

Spring 2024
BIOS 477/877
Bioinformatics and Molecular Evolution
Lecture 17

BIOS477/877 L17 - 1

1

TODAY'S TOPICS

➤ **MSA evaluation**

➤ **Conserved domain, pattern, profile**

- **PROSITE, PRINTS, Profile (PSSM)**
- **PSI-BLAST**

BIOS477/877 L17 - 2

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

Ref	Test	TCS=1/3
VA-T	-VAT	=33.3%
VA-G	-VAG	or 25% (w/ gap col.)
MCTG	M-CTG	SPS=5/9 =55.6%

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

- 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.ucl.ac.uk/download/QuanTest.tgz> (download only)
<http://bioinf.ucl.ac.uk/quantest2.tar> (download only)
- Sequence logo: graphical representation of a multiple alignment
Weblogo 3: <https://weblogo.threeplusone.com/>

BIOS477/877 L17 - 3

3

Benchmark alignment database: BALiBASE

<http://www.lbgj.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

BIOS477/877 L17 - 4

4

Other benchmark alignment database

- **HOMSTRAD: Homologous Structure Alignment Database**
<https://homstrad.mizuguchi-lab.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/>)
 Ianforno *et al.* (2014) Who watches the watchman? An appraisal of benchmarks for multiple sequence alignment.

BIOS477/877 L17 - 5

5

Multiple alignment quality

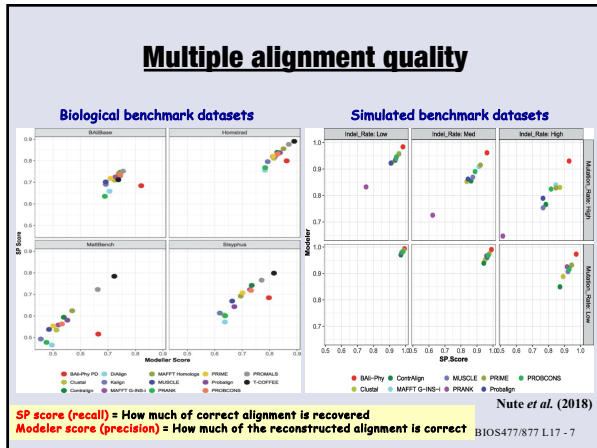
SP score:
 Sum-of-Pairs score
TCS score:
 Total-Column score

— CLUSTALW
 — Probcons
 — CLUSTAL-D
 — T-Coffee
 — DALIGN-TX
 — MAFFT
 — POA
 — Probalign
 — MUSCLE

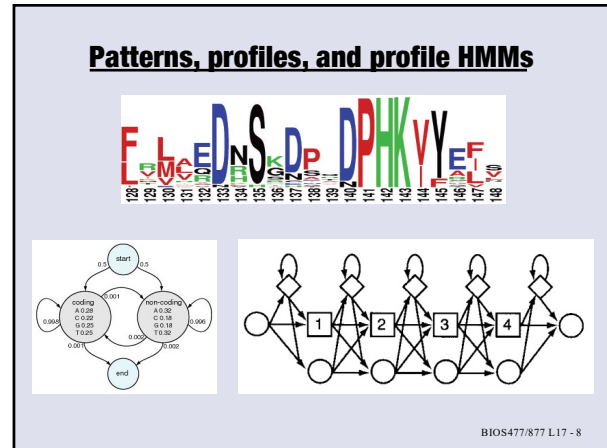
Pais *et al.* (2014)

BIOS477/877 L17 - 6

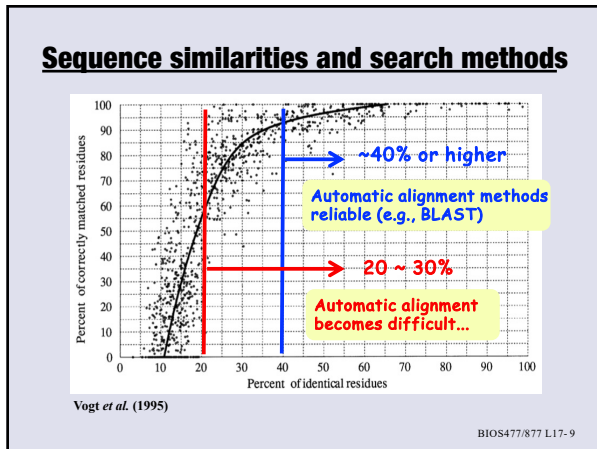
6



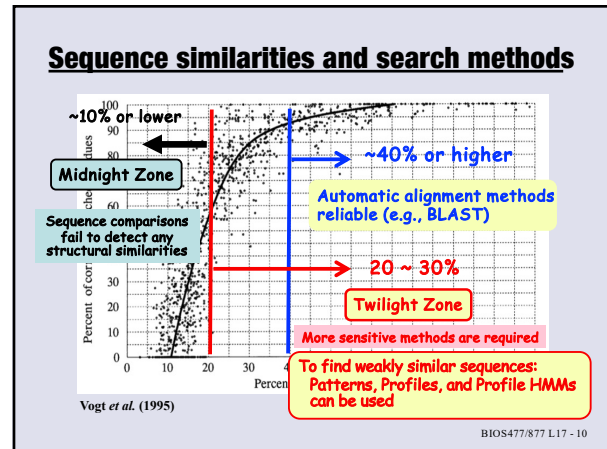
7



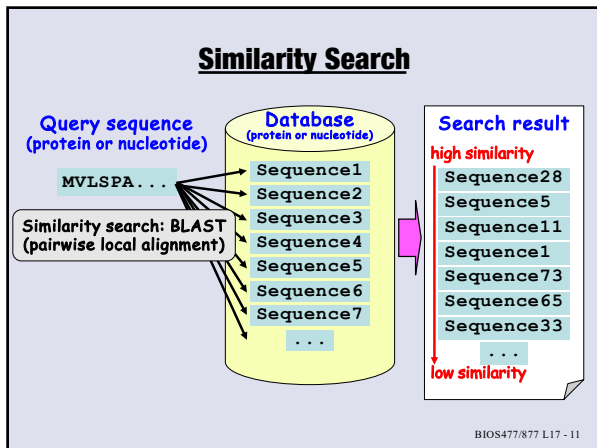
8



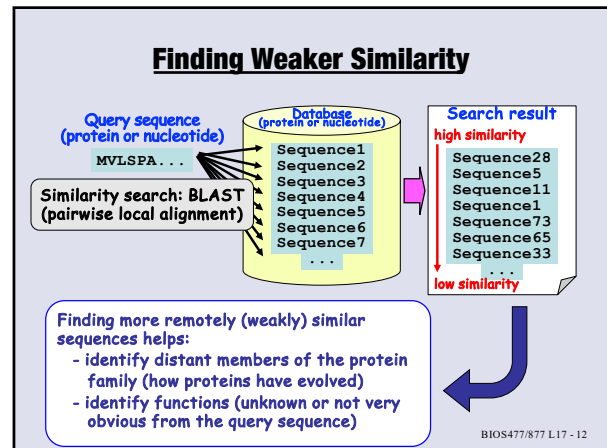
9



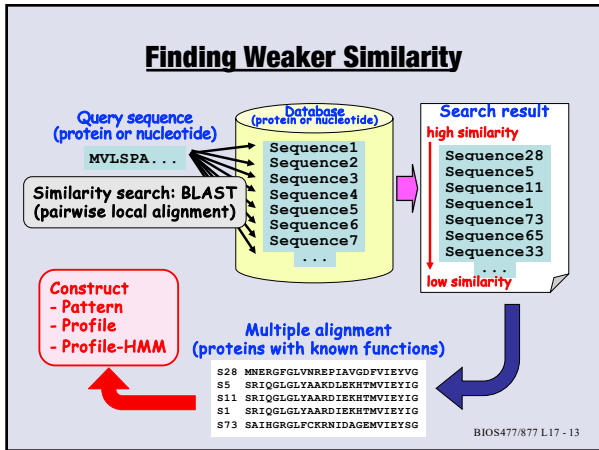
10



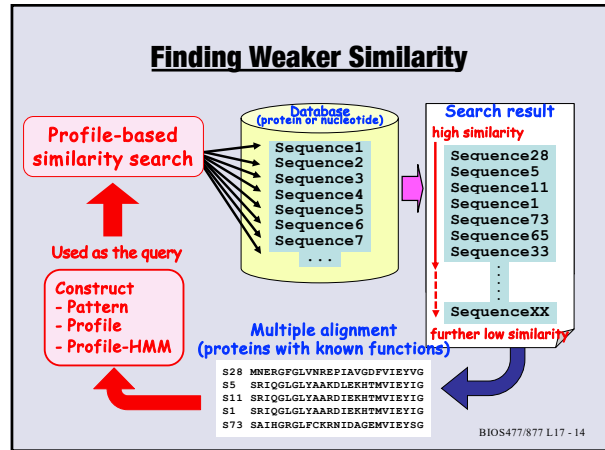
11



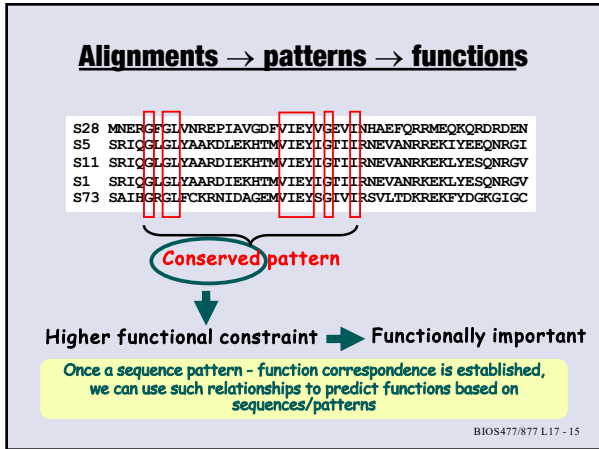
12



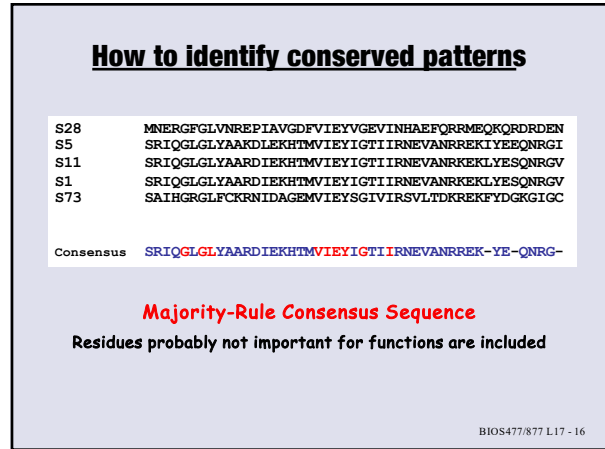
13



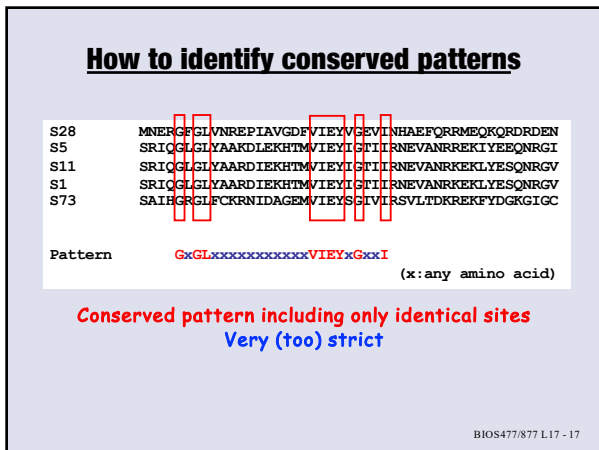
14



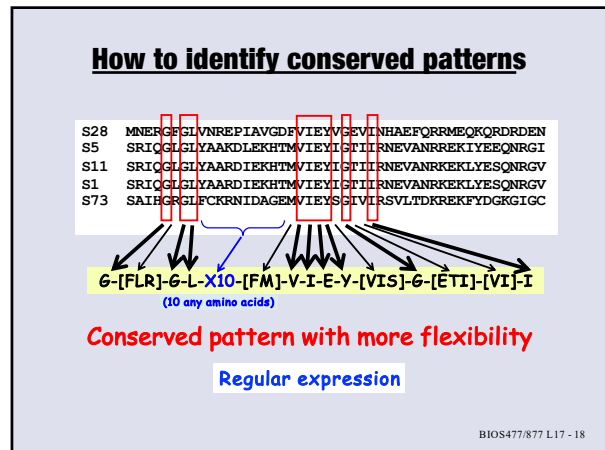
15



16



17



18

PROSITE Pattern Database

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

BIOS477/877 L17 - 19

19

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)

[GSTALIVMFYWC]-[GSTANCPDE]-[EDPKRH]-x(2)-[LIVMNOGA]-x(2)-[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSSH]-x(2)-[LIVM]

SH1A_FUGRU/131-147	SSIlh	caIALDRwaI
SH1A_HUMAN/122-138	SSIlh	caIALDRwaI
SH1A_MOUSE/122-138	SSIlh	caIALDRwaI
SH1A_PANTR/122-138	SSIlh	caIALDRwaI
SH1A_RAT/122-138	SSIlh	caIALDRwaI
SH1B_CAVPO/134-150	ASImh	cvIALDRwaI
SH1B_CRIGR/131-147	ASImh	cvIALDRwaI
SH1B_DIDMA/134-150	ASIlh	cvIALDRwaI
SH1B_FUGRU/119-135	SSIlh	caIALDRwaI
SH1B_HUMAN/135-151	ASIlh	cvIALDRwaI
SH1B_MOUSE/131-147	ASImh	cvIALDRwaI
SH1B_HUMAN/135-151	ASIlh	cvIALDRwaI
SH1B_RABIT/135-151	ASImh	cvIALDRwaI
SH1B_RAT/131-147	ASImh	cvIALDRwaI
SH1B_SPAEH/131-147	ASImh	cvIALDRwaI
SH1D_CANFA/124-140	ASIlh	cvIALDRwaI
SH1D_CAVPO/124-140	ASIlh	cvIALDRwaI
SH1D_FUGRU/122-138	ASIlh	cvIALDRwaI
SH1D_HUMAN/124-140	ASIlh	cvIALDRwaI
SH1D_MOUSE/121-137	ASIlh	cvIALDRwaI
SH1D_PIG/44-60	ASIlh	cvIALDRwaI
SH1D_RABIT/124-140	ASIlh	cvIALDRwaI
SH1D_RAT/121-137	ASIlh	cvIALDRwaI
SH1E_HUMAN/108-124	CSIlh	cvIALDRwaI
SH1E_PANTR/108-124	CSIlh	cvIALDRwaI
SH1E_PIG/55-71	CSIlh	cvIALDRwaI

BIOS477/877 L17 - 20

20

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)

[GSTALIVMFYWC]-[GSTANCPDE]-[EDPKRH]-x(2)-[LIVMNOGA]-x(2)-[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSSH]-x(2)-[LIVM]

SH1A_FUGRU/131-147	SSIlh	caIALDRwaI
SH1A_HUMAN/122-138	SSIlh	caIALDRwaI
SH1A_MOUSE/122-138	SSIlh	caIALDRwaI
SH1A_PANTR/122-138	SSIlh	caIALDRwaI
SH1A_RAT/122-138	SSIlh	caIALDRwaI
SH1B_CAVPO/134-150	ASImh	cvIALDRwaI
SH1B_CRIGR/131-147	ASImh	cvIALDRwaI
SH1B_DIDMA/134-150	ASIlh	cvIALDRwaI
SH1B_FUGRU/119-135	SSIlh	caIALDRwaI
SH1B_HUMAN/135-151	ASIlh	cvIALDRwaI
SH1B_MOUSE/131-147	ASImh	cvIALDRwaI
SH1B_HUMAN/135-151	ASIlh	cvIALDRwaI
SH1B_RABIT/135-151	ASImh	cvIALDRwaI
SH1B_RAT/131-147	ASImh	cvIALDRwaI
SH1B_SPAEH/131-147	ASImh	cvIALDRwaI
SH1D_CANFA/124-140	ASIlh	cvIALDRwaI
SH1D_CAVPO/124-140	ASIlh	cvIALDRwaI
SH1D_FUGRU/122-138	ASIlh	cvIALDRwaI
SH1D_HUMAN/124-140	ASIlh	cvIALDRwaI
SH1D_MOUSE/121-137	ASIlh	cvIALDRwaI
SH1D_PIG/44-60	ASIlh	cvIALDRwaI
SH1D_RABIT/124-140	ASIlh	cvIALDRwaI
SH1D_RAT/121-137	ASIlh	cvIALDRwaI
SH1E_HUMAN/108-124	CSIlh	cvIALDRwaI
SH1E_PANTR/108-124	CSIlh	cvIALDRwaI
SH1E_PIG/55-71	CSIlh	cvIALDRwaI

BIOS477/877 L17 - 21

21

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)

[GSTALIVMFYWC]-[GSTANCPDE]-[EDPKRH]-x(2)-[LIVMNOGA]-x(2)-[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSSH]-x(2)-[LIVM]

BIOS477/877 L17 - 22

22

PRINTS A compendium of protein fingerprints

PRINTS entry: 5HT1ARECEPTR

SH1ARECEPT1 GQNNNTASQPFPTGG GQNNNTSLRPFPTGG GQNNNTSPAFPETGG	SH1ARECEPT5 WRRCABNRVGFPCPTG CRGABNRVGFPCANG WRLVSKAGGALCANG	SH1ARECEPT2 GNVTSISNVFSYQVITS GNDTGLSNVFSYQVITS GNVTSISNVFSYQVITS	SH1ARECEPT6 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT3 RPFEDRSDFPACTISK RPFEDRSDFPACTISK RPFEDRSDFPACTISK	SH1ARECEPT7 APACLERNRNEAK VPACLERNRNEAK APASPFRNDRNEAK	SH1ARECEPT4 FRIRKTVKVRKGGAGTSLG FRIRKTVKVRKGGAGTSLG FRIRKTVKVRKGGAGTSLG	SH1ARECEPT8 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT9 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT10 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT11 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT12 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT13 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT14 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT15 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT16 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT17 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT18 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT19 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG	SH1ARECEPT20 AVRQGDGALELVIEVHRVG AVRQGDGDALELVIEVHRVG AVRQGDGDALELVIEVHRVG
--	---	--	--	---	---	--	--	--	---	---	---	---	---	---	---	---	---	---	---

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)

BIOS477/877 L17 - 23

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

BIOS477/877 L17 - 24

24

Profile: Position Specific Scoring Matrix

- EMBOSS: <http://bio.biomedicine.gu.se/emboss/> (Protein Profiles)
 - PROFPHYC: 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

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

Simple amino acid frequency

```

Name      mymatrix
Length    5
Maximum score 11
Thresh    75
Consensus RCEGH
1 0 0 0 0 0 0 0 1 ... 0 ... 0 ... 0 ...
2 0 0 0 3 0 0 0 0 ... 0 ... 0 ... 0 ...
3 0 0 0 0 2 0 0 0 ... 0 ... 0 ... 0 ...
4 1 0 0 0 0 2 0 0 ... 0 ... 0 ... 0 ...
5 0 0 0 0 0 0 2 0 ... 1 ... 0 ... 0 ...
A B C D E F G H I ... N ... Q R ...
(Asx) -> Asn & Asp
    
```

BIOS477/877 L17 -25

25

Profile: Position Specific Scoring Matrix

- EMBOSS: <http://bio.biomedicine.gu.se/emboss/> (Protein Profiles)
 - PROFPHYC: 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

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

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)
    
```

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

BIOS477/877 L17 -26

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

BIOS477/877 L17 -27

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

frequency of each AA at pos. 1

$W(1,b)$

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

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)
    
```

BIOS477/877 L17 -28

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

Frequency of Ala ($b=1$) at pos. 1

$W(1,A) = 0$

$Y(A,A) = 4$ (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 ...
    
```

BIOS477/877 L17 -29

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

Frequency of Arg ($b=2$) at pos. 1

$W(1,R) = 2/3$

$Y(A,R) = -1$ 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 ...
    
```

BIOS477/877 L17 -30

30

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

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

$M(1,A) = 0 * 4 + 0.67 * (-1) + 0 * (-2) + 0 * (-2) + \dots + 0.33 * (-2) + 0 * (-1) + \dots$
[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	...

BIOS477/877 L17 - 31

31

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

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

Why not 0?

$M(1,A) = 0.008 * 4 + 0.67 * (-1) + 0.008 * (-2) + 0.008 * (-2) + \dots + 0.33 * (-2) + 0.008 * (-1) + \dots$
[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	...

BIOS477/877 L17 - 32

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	...

BIOS477/877 L17 - 33

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!

BIOS477/877 L17 - 34

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)

BIOS477/877 L17 - 35

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)

BIOS477/877 L17 - 36

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
  
```

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	-4	-3	-1	-1	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	2	-4	2	2	5	0	-3	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	-1	-3	-2	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-4	-3	-4	4	2
L	-1	-3	-3	-4	-1	-3	-4	-3	-3	2	4

Pseudocount: 0.025/#seq is used

$M(1,A) = 0.008*4 + 0.67*(-1) + 0.008*(-2) + 0.008*(-2) + \dots + 0.33*(-2) + 0.008*(-1) + \dots$ [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	...

BIOS477/877 L17 -37

37

PROSITE Profile (PSSM) entry

```

ID G_PROTEIN_RECEP_F1_2; MATRIX.
AC P550262;
DT DEC-2001 (CREATED); DEC-2001 (DATA UPDATE); FEB-2004 (INFO UPDATE).
DE G-protein coupled receptors family 1 profile.
NA /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=259;
NA /DISJOINT: DEFINITION=PROTECT; R1=6; R2=5;
NA /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=1.9359; R2=0.0206056; TEXT='-LogE';
NA /CUT_OFF: LEVEL=0; SCORE=327; M_SCORE=8.5; MODE=1; TEXT='';
NA /CUT_OFF: LEVEL=-1; SCORE=227; M_SCORE=6.5; MODE=1; TEXT='';
NA /DEFAULT: D=...; B1=100; B2=100; M1=100; M2=100; M3=100; M4=100; M5=100;
NA /!;
  
```

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

BIOS477/877 L17 -38

38

Protein BLAST algorithms

Standard Protein BLAST

Enter Query Sequence

Or, upload file

Choose Search Set

Standard databases (or etc.):

Database: Non-redundant protein sequences (nr)

Program Selection

Quick BLASTP (Accelerated protein-protein BLAST)

PHI-BLAST (Pattern-Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

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.

BIOS477/877 L17 -39

39

PHI-BLAST: Pattern-Hit Initiated BLAST

Standard Protein BLAST

Enter Query Sequence

Or, upload file

Choose Search Set

Standard databases (or etc.):

Database: Non-redundant protein sequences (nr)

Program Selection

Quick BLASTP (Accelerated protein-protein BLAST)

PHI-BLAST (Pattern-Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

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

Use PROSITE pattern

BIOS477/877 L17 -40

40

DELTA-BLAST: Domain Enhanced Lookup Time Accelerated BLAST

Standard Protein BLAST

Enter Query Sequence

Or, upload file

Choose Search Set

Standard databases (or etc.):

Database: Non-redundant protein sequences (nr)

Program Selection

Quick BLASTP (Accelerated protein-protein BLAST)

PHI-BLAST (Pattern-Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

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

BIOS477/877 L17-41

41

PSI-BLAST: Position-Specific Iterated BLAST

Standard Protein BLAST

Enter Query Sequence

Or, upload file

Choose Search Set

Standard databases (or etc.):

Database: Non-redundant protein sequences (nr)

Program Selection

Quick BLASTP (Accelerated protein-protein BLAST)

PHI-BLAST (Pattern-Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

BIOS477/877 L17 -42

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)

BIOS477/877 L17 - 43

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)

BIOS477/877 L17 - 44

44

PSI-BLAST

Standard Protein BLAST

Enter Query Sequence

1st iteration = a regular BLASTP

Choose Search Set

Program Selection

- Quick BLASTP (Accelerated protein-protein BLAST)
- Rapid protein-protein BLAST
- PSI-BLAST (Position-Specific Iterated BLAST)**
- PSI-BLAST (Position-Specific Iterated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

BIOS477/877 L17 - 45

45

PSI-BLAST

➤ **Two different E-value thresholds**

Algorithm parameters

General Parameters

Max target sequences: 500

Expect threshold: 0.05

PSI-BLAST Threshold: 0.005

Scoring Parameters

Matrix: BLOSUM62

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only, Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM

PSI-BLAST Threshold: 0.005

BIOS477/877 L17 - 46

46

PSI-BLAST

Sequences producing significant alignments

Sequences with E-value BETTER than threshold

PSI-BLAST threshold (0.005)

Accession	Score	E-value	Bits	Positives	Identical	Accession	Score	E-value	Bits	Positives	Identical
osls_1.Drosophila melanogaster	638	100%	0.0	100.00%	308	NP_649618.1					
uncharacterized protein LOC8298927 (Drosophila simulans)	624	100%	0.0	97.40%	308	XP_020123230.1					
uncharacterized protein LOC122621875 (Drosophila simulans)	530	94%	0.0	95.21%	308	XP_043691819.1					
uncharacterized protein LOC0202945 (Drosophila simulans)	523	94%	0.0	93.10%	308	XP_019370261.1					
uncharacterized protein LOC8614088 (Drosophila simulans)	520	94%	0.0	95.21%	308	XP_020285023.1					
uncharacterized protein LOC120464595 (Drosophila simulans)	513	94%	0.0	95.60%	308	XP_03665195.1					
uncharacterized protein LOC8672733 (Drosophila simulans)	508	94%	0.0	94.52%	308	XP_020981116.1					
uncharacterized protein X7084_204136 (Drosophila simulans)	491	100%	0.0	97.80%	309	KAN6344985.1					
uncharacterized protein LOC12056402 (Drosophila simulans)	486	100%	0.0	96.87%	311	XP_03665196.1					
blastofast protein X5254_305420 (Drosophila simulans)	479	94%	2e-168	79.80%	302	KAN6358387.1					
uncharacterized protein LOC119581519 (Drosophila subobscura)	478	100%	0.0	80.71%	299	XP_037207720.1					
blastofast protein X5250_300824 (Drosophila simulans)	478	94%	1e-167	79.46%	305	KAN6361803.1					
blastofast protein LOC110178689 (Drosophila simulans)	476	94%	2e-167	79.46%	305	XP_020797975.1					
uncharacterized protein LOC120152821 (Drosophila simulans)	473	100%	1e-166	80.71%	308	XP_037207720.1					
blastofast protein X5254_307721 (Drosophila simulans)	474	94%	2e-166	80.47%	308	KAN6328672.1					
uncharacterized protein LOC108154668 (Drosophila simulans)	465	100%	0.0	75.32%	314	XP_017130665.1					
uncharacterized protein LOC108188251 (Drosophila simulans)	463	100%	1e-161	84.13%	311	XP_017130665.1					

BIOS477/877 L17 - 47

47

PSI-BLAST

Sequences with E-value WORSE than threshold

E-value lower than the PSI-BLAST threshold (0.005)

Accession	Score	E-value	Bits	Positives	Identical	Accession	Score	E-value	Bits	Positives	Identical
uncharacterized protein LOC120860383 (Drosophila simulans)	524	524	49%	0.001	25.47%	578	XP_045433686.1				
uncharacterized protein LOC120330559 (Drosophila simulans)	520	520	77%	0.001	24.40%	265	XP_037967024.1				
uncharacterized protein LOC113488740 isoform X1 (Trichobius nitidus)	520	520	81%	0.002	24.69%	299	XP_036734881.1				
uncharacterized protein LOC8614088 (Drosophila simulans)	520	520	78%	0.002	25.90%	268	XP_026720333.1				
blastofast protein LSTR_L1STR04646 (Laelobothrix striatellus)	516	516	78%	0.002	22.64%	278	Q2736556.1				
uncharacterized protein LOC113488740 isoform X2 (Trichobius nitidus)	516	516	81%	0.003	24.83%	296	XP_036734881.1				
uncharacterized protein LOC105386699 isoform X1 (Phlebotomus perniciosus)	504	504	77%	0.004	24.90%	266	XP_037967024.1				

Run PSI-BLAST iteration 2 with max number of sequences: 500

Sequences with E-value WORSE than threshold

E-value lower than the PSI-BLAST threshold (0.005)

But still lower than the E-value threshold for reporting (0.05)

Within the E-value threshold for reporting (0.05)

BIOS477/877 L17 - 48

48

PSI-BLAST

Run PSI-BLAST iteration 2 with max number of sequences 500

Sequences with E-value WORSE than threshold

select all 17 sequences selected

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI-BLAST	Used to build PSSM	Newly added
uncharacterized protein LOC122680183 (Harmonea axoridis)	Harmonea axoridis	52.4	52.4	49%	0.001	25.47%	276	XP_045473886.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X2 (Phyllaea xystolella)	Phyllaea xystolella	52.0	52.0	77%	0.001	24.60%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X1 (Tritochaea nitida)	Tritochaea nitida	52.0	52.0	81%	0.002	24.60%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC122677950 (Mericella lurtzia)	Mericella lurtzia	52.0	52.0	78%	0.002	25.90%	268	XP_045793033.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
hypothetical protein LSTR_L1STR00646 (Laeodithea strathellae)	Laeodithea strathellae	51.6	51.6	76%	0.002	22.64%	278	RZ239598.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X2 (Tritochaea nitida)	Tritochaea nitida	51.6	51.6	81%	0.003	24.83%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC103386919 isoform X1 (Phyllaea xystolella)	Phyllaea xystolella	50.4	50.4	77%	0.004	24.90%	266	XP_037967932.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

BIOS477/877 L17 - 49

PSM is constructed from these sequences and used for the next search

Check to include more sequences if necessary, but be very careful!

49

PSI-BLAST

Run PSI-BLAST iteration 2 with max number of sequences 500

Sequences with E-value WORSE than threshold

select all 17 sequences selected

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI-BLAST	Used to build PSSM	Newly added
uncharacterized protein LOC122680183 (Harmonea axoridis)	Harmonea axoridis	52.4	52.4	49%	0.001	25.47%	276	XP_045473886.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X2 (Phyllaea xystolella)	Phyllaea xystolella	52.0	52.0	77%	0.001	24.60%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X1 (Tritochaea nitida)	Tritochaea nitida	52.0	52.0	81%	0.002	24.60%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC122677950 (Mericella lurtzia)	Mericella lurtzia	52.0	52.0	78%	0.002	25.90%	268	XP_045793033.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
hypothetical protein LSTR_L1STR00646 (Laeodithea strathellae)	Laeodithea strathellae	51.6	51.6	76%	0.002	22.64%	278	RZ239598.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X2 (Tritochaea nitida)	Tritochaea nitida	51.6	51.6	81%	0.003	24.84%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC103386919 isoform X1 (Phyllaea xystolella)	Phyllaea xystolella	50.4	50.4	77%	0.004	24.90%	266	XP_037967932.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

BIOS477/877 L17 - 50

PSM 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

50

PSI-BLAST

Sequences producing significant alignments

500 sequences selected

Sequences with E-value BETTER than threshold

select all 500 sequences selected

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI-BLAST	Used to build PSSM	Newly added
uncharacterized protein LOC6567893 (Drosophila persimilis)	Drosophila ps.	422	422	100%	5e-144	75.00%	420	XP_028950866.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC19184548 (Drosophila persimilis)	Drosophila ps.	417	417	100%	5e-144	75.00%	304	XP_017145668.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC4891969 (Drosophila persimilis)	Drosophila ps.	421	421	100%	9e-144	74.68%	420	XP_001389828.5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC1296912 (Drosophila persimilis)	Drosophila ps.	408	408	94%	4e-140	97.81%	308	XP_001920291.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC125821873 (Drosophila persimilis)	Drosophila ps.	407	407	94%	9e-140	95.21%	308	XP_043655919.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ORF1, Drosophila melanogaster	Drosophila m.	406	406	94%	3e-139	100.00%	308	NP_649618.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

BIOS477/877 L17 - 51

51

PSI-BLAST

First iteration (BLASTP search)

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI-BLAST	Used to build PSSM	Newly added
uncharacterized protein LOC122680183 (Harmonea axoridis)	Harmonea axoridis	52.4	52.4	49%	0.001	25.47%	276	XP_045473886.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC103386919 isoform X2 (Phyllaea xystolella)	Phyllaea xystolella	52.0	52.0	77%	0.001	24.60%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X1 (Tritochaea nitida)	Tritochaea nitida	52.0	52.0	81%	0.002	24.60%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC122677950 (Mericella lurtzia)	Mericella lurtzia	52.0	52.0	78%	0.002	25.90%	268	XP_045793033.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
hypothetical protein LSTR_L1STR00646 (Laeodithea strathellae)	Laeodithea strathellae	51.6	51.6	76%	0.002	22.64%	278	RZ239598.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC13498740 isoform X2 (Tritochaea nitida)	Tritochaea nitida	51.6	51.6	81%	0.003	24.83%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC103386919 isoform X1 (Phyllaea xystolella)	Phyllaea xystolella	50.4	50.4	77%	0.004	24.90%	266	XP_037967932.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Second iteration

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI-BLAST	Used to build PSSM	Newly added
uncharacterized protein LOC103386919 isoform X2 (Phyllaea xystolella)	Phyllaea xystolella	151	151	84%	6e-46	73.92%	266	XP_037967934.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC103386919 isoform X1 (Phyllaea xystolella)	Phyllaea xystolella	149	149	84%	4e-39	72.51%	266	XP_037967932.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein D0YJ4_001208 (Sarcotraga bullata)	Sarcotraga s.	141	141	40%	2e-37	67.19%	44	TM055323.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

BIOS477/877 L17 - 52

- 1st iteration has the real E-values for the query.
- After the 2nd iteration, E-values are for the PSSM.

52

PSI-BLAST

First iteration (BLASTP search)

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI-BLAST	Used to build PSSM	Newly added
gains 1 (Drosophila melanogaster)	Drosophila melanogaster	838	838	100%	0.0	100.00%	308	NP_649618.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC5726927 (Drosophila simulans)	Drosophila sim.	824	824	100%	0.0	99.0%	308	XP_001920293.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC122671875 (Drosophila basalis)	Drosophila basalis	530	530	94%	0.0	99.0%	308	XP_043656819.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC6550244 (Drosophila areolaris)	Drosophila are.	523	523	94%	0.0	99.0%	308	XP_001979038.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC614060 (Drosophila sechelliae)	Drosophila sechelliae	520	520	94%	0.0	95.21%	308	XP_002082823.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
uncharacterized protein LOC12454180 (Drosophila santomea)	Drosophila santomea	513	513	94%	0.0	95.59%	308	XP_039495195.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

BIOS477/877 L17 - 53

E-values can become lower or higher with more iterations depending on the sequences included to build PSSM!

53