Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 17

1

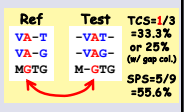
TODAY'S TOPICS

- MSA evaluation
- Conserved domain, pattern, profile
  - PROSITE, PRINTS, Profile (PSSM)
  - PSI-BLAST

2

Multiple alignment quality

- How to measure the quality of multiple alignments?
  - Sum of pairs score (SPS) and total column score (CS or TCS):
    - ➜ SPS=Proportion of correctly aligned AA pairs
    - ➜ TCS (CS) =Proportion of correctly aligned columns



[Programs available]
bali_score (from BAliBASE website)
qscore: http://drive5.com/qscore SPS, CS, Shift Score, *etc.*
Veralign: https://www.ibi.vu.nl/programs/veralignwww/

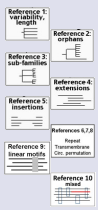
- T-coffee consistency based evaluation (TCS, iRMSD-APDB, Strike)
- MUMSA: average overlap score https://msa.sbc.su.se/cgi-bin/msa.cgi
- GUIDANCE2: guide-tree based alignment confidence https://taux.evolseq.net/guidance/
- Muscle5: ensemble MSA based accuracy https://www.drive5.com/muscle/
- AlignStat: MSA similarity/dissimilarity http://alignstat.science.latrobe.edu.au
- QuanTest: secondary structure prediction based
  http://www.bioinf.ucd.ie/download/QuanTest.tgz (download only)
  http://bioinf.ucd.ie/quantest2.tar (download only)
- Sequence logo: graphical representation of a multiple alignment
  Weblogo 3: https://weblogo.threeplusone.com/

3

Benchmark alignment database: BAliBASE

http://www.lbgi.fr/balibase/ (BAliBASE4)
Thompson *et al.* (1999); Thompson *et al.* (2005)

- 9 reference alignment sets
  - ➜ can be used to evaluate multiple alignment programs



- Reference 1: equidistant sequences with various levels of conservation
- Reference 2: families aligned with a highly divergent "orphan" sequence
- Reference 3: subfamilies with <25% residue identity between groups
- Reference 4: sequences with N/C-terminal extension
- Reference 5: internal insertions
- References 6, 7, 8: various protein families containing internal repeats, inversions, transmembrane regions, etc.
- References 9: linear motifs
- References 10: mixed

4

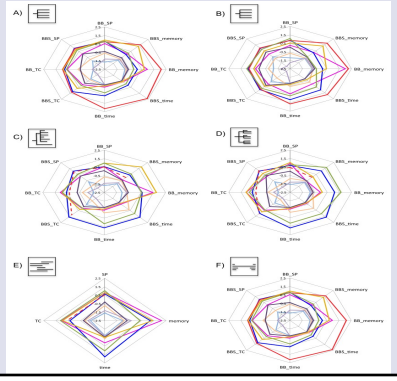
Other benchmark alignment database

- HOMSTRAD: Homologous Structure Alignment Database
  https://homstrad.mizuguchilab.org/homstrad/
  - ➜ A curated database of structure-based alignments for protein families
- PREFAB: Protein Reference Alignment Benchmark
  http://www.drive5.com/muscle/prefab.htm (MUSCLE website)
  - ➜ Automatically generated from structural pairwise alignment expanded with PSI-Blast
  - ➜ Collection of Benchmark alignment database is also available (BENCH)
    http://www.drive5.com/bench
- SABmark: Sequence and Structure Alignment Benchmark
  http://bioinformatics.vub.ac.be/databases/databases.html (no longer available?)
  - ➜ Structural alignments of fold groups
    (Twilight zone set, Superfamily set)

Edgar (2010) "Benchmark collection" (http://drive5.com/bench/)
Iantorno *et al.* (2014) Who watches the watchman? An appraisal of benchmarks for multiple sequence alignment.

5

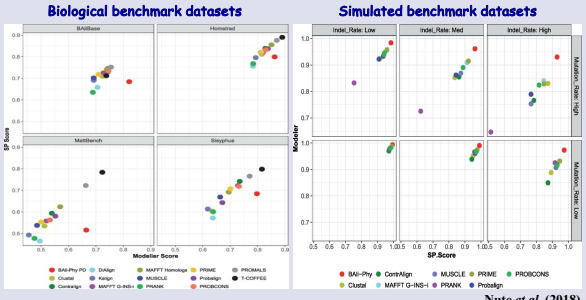
Multiple alignment quality



Pais *et al.* (2014)

6

Multiple alignment quality

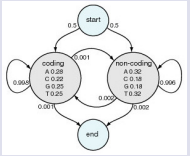


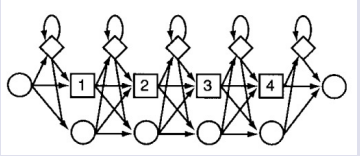
Nute *et al.* (2018)

SP score (recall) = How much of correct alignment is recovered
Modeler score (precision) = How much of the reconstructed alignment is correct

7

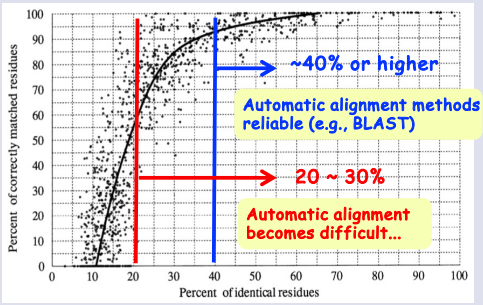
Patterns, profiles, and profile HMMs





8

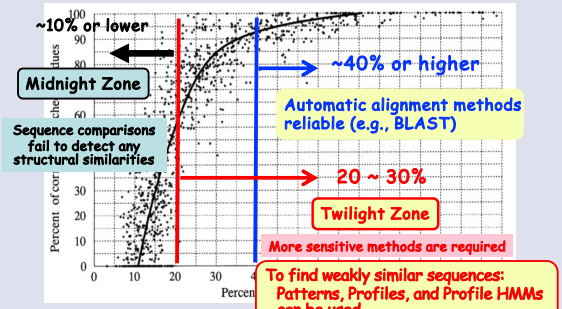
Sequence similarities and search methods



Vogt *et al.* (1995)

9

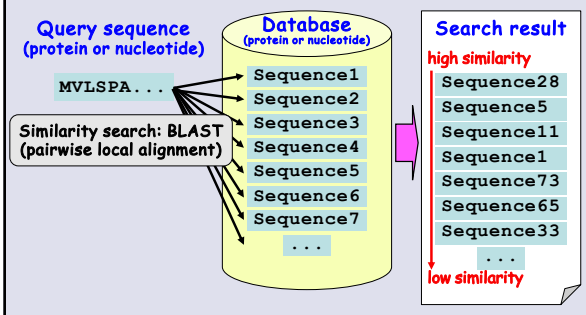
Sequence similarities and search methods



Vogt *et al.* (1995)

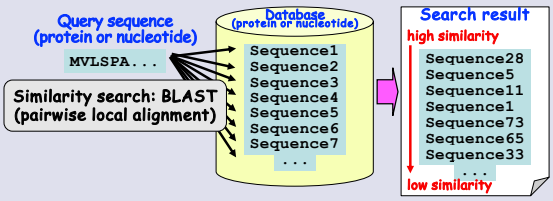
10

Similarity Search



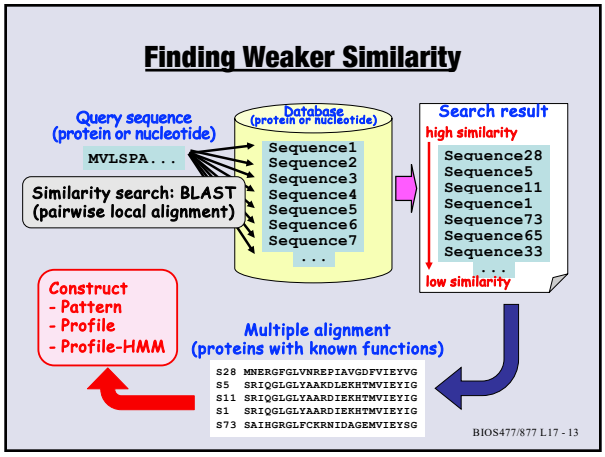
11

Finding Weaker Similarity

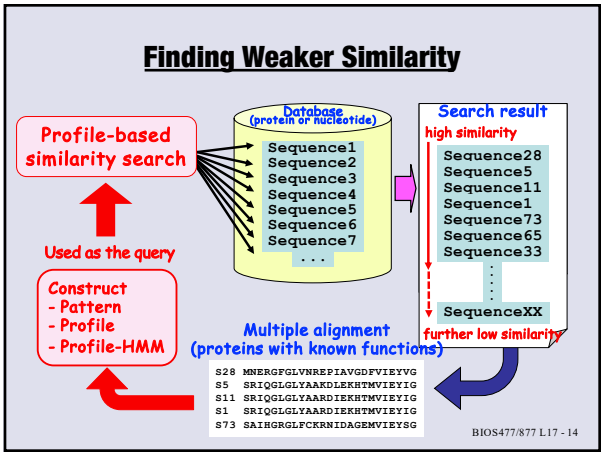


Finding more remotely (weakly) similar sequences helps:
- identify distant members of the protein family (how proteins have evolved)
- identify functions (unknown or not very obvious from the query sequence)

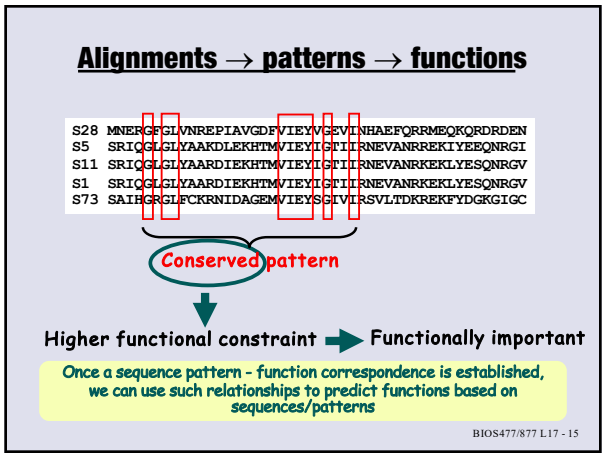
12



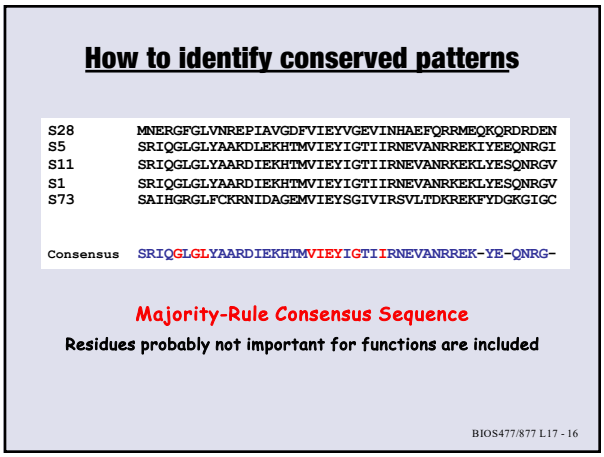
13



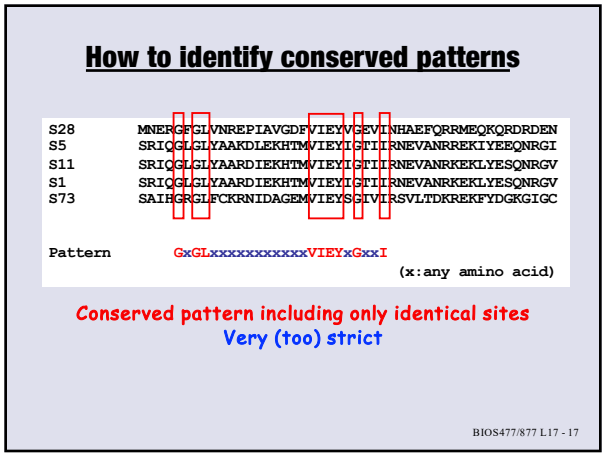
14



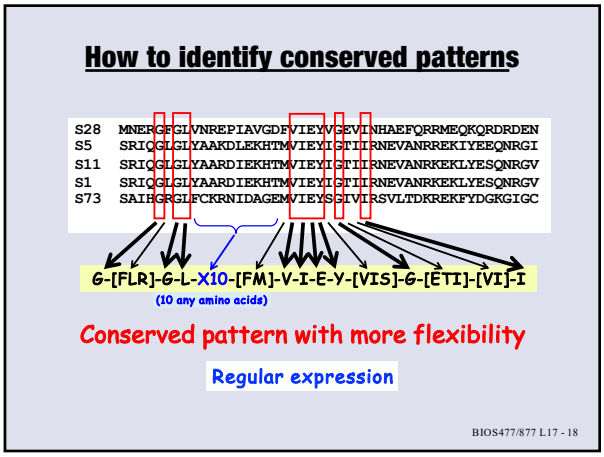
15



16



17



18

PROSITE Pattern Database

prosite — Database of protein domains, families and functional sites

➜ consists of biologically significant sites, patterns, and profiles
https://prosite.expasy.org/

- PROSITE pattern syntax is described in:
  https://prosite.expasy.org/prosuser.html - meth1

19

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]

```
5H1A_FUGRU/131-147      SSIlh caIAL R wai
5H1A_HUMAN/122-138      SSIlh caIAL R wai
5H1A_MOUSE/122-138      SSIlh caIAL R wai
5H1A_PANTR/122-138      SSIlh caIAL R wai
5H1A_RAT/122-138        SSIlh caIAL R wai
5H1B_CAVPO/134-150      ASImh cvIAL R wai
5H1B_CRIGR/131-147      ASImh cvIAL R wai
5H1B_DIDMA/134-150      ASIlh cvIAL R wai
5H1B_FUGRU/119-135      SSIlh cvIAL R wai
5H1B_HUMAN/135-151      ASIlh cvIAL R wai
5H1B_MOUSE/131-147      ASImh cvIAL R wai
5H1B_HUMAN/135-151      ASIlh cvIAL R wai
5H1B_RABIT/135-151      ASImh cvIAL R wai
5H1B_RAT/131-147        ASImh cvIAL R wai
5H1B_SPAEH/131-147      ASImh cvIAL R wai
5H1D_CANFA/124-140      ASIlh cvIAL R wai
5H1D_CAVPO/124-140      ASIlh cvIAL R wai
5H1D_FUGRU/122-138      ASIlh cvIAL R wai
5H1D_HUMAN/124-140      ASIlh cvIAL R wai
5H1D_MOUSE/121-137      ASIlh cvIAL R wai
5H1D_PIG/44-60          ASIlh cvIAL R wai
5H1D_RABIT/124-140      ASIlh cvIAL R wai
5H1D_RAT/121-137        ASIlh cvIAL R wai
5H1E_HUMAN/108-124      CSIlh cvIAL R wai
5H1E_PANTR/108-124      CSIlh cvIAL R wai
5H1E_PIG/55-71          CSIlh cvIAL R wai
```

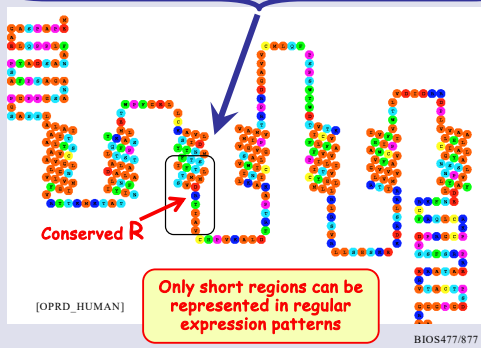

20

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]

```
5H1A_FUGRU/131-147      SSIlh caIAL R wai
5H1A_HUMAN/122-138      SSIlh caIAL R wai
5H1A_MOUSE/122-138      SSIlh caIAL R wai
5H1A_PANTR/122-138      SSIlh caIAL R wai
5H1A_RAT/122-138        SSIlh caIAL R wai
5H1B_CAVPO/134-150      ASImh cvIAL R wai
5H1B_CRIGR/131-147      ASImh cvIAL R wai
5H1B_DIDMA/134-150      ASIlh cvIAL R wai
5H1B_FUGRU/119-135      SSIlh cvIAL R wai
5H1B_HUMAN/135-151      ASIlh cvIAL R wai
5H1B_MOUSE/131-147      ASImh cvIAL R wai
5H1B_HUMAN/135-151      ASIlh cvIAL R wai
5H1B_RABIT/135-151      ASImh cvIAL R wai
5H1B_RAT/131-147        ASImh cvIAL R wai
5H1B_SPAEH/131-147      ASImh cvIAL R wai
5H1D_CANFA/124-140      ASIlh cvIAL R wai
5H1D_CAVPO/124-140      ASIlh cvIAL R wai
5H1D_FUGRU/122-138      ASIlh cvIAL R wai
5H1D_HUMAN/124-140      ASIlh cvIAL R wai
5H1D_MOUSE/121-137      ASIlh cvIAL R wai
5H1D_PIG/44-60          ASIlh cvIAL R wai
5H1D_RABIT/124-140      ASIlh cvIAL R wai
5H1D_RAT/121-137        ASIlh cvIAL R wai
5H1E_HUMAN/108-124      CSIlh cvIAL R wai
5H1E_PANTR/108-124      CSIlh cvIAL R wai
5H1E_PIG/55-71          CSIlh cvIAL R wai
```


21

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]



22

PRINTS — A compendium of protein fingerprints

PRINTS entry: 5HT1ARECEPTR

```
5HT1ARECEPTR1
GQGNNTTASQEPFGTGG
GQGNNTTTSLEPFGTGG
GQGNNTTSPPAPFETGG

5HT1ARECEPTR2
GNVTSISDVTFSYQVITS
GNDTGLSNVTFSYQVITS
GNTTGISDVTVSYQVITS

5HT1ARECEPTR3
RTPEDRSDPDACTISK
RAPEDRSNPNECTISK
RTPEDRSDPDACTISK

5HT1ARECEPTR4
FRIRKTVRKVEKKGAGTSLG
FRIRKTVKKVEKKGAGTSFG
FRIRKTVKKVEKTGADTRHG

5HT1ARECEPTR5
WRRCAENRAVGTPCTNG
CRRSAENRAVGTPCANG
WRLGVESKAGGACLCANG

5HT1ARECEPTR6
AVRQGDDEATLEVIEVHRVG
AVRQGEDDATLEVIEVHRVG
AVRQGDDGAALEVIEVHRVG

5HT1ARECEPTR7
APACLERKNERNAEAK
VPACLERKNERTAEAK
APASFERKNERNAEAK
```

7 motifs

Fingerprint:
- ➜ a group of conserved motifs used to characterize a protein family
- ➜ each motif is a profile
- ➜ Part of InterPro (PRINTS database is no longer updated)

23

Profile (PSSM)

- ➤ Profile: Position Specific Scoring Matrix (PSSM)
  - ➜ Constructed from multiple alignments
    - Short conserved domains (PRINTS, BLOCKS)
    - Protein families (PROSITE)
    - Results of similarity search (PSI-BLAST)
- ➤ More flexible than simple patterns
  - ➜ PSSM lists amino acid frequencies for each alignment position
- ➤ Profiles (PSSMs) can be used for database search to identify remote similarities

24

Profile: Position Specific Scoring Matrix

- **EMBOSS:** http://bio.biomedicine.gu.se/emboss/ (Protein Profiles)
  - PROPHECY: creates profiles from multiple alignment
  - PROFIT: scans a sequence/database with a profile
  - PROPHET: aligns a profile with sequence(s)

```
# Columns are amino acid counts A->Z
# Rows are alignment positions 1->n
Simple
Name            mymatrix
Length          5
Maximum score   11
Thresh          75
Consensus       RCEGH
1 0  0  0  0  0  0  1  0  ... 0  ... 0  2  ...
2 0  0  3  0  0  0  0  0  0  ... 0  ... 0  0  ...
3 0  0  0  0  2  0  0  0  0  ... 0  ... 1  0  ...
4 1  0  0  0  0  0  2  0  0  ... 0  ... 0  0  ...
5 0  0  0  0  0  0  0  2  0  ... 1  ... 0  0  ...
  A  B  C  D  E  F  G  H  I  ... N  ... Q  R  ...
```

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

Simple amino acid frequency

(Asx) →Asn & Asp

25

Profile: Position Specific Scoring Matrix

- **EMBOSS:** http://bio.biomedicine.gu.se/emboss/ (Protein Profiles)
  - PROPHECY: creates profiles from multiple alignment
  - PROFIT: scans a sequence/database with a profile
  - PROPHET: aligns a profile with sequence(s)

```
# Gribskov Protein Profile
# Columns are amino acids A->Z
# Last column is indel penalty
# Rows are alignment positions 1->n
...
Consensus       RCEGH
1 -2.33 0.00 -2.33  1.33  1.33  -3.67 -2.67  8.33 ... 11.67
2  3.00 0.00 15.00 -5.00 -6.00  -1.00  2.00 -1.00 ... -3.00
3  2.67 0.00 -6.00  8.67 12.00  -6.67  4.00  4.67 ...  1.33
4  9.00 0.00  2.33  5.00  4.33  -5.67 12.00 -1.67 ... -3.00
5  0.00 0.00 -1.67  4.67  4.33  -2.33  0.00 11.67 ...  3.67
     A     B     C     D     E     F     G     H   ...   R
   (Asx)
```

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

Non-zero values are used even for the AAs that are not found.

26

Profile: Position Specific Scoring Matrix

- **Gribskov profile** (Gribskov *et al.* 1987)
  - → Used in PROSITE profile
  - → Weighted scoring matrix for each position

Generated from:

- Multiple alignment
- Scoring matrix (*e.g.*, BLOSUM62)→$Y(a,b)$
- For amino acid a at position p:

$$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$$

$Y(a,b)$: Value in the scoring matrix for AA pair a and b
$W(p,b)$: Frequency of amino acid b in position p

27

Profile: Position Specific Scoring Matrix

- **Gribskov profile:** $M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$

Multiple alignment

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4

$W(1,b)$: frequency of each AA at pos. 1

$M(1,A) = \sum_{b=1}^{20} W(1,b) * Y(A,b)$

For amino acid A at position 1

```
1 -2.33 0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00 0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67 0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00 0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00 0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
     A     B     C     D   ...   H   ...   R   ...
   (Asx)
```


28

Profile: Position Specific Scoring Matrix

- **Gribskov profile:** $M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$

Multiple alignment

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4

$M(1,A) = \sum_{b=1}^{20} W(1,b) * Y(A,b)$

$W(1,A) = 0$
Frequency of Ala (b=1) at pos. 1

$Y(A,A) = 4$ (for b=1)

$M(1,A) = 0 * 4 + W(1,R) * Y(A,R) +$
$W(1,N) * Y(A,N) + W(1,D) * Y(A,D) +$
$... + W(1,H) * Y(A,H) + W(1,I) * Y(A,I) +$
$...$ [sum up for all 20 amino acids]

```
1 -2.33 0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00 0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67 0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00 0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00 0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
     A     B     C     D   ...   H   ...   R   ...
```


29

Profile: Position Specific Scoring Matrix

- **Gribskov profile:** $M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$

Multiple alignment

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4

$M(1,A) = \sum_{b=1}^{20} W(1,b) * Y(A,b)$

$W(1,R) = 2/3$
Frequency of Arg (b=2) at pos. 1

$Y(A,R) = -1$ for b=2

$M(1,A) = 0 * 4 + 0.67 * (-1) +$
$0 * (-2) + 0 * (-2) +$
$... + 0.33 * (-2) + 0 * (-1) +$
$...$ [sum up for all 20 amino acids]

```
1 -2.33 0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00 0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67 0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00 0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00 0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
     A     B     C     D   ...   H   ...   R   ...
```


30

Profile: Position Specific Scoring Matrix

➢ Gribskov profile: $M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$

But many amino acids have 0 counts! Should we just ignore them?

Multiple alignment

```
Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN
```

BLOSUM62

```
   A  R  N  D  C  Q  E  G  H  I  L
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1
R -1  5  0 -2 -3  1  0 -2  0 -3 -2
N -2  0  6  1 -3  0  0  0  1 -3 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1
Q -1  1  0  0 -3  5  2 -2  0 -3 -2
E -1  0  0  2 -4  2  5 -2  0 -3 -3
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2
```

$M(1,A) = 0 * 4 + 0.67 * (-1) +$
$0 * (-2) + 0 * (-2) +$
$... + 0.33 * (-2) + 0 * (-1) +$
$... \text{ [sum up for all 20 amino acids]}$

```
1 -2.33 0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00 0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67 0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00 0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00 0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
     A    B     C     D   ...   H   ...   R   ...
```


31

Profile: Position Specific Scoring Matrix

➢ Gribskov profile: $M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$

Instead of 0, a very small weight (e.g., 0.025/#seq) can be used.

Multiple alignment

```
Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN
```

BLOSUM62

```
   A  R  N  D  C  Q  E  G  H  I  L
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1
R -1  5  0 -2 -3  1  0 -2  0 -3 -2
N -2  0  6  1 -3  0  0  0  1 -3 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1
Q -1  1  0  0 -3  5  2 -2  0 -3 -2
E -1  0  0  2 -4  2  5 -2  0 -3 -3
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2
```

Why not 0?

$M(1,A) = 0.008*4 + 0.67*(-1) +$
$0.008*(-2) + 0.008*(-2) +$
$... + 0.33*(-2) + 0.008*(-1) +$
$... \text{ [sum up for all 20 amino acids]}$

```
1 -2.33 0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00 0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67 0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00 0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00 0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
     A    B     C     D   ...   H   ...   R   ...
```


32

Over-fitting problem in profile

```
Seq1 RDA
Seq2 REA
Seq3 REG
```

Position 1 has only Arg (R)
Position 2 has only Asp (D) & Glu (E)
Position 3 has only Ala (A) & Gly (G)

A Simple Model Can we find KEA or RET ?

```
Pos  A R N  D C Q  E  G H I L K T ...
 1   0 1 0  0 0 0  0  0 0 0 0 0 0 ...
 2   0 0 0 .3 0 0 .6  0 0 0 0 0 0 ...
 3  .6 0 0  0 0 0  0 .3 0 0 0 0 0 ...
```


33

Over-fitting problem in profile

```
Seq1 RDA
Seq2 REA
Seq3 REG
```

Position 1 has only Arg (R)
Position 2 has only Asp (D) & Glu (E)
Position 3 has only Ala (A) & Gly (G)

A Simple Model Can we find ~~KEA~~ or ~~RET~~ ?

```
Pos  A R N  D C Q  E  G H I L K T ...
 1   0 1 0  0 0 0  0  0 0 0 0 0 0 ...
 2   0 0 0 .3 0 0 .6  0 0 0 0 0 0 ...
 3  .6 0 0  0 0 0  0 .3 0 0 0 0 0 ...
```

No flexibility!

34

Pseudocount methods

➢ Simplest method (used with Gribskov profile)

$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$

$Y(a,b)$: Score for amino acid pair a and b (e.g., BLOSUM62)
$W(p,b) = n(p,b) / N$
$n(p,b)$: Number of amino acid b in the position p
N: Number of sequences

➜ Add a small constant to all the counts
$W(p,b) = \{n(p,b) + C\} / (N + C)$

C: pseudocount
(something small but not 0)

35

Pseudocount methods

➢ Simplest method
- Add a small constant to all the counts ($C = 0.1$, *etc.*)

➢ Substitution matrix dependent pseudocounts
- Proportional to scores, S_{ij}

➢ Dirichlet mixtures (Sjölander *et al.* 1996)
- Used in profile HMM
- Mixture of different types of pseudocounts
- Representing various context of protein sequences (*e.g.*, loop region, hydrophobic region)

36

Profile: Position Specific Scoring Matrix

➢ Gribskov profile: $M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$

Multiple alignment

```
Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN
```

Psuedocount: 0.025/#seq is used

BLOSUM62

```
   A  R  N  D  C  Q  E  G  H  I  L
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1
R -1  5  0 -2 -3  1  0 -2  0 -3 -2
N -2  0  6  1 -3  0  0  0  1 -3 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1
Q -1  1  0  0 -3  5  2 -2  0 -3 -2
E -1  0  0  2 -4  2  5 -2  0 -3 -3
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2
```

$M(1,A) = 0.008*4 + 0.67*(-1) +$
$0.008*(-2) + 0.008*(-2) +$
$... + 0.33*(-2) + 0.008*(-1) +$
$... [\text{sum up for all 20 amino acids}]$

```
1 -2.33  0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00  0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67  0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00  0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00  0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
   A     B     C     D    ...  H    ...  R    ...
```


37

PROSITE Profile (PSSM) entry

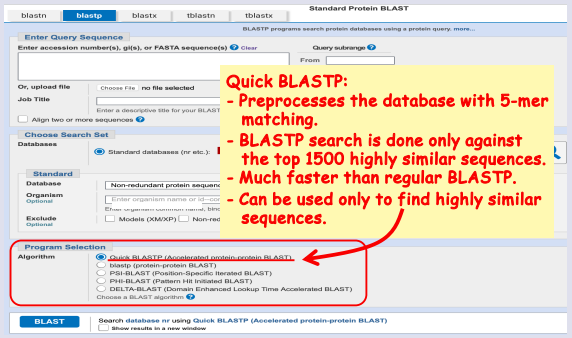
```
ID   G_PROTEIN_RECEP_F1_2; MATRIX.
AC   PS50262;
DT   DEC-2001 (CREATED); DEC-2001 (DATA UPDATE); FEB-2004 (INFO UPDATE).
DE   G-protein coupled receptors family 1 profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=259;
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=254;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=1.9359; R2=0.02006056; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=0; SCORE=327; N_SCORE=8.5; MODE=1; TEXT='!';
MA   /CUT_OFF: LEVEL=-1; SCORE=227; N_SCORE=6.5; MODE=1; TEXT='?';
MA   /DEFAULT: D=-[illegible] R1=-100; R1=-100; M1=-105; MM=-105; IM=-105; DM=-105; MM=1; MD=10;
MA   /I: B1=0; B1=-[illegible]
MA   /M: SY=' A  B  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y  Z
Pos1 /M: SY=' V'; M=-[illegible]
Pos2 /M: SY=' N'; M=-9,-33,-19,15,-2,-18,-[illegible],-18,-2,-26,-18,51,-20,-1,-2,9,1,-26,-38,-17,-2;
Pos3 /M: SY=' I'; M=-1,-21,-16,-26,-21,0,-16,-22,10,-22,8,-17,-22,-19,-20,-8,-3,10,-21,-7,-20;
Pos4 /M: SY=' L'; M=-6,-24,-17,-28,-21,8,-25,-20,-[illegible],-24,23,12,-21,-25,-20,-19,-17,-5,11,-17,0,-20;
Pos5 /M: SY=' V'; M=-1,-19,-16,-23,-21,-2,-24,-14,15,-[illegible],7,-16,-24,-18,-17,-7,-1,21,-26,-5,-20;
Pos6 /M: SY=' I'; M=-8,-28,-16,-33,-25,6,-31,-24,26,-26,24,15,-24,-26,-21,-23,-20,-8,20,-20,-1,-24;
Pos7 /M: SY=' I'; M=-6,-23,-19,-27,-21,5,-24,-16,8,-19,8,5,-20,-23,-17,-17,-15,-6,5,-2,7,-19;
Pos8 /M: SY=' V'; M=
Pos9 /M: SY=' I'; M=
Pos10 /M: SY=' F'; M=
Pos11 /M: SY=' R'; M=
Pos12 /M: SY=' K'; M=
Pos13 /M: SY=' R'; M=
Pos14 /M: SY=' R'; M=
Pos15 /M: SY=' A'; M=-8,-7,-21,-9,-3,-16,-16,-4,-14,7,-12,-1,-3,-14,3,11,-4,-4,-11,-23,-8,-1;
Pos16 /I: I=-[illegible]; MD=-23;
Pos17 /M: SY=' L'; M=-7,-17,-20,-19,-11,-2,-20,-11,1,-8,11,8,-14,-19,-8,-1,-14,-7,0,-18,-3,-10; D=-4;
      /I: I=-[illegible]MI=0; MD=-23; IM=0; DM=-23;
...
```

For each position, the likelihood of having each amino acid is listed
+ score: these amino acids are more likely to appear
- score: these amino acids are less likely to appear

Consensus sequence for each position

38

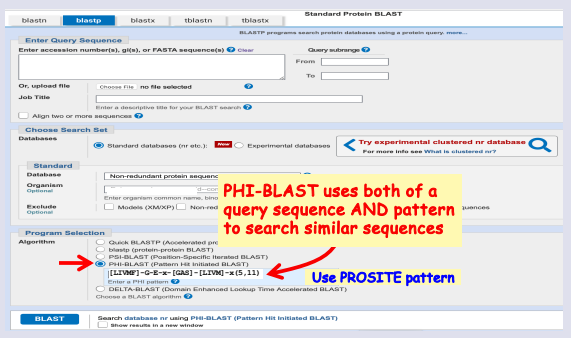
Protein BLAST algorithms



Quick BLASTP:
- Preprocesses the database with 5-mer matching.
- BLASTP search is done only against the top 1500 highly similar sequences.
- Much faster than regular BLASTP.
- Can be used only to find highly similar sequences.

39

PHI-BLAST: Pattern-Hit Initiated BLAST

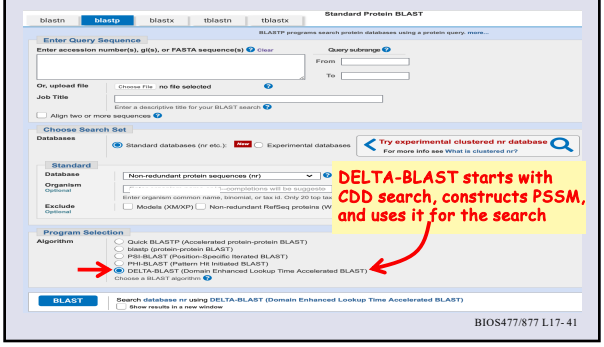


PHI-BLAST uses both of a query sequence AND pattern to search similar sequences

Use PROSITE pattern

40

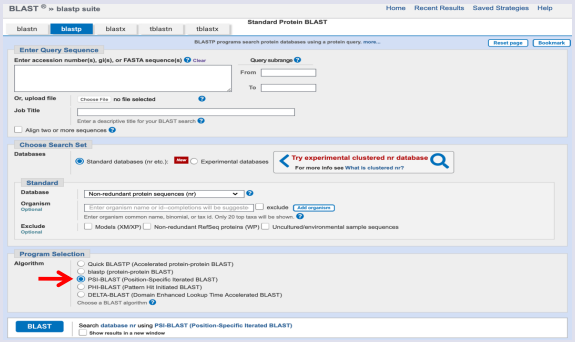
DELTA-BLAST: Domain Enhanced lookup Time Accelerated BLAST



DELTA-BLAST starts with CDD search, constructs PSSM, and uses it for the search

41

PSI-BLAST: Position-Specific Iterated BLAST



42

PSI-BLAST: Position-Specific Iterated BLAST

https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-2.html
(PSI-BLAST introduction)

https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html
(PSI-BLAST statistics)

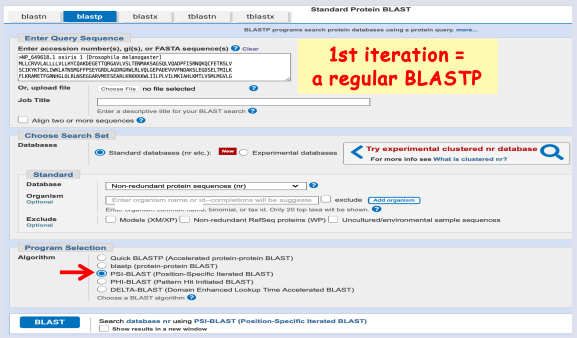
43

PSI-BLAST: Position-Specific Iterated BLAST

- PSI-BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi
  - → 1st iteration: a regular BLASTP search
    - Uses a scoring matrix (e.g., BLOSUM62)
  - → 2nd iteration
    - Multiple alignment is constructed from the highly similar hits
    - Positive-specific scoring matrix (PSSM) is constructed
    - Similarity search using the PSSM instead of the single query sequence
  - → 3rd, ..., iterations
    - Stop when no more new hit (or anytime)

44

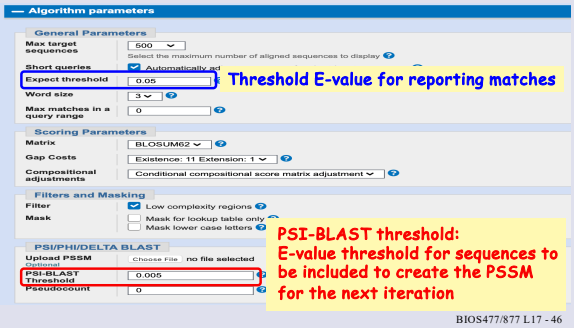
PSI-BLAST



45

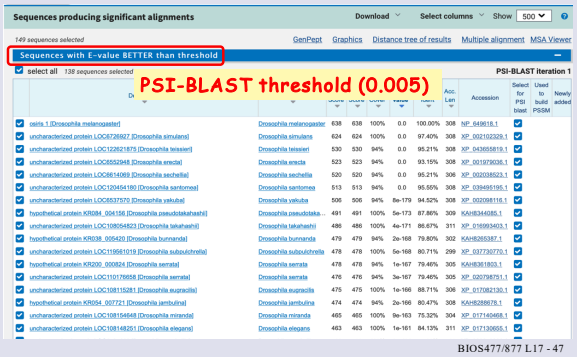
PSI-BLAST

- Two different E-value thresholds



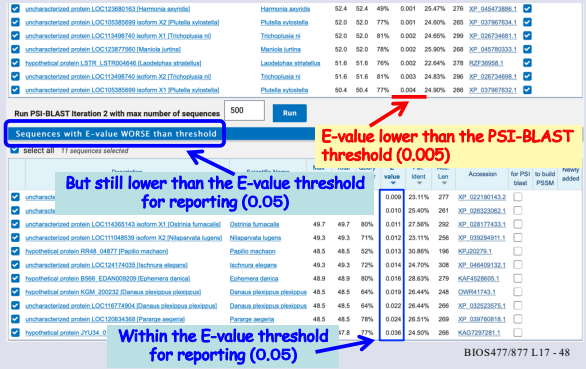
46

PSI-BLAST

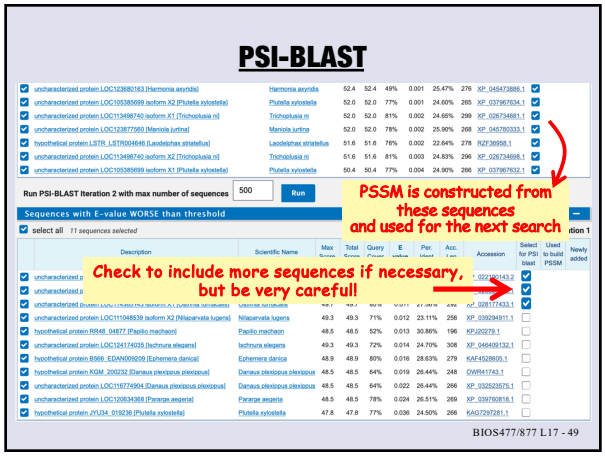


47

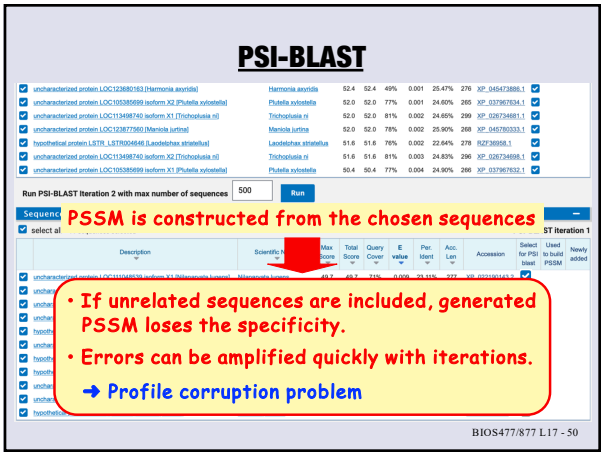
PSI-BLAST



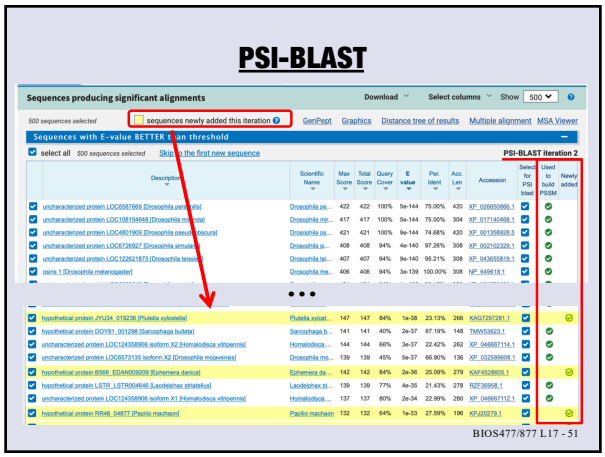
48


PSI-BLAST
PSSM is constructed from these sequences and used for the next search
Check to include more sequences if necessary, but be very careful!
BIOS477/877 L17 - 49

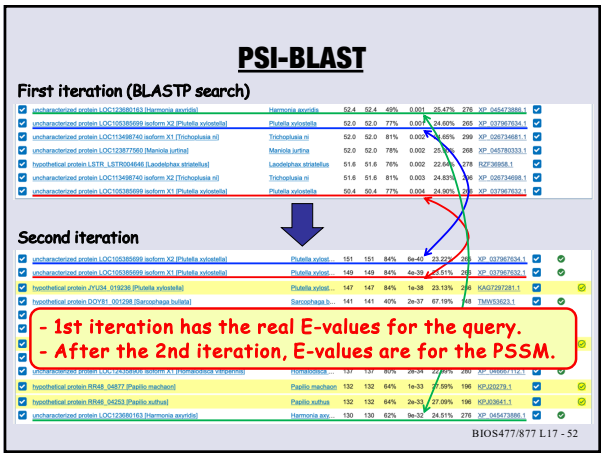
49


PSI-BLAST
PSSM is constructed from the chosen sequences
• If unrelated sequences are included, generated PSSM loses the specificity.
• Errors can be amplified quickly with iterations.
➜ Profile corruption problem
BIOS477/877 L17 - 50

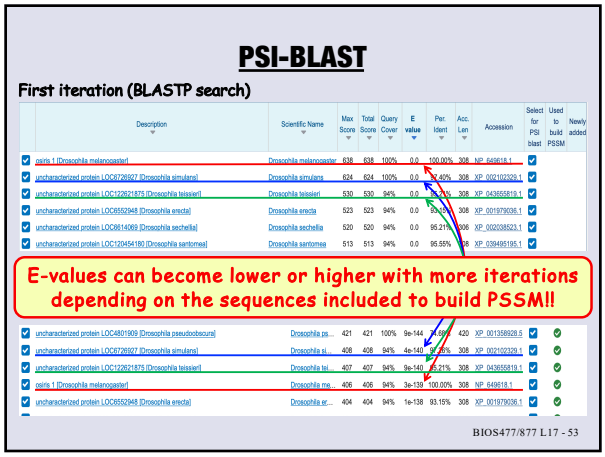
50


PSI-BLAST
Sequences producing significant alignments
sequences newly added this iteration
Sequences with E-value BETTER than threshold
PSI-BLAST iteration 2
BIOS477/877 L17 - 51

51


PSI-BLAST
First iteration (BLASTP search)
Second iteration
- 1st iteration has the real E-values for the query.
- After the 2nd iteration, E-values are for the PSSM.
BIOS477/877 L17 - 52

52


PSI-BLAST
First iteration (BLASTP search)
E-values can become lower or higher with more iterations depending on the sequences included to build PSSM!!
BIOS477/877 L17 - 53

53