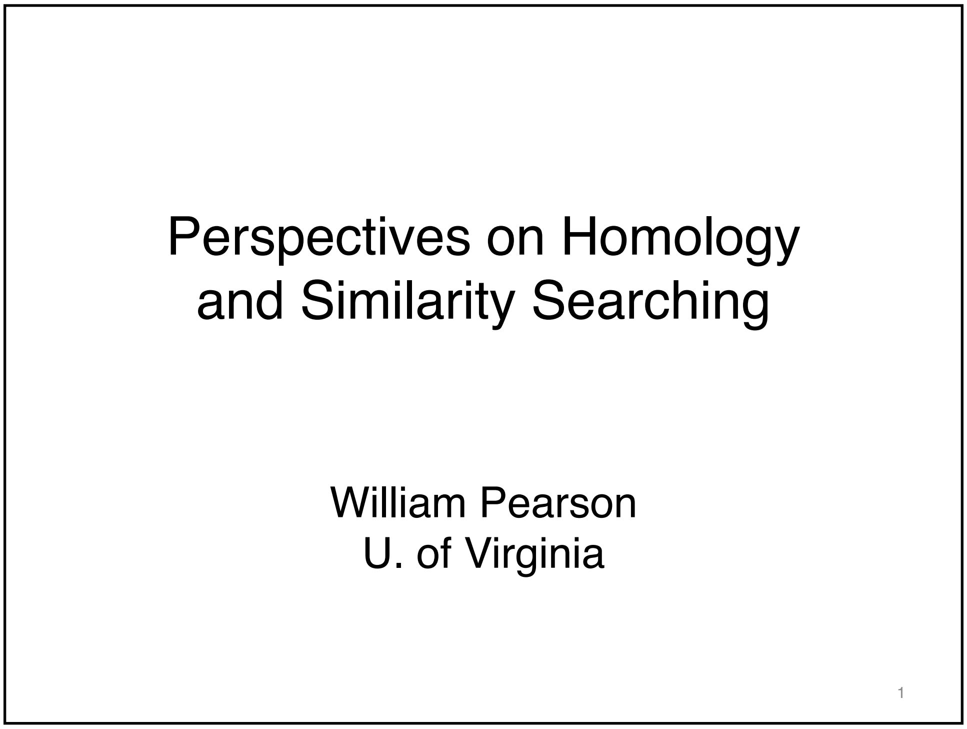
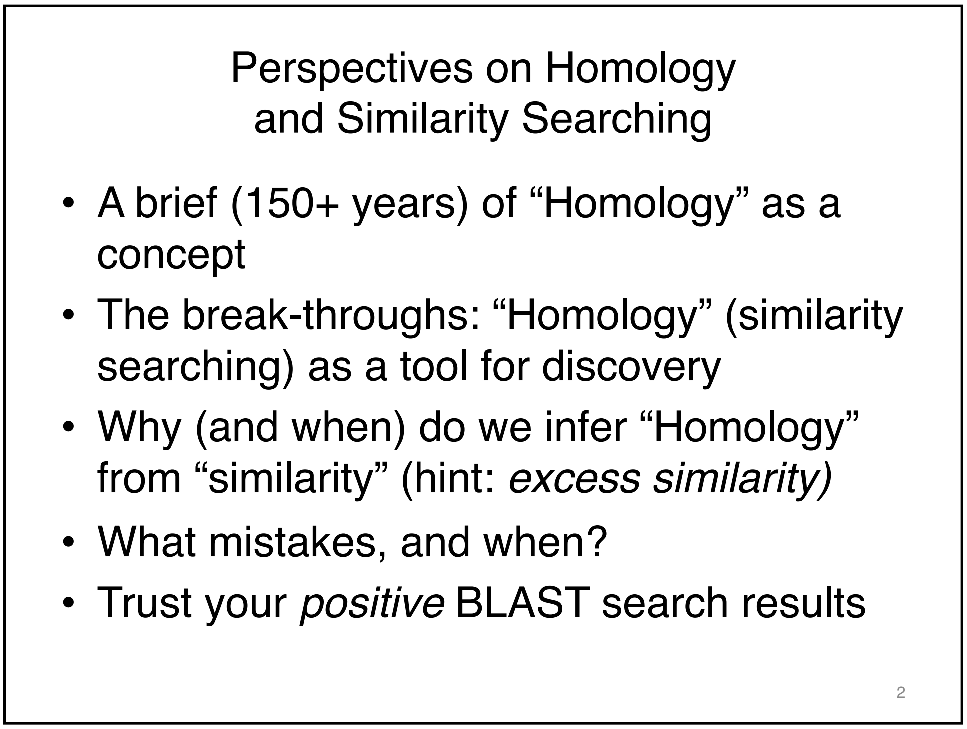


Perspectives on Homology and Similarity Searching

William Pearson
U. of Virginia

1

1

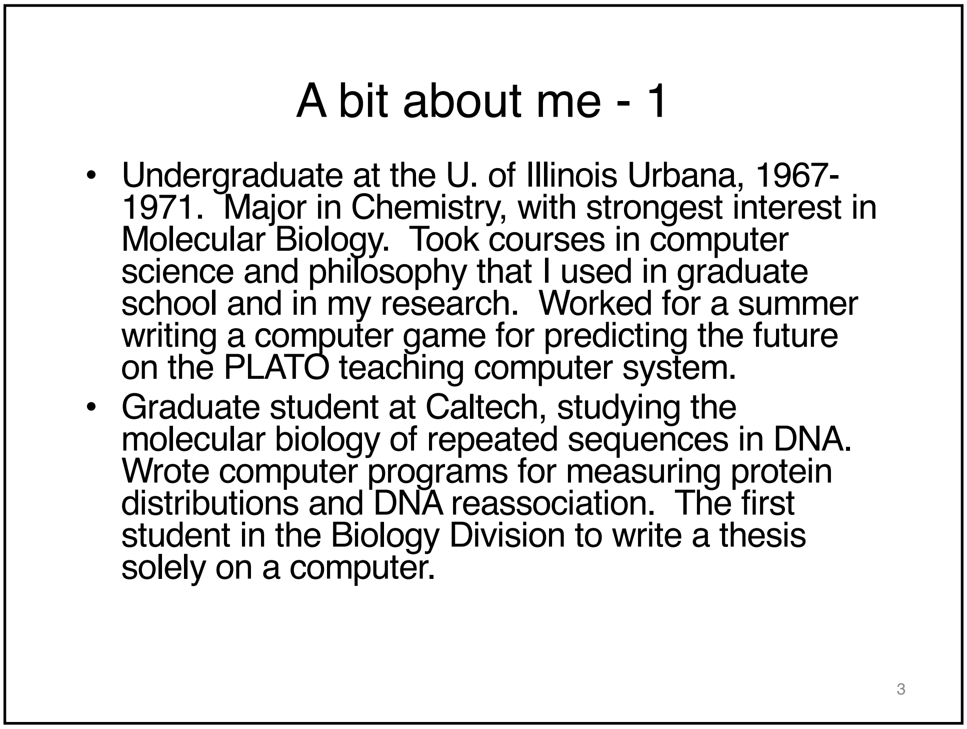


Perspectives on Homology and Similarity Searching

- A brief (150+ years) of “Homology” as a concept
- The break-throughs: “Homology” (similarity searching) as a tool for discovery
- Why (and when) do we infer “Homology” from “similarity” (hint: *excess similarity)*
- What mistakes, and when?
- Trust your *positive* BLAST search results

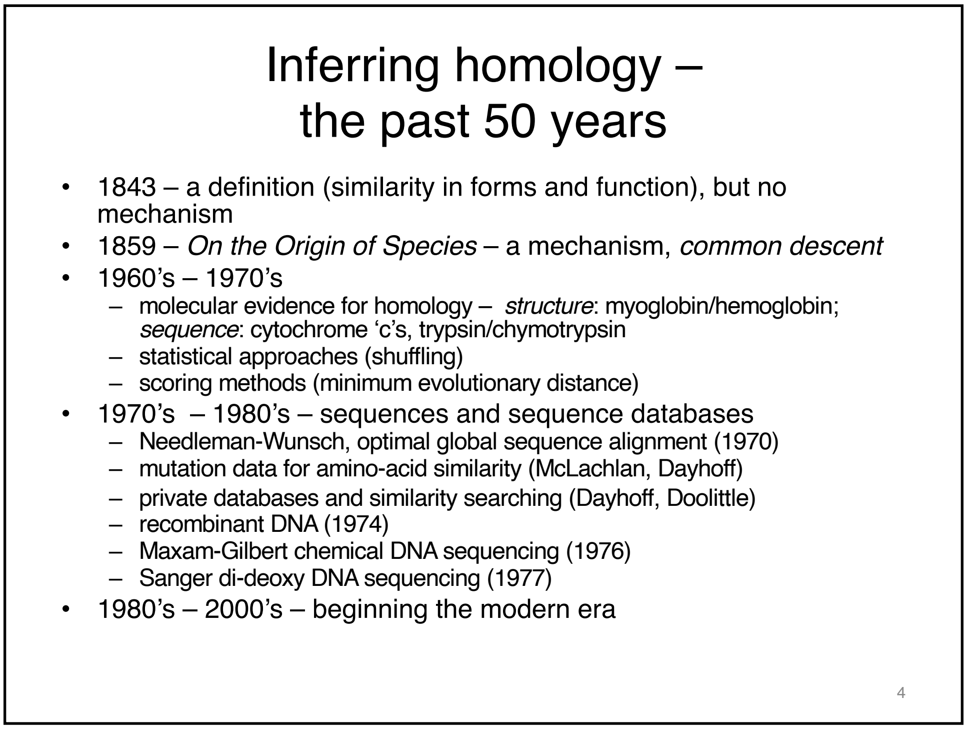
2

2



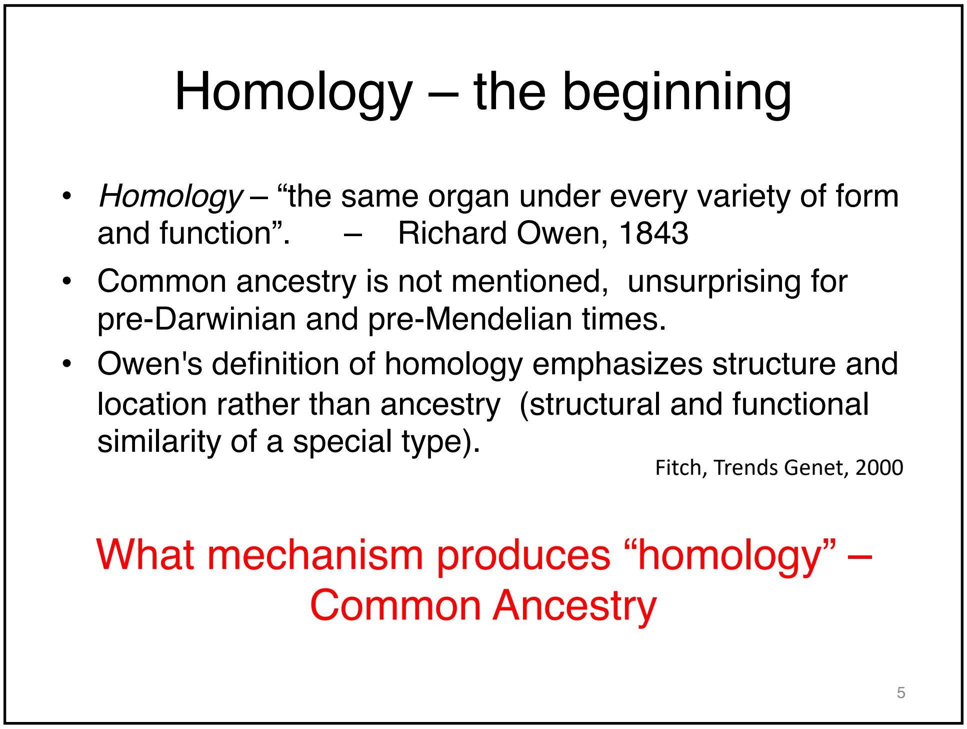
A bit about me - 1

- Undergraduate at the U. of Illinois Urbana, 1967-1971. Major in Chemistry, with strongest interest in Molecular Biology. Took courses in computer science and philosophy that I used in graduate school and in my research. Worked for a summer writing a computer game for predicting the future on the PLATO teaching computer system.
- Graduate student at Caltech, studying the molecular biology of repeated sequences in DNA. Wrote computer programs for measuring protein distributions and DNA reassociation. The first student in the Biology Division to write a thesis solely on a computer.



Inferring homology – the past 50 years

- 1843 – a definition (similarity in forms and function), but no mechanism
- 1859 – *On the Origin of Species* – a mechanism, *common descent*
- 1960's – 1970's
  - molecular evidence for homology – *structure*: myoglobin/hemoglobin; *sequence*: cytochrome 'c's, trypsin/chymotrypsin
  - statistical approaches (shuffling)
  - scoring methods (minimum evolutionary distance)
- 1970's – 1980's – sequences and sequence databases
  - Needleman-Wunsch, optimal global sequence alignment (1970)
  - mutation data for amino-acid similarity (McLachlan, Dayhoff)
  - private databases and similarity searching (Dayhoff, Doolittle)
  - recombinant DNA (1974)
  - Maxam-Gilbert chemical DNA sequencing (1976)
  - Sanger di-deoxy DNA sequencing (1977)
- 1980's – 2000's – beginning the modern era



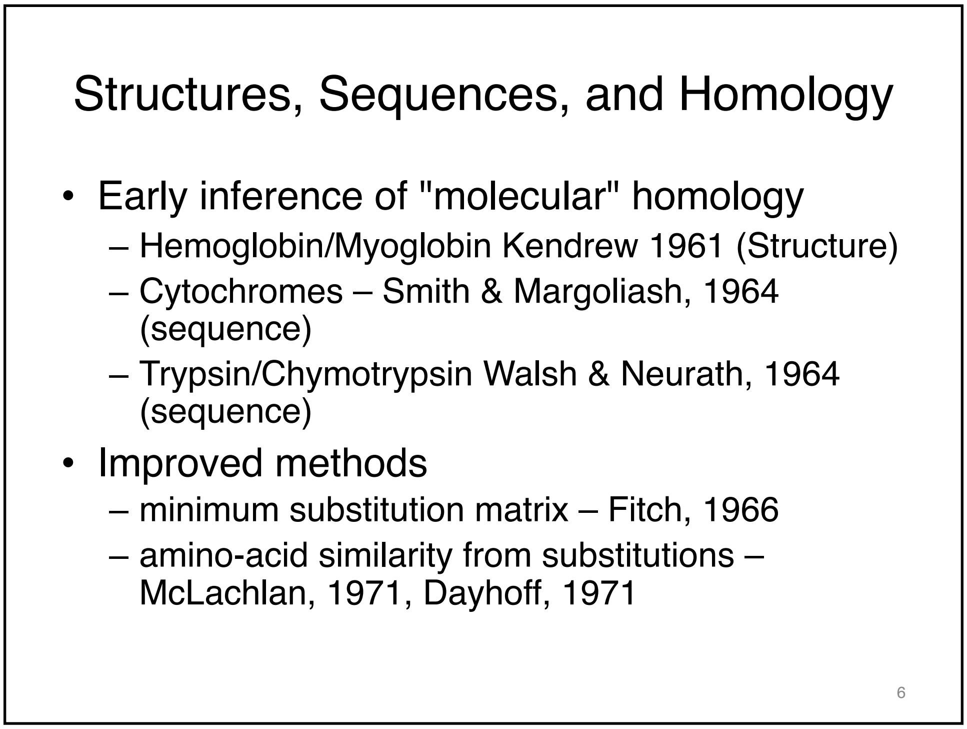
Homology – the beginning

- *Homology* – “the same organ under every variety of form and function”. – Richard Owen, 1843
- Common ancestry is not mentioned, unsurprising for pre-Darwinian and pre-Mendelian times.
- Owen's definition of homology emphasizes structure and location rather than ancestry (structural and functional similarity of a special type).

Fitch, Trends Genet, 2000

What mechanism produces “homology” – Common Ancestry

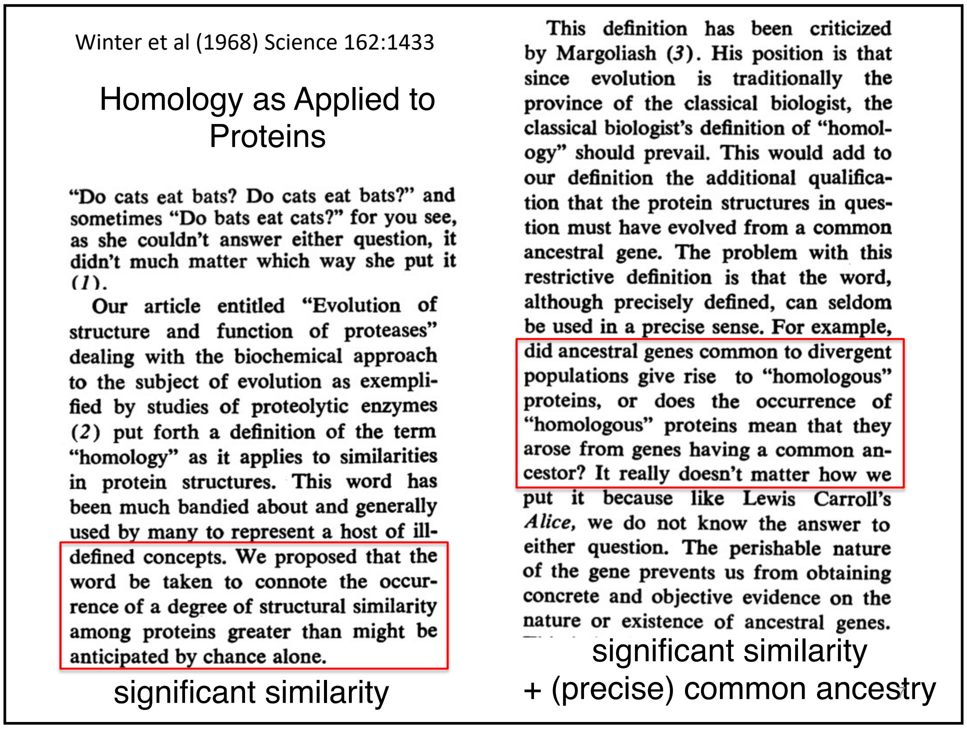
5



Structures, Sequences, and Homology

- Early inference of "molecular" homology
  - Hemoglobin/Myoglobin Kendrew 1961 (Structure)
  - Cytochromes – Smith & Margoliash, 1964 (sequence)
  - Trypsin/Chymotrypsin Walsh & Neurath, 1964 (sequence)
- Improved methods
  - minimum substitution matrix – Fitch, 1966
  - amino-acid similarity from substitutions – McLachlan, 1971, Dayhoff, 1971

6



Winter et al (1968) Science 162:1433

Homology as Applied to Proteins

"Do cats eat bats? Do cats eat bats?" and sometimes "Do bats eat cats?" for you see, as she couldn't answer either question, it didn't much matter which way she put it (1).

Our article entitled "Evolution of structure and function of proteases" dealing with the biochemical approach to the subject of evolution as exemplified by studies of proteolytic enzymes (2) put forth a definition of the term "homology" as it applies to similarities in protein structures. This word has been much bandied about and generally used by many to represent a host of ill-defined concepts. We proposed that the word be taken to connote the occurrence of a degree of structural similarity among proteins greater than might be anticipated by chance alone.

significant similarity

This definition has been criticized by Margoliash (3). His position is that since evolution is traditionally the province of the classical biologist, the classical biologist's definition of "homology" should prevail. This would add to our definition the additional qualification that the protein structures in question must have evolved from a common ancestral gene. The problem with this restrictive definition is that the word, although precisely defined, can seldom be used in a precise sense. For example, did ancestral genes common to divergent populations give rise to "homologous" proteins, or does the occurrence of "homologous" proteins mean that they arose from genes having a common ancestor? It really doesn't matter how we put it because like Lewis Carroll's *Alice*, we do not know the answer to either question. The perishable nature of the gene prevents us from obtaining concrete and objective evidence on the nature or existence of ancestral genes.

significant similarity
+ (precise) common ancestry

Inferring homology – the past 50 years

- 1980's – 1990 – beginning the modern era
  - *Identification of common molecular subsequences*. Smith and Waterman, 1981. optimal local alignments. (not practical for database searching, > 24 hr for a single search against 3,000 proteins)
  - Genbank – first freely available sequence database (DNA, 1983)
  - Break throughs – oncogenes and viral proteins as kinases and growth factors
  - *Rapid similarity searches of nucleic acid and protein data banks.* Wilbur and Lipman 1983. Heuristic, lookup-based DNA searches.
  - exon shuffling, oncogenes as growth factors
  - *Rapid and sensitive protein similarity searches.* Lipman and Pearson, 1985 –1985 – freely available protein databases
  - *Profile analysis* Gribskov et al, (1987)
  - *Improved tools for biological sequence comparison* Pearson and Lipman, 1988 – FASTA
  - *CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.* 1988, Higgins and Sharp
  - *A basic local alignment search tool* Altschul et al., (1990) *BLAST* – statistical thresholds to focus on homologs
  - *Gapped BLAST and PSI-BLAST* Altschul et al. (1997), faster and more sensitive
  - *Complete genomes*

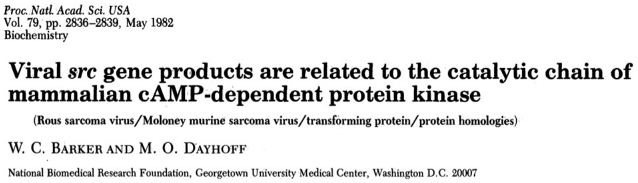
Perspectives on Homology and Similarity Searching

- A brief (150+ years) of “Homology” as a concept
- The break-throughs: “Homology” (similarity searching) as a tool for discovery
- Why (and when) do we infer “Homology” from “similarity” (hint: *excess similarity)*
- What mistakes, and when?
- Trust your *positive* BLAST search results

9

9

Surprises from sequence similarity searches



Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2836–2839, May 1982
Biochemistry

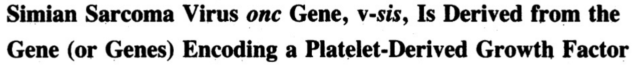
Viral *src* gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase

(Rous sarcoma virus/Moloney murine sarcoma virus/transforming protein/protein homologies)

W. C. Barker and M. O. Dayhoff

National Biomedical Research Foundation, Georgetown University Medical Center, Washington D.C. 20007

Proprietary database of ~3,000 protein sequences.
24 hr of VAX780 computer time ($100/hr)



Simian Sarcoma Virus *onc* Gene, v-*sis*, Is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor

Doolittle,R.F., Hunkapiller,M.W., Hood,L.E., Devare,S.G., Robbins,K.C., Aaronson,S.A. and Antoniades,H.N. (1983) *Science*, **221**, 275–277.

Proprietary database. (PDP-11??)

10

10

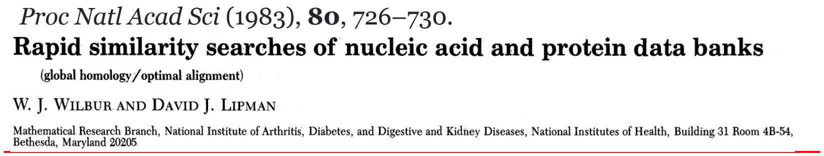
A bit about me - 2

- Post-doctoral fellow at Johns Hopkins (1978), where I went to learn how to clone recombinant DNA. Wrote computer programs for mapping restriction sites and assembling short DNA sequences (match, dmatch)
- Faculty member at U. of Virginia (1983). While waiting for my lab to be set up, worked with David Lipman to write the "FASTP" similarity searching, which became "FASTA", the predecessor to BLAST. Also cloned mouse and human glutathione S-transferases, discovered the human GSTM gene cluster, and the basis for the GSTM1 gene deletion.

11

11

40+ years of rapid sequence comparison



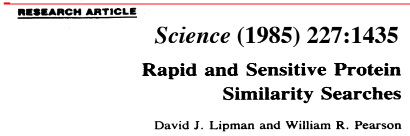
Proc Natl Acad Sci (1983), **80**, 726–730.

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



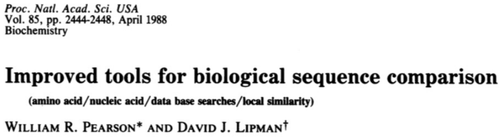
RESEARCH ARTICLE

Science **(1985) 227:1435**

Rapid and Sensitive Protein Similarity Searches

David J. Lipman and William R. Pearson

FASTP:
2,677 sequences



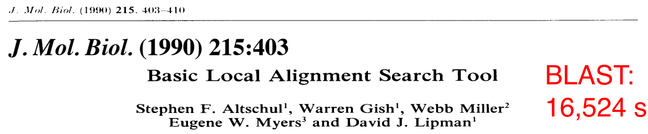
Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2444-2448, April 1988
Biochemistry

Improved tools for biological sequence comparison

(amino acid/nucleic acid/data base searches/local similarity)

WILLIAM R. PEARSON* AND DAVID J. LIPMAN†

FASTA:
4,253 sequences



J. Mol. Biol. (1990) 215, 403–410

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

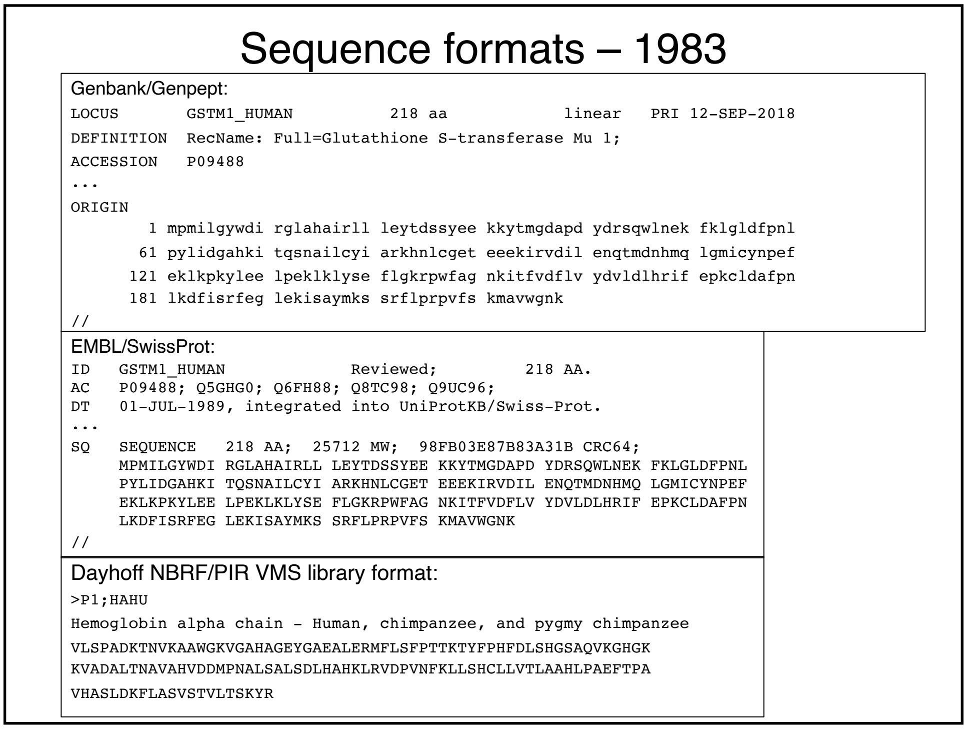
Basic Local Alignment Search Tool

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

BLAST:
16,524 sequences

12

12



Sequence formats – 1983

Genbank/Genpept:

```
LOCUS       GSTM1_HUMAN          218 aa            linear   PRI 12-SEP-2018
DEFINITION  RecName: Full=Glutathione S-transferase Mu 1;
ACCESSION   P09488
...
ORIGIN
        1 mpmilgywdi rglahairll leytdssyee kkytmgdapd ydrsqwlnek fklgldfpnl
       61 pylidgahki tqsnailcyi arkhnlcget eeekirvdil enqtmdnhmq lgmicynpef
      121 eklkpkylee lpeklklyse flgkrpwfag nkitfvdflv ydvldlhrif epkcldafpn
      181 lkdfisrfeg lekisaymks srflprpvfs kmavwgnk
//
```

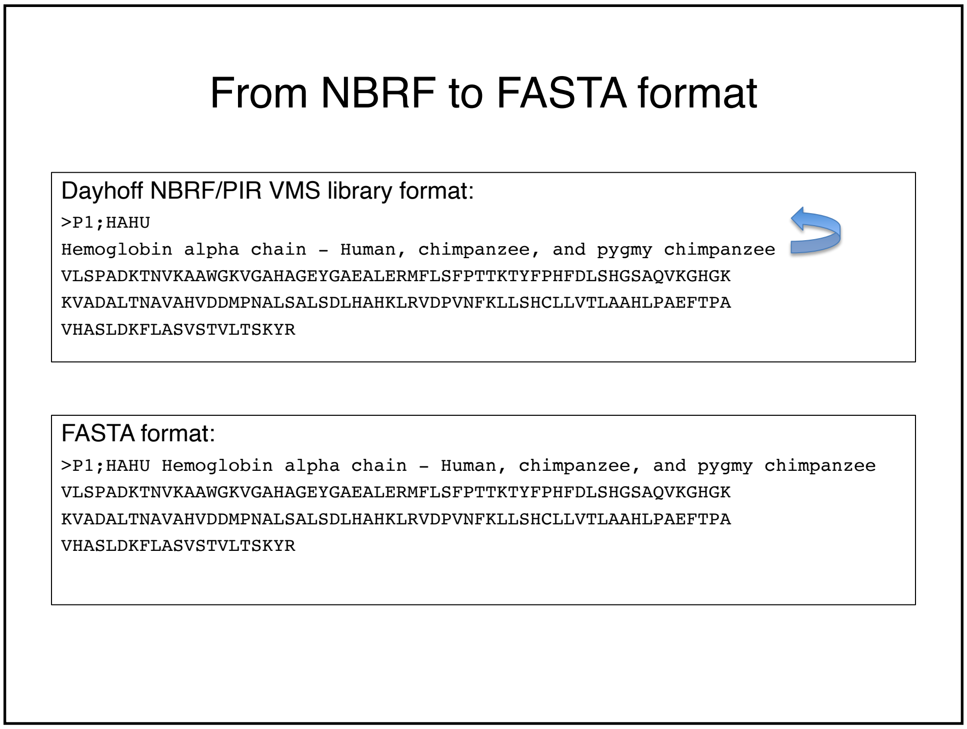
EMBL/SwissProt:

```
ID   GSTM1_HUMAN             Reviewed;         218 AA.
AC   P09488; Q5GHG0; Q6FH88; Q8TC98; Q9UC96;
DT   01-JUL-1989, integrated into UniProtKB/Swiss-Prot.
...
SQ   SEQUENCE   218 AA;  25712 MW;  98FB03E87B83A31B CRC64;
     MPMILGYWDI RGLAHAIRLL LEYTDSSYEE KKYTMGDAPD YDRSQWLNEK FKLGLDFPNL
     PYLIDGAHKI TQSNAILCYI ARKHNLCGET EEEKIRVDIL ENQTMDNHMQ LGMICYNPEF
     EKLKPKYLEE LPEKLKLYSE FLGKRPWFAG NKITFVDFLV YDVLDLHRIF EPKCLDAFPN
     LKDFISRFEG LEKISAYMKS SRFLPRPVFS KMAVWGNK
//
```

Dayhoff NBRF/PIR VMS library format:

```
>P1;HAHU
Hemoglobin alpha chain - Human, chimpanzee, and pygmy chimpanzee
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
```

13



From NBRF to FASTA format

Dayhoff NBRF/PIR VMS library format:

```
>P1;HAHU
Hemoglobin alpha chain - Human, chimpanzee, and pygmy chimpanzee
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
```

FASTA format:

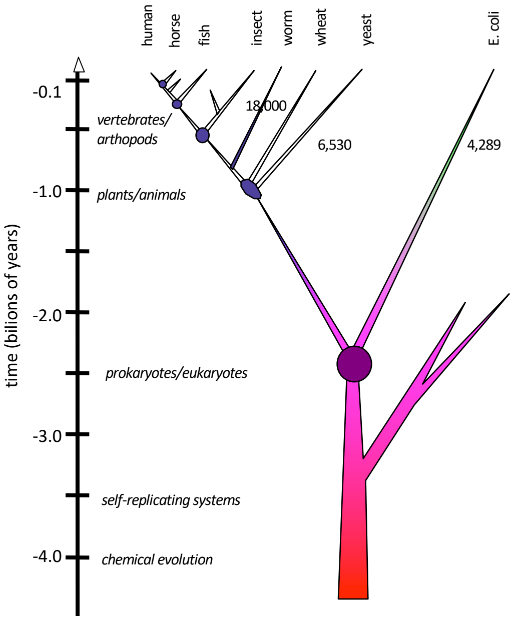
```
>P1;HAHU Hemoglobin alpha chain - Human, chimpanzee, and pygmy chimpanzee
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
```

14

Perspectives on Homology and Similarity Searching

- A brief (150+ years) of “Homology” as a concept
- The break-throughs: “Homology” (similarity searching) as a tool for discovery
- Why (and when) do we infer “Homology” from “similarity” (hint: *excess similarity)*
- What mistakes, and when?
- Trust your *positive* BLAST search results

Homologues share a common ancestor



Why do sequences look *similar*? Why do structures look *similar*?

Nebulae look similar because they are shaped by universal physical laws

convergence (not evolution or common ancestry)

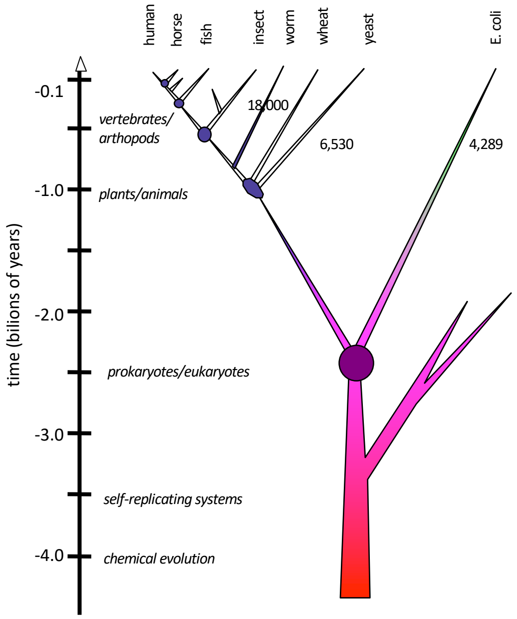
Organisms look similar when they share a common ancestors

beetle divergence:
300 Mya

Science
AAAS

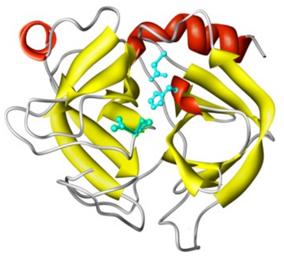
dog diversity:
15,000 yr

Homologues share a common ancestor

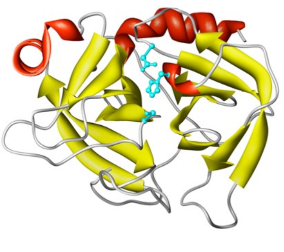


Proteins look similar when they share an ancestor

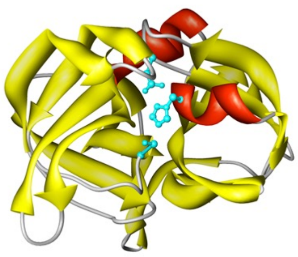
Homology <=> structural similarity
? sequence similarity



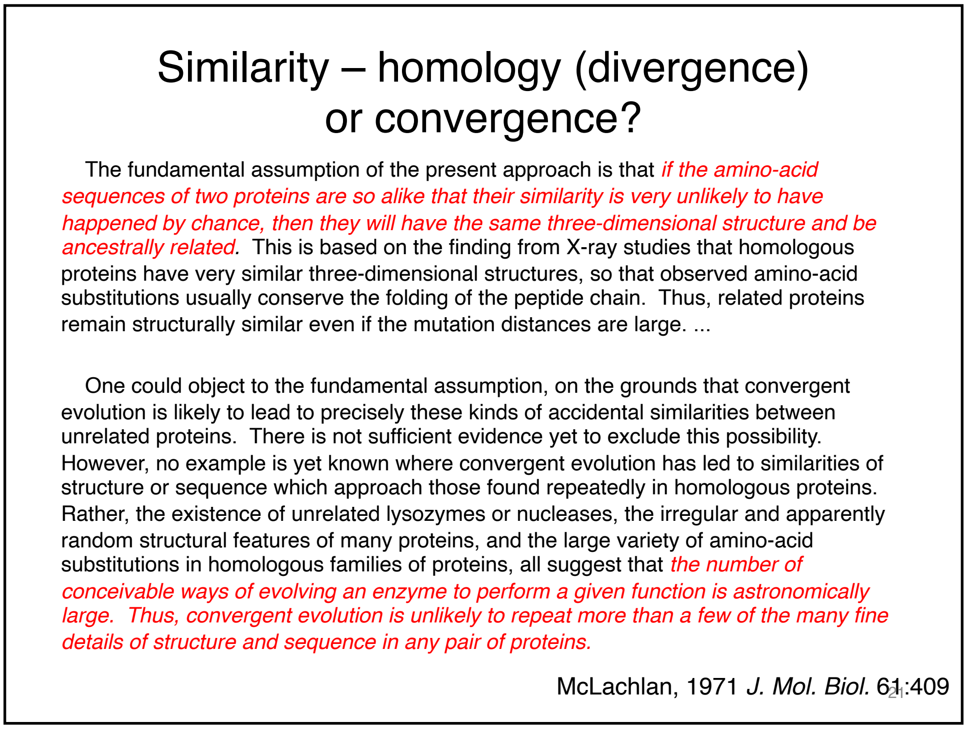
Bovine trypsin (5ptp)
Structure: E()< 10^{-23}; RMSD 0.0 A
Sequence: E()< 10^{-84} 100% 223/223



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



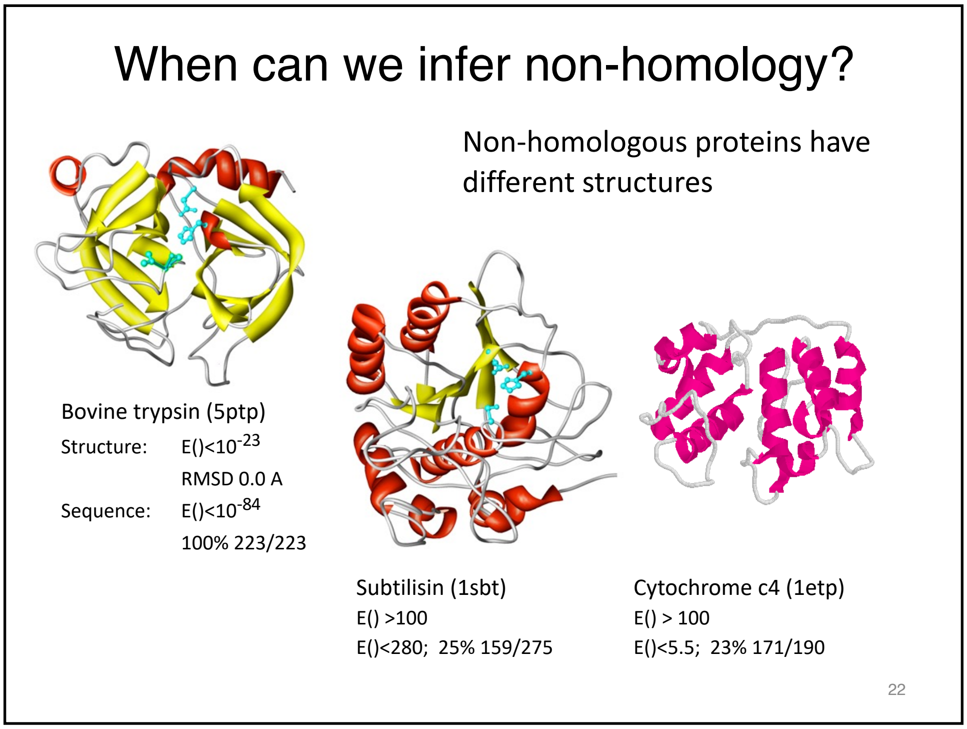
Similarity – homology (divergence) or convergence?

The fundamental assumption of the present approach is that *if the amino-acid sequences of two proteins are so alike that their similarity is very unlikely to have happened by chance, then they will have the same three-dimensional structure and be ancestrally related.* This is based on the finding from X-ray studies that homologous proteins have very similar three-dimensional structures, so that observed amino-acid substitutions usually conserve the folding of the peptide chain. Thus, related proteins remain structurally similar even if the mutation distances are large. ...

One could object to the fundamental assumption, on the grounds that convergent evolution is likely to lead to precisely these kinds of accidental similarities between unrelated proteins. There is not sufficient evidence yet to exclude this possibility. However, no example is yet known where convergent evolution has led to similarities of structure or sequence which approach those found repeatedly in homologous proteins. Rather, the existence of unrelated lysozymes or nucleases, the irregular and apparently random structural features of many proteins, and the large variety of amino-acid substitutions in homologous families of proteins, all suggest that *the number of conceivable ways of evolving an enzyme to perform a given function is astronomically large. Thus, convergent evolution is unlikely to repeat more than a few of the many fine details of structure and sequence in any pair of proteins.*

McLachlan, 1971 *J. Mol. Biol.* 61:409

21



When can we infer non-homology?

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)

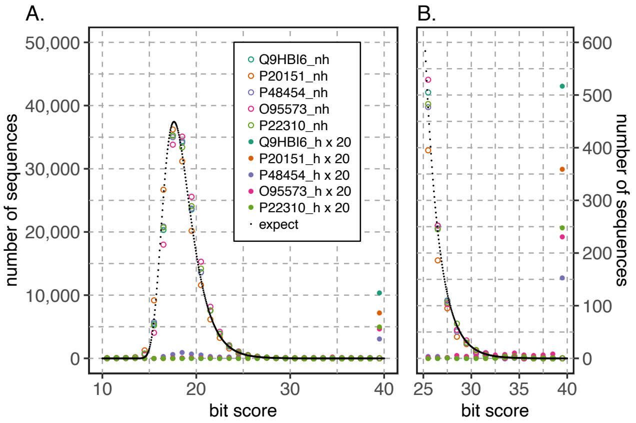
E() > 100

E()<5.5; 23% 171/190

22

22

Homology inferences are reliable because similarity statistics are accurate (I) (we know how unrelated sequences behave)



Distributions of similarity scores in searches with 5 human enzymes. Open circles (_nh) show scores for non-homologs. Closed circles show homolog (_h) scores.

23

23

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

With 100+ million protein sequences, it is easy to explore what can happen by chance
Proteins are not statistically similar by chance
Significant similarity implies common ancestry

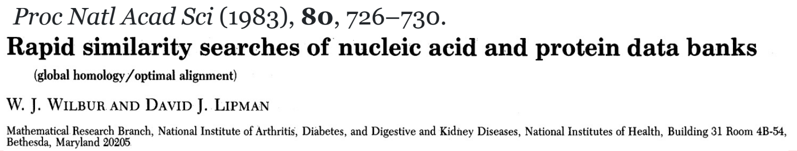
24

24

Inferring homology – the past 50 years

- 1990's – the genome deluge
  - 1995 *H. influenzae* – first complete bacterial genome
  - *Gapped BLAST and PSI-BLAST* Altschul et al. (1997), faster and more sensitive
  - 1997 – *E. coli, S. cerevisiae* genomes
  - 1999 – *C. elegans*
  - 2000 (March) – *D. melanogaster*
  - 2000 (June) – human genome draft sequence

40+ years of rapid sequence comparison



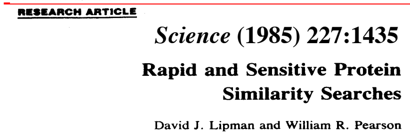
Proc Natl Acad Sci (1983), **80**, 726–730.

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



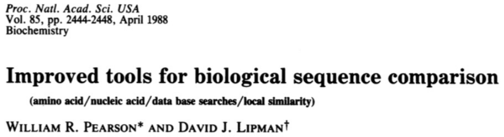
RESEARCH ARTICLE

Science (1985) 227:1435

Rapid and Sensitive Protein Similarity Searches

David J. Lipman and William R. Pearson

FASTP:
2,677 sequences



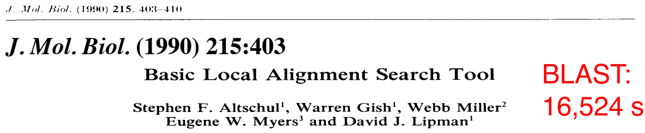
Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2444-2448, April 1988
Biochemistry

Improved tools for biological sequence comparison

(amino acid/nucleic acid/data base searches/local similarity)

William R. Pearson* and David J. Lipman†

FASTA:
4,253 sequences



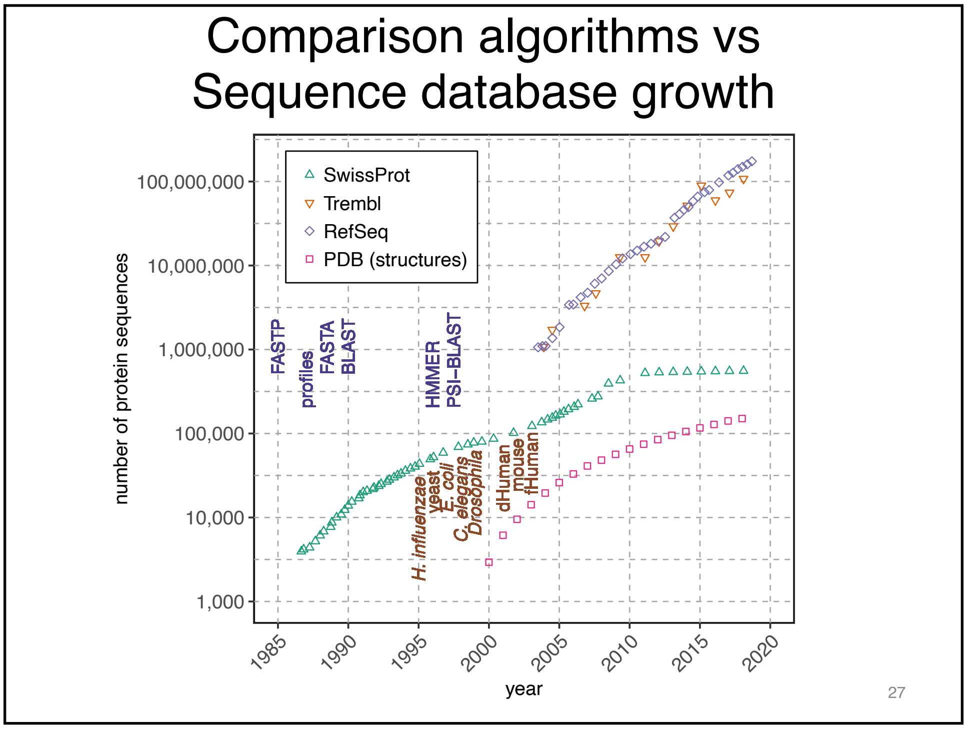
J. Mol. Biol. (1990) 215. 403–410

***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]

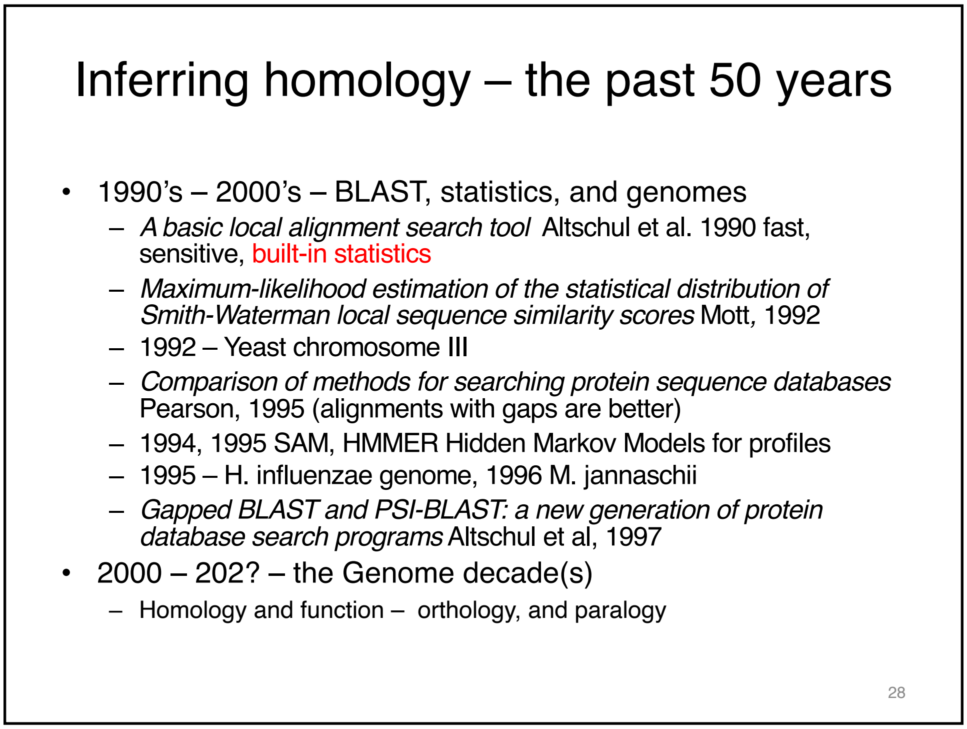
BLAST:
16,524 sequences



Comparison algorithms vs Sequence database growth

27

27

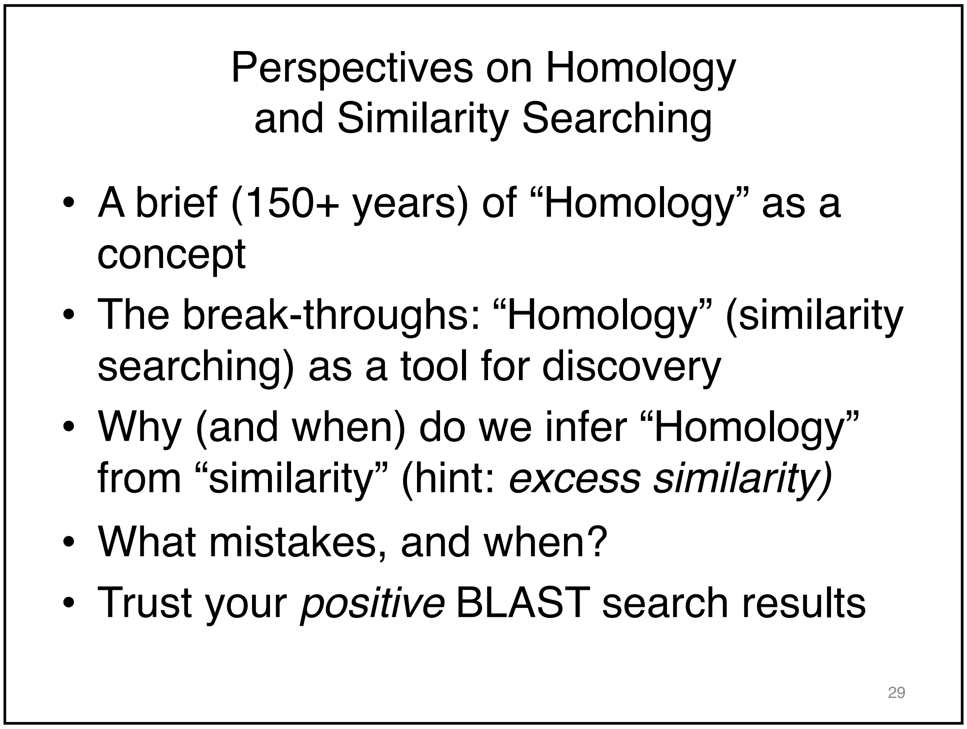


Inferring homology – the past 50 years

- 1990’s – 2000’s – BLAST, statistics, and genomes
  - *A basic local alignment search tool* Altschul et al. 1990 fast, sensitive, built-in statistics
  - *Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores* Mott, 1992
  - 1992 – Yeast chromosome III
  - *Comparison of methods for searching protein sequence databases* Pearson, 1995 (alignments with gaps are better)
  - 1994, 1995 SAM, HMMER Hidden Markov Models for profiles
  - 1995 – H. influenzae genome, 1996 M. jannaschii
  - *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs* Altschul et al, 1997
- 2000 – 202? – the Genome decade(s)
  - Homology and function – orthology, and paralogy

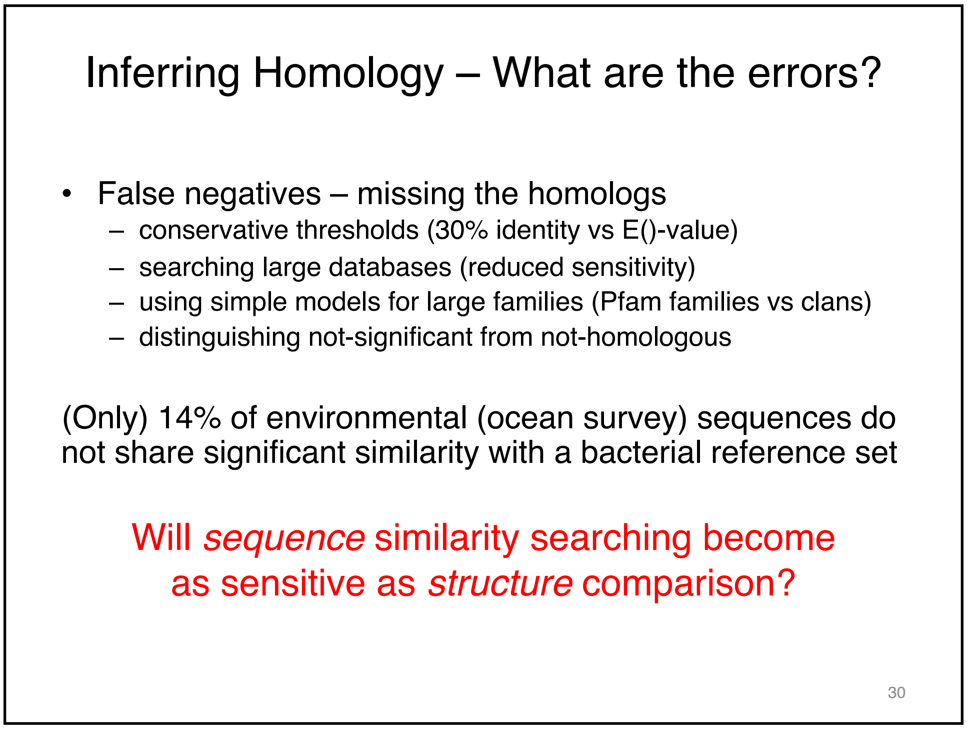
28

28



Perspectives on Homology and Similarity Searching

- A brief (150+ years) of “Homology” as a concept
- The break-throughs: “Homology” (similarity searching) as a tool for discovery
- Why (and when) do we infer “Homology” from “similarity” (hint: *excess similarity)*
- What mistakes, and when?
- Trust your *positive* BLAST search results



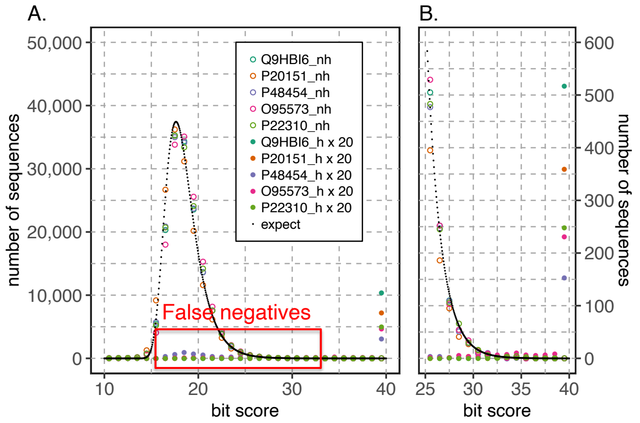
Inferring Homology – What are the errors?

- False negatives – missing the homologs
  - conservative thresholds (30% identity vs E()-value)
  - searching large databases (reduced sensitivity)
  - using simple models for large families (Pfam families vs clans)
  - distinguishing not-significant from not-homologous

(Only) 14% of environmental (ocean survey) sequences do not share significant similarity with a bacterial reference set

Will *sequence* similarity searching become as sensitive as *structure* comparison?

Homology inferences are reliable because similarity statistics are accurate (I) (we know how unrelated sequences behave)

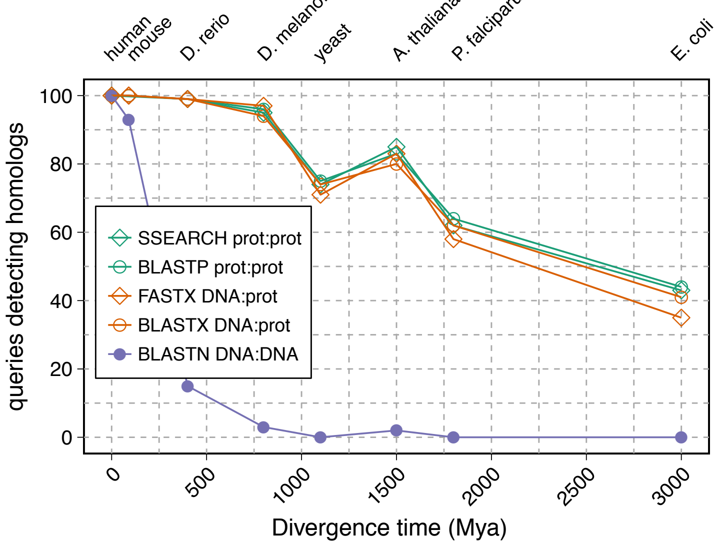


Distributions of similarity scores in searches with 5 human enzymes. Open circles (_nh) show scores for non-homologs. Closed circles show homolog (_h) scores.

31

31

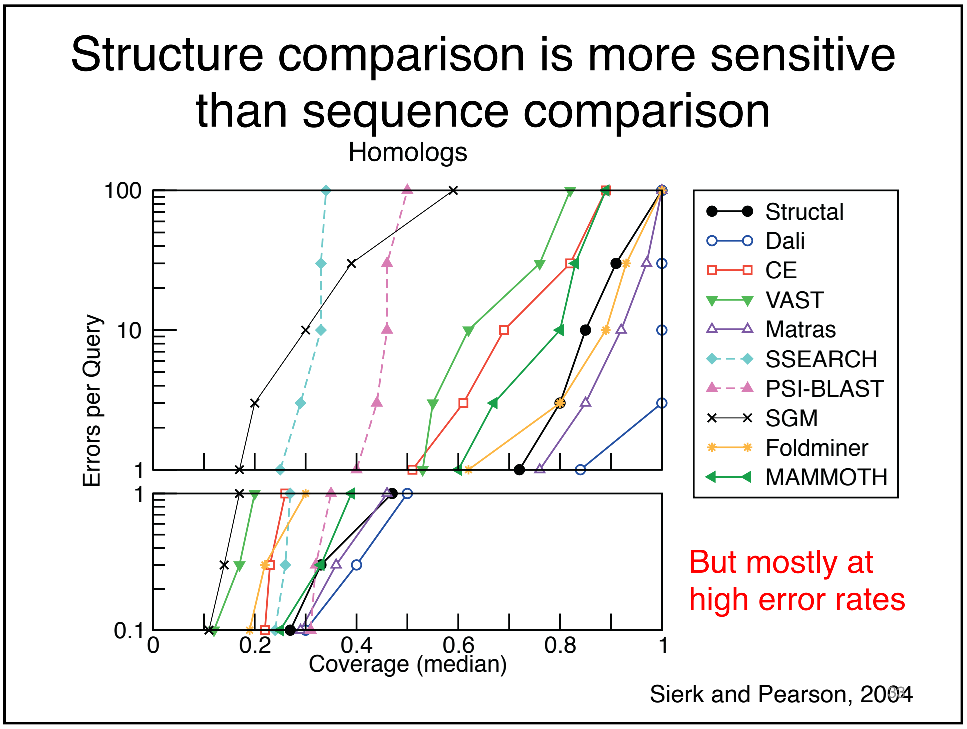
Human enzymes in other organisms (DNA vs protein comparison)



100 protein and mRNA sequences from human enzymes were compared to complete protein and mRNA sets from the indicated organisms.

32

32

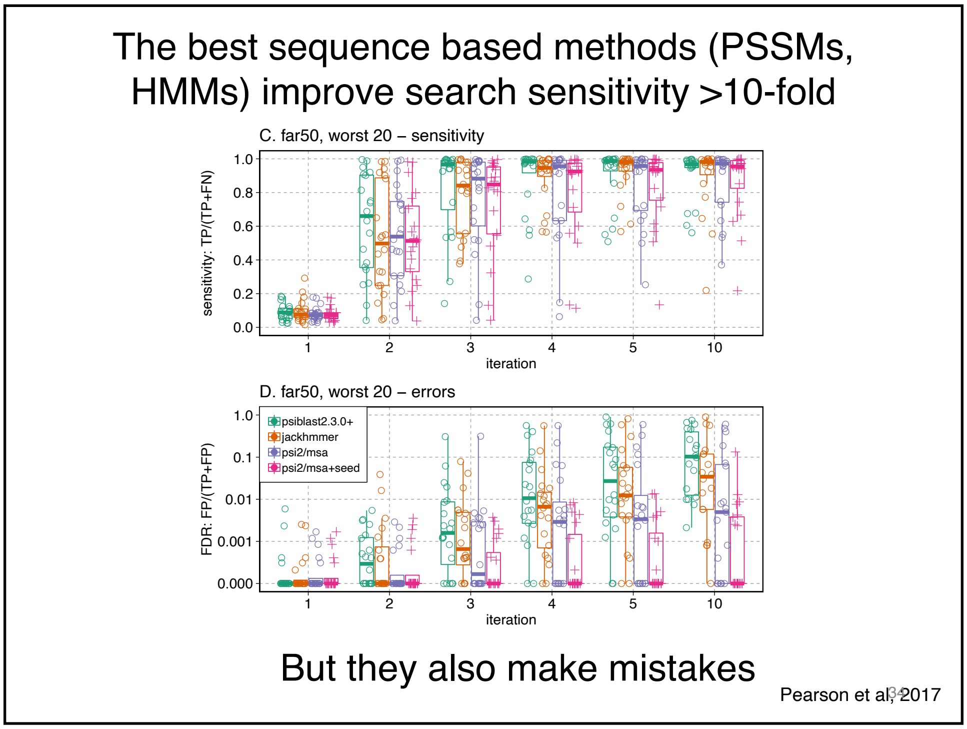


Structure comparison is more sensitive than sequence comparison

But mostly at high error rates

Sierk and Pearson, 2004

33



The best sequence based methods (PSSMs, HMMs) improve search sensitivity >10-fold

But they also make mistakes

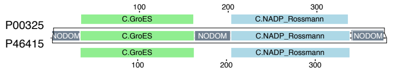
Pearson et al, 2017

34

Sensitive comparison methods miss homologs

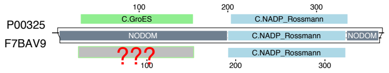
A. ADH1B_HUMAN (P00325) v. ADHX_DROME (P46415)

```
>>sp|P46415|ADHX_DROME Alcohol dehydrogenase class-3 OS=Drosophila melanogas  (379 aa)
qRegion:  34-161: 35-163 : score=379; bits=152.5; Id=0.532; Q=417.8 : C.GroES
qRegion: 203-335:205-339 : score=338; bits=136.0; Id=0.481; Q=368.4 : C.NADP_Rossmann
 Region:   2-33 :   3-34 : score=149; bits= 59.9; Id=0.750; Q=141.9 : NODOM
 Region:  34-161: 35-163 : score=379; bits=152.5; Id=0.532; Q=417.8 : C.GroES
 Region: 162-202:164-204 : score=179; bits= 72.0; Id=0.707; Q=178.6 : NODOM
 Region: 203-335:205-339 : score=338; bits=136.0; Id=0.481; Q=368.4 : C.NADP_Rossmann
 Region: 349-375:353-379 : score= 96; bits= 38.6; Id=0.519; Q= 78.6 : NODOM
Smith-Waterman score: 1161; 53.3% identity (74.9% similar) in379 aa overlap (2-375:3-379)
```



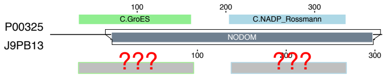
B. ADH1B_HUMAN (P00325) v. F7BAV9_MONDO (F7BAV9)

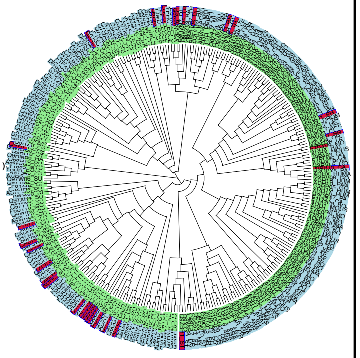
```
>>tr|F7BAV9|F7BAV9_MONDO Uncharacterized protein OS=Monodelphis domestica GN  (363 aa)
Smith-Waterman score: 1301; 57.5% identity (74.8% similar) in 381 aa overlap (7-373:1-361)
qRegion:  34-161: 28-149 : score=280; bits=113.1; Id=0.537; Q=301.6 :  C.GroES
qRegion: 203-335:191-323 : score=545; bits=220.1; Id=0.669; Q=620.5 :  C.NADP_Rossmann
 Region:   7-202:  1-190 : score=610; bits=246.4; Id=0.619; Q=697.7 :  NODOM
 Region: 203-336:191-324 : score=548; bits=221.3; Id=0.664; Q=624.3 :  C.NADP_Rossmann
 Region: 337-373:325-361 : score=143; bits= 57.8; Id=0.568; Q=135.8 :  NODOM
```



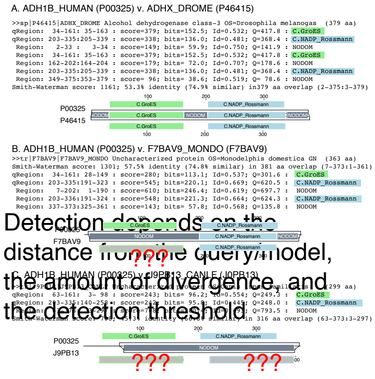
C. ADH1B_HUMAN (P00325) v. J9PB13_CANLF (J0PB13)

```
>>tr|J9PB13|J9PB13_CANLF Uncharacterized protein OS=Canis lupus familiaris    (299 aa)
qRegion:  63-161:  3- 98 : score=243; bits= 96.2; Id=0.554; Q=249.3 :  C.GroES
qRegion: 203-335:140-258 : score=242; bits= 95.8; Id=0.445; Q=248.0 :  C.NADP_Rossmann
 Region:  63-373:  3-297 : score=708; bits=280.4; Id=0.493; Q=793.5 :  NODOM
Smith-Waterman score: 708; 45.3% identity (66.8% similar) in 316 aa overlap (63-373:3-297)
```





Model-based methods miss homologs



A. ADH1B_HUMAN (P00325) v. ADHX_DROME (P46415)

```
>>sp|P46415|ADHX_DROME Alcohol dehydrogenase class-3 OS=Drosophila melanogas  (379 aa)
qRegion:  34-161: 35-163 : score=379; bits=152.5; Id=0.532; Q=417.8 : C.GroES
qRegion: 203-335:205-339 : score=338; bits=136.0; Id=0.481; Q=368.4 : C.NADP_Rossmann
 Region:   2-33 :   3-34 : score=149; bits= 59.9; Id=0.750; Q=141.9 : NODOM
 Region:  34-161: 35-163 : score=379; bits=152.5; Id=0.532; Q=417.8 : C.GroES
 Region: 162-202:164-204 : score=179; bits= 72.0; Id=0.707; Q=178.6 : NODOM
 Region: 203-335:205-339 : score=338; bits=136.0; Id=0.481; Q=368.4 : C.NADP_Rossmann
 Region: 349-375:353-379 : score= 96; bits= 38.6; Id=0.519; Q= 78.6 : NODOM
Smith-Waterman score: 1161; 53.3% identity (74.9% similar) in379 aa overlap (2-375:3-379)
```

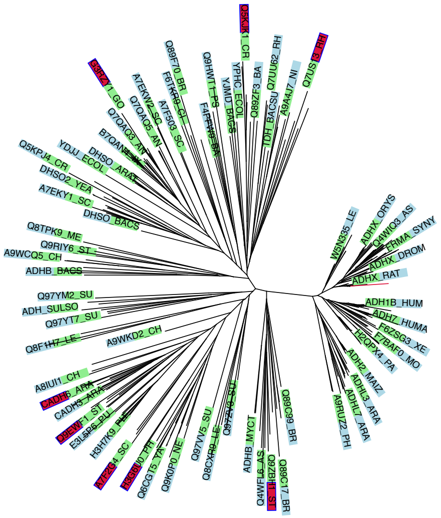
B. ADH1B_HUMAN (P00325) v. F7BAV9_MONDO (F7BAV9)

```
>>tr|F7BAV9|F7BAV9_MONDO Uncharacterized protein OS=Monodelphis domestica GN  (363 aa)
Smith-Waterman score: 1301; 57.5% identity (74.8% similar) in 381 aa overlap (7-373:1-361)
qRegion:  34-161: 28-149 : score=280; bits=113.1; Id=0.537; Q=301.6 :  C.GroES
qRegion: 203-335:191-323 : score=545; bits=220.1; Id=0.669; Q=620.5 :  C.NADP_Rossmann
 Region:   7-202:  1-190 : score=610; bits=246.4; Id=0.619; Q=697.7 :  NODOM
 Region: 203-336:191-324 : score=548; bits=221.3; Id=0.664; Q=624.3 :  C.NADP_Rossmann
 Region: 337-373:325-361 : score=143; bits= 57.8; Id=0.568; Q=135.8 :  NODOM
```

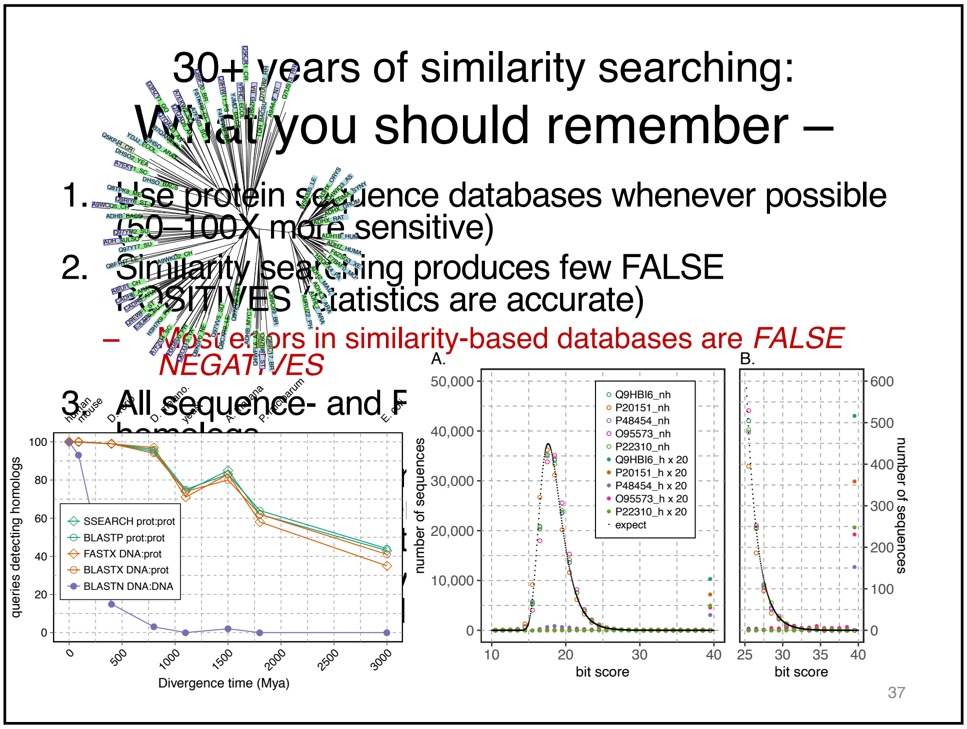
C. ADH1B_HUMAN (P00325) v. J9PB13_CANLF (J0PB13)

```
>>tr|J9PB13|J9PB13_CANLF Uncharacterized protein OS=Canis lupus familiaris    (299 aa)
qRegion:  63-161:  3- 98 : score=243; bits= 96.2; Id=0.554; Q=249.3 :  C.GroES
qRegion: 203-335:140-258 : score=242; bits= 95.8; Id=0.445; Q=248.0 :  C.NADP_Rossmann
 Region:  63-373:  3-297 : score=708; bits=280.4; Id=0.493; Q=793.5 :  NODOM
Smith-Waterman score: 708; 45.3% identity (66.8% similar) in 316 aa overlap (63-373:3-297)
```

Detection depends on the distance from the query/model, the amount of divergence, and the detection threshold



36



30+ years of similarity searching: What you should remember –

1. Use protein sequence databases whenever possible (50–100X more sensitive)
2. Similarity searching produces few FALSE POSITIVES (statistics are accurate)
   - Most errors in similarity-based databases are *FALSE NEGATIVES*
3. All sequence- and F
   homologs

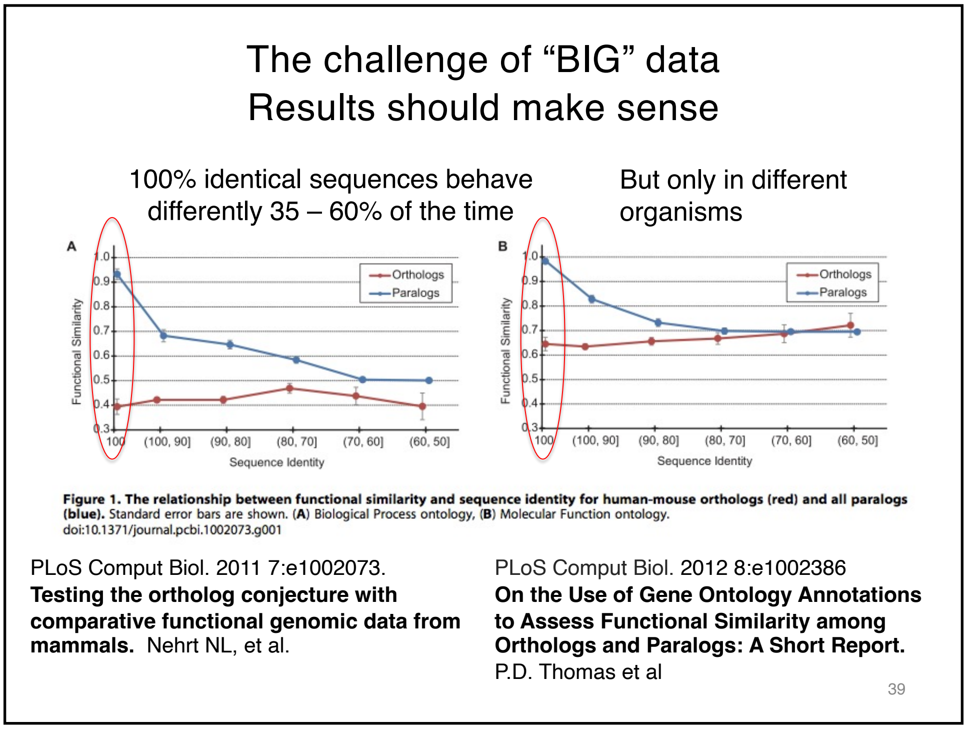
37

40+ years of similarity searching

- Lessons – what have we learned?
  - Darwin was right (we're all related)
  - life has been complex for a very long time (almost since the beginning!!)
  - very few constraints on protein sequences
- Challenges – what do we miss?
  - as protein space is sampled more densely, sequence comparison should become as sensitive as structure comparison
    - accurate alignments
    - accurate statistics
    - Hybrid PSSM/HMM – pairwise strategies
  - linking structure to function
    - accurate/robust measures of functional similarity
    - link functional information to alignments

38

38



The challenge of “BIG” data
Results should make sense

100% identical sequences behave differently 35 – 60% of the time

But only in different organisms

Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue). Standard error bars are shown. (**A**) Biological Process ontology, (**B**) Molecular Function ontology.
doi:10.1371/journal.pcbi.1002073.g001

PLoS Comput Biol. 2011 7:e1002073.
Testing the ortholog conjecture with comparative functional genomic data from mammals. Nehrt NL, et al.

PLoS Comput Biol. 2012 8:e1002386
On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report.
P.D. Thomas et al