[Blast][https://www.ncbi.nlm.nih.gov/books/NBK279684/]

BLAST (Basic Local Alignment Search Tool)是一套在蛋白质数据库或DNA数据库中进行相似性比较的分析工具；对一条或多条序列(可以是任何形式的序列)在一个或多个核酸或蛋白序列库中进行比对。

Usage

1. Create blast database

-dbtype: nucl 指定为核酸数据库

-parse_seqids: parse seqid for FASTA

-hash_index: create index of sequence hash values

```
makeblastdb -in blaKPC-2.fsa -hash_index -dbtype nucl -out databases/blaKPC-2
```

2. blastn，搜索核酸数据库

-evalue: 1e-3，表示10的-3次方，e表这是一种指数形式的计数方法。由数符、十进制数、阶码标志'E'或'e'以及阶符和阶码组成

S: S值表示两个序列的同源性，分值越高，表示它们之间的相似程度越大

E: E值表示S值的可靠性，表示随机条件下，其他序列能和目标序列相似程度大于S值的可能行，所以该值越小越好

`E=K*m*n(e^-lambada*S), K和lambada与数据库及算法有关，常量；m代表目标序列的长度，n代表数据库的大小，S就是前面提到的S值`

E值局限性：1. 当目标序列过小是，E值偏大，因无法的到较高S值；2. 在有gap情况下，两序列同源性高，但S值会下降；3. 序列非功能区域的有较低随机性，可能会导致两序列较高同源性

btop: 简述比对情况，数字表示匹配；GA表示G变成了A；-表示gap；* tblastx表示gap

-outfmt，0表示成对比较输出；6/7/8可额外指定输出格式

-sorthits，指定hits的排序，0表evalue，1表bit score，2表total score，3表一致性比率，4表覆盖度

-max_target_seqs，选择输出匹配数目，推荐5条或更多

```
blastn -evalue 1e-5 -max_target_seqs 1 -db databases/blaKPC-2 -query query.fasta -outfmt "6 qseqid qstart qend sseqid sstart send length pident mismatch gapopen bitscore evalue btop" > query_kp-2.blastn
```

3. tblastx，在蛋白数据库搜索DNA序列(nucletode vs nucletide by peptides)

首先同样构建nucl数据库，然后再构建一个prot数据库

```
makeblastdb -in blaKPC-2.fsa -hash_index -dbtype prot -out databases/blaKPC-2
```

最后比对，参数同blastn

-query_genecode, 指定query转录方式，微生物为11

-db_genecode，指定subject转录方法，微生物为11

```
tblastx -evalue 1e-3 -db databases/blaKPC-2 -query_gencode 11 -db_gencode 11 -query query.fasta -max_target_seqs 5 -outfmt "6 qseqid qlen qstart qend sseqid slen sstart send length pident mismatch gapopen bitscore evalue btop" > query_kpc-2.tblastx
```

Diamond

DIAMOND用于蛋白和转录后的DNA序列比对，用于测序数据：

- 成对比较蛋白和转录后到DNA序列，速度是BLAST的500到20000倍
- 可针对长read分析做移码比对
- 多种输出格式，包括BLAST pairwise, tabular和XML

Usage

使用方法类似blast

1. makedb，从fasta个数输入文件构建DIAMOND格式的参考数据库

-db/-d file 指定输出数据库文件

```
diamond makedb --in nr.faa -d nr
```

2. blastp，比对蛋白输入序列；blastx，比对转录后的DNA输入序列，默认输出BLAST tabular格式

--query-gencode，指定blastx用于转录的遗传密码子

--strand，both/plus/minus，指定query序列的链用于比对，默认为both

--min-orf/-l，忽略转录后包含小于该长度ORF的序列，设置为1将取消该参数作用

--sensitive，该比对方式敏感性更佳，用于长read比对；默认方式用于短read比对，例如30-40氨基酸序列，显著性比对>50 bits；—more-sensitive，更敏感比对

--frameshift/-F，针对DNA-vs-protein比对的移码罚分，数值在15作用。该参数推荐用于长，倾向于发生indel的序列，例如MinION reads，\和/对应转录方向+1/-1的移码

--matrix，打分矩阵，默认为BLOSUM62

--outfmt/-f，输出格式同blast参数，0表示BLAST pairwise格式，5表BLAST XML格式，6表示BLAST tabular可选参数，100表DIAMOND alignment archive(DAA)，该二进制格式可通过view命令生成其他输出格式

--evalue/-e，指定期待值(默认0.001)

--top，例如10，指定输出最大score到90% * 最大score之间的比对(setting this to 10 will report all alignments whose score is at most 10% lower than the best alignment score for a query)

--min-score，指定最小bit score

--id，指定最小一致性比例

--max-target-seqs，输出最大的匹配数目

```
diamond blastx -d nr -q reads.fna --id 90 --query-gencode 11 --max-target-seqs 1 --outfmt 6 qseqid qlen qstart qend sseqid slen sstart send length pident mismatch gapopen bitscore evalue btop -o matches.m8
```

默认输出tabular：`qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore.`

3. view，从DAA文件生成格式化输出

--daa/-a，指定输入DAA格式文件

--out/-o，指定输出文件，同样可采用--outfmt指定输出格式

4. dbinfo，查看数据库文件信息

The Genetic Codes

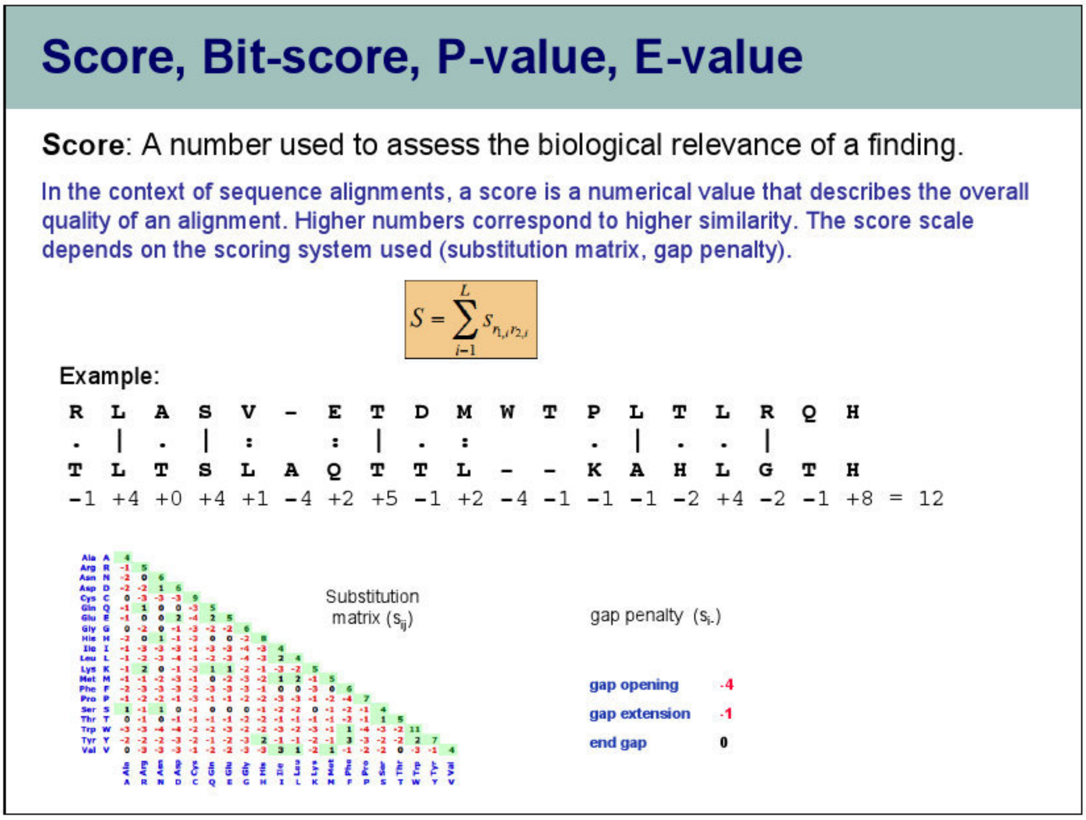
https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Systematic Range and Comments:

Table 11 is used for Bacteria, Archaea, prokaryotic viruses and chloroplast proteins. As in the standard code, initiation is most efficient at AUG. In addition, GUG and UUG starts are documented in Archaea and Bacteria . In E. coli, UUG is estimated to serve as initiator for about 3% of the bacterium's proteins. CUG is known to function as an initiator for one plasmid-encoded protein (RepA) in Escherichia coli. In addition to the NUG initiations, in rare cases Bacteria can initiate translation from an AUU codon as e.g. in the case of poly(A) polymerase PcnB and the InfC gene that codes for translation initiation factor IF3. The internal assignments are the same as in the standard code though UGA codes at low efficiency for Trp in Bacillus subtilis and, presumably, in Escherichia coli .

Query sequence type	Database sequence type	Alignment level type	What the program should be called	What the program is actually called
nucleotide	nucleotide	nucleotide	**blastNN**	**blastn**
peptide	peptide	peptide	**blastPP**	**blastp**
nucleotide	peptide	peptide	**blastNP**	**blastx**
peptide	nucleotide	peptide	**blastPN**	**tblastn**
nucleotide	nucleotide	peptide	**blastNNP**	**tblastx**

Evaluation criteria



Score, Bit-score, P-value, E-value

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

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty).

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

Example:

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

Substitution matrix (s_{ij})

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

gap penalty (s_{i-})

gap opening	-4
gap extension	-1
end gap	0

Score, Bit-score, P-value, E-value

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

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty).

Gap penalty	Alignment	Identity / Similarity	Gaps	Score
0	1 GTC-ATGCTA-GTCGT---GG---GTAGCATTTA-GCT-ATG-TGGG-GT 38 || ||||| |||| || |||| ||| | | ||| | | 1 -TCGATGCT-GGTCG-CAAGGCAAGTAG---TTATG-TCATGCT---AG- 39	27/50 (54.0%)	23/50	S=135
5	1 GTC-ATGCTAGTCG--TGGGTAGCATTTA-GCT-ATG-TGGGGT 38 || |||||.|||| .||..||.|.||| | | ||| |.| 1 -TCGATGCTGGTCGCAAGGCAAGTAGTTATG-TCATGCTAG--- 39	26/44 (59.1%)	11/44	S=67
10	1 ----------------------------GTCATGCTAGTCGTGGGTAGC 21 |||||||||| 1 TCGATGCTGGTCGCAAGGCAAGTAGTTATGTCATGCTAG----------- 39 22 ATTTAGCTATGTGGGGT 38 39 ----------------- 39	10/67 (14.9%)	57/67	S=50

Observations: If the gap penalty is too large, gaps are avoided and the sequences can not be properly aligned. If the gap penalty is too low, gaps are inserted everywhere to prevent mismatches. This does not produce any informative alignement. The "best" alignment is obtained for an intermediary gap penalty.

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

Score, Bit-score, P-value, E-value

Bit-score: A log-scaled version of a score.

In the context of sequence alignments (BLAST), the **bit-score S'** is a normalized score expressed in *bits* that lets you estimate the magnitude of the *search space* you would have to look through before you would expect to find an score as good as or better than this one by chance. Althshul proposes to following definition:

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

S is the raw score. Parameters λ and K depend on the substitution matrix and on the gap penalties (Altchul).

Ex: If the bit-score is 30, you would have to score, on average, about 2^{30} = 1 billion independent segment pairs to find a score this score by chance. Each additional bit doubles the size of the search space.

The bit-scores is thus a rescaled version of the raw alignment score that is *independent of the size of the search space*.

The **size of the search space** is proportional to the product of the query sequence length (*n*) * the sum of the lengths of the sequences in the database (*m*): ***N=n*m***. The size of the search space is then obtained by multiplying N by a coefficient *K* (Altschul).

Ex: When searching protein databases with protein queries, *K* is about 0.13. Thus, for a protein of length n=235 aa which is searched against a database of size m=12 496 420 aa, the size of the search space is equal to 0.13 * 235 * 12 496 420 = about 0.38 billion. In this case, a bit score of 30 (which corresponds to a space of 2^{30} = 1 billion) may have occurred by chance alone.

Score, Bit-score, P-value, E-value

P-value: Probability that an event occurs by chance.

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

P-val (S) = P(x ≥ S)

$$\begin{aligned} Pval_S^{MSP} &= Ke^{-\lambda S} \\ &= Ke^{-\ln(2)S' + \ln(K)} \\ &= 2^{-S'} \end{aligned}$$

This equation was derived from the EVD score distribution obtained from all pair alignments (see course).

E-value (Expectation value): correction of the *p-value* for multiple testing.

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

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

*E-val (S) = P-val (S) * N* where *N* is the size of the *search space* (*N = n*m* where *n* is the length of the query sequence and *m* is the length of the database).

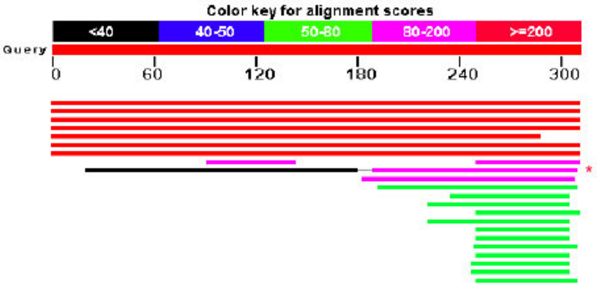
Score, Bit-score, P-value, E-value: example

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

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

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

BLAST (default parameters)



Results (output) of BLAST

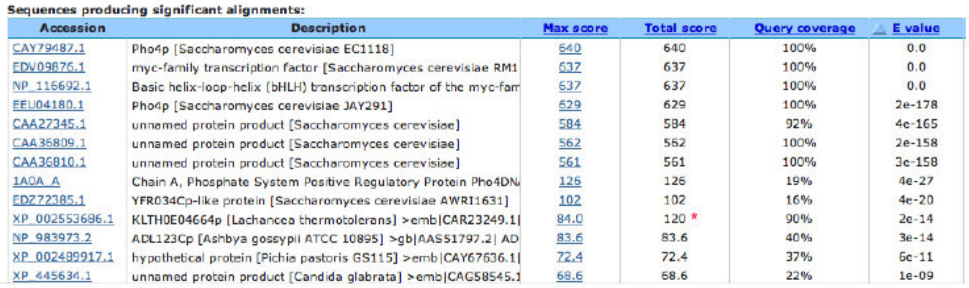
- The top segment displays the color key and the query based scale.
- The colored bars represent the actual HSPs. The position of each bar indicates the region of the query the HSP covers.
- The thin line (see *) indicates that the two HSPs are from the same sequence.
- Small vertical lines (not obtained here) indicate breaks, i.e., segments which are not connected in the actual alignment.

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

Score, Bit-score, P-value, E-value: example

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

Results (output) of BLAST



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
CAY79487.1	Pho4p [Saccharomyces cerevisiae EC1118]	640	640	100%	0.0
EDV09876.1	myc-family transcription factor [Saccharomyces cerevisiae RM1	637	637	100%	0.0
NP_116692.1	Basic helix-loop-helix (bHLH) transcription factor of the myc-far	637	637	100%	0.0
EEU04180.1	Pho4p [Saccharomyces cerevisiae JAY291]	629	629	100%	2e-178
CAA27345.1	unnamed protein product [Saccharomyces cerevisiae]	584	584	92%	4e-165
CAA36809.1	unnamed protein product [Saccharomyces cerevisiae]	562	562	100%	2e-158
CAA36810.1	unnamed protein product [Saccharomyces cerevisiae]	561	561	100%	3e-158
1A0A_A	Chain A, Phosphate System Positive Regulatory Protein Pho4DN	126	126	19%	4e-27
EDZ72385.1	YFR034Cp-like protein [Saccharomyces cerevisiae AWRI1631]	102	102	16%	4e-20
XP_002553686.1	KLTH0E04664p [Lachancea thermotolerans] >emb|CAR23249.1|	84.0	120 *	90%	2e-14
NP_983973.2	ADL123Cp [Ashbya gossypii ATCC 10895] >gb|AAS51797.2| AD	83.6	83.6	40%	3e-14
XP_002489917.1	hypothetical protein [Pichia pastoris GS115] >emb|CAY67636.1|	72.4	72.4	37%	6e-11
XP_445634.1	unnamed protein product [Candida glabrata] >emb|CAG58545.1	68.6	68.6	22%	1e-09

Max score = highest alignment score (bit-score) between the query sequence and the database sequence segment .

Total score = sum of alignment scores of all segments from the same database sequence that match the quary sequence (calculated over all segments). This score is different from the max score if several parts of the database sequence match different parts of the query sequence (see " * " in the example).

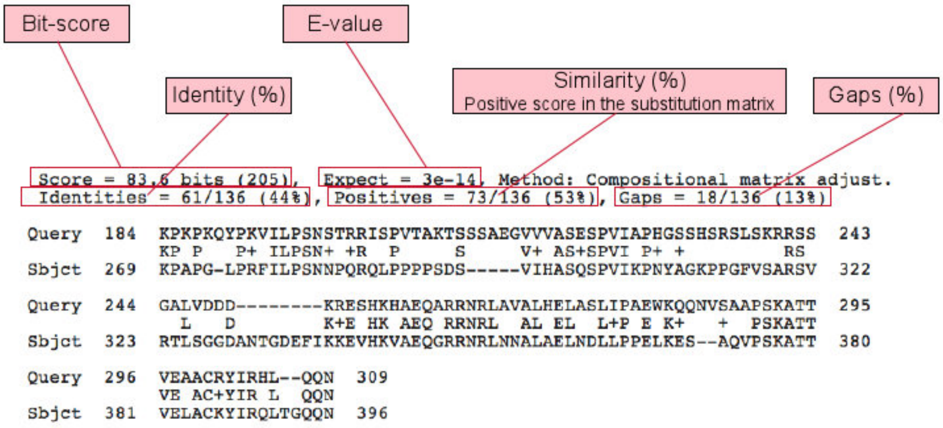
Query coverage = percent of the query length that is included in the aligned segments. This coverage is calculated over all segments (cf. total score).

E-value = number of alignments expected by chance with a particular score or better. The expect value is the default sorting metric and normally gives the same sorting order as Max score.

Score, Bit-score, P-value, E-value: example

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

Results (output) of BLAST

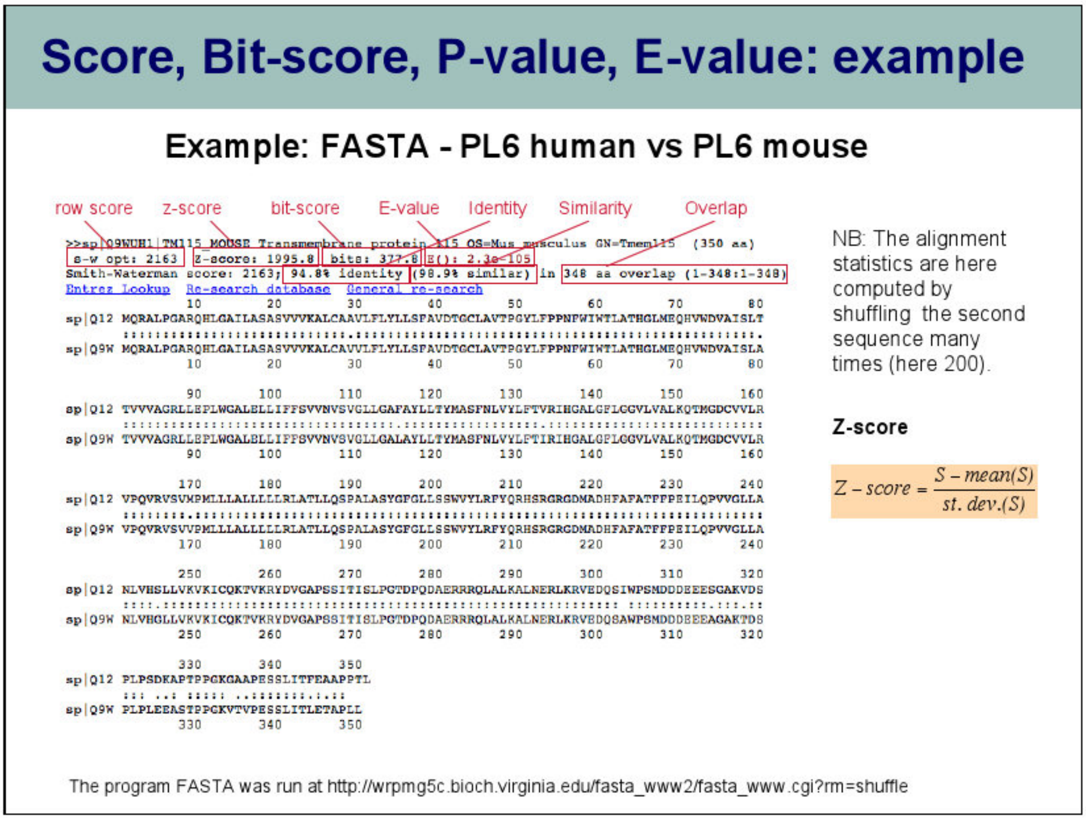


```
 Score = 83.6 bits (205),  Expect = 3e-14, Method: Compositional matrix adjust.
 Identities = 61/136 (44%), Positives = 73/136 (53%), Gaps = 18/136 (13%)

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

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

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



Score, Bit-score, P-value, E-value: example

Example: FASTA - PL6 human vs PL6 mouse

```
>>sp|Q9WUH1|TM115_MOUSE Transmembrane protein 115 OS=Mus musculus GN=Tmem115  (350 aa)
 s-w opt: 2163  Z-score: 1995.8  bits: 377.8 E(): 2.3e-105
Smith-Waterman score: 2163;  94.8% identity (98.9% similar) in 348 aa overlap (1-348:1-348)
Entrez Lookup  Re-search database  General re-search
               10        20        30        40        50        60        70        80
sp|Q12 MQRALPGARQHLGAILASASVVVKALCAAVLFLYLLSFAVDTGCLAVTPGYLFPPNFWIWTLATHGLMEQHVWDVAISLT
       ::::::::::::::::::::::::::::.::::::::::::::::::::::::::::::::::::::::::::::::::.
sp|Q9W MQRALPGARQHLGAILASASVVVKALCAVVLFLYLLSFAVDTGCLAVTPGYLFPPNFWIWTLATHGLMEQHVWDVAISLA
               10        20        30        40        50        60        70        80

               90       100       110       120       130       140       150       160
sp|Q12 TVVVAGRLLEPLWGALELLIFFSVVNVSVGLLGAFAYLLTYMASFNLVYLFTVRIHGALGFLGGVLVALKQTMGDCVVLR
       ::::::::::::::::::::::::::::::::::.:::::::::::::::::.:::::::::::::::::::::::::::
sp|Q9W TVVVAGRLLEPLWGALELLIFFSVVNVSVGLLGALAYLLTYMASFNLVYLFTIRIHGALGFLGGVLVALKQTMGDCVVLR
               90       100       110       120       130       140       150       160

              170       180       190       200       210       220       230       240
sp|Q12 VPQVRVSVMPMLLLALLLLLRLATLLQSPALASYGFGLLSSWVYLRFYQRHSRGRGDMADHFAFATFFPEILQPVVGLLA
       ::::::::.:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
sp|Q9W VPQVRVSVVPMLLLALLLLLRLATLLQSPALASYGFGLLSSWVYLRFYQRHSRGRGDMADHFAFATFFPEILQPVVGLLA
              170       180       190       200       210       220       230       240

              250       260       270       280       290       300       310       320
sp|Q12 NLVHSLLVKVKICQKTVKRYDVGAPSSITISLPGTDPQDAERRRQLALKALNERLKRVEDQSIWPSMDDDEEESGAKVDS
       ::::.::::::::::::::::::::::::::::::::::::::::::::::::::::::: ::::::::::.:::.::
sp|Q9W NLVHGLLVKVKICQKTVKRYDVGAPSSITISLPGTDPQDAERRRQLALKALNERLKRVEDQSAWPSMDDDEEEAGAKTDS
              250       260       270       280       290       300       310       320

              330       340       350
sp|Q12 PLPSDKAPTPPGKGAAPESSLITFEAAPPTL
       ::: ..: ::::: ..:::::::.:.::
sp|Q9W PLPLEEASTPPGKVTVPESSLITLETAPLL
              330       340       350
```

NB: The alignment statistics are here computed by shuffling the second sequence many times (here 200).

Z-score

$$Z-score = \frac{S - mean(S)}{st.\,dev.(S)}$$

The program FASTA was run at http://wrpmg5c.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle

[blastpgp][http://etutorials.org/Misc/blast/Part+V+BLAST+Reference/Chapter+13.+NCBI-BLAST+Reference/13.8+blastpgp+Parameters+PSI-BLAST+and+PHI-BLAST/]

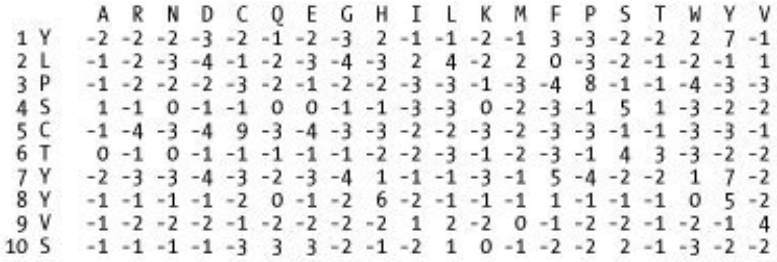
`blastpgp`用于PSI-BLAST和PHI-BLAST搜索. 对于标准的blasts, 这两个程序为更为敏感的蛋白blast搜索程序.

PSI-BLAST再搜索显著性hits时考虑位置特异性信息; PHI-BLAST使用pattern, 或profile来搜索比对.

PSI-BLAST

Position-specific iterated blast(`psiblast`), 通过制定的打分矩阵(scoring matrix)将query序列中的每个位置給予分值(基于搜索的连续迭代定义的比对情况). 指定的矩阵为位置特异性打分矩阵(position-specific scoring matrix, PSSM), 针对该位置的每一个氨基酸给予一个分值:

Figure 13-1. PSSM for the first 10 amino acids of the coelacanth HoxA11 protein



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

如图, 根据coelacanth Hoxa11蛋白(AAG39070)计算而来的PSSM. 查询序列位于左侧列, 每行是针对20种氨基酸给予的位置特异性分值. 针对1/7/8行的Y, 若是常规的blast算法, 这3个位置的Y将会拥有相同的分值.

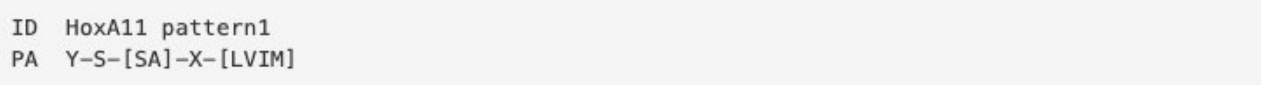
PSSM, 或为checkpoint文件, 是由PSI-BLAST内部生成的, 同时也可以使用参数`-c`导出为一个文件. 该参数很有用. 该checkpoint文件用于随后的PSI-BLAST(`blastpgp`)搜索, 或作为RPS-BLAST程序的数据库输入文件. 同时也用在blastall(a specifialized tblastn serach)搜索, `-p psitblastn`, 使用`-R <checkpoint file>`.

运行PSI-BLAST时, 参数`-j`需设置为大于1的值. 默认的`-j 1`表示不实用迭代搜索, 和单个BLASTP搜索一样. 通过`-j`参数指定最大的迭代数目, 程序会在出现convergence时停止. 当没有新的序列发现优于`-h`设置的阈值时, 搜索终止.

```
blastpgp -d nr -i my_protein -s T -j 5
blastpgp -d nr -i my_protein -R my_protein.ckp -d nr -j 5 -h 0.001
```

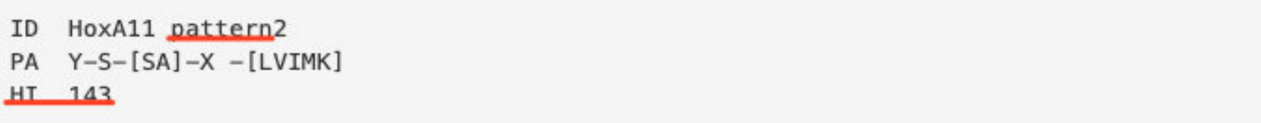
PHI-BLAST

Pattern-hit initiated BLAST, 使用输入序列定义pattern, 然后在蛋白数据中查询. 使用[PROSITE][http://ca.expasy.org/prosite/]格式定义pattern, 且用于比对的seed. 不同于用于seeding比对的words:



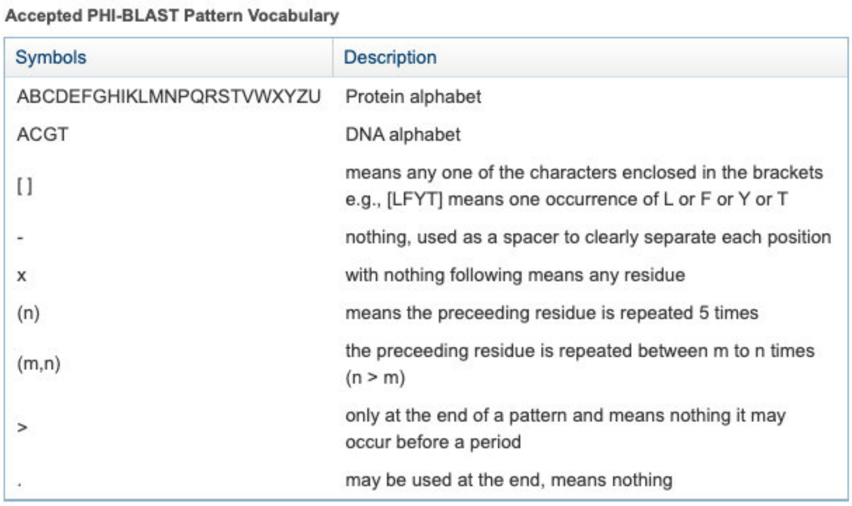
```
ID  HoxA11 pattern1
PA  Y-S-[SA]-X-[LVIM]
```

以ID起始的行, 随后两空格跟随pattern的名称. 下一行以PA起始, 随后两个空格, 接着为PROSITE格式pattern. dash(-)用于分隔字符, X表示任意字符, 方括号指定一个氨基酸的选择. 如果一个pattern出现超过一次, 可能需要限制其发生发的次数, 可通过HI(hit initiation)标签指定它在pattern文件中的位置. 例如, 指定143位置出现的pattern经被使用:



```
ID  HoxA11 pattern2
PA  Y-S-[SA]-X -[LVIMK]
HI  143
```

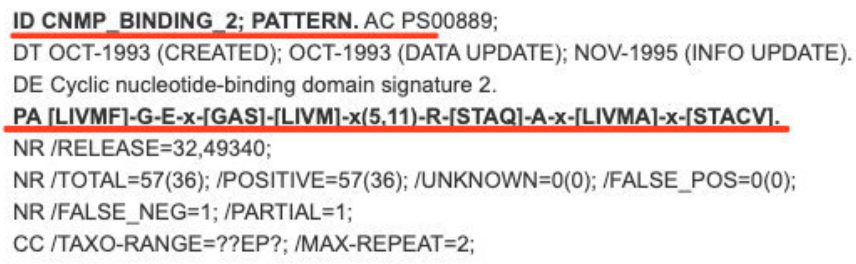
[规则][https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp#phi_pattern]:



Accepted PHI-BLAST Pattern Vocabulary

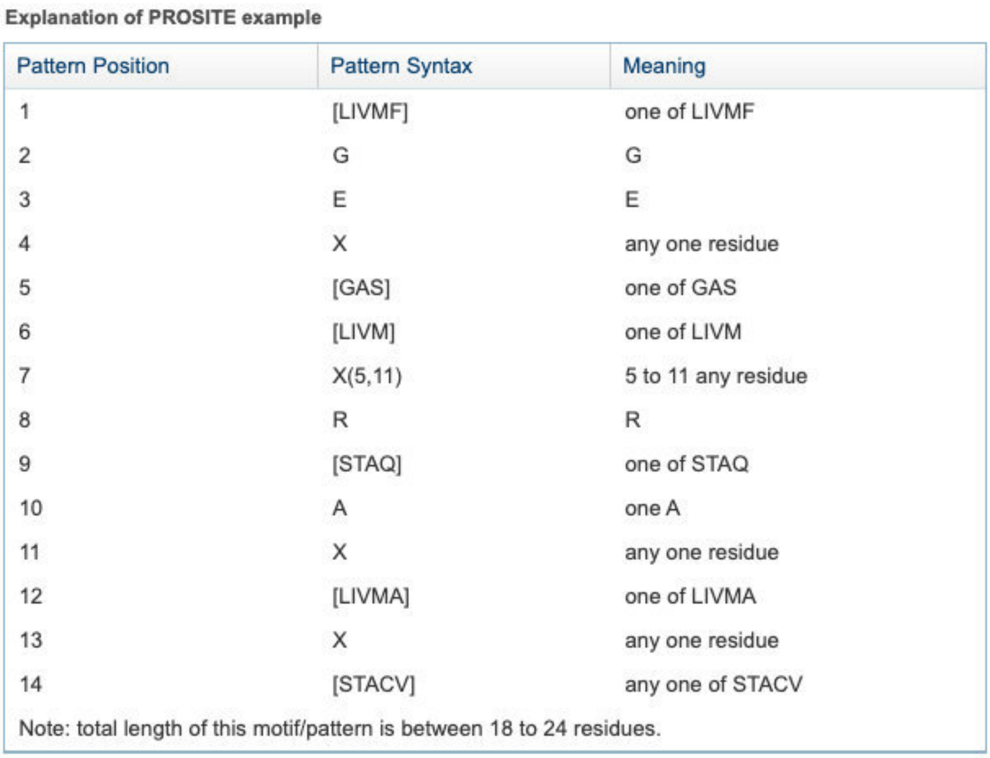
Symbols	Description
ABCDEFGHIKLMNPQRSTVWXYZU	Protein alphabet
ACGT	DNA alphabet
[ ]	means any one of the characters enclosed in the brackets e.g., [LFYT] means one occurrence of L or F or Y or T
-	nothing, used as a spacer to clearly separate each position
x	with nothing following means any residue
(n)	means the preceeding residue is repeated 5 times
(m,n)	the preceeding residue is repeated between m to n times (n > m)
>	only at the end of a pattern and means nothing it may occur before a period
.	may be used at the end, means nothing

例如:



```
ID CNMP_BINDING_2; PATTERN. AC PS00889;
DT OCT-1993 (CREATED); OCT-1993 (DATA UPDATE); NOV-1995 (INFO UPDATE).
DE Cyclic nucleotide-binding domain signature 2.
PA [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
NR /RELEASE=32,49340;
NR /TOTAL=57(36); /POSITIVE=57(36); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR /FALSE_NEG=1; /PARTIAL=1;
CC /TAXO-RANGE=??EP?; /MAX-REPEAT=2;
```

解释:

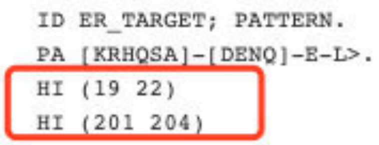


Explanation of PROSITE example

Pattern Position	Pattern Syntax	Meaning
1	[LIVMF]	one of LIVMF
2	G	G
3	E	E
4	X	any one residue
5	[GAS]	one of GAS
6	[LIVM]	one of LIVM
7	X(5,11)	5 to 11 any residue
8	R	R
9	[STAQ]	one of STAQ
10	A	one A
11	X	any one residue
12	[LIVMA]	one of LIVMA
13	X	any one residue
14	[STACV]	any one of STACV

Note: total length of this motif/pattern is between 18 to 24 residues.

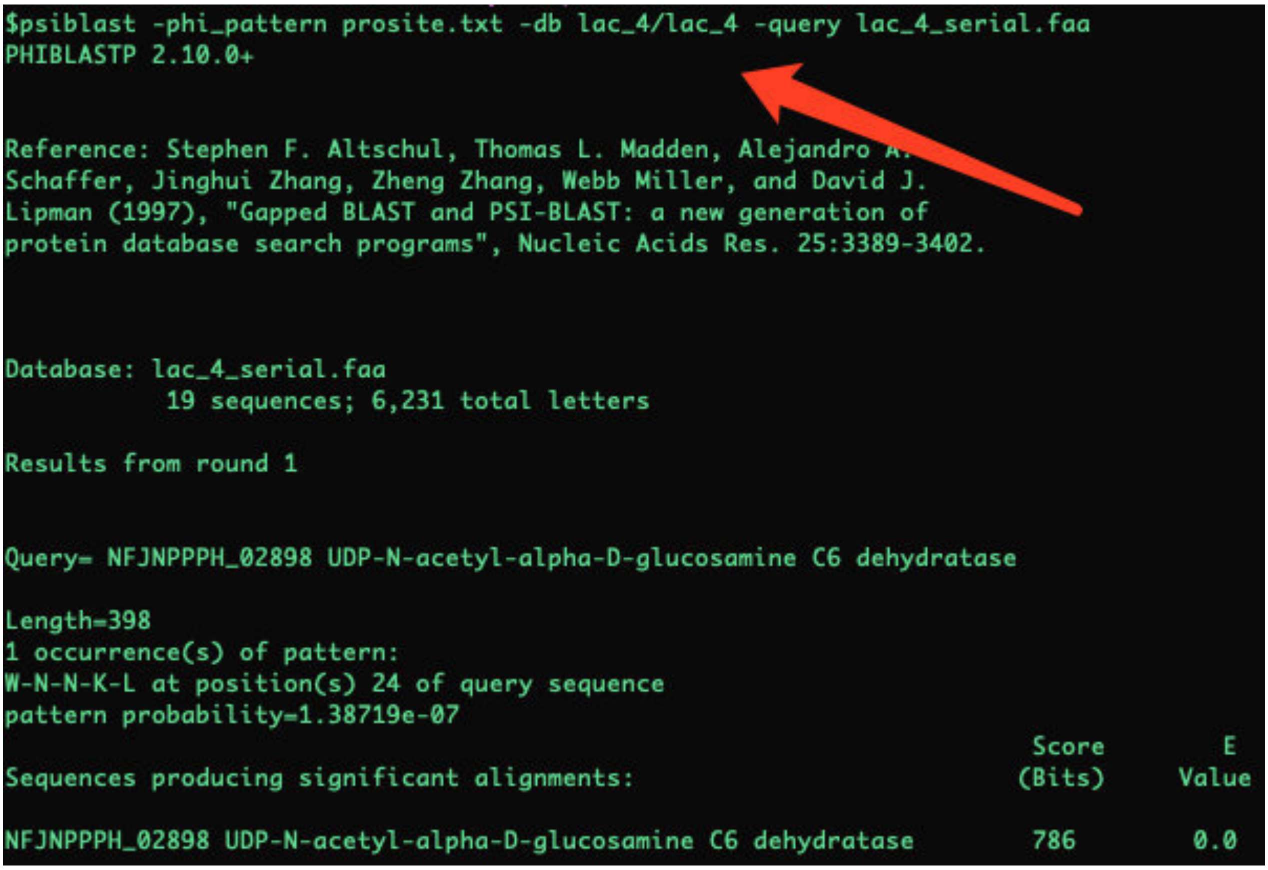
例如:



```
ID ER_TARGET; PATTERN.
PA [KRHQSA]-[DENQ]-E-L>.
HI (19 22)
HI (201 204)
```

解释: 使用HI指定pattern出现的位置区间一个为19-22; 第二个为201-204

[Practices][http://bioinf-hpc.ibun.unal.edu.co/cgi-bin/emboss/help/phiblast#input]



```
$psiblast -phi_pattern prosite.txt -db lac_4/lac_4 -query lac_4_serial.faa
PHIBLASTP 2.10.0+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: lac_4_serial.faa
           19 sequences; 6,231 total letters

Results from round 1


Query= NFJNPPPH_02898 UDP-N-acetyl-alpha-D-glucosamine C6 dehydratase

Length=398
1 occurrence(s) of pattern:
W-N-N-K-L at position(s) 24 of query sequence
pattern probability=1.38719e-07
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NFJNPPPH_02898 UDP-N-acetyl-alpha-D-glucosamine C6 dehydratase        786     0.0
```