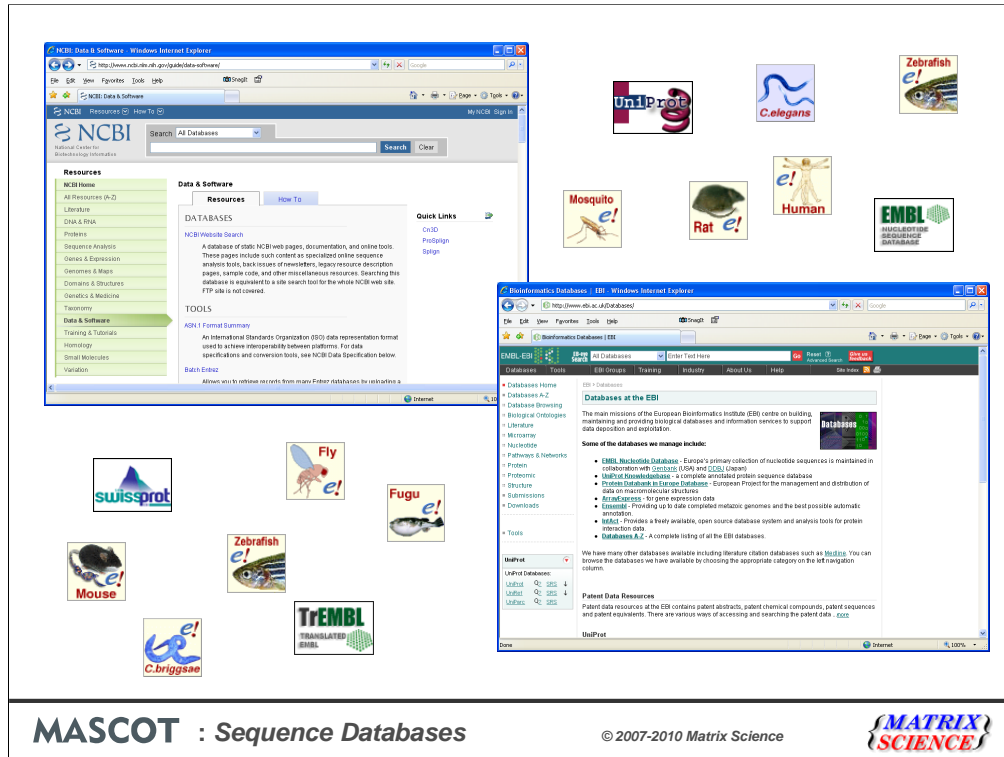


Sequence Databases

MASCOT

MATRIX
SCIENCE



MASCOT : Sequence Databases

© 2007-2010 Matrix Science

**MATRIX
SCIENCE**

When you install Mascot, it includes a copy of the Swiss-Prot protein database. However, it is almost certain that you and your colleagues will want to search other databases as well. There are very many to choose from, and Mascot allows you to have as many databases on-line for searching as you wish (limit of 64 in Mascot 2.2 and earlier).

Matrix Science doesn't supply sequence databases. Most databases are public domain, and there are a few sites that provide comprehensive database repositories. Two of the best known are NCBI and EBI. Here, you can download nr, GenBank, Swiss-Prot, EMBL, TrEMBL, etc.

For specialised databases, such as individual genomes, you may have to track down the FTP site of the group that is doing the sequencing.

Sequence Databases

Swiss-Prot (~500,000 entries)

- High quality, non-redundant; ideal for PMF & some MS/MS

NCBI nr, UniRef100 (~10,000,000 entries)

- Comprehensive, non-identical

UniRef90, UniRef50, etc.

- UniRef100 better for MS/MS; need explicit sequences

EST databases (>200,000,000 entries in translation)

- *Very* large and *very* redundant
- Not suitable for PMF

Sequences from a single genome

- Not suitable for PMF

MASCOT : *Sequence Databases*

© 2007-2010 Matrix Science



There are a huge number of database, and often it is not clear which is the appropriate one to choose for a search.

Swiss-Prot is acknowledged to be the best annotated database, but it is non-redundant, which is not ideal for MS/MS searches, where you often want explicit representations of every known sequence. Swiss-Prot is an ideal choice for PMF searches, where the loss of one or two peptides is not a concern.

The large, comprehensive, non-identical databases are the best choice for MS/MS searching where you don't want to miss any matches. NCBI nr and UniRef100 are the best examples of these databases, and contain similar sequences. If you search the non-redundant versions, you may miss some matches.

The EST databases are huge. Worth trying with high quality MS/MS data if a good match could not be found in a protein database. Not advisable for PMF, because many sequences correspond to protein fragments.

Single genome databases are good for protein characterisation using MS/MS data. You may want to include a contaminants database in the search, to ensure spectra from contaminants don't get mis-assigned to the target organism

NA Translation

K P I R L T A D L L A E T L Q A R R E W G P I F N I
 A S P S D # Q Q I S W Q K L Y K P E E S G G Q Y S T E
 Q A H Q T N S R S L G R N S T S Q K R V G A N I Q H
 CAAGCCCATCAGACTAACAGCAGATCTCTTGGCAGAACTCTACAGCCGAAGAGAGTGGGGCCAATATTCAACATT
 299200 299210 299220 299230 299240 299250 299260 299270
 TTCTGGGTAGTCTGATTGTCGCTAGAGAAACGCTTTGAGATGTTCTGGTCTTCTCTACCCCCGGTTATAAGTTGTAA
 C A W + V L L L D R P L F E V L W F L T P A L I + C E
 L G D S + C C I E Q C F S + L G S S L P P W Y E V N
 L G M L S V A S R K A S V R C A L L S H P G I N L M

```
Residue: FFLSSSSYY**CC*WLLLLPPPHHQRRRIIIMTTTTNNKSSRRVVVVAADDEEGGGG
Start: -----M-----
Base 1: TTTTNTTTTTTTTTTTCCCCCCCCCCCCCAAAAAAAAAAAGGGGGGGGGGGGGGGGGGGGGGG
Base 2: TTTTCCCCAAAGGGGTTCCTCCCAAAGGGGTTCCTCCCAAAGGGGTTCCTCCCAAAGGGG
Base 3: TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

* = stop

MASCOT : Sequence Databases

© 2007-2010 Matrix Science

**MATRIX
SCIENCE**

When we search a nucleic acid databases, Mascot always performs a 6 frame translation on the fly. That is, 3 reading frames from the forward strand and 3 reading frames from the complementary strand.

NA Translation

- Mascot translates on the fly in all 6 reading frames
- Translation starts from the beginning of the sequence, not from a start codon
- When a stop codon is encountered, inserts a gap and re-starts translation
- No attempt to resolve codon ambiguity
- Where taxonomy information is available, translation uses the correct genetic code.

The rules for NA translation in Mascot are

Translate the entire sequence, don't look for a start codon to begin

When a stop codon is encountered, leave a gap, and immediately re-start translation

There is no attempt to resolve ambiguous codons. For example, ACX can be translated as Threonine, because the identity of the last base is a don't care.

However, this is not done in Mascot.

Finally, all translations use the correct genetic code, as long as the taxonomy is known.

Which database?

- cmr.jcvi.org
- Microbial genomes
- *Helicobacter pylori*

The screenshot shows the JCVI CMR Genomes database interface. The page title is "Genomes: All Genomes". It displays a table of microbial genomes with columns for Organism Name, Kingdom, Taxon ID, Size, Complete Genome, and Sequencing Center. The table lists several bacterial genomes, including *Acidobacterium* and *Acidobacterium* species.

Organism Name (sort)	Kingdom (sort)	Taxon ID (sort)	Size (sort)	Complete Genome (sort)	Sequencing Center (sort)
<i>Acidobacterium</i> <i>marina</i> MBIC11017	Bacteria	229726	8.36 Mb	Yes	DOE Joint Genome Institute
<i>Acidobacterium</i> <i>lactuorum</i> PG-54	Bacteria	441760	1.49 Mb	Yes	BIPCM
<i>Acidobacterium</i> <i>sp.</i> JF-5	Bacteria	349163	3.96 Mb	Yes	DOE Joint Genome Institute
<i>Acidobacterium</i> <i>ferroplasma</i> ATCC 23270	Bacteria	243159	2.98 Mb	Yes	J. Craig Venter Institute
<i>Acidobacterium</i> <i>hacterium</i> F11045	Bacteria	204689	5.65 Mb	Yes	DOE Joint Genome Institute
<i>Acidobacterium</i> <i>cellulosum</i> 118	Bacteria	351807	2.44 Mb	Yes	DOE Joint Genome Institute
<i>Acidobacterium</i> <i>sp.</i> JF-5	Bacteria	387845	5.35 Mb	Yes	DOE Joint Genome Institute
<i>Acidobacterium</i> <i>sp.</i> JF-5	Bacteria	232721	4.58 Mb	Yes	DOE Joint Genome Institute
<i>Acidobacterium</i> <i>radiosolens</i> SK02	Bacteria	589318	3.28 Mb	No	J. Craig Venter Institute
<i>Acidobacterium</i> <i>sp.</i> ADP1	Bacteria	62977	3.59 Mb	Yes	Genoscope

MASCOT : Sequence Databases

© 2007-2010 Matrix Science

MATRIX
SCIENCE

Lets look at a typical small genome database.

JCVI, The J. Craig Venter Institute, has a list of completed microbial genomes.

Which database?

Assembled genomic DNA
sequence

h_pylori_26695.1con

Nucleic acid coding
sequences

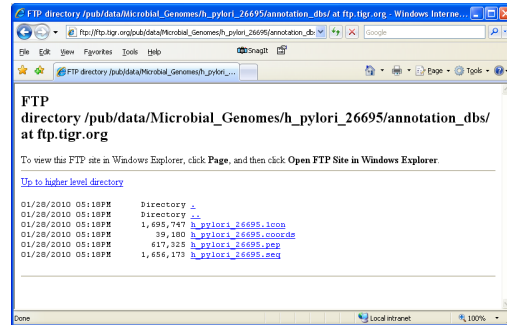
h_pylori_26695.seq

Coding sequences translated
to proteins

h_pylori_26695.pep

Table of co-ordinates for the
coding sequences in the
assembled chromosome

h_pylori_26695.coords



MASCOT : Sequence Databases

© 2007-2010 Matrix Science

MATRIX
SCIENCE

Follow the links for *Helicobacter pylori*, and these files are available for download.

- Assembled genomic DNA sequence
- Nucleic acid coding sequences
- Coding sequences translated to proteins
- Table of co-ordinates for the coding sequences in the assembled chromosome

If you are confident that the coding sequences and reading frames have been identified correctly, then the pep file would be the first choice for Mascot. It is possible to use the seq file, but this will result in slower searches, because Mascot has to translate each sequence in all six reading frames.

If you are not confident that the coding sequences and reading frames have been identified correctly, then you might wish to search the genomic DNA directly. *Helicobacter pylori* has a single chromosome, so h_pylori_26695.1con contains just one sequence, of length 1,667,867 bases.

Which database?

Assembled genomes

- Searching a database of one, (or a few), very long sequences is possible, but:
 - Mascot reports will be unwieldy
 - Memory inefficient
 - Better to split the sequence into segments
 - Small overlaps to ensure no peptide lost
 - Maintain frame numbering
- www.matrixscience.com/downloads/splitter.pl.gz

MASCOT : *Sequence Databases*

© 2007-2010 Matrix Science



Assembled genomes are not ideal for a Mascot search, because it would make the reports too unwieldy.

The longest human chromosome is chromosome 1 with 285 million base pairs

We don't know of any tools for reviewing the results which can handle 250 Mbp sequences.

Mascot requires a significant memory overhead to manipulate such long sequences, which means that unless you have a very large amount of RAM, the search is going to be using virtual memory ... i.e. swapping out to disk ... and run relatively slowly.

So, we recommend working with contigs or just chopping the chromosomes into more manageable lengths.

For efficient searching and reporting, the genomic DNA needs to be chopped into shorter sequences, with small overlaps to ensure no peptides are lost because they span a boundary. This is not a completely trivial task if you want to maintain the original forward and reverse frame numbering from chunk to chunk. A simple perl utility to split a long sequence can be downloaded from the Matrix Science web site.

Peptide Summary Report (Raft - 8) - Microsoft Internet Explorer

Address: http://hoala/mascot/cgi/master_results.pl?file=../data/20060120/F002123.dat

MASCOT Search Results

User : JSC
 Email :
 Search title : Raft - 8
 MS data file : C:\DOCUME~1\johnc\LOCALS~1\Temp\DisB6.tmp
 Database : Human IPI IPI_20060118 (57032 sequences; 24978932 residues)
 Timestamp : 20 Jan 2006 at 11:27:55 GMT

Significant hits:

IPI00007289	Tax_Id=9606 Alkaline phosphatase, placental type 1 precursor
IPI00290380	Tax_Id=9606 Alkaline phosphatase, placental-like precursor
IPI00013981	Tax_Id=9606 Proto-oncogene tyrosine-protein kinase YES
IPI00180292	Tax_Id=9606 Splice Isoform 5 of Brain-specific angiogenesis inhibitor 1-associated protein
IPI00228327	Tax_Id=9606 Keratin, type II cytoskeletal 1
IPI00383237	Tax_Id=9606 58 kDa protein
IPI00298622	Tax_Id=9606 Intestinal alkaline phosphatase precursor
IPI00019359	Tax_Id=9606 Keratin, type I cytoskeletal 9
IPI00019906	Tax_Id=9606 Splice Isoform 2 of Basigin precursor
IPI00298625	Tax_Id=9606 V-yes-1 Yamaguchi sarcoma Viral related oncogene homolog
IPI00179326	Tax_Id=9606 Insulin receptor tyrosine kinase substrate
IPI00026270	Tax_Id=9606 Carboxypeptidase M precursor
IPI00022624	Tax_Id=9606 Retinoic acid-induced protein 3
IPI00017184	Tax_Id=9606 EH-domain containing protein 1
IPI00021304	Tax_Id=9606 Keratin, type II cytoskeletal 2 epidermal
IPI00300725	Tax_Id=9606 Keratin, type II cytoskeletal 6A
IPI00220194	Tax_Id=9606 Solute carrier family 2, facilitated glucose transporter member 1
IPI00009505	Tax_Id=9606 Splice Isoform 1 of Beta-2-syntrophin
IPI00026059	Tax_Id=9606 Basic Beta 1 syntrophin
IPI00219365	Tax_Id=9606 Moesin
IPI00329115	Tax_Id=9606 Splice Isoform 1 of Protein C10orf47
IPI00455689	Tax_Id=9606 PREDICTED: similar to Keratin, type I cytoskeletal 18 (Cytokeratin 18) (K18) (C
IPI00455693	Tax_Id=9606 PREDICTED: similar to Keratin, type I cytoskeletal 18 (Cytokeratin 18) (K18) (C
IPI00004669	Tax_Id=9606 Polypeptide N-acetylgalactosaminyltransferase 2
IPI00026241	Tax_Id=9606 Bone marrow stromal antigen 2
IPI00028635	Tax_Id=9606 OTTHMP00000030901

MASCOT : Sequence Databases © 2007-2010 Matrix Science

MATRIX SCIENCE

To illustrate the features of the different types of database, we first searched a small dataset of a few hundred MS/MS spectra against a protein database, IPI human. We found significant matches to 28 human proteins

Peptide Summary Report (Raft - 8) - Microsoft Internet Explorer

Address: http://koala/mascot/cgi/master_results.pf?file=../data/20060120/F002125.dat

MASCOT Search Results

User : JSC
Email :
Search title : Raft - 8
MS data file : C:\DOCUMENT-1\johnc\LOCALS-1\Temp\DisB6.tmp
Database : EST_human_human_20060119 (45581712 sequences; 8078961710 residues)
Timestamp : 20 Jan 2006 at 13:51:44 GMT

Significant hits:

gi 47053565	BX458398 Homo sapiens PLACENTA Homo sapiens cDNA clone CS00E002YN21 5-PRIME, mRNA sequence
gi 82384040	DA832815 PLACE1 Homo sapiens cDNA clone PLACE1009867 5', mRNA sequence
gi 13341740	602507767F1 NIH_MGC_79 Homo sapiens cDNA clone IMAGE:4604921 5', mRNA sequence
gi 14051361	602631560F1 NCI_CGAP_Skn3 Homo sapiens cDNA clone IMAGE:4776638 5', mRNA sequence
gi 14461242	C83-GH0330-260101-650-F02 GH0330 Homo sapiens cDNA, mRNA sequence
gi 13910479	602627035F1 NCI_CGAP_Skn4 Homo sapiens cDNA clone IMAGE:4751984 5', mRNA sequence
gi 10351178	601434543F1 NIH_MGC_72 Homo sapiens cDNA clone IMAGE:3919498 5', mRNA sequence
gi 15932242	603034479F1 NIH_MGC_115 Homo sapiens cDNA clone IMAGE:5175739 5', mRNA sequence
gi 83080972	DB263154 UTERU2 Homo sapiens cDNA clone UTERU2021441 5', mRNA sequence
gi 80536324	DA607848 IMR322 Homo sapiens cDNA clone IMR322005139 5', mRNA sequence
gi 22683682	AGEHCOURT_7770784 NIH_MGC_70 Homo sapiens cDNA clone IMAGE:6021266 5', mRNA sequence
gi 20397858	AGEHCOURT_7549218 NIH_MGC_68 Homo sapiens cDNA clone IMAGE:6059273 5', mRNA sequence
gi 9127299	601111974F1 NIH_MGC_16 Homo sapiens cDNA clone IMAGE:3352650 5', mRNA sequence
gi 19180918	K-EST0096024 S22SHU16n1 Homo sapiens cDNA clone S22SHU16n1-96-C08 5', mRNA sequence
gi 12959439	EST178022 Colon carcinoma (HCC) cell line Homo sapiens cDNA 5' end similar to Yamaguchi sar
gi 13909289	602625586F1 NCI_CGAP_Skn4 Homo sapiens cDNA clone IMAGE:4750605 5', mRNA sequence
gi 10950367	AU125651 NT2RM4 Homo sapiens cDNA clone NT2RM4001963 5', mRNA sequence
gi 19359946	AGEHCOURT_6640963 NIH_MGC_99 Homo sapiens cDNA clone IMAGE:5434174 5', mRNA sequence
gi 2354003	nm76c02.s1 NCI_CGAP_Co9 Homo sapiens cDNA clone IMAGE:1074146 3' similar to TR:G1203820 G12
gi 80827804	DA585073 HLUNG2 Homo sapiens cDNA clone HLUNG2005215 5', mRNA sequence
gi 1843987	zr06e04.r1 Stratagene NT2 neuronal precursor 937230 Homo sapiens cDNA clone IMAGE:650718 5'
gi 52182059	BP267627 Sugano cDNA library, thyroid JTH Homo sapiens cDNA clone JTH07584, mRNA sequence
gi 82423761	DA833423 PLACE1 Homo sapiens cDNA clone PLACE1011132 5', mRNA sequence
gi 11643213	602185774F1 NIH_MGC_45 Homo sapiens cDNA clone IMAGE:4309987 5', mRNA sequence
gi 36346787	AGEHCOURT_15621164 NIH_MGC_147 Homo sapiens cDNA clone IMAGE:30531295 5', mRNA sequence
gi 18789274	AGEHCOURT_6575553 NIH_MGC_98 Homo sapiens cDNA clone IMAGE:5479414 5', mRNA sequence

Done Local intranet

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

With EST_human, we obtained almost the same results, just a couple of additional peptide matches. However, look at the hit-list on this report ... unlike the protein database search, it doesn't immediately communicate which proteins have been found. I'll return to this issue later.

Peptide Summary Report (Raft - 8) - Microsoft Internet Explorer

Address: http://hoala/mascot/cgi/master_results.pl?file=../data/20060120/F002125.dat

Select All Select None Search Selected ☐ Error tolerant Archive Report

1. [g1147053565](#) Mass: 34632 Score: 618 Queries matched: 11 Frame: 2
BX458398 Homo sapiens PLACENTA Homo sapiens cDNA clone CSODE002YN21 5-PRIME, mRNA sequence
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 22	460.17	918.32	918.51	-0.19	0	51	1.4	1	K.SVGGVVTTTR.V
<input checked="" type="checkbox"/> 23	462.68	923.35	923.51	-0.16	0	35	44	1	R.FPYVALSK.T
<input checked="" type="checkbox"/> 75	567.66	1133.30	1133.55	-0.25	0	53	1.2	1	R.GNEVISVMNR.A + Oxidation (M)
<input checked="" type="checkbox"/> 99	614.20	1226.39	1226.63	-0.25	0	30	1.9e+002	1	K.LGPELPLAMDR.F + Oxidation (M)
<input checked="" type="checkbox"/> 111	633.21	1304.41	1304.68	-0.28	0	69	0.025	1	K.GHFQTIQLSAAAR.F
<input checked="" type="checkbox"/> 137	726.18	1450.35	1450.65	-0.30	0	88	0.00033	1	R.NWYSADADVPASAR.Q
<input checked="" type="checkbox"/> 177	427.87	1707.46	1707.84	-0.38	0	71	0.013	1	R.VQHASPAQTYAHTVNR.N
<input checked="" type="checkbox"/> 215	975.81	1949.61	1950.02	-0.42	0	78	0.0024	1	K.NLILFLGDGMGVSTVTAAR.I + Oxidation (M)
<input checked="" type="checkbox"/> 226	1001.20	2000.39	2000.81	-0.42	0	84	0.00048	1	R.HGTPDPPEYDDYSQGGTR.L + Oxidation (M)
<input checked="" type="checkbox"/> 227	667.80	2000.39	2000.81	-0.41	0	(52)	0.93	1	R.HGTPDPPEYDDYSQGGTR.L + Oxidation (M)
<input checked="" type="checkbox"/> 298	901.59	2701.76	2702.30	-0.54	0	60	0.075	1	R.QEGCQDIATQLISNMDIDVILGGGR.K

2. [g1182384040](#) Mass: 26303 Score: 436 Queries matched: 8 Frame: 1
DA832815 PLACE1 Homo sapiens cDNA clone PLACE1009867 5', mRNA sequence
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 22	460.17	918.32	918.51	-0.19	0	51	1.4	1	K.SVGGVVTTTR.V
<input checked="" type="checkbox"/> 75	567.66	1133.30	1133.55	-0.25	0	53	1.2	1	R.GNEVISVMNR.A + Oxidation (M)
<input checked="" type="checkbox"/> 137	726.18	1450.35	1450.65	-0.30	0	88	0.00033	1	R.NWYSADADVPASAR.Q
<input checked="" type="checkbox"/> 177	427.87	1707.46	1707.84	-0.38	0	71	0.013	1	R.VQHASPAQTYAHTVNR.N

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

The master results report from the EST search looks pretty similar to the IPI search, except that the EST sequences are mostly shorter than full length proteins, so the peptide matches are more scattered. If we click on the protein accession number link

Mascot Search Results: Protein View

Match to: **gi|14051361** Score: **397**
602631560F1 HCl_CGAP_Skn3 Homo sapiens cDNA clone IMAGE:4776631
 Found in search of C:\DOCUME~1\johnc\LOCALS~1\Temp\DisB6.tmp
 Translated in frame 1

NB Matches were also found in other frames indicating a possible frame shift. Only matches in frame 1 are shown in this report

Show frame: **1**

Nominal mass (M_0): **31277**; Calculated pI value: **5.17**
 NCBI BLAST search of **gi|14051361** against nr
 Unformatted [sequence string](#) for pasting into other applications

Taxonomy: **Homo sapiens**

Fixed modifications: Carbamidomethyl (C)
 Variable modifications: Acetyl (N-term), Oxidation (M)
 Cleavage by Trypsin/P: cuts C-term side of KR
 Sequence Coverage: **23%**

Matched peptides shown in **Bold Red**

```

1 VRFLEQQNQV LQTKWELLQQ VDTSTRTHNL EPYFESFINN LRREVQLKS
51 DQSRLDSELK NMQDMVEDYR NKYEDEINKR TNAHEFVTI KKQVDGAYNT
101 KVDLQAKLDN LQCEIDFLTA LYQAELSQMG TQISETNVIL SHMNNRSLDL
131 DSLIAEVNAQ YEDIAQNSKA EAESLYQSKY EELQITAGR GDSVRNSKIE
201 TSLELRVVIQR LRSEIDNVKK QISNLQBSIS DAEQGENAS RHPRTS_NTW
251 RMPCSRPRQD LGRLLP_LP
  
```

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

We get a protein view. This is similar to the protein view for a protein database entry, except we have drop down list for the different translation frames. For this particular entry, most of the matches have been found in reading frame 1.

But, as so often happens, there is a frame shift in this entry, and there are additional matches in frame 3.

MASCOT : Sequence Databases

© 2007-2010 Matrix Science

MATRIX SCIENCE

Going back to the issue of the hit list and the descriptions not saying very much. There are several problems here. One is that EST databases usually have a huge amount of redundancy, which can make for very long reports. Another problem is that the sequences tend to be short, so we don't get much grouping of peptide matches into protein matches.

To address this problem, we can use the UniGene index from the National Center for Biotechnology Information to simplify the search results.

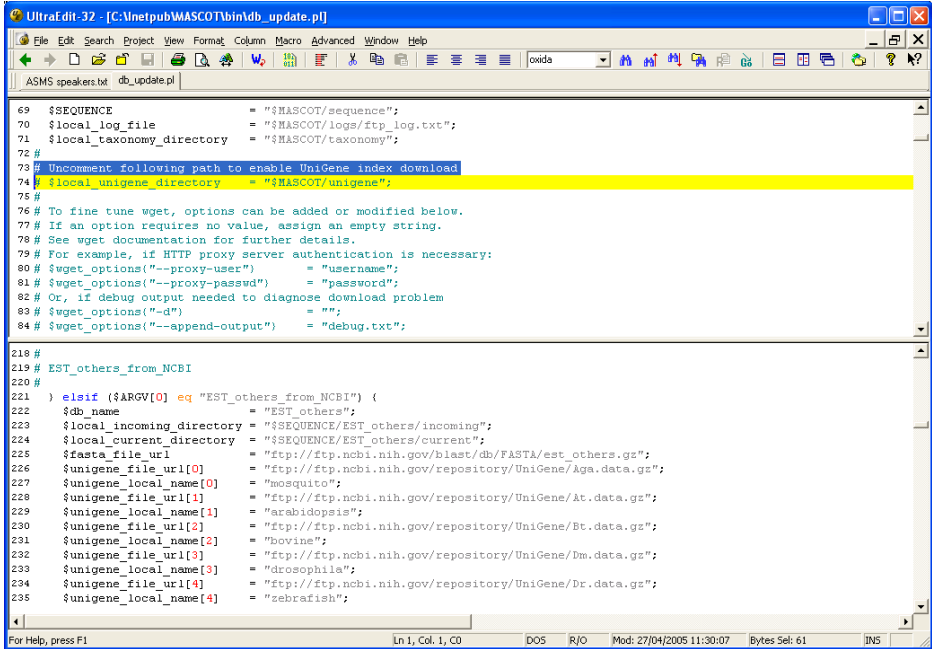
UniGene: An Organized View of the Transcriptome.

Each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location.

Species	UniGene Entries
Chordata	
Mammalia	
<i>Bos taurus</i> (cow)	42,957
<i>Canis lupus familiaris</i> (dog)	27,853
<i>Equus caballus</i> (horse)	8,348
<i>Homo sapiens</i> (human)	123,253
<i>Macaca fascicularis</i> (crab-eating macaque)	12,657
<i>Macaca mulatta</i> (rhesus monkey)	15,359
<i>Monodelphis domestica</i> (gray short-tailed opossum)	359
<i>Mus musculus</i> (mouse)	78,324
<i>Ornithorhynchus anatinus</i> (platypus)	1,831
<i>Oryctolagus cuniculus</i> (rabbit)	6,576
<i>Ovis aries</i> (sheep)	18,645
<i>Papio anubis</i> (olive baboon)	11,904
<i>Petromyscus maniculatus</i> (deer mouse)	10,429
<i>Pongo abelii</i> (Sumatran orangutan)	6,996
<i>Rattus norvegicus</i> (Norway rat)	63,427
<i>Rumex acetosa</i> (garden sorrel)	64,676

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

UniGene is not a sequence database, it is an index which is created by BLASTing GenBank sequences against themselves to cluster them into gene families.



```

UltraEdit-32 - [C:\inetpub\WASCOT\bin\db_update.pl]
File Edit Search Project View Format Column Macro Advanced Window Help
ASMS speakers.txt db_update.pl

69 $SEQUENCE           = "$MASCOT/sequence";
70 $local_log_file      = "$MASCOT/logs/ftp_log.txt";
71 $local_taxonomy_directory = "$MASCOT/taxonomy";
72 #
73 # Uncomment following path to enable UniGene index download
74 $local_unigene_directory = "$MASCOT/unigene";
75 #
76 # To fine tune wget, options can be added or modified below.
77 # If an option requires no value, assign an empty string.
78 # See wget documentation for further details.
79 # For example, if HTTP proxy server authentication is necessary:
80 # $wget_options("--proxy-user") = "username";
81 # $wget_options("--proxy-passwd") = "password";
82 # Or, if debug output needed to diagnose download problem
83 # $wget_options("-d") = "";
84 # $wget_options("--append-output") = "debug.txt";

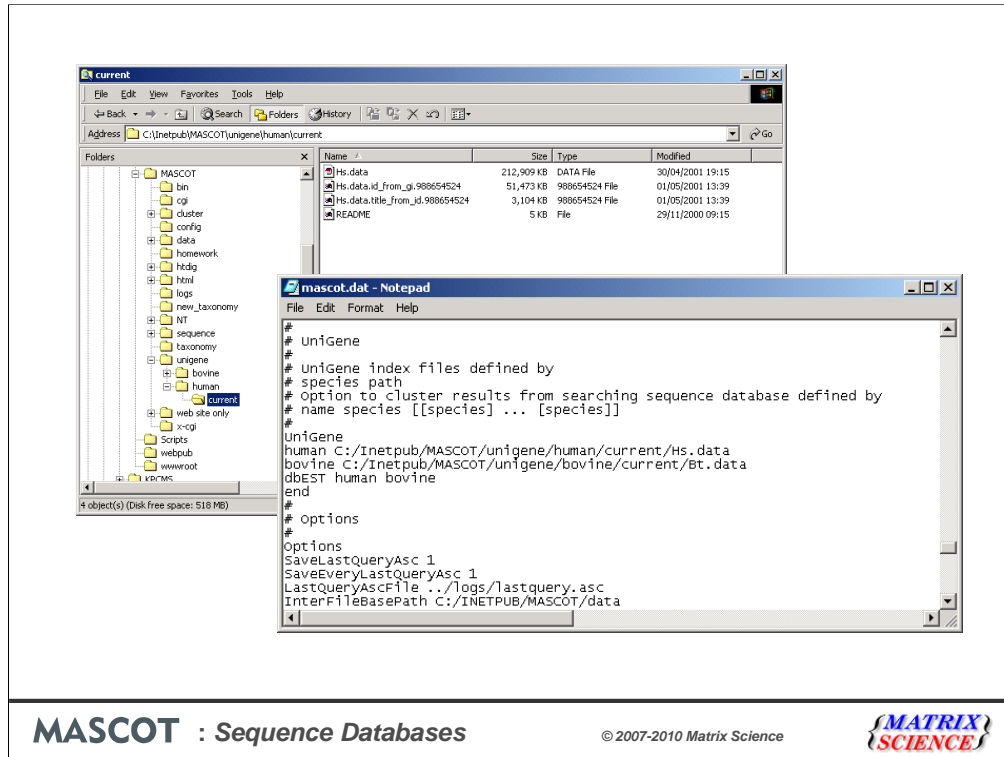
218 #
219 # EST_others_from_NCBI
220 #
221 ) elsif ($ARGV[0] eq "EST_others_from_NCBI") {
222     $db_name = "EST_others";
223     $local_incoming_directory = "$SEQUENCE/EST_others/incoming";
224     $local_current_directory = "$SEQUENCE/EST_others/current";
225     $fasta_file_url = "ftp://ftp.ncbi.nih.gov/blast/db/FASTA/est_others.gz";
226     $unigene_file_url[0] = "ftp://ftp.ncbi.nih.gov/repository/UniGene/aga.data.gz";
227     $unigene_local_name[0] = "mosquito";
228     $unigene_file_url[1] = "ftp://ftp.ncbi.nih.gov/repository/UniGene/at.data.gz";
229     $unigene_local_name[1] = "arabidopsis";
230     $unigene_file_url[2] = "ftp://ftp.ncbi.nih.gov/repository/UniGene/Bt.data.gz";
231     $unigene_local_name[2] = "bovine";
232     $unigene_file_url[3] = "ftp://ftp.ncbi.nih.gov/repository/UniGene/Dm.data.gz";
233     $unigene_local_name[3] = "drosophila";
234     $unigene_file_url[4] = "ftp://ftp.ncbi.nih.gov/repository/UniGene/Dr.data.gz";
235     $unigene_local_name[4] = "zebrafish";

```

For Help, press F1 Ln 1, Col 1, CO DOS R/O Mod: 27/04/2005 11:30:07 Bytes Sel: 61 INS

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

Unigene can be downloaded from the NCBI FTP site. The database update script, which will be covered in a later session, takes care of this automatically



To use a unigene index in Mascot, the data file for the species is downloaded and unpacked into a suitable directory structure

Then a few lines are added to the Mascot configuration file, mascot.dat.

Peptide Summary Report (Raft - 8) - Microsoft Internet Explorer

Address: http://koala/mascot/cgi/master_results.pf?file=../data/20060120/F002125.dat

MASCOT Search Results

User : JSC
 Email :
 Search title : Raft - 8
 MS data file : C:\DOCUMENT-1\johnc\LOCALS-1\Temp\DisB6.tmp
 Database : EST_human_human_20060119 (45581712 sequences; 8078961710 res
 Timestamp : 20 Jan 2006 at 13:51:44 GMT

Significant hits:

gi147053565	BX458398 Homo sapiens PLACENTA Homo sapiens cDNA
gi182384040	DA832815 PLACE1 Homo sapiens cDNA clone PLACE10
gi113341740	602507767F1 NIH_MGC_79 Homo sapiens cDNA clone
gi114051361	602631560F1 NCI_CGAP_Skn3 Homo sapiens cDNA clone
gi14461242	CM3-GH0330-260101-650-F02 GH0330 Homo sapiens cDNA clone
gi113910479	602627033F1 NCI_CGAP_Skn4 Homo sapiens cDNA clone
gi110351178	601434543F1 NIH_MGC_72 Homo sapiens cDNA clone
gi115932242	603034479F1 NIH_MGC_115 Homo sapiens cDNA clone
gi183080972	DB263154 UTERU2 Homo sapiens cDNA clone UTERU20
gi180536324	DA607848 IMR322 Homo sapiens cDNA clone IMR3220
gi122683682	AGEHCOURT_7770784 NIH_MGC_70 Homo sapiens cDNA clone IMAGE:6021266 5', mRNA sequence
gi120397858	AGEHCOURT_7549218 NIH_MGC_68 Homo sapiens cDNA clone IMAGE:6059273 5', mRNA sequence
gi19127299	601111974F1 NIH_MGC_16 Homo sapiens cDNA clone IMAGE:3352650 5', mRNA sequence
gi119180918	K-EST0096024 S22SHU16n1 Homo sapiens cDNA clone S22SHU16n1-96-C08 5', mRNA sequence
gi11259439	EST178022 Colon carcinoma (HCC) cell line Homo sapiens cDNA 5' end similar to Yamaguchi sar
gi113909289	602625586F1 NCI_CGAP_Skn4 Homo sapiens cDNA clone IMAGE:4750605 5', mRNA sequence
gi110950367	AU125651 NT2RM4 Homo sapiens cDNA clone NT2RM4001963 5', mRNA sequence
gi119359946	AGEHCOURT_6640963 NIH_MGC_99 Homo sapiens cDNA clone IMAGE:5434174 5', mRNA sequence
gi12354003	nm76c02.s1 NCI_CGAP_Co9 Homo sapiens cDNA clone IMAGE:1074146 3' similar to TR:G1203820 G12
gi180827804	DA585073 HLUNG2 Homo sapiens cDNA clone HLUNG2005215 5', mRNA sequence
gi11843987	zr06e04.r1 Stratagene NT2 neuronal precursor 937230 Homo sapiens cDNA clone IMAGE:650718 5'
gi152182059	BP267627 Sugano cDNA library, thyroid JTH Homo sapiens cDNA clone JTH07584, mRNA sequence
gi182423761	DA833423 PLACE1 Homo sapiens cDNA clone PLACE1011132 5', mRNA sequence
gi111643213	602185774F1 NIH_MGC_45 Homo sapiens cDNA clone IMAGE:4309987 5', mRNA sequence
gi136346787	AGEHCOURT_15621164 NIH_MGC_147 Homo sapiens cDNA clone IMAGE:30531295 5', mRNA sequence
gi118789274	AGEHCOURT_6575553 NIH_MGC_98 Homo sapiens cDNA clone IMAGE:5479414 5', mRNA sequence

UniGene index: human
 Max. number of: human
 Ions score cut-off: 0
 Sort unassigned: Decreasing Score
 Show statistics: ☐
 Require: ☐

MASCOT : Sequence Databases

© 2007-2010 Matrix Science

MATRIX SCIENCE

So, if Unigene is configured, we can select human from the drop-down list in the format controls

MASCOT Search Results

User : JSC
 Email :
 Search title : Raft - 8
 MS data file : C:\DOCUMENT-1\johnc\LOCALS-1\Temp\DisB6.tmp
 Database : EST_human_human_20060119 (45581712 sequences; 8078961710 residues)
 Timestamp : 20 Jan 2006 at 13:51:44 GMT

Significant hits:

- [Ms.284255](#) ALPP Alkaline phosphatase, placental (Regan isozyme)
- [Ms.333509](#) ALPPL2 Alkaline phosphatase, placental-like 2
- [Ms.194148](#) YES1 V-yes-1 Yamaguchi sarcoma viral oncogene homolog 1
- [Ms.808928](#) KRT1 Keratin 1 (epidermolytic hyperkeratosis)
- [Ms.128316](#) BAIAP2 BAI1-associated protein 2
- [Ms.198281](#) PKM2 Pyruvate kinase, muscle
- [Ms.501293](#) BSG Basigin (OK blood group)
- [Ms.439552](#) EEF1A1 Eukaryotic translation elongation factor 1 alpha 1
- [Ms.491767](#) LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
- [Ms.567251](#) CPM Carboxypeptidase M
- [Ms.514581](#) ACTG1 Actin, gamma 1
- [Ms.567541](#) LOC55971 BAI1-associated protein 2-like 1
- [Ms.498569](#) RPS2 RPL, U64 small nucleolar
- [Ms.433845](#) KRT5 Keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types)
- [Ms.433670](#) FTL Ferritin, light polypeptide
- [Ms.194691](#) GPCR5A G protein-coupled receptor, family C, group 5, member A
- [Ms.473721](#) SLC2A1 Solute carrier family 2 (facilitated glucose transporter), member 1
- [Ms.78771](#) PGK1 Phosphoglycerate kinase 1
- [Ms.487027](#) VIL2 Villin 2 (ezrin)
- [Ms.370895](#) RPN2 Ribophorin II
- [Ms.461117](#) SNTB2 Syntrophin, beta 2 (dystrophin-associated protein A1, 59kDa, basic component 2)
- [Ms.523774](#) EHD1 EH-domain containing 1
- [Ms.569397](#) TGFBI Transforming growth factor, beta-induced, 68kDa
- [Ms.528348](#) UBC Ubiquitin C
- [Ms.511325](#) PRSS1 Protease, serine, 1 (trypsin 1)
- [Ms.546286](#) RPS3 Ribosomal protein S3

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

Now, using the UniGene index as a lookup table, we can transform the results of an EST search.

This is now a much clearer picture, very similar to the protein database result. Please remember that we are not clustering the database sequences into consensus sequences prior to searching. This could lead to matches being missed. UniGene is being used after the search, to map one set of accessions to a more useful set.

MASCOT : Sequence Databases © 2007-2010 Matrix Science

When we look at individual hits in the report, we see the benefits of UniGene mapping. Here we have two hits from the EST search. They have no peptide matches in common, and the entry names also give no clue as to the protein function. However, when we look at the UniGene report, we find that these matches all belong to the same gene, for alkaline phosphatase.

Mascot Search Results: Protein View - Microsoft Internet Explorer

Address: http://hoala/mascot/cg/protein_view.pl?file=.../data/20060120/F002125.dat&hit=Hs%2e284255&px=1&protscore=1051.78&UNIGENE=human8_mudpit=99999999

Mascot Search Results

UniGene View

ID: Hs.284255
 TITLE: Alkaline phosphatase, placental (Regan isozyme)
 GENE: ALPP
 CYTOBAND: 2q37
 LOCUSLINK: 250
 HOMOL: YES
 EXPRESS: placenta ; mammary gland ; other ; colon ; mixed ; heart ; uterus
 RESTR_EXPR: placenta adult
 CHROMOSOME: 2
 STS: ACC=A002R47 UNISTS=27519
 STS: ACC=GBB:216786 UNISTS=156214
 PROTSIM: ORG=Escherichia coli; PROTI=1310910; PROTI=ref:NP_001623.2; PCT=100.00; ALN=535
 PROTSIM: ORG=Homo sapiens; PROTI=130742; PROTI=sp:P24823; PCT=73.46; ALN=535
 PROTSIM: ORG=Rattus norvegicus; PROTI=130746; PROTI=sp:P15693; PCT=72.50; ALN=525
 PROTSIM: ORG=Saccharomyces cerevisiae; PROTI=2117980; PROTI=pir:S69648; PCT=31.32; ALN=333
 SCOUNT: 104
 SEQUENCE: ACC=BC094743.1; NID=g63100303; PID=g63100304; SEQTYPE=mRNA
 SEQUENCE: ACC=BC009647.1; NID=g16307117; PID=g16307118; SEQTYPE=mRNA
 SEQUENCE: ACC=BC068501.1; NID=g46250428; PID=g46250429; SEQTYPE=mRNA
 SEQUENCE: ACC=NM_001632.2; NID=g13787194; PID=g13787195; SEQTYPE=mRNA
 SEQUENCE: ACC=M14170.1; NID=g178469; PID=g178470; SEQTYPE=mRNA
 SEQUENCE: ACC=M12551.1; NID=g178463; PID=g178464; SEQTYPE=mRNA
 SEQUENCE: ACC=M14169.1; NID=g178467; PID=g178468; SEQTYPE=mRNA
 SEQUENCE: ACC=AK075432.1; NID=g22761517; SEQTYPE=mRNA
 SEQUENCE: ACC=CR621490.1; NID=g50502297; SEQTYPE=HTC
 SEQUENCE: ACC=CR599162.1; NID=g50479969; SEQTYPE=HTC
 SEQUENCE: ACC=AF217992.1; NID=g10441914; SEQTYPE=mRNA
 SEQUENCE: ACC=AL576838.3; NID=g46255926; CLONE=CSOD1079YA12; END=3'; LID=13021; SEQTYPE=EST
 SEQUENCE: ACC=BX343738.2; NID=g46268692; CLONE=CSOD1017YB14; END=3'; LID=13021; SEQTYPE=EST

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

When you click on the accession number link of a unigene filtered report, you get full details for that particular gene family.

Human Genome Statistics

- 3×10^9 bases
(EST_human is $\sim 4 \times 10^9$ bases)
- 6×10^9 residues in 6 frame translation
- 99.75% of translated sequence is non-coding
- $\sim 1.5 \times 10^5$ tryptic limit peptides of 1500 Da \pm 0.5
- $\sim 6 \times 10^7$ no-enzyme peptides of 1500 Da \pm 0.5

We can also perform MS/MS searches on the raw genomic sequence data. Let's just look at some numbers for the assembled human genome.

The human genome assembly is approximately 3 billion bases, which makes it a little smaller than EST_human.

Since we must translate in all 6 reading frames, this corresponds to 6 billion amino acid residues.

In the human genome, only 1.5% of the sequence codes for proteins. This means that 99.75% of the 6 frame translation is non-coding and simply contributes to the background of random matches. This is a good test of the discrimination of the scoring scheme.

If we are matching MS/MS data from a tryptic peptide of nominal mass 1500 Da against the human genome, we are going to have to test 150 thousand peptides. Which sounds bad, but is not nearly as bad as the no-enzyme case where we have to test 60 million!

Entrez Genome view - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606

NCBI

NCBI Map Viewer

PubMed Nucleotide Protein Genome Gene Structure PopSet Taxonomy Help

Search for on chromosome(s) assembly All Find Advanced Search

Map Viewer

Map Viewer Home

Map Viewer Help

Human Maps Help

Release Notes

NCBI Resources

Genome Project

TaxPlot

Consensus Coding Sequence (CCDS)

ORC

Human Genome Resources

NCBI Handbook

RefSeq

Trace Archive (Watson)

Trace Archive (Vent)

Trace FTP (Personal Genomics)

Homo sapiens (human) genome view

Build 37.1 statistics [Switch to previous build](#)

[BLAST search the human genome](#)

1 2 3 4 5 6 7 8 9 10 11 12 13

14 15 16 17 18 19 20 21 22 X Y

Lineage: Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorhina, Catarrhini, Hominoidea, Homo, Homo sapiens

August 2009: NCBI released an updated version of the human genome reference genome assembly and updated annotation for all available assemblies. The reference assembly update includes modifications to all chromosomes and adds 9 alternate loci to the reference assembly definition; the updated assembly, named GRCh37, was provided by the [Genome Reference Consortium \(GRC\)](#). The previous version of the reference genome assembly, [NCBI Build 36.3](#), can still be accessed for Map Viewer display and for BLAST. For additional information about changes, statistics, and the status of the CCDS project please refer to:

- [Release Notes](#)
- [Statistics](#)

MASCOT : Sequence Databases

© 2007-2010 Matrix Science

MATRIX SCIENCE

You can download the human genome sequences from NCBI.

File Name	Size	Modified
hs_alt_HuRef_chr9.fa.gz	33,044,813	08/04/2009 12:00AM
hs_alt_HuRef_chr9.mfa.gz	35,362,972	08/04/2009 12:00AM
hs_alt_HuRef_chrX.agp.gz	439,186	08/04/2009 12:00AM
hs_alt_HuRef_chrX.fa.gz	41,167,877	08/04/2009 12:00AM
hs_alt_HuRef_chrX.mfa.gz	43,913,767	08/04/2009 12:00AM
hs_alt_HuRef_chrY.agp.gz	48,107	08/04/2009 12:00AM
hs_alt_HuRef_chrY.fa.gz	5,482,165	08/04/2009 12:00AM
hs_alt_HuRef_chrY.mfa.gz	5,834,008	08/04/2009 12:00AM
hs_ref_GRCh37_chr1.agp.gz	45,812	08/04/2009 12:00AM
hs_ref_GRCh37_chr1.fa.gz	67,847,132	08/04/2009 12:00AM
hs_ref_GRCh37_chr1.mfa.gz	72,558,118	08/04/2009 12:00AM
hs_ref_GRCh37_chr10.agp.gz	23,259	08/04/2009 12:00AM
hs_ref_GRCh37_chr10.fa.gz	39,588,518	08/04/2009 12:00AM
hs_ref_GRCh37_chr10.mfa.gz	42,330,042	08/04/2009 12:00AM
hs_ref_GRCh37_chr11.agp.gz	22,806	08/04/2009 12:00AM
hs_ref_GRCh37_chr11.fa.gz	39,520,708	08/04/2009 12:00AM
hs_ref_GRCh37_chr11.mfa.gz	42,270,563	08/04/2009 12:00AM
hs_ref_GRCh37_chr12.agp.gz	24,207	08/04/2009 12:00AM
hs_ref_GRCh37_chr12.fa.gz	39,220,368	08/04/2009 12:00AM
hs_ref_GRCh37_chr12.mfa.gz	41,945,418	08/04/2009 12:00AM
hs_ref_GRCh37_chr13.agp.gz	17,361	08/04/2009 12:00AM
hs_ref_GRCh37_chr13.fa.gz	28,984,000	08/04/2009 12:00AM
hs_ref_GRCh37_chr13.mfa.gz	31,028,361	08/04/2009 12:00AM
hs_ref_GRCh37_chr14.agp.gz	14,326	08/04/2009 12:00AM
hs_ref_GRCh37_chr14.fa.gz	26,666,936	08/04/2009 12:00AM
hs_ref_GRCh37_chr14.mfa.gz	28,505,883	08/04/2009 12:00AM
hs_ref_GRCh37_chr15.agp.gz	15,010	08/04/2009 12:00AM
hs_ref_GRCh37_chr15.fa.gz	24,698,438	08/04/2009 12:00AM
hs_ref_GRCh37_chr15.mfa.gz	26,397,015	08/04/2009 12:00AM
hs_ref_GRCh37_chr16.agp.gz	15,241	08/04/2009 12:00AM

MASCOT : Sequence Databases

© 2007-2010 Matrix Science



We chose the assembled chromosomes, 24 files. Although you could search this as a 24 entry database, this is not memory efficient, so we used the script mentioned earlier to split the chromosome sequences into overlapping segments of 12 kb

Peptide Summary Report (Raft - 8) - Microsoft Internet Explorer

Address: http://hoala/mascot/cg/master_results.pl?file=../data/20060120/F002128.dat

MASCOT Search Results

User : JSC
 Email :
 Search title : Raft - 8
 MS data file : C:\DOCUMENT-1\johnc\LOCALS-1\Temp\DisB6.tmp
 Database : MG 20040910 (1538466 sequences; 6214458308 residues)
 Timestamp : 20 Jan 2006 at 14:42:20 GMT

Significant hits:

chr2_19423	bases 233064001-233076121 Homo sapiens chromosome 2, complete sequence
chr2_19425	bases 2330888001-233100121 Homo sapiens chromosome 2, complete sequence
chr12_4280	bases 51348001-51360120 Homo sapiens chromosome 12, complete sequence
chr19_43	bases 528001-540121 Homo sapiens chromosome 19, complete sequence
chr22_2031	bases 24360001-24372122 Homo sapiens chromosome 22, complete sequence
chr18_61	bases 720001-732121 Homo sapiens chromosome 18, complete sequence
chr17_3082	bases 36972001-36984121 Homo sapiens chromosome 17, complete sequence
chr15_5858	bases 70284001-70296120 Homo sapiens chromosome 15, complete sequence
chr17_6392	bases 76692001-76704121 Homo sapiens chromosome 17, complete sequence
chr18_62	bases 732001-744121 Homo sapiens chromosome 18, complete sequence
chr15_5857	bases 70272001-70284120 Homo sapiens chromosome 15, complete sequence
chr12_1080	bases 12948001-12960120 Homo sapiens chromosome 12, complete sequence
chr12_4278	bases 51324001-51336120 Homo sapiens chromosome 12, complete sequence
chr20_2513	bases 30144001-30156120 Homo sapiens chromosome 20, complete sequence
chr1_3589	bases 43056001-43068120 Homo sapiens chromosome 1, complete sequence
chr7_8132	bases 97572001-97584121 Homo sapiens chromosome 7, complete sequence
chr17_1804	bases 21636001-21648121 Homo sapiens chromosome 17, complete sequence
chr11_5366	bases 64380001-64392121 Homo sapiens chromosome 11, complete sequence
chr5_1025	bases 123000001-123012121 Homo sapiens chromosome 5, complete sequence

Probability Based Mowse Score

Ions score is $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event.
 Individual ions scores > 63 indicate identity or extensive homology ($p < 0.05$).
 Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.

Address: http://hoala/mascot/cg/master_results.pl?file=../data/20060120/F002128.dat#H:19

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

This is the result of searching our data against the human genome assembly. If you thought the EST_human entry titles were uninformative, how much worse is this?

Peptide Summary Report (Raft - 8) - Microsoft Internet Explorer

Address http://hoala/mascot/cgi/master_results.pl?file=../data/20060120/F002128.dat

Select All Select None Search Selected ☐ Error tolerant Archive Report

1. [chr2_19423](#) Mass: 428948 Score: 854 Queries matched: 15
bases 233064001-233076121 Homo sapiens chromosome 2, complete sequence
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 22	460.17	918.32	918.51	-0.19	0	51	0.99	1	K.SVGGVVITR.V
<input checked="" type="checkbox"/> 23	462.68	923.35	923.51	-0.16	0	35		31	R.FPYVALSK.V
<input checked="" type="checkbox"/> 75	567.66	1133.30	1133.55	-0.25	0	53	0.81	1	R.GHEVISVHNR.A + Oxidation (M)
<input checked="" type="checkbox"/> 99	614.20	1226.39	1226.63	-0.25	0	30	1.2e+002	1	K.LGPEIPLANDR.F + Oxidation (M)
<input checked="" type="checkbox"/> 111	653.21	1304.41	1304.68	-0.28	0	69	0.018	1	K.GHFTIIGLSAAAR.F
<input checked="" type="checkbox"/> 137	726.18	1450.35	1450.65	-0.30	0	88	0.00023	1	R.HWYSDADVPASAR.Q
<input checked="" type="checkbox"/> 162	820.73	1639.44	1639.78	-0.33	0	103	5.5e-006	1	R.ALTEIMFDDAIER.A + Oxidation (M)
<input checked="" type="checkbox"/> 177	427.87	1707.46	1707.84	-0.38	0	71	0.0088	1	R.VQHASPAGTYHTVNR.N
<input checked="" type="checkbox"/> 179	864.29	1726.56	1726.93	-0.37	0	46	2.5	1	K.AYTVLLYGHGPGYVLK.D
<input checked="" type="checkbox"/> 210	956.24	1910.47	1910.86	-0.39	0	31		79	R.DSTLDPISLMENTEALR.L + 2 Oxidation (M)
<input checked="" type="checkbox"/> 226	1001.20	2000.39	2000.81	-0.42	0	84	0.00032	1	R.MGTPDPEYPDDYSQGGTR.L + Oxidation (M)
<input checked="" type="checkbox"/> 227	667.80	2000.39	2000.81	-0.41	0	(52)	0.62	1	R.MGTPDPEYPDDYSQGGTR.L + Oxidation (M)
<input checked="" type="checkbox"/> 269	790.22	2367.63	2368.13	-0.50	0	118	9.7e-008	1	R.QQSAVPLDEETHAGEDVAVFAR.G
<input checked="" type="checkbox"/> 318	1078.63	3232.88	3233.56	-0.69	0	(10)	2.6e+003	5	R.AGQLTSEEDTLSLVTADHSVFSFGGYPLR.G
<input checked="" type="checkbox"/> 319	809.24	3232.91	3233.56	-0.65	0	105	8.8e-007	1	R.AGQLTSEEDTLSLVTADHSVFSFGGYPLR.G

2. [chr2_19425](#) Mass: 436943 Score: 590 Queries matched: 10
bases 233088001-233100121 Homo sapiens chromosome 2, complete sequence
☐ Check to include this hit in error tolerant search or archive report

MASCOT : Sequence Databases

© 2007-2010 Matrix Science

MATRIX SCIENCE

If you click on an accession number link, for a protein view report, you can get either the standard protein view report or an alternative

GenBank format feature table - Microsoft Internet Explorer

Address: http://koala/mascot/cgi/protein_view.pl?file=../data/20060120/F002128.dat&hit=chr2_194238px=1&protscore=954.4567793343478_mudpit=10008_featuretablelength=10000

```

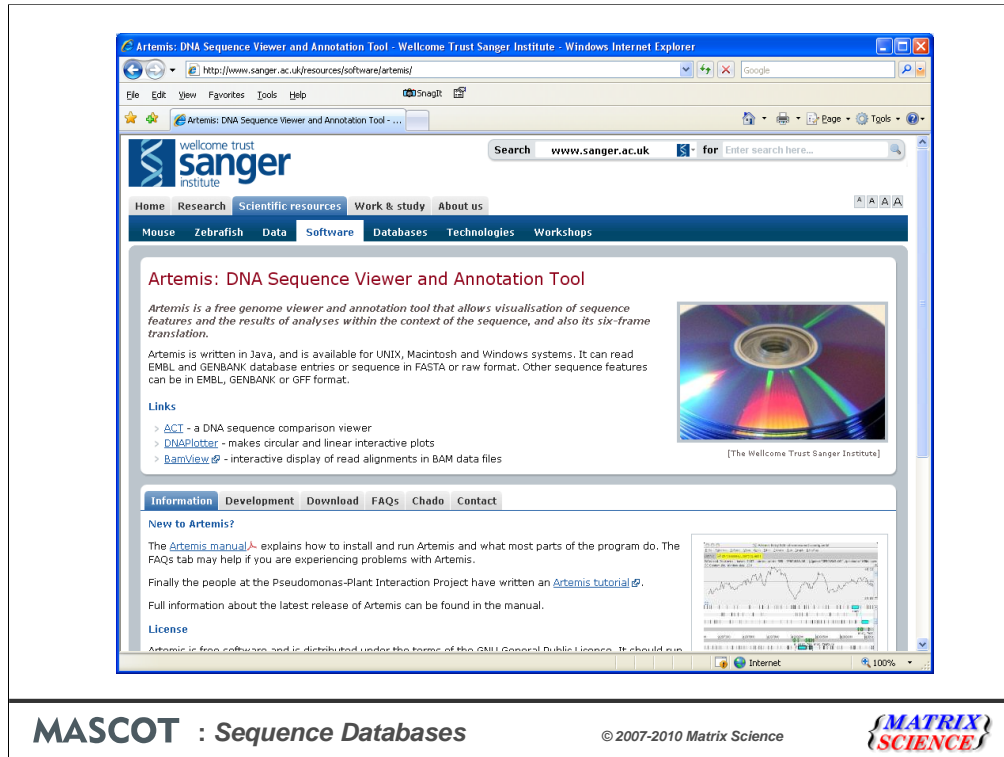
BLASTCDS      5797..5835
               /label=Q111
               /colour=2
               /note="Mascot match, query=111, mass=1304.68, score=69, rank=1, sequence=GNFQTIGLSAAAR"
               /blastp_file="../data/20060120/F002128.dat"
               /mass=1304.68
               /score=69
               /rank=1
               /translation="GNFQTIGLSAAAR"
BLASTCDS      6059..6097
               /label=Q137
               /colour=2
               /note="Mascot match, query=137, mass=1450.65, score=88, rank=1, sequence=NVYSDADVPAASAR"
               /blastp_file="../data/20060120/F002128.dat"
               /mass=1450.65
               /score=88
               /rank=1
               /translation="NVYSDADVPAASAR"
BLASTCDS      7099..7140
               /label=Q162
               /colour=2
               /note="Mascot match, query=162, mass=1639.78, score=103, rank=1, sequence=ALTETINFDDAIER"
               /blastp_file="../data/20060120/F002128.dat"
               /mass=1639.78
               /score=103
               /rank=1
               /translation="ALTETINFDDAIER"
BLASTCDS      6011..6058
               /label=Q177
               /colour=2
               /note="Mascot match, query=177, mass=1707.84, score=71, rank=1, sequence=VQHASPAGTYAHTVNR"
               /blastp_file="../data/20060120/F002128.dat"
               /mass=1707.84
               /score=71
               /rank=1
               /translation="VQHASPAGTYAHTVNR"

```

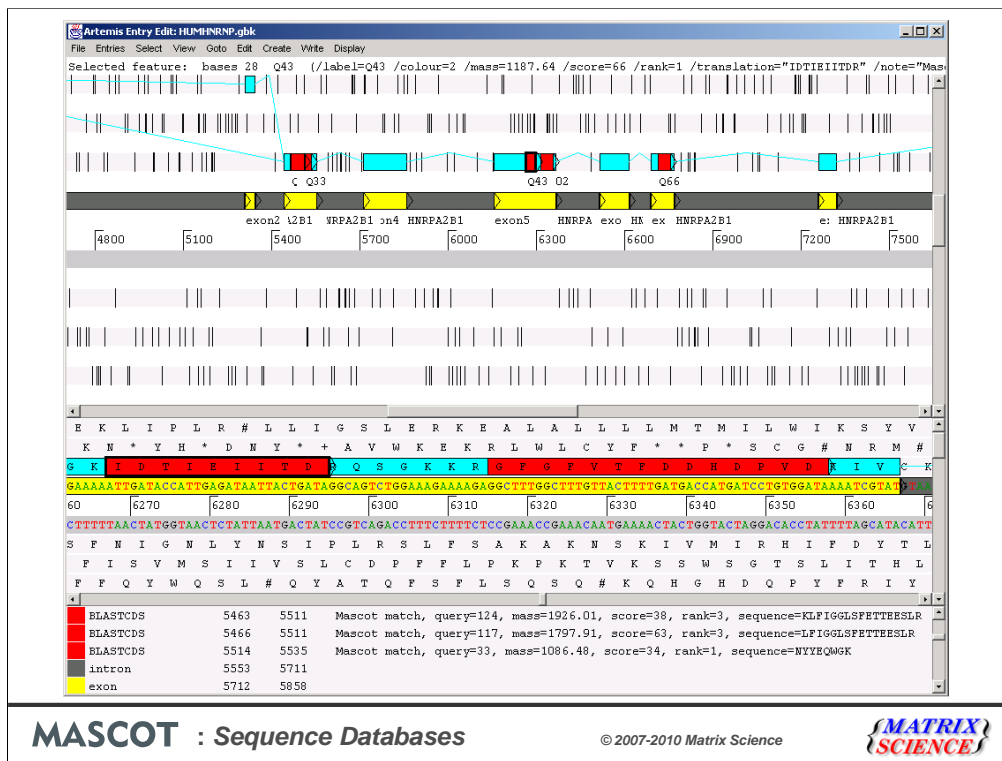
Done Local intranet

MASCOT : Sequence Databases © 2007-2010 Matrix Science **MATRIX SCIENCE**

This is the peptide match results formatted as an EMBL / GenBank format feature table. This may not look very friendly, but the advantage of this report is that it can be read into a standard genome browser



For example, one which we find works well is Artemis, a Java based genome browser developed and distributed by the Sanger Centre.



Here's the result of reading the feature table containing the Mascot peptide matches into Artemis. In the upper third, we have a low resolution view. This can be zoomed out to show an entire sequence as a single strip. We have the forward and complementary DNA strands, and the 6 frame translation. The vertical bars are stop codons. The yellow blocks are exons, while the blue blocks here are coding sequences. Individual Mascot peptide matches are shown in red. This particular gene has 8 peptide matches.

The middle third is a similar arrangement, but at high enough resolution to see individual bases and residues.

Finally, the lower third shows a tabular view of the feature table. When a match is selected, it is highlighted in all three views, and we can see the spectrum number, sequence, molecular weight, Mascot score, etc.

Not only does this allow us to zoom and pan around these extremely long sequences, it also allows us to view the peptide matches found by Mascot in the context of all the existing annotations. This gives us a powerful way to present the results of MS based searching complete genomes.

Human IPI vs. EST vs. Genome

Type of search : MS/MS Ion Search
Enzyme : Trypsin/P
Fixed modifications : [☑Carbamidomethyl \(C\)](#)
Variable modifications : [☑Acetyl \(N-term\)](#), [☑Phospho \(Y\)](#), [☑Phospho \(ST\)](#), [☑Oxidation \(M\)](#), [☑Gln->pyro-Glu \(N-term Q\)](#)
Mass values : Monoisotopic
Protein mass : Unrestricted
Peptide mass tolerance : ± 10 ppm ($\#^{13}\text{C} = 1$)
Fragment mass tolerance : ± 0.6 Da
Max missed cleavages : 1
Instrument type : ESI-TRAP
Number of queries : 8,797

Database	Size	Avg. 1% threshold	# matches @ 1% FDR
IPI_human 3.66	3.5 x 10 ⁷ residues	36	2961
EST_human 20100415	4.2 x 10 ⁹ bases	60	1899
Human Genome 20060306	3.1 x 10 ⁹ bases	60	1241

MASCOT : Sequence Databases

© 2007-2010 Matrix Science



All well and good, but which database gives the most matches? We searched a much larger dataset against all 3 databases. The data was the public iPRG2010 dataset distributed by ABRF.

There is a big drop in the number of matches between IPI_human and EST_human. The reason is mainly that EST_human is a much bigger database, by more than a factor of 100. This means that the score thresholds are approx 24 higher, and we lose all the weaker matches, that had scores between 36 and 60. Yes, there may be additional matches in EST, not found in IPI, but the net change is highly negative.

You can see at a glance that the human genome is even worse. This is not because of a still higher threshold; the database is very similar in size to EST_human. One reason is that a proportion of potential matches are missed because they are split across exon-intron boundaries. Based on average peptide length, approx 20% of matches would be lost for this reason. In this particular example, the difference is much larger than 20%. The other factor is that the human genome is only 1.5% coding sequence, and represents a single consensus genome. EST is 100% coding sequence and represents a wide range of SNPs and variants.

Human IPI vs. EST vs. Genome

- Searching complete chromosomes is possible, but unwieldy.
- Scoring statistics for assembled genome very similar to EST_human, but
 - the genome is a single consensus sequence, EST_human represents many variants
 - EST_human is 100% coding, HG assembly is 1.5% coding
 - lose approx 20% of matches because they straddle an exon - intron boundary
- In general, EST_human is a better choice
- References
 - Choudhary, J. S., Blackstock, W. P., Creasy, D. M. and Cottrell, J. S. (2001). Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics*, 1, 651-667.
 - Choudhary, J. S., Blackstock, W. P., Creasy, D. M. and Cottrell, J. S. (2001). Matching peptide mass spectra to EST and genomic DNA databases. *Trends in Biotechnology*, 19, S17-S22.

So, these are our conclusions for the human genome, and the same considerations probably hold for other large mammalian genomes.

Plant and bacterial genomes are a different matter. If the species is not well represented in the protein databases, there is a much stronger need to search EST or genomic databases