BLAST

- NCBI BLAST
- Basic Local Alignment Search Tool
- http://www.ncbi.nlm.nih.gov/BLAST/

Global versus local alignments

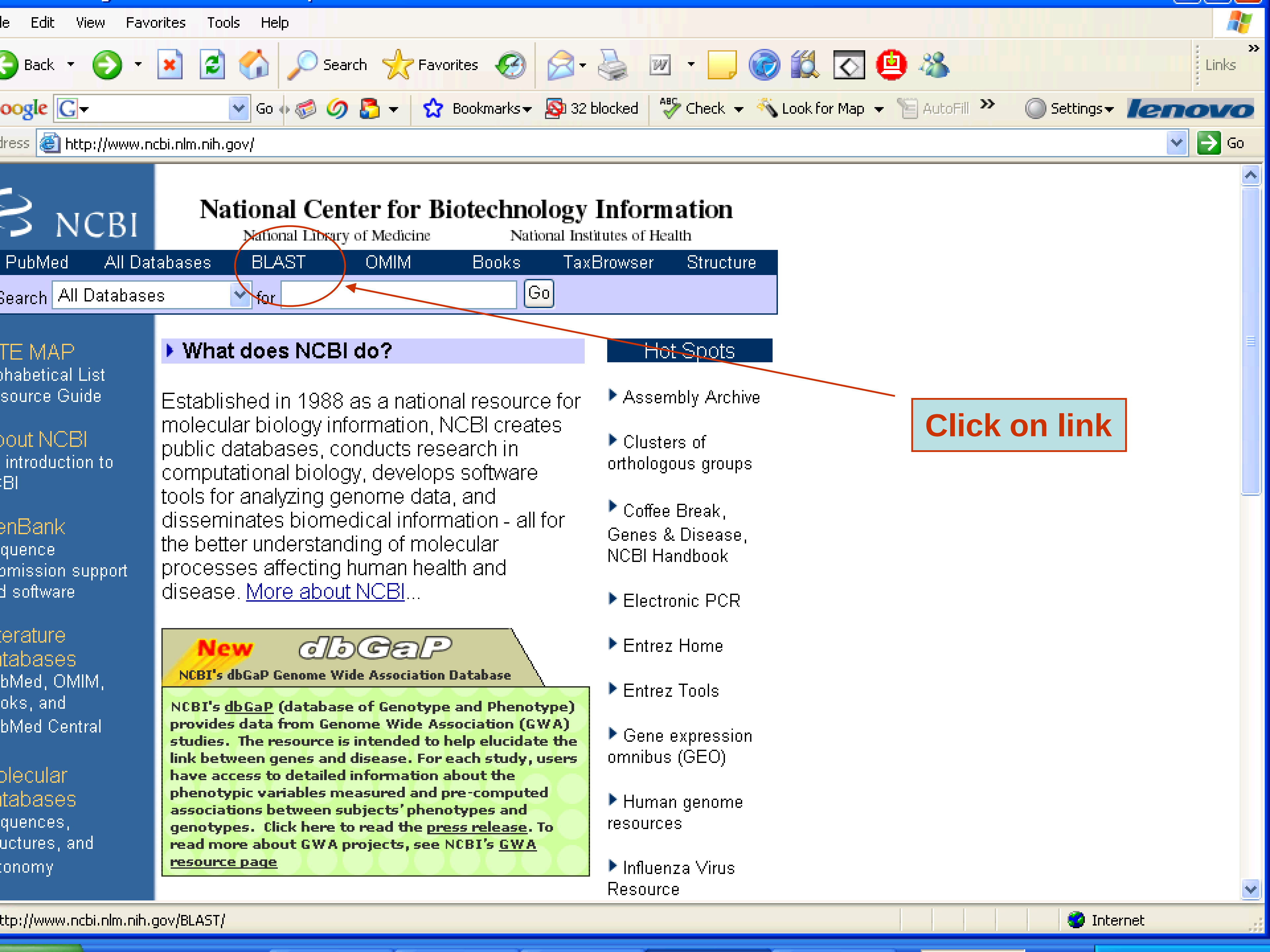
Global alignments:

- Attempt to align every residue in every sequence,
- Most useful when the sequences in the query set are similar and of roughly equal size.
- A general global alignment technique is called the Needleman-Wunsch algorithm

Local alignments:

- More useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context.
- The Smith-Waterman algorithm is a general local alignment method.

With sufficiently similar sequences, there is no difference between local and global alignments.



http://www.ncbi.nlm.nih.gov/

NCBI

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

TE MAP
phabetical List
source Guide

bout NCBI
introduction to
CBI

enBank
quence
bmission support
d software

terature
atabases
bMed, OMIM,
oks, and
bMed Central

olecular
atabases
quences,
uctures, and
xonomy

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More about NCBI...

New dbGaP

NCBI's dbGaP Genome Wide Association Database

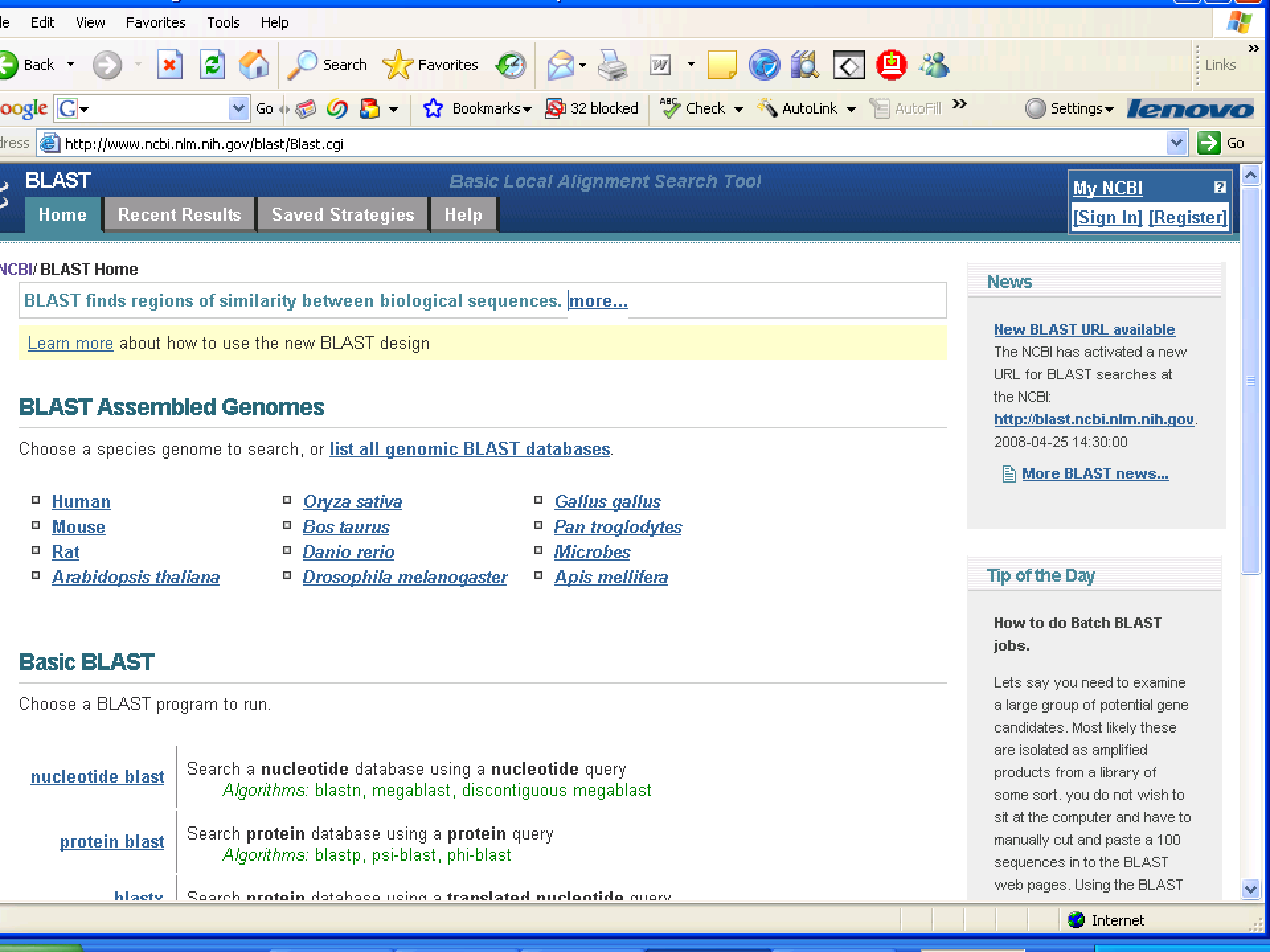
NCBI's dbGaP (database of Genotype and Phenotype) provides data from Genome Wide Association (GWA) studies. The resource is intended to help elucidate the link between genes and disease. For each study, users have access to detailed information about the phenotypic variables measured and pre-computed associations between subjects' phenotypes and genotypes. Click here to read the press release. To read more about GWA projects, see NCBI's GWA resource page

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- Influenza Virus Resource

ttp://www.ncbi.nlm.nih.gov/BLAST/

Click on link



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

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

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 *Algorithms:* blastn, megablast, discontiguous 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

News

New BLAST URL available

The NCBI has activated a new URL for BLAST searches at the NCBI: http://blast.ncbi.nlm.nih.gov.

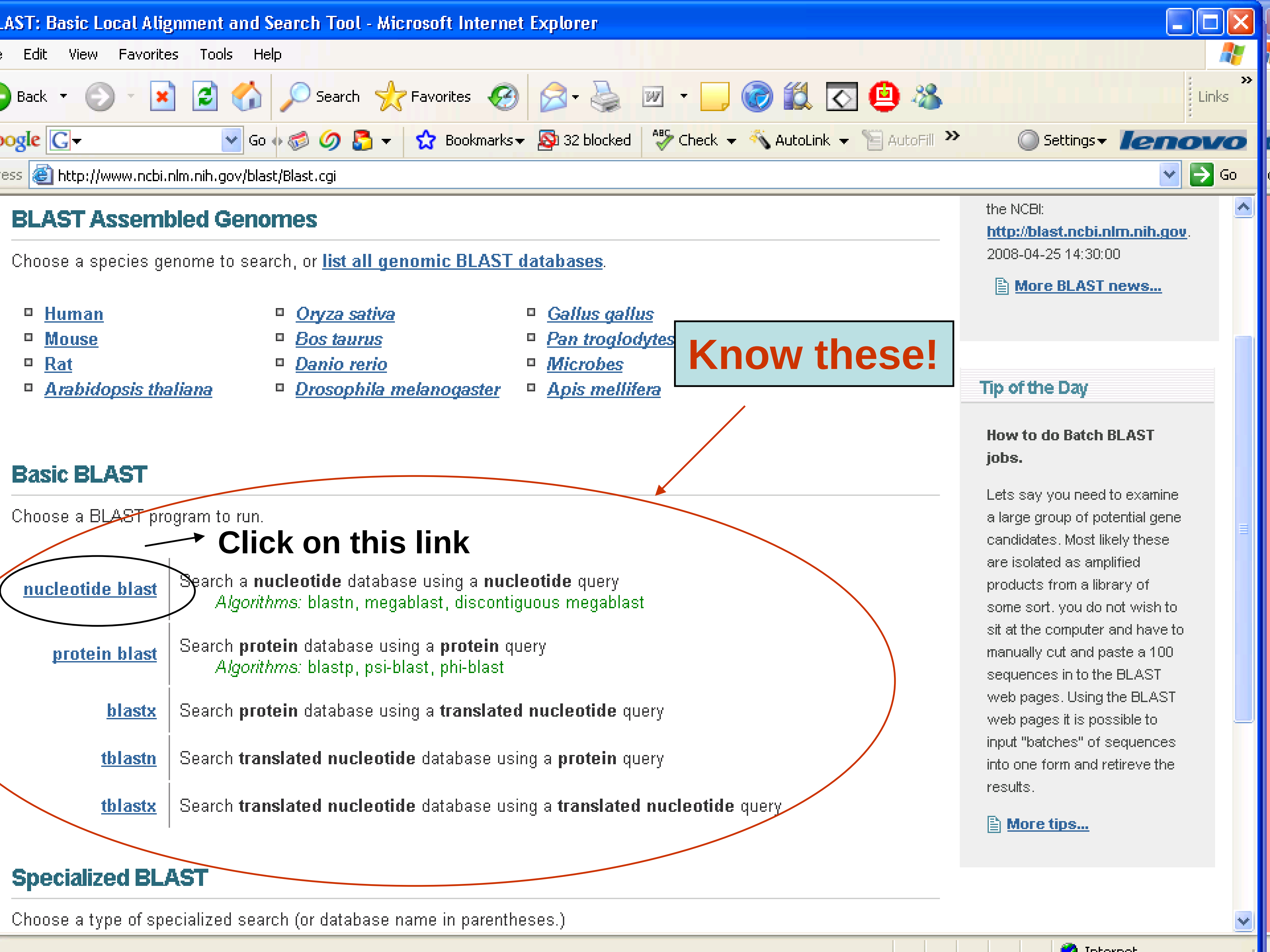
2008-04-25 14:30:00

More BLAST news...

Tip of the Day

How to do Batch BLAST jobs.

Lets say you need to examine a large group of potential gene candidates. Most likely these are isolated as amplified products from a library of some sort. you do not wish to sit at the computer and have to manually cut and paste a 100 sequences in to the BLAST web pages. Using the BLAST


LAST: Basic Local Alignment and Search Tool - Microsoft Internet Explorer
Edit View Favorites Tools Help
Back Search Favorites Links
Go Bookmarks 32 blocked Check AutoLink AutoFill Settings lenovo
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
Internet

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*

Know these!

Basic BLAST

Choose a BLAST program to run.

Click on this link

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

Specialized BLAST

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

the NCBI: http://blast.ncbi.nlm.nih.gov.
2008-04-25 14:30:00

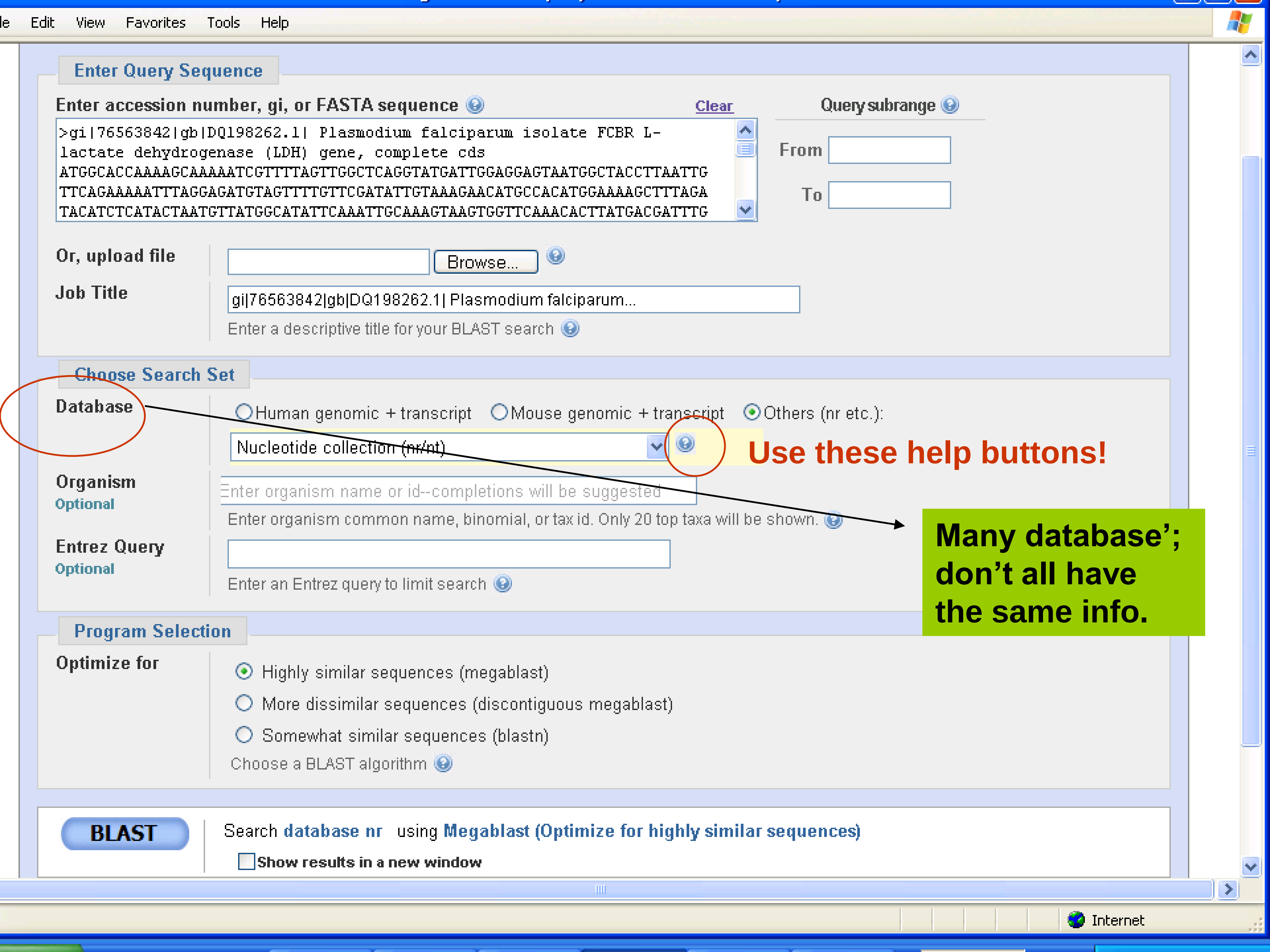
More BLAST news...

Tip of the Day

How to do Batch BLAST jobs.

Lets say you need to examine a large group of potential gene candidates. Most likely these are isolated as amplified products from a library of some sort. you do not wish to sit at the computer and have to manually cut and paste a 100 sequences in to the BLAST web pages. Using the BLAST web pages it is possible to input "batches" of sequences into one form and retireve the results.

More tips...



Enter Query Sequence

Enter accession number, gi, or FASTA sequence   Clear

```
>gi|76563842|gb|DQ198262.1| Plasmodium falciparum isolate FCBR L-
lactate dehydrogenase (LDH) gene, complete cds
ATGGCACCAAAAGCAAAAATCGTTTTAGTTGGCTCAGGTATGATTGGAGGAGTAATGGCTACCTTAATTG
TTCAGAAAAATTTAGGAGATGTAGTTTTGTTCGATATTGTAAAGAACATGCCACATGGAAAAGCTTTAGA
TACATCTCATACTAATGTTATGGCATATTCAAATTGCAAAGTAAGTGGTTCAAACACTTATGACGATTTG
```

Query subrange

From

To

Or, upload file   Browse...

Job Title: gi|76563842|gb|DQ198262.1| Plasmodium falciparum...

Enter a descriptive title for your BLAST search

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

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

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST   Search database nr using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

Use these help buttons!

Many database'; don't all have the same info.

Databases available for BLAST search

The BLAST pages offer several different databases for searching. Some of these, like SwissProt and PDB are complied outside of NCBI. Other like ecoli, dbEST and month, are subsets of the NCBI databases. Other "virtual Databases" can be created using the "Limit by Entrez Query" option.

Peptide Sequence Databases

Nr: All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF

Refseq: RefSeq protein sequences from NCBI's Reference Sequence Project.

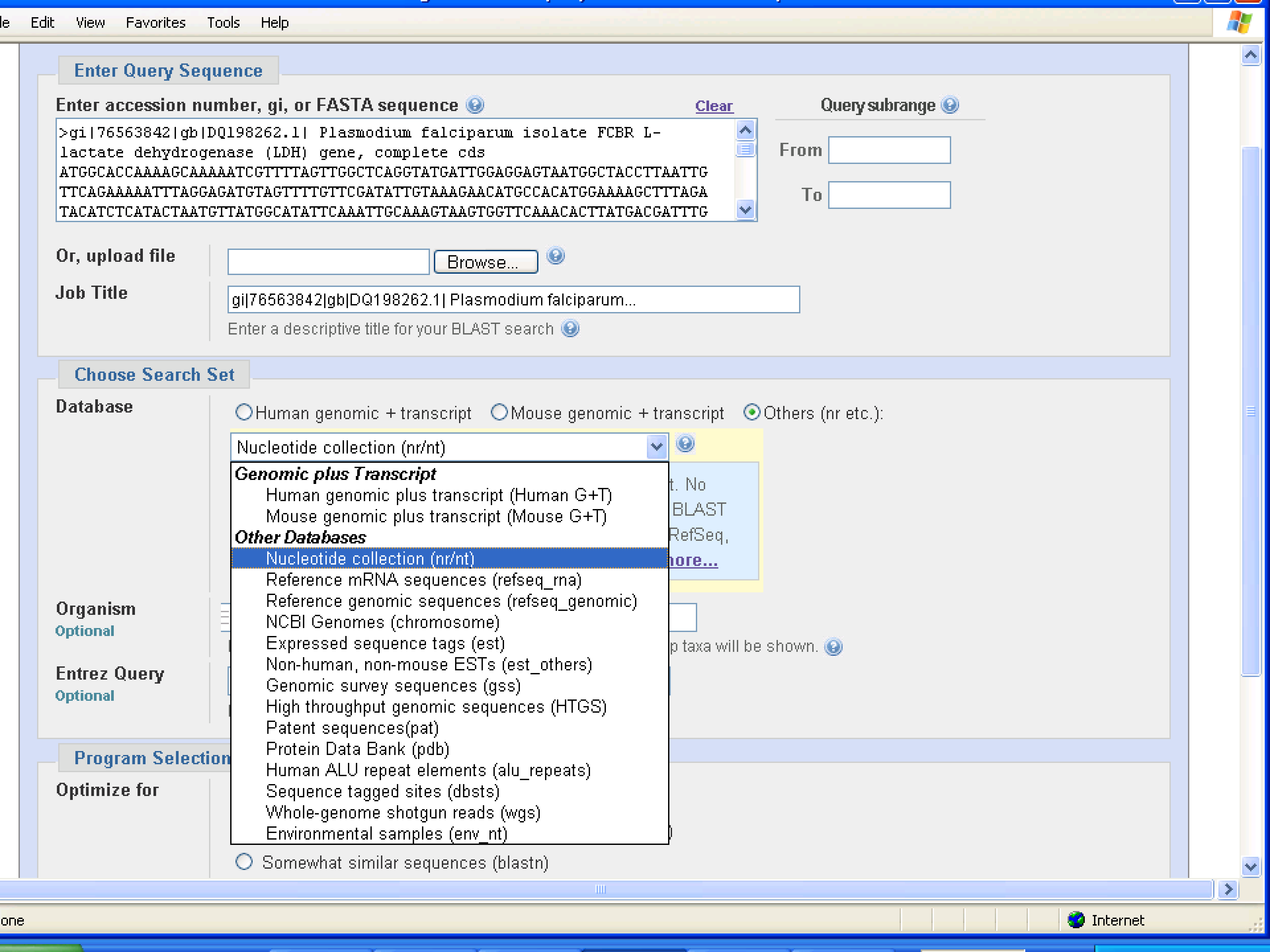
Swissprot: Last major release of the SWISS-PROT protein sequence database (no updates).

Pat: Proteins from the Patent division of GenPept.

pdb : Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank.

Month: All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.

env_nr: Protein sequences from environmental samples.



Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>gi|76563842|gb|DQ198262.1| Plasmodium falciparum isolate FCBR L-
lactate dehydrogenase (LDH) gene, complete cds
ATGGCACCAAAAGCAAAAATCGTTTTAGTTGGCTCAGGTATGATTGGAGGAGTAATGGCTACCTTAATTG
TTCAGAAAAATTTAGGAGATGTAGTTTTGTTCGATATTGTAAAGAACATGCCACATGGAAAAGCTTTAGA
TACATCTCATACTAATGTTATGGCATATTCAAATTGCAAAGTAAGTGGTTCAAACACTTATGACGATTTG
```

Query subrange

From

To

Or, upload file

Job Title: gi|76563842|gb|DQ198262.1| Plasmodium falciparum...

Enter a descriptive title for your BLAST search

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Genomic plus Transcript
- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases
- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun reads (wgs)
- Environmental samples (env_nt)

t. No
BLAST
RefSeq,
nore...

Organism
Optional

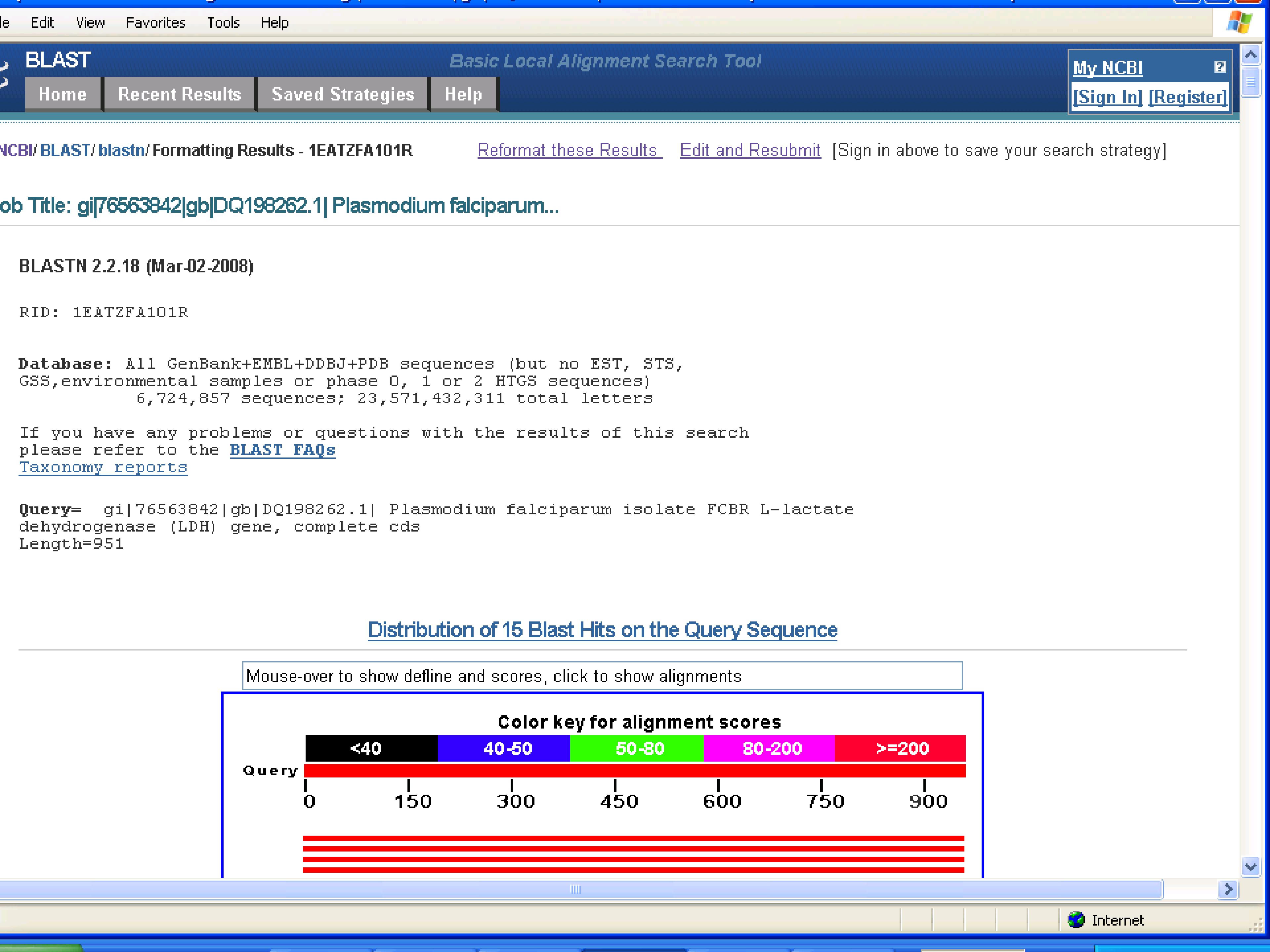
p taxa will be shown.

Entrez Query
Optional

Program Selection

Optimize for

○ Somewhat similar sequences (blastn)



NCBI/ BLAST/ blastn/ Formatting Results - 1EATZFA101R

[Sign in above to save your search strategy]

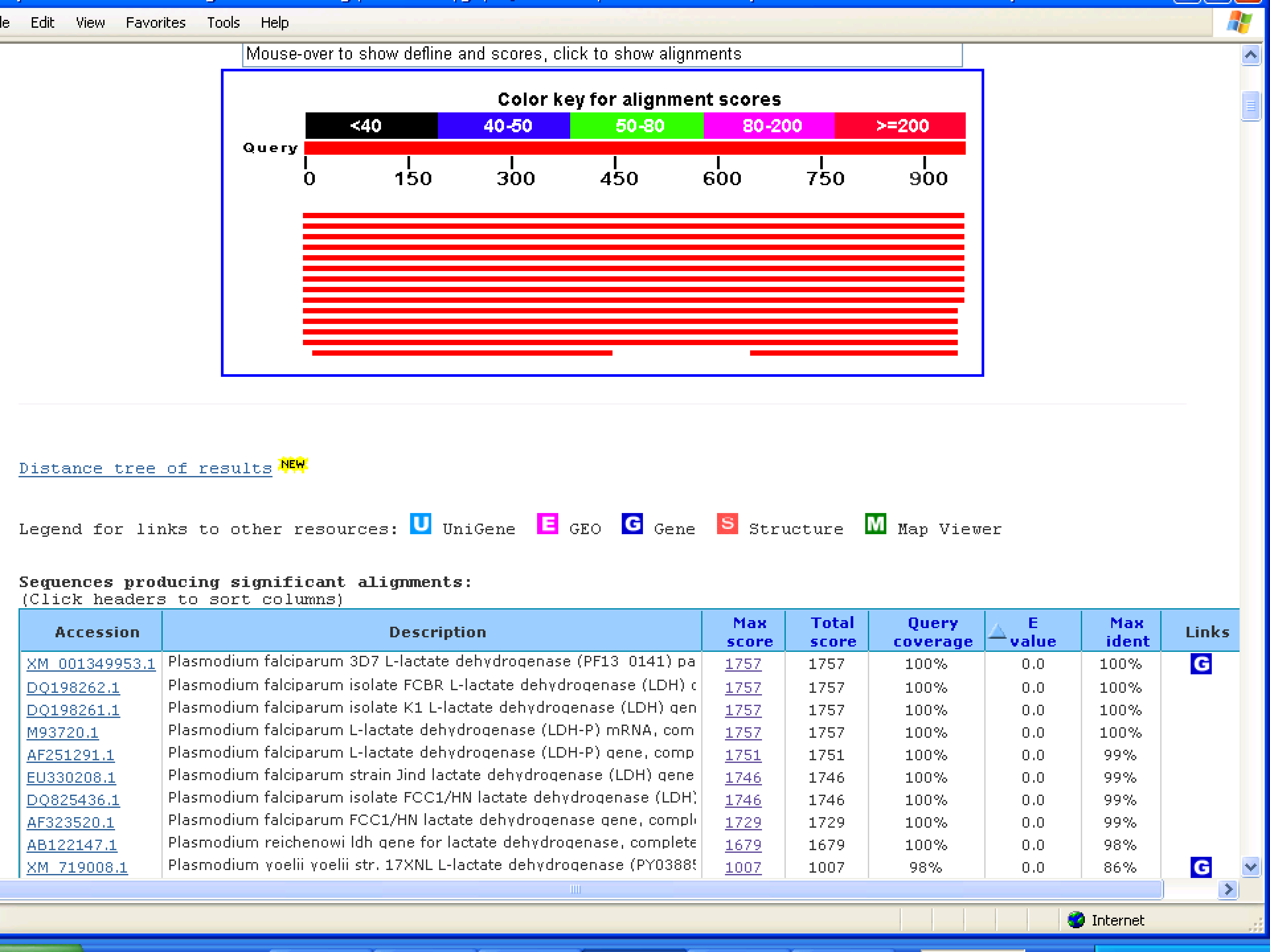
ob Title: gi|76563842|gb|DQ198262.1| Plasmodium falciparum...

BLASTN 2.2.18 (Mar-02-2008)

```
RID: 1EATZFA101R


Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           6,724,857 sequences; 23,571,432,311 total letters

If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports

Query=  gi|76563842|gb|DQ198262.1| Plasmodium falciparum isolate FCBR L-lactate
dehydrogenase (LDH) gene, complete cds
Length=951
```

Distribution of 15 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

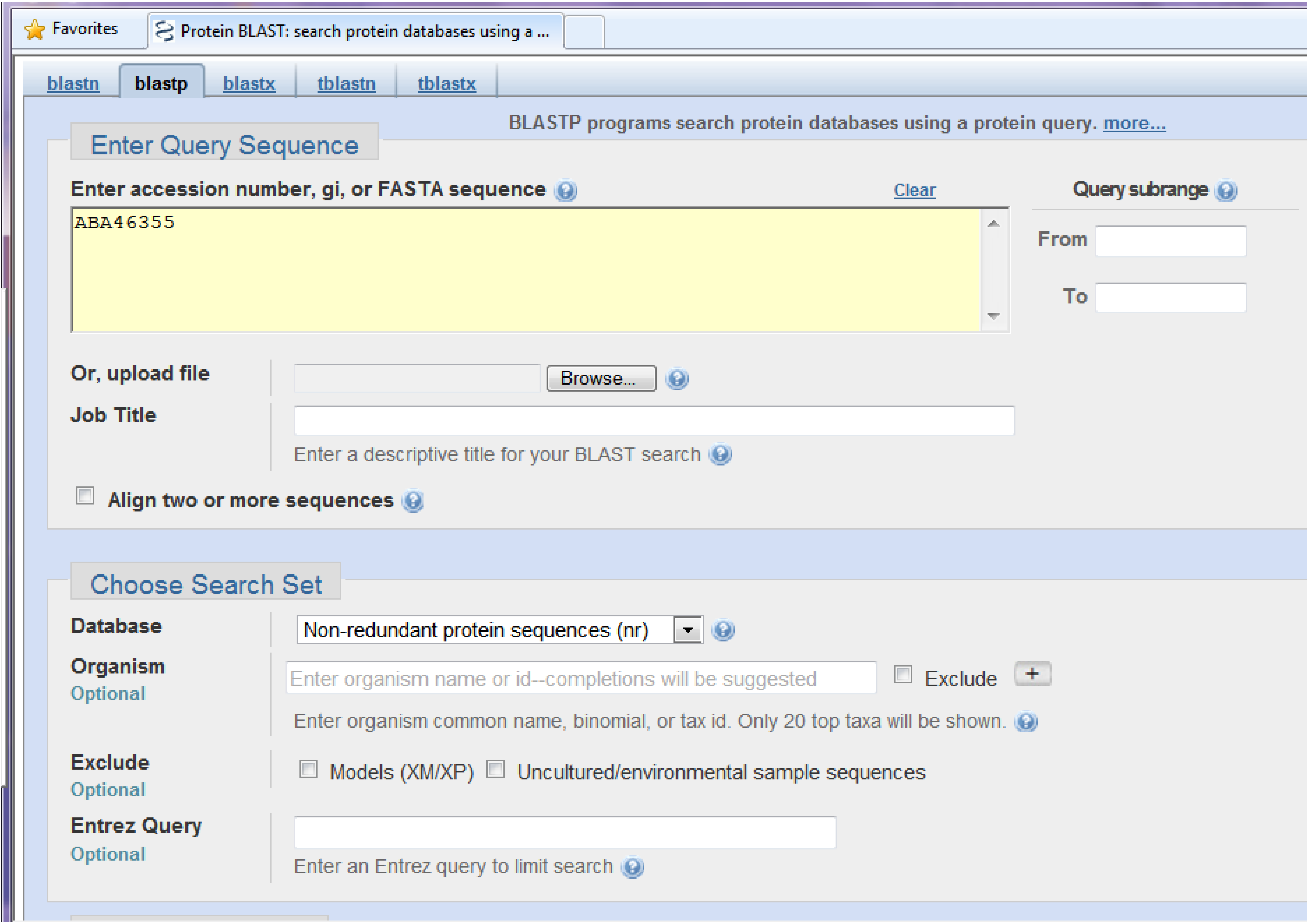

le
Edit
View
Favorites
Tools
Help
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
150
300
450
600
750
900
Internet

Distance tree of results NEW

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
XM_001349953.1	Plasmodium falciparum 3D7 L-lactate dehydrogenase (PF13_0141) pa	1757	1757	100%	0.0	100%	G
DQ198262.1	Plasmodium falciparum isolate FCBR L-lactate dehydrogenase (LDH) c	1757	1757	100%	0.0	100%	
DQ198261.1	Plasmodium falciparum isolate K1 L-lactate dehydrogenase (LDH) gen	1757	1757	100%	0.0	100%	
M93720.1	Plasmodium falciparum L-lactate dehydrogenase (LDH-P) mRNA, com	1757	1757	100%	0.0	100%	
AF251291.1	Plasmodium falciparum L-lactate dehydrogenase (LDH-P) gene, comp	1751	1751	100%	0.0	99%	
EU330208.1	Plasmodium falciparum strain Jind lactate dehydrogenase (LDH) gene	1746	1746	100%	0.0	99%	
DQ825436.1	Plasmodium falciparum isolate FCC1/HN lactate dehydrogenase (LDH)	1746	1746	100%	0.0	99%	
AF323520.1	Plasmodium falciparum FCC1/HN lactate dehydrogenase gene, compl	1729	1729	100%	0.0	99%	
AB122147.1	Plasmodium reichenowi ldh gene for lactate dehydrogenase, complete	1679	1679	100%	0.0	98%	
XM_719008.1	Plasmodium yoelii yoelii str. 17XNL L-lactate dehydrogenase (PY0388	1007	1007	98%	0.0	86%	G



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence: ABA46355

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude

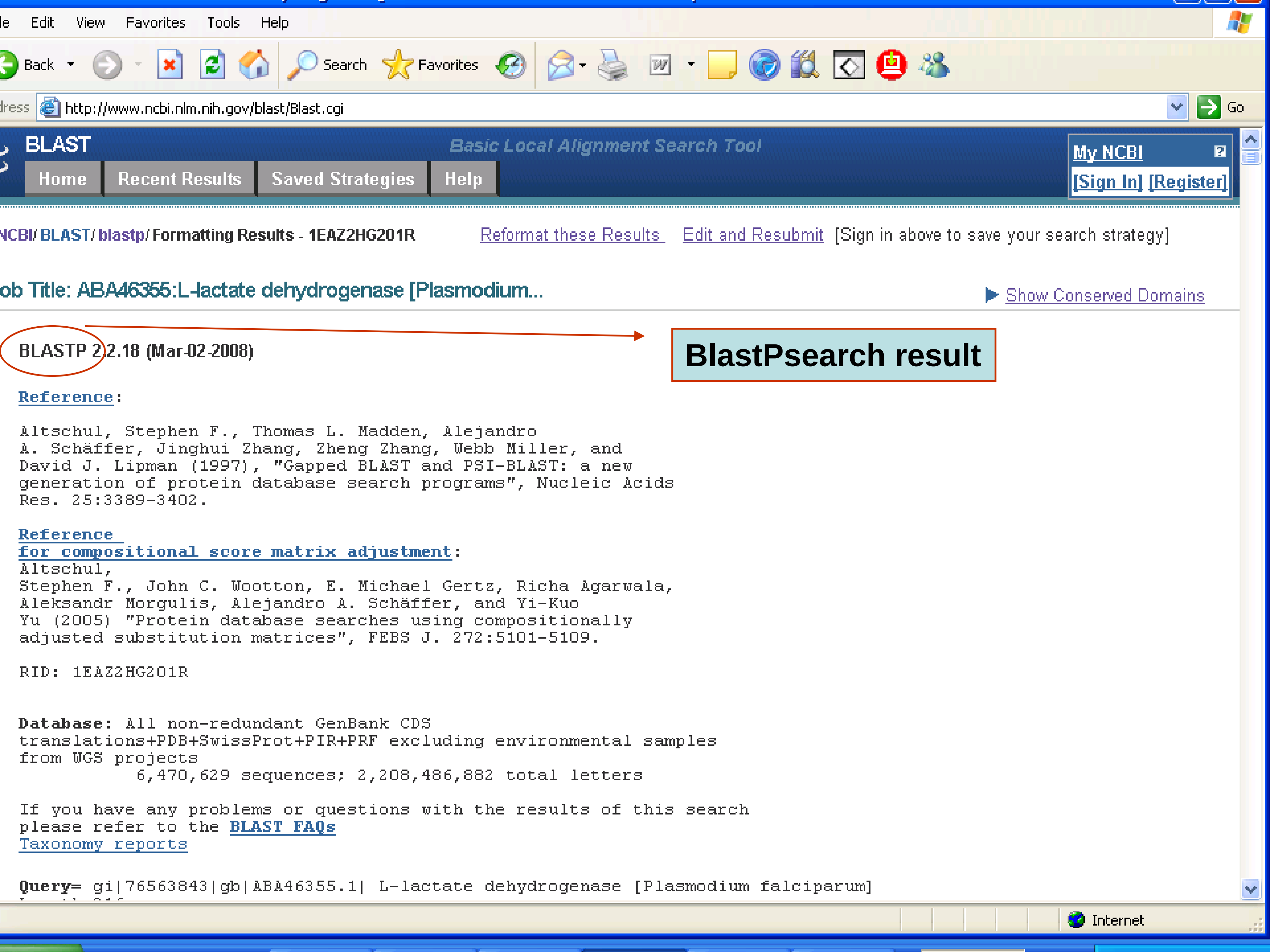
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search



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

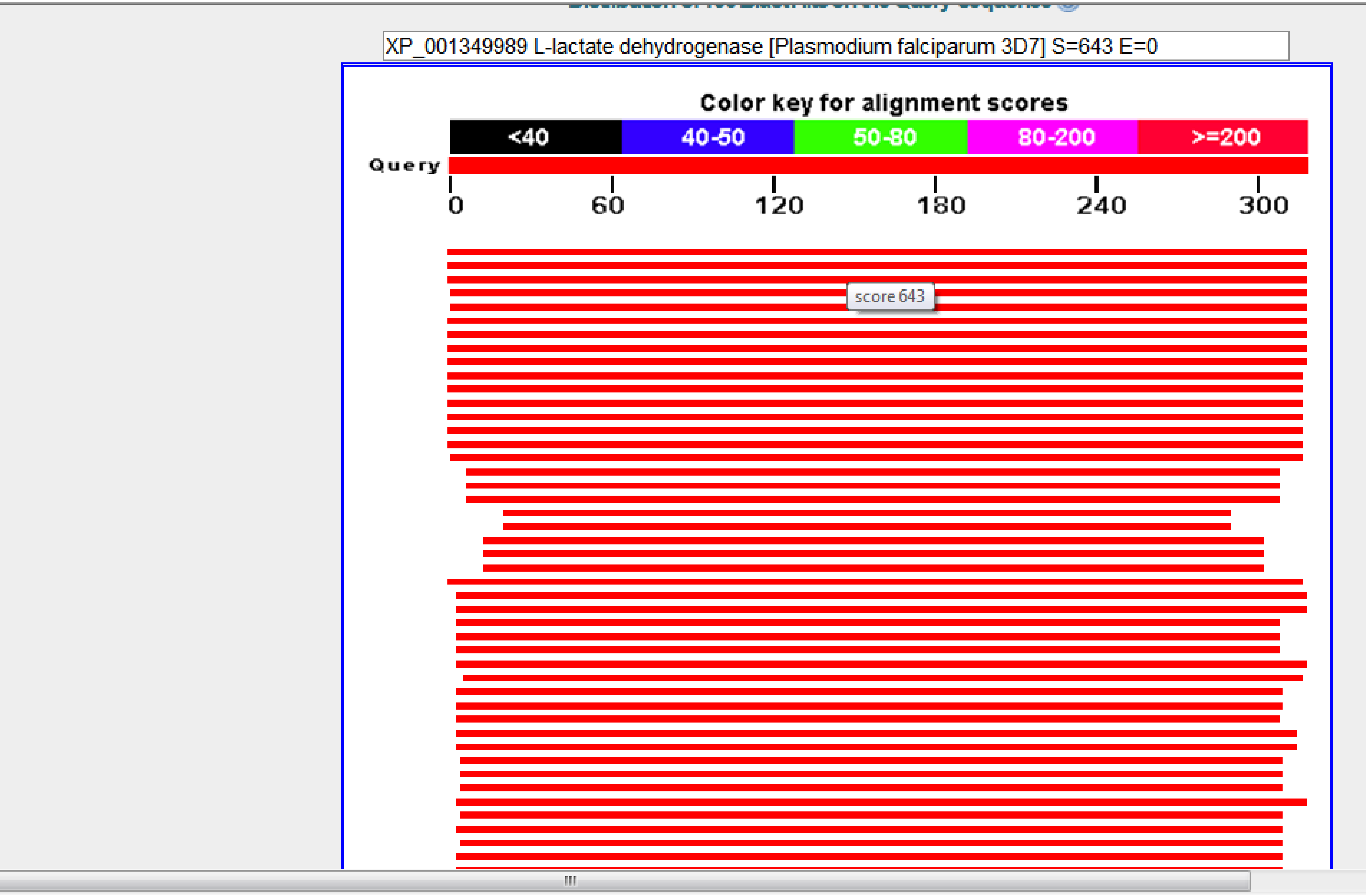
NCBI/ BLAST/ blastp/ Formatting Results - 1EAZ2HG201R   Reformat these Results   Edit and Resubmit [Sign in above to save your search strategy]

ob Title: ABA46355:L-lactate dehydrogenase [Plasmodium...   Show Conserved Domains

BLASTP 2.2.18 (Mar-02-2008)

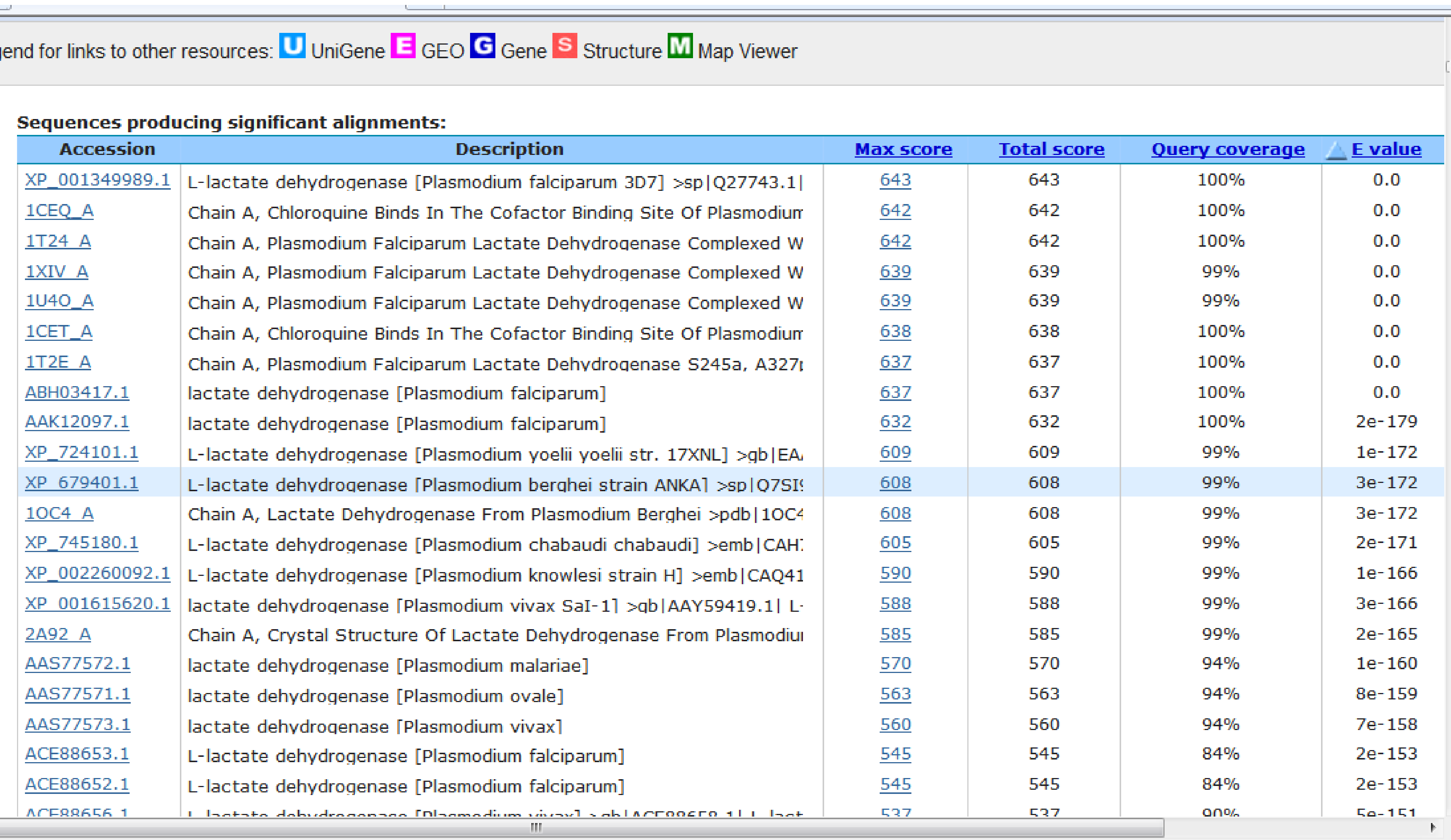
BlastPsearch result

```
Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro
A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic Acids
Res. 25:3389-3402.

Reference
for compositional score matrix adjustment:
Altschul,
Stephen F., John C. Wootton, E. Michael Gertz, Richa Agarwala,
Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo
Yu (2005) "Protein database searches using compositionally
adjusted substitution matrices", FEBS J. 272:5101-5109.

RID: 1EAZ2HG201R


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           6,470,629 sequences; 2,208,486,882 total letters

If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports


Query= gi|76563843|gb|ABA46355.1| L-lactate dehydrogenase [Plasmodium falciparum]
```



Distribution of 100 Blast Hits on the Query Sequence

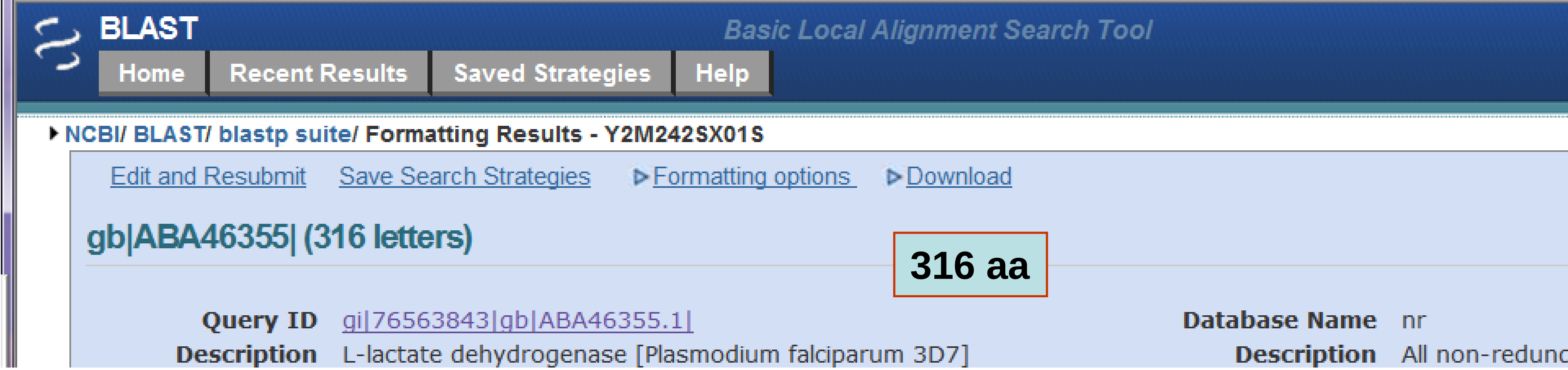
XP_001349989 L-lactate dehydrogenase [Plasmodium falciparum 3D7] S=643 E=0



gend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
XP_001349989.1	L-lactate dehydrogenase [Plasmodium falciparum 3D7] >sp|Q27743.1|	643	643	100%	0.0
1CEQ_A	Chain A, Chloroquine Binds In The Cofactor Binding Site Of Plasmodiur	642	642	100%	0.0
1T24_A	Chain A, Plasmodium Falciparum Lactate Dehydrogenase Complexed W	642	642	100%	0.0
1XIV_A	Chain A, Plasmodium Falciparum Lactate Dehydrogenase Complexed W	639	639	99%	0.0
1U4O_A	Chain A, Plasmodium Falciparum Lactate Dehydrogenase Complexed W	639	639	99%	0.0
1CET_A	Chain A, Chloroquine Binds In The Cofactor Binding Site Of Plasmodiur	638	638	100%	0.0
1T2E_A	Chain A, Plasmodium Falciparum Lactate Dehydrogenase S245a, A327	637	637	100%	0.0
ABH03417.1	lactate dehydrogenase [Plasmodium falciparum]	637	637	100%	0.0
AAK12097.1	lactate dehydrogenase [Plasmodium falciparum]	632	632	100%	2e-179
XP_724101.1	L-lactate dehydrogenase [Plasmodium yoelii yoelii str. 17XNL] >gb|EA	609	609	99%	1e-172
XP_679401.1	L-lactate dehydrogenase [Plasmodium berghei strain ANKA] >sp|Q7SI	608	608	99%	3e-172
1OC4_A	Chain A, Lactate Dehydrogenase From Plasmodium Berghei >pdb|1OC4	608	608	99%	3e-172
XP_745180.1	L-lactate dehydrogenase [Plasmodium chabaudi chabaudi] >emb|CAH	605	605	99%	2e-171
XP_002260092.1	L-lactate dehydrogenase [Plasmodium knowlesi strain H] >emb|CAQ41	590	590	99%	1e-166
XP_001615620.1	lactate dehydrogenase [Plasmodium vivax SaI-1] >gb|AAY59419.1| L-	588	588	99%	3e-166
2A92_A	Chain A, Crystal Structure Of Lactate Dehydrogenase From Plasmodiu	585	585	99%	2e-165
AAS77572.1	lactate dehydrogenase [Plasmodium malariae]	570	570	94%	1e-160
AAS77571.1	lactate dehydrogenase [Plasmodium ovale]	563	563	94%	8e-159
AAS77573.1	lactate dehydrogenase [Plasmodium vivax]	560	560	94%	7e-158
ACE88653.1	L-lactate dehydrogenase [Plasmodium falciparum]	545	545	84%	2e-153
ACE88652.1	L-lactate dehydrogenase [Plasmodium falciparum]	545	545	84%	2e-153
ACE88656.1	L-lactate dehydrogenase [Plasmodium vivax] >gb|ACE88658.1| L-lact	537	537	90%	5e-151

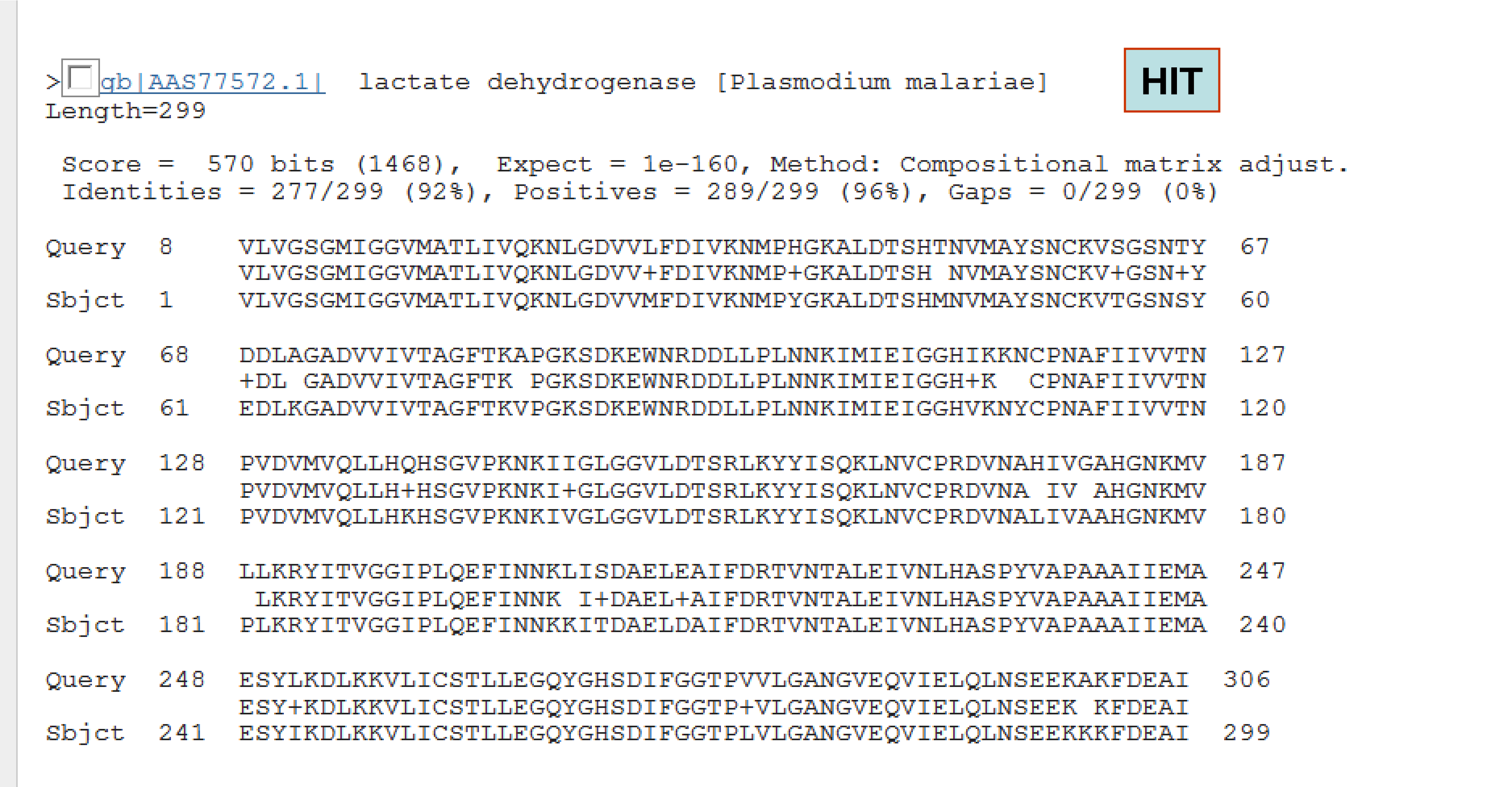


NCBI/ BLAST/ blastp suite/ Formatting Results - Y2M242SX01S

gb|ABA46355| (316 letters)

316 aa

Query ID	gi|76563843|gb|ABA46355.1|	**Database Name**	nr
Description	L-lactate dehydrogenase [Plasmodium falciparum 3D7]	**Description**	All non-redund



HIT

```
>☐gb|AAS77572.1|  lactate dehydrogenase [Plasmodium malariae]
Length=299

 Score =  570 bits (1468),  Expect = 1e-160, Method: Compositional matrix adjust.
 Identities = 277/299 (92%), Positives = 289/299 (96%), Gaps = 0/299 (0%)

Query  8    VLVGSGMIGGVMATLIVQKNLGDVVLFDIVKNMPHGKALDTSHTNVMAYSNCKVSGSNTY  67
            VLVGSGMIGGVMATLIVQKNLGDVV+FDIVKNMP+GKALDTSH NVMAYSNCKV+GSN+Y
Sbjct  1    VLVGSGMIGGVMATLIVQKNLGDVVMFDIVKNMPYGKALDTSHMNVMAYSNCKVTGSNSY  60

Query  68   DDLAGADVVIVTAGFTKAPGKSDKEWNRDDLLPLNNKIMIEIGGHIKKNCPNAFIIVVTN  127
            +DL GADVVIVTAGFTK PGKSDKEWNRDDLLPLNNKIMIEIGGH+K  CPNAFIIVVTN
Sbjct  61   EDLKGADVVIVTAGFTKVPGKSDKEWNRDDLLPLNNKIMIEIGGHVKNYCPNAFIIVVTN  120

Query  128  PVDVMVQLLHQHSGVPKNKIIGLGGVLDTSRLKYYISQKLNVCPRDVNAHIVGAHGNKMV  187
            PVDVMVQLLH+HSGVPKNKI+GLGGVLDTSRLKYYISQKLNVCPRDVNA IV AHGNKMV
Sbjct  121  PVDVMVQLLHKHSGVPKNKIVGLGGVLDTSRLKYYISQKLNVCPRDVNALIVAAHGNKMV  180

Query  188  LLKRYITVGGIPLQEFINNKLISDAELEAIFDRTVNTALEIVNLHASPYVAPAAAIIEMA  247
             LKRYITVGGIPLQEFINNK I+DAEL+AIFDRTVNTALEIVNLHASPYVAPAAAIIEMA
Sbjct  181  PLKRYITVGGIPLQEFINNKKITDAELDAIFDRTVNTALEIVNLHASPYVAPAAAIIEMA  240

Query  248  ESYLKDLKKVLICSTLLEGQYGHSDIFGGTPVVLGANGVEQVIELQLNSEEKAKFDEAI  306
            ESY+KDLKKVLICSTLLEGQYGHSDIFGGTP+VLGANGVEQVIELQLNSEEK KFDEAI
Sbjct  241  ESYIKDLKKVLICSTLLEGQYGHSDIFGGTPLVLGANGVEQVIELQLNSEEKKKFDEAI  299
```

On the BLAST result you find

1. References
2. Database
3. Query
   - the term that you asked.
4. A graphic display (coloured map) of the result
   - i.e. 15 BLASt hits on the query sequence.
   - Passing the mouse bar over the colour lists the sequence.
5. A hit list
   - showing the name of sequences similar to your query, ranked by similarity.

Hit list contains sequences producing significant alignments

1. Includes the **Accession number** e.g. DQ198262
   - The hyperlink takes you to the database entry containing the sequence.
2. Followed by a description of the sequence that was picked up
   - check carefully before getting too excited !
3. **SCORE** in bits.
   - A measure of the statistical significance of the alignment. The better the score the better the alignment. Matches BELOW 50 are unreliable.
4. **E-Value**. Expected value.
   - Measures the number of times you could have expected such a good match purely by chance.
   - A sequence with a value close to 0 i.e. 0.0000000000001 is a nearly identical sequence.
   - One is realistically looking for E-values smaller than 0.0001 or 10 -4.

5. % identity.
   - DNA: MORE than 75% identity is GOOD
   - PROTEINS: MORE than 25% identity is GOOD
   - Positives – fraction of residues that are similar (conserved).
   - Gaps – introduced to compensate for deletions/insertions
   - Space – no alignment
6. Length.
   - How LONG are the two segments that have been aligned. A short sequence is not very significant.
7. **ONE** sequence is **YOUR** sequence
   - i.e. query sequence. The other is the hit sequence**.**

SUMMARY.

A good alignment should contain not too many gaps and have a few sections of high similarity rather than one or two residues here and there.

E-value: Expect value

- Parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.
- Describes the random background noise that exists for matches between sequences.
- For example, an E-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size, one might expect to see one match with a similar score simply by chance.
- The lower the E-value, or the closer it is to "0", the higher is the "significance" of the match.

However,

- It is important to note that searches with short sequences can be virtually identical and have relatively high E-value.
- This is because the calculation of the E-value also takes into account the length of the query sequence. This is because shorter sequences have a high probability of occurring in the database purely by chance.
- For more information, see the following tutorial.
  - http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.section.622