# Sequence Similarity

## Protein Sequence Comparison and Protein Evolution

(What BLAST does/Why BLAST works)

William R. Pearson
www.people.virginia.edu/~wrp
wrp@virginia.edu

1

## *Sequence Similarity - Conclusions*

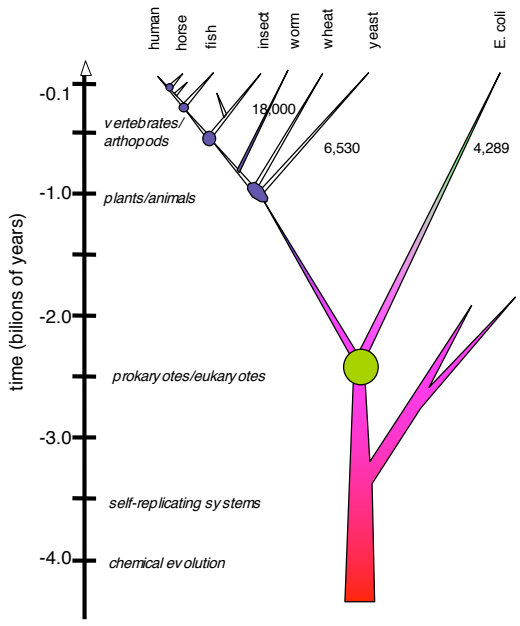
- Always compare Protein Sequences
- Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
- Homologous proteins share common structures, but not necessarily common functions
- Sequence statistical significance estimates are accurate (verify this yourself)$10^{-6} < E() < 10^{-3}$ is statistically significant
- Scoring matrices set evolutionary look back horizons - not every discovery is distant
- Structural and profile significance estimates are considerably less accurate that sequence comparison statistics

2

## Protein Evolution and Sequence Similarity

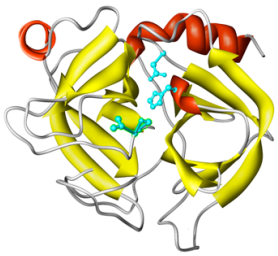
- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Sequence, Profile, and Structure Comparison

3

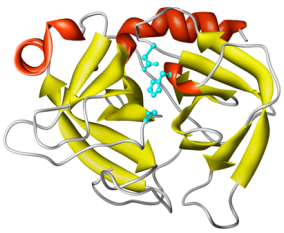



4

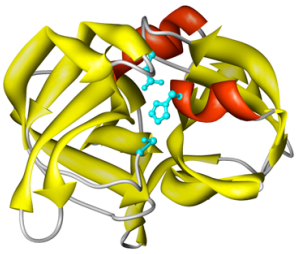
## Homology => structural similarity
## ? sequence similarity



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



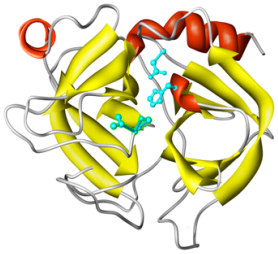
S. griseus trypsin (1sgt)
E()< $10^{-14}$ RMSD 1.6 A
E()< $10^{-19}$ 36%; 226/223



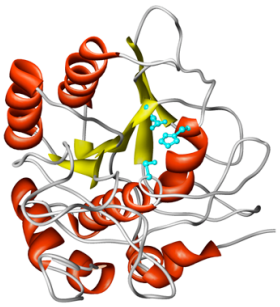
S. griseus protease A (2sga)
E()< $10^{-4}$; RMSD 2.6 A
E()< 2.6 25%; 199/181

5

## Non-homologous proteins have different structures



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



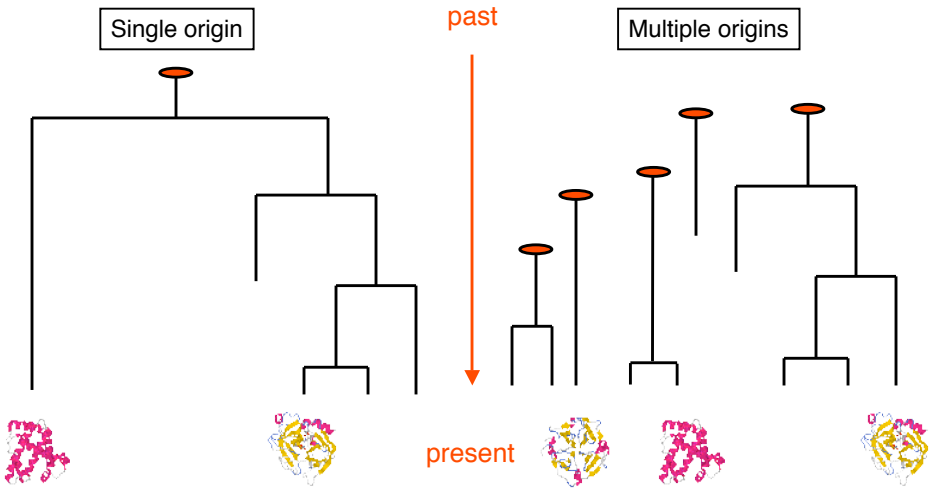
Subtilisin (1sbt)
E() >100
E()<280; 25% 159/275

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

6

*Are All Sequences Homologous?*

## No Homology without excess similarity




## What BLAST does:

Similarity ? <=> Homology

## Why BLAST works:

Statistical ? Biological
Significance <=> Significance

Divergence ? Convergence

8

## Some important dates in history

| | |
|---|---|
| Origin of the universe | –12[a] ±2 |
| Formation of the solar system | –4.6 ±0.4 |
| First self-replicating system | –3.5 ±0.5 |
| Prokaryotic-eukaryotic divergence | –2.5 ±0.3 |
| Plant-animal divergence | -1.0 |
| Invertebrate-vertebrate divergence | -0.5 |
| Mammalian radiation beginning | -0.1 |

[a]Billions of years ago

| Protein family | PAMs[a]/100 res. /$10^8$ years | Protein Lookback time[b] | |
|---|---|---|---|
| Pseudogenes | 400 | 45[c] | Primates,Rodents |
| Fibrinopeptides | 90 | 200 | Mammalian Radiation |
| Lactalbumins | 27 | 670 | Vertebrates |
| Ribonucleases | 21 | 850 | Animals |
| Hemoglobins | 12 | 1.5[d] | Plants/Animals |
| Acid Proteases | 8 | 2.3 | Prokayrotic/Eukarotic |
| Triosphosphate isomerase | 3 | 6 | Archaen |
| Glutamate dehydrogenase | 1 | 18 | ? |

[a]PAMs, point accepted mutations. [b]Useful lookback time, 360 PAMs, 15% identity. [c]Millions of years. [d]Billions of years.

9

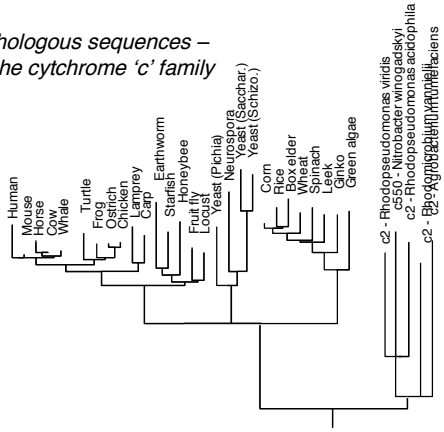
## E. coli proteins vs Human – Ancient Protein Domains

| 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 | Fumarase 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 |

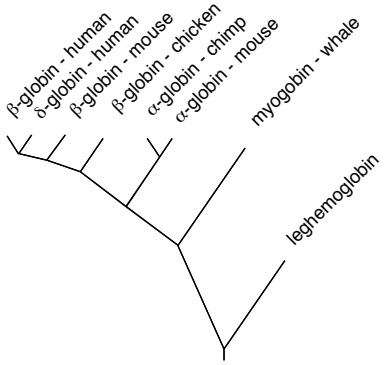
10

## Orthologs and Paralogs – Inferring Function

*Orthologous sequences – the cytchrome 'c' family*




*Paralogous genes – globins*

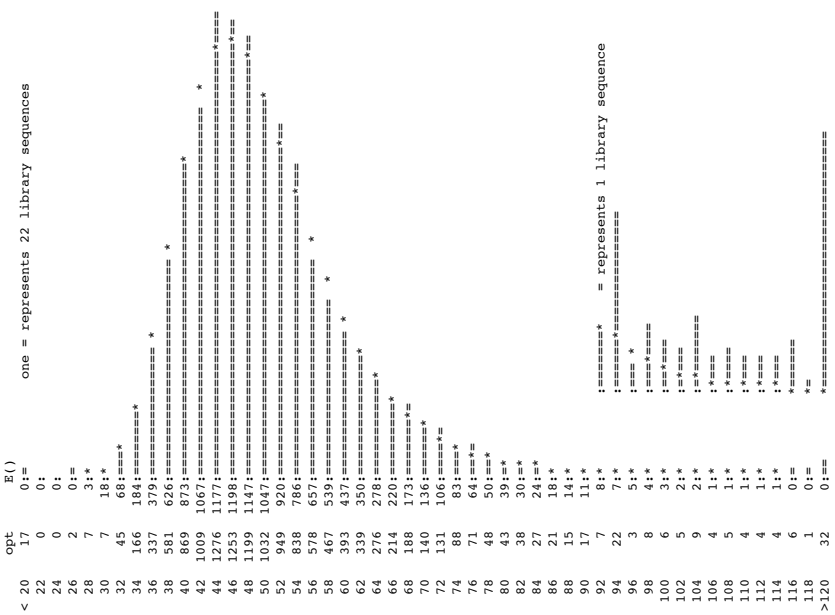



11

## 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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Sequence, Profile, and Structure Comparison

12




13

## 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

14

| The best scores are: | s-w | bits | E(14548) | % id | alen |
|---|---|---|---|---|---|
| PWHU6 H+-trans. ATP syn. - human mito. | 400 | 326.7 | 3.3e-90 | 1.000 | 226 |
| PWBO6 H+-trans. ATP syn. - cow mito. | 157 | 271.3 | 1.6e-73 | 0.779 | 226 |
| PWMS6 H+-trans. ATP syn. - mouse mito. | 118 | 262.4 | 7.6e-71 | 0.757 | 226 |
| PWXL6 H+-trans. ATP syn. - frog mito. | 745 | 177.3 | 3.1e-45 | 0.533 | 229 |
| PWFF6 H+-trans. ATP syn. - D. melanog. | 471 | 114.8 | 2.0e-26 | 0.378 | 222 |
| PWBY3 H+-trans. ATP syn. - yeast mito. | 438 | 107.3 | 4.4e-24 | 0.362 | 232 |
| PWAS6N H+-trans. ATP syn. - E. nidulans | 365 | 90.6 | 4.4e-19 | 0.304 | 230 |
| PWKQ6 H+-trans. ATP syn. - H. maydis | 353 | 87.9 | 3.0e-18 | 0.313 | 214 |
| PWWT6 H+-trans. ATP syn. - wheat mito. | 309 | 77.8 | 4.9e-15 | 0.292 | 233 |
| PWNT6M H+-trans. ATP syn. - tobacco | 309 | 77.8 | 5.0e-15 | 0.283 | 233 |
| PWZM6M H+-trans. ATP syn. - corn mito. | 283 | 71.9 | 2.2e-13 | 0.311 | 180 |
| LWEC6 H+-trans. ATP syn. - E. coli | 178 | 48.0 | 3.3e-06 | 0.237 | 236 |
| LWRZ6 H+-trans. ATP syn. - rice chloro. | 144 | 40.2 | 0.00063 | 0.242 | 231 |
| PWPMA6 H+-trans. ATP syn. - pea chloro. | 143 | 40.0 | 0.00074 | 0.250 | 232 |
| PWYBAA H+-trans. ATP syn. - Cyano. syn. | 142 | 39.7 | 0.00099 | 0.265 | 170 |
| PWSPA6 H+-trans. ATP syn. - spinach | 138 | 38.9 | 0.0016 | 0.238 | 231 |
| PWYCA6 H+-trans. ATP syn. - Synecho. | 127 | 36.3 | 0.0099 | 0.263 | 167 |
| LWNT6 H+-trans. ATP syn. - tobacco | 126 | 36.1 | 0.011 | 0.221 | 231 |
| LWLV6 H+-trans. ATP syn. - liverwort | 126 | 36.1 | 0.011 | 0.244 | 168 |
| PWEGAC H+-trans. ATP syn. - euglena | 123 | 35.4 | 0.018 | 0.257 | 214 |
| JQ0026 ATP/ADP translocase tlc1 - Ricket | 122 | 35.1 | 0.045 | 0.247 | 154 |
| S17420 ubiquinol--cytochrome-c reductase | 113 | 33.1 | 0.14 | 0.228 | 158 |
| QXBO2M NADH dehydrogenase (ubiquinone) | 107 | 31.7 | 0.32 | 0.261 | 211 |
| S17415 ubiquinol--cytochrome-c reductase | 105 | 31.3 | 0.49 | 0.277 | 137 |
| S17417 ubiquinol--cytochrome-c reductase | 104 | 31.0 | 0.57 | 0.277 | 137 |
| DNHUN2 NADH dehydrogenase (ubiquinone) | 103 | 30.8 | 0.61 | 0.201 | 149 |
| CBHU ubiquinol--cytochrome-c reductase | 102 | 30.6 | 0.79 | 0.268 | 205 |
| QRECAA aromatic amino acid trans. prot. | 103 | 30.8 | 0.82 | 0.234 | 111 |
| S17419 ubiquinol--cytochrome-c reductase | 101 | 30.3 | 0.92 | 0.234 | 158 |

15

```
>>LWEC6 H+-transporting ATP synthase (EC 3.6.1.34) protein 6 - Escherichia coli        (272 aa)
 s-w opt: 178  Z-score: 218.7  bits: 48.0 E(): 3.3e-06
Smith-Waterman score: 178;  23.729% identity (28.141% ungapped) in 236 aa overlap (8-222:45-264)

                    10        20        30        40        50        60        70        80
PWHU6         MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSLMLVSLIIFIA
                    :..  ..:::   ....::  .   ...  . ... :. . .: ... ..  :.. :.. . . .......
LWEC6   QNPPATFWTINIDSMFFSVVLGL---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSVKDMYHGKSKLIAPLALTIFVWVF
              40        50          60        70         80        90      100       110

                         90             100       110       120       130           140
PWHU6   TTNLLGLLP---------HSF-------TPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHFLPQGTPTPL-----IPMLVIIE
         ::. :::          : .       .:.......:.::. ..   ... : :  .... :  . :  :.     ::. .:.:
LWEC6   LMNLMDLLPIDLLPYIAEHVLGLPALRVVPSADVNVTLSMALGVF---ILILFYSIKMKGIGGFTKELTLQPFNHWAFIPVNLILE
          120       130       140       150          160       170       180         190

        150       160       170       180       190       210       220
PWHU6   TISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
        .::: .:..:..:: .:. ::.:.. ::..     :   :      :: :::.          .::..: .:. .::
LWEC6   GVSLLSKPVSLGLRLFGNMYAGELIFILIAGLLPWWSQWILNVPWAIFHILIIT---------LQAFIFMVLTIVYLSMASEEH
         200       210       220       230       240       250               260       270
```

16

## 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
```



## Where do scoring matrices come from?

Pam40

```
   A   R   N   D   E   I   L
A  8
R -9  12
N -4  -7  11
D -4 -13   3  11
E -3 -11  -2   4  11
I -6  -7  -7 -10  -7  12
L -8 -11  -9 -16 -12  -1  10
```

Pam250

```
   A   R   N   D   E   I   L
A  2
R -2   6
N  0   0   2
D  0  -1   2   4
E  0  -1   1   3   4
I -1  -2  -2  -2  -2   5
L -2  -3  -3  -4  -3   2   6
```

$q_{ij}$ : replacement frequency at PAM40, 250

$q_{R:N\,(40)} = 0.000435$ $p_R = 0.051$

$q_{R:N\,(250)} = 0.002193$ $p_N = 0.043$

$\lambda_2\ S_{ij} = \lg_2 (q_{ij}/p_i p_j)$ $\lambda_e\ S_{ij} = \ln(q_{ij}/p_i p_j)$ $p_R p_N = 0.002193$

$\lambda_2\ S_{R:N(\,40)} = \lg_2 (0.000435/0.00219) = -2.333$

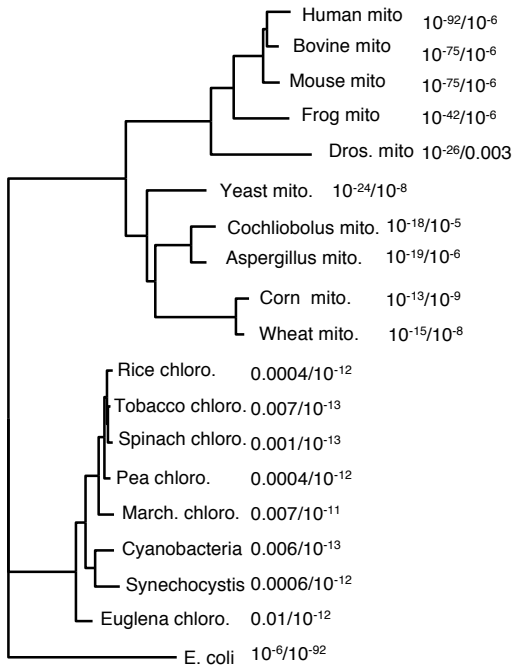
$\lambda_2 = 1/3$; $S_{R:N(\,40)} = -2.333/\lambda_2 = -7$

$\lambda\ S_{R:N(250)} = \lg 2\ (0.002193/0.002193) = 0$

- Scoring matrices can can be designed for different evolutionary distances (less=shallow; more=deep)
- Deep matrices allow more substitution

```
>PWEGAC H+-transporting ATP synthase (EC 3.6.1.34) chain a - Euglena gracilis chloroplast (252 aa)
 s-w opt: 123  Z-score: 151.6  bits: 35.4 E(): 0.018
Smith-Waterman score: 123;  25.701% identity (30.220% ungapped) in 214 aa overlap (21-222:50-243)

                         10        20        30        40        50        60            70
PWHU6             MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTK-GRT----WSLM
                                    .::: :  : : :.: :  . . ...: .:.:... .  . .:. :.     :  .
PWEGAC IANVEVGQHFYWSILGFQIHGQVLINSWIVILIIGF--LSIYTTKNL--TLVPANKQIFIELVTEFITDISKTQIGEKEYSKWVPY
      20        30        40        50          60          70        80        90      100

            80         90              100       110       120       130       140       150
PWHU6  LVSLIIFIATTNLLG-LLPHSFT--PTTQL---SMNLAMAIPLWAGTVIMGFRSKI-KNALAHFLPQGTPTPLIPMLVIIETISLL
       . ...:: ..:  : :.: ..   :. .:   .  ..  .  :   : .  : . . :..:..:     :::.. . :.: ..
PWEGAC IGTMFLFIFVSNWSGALIPWKIIELPNGELGAPTNDINTTAGLAILTSLAYFYAGLNKKGLTYFKKYVQPTPILLPINILEDFT--
           110       120       130       140       150       160       170       180

              160       170       180       190       200       210       220
PWHU6  IQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
        .:..:. :: .:: : .:.. .. :         .:  ::. . ::.: ..   .. ::: .:. : . :.
PWEGAC -KPLSLSFRLFGNILADELVVAVLVSL--------VP--LIVPVPLIFLGLF---TSGIQALIFATLSGSYIGEAMEGHH
       190        200       210                 220          230       240       250
```

19




20

## 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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Sequence, Profile, and Structure Comparison

21

## *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 |

22

Table 3: DNA and translated DNA similarity searches

| Taxonomic Group | blastx | blastn +3/-3 | blastn +1/-3 | |
|---|---|---|---|---|
| Bacteria eubacteria | | | | |
| . Proteobacteria proteobacteria | | | | |
| . . Gammaproteobacteria g-proteo. | | | | |
| . . . Enterobacteriaceae entero. | | | | |
| . . . . Shigella enterobacteria | | | | |
| . . . . . . . Shigella flexneri2a | 979 | 2165 | 2595 | enterobacteria |
| . . . . Escherichia coli CFT073 | 976 | 2130 | 2508 | enterobacteria |
| . . . . Escherichia coli O157:H7 | 959 | 2184 | 2642 | enterobacteria |
| . . . . Escherichia coli | 758 | 2253 | 2817 | enterobacteria |
| . . . . Edwardsiella tarda | 784 | 1102 | 180 | enterobacteria |
| . . Brucella melitensis 16M | 496 | 854 | 113 | a-proteobacter |
| . . Mesorhizobium loti | 60 | | | a-proteobacter |
| . . Bordetella bronchiseptica RB | 330 | 217 | | b-proteobacter |
| . . Geobacter metallireducens .. | 53 | | | d-proteobacter |
| . . Geobacter sulfurreducens PCA | 53 | | | d-proteobacter |
| . Prochlorococcus marinus MIT | 517 | 458 | | cyanobacteria |
| . Synechocystis sp. PCC 6803 ... | 466 | 284 | | cyanobacteria |
| . Clostridium perfringens str. 13 | 427 | | | eubacteria |
| . Streptomyces coelicolor A3(2). | 417 | | | high GC Gram+ |
| . Mycobacterium tuberculosis ... | 414 | 311 | | high GC Gram+ |
| . Listeria innocua ..... | 414 | 257 | | eubacteria |
| . Listeria monocytogenes ..... | 414 | 234 | | eubacteria |
| . Enterococcus faecium ..... | 411 | | | eubacteria |
| . Streptomyces avermitilis MA4680 | 409 | | | high GC Gram+ |
| . Lactococcus lactis ..... | 405 | 183 | | eubacteria |
| . Lactobacillus plantarum WCFS1. | 390 | 231 | | eubacteria |
| . Bacteroides thetaiotaomicronVPI | 387 | 233 | | CFB group bact |
| . Chloroflexus aurantiacus ..... | 72 | | | GNS bacteria |
| . Gloeobacter violaceus PCC 7421 | 48 | | | cyanobacteria |
| . Streptomyces viridifaciens ... | 45 | | | high GC Gram+ |
| . Clostridium tetani E88 ..... | 45 | | | eubacteria |

Bit scores from a `blastx` and `blastn` searches presented using the `BLAST` taxonomy summary option. The DNA sequence (`M84025`) encoding *E. coli* glutamate decarboxylase used to search the bacterial division of Genbank or Genpept. Species that contain a homolog with a bit score ≥ 45 (E() < $10^{-3}$ for `blastx`) are shown. The numbers under the `blastx` and `blastn` columns indicate the highest bit-score obtained for that taxonomic group.

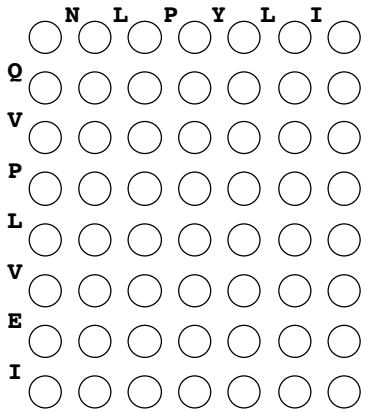
23

Table 6: Identification of anonymous DNA sequences at different evolutionary distances

| "unknown" DNA | excluded sequences | query length | E() thresh. | f'n found | Length | Coverage (X) |
|---|---|---|---|---|---|---|
| A. fulgidis | euryarch. | 10,000 | $10^{-6}$ | 1.00 | 0.785 | 632 |
| | | | $10^{-12}$ | 1.00 | 0.781 | 344 |
| | | 1,000 | $10^{-6}$ | 0.64 | 0.811 | 30 |
| | | | $10^{-12}$ | 0.64 | 0.748 | 21 |
| A. fulgidis 5% mut. | euryarch. | 10,000 | $10^{-6}$ | 1.00 | 0.657 | 260 |
| | | | $10^{-12}$ | 1.00 | 0.648 | 148 |
| | | 1,000 | $10^{-6}$ | 0.64 | 0.811 | 30 |
| | | | $10^{-12}$ | 0.64 | 0.748 | 21 |
| A. fulgidis | archaea | 10,000 | $10^{-6}$ | 1.00 | 0.725 | 607 |
| | | | $10^{-12}$ | 1.00 | 0.781 | 344 |
| | | 1,000 | $10^{-6}$ | 0.57 | 0.746 | 33 |
| | | | $10^{-12}$ | 0.52 | 0.733 | 21 |
| A. fulgidis 5% mut. | archaea | 10,000 | $10^{-6}$ | 1.00 | 0.553 | 240 |
| | | | $10^{-12}$ | 1.00 | 0.781 | 344 |
| | | 1,000 | $10^{-6}$ | 0.57 | 0.746 | 33 |
| | | | $10^{-12}$ | 0.52 | 0.733 | 21 |
| E. coli | bacteria | 10,000 | $10^{-6}$ | 1.00 | 0.430 | 102 |
| | | | $10^{-12}$ | 0.90 | 0.392 | 109 |
| | | 1,000 | $10^{-6}$ | 0.44 | 0.665 | 17 |
| | | | $10^{-12}$ | 0.36 | 0.682 | 11 |
| E. coli 5% mut. | bacteria | 10,000 | $10^{-6}$ | 0.90 | 0.375 | 92 |
| | | | $10^{-12}$ | 0.70 | 0.396 | 61 |
| | | 1,000 | $10^{-6}$ | 0.44 | 0.665 | 17 |
| | | | $10^{-12}$ | 0.36 | 0.682 | 11 |
| S. pyogenes | firmicutes | 10,000 | $10^{-6}$ | 1.00 | 0.723 | 570 |
| | | | $10^{-12}$ | 1.00 | 0.695 | 377 |
| S. pyogenes 5% mut. | firmicutes | 10,000 | $10^{-6}$ | 0.90 | 0.628 | 332 |
| | | | $10^{-12}$ | 0.90 | 0.524 | 195 |
| S. pyogenes | bacteria | 10,000 | $10^{-6}$ | 1.00 | 0.480 | 150 |
| | | | $10^{-12}$ | 0.90 | 0.475 | 89 |
| S. pyogenes 5% mut. | bacteria | 10,000 | $10^{-6}$ | 0.90 | 0.433 | 72 |
| | | | $10^{-12}$ | 0.80 | 0.433 | 43 |

24

## Smith-Waterman

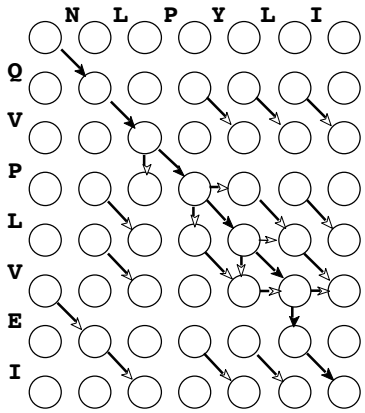



1. score every cell:

$S_{x,y} = max\{$
$S_{x-1,y-1} + match_{xy}$
$S_{x,y-1} - gappen$
$S_{x-1,y} - gappen$
0
$\}$

25

## Smith-Waterman




1. score every cell:

$S_{x,y} = max\{$
$S_{x-1,y-1} + match_{xy}$
$S_{x,y-1} - gappen$
$S_{x-1,y} - gappen$
0
$\}$

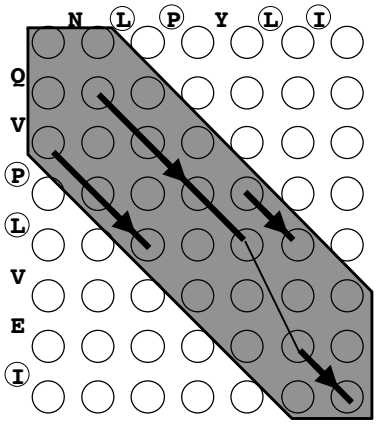
2. follow "traceback"

```
NLPYL-I
..: . :
QVPLVEI
```

Outcome: one continuous,optimal gapped alignment

26

# FASTA




1. Identify identical matches
   (length = *ktup*)
2. Extend along diagonal
   (local maximum)
3. Join diagonal segments (DP)
   (maintain linearity)
   (optimal sum score)

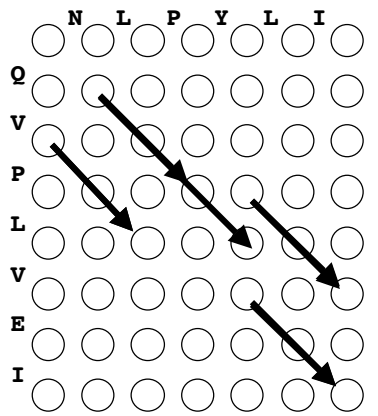
4. Banded Smith-Waterman

```
NLPYL-I
..: . :
QVPLVEI
```

Outcome: one continuous, near-optimal gapped alignment

27

# BLAST




1. neighborhood word hits
   (word length)
2. extend from diagonal ends
   (X-drop threshold)
3. report HSP linkages
   (maintain linearity)
   (probability)

```
NL    NLP   LI
.:    ..:   .:
PL    QVP   EI
```

Outcome: multiple HSPs, multiple linkages; only partially aligned

28

## More about scoring matrices ...

PAM series:

- Evolutionary model - extrapolated from PAM1
- PAM20: 20% change (mammals)
- PAM250: 250% change (<20% identity)
- Gap penalties should vary
- shallow matrices (PAM10-40) for short sequences and short distances

BLOSUM series

- Empirically determined, no extrapolation (no model)
- BLOSUM45-50 - distant (1/3 bits)
- BLOSUM80 -very highly conserved (not small change), high info/position
- BLOSUM62 - 1/2 bits

29

Changing Scoring Parameters

A. Search with MJ0050

| | BLOSUM50 -10/-2 | | | | BLOSUM62 -7/-1 | | | | BLOSUM62 -11/-1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **The best scores are:** | s-w | E() | %_id | alen | s-w | E() | %_id | alen | s-w | E() | %_id | alen |
| NP_416010 glutamate decarb. | 250 | e-11 | 24.9 | 401 | 216 | e-7 | 25.3 | 415 | 137 | e-8 | 22.9 | 332 |
| NP_417379 glycine decarb. | 169 | e-05 | 22.1 | 420 | 163 | 0.001 | 23.3 | 430 | 88 | 0.004 | 22.1 | 331 |
| NP_417025 aminotransferase | 122 | 0.02 | 23.6 | 254 | 119 | 0.12 | 24.5 | 257 | 76 | 0.04 | 23.7 | 118 |
| NP_414772 aminoacyl-his. | 110 | 0.15 | 23.4 | 188 | 108 | 0.74 | 23.2 | 311 | 57 | 6.9 | 23.4 | 188 |
| NP_415139 alkyl hydroperoxide | 99 | 1.1 | 26.9 | 156 | 104 | 1.5 | 24.5 | 233 | 62 | 2.0 | 28.9 | 97 |

B. Search with MJ1633

| | BLOSUM50 -10/-2 | | | | BLOSUM62 -7/-1 | | | | BLOSUM62 -11/-1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **The best scores are:** | s-w | E() | %_id | alen | s-w | E() | %_id | alen | s-w | E() | %_id | alen |
| NP_417809 KefB | 196 | e-06 | 28.2 | 177 | 162 | 0.02 | 27.3 | 176 | 143 | e-8 | 34.4 | 96 |
| NP_414589 K+ antiporter | 175 | e-04 | 25.4 | 142 | 141 | 0.2 | 24.7 | 166 | 131 | e-7 | 25.4 | 142 |
| NP_415011 transport protein | 133 | 0.03 | 23.2 | 142 | 113 | 4.4 | 23.2 | 142 | 89 | 0.005 | 23.2 | 142 |
| NP_417748 TrkA | 128 | 0.04 | 23.7 | 135 | 114 | 2.9 | 22.2 | 176 | 99 | e-3 | 21.8 | 133 |
| NP_416807 NAD(P) binding | 103 | 0.98 | 26.1 | 92 | | | | | 70 | 0.29 | 26.1 | 92 |

30

## Where do scoring matrices come from?

Pam40

| | A | R | N | D | E | I | L |
|---|---|---|---|---|---|---|---|
| A | 8 | | | | | | |
| R | -9 | 12 | | | | | |
| N | -4 | -7 | 11 | | | | |
| D | -4 | -13 | 3 | 11 | | | |
| E | -3 | -11 | -2 | 4 | 11 | | |
| I | -6 | -7 | -7 | -10 | -7 | 12 | |
| L | -8 | -11 | -9 | -16 | -12 | -1 | 10 |

Pam250

| | A | R | N | D | E | I | L |
|---|---|---|---|---|---|---|---|
| A | 2 | | | | | | |
| R | -2 | 6 | | | | | |
| N | 0 | 0 | 2 | | | | |
| D | 0 | -1 | 2 | 4 | | | |
| E | 0 | -1 | 1 | 3 | 4 | | |
| I | -1 | -2 | -2 | -2 | -2 | 5 | |
| L | -2 | -3 | -3 | -4 | -3 | 2 | 6 |

$q_{ij}$ : replacement frequency at PAM40, 250

$q_{R:N\ (40)} = 0.000435$ $\quad p_R = 0.051$

$q_{R:N\ (250)} = 0.002193$ $\quad p_N = 0.043$

$\lambda_2\ S_{ij} = \lg_2 (q_{ij}/p_i p_j)$ $\quad \lambda_e\ S_{ij} = \ln(q_{ij}/p_i p_j)$ $\quad p_R p_N = 0.002193$

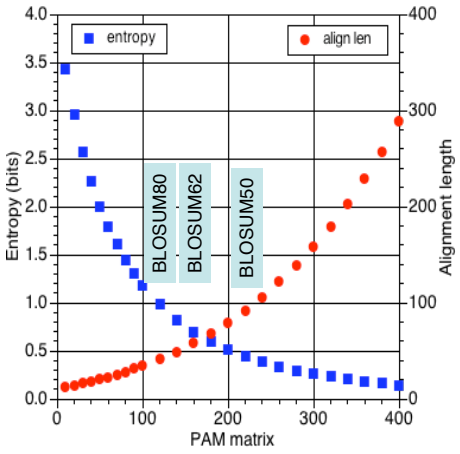
$\lambda_2\ S_{R:N(\ 40)} = \lg_2 (0.000435/0.00219) = -2.333$

$\lambda_2 = 1/3$; $S_{R:N(\ 40)} = -2.333/\lambda_2 = -7$

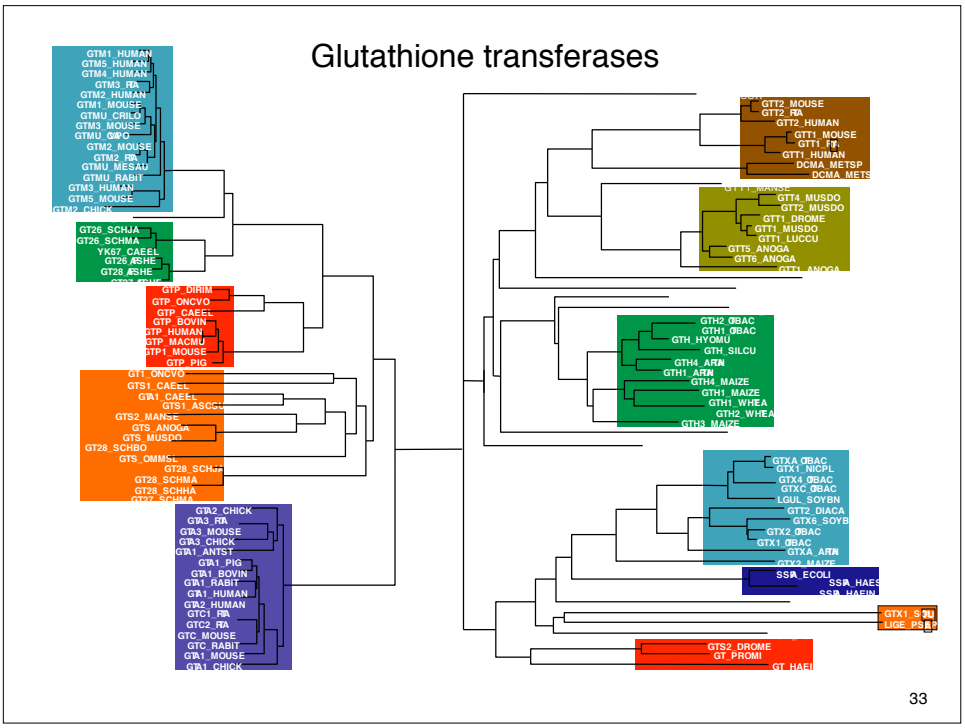
$\lambda\ S_{R:N(250)} = \lg2\ (0.002193/0.002193) = 0$

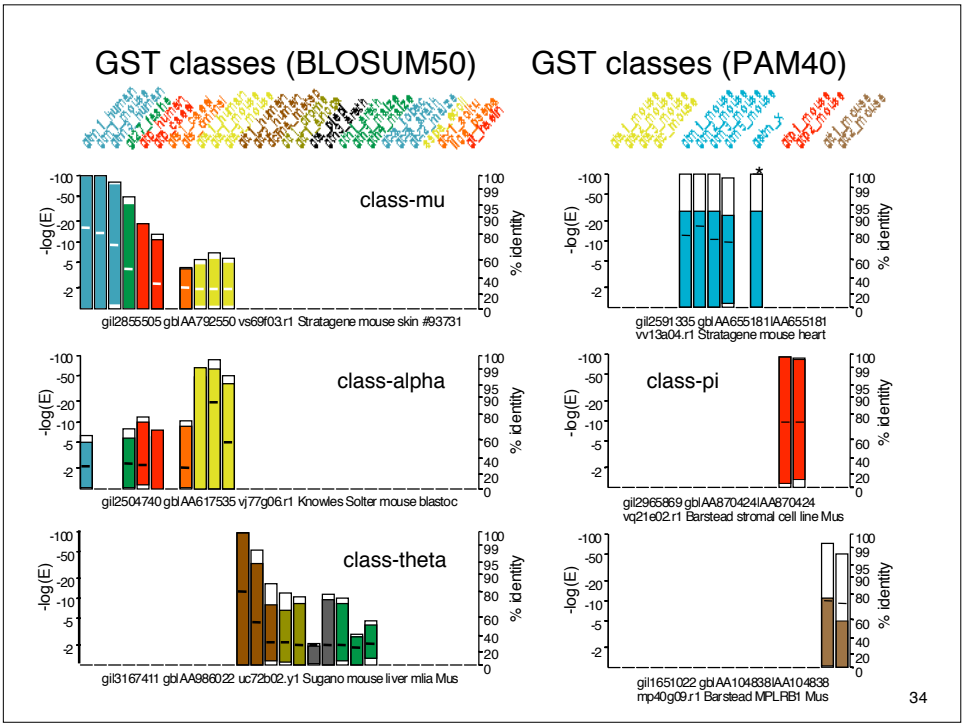
31

## PAM matrices and alignment length




32


Glutathione transferases
GTM1_HUMAN
GTM5_HUMAN
GTM4_HUMAN
GTM3_RA
GTM2_HUMAN
GTM1_MOUSE
GTMU_CRILO
GTM3_MOUSE
GTMU_CAVPO
GTM2_MOUSE
GTM2_RA
GTMU_MESAU
GTMU_RABIT
GTM3_HUMAN
GTM5_MOUSE
GTM2_CHICK
GT26_SCHJA
GT26_SCHMA
YK67_CAEEL
GT26_ASHE
GT28_ASHE
GTP_DIRIM
GTP_ONCVO
GTP_CAEEL
GTP_BOVIN
GTP_HUMAN
GTP_MACMU
GTP1_MOUSE
GTP_PIG
GT1_ONCVO
GTS1_CAEEL
GTA1_CAEEL
GTS1_ASCSU
GTS2_MANSE
GTS_ANOGA
GTS_MUSDO
GT28_SCHBO
GTS_OMMSE
GT28_SCHJA
GT28_SCHMA
GT28_SCHHA
GT27_SCHMA
GTA2_CHICK
GTA3_RA
GTA3_MOUSE
GTA3_CHICK
GTA1_ANTST
GTA1_PIG
GTA1_BOVIN
GTA1_RABIT
GTA1_HUMAN
GTA2_HUMAN
GTC1_RA
GTC2_RA
GTC_MOUSE
GTC_RABIT
GTA1_MOUSE
GTA1_CHICK
GTT2_MOUSE
GTT2_RA
GTT2_HUMAN
GTT1_MOUSE
GTT1_RA
GTT1_HUMAN
DCMA_METSP
DCMA_METS
GTT4_MUSDO
GTT2_MUSDO
GTT1_DROME
GTT1_MUSDO
GTT1_LUCCU
GTT5_ANOGA
GTT6_ANOGA
GTH2_ORYSA
GTH1_ORYSA
GTH_HYOMU
GTH_SILCU
GTH4_ARATH
GTH1_ARATH
GTH4_MAIZE
GTH1_MAIZE
GTH1_WHEAT
GTH2_WHEAT
GTH3_MAIZE
GTXA_ORYSA
GTX1_NICPL
GTX4_ORYSA
GTXC_ORYSA
LGUL_SOYBN
GTT2_DIACA
GTX6_SOYBN
GTX2_ORYSA
GTX1_ORYSA
GTXA_ARATH
GTX2_MAIZE
SSIA_ECOLI
SSIA_HAES
SSIA_HAEIN
GTX1_SCHJA
LIGE_PSEAE
GTS2_DROME
GT_PROMI
GT_HAEI
33


GST classes (BLOSUM50)
GST classes (PAM40)
class-mu
class-alpha
class-theta
class-pi
-log(E)
% identity
gil2855505 gbl AA792550 vs69f03.r1 Stratagene mouse skin #93731
gil2591335 gbl AA655181 IAA655181
vv13a04.r1 Stratagene mouse heart
gil2504740 gbl AA617535 vj77g06.r1 Knowles Solter mouse blastoc
gil2965869 gbl AA870424 IAA870424
vq21e02.r1 Barstead stromal cell line Mus
gil3167411 gbl AA986022 uc72b02.y1 Sugano mouse liver mlia Mus
gil1651022 gbl AA104838 IAA104838
mp40g09.r1 Barstead MPLRB1 Mus
34

## *Scoring Matrices - Summary*

- PAM and BLOSUM matrices greatly improve the sensitivity of protein sequence comparison – low identity with significant similarity
- PAM matrices have an evolutionary model - lower number, less divergence – lower=closer; higher=more distant
- BLOSUM matrices are sampled from conserved regions at different average identity – higher=more conservation
- Short alignments require shallow matrices
- Shallow matrices set maximum look-back time

35

## 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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Sequence, Profile, and Structure comparison

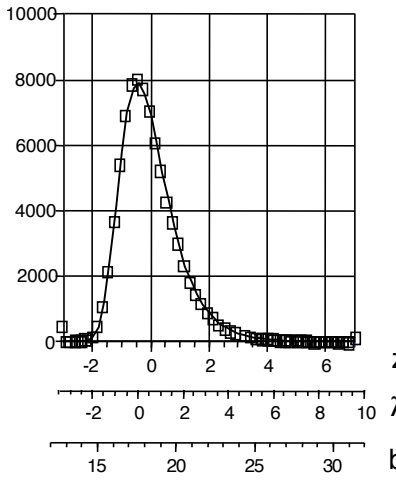
36

## 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

37

## Extreme value distribution




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

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

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

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

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

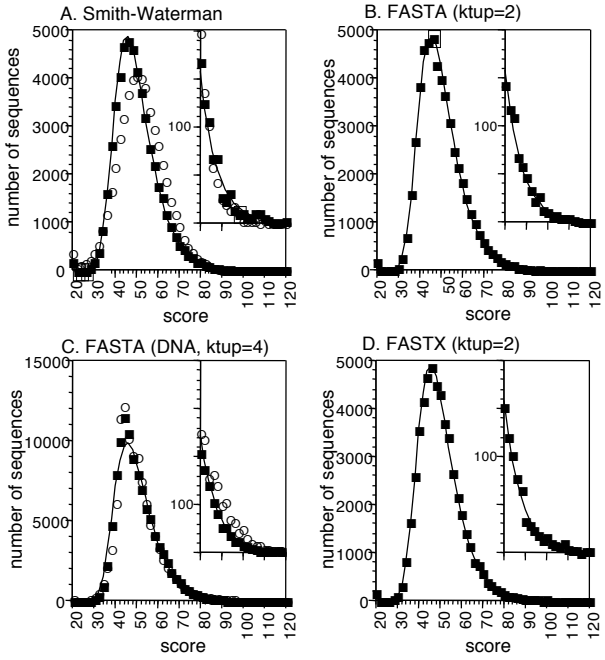
$P(B \text{ bits}) = m\ n\ 2^{-B}$

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

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

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

38




39

## Smith-Waterman (ssearch)

```
The best scores are:                      s-w bits E(115640)  %_id alen
GTM1_MOUSE Glutathione S-trans ( 218) 1497 363.5  2e-100  1.000  218
GTM2_CHICK Glutathione S-trans ( 220)  958 234.9 1.1e-61  0.619  218
GTP_HUMAN  Glutathione S-trans ( 210)  356  91.2 1.8e-18  0.308  211
PGD2_MOUSE Glutathione-req.    ( 199)  262  68.8 9.7e-12  0.319  204
GTA1_MOUSE Glutathione S-trans ( 223)  229  60.9 2.6e-09  0.284  225
SC1_OCTDO  S-crystallin 1 OL1  ( 215)  228  60.7 3.0e-09  0.269  219
GTS_MUSDO  Glutathione S-trans ( 241)  228  60.6 3.4e-09  0.264  201
GTS1_CAEEL Prob. Glut. S-trans ( 210)  220  58.8 1.1e-08  0.284  225
GTS_OMMSL  Glutathione S-trans ( 203)  196  53.0 5.5e-07  0.258  209
GTH3_ARATH Glutathione S-trans ( 215)  142  40.1  0.0045  0.310  126
GTT2_HUMAN Glutathione S-trans ( 244)  132  37.7   0.027  0.257  167
GT24_DROME Glutathione S-trans ( 216)  131  37.5   0.028  0.255  153
YFCG_ECOLI Hypothetical GST    ( 215)  112  33.0    0.64  0.235  187
YJY1_YEAST hypothetical 30.5   ( 261)  110  32.4   *1.1*  0.248  149
DCMA_METS1 dichloromethane DM  ( 267)  103  30.8     3.7  0.214  210
YA42_HAEIN Hypothetical prot.  ( 617)  108  31.7   *4.6*  0.283  120
GTO1_RAT   Glutathione trans   ( 241)  100  30.1     5.4  0.234  158
DP41_BACHD DNA polymerase I    ( 413)  104  30.8   *5.4*  0.234  184
GTH1_WHEAT Glutathione S-trans ( 229)   98  29.6     7.0  0.246  171
LGUL_SOYBN Lactoylglutathione  ( 219)   97  29.4     7.8  0.200  190
VP2_AHSV3  outer capsid prot   (1057)  108  31.5   *8.9*  0.205  200
GTH5_ARATH Glutathione S-trans ( 218)   96  29.2     9.2  0.258   66
DCMA_METSP dichloromethane DM  ( 288)   98  29.5     9.3  0.195  200
GTXA_ARATH Glutathione S-trans ( 224)   96  29.1     9.5  0.248  125
SLT_HAEIN  Putative soluble l  ( 593)  103  30.5   *9.9*  0.227  185
```

40

## Low gap penalties reduce sensitivity

```
The best scores are:                      s-w bits E(115640) %_id  alen
GTM1_MOUSE Glutathione S-tran  ( 218) 1497 164.0 2.3e-40  1.000  218
GTM2_CHICK Glutathione S-tran  ( 220)  958 107.5 2.4e-23  0.619  218
GTP_HUMAN  Glutathione S-tran  ( 210)  378  46.8 4.2e-05  0.308  211
PGD2_MOUSE Glutathione-req.    ( 199)  311  39.9  0.0048  0.319  204
GTA1_MOUSE Glutathione S-tran  ( 223)  296  38.1   0.019  0.313  233
SC1_OCTDO  S-crystallin 1 OL1  ( 215)  286  37.2   0.035  0.272  224
GTS_MUSDO  Glutathione S-tran  ( 241)  279  36.2   0.077  0.274  219
GTS_OMMSL  Glutathione S-tran  ( 203)  241  32.6    0.81  0.261  222
GTH3_ARATH Glutathione S-tran  ( 215)  190  27.1      38  0.293  198
GTT2_HUMAN Glutathione S-tran  ( 244)  189  26.7      55  0.271  210
GTT1_MUSDO Glutathione S-tran  ( 208)  183  26.4      58  0.276  199
MAAI_VIBCH Probable maleylace  ( 215)  184  26.5      58  0.235  247
YFCG_ECOLI Hypothetical GST-   ( 215)  184  26.5      58  0.246  224
GTXA_TOBAC prob. Glutathione   ( 220)  184  26.4      62  0.250  204
GTH1_WHEAT Glutathione S-tran  ( 229)  185  26.4      63  0.246  236
GTH7_ARATH Glutathione S-tran  ( 214)  180  26.1      77  0.254  228
T1MH_METJA Putative type I r   ( 558)  210  27.3     *85* 0.255  275
DP41_BACHD DNA polymerase I    ( 413)  200  26.8     *86* 0.244  234
GTH2_WHEAT Glutathione S-tran  ( 291)  188  26.3      90  0.247  251
```

41

## *FASTA search – low complexity regions*

Search with complete grou_drome:

```
The best scores are:                                          opt  bits  E(14548)
RGHUB1 GTP-binding regulatory protein beta-1 chai  ( 341)  237  46.6  3.5e-05
RGBOB1 GTP-binding regulatory protein beta-1 chai  ( 341)  237  46.6  3.5e-05
RGHUB3 GTP-binding regulatory protein beta-3 chai  ( 341)  233  46.0  5.2e-05
RGMSB4 GTP-binding regulatory protein beta-4 chai  ( 341)  232  45.8  5.7e-05
PIHUPF salivary proline-rich glycoprotein precurs  ( 252)  224  44.5 *0.00010*
RGFFB  GTP-binding regulatory protein beta chain   ( 347)  223  44.5  0.00014
PIRT3  acidic proline-rich protein precursor - rat ( 207)  199  40.8 *0.0011*
PIHUB6 salivary proline-rich protein precursor PR  ( 393)  203  41.6 *0.0012*
CGBO2S collagen alpha 2(I) chain - bovine (fragme  ( 403)  195  40.5 *0.0027*
WMBEW6 capsid protein - human herpesvirus 1 (stra  ( 636)  192  40.2 *0.0051*
W4WLB5 E4 protein - human papillomavirus type 5b   ( 246)  170  36.6 *0.024*
OZZQMY circumsporozoite protein precursor - Plasm  ( 368)  172  37.1 *0.026*
FOMVME gag polyprotein - murine leukemia virus (s  ( 537)  161  35.6 *0.10*
```

Search with seg-ed grou_drome: (low complexity regions removed)

```
The best scores are:                                          opt bits E(14548)
RGHUB3 GTP-binding regulatory protein beta-3 chai  ( 341)  233  56.5 3.6e-08
RGMSB4 GTP-binding regulatory protein beta-4 chai  ( 341)  232  56.3 4.1e-08
RGHUB2 GTP-binding regulatory protein beta-2 chai  ( 341)  228  55.5 7.2e-08
RGBOB1 GTP-binding regulatory protein beta-1 chai  ( 341)  225  54.9 1.1e-07
RGFFB  GTP-binding regulatory protein beta chain   ( 347)  223  54.5 1.5e-07
BVBYMS MSI1 protein - yeast (Saccharomyces cerevi  ( 423)  135  37.0 *0.033*
ERHUAH coatomer complex alpha chain homolog - hum (1225)  134  37.1 *0.088*
A28468 chromogranin A precursor - human            ( 458)  122  34.4 *0.21*
RGOOBE GTP-binding regulatory protein beta chain   ( 342)  120  33.9  0.22
```

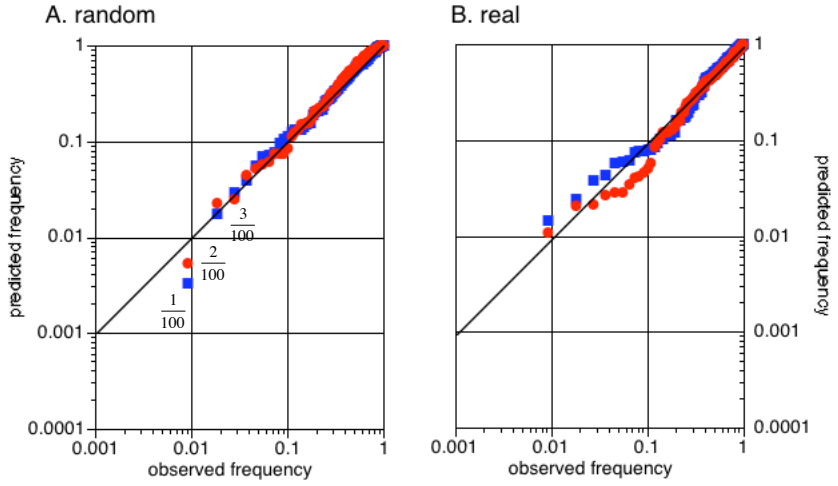
42

## pseg removes low-complexity regions

```
>gi|17380405|sp|P16371|GROU_DROME Groucho protein (Enhancer of split M9/10)

                              1-8     MYPSPVRH
                  paaggpppqgp 9-19
                             20-131   IKFTIADTLERIKEEFNFLQAQYHSIKLEC
                                      EKLSNEKTEMQRHYVMYYEMSYGLNVEMHK
                                      QTEIAKRLNTLINQLLPFLQADHQQQVLQA
                                      VERAKQVTMQELNLIIGQQIHA
                 qqvpggppqpmg 132-143
                             144-281  ALNPFGALGATMGLPHGPQGLLNKPPEHHR
                                      PDIKPTGLEGPAAAEERLRNSVSPADREKY
                                      RTRSPLDIENDSKRRKDEKLQEDEGEKSDQ
                                      DLVVDVANEMESHSPRPNGEHVSMEVRDRE
                                      SLNGERLEKPSSSGIKQE
               rppsrsgssssrstps 282-297
                             298-310  LKTKDMEKPGTPG
            akartptpnaaapapgvnpk 311-330
           qmmpqgpppagypgapyqrpa 331-351
                             352-719  DPYQRPPSDPAYGRPPPMPYDPHAHVRTNG
                                      IPHPSALTGGKPAYSFHMNGEGSLQPVPFP
                                      PDALVGVGIPRHARQINTLSHGEVVCAVTI
                                      SNPTKYVYTGGKGCVKVWDISQPGNKNPVS
                                      QLDCLQRDNYIRSVKLLPDGRTLIVGGEAS
                                      NLSIWDLASPTPRIKAELTSAAPACYALAI
                                      SPDSKVCFSCCSDGNIAVWDLHNEILVRQF
                                      QGHTDGASCIDISPDGSRLWTGGLDNTVRS
                                      WDLREGRQLQQHDFSSQIFSLGYCPTGDWL
                                      AVGMENSHVEVLHASKPDKYQLHLHESCVL
                                      SLRFAACGKWFVSTGKDNLLNAWRTPYGAS
                                      IFQSKETSSVLSCDISTDDKYIVTGSGDKK
                                      ATVYEVIY
```



## Protein Sequence Comparison Statistics are Accurate



## Statistical estimates from random shuffles

- BLAST estimates statistical significance from simulations of “normal” (average composition) proteins
- FASTA estimates statistical significance from the distribution of similarity scores obtained during the database search (selects 60,000 unrelated sequence scores from the database of *real* proteins)
- What if the sequences are different from most proteins, but similar to each other, e.g. membrane proteins?
- PRSS estimates statistical significance by producing hundreds of shuffled (random) sequences with the same length and composition, and then estimates λ and K from comparisons against those proteins

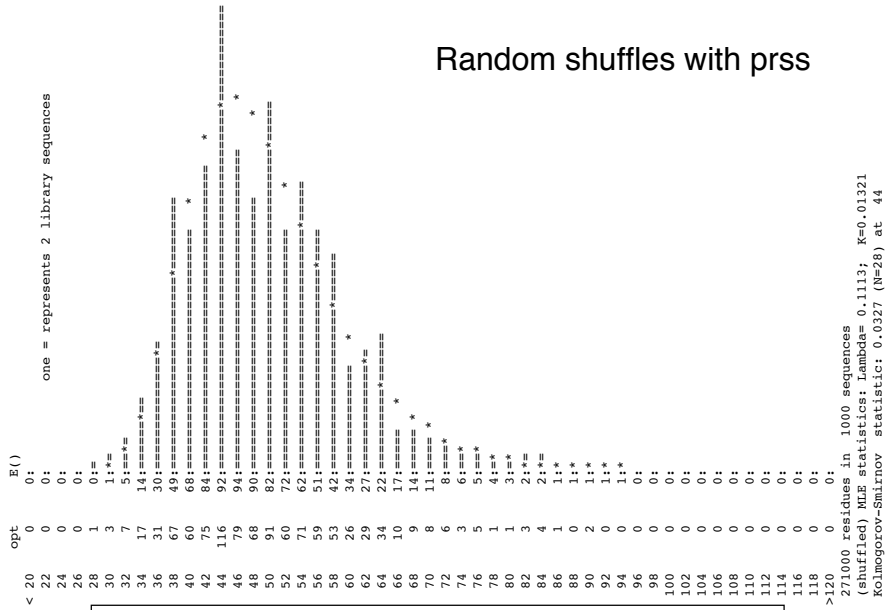
45

## prss - uniform and window shuffle

```
 >LWEC6_H+-transporting ATP synthase (EC 3.6.1.34) protein 6 - Escherichia coli
 MASENMTPQI YIGHHLNNLQ LDLRTFSLVD PQNPPATFWT INIDSMFFSV VLGLLFLVLF
 RSVAKKATSG VPGKFQTAIE LVIGFVNGSV KDMYHGKSKL IAPLALTIFV WVFLMNLMDL
 LPIDLLPYIA EHVLGLPALR VVPSADVNVT LSMALGVFIL ILFYSIKMKG IGGFTKELTL
 QPFNHWAFIP VNLILEGVSL LSKPVSLGLR LFGNMYAGEL IFILIAGLLP WWSQWILNVP
 WAIFHILIIT LQAFIFMVLT IVYLSMASEE H

 >lwec6_0 shuffled
 GMPISVLLFK PPEVLLVFLL SVMGTNFPAW GGFIMKGFKI VSFVGWVRFV AVAGHLALYK
 ITRDVNIVKS AVFGSALLHP LLLQLSELNL VFVNLLNIKI RTAYVHGMTL LSHIPLFPAS
 GEGVFSDMLM IITWNSASVL SGLDMFANIA LLGNPLLMTN IVIILQRKFI ATTKFSLADI
 HLHKQYSWDG MMSHTLIIFS ALELWVQNGD IFIPLNEYIL PFTLYVPNWL ITQALVVALV
 ELPGQQIDAE PLFLLPIPFS EKTWYGDIMF L

PRSS34 - 1000 shuffles;  uniform shuffle
 unshuffled s-w score: 178; bits(s=178|n_l=271): 34.8 p(178) < 2.005e-06
For 10000 sequences, a score >= 178 is expected 0.02005 times


 >lwec6_0 shuffled window: 10
 EDSMANTMPQ HQNILGYHLN DLRTSDFVLL FTQAPWPTPN SMNIDIVFSF VLLVLLFFGL
 SRGAVKATKS EQVTGIKFAP VVSGVILGFN HDKGMSLYKK VLPIIFLAAT DWLMNFVLLM
 IIDLYLLAPP ERVGHPLLAL APNVVVSVDT MLFLIGSALV IFSLMKGIKY TTIFGLEKGL
 QAWNFFPHIP NLSVEVGLLI GLPVRSSLKL MFLELAGNGY PFGILILILA SLINVWPWQW
 IAIIWTIFHL VQMTFFLAIL VSESELMIYA H

PRSS34 - 1000 shuffles;  window shuffle, window size: 20
 unshuffled s-w score: 178; bits(s=178|n_l=271): 34.5 p(178) < 2.601e-06
For 10000 sequences, a score >= 178 is expected 0.02602 times
```

46

## Random shuffles with prss




47

## Statistical estimates from random shuffles

| *algorithm* | closely related dopamine D2[a] | related thromboxane A2[b] | distantly related cAMP-1[c] | unrelated cytochrome oxidase[d] |
|---|---|---|---|---|
| Smith-Waterman | $3x10^{-9}$ | $2x10^{-4}$ | 0.01 | 0.57 |
| PRSS[e] | $8x10^{-10}$ | $10^{-4}$ | 0.007 | 0.45 |
| PRSS (window=20)[e] | $8x10^{-8}$ | 0.001 | 0.23 | 3.0 |

[a]D2DR_HUMAN, [b]TA2R_MOUSE, [c]CAR1_DICDI, [d]APPC_ECOLI
[e]after 1000 shuffles

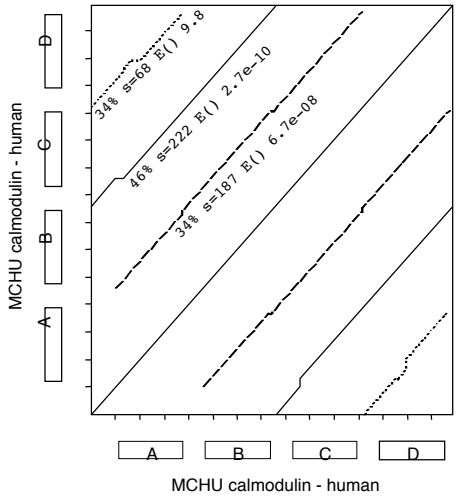
48

# Local alignments - calmodulin

```
 46.1% identity in 76 aa overlap (1-76:77-149); score:  222 E(10000): 2.7e-10
               10        20        30        40        50        60
mchu    MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADG
        : :  .::.:   .::: .::::::.: :.. ::  :: .::.. :. :...:: :.: ::
mchu    MKDTDSEEEI---REAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDG
          80           90        100       110       120       130
                70
mchu    NGTIDFPEFLTMMARK
        .: ... ::. ::. :
mchu    DGQVNYEEFVQMMTAK
            140

  34.3% identity in 105 aa overlap (11-111:47-147); score:  187 E(10000): 6.7e-08
                20        30         40        50        60
mchu    AEFKEAFSLFDKDGDGTITTKELGTVM-RSLGQNPTEAELQDMINEVDADGNGTIDFPEF
        ::... ..  : ::.:::   :. :.: :.. .. .: :... .   : :::: :.  :.
mchu    AELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAEL
          50        60        70        80        90       100
          70        80        90       100       110
mchu    ---LTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMT
           .: ...:. : . .: ::::    : ::.: ..  :. ..::
mchu    RHVMTNLGEKLTDEEVDEMIREA----DIDGDGQVNYEEFVQMMT
         110       120           130       140

  34.2% identity in 38 aa overlap (1-37:113-146); score:   68 E(10000):    9.8
               10        20        30
mchu    MADQLTEEQIAEF-KEAFSLFDKDGDGTITTKELGTVM
        ....::.:.. :. .::    : :::: .. .:.  .:
mchu    LGEKLTDEEVDEMIREA----DIDGDGQVNYEEFVQMM
             120           130       140
```



# Repeated domains with 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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Sequence, Profile, and Structure Comparison

51

## *BLAST* and *FASTA*
## Which program when?

Blast for proteins
Blast for speed
FASTA for DNA
FASTA for frameshifts
FASTA for accurate statistics
(protein and coding DNA)
SSEARCH for optimal
(be careful with PSI-BLAST)

52

## *Comparison programs in the FASTA3 package*

| | |
|---|---|
| `fasta` | Compare a protein sequence to a protein sequence database or a DNA sequence to a DNA sequence database using the FASTA algorithm. Search speed and selectivity are controlled with the *ktup* (wordsize) parameter. For protein comparisons, *ktup = 2* by default; *ktup =1* is more sensitive but slower. For DNA comparisons, *ktup=6* by default; *ktup=3* or *ktup=4* provides higher sensitivity; *ktup=1* should be used for oligonucleotides (DNA query lengths <= 20). |

| | |
|---|---|
| `ssearch` | Compare a protein sequence to a protein sequence database or a DNA sequence to a DNA sequence database using the Smith-Waterman algorithm. ssearch3 is about 10-times slower than FASTA3, but is more sensitive for full-length protein sequence comparison. |

| | |
|---|---|
| `fastx/ fasty` | Compare a DNA sequence to a protein sequence database, by comparing the translated DNA sequence in three frames and allowing gaps and frameshifts. `fastx3` uses a simpler, faster algorithm for alignments that allows frameshifts only between codons; `fasty3` is slower but produces better alignments with poor quality sequences because frameshifts are allowed within codons. |

53

## Which program when?

| Problem | Program | Explanation | Alternate |
|---|---|---|---|
| Identify unknown protein | (1) `fasta3` | General protein comparison. Use *ktup=2* (the unknown default) for speed; *ktup=1* for a more sensitive search. Search first against the smallest library likely to contain a homolog (i.e. SwissProt rather than Genpept). | `blastp/` |
| | (2)`ssearch3` | 10-50-fold slower than `fasta3` faster on Macs, but provides maximum sensitivity. No advantage for DNA comparisons. | `fasta3/ blastp` |
| | (3)`tfastx3/ tfasty3` | If a homolog cannot be found in the protein databases, check the DNA databases with `tfastx3` or `tfasty3`. `tfasty3` provides more accurate alignments, but is about 33% slower. | `tblastn/ tfasta`[a] |
| Identify structural DNA sequence | `fasta3` | If the DNA sequence encodes a protein, use protein sequence comparison first, then try translated protein sequence comparison (`fastx3`/`fasty3`). For repeated DNA sequences or structural RNAs, search first with *ktup=6* (the default), then *ktup=3*. Search with *ktup< 3* only for very short sequences (PCR primers). | `blastn` |
| Identify EST sequence | `fastx3/ fasty3` | Protein sequence comparison is far more sensitive than DNA comparison, so check first to see if the EST encodes a product homologous to a known protein. Current version searches forward strand only, so use `fastx3 -i` as well. | `fasta3/ blastx/ tblastx` |
| Confirm statistical significance | `prss3` | Use 500-2000 shuffles, and remember to normalize the statistical significance to the size of the database originally searched (typically 10,000 - 100,000 sequences). | |

[a]No longer recommended.

54

# Comparison of BLAST2 and FASTA3 Programs

| Program | | |
|---|---|---|
| BLAST | FASTA | Function |
| `blastp` | `fasta3` | General protein sequence similarity searches. **`blastp`** is faster and can show alignments between several domains in the same sequence. **`fasta3`** displays a Smith-Waterman final alignment and produces more accurate statistical estimates in some cases. |
| `blastn` | `fasta3` | DNA sequence comparison. **`blastn`** is highly optimized for speed; it uses a fixed word size (11 nucleotides) and scoring matrix that are inappropriate for some problems (e.g. searching for PCR primer matches). |
| `blastx` | `fastx3/ fasty3` | Compare a translated DNA to a protein sequence database. While **`blastx`** does six independent searches (one for each of the six frames), **`fastx3`** and **`fasty3`** effectively does a single forward (or backward) search, which allows frameshifts in computing the similarity score and alignments. As a result, **`fastx3`** and **`fasty3`** are more sensitive and can produce much better alignments than **`blastx`** when the DNA sequence has frameshift errors. |
| `tblastn` | `tfastx3/ tfasty3` | Compare a protein sequence to a DNA sequence database, translating in the three forward and reverse frames. Again, **`tfastx3`** and **`tfasty3`** provide more accurate alignments than **`tblastn`** when the DNA sequences have frameshift errors. |
| | `tblastx` | Compare a DNA query sequence to a DNA library, translating both sequences in all six frames and scoring using a protein substitution matrix (BLOSUM62). **`fasta3`** with *ktup=6* (the default) provides a similar function, but does not use a protein scoring matrix. |

55

# Scoring Matrices and Gap-penalties - *BLAST* vs *FASTA*

*BLAST*

- default scoring matrix: BLOSUM62 (1/2 bit)
- default gap penalty:
  -11 (open)/-1(extend)
  (lowest -9/-1, -8/-2)

*FASTA*

- default matrix:
  BLOSUM50 (1/3 bit)
- default gap penalty:
  old: -12 (first residue)/-2
  = new: -10 (open)/-2(ext)
- BLOSUM62 -7/-1
- PAM120 -16/-4
- PAM20 -24/-4

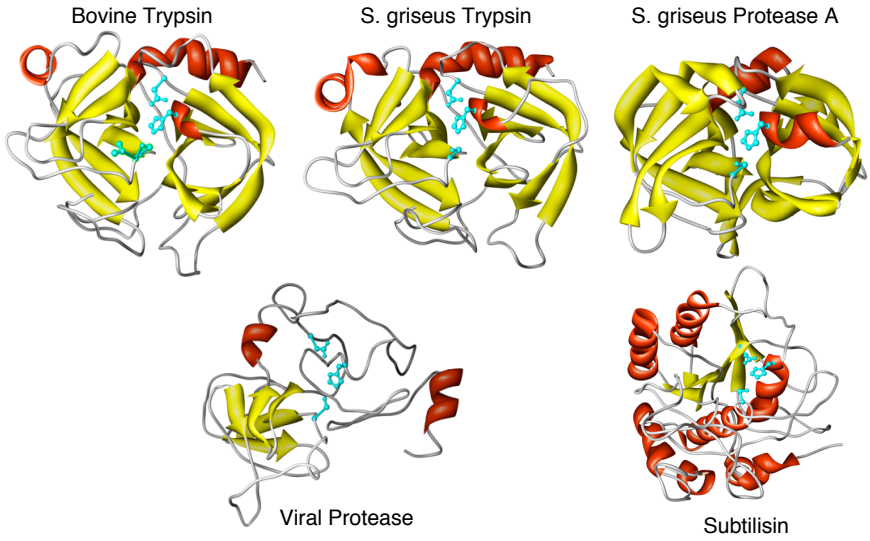
56

## 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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Sequence, Profile, and Structure Comparison

57

## Homologs, Topologs, and Convergence



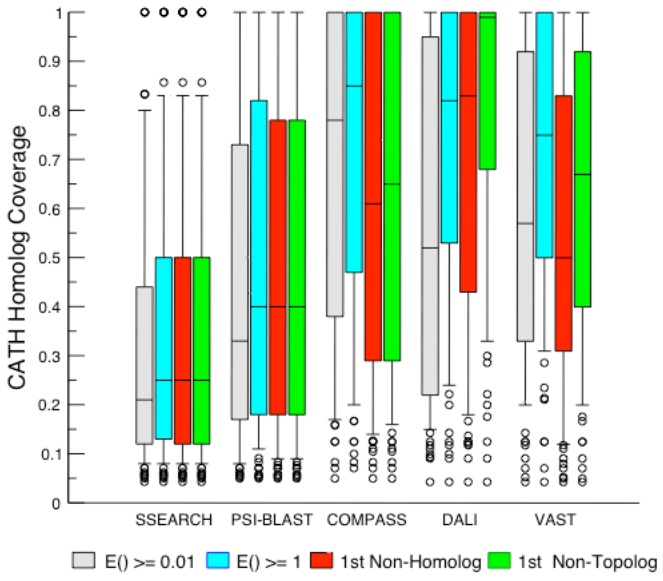

58

## Homlogy, Similarity, and Convergence – Serine Proteases

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

59

## Homologs found by different search methods




60

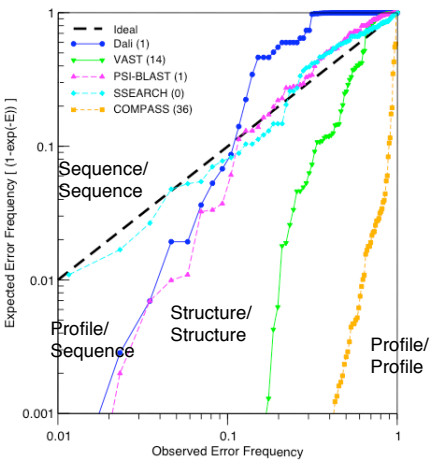
## Inferring Homology from Statistical Significance

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

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

61

Accuracy of statistical estimates




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

62

## Structure Comparison Statistics

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

63

## Sequence Similarity - Conclusions

- 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
- Protein 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
- Searching smaller libraries improves sensitivity
- Structural and profile significance estimates are considerably less accurate that sequence comparison statistics

64

## Discussion (exam) questions

1. What is the difference between similarity and homology? When does high identity not imply homology? What conclusions can be drawn from homology?
2. What is the range of an expectation value (E()-value)? If you compare a sequence to 50,000 random(unrelated) sequences, what should the expectation value for the highest of the 50,000 similarity scores be (on average)?
3. In a sequence similarity database search, you identify a statistically significant similarity (*E*()<0.005), but the alignment is relatively short (50 aa). How might you determine whether the alignment reflects a genuine homology, or a random sequence match?
4. What scoring matrix should be used to identify protein orthologs that have diverged over the past 100 My (e.g. human/mouse)?
5. When the *M. janaschii* genome was first sequenced, Venter and his colleagues stated that almost 60% of the open reading frames (proteins or genes) were novel to this organism. (For eubacterial like *E. coli* or *H. influenzae*, a similar number would be 20 - 40%.) On what would they base such a statement? Is it likely to be correct?

65