Course organization

- Introduction (Week 1-2)
  - Course introduction
  - A brief introduction to molecular biology
  - A brief introduction to sequence comparison
- Part I: Algorithms for Sequence Analysis (Week 3 - 11)
  - Chapter 1-3, Models and theories
    - » Probability theory and Statistics (Week 4)
    - » Algorithm complexity analysis (Week 5)
    - » Classic algorithms (Week 6)
    - » Lab: Linux and Perl
  - Chapter 4, Sequence alignment (week 7)
  - Chapter 5, Hidden Markov Models (week 8)
  - Chapter 6. Multiple sequence alignment (week 10)
  - Chapter 7. Motif finding (week 11)
  - Chapter 8. Sequence binning (week 11)
- Part II: Algorithms for Network Biology (Week 12 - 16)

Chapter 4:
Blast

Chaochun Wei

Fall 2014

Contents

- Reading materials
- Introduction to BLAST
- Inside BLAST
  - Algorithm
  - Karlin-Altschul Statistics

Reading

Karlin, S, and SF Altschul (1990), "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes", PNAS 87:2264-68
Altschul, SF, Gish, W, Miller, W, Myers, E, Lipman DJ (1990), "Basic Local Alignment Search Tool", J. Mol. Biol. 215:403-410

Supporting materials
Altschul, SF(1991), "Amino Acid substitution matrices from an information theoretic perspective", J. Mol. Biol. 219:555-65
Altschul, SF (1993), "A protein alignment scoring system sensitive at all evolution distances", J. Mol. Biol. 36:290-330
Altschul, SF, and W. Gish (1996), "Local alignment statistics", Methods Enzymol. 266:460-80
Altschul, SF, Bundschuh, R, Olsen, R, and T Hwa (2001). "The estimation of statistical parameters for local alignment score distributions", Nucl. Acids. Res. 29:351-61
Karlin, S, and SF Altschul (1993). "Applications and statistics for multiple high-scoring segments in molecular sequences". PNAS, 90:2264-68
Pearson, WR (1998), "Empirical statistical estimates for sequence similarity searches", J. Mol. Biol. 276:71-84.

Introduction to BLAST

- What is BLAST
  - Basic Local Alignment Search Tool
- Why BLAST
  - Quickly search a sequence database

Alignment in Real Life (20+ years ago)

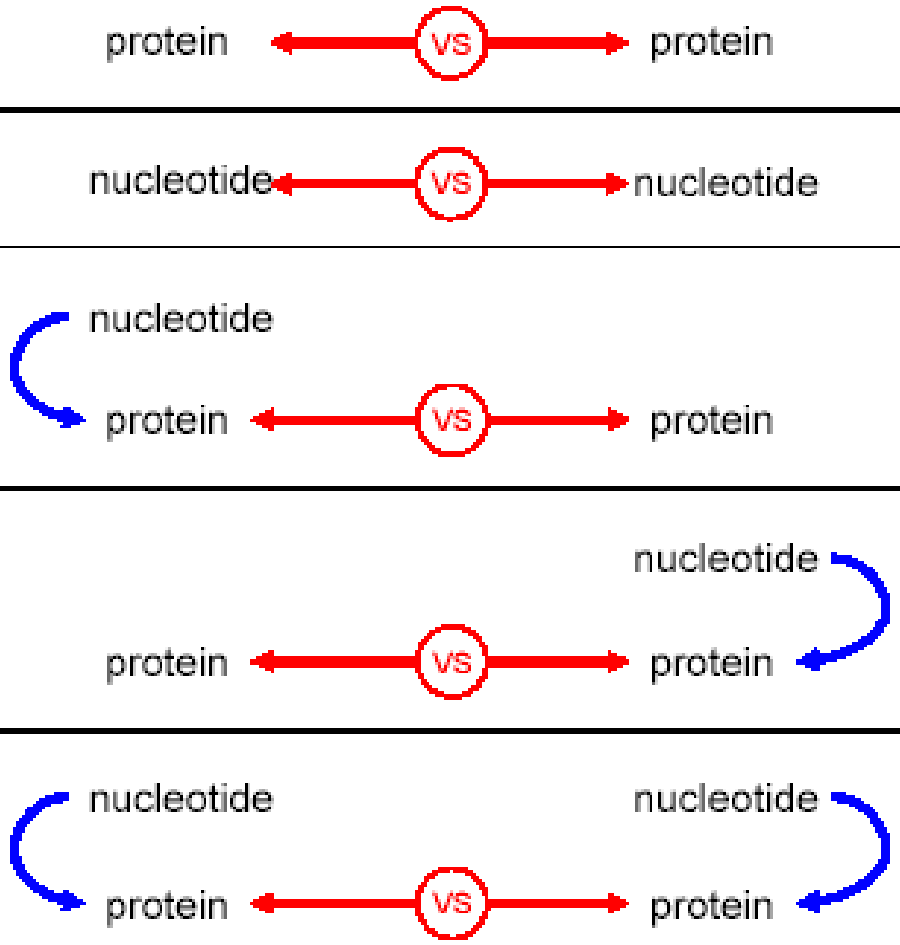
- One of the major uses of alignments is to find sequences in a database
- The current protein database contains about 10^8 residues!
  - Searching a 10^3 base long target sequence requires to evaluate about 10^{11} matrix cells…
  - … which will take about **three hours** in the rate of 10^7 evaluations per second.
  - Quite annoying when, say, 10^3 sequences are waiting to be searched. About **four months** will be required for completing the analysis!

Introduction to BLAST

- Different versions of BLAST
  - NCBI-BLAST
  - WU-BLAST (now AB-BLAST)

Different BLAST programs: according to the query and database

Program	Query	Database
blastp	protein	protein
blastn	nucleotide	nucleotide
blastx	nucleotide → protein	protein
tblastn	protein	nucleotide → protein
tblastx	nucleotide → protein	nucleotide → protein



```
BLASTP 3.0PE-AB [2009-10-30] [linux26-x64-I32LPF64 2009-11-17T18:52:53]

Copyright (C) 2009 Warren R. Gish.  All rights reserved.
Unlicensed use, reproduction or distribution are prohibited.
Advanced Biocomputing, LLC, licenses this software only for personal use
on a personally owned computer.

Reference:  Gish, W. (1996-2009) http://blast.advbiocomp.com

Query=  RU1A_HUMAN
        (282 letters)

Database:  /home/ccwei/courses/g_and_p/C.elegans/Proteome/ws_215.protein
           24,705 sequences; 10,879,267 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

K08D10.3        CE07355 WBGene00004386  locus:rnp-3     U1 small nucl...    378  3.2e-53   2
K08D10.4        CE28597 WBGene00004385  locus:rnp-2     U1 small nucl...    332  1.5e-51   2
C50D2.5 CE38492 WBGene00016808  status:Confirmed        UniProt:Q...    113  7.4e-08   1
F46A9.6 CE08260 WBGene00003172  locus:mec-8     mecanosensory ...    111  5.8e-07   2
R09B3.2 CE16307 WBGene00011155  RNA recognition motif. (ak...     91  2.6e-05   1
D2089.4b        CE30509 WBGene00004207  locus:ptb-1     status:Partia...     86  5.4e-05   2
T01D1.2g        CE41586 WBGene00001340  locus:etr-1     status:Confir...     95  6.5e-05   2
T23F6.4 CE18963 WBGene00004315  locus:rbd-1     RNA recognitio...     85  8.1e-05   2
T01D1.2a        CE12942 WBGene00001340  locus:etr-1     RNA-binding p...     95  9.0e-05   2
```

```
>K08D10.3       CE07355 WBGene00004386  locus:rnp-3     U1 small nuclear ribonucleoprotein
          A     status:Confirmed        UniProt:Q21323  protein_id:AAA98033.1
        Length = 217

 Score = 378 (138.1 bits), Expect = 3.2e-53, Sum P(2) = 3.2e-53
 Identities = 69/116 (59%), Positives = 89/116 (76%)

Query:     5 ETRPNHTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRSLKMRGQAFVIFKEVSS 64
             +  PNHTIY+NNLNEK+KKDELK+SL+ +F+QFG+I+ ++  R  KMRGQA ++FKEVSS
Sbjct:     3 DINPNHTIYVNNLNEKVKKDELKRSLHMVFTQFGEIIQLMSFRKEKMRGQAHIVFKEVSS 62

Query:    65 ATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKMKGTFVXXXXXXXXXXXXSQETPA 120
             A+NALR++QGFPFY KPMRIQYA+ DSD+I++ KGTFV              E PA
Sbjct:    63 ASNALRALQGFPFYGKPMRIQYAREDSDVISRAKGTFVEKRQKSTKIAKKPYEKPA 118

 Score = 179 (68.1 bits), Expect = 3.2e-53, Sum P(2) = 3.2e-53
 Identities = 33/77 (42%), Positives = 49/77 (63%)

Query:   206 PNHILFLTNLPEETNELMLSMLFNQFPGFKEVRLVPGRHDIAFVEFDNEVQAGAARDALQ 265
             PN+ILF +N+PE T    +  +F+QFPG +EVR +P   D AF+E+++E  +  AR AL
Sbjct:   141 PNNILFCSNIPEGTEPEQIQTIFSQFPGLREVRWMPNTKDFAFIEYESEDLSEPARQALD 200

Query:   266 GFKITQNNAMKISFAKK 282
              F+IT    + + FA K
Sbjct:   201 NFRITPTQQITVKFASK 217
```

Heuristic Search

- Rather than struggling to find the optimal alignment we may save a lot of time by employing heuristic algorithms
  - Execution time is much faster
  - May completely miss the optimal alignment

- Two important algorithms
  - BLAST
  - FASTA

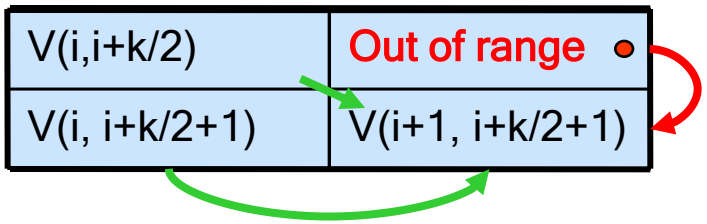
Basic Intuition 1: Seeds

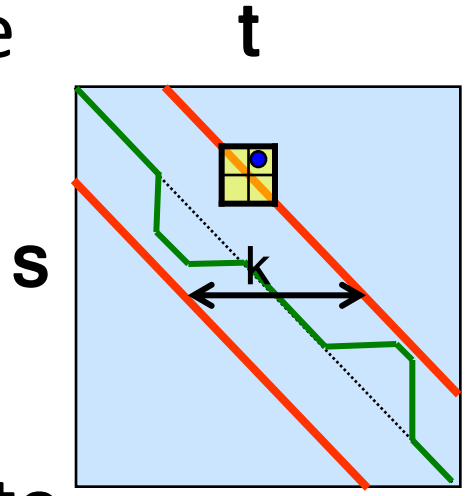
- **Observation:** Real-life matches often contain long strings with gap-less matches

- **Action:** Try to find significant gap-less matches and then extend them

Basic Intuition 2: Banded DP

- **Observation:** If the optimal alignment of **s** and **t** has few gaps, then path of the alignment will be close to diagonal



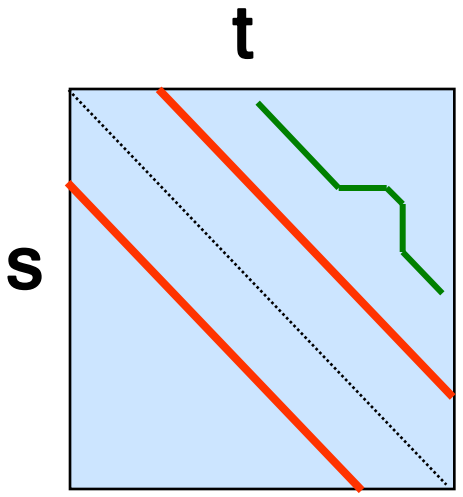


- **Action:** To find such a path, it suffices to search in a diagonal band of the matrix.
  - If the diagonal band consists of **k** diagonals (width **k**), then dynamic programming takes O(**kn**).
  - Much faster than $O(\mathbf{n}^2)$ of standard DP.

Banded DP for Local Alignment

- **Problem:** The banded diagonal needs not be the main diagonal when looking for a good local alignment
  - Also the case when the lengths of **`s`** and **`t`** are different

- **Solution:** Heuristically find potential diagonals and evaluate them using Banded DP



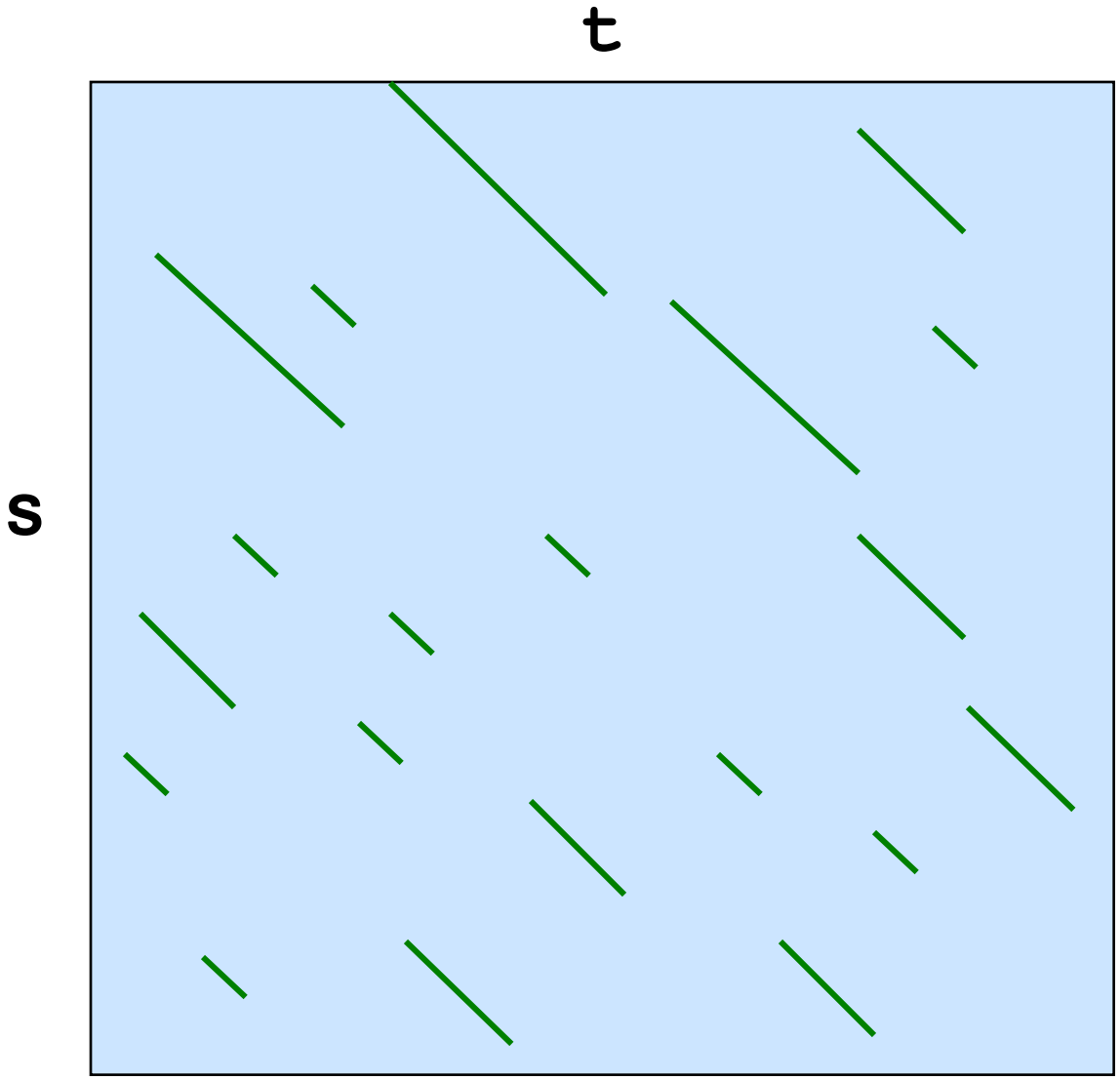
FASTA

- **Publication**
  - Pearson and Lipman, 1988
- **Input**
  - **Two sequences `s` and `t`**
  - **Parameter `ktup`** – defines the length of seeds.
    - Typically **`ktup`**=1-2 for proteins and **`ktup`**=4-6 for DNA/RNA
- **Output**
  - The best local alignment between **`s`** and **`t`**

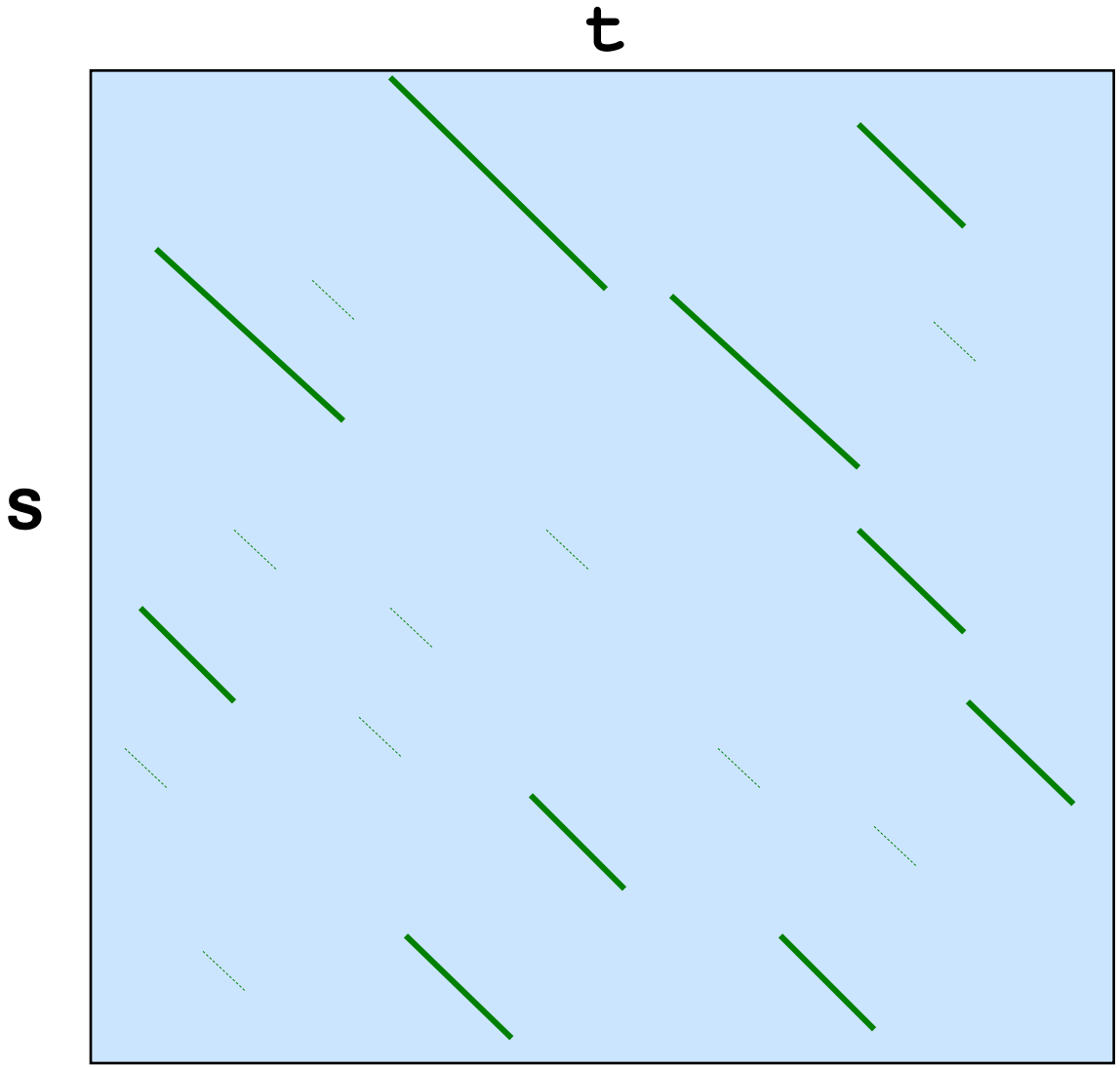
FASTA – Algorithm Outline

- Find regions in **`s`** and **`t`** containing high density of seeds
- Re-score the 10 regions with the highest scores using PAM matrix
- Eliminate segments that are unlikely to be part of alignments
- Optimize the best alignment using the banded DP algorithm

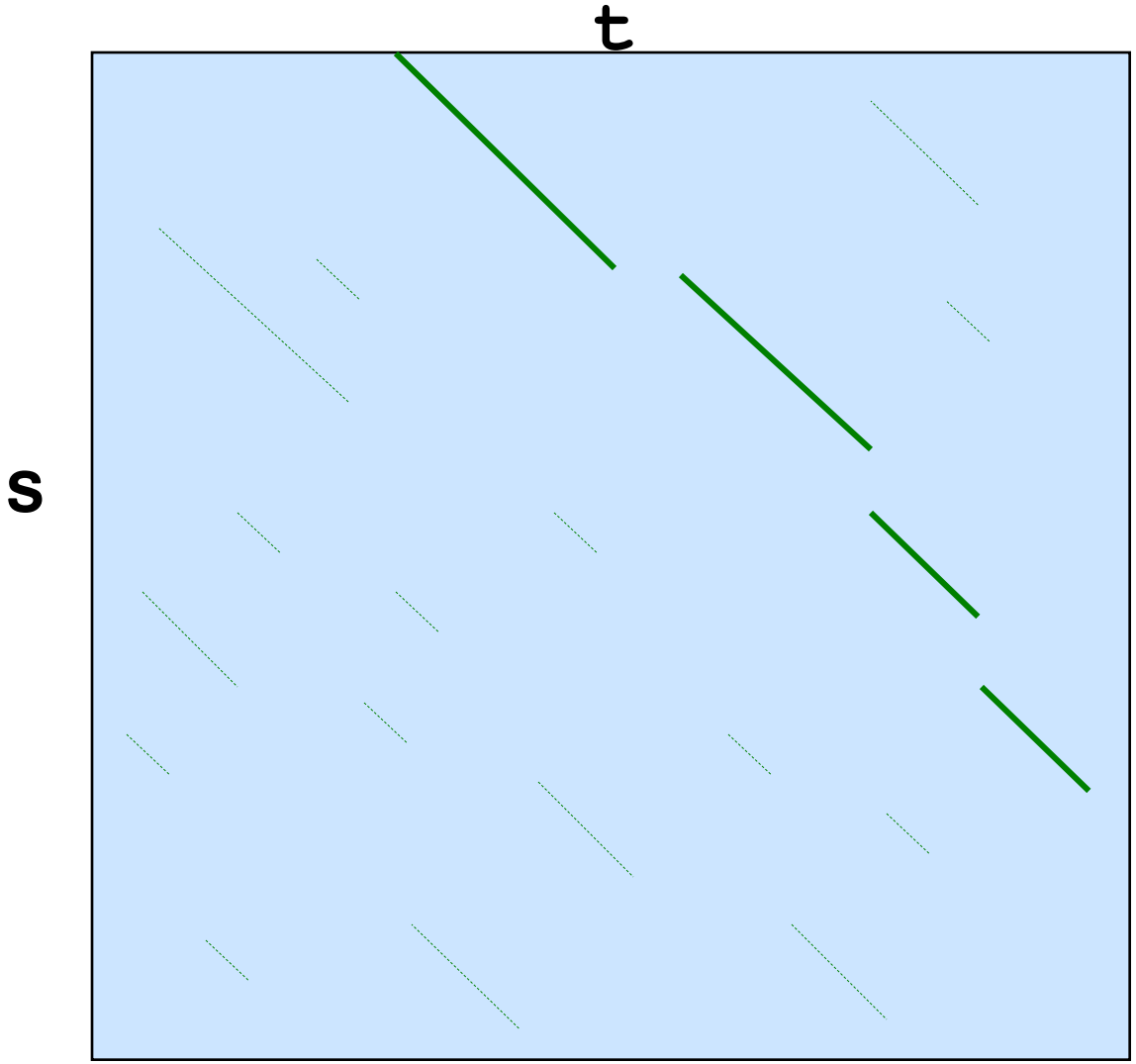
Step 1: Finding Seeds



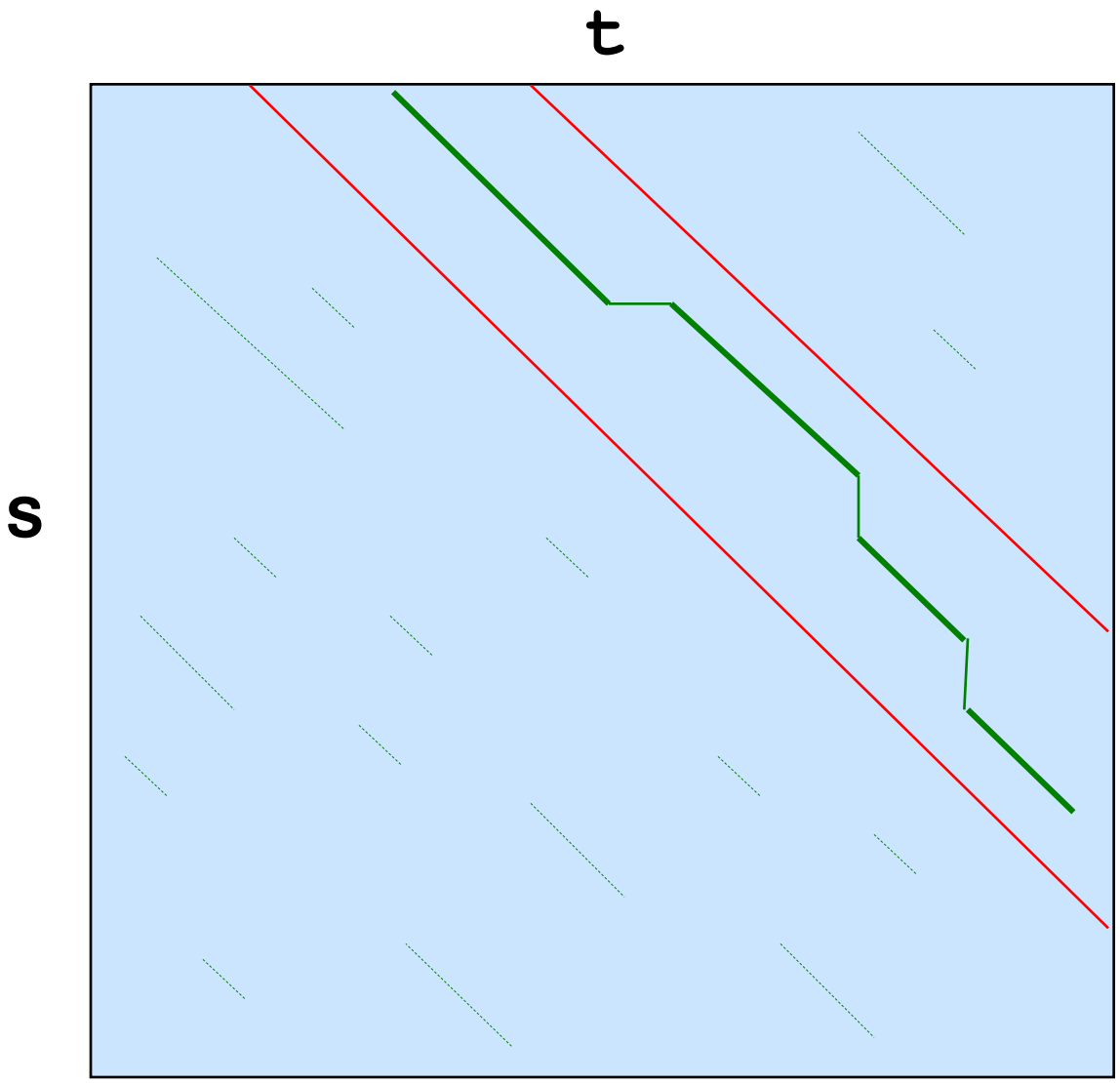
Step 2: Re-scoring Segments, Keeping Top 10



Step 3: Eliminating Unlikely Segments



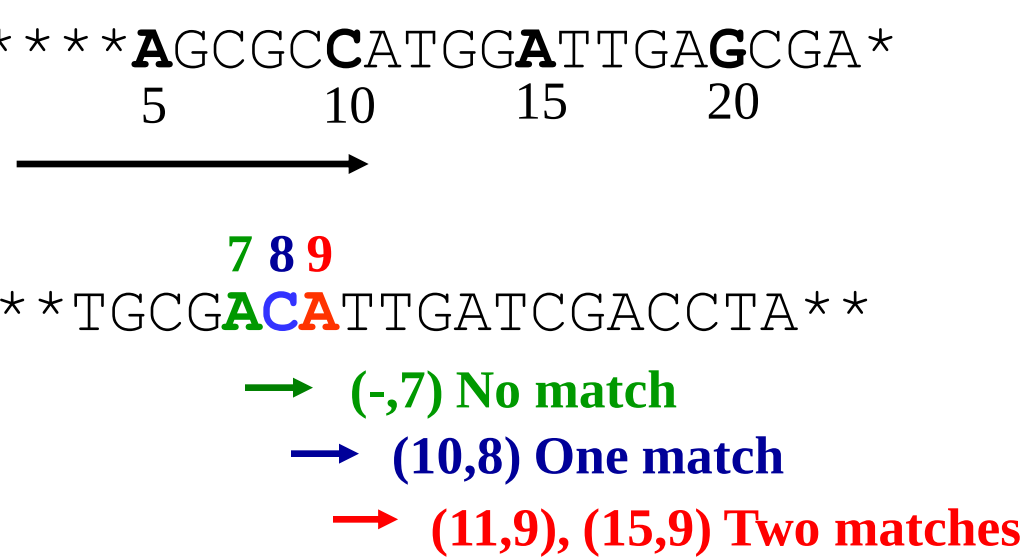
Step 4: Finding the Best Alignment



Finding Seeds Efficiently

- **Prepare an index table** of the database sequence **s** such that for any sequence of length **ktup**, one gets the list of its positions in **s**.
- **March on the query sequence t** while using the index table to list all matches with the database sequence **s**.

Index Table (ktup=2)	
AA	-
AC	-
AG	5, 19
AT	11, 15
CA	10
CC	9
CG	7 , 21
…	
TT	16



Connecting Seeds on the Same Diagonal

- The maximal size of the index table is $|\sum|^{\texttt{ktup}}$ where $\sum$ is the alphabet size (4 or 20).
  - For small **`ktup`**, the entire table is stored
- For large **`ktup`** values, one should keep only entries for tuples actually found in the database
  - In this case, hashing is needed
- Typical values of **`ktup`** are 1-2 for Proteins and 4-6 for DNA
- The index table is prepared for each database sequence ahead of users' matching requests, at compilation time.
  - Matching time is O(|**`t`**|·**`max`**{**`row_length`**})

Identifying Potential Diagonals

- Input: Sets of pairs
  - E.g, (6,4),(10,8),(14,12),(15,10),(20,4) …
- Task
  - Locate sets of pairs that are on the same diagonal.
- Method
  - Sort according to the difference *i-j.*
  - E.g, 6-4=2, 10-8=2, 14-12=2, 15-10=5, 20-4=16 …

FASTA Parameters

- *ktup* = 2 for proteins, 6 for DNA
- *init1* Score after rescanning with PAM250 (or other)
- *initn* Score after joining regions
- *opt* Score after Banded DP

Limits

- Local similarity might be missed because only 10 regions saved at *init1* stage.
- Non-identical conserved stretches may be overlooked

Basic Local Alignment Search Tool (BLAST)

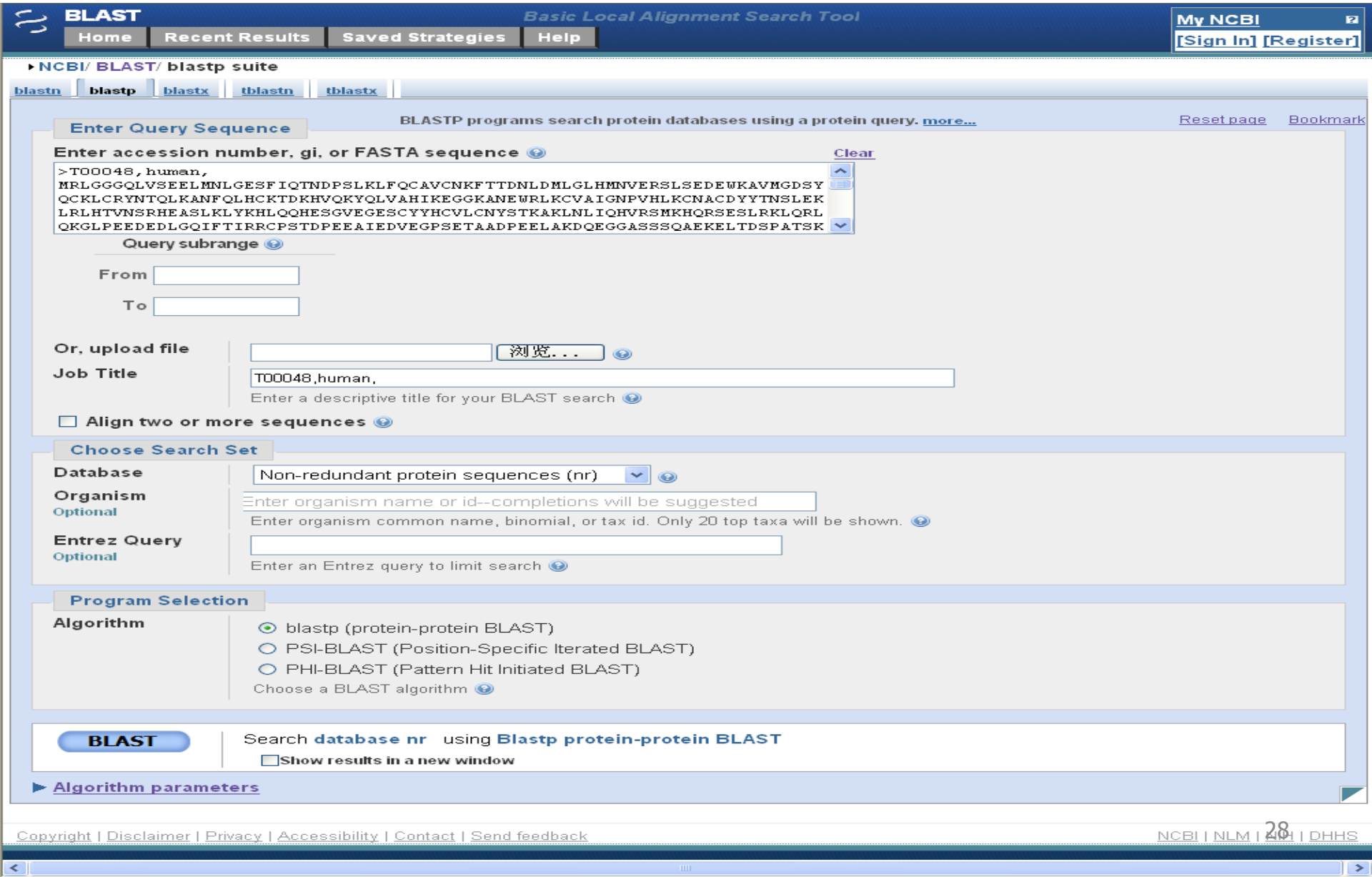
- **Publications:**
  - **Ungapped BLAST** – Altschul et al., 1990
  - **Gapped BLAST, PSI-BLAST** - Altschul et al., 1997

- **Input:**
  - **Query (target) sequence** – either DNA, RNA or Protein
  - **Scoring Scheme** – gap penalties, substitution matrix for proteins, identity/mismatch scores for DNA/RNA
  - **Word length `W`** – typical is `W`=3 for proteins and `W`=11 for DNA/RNA

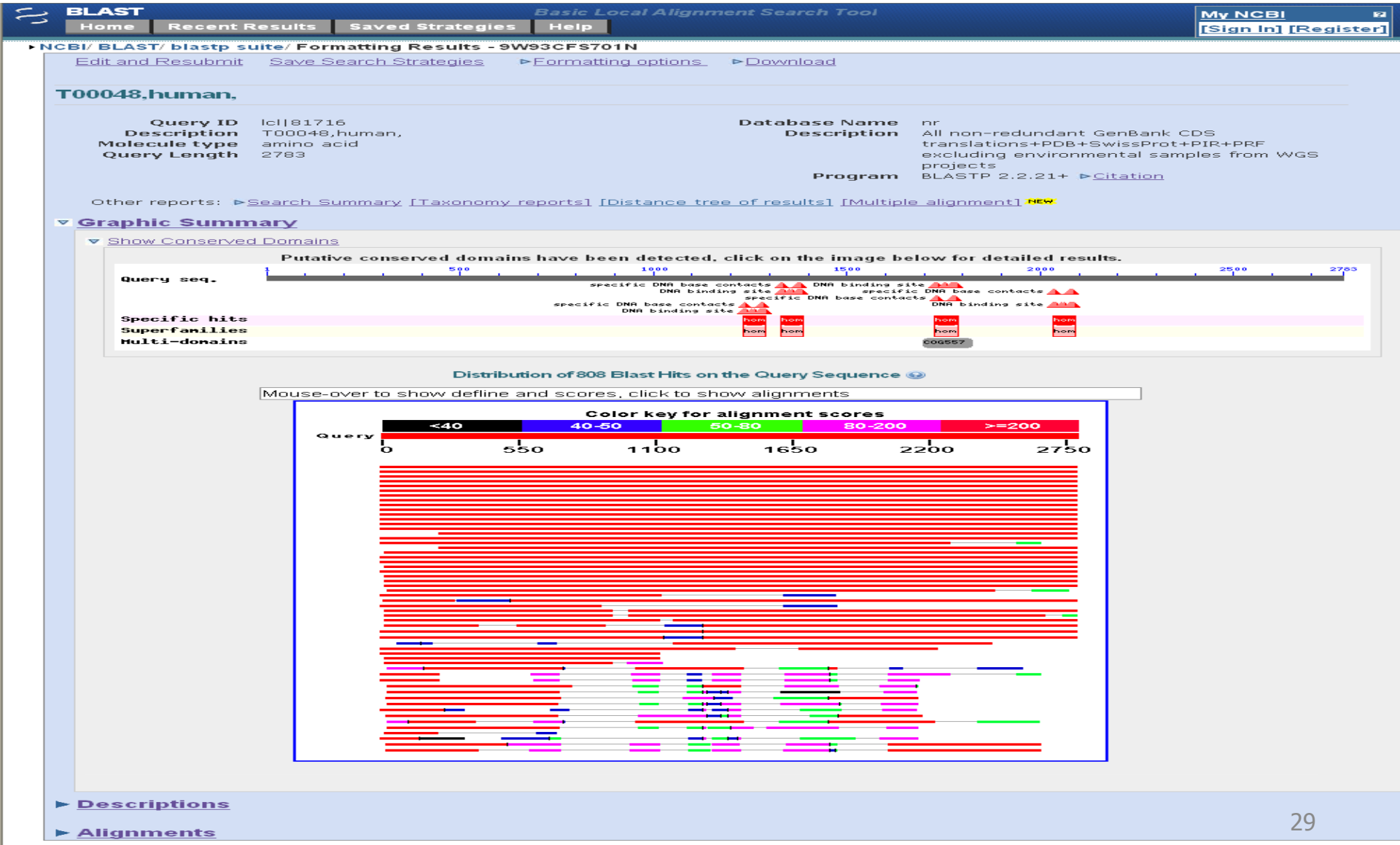
- **Output:**
  - Statistically significant matches

Running BLAST

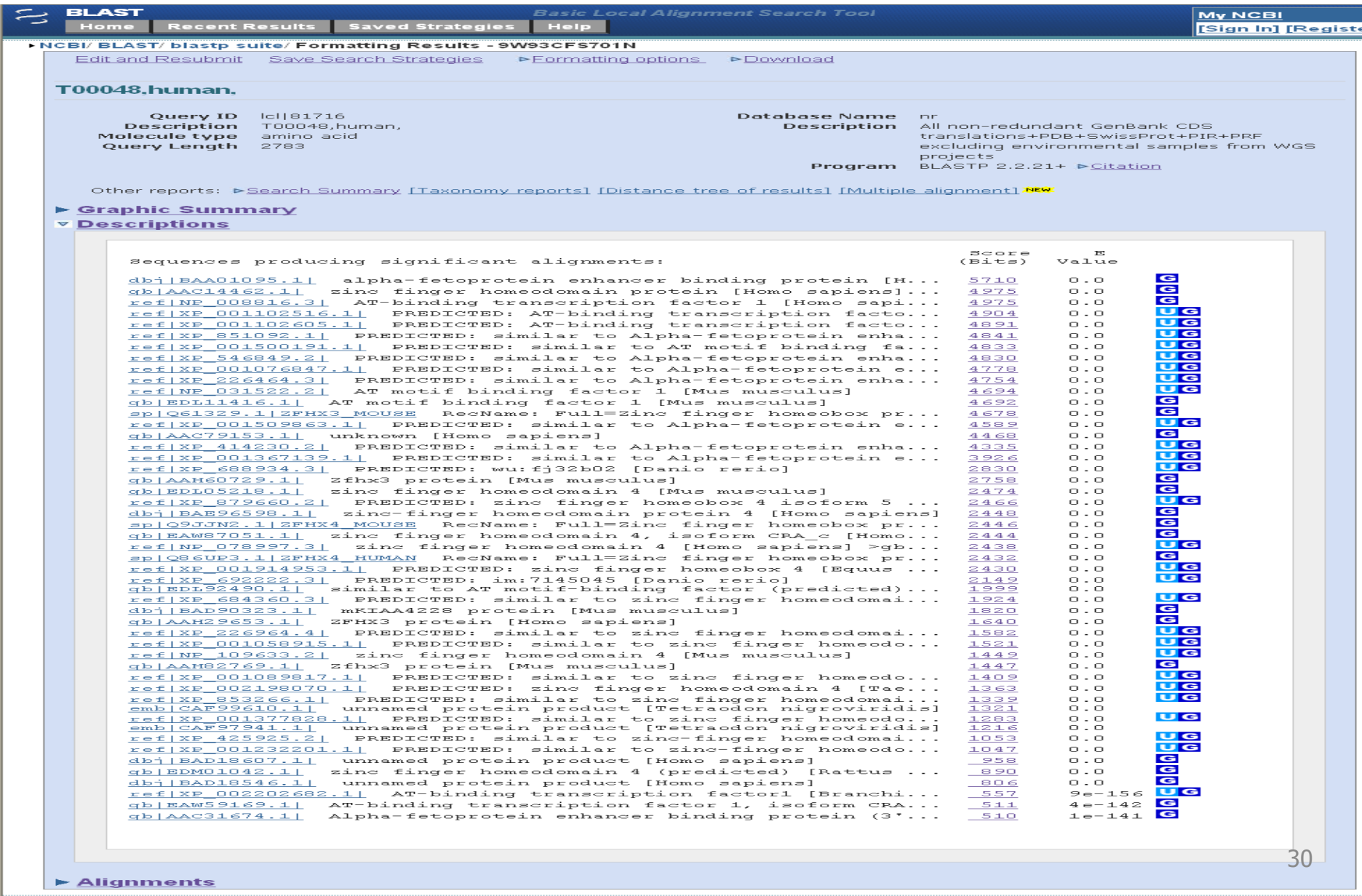
NCBI BLAST – web site



NCBI BLAST – result summary



NCBI BLAST use – predict function



BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help | My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - 9W93CFS701N

Edit and Resubmit | Save Search Strategies | ▸Formatting options | ▸Download

T00048.human.

Query ID	lcl|81716	**Database Name**	nr
Description	T00048,human,	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	2783	**Program**	BLASTP 2.2.21+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] NEW

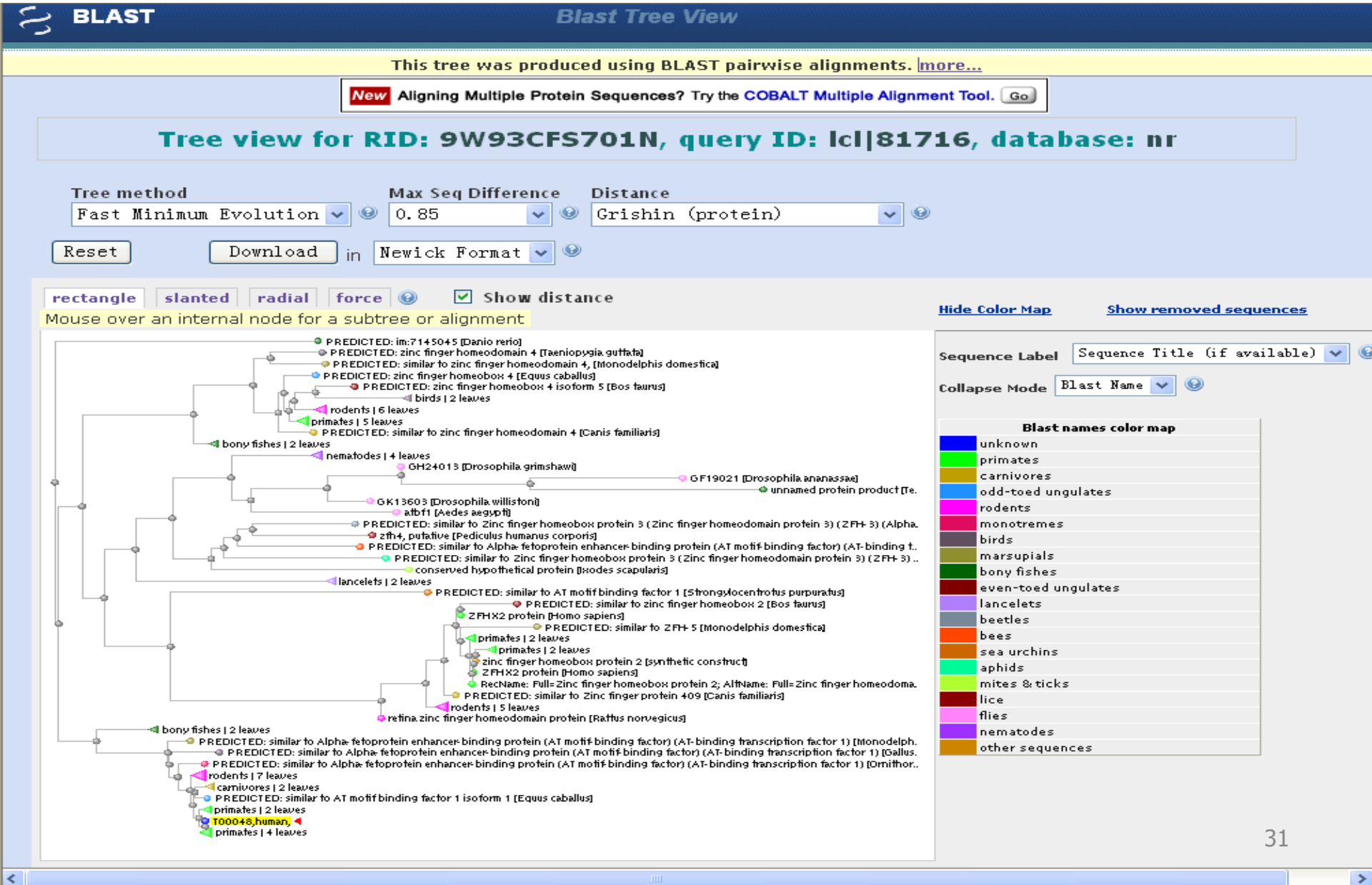
▸ **Graphic Summary**

▾ **Descriptions**

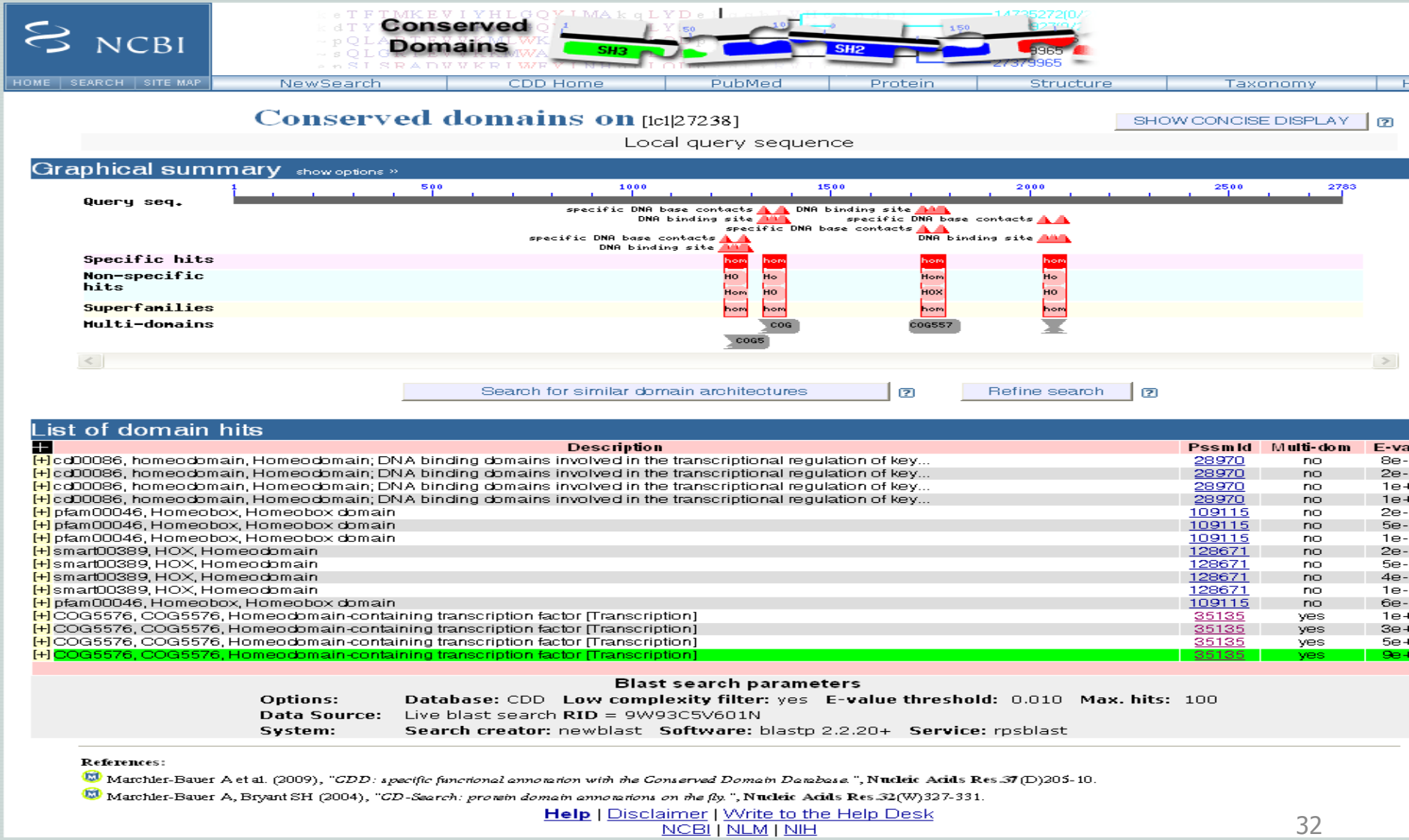
Sequences producing significant alignments:	Score (Bits)	E Value
dbj|BAA01095.1| alpha-fetoprotein enhancer binding protein [H...	5710	0.0
gb|AAC14462.1| zinc finger homeodomain protein [Homo sapiens]...	4975	0.0
ref|NP_008816.3| AT-binding transcription factor 1 [Homo sapi...	4975	0.0
ref|XP_001102516.1| PREDICTED: AT-binding transcription facto...	4904	0.0
ref|XP_001102605.1| PREDICTED: AT-binding transcription facto...	4891	0.0
ref|XP_851092.1| PREDICTED: similar to Alpha-fetoprotein enha...	4841	0.0
ref|XP_001500191.1| PREDICTED: similar to AT motif binding fa...	4833	0.0
ref|XP_546849.2| PREDICTED: similar to Alpha-fetoprotein enha...	4830	0.0
ref|XP_001076847.1| PREDICTED: similar to Alpha-fetoprotein e...	4778	0.0
ref|XP_226464.3| PREDICTED: similar to Alpha-fetoprotein enha...	4754	0.0
ref|NP_031522.2| AT motif binding factor 1 [Mus musculus]	4694	0.0
gb|EDL11416.1| AT motif binding factor 1 [Mus musculus]	4692	0.0
sp|Q61329.1|ZFHX3_MOUSE RecName: Full=Zinc finger homeobox pr...	4678	0.0
ref|XP_001509863.1| PREDICTED: similar to Alpha-fetoprotein e...	4589	0.0
gb|AAC79153.1| unknown [Homo sapiens]	4468	0.0
ref|XP_414230.2| PREDICTED: similar to Alpha-fetoprotein enha...	4335	0.0
ref|XP_001367139.1| PREDICTED: similar to Alpha-fetoprotein e...	3926	0.0
ref|XP_688934.3| PREDICTED: wu:fj32b02 [Danio rerio]	2830	0.0
gb|AAH60729.1| Zfhx3 protein [Mus musculus]	2758	0.0
gb|EDL05218.1| zinc finger homeodomain 4 [Mus musculus]	2474	0.0
ref|XP_879660.2| PREDICTED: zinc finger homeobox 4 isoform 5 ...	2466	0.0
dbj|BAB96598.1| zinc-finger homeodomain protein 4 [Homo sapiens]	2448	0.0
sp|Q9JJN2.1|ZFHX4_MOUSE RecName: Full=Zinc finger homeobox pr...	2446	0.0
gb|EAW87051.1| zinc finger homeodomain 4, isoform CRA_c [Homo...	2444	0.0
ref|NP_078997.3| zinc finger homeodomain 4 [Homo sapiens] >gb...	2438	0.0
sp|Q86UP3.1|ZFHX4_HUMAN RecName: Full=Zinc finger homeobox pr...	2432	0.0
ref|XP_001914953.1| PREDICTED: zinc finger homeobox 4 [Equus ...	2430	0.0
ref|XP_692222.3| PREDICTED: im:7145045 [Danio rerio]	2149	0.0
gb|EDI92490.1| similar to AT motif-binding factor (predicted)...	1999	0.0
ref|XP_684360.3| PREDICTED: similar to zinc finger homeodomai...	1924	0.0
dbj|BAD90323.1| mKIAA4228 protein [Mus musculus]	1820	0.0
gb|AAH29653.1| ZFHX3 protein [Homo sapiens]	1640	0.0
ref|XP_226964.4| PREDICTED: similar to zinc finger homeodomai...	1582	0.0
ref|XP_001058915.1| PREDICTED: similar to zinc finger homeodo...	1521	0.0
ref|NP_109633.2| zinc finger homeodomain 4 [Mus musculus]	1449	0.0
gb|AAH82769.1| Zfhx3 protein [Mus musculus]	1447	0.0
ref|XP_001089817.1| PREDICTED: similar to zinc finger homeodo...	1409	0.0
ref|XP_002198070.1| PREDICTED: zinc finger homeodomain 4 [Tae...	1363	0.0
ref|XP_853266.1| PREDICTED: similar to zinc finger homeodomai...	1339	0.0
emb|CAF99610.1| unnamed protein product [Tetraodon nigroviridis]	1321	0.0
ref|XP_001377828.1| PREDICTED: similar to zinc finger homeodo...	1283	0.0
emb|CAF97941.1| unnamed protein product [Tetraodon nigroviridis]	1216	0.0
ref|XP_425925.2| PREDICTED: similar to zinc-finger homeodomai...	1053	0.0
ref|XP_001232201.1| PREDICTED: similar to zinc-finger homeodo...	1047	0.0
dbj|BAD18607.1| unnamed protein product [Homo sapiens]	958	0.0
gb|EDM01042.1| zinc finger homeodomain 4 (predicted) [Rattus ...	890	0.0
dbj|BAD18546.1| unnamed protein product [Homo sapiens]	806	0.0
ref|XP_002202682.1| AT-binding transcription factor1 [Branchi...	557	9e-156
gb|EAW59169.1| AT-binding transcription factor 1, isoform CRA...	511	4e-142
gb|AAC31674.1| Alpha-fetoprotein enhancer binding protein (3'...	510	1e-141

▸ **Alignments**

NCBI BLAST use – infer evolutionary tree



NCBI BLAST use – construct families

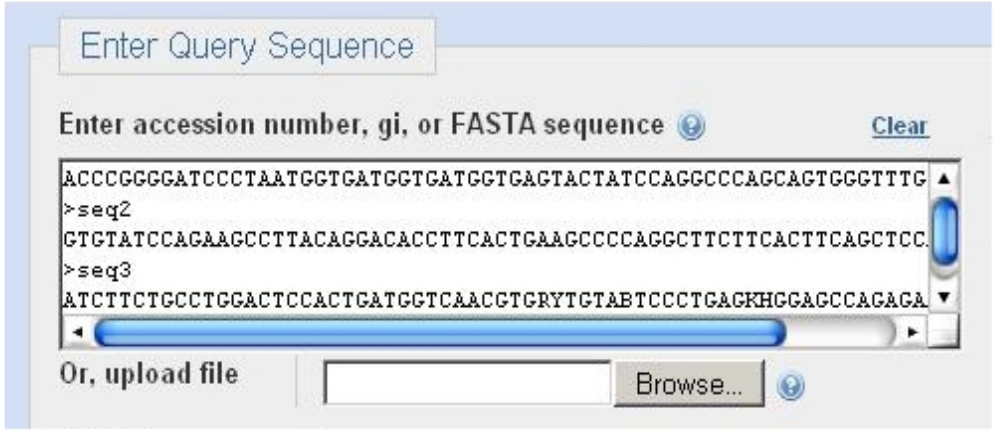


NCBI BLAST batch jobs

Batch BLAST jobs

(1) input "batches" of sequences into one form and retrieve the results

Select a BLAST search page form the main BLAST home page. Next you can either cut and paste multiple FASTA sequences from a text file into the main input box.



Or alternatively, you can use the browse button to import a local file from your computer.

Stand-alone BLAST

(1) NCBI standalone BLAST

You can retrieve BLAST execute files from NCBI ftp sites
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/

(2) The WU-BLAST

BLAST use – command line

BLAST use

(1) Make a formatted database to use

execute command : **formatdb (xdformat for WU-BLAST)**

input: fasta format sequences (database sequences)

output: formatted database , used by BLAST program

xdformat: create a WU-BLAST database

```
Purpose:  produce databases for BLAST in XDF (eXtended Database Format) from one
    or more input files in FASTA format; or report XDF databases to standard output
    in FASTA format.

  Create a database:
    xdformat [-p|-n] [options] fadb
    xdformat [-p|-n] -o xdbname [options] fadb...

  Append sequences to an existing database:
    xdformat [-p|-n] -a xdbname [options] fadb...

  Report the contents of existing database(s) to stdout in FASTA format:
    xdformat [-p|-n] -r [options] xdbname...

  Describe the contents of existing database(s):
    xdformat [-p|-n] -i xdbname...

  Verify the integrity of existing database(s):
    xdformat [-p|-n] -V xdbname...
```

BLAST use – command line

BLAST use

Carry out BLAST program

execute command : **blastn, blastp**

input: fasta sequences (query sequences), database, parameters

output : resulted alignment file

Blastn parameters

```
BLASTN 3.0PE-AB [2009-10-30] [linux26-x64-I32LPF64 2009-11-17T18:52:53]

Copyright (C) 2009 Warren R. Gish.  All rights reserved.
Unlicensed use, reproduction or distribution are prohibited.
Advanced Biocomputing, LLC, licenses this software only for personal use
on a personally owned computer.

Reference:  Gish, W. (1996-2009) http://blast.advbiocomp.com

Notice:  this program and its default parameter settings are optimized to find
nearly identical sequences rapidly.  To identify weak protein similarities
encoded in nucleic acid, use BLASTX, TBLASTN or TBLASTX.

Usage:

    BLASTN database queryfile [options]

Valid BLASTN options:  E, S, E2, S2, W, T, X, M, N, Y, Z, L, K, H, V and B
  -matrix <matrix-name>   use the specified scoring matrix (default matrix is
                    computed from M=+5 N=-4); be sure to consider changing the
                    default gap penalties when using a non-default scoring
                    system
```

Blastn parameters

```
-Q <s>    penalty score for a gap of length 1
-R <s>    penalty score for extending a gap by each letter after the first
-top     search only the top strand of the query
-bottom  search only the bottom strand of the query
-mformat <n>[,outfile]  specify alternate output format(s) (default 1)
-msgstyle <n>  specify alternate informatory message style (default 0)
-filter <method>  hard mask the query using the specified method (e.g.,
                  "seg", "xnu", "ccp", "dust" or "none")
-lcfilter    hard mask lower case letters in the query sequence
-lcmask      soft mask lower case letters in the query sequence
-topcomboN <n>  report this number of consistent (colinear) groups of HSPs
```

PART II inside into BLAST

Mathematic model of sequence alignment

Alphabet of biological sequence

- Nucleic acid sequence

  {A,T,C,G}
- Amino acid sequence {A,S,G,L,K,V,T,P,E,D,N,I,Q,R,F,Y,C,H,M,W}

Operation of sequence alignment

- Match (A,A)
- Replace (A,T)
- Delete (A, -)
- Insert (- , A)

Mathematic model of sequence alignment

How to define similarity between two sequences?

Distance

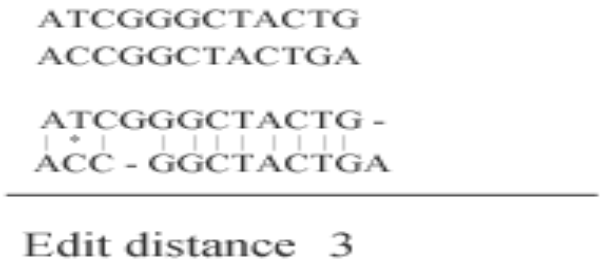
- Hamming distance

Mismatch number of two sequences with same length

- Edit distance

Operation number for one sequence transforming to another

s =	AAT	AGCAA	AGCACACA
t =	TAA	ACATA	ACACACTA
Hamming Distance(s,t)=	2	3	6



Mathematic model of sequence alignment

How to quantify the distance

Scoring

Simple scoring function

$$\begin{cases} \text{Match}(A, A) = 1 \\ \text{Substitution}(A, T) = 0 \\ \text{Delete}(A, -) = \text{Insert}(-, A) = -1 \end{cases}$$

Matrix for scoring

Matrix for nucleic acid sequence alignment

Matrix for amino acid sequence alignment

Mathematic model of sequence alignment

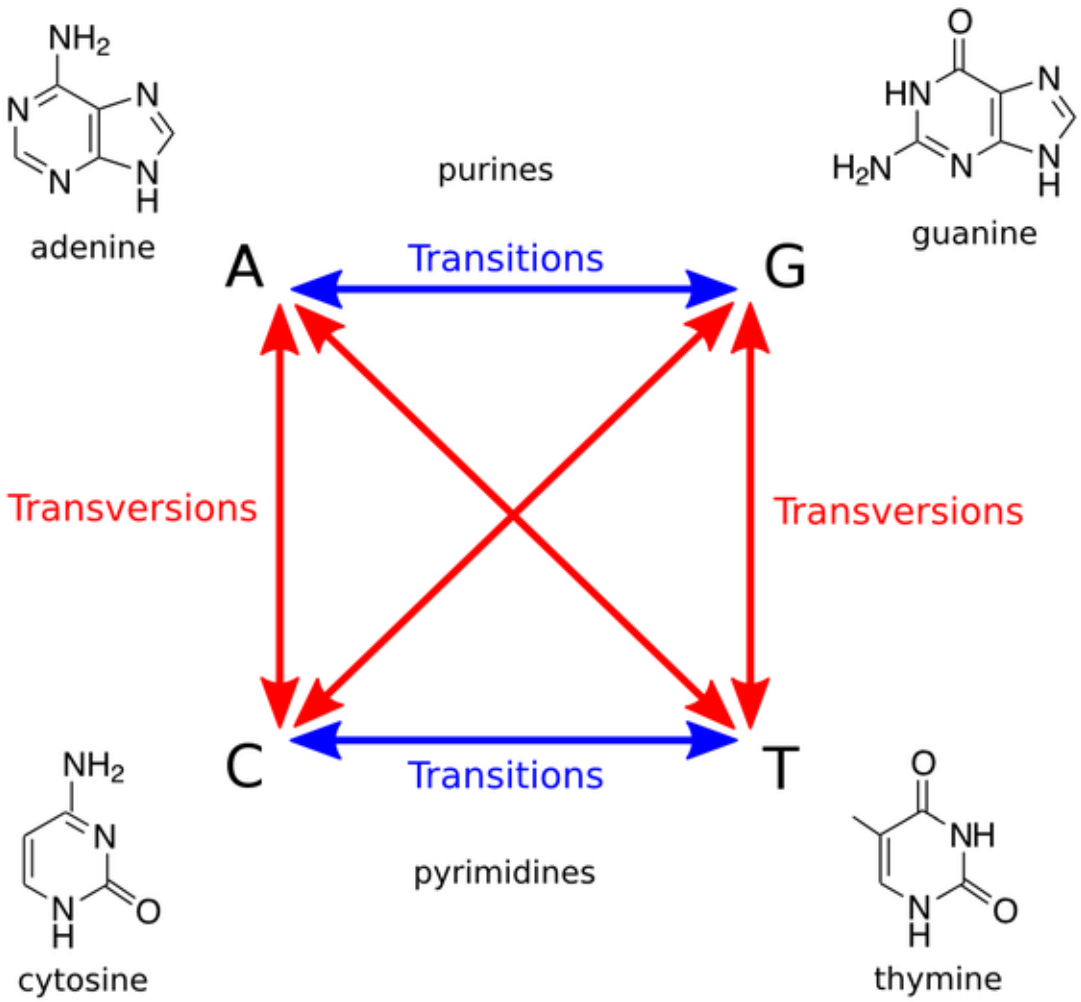
Matrix for nucleic ac

(1) equivalence matrix

(2) BLAST matrix

(3) transition-transversion m

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1



Mathematic model of sequence alignment

Matrix for amino acid sequence alignment

(1) identity matrix

(2) Point accepted mutation matrix (PAM)

(3) BLOSUM matrix

Mathematic model of sequence alignment

PAM70

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-4	-2	-1	-4	-2	-1	0	-4	-2	-4	-4	-3	-6	0	1	1	-9	-5	-1	-1	-1	-2	-11
R	-4	8	-3	-6	-5	0	-5	-6	0	-3	-6	2	-2	-7	-2	-1	-4	0	-7	-5	-4	-2	-3	-11
N	-2	-3	6	3	-7	-1	0	-1	1	-3	-5	0	-5	-6	-3	1	0	-6	-3	-5	5	-1	-2	-11
D	-1	-6	3	6	-9	0	3	-1	-1	-5	-8	-2	-7	-10	-4	-1	-2	-10	-7	-5	5	2	-3	-11
C	-4	-5	-7	-9	9	-9	-9	-6	-5	-4	-10	-9	-9	-8	-5	-1	-5	-11	-2	-4	-8	-9	-6	-11
Q	-2	0	-1	0	-9	7	2	-4	2	-5	-3	-1	-2	-9	-1	-3	-3	-8	-8	-4	-1	5	-2	-11
E	-1	-5	0	3	-9	2	6	-2	-2	-4	-6	-2	-4	-9	-3	-2	-3	-11	-6	-4	2	5	-3	-11
G	0	-6	-1	-1	-6	-4	-2	6	-6	-6	-7	-5	-6	-7	-3	0	-3	-10	-9	-3	-1	-3	-3	-11
H	-4	0	1	-1	-5	2	-2	-6	8	-6	-4	-3	-6	-4	-2	-3	-4	-5	-1	-4	0	1	-3	-11
I	-2	-3	-3	-5	-4	-5	-4	-6	-6	7	1	-4	1	0	-5	-4	-1	-9	-4	3	-4	-4	-3	-11
L	-4	-6	-5	-8	-10	-3	-6	-7	-4	1	6	-5	2	-1	-5	-6	-4	-4	-4	0	-6	-4	-4	-11
K	-4	2	0	-2	-9	-1	-2	-5	-3	-4	-5	6	0	-9	-4	-2	-1	-7	-7	-6	-1	-2	-3	-11
M	-3	-2	-5	-7	-9	-2	-4	-6	-6	1	2	0	10	-2	-5	-3	-2	-8	-7	0	-6	-3	-3	-11
F	-6	-7	-6	-10	-8	-9	-9	-7	-4	0	-1	-9	-2	8	-7	-4	-6	-2	4	-5	-7	-9	-5	-11
P	0	-2	-3	-4	-5	-1	-3	-3	-2	-5	-5	-4	-5	-7	7	0	-2	-9	-9	-3	-4	-2	-3	-11
S	1	-1	1	-1	-1	-3	-2	0	-3	-4	-6	-2	-3	-4	0	5	2	-3	-5	-3	0	-2	-1	-11
T	1	-4	0	-2	-5	-3	-3	-3	-4	-1	-4	-1	-2	-6	-2	2	6	-8	-4	-1	-1	-3	-2	-11
W	-9	0	-6	-10	-11	-8	-11	-10	-5	-9	-4	-7	-8	-2	-9	-3	-8	13	-3	-10	-7	-10	-7	-11
Y	-5	-7	-3	-7	-2	-8	-6	-9	-1	-4	-4	-7	-7	4	-9	-5	-4	-3	9	-5	-4	-7	-5	-11
V	-1	-5	-5	-5	-4	-4	-4	-3	-4	3	0	-6	0	-5	-3	-3	-1	-10	-5	6	-5	-4	-2	-11
B	-1	-4	5	5	-8	-1	2	-1	0	-4	-6	-1	-6	-7	-4	0	-1	-7	-4	-5	5	1	-2	-11
Z	-1	-2	-1	2	-9	5	5	-3	1	-4	-4	-2	-3	-9	-2	-2	-3	-10	-7	-4	1	5	-3	-11
X	-2	-3	-2	-3	-6	-2	-3	-3	-3	-3	-4	-3	-3	-5	-3	-1	-2	-7	-5	-2	-2	-3	-3	-11
*	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	1

PAM1=substitution matrix for aas mutation rate of 1%

PAM2=PAM1*PAM1

…

PAMN=PAM1^n

Clustering proteins with similarity above a certain threshold, then the substitution rates were counted from the multiple alignment

BLOck Substitution Matrix: BLOSUM

Mathematic model of sequence alignment

BLOSUM 62

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

Algorithm of BLAST

Motivation

(1) Speed up search process, reduce executive time

(2) effectively decrease store space

Feature

(1) Suit for huge data, especially for biological data

(2) Faster than Smith-Waterman algorithm

Algorithm of BLAST

- The main idea of BLAST is that there are often **high-scoring segment pairs (HSP)** contained in a statistically significant alignment.

- BLAST searches for high scoring sequence alignments between the query sequence and sequences in the database using a **heuristic approach** that approximates the Smith-Waterman algorithm

- the BLAST algorithm uses a heuristic approach that is less accurate than the Smith-Waterman but over **50 times faster**.

BLAST – Algorithm Outline

- **List all words** of length **`W`** that score at least **`T`** when aligned with the query sequence **`s`**
- **Scan the database `DB` for seeds**, namely words from the list that appear in sequences of **`DB`**
- **Find High Scoring Pairs (HSPs)** by extending the seeds in both directions. Keep best scoring HSPs
- **Combine several HSPs** using the banded DP algorithm

Step 1: Listing High Scoring Words of Length W

Word length **W**=3

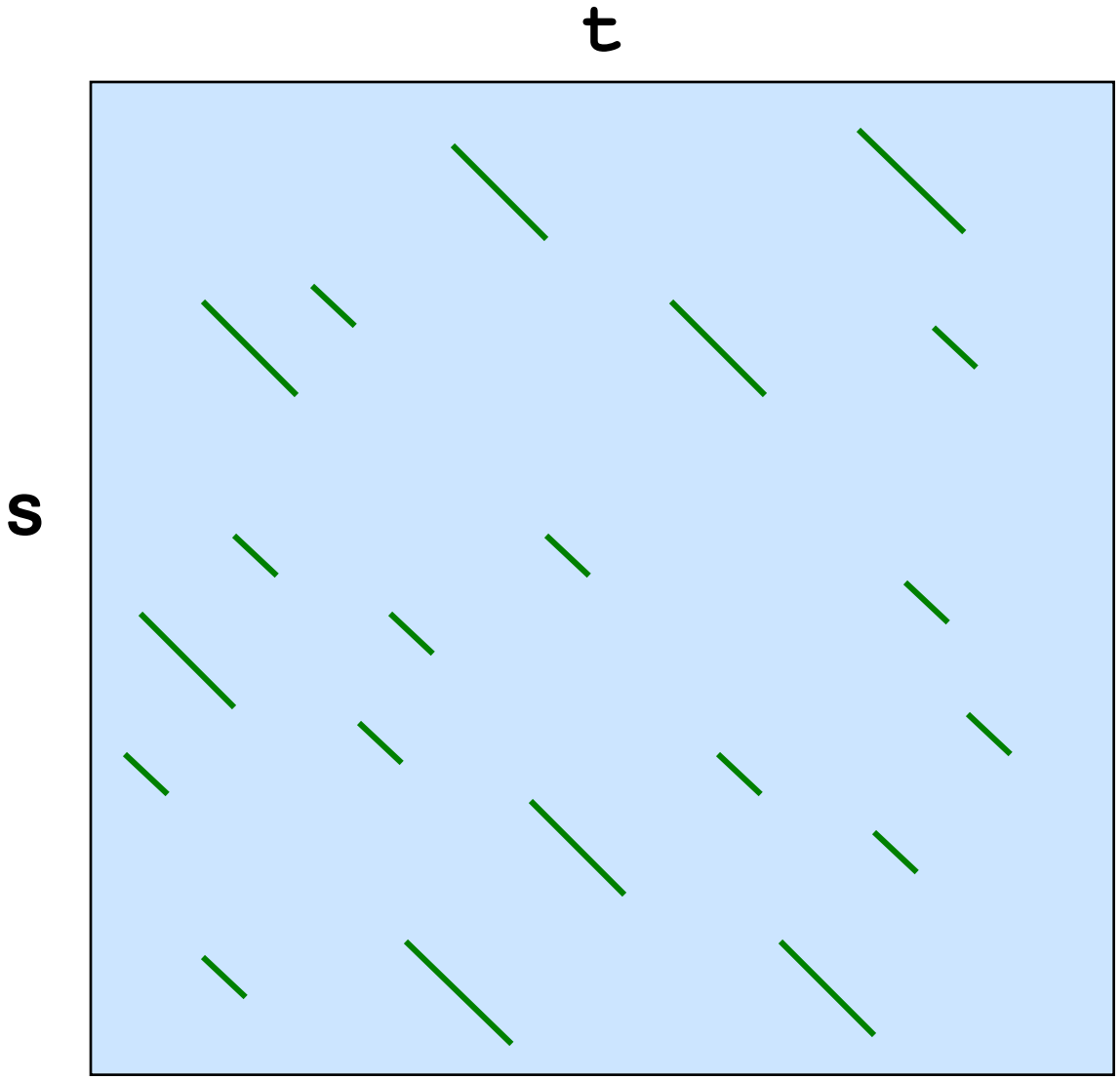
…GSVEDTTGSQSLAALLNKCKT**PQG**QRLVNQWIKQPLMDK…

High scoring words

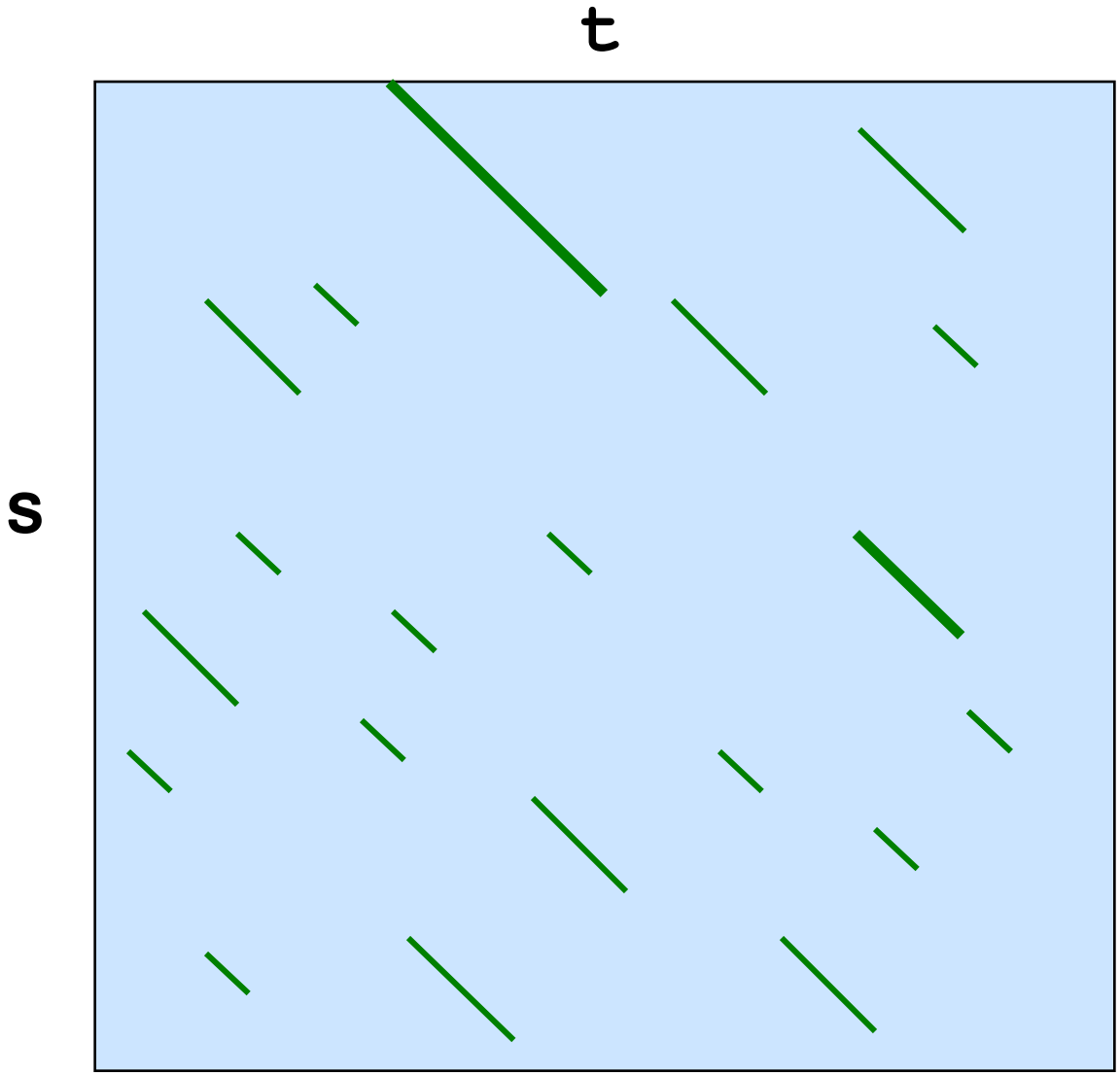
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12

Score threshold T=13

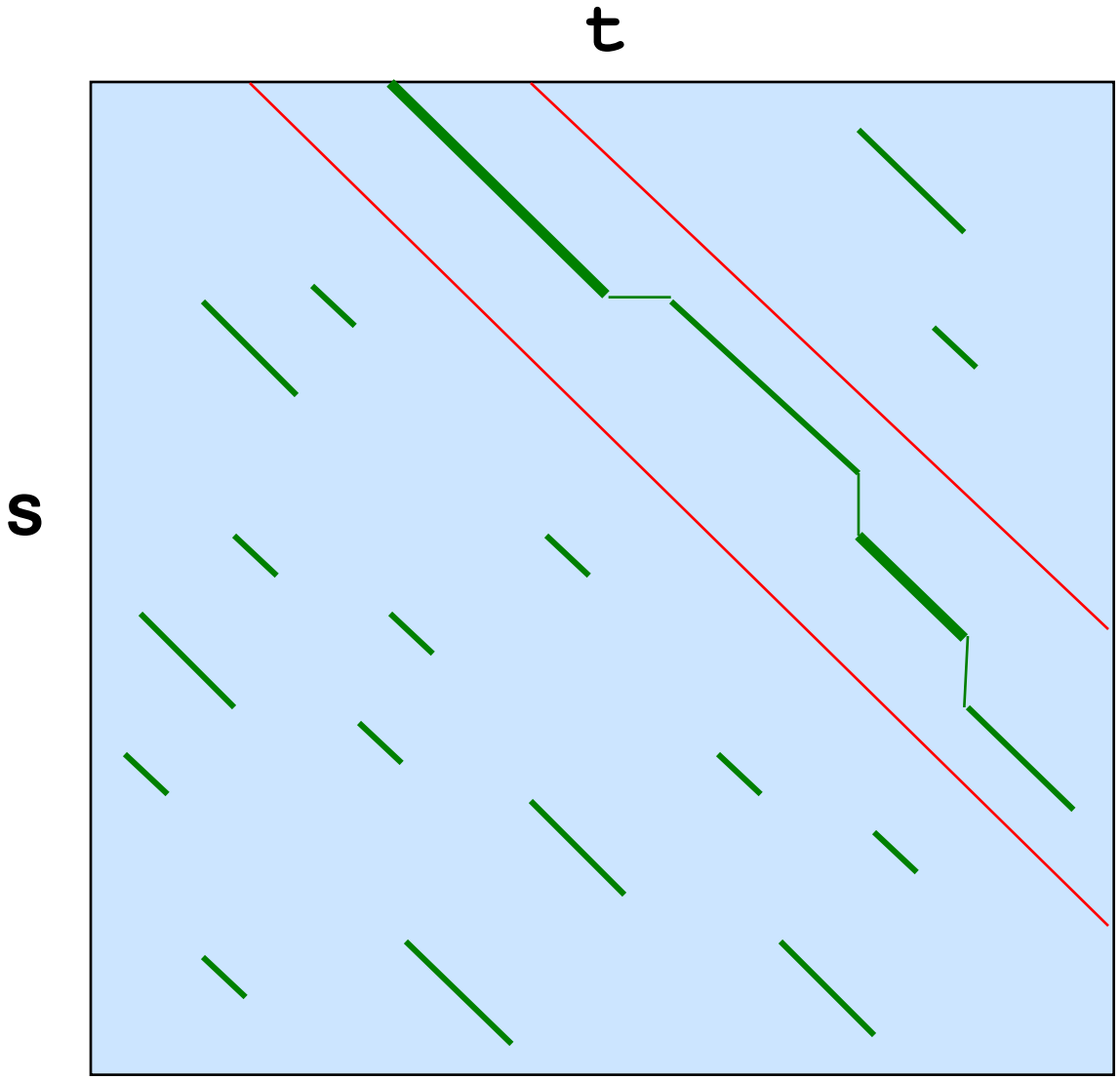
Step 2: Extracting Seeds



Step 3: Finding HSPs



Step 4: Combining HSPs



BLAST – Notes

- Listing words
  - Higher $\mathbf{T} \rightarrow$ lower sensitivity, faster execution time
- Extracting seeds
  - Use hash tables to make the process faster
- Finding HSPs
  - Only seeds located on the same diagonal with some other seeds located at a distance smaller than a threshold will be extended
- Gapped alignment
  - Will be triggered only for HSPs whose scores are higher than the threshold

Karlin-Altschul stastitics

If we search two sequences X and Y with a scoring matrix S_{ij} to identify the maximal-scoring segment pair, and if the following conditions hold:

1. The two sequences are i.i.d. and have respective background distributions P_X, and P_Y (can be the same),
2. The two sequence are effectively "long" or infinite and not too dissimilar in length,
3. The expected pairwise score sum_i,j $P_X(i)P_Y(j)S_{ij}$ is negative,
4. A positive score is possible, i.e. $P_X(i)P_Y(j)S_{ij}>0$ for some i and j.

Then Karlin-Altschul statistics tell us:

Karlin-Altschul statistics

- The maximal segment score has the close approximating distribution:

$$Prob(S > x) \approx= 1 - \exp(-K * \exp^{-\lambda * x})$$

where K and λ are constants that can be calculated according to

Karlin, S, and SF Altschul (1990), "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes", PNAS 87:2264-68

Karlin-Altschul statistics

- The scores in the scoring matrix are implicitly log-odds scores of the form:

$$S_{ij} = \log(Q_{ij} / (P_X(i) P_Y(j))) / \lambda$$

where Q_{ij} is the limiting target distribution of the letter pairs (i,j) in the MSP and λ is the unique positive-valued solution to the equation

$$\sum_{i,j} P_X(i) P_Y(j) e^{\lambda S_{ij}} = 1$$

- The expected frequency of chance occurrence of an MSP with score S or greater is:

$$E = KMNe^{-\lambda S}$$

Karlin-Altschul statistics

- Another way to express the scores in the scoring matrix:

$$S_{ij} = \log_b(Q_{ij} / (P_X(i) P_Y(j)))$$

where logarithms to some base b are used instead of Natural logarithms. Then λ is related to the base of the logarithms as follows:

$$\lambda \log_e b = 1$$

- The expected length of the MSP is

$$\boldsymbol{E(L) = \log(KMN)/H}$$

where H is the relative entropy of the target and background frequencies:

$$H = \sum_{i,j} (Q_{ij} \log(Q_{ij} / (P_X(i) P_Y(j))))$$

Karlin-Altschul statistics

- The expect score E of a database match is the number of times that an unrelated database sequence would obtain a score S higher than x by chance. (The relationship of P-value and E-value)

$$P \approx 1 - e^{-E}$$

- Normalized score for different database search

$$S' = \lambda S - \log K$$

then,

$$E = MNe^{-S'}$$

Karlin-Altschul statistics

- The "Edge Effect"

$$M' = M - E(L)$$

$$N' = N - E(L)$$

$$E' = KM'N'e^{-\lambda S}$$

Notes about the scores in Blast

- What does a big score mean?
- What you need to know about the scores
  - K, λ

Acknowledgement

- Some of the slides are from Dr. Guangyong Zheng, CAS