序列比对3

复习 双序列比对及算法

如果两个序列长度相同：

一致度（identity）= （一致字符的个数 / 全局比对长度）×100%

相似度（similarity）= （一致及相似的字符的个数 / 全局比对长度）×100%

序列1: CVHK-LA identity = (4/7)*100% = 57%
序列2: C-HKTIA similarity = ((4+1)/7)*100% = 71%

如果两个序列长度不相同：

一致度（identity）=（一致字符的个数 / 全局比对长度）×100%

相似度（similarity）=（一致及相似的字符的个数 / 全局比对长度）×100%

序列1: CVHKAT   identity = (4/6)*100% = 67%
序列2: CIHK-T   similarity = ((4+1)/6)*100% = 83%

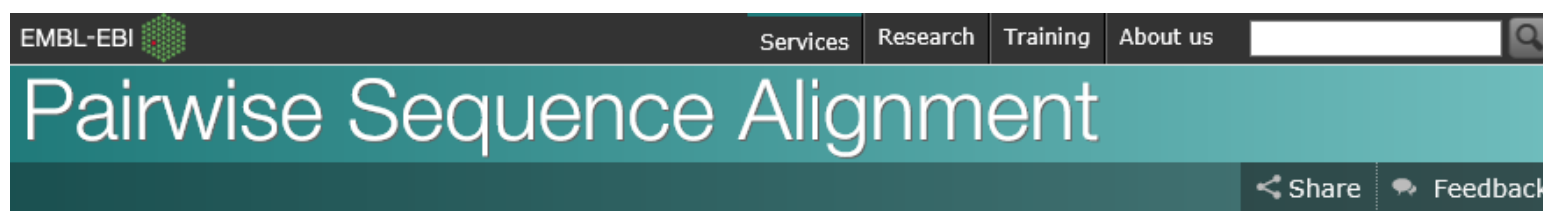无论两个序列长度是否相同，都要先做双序列全局比对，然后根据比对结果及比对长度计算它们的一致度和相似度。

EMBL双序列比对工具：全局比对工具

http://www.ebi.ac.uk/Tools/psa

EMBL-EBI

Services | Research | Training | About us

Pairwise Sequence Alignment

Share   Feedback

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. There are separate forms for protein or nucleotide sequences.

Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Genomic Alignment

Genomic alignment tools concentrate on DNA (or to DNA) alignments while accounting for characteristics present in genomic data.

EMBL双序列比对工具：全局比对工具

Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Needle ❓ (EMBOSS)

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

🔧 Protein 🔧 Nucleotide

Stretcher ❓ (EMBOSS)

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

🔧 Protein 🔧 Nucleotide

Genomic Alignment

Genomic alignment tools concentrate on DNA (or to DNA) alignments while accounting for characteristics present in genomic data.

Wise2DBA ❓

Wise2DBA (DNA Block Aligner) aligns two sequences under the assumption that the sequences share a number of colinear blocks of conservation separated by potentially large and varied lengths of DNA in the two sequences.

🔧 Launch Wise2DBA

GeneWise ❓

GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

🔧 Launch GeneWise

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

EMBL双序列比对工具：全局比对工具

Pairwise Sequence Alignment (PROTEIN)

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.
This is the form for protein sequences. Please go to the nucleotide form if you wish to align DNA or RNA sequences.

STEP 1 - Enter your protein sequences

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

```
>seq1_for_global_alignment
MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEK
EDGSILICLYESYFDPGKSISENIVSFIEKSYKSIFVLSPNFVQNEWCHYEFYFAHHNLF
HENSDHIILILLEPIPFYCIPTRYHKLKALLEKKAYLEWPKDRRKCGLFWANLRAAIN
```

Or, upload a file: [　　　　　　] [浏览...]

AND

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

```
>seq2_for_global_alignment
TTLDDPLGHMPERFDAFICYCPSDIQFVQEMIRQLEQTNYRLKLCVSDRDVLPGTCVWSI
ASELIEKRCRRMVVVSDDYLQSKECDFQTKFALSLSPGAHQKRLIPIKYKAMKKEFPSI
LRFITVCDYTNPCTKSWFWTRLAKALSLP
```
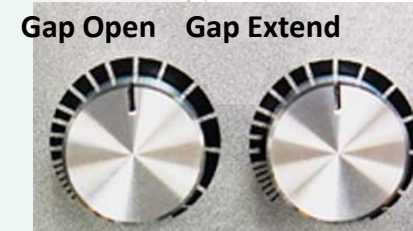
Or, upload a file: [　　　　　　] [浏览...]

global.fasta

STEP 2 - Set your pairwise alignment options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

[More options...] *(Click here, if you want to view or change the default settings.)*

EMBL双序列比对工具：全局比对工具

BLOSUM-80	BLOSUM-62	BLOSUM-45
PAM-1	PAM-120	PAM-250

亲缘关系较近的
序列之间的比对 ←——————————→ 亲缘关系较远的
序列之间的比对

Enter or paste your second protein sequenc

```
>seq2_for_global_alignment
TTLDDPLGHMPERFDAFICYCPSDIQFVQE
ASELIEKRCRRMVVVSDDYLQSKECDFQ
LRFITVCDYTNPCTKSWFWTRLAKALSLP
```

Or, upload a file: [   ] [浏览...]

STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	OUTPUT FORMAT
BLOSUM62 ▾	10 ▾	0.5 ▾	pair ▾

END GAP PENALTY	END GAP OPEN	END GAP EXTEND
false ▾	10 ▾	0.5 ▾

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

[Submit]

EMBL双序列比对工具：全局比对工具

```
#=======================================
# Aligned_sequences: 2
# 1: seq1_for_global_alignment
# 2: seq2_for_global_alignment
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 196
# Identity:      40/196 (20.4%)
# Similarity:    69/196 (35.2%)
# Gaps:          65/196 (33.2%)
# Score: 88.5
#=======================================

seq1_for_glob     1 MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNV--RFHAFISYSEHDSL     48
                      ||.:....::  ||.|||.|...|..
seq2_for_glob     1 ------------------------TTLDDPLGHMPERFDAFICYCPSDIQ     26

seq1_for_glob    49 WVKNELIPNLEKEDGSILICLYESYFDPGKSISENIVSFIEK-SYKSIFV     97
                     :|:  |:|..||.:..:|....:|.:......|||  ..:.:.|
seq2_for_glob    27 FVQ-EMIRQLEQTNYRLKLCVSDRDVLPGTCVWSIASELIEKRCRRMVVV     75

seq1_for_glob    98 LSPNFVQNEWCHYEFYFAHHNLFHENSDHIILILLEPIPFY--CIPTRYH    145
                     :|.::::|::.|.::..||          :.|.|.... .||.:|.
seq2_for_glob    76 VSDDYLQSKECDFQTKFA--------------LSLSPGAHQKRLIPIKYK    111

seq1_for_glob   146 KLKALLEKKAYLEWPKDRR---------KC--GLFWANLRAAIN--      178
                     .:|.        |:|...|        .|  ..||..|..|::
seq2_for_glob   112 AMKK--------EFPSILRFITVCDYTNPCTKSWFWTRLAKALSLP      149

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

	上下一致
:	上下相似
.	上下不相似
空格	字母对空位

Gap的类型及分值设置

Gap的类型及分值设置

```
#=======================================
# Aligned_sequences: 2
# 1: seq1_for_global_alignment
# 2: seq2_for_global_alignment
# Matrix: EBLOSUM62
# Gap_penalty: 1.0
# Extend_penalty: 5.0

# Length: 201
# Identity:      46/201 (22.9%)
# Similarity:    78/201 (38.8%)
# Gaps:          75/201 (37.3%)
# Score: 210.0
#=======================================

seq1_for_glob     1 MHHHHHHSSGVDLGTENLYFQSMKTT-QEQLKR--NVRFHAFISYSEHDS     47
                                      ||  .:.|..  . ||.|||.|...|
seq2_for_glob     1 ------------------------TTLDDPLGHMPE-RFDAFICYCPSD-     24

seq1_for_glob    48 L-WVKNELIPNLEKEDGSI-L-ICLYE-SYFDPGKS-IS-ENIVS-FIEK     90
                     : :|: |:|..|| :. :. | :|:.: ... || : :   :|.| .|||
seq2_for_glob    25 IQFVQ-EMIRQLE-QT-NYRLKLCVSDRDVL-PG-TCV-W-SIASELIEK     67

seq1_for_glob    91 -SYKSIFVLSPNFVQN-EWCHYEFYFAH-HNLFHENSDHI-ILI-LLEPI    135
                     ..:.:.|:|.:::|: | |.::..||   :| ...: |   .  |: ||
seq2_for_glob    68 RCRRMVVVVSDDYLQSKE-CDFQTKFA-L-SL-SPGA-H-Q-K-RLI-PI    108

seq1_for_glob   136 PFY-CIPTR-YHK-LKALLEKKA-YLEWPK-DRRKCGLFWANLRA-AIN-    178
                     . | .: .: :.. |: .: ... |.. | . | ..||..| | |::
seq2_for_glob   109 K-YKAM-KKEFPSILR-FI-TVCDYTN-P-C-T-K-SWFWTRL-AKALSL    148

seq1_for_glob   179 -    178

seq2_for_glob   149 P    149

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

gap open gap extend

分散的空位

Gap的类型及分值设置

```
#=======================================
# Aligned_sequences: 2
# 1: seq1_for_global_alignment
# 2: seq2_for_global_alignment
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 196
# Identity:     40/196 (20.4%)
# Similarity:   69/196 (35.2%)
# Gaps:         65/196 (33.2%)
# Score: 88.5
#=======================================

seq1_for_glob      1 MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNV--RFHAFISYSEHDSL    48
                                      ||.:....:: ||.|||.|....|..
seq2_for_glob      1 -----------------------TTLDDPLGHMPERFDAFICYCPSDIQ    26

seq1_for_glob     49 WVKNELIPNLEKEDGSILICLYESYFDPGKSISENIVSFIEK-SYKSIFV    97
                     :|: |:|..||:.:..:.:|:.:....||.:....||| ..:.:.|
seq2_for_glob     27 FVQ-EMIRQLEQTNYRLKLCVSDRDVLPGTCVWSIASELIEKRCRRMVVV    75

seq1_for_glob     98 LSPNFVQNEWCHYEFYFAHHNLFHENSDHIILILLEPIPFY--CIPTRYH   145
                     :|.::::|::.|.::..||              :.|.|.... .||.:|.
seq2_for_glob     76 VSDDYLQSKECDFQTKFA--------------LSLSPGAHQKRLIPIKYK   111

seq1_for_glob    146 KLKALLEKKAYLEWPKDRR---------KC--GLFWANLRAAIN--      178
                     .:|.        |:|...|         .|  ..||..|..|::
seq2_for_glob    112 AMKK--------EFPSILRFITVCDYTNPCTKSWFWTRLAKALSLP      149

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

gap open gap extend

集中的空位

Gap的类型及分值设置

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

>seq2_for_global_alignment
SEQUENCETTLDDPLGHMPERFDAFICYCPSDIQFVQEMIRQLEQTNYRLKLCVSDRDVL
PGTCVWSIASELIEKRCRRMVVVSDDYLQSKECDFQTKFALSLSPGAHQKRLIPIKYKA
MKKEFPSILRFITVCDYTNPCTKSWFWTRLAKALSLP

Or, upload a file: 浏览...

STEP 2 - Set your pairwise alignment options

MATRIX: BLOSUM62

GAP OPEN: 10
100
50
25
20
15
10
5
1

GAP EXTEND: 0.5
0.0005
0.001
0.05
0.1
0.2
0.4
0.5
0.6
0.8
1.0
5.0
10.0

OUTPUT FORMAT: pair

END GAP PENALTY: false
END GAP OPEN: 10

STEP 3 - Submit your job

Be notified by email (Tick this ... email when the results are available)

If you plan to use these servi... urse please contact us.

调整 gap open 和 gap extend以达到期望的比对结果

Gap Open   Gap Extend

11

EMBL双序列比对工具：局部比对工具

Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Water ❓ (EMBOSS)

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

🔧 **Protein** 🔧 Nucleotide

Matcher ❓ (EMBOSS)

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

🔧 Protein 🔧 Nucleotide

LALIGN ❓

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

🔧 Protein 🔧 Nucleotide

PromoterWise ❓

PromoterWise compares two DNA sequences allowing for inversions and translocations, ideal for promoters.

🔧 Launch PromoterWise

EMBL双序列比对工具：局部比对工具

STEP 1 - Enter your protein sequences

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

>Seq1
MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEK
EDGSILICLYESYFDPGKSISENIVSFIEKSYKSIFVLSPNFVQNEWCHYEFYFAHHNLF
HENSDHIILILLEPIPFYCIPTRYHKLKALLEKKAYLEWPKDRRKCGLFWANLRAAIN

Or, upload a file: [浏览...]

AND

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

>Seq2
GTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKEDGSILICLYESY
FDPGKEWCHYEFYFAHHNLFHENSDHIILILLEPIPFYCIPTRAAAAAAAAAAA

Or, upload a file: [浏览...]

STEP 2 - Set your pairwise alignment options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

[More options...] *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

[Submit]

local.fasta

EMBL双序列比对工具：局部比对工具

黑色是相似的部分，红色是不相似的部分

>Seq1
MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKED
GSILICLYESYFDPGKSISENIVSFIEKSYKSIFVLSPNFVQNEWCHYEFYFAHHNLFHENS
DHIILILLEPIPFYCIPTRYHKLKALLEKKAYLEWPKDRRKCGLFWANLRAAIN
>Seq2
GTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKEDGSILICLYESYFD
PGKEWCHYEFYFAHHNLFHENSDHIILILLEPIPFYCIPTRAAAAAAAAAAA

局部比对结果：

```
Seq1       22 GTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKEDG       71
              |||||||||||||||||||||||||||||||||||||||||||||||||
Seq2        1 GTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKEDG       50

Seq1       72 SILICLYESYFDPGKSISENIVSFIEKSYKSIFVLSPNFVQNEWCHYEFY      121
              ||||||||||||||                             |||||||||
Seq2       51 SILICLYESYFDPGK-------------------------EWCHYEFY       73

Seq1      122 FAHHNLFHENSDHIILILLEPIPFYCIPTR      151
              |||||||||||||||||||||||||||||||
Seq2       74 FAHHNLFHENSDHIILILLEPIPFYCIPTR      103
```

EMBL双序列比对工具：局部比对工具

全局比对与局部比对比较

局部比对

```
#=======================================
# Aligned_sequences: 2
# 1: Seq1
# 2: Seq2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 130
# Identity:      103/130 (79.2%)
# Similarity:    103/130 (79.2%)
# Gaps:          27/130 (20.8%)
# Score: 551.0
#=======================================
```

```
Seq1   14 GTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKEDG   63
          |||||||||||||||||||||||||||||||||||||||||||||||||
Seq2    1 GTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKEDG   50

Seq1   64 SILICLYESYFDPGKSISENIVSFIEKSYKSIFVLSPNFVQNEWCHYEFY  113
          ||||||||||||||                            ||||||||
Seq2   51 SILICLYESYFDPGK----------------------------EWCHYEFY   73

Seq1  114 FAHHNLFHENSDHIILILLEPIPFYCIPTR       143
          |||||||||||||||||||||||||||||
Seq2   74 FAHHNLFHENSDHIILILLEPIPFYCIPTR       103
```

全局比对

```
#=======================================
# Aligned_sequences: 2
# 1: Seq1
# 2: Seq2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 178
# Identity:      104/178 (58.4%)
# Similarity:    104/178 (58.4%)
# Gaps:          64/178 (36.0%)
# Score: 543.0
#=======================================
```

```
Seq1    1 MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWV   50
                       |||||||||||||||||||||||||||||||||||||
Seq2    1 -------------GTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWV   37

Seq1   51 KNELIPNLEKEDGSILICLYESYFDPGKSISENIVSFIEKSYKSIFVLSP  100
          |||||||||||||||||||||||||||
Seq2   38 KNELIPNLEKEDGSILICLYESYFDPGK---------------------   65

Seq1  101 NFVQNEWCHYEFYFAHHNLFHENSDHIILILLEPIPFYCIPTRYHKLKAL  150
                |||||||||||||||||||||||||||||||||||||||.....|.
Seq2   66 -----EWCHYEFYFAHHNLFHENSDHIILILLEPIPFYCIPTRAAAAAAA  110

Seq1  151 LEKKAYLEWPKDRRKCGLFWANLRAAIN       178
          ....
Seq2  111 AAAA------------------------       114
```

15

EMBL双序列比对工具：局部比对工具

EMBL双序列比对工具：局部比对工具

> 黑色是相似的部分，红色是不相似的部分

>Seq3
TTLDDPLGHMPERFDAFICYCPSDIQFVQEMIRQLEQTNYRLKLCVSDRDVLPGTCVWSI
ASELIEKRCRRMVVVVSDDLSLSPGAHQKRLIPIKYKAMKKEFPSILRFITVCDYTNPCT
KSWFWTRLAKALSLP
>Seq4
MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEK
EDGSILICLYESYFDPGKSISENIVSSNQNEWCHYEFYFAHHNLFHENSDHIILILLEPI
PFYCIPTRYHKLKALLEKKAYLEWPKDRRKCGLFWANLRAAIN

局部比对结果：

```
seq3    13 RFDAFICYCPSDIQFVQ-EMIRQLEQTNYRLKLCVSDRDVLPGTCV        57
           ||.||||.|...|..:|: |:|..||:.:..:.:|:.:....||..:
seq4    35 RFHAFISYSEHDSLWVKNELIPNLEKEDGSILICLYESYFDPGKSI        80
```

17

EMBL双序列比对工具：局部比对工具

全局比对与局部比对比较

局
部
比
对

```
#=======================================
#
# Aligned_sequences: 2
# 1: seq3
# 2: seq4
# Matrix: EBLOSUM62                      seq3    13 RFDAFICYCPSDIQFVQ-EMIRQLEQTNYRLKLCVSDRDVLPGTCV  57
# Gap_penalty: 10.0                                 ||.|||.|...|..:|: |:|..||:.:..:.:|:.:.....||.:
# Extend_penalty: 5.0                    seq4    35 RFHAFISYSEHDSLWVKNELIPNLEKEDGSILICLYESYFDPGKSI  80
#
# Length: 46
# Identity:       15/46 (32.6%)
# Similarity:     25/46 (54.3%)
# Gaps:            1/46 ( 2.2%)
# Score: 57.0
#
#
#=======================================
```

全
局
比
对

```
#=======================================
#                                        seq3     1 -----------------------TTLDDPLGHMPERFDAFICYCPSDIQ  26
#                                                                          ||.:....::  ||.|||.|....|..
# Aligned_sequences: 2                   seq4     1 MHHHHHHSSGVDLGTENLYFQSMKTTQEQLKRNV--RFHAFISYSEHDSL  48
# 1: seq3
# 2: seq4                                seq3    27 FVQ-EMIRQLEQTNYRLKLCVSDRDVLPGTCV------------WS----  59
# Matrix: EBLOSUM62                                :|: |:|..||:.:..:.:|:.:.....||..:            |.
# Gap_penalty: 10.0                      seq4    49 WVKNELIPNLEKEDGSILICLYESYFDPGKSISENIVSTSNQNEWCHYEF  98
# Extend_penalty: 0.5
#                                        seq3    60 --IASELIEKRCRRMVVVVSDDLSLSPGAHQKRLIPIKYKAMKK------ 101
# Length: 186                                        ....|..:....::::::.:.:..     ..||.:|..:|.
# Identity:       31/186 (16.7%)         seq4    99 YFAHHNLFHENSDHIILILLLEPIPF-------YCIPTRYHKLKALLEKKA 141
# Similarity:     57/186 (30.6%)
# Gaps:           73/186 (39.2%)         seq3   102 --EFPSILRFITVCDYTNPCTKSWFWTRLAKALSLP      135
# Score: 54.5                                        |:|...|         .|   ..||..|..|::
#                                        seq4   142 YLEWPKDRR---------KC--GLFWANLRAAIN--       164
#
#=======================================
```

其他在线双序列比对工具

软件名	比对类型	网址链接
EMBL	Global/Local	http://www.ebi.ac.uk/Tools/psa
PIR	Global	http://pir.georgetown.edu/pirwww/search/pairwise.shtml
Lalign	Global/Local	http://www.ch.embnet.org/software/LALIGN_form.html
LAGAN	Global	http://lagan.stanford.edu/lagan_web/index.shtml
AlignMe	Alignment of Membrane Proteins	http://www.bioinfo.mpg.de/AlignMe/AlignMe.html
MCALIGN	Alignment of non-coding DNA sequences	http://homepages.ed.ac.uk/eang33/mcalign/mcinstructions.html
Biotools	Global/Local	http://1.51.215.28/~gongj/biotools

其他在线双序列比对工具

Biotools双序列比对工具: http://101.76.246.94/~gongj/biotools

其他在线双序列比对工具

Pairwise Sequence Alignment (PROTEIN)

Result No.: 144746
Sequence Type: protein
Alignment Method: global
Gap Penalty: -3
Substitution Matrix: blosum60
Length of 1.Sequence: 8
Length of 2.Sequence: 8

Up Way:

D R R K C G L F

P C T K S W F W

Up Way:
Sequence Identity: 12.50%
Sequence Gaps : 87.50%
Alignment Score: -3

Down Way:

D R R K C G L F -

P C T K - S W F W

Down Way :
Sequence Identity: 22.22%
Sequence Gaps : 77.78%
Alignment Score: -3

BLAST搜索

✓ 复习：数据库中的序列相似性搜索

在游戏麻将泰坦中，你需要用眼睛从一推麻将牌中找出一对相同的麻将牌。

BLAST搜索

✓ 复习：数据库中的序列相似性搜索

对于一个蛋白质或核酸序列，你需要从序列数据库中找到与它相同或相似的序列。不可能再用眼睛去比较每一对序列，因为数据库中有太多序列，甚至用眼睛比较一对序列都是不可能做到的。

BLAST

UniProtKB
UniProt Knowledgebase

Swiss-Prot (550,552)
Manually annotated and reviewed.

TrEMBL (60,971,489)
Automatically annotated and not reviewed.

BLAST是如何工作的?

BLAST (Basic Local Alignment Search Tool，基本局部比对搜索工具) 是目前最常用的数据库搜索程序。

BLAST基本原理要点是片段对的概念。所谓片段对是指两个给定序列中的一对子序列，它们的长度相等，且可形成无空位的完全匹配。

```
KSD FGETS IV ATGG STE
    |||||    ||||
NE  FGETS LI ATGG NPA
```

BLAST首先找出探测序列和目标序列间所有的匹配程度超过一定阈值的序列片段对，然后对片段对根据给定的相似性阈值进行延伸，得到一定长度的相似性片段，最后给出高分值片段对（high-scoring pairs, HSPs）。改进后的BLAST允许空位的插入。

```
K SD FGETS  IV ATGG S TE
  ::  |||||  ::  ||||  :
  NE FGETS  LI ATGG N PA
```

高分值片段对

BLAST的种类

BLAST实际上是综合在一起的一组工具的统称，它不仅可用于直接对蛋白质序列数据库和核酸序列数据库进行搜索，而且可以将带搜索的核酸序列翻译成蛋白质序列后再进行搜索，或反之，以提高搜索效率。

Blastp：用蛋白质序列搜索蛋白质序列数据库

Blastn：用核酸序列搜索核酸序列数据库

Blastx：将核酸序列按6条链翻译成蛋白质序列后搜索蛋白质序列数据库

tblastn：用蛋白质序列搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成蛋白质序列后再搜索。

tblastx：将核酸序列按6条链翻译成蛋白质序列后搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成的蛋白质序列后再搜索。

根据搜索算法分：标准BLAST，PSI-BLAST，PHI-BLAST等。
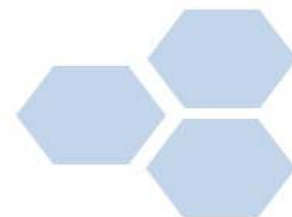
BLAST的种类

BLAST实际上是综合在一起的一组工具的统称，它不仅可用于直接对蛋白质序列数据库和核酸序列数据库进行搜索，而且可以将带搜索的核酸序列翻译成蛋白质序列后再进行搜索，或反之，以提高搜索效率。

Blastn： 用核酸序列搜索核酸序列数据库

Blastp： 用蛋白质序列搜索蛋白质序列数据库

核酸序列

核酸
数据库

BLASTn

序列按6条链翻译成蛋白质序列后搜索蛋白质序列数据库

质序列搜索核酸序列数据库，数据库中的核酸序列要按6
白质序列后再搜索。

序列按6条链翻译成蛋白质序列后搜索核酸序列数据库，
贝序列要按6条链翻译成的蛋白质序列后再搜索。

分：标准BLAST，PSI-BLAST，PHI-BLAST等。

BLAST的

蛋白质序列

蛋白质
数据库
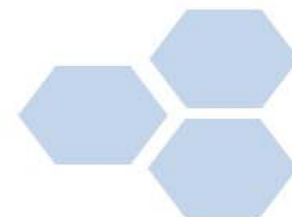
BLASTp

BLAST实际上是一组工具的统称，它不仅可用于直接对蛋白质序列数据库进行搜索，而且可以将带搜索的核酸序列翻译成进行搜索，或反之，以提高搜索效率。

Blastn：用核序列数据库

Blastp：用蛋白质序列搜索蛋白质序列数据库

Blastx：将核酸序列按6条链翻译成蛋白质序列后搜索蛋白质序列数据库

tblastn：用蛋白质序列搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成蛋白质序列后再搜索。

tblastx：将核酸序列按6条链翻译成蛋白质序列后搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成的蛋白质序列后再搜索。

根据搜索算法分：标准BLAST，PSI-BLAST，PHI-BLAST等。

BLAST的

BLAST实际上~~~~~~~~~~~不仅可用于直接对蛋白质序列数~~~~~~~~~~可以将带搜索的核酸序列翻译成~~~~~~提高搜索效率。

Blastp： 用蛋白~~

Blastn： 用核酸序列搜索核酸序列数据库

Blastx： 将核酸序列按6条链翻译成蛋白质序列后搜索蛋白质序列数据库

tblastn： 用蛋白质序列搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成蛋白质序列后再搜索。

tblastx： 将核酸序列按6条链翻译成蛋白质序列后搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成的蛋白质序列后再搜索。

根据搜索算法分： 标准BLAST，PSI-BLAST，PHI-BLAST等。

核酸序列 → 蛋白质序列

蛋白质
数据库

BLASTx

BLAST的...

蛋白质序列

核酸数据库 → 蛋白质数据库

tBLASTn

BLAST实际上...它不仅可用于直接对蛋白质序列数据...且可以将带搜索的核酸序列翻译成...以提高搜索效率。

Blastp： 用蛋白...

Blastn： 用核酸序列搜索核酸序列数据库

Blastx： 将核酸序列按6条链翻译成蛋白质序列后搜索蛋白质序列数据库

tblastn： 用蛋白质序列搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成蛋白质序列后再被搜索。

tblastx： 将核酸序列按6条链翻译成蛋白质序列后搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成的蛋白质序列后再被搜索。

根据搜索算法分：标准BLAST，PSI-BLAST，PHI-BLAST等。

BLAST的

核酸序列 → 蛋白质序列

核酸数据库 → 蛋白质数据库
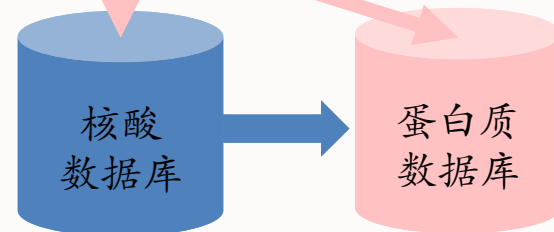
tBLASTx

BLAST实际上 ~~~ 它不仅可用于直接对蛋白质序列数据 ~~~ 且可以将带搜索的核酸序列翻译成 ~~~ 以提高搜索效率。

Blastp： 用蛋 ~~~

Blastn： 用核酸序列搜索核酸序列数据库

Blastx： 将核酸序列按6条链翻译成蛋白质序列后搜索蛋白质序列数据库

tblastn： 用蛋白质序列搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成蛋白质序列后再搜索。

tblastx： 将核酸序列按6条链翻译成蛋白质序列后搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成的蛋白质序列后再搜索。

根据搜索算法分： 标准BLAST，PSI-BLAST，PHI-BLAST等。

BLASTn

BLASTp

BLASTx
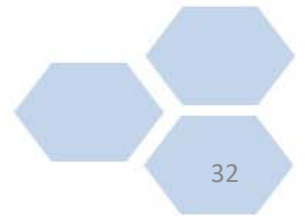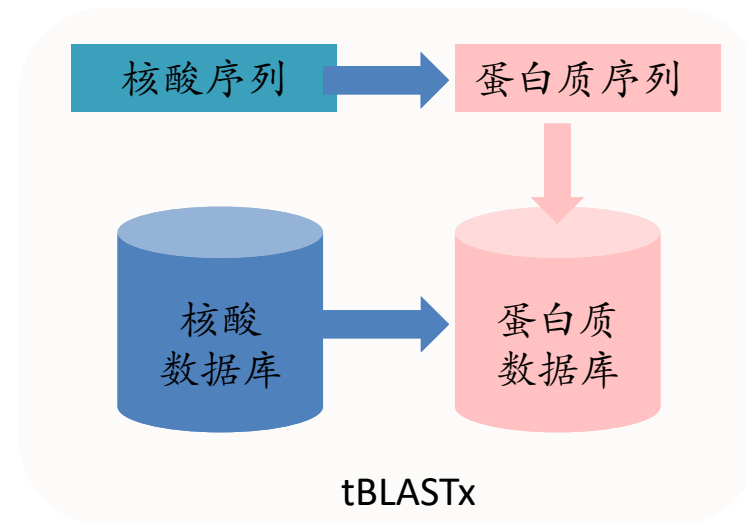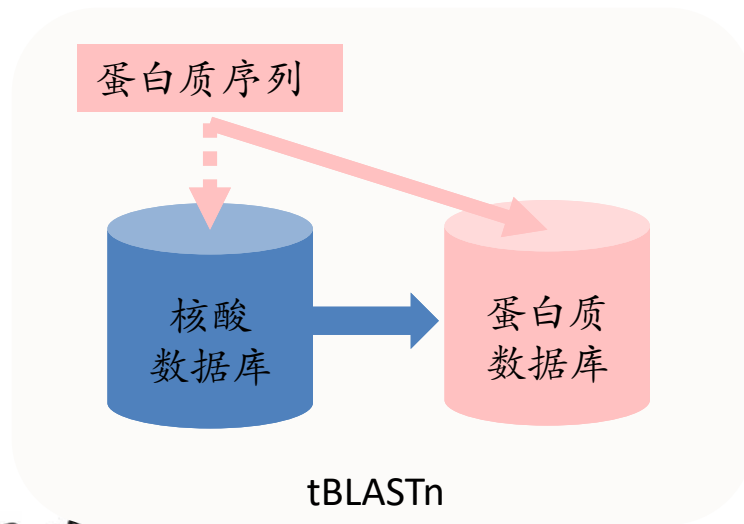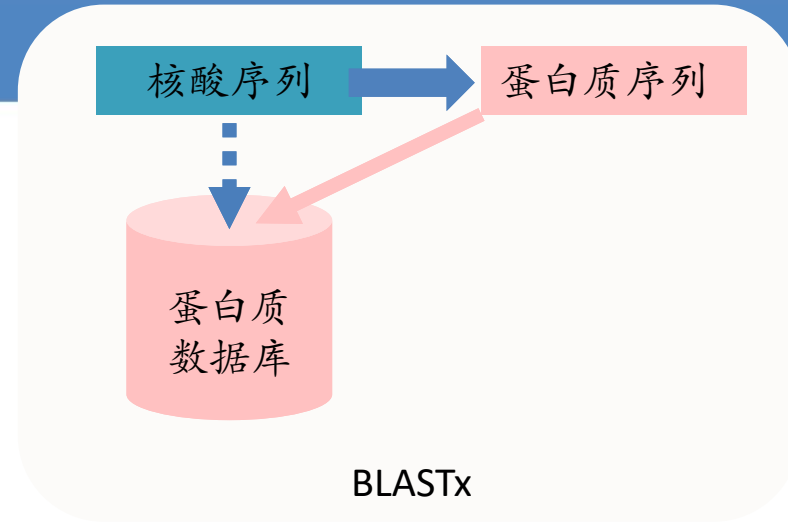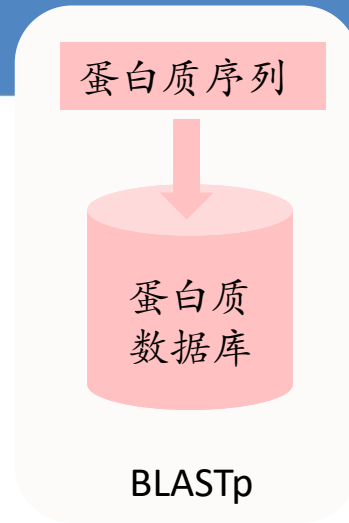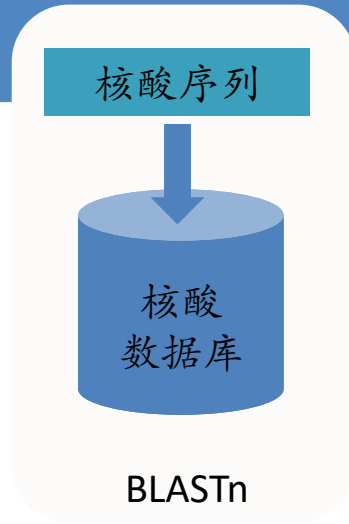
tBLASTn

tBLASTx

BLAST的种类

BLAST实际上是综合在一起的一组工具的统称，它不仅可用于直接对蛋白质序列数据库和核酸序列数据库进行搜索，而且可以将带搜索的核酸序列翻译成蛋白质序列后再进行搜索，或反之，以提高搜索效率。

Blastp：用蛋白质序列搜索蛋白质序列数据库

Blastn：用核酸序列搜索核酸序列数据库

Blastx：将核酸序列按6条链翻译成蛋白质序列后搜索蛋白质序列数据库

tblastn：用蛋白质序列搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成蛋白质序列后再搜索。

tblastx：将核酸序列按6条链翻译成蛋白质序列后搜索核酸序列数据库，数据库中的核酸序列要按6条链翻译成的蛋白质序列后再搜索。

根据搜索算法分：标准BLAST，PSI-BLAST，PHI-BLAST等。

NCBI：BLASTp

选择被搜索的数据库：用NCBI的BLAST工具搜索Swissprot数据库

第一部分：搜索任务描述

第二部分：图形化搜索结果

Graphic Summary

Show Conserved Domains

输入序列所属的蛋白质家族

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

Query seq.
Superfamilies

TIR

找到50个高分匹配片段(Hits)
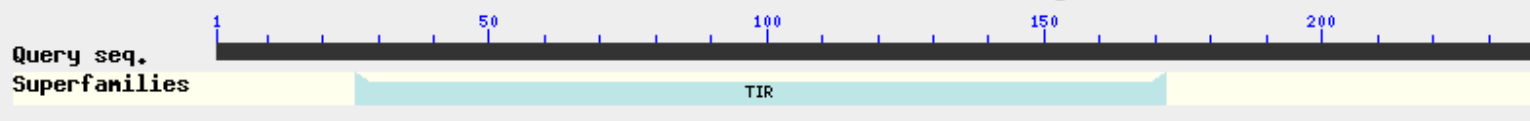
Distribution of 50 Blast Hits on the Query Sequence

Q9JLZ8 RecName: Full=Single Ig IL-1-related receptor; AltName:.. S=491 E=3.8e-175

Color key for alignment scores

| <40 | 40-50 | 50-80 | 80-200 | >=200 |

Query
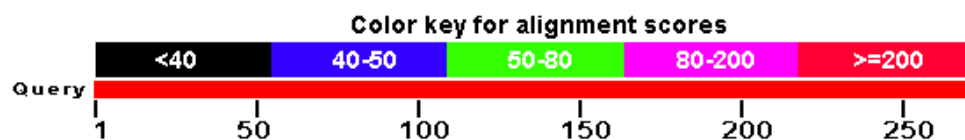
所有搜索到的Hits以不同颜色和长短的线列出

40

第三部分：序列信息列表

颜色 | 长短

NCBI：PSI-BLAST

有时基本的BLAST搜素还是不能满足需要。比如，你想通过一条蛋白质序列，搜罗出一个庞大的蛋白质家族。如果运行基本的BLAST搜素，你只能找到那些和探索序列十分相近的序列，而其他那些远亲就找不到了。换言之，你找到了你直接认识的朋友，而朋友的朋友都丢掉了。

PSI BLAST (Position-Specific Iterated BLAST，位点特异性迭代BLAST)

PSI-BLAST的特色是每次用位置特异权重矩阵（Position-Specific Scoring Matrix，PSSM）搜索数据库后再利用搜索的结果重新构建PSSM，然后用新的PSSM再次搜索数据库，如此反复（iteration）直至没有新的结果产生为止。

```
Seq1:   A    B    C    D
Seq2:   B    B    C    D
Seq3:   A    C    C    D
Seq4:   A    B    D    D
```

	1	2	3	4
A	75%	0	0	0
B	25%	75%	0	0
C	0	25%	75%	0
D	0	0	25%	100%

假设：A的朋友只有B
B的朋友除A外还有C

第二轮搜索结果（有新找到的序列）

直接跳到第一条新找到的序列，即第一条标黄的序列

NCBI：PHI-BLAST

PHI-BLAST (Pattern-Hit Initiated BLAST, 模式识别BLAST)：能找到与输入序列相似的并符合某种特定模式（pattern）的序列。

例如，N-糖基化位点基序（N-glycosylation site motif）总是符合以下特定模式：*以Asn开始，然后紧跟除了Pro之外的任何一个氨基酸，再紧跟Ser或者Thr，再紧跟除Pro外的任何一个氨基酸。* N{P}[ST]{P}

用正则表达式书写的符合模式：
{L}GEx[GAS][LIVM]x(3,7)

VGEAAMPRI yes   **VGEAAYPRI** no

MGEGLATA ?   **FGETAIII ?**

这种序列特征模式可能代表某个翻译后修饰的发生位点，也可以代表一个酶的活性位点，或者一个蛋白质家族的结构域、功能域。

其他BLAST

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Specialized searches

SmartBLAST	**Primer-BLAST**	**Global Align**	**CD-search**
Find proteins highly similar to your query	Design primers specific to your PCR template	Compare two sequences across their entire span (Needleman-Wunsch)	Find conserved domains in your sequence
GEO	**IgBLAST**	**VecScreen**	**CDART**
Find matches to gene expression profiles	Search immunoglobulins and T cell receptor sequences	Search sequences for vector contamination	Find sequences with similar conserved domain architecture
Targeted Loci	**Multiple Alignment**	**BioAssay**	**MOLE-BLAST**
Search markers for phylogenetic analysis	Align sequences using domain and protein constraints	Search protein or nucleotide targets in PubChem BioAssay	Establish taxonomy for uncultured or enviromental sequences

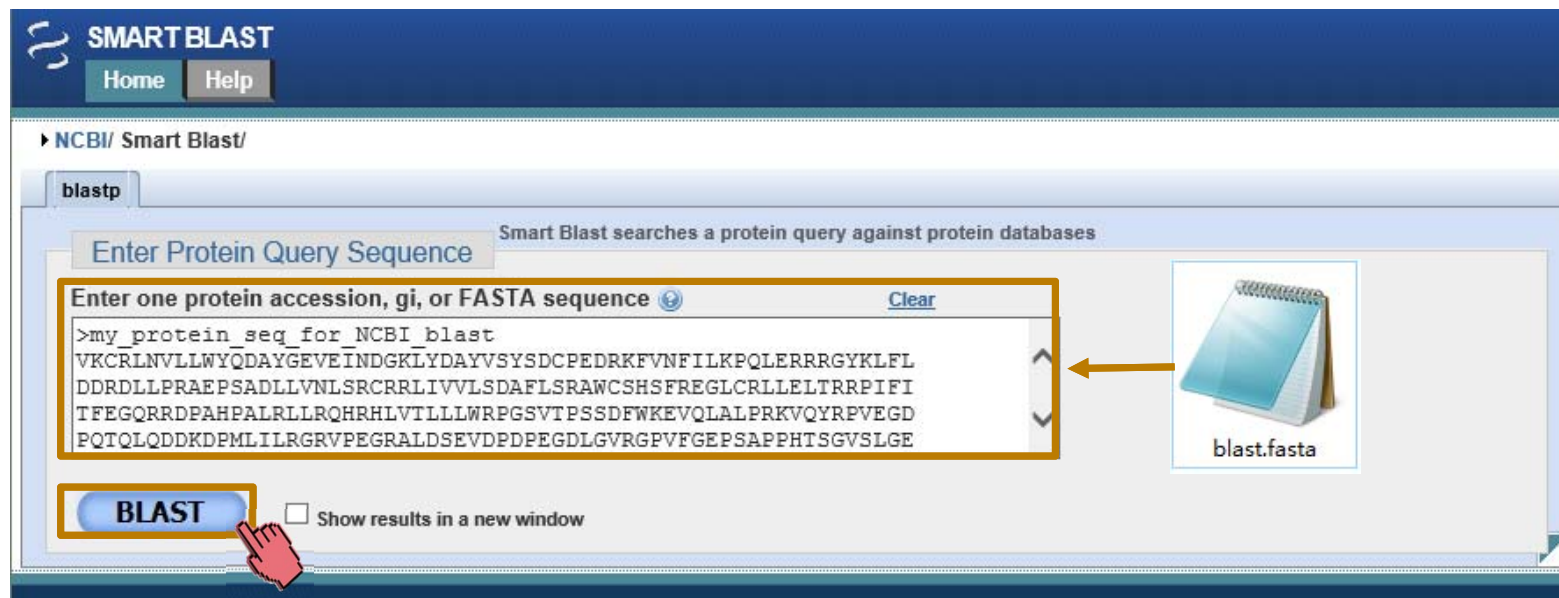PEOPLE'S MEDICAL PUBLISHING HOUSE

其他BLAST

超级简单明了的输入界面，是BLASTp的简约强化版

精简的搜索结果包括数据库中与输入序列最相似的三条序列，以及研究的最透彻的物种中可以展现一定的进化关系的最相似的两条序列。

其他BLAST

✓ 互联网上的免费搜索工具

位置	服务器	网址链接
USA	NCBI	http://www.ncbi.nlm.nih.gov/BLAST
Europe	ExPASy	http://web.expasy.org/blast
Europe	Uniprot	http://www.uniprot.org/blast/
Japan	DDBJ	http://blast.ddbj.nig.ac.jp

✓ **WU-BLAST** - WU 代表 Washington University. 比NCBI-BLAST更灵敏，在插入空位的算法上更灵活。

✓ **Smith and Waterman (SSEARCH):** 有点儿慢，但是比BLAST更准确。

✓ **FASTA:** 有点儿慢，但是对于DNA序列的比较比BLAST更准确。

✓ **BLAT:** 用于小的序列（如 cDNA 等）在大基因组中的搜索。