

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

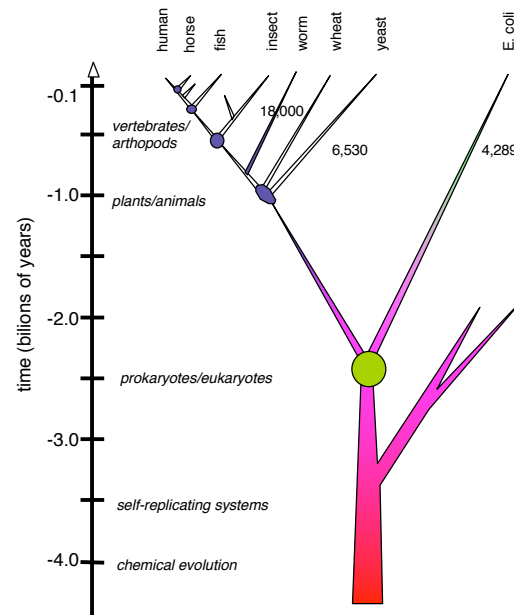
- Always compare Protein Sequences
- Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
- Homologous proteins share common structures, but not necessarily common functions
- Sequence statistical significance estimates are accurate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

2

Protein Evolution and Sequence Similarity

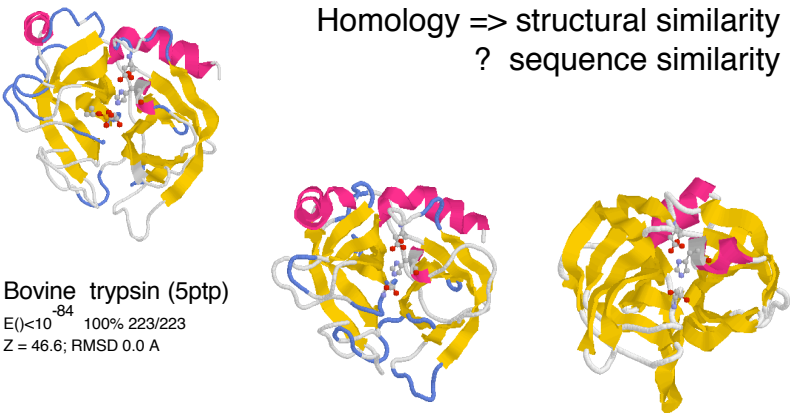
- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST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

3



4

Homology => structural similarity
? sequence similarity



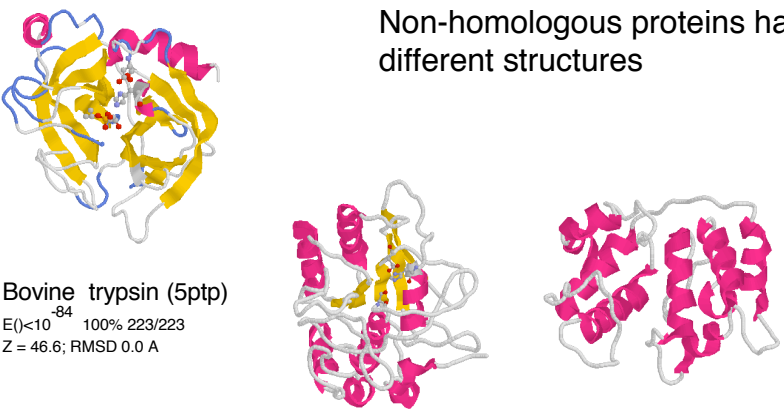
Bovine trypsin (5ptp)
 $E() < 10^{-84}$ 100% 223/223
 Z = 46.6; RMSD 0.0 Å

S. griseus trypsin (1sgt)
 $E() < 10^{-19}$ 36% 226/223
 Z = 31.3; RMSD 1.6 Å

S. griseus protease A (2sga)
 $E() < 2.6 \times 10^{-25}$ 25% 199/181
 Z = 13.7; RMSD 2.6 Å

5

Non-homologous proteins have
different structures

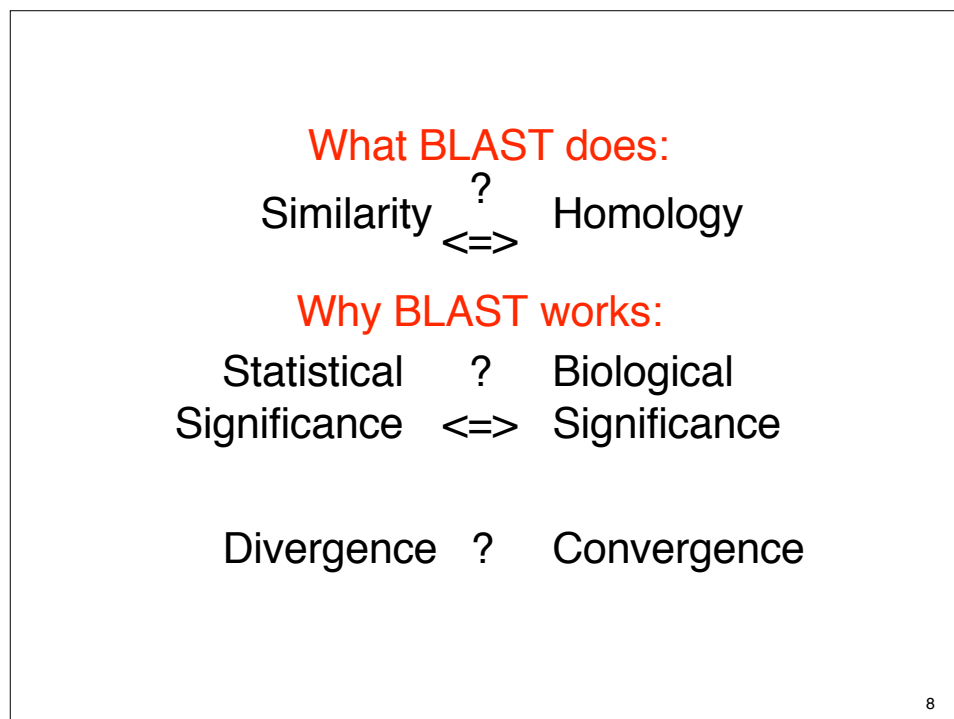
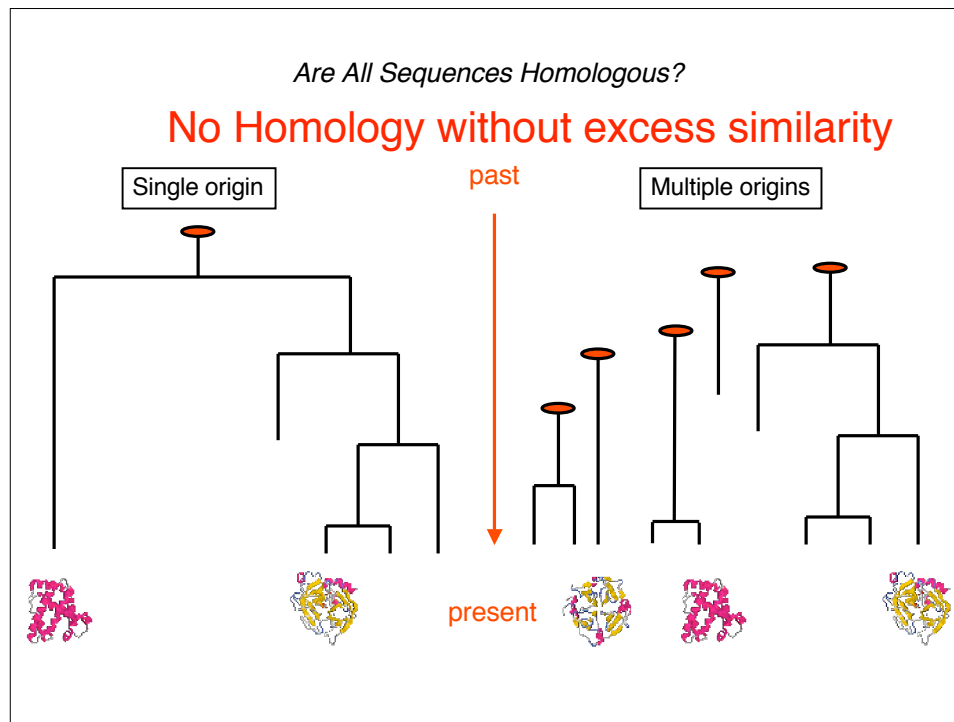


Bovine trypsin (5ptp)
 $E() < 10^{-84}$ 100% 223/223
 Z = 46.6; RMSD 0.0 Å

Subtilisin (1sbt)
 $E() < 280$ 25% 159/275
 Z < 2

Cytochrome c4 (1etp)
 $E() < 5.5$ 23% 171/190
 Z < 2

6



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

^aBillions 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	Prokaryotic/Eukarotic
Triosphosphate isomerase	3	6	Archaen
Glutamate dehydrogenase	1	18	?

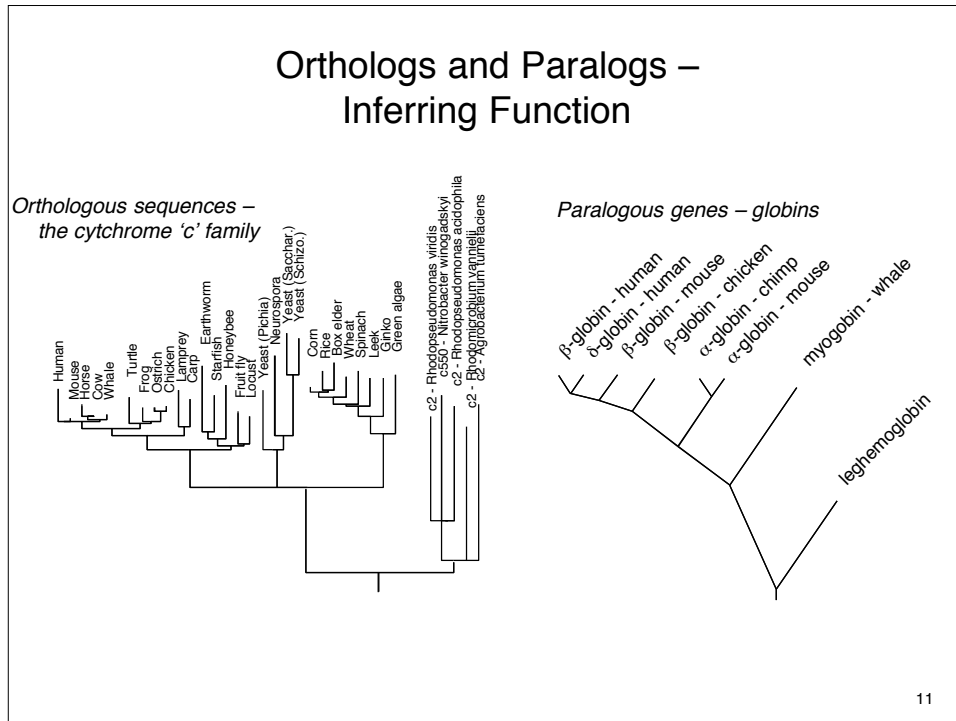
^aPAMs, point accepted mutations. ^bUseful lookback time, 360 PAMs, 15% identity. ^cMillions of years. ^dBillions of years.

9

Human proteins in E. coli

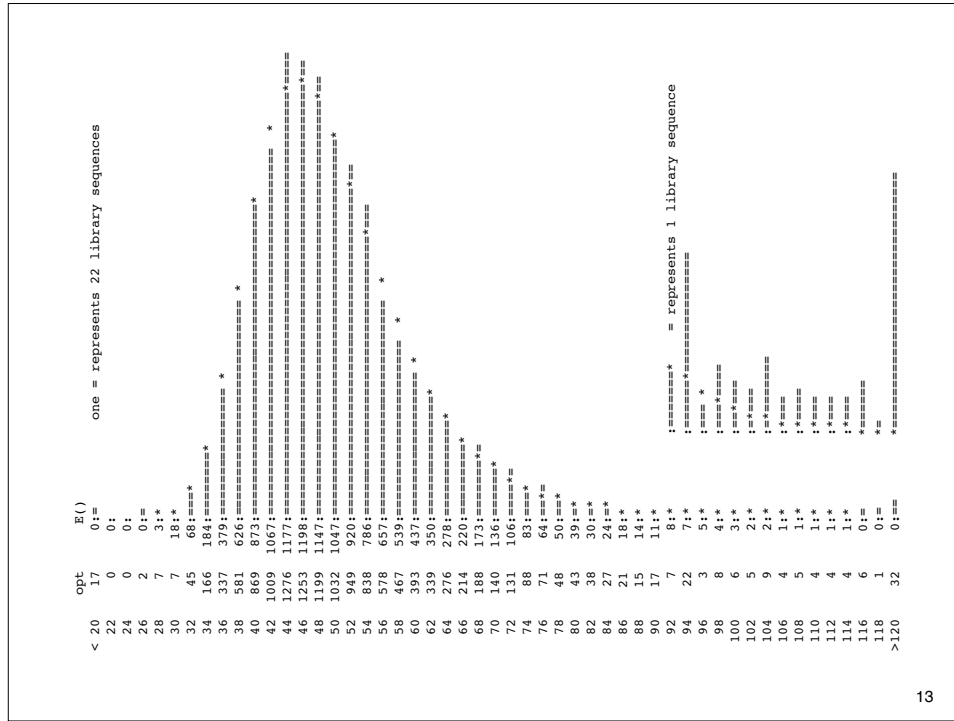
Description	length	E(4400)	%id	E(1700)	
IRE1_HUMAN iron-resp. element bind. prot.	1	837			
acnA aconitate hydratase	1	891	1.6e-195	53.4	1.2e-14
PHS_HUMAN glycogen phosphorylase	1	847			
glgP α-glucan-phosphorylase	1	815	4.0e-181	49.8	-
MUTA_HUMAN methylmalonyl-coA mutase	1	750			
sbm chromosome initiation factor	1	714	1.4e-178	59.3	-
G6PI_HUMAN glucose 6-P isomerase	1	558			
pgi glucose 6-P isomerase	1	549	2.2e-164	64.7	2.1e-14
CPSM_HUMAN carbamoyl-P isomerase	1	1500			
carB carbamoyl-P isomerase	1	1073	7.2e-162	40.3	2.2e-90
SYV_HUMAN valyl-tRNA synthetase	1	1263			
valS valyl-tRNA synthetase	1	951	2.2e-153	40.1	9.5e-72
ODO1_HUMAN 2-oxoglutarate DH E1	1	1002			
sucA 2-oxoglutarate DH E1	1	933	2.9e-143	39.1	-
GR75_HUMAN mito. stress-70 prot.	1	679			
dnaK DNA K protein (HSP70)	1	638	3.9e-138	60.4	-
DHSA_HUMAN succinate DH	1	664			
sdhA succinate DH	1	588	1.2e-126	55.2	1.3e-73
ATPB_HUMAN ATP synth. β-chain	1	529			
atpD ATP synth. F1 β-subunit	1	460	3.6e-123	71.7	9.5e-23
BGLR_HUMAN β-glucuronidase	1	651			
uida β-D-glucuronidase	1	603	1.4e-118	45.1	-
SYA_HUMAN alanyl-tRNA synthetase	1	968			
alaS alanyl-tRNA synthetase	1	876	4.7e-116	39.1	1.6e-38

10



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
- Alignment Algorithms
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Large scale comparison

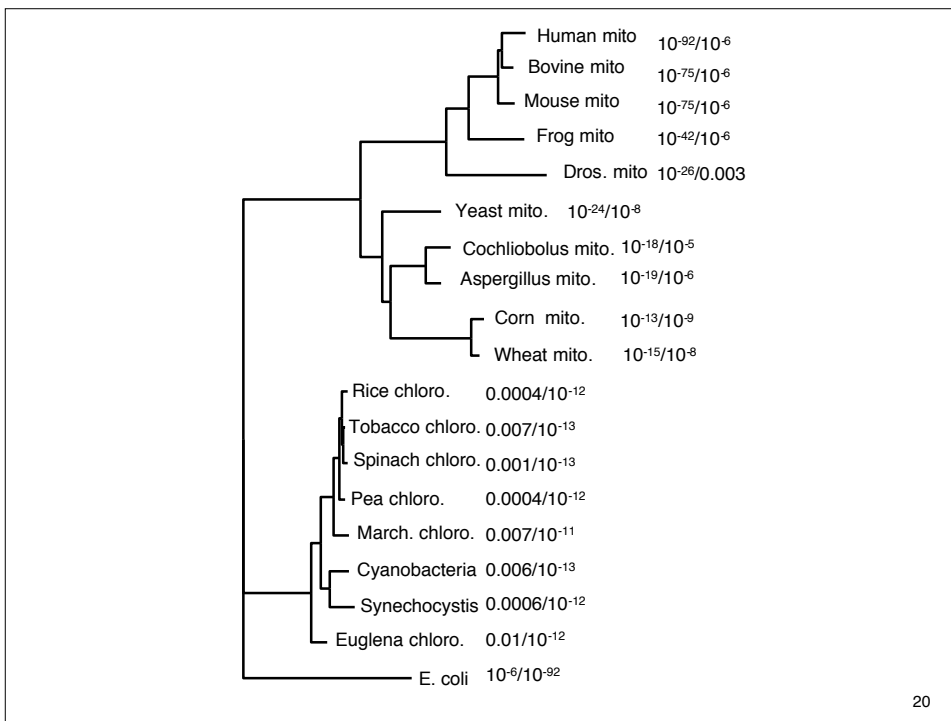
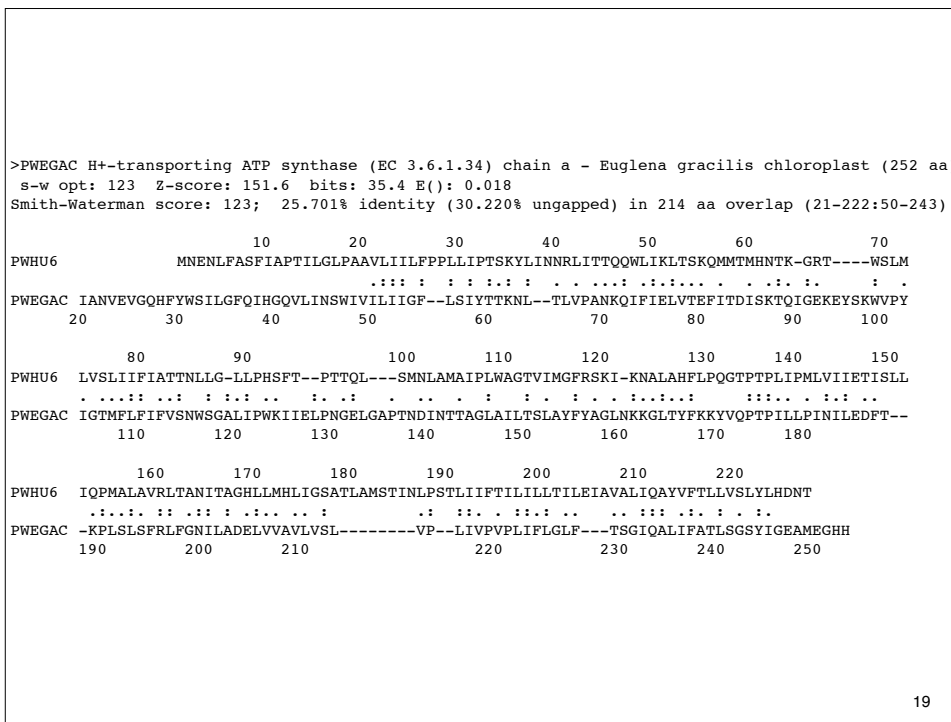


13

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



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?
- Large scale 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.leguminosarum <i>gstA</i>	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 mallei acetate 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

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
- **Alignment Algorithms**
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- BLAST and FASTA - which program when?
- Large scale comparison

23

Algorithms for Biological Sequence Comparison

algorithm	value calculated	scoring matrix	gap penalty	time required	
Needleman- and Wunsch	global similarity	arbitrary	penalty/gap q	$O(n^2)$	Needleman Wunsch, 1970
Sellers	(global) distance	unity	penalty/residue r k	$O(n^2)$	Sellers, 1974
Smith- Waterman, 1981 Waterman	local similarity	$S_{ij} < 0.0$	affine q + r k	$O(n^2)$	Smith and Gotoh, 1982
FASTA Pearson, 1985 Lipman, 1988	approx. local similarity	$S_{ij} < 0.0$	limited size q + r k	$O(n^2)/K$	Lipman and Pearson and
BLASTP 1990	maximum segment score	$S_{ij} < 0.0$	multiple segments	$O(n^2)/K$	Altshul et al.,
BLAST2.0 1997	approx. local	$S_{ij} < 0.0$	q+r k	$O(n^2)/K$	Altshul et al.,

24

The sequence alignment problem:

```

PMILGYWNVRL      PMILGYWNVRL      PM-ILGYWNVRL
:                : : : :      : : : :
PPYTIVYFPVRG     PPYTIVYFPVRG     PPYTIV-YFPVRG

PMILGYWNVRL      PMILGYWNVRL      PM-ILGYWNVRL
:                :. . . : : : : : : :
PPYTIVYFPVRG     PPYTIVYFPVRG     PPYTIV-YFPVRG

  P M I L G Y W N V R G L
P X
P X
Y  x          X x
T                x x
I    X x      x   x
V  x x x     X   x
Y  x        X x
F  x x x    x x
P X
V  x x x      X   x
R                X
G                X

Global:
-PMILGYWNVRL
:. . . : : :
PPYTIVYFPVRG-

Local:
AAAAAAPMILGYWNVRLBBBBB
:. . . : : :
XXXXXXXXPPYTIVYFPVRGYYYYYY
    
```

25

Algorithms for Biological Sequence Comparison

		Global	Local	Distance
HBHU vs HBHU	Hemoglobin beta-chain - human	725	725	0
	HAHU Hemoglobin alpha-chain - human	314	322	152
	MYHU Myoglobin - Human	121	166	212
	GPYL Leghemoglobin - Yellow lupin	8	43	239
	LZCH Lysozyme precursor - Chicken	-107	32	220
	NRBO Pancreatic ribonuclease - Bovine	-124	31	280
	CCHU Cytochrome c - Human	-160	26	321
MCHU vs MCHU	Calmodulin - Human	671	671	0
	TPHUCS Troponin C, skeletal muscle	395	438	161
	PVPK2 Parvalbumin beta - Pike	-57	115	313
	CIHUH Calpain heavy chain - Human	-2085	100	2463
	AQJFNV Aequorin precursor - Jelly fish	-65	76	391
	KLSWM Calcium binding protein - Scallop	-89	52	323
QRHULD vs EGMSMG	EGF precursor	-591	655	2549

26

Smith-Waterman

		N	L	P	Y	L	I
Q	○	○	○	○	○	○	○
V	○	○	○	○	○	○	○
P	○	○	○	○	○	○	○
L	○	○	○	○	○	○	○
V	○	○	○	○	○	○	○
E	○	○	○	○	○	○	○
I	○	○	○	○	○	○	○

1. score every cell:

$$S_{x,y} = \max \left\{ \begin{array}{l} S_{x-1,y-1} + \text{match}_{xy} \\ S_{x,y-1} - \text{gappen} \\ S_{x-1,y} - \text{gappen} \\ 0 \end{array} \right.$$

Global and Local Alignment Paths

Global

	A	B	D	D	E	F	G	H	I
A	\	\	\	\	\	\	\	\	\
B	1	-1	-1	-1	-1	-1	-1	-1	-1
D	-1	2	0	-2	-2	-2	-2	-2	-2
E	-1	0	3	1	-1	-3	-3	-3	-3
G	-1	-2	1	2	2	0	-2	-4	-4
K	-1	-2	-1	0	1	1	1	-1	-3
H	-1	-2	-3	-2	-1	0	0	0	-2
I	-1	-2	-3	-4	-3	-2	-1	1	-1
	-1	-2	-3	-4	-5	-4	-3	-1	2

Optimum global alignment (score: 2)
 A B D D E F G H I (top)
 A B D - E G K H I (side)
 or A B - D E G K H I

Local

	A	B	D	D	E	F	G	H	I
A	\	\	\	\	\	\	\	\	\
B	1	0	0	0	0	0	0	0	0
D	0	2	0	0	0	0	0	0	0
E	0	0	3	1	0	0	0	0	0
G	0	0	1	2	2	0	0	0	0
K	0	0	0	0	1	1	1	0	0
H	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	1	0
	0	0	0	0	0	0	0	0	2

Optimal local alignment (score 3):
 A B D (top)
 A B D (side)

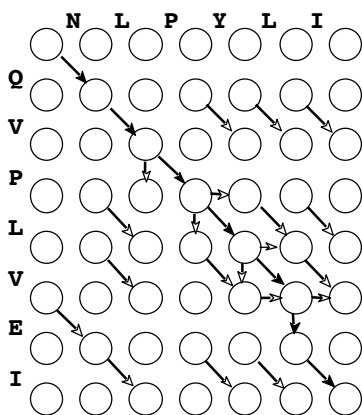
Algorithms for Global and Local Similarity Scores

```

Global:
  S(0,0) ← 0
  for j ← 1 to N do
    S(0,j) ← S(0,j-1) + σ(  $\bar{b}_j$  )
  for i ← 1 to M do
    [ S(i,0) ← S(i-1,0) + σ(  $a_i$  )
      for j ← 1 to N do
        S(i,j) ← max[S(i-1,j-1) + σ(  $\frac{a_i}{b_j}$  ), S(i-1,j) + σ(  $\frac{a_i}{-}$  ), S(i,j-1) + σ(  $\bar{b}_j$  ) ]
      ]
  write "Global similarity score is" S(M,N)

Local:
  best ← 0
  for j ← 1 to N do
    S'(0,j) ← 0
  for i ← 1 to M do
    [ S'(i,0) ← 0
      for j ← 1 to N do
        [ S'(i,j) ← max[0, S'(i-1,j-1) + σ(  $\frac{a_i}{b_j}$  ), S'(i-1,j) + σ(  $\frac{a_i}{-}$  ), S'(i,j-1) + σ(  $\bar{b}_j$  ) ]
          best ← max(S'(i,j), best)
        ]
      ]
  write "Local similarity score is" best
    
```

Smith-Waterman



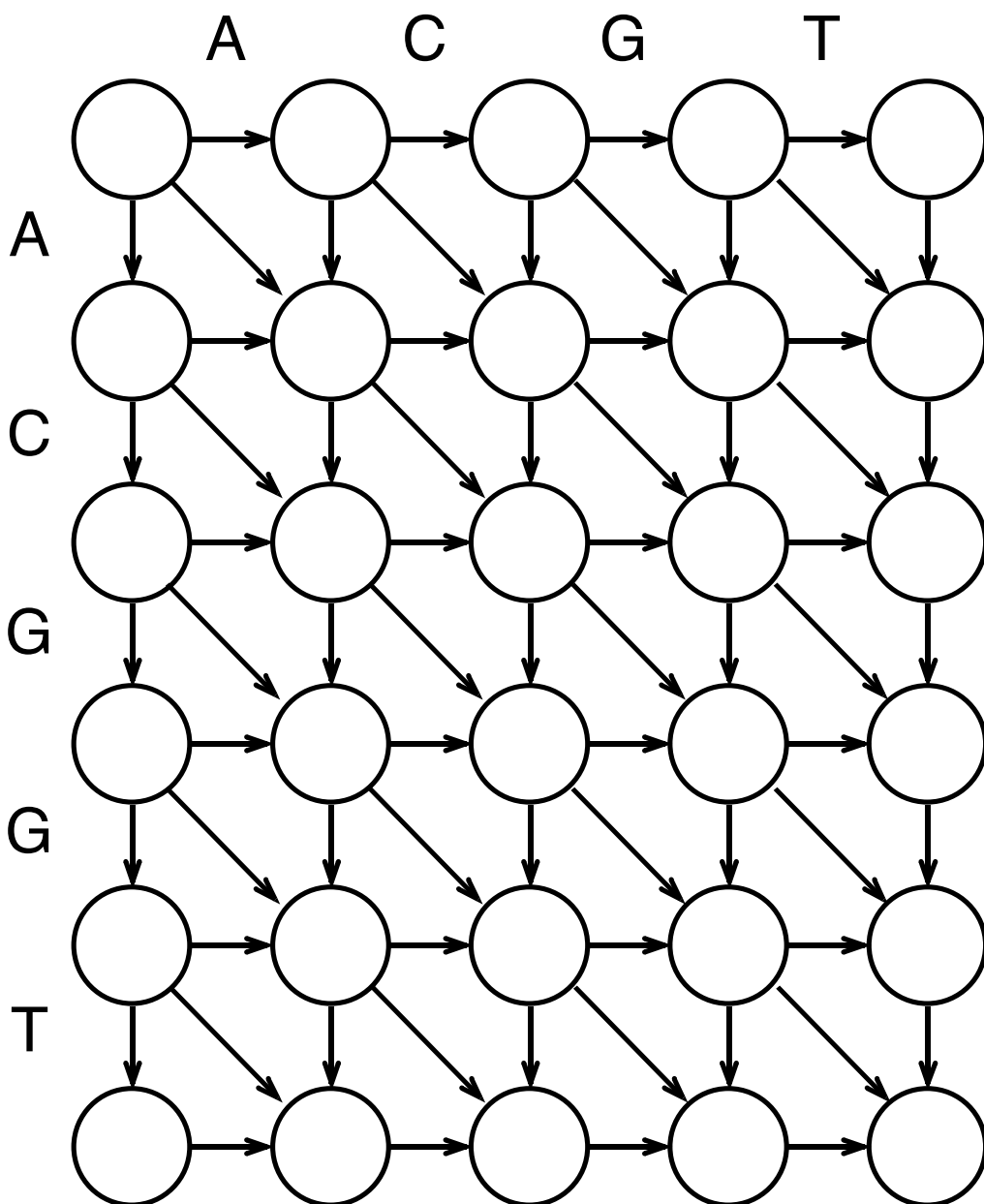
- score every cell:

$$S_{x,y} = \max \{ \begin{array}{l} S_{x-1,y-1} + \text{match}_{xy} \\ S_{x,y-1} - \text{gappen} \\ S_{x-1,y} - \text{gappen} \\ 0 \end{array} \}$$
- follow "traceback"

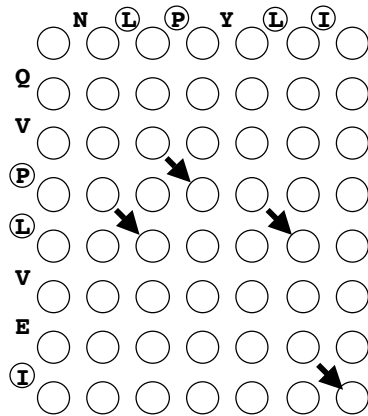
NLPYL-I
 .. : . :
 QVPLVEI

Outcome: one continuous, optimal gapped alignment

+1 : match
-1 : mis-match
-2 : gap



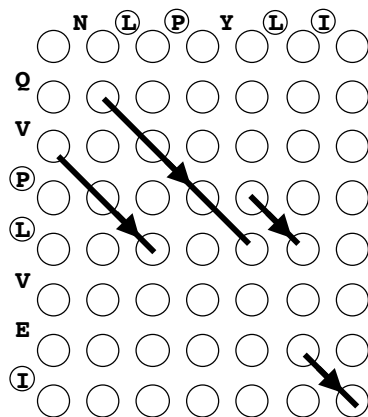
FASTA



1. Identify identical matches
(length = $ktup$)

31

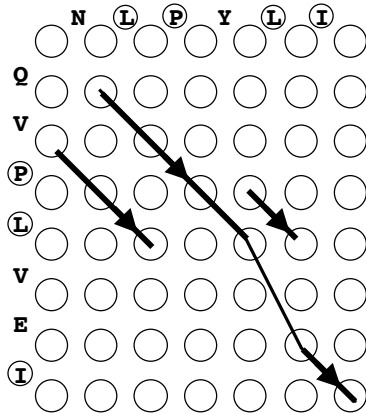
FASTA



1. Identify identical matches
(length = $ktup$)
2. Extend along diagonal
(local maximum)

32

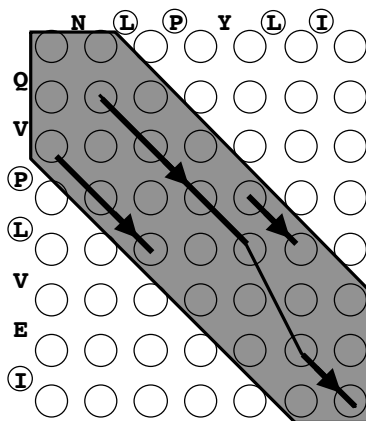
FASTA



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

33

FASTA



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

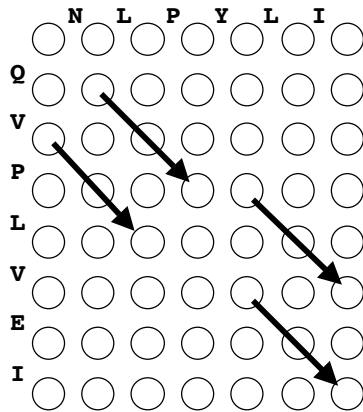
4. Banded Smith-Waterman

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

Outcome: one continuous, near-optimal gapped alignment

34

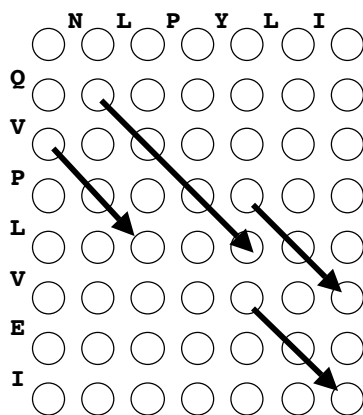
BLAST



1. neighborhood word hits (word length)

35

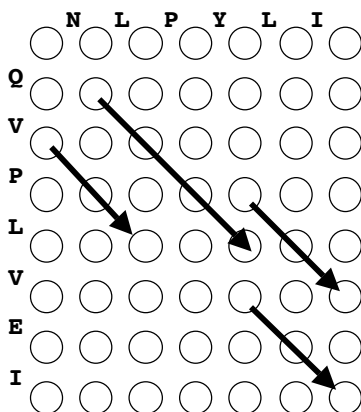
BLAST



1. neighborhood word hits (word length)
2. extend from diagonal ends (X-drop threshold)

36

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

37

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

38

Where do scoring matrices come from?

$$\lambda S = \ln \frac{q_{ij}}{p_i p_j}$$

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

$q_{R:N(40)} = 0.000435$

$p_R = 0.051$

$q_{R:N(250)} = 0.002193$

$p_N = 0.043$

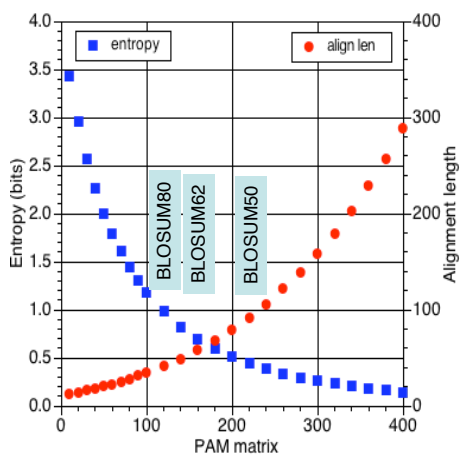
$\lambda_2 S_{ij} = \lg_2 (q_{ij}/p_i p_j)$ $\lambda_e S_{ij} = \ln (q_{ij}/p_i p_j)$ $p_R p_N = 0.002193$

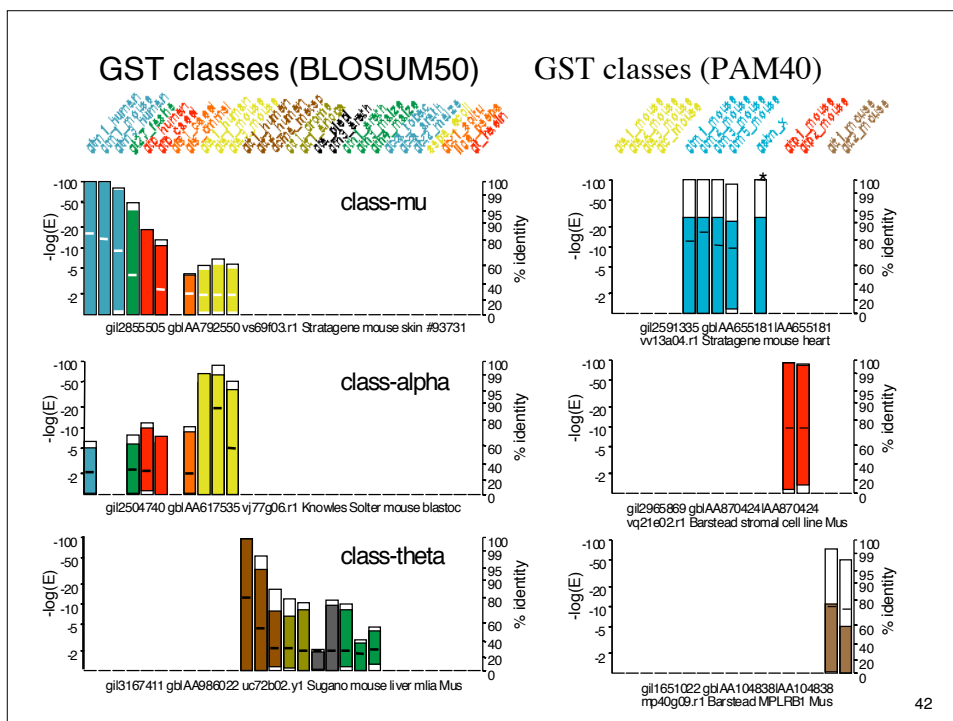
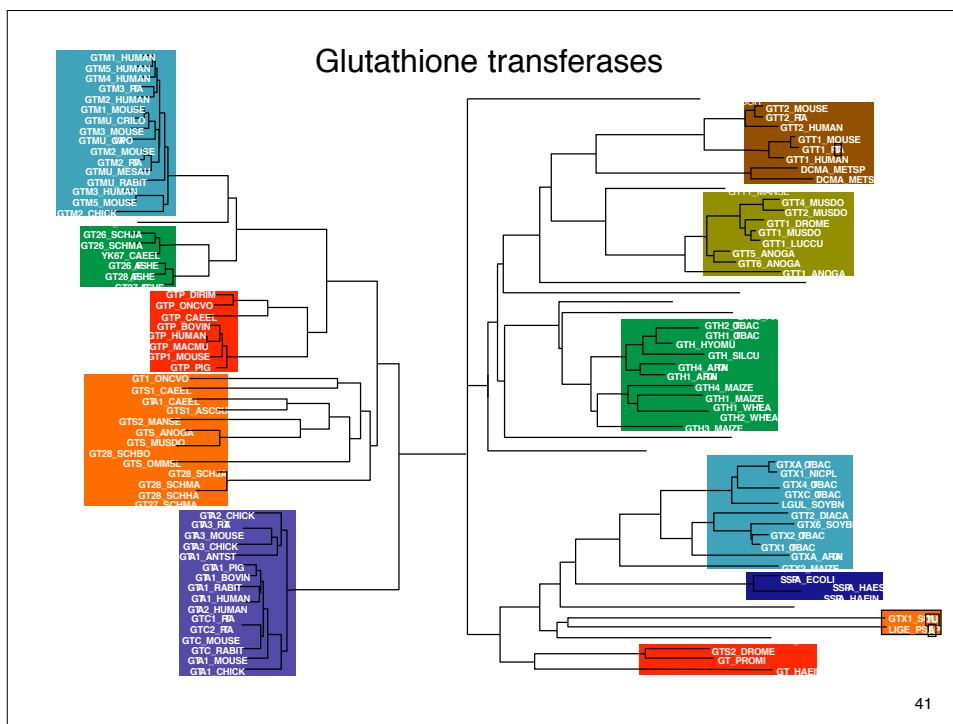
$\lambda_2 S_{R:N(40)} = \lg_2 (0.000435/0.00219) = -2.333$

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

$\lambda S_{R:N(250)} = \lg_2 (0.002193/0.002193) = 0$

PAM matrices and alignment length





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

43

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?
- Large scale comparison

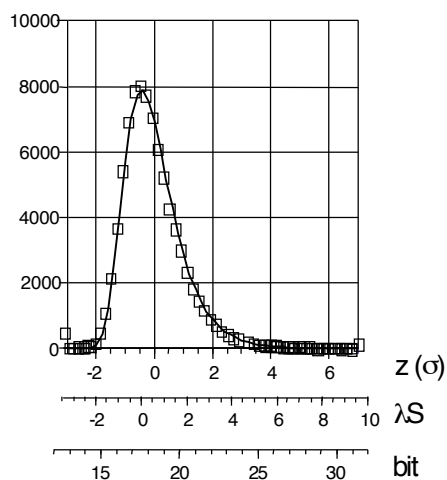
44

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

45

Extreme value distribution



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

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

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

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

$$E(S' > x \text{ ID}) = P D$$

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

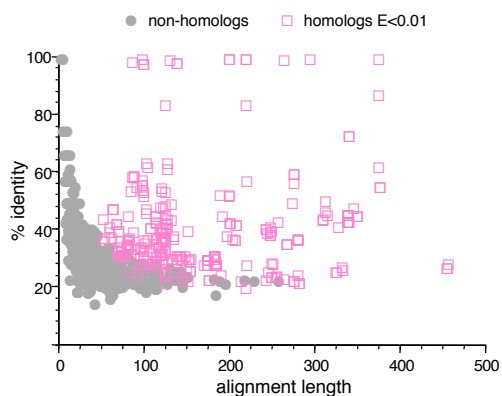
$$P(40 \text{ bits}) = 1.5 \times 10^{-7}$$

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

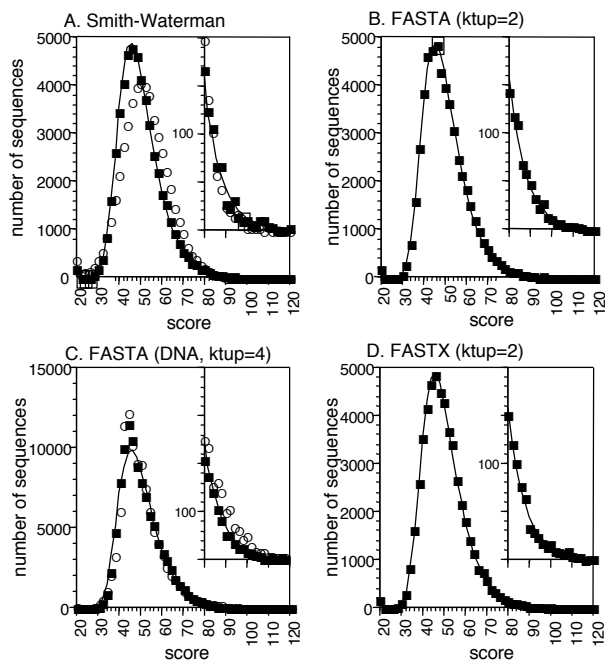
$$E(40 \mid D=2E6) = 0.3$$

46

Sequence identity and statistical significance



47



48

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 1	(593)	103	30.5	*9.9*	0.227	185

49

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

50

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

51

pseg removes low-complexity regions

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

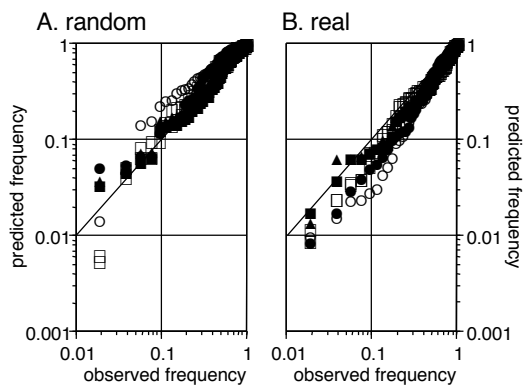
```

          1-8  MYPSVVRH
paagppppgpp 9-19
                20-131 IKFTIADTLERIKKEEFNQLQAQYHSIKLEC
                EKLSNEKTEMQRHYVMYEMSYGLNVEMHK
                QTEIAKRLNLTINQLLPLQADHQQVLQA
                VERAKQVTMQELNLIIGQQIHA
qqvppppppmq 132-143
                144-281 ALNPFALGATMGLPHGPQGLLNKPEHHR
                PDIKPTGLEPAAAEERLNSVSPADREKY
                RTRSPLDIENDSKRRKDEKLOEDEGEKSDQ
                DLVVDVANEMESHSPRNGEHVSMVDRDRE
                SLNGERLEKPSSSGIKQE
rppsrsgsssrstps 282-297
                298-310 LKTKDMEKPGTPG
akartptnaaapgvnpk 311-330
qmmpqppppagypgpyqrp 331-351
                352-719 DPYQRPPSDPAYGRPPMPYDPHAHVRTNG
                IPHPSALTGGKPAYSFHMNGEGLQVPVFP
                PDALVGVGIPRHARQINTLSHGEVVCVAVTI
                SNPTKYVYTGKGCVKVWDISQPGNKNPVS
                QLDCLQRDNYIRSVKLLPDGRTLIVGGEAS
                NLSIWDLASPTPRIKAELTSAAAPACYALAI
                SPDSKVCFSCCSDGNIAVWDLHNEILVRQF
                QGHTDGASCIDISPDGSRSLWTGGLDNTVRS
                WDLREGRLQQHDFSSQIFSLGYCPTGDWL
                AVGMENSHVEVLHASKPKYQLHLHESCVL
                SLRFAACGKWFVSTGKDLLNAWRTPYGAS
                IFQSKETSSVLSCDISTDDKIIVTGS GDKK
                ATVYEVII

```

52

Smith-Waterman statistics



53

Statistical estimates from random shuffles

<i>algorithm</i>	closely related dopamine D2 ^a	related thromboxane A2 ^b	distantly related cAMP-1 ^c	unrelated cytochrome oxidase ^d
Smith- Waterman	3×10^{-9}	2×10^{-4}	0.01	0.57
PRSS ^e	8×10^{-10}	10^{-4}	0.007	0.45
PRSS (window=20) ^e	8×10^{-8}	0.001	0.23	3.0
fasta, <i>ktup=1</i>	3×10^{-9}	10^{-4}	0.02	0.39
fasta, <i>ktup=2</i>	2×10^{-6}	10^{-4}	2.2	0.36
BLASTP	2×10^{-22}	0.07	>1.0	>1.0

^aD2DR_HUMAN, ^bTA2R_MOUSE, ^cCAR1_DICDI, ^dAPPC_ECOLI
^eafter 1000 shuffles

54

prss - uniform and window shuffle

```
>lweec6 H+-transporting ATP synthase (EC 3.6.1.34) protein 6 - Escherichia coli
MASENMTPOD YIGHHLNQLD LDLRTFSLVD PQNPPATFWT INIDSMFFSV VLGLLFLVLF
RSVAKKATSG VPGKFQTAIE LVIGFVNGSV KDMYHGKSKL IAPLALTIFV WVFLMNLMDL
LPIDILPYTA EHVGLPALR VVPSADVNVV LSMALGVFIL ILFYSIKMKG IGGFTKELTL
QPFNHWFAPV VNLILEGVSL LSKPVSLGLR LFGNMYAGEL IFILIAGLLP WWSQWILNVP
WAIFHILIT LQAFIFMVLV IVYLSMASEE H
```

```
>lweec6_0 shuffled
GMPISVLLFK PPEVLLVPLL SVMGTNPPAW GGFIMKGFPI VSFVGVVRFV AVAGHLALYK
ITRDVNIKVS AVFGSALLHP LLLQLSELNL VFNLLNIKI RTAYVHGRTL LSHIPLFPAS
GEGVFSMDML IITWNSASVL SGLDMFANIA LLGNPLMTN IVIILQRKFI ATTKFSLADI
HLHKQYSWDG MMSHTLIIFS ALELVQVQND IFIPLNEYIL PFTLYVNPWL ITQALVVALV
ELPGQQIDAE PLFLPLPIPS ERTWYGDIMF L
```

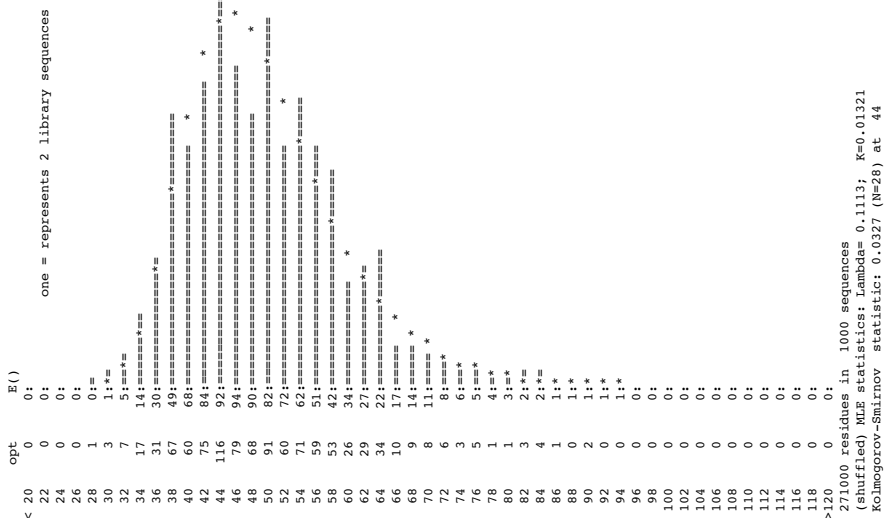
PRSS34 - 1000 shuffles; uniform shuffle
 unshuffled s-w score: 178; bits(s=178|n_l=271): 34.8 p(178) < 2.005e-06
 For 10000 sequences, a score >= 178 is expected 0.02005 times

```
>lweec6_0 shuffled window: 10
EDSMANTMPO HQNILGYHLN DLRTSDFVLL FTQAPWPTPN SMNIDIVFSF VLLVLLFFGL
SRGAVKATKS EQVTGKIFAP VVSGVILGFN HDKGMSLYKK VLPPIFLAAT DWLMNFVLLM
IIDLYLLAPP ERVGHPLLAL APNVVVSVDV MLFLIGSALV IFSLMKGIKY TTIFGLEKGL
QAWNFFPHIP NLSVEVGLLI GLPVRSSLKL MFLELAGNGY PFGILILILA SLINVWPQWQ
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

55

Random shuffles with prss



56

Statistical estimates from random shuffles

<i>algorithm</i>	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
fasta, <i>ktup</i> =1	3x10 ⁻⁹	10 ⁻⁴	0.02	0.39
fasta, <i>ktup</i> =2	2x10 ⁻⁶	10 ⁻⁴	2.2	0.36
BLASTP	2x10 ⁻²²	0.07	>1.0	>1.0

^aD2DR_HUMAN, ^bTA2R_MOUSE, ^cCAR1_DICDI, ^dAPPC_ECOLI
^eafter 1000 shuffles

57

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  MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQDMINEVDADG
      : : . . . . : . . . . . . . . . . . . . . . . . . . . . . . . . . . .
mchu  MKDTSDEEET---REAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDG
      80      90      100      110      120      130

      70
mchu  NGTIDFPEFLTMMARK
      . : . . . . . . . . . . . . . . . . . . . . . . . . . . . .
mchu  DGQVNYEEFVQMMTAK
      140

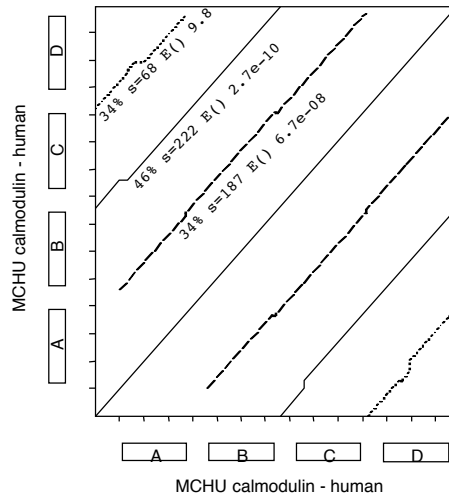
34.3% identity in 105 aa overlap (11-111:47-147); score: 187 E(10000): 6.7e-08
      20      30      40      50      60
mchu  AEFKEAFSLFDKDGDTITTKELGTVM-RSLGQNPTEAELQDMINEVDADGNGTIDFPEF
      : . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
mchu  AELQDMINEVDADGNGTIDFPEFLTMMARKMKDTSDEEETIREAFRVFDKDGNGYISAAEL
      50      60      70      80      90      100
      70      80      90      100      110
mchu  ---LTMMARKMKDTSDEEETIREAFRVFDKDGNGYISAAELRHVMT
      . : . . . . . . . . . . . . . . . . . . . . . . . . . . . .
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-KEAFSLFDKDGDTITTKELGTVM
      . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
mchu  LGEKLTDEEVDEMIREA----DIDGDGQVNYEEFVQMM
      120      130      140

```

58

Repeated domains with local alignments



59

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?**
- Large scale comparison

60

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)

61

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.

62

Comparison programs in the FASTA3 package

tfastx3/ Compare a protein sequence to a DNA sequence
tfasty3 database, calculating similarities with frameshifts
 to the forward and reverse directions

tfasta3 DO NOT USE - tfastx/y are preferred because
 they calculate similarity over frameshifts

fastf3/ Compare a mixed peptide sequence to a protein
tfastf3 sequence database. Deconvolute the mixture of
 peptides produced by CNBr and sequenced
 without separation

fasts3/ Search with unordered short (4 - 10 residue)
tfasts3 sequences, as obtained from MS/MS

63

FASTA - library formats

```

FASTA      >comment
           sequence

Genbank    LOCUS
flatfile   DEFINITION . . .
           ORIGIN
           acgtacgtac

acgtacgtac

PIR/GCG    >P1;locusid
GCGBIN     description
           sequence mpmvlifgsk

BLAST1.4   *.psq,.phr,pid,
*.nsq,nhr,nid
BLAST2.0

EMBL       ID/DE/SQ

mysql      SELECT prot.gid, prot.seq
           FROM prot,swissprot
           WHERE
           prot.gid=swissprot.gid;
  
```

Library/library comparison in parallel

Programs: (use Parallel Virtual Machine - PVM or Message Passing Interface - MPI)

```

pv34compfa mp34compfa fasta34
pv34compfx mp34compfx fastx34
pv34comptfx mp34comptfx tfastx34
pv34compsw mp34compsw ssearch34
...
  
```

Library vs library comparison:

```

pv34compfa -m 9 -E 1e-3 -d 0
           query.sql swissprot
  
```

64

Which program when?	Problem	Program	Explanation	Alternative
Identify unknown protein		(1) <code>fasta3</code>	General protein comparison. Use <code>ktup=2</code> (the unknown default) for speed; <code>ktup=1</code> for a more sensitive search. Search first against the smallest library likely to contain a homolog (i.e. SwissProt rather than Genpept).	<code>blastp</code>
		(2) <code>ssearch3</code>	10-50fold slower than <code>fasta3</code> , but provides maximum sensitivity. No advantage for DNA comparisons.	<code>fasta3/blastp</code>
		(3) <code>tfastx3/ tfasty3</code>	If a homolog cannot be found in the protein databases, check the DNA databases with <code>tfastx3</code> or <code>tfasty3</code> . <code>tfasty3</code> provides more accurate alignments, but is about 33% slower.	<code>tblastn/ tfasta^d</code>
Identify structural DNA sequence		<code>fasta3</code>	If the DNA sequence encodes a protein, use protein sequence comparison first, then try translated protein sequence comparison (<code>fastx3/fasty3</code>). For repeated DNA sequences or structural RNAs, search first with <code>ktup=6</code> (the default), then <code>ktup=3</code> . Search with <code>ktup< 3</code> only for very short sequences (PCR primers).	<code>blastn</code>
Identify EST sequence		<code>fastx3/ fasty3</code>	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.	<code>fasta3/ blastx/ tblastx</code>
Identify new orthologs		<code>tfastx3/ tfasty3</code>	If possible, search EST sequences from the same species. Use low/close MDM20 scoring matrices to detect close relationships and avoid distant relationships. Confirm statistical significance.	<code>tblastn/ tblastx</code>

65

BLAST2 vs FASTA3	Program		Function
	BLAST	FASTA	
	<code>blastp</code>	<code>fasta3</code>	General protein sequence similarity searches. <code>blastp</code> is faster and can show alignments between several domains in the same sequence. <code>fasta3</code> displays a Smith-Waterman final alignment and produces more accurate statistical estimates in some cases.
	<code>blastn</code>	<code>fasta3</code>	DNA sequence comparison. <code>blastn</code> 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). <code>blastn</code> searches with both strands of a DNA sequence. <code>fasta3</code> does not; two searches (<code>fasta3</code> and <code>fasta3 -i</code>) are required. ^a
	<code>blastx</code>	<code>fastx3/ fasty3</code>	Compare a translated DNA to a protein sequence database. While <code>blastx</code> does six independent searches (one for each of the six frames), <code>fastx3</code> and <code>fasty3</code> effectively does a single forward (or backward) search, which allows frameshifts in computing the similarity score and alignments. As a result, <code>fastx3</code> and <code>fasty3</code> are more sensitive and can produce much better alignments than <code>blastx</code> when the DNA sequence has frameshift errors. <code>blastx</code> searches in the forward and reverse frames; <code>fastx3/fasty3</code> searches only in the forward or the reverse (<code>fasty3 -i</code>) frame.
	<code>tblastn</code>	<code>tfastx3/ tfasty3/ tfasta</code>	Compare a protein sequence to a DNA sequence database, translating in the three forward and reverse frames. Again, <code>tfastx3</code> and <code>tfasty3</code> provide more accurate alignments than <code>tblastn</code> or <code>tfasta</code> when the DNA sequences have frameshift errors.
		<code>tblastx</code>	Compare a DNA query sequence to a DNA library, translating both sequences in all six frames and scoring using a protein substitution matrix (BLOSUM62). <code>fasta3</code> with <code>ktup=6</code> (the default) provides a similar function, but does not use a protein scoring matrix.

66

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

67

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?
- Large scale comparison

68

Database driven large scale comparison

- Build relational database with ORF coordinates on contigs
- Missassembly: ORFs with high identity to several contigs
- Take query sequences from intergenic regions
- Search comprehensive databases

69

Intergenic FASTX hits (examples)

```

288>>@C:761 00050.Contig331 1134 aa
FASTX (3.39 May 2001) function [optimized, BL50 matrix (o=15:-5:-1)xS] ktup: 2
  join: 39, opt: 33, open/ext: -14/ -2 shift: -20, width: 16
The best scores are:
          bits E(99359) %_id  alen  an0  ax0  pn0  px0  anl  axl  fs
gi|1170693|KPRS_HAEIN RIBOSE-PHOSPHATE PYROPHO ( 315) [r] 206 3.1e-52 0.484 860 1869 1015 761 2655 32 311 0
gi|11132927|KPRS_BUCAI RIBOSE-PHOSPHATE PYROPH ( 315) [r] 205 5.9e-52 0.472 854 1872 1015 761 2655 31 310 0
gi|12229873|KPRS_HELPJ RIBOSE-PHOSPHATE PYROPH ( 318) [r] 201 9.2e-51 0.470 833 1869 1021 761 2655 41 312 0
gi|2829453|KPRS_HELPY RIBOSE-PHOSPHATE PYROPHO ( 318) [r] 201 1.3e-50 0.466 831 1869 1021 761 2655 41 312 0
gi|2506803|KPRS_ECOLI RIBOSE-PHOSPHATE PYROPHO ( 315) [r] 200 2.4e-50 0.458 831 1893 1015 761 2655 24 310 0
gi|1170694|KPRS_SALTY RIBOSE-PHOSPHATE PYROPHO ( 315) [r] 199 3.9e-50 0.454 828 1893 1015 761 2655 24 310 0
gi|11132834|KPR3_ARATH PROBABLE RIBOSE-PHOSPHA ( 386) [r] 191 1.3e-47 0.429 791 1875 1012 761 2655 104 385 0
...

541>>@C:2317 00050.Contig616 1240 aa
FASTX (3.39 May 2001) function [optimized, BL50 matrix (o=15:-5:-1)xS] ktup: 2
  join: 40, opt: 35, open/ext: -14/ -2 shift: -20, width: 16
The best scores are:
          bits E(99359) %_id  alen  an0  ax0  pn0  px0  anl  axl  fs
gi|1173444|SKD1_MOUSE SKD1 PROTEIN ( 444) [f] 233 4.5e-60 0.504 363 2479 3536 2317 5873 8 367 1
gi|1706647|VPS4_YEAST VACUOLAR PROTEIN SORTING( 437) [f] 226 4.7e-58 0.519 324 2578 3521 2317 5873 39 360 1
gi|9087199|SKD1_HUMAN SKD1 PROTEIN ( 444) [f] 226 4.8e-58 0.490 363 2479 3536 2317 5873 8 367 1
gi|1173445|SKD1_SCHPO SUPPRESSOR PROTEIN OF BEM( 432) [f] 219 5e-56 0.482 342 2509 3521 2317 5873 17 356 1
gi|12230605|SPAS_MOUSE SPASTIN ( 504) [f] 171 1.6e-41 0.476 271 2704 3506 2317 5873 189 453 1
gi|12230611|SPAS_HUMAN SPASTIN ( 616) [f] 169 7.9e-41 0.469 271 2704 3506 2317 5873 301 565 1
gi|462591|MEI1_CAEL MEIOTIC SPINDLE FORMATION( 472) [f] 141 2.8e-32 0.411 280 2710 3497 2317 5873 137 414 1
...

```

70

Large-scale alignment summaries

vs Fungi:

2>>00050.Contig2_1008_160 00050.Contig2_>1008_160 283 aa
11496243 residues in 27113 sequences

FASTA (3.40 July 2001) function [optimized, BL50 matrix (15:-5)xS] ktup: 1
join: 42, opt: 30, open/ext: -10/ -2, width: 32

The best scores are:

	opt	bits	E(27113)	%_id	sw	alen	an0	ax0	an1	ax1	
gi 11282558	hypothetical protein SP(601)	345	91	2.6e-18	0.300	369	253	27	275	48	290
gi 6321996	Yhr202wp [Saccharomyces (602)	345	91	2.6e-18	0.306	351	252	27	263	44	286
gi 12056489	putative 5' nucleotidas(601)	266	72	1.3e-12	0.289	343	253	27	275	48	290
gi 7490856	hypothetical protein SP (635)	220	61	2.8e-09	0.272	351	287	4	275	7	281
gi 6474524	Hypothetical protein [S (130)	162	46	1.2e-05	0.422	162	64	27	89	65	127

vs Metazoa

12>>00050.Contig21_1_879 293 aa
83266113 residues in 259343 sequences

FASTA (3.40 July 2001) function [optimized, BL50 matrix (15:-5)xS] ktup: 1
join: 42, opt: 30, open/ext: -10/ -2, width: 32

The best scores are:

	opt	bits	E(259343)	%_id	sw	alen	an0	ax0	an1	ax1	
gi 6715146	proprotein convertase (1323)	224	51	3.8e-05	0.261	242	272	13	269	1023	1273
gi 423565	serine proteinase (EC (1548)	224	51	4.2e-05	0.245	263	322	1	271	1087	1389
gi 12644383	PROPROTEIN CONVERTAS (1877)	224	51	4.7e-05	0.245	263	322	1	271	1416	1718
gi 11359826	furin (EC 3.4.21.75 (1299)	219	50	7e-05	0.243	250	288	6	269	935	1200
gi 14249600	hypothetical protein (441)	211	48	9.8e-05	0.256	255	273	13	266	120	372
gi 12274831	ba88M19.1 (EGF-like- (401)	205	47	0.0002	0.246	210	211	65	257	25	228

71

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

72