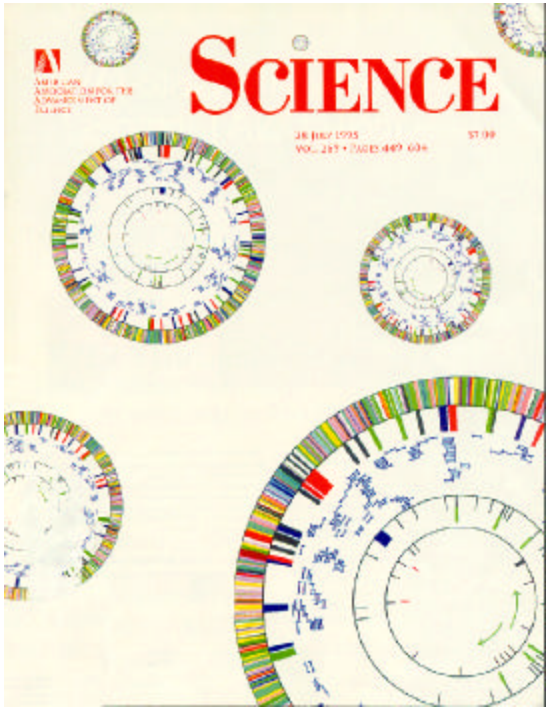
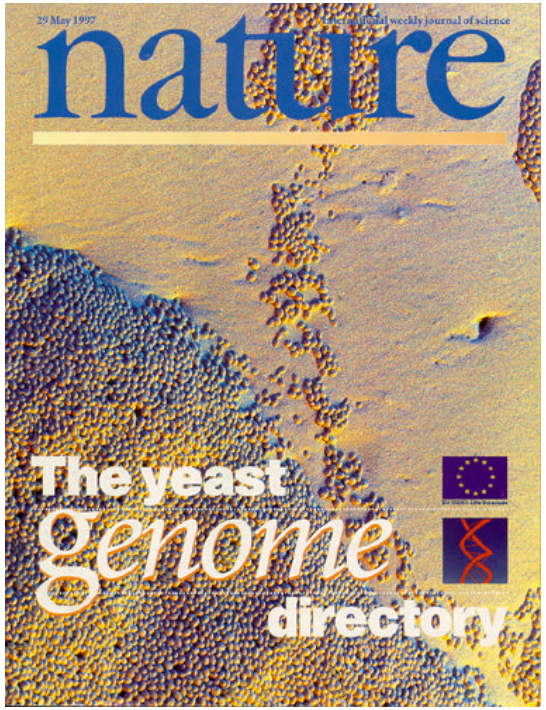
Using a Finite Parts List to Understand Complexity:

Analysis of Whole-Genome Expression Patterns in terms of Protein Folds

Mark Gerstein

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



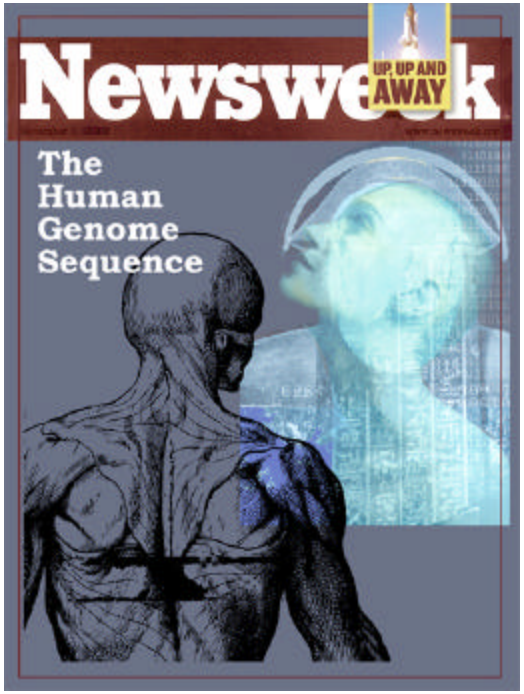


1995

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

1997

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



1998

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

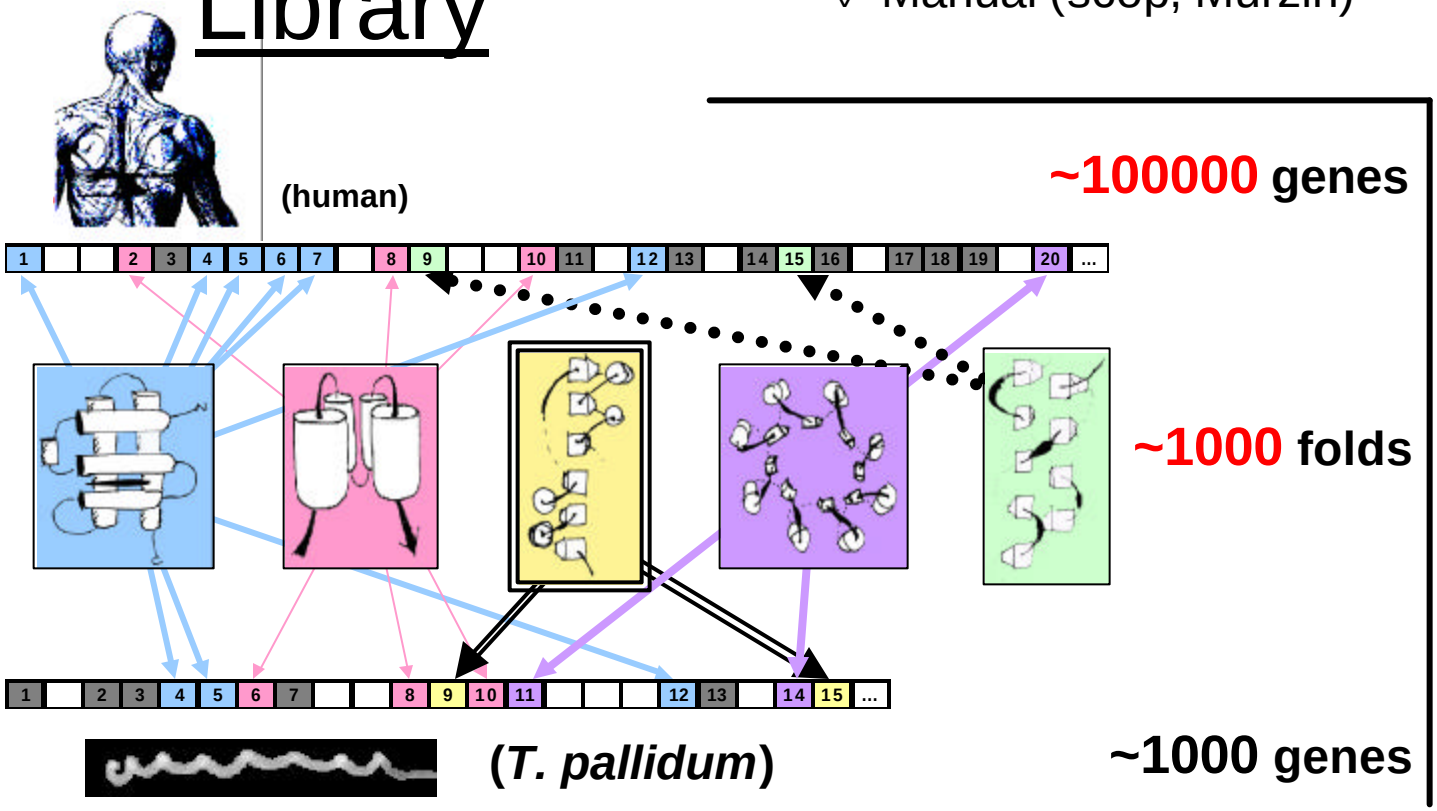
2000?

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

The World of Structures is also Finite: A Fold Library

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

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



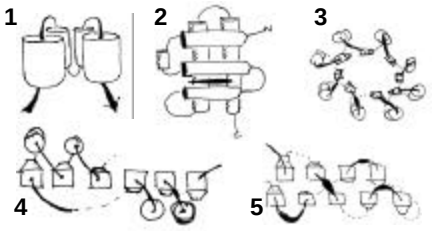
Using a Finite Parts List to Understand Complexity: Analysis of Whole-Genome Expression Patterns in terms of Protein Folds

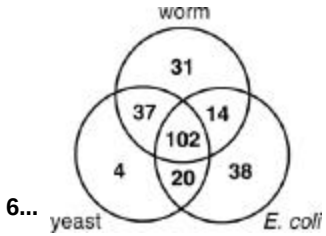
1 Past: Comparing Genomes in terms of Protein Folds

Fold Library, Shared and Unique Folds (Venn, fold tree), Common Folds (top-10), Repeated βαβ, Relation of Structural Class to Functional Class, Bias Problem, Prediction (esp. TMs)

2 Preliminary: Integrating Expression Data into the Analysis

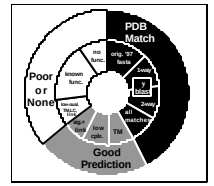
Brown lab yeast data, Top fold by expression, 12-TMs exp. Change, Functional Class Correlated with Expression Change



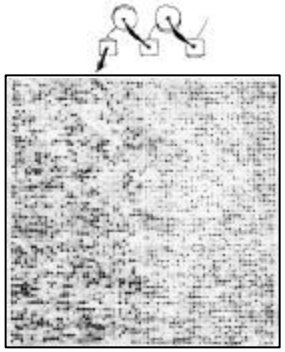


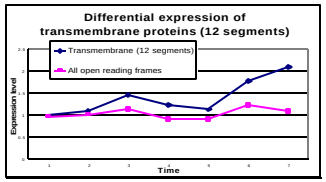
SCOP

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		

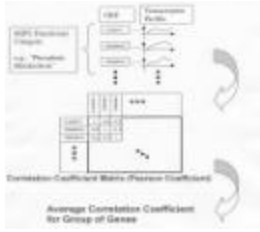


Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	125d	6	6	7
7-bladed beta-propeller	2bbk-H	7	8	16
TIM-barrel	1byb	8	5	6
like Ferrodoxin	1fxd	9	7	10
DNA-binding 3-helix bundle	1enh	10	30	36
...	...			
GroES-like	1lep-A	17	10	9
...	...			
like HSP70, Ct-dom.	1dkz-A	20	11	8



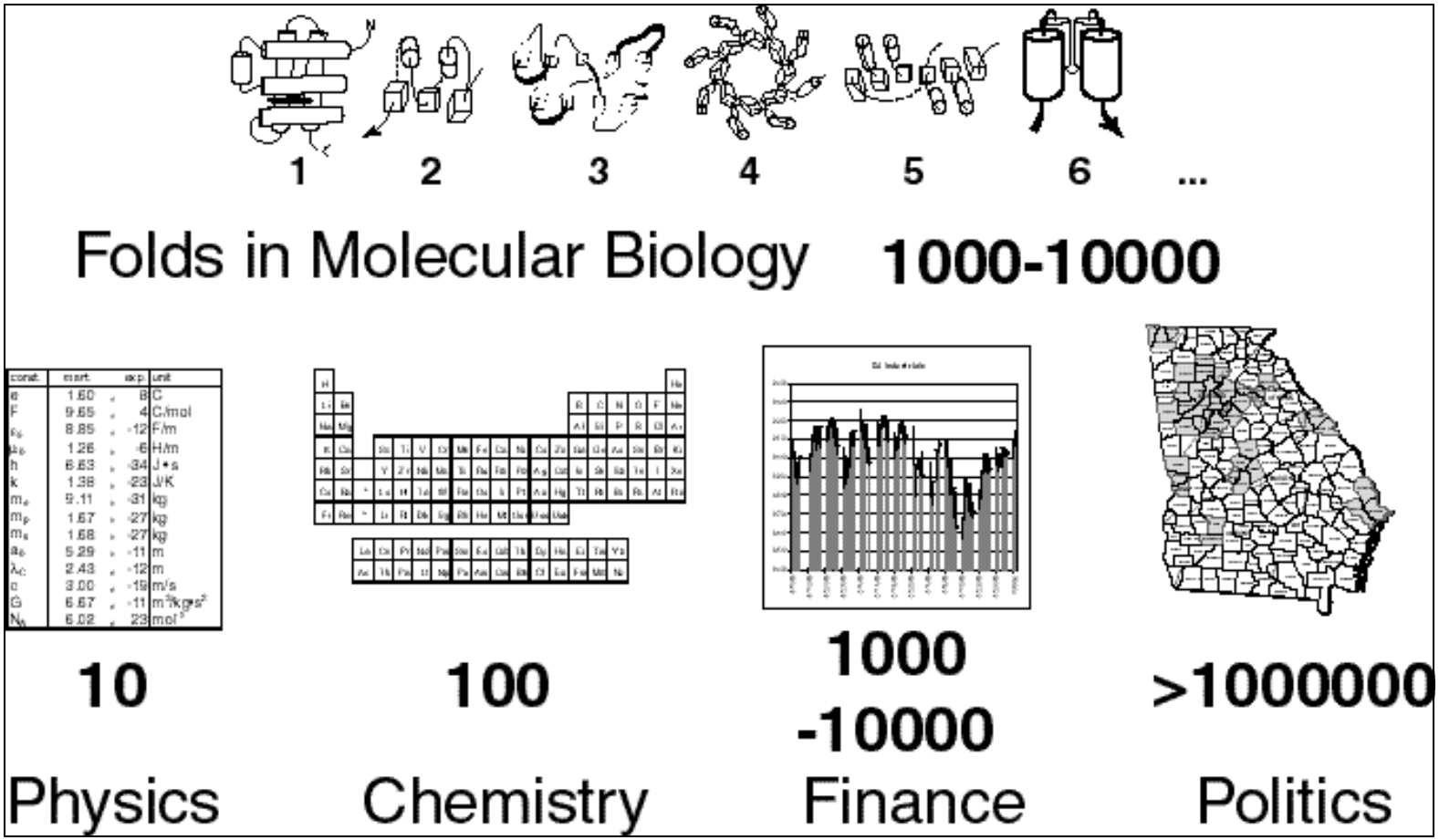


http://bioinfo.mbb.yale.edu/genome



Fold Library vs. Other Fundamental Data structures

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



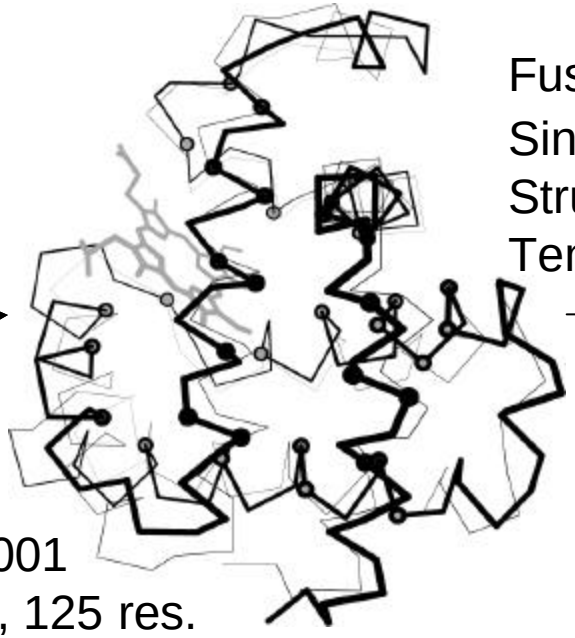
(Large than physics and chemistry, Similar to Finance (Exact Finite Number of Objects (3,056 on NYSE by 1/98), descrip. by Standardized Statistics (even abbrevs, INTC) and groups (sectors)) Smaller than Social Surveys, Indefinite Number of People, Not Well Defined Vocabulary and statistics.

Building a Library of Know Folds

Hb

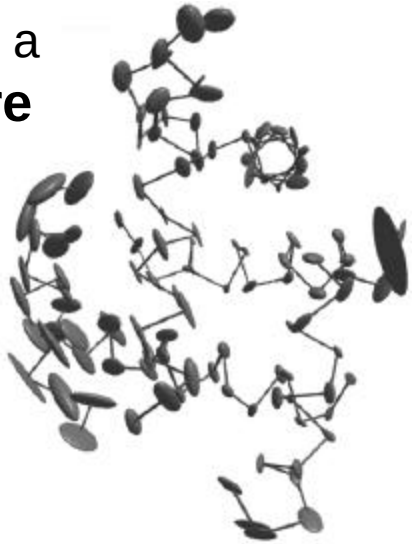
Alignment
of Individual
Structures

Mb

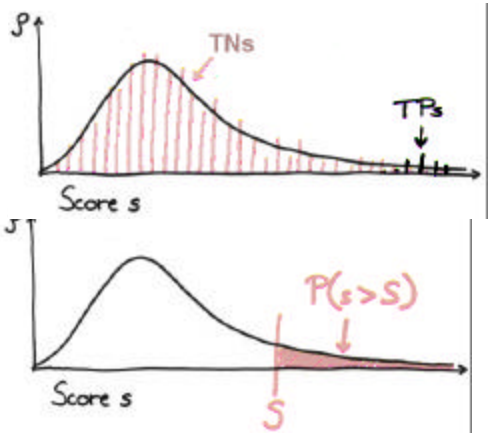


P<.000001
RMS~1, 125 res.

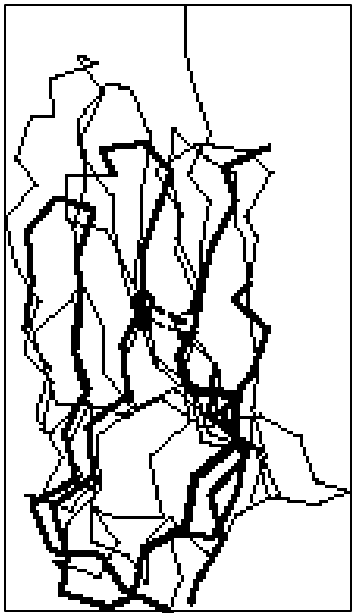
Fusing into a
Single **Core**
Structure
Template



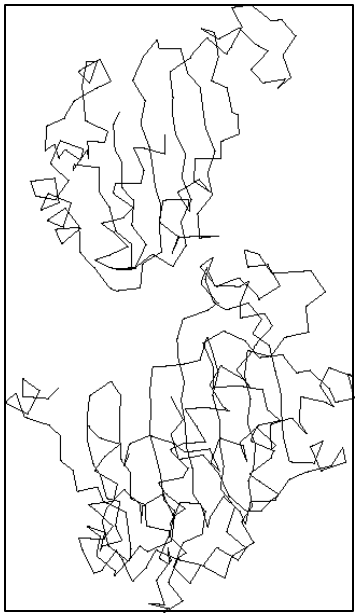
Statistics
to Establish
Relation-
ships
(P-values)



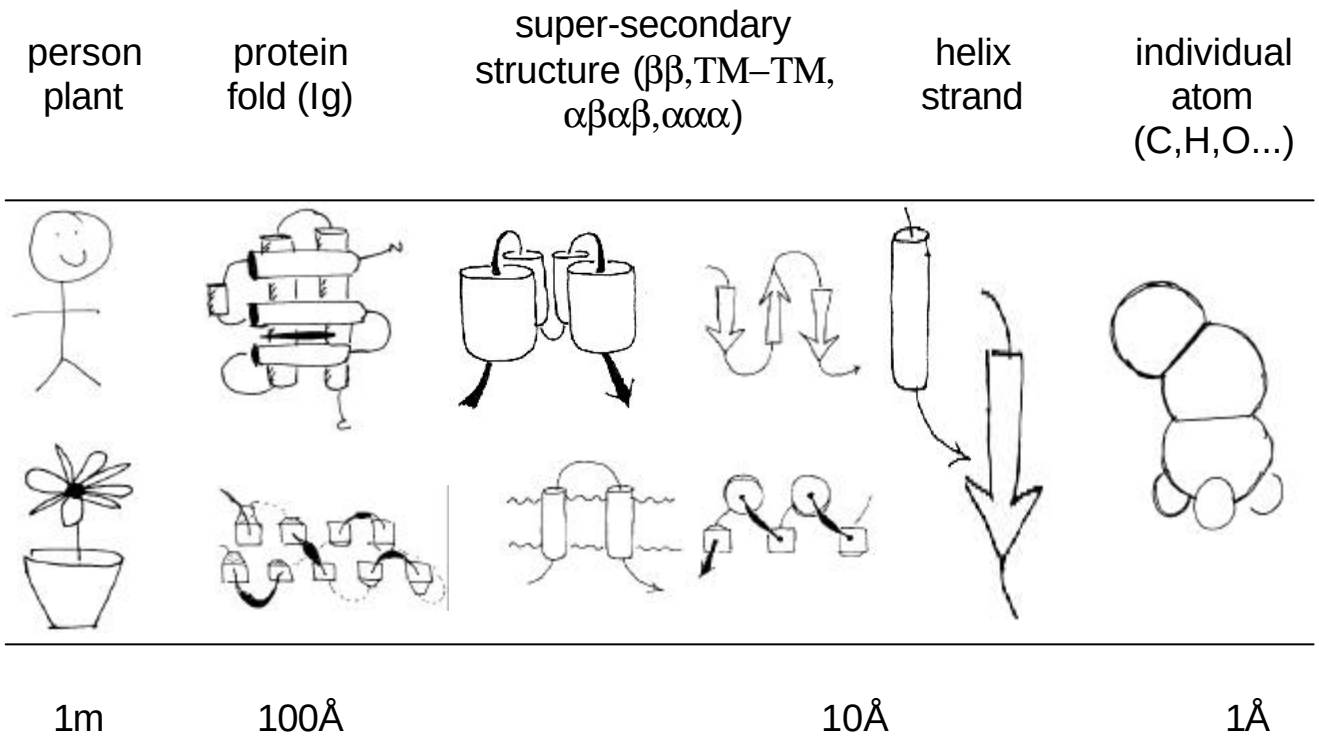
P<.001
RMS~3
100 res

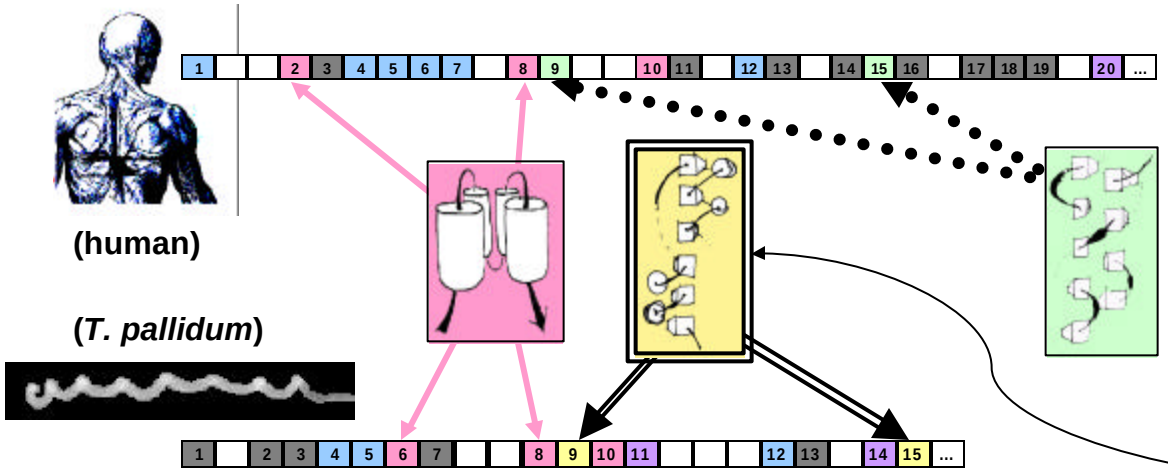


P~1



At What Structural Resolution Are Organisms Different?

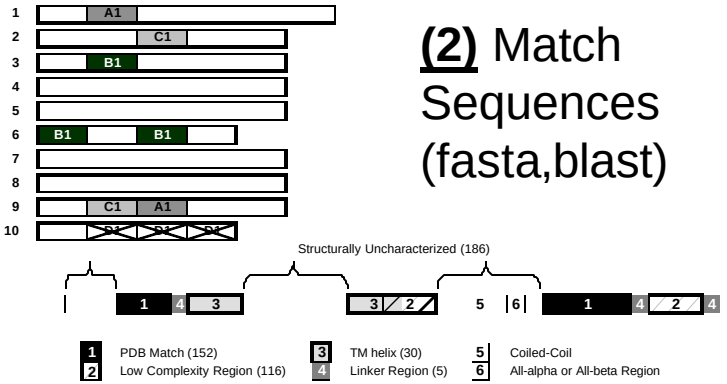




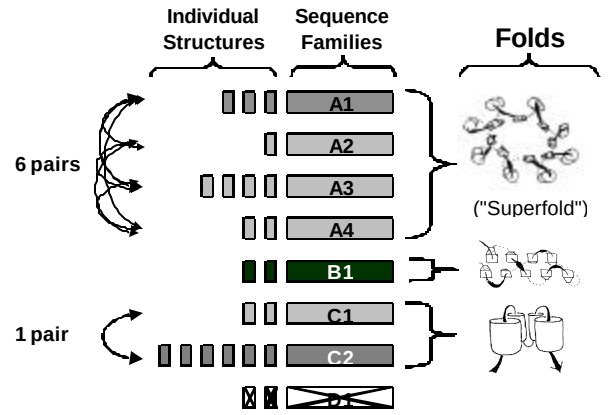
Practical Relevance

(Pathogen only folds as possible targets)

Cross-Reference: Folds→Sequences → Organisms

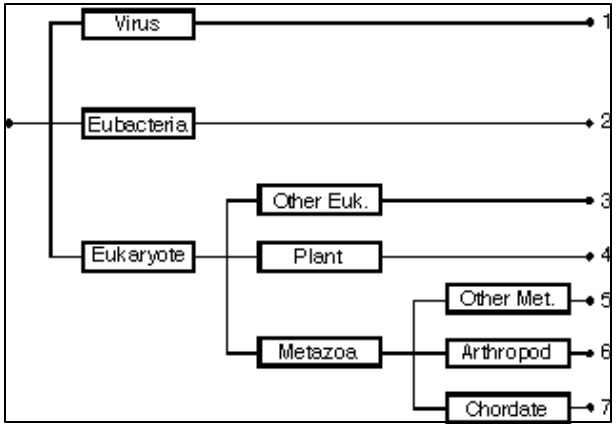


(2) Match Sequences (fasta,blast)



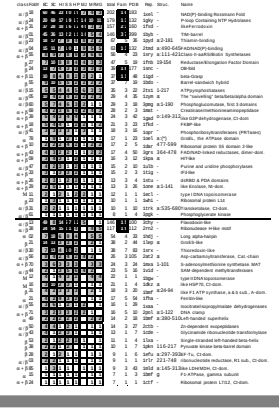
(1) Structures in Folds (scop)

(3) Organize Sequences by Genome or Taxon



3 + 5

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

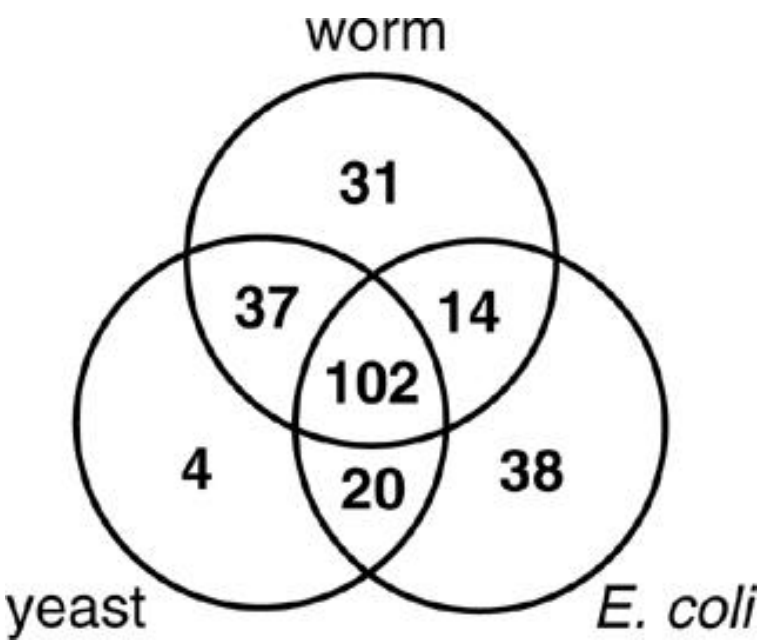


(4) Results in "Fold Table"

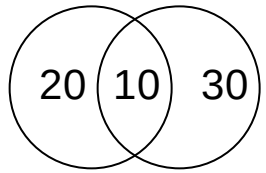
class	Fold#	EC	SC	HI	SS	HP	MJ	MP	MG	total	Fam.	PDB	Rep.	Struc.	Name
α/β	18	60	46	23	40	19	7	4	3	202	16	183	1xel	-	NAD(P)-bindi
α/β	24	20	69	17	19	17	16	10	11	179	13	132	1gky	-	P-loop Contai
α+β	31	37	28	18	16	12	40	3	3	157	23	160	1fxd	-	like Fe odox
α/β	01	45	36	13	22	11	10	5	4	146	37	399	1byb	-	TIM-ba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

Shared Folds in Initial Genomes

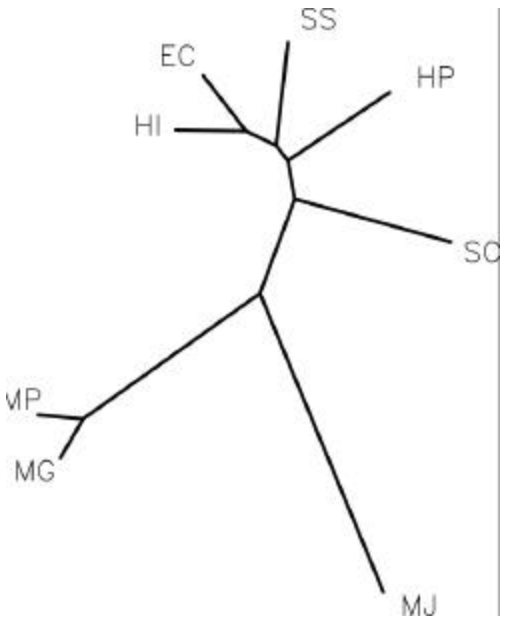
How 339 Folds are Distributed in 3 genomes?



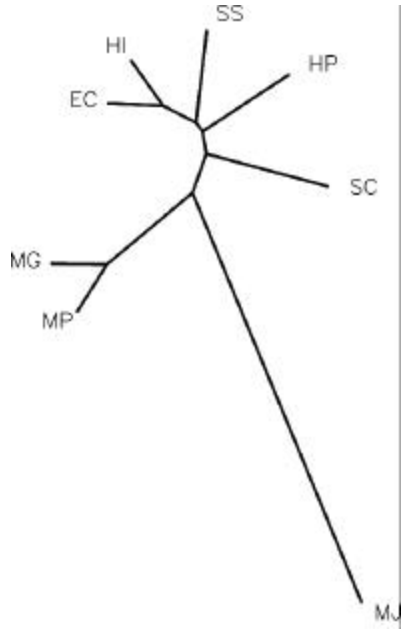
Cluster 8 genomes (Dist ~ # shared folds)



D=10/(20+10+30)



Fold Tree

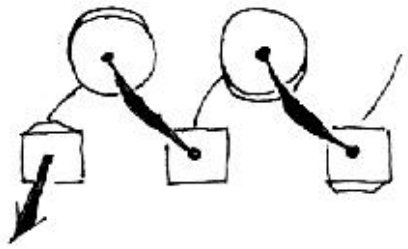


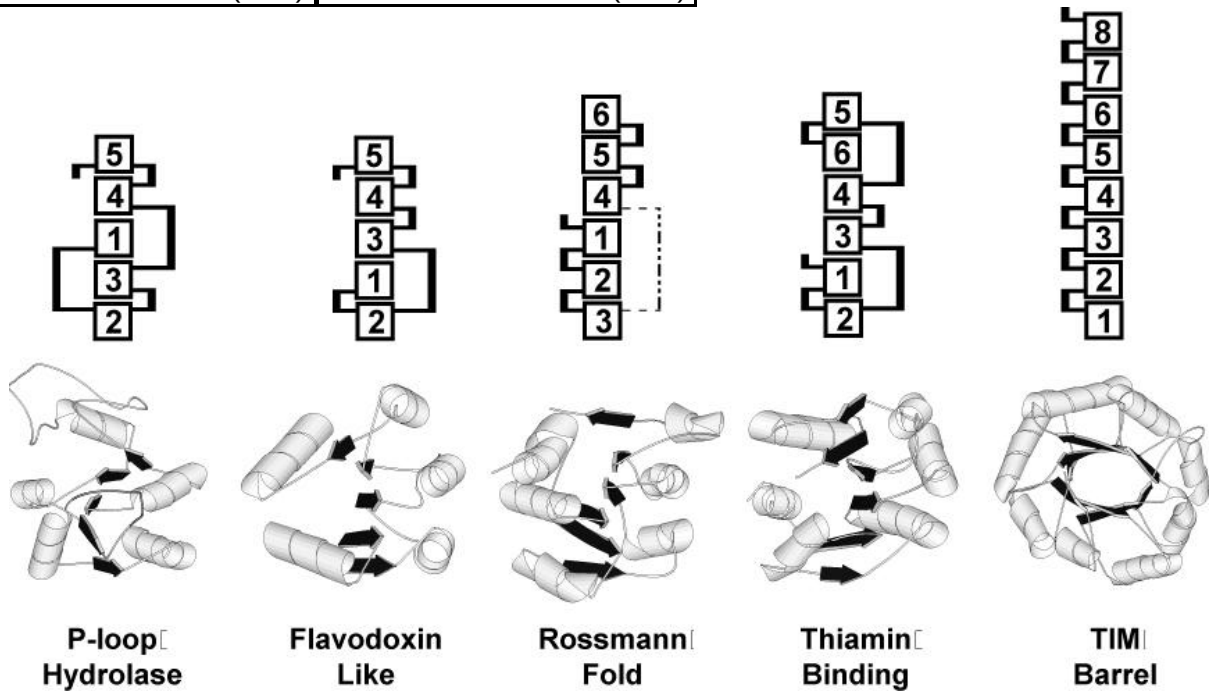
"Classic" Tree

Most Common Folds

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

βαβ unit common in microbial genomes

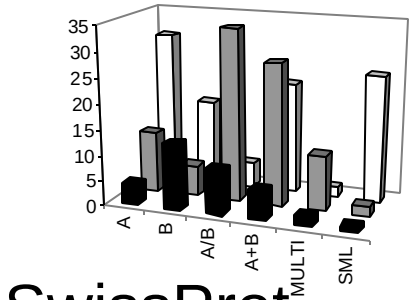




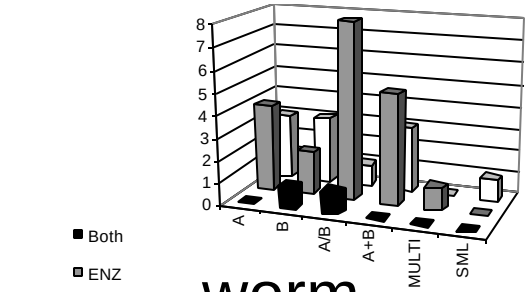
Relate Overall Structural Classes to Functional Classes

SCOP

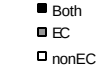
ENZYME	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

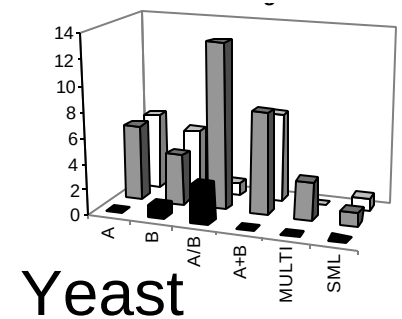


SwissProt

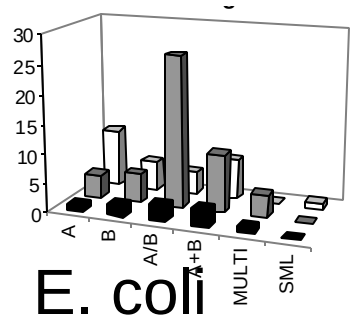


worm





Yeast



E. coli

MIPS Yeast Functional Catalog (Mewes)

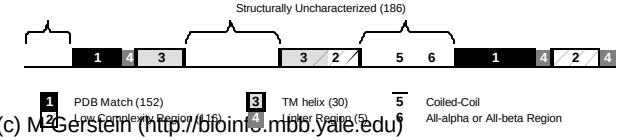
SCOP

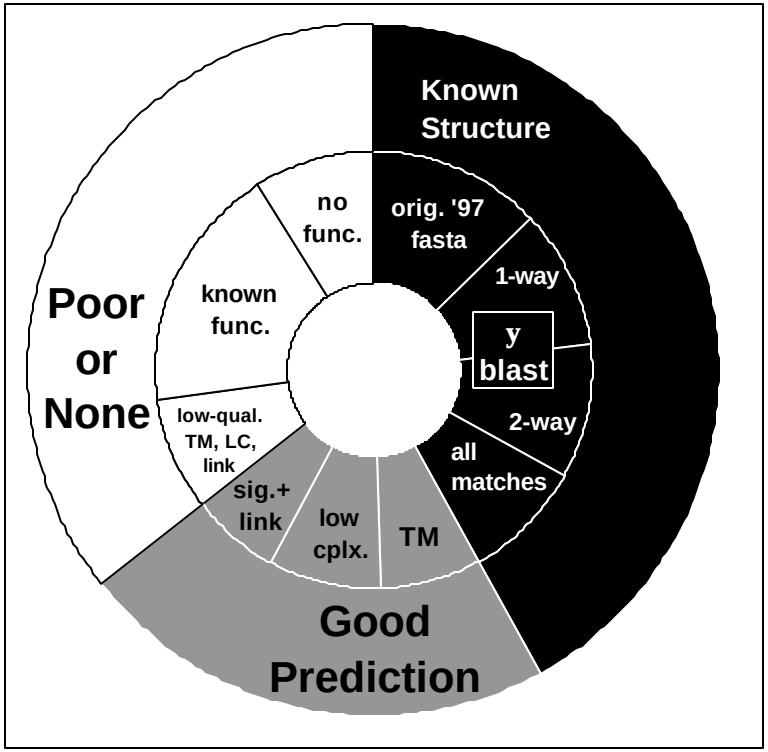
MIPS Functional Cat.		A	B	A/B	A+B	MULTI	SML
metabolism	1	3.5	2.3	10	4.5	1.3	0.8
energy	2	1.1	1.2	5	1.5	0.3	0.2
growth, div., DNA syn.	3	4.9	3.6	4	4.5	1.8	1.2
transcription	4	1.5	1.3	2.2	1.5	0.5	0.8
protein synthesis	5	1	0.9	0.7	1.3	0.3	0.2
protein targetting	6	1.2	1.7	2	1.6	0.5	0.3
transport facilitation	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	

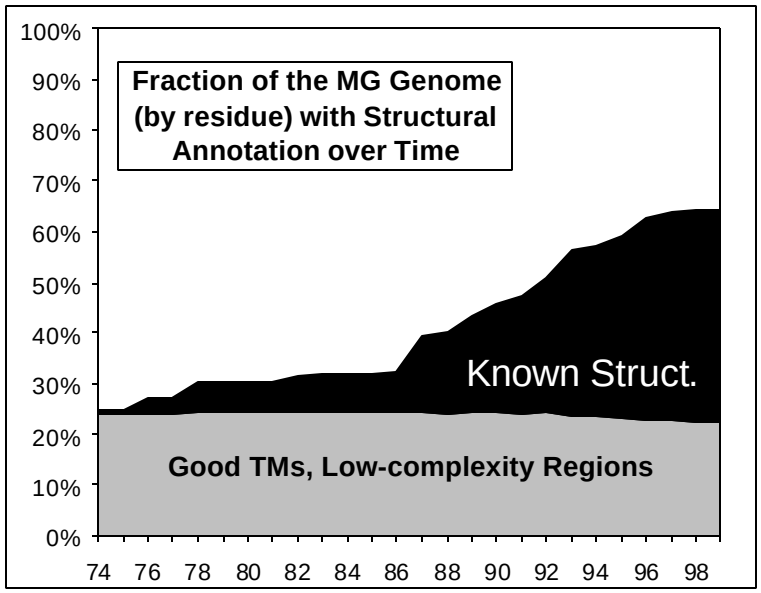
Bias Problem → Prediction

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

Same Sampling Issues with US Census!!



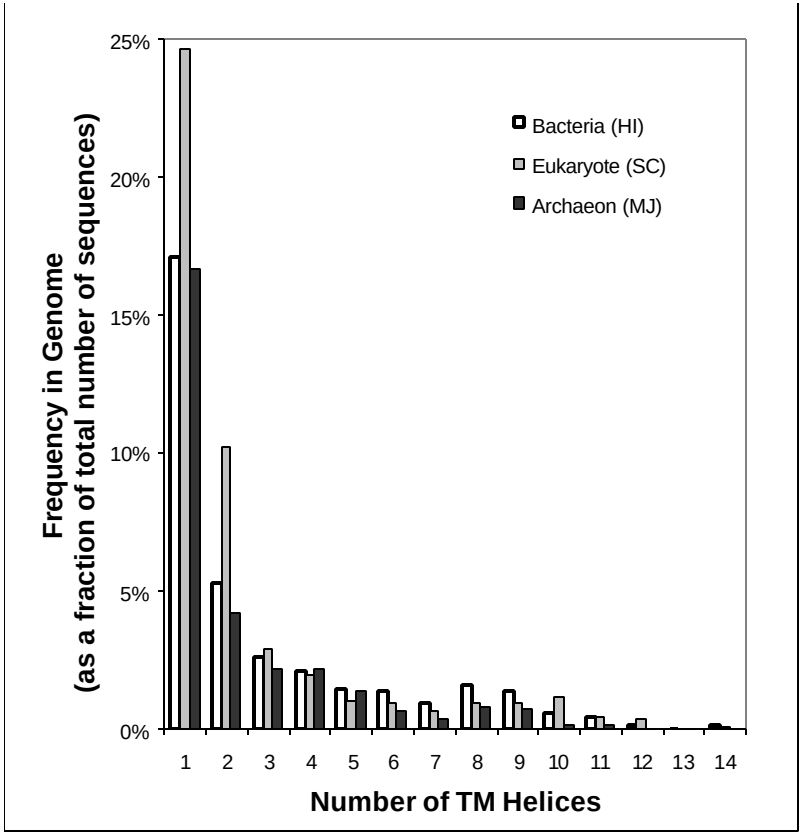


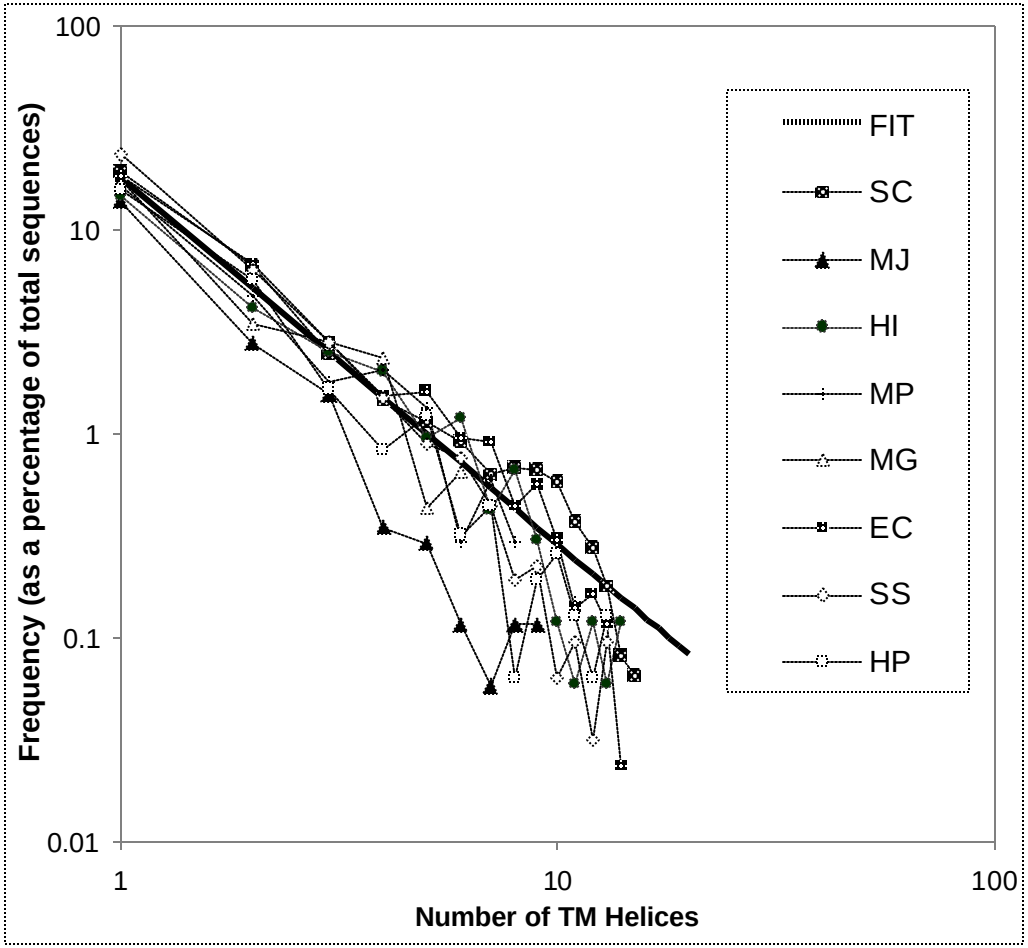


TM-helix "prediction"

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

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





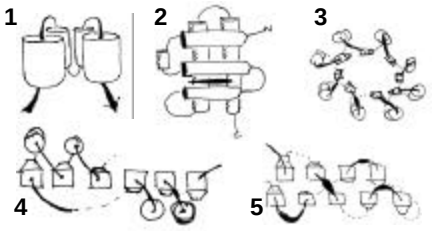
Using a Finite Parts List to Understand Complexity: Analysis of Whole-Genome Expression Patterns in terms of Protein Folds

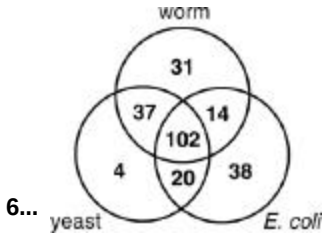
1 Past: Comparing Genomes in terms of Protein Folds

Fold Library, Shared and Unique Folds (Venn, fold tree), Common Folds (top-10), Repeated βαβ, Relation of Structural Class to Functional Class, Bias Problem, Prediction (esp. TMs)

2 Preliminary: Integrating Expression Data into the Analysis

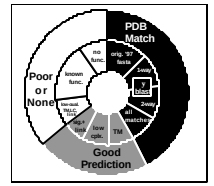
Brown lab yeast data, Top fold by expression, 12-TMs exp. Change, Functional Class Correlated with Expression Change



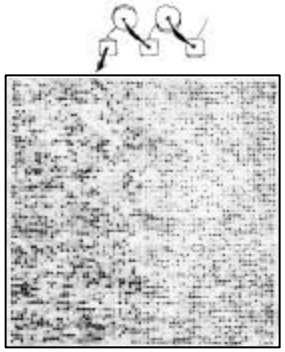


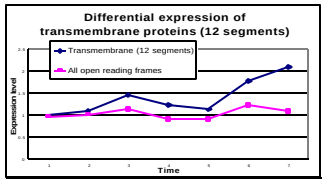
SCOP

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		

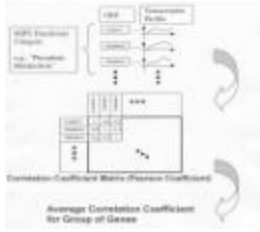


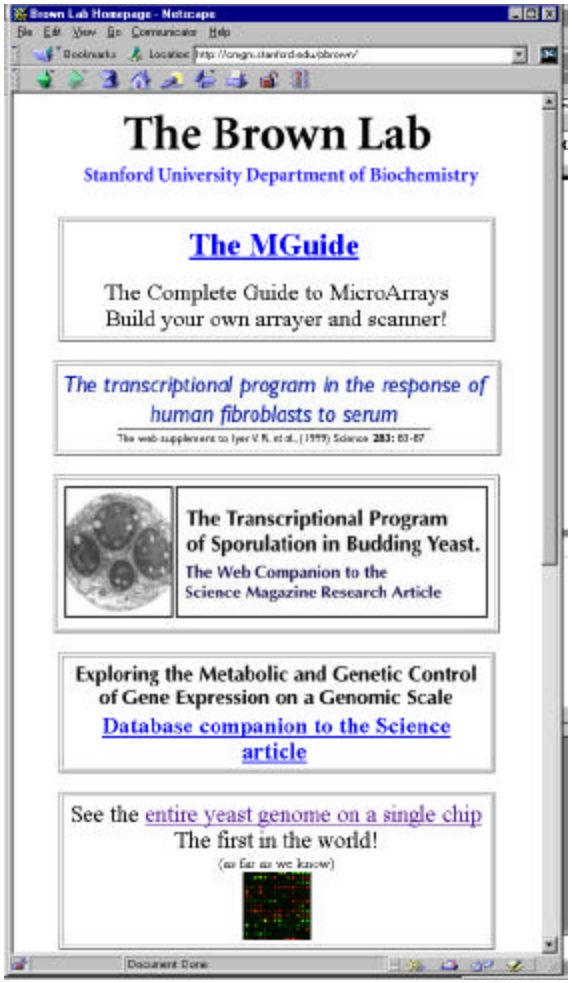
Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	125d	6	6	7
7-bladed beta-propeller	2bbk-H	7	8	16
TIM-barrel	1byb	8	5	6
like Ferrodoxin	1fxd	9	7	10
DNA-binding 3-helix bundle	1enh	10	30	36
...	...			
GroES-like	1lep-A	17	10	9
...	...			
like HSP70, Ct-dom.	1dkz-A	20	11	8



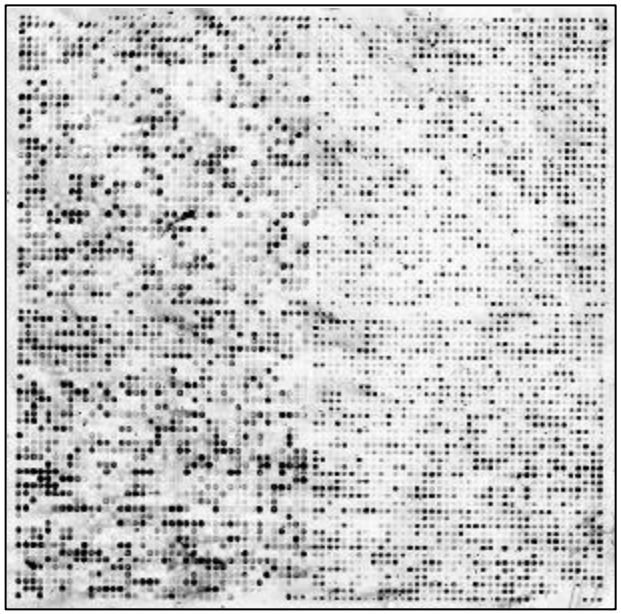


http://bioinfo.mbb.yale.edu/genome





Integrate Gene Expression Data into Folds in Genome Analysis



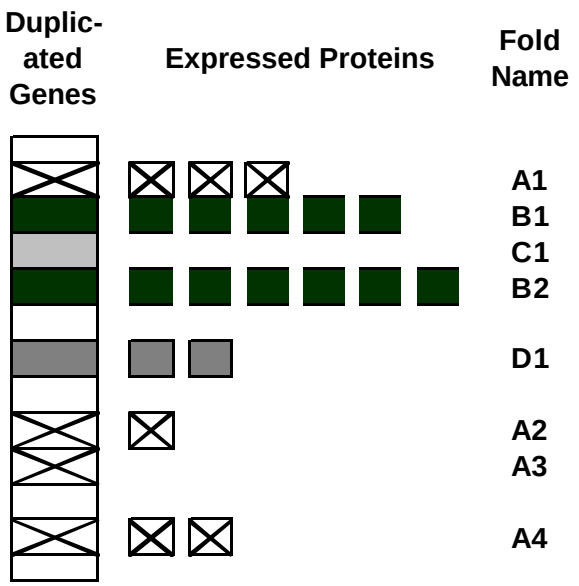
Yeast Expression Data Principally from Brown lab site (1st available!).

Also: SAGE data, Church lab data, Snyder lab transposon data, Young lab data

Caveat: With GeneChips, only rough ORF-to-ORF comparison possible, 1-ORF comparisons along timecourse better. SAGE and transposon data better for this regard. Mostly we just aggregate the data.

Data from: "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale". DeRisi et al. Science 278:680. Also, Chu *et al. Science* 1998 October 23; 282: 699-705

Top-10 Folds according to Expression



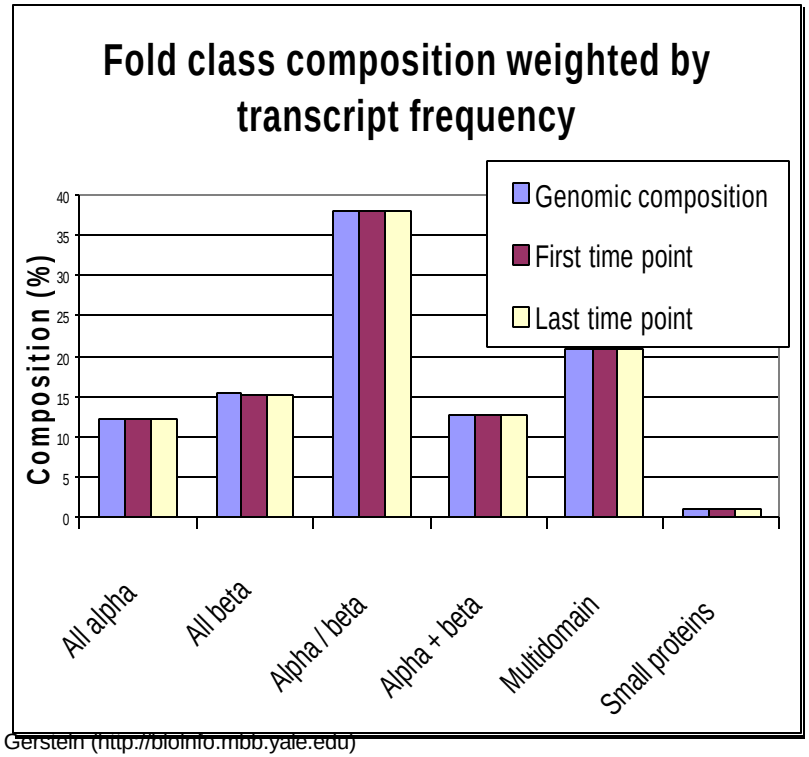
Tot.	Dup.	Exp.
A	4	6
B	2	11
C	1	0
D	1	2

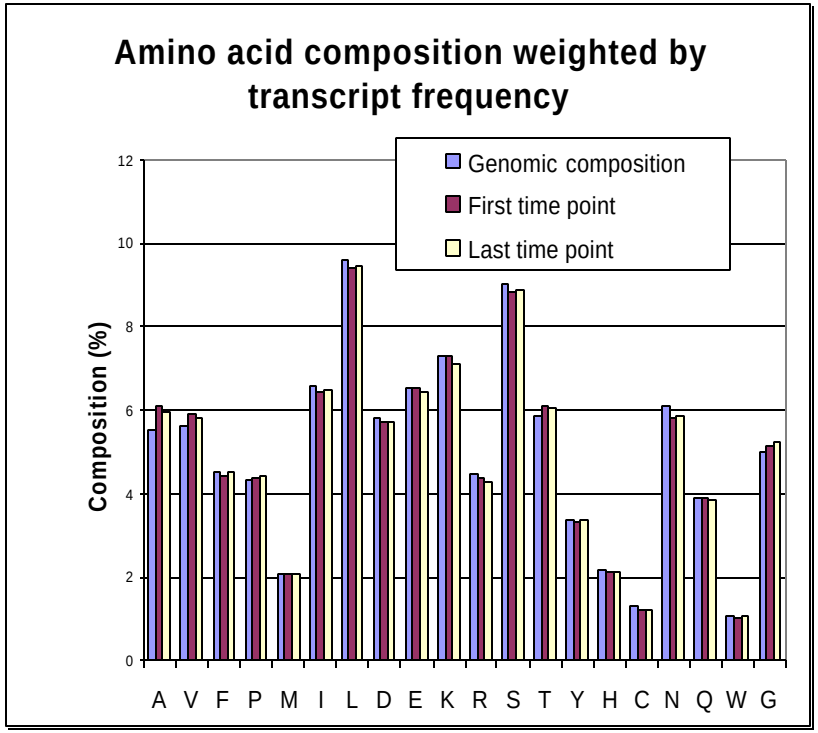
- Previous top-10 measures duplication
- Now weight by expression using data from Brown et al.

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

Expression not related to Overall Fold Class or Overall Composition

- Fold class composition weighted by transcript frequency does **not** change during differential expression of genes.
- Amino acid composition weighted by transcript frequency does **not** change during differential expression of genes.





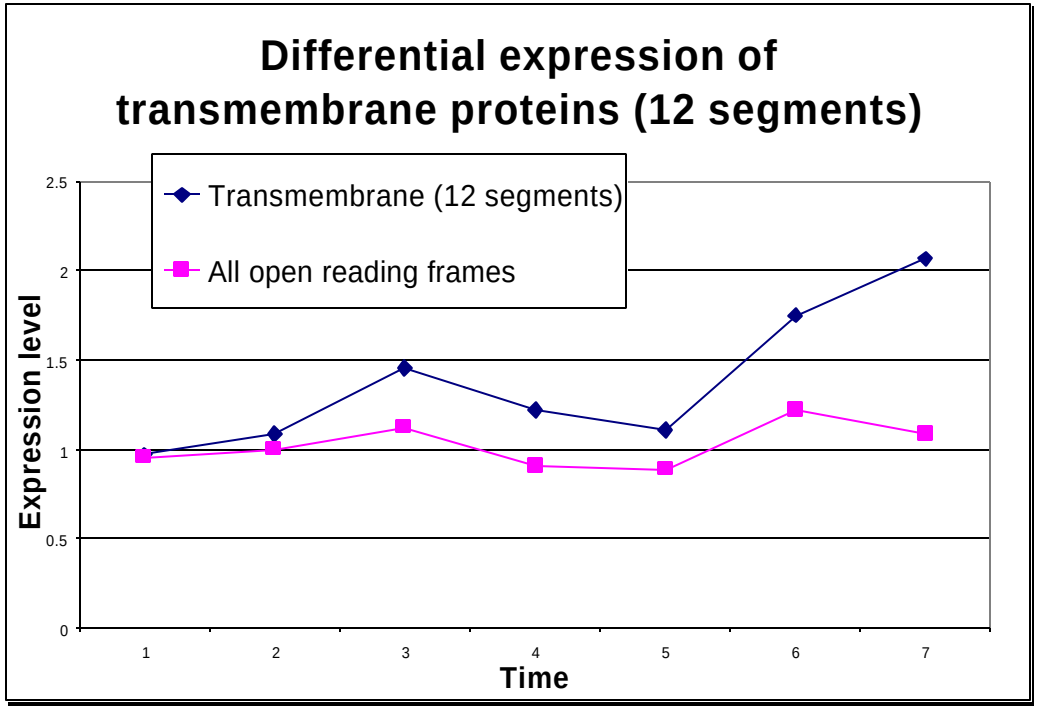
Different Classes of Membrane Proteins Have Different Changes in Expression Level (esp. 12 TMs)

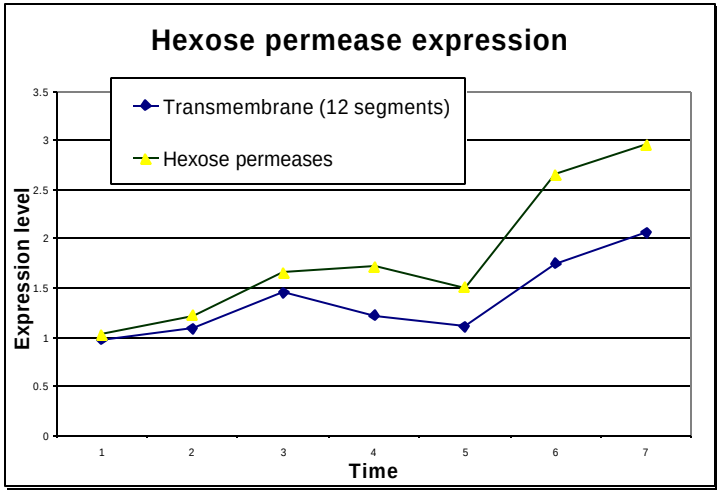
Most Expressed TMs in aerobic conditions

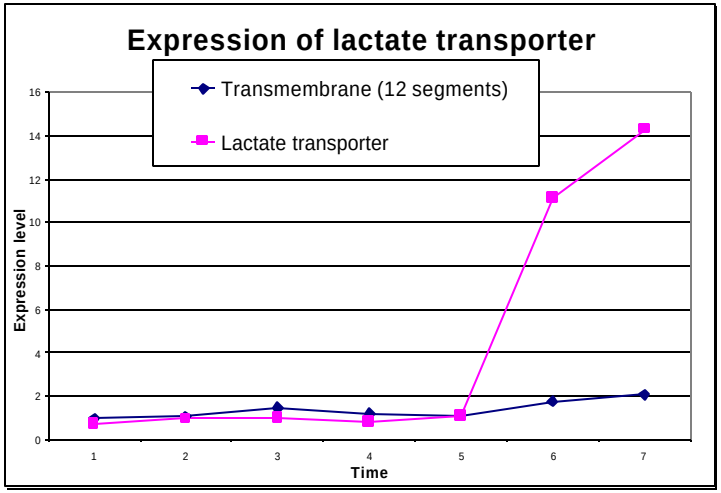
ORF	TMs
YHR078W	4
YGL008C	6
YBR012W-B	2
YLR340W	2
YPL131W	2
YHR099W	2
YMR205C	2
YHR216W	2
YLR432W	2
YIL075C	5

Most Expressed TMs in anaerobic conditions

ORF	TMs
YPR149W	4
YDR343C	9
YDR342C	9
YKL217W	7
YHR096C	9
YBR116C	2
YIL088C	6
YBR012W-B	2
YBR054W	7
YBR218C	2





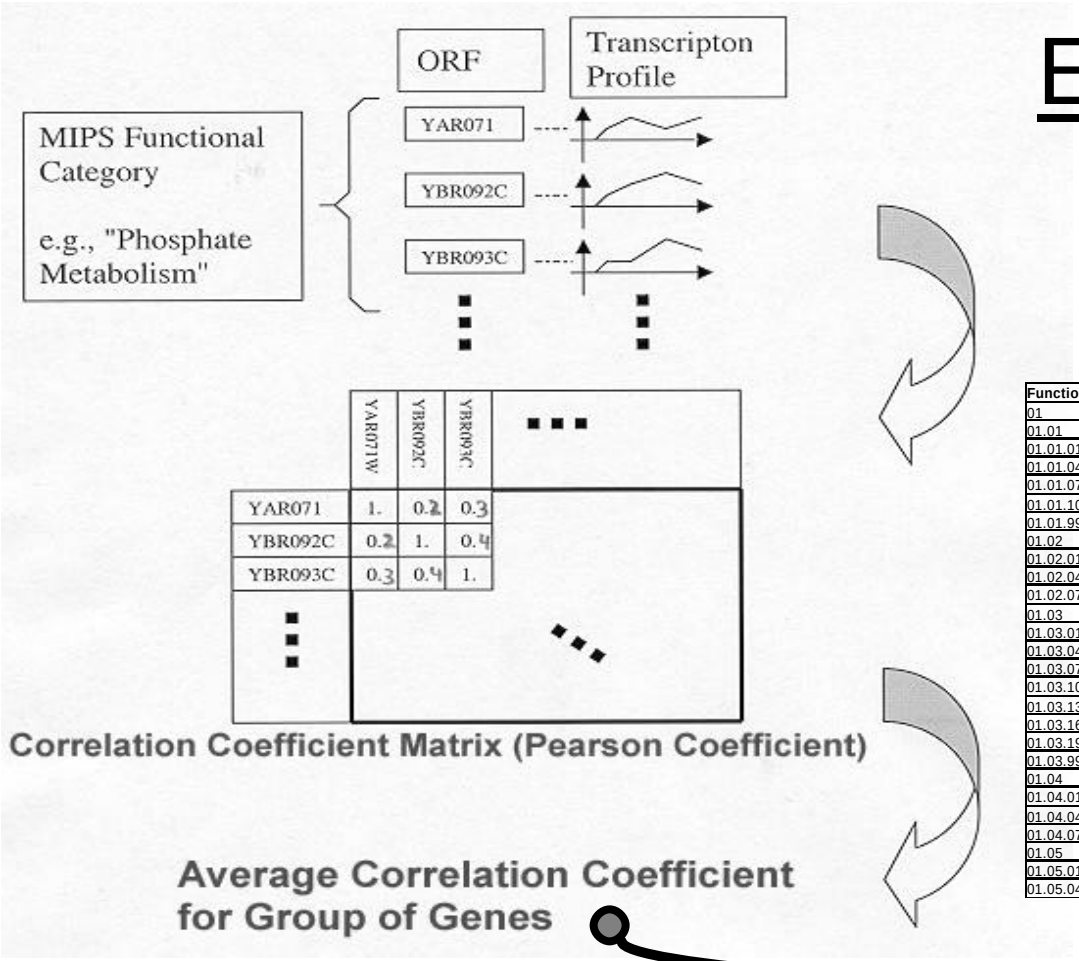


Column gives the expression in aerobic conditions (high sugar, second time-series data point in DeRisi et al.), and other column, in anaerobic conditions (low sugar, high ethanol, last time-series data point in DeRisi et al.). 9 hexose permeases, 1 lactate transporter.

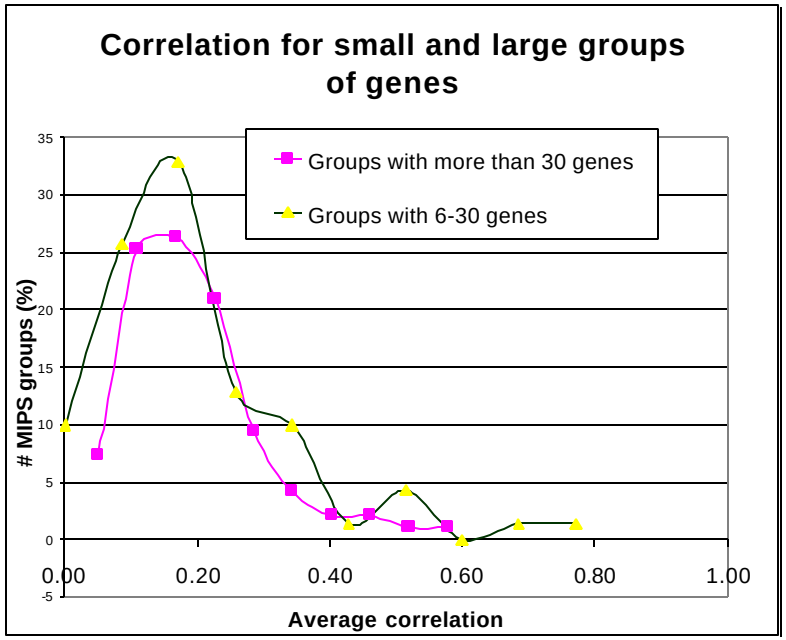
Functional category number	Function	Average correlation	# ORFs
01	METABOLISM	0.1001	1005
01.01	amino-acid metabolism	0.1488	199
01.01.01	amino-acid biosynthesis	0.239	114
01.01.04	regulation of amino-acid metabolism	0.23	32

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

Correlate with Expression Level with 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	36
01.01.99	other amino-acid metabolism activities	0.2205	4
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	134
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	1
01.03.16	polynucleotide degradation	0.2461	0
01.03.19	nucleotide transport	0.1187	12
01.03.99	other nucleotide-metabolism activities	-0.0328	7
01.04	phosphate metabolism	0.1348	31
01.04.01	phosphate utilization	0.16	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



Results from Analysis of Correlation of Functional Class and Expression

Highest Correlations

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

Functional category numbe	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

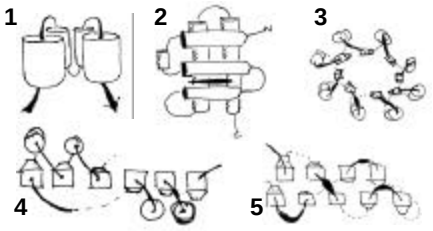
Using a Finite Parts List to Understand Complexity: Analysis of Whole-Genome Expression Patterns in terms of Protein Folds

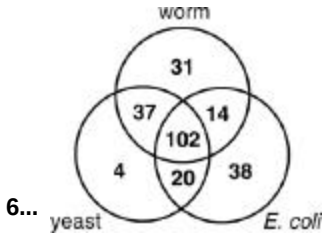
1 Past: Comparing Genomes in terms of Protein Folds

Fold Library, Shared and Unique Folds (Venn, fold tree), Common Folds (top-10), Repeated βαβ, Relation of Structural Class to Functional Class, Bias Problem, Prediction (esp. TMs)

2 Preliminary: Integrating Expression Data into the Analysis

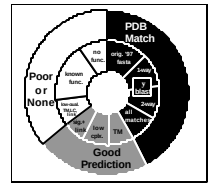
Brown lab yeast data, Top fold by expression, 12-TMs exp. Change, Functional Class Correlated with Expression Change



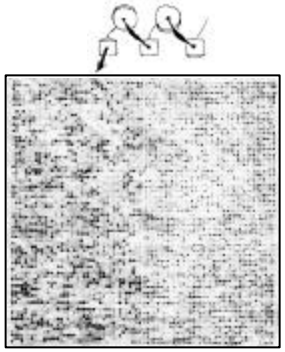


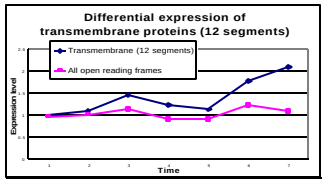
SCOP

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		

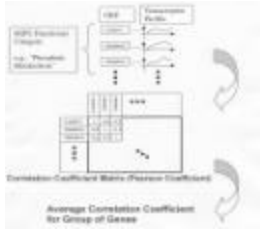


Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	125d	6	6	7
7-bladed beta-propeller	2bbk-H	7	8	16
TIM-barrel	1byb	8	5	6
like Ferrodoxin	1fxd	9	7	10
DNA-binding 3-helix bundle	1enh	10	30	36
...	...			
GroES-like	1lep-A	17	10	9
...	...			
like HSP70, Ct-dom.	1dkz-A	20	11	8





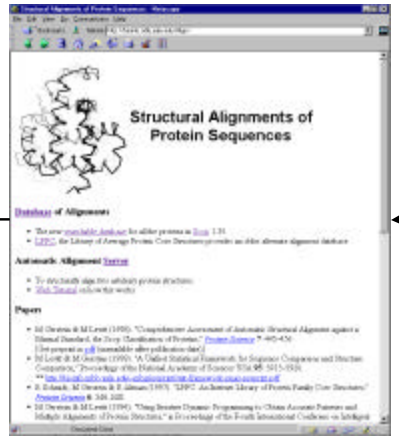
http://bioinfo.mbb.yale.edu/genome

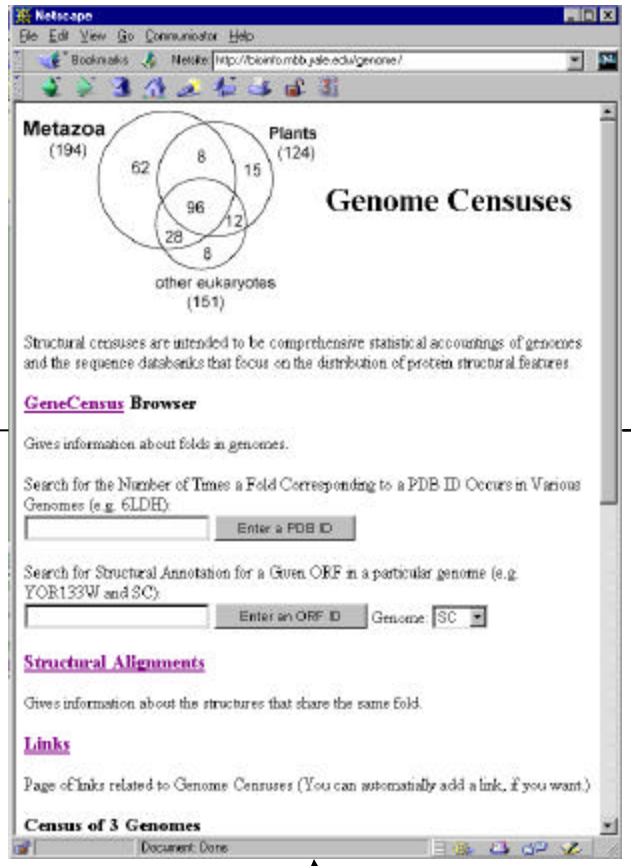


GeneCensus

Alignment Database

Alignment Server

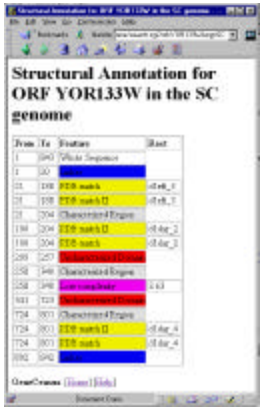




Detailed Tables

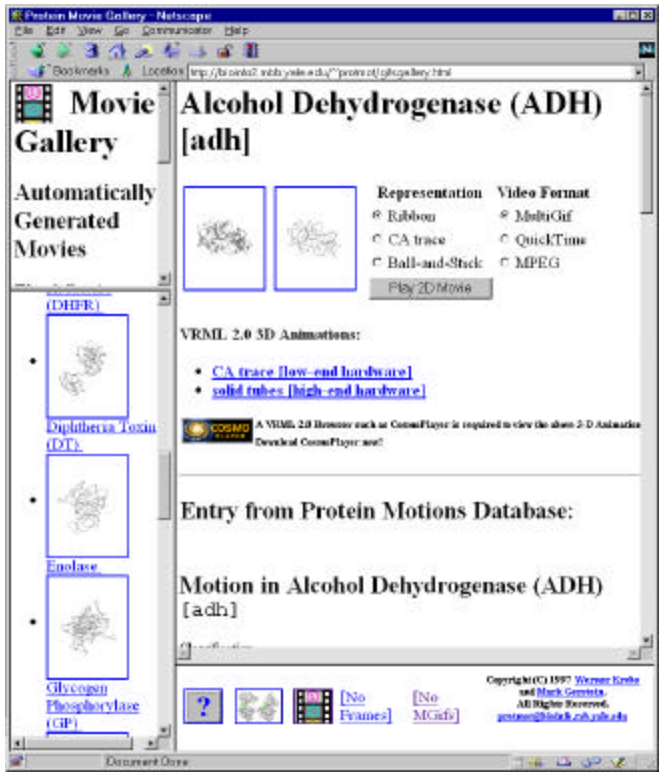
PDB Query

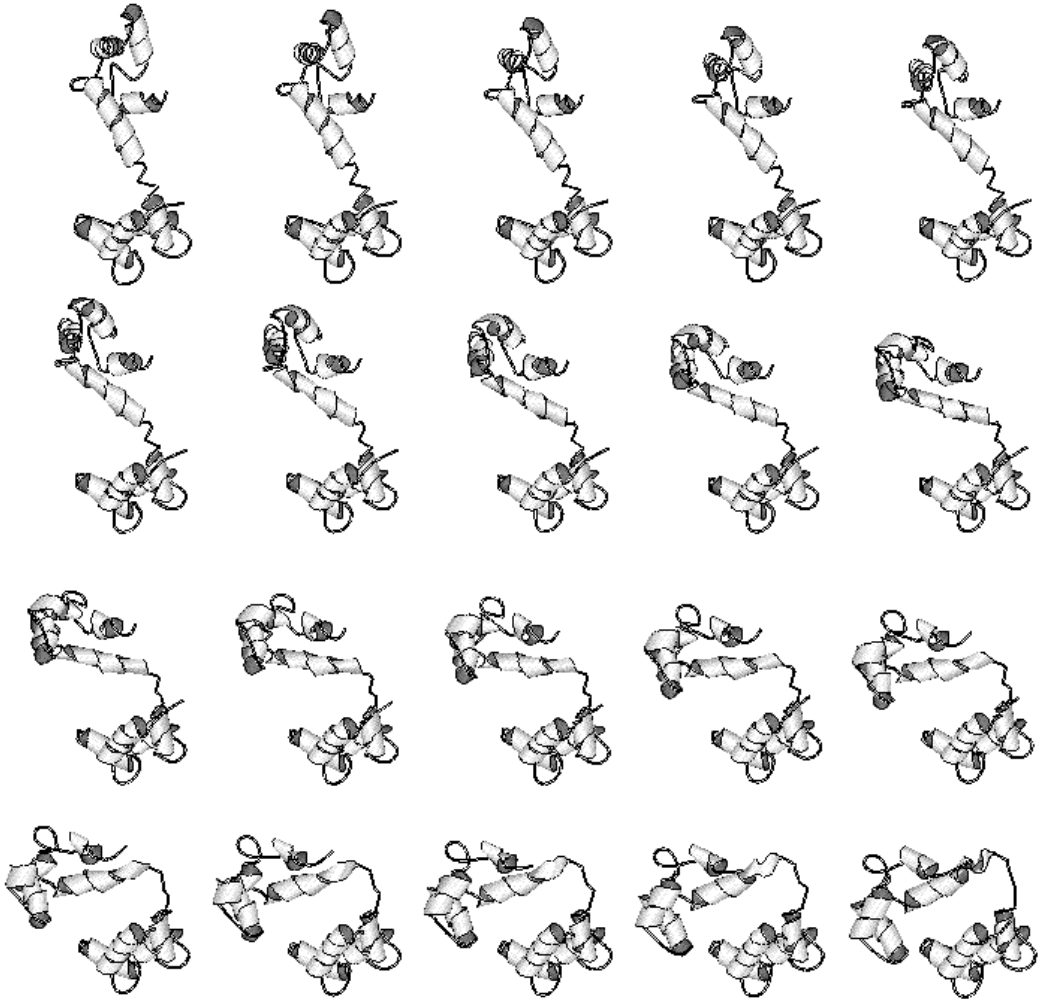
ORF Query



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

“Morph” Movies of Protein Motions





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

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

Acknowledgements: **Ronald Jansen**

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

ONR (Bright)

PhRMA
NSF