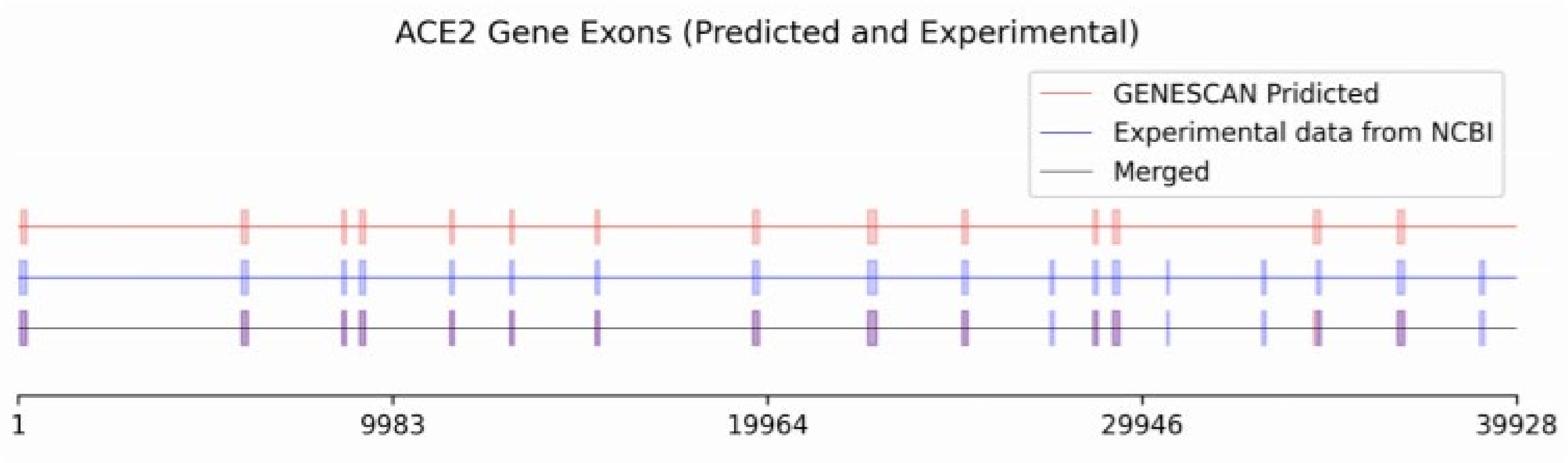
Practical 4

pairwise alignment

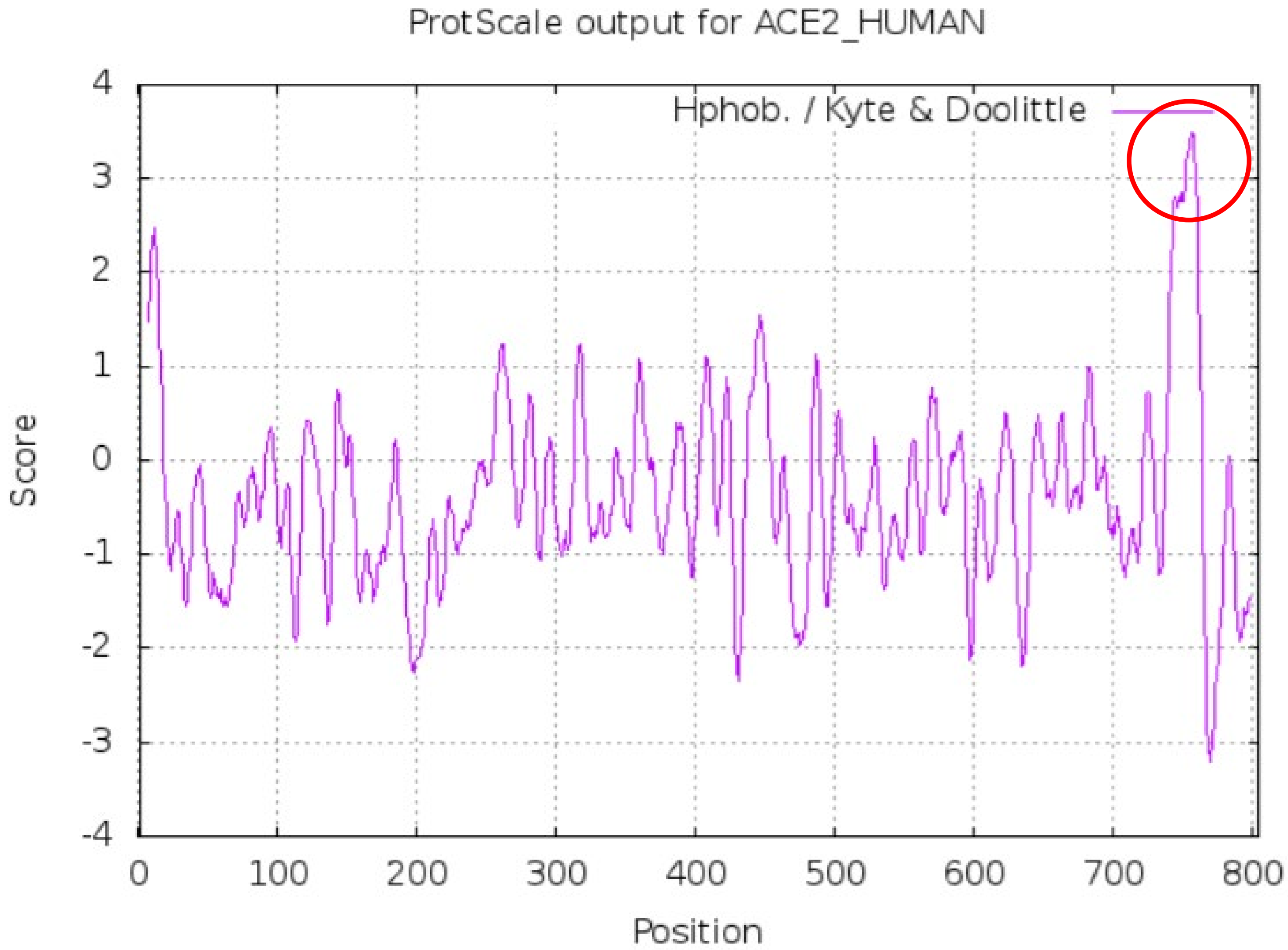
唐凯临

2021.4

REVIEW


ACE2 Gene Exons (Predicted and Experimental)
GENESCAN Pridicted
Experimental data from NCBI
Merged
1
9983
19964
29946
39928

REVIEW


ProtScale output for ACE2_HUMAN
Hphob. / Kyte & Doolittle
Score
4
3
2
1
0
-1
-2
-3
-4
0
100
200
300
400
500
600
700
800
Position

REVIEW

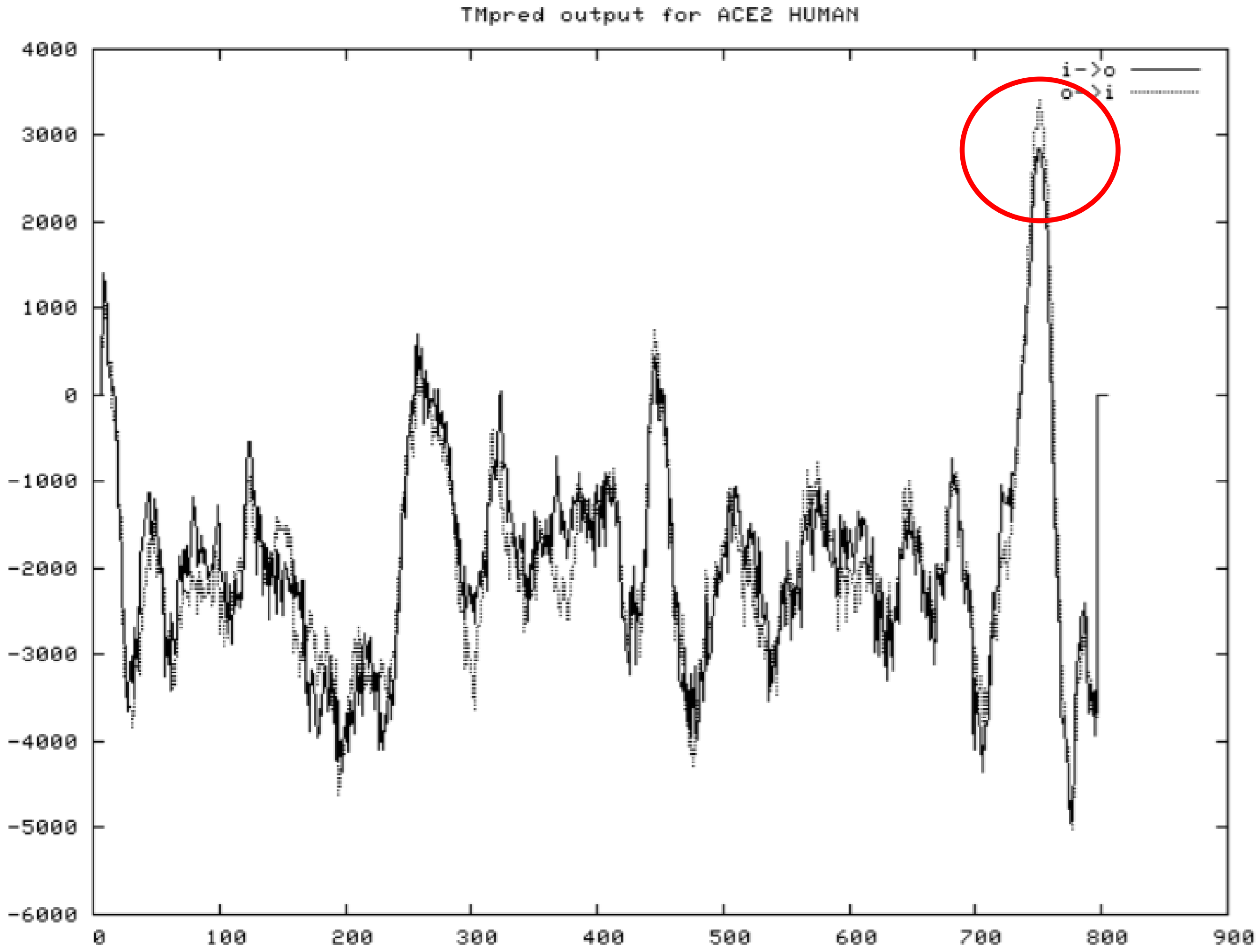
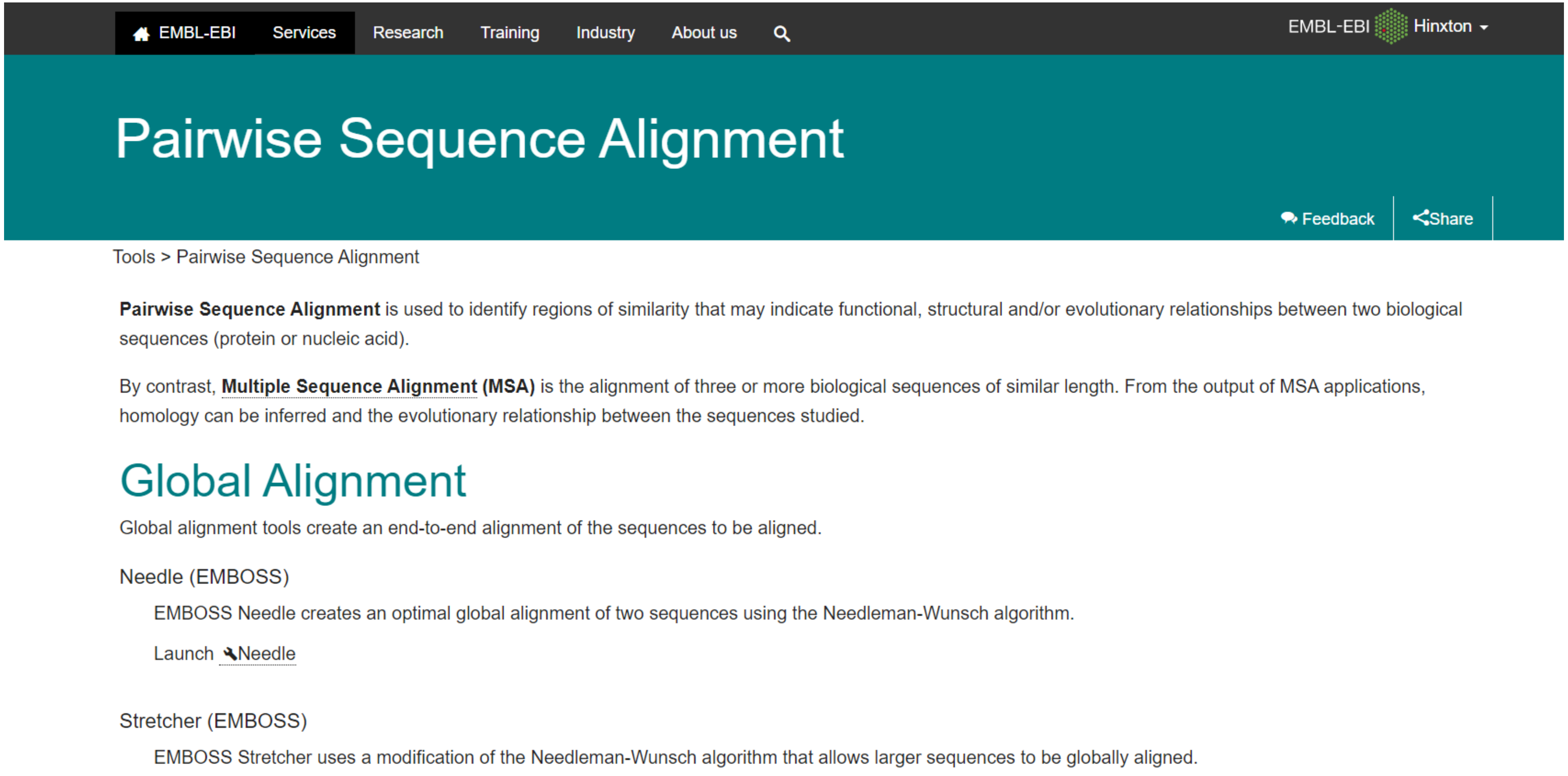


Figure 4: TMpred output for ACE2_HUMAN

EMBL全局双序列比对工具

- https://www.ebi.ac.uk/Tools/psa/



Pairwise Sequence Alignment

Tools > Pairwise Sequence Alignment

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

By contrast, **Multiple Sequence Alignment (MSA)** is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned.

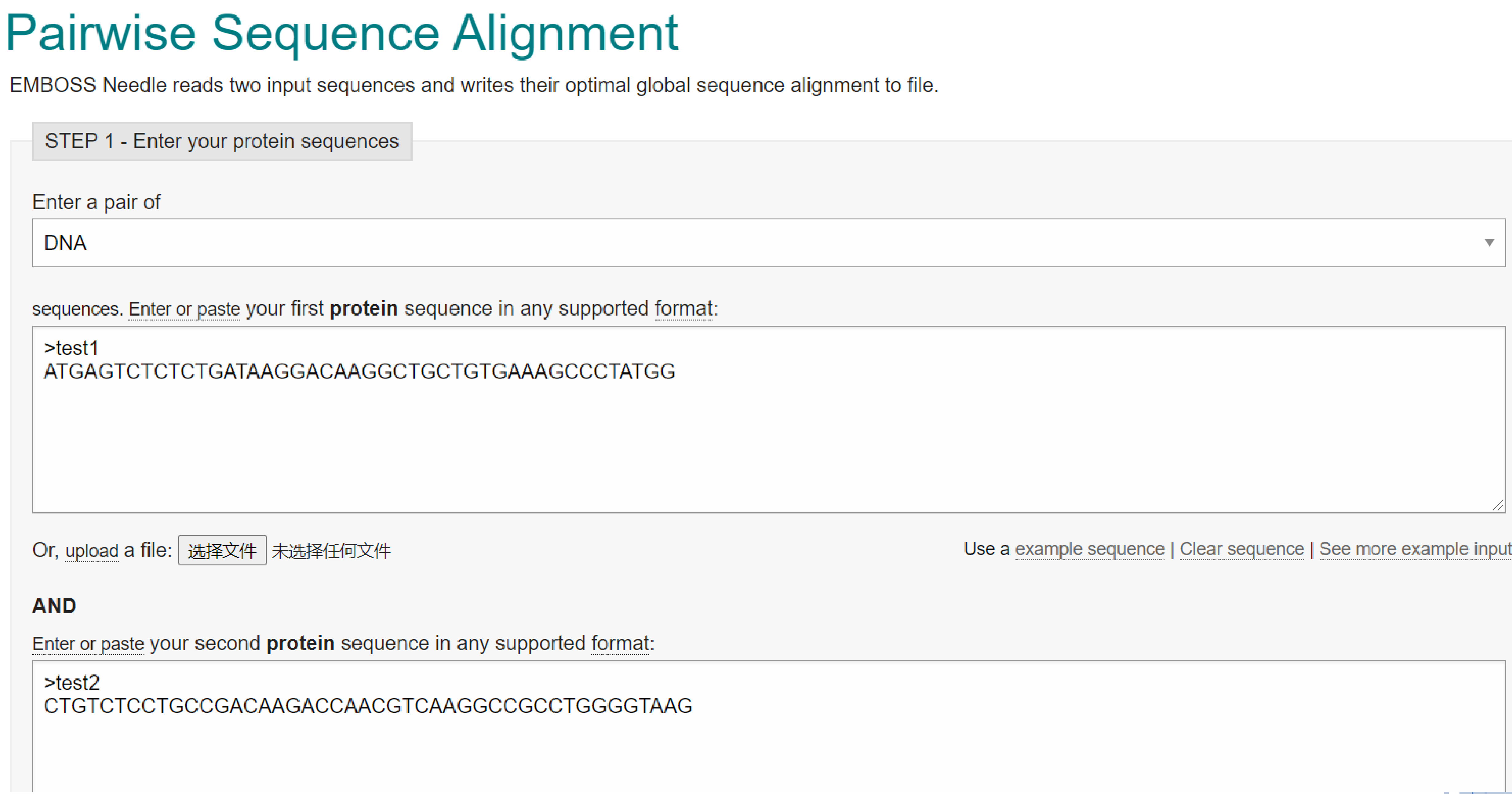
Needle (EMBOSS)

EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

Launch Needle

Stretcher (EMBOSS)

EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.



Pairwise Sequence Alignment

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

STEP 1 - Enter your protein sequences

Enter a pair of

DNA ▾

sequences. Enter or paste your first **protein** sequence in any supported format:

```
>test1
ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG
```

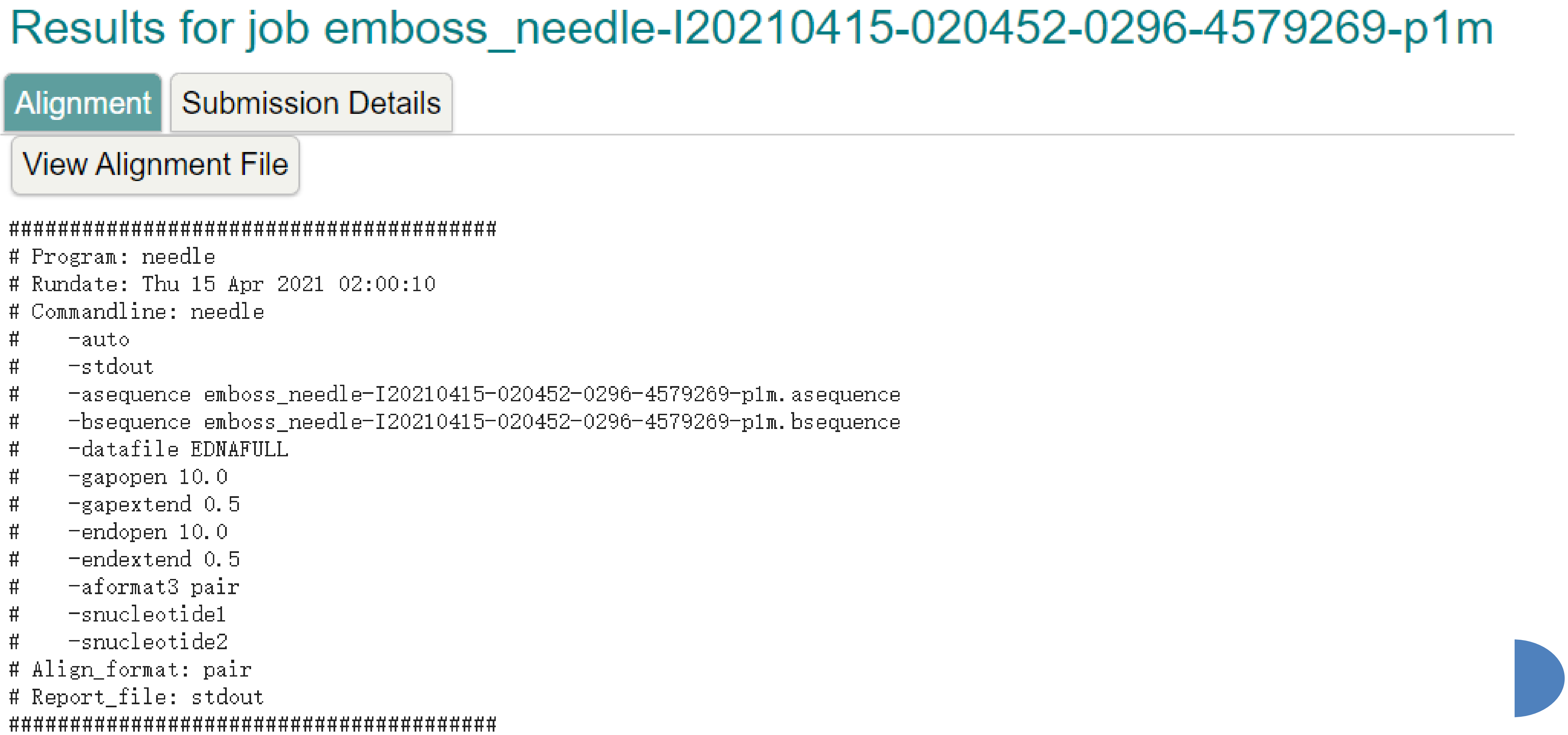
Or, upload a file: 选择文件 未选择任何文件

Use a example sequence | Clear sequence | See more example input

AND

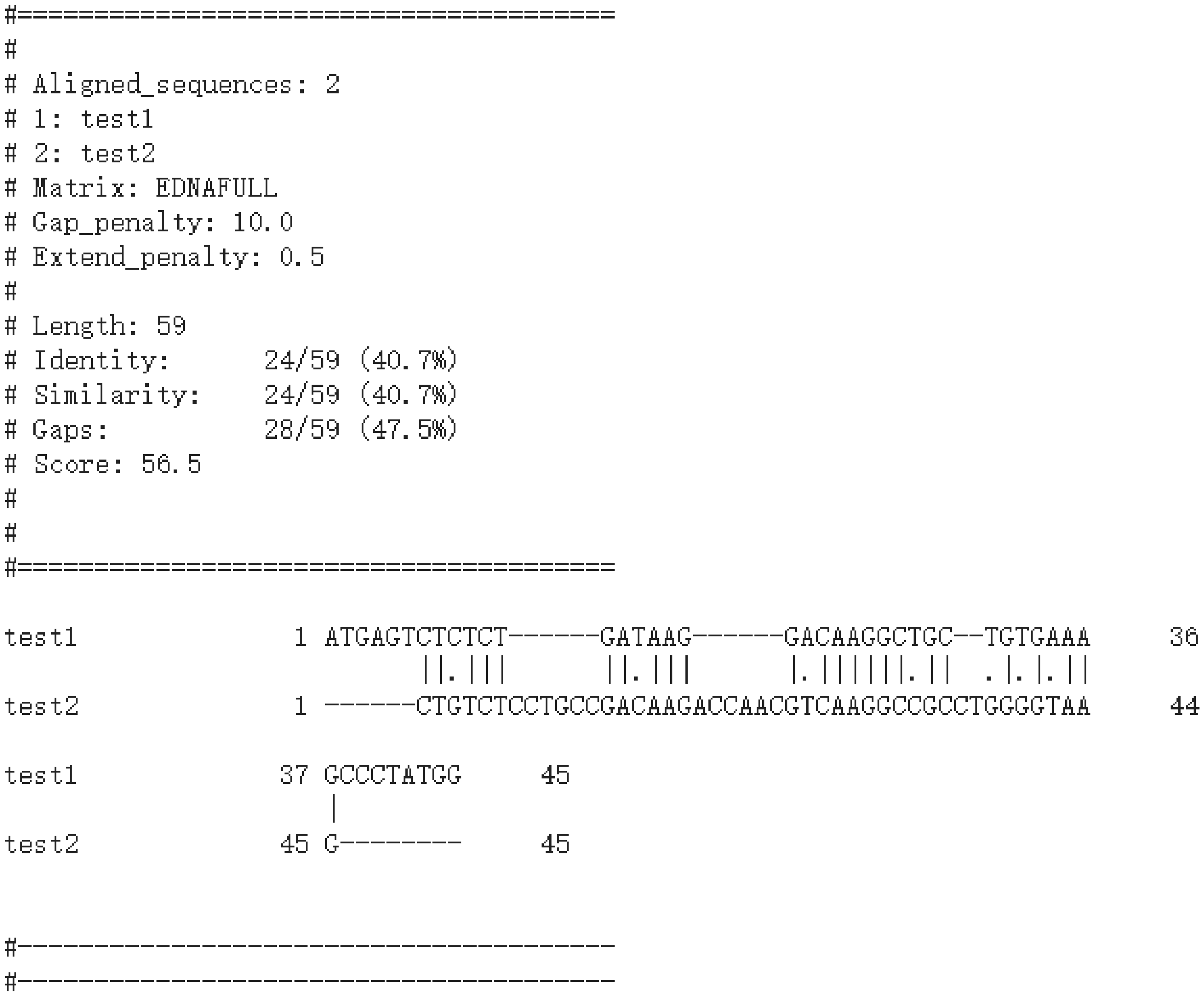
Enter or paste your second **protein** sequence in any supported format:

```
>test2
CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG
```

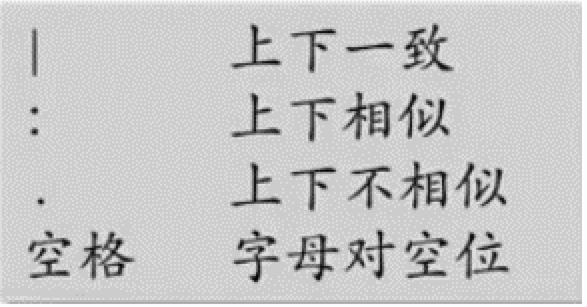


Results for job emboss_needle-I20210415-020452-0296-4579269-p1m

```
########################################
# Program: needle
# Rundate: Thu 15 Apr 2021 02:00:10
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20210415-020452-0296-4579269-p1m.asequence
#    -bsequence emboss_needle-I20210415-020452-0296-4579269-p1m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################
```



```
#=======================================
#
# Aligned_sequences: 2
# 1: test1
# 2: test2
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 59
# Identity:      24/59 (40.7%)
# Similarity:    24/59 (40.7%)
# Gaps:          28/59 (47.5%)
# Score: 56.5
#
#
#=======================================

test1              1 ATGAGTCTCTCT------GATAAG------GACAAGGCTGC--TGTGAAA     36
                           ||.|||      ||.|||      |.||||||.||  .|.|.||
test2              1 ------CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGTAA     44

test1             37 GCCCTATGG     45
                     |
test2             45 G--------     45


#---------------------------------------
#---------------------------------------
```



|	上下一致
:	上下相似
.	上下不相似
空格	字母对空位

QUESTION

- 已知要比对的两条序列是同源序列，猜测他们结构和功能类似。其中一条序列的结构已知，另一条未知。
- 如果序列比对，用其中已知结构的序列做模板，来预测另一个序列的结构。如何调整参数?

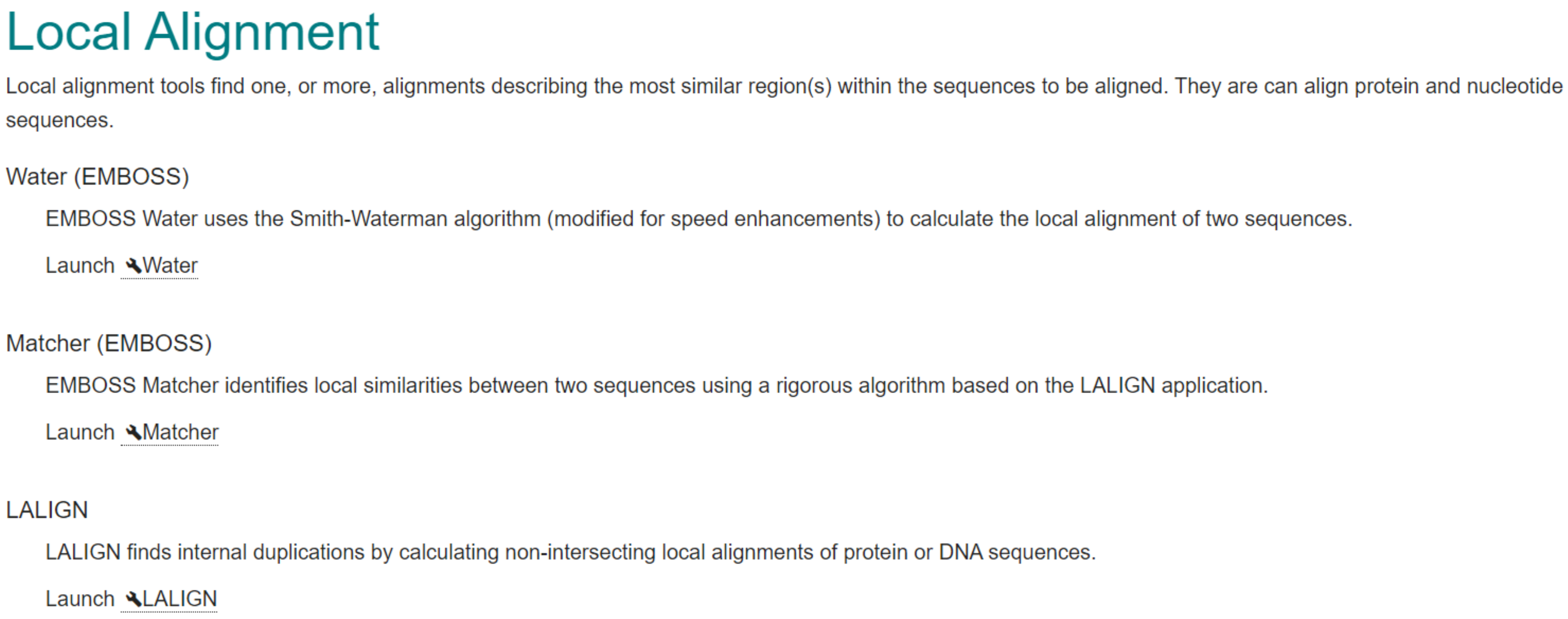
- Score matrix：Blosum大，PAM小
- Gap：gap开头小，延伸大。

QUESTION

- 已知比对的两条序列绝大部分区域都很相似，但是其中一条序列的一个功能区在另一条序列中是缺失的。
- 想要通过序列比对把这个功能区找出来。

- GAP：gap开头大，延伸小

EMBL局部双序列比对工具



Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. They are can align protein and nucleotide sequences.

Water (EMBOSS)

EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

Launch Water

Matcher (EMBOSS)

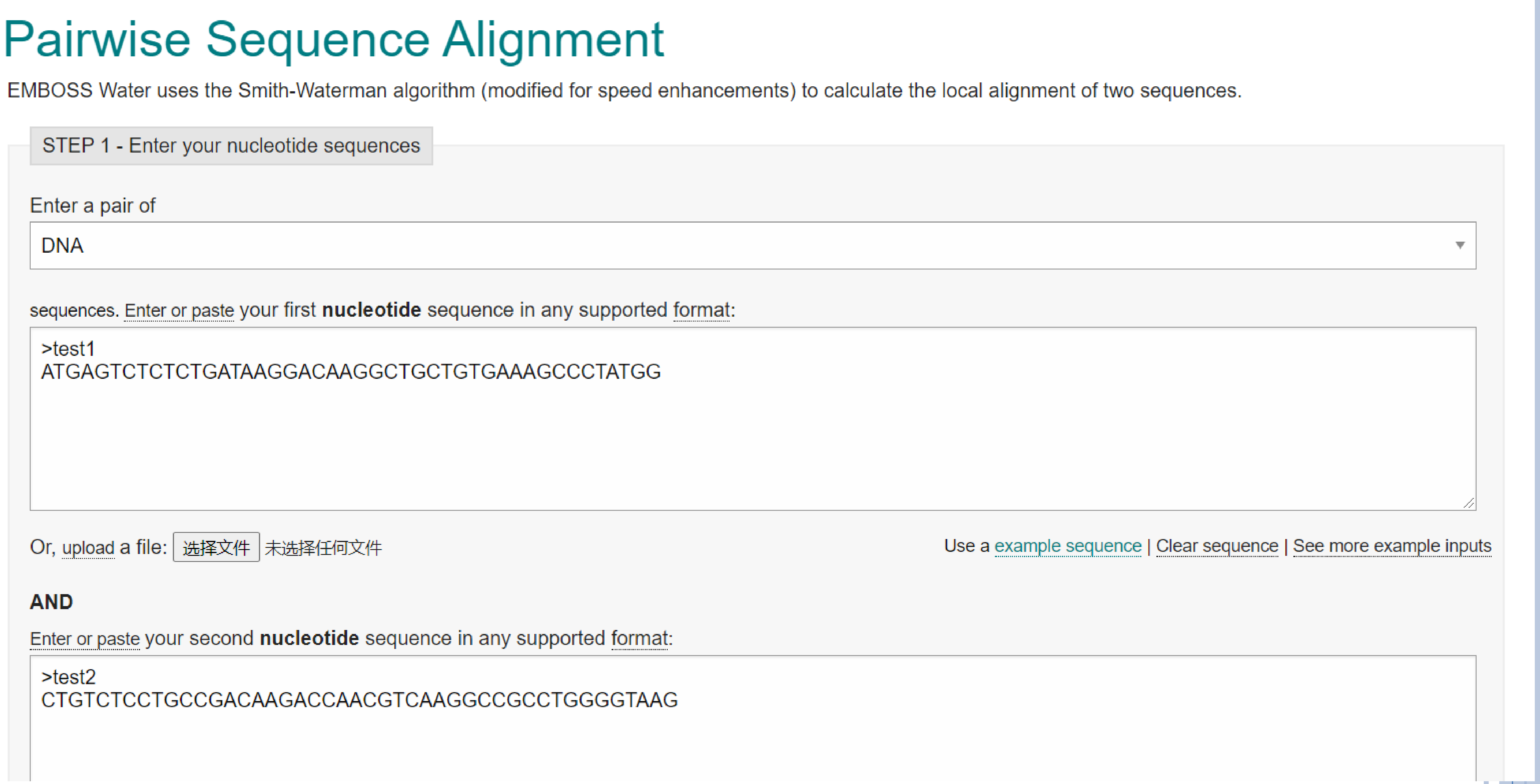
EMBOSS Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.

Launch Matcher

LALIGN

LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.

Launch LALIGN



Pairwise Sequence Alignment

EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

STEP 1 - Enter your nucleotide sequences

Enter a pair of

DNA

sequences. Enter or paste your first **nucleotide** sequence in any supported format:

```
>test1
ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG
```

Or, upload a file: 选择文件 未选择任何文件
Use a example sequence | Clear sequence | See more example inputs

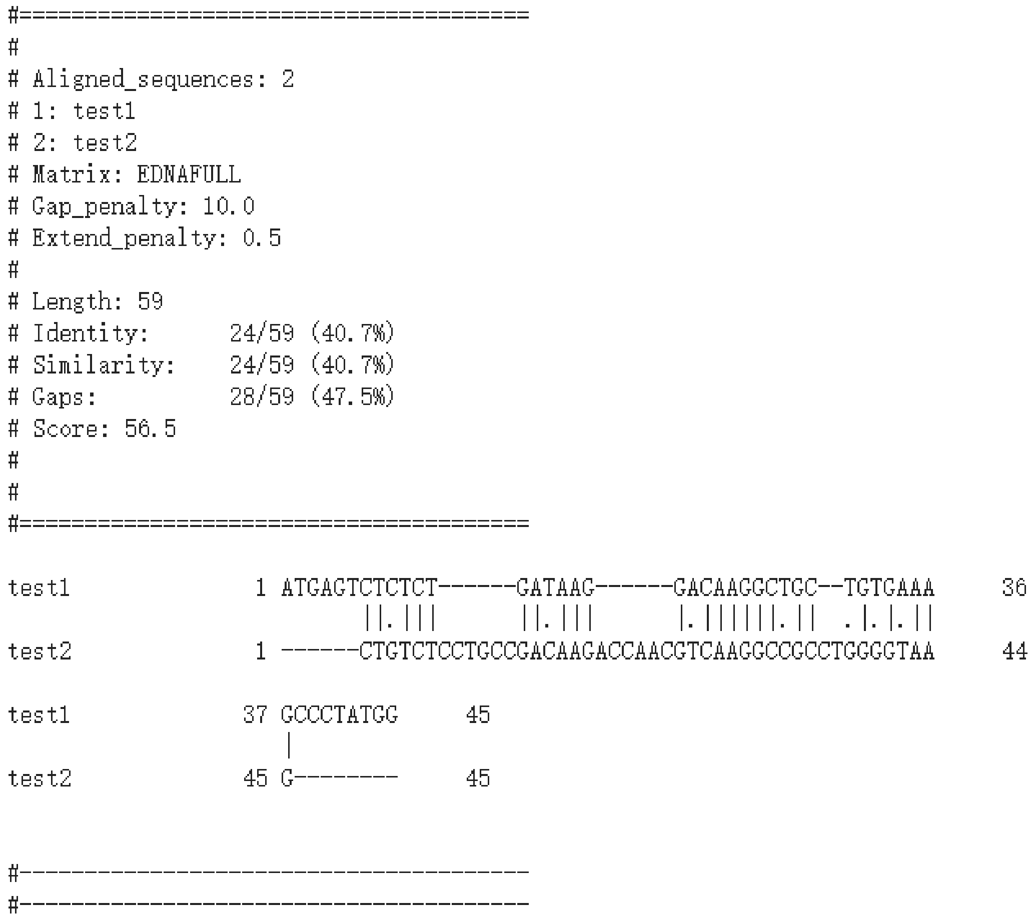
AND

Enter or paste your second **nucleotide** sequence in any supported format:

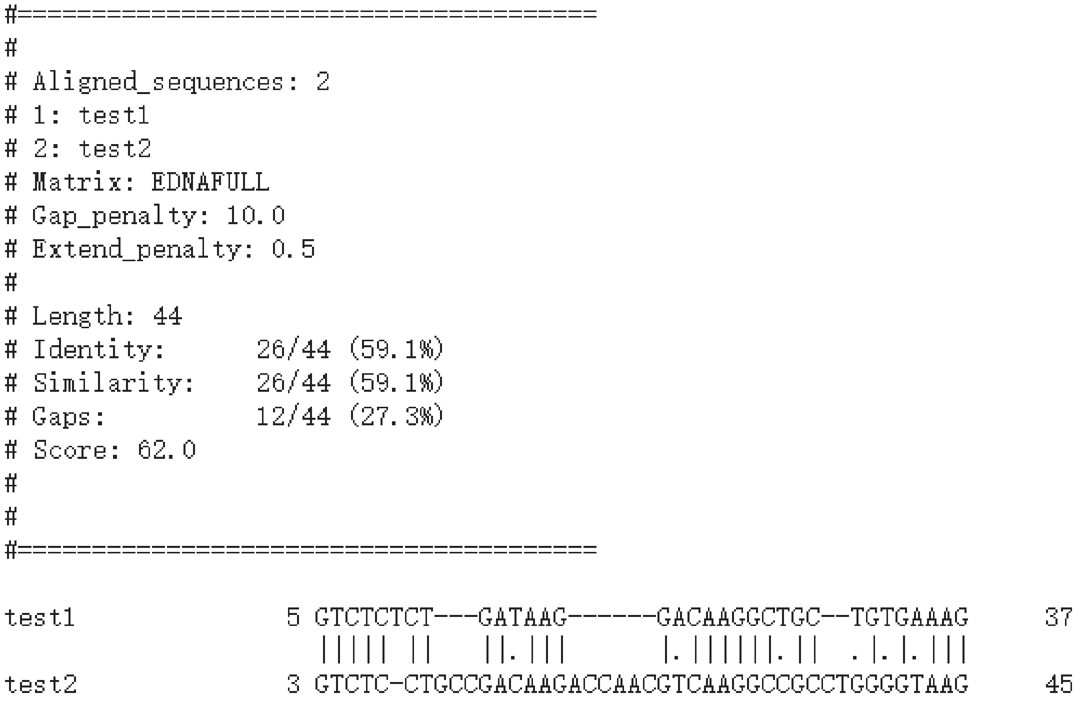
```
>test2
CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG
```

```
########################################
# Program: water
# Rundate: Thu 15 Apr 2021 02:39:04
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20210415-023900-0380-40159559-p2m.asequence
#    -bsequence emboss_water-I20210415-023900-0380-40159559-p2m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: test1
# 2: test2
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 44
# Identity:      26/44 (59.1%)
# Similarity:    26/44 (59.1%)
# Gaps:          12/44 (27.3%)
# Score: 62.0
#
#
#=======================================

test1              5 GTCTCTCT---GATAAG------GACAAGGCTGC--TGTGAAAG     37
                     ||||| ||   ||.|||      |.||||||.||  .|.|.|||
test2              3 GTCTC-CTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG     45
```

有GAP的情况：默认参数

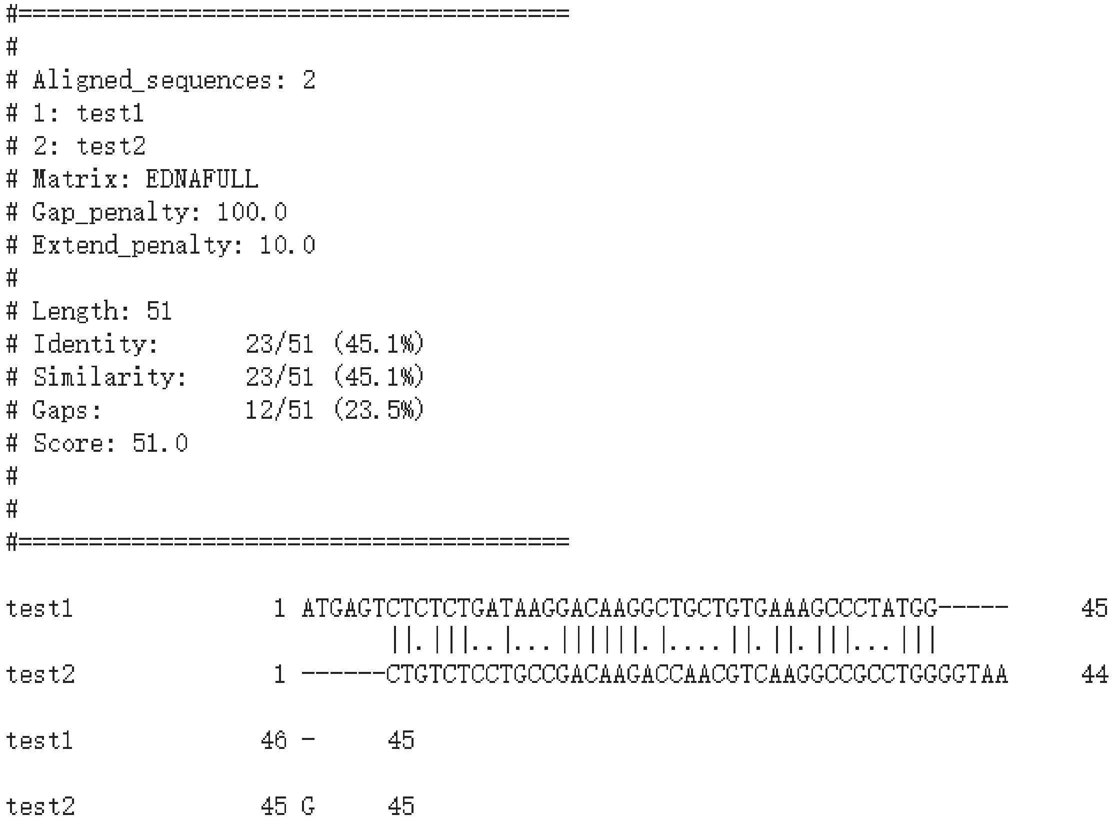


```
#=======================================
#
# Aligned_sequences: 2
# 1: test1
# 2: test2
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 59
# Identity:      24/59 (40.7%)
# Similarity:    24/59 (40.7%)
# Gaps:          28/59 (47.5%)
# Score: 56.5
#
#
#=======================================

test1              1 ATGAGTCTCTCT------GATAAG------GACAAGGCTGC--TGTGAAA     36
                           ||.|||      ||.|||      |.||||||.||  .|.|.||
test2              1 ------CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAA     44

test1             37 GCCCTATGG     45
                     |
test2             45 G--------     45


#---------------------------------------
#---------------------------------------
```

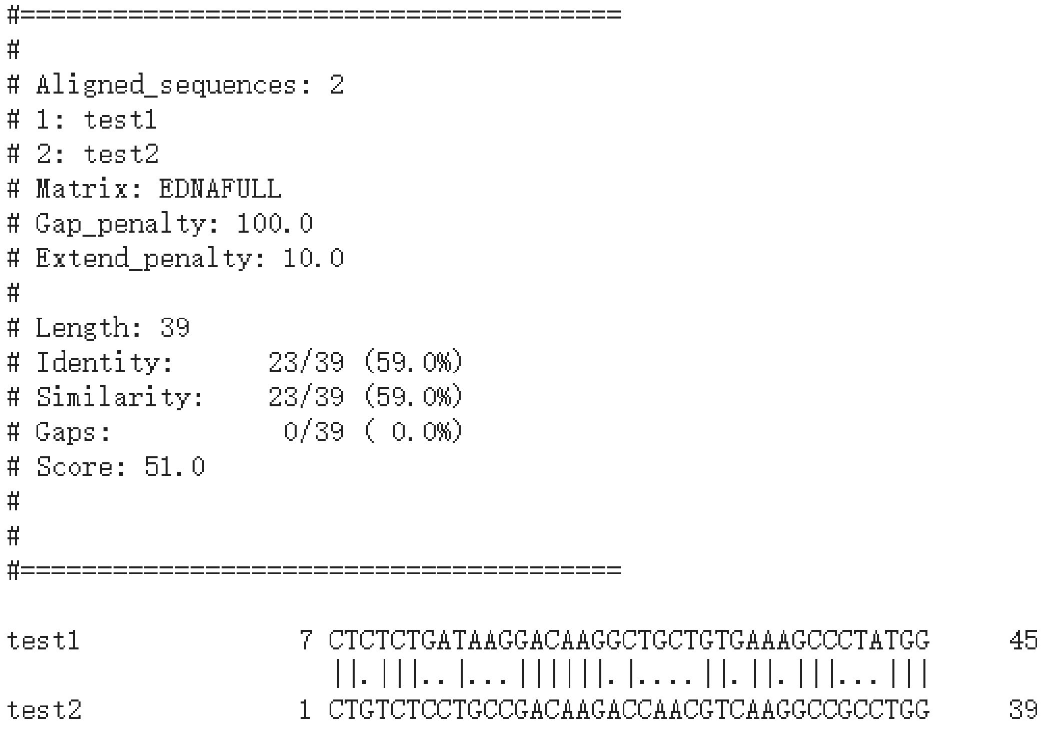


```
#=======================================
#
# Aligned_sequences: 2
# 1: test1
# 2: test2
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 44
# Identity:      26/44 (59.1%)
# Similarity:    26/44 (59.1%)
# Gaps:          12/44 (27.3%)
# Score: 62.0
#
#
#=======================================

test1              5 GTCTCTCT---GATAAG------GACAAGGCTGC--TGTGAAAG     37
                     ||||| ||   ||.|||      |.||||||.||  .|.|.|||
test2              3 GTCTC-CTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG     45
```

增加GAP罚分



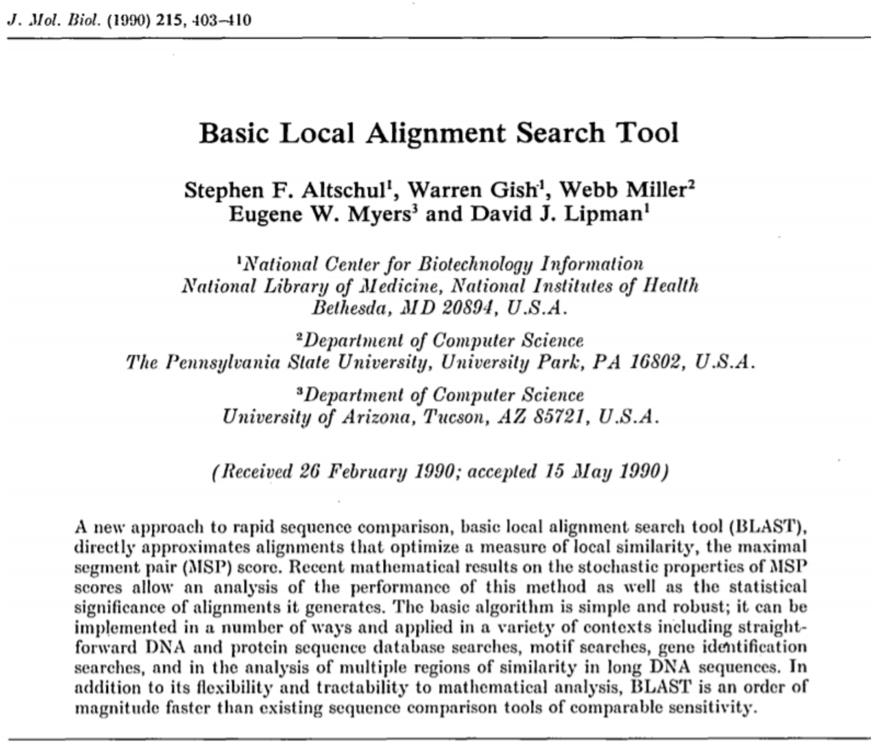
```
#=======================================
#
# Aligned_sequences: 2
# 1: test1
# 2: test2
# Matrix: EDNAFULL
# Gap_penalty: 100.0
# Extend_penalty: 10.0
#
# Length: 51
# Identity:      23/51 (45.1%)
# Similarity:    23/51 (45.1%)
# Gaps:          12/51 (23.5%)
# Score: 51.0
#
#
#=======================================

test1                  1 ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG-----     45
                               ||.|||..|...||||||.|....||.||.|||...|||
test2                  1 ------CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAA     44

test1                 46 -      45

test2                 45 G      45
```



```
#=======================================
#
# Aligned_sequences: 2
# 1: test1
# 2: test2
# Matrix: EDNAFULL
# Gap_penalty: 100.0
# Extend_penalty: 10.0
#
# Length: 39
# Identity:      23/39 (59.0%)
# Similarity:    23/39 (59.0%)
# Gaps:           0/39 ( 0.0%)
# Score: 51.0
#
#
#=======================================

test1                  7 CTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG     45
                         ||.|||..|...||||||.|....||.||.|||...|||
test2                  1 CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGG     39
```

BLAST

- More than 60K cites respectively



J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

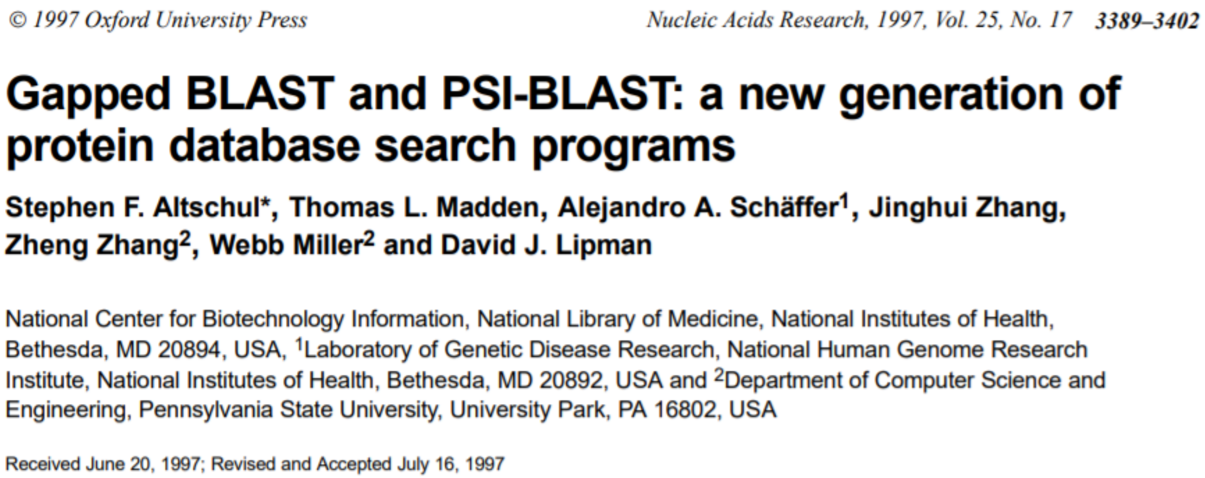
[1]*National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]*Department of Computer Science*
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]*Department of Computer Science*
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.



© 1997 Oxford University Press *Nucleic Acids Research, 1997, Vol. 25, No. 17* ***3389–3402***

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer[1], Jinghui Zhang, Zheng Zhang[2], Webb Miller[2] and David J. Lipman

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [1]Laboratory of Genetic Disease Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA and [2]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

Received June 20, 1997; Revised and Accepted July 16, 1997

1. Stephen F. Altschul, et.al., Basic local alignment search tool, Journal of Molecular Biology, 1990, 215(3): 403-410

2. Stephen F. Altschul, et.al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 1997, 25(17): 3389–3402

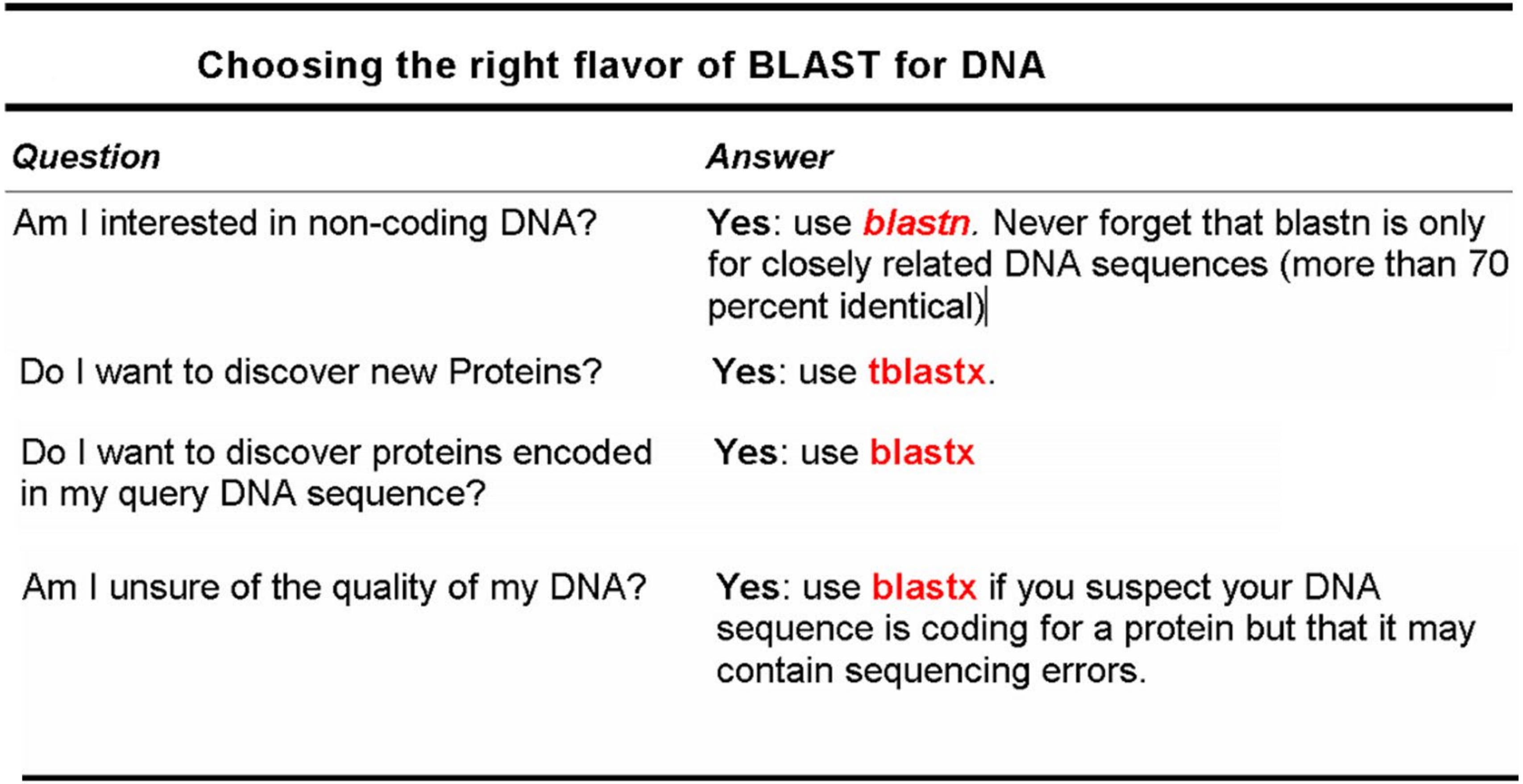
BLAST Flavors

Query Sequence	Search Database	BLAST Program	Sequence Comparison	BLAST output
DNA	nucleotide	**blastn**	compare query nucleotide against nucleotide db	Nucleotide
DNA	protein	**blastx**	translate query seq in all reading frames into amino acids, then compare with protein db	Amino acid
DNA	nucleotide	**tblastx**	translate both query & db seq in all reading frames, then compare between protein seqs	Amino acid
Protein	protein	**blastp**	compare query protein against protein db	Amino acid
Protein	nucleotide	**tblastn**	translate db nucleotide seq in all reading frames, then compare between protein seqs	Amino acid

How to remember?

- when you have "X" after "blast" – the query is translated
- when you have "T" before "blast"– the database is translated

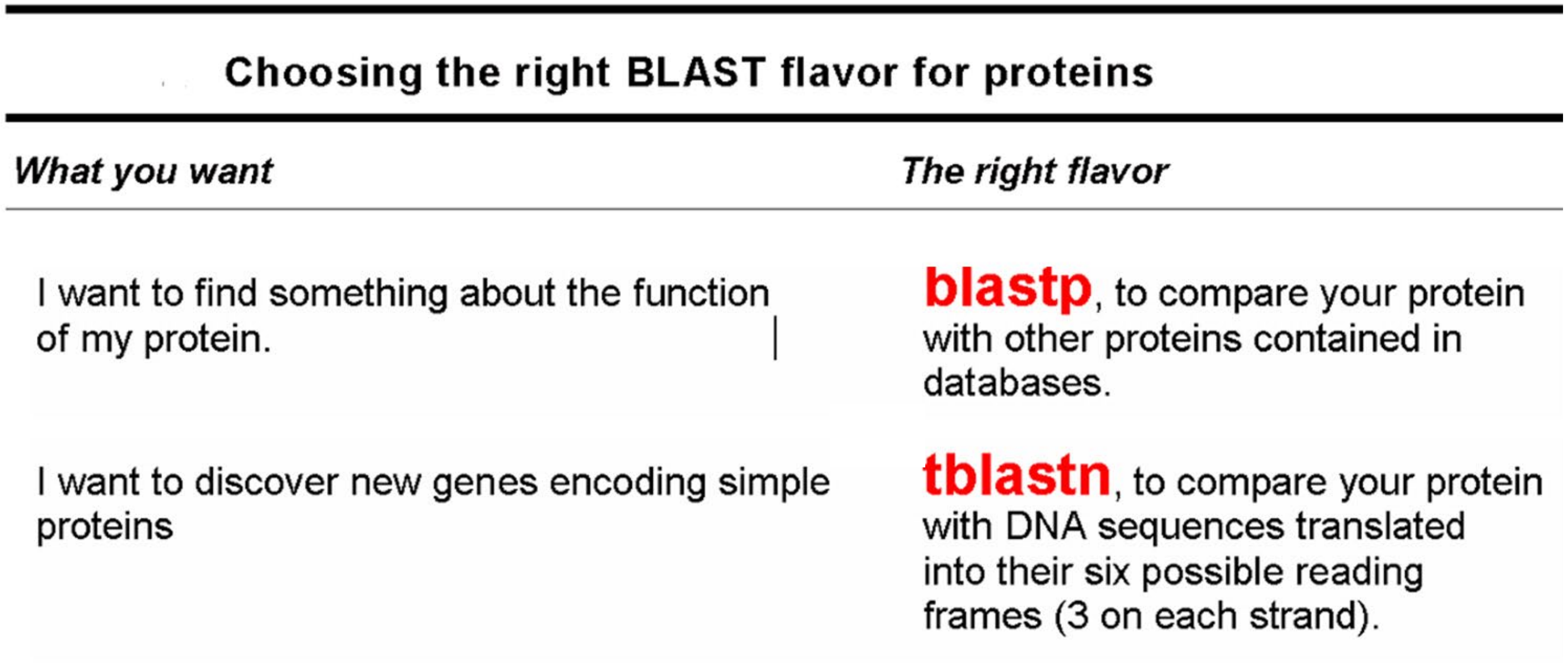
How to choose



Choosing the right flavor of BLAST for DNA

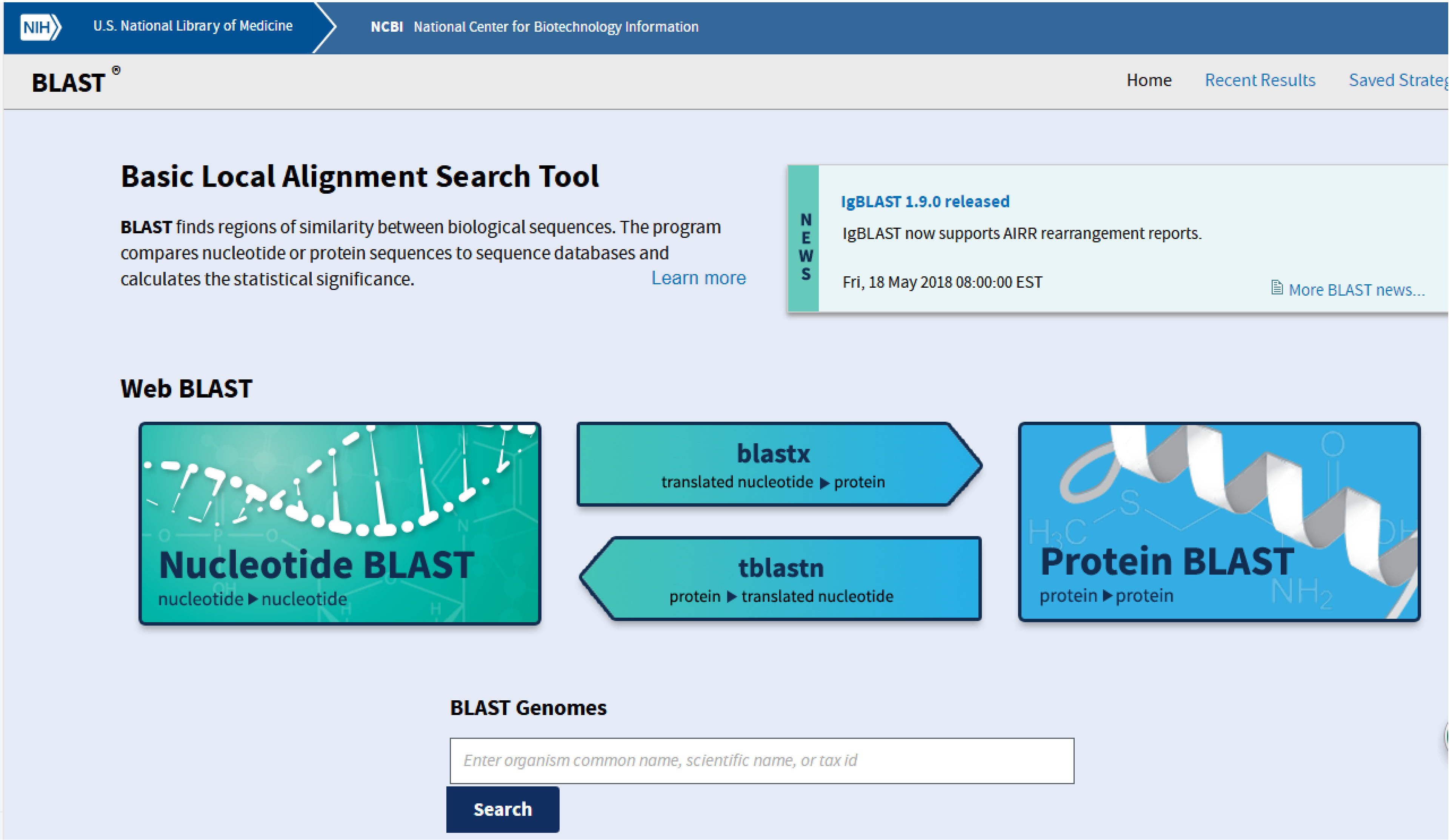
Question	*Answer*
Am I interested in non-coding DNA?	**Yes**: use ***blastn***. Never forget that blastn is only for closely related DNA sequences (more than 70 percent identical)
Do I want to discover new Proteins?	**Yes**: use **tblastx**.
Do I want to discover proteins encoded in my query DNA sequence?	**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.

How to choose



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).



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST ®
Home
Recent Results
Saved Strateg

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

IgBLAST 1.9.0 released

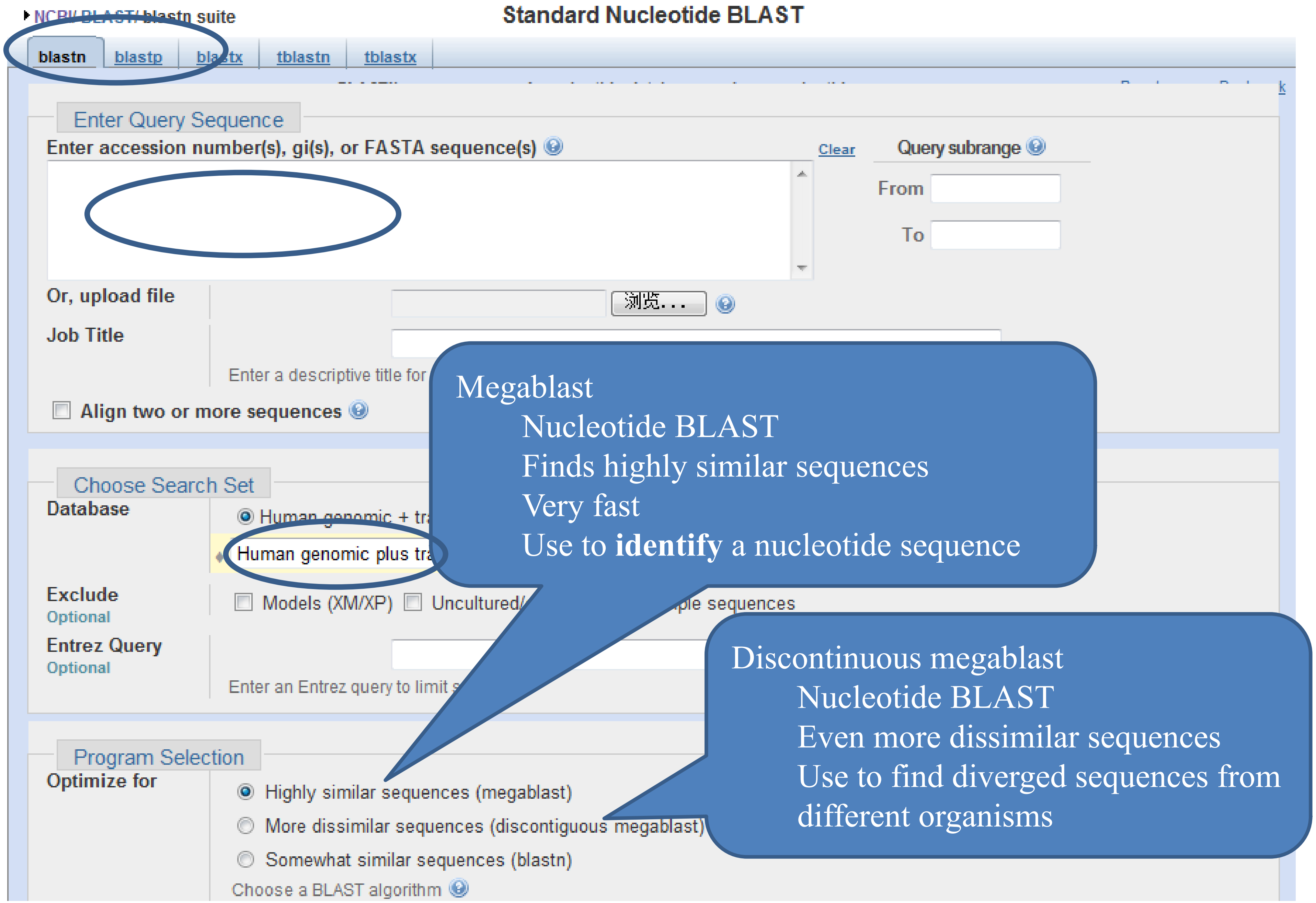
IgBLAST now supports AIRR rearrangement reports.

Fri, 18 May 2018 08:00:00 EST

More BLAST news...

Web BLAST

BLAST Genomes



▸ NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

Enter Query Sequence

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

Query subrange

From

To

Or, upload file 浏览...

Job Title

Enter a descriptive title for

□ Align two or more sequences

Choose Search Set

Database ● Human genomic + tr

Human genomic plus tr

Exclude Optional □ Models (XM/XP) □ Uncultured/ ...ple sequences

Entrez Query Optional

Enter an Entrez query to limit s

Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

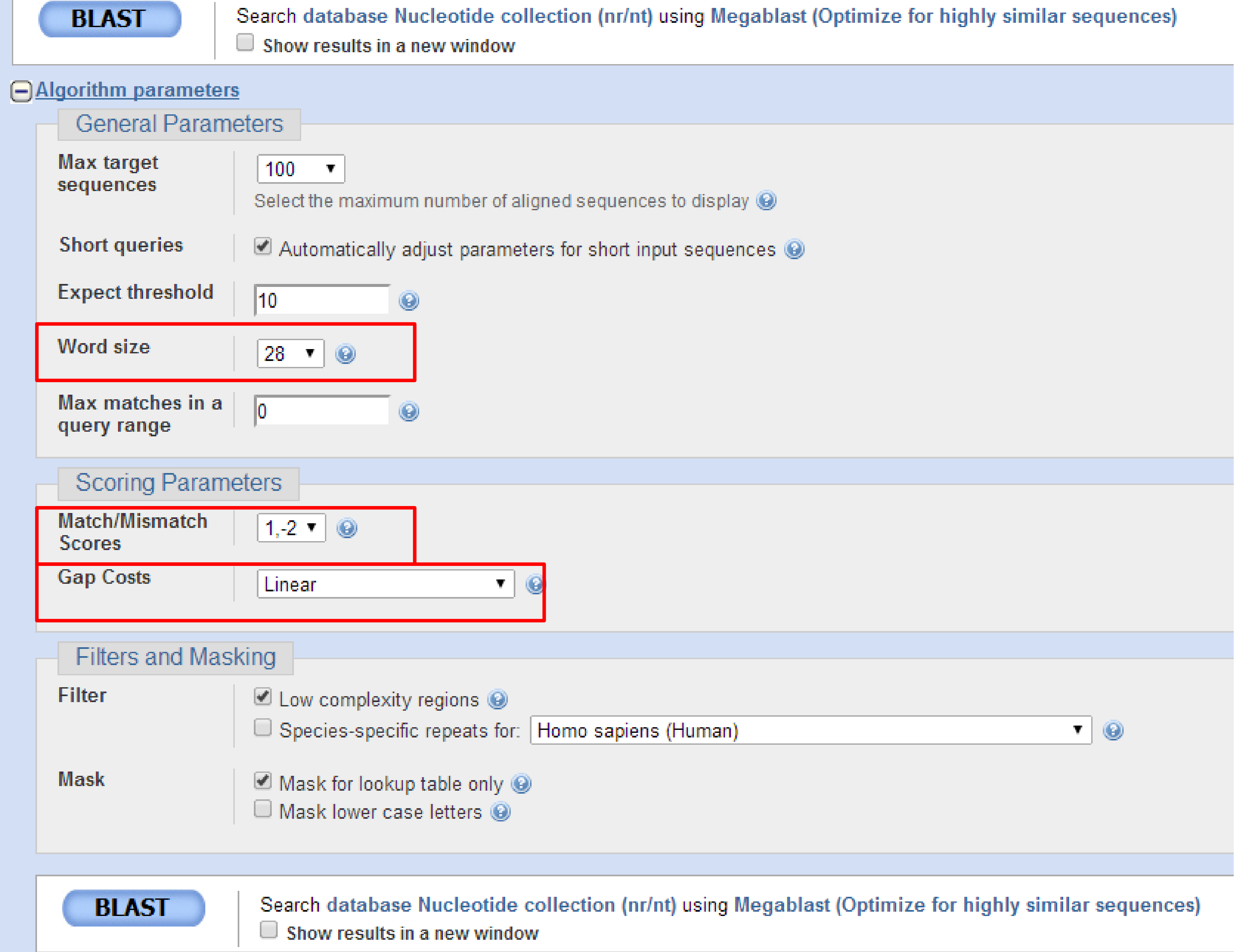
Megablast

- Nucleotide BLAST
- Finds highly similar sequences
- Very fast
- Use to **identify** a nucleotide sequence

Discontinuous megablast

- Nucleotide BLAST
- Even more dissimilar sequences
- Use to find diverged sequences from different organisms

Megablast



Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ 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: 10

Word size: 28

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

Filters and Masking

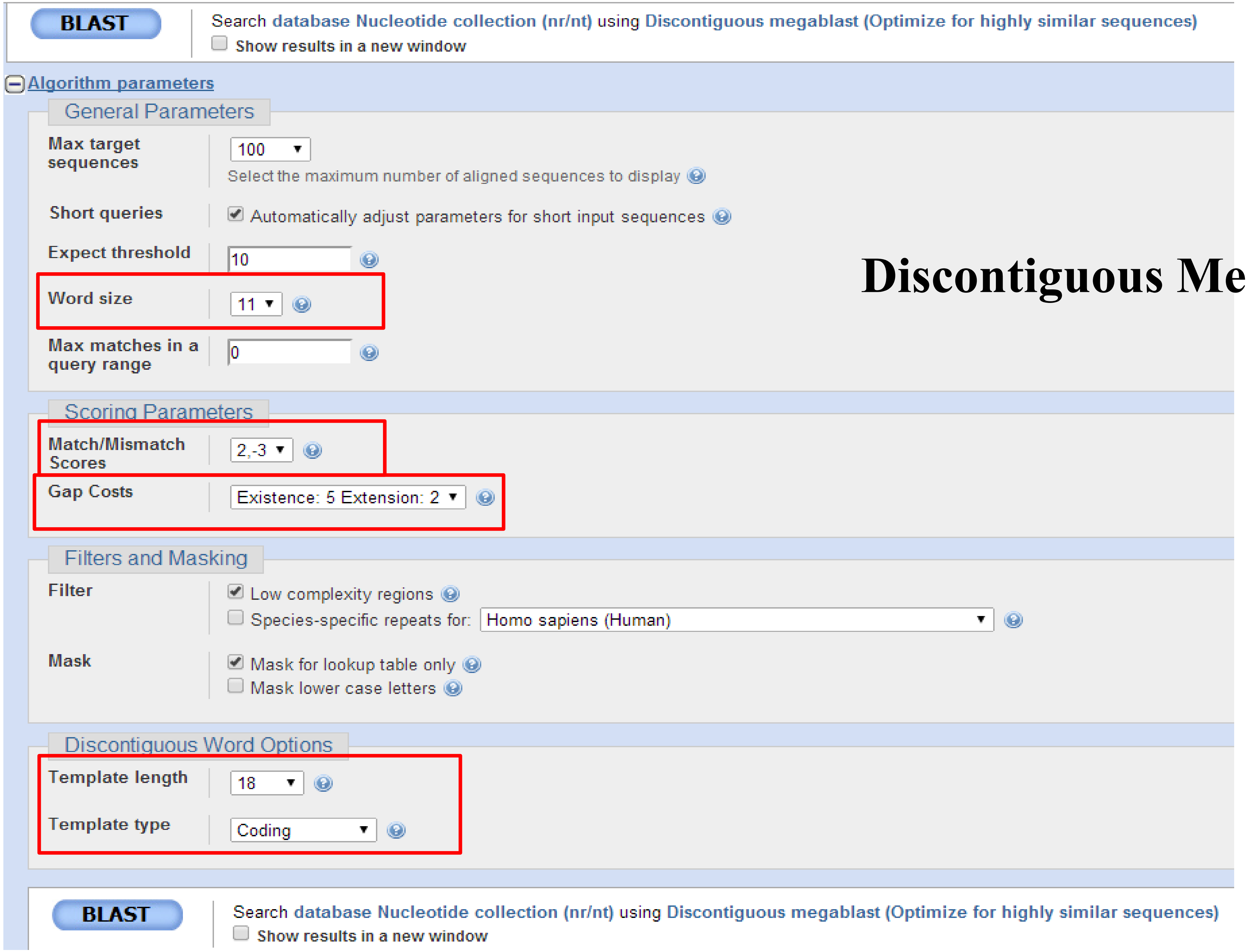
Filter: ☑ Low complexity regions
☐ Species-specific repeats for: Homo sapiens (Human)

Mask: ☑ Mask for lookup table only
☐ Mask lower case letters

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

Discontiguous Megablast



Search database Nucleotide collection (nr/nt) using Discontiguous megablast (Optimize for highly similar sequences)

□ 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: 10

Word size: 11

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter: ☑ Low complexity regions
□ Species-specific repeats for: Homo sapiens (Human)

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

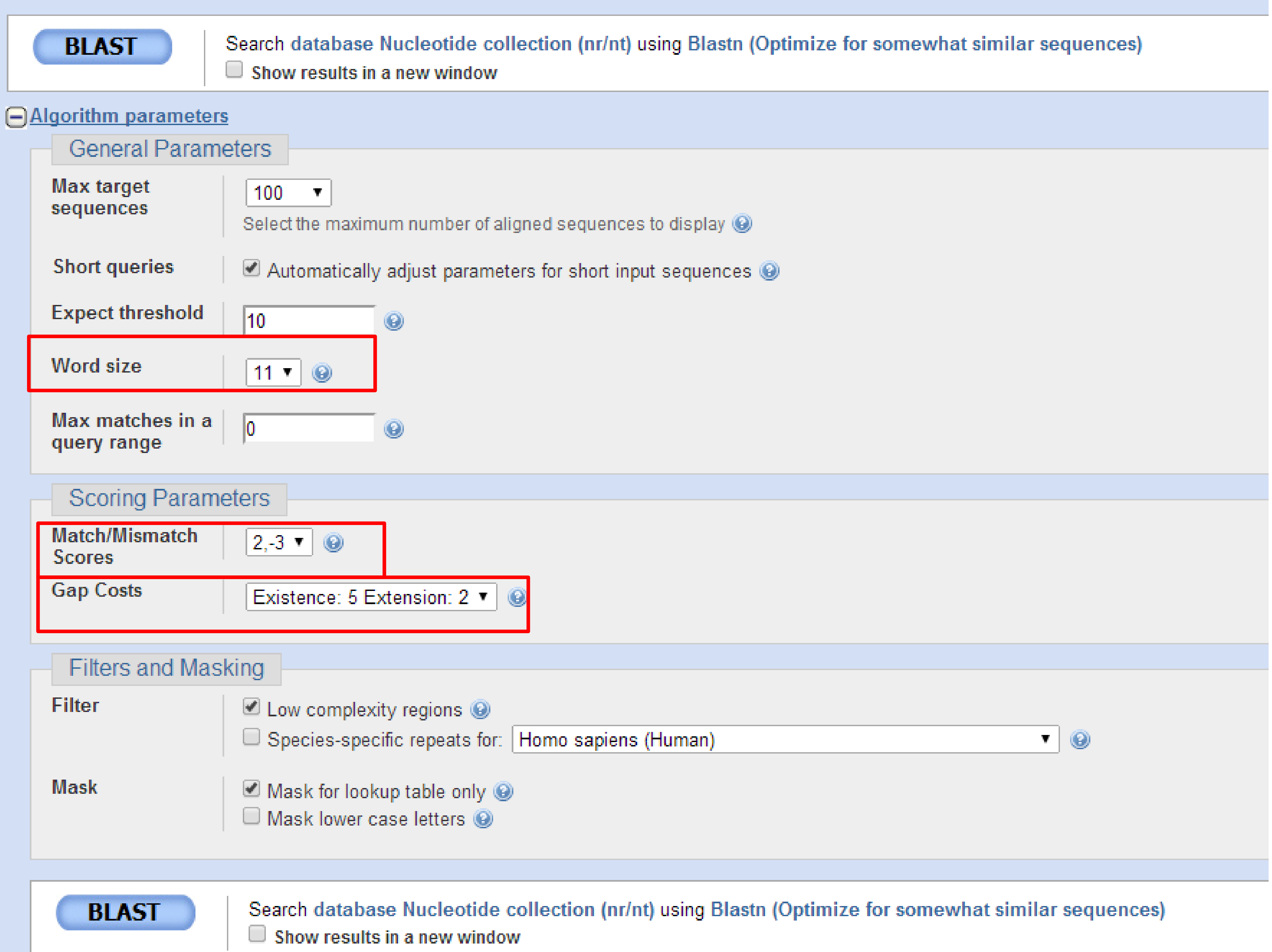
Discontiguous Word Options

Template length: 18

Template type: Coding

Search database Nucleotide collection (nr/nt) using Discontiguous megablast (Optimize for highly similar sequences)

□ Show results in a new window



BLAST | Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
□ 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: 10

Word size: 11 ▼

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 2,-3 ▼

Gap Costs: Existence: 5 Extension: 2 ▼

Filters and Masking

Filter: ☑ Low complexity regions
□ Species-specific repeats for: Homo sapiens (Human) ▼

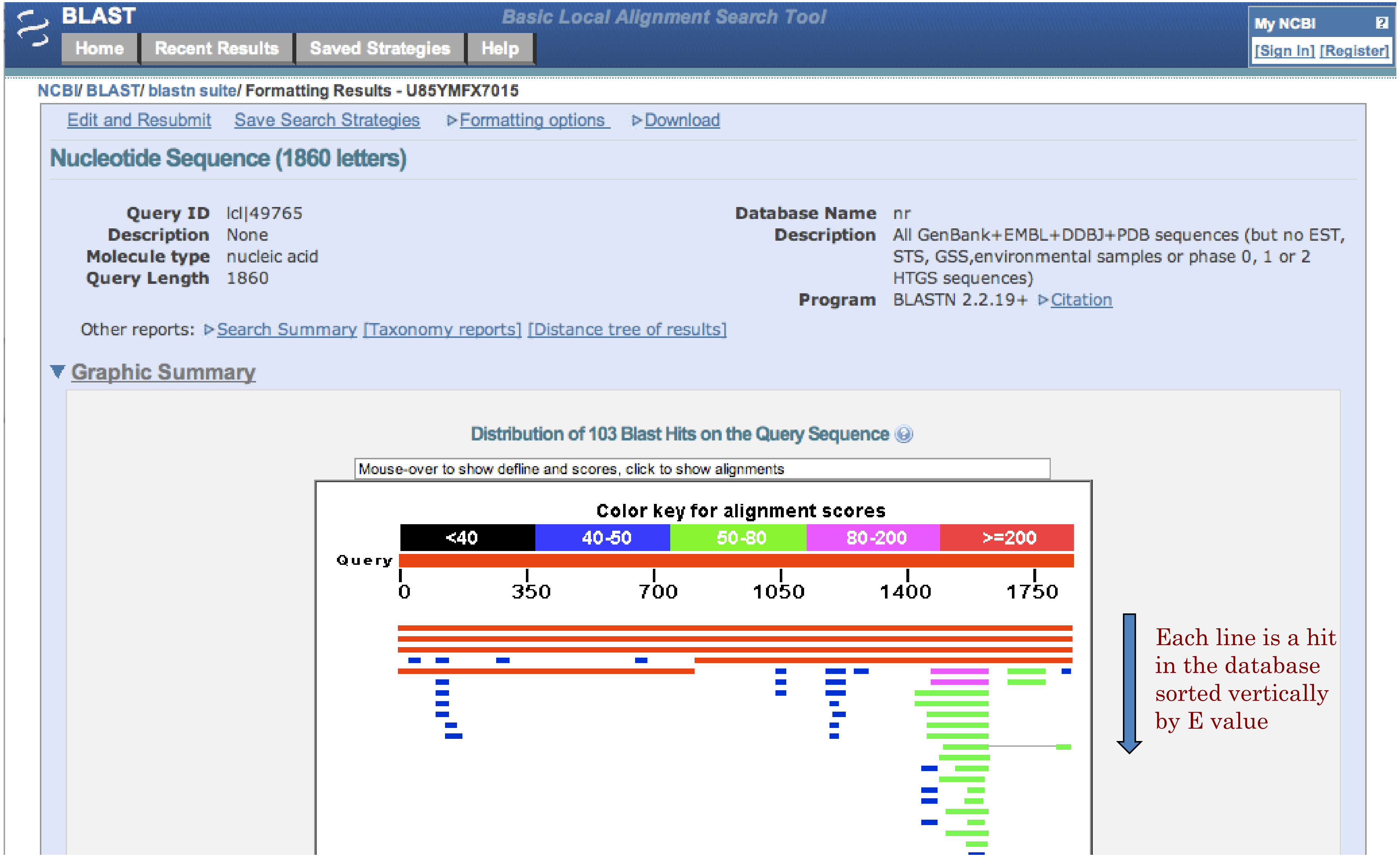
Mask: ☑ Mask for lookup table only
□ Mask lower case letters

BLAST | Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
□ Show results in a new window

Blastn

LIMIT BY ENTREZ QUERY

- protease NOT hiv1[organism]
- 1000:2000[slen]
- Mus musculus[organism] AND biomol_mrna[properties]
- 10000:100000[mlwt]
- all[filter] NOT environmental sample[filter] NOT metagenomes[orgn]



NCBI/ BLAST/ blastn suite/ Formatting Results - U85YMFX7015

Nucleotide Sequence (1860 letters)

Query ID	lcl|49765	**Database Name**	nr
Description	None	**Description**	All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
Molecule type	nucleic acid		
Query Length	1860		
		Program	BLASTN 2.2.19+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results]

▼ Graphic Summary

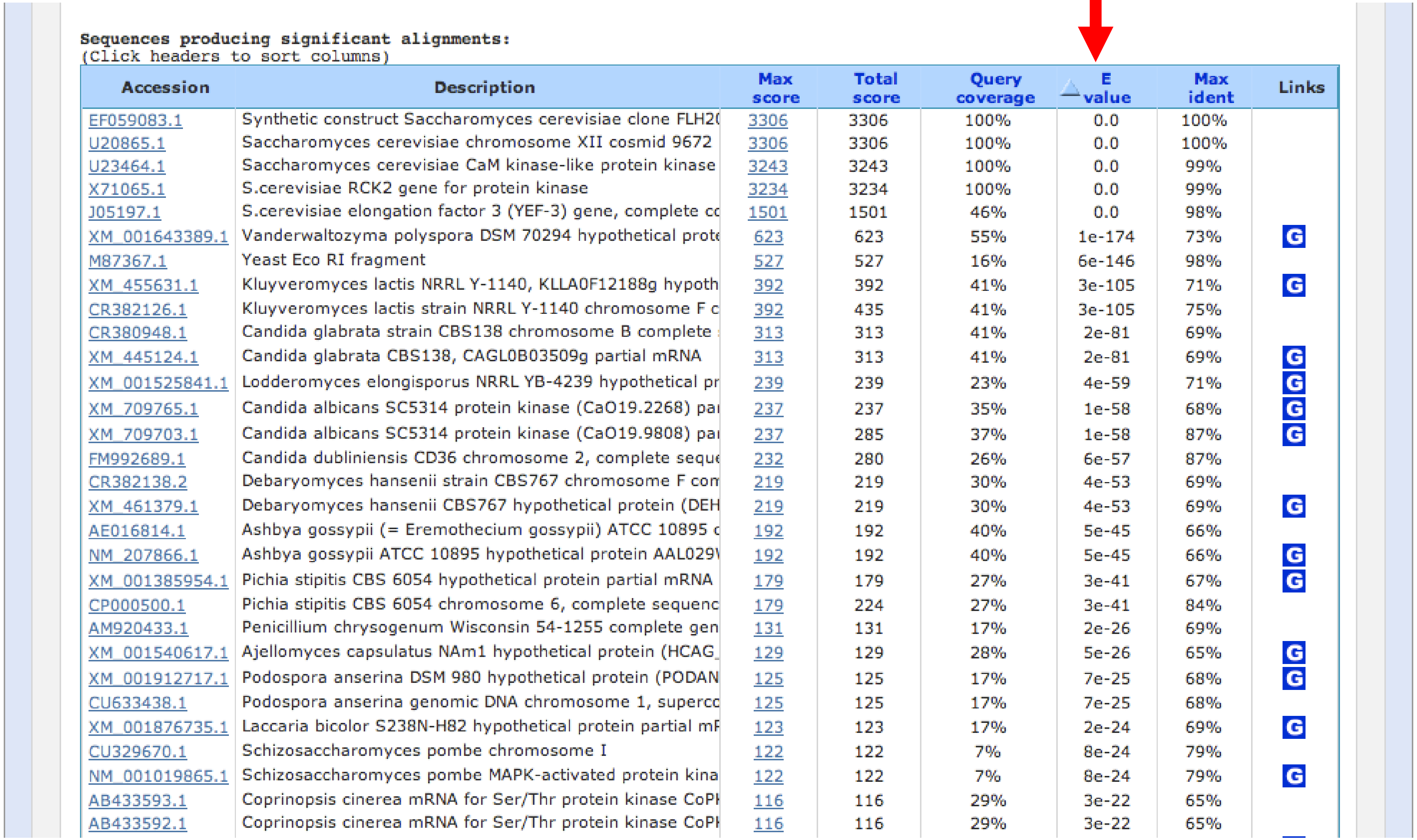
Distribution of 103 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Each line is a hit in the database sorted vertically by E value

Colored rectangles along the X axis show where in the query sequence
a similarity in the database has been found. Color indicates degree of similarity

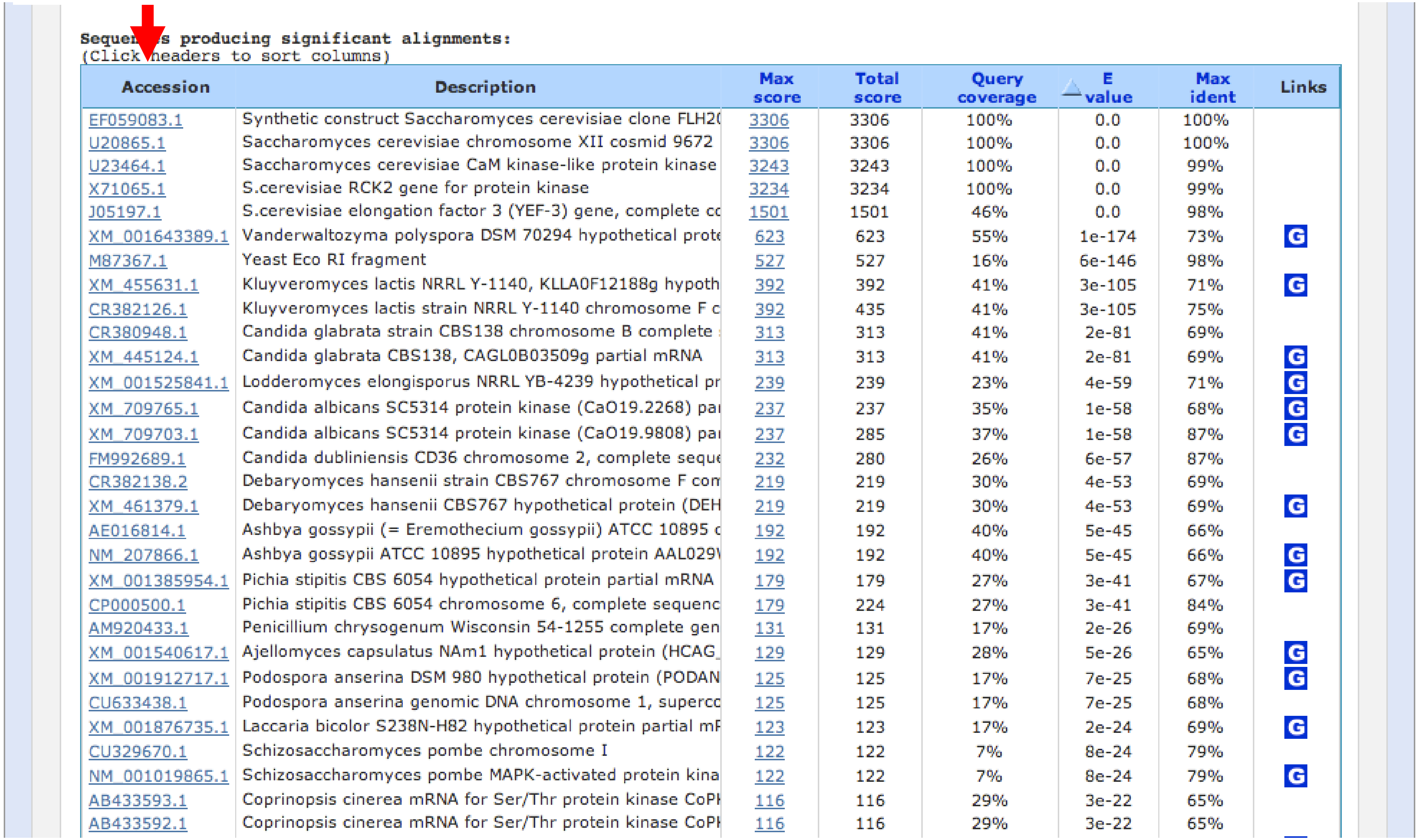
Output sorted by E value



Sequences producing significant alignments:
(Click headers to sort columns)

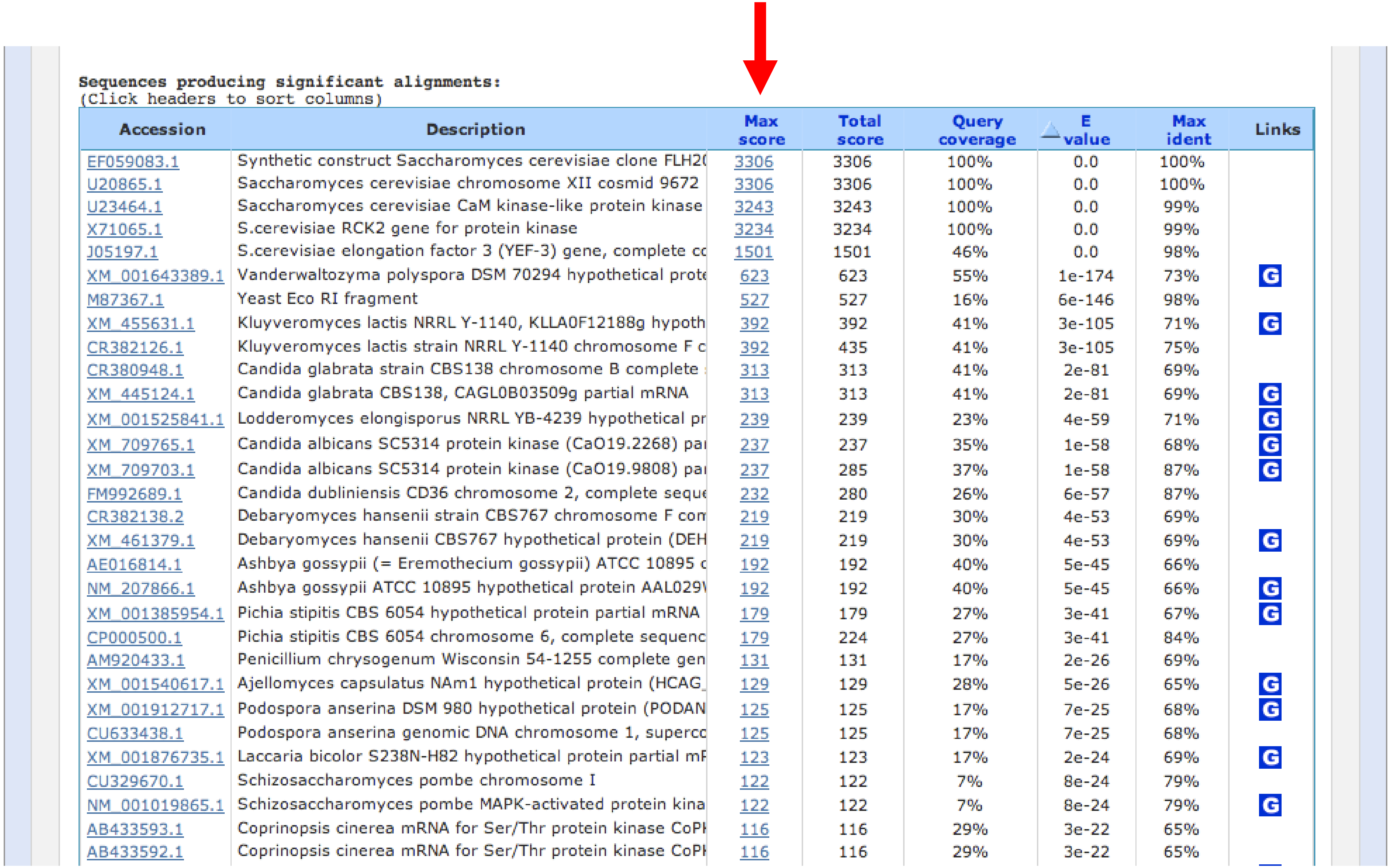
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
EF059083.1	Synthetic construct Saccharomyces cerevisiae clone FLH2	3306	3306	100%	0.0	100%	
U20865.1	Saccharomyces cerevisiae chromosome XII cosmid 9672	3306	3306	100%	0.0	100%	
U23464.1	Saccharomyces cerevisiae CaM kinase-like protein kinase	3243	3243	100%	0.0	99%	
X71065.1	S.cerevisiae RCK2 gene for protein kinase	3234	3234	100%	0.0	99%	
J05197.1	S.cerevisiae elongation factor 3 (YEF-3) gene, complete cc	1501	1501	46%	0.0	98%	
XM_001643389.1	Vanderwaltozyma polyspora DSM 70294 hypothetical prote	623	623	55%	1e-174	73%	G
M87367.1	Yeast Eco RI fragment	527	527	16%	6e-146	98%	
XM_455631.1	Kluyveromyces lactis NRRL Y-1140, KLLA0F12188g hypoth	392	392	41%	3e-105	71%	G
CR382126.1	Kluyveromyces lactis strain NRRL Y-1140 chromosome F c	392	435	41%	3e-105	75%	
CR380948.1	Candida glabrata strain CBS138 chromosome B complete	313	313	41%	2e-81	69%	
XM_445124.1	Candida glabrata CBS138, CAGL0B03509g partial mRNA	313	313	41%	2e-81	69%	G
XM_001525841.1	Lodderomyces elongisporus NRRL YB-4239 hypothetical pr	239	239	23%	4e-59	71%	G
XM_709765.1	Candida albicans SC5314 protein kinase (CaO19.2268) pa	237	237	35%	1e-58	68%	G
XM_709703.1	Candida albicans SC5314 protein kinase (CaO19.9808) pa	237	285	37%	1e-58	87%	G
FM992689.1	Candida dubliniensis CD36 chromosome 2, complete seque	232	280	26%	6e-57	87%	
CR382138.2	Debaryomyces hansenii strain CBS767 chromosome F con	219	219	30%	4e-53	69%	
XM_461379.1	Debaryomyces hansenii CBS767 hypothetical protein (DEH	219	219	30%	4e-53	69%	G
AE016814.1	Ashbya gossypii (= Eremothecium gossypii) ATCC 10895 c	192	192	40%	5e-45	66%	
NM_207866.1	Ashbya gossypii ATCC 10895 hypothetical protein AAL029\	192	192	40%	5e-45	66%	G
XM_001385954.1	Pichia stipitis CBS 6054 hypothetical protein partial mRNA	179	179	27%	3e-41	67%	G
CP000500.1	Pichia stipitis CBS 6054 chromosome 6, complete sequenc	179	224	27%	3e-41	84%	
AM920433.1	Penicillium chrysogenum Wisconsin 54-1255 complete gen	131	131	17%	2e-26	69%	
XM_001540617.1	Ajellomyces capsulatus NAm1 hypothetical protein (HCAG_	129	129	28%	5e-26	65%	G
XM_001912717.1	Podospora anserina DSM 980 hypothetical protein (PODAN	125	125	17%	7e-25	68%	G
CU633438.1	Podospora anserina genomic DNA chromosome 1, superco	125	125	17%	7e-25	68%	
XM_001876735.1	Laccaria bicolor S238N-H82 hypothetical protein partial mF	123	123	17%	2e-24	69%	G
CU329670.1	Schizosaccharomyces pombe chromosome I	122	122	7%	8e-24	79%	
NM_001019865.1	Schizosaccharomyces pombe MAPK-activated protein kina	122	122	7%	8e-24	79%	G
AB433593.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoPk	116	116	29%	3e-22	65%	
AB433592.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoPk	116	116	29%	3e-22	65%	

Link to GenBank file



Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
EF059083.1	Synthetic construct Saccharomyces cerevisiae clone FLH20	3306	3306	100%	0.0	100%	
U20865.1	Saccharomyces cerevisiae chromosome XII cosmid 9672	3306	3306	100%	0.0	100%	
U23464.1	Saccharomyces cerevisiae CaM kinase-like protein kinase	3243	3243	100%	0.0	99%	
X71065.1	S.cerevisiae RCK2 gene for protein kinase	3234	3234	100%	0.0	99%	
J05197.1	S.cerevisiae elongation factor 3 (YEF-3) gene, complete cc	1501	1501	46%	0.0	98%	
XM_001643389.1	Vanderwaltozyma polyspora DSM 70294 hypothetical prote	623	623	55%	1e-174	73%	G
M87367.1	Yeast Eco RI fragment	527	527	16%	6e-146	98%	
XM_455631.1	Kluyveromyces lactis NRRL Y-1140, KLLA0F12188g hypoth	392	392	41%	3e-105	71%	G
CR382126.1	Kluyveromyces lactis strain NRRL Y-1140 chromosome F c	392	435	41%	3e-105	75%	
CR380948.1	Candida glabrata strain CBS138 chromosome B complete	313	313	41%	2e-81	69%	
XM_445124.1	Candida glabrata CBS138, CAGL0B03509g partial mRNA	313	313	41%	2e-81	69%	G
XM_001525841.1	Lodderomyces elongisporus NRRL YB-4239 hypothetical pr	239	239	23%	4e-59	71%	G
XM_709765.1	Candida albicans SC5314 protein kinase (CaO19.2268) pa	237	237	35%	1e-58	68%	G
XM_709703.1	Candida albicans SC5314 protein kinase (CaO19.9808) pa	237	285	37%	1e-58	87%	G
FM992689.1	Candida dubliniensis CD36 chromosome 2, complete seque	232	280	26%	6e-57	87%	
CR382138.2	Debaryomyces hansenii strain CBS767 chromosome F com	219	219	30%	4e-53	69%	
XM_461379.1	Debaryomyces hansenii CBS767 hypothetical protein (DEH	219	219	30%	4e-53	69%	G
AE016814.1	Ashbya gossypii (= Eremothecium gossypii) ATCC 10895 c	192	192	40%	5e-45	66%	
NM_207866.1	Ashbya gossypii ATCC 10895 hypothetical protein AAL029\	192	192	40%	5e-45	66%	G
XM_001385954.1	Pichia stipitis CBS 6054 hypothetical protein partial mRNA	179	179	27%	3e-41	67%	G
CP000500.1	Pichia stipitis CBS 6054 chromosome 6, complete sequenc	179	224	27%	3e-41	84%	
AM920433.1	Penicillium chrysogenum Wisconsin 54-1255 complete gen	131	131	17%	2e-26	69%	
XM_001540617.1	Ajellomyces capsulatus NAm1 hypothetical protein (HCAG_	129	129	28%	5e-26	65%	G
XM_001912717.1	Podospora anserina DSM 980 hypothetical protein (PODAN	125	125	17%	7e-25	68%	G
CU633438.1	Podospora anserina genomic DNA chromosome 1, superco	125	125	17%	7e-25	68%	
XM_001876735.1	Laccaria bicolor S238N-H82 hypothetical protein partial mF	123	123	17%	2e-24	69%	G
CU329670.1	Schizosaccharomyces pombe chromosome I	122	122	7%	8e-24	79%	
NM_001019865.1	Schizosaccharomyces pombe MAPK-activated protein kina	122	122	7%	8e-24	79%	G
AB433593.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoPł	116	116	29%	3e-22	65%	
AB433592.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoPł	116	116	29%	3e-22	65%	

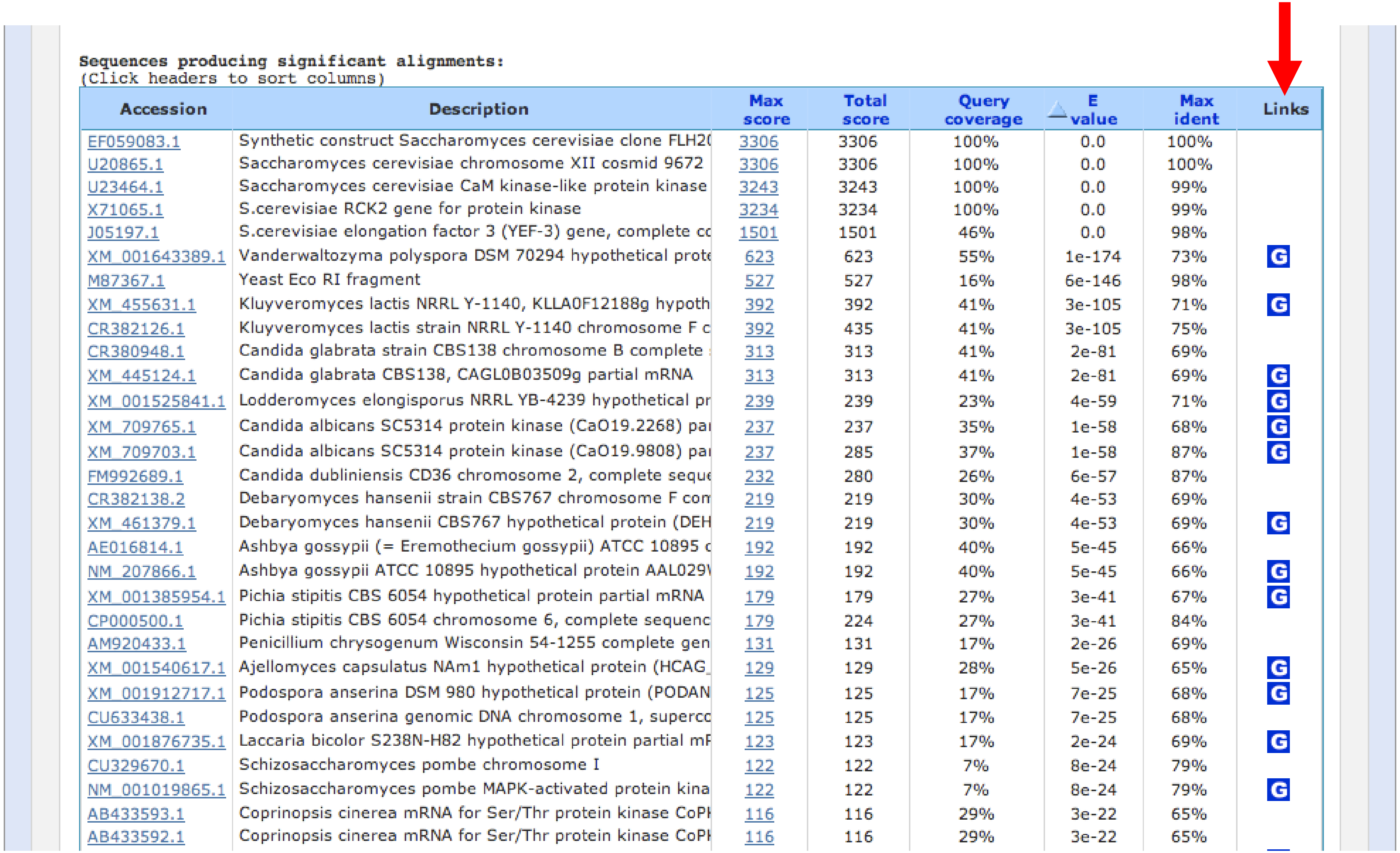


Link to alignment

Sequences producing significant alignments:
(Click headers to sort columns)

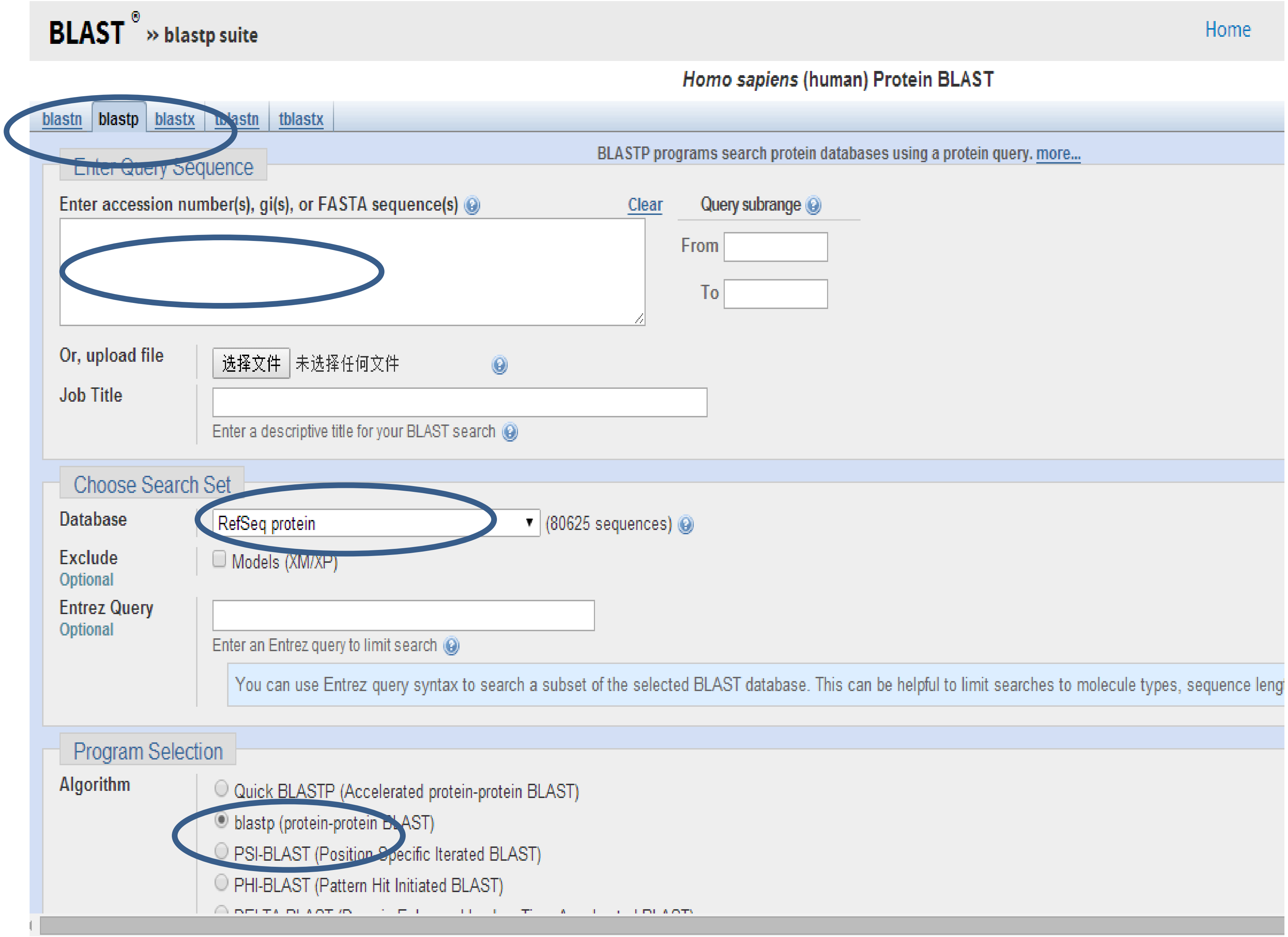
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
EF059083.1	Synthetic construct Saccharomyces cerevisiae clone FLH2	3306	3306	100%	0.0	100%	
U20865.1	Saccharomyces cerevisiae chromosome XII cosmid 9672	3306	3306	100%	0.0	100%	
U23464.1	Saccharomyces cerevisiae CaM kinase-like protein kinase	3243	3243	100%	0.0	99%	
X71065.1	S.cerevisiae RCK2 gene for protein kinase	3234	3234	100%	0.0	99%	
J05197.1	S.cerevisiae elongation factor 3 (YEF-3) gene, complete c	1501	1501	46%	0.0	98%	
XM_001643389.1	Vanderwaltozyma polyspora DSM 70294 hypothetical prot	623	623	55%	1e-174	73%	G
M87367.1	Yeast Eco RI fragment	527	527	16%	6e-146	98%	
XM_455631.1	Kluyveromyces lactis NRRL Y-1140, KLLA0F12188g hypoth	392	392	41%	3e-105	71%	G
CR382126.1	Kluyveromyces lactis strain NRRL Y-1140 chromosome F c	392	435	41%	3e-105	75%	
CR380948.1	Candida glabrata strain CBS138 chromosome B complete	313	313	41%	2e-81	69%	
XM_445124.1	Candida glabrata CBS138, CAGL0B03509g partial mRNA	313	313	41%	2e-81	69%	G
XM_001525841.1	Lodderomyces elongisporus NRRL YB-4239 hypothetical pr	239	239	23%	4e-59	71%	G
XM_709765.1	Candida albicans SC5314 protein kinase (CaO19.2268) pa	237	237	35%	1e-58	68%	G
XM_709703.1	Candida albicans SC5314 protein kinase (CaO19.9808) pa	237	285	37%	1e-58	87%	G
FM992689.1	Candida dubliniensis CD36 chromosome 2, complete seque	232	280	26%	6e-57	87%	
CR382138.2	Debaryomyces hansenii strain CBS767 chromosome F com	219	219	30%	4e-53	69%	
XM_461379.1	Debaryomyces hansenii CBS767 hypothetical protein (DEH	219	219	30%	4e-53	69%	G
AE016814.1	Ashbya gossypii (= Eremothecium gossypii) ATCC 10895 c	192	192	40%	5e-45	66%	
NM_207866.1	Ashbya gossypii ATCC 10895 hypothetical protein AAL029	192	192	40%	5e-45	66%	G
XM_001385954.1	Pichia stipitis CBS 6054 hypothetical protein partial mRNA	179	179	27%	3e-41	67%	G
CP000500.1	Pichia stipitis CBS 6054 chromosome 6, complete sequenc	179	224	27%	3e-41	84%	
AM920433.1	Penicillium chrysogenum Wisconsin 54-1255 complete gen	131	131	17%	2e-26	69%	
XM_001540617.1	Ajellomyces capsulatus NAm1 hypothetical protein (HCAG_	129	129	28%	5e-26	65%	G
XM_001912717.1	Podospora anserina DSM 980 hypothetical protein (PODAN	125	125	17%	7e-25	68%	G
CU633438.1	Podospora anserina genomic DNA chromosome 1, superco	125	125	17%	7e-25	68%	
XM_001876735.1	Laccaria bicolor S238N-H82 hypothetical protein partial m	123	123	17%	2e-24	69%	G
CU329670.1	Schizosaccharomyces pombe chromosome I	122	122	7%	8e-24	79%	
NM_001019865.1	Schizosaccharomyces pombe MAPK-activated protein kina	122	122	7%	8e-24	79%	G
AB433593.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoP	116	116	29%	3e-22	65%	
AB433592.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoP	116	116	29%	3e-22	65%	

Link to Entrez Gene



Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
EF059083.1	Synthetic construct Saccharomyces cerevisiae clone FLH2(	3306	3306	100%	0.0	100%	
U20865.1	Saccharomyces cerevisiae chromosome XII cosmid 9672	3306	3306	100%	0.0	100%	
U23464.1	Saccharomyces cerevisiae CaM kinase-like protein kinase	3243	3243	100%	0.0	99%	
X71065.1	S.cerevisiae RCK2 gene for protein kinase	3234	3234	100%	0.0	99%	
J05197.1	S.cerevisiae elongation factor 3 (YEF-3) gene, complete cc	1501	1501	46%	0.0	98%	
XM_001643389.1	Vanderwaltozyma polyspora DSM 70294 hypothetical prote	623	623	55%	1e-174	73%	G
M87367.1	Yeast Eco RI fragment	527	527	16%	6e-146	98%	
XM_455631.1	Kluyveromyces lactis NRRL Y-1140, KLLA0F12188g hypoth	392	392	41%	3e-105	71%	G
CR382126.1	Kluyveromyces lactis strain NRRL Y-1140 chromosome F c	392	435	41%	3e-105	75%	
CR380948.1	Candida glabrata strain CBS138 chromosome B complete	313	313	41%	2e-81	69%	
XM_445124.1	Candida glabrata CBS138, CAGL0B03509g partial mRNA	313	313	41%	2e-81	69%	G
XM_001525841.1	Lodderomyces elongisporus NRRL YB-4239 hypothetical pr	239	239	23%	4e-59	71%	G
XM_709765.1	Candida albicans SC5314 protein kinase (CaO19.2268) pa	237	237	35%	1e-58	68%	G
XM_709703.1	Candida albicans SC5314 protein kinase (CaO19.9808) pa	237	285	37%	1e-58	87%	G
FM992689.1	Candida dubliniensis CD36 chromosome 2, complete seque	232	280	26%	6e-57	87%	
CR382138.2	Debaryomyces hansenii strain CBS767 chromosome F com	219	219	30%	4e-53	69%	
XM_461379.1	Debaryomyces hansenii CBS767 hypothetical protein (DEH	219	219	30%	4e-53	69%	G
AE016814.1	Ashbya gossypii (= Eremothecium gossypii) ATCC 10895 c	192	192	40%	5e-45	66%	
NM_207866.1	Ashbya gossypii ATCC 10895 hypothetical protein AAL029\	192	192	40%	5e-45	66%	G
XM_001385954.1	Pichia stipitis CBS 6054 hypothetical protein partial mRNA	179	179	27%	3e-41	67%	G
CP000500.1	Pichia stipitis CBS 6054 chromosome 6, complete sequenc	179	224	27%	3e-41	84%	
AM920433.1	Penicillium chrysogenum Wisconsin 54-1255 complete gen	131	131	17%	2e-26	69%	
XM_001540617.1	Ajellomyces capsulatus NAm1 hypothetical protein (HCAG_	129	129	28%	5e-26	65%	G
XM_001912717.1	Podospora anserina DSM 980 hypothetical protein (PODAN	125	125	17%	7e-25	68%	G
CU633438.1	Podospora anserina genomic DNA chromosome 1, superco	125	125	17%	7e-25	68%	
XM_001876735.1	Laccaria bicolor S238N-H82 hypothetical protein partial mF	123	123	17%	2e-24	69%	G
CU329670.1	Schizosaccharomyces pombe chromosome I	122	122	7%	8e-24	79%	
NM_001019865.1	Schizosaccharomyces pombe MAPK-activated protein kina	122	122	7%	8e-24	79%	G
AB433593.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoPI	116	116	29%	3e-22	65%	
AB433592.1	Coprinopsis cinerea mRNA for Ser/Thr protein kinase CoPI	116	116	29%	3e-22	65%	



BLAST® » blastp suite

Home

Homo sapiens (human) 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

Choose Search Set

Database: RefSeq protein (80625 sequences)

Exclude Optional: □ Models (XM/XP)

Entrez Query Optional

Enter an Entrez query to limit search

You can use Entrez query syntax to search a subset of the selected BLAST database. This can be helpful to limit searches to molecule types, sequence leng

Program Selection

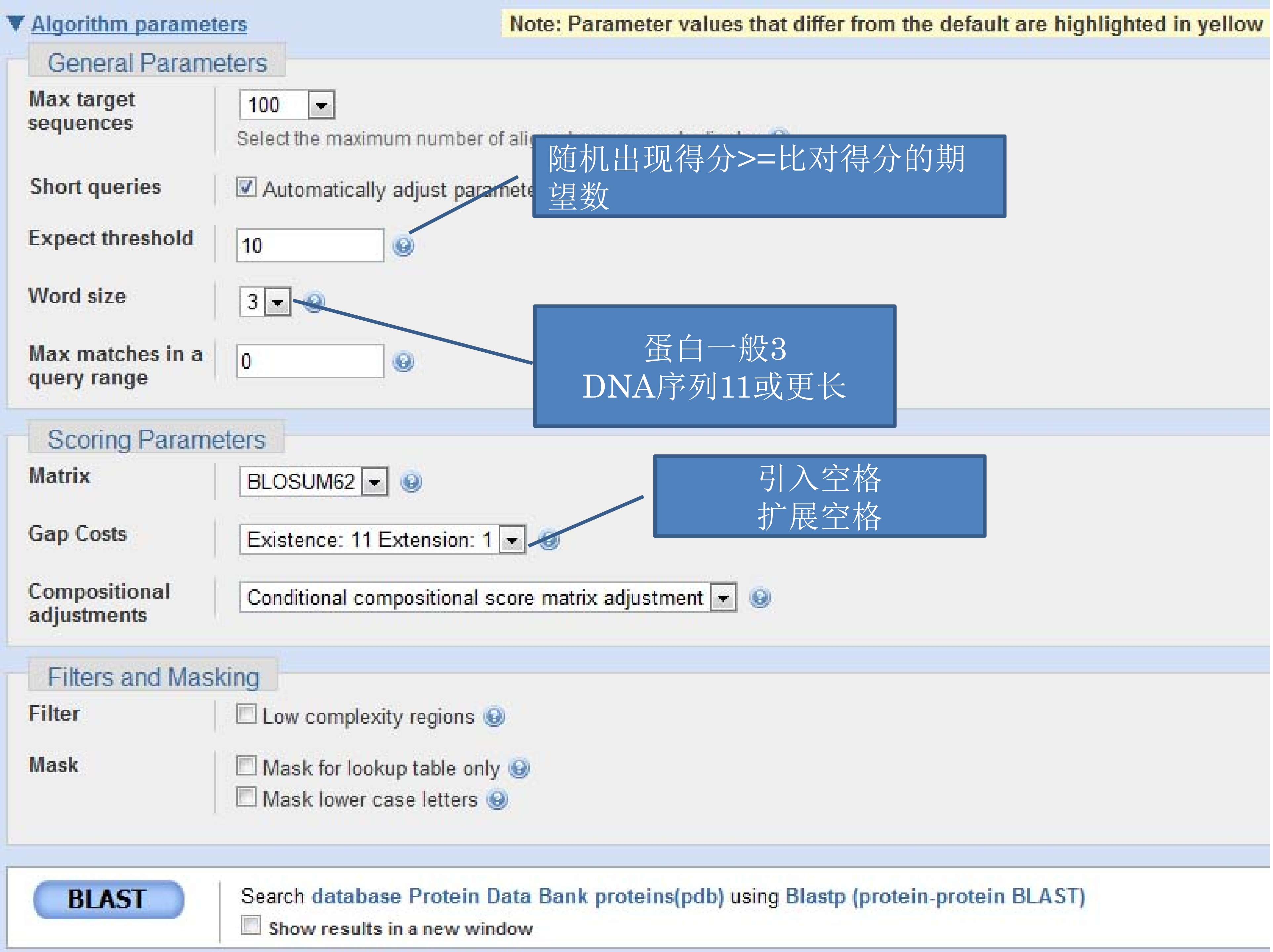
Algorithm

○ Quick BLASTP (Accelerated protein-protein BLAST)

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)



▼ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 100

Select the maximum number of ali

Short queries: ☑ Automatically adjust paramete

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

BLAST

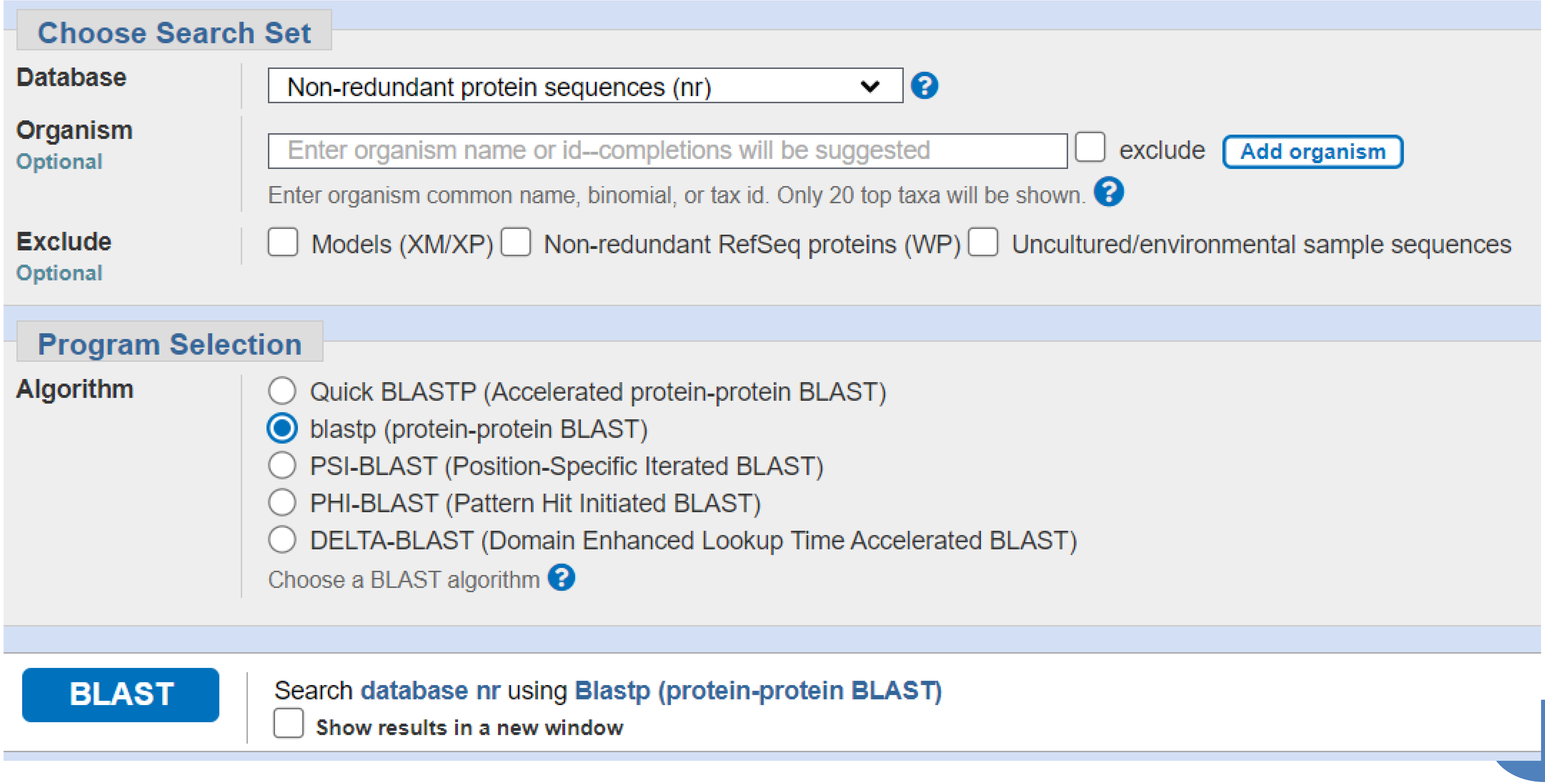
Search database Protein Data Bank proteins(pdb) using Blastp (protein-protein BLAST)

□ Show results in a new window

随机出现得分>=比对得分的期望数

蛋白一般3
DNA序列11或更长

引入空格
扩展空格



Choose Search Set

Database: Non-redundant protein sequences (nr)

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

Filtering and Masking

- Filtering is only applied to the query sequence (or its translation products), not to database sequences.

- Filtering can eliminate statistically significant but biologically uninteresting reports from the blast output (e.g., hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.

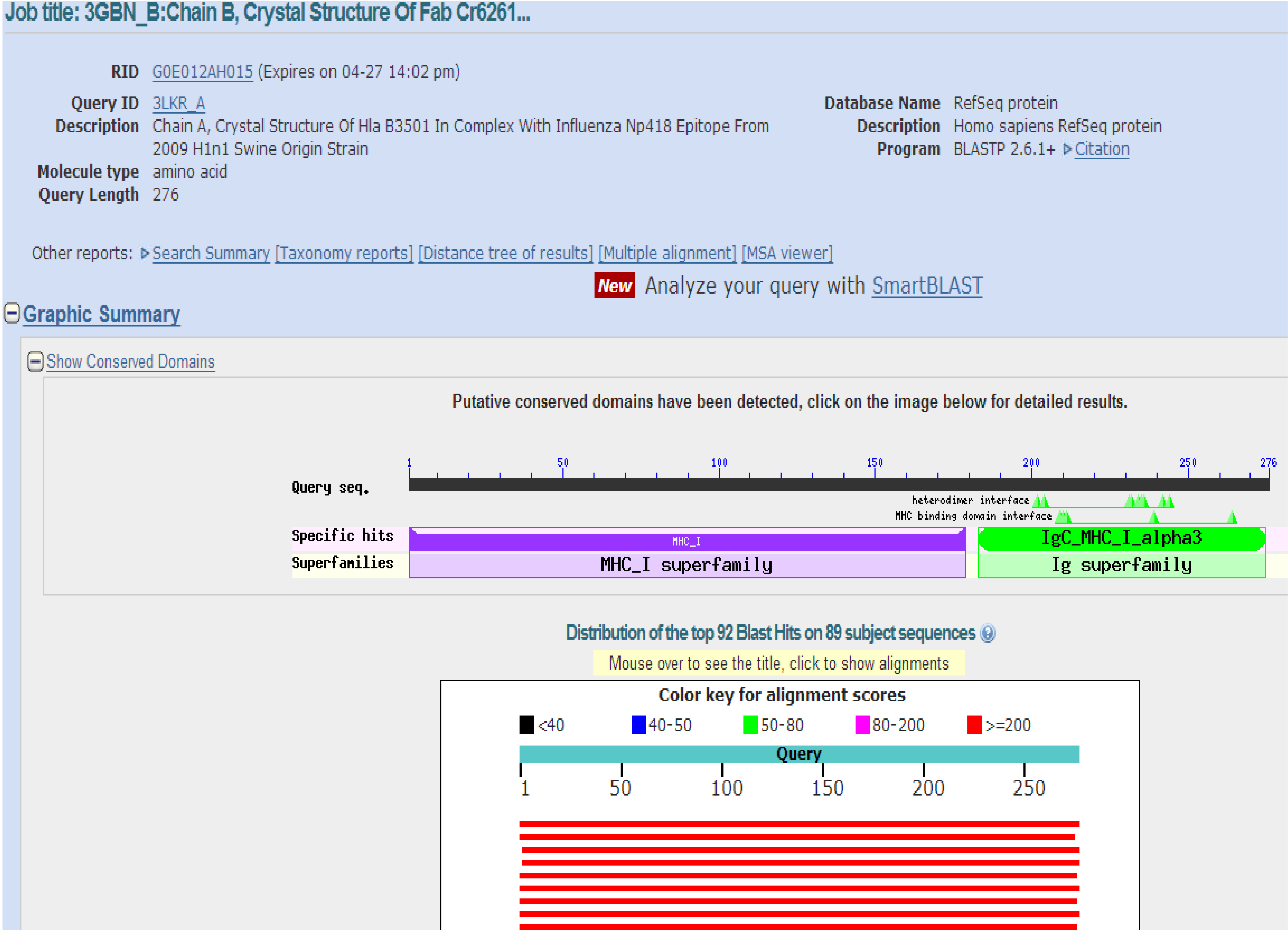
Filtering and Masking

- Filter (Human repeats) This option masks Human repeats (LINE's, SINE's, plus retroviral repeasts) and is useful for human sequences that may contain these repeats.
- Filtering for repeats can increase the speed of a search especially with very long sequences (>100 kb) and against databases which contain large number of repeats (htgs).
- This filter should be checked for genomic queries to prevent potential problems that may arise from the numerous and often spurious matches to those repeat elements.

Filtering and Masking

- Mask for lookup table only
  - Avoids matches to low-complexity sequences or repeats.
  - The BLAST extensions are performed without masking and so they can be extended through low-complexity sequence.

- Mask lower case.
  - Enter a query in the fasta format using upper case letters for the search, using lower case letters for filtering.



Job title: 3GBN_B:Chain B, Crystal Structure Of Fab Cr6261...

RID G0E012AH015 (Expires on 04-27 14:02 pm)

Query ID 3LKR_A
Description Chain A, Crystal Structure Of Hla B3501 In Complex With Influenza Np418 Epitope From 2009 H1n1 Swine Origin Strain
Molecule type amino acid
Query Length 276

Database Name RefSeq protein
Description Homo sapiens RefSeq protein
Program BLASTP 2.6.1+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

⊟Graphic Summary

⊟Show Conserved Domains

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

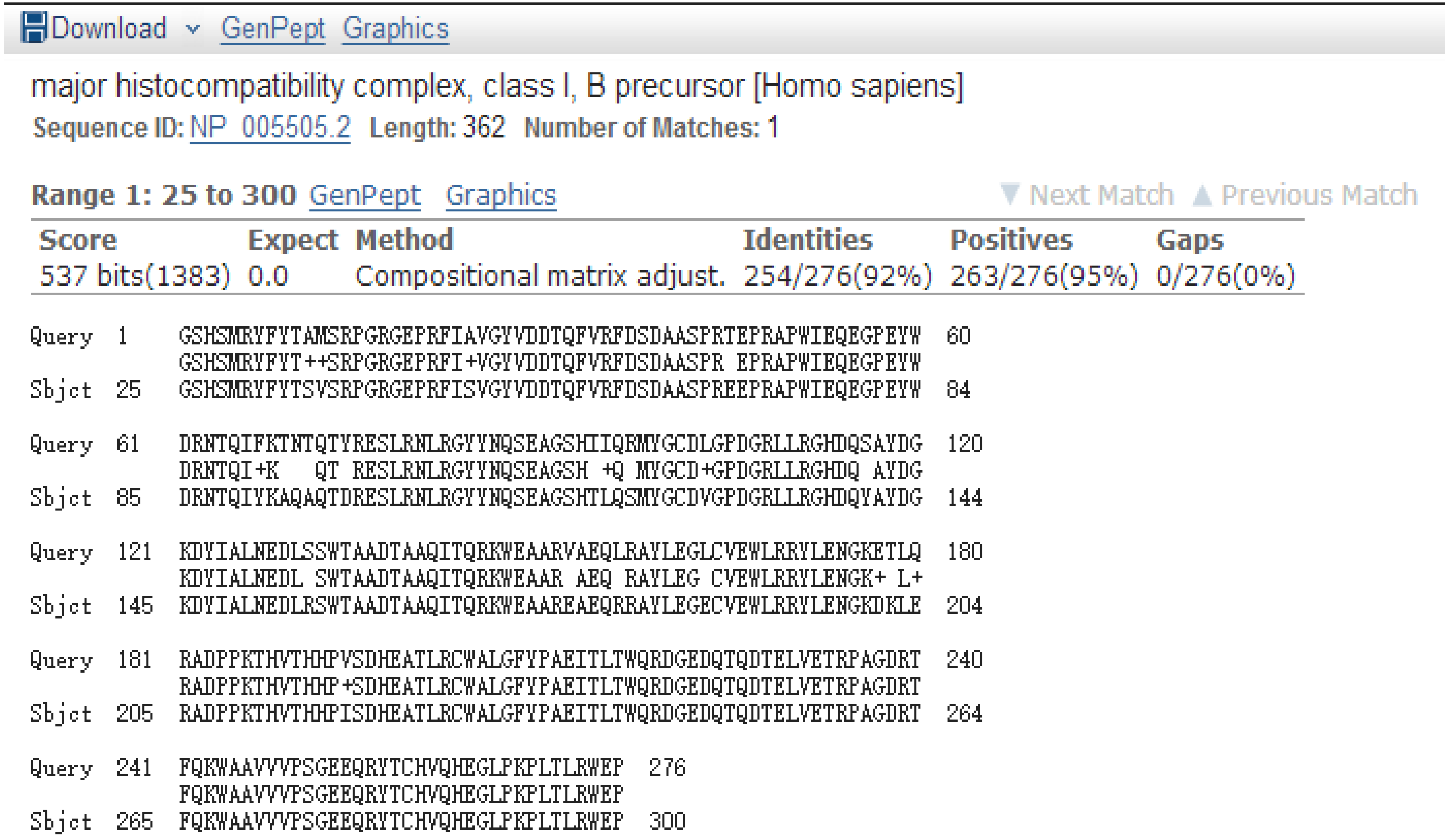
Distribution of the top 92 Blast Hits on 89 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

⬛ <40 🟦 40-50 🟩 50-80 🟪 80-200 🟥 >=200

Blast results page : **Alignments**



Download ⌄ GenPept Graphics

major histocompatibility complex, class I, B precursor [Homo sapiens]

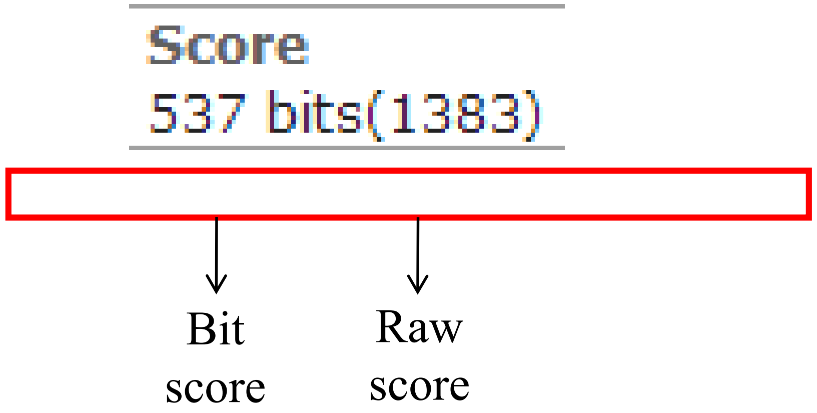
Sequence ID: NP_005505.2 Length: 362 Number of Matches: 1

Range 1: 25 to 300 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
537 bits(1383)	0.0	Compositional matrix adjust.	254/276(92%)	263/276(95%)	0/276(0%)

```
Query  1    GSHSMRYFYTAMSRPGRGEPRFIAVGYVDDTQFVRFDSDAASPRTEPRAPWIEQEGPEYW  60
            GSHSMRYFYT++SRPGRGEPRFI+VGYVDDTQFVRFDSDAASPR EPRAPWIEQEGPEYW
Sbjct  25   GSHSMRYFYTSVSRPGRGEPRFISVGYVDDTQFVRFDSDAASPREEPRAPWIEQEGPEYW  84

Query  61   DRNTQIFKTNTQTYRESLRNLRGYYNQSEAGSHIIQRMYGCDLGPDGRLLRGHDQSAYDG  120
            DRNTQI+K   QT RESLRNLRGYYNQSEAGSH +Q MYGCD+GPDGRLLRGHDQ AYDG
Sbjct  85   DRNTQIYKAQAQTDRESLRNLRGYYNQSEAGSHTLQSMYGCDVGPDGRLLRGHDQYAYDG  144

Query  121  KDYIALNEDLSSWTAADTAAQITQRKWEAARVAEQLRAYLEGLCVEWLRRYLENGKETLQ  180
            KDYIALNEDL SWTAADTAAQITQRKWEAAR AEQ RAYLEG CVEWLRRYLENGK+ L+
Sbjct  145  KDYIALNEDLRSWTAADTAAQITQRKWEAAREAEQRRAYLEGECVEWLRRYLENGKDKLE  204

Query  181  RADPPKTHVTHHPVSDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDRT  240
            RADPPKTHVTHHP+SDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDRT
Sbjct  205  RADPPKTHVTHHPISDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDRT  264

Query  241  FQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWEP  276
            FQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWEP
Sbjct  265  FQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWEP  300
```

ANATOMY OF AN ALIGNMENT



Score
537 bits(1383)

Bit score

Raw score

- **Score** provides alignment score in both normalized (bits)and raw (in the bracket) form
- **E-value** measures the reliability of the score, which refers to the number of hits with a score **equal to or better** than the alignment score that would be "expected" **by chance.**
- E-value: 9e-78 = 9 * 10^{-78}

SCORE & E

- S值表示两序列的相似性，分值越高表明它们之间相似的程度越大。

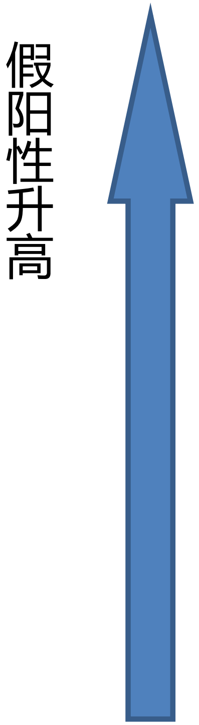
- E值就是S值可靠性的评价。**它表明在随机的情况下，其它序列与目标序列相似度要大于这条显示的序列的可能性。**所以它的分值越低越好。

BLAST搜索的统计学显著性

- 对于两个随机序列s和t，随机观察到比对得分大于等于x的概率：
  - $P\ (s \geq x) = 1 - \exp(\ -Kste^{-\lambda x}\ )$

- BLAST返回比对得分大于阈值S的期望值为：
  - $E = -Kste^{-\lambda S}$
  - 随着S的增加，E值呈指数下降，比对随机发生的可能性就接近于0（阈值越高，序列相似就越可信）
  - 数据库的大小和探测序列的长度影响比对随机发生的可能性 （序列越长，序列相似就越可信）

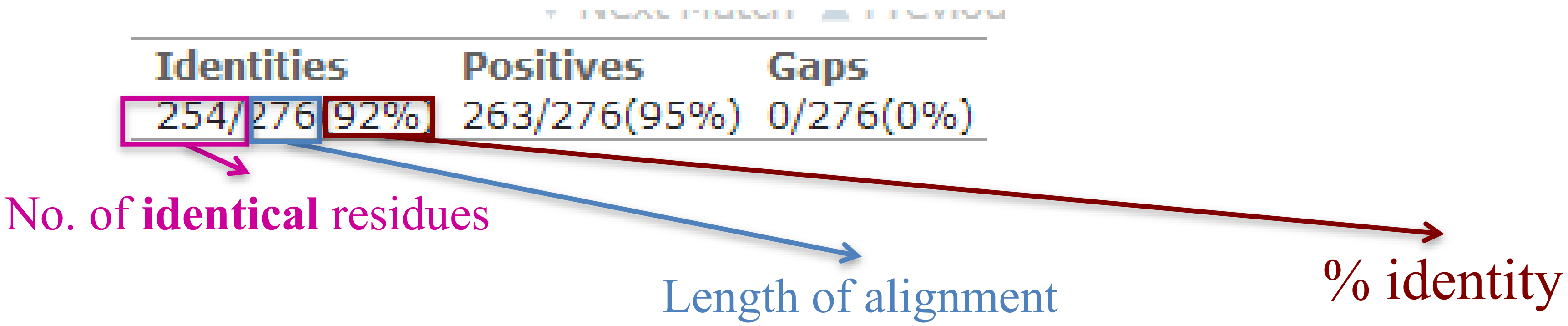
BLAST搜索的统计学显著性



假阳性升高

E	**P**
10	0.99995
5	0.99326
2	0.86466
1	0.63212
0.1	0.09516
0.05	0.04877
0.001	0.0009995
0.0001	0.0001

ANATOMY OF AN ALIGNMENT

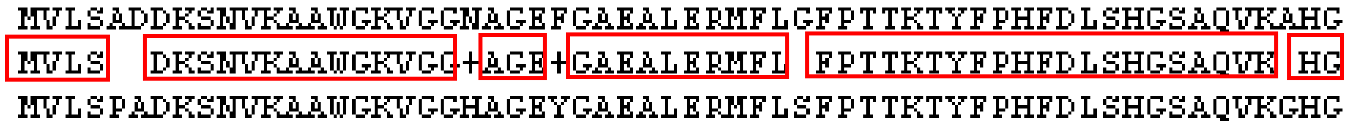


Identities	Positives	Gaps
254/276(92%)	263/276(95%)	0/276(0%)

No. of **identical** residues

Length of alignment

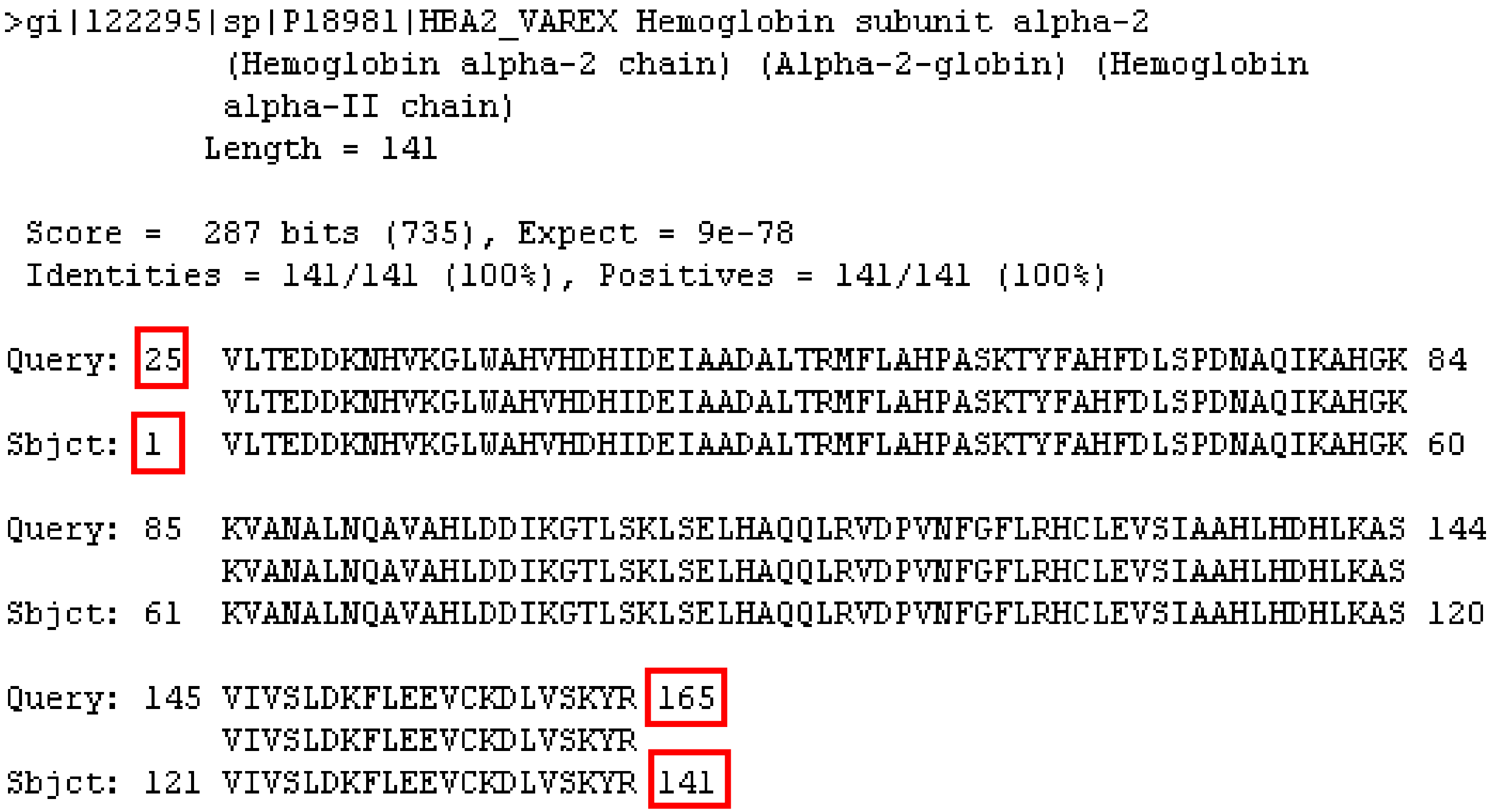
% identity



```
MVLSADDKSNVKAAWGKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
MVLS  DKSNVKAAWGKVGG+AGE+GAEALERMFL FPTTKTYFPHFDLSHGSAQVK HG
MVLSPADKSNVKAAWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
```

- **Identities** (% identity) provides the fraction of number of identical residues (boxed in red above) over the total length of alignment
- **Positives** (% positives) provides the fraction of positive residues (**number of identical residues + number of similar residues with the "+" sign**) over the length of the alignment

ANATOMY OF AN ALIGNMENT



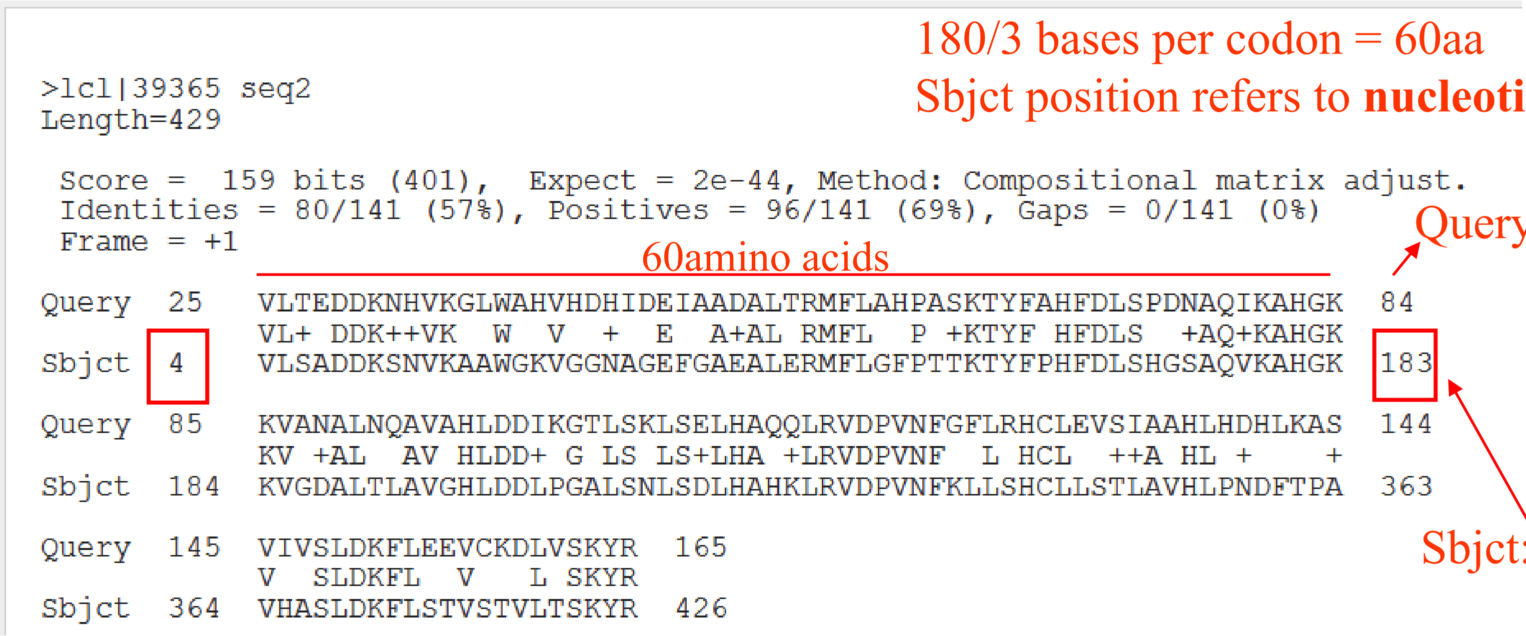
```
>gi|122295|sp|P18981|HBA2_VAREX Hemoglobin subunit alpha-2
           (Hemoglobin alpha-2 chain) (Alpha-2-globin) (Hemoglobin
           alpha-II chain)
          Length = 141

 Score =  287 bits (735), Expect = 9e-78
 Identities = 141/141 (100%), Positives = 141/141 (100%)

Query: 25   VLTEDDKNHVKGLWAHVHDHIDEIAADALTRMFLAHPASKTYFAHFDLSPDNAQIKAHGK 84
            VLTEDDKNHVKGLWAHVHDHIDEIAADALTRMFLAHPASKTYFAHFDLSPDNAQIKAHGK
Sbjct: 1    VLTEDDKNHVKGLWAHVHDHIDEIAADALTRMFLAHPASKTYFAHFDLSPDNAQIKAHGK 60

Query: 85   KVANALNQAVAHLDDIKGTLSKLSELHAQQLRVDPVNFGFLRHCLEVSIAAHLHDHLKAS 144
            KVANALNQAVAHLDDIKGTLSKLSELHAQQLRVDPVNFGFLRHCLEVSIAAHLHDHLKAS
Sbjct: 61   KVANALNQAVAHLDDIKGTLSKLSELHAQQLRVDPVNFGFLRHCLEVSIAAHLHDHLKAS 120

Query: 145 VIVSLDKFLEEVCKDLVSKYR 165
            VIVSLDKFLEEVCKDLVSKYR
Sbjct: 121 VIVSLDKFLEEVCKDLVSKYR 141
```

- Local alignment **start** and **end position** for query and subject sequences

Compare a query protein with a DNA subject sequence

BLAST Program: **tblastn**

- ➢ Translate the subject DNA into amino acids (6-frame)
- ➢ Blast output alignment: protein



```
>lcl|39365 seq2
Length=429

 Score =  159 bits (401),  Expect = 2e-44, Method: Compositional matrix adjust.
 Identities = 80/141 (57%), Positives = 96/141 (69%), Gaps = 0/141 (0%)
 Frame = +1

Query  25   VLTEDDKNHVKGLWAHVHDHIDEIAADALTRMFLAHPASKTYFAHFDLSPDNAQIKAHGK  84
            VL+ DDK++VK  W  V  +  E  A+AL RMFL  P +KTYF HFDLS  +AQ+KAHGK
Sbjct  4    VLSADDKSNVKAAWGKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHGK  183

Query  85   KVANALNQAVAHLDDIKGTLSKLSELHAQQLRVDPVNFGFLRHCLEVSIAAHLHDHLKAS  144
            KV +AL  AV HLDD+ G LS LS+LHA +LRVDPVNF  L HCL  ++A HL +    +
Sbjct  184  KVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPA  363

Query  145  VIVSLDKFLEEVCKDLVSKYR  165
            V  SLDKFL  V   L SKYR
Sbjct  364  VHASLDKFLSTVSTVLTSKYR  426
```

180/3 bases per codon = 60aa

Sbjct position refers to **nucleotide** position

60amino acids

Query: 84-25+1= 60

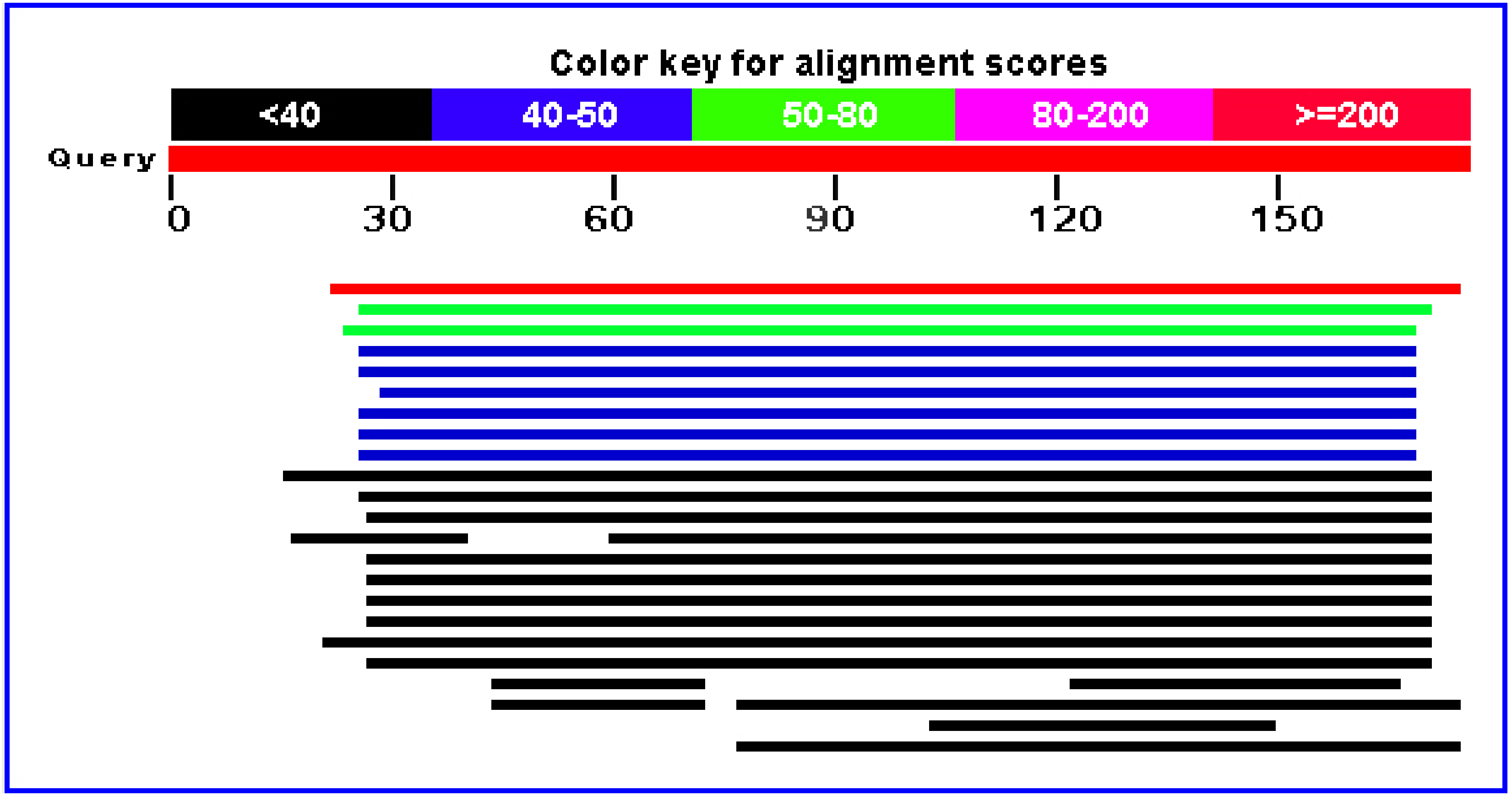
Sbjct: 183-4+1= 180

Insertion of extra nucleotides in the subject DNA will cause **frame shift**, then affect translation & alignment

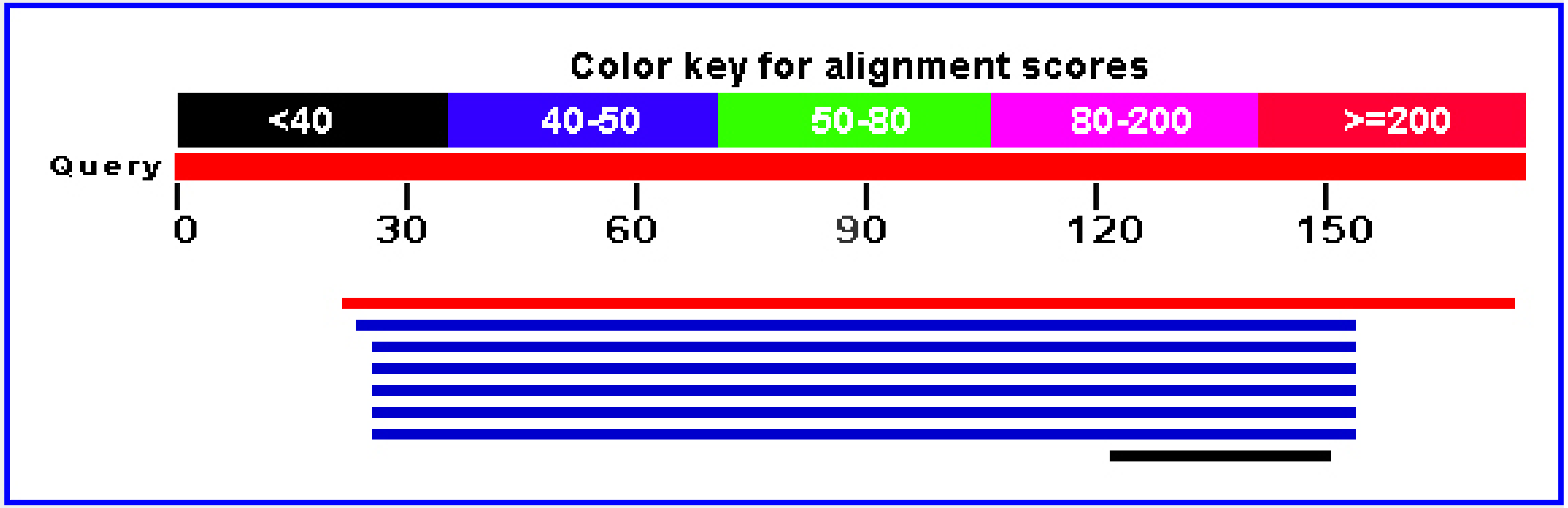
BLAST 应用实例2：脂质运载蛋白

- 改变打分矩阵对结果的影响
  - 脂质运载蛋白： sp|P31025
  - Blosum62
  - PAM30

使用BLOSUM62矩阵搜索



使用PAM30矩阵搜索



Some rules to note when inferring homology

- Similarity can be indicative of homology

- Generally, if two sequences are significantly similar over entire length they are likely homologous

- You cannot measure homology - you cannot say two sequences are 90% homologous; instead, based on the similarity you infer whether they are homologous or not.

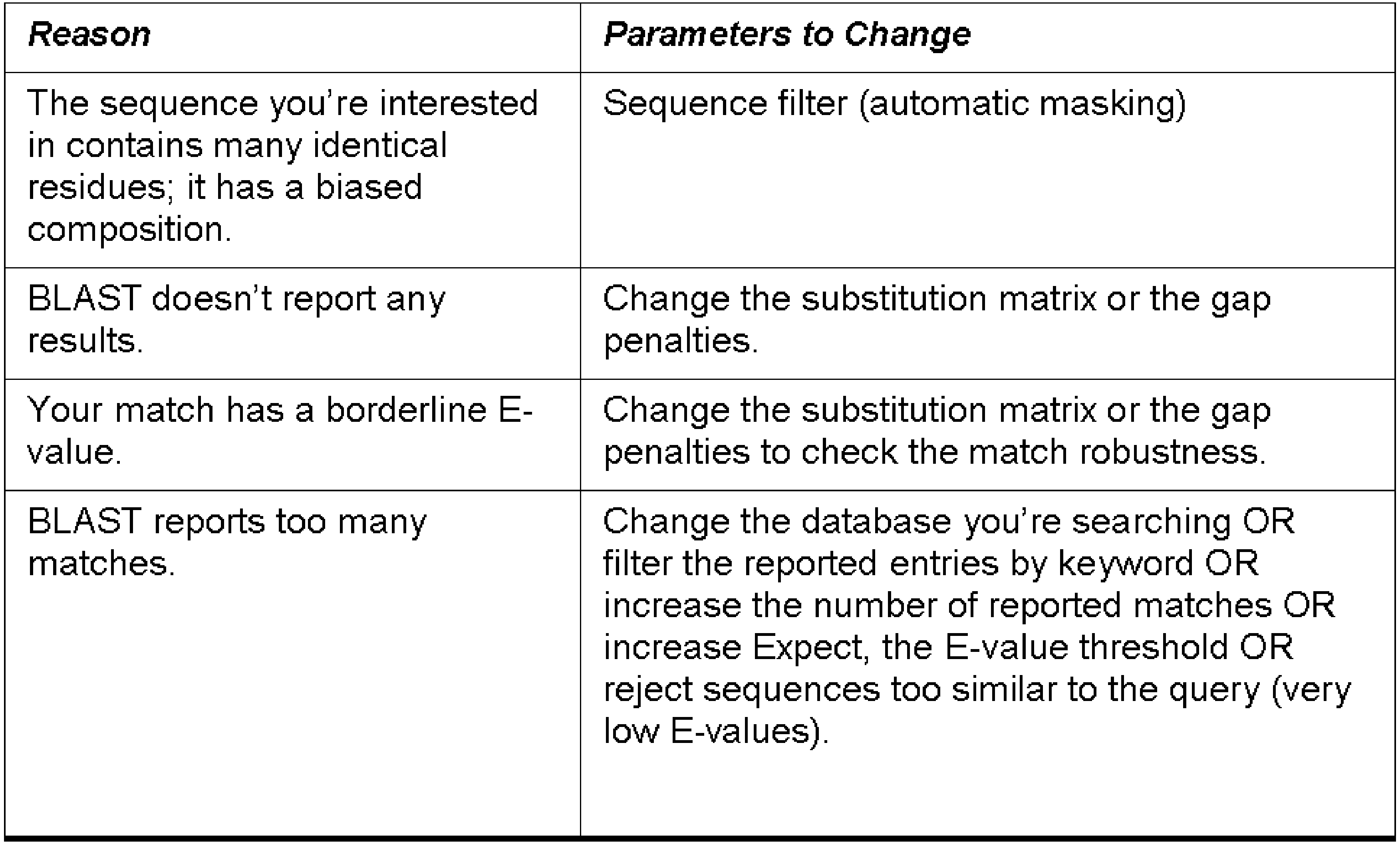
nr/nt, Refseq & Swissprot

- nr/nt database contains ALL known sequences reported at NCBI
- NCBI created two databases called RefSeq_Protein and RefSeq_Genomic, designed to reduce duplication in nr/nt by selecting unique representative sequences for each locus
- Swissprot or Uniprot is a database of highly curated protein sequences , representing an effort to annotate/enrich all the protein sequence records in nr

Blast2Seq

- BLAST 2 Sequences (bl2seq) - aligns two sequences of your choice
  - The sequence you input in the first text box is treated as the **query** sequence
  - The sequence you input in the second text box is treated as a **subject** sequence (“imaginary” database)
  - Hence, even though you are comparing only two sequences, the different **blast flavors** can also be applied for different query tasks
  - Also provides a graphical representation of the alignment

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).