BLAST & PSI-BLAST

Lecture outline

- BLAST
- PSI-BLAST

Sequence database similarity searching

- A process involves submission of a query sequence and performing a pairwise comparison of the query sequence with all individual sequences in a database.
- Requirements for database searching:

  ① Sensitivity – the ability to find as many correct hits as possible

  ② Specificity – the ability to exclude incorrect hits

  ③ Speed – the time it takes to get results from database searches
- A compromise between the three criteria often has to be made.

Why search sequence database

- Fundamental to understanding the relatedness of any query sequence to other known proteins or DNA sequences in the database.
- One of the most effective ways to assign putative functions to newly determined sequences.
- Applications include
  - identifying homologs
  - discovering new genes or proteins
  - discovering variants of genes or proteins
  - exploring protein structure and function

Dynamic programming and heuristic algorithms

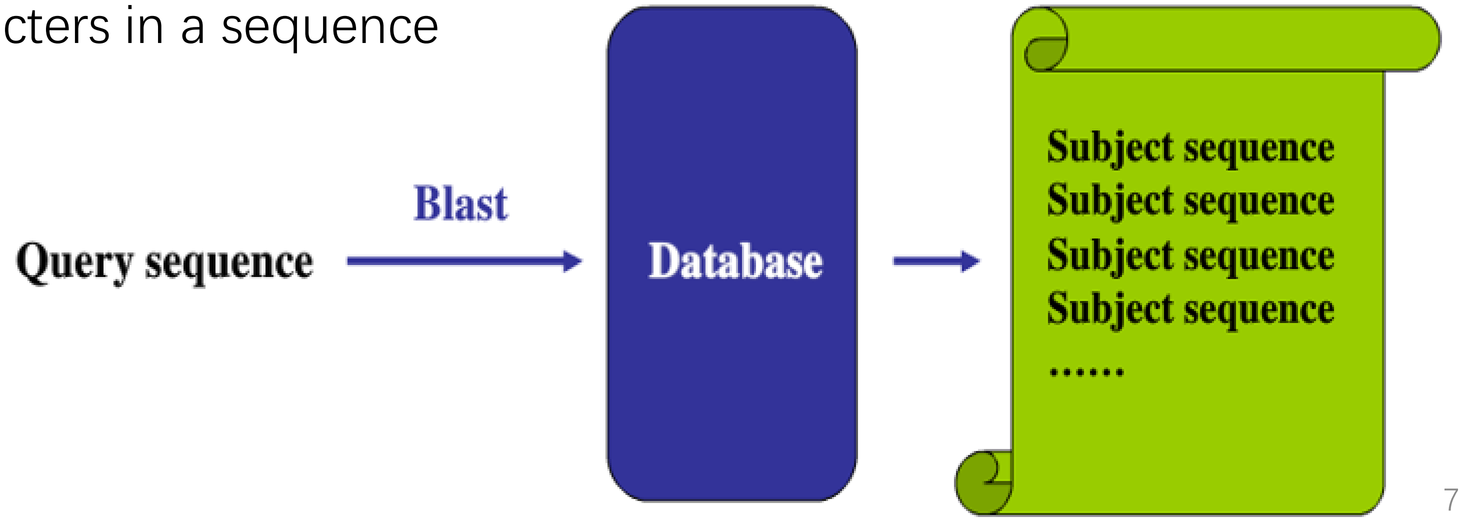
- Dynamic programming
  - An exhaustive method and computationally very intensive
  - Guaranteed to find the optimal solution
  - Too slow and impractical when computational resources are limited
- Heuristic method
  - Much faster than dynamic programming
  - Not guaranteed to find the optimal solution
  - It is often used because of the need for obtaining results within a realistic time frame without significantly sacrificing the accuracy of the computational output.

BLAST

BLAST-Basic Local Alignment Search Tool

BLAST-Basic Local Alignment Search Tool

- Developed in 1990
- BLAST uses **heuristics** to align a query sequence with all sequences in a database.
- The heuristic algorithms perform faster searches because they examine only a fraction of the possible alignments examined in regular dynamic programming.
- Good balance of speed, sensitivity and specificity
- **Basic assumption** – two related sequences must have at least one **word** in common
- **Word**: a short string of characters in a sequence



① Choose the query sequence

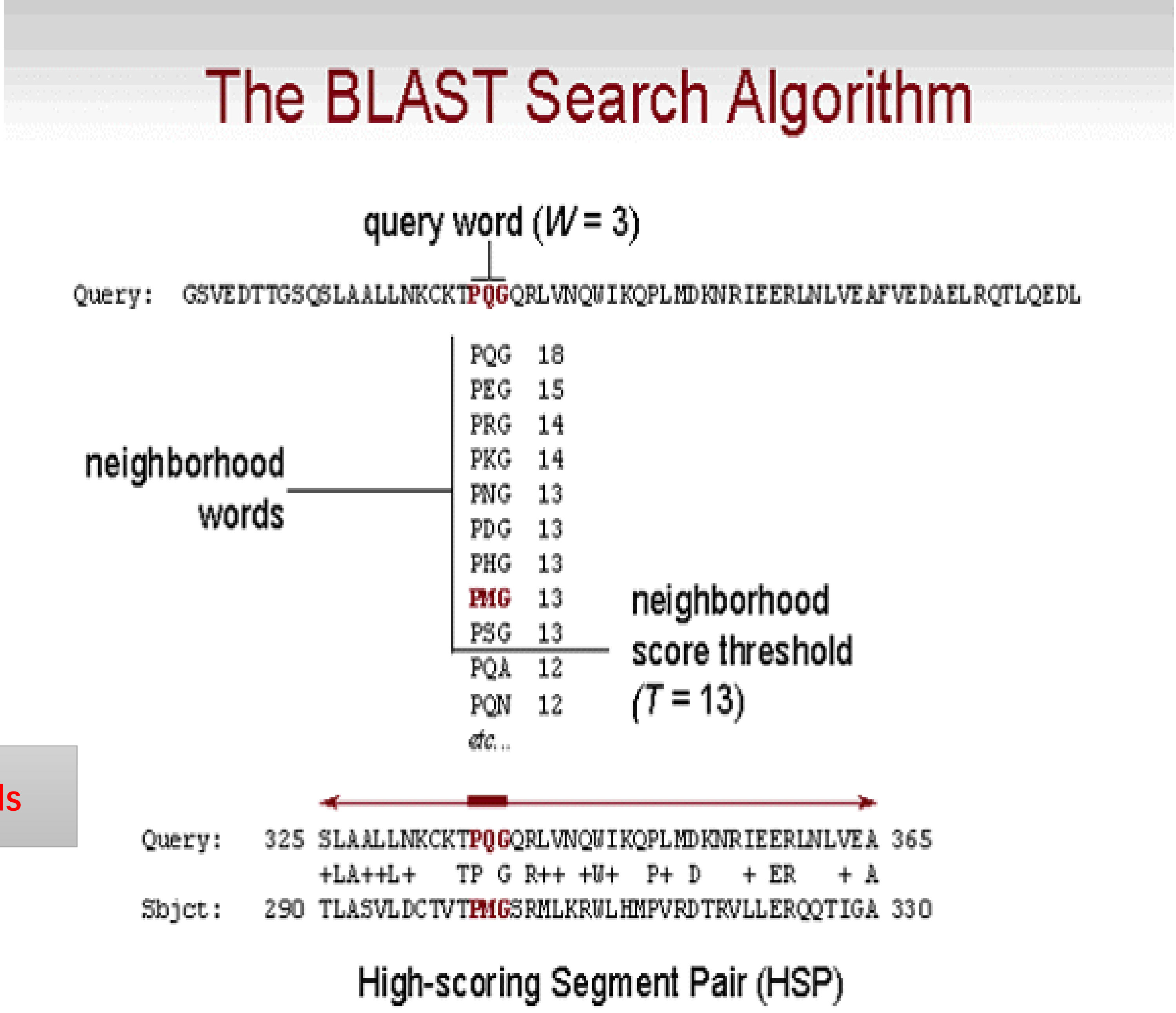
② Create a list of words from the query sequence

Each **word** is typically 6 consecutive residues for protein sequences and 28 consecutive bases for DNA sequences.

GSVEDTTGSQSL. . .

```
GSV
 SVE
  VED
   EDT
    DTT
     TTG
      TGS
       GSQ
        QSL
      . . .
```

Query words



The BLAST Search Algorithm

Query: GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++  +W+  P+ D    +ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

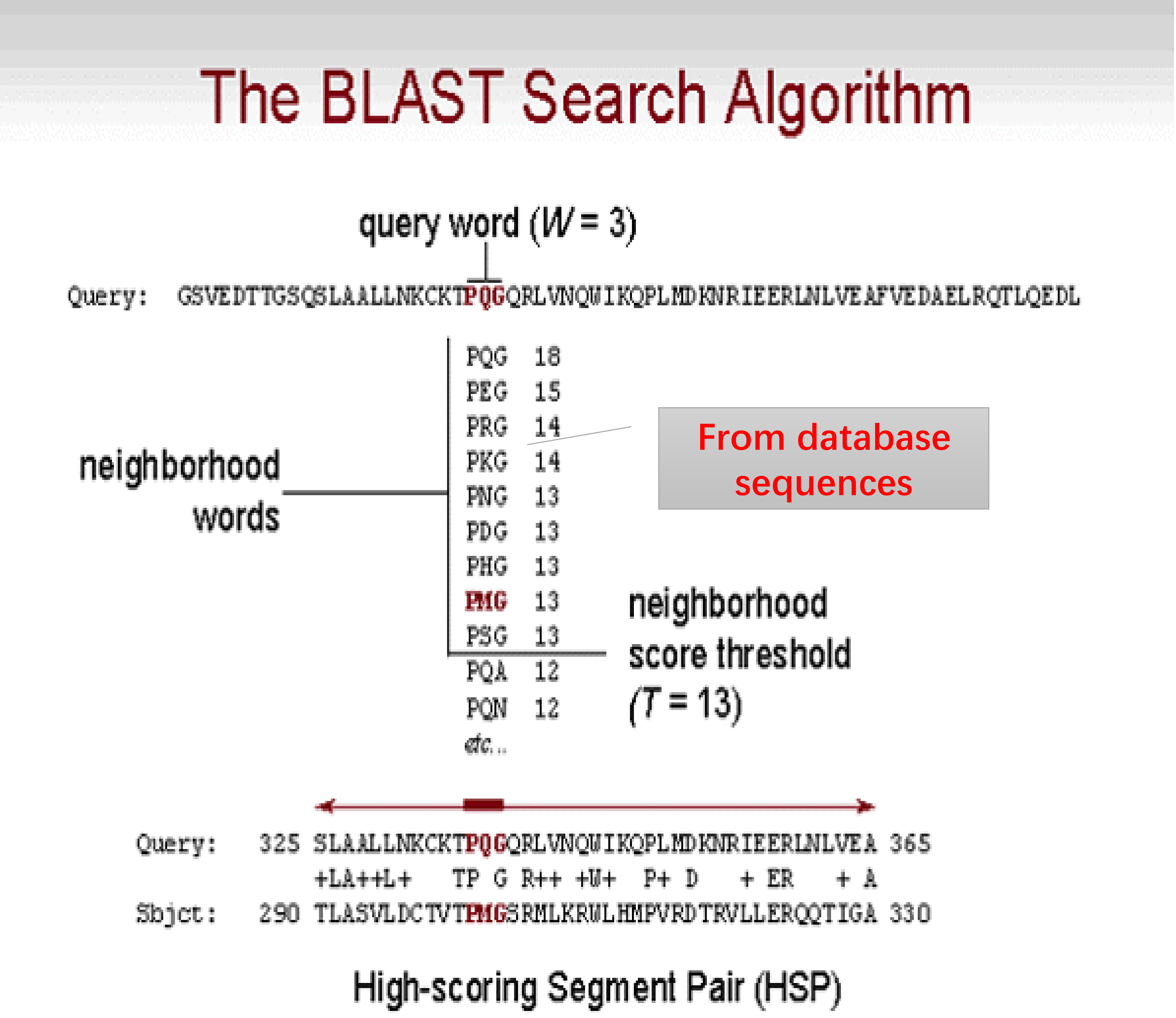
High-scoring Segment Pair (HSP)

① Choose the query sequence

② Create a list of words from the query sequence

③ Scan each database sequence for the occurrence of the query words

- The matching of the words is scored by a given substitution matrix.
- A word is considered a match if its score >= *T*



The BLAST Search Algorithm

query word (*W* = 3)

```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

neighborhood score threshold (*T* = 13)

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+    TP G R++ +W+  P+ D    + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

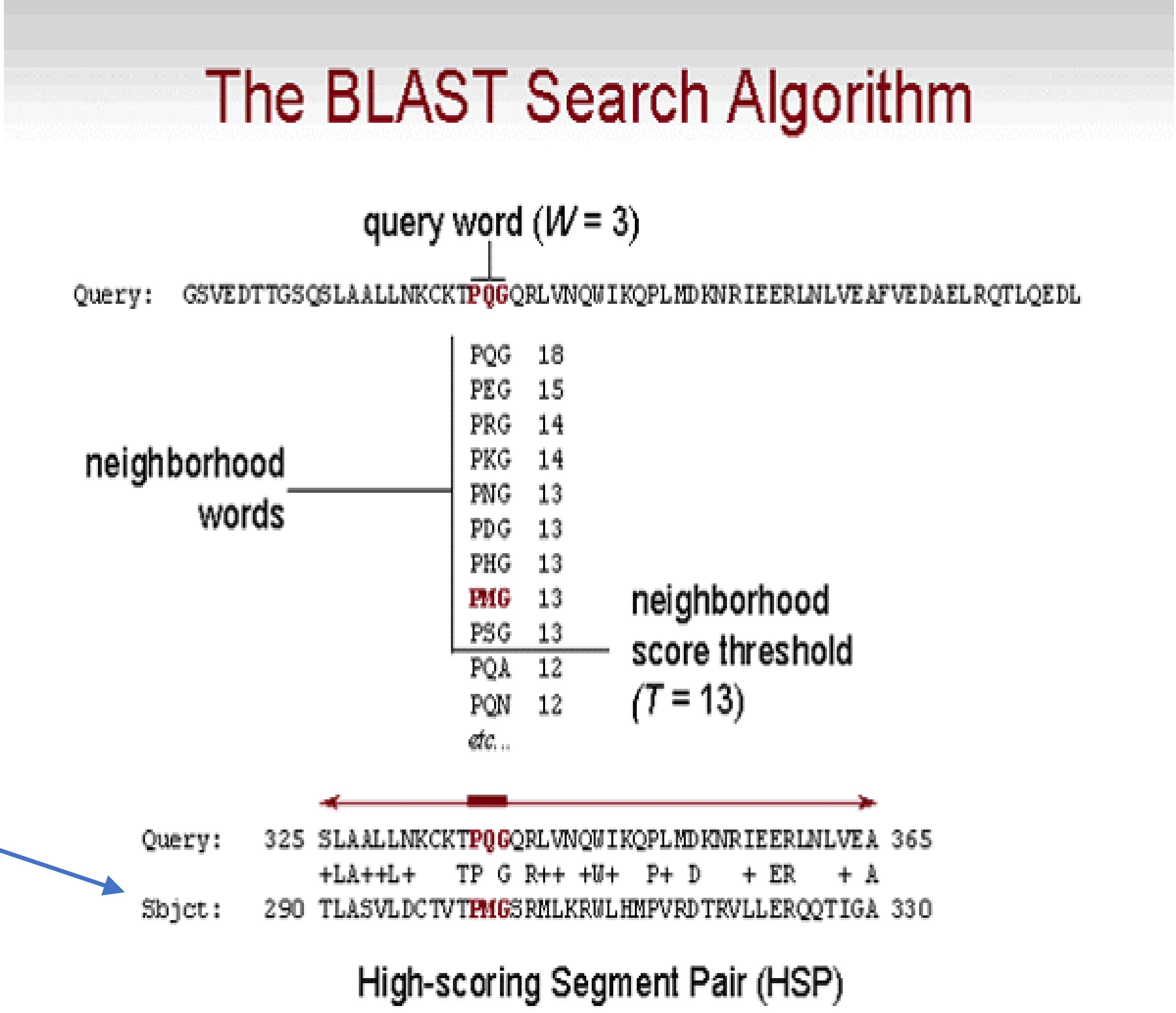
High-scoring Segment Pair (HSP)

① Choose the query sequence

② Create a list of words from the query sequence

③ Scan each database sequence for the occurrence of the query words

④ Identify database sequences containing the neighborhood words



The BLAST Search Algorithm

```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

```
Query:   325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
             +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:   290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

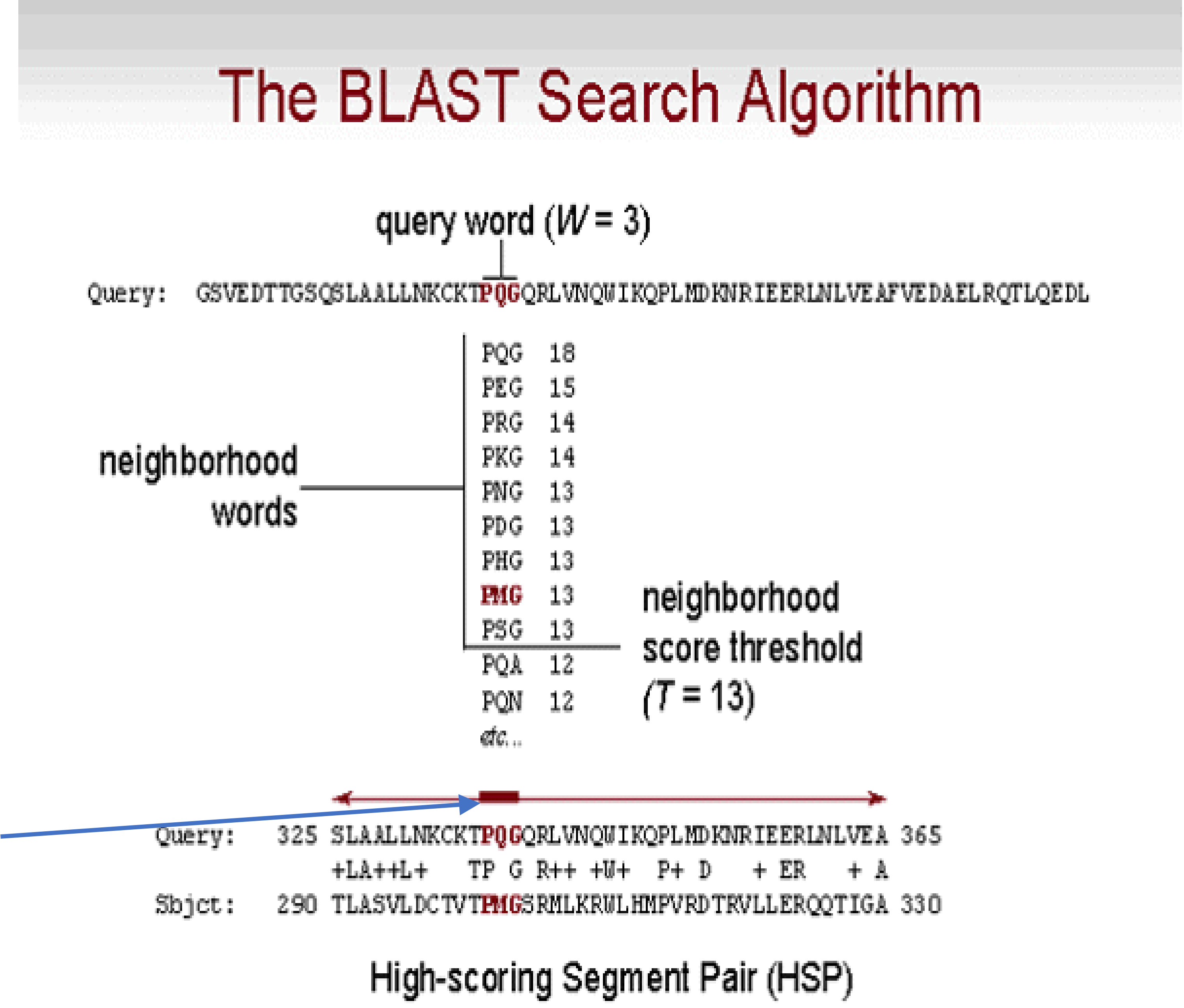
① Choose the query sequence

② Create a list of words from the query sequence

③ Search neighborhood words from a sequence database

④ Identify database sequences containing the neighborhood words

⑤ For each **match**, extend ungapped alignment in both directions while counting the alignment score



The BLAST Search Algorithm

query word (*W* = 3)

```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

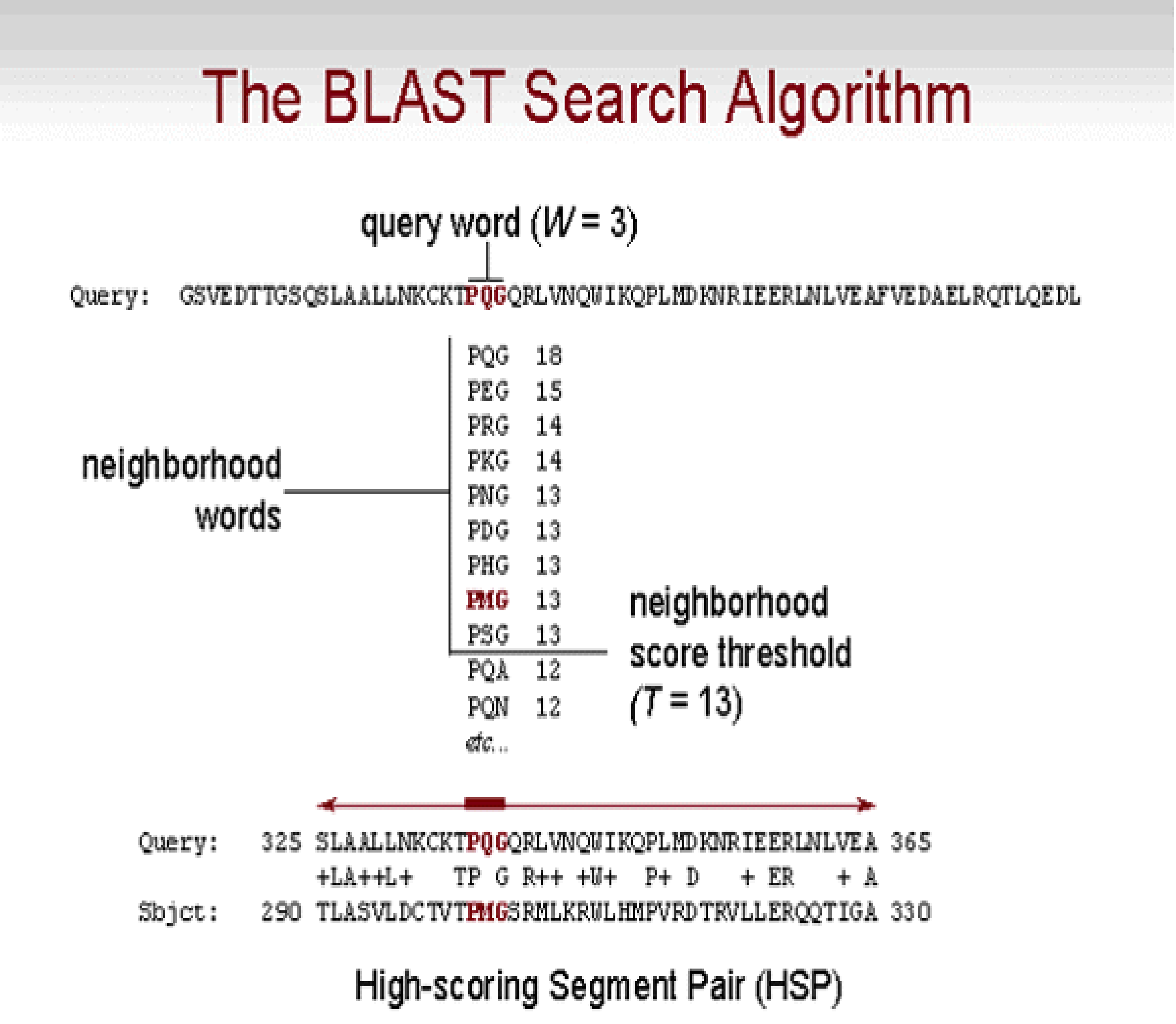
neighborhood score threshold (*T* = 13)

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+    TP G R++ +W+  P+ D    + ER    + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

⑥ The extension continues until the score of the alignment drops below a threshold due to mismatches. The resulting contiguous aligned segment pair without gaps is called **high-scoring segment pair** (**HSP**)

⑦ Return the HSPs – local alignments between the query and database sequences



The BLAST Search Algorithm

```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc. ..	

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

BLAST-Basic Local Alignment Search Tool

- Using a heuristic method, BLAST finds similar sequences by locating short matches (i.e. hit words) between the two sequences.
- For each short match, it extends ungapped alignment in both directions to make local alignments.
- The BLAST algorithm is less accurate than the Smith-Waterman algorithm but over 50-100 times faster.

Statistical significance

- The BLAST output provides a list of pairwise sequence matches ranked by statistical significance.
- In BLAST searches, this statistical indicator is known as the **E-value**.
- E-value (expectation value) – the chance that the match could be random
- **The lower the E-value, the more significant the match.**

E-value

$$E = m \times n \times P$$

where m is the total number of residues in a database

n is the number of residues in the query sequence

P is the probability that an HSP alignment is a result of random chance

- e.g. if E=10, 10 matches with scores this high are expected to be found by chance

The E-value is dependent on the length of the database!!

As the database grows, the E-value for a given sequence match also increases.

Previously detected homologs may be lost as the database enlarges!!

Bit score

- Another prominent statistical indicator used in addition to the E-value in a BLAST output.
- Normalized based on the raw pairwise alignment score.

$$S' = (\lambda \times S - \ln K)/ \ln 2$$

where λ is the Gumble distribution constant, S is the raw alignment score, and K is a constant associated with the scoring matrix used.

- The higher the bit score, the more highly significant the match is.
- Measures sequence similarity independent of query sequence length and database size.

A simple example

- A database containing a total of 10^12 residues
- A query sequence of 100 residues
- A *P*-value for the ungapped HSP region in one of the database matches is 10^(-20)

- $E = 100 \times 10^{12} \times 10^{-20}$

  $= 10^{-6}$

- This indicates that the probability of this database sequence match occurring due to random chance is 10−6

Empirical interpretation of E-value

- **$E = 0$** – means that the two sequences are statistically **identical**
- **$E < 10^{-4}$** – the database match can be considered homologous or related to the query sequence

BLAST programs

Program	Database (Subject)	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nt. → Protein
TBLASTN	Nt. → Protein	Protein
TBLASTX	Nt. → Protein	Nt. → Protein

BLAST programs

- **BLASTN** queries **nucleotide** sequences with a **nucleotide** sequence database.
- **BLASTP** uses **protein** sequences as queries to search against a **protein** sequence database.
- **BLASTX** uses **nucleotide** sequences as queries and translates them to produce translated protein sequences, which are used to query a **protein** sequence database.
- **TBLASTN** queries **protein** sequences to a **nucleotide** sequence database with the sequences translated to proteins.
- **TBLASTX** uses **nucleotide** sequences, which are translated to proteins, to search against a **nucleotide** sequence database that has all the sequences translated to proteins as well.

Four steps to a BLAST search

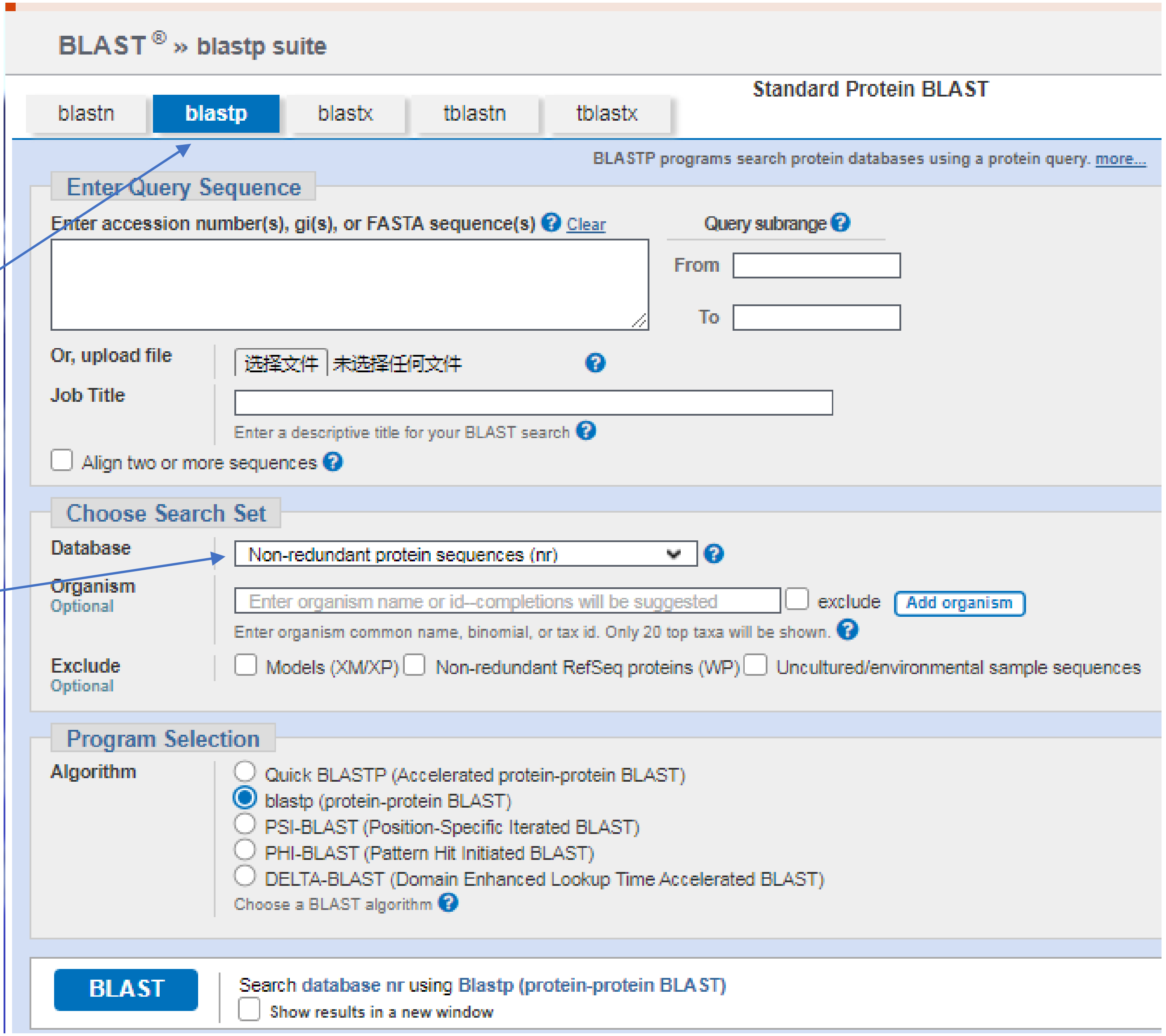
(1) Choose the sequence (query)

(2) Select the BLAST program

(3) Choose the database to search

(4) Choose optional parameters

Then click "BLAST"



BLAST® » blastp suite

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

Query subrange ?

From

To

Or, upload file 选择文件 未选择任何文件 ?

Job Title

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

Choose Search Set

Database: Non-redundant protein sequences (nr) ?

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ?

Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm

○ Quick BLASTP (Accelerated protein-protein BLAST)

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

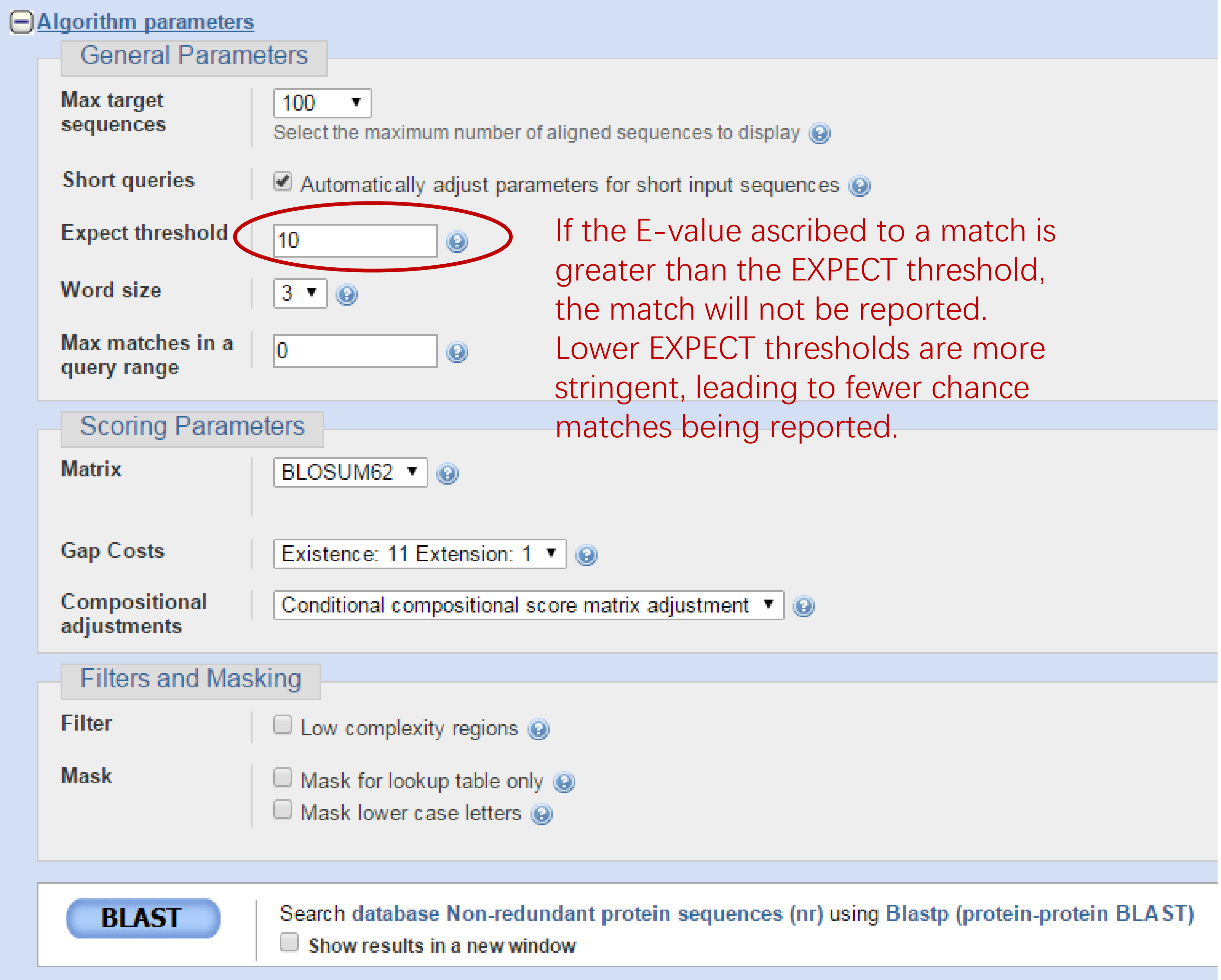
Choose a BLAST algorithm ?

BLAST

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

http://blast.ncbi.nlm.nih.gov/Blast.cgi



⊟ Algorithm parameters

General Parameters

Max target sequences: 100 ▼
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3 ▼

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62 ▼

Gap Costs: Existence: 11 Extension: 1 ▼

Compositional adjustments: Conditional compositional score matrix adjustment ▼

Filters and Masking

Filter: □ Low complexity regions

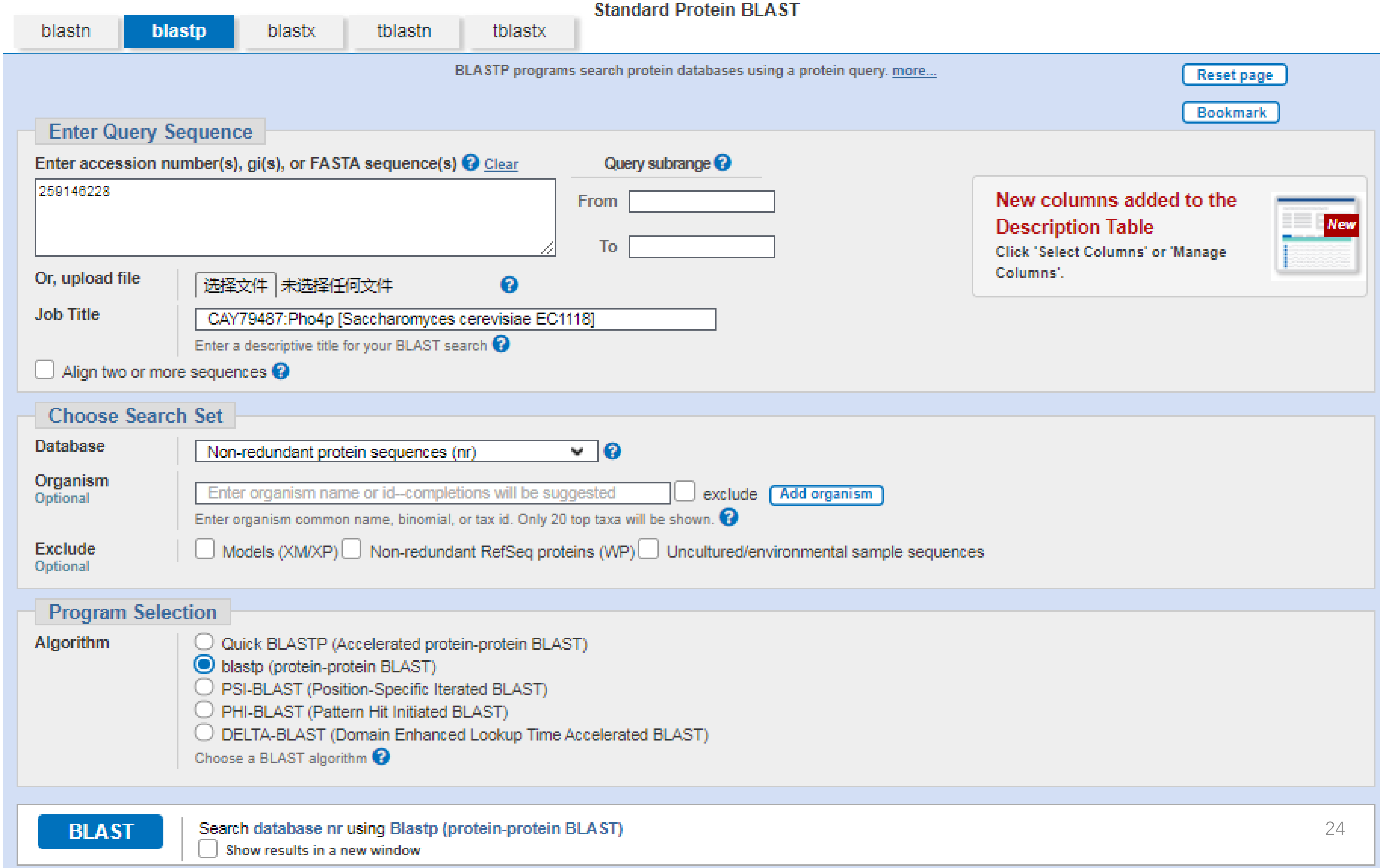
Mask: □ Mask for lookup table only
□ Mask lower case letters

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
□ Show results in a new window

If the E-value ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported.

The BLAST output – hit list

- BLAST lists the best matches (hits)
    - For each hit, BLAST provides:
        - Accession number – links to Genbank flatfile
        - Description
        - E-value
            - An indicator of how good a match to the query sequence
        - Score
            - Link to an alignment



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

259146228

Query subrange ?

From

To

Or, upload file 选择文件 未选择任何文件 ?

Job Title: CAY79487:Pho4p [Saccharomyces cerevisiae EC1118]

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

New columns added to the Description Table

Click 'Select Columns' or 'Manage Columns'.

Choose Search Set

Database: Non-redundant protein sequences (nr) ?

Organism
Optional

Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ?

Exclude
Optional

□ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm

- ○ Quick BLASTP (Accelerated protein-protein BLAST)
- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)
- ○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm ?

Search database nr using Blastp (protein-protein BLAST)

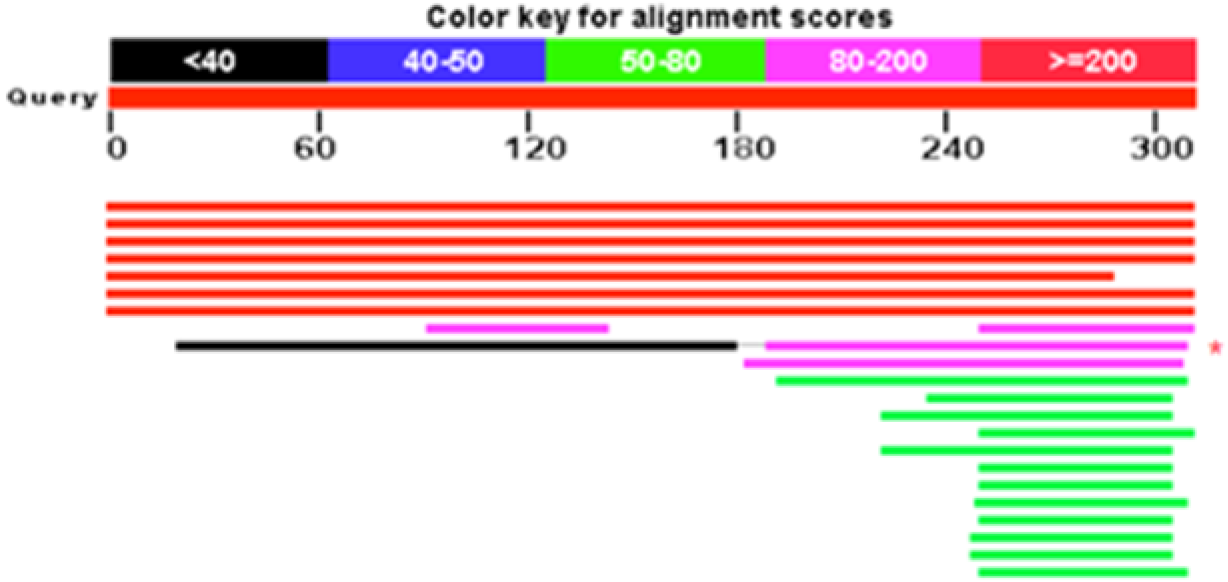
□ Show results in a new window

Example: BLAST - Pho4p (*S. cerevisiae*)

```
>gi|259146228|emb|CAY79487.1| Pho4p [Saccharomyces cerevisiae EC1118]
MGRTTSEGIHGFVDDLEPKSSILDKVGDFITVNTKRHDGREDFNEQNDELNSQEHHNSSENGNENENEQD
SLALDDLDRAFELVEGMDMDWMMPSHAHHSPATTATIKPRLLYSPLIHTQSAVPVTISPNLVATATSTTS
ANKVTKNKSNSSPYLNKRRGKPGPDSATSLFELPDSVIPTPKPKPKPKQYPKVILPSNSTRRISPVTAKT
SSSAEGVVVASESPVIAPHGSSHSRSLSKRRSSGALVDDDKRESHKHAEQARRNRLAVALHELASLIPAE
WKQQNVSAAPSKATTVEAACRYIRHLQQNVST
```

Query (input) sequence
(Pho4p from *S. cerevisiae*)

BLAST (default parameters)



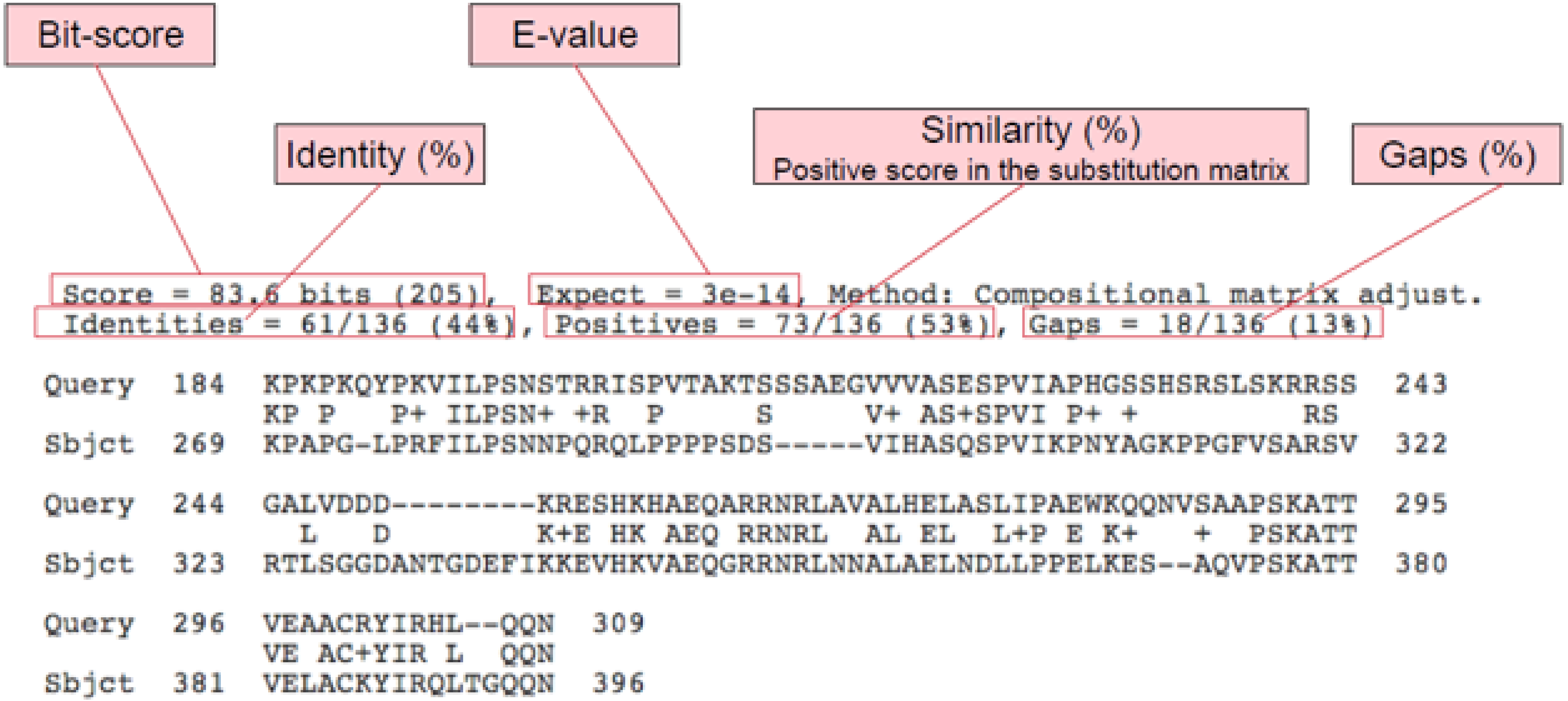
Results (output) of BLAST

- The top segment displays the color key and the query based scale.
- The colored bars represent the actual HSPs. The position of each bar indicates the region of the query the HSP covers.
- The thin line (see *) indicates that the two HSPs are from the same sequence.

Explanation of Output of a BLAST Search:
http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/new_view.html

Example: BLAST - Pho4p (*S. cerevisiae*)

Results (output) of BLAST



Bit-score

E-value

Identity (%)

Similarity (%)
Positive score in the substitution matrix

Gaps (%)

```
 Score = 83.6 bits (205),  Expect = 3e-14, Method: Compositional matrix adjust.
 Identities = 61/136 (44%), Positives = 73/136 (53%), Gaps = 18/136 (13%)

Query  184  KPKPKQYPKVILPSNSTRRISPVTAKTSSSAEGVVVASESPVIAPHGSSHSRSLSKRRSS  243
            KP P   P+ ILPSN+ +R  P     S     V+ AS+SPVI P+ +         RS
Sbjct  269  KPAPG-LPRFILPSNNPQRQLPPPPSDS-----VIHASQSPVIKPNYAGKPPGFVSARSV  322

Query  244  GALVDDD--------KRESHKHAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATT  295
              L   D        K+E HK AEQ RRNRL  AL EL  L+P E K+   +   PSKATT
Sbjct  323  RTLSGGDANTGDEFIKKEVHKVAEQGRRNRLNNALAELNDLLPPELKES--AQVPSKATT  380

Query  296  VEAACRYIRHL--QQN  309
            VE AC+YIR L  QQN
Sbjct  381  VELACKYIRQLTGQQN  396
```

BLAST makes suggestions

- BLAST takes a query sequence
- Compares it with all sequences in the databases
  - By constructing local alignments
- Lists those that appear to be similar to the query sequence
  - The “hit list”
- BLAST doesn’t offer a direct binary decision on whether two sequences are related or not. Just Tells you why it thinks they are homologs
  - E-value
  - Bit score
  - Graphic display
  - Alignment

BLAST makes suggestions, You draw the conclusions!

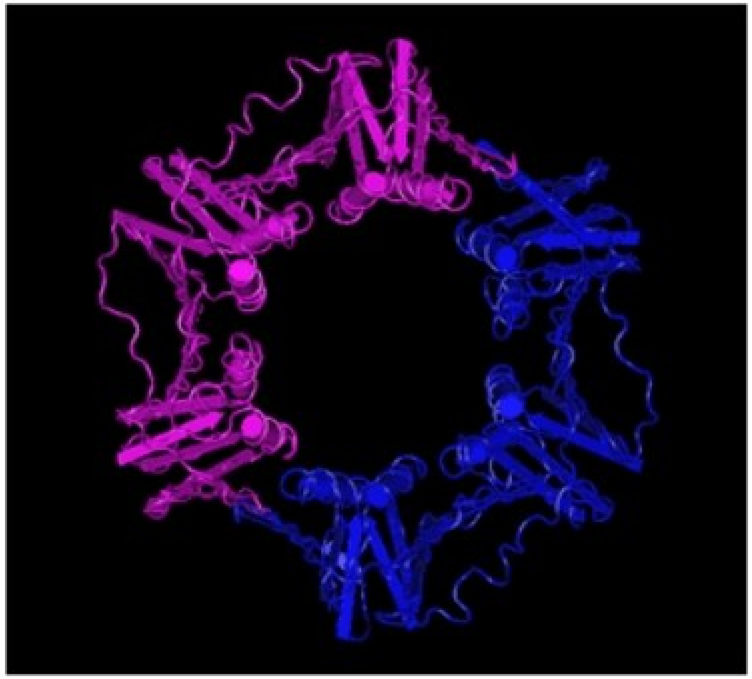
- Look at E-value
- Look at graphic display
- Look at alignment

- Make your best guess!

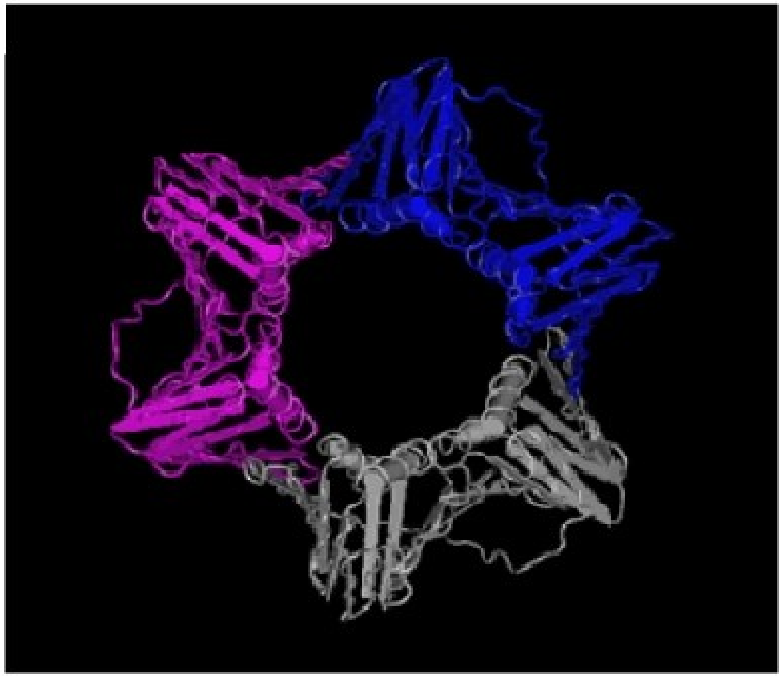
PSI-BLAST

Position specific iterated BLAST

Distant homology detection - example



Escherichia coli DNA polymerase III β-subunit
PDB accession number 2POL



Human proliferating cell nuclear antigen (PCNA)
PDB accession number 1AXC

The conventional BLASTp program detects no sequence similarity between these two proteins. This distant sequence similarity, however, can be detected by PSI-BLAST.

Position specific iterated BLAST: PSI-BLAST

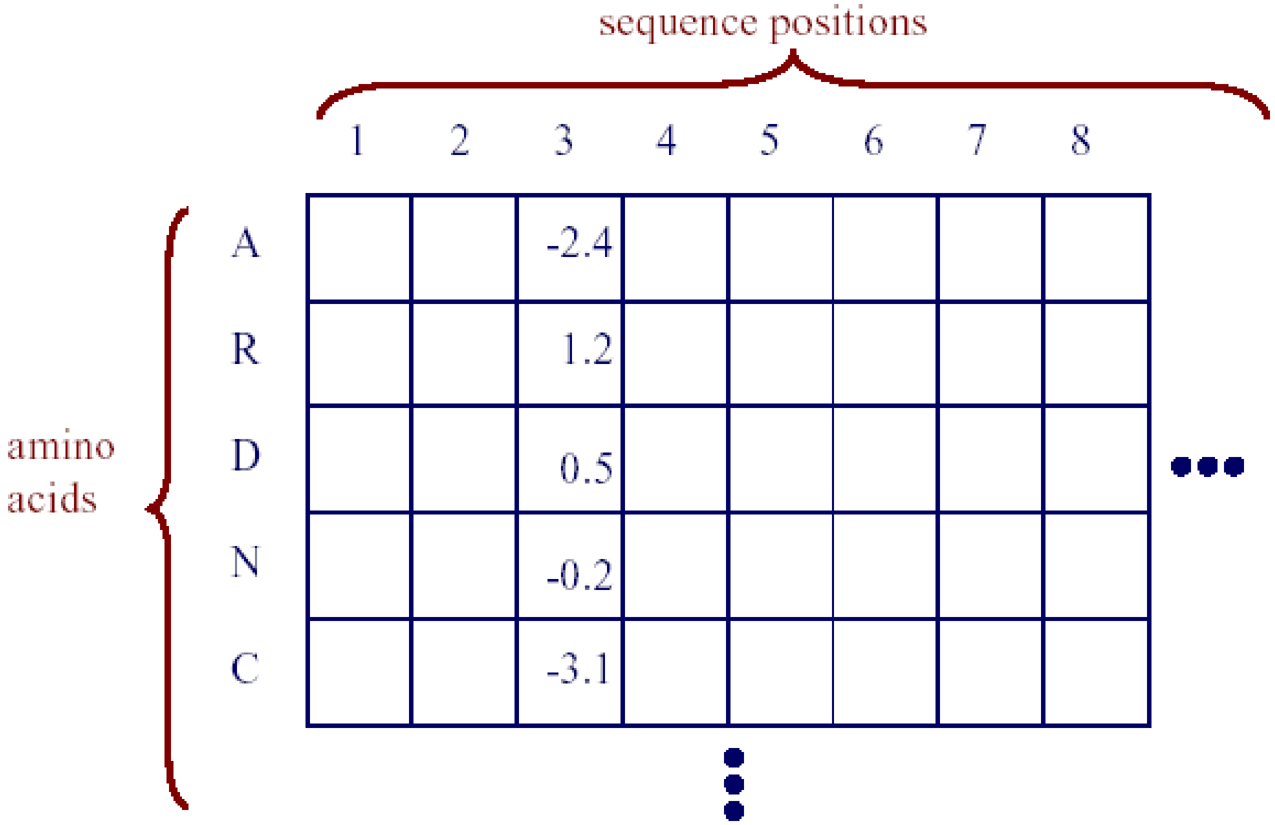
- A **protein** sequence **profile** search program
- The program first performs a **BLASTP** database search. The PSI-BLAST program uses the information from any significant alignments returned to construct a **position-specific score matrix (PSSM)**, which replaces the query sequence for the next round of database searching.
- PSI-BLAST may be iterated until no new significant alignments are found.

Position specific iterated BLAST: PSI-BLAST

- At this time PSI-BLAST may be used only for comparing **protein queries** with **protein databases**.
- More sensitive than standard BLAST to look for sequence homologues
- Used for distant homology detection
- Results should be interpreted carefully – Can include false homologous!!

Position-specific scoring matrix - PSSM

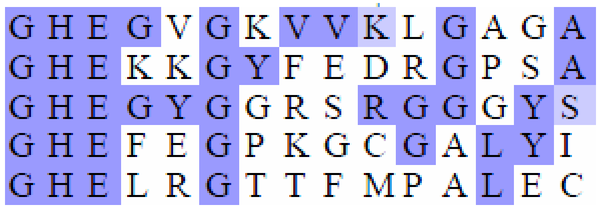
- A PSSM or profile is an n by m matrix, where n is the size of the alphabet, and m is the length of the sequence.
- The entry at (i, j) is the score assigned by the PSSM to letter i at the jth position.



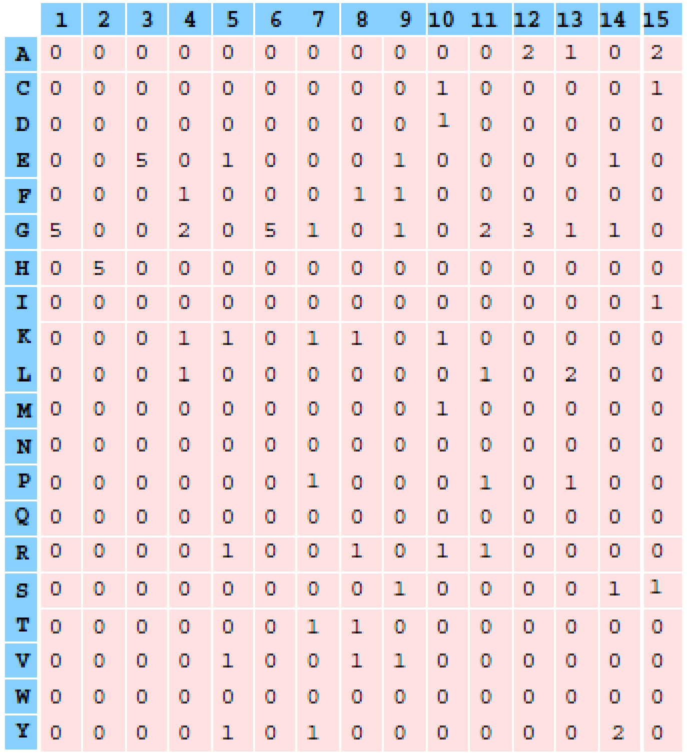
	1	2	3	4	5	6	7	8
A			-2.4					
R			1.2					
D			0.5					
N			-0.2					
C			-3.1					

How to build a PSSM – a simple example

- The profile contains scores that describe the frequency of each of the twenty amino acid residues to be at each profile position.



```
G H E G V G K V V K L G A G A
G H E K K G Y F E D R G P S A
G H E G Y G G R S R G G G Y S
G H E F E G P K G C G A L Y I
G H E L R G T T F M P A L E C
```

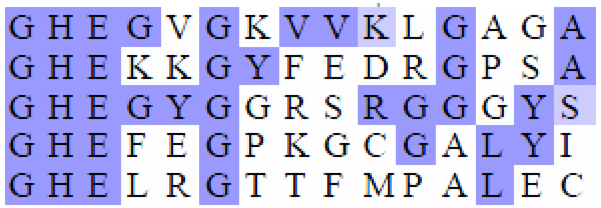


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3	1	1	0
H	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
T	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0	0	2	0

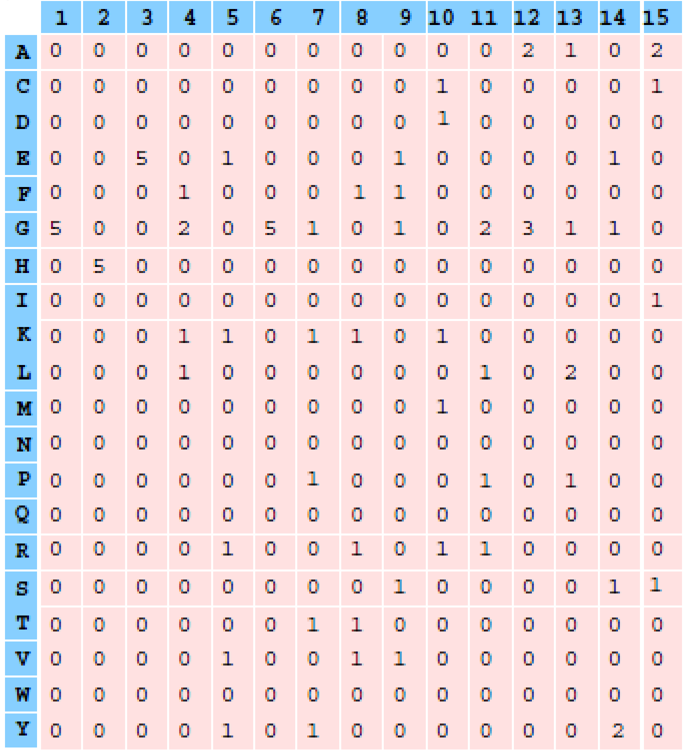
- Column 1: $f_{A,1} = \frac{0}{5} = 0,\ f_{G,1} = \frac{5}{5} = 1,\ \ldots$
- Column 2: $f_{A,2} = \frac{0}{5} = 0,\ f_{H,2} = \frac{5}{5} = 1,\ \ldots$
- ...
- Column 15: $f_{A,15} = \frac{2}{5} = 0.4,\ f_{C,15} = \frac{1}{5} = 0.2,\ \ldots$

How to build a PSSM – a simple example

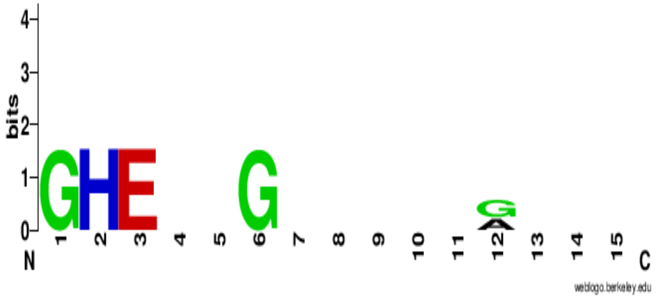
- The logo graphically displays the sequence conservation at a particular position in the alignment of sequences.
- The height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position.



```
GHEGVGKVVKLGAGA
GHEKKGYFEDRGPSA
GHEGYGGRSRGGGYS
GHEFEGPKGCGALYI
GHELRGTTFMPALEC
```



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3	1	1	0
H	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
T	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0	0	2	0

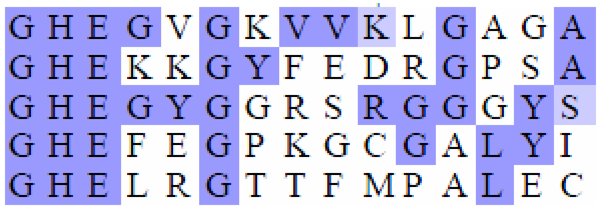


http://weblogo.berkeley.edu/logo.cgi

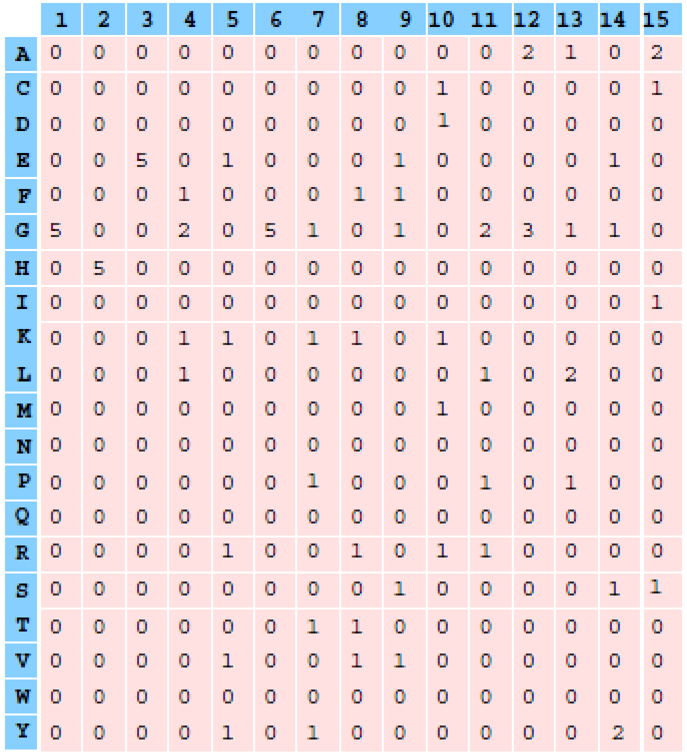
- Column 1: $f_{A,1} = \frac{0}{5} = 0,\ f_{G,1} = \frac{5}{5} = 1,\ \ldots$
- Column 2: $f_{A,2} = \frac{0}{5} = 0,\ f_{H,2} = \frac{5}{5} = 1,\ \ldots$
- ...
- Column 15: $f_{A,15} = \frac{2}{5} = 0.4,\ f_{C,15} = \frac{1}{5} = 0.2,\ \ldots$

How to build a PSSM – a simple example

- The PSSM captures the conservation pattern in alignment and stores it as a matrix of scores for each position in the alignment.
- In the PSSM, high conserved positions receive high scores and weakly conserved positions receive low scores.

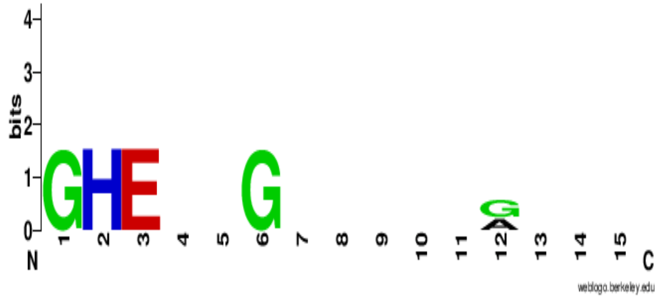


```
G H E G V G K V V K L G A G A
G H E K K G Y F E D R G P S A
G H E G Y G G R S R G G G Y S
G H E F E G P K G C G A L Y I
G H E L R G T T F M P A L E C
```



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3	1	1	0
H	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
T	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0	0	2	0

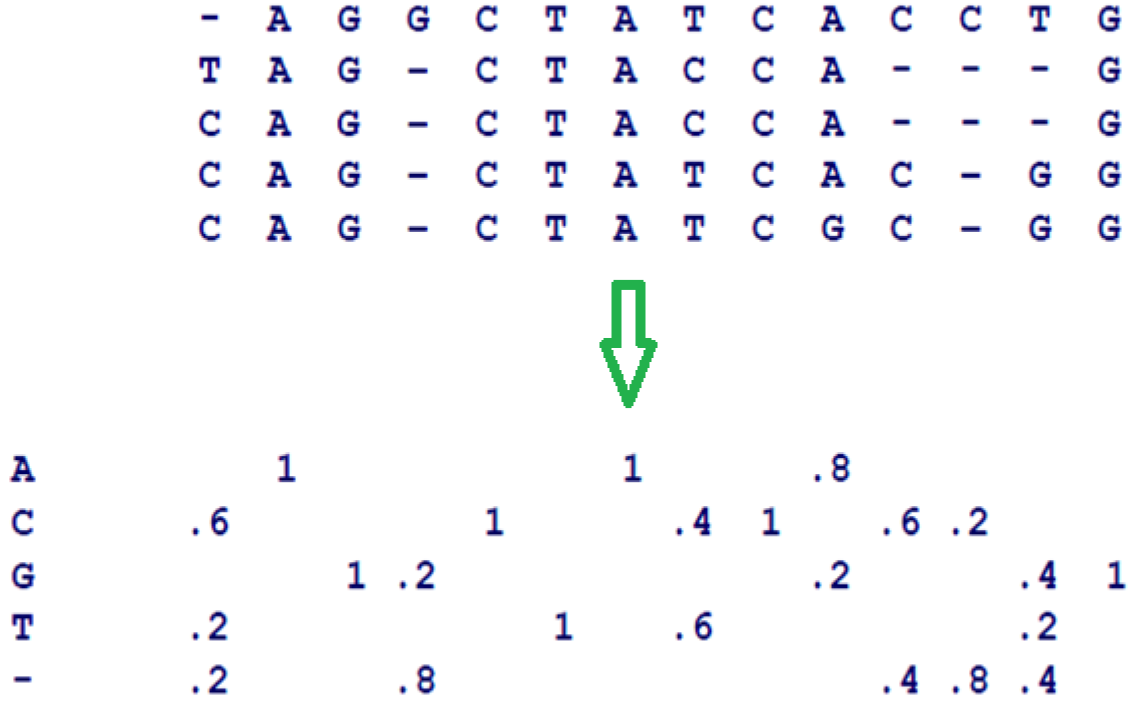
- Column 1: $f_{A,1} = \frac{0}{5} = 0,\ f_{G,1} = \frac{5}{5} = 1,\ \ldots$
- Column 2: $f_{A,2} = \frac{0}{5} = 0,\ f_{H,2} = \frac{5}{5} = 1,\ \ldots$
- …
- Column 15: $f_{A,15} = \frac{2}{5} = 0.4,\ f_{C,15} = \frac{1}{5} = 0.2,\ \ldots$



http://weblogo.berkeley.edu/logo.cgi

PSI-BLAST principle

1. A standard BLASTP search is performed against a database
2. Multiple sequence alignment (MSA) of significant alignments
3. A PSSM is constructed automatically from a MSA to captures the conservation pattern in alignment. High conserved positions receive high scores and weakly conserved positions receive low scores.



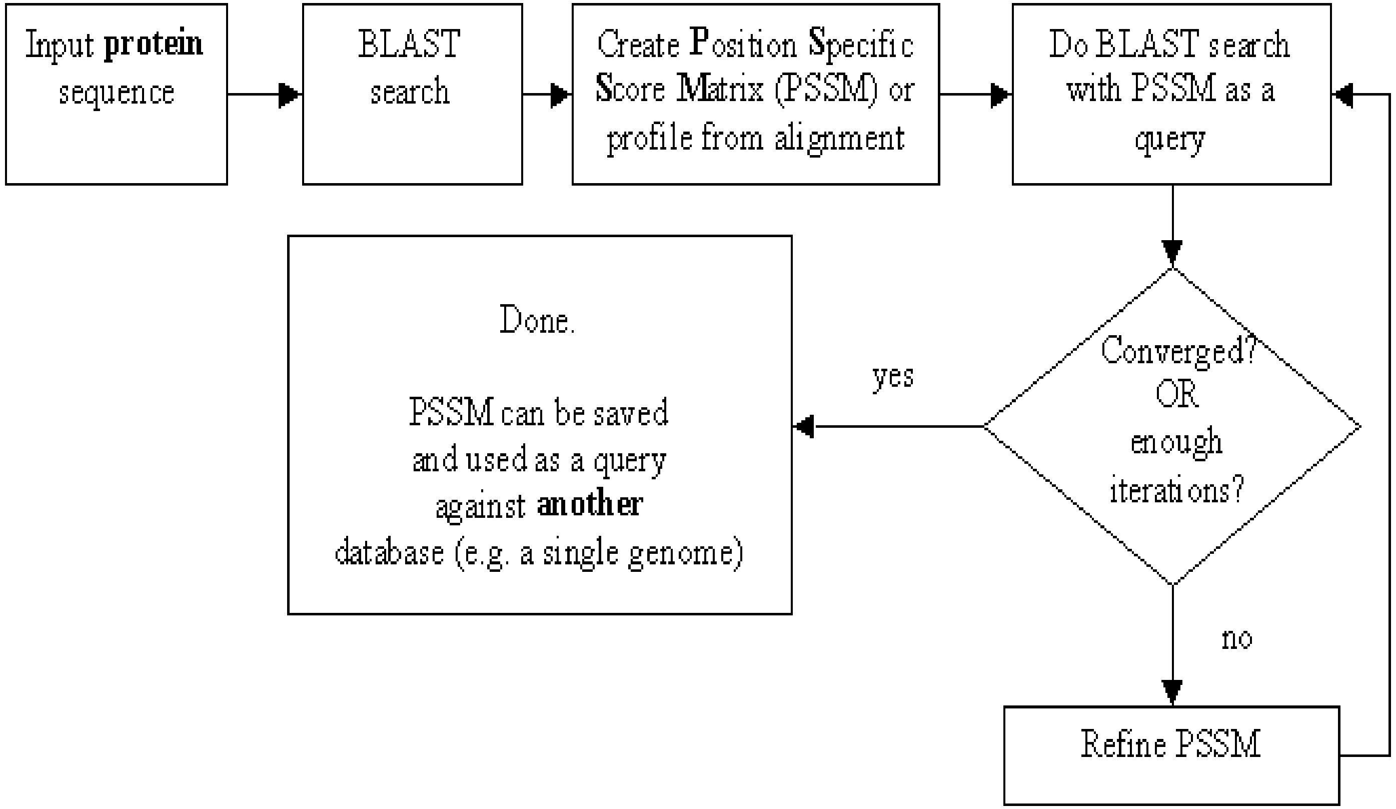
```
- A G G C T A T C A C C T G
T A G - C T A C C A - - - G
C A G - C T A C C A - - - G
C A G - C T A T C A C - G G
C A G - C T A T C G C - G G
```

A		1					1			.8				
C	.6				1			.4	1		.6	.2		
G			1	.2						.2			.4	1
T	.2					1		.6					.2	
-	.2			.8							.4	.8	.4	

PSI-BLAST principle

4. The PSSM replaces the query sequence for a further search of the database to detect sequences that match the conservation pattern specified by the PSSM.
5. Estimate the statistical significance of the new found sequences: E-value

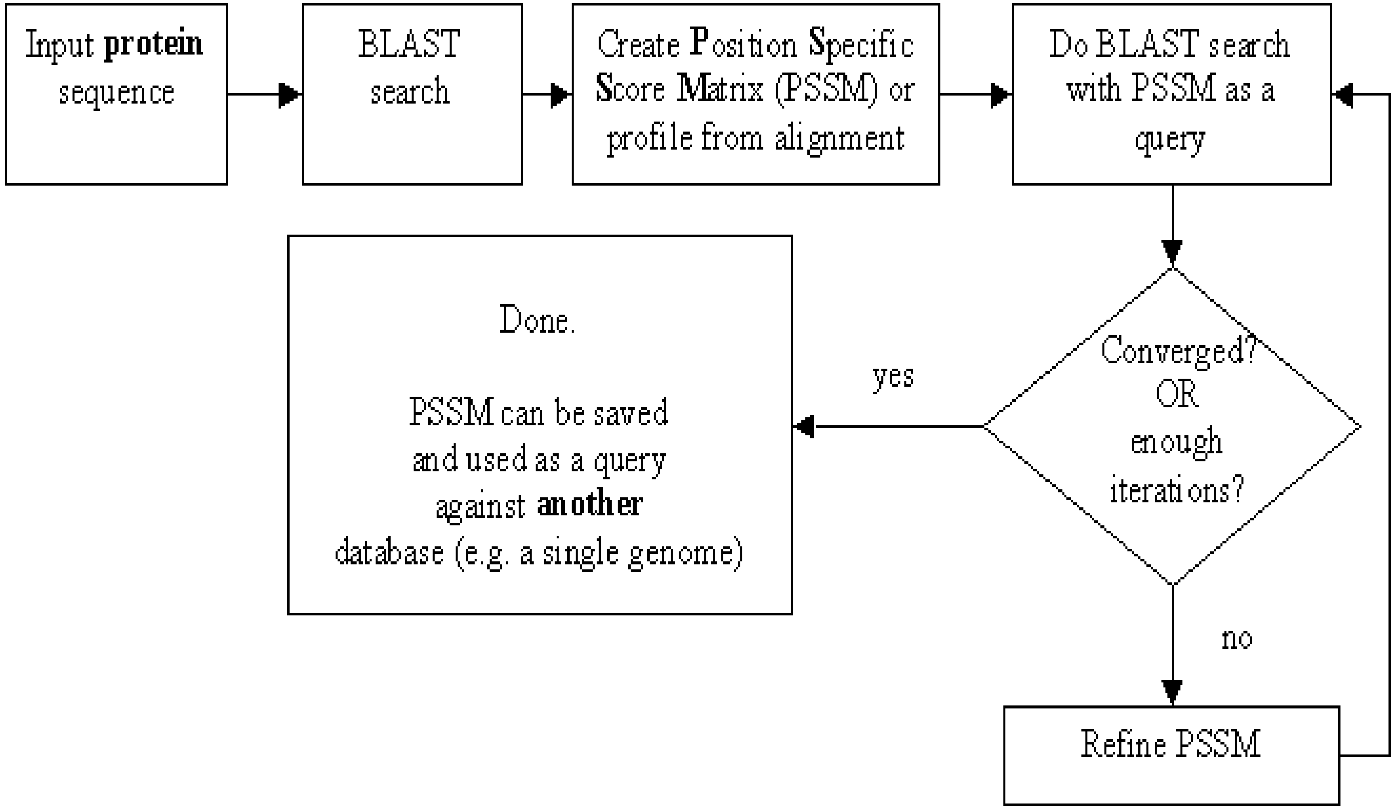
   Default E-value = 0.005


Input **protein** sequence
BLAST search
Create Position Specific Score Matrix (PSSM) or profile from alignment
Do BLAST search with PSSM as a query
Converged? OR enough iterations?
yes
Done.
PSSM can be saved and used as a query against **another** database (e.g. a single genome)
no
Refine PSSM

© Olga Zhaxybayeva

PSI-BLAST principle

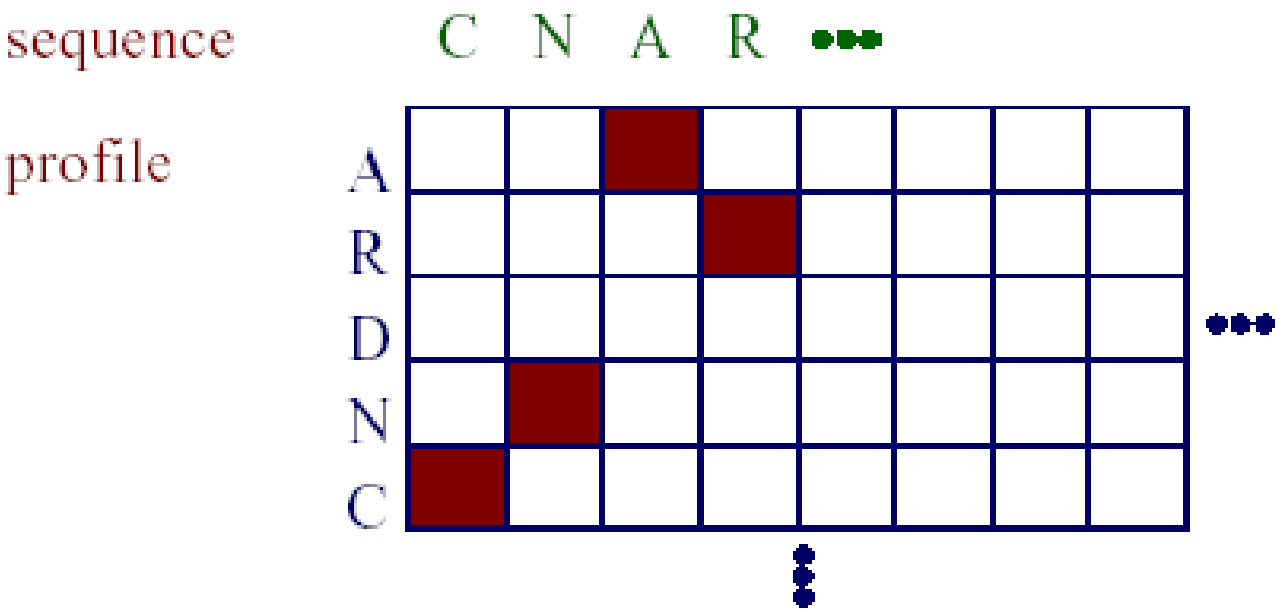
6. Add the new found significant sequences to build a new MSA and PSSM, and steps 2 to 5 can be repeated.
7. Iteration continues until no new significant sequences are found or user decides to stop.


Input **protein** sequence
BLAST search
Create Position Specific Score Matrix (PSSM) or profile from alignment
Do BLAST search with PSSM as a query
Converged? OR enough iterations?
yes
Done.
PSSM can be saved and used as a query against **another** database (e.g. a single genome)
no
Refine PSSM

Position specific iterated BLAST: PSI-BLAST

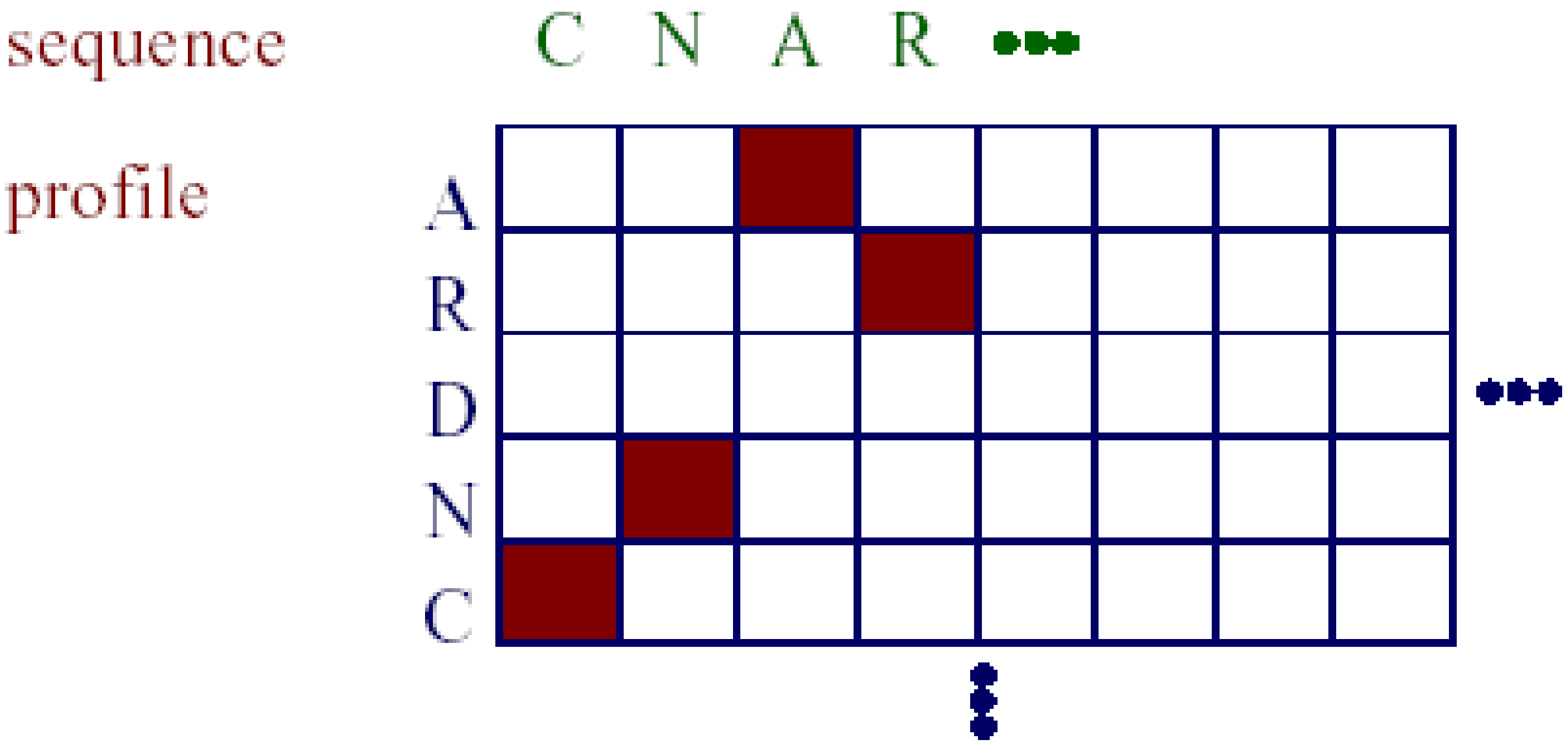
- basic idea
  - use results from BLASTP query to construct a *profile matrix*
  - search database with a profile matrix instead of query sequence
- Iterate
- The iterative profile generation process makes PSI-BLAST far more capable of detecting distant sequence similarities than a single query alone in BLASTP, because it combines the underlying conservation information from a range of related sequence into a single score matrix.

Profile representation of multiple sequences



- Earlier, we were aligning a **sequence against a sequence**
- Can we align a **sequence against a profile** in order to find out whether this sequence belongs to that protein family?

Align profile matrix to a sequence



The score of a profile assigned to a sequence = the sum of the scores for the residues in the sequence at each position in the PSSM

Amino acid /Position	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	-999.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	-999.0	-1.3	-1.3	-2.4	0.0	-2.7	-2.0	0.0	-1.9
E	-999.0	0.1	-1.2	-0.4	0.0	-2.4	-0.6	0.0	-1.9
F	0.0	0.8	0.8	0.1	0.0	-2.1	0.3	0.0	-0.4
G	-999.0	0.5	0.2	-0.7	0.0	-0.3	-1.1	0.0	-0.8
H	-999.0	0.8	0.2	-0.7	0.0	-2.2	0.1	0.0	-1.1
I	-1.0	1.1	1.5	0.5	0.0	-1.9	0.6	0.0	0.7
K	-999.0	1.1	0.0	-2.1	0.0	-2.0	-0.2	0.0	-1.7
L	-1.0	1.0	1.0	0.9	0.0	-2.0	0.3	0.0	0.5
M	-1.0	1.1	1.4	0.8	0.0	-1.8	0.1	0.0	0.1
N	-999.0	0.8	0.5	0.0	0.0	-1.1	0.1	0.0	-1.2
P	-999.0	-0.5	0.3	-1.9	0.0	-0.2	0.1	0.0	-1.1
Q	-999.0	1.2	0.0	0.1	0.0	-1.8	0.2	0.0	-1.6
R	-999.0	2.2	0.7	-2.1	0.0	-1.8	0.1	0.0	-1.0
S	-999.0	-0.3	0.2	-0.7	0.0	-0.6	-0.2	0.0	-0.3
T	-999.0	0.0	0.0	-1.0	0.0	-1.2	0.1	0.0	-0.2
V	-1.0	2.1	0.5	-0.1	0.0	-1.1	0.7	0.0	0.3
W	0.0	-0.1	0.0	-1.8	0.0	-2.4	-0.1	0.0	-1.4
Y	0.0	0.9	0.8	-1.1	0.0	-2.0	0.5	0.0	-0.9

From the above PSSM calculating score of peptide "FSDFCVGHY“:

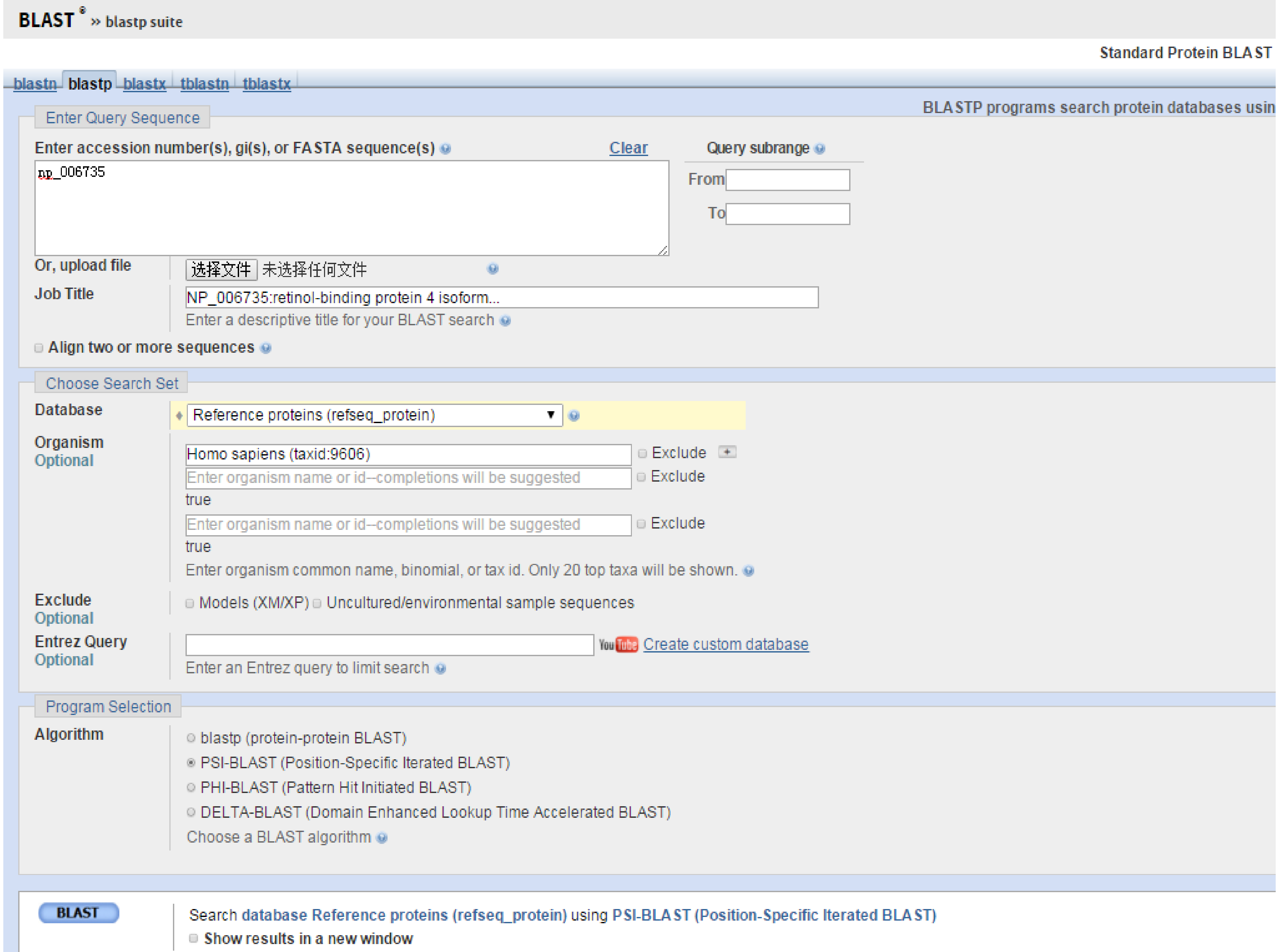
Amino acid /Position	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	-999.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0
D	-999.0	-1.3	**-1.3**	-2.4	0.0	-2.7	-2.0	0.0	-1.9
E	-999.0	0.1	-1.2	-0.4	0.0	-2.4	-0.6	0.0	-1.9
F	**0.0**	0.8	0.8	**0.1**	0.0	-2.1	0.3	0.0	-0.4
G	-999.0	0.5	0.2	-0.7	0.0	-0.3	**-1.1**	0.0	-0.8
H	-999.0	0.8	0.2	-0.7	0.0	-2.2	0.1	**0.0**	-1.1
I	-1.0	1.1	1.5	0.5	0.0	-1.9	0.6	0.0	0.7
K	-999.0	1.1	0.0	-2.1	0.0	-2.0	-0.2	0.0	-1.7
L	-1.0	1.0	1.0	0.9	0.0	-2.0	0.3	0.0	0.5
M	-1.0	1.1	1.4	0.8	0.0	-1.8	0.1	0.0	0.1
N	-999.0	0.8	0.5	0.0	0.0	-1.1	0.1	0.0	-1.2
P	-999.0	-0.5	0.3	-1.9	0.0	-0.2	0.1	0.0	-1.1
Q	-999.0	1.2	0.0	0.1	0.0	-1.8	0.2	0.0	-1.6
R	-999.0	2.2	0.7	-2.1	0.0	-1.8	0.1	0.0	-1.0
S	-999.0	**-0.3**	0.2	-0.7	0.0	-0.6	-0.2	0.0	-0.3
T	-999.0	0.0	0.0	-1.0	0.0	-1.2	0.1	0.0	-0.2
V	-1.0	2.1	0.5	-0.1	0.0	-1.1	0.7	0.0	0.3
W	0.0	-0.1	0.0	-1.8	0.0	-2.4	-0.1	0.0	-1.4
Y	0.0	0.9	0.8	-1.1	0.0	-2.0	0.5	0.0	**-0.9**

From the above PSSM calculating score of peptide "FSDFCVGHY“:

0-0.3-1.3+0.1+0-1.1-1.1+0-0.9=-4.6

Amino acid /Position	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	-999.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	-999.0	-1.3	-1.3	-2.4	0.0	-2.7	-2.0	0.0	-1.9
E	-999.0	0.1	-1.2	-0.4	0.0	-2.4	-0.6	0.0	-1.9
F	0.0	0.8	0.8	0.1	0.0	-2.1	0.3	0.0	-0.4
G	-999.0	0.5	0.2	-0.7	0.0	-0.3	-1.1	0.0	-0.8
H	-999.0	0.8	0.2	-0.7	0.0	-2.2	0.1	0.0	-1.1
I	-1.0	1.1	1.5	0.5	0.0	-1.9	0.6	0.0	0.7
K	-999.0	1.1	0.0	-2.1	0.0	-2.0	-0.2	0.0	-1.7
L	-1.0	1.0	1.0	0.9	0.0	-2.0	0.3	0.0	0.5
M	-1.0	1.1	1.4	0.8	0.0	-1.8	0.1	0.0	0.1
N	-999.0	0.8	0.5	0.0	0.0	-1.1	0.1	0.0	-1.2
P	-999.0	-0.5	0.3	-1.9	0.0	-0.2	0.1	0.0	-1.1
Q	-999.0	1.2	0.0	0.1	0.0	-1.8	0.2	0.0	-1.6
R	-999.0	2.2	0.7	-2.1	0.0	-1.8	0.1	0.0	-1.0
S	-999.0	-0.3	0.2	-0.7	0.0	-0.6	-0.2	0.0	-0.3
T	-999.0	0.0	0.0	-1.0	0.0	-1.2	0.1	0.0	-0.2
V	-1.0	2.1	0.5	-0.1	0.0	-1.1	0.7	0.0	0.3
W	0.0	-0.1	0.0	-1.8	0.0	-2.4	-0.1	0.0	-1.4
Y	0.0	0.9	0.8	-1.1	0.0	-2.0	0.5	0.0	-0.9

From the above PSSM calculating score of peptide “FVPEFSAAM“:



BLAST® » blastp suite

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

BLASTP programs search protein databases usin

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

np_006735

Query subrange

From

To

Or, upload file 选择文件 未选择任何文件

Job Title NP_006735:retinol-binding protein 4 isoform...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Reference proteins (refseq_protein)

Organism Optional: Homo sapiens (taxid:9606) □ Exclude

Enter organism name or id--completions will be suggested □ Exclude

true

Enter organism name or id--completions will be suggested □ Exclude

true

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional: Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

○ blastp (protein-protein BLAST)

● PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database Reference proteins (refseq_protein) using PSI-BLAST (Position-Specific Iterated BLAST)

□ Show results in a new window

Review

Essential Bioinformatics:

- Chapter four