序列分析

序列比对

- 序列比对的理论基础是进化学说，如果两个序列之间具有足够的相似性，就推测二者可能有共同的进化祖先，经序列内残基或者序列片段的**替换**、**插入**、**缺失**等遗传编译过程分别演化而来。
- 相似性高并不一定来自同一祖先。


similarity
homology
structure
function
evolution
……

核苷酸比对VS蛋白比对

- 一般从进化意义上分析,蛋白比对比核苷酸更可取
  - 蛋白序列大约可追溯10亿年前的祖先
  - 核苷酸序列大约可追溯6亿年
- 确定DNA一致性,多态性用核苷酸比对

提出问题

- 已知数条蛋白质序列，
- 我们能获得什么信息？

A S R L Y K A A

A S R L Y K S A

S R P Y T K A A

怎么做？

- 评估不同氨基酸的相似性
- 如何序列序列配对

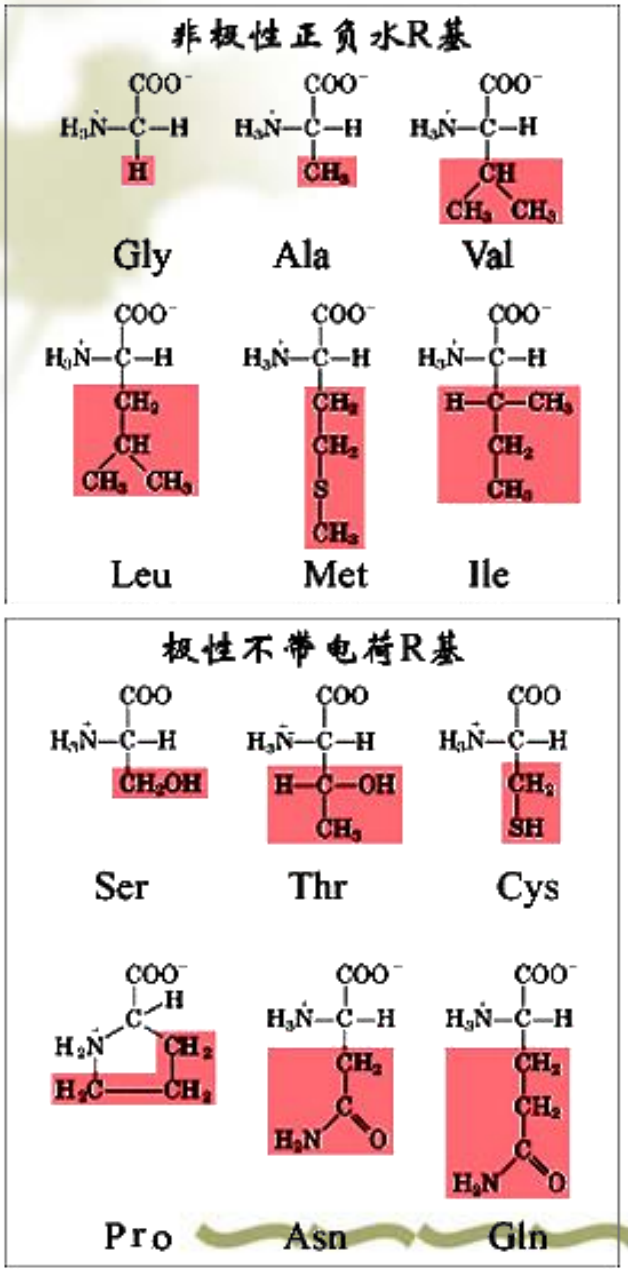
氨基酸的相似性

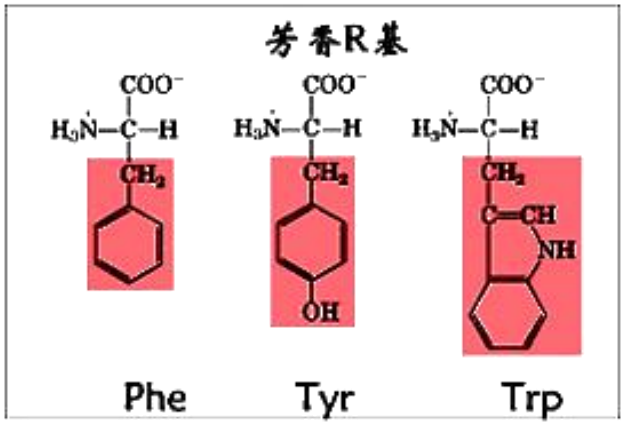
- （i）等价矩阵

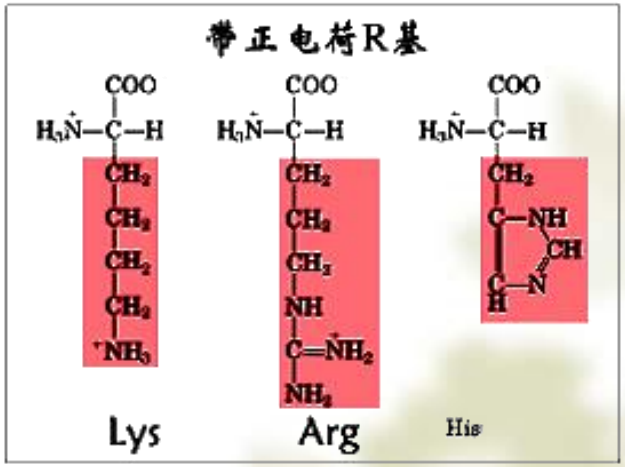
$$R_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

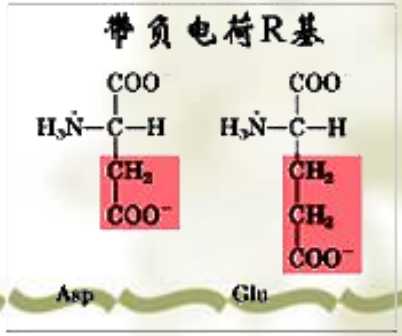
- （ii）疏水矩阵
- （iii）氨基酸突变代价矩阵GCM
- （iv）PAM矩阵
- （v） BLOSUM矩阵

氨基酸的相似性


非极性正负水R基
Gly
Ala
Val
Leu
Met
Ile
极性不带电荷R基
Ser
Thr
Cys
Pro
Asn
Gln


芳香R基
Phe
Tyr
Trp


带正电荷R基
Lys
Arg
His


带负电荷R基
Asp
Glu

PAM矩阵（Point Accepted Mutation）

- 意义:在蛋白质中被自然选择接受的单个氨基酸替换
- 1978年,Dayhoff 等研究了71个近似蛋白的1572个突变,ASN, SER等突变频繁, TRP,CYS等很少突变,由此统计了氨基酸两两突变频率表 。
- 从概率到对数比值:
  - $S(a, b)=10*\lg(M_{ab}/P_b)$
  - M_{ab}: a 突变到b的概率
  - P_b: b在所有统计数据中的概率
- 概率矩阵乘法,外推远缘概率值PAM1>>PAM250

PAM 矩阵

- 缺点：一旦PAM1的矩阵有效地误 差，那么自乘250后得到的PAM250矩阵的误差就会变得很大。

- 针对不同的进化距离采用PAM 矩阵
  - 序列相似度 = 40% 50% 60%
  - 打分矩阵 = PAM120 PAM80 PAM60
  - PAM250 → 14% ~ 27%
  - 这类矩阵里列出同源蛋白质在进化过程中氨基酸变化的可能性。

BLOSUM矩阵

- BLOSUM矩阵： 是目前用得较多的打分矩阵，它是Henikoff根据BLOCK数据库中蛋白质序列的高度保守部分的alignment而得到，是现在很多软件的首选矩阵，最常用的是BLOSUM62。

- Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

两种打分矩阵的比较

- 二者目标都是对进化上相似的序列打分
- 差别在具体实现方法上：
  - PAM矩阵基于近似序列的统计，从进化距离近的数据外推到远。
  - BLOSUM矩阵基于蛋白家族中的多序列比对统计。

- The two result in the same scoring outcome, but use differing methodologies. BLOSUM directly look at mutations in motifs of related sequences while PAM's extrapolate evolutionary information based on closely related sequences.

使用区别

- PAM-n中，n 越小，表示氨基酸变异的可能性越小；相似的序列之间比较应该选用n值小的矩阵，不太相似 的序列之间比较应该选用n值大的矩阵。PAM-250用于约20%相同序列之间的比较。

- BLOSUM-n中，n越小，表示氨基酸相似的可能性越小；相似的序列之间比较应该选用 n 值大的矩阵，不太相似的序列之间比较应该选 用n值小的矩阵。BLOSUM-62用来比较62%相似度的序列，BLOSUM-80用来比较80%左右的序列。

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

BLOSUM 62

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

序列联配(alignment)

当允许有gap时，考虑两条长度分别为m和n的序列，可能的alignment方式数目是：

$$\frac{(m+n)!}{m!\,n!}$$

当 m=n=4 时，有　70 种alignment方式；

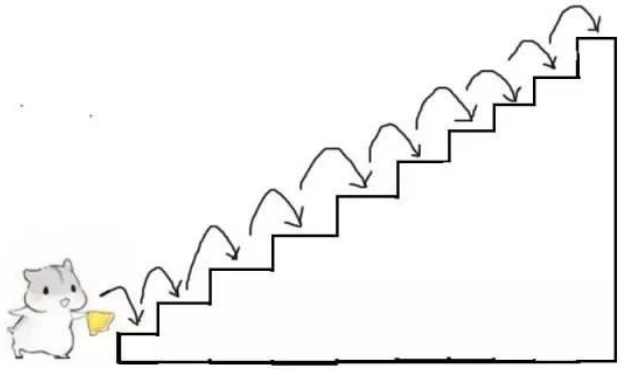
当 m=n=8 时，有12870 种alignment方式；

。。。。。。

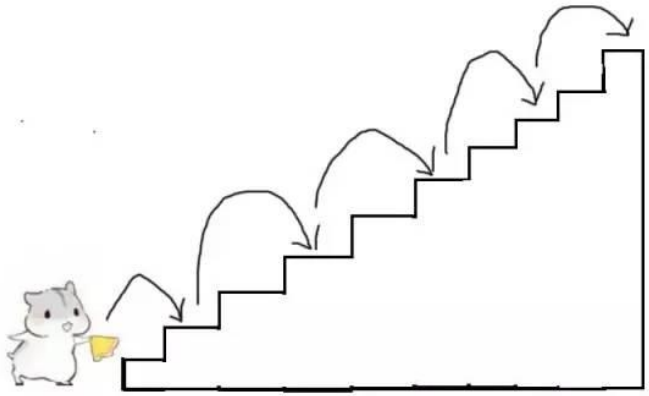
显然，对于实际的蛋白质或DNA序列来说，用这种穷举法是行不通的，而到目前为止还没有其他可保证找到最佳的alignment的方法。 **Needleman-Wunsch的算法**是目前公认的最有效的近似方法。

动态规划的例子

- **题目：**
- 有一座高度是**10**级台阶的楼梯，从下往上走，每跨一步只能向上**1**级或者**2**级台阶。要求用程序来求出一共有多少种走法。



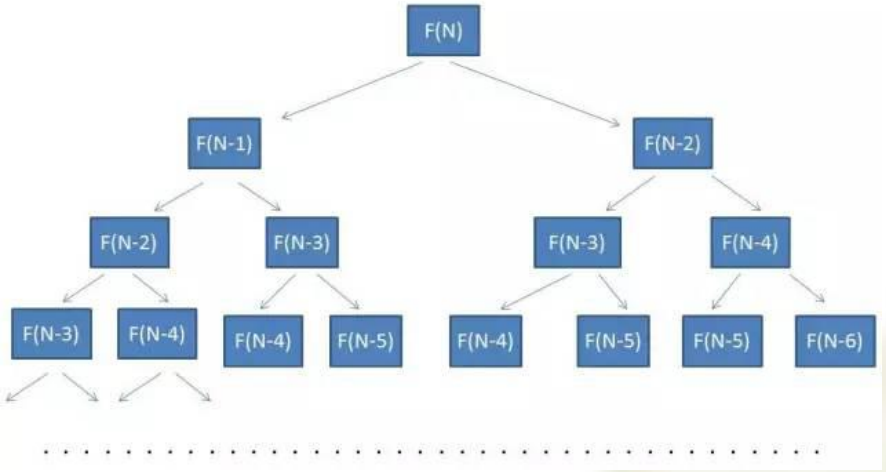
- 比如，每次走1级台阶，一共走10步，这是其中一种走法。我们可以简写成 1,1,1,1,1,1,1,1,1,1。

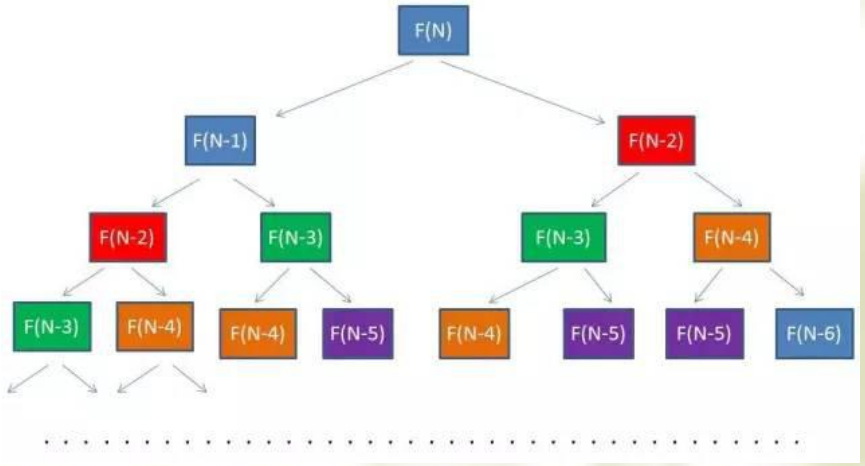


- 再比如，每次走2级台阶，一共走5步，这是另一种走法。我们可以简写成 2,2,2,2,2。

递归算法的思路

- F(1) = 1;
- F(2) = 2;
- F(n) = F(n-1)+F(n-2)
  （ n>=3 ）





动态规划的思路

- F(1) = 1;
- F(2) = 2;
- F(n) = F(n-1)+F(n-2)（n>=3）

- 要点:
- 1)边界
- 2)递推关系(状态转移公式)
- 3)最优子结构

台阶数	1	2	3	4	5	6	7	8	9
走法数	1	2	3						

台阶数	1	2	3	4	5	6	7	8	9
走法数	1	2	3	5					

台阶数	1	2	3	4	5	6	7	8	9
走法数	1	2	3	5	8				

NW的序列比对方法

问题分解：

例: 对序列 P 和 G 进行比对

P
G

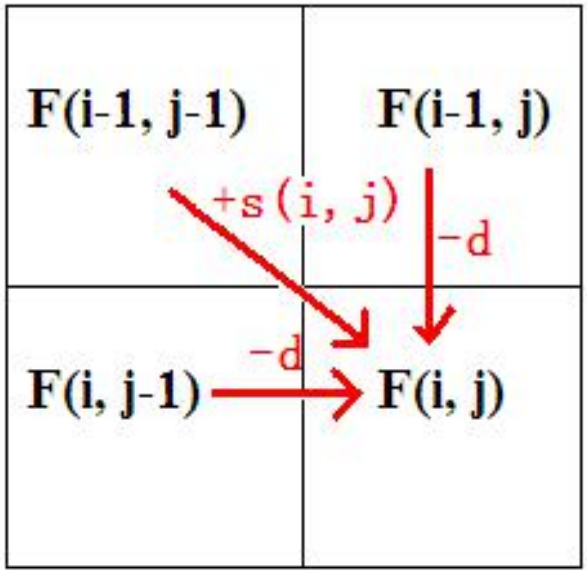
P
^ G

^ P
G

NW的序列比对方法

条件: 1. 两条序列长度相差不多; 2. 两条序列确实存在较 高的同源性。

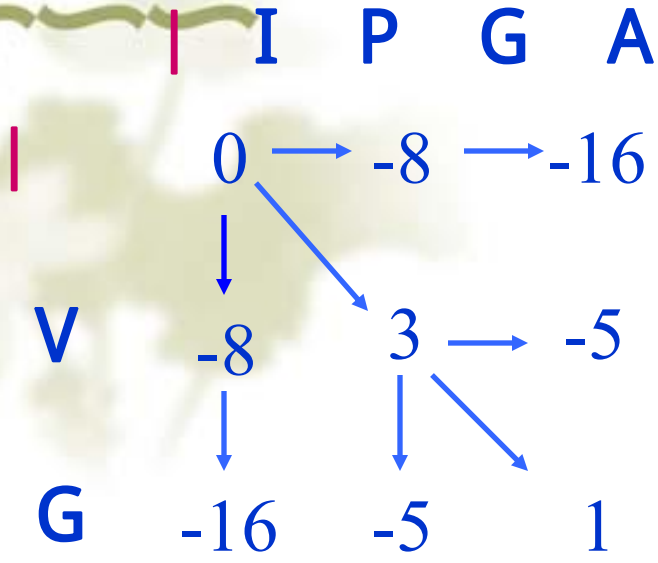
方法:



$$F(i,j)=\max\begin{cases}F(i-1,j)-d\\F(i-1,j-1)+s(i,j)\\F(i,j-1)-d\end{cases}$$

文献: J. Mol. Biol. 48: 443-452, 1970

J. Mol. Biol. 162: 705-708, 1982



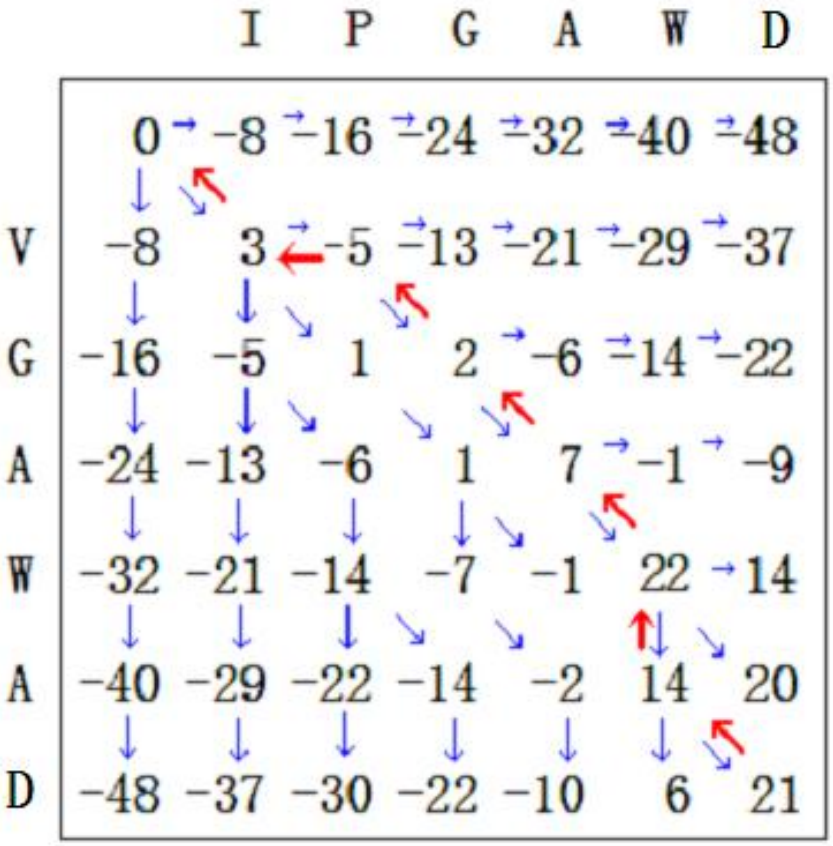
BLOSUM-45 (删减)

	I	P	G	A	W	D
V	3	-3	-3	0	-3	-3
G	-4	-2	7	0	-2	-1
W	-2	-3	-2	-2	15	-4
A	-1	-1	0	5	-2	-2
D	-4	-1	-1	-2	-4	7

例: 用BLOSUM45对序列 IPGAWD 和 VGAWAD 进行全局alignment。

BLOSUM-45 (删减)

	I	P	G	A	W	D
V	3	-3	-3	0	-3	-3
G	-4	-2	7	0	-2	-1
W	-2	-3	-2	-2	15	-4
A	-1	-1	0	5	-2	-2
D	-4	-1	-1	-2	-4	7



Traceback(回溯)：从右下角往左上角回溯，如果当前分值：

1. 来自对角线，对应的两个氨基酸对齐；
2. 来自上方，“横”序列加gap;
3. 来自左边，“纵”序列加gap.

可能存在异议情况。

IPGAW-D

V-GAWAD

序列比对软件
FASTA 和 BLAST

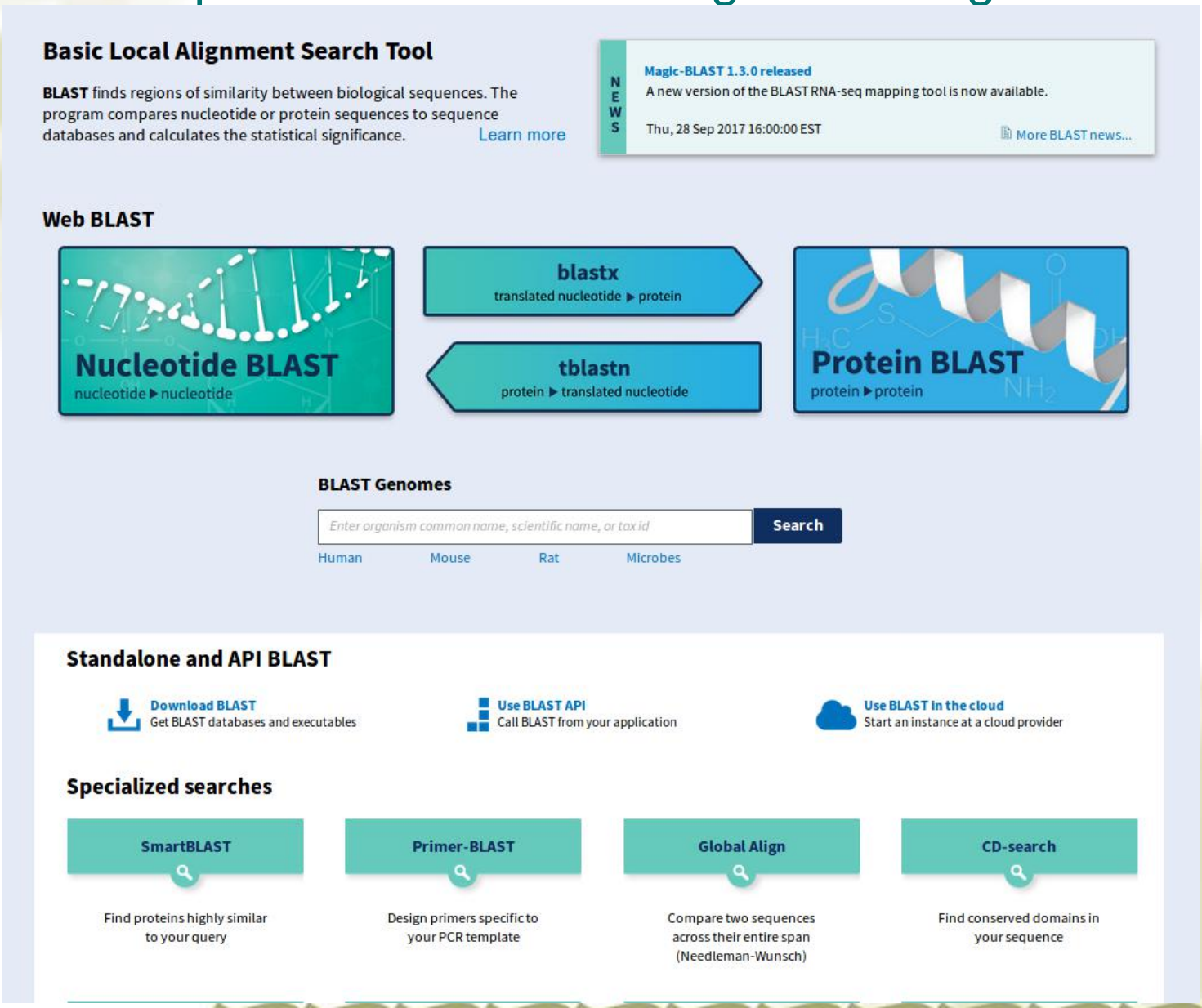
- FASTA和BLAST是目前功能最全，使用最广的同源性数据库搜索软件包。它们在Needleman 的动态算法的基础上做了很多技术上的改进, 如采用启发性算法, 使得在精确度牺牲较小的情况下, 速度快了很多。

- FASTA 是 D.J. Lipman and W.R. Person 在1985年提出一个全局联配算法。{ Science 227, 1435-1441, 1985; PNAS 85, 2444-2448, 1988}

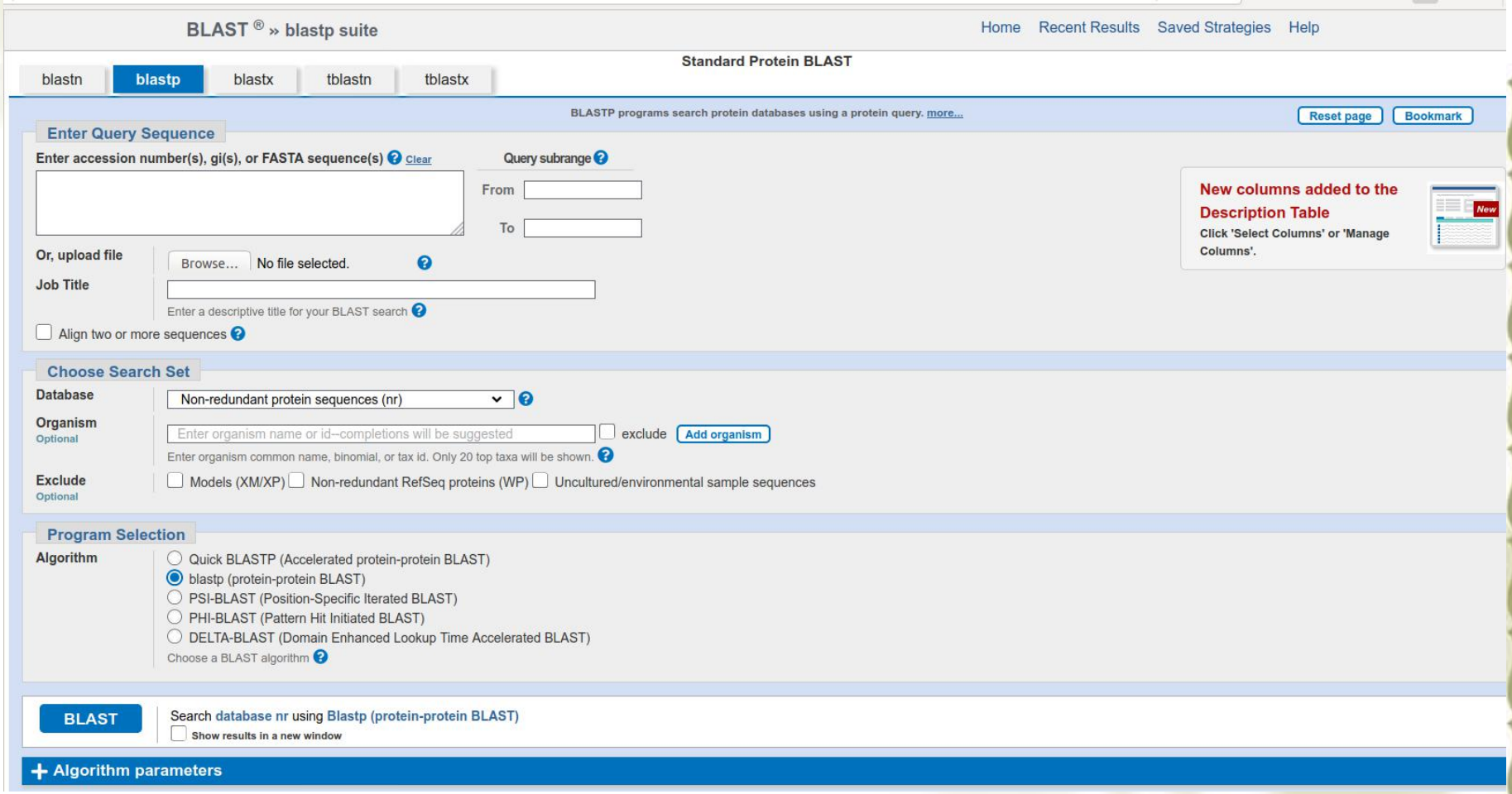
- BLAST(Basic Local Alignment Search Tool) 是 D.J. Lipman 和 S.F. Altschul等人1990年提出的，最初被设计用于序列局部比对。 {J. Mol. Biol. 215, 403-410, 1990}

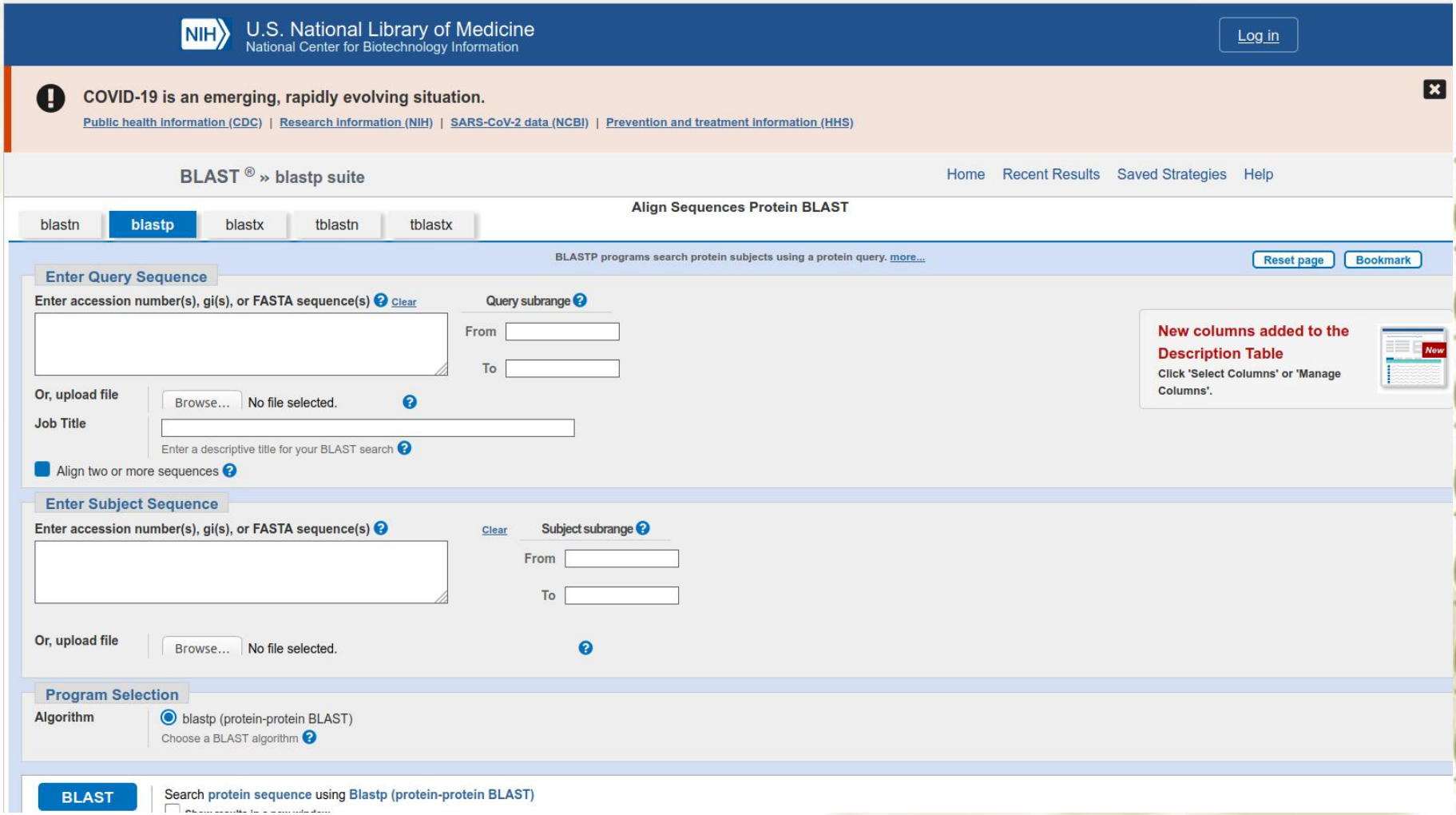
- 两个算法都经过多次改进，变得越来越相象。

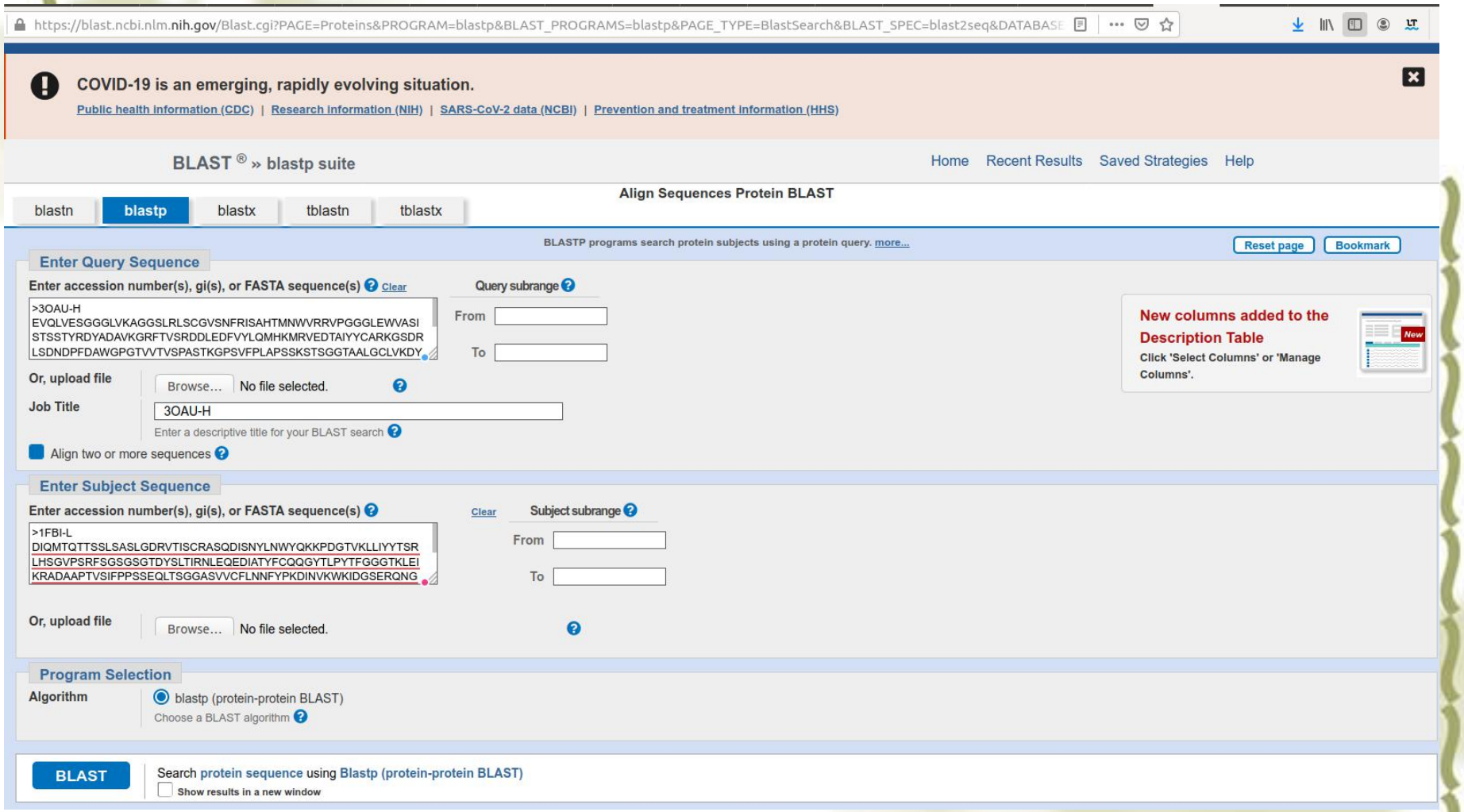
https://blast.ncbi.nlm.nih.gov/Blast.cgi

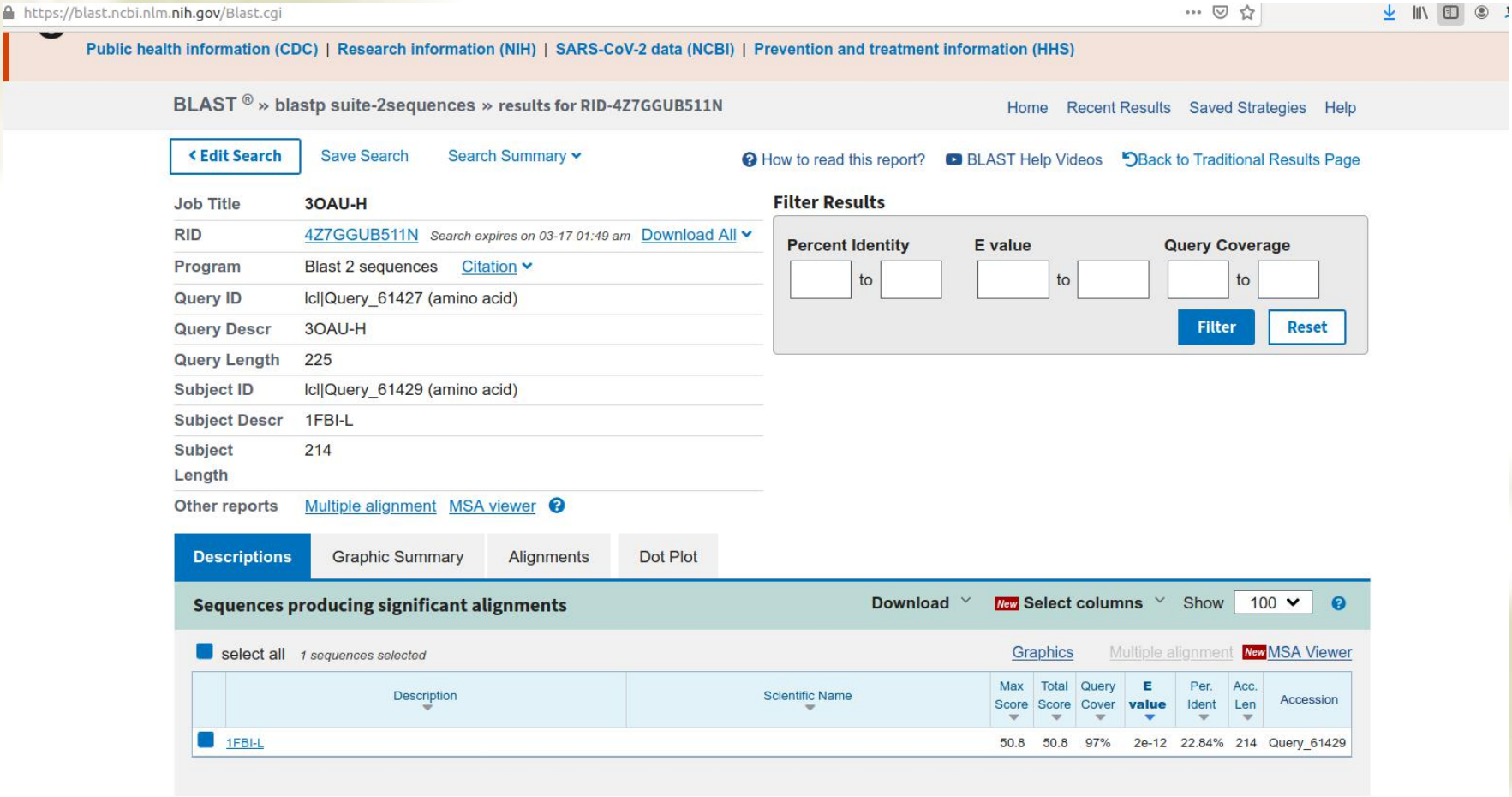


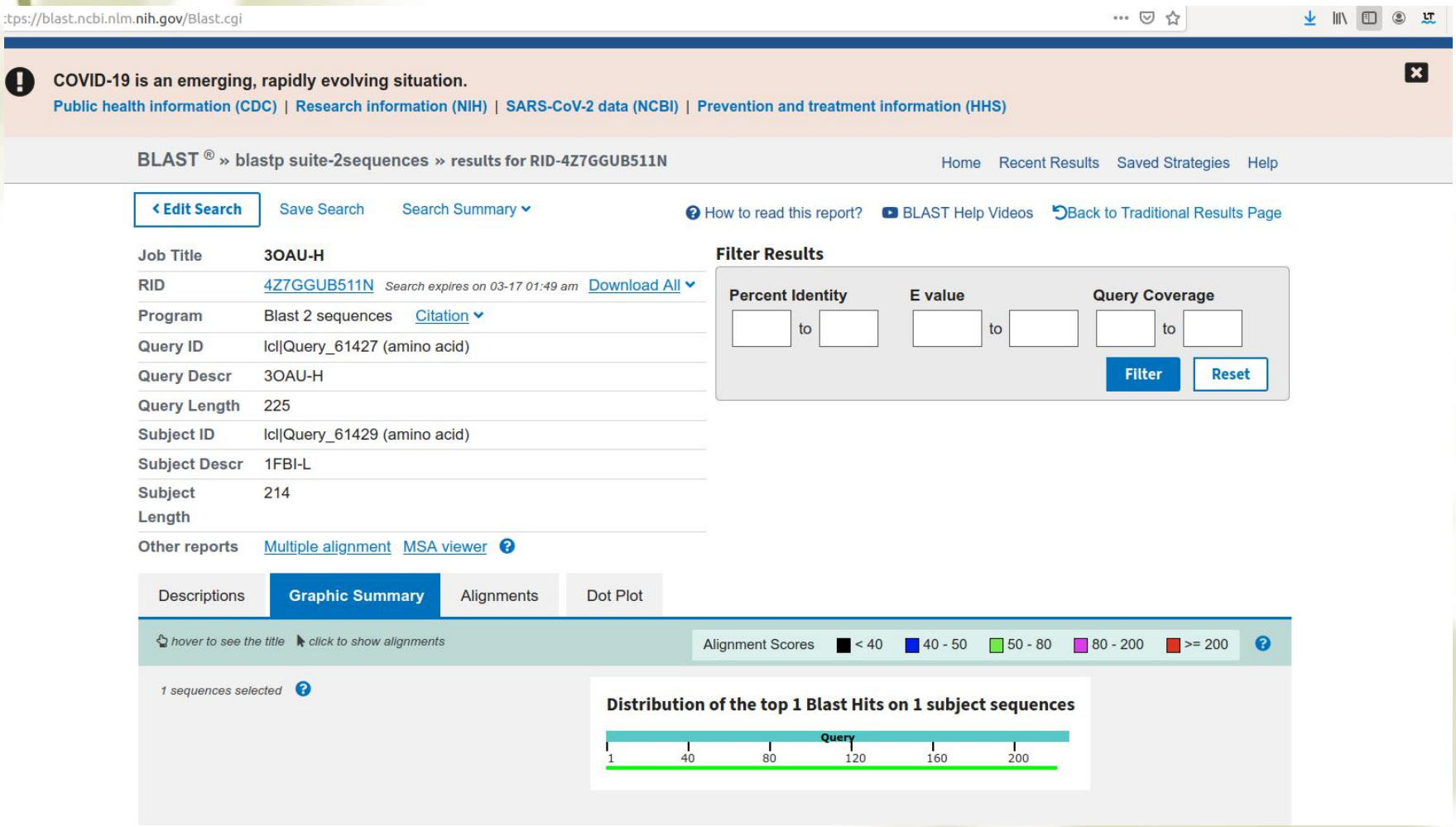
blastp


BLAST ® » blastp suite
Home
Recent Results
Saved Strategies
Help
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
New columns added to the Description Table
Click 'Select Columns' or 'Manage Columns'.
New
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
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Non-redundant RefSeq proteins (WP)
Uncultured/environmental sample sequences
Program Selection
Algorithm
Quick BLASTP (Accelerated protein-protein BLAST)
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database nr using Blastp (protein-protein BLAST)
Show results in a new window
+ Algorithm parameters


U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
COVID-19 is an emerging, rapidly evolving situation.
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)
BLAST ® » blastp suite
Home Recent Results Saved Strategies Help
Align Sequences Protein BLAST
blastn blastp blastx tblastn tblastx
BLASTP programs search protein subjects using a protein query. more...
Reset page Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
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
New columns added to the Description Table
Click 'Select Columns' or 'Manage Columns'.
New
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
Or, upload file
Browse... No file selected.
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)


https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&DATABASE
COVID-19 is an emerging, rapidly evolving situation.
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)
BLAST ® » blastp suite
Home Recent Results Saved Strategies Help
Align Sequences Protein BLAST
blastn blastp blastx tblastn tblastx
BLASTP programs search protein subjects using a protein query. more...
Reset page Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
>3OAU-H
EVQLVESGGGLVKAGGSLRLSCGVSNFRISAHTMNWVRRVPGGGLEWVASI
STSSTYRDYADAVKGRFTVSRDDLEDFVYLQMHKMRVEDTAIYYCARKGSDR
LSDNDPFDAWGPGTVVTVSPASTKGPSVFPLAPSSKSTSGGTAALGCLVKDY
Query subrange
From
To
Or, upload file
Browse... No file selected.
Job Title
3OAU-H
Enter a descriptive title for your BLAST search
Align two or more sequences
New columns added to the Description Table
Click 'Select Columns' or 'Manage Columns'.
New
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
>1FBI-L
DIQMTQTTSSLSASLGDRVTISCRASQDISNYLNWYQKKPDGTVKLLIYYTSR
LHSGVPSRFSGSGSGTDYSLTIRNLEQEDIATYFCQQGYTLPYTFGGGTKLEI
KRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNG
Subject subrange
From
To
Or, upload file
Browse... No file selected.
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window


https://blast.ncbi.nlm.nih.gov/Blast.cgi
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)
BLAST ® » blastp suite-2sequences » results for RID-4Z7GGUB511N
Home Recent Results Saved Strategies Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 3OAU-H
RID 4Z7GGUB511N Search expires on 03-17 01:49 am Download All
Program Blast 2 sequences Citation
Query ID lcl|Query_61427 (amino acid)
Query Descr 3OAU-H
Query Length 225
Subject ID lcl|Query_61429 (amino acid)
Subject Descr 1FBI-L
Subject Length 214
Other reports Multiple alignment MSA viewer
Filter Results
Percent Identity to
E value to
Query Coverage to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Dot Plot
Sequences producing significant alignments
Download
New Select columns
Show 100
select all 1 sequences selected
Graphics
Multiple alignment
New MSA Viewer
Description
Scientific Name
Max Score
Total Score
Query Cover
E value
Per. Ident
Acc. Len
Accession
1FBI-L 50.8 50.8 97% 2e-12 22.84% 214 Query_61429


:tps://blast.ncbi.nlm.nih.gov/Blast.cgi
COVID-19 is an emerging, rapidly evolving situation.
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)
BLAST ® » blastp suite-2sequences » results for RID-4Z7GGUB511N
Home Recent Results Saved Strategies Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 3OAU-H
RID 4Z7GGUB511N Search expires on 03-17 01:49 am Download All
Program Blast 2 sequences Citation
Query ID lcl|Query_61427 (amino acid)
Query Descr 3OAU-H
Query Length 225
Subject ID lcl|Query_61429 (amino acid)
Subject Descr 1FBI-L
Subject Length 214
Other reports Multiple alignment MSA viewer
Filter Results
Percent Identity to
E value to
Query Coverage to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Dot Plot
hover to see the title
click to show alignments
Alignment Scores < 40 40 - 50 50 - 80 80 - 200 >= 200
1 sequences selected
Distribution of the top 1 Blast Hits on 1 subject sequences
Query
1 40 80 120 160 200

Length

Other reports Multiple alignment MSA viewer

Descriptions | Graphic Summary | **Alignments** | Dot Plot

Alignment view: Pairwise | Restore defaults | Download

1 sequences selected

Download | Graphics | Next | Previous | Descriptions

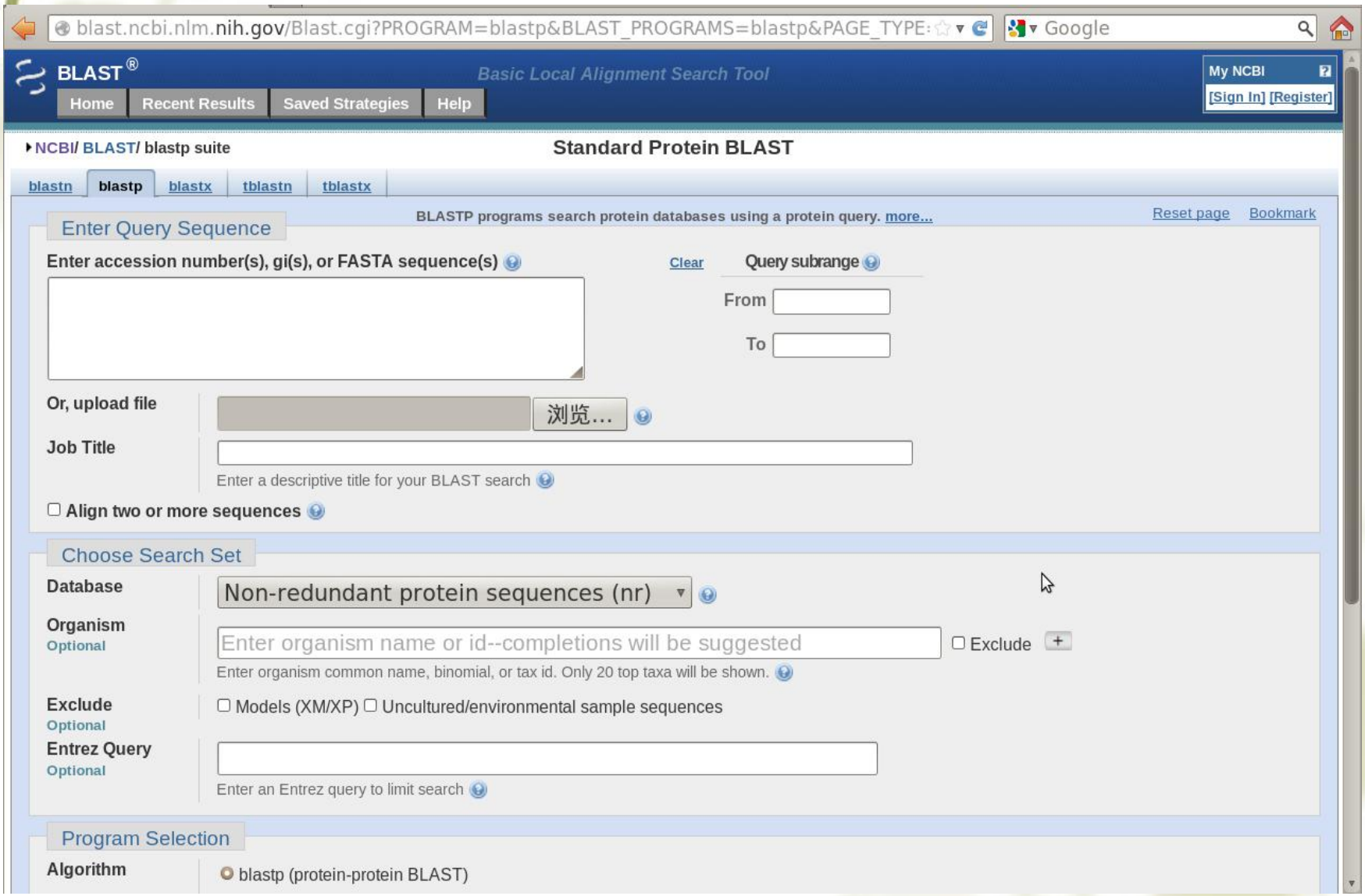
1FBI-L

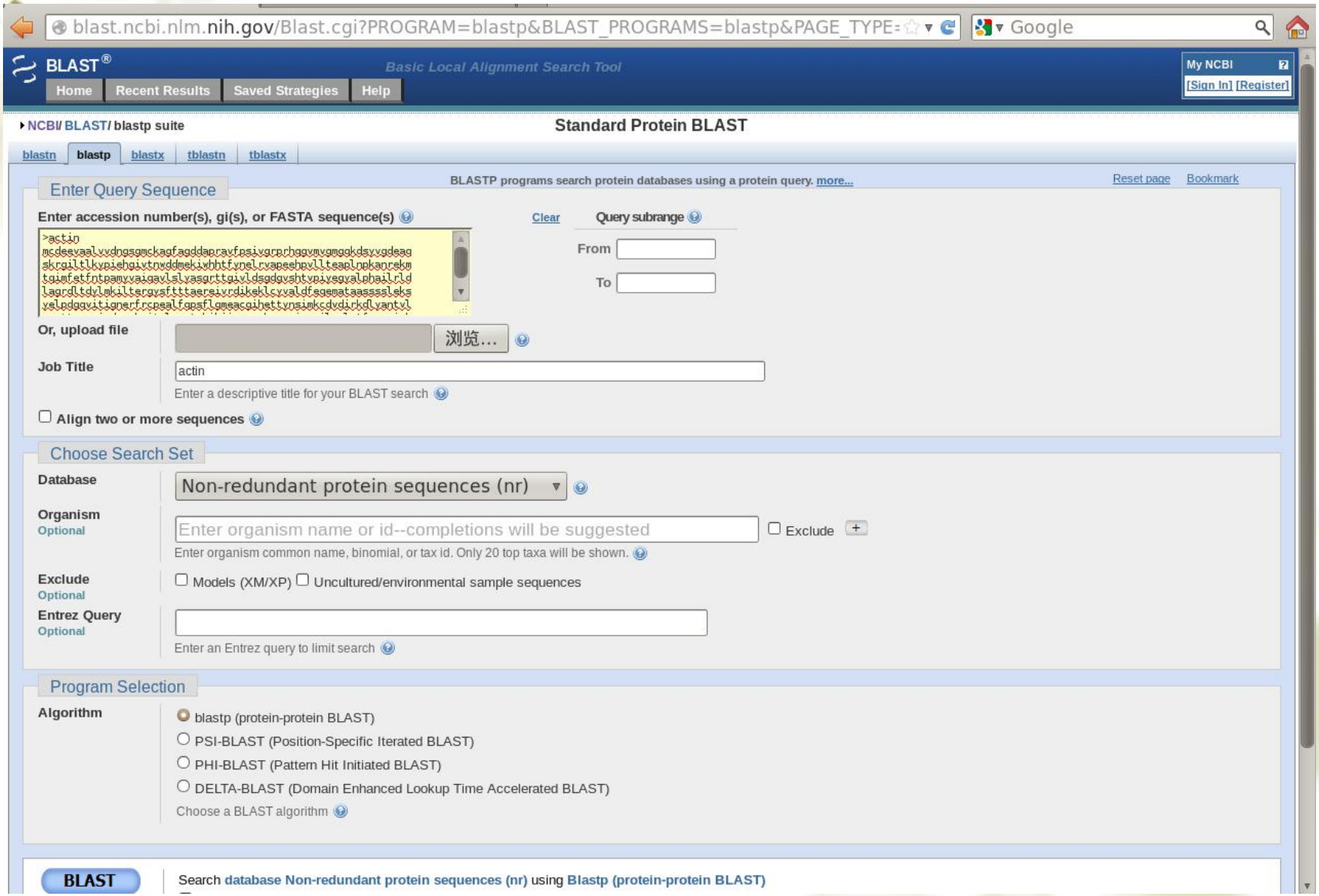
Sequence ID: Query_61429 Length: 214 Number of Matches: 1

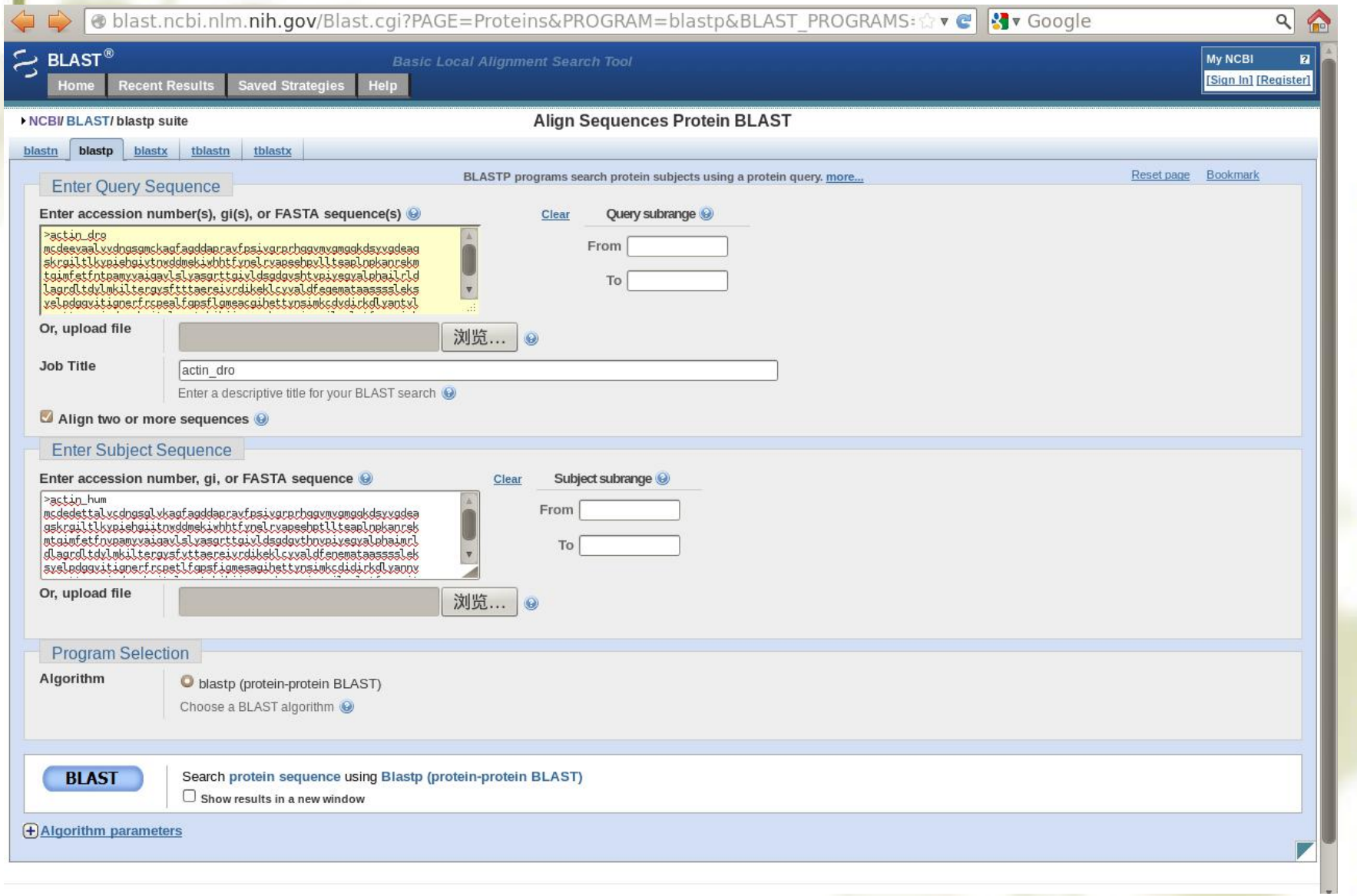
Range 1: 1 to 207 Graphics | Next Match | Previous Match

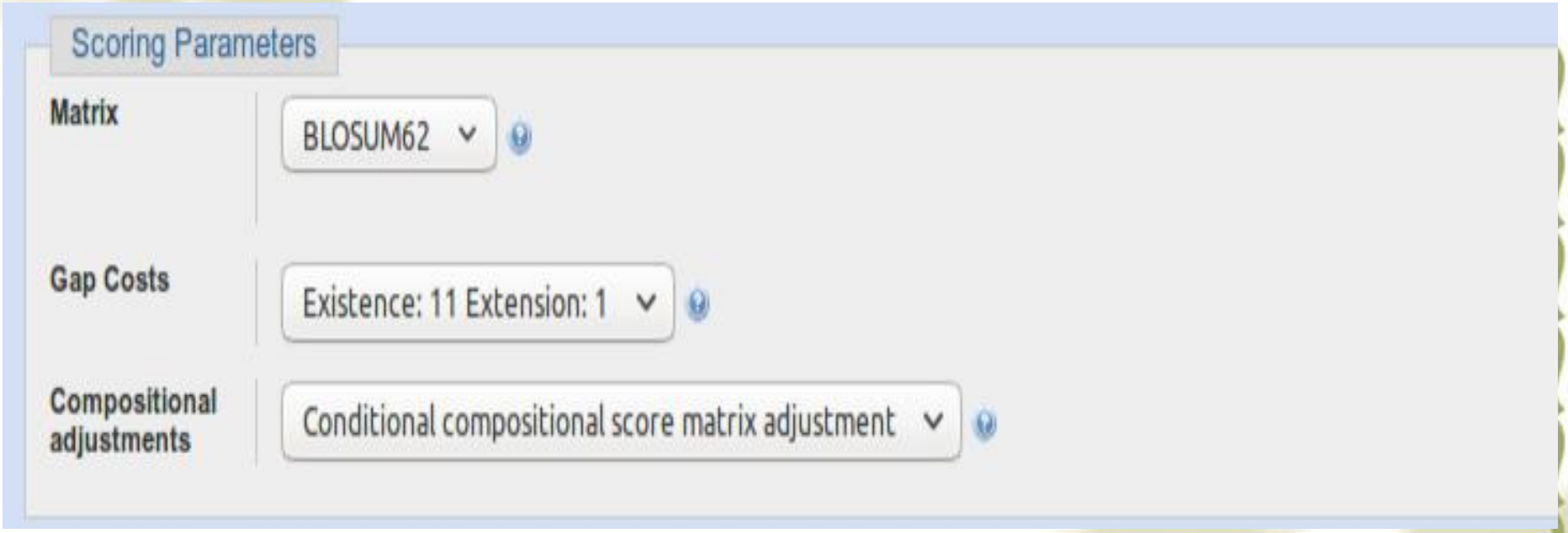
Score	Expect	Method	Identities	Positives	Gaps
50.8 bits(120)	2e-12	Compositional matrix adjust.	53/232(23%)	97/232(41%)	38/232(16%)

```
Query  1    EVQLVESGGGLVKAGG-----SLRLSCGVSNFRISAHTMNWVRRVPGGGLEWVASISTSS  55
            ++Q+ ++   L  + G     S R S  +SN+      +NW ++ P G ++ +   ++
Sbjct  1    DIQMTQTTSSLSASLGDRVTISCRASQDISNY------LNWYQKKPDGTVKLLIYYTSR-  53

Query  56   TYRDYADAVKGRFTVSRDDLEDFVYLQMHKMRVEDTAIYYCARKGSDRLSDNDPFDAWGP  115
                    V  RF+ S    +    L +  +  ED A Y+C +  +       P+  +G
Sbjct  54   ----LHSGVPSRFSGSGSGTD--YSLTIRNLEQEDIATYFCQQGYTL------PY-TFGG  100

Query  116  GTVVTVSPASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHT  175
            GT + +  A    P+V    PSS+  + G A++ C + +++P+ + V W       G
Sbjct  101  GTKLEIKRADA-APTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWK----IDGSER  155

Query  176  FPAVLQS-------SGLYSLSSVVTVPSSSLGTQ-TYICNVNHKPSNTKVDK  219
               VL S           YS+SS +T+          +Y C   HK S + + K
Sbjct  156  QNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIVK  207
```


blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE:
Google
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
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 your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Algorithm
blastp (protein-protein BLAST)


blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=
Google
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>actin
Or, upload file
浏览...
Job Title
actin
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)


blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS:
Google
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein subjects using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>actin_dro
Or, upload file
浏览...
Job Title
actin_dro
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number, gi, or FASTA sequence
Clear
Subject subrange
From
To
>actin_hum
Or, upload file
浏览...
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters


Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment

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

Google

BLAST® *Basic Local Alignment Search Tool*

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite-2sequences/ Formatting Results - MJ8DTGSE11R

Edit and Resubmit | Save Search Strategies | Formatting options | Download | How to read this page | Blast report description

Blast 2 sequences

actin_dro

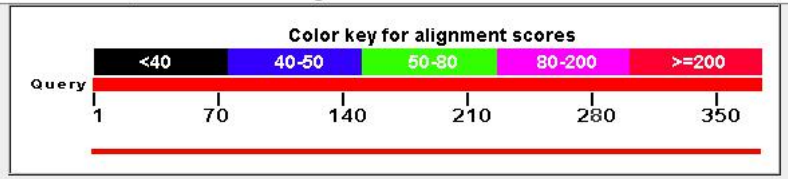
Query ID	lcl|36237	**Subject ID**	36239
Description	actin_dro	**Description**	actin_hum
Molecule type	amino acid	**Molecule type**	amino acid
Query Length	376	**Subject Length**	377
		Program	BLASTP 2.2.27+ Citation

Other reports: Search Summary [Taxonomy reports] [Multiple alignment]

Graphic Summary

Distribution of 1 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Dot Matrix View

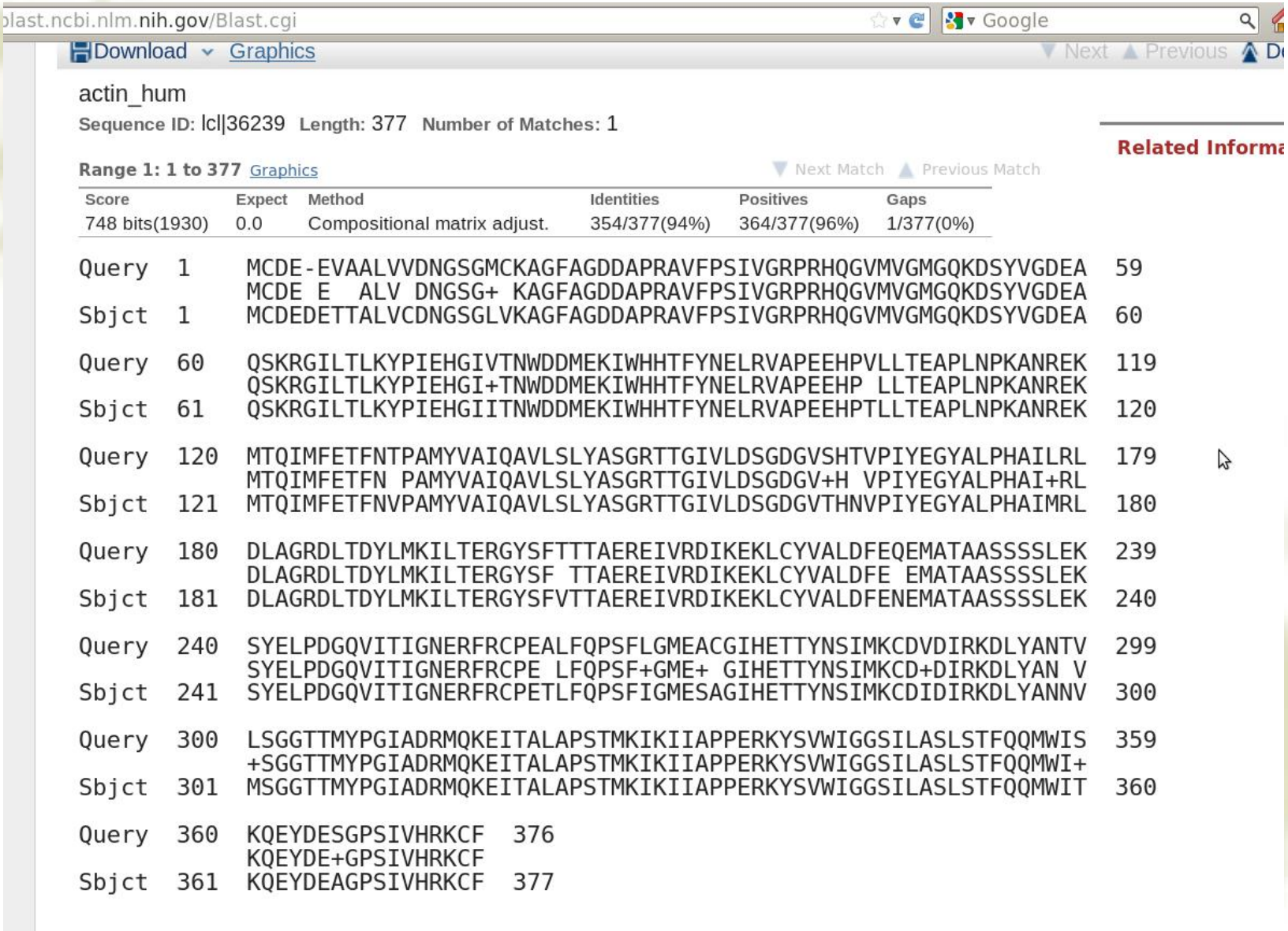
Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | Graphics | Multiple alignment

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
☐	actin_hum	748	748	100%	0.0	94%	36239


last.ncbi.nlm.nih.gov/Blast.cgi
Google
Download
Graphics
Next
Previous
actin_hum
Sequence ID: lcl|36239 Length: 377 Number of Matches: 1
Related Informa
Range 1: 1 to 377 Graphics
Next Match
Previous Match
Score Expect Method Identities Positives Gaps
748 bits(1930) 0.0 Compositional matrix adjust. 354/377(94%) 364/377(96%) 1/377(0%)
Query 1 MCDE-EVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEA 59
MCDE E ALV DNGSG+ KAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEA
Sbjct 1 MCDEDETTALVCDNGSGLVKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEA 60
Query 60 QSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREK 119
QSKRGILTLKYPIEHGI+TNWDDMEKIWHHTFYNELRVAPEEHP LLTEAPLNPKANREK
Sbjct 61 QSKRGILTLKYPIEHGIITNWDDMEKIWHHTFYNELRVAPEEHPTLLTEAPLNPKANREK 120
Query 120 MTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRL 179
MTQIMFETFN PAMYVAIQAVLSLYASGRTTGIVLDSGDGV+H VPIYEGYALPHAI+RL
Sbjct 121 MTQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRL 180
Query 180 DLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEK 239
DLAGRDLTDYLMKILTERGYSF TTAEREIVRDIKEKLCYVALDFE EMATAASSSSLEK
Sbjct 181 DLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEK 240
Query 240 SYELPDGQVITIGNERFRCPEALFQPSFLGMEACGIHETTYNSIMKCDVDIRKDLYANTV 299
SYELPDGQVITIGNERFRCPE LFQPSF+GME+ GIHETTYNSIMKCD+DIRKDLYAN V
Sbjct 241 SYELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNV 300
Query 300 LSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWIS 359
+SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWI+
Sbjct 301 MSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWIT 360
Query 360 KQEYDESGPSIVHRKCF 376
KQEYDE+GPSIVHRKCF
Sbjct 361 KQEYDEAGPSIVHRKCF 377

序列比对结果的解读

- 假设A-A得1分，A-B得-1分

- AAAA
- AAAA
- 得分4分

- AAAAAA
- AAABAA
- 得分4分

- 这两对序列，得分相同是什么意义？
- 这两对序列，谁的序列相似性更可靠？

Blast统计学显著性：E-value

称为期望值，计算公式为：

$$E=K\cdot m\cdot n\cdot e^{-\lambda S}$$

- 其意义为：若查询序列为一条随机序列，对于同样的配对空间，有希望找到 E 条相似性得分为 S 的序列。
- m，n 是序列长度，K和λ参数与打分系统有关，一般通过Monte Carlo模拟得出。
- 显然，E的值越小，说明结果越有意义。当 E小于0.05时，E值可近似为统计显著性P值。

关于E值的一些经验规则:

- E 值小于 0.01的序列可以认定为同源序列; E值介于1和10之间的序列也是一些值得注意的序列。
- 进行蛋白质序列同源搜索时，E值上限的默认设置是10.0，进行核酸序列同源搜索时E值上限的默认设置是2.0 。
- 根据自己的需要设置E值上限。