<u>Analysis of Genomes &</u> <u>Transcriptomes in terms of the</u> <u>Occurrence of Parts and Features:</u>

Surveys of a Finite Parts List

Mark Gerstein Molecular Biophysics & Biochemistry and Computer Science, Yale University

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, J Qian, A Drawid, D Greenbaum, R Jansen

> Transcriptome 2000, Paris 8 November 2000



Simplifying the Complexity of Genomes: Global Surveys of a

Finite Set of Parts from Many Perspectives



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

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



Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

1 Using Parts to Interpret Genomes. Shared and/or unique parts. Venn Diagrams, <u>Fold tree</u> with all-β diff. Ortholog tree. Top-10 folds.

2 Using Parts to Interpret Pseudogenomes. In worm, top Ψ–folds (DNAse, hydrolase) v top-folds (Ig). chr. IV enriched, dead and dying families (90YG v 1G)

3 Using Parts to Interpret Transcriptomes: Expression & Structure. Top-10

parts in mRNA. <u>Enriched</u> in transcriptome: $\alpha\beta$ folds, energy, synthesis,TIM fold, VGA. <u>Depleted</u>: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

5 Expression & Function.

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



bioinfo.mbb.yale.edu

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder



PDB Report - Netscape

Results in Old Format

Fold 45 38 43 30 84 123 28 51 93 88 77 35 22

Occurrences

See Matches

in Genome

Superfamily

Occurrences

See Matches

in Genome

Edit View Go Communicator Help

🜮 Bookmarks 🛛 🦧 Location: http://bioinfo.mbb.yale.edu/genome/search.cgi?pdb=3

Formatted Text in Popup Window

Data Genome Atul Mjan Mthe Phor Scer Cele Aaeo Syne Ecol Bsub Mtub Hinf Hpyl Mgen Mpne Bbur Tpal Ctra Cone Roro

Chain. a Domain.

Genome Occurrence of 3tim's Fold

ID 3tim

Document Done

 - 🗆 ×

What's Related

Superfamily 3.1.11

See Matches in Genome

20 CA

SCOP Fold Number, 3.001

 12 20 15 19

 Disclaimer

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



<u>Cluster Trees Grouping Initial</u> <u>Genomes on Basis of Shared Folds</u>



9 (C

Distribution of Folds in Various Classes



<u>Compare with Ortholog Occurrence Trees</u> (Part = ortholog v fold)



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

Common Folds in Genome, Varies Betw. Genomes

Depen method (new to Blast, I get sha	ds on compariso d, DB, sfams v fo op superfamilies Intersection of to ared and commo	on olds, &c via ψ- op-10 to on)	Top-10 V	Vorm	Folds	class	num. matches in worm genome (N)	frac. all worm dom. (F)	in EC?	in SC?
 2		\rightarrow	lg			В	830	1.7%		
	D & A A	A)	Knottins			SML	565	1.1%		
		10 Mg 1990	Protein kinases	s (cat. co	ore)	MULT	472	0.9%		
			C-type lectin-lil	<e< td=""><td></td><td>A+B</td><td>322</td><td>0.6%</td><td></td><td></td></e<>		A+B	322	0.6%		
•	Maria		corticoid recep	. (DNA-ł	oind dom.)	SML	276	0.5%		
$\boldsymbol{\Lambda}$			Ligand-bind do	m. nuc.	receptor	Α	257	0.5%		
_			alpha-alpha su	perhelix		Α	247	0.5%		
			C2H2 Zn finge	r		SML	239	0.5%		
$\mathbf{\Delta}$	S S S S S S S S S S S S S S S S S S S		P-loop NTP Hy	rdrolase		A/B	235	0.5%		
\mathbf{V}			Ferrodoxin			A+B	207	0.4%		
	M. genitalium	B. subtilis	E. coli		M. thermo- autotrophicum	A. fulgi	dus	5	3. cerevisiae	
ank 1	Superfamily #	Superfamily (P-loop	Superfamily #	Rank	Superfamily #	Superfa	mily #	1	P-loop	# 249
2	SAM methyl- 16	Nydrolyase Mossmann 16	Rossmann 158	1	hydrolyase 93 Phosphate-	A hydroly Rossm	ann 104	2 X P	nydrofyase rotein kinase	123
3	Rossmann 13	Phosphate- binding barrel 7	9 Phosphate- binding barrel 64	3	binding barrel binding barrel Bossmann 53	doms Phosph	ate- 56	3 8	Rossmann domain	90

Δ	P-loop hydrolase	60	Δ	hydrolyase	173	Δ	P-loop hydrolase	191	1	Α	P-loop hydrolyase	93	Α	P-loop hydrolyase	118	105	Δ	hydrolyase	249
0	SAM methyl- transferase	16	\otimes	Rossmann domain	165	8	Rossmann domain	158	2	•	Phosphate- binding barrel	54	ø	Rossmann domain	104	2	х	Protein kinase	123
8	Rossmann domain	13		Phosphate- binding barrel	79		Phosphate- binding barrel	64	3	8	Rossmann domains	53		Phosphate- binding barrel	56	3	⊗	Rossmann domain	90
439832	Class I synthetase	12	٠	PLP-transferase	44	٠	PLP-transferase	38	4	ð	Ferredoxins	48	Ô	Ferredoxins	49	4		RNA-binding domain	75
	Class II synthetase	11	*	CheY-like domain	36	*	CheY-like domain	36	5	ò	SAM methyl- tranferase	17	Ò	SAM methyl- tranferase	24	5	0	SAM methyl- transferase	63
	Nucleic acid binding dom.	11	0	SAM methyl- transferase	30	٥	Ferredoxins	35	6	٠	PLP-transferases	15	٠	PLP-transferases	18	6		Ribonuclease H- like	57
S.		479	100.00		4268			4268	Total ORFs	<u> </u>		1869	Ť		2409	Total ORFs			6218
26		105			465			458	with Common	(e		252			309	with Common			560
÷.		(22%)			(11%)	-		(11%)	Superfamilies	3		(14%)	-		(13%)	Superfamilies	3		(9%)
	8	 P-loop hydrolase SAM methyl- transferase Rossmann domain Class I synthetase Class II synthetase Nucleic acid binding dom. 	P-loop hydrolase 60 SAM methyl- transferase 16 Rossmann 13 Class I 12 Class I 12 Class II 11 Nucleic acid binding dom. 11 479 105 (22%)	P-loop hydrolase 60 SAM methyl- transferase 16 Sesmann 13 Class I 12 Class I 12 Class I 12 Class I 11 synthetase 11 Nucleic acid binding dom. 11	P-loop hydrolase SAM methyl- transferase SAM	P-loop hydrolase 60 A hydrolyase 173 SAM methyl- transferase 16 Nossmann domain 165 Rossmann domain 13 Phosphate- binding barrel 79 Class I synthetase 12 PLP-transferase 44 Class II synthetase 11 SAM methyl- transferase 30 Nucleic acid binding dom. 11 SAM methyl- transferase 30 479 4268 105 465 (22%) (11%)	P-loop hydrolase 60 A hydrolyase 173 A SAM methyl- transferase 16 Bossmann domain 165 S Rossmann domain 13 Phosphate- binding barrel 79 ● Class I synthetase 11 PLP-transferase 44 ● Class II synthetase 11 SAM methyl- transferase 30 ● Vucleic acid binding dom. 11 SAM methyl- transferase 30 ● 479 4268 465 465 465 465	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 165 Rossmann domain Phosphate- binding barrel 79 Phosphate- binding barrel Phosphate- binding barrel Class I synthetase 12 PLP-transferase 44 PLP-transferase PLP-transferase Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins Ferredoxins 479 4268 465 11% 11% 11% 11% 11%	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 165 Rossmann domain 158 Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 479 4268 4268 4268 4268	A P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 SAM methyl- transferase 16 Bossmann domain 165 Rossmann domain 165 Rossmann domain 158 2 Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Class II synthetase 11 CheY-like domain 36 ‡ CheY-like domain 36 5 Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 105 465 458 with Common (22%) (11%) (11%) Superfamilies	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 ● Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Image: Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Image: CheY-like domain 36 5 Image: CheY-like domain 36	P-loop hydrolase 60 A hydrolyase 173 P-loop hydrolase 191 1 P-loop hydrolyase SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 9 Phosphate- binding barrel 64 3 Rossmann domains 158 2 Rossmann domains 165 Rossmann domain 158 2 Rossmann domains 8 8 4 Phosphate- binding barrel 64 3 Rossmann domains 8 8 4 Percedoxins 8 8 4 Percedoxins 8 8 8 4 Percedoxins 8 8 8 8 9 9 9 10 <td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolase 93 SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17 Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 15 479 4268 4268 4268 With Common 252 (122) (115) (115) Superfamilies (14%)</td> <td>A P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop 93 A SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 S Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 • Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 • Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17 • Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 15 • 479 4268 4268 4268 Total ORFs 1869 • • • • • • • • • • • • • • • • •</td> <td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 P-loop 93 A P-loop hydrolyase 11 A P-loop hydrolyase 93 A P-loop hydrolyase hydrolyase Possmann hydrolyase 93 A hydrolyase hydrolyase<td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 93 A P-loop hydrolyase 118 SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 104 Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Phosphate- binding barrel 56 Class I synthetase 11 RoeY-like domain 36 4 Ferredoxins 48 Ferredoxins 49 Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17 SAM methyl- tranferase 24 Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 15 PLP-transferases 18 105 465 4268 4268 tranferase 15 PLP-transferases 16 111%) (11%) (11%)</td><td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 118 Holop hydrolyase 104 2 Class I 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Ferredoxins 49 4 Class II 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17</td><td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 118 P-loop hydrolyase <th< td=""><td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 A hydrolyase 93 A hydrolyase 118 I A hydrolyase SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 104 2 X Protein kinase Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Phosphate- binding barrel 56 3 Rossmann domain Rossmann domain 104 2 X Protein kinase Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Ferredoxins 49 4 SAM methyl- transferase 24 5 SAM methyl- transferase 18 6 Ribonuclease H- like Ucleic acid binding</td></th<></td></td>	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolase 93 SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17 Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 15 479 4268 4268 4268 With Common 252 (122) (115) (115) Superfamilies (14%)	A P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop 93 A SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 S Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 • Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 • Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17 • Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 15 • 479 4268 4268 4268 Total ORFs 1869 • • • • • • • • • • • • • • • • •	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 P-loop 93 A P-loop hydrolyase 11 A P-loop hydrolyase 93 A P-loop hydrolyase hydrolyase Possmann hydrolyase 93 A hydrolyase hydrolyase <td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 93 A P-loop hydrolyase 118 SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 104 Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Phosphate- binding barrel 56 Class I synthetase 11 RoeY-like domain 36 4 Ferredoxins 48 Ferredoxins 49 Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17 SAM methyl- tranferase 24 Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 15 PLP-transferases 18 105 465 4268 4268 tranferase 15 PLP-transferases 16 111%) (11%) (11%)</td> <td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 118 Holop hydrolyase 104 2 Class I 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Ferredoxins 49 4 Class II 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17</td> <td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 118 P-loop hydrolyase <th< td=""><td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 A hydrolyase 93 A hydrolyase 118 I A hydrolyase SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 104 2 X Protein kinase Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Phosphate- binding barrel 56 3 Rossmann domain Rossmann domain 104 2 X Protein kinase Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Ferredoxins 49 4 SAM methyl- transferase 24 5 SAM methyl- transferase 18 6 Ribonuclease H- like Ucleic acid binding</td></th<></td>	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 93 A P-loop hydrolyase 118 SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 104 Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Phosphate- binding barrel 56 Class I synthetase 11 RoeY-like domain 36 4 Ferredoxins 48 Ferredoxins 49 Class II synthetase 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17 SAM methyl- tranferase 24 Nucleic acid binding dom. 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 15 PLP-transferases 18 105 465 4268 4268 tranferase 15 PLP-transferases 16 111%) (11%) (11%)	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 118 Holop hydrolyase 104 2 Class I 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Ferredoxins 49 4 Class II 11 SAM methyl- transferase 30 Ferredoxins 35 6 PLP-transferases 17	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 1 A P-loop hydrolyase 118 P-loop hydrolyase <th< td=""><td>P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 A hydrolyase 93 A hydrolyase 118 I A hydrolyase SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 104 2 X Protein kinase Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Phosphate- binding barrel 56 3 Rossmann domain Rossmann domain 104 2 X Protein kinase Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Ferredoxins 49 4 SAM methyl- transferase 24 5 SAM methyl- transferase 18 6 Ribonuclease H- like Ucleic acid binding</td></th<>	P-loop hydrolase 60 A hydrolyase 173 A P-loop hydrolase 191 A hydrolyase 93 A hydrolyase 118 I A hydrolyase SAM methyl- transferase 16 Rossmann domain 165 Rossmann domain 158 2 Phosphate- binding barrel 54 Rossmann domain 104 2 X Protein kinase Rossmann domain 13 Phosphate- binding barrel 79 Phosphate- binding barrel 64 3 Rossmann domains 53 Phosphate- binding barrel 56 3 Rossmann domain Rossmann domain 104 2 X Protein kinase Class I synthetase 12 PLP-transferase 44 PLP-transferase 38 4 Ferredoxins 48 Ferredoxins 49 4 SAM methyl- transferase 24 5 SAM methyl- transferase 18 6 Ribonuclease H- like Ucleic acid binding

Eubacteria

Archaea

(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu 12

Yeast







ARTICLES

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 high 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 inter-

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









P-loop hydrolase



Flavodoxin

like





Rossmann Fold





Thiamin Binding

Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

1 Using Parts to Interpret Genomes. Shared and/or unique parts. Venn Diagrams, <u>Fold tree</u> with all-β diff. Ortholog tree. Top-10 folds.

2 Using Parts to Interpret Pseudogenomes. In worm, top Ψ–folds (DNAse, hydrolase) v top-folds (Ig). chr. IV enriched, dead and dying families (90YG v 1G)

3 Using Parts to Interpret Transcriptomes: Expression & Structure. Top-10

parts in mRNA. <u>Enriched</u> in transcriptome: $\alpha\beta$ folds, energy, synthesis,TIM fold, VGA. <u>Depleted</u>: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

5 Expression & Function.

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



bioinfo.mbb.yale.edu

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder

Pseudogenomics: Surveying "Dead" Parts



Folds in Pseudogenes



1,155 (30% of

ΨG)

Intronic

pseudogenes *

30%

351

704

61%

frameshift in mid-domain

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	***	d21bd 1.95 Nuc. receptor ligand-binding domain
2 (555)	6	er A	d1dec 7.3 Knottin	7 (243)	34	hu	d1a17 1.91 Alpha/alpha superhelix
3 (434)	3	AN CONTRACT	d31ck 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	S	d1zfo 7.33 Glucocorticoid receptor DNA- binding dom.	10 (197)	13	Ŵ	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	œ۶	d1dec 7.3 Knottin
2 (32)	1		d1ajw 2.1 Immuno- globulin	7 (17)	5	S	d1zfo 7.33 Glucocorticoid receptor DNA- binding dom.
3 (27)	3		d31ck 5.1 Protein kinase	8 (15)	6	**	d21bd 1.95 Nuc. receptor ligand-binding domain
4 (25)	11		d1cv1 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





(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu 20



<u>Completely Dead</u> <u>Families</u>

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

<u>Amino Acid Composition of Pseudogenes is</u> <u>Midway between Proteins and Random</u>



Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

1 Using Parts to Interpret Genomes. Shared and/or unique parts. Venn Diagrams, <u>Fold tree</u> with all-β diff. Ortholog tree. Top-10 folds.

2 Using Parts to Interpret Pseudogenomes. In worm, top Ψ–folds (DNAse, hydrolase) v top-folds (Ig). chr. IV enriched, dead and dying families (90YG v 1G)

3 Using Parts to Interpret Transcriptomes: Expression & Structure. Top-10

parts in mRNA. <u>Enriched</u> in transcriptome: $\alpha\beta$ folds, energy, synthesis,TIM fold, VGA. <u>Depleted</u>: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

5 Expression & Function.

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



bioinfo.mbb.yale.edu

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder Cell, Vol. 95, 717-728, November 25, 1998, Copyright ±1998 by Cell Press

Dissecting the Regulatory Circuitry of a Eukaryotic Genome

Frank C. P. Holstage, "Erra G. Jennings," John J. Wyrick, "Tong lim Lee," Christoph J. Hongartner," Michael R. Green, J. Todd R. Golub, "Entits 5. Lander," and Richard A. Young"11 "Whitehind Institute for Biomedical Research Cambridge, Nassachusetts 02142 "Department of Biology Massachusetts Institute of Technology Cambridge, Massachusetts 02133 "Hoseard Hughes Medical Institute Program in Motocular Madicine Umensity of Massachusetts Medical Center



Young, Church... Affymetrix GeneChips Abs. Exp.

specific transcription factors, a novel mechan monthly break when a realised is compared to be mapp 3.2 + 4 + 4 1 of some other gone's altered mRNA levels. The 50 I with that obtained by its inactivactor. Comparison of the two data sets sonce we have identified that receive Wedli function to The Brown Lab to with examplerit riscas kinetics in the same extent as Reis! function as these at which lose Technology, Prote conclar-associated transcriptional regulators are kely to function through interactions with Medil. alysis on the web site for details). For Stanford University Department of Biochemistry and by to classifien of Modil where Seb5 is a component of the Sebimediator camp an could be made, the mRHAs of 666 shose function is also not known (Thompson of a with similar kinetics in the MirdS and 1907: Way of al., 1904: Knieske and Voume TWO Kam int al. 1994: Knocke and Young, 1994; Inter gartner et al. 1996; Myars et al. 1990). Te determine the garome-while dependence of game expression in Set6, a sizale lacking an SMSH game and its weld-type Thus, the expression of 1016 of yeast anderst on ModG as they are on Rpb1. The MGuide are most likely to have a direct require The Complete Guide to MicroArrays knotion. The genus whose transcript counterpart, were compared boo the web site for d the Robil kinetics could have a direct, sevent for Medil function, on the effects balled information). The results indicate that 10% of all genes require SrbS function for their expression. With the SRUS delifion strain and other constitutive materits. Build your own arrayer and scanner! control ato a sacondary consequence The transcriptional program in the response human fibroblasts to serum Also: The Transcription of Sporulation in The Web Company Brown,

Brown, marrays, Rel. Exp. over Timecourse <u>Also</u>: SAGE (mRNA); 2D gels for Protein Abundance (Aebersold, Futcher)

<u>Gene Expression</u> <u>Datasets: the Yeast</u> <u>Transcriptome</u>

Yeast Expression Data: 6000 levels! Integrated Gene Expression Analysis System: X-ref. Parts and Features against expression data...



and pair when succed potential at the set of the set of

Op. the exclusion of explanation from the NEAL Star. Op. the exclusion of explanation is advanted by wrange bases, not 114-57 memory benchmarks uport and physical procession and oracle and basics of the is induced by a black or mender, and ware 144-57 age data series.

<u>Common Parts: the Transcriptome</u>

	×14					Co	mpositic	on				F	Rank				
		Fold		Fold Class	Rep. PDB	Genome [%]	Transcriptome [%]	Rel. Diff. [%]		Genome	Young	Samson	Church-alpha Church-alpha Church-aal	Church-heat	SAGE-GM	SAGE-L SAGE-S	
5	(R) 1 🔪 -	TIM barrel		oڧ	1 hyb	4 2	83	+98		5	1	1 1	1 1	1	1	1 1	
		P-loop NTP hydrol	ases	α/β	laky	5.8	5.2	-11	+	3	2	2 4	44	5	5	6 7	
_	🛌 🖉 📕	Ferredoxin like		α ιβ	lfxd	3.9	3.4	-14	•	6	3	7 1	1 9 8	10	4	10 11]
$\mathbf{\Sigma}$	M /• [Rossmann fold		α/β	1xel	3.3	3.3	0		8	4	3 3	3 3 2	2	19	15 9	
		7-bladed beta-prop	beller	β	1mda*	6.4	2.9	-55		2	5	4 5	566	7	9	9 16	3
0		aplha-alpha superl	nelix	α	2bct	4.4	2.7	-37	-	4	6	11 1	5 16 12	12	8	58	
		Thioredoxin fold		α/β	2trx	1.7	2.7	+63	+	14	7	6 8	B 2 5	4	11	10 6	
		G3P dehydrogena	se-like	Cttβ	ldrw†	0.2	2.7	+1316	÷	81	8	12	2 5 3	3	35	19 30)
		HSP70 C-term_fra	ament	multi	liga idky	0.6	2.0	+348	÷	36	9	10 2	1 9 18 7 11 16	21	82 1	122 12 25 56	5
	/ É		ginon	mana	rany	0.0	2.0	1201	+	51	10	10 1		12	10	20 00	'
	/ _•[Leu-zipper		α	1zta	3.8	2.1	-46	-	7	15	8 1	4 21 15	19	21	20 33	3
		Protein kinases (ca	at. core)	multi	lhcl	6.8	1.6	-77	-	1	18	19 5	9 16 11	15	13	16 17	7
		alpha/beta hydrola	ses	α/β	2ace	2.2	0.9	-62	-	10	32	31 2	5 26 21	23	26	26 26	3
	1®18	Zn2/C6 DNA-bind.	dom.	sml	1aw6	2.6	0.3	-89		9	75	94 2	7 50 32	40	48	39 50)
4	Feature <i>F</i> is Folds, in particular the TIM-barrel (3.1)	Number of TIM-barrel fold matches in yeast genome	Number matches all folds i yeast ge	of with n nome	Genom compo TIM-ba fold ma	ne sition of arrel atches	Numbe TIM-ba fold ma weighte express	r of rrel tches ed by sion	Nu ma all we ex	umber atches folds eighteo pressi	of with I by on	Transo compo TIM-ba fold ma	criptome osition of arrel atches	Rel enr TIM mat	ative ichm 1-bar tche: 1scrij	ent of rel s in ptome	F
	Spec. Num.	65	156	50	4.2	2%	38	39		470)9	8.	3%	5	97.8	8%	

7®

		Fre	eq.	С	ha	ng	e	ъ			Folds
A CONTRACTOR	Fold of	Genome	Transcriptome	CDC28	CDC15	Diauxic Shift	Sporulation	E. co <i>li</i> heat sho	Rep. PDB	Folds	that
1000 0	Protein kinases (cat. core)	1	18	94	98	139	60	100	1p38	K KS	> <u>change</u>
	β-propeller	2	5	160	108	109	82		1mda	1000 miles	
(all)	P-loop NTP hydrolases	3	2	100	88	91	57	39	1gky	n.	
r l	α-α superhelix	4	6	136	90	110	44	55	2bct		10
The /	TIM-barrel	5	1	58	57	39	24	81	1byb	•/	<u>in</u>
80,0 9000	Ferredoxin-like	6	3	135	61	63	70	144	1fxd	•	· -
\$\$/ 	Rossmann fold	8	4	55	99	43	56	92	1xel		trequency
	Ribonucleotide reductase (R1)	100	143	1	+		-	35	1rlr	—	
	ATPase dom. of HSP90	100	91	2	4	72	73	2	1ah6	•	are
	Homing endonuclease-like	130	164	3	136	85	175	- 41	1af5	\backslash	
-	Aminoacid dehydrogenases; dim. dom		-	4	169	121	3	51	1hup	(D)	not
BK-	DNA topo I (N-term)	1		175	1	148	126		lois	ACON.	<u>1101</u>
	DNA clamp	130	115	8	2	87	11	60	2pol	189	common
Ð	Metallothionein	100	14	89	3	33	12		1mhu		
	Phosphoenolpyruvate carboxykinase	130	190	51	26	1	98	169	1ayl		
A CONTRACTOR	Citrate synthase	81	120	14	8	2	28	51	1csh		\geq
CTES .	N-carbamoylsarcosine amidohydrolase	130	112	9	8	3	138	118	1nba		
6 1 V	TBP-like	81	91	46	38	4	75	100	1bv1	States	
	5'-3' exonuclease	67	150	32	125	162	1	157	1tfr	HAR	
	α/α toroid	62	132	169	145	114	2	100	1gai	~ V. 3 8 8	Changing
~ & ~	Cyclin-like	20	61	20	15	129	4		1vin		
and a	ATPase domain of GroEL	36	34	183	143	61	151	1	1aon	İ	Folds
- Carlos	Head domain of GrpE	130	135	196	31	165	165	3	1dkg		<u> </u>
	HSP70 (C-term)	31	10	16	11	55	117	4	1dkz		

Most Versatile Folds – Relation to Interactions



Composition of Transcriptome in terms of



Composition of Genome vs. Transcriptome

	$\sum_{\text{off } i} n_i(F)$	$\sum_{F} \sum_{\text{orf } i} n_i(F)$	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 <i>F</i> 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 <i>F</i> 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



Relating the Transcriptome to Cellular Protein Abundance (Translatome)



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

1



Simple story is translatome is enriched in same way as transcriptome

<u>Amino Acid Enrichment –</u> <u>Complexities</u>



1

(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu 34

Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

1 Using Parts to Interpret Genomes. Shared and/or unique parts. Venn Diagrams, <u>Fold tree</u> with all-β diff. Ortholog tree. Top-10 folds.

2 Using Parts to Interpret Pseudogenomes. In worm, top Ψ–folds (DNAse, hydrolase) v top-folds (Ig). chr. IV enriched, dead and dying families (90YG v 1G)

3 Using Parts to Interpret Transcriptomes: Expression & Structure. Top-10

parts in mRNA. <u>Enriched</u> in transcriptome: $\alpha\beta$ folds, energy, synthesis,TIM fold, VGA. <u>Depleted</u>: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

5 Expression & Function.

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



bioinfo.mbb.yale.edu

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder



Expression Level is Related to Localization





~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.

<u>Bayesian</u> <u>System for</u> <u>Localizing</u> <u>Proteins</u>

loc=



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



18 Features: Expression Level (absolute and fluctuations), signal seq., KDEL, NLS, Essential?, aa composition <u>Bayesian</u> <u>System for</u> <u>Localizing</u> <u>Proteins</u>

loc=



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



(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu 4







Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

1 Using Parts to Interpret Genomes. Shared and/or unique parts. Venn Diagrams, <u>Fold tree</u> with all-β diff. Ortholog tree. Top-10 folds.

2 Using Parts to Interpret Pseudogenomes. In worm, top Ψ–folds (DNAse, hydrolase) v top-folds (Ig). chr. IV enriched, dead and dying families (90YG v 1G)

3 Using Parts to Interpret Transcriptomes: Expression & Structure. Top-10

parts in mRNA. <u>Enriched</u> in transcriptome: $\alpha\beta$ folds, energy, synthesis,TIM fold, VGA. <u>Depleted</u>: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

5 Expression & Function.

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



bioinfo.mbb.yale.edu

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder







			Exper	iment	
		Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo- rulation
	Cell growth, di- vision & 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
agor	Intracellular transport	>4	>4	0	0
cate	Ionic homeostasis	>4	>4	0	0.8
IPS	Metabolism	>4	>4	0	0
Σ	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

	Fractio	on of sig	nificant	groups	Total #
	CDC28	CDC15	Diauxic Shift	Sporu- lation	groups
MIPS 1	63%	81%	19%	13%	16
MIPS 2	50%	63%	17%	13%	102
MIPS 3	23%	33%	5%	4%	73
"Energy" (2 nd level)	40%	60%	20%	0%	10
SOM	93%	-	-	-07	30
Clustering		8	0%		25

						Correlation:
1	54		Exper	iment		
[Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo- rulation	Always Significant
	Respiration	>4	>4	>4	3.4	Significant
	TCA pathway	>4	>4	>4	0.6	
	Glycogen, trehalose metabolism	>4	>4	1.2	0.7	Sometimes
≥	Glycolysis	>4	>4	0.9	2.1	Significant
atego	Gluconeogenesis	3.7	>4	0.1	1.7	(depends
Sci	Glyoxylate cycle	1.6	0.7	3.0	2.3	on expt.)
M	Pentose-phosphate pathway	1.5	0.8	0	0.6	
	Fermentation	1.3	>4	0	2.2	Novor
	Other energy generation activities	0.7	0.1	0.1	0.2	
1	Beta-oxidation of fatty acids	0.5	0.4	0.4	0.2	Significant
2 2	sod o	nΓ		tril	hut	ione
Ja	<u>360 0</u>		713		Jui	<u>.iuiis,</u>
	Cor	rel	ati	on	of	

Established Functional

Categories, Computer

Clusterings

orrelation:

metimes gnificant Never gnificant (c) Mark Gerstein, 2000, Yale, bioinf 49



(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu 50



Can we define FUNCTION 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'



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



VS

YPD + 8mM caffeine	Caff				۱۸	16
Cycloheximide hypersensitivity: YPD + 0.08 ?g/ml cycloheximide at 30°C	Cyc ^s				V	VI
White/ red color on YPD	W/R					
YPGlycerol	YPG				Ph	∩ r
Calcofluor hypersensitivity: YPD + 12 ?g/ml calcofluor at 30°C	Calc ^s			٨		
YPD + 46 ?g/ml hygromycin at 30°C	Hyg				Tuene	
YPD + 0.003% SDS	SDS			0	Trans	pos
Benomyl hypersensitivity: YPD + 10 ?g/ml benomyl	Ben ^s			Ö	each	yea
YPD + 5-bromo-4-chloro-3-indolyl phosphate 37°C	BCIP			д +	affect	ed i
YPD + 0.001% methylene blue at 30°C	MB			t F	nhonc	stur
Benomyl resistance: YPD + 20 ?g/ml benomyl	Ben ^R			0 H	phenc	лур
YPD at 37°C	YPD ³⁷			Ø	be tre	ate
YPD + 2 mM EGTA	EGTA				data	
YPD + 0.008% MMS	MMS			V		
YPD + 75 mM hydroxyurea	HU					- 1
YPD at 11°C (COLD)	YPD ¹¹					- 1
Calcofluor resistance: YPD + 66.7 ?g/ml calcofluor at 30°C	Calc ^R					
Cycloheximide resistance: YPD + 0.3 ?g/ml	Cyc ^R					- 1
Hyperhaploid invasive growth mutants	HHIG					- H
YPD + 0.9 M NaCl	NaCl					
M Snyder	ected	YBR102c	YER021W	Affe	ected	l
Cor	ndition			by	Cold	

<u>nole Genome</u> notype Profiles

son insertions into (almost) ast gene to see how yeast is in 20 conditions. Generates a be pattern vector, which can ed 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," crossref. to MIPs Functional Classes

(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

54

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



Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

1 Using Parts to Interpret Genomes. Shared and/or unique parts. Venn Diagrams, <u>Fold tree</u> with all-β diff. Ortholog tree. Top-10 folds.

2 Using Parts to Interpret Pseudogenomes. In worm, top Ψ–folds (DNAse, hydrolase) v top-folds (Ig). chr. IV enriched, dead and dying families (90YG v 1G)

3 Using Parts to Interpret Transcriptomes: Expression & Structure. Top-10

parts in mRNA. <u>Enriched</u> in transcriptome: $\alpha\beta$ folds, energy, synthesis,TIM fold, VGA. <u>Depleted</u>: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

5 Expression & Function.

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



bioinfo.mbb.yale.edu

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder





6 RA	NKINGS 30 H	lighest Va	lues	cele	0 <u>cc</u>	urrend	es - N	etsca	pe _			_		_				_					N	on ze	ro hits	
jie E	<u>i</u> dit ⊻iew <u>G</u> o	Communi	cator	Help	10.25		1960		1000	25623	(AND)	100		2.56	120720	12450	02805	1245					R	ank	again	
YAL	E GERS	Local	ion hit B	p://bio	Fol	d Or	edu/ali	ren	kings/?	updow	n=up&s	ubcate	egory=F	olds&ca	slegory	=OCCUI	iencest	irankbj		d Ir C	<mark>lajw</mark> muun lass: .	il oglo All b	.002.001 bulin-like eta protei	beta ns	sandwich	
	• <u>Cae</u>	enorhai Fold	bditi	s el	egan Alignu	ns ment:		0	ccur	rend	es ir	1 Ge	nom	es			⊙ Fol	ds		d K C	<mark>1 dec</mark> nottis lass: 1	ins (S Smal	. 007 003 mall inhib 1 Proteins	itors		
⊙ 30 : igheat	Exp Highert O 100 5 Ranked	interssion Andress O I	op Hulf				Ом	dle 30	Омы	(die 100	© Осс О ми	urren. Me Hal	ces £		C	Осси	arence 9 Lowest	s Fold		d P C	<mark>31ck</mark> rotein lass: 1	:1 kina Mult	005.001 ases (PK) i-domain	(alph	a and beta)	
Out	Representative	20 Genomes	Aful	Mian	Mille	Phor	Scer	Cele	Azes	Syne	Ecol	Bask	Mtub	Him	Hext	Maen	Mone	Bbur		P	rotem	IS .				-
of 248	Domain	\Leftrightarrow	⇔	⇔	⇔	⇔	⇔	9	⇔	⇔	⇔	⇔	⇔	⇔	⇔	⇔	⇔	⇔		dic	1tsg	il.	004.105			
1	dlajv	881	2	2	2	0	4	829	2	11	18	2	2	0	2	0	0	2		č	lass:	Alph	a plus be	ta pr	oteins	1
2	didec	576	0	0	0	3	3	558	1	1	0	2	1	2	0	1	1	2				сэ. 				E
3	d31ck	636	2	3	2	3	129	454	1	13	1	3	15	0	1	1	1	0		o c	d1zto :1.007.033 Obsessetiseid recenter like					
4	ditsg	323	0	0	0	0	1	322	0	0	0	0	0	0	0	0	0	0		a	(DNA-binding domain)					
5	dizfo	297	0	0	0	1	10	284	0	0	1	0	0	0	0	0	1	0		C	lass: :	Smal	1 Proteins			2
6	d2 1bd	257	0	0	0	0	0	257	0	0	0	0	0	0	0	0	0	0		d	2lbd	:1	.001.095			
7	d1a17	494	0	21	7	4	114	250	25	42	6	1	0	1	0	0	0	6		L	igand	-bin	ding dom	in of	nuclear	
8	d1sp2	321	1	2	0	2	77	238	0	0	0	0	0	1	0	0	0	0		c	lass:	Alla	lpha prote	ins		
9	didai	1301	56	62	61	61	130	223	57	68	73	64	54	51	46	34	36	69			1.17		001.001			
10	d2aw0	1426	93	87	583	21	114	207	31	38	83	27	44	34	21	3	3	6		alpha-alpha superhelix						
11	dicvl	420	13	1	2	5	39	175	3	24	19	33	70	5	1	4	5	2		Class: All alpha proteins						
12	digotb	323	0	0	0	0	140	168	0	10	0	0	1	1	0	0	0	0								H
13	ditop	138	0	0	0	0	0	138	0	0	0	0	0	0	0	0	0	0		₫	1sp2	_:1	.007.031			
14	dleny	785	31	16	26	19	72	132	33	59	79	98	126	25	21	4	5	5		C	lassic lass: 1	zino Smal	finger Proteins			
15	<u>d1a5j 1</u>	315	6	10	6	7	31	130	4	15	24	23	29	13	4	1	1	1	3	-			0	Docu	ment: Done	
16	d1aj2	858	45	38	43	30	84	123	28	51	93	88	77	35	22	7	12	20	15	19	19	9	1.003.001	α/β	beta/alpha (TIM
16	dibor	152	0	0	0	0	29	123	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.007.037	s	RING finger	don
18	dildr	117	0	0	0	0	0	116	0	0	0	0	0	0	0	0	0	0	0	1	0	0	<u>1.007.011</u>	s	Ligand-bind low-density receptor	ling (
19	direc	129	0	0	0	0	14	112	0	1	0	0	0	0	2	0	0	0	0	0	0	0	1.001.037	α.	EF Hand-like	e
20	dipty	122	0	0	0	0	10	111	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1.003.037	α/β	(Phosphotyn phosphatase	rosir es II
							the second se																			

Edit View <u>G</u> o <u>Communicator</u> Help				lon di	A 808 920						
LE GERSTEIN LAB		searc	search GO CO A BEE AN								
Rankings ²											
View the first 30 folds View the entire table		fold occurrence in Scer	fold occurrence in Cele	fold percentage in reference genome	fold percentage in reference transcriptome						
Max of all		140	829	11.3	11.1						
Min of all		0	0	0.1	0						
Average		6.9	22.9	0.5	0.5						
Non zero hits		215	247	213	134						
Rank again		\$	_	\Rightarrow	\Rightarrow						
dlajw_:1.002.001 Immunoglobulin-like beta-sandwich Class: All beta proteins	¥.	4	829	0.2	0.2						
dldec :1 007 003 Knottins (Small inhibitors Class: Small Proteins	e P	3	556	÷.	i.						
d3lck:1 005 001 Protein kinases (PK) Class: Multi-domain (alpha and beta) proteins	AND	129	454	6.4	0.9						
<mark>11tsg</mark> :1.004.105 -type lectim-like Jass: Alpha plus beta proteins	**	1	322	#12 200							
<mark>11zfo11007.033</mark> Hucocorticoid receptor-like DNA-binding domain) Jass: Small Proteins	Š	10	284	0.4	0.1						
d2lbd:1.001.095 Ligand-binding domain of nuclear receptor Class: All alpha proteins	*	0	257	(i)	4						
dlal7_:1.001.091 alpha-alpha superhelix Class: All alpha proteins	hi:	114	250	4.3	2.3						
dlsp2:1.007.031 Classic zinc finger Class: Small Proteins	-	77	238	1.4	0.2						
Document: Done					1 d9 🖬						
19 9 1.003.001 a/b beta/alpha (T 0 0 1.007.037 s RING finger d 0 0 1.007.011 s Ligand-bindit 0 0 1.007.011 s Ligand-bindit	IM)-barrel Iomain ng domain of ipoprotein	R	ank F	Folds	by G						

13.037 α/β (Phosphotyrosine) protein phosphatases II

] 🐘 🚨 🗗 🖪 🖋]

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



ne Occurrence, Expression, Fold Clustering, Length, &c

Surveying a Finite PartsList from Many Perspective



(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu 58

<u>GeneCensus Dynamic</u> <u>Tree Viewers</u>

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

The law of	GeneCensus TreeServe	er - Netscape			-100v		X	GeneCensus TreeServe	ar - Netscape					
Image: The stand in the stand procession of the	Elle Edit View Go Comm	nunicator <u>H</u> elp					E	ile <u>E</u> dit <u>V</u> iew <u>G</u> o <u>C</u> omm	nunicator <u>H</u> elp					
VALE GERSTEIN LAB Pranch Op 34 A BLA OVE GENECENSUS Vale GERSTEIN LAB Pranch Op 34 A BLA OVE GENECENSUS Vale GERSTEIN LAB Pranch Op 34 A BLA OVE Genecensus Community and the second of the second o	👔 🌿 Bookmarks 🥠 Lo	cation: http://bioi	info.mbb.yale.edu/genecensus/l	ee.cgi?view=1&type=comp-d	śrł-per	👻 🎲 What's Related 👔	1	🏾 🌿 Bookmarks 🏼 🎄 Li	cation: http://b	ioinfo.mbb.yale	.edu/genecensu	s/		Vhat's Related
	YALE GERSTEIN	AB			search G	o dio 📥 🧱 🎘 🛛 🖬		YALE GERSTEIN	AB			s	earch	💿 ॐ 本 🚟 🕷 😗 🖥
Constant of the constant	GENECENS	SUS						GENECENS	SUS					
Devention de la constanta de			ribosome folds	stams COG	is composition ORFs					ribo	some fol	ds stams COGs	composition ORFs	/
Tree Based On Dinucleotide Composition Percentages Tree desced on Dinucleotide Composition Percentages Tree desced on Dinucleotide Composition Percentages Dimension Paradia Bill Bandwide Bill Bandwi			UINUCIRODAR ZUA _ AMI	no And 2014 <u>pinudeonde 2</u>	ay <u>Amino Asigʻzy</u>						<u>LSU N</u>	<u>nn _ 8 550 irnn _20 550 irnn</u> _	_	
Tree Based On Dinucleotide Composition Percentages Littereter Transfer Tree Journal of Dinasterine Jacobie Bold			20 0001149 20 99 20 14	N CARING CRADE	<u>o ami 1877 o ami</u>			Tran D		- Carl		imilarity of Omo	II. Cubunit Diba	annal DNA
Interference intervery constantion bill accente intervery Intervery <th></th> <th>Tree Do</th> <th>and On Dinua</th> <th>actida Comp</th> <th>anitian Daraantag</th> <th></th> <th></th> <th>Tree Ba</th> <th>ased O</th> <th>n Sequ</th> <th>ience S</th> <th>imilarity of Sma</th> <th>II Subunit Ribo</th> <th>somal RNA</th>		Tree Do	and On Dinua	actida Comp	anitian Daraantag			Tree Ba	ased O	n Sequ	ience S	imilarity of Sma	II Subunit Ribo	somal RNA
Tre Licentral Oppoint Careton unerer inspiration to account account of the control of the contro		пее ва	ised On Dinuci	eotide Comp	osition Percentag	es					SelectView: <mark>†</mark>	axonomy composition fold s	superfamily oldCOG newCC	G rRNAtree
The identified Departure Devalue Deval			El SelectView: tax	enomy <u>composition</u> fold	d superfamily oldCOG newCOG	rRNAtree		Tree [download]	Organism	Yale Resources	Superkingdom	Kingdom	Phylum	Class
Asso Galance Asso Association	Tree [download]	Organism _R	Tale Superkingdom	Kingdom	Phylum	Class			- Cele		E	Metazoa	Nematoda	Chromadorea
Phone D Expressionality Termicolation Ter		- Azeo	• •	Aquificales	Aquificaceae	Aquifex	Ш.		Scer		Ē	Fungi	Ascomycota	Hemiascomycetes
Basis Emission Basiliux/Lastidum, group Basiliux/Statidutoscocu, gro Aulu A Euryacheesta Muturolitacteristica Muturolitacteristica Aulu A Euryacheesta Muturolitacteristica Muturolitacteristica Muturolitacteristica Aulu A Euryacheesta Muturolitacteristica Muturolitacteristica Muturolitacteristica Muturolitacteristica B Descendance Adulitata Adulitata Adulitata Muturolitacteristica B Euroacheesta Muturolitacteristica Adulitata Muturolitacteristica B Descendance Adulitata Adulitata Muturolitacteristica Basilitacian Adulitata Adulitata Adulitata Muturolitacteristica Basilitacian Basilitacian Adulitata Adulitata Muturolitacteristica Basilitataciana <t< th=""><th></th><th>-Phor</th><th></th><th>Eurvarchaeota</th><th>Thermococcales</th><th>Thermococcaceae</th><th></th><th>,</th><th>- Mian</th><th>ě</th><th></th><th>Eurvarchaeota</th><th>Methanococcales</th><th>Methanococcaceae</th></t<>		-Phor		Eurvarchaeota	Thermococcales	Thermococcaceae		,	- Mian	ě		Eurvarchaeota	Methanococcales	Methanococcaceae
Syme Encreachaeta Crassodatina Symechocystia 4 Aut A Eurarchaeta Anthanodacteriales 4 Aut A Eurarchaeta Anthanodacteriales 4 Chaesadobaeta Acchaesadobaeta Acchaesadobaeta 4 Chaesadobaeta Acchaesadobaeta Acchaesadobaeta 4 Chaesadobaeta Acchaesadobaeta Acchaesadobaeta 4 Chaesadobaeta Acchaesadobaeta Acchaesadobaeta 4 Chaesadobaeta Beinacotaeta Acchaesadobaeta 4 Chaesadobaeta Beinacotaeta Acchaesadobaeta 4 Chaesadobaeta Beinacotaeta Actinodacteridae 4 Chaesadobaeta Beinacotae Actinodacteridae 4 Chaesadobaeta Beinacotaeta Actinodacteridae 4 Chaesadobaeta Beinacotaeta Actinodacteridae 4 Chaesadobaeta Actinodacteridae Beinacotaeta Actinodacteridae 4 Chaesadobaeta Beinacotaeta Actinodacteridae Actinodacteridae 4 Chaesadobaeta Beinacotaeta Actinodacteridae Actinodacteridae 4 Chaesadobaeta Beinacotaeta Beinacotaetaeta Beinacotaetaeta		Bsub	i ii	Firmicutes	Bacillus/Clostridium group B	acillus/Staphylococcus gro			Phor	i i		Eurvarchaeota	Thermococcales	Thermococcaceae
And Anthacquidates Anthacquidates </th <th></th> <th>-Svne</th> <th></th> <th>CvanoBacteria</th> <th>Chroococcales</th> <th>Synechocystis</th> <th>Ш.</th> <th> I</th> <th>Mthe</th> <th></th> <th></th> <th>Eurvarchaeota</th> <th>MethanoBacteriales</th> <th>MethanoBacteriaceae</th>		-Svne		CvanoBacteria	Chroococcales	Synechocystis	Ш.	I	Mthe			Eurvarchaeota	MethanoBacteriales	MethanoBacteriaceae
Constantiales Constan		- Aful		Eurvarchaeota	Archaeoplobales	Archaeoglobaceae	Ш.		LAful			Furvarchaeota	Archaenninhales	Archaeoniohaceae
Burnischastalise Seinschastalise		Ecol		ProteoBacteria	gamma subdivision	EnteroBacteriaceae	Ш.	Ч	- 4360			Aquificales	Aquificaceae	Aquifex
Actionalization de la constituita de la con	111 76	Tpal		Spirochaetales	Spirochaetaceae	Treponema	Ш.		Mtub			Firmicutes	ActinoBacteria	Actinobacteridae
Image: Section of the section of th		Mthe		Eurvarchaenta	MethanoBacteriales	MethanoBacteriaceae			Beub			Firmicutes	Bacillue/Clostridium_group	n Bacillus/Stanbylococcus are
Image: B Emissions Balling/Classicium, group Hamissions Hamissions Hamission Hamision Hamission Hamision<		Mtub		Firmicutes	ActinoBacteria	Actionhacteridae		<u> </u>	Mone			Eirmicutes	Bacillus/Clostridium_group	n Mellicutos
Image:		-Mone		Firmicutes	Bacillus/Clostridium, aroun	Mollicutes	Ш.		Maan			Firmicutes	Bacillus/Clostridium_group	n Mellicutes
Understand Construction C		Scer	Ē	Funni	Ascomycota	Hemissromyretes	Ш.	U.	How			Protoe Postorio	Dacings/Clostingium_group	2 Molicology group
Import Bit constraints International constraints	1 4	Hov		ProtonBacteria	ansilon subdivision	Helicobacter, group		1	- Ppro			ProteoDacteria DistanDestaria	epsilon_subdivision	<u>Disketteislee</u>
Image: Second				ProtocBacteria	apmma, subdivision	Posteurellacease	Ш.		- Epro			ProteoDacteria DistasPastaria	alpha_subdwsion	<u>Rickensiales</u>
Under Organisms that are not in the Tree Organisms that are not in the Tree Pfa2 E Abreolata Apper A Cunvariabaeta Apper A Cunvariabaeta Apper A Sinchaetales	L			Metazoa	Nematoda	Chromodorea	Ш.		Ecol		<u> </u>	ProteoDacteria	gamma_subdivision	<u>emerobacteriaceae</u>
Immedia Declanation products Methanococcales Methanococcales Eonetia Milin A Europationani gues Methanococcales B Spinochastales		Mgen		Firmicular	Bacillue/Clastidium aroun	Mallicutes	Ш.	-	- Hint			Proteobacteria	gamma_supdivision	Pasteurellaceae
Image: Super-Decemption Construction International constraints Spinochaetales Spinochaetale Spinochaetale Spinochaetales Spinochaetales Spinochaetale Spinochaetales Spinochaetales Spinochaetales Spinochaetale Spinochaetales Spinochaetales Spinochaetales Spinochaetales Spinochaetales Spinochaetales Spinochaetales		- Misn		Euroarchaeata	Methanocaccalae	Methanococcacaaa		l r	Tral			Spirochaetales	Spirochaetaceae	Dorrella
Image: Constraint of the state of the s		Bhur		Spirachaetales	Sninchaetacaaa	Borrelia			- <u>Ipai</u>		B	Spirochaetales	Spirochaetaceae	Treponema
Image: Construction of the construc		LBoro		ProtenBacteria	aloha subdivision	Picketteislee		L	Syne			CyanoBactena	Chlorodicales	Synechocystis
Option Organisms Chlamydiaesa Chlamydia Organisms that are not in the Tree Organisms that are not in the Tree Organisms that are not in the Tree Pfa2 E Alveolata Apicomplexa Haemosporida Hpv2 B ProteoBacteria epsilon_subdivision Helicobacter_group Hpv2 B ProteoBacteria epsilon_subdivision Helicobacter_group Lei1 E Euglenozoa Kinetoplastida Trypanosomatidae Aper A Crenarchaeota Desulfurococcales Desulfurococcales Pfa3 E Abreolata Apicomplexa Haemosporida M Drad B Thermotogales N/A Pfa3 E Abreolata Apicomplexa Haemosporida M Document. Done Document. Done Document. Done Discource to the tree Discource to the tree		Cone		Chlamydiales	Chlamydiaceae	Chlamydonhila	Ш.		Chre			Chlamydiales	Chlamydiaceae	Chlamydophila
Organisms that are not in the Tree Organisms that are not in the Tree Pfa2 E Abreolata Apicomplexa Haemosporida Hpv2 B ProteoBacteria epsilon_subdivision Helicobacter_group Hpv2 B ProteoBacteria epsilon_subdivision Helicobacter_group Lei1 E Euglenozoa Kinetoplastida Trypanosomatidae Aper A Crenarchaeota Desulfurococcales Desulfurococcales Aper A Crenarchaeota Desulfurococcales Desulfurococcales Pfa3 E Abreolata Apicomplexa Haemosporida Menosporida Maemosporida Maemosporida Maemosporida Pfa3 E Abreolata Apicomplexa Haemosporida Menosporida Maemosporida Maemosporida Maemosporida Maemosporida Menosporida Maemosporida Maemosporida Maemosporida Maemosporida Maemosporida Maemosporida Maemosporida Maemosporida Maemosporida Menosporida Maemosporida Maemosporida Maemosporida Maemosporida <th></th> <th>Ctra</th> <th></th> <th>Chlamydiales</th> <th>Chlamydiaceae</th> <th>Chlamydia</th> <th></th> <th></th> <th>- <u>Ctra</u></th> <th></th> <th>Organia</th> <th><u>Chiamyolales</u></th> <th>Chiamydiaceae</th> <th>Chiamydia</th>		Ctra		Chlamydiales	Chlamydiaceae	Chlamydia			- <u>Ctra</u>		Organia	<u>Chiamyolales</u>	Chiamydiaceae	Chiamydia
Pfa2 E Aveolata Apticomplexa Haemosponda Hpv2 B ProteoBacteria epsilon_subdivision Helicobacter_group Hpv2 B ProteoBacteria epsilon_subdivision Helicobacter_group Lei1 E Euglenozoa Kinetoplastida Trypanosomatidae Aper A Crenarchaeota Desulfurococcales Desulfurococcales Aper A Crenarchaeota Desulfurococcales Desulfurococcales Tmar B Thermotogales Thermotogales N/A Pfa3 E Abreolata Apicomplexa Haemosporida Drad B Thermotogales Haemosporida Deinococcus group Deinococcus group Document. Done Sk Sk Sk Sk Sk Drad B Thermotogales Deinococcus group		oua	Organism	s that are not in th	e Tree	<u>omaniyana</u>			D6-0		Urganis	ams that are not in the	Tree	Hermonolde
Internet concentration Internet concentraticon Internet concentration <thi< th=""><th></th><th>Pfa2</th><th>E</th><th>Alveniste</th><th>Anicomoleva</th><th>Haemosnorida</th><th></th><th></th><th>Praz</th><th></th><th>E</th><th>Alveolata</th><th>Apicomplexa</th><th>Haemosponoa</th></thi<>		Pfa2	E	Alveniste	Anicomoleva	Haemosnorida			Praz		E	Alveolata	Apicomplexa	Haemosponoa
Lei Contractarente approximation approximati		Hpv2		ProtenBacteria	ensilon subdivision	Helicobacter oroup			Hpy2		B	ProteoBacteria	epsilon_subdivision	Helicobacter_group
Aper Desulfurococcales Desulfurococcales Aper Aper Aper Aper Aper A Crenarchaeota Desulfurococcales Tmar B Thermotogales Thermotogales Thermotogales Thermotogales N/A Pfa3 E Abreolata Apicomplexa Haemosporida Drad B Thermus/Deinococcus group Deinococcules Deinococcus Deinococcus V		Lei1	Ē	Euglephica	Kingtonlactida	Trypanosonatidae			Len			Euglenozoa	Kinetoplastida	<u>Trypanosomatidae</u>
Imar B Intermotogales Intermotogales Intermotogales Intermotogales Intermotogales Pfa3 E Abreolata Apicomplexa Haemosporida Imar B Intermotogales Intermotogales Intermotogales Imar B Intermotogales Intermotogales Intermotogales Imar B Intermotogales Intermotogales Intermotogales Imar B Intermotogales Intermotogales Haemosporida Imar Drad B Thermotogales Haemosporida Imar Drad B Intermotogales Haemosporida Imar Document Drad B Intermotogales Haemosporida		Aper		Crenarchaeota	Desulfuroconcales	Desulfurococcaceae			Aper		A	Crenarchaeota	Desunurococcales	Desulturococcaceae
Pfa3 E Abreolata Apicomplexa Haemosponda Image: Specific and the specific and t	1	Tmar		Thermotogales	Thermologa	N/A			I mar		Ë	Inermotogales	Inermotoga	<u>N/A</u>
Image: Second and the second and t		Pfa3		Alveolata	Anicomolexa	Haemosonida	-1		Pras			<u>Alveolata</u>	Apicomplexa	Haemosporida
📓 🕒 Document Done 🛛 🐝 😂 🕼 🖓 🖓 👘 Document Done	4	FIQ	■ [5]	CHARMEN BURNEY	Chicouthieva	The mosperior		đ	Drad	•	8	Thermus/Deinococcus_grou	p Deinococcales	Deinococcus
		Document Done			E 🦓	6 🗳 🐠 🖪 🥩	//		Document: Don	e			3	🦗 🗳 📣 🖬 🗭 📕

Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

1 Using Parts to Interpret Genomes. Shared and/or unique parts. Venn Diagrams, <u>Fold tree</u> with all-β diff. Ortholog tree. Top-10 folds.

2 Using Parts to Interpret Pseudogenomes. In worm, top Ψ–folds (DNAse, hydrolase) v top-folds (Ig). chr. IV enriched, dead and dying families (90YG v 1G)

3 Using Parts to Interpret Transcriptomes: Expression & Structure. Top-10

parts in mRNA. <u>Enriched</u> in transcriptome: $\alpha\beta$ folds, energy, synthesis,TIM fold, VGA. <u>Depleted</u>: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

5 Expression & Function.

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



bioinfo.mbb.yale.edu

H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder