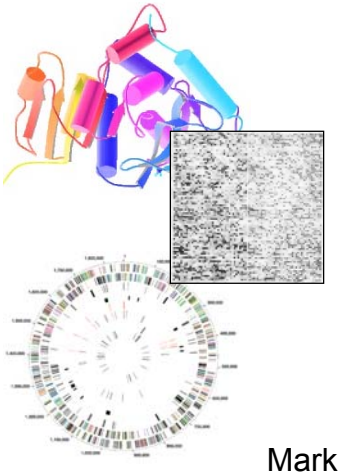
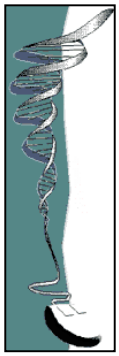
BIOINFORMATICS
Surveys


Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a

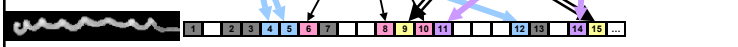
Large-scale Database Surveys (contents)

- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Function Classification
- Cross-tabulation, folds and functions
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection

Simplifying the Complexity of Genomes: Global Surveys of a Finite Set of Parts from Many Perspectives

(human)

(*T. pallidum*)

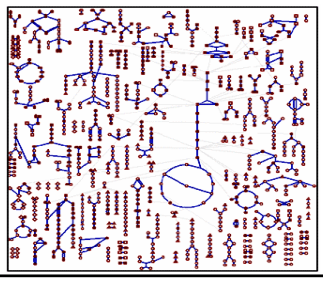


~100000 genes

~1000 folds

~1000 genes

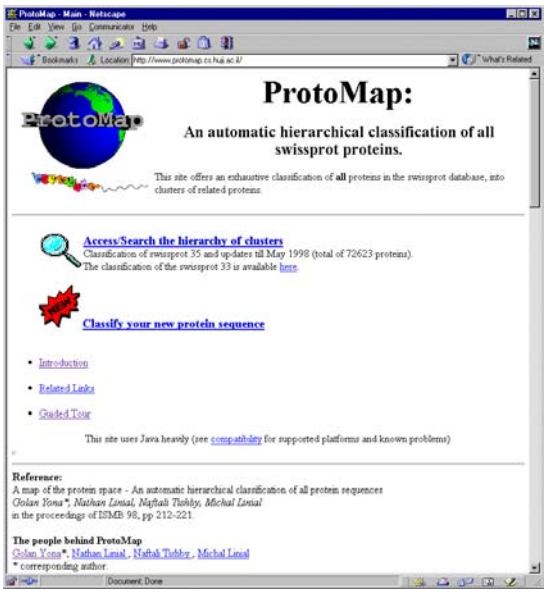
Same logic for sequence families, blocks, orthologs, motifs, pathways, functions....



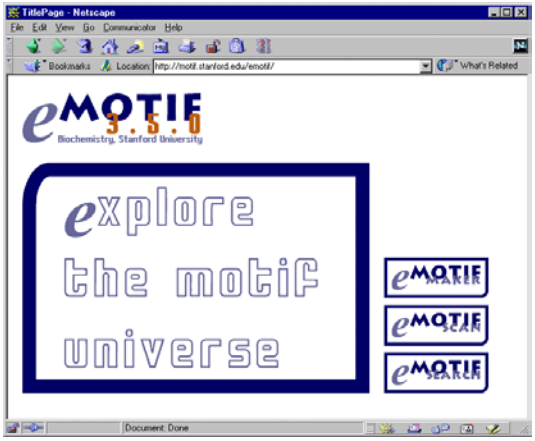
Functions picture from www.fruitfly.org/~suzi (Ashburner); Pathways picture from, ecocyc.pangeasystems.com/ecocyc (Karp, Riley). Related resources: COGS, ProDom, Pfam, Blocks, Domo, WIT, CATH, Scop....

Extra

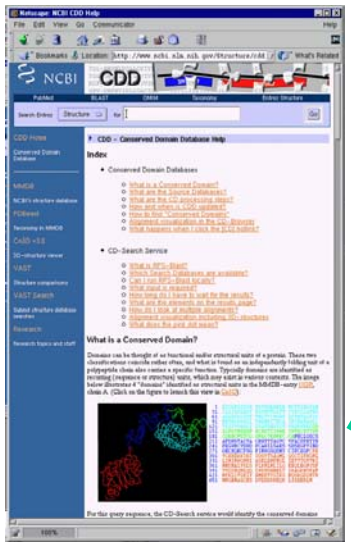
Part = Homolog

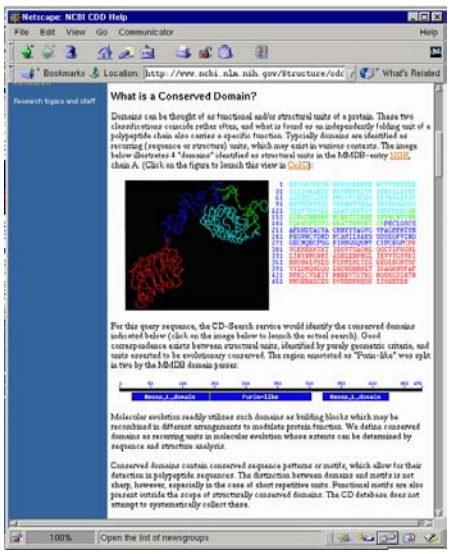


Part = Motif

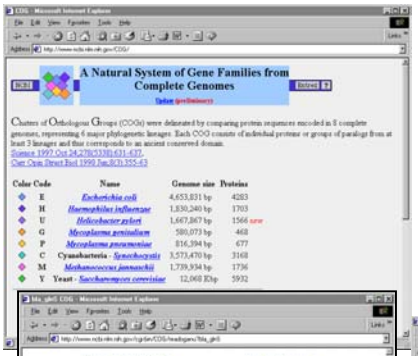


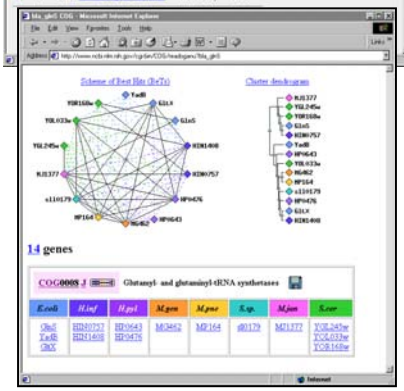
Part = Conserved Domains

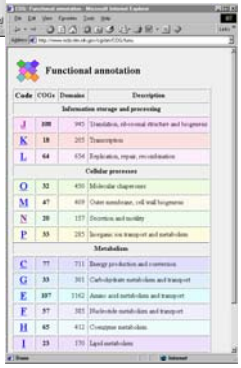




Part = Ortholog
COGs - Orthologs







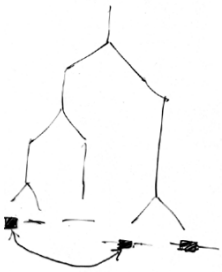
Ortholog ~ gene with precise same role in diff. organism, directly related by descent from a common ancesor

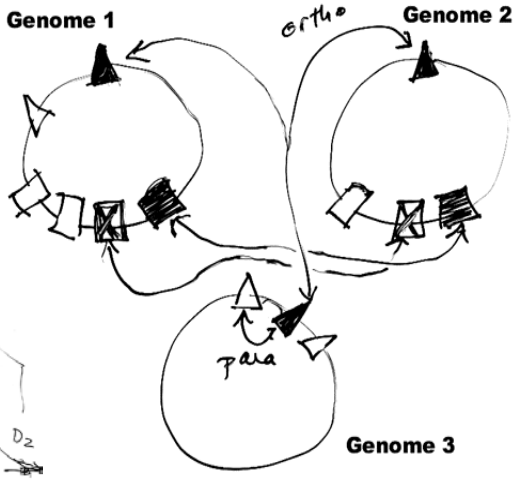
vs
Paralog

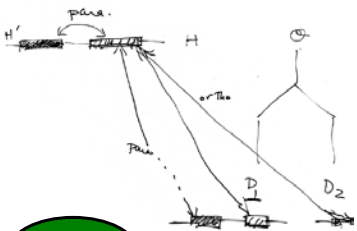
Ortholog, homolog, fold

(Lipman, Koonin, NCBI)

What is an Ortholog?





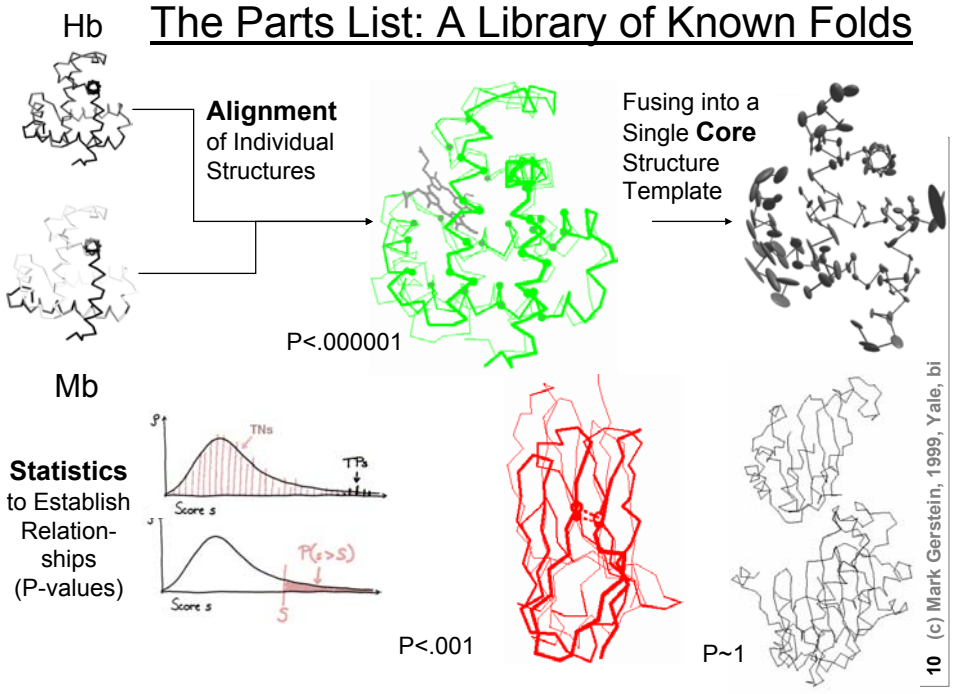


Core

Large-scale Database Surveys (contents)

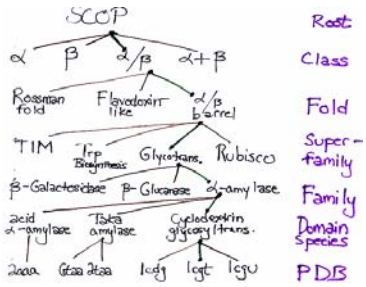
- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Function Classification
- Cross-tabulation, folds and functions
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection

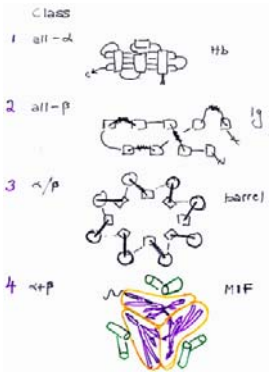
The Parts List: A Library of Known Folds



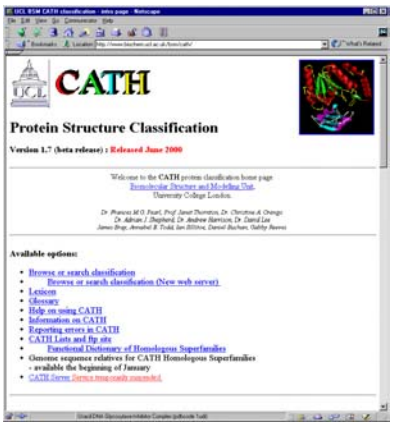
Fold Classifications

- Scop
  - ◊ Chothia, Murzin (Cambridge)
  - ◊ Manual classification, auto-alignments available
  - ◊ Evolutionary clusters
- Cath
  - ◊ Thornton (London)
  - ◊ semi-automatic classification with alignments
  - ◊ class, arch, topo., homol.
- FSSP
  - ◊ Sander, Holm (Cambridge)
  - ◊ totally automatic with DALI
  - ◊ objective but not always interpretable clusters
- VAST



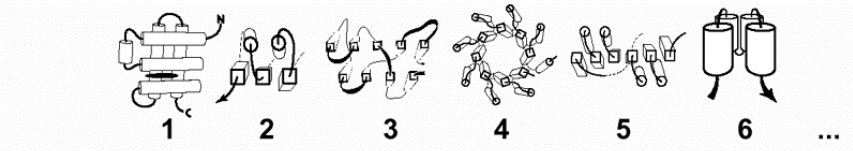


Part = Fold

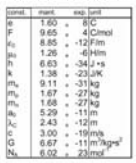


Fold Library vs. Other Fundamental Data structures

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

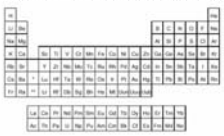


Folds in Molecular Biology **1000-10000**



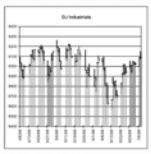
10

Physics



100

Chemistry



1000 -10000

Finance

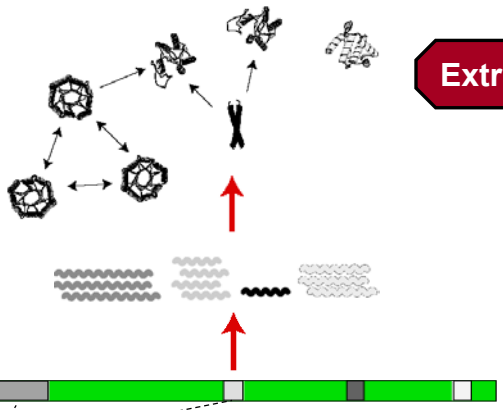
>1000000

Politics

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

The Next Step: Post-genomic Challenges

Extra



↑ #1: Understanding Protein Function on a Genomic Scale

Large-scale Biochemistry: Expression, Structural Genomics, Protein Interactions

Initial Step: genome sequence & genes

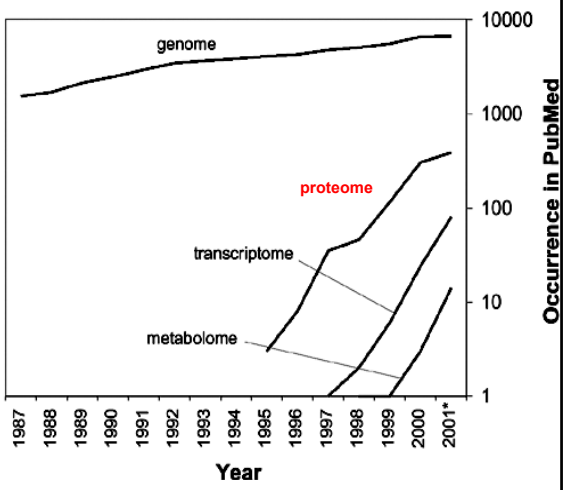
#2: Understanding the Meaning of Intergenic Regions

Evolutionary Implications as a Graveyard: Pseudogenes, Regulatory Regions, Repeats.

Term	Google Hits	PubMed Hits	1st PubMed Hit Year
Genome	**~1880000**	**66171**	**1932 ****
Proteome	**~63,000**	**703**	**1995**
Transcriptome	**3520**	**72**	**1997**
Physiome	2980	15	1997
Metabolome	**349**	**12**	**1998**
Phenome	4980	6	1995
Morphome	238	2	1996
Interactome	56	2	1999
Glycome	46	1	2000
Secretome	21	1	2000
Ribonome	1	1	2000
Orfeome	42	-	-
Regulome	18	-	-
Cellome	17	-	-
Operome	8	-	-
Transportome	1	-	-
Functome	1	-	-

An 'Omic Language to Describe the Next Steps

Extra



proteomics

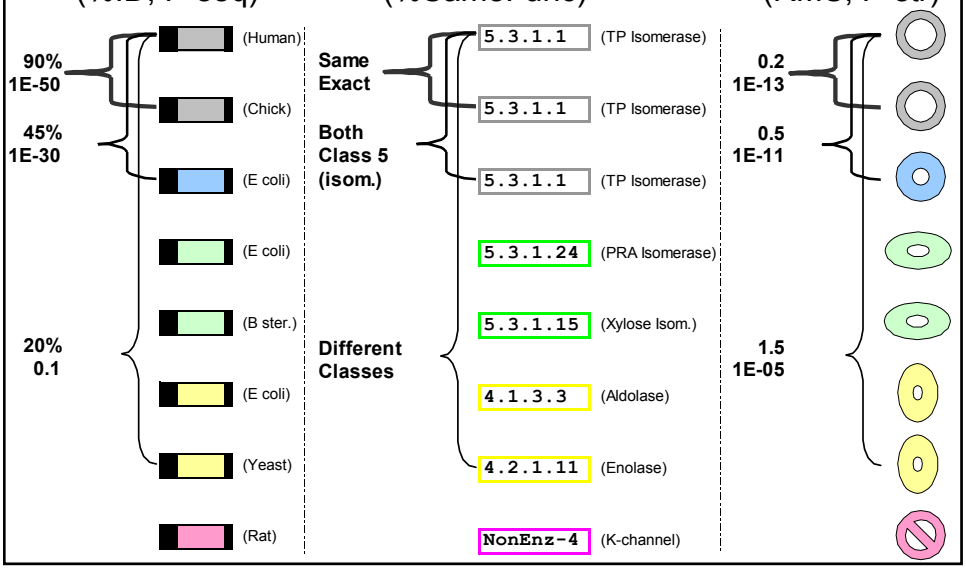
Integrating hetergeneous 'omic information through proteins: families, folds, locations, functions, interactions, pseudogenes…

Large-scale Database Surveys (contents)

- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Function Classification
- Cross-tabulation, folds and functions
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection

Annotation Transfer: TIM ex.

Extra



Chothia & Lesk, 1986 -- 32 points

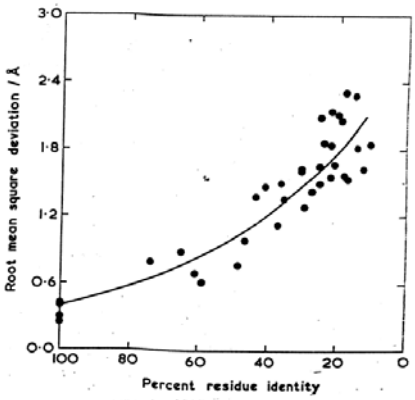


Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

EMBO J **4**: 823 (1986)

"The relation between the divergence of sequence and structure in proteins"

32 pairs of homologous proteins

RMS, percent identity

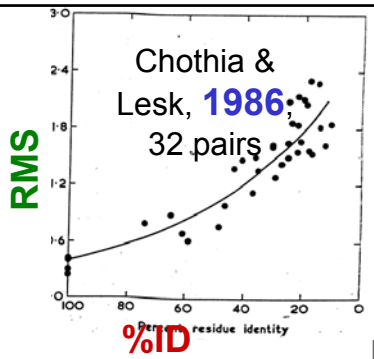
$\Delta = 0.40\ e^{1.87H}$

Now redo with >16,000 pairs in scop + auto-alignments (pdb95d)....

Extra

Structural Divergence vs. Sequence Divergence

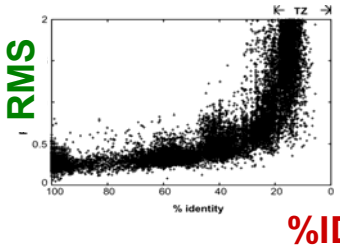
Chothia & Lesk, **1986**, 32 pairs

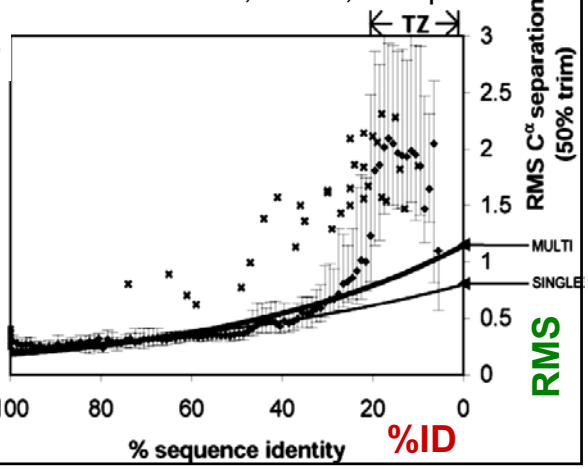


Sequence Diverg. (**%ID, P_{seq}**)
Structural Diverg. (**RMS, P_{str}**)

Extra

Wilson et al., **2000**, 16K pairs





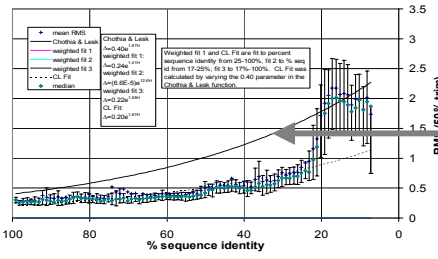
C&L Δ=.4 exp(1.9 H)
Multi: Δ= .2 exp(1.3 H)
Single: Δ= .2 exp(1.9 H)

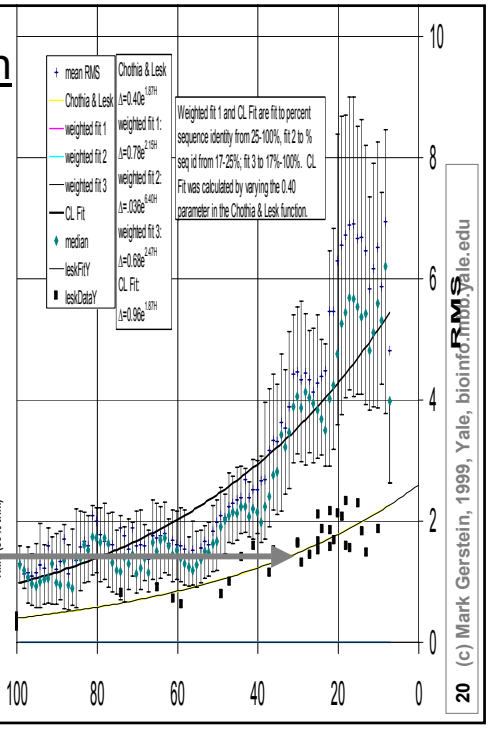
Related work done by Thornton, Pearson, Brenner

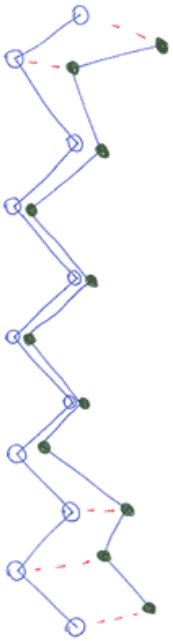
Extra

Problems with RMS

- Dominated by worst-fitting atoms
- Trimming is arbitrary (50%)
- "Bunching up" between 20% and 0% identity







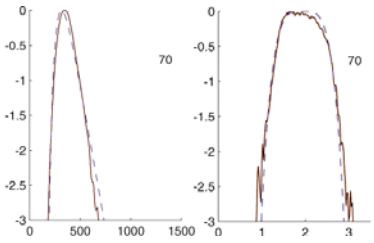
Problems with RMS and %ID

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

S_{str} RMS

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

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

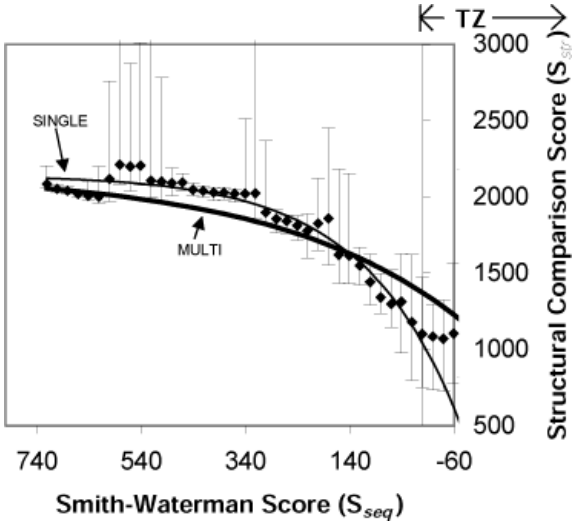


Structural Comp. Score vs.Smith-Waterman Score

overcomes zero bunching, trimming problem

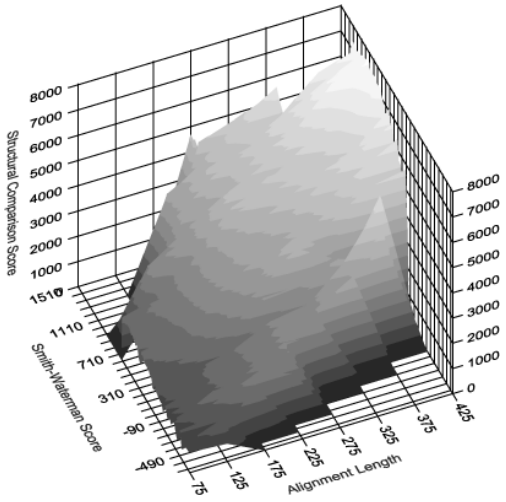
Sstr = 100(21 - 11 exp (-0.0054 SWS)

Extra



Problems with Structural Alignment Score

Different Lengths give different scores.
Scores follow equation of the form:
$y=An+Mx+B$

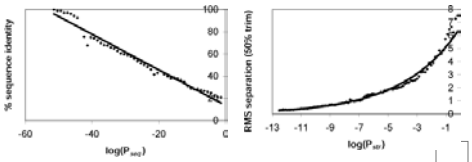


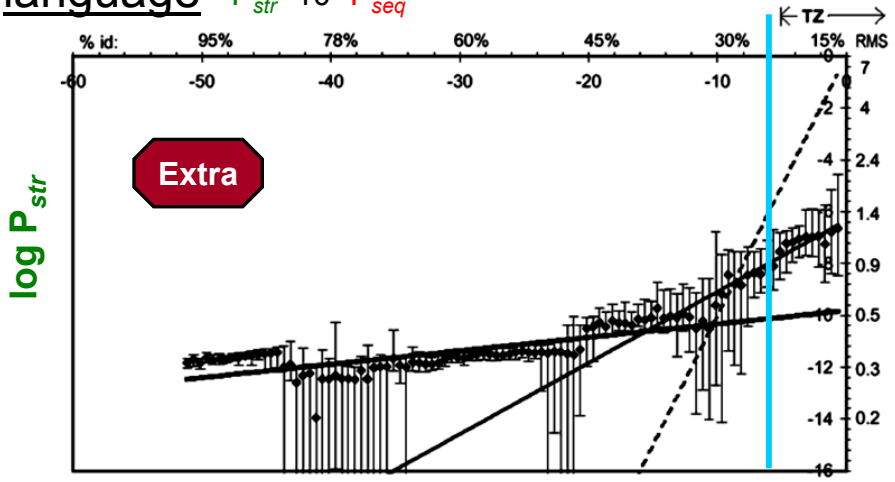
Extra

Modern statistical language

Not in TZ
$P_{str}=10^{-10}P_{seq}^{.05}$

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





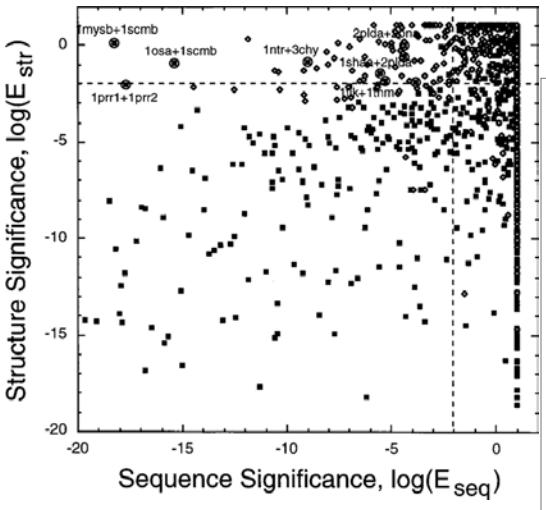
Extra

(overcomes length dependency)

Focus on Twilight Zone

- Sequence Sig. without structure signif.
  - ◊ Protein motions
  - ◊ small proteins
  - ◊ low-res, NMR
- Struc. Sig. without Seq. signif.
  - ◊ More in bottom-right than top-left

Extra



Relationship of Similarity in Sequence & Structure - Summary

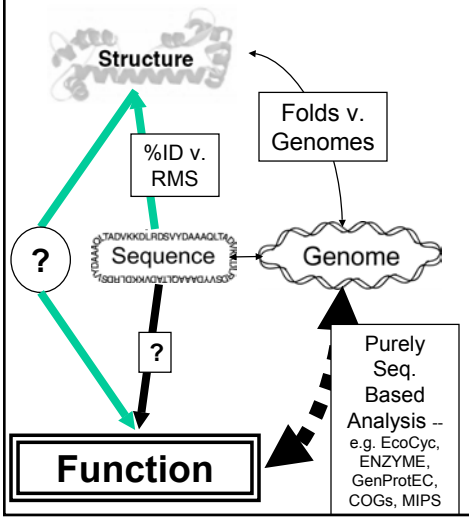
Extra

	Sequence Similarity	Structural Similarity	Features	Limitations
Traditional Scores	Percent sequence identity	RMS C^{α} separation	Well understood, in use	RMS depends most highly on worst matches, requiring arbitrary trimming
Aligment Similarity Scores	S_{seq}	S_{str}	Analogous similarity scores, S_{str} depends most highly on best matches	Dependence on alignment length
Modern Probabilistic Scores	P_{seq}	P_{str}	Statistical significance, unified framework for different comparisons	Not as familiar as RMS and percent identity, some residual length-dependency

Large-scale Database Surveys (contents)

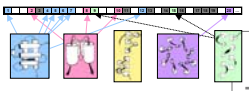
- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Function Classification
- Cross-tabulation, folds and functions
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection

Adding Structure to Functional Genomics, Function to Structural Genomics

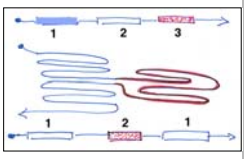


Why Structure? Do we really need it?

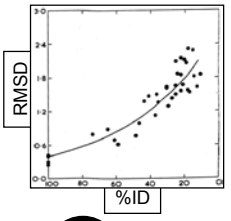
1 Most Highly Conserved



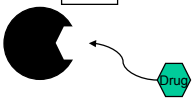
2 Precisely Defined Modules



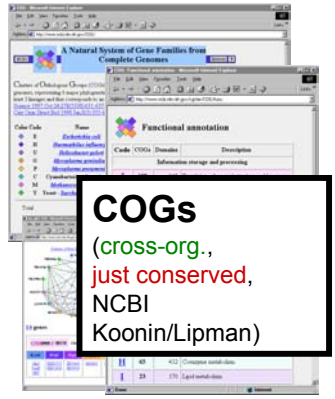
3 Seq. ⇔ Struc. Clearer than Seq. ⇔Func.



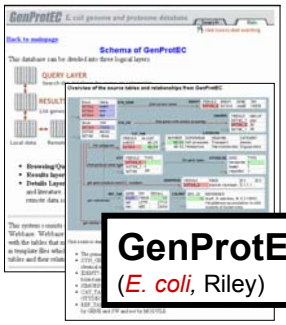
4 Link to Chemistry, Drugs



Functional Classification

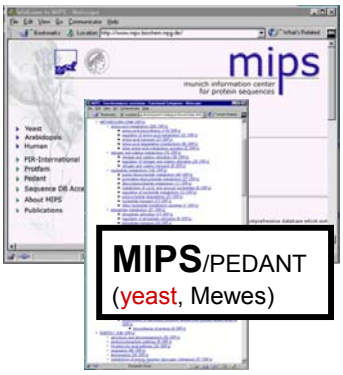


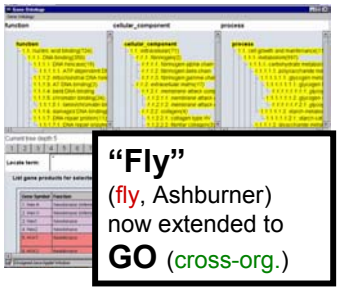
COGs
(cross-org.,
just conserved,
NCBI
Koonin/Lipman)



GenProtEC
(*E. coli*, Riley)

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





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

MIPS/PEDANT
(yeast, Mewes)

Also:

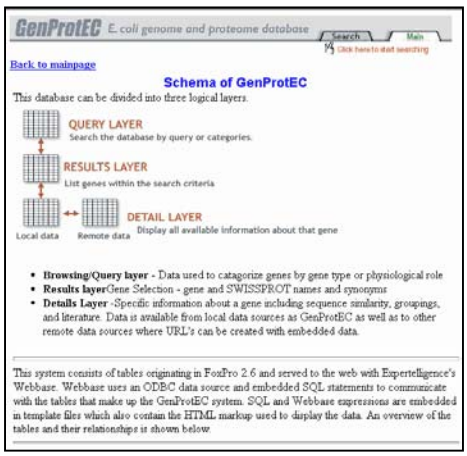
Other SwissProt Annotation

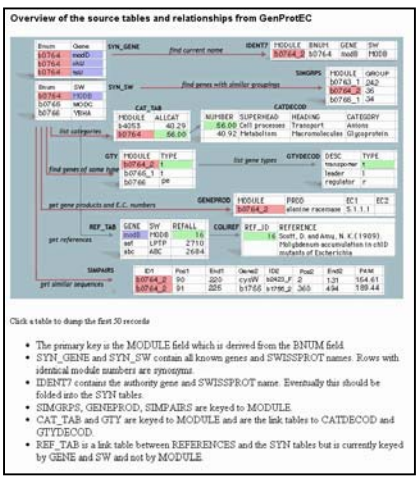
WIT, KEGG (just pathways)

TIGR EGAD (human ESTs)

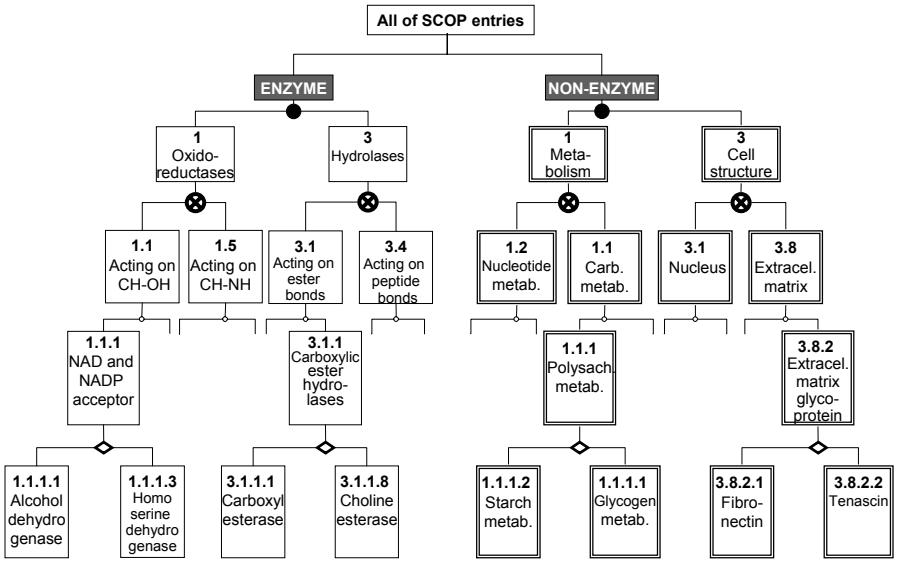
GenProtEC - Functional Classification

the E. coli database
http://genprotec.mbl.edu/start





Hierarchy of Protein Functions



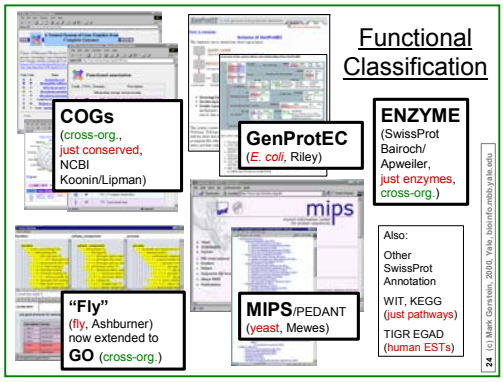
Can we define FUNCTION?

Problems defining function:

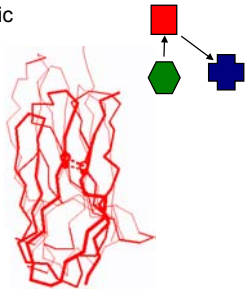
Multi-functionality: 2 functions/protein (also 2 proteins/function)

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

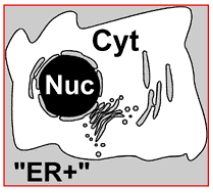
Non-systematic Terminology:
'suppressor-of-white-apricot' & 'darkener-of-apricot'



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



vs.



Stange Gene Names

They have been called aloof, spineless, uncoordinated, strung out, two-faced, puny and spastic-and by some of the most respected scientists in the world. Who are receiving such insults? Genes, which their discoverers are free to name as they wish.

Many of the most outlandish labels come from the study of Drosophila, the fruit flies. **Yippee**, a novel fly gene, which was described in the August issue of the journal Insect Molecular Biology, is one recent example. It was named for the reaction of Katarina Roxström-Lindquist, a graduate student at the University of Stockholm, upon cloning yippee. If she has a good result, Katarina has a habit of writing "yippee" in the margin of her notebook," explains Ingrid Faye, Katarina?s thesis advisor.

In such cases, the appellation says more about the scientist than the gene. Star Trek aficionados surface with names like **vulcan** and **klingon**. For an avid baseball fan, **stranded at second** is a clear way to describe a mutant that dies during development, usually in the second larval stage. Even liquor preferences make their way out of the cabinet and into the literature with genes such as **grappa**.

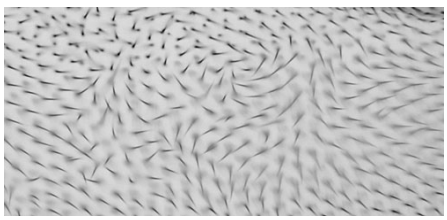
But few names are so closely tied to popular tastes or culture. Scientists more often rely on scriptural, literary or historical sources to devise a colorful term that describes what they see under the microscope.

A gene that affects female fertility was dubbed **sarah** after the wife of Abraham who was infertile for many years but eventually bore a child. One mutation that causes fly embryos to remain headless affects the exuperantia gene, named after a famous slave in ancient Europe who was beheaded as a result of his faith. In a recent issue of the Journal of Cell Biology, Daniel St. Johnston and colleagues reported on **barentsz,** which they named after a Dutch explorer who froze to death in the ice near the North Pole. Why? Because the mutant blocks the movement of a key messenger RNA, causing it to get stuck in the wrong place. **Agoraphobic** refers to a mutant for which the larvae look normal but never crawl out of the egg shell.

Stephen Crews, a professor of biochemistry at the University of North Carolina at Chapel Hill, agrees: "If genes are named in a clever manner, then that probably helps you remember." He followed this prescription in 1987 when he named a Drosophila gene **single-minded** after the visual effect of the mutant morphology. Flies with mutations in this gene possess a single bundle of axons in their nervous systems instead of two. He had also considered using simple-minded but abandoned that label because the name could have been taken as offensive, especially if the function of the equivalent gene in people proved somehow related to a mental disability. His concern was well founded. Recent studies have implicated one of the two human single-minded genes in Down syndrome.

Political correctness has indeed been an issue in the past. In 1963 a mutant fly gene was discovered that caused males to court other males. The assigned gene name of fruity was eventually changed to fruitless after much public disapproval. A similar situation arose more recently, when scientists at Princeton University found mutations in flies that caused them to be learning-defective or, in the vernacular of the investigators, "negged out." They therefore named the corresponding genes after vegetables-cabbage, rutabaga, radish and turnip-which some scientists found objectionable.

Strange Gene Names 2



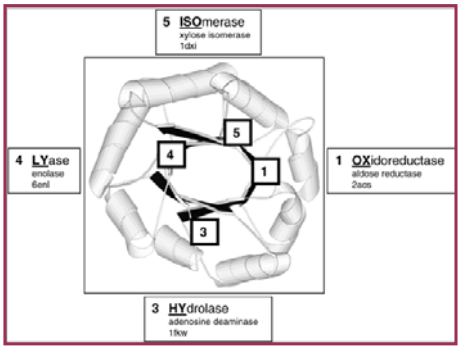
how beautiful it was," says Adler. "Under the microscope, you can see swirling patterns, rather than all the hairs pointing in the same direction as in wildtype." It immediately brought to mind **Starry Night**, the painting by Van Gogh." So when he isolated a similar gene that same year, he naturally enough named it **Van Gogh**. Adler asserts that with a descriptive designation, the connection becomes more personal and subtle between the name of the gene and its visual effects.

Fortunately such troubles rarely arise, in part because most genes do not affect the behavior or appearance of the organism so directly. When obvious clues to the action of the gene are lacking, geneticists often pick a name based on the inferred function of the gene product. **Redtape** is the most recent in a series of designations given to genes which, when mutated, block transport along axons. The predecessors of redtape include **roadblock**, **gridlock** and **Sunday driver**, all emanating from the lab of Lawrence Goldstein at the University of California, San Diego. However, names based on presumed function are often not particularly creative-sometimes they are even misleading. For example, one human gene, first described in 1992, is aryl hydrocarbon receptor nuclear translocator, a mouthful that is hard to remember except by its acronym, **ARNT**. Worse, recent studies have shown that ARNT might not act as an aryl hydrocarbon receptor translocator after all, suggesting it might soon be due for aechristening.

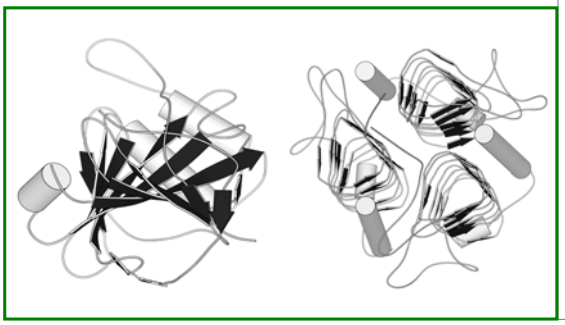
Perhaps the new name will be shorter and more memorable. Or maybe not. Even the author of starry night and Van Gogh concedes that not all genes can be designated so inventively. As Adler says, "There are a lot of genes, and only so many names you can come up with."-Marla Vacek

End of class 2002,11.20
(Bioinfo-12)
[starting in databases]

Fold-Function Combinations #1



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

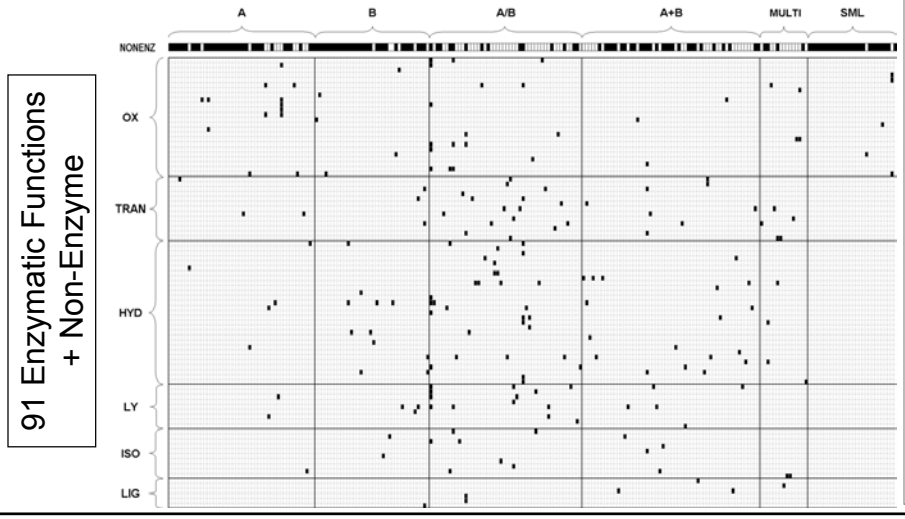


Two Different Folds Catalyze the Same Reaction -- e.g. Carbonic Anhydrases (4.2.1.1)

Fold-Function Combinations Cross-Tabulation

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

229 Folds



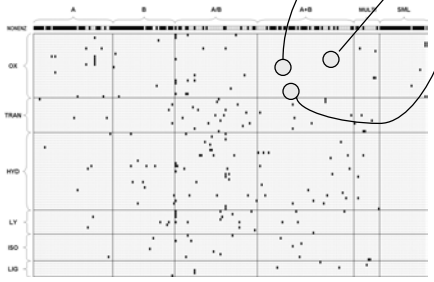
91 Enzymatic Functions + Non-Enzyme

Fold-Function Combinations Cross-Tabulation Summary Diagram

	A	B	A/B	A+B	MULTI	SML	sum
NONENZ	34	30	14	28	4	26	136
OX	13	5	17	3	4	5	47
TRAN	3	3	16	8	5		35
HYD	4	11	30	18	4		67
LY	2	3	13	5			23
ISO	1	2	7	4	2		16
LIG		1	2	3	1		7
sum	57	55	99	69	20	31	331

3

Core



SCOP

ENZYME

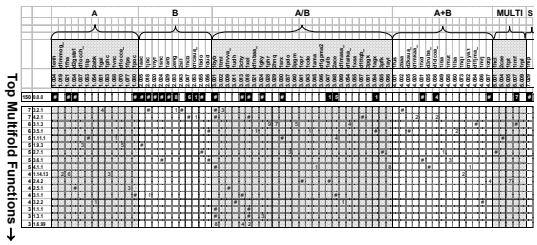
	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

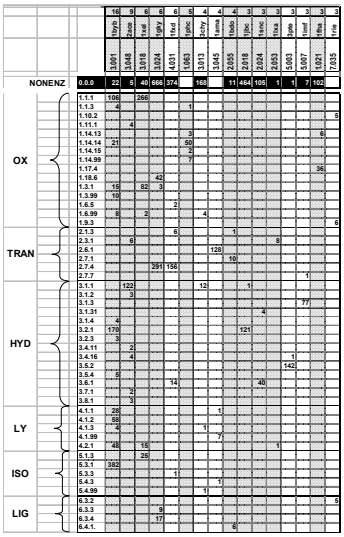
[Similar analysis in Martin et al. (1998), *Structure* 6: 875]

The Most Versatile Folds, Versatile Functions

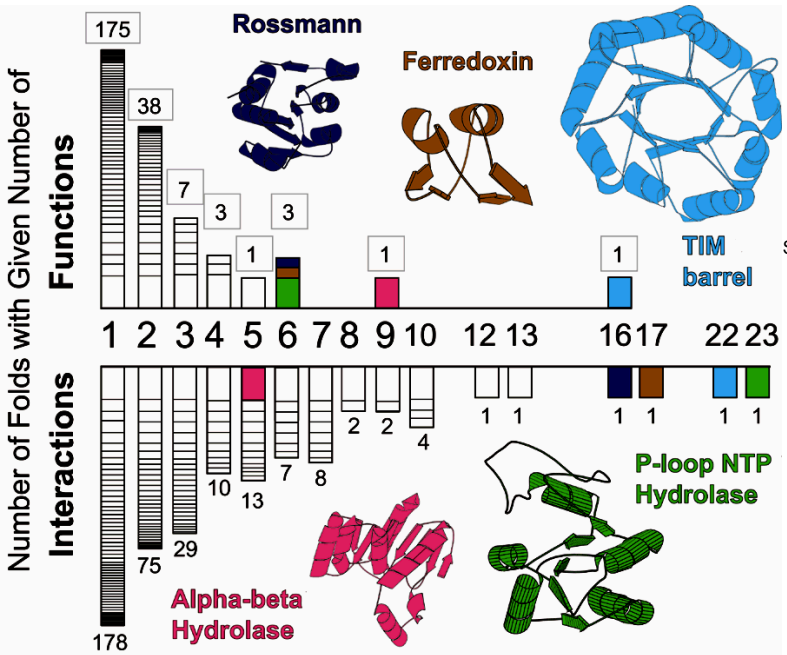
Top-4 Functions: Glycosidases, carboxy-lyases, phosphoric monoester hydrolases, linear monoester hydrolases (3.2.1, 4.2.1 3.1.3, 3.5.1)

Top-5 Folds: TIM-barrel (16), alpha-beta hydrolase fold (9), Rossmann fold (6), P-loop NTP hydrolase fold (6), Ferrodoxin fold (6)





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

Compare Classifications and Genomes

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

Extra

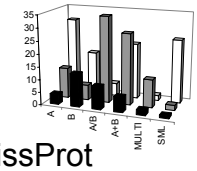
SCOP

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

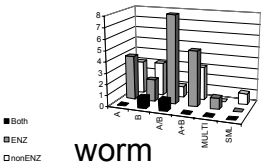
CATH (Thornton)

CATH

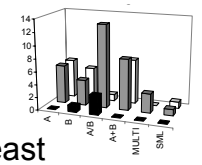
ENZYME	A	B	AB
NONENZ	10	9.0	15
OX	5.1	5.1	10
TRAN		1.3	13
HYD	2.6	1.3	14
LY		2.6	1.3
ISO	1.3	1.3	5.1
LIG			1.3



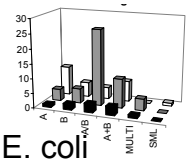
SwissProt



worm



Yeast



E. coli

MIPS YFC (Mewes)

SCOP

MIPS Functional Cat.		A	B	A/B	A+B	MULTI	SML
metabolism	1	3.5	2.3	10	4.5	1.5	0.8
energy	2	1.1	1.2	5	1.9	0.3	0.2
growth, dv., DNA syn.	3	4.9	3.6	4	4.5	1.6	1.2
transcription	4	1.5	1.3	2.2	1.6	0.5	0.8
protein synthesis	5	1	0.9	0.7	1.3	0.3	0.2
protein targeting	6	1.4	1.7	2	1.6	0.5	0.3
transport facilitation	7	0.9	0.5	0.7	0.6	0.4	
intracellular transport	8	1.6	2.1	1.6	0.6	1	
cellular biogenesis	9	0.9	0.7	1.2	0.3	0.3	0.1
signal transduction	10	1	1	1.1	0.3	0.7	0.3
cell rescue, defense	11	1.5	1	2.6	1.6	0.7	0.5
toxic homeostatis	13	0.5	0.3	0.4	0.4	0.2	

COGs vs SCOP: Different Structure Function Relationships for Most Conserved Proteins

Extra

All Yeast COGs

SCOP

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

Most Conserved COGs

SCOP

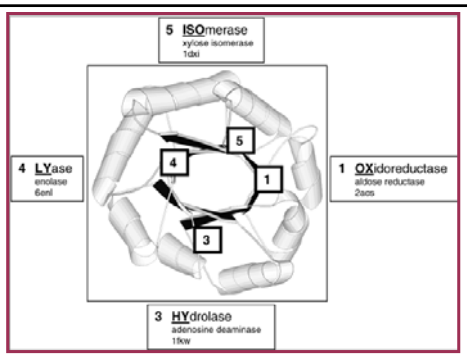
		A	B	A/B	A+B	MULTI	SML
Metabolism	C			7.2	2.9		
	E	1.4		1.4	1.4		
	F			2.9			
	G			4.3	1.4		
	H	1.4	2.9		1.4		
	I						
Information Storage & Processing	J	8.7	7.2	7.2	10	1.4	1.4
	K						
	L					1.4	
Cellular Processes	M						
	N	1.4		1.4			
	O	2.9		7.2	2.9		
	P		1.4		2.9	1.4	

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

Extra

From here to end of Surveys all is "extra" unless otherwise marked.

Fold-Function Combinations #2



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

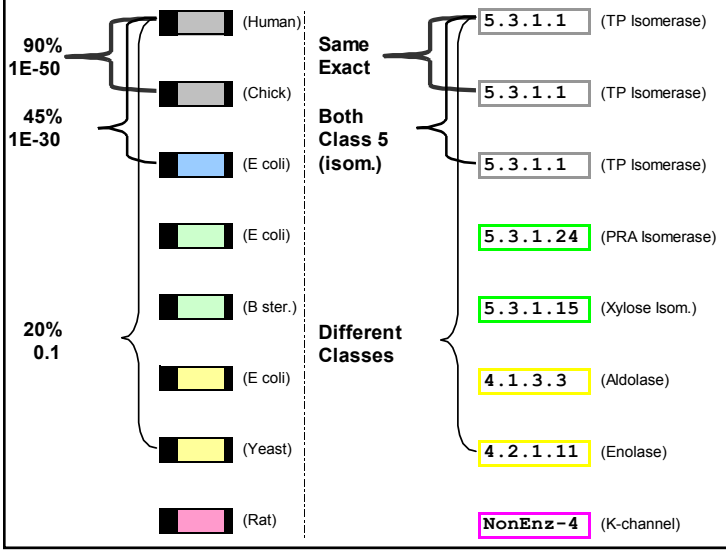
at what degree of divergence?
Sequence Diverg. (**%ID, P_{seq}**)
Structural Diverg. (**RMS, P_{str}**)
Functional Diverg. (**%SameFunc**)

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

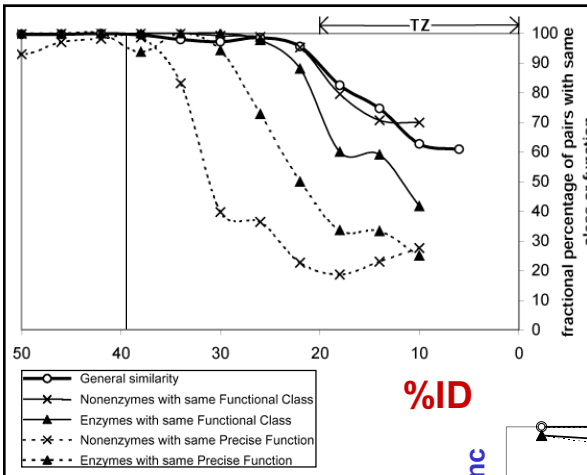
89%	Human	TP Isomerase	5.3.1.1
	Chick	TP Isomerase	5.3.1.1
45%	E coli	TP Isomerase	5.3.1.1
	E coli	PRA Isomerase	5.3.1.24
	B ster.	Xylose Isomerase	5.3.1.5
~20%	E coli	Aldolase	4.1.3.3
	Yeast	Enolase	4.2.1.11
	Rat	K-channel B-sub.	NON-ENZ
	Photobact.	Flavoprotein?	NON-ENZ

Same Exact Func.
Both Class 5
Completely Different
Same Exact Func.

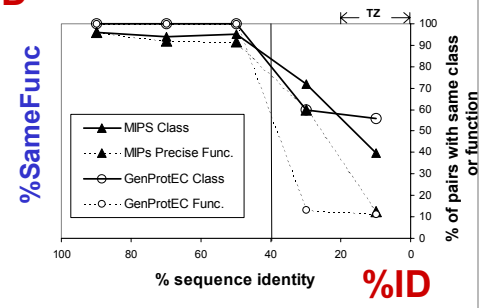
Annotation Transfer: TIM ex.

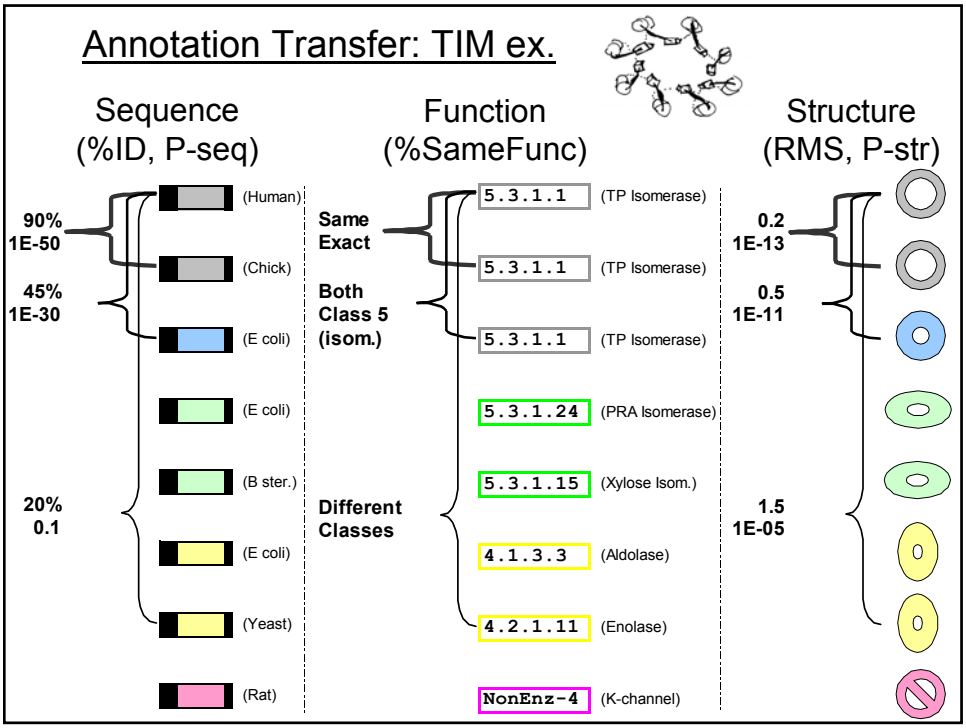


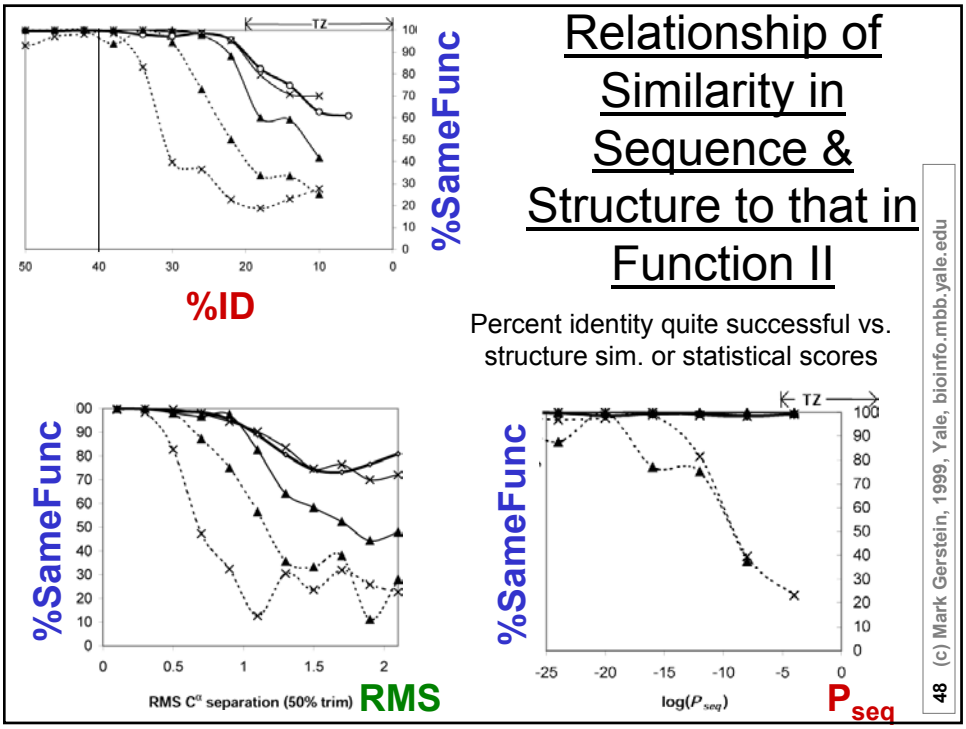
Relationship of Similarity in Sequence & Structure to that in Function



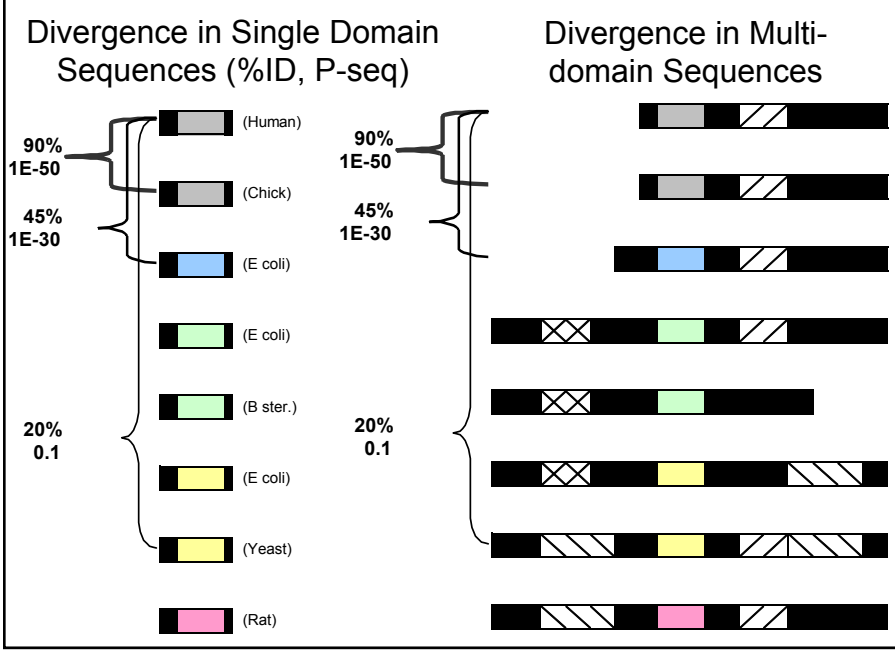
See at what %ID have diff. function (both broad & precise). Use 4 func. classifications -- ENZYME, FLY (+extra), MIPS, GenProtEC



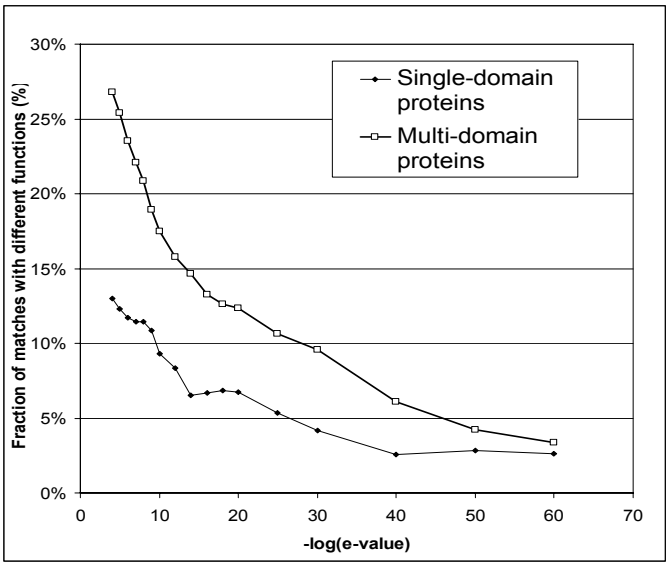

Annotation Transfer: TIM ex.
Sequence (%ID, P-seq)
Function (%SameFunc)
Structure (RMS, P-str)
90% 1E-50
45% 1E-30
20% 0.1
(Human)
(Chick)
(E coli)
(E coli)
(B ster.)
(E coli)
(Yeast)
(Rat)
Same Exact
Both Class 5 (isom.)
Different Classes
5.3.1.1 (TP Isomerase)
5.3.1.1 (TP Isomerase)
5.3.1.1 (TP Isomerase)
5.3.1.24 (PRA Isomerase)
5.3.1.15 (Xylose Isom.)
4.1.3.3 (Aldolase)
4.2.1.11 (Enolase)
NonEnz-4 (K-channel)
0.2 1E-13
0.5 1E-11
1.5 1E-05


Relationship of Similarity in Sequence & Structure to that in Function II
%SameFunc
%ID
TZ
Percent identity quite successful vs. structure sim. or statistical scores
%SameFunc
RMS
RMS Cα separation (50% trim)
%SameFunc
Pseq
log(Pseq)
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu
48

Sequence Divergence of Multidomain Proteins



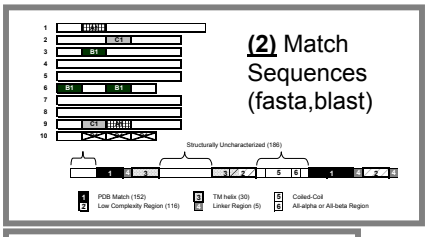
Multi-domain proteins have greater divergence in function with sequence

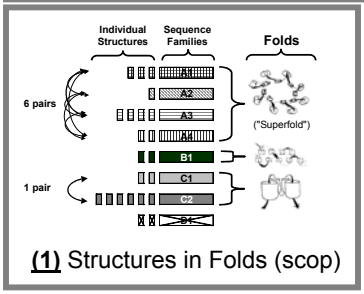


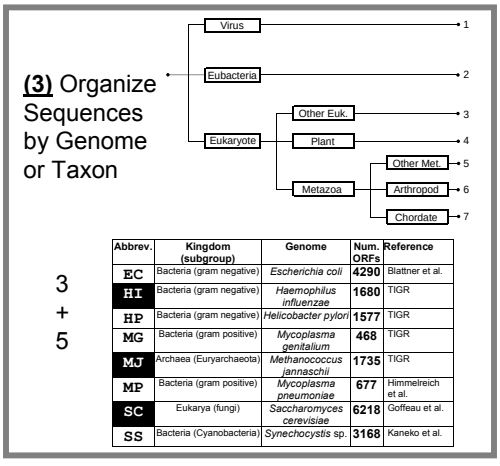
Large-scale Example: Census DB

- 9 Genome Comparison
- 1437 Relational Tables
- 442 Mb
- Simple ASCII Layout

Cross-Reference: Folds→Sequences → Organisms

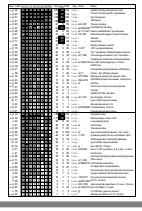






3 + 5

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



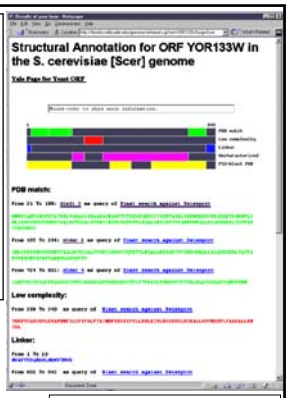
(4) Results in "Fold Table"

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

Integrated Analysis System: X-ref Parts with Genomes

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

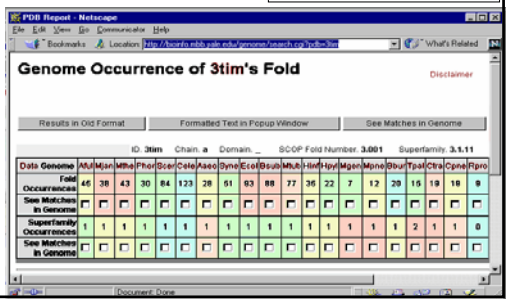
finding parts in genome sequences **blast**, ψ-blast, **fasta**, TM, low-complexity, &c (Altschul, Pearson, Wooton)



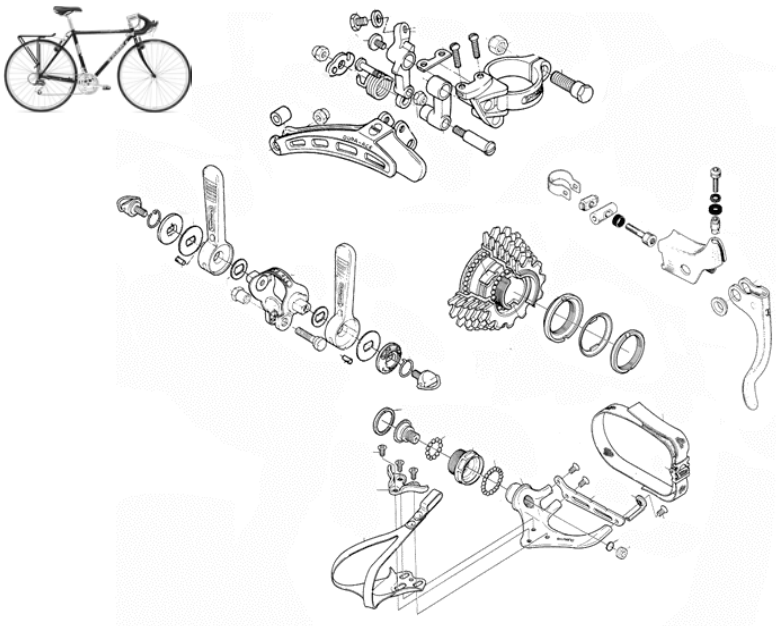
part occurrence profiles

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

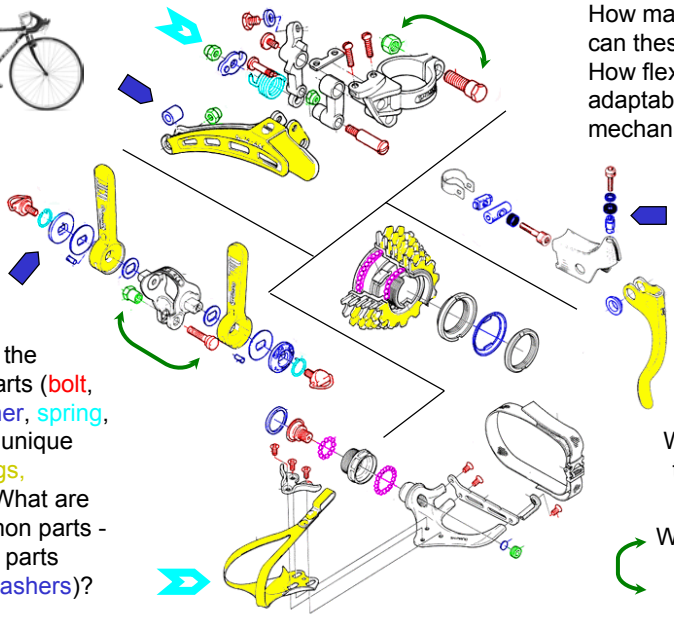
Remington, Matthews '80; **Taylor, Orengo '89**, '94; Thornton, CATH; Artymiuk, Rice, Willett '89; Sali, Blundell, '90; Vriend, Sander '91; Russell, Barton '92; **Holm, Sander '93+ (FSSP)**; Godzik, Skolnick '94; Gibrat, Bryant '96 (VAST); F Cohen, '96; Feng, Sippl '96; G Cohen '97; Singh & Brutlag,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



A Parts List Approach to Bike Maintenance



A Parts List Approach to Bike Maintenance

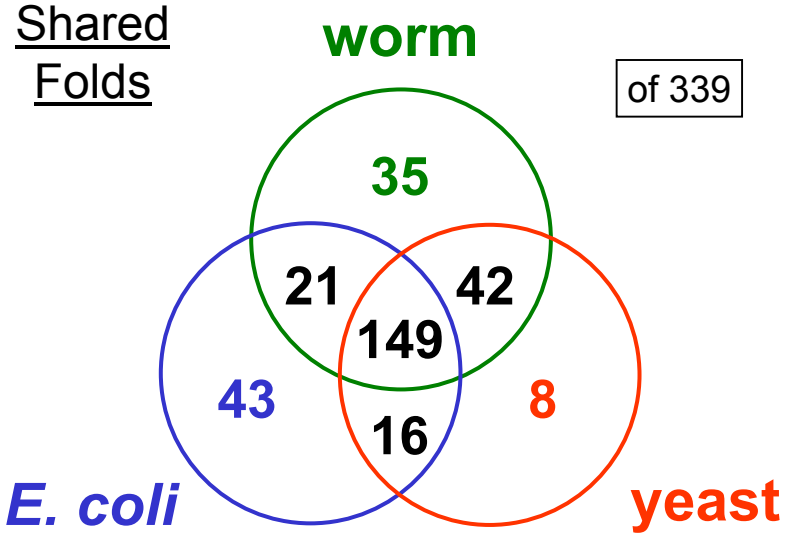


How many roles can these play? How flexible and adaptable are they mechanically?

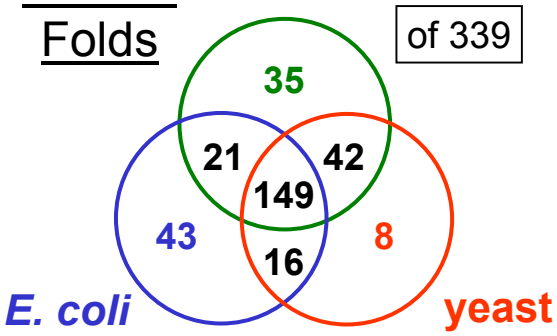
What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

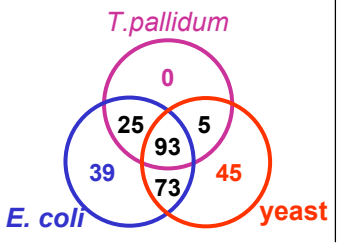
Where are the parts located? Which parts interact?

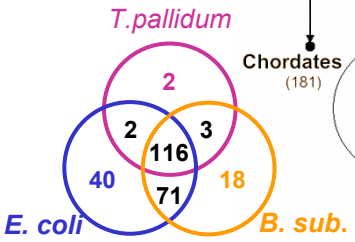
Shared Folds

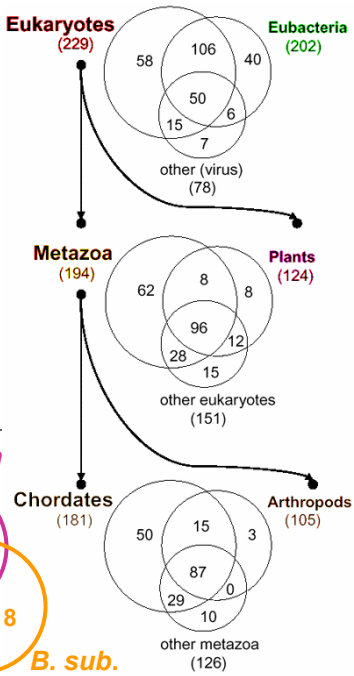


Shared Folds



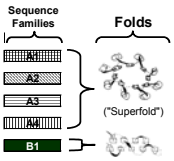






Patterns of Folds Usage in 8 Genomes

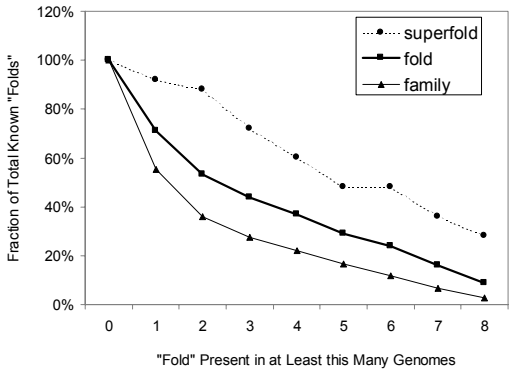
	fold	fam.	super fold
total in PDB	338	990	25
in at least one of 8 genomes	240	547	23
present in this many genomes			
1	60	192	1
2	32	82	4
3	23	54	3
4	27	53	3
5	17	50	0
6	27	49	3
7	24	41	2
8	30	26	7



Superfold = fold that allows many non-homologous seq. (Thornton)

```
ESHSHMMM (##)  ESHSHMMM (##)  ESHSHMMM (##)  ESHSHMMM (##)  ESHSHMMM (##)
CCISPJPG       CCISPJPG       CCISPJPG       CCISPJPG       CCISPJPG

11111111 (30)  .1.....  (23)  1.......  (19)  11111.11 (16)  111111..  (16)
1111....  (09)  11111...  (08)  1.1.....  (08)  1.111.11 (06)  11......  (06)
...1....  (06)  1.11....  (05)  .1.1....  (05)  1.111...  (04)  11.1....  (04)
.1...1..  (04)  ..1.....  (04)  111111.1  (03)  1111111.  (03)  1111..11  (03)
1111.1..  (03)  .....1..  (03)  1111.111  (02)  111...11  (02)  111.11..  (02)
1.11.1..  (02)  ..111...  (02)  .1.11...  (02)  1..1.1..  (02)  1.1..1..  (02)
111.....  (02)  .11.....  (02)  ......1.  (02)  ....1...  (02)  111..111  (01)
111.1.11  (01)  1.111..1  (01)  1.1111..  (01)  .1.1..11  (01)  .1.11.1.  (01)
.11.1..1  (01)  1....111  (01)  1..111..  (01)  1.1...11  (01)  1.1..11.  (01)
11....11  (01)  11.1.1..  (01)  11.11...  (01)  111..1..  (01)  111.1...  (01)
.11...1.  (01)  1.....11  (01)  1...11..  (01)  1.1.1...  (01)  ......11  (01)
....1..1  (01)  ...1.1..  (01)  ...11...  (01)  ..1.1...  (01)  .1....1.  (01)
1....1..  (01)  .......1  (01)
```

Whole Genome Trees

The Genomic Tree as Revealed from Whole Proteome Comparisons

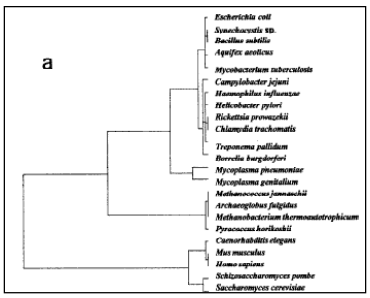
Fredj Tekaia,[1,3] Antonio Lazcano,[2] and Bernard Dujon[1]

[1]Unité de Génétique Moléculaire des Levures (URA1300 Centre National de la Recherche Scientifique (CNRS) and UFR927 University Pierre and Marie Curie), Institut Pasteur, 75724 Paris Cedex 15, France; [2]Facultad de Ciencias, UNAM, Apdo. Cd. Universitaria, 04510 Mexico City, Mexico

The availability of a number of complete cellular genome sequences allows the development of organisms' classification, taking into account their genome content, the loss or acquisition of genes, and overall gene similarities as signatures of common ancestry. On the basis of correspondence analysis and hierarchical classification methods, a methodological framework is introduced here for the classification of the available 20 completely sequenced genomes and partial information for *Schizosaccharomyces pombe*, *Homo sapiens*, and *Mus musculus*. The outcome of such an analysis leads to a classification of genomes that we call a genomic tree. Although these trees are phenograms, they carry with them strong phylogenetic signatures and are remarkably similar to 16S-like rRNA-based phylogenies. Our results suggest that duplication and deletion events that took place through evolutionary time were globally similar in related organisms. The genomic trees presented here place the Archaea in the proximity of the Bacteria when the whole gene content of each organism is considered, and when ancestral gene duplications are eliminated. Genomic trees represent an additional approach for the understanding of evolution at the genomic level and may contribute to the proper assessment of the evolutionary relationships between extant species.

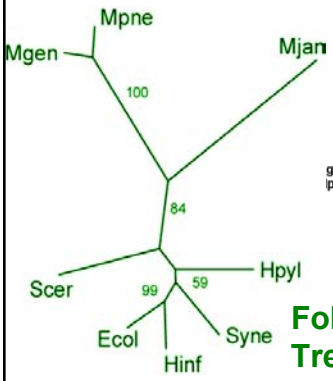
The determination of complete genome sequences from ≥20 organisms offers an unprecedented opportu- horizontal transfer (which may have been more intense during early cellular evolution, Woese 1998), un-

9:550–557 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

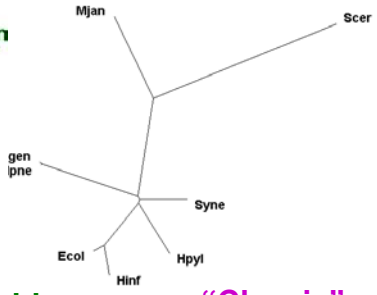


orthologs, homologs, folds, motifs

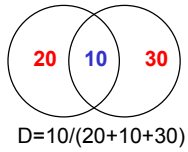
Cluster Trees Grouping Initial Genomes on Basis of Shared Folds



Fold Tree



"Classic" Tree

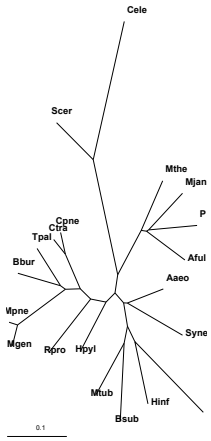


D=10/(20+10+30)

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

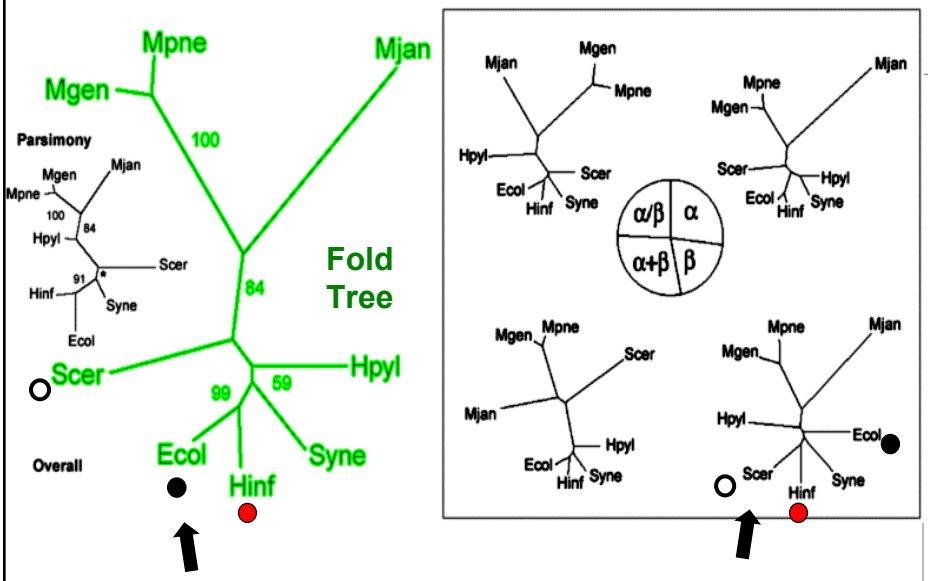
D = shared fold dist. betw. 2 genomes

T= total # folds in both



20 Genomes

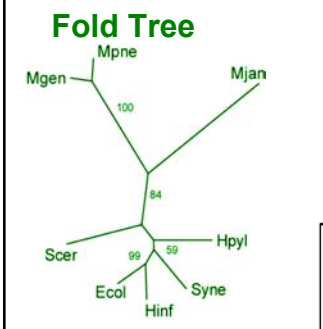
Distribution of Folds in Various Classes

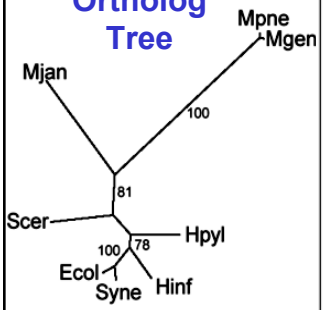


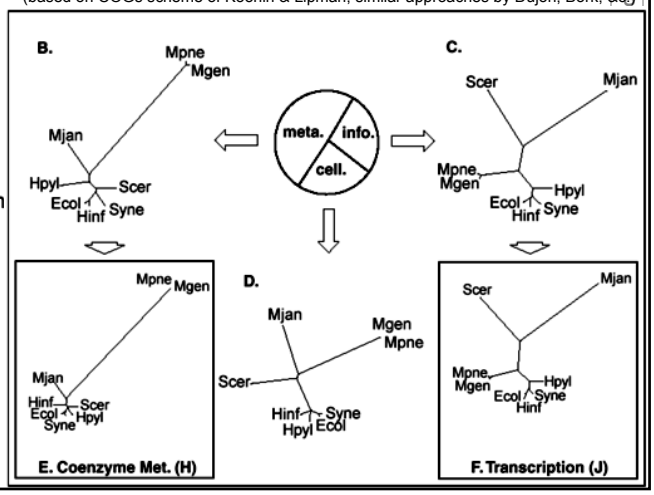
Unusual distribution of all-beta folds

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

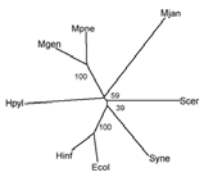
(based on COGs scheme of Koonin & Lipman, similar approaches by Dujon, Bork, &c)



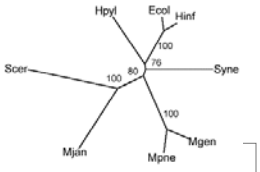




Compare with trees on spectrum of "levels": single-gene trees, whole-genome composition trees

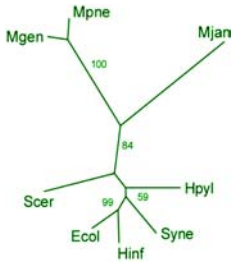


TIM

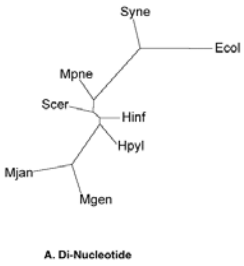


ribosomal protein

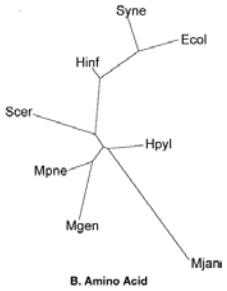
Single-gene Trees



Fold Tree



A. Di-Nucleotide



B. Amino Acid

Ortholog Tree

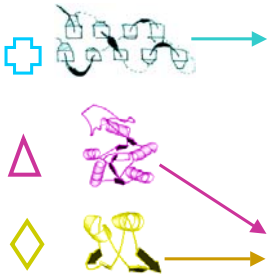
"Classic" Tree

AA & di-NT Composition Trees

(S Karlin)

Common Folds in Genome, Varies Betw. Genomes

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



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

Eubacteria

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

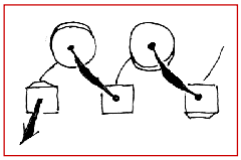
Archaea

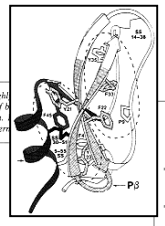
Rank	*M. thermo-autotrophicum* Superfamily	#	*A. fulgidus* Superfamily	#
1	P-loop hydrolyase	93	P-loop hydrolyase	118
2	Phosphate-binding barrel	54	Rossmann domain	104
3	Rossmann domains	53	Phosphate-binding barrel	56
4	Ferredoxins	48	Ferredoxins	49
5	SAM methyl-tranferase	17	SAM methyl-tranferase	24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

Yeast

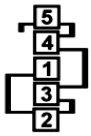
Rank	*S. cerevisiae* Superfamily	#
1	P-loop hydrolyase	249
2	Protein kinase	123
3	Rossmann domain	90
4	RNA-binding domain	75
5	SAM methyl-transferase	63
6	Ribonuclease H-like	57
Total ORFs		6218
with Common Superfamilies		560 (9%)

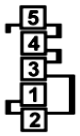
Common, Shared Folds: βαβ structure

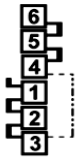


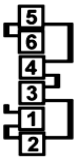


All share α/β structure with repeated R.H. βαβ units connecting adjacent strands or nearly so (18+4+2 of 24)

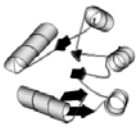


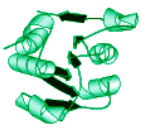


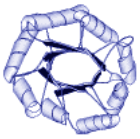




HI, MJ, SC vs scop 1.32







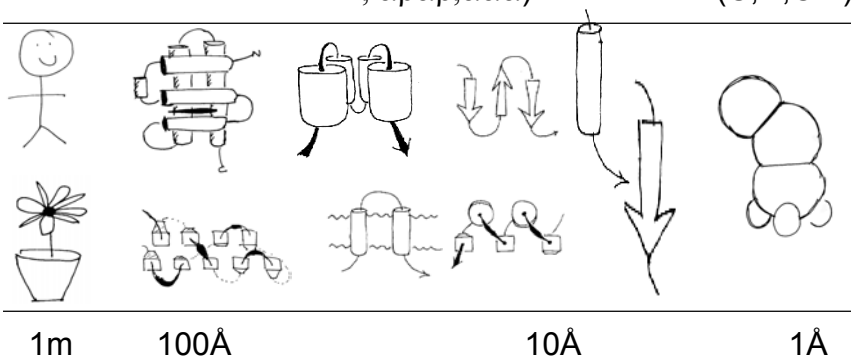
P-loop hydrolase | **Flavodoxin like** | **Rossmann Fold** | **Thiamin Binding** | **TIM-barrel**

What are the most common folds: Overall? In plants? In animals?

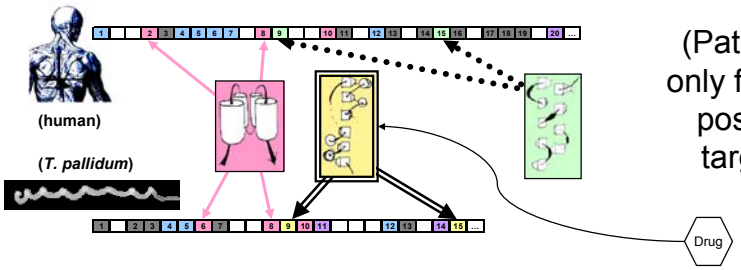
Example Structure (PDB)	Class	Fold Name	Num. Seq.	Families	Num. of Sequences: Total	Eukaryote: Virus	Eubacteria	Plant	Metazoan	Other
Totals			719		37706	3139	7032	4960	19319	1828
Overall Top-10										
1REI-A	β	Immunoglobulin-like	32		13	0	1	0	25	0
6TIM-B	α/β	TIM-barrel	29		6	0	7	20	2	13
1ATP-E	O	Protein Kinases (catalytic core)	1		4	3	0	3	6	6
1FXD	O	Ferredoxin-like	17		4	2	2	17	0	8
1AKE-A	α/β	NTP Hydrolases containing P-loop	9		3	0	5	3	2	7
1HDD-C	α	DNA-binding 3-helical bundle	13		3	0	0	2	5	0
2HSD-A	α/β	Rossmann Fold (NAD binding)	11		3	0	7	3	1	3
1MBD	α	Globin-like	3		2	0	1	0	4	1
2RN2	α/β	like Ribonuclease H	15		2	0	1	2	1	5
1ZNF	S	Classic Zinc Finger	2		1	0	0	0	3	1
Sequence Family Top-11										
1REI-A	β	Immunoglobulin-like	32		13	0	1	0	25	0
6TIM-B	α/β	TIM-barrel	29		6	0	7	20	2	13
1FXD	O	Ferredoxin-like	17		4	2	2	17	0	8
2RN2	α/β	like Ribonuclease H	15		2	0	1	2	1	5
1PYP	β	OB-fold	15		0	0	1	0	0	0
1PTX	S	Small inhibitors, toxins, lectins	14		0	0	0	3	0	0
2TBV-C	β	Viral coat and capsid proteins	14		1	12	0	0	0	0
1HDD-C	α	DNA-binding 3-helical bundle	13		3	0	0	2	5	0
2HSD-A	α/β	Rossmann Fold (NAD binding)	11		3	0	0	7	3	1	3
1RCF	α/β	Flavodoxin-like	11		0	0	0	4	0	0	0
1RCB	α	4-helical cytokines	11		0	0	0	0	2	0

Fold Name	Number	Percent of Sequences: Virus	Eubacteria	Eukaryote: Plant	Metazoan	Other
Plant Top-10						
α/β TIM-barrel	29	6	»	7	20	2	13
O like Ferredoxin	17	4	2	2	17	»	8
α/β NTP Hydrolases containing P-loop	9	3	»	5	3	2	7
O Protein Kinases (catalytic core)	1	4	3	»	3	6	6
S Small inhibitors, toxins, lectins	14	»			3	»	»
α/β Rossmann Fold (NAD binding)	11	3	»	7	3	1	3
O RuBisCO (small subunit)	1	»		»	2		»
β like Concanavalin A	6	»		»	2	»	2
α like Hydrophobic Seed Protein	2	»			2		
α/β like Ribonuclease H	15	2	5	1	2	1	5
Metazoan Top-10							
β like Immunoglobulin	32	13	»	1	»	25	»
O Protein Kinases (catalytic core)	1	4	3	»	3	6	6
α DNA-binding 3-helical bundle	13	3	»	»	2	5	»
α like Globin	3	2		1	»	4	1
S Classic Zinc Finger	2	1	»			3	1
α/β NTP Hydrolases containing P-loop	9	3	»	5	3	2	7
β Trypsin-like serine proteases	4	1	1	»		2	»
α Cytochrome P450	1	1		»	»	2	1
S like Glucocort. receptor (DNA-binding)	4	1		»		2	»
α EF-hand	3	1		»	1	2	1

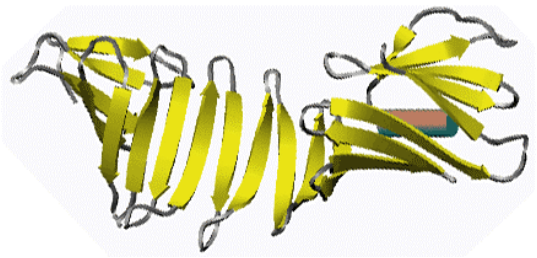
At What Structural Resolution Are Organisms Different?



Practical Relevance of Structural Genomics



(Pathogen only folds as possible targets)



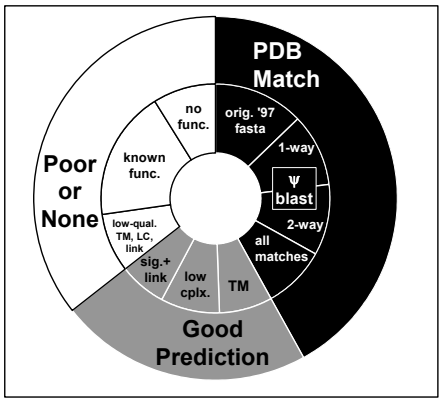
- OspA protein
  - ◊ in Lyme-disease spirochete *B. burgdorferi*
  - ◊ previously identified as the antigen for vaccine
  - ◊ has novel fold (C Lawson)

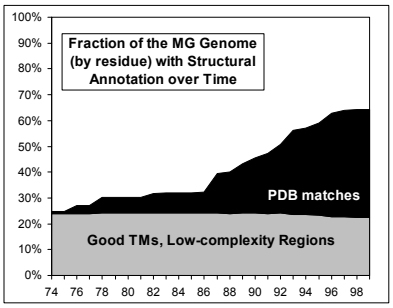
Large-scale Database Surveys (contents)

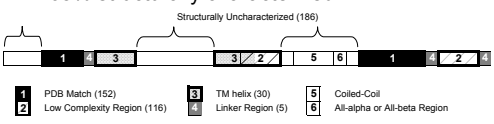
- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection
- Function Classification
- Cross-tabulation, folds and functions

Know All Folds in a Genome: How are we doing on MG?

- MG smallest genome with 479 ORFs
- Separate PDB Match, TMs, LC (SEG), linkers
- How many residues in genome matched by known folds, in 1975, '76, '77...'00...'50
- The impact of PSI-blast in comparison to pairwise methods
  - ◊ Two way PSI-blast gives an improvement (genome vs PDB, PDB vs. genome)
- Union of many sets of PDB matches finds >40% of a.a. and more than half the ORFs (242/479)
  - ◊ (Eisenberg, Godzik, Bork, Koonin, Frishman)
- ~65% structurally characterized

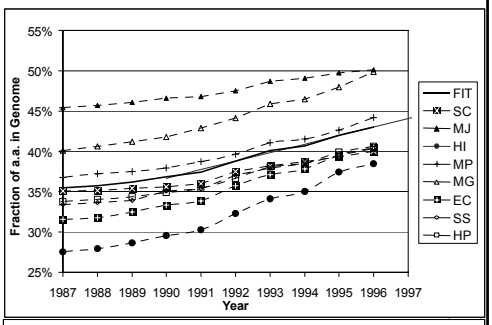


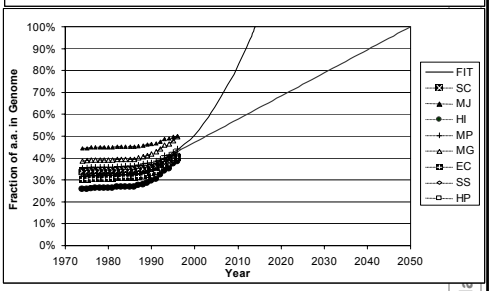


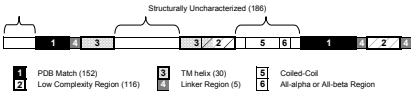


Know All Folds in Genome: MG Optimistic → Prediction

- Just use one pairwise method for matching
- Multiple, big genomes (e.g. SC)

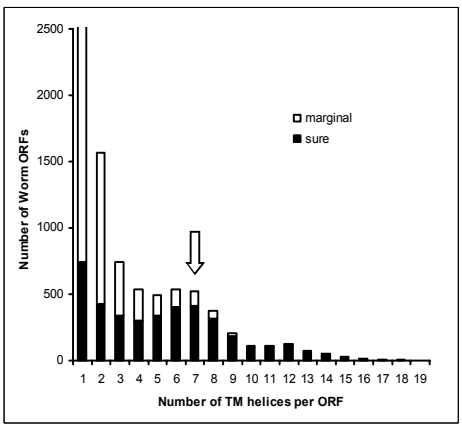


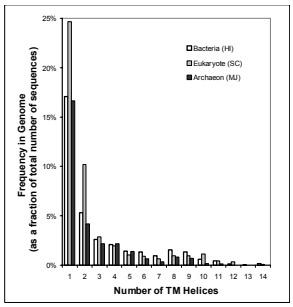


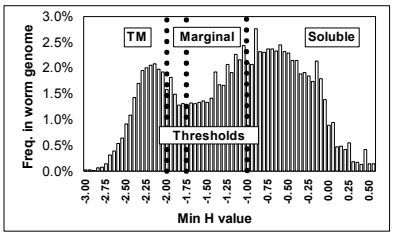


TM-helix "prediction"

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

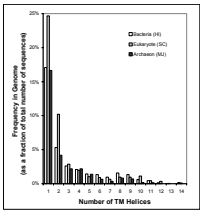


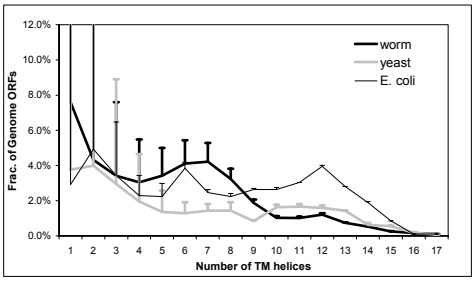




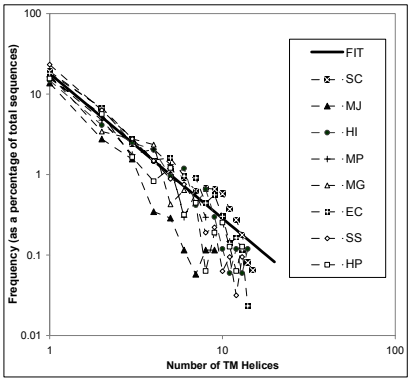
Comparative Genomics of Membrane Proteins

- Yeast has more mem. prots., esp. 2-TMs
- Similar conclusions to others: von Heijne, Rost, Jones, &c.



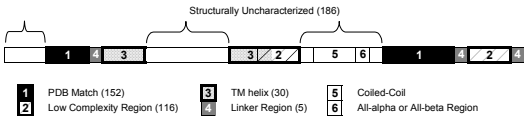


- Overall, no strong preference for particular supersecondary structures
  - ◊ Freq. of Number of TM helixes follows a Zipf-like law: $F=1/[5n^2]$
- In detail, worm has a peak for 7-TMs and *E. coli* for 12-TMs



2º Structure Prediction

- Bulk prediction of 2º struc. in genomes
- Same fraction of α and β (by element, half each)
- Both overall and only for unknown soluble proteins.



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

Fraction of residues Predicted to be in...	strand	helix
Avg	**17%**	**39%**
SD	**1%**	**2%**
EC	17%	39%
HI	16%	41%
HP	15%	42%
MG	17%	39%
MJ	19%	37%
MP	17%	39%
SC	17%	34%
SS	16%	38%

Not expected since……

Different Amino Acid Composition Should Give Different 2º Structure

Each a.a. has different propensity for local structure
->
Different Compositions (K from 4.4 in EC to 10.4 in MJ, Q too)
->
Different Local Structure (but compensation?)

Propensities from Regan (beta) and Baldwin (alpha)

	Amino Acid Composition								Propensity (kcal/mole)		
	EC	HI	SS	SC	HP	MP	MG	MJ	TM-hlx	helix	strand
K	4.4	6.3	4.2	7.3	8.9	8.6	9.5	10.4	8.8	-1.5	-0.4
C	1.2	1.0	1.0	1.3	1.1	.8	.8	1.3	-2	-1.1	-0.8
R	5.5	4.5	5.1	4.5	3.5	3.5	3.1	3.8	12.3	-1.9	-0.4
N	4.0	4.9	4.0	6.1	5.9	6.2	7.5	5.3	4.8	-1	-0.5
Q	4.4	4.6	5.6	3.9	3.7	5.4	4.7	1.5	4.1	-1.3	-0.4
A	9.5	8.2	8.5	5.5	6.8	6.7	5.6	5.5	-1.6	-1.9	0
I	6.0	7.1	6.3	6.6	7.2	6.6	8.2	10.5	-3.1	-1.2	-1.3
H	2.3	2.1	1.9	2.2	2.1	1.8	1.6	1.4	3	-1.1	-0.4
S	5.8	5.8	5.8	9.0	6.8	6.5	6.6	4.5	-0.6	-1.1	-0.9
M	2.8	2.4	2.0	2.1	2.2	1.6	1.5	2.2	-3.4	-1.4	-0.9
P	4.4	3.7	5.1	4.3	3.3	3.5	3.0	3.4	0.2	3	>3.0
G	7.4	6.6	7.4	5.0	5.8	5.5	4.6	6.3	-1	0	1.2
F	3.9	4.5	4.0	4.5	5.4	5.6	6.1	4.2	-3.7	-1	-1.1
E	5.7	6.5	6.0	6.5	6.9	5.7	5.7	8.7	8.2	-1.2	-0.2
Y	2.9	3.1	2.9	3.4	3.7	3.2	3.2	4.4	0.7	-1.2	-1.6
V	7.1	6.7	6.7	5.6	5.6	6.5	6.1	6.9	-2.6	-0.8	-0.9
T	5.4	5.2	5.5	5.9	4.4	6.0	5.4	4.0	-1.2	-0.6	-1.4
D	5.1	5.0	5.0	5.8	4.8	5.0	4.9	5.5	9.2	-1	0.9
L	10.6	10.5	11.4	9.6	11.2	10.3	10.7	9.5	-2.8	-1.6	-0.5
W	1.5	1.1	1.6	1.0	.7	1.2	1.0	.7	-1.9	-1.1	-1
total propensity											
α	-1.00	-1.02	-0.96	-1.00	-1.05	-1.03	-1.05	-1.01			
β	-0.27	-0.33	-0.26	-0.36	-0.37	-0.38	-0.42	-0.36			

Supersecondary structure words

- Look at super-secondary patterns ("words" such as αα or βαβ) in predictions
- Compare observed freq. with expected freq.
  odds = f(αβ)/f(α)f(β)
  (Freq. Words, Karlin)
- Do have differences between genomes (and PDB) here

HI more αα, ααα, αααα ...

SC more ββ, βββ, βββββ...

MJ more αβαβ, βαβα …

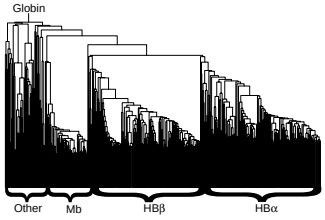
Super-Secondary Structure "Word"	Maximum Difference between 3 Genomes	Relative Abundance (Odds Ratio) HI	MJ	SC	PDB
ββ	26%	0.96	1.06	1.24	1.22
αα	15%	0.97	0.85	0.83	0.85
αβ	10%	1.09	1.09	0.99	0.95
βα	7%	0.98	1.00	0.93	0.99
βββ	41%	0.96	1.15	**1.46**	1.62
ααα	19%	**1.01**	0.83	0.84	0.92
αβα	18%	1.04	1.03	0.87	1.16
ααβ	15%	1.03	0.97	0.89	0.70
βαβ	12%	1.15	**1.24**	1.10	1.19
βαα	11%	0.93	0.87	0.83	0.78
ββα	9%	0.90	0.94	0.99	0.82
αββ	6%	0.97	0.98	1.03	0.80
ββββ	54%	1.03	1.35	**1.78**	2.28
αααα	29%	**1.10**	0.82	0.89	1.18
βββα	25%	0.85	0.94	1.10	0.98
βαβα	23%	1.11	**1.18**	0.94	1.48
αβαβ	21%	1.21	**1.23**	0.99	1.39
αβαα	21%	1.00	0.95	0.81	1.00
…	…	…	…	…	…

Large-scale Database Surveys (contents)

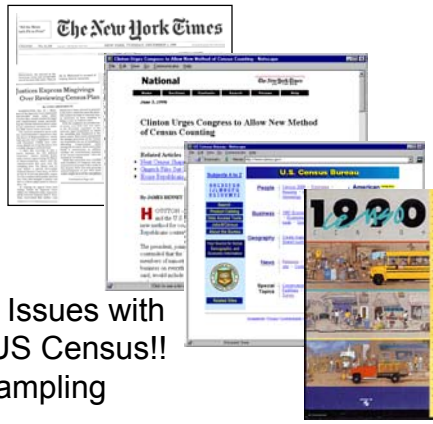
- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection
- Function Classification
- Cross-tabulation, folds and functions

An Issue with Fold Counting: Biases in the Databanks

Example Structure (PDB)	Fold Name	Percentage of known folds in genome	Rank in eubacterial Top-10
Top-10 in a bacterial genome (H. influenzae)			
2HSD-A	Rossmann Fold (NAD binding)	9.6	1
1AKE-A	NTP Hydrolases containing P-loop	5.7	3
1RCF	Flavodoxin-like	5.1	4
6TIM-B	TIM-barrel	4.5	2
1FXD	Ferredoxin-like	4.2	5
2RN2	like Ribonuclease H	3.0	16
1SBP	like Periplasmic binding protein (class II)	3.0	11
2DRI	like Periplasmic binding protein (class I)	3.0	19
1SRY-*	Class II aaRS and biotin synthetases	2.7	50
1PYP	OB-fold	2.7	9



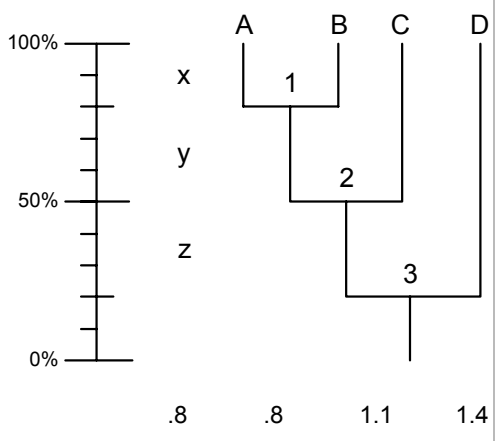
- Over-representation of certain species and functions in the databanks (e.g. human v. plant globins, Ig's)
  - Nevertheless HI top-10 like eubacterial top-10
- PDB small, biased sample of genome (6-12%)
- Diff. numbers with diff. comparison sensitivity
  - FASTA, HMM, &c
  - Some Correction with Seq. Weighting, Diff. Sampling
  - Uniform sampling is better than high sensitivity for some and low for others (ψ-blast problem)
  - Best to avoid FPs than FNs for Venn



Same Issues with Real US Census!! Sampling

Using a Tree to Correct for Biases

- Databank has biases.
- Assuming "fair" distribution spreads sequences uniformly through "space", want to weight sequences:
  - ◊ over-represented, down (mammal)
  - ◊ under-represented, up (plant & NV)
- Weights derived from a tree
  - ◊ Length of an unshared branch is allotted directly to sequence
  - ◊ Length of a shared branch is divided proportionally among sequences



Other schemes (Argos, Sander)

Class Notes

Paper topics due by end of week
Brief email to MG, DG, JS
(1 sentence to 1 paragraph)
We'll respond with a thumbs up or down

Probably won't get to simulation

New datamining notes

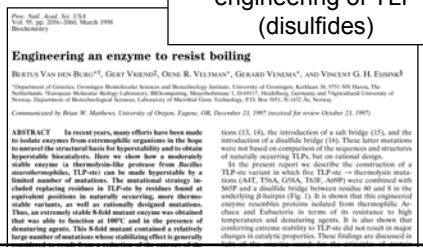
Different Perspectives on Protein Thermostability

In depth focus on single molecule vs. broad view of many (all?) proteins. Anectdotal vs. Comprehensive (the genomic perspective)

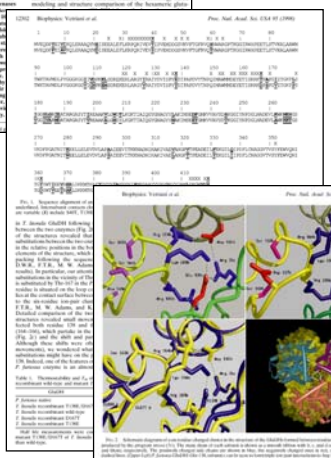
Protein thermostability above 100°C: A key role for ionic interactions

Ion pairs in GluDHs

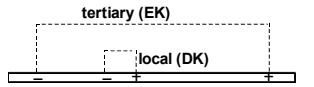
Change in entropy of unfolded state in engineering of TLP (disulfides)



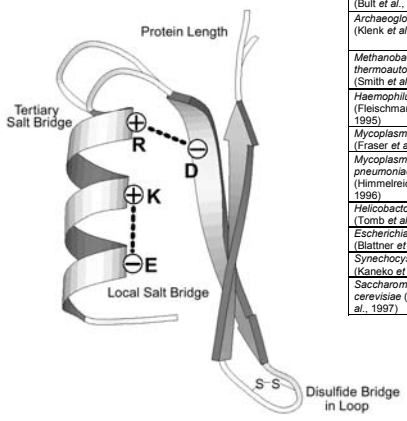
Engineering an enzyme to resist boiling



Thermostability: Analyzing a few Factors with Genome Comparison



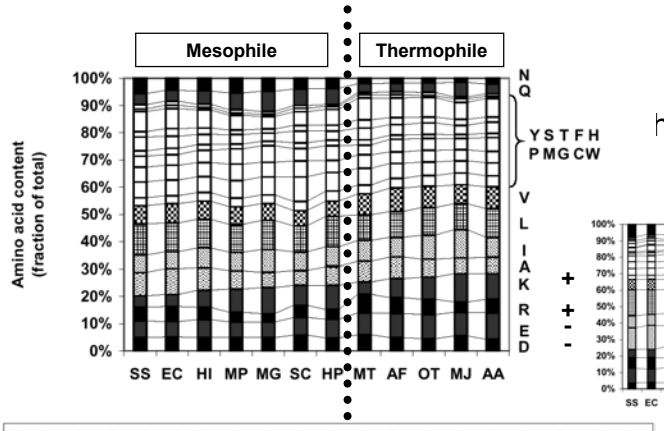
CEEEEHHHHHHHHHCCEEEEEEEEECC
CM**E**APAGNI**D**II**K**AGMKSPVQLTV**K**NDT



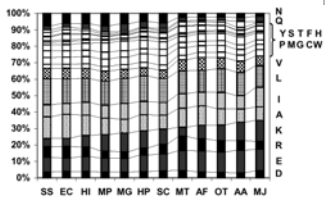
Organism	Category	Genome Abbreviation	# of Proteins	Physiological condition
Pyrococcus horikoshii (Strain OT3) (Kawarabayasi *et al.*, 1998)	archaea	OT	2061	98°C, anaerobe
Aquifex aeolicus (Deckert *et al.*, 1998)	eubacteria, gram negative	AA	1522	95°C
Methanococcus janaschii (Bult *et al.*, 1996)	archaea	MJ	1735	85°C, anaerobe
Archaeoglobus fulgidus (Klenk *et al.*, 1997)	archaea	AF	2409	83°C, anaerobe
Methanobacterium thermoautotrophicum (Smith *et al.*, 1997)	archaea	MT	1869	65°C, anaerobe
Haemophilus influenzae (Fleischmann *et al.*, 1995)	eubacteria, gram negative	HI	1680	mesophilic temp.
Mycoplasma genitalium (Fraser *et al.*, 1995)	eubacteria, gram positive	MG	470	mesophilic temp.
Mycoplasma pneumoniae (Himmelreich *et al.*, 1996)	eubacteria, gram positive	MP	677	mesophilic temp.
Helicobactor pylori (Tomb *et al.*, 1997)	eubacteria, gram negative	HI	1590	mesophilic temp.
Escherichia coli (Blattner *et al.*, 1997)	eubacteria, gram negative	EC	4288	mesophilic temp.
Synechocystis sp. (Kaneko *et al.*, 1996)	cyanobacteria	SS	3168	mesophilic temp.
Saccharomyces cerevisiae (Goffeau *et al.*, 1997)	eukaryote, fungus	SC	6218	mesophilic temp.

Composition Analysis of the Proteome

More Charged Residues in Thermophiles, Suggestive of Salt Bridges



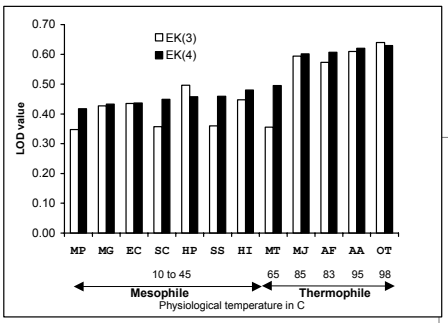
predicted helices in all ORFs

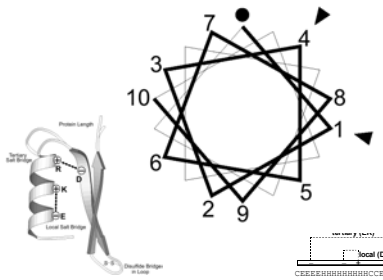


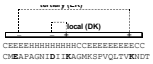
whole genome

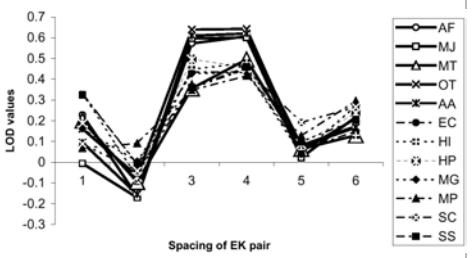
1-4 Spacing of Charged Residues More than Expected in Thermophile Helices ⇒ Salt Bridges

Quantify with LOD score
LOD = log (observed/expected)
For inst.,
expected[EK(4)] ~ f(E)*f(K)
LOD > 0, greater than expected

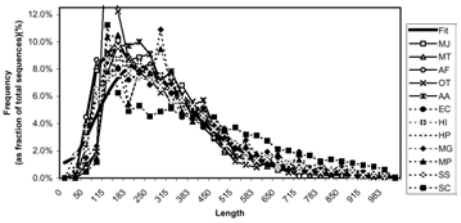


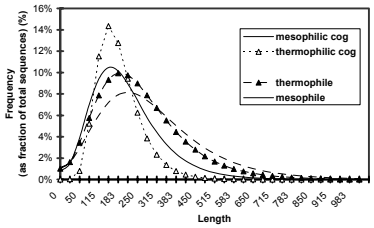




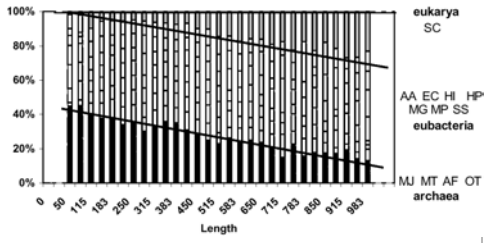


Sequence Length Doesn't Completely Relate to Thermostability





Simple distributions of sequence length have thermophiles shorter (Eisenberg)

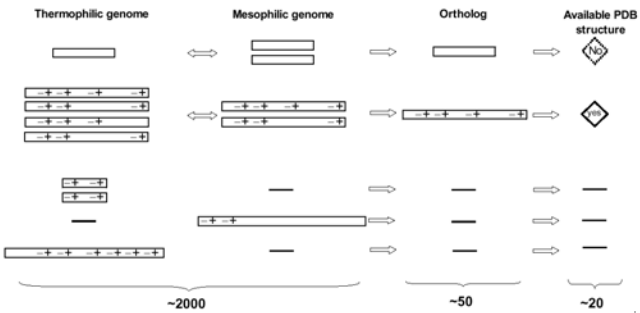


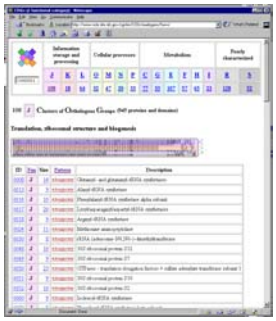
But this neglects special case of AA (eubacterial thermophile): archeal sequences shorter

Average	ARCHAEA ~281				BACTERIA ~326							EUKARYAOTE ~445	
Genome	MJ	MT	AF	OT	AA	EC	HI	HP	MG	MP	SS	SC	CE
Length	286	281	274	283	317	316	300	312	363	350	326	466	423

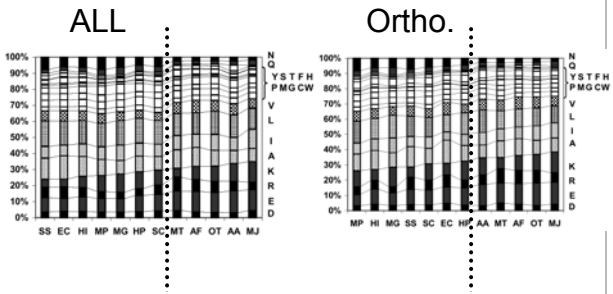
Controlling for Biases: Stratified Sample

Correct for duplications, repeats, unique families; Extend COGs to get 52 ortholog families





(COGs, Lipman, Koonin)

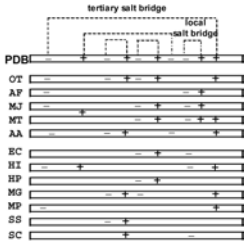


Meso, MT AF OT ***AA*** MJ Meso, ***AA*** MT AF OT MJ

Controls II: Known Structures, Random Genomes

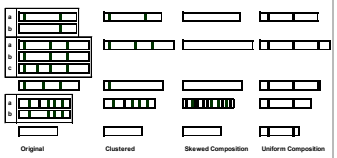
3D Structures

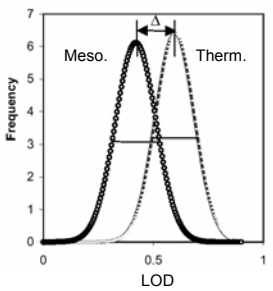
For orthologs of known structure: map tertiary salt bridges onto multiple alignment and look at conservation in Therm. vs. Meso.



COG		Cat.	PDB	Therm. Avg. SB	Meso Avg. SB	Diff.	
49	J	ribosomal	1rss	5.6	3.1	3	+
80	J	ribosomal	1aci	0.8	0.7	0.1	
81	J	ribosomal	1ad2	6.4	4.3	2.1	+
91	J	ribosomal	1bxe	1.8	0.9	0.9	
93	J	ribosomal	1whi	3	1.9	1.1	+
96	J	ribosomal	1sei	2	2.1	-0.1	
98	J	ribosomal	1pkp	0.6	1.7	-1.1	-
184	J	ribosomal	1a32	1.8	1.9	-0.1	
186	J	ribosomal	1rip	0.4	0.9	-0.5	
16	J	synthetase	1pys	7.6	2.6	5	+
124	J	synthetase	1ady	9.6	6.1	3.5	+
162	J	synthetase	2ts1	3.8	3.3	0.5	
30	J	other	1yub	5	5.3	-0.3	
125	F	other	1tmk	0.8	0.4	0.4	
149	C	other	1btm	3	4.3	-1.3	-
541	N	other	1fts	3.6	3.4	0.2	
112	E	other	1cj0	6.2	4.6	1.6	+
552	N	other	1ffh	4.2	4.6	-0.4	

Random Sampling: Make up random thermo. and meso. genomes, see what distribution of each statistic is





End of class on 11.27

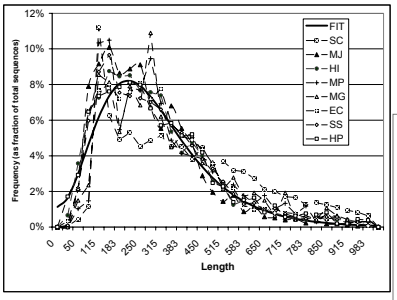
How Representative are the Known Structures of the Proteins in a Complete Genome? The issue of Bias

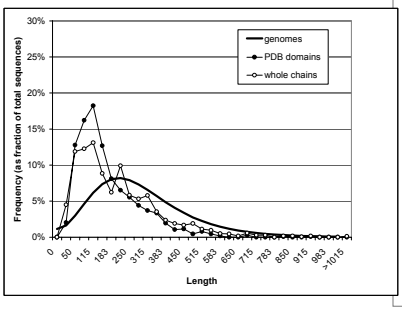
Assess 2º,TM predictions
- (+) comprehensive, statistical
- (-) predictions inaccurate (~65%)
- (-) extrapolate from PDB (esp. TM), domain problem

Is prediction (extrapolation) based on known structures justified?

Length: Genomes Sequences are longer than those in Known Structures

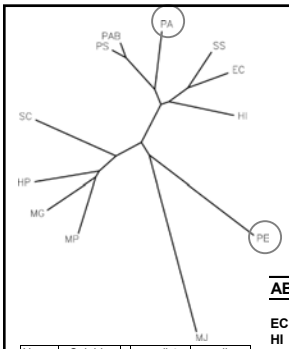
340 aa for avg. genome seq.
(470 aa for yeast)
205 aa for PDB chain
~160 aa for PDB domain





Amino Acid Composition

How Representative are the Known Structures of the Proteins in Complete Genome?



Name	Soluble PDB	= all-β	+ all-α
A	8.40%	6.8%	9.2%
C	1.72%	1.6%	1.4%
D	5.91%	5.9%	5.8%
E	6.29%	5.2%	7.3%
F	3.94%	4.2%	4.2%
G	7.79%	8.4%	6.4%
H	2.19%	2.1%	2.2%
I	5.54%	5.4%	5.1%
K	6.02%	5.6%	6.5%
L	8.37%	7.3%	9.6%
M	2.15%	1.7%	2.4%
N	4.57%	5.3%	4.4%
P	4.70%	5.1%	4.4%
Q	3.73%	3.5%	4.2%
R	4.78%	4.2%	5.4%
S	5.97%	7.2%	5.7%
T	5.87%	7.2%	5.2%
V	6.96%	7.6%	5.7%
W	1.46%	1.7%	1.5%
Y	3.64%	3.8%	3.5%

ABS.

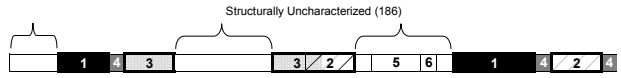
	rms	K	I	C	Q	W	N	F	L	G	A	P	S	R	H	M	E	D	T	Y	V
EC		4.4	6.0	1.2	4.4	1.5	4.0	3.9	10.6	7.4	9.5	4.4	5.8	5.5	2.3	2.8	5.7	5.1	5.4	2.9	7.1
HI		6.3	7.1	1.0	4.6	1.1	4.9	4.5	10.5	6.6	8.2	3.7	5.8	4.5	2.1	2.4	6.5	5.0	5.2	3.1	6.7
SS		4.2	6.3	1.0	5.6	1.6	4.0	4.0	11.4	7.4	8.5	5.1	5.8	5.1	1.9	2.0	6.0	5.0	5.5	2.9	6.7
SC		7.3	6.6	1.3	3.9	1.0	6.1	4.5	9.6	5.0	5.5	4.3	9.0	4.5	2.2	2.1	6.5	5.8	5.9	3.4	5.6
HP		8.9	7.2	1.1	3.7	.7	5.9	5.4	11.2	5.8	6.8	3.3	6.8	3.5	2.1	2.2	6.9	4.8	4.4	3.7	5.6
MP		8.6	6.6	.8	5.4	1.2	6.2	5.6	10.3	5.5	6.7	3.5	6.5	3.5	1.8	1.6	5.7	5.0	6.0	3.2	6.5
MG		9.5	8.2	.8	4.7	1.0	7.5	6.1	10.7	4.6	5.6	3.0	6.6	3.1	1.6	1.5	5.7	4.9	5.4	3.2	6.1
MJ		10.4	10.5	1.3	1.5	.7	5.3	4.2	9.5	6.3	5.5	3.4	4.5	3.8	1.4	2.2	8.7	5.5	4.0	4.4	6.9
AVG		7.5	7.3	1.1	4.2	1.1	5.5	4.8	10.5	6.1	7.0	3.8	6.4	4.2	1.9	2.1	6.5	5.1	5.2	3.3	6.4
SD		2.3	1.4	.2	1.3	.3	1.2	.8	.7	1.0	1.5	.7	1.3	.9	.3	.4	1.0	.3	.7	.5	.6

Diff.

		K	I	C	Q	W	N	F	L	G	A	P	S	R	H	M	E	D	T	Y	V
EC	16	-25	8	-29	19	7	-15	-2	28	-6	13	-5	-3	16	3	28	-7	-14	-7	-22	1
HI	17	8	27	-38	24	-21	6	12	26	-15	-2	-20	-2	-6	-7	10	5	-17	-11	-14	-4
SS	20	-29	13	-39	49	9	-13	1	37	-6	1	11	-3	6	-15	-8	-2	-16	-6	-20	-4
SC	21	24	18	-21	5	-27	31	14	15	-36	-34	-7	51	-7	-2	-4	5	-4	0	-8	-20
HP	27	52	29	-34	0	-51	27	36	34	-26	-18	-29	14	-28	-4	2	11	-20	-25	1	-20
MP	28	45	18	-55	44	-17	35	41	24	-29	-20	-25	8	-27	-18	-28	-8	-17	2	-11	-7
MG	36	61	48	-50	27	-32	62	53	28	-41	-33	-36	11	-35	-28	-30	-8	-18	-8	-11	-12
MJ	38	77	88	-23	-61	-49	14	6	14	-19	-35	-28	-25	-20	-35	1	40	-8	-31	20	-2
AVG		26	31	-36	13	-23	19	20	26	-22	-16	-17	6	-13	-13	-4	4	-14	-11	-8	-9
RMS		45	39	38	35	31	30	28	27	25	24	23	21	21	18	18	16	15	15	15	11

Composition of Different Regions of Genomes

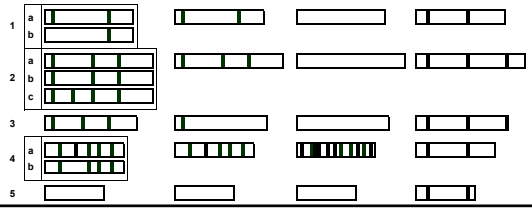
- Are composition differences uniform?
- Resampling
- Non-globular regions differ most in occurrence and composition
- Remove Repetitive Regions (SEG)



1 PDB Match (152)
2 Low Complexity Region (116)
3 TM helix (30)
4 Linker Region (5)
5 Coiled-Coil
6 All-alpha or All-beta Region

	AVG	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Statistics for Amino Acids										
Total Number	775998		1358465	505279	500616	170400	497968	237905	2900670	1033450
Fraction Masked by...										
PDB Match	8.7%	3.7%	11.1%	13.7%	8.8%	12.9%	7.1%	9.7%	6.2%	9.0%
Non-globular Region	21.7%	**6.9%**	16.7%	13.9%	22.2%	28.2%	35.1%	24.7%	23.9%	20.5%
TM-helix	4.9%	1.4%	7.3%	6.1%	4.8%	3.8%	2.9%	4.5%	5.2%	5.9%
Linker Region	5.1%	0.4%	5.3%	4.8%	4.8%	5.0%	5.0%	5.2%	4.6%	5.1%
Fraction Remaining										
Uncharacterized	59.7%	8.9%	59.6%	61.5%	59.4%	50.2%	49.9%	55.8%	60.0%	59.6%

	AVG	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Overall	23%	10%	16%	17%	27%	36%	38%	28%	21%	20%
PDB Match	18%	9%	12%	14%	24%	27%	34%	20%	12%	15%
Non Globular Region	**36%**	13%	32%	33%	39%	50%	52%	40%	42%	35%
TM-helix	**49%**	15%	55%	53%	55%	57%	55%	56%	56%	51%
Linker Region	27%	10%	22%	24%	29%	39%	33%	35%	21%	25%
Uncharacterized Region	23%	6%	15%	17%	26%	34%	32%	27%	20%	19%



PDB	Select	length	class	name
1sty	-	137	β	Staph nuclease
1cgp	a:9-137	129	β	CAP
1bgh	-	85	β	Gene V protein
1pht	-	83	β	SH3 domain
1tpf	a:	250	α/β	TIM
1wsy	a:	248	α/β	Trp Synthase
8dfr	-	186	α/β	DHFR
2rn2	-	155	α/β	Ribonuclease H
1brs	d:	87	α/β	Barstar
1gbs	-	185	α+β	Hen Lyzozyme
1l91	-	162	α+β	T4 lysozyme
1931	-	129	α+β	alpha-Lactabumin
7rsa	-	124	α+β	RNAse A
1brn	l:	108	α+β	Barnase
1fkd	-	107	α+β	FK506
9rnt	-	104	α+β	RNAse T1
1sha	a:	103	α+β	SH2 domain
1ubi	-	76	α+β	Ubiquitin
1cse	i:	63	α+β	CI-2 inhibitor
1igd	-	61	α+β	B1 domain
1mbd	-	153	α	Globin
1hrc	-	105	α	Cytochrome c
2wrp	r:	104	α	Trp Repressor
1lli	a:	89	α	Cro Repressor
1cop	d:	66	α	Lambda Repressor
1rpo	-	61	α	ROP
1myk	a:	47	α	Arc Repressor
2zta	a:	31	α	GCN4 zipper
1btl	-	263	M	beta-Lactamase
1bpi	-	58	S	BPTI
AVG		116		

Name	Hydroph. Polar	Soluble PDB	biophys. proteins	Rel. Diff.
		PS	BP	BP/PS -1
P	**H**	4.7%	3.7%	-21%
F	**H**	4.0%	3.2%	-19%
M	**H**	2.1%	1.8%	-16%
D	*P*	6.0%	5.1%	-16%
V	**H**	7.0%	6.2%	-12%
C	**H**	1.7%	1.5%	-9%
S	*P*	6.0%	5.7%	-5%
G	.	7.8%	7.7%	-1%
I	**H**	5.6%	5.5%	-1%
N	*P*	4.6%	4.6%	0%
W	**H**	1.4%	1.5%	1%
T	*P*	5.8%	6.0%	2%
L	**H**	8.4%	8.7%	5%
A	.	8.4%	8.8%	6%
Y	.	3.7%	3.9%	6%
H	*P*	2.2%	2.4%	6%
Q	*P*	3.7%	4.0%	6%
R	*P*	4.8%	5.2%	9%
E	*P*	6.2%	7.0%	13%
K	*P*	5.9%	7.7%	30%

Biophysical Proteins

Proteins that inform our view of the folding process -- as compared to the PDB.

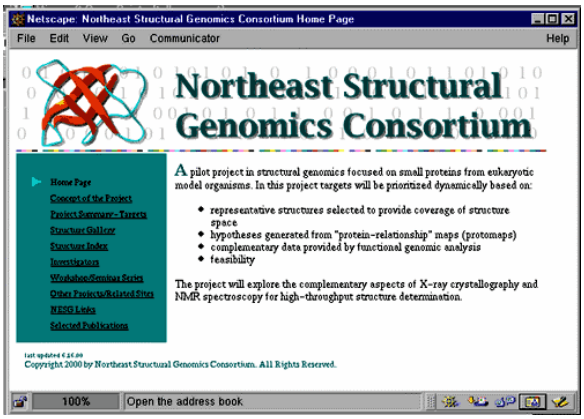
Shorter (116 v 161)

Fewer hydrophobes

Large-scale Database Surveys (contents)

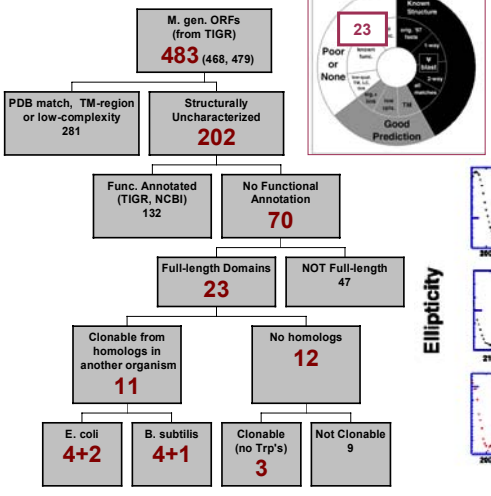
- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection
- Function Classification
- Cross-tabulation, folds and functions

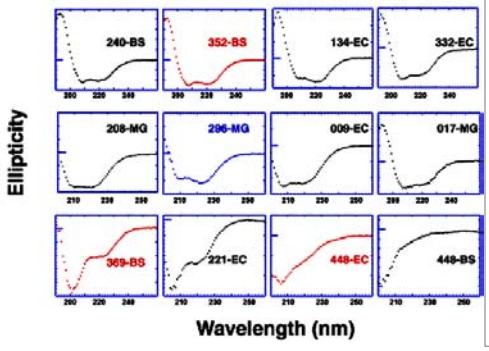
nesg.org



G Montelione

Finding Unusual Proteins for Expt. Structural Genomics

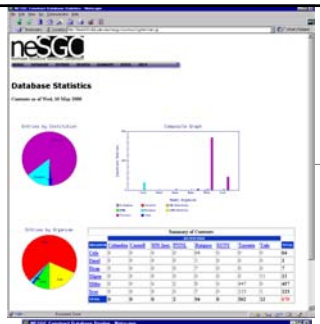




- **Prospective Target Selection**
- Identify Proteins in *M. genitalium* that are most atypical structurally (hardest)
- Characterize biophysically by CD (do they fold normally?)

L Regan

Tracking Database



Large-scale Database Surveys (contents)

- Fold Library
- Parts Lists: homologs, motifs, orthologs, folds
- Overall Sequence-structure Relationships, Annotation Transfer
- Parts in Genomes, shared & common folds
- Genome Trees
- Extent of Fold Assignment: the Bias Problem
- Bulk Structure Prediction
- The Genomic vs. Single-molecule Perspective
- Understanding Biases in Sampling
- Relationship to experiment: LIMS, target selection
- Function Classification
- Cross-tabulation, folds and functions