Margaret Dayhoff

(1925-1983)

1945 - BA in Mathematics at NYU

1948 – PhD in Quantum Chemistry (Prof. George Kimball) from Columbia Univ. **"Punched Card Calculation of Resonance Energies"** *J. Chem. Phys*. 17,

1959 – National Biomedical Research Foundation (later part of Georgetown University)

Computational aids for protein sequence determination

Origins of Life

1965 Protein Atlas (65 proteins)

1980 - President of Biophysical Society

Science. 1966 Apr 15;152(3720):363-366.

Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences.

Eck RV, Dayhoff MO.

The structure of present-day ferredoxin, with its simple, inorganic active site and its functions basic to photon-energy utilization, suggests the incorporation of its prototype into metabolism very early during biochemical evolution, even before complex proteins and the complete modern genetic code existed. The information in the amino acid sequence of ferredoxin enables us to propose a detailed reconstruction of its evolutionary history. Ferredoxin has evolved by doubling a shorter protein, which may have contained only eight of the simplest amino acids. This shorter ancestor in turn developed from a repeating sequence of the amino acids alanine, aspartic acid or proline, serine, and glycine. We explain the persistence of living relics of this primordial structure by invoking a conservative principle in evolutionary biochemistry: The processes of natural selection severely inhibit any change a well-adapted system on which several other essential components depend.

1. A D S G

2. A D S G A D S G A D S G A D D S G A D S G A D S G A D S G

3. A D S D A D S C V D C G A C A S V C P V G A P S Q G D S G

4. A D S D A D S C V D C G A C A S V C P V G A P S Q G D S G A D S D A D S C V D C G A C A S V C P V G A P S Q G D S G

5. A Q K I A D S C V S C G A C A S E C P V N A I S Q G D S I F V I D A D T C I D C G N C A N V C P V G A P V Q E

Fig. 3. Proposed origin and evolution of ferredoxin (see text for fuller details). Row 1: Originally, in an extremely primitive organism, a short sequence of four of the simplest amino acids (alanine, aspartic acid, serine, and glycine) could be produced. Row 2: This sequence lengthened by doubling of the genetic material, and one discontinuity occurred (underlined). Row 3: The genetic code becoming more versatile, mutations (underlined) occurred, but only to relatively simple amino acids (the same four, plus cysteine, valine, proline, and glutamine). Iron sulfide was attached to the cysteines, which constituted the "active site" of the respiratory function of this primitive ferredoxin. This configuration still persists. Row 4: By "chromosome" aberration, the whole chain doubled. Row 5: The present more intricate genetic code having evolved, further mutations (underlined) to more complex amino acids occurred. The last three links were deleted. The result was the present sequence of ferredoxin from *C. pasteurianum* (*4*).

a radical change. We predict that when the three-dimensional structure of ferredoxin is worked out, evidence will be found for the previous stage, with its two identical, cooperating, shorter chains. The three end units may have

length. If so, we may expect to see evidences of duplication in other protein sequences, when ways of recognizing distant homologous relationships become more precise than the mere counting of the few identical amino acids remaining. The diheme peptide of *Chromatium* may possibly be such a case

Such ancient systems are extremely conservative, because so many diverse later reactions have become intricately dependent on them that they are no longer "free" to evolve. A mutational change which might be beneficial in one way, in almost every case would be a strong disadvantage in many other ways. When such a mutation occurred, the process of natural selection would therefore reject it. This conservative principle enables us to comprehend why ferredoxin from a living organism could still retain detectable details of its ancient origin.

Thus, in organisms still living there may exist biochemical relics of the era encompassing the origin and evolution of the genetic mechanism. Determina-

Science. 1966 Apr 15;152(3720):363-366.

Biochem Biophys Res Commun. 1970 May 22;39(4):757-65.

The occurrence in proteins of the tripeptides Asn-X-Ser and Asn-X-Thr and of bound carbohydrate.

Hunt LT, Dayhoff MO.

The 101 occurrences of the tripeptides Asn-X-Ser and Asn-X-Thr in the available protein sequence data are tabulated; carbohydrate is found, attached to the asparagine, in not more than 20 of the 101 tripeptides. A statistical analysis of the data from all completely sequenced proteins shows that the observed frequency of occurrence of the two kinds of tripeptides is only about 65% of the expected.

This lowered frequency is evidence for a newly postulated kind of limitation—which we call a "restricted sequence"—imposed by natural selection on the primary structure of proteins.

We suggest that the frequency of occurrence of the Asn-X-Ser/Thr tripeptides in the available protein sequences, which is considerably lower than expected, reflects a restriction by natural selection on the occurrence of the two tripeptides in proteins. Selection would reject a protein which acquired the tripeptide(s) by mutation, if carbohydrate, bound to the tripeptide by the enzyme, subsequently interfered with a normal interaction or function of the protein.

Many of the sequenced proteins were orthologs from different organisms

J Mol Evol. 1973;2(2-3):99-116.

Eukaryote evolution: a view based on cytochrome c sequence data.

McLaughlin PJ, Dayhoff MO.

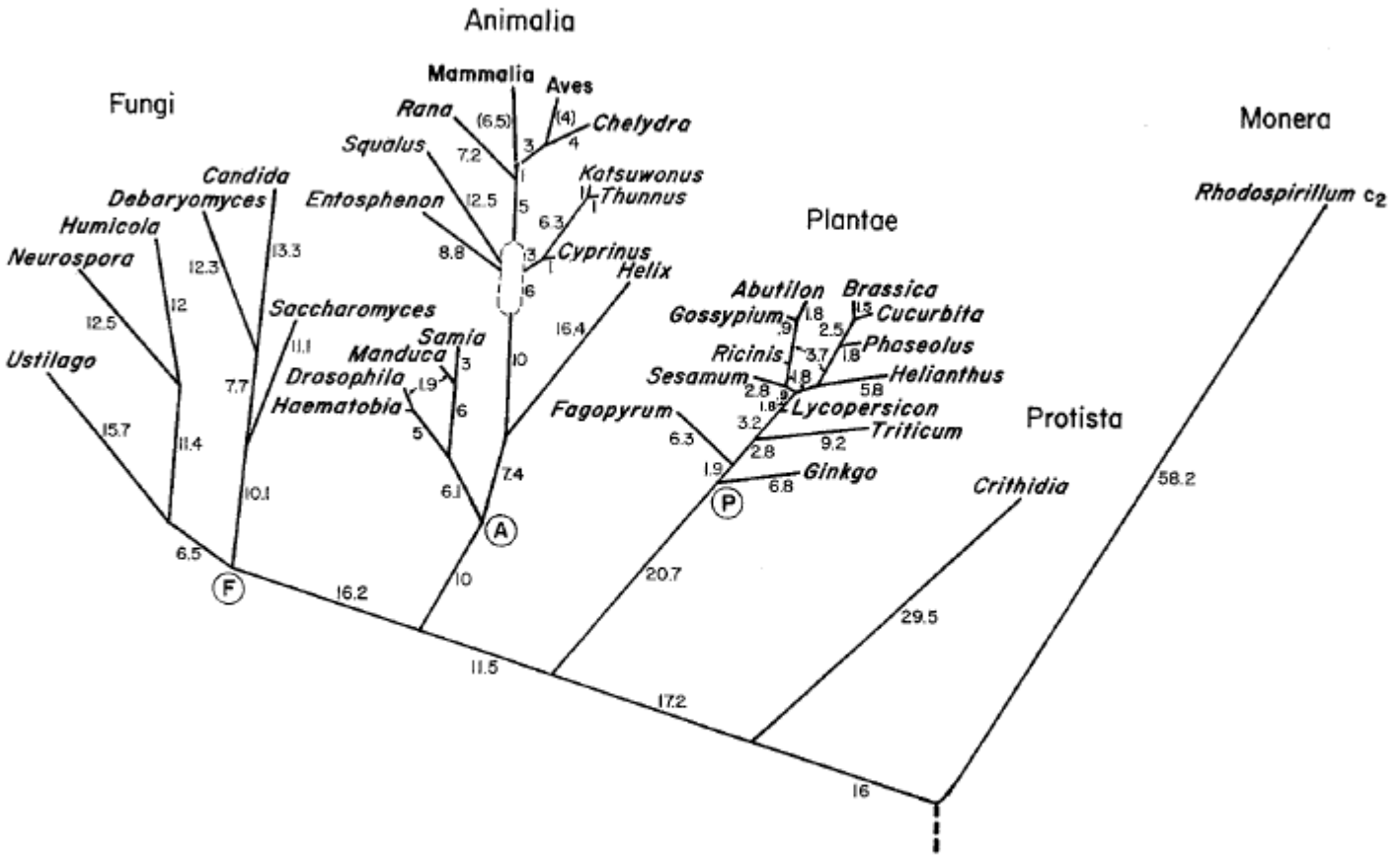


Fig. 2. The detailed cytochrome *c* evolutionary tree. The order of branching for the five kingdoms is the same as configuration 1 in Fig. 4. The progression of time is toward the top of the tree. The lengths of the branches are drawn in proportion to the numbers beside the branches, which are PAMs or Accepted Point Mutations estimated to have occurred on these branches

Searching the growing sequence database…

Biochem Biophys Res Commun. 1974 Oct 8;60(3):1020-8.
Epidermal growth factor: internal duplication and probable relationship to pancreatic secretory trypsin inhibitor.
Hunt LT, Barker WC, Dayhoff MO.

Biochem Biophys Res Commun. 1976 Apr 19;69(4):852-9.
Sequence similarity between cholera toxin and glycoprotein hormones: implications for structure activity relationship and mechanism of action.
Ledley FD, Mullin BR, Lee G, Aloj SM, Fishman PH, Hunt LT, Dayhoff MO, Kohn LD.

Biochem Biophys Res Commun. 1980 Jul 31;95(2):864-71.
A surprising new protein superfamily containing ovalbumin, antithrombin-III, and alpha 1-proteinase inhibitor.
Hunt LT, Dayhoff MO.

Rapid similarity searches of nucleic acid and protein data banks.

Wilbur WJ, Lipman DJ.

Proc Natl Acad Sci U S A 1983 Feb;80(3):726-30

With the development of large data banks of protein and nucleic acid sequences, the need for efficient methods of searching such banks for sequences similar to a given sequence has become evident. We present an algorithm for the global comparison of sequences based on matching k-tuples of sequence elements for a fixed k. The method results in substantial reduction in the time required to search a data bank when compared with prior techniques of similarity analysis, with minimal loss in sensitivity. The algorithm has also been adapted, in a separate implementation, to produce rigorous sequence alignments. Currently, using the DEC KL-10 system, we can compare all sequences in the entire Protein Data Bank of the National Biomedical Research Foundation with a 350-residue query sequence in less than 3 min and carry out a similar analysis with a 500-base query sequence against all eukaryotic sequences in the Los Alamos Nucleic Acid Data Base in less than 2 min.

Cancer Gene Meets Its Match

New York Times July 3, 1983

The New York Times

Waterfield MD et al., Nature 1983 Jul 7;304(5921):35-39
Doolittle RF et al., Science 1983 Jul 15;221(4607):275-277

```
v-sis: 6   QGDPIPEELYKMLSGHSIRSFDDLQRLLQGDSGKEDGAELDLNMTRSHSGGELESLARGK 65
           QGDPIPEELY+MLS HSIRSFDDLQRLL GD G+EDGAELDLNMTRSHSGGELESLARG+
PDGF : 10  QGDPIPEELYEMLSDHSIRSFDDLQRLLHGDPGEEDGAELDLNMTRSHSGGELESLARGR 69

v-sis: 66  RSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNVQ 125
           RSLGSL++AEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNVQ
PDGF : 70  RSLGSLTIAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNVQ 129

v-sis: 126 CRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCEIVAAARAVTRSPGTSQEQR 185
           CRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCE VAAAR VTRSPG SQEQR
PDGF : 130 CRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCETVAAARPVTRSPGGSQEQR 189

v-sis: 186 AKTTQSRVTIRTVRVRRPPKGKHRKCKHTHDKTALKETLGA 226
           AKT Q+RVTIRTVRVRRPPKGKHRK KHTHDKTALKETLGA
PDGF : 190 AKTPQTRVTIRTVRVRRPPKGKHRKFKHTHDKTALKETLGA 230
```

"Now a serendipitous computer search has matched it with the product of a gene that causes cell growth to run amok - a cancer gene found in a monkey virus. The discovery, which will be reported this month in the journals Science and Nature, may provide a key link in the chain of events that causes cancer."

V-sis and Platelet-Derived Growth Factor (PDGF)

An earlier, more subtle discovery…

Viral src gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase Barker WC, **Dayhoff MO**. PNAS 1982 May;79(9):2836-2839

```
Query: 113 YAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKR---VKGRTWT---LC 166
           Y+  +V    +LHS  +++ DLKP N+LI +Q   +++DFG +++   ++GR  +   +
Sbjct: 125 YSLDVVNGLLFLHSQSILHLDLKPANILISEQDVCKISDFGCSQKLQDLRGRQASPPHIG 184

Query: 167 GTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVR 223
           GT  + APEI+  +      D ++ G+ +++M     P ++ P  +   +V+  +R
Sbjct: 185 GTYTHQAPEILKGEIATPKADIYSFGITLWQMTTREVP-YSGEPQYVQYAVVAYNLR 240
```

Biology not Algorithms

- compare proteins, not DNA
- must detect similar amino acids not just identities

vary greatly in their mutability; 55% of the
52% of the cysteines and 27% of the glycines
unchanged, but only 6% of the highly muta-

From the series of distance-dependent mutation probability matrices, we can compute detailed answers to the
question "How does the evolutionary process affect

Comparison of Scoring Matrices

We have compared a number of scoring matrices using ALIGN. The results of these comparisons, involving a broad selection of pairs of related sequences, are listed in Table 24. MDM_{78} gives the highest average score, although it does not always give the highest score for a particular comparison. It is the only matrix that consistently detects relatedness (scores ≥ 3.0 SD) for the entire range of sequences tested. In the comparison of antibacterial substance A with neocarzinostatin, GCM and UM give better scores than MDM_{78}. This may be due to the conservation of what are usually more mutable amino acids. In the other comparisons, matrices based on mutation data perform better, and MDM_{78} usually gives the strongest indication of relatedness. The average scores are shown at the bottom of the table. The score using MDM_{78} is 1 SD better than that using AAAM, 2 SD better than that using GCM and almost 3 SD better than that using UM.

Table 25 shows segment comparison scores for a broad range of sequences including tests for internal duplications using different scoring matrices. Again, on occasion another matrix gives a better score, but only MDM_{78} consistently indicates known relationships between sequences; of the scoring matrices that we tested, it is clearly the best. The average score using it is 2.5 SD better than that for any of the other matrices.

In order to ascertain whether either ALIGN or RELATE produces false-positive results with any of the scoring matrices we tested, we examined 28 pairs of unrelated proteins. Neither program gave false-positive results with any of the matrices. The mean alignment score for the 28 comparisons was between 0.2 and −0.2 for all four matrices. The mean segment comparison score for the 28 pairs was between 0.3 and −0.4 for all four matrices. All of these trials were based on 100 randomized sequence comparisons.

Comparison of MDM_{78} with Its Predecessors

Using a variety of distantly related sequences, we have compared the results using the recently derived MDM_{78}, the two previous mutation data matrices, MDM_{67} and MDM_{69}, based on one-fourth and one-half as much data, respectively, and components of MDM_{78}: the diagonal elements alone, with all off-diagonal elements equal to zero, and the off-diagonal elements, with the diagonal

Table 24
Comparison of Matrices for Calculating Alignment Scores

Sequences Compared	Score (in SD units) Obtained with UM	GCM	AAAM	MDM_{78}
Antibacterial substance A — *Streptomyces* vs. Neocarzinostatin — *Streptomyces*	3.1	3.2	2.6	2.9
Ferredoxin — *Clostridium pasteurianum* vs. Ferredoxin — *Spirulina maxima*	0.1	1.6	1.8	3.4
Hemoglobin alpha — Human vs. Myoglobin — Human	5.8	6.6	9.9	10.7
Hemoglobin alpha — Human vs. Globin CTT-III — Midge larva	2.0	2.4	3.2	3.5
Cytochrome c — Horse vs. Cytochrome c_6 — *Spirulina*	4.5	4.3	7.3	6.1
Cytochrome c — Horse vs. Cytochrome c_{553} — *Desulfovibrio*	0.2	0.4	0.4	3.9
Beta$_2$-microglobulin — Human vs. Ig mu chain C4 homology region — Human Gal	3.6	3.3	4.7	4.8
Ig mu chain C4 homology region — Human Gal vs. Ig epsilon chain C4 homology region — Human Nd	4.7	9.0	9.2	12.1
Average score	3.0	3.9	4.9	5.9

In these comparisons, we used values for the gap penalty (P) and the matrix bias (B) that have been useful for a broad selection of sequence comparisons in our experience, typically 60 and 60 for MDM_{78} (Figure 85), 1 and 1 for GCM, and 0.3 and 0.3 for UM. In the comparison of antibacterial substance A with neocarzinostatin, a bias of 20 and a penalty of 80 were used with MDM_{78} because these are more typical choices for detecting very distant sequence relationships. In the comparisons using AAAM, for which our experience is limited, we varied B from −2 to +4; P was chosen to be 6 and 8. These values produced alignments that were similar in numbers of gaps and gap length to alignments using the other scoring matrices; they produced scores that were statistically indistinguishable from one another. For the above values, we used P = 6 and B = −2. Three hundred randomized sequence comparisons were used in determining scores for AAAM and MDM_{78}; thus, the estimated percent standard deviations of these scores are 4%. UM and GCM scores were calculated using 100 randomized sequence comparisons; thus, the estimated percent standard deviations of these scores are 7%.

Table 23
Correspondence between Observed Differences

Amino acid pairs with scores above 1 replace each other more often as alternatives in related sequences than

Table 25
Comparison of Matrices for Calculating Segment Comparison Scores

Sequences Compared	Score (in SD units) Obtained with UM	GCM	AAAM	MDM_{78}
Cytochrome c_6 — *Monochrysis* vs. Cytochrome c_2 — *Rhodospirillum*	4.7	3.1	2.5	3.5
Azurin — *Bordetella* vs. Plastocyanin — French bean	1.6	2.8	3.1	4.1
Ferredoxin — *Clostridium pasteurianum* vs. Ferredoxin — *Desulfovibrio*	3.9	3.1	4.3	6.0
Troponin C — Rabbit vs. Parvalbumin — Pike	7.6	8.3	8.0	10.2
Troponin C — Rabbit vs. Myosin A1 light chain — Rabbit	8.0	9.3	6.7	15.1
Internal Duplication				
Tropomyosin alpha chain — Rabbit	5.9	4.0	3.6	8.3
Protease inhibitor, submandibular gland — Dog	4.1	3.6	5.3	7.9
Cytochrome c_3 — *Desulfovibrio gigas*	0.5	1.3	0.7	3.9
Ferredoxin — *C. pasteurianum*	7.8	5.9	7.1	7.7
Average score	4.9	4.6	4.6	7.4

In the cytochrome c_3-c_3 and in the ferredoxin internal duplication comparisons, a segment length of 15 residues was used; in the other comparisons, we used a segment length of 20 residues. Three hundred randomized sequence comparisons were used in calculating scores for AAAM and MDM_{78}; thus, the percent standard deviations for these scores are 4%. One hundred comparisons were used for UM and GCM; thus, their percent standard deviations are 7%.

Table 26
Comparison of Mutation Data Matrices for Calculating Alignment Scores

Sequences Compared	Scores (in SD units) Obtained with MDM_{67}	MDM_{69}	MDM_{78}	UM	Diagonal Only MDM_{78}	Off-diagonal and Averaged Diagonal MDM_{78}
Antibacterial substance A — *Streptomyces* vs. Neocarzinostatin — *Streptomyces*	2.0	2.4	2.9	3.1	1.4	1.8
Ferredoxin — *Clostridium pasteurianum* vs. Ferredoxin — *Spirulina maxima*	2.6	2.6	3.4	0.1	2.7	2.7
Hemoglobin alpha — Human vs. Myoglobin — Human	9.9	9.7	10.7	5.8	9.9	10.3
Hemoglobin alpha — Human vs. Globin CTT-III — Midge larva	2.6	2.4	3.5	2.0	0.9	3.5
Cytochrome c — Horse vs. Cytochrome c_6 — *Spirulina*	5.6	5.4	6.1	4.5	5.6	5.8
Cytochrome c — Horse vs. Cytochrome c_{553} — *Desulfovibrio*	3.8	3.9	3.9	0.2	2.0	2.8
Beta$_2$-microglobulin — Human vs. Ig mu chain C4 homology region — Human Gal	3.3	2.8	4.8	3.6	3.9	4.8
Ig mu chain C4 homology region — Human Gal vs. Ig epsilon chain C4 homology region — Human Nd	10.1	11.5	12.1	4.7	11.2	11.9
Average score	5.0	5.1	5.9	3.0	4.7	5.5

In the comparison of antibacterial substance A with neocarzinostatin, a matrix bias of 20 and a gap penalty of 80 were used with MDM_{78} and its derivatives. In the other comparisons with MDM_{78}, a bias of 60 and a penalty of 60 were used. A penalty of 6 and a bias of 6 were used with MDM_{67} and MDM_{69} because their elements are expressed to one significant figure less than MDM_{78}. Three hundred random comparisons were used in determining scores for all matrices except UM, for which 100 random comparisons were made. A penalty of 0.3 and bias of 0.3 were used with UM.

Science. 1985 Mar 22;227(4693):1435-41.

Rapid and sensitive protein similarity searches.

Lipman DJ, Pearson WR.

An algorithm was developed which facilitates the search for similarities between newly determined amino acid sequences and sequences already available in databases. Because of the algorithm's efficiency on many microcomputers, sensitive protein database searches may now become a routine procedure for molecular biologists. The method efficiently identifies regions of similar sequence and then scores the aligned identical and differing residues in those regions by means of an amino acid replaceability matrix. This matrix increases sensitivity by giving high scores to those amino acid replacements which occur frequently in evolution. The algorithm has been implemented in a computer program designed to search protein databases very rapidly. For example, comparison of a 200-amino-acid sequence to the 500,000 residues in the National Biomedical Research Foundation library would take less than 2 minutes on a minicomputer, and less than 10 minutes on a microcomputer (IBM PC).

Use Wilbur-Lipman for initial guesses, then rescore using Dayhoff Matrix

J Mol Biol. 1990 Oct 5;215(3):403-10.

Basic local alignment search tool.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

Maximal Segment Pair was local ungapped optimal alignment – using Dayhoff Matrix

Karlin's statistics for MSP's allowed direct use of Dayhoff matrix and now could assess odds for the guessing…

Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. altschul@ncbi.nlm.nih.gov

The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. For protein comparisons, a variety of definitional, algorithmic and statistical refinements described here permits the execution time of the BLAST programs to be decreased substantially while enhancing their sensitivity to weak similarities. A new criterion for triggering the extension of word hits, combined with a new heuristic for generating gapped alignments, yields a gapped BLAST program that runs at approximately three times the speed of the original. In addition, a method is introduced for automatically combining statistically significant alignments produced by BLAST into a position-specific score matrix, and searching the database using this matrix. The resulting Position-Specific Iterated BLAST (PSI-BLAST) program runs at approximately the same speed per iteration as gapped BLAST, but in many cases is much more sensitive to weak but biologically relevant sequence similarities. PSI-BLAST is used to uncover several new and interesting members of the BRCT superfamily.

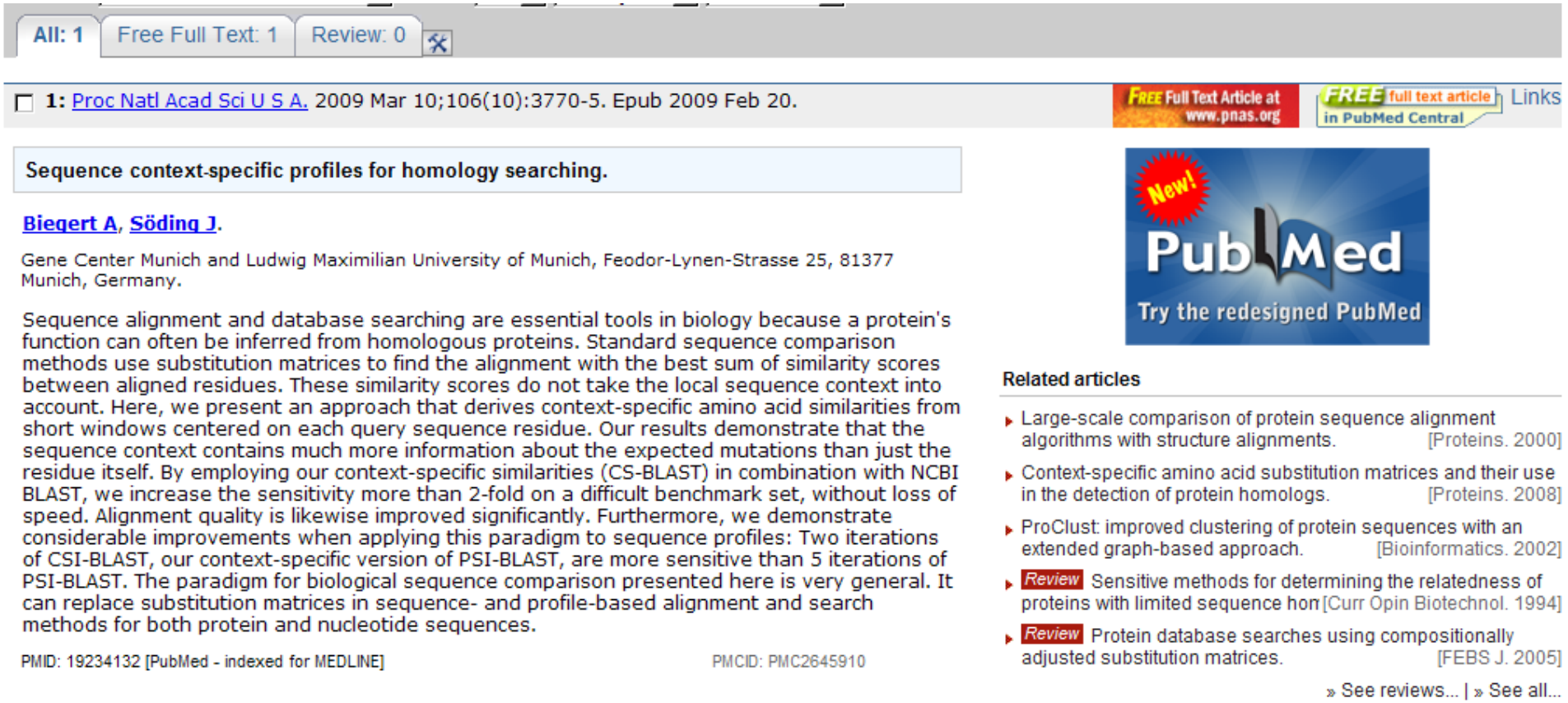
Everyone was publishing papers about methods more sensitive but significantly slower than BLAST – why not just search the database multiple times using hits to improve model?

Generate a protein family-specific & position-specific similarity matrix…

	ref|YP_003256739.1|	hypothetical protein D11S_2169 [Aggregati...	158	4e-37	G
	ref|YP_164090.1|	hypothetical protein B3ORF53 [Pseudomonas ph...	153	1e-35	G
NEW	ref|YP_001294543.1|	hypothetical protein ORF035 [Pseudomonas ...	151	5e-35	G
	ref|YP_419838.1|	hypothetical protein amb0475 [Magnetospirill...	147	1e-33	G
	ref|NP_918979.1|	conserved tail assembly protein [Burkholderi...	143	2e-32	G
	ref|YP_950469.1|	hypothetical protein DMS3-45 [Pseudomonas ph...	137	8e-31	G
	ref|YP_002332471.1|	hypothetical protein PPMP29_gp46 [Pseudom...	137	1e-30	G
	ref|YP_002491724.1|	putative phage associated protein [Anaero...	137	1e-30	G
	ref|NP_872756.1|	hypothetical protein HD0150 [Haemophilus duc...	136	2e-30	G
	ref|YP_001469175.1|	hypothetical protein PMV22_orf50 [Phage M...	135	3e-30	G
	ref|NP_938257.1|	hypothetical protein D3112p50 [Pseudomonas p...	135	3e-30	G
	ref|YP_002332358.1|	hypothetical protein PPMP38_gp46 [Pseudom...	135	4e-30	G
	ref|YP_002439195.1|	hypothetical protein PLES_15911 [Pseudomo...	135	4e-30	G
	ref|ZP_05112773.1|	hypothetical protein SADFL11_658 [Labrenzi...	131	6e-29	
NEW	ref|YP_421142.1|	hypothetical protein amb1779 [Magnetospirill...	119	3e-25	G
NEW	ref|YP_747669.1|	hypothetical protein Neut_1458 [Nitrosomonas...	118	5e-25	G
	ref|YP_001405889.1|	hypothetical protein CHAB381_0286 [Campyl...	117	8e-25	G
	ref|ZP_05134331.1|	conserved hypothetical protein [Stenotroph...	117	9e-25	
	ref|ZP_03543676.1|	conserved hypothetical protein [Comamonas ...	117	9e-25	
NEW	ref|YP_436735.1|	hypothetical protein HCH_05652 [Hahella chej...	114	9e-24	G
NEW	ref|YP_966527.1|	hypothetical protein Dvul_1080 [Desulfovibri...	113	2e-23	G
	ref|YP_002521027.1|	hypothetical protein RSKD131_4094 [Rhodob...	113	2e-23	G
NEW	ref|ZP_05845046.1|	conserved hypothetical protein [Rhodobacte...	112	4e-23	
NEW	ref|YP_001293431.1|	hypothetical protein ORF024 [Pseudomonas ...	101	5e-20	G
NEW	ref|ZP_03724513.1|	hypothetical protein ObacDRAFT_9012 [Opitu...	93.5	2e-17	
	ref|YP_167938.1|	hypothetical protein SPO2730 [Ruegeria pomer...	93.5	2e-17	G
NEW	ref|YP_001971684.1|	putative phage tail assembly protein [Ste...	93.1	3e-17	G
NEW	ref|ZP_05403239.2|	phage conserved hypothetical protein [Mits...	91.2	1e-16	
NEW	ref|YP_001354472.1|	hypothetical protein mma_2782 [Janthinoba...	88.1	8e-16	G
NEW	ref|NP_873087.1|	hypothetical protein HD0534 [Haemophilus duc...	87.3	1e-15	G
NEW	ref|YP_001354412.1|	hypothetical protein mma_2722 [Janthinoba...	86.2	3e-15	G
NEW	ref|ZP_06066190.1|	predicted protein [Acinetobacter junii SH205]	86.2	3e-15	
NEW	ref|ZP_05715341.1|	hypothetical protein VMD_03870 [Vibrio mim...	84.3	1e-14	
	ref|YP_246329.1|	hypothetical protein RF_0313 [Rickettsia fel...	81.6	8e-14	G

Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.

Improving sequence similarity searching…



All: 1 | Free Full Text: 1 | Review: 0

1: Proc Natl Acad Sci U S A. 2009 Mar 10;106(10):3770-5. Epub 2009 Feb 20.

FREE Full Text Article at www.pnas.org | FREE full text article in PubMed Central | Links

Sequence context-specific profiles for homology searching.

Biegert A, Söding J.

Gene Center Munich and Ludwig Maximilian University of Munich, Feodor-Lynen-Strasse 25, 81377 Munich, Germany.

Sequence alignment and database searching are essential tools in biology because a protein's function can often be inferred from homologous proteins. Standard sequence comparison methods use substitution matrices to find the alignment with the best sum of similarity scores between aligned residues. These similarity scores do not take the local sequence context into account. Here, we present an approach that derives context-specific amino acid similarities from short windows centered on each query sequence residue. Our results demonstrate that the sequence context contains much more information about the expected mutations than just the residue itself. By employing our context-specific similarities (CS-BLAST) in combination with NCBI BLAST, we increase the sensitivity more than 2-fold on a difficult benchmark set, without loss of speed. Alignment quality is likewise improved significantly. Furthermore, we demonstrate considerable improvements when applying this paradigm to sequence profiles: Two iterations of CSI-BLAST, our context-specific version of PSI-BLAST, are more sensitive than 5 iterations of PSI-BLAST. The paradigm for biological sequence comparison presented here is very general. It can replace substitution matrices in sequence- and profile-based alignment and search methods for both protein and nucleotide sequences.

PMID: 19234132 [PubMed - indexed for MEDLINE] PMCID: PMC2645910

New! PubMed Try the redesigned PubMed

Related articles

- Large-scale comparison of protein sequence alignment algorithms with structure alignments. [Proteins. 2000]
- Context-specific amino acid substitution matrices and their use in the detection of protein homologs. [Proteins. 2008]
- ProClust: improved clustering of protein sequences with an extended graph-based approach. [Bioinformatics. 2002]
- Review Sensitive methods for determining the relatedness of proteins with limited sequence hon [Curr Opin Biotechnol. 1994]
- Review Protein database searches using compositionally adjusted substitution matrices. [FEBS J. 2005]

» See reviews... | » See all...

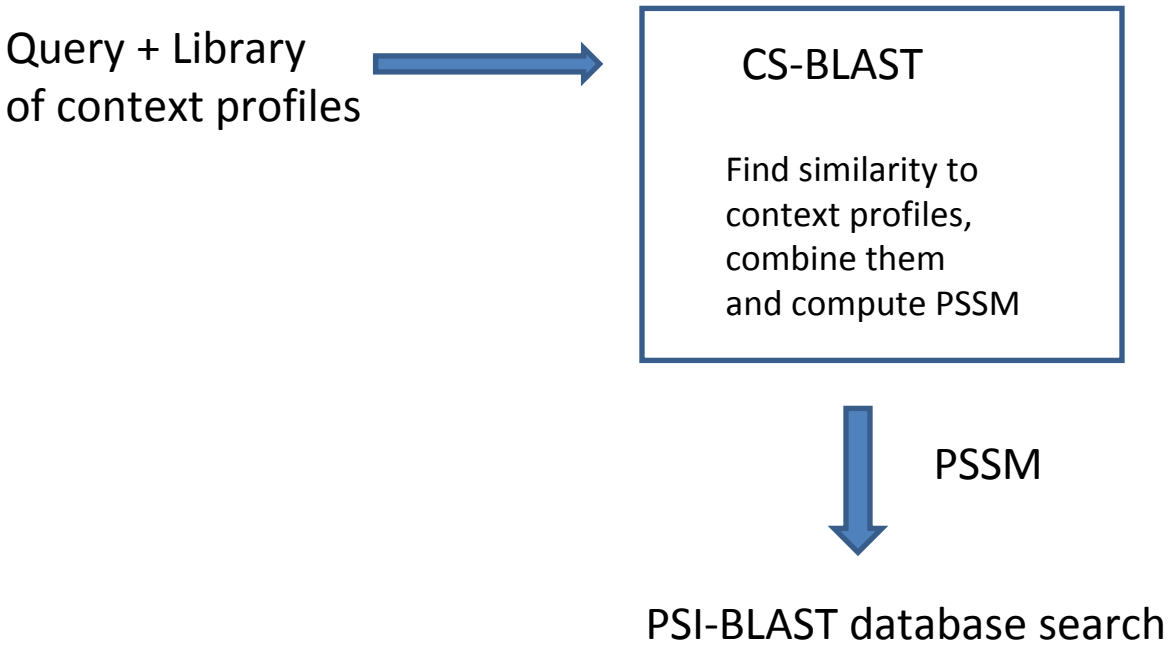
Why is it better?

- general structural information

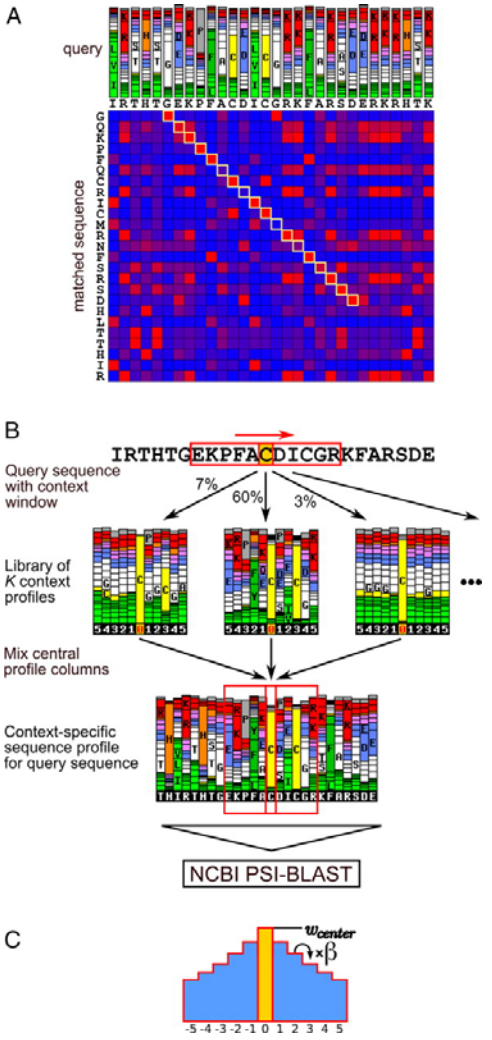
OR

- protein family information

CS-BLAST

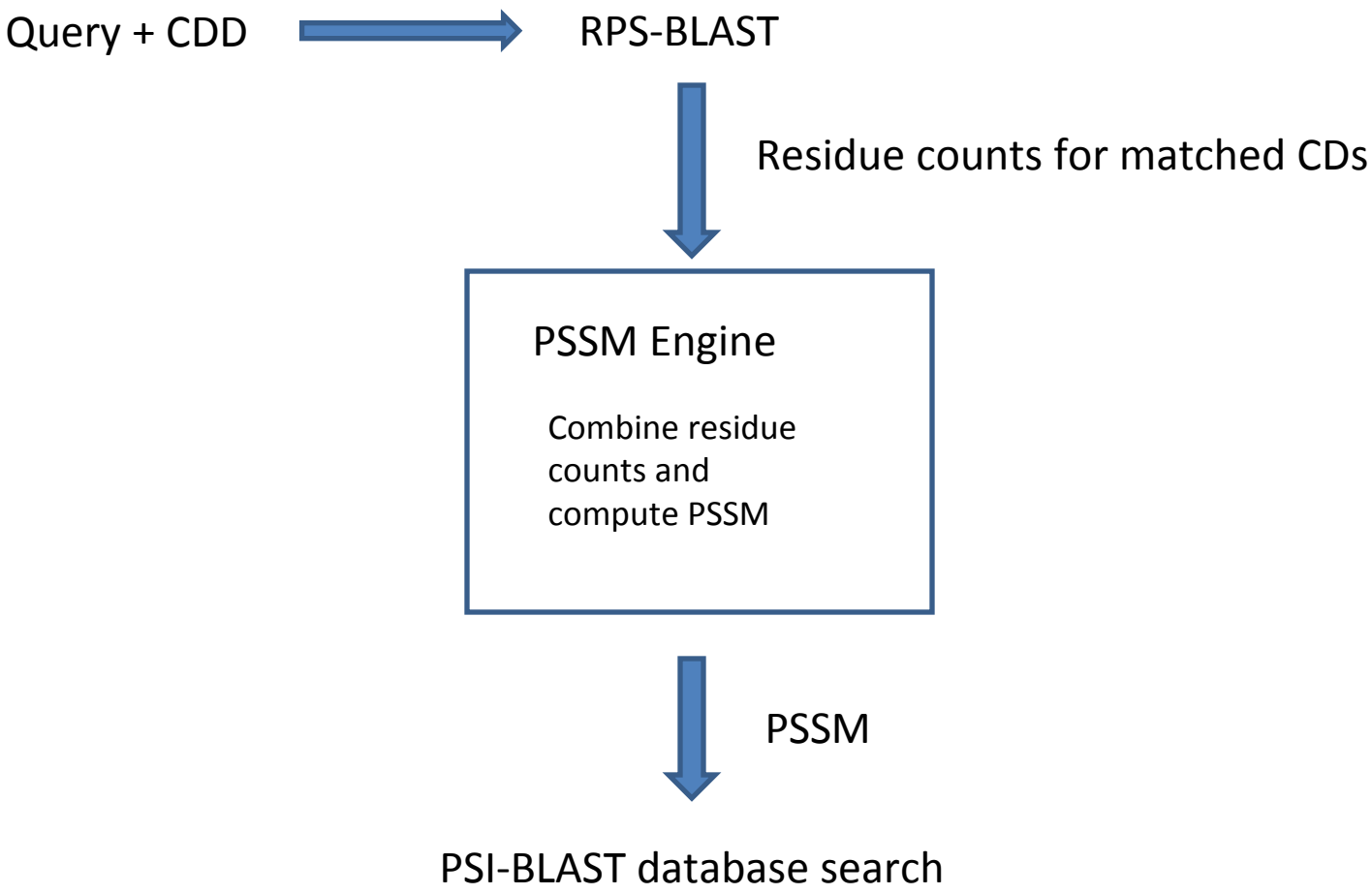

Query + Library
of context profiles
CS-BLAST
Find similarity to
context profiles,
combine them
and compute PSSM
PSSM
PSI-BLAST database search

Method of context-specific sequence comparison



Biegert A, Söding J PNAS 2009;106:3770-3775

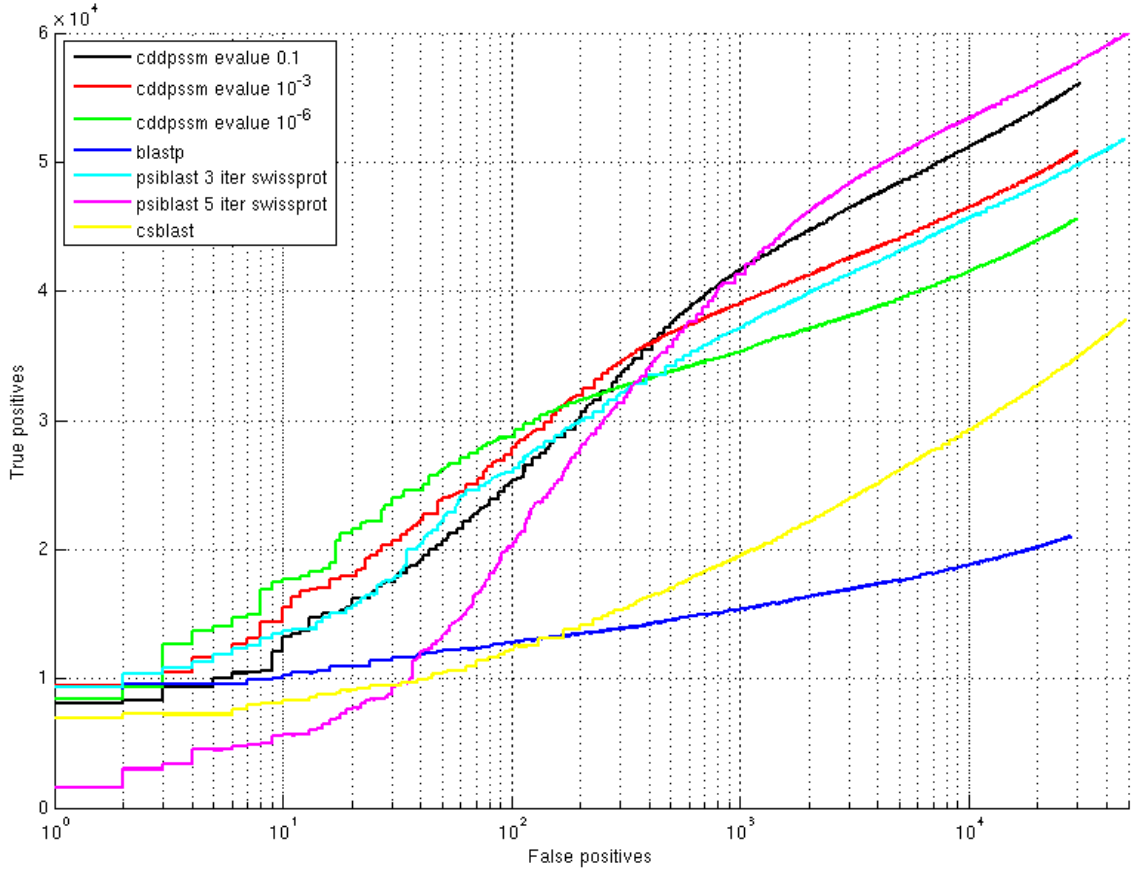
CDD-PSSM


Query + CDD
RPS-BLAST
Residue counts for matched CDs
PSSM Engine
Combine residue counts and compute PSSM
PSSM
PSI-BLAST database search

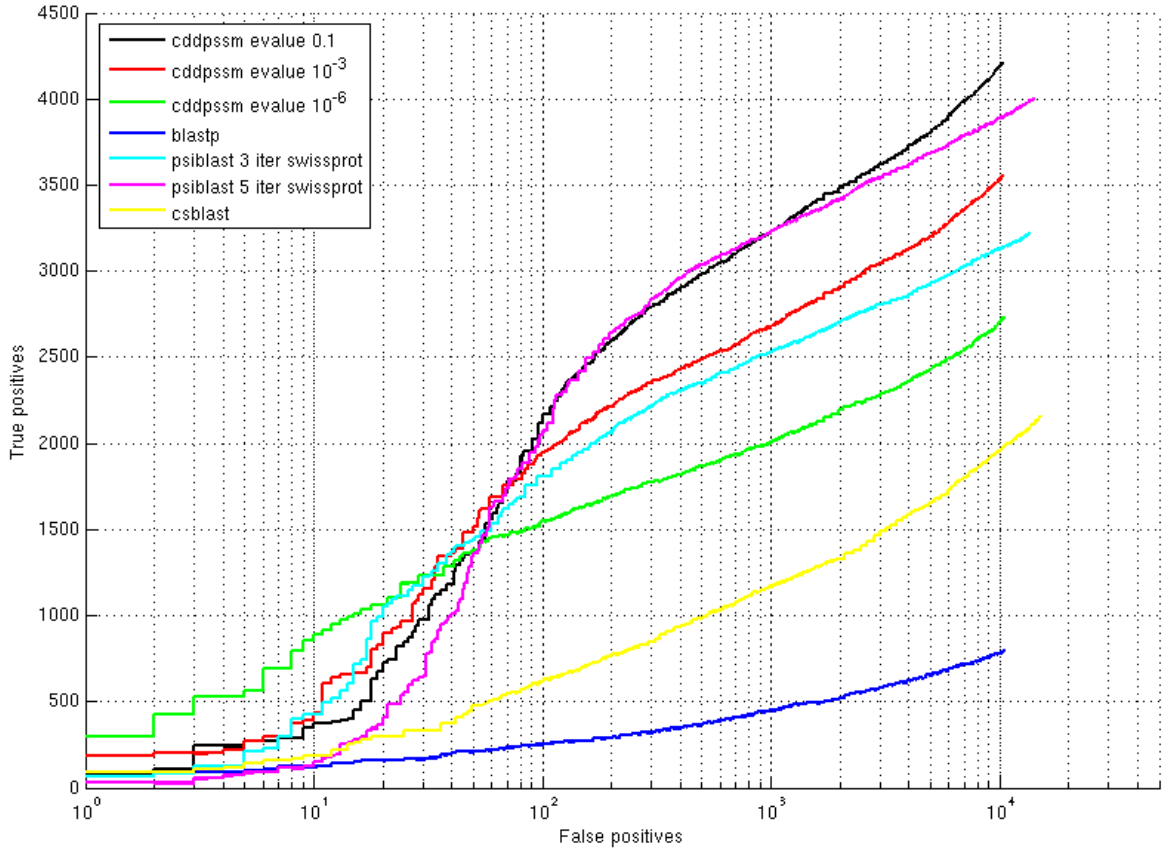
Experiments

- Hits that belong to the same superfamily as query are considered true positives
- Hits with the same fold as query but different superfamily are ignored
- All other hits are considered false positives

True vs. false positives for SCOP/ASTRAL 1.75 (9705 queries)



True vs. false positives for queries from SCOP families of size 1 in CS-BLAST benchmark (1874 queries)



How often would one find matches?

How many protein families would there be?

Prior to the genome project, there was only a small percentage of genes from the genomes of a number of evolutionarily distant organisms (e.g. human, fly, yeast, e.coli).

Unexpected similarities should be extremely rare.

Hubris, the Genome Project, and Protein Families

Chothia, C. (1992). **One thousand families for the molecular biologist**. Nature, 357, 543-544.

Green P, Lipman D, Hillier L, Waterson R, States,D, and Claverie JM (1993). **Ancient Conserved Regions in New Gene Sequences and the Protein Databases.** Science, 259, 1711-1716.

ACR = similarity detected between sequences from distantly related organisms

Ancient Conserved Regions in New Gene Sequences and the Protein Databases

Philip Green,* David Lipman, LaDeana Hillier, Robert Waterston, David States, Jean-Michel Claverie

Sets of new gene sequences from human, nematode, and yeast were compared with each other and with a set of *Escherichia coli* genes in order to detect ancient evolutionarily conserved regions (ACRs) in the encoded proteins. Nearly all of the ACRs so identified were found to be homologous to sequences in the protein databases. This suggests that currently known proteins may already include representatives of most ACRs and that new sequences not similar to any database sequence are unlikely to contain ACRs. Preliminary analyses indicate that moderately expressed genes may be more likely to contain ACRs than rarely expressed genes. It is estimated that there are fewer than 900 ACRs in all.

Understanding the functions and structures of the array of proteins expressed in living organisms is a fundamental goal of molecular biology. Our hope of attaining this goal stems largely from the unifying theme of shared evolutionary ancestry: related organisms have similar proteins and, within an organism, different proteins of related function are often wholly or partly similar in sequence, reflecting gene duplication and exon shuffling (*1*) during evolution. Such similarities can provide important functional insights, and consequently an important step in characterizing any newly sequenced gene is to compare its encoded protein sequence with the protein sequence databases in order to look for conserved regions shared with known proteins.

The present study uses extensive new sets of gene sequences to address several general questions about conserved regions: how many of these regions exist, what fraction has been discovered, and what proportion and types of proteins contain them. We focus on ancient conserved regions, or ACRs, detected through similarities between proteins from distantly related organisms. Over long evolutionary periods the less constrained portions of the sequences will have significantly diverged; consequently, the regions of similarity are usually those of greatest structural or functional significance. ACRs often correspond to specific domains (or motifs) present in a variety of proteins, such as zinc-finger DNA binding domains (*2*), or to enzyme active sites, but they can also comprise most or all of the sequence of a single highly conserved protein or protein family, such as actins and histones. Conserved regions of all of these types have been extensively cataloged (*3*, *4*). Because the degree of similarity between two related proteins reflects not only the amount of time since their last common ancestor but also their rates of sequence evolution, which can vary greatly for different proteins (*5*), not all proteins need contain ACRs.

The precise definition of an ACR depends on its required age and distribution among organisms and on the method used to detect sequence similarities. The present study involves ACRs that antedate the radiation of the major animal phyla [some 580 to 540 million years ago (*6*)] and that are present in diverse eukaryotes. We detected similarities by using the sequence alignment program BLAST (*7*) with a score cutoff sufficiently high to distinguish confidently true homologies from background in database searches (*8*). Figure 1 shows a representative BLAST alignment at this score level. Typically, a BLAST comparison of two related proteins reveals several (gap-free) aligned segments, separated by unaligned regions; in such cases we considered the entire collection of aligned segments to constitute a single conserved region, provided the segments always tended

P. Green, L. Hillier, and R. Waterston are in the Genetics Department, Washington University Medical School, St. Louis, MO 63110. D. Lipman, D. States, and J.-M. Claverie are at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

*To whom correspondence should be addressed.

Lots of new sequence data – *how many conserved protein families do we find that are not already in the databases?*

Sets compared	Matching Sequences	ACRs	ACRs in database
worm ESTs, **human** ESTs	77, 66	34	31 (91%)
worm ESTs, **yeast** ORFs	23, 13	9	8 (89%)
worm genes, **human** ESTs	17, 17	12	12 (100%)
worm genes, **yeast** ORFs	6, 4	4	3 (75%)
human ESTs, **yeast** ORFs	14, 13	10	10 (100%)

~1000 different ACR's

and multiply represented *C. elegans* ESTs. [Doubly triply represented ESTs at least two others, by the ed into those with database ACRs and those without and within each subgroup matches with the other described [(*8*) and Table 1].

EST representation		
Single	Double	Triple
310 (675)	144 (154)	116 (73)
34 (1)	23 (2)	17 (0)
15 (1)	5 (0)	2 (0)
67 (1)	46 (0)	22 (0)
92 (41)	59 (13)	46 (24)
259 (494)	115 (105)	92 (54)

likely considerable variation among ACRs, with some represented only once and others represented many times; a more detailed picture will emerge as the sequencing projects progress. It will also be of interest to learn what proportion of the ACRs are specific to metazoans.

Expression Level and Degree of Conservation

To better understand the impact of expression level bias in the EST sets, we looked for a possible relation between expression level and ACR presence. Because detailed expression data on these clones are not yet available, we assumed that to a first approximation genes represented in multiple independent clones in the cDNA libraries are, on average, expressed at higher levels than singly represented genes. Analyses were confined to the C. *elegans* ESTs (*29*), which were classified as singly represented (not overlapping any other EST) or multiply represented (overlapping at least one other EST). We found (Table 4) that database ACRs are present in a substantially higher fraction of the multiply represented ESTs (260/487, or 53%) than of the singly represented ESTs (310/985, or 31%). A similar trend holds for the C. *elegans* ACRs detected by similarity to the other sequence sets (*30*). Moreover, multiply represented ESTs have generally higher similarity scores with their distant homologs in the database than do singly represented ESTs (Fig. 2). The higher proportion of ACRs among multiply represented ESTs thus appears to be at least in part a consequence of their generally stronger similarities with distantly related genes and cannot simply be explained by a bias in the database itself toward moderately to highly expressed genes (*31*).

These results suggest that moderately expressed proteins have, on average, been more highly conserved in sequence over long evolutionary periods than have rarely expressed ones and in particular are more likely to contain ACRs. This is presumably

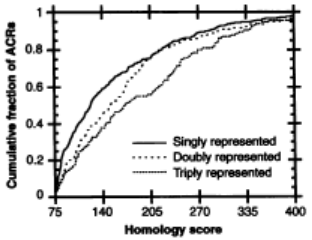


Fig. 2. Distribution of homology scores for database ACRs in singly and multiply represented *C. elegans* ESTs. For each EST having a cross-phylum match against SWISS-PROT, the average score of all such matches was taken to indicate the degree of conservation of the corresponding ACR. The cumulative fraction of ACRs having average scores less than a given value is plotted. Relatively more of the multiply represented ESTs have average scores exceeding any given value.

attributable in part to higher selective pressures to optimize the activities and structures of these proteins and to minimize undesired interactions with other cellular components. Given the indirectness of our method of assessing expression level, more detailed expression data on these clones will be required to confirm and accurately quantify this correlation.

Sequences Without ACRs

An early finding of the genome sequencing projects was that the majority of genes are not similar to anything in the databases (*11*, *12*). It has usually been assumed that this reflects the relative incompleteness of the databases rather than the absence of highly conserved regions in these genes. This assumption now appears incorrect. Because 30% or fewer of the genes in the genomic sets we analyzed contain database ACRs, and perhaps 85% of ACRs are present in the databases, the fraction of genes that contain ACRs is roughly 40% (0.30/0.85) or less. The other 60%—or over 90% of those sequences that are not currently similar to a distantly related sequence in the databases—do not have ACRs and must therefore correspond to proteins or protein regions that either evolved more recently than the metazoan radiation or evolved prior to it but have not been strongly conserved (*5*). In either case, they are unlikely to have strong similarities to any genes from distantly related organisms. For these sequences, homologies will be detectable only with the use of more sensitive methods of analysis or by comparisons with genes from more closely related organisms.

Many of these genes may have ancient functions despite their lack of sequence conservation. It is unlikely that the sequence requirements for a minimally active protein of any given function could be particularly stringent; otherwise, given the improbability of a specific sequence of any significant length arising solely by chance mutation, an appropriate substrate for selection to begin acting upon would never have arisen. Although optimization of activity can entail much more stringent sequence requirements, such optimization may only have been strongly selected for in a minority of the proteins in an organism. Thus, the majority of protein sequences may be relatively unconstrained and as a result may be drifting too rapidly to retain detectable similarities over long evolutionary periods. For this reason, one should not assume that ACRs necessarily represent all of the ancestral functional domains. Nor do they correspond to the universe of ancestral exons (*32*) because the majority of exons do not appear to be highly conserved. In fact, the differential rate of evolution of different protein regions considerably complicates the task of estimating the ancestral exon number.

In summary, it appears that the number of ACRs is relatively small—far smaller than the number of genes in a eukaryotic genome—and that most ACRs are represented among currently known proteins. We would emphasize, however, that more sequence data will be required to improve our understanding of conserved protein regions. The estimates above suggest that roughly one-third of ACRs have not yet been discovered because they are represented in only one phylum (or not at all) in the current databases. Detection of less highly conserved ACRs may only be possible when they are represented in multiple distantly related sequences. Finally, to increase our understanding of sequences that lack ACRs, it will be important to acquire sequence information from closely related organisms.

REFERENCES AND NOTES

1. W. A. Gilbert, *Science* 228, 823 (1985).
2. J. M. Berg, *ibid.* 232, 485 (1986).
3. A. Bairoch, *Nucleic Acids Res.* 20, 2013 (1992).
4. S. Henikoff and J. G. Henikoff, *ibid.* 19, 6565 (1991).
5. R. F. Doolittle, *Protein Sci.* 1, 191 (1992).
6. A. H. Knoll, *Science* 256, 622 (1992).
7. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. Lipman, *J. Mol. Biol.* 215, 403 (1990). BLAST has been shown [W. R. Pearson, *Genomics* 11, 635 (1991); S. F. Altschul *et al.*, *ibid.*, p. 408] to have comparable sensitivity to two other commonly used sequence alignment methods, FASTA [W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* 85, 2444 (1988)] and the Smith-Waterman dynamic programming method (*23*). In the present study, all BLAST comparisons were done with the use of conceptual translations of the DNA sequences.
8. Matches with scores of at least 75 obtained with the PAM120 matrix [M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC, 1979), vol.

ACR's more likely for genes with higher expression (i.e. lower propensity for gene loss)

Gene expression level positively correlated with higher similarity scores (i.e. negatively correlated with evolutionary rate)

A significant fraction of the genes of an organism have a relatively high evolutionary rate…

Earliest Estimates of Number of Protein Families - ~1000

- Zuckerkandl,E. (1974) Accomplissement et perspectives de la paleogenetique chimique. In: Ecole de Roscoff –1974, p. 69. Paris:CNRS.

  "The appearance of new structures and functions in proteins during evolution", J. Mol. Evol. 7, 1-57 (1975).

- Dayhoff, M.O. (1974) Federation Proceedings 33, 2314.

  "The origin and evolution of protein superfamilies", Fed.Proc. 35, 2132-2138 (1976).

Atlas of Protein Sequence and Structure, Vol. 5, Supplement 3 (1978) pg. 10:

"It has been estimated that in humans there are approximately 50,000 proteins of functional or medical importance. … A landmark of molecular biology will occur when one member of each superfamily has been elucidated. At the present rate of 25 per year, this will take less than 15 years."

Fed Proc. 1976 Aug;35(10):2132-8.

The origin and evolution of protein superfamilies.

Dayhoff MO.

The organization of proteins into superfamilies based primarily on their sequences is introduced: examples are given of the methods used to cluster the related sequences and to elucidate the evolutionary history of the corresponding genes within each superfamily. Within the framework of this organization, the amount of sequence information currently and potentially available in all living forms can be discussed. The 116 superfamilies already sampled reflect possibly 10% of the total number. There are related proteins from many species in all of these superfamilies, suggesting that the origin of a new superfamily is rare indeed. The proteins so far sequenced are so rigorously conserved by the evolutionary process that we would expect to recognize as related descendants of any protein found in the ancestral vertebrate. The evolutionary history of the thyrotropin-gonadotropin beta chain superfamily is discussed in detail as an example. Some proteins are so constrained in structure that related forms can be recognized in prokaryotes and eukaryotes. Evolution in these superfamilies can be traced back close to the origin of life itself. From the evolutionary tree of the c-type cytochromes the identity of the prokaryote types involved in the symbiotic origin of mitochondria and chloroplasts begins to emerge.

Group	Criteria for Clustering Sequences	Identification Of Cluster
Superfamilies	Probability of Similarity by Chance<10^{-6}	Number
Families	<50% different	Letter
Subfamilies	<20% different	Paragraph
Atlas entries	<5% different	semicolon

Not particularly *evolutionary* perspective but only tiny sample from a number of different organisms…

69. Thymopoietin II
A. Thymopoietin II
Bovine
70. Thymosin alpha$_1$
A. Thymosin alpha$_1$
Bovine
71. Calcitonin
A. Calcitonins
Human; rat
Pig; bovine, sheep
Eel; salmon 1; salmon 2 and 3
72. Parathyrin
A. Parathyrin
Bovine; pig
73. Glucagon related
A. Glucagon
Pig, bovine, Arabian camel, human, rabbit, rat; duck, turkey, chicken
B. Gastric inhibitory polypeptide
Pig
C. Secretin
Pig
D. Vasoactive intestinal peptide
Pig; chicken
E. Pancreatic hormone
Chicken
F. Pancreatic hormone
Bovine
74. Motilin
A. Motilin
Pig
75. Proinsulin related
A. Insulin
Human[5], rabbit, hamster, pig[5], horse[5], elephant, bovine[5], sheep[5], camel, goat, sei whale, sperm whale, finback whale, dog, spiny mouse, rate 1[5] and 2[5], mouse 1 and 2; chicken, turkey; duck[5]; rattlesnake
Guinea pig[5]
Coypu
Cod, toadfish 1 and 2; angler fish, tuna 2; bonito
Atlantic hagfish
B. Insulin-like growth factors
I Human
II Human
C. Relaxin
Pig
76. Gastrin related
A. Gastrin
Human; pig
B. Cholecystokinin-pancreozymin
Pig
77. Paragonial peptide
A. Paragonial peptide PS-1
Fruit fly

[5] For those species indicated, the C-peptide, and in some cases the complete proinsulin sequence, is known.

Toxins

78. Snake venom toxins (proteroglyphs)
A. Long neurotoxins
Formosan banded krait 1
Broad-banded blue sea snake 1
Middle Asian cobra 1
Forest cobra 1
King cobra 2; king cobra 1
Forest cobra 2; Ethiopian cobra 1; Cape cobra 1; Thailand cobra 1; Indian cobra 1
Jameson's mamba 1
West African green mamba 1; W. African green mamba 2
Black mamba 1; black mamba 2
B. Venom proteins
Banded Egyptian cobra CM-13b; forest cobra S_4C_{11}
C. Short toxin 1
Green mamba
D. Short toxins 2
Green mamba
West African green mamba
E. Short neurotoxins
Black mamba 1; Jameson's mamba 1, West African green mamba 1
Banded Egyptian cobra 3 and 4
Banded Egyptian cobra 2, Cape cobra 1; forest cobra 1; ringhals 1; blackneck spitting cobra 1; Middle Asian cobra 1, *Naja naja philippinensis* 1, *Naja naja samarensis* 1; Cape cobra 2, banded Egyptian cobra 1; Formosan cobra 1; ringhals 2
Broad-banded blue sea snake 1
Beaked sea snake 1, yellow-bellied sea snake 1
Reef sea snake 1
F. Cytotoxins
Indian cobra 2, Formosan cobra 3; Middle Asian cobra 2; Formosan cobra 2 and 4; Indian cobra 1; forest cobra 1; Cambodian cobra 1, Formosan cobra 1; Middle Asian cobra 1; Mozambique cobra 4; banded Egyptian cobra 10; banded Egyptian cobra 9; banded Egyptian cobra 1, Cape cobra 1; Cape cobra 3; banded Egyptian cobra 3 and 8; banded Egyptian cobra 4; banded Egyptian cobra 2; Cape cobra 2; banded Egyptian cobra 5, 6, and 7
Mozambique cobra 1, blackneck spitting cobra 1; Mozambique cobra 2; Mozambique cobra 3
Ringhals 1
Banded Egyptian cobra 11
Forest cobra 3; forest cobra 2
79. Snake venom toxin (solenoglyphs)
A. Crotamine
South American rattlesnake
80. Arthropod neurotoxins
A. Neurotoxin
I Scorpion (*Androctonus*)
B. Neurotoxin
II Scorpion (*Androctonus*)
C. Neurotoxins
I North American scorpion
1 North American scorpion; 2 N. Am. scorpion; 3 N. Am. scorpion
D. Mast-cell degranulating peptide
Honey bee
81. Hemolytic peptides
A. Melittins
Major Honey bee[6], major Indian bee; Ceylon bee; minor honey bee; free-nesting bee
B. Bombinin
Unks
82. Heteronemertine worm neurotoxin
A. Neurotoxin B-IV
Heteronemertine worm
83. Sea anemone toxin
A. Toxins
II *Anemonia sulcata*; Anopleurin A *Anthopleura xanthogrammica*
84. Plant toxins
A. Mistletoe toxins
Viscotoxins A2, B, 1-PS European mistletoe; viscotoxin A3 European mistletoe
Phoratoxin California mistletoe
B. Purothionins
A-I Wheat; A-II wheat
85. Antibacterial proteins
A. Antibacterial substance A
Streptomyces carzinostaticus F41
B. Neocarzinostatin
Streptomyces carzinostaticus F41
86. Enterotoxin
A. Enterotoxin B
Staphylococcus aureus S6
87. Cholera enterotoxin beta chain
A. Cholera enterotoxin beta chain
Vibrio cholerae

[6] The complete promelittin sequence is known.

Immunoglobulin Related Proteins

88. Immunoglobulin variable (V) regions
A. Ig kappa chain V regions
Human Ni — Subgroup I
Human Ag; Au; Bi; Car; Dee; Eu; Gal; Hau; Ka; Lay; Ou; Rei; Roy; Scw — Subgroup I
Human Cum; Fr; Mil; Tew — Subgroup II
Human B6; Pom; Ti — Subgroup III
Human Len — Subgroup IV
Mouse MOPC 21; MPC 11
Mouse MOPC 41[7]
Mouse MOPC 173
Mouse 70; MOPC 321
Rat S211
Rabbit 2717
Rabbit 3315
Rabbit 3368
Rabbit 3374; 4135; BS-1; BS-5; K-25
Rabbit 3547
Rabbit K16-167
B. Ig lambda chain V regions, human
Ha — Subgroup I
New — Subgroup I
Newm — Subgroup I
Vor — Subgroup I
Bo; Bur; Mcg; Vil — Subgroup II
Nei — Subgroup II
Tro; Boh — Subgroup II
Sh — Subgroup III
Bau; X — Subgroup IV
Kern — Subgroup IV
Del — Subgroup V
C. Ig lambda chain V regions
Mouse MOPC 104E[7], J558, S104, S178; MOPC 315
Pig
D. Ig heavy chain V regions, human subgroup I
Eu
Nd
E. Ig heavy chain V regions, human subgroup II
Cor
Daw
He
Ou
F. Ig heavy chain V region, human subgroup II
Newm

[7] The precursor sequence is also known.

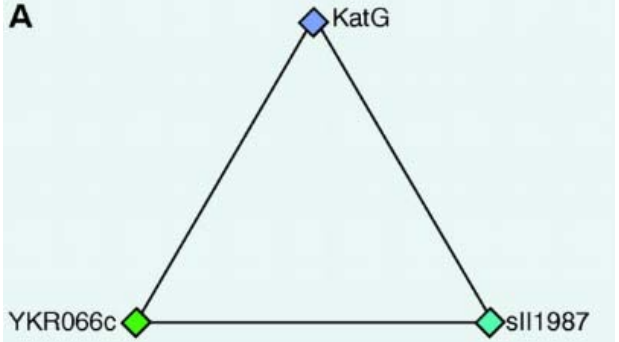
Science. 1997 Oct 24;278(5338):631-7.

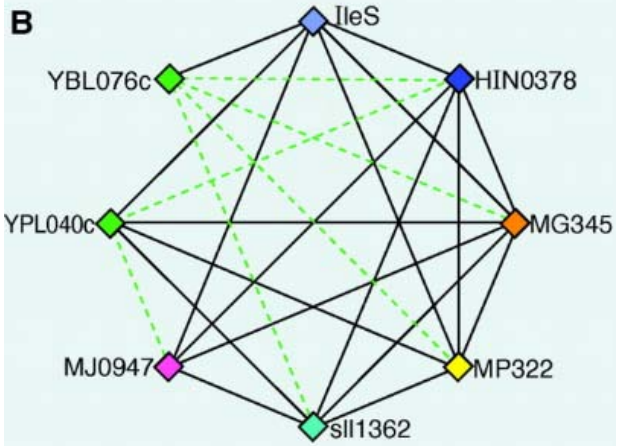
A genomic perspective on protein families.

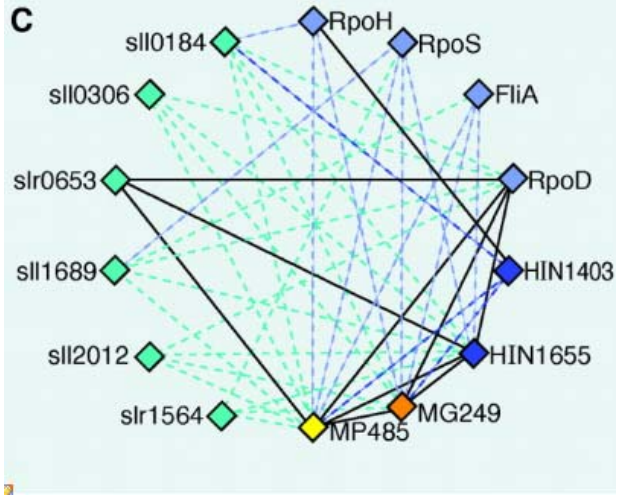
Tatusov RL, Koonin EV, Lipman DJ.
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

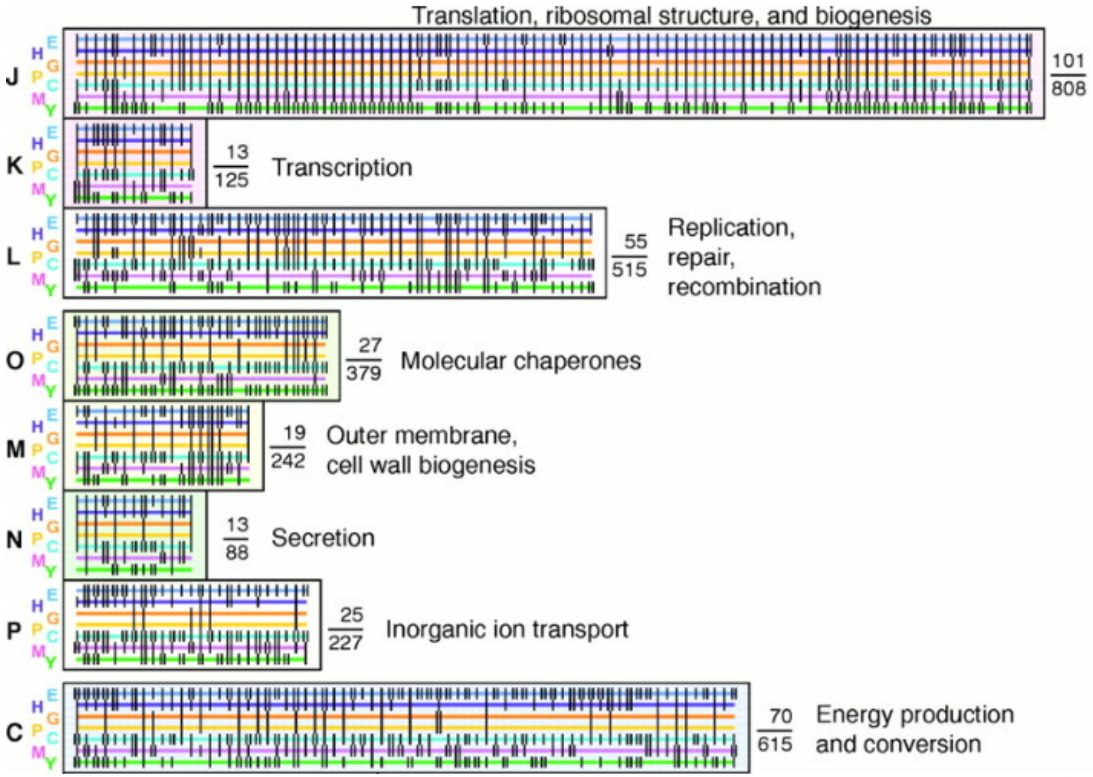
In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Comparison of proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

More sequence data should make the job of annotation easier…









Bacteria+Eukarya +Archaea		Bacteria+Eukarya		Bacteria+Archaea		Bacteria only	
Pattern	COGs	Pattern	COGs	Pattern	COGs	Pattern	COGs
eh___cmy	119	eh___c_y	80	eh___cm_	52	ehgpc___	53
ehgpcmy	114	ehgpc_y	66	e____cm_	43	e_gpc__	5
e____cmy	37	e____c_y	56	ehgpcm_	15	eh_pc__	2
eh____my	18	ehgp__y	5	e_gpcm_	4		
_____cmy	13	e_gpc_y	2	_h__cm_	3		
e_____my	7	e__p__y	1	eh_p_m_	2		
__gpcmy	4	e_gp__y	1	ehgp_m_	2		
_h____my	2	eh_pc_y	1	e_gp_m_	1		
eh_p_my	2	_h__c_y	1				
ehgp_my	2	__gpc_y	1				
e_gpcmy	2	_hgp__y	1				
__gp_my	1						
e_gp_my	1						
eh_pcmy	1						
Sum	323		215		122		60
COGs(%)	45		30		17		8

Have we been asking the question correctly?

How many protein families would there be?

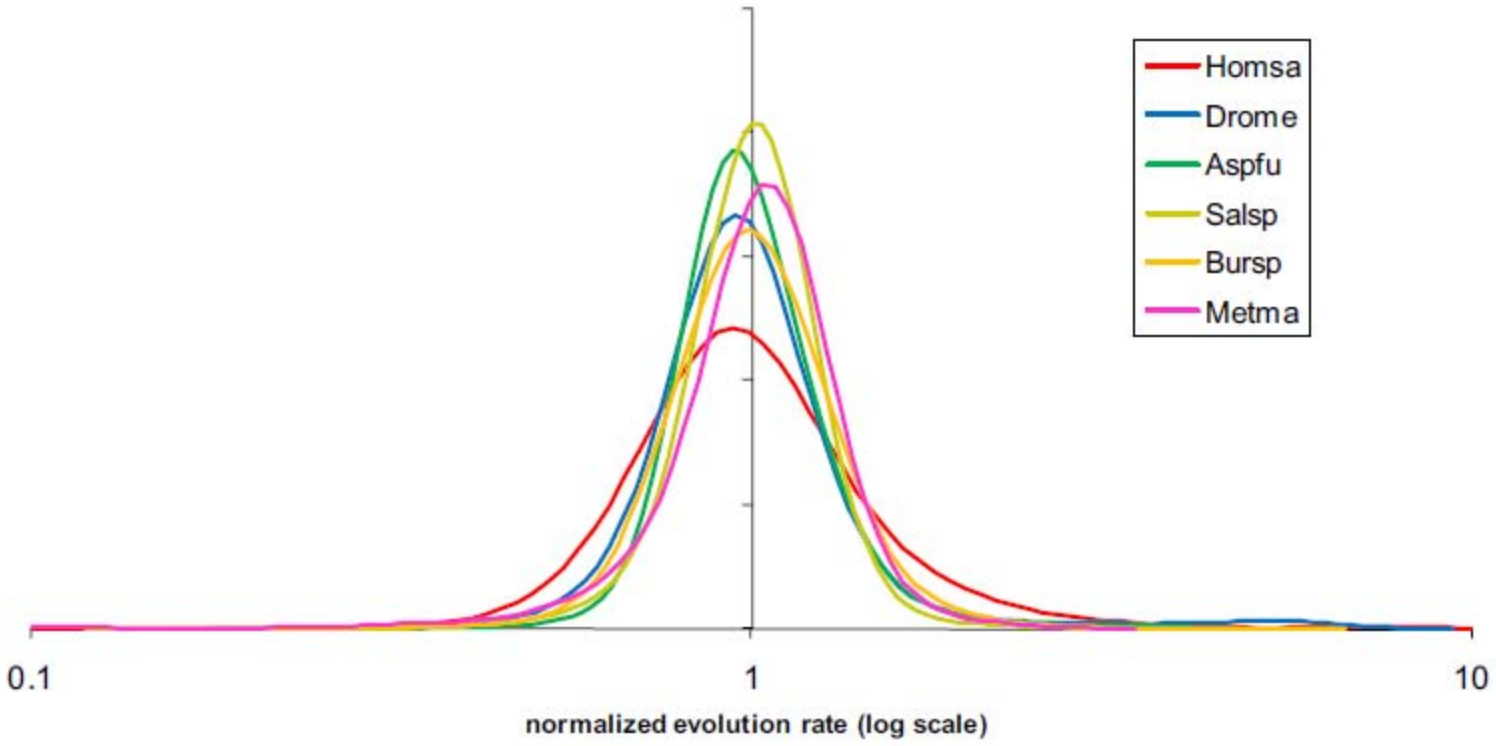
Or

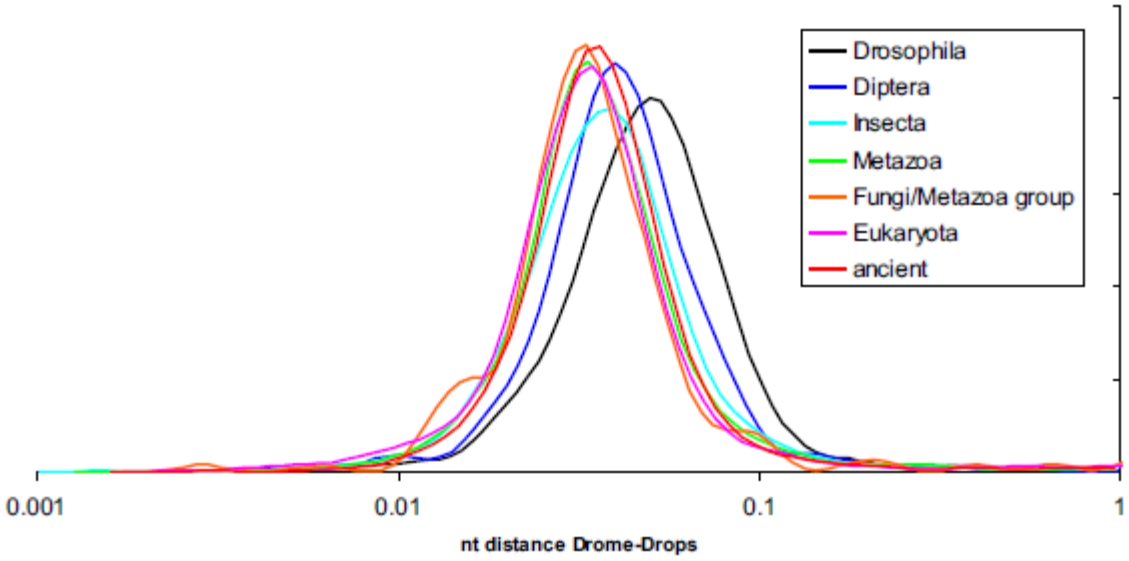
How many distant taxa?

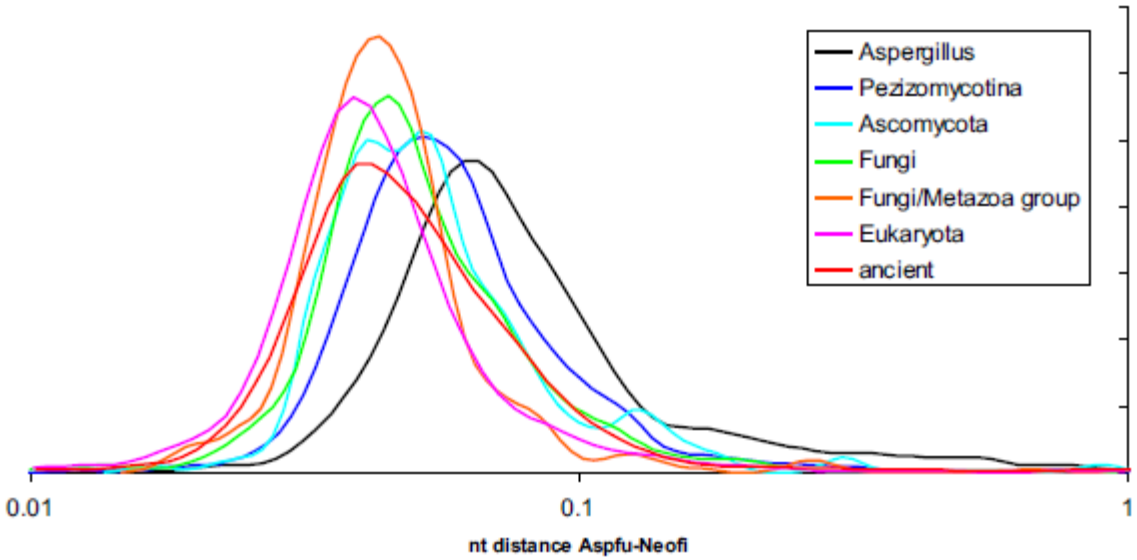
Proc Natl Acad Sci U S A. 2009 May 5;106(18):7273-80. Epub 2009 Apr 7.

Inaugural Article: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages.

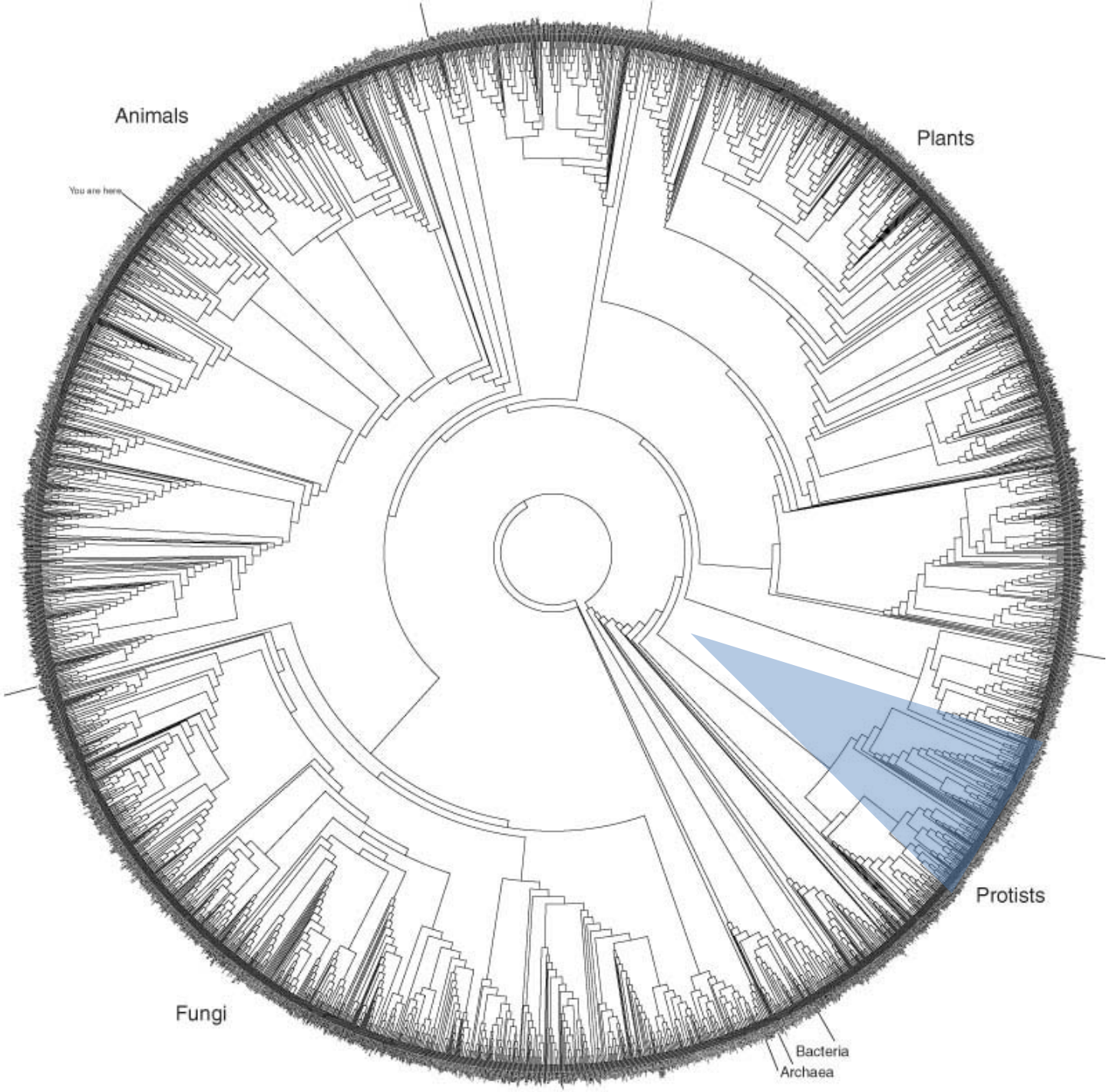
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ.





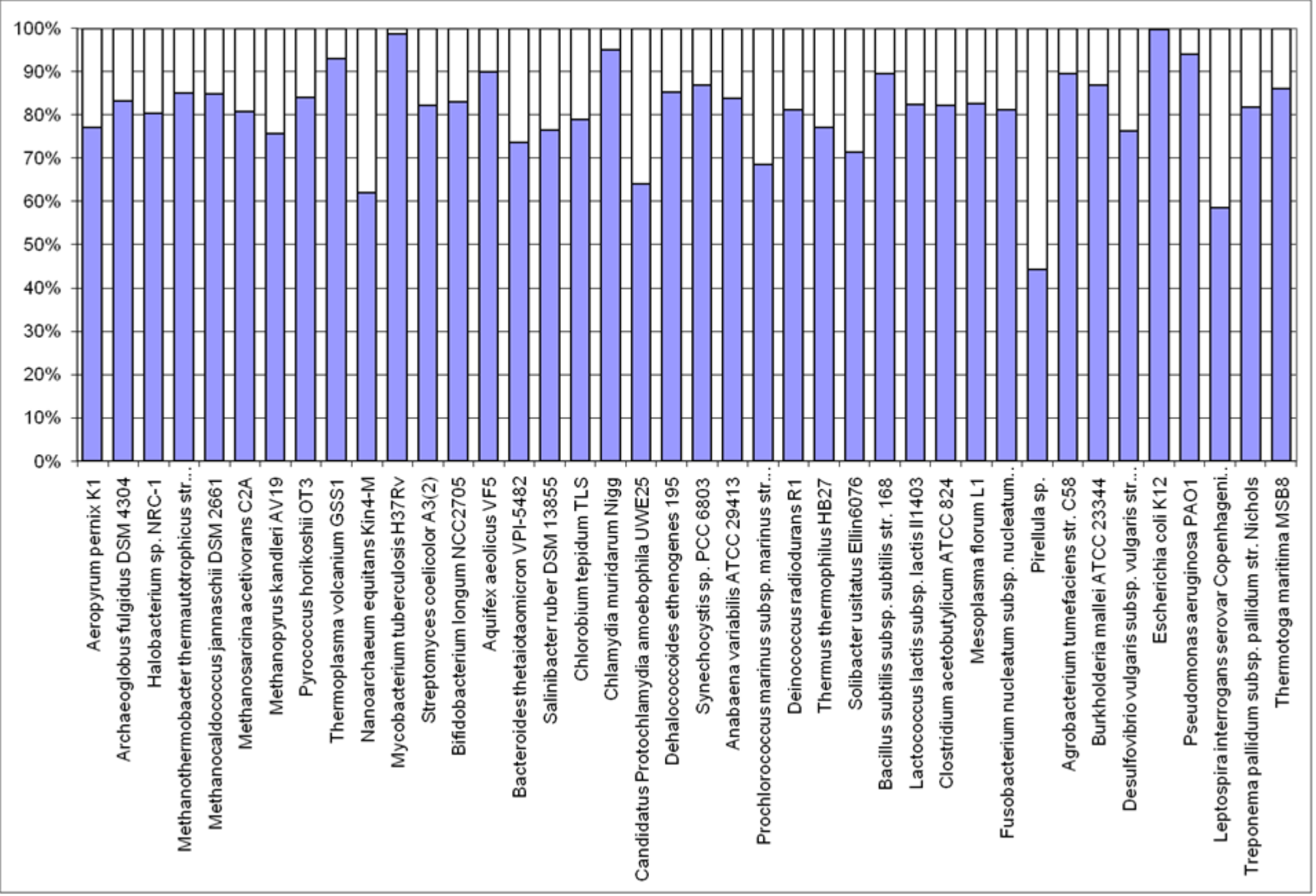


Slowly evolving proteins are being generated all the time…

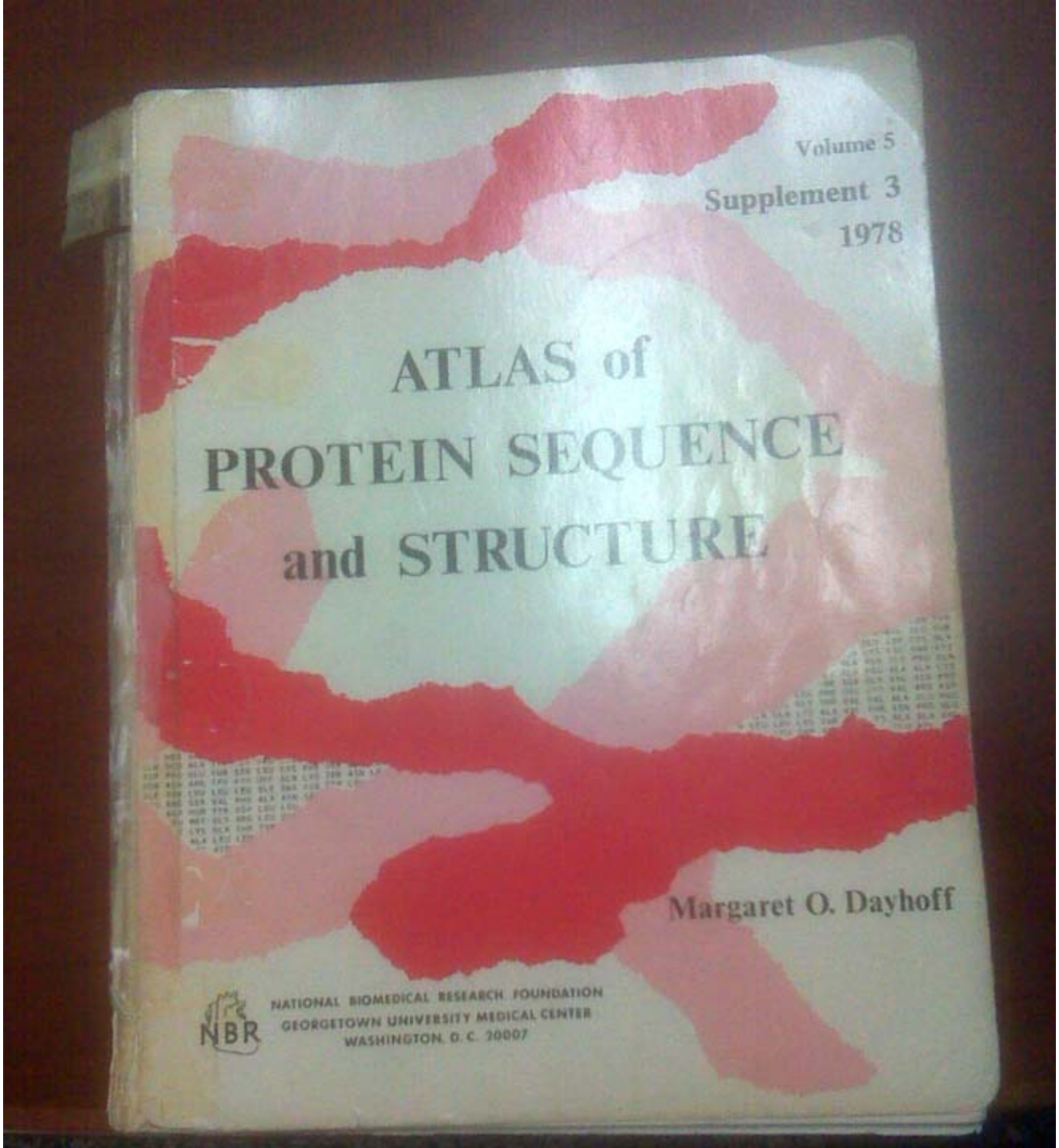


D.
HILLIS/UNIVERSITY
OF TEXAS, AUSTIN

Science, 2003,
300:1692-1697



Coverage of bacterial and archaeal genomes in the EggNOG (new COGs) database
Jensen et al. Nucleic Acids res. 2008, 36: D250-354; figure: Wolf-Koonin, unpublished


Volume 5
Supplement 3
1978
ATLAS of
PROTEIN SEQUENCE
and STRUCTURE
Margaret O. Dayhoff
NBR
NATIONAL BIOMEDICAL RESEARCH FOUNDATION
GEORGETOWN UNIVERSITY MEDICAL CENTER
WASHINGTON, D. C. 20007

ATLAS OF PROTEIN SEQUENCE AND STRUCTURE

Volume 5

SUPPLEMENT 3

1978

Library of Congress Card Catalogue Number 65-29342
ISBN 0-912466-07-3

Copies may be obtained from the publisher:

The National Biomedical Research Foundation
Post Office Box 629, Silver Spring, Maryland 20901

Also available from the publisher:

Protein Sequence Data Tape
Protein Segment Dictionary 78
Protein Data Search Services
Atlas of Protein Sequence and Structure
- Volume 5
- Supplement 1
- Supplement 2

Atlas Staff

Editor **Margaret O. Dayhoff, *Ph***

Senior Scientific Staff
Lois T. Hunt, *Ph.D.*
Winona C. Barker, *Ph.D.*
Robert M. Schwartz, *Ph.D.*
Charity L. Young, *Ph.D.*

Research Assistants
Lynne K. Ketcham, *M.S.*
Susan Hurst-Calderone, *M.S.*
Clare Marie Tomaselli, *B.S.*

Senior Analyst
Bruce C. Orcutt, *Ph.D.*

Editorial Coordinator
Margaret C. Blomquist, *B.*

Editorial Assistant
Cathy White, *B.S.*

Illustrator
Karen Cali Lawson

16 Contractile System Proteins

W.C. Barker, L.K. Ketcham, and M.O. Dayhoff

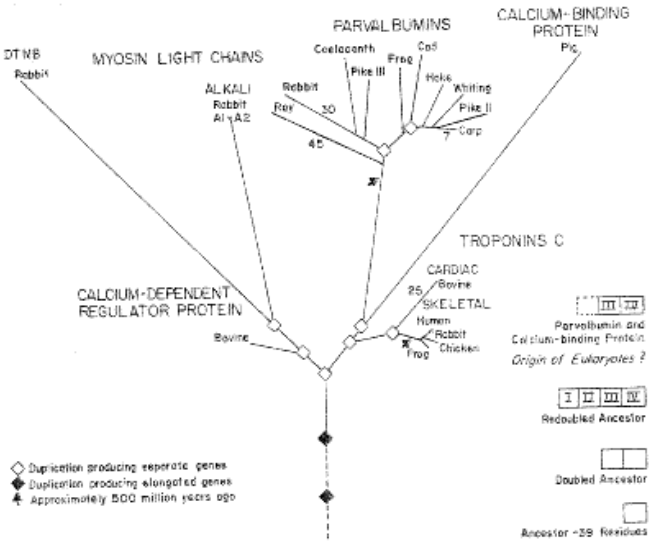


Figure 57. Evolutionary tree of the troponin C superfamily. The earliest events shown are two internal duplications that produced a gene four times as long as the ancestral gene, which probably coded for a calcium-binding peptide of about 39 amino acids. These internal duplications and one or more of the subsequent duplications to produce separate genes probably occurred in prokaryote ancestors. An early duplication gave rise to the two major branches of the tree. On one side parvalbumin and the calcium-binding protein diverged together from troponin C; these genes then must have lost the amino-terminal portion of the redoubled ancestor. Because the line to bovine cardiac troponin C arose before the divergence of skeletal muscle troponin C of frog, chicken, and rabbit from one another, a gene duplication is represented rather than a species divergence. The rate of change of troponin C is estimated to be 1.5 accepted point mutations per 100 residues per 100 million years. If troponin C has been changing at this unusually slow rate since the divergence of the cardiac and skeletal muscle forms, the gene duplications that allowed the specialization of cardiac and skeletal muscle may have occurred a billion years ago. On the other major branch of the tree, the first duplication produced the genes for the ancestral myosin light chain and for the calcium-dependent regulator protein. This protein is found in many tissues and it regulates various calcium-dependent events such as secretion, movement, cell division, and metabolic activity. Because it is the most slowly changing protein of this superfamily, its function probably corresponds most closely to the function of the common ancestor of these proteins. The next duplication to occur gave rise to the two main types of myosin light chains. The myosin A1 light chain is of recent origin as it is less than 4% different from the A2 chain, except for an amino-terminal 41-residue segment of unusual composition, which was not counted in constructing the tree. This tree was derived from matrices of estimated numbers of amino acid replacements between the sequences. It is a composite of topologies determined for the parvalbumins alone, for the sequences with four homology regions, and for half-chains of these compared with the parvalbumins and calcium-binding protein. By aligning the shorter sequences with halves of the longer sequences and constructing topologies that separately reflect the evolution of both halves of the longer chains, we determined where the point of earliest time was located and therefore where to place the trunk on the tree; the order of divergence of the branches to the parvalbumins and calcium-binding protein then became clear. A very slightly smaller tree was obtained by interchanging the branches to frog and chicken skeletal muscle troponin C, an arrangement that disagrees with accepted evidence on the order of divergence of these species. The branching order of the fish parvalbumins is not well resolved and also does not conform to that expected from biological evidence; it is clear that several duplications of the parvalbumin gene have occurred in these species. Only the two most clearly established duplications are shown. The branch lengths are proportional to the inferred number of mutations per 100 residues; these numbers are shown for several branches. The lengths of very long branches and of the internodal distances between such branches are rough estimates.

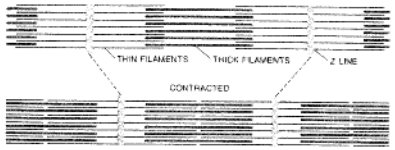


Figure 58. Structure of striated muscle fiber. Thin filaments extend from the *Z* lines, which are flat structures composed of protein. The thick filaments lie between the thin filaments, centered between *Z* lines. When the muscle contracts, the thin filaments slide past the thick filaments. This figure was taken, with permission, from "The Cooperative Action of Muscle Proteins," J.M. Murray and A. Weber. Copyright © February 1974 by Scientific American, Inc. All rights reserved.

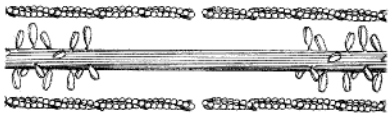


Figure 59. Thick and thin filaments of muscle. Extending from the thick filaments are the double heads of myosin molecules. These heads form crossbridges that interact with actin molecules in the thin filaments. During contraction the myosin heads attach, change orientation, and detach in such a way as to move the thin filaments relative to the thick filaments. This figure was taken, with permission, from "The Cooperative Action of Muscle Proteins," J.M. Murray and A. Weber. Copyright © February 1974 by Scientific American, Inc. All rights reserved.

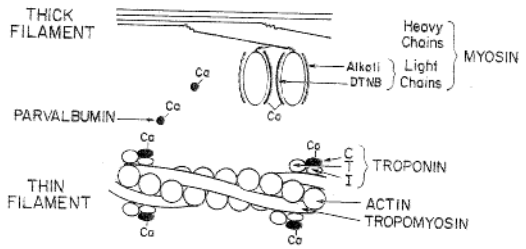


Figure 60. Proteins of the contractile element of skeletal muscle. Those that bind calcium are shown as solid black. Portions of the thin and thick filaments, including one crossbridge, are shown. The thick filaments contain myosin molecules, which consist of two heavy and four light chains. At one end of each heavy chain is a globular structure called the head, which contains the sites of actin binding and ATPase activity. Two types of light chains are associated with each head. The DTNB light chains bind calcium. The thin filaments are more complex. Tropomyosin molecules, which are coiled coils of two very similar chains, lie in the grooves between the two helical strands of actin monomers. Each tropomyosin molecule spans seven actin monomers. A troponin complex is associated with each tropomyosin molecule. Troponin T is responsible for the binding of the complex to tropomyosin. Troponin I inhibits the interaction of actin and myosin when the muscle is at rest. Troponin C is responsible for the regulation of muscle contraction by calcium ions. When the nerve impulse reaches the muscle cell, calcium is released from the sarcoplasmic reticulum and binds to troponin C, causing a conformational change in the relationship of these proteins such that actin and myosin are able to react with each other. Parvalbumin is a soluble calcium-binding protein that may play a role in the modulation of calcium ion concentrations in muscle cells. (Adapted from Figure 1 in Kendrick-Jones, J., and Jakes, R., Trends Biochem. Sci. 1, 281-284, 1976.)

NUMBER OF DIFFERENCES

	1	2	3	4	5	6	7	8	9	
CA-DEPENDENT REGULATOR 1 BOVINE		107	94	93	77	75	74	76	72	1
MYOSIN DTNB CHAIN 2 RABBIT	70		118	117	120	115	115	115	114	2
MYOSIN ALKALI CHAIN 3 A1 RABBIT	61	77		7	109	106	106	105	103	3
4 A2 RABBIT	61	76	5		110	105	105	104	102	4
TROPONIN C 5 CARDIAC BOVINE	51	78	71	71		50	50	44	48	5
6 SKELETAL HUMAN	50	76	69	69	33		1	15	13	6
7 SKELETAL RABBIT	49	76	69	69	33	1		15	13	7
8 SKELETAL CHICKEN	50	76	69	68	29	10	10		16	8
9 SKELETAL FROG	48	75	67	67	32	9	9	11		9

PERCENT DIFFERENCE

Matrix 37. Troponin C superfamily, longer sequences. This matrix is based on Alignment 35, with positions 1-38 and 200-205 omitted.

NUMBER OF DIFFERENCES

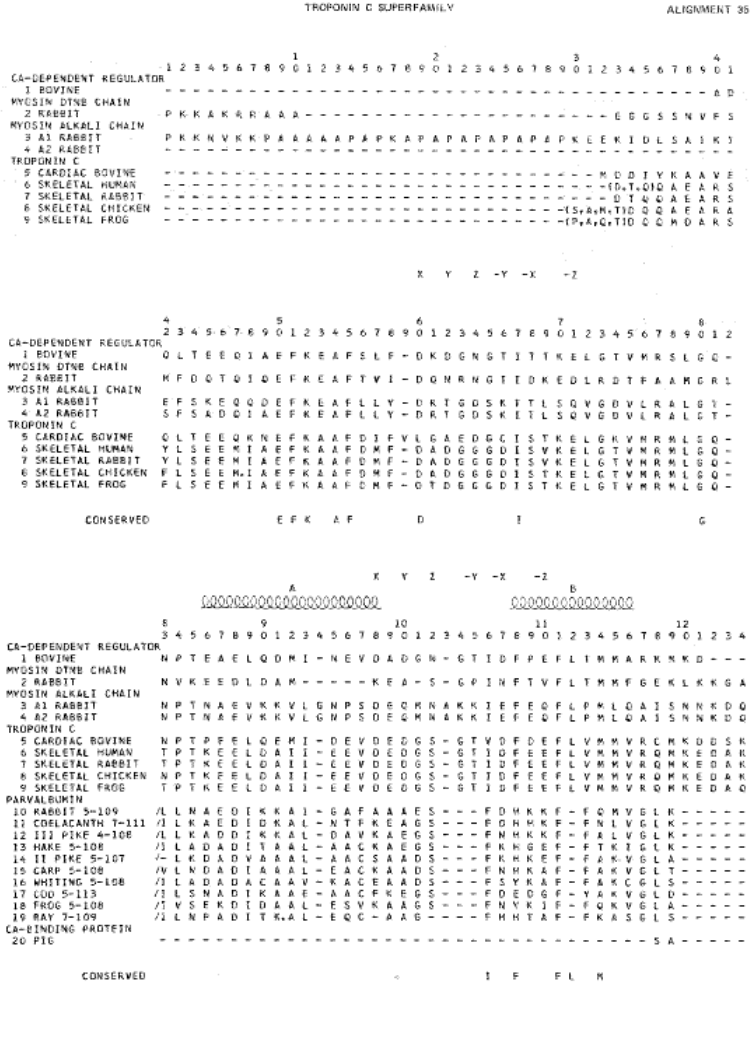
	10	11	12	13	14	15	16	17	18	19	20	
PARVALBUMIN 10 RABBIT		46	42	55	53	47	58	57	52	59	92	10
11 COELACANTH	41		44	53	58	51	56	55	53	55	96	11
12 III PIKE	39	40		49	50	43	52	45	49	64	94	12
13 HAKE	50	48	45		32	27	29	40	37	55	93	13
14 II PIKE	49	52	46	30		24	31	45	43	58	91	14
15 CARP	43	46	39	25	22		23	36	35	55	94	15
16 WHITING	53	50	48	27	29	21		40	39	60	94	16
17 COD	52	50	41	37	41	33	37		47	63	92	17
18 FROG	48	48	45	34	40	32	36	43		55	89	18
19 RAY	53	50	58	50	52	50	54	57	50		94	19
CA-BINDING PROTEIN 20 PIG	83	85	85	84	83	85	85	83	80	85		20

PERCENT DIFFERENCE

Matrix 38. Parvalbumins and calcium-binding protein. This matrix is based on an alignment similar to Alignment 35 but including the entire sequences except for four carboxyl-terminal residues of cod parvalbumin and calcium-binding protein.

TROPONIN C SUPERFAMILY

ALIGNMENT 35



Alignment 35. Troponin C superfamily. The proteins are grouped into families of sequences that are generally less than 50% different from one another. Conserved amino acids, shown beneath the alignment, are those found in more than half of the sequences in more than half of the families. All of the proteins shown were isolated from striated muscle except the calcium-binding protein from pig intestinal mucosa and the calcium-dependent regulator protein from bovine brain. Calcium-dependent regulator protein and skeletal muscle troponin C bind four calcium ions, cardiac muscle troponin C binds three calcium ions, parvalbumin binds two, and myosin DTNB light chain and intestinal calcium-binding protein each bind one. The myosin A1 and A2 light chains do not bind calcium. The positions of the proposed calcium ligands, which form the vertices of an octahedron, are designated by the letters X, Y, Z, -Y, -X, and -Z above the alignment. These letters represent the axes upon which the ligands fall. The six helical regions of carp parvalbumin are indicated by the coiled lines above the alignment and are labeled A-F. All of the sequences except coelacanth parvalbumin are acetylated or blocked in an undetermined manner at the amino end. A few residues at the beginning of the parvalbumins have been omitted.

● FLAGELLIN - BACILLUS SUBTILIS 168

DELANGE, R.J., CHANG, J.Y., SHAPER, J.H., AND GLAZER, A.N., J. BIOL. CHEM. 251, 705-711, 1976 (COMPLETE SEQUENCE WITH EXPERIMENTAL DETAILS)

CHANG, J.Y., DELANGE, R.J., SHAPER, J.H., AND GLAZER, A.N., J. BIOL. CHEM. 251, 695-700, 1976 (CNBR PEPTIDES)

SHAPER, J.H., DELANGE, R.J., MARTINEZ, R.J., AND GLAZER, A.N., J. BIOL. CHEM. 251, 701-704, 1976 (TRYPTIC PEPTIDES)

SEE THE ATLAS, VOL.5, SUPPL.2, P.250.

● TROPONYOSIN ALPHA CHAIN, SKELETAL MUSCLE - RABBIT

```
          5         10        15        20        25        30
  1 M D A I K K K M Q M L K L D K E N A L D R A E Q A E A D K K
 31 A A E D R S K Q L E D E L V S L Q K K L K G T E D E L D K Y
 61 S E A L K D A Q E K L E L A E K K A T D A E A D V A S L N R
 91 R I Q L V E E E L D R A Q E R L A T A L Q K L E E A E K A A
121 D E S E R G M K V I E S R A Q K D E E K M E I Q E I Q L K E
151 A K H I A E D A D R K Y E E V A R K L V I I E S D L E R A E
181 E R A E L S E G K C A E L E E E L K T V T N N L K S L E A Q
211 A E K Y S Q K E D K Y E E E I K V L S D K L K E A E T R A E
241 F A E R S V T K L E K S I D D L E D E L Y A Q K L K Y K A I
271 S E E L D H A L N D M T S I
```

COMPOSITION

36	ALA	A	14	GLN	Q	33	LEU	L	15	SER	S
14	ARG	R	56	GLU	E	39	LYS	K	8	THR	T
5	ASN	N	3	GLY	G	6	MET	M	0	TRP	W
24	ASP	D	2	HIS	H	1	PHE	F	6	TYR	Y
1	CYS	C	42	ILE	I	0	PRO	P	9	VAL	V

MOL. WT. UNMOD. CHAIN = 32,680 NUMBER OF RESIDUES = 284

STONE, D., AND SMILLIE, L.B., J. BIOL. CHEM. 253, 1137-1140, 1978 (COMPLETE SEQUENCE WITH EXPERIMENTAL DETAILS AND REVISION)
THE RESIDUE AT POSITION 24 IS GLN, NOT GLU.

SODEK, J., HODGES, R.S., AND SMILLIE, L.B., J. BIOL. CHEM. 253, 1129-1136, 1978 (SEQUENCE OF RESIDUES 142-284 WITH EXPERIMENTAL DETAILS)
THE SEQUENCE WAS DETERMINED ON A MIXED POPULATION OF TROPOMYOSIN CHAINS. AT 14 POSITIONS WHERE HETEROGENEITY WAS OBSERVED, THE AMINO ACID FOUND IN HIGHEST YIELD WAS ASSUMED TO BE CHARACTERISTIC OF THE ALPHA CHAIN.

STONE, D., SODEK, J., JOHNSON, P., AND SMILLIE, L.B., IN PROC. 9TH FEBS MTG., PP.125-136, PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST, 1974 (COMPLETE SEQUENCE, PRELIMINARY REPORT)
THE AMINO END IS ACETYLATED.

THE MOLECULE IS A COILED COIL OF TWO SIMILAR HELICAL CHAINS. THE SEQUENCE EXHIBITS A PROMINENT SEVEN-RESIDUE PERIODICITY.

● TROPONIN C, SKELETAL MUSCLE - HUMAN

ROMERO-HERRERA, A.E., CASTILLO, O., AND LEHMANN, H., J. MOL. EVOL. 8, 251-270, 1976 (SEQUENCE WITH EXPERIMENTAL DETAILS)
THE MAJOR COMPONENT OF HUMAN SKELETAL MUSCLE TROPONIN C DIFFERS FROM THAT OF RABBIT ONLY IN HAVING 112-PRO. RESIDUES 1-86 OF A MINOR COMPONENT APPEAR TO DIFFER FROM BOVINE CARDIAC TROPONIN C ONLY IN HAVING 62-GLU. PEPTIDES CORRESPONDING TO THE REMAINDER OF THE SEQUENCE WERE NOT FOUND.
THE AMINO END IS BLOCKED IN BOTH COMPONENTS.

SEE THE ATLAS, VOL.5, SUPPL.2, P.250, FOR THE RABBIT SEQUENCE.

● TROPONIN C, SKELETAL MUSCLE - RABBIT

COLLINS, J.H., GREASER, M.L., POTTER, J.D., AND HORN, M.J., J. BIOL. CHEM. 252, 6356-6362, 1977 (COMPLETE SEQUENCE WITH EXPERIMENTAL DETAILS)
THE SEQUENCE IS AS SHOWN IN THE ATLAS, VOL.5, SUPPL.2, P.250, WITH ALL PUNCTUATION REMOVED.

● TROPONIN C, SKELETAL MUSCLE - CHICKEN

```
          5         10        15        20        25        30
  1(S,A,M,T)D Q Q A E A R A F L S E E M.I A E F K A A F D M F D
 31 A D G G G D I S T K E L G T V M N R M L G Q N P T K E E L D A
 61 I I E E V D E D G S G T I D F E E F L V M M V R Q M K E D A
 91 K G K S E E E L A D C F R I F D K N A D G F I D I E E L G E
121 I L R A T G E H V T E E D I E D L M K D S D K N N D G R I D
151 F D E F L K M M E G V Q
```

COMPOSITION

13	ALA	A	5	GLN	Q	10	LEU	L	6	SER	S
6	ARG	R	25	GLU	E	10	LYS	K	7	THR	T
4	ASN	N	13	GLY	G	11	MET	M	0	TRP	W
21	ASP	D	1	HIS	H	11	PHE	F	0	TYR	Y
1	CYS	C	11	ILE	I	1	PRO	P	6	VAL	V

MOL. WT. UNMOD. CHAIN = 18,245 NUMBER OF RESIDUES = 162

WILKINSON, J.M., FEBS LETT. 70, 254-256, 1976 (SEQUENCE, PRELIMINARY REPORT)
THE AMINO END IS BLOCKED.

The origin of the single-letter code for the amino acids

The origin of the single-letter code for the amino acids is of historical interest, and in fact, this story may help the student to learn the code. The reason for the code is simple enough–in the very early days of bioinformatics, the very fastest computers were in fact, rather clunky. Dr. Margaret Oakley Dayhoff, arguably the founder of the field of bioinformatics, shortened the code from the three letter designations to the single letter code in an effort to reduce the size of the data files needed to describe the sequence of amino acids in a protein. The listing of amino acids, the three letter and single letter code, and the explanation for the choice of the single letter is given below. Note that there are 20 amino acids commonly found in proteins, and 26 letters in the alphabet. As a result, most of the letters are used.

To develop a single-letter code for the amino acids, Dr. Dayhoff attempted to make the code as easy to remember as possible. Of course, if the name of each amino acid began with a different letter, the code would be simple indeed. For 6 of the amino acids, the first letter of the name is unique, making the code simple. These are:

Amino Acid	3 letter code	Single letter code	Explanation
Cysteine	Cys	C	First letter of the name
Histidine	His	H	First letter of the name
Isoleucine	Ile	I	First letter of the name
Methionine	Met	M	First letter of the name
Serine	Ser	S	First letter of the name
Valine	Val	V	First letter of the name

For the other amino acids, the first letter of the name is not unique to a single amino acid, so Dr. Dayhoff assigned the letters A, G, L, P and T to the amino acids Alanine, Glycine, Leucine, Proline and Threonine, respectively, which occur more frequently in proteins than do the other amino acids having the same first letters.

Amino Acid	3 letter code	Single letter code	Explanation
Alanine	Ala	A	First letter of the name
Glycine	Gly	G	First letter of the name
Leucine	Leu	L	First letter of the name
Proline	Pro	P	First letter of the name
Threonine	Thr	T	First letter of the name

Some of the other amino acids are phonetically suggestive.

Amino Acid	3 letter code	Single letter code	Explanation
Arginine	Arg	R	aRginine
Phenylalanine	Phe	F	Fenylalanine
Tyrosine	Tyr	Y	tYrosine
Tryptophan	Trp	W	tWiptophan (or, contains Double ring)

http://www.biology.arizona.edu/biochemistry/problem_sets/aa/Dayhoff.html

Comp. Biochem. Physiol., Vol. 62B. pp. 1-5.

0305-0491/79/0115-0001$02.00/0

Margaret Dayhoff & Systems Biology…

MINIREVIEW

EVOLUTION OF HOMOLOGOUS PHYSIOLOGICAL MECHANISMS BASED ON PROTEIN SEQUENCE DATA

W. C. BARKER and M. O. DAYHOFF
National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C. 20007, U.S.A.

(*Received* 3 *May* 1978)

Abstract—1. Genetic duplications can give rise to homologous physiological mechanisms that include structurally related protein components. There are many such examples of related proteins within the human body.

2. Evolutionary histories showing the origins and subsequent divergences of these distantly related proteins can be derived from the protein sequences and correlated with the functional characteristics of these proteins.

3. The hormones related to glucagon provide an example of homology of physiological mechanisms and emergence of new functions subsequent to gene duplications.

4. The proteins related to troponin C illustrate the participation of distantly related proteins in the same mechanism (muscle contraction), the relationship of proteins characteristic of a specialized tissue to proteins found in all eukaryote cells, and the correlation of genetic duplications with the evolutionary appearance of different types of muscle.

HOMOLOGOUS PHYSIOLOGICAL MECHANISMS

Gene duplications in ancestral species have led to the presence of distantly related proteins in present-day organisms. These duplications provided the potential for major evolutionary advances including the emergence of new physiological mechanisms homologous (evolutionarily related) to existing mechanisms. A duplication may involve the entire genome, an individual chromosome, part of a chromosome, a single gene, or part of a gene (Ohno, 1970). Thereafter, the independently accumulating genetic changes will pro-

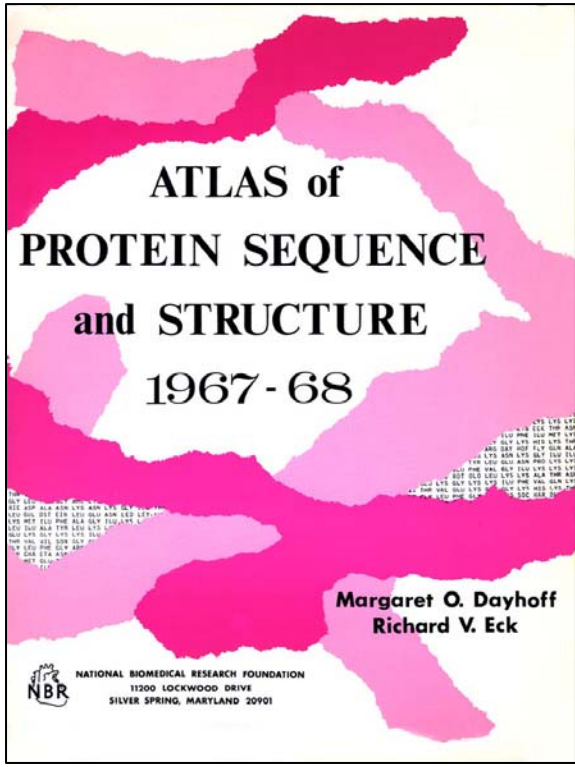
genome in ways that are to a considerable extent essential for the orderly differentiation and proper functioning of the mechanism. This genetic organization is also a result of an evolutionary history that includes different types of duplications, point mutations and crossover events. Entire mechanisms duplicate when a genome duplicates and perhaps also when a chromosome duplicates. Duplication of single genes produces related genes tandemly arranged on the same chromosome. These genes may evolve to produce proteins that appear serially during development, as do the epsilon, gamma, delta and beta chains

Orig Life. 1982 Mar;12(1):81-91.

Evolution of major metabolic innovations in the precambrian.

Barnabas J, Schwartz RM, Dayhoff MO.

A combination of the information on the metabolic capabilities of prokaryotes with a composite phylogenetic tree depicting an overview of prokaryote evolution based on the sequences of bacterial ferredoxin, 2Fe-2S ferredoxin, 5S ribosomal RNA, and c-type cytochromes shows three zones of major metabolic innovation in the Precambrian. The middle of these, which reflects the genesis of oxygen-releasing photosynthesis and aerobic respiration, links metabolic innovations of the anaerobic stem on the one hand and, on the other, proliferation of aerobic bacteria and the symbiotic associations leading to the eukaryotes. We consider especially those pathways where information on the structure of the enzymes is known. Halobacterium and Thermoplasma (archaebacteria) do not belong to a totally independent line on the basis of the composite tree but branch from the eukaryote cytoplasmic line.



Margaret Dayhoff