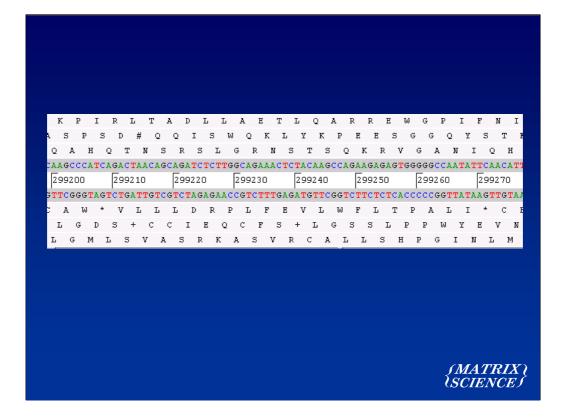
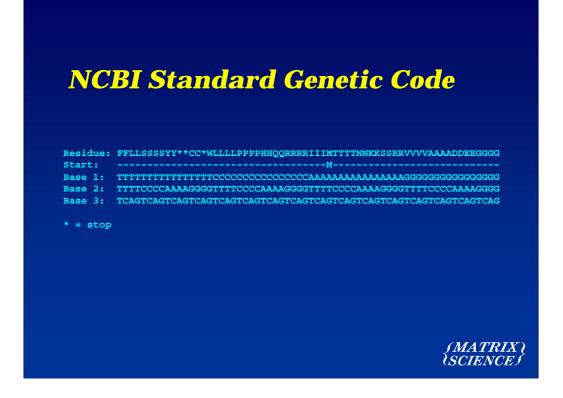
Searching Nucleotide Databases





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.



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)

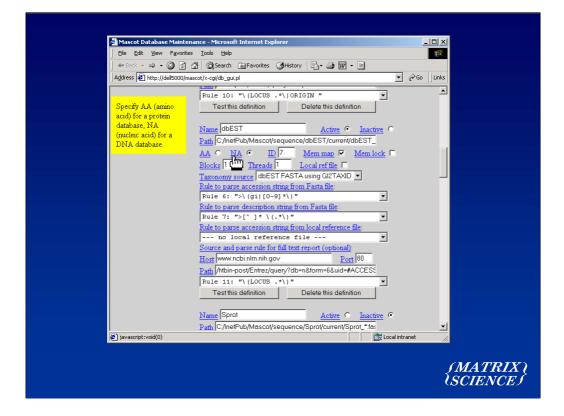
The rules for Nucleic Acid translation in Mascot are:

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.

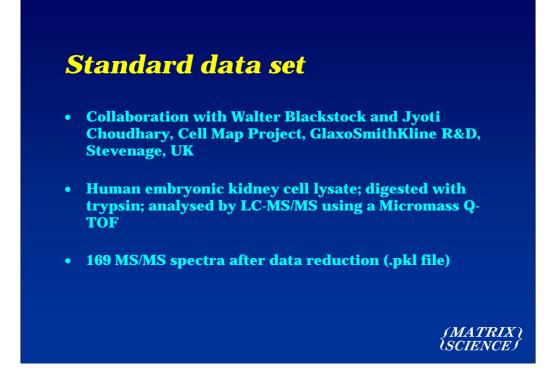
Finally, all translations use the standard genetic code. Ideally, we would use species specific code where a sequence has a known taxonomy. But, again, this is not done at present.



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

Mascot Database Maintenance - Microsoft Internet Explorer	
Eile Edit View Favorites Iools Help	
] 4= Back • → - ③ 🗿 🚮 ③Search 📾 Favorites ⑤ History 🛂 • 🎒 🐨 - 🗐	
Address Addres	▼ 🖗 Go 🛛 Links
MailTransport 0	
MonitorEmailCheckFreq 300	
JobCrashedTimeout 36000	
MonitorTestTimeout 12000	
LogoImageFile/images/88x31_logo_white.gif	
ErrorLogFile/logs/errorlog.txt	
SearchLogFile/logs/searches.log	
MonitorLogFile/logs/monitor.log	_
NTMonitorGroup Administrators	_
NTIUserGroup Guests	
ForkForUnixApache	
IgnoreIonsScoreBelow 0	
MaxSequenceLen 50000000	
MaxConcurrentsearches 10	
CentroidWidth 0.25	
FormVersion 1.01	
ReportNumberChoices 5,10,20,30,40,50,100,200	
ShowSubSets 0	
RequireBoldRed 0	
ReviewColWidths 7.8.8.27.30,100,32,25,4,13,2,4,6,16	
Add New Option Return to database definitions	
	•
	Eccal intranet

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



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/m	ascot/cgi/master_re	ssults.pl?file=/data/20001016/F003980.dat - Microsoft Internet Explorer
<u>Eile E</u> dit <u>V</u> iew	F <u>a</u> vorites <u>T</u> ools <u>H</u> el	p 🔤
🕁 Back 🔹 🔿 👻 🌘	🗿 👔 🚮 🔞 Search	h 🗃 Favorites 🎯 History 🔄 🚽 📰 👻 📄
Address 🙋 http://del	l5000/mascot/cgi/master	_results.pl?file=/data/20001016/F003980.dat#Hit1
,		
(MATRIX)	laraat S.	earch Results
(SCIENCE)	lascot Se	carch Results
User	: JSC	
user Email	: JSC@work	
Search title	: Annexin mix	·
MS data file		cst data\Glaxo\qtof10348.pkl
Database	•	21 (508120 sequences; 156794043 residues)
Timestamp) at 13:38:06 GMT
Significant hi	ts: LUHU	annexin I – human
	AAC52068	HSTALDR3 NID: - Homo sapiens
	AAB19866	S57440S13 NID: - Rattus sp.
	AAC78495	OCU24656 NID: - Oryctolagus cuniculus
	1TGSZ	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
	<u>Q9TTV2</u>	VITAMIN D RESPONSE ELEMENT BINDING PROTEIN Saguinus oedipus (Cotton-top tamarin).
	<u>Q9XSY6</u>	HNRNP A/B RELATED PROTEIN (FRAGMENT) Felis silvestris catus (Cat).
	<u>S60335</u>	TGF-beta receptor interacting protein 1 - human
	Q9QZD9	TGF-BETA RECEPTOR BINDING PROTEIN Mus musculus (Mouse).
	<u>B38611</u>	casein kinase II (EC 2.7.1) alpha' chain - chicken
	B35838	casein kinase II (EC 2.7.1) alpha' chain - human
	<u>S55282</u> AAB47721	isocitrate dehydrogenase (NAD+) (EC 1.1.1.41) alpha chain precursor - human Unggruppi V NTM - Argent cariere
	S41754	HUMKRT1X NID: - Homo sapiens CRKL protein - human
	KRHU2	keratin 1, type II, cytoskeletal - human
	LUGP1	annexi I - quinea piq
	BAA37117	AB001915 NID: - Homo sapiens
	S40776	ribonucleoprotein - African clawed frog
	CAA64477	SSANNEXNI NID: - Sus scrofa
	AAA59468	HUMERTIOA NID: - Homo sapiens
	PC4375	telomeric and tetraplex DNA binding protein gTBP42 V - rat (fragment)
	JC5660	hepatoma-derived growth factor - mouse
	1HA11	hnrnp al hnrnp al (rbdl, rbd2) hnrnp al 1-184, fragment 1 - human
	152962	FBRNP – human
	093446	ANNEXIN MAX3 Oryzias latipes (Medaka fish).
	G3P2_HUMAN	GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (EC 1.2.1.12) Homo sapiens (Human).
1		Cff Local intranet

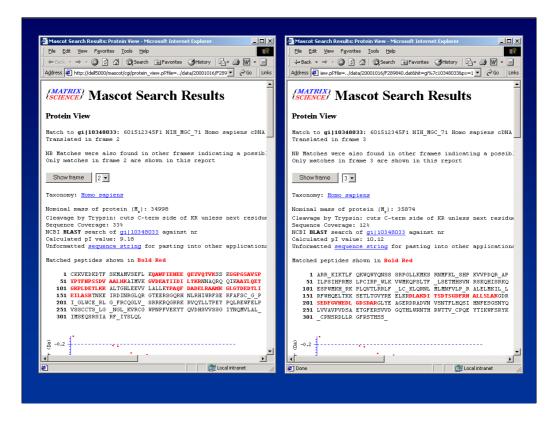
First of all, we searched the data against a comprehensive nonidentical protein database, MSDB. We found significant matches to 22 human proteins ... and one non-human, our frequent flyer, bovine trypsin.

Mascot Search Results - Microsoft Int	ernet Explorer
Eile Edit View Favorites Tools H	
🗘 Back 🔹 🔿 🖌 🙆 👔 🖓 🥘 Sear	rch 📾 Favorites 🏼 🔞 History 🛛 🖳 🖉 🐨 🕶 🚍
Address 🛃 http://dell5000/mascot/cgi/maste	er_results.pl?file=/data/20001016/F289840.dat
(MATRIX) SCIENCE/ Mascot S	earch Results
User : JSC	
Email : jcottrell@	<pre>matrixscience.com</pre>
Search title : Annexin	
MS data file : U:\Mascot	test data\Glaxo\qtof10348.pkl
Database : dbEST 2000)1001 (35277150 seguences; 4704334140 residues)
	ens (human) (14927850 seguences)
-	00 at 21:15:15 GMT
	33 601512345F1 NIH MGC 71 Homo sapiens cDNA clone IMAGE:3913811 5'
	40 601512293F1 NIH_MGC_71 Homo sapiens cDNA clone DMAGE:3913822 5'
<u>gi 1034061</u>	
gi 6571609	
gi 1033082	
gi 9145520	
<u>gi 9891351</u> qi 1034530	
qi 1019866	
gi 1034720	
gi 1031978	
gi 1033234	
gi 1009933	
gi 1014535	
gi 9155988	
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	5 RC1-CT0249-030300-026-b03 CT0249 Homo sapiens cDNA
gi 1015364	7 601556364T1 NIH_MGC_58 Homo sapiens cDNA clone IMAGE:3826407 3'
gi 1031914	
gi 6798648	
gi 1032794	
gi 9898273	
<u>gi 1040110</u>	
<u>gi 8906307</u>	
gi 8627305	
gi 1014628	
gi 8114235	5 PM3-DT0064-260300-002-g12 DT0064 Homo sapiens cDNA

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.

PBak → → · · · · · · · · · · · · · · · · ·		View Favorites	<u>T</u> ools <u>H</u> elp						
witch to Protein Summary Report o create a bookmark for this report, right click this link: <u>Peptide Summary Report (Annexin)</u> huster matches using UniGene index for human bovine SelectAll SelectNone Search Selected Archive Report . <u>gri 10348033</u> Mass: 35874 Total score: 700 Peptides matched: 14 601512345F1 NIH_MGC_71 Homo sapiens cDNA clone INAGE:3913811 5' Check to include this hit in archive report Query Observed Mr(expt) Mr(calc) Delta Miss Score Rank Peptide 7 12 415.19 828.36 828.51 -0.14 0 33 1 NALLSLAK 7 45 607.16 1212.31 1212.53 -0.22 0 69 1 TRAGEDADELR 7 52 631.70 1261.38 1261.59 -0.22 0 69 1 TRAGEDADELR 7 52 631.70 1261.38 1362.66 -0.27 0 73 1 GVDEATIIDILTK 7 91 515.20 1542.58 1542.66 -0.28 1 46 1 GVDEATIIDILTK 7 92 851.77 1701.52 1701.88 -0.36 0 82 1 GLATTIDILTK 7 92 877.49 1639.45 1639.77 -0.32 1 (41) 1 DLAKDITSDTSGDFR 7 10 881.77 1701.52 1701.88 -0.36 0 82 1 GLATEDTLEILASR 105 870.21 1738.41 1738.73 -0.32 0 82 2 SEDEGVNEDLEDSGAR + 1 Methyl ester (DE) 7 13 707.22 2118.63 2119.08 -0.45 1 35 2 ANIKGLEGTDEDTLEILASR 131 707.22 2118.63 2119.08 -0.45 1 35 2 ANIKGLEGTDEDTLEILASR + 1 Oxidation (M) 7 132 1062.33 2122.64 2122.95 -0.35 0 (72) 1 QANFIENEEQEVYQTVK + 1 Pyro-glu (N-term Q) 7 13 1 QANFIENEEQEVYQTVK	⇒Back 💌	ə - 🙆 🙆 🖞	🖞 🔕 Search	😹 Favorites 🛛 🕄	History	à- 🎒	w - 1		
o create a bookmark for this report, right click this link: <u>Peptide Summary Report (Annexin</u>) huter matches using UniGene index for human <u>bovine</u> Select All Select None Search Selected Archive Report . <u>gri 10348033</u> Mass: 35874 Total score: 700 Peptides matched: 14 60151234571 NII_MCC_71 Homo septens cDNA clone IMAGE:3913811 5' Check to include this hit in archive report Query Observed Mr(expt) Mr(calc) Delta Miss Score Rank Peptide 2 12 415.19 828.36 828.51 -0.14 0 33 1 NALLSLAK 2 45 607.16 1212.31 1212.53 -0.21 0 70 1 DITSDTSOPER 2 53 631.70 1261.38 1261.59 -0.22 0 69 1 TFAQFDADELR 2 69 694.25 1386.49 1386.76 -0.27 0 73 1 GVDEATIIDILITK 2 98 547.49 1639.43 1639.77 -0.32 1 (41) 1 DLAKDITSDTSGPER 2 98 547.49 1639.43 1639.77 -0.32 1 (41) 1 DLAKDITSDTSGPER 2 99 547.29 1532.48 1542.66 -0.20 1 46 1 GUTPEATIDILITKR 2 98 547.49 1639.43 1639.77 -0.32 1 (41) 1 DLAKDITSDTSGPER 2 103 651.77 1701.52 1701.88 -0.36 0 82 1 GLGTDEDTLIEILASR + 1 Methyl ester (DE) 3 13 707.22 2118.63 2119.08 -0.45 1 35 2 AMMKGLGTDEDTLIELASR + 1 Oxidation (M) 3 13 707.22 2118.63 2119.08 -0.35 0 (72) 1 QMFTENEEQEVYQTVK + 1 Pyro-glu (N-term Q) 3 13 1070.83 2139.64 2140.01 -0.37 0 84 1 QMFTENEEQEVYQTVK	dress 🧧	http://dell5000/mas	cot/cgi/master_re	sults.pl?file=/dat	a/20001016/	289840.	dat		🗾 🔗 GO 🗍 L
b create a bookmark for this report, right click this link: <u>Peptide Summary Report (Annexin</u>) tuster matches using UniGene index for human bovine Select All Select None Search Selected Archive Report . <u>gri110348033</u> Mass: 35874 Total score: 700 Peptides matched: 14 60151234571 NIH MCC_71 Homo septens cDNA clone IMAGE:3913811 5' Check to include this hit in archive report Query Observed Mr(expt) Mr(calc) Delta Miss Score Rank Peptide 2 12 415.19 828.36 828.51 -0.14 0 33 1 NALLSLAK 2 15 607.16 1212.31 1212.53 -0.21 0 70 1 DITSDTSOPER 2 53 631.70 1261.38 1261.59 -0.22 0 69 1 TFAQFDADELR 2 59 694.25 1386.49 1386.76 -0.27 0 73 1 GVDEATIIDILITK 2 59 694.25 1386.49 1386.76 -0.27 0 73 1 GVDEATIIDILITK 2 59 59.75 1639.48 1639.77 -0.32 1 (41) 1 DLAKDITSDTSOPER 2 59 694.71 1738.41 1738.73 -0.32 0 82 1 GLGTPEDTILEILASR + 1 Methyl ester (DE) 3 13 707.22 2118.63 2119.08 -0.45 1 35 2 AMKGLGTDEDTLIELLASR + 1 Oxidation (M) 3 13 1070.83 2129.64 2140.01 -0.37 0 84 1 QMFIENEEQEVYQTVK									
uster matches using UniGene index for human bovine Select All Select None Search Selected Archive Report gri 10346033 Mass: 35874 Total score: 700 Peptides matched: 14 60151234571 NHR MCC_71 Homo sepiens cDNA clone INAGE:3913811 5' Check to include this hit in archive report Query Observed Mr(expt) Mr(calc) Delta Miss Score Rank Peptide 7 12 415.19 828.36 828.51 -0.14 0 33 1 NALLSLAK 7 45 607.16 1212.31 1212.53 -0.21 0 70 1 DITSDTSGDFR 7 52 631.70 1261.38 1261.59 -0.22 0 69 1 TPAQFDADELR 7 52 631.70 1261.38 1261.59 -0.22 0 69 1 TPAQFDADELR 7 59 694.25 1366.49 1386.76 -0.27 0 73 1 GVDEATIDILTK 7 59 694.25 1366.49 1386.76 -0.28 1 46 1 GVDEATIDILTK 7 59 694.25 1369.77 -0.32 1 (41) 1 DLAKDITSDTSGDFR 7 99 620.75 1639.48 1639.77 -0.29 1 52 1 DLAKDITSDTSGDFR 7 99 620.75 1639.48 1639.77 -0.29 1 52 1 DLAKDITSDTSGDFR 7 99 620.75 1639.48 1639.77 -0.29 1 52 1 DLAKDITSDTSGDFR 7 103 851.77 1701.52 1701.88 -0.36 0 82 1 GLOTDEDTLIEILASR 7 105 870.21 1738.41 1738.73 -0.32 0 82 2 SEDFGVMEDLGDSDAR + 1 Methyl ester (DE) 7 123 476.92 1903.67 1904.03 -0.36 1 22 1 AAYLOETGKPLDETLKK 131 707.22 2118.63 2119.08 -0.45 1 35 2 AAMKGLGTDEDTLIEILASR + 1 0xidation (M) 7 132 1062.33 2122.64 2122.84 -0.35 0 (72) 1 QAMFIENEEQEVYQTVK + 1 Pyro-glu (N-term Q) 7 133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAMFIENEEQEVYQTVK	vitch to l	Protein Summar	<u>y Report</u>						
uster matches using UniGene index for human bovine Select All Select None Search Selected Archive Report gri 10346033 Mass: 35874 Total score: 700 Peptides matched: 14 60151234571 NIH ROC_71 Homo sepiens cDNA clone INAGE: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 59 694.25 1366.49 1386.76 -0.27 0 73 1 GVDEATIIDILTK 591 515.20 1542.58 1542.66 -0.28 1 46 1 GVDEATIIDILTK 592 654.749 1639.45 1639.77 -0.32 1 (41) 1 DLAKDITSDTSGDFR 593 631.77 1701.52 1701.88 -0.36 0 82 1 GLOTEDTILIELLASR 503 631.77 1701.52 1701.88 -0.36 0 82 1 GLOTEDTILIELLASR 504 670.21 1738.41 1738.73 -0.32 0 82 2 SEDFGVMEDGEDSAR + 1 Methyl ester (DE) 5123 476.92 1903.67 1904.03 -0.36 1 22 1 AAN(EGRPLDETLIELLASR 513 707.22 2118.63 2119.08 -0.45 1 35 2 AAN(EGRPDETLIELLASR + 1 0xidation (M) 5133 1070.83 2129.64 2140.01 -0.37 0 84 1 QAMFIENEEQEVQTVK	create :	a hookmark for	this report rig	ht click this link	·· Pentide :	Summar	w Repor	t (Annex	in)
Select All Select None Search Selected Archive Report gril 10346033 Mass: 35874 Total score: 700 Peytides matched: 14 601512345F1 NIR MCC_71 Homo septens cDNA clone IMAGE:3913811 5' Check to include this hit in archive report Total score: 700 Peytides matched: 14 601512345F1 NIR MCC_71 Homo septens cDNA clone IMAGE:3913811 5' Query Observed Mr(expt) Mr(calc) Pelta Miss Score Rank Peytide 12 415.19 828.36 828.51 -0.14 0 33 1 MALLSLAK 45 607.16 1212.31 1212.53 -0.22 0 69 1 TRAGEDADELR 52 631.70 1261.38 1261.59 -0.22 0 69 1 TRAGEDADELR 59 694.25 1366.49 1366.76 -0.27 0 73 1 GVDEATIIDILTK 91 515.20 1542.66 -0.28 1 46 1 GVDEATIIDILTKR 92 820.75 1639.48 1639.77 -0.32 1 411 <td>, or other</td> <td></td> <td></td> <td></td> <td><u> </u></td> <td></td> <td>7 2000 01</td> <td></td> <td></td>	, or other				<u> </u>		7 2000 01		
g1 10348033 Mass: 35874 Total score: 700 Peptides matched: 14 G01512345F1 NIH_MCC_71 Homo septens cDNA clone IMAGE:3913811 5' Check to include this hit in archive report Query Observed Mr(capt) 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 1386.76 -0.27 0 73 1 GUDEATIIDILTK 91 515.20 1542.58 1542.66 -0.22 1 46 1 GUDEATIIDILTK 92 517.49 1639.45 1639.77 -0.32 1 411 1 DLAKDITSDTSGDFR 103 851.77 1701.52 1701.88 -0.36 0 82 1 GLOTDEDTILEILASR 105 870.21 1738.41 1738.73 <td>uster ma</td> <td>tches using Uni</td> <td>Gene index for</td> <td><u>human</u> <u>bovine</u></td> <td></td> <td></td> <td></td> <td></td> <td></td>	uster ma	tches using Uni	Gene index for	<u>human</u> <u>bovine</u>					
gil10348033 Mass: 35874 Total score: 700 Peptides matched: 14 601512345F1 NIH_NGC_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 59 694.25 1386.76 -0.27 0 73 1 GUDEATIIDILTK 91 515.20 1542.58 1542.86 -0.22 1 46 1 GUDEATIIDILTK 92 620.75 1639.45 1639.77 -0.32 1 41 1 DLAKDITSDTSGDFR 193 851.77 1701.88 -0.36 0 82 1 GLOTDEDTILIELIASR 195 870.21 1738.41 1738.73	o								
G01512345F1 NIH_MGC_71 Homo sapiens cDNA clone THAGE: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 DITSDTSOFR 53 631.70 1261.38 1261.59 -0.22 0 69 1 TPAGFDADELR 91 515.20 1386.76 -0.27 0 73 1 GVDEATIIDILTK 92 620.75 1639.45 1639.77 -0.32 1 41 1 DLAKDITSDTSGDFR 103 651.77 1701.52 1701.88 -0.36 0 82 1 GLOTEDTLIEILASR 105 670.21 1738.41 1738.73 -0.32 0 82 SEPEGVMEDLGESDAR + 1 Methyl ester (DE) 113 707.22 186.37 -0.36 2 SEPEGVMEDLGEDSDAR + 1 Methyl ester (DE)	Select A	Select I	vone	Search Selecte	a	Arch	nive Rep	ort	
G01512345F1 NIH_MGC_71 Homo sapiens cDNA clone THAGE: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 DITSDTSOFR 53 631.70 1261.38 1261.59 -0.22 0 69 1 TPAGFDADELR 91 515.20 1386.76 -0.27 0 73 1 GVDEATIIDILTK 92 620.75 1639.45 1639.77 -0.32 1 41 1 DLAKDITSDTSGDFR 103 651.77 1701.52 1701.88 -0.36 0 82 1 GLOTEDTLIEILASR 105 670.21 1738.41 1738.73 -0.32 0 82 SEPEGVMEDLGESDAR + 1 Methyl ester (DE) 113 707.22 186.37 -0.36 2 SEPEGVMEDLGEDSDAR + 1 Methyl ester (DE)	gi110	348033 Mas	s: 35874	Total score	: 700 1	entid	es mate	hed: 1	4
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 DITSDTSGPR 53 631.70 1261.38 1261.59 -0.22 0 69 1 TPA0FDADELR 59 694.25 1386.49 1386.76 -0.27 0 73 1 GVDEATIIDILTK 91 515.20 1542.58 1542.66 -0.22 1 46 1 GVDEATIIDILTK 92 820.75 1639.48 1639.77 -0.32 1 (41) 1 DLAKDITSDTSGPR 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 SEDFGVMEDL@SDAR + 1 Met									
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 1386.76 -0.27 0 73 1 GVDEATIIDILTK 91 515.20 1542.58 1542.66 -0.28 1 46 1 GVDEATIIDILTK 92 520.75 1639.77 -0.32 1 (41) DLAKDITSDTSGDFR 92 820.75 1639.48 1639.77 -0.29 1 52 1 DLAKDITSDTSGDFR 103 851.77 1701.82 1701.88 -0.36 0 82 1 GLGTDEDTLIEILASR 105 870.21 1738.41 1738.73 -0.32 0 82 2 SEPEGVMEDLGSDSAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.35 0 72 1 AAKELGTDEDTLIEI	Check	to include	this hit in	archive rep	port				
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 DITSDTSGDPR 53 631.70 1261.38 1261.59 -0.22 0 69 1 TPAQPDADELR 69 694.25 1386.76 -0.27 0 73 1 GVDEATIIDILTK 91 515.20 1542.58 1542.86 -0.28 1 46 1 GVDEATIIDILTK 92 820.75 1639.47 -0.32 1 (41) DLAKDITSDTSGDFR 92 820.75 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 SEPEVWELGBSDAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.36 22 1 AAKELGTDEDTLIEILASR + 1	_						_		
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 1366.49 1366.76 -0.27 0 73 1 GVDEATIIDILTK 91 515.20 1542.58 1542.66 -0.28 1 46 1 GVDEATIIDILTK 92 547.49 1639.45 1639.77 -0.22 1 41 1 DLAKDITSDTSGDFR 93 820.75 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 SEDFGVMEDLGSDAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.36 22 1 AAMKGLGTDEDTLIEILASR 1 131 707.22 2118.63 2119.08 -0.45									-
53 631.70 1261.38 1261.59 -0.22 0 69 1 TPAQFDADELR 59 694.25 1386.49 1386.76 -0.27 0 73 1 GVDEATIIDILTK 91 515.20 1542.58 1542.66 -0.28 1 46 1 GVDEATIIDILTK 92 547.49 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 SEDFGVMEDLGDSDAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.36 22 1 AMKGLGTDEDTLIEILASR 131 707.22 2118.63 2119.08 -0.45 1 35 2 AMKGLGTDEDTLIEILASR + 1 0xidation (M) 132 1062.33 2122.64 2122.99 -0.35									
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 92 515.20 1542.58 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 1701.86 -0.36 0 82 1 GUDTEDTLIEILASR 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 AAXLGETGEDTLIETLASR 131 707.22 2118.63 2119.08 -0.45 1 35 2 AAMKGETGEDTLIETLASR + 1 0xidation (M) 132 1062.33 2122.64 2122.98 -0.35 0 (72) 1 QAHFIENEEQEYQTVK + 1 Pyro-glu (N-term Q) 133 1070.83 2139.64 <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>									
91 515.20 1542.58 1542.86 -0.28 1 46 1 GVDEATIIDILTKR 96 547.49 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 SEGFWAEDLGDSDAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.36 1 22 1 AAKUGETGENDELEKK 131 707.22 118.63 2119.08 -0.45 1 35 2 AAKKGETOEDTLIEILASR + 1 0xidation (M) 132 1062.33 2122.64 2122.98 -0.35 0 (72) 1 QAMFIENEEQEYVQTVK + 1 Pyro-glu (N-term Q) 133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAMFIENEEQEYVQTVK									-
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 GLGTBEDTLIEILASR 105 870.21 1738.41 1738.73 -0.32 0 82 2 SEFEWWEDLGSDSDAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.36 1 22 1 AAYLQETGKPLDETLKK 131 707.22 118.63 2119.08 -0.45 1 35 2 AAMKGLGTDEDTLIEILASR + 1 0xidation (M) 132 1062.33 2122.64 2122.98 -0.35 0 (72) 1 QAMFIENEEQEXVQTVK + 1 Pyro-glu (N-term Q) 133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAMFIENEEQEXVQTVK									
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 SEDFGVMEDLGSDSAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.36 1 22 1 AAMK0LGTGKPLDETLKK 131 707.22 2118.63 2119.08 -0.45 1 35 2 AAMK0LGTGKPLDETLKK 132 1062.33 2122.64 2122.98 -0.35 0 (72) 1 QAWFIENEEQEVQTVK + 1 Pyro-glu (N-term Q) 133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAWFIENEEQEVQTVK	_							-	
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 SEDFGVMEDLGDSDAR + 1 Methyl ester (DE) 123 476.92 1903.67 1904.03 -0.36 1 22 1 AAMKGLGTDEDTLIEILASR 131 707.22 2118.63 2119.08 -0.45 1 35 2 AAMKGLGTDEDTLIEILASR + 1 0xidation (M) 132 1062.33 2122.64 2122.99 -0.35 0 (72) 1 QAMFIENEEQEVQTVK + 1 Pyro-glu (N-term Q) 133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAMFIENEEQEVQTVK	_								
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 0xidation (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	_							-	
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 0xidation (M) 132 1062.33 2122.64 2122.98 -0.35 0 (72) 1 QAMFIENEEQEYQQTVK + 1 Pyro-glu (N-term Q) 133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAMFIENEEQEYQQTVK									
131 707.22 2118.63 2119.08 -0.45 1 35 2 AAMKGLGTDEDTLIEILASR + 1 0xidation (M) 132 1062.33 2122.64 2122.98 -0.35 0 (72) 1 QAWFIENEEQEXVQTVK + 1 Pyro-glu (N-term Q) 133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAWFIENEEQEYVQTVK									
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									-
133 1070.83 2139.64 2140.01 -0.37 0 84 1 QAWFIENEEQEYVQTVK	131		2122.64			0	(72)	1	
		1070.83	2139.64	2140.01	-0.37	0	84	1	
	132		2354.72	2355.15	-0.43	0	66	1	GGP GSAV SPYP TFNP SSDVAAL HK
	<u>132</u> <u>133</u>	785.91							
gi 10347940 Mass: 29372 Total score: 694 Peptides matched: 13	<u>132</u> <u>133</u> <u>149</u>		s: 29372	m-+-7					

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...

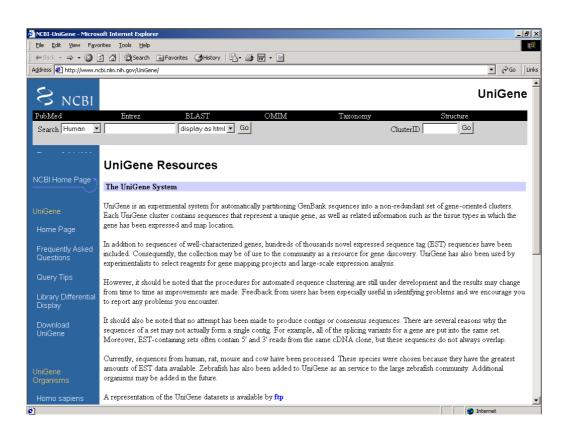


we get a protein view. This is similar to the protein view for a protein database entry, except we have drop down list for the different translation frames. For this particular entry, most of the matches have been found in reading frame 2.

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

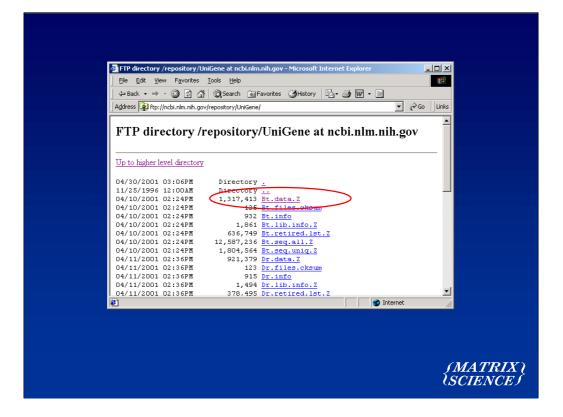
Mascot Search Res	ts - Microsoft Internet Explorer	_ 8
File Edit View F	vorites Tools Help	1
standar → → → 🐼	🕅 🚮 🔞 Search 🗟 Favorites 🎲 History 🖏 - 🚔 👿 - 🚍	
		🔗 Go 🛛 Lin
Agaress 😰 http://deils	00/mascot/cgi/master_results.pl?file=/data/20001016/F289840.dat	(r co j un
(<u>matrix</u>) (science) N	ascot Search Results	
User	: JSC	
Email	: jcottrell@matrixscience.com	
Search title	: Annexin	
MS data file	: U:\Mascot test data\Glaxo\gtof10348.pkl	
Database	: dbEST 20001001 (35277150 sequences; 4704334140 residues)	
Taxonomy	: Homo sapiens (human) (14927850 sequences)	
Timestamp Significant bit	: 16 Oct 2000 at 21:15:15 GMT : gi 10348033 601512345F1 NIH MGC 71 Homo sapiens cDNA clone IMAGE:3913811 5'	
significant hit	(1)10347940 60151229311 NTL MCC 71 Homo sapiens CDNA CLONE IMAGE: 3913822 5'	
	ri 10340616 601509784F1 NIH MGC 71 Homo sapiens cDNA clone IMAGE:3911108 5'	
	ri 6571609 xb37e04.y1 NCI CGAP Lu31 Homo sapiens cDNA clone IMAGE:2578494 5' similar to SW:ANX1	HUMAN PO
	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 NTH MGC 71 Homo sapiens cDNA clone DMAGE: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 10332342 601508038F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3909657 5' gi 10099330 RC3-HT0649-150600-023-e08 HT0649 Homo sapiens cDNA	
	gi 101325359 KC3-INIO43-150000-0225-000 HI0039 Homo sapiens cDNa gi 10145359 601565022F1 NTH MGC 21 Homo sapiens cDNa Clone IMAGE:3840124 5'	
	ri 9155988 601156847F1 NIH MGC 21 Homo sapiens cDNA clone IMAGE:3140530 5'	
	ri 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 NTH 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 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:M552	68 CASETN
	gi 6627305 RC3-HT047/D-120200-013-q02 HT0470 Homo sapiens cDNA	oo onsisiir
	gi 101/46282 6015702851 NTH MGC 21 Homo sapiens cDNA Clone IMAGE:3844988 5'	
	gi 8114235 PM3-DT0064-260300-002-g12 DT0064 Homo sapiens cDNA	
1		Þ
	·····································	

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.

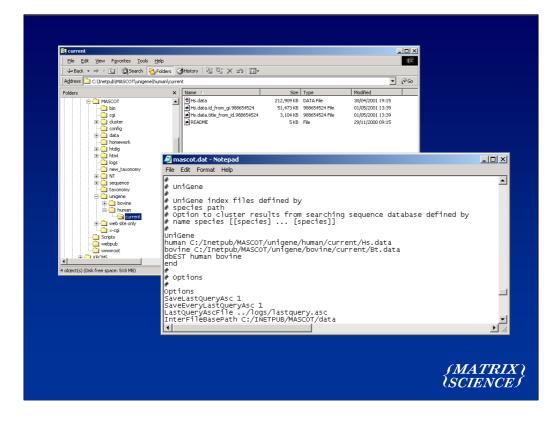


To address this problem, we have recently started using the UniGene index from the National Center for Biotechnology Information to simplify the search results.

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.

_	ults - Microsoft Intern	et explorer	_ 8
<u>File E</u> dit <u>V</u> iew F	<u>avorites T</u> ools <u>H</u> elp		
← Back - → - 🧭) 🗿 🚮 🥘 Search	🕋 Favorites 🛛 🔞 History 🛛 🛃 📰 📃	
Address 😿 C:\Docume	ents and Settings\JohnC\E	esktop\UniGene.mht	▼ 🖉 Go Lir
(<u>matrix</u>) (science)	lascot Se	arch Results	
User	: JSC		
Email		trixscience.com	
Search title	: Annexin		
IS data file		st data\Glaxo\qtof10348.pkl	
)atabase Taxonomy		01 (35277150 sequences; 4704334140 residues) (human) (14927850 sequences)	
limestamp		at 21:15:15 GMT	
Significant hi		ANXA1 annexin A1	
agnizzation in	Hs.77290	TALDO1 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	
	<u>Hs.82201</u>	CSNK2A2 casein kinase 2, alpha prime polypeptide	
	Hs.250616	IDH3A isocitrate dehydrogenase 3 (NAD+) alpha	
	<u>Hs.80828</u> Hs.278572	KRT1 keratin 1 (epidermolytic hyperkeratosis) 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)	
	<u>Hs.181165</u>	EEF1A1 eukaryotic translation elongation factor 1 alpha 1	
	gi 7306319	RCO-BT0387-170100-011-a08 BT0387 Homo sapiens cDNA	
	<u>Hs.89525</u> Hs.289109	HDGF hepatoma-derived growth factor (high-mobility group protein 1-like) DDAH1 dimethylarginine dimethylaminohydrolase 1	
	Hs. 217493	ANXA2 annexin A2	
	Hs.249495	HNRPA1 heterogeneous nuclear ribonucleoprotein A1	
	Hs.249247	FBRMP 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	
	<u>gi 10244362</u>	RCO-ANOO40-200700-022-f03 ANOO40 Homo sapiens cDNA	
1) b

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.

Mascot Search Results - Microsof	Internet Explorer
Ele Edit View Favorites Tools	Heb BB
] ↓= Back • → • ② ③ △ △ ◎	Search 👜 Favorites 🎯 History 🔄 - 🎒 🕅 - 📄 🔰
RCO-PT0006-271199-01: Check to include this Query Observed Mr(1281 Total score: 169 Peptides matched: 4 -006 PT0006 Homo sapiens cDNA nit in archive report expt) Mr(calc) pettode Rank Peptide 0.01 93.04 -06 0.01 90.04 -06 pettode - -
50 414.17 123	🗿 Mascot Search Results - Microsoft Internet Explorer
	Ple Edit View Favorites Tools Help
	↓ Back + → - ② ② ③ ④ ③ ② Search ⓐ Favorites ③ History ⑤ - ④ 酚 - ● □ Linis
11692082 Hase: 13131 RC0-F0006-27119-01 RC0-F006-27119-01 ICO-F006-27119-01 ICO-F006-27119-01 Done ICO-F006-27119-01 Done ICO-F006-27119-01 ICO-F006-27119-01 ICO-F006-27119-01	Query Observed Hr(calc) Delta Hiss Score Rank Peptide 1 430.11 428.10 428.19 0.91 0 6 6 DARP + 1 Acetyl (N-term); 1 Methyl ester (DE) V 5 402.17 602.33 802.43 -0.12 0 33 1 TVSLGAGAK V 11 426.14 303.11 930.47 -0.16 0 6 10 OPSSVD1K 50 414.17 1239.40 1239.69 -0.21 2 18 4 SAPG66SKVPQKK V 115 910.23 108.44 188.43 -0.36 0 63 1 MTDQEAIQUEWQRK 1 1 V 112 108.44 1084.47 1834.37 1024.20 -0.46 0 70 1 MSVUPTVSLOEFITEPVVLR + 1 0xidation (H) V 140 122.08 2241.75 2242.20 -0.46 0 70 1 MSVUPTVSLOEFITEPVVLR + 1 0xidation (H) V 1402 122.28
	Image: Contract of the second seco
	20 Done 2242.20 -0.45 0 (52) 1 MSVQPTVSL66FEITPPVVLR + 1 0xidation (H)
	3.85 2944.45 -0.60 1 (45) 1 TVSLGAGAKORELHUVEAEANOUVEGSPIK + 1 0xidation (H) 3.85 2944.45 -0.60 1 46 2 TVSLGAGAKOELHUVEAEANOUVEGSPIK + 1 0xidation (H)
	Total score: 166 Peptides matched: 6 -c03 DT0071 Homo sapiens cDNA

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.

	Results: Protein View - Microsoft Internet Explorer
<u>File Edit Viev</u>	/ Favorites Iools Help
⇔Back 🔹 ⇒	🖉 🖉 🖄 🕲 Search 🖻 Favorites 🎯 History 🔄 🖓 🚽 🔄
Address 🙋 http:,	/dell5000/mascot/cgi/protein_view.pl?file=/data/20001016/fr289840.dat&hit=Hs%2e78225&px=18UNIGENE=human 🔽 🔗 Go 🗍 Un
{MATRIX}	Mascot Search Results
UniGene Vi	ew
ID	Hs.78225
TITLE	annexin Å1
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
FXMAP	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
SEOHENCE	ACC=AM973706: NTD=08164890: LTD=3051

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 10⁹ bases
- (dbEST was 3.4 x 10⁹ bases on 1 May 2001)
- ~ 6 x 10⁹ residues in 6 frame translation
- 99.75% of translated sequence is non-coding
- 1.5 x 10⁵ tryptic limit peptides of 1500 Da ± 0.5
- 6 x 10⁷ no-enzyme peptides of 1500 Da ± 0.5

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

(MATRIX) SCIENCE (

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...

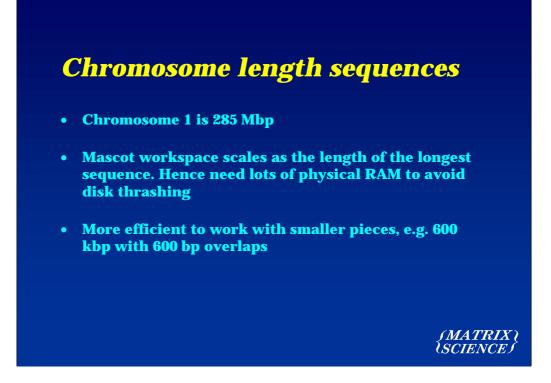
but is not nearly as bad as the no-enzyme case where we have to test 60 million.

http://genome.cse.ucsc.edu/ - Microsoft Internet Explorer		181
File Edit View Favorites Tools Help	_	
ber gent gent gent gent gent gent gent gent		
ddress 🖗 http://genome.cse.ussc.edu/	▼ <i>∂</i> Go	Lin
aness 🔄 inde Walancine research and	<u> </u>]
Human Genome Project Working Draft at UCSC		
Genome BrowserBLAT SearchDownload SequenceDownload AnnotationsAbout the Draft	Email Group	
Iuman Genome Browser		
 Oct. 7, 2000 freeze - Most complete browser, based on the Jan. 9, 2001 assembly of the October freeze. Sept. 5, 2000 freeze - Browser based on September assembly. July 17, 2000 freeze - Lots of tracks, some not on more recent browsers, but older assembly. BLAT Search the Human Genome Dec. 12, 2000 freeze - New assembly, annotations just beginning. Centromeres fixed Oct. 7, 2000 freeze - Most completely annotated assembly. Downloading the Human Genome 		
 Mirror Sites. Dec. 12, 2000 full data set. Centromeres fixed Dec. 12, 2000 data set by chromosome. Centromeres fixed Oct. 7, 2000 full data set Oct. 7, 2000 full data set et y chromosome Sept. 5, 2000 full as set by chromosome July 17, 2000 full data set Archive of older assemblies 		

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

le <u>E</u> dit <u>V</u> iew F <u>a</u> vorites <u>T</u> ool:	s <u>H</u> elp			
Back 🔹 🔿 🚽 🙆 🚮 🖗	🕽 Search 🝙 Favorites 🏼 🕥 Histo	ory 🛛 🗳 🎒 🚺] • 🖃	
ress 🙋 http://genome.cse.ucsc.ed	du/goldenPath/12dec2000/bigZips/			▼ @Go
ndex of /gold	enPath/12de	c2000/l	bigZips	
Name	Last modified	<u>Size</u> <u>Des</u>		
Parent Directory	29-Mar-2001 03:38	-	1 file per chromosome	
chromAgp.zip	05-Apr-2001 16:18	4.7M		
chromFa.zip	05-Apr-2001 16:58	801M		
chromFaMasked.zip	05-Apr-2001 16:22	508M		
chromOut.zip	05-Apr-2001 17:45	71.7M		
contigAgp.zip	29-Mar-2001 03:51	4.7M		
contigFa.zip	31-Mar-2001 21:26	799M		
contigFaMasked.zip	31-Mar-2001 21:37	506M		
<u>liftAll.zip</u>	11-Apr-2001 10:32	12 k	1 file per contig	
is directory contains dered and oriented acc ngerprint map and geno tween fragments and br d pairs. Chromosomes 2 obtained from NCBI. T les:	ording to the corresp me layout from Wash. idging mRNA, EST, pla 1 and 22 are the fini	onding U., taking i smid and BAC shed version	3	
	ng draft sequence one clone layout) in Fast or each chromosome an	a format. U		
ntigFaMasked.zip - Rep	eatMasked version of	contigFa.zip	(with	

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.



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.

Pie Edit Vew Favorites Tools telpt Version Provides Tools telpt Version Proversion Proversitelpt		
Pack + Pack	Mascot Search Results - Microsoft Internet Explorer	X
<pre>wdws i http://dwf5000/macod/cg/mac</pre>	Ejle Edit View Favorites Iools Help	
Market Mascot Search Results User :	♀ Back マ ⇒ → 🙆 🕼 🚮 🛞 Search 📾 Favorites 🎯 History 🔄 र 🚔 🐨 र 🚍	
Market Mascot Search Results User :	Address 🚯 http://dell5000/mascob/col/mascer_results.pl?file=/data/20001207/F001327.dat	▼ 🗟 Go Links
User : JSC Bmail : JSC@vork Search tille : Annexin mix: Oct 7th St data file : U:Wascot test data(Glasco)qtof10348.pkl Database : H6 (34668 sequences; 6913762562 residues) Timestamp : B Dec 2000 at 09:30:44 dff Significant hits chr9_123 73200001-73800598 chr16_112 6560001-70200599 chr16_112 6560001-70200599 chr16_112 6600001-70200599 chr16_112 66400001-65000597 chr16_0 35400001-36000597 chr12_31 11400001-18000597 chr12_31 17400001-3000598 chr12_31 17400001-18000597 chr2_321 16600001-82200599 chr2_321 16600001-82200599 chr2_321 16600001-82200599 chr2_321 16600001-82200599 chr2_321 16600001-82200599 chr12_98 58200001-58800598 chr91_92 56200001-9200598 chr91_93 56200001-9200598 chr91_93 56200001-9200598 chr9_1_93 56200001-9200598 chr9_1_93 56200001-9200		
Binal : : JSC@vork Search title : : WiMascot test data[Glaxo]qtof10348.pkl BX data file : : WiMascot test data[Glaxo]qtof10348.pkl Database : : NC (34668 sequences; 6913762562 residues) Timestamp : 8 Dec 2000 at 09:50:44 dRT Significant hits : ohr9:123 73200001-73800598 ohr12:12 73200001-7600598 ohr12:12 7560001-76200599 ohr13:127 7560001-76200599 ohr15.127 ohr13:127 7560001-75200599 ohr15.127 ohr13:127 7560001-75200599 ohr15.127 ohr2.173 10440001-105000597 ohr13:12 ohr2.16 64400001-2000599 ohr2.18 211800001-3200599 ohr2.312 10660001-197200599 ohr2.312 10660001-187200597 ohr2.312 10660001-187200599 ohr2.312 10660001-187200599 ohr2.192 11400001-8000599 ohr2.192 10600001-12000599 ohr2.193 53200001-38000599 ohr2.194 1500001-9200599 ohr2.195 12400001-124000599 ohr2.195 12400001-12400599 ohr2.194 12800001-2000599	(MATRIX) Mascot Search Results	
Search title : Annexin mix: Oct 7th MS data file : U:\Wascot test data[Gax0\qtof10348.pk] Database : NG (34668 sequences; 6913762562 residues) Timestamp : 8 Dec 2000 at 09:30:44 GMT Significant hits: chr9[12] 7320001-73800598 chr7_46 2700001-27600598 chr12_94 5580001-56000599 chr15_175 10440001-105000599 chr15_127 7560001-76200599 chr12_12 7660001-70200599 chr16 115 6640001-69000597 chr10 132 7660001-72200599 chr2 314 211800001-212400597 chr2 6 3900001-36000597 chr2 6 3900001-36000599 chr2 12 166600001-18200599 chr2 131 1740001-18000597 chr2 6 3900001-38000599 chr2 132 166600001-18200599 chr2 131 132 76600001-18200599 chr2 132 186600001-18200599 chr2 132 186600001-18200599 chr2 132 186600001-18200599 chr2 132 18600001-18200599 chr2 132 1800001-18200599 chr2 132 5000001-18000598 chr2 132 7600001-18200599 chr2 132 7600001-18200599 chr2 200 599 chr2 200 590 chr2 200 590 chr2 200 599 chr2 200 chr2 200 590 chr2 200 599 chr2 200 chr2 200 590 chr2 200 599 chr2 200 chr2 200 ch	User : JSC	
<pre>MS data file : U:\Mascot test data\Glaxo\qtof10348.pkl Database : HG (34660 sequences; 6913762562 residues) Timestamy : B Dec 2000 at 09:50:44 dMT Significant hits : ohr9_123 73200001-73800598 ohr16_117 6960001-76200599 ohr5_175 104400001-76200599 ohr5_175 104400001-76200599 ohr6_115 68400001-76200599 ohr6_115 68400001-212400597 ohr6_115 68400001-212400597 ohr6_230 17400001-10500599 ohr6_70 at1400001-10500599 ohr6_71 05600001-105200599 ohr6_72 116600001-102200599 ohr6_72 116600001-102200599 ohr6_72 118600001-105200599 ohr6_70 at1400001-42000599 ohr15_164 97800001-9200599 ohr15_164 97800001-9200599 ohr13_129 76600001-9200599 ohr13_129 76800001-91200598 ohr13_129 76800001-91200598 ohr13_129 76800001-91200599 ohr13_129 76800001-91200599 ohr13_00001-91200599 ohr13_120 76800001-91200599 ohr13_120 76800001-91200599 ohr13_120 76800001-91200599 ohr13_120 76800001-91200599 ohr13_120 76800001-91200599 ohr13_120 76800001-91200599 ohr13_120 76800001-91200599 ohr13_00001-91200599 ohr13_00001-91200599 ohr13_120 76800001-91200599 ohr13_120 76800001-91200590 o</pre>	Email : JSC@work	
Database : MC (34668 sequence; 6913762562 residues) Timestamp : 8 Dec 2000 at 09:50:44 dMT Significant hits: drv7 46 27000001-27600598 drv16 117 65600001-70200599 drv15 127 7560001-76200599 drv16 117 65600001-70200599 drv1 10 33400001-36000597 drv1 10 33400001-36000597 drv1 11 32 78600001-79200599 drv2 354 211800001-18000597 drv7 66 39000001-305098 drv1 0 177 105600001-105200599 drv2 312 186600001-105200599 drv1 2 314 1400001-18000597 drv7 66 39000001-305098 drv1 0 177 105600001-105200599 drv1 2 314 1400001-18000599 drv1 2 315 211800001-18000599 drv1 2 315 21800001-18000599 drv1 2 315 31600001-9200599 drv1 2 315 31600001-9200599 drv1 2 315 31600001-9200599 drv1 2 315 31600001-9200599 drv1 2 96 5820001-58800598 drv1 1 137 61600001-9200599 drv1 2 96 5800001-91200599 drv1 1 137 61600001-71700599 drv1 2 125 128400001-12800597 drv1 2 15 128400001-12800597 drv1 2 15 128400001-12800599 drv1 2 00 5800000000000000000000000000000000	Search title : Annexin mix: Oct 7th	
Timestamp : 8 Dec 2000 at 09:50:44 GMT Significant hits: chr9 123 7320001-73800598 chr12 94 55800001-56400598 chr12 94 55800001-76200599 chr15 127 75600001-76200599 chr15 127 75600001-76200599 chr2 135 68400001-6900597 chr6 113 68400001-29200599 chr2 354 211800001-29200599 chr2 354 211800001-29200599 chr2 354 211800001-3900597 chr2 354 211800001-3900597 chr2 10 5600001-39600597 chr2 312 116600001-180200599 chr2 312 105600001-39600599 chr2 113 05600001-39600599 chr2 12 10 5600001-39600599 chr2 12 10 5600001-39600599 chr2 12 11 05600001-39600599 chr2 12 11 05600001-39600599 chr2 12 11 05600001-39600599 chr2 12 11 05600001-39600599 chr2 12 11 0560001-39600599 chr2 12 11 0560001-39600599 chr2 12 11 0560001-39600599 chr2 12 11 0560001-39600599 chr2 13 12 9 7680001-3800599 chr13 129 7680001-19200599 chr13 129 7680001-19200599 chr13 129 7680001-19200599 chr13 129 12840001-124800597 chr12 128 128 128 128 128 128 128 128 128 1		
Significant hits: chr9_123 73200001-73800598 chr7_2_6 27000001-27600598 chr12_94 55800001-56400598 chr15_127 75600001-75200599 chr5_175 104400001-105000599 chr8_115 68400001-69000597 chr10_132 78600001-79200399 chr2_354 211800001-132400597 chr22_30 17400001-10600597 chr22_310 17400001-10600599 chr2_312 186600001-106200599 chr2_312 186600001-106200599 chr10_177 105500001-106200599 chr12_99 5802001-58800598 chr10_15164 97800001-9200399 chr13_164 97800001-9200399 chr13_129 76800001-9200598 chr11_137 81600001-8200599 chr13_129 76800001-77400599 chr13_129 76800001-77400599 chr12_15 128400001-124800597 chr2_209 124200001-9200599 chr2_15 128400001-124800599 chr2_15 128400001-124800599 chr2_15 128400001-7400599 chr2_15 128400001-12800599 chr2_0(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).		
chr7_46 27000001-27600598 chr12_94 5500001-5600059 chr13_17 6960001-70200599 chr15_175 10440001-105000598 chr1_60 35400001-36000599 chr1_61 35400001-70200599 chr1_61 35400001-36000597 chr1_61 35400001-70200599 chr2_354 21800001-72400597 chr2_31 11800001-180200597 chr2_31 11800001-39600598 chr2_31 118060001-180200599 chr2_312 18660001-180200599 chr2_312 18660001-80200599 chr12_98 58200001-58800598 chr11_137 81600001-91200598 chr11_137 81600001-91200599 chr12_98 58200001-128200599 chr12_15 128400001-128400597 chr2_1058 124200001-12800599 chr12_125 128400001-12800599 chr12_125 128400001-12800599 chr2_1051 128400001-12800599 chr2_206 1242000599 chr2_215 128400001-12800599 chr2_206		
shr12 94 5500001-56400598 shr16 117 6960001-76200599 shr15 127 7560001-76200599 shr16 115 6400001-105000599 shr16 115 68400001-69200599 shr2 314 211800001-212400597 shr2 334 211800001-18000597 shr2 330 17400001-18000597 shr2 331 211800001-18000597 shr2 331 211800001-18000597 shr2 331 211800001-18000597 shr2 331 211800001-18000597 shr2 332 16600001-180200599 shr2 312 186600001-187200599 shr6 70 41400001-42000599 shr2 9 9600001-91200598 shr13 129 76800001-97400599 shr13 129 76800001-77400599 shr13 129 76800001-77400599 shr13 129 76800001-77400599 shr13 129 76800001-72400599 shr13 129		
shr15 127 75600001-76200599 shr1 60 3540001-36000599 shr1 61 3540001-36000597 shr1 61 3540001-36000597 shr1 61 3540001-3200599 shr1 7 10560001-722400597 shr2 354 21180001-18000597 shr2 354 21180001-39600597 shr2 354 21180001-18000597 shr2 350 17400001-36000597 shr2 312 105600001-39600599 shr2 312 105600001-39600599 shr4 70 1400001-42000599 shr4 70 1400001-38000599 shr4 70 1400001-58000599 shr4 70 1560001-32200599 shr4 71 329 76800001-77400599 shr4 1329 76800001-77400599 shr4 215 128400001-128000597 shr4 215 128400001-128000599 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).		
chr5_175 104400001-105000598 chr1_60 35400001-36000599 chr16_113 68400001-69000597 chr16_132 78600001-79200599 chr2354 211800001-1212400597 chr22_30 17400001-18000597 chr2_310 17400001-18000597 chr2_310 17400001-18000597 chr2_312 186600001-187200599 chr10_177 105600001-187200597 chr2_312 186600001-187200597 chr2_312 98600001-98400599 chr13_164 97800001-98400599 chr13_127 96800001-91200598 chr11_137 81600001-82200599 chr11_137 1600001-124800557 chr11_129 76800001-77400599 chr12_125 128400001-124800559 chr1_215 128400001-128000599	<u>chr16 117</u> 69600001-70200599	
shr1 60 35400001-3600599 chr2 113 68400001-79200599 chr2 354 211800001-79200599 chr2 354 211800001-79200599 chr2 354 211800001-79200599 chr2 354 211800001-79200599 chr2 354 211800001-165200599 chr2 312 18660001-165200599 chr12 177 105600001-165200599 chr12 96 58200001-58800599 chr12 98 58200001-58800598 chr12 98 58200001-58800598 chr12 98 58200001-79200599 chr13 129 76800001-79200599 chr12 137 128400001-12800597 chr12 132 128400001-128000597 chr12 131 12840000597 chr12 213 128400001-128000599		
chr8_115 68400001-69000597 chr10_132 78600001-79200599 chr22_30 17400001-18000597 chr22_31 17400001-106200599 chr10_177 105600001-106200599 chr2<312		
chr10_132 78600001-79200599 chr2 354 211800001-212400597 chr7 66 39000001-39600598 chr7 66 39000001-39600598 chr7 10560001-106200599 chr2 chr2 110560001-1106200599 chr2 chr2 186600001-187200599 chr2 chr2 1100001-28200599 chr1 chr1 15164 97800001-9800598 chr1 129 90600001-91200598 chr1 137 81600001-82200599 chr1 137 81600001-82200599 chr1 137 81600001-9200599 chr1 137 81600001-124800597 chr2 286 124200001-128000599 chr2 128 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).		
shr2 354 211800001-212400397 shr22 30 17400001-18000397 shr22 30 17400001-39600398 shr22 312 10560001-106200399 shr2 312 10660001-187200399 shr2 312 10660001-187200399 shr2 312 10660001-187200399 shr2 312 10660001-187200399 shr2 312 10660001-98400399 shr15 164 97800001-98400399 shr12 32 90600001-98400398 shr2 132 96600001-98400399 shr13 129 76800001-77400599 shr13 129 76800001-77400599 shr2 200 124200001-124800397 shr13 129 76800001-72400599 shr13 129 76800001-72400599 shr12 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).		
chr7_66 39000001-39600598 chr10_177 10560001-106200599 chr2_312 18660001-187200597 chr6_70 41400001-42700599 chr12_98 58200001-58800598 chr12_98 58200001-58200599 chr12_98 58200001-58200599 chr12_98 58200001-91200598 chr13_129 76800001-97200599 chr13_129 76800001-77400599 chr2_13 124200001-124800597 chr12_13 124400001-128000599 chr2_13 124400001-128000599 chr2_13 124400001-128000599 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).		
chr10_177 105600001-106200399 chr2_312 106600001-187200397 chr6_70 41400001-42000399 chr13_164 97800001-98400399 chr14_296 5820001-58800598 chr13_129 96600001-91200599 chr13_129 76800001-77400399 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr13_129 76800001-724000599 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).		
chr2_312 186600001-187200597 chr6_70 41400001-42000599 chr12_90 58200001-5800598 chr12_90 58200001-5800598 chr11_137 6160001-62200599 chr12_139 76800001-77400599 chr12_13 124200001-124800597 chr12_13 124200001-124800597 chr12_13 124200001-124800597 chr12_13 128400001-129000599		
chré 70 41400001-42000599 chr15_164 97800001-98800598 chr29_152 90600001-98200598 chr29_152 90600001-91200598 chr13_129 76800001-77400599 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr13_129 76800001-72400599 chr12_129 124200001-124800597 chr12_15 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).		
chr15 164 97800001-98400599 chr12 98 58200001-58800598 chr2 98 9060001-91200598 chr11 137 81600001-92200599 chr2 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).		
chr12 99 \$8200001-5800598 chr9 132 90600001-91200598 chr11 137 81600001-82200599 chr13 129 76800001-77400599 chr12 215 124200001-124800597 chr12 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).		
chr11137 81600001-82200599 chr13129 76800001-72400599 chr1215 124200001-124800597 chr1215 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).		
chr13 129 76800001-77400599 chr2 208 124200001-124800597 chr1 215 128400001-12900599 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).		
chrX 208 chr1 213 124200001-124800597 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).		
chr1_213 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).		
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).		
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).		
Individual ions scores > 68 indicate identity or extensive homology ($p < 0.05$).	Probability Based Mowse Score	
Individual ions scores > 68 indicate identity or extensive homology ($p < 0.05$).	Constraint (1987)	
	induvidual ions scores ∼ os inducate identity or extensive homology (p<0.00).	
] Done	Local intranet

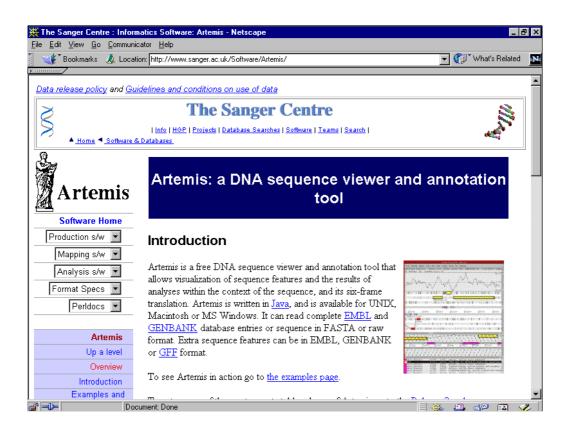
This is the result of searching our standard data set against the unmasked human genome assembly.

_		rch Results - Mie		Explorer							8 ×
		View Favorites				1 0	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,				
ψE	iack 🔻 =	* • 🙆 🙆 🖆	} 🖓 Search [🚡 Favorites 🛛 🕃	History	à• 🎒					
Addre	iss 🙋 hi	tp://dell5000/mase	:ot/cgi/master_res	ults.pl?file=/data	a/20001207/F	001327.	dat				Links
	chr9 1	22 Tat	-1 4	91 Peptide		4. 12					
		not retriev			s matche	u . 15					
	heck t	o include t	his hit in	archive rep	ort						
											_
	luery	Observed	Mr(expt)	Mr(calc)			Score		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		
☑	<u>93</u>	775.76	1549.50	1549.81	-0.31	0	69	1	GTDVNVFNTILTTR		
☑	<u>98</u>	547.49	1639.45	1639.77	-0.32	1	(41)	1	DLAKDITSDTSGDFR		
•	<u>99</u>	820.75	1639.48	1639.77	-0.29	1	52	1	DLAKDITSDTSGDFR		
	<u>101</u>	840.24	1678.47	1677.90	0.57	1	19	7	KGTDVNVFNTILTTR		
✓	<u>103</u>	851.77	1701.52	1701.88	-0.36	0	82	1	GLGTDEDTLIEILASR		
\checkmark	105	870.21	1738.41	1738.73	-0.32	0	99	1	SEDF GVNEDLAD SDAR		
V	<u>149</u>	785.91	2354.72	2355.15	-0.43	0	66	1	GGP GSAV SPYP TFNP S SDVAALHK		
											-
	chr7_4	6 Tota not retriev		04 Peptide	s matche	d: 7					
_		o include t		-	nort.						
	luery	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide		
	1	430.11	429.10	429.22	-0.12	0	7	4	DAPK		
•	<u>33</u>	544.15	1086.29	1086.48	-0.19	0	34	1	NYYEQWGK		
•	<u>43</u>	594.71	1187.41	1187.64	-0.23	0	66	1	IDTIEIITDR		
☑	66	689.19	1376.36	1376.62	-0.26	0	44	1	GGGGNF GP GP GSNFR		
~	102	565.82	1694.44	1694.76	-0.32	0	42	1	GFGFVTFDDHDPVDK		
ন	117	899.82	1797.63	1797.91	-0.29	0	63	1	LFIGGLSFETTEESLR		_
											▶
] Dor	ne									🚝 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.

F149977.mas - WordF	ad
ile <u>E</u> dit ⊻iew Insert F	
D e 1 -	
	/ Jiasep IIIc/ uata/20000310/11755/1.uat /mags=1549.79
	/ acore=54
	/rank=1
	/translation="AGANLFELENFVAR"
BLASTCDS	170915170951
	/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
BLASTCDS	/translation="EFLVWAVNDAIER" 12888.12927
DLASICOS	/12000.1292/ /1abel=070
	/ raber-gro / colour=2
	/orice="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(23999812400026)
	/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
	/fank-1 /translation="ELPGGALTLVNDAGMR"
BLASTCDS	28774602877499
DEADICED	/1abl=073
or Help, press F1	, Table 213

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.

Artemis Entry Ed File Entries Select		<mark>gbk</mark> Edit Crea	te VAkite	Dienley															_	
										,										
Selected featu	re: bases		43 (/1	abel=	243 /		ar=2 /	mass=1	187.64	/scor	e=66 ,	/rank	$=1/t_1$	ansla	tion=	"IDTI		DR",	/note	="Mas
-++-+																				
			C 033			╟╢╴			Q43				} 66				-	++-		Π
	_	N	C 200	N		N	_		245	56			N	_	_	_		_	_	-
		V		/		7		exc	-	/	VV	_ <u>/</u>	V	0.0.4			VV			
E	E	exon2		IRPA2B		HNRE					exo E		HNRP	-		E		INRPA		
4800	5100	5	100	5	700		600	0	630	0	6	500		6900		72	200		750	
																1.11		1.0.01		_
•					_	-						-								• •
EKLI	PLR	# L	L I	G S	L	Е	R K	ΕÅ	LA	г	L L	L	МТ	M :	ΓL	W	IK	s	Y	v _
K N * Y	H * D	N Y	* +	A	V W	К	E K	R I	LW	г с	Y F	*	* F	*	s c	G	#	N R	M	#
G K I D I	IE	E I '	T D) Q	S	G K	K	R G	F G	F N	7 Т	F D	D	H D	Р	V D	×.	I	v c	K
GAAAAATTGATAC	CATTGAGA?	TAATTA	CTGATA	GGCAG	TCTG	GAAA	GAAAA	GAGGC:	TTTGGC	TTTGT	TACTI	TTGA	TGACO	ATGAT	FCCTG	TGGA	TAAA	ATCG	TAT	TAA
60 6270	6280	1	6290		6300		631	n	6320		6330		634	n	63	50		6360		6
CTTTTTAACTATG			1		AGAC	<u>ر.</u>			AAACCG	*****	_		ACTICS	macmi			8 mmm			8 mm
	W T Y		SACIAI		NOAC	~ .		C 1000				w T		un r		NCCI.		TNOC		
SFNIG	и г	C N	5 I	ь т	R	э Г	F.	S A	к А	K I	1 5	V I	v	I IV	R	п І	F.	D	I T	ц
FISV	M S I	I V	S L	С	D P	F	FL	PI	K P	к т	VF	s	S W	S	G T	S	L	ΙT	Н	L
FFQY	W Q S	L #	Q Y	A T	Q	F	S F	L S	QS	Q	# K	Q	H G	ΗI	Q	P	Y F	R	I	Y
•																				• •
BLASTCDS	54		5511						mass=19						-					
BLASTCDS	54		5511						mass=1'										reesl	R
BLASTCDS	55		5535	Masco	ot mat	tch,	query	=33, m	ass=108	36.48,	score	=34,	rank=	1, se	quenc	e=NYY	EQWG	K		
intron	55	53 5	5711																	
exon	57	12 !	5858																	-

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.

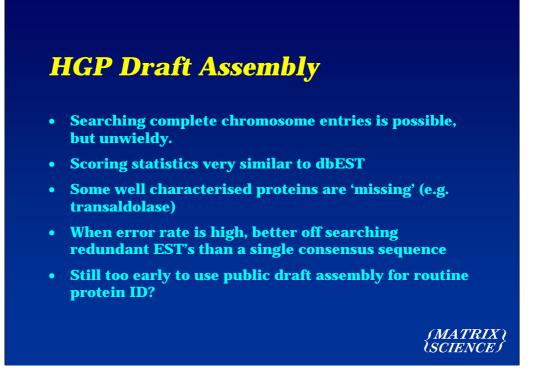
Not only does this allow us to zoom and pan around these extremely long sequences, it also allows us to view the peptide matches found by Mascot in the context of all the existing annotations. This gives us a powerful way to present the results of MS based searching complete genomes.

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

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.



The main conclusions of our database comparison study are listed.



http://www.matrixscience.com