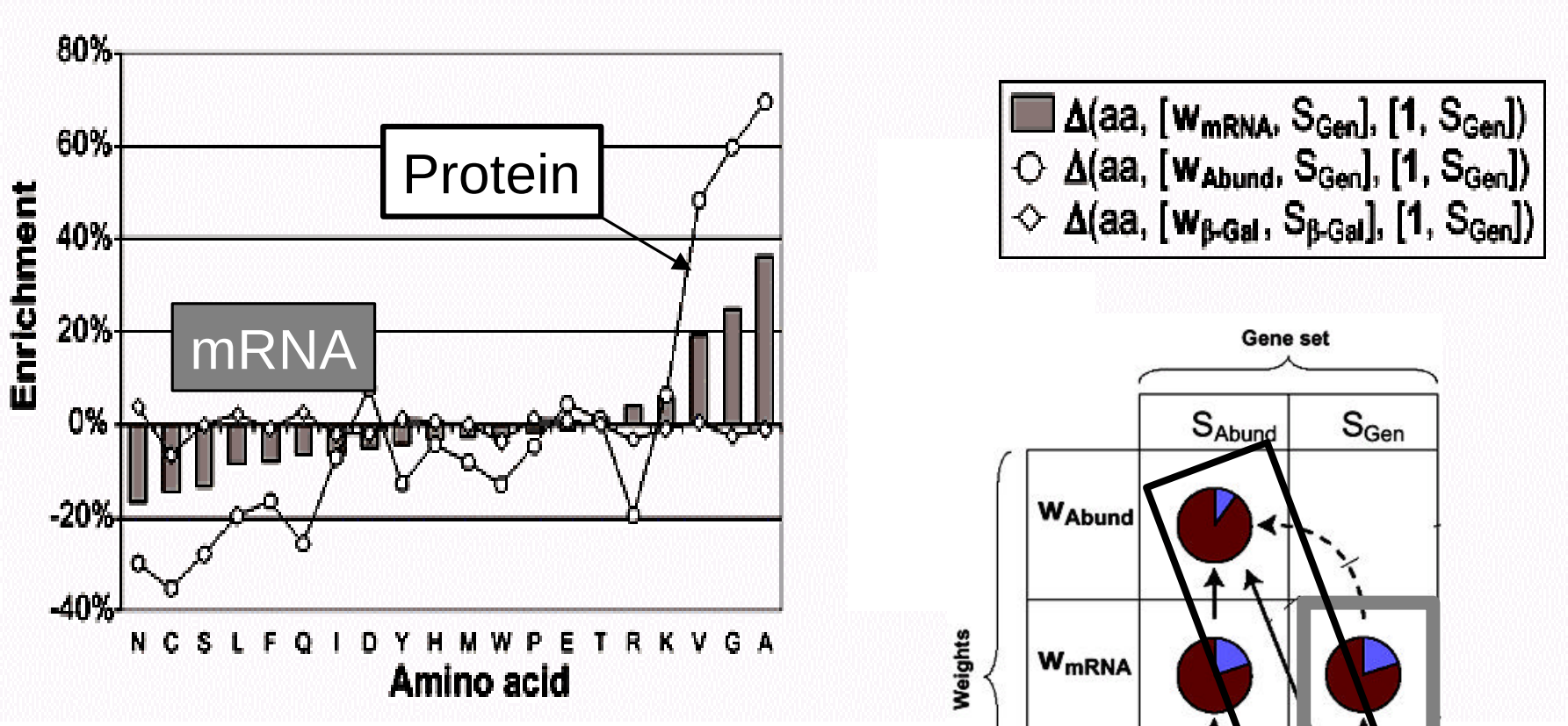
mRNA and protein abundance related, roughly

~150 protein abundance values from merging results of 2D gel expts. of Aebersold & Futcher



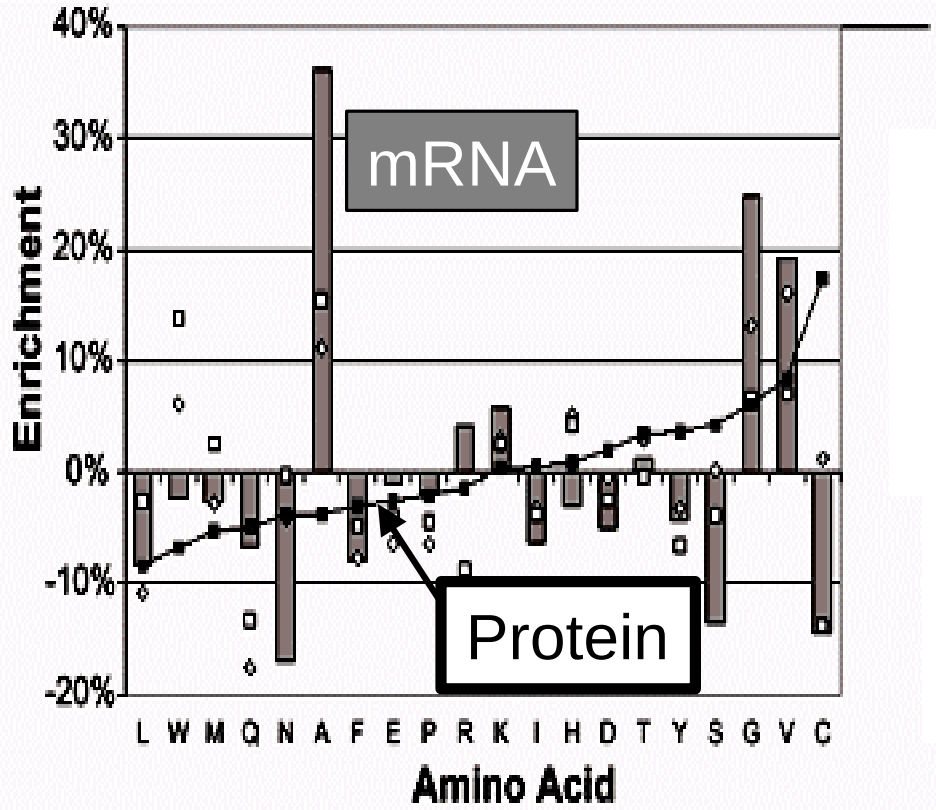
mRNA values for same 150 genes from merging and scaling 6 yeast expressions

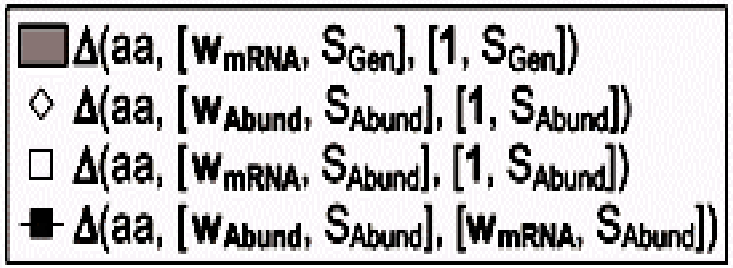
Amino Acid Enrichment

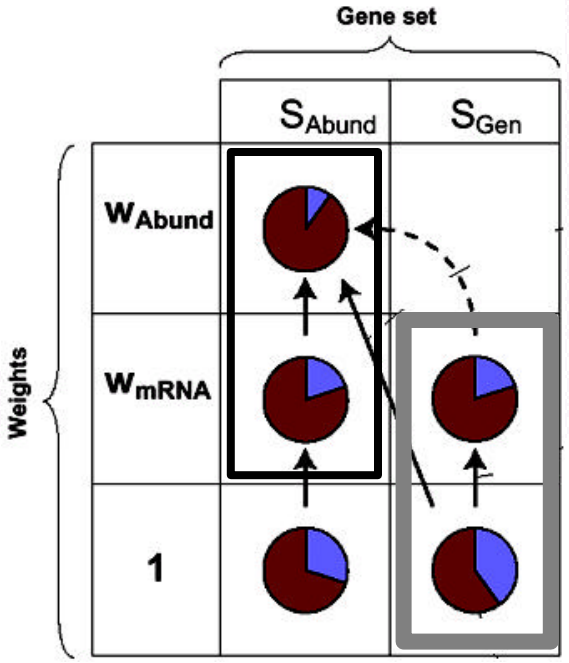


Simple story is translatome is enriched in same way as transcriptome

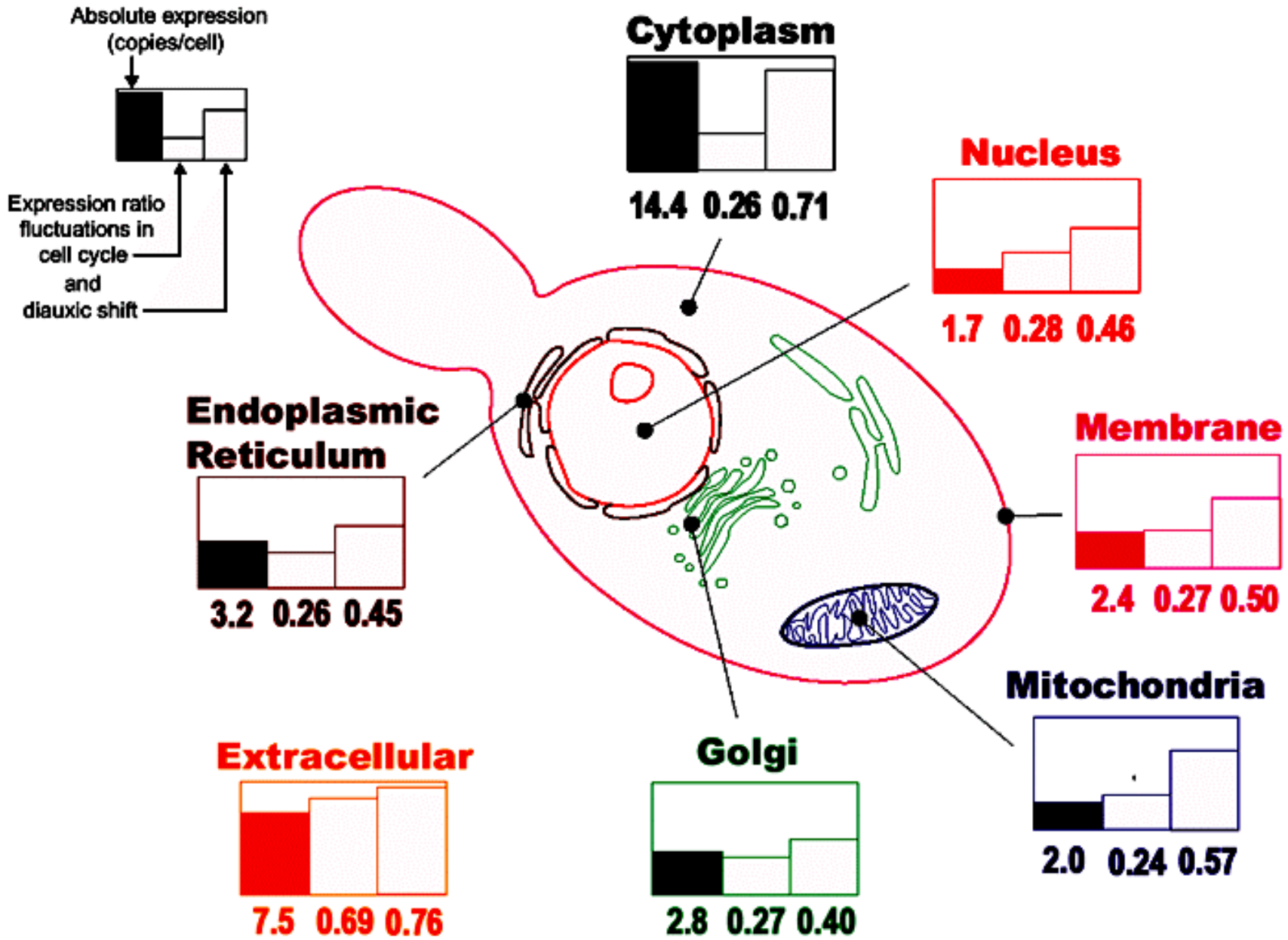
Amino Acid Enrichment – Complexities



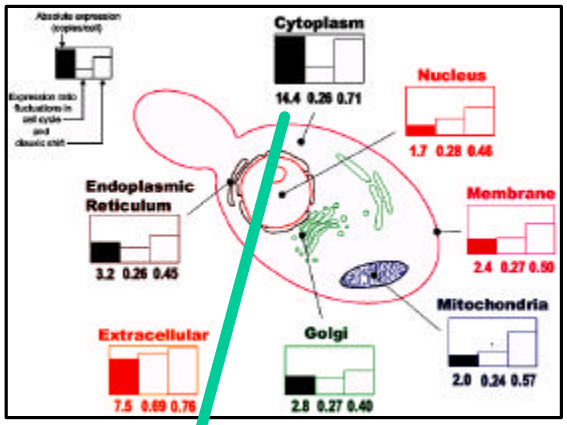


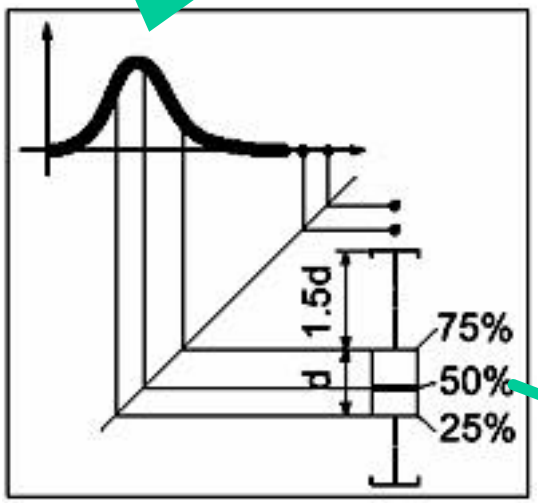


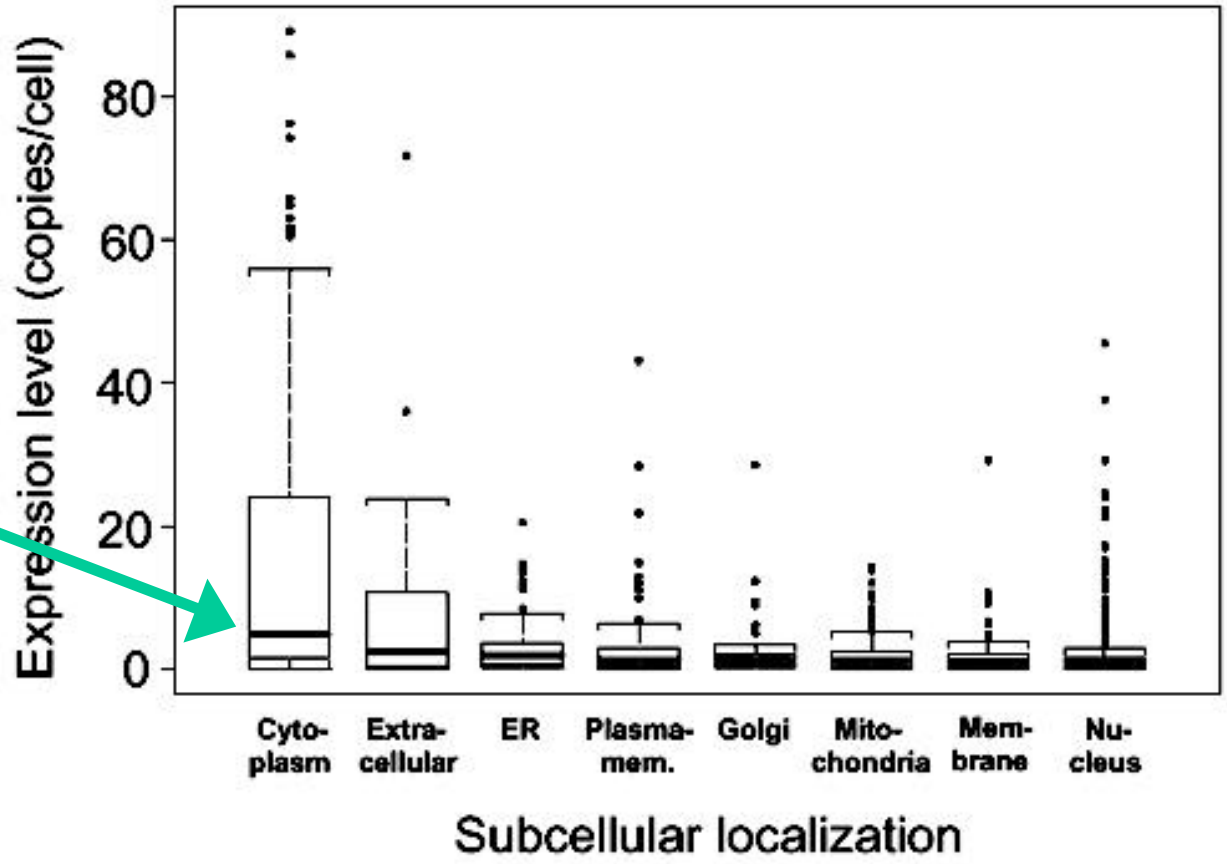
Expression Level is Related to Localization



Distributions of Expression Levels

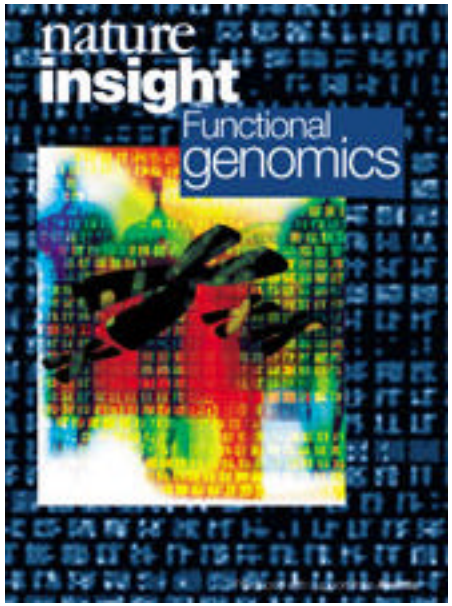






~6000 yeast genes with expression levels

but only ~2000 with localization….



insight review articles

Genomics, gene expression and DNA arrays

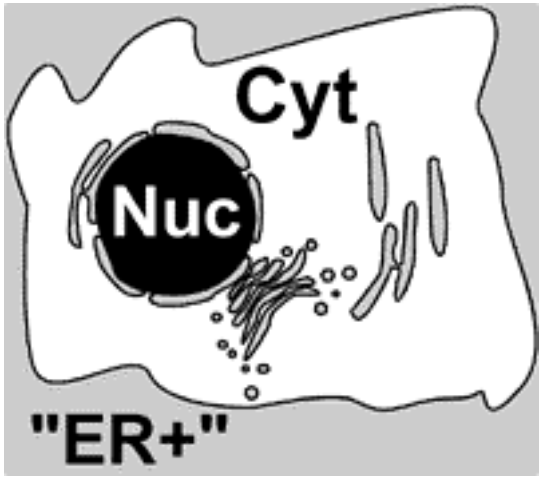
David J. Lockhart & Elizabeth A. Winzeler

Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, USA

Experimental genomics in combination with the growing body of sequence information promise to revolutionize the way cells and cellular processes are studied. Information on genomic sequence can be used experimentally with high-density DNA arrays that allow complex mixtures of RNA and DNA to be interrogated in a parallel and quantitative fashion. DNA arrays can be used for many different purposes, most prominently to measure levels of gene expression (messenger RNA abundance) for tens of thousands of genes simultaneously. Measurements of gene expression and other applications of arrays embody much of what is implied by the term (genomics); they are broad in scope, large in scale, and take advantage of all available sequence information for experimental design and data interpretation in pursuit of biological understanding.

Bayesian System for Localizing Proteins

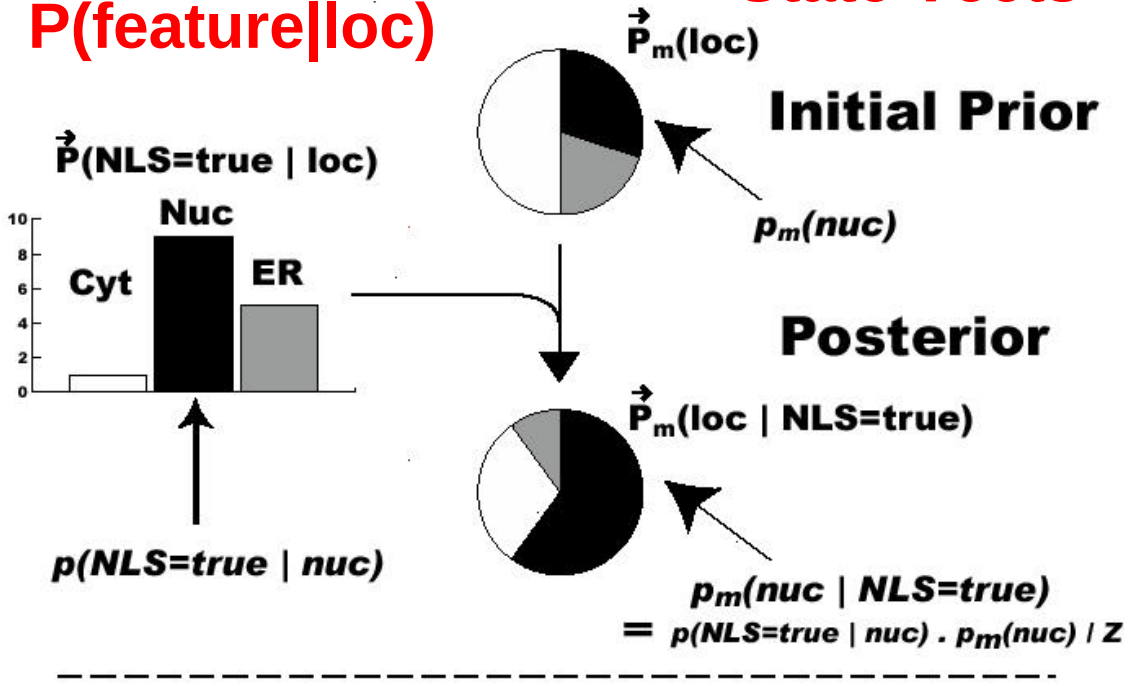
loc=



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.

Feature Vects
P(feature|loc)

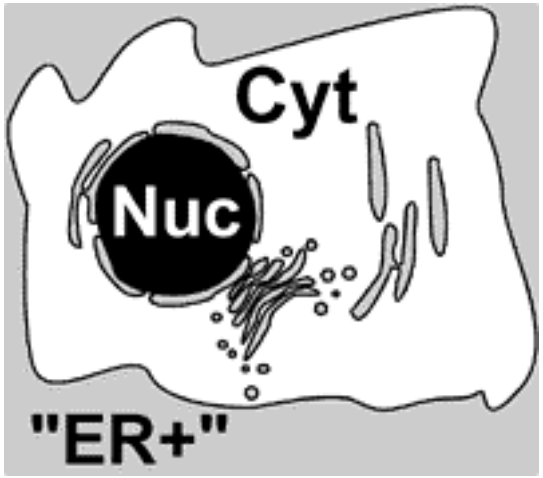
State Vects



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

Bayesian System for Localizing Proteins

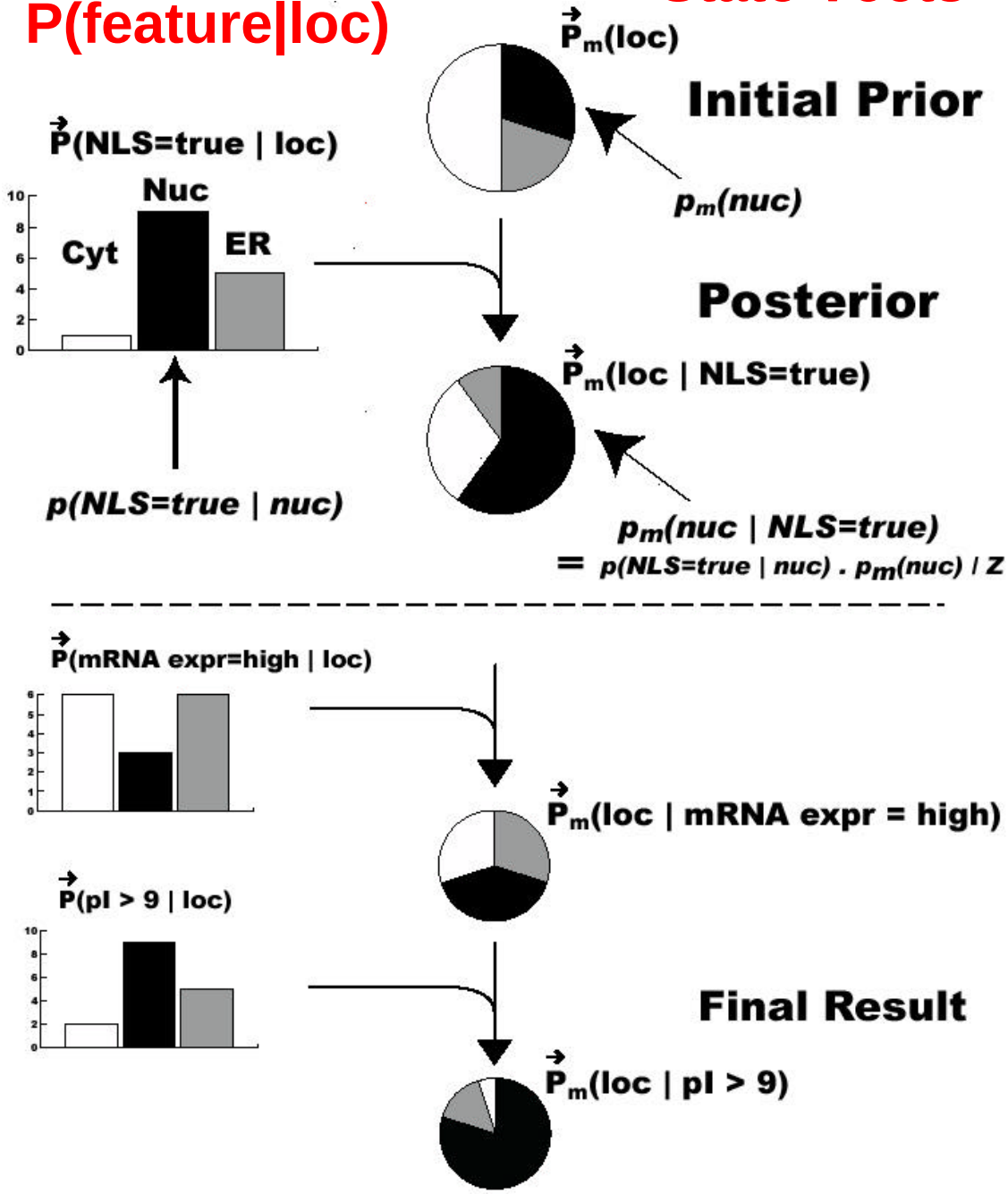
loc=



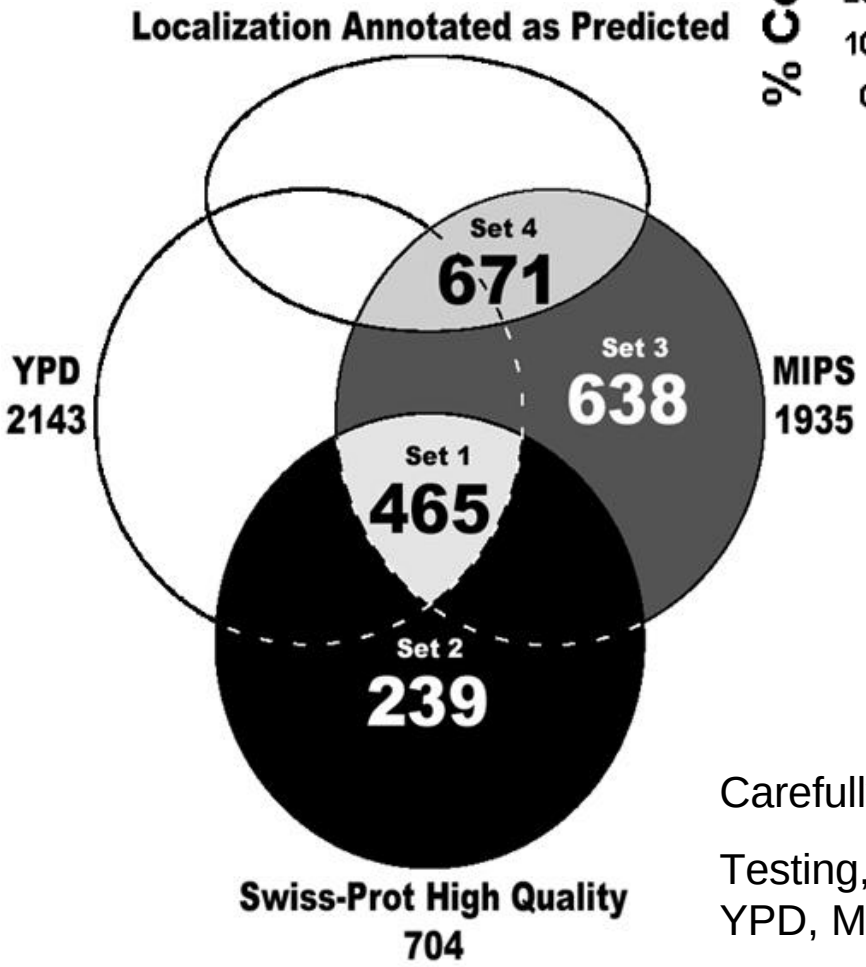
Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.

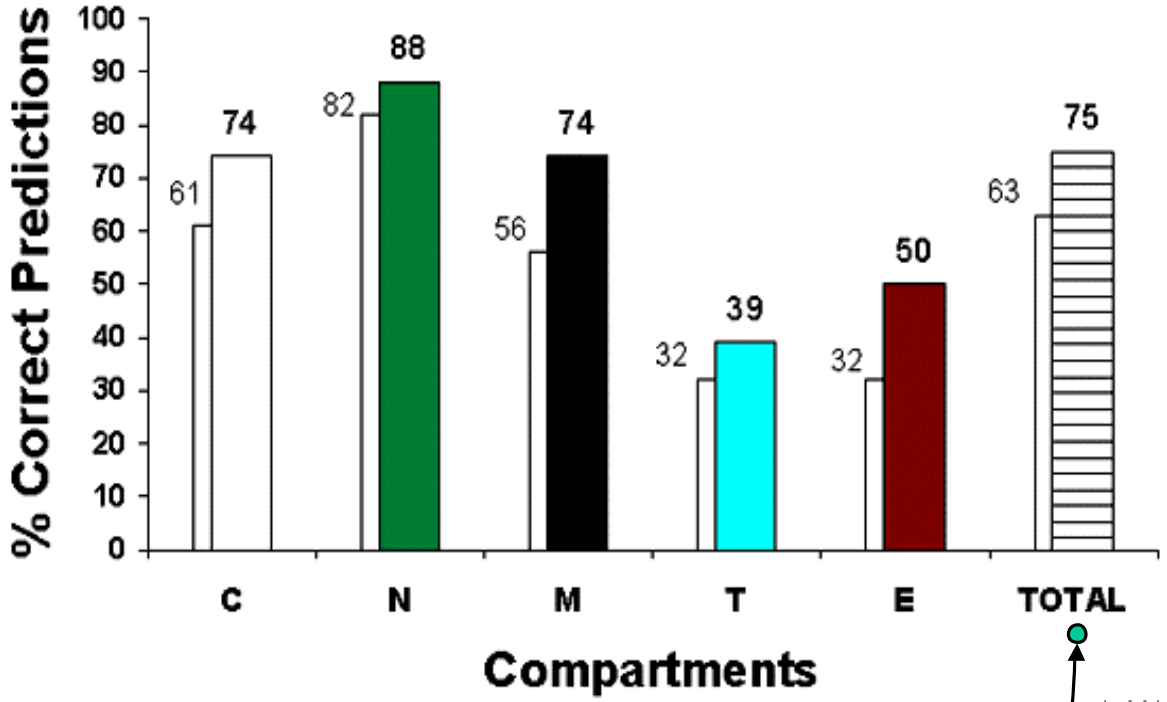
Feature Vects
P(feature|loc)

State Vects



Results on Testing Data





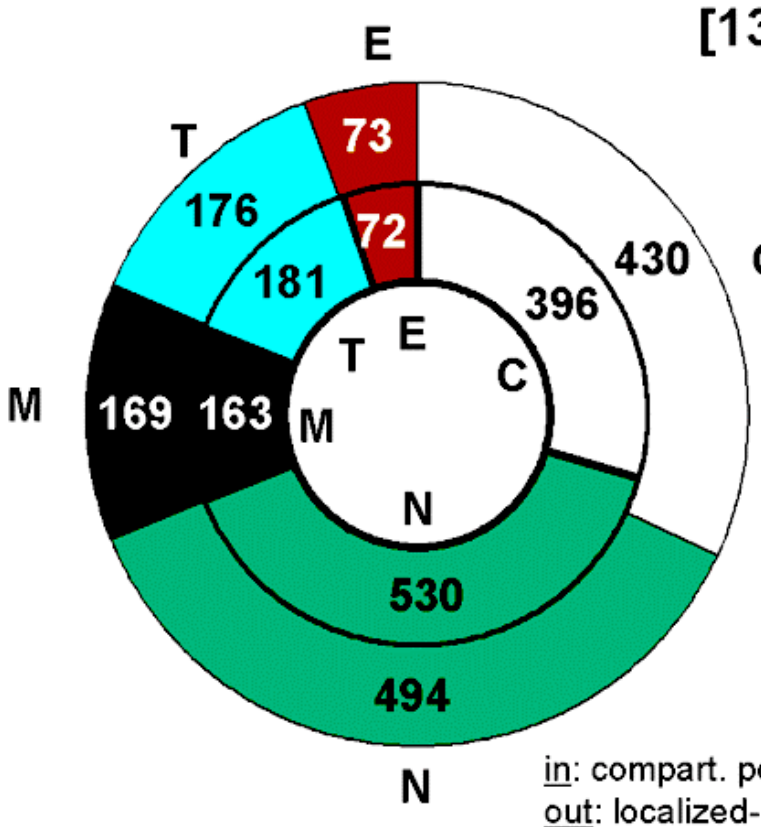
Individual proteins: 75% with cross-validation

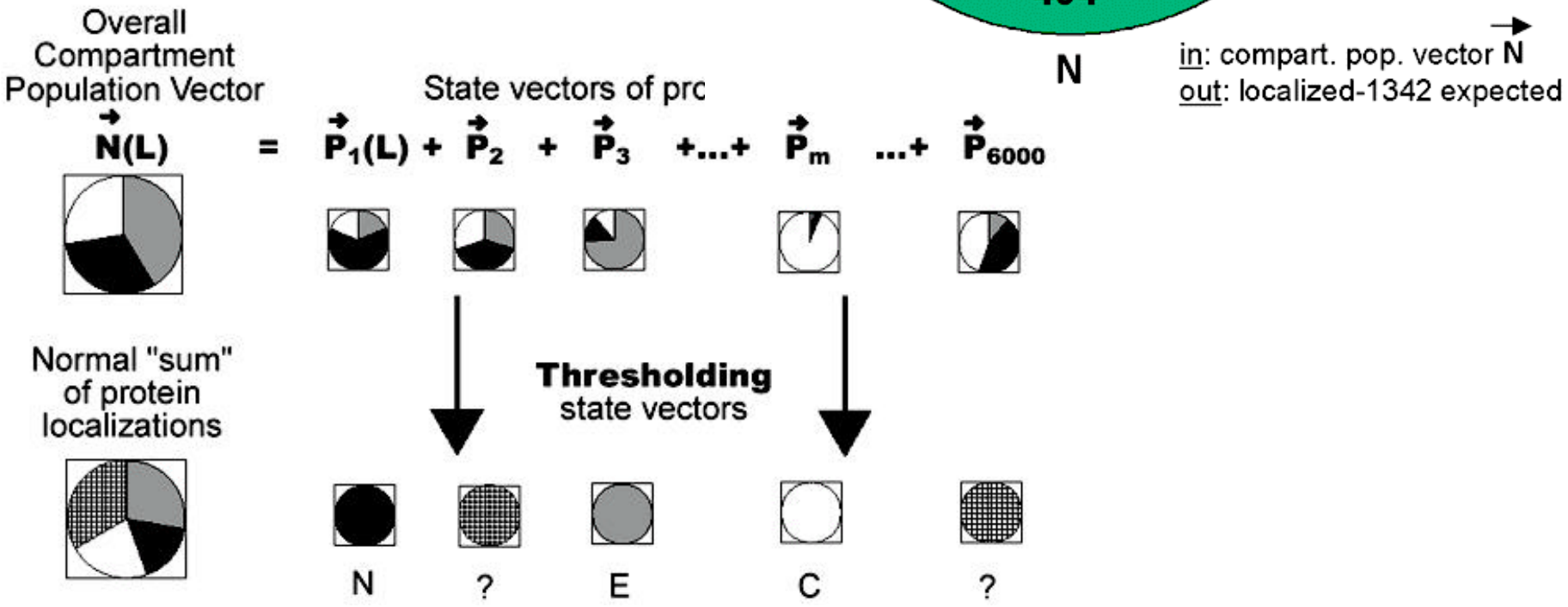
Carefully clean training dataset to **avoid circular logic**

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

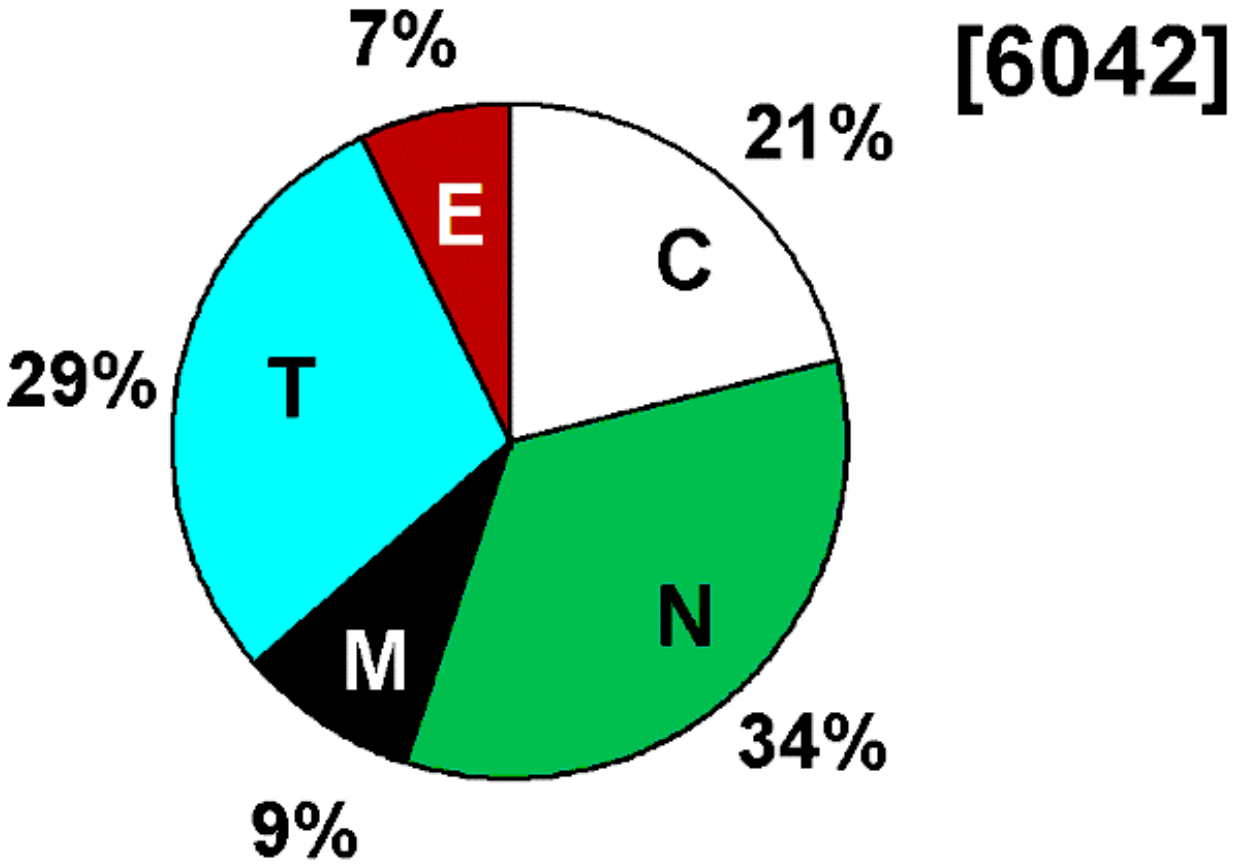
Results on Testing Data #2

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

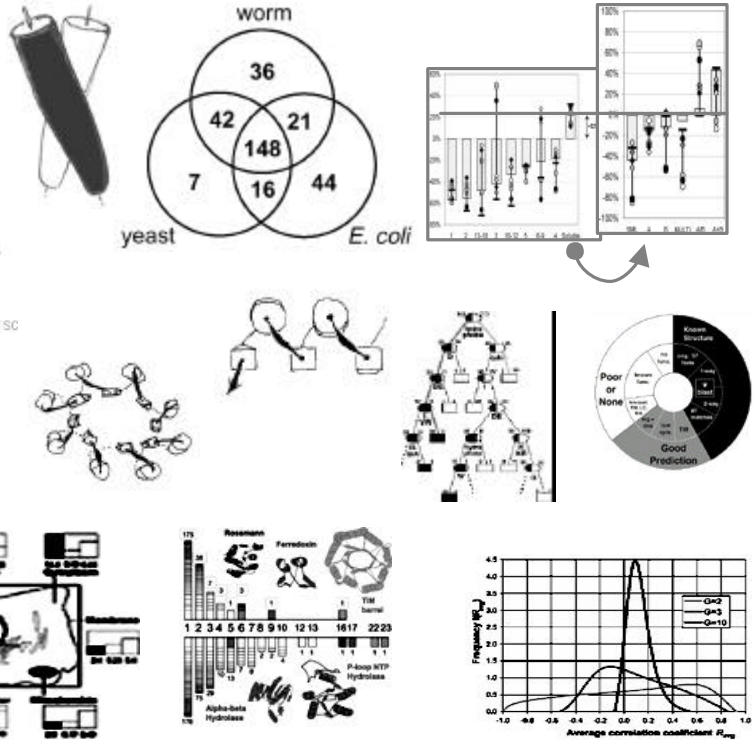




Extrapolation to Compartment Populations of Whole Yeast Genome: ~4000 predicted + ~2000 known



Comparative Genomics: Surveys of a Finite Parts List



1 Using Folds to Interpret Genomes.

Fold Library background. Shared and/or unique parts. Venn Diagram, Fold tree with all-β diff. Ortholog tree. Horizontal Transfer. Commor Parts: Top-10 folds with βαβ. Common Ψ-fold.

- **Tricky Issues →Expt. SG.** Extent of fold assignment (MG, 65%, 98). Predictions. Biases. NESG.org Construct DB. Datamining this w/ decision trees. Selecting weird MG CD targets.
- **Folds & Functions.** Roles/part? How many folds/func? 331 of ~20K combinations. Mostly 1 func/fold, but some versatile scaffolds -- TIM most versatile. Similar for interactions. Func. Divergence vs. Seq. & Struc. Diverg.

4 Using Folds to Interpret Expression Data.

Top-10 parts in other terms. Enriched in transcriptome: VGA, αβ folds, energy, synthesis, Cytoplasmic, TIM fold. Depleted: NS, long, TMs, transport, transcription, Nuclear, Leu-zip fold. Bayesian localizer.

5 Tricky Issues: Relating Expression to Function.

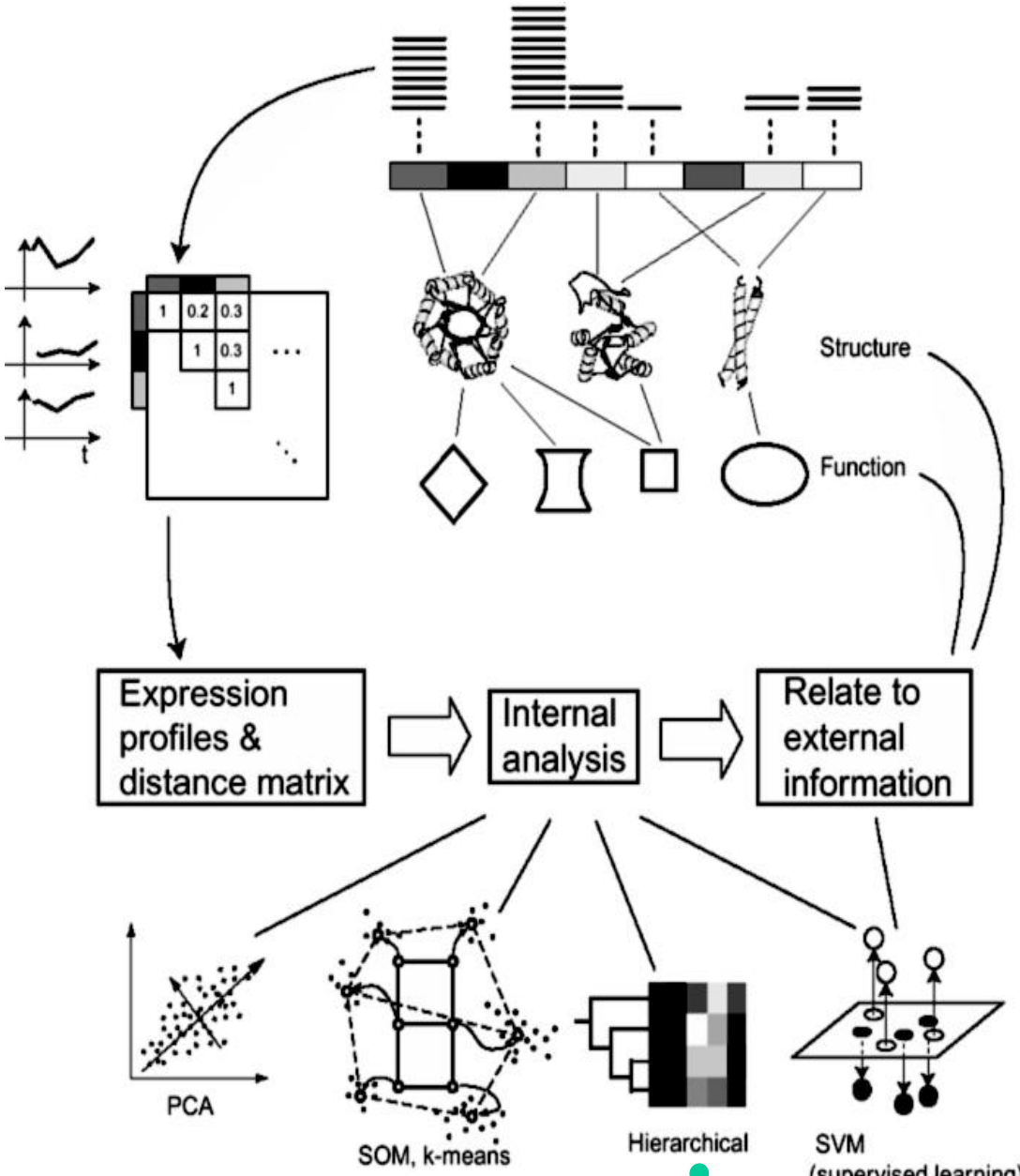
Expression relates to structure & localization but to function, globally? Weak relation to protein-protein interactions.

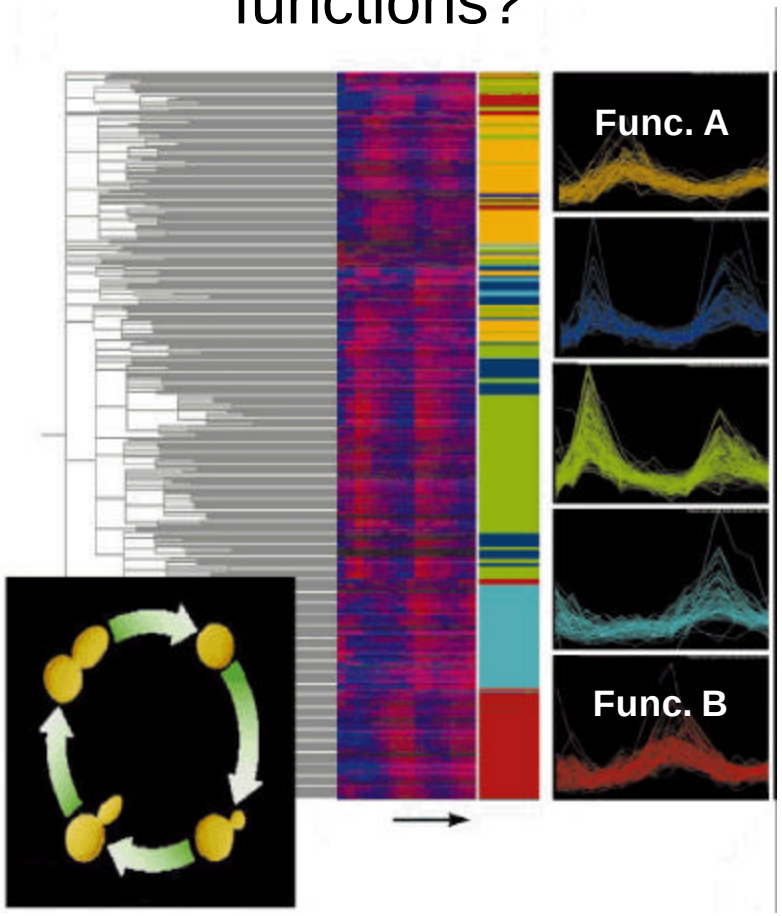
H Hegyi, J Lin, B Stenger, N Echols, P Bertone, J Qian, L Regan, S Balasubramanian, V Alexandrov, G Montelione, A Edwards, C Wilson, Y Kluger, C Arrowsmith, A Drawid, R Jansen, D Greenbaum, S Teichmann, P Harrison

bioinfo.mbb.yale.edu

Do Expression Clusters Relate to Protein Function?

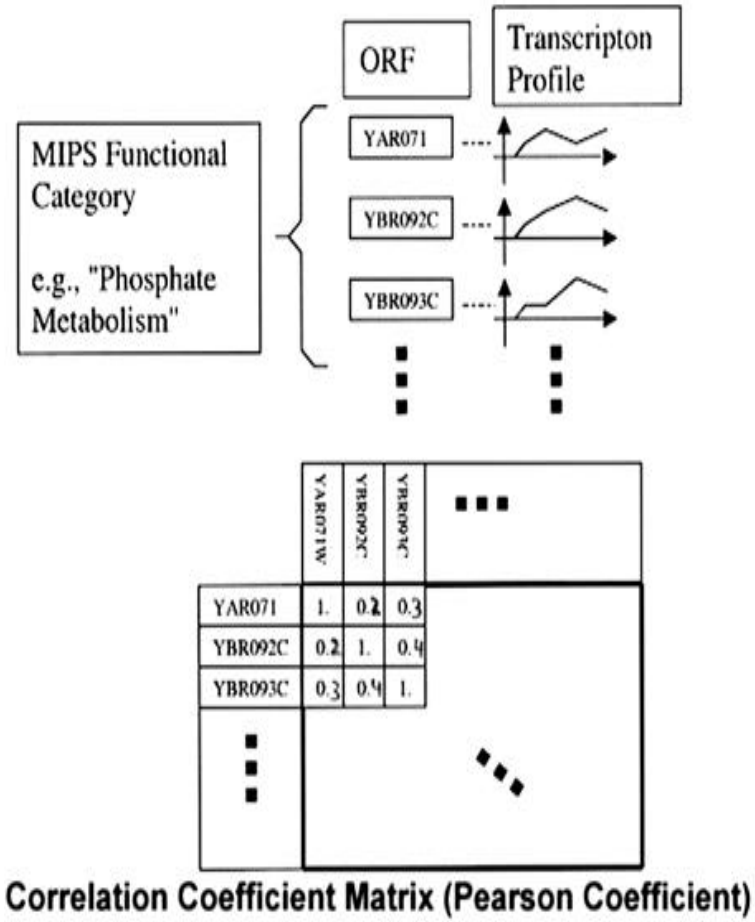
Can they predict functions?





- Clustering of expression profiles
- Grouping functionally related genes together (?)
- **Botstein (Eisen),** Lander, Haussler, and Church groups, Eisenberg

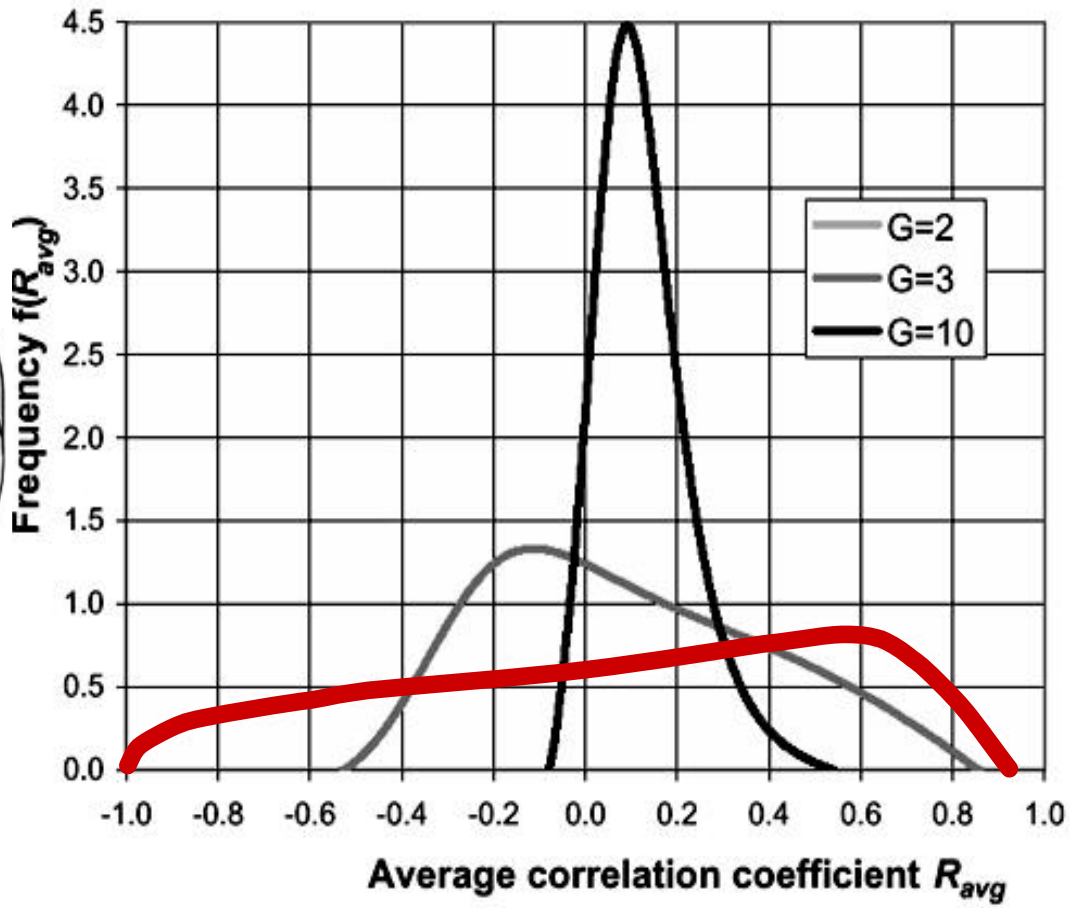
Distributions of Gene Expression Correlations, for All Possible Gene Groupings



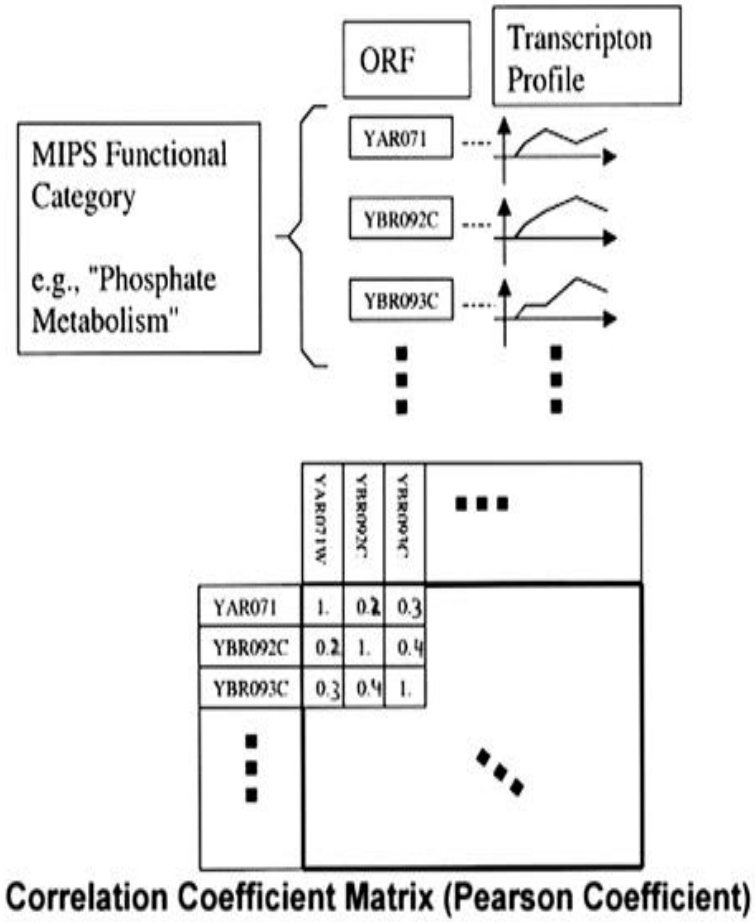
Average Correlation Coefficient for Group of Genes

Sample for Diauxic shift Expt. (Brown),

Ex. $R_{avg,G=3}$ =
[R(gene-1,gene-3) + R(gene-1,gene-4)
+
R(gene-5,gene-7)] / 3



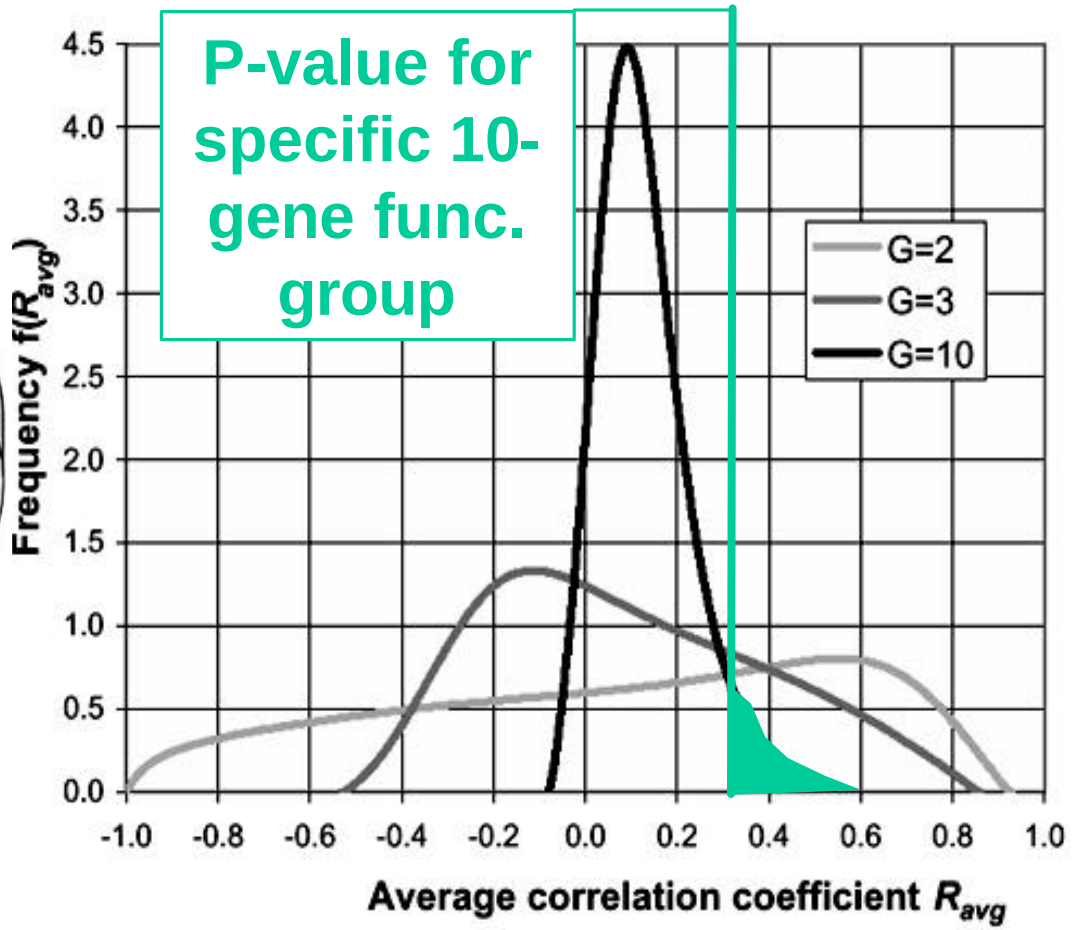
Distributions of Gene Expression Correlations, for All Possible Gene Groupings 2



Average Correlation Coefficient for Group of Genes

Sample for Diauxic shift Expt. (Brown),

Ex. $R_{avg,G=3}$ =
[R(gene-1,gene-3) + R(gene-1,gene-4)
+
R(gene-5,gene-7)] / 3



MIPS category	Experiment: Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo-rulation
Cell growth, division & DNA syn.	>4	>4	>4	>4
Protein synthesis	>4	>4	>4	>4
Transcription	>4	>4	>4	1.6
Cellular organization	>4	>4	0.3	0.3
Energy	>4	>4	0.1	0.9
Cell rescue, defense, death	>4	>4	0	0
Intracellular transport	>4	>4	0	0
Ionic homeostasis	>4	>4	0	0.8
Metabolism	>4	>4	0	0
Transport facilitation	>4	>4	0	0
Signal transduction	2.5	1.6	0.1	0.6
Unclassifield	2.3	>4	0	0
Cellular biogenesis	2.0	>4	0.4	0.2
Protein destination	0.3	>4	0.2	0.6
Retrotransposon & plasmid	0	2.8	1.9	1.0

MIPS category	Experiment: Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo-rulation
Respiration	>4	>4	>4	3.4
TCA pathway	>4	>4	>4	0.6
Glycogen, trehalose metabolism	>4	>4	1.2	0.7
Glycolysis	>4	>4	0.9	2.1
Gluconeogenesis	3.7	>4	0.1	1.7
Glyoxylate cycle	1.6	0.7	3.0	2.3
Pentose-phosphate pathway	1.5	0.8	0	0.6
Fermentation	1.3	>4	0	2.2
Other energy generation activities	0.7	0.1	0.1	0.2
Beta-oxidation of fatty acids	0.5	0.4	0.4	0.2

Correlation:

Always Significant

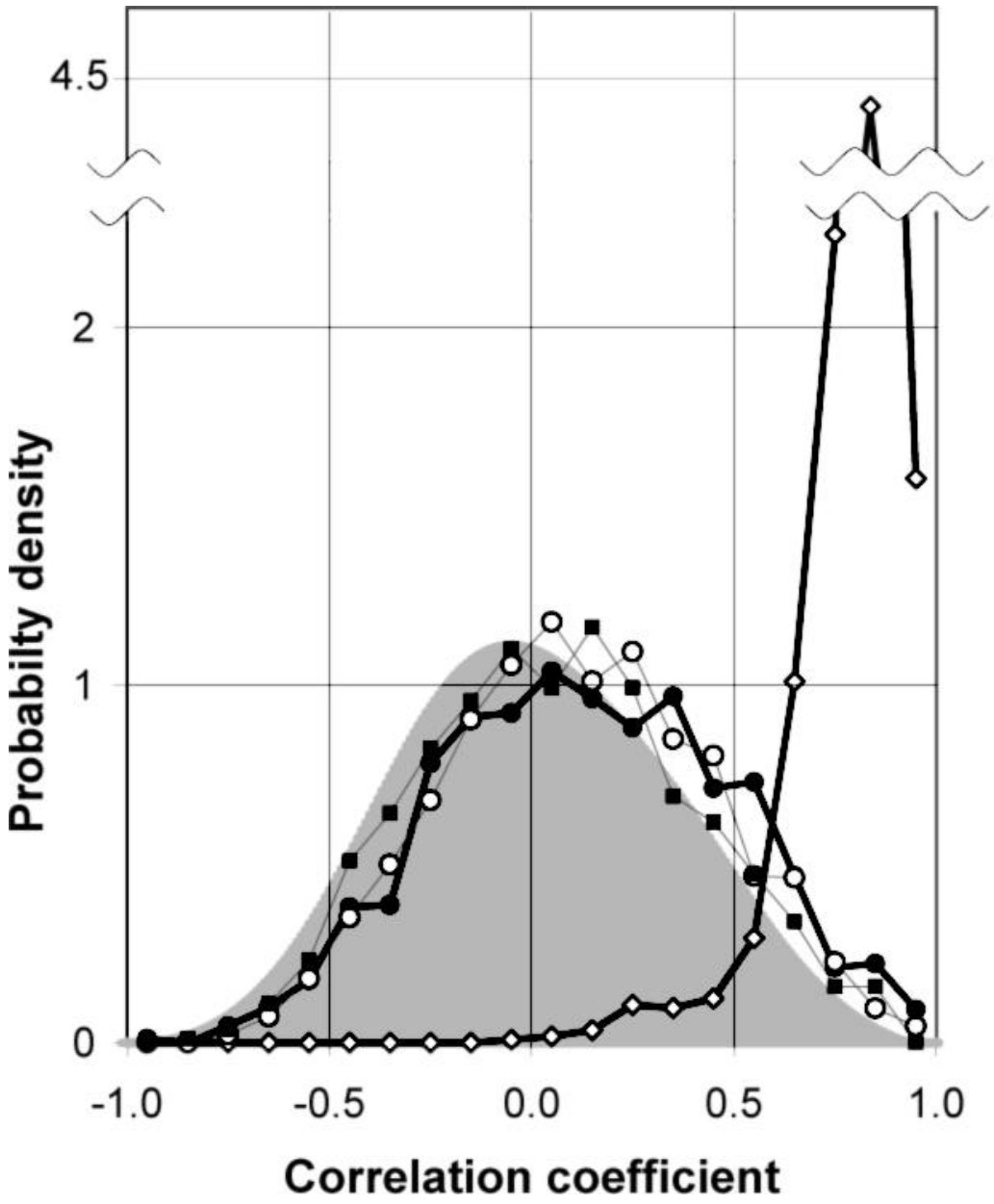
Sometimes Significant (depends on expt.)

Never Significant

	Fraction of significant groups: CDC28	CDC15	Diauxic Shift	Sporulation	Total # groups
MIPS 1	63%	81%	19%	13%	16
MIPS 2	50%	63%	17%	13%	102
MIPS 3	23%	33%	5%	4%	73
"Energy" (2nd level)	40%	60%	20%	0%	10
SOM	93%	-	-	-	30
Hierarch. Clustering	80%				25

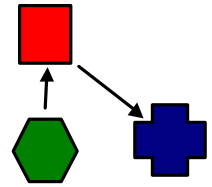
Based on Distributions, Correlation of Established Functional Categories, Computer Clusterings

Protein-Protein Interactions & Expression



between selected expression timecourses in CDC28 expt. (Davis)

Use same formalism to assess how closely related expression timecourses to sets of known p-p interactions

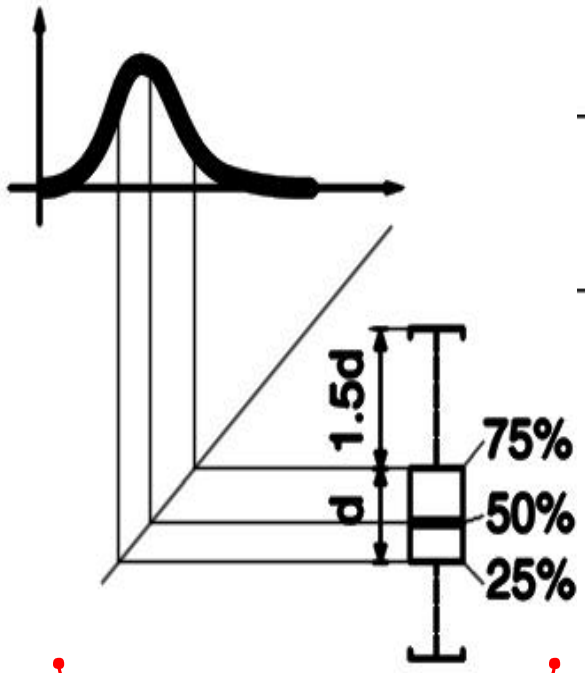


Sets of interactions

Random (cell cycle CDC28) (all pairs) (control)

physical
genetic
Y2H
(from MIPS)
(Uetz et al.)

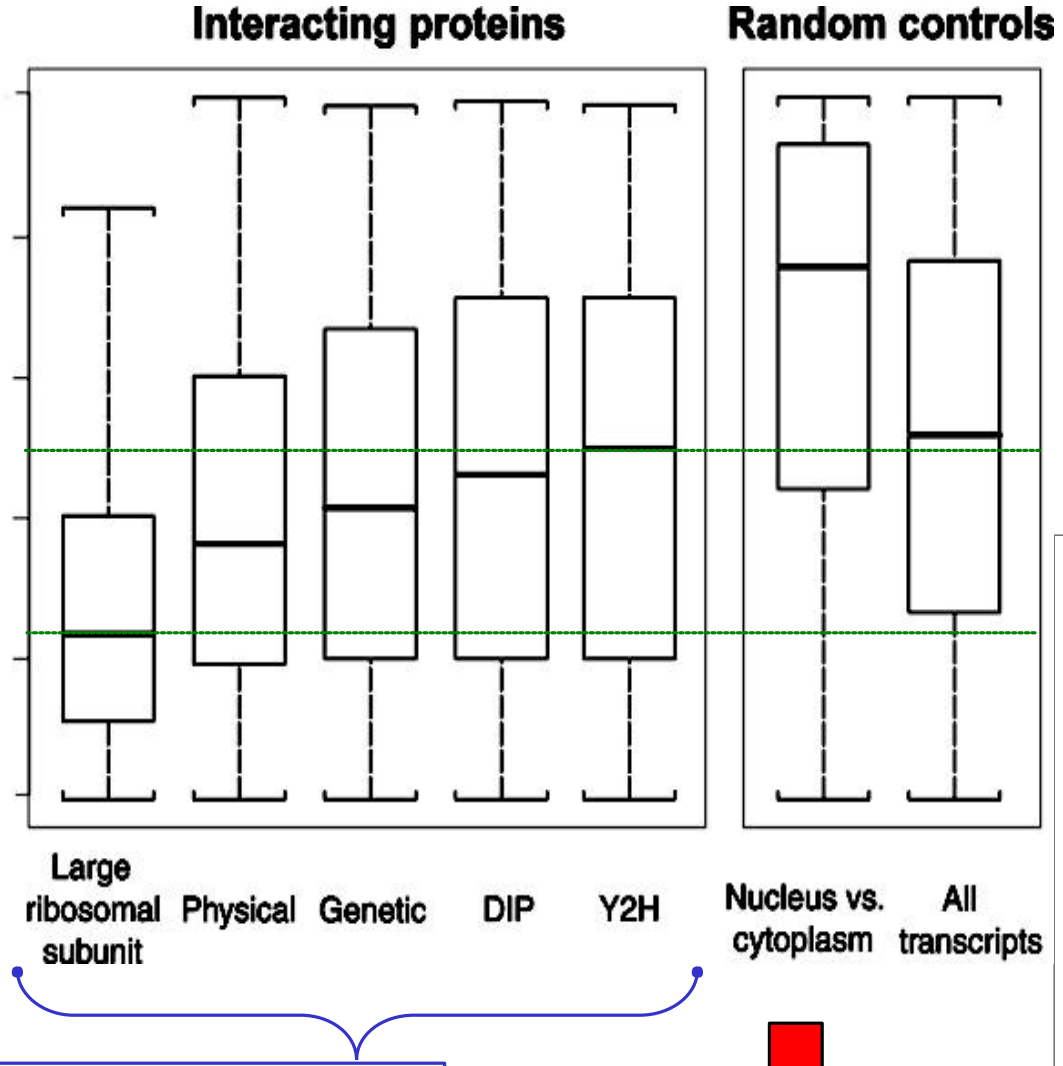
Large ribosomal subunit
(strong interaction, clearly diff.)

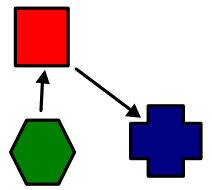


Distribution of Normalized Expression Levels

for

Sets of Interacting Proteins





Relation of P-P Interactions to Abs. Expression Level

Can we define FUNCTION well enough to relate to expression?

Fold, Localization, Interactions & Regulation are attributes of proteins that are much more clearly defined

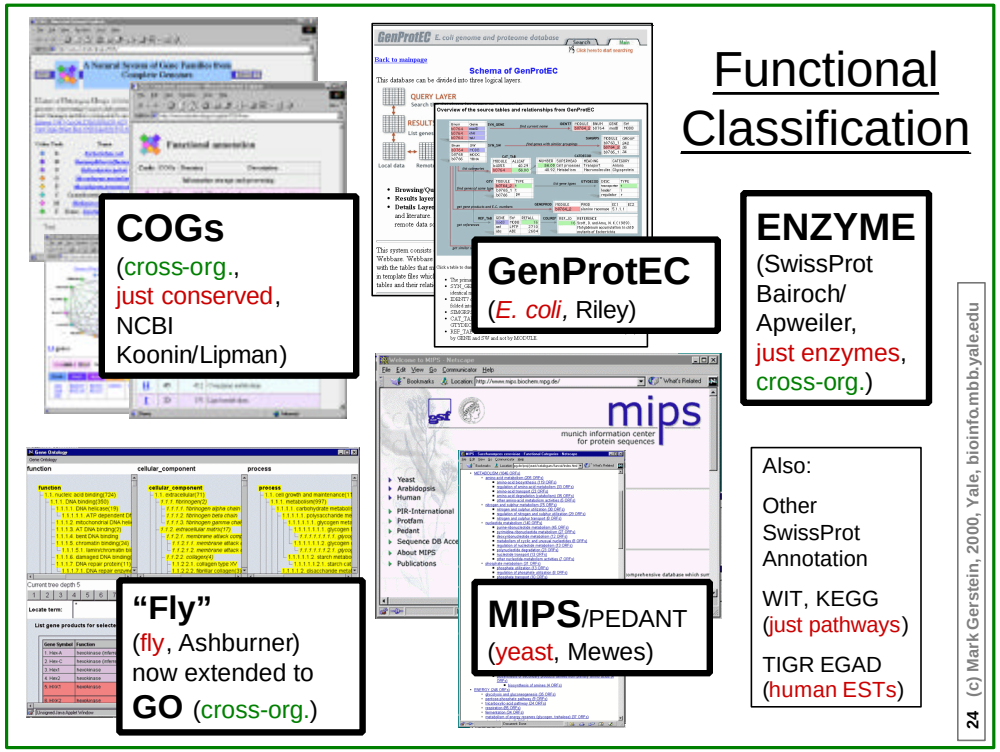
Problems defining function:

Multi-functionality: 2 functions/protein (also 2 proteins/function)

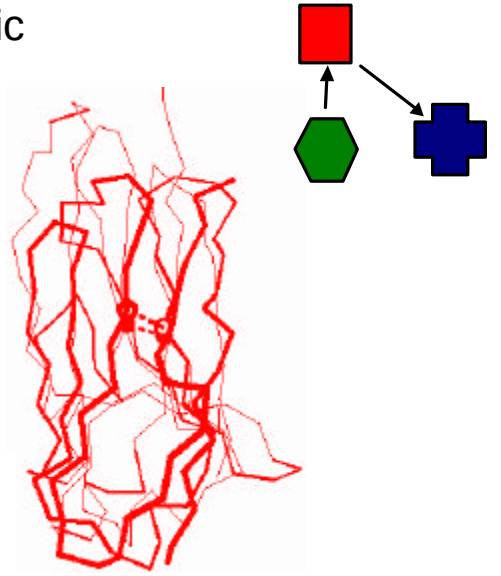
Conflating of Roles: molecular action, cellular role, phenotypic manifestation.

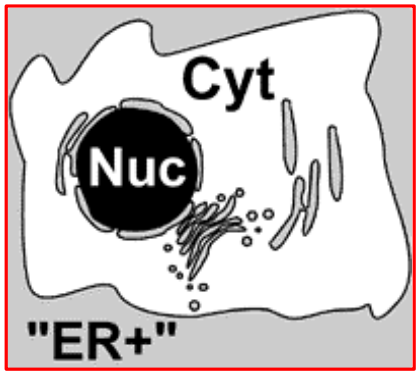
Non-systematic Terminology:

'suppressor-of-white-apricot' & 'darkener-of-apricot'



vs.

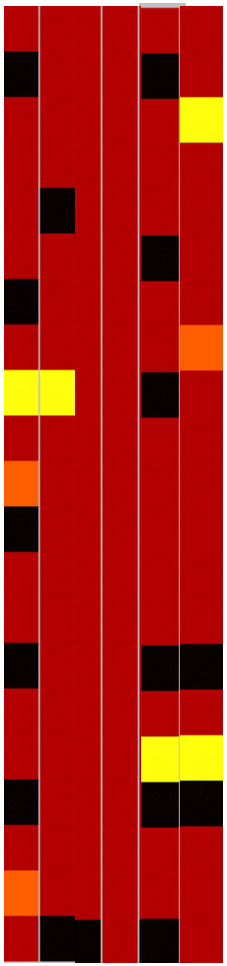




Whole Genome Phenotype Profiles

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**

YPD + 8mM caffeine	**Caff**
Cycloheximide hypersensitivity: YPD + 0.08 ?g/ml cycloheximide at 30°C	**Cyc^S**
White/ red color on YPD	**W/R**
YPGlycerol	**YPG**
Calcofluor hypersensitivity: YPD + 12 ?g/ml calcofluor at 30°C	**Calc^S**
YPD + 46 ?g/ml hygromycin at 30°C	**Hyg**
YPD + 0.003% SDS	**SDS**
Benomyl hypersensitivity: YPD + 10 ?g/ml benomyl	**Ben^S**
YPD + 5-bromo-4-chloro-3-indolyl phosphate 37°C	**BCIP**
YPD + 0.001% methylene blue at 30°C	**MB**
Benomyl resistance: YPD + 20 ?g/ml benomyl	**Ben^R**
YPD at 37°C	**YPD^{37}**
YPD + 2 mM EGTA	**EGTA**
YPD + 0.008% MMS	**MMS**
YPD + 75 mM hydroxyurea	**HU**
YPD at 11°C (COLD)	**YPD^{11}**
Calcofluor resistance: YPD + 66.7 ?g/ml calcofluor at 30°C	**Calc^R**
Cycloheximide resistance: YPD + 0.3 ?g/ml cycloheximide	**Cyc^R**
Hyperhaploid invasive growth mutants	**HHIG**
YPD + 0.9 M NaCl	**NaCl**



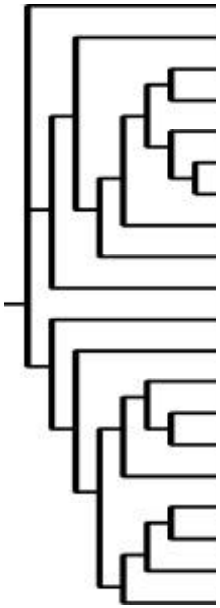
YBR102c YAL009c YBR01w YCL029c YMR009c YER021w

Affected by Another Condition

WT

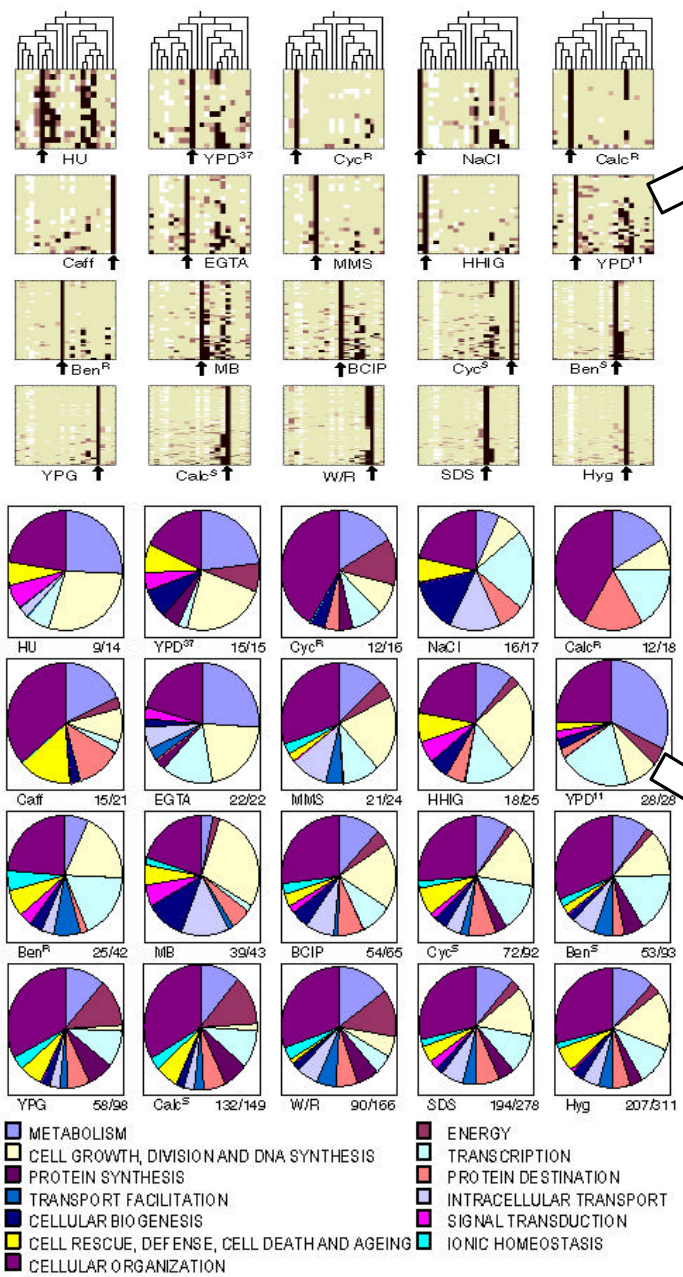
Affected by Cold

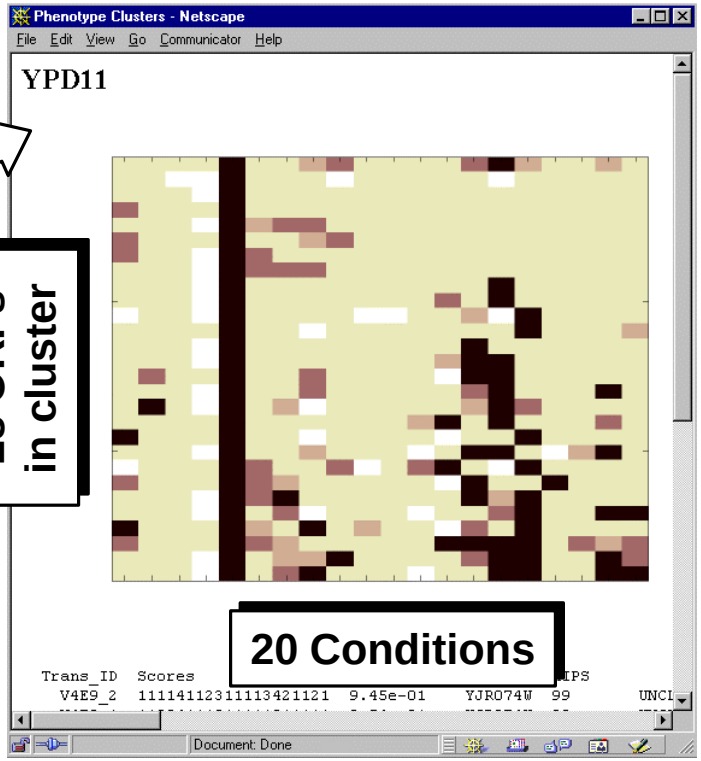
M Snyder

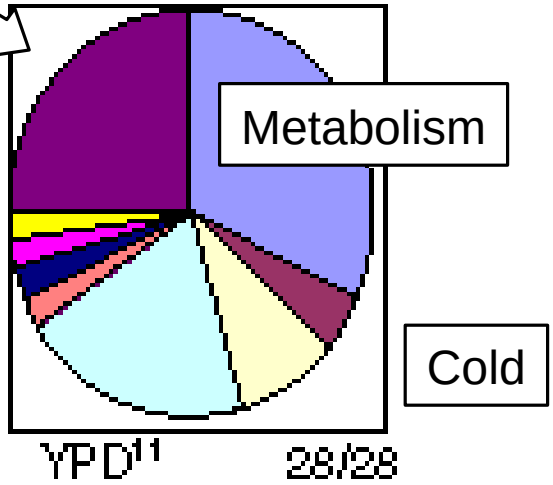


Clustering Conditions

Phenotype ORF Clusters from Transposon Expt.







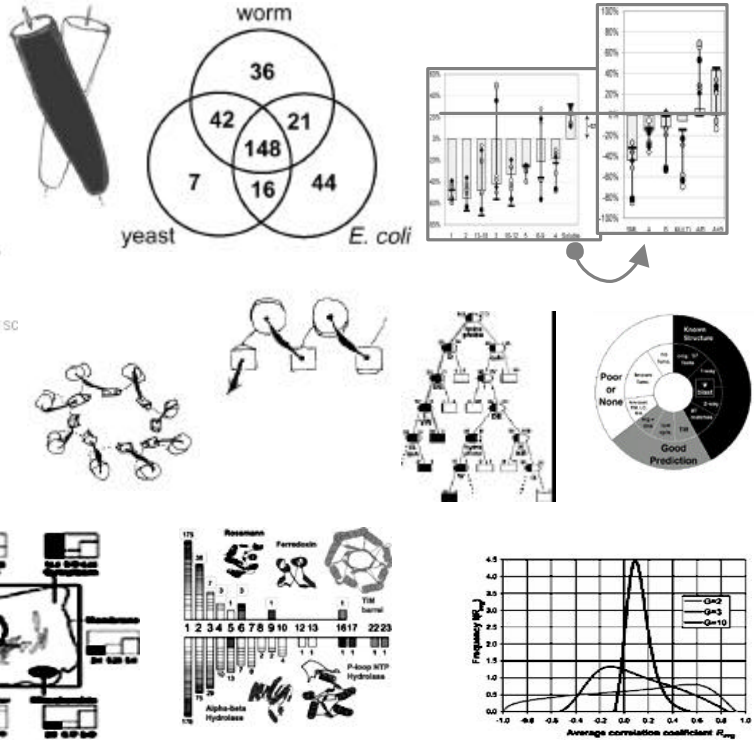
Transposon insertions into (almost) each yeast gene to see how yeast is affected in 20 conditions. Generates a phenotype pattern vector, which can be treated **similarly to expression data**

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

M Snyder,
A Kumar,
et al….

Comparative Genomics: Surveys of a Finite Parts List



1 Using Folds to Interpret Genomes.

Fold Library background. Shared and/or unique parts. Venn Diagram, Fold tree with all-β diff. Ortholog tree. Horizontal Transfer. Commor Parts: Top-10 folds with βαβ. Common Ψ-fold.

- **Tricky Issues →Expt. SG.** Extent of fold assignment (MG, 65%, 98). Predictions. Biases. NESG.org Construct DB. Datamining this w/ decision trees. Selecting weird MG CD targets.
- **Folds & Functions.** Roles/part? How many folds/func? 331 of ~20K combinations. Mostly 1 func/fold, but some versatile scaffolds -- TIM most versatile. Similar for interactions. Func. Divergence vs. Seq. & Struc. Diverg.

4 Using Folds to Interpret Expression Data.

Top-10 parts in other terms. Enriched in transcriptome: VGA, αβ folds, energy, synthesis, Cytoplasmic, TIM fold. Depleted: NS, long, TMs, transport, transcription, Nuclear, Leu-zip fold. Bayesian localizer.

5 Tricky Issues: Relating Expression to Function.

Expression relates to structure & localization but to function, globally? Weak relation to protein-protein interactions.

H Hegyi, J Lin, B Stenger, N Echols, P Bertone, J Qian, L Regan, S Balasubramanian, V Alexandrov, G Montelione, A Edwards, C Wilson, Y Kluger, C Arrowsmith, A Drawid, R Jansen, D Greenbaum, S Teichmann, P Harrison

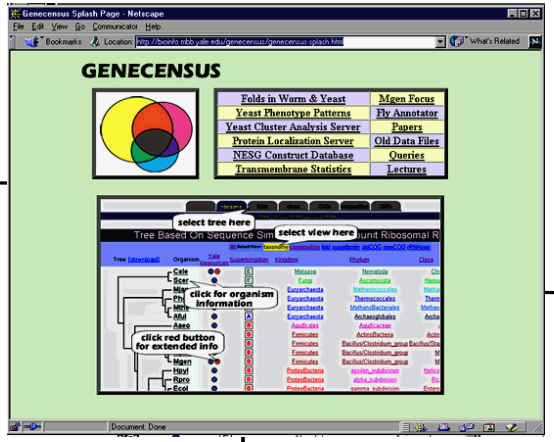
bioinfo.mbb.yale.edu

GeneCensus

bioinfo.mbb.yale.edu

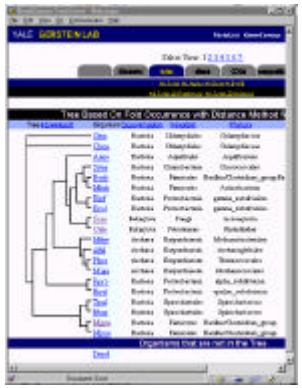
Alignment Database

Alignment Server

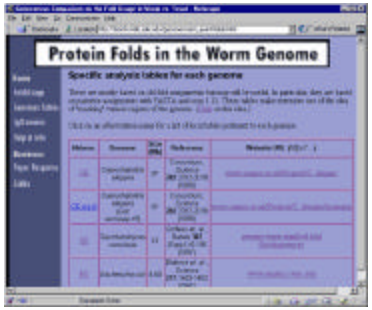


Ranks

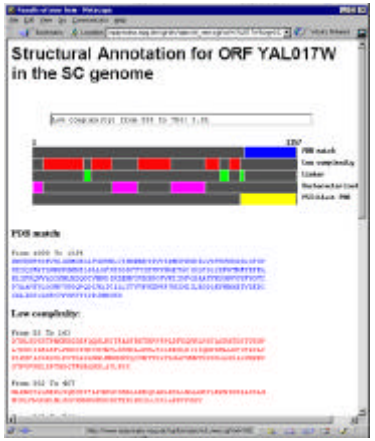
Trees



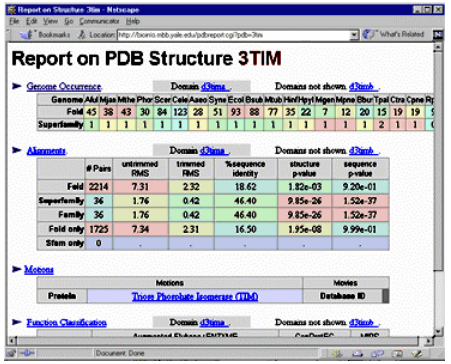
Detailed Tables



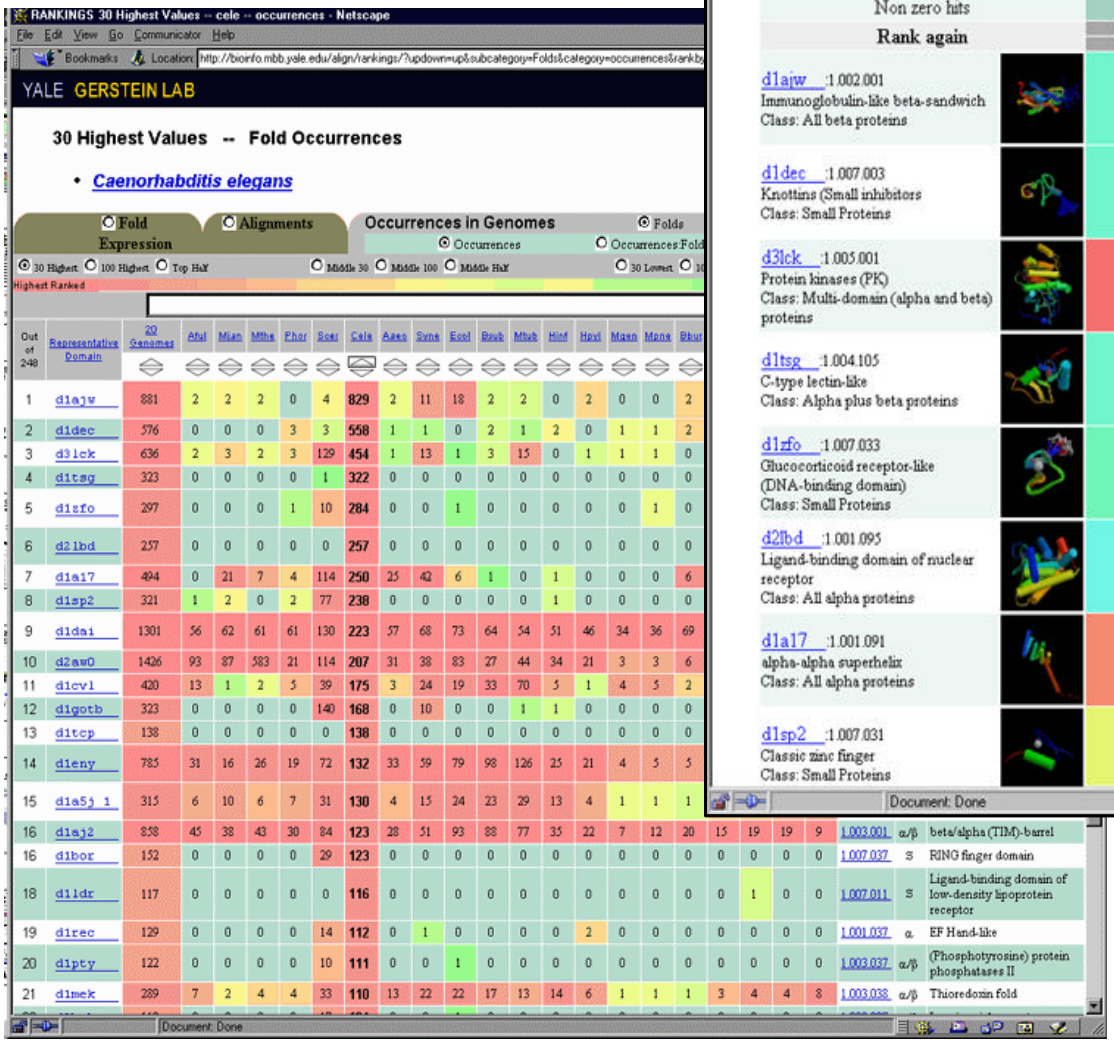
ORF Query

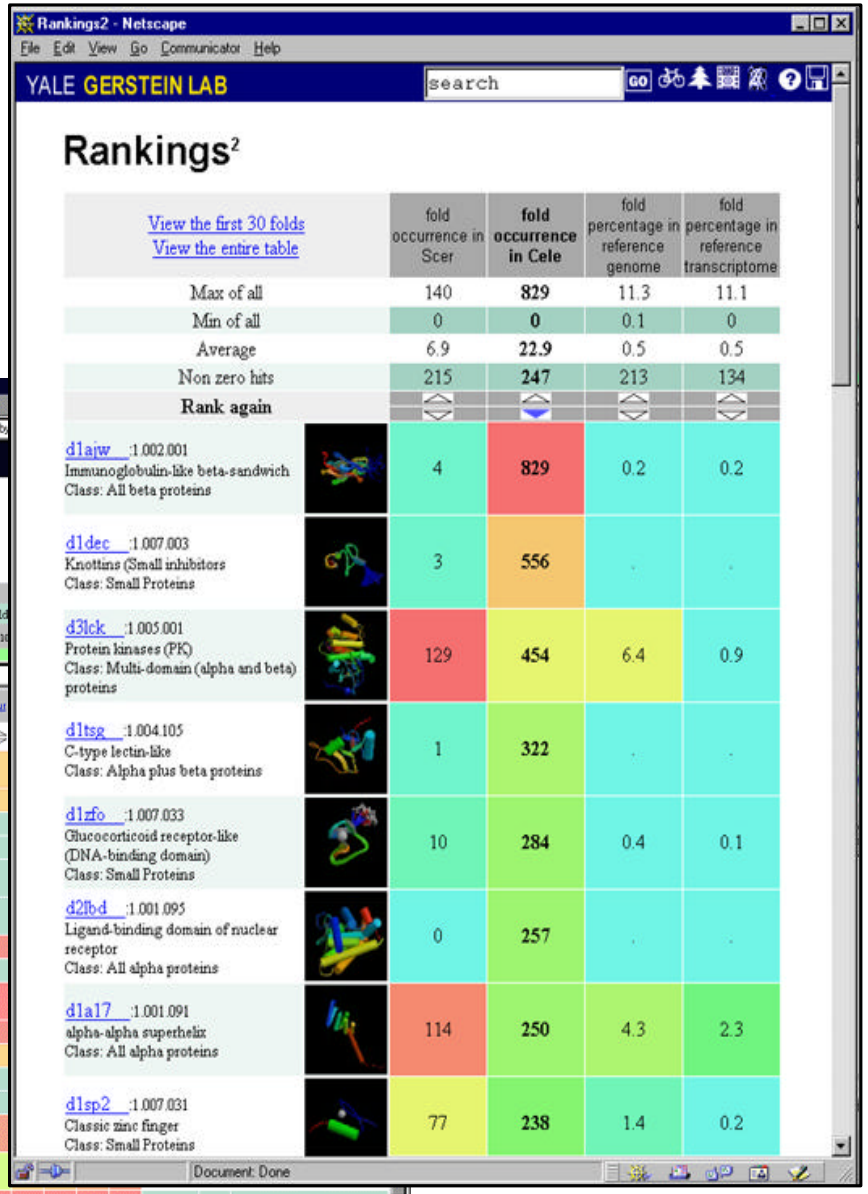


PDB Query



PartsList Ranking Viewers

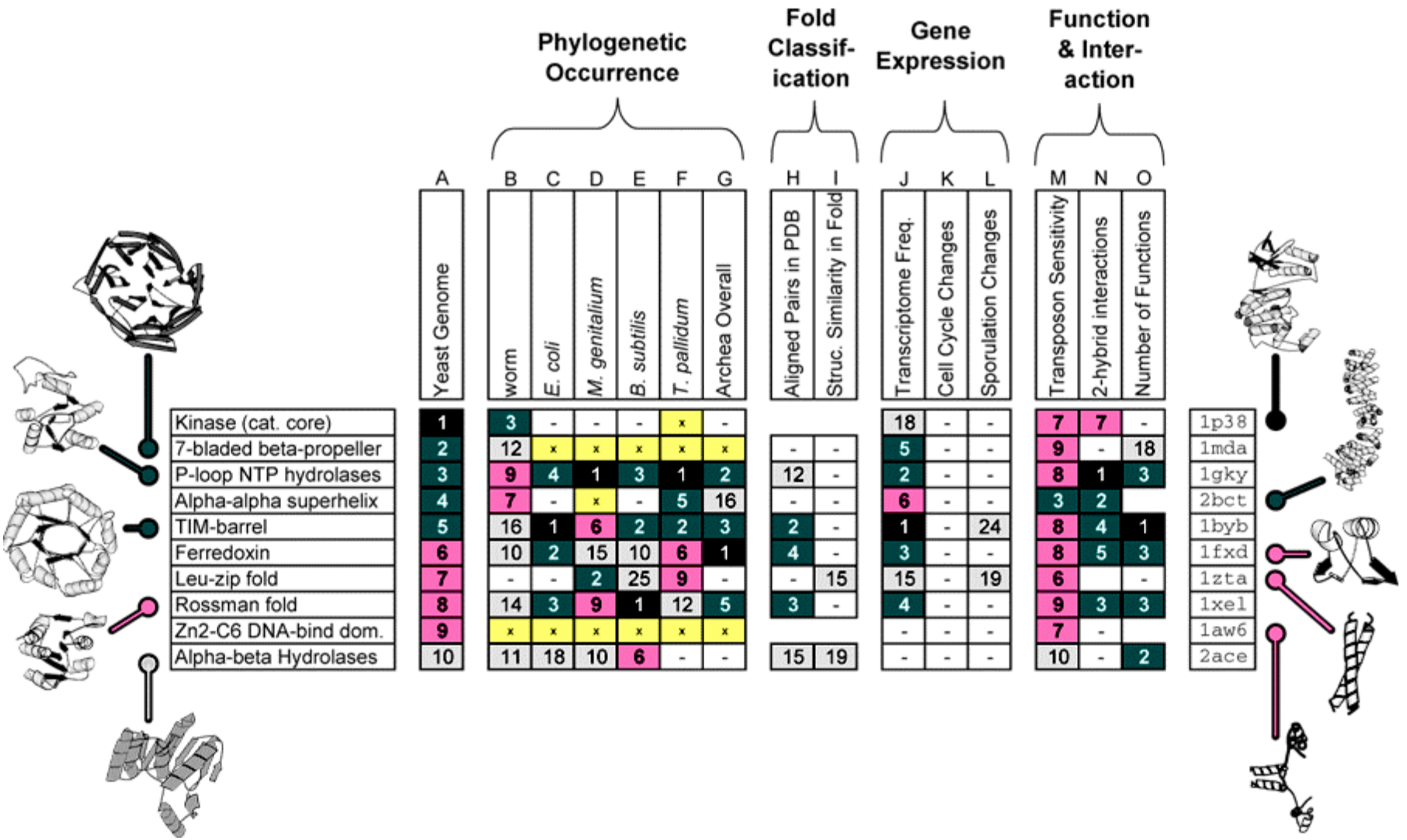




Rank Folds by Genome Occurrence, Expression, Fold Clustering, Length, &c

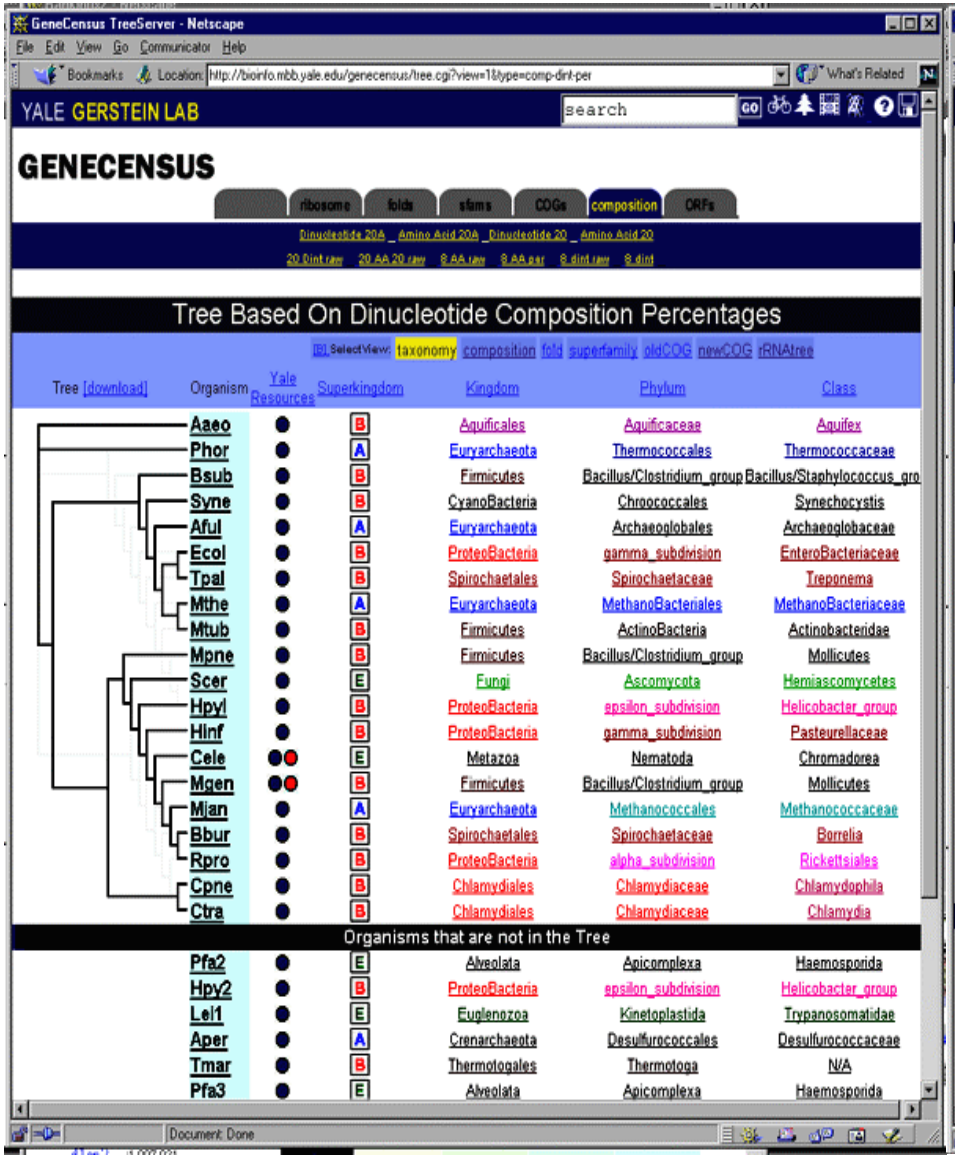
J Qian, B Stenger, J Lin....

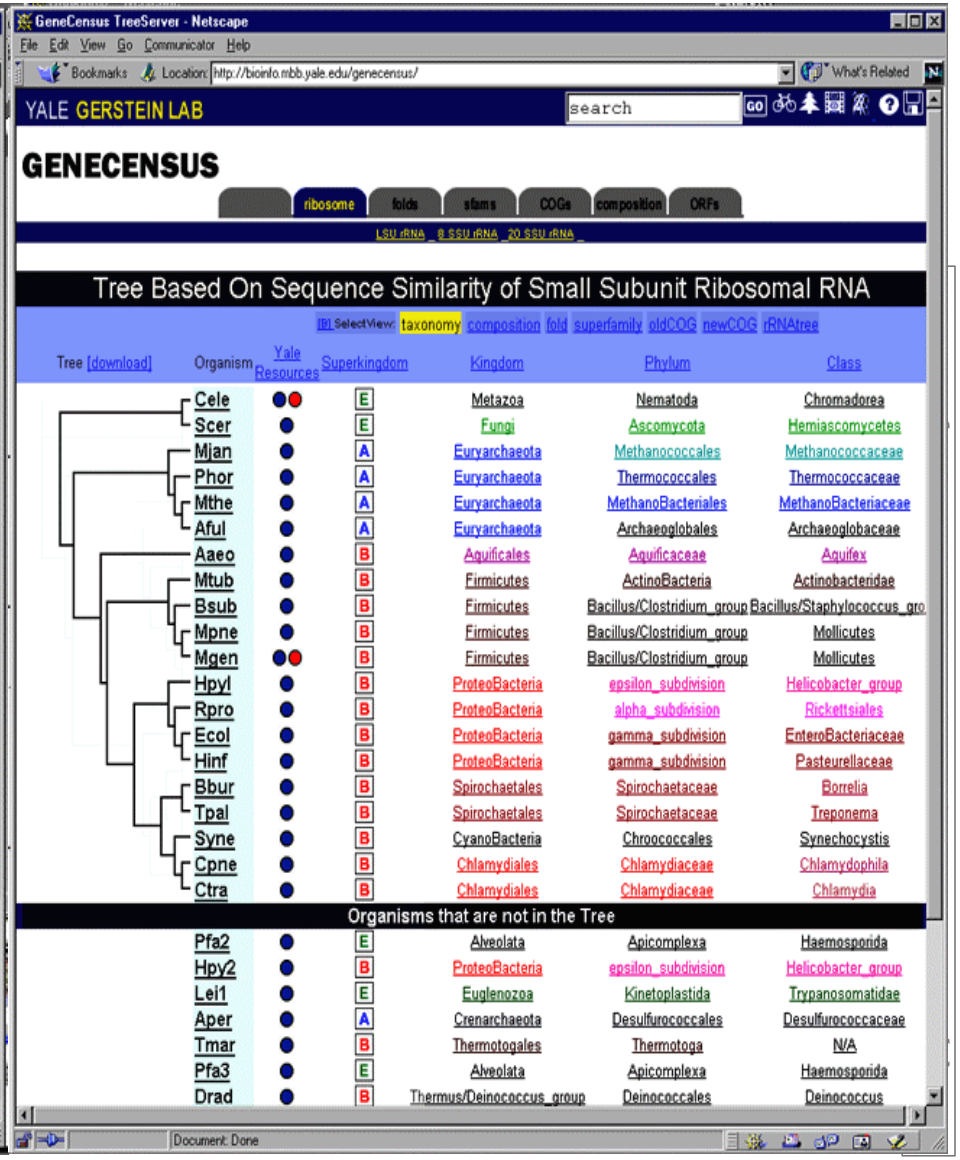
Surveying a Finite PartsList from Many Perspective



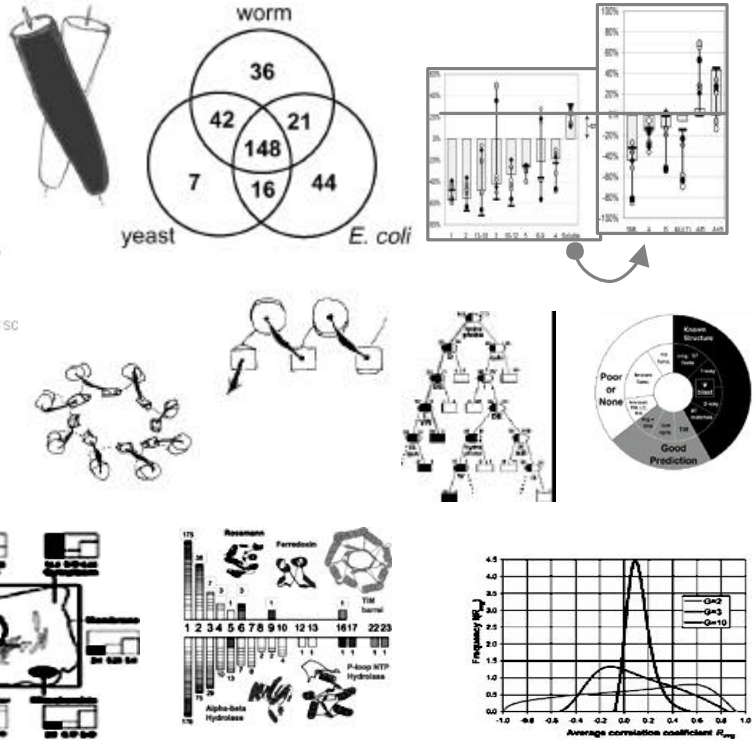
GeneCensus Dynamic Tree Viewers

Recluster organisms based on folds, composition, &c and compare to traditional taxonomy





Comparative Genomics: Surveys of a Finite Parts List



1 Using Folds to Interpret Genomes.

Fold Library background. Shared and/or unique parts. Venn Diagram, Fold tree with all-β diff. Ortholog tree. Horizontal Transfer. Commor Parts: Top-10 folds with βαβ. Common Ψ-fold.

- **Tricky Issues →Expt. SG.** Extent of fold assignment (MG, 65%, 98). Predictions. Biases. NESG.org Construct DB. Datamining this w/ decision trees. Selecting weird MG CD targets.
- **Folds & Functions.** Roles/part? How many folds/func? 331 of ~20K combinations. Mostly 1 func/fold, but some versatile scaffolds -- TIM most versatile. Similar for interactions. Func. Divergence vs. Seq. & Struc. Diverg.

4 Using Folds to Interpret Expression Data.

Top-10 parts in other terms. Enriched in transcriptome: VGA, αβ folds, energy, synthesis, Cytoplasmic, TIM fold. Depleted: NS, long, TMs, transport, transcription, Nuclear, Leu-zip fold. Bayesian localizer.

5 Tricky Issues: Relating Expression to Function.

Expression relates to structure & localization but to function, globally? Weak relation to protein-protein interactions.

H Hegyi, J Lin, B Stenger, N Echols, P Bertone, J Qian, L Regan, S Balasubramanian, V Alexandrov, G Montelione, A Edwards, C Wilson, Y Kluger, C Arrowsmith, A Drawid, R Jansen, D Greenbaum, S Teichmann, P Harrison

bioinfo.mbb.yale.edu