From Sequences to Science – New Perspectives on Protein Sequences and Structures

William R. Pearson
U. of Virginia

Department of Biochemistry and Molecular Genetics

20 Years of Biological Sequence Comparison

Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 726–730, February 1983
Biochemistry

***PNAS* (1983) 80:726**

Rapid similarity searches of nucleic acid and protein data banks

(global homology/optimal alignment)

W. J. WILBUR AND DAVID J. LIPMAN

Mathematical Research Branch, National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, National Institutes of Health, Building 31 Room 4B-54, Bethesda, Maryland 20205

RESEARCH ARTICLE

***Science* (1985) 227:1435**

Rapid and Sensitive Protein Similarity Searches

David J. Lipman and William R. Pearson

J. Mol. Biol. (1990) 215, 403–410

***J. Mol. Biol.* (1990) 215:403**

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

Department of Biochemistry and Molecular Genetics

From Sequences to Science –

- Homology is inferred from e*xcess* similarity – *A statistical perspective*
- Sequence similarity statistics are very accurate – *Unrelated sequences have scores indistinguishable from random sequences*
- Structure comparison and homology – a different perspective – *Can structures be "random" ?*
- Is protein folding difficult?
  - Different Structures for Similar functions
  - Are protein Sequences random?
  - Are all proteins built from small domains?

Department of Biochemistry and Molecular Genetics

New Perspectives on Sequence and Structure Comparison

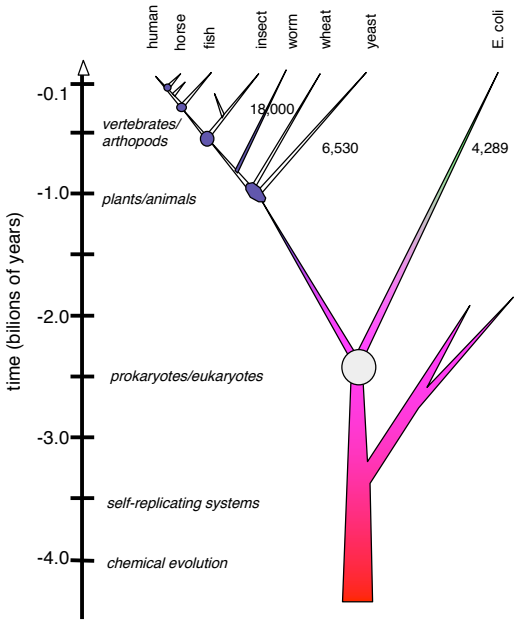
- Protein sequence comparison reliably identifies homologs (but not non-homologs) because *unrelated* sequences behave like *random* sequences – *Not random → homologous*
- Protein structure comparison can identify additional homologs, at the cost of *high false-positive* rates
- Structure comparison methods have less presumption that unrelated structures behave like random structures – *structure comparison statistics are very unreliable*
- Protein structures can be considered *random* – new structures with similar functions appear frequently, the distribution of protein words is random, and long domains emerged quickly
- The current protein universe samples a small fraction of possible protein structures

Department of Biochemistry and Molecular Genetics

From Sequences to Science –

- Homology is inferred from e*xcess* similarity – *A statistical perspective*
- Sequence similarity statistics are very accurate – *Unrelated sequences have scores indistinguishable from random sequences*
- Structure comparison and homology – a different perspective – *Can structures be "random" ?*
- Is protein folding difficult?
  - Different Structures for Similar functions
  - Are protein Sequences random?
  - Are all proteins built from small domains?

Department of Biochemistry and Molecular Genetics



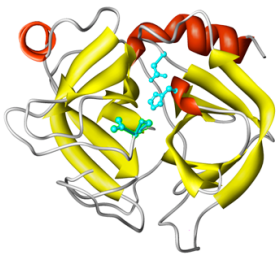
Department of Biochemistry and Molecular Genetics

E. coli proteins vs Human – Ancient Protein Domains

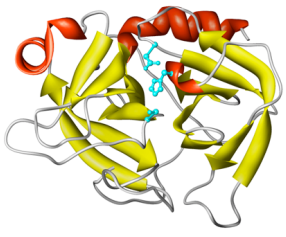
expect	%_id	alen	E coli descr	Human descr	sp_name
2.7e-206	53.8	944	glycine decarboxylase, P	Glycine dehydrogenase [de	GCSP_HUMAN
1.2e-176	59.5	706	methylmalonyl-CoA mutase	Methylmalonyl-CoA mutase,	MUTA_HUMAN
3.8e-176	50.6	803	glycogen phosphorylase [E	Glycogen phosphorylase, l	PHS1_HUMAN
9.9e-173	55.6	1222	B12-dependent homocystein	5-methyltetrahydrofolate-	METH_HUMAN
1.8e-165	41.8	1031	carbamoyl-phosphate synth	Carbamoyl-phosphate synth	CPSM_HUMAN
5.6e-159	65.7	542	glucosephosphate isomeras	Glucose-6-phosphate isome	G6PI_HUMAN
8.1e-143	53.7	855	aconitate hydrase 1 [Esch	Iron-responsive element b	IRE1_HUMAN
2.5e-134	73.0	459	membrane-bound ATP syntha	ATP synthase beta chain,	ATPB_HUMAN
3.3e-121	55.8	550	succinate dehydrogenase,	Succinate dehydrogenase [	DHSA_HUMAN
1.5e-113	60.6	401	putative aminotransferase	Cysteine desulfurase, mit	NFS1_HUMAN
4.4e-111	60.9	460	fumarase C= fumarate hydr	Fumarate hydratase, mitoc	FUMH_HUMAN
1.5e-109	56.1	474	succinate-semialdehyde de	Succinate semialdehyde de	SSDH_HUMAN
3.6e-106	44.7	789	maltodextrin phosphorylas	Glycogen phosphorylase, m	PHS2_HUMAN
1.4e-102	53.1	484	NAD+-dependent betaine al	Aldehyde dehydrogenase, E	DHAG_HUMAN
3.8e-98	53.0	449	pyridine nucleotide trans	NAD(P) transhydrogenase,	NNTM_HUMAN
5.8e-96	49.9	489	glycerol kinase [Escheric	Glycerol kinase, testis s	GKP2_HUMAN
2.1e-95	66.8	328	glyceraldehyde-3-phosphat	Glyceraldehyde 3-phosphat	G3P2_HUMAN
5.0e-91	62.5	368	alcohol dehydrogenase cla	Alcohol dehydrogenase cla	ADHX_HUMAN
6.7e-91	56.5	393	protein chain elongation	Elongation factor Tu, mit	EFTU_HUMAN
9.5e-91	56.6	392	protein chain elongation	Elongation factor Tu, mit	EFTU_HUMAN
2.2e-89	59.1	369	methionine adenosyltransf	S-adenosylmethionine synt	METK_HUMAN
6.5e-88	53.3	422	enolase [Escherichia coli	Alpha enolase (2-phospho-	ENOA_HUMAN
9.2e-88	43.3	536	NAD-linked malate dehydro	NADP-dependent malic enzy	MAOX_HUMAN
7.3e-86	55.5	389	2-amino-3-ketobutyrate Co	2-amino-3-ketobutyrate co	KBL_HUMAN
5.2e-83	44.4	543	degrades sigma32, integra	AFG3-like protein 2 (Para	AF32_HUMAN

Department of Biochemistry and Molecular Genetics

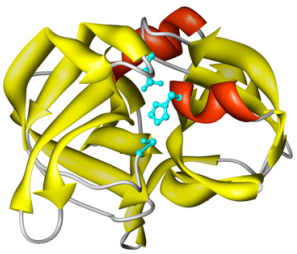
Homology => structural similarity
? sequence similarity



Bovine trypsin (5ptp)
Structure: E()< 10^{-23}; RMSD 0.0 A
Sequence: E()< 10^{-84}
100% 223/223



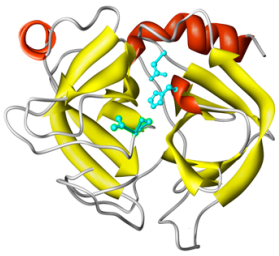
S. griseus trypsin (1sgt)
E()< 10^{-14} RMSD 1.6 A
E()< 10^{-19} 36%; 226/223



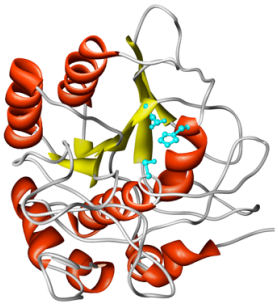
S. griseus protease A (2sga)
E()< 10^{-4}; RMSD 2.6 A
E()< 2.6 25%; 199/181

Department of Biochemistry and Molecular Genetics

Non-homologous proteins have different structures



Bovine trypsin (5ptp)
Structure: E()<10^{-23}
RMSD 0.0 A
Sequence: E()<10^{-84}
100% 223/223

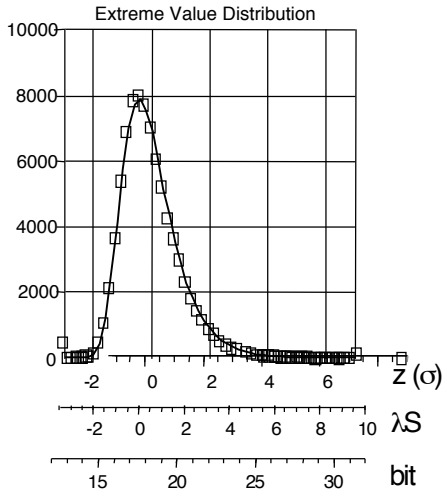


Subtilisin (1sbt)
E() >100
E()<280; 25% 159/275

Cytochrome c4 (1etp)
E() > 100
E()<5.5; 23% 171/190

Department of Biochemistry and Molecular Genetics

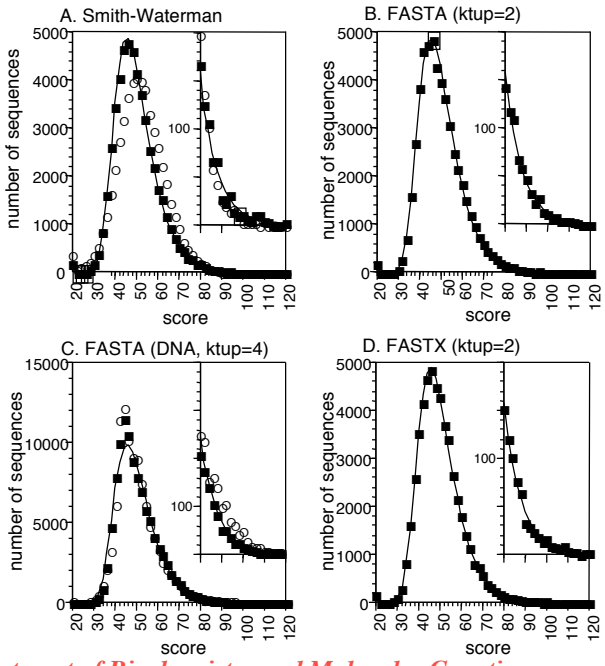
What is Excess Similarity? – Statistical Models of Similarity Scores



$S' = \lambda S_{raw} - \ln K\ m\ n$
$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$
$P(S'>x) = 1 - \exp(-e^{-x})$
$P(S_{bit} > x) = 1 - \exp(-mn2^{-x})$
$E(S'>x \mid D) = P\ D$

$P(B \text{ bits}) = m\ n\ 2^{-B}$
$P(40 \text{ bits}) = 1.5 \times 10^{-7}$
$E(40 \mid D=4000) = 6 \times 10^{-4}$
$E(40 \mid D=2E6) = 0.3$

Department of Biochemistry and Molecular Genetics



Inferring Homology from Statistical Significance

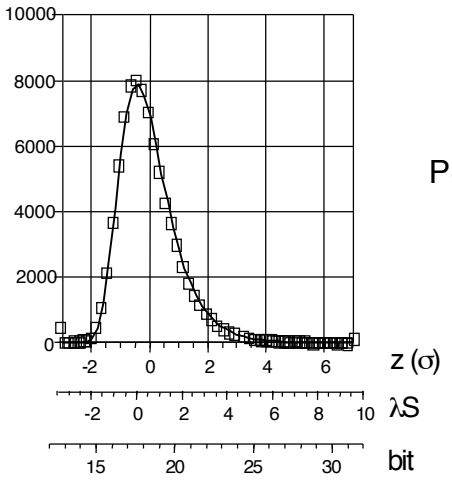
- Real *UNRELATED* sequences have similarity scores that are indistinguishable from *RANDOM* sequences
- If a similarity is NOT *RANDOM,* then it must be NOT *UNRELATED*
- Therefore, NOT *RANDOM* (statistically significant) similarity must reflect *RELATED* sequences

From Sequences to Science –

- Homology is inferred from e*xcess* similarity – *A statistical perspective*
- Sequence similarity statistics are very accurate – *Unrelated sequences have scores indistinguishable from random sequences*
- Structure comparison and homology – a different perspective – *Can structures be "random" ?*
- Is protein folding difficult?
  - Different Structures for Similar functions
  - Are protein Sequences random?
  - Are all proteins built from small domains?

Department of Biochemistry and Molecular Genetics

Extreme value distribution



$S' = \lambda S_{raw} - \ln K\ m\ n$

$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$

$P(S'>x) = 1 - \exp(-e^{-x})$

$P(S_{bit} > x) = 1 - \exp(-mn2^{-x})$

$E(S'>x \mid D) = P\ D$

$P(B \text{ bits}) = m\ n\ 2^{-B}$

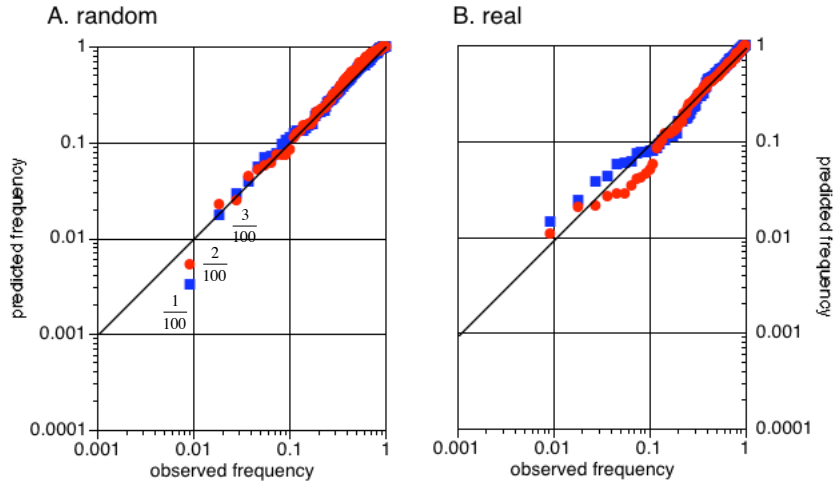
$P(40 \text{ bits}) = 1.5 \times 10^{-7}$

$E(40 \mid D=4000) = 6 \times 10^{-4}$

$E(40 \mid D=2E6) = 0.3$

Department of Biochemistry and Molecular Genetics

Protein Sequence Comparison Statistics are Accurate



Inferring Homology from Statistical Significance

- Real *UNRELATED* sequences have similarity scores that are indistinguishable from *RANDOM* sequences
- If a similarity is NOT *RANDOM*, then it must be NOT *UNRELATED*
- Therefore, NOT *RANDOM* (statistically significant) similarity must reflect *RELATED* sequences

From Sequences to Science –

- The Inference of Homology from Significant Similarity – *A statistical perspective*
- The Accuracy of Sequence Similarity Statistics
- Structure Comparison and Homology – a different perspective
- Is protein folding difficult?
  - Different Structures for Similar functions
  - Are protein Sequences random?
  - Are all proteins built from small domains?

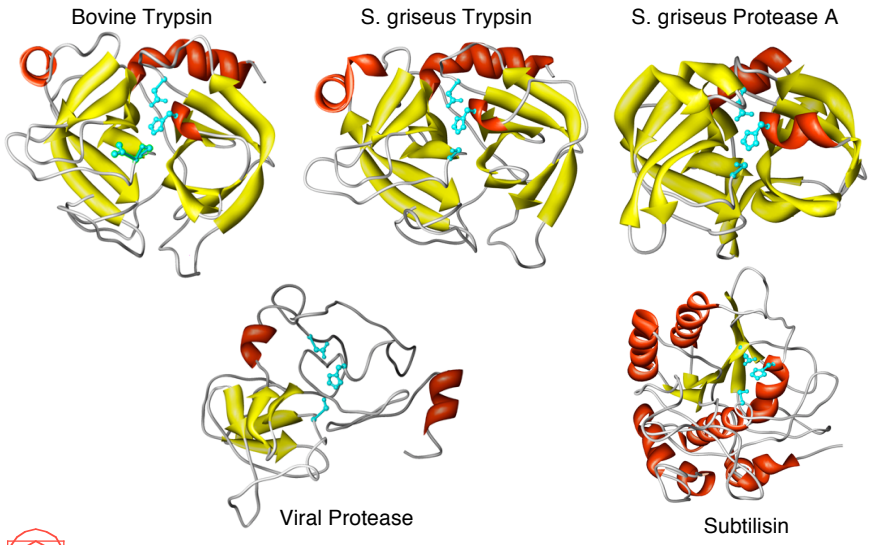
Department of Biochemistry and Molecular Genetics

Homology from Similarity – Sequence or Structure?

- Structure comparison is the "gold standard" for establishing homology – structures change more slowly than sequence
- Structure comparison problems –
  - Structures are not unique (differ by > 1.5 Å for identical sequences)
  - No optimal alignment algorithm
  - Poor understanding of statistics - no "random" structures
- Statistical significance of structural similarity rarely quantified - homology vs analogy (convergence).

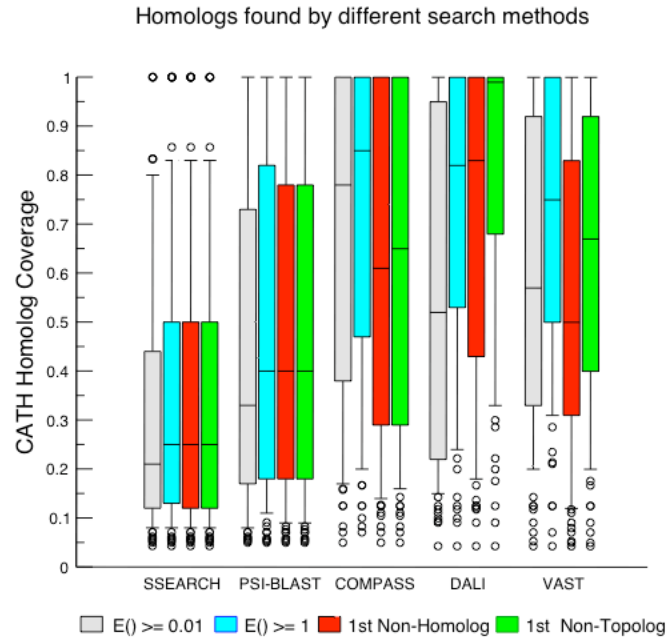
Department of Biochemistry and Molecular Genetics

Homologs, Topologs, and Convergence



Homlogy, Similarity, and Convergence – Serine Proteases

		CATH Homology			Topology	Convergent
		Bovine Trypsin	S. griseus Trypsin	S. griseus Protease A	Viral Protease	Subtilisin
		5PTP vs. :	1SGT	2SGA	1BEF	1SBT
Structure/ Structure →	**Dali**	Z	32.7	13.7	8.8	<2
		E(2775)	10^{-14}	10^{-4}	0.02	>100
		N_{align} (%id)	209 (34)	147 (19)	131 (10)	N/A
		RMSD (Å)	1.4	2.8	2.9	N/A
→	**VAST**	E(2775)	10^{-21}	0.017 [a]	1.94	N/A
		N_{align} (%id)	208 (34)	130 (22)	122 (14)	N/A
		RMSD (Å)	1.5	2.3	2.8	N/A
	COMPASS	E(10000)	10^{-114}	10^{-13}	0.056	13 ← Profile/ Profile
Profile/ Sequence →	**PSI-BLAST**	E(2775)	10^{-48}	2.5	>10	>10
		N_{align}	231	40	N/A	N/A
	SSEARCH	E(10000)	10^{-19}	2.6	>10	>10 ← Sequence/ Sequence
		N_{align} (%id)	223 (36)	181 (25)	68 (33)	159 (25)

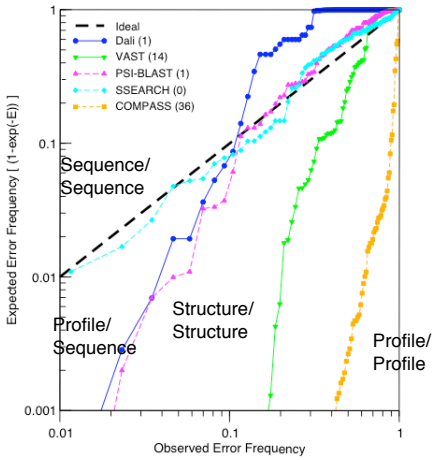


Inferring Homology from Statistical Significance

- Real *UNRELATED* sequences have similarity scores that are indistinguishable from *RANDOM* sequences
- If a similarity is NOT *RANDOM,* then it must be NOT *UNRELATED*
- Therefore, NOT *RANDOM* (statistically significant) similarity must reflect *RELATED* sequences

1. *Should Unrelated Structures have E() ≥ 1 ?*
2. *Are there "chance" Structural Similarities?*

Accuracy of statistical estimates



- SSEARCH (Smith-Waterman) provides very accurate statistical estimates
- PSI-BLAST and Dali provide estimates that off by 10–100-fold
- Other structure comparison methods provide wild over estimates of statistical significance – *BEWARE of claims of significant structural similarity*

Department of Biochemistry and Molecular Genetics

Structure Comparison Statistics

- Most structure comparison methods report very significant structural similarity for non-homologous proteins (*unrelated ≠ random*)
- These significance estimates are used to infer *ancient domain homologies*, which are preferred to *multiple independent origins*
- Dali produces relatively accurate estimates, and is one of the most sensitive search methods – thus, unrelated structures may be random
- If structural similarity can be random, there may be many more possible structures than existing ones

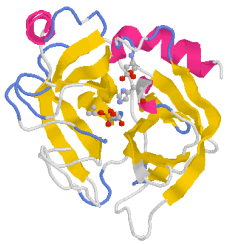
Department of Biochemistry and Molecular Genetics

From Sequences to Science –

- Homology is inferred from *excess* similarity – *A statistical perspective*
- Sequence similarity statistics are very accurate – *Unrelated sequences have scores indistinguishable from random sequences*
- Structure comparison and homology – a different perspective – *Can structures be "random" ?* *YES*
- Is protein folding difficult?
  - Different Structures for Similar functions
  - Are protein Sequences random?
  - Are all proteins built from small domains?

Department of Biochemistry and Molecular Genetics

Independent Emergence of Function – 2 distinct serine proteases

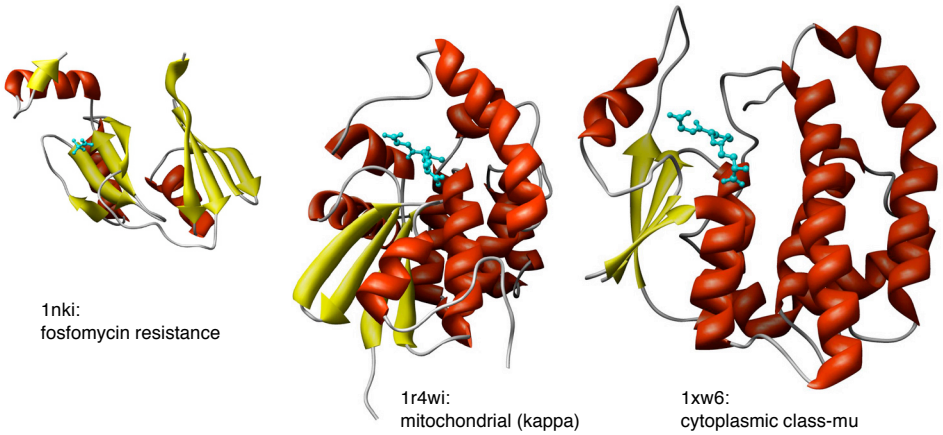


Bovine trypsin (5ptp)
Structure: E()<10^{-23}
RMSD 0.0 A
Sequence: E()<10^{-84}
100% 223/223

Subtilisin (1sbt)
E() > 100
E()<280; 25% 159/275

Department of Biochemistry and Molecular Genetics

Independent Emergence of Function – Glutathione Transferase Activity



Department of Biochemistry and Molecular Genetics

Re-creating functions from convergent structures

- Subtilisin/Trypsin (same chemical structure)
- Other protease families
- Glutathione transferases (at least 4 independent events)
- Non-orthologous displacement – replacing a homolog with a non-homolog

Nature frequently re-creates an existing function

Department of Biochemistry and Molecular Genetics

From Sequences to Science –

- Homology is inferred from excess similarity – *A statistical perspective*
- Sequence similarity statistics are very accurate – *Unrelated sequences have scores indistinguishable from random sequences*
- Sequence comparison and structure comparison – *sensitivity and statistics –*
- Structure comparison and homology – a different perspective – *Can structures be "random" ?*
- Is protein folding difficult?
  - Different Structures for Similar functions
  - Are protein Sequences random?
  - Are all proteins built from small domains?

Department of Biochemistry and Molecular Genetics

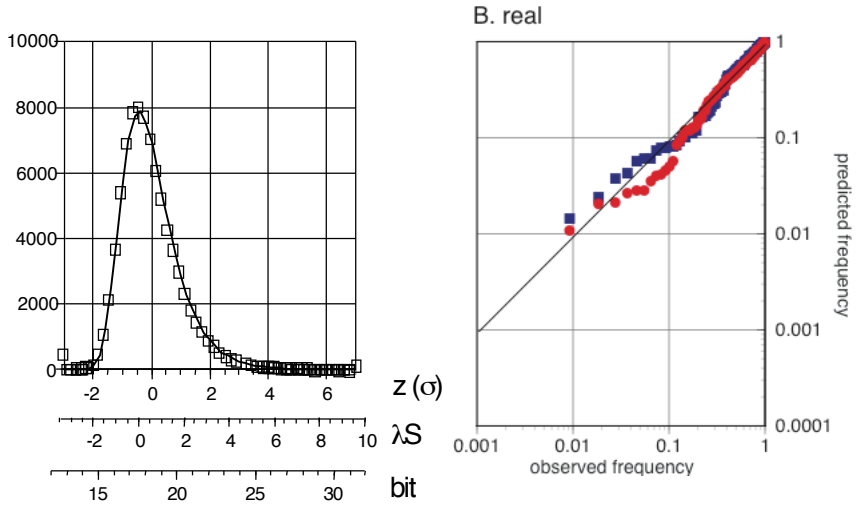
Different Perspectives on Protein Folding and Protein Structure

- Protein folding is very hard
- $<10^{-23}$ can fold
- Very few sequences can fold (sequence space is sparse)
- Common protein folds reflect convergent evolution to a small number of stable structures
- Most possible structures have been seen

- Protein folding is difficult, but not nearly impossible (sequence space is uniformly populated)
- Common structures reflect evolutionary history and nature's trials, not an exhaustive search
- New genomes typically have 20% "new" proteins
- New sequences will have new structures

Department of Biochemistry and Molecular Genetics

Sequence Similarity Scores are Random

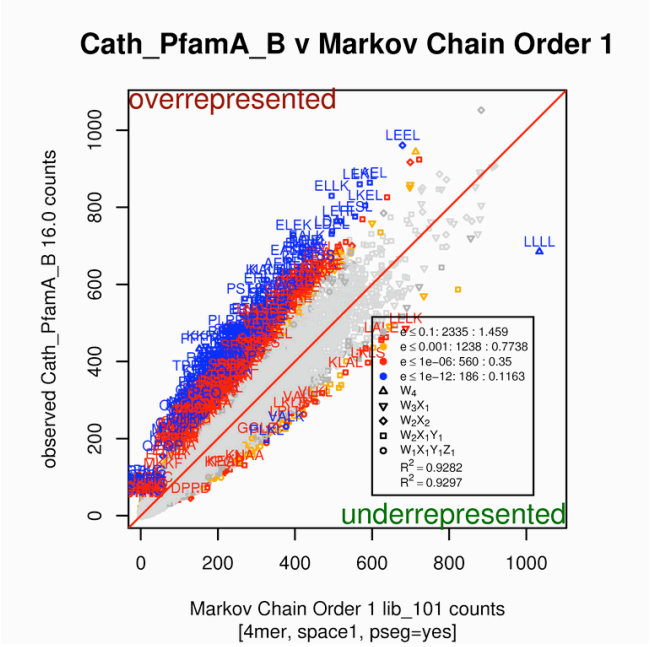


Protein sequence space – sparse or uniform?

- Examine small peptide words in proteins from a non-redundant library of sequences.
- For short n-mers (3 – 5 amino acids), one can compare frequencies from real and random sequences
  - 3 aa ($20^3 = 10^{3.9}$ = 8,000 unique 3mers)
  - 4 aa ($20^4 = 10^{5.2}$ = 160,000 unique 4mers)
  - 5 aa ($20^5 = 10^{3.9}$ = 3,200,000 unique 5mers)
- Can real proteins be distinguished from random sequences based on oligo-peptide (word) counts?

One (protein) family - One vote

- Pfam-A ver 12.0 7316 domains
- Remove fragment domains: 7006
- Single Linkage clustering E()< 1e-03: 6,956 domains
- ~1.5 million aa



Department of Biochemistry and Molecular Genetics

K_1 characterizes the complexity of a word

$$\Omega = \frac{L!}{\prod_{i=1}^{N} n_i!} \quad where\ L = 4,\ N = 20$$

$$K = \frac{1}{L}\log_2 \Omega$$

$$\Omega = \frac{4!}{3! \times 1!} = 4$$

$$K_{3,1} = \frac{1}{4}\log_2 4 = 0.5$$

Type	Number	Percent	K_1
W_4	20	0.0125	0
W_3X_1 GGPG	1,520	0.95	0.5
W_2X_2	1,140	0.71	0.65
$W_2X_1Y_1$	41,040	25.7	0.90
$W_1X_1Y_1Z_1$	116,280	72.7	1.15

Department of Biochemistry and Molecular Genetics

Summary of exceptional words by complexity type at *E* () < 0.001

Type	Number	1	%	2	4
W_4	20	13	65	11	16
W_3X_1	1,520	369	24	289	323
W_2X_2	1,140	168	15	137	151
$W_2X_1Y_1$	41,040	1,436	3.5	1,025	676
$W_1X_1Y_1Z_1$	116,280	661	0.6	579	75
Total	160,000	2647		2,041	1,241

Department of Biochemistry and Molecular Genetics

Are protein sequences random?

- Simply counting n-mer's (3, 4, 5-) from comprehensive databases shows significant differences from random
- These differences reflect family preferences for n-mers
- When each protein family is counted once, some n-mers are over-represented, but less than 0.6% when di-amino acid frequencies are preserved
- When low complexity regions are removed < 700/116,000 4-mers are over-represented

Yes

Department of Biochemistry and Molecular Genetics

New Perspectives on Sequence and Structure Comparison

- Protein sequence comparison reliably identifies homologs (but not non-homologs) because *unrelated* sequences behave like *random* sequences – *Not random → homologous*
- Protein structure comparison can identify additional homologs, at the cost of *high false-positive* rates
- Structure comparison methods have less presumption that unrelated structures behave like random structures – *structure comparison statistics are very unreliable*
- Protein structures can be considered *random* – new structures with similar functions appear frequently, the distribution of protein words is random, and long domains emerged quickly
- The current protein universe samples a small fraction of possible protein structures

Department of Biochemistry and Molecular Genetics

From Sequences to Science –

- Homology is inferred from *excess* similarity – *A statistical perspective*
- Sequence similarity statistics are very accurate – *Unrelated sequences have scores indistinguishable from random sequences*
- Sequence comparison and structure comparison – *sensitivity and statistics –*
- Structure comparison and homology – a different perspective – *Can structures be "random" ?*
- Is protein folding difficult?
  - Different Structures for Similar functions
  - Are protein Sequences random?
  - Are all proteins built from small domains?

Department of Biochemistry and Molecular Genetics

E. coli proteins vs Human – Ancient Protein Domains

expect	%_id	alen	E coli descr	Human descr	sp_name
2.7e-206	53.8	944	glycine decarboxylase, P	Glycine dehydrogenase [de	GCSP_HUMAN
1.2e-176	59.5	706	methylmalonyl-CoA mutase	Methylmalonyl-CoA mutase,	MUTA_HUMAN
3.8e-176	50.6	803	glycogen phosphorylase [E	Glycogen phosphorylase, l	PHS1_HUMAN
9.9e-173	55.6	1222	B12-dependent homocystein	5-methyltetrahydrofolate-	METH_HUMAN
1.8e-165	41.8	1031	carbamoyl-phosphate synth	Carbamoyl-phosphate synth	CPSM_HUMAN
5.6e-159	65.7	542	glucosephosphate isomeras	Glucose-6-phosphate isome	G6PI_HUMAN
8.1e-143	53.7	855	aconitate hydrase 1 [Esch	Iron-responsive element b	IRE1_HUMAN
2.5e-134	73.0	459	membrane-bound ATP syntha	ATP synthase beta chain,	ATPB_HUMAN
3.3e-121	55.8	550	succinate dehydrogenase,	Succinate dehydrogenase [	DHSA_HUMAN
1.5e-113	60.6	401	putative aminotransferase	Cysteine desulfurase, mit	NFS1_HUMAN
4.4e-111	60.9	460	fumarase C= fumarate hydr	Fumarate hydratase, mitoc	FUMH_HUMAN
1.5e-109	56.1	474	succinate-semialdehyde de	Succinate semialdehyde de	SSDH_HUMAN
3.6e-106	44.7	789	maltodextrin phosphorylas	Glycogen phosphorylase, m	PHS2_HUMAN
1.4e-102	53.1	484	NAD+-dependent betaine al	Aldehyde dehydrogenase, E	DHAG_HUMAN
3.8e-98	53.0	449	pyridine nucleotide trans	NAD(P) transhydrogenase,	NNTM_HUMAN
5.8e-96	49.9	489	glycerol kinase [Escheric	Glycerol kinase, testis s	GKP2_HUMAN
2.1e-95	66.8	328	glyceraldehyde-3-phosphat	Glyceraldehyde 3-phosphat	G3P2_HUMAN
5.0e-91	62.5	368	alcohol dehydrogenase cla	Alcohol dehydrogenase cla	ADHX_HUMAN
6.7e-91	56.5	393	protein chain elongation	Elongation factor Tu, mit	EFTU_HUMAN
9.5e-91	56.6	392	protein chain elongation	Elongation factor Tu, mit	EFTU_HUMAN
2.2e-89	59.1	369	methionine adenosyltransf	S-adenosylmethionine synt	METK_HUMAN
6.5e-88	53.3	422	enolase [Escherichia coli	Alpha enolase (2-phospho-	ENOA_HUMAN
9.2e-88	43.3	536	NAD-linked malate dehydro	NADP-dependent malic enzy	MAOX_HUMAN
7.3e-86	55.5	389	2-amino-3-ketobutyrate Co	2-amino-3-ketobutyrate co	KBL_HUMAN
5.2e-83	44.4	543	degrades sigma32, integra	AFG3-like protein 2 (Para	AF32_HUMAN

Department of Biochemistry and Molecular Genetics

Ancient conserved domains

Accession #	NAME	LENGTH	Domain Coverage	TREE	PDB %id	Spriochaetales b.burg	alphaproteo e.coli	gammaproteo R.prok	Cyanobacteria Synecho	Gram_positive B.Subtillius	Archaea Pyrob.aero	Archaea M.jan
P10809	60 kDa heat shock protein,	573	cpn60_TCP1 0.86	Tree	PDB 0.532	2.9e-109 0.497	5.7e-104 0.509	1.5e-118 0.543	1.3e-100 0.430	8e-111 0.514	-	-
P07954	Fumarate hydratase, mitoch	510	lyase_1 0.86	Tree	PDB 0.664	-	3.4e-111 0.606	1.5e-120 0.628	9e-116 0.587	1.8e-113 0.578	2.1e-52 0.371	-
P00395	Cytochrome c oxidase polyp	513	COX1 0.89	Tree	PDB 0.916	-	5.9e-77 0.411	2.7e-143 0.575	1.2e-96 0.436	5e-90 0.408	1.1e-68 0.366	-
NP_112486	tRNA-guanine transglycosyl	403	TGT 0.59	Tree	PDB 0.432	1.8e-52 0.366	2.8e-64 0.434	2.1e-59 0.418	1e-65 0.468	3.8e-75 0.463	-	4e-09 0.250
P04424	Argininosuccinate lyase (E	464	lyase_1 0.95	Tree	PDB 0.996	-	6.7e-77 0.451	-	2.1e-77 0.435	1.9e-86 0.443	4e-23 0.261	7.6e-51 0.333
P47895	Aldehyde dehydrogenase 6 (	512	aldedh 0.92	Tree	PDB 0.728	-	7.6e-61 0.335	-	2.8e-51 0.364	5.5e-73 0.408	4e-68 0.383	4.5e-61 0.364
P13716	Delta-aminolevulinic acid	330	ALAD 0.97	Tree	PDB 1.000	-	7.7e-49 0.429	2.2e-44 0.364	5.6e-51 0.394	7.1e-58 0.448	4.6e-53 0.465	6.5e-54 0.432

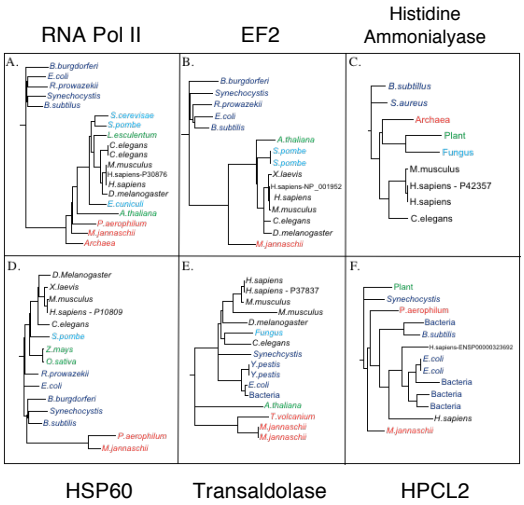
Department of Biochemistry and Molecular Genetics

Identification of Ancient Long One-Domain Proteins

Protein Classifications	Number
Long Human Families with a Prokaryotic Homolog	1033
Long Human Families with One Sequence Domain	330
Ancient Long One-Domain Families (Consistent Tree) (ALOD)	249
ALOD Families with GO Annotations	179
ALOD Families with Solved Structure Homolog (SSH)	164
ALOD Families with SSH Having Cath/Scop Classifications	132
ALOD Families with SSH Having One CATH/SCOP Domain	58

Department of Biochemistry and Molecular Genetics

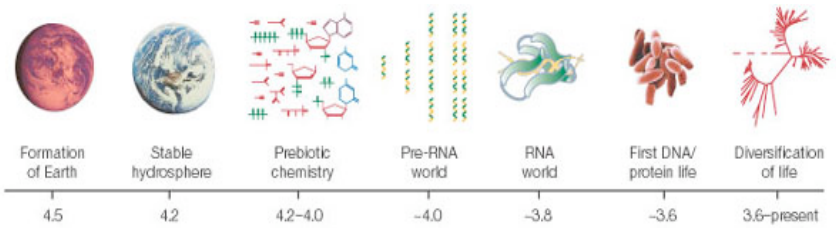
Ancient Domains, or Horizontal Transfer?



Ancient Domains have diverse functions

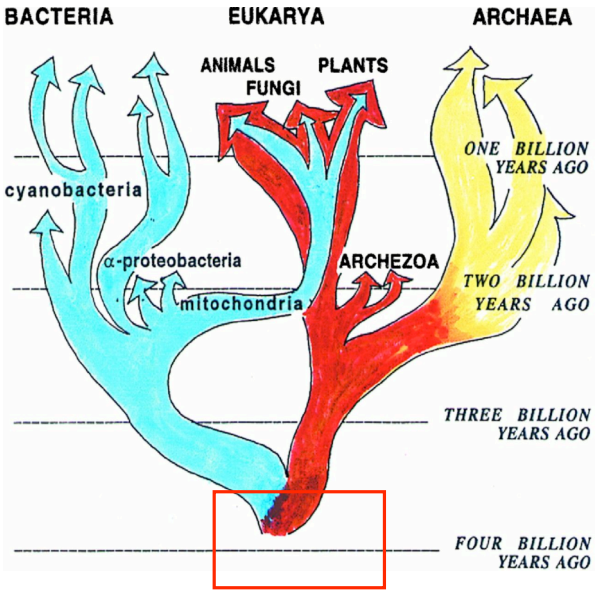
GO Accession	GO Term	Mean Random Proteins with GO Term	ALODs with GO Term	Poisson P()
GO.0005739	mitochondrion	16	32	8.4E-05
GO.0009058	biosynthesis	30	49	7.0E-04
GO.0009056	catabolism	16	28	1.6E-03
GO.0006091	energy pathways	8	16	1.9E-03
GO.0005975	carbohydrate metabolism	12	22	2.3E-03
GO.0006519	amino acid metabolism	7	14	4.2E-03
GO.0008152	metabolism	123	144	2.7E-02
GO.0006260	DNA replication	5	7	NS
GO.0006412	protein biosynthesis	13	16	NS
GO.0006139	nucleic acid metabolism	27	29	NS
GO.0006629	lipid metabolism	11	12	NS
GO.0006350	transcription	9	9	NS
GO.0006118	electron transport	11	11	NS
GO.0016020	membrane	48	40	NS
GO.0019538	protein metabolism	51	38	NS
GO.0007165	signal transduction	24	4	NS

Life evolved quickly



Joyce (2002) *Nature* **418**:214

Last Universal Common Ancestor (LUCA)
Most pathways were established by this time.



New Perspectives on Sequence and Structure Comparison

- Protein sequence comparison reliably identifies homologs (but not non-homologs) because *unrelated* sequences behave like *random* sequences – *Not random → homologous*
- Protein structure comparison can identify additional homologs, at the cost of *high false-positive* rates
- Structure comparison methods have less presumption that unrelated structures behave like random structures – *structure comparison statistics are very unreliable*
- Protein structures can be considered *random* – new structures with similar functions appear frequently, the distribution of protein words is random, and long domains emerged quickly
- The current protein universe samples a small fraction of possible protein structures

Department of Biochemistry and Molecular Genetics

Are Sequences that Fold Rare?

1. Does the number of Protein *Folds* indicate the number of Protein *Families* ?
   - Does structural similarity imply Homology? Yes/No Significant similarity implies homology, but much fold similarity may reflect analogy, or convergence.
   - Does the abundance of common folds suggest saturation of "fold-able" proteins? No – new protein families can emerge that share common folds
2. Is Protein folding so difficult that fold-able units must be small
   - Are all proteins build from small, fold-able, domains? No – A substantial fraction of ancient domains (25 - 30%) are long (200 - 500+ aa)

Department of Biochemistry and Molecular Genetics

Acknowledgements

Mike Sierk
Dan Lavelle
Mike Smoot
Aaron Gussman
Fitz Elliot
Brandi Cantarel
Anne Westbrook
Justin Reese

Department of Biochemistry and Molecular Genetics

www.cshl.org

COMPUTATIONAL GENOMICS

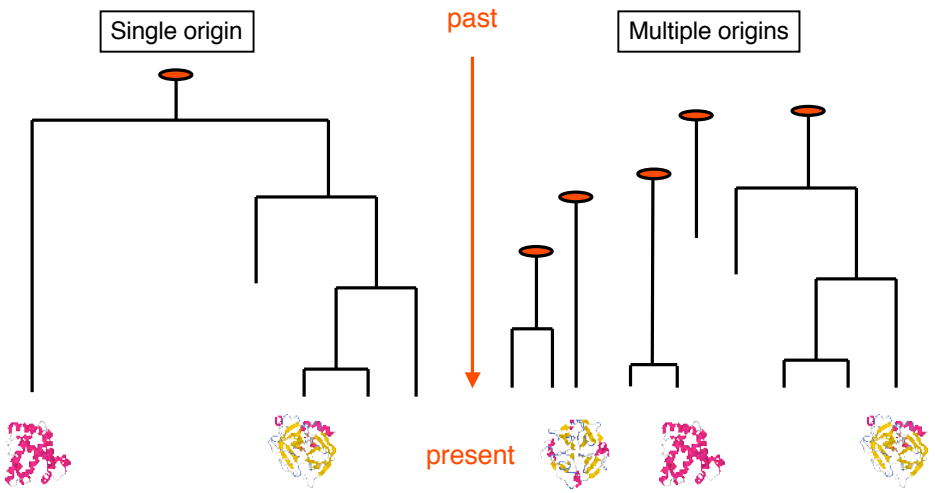
OCTOBER 31 – NOVEMBER 5, 2001
Application Deadline: July 15, 2001

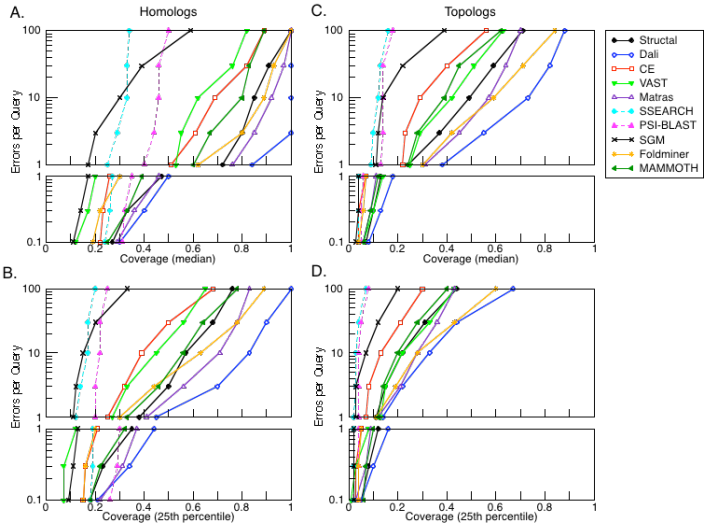
Instructors:
Pearson, William, Ph.D., *University of Virginia, Charlottesville, Va*
Smith, Randall, Ph.D., *SmithKline Beecham Pharmaceuticals, King of Prussia, PA*

Department of Biochemistry and Molecular Genetics

Are All Sequences Homologous?

No Homology without excess similarity





- When misclassification bias is removed, sequence comparison methods still perform well at low error rates.
- At modest error rates (1 - 2/ query), structure comparison methods perform very well, on average