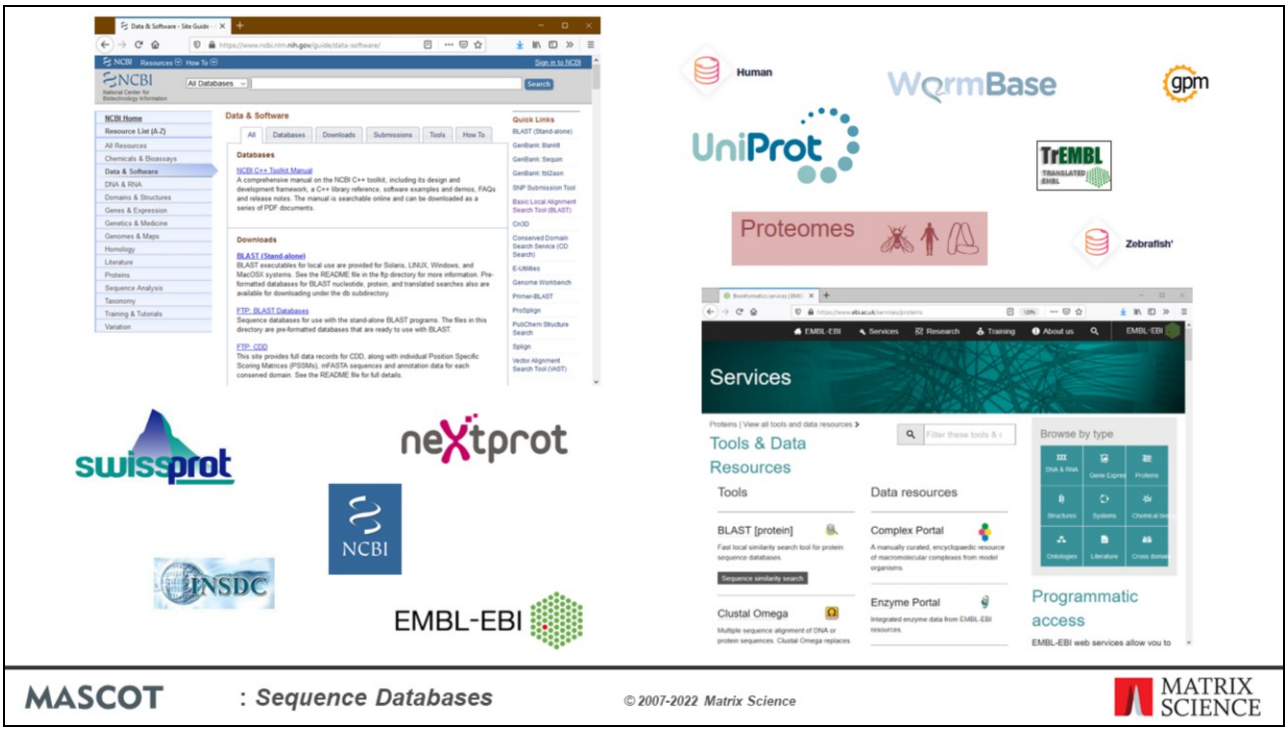
Sequence Databases

MASCOT

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.

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 (~564,000 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 ~340,000,000 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

MASCOT : *Sequence Databases* 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.

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 compressive. 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 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 UniProt100, 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
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  F
 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
CAAGCCCATCAGACTAACAGCAGATCTCTTGGCAGAAACTCTACAAGCCAGAAGAGAGTGGGGGCCAATATTCAACATT
 299200    299210    299220    299230    299240    299250    299260    299270
GTTCGGGTAGTCTGATTGTCGTCTAGAGAACCGTCTTTGAGATGTTCGGTCTTCTCTCACCCCCGGTTATAAGTTGTAA
 A  W  *  V  L  L  L  D  R  P  L  F  E  V  L  W  F  L  T  P  A  L  I  *  C  F
 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:  FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Start:    ----------------------------------M----------------------------
Base 1:   TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base 2:   TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base 3:   TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG

* = stop
```

MASCOT : *Sequence Databases* 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.

MASCOT : *Sequence Databases* MATRIX SCIENCE

The rules for NA translation in Mascot are

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

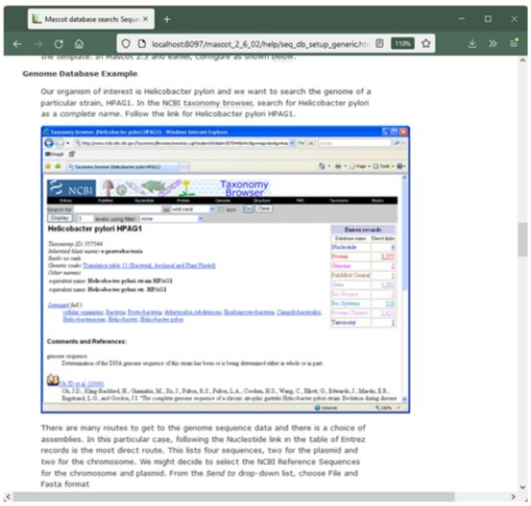
When a stop codon is encountered, leave a gap, and immediately re-start translation

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* 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 - http://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
  - www.matrixscience.com/downloads/splitter.pl.gz

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

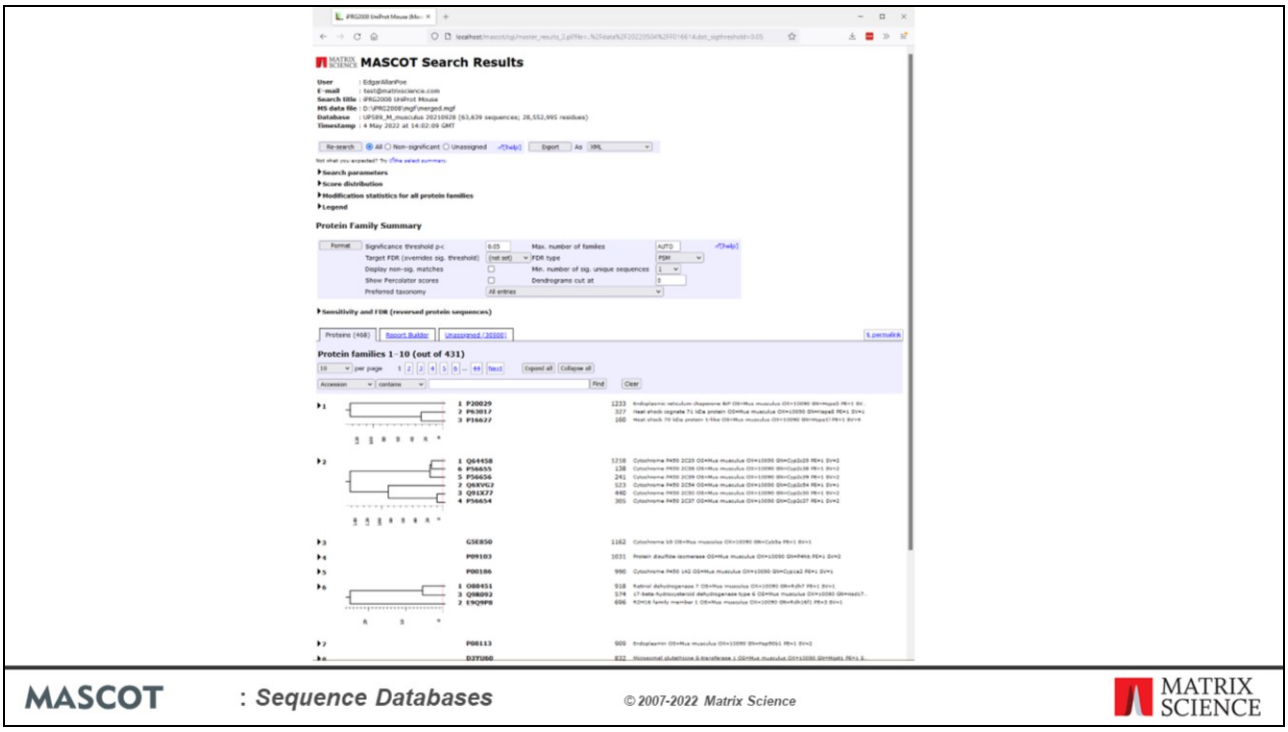
The longest human chromosome is chromosome 1 with 285 million base pairs

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.



To illustrate the features of the different types of database, we first searched a very small dataset of 33 thousand MS/MS spectra against a protein database, the Uniprot complete mouse proteome. There are significant matches to some 431 Mouse proteins.

MASCOT : Sequence Databases © 2007-2022 Matrix Science MATRIX SCIENCE

With Mus_EST, we obtained a very similar set of peptide matches. However, look at the hit-list. Unlike the protein database search, it doesn't immediately communicate which proteins have been found. I'll return to this issue later.

		Score	Mass	Matches	Sequences	emPAI	F	
5.1	AW012478	383	15561	23 (23)	4 (4)	3.58	2	AW012478.1 uo05d11.y1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:2582517 5' simil...
	53 samesets of AW012478							
5.2	AI526761	201	25379	14 (14)	5 (5)	1.56		AI526761.1 uj42e11.y1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:1922636 5' similar...
5.3	AA238951	171	23932	6 (6)	2 (2)	0.49	1	AA238951.1 mx94d12.r1 Soares mouse NML Mus musculus cDNA clone IMAGE:694007 5' similar to gb...
	20 samesets of AA238951							
5.4	AI047293	140	33937	17 (17)	5 (5)	1.02	1	AI047293.1 ud64e07.y1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:1450692 5' simila...
	2 samesets of AI047293							
5.5	AI132230	122	31131	9 (9)	3 (3)	0.59	6	AI132230.1 ue32e09.x1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:1482088 3' simila...
	2 samesets of AI132230							
5.6	AA882179	64	25761	2 (2)	2 (2)	0.45	2	AA882179.1 vx38g10.r1 Stratagene mouse lung 937302 Mus musculus cDNA clone IMAGE:1277538 5'...
	2 samesets of AA882179							
5.7	AI529694	63	20243	6 (6)	2 (2)	0.60	2	AI529694.1 ui81c11.y1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:1888820 5' similar...

Redisplay All None

56 peptide matches (21 non-duplicate, 35 duplicate)

Auto-fit to window

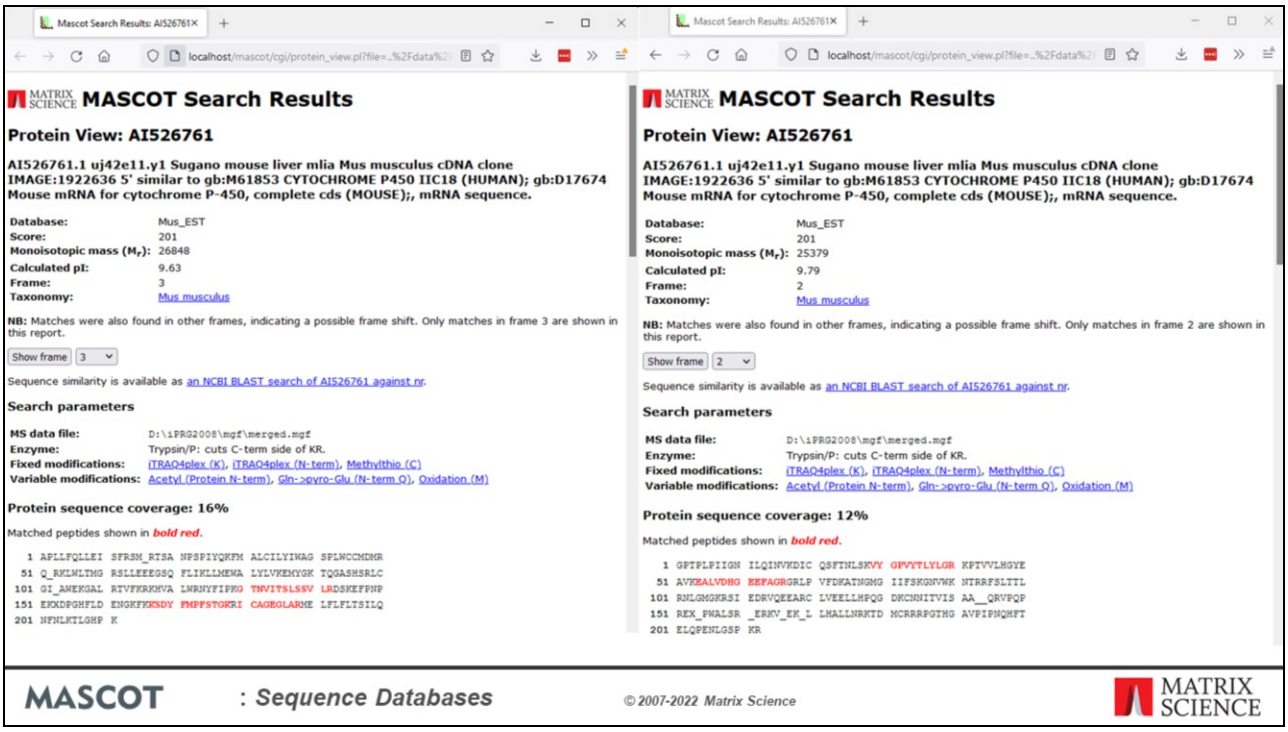
Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	3	4	5	6	7	Peptide
4447	2	521.2416	1560.7029	1559.8187	0.8842	0	59	0.00065	1	U	■							K.NISQSFTNFSK.A
4705		525.4566	1573.3479	1572.7654	0.5824	0	71	0.00026	1	U		■						K.EALVDHGEEFAGR.G
5549	4	540.3247	1078.6349	1078.5385	0.0964	0	54	0.0091	1			■		■	■		■	R.ICAGEGLAR.M
5603	2	541.3661	1080.7177	1080.6059	0.1118	0	44	0.027	1	U				■				K.YPDVTAK.V
7780		577.9297	1153.8449	1153.6045	0.2404	0	49	0.032	1	U	■							R.GSFPMAEK.I
8340		586.0058	1169.9970	1169.5994	0.3976	0	33	0.028	1	U	■							R.GSFPMAEK.I + Oxidation (M)
18114	2	739.5340	1477.0534	1476.8634	0.1900	0	82	7.2e-05	1	U					■			K.GTTVITSLSSVLR.D
18138		739.6176	1477.2207	1476.8108	0.4100	0	63	0.0026	1	U				■				R.DFIDYYLIK.Q
18443	2	745.9500	1489.8854	1489.8586	0.0268	0	73	0.00021	1	U		■						K.GTNVITSLSSVLR.D
19465	1	510.6492	1528.9257	1528.7756	0.1500	0	59	0.0019	1	U			■					K.EALVDHGDVFAGR.G
19473		765.5187	1529.0228	1528.7756	0.2472	0	93	3.3e-06	1	U			■					K.EALVDHGDVFAGR.G
19851	2	773.1436	1544.2725	1543.8521	0.4205	0	61	0.00023	1			■	■					VYGPVYTLYLGR.K
20247	7	781.1422	1560.2698	1559.8187	0.4511	0	90	5.3e-06	1	U	■							K.NISQSFTNFSK.A
20594		787.5444	1573.0743	1572.7654	0.3089	0	43	0.049	1	U		■						K.EALVDHGEEFAGR.G
21390	6	536.3278	1605.9617	1605.8232	0.1384	1	46	0.016	1	U	■							K.EALIDRGEEFAGR.G
21414	1	804.1063	1606.1980	1605.8232	0.3747	1	45	0.016	1	U	■							K.EALIDRGEEFAGR.G
22532		554.6824	1661.0254	1660.7814	0.2440	0	48	0.014	1	U						■		K.EALIDRGEEFSDR.G
23907		868.6286	1735.2427	1734.8402	0.4025	0	61	0.0032	1		■					■		R.VQEEAQCLVEELR
25549		614.1602	1839.4589	1838.9602	0.4986	1	42	0.038	1			■		■	■			K.ESDYFMFFSTGK.R
26750	6	645.2635	1932.7686	1932.0772	0.6914	0	50	0.0017	1	U				■				K.GTTVITSLSSVLHDSK.E
30244		776.1600	2325.4582	2325.2824	0.1758	0	43	0.023	1	U							■	R.FIDLIFNSLPHEVTSDIK.F

36 subsets and intersections (215 subset proteins in total)

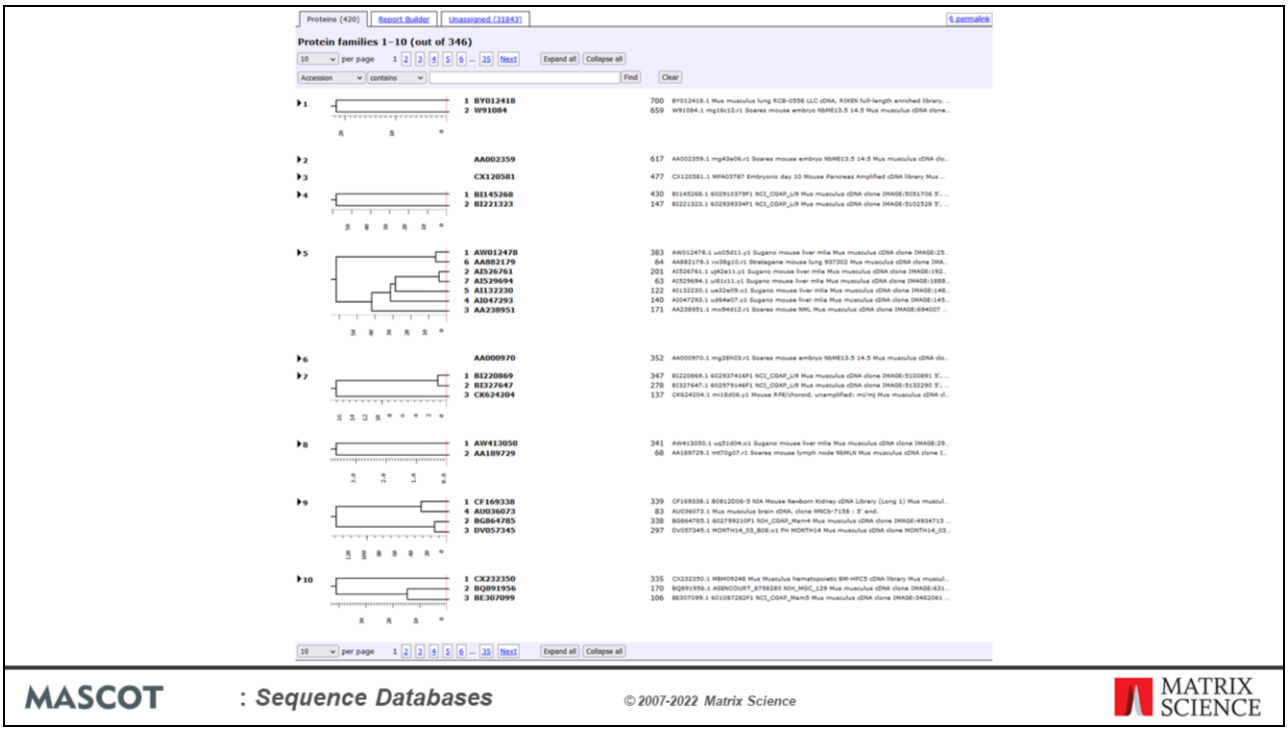
	Score	Mass	Subset of	
BI247578	378	36187	S.1	BI247578.1 602960035F1 NCI_CGAP_U9 Mus musculus cDNA clone IMAGE:5125617 5', mRNA sequence.
AA575398	370	16504	S.1	AA575398.1 vl87b09.r1 Stratagene mouse diaphragm (#937303) Mus musculus cDNA clone IMAGE:987641 5' simil...
22 samesets of AA575398				
AI097827	365	19139	S.1	AI097827.1 ue36h05.y1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:1482489 5' similar to gb:D1767...
2 samesets of AI097827				
BF234254	348	38548	S.1	BF234254.1 602026142F1 NCI_CGAP_U9 Mus musculus cDNA clone IMAGE:4161347 5', mRNA sequence.
7 samesets of BF234254				
AI529126	344	21653	S.1	AI529126.1 ui60f01.y1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:1886809 5' similar to gb:D17674...
AI785827	335	13453	S.1	AI785827.1 ui79g03.y1 Sugano mouse liver mlia Mus musculus cDNA clone IMAGE:1888660 5' similar to gb:M6185...

MASCOT : Sequence Databases © 2007-2022 Matrix Science MATRIX SCIENCE

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

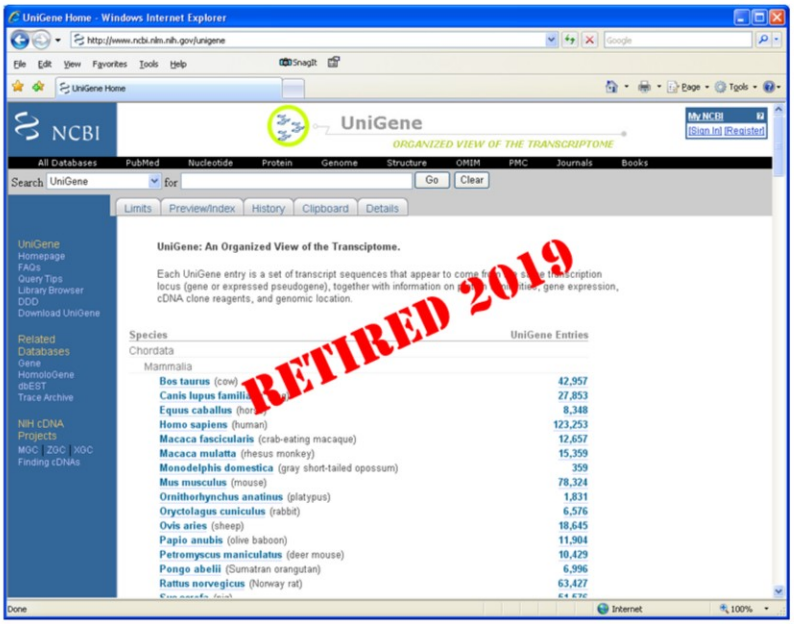


We get a protein view. This is similar to the protein view for a protein database entry, except we have drop down list for the different translation frames. For this particular entry, most of the matches have been found in reading frame 3.
But, as so often happens, there is a frame shift in this entry, and there is an additional match in frame 2.

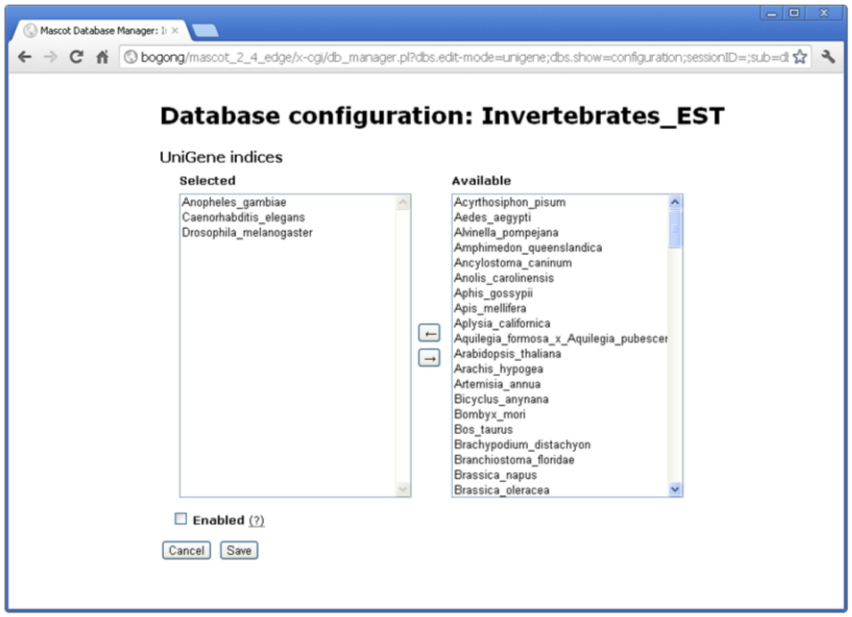


Going back to the issue of the hit list and the descriptions not saying very much. There are several problems here. One is that EST databases usually have a huge amount of redundancy, which can make for very long reports. Another problem is that the sequences tend to be short, so we don't get much grouping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



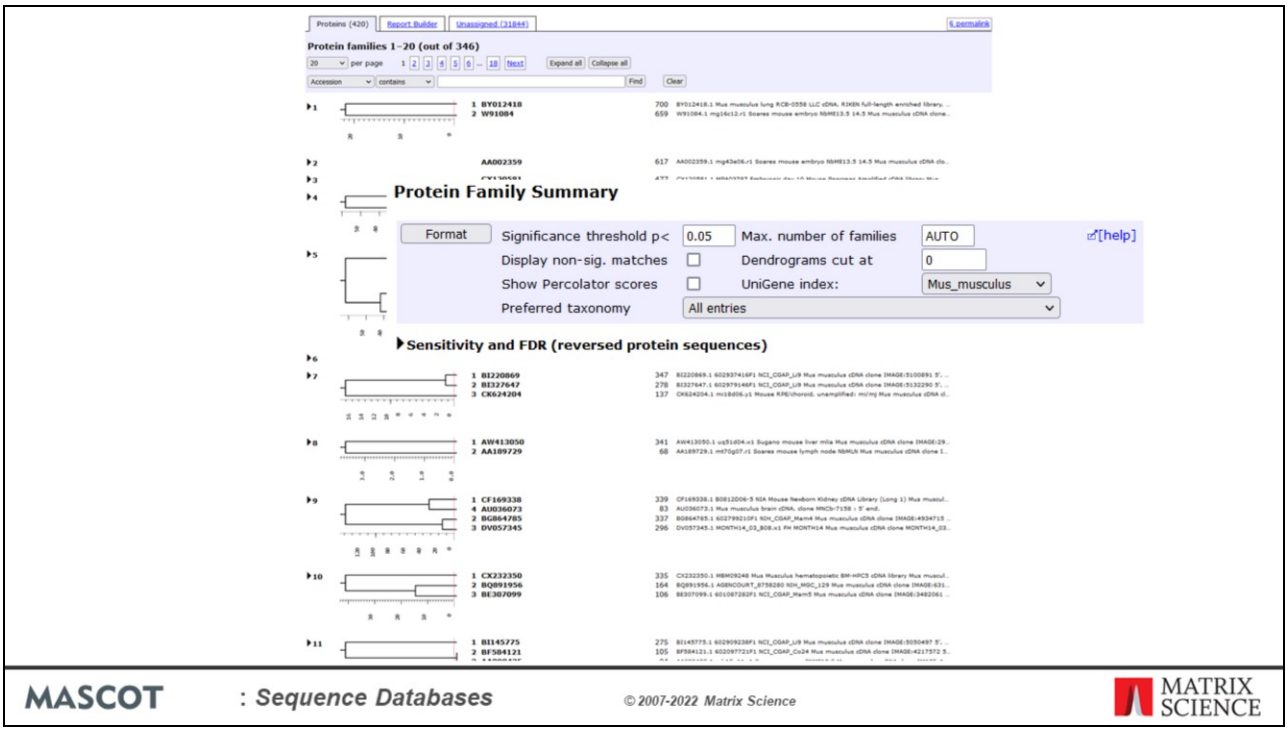
UniGene is not a sequence database, it is an index which is created by BLASTing GenBank sequences against themselves to cluster them into gene families. NCBI retired Unigene indexes in 2019. Mascot Server 2.7 still supports the feature and the indexes are still available for download.



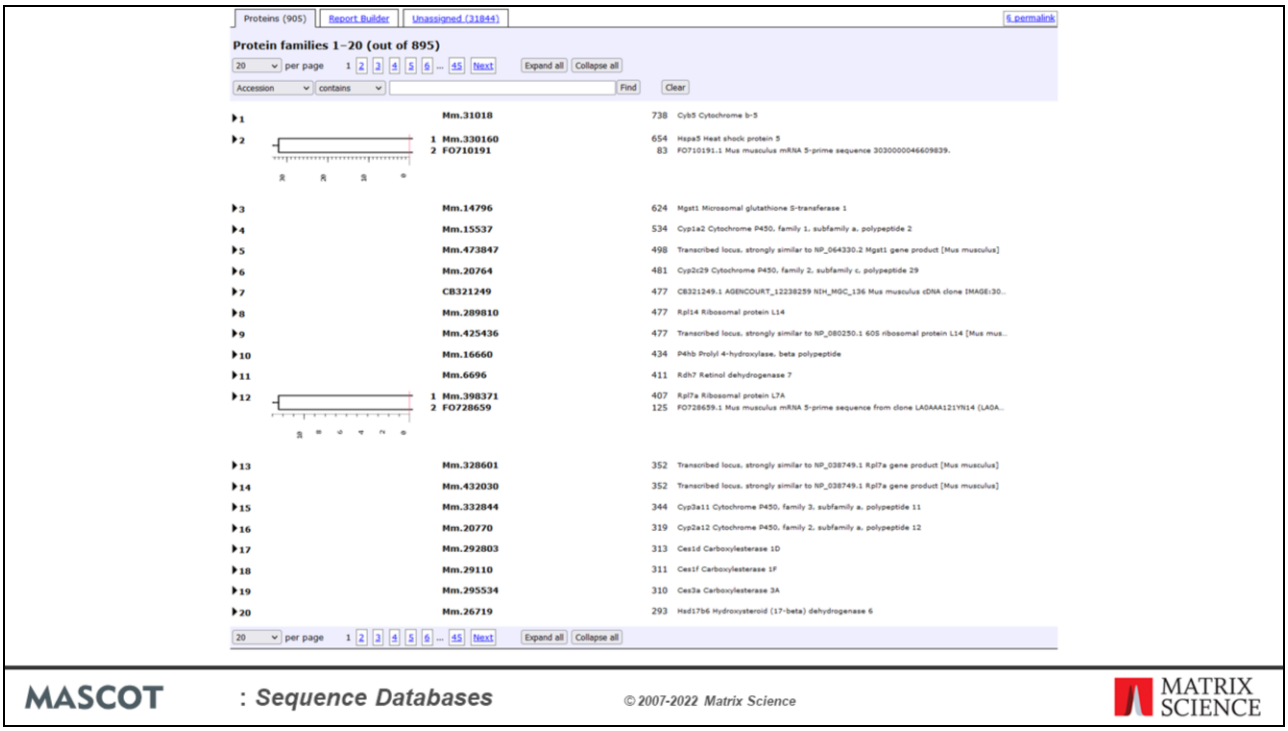
MASCOT : *Sequence Databases* MATRIX SCIENCE

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

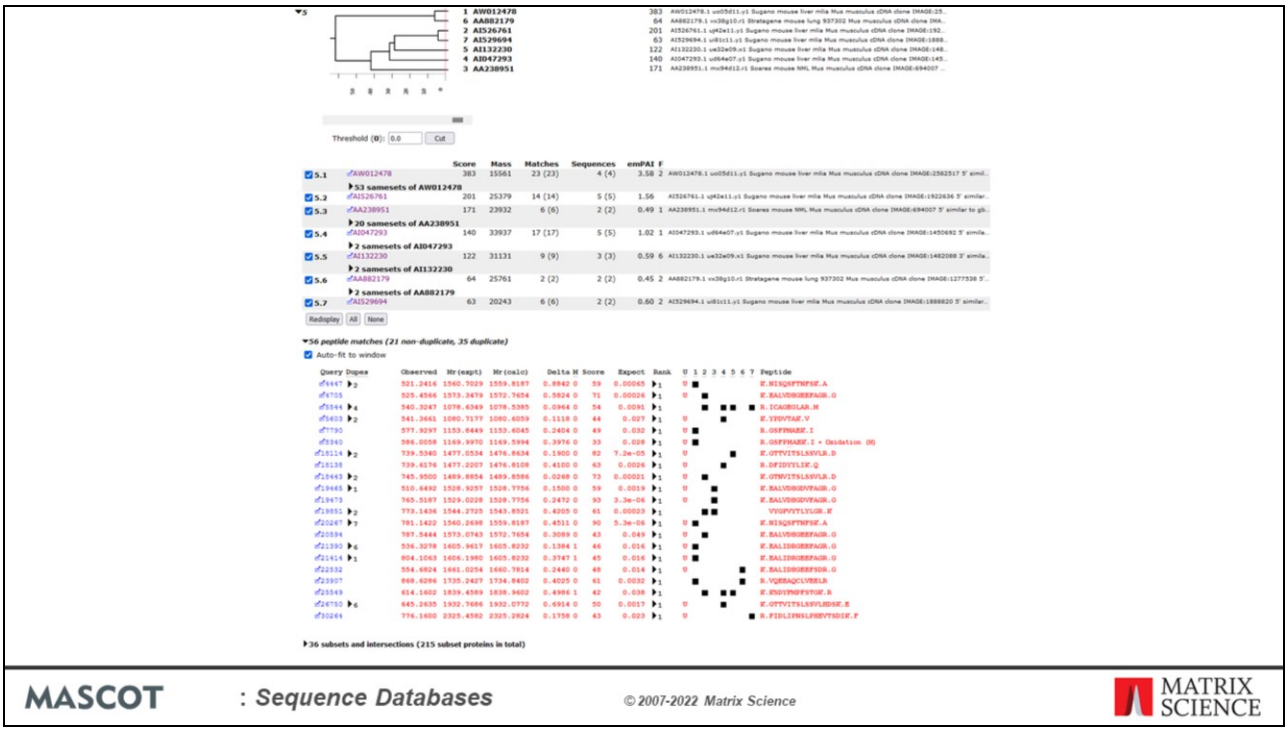


When Unigene is configured, we can select Mus_musculus from the drop-down list in the format controls



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.



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 | Unassigned (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 2 FO710191	654 83	Hspa5 Heat shock protein 5 FO710191.1 Mus musculus mRNA 5-prime sequence 3030000046609839.
3	Mm.14796	624	Mgst1 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_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)			Cyp2c29 Cytochrome P450, family 2, subfamily c, polypeptide 29

40 peptide matches (12 non-duplicate, 28 duplicate)

Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	Peptide
4447	2	521.2416	1560.7029	1559.8187	0.8842	0	59	0.00065	1	U	K.NISQSFTNFSK.A
5544	4	540.3247	1078.6349	1078.5385	0.0964	0	54	0.0091	1	U	R.ICAGEGLAR.M
5603	2	541.3661	1080.7177	1080.6059	0.1118	0	44	0.027	1	U	K.YPDVTAK.V
7790		577.9297	1153.8449	1153.6045	0.2404	0	49	0.032	1	U	R.GSFFMAEK.M
8340		586.0058	1169.9970	1169.5994	0.3976	0	33	0.028	1	U	R.GSFFMAEK.M + Oxidation (M)
18138		739.6176	1477.2207	1476.8108	0.4100	0	63	0.0026	1	U	R.DFIDYYLIK.Q
20267	7	781.1422	1560.2698	1559.8187	0.4511	0	90	5.3e-06	1	U	K.NISQSFTNFSK.A
21390	6	536.3278	1605.9617	1605.8232	0.1384	1	46	0.016	1	U	K.EALIDRGEEFAGR.G
21414	1	804.1063	1606.1980	1605.8232	0.3747	1	45	0.016	1	U	K.EALIDRGEEFAGR.G
23907		868.6286	1735.2427	1734.8402	0.4025	0	61	0.0032	1	U	R.VQEEAQCLVEELR.R
25549		614.1602	1839.4589	1838.9602	0.4986	1	42	0.038	1	U	K.KSDYFMPFSTGK.R
26750	6	645.2635	1932.7686	1932.0772	0.6914	0	50	0.0017	1	U	K.GTTVITSLSSVLHDSK.E

1 subset or intersection (1 subset protein in total)

MASCOT : Sequence Databases © 2007-2022 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 ~ 2.2×10^9 bases)
- 5.4×10^9 residues in 6 frame translation
- 99.75% of translated sequence is non-coding
- ~ 1.5×10^5 tryptic limit peptides of 1500 Da ± 0.5
- ~ 6×10^7 no-enzyme peptides of 1500 Da ± 0.5

MASCOT : *Sequence Databases* MATRIX SCIENCE

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.

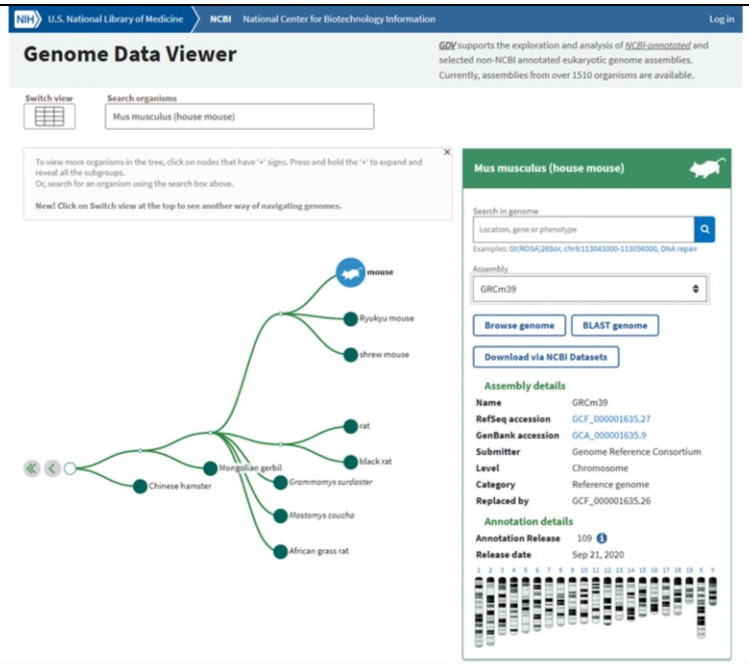
The mouse genome assembly is approximately 2.7 billion bases, which makes it a little larger than Mus_EST.

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

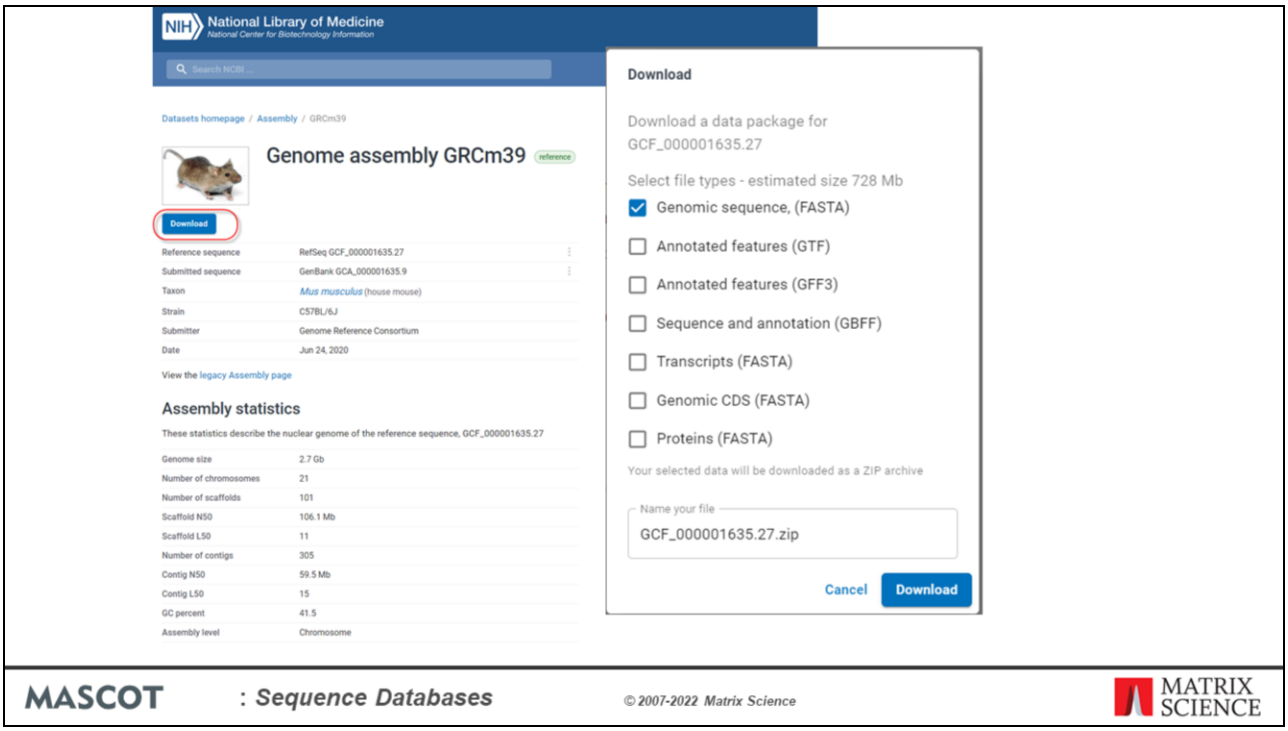
In the mouse genome, only 1.5% of the sequence codes for proteins. This means that 99.75% of the 6 frame translation is non-coding and simply contributes to the background of random matches. This is a good test of the discrimination of the scoring scheme.

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!

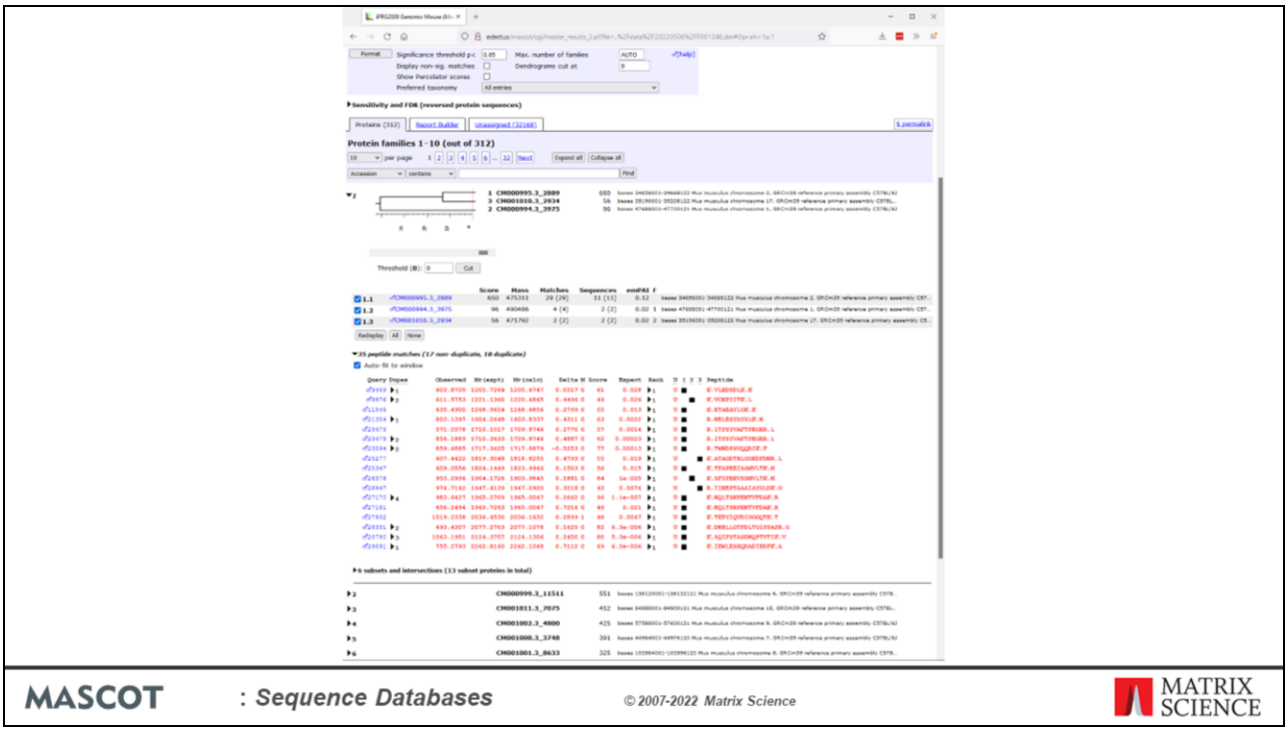


You can download the mouse genome sequences from NCBI.

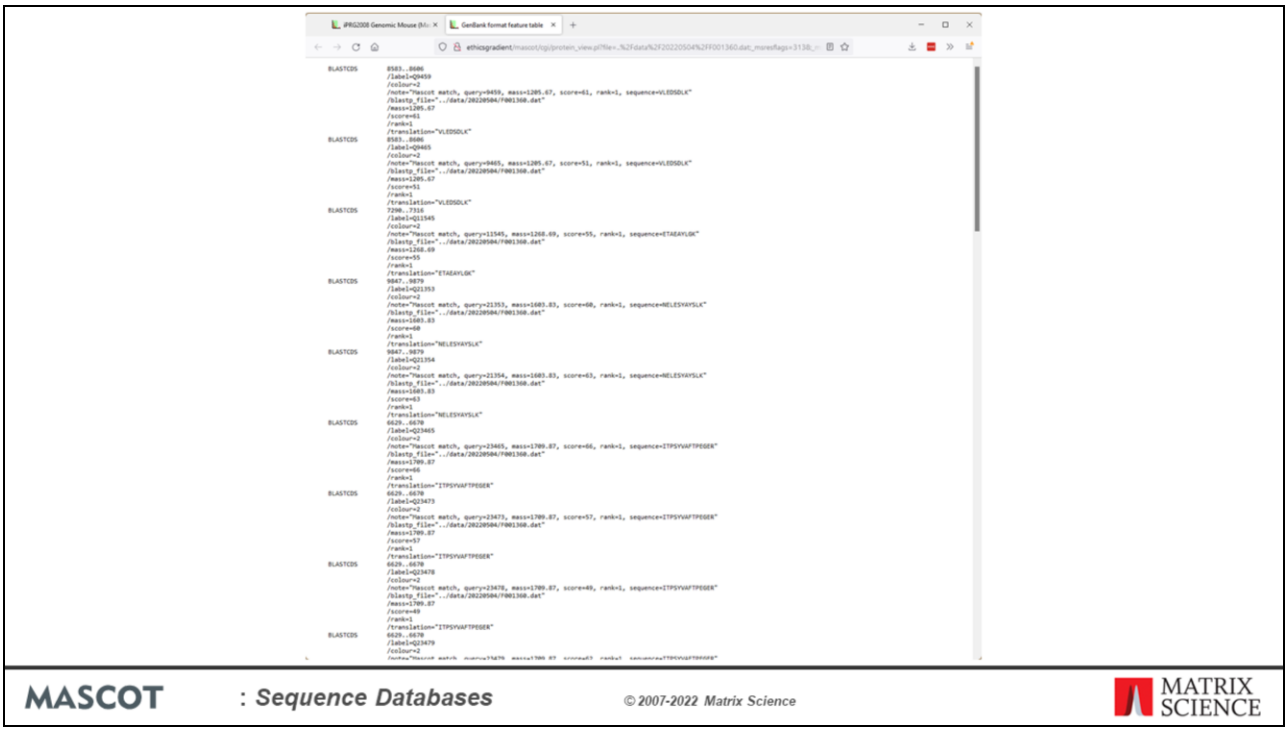


We chose the assembled chromosomes, 24 files. Although you could search this as a 24 entry database, this is not memory efficient, so we used the script mentioned earlier to split the chromosome sequences into overlapping segments of 12 kb

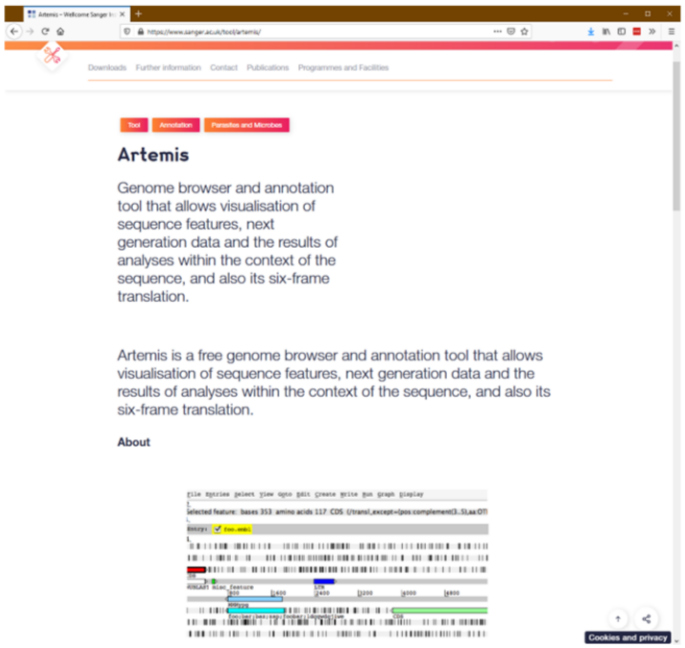
This is the result of searching our data against the mouse genome assembly. If you thought the Mus_EST entry titles were uninformative, how much worse is this?



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

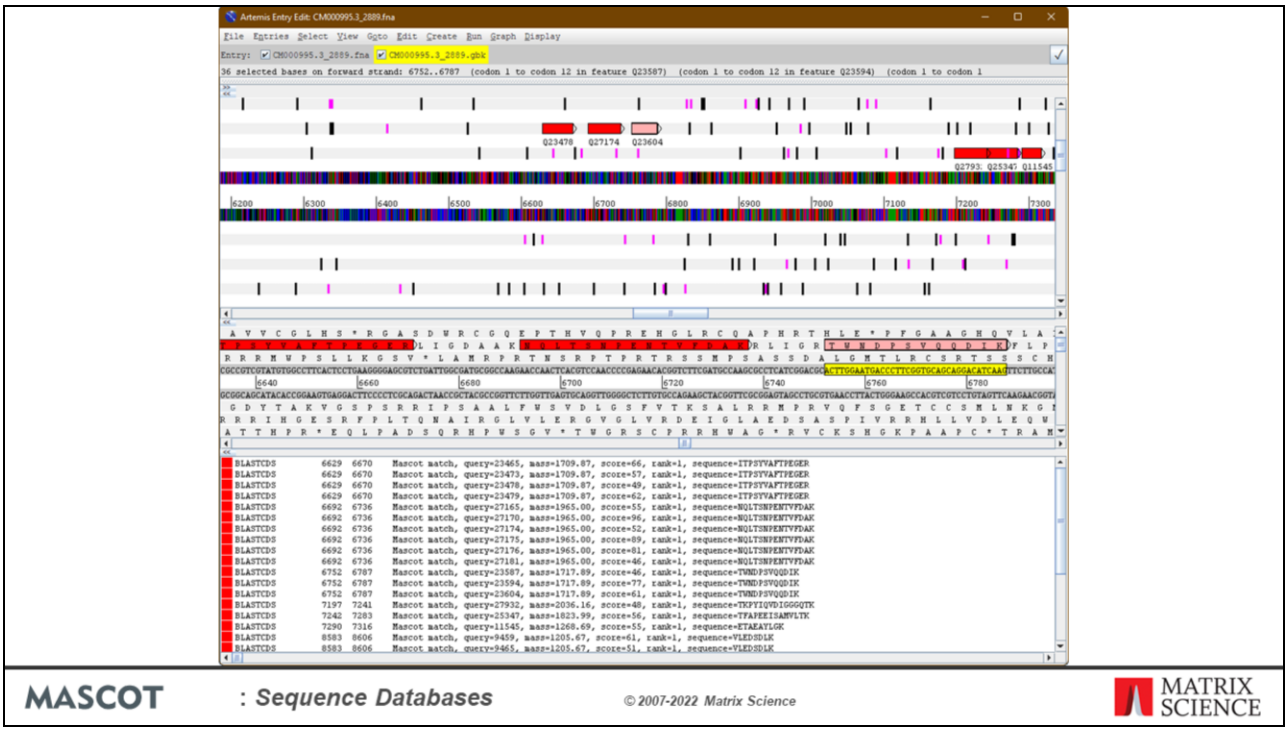


This is the peptide match results formatted as an EMBL / GenBank format feature table. This may not look very friendly, but the advantage of this report is that it can be read into a standard genome browser.
To enable this feature add the “FeatureTableLength” parameter to the options section of the mascot.dat 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.



MASCOT : *Sequence Databases* 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	: iTRAQ4plex (K), iTRAQ4plex (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 in residues	Average score threshold	Number of PSMs (1% FDR)
Uniprot mouse	2. 8 x 10^7	37	1834
EST mouse	4.5 x 10^9	59	675
Mouse genome	5.5 x 10^9	59	548

MASCOT : *Sequence Databases* 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.

neXtProt *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