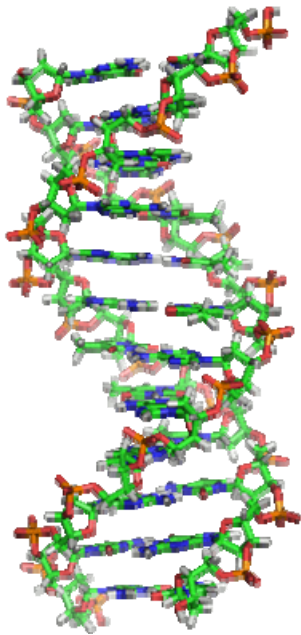


Basic Local Alignment and Search Tool (BLAST)



Database searching

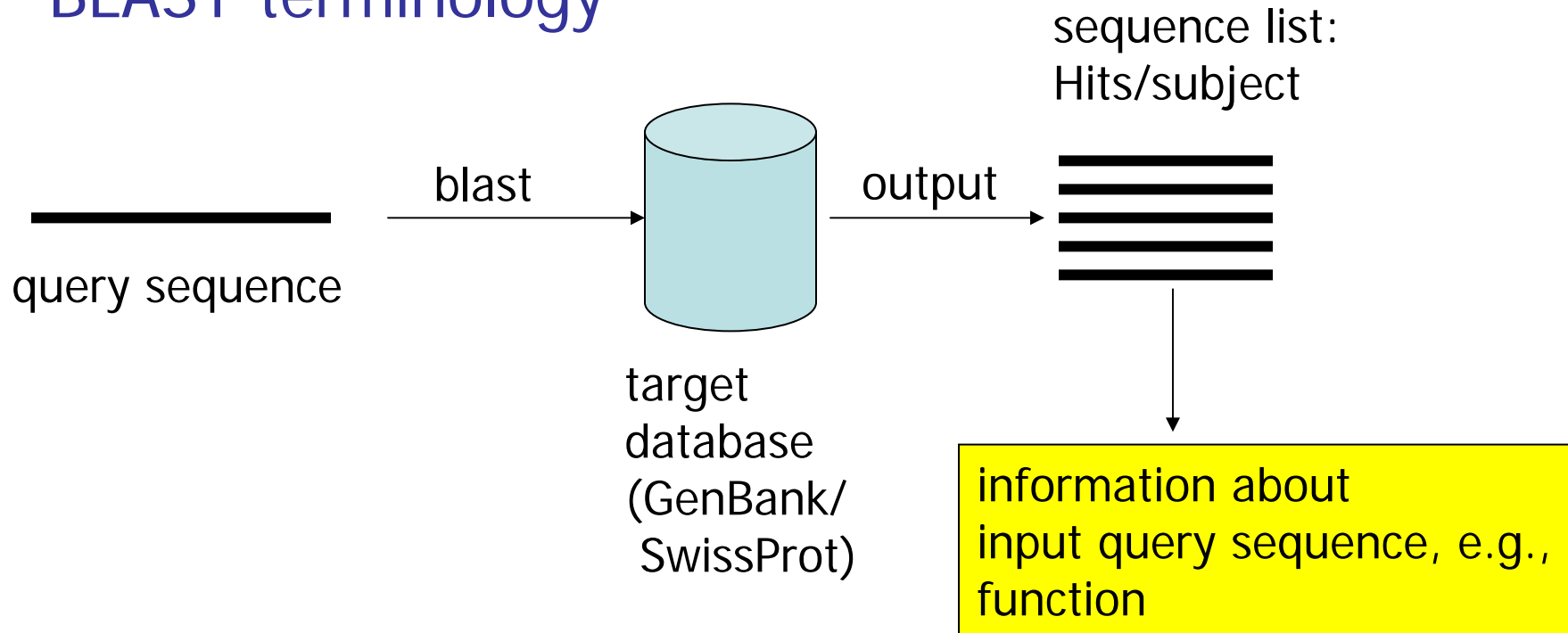
10-02-2009

Barbera van Schaik

Why use BLAST?

- Dynamic Programming is not suitable for comparing a query sequence against a database
 - Takes too much time!
- BLAST is a heuristic method to find the highest locally optimal alignments
 - BLAST improved overall speed of searches
 - BLAST maintains good sensitivity

BLAST terminology



The aim of a database (blast) search is to discover sequence homology on basis of sequence similarity

BLAST returns similar sequences, not necessarily biological similar sequences

BLAST variants

Sequence type	nucleotide database	protein database
nucleotide query	blastn/tblastx	blastx
amino acid query	tblastn	blastp

blastn: finds NT sequences similar to your NT sequence

blastp: finds AA sequences similar to your AA sequence

blastx: finds AA sequences similar to translation of your NT sequence (if you cannot recognize an ORF)

tblastn: translate AA sequence and searches against NT database (for finding pseudogenes)

tblastx: keep computers busy

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

Basic BLAST

Choose a BLAST program to run.

[nucleotide blast](#)

Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast

[protein blast](#)

Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast

[blastx](#)

Search **protein** database using a **translated nucleotide** query

[tblastn](#)

Search **translated nucleotide** database using a **protein** query

[tblastx](#)

Search **translated nucleotide** database using a **translated nucleotide** query

Web interface changes now and then

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

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- ▣ [Human](#)
- ▣ [Mouse](#)
- ▣ [Rat](#)
- ▣ [Arabidopsis thaliana](#)
- ▣ [Oryza sativa](#)
- ▣ [Bos taurus](#)
- ▣ [Danio rerio](#)
- ▣ [Drosophila melanogaster](#)
- ▣ [Gallus gallus](#)
- ▣ [Pan troglodytes](#)
- ▣ [Microbes](#)
- ▣ [Apis mellifera](#)

BLAST

[Help](#)

The Map Viewer provides a wide variety of genome mapping and sequencing data. [More..](#)

▼ Vertebrates (16)			
▼ Mammals (14)			
▼ Primates (3)			
Scientific name	Common name	Build	Tools
<i>Homo sapiens</i>	human	Build 36.3	Q B Cr G
		Build 35.1	Q B Cr
<i>Macaca mulatta</i>	rhesus macaque	Build 1.1	Q B G
<i>Pan troglodytes</i>	chimpanzee	Build 2.1	Q B G
▼ Rodents (2)			
Scientific name	Common name	Build	Tools
<i>Mus musculus</i>	laboratory mouse	Build 37.1	Q B Cr G
		Build 36.1	Q B
<i>Rattus norvegicus</i>	rat	RGSC v3.4	Q B G
▶ Monotremes (1)			
▶ Marsupials (1)			
▶ Other Mammals (7)			
▶ Other Vertebrates (2)			
▶ Invertebrates (12)			
▶ Protoczoa (18)			

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
 - Search [trace archives](#)
 - Find [conserved domains](#) in your sequence (cds)
 - Find sequences with similar [conserved domain architecture](#) (cdart)
 - Search sequences that have [gene expression profiles](#) (GEO)
 - Search [immunoglobulins](#) (IgBLAST)
 - Search for [SNPs](#) (snp)
 - Screen sequence for [vector contamination](#) (vecscreen)
 - [Align](#) two sequences using BLAST (bl2seq)
 - Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
-

BLASTing a sequence at NCBI – programs

The screenshot shows the NCBI BLAST website. At the top, there is a navigation bar with the BLAST logo and the text "Basic Local Alignment Search Tool". Below this are tabs for "Home", "Recent Results", "Saved Strategies", and "Help". On the right side of the navigation bar, there is a "My NCBI" section with "Sign In" and "Register" links. The main content area is titled "NCBI/BLAST Home" and contains a description of BLAST: "BLAST finds regions of similarity between biological sequences. more...". Below this is a promotional banner for "Primer-BLAST" with a "Go" button. The "BLAST Assembled Genomes" section lists various species genomes for search, including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. The "Basic BLAST" section lists five programs: nucleotide blast, protein blast (circled in red), blastx, tblastn, and tblastx, each with a brief description and a list of algorithms. On the right side, there are two sidebar sections: "News" with a link to "Align Sequences with BLAST" and "Tip of the Day" with a link to "How to Search Custom Databases in Web-Blast Using Entrez Queries".

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

► NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New Designing or Testing PCR Primers? Try your search in **Primer-BLAST**. [Go](#)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- Human
- Mouse
- Rat
- Arabidopsis thaliana
- Oryza sativa
- Bos taurus
- Danio rerio
- Drosophila melanogaster
- Gallus gallus
- Pan troglodytes
- Microbes
- Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

News

[Align Sequences with BLAST](#)

A new BI2seq functionality has been added to the standard BLAST pages that allows you to align a query against a set of subject sequences.
2008-09-04 12:56:52

[More BLAST news...](#)

Tip of the Day

How to Search Custom Databases in Web-Blast Using Entrez Queries.

A powerful feature of the BLAST Web interface is the ability to limit BLAST searches to a subset of any database using a standard Entrez query. Skillful use of Entrez queries allows the equivalent of on-the-fly construction of databases of exact composition.

[More tips...](#)

BLASTing a sequence at NCBI – enter accession

BLAST *Basic Local Alignment Search Tool* My NCBI [Sign In](#) [Register](#)

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

▶ [NCBI/BLAST/blastp suite](#)

[blastn](#) **[blastp](#)** [blastx](#) [tblastn](#) [tblastx](#)

BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#) Query subrange

From To

Or, upload file [Browse...](#)

Job Title
Enter a descriptive title for your BLAST search

Blast 2 sequences

Choose Search Set

Database

Organism Optional
Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

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)

Database choice

Protein databases

Also good for protein coding nucleotide queries

Choose a non-redundant database

Nucleotide databases

Non-redundant database

Filter on organism / other Entrez query

BLASTing a sequence at NCBI – parameters

Choose a BLAST algorithm

BLAST Search **database nr** using **Blastp (protein-protein BLAST)**
 Show results in a new window

▼ Algorithm parameters

General Parameters

Max target sequences
Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Scoring Parameters

Matrix

Gap Costs

Compositional adjustments

Filters and Masking

Filter Low complexity regions

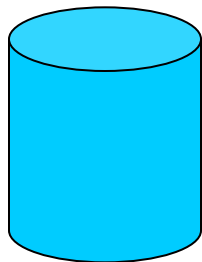
Mask Mask for lookup table only
 Mask lower case letters

Blast algorithm: step 1

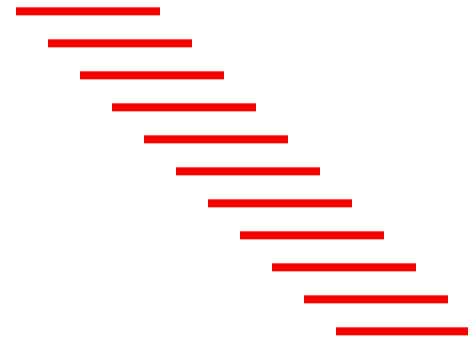
protein query sequence



protein database

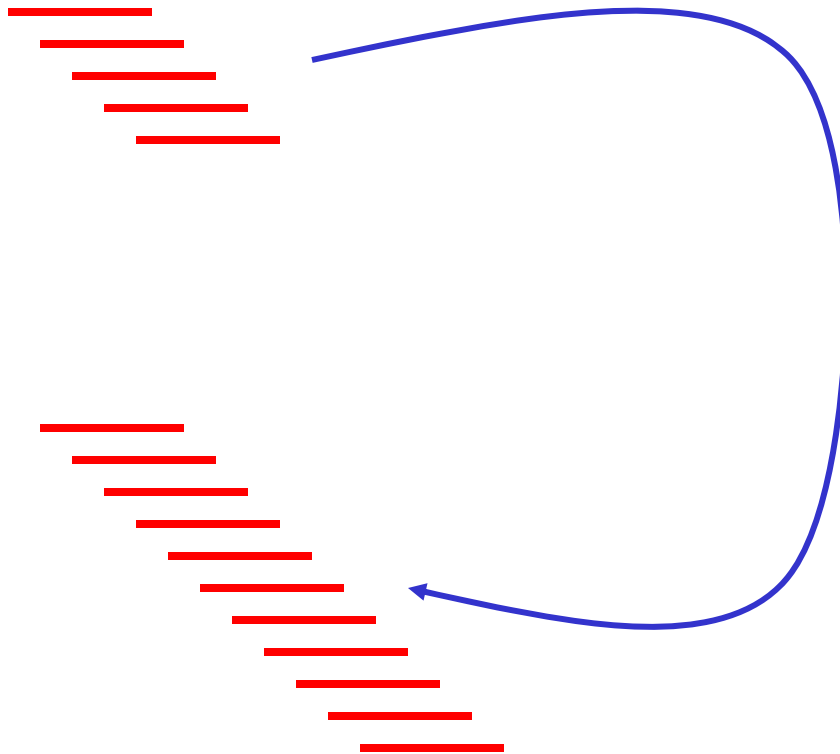


compile list of 'words'
of length W



Blast algorithm: step 2

Initial search



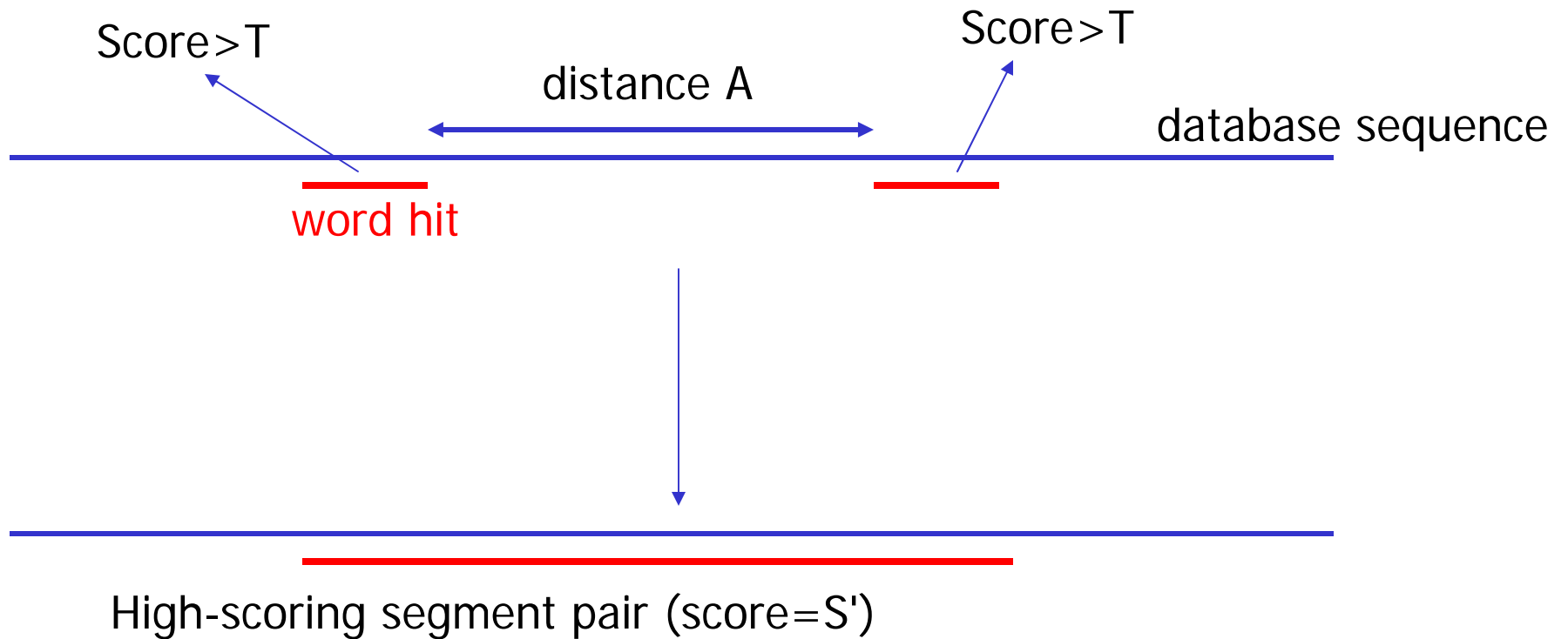
Use PAM/BLOSUM matrix

Find word of length 'W' that scores at least 'T' (T=11)

Exact matches only

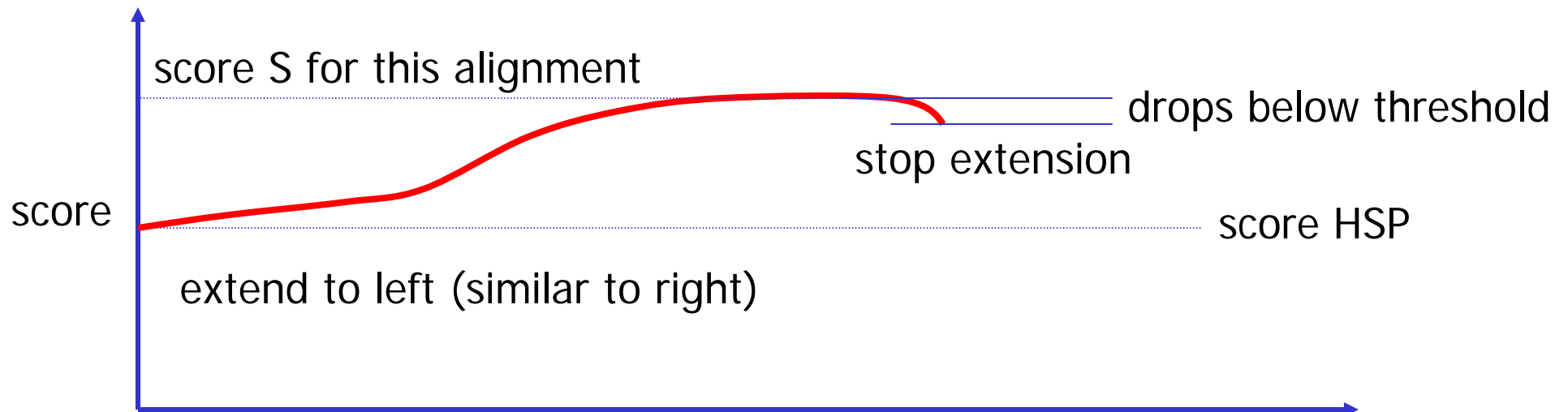
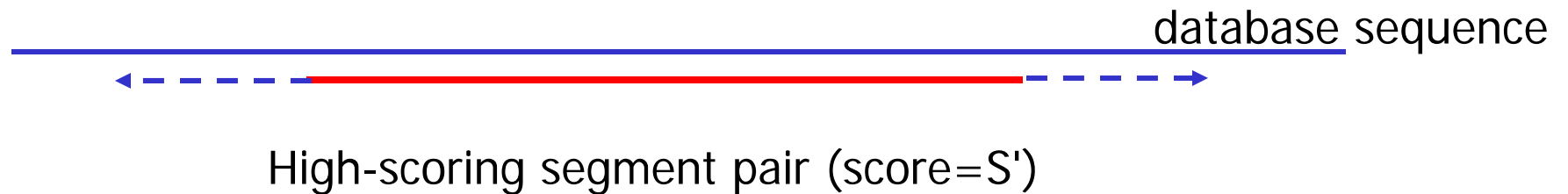
The parameter T dictates the speed and sensitivity
-increasing T increases speed,
decreases sensitivity

Join words on same diagonal (ungapped)



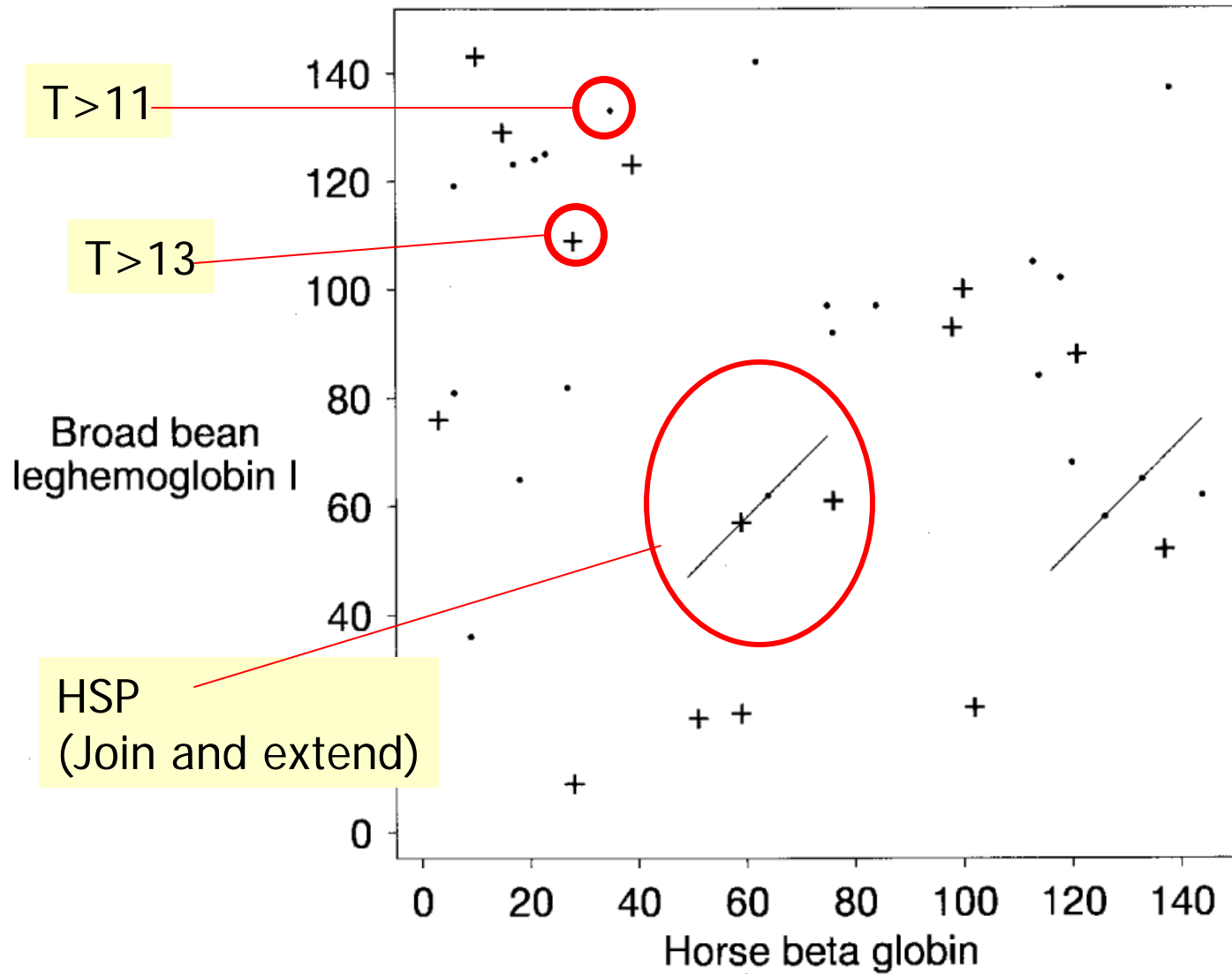
Join words on same diagonal

Extend HSP until score drops small amount below highest score of shorter alignment

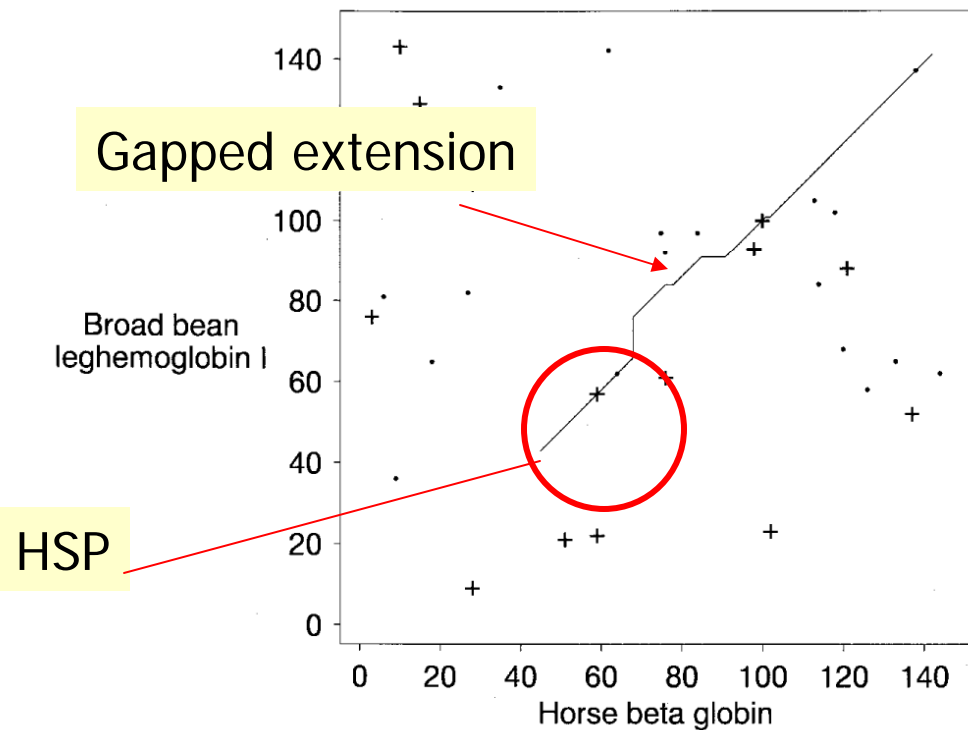


If $S > \text{threshold}$ (based on random sequences) then keep HSP

Finding HSP's



Trigger gapped extension



```

Leghemoglobin 43 FSFLKDSAGVVDSPKLGHAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS----- 90
                  F L + V+ +PK+ AH +KV                      L + GE V LD G+
Beta globin   45 FGDLSNPGAVMGNPKVKAHGKKV-----LHSFGEGVHHLNLDNLKGTFAALSE 90
    
```

```

Leghemoglobin 91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                  +H K +DP +F ++ L+ + G ++ EL A+++ G+A A+
Beta globin   91 LHCDKLHVDPENFRLLGNVLVVVLLARHFGKDFTPPELQASYQKVVAGVANAL 141
    
```

BLASTing a sequence at NCBI – parameters

Choose a BLAST algorithm

BLAST Search **database nr** using **Blastp (protein-protein BLAST)**
 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: 3

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

Masking of sequences – low complexity

Low complexity repeats in genome

Many amino-acid “stretches” in proteins

BLAST recognizes these regions as similar

but, they are NOT biologically related

Masking of sequences – highly abundant sequences

First query sequence against database that contains domains representative of large sequence families

- Alu repeats
- Protein kinase catalytic domains
- Vector sequences

Then mask these domains in the query sequence and continue search


Masking option replaces these regions with XXXXXXXX

When do you change the 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	The substitution matrix or the gap penalties
Your match has a borderline e-value	The substitution matrix or the gap penalties to check the match's robustness
BLAST reports too many matches	The database you're searching OR filter the reported entries by keyword OR increase the nr of reported matches OR increase Expect (the e-value threshold) OR reject sequences too similar to the query (those with very low e-values)

Parameters are already optimized



BLASTing a sequence at NCBI – parameters


Choose a BLAST algorithm 


BLAST Search **database nr** using **Blastp (protein-protein BLAST)**
 Show results in a new window


▼ **Algorithm parameters**

General Parameters


Max target sequences: 
Select the maximum number of aligned sequences to display 


Short queries: Automatically adjust parameters for short input sequences 


Expect threshold: 

Word size: 


Scoring Parameters



Matrix: 

Gap Costs: 

Compositional adjustments: 

Filters and Masking

Filter: Low complexity regions 

Mask: Mask for lookup table only 
 Mask lower case letters 

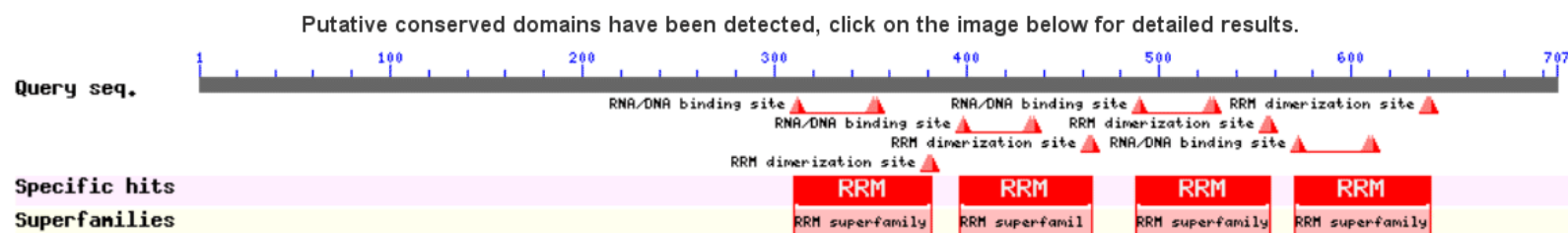
BLASTing a sequence at NCBI – job status

BLAST Basic Local Alignment Search Tool My NCBI [\[Sign In\]](#) [\[Register\]](#)

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

▶ [NCBI/BLAST/blastp suite/](#) [Formatting Results - SVTZB9YT011](#) [\[Formatting options\]](#)

Job Title: [gij128843|sp|P09405.2|NUCL_MOUSE](#) RecName:...



Request ID	SVTZB9YT011
Status	Searching
Submitted at	Sat Feb 7 14:58:35 2009
Current time	Sat Feb 7 14:58:48 2009
Time since submission	00:00:13

This page will be automatically updated in 6 seconds

If it takes too long: try another BLAST server

Country / continent	Program	URL
USA	BLAST / PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Europe	BLAST	http://www.expasy.ch/tools/blast/
Europe	BLAST	http://www.ch.embnet.org/software/bBLAST.html
Europe	BLAST	http://www.ebi.ac.uk/blast
Japan	BLAST / PSI-BLAST	http://blast.ddbj.nig.ac.jp/top-e.html

Warning: different database (versions) !

BLASTing a sequence at NCBI – blast summary

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - SVTZB9YT011

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

gi|128843|sp|P09405.2|NUCL_MOUSE RecName:...

Query ID	Id 22157	Database Name	nr
Description	gi 128843 sp P09405.2 NUCL_MOUSE RecName: Full=Nucleolin; AltName: Full=Protein C23	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid	Program	BLASTP 2.2.19+ Citation
Query Length	707		

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Related Structures](#)

Graphic Summary

Show Conserved Domains

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

Query seq. 1 100 200 300 400 500 600 707

RNA/DNA binding site RNA/DNA binding site RRM dimerization site

RNA/DNA binding site RRM dimerization site

RRM dimerization site RNA/DNA binding site

RRM dimerization site

Specific hits RRM RRM RRM RRM

Superfamilies RRM superfamily RRM superfamil RRM superfamily RRM superfamily

Distribution of 172 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Color key for alignment scores

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

Query 0 100 200 300 400 500 600 700

BLASTing a sequence at NCBI – used parameters

Other reports: [▼ Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#) [\[Related Structures\]](#)

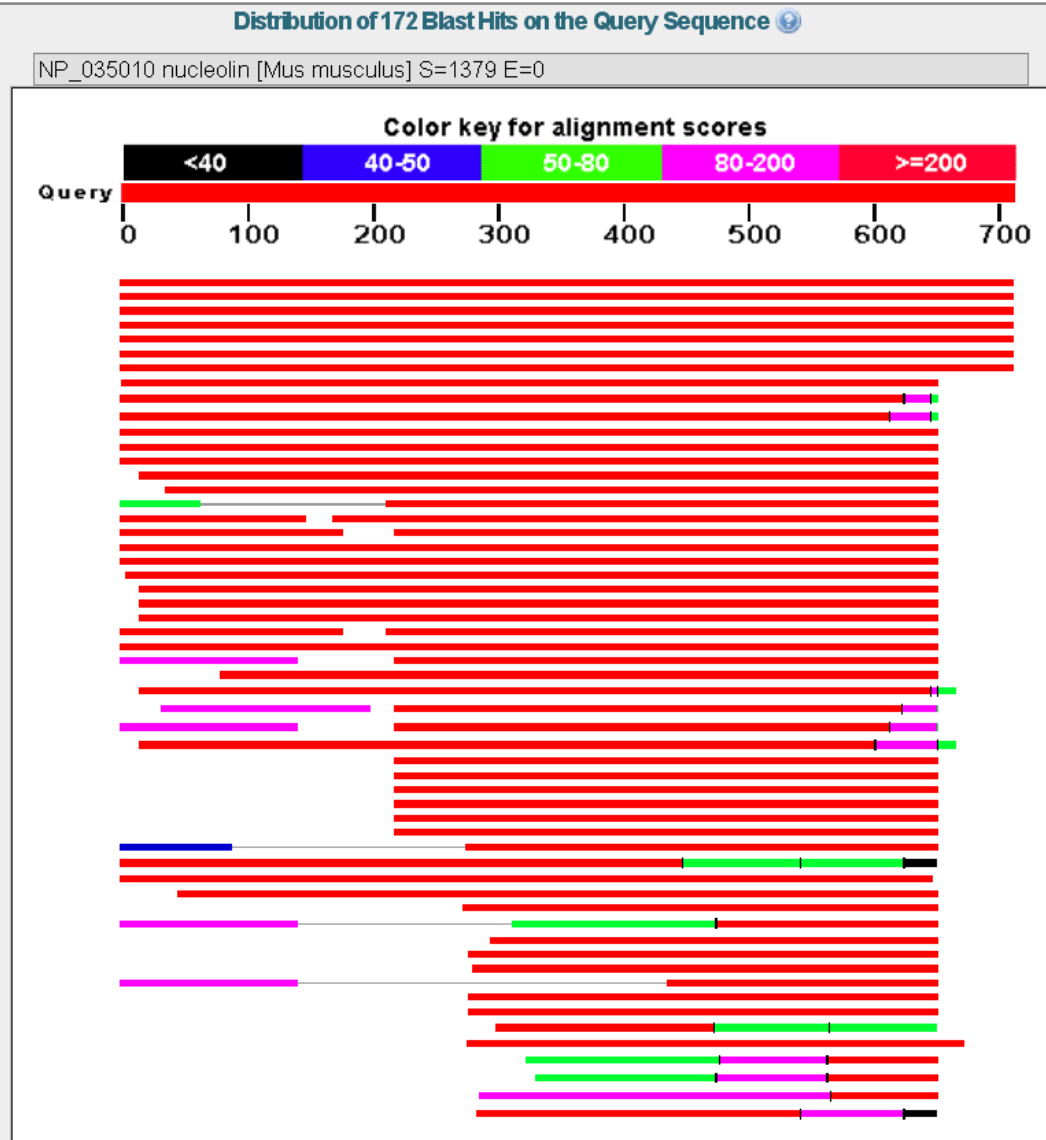
Search Parameters	
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Threshold	11
Composition-based stats	2
Filter string	F
Genetic Code	1
Window Size	40

Database	
Posted date	Feb 6, 2009 5:53 PM
Number of letters	2,699,408,701
Number of sequences	7,831,890
Entrez query	none

Karlin-Altschul statistics		
Params	Ungapped	Gapped
Lambda	0.302661	0.267
K	0.127079	0.041
H	0.344587	0.14

Results Statistics	
Length adjustment	143
Effective length of query	564
Effective length of database	1579448431
Effective search space	890808915084
Effective search space used	890808915084

BLASTing a sequence at NCBI – graphical display



BLASTing a sequence at NCBI – hit list

▼ Descriptions

Sequences producing significant alignments:

		Score (Bits)	E Value
ref NP_035010.3 	nucleolin [Mus musculus] >sp P09405.2 NUCL_M...	1379	0.0
dbj BAE36484.1 	unnamed protein product [Mus musculus]	1378	0.0
dbj BAE38940.1 	unnamed protein product [Mus musculus]	1378	0.0
dbj BAE40448.1 	unnamed protein product [Mus musculus] >dbj B...	1375	0.0
dbj BAC26311.1 	unnamed protein product [Mus musculus]	1373	0.0
gb AAH05460.1 	Nucleolin [Mus musculus]	1371	0.0
dbj BAC27474.1 	unnamed protein product [Mus musculus]	1363	0.0
gb EDL40224.1 	nucleolin, isoform CRA_e [Mus musculus]	1009	0.0
gb EDL40223.1 	nucleolin, isoform CRA_d [Mus musculus]	966	0.0
gb EDL40222.1 	nucleolin, isoform CRA_c [Mus musculus]	942	0.0
sp P13383.3 NUCL_RAT	RecName: Full=Nucleolin; AltName: Full=P...	941	0.0
ref NP_036881.2 	nucleolin [Rattus norvegicus] >gb AAH85751.1...	941	0.0
sp P08199.2 NUCL_MESAU	RecName: Full=Nucleolin; AltName: Full...	919	0.0
gb EDL75577.1 	nucleolin, isoform CRA_b [Rattus norvegicus]	912	0.0
gb AAA36966.1 	nucleolin, C23	893	0.0
gb EDL40220.1 	nucleolin, isoform CRA_a [Mus musculus]	797	0.0
dbj BAC34476.1 	unnamed protein product [Mus musculus]	796	0.0
gb EDL40221.1 	nucleolin, isoform CRA_b [Mus musculus]	786	0.0
gb AAD56625.1 AF151373_1	nucleolin-related protein NRP [Rattu...	781	0.0
sp Q4R4J7.3 NUCL_MACFA	RecName: Full=Nucleolin >dbj BAE00345....	768	0.0
ref XP_001116949.1 	PREDICTED: similar to nucleolin [Macaca m...	762	0.0
ref XP_861643.1 	PREDICTED: similar to nucleolin-related prot...	761	0.0
ref XP_861613.1 	PREDICTED: similar to nucleolin-related prot...	761	0.0
ref XP_850477.1 	PREDICTED: similar to nucleolin-related prot...	761	0.0
gb EDL75581.1 	nucleolin, isoform CRA_e [Rattus norvegicus]	756	0.0
ref XP_516145.2 	PREDICTED: hypothetical protein [Pan troglod...	755	0.0
gb EDL75579.1 	nucleolin, isoform CRA_d [Rattus norvegicus] >...	748	0.0
ref XP_001495211.2 	PREDICTED: nucleolin [Equus caballus]	727	0.0
ref XP_861582.1 	PREDICTED: similar to nucleolin-related prot...	701	0.0



How often would this hit have occurred by chance?

Rule of thumb:
E-value < 0.0001

BLASTing a sequence at NCBI

```
> gb|AAF62554.1 | G nucleolin [Oncorhynchus mykiss]
Length=255

GENE ID: 100135911 LOC100135911 | nucleolin [Oncorhynchus mykiss]

Score = 239 bits (610), Expect = 7e-61, Method: Compositional matrix adjust.
Identities = 133/260 (51%), Positives = 182/260 (70%), Gaps = 11/260 (4%)

Query 283 KKEMTRKQKEAPEAKKQKVEGSEPTTFFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVD 342
          K++  +KE P AKK K SE F LFIGNLN NK +E+K A++ F+K +L V D
Sbjct 2 KRKADNKKETPPAKKAK---SESDDTFCFLFIGNLSNKDFDEIKEALAAFFSKKNLEVD 58

Query 343 VRTGTNRKFGYVDFESAEDLEKALELTGLKVFGNEIKLEKPKGR----DSKKVRAARTLL 398
          VR G ++KFGYV+F SAED++ A+EL G K G E+K++K + + + KK R ARTL
Sbjct 59 VRLGASKKFGYVEFASAEDMQTAMELNGKKCMGQELKMDKARSKGNSQEEKKDRDARTLF 118

Query 399 AKNLSFNITEDELKEVFEDAMEIRL-VSQDGKSKGIAYIEFKSEADAENLLEEKQGAIED 457
          KNL F+ TED+LKEVF +A+EIR+ QDG ++GIAYI EK+EA A+K L E QGA++
Sbjct 119 VKNLPEFSATEDDLKEVFANAVEIRIPTGQDGSNRGIAYIAFKTEAMADKMLTEAQGADVQ 178

Query 458 GRSVSLYYTGEGKQRQERTGKTSTWSGESKTLVLSNLSYSATKETLEEVFEKATFIKVPQ 517
          GRS+ + YTG K Q+ R + + ESKTL+++NLSYSAT+++L+ FE A I+VPQ
Sbjct 179 GRSIMVDYTGIKSQKGRP--PAQAAAESKTLIVNLSYSATEDSLQSAFEGAVSIRVPQ 236

Query 518 NPHGKPKGYAFIEFASFEDA 537
          N +G+PKG+AF+EF S E A
Sbjct 237 N-NGRPGGFAFVEFESAEXA 255

Score = 99.8 bits (247), Expect = 8e-19, Method: Compositional matrix adjust.
Identities = 76/242 (31%), Positives = 118/242 (48%), Gaps = 29/242 (11%)

Query 396 TLLAKNLSFNITEDELKEVFE-----DAMEIRLVSDGKSKGIAYIEFKSEADAEN 447
          L NL+ N DE+KE + ++RL G SK Y+EF S D +
Sbjct 26 CLFIGNLSNPKDFDEIKEALAAFFSKKNLEVDVRL----GASKKFGYVEFASAEDMQTA 81

Query 448 LEEKQGAIDGRSVSLYYTGEGKQRQERTGKTSTWSGESKTLVLSNLSYSATKETLEEVF 507
          +E G + G+ + + KG QE +++TL + NL +SAT++ L+EVF
Sbjct 82 ME-LNGKKCMGQELKMDKARSKGNSQEEKKDR-----DARTLFVKNLPEFSATEDDLKEVF 135

Query 508 EKATFIKVPQNPHGKPKGYAFIEFASFEDAENLNSCNKMEIEGRTIRLELQGSNSR--- 564
          A I++P G +G A+I F + A + L +++GR+I ++ G S+
Sbjct 136 ANAVEIRIPTGQDGSNRGIAYIAFKTEAMADKMLTEAQGADVQGRSIMVDYTGIKSQKGG 195

Query 565 -----SQPSKTLFVKGLSEDTEETLKESEFEGSVRARIVTDRETGSSKGFVDFNSEE 618
          + SKTL V LS TE++L+ +FEG+V R+ + G KGF FV+F S E
Sbjct 196 RPPAQAAAESKTLIVNLSYSATEDSLQSAFEGAVSIRV--PQNNRPGGFAFVEFESAE 253

Query 619 DA 620
          A
Sbjct 254 XA 255
```

Alternatives for homology searches

Country / continent	Program	Address
USA	FASTA	http://fasta.bioch.Virginia.edu/fasta
Europe	FASTA	http://www.ebi.ac.uk/fasta33
Europe	SSEARCH	http://www.ch.embnet.org/software/GMFDF_form.html
Japan	SSEARCH / FASTA	http://www.ddbj.nig.ac.jp/search/search-e.html
USA	BLAT	http://genome.ucsc.edu/

