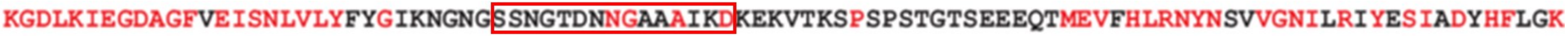
Heuristic approaches

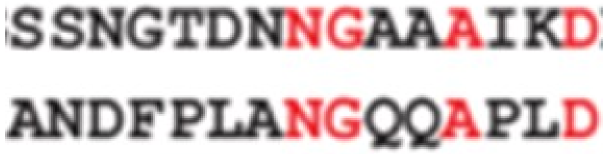
Basic Local Alignment Search Tool (BLAST): some terms

Segment: is a substring of a sequence.



KGDLKIEGDAGFVEISNLVLYFYGIKNGNGSSNGTDNNGAAAIKDKEKVTKSPSPSTGTSEEEQTMEVFHLRNYNSVVGNILRIYESIADYHFLGK

Segment pair: is a pair of segments of equal length from two sequences (gapless alignment).



SSNGTDNNGAAAIKD

ANDFPLANGQQAPLD

Locally maximal segment: is a segment whose alignment score (without gaps) can not be improved by shortening or extending it.

Maximum segment pair (MSP): in two sequences S and T, is a segment with the maximum score over all segment pairs in S and T.

High scoring pairs (HSP): are MSP with score higher than a given cutoff C.

Steps of BLAST Algorithm

Let us say that the query sequence given is: **QLNFSAGW**

Step 1: Find out all the words of length w in the given query sequence.

Let us take w = 2, then we will have the following words: **QL, LN, NF, FS, SA, AG and GW** (Total number of words: L-w+1 where L is the length of the query sequence and w is word size).

Steps of BLAST Algorithm

Step 2: Find out all the words having a score of at least T.

Let us take T=9 and the scores of each word are
QL=9
LN=10
NF=12
FS=10
SA=8
AG=10
GW=17

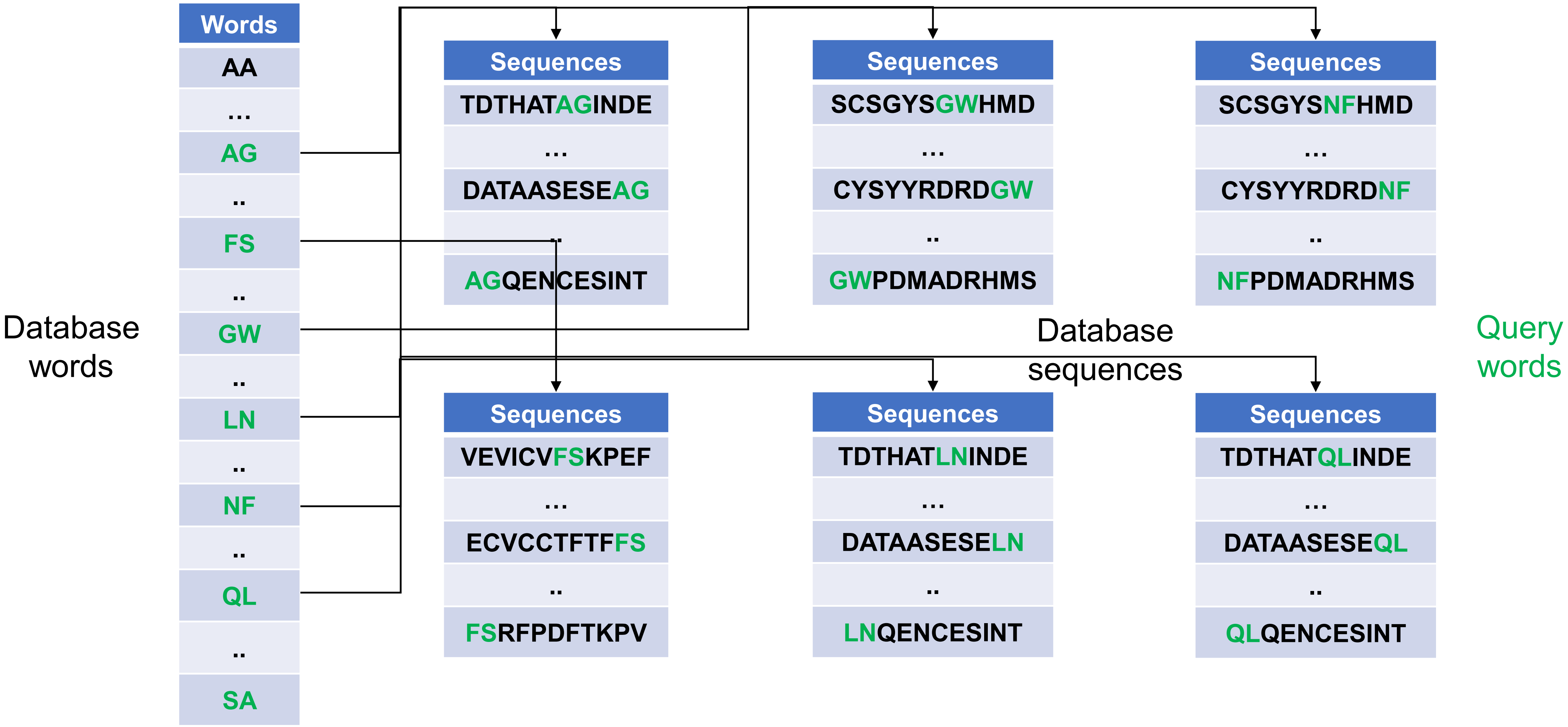
BLOSUM 62 scoring matrix

(positive values are shaded)

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

Steps of BLAST Algorithm

Step 3: Search (scan) the database for all occurrence of query words. To do that, index database sequences into table of words (pre-compute this). Then, index query words into the table (at the query time).



Database words

Words
AA
…
AG
..
FS
..
GW
..
LN
..
NF
..
QL
..
SA

Sequences
TDTHATAGINDE
…
DATAASESEAG
..
AGQENCESINT

Sequences
SCSGYSGWHMD
…
CYSYYRDRDGW
..
GWPDMADRHMS

Sequences
SCSGYSNFHMD
…
CYSYYRDRDNF
..
NFPDMADRHMS

Database sequences

Query words

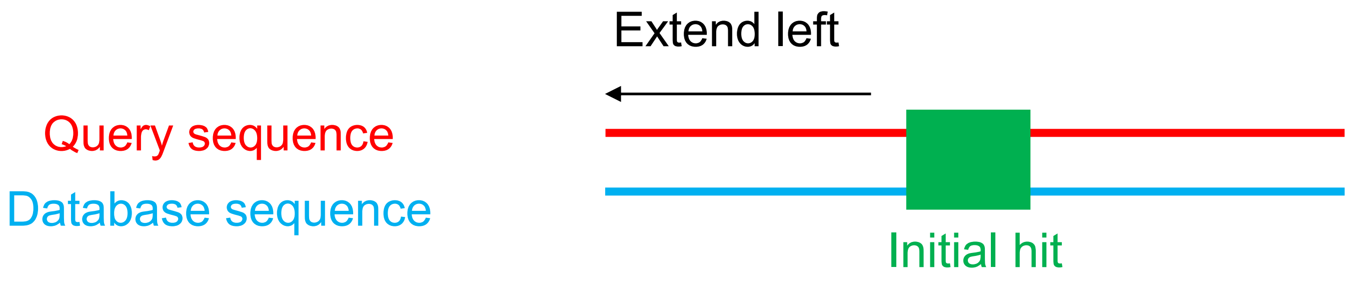
Sequences
VEVICVFSKPEF
…
ECVCCTFTFFS
..
FSRFPDFTKPV

Sequences
TDTHATLNINDE
…
DATAASESELN
..
LNQENCESINT

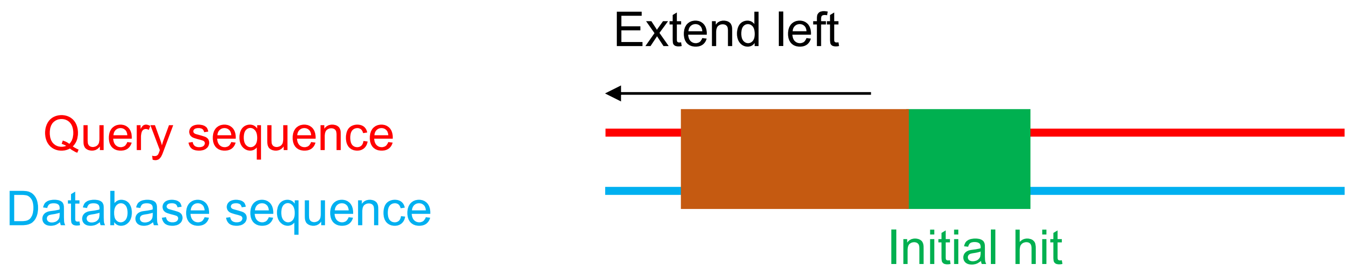
Sequences
TDTHATQLINDE
…
DATAASESEQL
..
QLQENCESINT

Steps of BLAST Algorithm

Step 4: Extend the hit: extend until score starts decreasing (gapless).

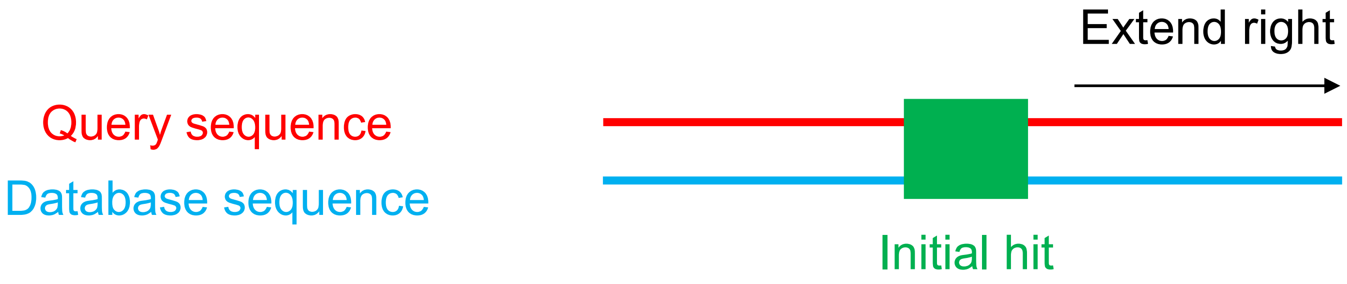


Return high scoring segment pair: return a segment pair having at least a score of S (let us say), a score cut-off.

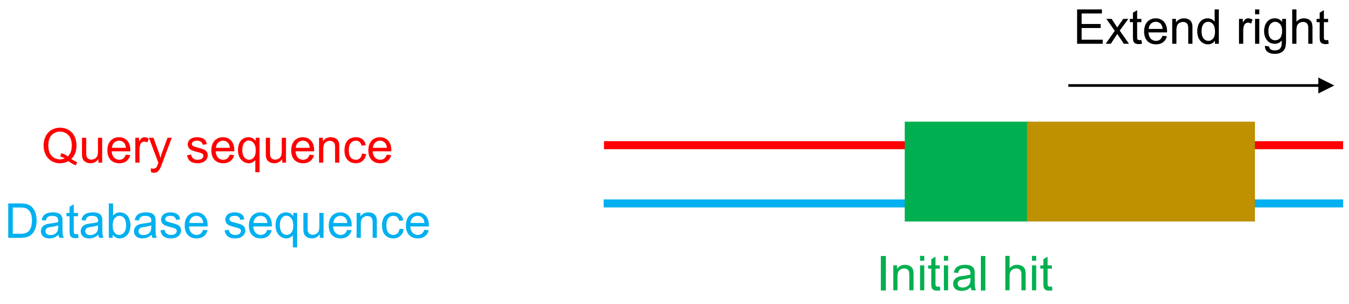


Steps of BLAST Algorithm

Step 5: Extend the hit: repeat the step 4 in the right direction.



Return high scoring segment pair: return a segment pair having at least a score of S (let us say), a score cut-off.

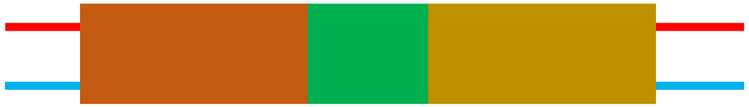


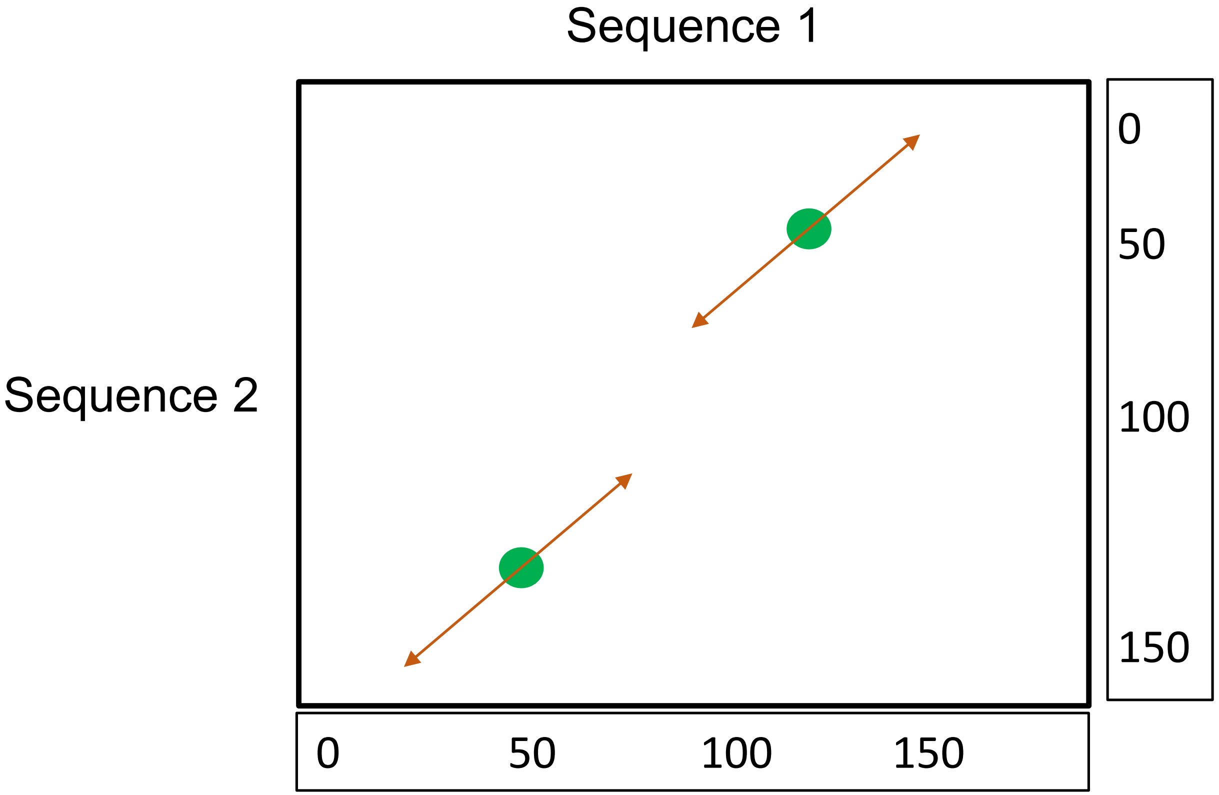
Steps of BLAST Algorithm

Step 6: Join the extension: this gives a gapless sequence alignment.

Query sequence

Database sequence

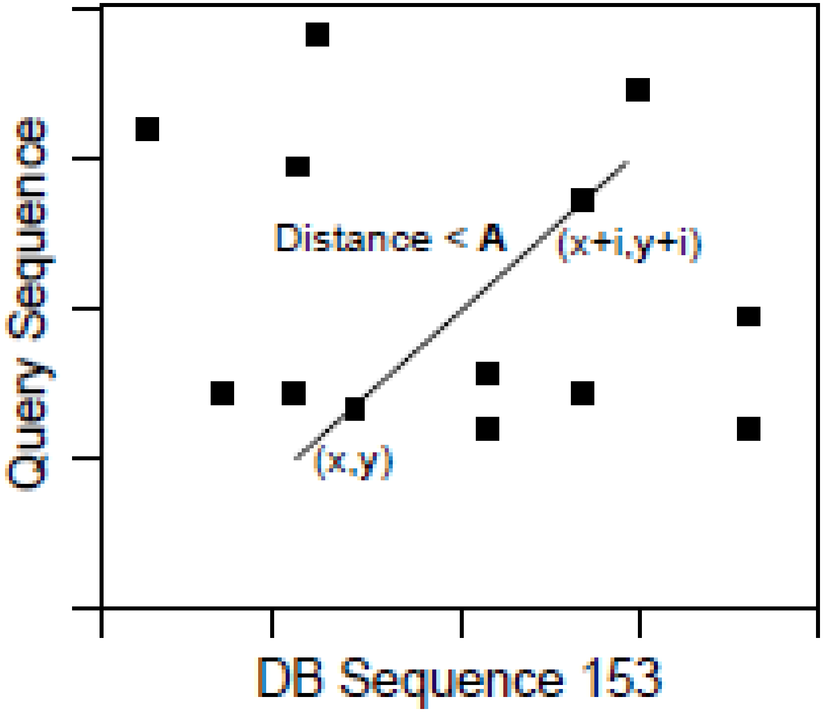




Gapped BLAST

Gapped extension

- Extensions can be triggered only by two or more hits on the same diagonal.
- Hits must also be less than a distance ***D*** (let us say) from each other to trigger extension.
- Typically the Dynamic Programming sequence alignment method is applied to find gapped alignment.



Advantages

- Serves to reduce the number of extensions.
- Gapped BLAST is more sensitive and selective than the original, ungapped BLAST.

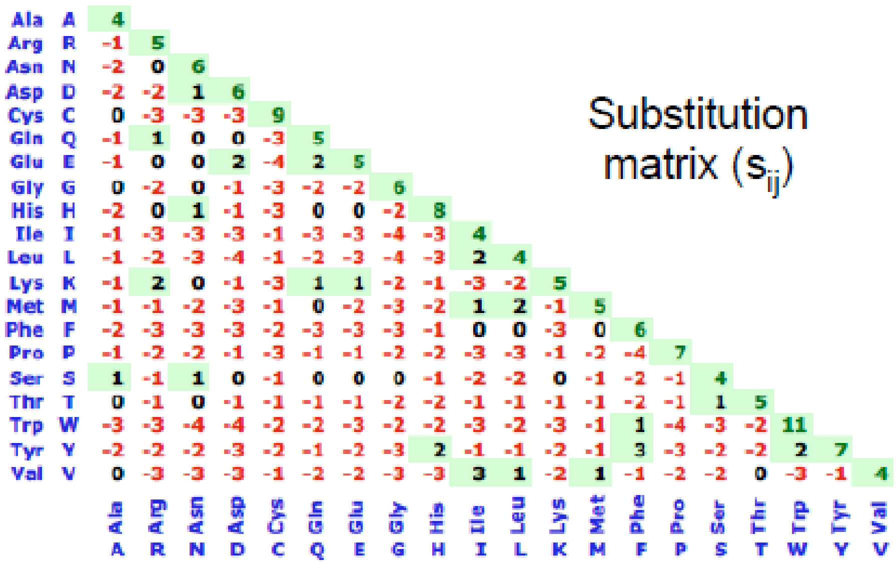
BLAST: some concepts

The central idea of the BLAST algorithm is that a statistically significant alignment is likely to contain a high-scoring pair of aligned words.

Score: A number used to assess the biological relevance of a finding.

$$\boxed{S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}}$$

```
R   L   A   S   V   -   E   T   D   M   W   T   P   L   T   L   R   Q   H
.   |   .   |   :       :   |   .   :           .   |   .   .   |
T   L   T   S   L   A   Q   T   T   L   -   -   K   A   H   L   G   T   H
-1  +4  +0  +4  +1  -4  +2  +5  -1  +2  -4  -1  -1  -1  -2  +4  -2  -1  +8  =  12
```



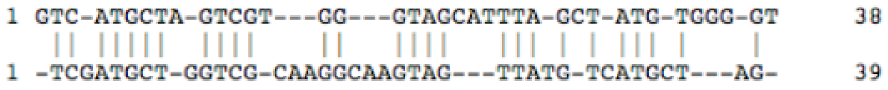
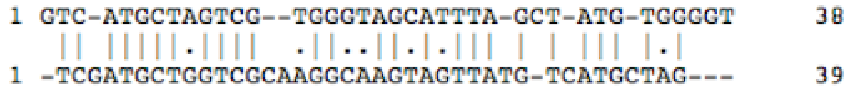
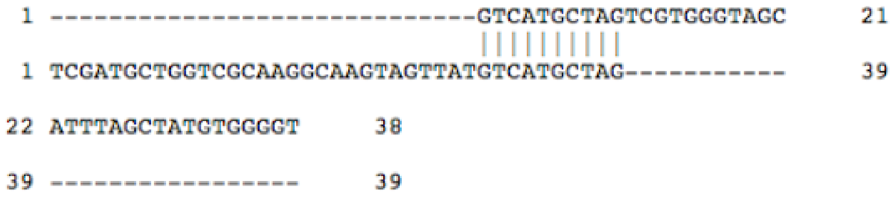
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Substitution matrix (s_{ij})

Gap penalty

Gap opening = -4
Gap extension = -1
End gap = 0

BLAST: some concepts

Gap penalty	Alignment	Identity / Similarity	Gaps	Score
0		27/50 (54.0%)	23/50	S=135
5		26/44 (59.1%)	11/44	S=67
10		10/67 (14.9%)	57/67	S=50

```
1 GTC-ATGCTA-GTCGT---GG---GTAGCATTTA-GCT-ATG-TGGG-GT     38
   || |||||  ||||    ||   ||||   ||| | | ||| |    |
1 -TCGATGCT-GGTCG-CAAGGCAAGTAG---TTATG-TCATGCT---AG-     39
```

```
1 GTC-ATGCTAGTCG--TGGGTAGCATTTA-GCT-ATG-TGGGGT     38
   || |||||.||||  .||..||.|.||| | | ||| |.|
1 -TCGATGCTGGTCGCAAGGCAAGTAGTTATG-TCATGCTAG---     39
```

```
 1 -----------------------------GTCATGCTAGTCGTGGGTAGC     21
                                ||||||||||
 1 TCGATGCTGGTCGCAAGGCAAGTAGTTATGTCATGCTAG-----------     39

22 ATTTAGCTATGTGGGGT     38

39 -----------------     39
```

Remark: The scores of these different alignments can not be compared (neither used to select the best alignment) because their scale depends on the gap penalty.

Assessing the significance of sequence alignments

To facilitate calculations, a sequence alignment score S may also be normalized to produce a score S’ (also known as bit score):

$$S' = \lambda S - \ln Kmn$$

The **bit-score (S’)** is a normalized score expressed in *bits* that lets you estimate the magnitude of the *search space* you would have to look through before you would expect to find an score as good as or better than this one by chance.

P-value: Probability that an event occurs by chance. In the context of sequence alignments, the *P-value* associated to a score S is the probability to obtain by chance a score **x** at least equal to S.

$$P(S) = P(x \geq S) = Ke^{-\lambda S} = Ke^{-\left(S' \ln(2) + \ln(K)\right)} = 2^{-S'}$$

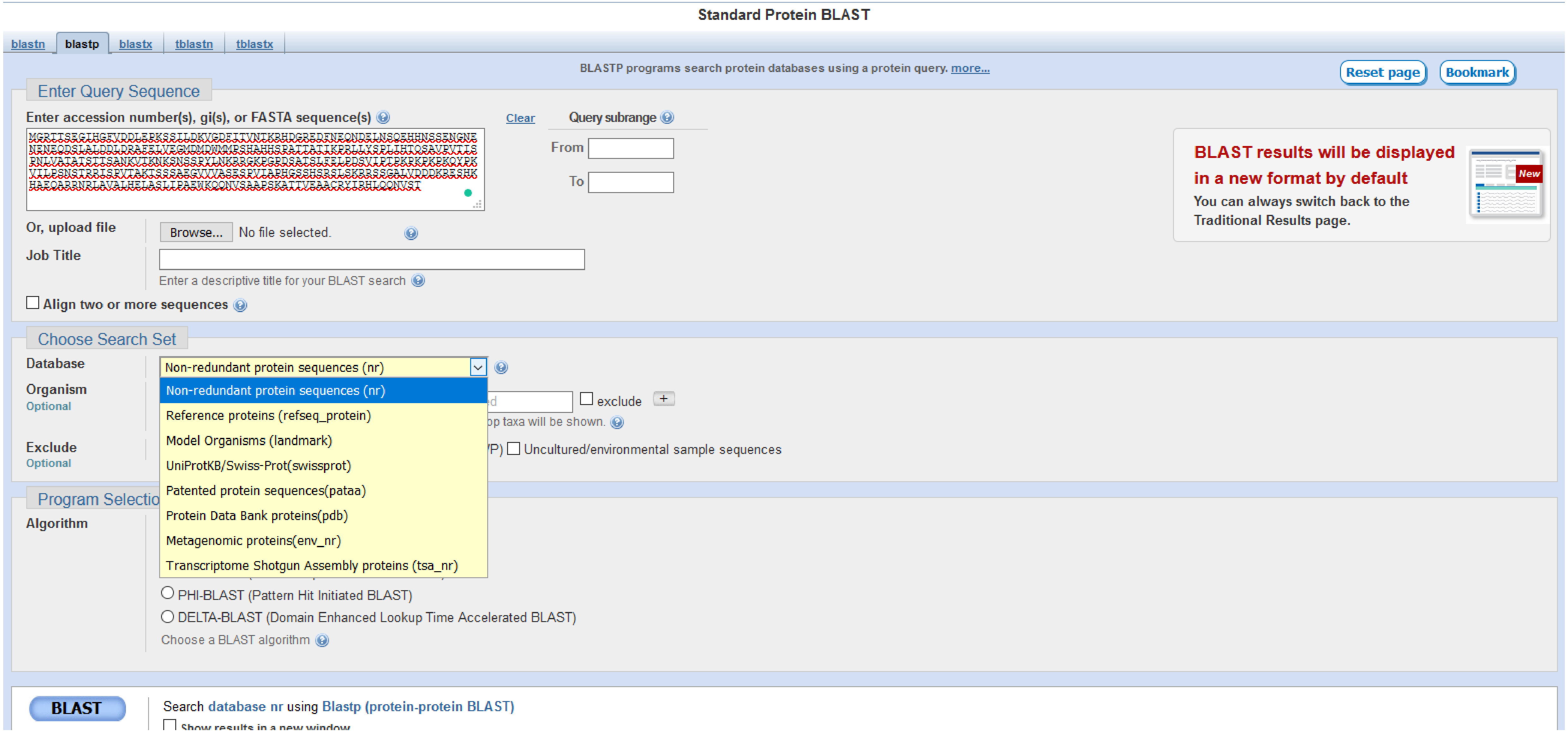
Assessing the significance of sequence alignments

E-value: Correction of the *P-value* for multiple testing. In the context of database searches, the E-value (associated to a score S) is the number of distinct alignments, with a score equivalent to or better than S, that are expected to occur in a database search by chance. The lower the E-value, the more significant the score is.

$$
\begin{aligned}
E &= mn \cdot Pval \\
&= Kmne^{-\lambda S} \\
&= NKe^{-\lambda S} \\
&= N/2^{S'}
\end{aligned}
$$

N = size of the search space (n x m).

BLAST Run Example



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)

```
MGRTTSEGIHGFVDDLEPKSSILDKVGDFITVNTKRHDGREDFNEQNDELNSQEHHNSSENGNE
NENEQDSLALDDLDRAFELVEGMDMDWMMPSHAHHSPATTATIKPRLLYSPLIHTQSAVPVTIS
PNLVATATSTTSANKVTKNKSNSSPYLNKRRGKPGPDSATSLFELPDSVIPTPKPKPKPKQYPK
VILPSNSTRRISPVTAKTSSSAEGVVVASESPVIAPHGSSHSRSLSKRRSSGALVDDDKRESHK
HAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATTVEAACRYIRHLQQNVST
```

Query subrange

From

To

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Or, upload file No file selected.

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Model Organisms (landmark)
- UniProtKB/Swiss-Prot(swissprot)
- Patented protein sequences(pataa)
- Protein Data Bank proteins(pdb)
- Metagenomic proteins(env_nr)
- Transcriptome Shotgun Assembly proteins (tsa_nr)

Organism Optional

d □ exclude +

p taxa will be shown.

Exclude Optional

P) □ Uncultured/environmental sample sequences

Program Selectio

Algorithm

○ PHI-BLAST (Pattern Hit Initiated BLAST)

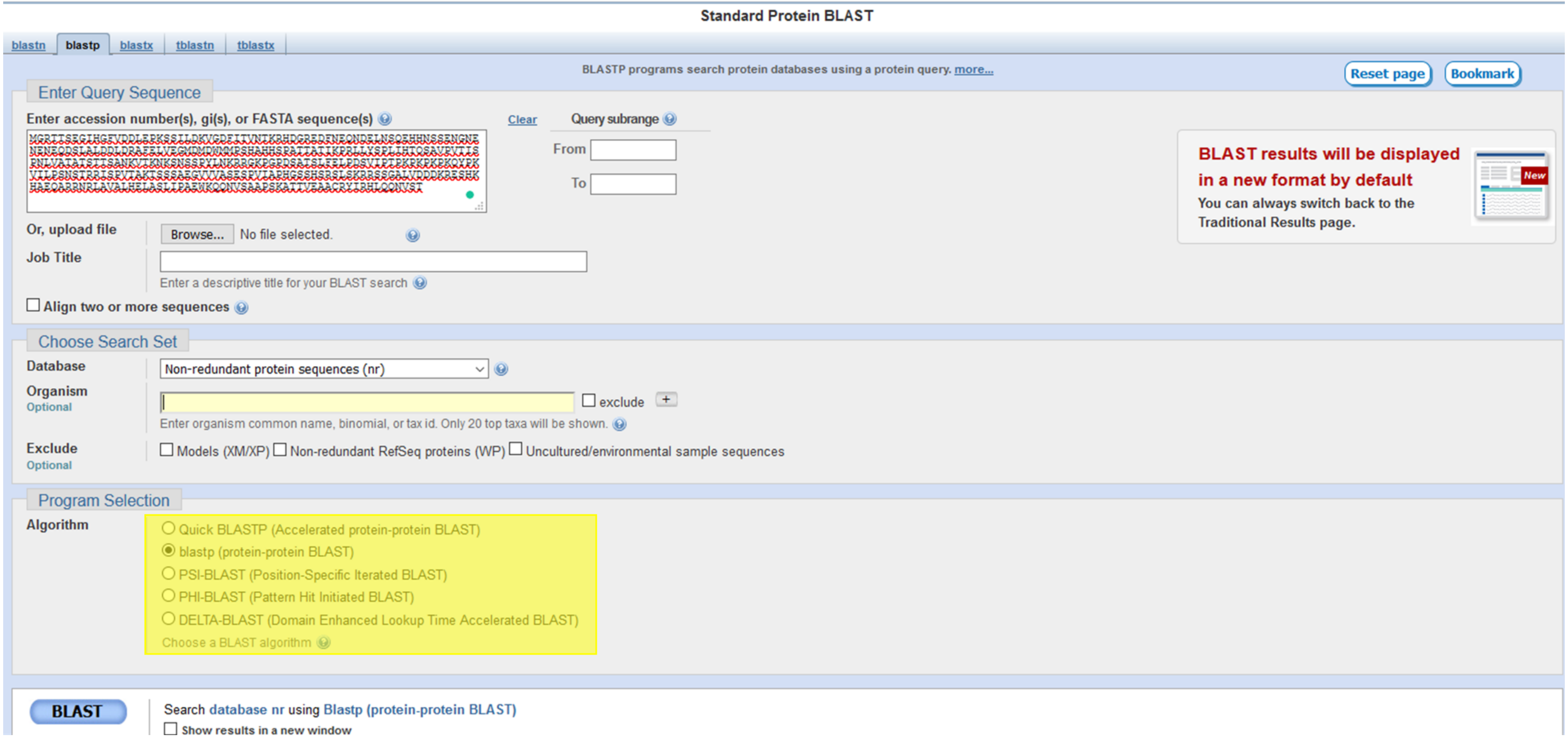
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

BLAST Run Example



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

```
MGRTTSEGIHGFVDDLEPKSSILDKVGDFITVNTKRHDGREDFNEQNDELNSQEHHNSSENGNE
NENEQDSLALDDLDRAFELVEGMDMDWMMPSHAHHSPATTATIKPRLLYSPLIHTQSAVPVTIS
PNLVATATSTTSANKVTKNKSNSSPYLNKRRGKPGPDSATSLFELPDSVIPTPKPKPKPKQYPK
VILPSNSTRRISPVTAKTSSSAEGVVVASESPVIAPHGSSHSRSLSKRRSSGALVDDDKRESHK
HAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATTVEAACRYIRHLQQNVST
```

Query subrange

From

To

Or, upload file Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional

□ exclude +

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

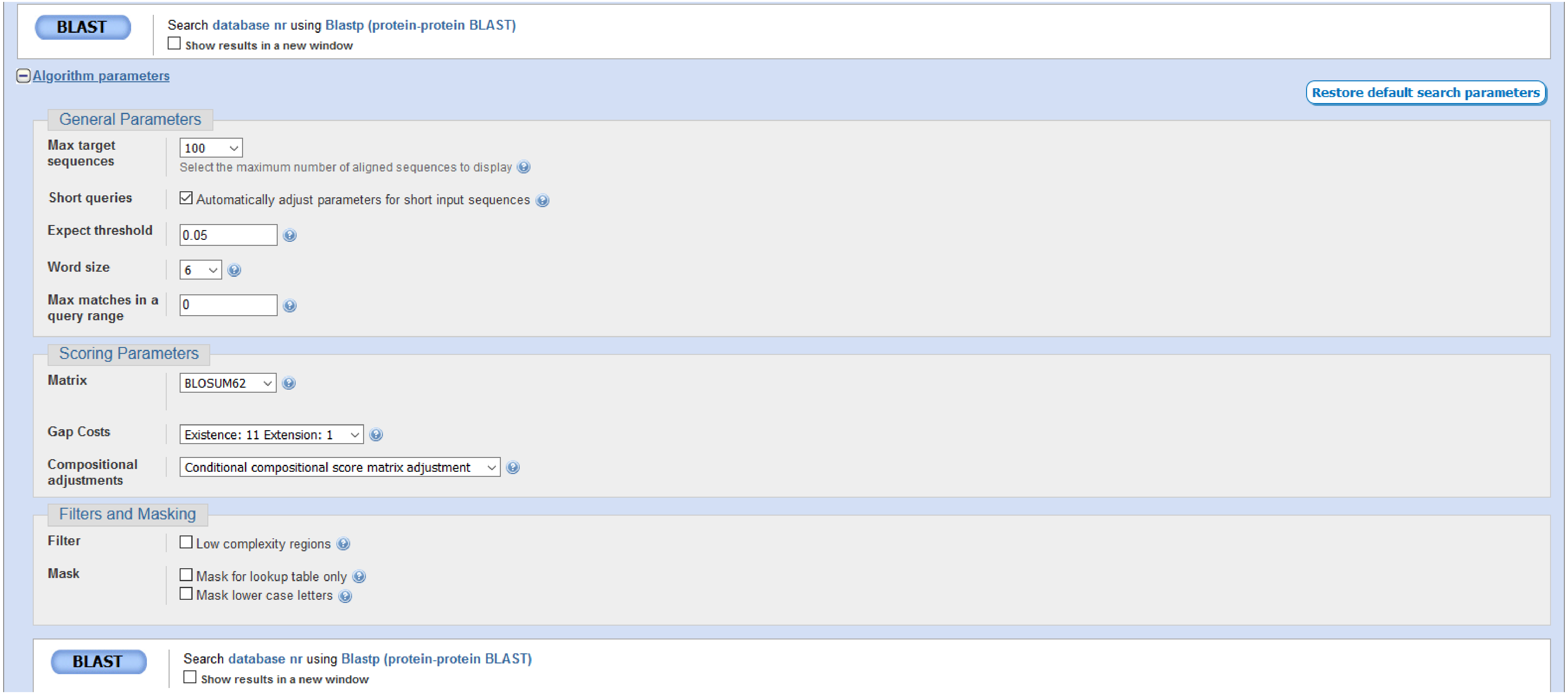
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

BLAST Run Example



Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

⊟ 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: 0.05

Word size: 6

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

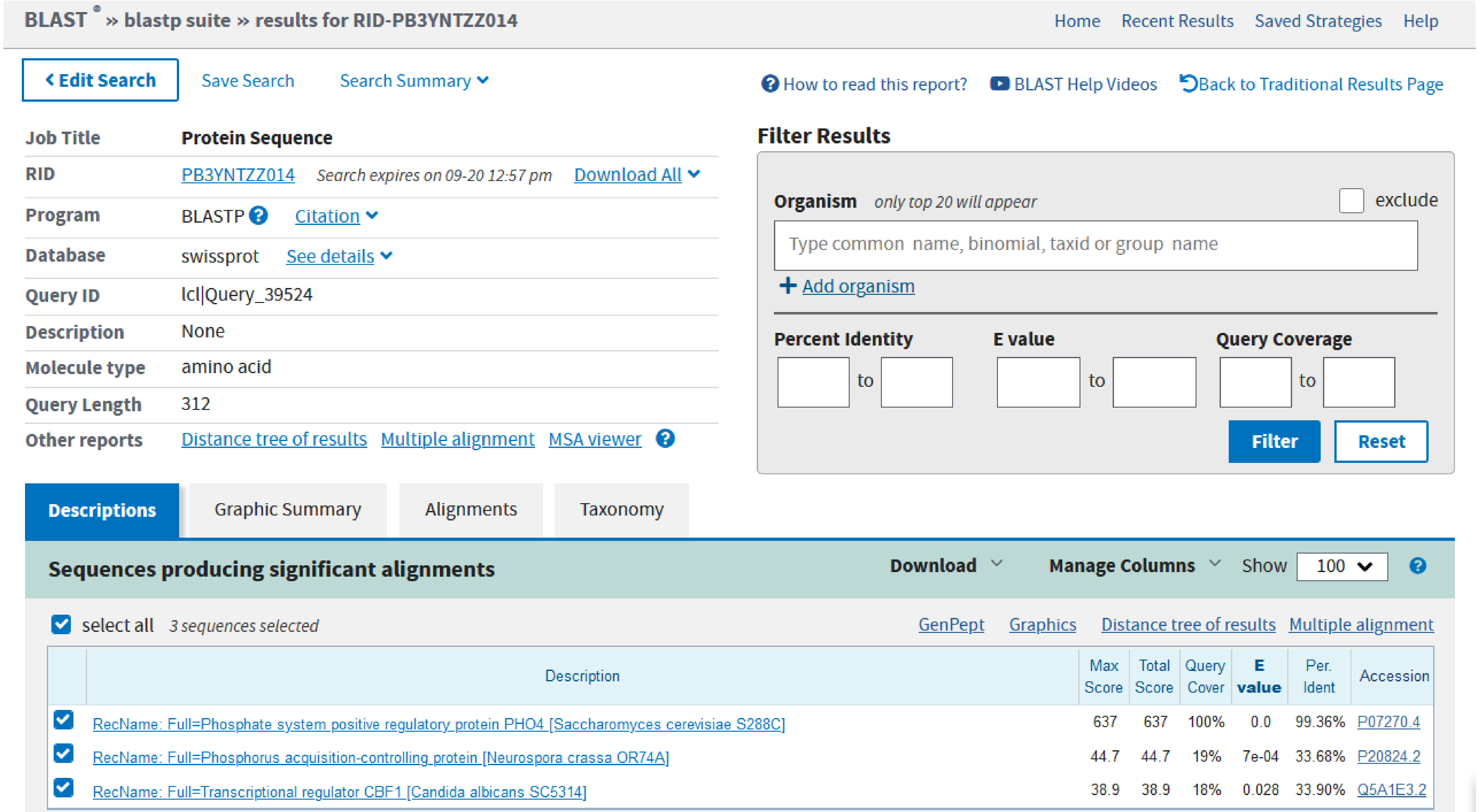
Mask: □ Mask for lookup table only

□ Mask lower case letters

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

BLAST Run Example



Job Title	Protein Sequence
RID	PB3YNTZZ014 *Search expires on 09-20 12:57 pm* Download All ⌄
Program	BLASTP Citation ⌄
Database	swissprot See details ⌄
Query ID	lcl|Query_39524
Description	None
Molecule type	amino acid
Query Length	312
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity ___ to ___ **E value** ___ to ___ **Query Coverage** ___ to ___

Filter Reset

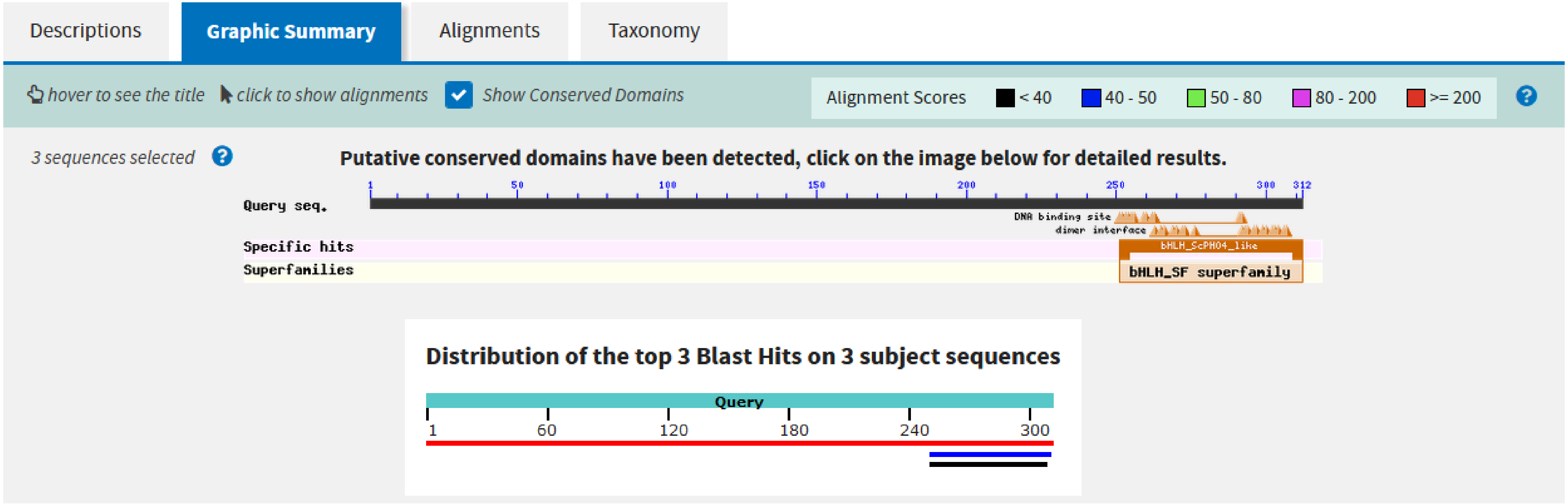
Sequences producing significant alignments

Show 100

☑ select all *3 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	RecName: Full=Phosphate system positive regulatory protein PHO4 [Saccharomyces cerevisiae S288C]	637	637	100%	0.0	99.36%	P07270.4
☑	RecName: Full=Phosphorus acquisition-controlling protein [Neurospora crassa OR74A]	44.7	44.7	19%	7e-04	33.68%	P20824.2
☑	RecName: Full=Transcriptional regulator CBF1 [Candida albicans SC5314]	38.9	38.9	18%	0.028	33.90%	Q5A1E3.2

BLAST Run Example



☑ Show Conserved Domains

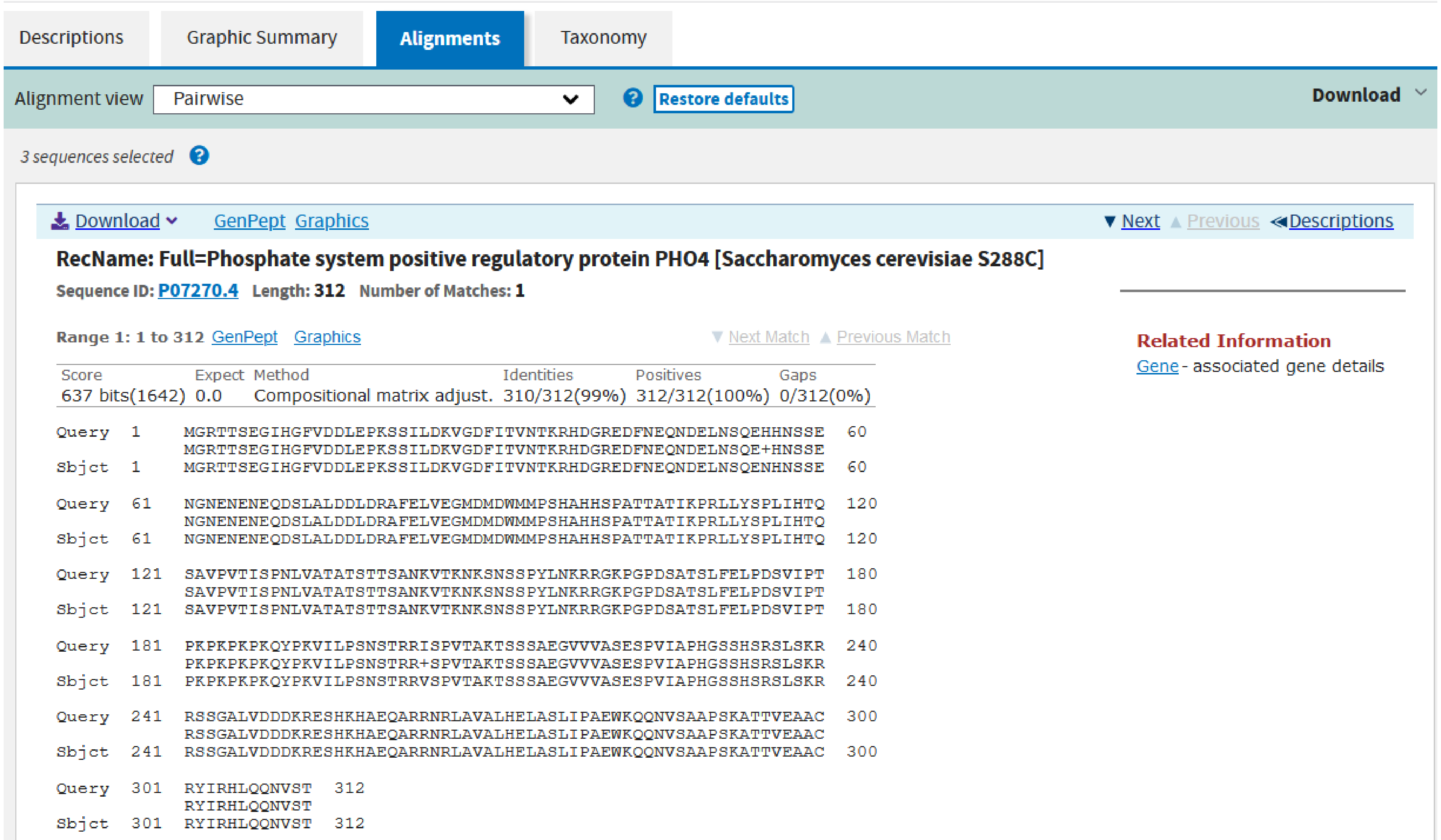
Alignment Scores ■ < 40 🟦 40 - 50 🟩 50 - 80 🟪 80 - 200 🟥 >= 200

3 sequences selected

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 3 Blast Hits on 3 subject sequences

BLAST Run Example



Alignment view: Pairwise

3 sequences selected

RecName: Full=Phosphate system positive regulatory protein PHO4 [Saccharomyces cerevisiae S288C]

Sequence ID: P07270.4 **Length:** 312 **Number of Matches:** 1

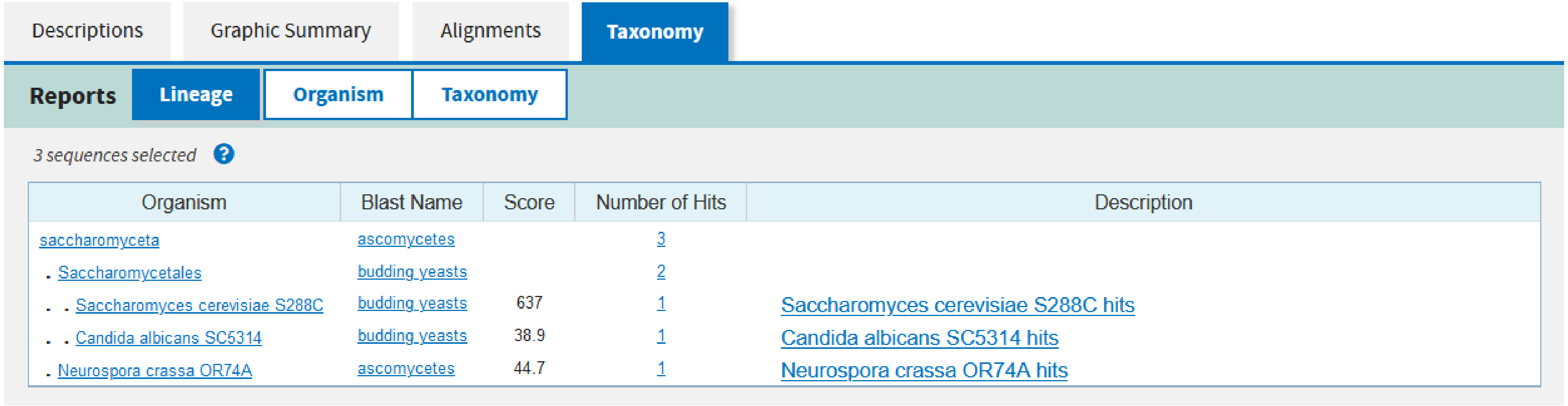
Range 1: 1 to 312 GenPept Graphics

Related Information

Gene - associated gene details

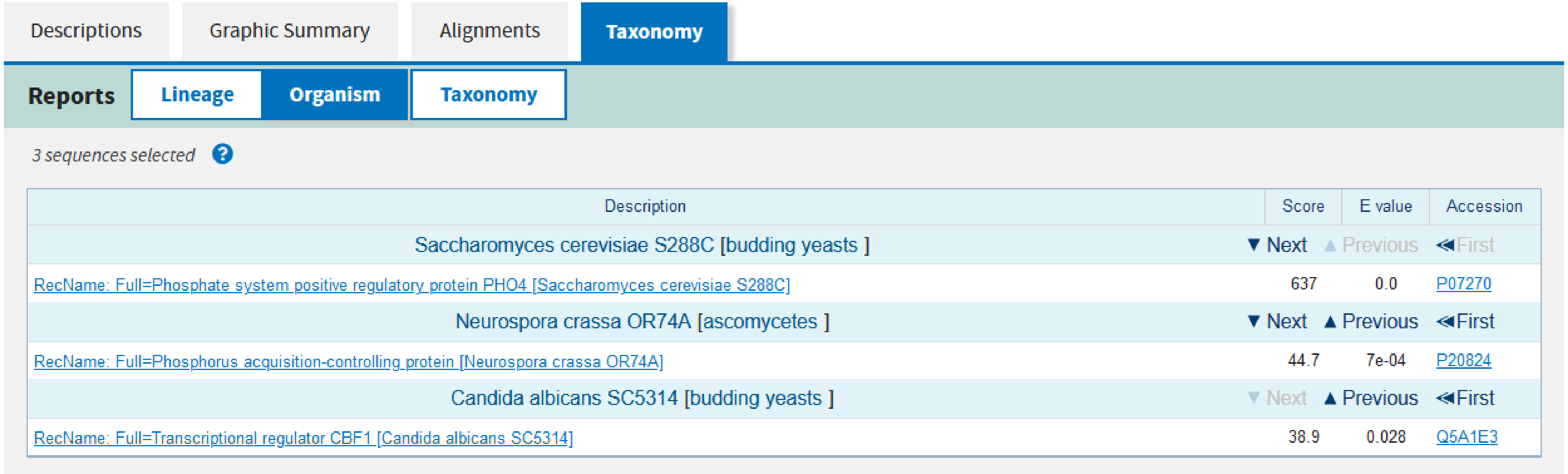
```
Score            Expect  Method                         Identities     Positives       Gaps
637 bits(1642)   0.0     Compositional matrix adjust.   310/312(99%)   312/312(100%)   0/312(0%)

Query  1    MGRTTSEGIHGFVDDLEPKSSILDKVGDFITVNTKRHDGREDFNEQNDELNSQEHHNSSE  60
            MGRTTSEGIHGFVDDLEPKSSILDKVGDFITVNTKRHDGREDFNEQNDELNSQE+HNSSE
Sbjct  1    MGRTTSEGIHGFVDDLEPKSSILDKVGDFITVNTKRHDGREDFNEQNDELNSQENHNSSE  60

Query  61   NGNENENEQDSLALDDLDRAFELVEGMDMDWMMPSHAHHSPATTATIKPRLLYSPLIHTQ  120
            NGNENENEQDSLALDDLDRAFELVEGMDMDWMMPSHAHHSPATTATIKPRLLYSPLIHTQ
Sbjct  61   NGNENENEQDSLALDDLDRAFELVEGMDMDWMMPSHAHHSPATTATIKPRLLYSPLIHTQ  120

Query  121  SAVPVTISPNLVATATSTTSANKVTKNKSNSSPYLNKRRGKPGPDSATSLFELPDSVIPT  180
            SAVPVTISPNLVATATSTTSANKVTKNKSNSSPYLNKRRGKPGPDSATSLFELPDSVIPT
Sbjct  121  SAVPVTISPNLVATATSTTSANKVTKNKSNSSPYLNKRRGKPGPDSATSLFELPDSVIPT  180

Query  181  PKPKPKPKQYPKVILPSNSTRRISPVTAKTSSSAEGVVVASESPVIAPHGSSHSRSLSKR  240
            PKPKPKPKQYPKVILPSNSTRR+SPVTAKTSSSAEGVVVASESPVIAPHGSSHSRSLSKR
Sbjct  181  PKPKPKPKQYPKVILPSNSTRRVSPVTAKTSSSAEGVVVASESPVIAPHGSSHSRSLSKR  240

Query  241  RSSGALVDDDKRESHKHAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATTVEAAC  300
            RSSGALVDDDKRESHKHAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATTVEAAC
Sbjct  241  RSSGALVDDDKRESHKHAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATTVEAAC  300

Query  301  RYIRHLQQNVST  312
            RYIRHLQQNVST
Sbjct  301  RYIRHLQQNVST  312
```

BLAST Run Example



3 sequences selected ?

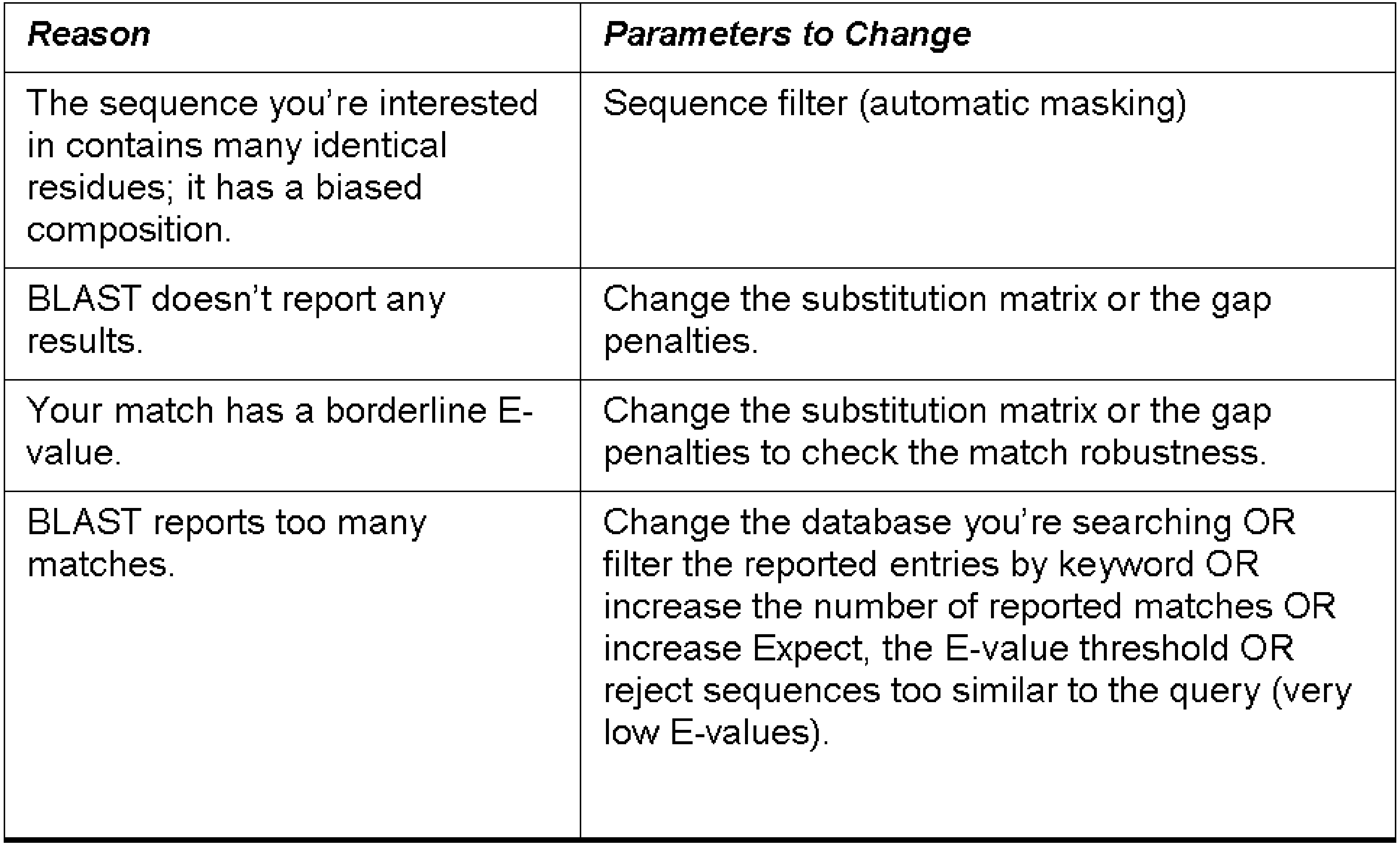
Organism	Blast Name	Score	Number of Hits	Description
saccharomyceta	ascomycetes		3	
. Saccharomycetales	budding yeasts		2	
. . Saccharomyces cerevisiae S288C	budding yeasts	637	1	Saccharomyces cerevisiae S288C hits
. . Candida albicans SC5314	budding yeasts	38.9	1	Candida albicans SC5314 hits
. Neurospora crassa OR74A	ascomycetes	44.7	1	Neurospora crassa OR74A hits



3 sequences selected ?

Description	Score	E value	Accession
Saccharomyces cerevisiae S288C [budding yeasts]	▼ Next	▲ Previous	◀◀First
RecName: Full=Phosphate system positive regulatory protein PHO4 [Saccharomyces cerevisiae S288C]	637	0.0	P07270
Neurospora crassa OR74A [ascomycetes]	▼ Next	▲ Previous	◀◀First
RecName: Full=Phosphorus acquisition-controlling protein [Neurospora crassa OR74A]	44.7	7e-04	P20824
Candida albicans SC5314 [budding yeasts]	▼ Next	▲ Previous	◀◀First
RecName: Full=Transcriptional regulator CBF1 [Candida albicans SC5314]	38.9	0.028	Q5A1E3

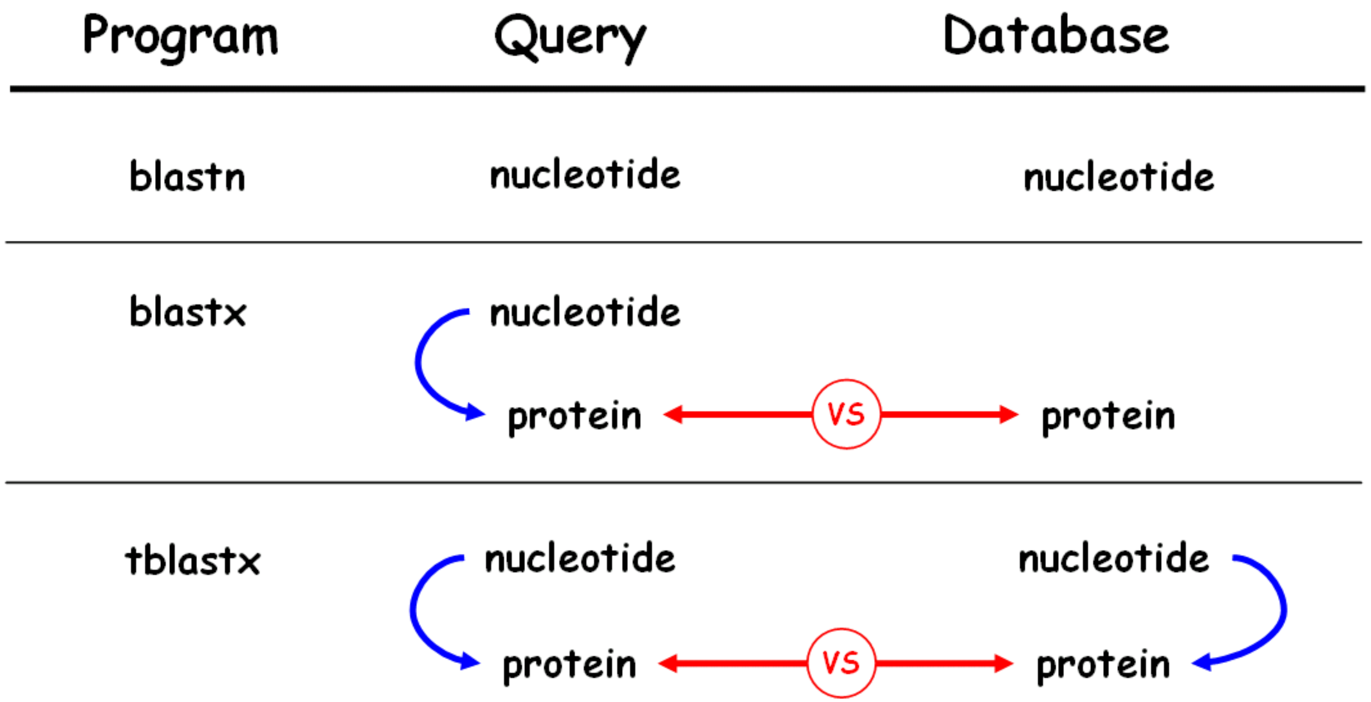
Some Reasons for Changing the Default Parameters



Reason	*Parameters to Change*
The sequence you're interested in contains many identical residues; it has a biased composition.	Sequence filter (automatic masking)
BLAST doesn't report any results.	Change the substitution matrix or the gap penalties.
Your match has a borderline E-value.	Change the substitution matrix or the gap penalties to check the match robustness.
BLAST reports too many matches.	Change the database you're searching OR filter the reported entries by keyword OR increase the number of reported matches OR increase Expect, the E-value threshold OR reject sequences too similar to the query (very low E-values).

BLAST Programs

Basic BLAST	Nucleotide BLAST	blastn, megablast, discontiguous megablast	Search a nucleotide database using a nucleotide query
	Protein BLAST	blastp, psi-blast, phi-blast	Search protein database using a protein query
		blastx	Search protein database using a translated nucleotide query
		tblastn	Search translated nucleotide database using a protein query
		tblastx	Search translated nucleotide database using a translated nucleotide query



Program Query Database

blastn nucleotide nucleotide

blastx nucleotide

protein vs protein

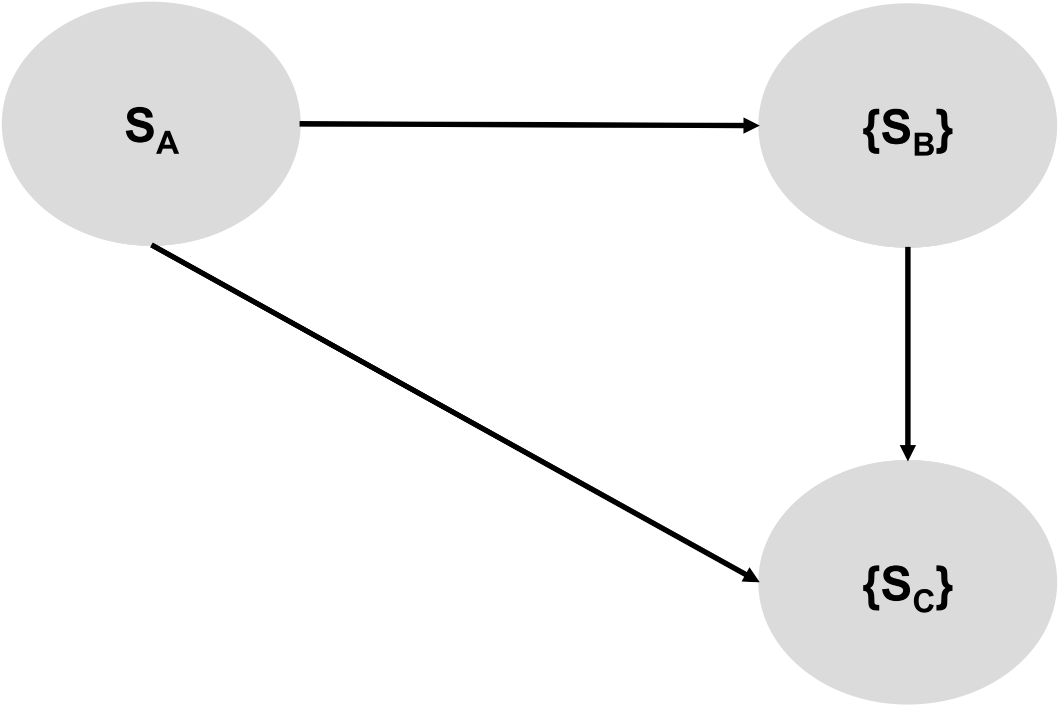
tblastx nucleotide nucleotide

protein vs protein

PSI-BLAST

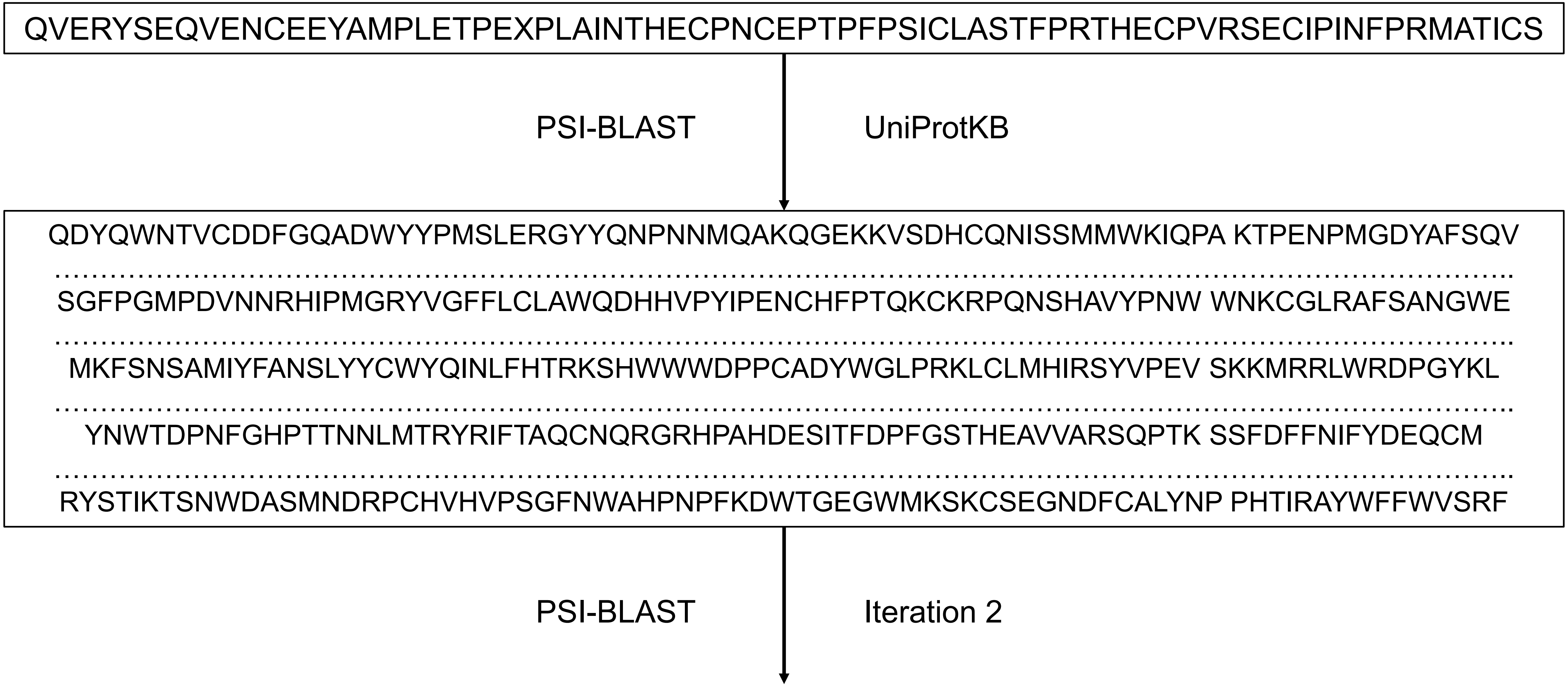
- PSI-BLAST: Position Specific Iterative BLAST

- Used for distant (or remote) homology detection



PSI-BLAST

- PSI-BLAST: Working methodology



QVERYSEQVENCEEYAMPLETPEXPLAINTHECPNCEPTPFPSICLASTFPRTHECPVRSECIPINFPRMATICS

PSI-BLAST

UniProtKB

QDYQWNTVCDDFGQADWYYPMSLERGYYQNPNNMQAKQGEKKVSDHCQNISSMMWKIQPA KTPENPMGDYAFSQV

…………………………………………………………………………………………………..

SGFPGMPDVNNRHIPMGRYVGFFLCLAWQDHHVPYIPENCHFPTQKCKRPQNSHAVYPNW WNKCGLRAFSANGWE

…………………………………………………………………………………………………..

MKFSNSAMIYFANSLYYCWYQINLFHTRKSHWWWDPPCADYWGLPRKLCLMHIRSYVPEV SKKMRRLWRDPGYKL

…………………………………………………………………………………………………..

YNWTDPNFGHPTTNNLMTRYRIFTAQCNQRGRHPAHDESITFDPFGSTHEAVVARSQPTK SSFDFFNIFYDEQCM

…………………………………………………………………………………………………..

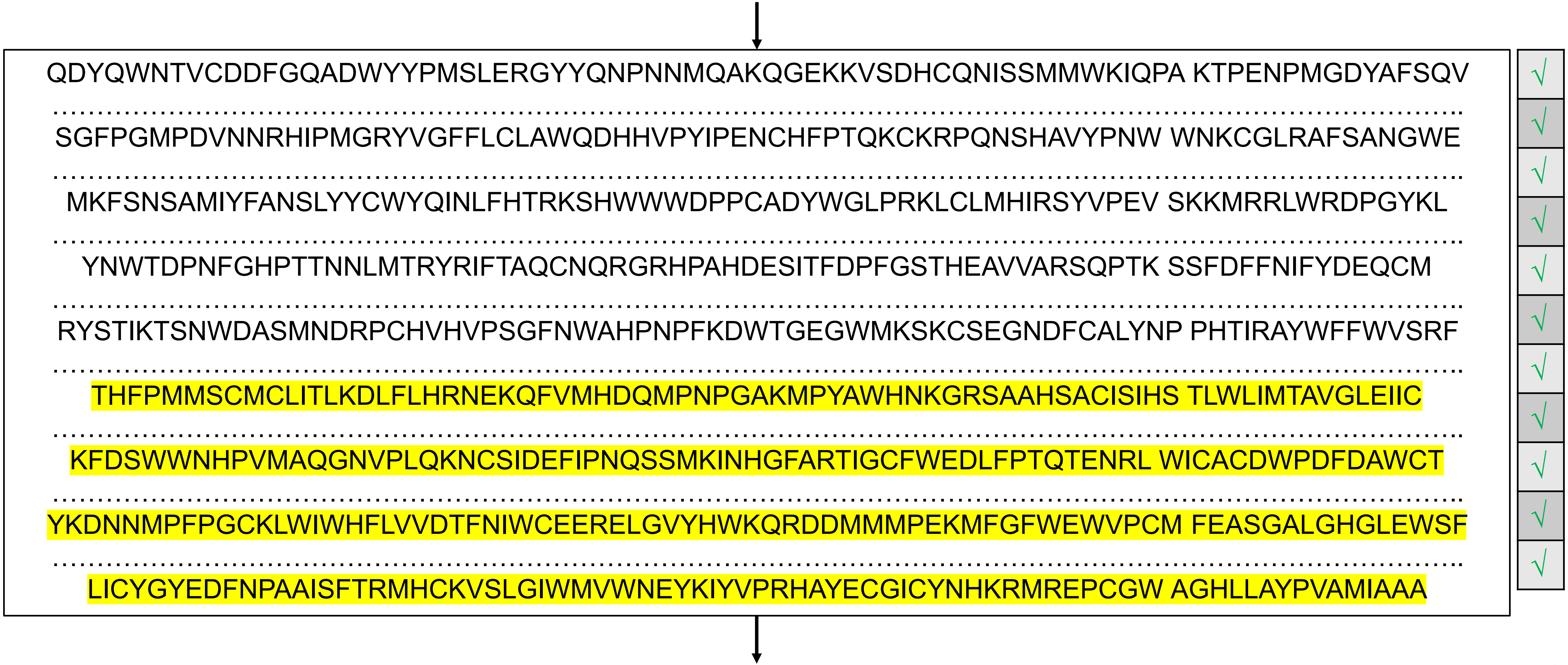
RYSTIKTSNWDASMNDRPCHVHVPSGFNWAHPNPFKDWTGEGWMKSKCSEGNDFCALYNP PHTIRAYWFFWVSRF

PSI-BLAST

Iteration 2

PSI-BLAST

- PSI-BLAST: Working methodology



QDYQWNTVCDDFGQADWYYPMSLERGYYQNPNNMQAKQGEKKVSDHCQNISSMMWKIQPA KTPENPMGDYAFSQV

....

SGFPGMPDVNNRHIPMGRYVGFFLCLAWQDHHVPYIPENCHFPTQKCKRPQNSHAVYPNW WNKCGLRAFSANGWE

....

MKFSNSAMIYFANSLYYCWYQINLFHTRKSHWWWDPPCADYWGLPRKLCLMHIRSYVPEV SKKMRRLWRDPGYKL

....

YNWTDPNFGHPTTNNLMTRYRIFTAQCNQRGRHPAHDESITFDPFGSTHEAVVARSQPTK SSFDFFNIFYDEQCM

....

RYSTIKTSNWDASMNDRPCHVHVPSGFNWAHPNPFKDWTGEGWMKSKCSEGNDFCALYNP PHTIRAYWFFWVSRF

....

THFPMMSCMCLITLKDLFLHRNEKQFVMHDQMPNPGAKMPYAWHNKGRSAAHSACISIHS TLWLIMTAVGLEIIC

....

KFDSWWNHPVMAQGNVPLQKNCSIDEFIPNQSSMKINHGFARTIGCFWEDLFPTQTENRL WICACDWPDFDAWCT

....

YKDNNMPFPGCKLWIWHFLVVDTFNIWCEERELGVYHWKQRDDMMMPEKMFGFWEWVPCM FEASGALGHGLEWSF

....

LICYGYEDFNPAAISFTRMHCKVSLGIWMVWNEYKIYVPRHAYECGICYNHKRMREPCGW AGHLLAYPVAMIAAA

PSI-BLAST

- PSI-BLAST: Working methodology

QDYQWNTVCDDFGQADWYYPMSLERGYYQNPNNMQAKQGEKKVSDHCQNISSMMWKIQPA KTPENPMGDYAFSQV

....

SGFPGMPDVNNRHIPMGRYVGFFLCLAWQDHHVPYIPENCHFPTQKCKRPQNSHAVYPNW WNKCGLRAFSANGWE

....

MKFSNSAMIYFANSLYYCWYQINLFHTRKSHWWWDPPCADYWGLPRKLCLMHIRSYVPEV SKKMRRLWRDPGYKL

....

YNWTDPNFGHPTTNNLMTRYRIFTAQCNQRGRHPAHDESITFDPFGSTHEAVVARSQPTK SSFDFFNIFYDEQCM

....

RYSTIKTSNWDASMNDRPCHVHVPSGFNWAHPNPFKDWTGEGWMKSKCSEGNDFCALYNP PHTIRAYWFFWVSRF

....

THFPMMSCMCLITLKDLFLHRNEKQFVMHDQMPNPGAKMPYAWHNKGRSAAHSACISIHS TLWLIMTAVGLEIIC

....

KFDSWWNHPVMAQGNVPLQKNCSIDEFIPNQSSMKINHGFARTIGCFWEDLFPTQTENRL WICACDWPDFDAWCT

....

YKDNNMPFPGCKLWIWHFLVVDTFNIWCEERELGVYHWKQRDDMMMPEKMFGFWEWVPCM FEASGALGHGLEWSF

....

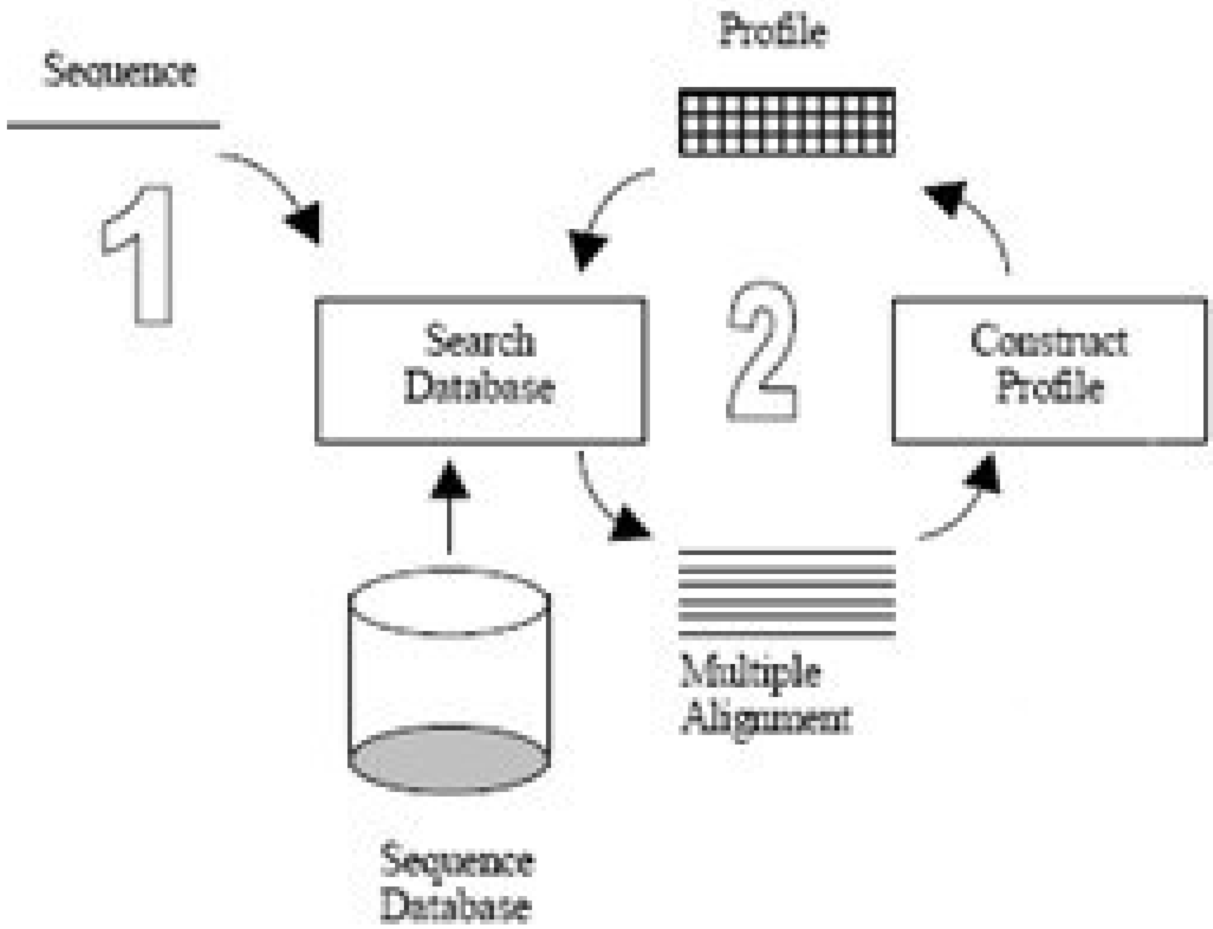
LICYGYEDFNPAAISFTRMHCKVSLGIWMVWNEYKIYVPRHAYECGICYNHKRMREPCGW AGHLLAYPVAMIAAA

....

THLAPYKYMMLYMKFDEGVRLILKGHEIQACFPTFSRNWGCFTQASTRQGCMGYWSRKIK DMIHCCNTHACMHLS

PSI-BLAST

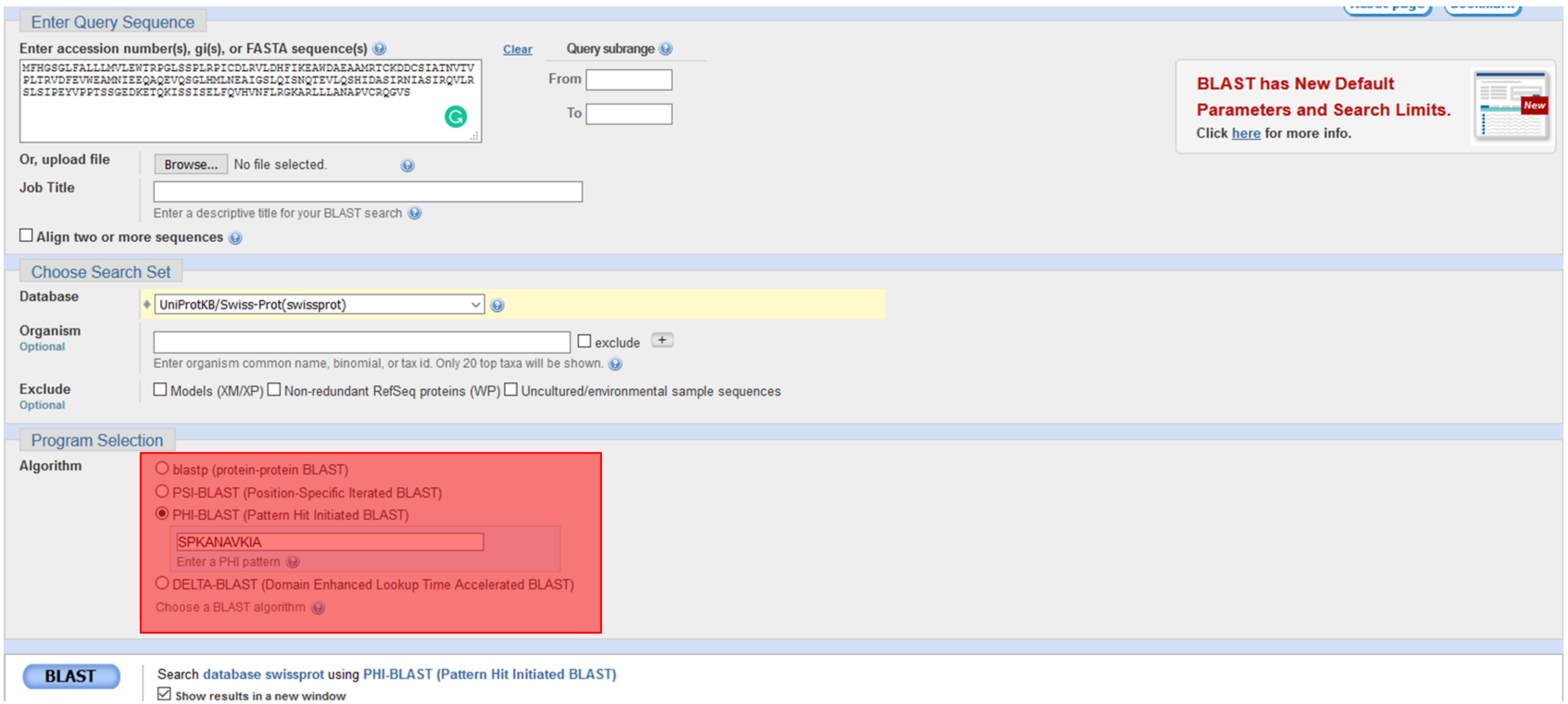
- PSI-BLAST: Working methodology



PHI-BLAST

- PHI-BLAST: Pattern Hit Initiated BLAST

- Idea is that many proteins contain a signature of sequences which can be utilized to search for homologous sequences containing the motif.



Enter Query Sequence

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

```
MFHGSGLFALLLMVLEWTRPGLSSPLRPICDLRVLDHFIKEAWDAEAAMRTCKDDCSIATNVTV
PLTRVDFEVWEAMNIEEQAQEVQSGLHMLNEAIGSLQISNQTEVLQSHIDASIRNIASIRQVLR
SLSIPEYVPPTSSGEDKETQKISSISELFQVHVNFLRGKARLLLANAPVCRQGVS
```

Query subrange

From

To

BLAST has New Default Parameters and Search Limits. Click here for more info.

Or, upload file: Browse... No file selected.

Job Title:

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: UniProtKB/Swiss-Prot(swissprot)

Organism (Optional): □ exclude

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:

○ blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

● PHI-BLAST (Pattern Hit Initiated BLAST)

SPKANAVKIA

Enter a PHI pattern

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

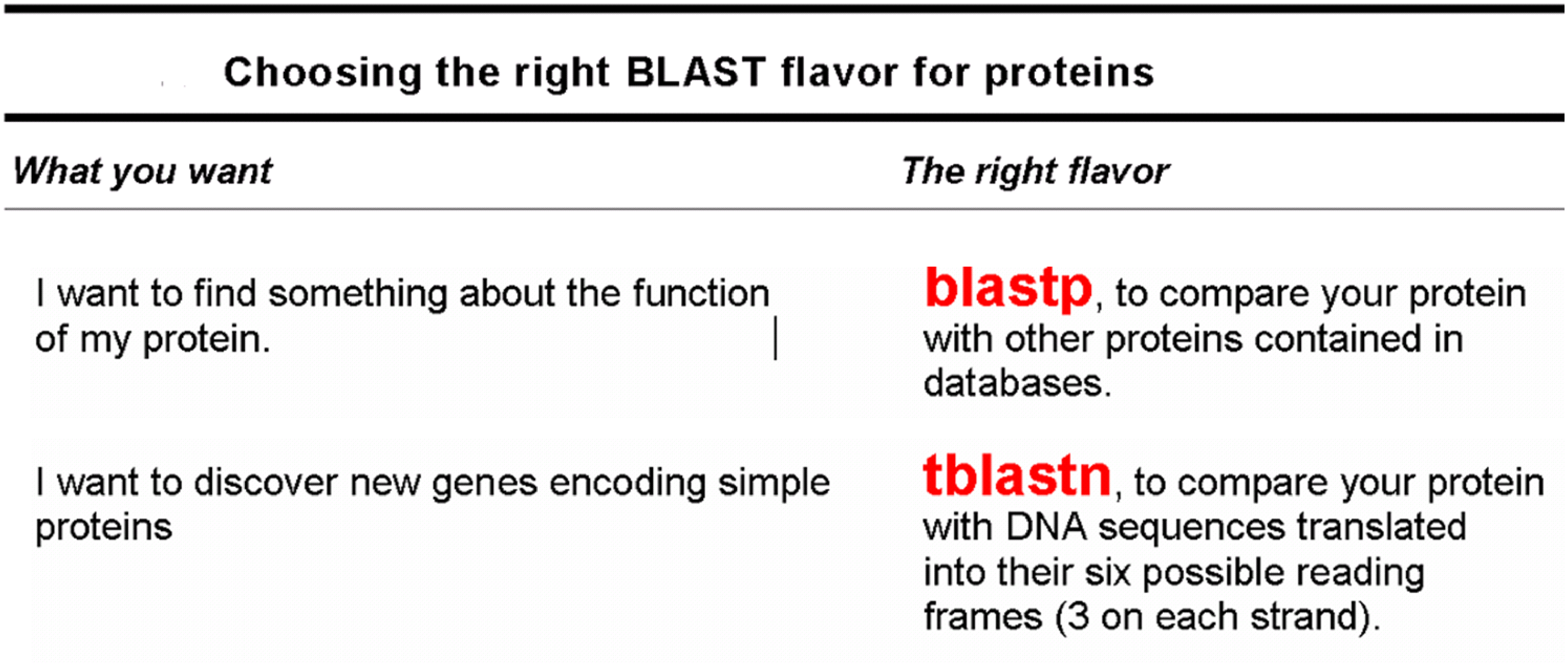
Choose a BLAST algorithm

Search database swissprot using PHI-BLAST (Pattern Hit Initiated BLAST)

☑ Show results in a new window

- The output of the PHI-BLAST is the same as that of the PSI-BLAST except that the position of the signature is highlighted in each of the alignments.

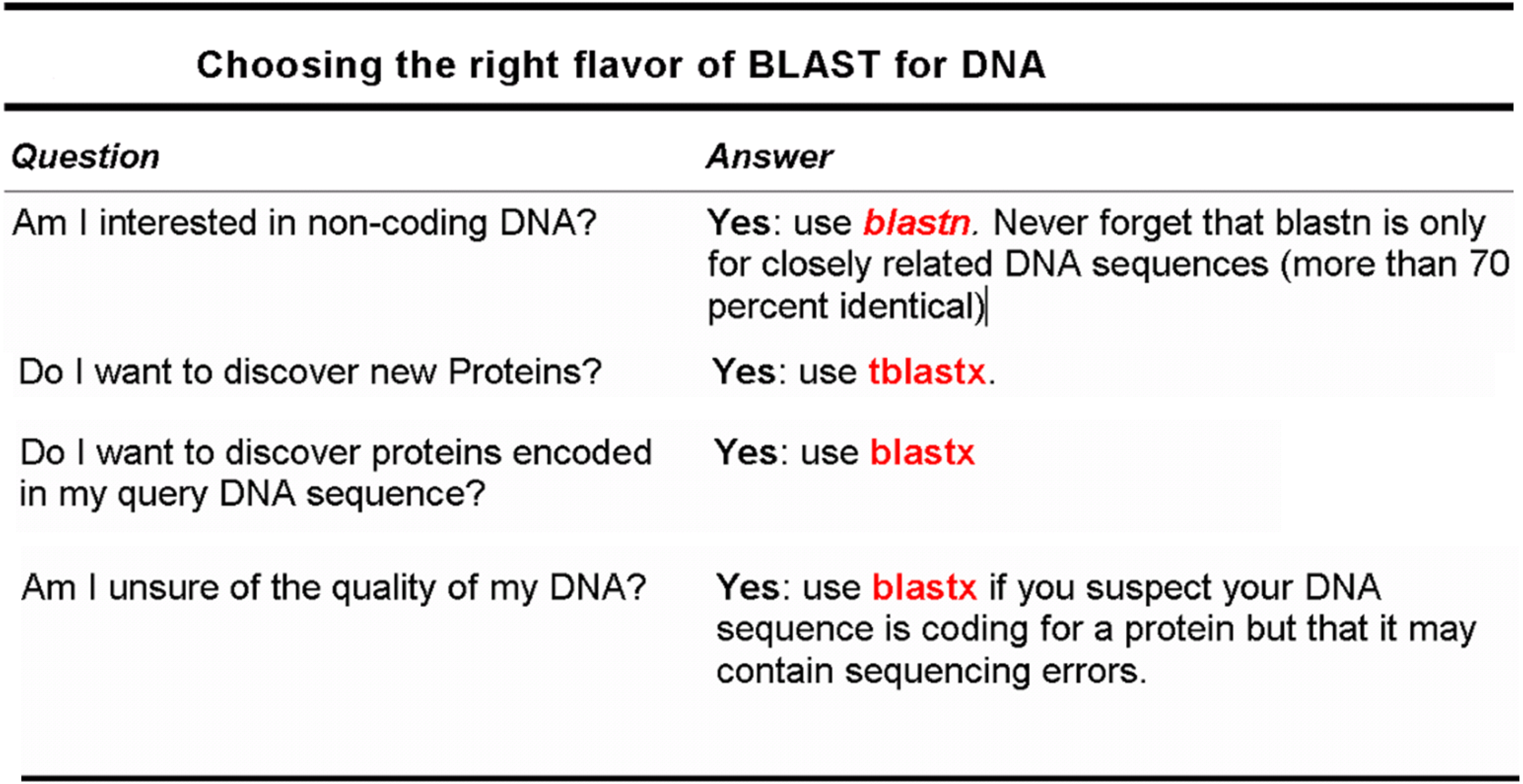
BLASTing a Protein Sequence



Choosing the right BLAST flavor for proteins

What you want	*The right flavor*
I want to find something about the function of my protein.	**blastp**, to compare your protein with other proteins contained in databases.
I want to discover new genes encoding simple proteins	**tblastn**, to compare your protein with DNA sequences translated into their six possible reading frames (3 on each strand).

Asking the Right Question with BLAST

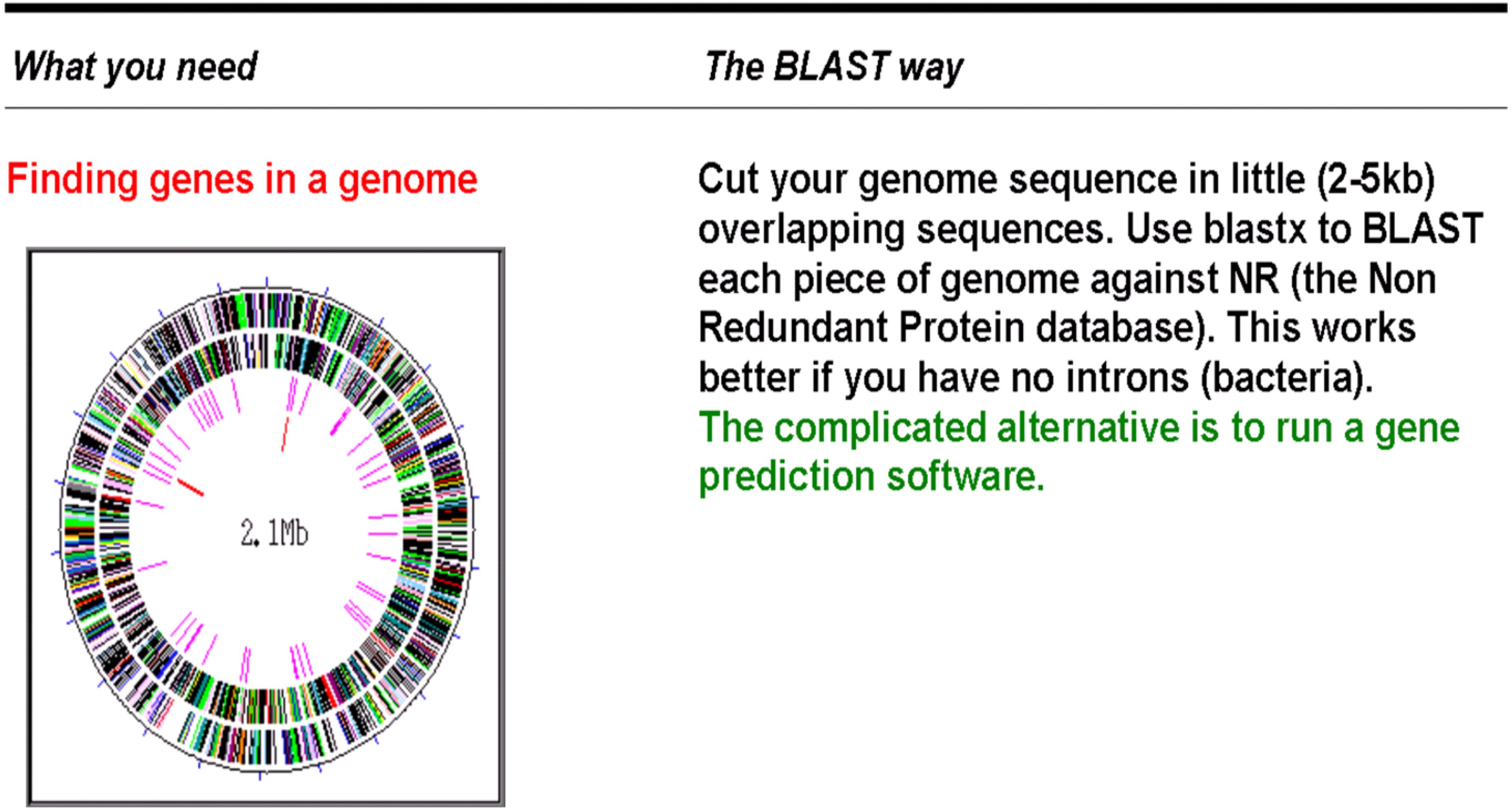


Choosing the right flavor of BLAST for DNA

Question	*Answer*
Am I interested in non-coding DNA?	**Yes**: use ***blastn.*** Never forget that blastn is only for closely related DNA sequences (more than 70 percent identical)
Do I want to discover new Proteins?	**Yes**: use **tblastx**.
Do I want to discover proteins encoded in my query DNA sequence?	**Yes**: use **blastx**
Am I unsure of the quality of my DNA?	**Yes**: use **blastx** if you suspect your DNA sequence is coding for a protein but that it may contain sequencing errors.

The BLAST Way of Doing Things

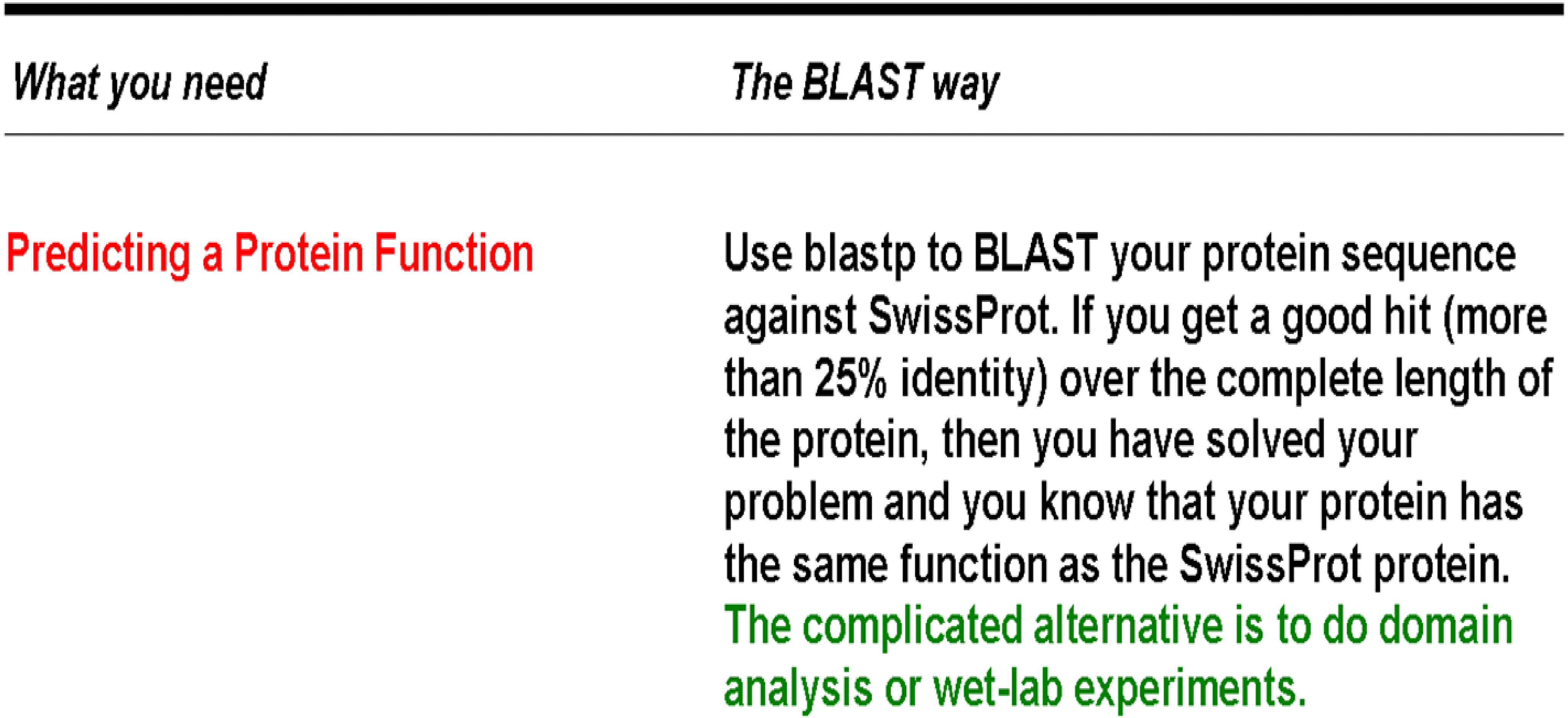
Gene-hunting with BLAST



What you need	*The BLAST way*
Finding genes in a genome	Cut your genome sequence in little (2-5kb) overlapping sequences. Use blastx to BLAST each piece of genome against NR (the Non Redundant Protein database). This works better if you have no introns (bacteria). The complicated alternative is to run a gene prediction software.

The BLAST Way of Doing Things

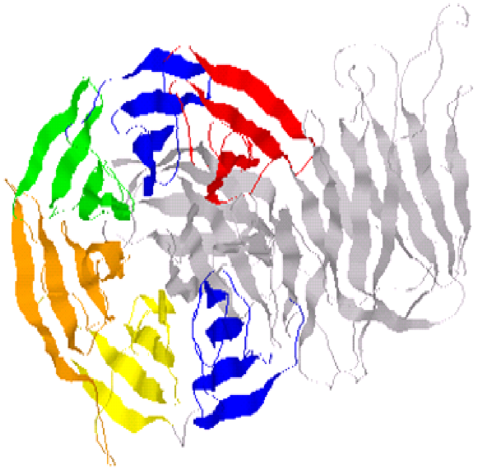
Predicting protein function with BLAST



What you need	*The BLAST way*
Predicting a Protein Function	Use blastp to BLAST your protein sequence against SwissProt. If you get a good hit (more than 25% identity) over the complete length of the protein, then you have solved your problem and you know that your protein has the same function as the SwissProt protein. The complicated alternative is to do domain analysis or wet-lab experiments.

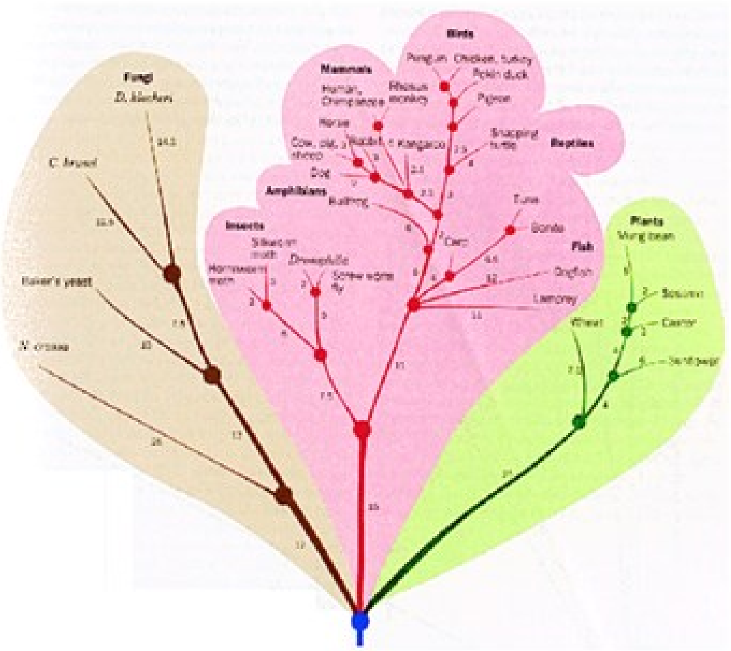
The BLAST Way of Doing Things

Structural analysis with BLAST

What you need	*The BLAST way*
Predicting a Protein 3D structure 	Use blastp to BLAST your protein against PDB (the database of protein structure). If you get a good hit, (more than 25% identity), then you know that your protein and this good hit have a similar 3D structure. The complicated alternative is to do homology modeling, Xray or NMR analysis of your protein.

The BLAST Way of Doing Things

Gathering members of a protein family

What you need	*The BLAST way*
Finding a protein family members  Fungi D. blachen C. bruset Baker's yeast N. crassa Mammals Human, Chimpanzee Rhesus monkey Horse Cow, pig, sheep Rabbit Kangaroo Dog Birds Penguin Chicken, turkey Pekin duck Pigeon Snapping turtle Reptiles Amphibians Bullfrog Insects Silkworm moth Drosophila Hornworm moth Screw worm fly Tuna Bonito Carp Fish Dogfish Lamprey Plants Mung bean Sesame Wheat Castor Sunflower	Use blastp (or its more powerful cousin Psi-BLAST) and run it on NR the non-redundant protein family. Once you have all the members of the family, you can make a multiple sequence alignment and draw a phylogenic tree. The Complicated alternative is to use PCR for Clonning your sequences

Thank You