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

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

For specialised databases, such as individual genomes, you may have to track down the FTP site of the group that is doing the sequencing.
There are a huge number of database, and often it is not clear which is the appropriate one to choose for a search.

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

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

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

Single genome databases are good for protein characterisation using MS/MS data. You may want to include a contaminants database in the search, to ensure spectra from contaminants don’t get mis-assigned to the target organism.
When we search a nucleic acid databases, Mascot always performs a 6 frame translation on the fly. That is, 3 reading frames from the forward strand and 3 reading frames from the complementary strand.
NA Translation

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

The rules for NA translation in Mascot are
Translate the entire sequence, don't look for a start codon to begin
When a stop codon is encountered, leave a gap, and immediately re-start translation
There is no attempt to resolve ambiguous codons. For example, ACX can be translated as Threonine, because the identity of the last base is a don't care. However, this is not done in Mascot.
Finally, all translations use the correct genetic code, as long as the taxonomy is known.
Which database?

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

Let's look at a typical small genome database.

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

Assembled genomic DNA sequence
  h_pylori_26695.1con
Nucleic acid coding sequences
  h_pylori_26695.seq
Coding sequences translated to proteins
  h_pylori_26695.pep
Table of co-ordinates for the coding sequences in the assembled chromosome
  h_pylori_26695.coords

Follow the links for Helicobacter pylori, and these files are available for download.
  • Assembled genomic DNA sequence
  • Nucleic acid coding sequences
  • Coding sequences translated to proteins
  • Table of co-ordinates for the coding sequences in the assembled chromosome

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

If you are not confident that the coding sequences and reading frames have been identified correctly, then you might wish to search the genomic DNA directly. Helicobacter pylori has a single chromosome, so h_pylori_26695.1con contains just one sequence, of length 1,667,867 bases.
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 small dataset of a few hundred MS/MS spectra against a protein database, IPI human. We found significant matches to 28 human proteins.
With EST_human, we obtained almost the same results, just a couple of additional peptide matches. However, look at the hit-list on this report … unlike the protein database search, it doesn’t immediately communicate which proteins have been found. I’ll return to this issue later.
The master results report from the EST search looks pretty similar to the IPI search, except that the EST sequences are mostly shorter than full length proteins, so the peptide matches are more scattered. If we click on the protein accession number link...
We get a protein view. This is similar to the protein view for a protein database entry, except we have drop down list for the different translation frames. For this particular entry, most of the matches have been found in reading frame 1.

But, as so often happens, there is a frame shift in this entry, and there are additional matches in frame 3.
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. The database update script, which will be covered in a later session, takes care of this automatically.
To use a unigene index in Mascot, the data file for the species is downloaded and unpacked into a suitable directory structure.

Then a few lines are added to the Mascot configuration file, mascot.dat.
So, if Unigene is configured, we can select human 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.
When we look at individual hits in the report, we see the benefits of UniGene mapping. Here we have two hits from the EST search. They have no peptide matches in common, and the entry names also give no clue as to the protein function. However, when we look at the UniGene report, we find that these matches all belong to the same gene, for alkaline phosphatase.
When you click on the accession number link of a unigene filtered report, you get full details for that particular gene family.
We can also perform MS/MS searches on the raw genomic sequence data. Let’s just look at some numbers for the assembled human genome.

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

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

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

If we are matching MS/MS data from a tryptic peptide of nominal mass 1500 Da against the human genome, we are going to have to test 150 thousand peptides. Which sounds bad, but is not nearly as bad as the no-enzyme case where we have to test 60 million!
You can download the human 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 human genome assembly. If you thought the EST human 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.
For example, one which we find works well is Artemis, a Java based genome browser developed and distributed by the Sanger Centre.
Here’s the result of reading the feature table containing the Mascot peptide matches into Artemis. In the upper third, we have a low resolution view. This can be zoomed out to show an entire sequence as a single strip. We have the forward and complementary DNA strands, and the 6 frame translation. The vertical bars are stop codons. The yellow blocks are exons, while the blue blocks here are coding sequences. Individual Mascot peptide matches are shown in red. This particular gene has 8 peptide matches.

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

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

Not only does this allow us to zoom and pan around these extremely long sequences, it also allows us to view the peptide matches found by Mascot in the context of all the existing annotations. This gives us a powerful way to present the results of MS based searching complete genomes.
All well and good, but which database gives the most matches? We searched a much larger dataset against all 3 databases. The data was the public iPRG2010 dataset distributed by ABRF.

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

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

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

References

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

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