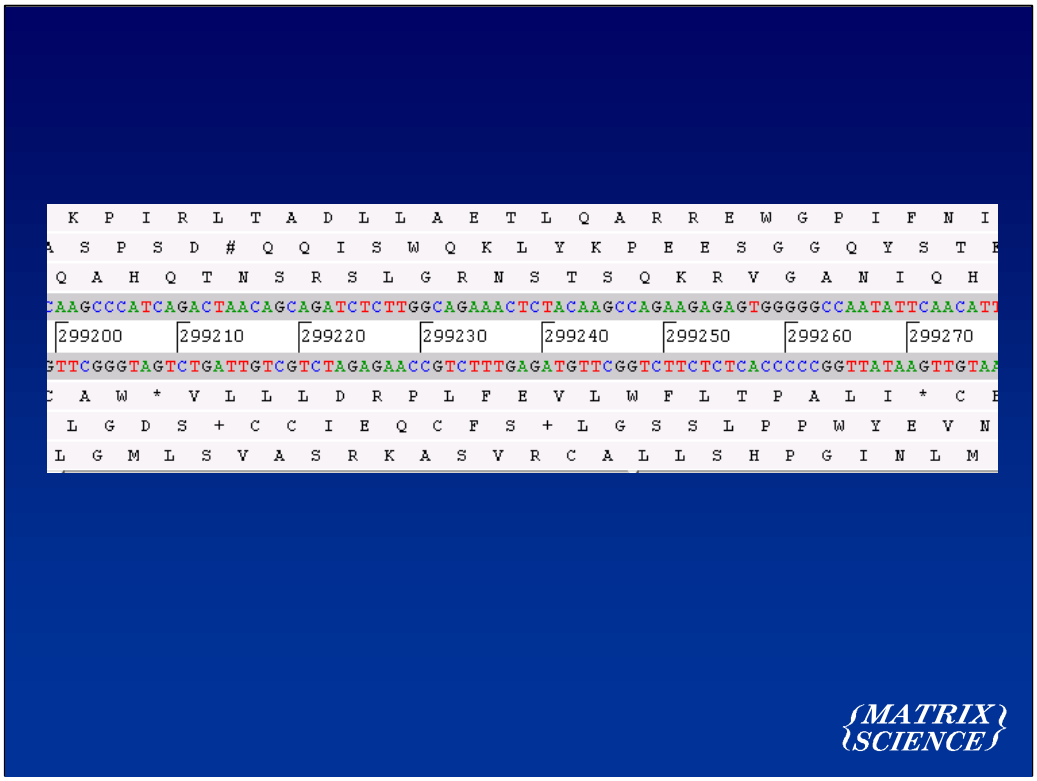
Searching Nucleotide Databases

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.

NCBI Standard Genetic Code

```
Residue: FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Start:   ----------------------------------M---------------------------
Base 1:  TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base 2:  TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base 3:  TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG

* = stop
```

MATRIX SCIENCE

Translation uses the standard genetic code, shown here.

NA Translation

- Always translate in all 6 reading frames
- Translation starts from the beginning of the sequence, not from a start codon
- When a stop codon is encountered, insert a gap and re-start translation.
- No attempt to resolve codon ambiguity.
- All translation uses the NCBI standard genetic code.

MATRIX SCIENCE

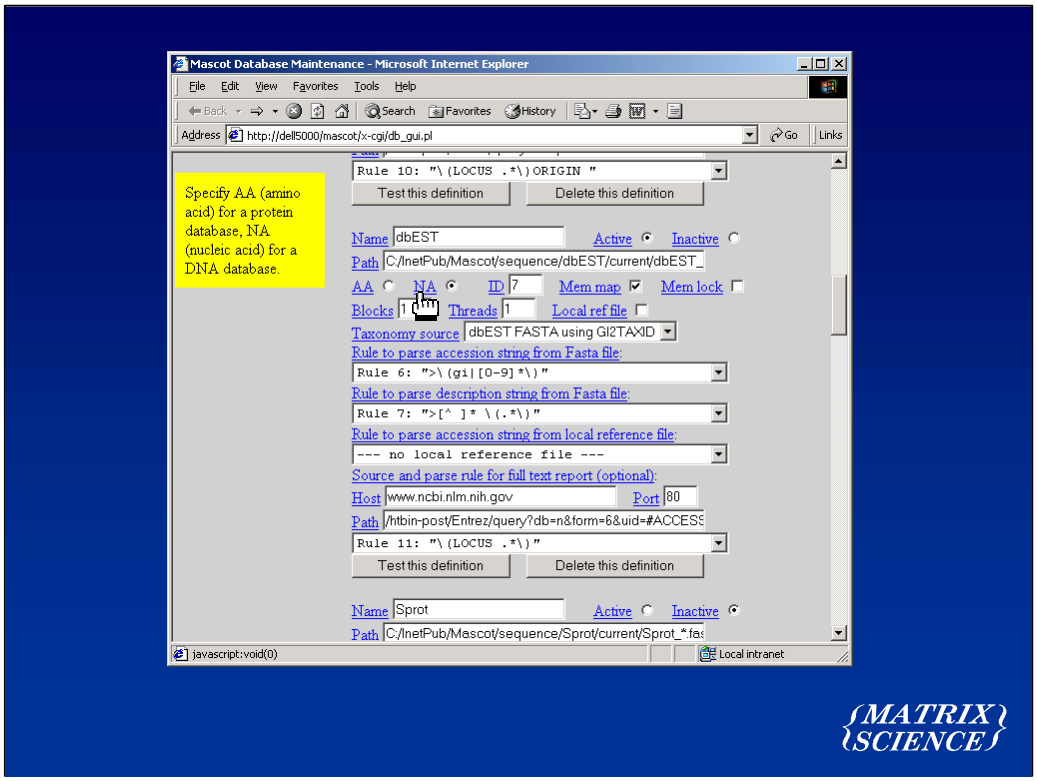
The rules for Nucleic Acid translation in Mascot are:

we translate the entire sequence, we don't look for a start codon.

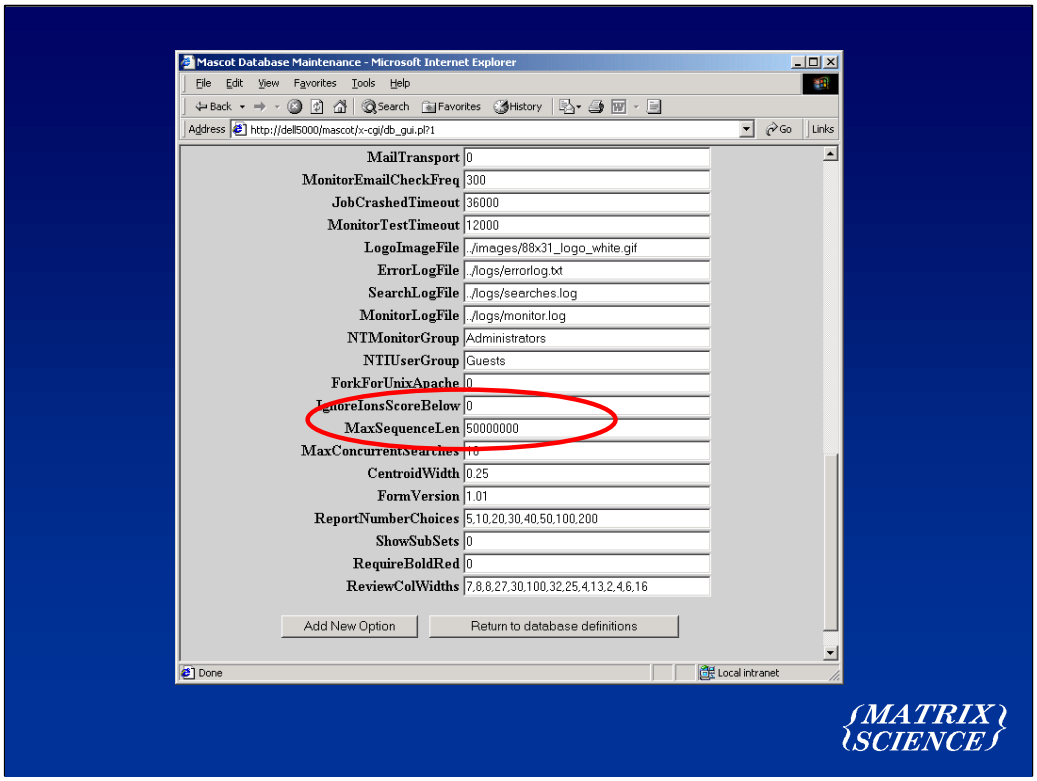
When a stop codon is encountered, we leave a gap, and immediately re-start translation.

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 the current code.

Finally, all translations use the standard genetic code. Ideally, we would use species specific code where a sequence has a known taxonomy. But, again, this is not done at present.



Setting up a nucleic acid database in Mascot is no different from setting up a protein database. The only two things to watch are that the database type is specified as NA



And, if the sequences are very long, you may need to increase the upper limit on the sequence length of individual entries.

Standard data set

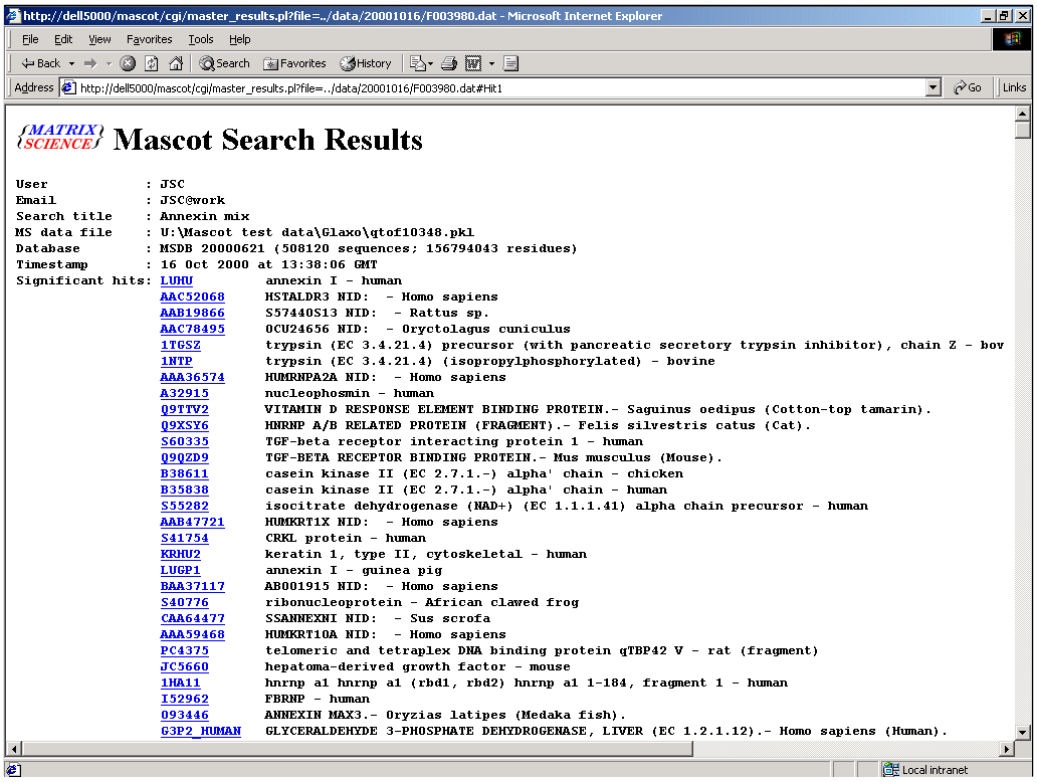
- **Collaboration with Walter Blackstock and Jyoti Choudhary, Cell Map Project, GlaxoSmithKline R&D, Stevenage, UK**
- **Human embryonic kidney cell lysate; digested with trypsin; analysed by LC-MS/MS using a Micromass Q-TOF**
- **169 MS/MS spectra after data reduction (.pkl file)**

MATRIX SCIENCE

Here are examples of searching the same data set against protein, EST and genomic sequence databases. This data set was generated by Jyoti Choudhary and Walter Blackstock of GlaxoSmithKline.

They generated a high quality LC-MS/MS data from a tryptic digest of whole cell lysate from human embryonic kidney cells.

After data reduction, we were left with 169 MS/MS spectra.



http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F003980.dat - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

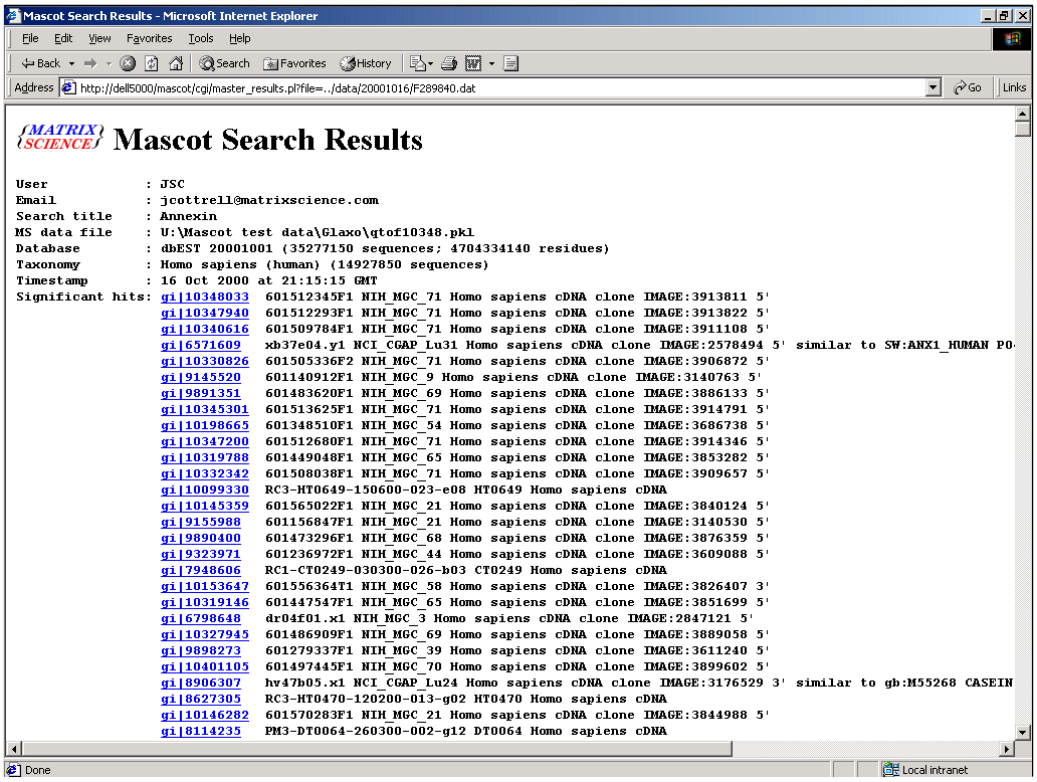
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F003980.dat#Hit1 Go Links

Mascot Search Results

```
User             : JSC
Email            : JSC@work
Search title     : Annexin mix
MS data file     : U:\Mascot test data\Glaxo\qtof10348.pkl
Database         : MSDB 20000621 (508120 sequences; 156794043 residues)
Timestamp        : 16 Oct 2000 at 13:38:06 GMT
Significant hits: LUHU        annexin I - human
                  AAC52068    HSTALDR3 NID:  - Homo sapiens
                  AAB19866    S57440S13 NID:  - Rattus sp.
                  AAC78495    OCU24656 NID:  - Oryctolagus cuniculus
                  1TGS2       trypsin (EC 3.4.21.4) precursor (with pancreatic secretory trypsin inhibitor), chain Z - bov
                  1NTP        trypsin (EC 3.4.21.4) (isopropylphosphorylated) - bovine
                  AAA36574    HUMRNPA2A NID:  - Homo sapiens
                  A32915      nucleophosmin - human
                  Q9TTV2      VITAMIN D RESPONSE ELEMENT BINDING PROTEIN.- Saguinus oedipus (Cotton-top tamarin).
                  Q9XSY6      HNRNP A/B RELATED PROTEIN (FRAGMENT).- Felis silvestris catus (Cat).
                  S60335      TGF-beta receptor interacting protein 1 - human
                  Q9QZD9      TGF-BETA RECEPTOR BINDING PROTEIN.- Mus musculus (Mouse).
                  B38611      casein kinase II (EC 2.7.1.-) alpha' chain - chicken
                  B35838      casein kinase II (EC 2.7.1.-) alpha' chain - human
                  S55282      isocitrate dehydrogenase (NAD+) (EC 1.1.1.41) alpha chain precursor - human
                  AAB47721    HUMKRT1X NID:  - Homo sapiens
                  S41754      CRKL protein - human
                  KRHU2       keratin 1, type II, cytoskeletal - human
                  LUGP1       annexin I - guinea pig
                  BAA37117    AB001915 NID:  - Homo sapiens
                  S40776      ribonucleoprotein - African clawed frog
                  CAA64477    SSANNEXNI NID:  - Sus scrofa
                  AAA59468    HUMKRT10A NID:  - Homo sapiens
                  PC4375      telomeric and tetraplex DNA binding protein qTBP42 V - rat (fragment)
                  JC5660      hepatoma-derived growth factor - mouse
                  1HA11       hnrnp a1 hnrnp a1 (rbd1, rbd2) hnrnp a1 1-184, fragment 1 - human
                  I52962      FBRNP - human
                  O93446      ANNEXIN MAX3.- Oryzias latipes (Medaka fish).
                  G3P2_HUMAN  GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (EC 1.2.1.12).- Homo sapiens (Human).
```

Local intranet

First of all, we searched the data against a comprehensive non-identical protein database, MSDB. We found significant matches to 22 human proteins … and one non-human, our frequent flyer, bovine trypsin.



Mascot Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

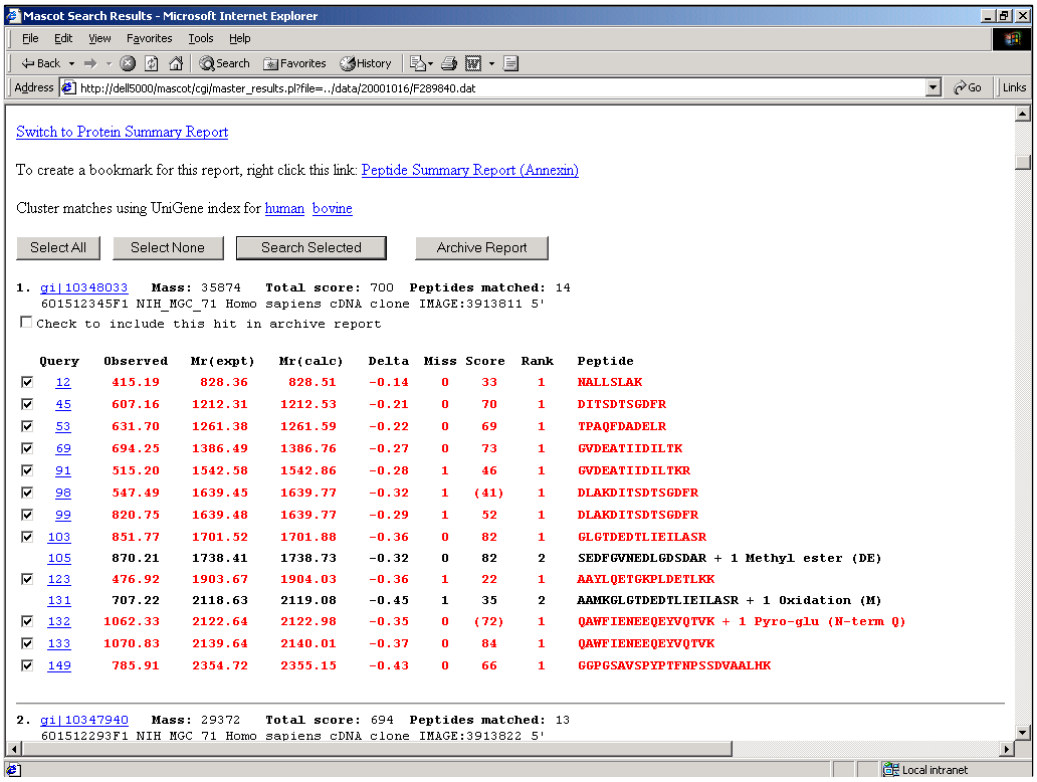
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F289840.dat Go Links

Mascot Search Results

```
User             : JSC
Email            : jcottrell@matrixscience.com
Search title     : Annexin
MS data file     : U:\Mascot test data\Glaxo\qtof10348.pkl
Database         : dbEST 20001001 (35277150 sequences; 4704334140 residues)
Taxonomy         : Homo sapiens (human) (14927850 sequences)
Timestamp        : 16 Oct 2000 at 21:15:15 GMT
Significant hits: gi|10348033  601512345F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3913811 5'
                  gi|10347940  601512293F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3913822 5'
                  gi|10340616  601509784F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3911108 5'
                  gi|6571609   xb37e04.y1 NCI_CGAP_Lu31 Homo sapiens cDNA clone IMAGE:2578494 5' similar to SW:ANX1_HUMAN P0
                  gi|10330826  601505336F2 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3906872 5'
                  gi|9145520   601140912F1 NIH_MGC_9 Homo sapiens cDNA clone IMAGE:3140763 5'
                  gi|9891351   601483620F1 NIH_MGC_69 Homo sapiens cDNA clone IMAGE:3886133 5'
                  gi|10345301  601513625F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3914791 5'
                  gi|10198665  601348510F1 NIH_MGC_54 Homo sapiens cDNA clone IMAGE:3686738 5'
                  gi|10347200  601512680F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3914346 5'
                  gi|10319788  601449048F1 NIH_MGC_65 Homo sapiens cDNA clone IMAGE:3853282 5'
                  gi|10332342  601508038F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3909657 5'
                  gi|10099330  RC3-HT0649-150600-023-e08 HT0649 Homo sapiens cDNA
                  gi|10145359  601565022F1 NIH_MGC_21 Homo sapiens cDNA clone IMAGE:3840124 5'
                  gi|9155988   601156847F1 NIH_MGC_21 Homo sapiens cDNA clone IMAGE:3140530 5'
                  gi|9890400   601473296F1 NIH_MGC_68 Homo sapiens cDNA clone IMAGE:3876359 5'
                  gi|9323971   601236972F1 NIH_MGC_44 Homo sapiens cDNA clone IMAGE:3609088 5'
                  gi|7948606   RC1-CT0249-030300-026-b03 CT0249 Homo sapiens cDNA
                  gi|10153647  601556364T1 NIH_MGC_58 Homo sapiens cDNA clone IMAGE:3826407 3'
                  gi|10319146  601447547F1 NIH_MGC_65 Homo sapiens cDNA clone IMAGE:3851699 5'
                  gi|6798648   dr04f01.x1 NIH_MGC_3 Homo sapiens cDNA clone IMAGE:2847121 5'
                  gi|10327945  601486909F1 NIH_MGC_69 Homo sapiens cDNA clone IMAGE:3889058 5'
                  gi|9898273   601279337F1 NIH_MGC_39 Homo sapiens cDNA clone IMAGE:3611240 5'
                  gi|10401105  601497445F1 NIH_MGC_70 Homo sapiens cDNA clone IMAGE:3899602 5'
                  gi|8906307   hv47b05.x1 NCI_CGAP_Lu24 Homo sapiens cDNA clone IMAGE:3176529 3' similar to gb:M55268 CASEIN
                  gi|8627305   RC3-HT0470-120200-013-g02 HT0470 Homo sapiens cDNA
                  gi|10146282  601570283F1 NIH_MGC_21 Homo sapiens cDNA clone IMAGE:3844988 5'
                  gi|8114235   PM3-DT0064-260300-002-g12 DT0064 Homo sapiens cDNA
```

Done Local intranet

With dbEST, we obtained almost the same results, just a couple of additional peptide matches. However, unlike the protein database search, it doesn't immediately communicate which proteins have been found.



Mascot Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F289840.dat Go Links

Switch to Protein Summary Report

To create a bookmark for this report, right click this link: Peptide Summary Report (Annexin)

Cluster matches using UniGene index for human bovine

Select All | Select None | Search Selected | Archive Report

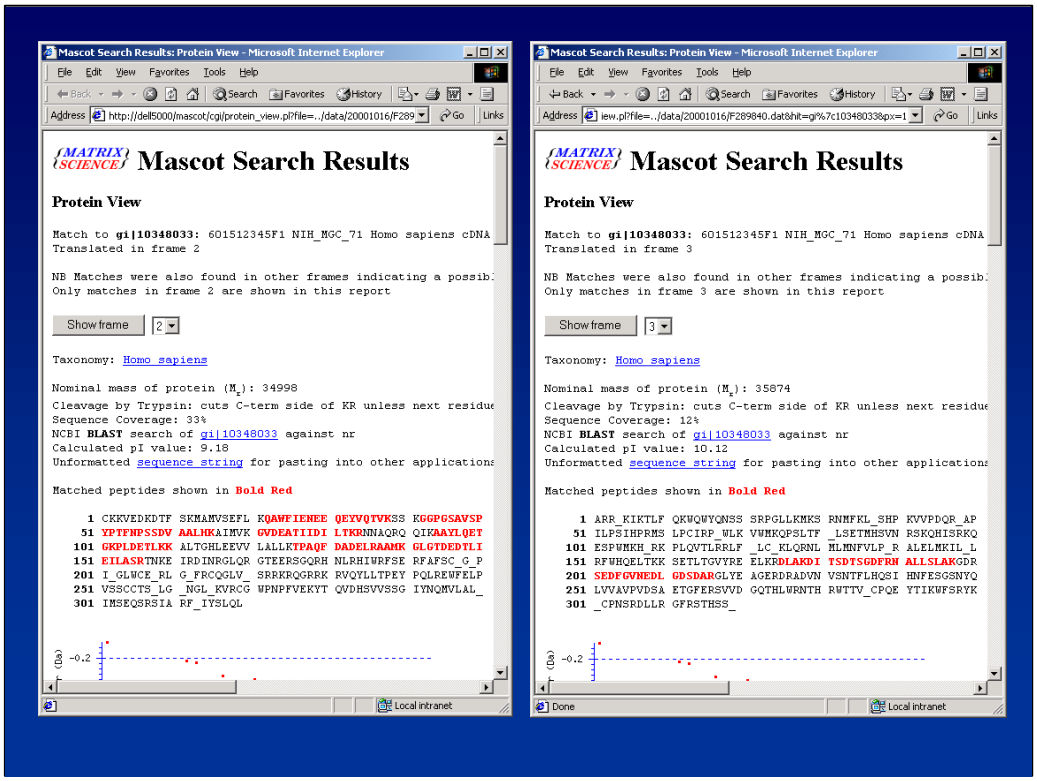
1. gi|10348033 **Mass:** 35874 **Total score:** 700 **Peptides matched:** 14
601512345F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3913811 5'
Check to include this hit in archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
12	415.19	828.36	828.51	-0.14	0	33	1	NALLSLAK
45	607.16	1212.31	1212.53	-0.21	0	70	1	DITSDTSGDFR
53	631.70	1261.38	1261.59	-0.22	0	69	1	TPAQFDADELR
69	694.25	1386.49	1386.76	-0.27	0	73	1	GVDEATIIDILTK
91	515.20	1542.58	1542.86	-0.28	1	46	1	GVDEATIIDILTKR
98	547.49	1639.45	1639.77	-0.32	1	(41)	1	DLAKDITSDTSGDFR
99	820.75	1639.48	1639.77	-0.29	1	52	1	DLAKDITSDTSGDFR
103	851.77	1701.52	1701.88	-0.36	0	82	1	GLGTDEDTLIEILASR
105	870.21	1738.41	1738.73	-0.32	0	82	2	SEDFGVNEDLGDSDAR + 1 Methyl ester (DE)
123	476.92	1903.67	1904.03	-0.36	1	22	1	AAYLQETGKPLDETLKK
131	707.22	2118.63	2119.08	-0.45	1	35	2	AAMKGLGTDEDTLIEILASR + 1 Oxidation (M)
132	1062.33	2122.64	2122.98	-0.35	0	(72)	1	QAWFIENEEQEYVQTVK + 1 Pyro-glu (N-term Q)
133	1070.83	2139.64	2140.01	-0.37	0	84	1	QAWFIENEEQEYVQTVK
149	785.91	2354.72	2355.15	-0.43	0	66	1	GGPGSAVSPYPTFNPSSDVAALHK

2. gi|10347940 **Mass:** 29372 **Total score:** 694 **Peptides matched:** 13
601512293F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3913822 5'

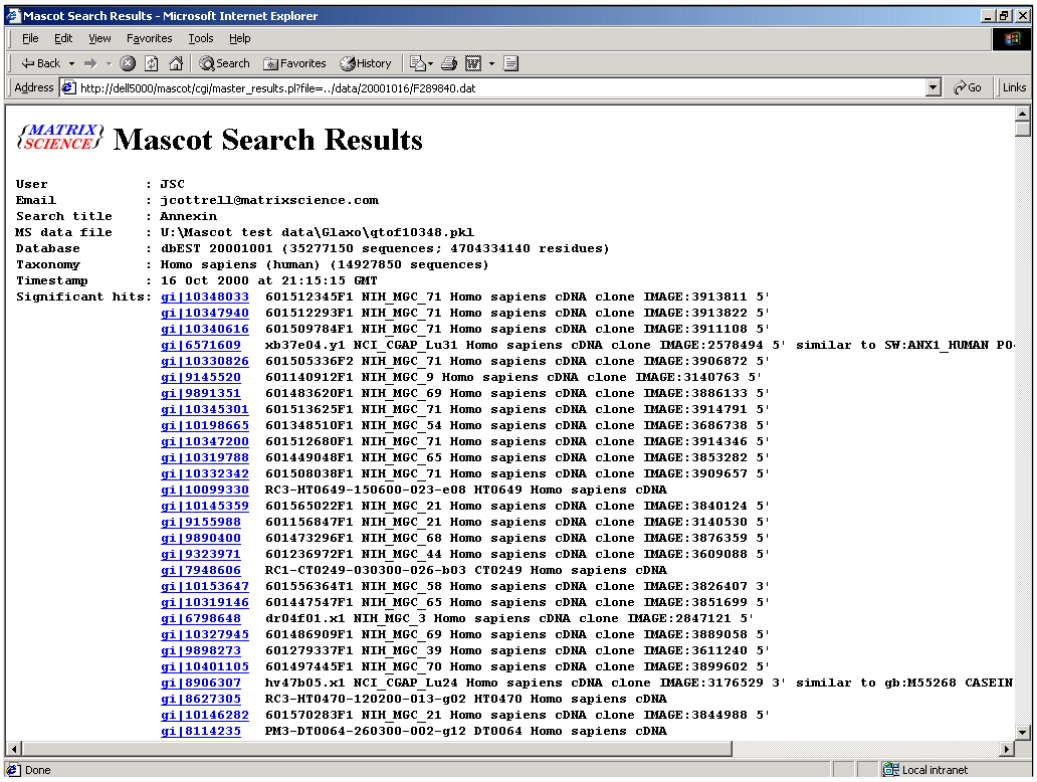
Local intranet

The master results report from the EST search looks pretty similar to the MSDB 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 2.

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



Mascot Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

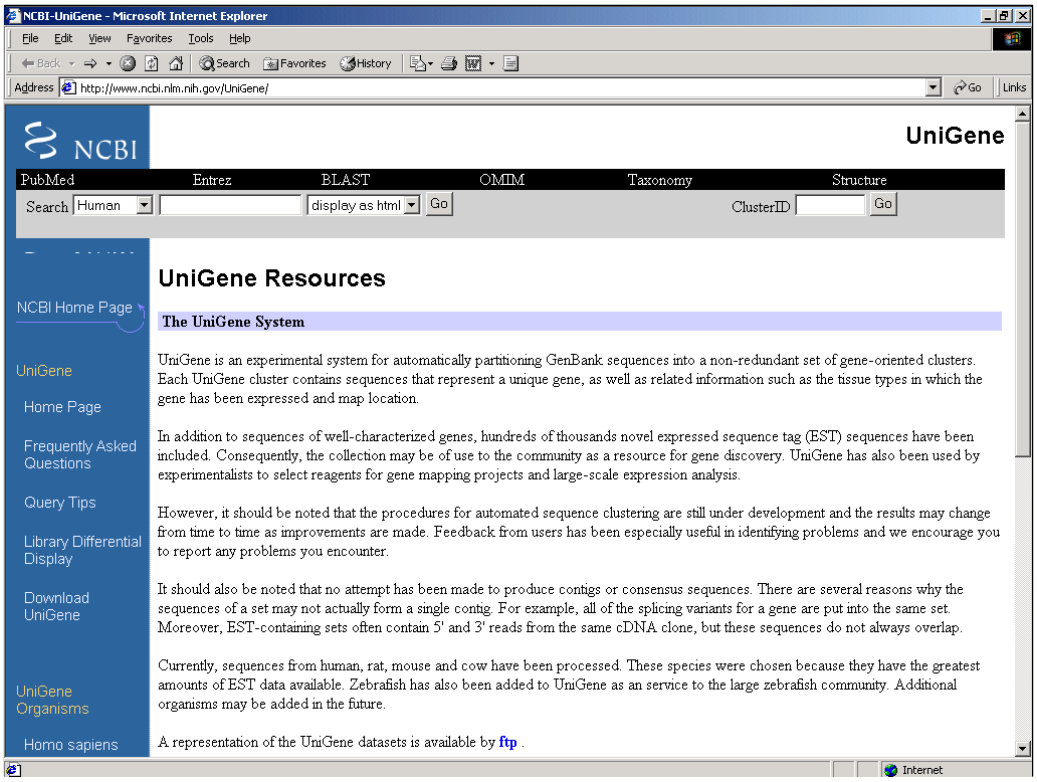
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F289840.dat Go Links

Mascot Search Results

```
User             : JSC
Email            : jcottrell@matrixscience.com
Search title     : Annexin
MS data file     : U:\Mascot test data\Glaxo\qtof10348.pkl
Database         : dbEST 20001001 (35277150 sequences; 4704334140 residues)
Taxonomy         : Homo sapiens (human) (14927850 sequences)
Timestamp        : 16 Oct 2000 at 21:15:15 GMT
Significant hits: gi|10348033  601512345F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3913811 5'
                  gi|10347940  601512293F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3913822 5'
                  gi|10340616  601509784F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3911108 5'
                  gi|6571609   xb37e04.y1 NCI_CGAP_Lu31 Homo sapiens cDNA clone IMAGE:2578494 5' similar to SW:ANX1_HUMAN P0
                  gi|10330826  601505336F2 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3906872 5'
                  gi|9145520   601140912F1 NIH_MGC_9 Homo sapiens cDNA clone IMAGE:3140763 5'
                  gi|9891351   601483620F1 NIH_MGC_69 Homo sapiens cDNA clone IMAGE:3886133 5'
                  gi|10345301  601513625F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3914791 5'
                  gi|10198665  601348510F1 NIH_MGC_54 Homo sapiens cDNA clone IMAGE:3686738 5'
                  gi|10347200  601512680F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3914346 5'
                  gi|10319788  601449048F1 NIH_MGC_65 Homo sapiens cDNA clone IMAGE:3853282 5'
                  gi|10332342  601508038F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3909657 5'
                  gi|10099330  RC3-HT0649-150600-023-e08 HT0649 Homo sapiens cDNA
                  gi|10145359  601565022F1 NIH_MGC_21 Homo sapiens cDNA clone IMAGE:3840124 5'
                  gi|9155988   601156847F1 NIH_MGC_21 Homo sapiens cDNA clone IMAGE:3140530 5'
                  gi|9890400   601473296F1 NIH_MGC_68 Homo sapiens cDNA clone IMAGE:3876359 5'
                  gi|9323971   601236972F1 NIH_MGC_44 Homo sapiens cDNA clone IMAGE:3609088 5'
                  gi|7948606   RC1-CT0249-030300-026-b03 CT0249 Homo sapiens cDNA
                  gi|10153647  601556364T1 NIH_MGC_58 Homo sapiens cDNA clone IMAGE:3826407 3'
                  gi|10319146  601447547F1 NIH_MGC_65 Homo sapiens cDNA clone IMAGE:3851699 5'
                  gi|6798648   dr04f01.x1 NIH_MGC_3 Homo sapiens cDNA clone IMAGE:2847121 5'
                  gi|10327945  601486909F1 NIH_MGC_69 Homo sapiens cDNA clone IMAGE:3889058 5'
                  gi|9898273   601279337F1 NIH_MGC_39 Homo sapiens cDNA clone IMAGE:3611240 5'
                  gi|10401105  601497445F1 NIH_MGC_70 Homo sapiens cDNA clone IMAGE:3899602 5'
                  gi|8906307   hv47b05.x1 NCI_CGAP_Lu24 Homo sapiens cDNA clone IMAGE:3176529 3' similar to gb:M55268 CASEIN
                  gi|8627305   RC3-HT0470-120200-013-g02 HT0470 Homo sapiens cDNA
                  gi|10146282  601570283F1 NIH_MGC_21 Homo sapiens cDNA clone IMAGE:3844988 5'
                  gi|8114235   PM3-DT0064-260300-002-g12 DT0064 Homo sapiens cDNA
```

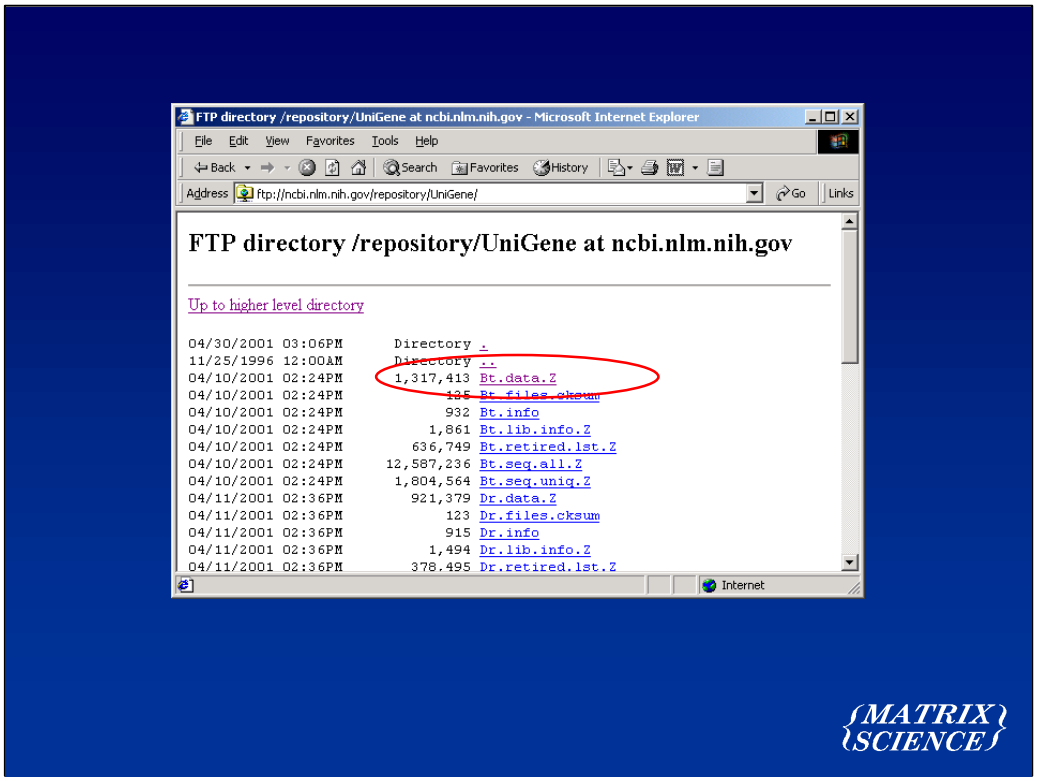
Done Local intranet

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.

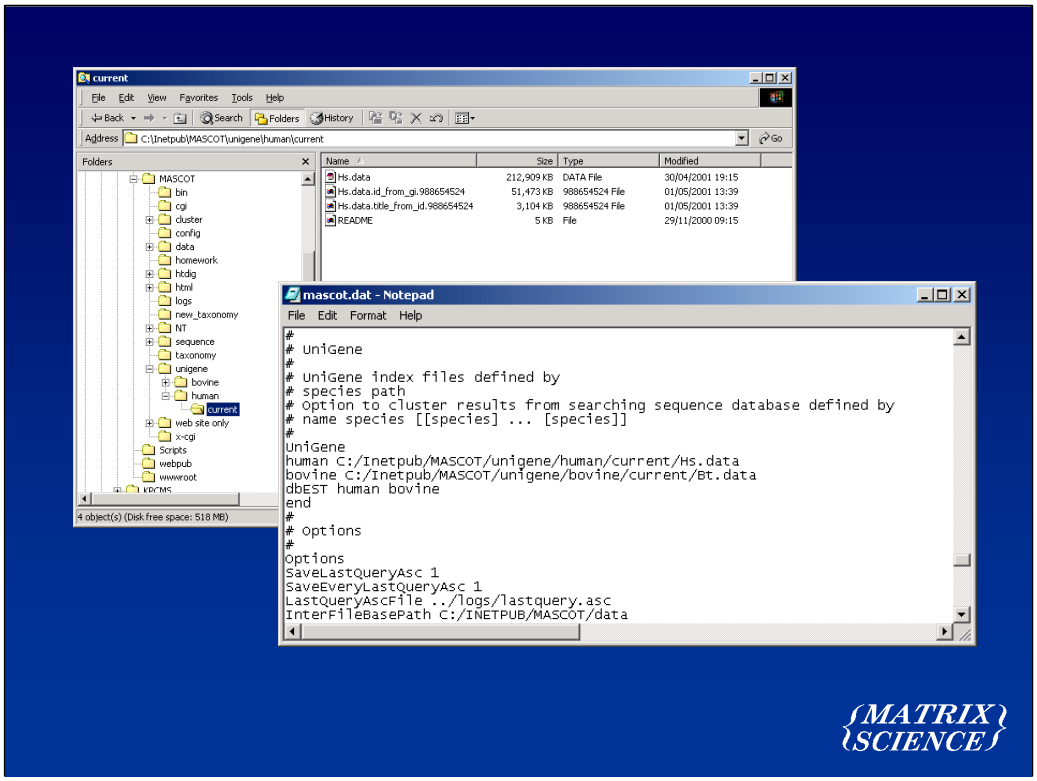


To address this problem, we have recently started using the UniGene index from the National Center for Biotechnology Information to simplify the search results.

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

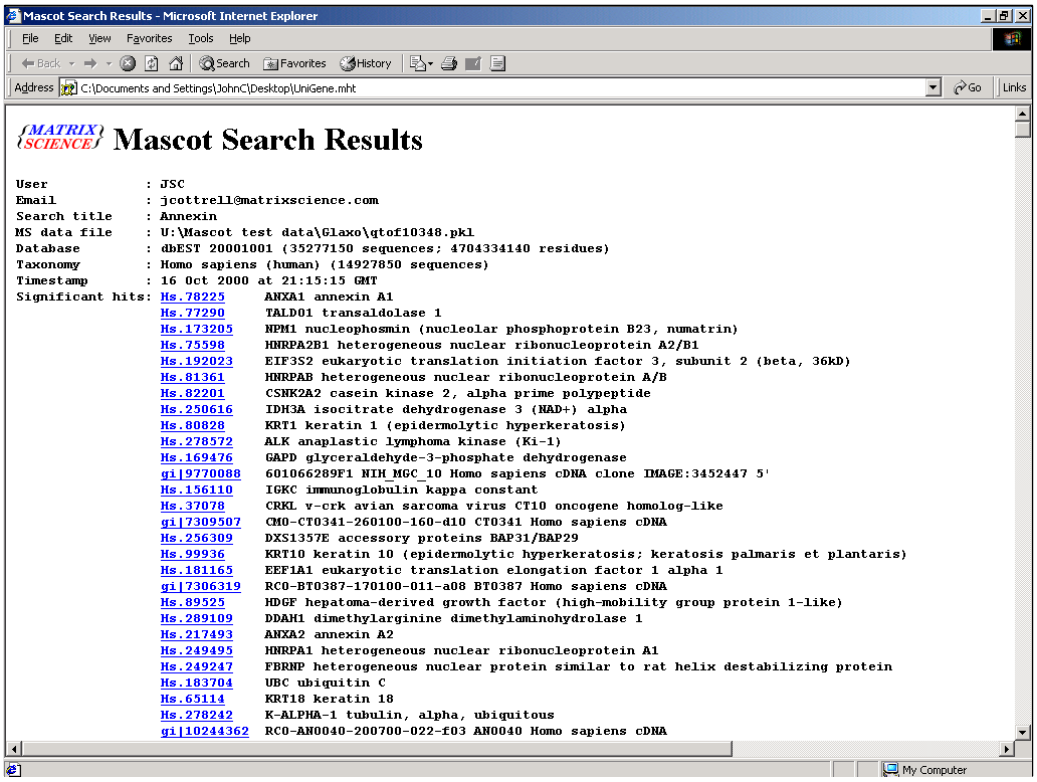


Unigene can be downloaded from the NCBI FTP site. Several important species are available: human, mouse, rat, cow and zebra fish.



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.



Mascot Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

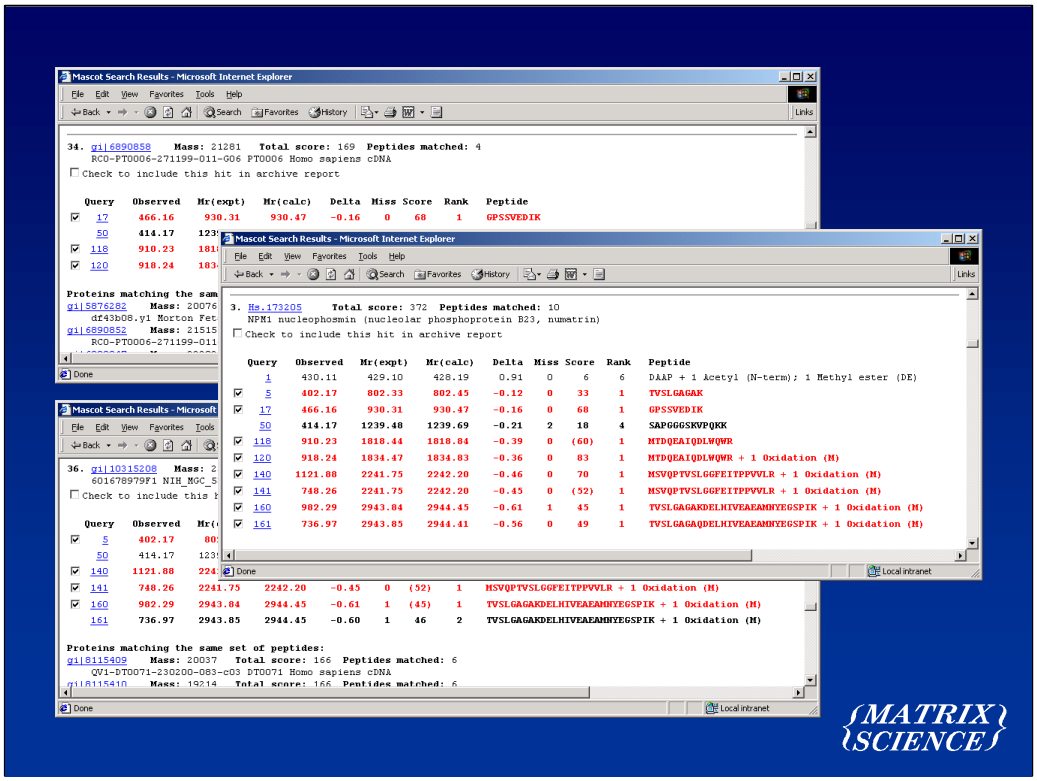
Address C:\Documents and Settings\JohnC\Desktop\UniGene.mht

Mascot Search Results

```
User             : JSC
Email            : jcottrell@matrixscience.com
Search title     : Annexin
MS data file     : U:\Mascot test data\Glaxo\qtof10348.pkl
Database         : dbEST 20001001 (35277150 sequences; 4704334140 residues)
Taxonomy         : Homo sapiens (human) (14927850 sequences)
Timestamp        : 16 Oct 2000 at 21:15:15 GMT
Significant hits: Hs.78225      ANXA1 annexin A1
                  Hs.77290      TALD01 transaldolase 1
                  Hs.173205     NPM1 nucleophosmin (nucleolar phosphoprotein B23, numatrin)
                  Hs.75598      HNRPA2B1 heterogeneous nuclear ribonucleoprotein A2/B1
                  Hs.192023     EIF3S2 eukaryotic translation initiation factor 3, subunit 2 (beta, 36kD)
                  Hs.81361      HNRPAB heterogeneous nuclear ribonucleoprotein A/B
                  Hs.82201      CSNK2A2 casein kinase 2, alpha prime polypeptide
                  Hs.250616     IDH3A isocitrate dehydrogenase 3 (NAD+) alpha
                  Hs.80828      KRT1 keratin 1 (epidermolytic hyperkeratosis)
                  Hs.278572     ALK anaplastic lymphoma kinase (Ki-1)
                  Hs.169476     GAPD glyceraldehyde-3-phosphate dehydrogenase
                  gi|9770088    601066289F1 NIH_MGC_10 Homo sapiens cDNA clone IMAGE:3452447 5'
                  Hs.156110     IGKC immunoglobulin kappa constant
                  Hs.37078      CRKL v-crk avian sarcoma virus CT10 oncogene homolog-like
                  gi|7309507    CM0-CT0341-260100-160-d10 CT0341 Homo sapiens cDNA
                  Hs.256309     DXS1357E accessory proteins BAP31/BAP29
                  Hs.99936      KRT10 keratin 10 (epidermolytic hyperkeratosis; keratosis palmaris et plantaris)
                  Hs.181165     EEF1A1 eukaryotic translation elongation factor 1 alpha 1
                  gi|7306319    RC0-BT0387-170100-011-a08 BT0387 Homo sapiens cDNA
                  Hs.89525      HDGF hepatoma-derived growth factor (high-mobility group protein 1-like)
                  Hs.289109     DDAH1 dimethylarginine dimethylaminohydrolase 1
                  Hs.217493     ANXA2 annexin A2
                  Hs.249495     HNRPA1 heterogeneous nuclear ribonucleoprotein A1
                  Hs.249247     FBRNP heterogeneous nuclear protein similar to rat helix destabilizing protein
                  Hs.183704     UBC ubiquitin C
                  Hs.65114      KRT18 keratin 18
                  Hs.278242     K-ALPHA-1 tubulin, alpha, ubiquitous
                  gi|10244362   RC0-AN0040-200700-022-f03 AN0040 Homo sapiens cDNA
```

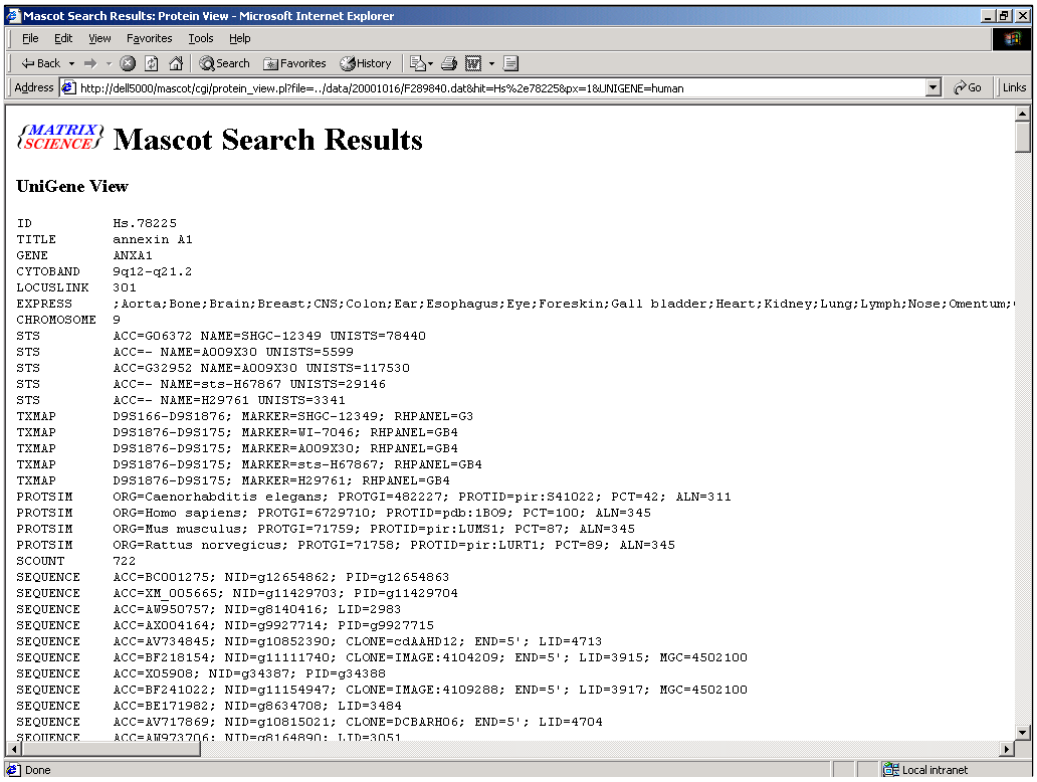
Now, using the UniGene index as a lookup table, we can transform the results of a dbEST search.

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 simplify the results.



When we look further down the report, at details of individual matches, we see the benefits of clustering the ESTs. Here we have two groups of matches from the dbEST search. These matches are well down the report, and appear to have little in common apart from a very weak match to query 50. There is no particular reason to connect these two hits.

However, when we look at the unigene report, we find that these matches all belong to the same gene, nucleophosmin. And, of course, because this gene now has 10 matches, it is listed near the top of the report.



Mascot Search Results: Protein View - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://dell5000/mascot/cgi/protein_view.pl?file=../data/20001016/F289840.dat&hit=Hs%2e782258&px=1&UNIGENE=human Go Links

MATRIX SCIENCE **Mascot Search Results**

UniGene View

```
ID          Hs.78225
TITLE       annexin A1
GENE        ANXA1
CYTOBAND    9q12-q21.2
LOCUSLINK   301
EXPRESS     ;Aorta;Bone;Brain;Breast;CNS;Colon;Ear;Esophagus;Eye;Foreskin;Gall bladder;Heart;Kidney;Lung;Lymph;Nose;Omentum;
CHROMOSOME  9
STS         ACC=G06372 NAME=SHGC-12349 UNISTS=78440
STS         ACC=- NAME=A009X30 UNISTS=5599
STS         ACC=G32952 NAME=A009X30 UNISTS=117530
STS         ACC=- NAME=sts-H67867 UNISTS=29146
STS         ACC=- NAME=H29761 UNISTS=3341
TXMAP       D9S166-D9S1876; MARKER=SHGC-12349; RHPANEL=G3
TXMAP       D9S1876-D9S175; MARKER=WI-7046; RHPANEL=GB4
TXMAP       D9S1876-D9S175; MARKER=A009X30; RHPANEL=GB4
TXMAP       D9S1876-D9S175; MARKER=sts-H67867; RHPANEL=GB4
TXMAP       D9S1876-D9S175; MARKER=H29761; RHPANEL=GB4
PROTSIM     ORG=Caenorhabditis elegans; PROTGI=482227; PROTID=pir:S41022; PCT=42; ALN=311
PROTSIM     ORG=Homo sapiens; PROTGI=6729710; PROTID=pdb:1BO9; PCT=100; ALN=345
PROTSIM     ORG=Mus musculus; PROTGI=71759; PROTID=pir:LUMS1; PCT=87; ALN=345
PROTSIM     ORG=Rattus norvegicus; PROTGI=71758; PROTID=pir:LURT1; PCT=89; ALN=345
SCOUNT      722
SEQUENCE    ACC=BC001275; NID=g12654862; PID=g12654863
SEQUENCE    ACC=XM_005665; NID=g11429703; PID=g11429704
SEQUENCE    ACC=AW950757; NID=g8140416; LID=2983
SEQUENCE    ACC=AX004164; NID=g9927714; PID=g9927715
SEQUENCE    ACC=AV734845; NID=g10852390; CLONE=cdAAHD12; END=5'; LID=4713
SEQUENCE    ACC=BF218154; NID=g11111740; CLONE=IMAGE:4104209; END=5'; LID=3915; MGC=4502100
SEQUENCE    ACC=X05908; NID=g34387; PID=g34388
SEQUENCE    ACC=BF241022; NID=g11154947; CLONE=IMAGE:4109288; END=5'; LID=3917; MGC=4502100
SEQUENCE    ACC=BE171982; NID=g8634708; LID=3484
SEQUENCE    ACC=AV717869; NID=g10815021; CLONE=DCBARH06; END=5'; LID=4704
SEQUENCE    ACC=AW973706; NID=g8164890; LID=3051
```

Done Local intranet

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

Human genome

- ~ 3 x 10^9 bases
- (dbEST was 3.4 x 10^9 bases on 1 May 2001)
- ~ 6 x 10^9 residues in 6 frame translation
- 99.75% of translated sequence is non-coding
- 1.5 x 10^5 tryptic limit peptides of 1500 Da ± 0.5
- 6 x 10^7 no-enzyme peptides of 1500 Da ± 0.5

MATRIX SCIENCE

We can also perform MS/MS searches on continuous raw genomic sequence data. The recent availablilty of a draft assembly of the human genome has made this a focus of great interest. Let's just look at some numbers.

The human genome assembly is approximately 3 billion bases which makes it similar in size to dbEST.

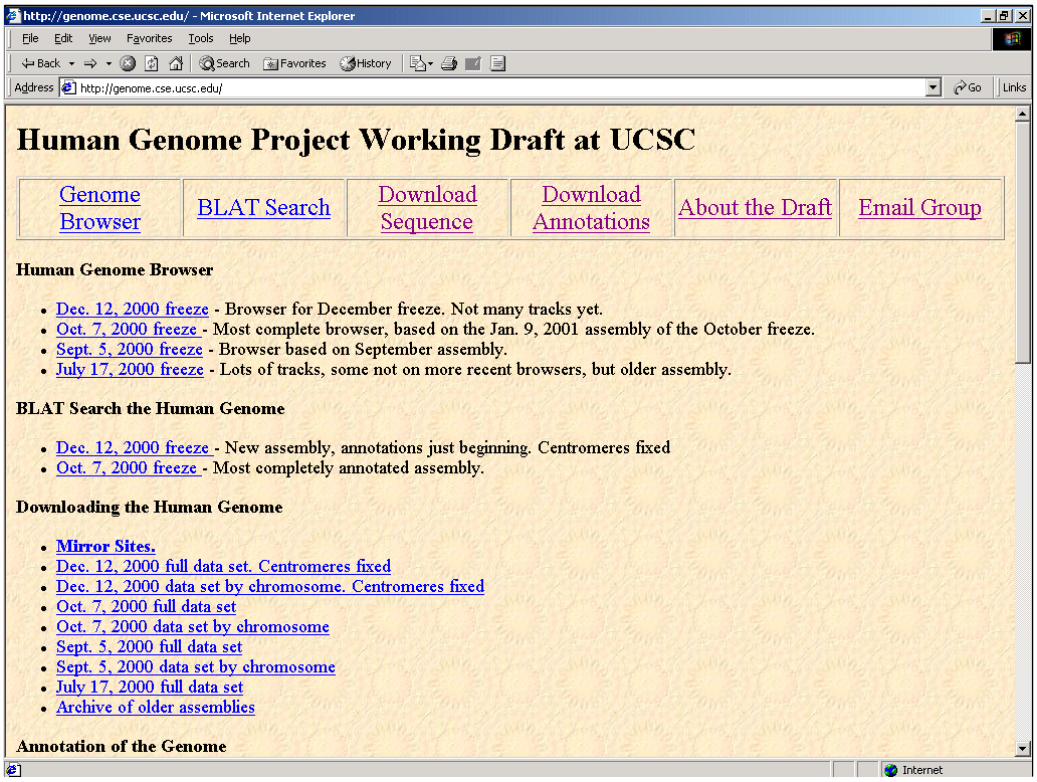
Since we must translate in all 6 reading frames, this corresponds to 6 billion amino acid residues.

In the human genome, only 1.5% of the sequence codes for proteins. Conversely, 99.75% of the translated sequence is non-coding and simply contributes to the background of random matches. This is a severe test of the discrimination of the scoring scheme.

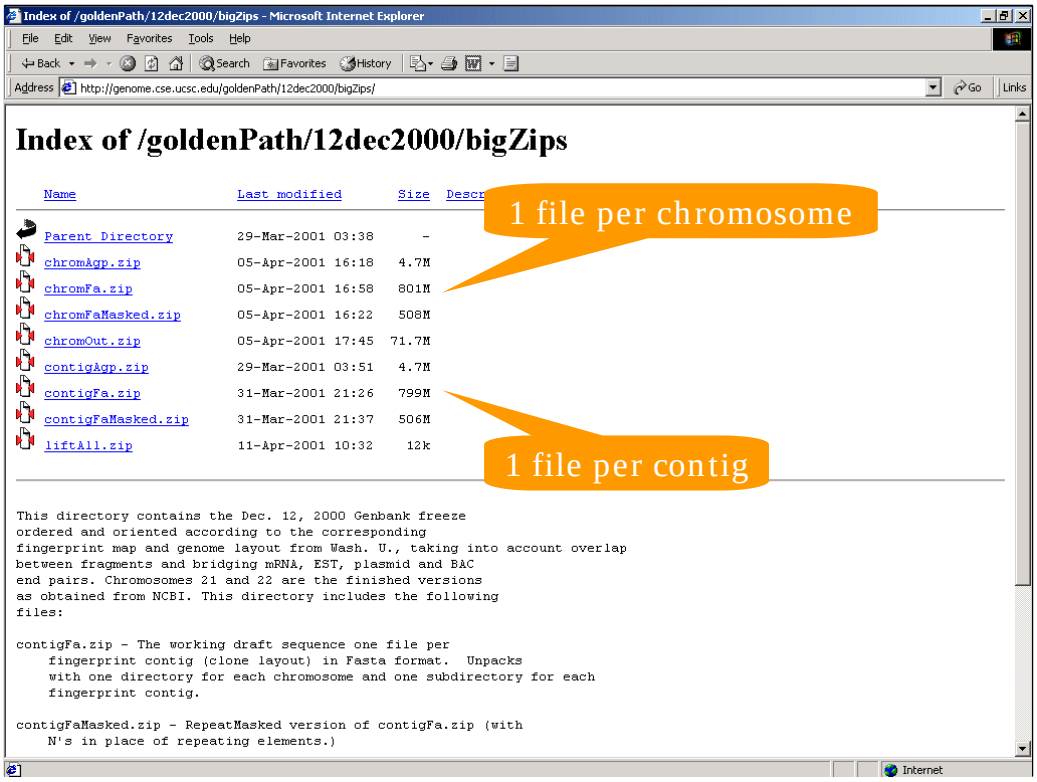
This is only slightly worse than dbEST. Although dbEST is essentially all coding sequences, after translation in 6 frames, we know that 83% (5/6) must be junk.

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.



The draft assembly can be downloaded from UC Santa Cruz GoldenPath web site



You can download the assembly as chromosome length sequences or as collections of contigs. Searching the chromosome length sequences in Mascot is possible, but not advisable.

Chromosome length sequences

- Chromosome 1 is 285 Mbp
- Mascot workspace scales as the length of the longest sequence. Hence need lots of physical RAM to avoid disk thrashing
- More efficient to work with smaller pieces, e.g. 600 kbp with 600 bp overlaps

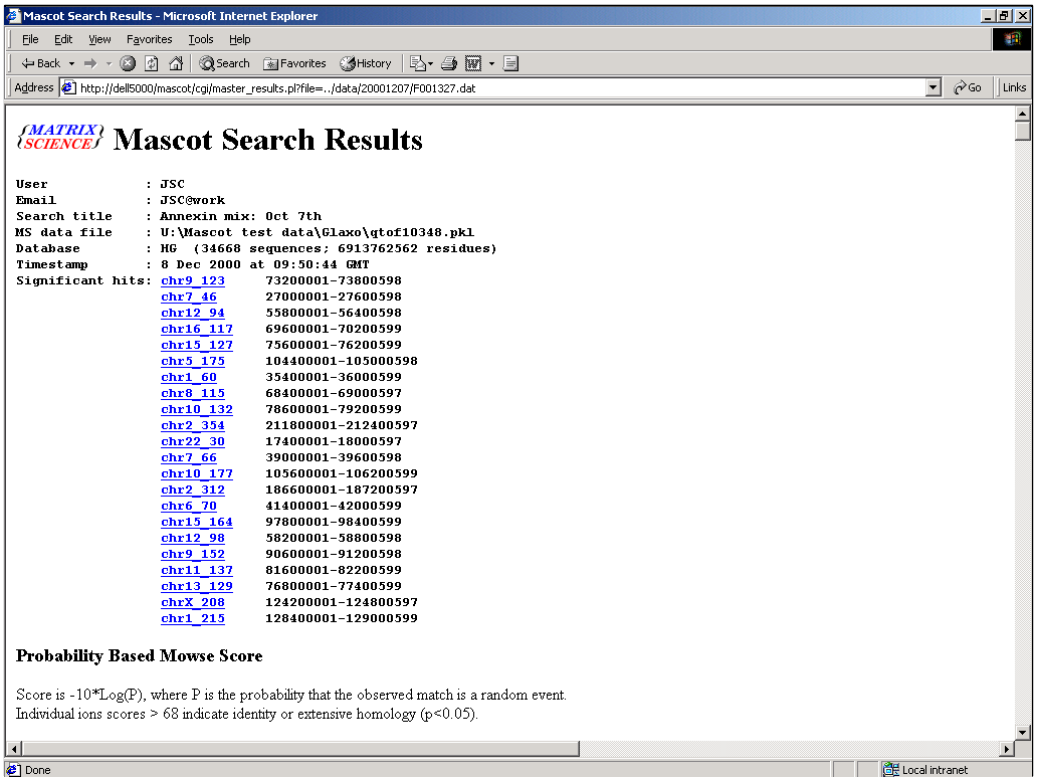
MATRIX SCIENCE

The longest human chromosome is chromosome 1, 285 million bp.

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 with small overlaps.

In any case, we don't know of any tools for reviewing the results which can handle 250 Mbp sequences.



Mascot Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001207/F001327.dat Go Links

Mascot Search Results

```
User             : JSC
Email            : JSC@work
Search title     : Annexin mix: Oct 7th
MS data file     : U:\Mascot test data\Glaxo\qtof10348.pkl
Database         : HG  (34668 sequences; 6913762562 residues)
Timestamp        : 8 Dec 2000 at 09:50:44 GMT
Significant hits: chr9_123      73200001-73800598
                  chr7_46       27000001-27600598
                  chr12_94      55800001-56400598
                  chr16_117     69600001-70200599
                  chr15_127     75600001-76200599
                  chr5_175      104400001-105000598
                  chr1_60       35400001-36000599
                  chr8_115      68400001-69000597
                  chr10_132     78600001-79200599
                  chr2_354      211800001-212400597
                  chr22_30      17400001-18000597
                  chr7_66       39000001-39600598
                  chr10_177     105600001-106200599
                  chr2_312      186600001-187200597
                  chr6_70       41400001-42000599
                  chr15_164     97800001-98400599
                  chr12_98      58200001-58800598
                  chr9_152      90600001-91200598
                  chr11_137     81600001-82200599
                  chr13_129     76800001-77400599
                  chrX_208      124200001-124800597
                  chr1_215      128400001-129000599
```

Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 68 indicate identity or extensive homology (p<0.05).

Done Local intranet

This is the result of searching our standard data set against the unmasked human genome assembly.

Mascot Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001207/F001327.dat Go Links

1. chr9_123 **Total score:** 491 **Peptides matched:** 13

Could not retrieve title string

☐ Check to include this hit in archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
1	430.11	429.10	429.22	-0.12	0	7	4	DAPK
11	415.19	828.36	828.50	-0.13	0	32	6	VLDLELK
12	415.19	828.36	828.51	-0.14	0	33	2	NALLSLAK
☑ 45	607.16	1212.31	1212.53	-0.21	0	70	1	DITSDTSGDFR
☑ 53	631.70	1261.38	1261.59	-0.22	0	69	1	TPAQFDADELR
☑ 65	457.85	1370.54	1370.77	-0.23	1	46	1	VLDLELKGDIEK
☑ 93	775.76	1549.50	1549.81	-0.31	0	69	1	GTDVNVFNTILTTR
☑ 98	547.49	1639.45	1639.77	-0.32	1	(41)	1	DLAKDITSDTSGDFR
☑ 99	820.75	1639.48	1639.77	-0.29	1	52	1	DLAKDITSDTSGDFR
101	840.24	1678.47	1677.90	0.57	1	19	7	KGTDVNVFNTILTTR
☑ 103	851.77	1701.52	1701.88	-0.36	0	82	1	GLGTDEDTLIEILASR
☑ 105	870.21	1738.41	1738.73	-0.32	0	99	1	SEDFGVNEDLADSDAR
☑ 149	785.91	2354.72	2355.15	-0.43	0	66	1	GGPGSAVSPYPTFNPSSDVAALHK

2. chr7_46 **Total score:** 204 **Peptides matched:** 7

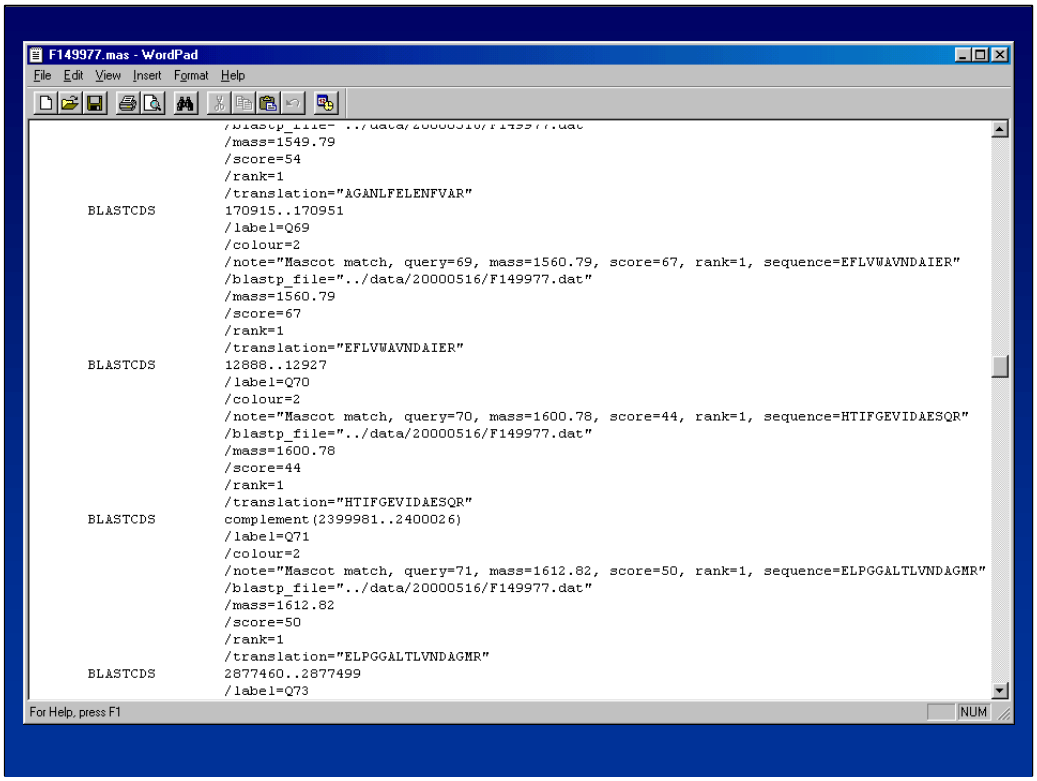
Could not retrieve title string

☐ Check to include this hit in archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
1	430.11	429.10	429.22	-0.12	0	7	4	DAPK
☑ 33	544.15	1086.29	1086.48	-0.19	0	34	1	NYYEQWGK
☑ 43	594.71	1187.41	1187.64	-0.23	0	66	1	IDTIEIITDR
☑ 66	689.19	1376.36	1376.62	-0.26	0	44	1	GGGGNFGPGPGSNFR
☑ 102	565.82	1694.44	1694.76	-0.32	0	42	1	GFGFVTFDDHDPVDK
☑ 117	899.82	1797.63	1797.91	-0.29	0	63	1	LFIGGLSFETTEESLR

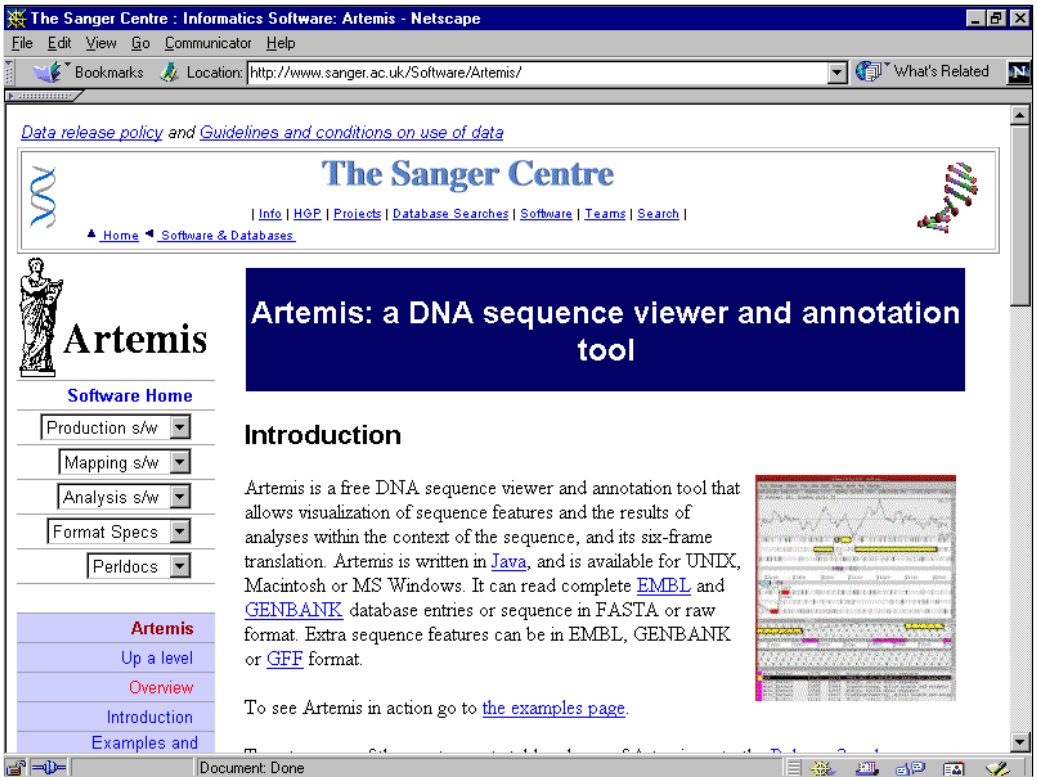
Done Local intranet

The tabular reports which we and others have developed for reporting results from protein and dbEST databases are just not suitable for a sequence the length and complexity of a human chromosome.

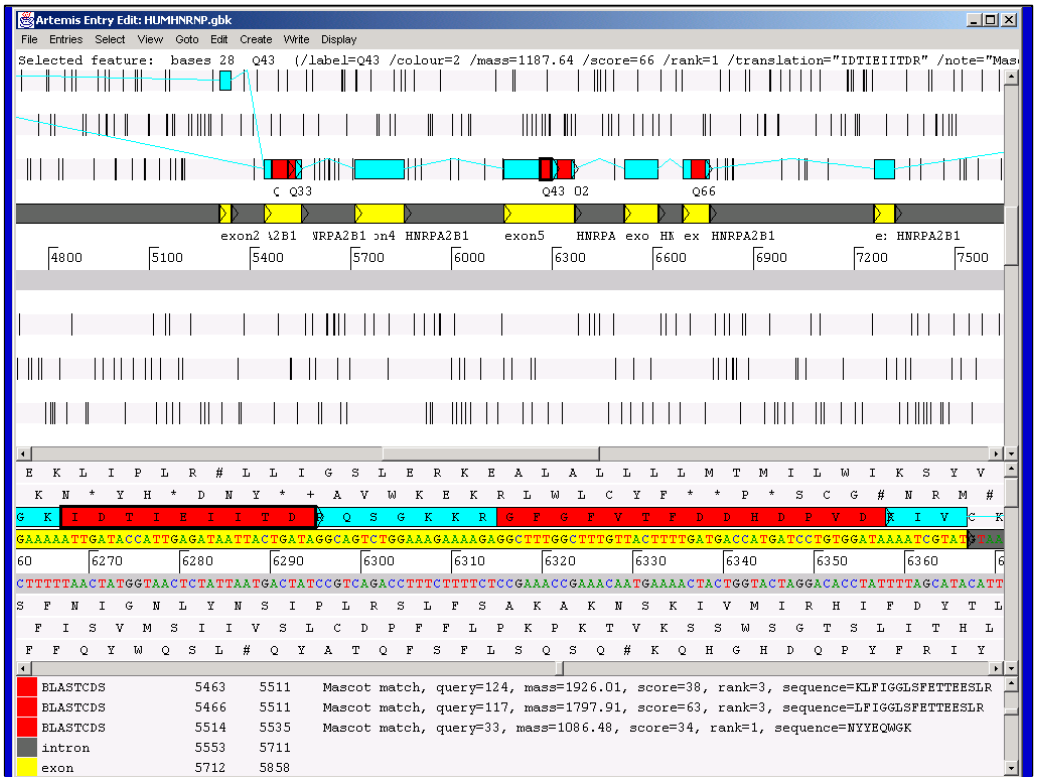


```
                     /mass=1549.79
                     /score=54
                     /rank=1
                     /translation="AGANLFELENFVAR"
     BLASTCDS        170915..170951
                     /label=Q69
                     /colour=2
                     /note="Mascot match, query=69, mass=1560.79, score=67, rank=1, sequence=EFLVWAVNDAIER"
                     /blastp_file="../data/20000516/F149977.dat"
                     /mass=1560.79
                     /score=67
                     /rank=1
                     /translation="EFLVWAVNDAIER"
     BLASTCDS        12888..12927
                     /label=Q70
                     /colour=2
                     /note="Mascot match, query=70, mass=1600.78, score=44, rank=1, sequence=HTIFGEVIDAESQR"
                     /blastp_file="../data/20000516/F149977.dat"
                     /mass=1600.78
                     /score=44
                     /rank=1
                     /translation="HTIFGEVIDAESQR"
     BLASTCDS        complement(2399981..2400026)
                     /label=Q71
                     /colour=2
                     /note="Mascot match, query=71, mass=1612.82, score=50, rank=1, sequence=ELPGGALTLVNDAGMR"
                     /blastp_file="../data/20000516/F149977.dat"
                     /mass=1612.82
                     /score=50
                     /rank=1
                     /translation="ELPGGALTLVNDAGMR"
     BLASTCDS        2877460..2877499
                     /label=Q73
```

For very long DNA sequences, such as this, what we have done is to switch from our standard protein view report to outputting the peptide match results as an EMBL / GenBank format feature table. This may not look very friendly, but the advantage is that this report can now be read into a standard genome browser.



Having generated a feature table, we can now use a genome browser to view it. One which we find works well is Artemis, a Java based genome browser developed and distributed by the Sanger Centre.



Here is an Artemis screenshot showing three views of a portion of the genome. 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.

Key	Category	MSDB	dbEST	HG
a	Top match with significant ions score	74	56	33
b	Top match, but ions score not significant	26	37	13
c	Not top match and ions score not significant	10	11	11
d	No match because of higher scoring non-significant matches	-	6	11
e	No match because peptide sequence not found in MSDB	4	-	-
f	No match because peptide sequence not found in dbEST	-	2	-
g	No match because coding sequence substantially missing from HG	-	-	15
h	No match because coding sequence poorly aligned in HG	-	-	10
i	No match because peptide spans exon / intron boundary in HG	-	-	19
k	No match because peptide results from non-tryptic post-translational processing	-	2	2

MATRIX SCIENCE

When we make a detailed comparison of the results from searching the same data against the three different types of database, the major differences are caused by two factors.

First, the human genome assembly is only a draft assembly and, at the time we did this study, there were complete mRNA's which were either poorly aligned or even missing. This accounted for 25 "missing" peptide matches. Obviously, this situation will change over the coming months as the assembly is refined.

The second factor will not change, if we choose to search the raw genomic sequence. Approximately one quarter of peptide matches are missed because they span exon / intron boundaries. This is not a severe problem if we have multiple peptides from the protein, but is clearly a limitation.

HGP Draft Assembly

- Searching complete chromosome entries is possible, but unwieldy.
- Scoring statistics very similar to dbEST
- Some well characterised proteins are 'missing' (e.g. transaldolase)
- When error rate is high, better off searching redundant EST's than a single consensus sequence
- Still too early to use public draft assembly for routine protein ID?

MATRIX SCIENCE

The main conclusions of our database comparison study are listed.

MATRIX
SCIENCE

http://www.matrixscience.com