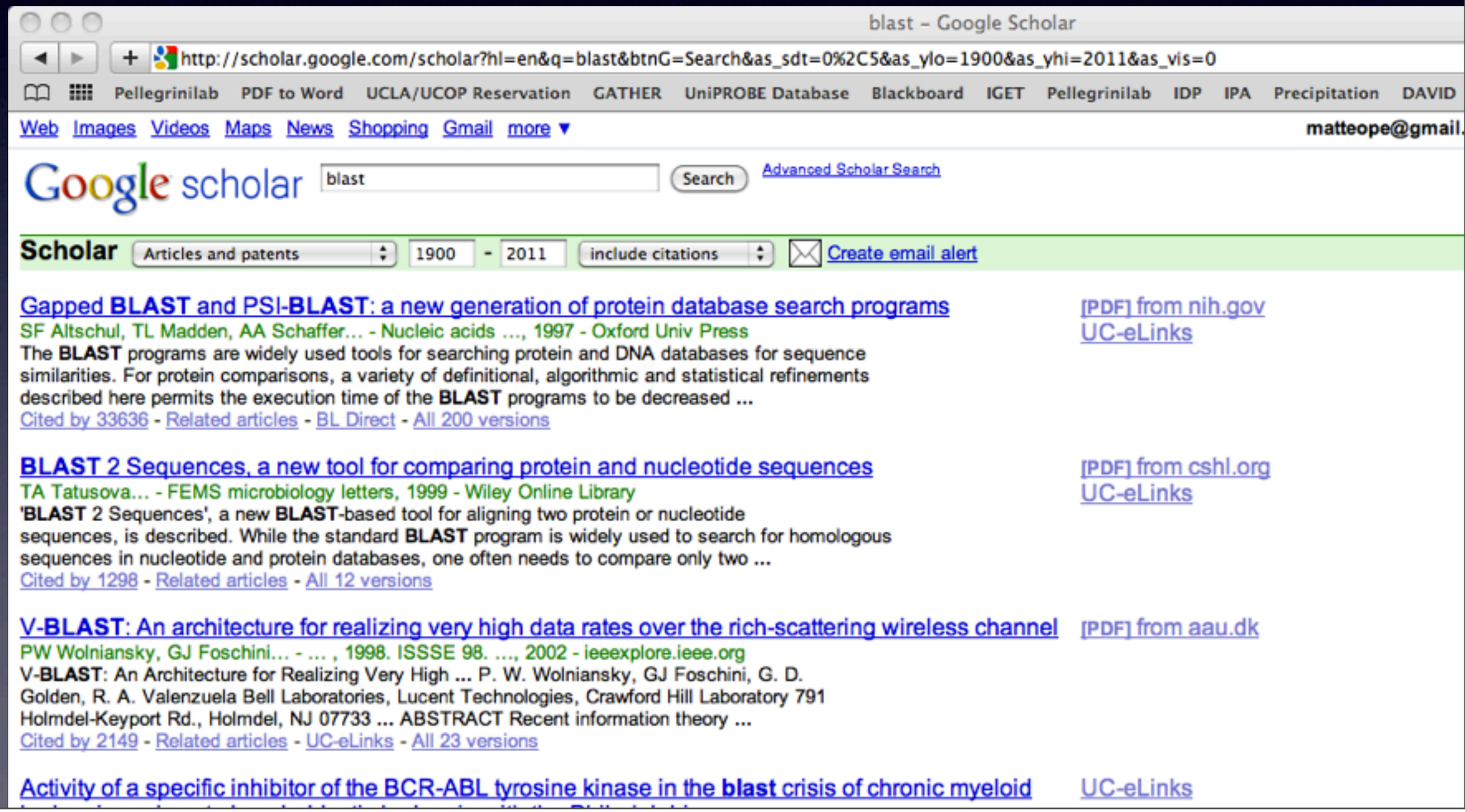
BLAST

MCDB 187

BLAST

- Basic Local Alignment Sequence Tool
- Uses shortcut to compute alignments of a sequence against a database very quickly
- Typically takes about a minute to align a sequence against a database of one million sequences
- Accurately computes the statistical significance of the alignment

One of most highly cited papers



Global vs Local Alignments

- In 1981 Smith and Waterman introduced a modification to the Needleman-Wunsch algorithm to search for optimal alignment fragments
- Positions in the alignment matrix with negative scores are set to zero, and are considered new start positions

Definitions

- Maximum Segment Pair (MSP): highest scoring pair of identical length segments from 2 sequences
- Word pair: segment pair of fixed length w
- T: minimum word pair score

Algorithm

- Pre-compute all words that score above T against the query sequence
- Search for these words in a database
- For each match, extend the alignment to generate the maximum segment pair
- Compute statistics of the resulting score

Step 1 : break up sequence into words

- ATCGTCTATTCCCGG
- w=4 letter words
  - ATCG
  - TCGT
  - CGTC
  - etc.

Step 2: find high scoring matches

- Start with word 1: ATCG
- score identities as +1, and mismatches a 0
- find all words that have a score of at least T=3
  - ATCA
  - ATCT
  - ATCC
  - etc.

Search for high scoring words in database

- Find all sequences in the database that contain a high scoring word
- For example:
- CCGGCT**ATCA**TCTATTCCCGGTTCG

Extend ALignments Around matching word

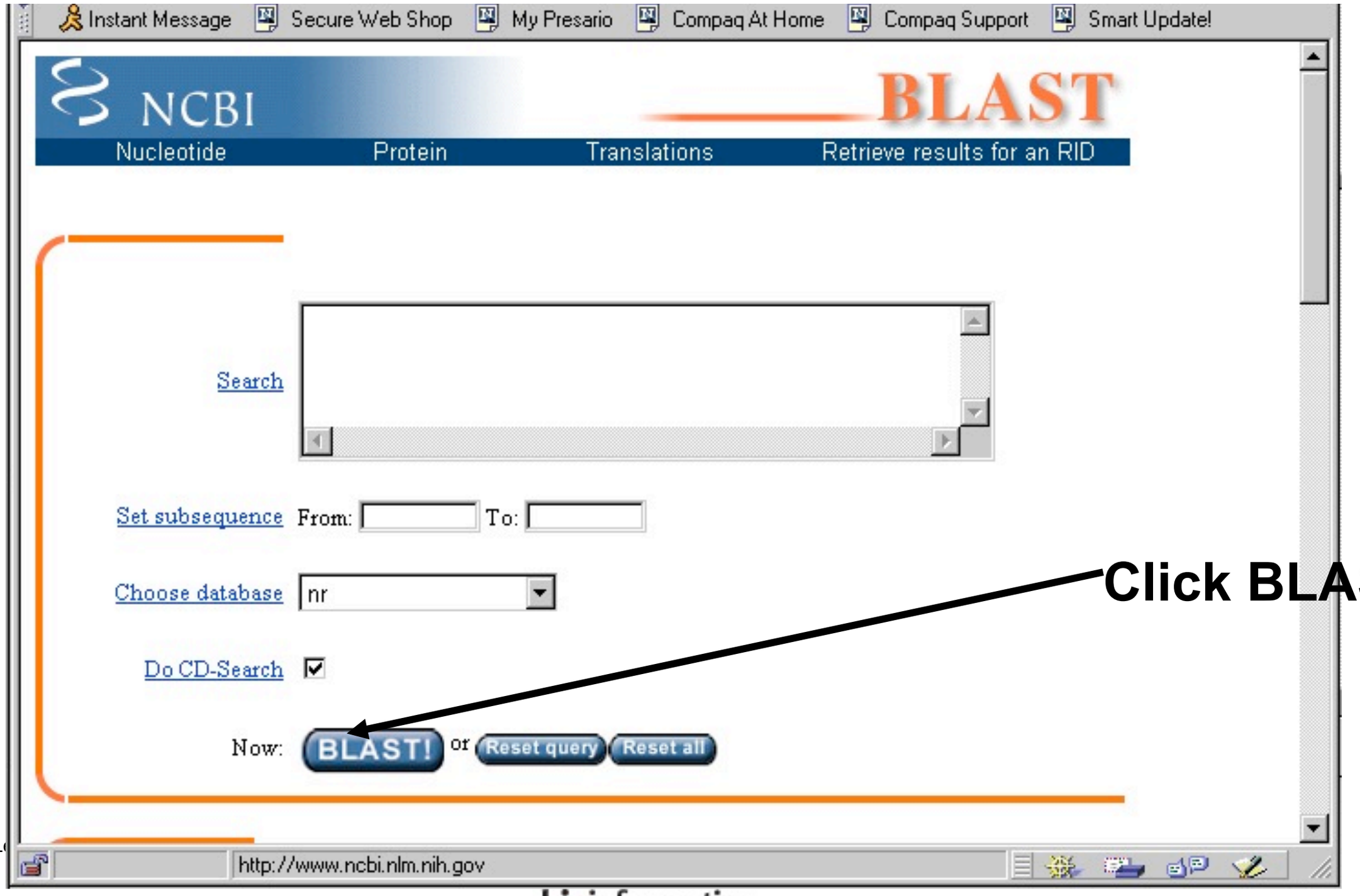
- CCGGCT**ATCA**TCTATTCCCGGTTCG
- **ATCG**TCTATTCCCGG
- The Yellow seqeunce alignmnet corresponds to the MSP wth score S


NCBI BLAST - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/BLAST/
Search
Mail AIM Home Radio Netscape Search Bookmarks Instant Message WebMail Radio People Yellow Pages Download Calendar
NCBI BLAST
NCBI
BLAST
PubMed Entrez BLAST OMIM Taxonomy Structure
Info
FAQs
News
References
Credits
Education
Program selection guide
Tutorial
URL API guide
Download
Executables
Databases
Source code
Support
Helpdesk
Mailing list
Nucleotide
Discontiguous megablast
Megablast
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
PHI- and PSI-BLAST
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Search by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat
Fugu rubripes, zebrafish
Insects, nematodes, plants, fungi, malaria
Microbial genomes, other eukaryotic genomes
Special
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
Meta
Retrieve results by RID
Get this page with javascript-free links
Disclaimer
Privacy statement
Accessibility
Valid XHTML 1.0, CSS.
Done
start
T-COFFEE server - N...
NCBI BLAST - Netscape
postdoc - Inbox for d...
Bioinfo3.1
10:11 AM

Different Flavours of BLAST

- **BLASTP** - protein query against protein DB
- **BLASTN** - DNA/RNA query against GenBank (DNA)
- **BLASTX** - 6 frame trans. DNA query against proteinDB
- **TBLASTN** - protein query against 6 frame GB transl.
- **TBLASTX** - 6 frame DNA query to 6 frame GB transl.
- **PSI-BLAST** - protein 'profile' query against protein DB
- **PHI-BLAST** - protein pattern against protein DB

Running NCBI BLAST



MT0895

- MMKIQIYGTGCANCQMLEKNAREAVKELGID AEFEKIKEMDQILEAGLTALPGLAVDGELKI MGRVASKEEIKKILS

Formatting Results

Lec

BLAST Format Options

Format

Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

Number of: Descriptions 100 Alignments 50

Alignment view: Pairwise

Format for PSI-BLAST ☐ with inclusion threshold: 0.005

Limit results by entrez query: or select from: All organisms

Expect value range:

Layout: Two Windows Formatting options on page with results: None

Autoformat: Semi-auto

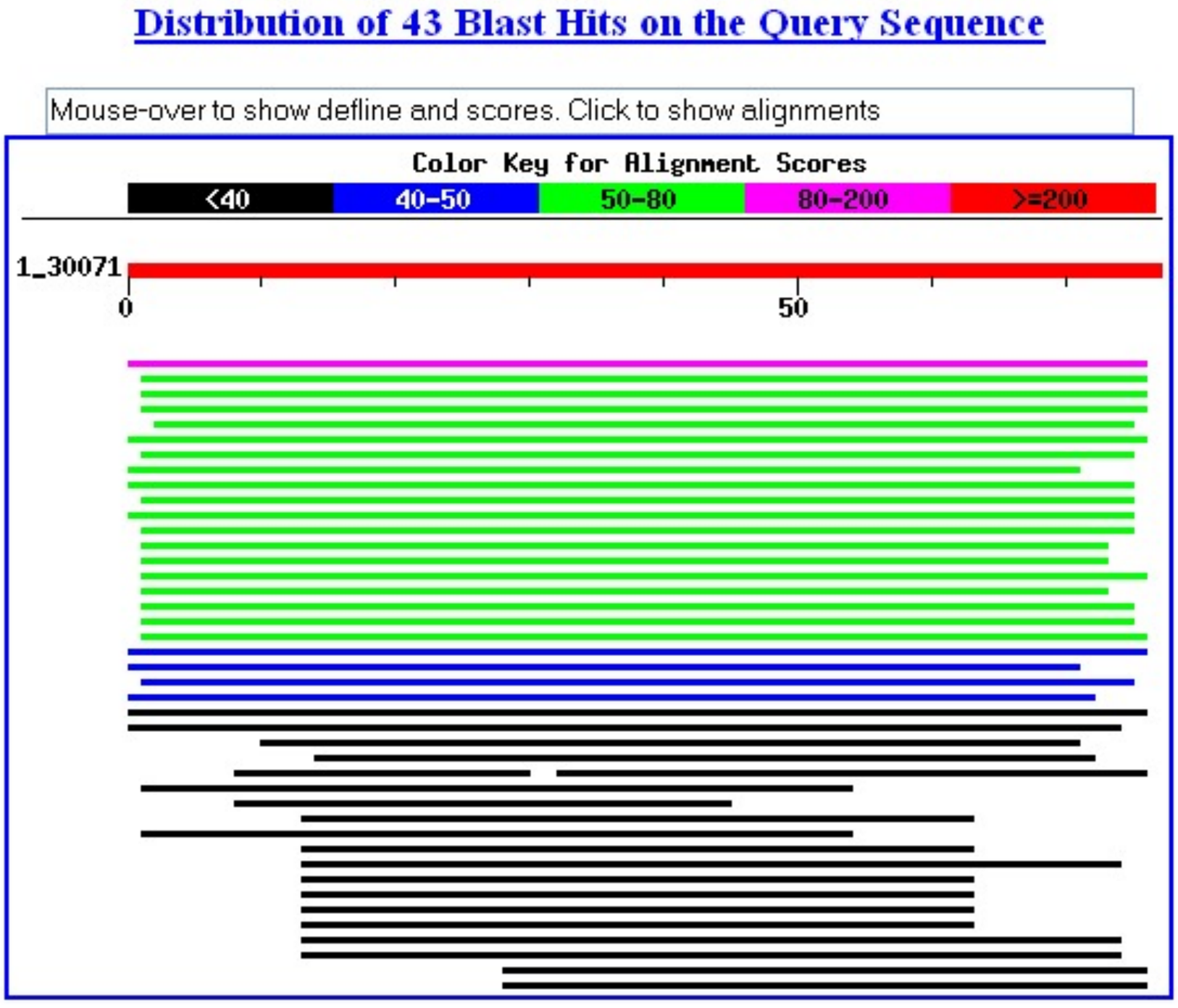
BLAST! or Reset all

Get the URL with preset values ? Get URL

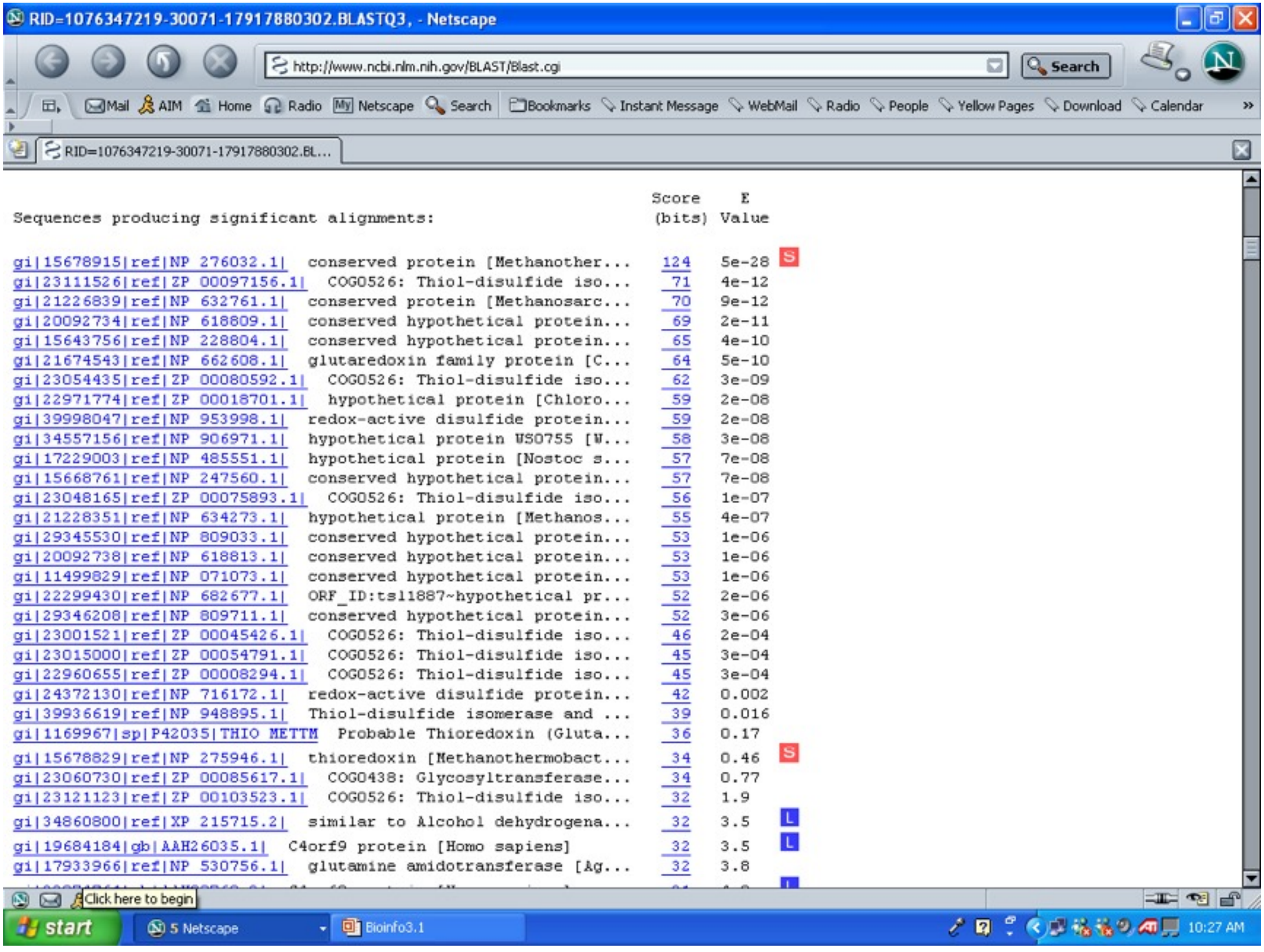
bioinformatics.ca

BLAST Output

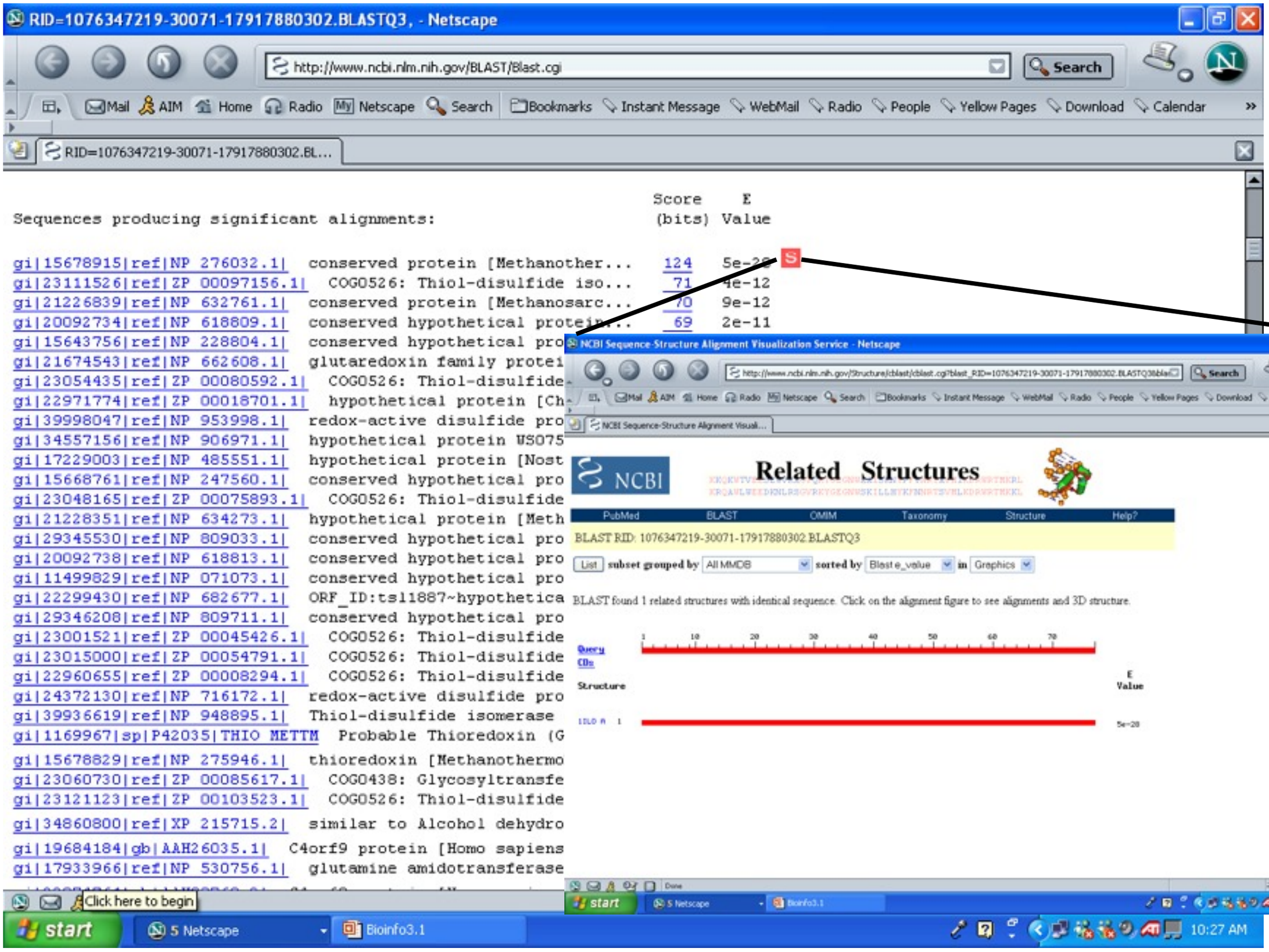
BLAST Output



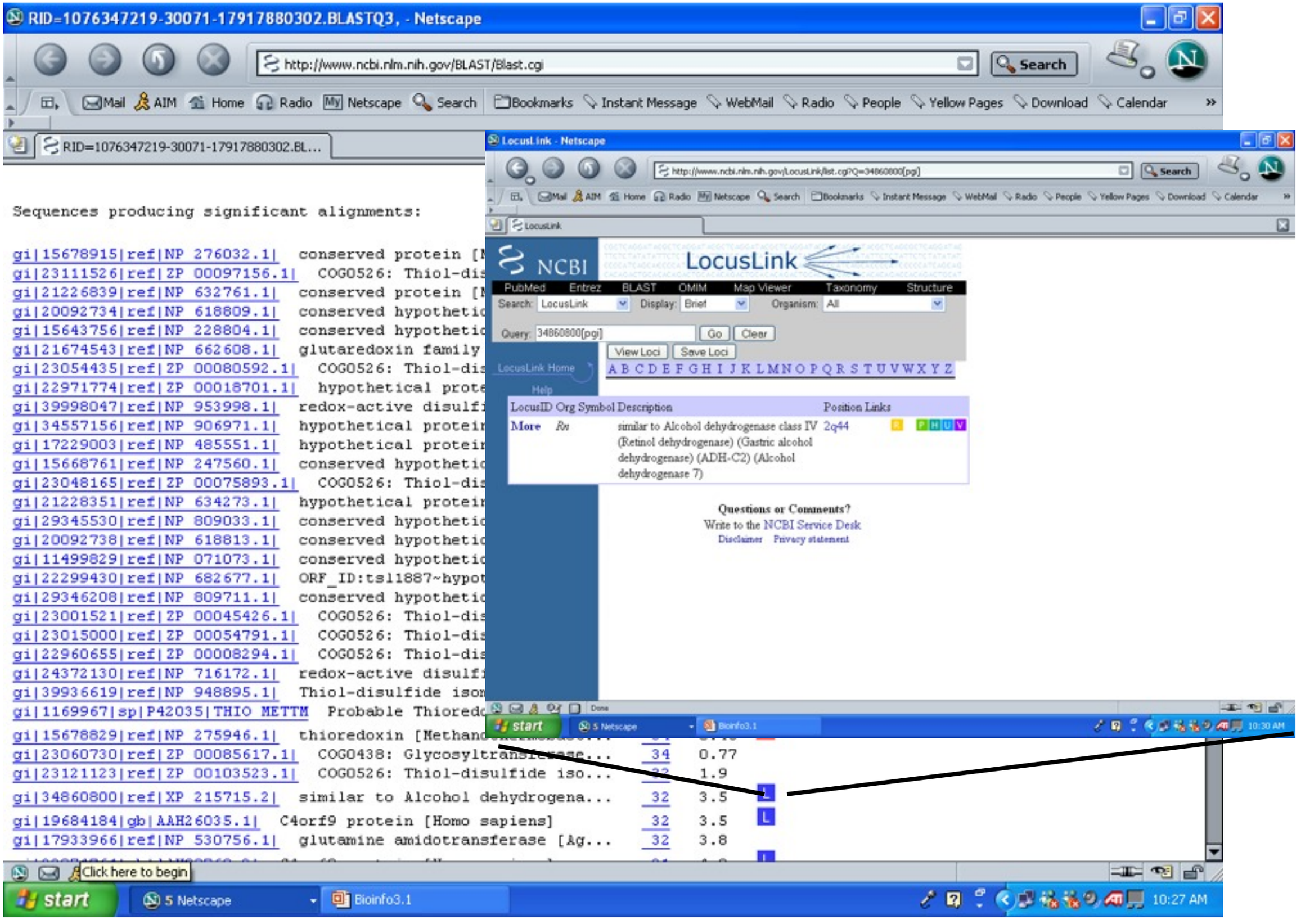
BLAST Output



BLAST Output



BLAST Output



BLAST Output

```
 gi|2621990|gb|AAB85393.1|    conserved protein [Methanothermobacter thermautotrophicus str.
          Delta H]
          Length = 77

 Score =  124 bits (310), Expect = 5e-28
 Identities = 77/77 (100%), Positives = 77/77 (100%)

Query: 1  MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL 60
          MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL
Sbjct: 1  MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL 60

Query: 61 KIMGRVASKEEIKKILS 77
          KIMGRVASKEEIKKILS
Sbjct: 61 KIMGRVASKEEIKKILS 77
```

```
>gi|23111526|ref|ZP_00097156.1|    COG0526: Thiol-disulfide isomerase and thioredoxins
          [Desulfitobacterium hafniense]
          Length = 76

 Score = 71.2 bits (173), Expect = 4e-12
 Identities = 40/76 (52%), Positives = 57/76 (75%)

Query: 2  MKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGELK 61
          M I+I GTGCANC+ LE NA+EA+KELG+DA  EK++++  I+  G+   P L V+ ++K
Sbjct: 1  MVIKILGTGCANCKKLEANAKEAIKELGLDAVVEKVEDLQAIMAYGVMKTPALVVNEQVK 60

Query: 62 IMGRVASKEEIKKILS 77
          +MG+V S EEIKK L+
Sbjct: 61 VMGKVLSAEEIKKYLN 76
```

bioinformatics.ca

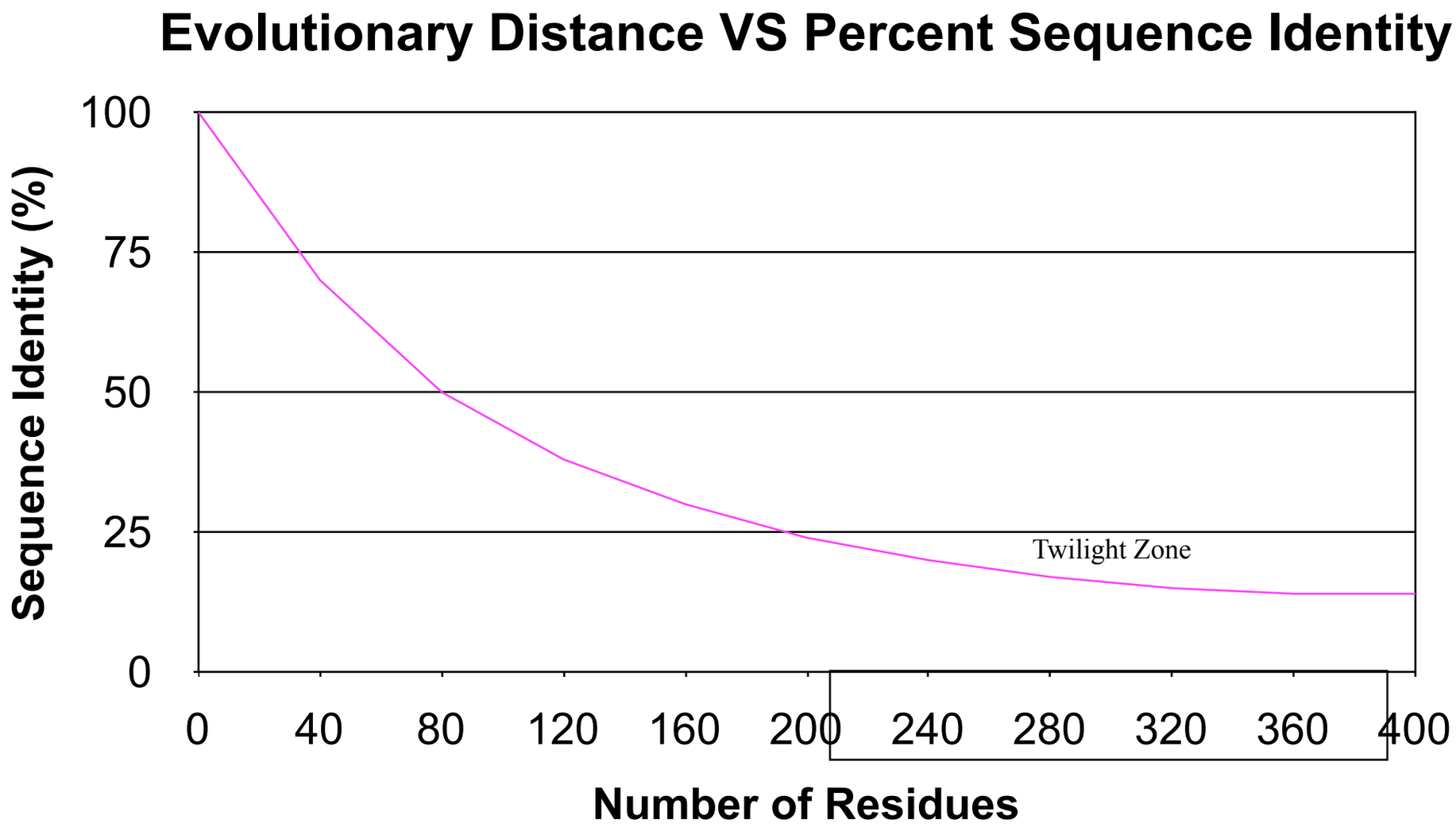
BLAST Parameters

- **Identities** - No. & % exact residue matches
- **Positives** - No. and % similar & ID matches
- **Gaps** - No. & % gaps introduced
- **Score** - Summed HSP score (S)
- **Bit Score** - a normalized score (S')
- **Expect (E)** - Expected # of chance HSP aligns
- **P** - Probability of getting a score > X
- **T** - Minimum word or k-tuple score **(Threshold)**

BLAST - Rules of Thumb

- Expect (E-value) is equal to the number of BLAST alignments with a given Score that are expected to be seen simply due to chance
- Don't trust a BLAST alignment with an Expect score > 0.01 (Grey zone is between 0.01 - 1)
- Expect and Score are related, but Expect contains more information. Note that %Identies is more useful than the bit Score
- Recall Doolittle's Curve (%ID vs. Length, next slide) %ID > 30 - numres/50
- If uncertain about a hit, perform a PSI-BLAST search

Doolittle's Curve



Statistics

- Probability of observing a score greater than S

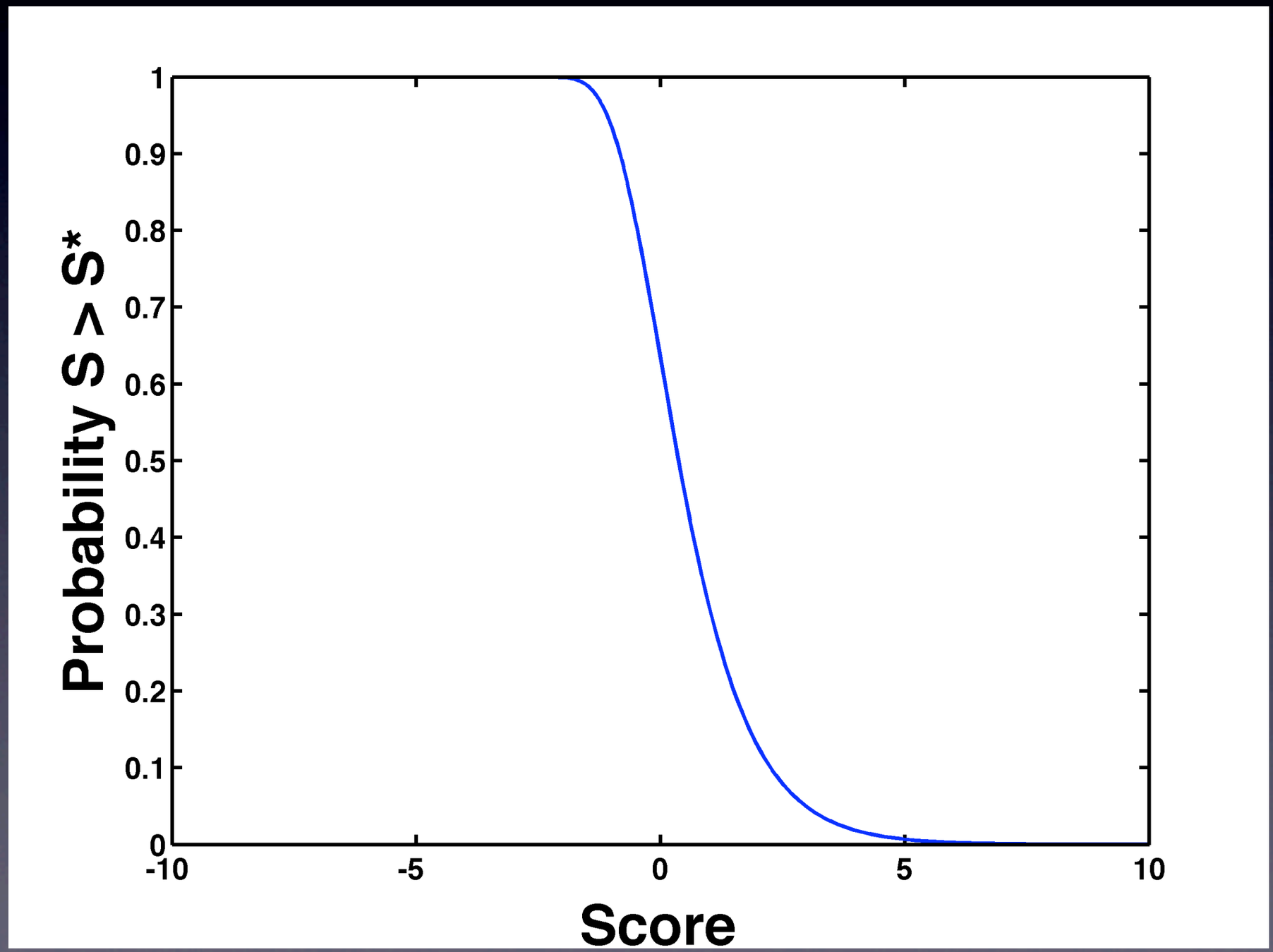
- Probability of observing multiple MSPs greater than S

$$P = 1 - e^{-y}$$

$$y = Kmne^{-\lambda S}$$

$$1 - e^{-y} \sum_{i=0}^{c-1} \frac{y^i}{i!}$$

Extreme Value Distribution



Scoring Matrices

- **BLOSUM Matrices**
  - Developed by Henikoff & Henikoff (1992)
  - BLOcks SUbstitution Matrix
  - Derived from the BLOCKS database
- **PAM Matrices**
  - Developed by Schwarz and Dayhoff (1978)
  - Point Accepted Mutation
  - Derived from manual alignments of closely related proteins

Conclusions

- **BLAST is the most important program in bioinformatics (maybe all of biology)**
- **BLAST is based on sound statistical principles (key to its speed and sensitivity)**
- **A basic understanding of its principles is key for using/interpreting BLAST output**
- **Use NBLAST or MEGABLAST for DNA**
- **Use PSI-BLAST for protein searches**