

Sequence Databases

MASCOT

 MATRIX
SCIENCE

The collage features several logos and screenshots of biological databases and tools. At the top left is a screenshot of the NCBI 'Data & Software' page. To its right are logos for Human, WormBase, gpm, UniProt, TrEMBL, Proteomes, and Zebrafish. Below these are logos for swissprot, nextprot, NCBI, and EMBL-EBI. On the right side is a screenshot of the EMBL-EBI 'Services' page. At the bottom, the Mascot logo is displayed with the text ': Sequence Databases' and '© 2007-2023 Matrix Science', followed by the Matrix Science logo.

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.

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 (~568,000 entries)

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

UniProt proteome database (size varies by species)

- >300K proteomes of which 18K are reference proteomes
- Quality varies depending on popularity of species

NCBIprot, UniRef100 (NCBIprot ~520,000,000 entries)

- Comprehensive, non-identical

EST databases (>400,000,000 entries in translation)

- Very large and very redundant
- Not suitable for PMF

Sequences from a single genome

- Not suitable for PMF

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

SwissProt is acknowledged to be the best annotated database, and is non-redundant, making it an ideal choice for PMF searches, where the loss of one or two peptides is not a concern. SwissProt is also a good choice for MS/MS of a well characterised organism, such as human or mouse or yeast.

UniProt proteome database for the species of interest are an excellent database to choose especially if the species is of research importance, Human, Rat, Mouse, E. Coli etc as they will be well annotated and comprehensive. For less commonly analysed species they can still be a good resource that is a smaller database to search than, say, all of green plants in NCBIprot. The Uniprot proteomes are based on the translation of a completely sequenced genome and will normally include sequences that derive from extra-chromosomal elements such as plasmids or organellar genomes in organisms. Some proteomes may also include protein sequences based on high quality cDNAs. The raw sequence data comes from translations of genome sequence submissions to the International Nucleotide Sequence Database Consortium (INSDC). Proteomes with a Benchmarking Universal Single-Copy Orthologs (BUSCO) complete score above 95% considered good.

The comprehensive, non-identical databases are a good choice for MS/MS searching if you don't want to miss any matches. After NCBI changed the accession number formatting in 2017 the nr database definition is now called NCBIprot on Mascot

Server.

NCBIprot and UniRef100 both aim to include explicit representations of all known protein sequences. However, they are huge, over 300 million entries so take a long time to search. Plus, only the best quality data will obtain matches when searching the whole database. There are some non-redundant versions of UniRef, such as UniRef90 and UniRef50. If you search these databases you may miss some matches.

If the genome of your organism of interest has not been sequenced, it won't be represented in the protein databases, but there may be lots of Expressed Sequence Tags (ESTs). Not advisable for PMF, because many sequences correspond to protein fragments.

Single genome databases can sometimes be useful for MS/MS searches. You will want to include a contaminants database in the search, to ensure spectra from contaminants don't get mis-assigned to the target organism.

(Entry counts from mid 2022)

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
 S P S D # Q Q I S W Q K L Y I P E E S G G Q Y S T H
 Q A H Q T N Q S R S L G R N S T S Q K R V G A N I Q H
 AAGGCCCTCAGACTAAACAGCAGATCTCTTGGCAGAACTCTACAGCCGAAGAGAGTGGGGGCCAATATTCAACATT
 [299200] [299210] [299220] [299230] [299240] [299250] [299260] [299270]
 TTCGGGTAGTCTGATTGCTCTAGAGAACCGTCTTTGAGATGTTCCGGTCTTCTCTCACCCTGGTTATAAGTTGTAA
 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: FFLSSSSSY*CC*WLLLLPPPHHQRRRIIIMTTTTNNKKSSRRVVVAAADDEEGGGG
Start: -----M-----
Base 1: TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAGGGGGGGGGGGGGGGGGGG
Base 2: TTTTCCCAAAAGGGGTTCCTCCCAAAGGGGTTCCTCCCAAAGGGGTTCCTCCCAAAGGGG
Base 3: TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

* = stop

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.

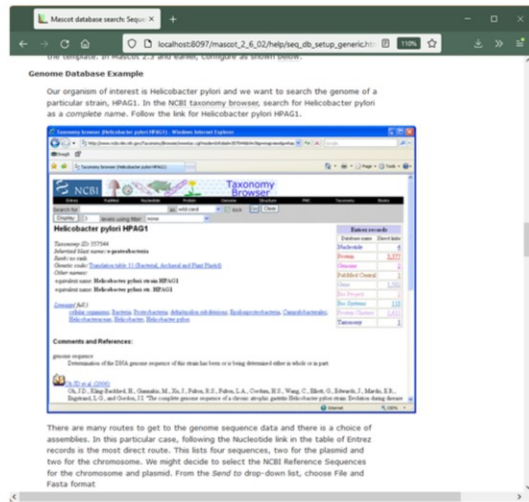
When a stop codon is encountered, leave a gap, and immediately re-start translation.

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.

Single Genome Data

Mascot help pages describe how to navigate NCBI web site



MASCOT

: Sequence Databases

© 2007-2023 Matrix Science

**MATRIX
SCIENCE**

All the genomes in GenBank are translated into protein sequences in NCBIprot. Usually, this is the simplest option for a Mascot search. But, if you are not confident that the coding sequences and reading frames have been identified correctly, or you are looking for something unusual, you might wish to search the genomic DNA directly. The Mascot help page for a generic database describes how to locate and download different types of sequence data, including genomic DNA -

https://www.matrixscience.com/help/seq_db_setup_generic.html

Single Genome Data

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
- <https://www.matrixscience.com/downloads/splitter.pl.gz>

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

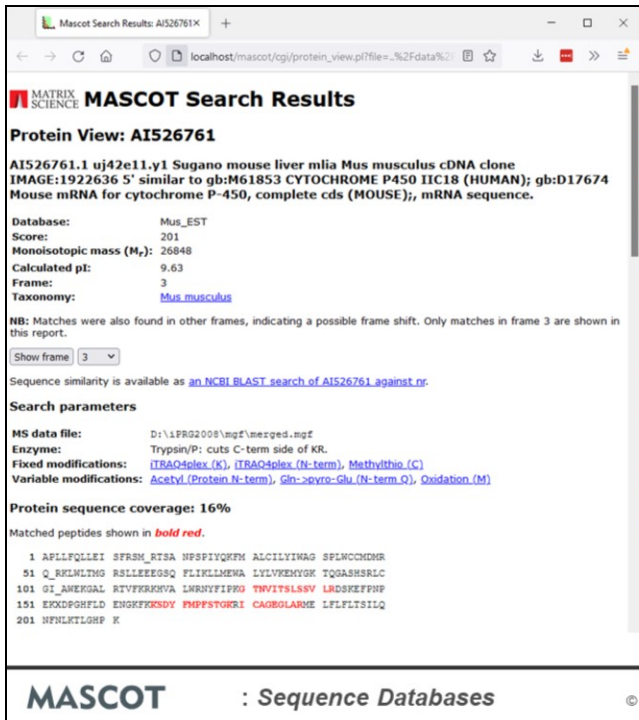
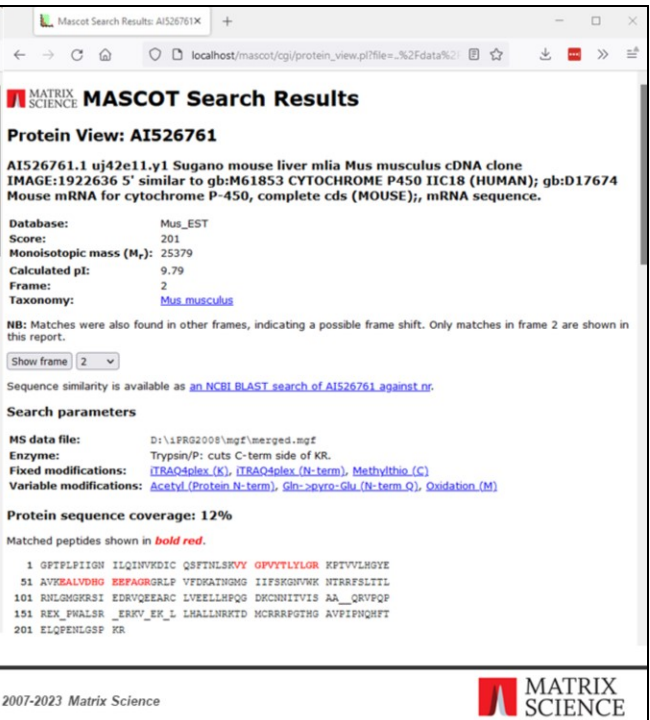
The longest human chromosome is chromosome 1 with 285 million base pairs. We don't know of any tools for reviewing the results which can handle 250 Mbp sequences.

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.

	Score	Mass	Matches	Sequences	emPAI	F										
5.1	383	15561	23 (23)	4 (4)	3.58	2	AW012478.1 uc05d11.y1 Sugano mouse liver m1a Mus musculus cDNA clone IMAGE:282317 5' siml...									
P53 sameasets of AW012478																
5.2	201	25379	14 (14)	5 (5)	1.56		A1524761.1 uc42d11.y1 Sugano mouse liver m1a Mus musculus cDNA clone IMAGE:192234 5' siml...									
5.3	171	23932	6 (6)	2 (2)	0.49	1	A4238951.1 uc04d12.y1 Soares mouse liver m1a Mus musculus cDNA clone IMAGE:694057 5' siml to gh...									
P20 sameasets of A4238951																
5.4	140	33937	17 (17)	5 (5)	1.02	1	A1047293.1 uc04d07.y1 Sugano mouse liver m1a Mus musculus cDNA clone IMAGE:1450492 5' siml...									
P2 sameasets of A1047293																
5.5	122	31131	9 (9)	3 (3)	0.59	6	A1132230.1 uc32d09.x1 Sugano mouse liver m1a Mus musculus cDNA clone IMAGE:1482088 5' siml...									
P2 sameasets of A1132230																
5.6	64	25761	2 (2)	2 (2)	0.45	2	A4882179.1 uc38d10.y1 Stratagene mouse lung 937302 Mus musculus cDNA clone IMAGE:1277938 5'...									
P2 sameasets of A4882179																
5.7	63	20243	6 (6)	2 (2)	0.60	2	A1529694.1 uc01c11.y1 Sugano mouse liver m1a Mus musculus cDNA clone IMAGE:1888820 5' siml...									
P2 sameasets of A1529694																
▼56 peptide matches (21 non-duplicate, 35 duplicate)																
Auto-fit to window																
Query Index	Observed	Mr (exp1)	Mr (calc)	Delta M	Score	Expect	Rank	0	1	2	3	4	5	6	7	Peptide
f4447 P3	521.2416	1560.7029	1559.8187	0.8842	0	59	0.00065	P1	0							K.NISQPTWTFSE.A
f4705	525.4566	1573.3479	1572.7654	0.5824	0	71	0.00028	P1	0							K.SALVONHDFPAG.G
f5144 P4	540.3247	1678.6349	1678.5385	0.0964	0	54	0.0091	P1	0							R.SCARGLAR.H
f5615 P2	541.3646	1680.7179	1680.4059	0.3118	0	44	0.007	P1	0							K.DYPTM.V
f7790	577.9297	1553.8449	1553.6045	0.2404	0	49	0.030	P1	0							R.OFFPHARE.I
f8340	586.6058	1169.9970	1169.5994	0.3976	0	33	0.028	P1	0							R.OFFPHARE.I - Oxidation (M)
f15114 P2	739.5340	1477.0534	1476.8634	0.1900	0	82	7.2e-05	P1	0							K.OTTVITSLSVLR.D
f18138	739.6174	1477.2207	1476.8108	0.4100	0	63	0.0026	P1	0							R.DFIDVYLK.Q
f19443 P2	745.4500	1489.8954	1489.5064	0.3846	0	73	0.0021	P1	0							K.OTTVITSLSVLR.D
f19465 P1	510.4492	1528.9257	1528.7756	0.1500	0	59	0.0019	P1	0							K.SALVONHDFPAG.G
f19473	745.5187	1529.0228	1528.7756	0.2472	0	93	3.3e-06	P1	0							K.SALVONHDFPAG.G
f19851 P2	773.1436	1544.2725	1543.8521	0.4205	0	61	0.0023	P1	0							VYGVVYLYLR.K
f20247 P7	781.1422	1560.2498	1559.8187	0.4311	0	80	5.3e-06	P1	0							K.NISQPTWTFSE.A
f20594	787.3444	1573.0743	1572.7654	0.2689	0	43	0.049	P1	0							K.SALVONHDFPAG.G
f21390 P6	536.3278	1605.9617	1605.8232	0.1384	1	44	0.014	P1	0							K.SALVONHDFPAG.G
f21414 P1	804.1063	1606.1980	1605.8232	0.3747	1	45	0.014	P1	0							K.SALVONHDFPAG.G
f22532	554.6924	1663.0254	1660.7614	0.2440	0	49	0.014	P1	0							K.SALVONHDFPAG.G
f23907	849.6286	1735.2427	1734.8402	0.4025	0	61	0.0030	P1	0							R.VQERACLVLR.A
f25449	614.1602	1639.4589	1638.9402	0.5187	1	42	0.036	P1	0							K.DYPTM.V
f26750 P6	645.2435	1932.7686	1932.0772	0.6914	0	50	0.0017	P1	0							K.OTTVITSLSVLR.D
f50244	776.1600	2325.4582	2325.2824	0.1758	0	43	0.023	P1	0							R.FIDILPMLPHEVTSIDK.F
▼36 subsets and intersections (215 subset proteins in total)																
</																

MASCOT Search Results

Protein View: AI526761

AI526761.1 uj42e11.y1 Sugano mouse liver mlaia Mus musculus cDNA clone
 IMAGE:1922636 5' similar to gb:M61853 CYTOCHROME P450 IIC18 (HUMAN); gb:D17674
 Mouse mRNA for cytochrome P-450, complete cds (MOUSE);, mRNA sequence.

Database: Mus_EST
 Score: 201
 Monoisotopic mass (M_r): 26848
 Calculated pI: 9.63
 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 AI526761 against nr](#).

Search parameters

MS data file: D:\IPRO2008\mgf\merged.mgf
 Enzyme: Trypsin/P: cuts C-term side of KR.
 Fixed modifications: [ITRAQ4plex \(K\)](#), [ITRAQ4plex \(N-term\)](#), [Methylation \(C\)](#)
 Variable modifications: [Acetyl \(Protein N-term\)](#), [Gln->pyro-Glu \(N-term Q\)](#), [Oxidation \(M\)](#)

Protein sequence coverage: 16%

Matched peptides shown in **bold red**.

```

1  APLFLQLLEI SFRSM_RISA NPSPIYQKFM ALCLLYINAG SFLWCKHMR
51 Q_RKLMLTMS RSLLEERGSQ FLIKILMENA LVLVEMYGK TQASASRLC
101 QI_ANEKSGAL RTVFERKHVA LNRWYFIYFG THVITSLSSV LRDSKEFFHP
151 EKXDPGHFLD ENKFKQSDY PNPFTGKRI CAGBGLARKE LFLFLTSILQ
201 NFNKLTLGHP K
      
```

MASCOT Search Results

Protein View: AI526761

AI526761.1 uj42e11.y1 Sugano mouse liver mlaia Mus musculus cDNA clone
 IMAGE:1922636 5' similar to gb:M61853 CYTOCHROME P450 IIC18 (HUMAN); gb:D17674
 Mouse mRNA for cytochrome P-450, complete cds (MOUSE);, mRNA sequence.

Database: Mus_EST
 Score: 201
 Monoisotopic mass (M_r): 25379
 Calculated pI: 9.79
 Frame: 2
 Taxonomy: [Mus musculus](#)

NB: Matches were also found in other frames, indicating a possible frame shift. Only matches in frame 2 are shown in this report.

Show frame: 2

Sequence similarity is available as [an NCBI BLAST search of AI526761 against nr](#).

Search parameters


MS data file: D:\IPRO2008\mgf\merged.mgf
 Enzyme: Trypsin/P: cuts C-term side of KR.
 Fixed modifications: [ITRAQ4plex \(K\)](#), [ITRAQ4plex \(N-term\)](#), [Methylation \(C\)](#)
 Variable modifications: [Acetyl \(Protein N-term\)](#), [Gln->pyro-Glu \(N-term Q\)](#), [Oxidation \(M\)](#)

Protein sequence coverage: 12%

Matched peptides shown in **bold red**.


```

1  GPTFLPIQGH ILQINWYDIC QSFTHLSKPY GPVYTLVGR KPTVVLHGYE
51 AVYBALVWNG KRFAGRGRLP VFDKATHWNG IIFKQKFWK WTRFSLTL
101 RNLQNGFVRSI EDVQEEARC LVEELNPGQ DKCNHITVIS AA_QRVQGP
151 REX_PHALSR _ERKV_EK_L LKALLNRKTD MCRARPOTNG AVTIPHQWFT
201 ELQFENLGSP KR
      
```

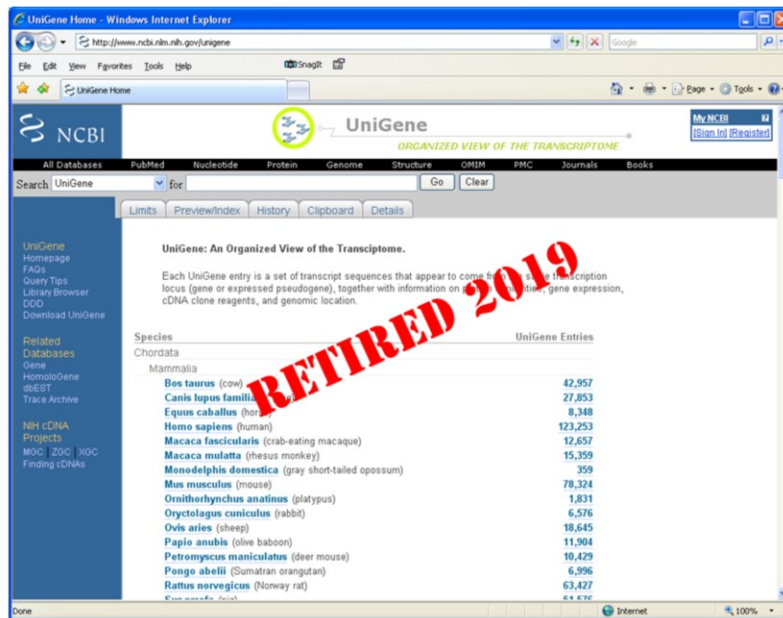


: Sequence Databases

© 2007-2023 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.

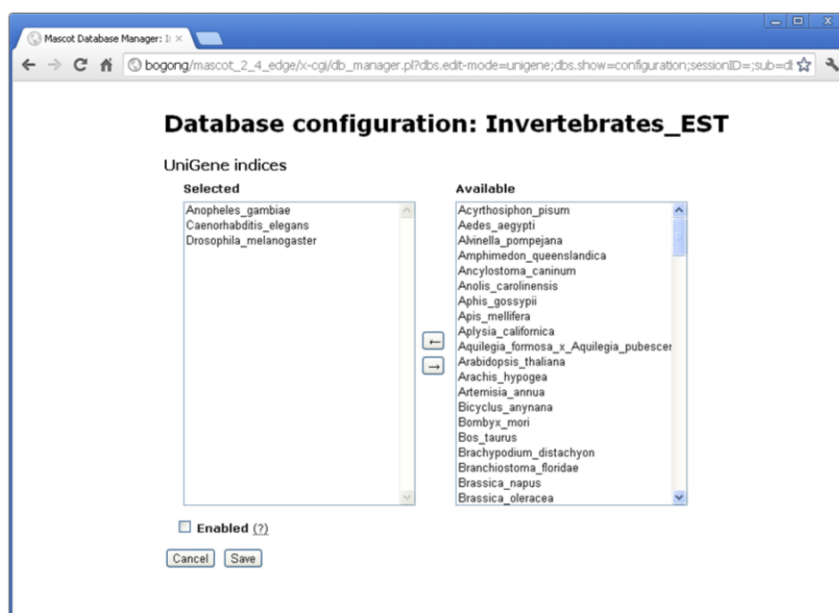


MASCOT : Sequence Databases

© 2007-2023 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. NCBI retired UniGene indexes in 2019. Mascot Server 2.8 still supports the feature and the indexes are still available for download.



UniGene index files can be downloaded manually from the NCBI FTP site, but if you are using Mascot 2.4 or later, UniGene is predefined for the EST databases from both NCBI and EMBL. If enabled, index files will be downloaded automatically whenever the FASTA file is updated.

If using Mascot 2.3 or earlier, you have to make configuration changes in the database update script and mascot.dat. Details can be found in Chapter 6 of the manual and in the Mascot help page for NCBI EST.

Format	Significance threshold $p <$	0.05	Max. number of families	AUTO	[help]
	Target FDR (overrides sig. threshold)	(not set) ▼	FDR type	PSM ▼	
	Display non-sig. matches	<input type="checkbox"/>	Min. number of sig. unique sequences	1 ▼	
	Show Percolator scores	<input type="checkbox"/>	Dendrograms cut at	0	
	UniGene index:	None ▼			
	Preferred taxonomy	None ▼			
		Mus_musculus			

► Sensitivity and FDR (reversed protein sequence): ,

When UniGene is configured, we can select Mus_musculus from the drop-down list in the format controls.

Proteins (905)
Report Builder
Unassigned (11844)
[permalink](#)

Protein families 1–20 (out of 895)

20 per page
1
2
3
4
5
6
...
45
Next
Expand all
Collapse all

Accession contains Find Clear

1	Mm.31018	738	Cytb5 Cytochrome b-5
2	1 Mm.330160 2 F0710191	654	Hspa3 Heat shock protein 3
		83	F0710191.1 Mus musculus mRNA 5-prime sequence 303000004660929.
3	Mm.14796	624	Mgat1 Microsomal glutathione S-transferase 1
4	Mm.15537	534	Cyp1a2 Cytochrome P450, family 1, subfamily a, polypeptide 2
5	Mm.473847	498	Transcribed locus, strongly similar to NP_044330.2 Mgat1 gene product [Mus musculus]
6	Mm.20764	481	Cyp2d9 Cytochrome P450, family 2, subfamily c, polypeptide 29
7	CB321249	477	CB321249.1 AGENCOURT_12238239 NM_MSC_136 Mus musculus cDNA clone IMAGE30...
8	Mm.289810	477	Rpl14 Ribosomal protein L14
9	Mm.425436	477	Transcribed locus, strongly similar to NP_080230.1 60S ribosomal protein L14 [Mus mus...
10	Mm.16660	434	P4hb Poly(4-hydroxylase, beta polypeptide
11	Mm.6696	411	Rdh7 Retinol dehydrogenase 7
12	1 Mm.398371 2 F0728659	407	Rpl7a Ribosomal protein L7A
		125	F0728659.1 Mus musculus mRNA 5-prime sequence from clone LADAA121YR14 (LADA...
13	Mm.328601	352	Transcribed locus, strongly similar to NP_038749.1 Rpl7a gene product [Mus musculus]
14	Mm.432030	352	Transcribed locus, strongly similar to NP_038749.1 Rpl7a gene product [Mus musculus]
15	Mm.332844	344	Cyp2a11 Cytochrome P450, family 2, subfamily a, polypeptide 11
16	Mm.20770	319	Cyp2a12 Cytochrome P450, family 2, subfamily a, polypeptide 12
17	Mm.292803	313	Ces1d Carboxylesterase 1D
18	Mm.29110	311	Ces1f Carboxylesterase 1F
19	Mm.295534	310	Ces3e Carboxylesterase 3A
20	Mm.26719	293	Hsd17b6 Hydroxysteroid (17-beta) dehydrogenase 6

20 per page
1
2
3
4
5
6
...
45
Next
Expand all
Collapse all

MASCOT

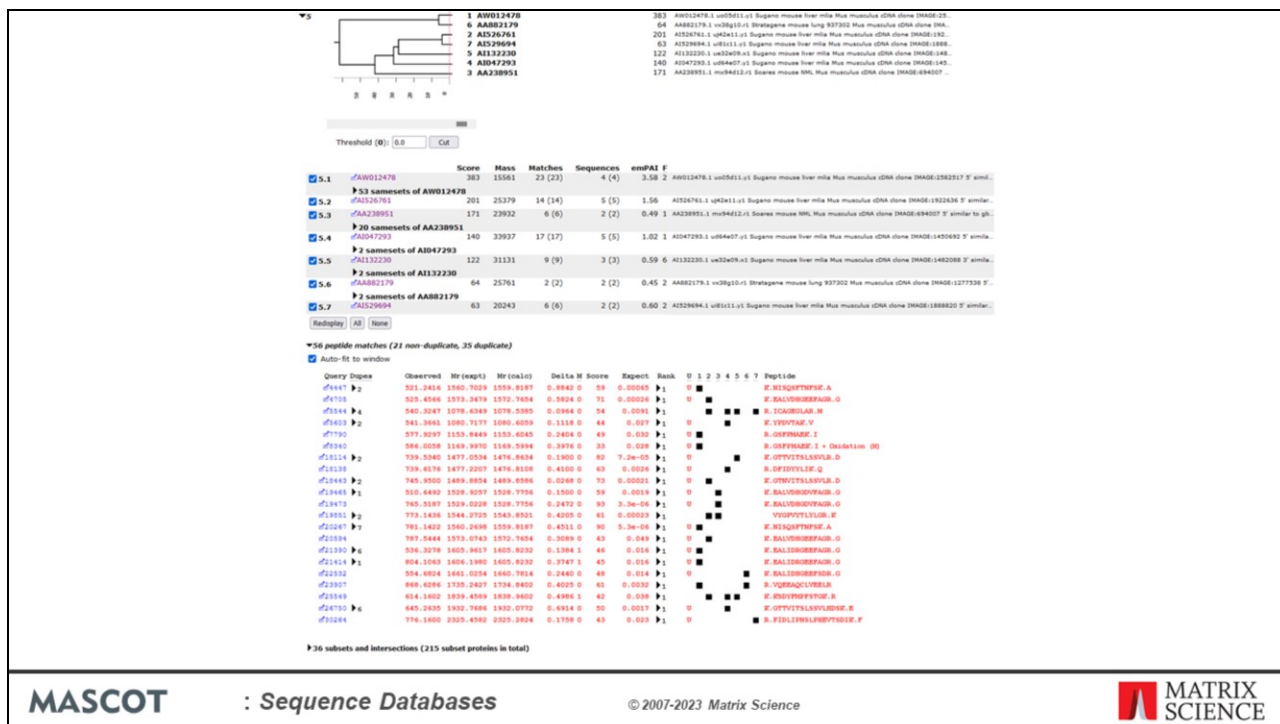
: Sequence Databases

© 2007-2023 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

© 2007-2023 Matrix Science



The protein family summary groups entries together, but it can only connect overlapping entries which have at least one shared peptide match, so it will sometimes fail.

There are seven proteins entries grouped together in protein family 5 from the EST search. The entry names give no clue as to the protein function.

Proteins (905)
Report Builder
Unsigned (31844)
[Permalink](#)

Protein families 1–20 (out of 895)
20 per page
1 2 3 4 5 6 ... 45 Next
Expand all Collapse all

Accession contains Find Clear

1 Mm.31018 738 Cyb5 Cytochrome b-5

2 1 Mm.330160 654 Hspa5 Heat shock protein 5

2 FO710191 83 FO710191:1 Mus musculus mRNA 5-prime sequence 3030000046609839.

3 Mm.14796 624 Mgst1 Mitochondrial glutathione S-transferase 1

4 Mm.15537 534 Cyp1a2 Cytochrome P450, family 1, subfamily a, polypeptide 2

5 Mm.473847 498 Transcribed locus, strongly similar to NP_064330.2 Mgst1 gene product [Mus musculus]

6 Mm.20764 481 Cyp2c29 Cytochrome P450, family 2, subfamily c, polypeptide 29

	Score	Mass	Matches	Sequences	emPAI	F
6.1 Mm.20764	481		0	40 (40)	9 (9)	

40 peptide matches (12 non-duplicate, 28 duplicate)
Auto-fit to window

Query	Dupes	Observed	Hr (expt)	Hr (calc)	Delta	M	Score	Expect	Rank	U	Peptide
M4447	2	521.2416	1560.7029	1559.8187	0.8842	0	59	0.00065	1	U	K.NISQSFTNFSK.A
M5544	4	540.3247	1078.6349	1078.5385	0.0964	0	54	0.0091	1	U	R.ICAGEGLAR.M
M403	2	541.3661	1080.7177	1080.6059	0.1118	0	44	0.027	1	U	K.YPDVTAQ.V
M7790		577.9297	1153.8449	1153.6045	0.2404	0	49	0.032	1	U	R.GSFFMAEK.M
M5340		586.0058	1169.9970	1169.5994	0.3976	0	33	0.028	1	U	R.GSFFMAEK.M + Oxidation (D)
M18138		739.6176	1477.2207	1476.8108	0.4100	0	63	0.0026	1	U	R.DFIDYLLIK.Q
M20247	7	781.1422	1560.2698	1559.8187	0.4511	0	90	5.3e-06	1	U	K.NISQSFTNFSK.A
M21390	6	536.3278	1605.9617	1605.8232	0.1384	1	46	0.016	1	U	K.EALIDRGEFAQR.G
M21414	1	804.1063	1606.1980	1605.8232	0.3747	1	45	0.016	1	U	K.EALIDRGEFAQR.G
M23907		868.6286	1735.2427	1734.8402	0.4025	0	61	0.0032	1	U	R.VQREAGCLVEELR.R
M25549		614.1602	1839.4589	1838.9602	0.4986	1	42	0.038	1	U	K.RSDYHFFSTQK.R
M24750	6	645.2635	1932.7686	1932.0772	0.6914	0	50	0.0017	1	U	K.GTTVTITSLSSVLDSK.E

1 subset or intersection (1 subset protein in total)

MASCOT : Sequence Databases
© 2007-2023 Matrix Science
MATRIX SCIENCE

However, when we look at the UniGene report, we find that many of these matches all belong to the same gene, for Cytochrome P450.

In this case there was some over grouping of proteins with shared peptides, and these have been split off into separate protein families.

The other advantage of UniGene is that it gives us the more useful descriptions.

Mouse Genome Statistics

- **2.7×10^9 bases**

(Mus_EST is $\sim 2.2 \times 10^9$ bases)

- **5.4×10^9 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 mouse genome.

The mouse genome assembly is approximately 2.7 billion bases, which makes it a little larger than Mus_EST.

Since we must translate in all 6 reading frames, this corresponds to 5.4 billion amino acid residues.

In the mouse 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 mouse 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!

U.S. National Library of Medicine

National Center for Biotechnology Information

Log in

Genome Data Viewer

GDV supports the exploration and analysis of *NCBI-annotated* and selected non-NCBI annotated eukaryotic genome assemblies. Currently, assemblies from over 1510 organisms are available.

Switch view

Search organisms

Mus musculus (house mouse)

To view more organisms in the tree, click on nodes that have "+" signs. Press and hold the "+" to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.

Mus musculus (house mouse)

Search in genome

Location, gene or phenotype
 Examples: GRCm39, chr1:113043000-113056000, DNA repair

Assembly

GRCm39

Assembly details

Name

GRCm39

RefSeq accession

GCF_000001635.27

GenBank accession

GCA_000001635.9

Submitter

Genome Reference Consortium

Level

Chromosome

Category

Reference genome

Replaced by

GCF_000001635.26

Annotation details

Annotation Release

109


Release date

Sep 21, 2020


MASCOT : Sequence Databases

© 2007-2023 Matrix Science

You can download the mouse genome sequences from NCBI.


National Library of Medicine
National Center for Biotechnology Information

[Datasets homepage](#) / [Assembly](#) / [GRCm39](#)



Genome assembly GRCm39
reference

Download

Reference sequence	RefSeq GCF_000001635.27
Submitted sequence	GenBank GCA_000001635.9
Taxon	<i>Mus musculus</i> (house mouse)
Strain	C57BL/6J
Submitter	Genome Reference Consortium
Date	Jun 24, 2020

[View the legacy Assembly page](#)

Assembly statistics

These statistics describe the nuclear genome of the reference sequence, GCF_000001635.27

Genome size	2.7 Gb
Number of chromosomes	21
Number of scaffolds	101
Scaffold N50	106.1 Mb
Scaffold L50	11
Number of contigs	305
Contig N50	59.5 Mb
Contig L50	15
GC percent	41.5
Assembly level	Chromosome

Download

Download a data package for GCF_000001635.27

Select file types - estimated size 728 Mb

- ☒ Genomic sequence, (FASTA)
- ☐ Annotated features (GTF)
- ☐ Annotated features (GFF3)
- ☐ Sequence and annotation (GBFF)
- ☐ Transcripts (FASTA)
- ☐ Genomic CDS (FASTA)
- ☐ Proteins (FASTA)

Your selected data will be downloaded as a ZIP archive

Name your file

Cancel
Download

MASCOT
: *Sequence Databases*

© 2007-2023 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

MASCOT Search Results

User: [redacted]
 E-mail: [redacted]
 Search title: IPKG2008 Genomic Mouse
 MS data file: D:\IPKG2008\mgf\merged.mgf
 Database: Mus musculus (NCBI) genome (20040425) (3,364,368 sequences; 5,535,816,732 residues)
 Timestamp: 4 May 2022 at 15:08:34 GMT

Search: As:

Not what you expected? Try [flex select summary](#).

Search parameters

Score distribution

Qualification statistics for all protein families

Legend

Protein Family Summary

Format: Significance threshold p: Max. number of families:
 Target FDR (overlaid sig. threshold): FDR type:
 Display non-sig. matches: ☐ Min. number of sig. unique sequences:
 Show Peptide scores: ☐ Dendrogram cut at:
 Preferred taxonomy:

Identifiability and FDR (reversed protein sequences)

Proteins (332)

Protein families 1-10 (out of 312)

10 per page: 1 2 3 4 5 6 7 8 9 10 Next Export all Collapse all

Accession	Options	Find	Ctrl
1	CH000995.3_2889	650	Access 0407001-0407001 Mus musculus chromosome 5, BRIN29 reference primer.
2	CH000995.3_2914	16	Access 0407001-0407001 Mus musculus chromosome 5, BRIN29 reference primer.
3	CH000995.3_2915	86	Access 0407001-0407001 Mus musculus chromosome 5, BRIN29 reference primer.
4	CH000995.3_11511	551	Access 12012011-12012011 Mus musculus chromosome 5, BRIN29 reference primer.
5	CH000995.3_1075	412	Access 0407001-0407001 Mus musculus chromosome 5, BRIN29 reference primer.
6	CH000995.3_4800	425	Access 0407001-0407001 Mus musculus chromosome 5, BRIN29 reference primer.
7	CH000995.3_2748	291	Access 0407001-0407001 Mus musculus chromosome 5, BRIN29 reference primer.
8	CH000995.3_2633	275	Access 12012011-12012011 Mus musculus chromosome 5, BRIN29 reference primer.
9	CH000995.3_10044	304	Access 12012011-12012011 Mus musculus chromosome 5, BRIN29 reference primer.
10	CH000995.3_2926	288	Access 0407001-0407001 Mus musculus chromosome 5, BRIN29 reference primer.
11	CH000995.3_12642	129	Access 12012011-12012011 Mus musculus chromosome 5, BRIN29 reference primer.
12	CH000995.3_12131	276	Access 12012011-12012011 Mus musculus chromosome 5, BRIN29 reference primer.
13	CH000995.3_10038	276	Access 12012011-12012011 Mus musculus chromosome 5, BRIN29 reference primer.

Not what you expected? Try [flex select summary](#).

MASCOT : Sequence Databases

© 2007-2023 Matrix Science

MATRIX SCIENCE

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

ethcogradient:mascot.cgi?master_results_2.phtml=62f4a4%2F20220504%2F001360.dat_sighrthreshold=0.054

Format: Significance threshold p-c: 0.05 Max. number of families: AUTO [?help]
 Target FDR (overrides sig. threshold): [not set] FDR type: PSM
 Display non sig. matches: [] Min. number of sig. unique sequences: 1
 Show Percolator scores: [] Dendrograms cut at: 0
 Preferred taxonomy: All entries

Sensitivity and FDR (reversed protein sequences)

Proteins (332) Report Builder Unassigned (32488) 1, normal

Protein families 1-10 (out of 312)

10 per page 1 2 3 4 5 6 Next Expand all Collapse all Clear

Accession contains Find

▼ 1

1 CH000995.3_2889
 2 CH001010.3_2934
 2 CH000994.3_2975

650 bases 3463001-3468122 Mus musculus chromosome 2, GRCh38 reference primary assembly, CDS.
 56 bases 3519601-3520822 Mus musculus chromosome 17, GRCh38 reference primary assembly, CDS.
 56 bases 4768001-4770121 Mus musculus chromosome 1, GRCh38 reference primary assembly, CDS.

Threshold (B): 0 Cut

Accession	Score	Mass	Matches	Sequences	empAI F	Database
1.1 CH000995.3_2889	650	475212	29 (29)	11 (11)	0.12	bases 3463001-3468122 Mus musculus chromosome 2, GRCh38 reference primary assembly, CDS.
1.2 CH000994.3_2975	96	490486	4 (4)	2 (2)	0.02	bases 4768001-4770121 Mus musculus chromosome 1, GRCh38 reference primary assembly, CDS.
1.3 CH001010.3_2934	56	471792	2 (2)	2 (2)	0.02	bases 3519601-3520822 Mus musculus chromosome 17, GRCh38 reference primary assembly, CDS.

Redisplay All None

▼ 35 peptide matches (17 non-duplicate, 18 duplicate)

Auto-fit to window

Query Tripe	Observed	Mr (aa)	Mr (aa)	Delta M Score	Expect	Rank	1	2	3	Peptide
df6119 P1	803.8705	1205.7284	1205.4747	0.0558	0	41	0.028	P1	0	K.VLSDHGLP.E
df6176 P2	411.5753	1223.1340	1220.4840	0.4494	0	49	0.026	P1	0	K.VGSDITPL.L
df1144 P1	635.4902	1248.9454	1248.4854	0.2759	0	55	0.013	P1	0	K.FHSDHGLP.E
df1114 P1	803.1397	1404.2448	1403.8337	0.4311	0	43	0.0022	P1	0	K.HSLSDHGLP.E
df2473 P1	571.0578	1710.1521	1709.4746	0.2770	0	57	0.0014	P1	0	K.LTPSDHGLP.E
df2473 P2	804.1669	1710.1620	1709.4746	0.4887	0	60	0.0003	P1	0	K.LTPSDHGLP.E
df2534 P1	656.4864	1717.3425	1717.4879	-0.5253	0	77	0.0013	P1	0	K.HSDHGLP.E
df1277 P1	607.4422	1819.3048	1818.8205	0.4793	0	55	0.019	P1	0	K.HSDHGLP.E
df2487 P1	609.0594	1824.1449	1823.4948	0.1503	0	94	0.015	P1	0	K.HSDHGLP.E
df2476 P1	903.0935	1904.1705	1903.8845	0.1881	0	94	1e-05	P1	0	K.HSDHGLP.E
df2487 P1	974.7142	1947.4139	1947.0320	0.3218	0	43	0.0074	P1	0	K.HSDHGLP.E
df1715 P1	563.4427	1949.2705	1949.0247	0.2842	0	94	1.1e-07	P1	0	K.HSDHGLP.E
df1715 P2	636.2484	1949.2705	1949.0247	0.7218	0	44	0.001	P1	0	K.HSDHGLP.E
df1715 P3	1019.2338	2036.4530	2036.1630	0.2899	0	48	0.0047	P1	0	K.HSDHGLP.E
df1715 P4	630.4307	2077.2703	2077.1076	0.1625	0	82	4.3e-06	P1	0	K.HSDHGLP.E
df1715 P5	1063.1591	2128.2701	2128.1264	0.2450	0	60	5.3e-06	P1	0	K.HSDHGLP.E
df1715 P6	755.2793	2242.8140	2242.1048	0.7112	0	49	6.2e-06	P1	0	K.HSDHGLP.E

MASCOT

: Sequence Databases

© 2007-2023 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

MASCOT : Sequence Databases © 2007-2023 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.

To enable this feature add the “FeatureTableLength” parameter to the options section of the mascot.dat file using the Configuration Editor->Configuration Options editor or a text editor. Set the value to less than the number of bases that the genomic was split into. A FeatureTableLength 10000 is a good value.

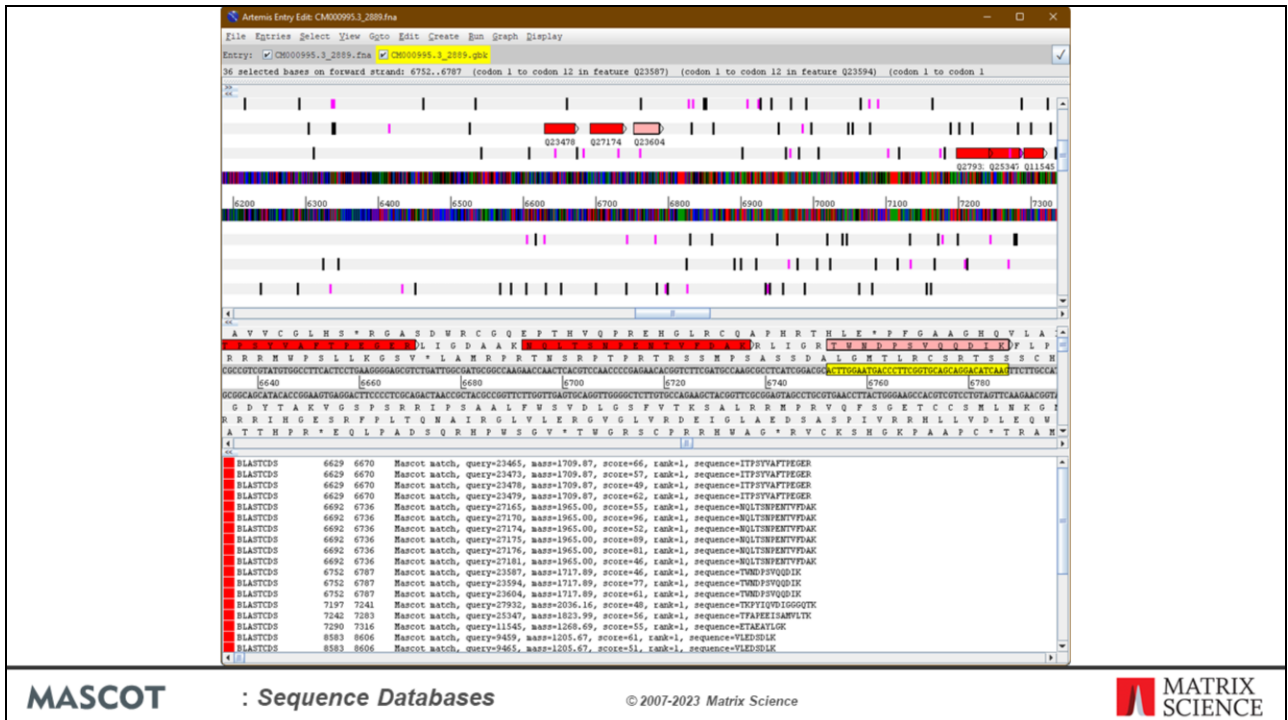
The screenshot displays the Artemis genome browser interface. At the top, there is a navigation bar with links: Downloads, Further information, Contact, Publications, and Programmes and Facilities. Below this, a red banner contains the text "Tool Annotations Features and Modules". The main heading is "Artemis", followed by a description: "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." Below this, a paragraph states: "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." An "About" section is also present. The bottom part of the screenshot shows a feature table with columns for "Feature", "Start", "End", "Name", "Gene", and "Accession". The table contains several rows of data, including "Gene", "Feature", "Start", "End", "Name", "Gene", and "Accession".

MASCOT : Sequence Databases

© 2007-2023 Matrix Science

MATRIX SCIENCE

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 start (purple) and stop (black) codons.. Individual Mascot peptide matches are shown in red or pink when selected. This particular gene has 11 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.

Mouse UniProt vs. EST vs. Genome

▼ Search parameters

Type of search	: MS/MS Ion Search
Enzyme	: Trypsin/P
Fixed modifications	: O^6 TRAQ4plex (K), O^6 TRAQ4plex (N-term), O^6 Methylthio (C)
Variable modifications	: O^6 Acetyl (Protein N-term), O^6 Gln->pyro-Glu (N-term Q), O^6 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 in residues	Average score threshold	Number of PSMs (1% FDR)
Uniprot mouse	2.8×10^7	37	1834
EST mouse	4.5×10^9	59	675
Mouse genome	5.5×10^9	59	548

MASCOT

: Sequence Databases

© 2007-2023 Matrix Science



All well and good, but which database gives the most matches? We searched a larger dataset against all 3 databases. The data was the public iPRG2008 dataset distributed by ABRF.

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 22 higher, and we lose all the weaker matches, that had scores between 37 and 59. Yes, there may be additional matches in EST, not found in Uniprot, 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; although the database is slight larger than Mus_EST the thresholds are the same. 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 just under 20% at 18.8%. The other factor 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.

Mouse UniProt vs. EST vs. Genome

- **Searching complete chromosomes is possible, but unwieldy.**
- **Scoring statistics for assembled genome very similar to Mus_EST, but**
 - the genome is a single consensus sequence, Mus_EST represents many variants
 - Mus_EST is 100% coding, MG assembly is 1.5% coding
 - lose approx 20% of matches because they straddle an exon - intron boundary

• **In general, Mus_EST 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 mouse 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.