

Mixture Mode for Peptide Mass Fingerprinting

ASMS 2003

MATRIX
SCIENCE

Mixture Mode: New in Mascot 1.9

- All peptide mass fingerprint searches now test for the possibility that the sample is a mixture of proteins.
- Mascot will automatically detect and report statistically significant protein mixtures of up to six components

ASMS 2003



One of the new features in Mascot 1.9 was a mixture mode for peptide mass fingerprints. Although it is always preferable to use MS/MS when dealing with a complex mixture or looking for a minor component, many Mascot users had requested us to improve the handling of a simple mixture in a peptide mass fingerprint search.

The way this has been implemented is transparent to the user. Mascot automatically looks for mixtures of up to six components and will report any that are statistically significant.

Mascot Search Results - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/master_results.pl?file=../data/20030602/FIncObTE.dat

Probability Based Mowse Score

Score is $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event.
 Protein scores greater than 73 are significant ($p < 0.05$).

Concise Protein Summary Report

[Switch to full Protein Summary Report](#)

To create a bookmark for this report, right click this link: [Concise Summary Report \(../data/20030602/FIncObTE.dat\)](#)

Re-Search All Search Unmatched

- Mixture** **Total score: 254** **Peptides matched: 18**
 Components (only one family member shown for each component):

B53193	Mass: 48000	Total score: 144	Peptides matched: 11
hedgehog homolog vhh-1 - rat			
CAA57179	Mass: 46883	Total score: 85	Peptides matched: 7
HSKKA2 NID: - Homo sapiens			
- | | | | |
|------------------------------|--------------------|-------------------------|-----------------------------|
| B53193 | Mass: 48000 | Total score: 144 | Peptides matched: 11 |
| hedgehog homolog vhh-1 - rat | | | |

Done Internet

In this example, which shows a concise protein summary report, we have an example of a simple mixture. Two proteins of similar size - possibly co-eluting on a gel. If we want to see more detail, just click the link for the full protein summary report

Mascot Search Results - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/master_results.pl?file=../data/20030602/F1ncObTE.dat&REPTYPE=protein

15. Q8M3V7	14025	77	Sonic hedgehog (Fragment).- Suncus murinus (House shrew) (Musk shrew).
16. Q9XSI6	16008	74	Sonic hedgehog (Fragment).- Bos taurus (Bovine).
17. Q8Q6N4	11518	61	Sonic hedgehog (Fragment).- Scyliorhinus canicula (Spotted dogfish) (Spotted catsh).
18. Q89800	113391	58	Dihydropyrimidine dehydrogenase (EC 1.3.1.2).- Rattus norvegicus (Rat).
19. Q9W6C1	15856	57	Sonic hedgehog protein (Fragment).- Eleutherodactylus coqui (Puerto Rican coqui).
20. Q9YGV7	16760	54	Sonic hedgehog (Fragment).- Ambystoma mexicanum (Axolotl).

Results List

1. Mixture 1 Score: 254
 Components: 1. [B53193](#) hedgehog homolog vhh-1 - rat
 2. [CAA57179](#) HSHKA2 NID: - Homo sapiens

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Comp	Peptide
708.38	707.37	707.38	-0.00	362	367	0	1	AFAPFR
710.37	709.36	709.36	0.00	159	164	0	1	YGHLAR
712.42	711.41	711.41	-0.00	295	300	0	1	VRPGQR
759.36	758.35	758.36	-0.00	218	224	0	1	DLSPGDR
805.48	804.47	804.47	-0.00	40	46	0	1	LTPLAYK
977.49	976.48	976.48	0.00	146	154	0	1	AVDITTSDR
1003.52	1002.51	1002.52	-0.00	124	132	0	2	MVNVIDNAK
1006.41	1005.40	1005.40	0.00	89	97	0	1	DEENTGADR
1227.67	1226.66	1226.66	0.00	150	160	0	2	QLVEADINGLR
1287.79	1286.78	1286.78	0.00	368	380	0	1	LAHALLAALAPAR
1445.69	1444.68	1444.68	0.00	73	84	0	2	QLFQENAELESR
1464.68	1463.67	1463.67	-0.00	291	303	0	2	DSLNTLTSEAR
1502.79	1501.78	1501.78	0.00	234	245	0	1	LLYSDFLFLDR
1622.87	1621.86	1621.87	-0.00	206	220	0	2	LNIEVDAAPVLDLR
1776.85	1775.84	1775.84	0.00	165	179	0	1	LAVEAGFDVVYTESK
1851.99	1850.98	1850.98	-0.00	275	290	0	2	TVNTLELELAQHSRLR
2469.03	2468.02	2468.03	-0.00	125	145	0	1	VTEGWDEDGHSEESLHYEGR
2475.17	2474.16	2474.16	-0.00	85	105	0	2	IQEASHSQVLTHTPDYQSHFR

No match to: 930.55, 1126.57

2. [B53193](#) Mass: 48000 Score: 144
 hedgehog homolog vhh-1 - rat

Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
708.38	707.37	707.38	-0.00	362	367	0	AFAPFR
710.37	709.36	709.36	0.00	159	164	0	YGHLAR
712.42	711.41	711.41	-0.00	295	300	0	VRPGQR

Isn't that pretty? Matches to individual protein are colour coded.

Mascot is scoring the match for the complete set of experimental mass values to the in silico digest products of the putative protein mixture. It isn't a subtractive approach, where we compare the experimental mass values to one protein, remove the matched values, and then compare the remainder with the next protein. It would be very difficult to provide a true probability-based score for the subtractive approach. Also, it is less accurate because, in large data sets, there are likely to be mass values that could match to more than one of the proteins in the mixture.

Mascot Search Results - Microsoft Internet Explorer			
Address: http://www.matrixscience.com/cgi/master_results.pl?file=../data/20030602/F1ncOz5a.dat			
1. Mixture	Total score: 287	Peptides matched: 32	
Components (only one family member shown for each component):			
gi 1363944	Mass: 51206	Total score: 184	Peptides matched: 19
keratin 16, type I, cytoskeletal - human			
gi 15559584	Mass: 59981	Total score: 102	Peptides matched: 14
Similar to keratin 6A [Homo sapiens]			
2. Mixture	Total score: 286	Peptides matched: 32	
Components (only one family member shown for each component):			
gi 1363944	Mass: 51206	Total score: 184	Peptides matched: 19
keratin 16, type I, cytoskeletal - human			
gi 1346349	Mass: 60030	Total score: 101	Peptides matched: 14
Keratin, type II cytoskeletal 6F (Cytokeratin 6F) (CK 6F) (K6F keratin)			
3. Mixture	Total score: 198	Peptides matched: 26	
Components (only one family member shown for each component):			
gi 15559584	Mass: 59981	Total score: 102	Peptides matched: 14
Similar to keratin 6A [Homo sapiens]			
gi 15431310	Mass: 51589	Total score: 100	Peptides matched: 13
keratin 14; cytokeratin 14 [Homo sapiens]			
4. Mixture	Total score: 197	Peptides matched: 26	
Components (only one family member shown for each component):			
gi 1346349	Mass: 60030	Total score: 101	Peptides matched: 14
Keratin, type II cytoskeletal 6F (Cytokeratin 6F) (CK 6F) (K6F keratin)			
gi 15431310	Mass: 51589	Total score: 100	Peptides matched: 13
keratin 14; cytokeratin 14 [Homo sapiens]			
5.	gi 1363944	Mass: 51206	Total score: 184 Peptides matched: 19
keratin 16, type I, cytoskeletal - human			
	gi 24430192	Mass: 51236	Total score: 184 Peptides matched: 19
keratin 16; keratin, type I cytoskeletal 16; cytokeratin 16 [Homo sapiens]			
	gi 186685	Mass: 50668	Total score: 100 Peptides matched: 14
keratin type 16			
	gi 129739476	Mass: 18911	Total score: 51 Peptides matched: 6

Even though we try to present the results as simply and clearly as possible, there are times when things get complicated. In this example, we have a mixture of keratins. Trouble is, there are many homologous keratins in the database. This means that there are several possible combinations of high scoring proteins. Hit 1 is a mix of keratin 16 and keratin 6A while hit 2 is a mixture of the same keratin 16 combined with keratin 6F. Hits 3 and 4 are lower scoring, but still statistically significant improvements over the score for a single component at hit 5. To minimise the complexity, we only show a single family member for each component of a mixture. Otherwise, we would have a much larger number of permutations. The good news is that this example is pretty much a worst case. Most results are much simpler

Mascot Search Results - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/master_results.pl?file=../data/20030602/FTnofnet.dat&REPTYPE=concise

Probability Based Mowse Score

Score is $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event. Protein scores greater than 58 are significant ($p < 0.05$).

A mixture will only be reported if it is statistically significant

Otherwise, the hit list would be swamped with possible mixtures

Concise Protein Summary Report

[Switch to full Protein Summary Report](#)

To create a bookmark for this report, right click this link: [Concise Summary Report \(A.H. 90k\)](#)

Re-Search All Search Unmatched

- Mixture** **Total score: 148** **Peptides matched: 21**

Components (only one family member shown for each component):

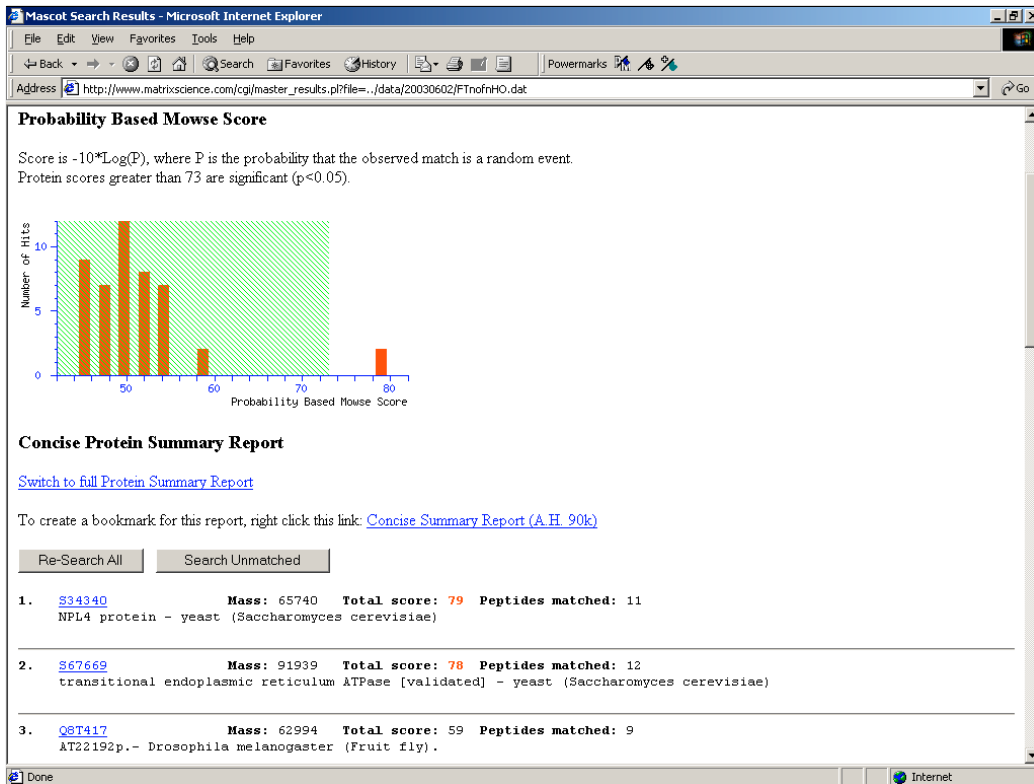
S34340	Mass: 65740	Total score: 79	Peptides matched: 11
NPL4 protein - yeast (Saccharomyces cerevisiae)			
S67669	Mass: 91939	Total score: 78	Peptides matched: 12
transitional endoplasmic reticulum ATPase [validated] - yeast (Saccharomyces cerevisiae)			
- [S34340](#) Mass: 65740 Total score: 79 Peptides matched: 11

NPL4 protein - yeast (Saccharomyces cerevisiae)

As I mentioned earlier, a mixture will only be reported if it is statistically significant. What does this mean?

If you think about it, you can always increase the number of matches by throwing in additional proteins. In this example, we have a mixture of two proteins that match 21 experimental mass values out of 32. Imagine we have another protein in the database that could match 1 or 2 more of the unmatched values. Chances are, the score would go down, because getting 23 or 24 matches from 3 proteins is statistically less interesting than getting 21 matches from 2 proteins. But what about a third protein that adds 3 or 4 or 5 more matches? At some point, the score will start increasing. Should we report this as a possible mixture?

Probably not, because in a database of (say) 1 million proteins, there are 10^{18} combinations of 3 proteins. So, it is essential to limit the results to mixtures that are statistically significant. Otherwise, the hit list becomes swamped with meaningless combinations that have marginally higher scores.



To demonstrate this, we take the previous search, which used a taxonomy filter of fungi, and repeat it against the complete MSDB database.

Even though the matches are identical, we no longer report a mixture because the significance threshold is higher for a search of the complete database, so the mixture is no longer significant

Mascot Search Results - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/master_results.pl?file=../data/20030602/FTncOeYn.dat

Probability Based Mowse Score

Score is $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event.
Protein scores greater than 52 are significant ($p < 0.05$).

Search Parameters

Type of search : Peptide Mass Fingerprint
 Enzyme : Trypsin
 Mass values : Monoisotopic
 Protein Mass : Unrestricted
 Peptide Mass Tolerance : ± 0.2 Da
 Peptide Charge State : Mr
 Max Missed Cleavages : 1
 Number of queries : 88

Concise Protein Summary Report

[Switch to full Protein Summary Report](#)

To create a bookmark for this report, right click this link: [Concise Summary Report \(../data/20030602/FTncOeYn.dat\)](#)

Re-Search All Search Unmatched

- Mixture** **Total score: 214** **Peptides matched: 32**
 Components (only one family member shown for each component):
[A24709](#) **Mass: 44341** **Total score: 86** **Peptides matched: 14**
 ketol-acid reductoisomerase (EC 1.1.1.86) ILV5 - yeast (Saccharomyces cerevisiae)
[1CNP](#) **Mass: 33003** **Total score: 71** **Peptides matched: 10**
 cytochrome-c peroxidase (EC 1.11.1.5) mutant (W191G) (with 1,2-dimethylimidazole) - yeast (Saccharomyces cerevisiae)
[DSBYN](#) **Mass: 25758** **Total score: 64** **Peptides matched: 9**
 superoxide dismutase (EC 1.15.1.1) (Mn) precursor [validated] - yeast (Saccharomyces cerevisiae)

Done Internet

What about more complex mixtures? Well, although Mascot is looking for the possibility of up to a 6 component mixture, we haven't yet seen any real examples of mixtures with more than 3 components.

This shows an example of a real 3 component mixture. As you can see, the mass tolerance is nothing special, and we are only matching a total of 32 values out of 88, yet we have an excellent score for the mixture

Mascot Search Results - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/master_results.pl?file=../data/20030602/FTncOeYn.dat&REPTYPE=protein

1. Mixture 1 Score: 214

Components: 1. A24709 ketol-acid reductoisomerase (EC 1.1.1.86) ILV5 - yeast (Saccharomyces cerevisiae)
 2. 1CMP cytochrome-c peroxidase (EC 1.11.1.5) mutant (W191G) (with 1,2-dimethylimidazole) - yeast (Saccharomyces cerevisiae)
 3. DSBYN superoxide dismutase (EC 1.15.1.1) (Mn) precursor [validated] - yeast (Saccharomyces cerevisiae)

Observed	Mr (expt)	Mr (calc)	Delta	Start	End	Miss	Comp	Peptide
721.43	721.43	721.44	-0.01	241	246	0	2	YLSIVK
784.43	784.43	784.47	-0.04	28	34	0	3	VTLPLDK
788.35	788.35	788.39	-0.04	285	291	0	2	TLEEQGL
797.35	797.35	797.43	-0.08	71	76	0	1	LDDYFK
819.29	819.29	819.39	-0.10	377	382	0	1	NMEIWK
882.57	882.57	882.50	0.07	88	94	0	2	FLEPIHK
890.47	890.47	890.52	-0.05	19	26	0	2	VYNAIALK
916.48	916.48	916.53	-0.05	5	13	1	3	TAAANLTKK
1020.51	1020.51	1020.53	-0.02	276	284	0	2	DAPSPFIFK
1054.65	1054.65	1054.57	0.08	69	76	1	1	EKLDDYFK
1057.55	1057.55	1057.60	-0.05	90	98	0	3	MIAIQQNIK
1115.67	1115.67	1115.62	0.05	368	376	1	1	LEKELDTIR
1128.61	1128.61	1128.61	0.00	214	222	0	3	AIWNVVWVK
1141.55	1141.55	1141.58	-0.03	116	126	0	1	AAIEDGVVPGK
1148.65	1148.65	1148.61	0.04	127	136	0	1	NLFTVEDAIK
1168.62	1168.62	1168.66	-0.04	98	108	0	1	DNGLNVIIGVR
1185.63	1185.63	1185.69	-0.06	89	98	1	3	KMIAIQQNIK
1294.67	1294.67	1294.71	-0.04	164	176	0	2	EVVALNGAHALGK
1304.71	1304.71	1304.71	0.00	127	137	1	1	NLFTVEDAIKR
1354.65	1354.65	1354.62	0.03	355	365	0	1	SLEFNSQPDYR
1431.85	1431.85	1431.79	0.06	151	164	0	3	LAGVQSGMAFIVK
1528.80	1528.80	1528.77	0.03	166	178	0	1	TLVYSHGSPVFK
1576.83	1576.83	1576.80	0.03	132	145	0	3	AIDEQFGSLDELIK
1609.85	1609.85	1609.77	0.08	212	226	0	1	GINSYAVMNDVTGK
1636.81	1636.81	1636.75	0.06	73	87	0	2	EFNDPSNAGLQNGFK
1751.87	1751.87	1751.91	-0.04	333	347	0	1	NALKPVFDLYESTK
1957.89	1957.89	1957.92	-0.03	224	240	0	2	SGYHMLPTDYSLIQDFK
2098.21	2098.21	2098.17	0.04	179	197	1	1	DLTHVEPPKDLVILVAFK
2243.17	2243.17	2243.10	0.07	77	97	0	1	NDTFALIGYSGYGGGLNLR
2285.21	2285.21	2285.10	0.11	27	45	1	2	LREDDYDNNYIGYSPVLR
2480.27	2480.27	2480.20	0.07	35	55	0	3	WDFGALFPYISGQINELHYTK
2779.51	2779.51	2779.35	0.16	95	120	0	2	EFPWISSGDLFSLGGVTAVQEMQGGK

No match to: 469.72, 801.45, 805.37, 832.41, 834.39, 841.49, 844.37, 863.41, 869.65, 903.57, 912.56, 939.52, 958.53, 974.55,

The limitation we're up against is that the components need to be reasonably well balanced. In this case, the matches are distributed 14 - 10 - 9.

The statistics of the process make it difficult or impossible to pick up a minor component, and the situation gets worse as the number of components increases.

At the end of the day, the only way to handle a complex mixture and detect minor components is with MS/MS. However, we hope you agree that mixture mode is a useful addition to the Mascot peptide mass fingerprint search