Chapter 4.
Sequence Comparison

Contents

1. Sequence comparison
2. Sequence alignment
3. Sequence mapping

Reading materials

Required

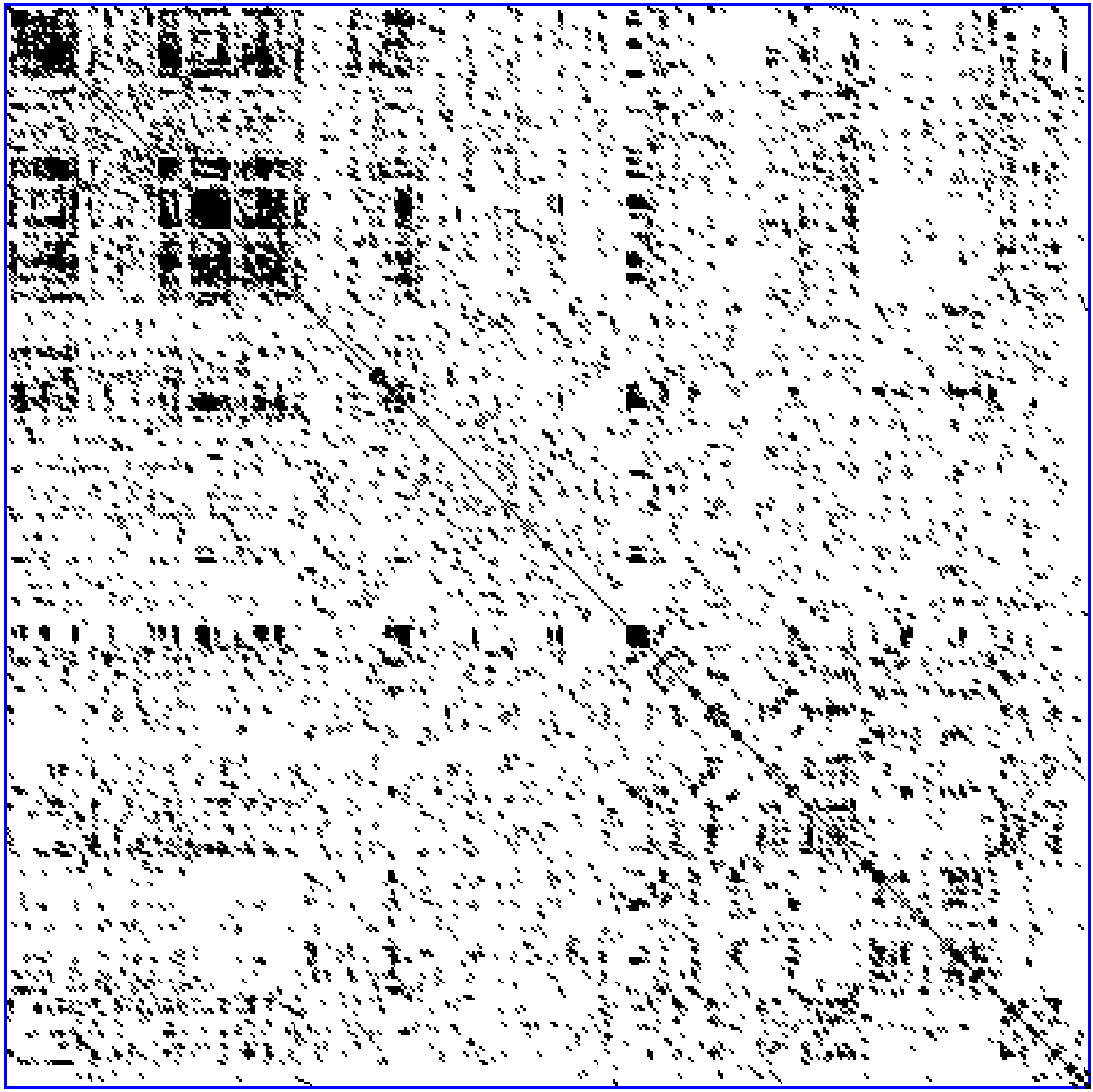
1. "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Needleman, SB and Wunsch, CD. J. Mol. Biol. 48:443-453, 1970
2. "Identification of Common Molecular Subsequences", Smith, TF and Waterman, MS. J. Mol. Biol. 147: 195-197, 1981
   The Smith/Waterman algorithm

3. https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Other recommended background:

1. "An improved algorithm for matching biological sequences", Gotoh, O. J. Mol. Biol. 162:705-708, 1982
   The efficient form of the Needleman/Wunsch and Smith/Waterman algorithms.
2. "Optimal alignment in linear space", Myers, E. W. and Miller, W. CABIOS 4: 11-17, 1988.
   More advanced reading: a divide and conquer method to reduce the memory cost from O(n^2) to O(n)

The simple but powerful dot plot



A DNA dot plot of a human zinc finger transcription factor (GenBank ID NM_002383), showing regional self-similarity

Sequence comparison algorithms

- Simple identity (as in C's strcmp())
- Hashing
- Longest common substring

Longest common substring

	Δ	C	A	G	C	C	U	C	G	C	U	U	A	G
Δ	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
A	0·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0
A	0·0	0·0	1·0	0·7	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·7
U	0·0	0·0	0·0	0·7	0·3	0·0	1·0	0·0	0·0	0·0	1·0	1·0	0·0	0·7
G	0·0	0·0	0·0	1·0	0·3	0·0	0·0	0·7	1·0	0·0	0·0	0·7	0·7	1·0
C	0·0	1·0	0·0	0·0	2·0	1·3	0·3	1·0	0·3	2·0	0·7	0·3	0·3	0·3
C	0·0	1·0	0·7	0·0	1·0	3·0	1·7	1·3	1·0	1·3	1·7	0·3	0·0	0·0
A	0·0	0·0	2·0	0·7	0·3	1·7	2·7	1·3	1·0	0·7	1·0	1·3	1·3	0·0
U	0·0	0·0	0·7	1·7	0·3	1·3	2·7	2·3	1·0	0·7	1·7	2·0	1·0	1·0
U	0·0	0·0	0·3	0·3	1·3	1·0	2·3	2·3	2·0	0·7	1·7	2·7	1·7	1·0
G	0·0	0·0	0·0	1·3	0·0	1·0	1·0	2·0	3·3	2·0	1·7	1·3	2·3	2·7
A	0·0	0·0	1·0	0·0	1·0	0·3	0·7	0·7	2·0	3·0	1·7	1·3	2·3	2·0
C	0·0	1·0	0·0	0·7	1·0	2·0	0·7	1·7	1·7	3·0	2·7	1·3	1·0	2·0
G	0·0	0·0	0·7	1·0	0·3	0·7	1·7	0·3	2·7	1·7	2·7	2·3	1·0	2·0
G	0·0	0·0	0·0	1·7	0·7	0·3	0·3	1·3	1·3	2·3	1·3	2·3	2·0	2·0

FIG. 1. H_{ij} matrix generated from the application of eqn (1) to the sequences A-A-U-G-C-C-A-U-U-G-A-C-G-G and C-A-G-C-C-U-C-G-C-U-U-A-G. The underlined elements indicate the trackback path from the maximal element 3·30.

Smith and Waterman, JMB, 1981, 147, 195-197

Analysis of algorithms and big-O notation

Measure the Complexity of an algorithm: O()

- strcmp: O(n)
- longest common substring: O(nm)

Pattern matching algorithms

- Brute force
- Knuth/Morris/Pratt: a finite state automata solution
- Regular expressions and nondeterministic finite state automata

Dynamic programming sequence alignment algorithms

- Needleman/Wunsch global alignment
- Smith/Waterman local alignment
- Linear and affine gap penalties

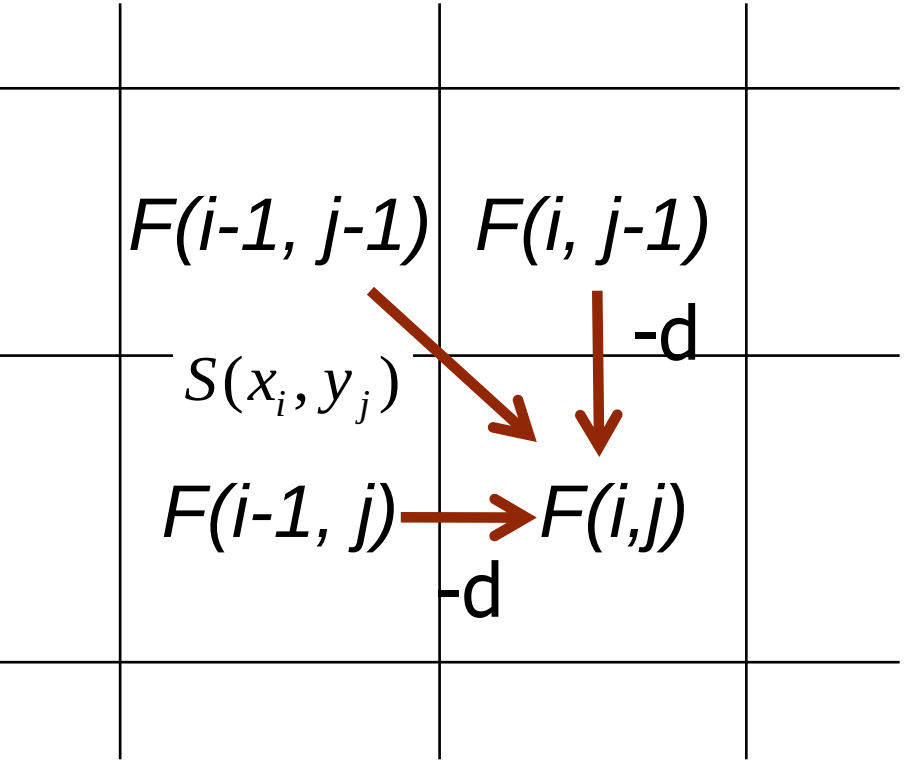
Needleman/Wunsch global alignment (1970)

- Two sequences $X = x_1 \ldots x_n$ and $Y = y_1 \ldots y_m$
- Let $F(i, j)$ be the optimal alignment score *of* $X_{1 \ldots i}$ of X up to x_i and $Y_{1 \ldots j}$ of Y up to Y_j *(*$0 \le i \le n$, $0 \le j \le m$*), then we have*

$$F(0,0) = 0$$

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

Needleman/Wunsch global alignment (1970)



$$F(0,0) = 0$$

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

Smith/Waterman local alignment (1981)

- Two sequences $X = x_1 \ldots x_n$ and $Y = y_1 \ldots y_m$
- Let $F(i, j)$ be the optimal alignment score *of* $X_{1 \ldots i}$ of X up to x_i and $Y_{1 \ldots j}$ of Y up to Y_j *(*$0 \le i \le n$, $0 \le j \le m$*), then we have*

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} 0 \\ F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

Linear and affine gap penalties

- Linear: $w(k) = k\,d$
- Affine: $w(k) = d + (k-1)\,e$
- Let $M(i,j)$, $I_x(i,j)$, $I_y(i,j)$ be the best scores up to (i,j):
  - $M(i,j)$: x_i is aligned to y_j;
  - $I_x(i,j)$: x_i is aligned to a gap;
  - $I_y(i,j)$ y_j is aligned to a gap

then we have

$$M(i,j) = \max\begin{cases} M(i-1,j-1) + s(x_i, y_j), \\ I_x(i-1,j-1) + s(x_i, y_j), \\ I_y(i-1,j-1) + s(x_i, y_j); \end{cases}$$

$$I_x(i,j) = \max\begin{cases} M(i-1,j) - d, \\ I_x(i-1,j) - e; \end{cases}$$

$$I_y(i,j) = \max\begin{cases} M(i,j-1) - d, \\ I_y(i,j-1) - e. \end{cases}$$

Contents

1. Sequence comparison
2. Sequence alignment
3. Sequence mapping

Sequence Alignment

BLAST

Blast

- **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
  - **B**asic **L**ocal **A**lignment **S**earch **T**ool
- Why BLAST
  - Quickly search a sequence database

Alignment in Real Life (30 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 long target sequence requires to evaluate about 10^{11} matrix cells…
  - … which will take about three hours in the rate of 10^6 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
```

WARNING: HSPs involving 198 database sequences were not reported due to the limiting value of parameter B = 250.

NOTE: You may want to consider using a low-complexity sequence filter to reduce the number of spurious matches that may be appearing in the output. See the filter option at http://blast.advbiocomp.com/doc/parameters.html#filter.

Parameters:

ctxfactor=1.00
E=10

Query Frame	MatID	Matrix name	As Used Lambda	As Used K	As Used H	Computed Lambda	Computed K	Computed H
+0	0	BLOSUM62	0.318	0.135	0.401	same	same	same
		Q=9,R=2	0.244	0.0300	0.180	n/a	n/a	n/a

Query Frame	MatID	Length	Eff.Length	E	S	W	T	X	E2	S2
+0	0	282	282	9.9	67	3	11	22	0.38	34
					43				0.42	37

Statistics:

Statistics:

Database: ../C.elegans/Proteome/ws_215.protein
Title: ws_215.protein
Posted: 4:57:53 PM CST Mar 9, 2017
Created: 9:21:18 PM CST Oct 27, 2016
Format: XDF-1
# of letters in database: 10,879,267
# of sequences in database: 24,705
# of database sequences satisfying E: 448
No. of states in DFA: 619 (141 KB)
Total size of DFA: 361 KB (2136 KB)
Time to generate neighborhood: 0.00u 0.00s 0.00t Elapsed: 00:00:00
No. of threads or processors used: 1
Search cpu time: 2.22u 0.00s 2.22t Elapsed: 00:00:03
Total cpu time: 2.27u 0.00s 2.27t Elapsed: 00:00:03
Start: Thu Mar 16 17:08:03 2017 End: Thu Mar 16 17:08:06 2017
NOTES ISSUED: 1
WARNINGS ISSUED: 1

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

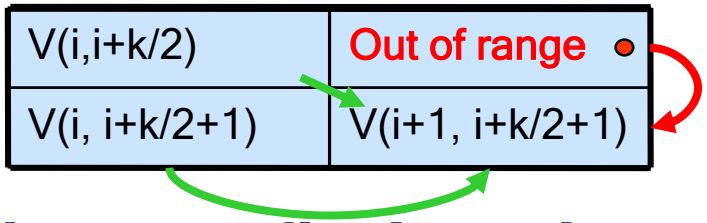
Basic Intuition 1: Seeds

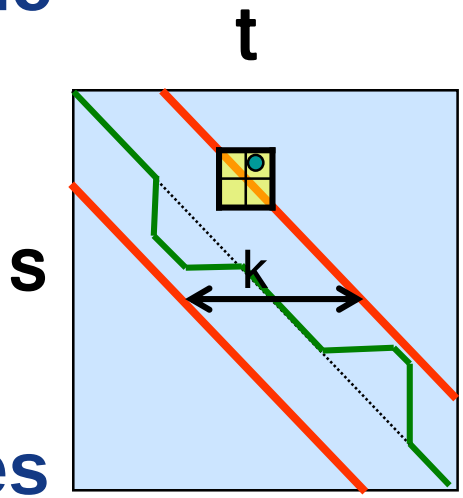
- **Observation: Real-life matches often contain long strings with gap-less matches**

- **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**



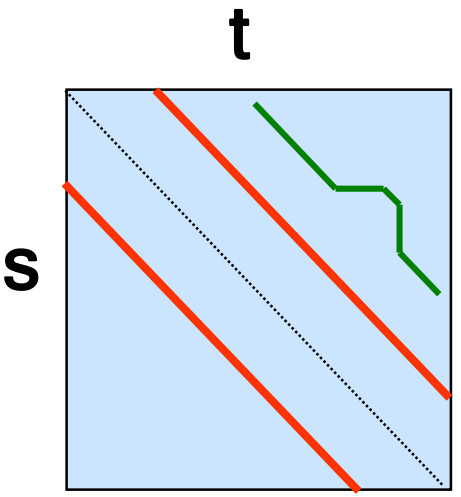


- **Action: To find such a path, it suffices to search in a diagonal band of the matrix.**
  - **If the diagonal band consists of `k` diagonals (width `k`), then dynamic programming takes $O(kn)$.**
  - **Much faster than $O(n^2)$ of standard DP.**

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**

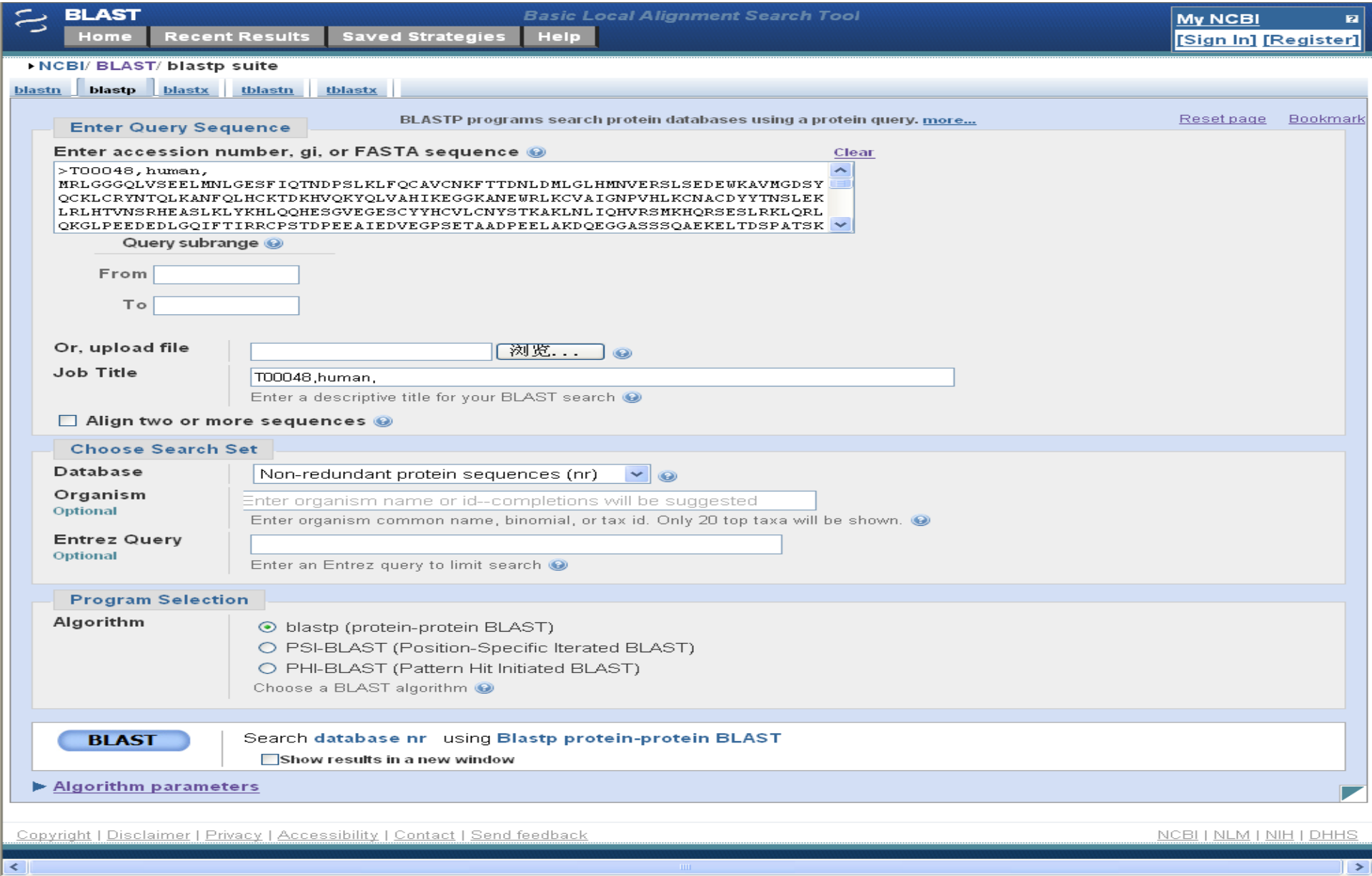


Basic Local Alignment Search Tool (BLAST)

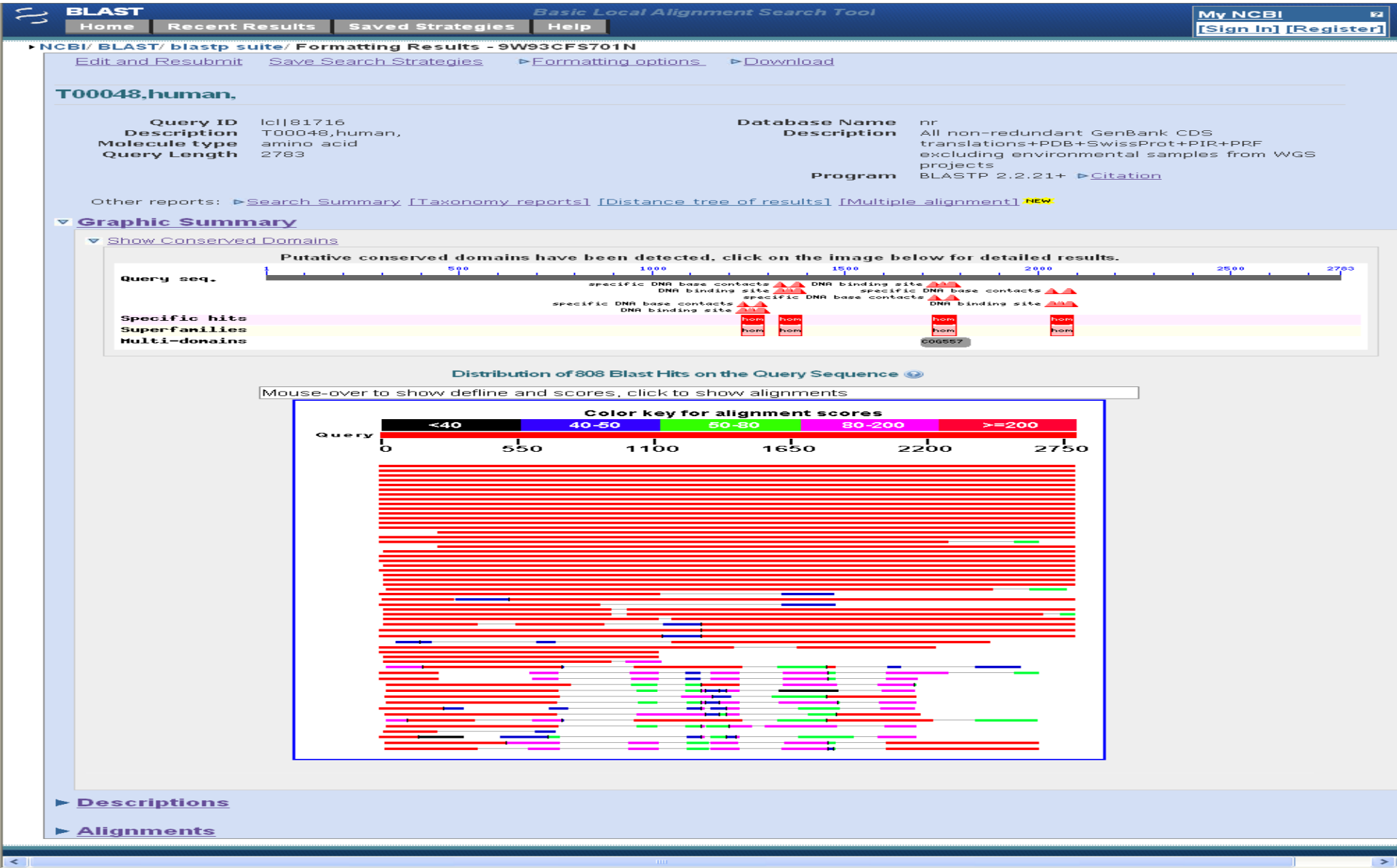
- **Publications:**
  - Ungapped BLAST – Altschul et al., 1990
  - Gapped BLAST, PSI-BLAST - Altschul et al., 1997
- **Input:**
  - Query (target) sequence – either DNA, RNA or Protein
  - Scoring Scheme – gap penalties, substitution matrix for proteins, identity/mismatch scores for DNA/RNA
  - Word 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 – 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

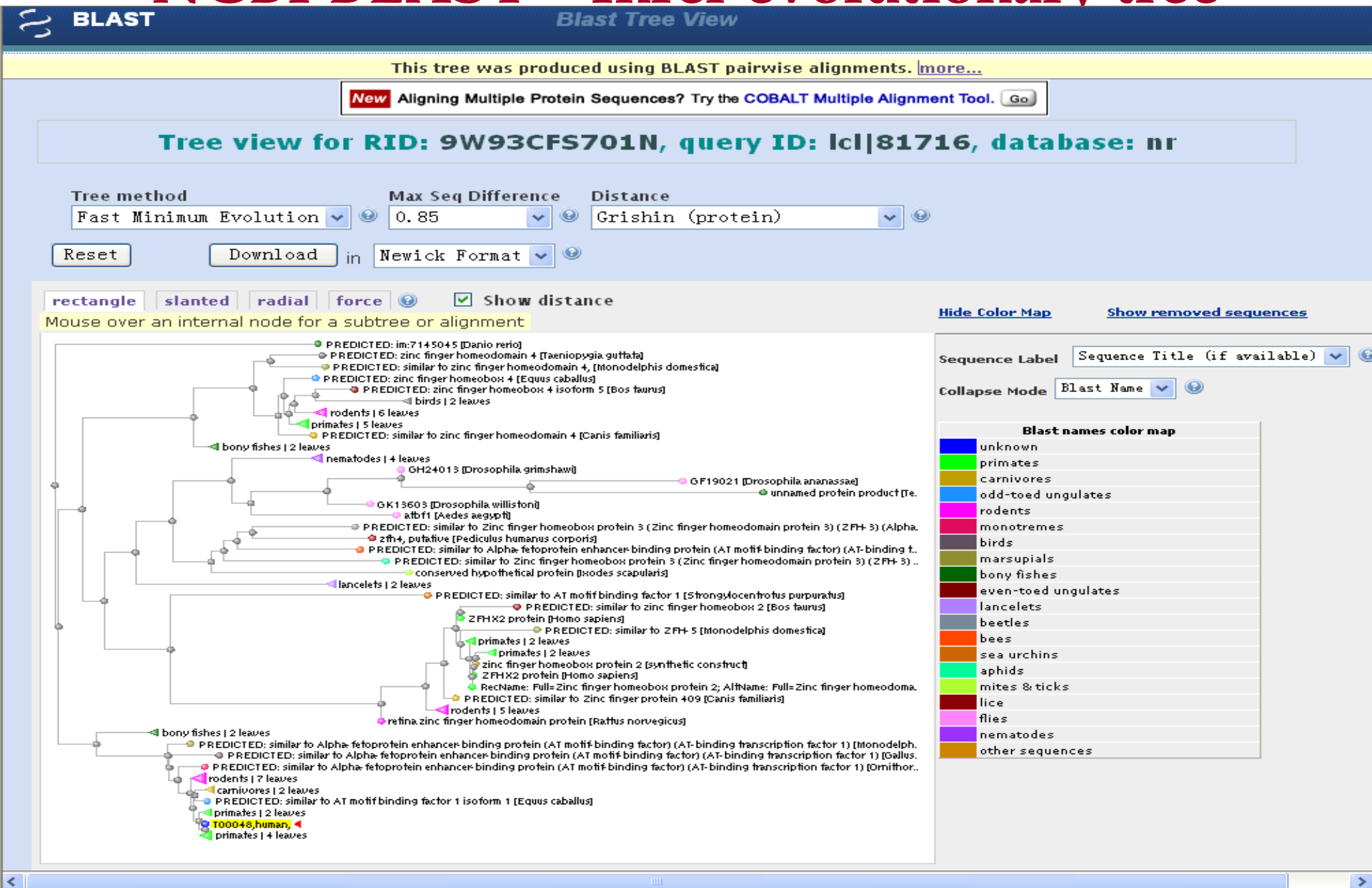
▶ Graphic Summary

▼ Descriptions

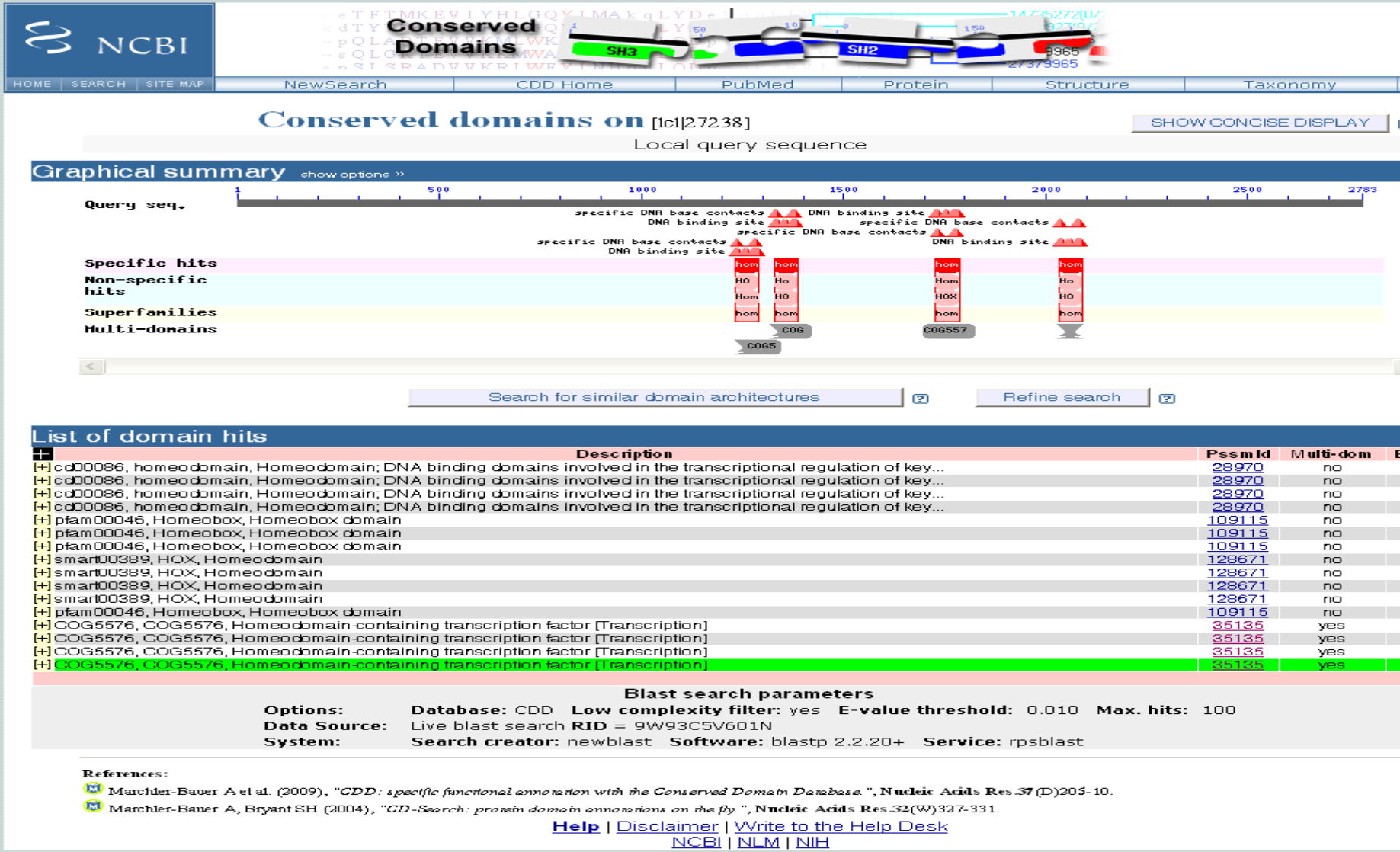
Sequences producing significant alignments:	Score (Bits)	E Value	
dbj|BAA01095.1| alpha-fetoprotein enhancer binding protein [H...	5710	0.0	G
gb|AAC14462.1| zinc finger homeodomain protein [Homo sapiens]...	4975	0.0	G
ref|NP_008816.3| AT-binding transcription factor 1 [Homo sapi...	4975	0.0	G
ref|XP_001102516.1|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

▶ Alignments

NCBI BLAST – Infer evolutionary tree



NCBI BLAST – construct families

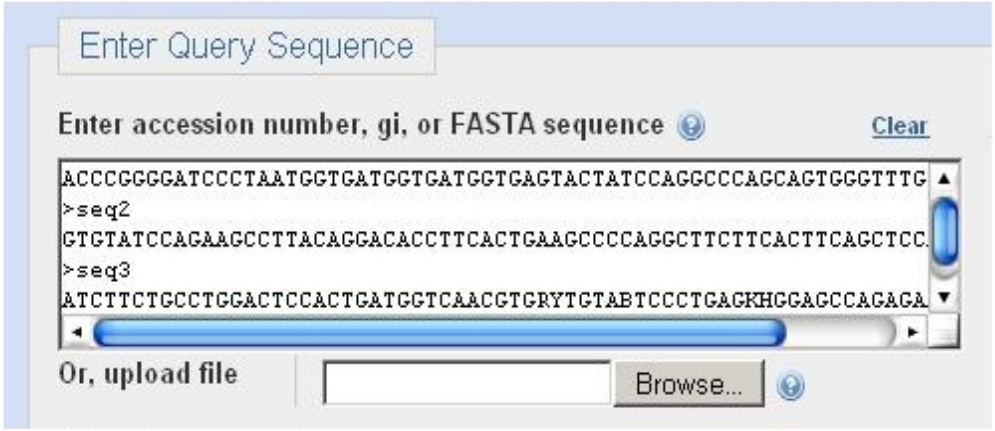


NCBI BLAST batch jobs

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

Run BLAST in command line

BLAST

(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...
```

Run BLAST in 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
```

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

```
ATCGGGCTACTG
ACCGGCTACTGA

ATCGGGCTACTG -
|  *  | |||||||||
ACC - GGCTACTGA
```

Edit distance 3

Mathematic model of sequence alignment

How to quantify the distance

Scoring

Simple scoring function

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

Matrix for scoring

Matrix for nucleic acid sequence alignment

Matrix for amino acid sequence alignment

Mathematic model of sequence alignment

Matrix for nucleic acid sequence alignment

(1) equivalence matrix

(2) BLAST matrix

(3) transition-transversion matrix

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

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

	A	T	C	G
A	1	-5	-5	-1
T	-5	1	-1	-5
C	-5	-1	1	-5
G	-1	-5	-5	1

Mathematic model of sequence alignment

Matrix for amino acid sequence alignment

(1) equivalence 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

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

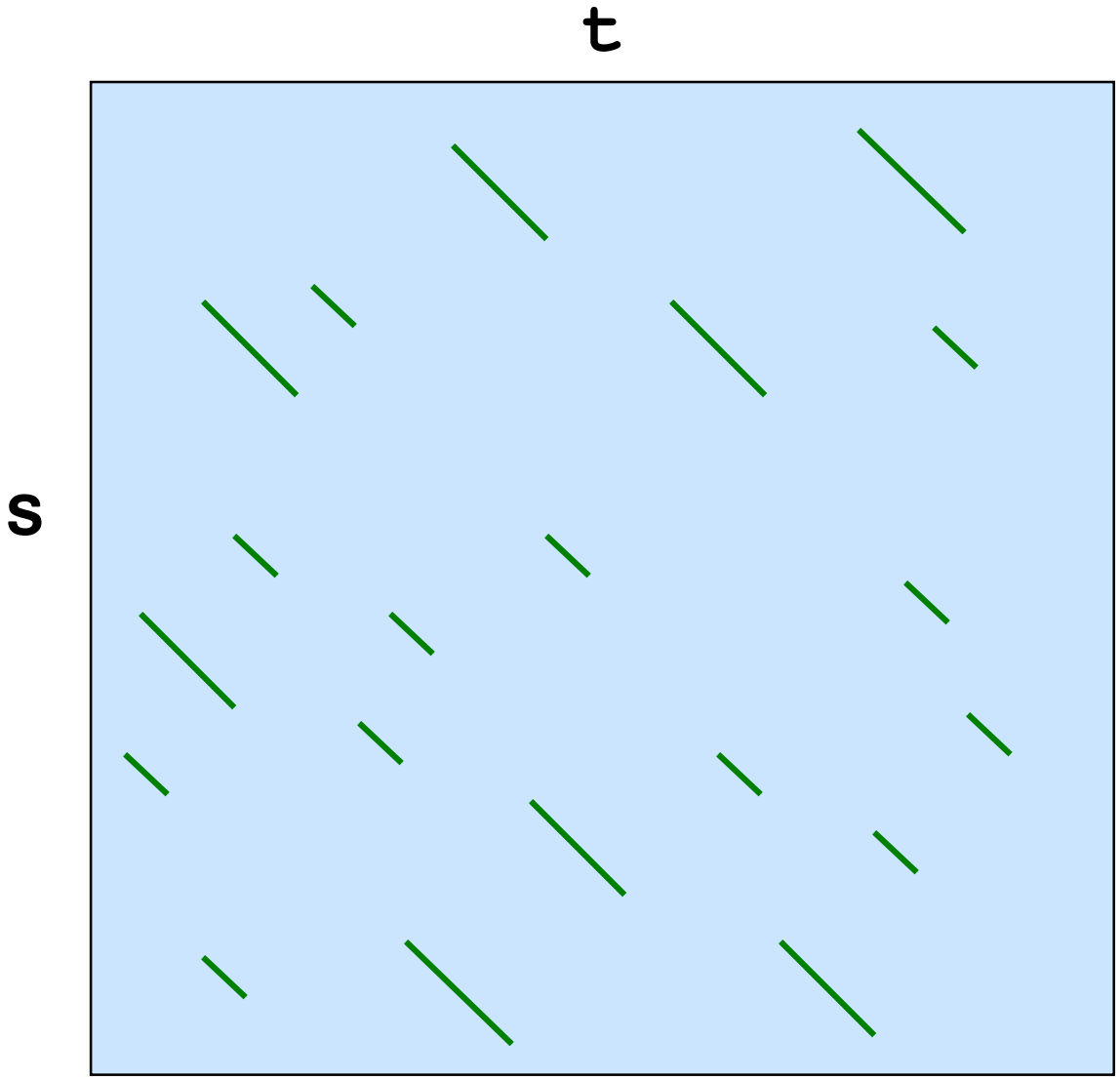
...GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDK...

High scoring words

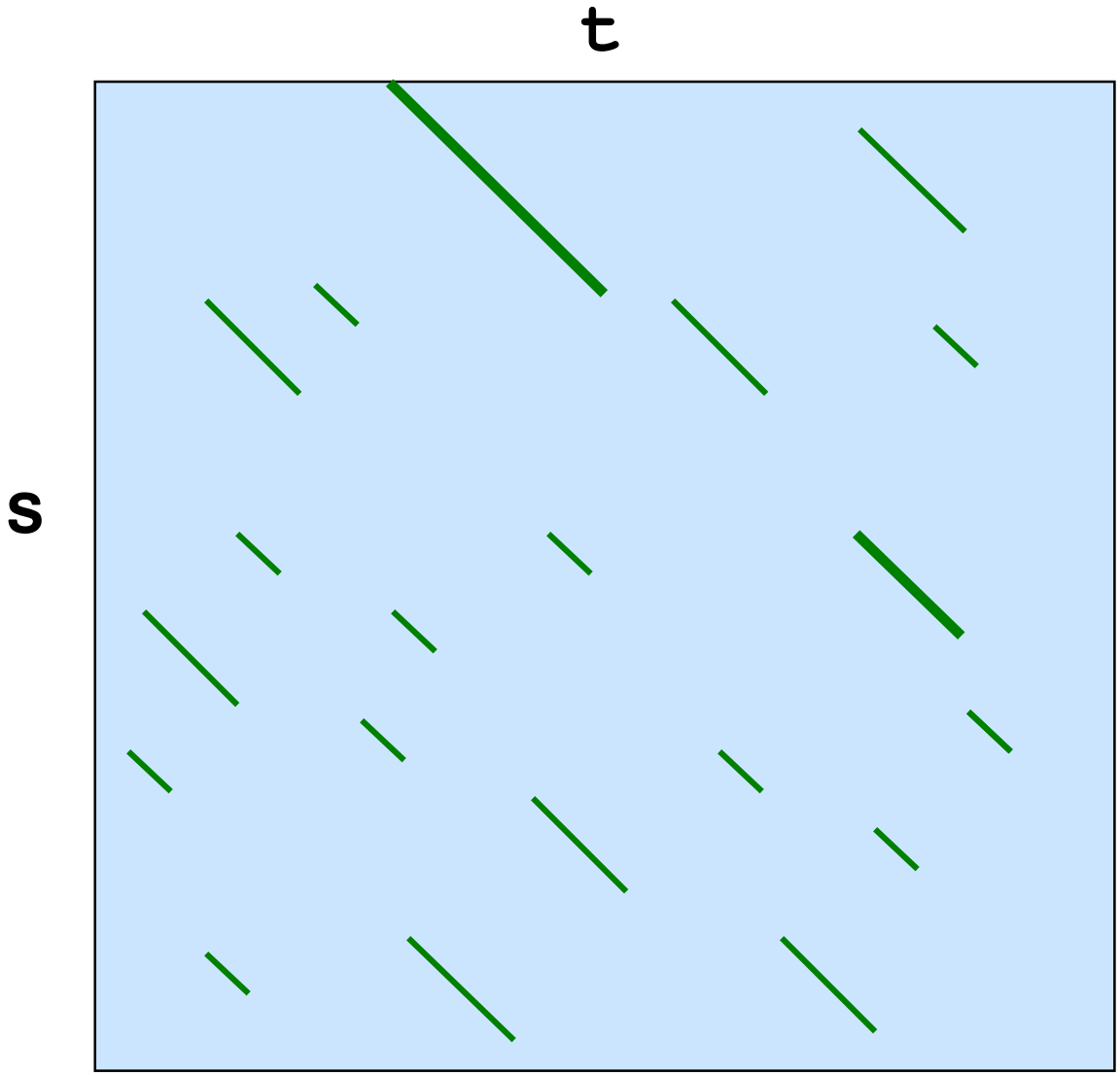
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12

Score threshold T=13

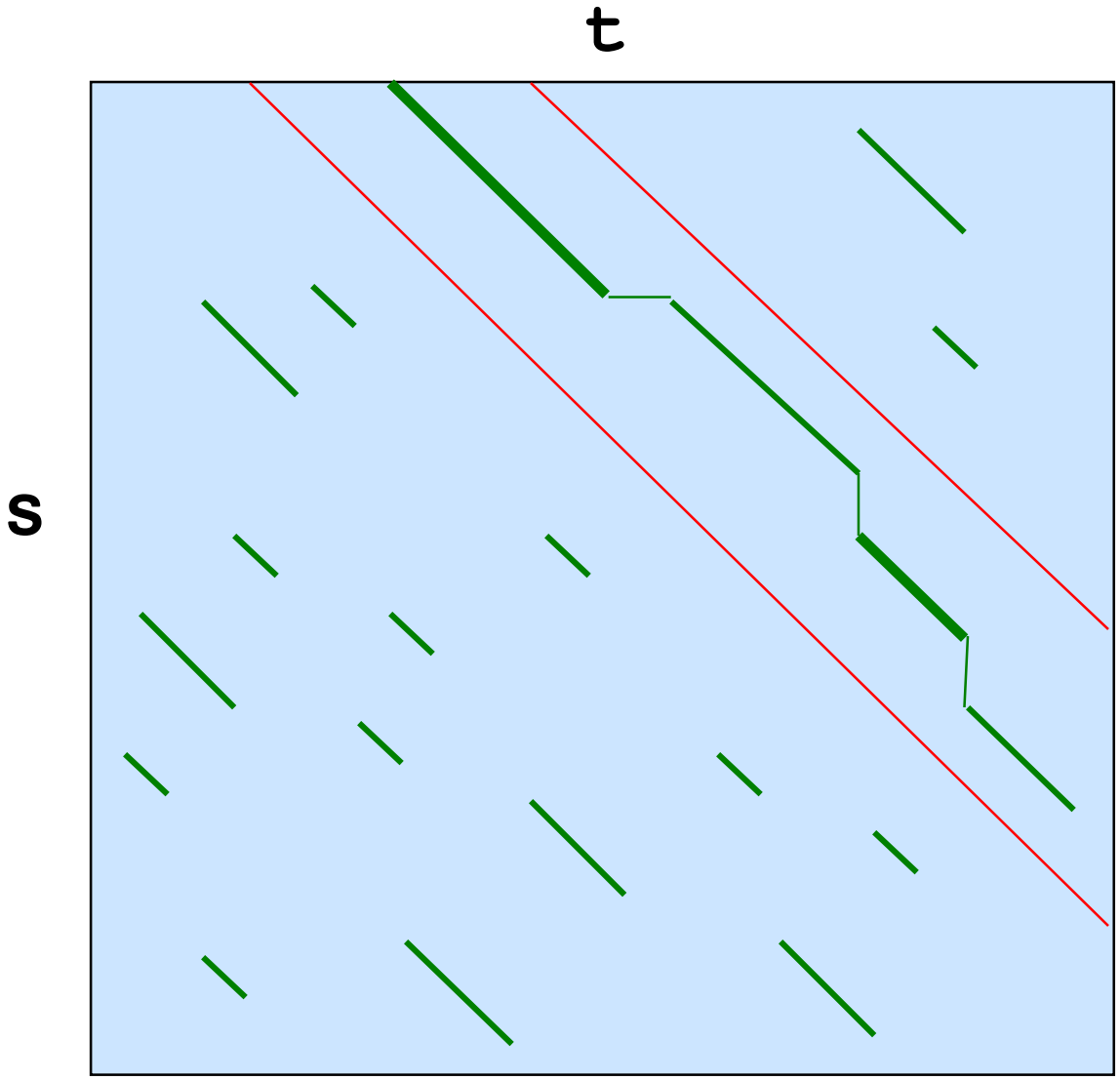
Step 2: Extracting Seeds



Step 3: Finding HSPs



Step 4: Combining HSPs



BLAST – Notes

- Listing words
  - Higher $T \rightarrow$ 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

https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

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:

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

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

$$E(L) = \log(KMN)/H$$

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

Sequence alignment or mapping

Chaochun Wei

Spring 2018

Sequence alignment tools other than BLAST

- BLAT
- Bowtie
- BWA

Reading

- Kent, WJ (2002), "BLAT– the BLAST-like alignment tool", Genome Research, 12(4):656-64
- Blat FAQ: http://genome.ucsc.edu/FAQ/FAQblat.html

- Trapnell, C. and Salzberg, S. (2009), "How to map billions of short reads onto genomes", Nature Biotechnology, 27(5)455-457
- Li, Ruiqiang, Li, Yingrui, Kristiansen, Karsten, and Wang, Jun (2008), "SOAP:short oligonucleotide alignment program", Bioinformatics, 24(5)713-714.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009), "Ultrafast and memory-efficient alignment of short DNA sequence to the human genome", Genome Biology,10:R25.
- Trapnell, C., Pachter, L., and Salzberg, S. (2009), "TopHat: discovering splice junctions with RNA-seq", Bioinformatics, 25(9)1105-1111.

BLAT: Blast-Like Alignment Tool

- Not BLAST
- Indexed on database， stored in memory
  - Need ~1G memory for human genome
- Need some extra time for database initialization (index)
- Can be 500 times faster than BLAST
- Can display results in the UCSC genome browser

Kent, WJ (2002), “BLAT– the BLAST-like alignment tool”, Genome Research, 12(4):656-64

Blat FAQ: http://genome.ucsc.edu/FAQ/FAQblat.html

BLAT

- **Designed to quickly find**
  - **DNA sequences of 95% and greater similarity of length 25 bases or more.**
  - **Protein sequences of 80% and greater similarity of length 20 amino acids or more.**
- **In practice**
  - **DNA BLAT works well on primates, and**
  - **protein blat on land vertebrates**

BLAT—The BLAST-Like Alignment Tool

Timing of BLAT vs.WU-TBLASTX on a Data Set of 1000 Mouse Reads against a RepeatMasked Human Chromosome 22

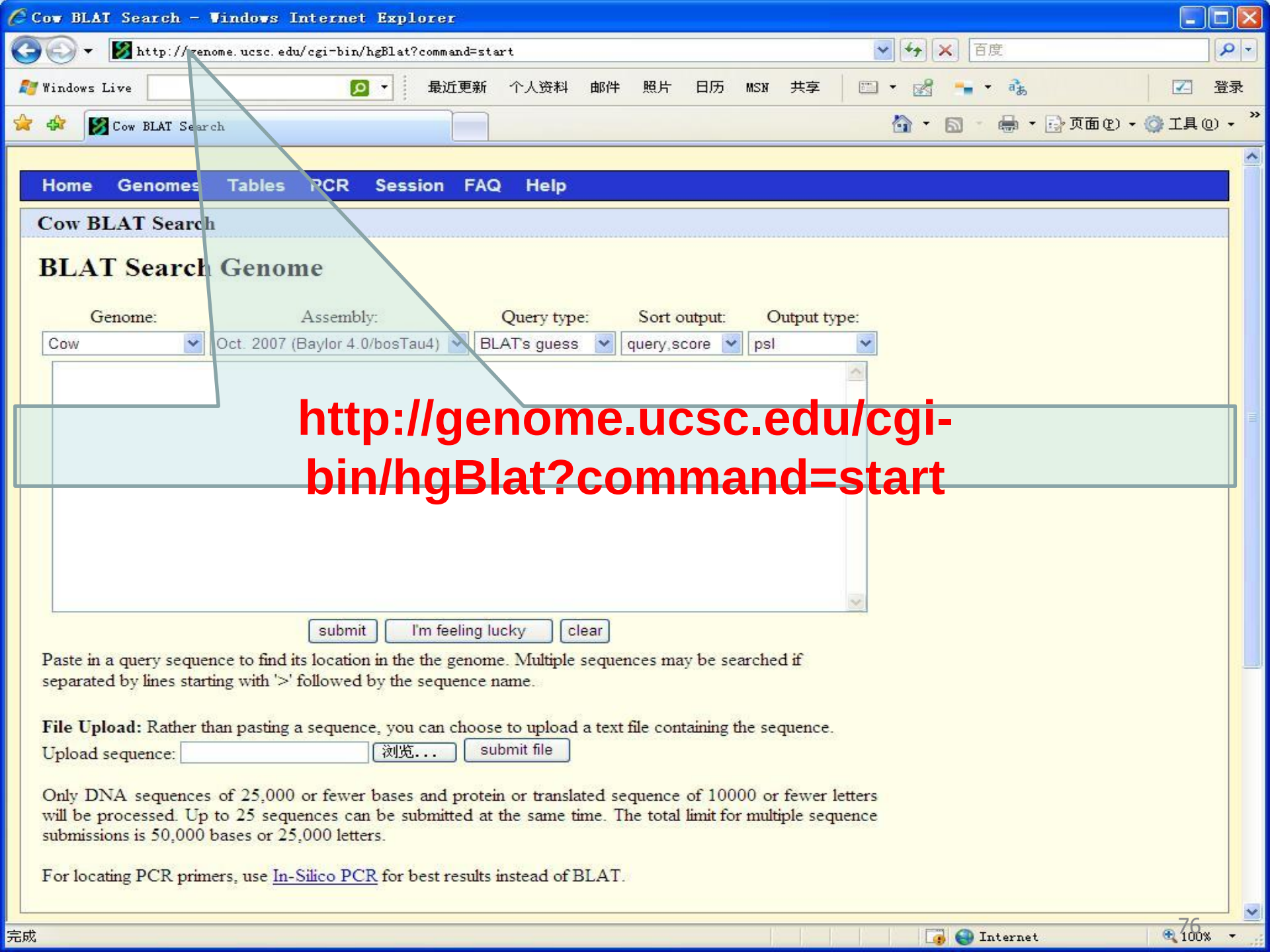
Method	K	N	Matrix	Time
WU-TBLASTX	5	1	+15/−12	2736 s
WU-TBLASTX	5	1	BLOSUM62	2714 s
BLAT	5	1	+2/−1	61 s
BLAT	4	2	+2/−1	37 s

K: the size of the perfectly matching as a seed for an alignment
N: the number of hits in a gapless 100-aa window required to trigger a detailed alignment.
Matrix: column describes the match/mismatch scores or the substitution score matrix used.

Comparison of sequencing platforms (2019.2)

Platforms	Sanger	454	HiSeq X Ten *	MiSeq *	NovaSeq 6000*	PacBio Sequel **	Nanopore
Read length	650-1100	150-1000	150	36-300	2x250	Average 30k	Up to 2 Mb
# of reads/run	96	0.4-2M	5.3-6 B	15M – 25M	32-40B	~500k	Up to 500
Error rate	10^-3	<10^-2	~10^-3	~10^-3	~10^-3	~1%	Varies
Cost ($/Mbp)	5000	~5	<0.01	~0.5	<0.001	~0.3	~0.1
Time/run	~3 hours	~7 hours	<3 days	4-56 hours	13-38hr	< 20 hours	As little as 5 mins
Throughput	100Kb	~1Gb	1.6-1.8Tb	540Mb-15Gb	4.8-6Tb	Up to 20 Gb	10-30Gb

* https://www.illumina.com/systems/
** https://www.pacb.com/products-and-services/sequel-system/


Cow BLAT Search - Windows Internet Explorer
http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
Cow BLAT Search
Home Genomes Tables PCR Session FAQ Help
Cow BLAT Search
BLAT Search Genome
Genome:
Assembly:
Query type:
Sort output:
Output type:
Cow
Oct. 2007 (Baylor 4.0/bosTau4)
BLAT's guess
query,score
psl
submit
I'm feeling lucky
clear
Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.
File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence:
浏览...
submit file
Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.
For locating PCR primers, use In-Silico PCR for best results instead of BLAT.
http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

How to map billions of short reads onto genomes

Cole Trapnell & Steven L Salzberg

Mapping the vast quantities of short sequence fragments produced by next-generation sequencing platforms is a challenge. What programs are available and how do they work?

A new generation of DNA sequencers that can rapidly and inexpensively sequence billions of bases is transforming genomic science. These new machines are quickly becoming the technology of choice for whole-genome sequencing and for a variety of sequencing-based assays, including gene expression, DNA-protein interaction, human resequencing and RNA splicing studies[1–3]. For example, the RNA-Seq protocol, in which processed mRNA is converted to cDNA and sequenced, is enabling the identification of previously unknown genes and alternative splice variants; the ChIP-Seq approach, which sequences immunoprecipitated DNA fragments bound to proteins, is revealing networks of interactions between transcription factors and DNA regulatory elements[4]; and the whole-genome sequencing of tumor cells is uncovering previously unidentified cancer-

Table 1 A selection of short-read analysis software

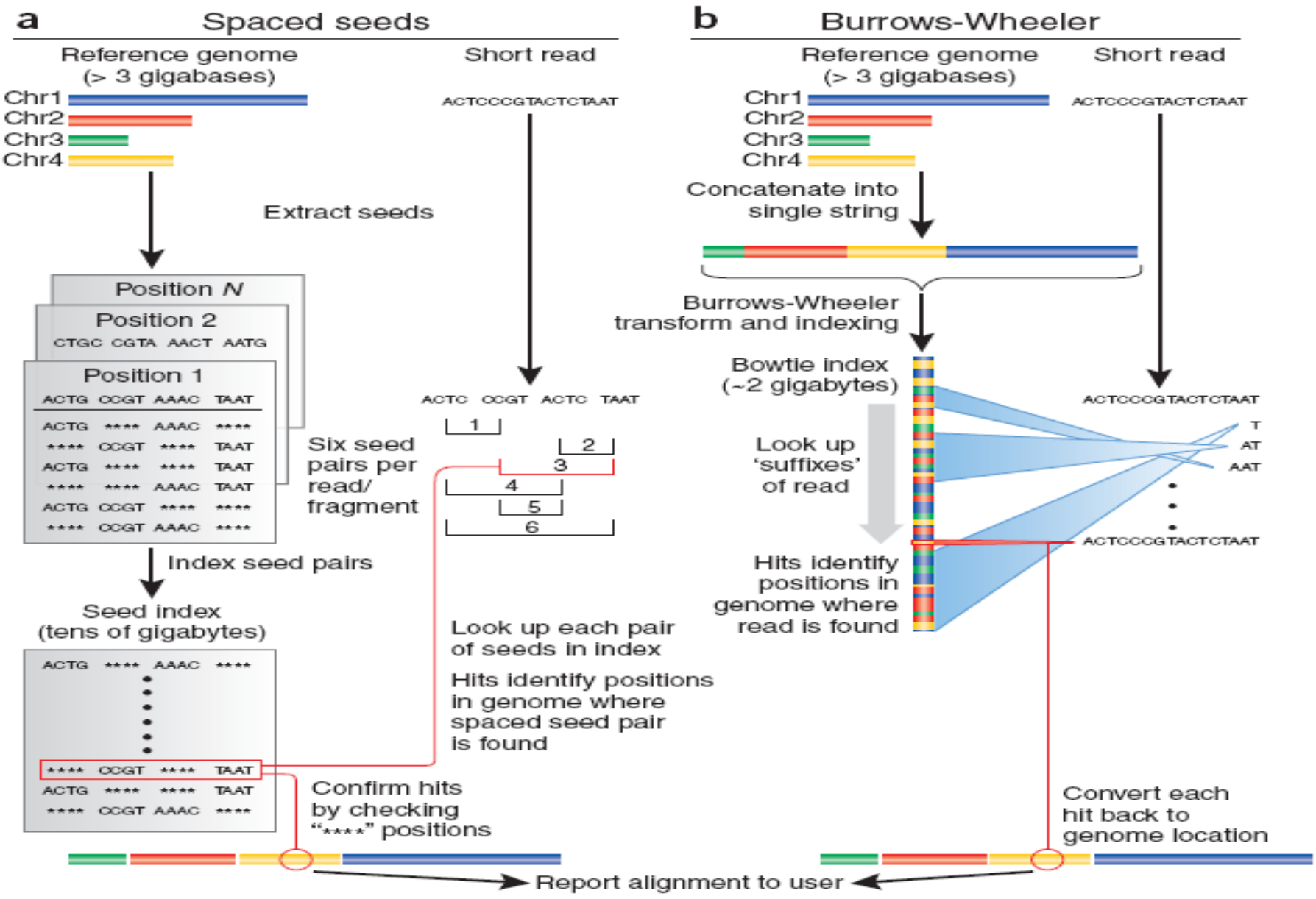
Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

In this case, to make sense of the reads, their positions within the reference sequence must be determined. This process is known as aligning or 'mapping' the read to the reference. In one version of the mapping problem, reads must be aligned without allowing large gaps in to understand why the mapping problems are computationally difficult, which difficulties have been overcome and what challenges and opportunities remain.

Challenges of mapping short reads

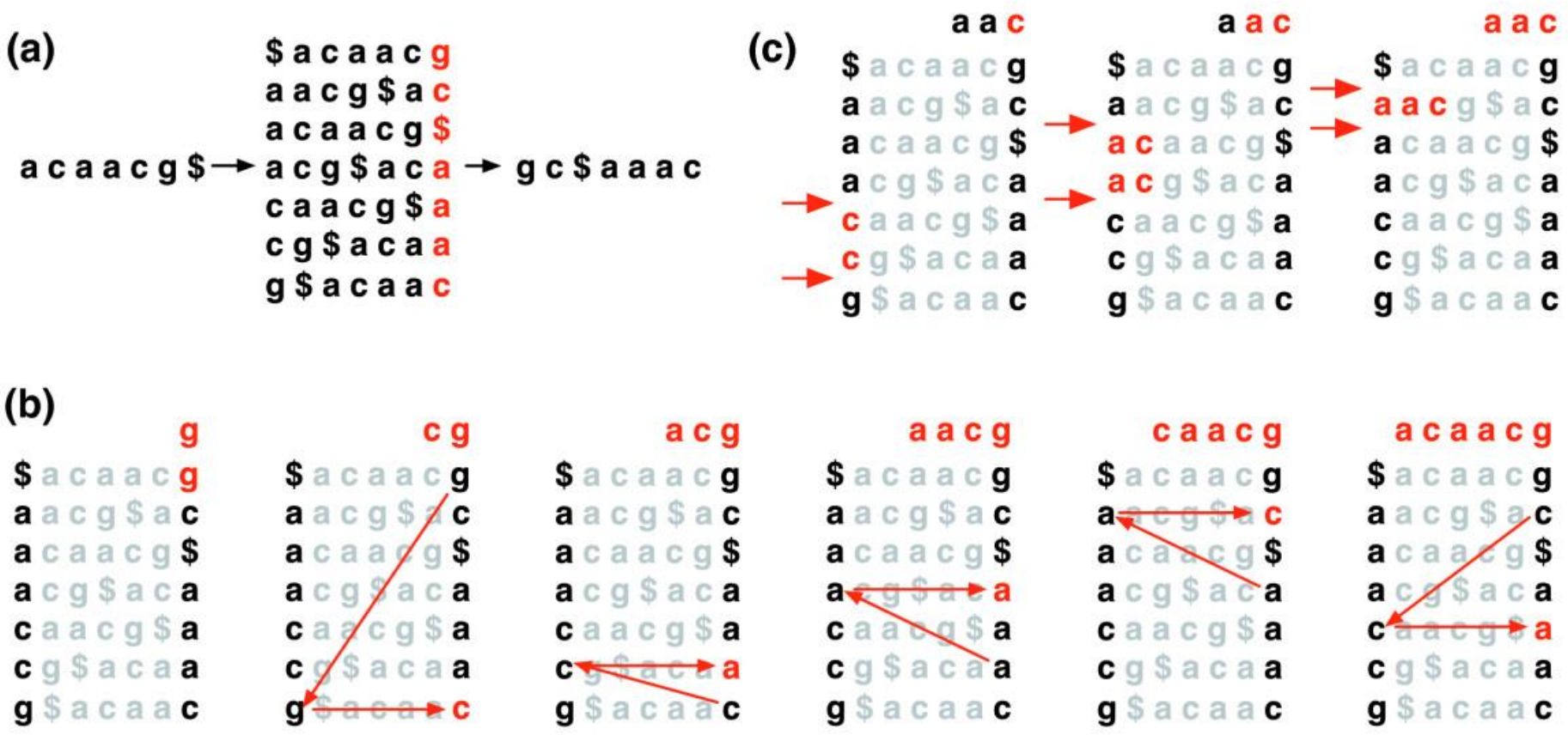
Sequence alignment/mapping

- Aligning (mapping) billions of short reads
  - Bowtie
  - SOAP
  - BWA
  - Tophat



Algorithms (a) based on spaced-seed indexing; (b) based on Burrows-Wheeler transform

Bowtie (Burrows-Wheeler transform)



(a) The Burrows-Wheeler matrix and transformation for 'acaacg'. **(b)** Steps taken by EXACTMATCH to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'. **(c)** UNPERMUTE repeatedly applies the last first (LF) mapping to recover the original text (in red on the top line) from the Burrows-Wheeler transform (in black in the rightmost column).

Langmead *et al. Genome Biology* 2009 **10**:R25 doi:10.1186/gb-2009-10-3-r25

Bowtie versus SOAP v1.10 and Maq v0.6.6

	Platform	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
Bowtie -v 2	Server	15 m 7 s	15 m 41 s	33.8	1,149	-	67.4
SOAP		91 h 57 m 35 s	91 h 47 m 46 s	0.10	13,619	351×	67.3
Bowtie	PC	16 m 41 s	17 m 57 s	29.5	1,353	-	71.9
Maq		17 h 46 m 35 s	17 h 53 m 7 s	0.49	804	59.8×	74.7
Bowtie	Server	17 m 58 s	18 m 26 s	28.8	1,353	-	71.9
Maq		32 h 56 m 53 s	32 h 58 m 39 s	0.27	804	107×	74.7

The performance and sensitivity when aligning 8.84 M reads from the 1,000 Genome project (NCBI Short Read Archive: SRR001115) trimmed to 35 base pairs.

Langmead *et al. Genome Biology* 2009 **10**:R25 doi:10.1186/gb-2009-10-3-r25

Comparison of short-read mapping methods

Software	data base size	Index time	Index size	Reads	time	Result size	Mapped
MAQ	70M	--	34.3M	20,000	112s	270.8K	19,576
SOAP	70M	91.46s	792.8M	20,000	0.46s	1.9M	12,564/ 12,954
SOAP	70M	91.46s	792.8M	3,251,337	83.97s	316.1M	2,055,104/ 2,120,025
bowtie	70M	220s	81.9M	20,000	<1s	2.7M	18,958/ 19,573
bowtie	70M	220s	81.9M	3,251,337	77.2s	431.7M	3,086,705/ 3,184,978

Reads: simulated by MetaSim version 0.9.1

Index: converted from virus sequences from NCBI

Comparison of short-read mapping methods

		GPU												CPU											
Dataset	Volume (Gbp)	SOAP3-dp (Full SA)			SOAP3			BarraCUDA			CUSHAW			BWA			Bowtie2			SeqAlto			GEM			CUSHAW2		
		%	Time (s)	Mem.	%	Time (fold)	Mem.	%	Time (fold)	Mem.	%	Time (fold)	Mem.	%	Time (fold)	Mem.	%	Time (fold)	Mem.	%	Time (fold)	Mem.	%	Time (fold)	Mem.	%	Time (fold)	Mem.
realYHPE100	12.24	98.12%	1,079	19	−10.60%	2.63	21.1	−6.39%	14.03	4.1	−3.13%	46.79	2.7	−4.14%	16.03	4.9	−3.87%	12.04	3.5	−0.48%	14.25	7.2	−2.86%	3.54	4.5	−1.39%	12.09	3.6
realYHPE150	56.23	97.16%	6,835	19.6	−28.33%	0.64	22.7	−10.99%	7.22	4.3	−10.11%	24.57	3.1	−8.05%	15.26	5.0	−7.45%	7.82	3.5	−3.65%	17.69	7.2	−5.85%	3.76	5	−6.47%	8.04	3.6
SRR211279	5.07	97.21%	439	18.4	−6.83%	1.05	20.8	−5.08%	13.42	4.1	−2.22%	54.22	2.1	−2.81%	16.27	4.9	−0.69%	12.00	3.5	−0.48%	17.29	7.2	−2.44%	4.11	4.6	−2.14%	12.03	3.6

The percentage of reads aligned and time consumption of aligners other than SOAP3-dp are recorded as the difference or ratio based on SOAP3-dp's figures. The '%' column represents 'Properly paired' for PE reads. 'Mem.' represents the peak memory consumption in gigabytes during alignment.

doi:10.1371/journal.pone.0065632.t001

Luo R, Wong T, Zhu J, Liu C-M, et al. (2013) SOAP3-dp: Fast, Accurate and Sensitive GPU-Based Short Read Aligner. PLoS ONE 8(5): e65632. doi:10.1371/journal.pone.0065632
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0065632

Summary

- **Long sequence alignment**
  - **BLAST (accurate, but slow)**
  - **BLAT (faster)**
- **Short read mapping**
  - **Bowtie (fast)**
  - **BWA (fast)**
  - **SOAP3 (faster)**

Acknowledgement

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