# Computational Genomics
# Protein Evolution / Similarity Searching

## What BLAST Does / Why BLAST works

Bill Pearson
wrp@virginia.edu

1

## Protein Evolution/ Similarity Searching

- 9:00 – Homology and Expectation value
- 11:00 – Similarity searching Workshop I
- 1:30 – safety overview/discussion
- 2:00 – Practical Similarity Searching, improving sensitivity
- 3:15 – Workshop II – More scoring matrices/over-extension
- 4:30 – Wine and Cheese / Poster session

2

## Effective Similarity Searching

1. Always search protein databases (possibly with translated DNA)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search

---

1. Search smaller (comprehensive) databases
2. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
3. All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss

## *Sequence Similarity - Conclusions*

- *Homologous* sequences share a common ancestor, but most sequences are *non-homologous*
- Always compare Protein Sequences
- Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
- Homologous proteins share common structures, but not necessarily common functions
- Sequence statistical significance estimates are accurate (verify this yourself)$10^{-6} < E() < 10^{-3}$ is statistically significant
- Scoring matrices set evolutionary look back horizons - not every discovery is distant
- PSI-BLAST can be more sensitive, but with lower statistical accuracy

4

*Establishing homology from statistically significant similarity*

Why BLAST works

- For most proteins, homologs are easily found over long evolutionary distances (500 My – 2 By) using standard approaches (BLAST, FASTA)
- Difficult for distant relationships or very short domains
- Most default search parameters are optimized for distant relationships and work well

5

Protein Evolution and Sequence Similarity

**Similarity Searching I**

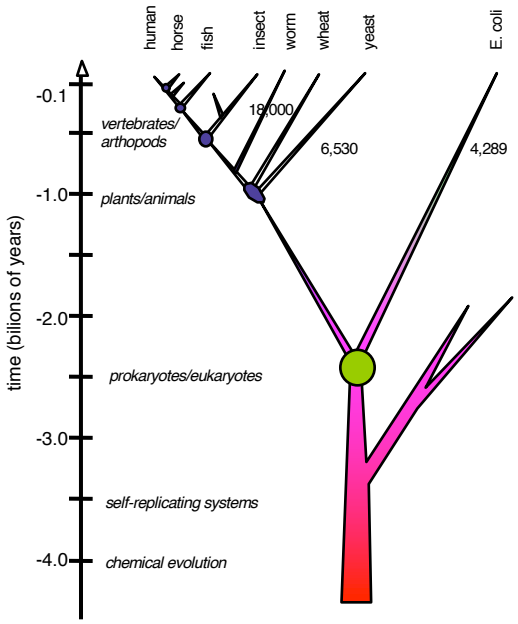
- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison

**Similarity Searching II**

- More effective similarity searching
  - Smaller databases
  - Appropriate scoring matrices
  - Using annotation/domain information

6

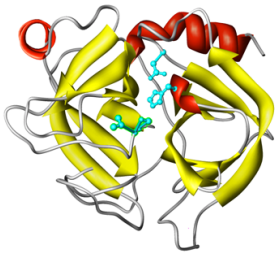
# Homologues share a common ancestor



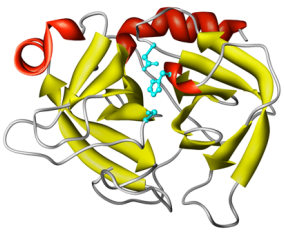

7

# When do we infer homology?

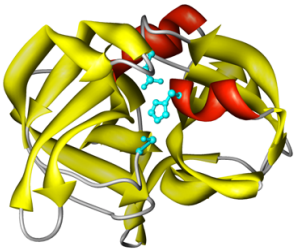
Homology <=> structural similarity
? sequence similarity



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



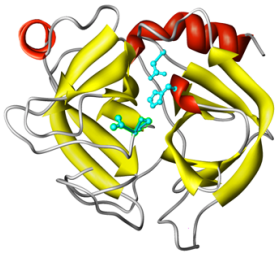
S. griseus trypsin (1sgt)
E()< $10^{-14}$ RMSD 1.6 A
E()< $10^{-19}$ 36%; 226/223



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

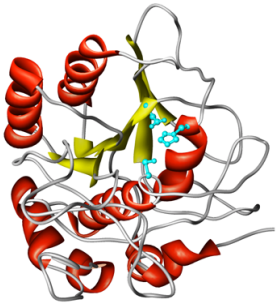
8

## 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

9

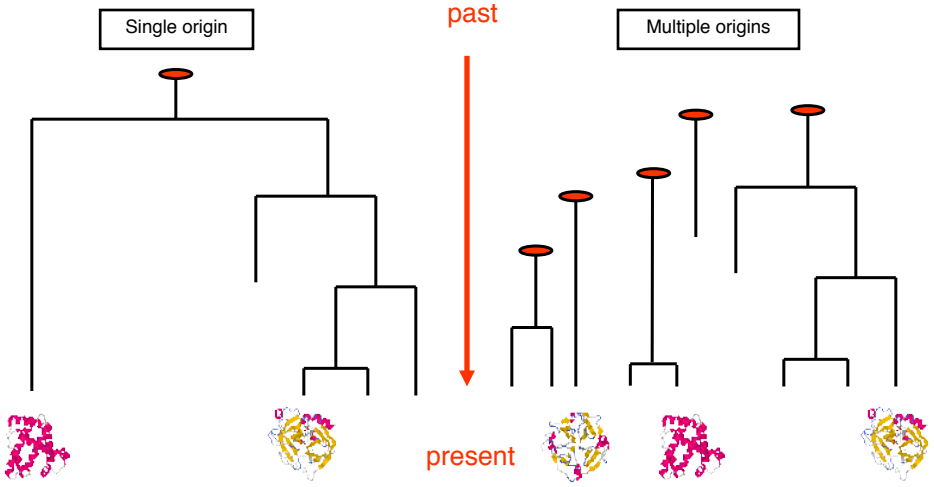
## Homology is confusing I: Homology defined Three(?) Ways

- 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

10

## 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).

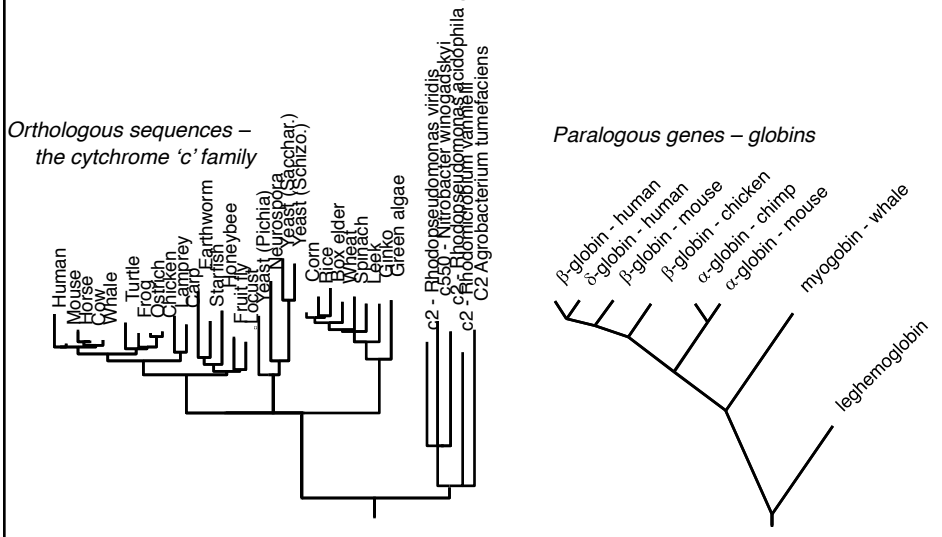
12

## 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 |

13

## Orthologs and Paralogs – Inferring Function



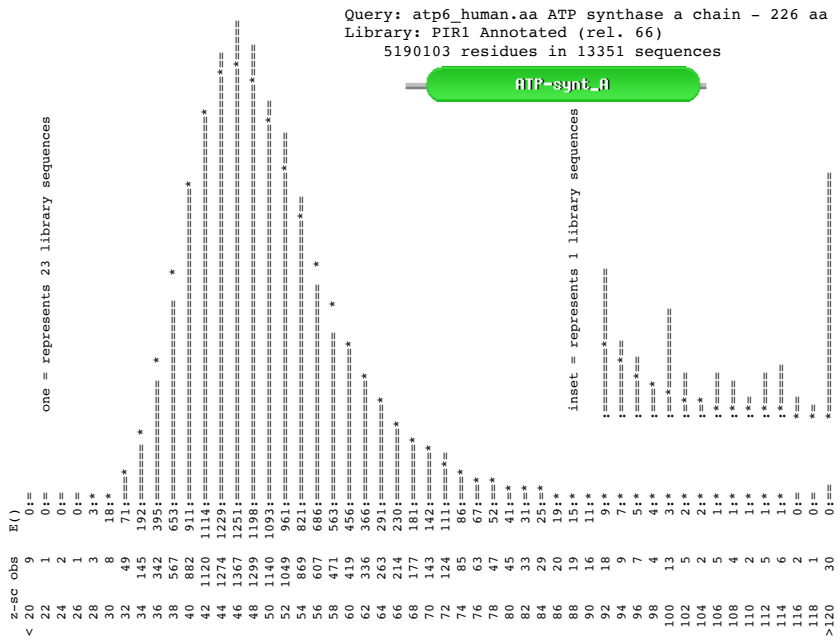

14

# Protein Evolution and Sequence Similarity

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

- More effective similarity searching
  - Smaller databases
  - Appropriate scoring matrices
  - Using annotation/domain information

15




16

# 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

17

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 |

18


```
>>sp|P0AB98|ATP6_ECOLI ATP synthase a chain (ATPase protein 6) g  (271 aa)
 s-w opt: 178  Z-score: 218.2  bits: 47.9 E(): 3.2e-06
Smith-Waterman score: 178; 23.3% identity (58.5% similar) in 236 aa overlap (8-222:45-264)

                                                 10        20        30        40
human                                          MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQ
                                                  :..  ..:::   ....::  .   ...  . ... :. .
E coli NMTPQDYIGHHLNNLQLDLRTFSLVDPQNPPATFWTINIDSMFFSVVLGL---LFLVLFRSVAKKATSG-VPGKFQTAIE
          10        20        30        40        50           60        70         80

       50        60        70        80                90              100       110
human  WLIKLTSKQMMTMHNTKGRTWSLMLVSLIIFIATTNLLGLLP---------HSF-------TPTTQLSMNLAMAIPLWAG
        .: ... ..  :.. :..  . . .......   ::. :::         : .       .:.......:.::. ..
E coli LVIGFVNGSVKDMYHGKSKLIAPLALTIFVWVFLMNLMDLLPIDLLPYIAEHVLGLPALRVVPSADVNVTLSMALGVF--
              90       100       110       120       130       140       150

            120       130           140       150       160       170       180
human  TVIMGFRSKIKNALAHFLPQGTPTPL-----IPMLVIIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINL
        ... : :   .... :  . :  :.     ::. .:.: .::: .:..:..:: .:. ::.:.. ::..     :   :
E coli -ILILFYSIKMKGIGGFTKELTLQPFNHWAFIPVNLILEGVSLLSKPVSLGLRLFGNMYAGELIFILIAGLLPWWSQWIL
       160       170       180       190       200       210       220       230

        190       200       210       220
human  PSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
          :: :::.          .::..: .:. .::
E coli NVPWAIFHILIIT---------LQAFIFMVLTIVYLSMASEEH
       240       250               260       270
```



# The PAM250 matrix

| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

frequency of replacement in homologs

$$\lambda S = \log\left(\frac{q_{ij}}{p_i p_j}\right)$$

frequency of alignment by chance

```
>>sp|P30391|ATPI_EUGGR Chloroplast ATP synthase a chain precursor   (251 aa)
 s-w opt: 123  Z-score: 151.3  bits: 35.4 E(): 0.017
Smith-Waterman score: 123; 25.7% identity (57.9% similar) in 214 aa overlap (21-222:50-243)

                                      10        20        30        40        50        60
human                          MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTM
                                                .::: :  : : :.: :  . . ...: .:.:... .  .
Euglena VNMFISGIFQIANVEVGQHFYWSILGFQIHGQVLINSWIVILIIGF--LSIYTTKNL--TLVPANKQIFIELVTEFITDI
       10        20        30        40        50          60          70        80

                    70        80         90           100       110       120
human   HNTK-GRT----WSLMLVSLIIFIATTNLLG-LLPHSFT--PTTQL---SMNLAMAIPLWAGTVIMGFRSKI-KNALAHF
         .:. :.     :  .. ....:: ..:  : :.: ..   :. .:   . ..  .  :   : .  : . . :..:..:
Euglena SKTQIGEKEYSKWVPYIGTMFLFIFVSNWSGALIPWKIIELPNGELGAPTNDINTTAGLAILTSLAYFYAGLNKKGLTYF
           90        100       110       120       130       140       150       160

       130       140       150       160       170       180       190       200
Human   LPQGTPTPLIPMLVIIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLLTILEIAVAL
            :::..  . :.: ..   .:..:. :: .:: : .:.. .. :         .:  ::. . ::.: ..   ..
Euglena KKYVQPTPILLPINILEDFT---KPLSLSFRLFGNILADELVVAVLVSL--------VP--LIVPVPLIFLGLF---TSG
          170       180          190       200       210                 220

       210       220
human   IQAYVFTLLVSLYLHDNT
         ::: .:. : . :.
Euglena IQALIFATLSGSYIGEAMEGHH
      230       240       250
```



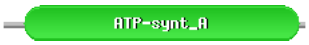
```
                  Query: atp6_human.aa ATP synthase a chain - 226 aa
                  Library: 5190103 residues in 13351 sequences

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
```

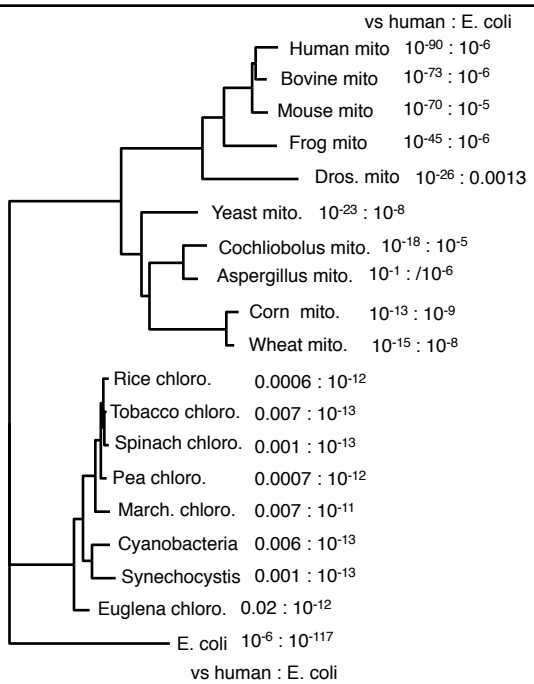
Query: atp6_ecoli.aa ATP synthase a - 271 aa
Library: 5190103 residues in 13351 sequences

| 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 |

23

# Homology is Transitive (on domains)



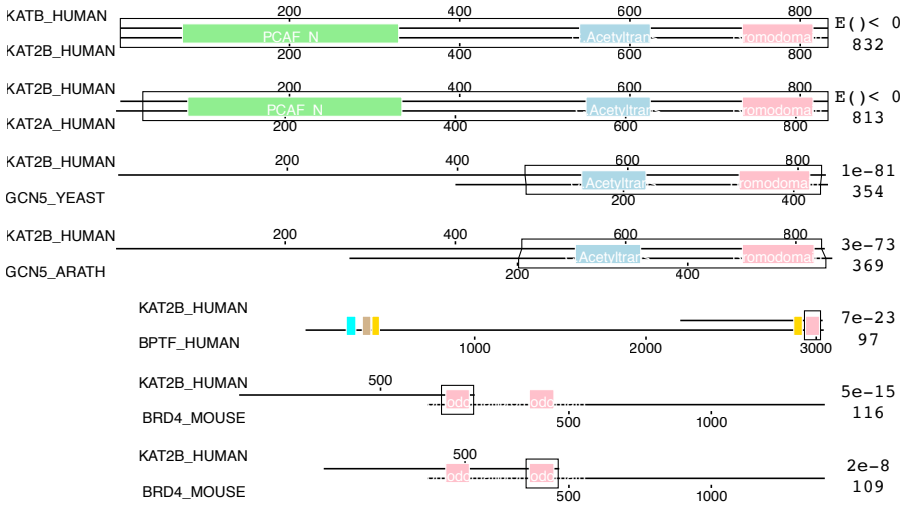

24

# 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 |

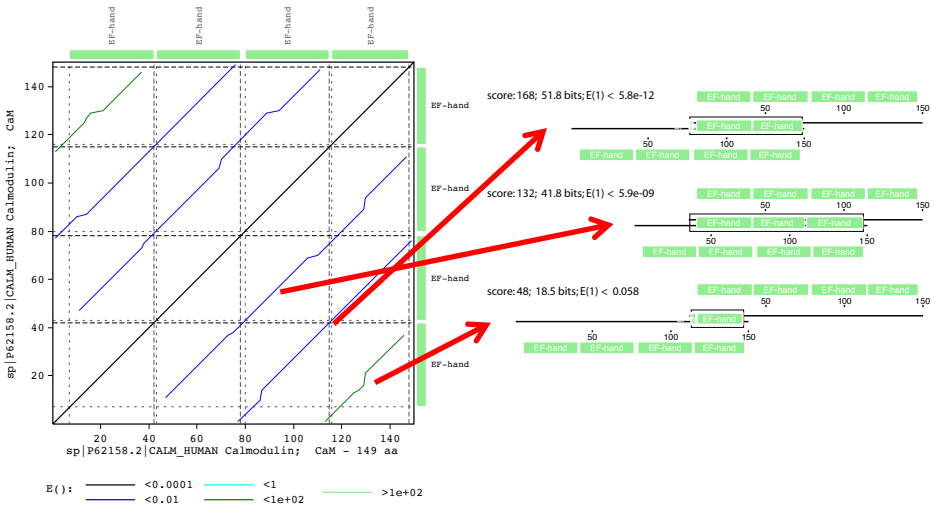
25

# Homology and Domains – Histone acetyltransferase KAT2B




26

## LALIGN – Identifying mobile domains: mobile (duplicated) domains in local alignments




## Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- More effective similarity searching
  - Smaller databases
  - Appropriate scoring matrices
  - Using annotation/domain information

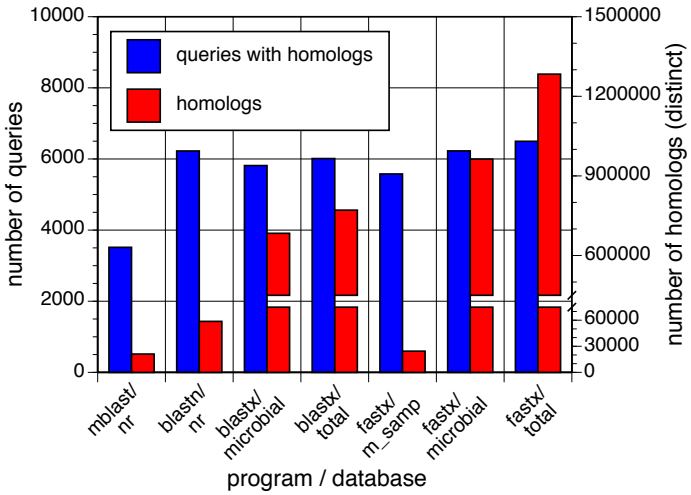
28

## *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 |

29

## Improving search strategies (windshield splatter metagenomics)




- always use protein/translated DNA comparisons
- smaller databases are more sensitive

## Effective Similarity Searching

1. Always search protein databases (possibly with translated DNA)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search (proteins)

---

1. Search smaller (comprehensive) databases
2. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (vertebrates, α-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
3. All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss

## Computer lab: fasta.bioch.virginia.edu/mol_evol

- Significant hits are homologous
- Non-significant hits? Homologous or not?
- Are *all* aligned residues homologous
- Are *unaligned* residues non-homologous
- Are domains really missing?

32