Protein Evolution and Similarity Searching
Establishing Homology

Biol4230 Tues, Jan 23, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Goals of today's lecture:
- a quick overview of protein structure
- why search for homologs?
- understand why and how homology is inferred; the meaning of "expectation value"
- significance => homology, but no-significance ≠> non-homology
- understand sequence similarity, and why protein comparison is more sensitive than DNA sequence comparison

To learn more:

- Pevsner, Ch. 3
- Recombinant DNA, Ch. 12
- Pearson, (2000) "Protein Evolution and Sequence Comparison" ISMB2000 tutorial (collab)
- Koonin and Galperin (2003) Sequence – Evolution – Function www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=sef Chapter 4, section 4.2, Principles of sequence similarity searches (collab)
- Doolittle (1981) "Similar amino acid sequences: Chance or common ancestry" Science 214:149-159
- Computer exercises
  http://fasta.bioch.virginia.edu/biol4230/blast_demo.html

Homology, similarity, and protein structure

- Central dogma: DNA ➔ RNA ➔ proteins
- Mutations and rearrangements in DNA cause changes in protein structure
- Genome sequences (DNA) determine protein *sequence*
- Protein *sequence* determines protein *structure*
  - we cannot (accurately) predict structure from sequence
- Protein *structure* determines protein *function*
  - we cannot (accurately) predict function from structure
- Biologists (and physicians) care about *function*

From *sequence* to *structure*

DNA

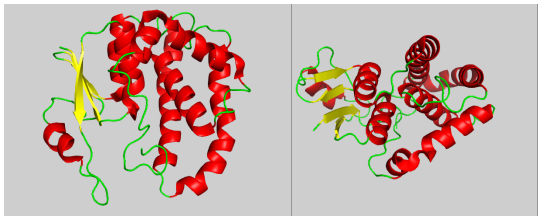
mRNA

```
LOCUS       NM_000561              1266 bp    mRNA    linear   PRI 25-MAY-2014
DEFINITION  Homo sapiens glutathione S-transferase mu 1 (GSTM1), transcript
            variant 1, mRNA.
ACCESSION   NM_000561
```

protein (sequence)

```
>ref|NP_000552.2| GSTM1 (human)
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTM
GDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKI
TQSNAILCYIARKHNLCGETEEEKIRVDILENQTM
DNHMQLGMICYNPEFEKLKPKYLEELPEKLKLYSE
FLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCL
DAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFS
KMAVWGNK
```

protein (structure, 1XW6)

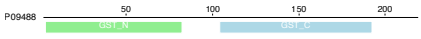


From *sequence* to *structure*

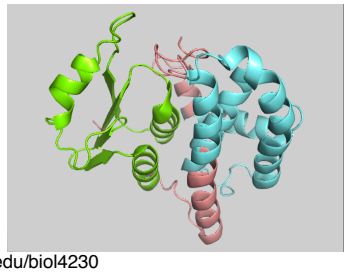
- Protein 3-D structures contain simpler, regularly repeating patterns of H-bonding called secondary structure:
  - alpha-helices
  - beta-strands (beta-sheets)
- Many proteins are build from evolutionarily mobile (structurally compact) domains (modules)

protein (sequence)

```
>ref|NP_000552.2| GSTM1 (human)
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTM
GDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKI
TQSNAILCYIARKHNLCGETEEEKIRVDILENQTM
DNHMQLGMICYNPEFEKLKPKYLEELPEKLKLYSE
FLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCL
DAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFS
KMAVWGNK
```



protein (structure, 1XW6)



Why do we search? How well do we search?

- Why search?
  - identify "similar" proteins (similar sequence?, similar structure?, similar function?)
  - what level of *sequence* similarity guarantees *structural* or *functional* similarity?
- How well do we search?
  - sensitivity: do we find all similar structures? functions?
  - specificity (selectivity): do all sequences this similar have similar structure? function?

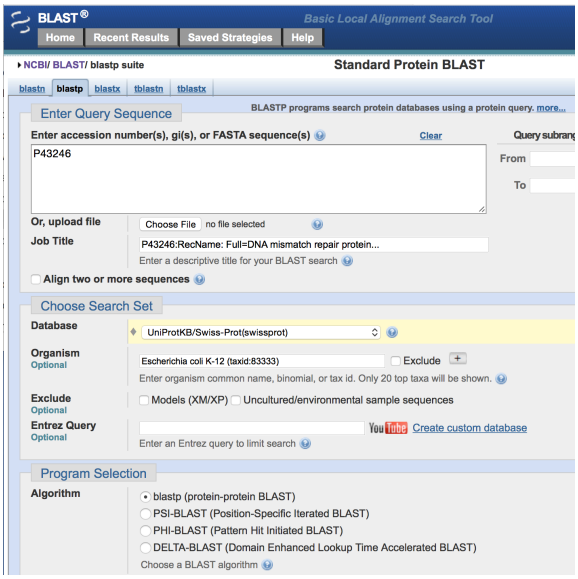
Is there a one-to-one mapping from sequence to structure? *yes*
Is there a one-to-one mapping from structure to function? *no*

Homologous proteins (proteins that evolved from a common ancestor) *always* have similar structures, and *sometimes* have similar functions.

Why do we search?

- (1993 – individual genes) Hereditary non-polyposis colon cancer (HNPCC). Is MSH2 related to an existing gene with known function?
  - if related, is it likely to have the same function?
- (2015 – whole genomes) I've isolated a new bacteria that makes a revolutionary antibiotic
  - which bacterial genes produce the antibiotic?
  - are those genes found in other bacteria?

(1993) MSH2 homolog in E. coli?



(1993) MSH2 homolog in E. coli?

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=DNA mismatch repair protein MutS [Escherichia coli K-12]	270	270	58%	1e-77	33%	P23909.1
RecName: Full=Uncharacterized protein YcgG [Escherichia coli K-12]	30.8	30.8	20%	0.54	22%	P75995.2
RecName: Full=Protein sirB1 [Escherichia coli O157:H7]	28.5	28.5	4%	2.3	39%	P0AGM5.1
RecName: Full=tRNA modification GTPase MnmE [Escherichia coli K-12]	28.5	28.5	5%	2.8	38%	P25522.3
RecName: Full=DNA polymerase I; Short=POL I [Escherichia coli K-12]	28.1	28.1	7%	4.1	29%	P00582.1
RecName: Full=Putative inner membrane metabolite transport protein YdfJ [Escherichia coli K-12]	27.7	27.7	15%	4.4	29%	P77228.1
RecName: Full=Stringent starvation protein A [Escherichia coli O157:H7]	26.6	26.6	4%	7.9	35%	P0ACA3.2
RecName: Full=Cell division protein ZapD; AltName: Full=Z ring-associated protein D [Escherichia	26.6	26.6	4%	9.1	32%	P36680.2

Which of these proteins have the same structure
Which have the same function?

(1993) MSH2 homolog in E. coli?

Download | GenPept | Graphics

RecName: Full=DNA mismatch repair protein MutS [Escherichia coli K-12]

Sequence ID: sp|P23909.1|MUTS_ECOLI Length: 853 Number of Matches: 1

See 2 more title(s)

Range 1: 269 to 797 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
270 bits(691)	1e-77	Compositional matrix adjust.	183/560(33%)	284/560(50%)	40/560(7%)

```
Query  302  LDIAAVRALNLFQ---GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEE  358
            +D A  R L + Q   G  E+T     LA++L+   TP G R++ +W+  P+ D   + E
Sbjct  269  MDAATRRNLEITQNLAGGAENT-----LASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLE  323

Query  359  RLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAKKFQRQAANLQDCYRLYQGINQLPNVI  418
            R   + A      +    LQ  +LR+  DL R+  +   + A  +D  R+      QLP +
Sbjct  324  RQQTIGAL---QDFTAGLQP-VLRQVGDLERILARLALRTARPRDLARMRHAFQQLPELR  379

Query  419  QALEKHEGKHQKLLLAVFVTPLTDLRSDFSKFQEMIETTLDMDQVENHEFLVK------P  472
              LE           V   P+  LR     +F E+     L+    ++     LV+
Sbjct  380  AQLE-----------TVDSAPVQALREKMGEFAEL-RDLLERAIIDTPPVLVRDGGVIAS  427

Query  473  SFDPNLSELREIMNDLEKKMQSTLISAARDLGLDPGKQIKLDSSAQFGYYFRVTCKEEKV  532
             ++  L E R + +      ++    +      GLD    +K+  +A  GYY +++  +  +
Sbjct  428  GYNEELDEWRALADGATDYLERLEVRERERTGLD---TLKVGFNAVHGYYIQISRGQSHL  484

Query  533  LRNNKNFSTVDIQKNGVKFTNSKLTSLNEEYTKNKTEYEEAQDAIVKEIVNISSGYVEPM  592
                 N+      KN  ++    +L     ++K +     +  + +E+ ++    ++E +
Sbjct  485  --APINYMRRQTLKNAERYIIPELKEYEDKVLTSKGKALALEKQLYEELFDLLLPHLEAL  542

Query  593  QTLNDVLAQLDAVVSFAHVSNGAPVPYVRPAILEKGQGRIILKASRHACVEVQDEIAFIP  652
            Q     LA+LD +V+ A       + Y  P  ++K   RI     RH  VE      FI
Sbjct  543  QQSASALAELDVLVNLAE--RAYTLNYTCPTFIDKPGIRI--TEGRHPVVEQVLNEPFIA  598

Query  653  NDVYFEKDKQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESAEVSIVDCILARV  712
            N +     ++M  IITGPNMGGKSTY+RQT +I LMA IG +VP +  E+  +D I  RV
Sbjct  599  NPLNLSPQRRML-IITGPNMGGKSTYMRQTALIALMAYIGSYVPAQKVEIGPIDRIFTRV  657

Query  713  GAGDSQLKGVSTFMAEMLETASILRSATKDSLIIIDELGRGTSTYDGFGLAWAISEYIAT  772
            GA D    G STFM EM ETA+IL +AT+ SL+++DE+GRGTSTYDG  LAWA +E +A
Sbjct  658  GAADDLASGRSTFMVEMTETANILHNATEYSLVLMDEIGRGTSTYDGLSLAWACAENLAN  717

Query  773  KIGAFCMFATHFHELTALANQIPTVNNLHVTALTTEETLTMLYQVKKGVCDQSFGIHVAE  832
            KI A  +FATH+ ELT L  ++  V N+H+ AL   +T+  ++ V+ G    +S+G+ VA
Sbjct  718  KIKALTLFATHYFELTQLPEKMEGVANVHLDALEHGDTIAFMHSVQDGAASKSYGLAVAA  777

Query  833  LANFPKHVIECAKQKALELE  852
            LA  PK VI+ A+QK  ELE
Sbjct  778  LAGVPKEVIKRARQKLRELE  797
```

Are all parts of the alignment equally similar?

(2015) what is YCGG_ECOLI?

Probably a di-GMP-phosphodiesterase

```
The best scores are:                                      s-w bits E(459565) %_id  %_sim
alen
sp|P75995.2|YCGG_ECOLI Uncharacterized protein YcgG [Es ( 507) 2518 985.5       0 1.000 1.000  507
sp|P21514.2|YAHA_ECOLI Cyclic di-GMP phosphodiesterase  ( 362)  512 204.4 2.5e-51 0.442 0.725  251
sp|P76446.1|RTN_ECOLI Protein Rtn [Escherichia coli K-1 ( 518)  443 177.4 4.7e-43 0.297 0.631  444
sp|P76261.2|ADRB_ECOLI Putative cyclic-di-GMP phosphodi ( 532)  409 164.2 4.7e-39 0.277 0.579  523
sp|P76129.4|DOSP_ECOLI Oxygen sensor protein DosP; Dire ( 799)  370 148.9 2.8e-34 0.349 0.689  238
sp|Q9I310.1|Y1727_PSEAE Uncharacterized signaling prote ( 685)  362 145.8   2e-33 0.353 0.689  235
sp|P55552.1|Y4LL_RHISN Uncharacterized protein y4lL [Si ( 827)  359 144.6 5.7e-33 0.344 0.652  250
sp|P32701.2|YJCC_ECOLI Putative cyclic-di-GMP phosphodi ( 528)  351 141.6 2.9e-32 0.314 0.675  277
sp|Q9HYT3.1|Y3311_PSEAE Uncharacterized signaling prote ( 783)  350 141.1   6e-32 0.343 0.695  239
sp|Q55434.1|PHY2_SYNY3 Phytochrome-like protein cph2; B (1276)  346 139.4 3.1e-31 0.367 0.676  256
sp|Q9ABX9.1|Y091_CAUCR Uncharacterized signaling protei ( 809)  337 136.1 2.1e-30 0.380 0.662  237
sp|P77334.1|GMR_ECOLI Cyclic di-GMP phosphodiesterase G ( 661)  332 134.2 6.4e-30 0.340 0.685  235
sp|O34311.2|YKOW_BACSU Signaling protein YkoW [Bacillus ( 800)  310 125.5   3e-27 0.307 0.641  251
sp|P64830.1|Y1392_MYCBO Uncharacterized protein Mb1392c ( 307)  297 120.7 3.3e-26 0.343 0.628  239
sp|P37649.3|YHJK_ECOLI Protein YhjK [Escherichia coli K ( 662)  283 115.1 3.5e-24 0.313 0.647  249
sp|P75800.1|YLIE_ECOLI Putative cyclic di-GMP phosphodi ( 782)  271 110.4 1.1e-22 0.293 0.636  239
sp|Q8EJM6.1|PDEB_SHEON Cyclic di-GMP phosphodiesterase  ( 856)  267 108.8 3.6e-22 0.294 0.664  235
sp|Q9KU26.1|MBAA_VIBCH Biofilm architecture maintenance ( 791)  242 99.1 2.8e-19 0.252 0.646  254
sp|Q9JMT8.1|YUAB_ECOLI Uncharacterized HTH-type transcr ( 353)  237 97.3 4.2e-19 0.249 0.635  241
sp|P77172.1|YFGF_ECOLI Cyclic di-GMP phosphodiesterase  ( 747)  240 98.3 4.5e-19 0.280 0.628  261
sp|Q9KVL2.1|CDPA_VIBCH Cyclic di-GMP phosphodiesterase  ( 829)  166 69.5 2.4e-10 0.250 0.621  232
sp|O35014.1|YKUI_BACSU Uncharacterized EAL-domain conta ( 407)  137 58.3 2.6e-07 0.246 0.560  232
sp|P37646.3|YHJH_ECOLI Cyclic di-GMP phosphodiesterase  ( 255)  125 53.8 3.9e-06 0.296 0.653   98
sp|P14203.1|YUXH_BACSU Uncharacterized protein YuxH [Ba ( 409)  103 45.1  0.0026 0.278 0.538  169
sp|P75990.1|YCGF_ECOLI Blue light- and temperature-regu ( 403)   93 41.2   0.037 0.247 0.614  166
sp|P13518.2|CSRD_ECOLI RNase E specificity factor CsrD; ( 646)   92 40.7   0.085 0.198 0.563  222
```


Protein Evolution and Sequence Similarity

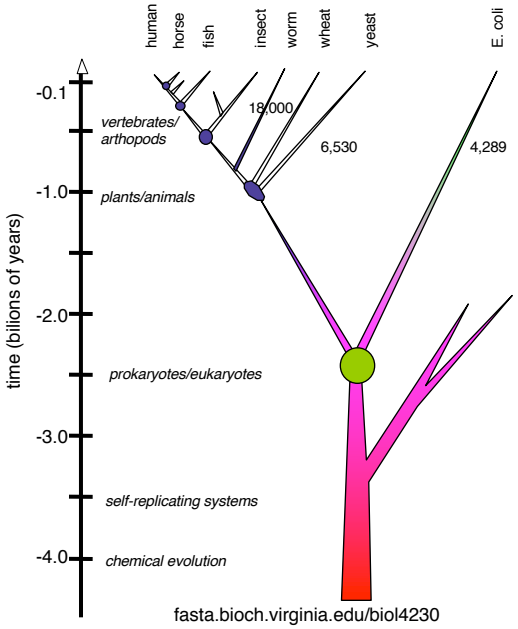
Similarity Searching I

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison

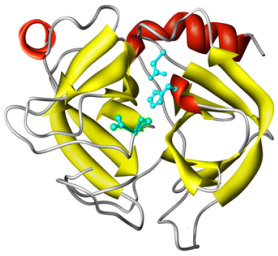
Similarity Searching II

- Alignment algorithms
- What are the problems?
  - missed homologs (false negatives, sensitivity)
  - ?false positives? (specificity)
- What do the statistics mean?
- How can we change behavior (scoring matrices)

Homologues share a common ancestor

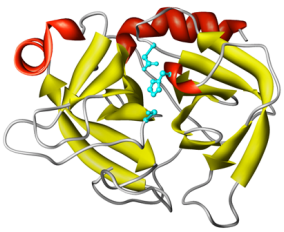


When do we infer homology?

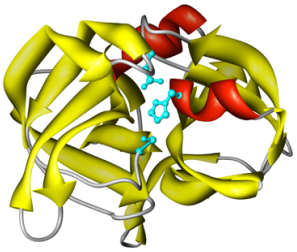


Bovine trypsin (5ptp)
Structure: E()< 10^{-23}; RMSD 0.0 A
Sequence: E()< 10^{-84} 100% 223/223

Homology <=> structural similarity
? sequence similarity

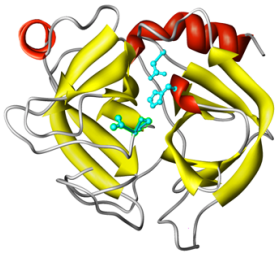


S. griseus trypsin (1sgt)
E()< 10^{-14} RMSD 1.6 A
E()< 10^{-19} 36%; 226/223



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

When can we infer non-homology?

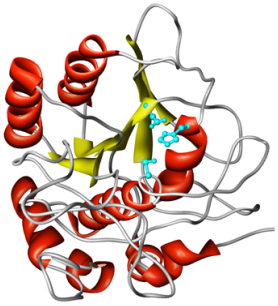


Bovine trypsin (5ptp)

Structure: $E()<10^{-23}$
RMSD 0.0 A

Sequence: $E()<10^{-84}$
100% 223/223

Non-homologous proteins have different structures



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

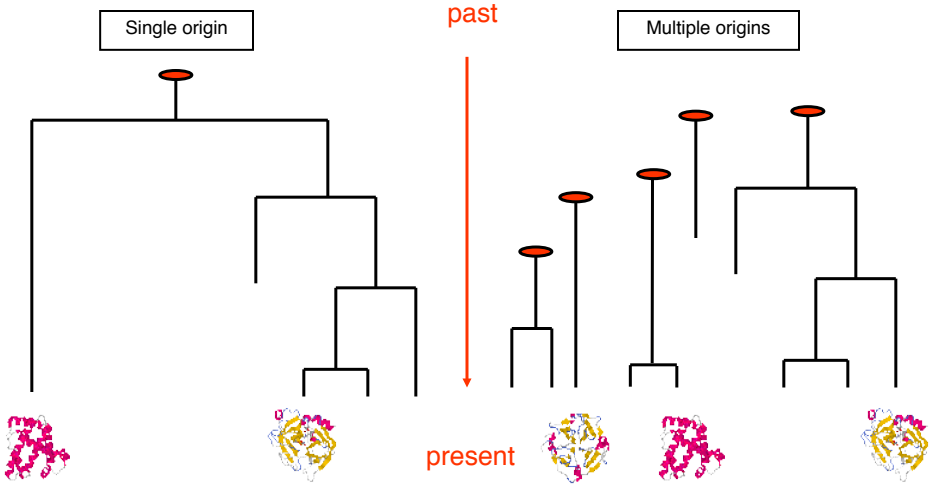
Cytochrome c4 (1etp)
E() > 100
E()<5.5; 23% 171/190

Homology is confusing I: Homology defined Three(?) Ways

- Proteins/genes/DNA that share a common ancestor
- Specific positions/columns in a multiple sequence alignment that have a 1:1 relationship over evolutionary history
  - sequences are *50% homologous* ???
- Specific (morphological/functional) characters that share a recent divergence (clade)
  - bird/bat/butterfly wings are/are not homologous

Homology is confusing II: Are All Sequences Homologous?

No Homology without excess similarity



Homology from sequence similarity

- Sequences are inferred to share a common ancestor based on statistically significant *excess* similarity. Any evidence of *excess* similarity can be used to infer homology
- Lack of sequence evidence *cannot* be used to infer non-homology.
  - Proteins with different structures are non-homologous
- There are always two alternative hypotheses: homology (common ancestry), or independence – one must weigh the evidence for each hypothesis (independence is the *null* hypothesis).

E. coli proteins vs Human – Ancient Protein Domains

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

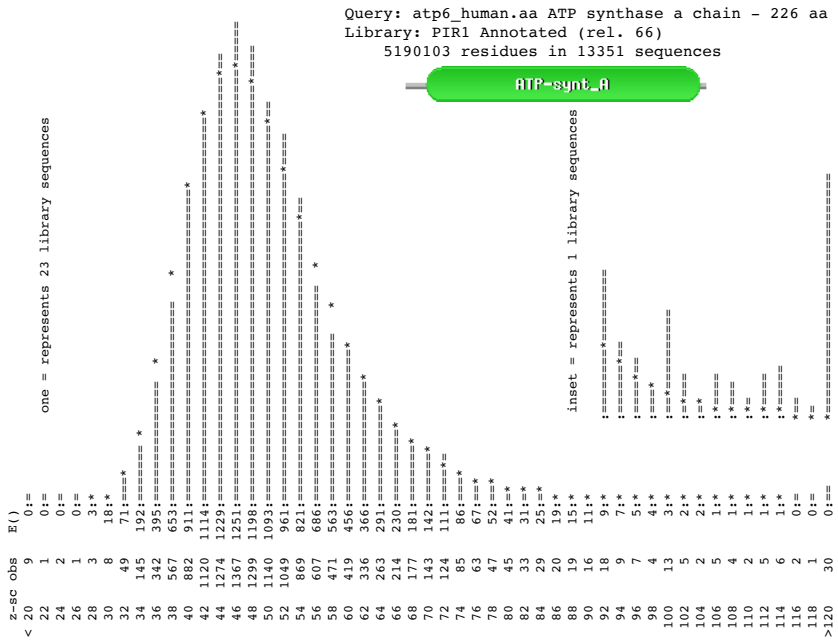
Protein Evolution and Sequence Similarity

Similarity Searching I

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison

Similarity Searching II

- Alignment algorithms
- What are the problems?
  - missed homologs (false negatives, sensitivity)
  - ?false positives? (specificity)
- What do the statistics mean?
- How can we change behavior (scoring matrices)



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

```
                    Query: atp6_human.aa ATP synthase a chain - 226 aa
                    Library: 5190103 residues in 13351 sequences

  The best scores are:                          ( len)  s-w bits E(13351) %_id  %_sim  alen
  sp|P00846|ATP6_HUMAN ATP synthase a chain (AT ( 226) 1400 325.8 5.8e-90 1.000 1.000  226
  sp|P00847|ATP6_BOVIN ATP synthase a chain (AT ( 226) 1157 270.5 2.5e-73 0.779 0.951  226
  sp|P00848|ATP6_MOUSE ATP synthase a chain (AT ( 226) 1118 261.7 1.2e-70 0.757 0.916  226
  sp|P00849|ATP6_XENLA ATP synthase a chain (AT ( 226)  745 176.8 4.0e-45 0.533 0.847  229
  sp|P00851|ATP6_DROYA ATP synthase a chain (AT ( 224)  473 115.0 1.7e-26 0.378 0.721  222
  sp|P00854|ATP6_YEAST ATP synthase a chain pre ( 259)  428 104.7 2.3e-23 0.353 0.694  232
  sp|P00852|ATP6_EMENI ATP synthase a chain pre ( 256)  365  90.4 4.8e-19 0.304 0.691  230
  sp|P14862|ATP6_COCHE ATP synthase a chain (AT ( 257)  353  87.7 3.2e-18 0.313 0.650  214
  sp|P68526|ATP6_TRITI ATP synthase a chain (AT ( 386)  309  77.6 5.1e-15 0.289 0.651  235
  sp|P05499|ATP6_TOBAC ATP synthase a chain (AT ( 395)  309  77.6 5.2e-15 0.283 0.635  233
  sp|P07925|ATP6_MAIZE ATP synthase a chain (AT ( 291)  283  71.7 2.3e-13 0.311 0.667  180
  sp|P0AB98|ATP6_ECOLI ATP synthase a chain (AT ( 271)  178  47.9 3.2e-06 0.233 0.585  236
  sp|P0C2Y5|ATPI_ORYSA Chloroplast ATP synth (A ( 247)  144  40.1 0.00062 0.242 0.580  231
  sp|P06452|ATPI_PEA Chloroplast ATP synthase a ( 247)  143  39.9 0.00072 0.250 0.586  232
  sp|P27178|ATP6_SYNY3 ATP synthase a chain (AT ( 276)  142  39.7 0.00095 0.265 0.571  170
  sp|P06451|ATPI_SPIOL Chloroplast ATP synthase ( 247)  138  38.8  0.0016 0.242 0.580  231
  sp|P08444|ATP6_SYNP6 ATP synthase a chain (AT ( 261)  127  36.3  0.0095 0.263 0.557  167
  sp|P69371|ATPI_ATRBE Chloroplast ATP synthase ( 247)  126  36.0  0.01   0.221 0.571  231
  sp|P06289|ATPI_MARPO Chloroplast ATP synthase ( 248)  126  36.0  0.011  0.240 0.575  167
  sp|P30391|ATPI_EUGGR Chloroplast ATP synthase ( 251)  123  35.4  0.017  0.257 0.579  214

  sp|P19568|TLCA_RICPR ADP,ATP carrier protein  ( 498)  122  35.0  0.043  0.243 0.579  152
  sp|P24966|CYB_TAYTA Cytochrome b              ( 379)  113  33.0  0.13   0.234 0.532  158
  sp|P03892|NU2M_BOVIN NADH-ubiquinone oxidored ( 347)  107  31.7  0.31   0.261 0.479  211
  sp|P68092|CYB_STEAT Cytochrome b              ( 379)  104  31.0  0.54   0.277 0.547  137
  sp|P03891|NU2M_HUMAN NADH-ubiquinone oxidored ( 347)  103  30.8  0.58   0.201 0.537  149
  sp|P00156|CYB_HUMAN Cytochrome b              ( 380)  102  30.5  0.74   0.268 0.585  205
  sp|P15993|AROP_ECOLI Aromatic amino acid tr   ( 457)  103  30.7  0.78   0.234 0.622  111
  sp|P24965|CYB_TRANA Cytochrome b              ( 379)  101  30.3  0.87   0.234 0.563  158
  sp|P29631|CYB_POMTE Cytochrome b              ( 308)   99  29.9  0.95   0.274 0.584  113
  sp|P24953|CYB_CAPHI Cytochrome b              ( 379)   99  29.8  1.2    0.236 0.564  140
```

fasta.bioch.virginia.edu/biol4230 23


```
>sp|P00846|ATP6_HUMAN ATP synthase subunit a; F-ATPase protein 6;      Length=226
 vs:
>sp|P0AB98|ATP6_ECOLI ATP synthase subunit a; ATP synthase F0 subunit; Length=271
 Score = 47.9 bits (178),  Expect = 3e-06
 Identities = 55/199 (27%), Positives = 113/199 (56%), Gaps = 37/199 (18%)

Query  8    SFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSLML 72
            S     +LGL   ++++LF  +    +   +  ++ T  + +I + +  +  M++ K +  + +
Sbjct  45   SMFFSVVLGL---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSVKDMYHGKSKLIAPLA 105

Query  73   VSLIIFIATTNLLGLLP---------HSF-------TPTTQLSMNLAMAIPLWAGTVIMGFRSKI 121
            +++ +++   NL+ LLP         H +        P+  +++ L+MA+ ++   +++ F S
Sbjct  106  LTIFVWVFLMNLMDLLPIDLLPYIAEHVLGLPALRVVPSADVNVTLSMALGVF---ILILFYSIK 167

Query  122  KNALAHFLPQGTPTPL-----IPMLVIIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAM 181
               +  F  + T  P+     IP+ +I+E +SLL +P++L +RL  N+ AG L+  LI
Sbjct  168  MKGIGGFTKELTLQPFNHWAFIPVNLILEGVSLLSKPVSLGLRLFGNMYAGELIFILIAGLLPWW 232

Query  182  STINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYL 222
            S   L     IF ILI+          +QA++F +L  +YL
Sbjct  233  SQWILNVPWAIFHILIIT---------LQAFIFMVLTIVYL 264
```

fasta.bioch.virginia.edu/biol4230 24

The PAM250 matrix

```
Cys  12
Ser   0  2
Thr  -2  1  3
Pro  -1  1  0  6
Ala  -2  1  1  1  2
Gly  -3  1  0 -1  1  5
Asn  -4  1  0 -1  0  0  2
Asp  -5  0  0 -1  0  1  2  4
Glu  -5  0  0 -1  0  0  1  3  4
Gln  -5 -1 -1  0  0 -1  1  2  2  4
His  -3 -1 -1  0 -1 -2  2  1  1  3  6
Arg  -4  0 -1  0 -2 -3  0 -1 -1  1  2  6
Lys  -5  0  0 -1 -1 -2  1  0  0  1  0  3  5
Met  -5 -2 -1 -2 -1 -3 -2 -3 -2 -1 -2  0  0  6
Ile  -2 -1  0 -2 -1 -3 -2 -2 -2 -2 -2 -2 -2  2  5
Leu  -6 -3 -2 -3 -2 -4 -3 -4 -3 -2 -2 -3 -3  4  2  6
Val  -2 -1  0 -1  0 -1 -2 -2 -2 -2 -2 -2 -2  2  4  2  4
Phe  -4 -3 -3 -5 -4 -5 -4 -6 -5 -5 -2 -4 -5  0  1  2 -1  9
Tyr   0 -3 -3 -5 -3 -5 -2 -4 -4 -4  0 -4 -4 -2 -1 -1 -2  7 10
Trp  -8 -2 -5 -6 -6 -7 -4 -7 -7 -5 -3  2 -3 -4 -5 -2 -6  0  0 17
      C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

Scoring Matrix summary:
(1) Used to produce alignment score
(2) Identities always positive, but some (rare, conserved) more positive than others.
(3) Similar amino-acids also positive
(4) Most aligned pairs get negative scores


```
>sp|P00846|ATP6_HUMAN ATP synthase subunit a; F-ATPase protein 6  vs:
>sp|P30391|ATPI_EUGGR ATP synthase subunit a, chloroplastic; ATP synthase F0 sub
Length=251

 Score = 35.4 bits (123),  Expect = 0.02
 Identities = 55/182 (30%), Positives = 101/182 (55%), Gaps = 32/182 (17%)

Query  21   VLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTK-GRT----WSLMLVSLIIFIA 80
            +LII F  L I T+K L    +   +Q +I+L ++ +  +  T+ G      W   + ++ +FI
Sbjct  50   ILIIGF--LSIYTTKNL--TLVPANKQIFIELVTEFITDISKTQIGEKEYSKWVPYIGTMFLFIF 110

Query  81   TTNLLG-LLPHSFT--PTTQL---SMNLAMAIPLWAGTVIMGFRSKI-KNALAHFLPQGTPTPLI 138
             +N  G L+P      P  +L   + ++     L   T +  F + + K  L +F      PTP++
Sbjct  111  VSNWSGALIPWKIIELPNGELGAPTNDINTTAGLAILTSLAYFYAGLNKKGLTYFKKYVQPTPIL 175

Query  139  PMLVIIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILE 203
              + I+E  +   +P++L+ RL  NI A  L++ ++ S         +P  LI+   LI+L ++
Sbjct  176  LPINILEDFT---KPLSLSFRLFGNILADELVVAVLVSL--------VP--LIVPVPLIFLGLF- 226

Query  204  IAVALIQAYVFTLLVSLYL 222
               + IQA +F  L   Y+
Sbjct  227  --TSGIQALIFATLSGSYI 243
```

Query: atp6_human.aa ATP synthase a chain - 226 aa
Library: 5190103 residues in 13351 sequences

The best scores are:	(len)	s-w	bits	E(13351)	%_id	%_sim	alen
sp|P00846|ATP6_HUMAN ATP synthase a chain (AT	(226)	1400	325.8	5.8e-90	1.000	1.000	226
sp|P00847|ATP6_BOVIN ATP synthase a chain (AT	(226)	1157	270.5	2.5e-73	0.779	0.951	226
sp|P00848|ATP6_MOUSE ATP synthase a chain (AT	(226)	1118	261.7	1.2e-70	0.757	0.916	226
sp|P00849|ATP6_XENLA ATP synthase a chain (AT	(226)	745	176.8	4.0e-45	0.533	0.847	229
sp|P00851|ATP6_DROYA ATP synthase a chain (AT	(224)	473	115.0	1.7e-26	0.378	0.721	222
sp|P00854|ATP6_YEAST ATP synthase a chain pre	(259)	428	104.7	2.3e-23	0.353	0.694	232
sp|P00852|ATP6_EMENI ATP synthase a chain pre	(256)	365	90.4	4.8e-19	0.304	0.691	230
sp|P14862|ATP6_COCHE ATP synthase a chain (AT	(257)	353	87.7	3.2e-18	0.313	0.650	214
sp|P68526|ATP6_TRITI ATP synthase a chain (AT	(386)	309	77.6	5.1e-15	0.289	0.651	235
sp|P05499|ATP6_TOBAC ATP synthase a chain (AT	(395)	309	77.6	5.2e-15	0.283	0.635	233
sp|P07925|ATP6_MAIZE ATP synthase a chain (AT	(291)	283	71.7	2.3e-13	0.311	0.667	180
sp|P0AB98|ATP6_ECOLI ATP synthase a chain (AT	(271)	178	47.9	3.2e-06	0.233	0.585	236
sp|P0C2Y5|ATPI_ORYSA Chloroplast ATP synth (A	(247)	144	40.1	0.00062	0.242	0.580	231
sp|P06452|ATPI_PEA Chloroplast ATP synthase a	(247)	143	39.9	0.00072	0.250	0.586	232
sp|P27178|ATP6_SYNY3 ATP synthase a chain (AT	(276)	142	39.7	0.00095	0.265	0.571	170
sp|P06451|ATPI_SPIOL Chloroplast ATP synthase	(247)	138	38.8	0.0016	0.242	0.580	231
sp|P08444|ATP6_SYNP6 ATP synthase a chain (AT	(261)	127	36.3	0.0095	0.263	0.557	167
sp|P69371|ATPI_ATRBE Chloroplast ATP synthase	(247)	126	36.0	0.01	0.221	0.571	231
sp|P06289|ATPI_MARPO Chloroplast ATP synthase	(248)	126	36.0	0.011	0.240	0.575	167
sp|P30391|ATPI_EUGGR Chloroplast ATP synthase	(251)	123	35.4	0.017	0.257	0.579	214
sp|P19568|TLCA_RICPR ADP,ATP carrier protein	(498)	122	35.0	0.043	0.243	0.579	152
sp|P24966|CYB_TAYTA Cytochrome b	(379)	113	33.0	0.13	0.234	0.532	158
sp|P03892|NU2M_BOVIN NADH-ubiquinone oxidored	(347)	107	31.7	0.31	0.261	0.479	211
sp|P68092|CYB_STEAT Cytochrome b	(379)	104	31.0	0.54	0.277	0.547	137
sp|P03891|NU2M_HUMAN NADH-ubiquinone oxidored	(347)	103	30.8	0.58	0.201	0.537	149
sp|P00156|CYB_HUMAN Cytochrome b	(380)	102	30.5	0.74	0.268	0.585	205
sp|P15993|AROP_ECOLI Aromatic amino acid tr	(457)	103	30.7	0.78	0.234	0.622	111
sp|P24965|CYB_TRANA Cytochrome b	(379)	101	30.3	0.87	0.234	0.563	158
sp|P29631|CYB_POMTE Cytochrome b	(308)	99	29.9	0.95	0.274	0.584	113
sp|P24953|CYB_CAPHI Cytochrome b	(379)	99	29.8	1.2	0.236	0.564	140

fasta.bioch.virginia.edu/biol4230 27

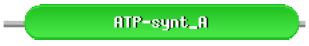
Query: atp6_ecoli.aa ATP synthase a - 271 aa
Library: 5190103 residues in 13351 sequences

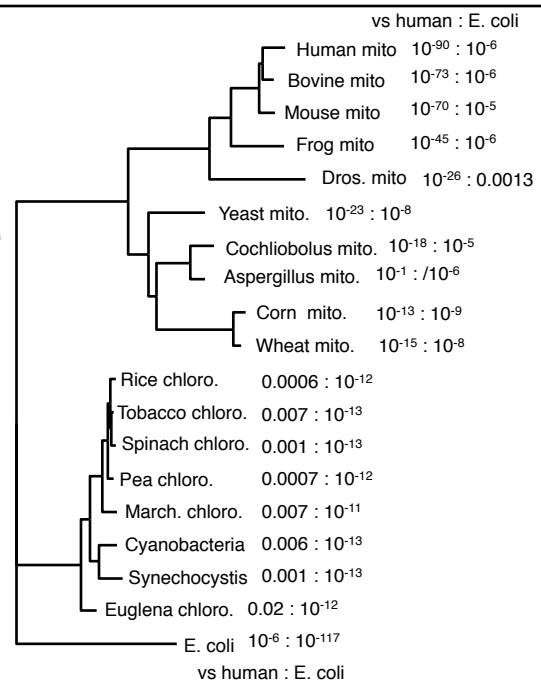
The best scores are:	(len)	s-w	bits	E(13351)	%_id	%_sim	alen
sp|P0AB98|ATP6_ECOLI ATP synthase a chain (AT	(271)	1774	416.8	3.e-117	1.000	1.000	271
sp|P06451|ATPI_SPIOL Chloroplast ATP synthase	(247)	274	70.4	5.8e-13	0.270	0.616	211
sp|P69371|ATPI_ATRBE Chloroplast ATP synthase	(247)	271	69.7	9.3e-13	0.270	0.607	211
sp|P08444|ATP6_SYNP6 ATP synthase a chain (AT	(261)	271	69.7	9.9e-13	0.267	0.600	240
sp|P06452|ATPI_PEA Chloroplast ATP synthase a	(247)	266	68.5	2.1e-12	0.274	0.614	223
sp|P30391|ATPI_EUGGR Chloroplast ATP synthase	(251)	265	68.3	2.5e-12	0.298	0.596	225
sp|P0C2Y5|ATPI_ORYSA Chloroplast ATP synthase	(247)	260	67.2	5.4e-12	0.259	0.603	239
sp|P27178|ATP6_SYNY3 ATP synthase a chain (AT	(276)	260	67.1	6.1e-12	0.264	0.578	258
sp|P06289|ATPI_MARPO Chloroplast ATP synthase	(248)	250	64.8	2.7e-11	0.261	0.621	211
sp|P07925|ATP6_MAIZE ATP synthase a chain (AT	(291)	215	56.7	8.7e-09	0.259	0.578	232
sp|P68526|ATP6_TRITI ATP synthase a chain (AT	(386)	209	55.3	3.1e-08	0.259	0.603	239
sp|P00854|ATP6_YEAST ATP synthase a chain pre	(259)	204	54.2	4.5e-08	0.235	0.578	277
sp|P05499|ATP6_TOBAC ATP synthase a chain (AT	(395)	189	50.7	7.8e-07	0.220	0.582	268
sp|P00846|ATP6_HUMAN ATP synthase a chain (AT	(226)	178	48.2	2.5e-06	0.237	0.589	236
sp|P00852|ATP6_EMENI ATP synthase a chain pre	(256)	178	48.2	2.8e-06	0.209	0.590	244
sp|P00849|ATP6_XENLA ATP synthase a chain (AT	(226)	173	47.1	5.5e-06	0.261	0.630	165
sp|P00847|ATP6_BOVIN ATP synthase a chain (AT	(226)	172	46.8	6.5e-06	0.233	0.581	236
sp|P14862|ATP6_COCHE ATP synthase a chain (AT	(257)	171	46.6	8.7e-06	0.204	0.608	265
sp|P00848|ATP6_MOUSE ATP synthase a chain (AT	(226)	166	45.5	1.7e-05	0.259	0.617	193
sp|P00851|ATP6_DROYA ATP synthase a chain (AT	(224)	139	39.2	0.0013	0.225	0.549	253
sp|P24962|CYB_STELO Cytochrome b	(379)	125	35.9	0.021	0.223	0.575	193
sp|P09716|US17_HCMVA Hypothetical protein HVL	(293)	109	32.3	0.21	0.260	0.565	131
sp|P68092|CYB_STEAT Cytochrome b	(379)	109	32.2	0.27	0.211	0.562	194
sp|P24960|CYB_ODOHE Cytochrome b	(379)	104	31.1	0.61	0.210	0.555	200
sp|P03887|NU1M_BOVIN NADH-ubiquinone oxidored	(318)	98	29.7	1.3	0.287	0.545	167
sp|P24992|CYB_ANTAM Cytochrome b	(379)	99	29.9	1.4	0.192	0.565	193

Similarity score (and significance) depends on the query perspective

fasta.bioch.virginia.edu/biol4230 28

Homology is Transitive (on domains)

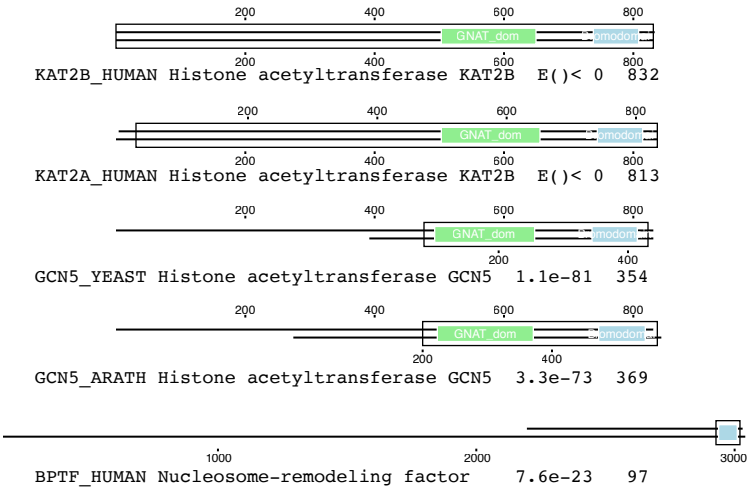




Homology and Domains – Histone acetyltransferase KAT2B

```
The best scores are:                                      s-w bits E(454402) %_id  %_sim  alen
KAT2B_HUMAN Histone acetyltransferase KAT2B ( 832) 3820 1456.       0 1.000 1.000  832
KAT2A_HUMAN Histone acetyltransferase KAT2A ( 837) 2747 1049.       0 0.721 0.870  813
GCN5_SCHPO Histone acetyltransferase gcn5   ( 454)  867 334.7   3e-90 0.483 0.768  354
GCN5_YEAST Histone acetyltransferase GCN5   ( 439)  792 306.2 1.1e-81 0.469 0.760  354
GCN5_ORYSJ Histone acetyltransferase GCN5   ( 511)  760 294.0 5.9e-78 0.436 0.755  376
GCN5_ARATH Histone acetyltransferase GCN5;  ( 568)  719 278.4 3.3e-73 0.434 0.740  369
BPTF_HUMAN Nucleosome-remodeling factor sub (3046)  286 113.6 7.6e-23 0.495 0.804   97
NU301_DROME Nucleosome-remodeling factor su (2669)  276 109.8 9.1e-22 0.511 0.819   94
CECR2_HUMAN Cat eye syndrome critical regio (1484)  232  93.2   5e-17 0.371 0.790  105
BRD4_HUMAN Bromodomain-containing protein 4 (1362)  214  86.4 5.2e-15 0.379 0.698  116
BRD4_MOUSE Bromodomain-containing protein 4 (1400)  214  86.4 5.3e-15 0.379 0.698  116
BAZ2A_HUMAN Bromodomain adjacent to zinc fi (1905)  211  85.2 1.7e-14 0.382 0.683  123
BAZ2A_XENLA Bromodomain adjacent to zinc fi (1698)  206  83.3 5.5e-14 0.350 0.684  117
FSH_DROME Homeotic protein female sterile;  (2038)  205  82.9 8.8e-14 0.341 0.667  129
BAZ2A_MOUSE Bromodomain adjacent to zinc fi (1889)  204  82.5   1e-13 0.368 0.680  125
BRDT_MACFA Bromodomain testis-specific prot ( 947)  197  80.0   3e-13 0.367 0.697  109
BRD3_HUMAN Bromodomain-containing protein 3 ( 726)  194  78.9 4.9e-13 0.362 0.664  116
```

Homology and Domains – Histone acetyltransferase KAT2B



Protein Evolution and Sequence Similarity

Similarity Searching I

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison

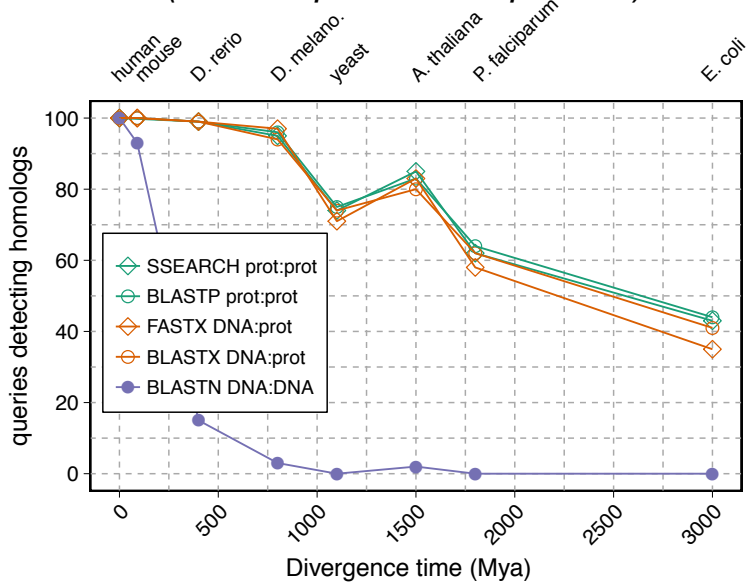
Similarity Searching II

- Alignment algorithms
- What are the problems?
  - missed homologs (false negatives, sensitivity)
  - ?false positives? (specificity)
- What do the statistics mean?
- How can we change behavior (scoring matrices)

DNA vs protein sequence comparison

The best scores are:		DNA E(188,018)	tfastx3 E(187,524)	prot. E(331,956)
DMGST	D.melanogaster GST1-1	1.3e-164	4.1e-109	1.0e-109
MDGST1	M.domestica GST-1 gene	2e-77	3.0e-95	1.9e-76
LUCGLTR	Lucilia cuprina GST	1.5e-72	5.2e-91	3.3e-73
MDGST2A	M.domesticus GST-2 mRNA	9.3e-53	1.4e-77	1.6e-62
MDNF1	M.domestica nf1 gene. 10	4.6e-51	2.8e-77	2.2e-62
MDNF6	M.domestica nf6 gene. 10	2.8e-51	4.2e-77	3.1e-62
MDNF7	M.domestica nf7 gene. 10	6.1e-47	9.2e-77	6.7e-62
AGGST15	A.gambiae GST mRNA	3.1e-58	4.2e-76	4.3e-61
CVU87958	Culicoides GST	1.8e-41	4.0e-73	3.6e-58
AGG3GST11	A.gambiae GST1-1 mRNA	1.5e-46	2.8e-55	1.1e-43
BMO6502	Bombyx mori GST mRNA	1.1e-23	8.8e-50	5.7e-40
AGSUGST12	A.gambiae GST1-1 gene	2.3e-16	4.5e-46	5.1e-37
MOTGLUSTRA	Manduca sexta GST	5.7e-07	2.5e-30	8.0e-25
RLGSTARGN	R.legominosarum *gstA*	0.0029	3.2e-13	1.4e-10
HUMGSTT2A	H. sapiens GSTT2	0.32	3.3e-10	2.0e-09
HSGSTT1	H.sapiens GSTT1 mRNA	7.2	8.4e-13	3.6e-10
ECAE000319	E. coli hypothet. prot.	—	4.7e-10	1.1e-09
MYMDCMA	Methyl. dichlorometh. DH	—	1.1e-09	6.9e-07
BCU19883	Burkholderia maleylacetate red.	—	1.2e-09	1.1e-08
NFU43126	Naegleria fowleri GST	—	3.2e-07	0.0056
SP505GST	Sphingomonas paucim	—	1.8e-06	0.0002
EN1838	H. sapiens maleylaceto. iso.	—	2.1e-06	5.9e-06
HSU86529	Human GSTZ1	—	3.0e-06	8.0e-06
SYCCPNC	Synechocystis GST	—	1.2e-05	9.5e-06
HSEF1GMR	H.sapiens EF1g mRNA	—	9.0e-05	0.00065

Detectable homologs to human enzymes (DNA vs protein comparison)



Why is protein comparison more sensitive?

- Larger alphabet: 20 aa vs 4 nt, means long alignments less likely by chance
- similarity scoring matrix
  - proteins have BLOSUM62: L ~ (V,I)
  - DNA typically match/mismatch A ≠ G
  - in 3rd codon position, DNA mismatch can be amino acid identity
- Smaller databases
- Better statistics
  - for proteins, E() < 0.001 is 1/1000 (unrelated looks like random)
  - for DNA, E() < 10^{-10} a more reliable threshold (unrelated doesn't always look random)

Computer lab: fasta.bioch.virginia.edu/biol4230/blast_demo.html

- Significant hits are homologous
- Non-significant hits? Homologous or not?
- Are *all* aligned residues homologous
- Are *unaligned* residues non-homologous
- Are domains really missing?
- Run a search from the command line