Integrative Genomics:

Surveys of a Finite Parts List

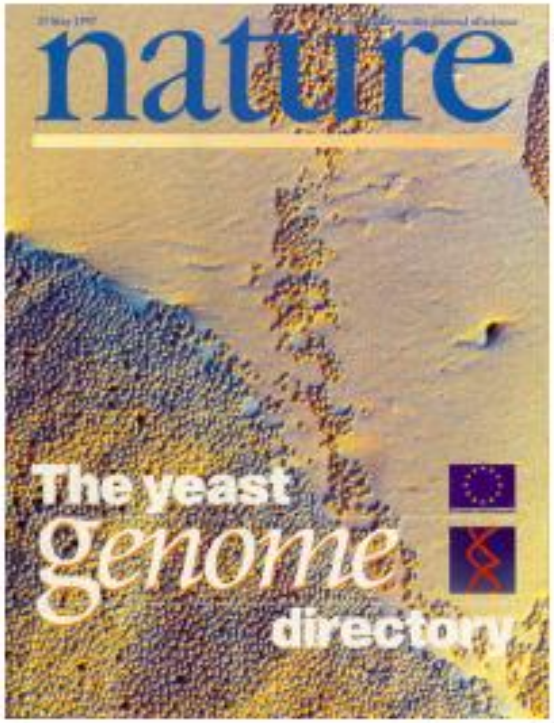
Mark Gerstein

H Hegyi, J Lin, N Echols,
P Harrison, A Drawid, M Levitt, R Jansen, D
Greenbaum, M Snyder, S Teichmann, P Bertone,
B Stenger, V Alexandrov, J Tsai,
C Wilson, J Qian, W Krebs

Talk at Stanford U, 16 April 2001

1995

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



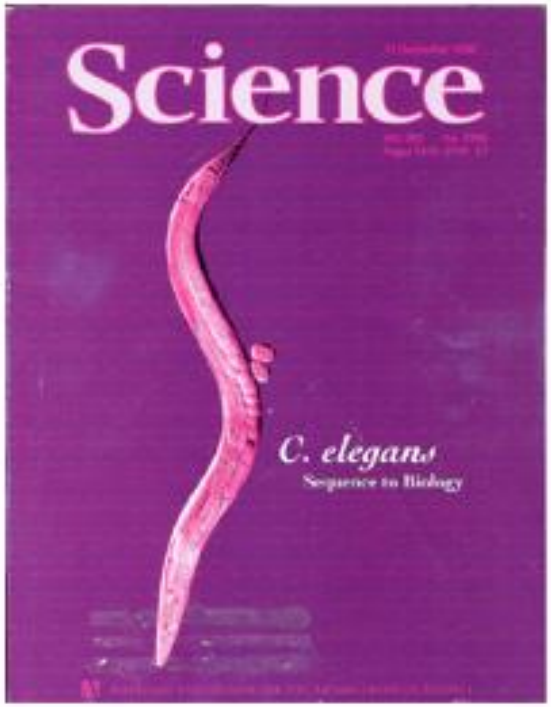
Genomes highlight the **Finiteness** of the "Parts" in Biology

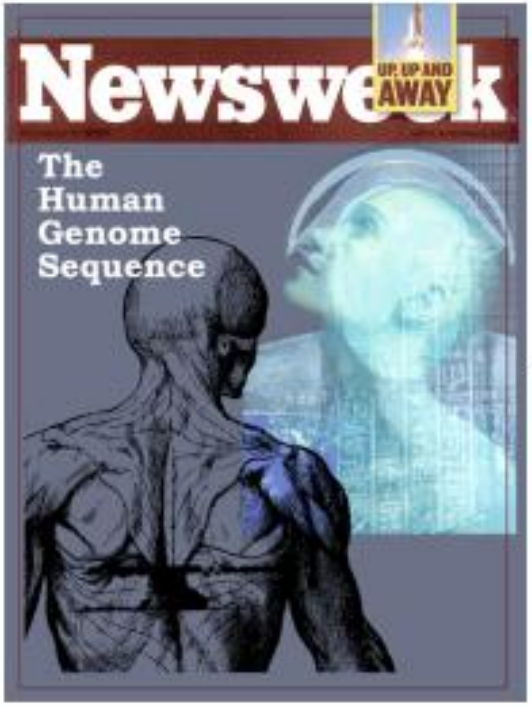
1997

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

1998

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





↔

2000

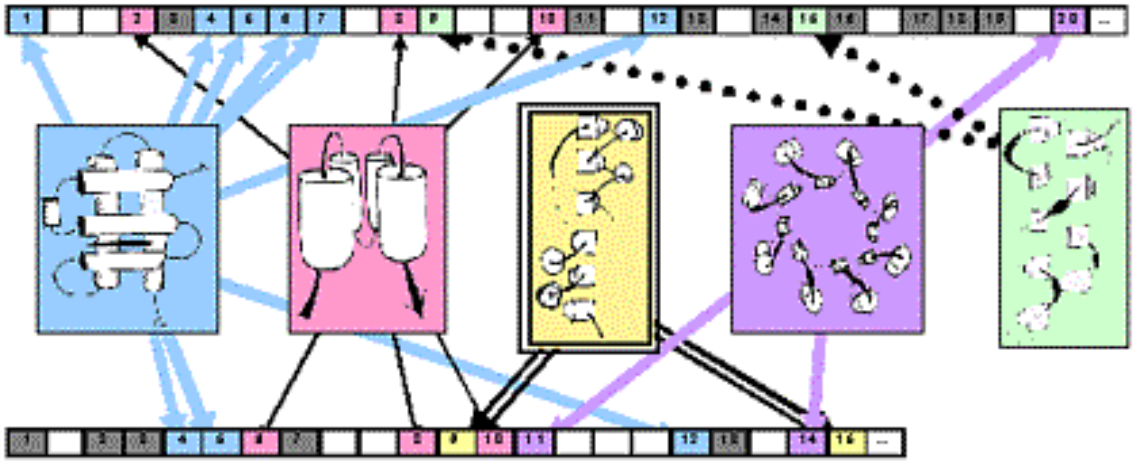
Human,
~3 Gb,
~30K genes

'98 spoof real thing, Apr '00

Global Surveys of a Finite Set of Parts from Many Perspectives

(human)

(*T. pallidum*)

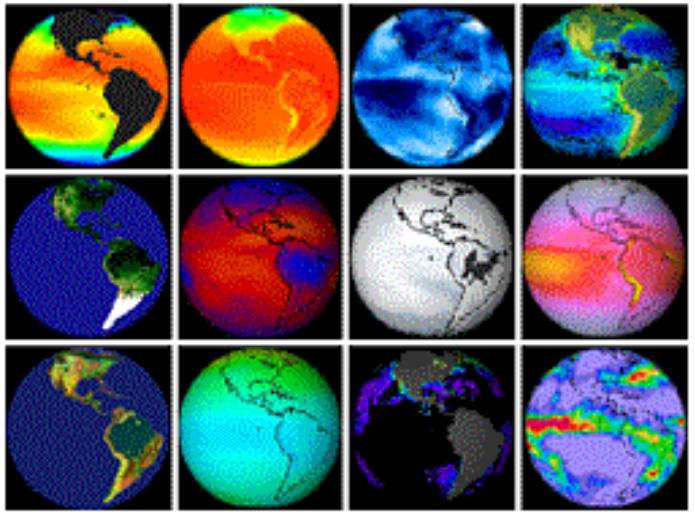


~30000 genes

~1000 folds

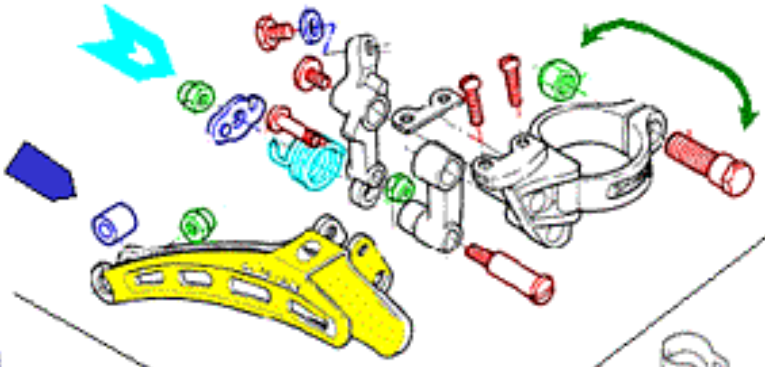
~1000 genes

World of structures is even more finite than that of sequences, providing simplification…

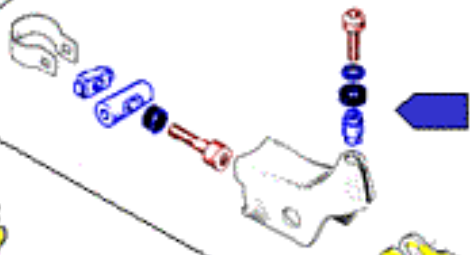


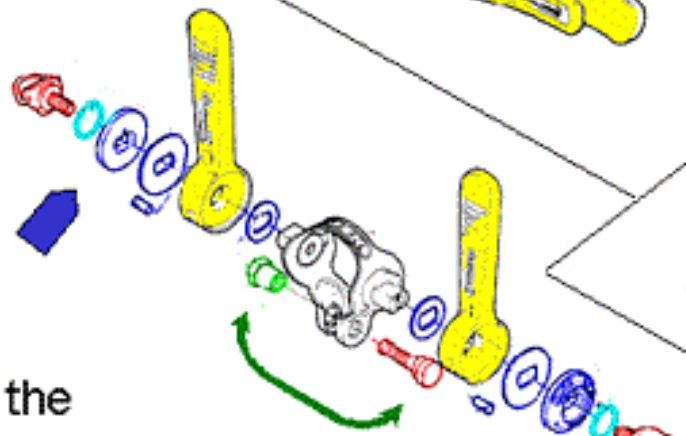
A Parts List Approach to Bike Maintenance

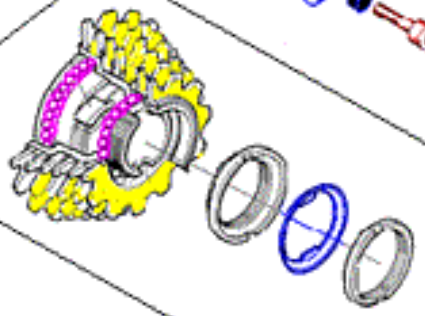
A Parts List Approach to Bike Maintenance

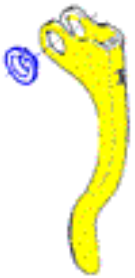


How many roles can these play? How flexible and adaptable are they mechanically?

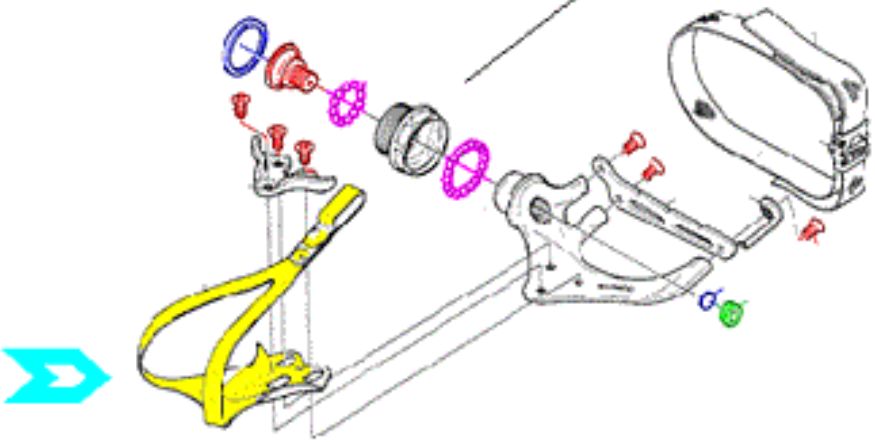






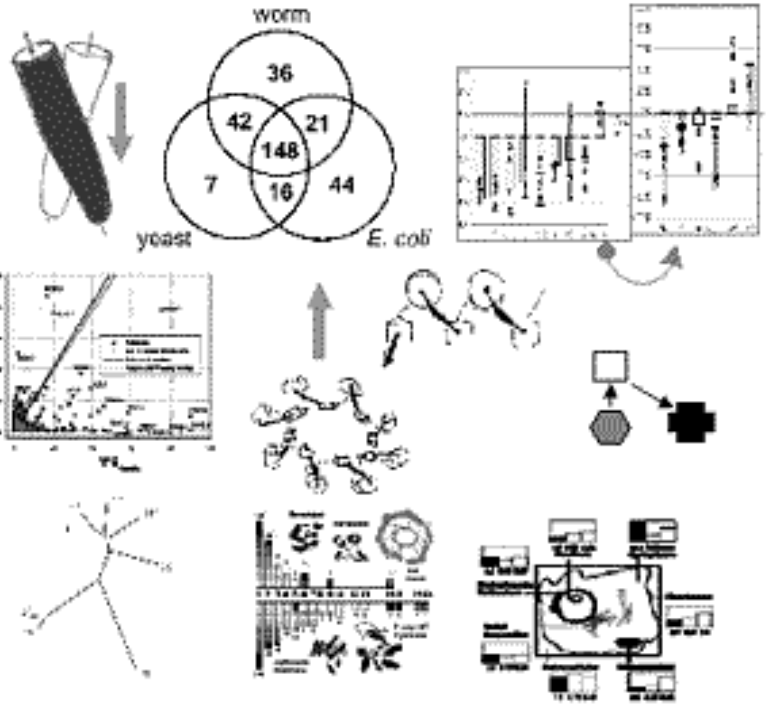


What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?



Where are the parts located? Which parts interact?

Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

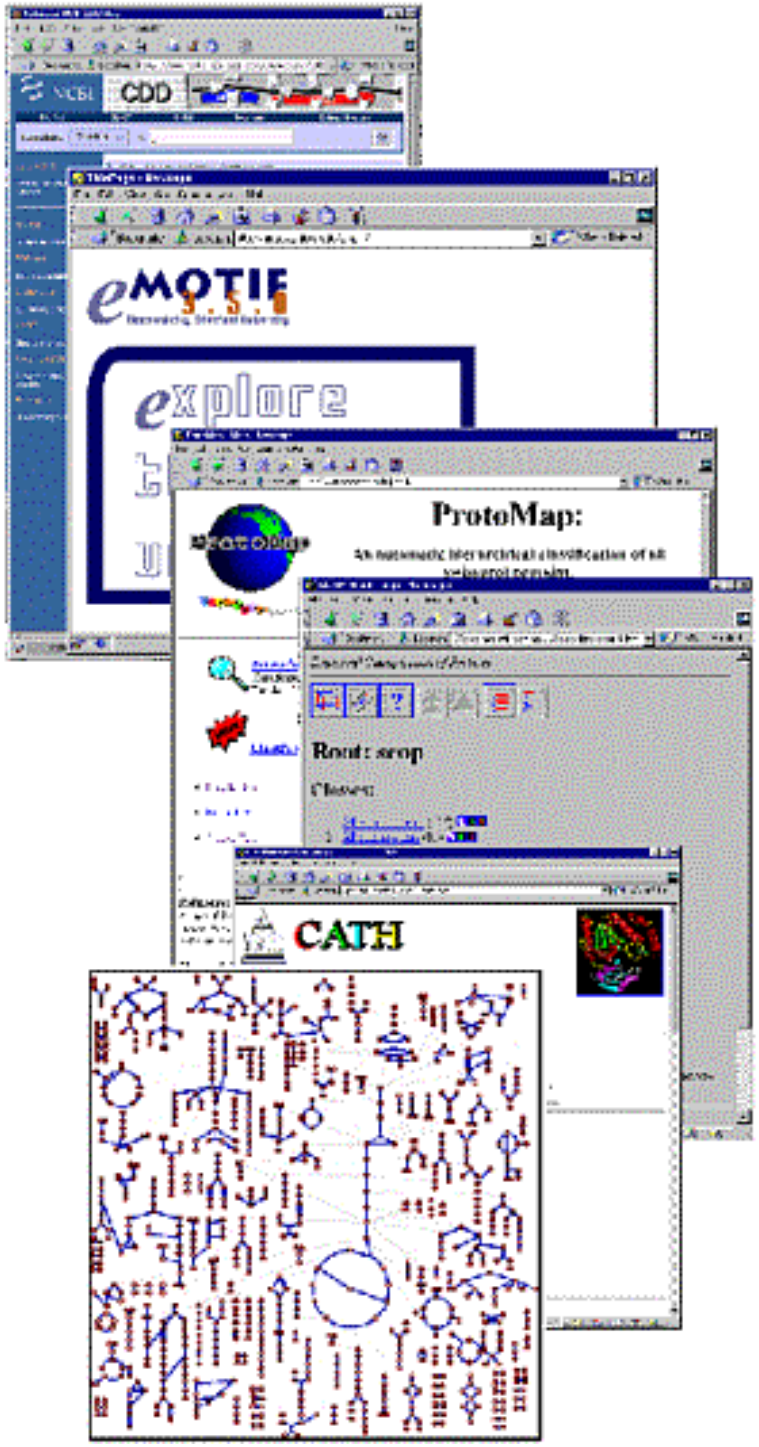
Using Parts & Categories to Mine Expression Data

Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

Integrated Views based on Parts

H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

bioinfo.mbb.yale.edu



What is a Proteomic "Part" or Category?

Result of **grouping together** sequences & structures

Many **non-unique but self-consistent** definitions

Folds: most intuitive & fundamental parts (focus here)

Same logic for seq. families (homologs), orthologs, blocks, motifs.... even pathways, functional systems

Much **previous work...**

(CATH, scop, Blocks, FSSP, Interpro, eMotif, Prosite, CDD, Pfam, ProtoMap, 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) (Functions picture from www.fruitfly.org/~suzi (Ashburner); Pathways picture from, ecocyc.pangeasystems.com/ecocyc (Karp, Riley).)

Folds as the best parts, the final annotation for the human genome. Why?

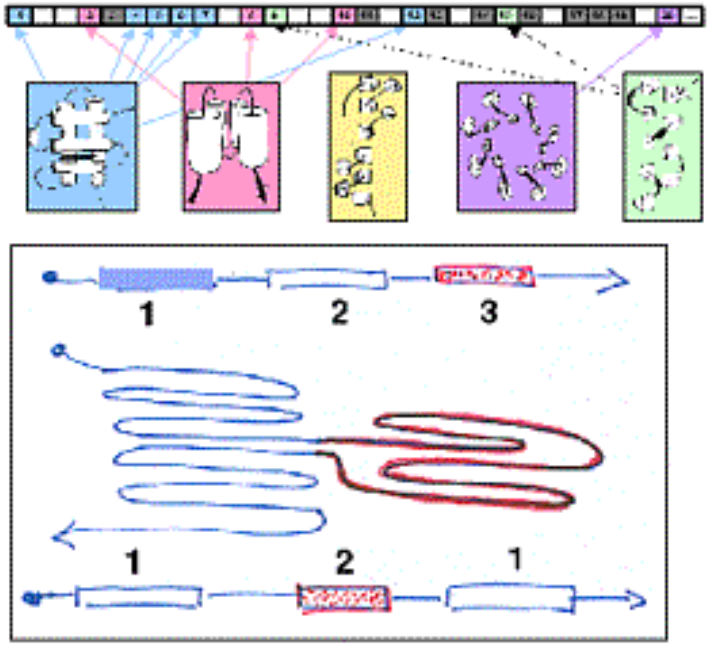
1 Most Highly Conserved

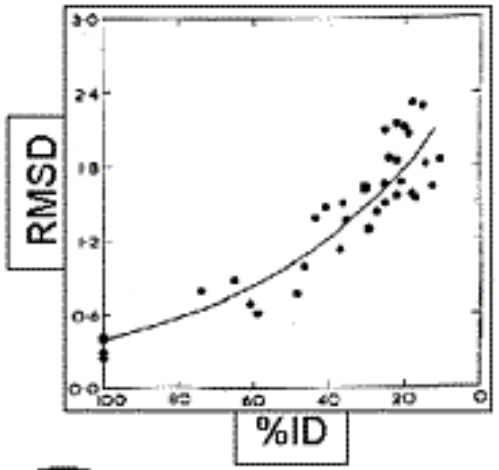
2 Precisely Defined Modules

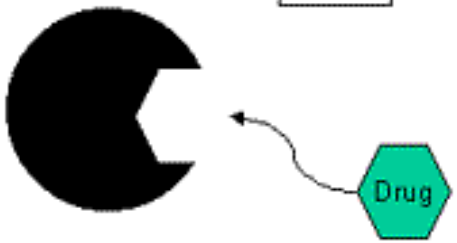
3 Seq. ⇔ Struc. Clearer than Seq. ⇔Func.

4 Link to Chemistry, biochemical function, drugs

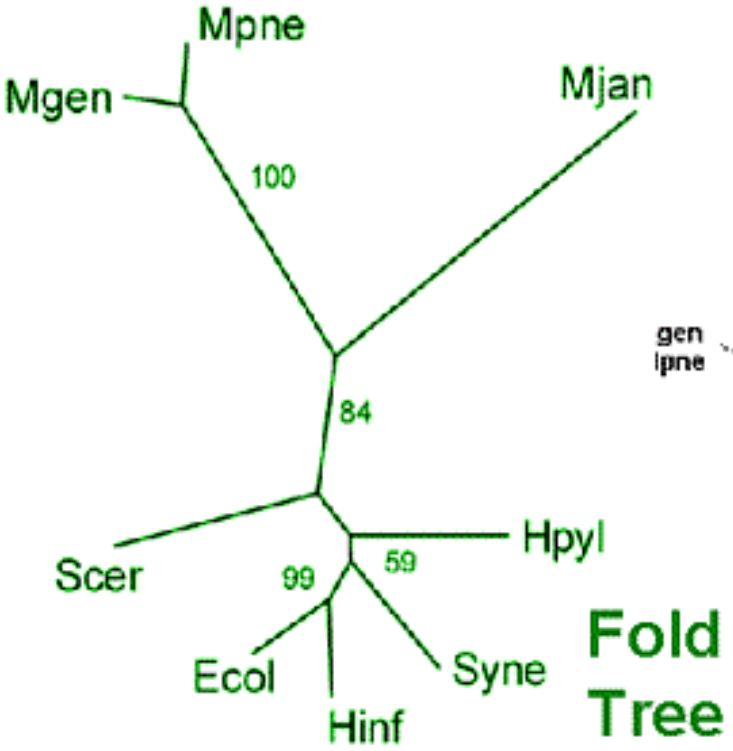
Here mostly
scop+astral+auto-alignments
(Chothia, Levitt, Brenner, Murzin)



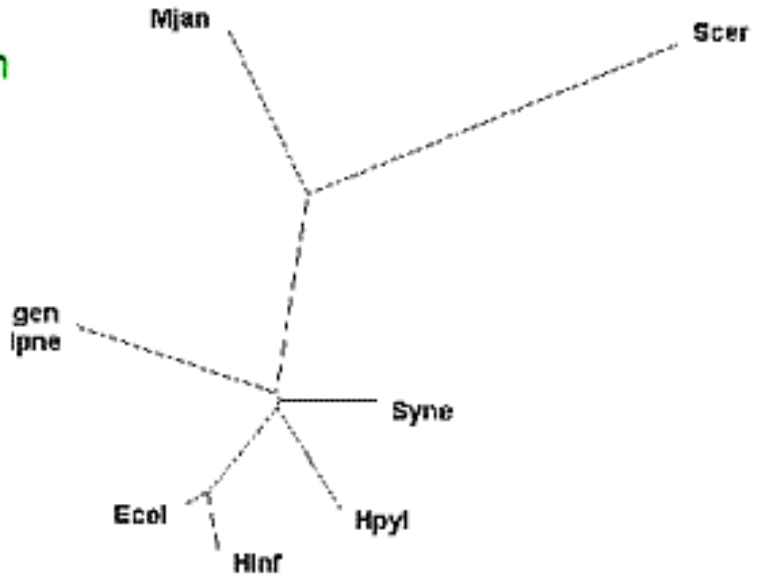




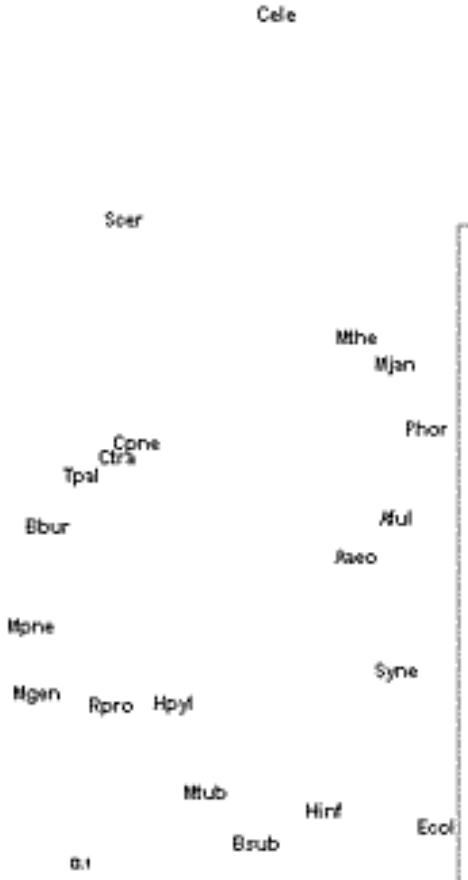
Cluster Trees Grouping Initial Genomes on Basis of Shared Folds

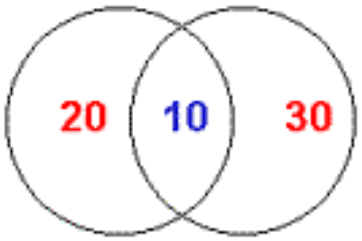


Fold Tree



"Classic" Tree



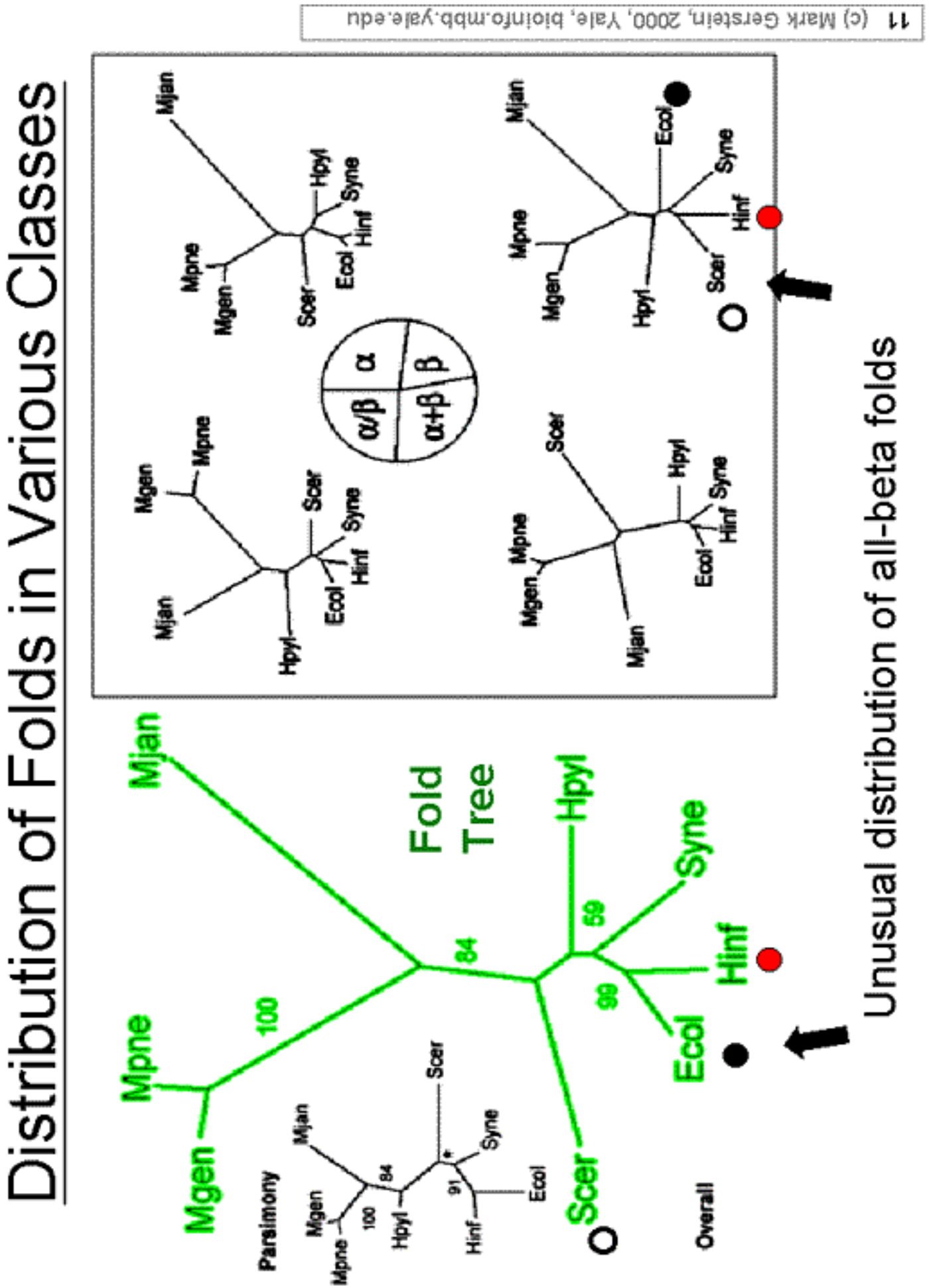


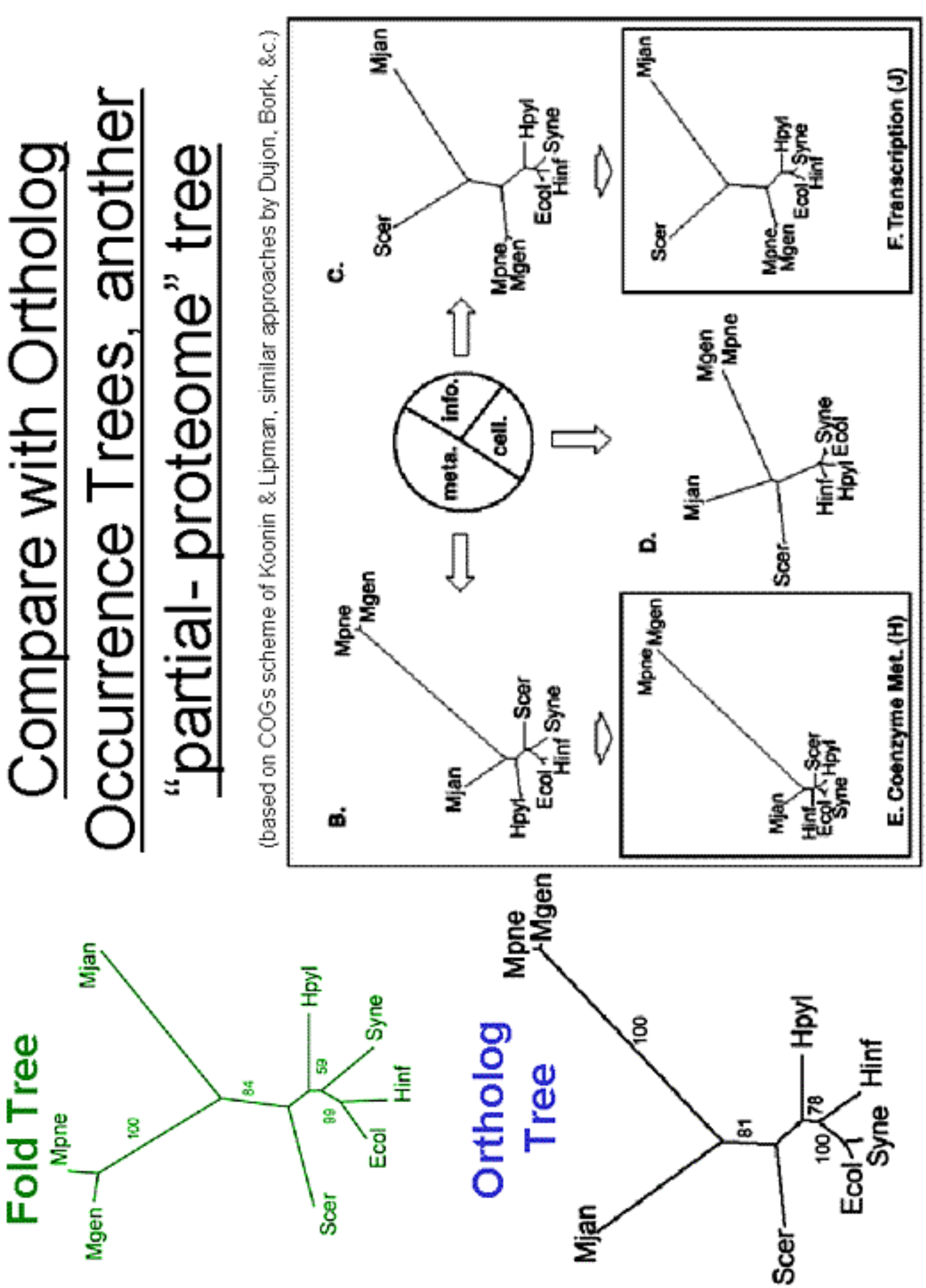
D=10/(20+10+30)

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

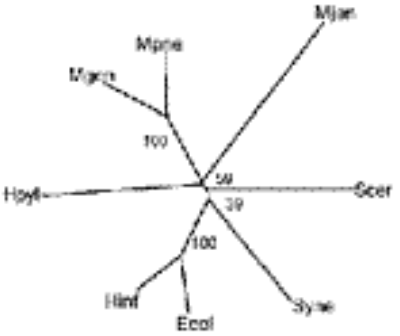
D = shared fold dist. betw. 2 genomes

T= total # folds in both

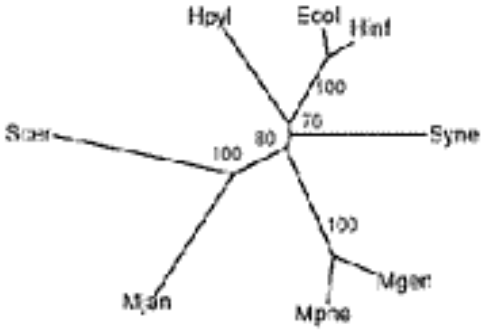

Distribution of Folds in Various Classes
11 (c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu
Mjan
Mpne
Mgen
Scer
Hpyl
Syne
Hinf
Ecol
α
β
α/β
α+β
Fold Tree
Parsimony
Overall
100
84
99
91
Unusual distribution of all-beta folds


Compare with Ortholog Occurrence Trees, another "partial- proteome" tree
(based on COGs scheme of Koonin & Lipman, similar approaches by Dujon, Bork, &c.)
Fold Tree
Mpne
Mgen
Mjan
Scer
Hpyl
Ecol
Syne
Hinf
100
84
59
89
Orthology Tree
Mpne
Mgen
Mjan
Scer
Hpyl
Ecol
Syne
Hinf
100
81
78
100
B.
C.
D.
E. Coenzyme Met. (H)
F. Transcription (J)
meta.
info.
cell.

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

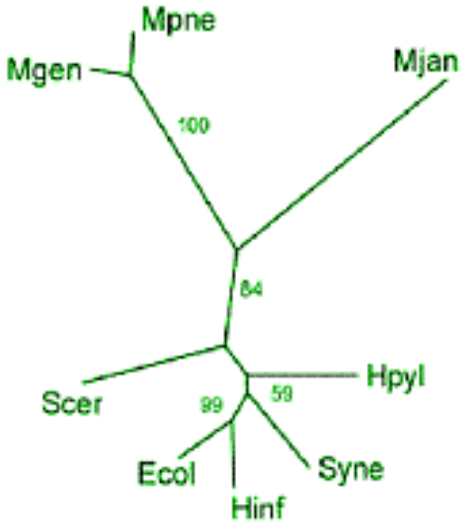


TIM

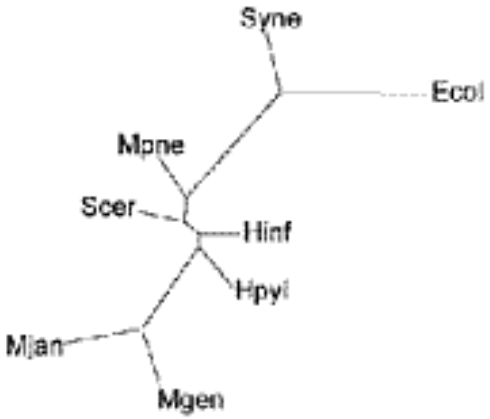


ribosomal protein

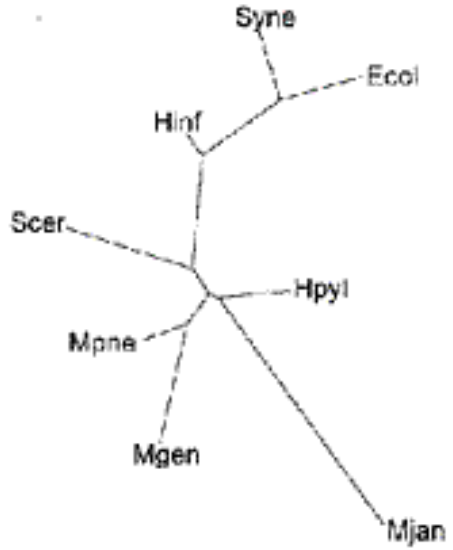
Single-gene Trees



Fold Tree



A. Di-Nucleotide



B. Amino Acid

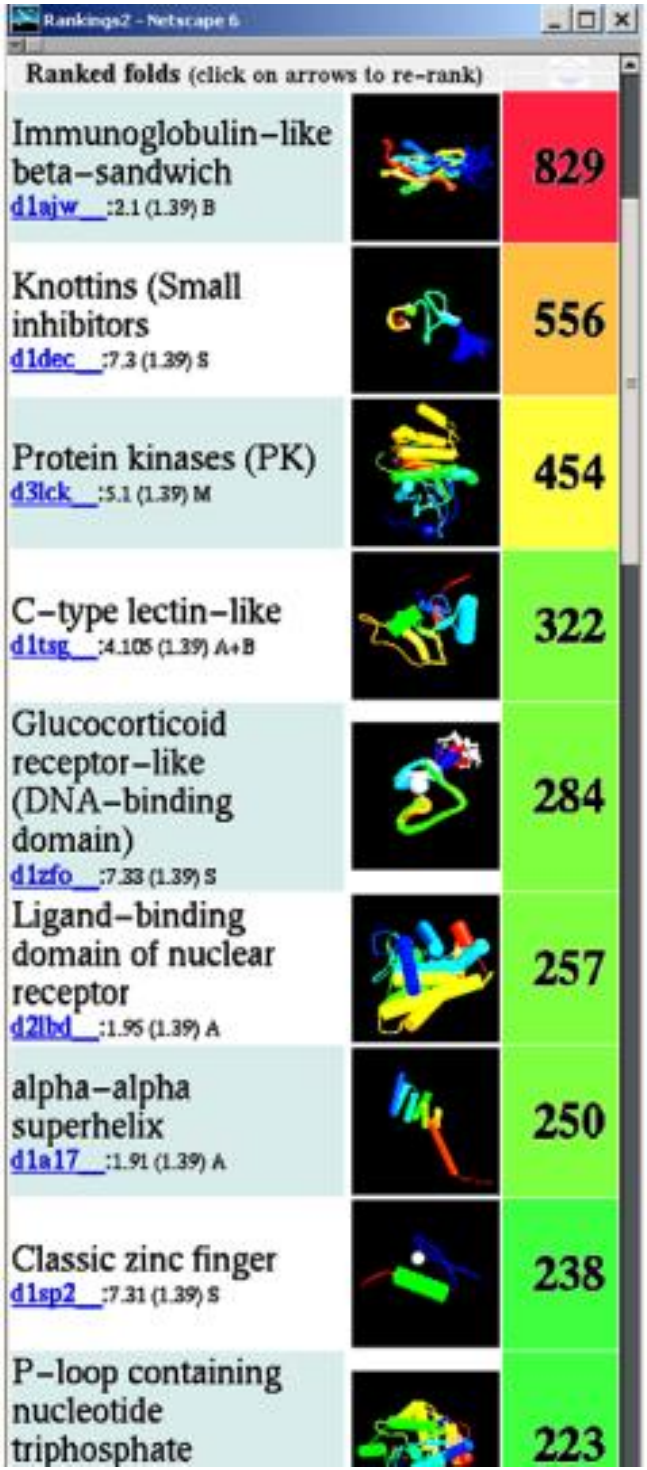
Ortholog Tree

"Classic" Tree

AA & di-NT Composition Trees

(S Karlin)

Common Folds



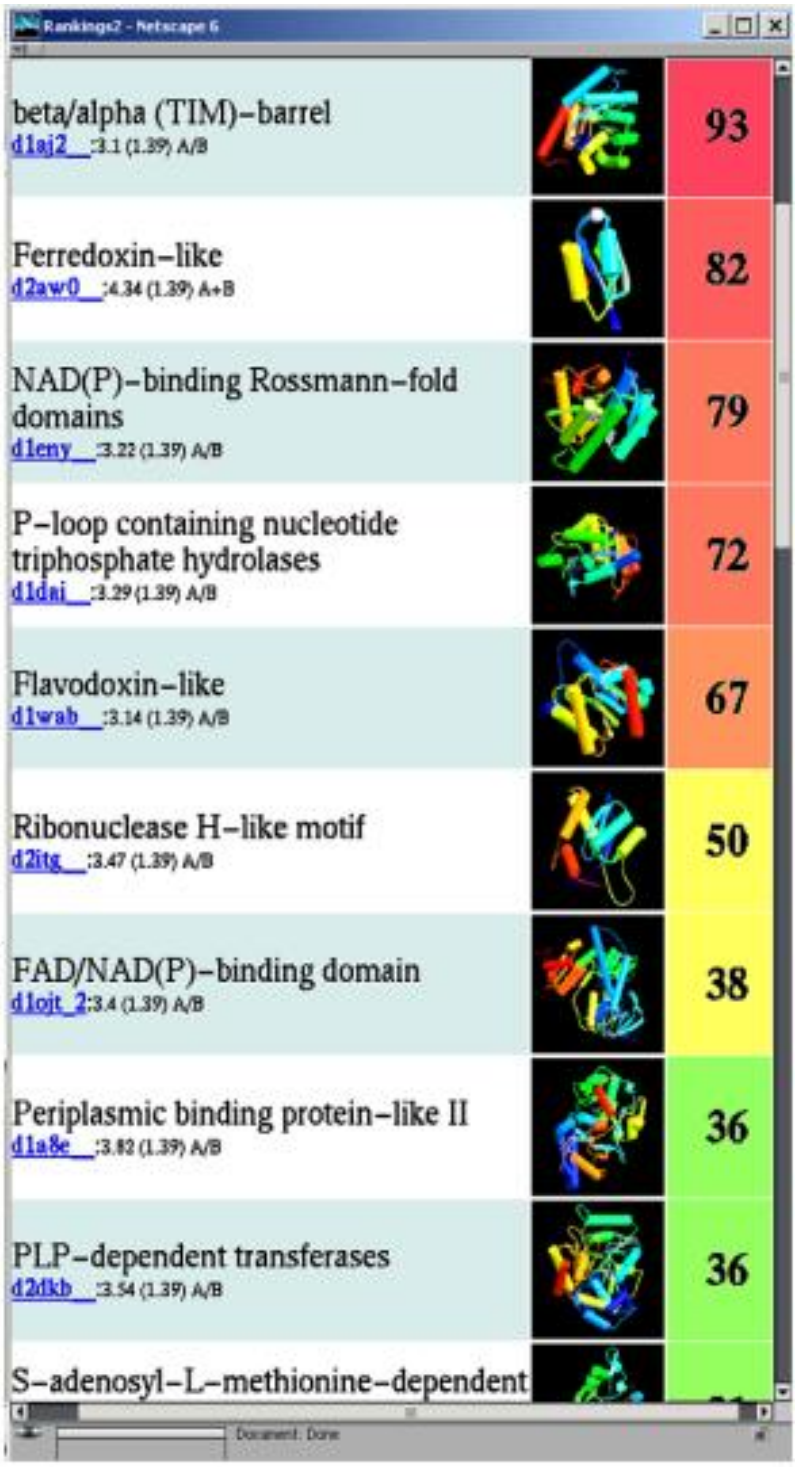
⇦ **Worm**

Extracellular, signaling and trans. factor folds (Ig, kinase, lectin, nuc. receptor, ZnF)

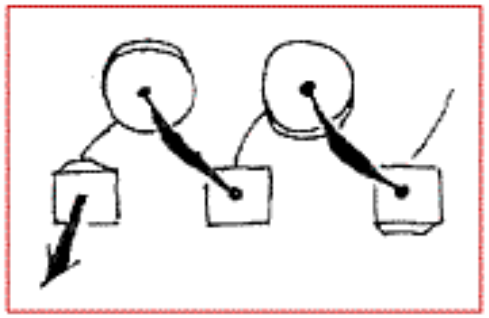
E. coli ⇨

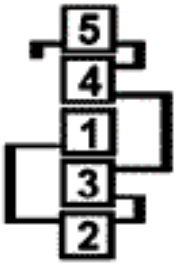
Metabolic Folds (TIM, Ferrodoxin, Rossman, P-loop hydrolase, FAD-binding)

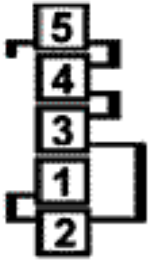
(partslist.org)

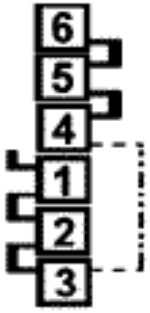


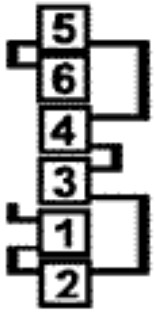
Common, Shared Folds: βαβ structure



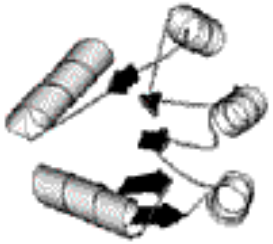






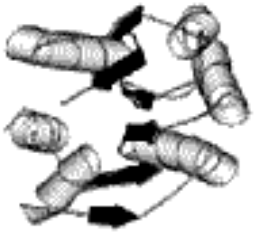


P-loop hydrolase



Flavodoxin like

Rossmann Fold



Thiamin Binding

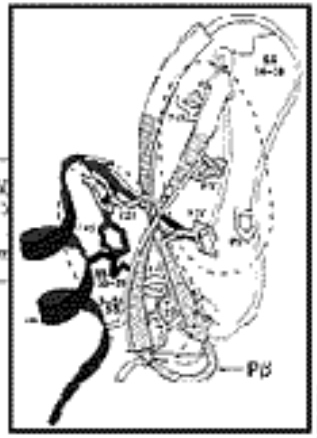
TIM-barrel

A peptide model of a protein folding intermediate

Terrence G. Oas & Peter S. Kim

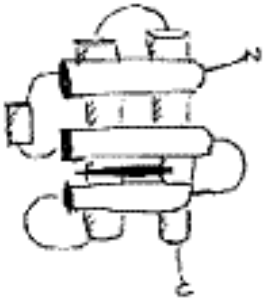
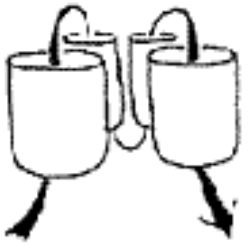
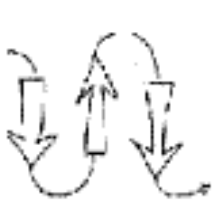
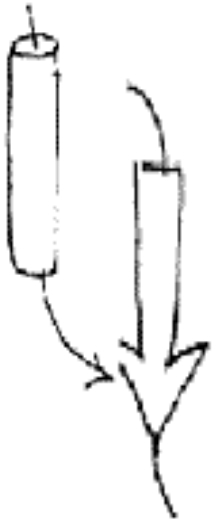
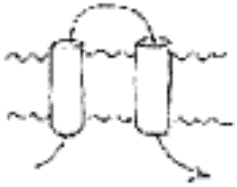
Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

336: 42



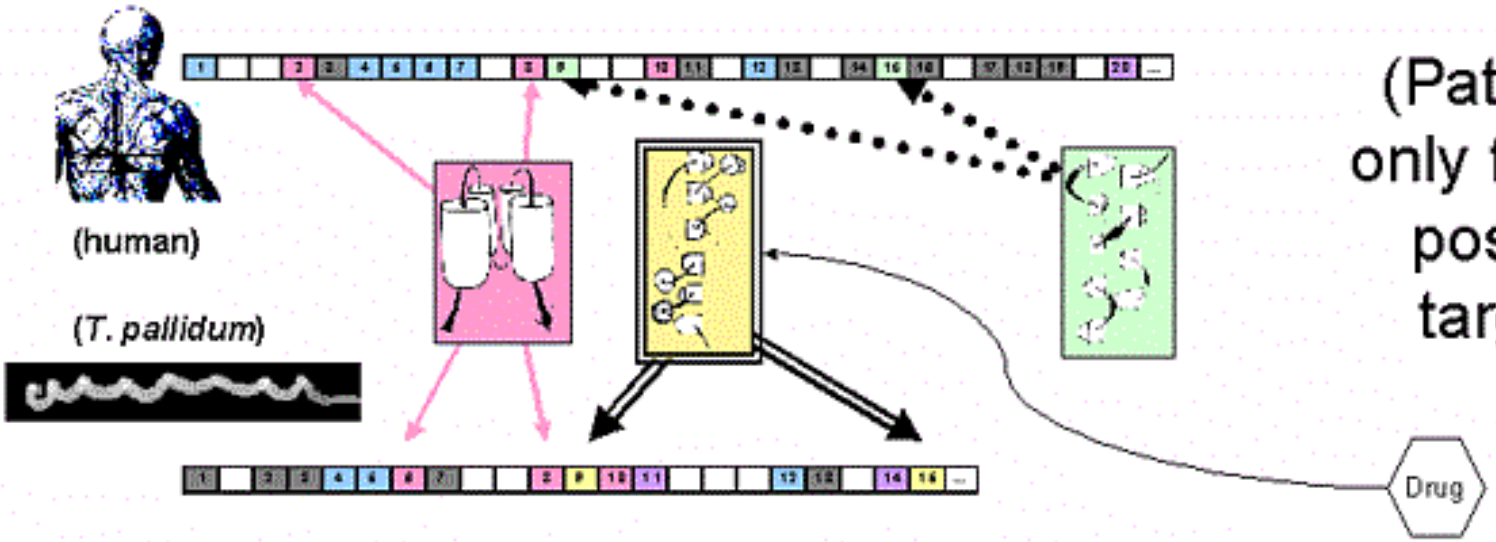
HI, MJ, SC vs scop 1.32. All share α/β structure with repeated R.H. βαβ units connecting adjacent strands or nearly so (18+4+2 of 24)

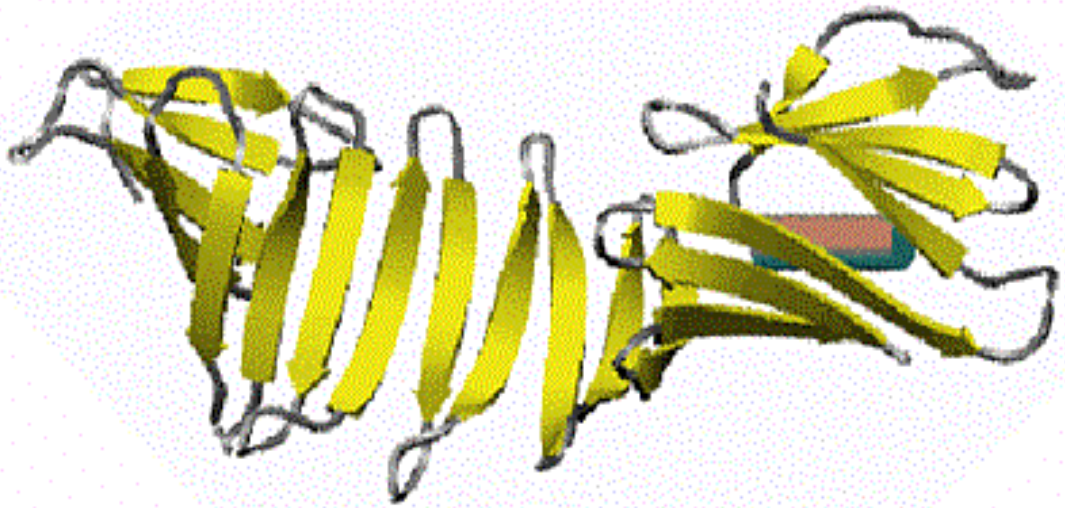
At What Structural Resolution Are Organisms Different?

person plant	protein fold (lg)	super-secondary structure ($\beta\beta$,TM–TM, $\alpha\beta\alpha\beta$,$\alpha\alpha\alpha$)	helix strand	individual atom (C,H,O...)
					
					

1m 100Å 10Å 1Å

Practical Relevance of Structural Genomics

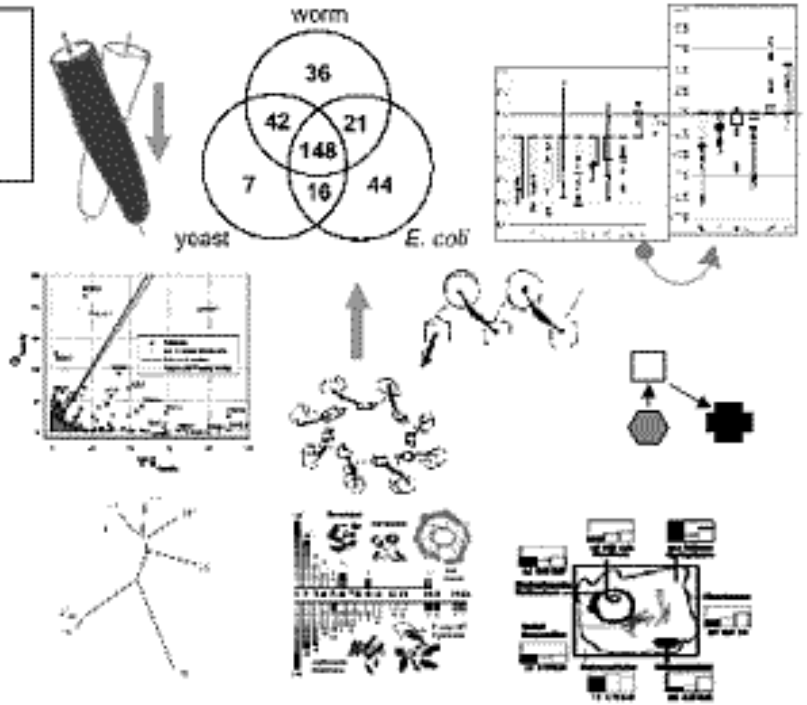




OspA protein

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

Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

Using Parts & Categories to Mine Expression Data

Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

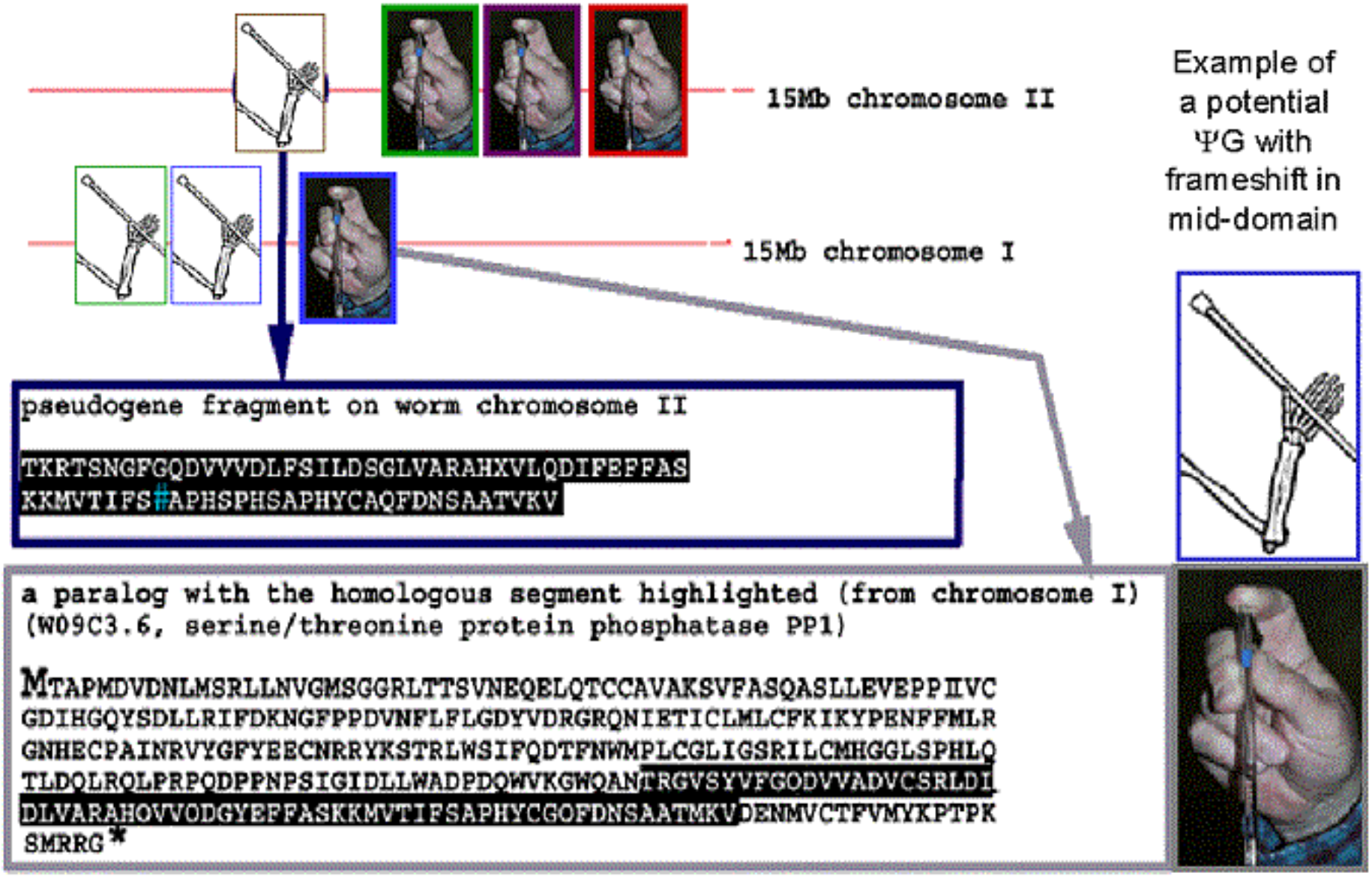
Integrated Views based on Parts

H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

bioinfo.mbb.yale.edu

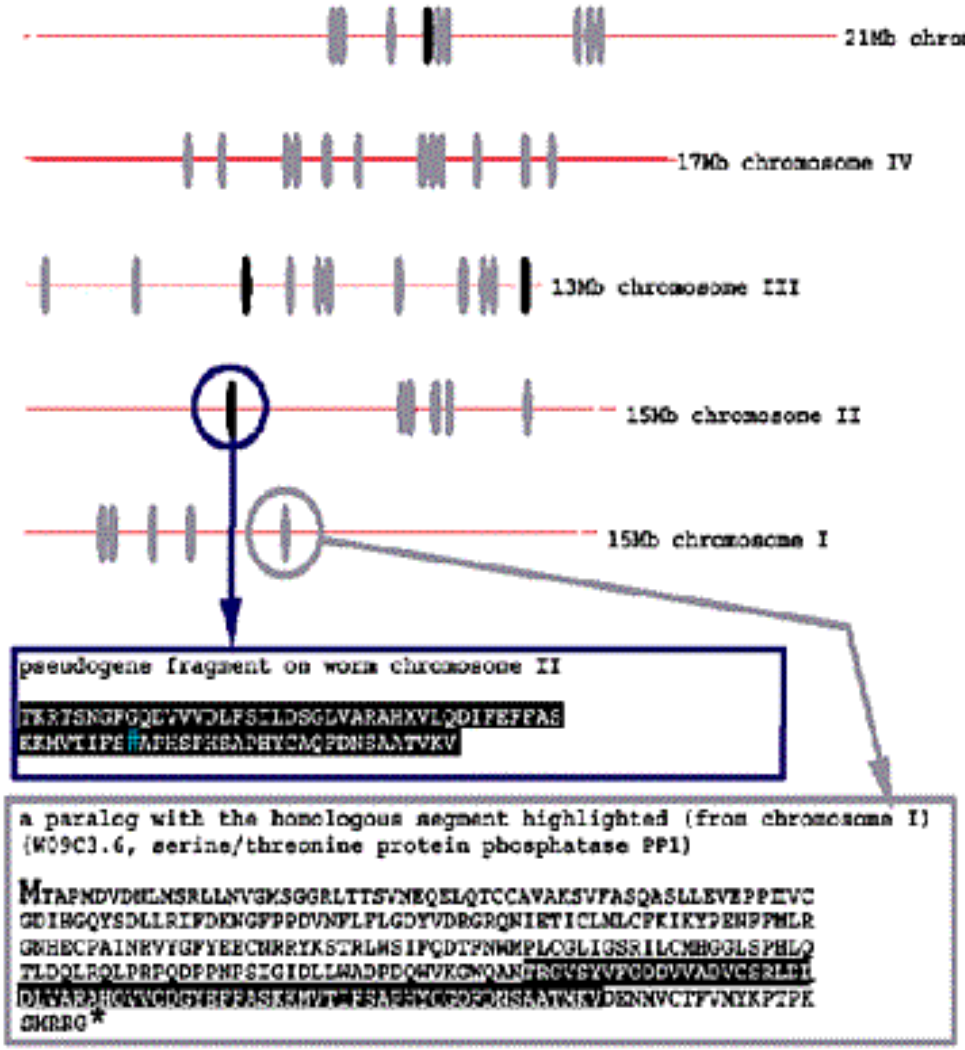
Pseudogenomics: Surveying "Dead" Parts
15Mb chromosome II
15Mb chromosome I
MTAPMDVDNLMSRLLNVGMSGGRLTTSVNEQELQTCCAVAKSVFASQASLLEVEPPIIVC
GDIHGQYSDLLRIFDKNGFPPDVNFLFLGDYVDRGRQNIETICLMLCFKIKYPENFFMLR
GNHECPAINRVYGFYEECNRRYKSTRLWSIFQDTFNWMPLCGLIGSRILCMHGGLSPHLQ
TLDQLRQLPRPQDPPNPSIGIDLLWADPDQWVKGWQANTRGVSYVFGQDVVADVCSRLDI
DLVARAHQVVQDGYEFFASKKMVTIFSAPHYCGQFDNSAATMKVDENMVCTFVMYKPTPK
SMRRG*
19 (c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

Pseudogenomics: Surveying "Dead" Parts



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

Folds in Pseudogenes



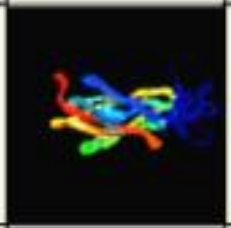
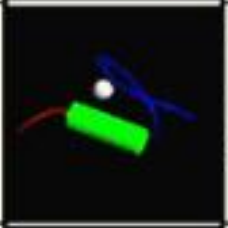
Example of a potential ΨG with frameshift in mid-domain

ΨG identification pipeline to Summary of Pseudogenes in worm

G=19K G_E=8K ΨG=4K (2K)

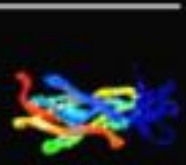
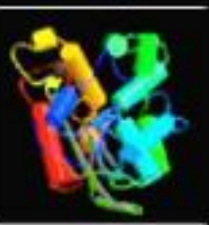
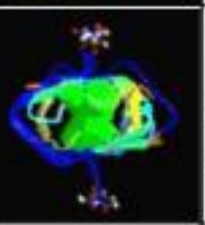
	Category	Total number	Number for genes with EST match	Genes with EST match as percentage of *Category*	Number for genes in paralog families with EST match	Genes in paralog families with EST match as percentage *of Category*
Genes	Total	18,576 (G)	7,829 (G_E)	42%	13,417 (G_P)	72%
	Singletons	5,913	2,788	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

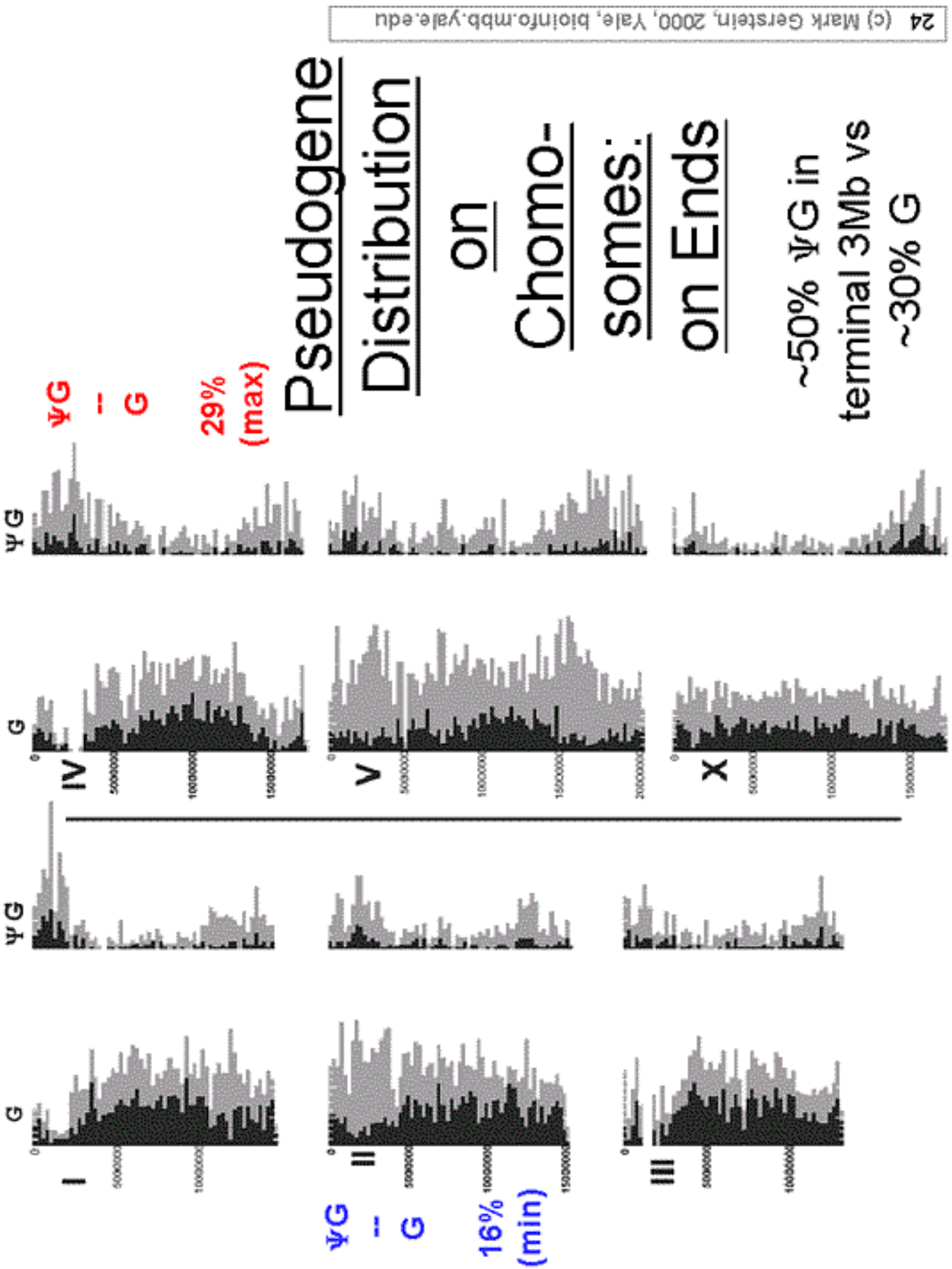
G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
1 (769)	2		d1ajw__ 2.1 Immuno-globulin	6 (246)	8		d2lbd__ 1.95 Nuc. receptor ligand-binding domain
2 (555)	6		d1dec__ 7.3 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

[22] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

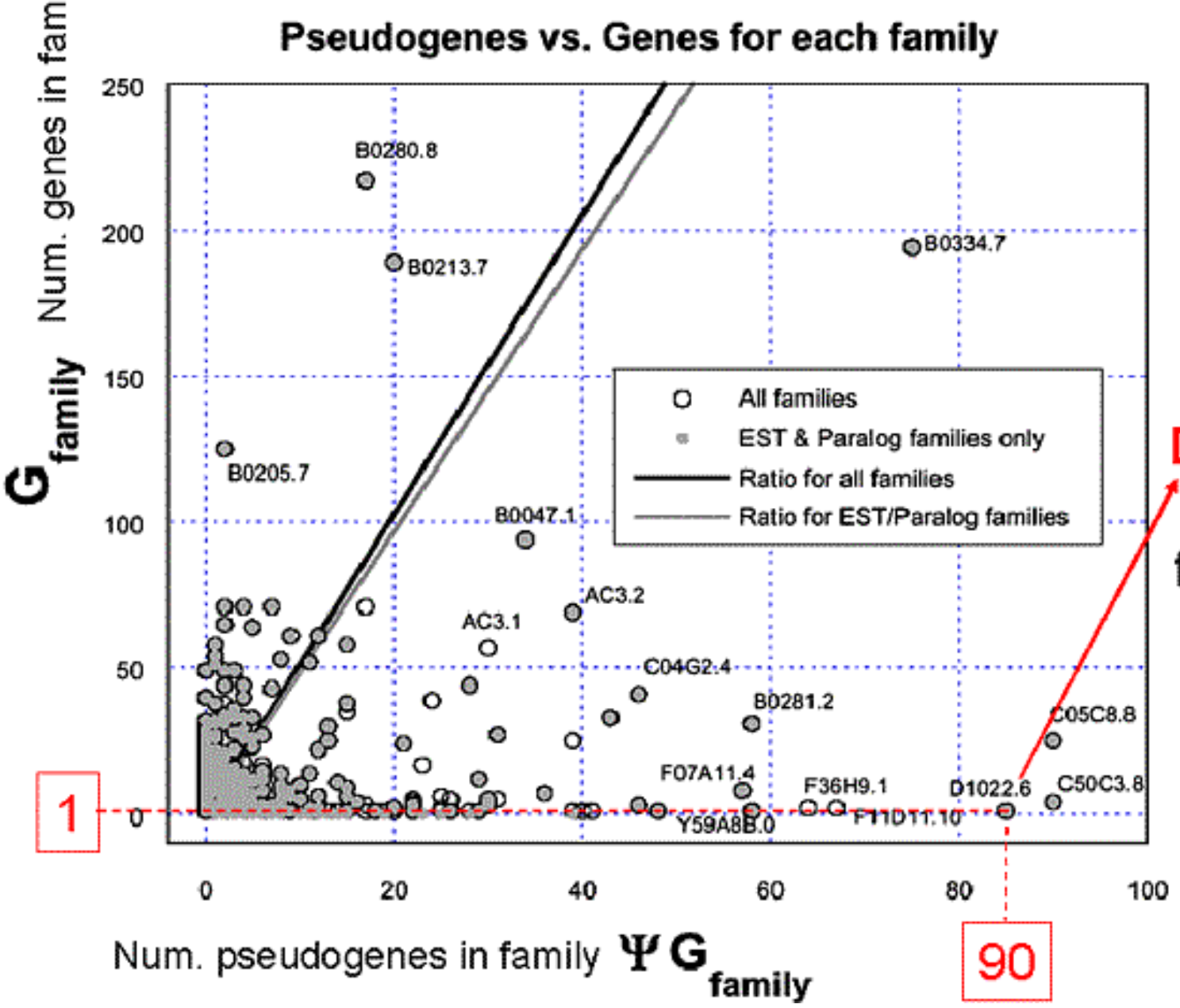
Most Common Worm "Pseudofolds" #2

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

[23] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List


24 (c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu
Pseudogene Distribution on Chromosomes: on Ends
~50% ΨG in terminal 3Mb vs ~30% G
ΨG - G
29% (max)
ΨG - G
16% (min)
ΨG
G
I
II
III
IV
V
X

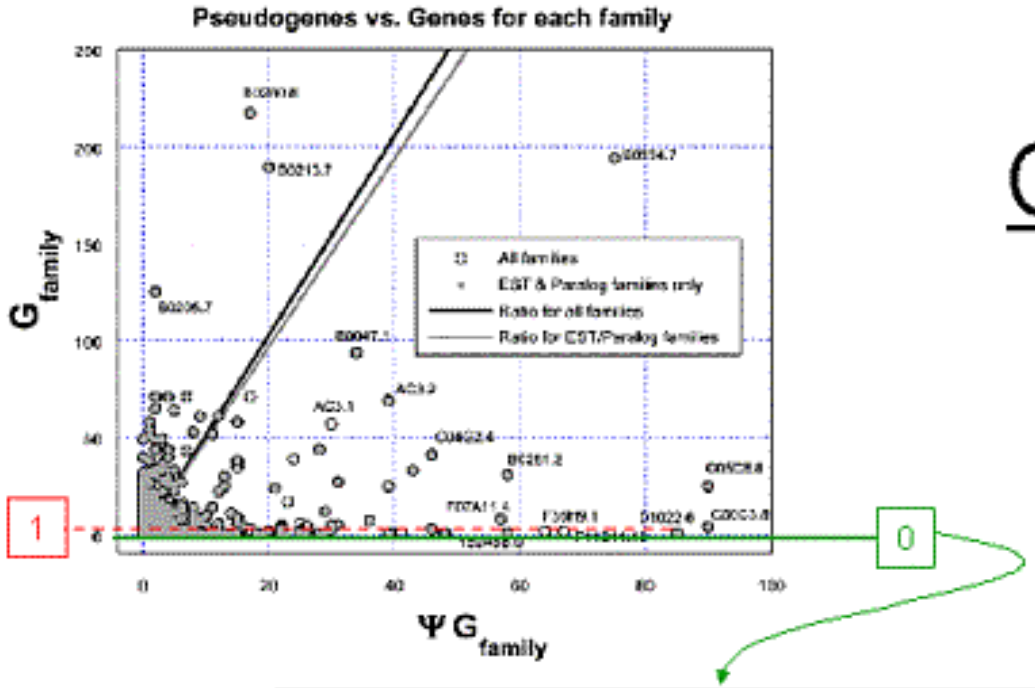
Decayed Lines of Genes?



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

[25] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

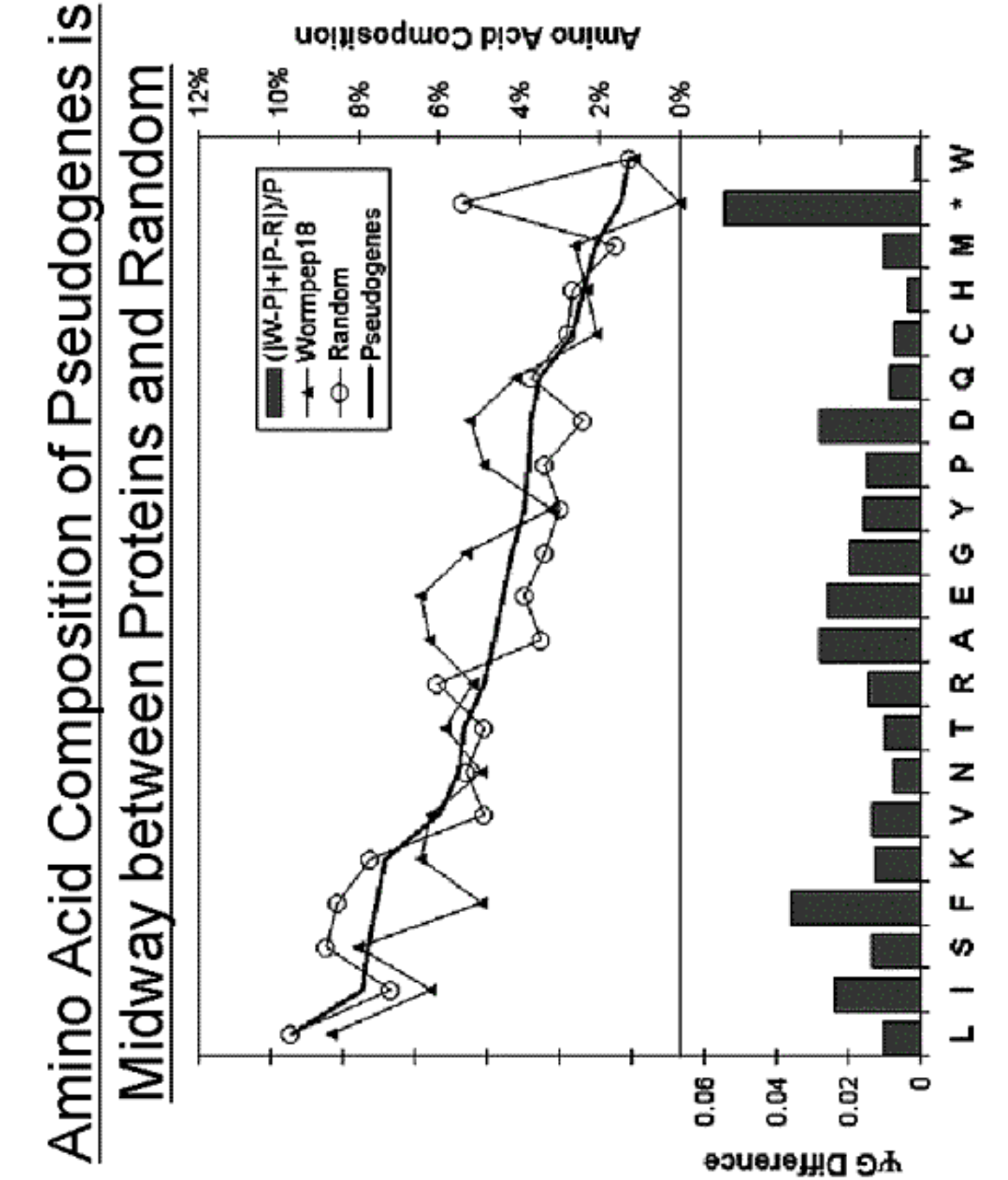
Completely Dead Families



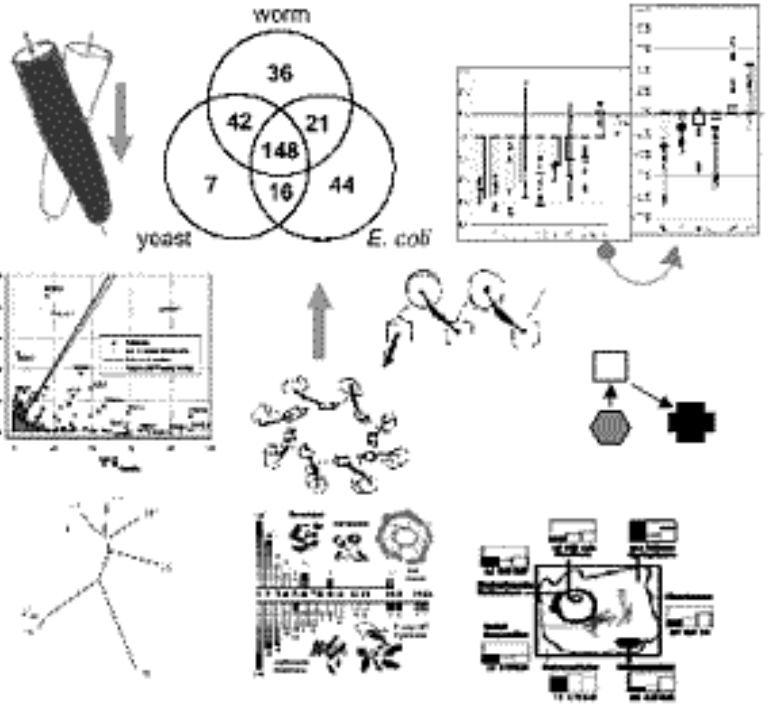
Rank	Number matches	Organism of closest match*	PROTOMAP family representative	Notes on representative
#1	7 *******	Yeast	YJA7_YEAST	Hypothetical protein in yeast
#2 =	5 *****	Human	XPD_MOUSE	Xeroderma pigmentosum group D complementing protein
#2 =	5 *****	Cow	CPSA_BOVIN	Cleavage and polyadenylation specificity 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

[26] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

Amino Acid Composition of Pseudogenes is Midway between Proteins and Random



Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

Using Parts & Categories to Mine Expression Data

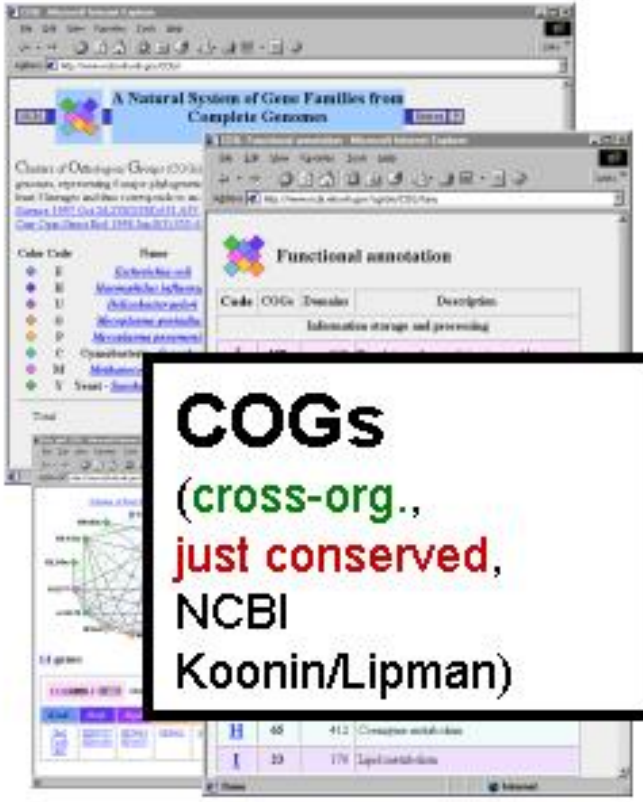
Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

Integrated Views based on Parts

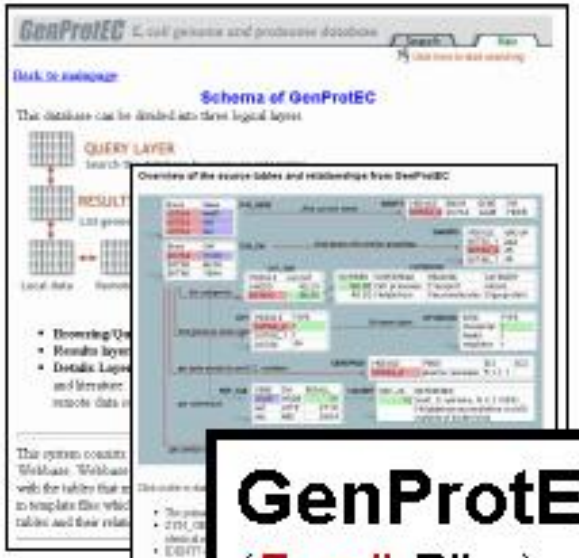
H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

bioinfo.mbb.yale.edu

Functional Classification



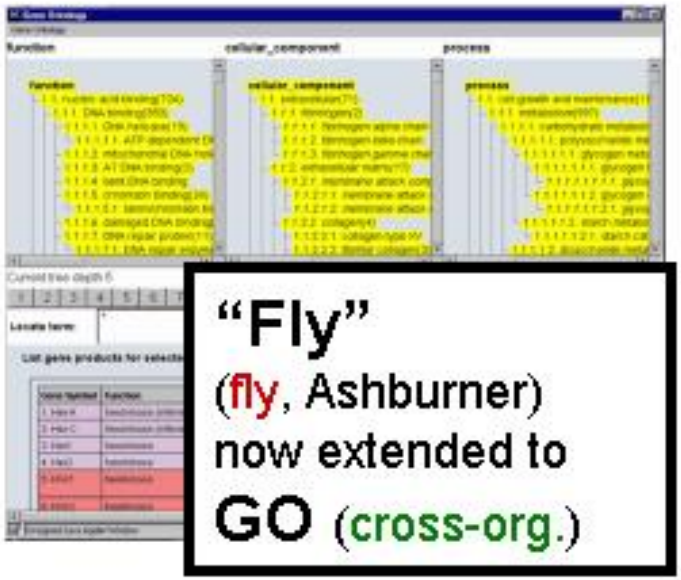
COGs
(cross-org.,
just conserved,
NCBI
Koonin/Lipman)



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)

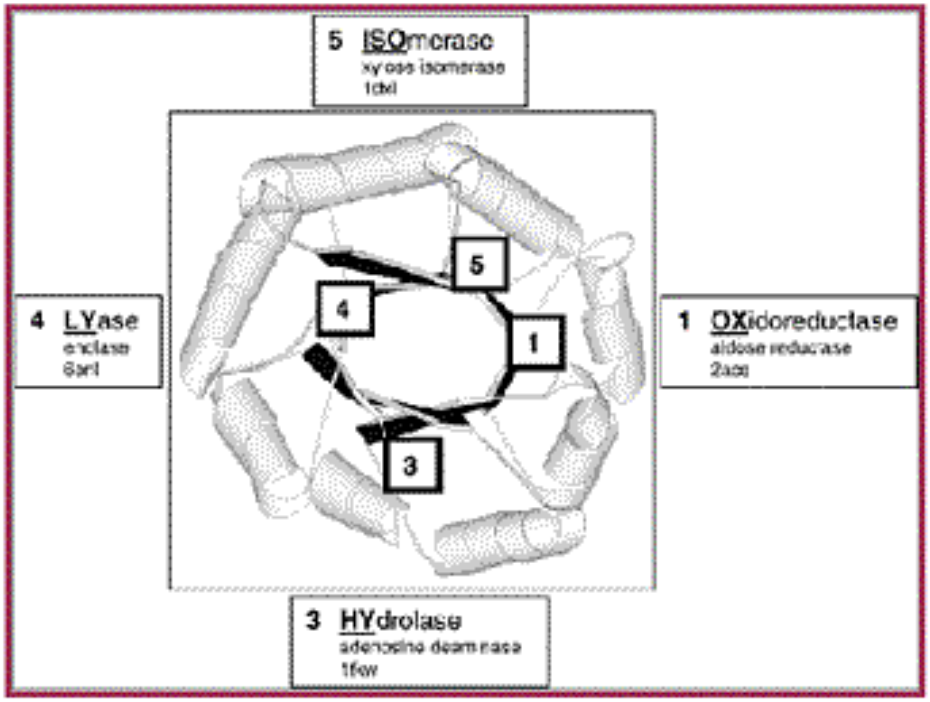
TIGR EGAD
(human ESTs)

SGD (yeast)

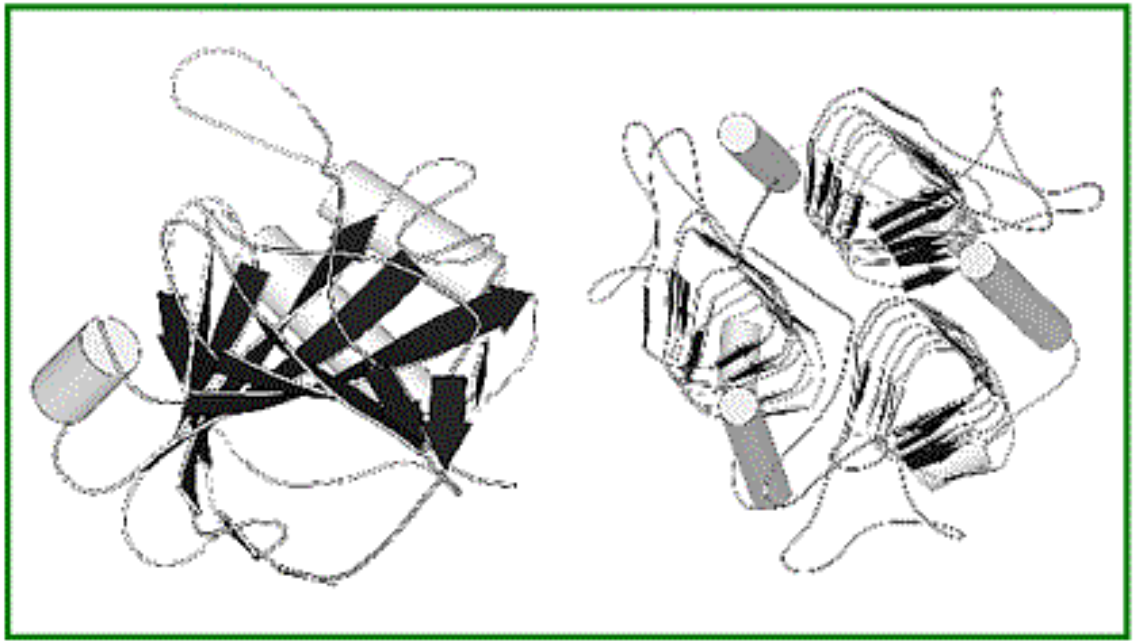
[29] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

Fold-Function Combinations

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



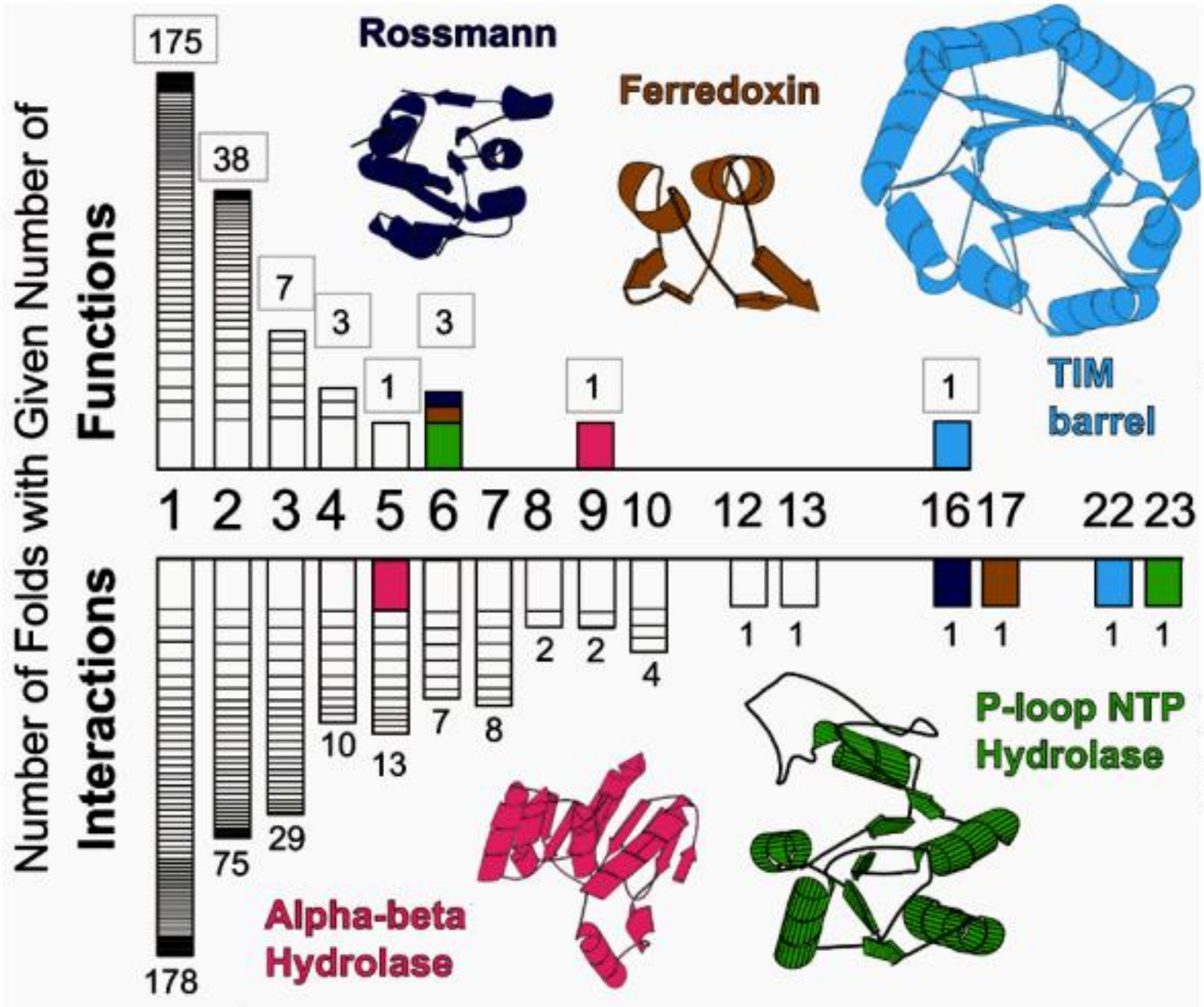
Two Different Folds Catalyze the Same Reaction -- e.g. Carbonic Anhydrases (4.2.1.1)



Fold-Function Combinations
~20K (=92x229) Possible,
331 Observed
229 Folds
91 Enzymatic Functions
+ Non-Enzyme
NONENZ
OX
TRAN
HYD
LY
ISO
LIG
A
B
A/B
A+B
MULTI
SML
31 (c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

Most Versatile Folds
Rossmann
Ferredoxin
TIM barrel
P-loop NTP Hydrolase
Alpha-beta Hydrolase
Similar results Martin et al. (1998)
Number of Folds with Given Number of Functions
1 2 3 4 5 6 7 8 9 10 12 13 16 17 22 23
175
38
7
3
3
1
1
1
1
32 (c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

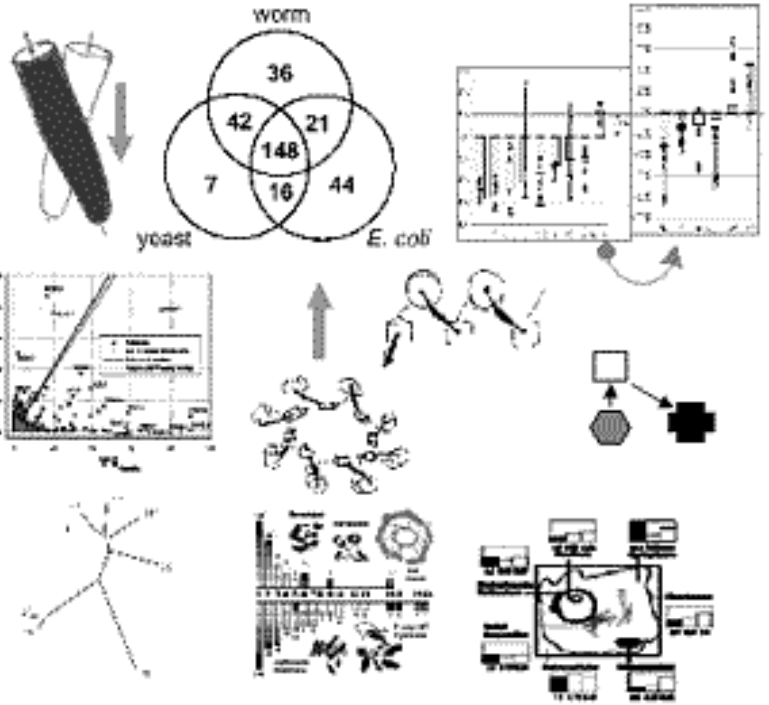
Most Versatile Folds – Relation to Interactions



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

Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

Using Parts & Categories to Mine Expression Data

Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

Integrated Views based on Parts

H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

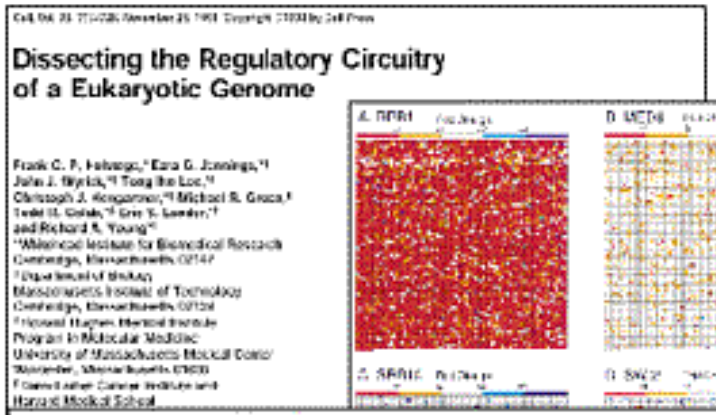
bioinfo.mbb.yale.edu

Expression Datasets

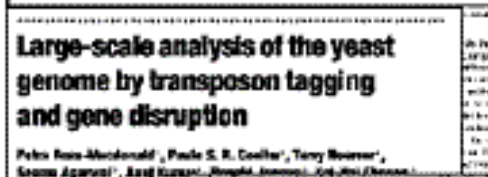
Yeast Expression Data

Individual data points are **noisy but average with parts** and categories to simplify

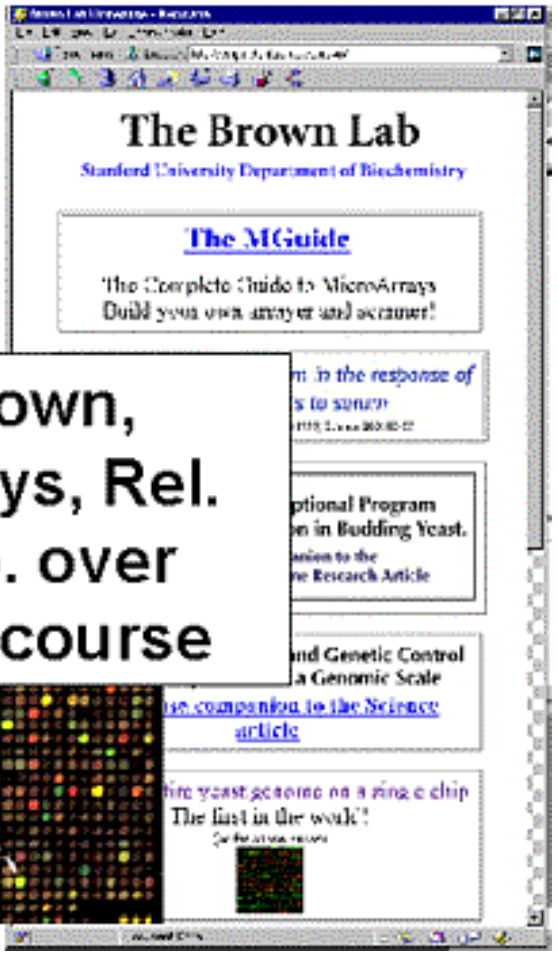
Also:
SAGE (mRNA); 2D gels for Protein Abundance (Aebersold, Futcher)

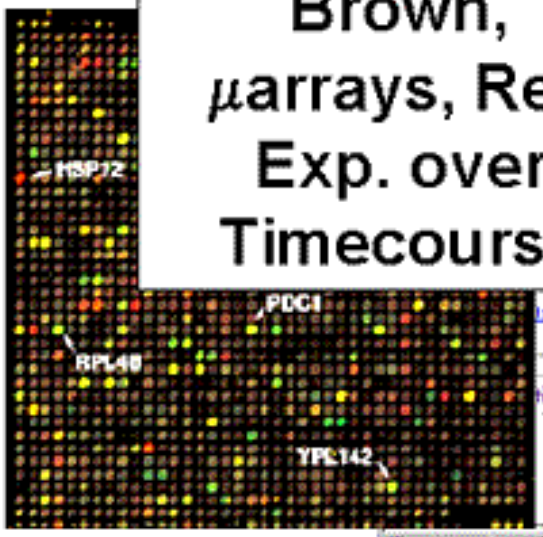

Dissecting the Regulatory Circuitry of a Eukaryotic Genome

Young, Affymetrix GeneChips Abs. Exp.


Large-scale analysis of the yeast genome by transposon tagging and gene disruption

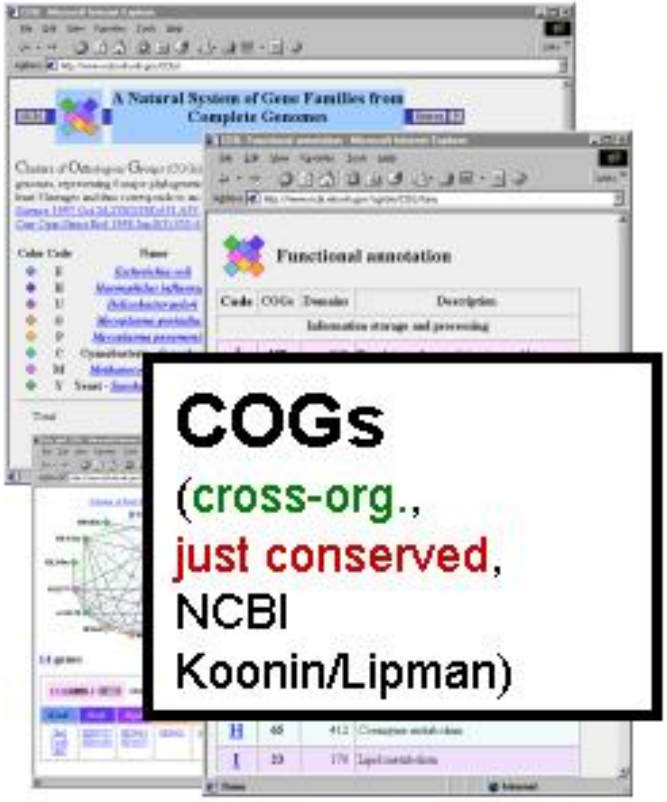
Snyder, Transposons, Protein Abundance



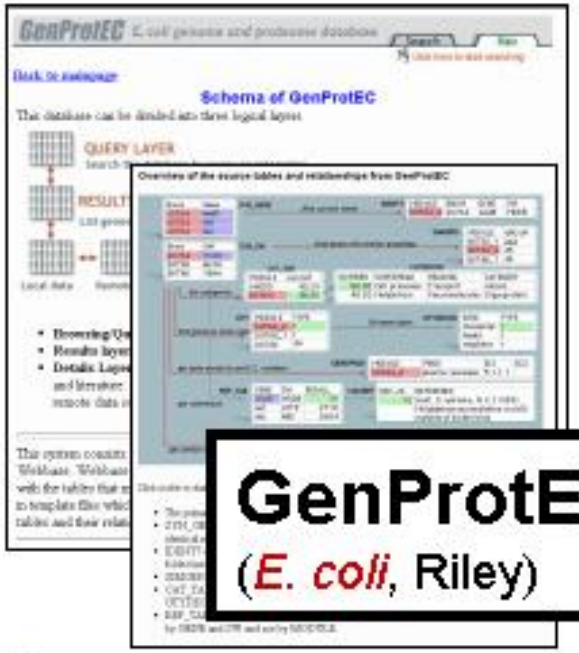


Brown, μarrays, Rel. Exp. over Timecourse

Functional Classification

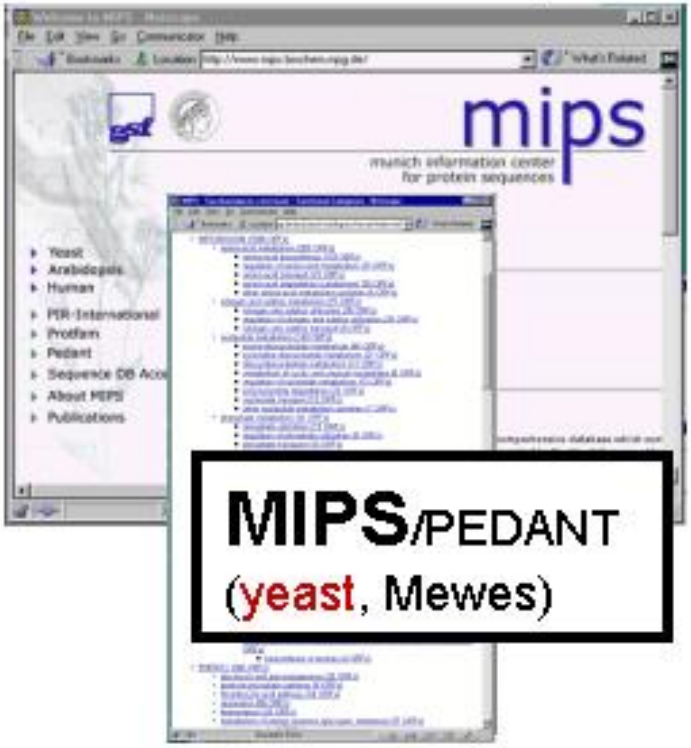


COGs
(cross-org.,
just conserved,
NCBI
Koonin/Lipman)

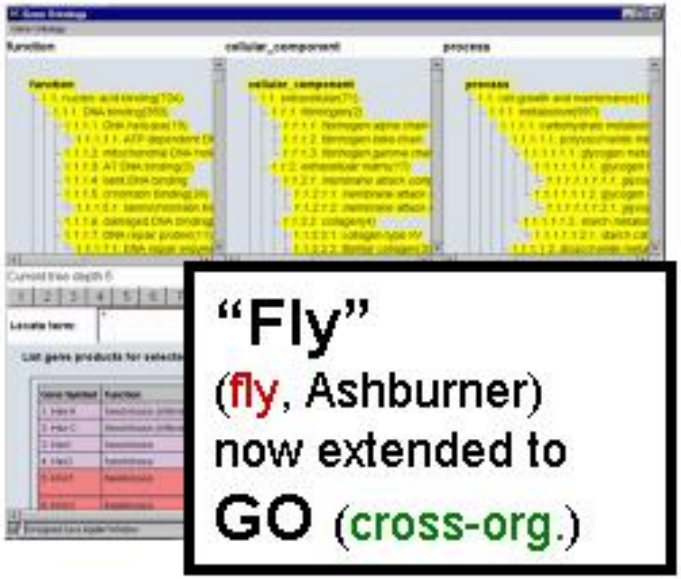


GenProtEC
(*E. coli*, Riley)

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)

[36] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

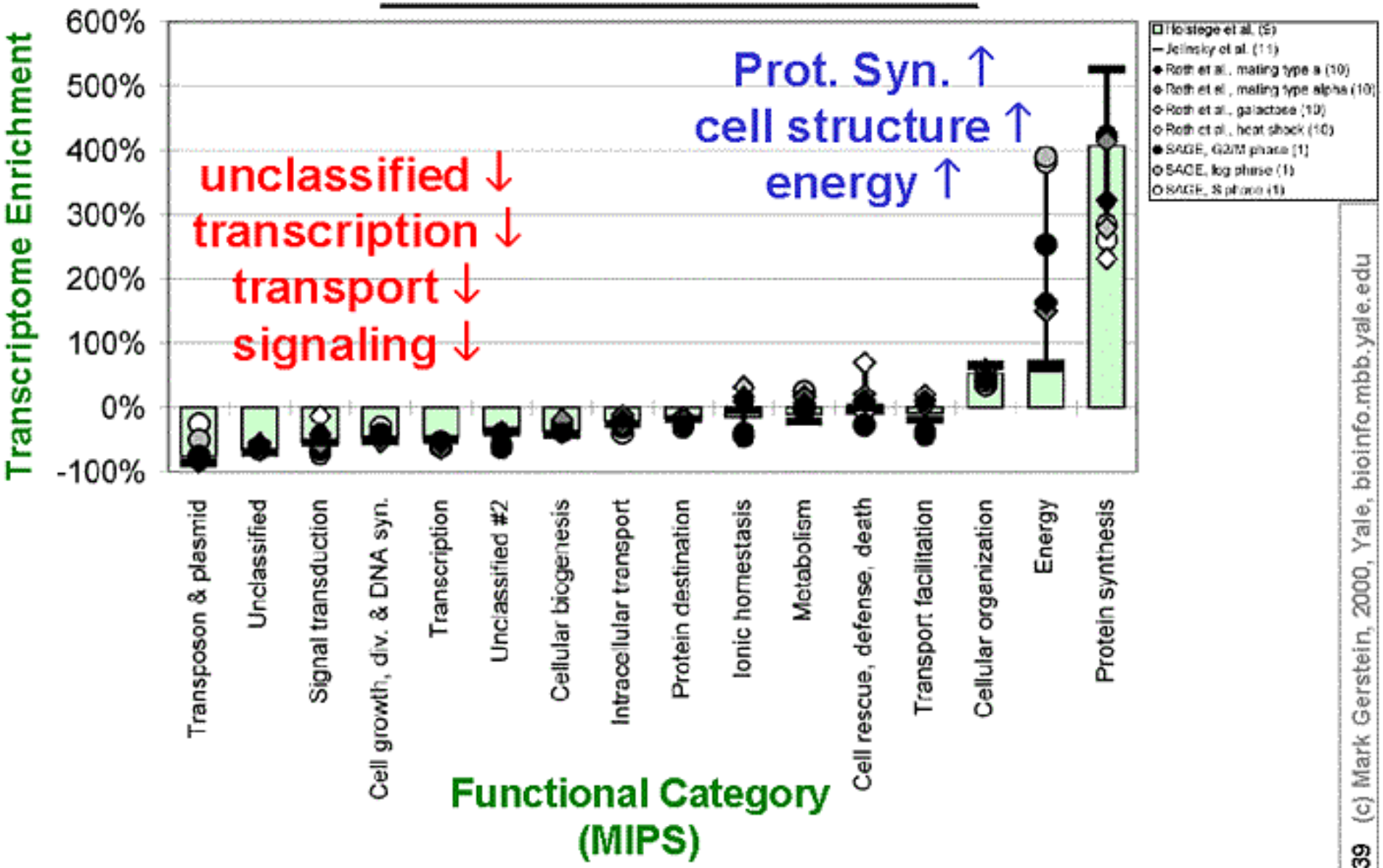
Transcriptome Composition

Common Folds in the Transcriptome

(partslist.org)

	Yeast Genome Rank	Trans-criptome Rank
beta/alpha (TIM)-barrel d1aj2__ :3.1 (1.39) A/B	5	1
P-loop containing nucleotide triphosphate hydrolases d1dai__ :3.29 (1.39) A/B	3	2
Ferredoxin-like d2awO__ :4.34 (1.39) A+B	6	3
NAD(P)-binding Rossmann-fold domains d1eny__ :3.22 (1.39) A/B	8	4
7-bladed beta-propeller d1gotb_ :2.51 (1.39) B	2	5
alpha-alpha superhelix d1a17__ :1.91 (1.39) A	4	6
Thioredoxin fold d1mek__ :3.38 (1.39) A/B	14	7
Glyceraldehyde-3-phosphate dehydrogenase-like d1drw_2 :4.42 (1.39) A+B	78	8
beta-Grasp d1tif__ :4.11 (1.39) A+B	36	9
Heat shock protein 70kD (HSP70) d1dkza_ :5.17 (1.39) M	31	10
Reductase/isomerase/elongation factor common domain d1dar_1 :2.32 (1.39) B	60	11
Serpins d1psi__ :5.2 (1.39) M	78	12
N-terminal nucleophile aminohydrolases (Ntn hydrolases) d1neda_ :4.95 (1.39) A+B	18	13
Metallothionein d1aoo__ :7.38 (1.39) S	96	14
Oligomers of long helices d2ifo__ :1.105 (1.39) A (Leu-zip)	7	15
DNA/RNA polymerases d1har__ :5.9 (1.39) M	25	16
Ribonuclease H-like motif d2itg__ :3.47 (1.39) A/B	15	17
Protein kinases (PK) d3lck__ :5.1 (1.39) M	1	18
Class II aaRS and biotin synthetases d1pysa_ :4.61 (1.39) A+B	21	19
PLP-dependent transferases d2dkb__ :3.54 (1.39) A/B	12	20
DNA/RNA-binding 3-helical bundle d1a5j_1 :1.4 (1.39) A	15	21

Composition of Transcriptome in terms of Functional Classes



[39] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

Common Folds in the Transcriptome #2

(partslist.org)

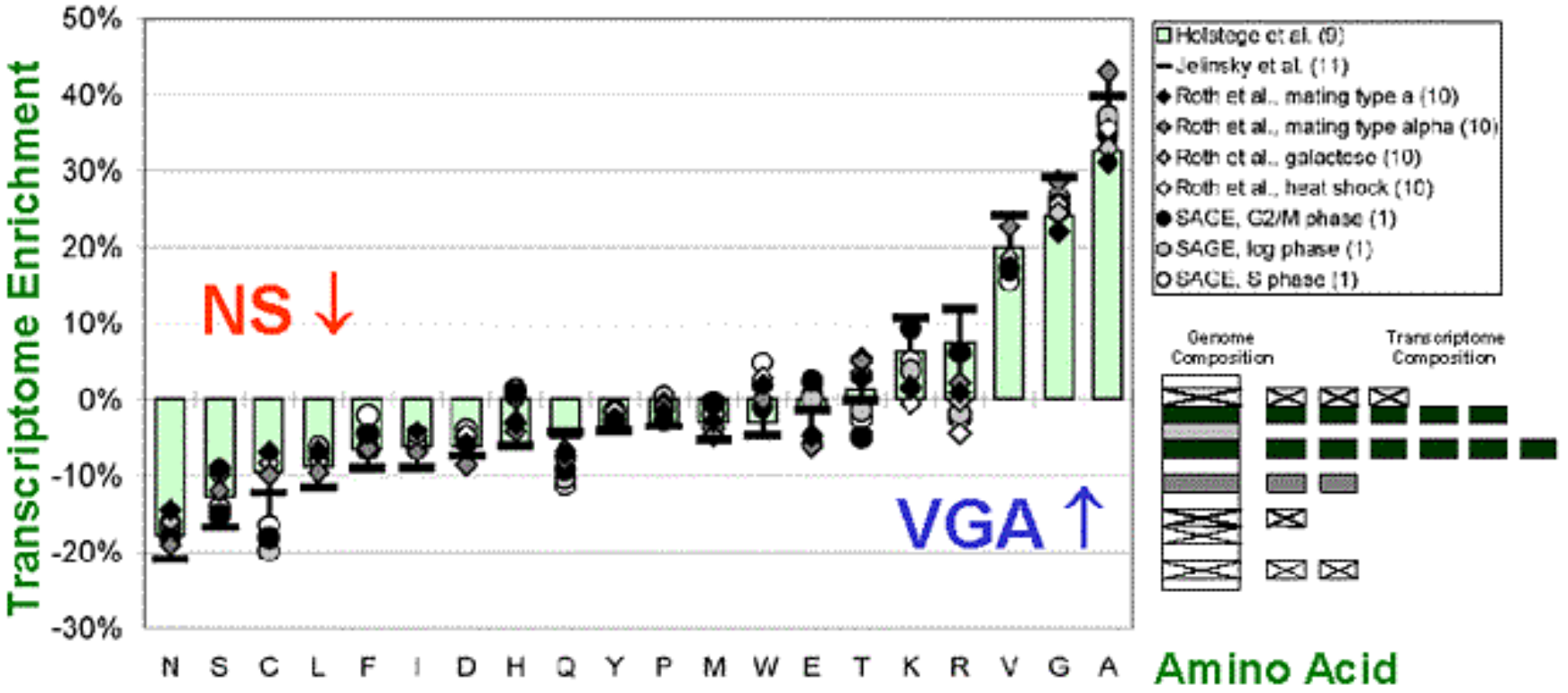
	Yeast Genome Rank	Trans-criptome Rank	
beta/alpha (TIM)-barrel d1aj2__ :3.1 (1.39) A/B	5	1	<Met.↑
P-loop containing nucleotide triphosphate hydrolases d1dai__ :3.29 (1.39) A/B	3	2	<Met.↑
Ferredoxin-like d2awO__ :4.34 (1.39) A+B	6	3	<Met.↑
NAD(P)-binding Rossmann-fold domains d1eny__ :3.22 (1.39) A/B	8	4	<Met.↑
7-bladed beta-propeller d1gotb_ :2.51 (1.39) B	2	5	↓Tfac.>
alpha-alpha superhelix d1a17__ :1.91 (1.39) A	4	6	
Thioredoxin fold d1mek__ :3.38 (1.39) A/B	14	7	
Glyceraldehyde-3-phosphate dehydrogenase-like d1drw_2 :4.42 (1.39) A+B	78	8	↓Sign.> <Met.↑
beta-Grasp d1tif__ :4.11 (1.39) A+B	36	9	
Heat shock protein 70kD (HSP70) d1dkza_ :5.17 (1.39) M	31	10	↓Tfac.>

	Yeast Genome Rank	Trans-criptome Rank
Reductase/isomerase/elongation factor common domain d1dar_1 :2.32 (1.39) B	60	11
Serpins d1psi__ :5.2 (1.39) M	78	12
N-terminal nucleophile aminohydrolases (Ntn hydrolases) d1neda_ :4.95 (1.39) A+B	18	13
Metallothionein d1aoo__ :7.38 (1.39) S	96	14
Oligomers of long helices d2ifo__ :1.105 (1.39) A (Leu-zip)	7	15
DNA/RNA polymerases d1har__ :5.9 (1.39) M	25	16
Ribonuclease H-like motif d2itg__ :3.47 (1.39) A/B	15	17
Protein kinases (PK) d3lck__ :5.1 (1.39) M	1	18
Class II aaRS and biotin synthetases d1pysa_ :4.61 (1.39) A+B	21	19
PLP-dependent transferases d2dkb__ :3.54 (1.39) A/B	12	20
DNA/RNA-binding 3-helical bundle d1a5j_1 :1.4 (1.39) A	15	21

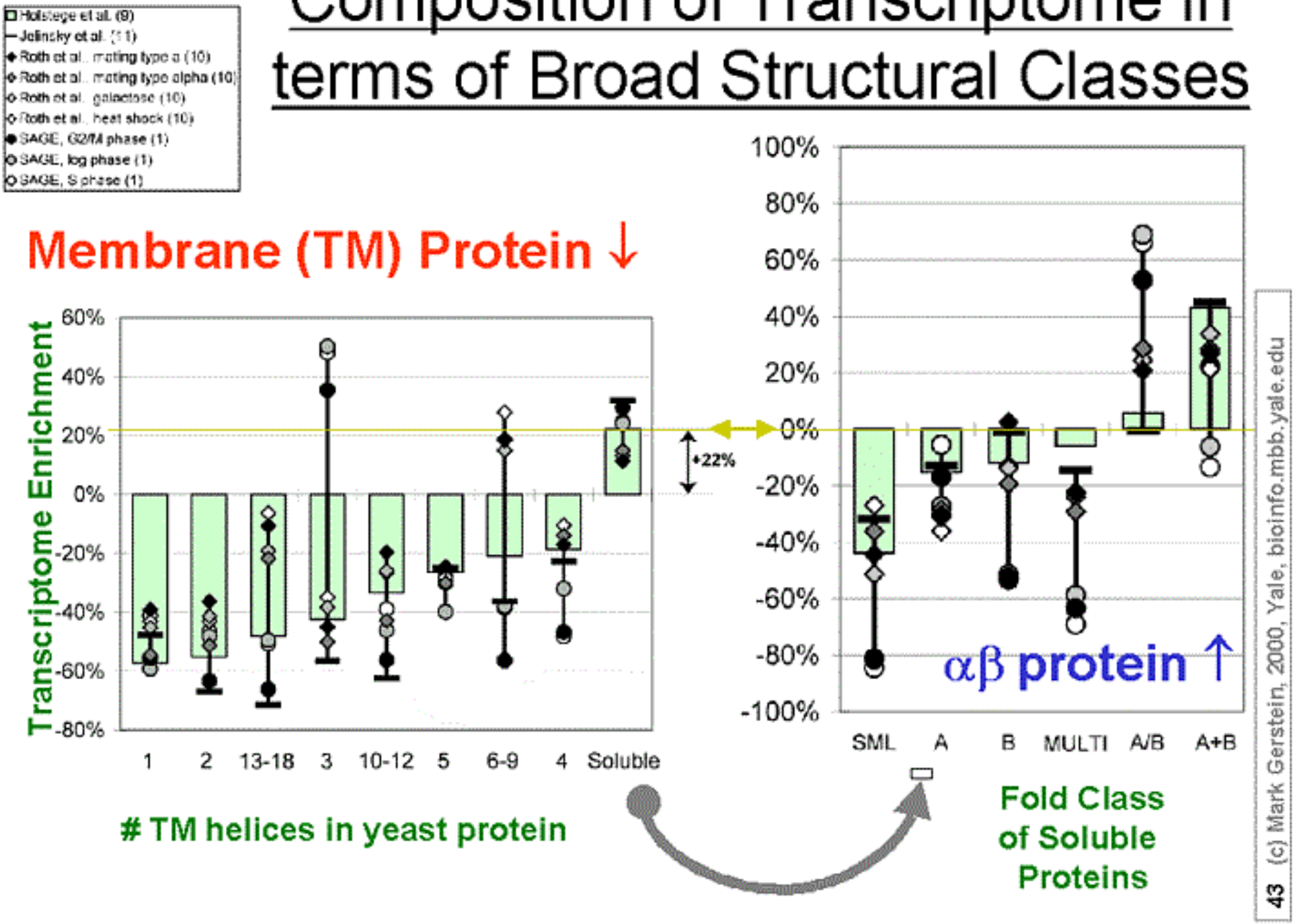
Folds that change a lot in frequency
are not common
Changing Folds
Common Folds
Freq.
Change
Fold of
Genome
Transcriptome
CDC28
CDC15
Diauxic Shift
Sporulation
E. coli heat shock
Rep. PDB
Protein kinases (cat. core)
β-propeller
P-loop NTP hydrolases
α–α superhelix
TIM-barrel
Ferredoxin-like
Rossmann fold
Ribonucleotide reductase (R1)
ATPase dom. of HSP90
Homing endonuclease-like
Aminoacid dehydrogenases; dim. dom
DNA topo I (N-term)
DNA clamp
Metallothionein
Phosphoenolpyruvate carboxykinase
Citrate synthase
N-carbamoylsarcosine amidohydrolase
TRP-like
5'-3' exonuclease
α/α toroid
Cyclin-like
ATPase domain of GroEL
Head domain of GrpE
HSP70 (C-term)
1p38
1nda
1gky
2bct
1byb
1fxd
1xei
1rlr
1ah6
1af5
1hnp
1ois
2pol
1nhu
1ayl
1csh
1nba
1hvl
1tfz
1gai
1vin
1aon
1dkg
1dkz

Composition of Genome vs. Transcriptome

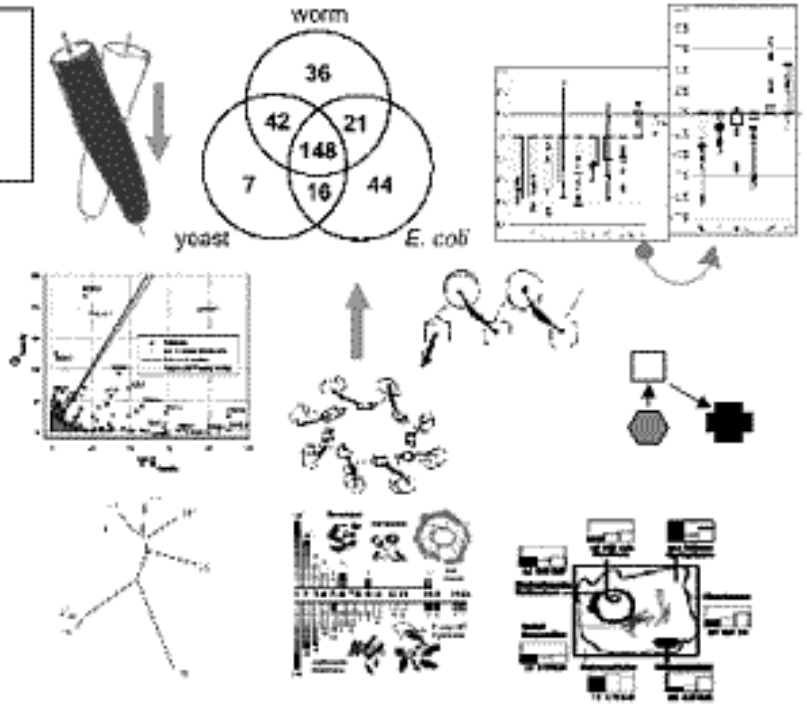
	$\sum_{\text{def } i} n_i(F)$	$\sum_{i'} \sum_{\text{def } i} n_i(F)$	$G(F)$	$\sum_{\text{def } i} e_i n_i(F)$	$\sum_{i'} \sum_{\text{def } i} e_i n_i(F)$	$T(F)$	$D(F)$
Feature *F* is Amino acids, in particular Ala	Number of Ala in yeast	Number of amino acids in yeast	Genome composition of Ala in yeast	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%**



Composition of Transcriptome in terms of Broad Structural Classes



Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

Using Parts & Categories to Mine Expression Data

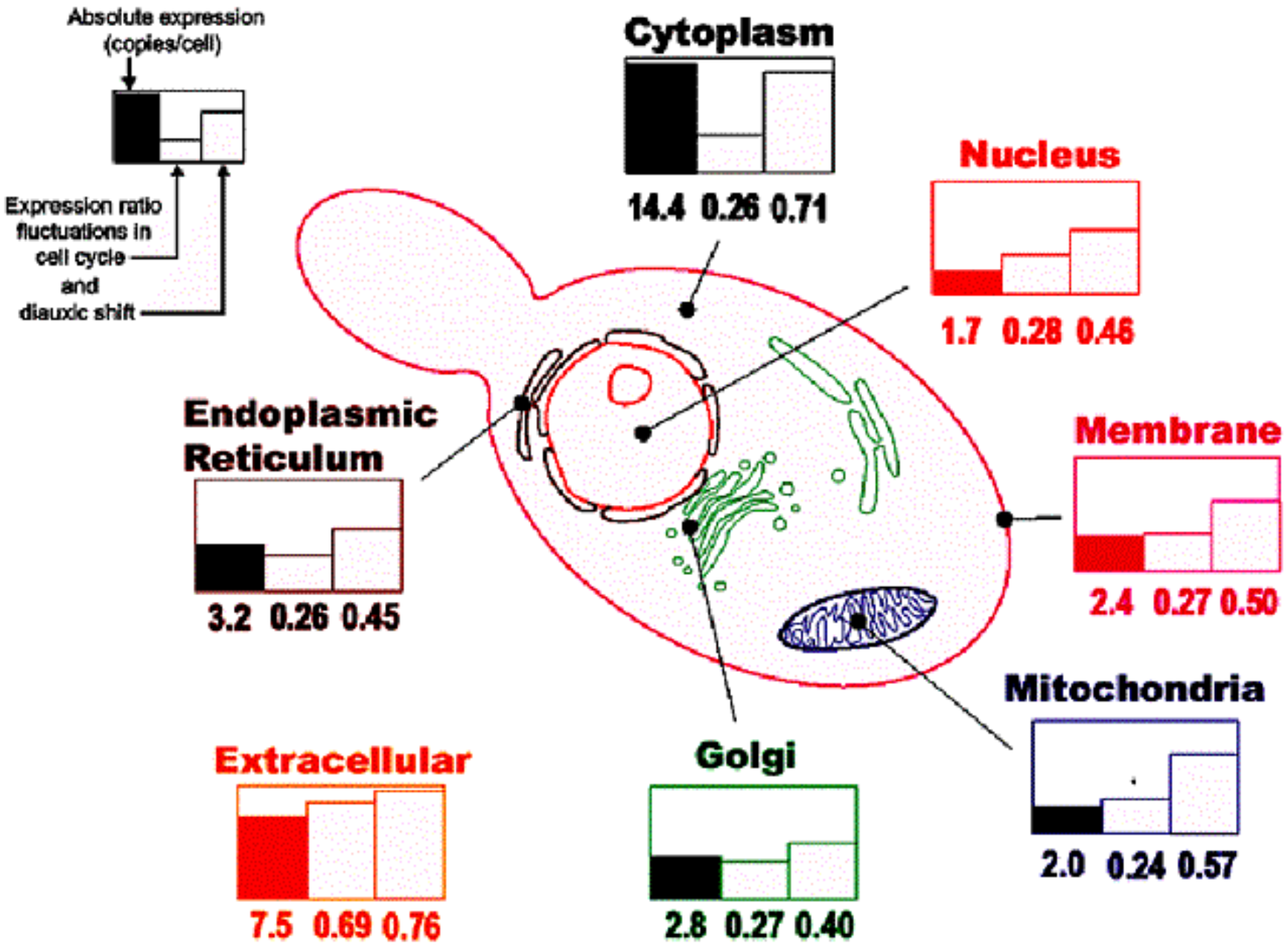
Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

Integrated Views based on Parts

H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

bioinfo.mbb.yale.edu

Expression Level is Related to Localization



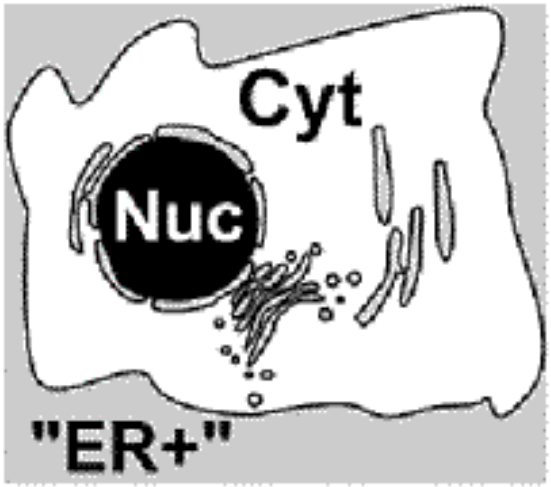
Distributions of Expression Levels

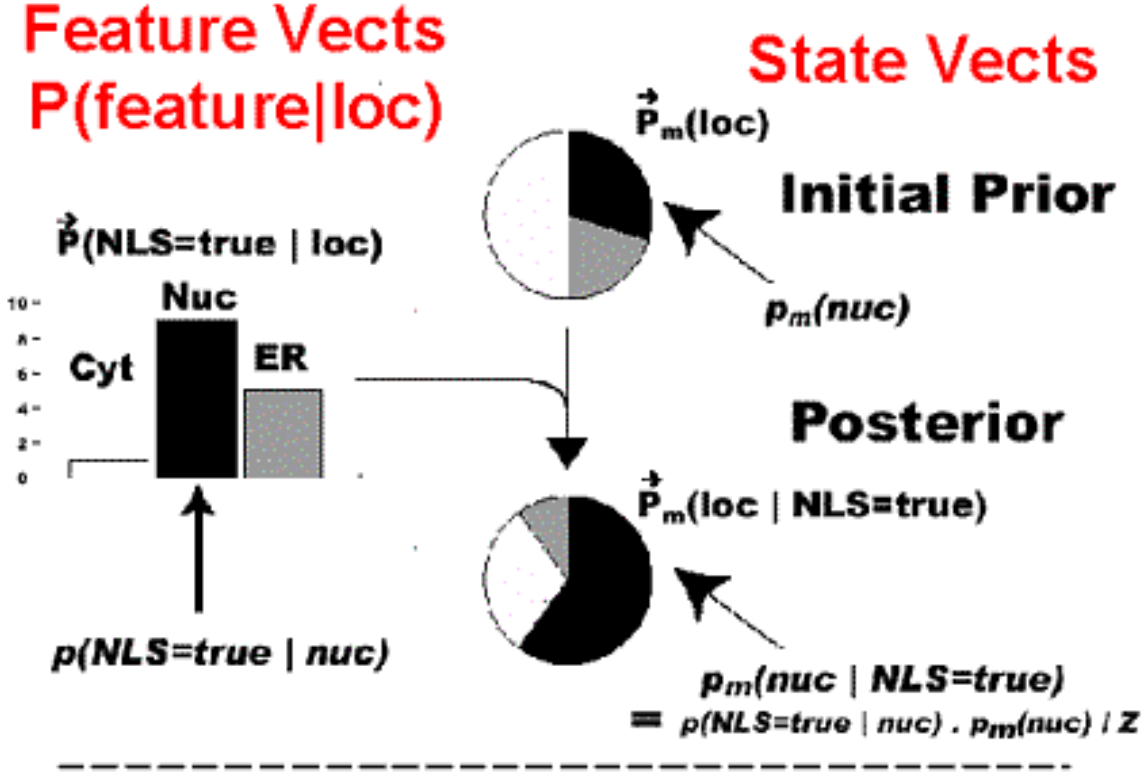
Bayesian System for Localizing Proteins

~6000 yeast genes
with expression levels but only
~2000 with known localization

Bayesian System for Localizing Proteins

loc=



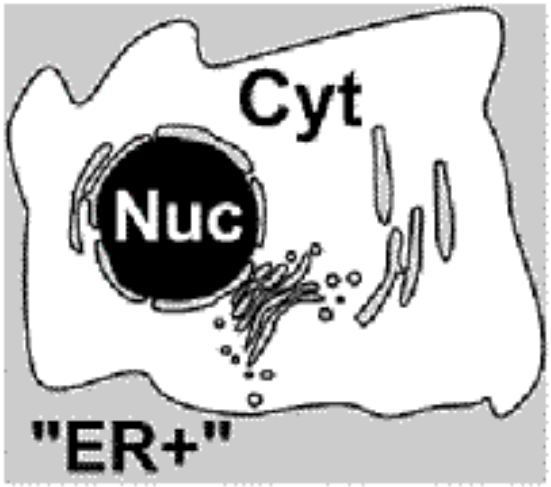


18 Features: Expression Level (abs. & fluct.), signal seq., KDEL, NLS, Essential?, composition

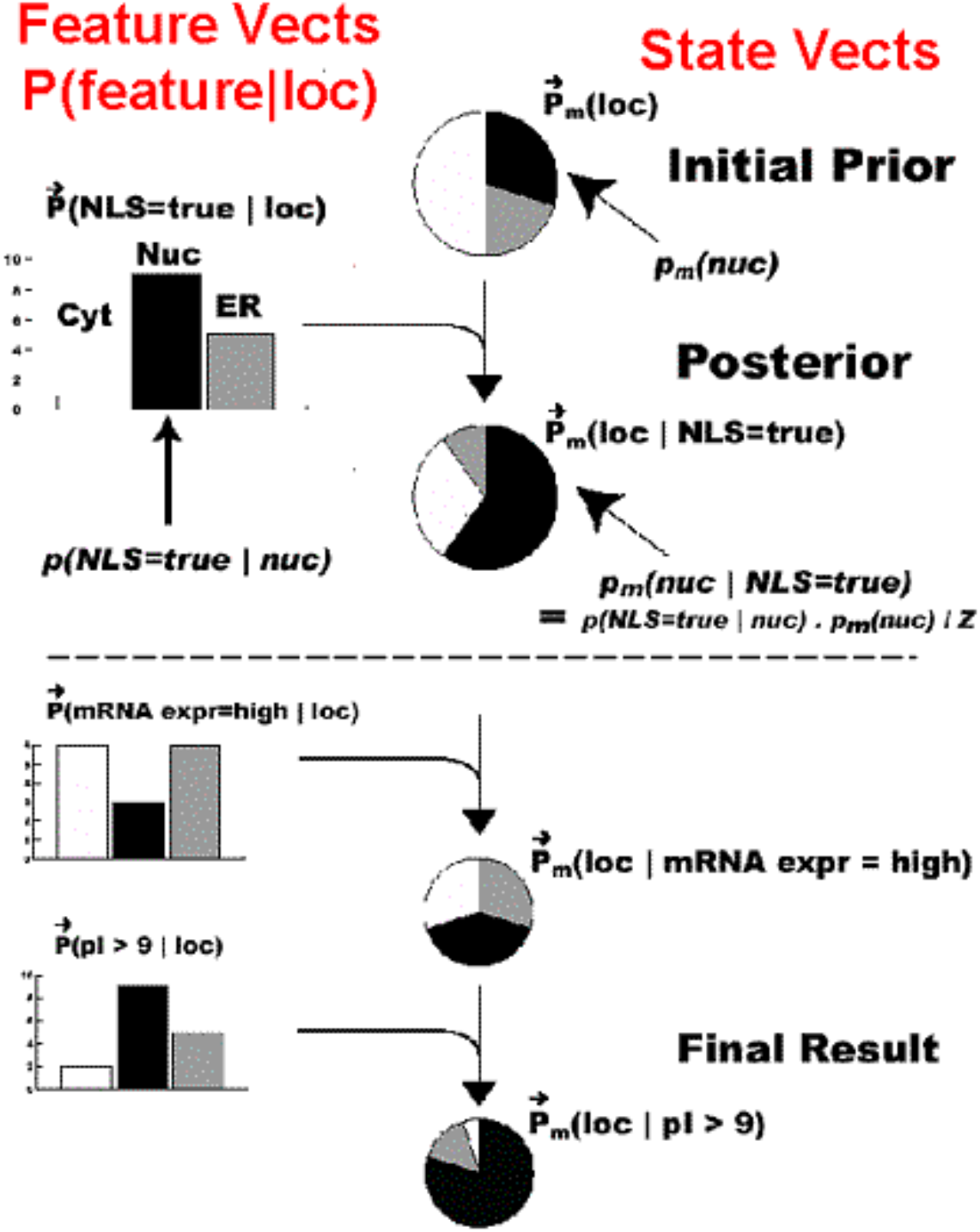
[48] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

Bayesian System for Localizing Proteins

loc=

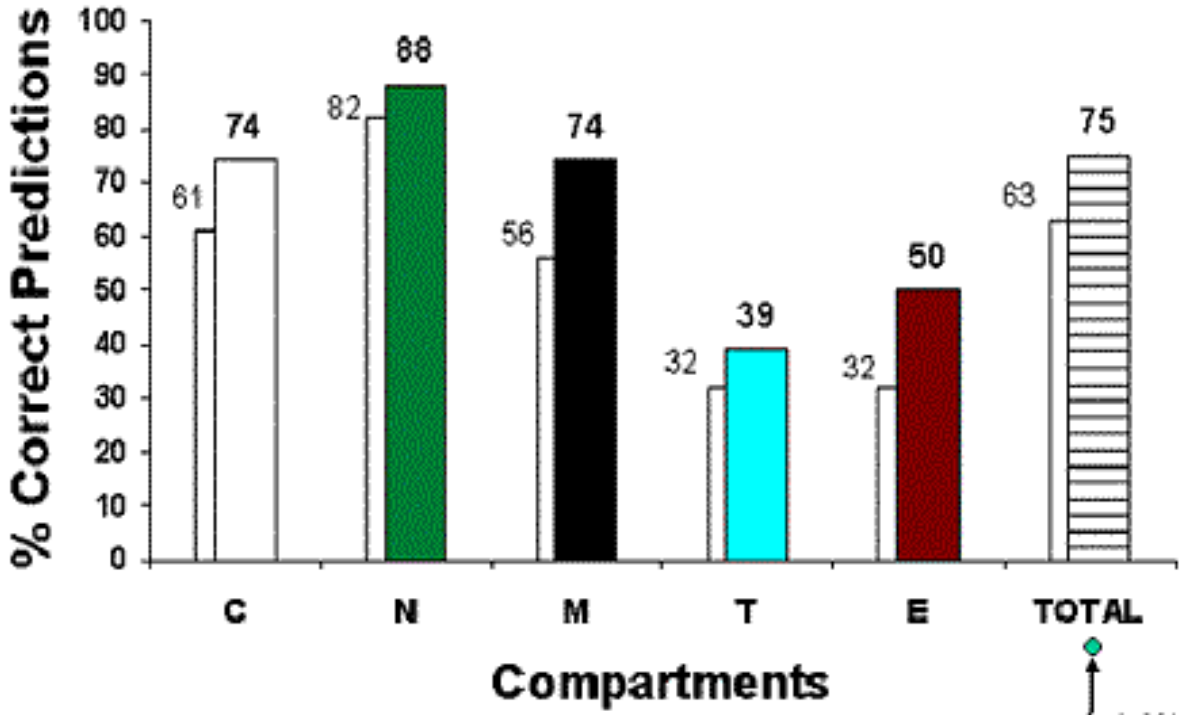


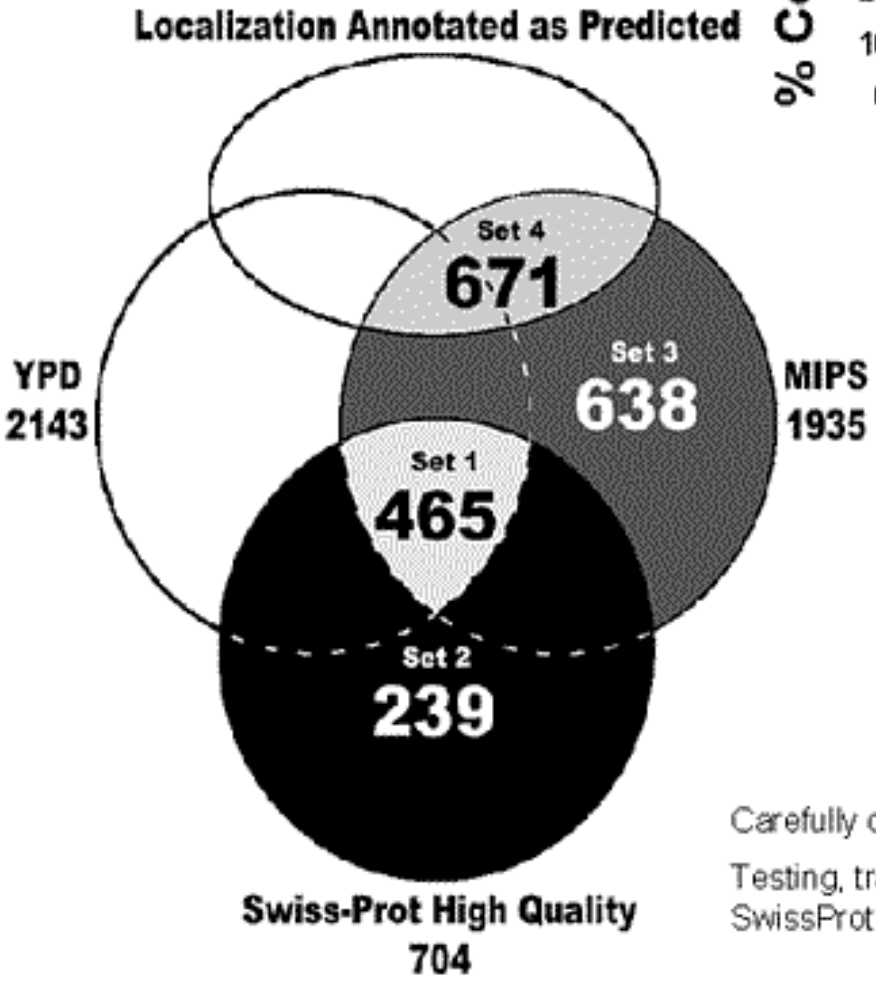
Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.



[49] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

Results on Testing Data





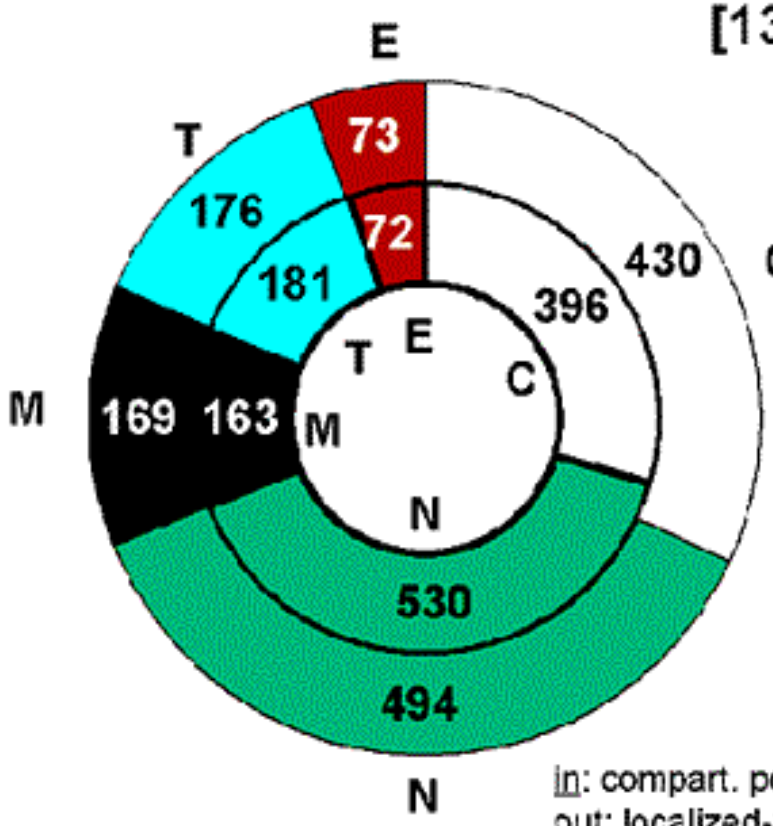
Individual proteins: 75%
with cross-validation

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.



Overall Compartment Population Vector

$\vec{N}(L) = \vec{P}_1(L) + \vec{P}_2 + \vec{P}_3 + \ldots + \vec{P}_m + \ldots + \vec{P}_{6000}$

State vectors of prc

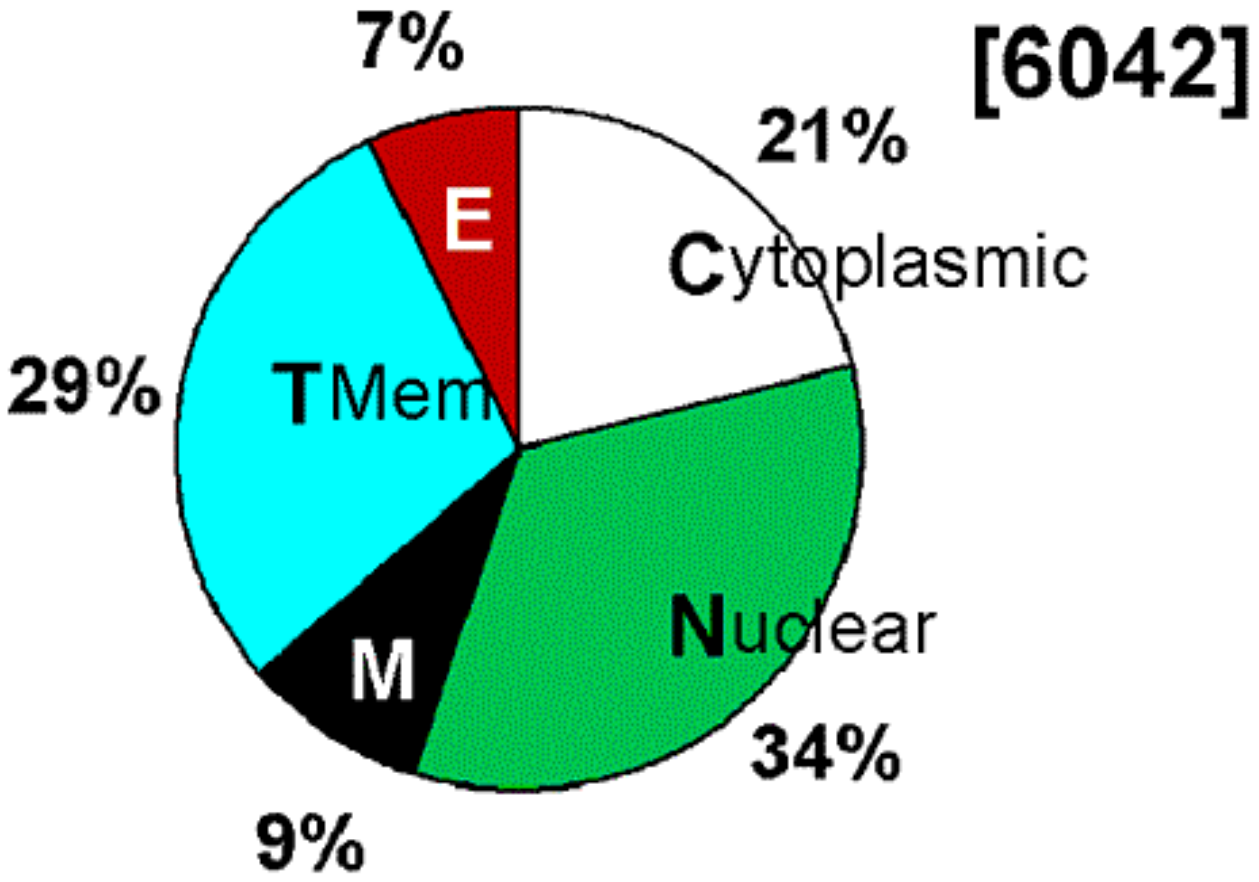
Normal "sum" of protein localizations

Thresholding state vectors

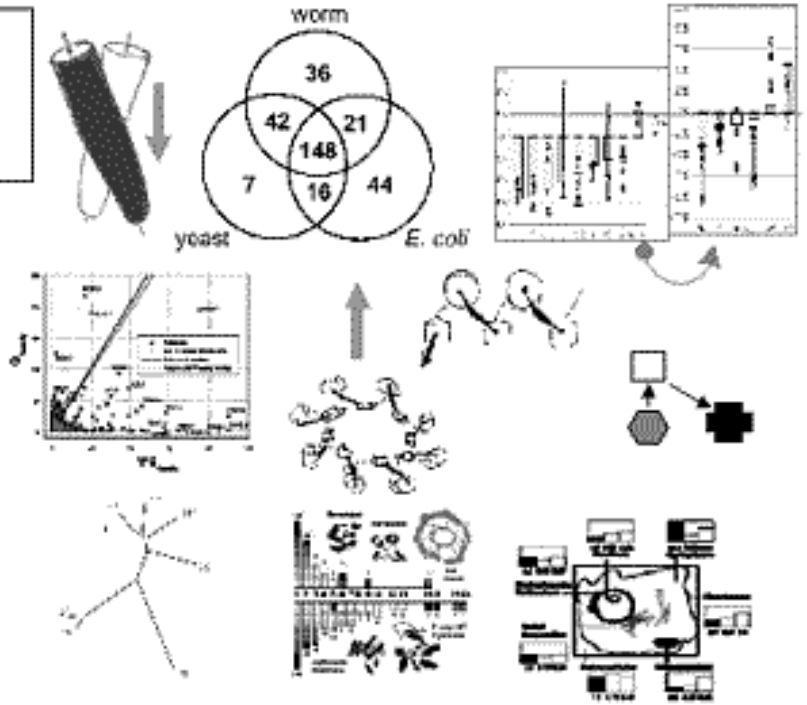
N ? E C ?

Gives **96%** pop. similarity.

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



Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

Using Parts & Categories to Mine Expression Data

Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

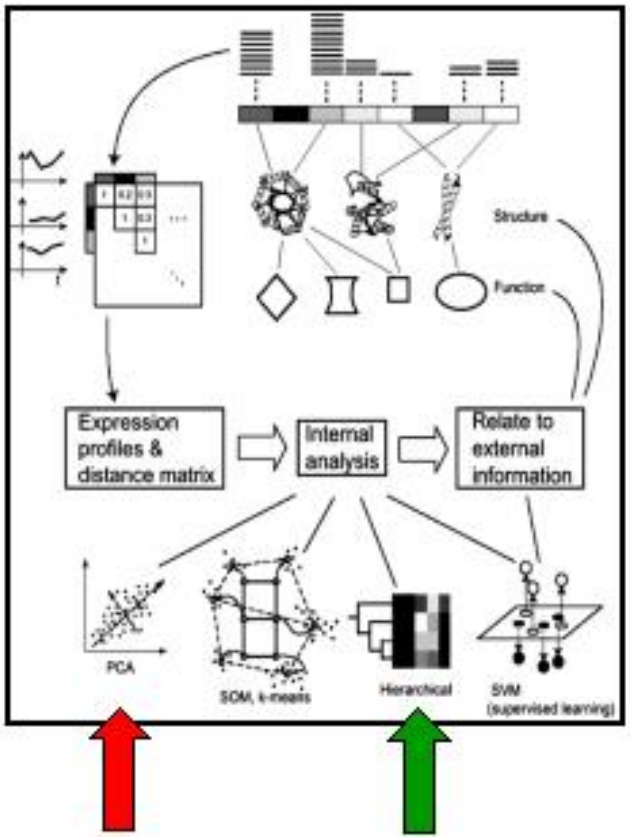
Integrated Views based on Parts

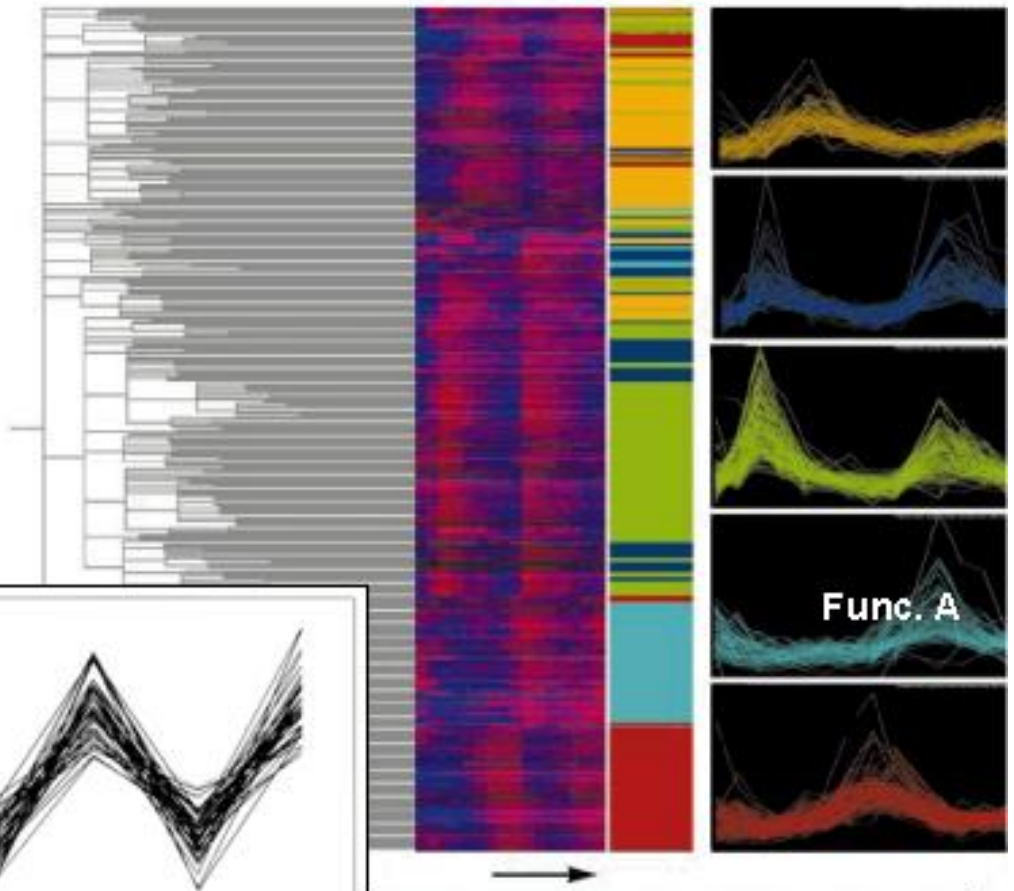
H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

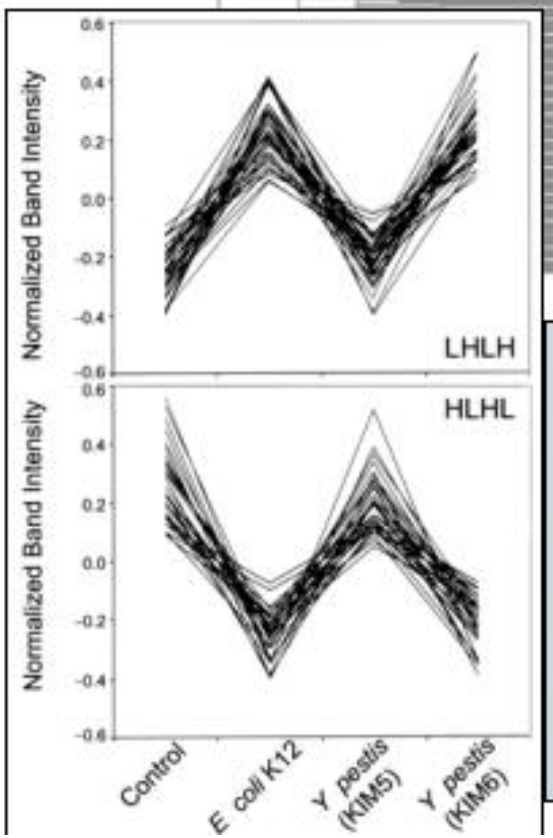
bioinfo.mbb.yale.edu

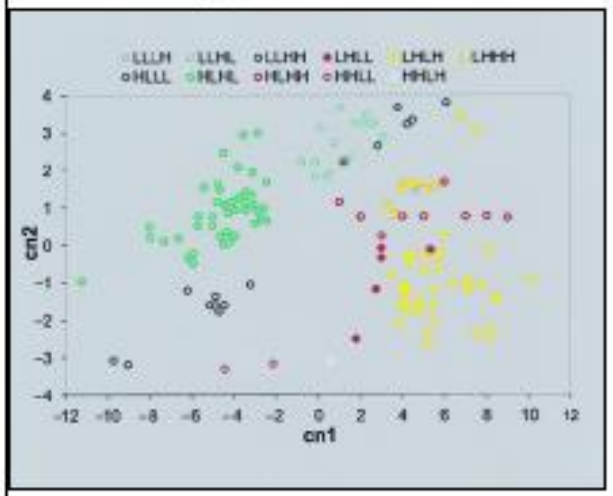
Do Expression Clusters Relate to Protein Function?
Botstein-Brown (Eisen), Altman, Lander, Haussler, Church, & Eisenberg groups
Structure
Function
Relate to external information
Internal analysis
Expression profiles & distance matrix
SVM (supervised learning)
Hierarchical
SOM, k-means
PCA

Examples: Hierarchical Clustering & PCA, Do they relate to Function?



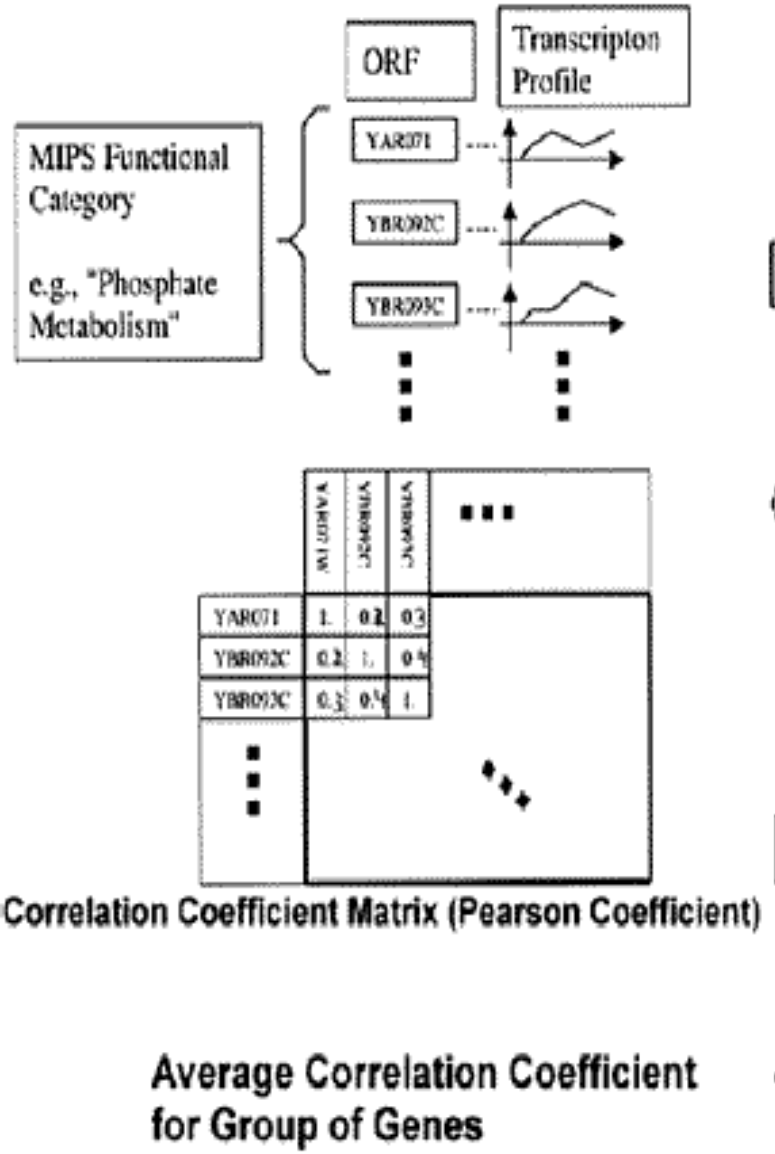


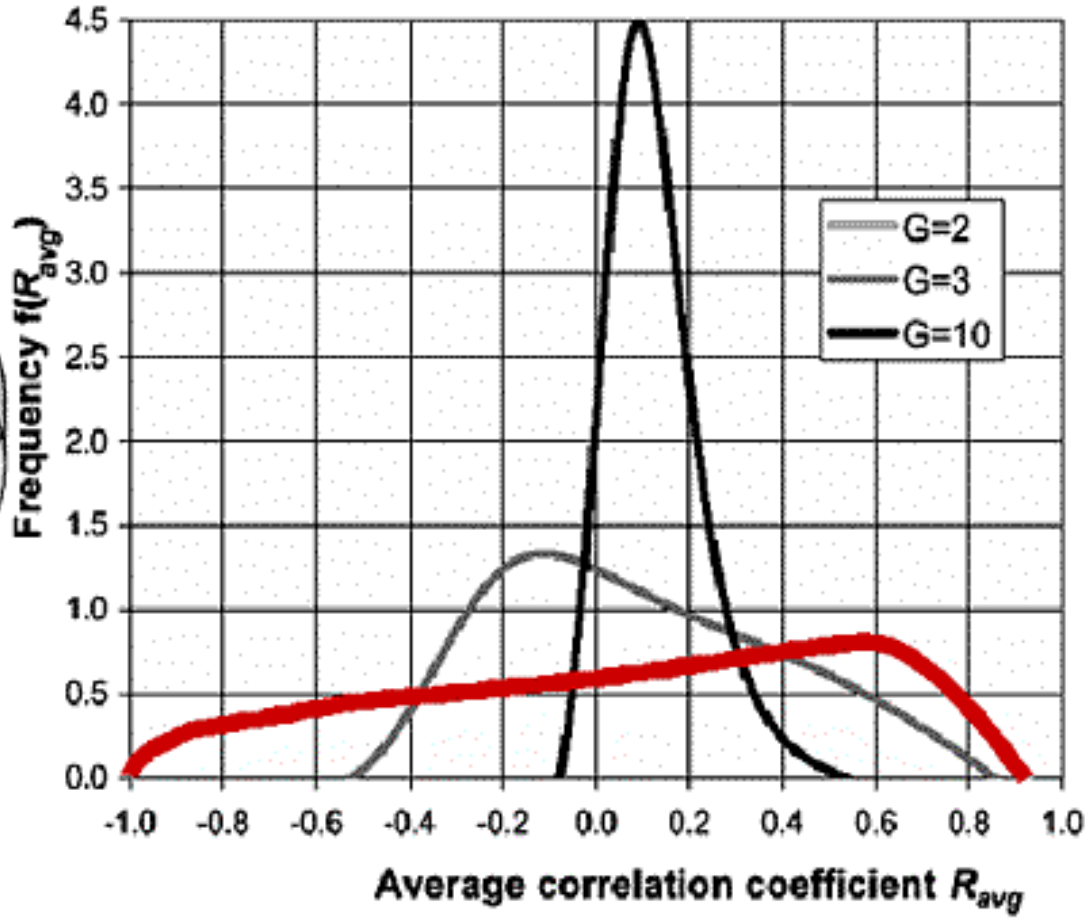




[55] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

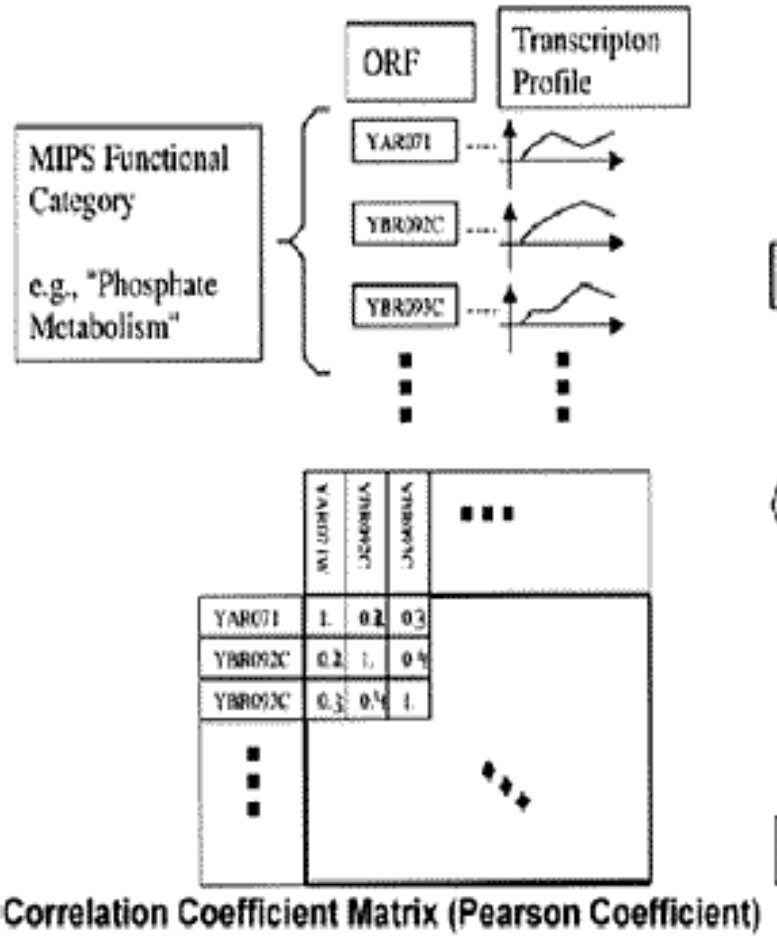
Distributions of Gene Expression Correlations, for All Possible Gene Groupings





[56] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

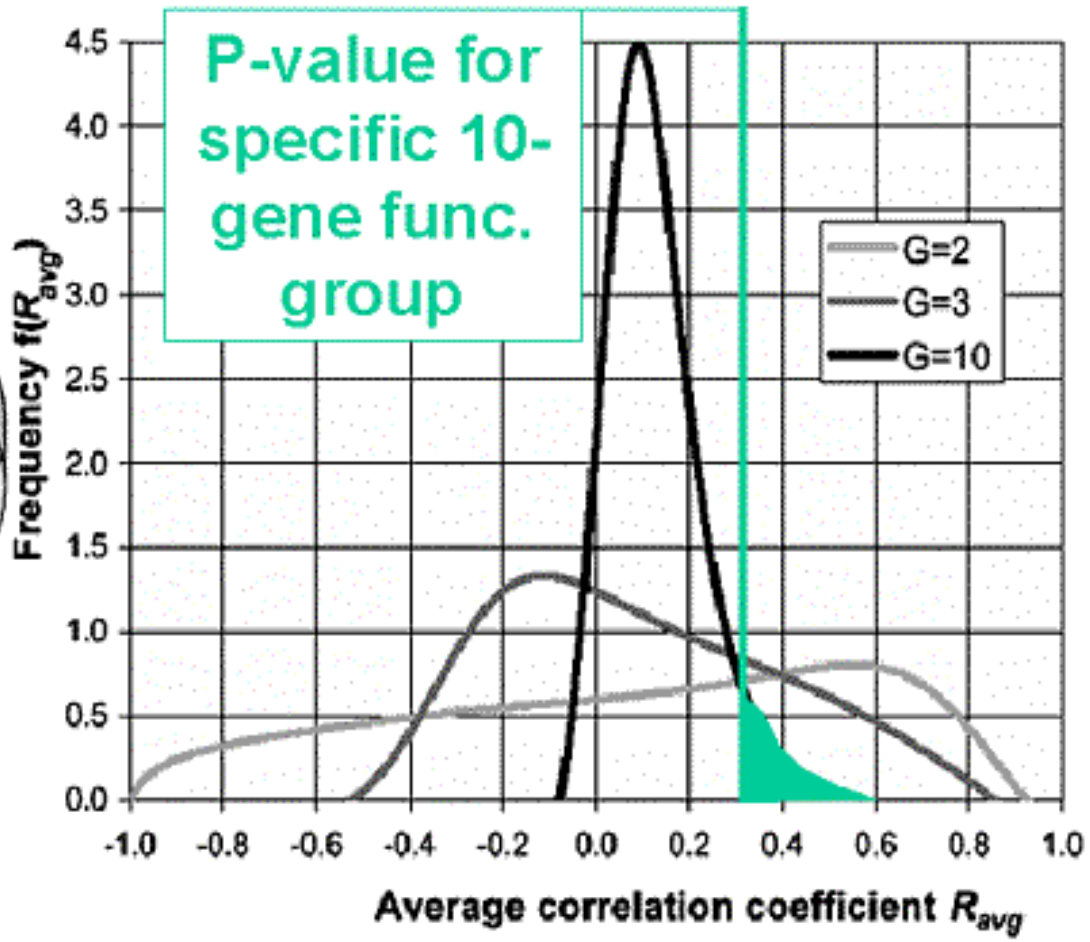
Distributions of Gene Expression Correlations, for All Possible Gene Groupings #2

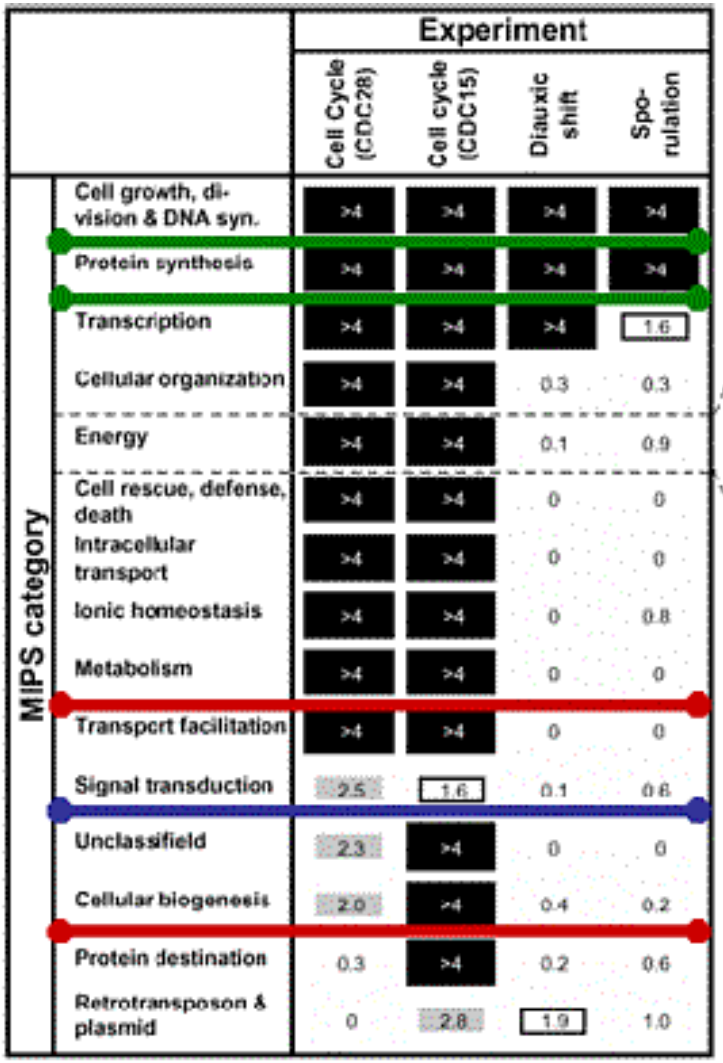


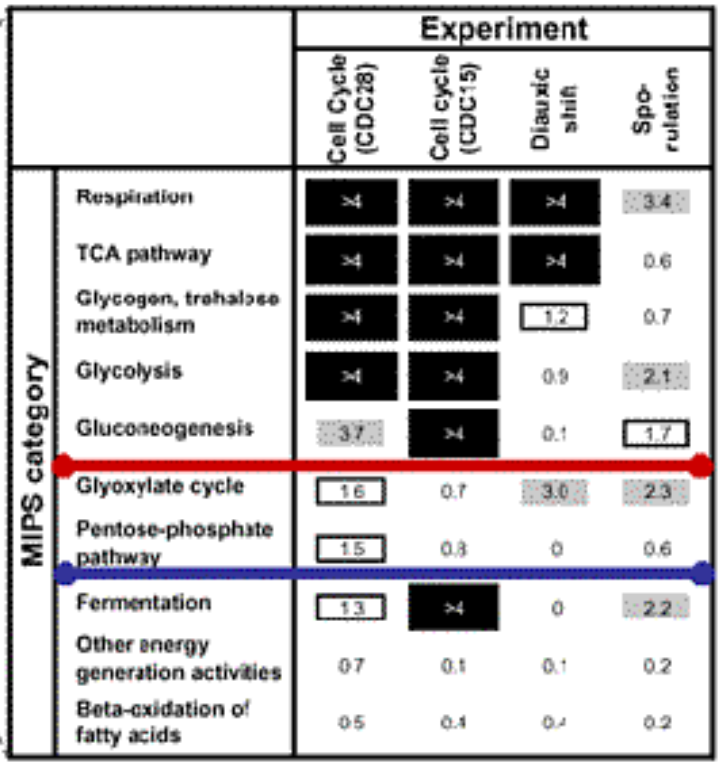
Average Correlation Coefficient for Group of Genes

Sample for Diauxic shift Expt. (Brown),

Ex. R_{avg0-3}=
[R(gene-1,gene-3) + R(gene-1,gene-4) + R(gene-5,gene-7)] / 3







Correlation:

Always Significant

Sometimes Significant (depends on expt.)

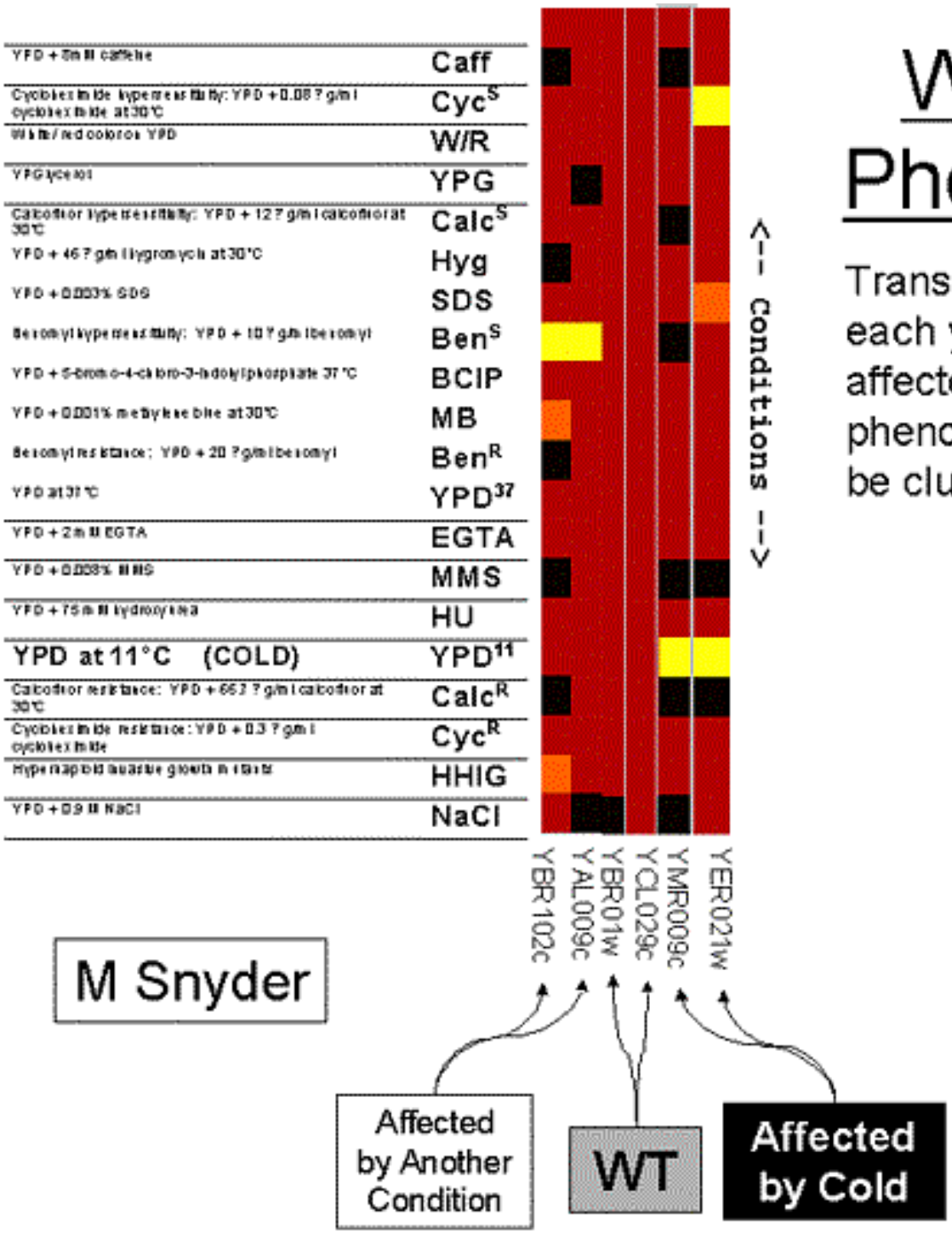
Never Significant

	Fraction of significant groups				Total # groups
	CDC28	CDC15	Diauxic Shift	Sporu-lation	
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

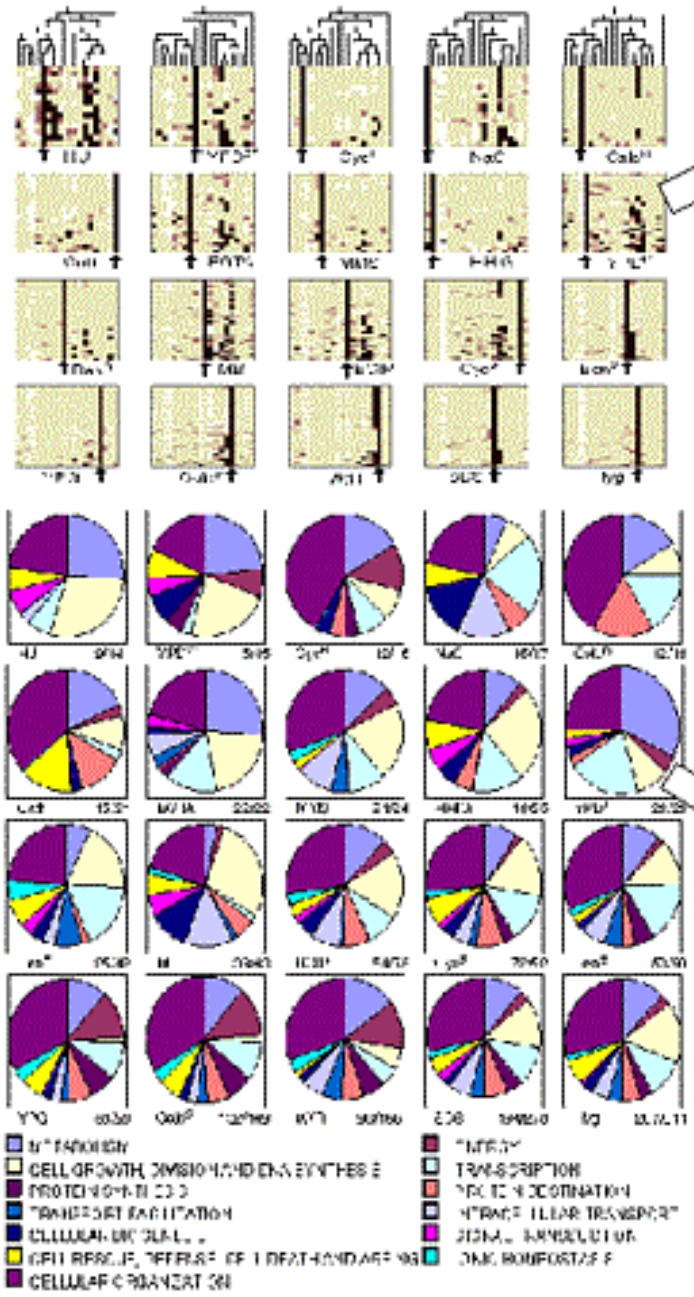
Based on Distributions, Correlation of Established Functional Categories, Computer Clusterings

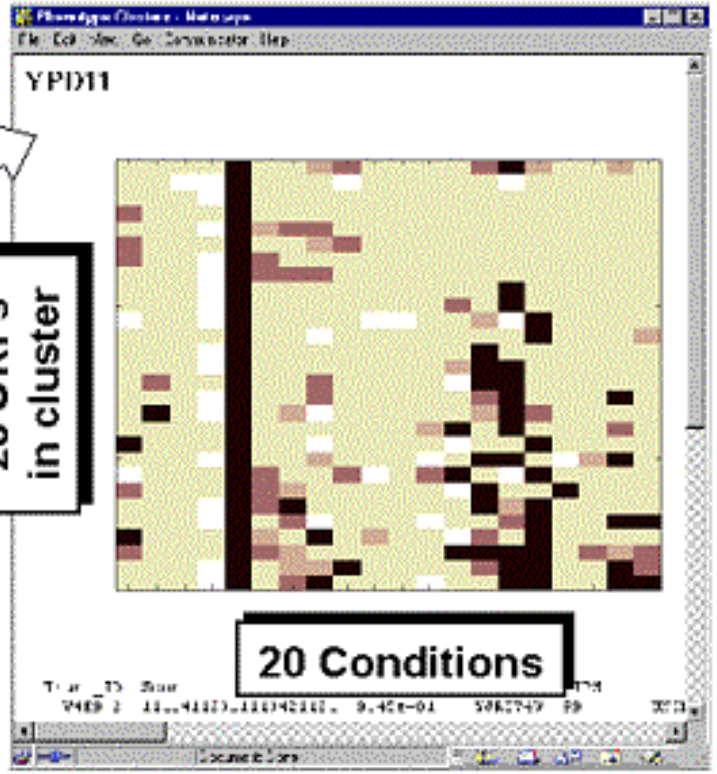
Whole Genome Phenotype Profiles

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



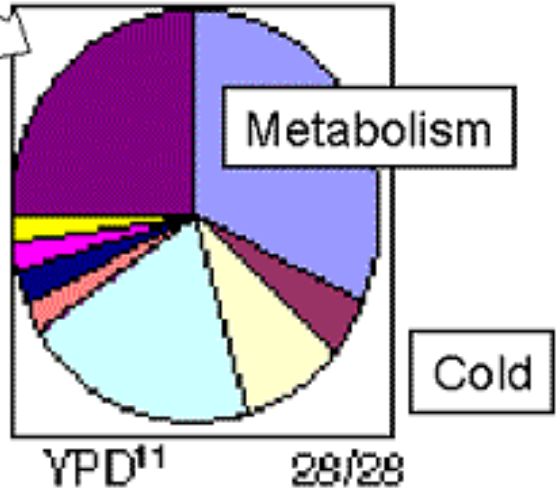
Phenotype ORF Clusters from Transposon Expt.





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**

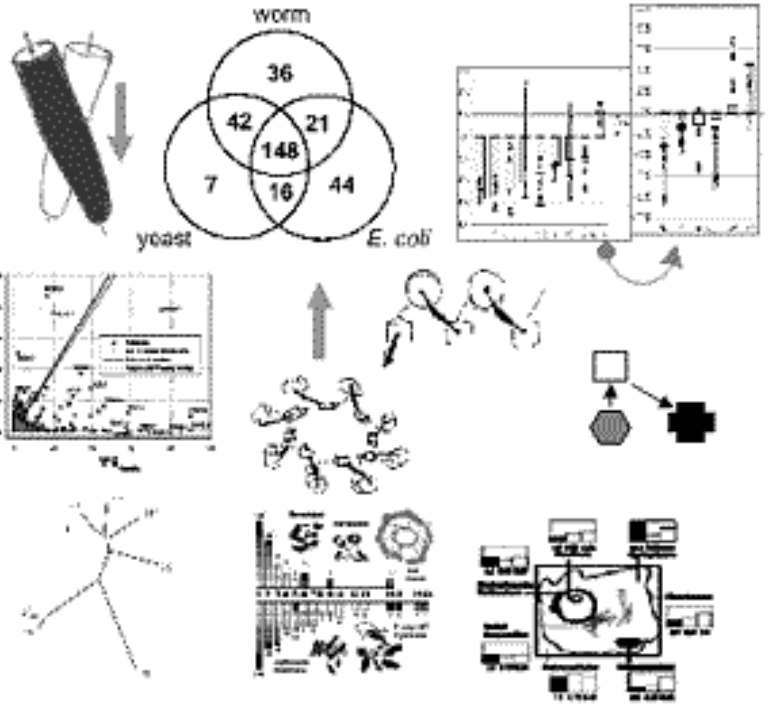
k-means clustering of ORFs based on "phenotype patterns," cross-ref. to MIPs Functional Classes



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

M Snyder, A Kumar, et al....

Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

Using Parts & Categories to Mine Expression Data

Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

Integrated Views based on Parts

H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

bioinfo.mbb.yale.edu

[62] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

[63] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

[64] Dr. M. Gerstein, Yale (ITP 2/28/01) Integrative Genomics: Surveys of a Finite Parts List

Can FUNCTION be defined well enough to relate to expression?

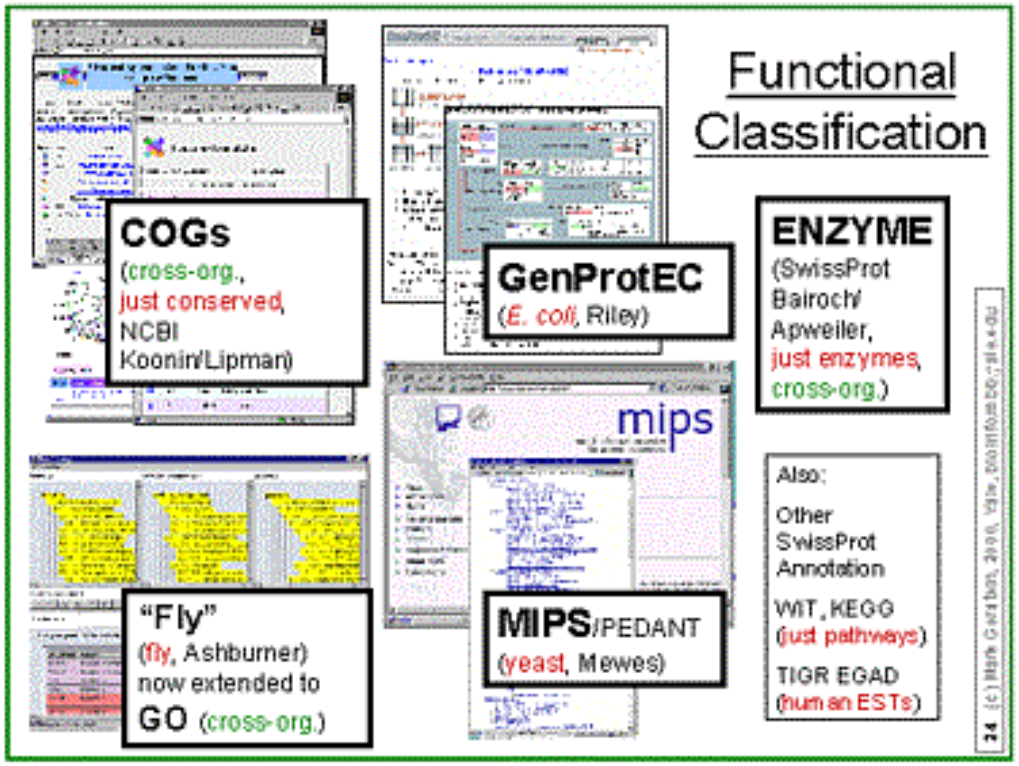
Problems defining function:

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

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

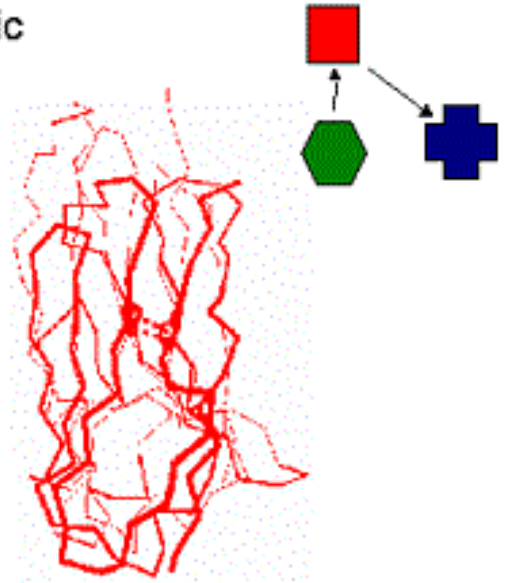
Non-systematic Terminology:

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

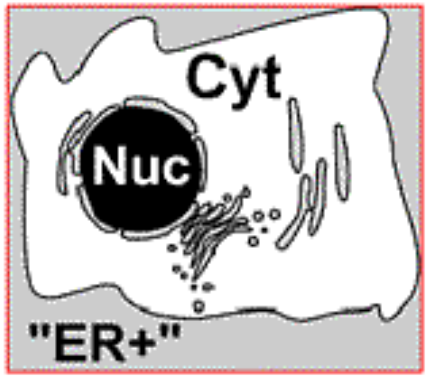


More clearly defined attributes of proteins

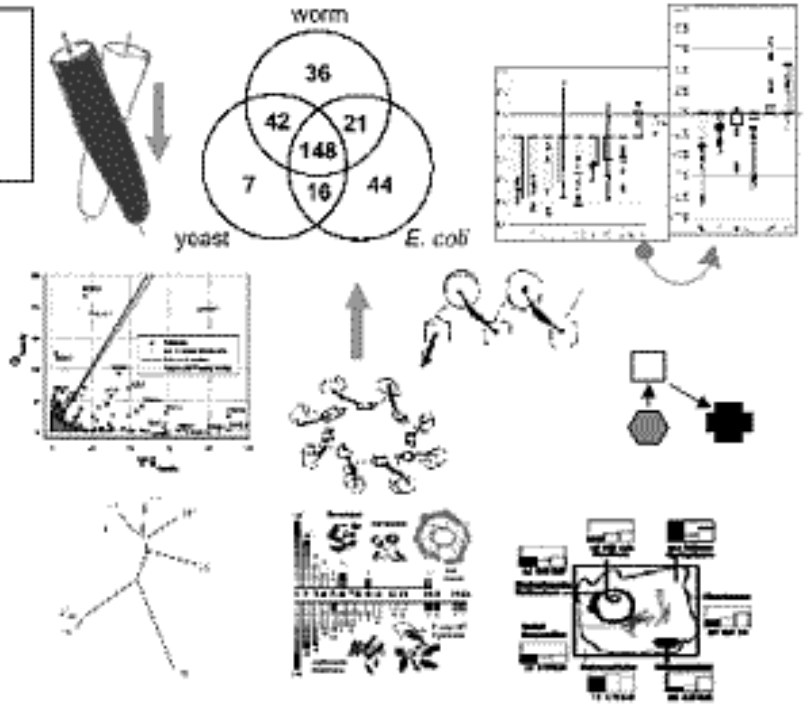
Fold, Localization, Interactions & Regulation



vs.



Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

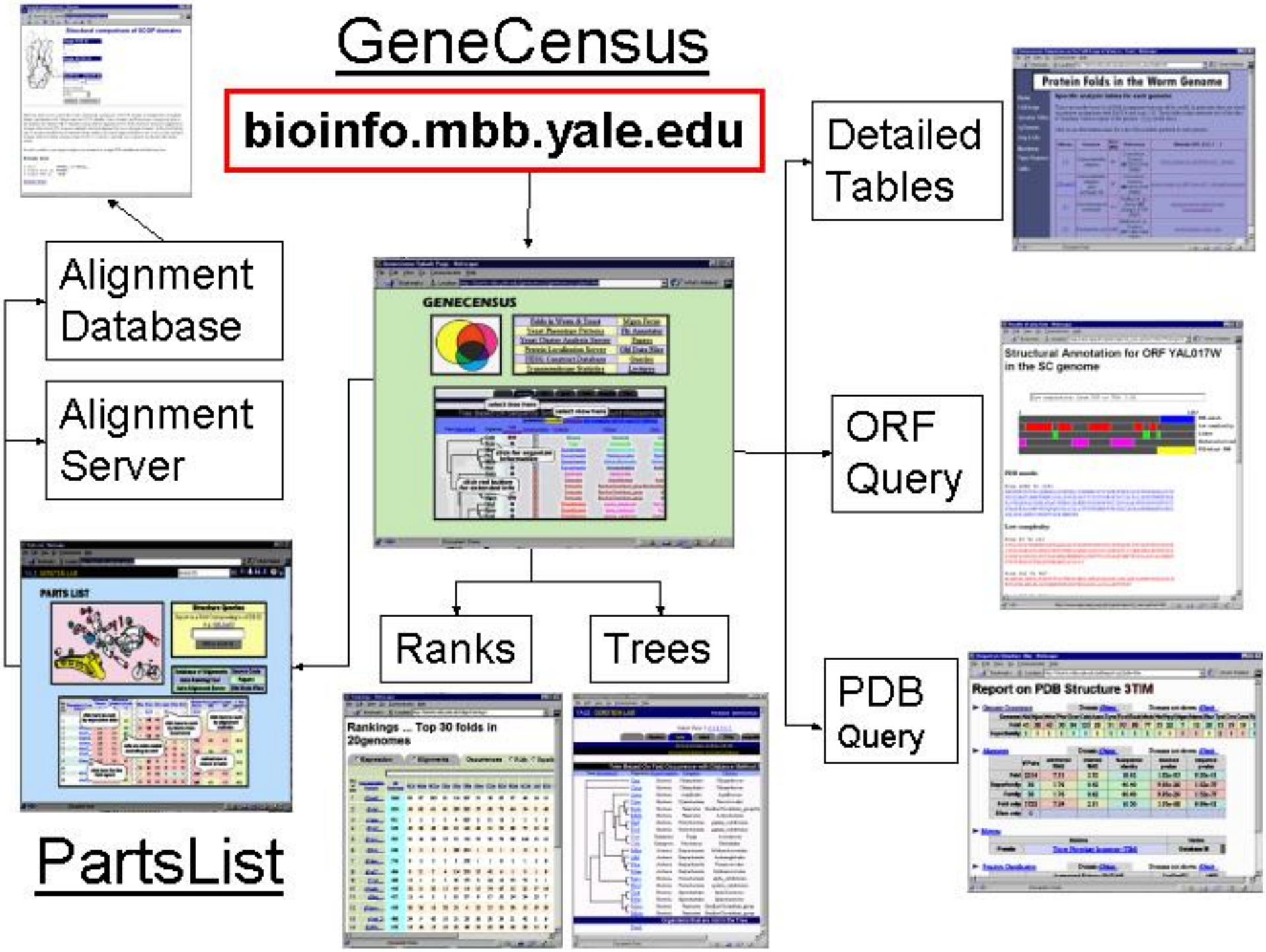
Using Parts & Categories to Mine Expression Data

Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

Integrated Views based on Parts

H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

bioinfo.mbb.yale.edu



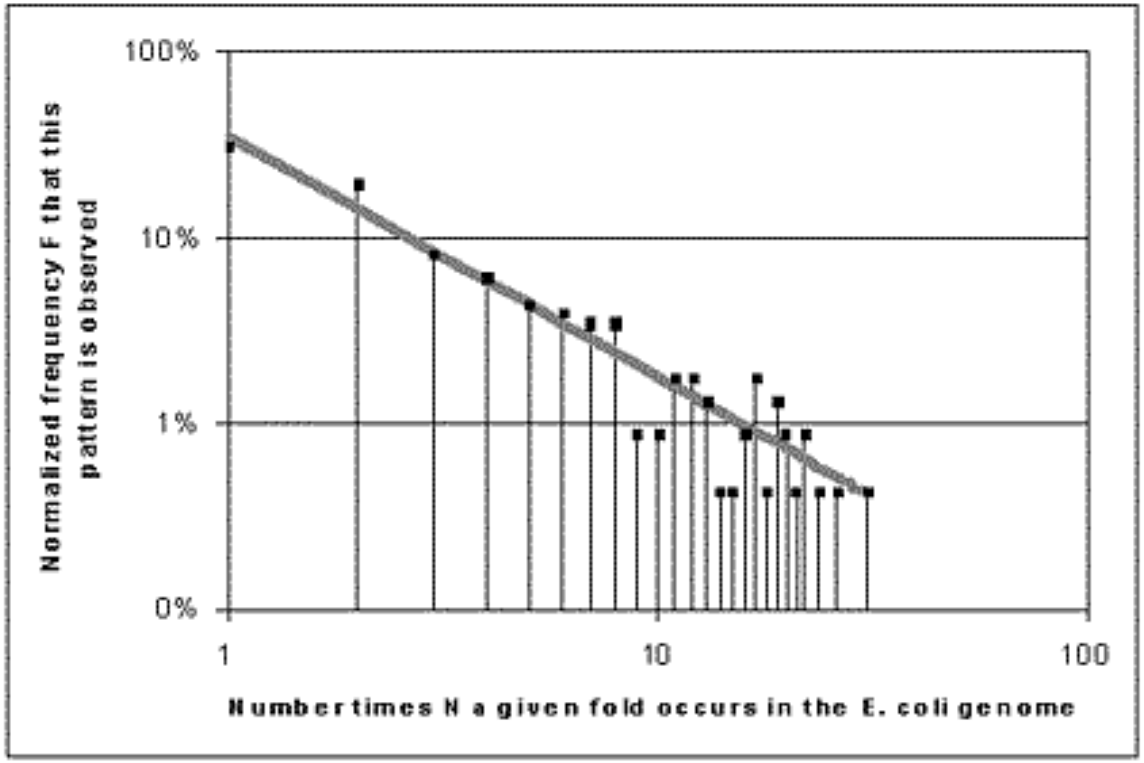
GeneCensus Dynamic Tree Viewers

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

Surveying a Finite PartsList from Many Perspectives

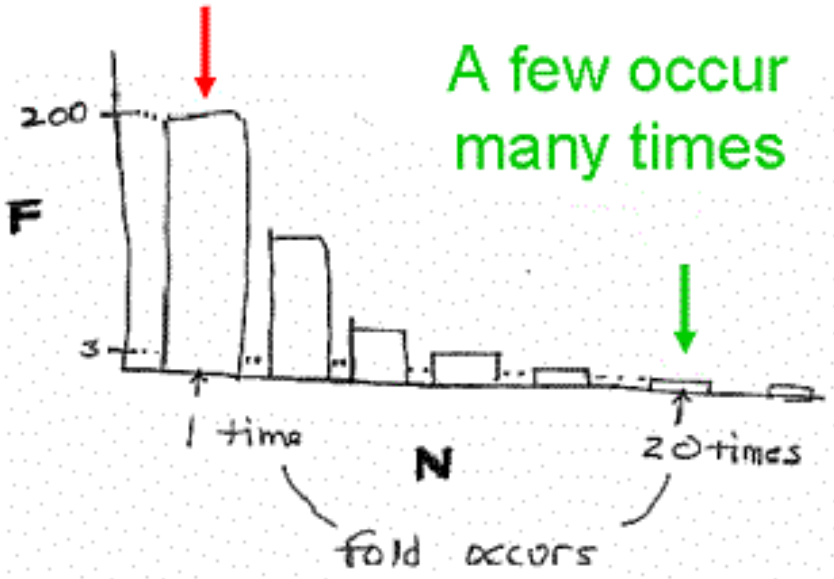
Insert Animations

Part Occurrence Frequencies Follow a Power Law



Most occur just once

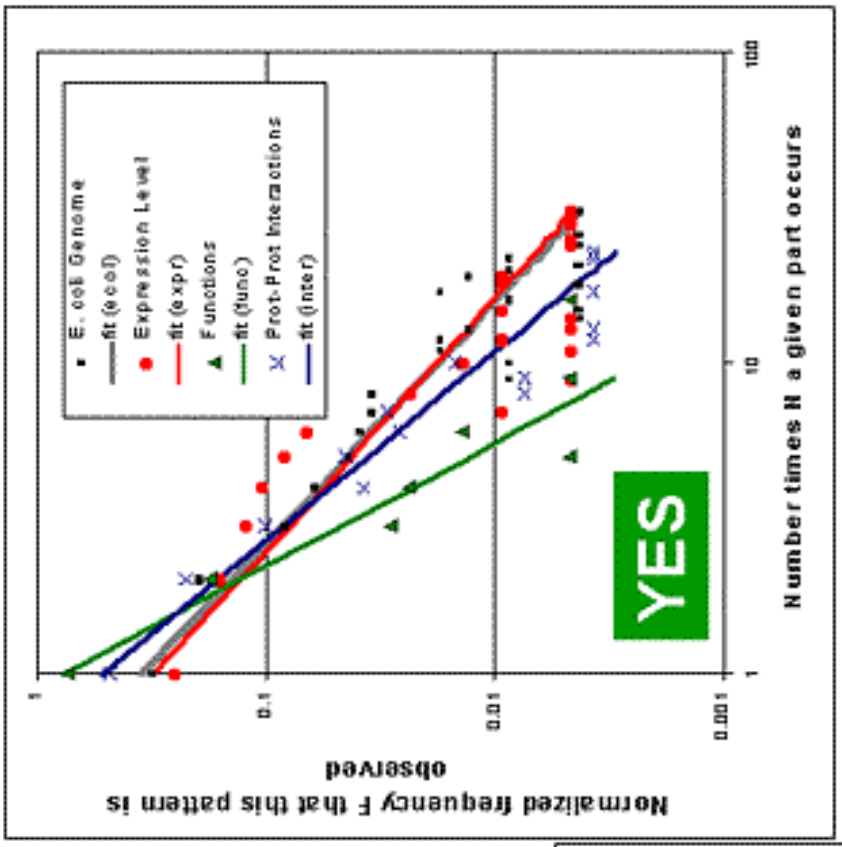
A few occur many times

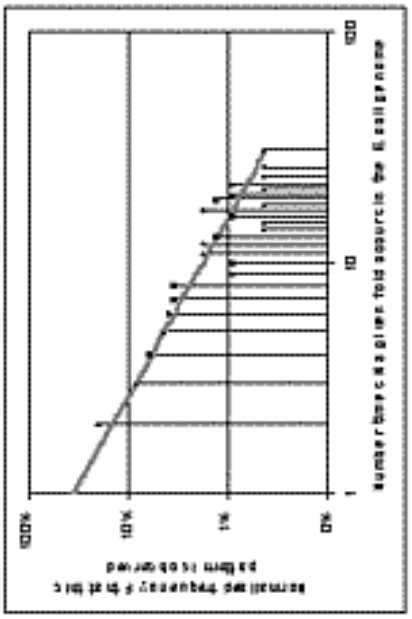


Overall falloff follows a Power Law (Zipf law, "80-20" rule)

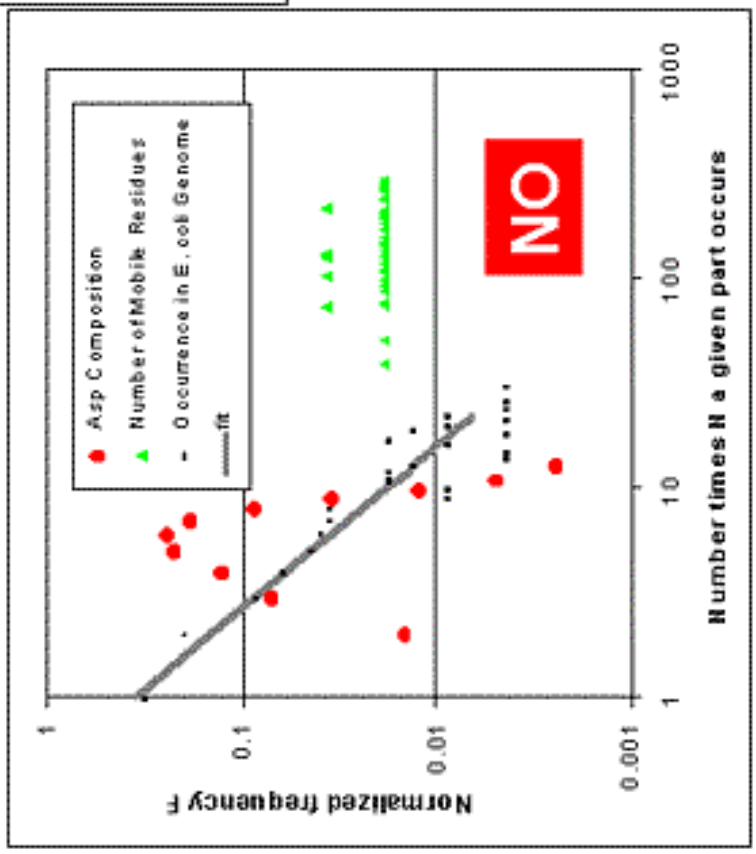
$$F = AN^{-B}$$

(A=.35, B=1.2 above)





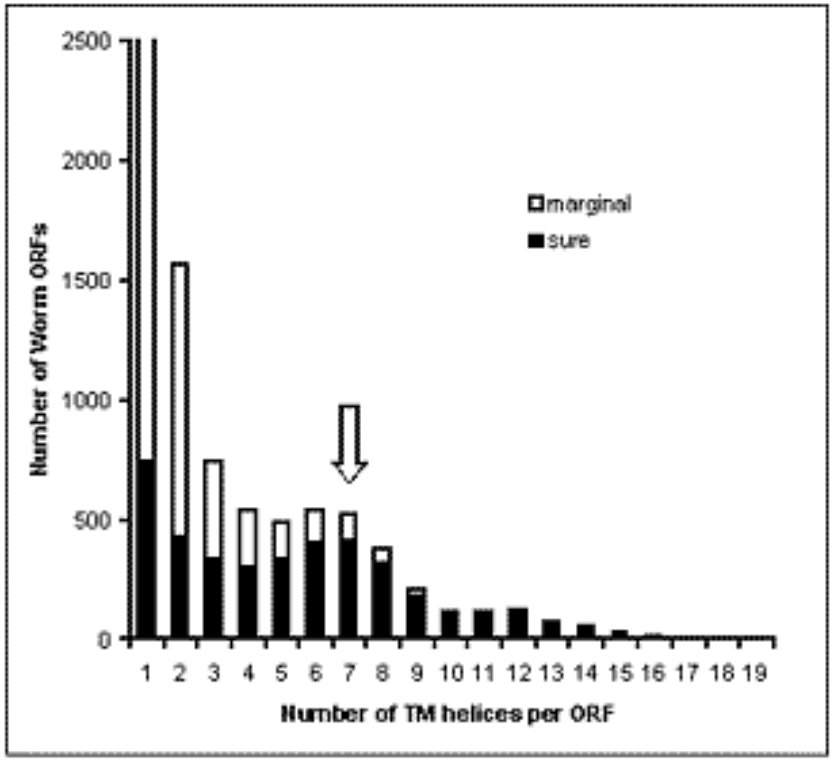
Many (but not all)
Part Quantities
Follow Power Laws

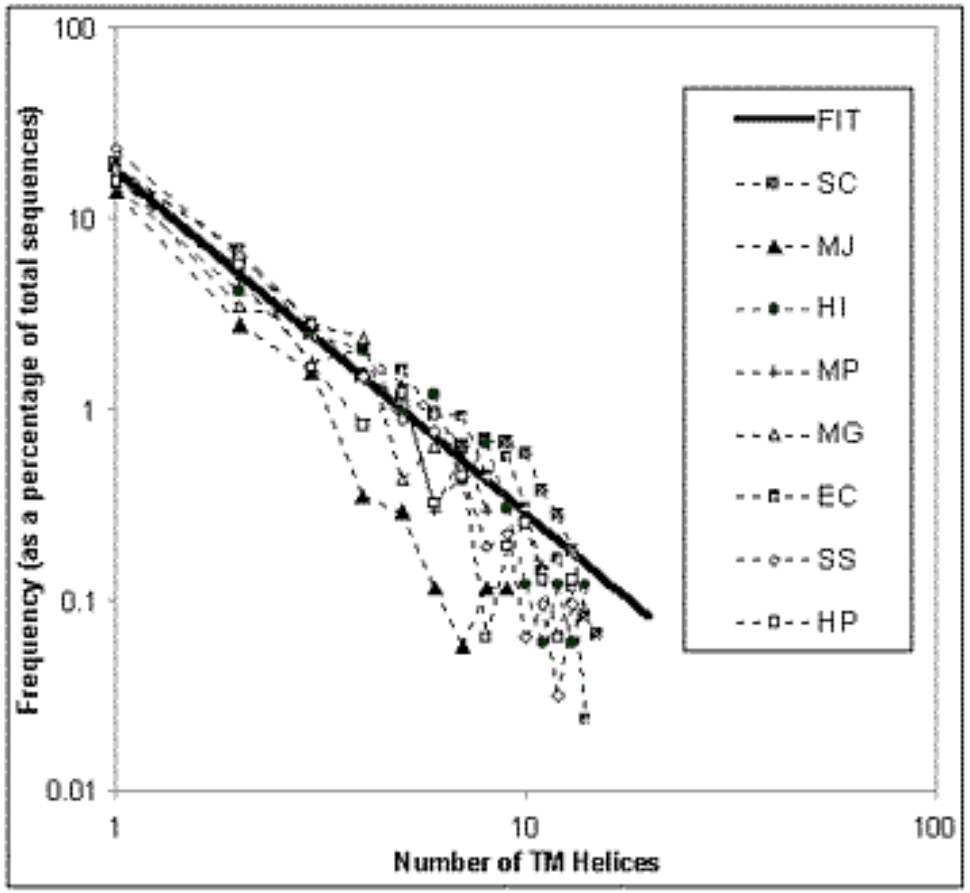


Occurrence of TM-helices follow Zipf Law

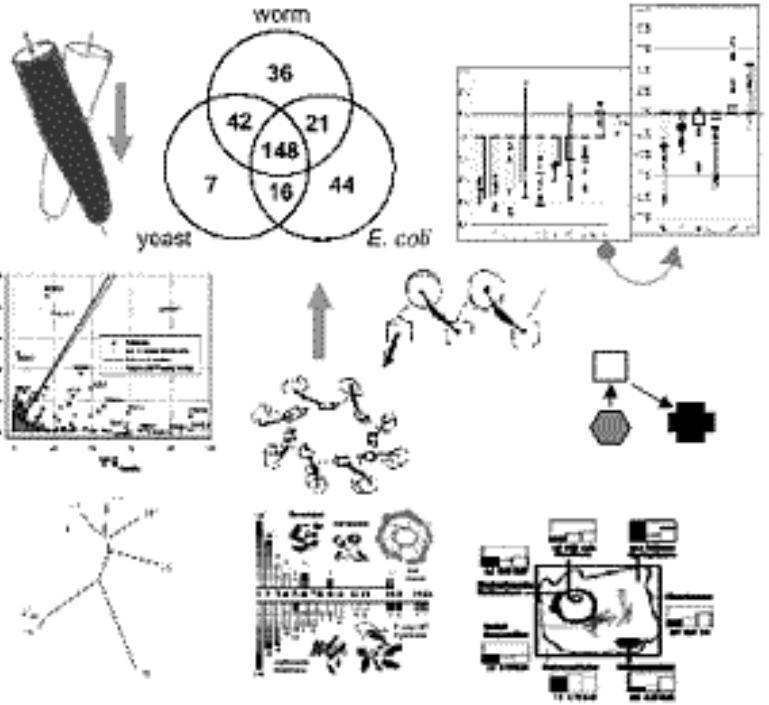
- TM identification (KD, GES).
  Count number with 2 peaks, 3 peaks, &c.
- Similar conclusions to others: von Heijne, Rost, Jones, &c.

- Divide into sure and marginal
  (Boyd & Beckwith's criteria)
- **$F = 1/[5N^2]$**





Integrative Genomics: Surveys of a Finite Parts List



Using Parts to Interpret Genomes

Shared & Common parts: Venn Diag. Whole-genome trees, top-10 with βαβ.

Ψ-genes

Folds/func? A few versatile scaffolds (TIM).

Using Parts & Categories to Mine Expression Data

Folds: Top-10 in expression (TIM)
Localization: Bayesian framework
Function: Is there a relation?
Interactions: Permanent cplx. vs other types

Integrated Views based on Parts

H Hegyi, J Lin, N Echols, P Harrison, M Levitt, C Wilson, R Das, A Drawid, R Jansen, D Greenbaum, M Snyder, S Teichmann, P Bertone, B Stenger, J Tsai, C Wilson, V Alexandrov, J Qian, W Krebs, M Snyder

bioinfo.mbb.yale.edu