

Searching Nucleic Acid Databases

MASCOT

{MATRIX}
{SCIENCE}

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
- Matches to all 6 frames treated as part of the same entry
- 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

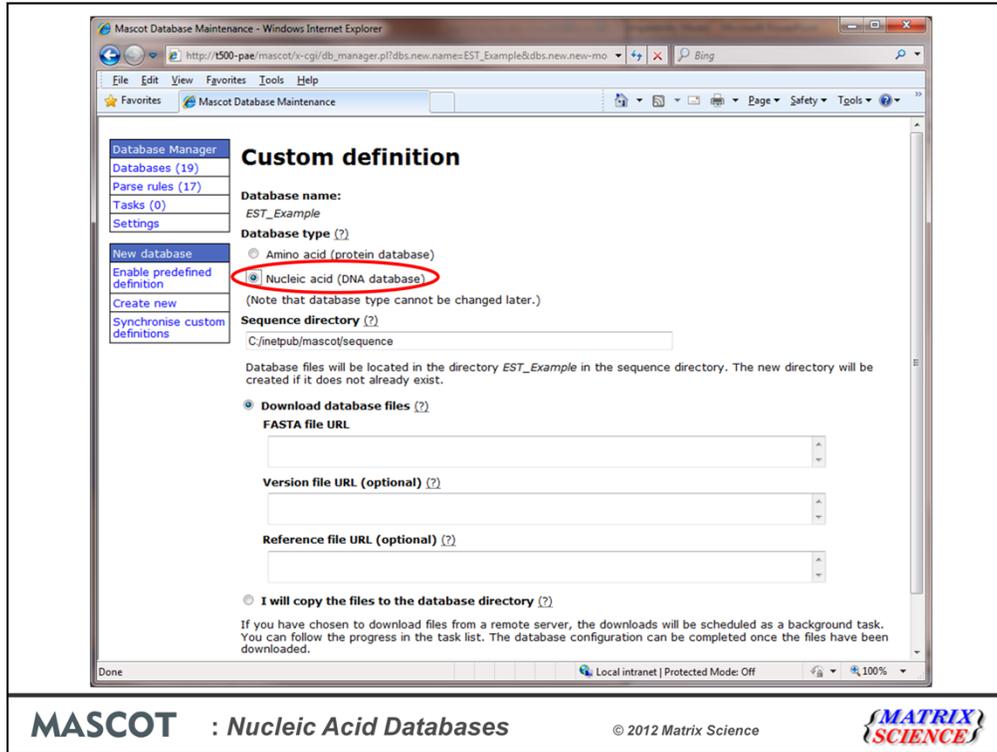
When a stop codon is encountered, leave a gap, and immediately re-start translation

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.

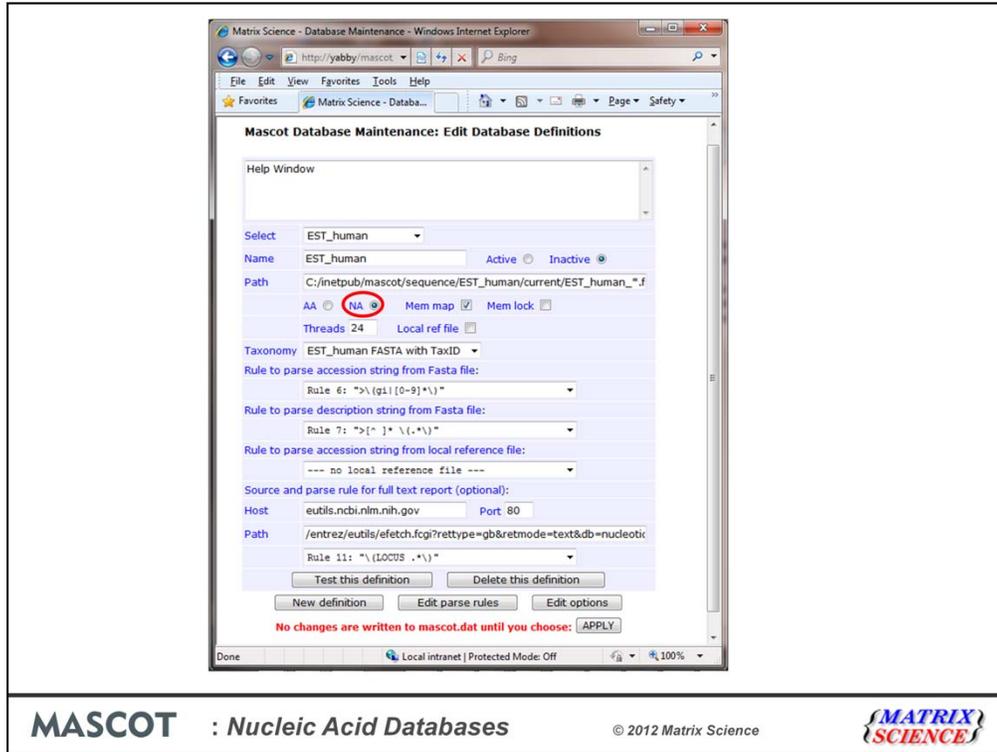
However, this is not done in Mascot.

It is better to allow Mascot to search the nucleic acid database directly rather than searching a pre-translated database because Mascot will treat matches from all 6 frames as part of the same entry, so matches will remain grouped together even if there is a frame shift.

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



Setting up a nucleic acid database in Mascot is no different from setting up a protein database. The only thing to watch is that the database type is specified as NA. This is where you do it in Mascot 2.4 in the new Database Manager when you are setting up a custom definition.



And this is where you do this in Mascot 2.3 and earlier on the database maintenance page.

MASCOT : Nucleic Acid Databases

© 2012 Matrix Science

MATRIX
SCIENCE

Select Summary Report (PRG2008 UniProt Mouse) - Mozilla Firefox

Select Summary Report (PRG2008 UniPr...

logon/mascot 2.4.0.64/cgi/master_results.pl?file=%2Fdata%2F20120430%2F001431.dat&_querylists=all&SEPTYPE=select16_s...

Most Visited Getting Started Latest Headlines Google Overview (Java 2 Platf... The World Clock - T... Overview (Java Platfor... Matrix Science - Home Matrix Science - Datab...

MATRIX
SCIENCE Mascot Search Results

User : PAE
 Email :
 Search title : iPRG2008 UniProt Mouse
 MS data file : D:\iPRG2008\mgf\merged.mgf
 Database : UniProt_mouse mouse_20120418 (55190 sequences; 26879425 residues)
 Timestamp : 30 Apr 2012 at 10:33:41 GMT
 Enzyme : Trypsin/P
 Fixed modifications : iTRAQ4plex (K), iTRAQ4plex (N-term), Methylthio (C)
 Variable modifications : Acetyl (Protein N-term), Gln->pyro-Gln (N-term Q), Oxidation (M)
 Mass values : Monoisotopic
 Protein Mass : Unrestricted
 Peptide Mass Tolerance : ± 0.9 Da
 Fragment Mass Tolerance : ± 0.6 Da
 Max Missed Cleavages : 1
 Instrument type : ESI-TRAP
 Number of queries : 33191
 Protein hits :

F20029	78 kDa glucose-regulated protein OS=Mus musculus GN=Hsp95 FE=1 SV=3
Q38550	Cytochrome b-5, isoform cB.2 OS=Mus musculus GN=Cyb5 FE=3 SV=1
Q64458	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 FE=1 SV=2
F09103	Protein disulfide-isomerase OS=Mus musculus GN=P4hb FE=1 SV=2
F00186	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 FE=1 SV=1
Q88451	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 FE=2 SV=1
F08113	Endoplasmic reticulum chaperone protein OS=Mus musculus GN=Hsp90b1 FE=1 SV=2
D3YV60	Uncharacterized protein OS=Mus musculus GN=Mgac1 FE=4 SV=1
D3YX74	Uncharacterized protein OS=Mus musculus GN=Om5619 FE=4 SV=1
F56593	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 FE=1 SV=2
F47963	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 FE=2 SV=3
E9Q9F8	Protein Rdn9 OS=Mus musculus GN=Rdn9 FE=3 SV=1
E9Q9T0	Uncharacterized protein OS=Mus musculus GN=Om5730 FE=3 SV=1
Q63880	Carboxylesterase 3A OS=Mus musculus GN=Ces3a FE=1 SV=2
D3Z041	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acl1 FE=4 SV=1
Q64459	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 FE=1 SV=1
Q10899	Protein Ugt2b5 OS=Mus musculus GN=Ugt2b5 FE=2 SV=1
Q9D8E6	60S ribosomal protein L4 OS=Mus musculus GN=Rpl4 FE=1 SV=3
F27773	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 FE=1 SV=2
Q9WV74	Carboxylesterase 1D OS=Mus musculus GN=Ces1d FE=1 SV=1

To illustrate the features of the different types of database, we searched the public iPRG2008 dataset distributed by ABRF against a protein database, UniProt mouse. We found significant matches to 438 mouse proteins.

MASCOT : Nucleic Acid Databases © 2012 Matrix Science **MATRIX SCIENCE**

Select Summary Report (PRG2008 EST_mouse) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Select Summary Report (PRG2008 EST_...

bogon/mascot_2_4_0_64/cgi/master_results.pl?file...%2Fdata%2F20120430%2F001432.dat&querylists=all&REPTYPE=select&sig...

Most Visited Getting Started Latest Headlines Google Overview (Java 2 Platf... The World Clock - T... Overview (Java Platfor... Matrix Science - Home Matrix Science - Datab...

MATRIX SCIENCE Mascot Search Results

User : PAE
 Email :
 Search title : 1PRG2008 EST_mouse
 MS data file : D:\1PRG2008\mgf\merged.mgf
 Database : EST_mouse mouse_20120429 (29121420 sequences: 4480529140 residues)
 Timestamp : 30 Apr 2012 at 12:00:29 GMT
 Enzyme : Trypsin/P
 Fixed modifications : iTRAQ4plex (K), iTRAQ4plex (N-term), Methylthio (C)
 Variable modifications : Acetyl (Protein N-term), Gln->pyro-Gln (N-term Q), Oxidation (M)
 Mass values : Monoisotopic
 Protein Mass : Unrestricted
 Peptide Mass Tolerance : ± 0.9 Da
 Fragment Mass Tolerance : ± 0.6 Da
 Max Missed Cleavages : 1
 Instrument type : ESI-TRAP
 Number of queries : 33191
 Protein hits :

gi11288567	ms24602.r1 Soares mouse p3NMF19.5 Mus musculus cDNA clone IMAGE:330363 5' similar to gb:NM22865 CYTOCHROME B5 (HUMAN);
gi11542754	mg16cl2.r1 Soares mouse embryo NmME13.5 14.5 Mus musculus cDNA clone IMAGE:423958 5' similar to gb:NM22865 CYTOCHROME
gi11284419	ms42b07.r1 Soares mouse p3NMF19.5 Mus musculus cDNA clone IMAGE:313333 5' similar to gb:J03746 GLUTATHIONE S-TRANSFER
gi126139681	BY034238 RIKEN full-length enriched, 14 days embryo liver Mus musculus cDNA clone I530010L01 5', mRNA sequence
gi129516299	AGENCOURT_12748949 NIH_MGC_178 Mus musculus cDNA clone IMAGE:30297876 5', mRNA sequence
gi12455723	VM94702.r1 Knowles Solter mouse blastocyst B1 Mus musculus cDNA clone IMAGE:1005919 5' similar to SW:R14A_YEAST P861C
gi114605269	602910379F1 NCI_CGAP_L19 Mus musculus cDNA clone IMAGE:5051706 5', mRNA sequence
gi114215323	602799210F1 NCI_CGAP_Mam4 Mus musculus cDNA clone IMAGE:4934715 5', mRNA sequence
gi11564576	ms10d12.r1 StraCapene mouse diaphragm (4937303) Mus musculus cDNA clone IMAGE:521111 5' similar to gb:NM61855 CYTOCHR
gi130198940	AGENCOURT_13680746 NIH_MGC_177 Mus musculus cDNA clone IMAGE:30309176 5', mRNA sequence
gi16938985	uq51d04.x1 Supano mouse liver mlia Mus musculus cDNA clone IMAGE:2921959 3' similar to gb:J04449 CYTOCHROME P450 I1IF
gi11309390	mc19f08.r1 Soares mouse p3NMF19.5 Mus musculus cDNA clone IMAGE:348999 5' similar to gb:NM146689_cds1 Mouse surfeit loc
gi156884421	HBM03503 Mus Musculus hematopoietic BM-HPC5 cDNA library Mus musculus cDNA 5', mRNA sequence
gi138167193	A0312A06-5 NIA House Trophoblast Stem Cell cDNA Library (Long 1) Mus musculus cDNA clone NIA:A0312A06 IMAGE:30734309
gi112462586	602335095F2 NCI_CGAP_Mam1 Mus musculus cDNA clone IMAGE:4458430 5', mRNA sequence
gi150073024	BP753136 mouse (C57BL/6) pancreatic islet library with recombination-based method Mus musculus cDNA clone msa01045 5'
gi1217512017	AGENCOURT_11293707 NIH_MGC_164 Mus musculus cDNA clone IMAGE:30147226 5', mRNA sequence
gi114432448	ui65d09.y1 Supano mouse liver mlia Mus musculus cDNA clone IMAGE:1887281 5' similar to gb:NM55053 CYTOCHROME P450 IA2
gi114300393	602821413F1 NCI_CGAP_Mam6 Mus musculus cDNA clone IMAGE:4950242 5', mRNA sequence
gi176384643	MONTH14_03_R04.v1.PK MONTH14 Mus musculus cDNA clone MONTH14_03_R04 similar to RefSeq:3095 protein 5 (mammalian)

With EST_mouse, we obtained almost the same results, just a couple of additional peptide matches. However, look at the hit-list on this report ... we have a very long match list because there is a large amount of redundancy in the EST databases and unlike the protein database search, most of the matched entries don't immediately communicate which proteins have been found. I'll return to these issues later.

Select Summary Report (PRG2008_EST_mouse) - Mozilla Firefox

Significance threshold $p < 0.05$ Max. number of hits AUTO Show Percolator scores

Standard scoring MudPIT scoring Ions score or expect cut-off 0.5 Show sub-sets 0

Show pop-ups Suppress pop-ups Require bold red

Re-Search All queries Unassigned Below homology threshold Below identity threshold

1. [gi13288567](#) Mass: 18637 Score: 631 Matches: 19(15) Sequences: 4(4) eMPAI: 1.96 Frame: 1
 mb24f02.r1 Soares mouse p3NMF19.5 Mus musculus cDNA clone IMAGE:330363 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN); mRNA sequence

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
18020	738.0084	1474.0022	1473.7959	0.2064	0	74	0.0018	1	1	K.YVTLREIQK.H 18018 18021 18023 18025
20552	787.0462	1572.0778	1571.8065	0.2712	0	76	0.0011	1	1	K.TVIIGELHFDOR.S
20563	525.0861	1572.2366	1571.8065	0.4301	0	(57)	0.077	1	1	K.TVIIGELHFDOR.S 20551
22386	828.5314	1655.0483	1654.8437	0.2047	0	62	0.023	1	1	K.FLEHFQGEVLR.E
30413	1175.6897	2349.3647	2349.0227	0.3420	0	(100)	1.9e-06	1	1	R.EQAQGDATENFEDVQHSYDAR.E 30427
30415	784.1341	2349.3805	2349.0227	0.3578	0	104	6.4e-07	1	1	R.EQAQGDATENFEDVQHSYDAR.E 30412 30417 30423 30426 30428 30429 30430

Proteins matching the same set of peptides:

[gi13369609](#) Mass: 18789 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 3
 md85e07.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone IMAGE:375204 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN); mRNA sequence

[gi11497280](#) Mass: 20396 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 1
 ml25d08.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone IMAGE:464559 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN); mRNA sequence

[gi13509570](#) Mass: 18741 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 2
 ml54q13.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone IMAGE:467399 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN); mRNA sequence

[gi11520042](#) Mass: 20179 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 3
 ml54q13.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone IMAGE:467399 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN); mRNA sequence

[gi11531387](#) Mass: 20628 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 1
 ml54b07.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone IMAGE:479965 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN); mRNA sequence

[gi11539480](#) Mass: 20238 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 1
 me95f06.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone IMAGE:403331 5' similar to gb:M22865 CYTOCHROME B5 (HUMAN); mRNA sequence

[gi13564396](#) Mass: 18664 Score: 631 Matches: 19(15) Sequences: 4(4) Frame: 1

MASCOT : Nucleic Acid Databases © 2012 Matrix Science **MATRIX SCIENCE**

The master results report from the EST search looks pretty similar to the Uniprot Mouse 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: gi|9156892
 601099741F1 NCL_CGAP_Lu29 Mus musculus cDNA clone IMAGE:3492011 5', mRNA sequence

Database: EST_mouse
 Score: 323
 Nominal mass (M_r): 22669
 Calculated pI: 10.75
 Frame: 3
 Taxonomy: [Mus musculus](#)

NB: Matches were also found in other frames indicating a possible frame shift. Only matches in frame 3 are shown in this report.

Show frame: 3

Sequence similarity is available as an [NCBI BLAST search of gi|9156892 against nr](#).

Search parameters

MS data file: D:\FRQ2008\mgf\merged.mgf
 Enzyme: Trypsin/P; cuts C-term side of KR.
 Fixed modifications: [ITRAQ4plex \(K\)](#), [ITRAQ4plex \(N-term\)](#), [Methylthio \(C\)](#)
 Variable modifications: [Acetyl \(Protein N-term\)](#), [Gln->pyro-Glu \(N-term Q\)](#), [Oxidation \(M\)](#)

Protein sequence coverage: 26%

Matched peptides shown in **bold red**.

```

1 KAKGRKVA PAVVDSQEAK KVVNPLFER FKHFGQDI QPERDLRFV
51 KMFYIQLQR QRALLYKRLK VVFAINQFTQ ALDRQTATQL LKLAHKYRFE
101 TKQEKRHKAG WAPLLEKEL LAKATSQQLD HLSSEQESIQ SPFWWRTRRL
151 SWW_LPHM_T PIELGGFS
  
```

Unformatted sequence string: [168 residues](#) (for pasting into other applications).

MASCOT : Nucleic Acid Databases © 2012 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 3.

But, as so often happens, there is a frame shift in this entry, and there is an additional match in frame 2.

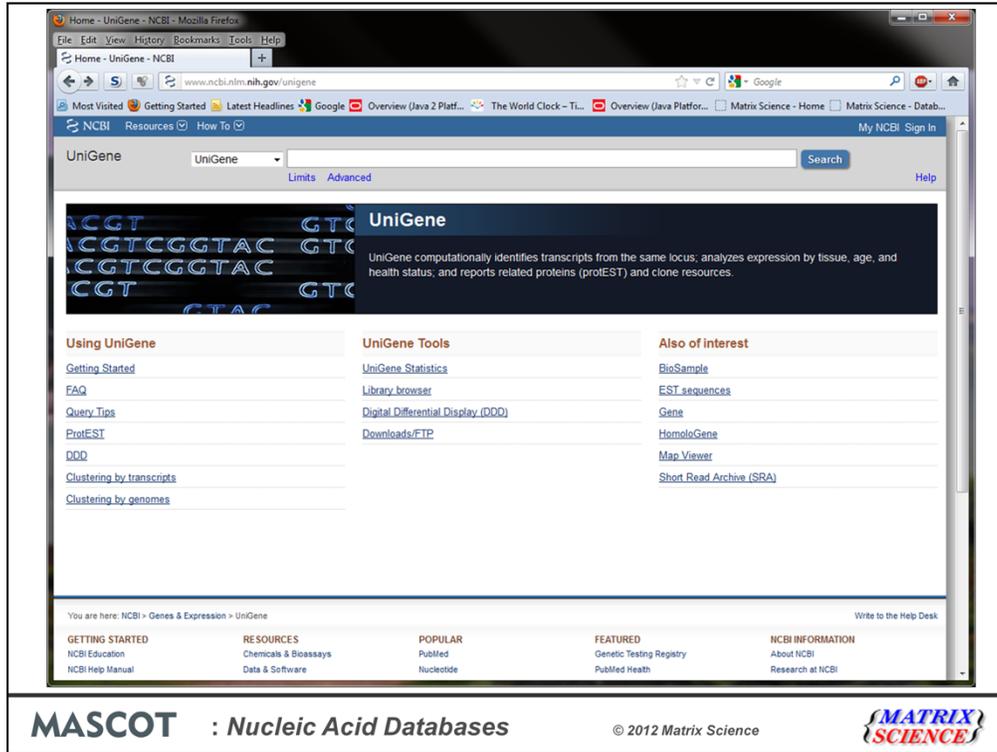
Mascot Search Results

User : PAE
 Search title : 1PRG2008 EST_mouse
 MS data file : D:\1PRG2008\mgf\merged.mgf
 Database : EST_mouse mouse_20120429 (29121420 sequences: 4480529140 residues)
 Timestamp : 30 Apr 2012 at 12:00:29 GMT
 Enzyme : Trypsin/P
 Fixed modifications : iTRAQ4plex (K), iTRAQ4plex (N-term), Methylthio (C)
 Variable modifications : Acetyl (Protein N-term), Gln->pyro-Gln (N-term Q), Oxidation
 Mass values : Monoisotopic
 Protein Mass : Unrestricted
 Peptide Mass Tolerance : ± 0.9 Da
 Fragment Mass Tolerance : ± 0.6 Da
 Max Missed Cleavages : 1
 Instrument type : ESI-TRAP
 Number of queries : 33191
 Protein hits :
 gi|11288567|mb24602.r1|Soares mouse p3NMF19.5 Mus musculus
 gi|11542754|mg16cl2.r1|Soares mouse embryo NME13.5 14.5
 gi|11284419|ma42b07.r1|Soares mouse p3NMF19.5 Mus musculus
 gi|126139681|BV034238|RIKEN full-length enriched, 14 days embryo liver Mus musculus cDNA clone B330210L01.5', mRNA sequence
 gi|129516299|AGENCOURT_12748945|NIH_MGC_179 Mus musculus cDNA clone IMAGE:130297876 5', mRNA sequence
 gi|12455723|VNS4202.r1|Knobler Solter mouse blastocyst B1 Mus musculus cDNA clone IMAGE:1005919 5' similar to SW:R14A_YEAST P361C
 gi|114605269|602910379F1|NCI_CGAP_L19 Mus musculus cDNA clone IMAGE:5051706 5', mRNA sequence
 gi|114215323|602799210F1|NCI_CGAP_Mam4 Mus musculus cDNA clone IMAGE:4934715 5', mRNA sequence
 gi|11564576|m10d12.r1|StraCepene mouse diaphragm (4937303) Mus musculus cDNA clone IMAGE:521111 5' similar to gb:M61855 CYTOCHROME
 gi|130199940|AGENCOURT_13680746|NIH_MGC_177 Mus musculus cDNA clone IMAGE:30309176 5', mRNA sequence
 gi|6938985|uq51d04.x1|Supano mouse liver mlia Mus musculus cDNA clone IMAGE:2921959 3' similar to gb:J04449 CYTOCHROME P450 I1IP
 gi|11309390|mc19f08.r1|Soares mouse p3NMF19.5 Mus musculus cDNA clone IMAGE:348999 5' similar to gb:M14689_cds1 Mouse surfeit loc
 gi|156894421|HBM03503|Mus Musculus hematopoietic BM-BMPC5 cDNA Library Mus musculus cDNA 5', mRNA sequence
 gi|138167193|A0312A06-5|NIA House Trophoblast Stem Cell cDNA Library (Long 1) Mus musculus cDNA clone NIA:A0312A06 IMAGE:30734309
 gi|112462586|602335095F2|NCI_CGAP_Mam1 Mus musculus cDNA clone IMAGE:4458430 5', mRNA sequence
 gi|150073024|BP753136|mouse (C57BL/6) pancreatic islet library with recombination-based method Mus musculus cDNA clone mia01045 5'
 gi|1217512017|AGENCOURT_11293707|NIH_MGC_164 Mus musculus cDNA clone IMAGE:30147226 5', mRNA sequence
 gi|14443248|u165d09.y1|Supano mouse liver mlia Mus musculus cDNA clone IMAGE:1887281 5' similar to gb:M55053 CYTOCHROME P450 IA2
 gi|114300393|602821413F1|NCI_CGAP_Mam6 Mus musculus cDNA clone IMAGE:4950242 5', mRNA sequence
 gi|176384643|MONTH14_03_B08.v1.PK|MONTH14 Mus musculus cDNA clone MONTH14_03_B08 similar to RefSeq:3049 protein 5' (mammalian)

MASCOT : Nucleic Acid Databases © 2012 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 is not a sequence database, it is an index which is created by BLASTing GenBank sequences against themselves to cluster them into gene families.

UniGene can be downloaded from the NCBI FTP site. In Mascot 2.4, the database manager automatically downloads and configures unigene indexes for pre-defined databases such as EST Human.

For Mascot 2.3 and earlier, downloading can be automated using the db_update.pl script. Some manual configuration of Mascot.dat is then required in order to enable the indexes.

Mascot Search Results

User : PAE
 Email :
 Search title : 1PRG2008 EST_mouse
 MS data file : D:\1PRG2008\mgf\merged.mgf
 Database : EST_mouse mouse_20120429 (29121420 sequences: 4480529140 residues)
 Timestamp : 30 Apr 2012 at 12:00:29 GMT
 Enzyme : Trypsin/P
 Fixed modifications : iTRAQ4plex (K), iTRAQ4plex (N-term), Methylthio (C)
 Variable modifications : Acetyl (Protein N-term), Glu-pyro-Glu (N-term Q), Oxidation
 Mass values : Monoisotopic
 Protein Mass : Unrestricted
 Peptide Mass Tolerance : ± 0.9 Da
 Fragment Mass Tolerance : ± 0.6 Da
 Max Missed Cleavages : 1
 Instrument type : ESI-TRAP
 Number of queries : 33191
 Protein hits :
 gi|11288567|mb24602.r1|Soares mouse p3NMF19.5 Mus musculus
 gi|11542754|mg16cl2.r1|Soares mouse embryo NME13.5 14.5
 gi|11284419|ma42b07.r1|Soares mouse p3NMF19.5 Mus musculus
 gi|126139681|BY034238|RIKEN full-length enriched, 14 days embryo liver Mus musculus cDNA clone B337012UL1 5', mRNA sequence
 gi|129516299|AGENCOURT_12748945|NIH_MGC_178 Mus musculus cDNA clone IMAGE:30297876 5', mRNA sequence
 gi|12455729|VNS4Z02.i|Knovles Solter Mouse blastocyst B1 Mus musculus cDNA clone IMAGE:1005915 5' similar to SW:R14A_YEAST P861C
 gi|114605269|602910379F1|NCI_CGAP_L19 Mus musculus cDNA clone IMAGE:5051706 5', mRNA sequence
 gi|114215323|602799210F1|NCI_CGAP_Mam4 Mus musculus cDNA clone IMAGE:4934715 5', mRNA sequence
 gi|11564576|ml10d12.r1|StraCape mouse diaphragm (4937303) Mus musculus cDNA clone IMAGE:521111 5' similar to gb:M61855 CYTOCHROME
 gi|130198940|AGENCOURT_13680746|NIH_MGC_177 Mus musculus cDNA clone IMAGE:30309176 5', mRNA sequence
 gi|16938985|uq51d04.x1|Supano mouse liver mlia Mus musculus cDNA clone IMAGE:2921959 3' similar to gb:J04449 CYTOCHROME P450 I11P
 gi|11309390|mc19F08.r1|Soares mouse p3NMF19.5 Mus musculus cDNA clone IMAGE:348999 5' similar to gb:M14689_cds1 Mouse surfeit loc
 gi|156884421|HBM03503|Mus Musculus hematopoietic BM-HPCS cDNA library Mus musculus cDNA 5', mRNA sequence
 gi|138167193|A0312A06-5|NIA Mouse Trophoblast Stem Cell cDNA Library (Long 1) Mus musculus cDNA clone NIA:A0312A06 IMAGE:30734309
 gi|112482586|602335095F2|NCI_CGAP_Mam1 Mus musculus cDNA clone IMAGE:4458430 5', mRNA sequence
 gi|150073024|BP753136|mouse (C57BL/6) pancreatic islet library with recombination-based method Mus musculus cDNA clone mia01045 5'
 gi|127512017|AGENCOURT_11293707|NIH_MGC_164 Mus musculus cDNA clone IMAGE:30147226 5', mRNA sequence
 gi|114432448|ui65409.y1|Supano mouse liver mlia Mus musculus cDNA clone IMAGE:1887281 5' similar to gb:M55053 CYTOCHROME P450 I2A2
 gi|114300393|602821413F1|NCI_CGAP_Mam6 Mus musculus cDNA clone IMAGE:4950242 5', mRNA sequence
 ml176384643|MONTH14_03_R08.v1.FH|MONTH14 Mus musculus cDNA clone MONTH14_03_R08 similar to RefSeq:3095 protein 5' (mammalian)

Format Controls

UniGene index: None
 Max. number: Mus musculus
 Ions score or expect cut-off: 0
 Sort unassigned: Decreasing Score

MASCOT : Nucleic Acid Databases © 2012 Matrix Science **MATRIX SCIENCE**

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

MASCOT : Nucleic Acid Databases © 2012 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 : Nucleic Acid Databases © 2012 Matrix Science **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 – hits 13 and 120. They don't have any shared peptide matches, 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 Prolyl 4-hydroxylase.

MASCOT Search Results

UniGene View

ID [Nm_1](#)

TITLE S100 calcium binding protein A10 (calpactin)

GENE S100a10

CYTOBAND 3 F1-F2|3 41.7 cM

GENE_ID 20194

LOCUSLINK 20194

HOMOL YES

EXPRES adipose tissue| blood| bone marrow| brain| connective tissue| dorsal root ganglion| embryonic tissue| extraembryonic tissue| eye| heart| inner

CHROMOSOME 3

STS ACC=RM125510 UNISTS=162328

STS ACC=M16465 UNISTS=178878

STS ACC=RM14905 UNISTS=161130

STS ACC=RM128467 UNISTS=211775

STS ACC=SI00a10 UNISTS=480307

STS ACC=SI00a10 UNISTS=463493

STS ACC=SI00a10 UNISTS=507204

PROTSIM ORG=10090; PROTI=4677833; PROTI=MF_033138.1; FCT=100.00; ALN=95

PROTSIM ORG=28377; PROTI=32730010; PROTI=XP_003229717.1; FCT=91.43; ALN=94

PROTSIM ORG=9886; PROTI=126723603; PROTI=MF_001075632.1; FCT=93.81; ALN=95

PROTSIM ORG=9604; PROTI=4506761; PROTI=MF_002987.1; FCT=93.81; ALN=95

PROTSIM ORG=13614; PROTI=126313738; PROTI=XP_001366907.1; FCT=92.78; ALN=95

PROTSIM ORG=9615; PROTI=73981599; PROTI=XP_533060.2; FCT=91.30; ALN=95

PROTSIM ORG=9913; PROTI=2780715; PROTI=MF_77078.1; FCT=93.81; ALN=95

PROTSIM ORG=8355; PROTI=147901788; PROTI=MF_001087281.1; FCT=63.74; ALN=90

PROTSIM ORG=9258; PROTI=149405573; PROTI=MF_001513822.1; FCT=51.55; ALN=96

PROTSIM ORG=7955; PROTI=47088671; PROTI=MF_598168.1; FCT=52.00; ALN=98

PROTSIM ORG=9796; PROTI=255522873; PROTI=MF_001157339.1; FCT=93.81; ALN=95

PROTSIM ORG=9544; PROTI=76126457; PROTI=MF_001028124.1; FCT=93.81; ALN=95

PROTSIM ORG=8364; PROTI=58332722; PROTI=MF_001011436.1; FCT=65.22; ALN=90

PROTSIM ORG=7998; PROTI=318056044; PROTI=MF_001187971.1; FCT=51.00; ALN=99

PROTSIM ORG=0116; PROTI=3382079; PROTI=MF_112376.1; FCT=96.84; ALN=93

PROTSIM ORG=9031; PROTI=45382861; PROTI=MF_590837.1; FCT=88.66; ALN=95

PROTSIM ORG=9813; PROTI=31124503; PROTI=MF_00328806.1; FCT=93.81; ALN=95

PROTSIM ORG=9103; PROTI=324915895; PROTI=XP_003204247.1; FCT=88.66; ALN=95

PROTSIM ORG=8129; PROTI=348523989; PROTI=MF_003452494.1; FCT=53.12; ALN=94

PROTSIM ORG=9381; PROTI=332810316; PROTI=MF_3113807.1; FCT=93.81; ALN=95

SCOUNT 354

SEQUENCE ACC=C461462.1; HID=2481764; CLON=IMAGE:678474; END=5'; LID=12110; SEQTYPE=EST; TRACE=18140983

SEQUENCE ACC=C857516.1; HID=29498246; CLON=IMAGE:30295364; END=5'; LID=12739; SEQTYPE=EST; TRACE=196933136

SEQUENCE ACC=C8566164.1; HID=29485694; CLON=IMAGE:30294362; END=5'; LID=12615; SEQTYPE=EST; TRACE=196933979

SEQUENCE ACC=DV058483.1; HID=76380766; CLON=IMAGE:04_M0; LID=18145; SEQTYPE=EST

SEQUENCE ACC=DV060885.1; HID=76388183; CLON=IMAGE:04_M0; LID=18147; SEQTYPE=EST

SEQUENCE ACC=DV060885.1; HID=76388183; CLON=IMAGE:04_M0; LID=18147; SEQTYPE=EST

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

Genomic databases

- **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

We can also perform MS/MS searches on the raw genomic sequence data.

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

For example, 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.

Mus musculus (house mouse)

The laboratory mouse is a major model organism for basic mammalian biology, human disease, and genome evolution, and its genome has been sequenced.

Lineage: Eukaryote(4461), Metazoa(227), Chordata(24), Craniata(4), Vertebrata(23), Euteleostomi(23), Mammalia(11), Eutheria(10), Euarchontoglires(10), Glires(10), Rodentia(2), Sciurognathi(2), Muridae(1), Murinae(1), Mus(2), Mus(2), Mus musculus(1)

The mouse is one of the major organisms for modeling human disease and comparative genome analysis. There are over 450 inbred strains of mice, providing a wealth of different genotypes and phenotypes for genetic and other studies. In addition, thousands of spontaneous, radiation- or chemically-induced, and transgenic mutants provide potential models [here](#).

Organism Overview See also: [Genome List](#) [Organella List](#)

Chromosomes Click on chromosome name to see map/trace

Type	Count
RefSeq Genome	4
Clone ends	1
Epigenomics	34
Genome resequencing	10
Map	1
Other	123
Proteomes	10
Targeted Locus (L-loc)	1
Transcriptome or Gene expression	4002
Variation	69

Assembly and Annotation other assemblies are available

Assembly Name	EGSC(17)
Last sequence update	23-Feb-2012
Highest level of assembly	Chromosome
Size (total bases)	2,716,349,162
Number of genes	34,256
Number of proteins	27,851

Mitochondrial Genome

Last record update	21-Oct-2010
Last sequence update	09-Sep-2003
Size	16,299
Number of genes	37
Number of proteins	13

Related BioProjects

Publications

- Intragenic enhancers act as alternative promoters [Mol Cell 2012]
- High-efficiency of DrosT imparts leukemia stem cell function through dera [Genes Dev 2012]
- Deep-sequencing identification of the genomic targets of the cyclin-dependent kinase [Nat Genet 2011]

Search details

"Mus musculus" [Organism]

External Resources

- Broad Institute
- ODR Mouse Genotyping
- Cetera
- EU RMI map
- Ensembl Genome Browser

MASCOT : Nucleic Acid Databases © 2012 Matrix Science **MATRIX SCIENCE**

You can download the mouse genome sequences from the NCBI.

Index of ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/Assembled_chromosomes/seq/

Up to higher level directory

Name	Size	Last Modified
mm_assembly.Chromosome.chr1.fasta.gz	54125 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr2.fasta.gz	58146 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr3.fasta.gz	36381 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr4.fasta.gz	39939 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr5.fasta.gz	33171 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr6.fasta.gz	35484 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr7.fasta.gz	30983 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr8.fasta.gz	33021 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr9.fasta.gz	32543 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr10.fasta.gz	34834 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr11.fasta.gz	32372 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr12.fasta.gz	34629 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr13.fasta.gz	28878 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr14.fasta.gz	28866 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr15.fasta.gz	27144 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr16.fasta.gz	29073 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr17.fasta.gz	26029 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr18.fasta.gz	26627 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr19.fasta.gz	25064 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr20.fasta.gz	26812 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr21.fasta.gz	26862 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chrX.fasta.gz	17843 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chrY.fasta.gz	50402 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chrZ.fasta.gz	53003 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr1.fasta.gz	45109 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr2.fasta.gz	48290 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr3.fasta.gz	43276 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr4.fasta.gz	44800 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr5.fasta.gz	40350 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr6.fasta.gz	42019 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr7.fasta.gz	41488 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr8.fasta.gz	44372 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr9.fasta.gz	39993 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr10.fasta.gz	39575 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr11.fasta.gz	35530 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr12.fasta.gz	38033 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr13.fasta.gz	33966 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr14.fasta.gz	36270 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr15.fasta.gz	39000 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr16.fasta.gz	41776 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr17.fasta.gz	338 KB	09/03/2012 22:30:00
mm_assembly.Chromosome.chr18.fasta.gz	574 KB	09/03/2012 22:30:00

MASCOT : Nucleic Acid Databases © 2012 Matrix Science **MATRIX SCIENCE**

So for the mouse genome, we chose the assembled chromosomes - 21 files (using the GRCm38 reference assembly). Although you could search this as a 21 entry database, as I just mentioned this is not memory efficient, so we used the splitter script to split the chromosome sequences into overlapping segments of 12 kb

Select Summary Report (PRG2008 Mouse_genome) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Select Summary Report (PRG2008 Mouse_genome)

http://bogong/mascot_2_4_0_64/cgi/master_results.pl?file=%2Fdata%2F20120430%2F001430.dat&querylists=all&REPTYPE=select&sig

Most Visited Getting Started Latest Headlines Google Overview (Java 2 Platf... The World Clock - T... Overview (Java Platfor... Matrix Science - Home Matrix Science - Datab...

MASCOT Search Results

User : PAE
 Email :
 Search title : IPRG2008 Mouse genome
 MS data file : D:\IPRG2008\wgT\merged.mgf
 Database : Mouse_genome_genome_20120123 (1362828 sequences: 5505027542 residues)
 Timestamp : 30 Apr 2012 at 13:56:22 GMT
 Enzyme : Trypsin/P
 Fixed modifications : iTRAQ4plex (K), iTRAQ4plex (N-term), Methylthio (C)
 Variable modifications : Acetyl (Protein N-term), Gln-pyro-Gln (N-term Q), Oxidation (M)
 Mass values : Monoisotopic
 Protein Mass : Unrestricted
 Peptide Mass Tolerance : ± 0.9 Da
 Fragment Mass Tolerance : ± 0.6 Da
 Max Missed Cleavages : 1
 Instrument type : ESI-TRAP
 Number of queries : 33191
 Protein hits :

gi ref inc	base	description
gi 1372099108 ref inc 000068.71 2898	bases 34764001-34776120	Mus musculus strain C57BL/6J chromosome 2, GRCh38 C57BL/6J
gi 1372099104 ref inc 000072.61 11513	bases 138144001-138156120	Mus musculus strain C57BL/6J chromosome 6, GRCh38 C57BL/6J
gi 1372099092 ref inc 000084.61 7073	bases 84864001-84876120	Mus musculus strain C57BL/6J chromosome 18, GRCh38 C57BL/6J
gi 1372099101 ref inc 000075.61 4807	bases 57672001-57684121	Mus musculus strain C57BL/6J chromosome 9, GRCh38 C57BL/6J
gi 1372099103 ref inc 000073.61 3778	bases 45324001-45336121	Mus musculus strain C57BL/6J chromosome 7, GRCh38 C57BL/6J
gi 1372099102 ref inc 000074.61 8573	bases 102864001-102876122	Mus musculus strain C57BL/6J chromosome 8, GRCh38 C57BL/6J
gi 1372099100 ref inc 000076.61 10658	bases 127884001-127896122	Mus musculus strain C57BL/6J chromosome 10, GRCh38 C57BL/6J
gi 1372099101 ref inc 000075.61 6025	bases 72288001-72300121	Mus musculus strain C57BL/6J chromosome 9, GRCh38 C57BL/6J
gi 1372099099 ref inc 000077.61 10047	bases 120552001-120564120	Mus musculus strain C57BL/6J chromosome 11, GRCh38 C57BL/6J
gi 1372099098 ref inc 000078.61 6303	bases 75624001-75636121	Mus musculus strain C57BL/6J chromosome 12, GRCh38 C57BL/6J
gi 1372099105 ref inc 000071.61 12156	bases 145860001-145872121	Mus musculus strain C57BL/6J chromosome 5, GRCh38 C57BL/6J
gi 1372099102 ref inc 000074.61 8745	bases 104928001-104940122	Mus musculus strain C57BL/6J chromosome 8, GRCh38 C57BL/6J
gi 1372099102 ref inc 000074.61 8755	bases 105048001-105060122	Mus musculus strain C57BL/6J chromosome 8, GRCh38 C57BL/6J
gi 1372099102 ref inc 000074.61 8729	bases 104736001-104748122	Mus musculus strain C57BL/6J chromosome 8, GRCh38 C57BL/6J
gi 1372099102 ref inc 000074.61 8731	bases 104760001-104772122	Mus musculus strain C57BL/6J chromosome 8, GRCh38 C57BL/6J
gi 1372099103 ref inc 000073.61 2252	bases 27024001-27036121	Mus musculus strain C57BL/6J chromosome 7, GRCh38 C57BL/6J
gi 1372099105 ref inc 000071.61 7244	bases 86916001-86928121	Mus musculus strain C57BL/6J chromosome 5, GRCh38 C57BL/6J
gi 1372099097 ref inc 000079.61 1969	bases 23616001-23628121	Mus musculus strain C57BL/6J chromosome 13, GRCh38 C57BL/6J
gi 1372099102 ref inc 000074.61 7765	bases 9316001-93180122	Mus musculus strain C57BL/6J chromosome 8, GRCh38 C57BL/6J
gi 1372099095 ref inc 000081.61 6868	bases 82404001-82416121	Mus musculus strain C57BL/6J chromosome 15, GRCh38 C57BL/6J
gi 1372099103 ref inc 000073.61 11788	bases 145844001-145856121	Mus musculus strain C57BL/6J chromosome 7, GRCh38 C57BL/6J

MASCOT : Nucleic Acid Databases © 2012 Matrix Science **MATRIX SCIENCE**

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

MASCOT : Nucleic Acid Databases © 2012 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

```

BLASTCDS      complement(11200..11223)
               /label=Q97
               /colour=2
               /note="Mascot match, query=97, mass=1086.48, score=31, rank=4, sequence=NYYEQWGR"
               /blastp_file="../../data/20120501/F001462.dat"
               /mass=1086.48
               /score=31
               /rank=4
               /translation="NYYEQWGR"
BLASTCDS      complement(11200..11223)
               /label=Q98
               /colour=2
               /note="Mascot match, query=98, mass=1086.48, score=31, rank=4, sequence=NYYEQWGR"
               /blastp_file="../../data/20120501/F001462.dat"
               /mass=1086.48
               /score=31
               /rank=4
               /translation="NYYEQWGR"
BLASTCDS      complement(11200..11223)
               /label=Q99
               /colour=2
               /note="Mascot match, query=99, mass=1086.48, score=31, rank=4, sequence=NYYEQWGR"
               /blastp_file="../../data/20120501/F001462.dat"
               /mass=1086.48
               /score=31
               /rank=4
               /translation="NYYEQWGR"
BLASTCDS      complement(10565..10594)
               /label=Q127
               /colour=2
               /note="Mascot match, query=127, mass=1187.64, score=69, rank=1, sequence=IDTIEIITDR"
               /blastp_file="../../data/20120501/F001462.dat"
               /mass=1187.64
               /score=69
               /rank=1
               /translation="IDTIEIITDR"
BLASTCDS      complement(10565..10594)
               /label=Q128
               /colour=2
               /note="Mascot match, query=128, mass=1187.64, score=69, rank=1, sequence=IDTIEIITDR"

```

MASCOT : Nucleic Acid Databases © 2012 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

Artemis: Genome Browser and Annotation Tool

Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation.

Artemis is written in Java, and is available for UNIX, Macintosh and Windows systems. It can read EMBL and GENBANK database entries or sequence in FASTA, indexed FASTA or raw format. Other sequence features can be in EMBL, GENBANK or GFF format.

Links

- > [ACT](#) - a DNA sequence comparison viewer
- > [DNAPlotter](#) - makes circular and linear interactive plots
- > [BamView](#) - interactive display of read alignments in BAM data files

Information | Development | Download | FAQs | Chado | Courses | Contact

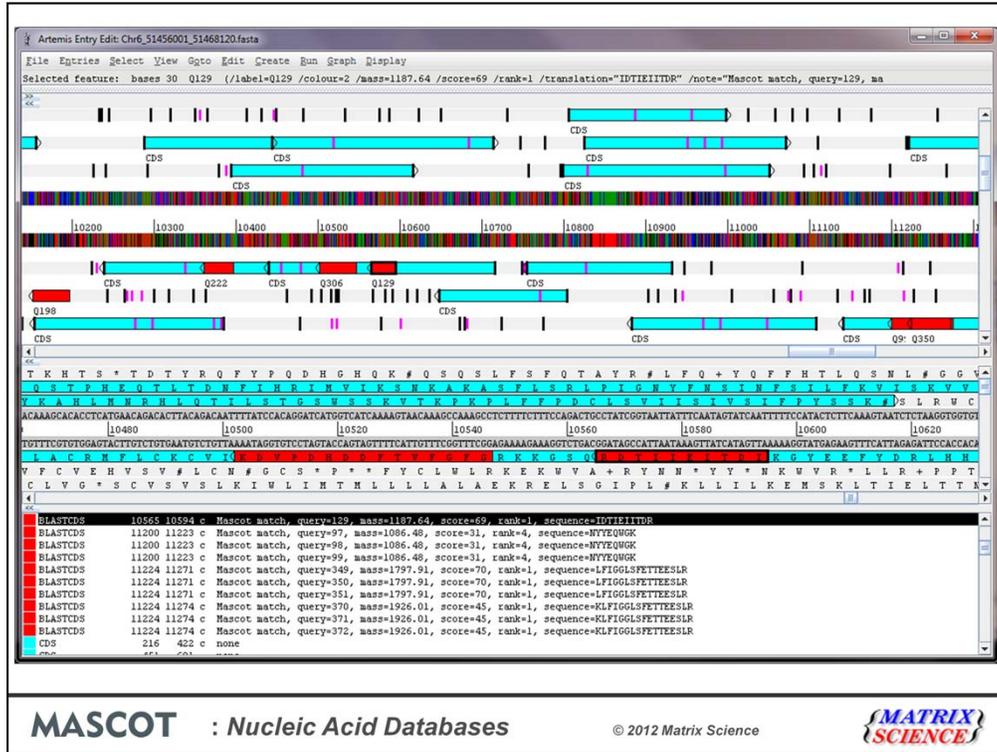
New to Artemis?

The [Artemis manual](#) explains how to install and run Artemis and what most parts of the program do. The FAQs may help if you are experiencing problems with Artemis. Also an [Artemis poster](#) gives an overview of browsing genomes and visualisation of next generation data in Artemis. There are also use case examples of [browsing next generation sequence data](#).

Full information about the latest release of Artemis can be found in the manual and the current [release notes](#).

MASCOT : Nucleic Acid Databases © 2012 Matrix Science {MATRIX SCIENCE}

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 black vertical bars are stop codons, the purple bars are start codons. The blue blocks are open reading frames. Individual Mascot peptide matches are shown in red. This particular gene has 5 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.

Protein vs. EST vs. Genome

▼ Search parameters

Type of search	: MS/MS Ion Search
Enzyme	: Trypsin/P
Fixed modifications	: ∅TRAQ4plex (K), ∅TRAQ4plex (N-term), ∅Methylthio (C)
Variable modifications	: ∅Acetyl (Protein N-term), ∅Gln->pyro-Glu (N-term Q), ∅Oxidation (M)
Mass values	: Monoisotopic
Protein mass	: Unrestricted
Peptide mass tolerance	: ± 0.9 Da
Fragment mass tolerance	: ± 0.6 Da
Max missed cleavages	: 1
Instrument type	: ESI-TRAP
Number of queries	: 33,191

Database	Size	Avg. 5% identity threshold	#matches > threshold
UniProt_mouse_20120418	2.7 x 10 ⁷	37	1910
EST_mouse_20120429	4.5 x 10 ⁹	60	465
Mouse_genome_20120123	2.7 x 10 ⁹	60	367

MASCOT : Nucleic Acid Databases

© 2012 Matrix Science

MATRIX
SCIENCE

All well and good, but which database gives the most matches?

As you can see in the table, there is a big drop in the number of matches between Uniprot_mouse and EST_mouse. The reason is mainly that EST_mouse is a much bigger database, by more than a factor of 100. This means that the score thresholds are approx 23 higher, and we lose all the weaker matches, that had scores between 37 and 60. Yes, there may be additional matches in EST, not found in Uniprot mouse, but the net change is highly negative.

You can see at a glance that the mouse genome is even worse. This is not because of a still higher threshold; the database is similar in size to EST_mouse. 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 approximately 21%. The other factor than can affect this is that the mouse 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.

Protein vs. EST vs. Genome

- Searching complete chromosomes is possible, but unwieldy.
- Scoring statistics for assembled genome very similar to EST_mouse, but
 - the genome is a single consensus sequence, EST_mouse represents many variants
 - EST_mouse is 100% coding, Genome assembly is ~1.5% coding
 - lose approx 20% of matches because they straddle an exon - intron boundary
- In general, EST_mouse is a better choice
- References (for human genome)
 - 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 mouse genome, and the same considerations probably hold for other large mammalian genomes (and certainly for the human genome).

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

Entrez records

Database name	Subtree links	Direct links
Nucleotide	3,440	-
Nucleotide EST	567,880	-
Nucleotide GSS	46,391	-
Protein	2,269	1
Genome	1	-
Popset	64	61
GEO Datasets	190	3
UniGene	24,930	-
UniSTS	139	-
PubMed Central	1,024	331
Gene	140	-
SRA Experiments	8	-
Probe	195	-
Bio Project	24	-
Bio Sample	248	-
Protein Clusters	70	-
Taxonomy	117	1

MASCOT : Nucleic Acid Databases © 2012 Matrix Science **MATRIX SCIENCE**

Here we're looking at the overall numbers for all the species of the genus Citrus in Genbank – and as you can see we have very few protein entries, but some genome data from a few species and reasonably large number of ESTs. So if you were working on a Citrus species, you would probably need to be searching a nucleic acid database.

Primary sequence variants

- Protein database
 - Look for all single base substitutions
 - No attempt to find single base insertions & deletions because of frame shifts
- Nucleic acid database
 - Look for all single base substitutions, insertions & deletions

An important reason for why you may wish to search against a nucleic acid database is to look for sequence variants, such as single nucleotide polymorphisms (SNPs) or sequencing errors using the error tolerant search in Mascot.

For a protein database, we can't look for the consequences of inserted or deleted bases, because these give rise to frame shifts, and the entire sequence changes from that point on, but if we're searching a nucleic acid database, we can look for all single base substitutions, insertions and deletions.

MASCOT Search Results

Peptide View

MS/MS Fragmentation of **GVDEATHDILTK**
 Found in [gi|372099091|ref|NC_000085.6|_1699](#) in **Mouse_genome**, bases 20376001-20388122 Mus musculus strain C57BL/6J
 chromosome 19, GRcm38 C57BL/6J

Translated in frame 6, insertion of G2 found by error tolerant search
 GTGTGGATGA AGCAACCATC ATTGACATTC TTACCAAG - original
 GGTGTGGATG AAGCAACCAT GATTGACATT CTTACCAAG - modified

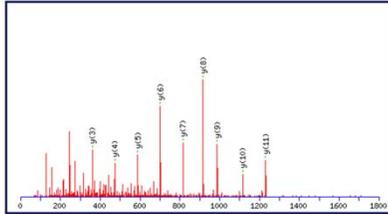
Match to Query 205: 1386.494048 from(694.254300,2+) intensity(46.8095) index(443)
 Data file 500.pkl

Click mouse within plot area to zoom in by factor of two about that point

Or, Plot from 0 to 1800 Da Full range

Label all possible matches Label matches used for scoring

Show Y-axis



Monoisotopic mass of neutral peptide Mr(calc): 1386.7606
 Variable modifications:
 : NA_INSERTION
 Ions Score: 73 Expect: 0.0013
 Matches : 9/114 fragment ions using 14 most intense peaks [\(help\)](#)

#	b	b ⁺⁺	b ⁰	b ⁰⁺⁺	Seq.	y	y ⁺⁺	y [*]	y ^{++*}	y ⁰	y ⁰⁺⁺	#
1	58.0287	29.5180			G							13
2	157.0972	79.0522			V	1330.7464	665.8769	1313.7199	657.3636	1312.7359	656.8716	12
3	272.1241	136.5657	254.1135	127.5604	D	1231.6780	616.3426	1214.6515	607.8294	1213.6674	607.3374	11
4	401.1667	201.0870	383.1561	192.0817	E	1116.6511	558.8292	1099.6245	550.3159	1098.6405	549.8239	10
5	472.2038	236.6055	454.1932	227.6003	A	987.6085	494.3079	970.5819	485.7946	969.5979	485.3026	9
6	573.2515	287.1294	555.2499	278.1241	T	916.5714	458.7893	899.5448	450.2760	898.5608	449.7840	8
7	686.3355	343.6714	668.3250	334.6661	I	815.5237	408.2655	798.4971	399.7522	797.5131	399.2602	7
8	799.4196	400.2134	781.4090	391.2082	I	702.4396	351.7234	685.4131	343.2102	684.4291	342.7182	6
9	914.4466	457.2269	896.4360	448.7216	D	589.3556	295.1814	572.3290	286.6681	571.3450	286.1761	5
10	1027.5306	514.2689	1009.5201	505.2637	I	474.3286	237.6679	457.3021	229.1547	456.3180	228.6627	4
11	1140.6147	570.8110	1122.6041	561.8057	L	361.2445	181.1259	344.2180	172.6126	343.2340	172.1206	3
12	1241.6624	621.3348	1223.6518	612.3295	T	248.1605	124.5839	231.1339	116.0706	230.1499	115.5786	2
13					K	147.1128	74.0600	130.0863	65.5468			1

MASCOT : Nucleic Acid Databases

© 2012 Matrix Science



We have a nice example of a possible base insertion.