Stupeň podobnosti dvou sekvencí

SEKVENCE A: MASAQSFYLL
SEKVENCE B: MASGQWLLAS

Které oblasti A a B jsou si nejvíce podobné ?

Jsou si A a B víc podobné než A a C ?

Která ze sekvencí X1,...,Xn je nejpodobnější A ?

Jaká je pravděpodobnost výskytu dané podobnosti v náhodné sekvenci ?

Existují už publikované homology sekvence A ?

Jaká je funkce A ?

Stupeň podobnosti dvou sekvencí

IDENTITA

```
MASAQSFYLL
||||||||||
MASAQSFYLL
```

SUBSTITUCE

```
MASAQSFYLL      MASAQSFYLL
||||||:|||      |||||| |||
MASAQSWYLL      MASAQSTYLL
```

INZERCE/DELECE

```
MASAQSFYLL
|||||| |||
MASAQS-YLL
```

TRANSPOZICE

```
MASAQSFYLL
|||||  |||
MASAQFSYLL
```

Stupeň podobnosti dvou sekvencí

Netriviální hodnocení substitucí
u proteinů
(matice PAM250)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Stupeň podobnosti dvou sekvencí

```
      A  G  A  T  A
   0 -2 -4 -6 -8 -10
A -2  2  0 -2 -3 -5
G -4 -1  ?
T -6
C -8
A -10

INDEL=-2 IDENT=2 SUBST=-1

AGATA
||  |    S=4
AGTCA
```

BLAST (basic local alignment search tool)

Co když jsou sekvence dlouhé a máme jich několik milionů ?

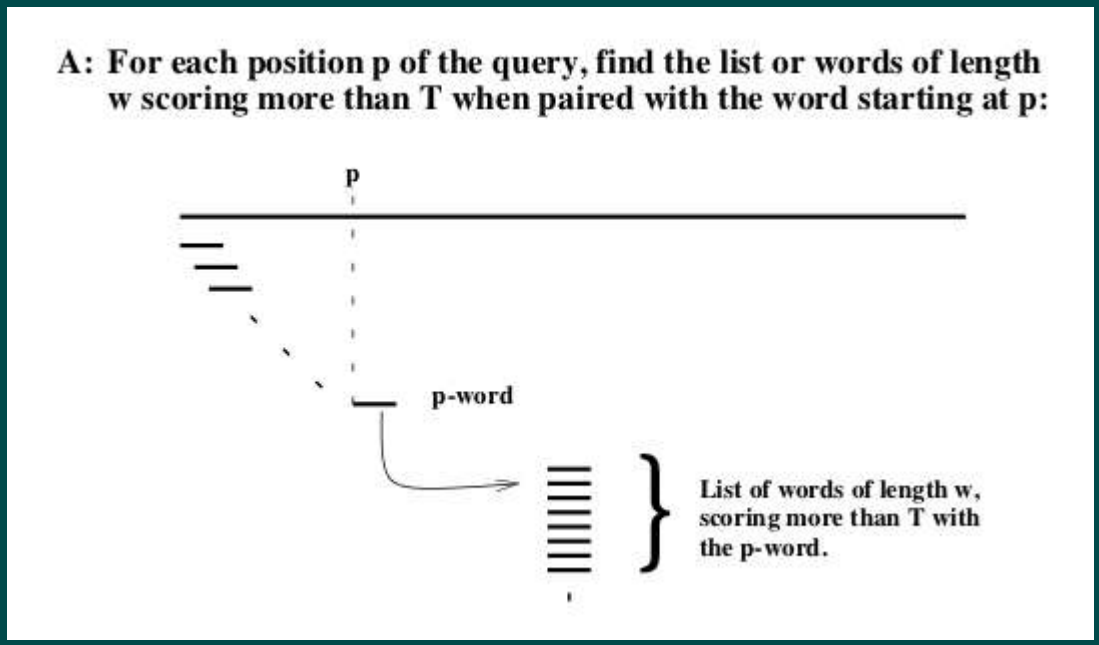
DP nestačí, výpočty trvají příliš dlouho. Alternativou výpočtu by byl předpočítaný soubor výskytu různých slov v databázi (index). Problémem indexu je, že je pro dlouhá slova nezvladatelný objemově. Existuje např.

$20^8 = 25\ 600\ 000\ 000$

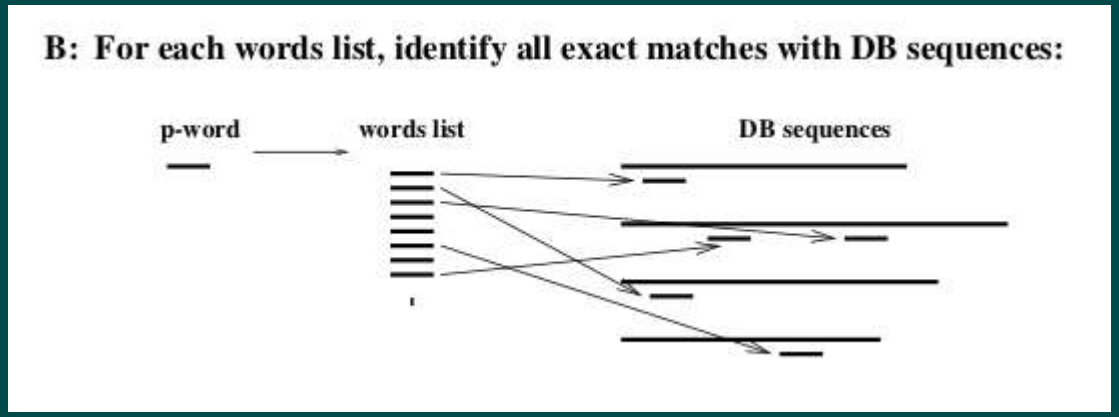
různých uspořádání osmi aminokyselin v řetězci, několik způsobů hodnocení podobnosti atd.

Kompromisem je heuristické řešení. Nalezení tzv. “seeds”, výskytu krátkých řetězců a hledání podobnosti DP algoritmem jenom v jejich blízkosti.

BLAST (basic local alignment search tool)

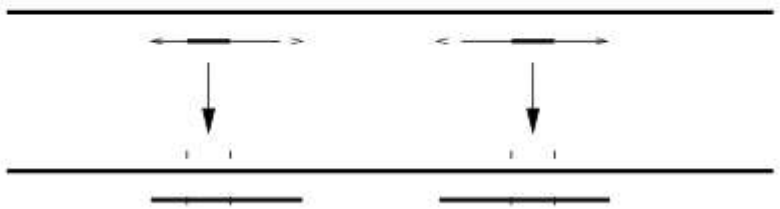


BLAST (basic local alignment search tool)

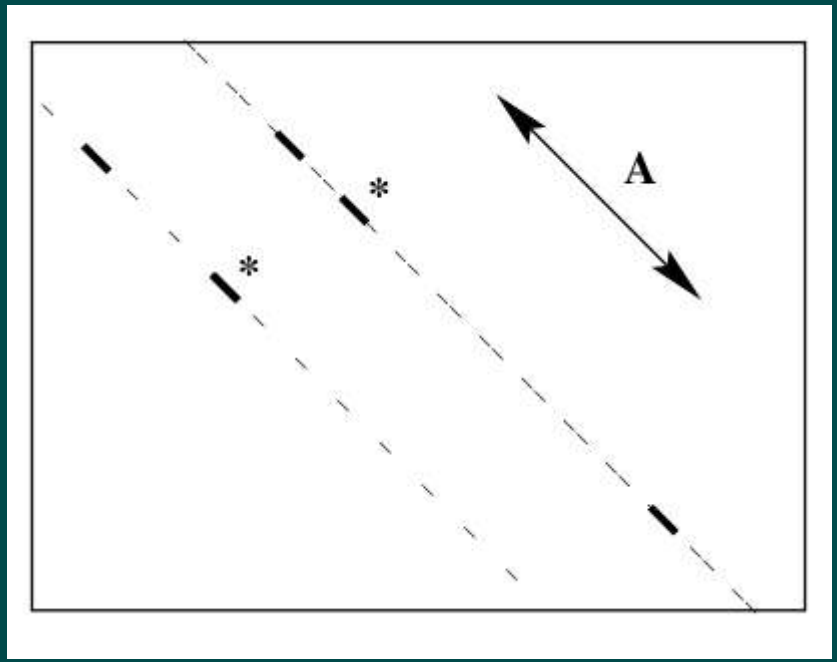


BLAST (basic local alignment search tool)

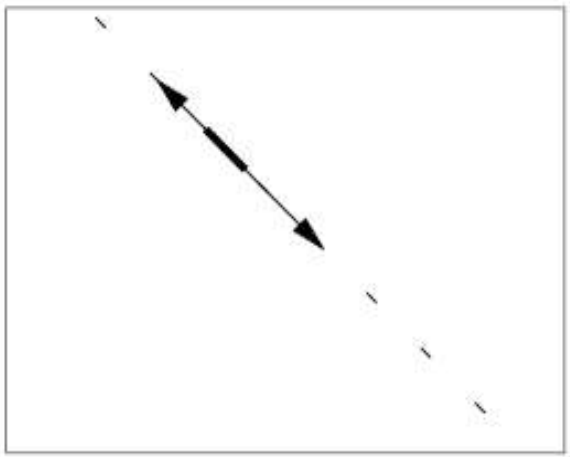
C: For each word match («hit»), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.



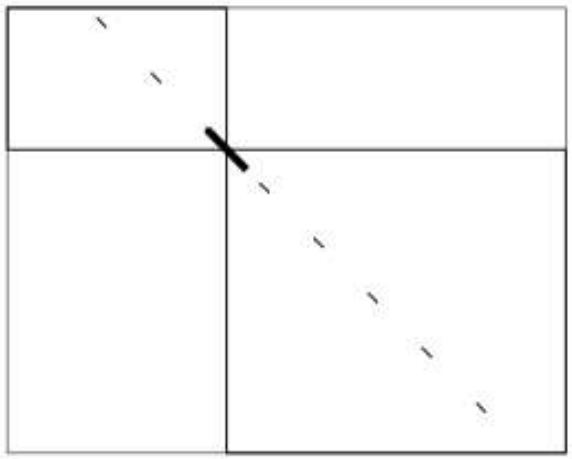
BLAST (basic local alignment search tool)



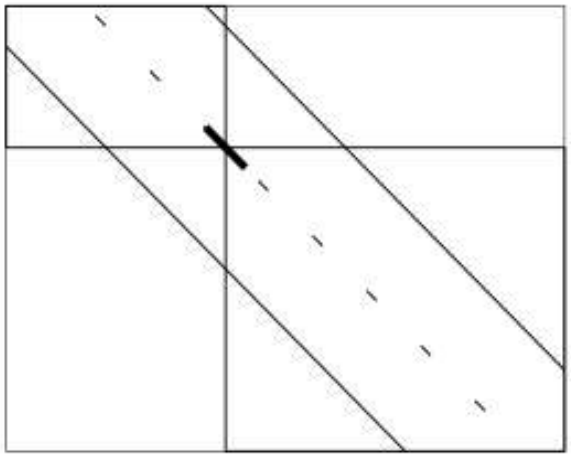
BLAST (basic local alignment search tool)



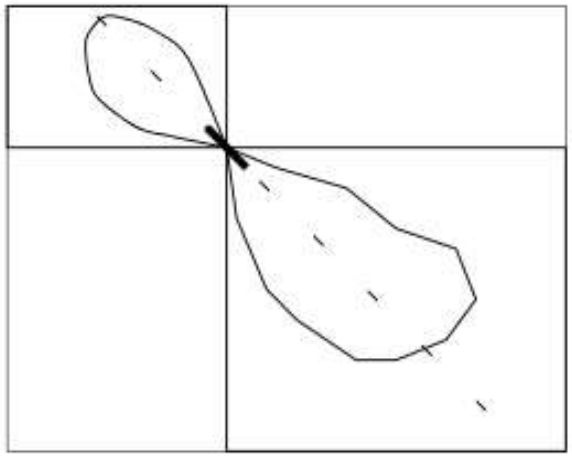
Ungapped extension



Gapped extension by full DP

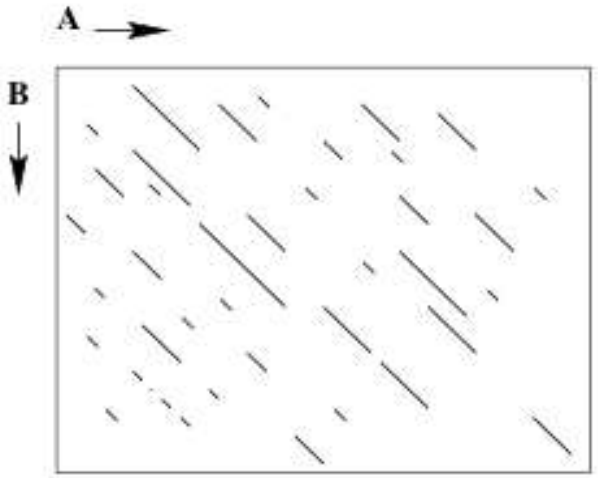


Gapped extension by «banded DP»

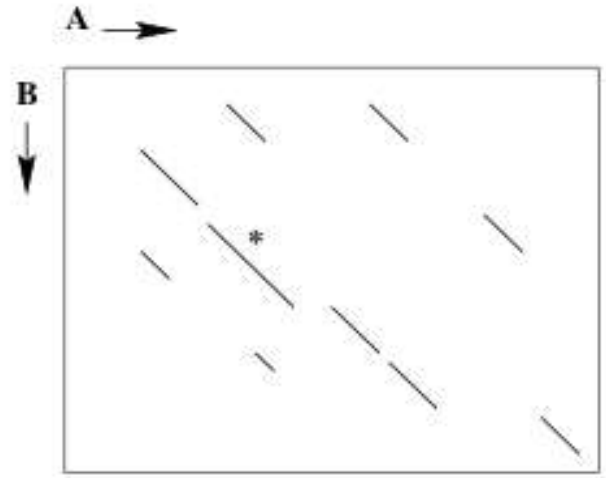


Gapped extension by «score-limited DP»

FASTA

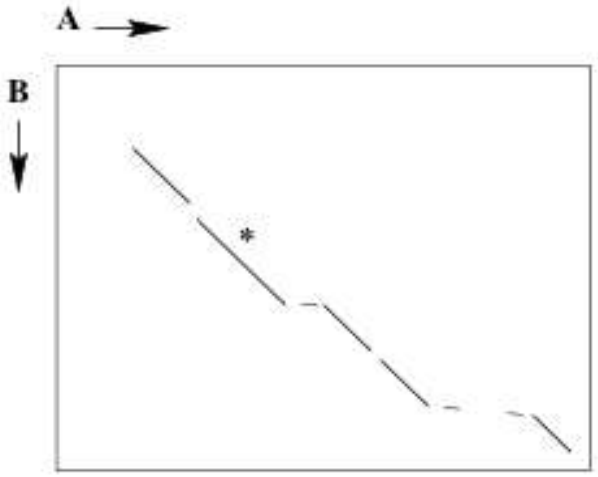


Identify all k-tuple matches



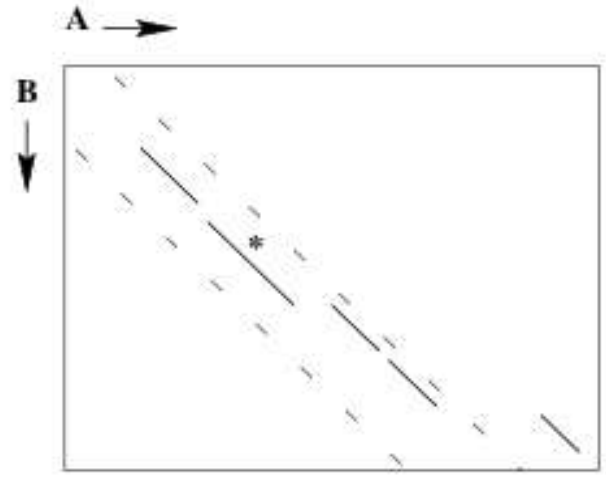
Re-score the 10 best scoring regions using a scoring matrix

→ Init1 score



Apply joining procedure

→ Initn score



Apply limited DP

→ Opt score

BLAST (basic local alignment search tool)

P-VALUE . . . P(score>S)

PRAVDĚPODOBNOST VÝSKYTU PODOBNOSTI VĚTŠÍ NEŽ S V NÁHODNÝCH SEKVENCÍCH URČITÉ DÉLKY

$$P(MSP(M,N)>S) = 1 - \exp(-Kmn.\exp(-lambda.S))$$

E-VALUE

OČEKÁVANÝ POČET PODOBNOSTÍ KDE score>S

$$Kmn.\exp(-lambda.S)$$

BLAST (basic local alignment search tool)

PAM150
Percent Accepted Mutations

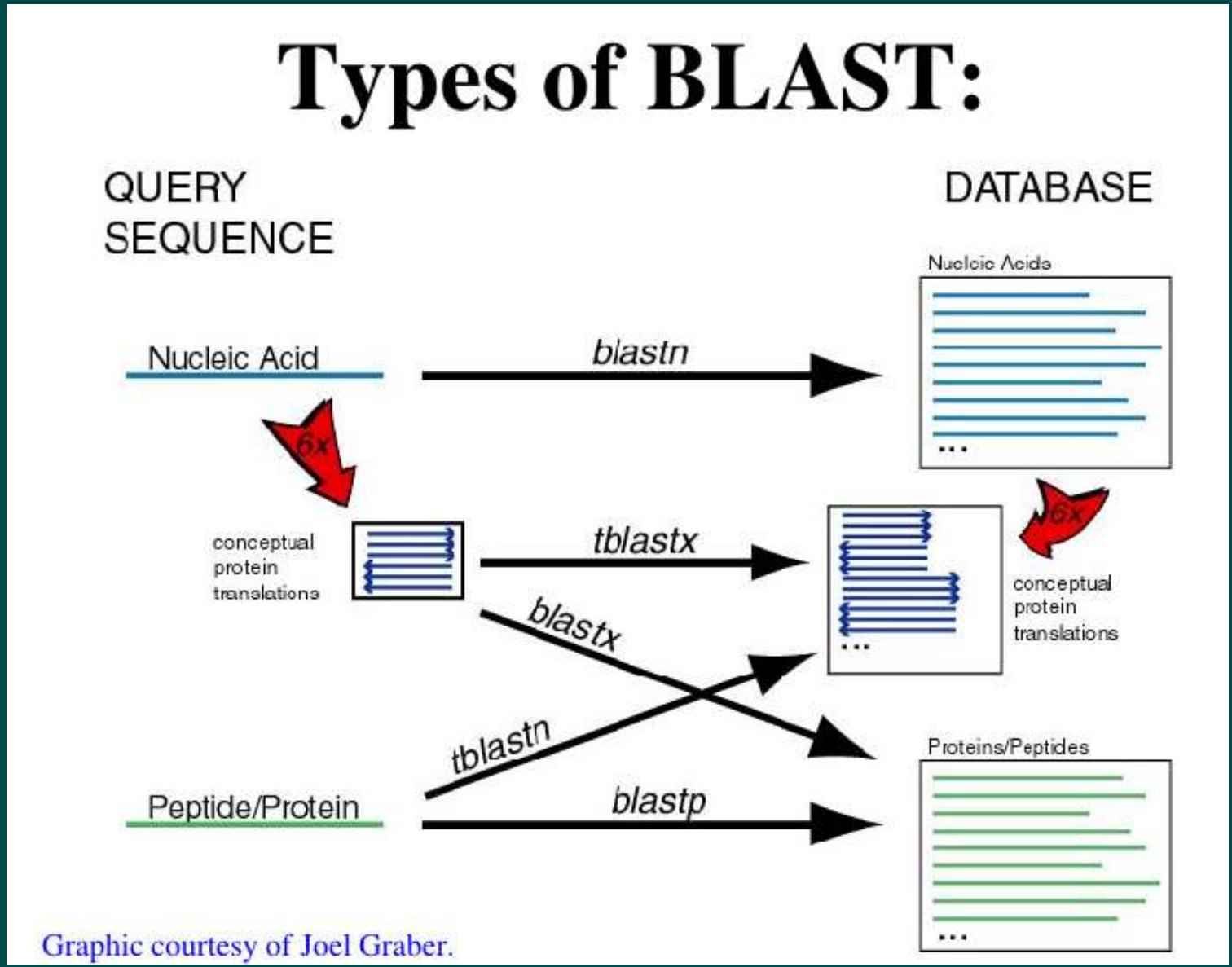
Substituční matice odvozena z předpokladu 150 mutací na 100 pozic sekvence

BLOSUM65
BLOck SUbstitution Matrix

Substituční matice odvozena ze sekvencí se 65% identitou

PAM120 <=> BLOSUM80
PAM250 <=> BLOSUM45

BLAST (basic local alignment search tool)



BLAST (basic local alignment search tool)

```
>gi|50757596|ref|XP_425354.1|similar to protein kinase
 Length = 613

 Score = 50.4 bits (119), Expect(2) = 2e-17
 Identities = 26/54 (48%), Positives = 36/54 (66%), Gaps = 1/54 (1%)

Query: 740 YVMVLEYANEGNLREYLEK-KFDTLQWENKIQMALDITRGLLCLHSRNIIHRDL 582
           Y +V EY +EG+LR YL K +  +L  +  I  ALDI RG+  +HS+ +IHRDL
Sbjct: 250 YCVVTEYLSEGSLRAYLHKLERKSLPLQKLIAFALDIARGMEYIHSQGVIHRDL 303
```

BLAST (basic local alignment search tool)

BLAST (NCBI-BLAST WU-BLAST)
   BLASTN BLASTP BLASTX TBLASTN TBLASTX
MEGABLAST
PSI-BLAST
PHI-BLAST
SNPBLAST
BLASTZ

BLAST (basic local alignment search tool)

BLAT
SESAM
PATTERN_HUNTER
PSST
PRIMEX