

Statistical Validation of Database Search Results

Alexey Nesvizhskii

nesvi@systemsbiology.org

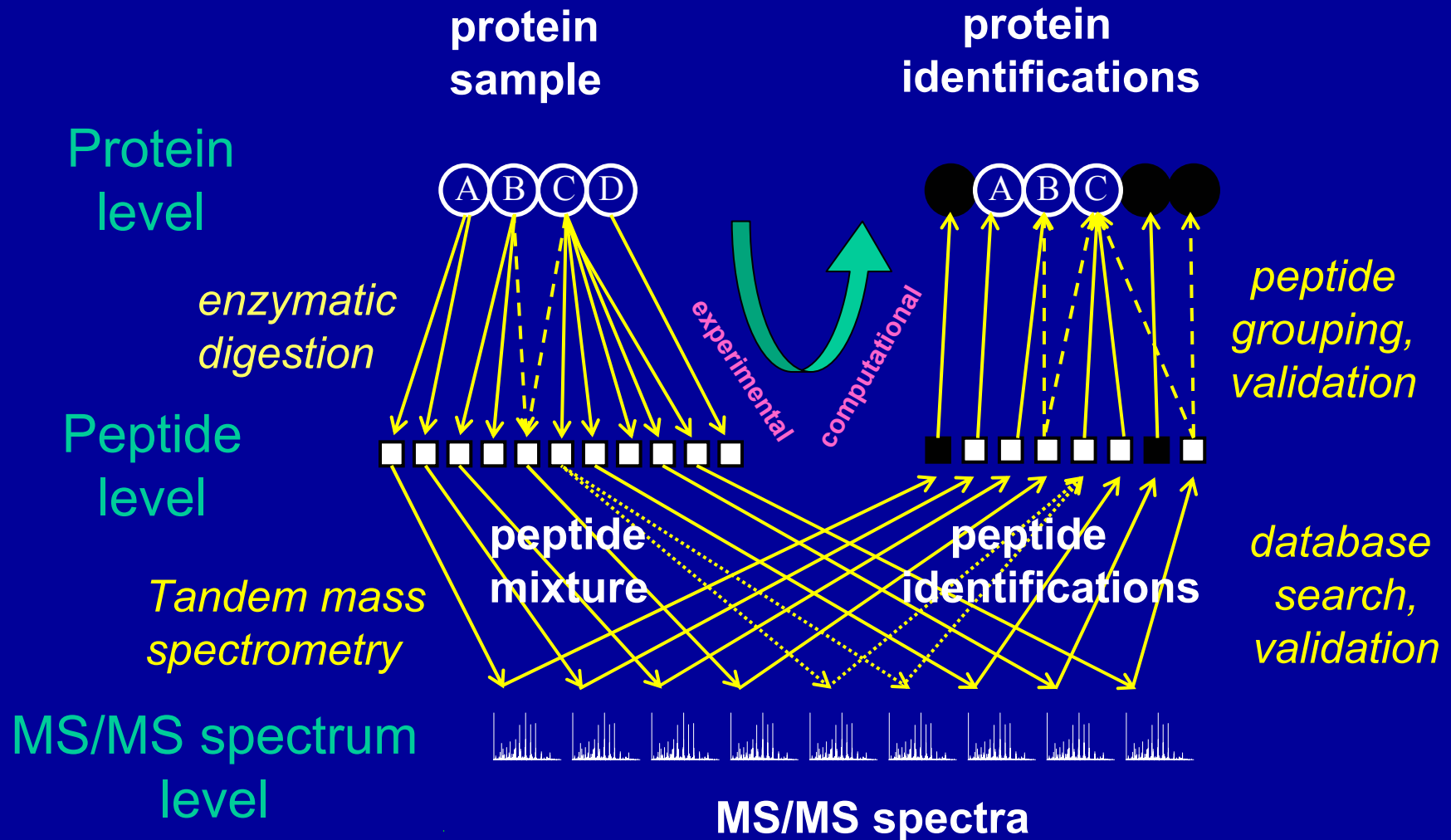
Institute for Systems Biology
Seattle, WA
U.S.A.

Standards in Proteomics
January 4, 2005

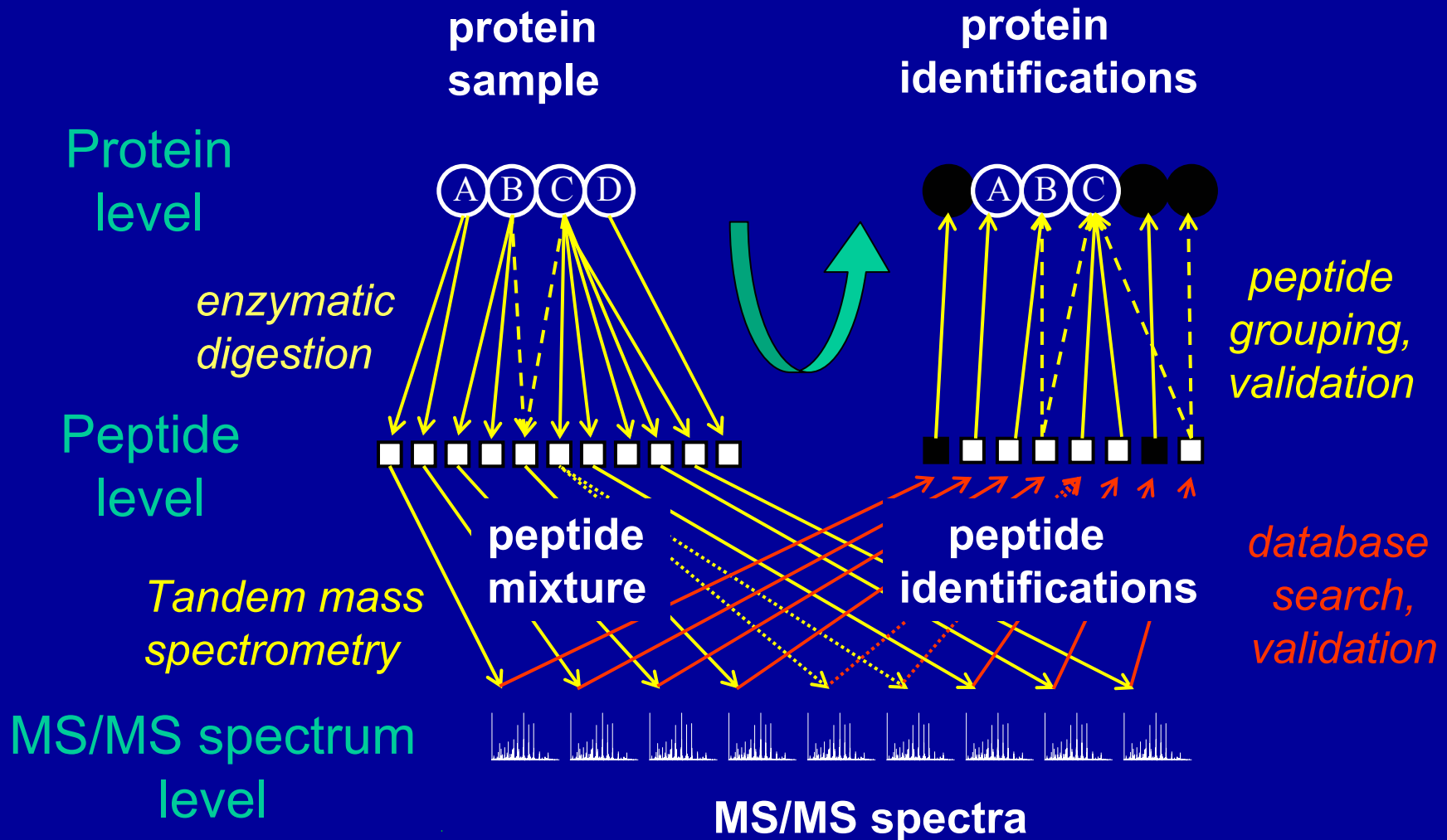
Outline

- High throughput MS/MS-based proteomics
- Peptide identification and validation
 - p-value, expectation value
 - Probabilities, error rates
- Protein Identification
 - error rate inflation (due to non-random peptide grouping)
 - protein inference problem (degenerate peptides)

Shotgun Protein Identification

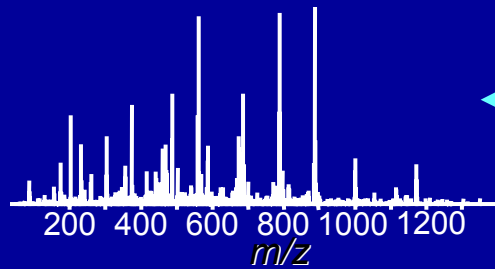


Shotgun Protein Identification



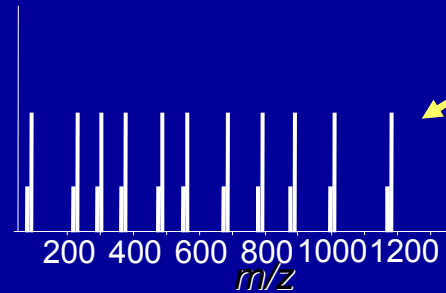
MS/MS Database Search

Acquired MS/MS
spectrum



Sequence Database

Theoretical spectrum



ISLLDAQSAPLR
VVEELCPTPEGK
DLLLQWCWENGK
ECDVVSNTIIAEK
GDAVFVIDALNR
VPTPNVSVVDLTNR
SYLFCMENSAEK
PEQSDLRSWTAK

correlate

similarity score

best matching database peptide

Algorithms: SEQUEST, Mascot, Sonar, SpectrumMill, ...

best matching database peptide can be correct or
incorrect identification

Validation of Peptide Identifications

Netscape: INTERACT by J.Eng, Institute for Systems Biology

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://regis.systemsbio.net/data/search/akeller/HINF_PLUS_CONTROL/interact.htm What's Related

Red Hat Network Training Support Software Hardware Developers Embedded Search Documentation Downloads Shop

sort/Restore: Apply filtering below ... no sort or restore Help

FILE: [/data/search/akeller/HINF_PLUS_CONTROL/interact-da] Tryptic: 1 2 MaxMissed: 3 DelRows: Xcorr: +1 +2 +3 dCn: RSp: InclRA: MarkRA: XPRESS: GO

724	/sergei digest A full 01.1818.1815.2	1945.2	(-0.5)	3.2763	0.263	559.9	1	16/32	SP P02754 LACB BOVIN	R.VYVEELKPTPEGDLEIL.L
969	/sergei digest A full 01.2359.2361.3	2854.3	(-1.2)	3.2389	0.312	447.1	1	25/36	SP P00321 CAM2 BOVIN	R.TLNFHAEQPELLMLANWPAQPLK.N
328	/sergei digest A full 01.1063.1065.2	1582.8	(-0.0)	3.2171	0.384	1508.4	1	18/24	SP P00321 CAM2 BOVIN	K.FRAELHLVHNTK.V
340	/sergei digest A full 01.1031.1033.2	974.1	(+0.2)	3.2106	0.316	454.9	1	14/15	SP P00321 CAM2 BOVIN	K.VLDLADSIK.T
796	/sergei digest A full 01.1949.1949.2	2215.5	(-0.2)	3.2058	0.435	531.2	1	21/36	SP P02754 LACB BOVIN	V.VYVEELKPTPEGDLEIL.L
1057	/sergei digest A full 01.2573.2575.3	4013.3	(-2.2)	3.1870	0.172	179.7	161	24/135	SW.K205.HUMAN	G.AGAGGGVGFGGAGGSSFGFGAGGAGGFGSLGG
1278	/sergei digest A full 01.3508.3505.3	2709.1	(+1.5)	3.1834	0.198	686.4	1	27/100	SP P02754 LACB BOVIN	K.VAGTVYSLANASDSLILDAQSAPLR.V
523	/sergei digest A full 01.1417.1425.2	1610.9	(-0.3)	3.1691	0.388	1389.3	1	20/26	SP P00489 PMS2 RABIT	K.VHCNPNLSLFDVQVK.R
60	/sergei digest A full 01.0579.0581.2	1246.3	(+0.2)	3.1600	0.383	1196.1	1	18/20	SP P02754 LACB BOVIN	R.TPEVDDEALEK.F
612	/sergei digest A full 01.1587.1589.3	2204.6	(+0.3)	3.1594	0.323	1098.0	1	28/80	SP Q29443 TRFE BOVIN	K.LKGGADAMSLDGGSLYIAGK.C
1360	/sergei digest A full 01.3977.3985.2	2709.1	(-0.1)	3.1502	0.445	494.4	1	14/50	SP P02754 LACB BOVIN	K.VAGTVYSLANASDSLILDAQSAPLR.V
922	/sergei digest A full 01.2247.2253.3	3966.5	(+2.1)	3.1231	0.035	211.9	53	21/135	HI0583	R.LQPTLADLPLFGEVGVDTYVCAPLKRAIQQ
300	/sergei digest A full 01.1017.1021.2	1640.9	(-0.2)	3.1126	0.468	406.3	1	18/28	SP P02769 ALBU BOVIN	R.KVPQUSTPTLVVSR.S
1026	/sergei digest A full 01.2503.2509.3	2854.3	(+0.8)	3.0998	0.206	371.4	2	23/36	SP P00321 CAM2 BOVIN	R.TLNFHAEQPELLMLANWPAQPLK.N
846	/sergei digest A full 01.2079.2079.2	1551.8	(-0.4)	3.0660	0.357	573.8	1	18/26	SP Q29443 TRFE BOVIN	R.TAGVNIPLMGLYSK.I
753	/sergei digest A full 01.1855.1871.3	2031.3	(+1.6)	3.0715	0.427	1025.6	1	27/76	SP P02754 LACB BOVIN	V.SLANASDSLILDAQSAPLR.V
576	/sergei digest A full 01.1513.1515.3	1679.9	(+0.1)	3.0641	0.289	468.0	1	21/52	SP P00489 PMS2 RABIT	R.LKGGADAMSLDGGSLYIAGK.C
69	/sergei digest A full 01.0603.0605.3	1250.4	(+0.4)	3.0525	0.115	1575.4	1	23/36	SP P02769 ALBU BOVIN	R.TPEVDDEALEK.F
1022	/sergei digest A full 01.2435.2439.3	2241.5	(-0.3)	3.0417	0.022	486.2	23	22/76	HI0053	G.SRTSKGEFFALDMLATNKI.N
935	/sergei digest A full 01.1759.1761.2	1890.1	(-0.2)	3.0407	0.420	462.0	1	17/28	SP P02769 ALBU BOVIN	R.HPTFYVAPLLYVANK.V
215	/sergei digest A full 01.0878.0875.2	1605.8	(-0.2)	3.0358	0.442	555.7	1	15/26	SP Q29443 TRFE BOVIN	K.DNPOTHYVAVAVK.K
1385	/sergei digest A full 01.4038.4101.2	2709.1	(+0.6)	3.0143	0.212	536.9	1	17/50	SP P02754 LACB BOVIN	K.VAGTVYSLANASDSLILDAQSAPLR.V
1126	/sergei digest A full 01.2765.2771.3	4062.5	(+1.2)	2.9953	0.230	535.1	1	23/148	HI1574	I.DLITLDDALRSQGVSDGVGFAYLRAELSNNT
106	/sergei digest A full 01.0538.0535.2	1194.4	(+0.1)	2.9868	0.278	330.3	1	15/12	SP P02754 LACB BOVIN	K.VYVLTDTYK.V
642	/sergei digest A full 01.1649.1649.3	2913.4	(-0.3)	2.9659	0.180	723.7	1	30/100	SP P02555 CAB2 BOVIN	V.WHOPHPLPTVHFFPQSVLSLSQSK.V
323	/sergei digest A full 01.1051.1053.3	1636.8	(+0.6)	2.9546	0.355	374.5	1	27/52	SP P02754 LACB BOVIN	R.TPEVDDEALEK.F
795	/sergei digest A full 01.1965.1967.3	2336.6	(+2.4)	2.9429	0.056	362.6	12	20/80	HI0437	F.TTDAVKIATAPAFRVNEKTEN.I
1088	/sergei digest A full 01.2645.2647.3	3336.8	(+1.7)	2.9239	0.096	130.0	79	23/128	HI1136	L.AQLSGGVKARCKOVHAGGAGGAGGVKLVQDV
132	/sergei digest A full 01.0729.0723.3	1342.6	(+0.3)	2.9069	0.228	624.1	1	19/40	SP P00634 PPE EGOLI	A.KTPENPLDNE.A
645	/sergei digest A full 01.1653.1655.2	1754.9	(-0.8)	2.8966	0.428	1130.7	1	17/26	SP I44006 GSP RABIT	K.LIQTWDEPQSNR.V
765	/sergei digest A full 01.1833.1837.3	2333.8	(+2.8)	2.8552	0.003	644.3	1	24/76	HI0452	A.TLPEDEHLADGKLLPLVLLK.L
310	/sergei digest A full 01.1031.1033.2	1066.2	(-0.1)	2.8501	0.212	874.5	1	14/16	SP P02754 LACB BOVIN	K.VYVLTDTYK.K
1018	/sergei digest A full 01.2437.2439.3	3325.8	(+2.2)	2.8437	0.210	246.5	2	22/108	SP P00321 CAM2 BOVIN	V.SSQMLKFTLTLNFHAEQPELLMLANWVR.P
327	/sergei digest A full 01.1057.1061.3	1582.8	(+0.8)	2.8489	0.276	1076.8	1	31/48	SP P00321 CAM2 BOVIN	K.FRAELHLVHNTK.V
686	/sergei digest A full 01.1731.1733.3	2264.6	(+0.1)	2.8453	0.065	384.2	6	21/72	HI0134	P.LNFFGNCNSPQYKNTFKK.F
731	/sergei digest A full 01.1955.1957.3	3077.9	(+0.8)	2.8390	0.111	120.8	8	19/104	HI0635	A.DLKTTHPFAIGALLVWGKLNKLVNRH.C
353	/sergei digest A full 01.1123.1125.3	1636.8	(-0.3)	2.8353	0.272	580.5	1	22/52	SP P02754 LACB BOVIN	R.TPEVDDEALEK.F
145	/sergei digest A full 01.0755.0757.2	1194.4	(-0.3)	2.8348	0.242	331.9	1	15/18	SP P02754 LACB BOVIN	K.VYVLTDTYK.V
721	/sergei digest A full 01.1805.1809.3	3491.9	(-1.9)	2.8180	0.072	163.1	8	20/120	HI1188	Q.KVKCKEFGFGTVINVEGSEHNTRLQIAFQAO
875	/sergei digest A full 01.2147.2149.3	3518.3	(+1.0)	2.8168	0.089	211.1	12	22/135	HI1015	L.GSICGVIGLVGMVGVLEVSAGAPKMAVSFI
427	/sergei digest A full 01.1243.1245.3	2909.3	(-1.0)	2.8109	0.033	204.5	125	23/100	HI1237	T.HVLOGGFPRPADNKLGLQPNFEGIN.P
770	/sergei digest A full 01.1905.1911.3	2351.9	(+1.2)	2.8101	0.177	268.9	132	19/38	HI1065	I.GIAETHPSPVPLFARLGR.V
926	/sergei digest A full 01.2283.2255.3	2872.1	(-2.0)	2.7800	0.238	524.0	1	25/112	HI0254	T.VGATVVASADLPGSSVTVGSAVSTIG.N
753	/sergei digest A full 01.1879.1883.3	3473.0	(-2.4)	2.7784	0.011	112.1	11	23/124	SW.TRPE BOVIN	I.VYHVSNTNLNNDIMLTKLSAASLNSEVAS
1300	/sergei digest A full 01.3637.3643.3	3519.9	(-2.4)	2.7632	0.114	270.7	13	20/132	HI0906	K.AEALGEPVGAULVDDARHITGEGWNLSTVO
1242	/sergei digest A full 01.3131.3133.3	3847.3	(-2.6)	2.7632	0.032	184.7	202	19/140	HI0528	Q.LATHGVLSQSFVRILQADPTLVGNLNAVTR
605	/sergei digest A full 01.1579.1579.2	1441.6	(+0.7)	2.7514	0.481	934.6	1	17/22	SP P00489 PMS2 RABIT	K.LLSYVDDAEPIK.D
749	/sergei digest A full 01.1851.1853.3	2327.7	(+0.7)	2.7463	0.176	449.9	8	25/152	SW.KLIGT HUMAN	S.DBNHPTLDEPLAKNK.N
1361	/sergei digest A full 01.3977.3985.2	4062.8	(+0.2)	2.7463	0.118	355.3	3	25/152	HI0219	I.PPILQTPVVVVFPLFQEVWGVSK.V
1053	/sergei digest A full 01.2579.2587.2	2462.0	(-0.6)	2.7408	0.311	374.1	1	15/44	SP P02566 CAB2 BOVIN	I.PPILQTPVVVVFPLFQEVWGVSK.V

2
5
1
3
9
4

spectrum scores protein peptide

output from
MS/MS database
search algorithms

thousands of
spectra

- which peptide
identification are
correct?
- how to filter data?

Threshold Model

Netscape: INTERACT by J.Eng. Institute for Systems Biology

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://regis.systemsbiology.net/data/search/akeller/HINF_PLUS_CONTROL/interact.htm What's Related

Red Hat Network Training Support Software Hardware Developers Embedded Search Documentation Downloads Shop

Sort/Restore: Apply filtering below ... no sort or restore Help

FILE: Tryptic: 1 2 MaxMissed: 3 DelRows:

Xcorr: dCn: RSp: InclAA: MarkAA: XPRESS:

724	/sergei digest A full 01 1813 1815.2	1945.2	(-0.5)	3.2763	0.263	559.9	1	16/32	SP1P02754 LACB BOVIN	R. VYVEELKPTPEGDLEIL L
969	/sergei digest A full 01 2359 2361.3	2854.3	(-1.2)	3.2389	0.312	447.1	1	25/36	SP1P00321 CAM2 BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
328	/sergei digest A full 01 1063 1065.2	1582.8	(-0.0)	3.2171	0.384	1508.4	1	18/24	K. TARELMLVWNTK Y	
340	/sergei digest A full 01 1031 1033.2	974.1	(+0.2)	3.2106	0.316	454.9	1	14/16	SP1P00321 CAM2 BOVIN	K. VLDALDSIK T
786	/sergei digest A full 01 1949 1949.2	2215.5	(-0.2)	3.2068	0.435	531.2	1	21/36	SP1P02754 LACB BOVIN	V. VYVEELKPTPEGDLEIL L
1057	/sergei digest A full 01 2573 2575.3	4013.3	(-2.2)	3.1870	0.172	179.7	161	24/135	SW. KZ05 HUMAN	
1278	/sergei digest A full 01 3508 3505.3	2709.1	(+1.5)	3.1834	0.198	686.4	1	27/100	K. VAGTYVSLAWASDISLLDAQSAFLR V	
523	/sergei digest A full 01 1417 1425.2	1610.9	(-0.3)	3.1601	0.388	1389.3	1	20/26	SP1P0489 PMS2 RABIT	V. VYVEELKPTPEGDLEIL L
60	/sergei digest A full 01 0579 0581.2	1246.3	(+0.2)	3.1600	0.383	1196.1	1	18/20	SP1P02754 LACB BOVIN	R. TPEVDEALEK F
612	/sergei digest A full 01 1587 1589.3	2704.6	(+0.3)	3.1594	0.323	1098.0	1	28/80	K. IMKGEADAMSLDGGVLYIAC C	
1360	/sergei digest A full 01 3977 3985.2	3966.5	(+0.2)	3.1511	0.111	111.5	53	14/50	SP1P02754 LACB BOVIN	K. VAGTYVSLAWASDISLLDAQSAFLR V
322	/sergei digest A full 01 2247 2253.3	3966.5	(+0.2)	3.1511	0.111	111.5	53	21/135	MI0859	
900	/sergei digest A full 01 1017 1021.2	1640.9	(-0.4)	3.0938	0.206	371.4	2	18/28	SP1P02769 ALEU BOVIN	R. LUTFALLDIFGPGVDPYTGAPPLKRAIC
1026	/sergei digest A full 01 2503 2509.3	2854.3	(+0.8)	3.0938	0.206	371.4	2	23/36	SP1P00321 CAM2 BOVIN	R. KVP0USTPTLVEVSR S
846	/sergei digest A full 01 2079 2079.2	1551.8	(-0.4)	3.0860	0.357	573.8	1	18/26	SP1P02944 TRFE BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
753	/sergei digest A full 01 1853 1871.3	2031.3	(+1.6)	3.0715	0.427	1025.6	1	27/75	SP1P02754 LACB BOVIN	R. TAGNMLPGLLYSK I
576	/sergei digest A full 01 1513 1515.3	1679.9	(+0.1)	3.0691	0.289	468.0	1	21/52	SP1P0489 PMS2 RABIT	V. SLAWASDISLLDAQSAFLR V
69	/sergei digest A full 01 0503 0505.3	1250.4	(+0.4)	3.0525	0.115	575.4	1	23/36	SP1P02769 ALEU BOVIN	R. GBEVTSDDIDLER L
1022	/sergei digest A full 01 2435 2439.3	2241.5	(-0.3)	3.0417	0.022	486.2	23	22/75	MI0053	R. TPEVDEALEK F
635	/sergei digest A full 01 1759 1761.2	1890.1	(-0.2)	3.0407	0.420	462.0	1	17/28	SP1P02769 ALEU BOVIN	K. VYVEELKPTPEGDLEIL L
215	/sergei digest A full 01 0873 0875.2	1605.8	(-0.2)	3.0358	0.442	555.7	1	15/26	SP1P02944 TRFE BOVIN	R. HPTFYVAPLLYANK Y
1385	/sergei digest A full 01 4033 4101.2	2709.1	(+0.6)	3.0143	0.212	536.9	1	17/50	SP1P02754 LACB BOVIN	K. DNP0THYVAVVUK K
1106	/sergei digest A full 01 0583 0635.2	1194.4	(+0.1)	2.9868	0.273	930.3	1	15/18	SP1P02754 LACB BOVIN	K. VAGTYVSLAWASDISLLDAQSAFLR V
642	/sergei digest A full 01 1649 1649.3	2913.4	(-0.3)	2.9659	0.180	723.7	1	30/100	SP1P02656 CAB BOVIN	K. VYVEELKPTPEGDLEIL L
323	/sergei digest A full 01 1051 1053.3	1636.8	(+0.6)	2.9546	0.355	974.5	1	27/52	SP1P02754 LACB BOVIN	W. MHOPHOPLEPTVMFPQSVLSLSQSK V
795	/sergei digest A full 01 1965 1967.3	2336.6	(+2.4)	2.9429	0.056	362.6	12	20/80	SP1P02754 LACB BOVIN	R. TPEVDEALEK F
1088	/sergei digest A full 01 2647 2647.3	3336.8	(+1.7)	2.9239	0.096	130.0	70	23/128	MI1136	F. TAGNMLPGLLYSK I
132	/sergei digest A full 01 0723 0723.3	1342.6	(+0.3)	2.9069	0.228	624.1	1	19/40	SP1P00634 PPE BOVLI	L. AOLSGGVKACQVHAGRGKAGGVLYOVU
645	/sergei digest A full 01 1653 1659.3	1754.9	(-0.8)	2.8986	0.428	1190.3	1	17/26	SP1P46406 GSP RABIT	R. RTPEVDEALEK F
765	/sergei digest A full 01 1833 1837.3	2333.8	(+2.8)	2.8552	0.003	644.3	1	24/75	MI0457	K. ILISVYDDEAFIR F
310	/sergei digest A full 01 1031 1033.2	1066.2	(-0.1)	2.8501	0.212	74.5	1	14/15	SP1P02754 LACB BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
1018	/sergei digest A full 01 2487 2489.3	3336.8	(+2.2)	2.8437	0.210	246.5	2	22/108	SP1P00321 CAM2 BOVIN	P. TLNFHAEPEPELLMLAWPEAPLIK N
327	/sergei digest A full 01 1057 1051.3	1582.8	(+0.8)	2.8489	0.239	1076.8	1	31/48	SP1P00321 CAM2 BOVIN	R. TARELMLVWNTK Y
686	/sergei digest A full 01 1731 1733.3	2264.6	(-0.1)	2.8459	0.065	384.2	6	21/72	MI0134	P. TLNFHAEPEPELLMLAWPEAPLIK N
791	/sergei digest A full 01 1935 1937.3	3077.9	(+0.3)	2.8390	0.111	120.8	8	19/104	MI0635	R. TLNFHAEPEPELLMLAWPEAPLIK N
359	/sergei digest A full 01 1123 1125.3	1636.8	(-0.5)	2.8355	0.272	580.5	1	22/52	SP1P02754 LACB BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
145	/sergei digest A full 01 0755 0757.2	1194.4	(-0.3)	2.8348	0.242	931.9	1	15/18	SP1P02754 LACB BOVIN	K. VYVEELKPTPEGDLEIL L
721	/sergei digest A full 01 1805 1809.3	3491.9	(-1.9)	2.8180	0.072	163.1	8	20/120	MI1188	R. TLNFHAEPEPELLMLAWPEAPLIK N
875	/sergei digest A full 01 2147 2149.3	3518.3	(+1.0)	2.8131	0.089	211.1	12	22/135	MI1015	P. TLNFHAEPEPELLMLAWPEAPLIK N
427	/sergei digest A full 01 1243 1245.3	2909.3	(-1.9)	2.8109	0.093	204.5	125	23/100	MI1237	L. AOLSGGVKACQVHAGRGKAGGVLYOVU
770	/sergei digest A full 01 1905 1911.3	2351.9	(-0.2)	2.8101	0.197	269.3	132	19/38	MI1066	R. TLNFHAEPEPELLMLAWPEAPLIK N
926	/sergei digest A full 01 2233 2235.3	2872.2	(-2.0)	2.7800	0.077	524.0	1	25/112	MI0254	R. TLNFHAEPEPELLMLAWPEAPLIK N
753	/sergei digest A full 01 1879 1883.3	3470.9	(-2.4)	2.7784	0.011	112.1	11	23/124	SW. TRPV BOVIN	T. TLNFHAEPEPELLMLAWPEAPLIK N
1300	/sergei digest A full 01 3637 3643.3	3519.9	(-2.4)	2.7632	0.114	20.7	13	20/132	MI0906	R. TLNFHAEPEPELLMLAWPEAPLIK N
1242	/sergei digest A full 01 3131 3133.3	3847.3	(-2.6)	2.7632	0.032	18.7	202	19/140	MI0528	R. TLNFHAEPEPELLMLAWPEAPLIK N
605	/sergei digest A full 01 1579 1579.3	1441.6	(+0.7)	2.7514	0.481	954.8	1	17/22	SP1P0489 PMS2 RABIT	R. TLNFHAEPEPELLMLAWPEAPLIK N
749	/sergei digest A full 01 1327 1327.3	2327.7	(+0.7)	2.7508	0.176	490.9	2	19/38	SW. KZ05 HUMAN	R. TLNFHAEPEPELLMLAWPEAPLIK N
1361	/sergei digest A full 01 3879 3885.3	4052.8	(+0.2)	2.7469	0.118	355.3	1	25/152	MI0635	R. TLNFHAEPEPELLMLAWPEAPLIK N
1053	/sergei digest A full 01 2579 2587.2	2422.0	(-0.6)	2.7408	0.311	374.1	1	15/44	SP1P02666 CAB BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
279	/sergei digest A full 01 0977 0979.2	1246.3	(+0.2)	2.7400	0.195	529.9	1	15/20	SP1P02754 LACB BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
585	/sergei digest A full 01 1539 1541.2	2111.1	(-0.3)	2.7258	0.188	20.7	13	20/34	SP1P02754 LACB BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
172	/sergei digest A full 01 0739 0739.3	2111.1	(-0.3)	2.7258	0.188	20.7	13	20/34	SP1P02754 LACB BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
507	/sergei digest A full 01 1339 1339.2	2081.3	(+2.3)	2.7195	0.081	622.4	12	34/38	SP1P02666 CAB BOVIN	R. TLNFHAEPEPELLMLAWPEAPLIK N
134	/sergei digest A full 01 0819 0823.3	2202.7	(+2.3)	2.7189	0.004	416.9	144	18/72	MI0936	R. TLNFHAEPEPELLMLAWPEAPLIK N
953	/sergei digest A full 01 2323 2327.3	2422.0	(-2.8)	2.7153	0.054	275.9	45	21/104	MI0508	R. TLNFHAEPEPELLMLAWPEAPLIK N

correct

incorrect

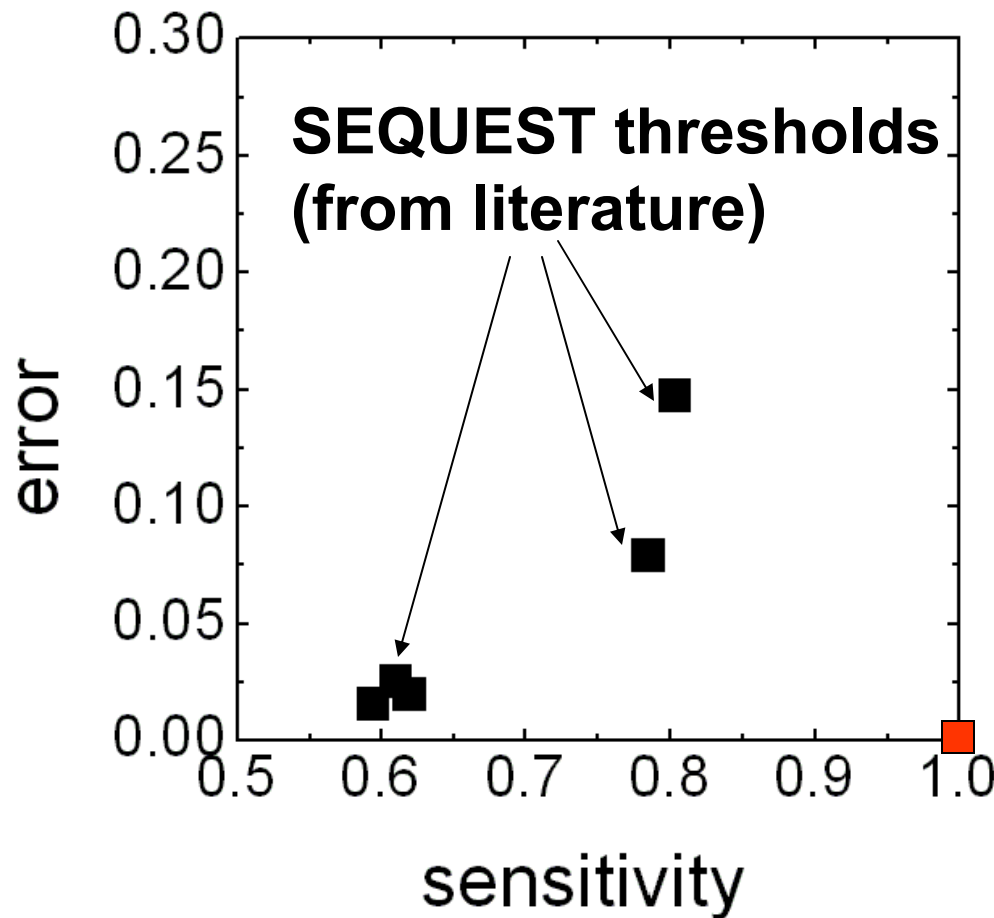
sort by search score

threshold

SEQUENT:
Xcorr > 2.0
 $\Delta C_n > 0.1$

MASCOT:
Ion Score > 30

Threshold Model: Bad Discrimination and Inconsistency



test data (18 proteins): OMICS 6(2), 207 (2002)

Sensitivity:
fraction of all
correct results
passing filter

Error Rate:
fraction of all
results passing
filter that are
incorrect

Ideal Spot

Difficulties in Interpreting Peptide Identifications based on MS/MS

Applies to both SEQUEST and Mascot (as it is used in practice) and, to large degree, to more recent tools

- No 'useful' measures of confidence

(Mascot: 'identity threshold' guideline is not practical and rarely used)

- Different criteria used to filter data
- Unknown and variable false positive error rates

Just as assignment of quality scores to each base in DNA sequencing was essential for the genome sequencing programs, statistical models for estimating the accuracy of peptide and protein identifications are crucial for the success of high throughput proteomics

Statistical Validation

- p-values or expectation values

used, e.g., in sequence similarity searching

- Probabilities (Bayes)

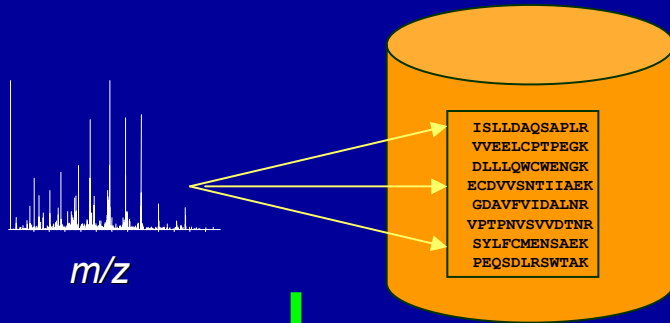
based on the ratio of two distributions (correct and incorrect) derived from the data (entire dataset)

used, e.g., in information retrieval (relevant vs. non-relevant documents)

Expectation Values (empirical model)

Spectrum

Database



$$p(s_m) = \sum_{s \geq s_m} P(s)$$

probability to get score $s \geq s_m$ by chance

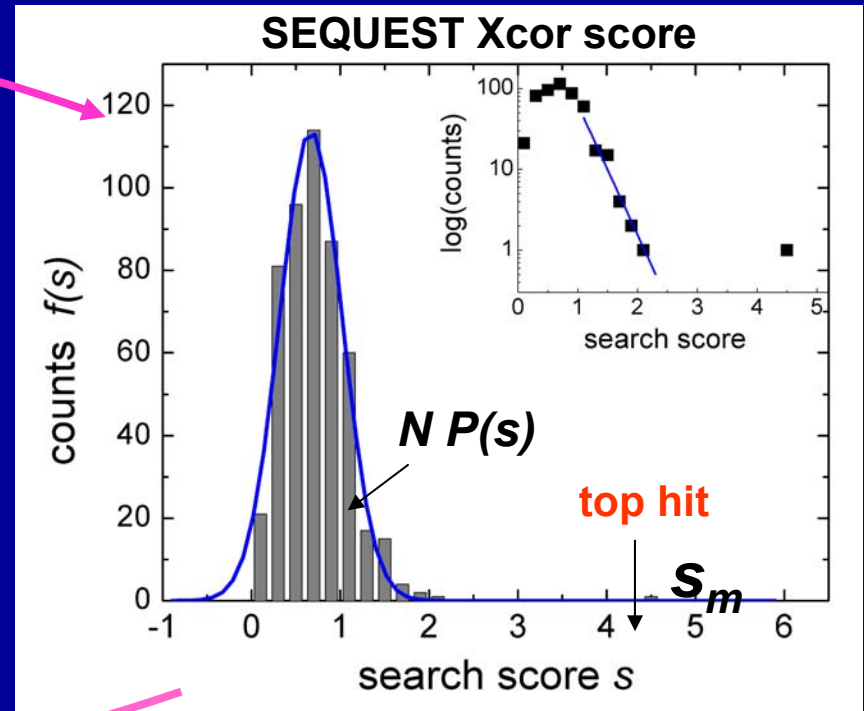
$$E(s_m) = N \sum_{s \geq s_m} P(s)$$

expected number of random matches with $s \geq s_m$

Rank	Peptide	Score
1	ISLLDAQSAPLR	4.5
2	VVEELCTPEGK	2.1
3	DLLQWCWENGK	2.0
4	ECDVVSNTIIAEK	1.9
5	GDAVFVIDALNR	1.7
6	VPTPNVSVVTNR	1.6
7	SYLFCMEAEK	1.6
8	PEQSDLRSWTAK	1.5
...	...	...

N peptides

p
10^{-5}
0.11
0.23
0.52
0.72
0.86
0.86
0.94
...

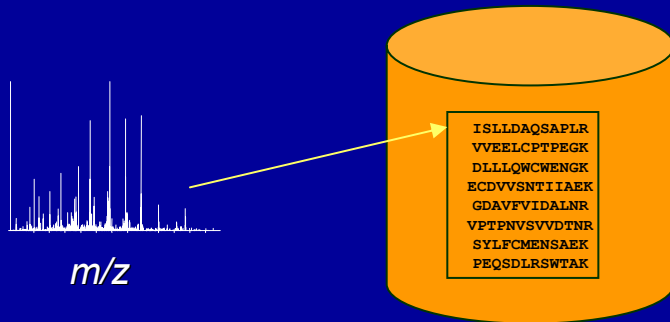


Fenyó & Beavis *Anal. Chem.* (2003)

Expectation Values (explicit model)

Spectrum

Database



Sadygov & Yates *Anal. Chem.* (2003)

Geer et al. *J. Proteome Res.* (2004)

$$P(s) = \frac{\mu^s}{s!} \exp(-\mu)$$

Poisson
distribution

μ : function of mass tolerance,
number experimental peaks
number of calculated ions
mass, charge

s : number of matched peaks

probability to get score
 $s \geq s_m$ by chance

$$p(s_m) = \sum_{s \geq s_m} P(s)$$

Geer et al.
(upper bound)

$$E(s_m) = N(1 - (1 - \sum_{s \geq s_m} P(s))^N) \approx N^2 \sum_{s \geq s_m} P(s)$$

expected number of
random matches
with $s \geq s_m$

$$E(s_m) = N \sum_{s \geq s_m} P(s)$$

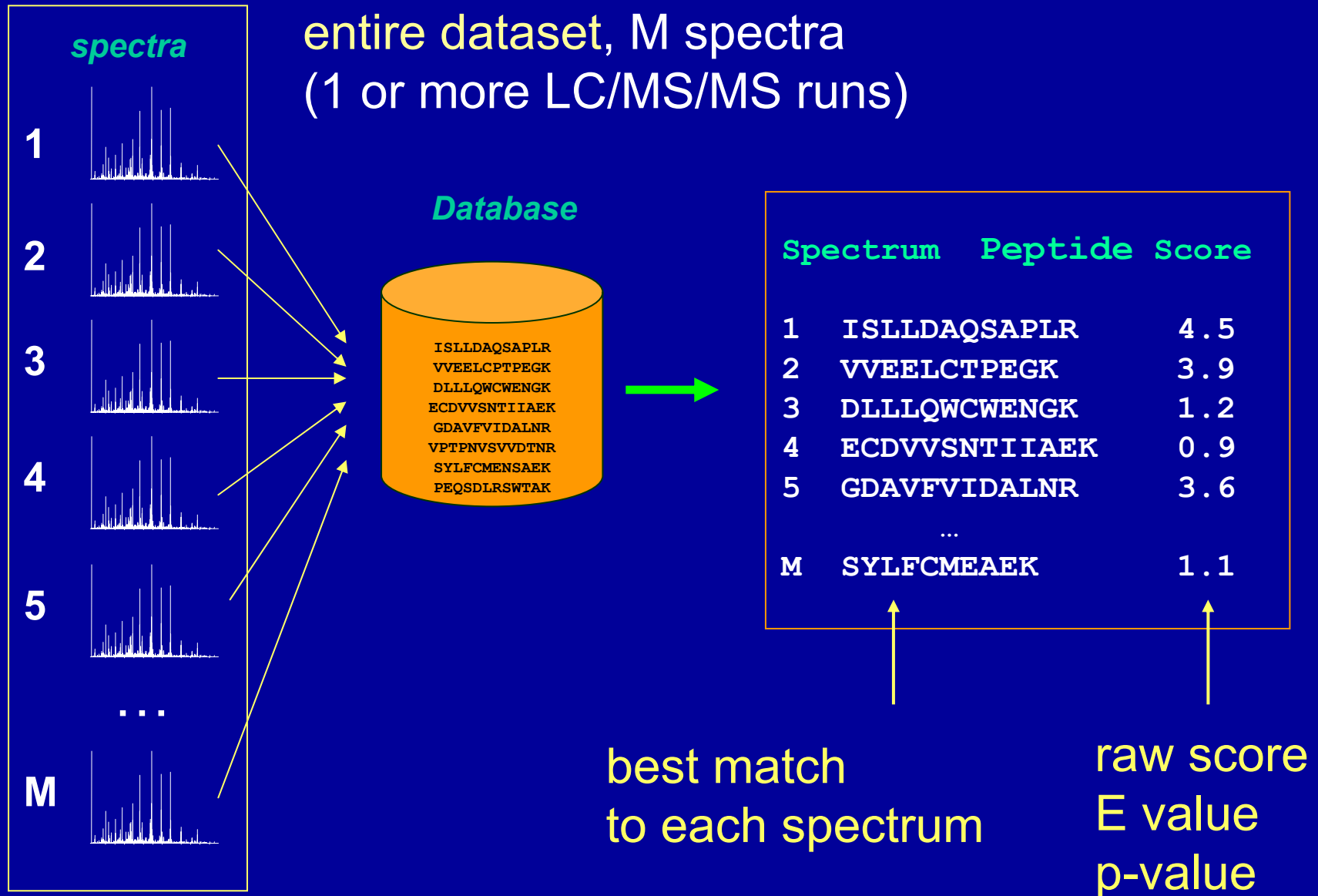
Sadygov et al.
(lower bound)

Expectation Values (or p-values): Limitations

1. P-values or E-values are not well suited for the analysis of large-scale datasets (do not allow estimation of error rates as a function of filtering threshold)

see, e.g., recent papers by Tsibshirani and others on the subject of p-values vs. False Discovery Rate (FDR) approach
2. Difficult to take advantage of other useful information (e.g., number of missed cleavages, peptide retention time)
3. Need to compute protein probabilities by combining probabilities of peptides corresponding to the same protein. Whether peptide expectation values can be used for that purpose is not clear

Modeling Large-Scale Datasets



Statistical Model for Computing Peptide Probabilities (PeptideProphet)

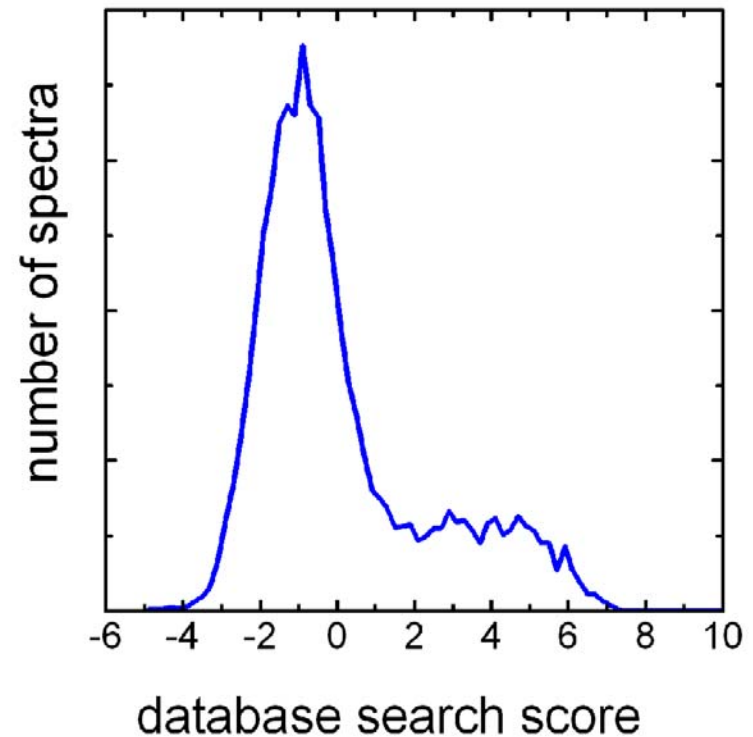
entire dataset:

Spectrum	Peptide	Score
1	ISLLDAQSAPLR	4.5
2	VVEELCTPEGK	3.9
3	DLLLQWCWENGK	1.2
4	ECDVVSNTIIAEK	0.9
5	GDAVFVIDALNR	3.6
...		
M	SYLFCMEAEK	1.1

spectrum

best match

score



A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold *Anal. Chem.* 74, 5383 (2002)

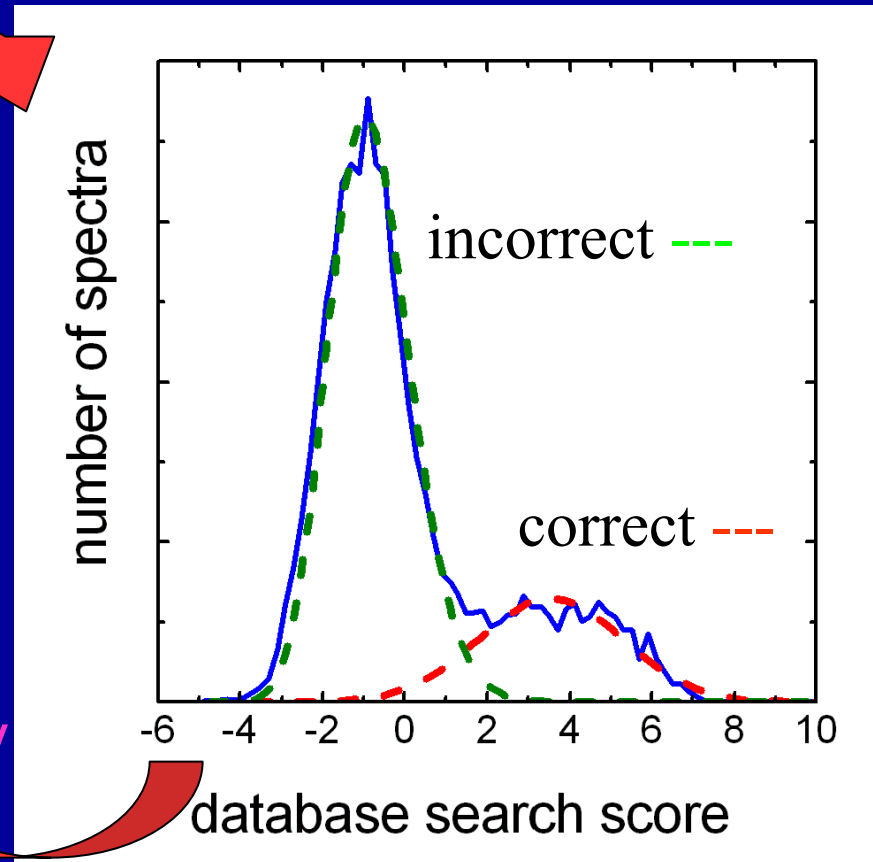
Statistical Model for Computing Peptide Probabilities (PeptideProphet)

entire dataset:

Spectrum	Peptide	Score	
1	ISLLDAQSAPLR	4.5	1.00
2	VVEELCTPEGK	3.9	0.99
3	DLLLQWCWENGK	1.2	0.11
4	ECDVVSNTIIAEK	0.9	0.00
5	GDAVFVIDALNR	3.6	0.87
...	...	...	...
M	SYLFCMEAEK	1.1	0.02

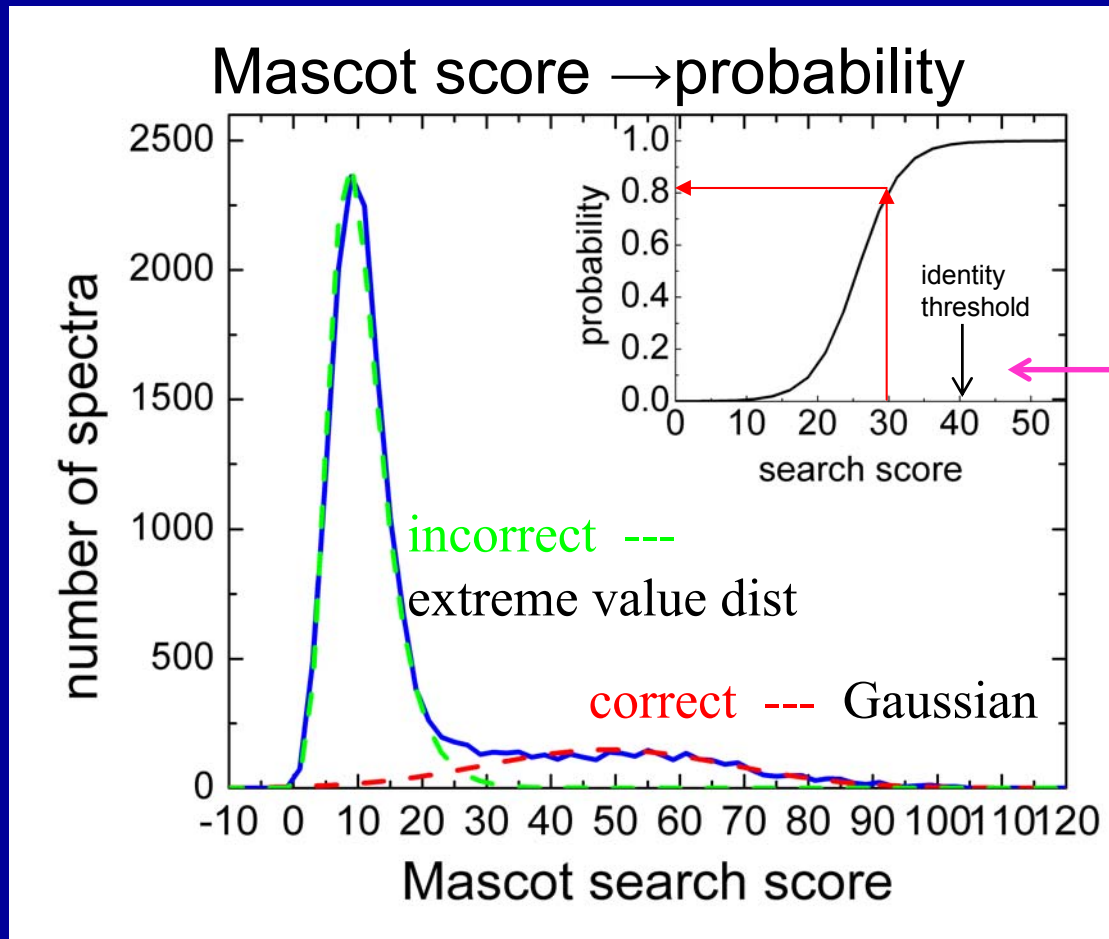
probability

'unsupervised clustering'



EM mixture model algorithm learns the most likely distributions among correct and incorrect peptide assignments given the observed data

Illustration: Assigning Probabilities to Mascot Search Results



distributions are learned from the data

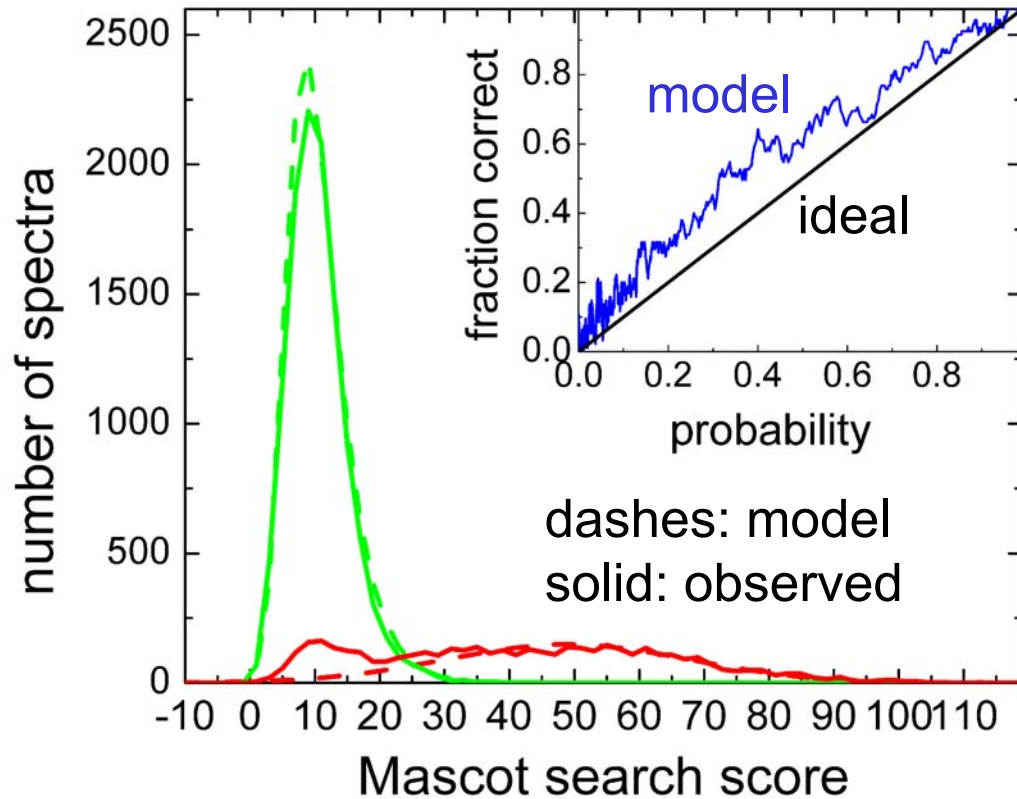
conversion of Mascot search scores into probabilities

To address a common misunderstanding:

distribution parameters ARE NOT determined using a control dataset of 18 proteins, or any other training dataset for that matter. They are learned from each analyzed dataset anew using the EM mixture model algorithm

H. Influenzae, membrane fraction, 15 LC/MS/MS runs (~30,000 spectra)

Accuracy of Learned Distributions and Computed Probabilities



database searched:

Human

H. Influenzae

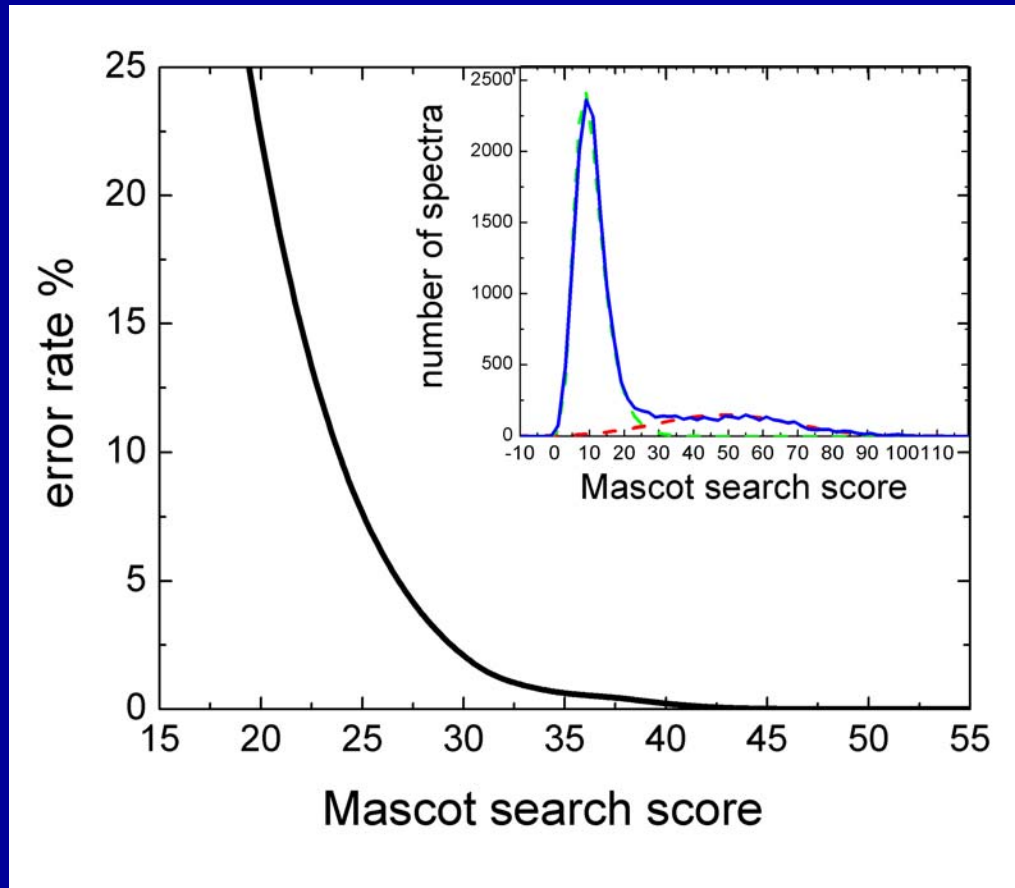
size ratio: ~ 20:1

For those familiar with “reverse database search” approach:
This is an equivalent of appending 20 randomized databases of equal size.

Method is accurate

H. Influenzae, membrane fraction, 15 LC/MS/MS runs
~30,000 spectra

Estimation of False Positive Error Rates



Computed probabilities allow to estimate sensitivity and false positive error rate, e.g.

$$\langle Err \rangle = \frac{\sum_{\{i|p_i \geq p_T\}} (1 - p_i)}{\sum_{\{i|p_i \geq p_T\}} 1}$$

Use model predictions to achieve desired balance between sensitivity and the error rates

no need to do “reverse database searching”
to estimate the error rates

Information Useful for Validation

database search scores

+

peptide sequence properties:

- number of tryptic termini
- mass accuracy
- number of missed cleavages
- presence of a specific a.a. (C if ICAT)
- presence of a specific sequence motif (NxS/T)

peptide separation coordinates:

- elution time (reverse-phase LC)
- pI value (isoelectric focusing)

PeptideProphet statistical model can use all
this information → method is more discriminative

PeptideProphet Software

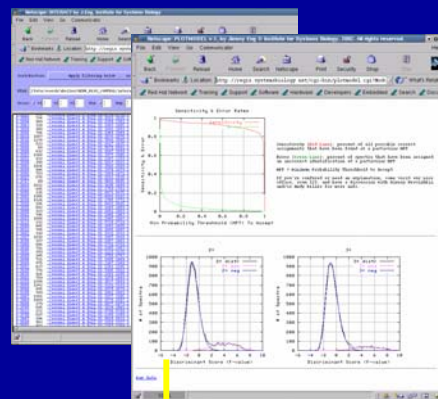


LCQ TOF-TOF Q-TOF
SEQUEST MASCOT

hard to compare
datasets, unknown
error rates

Also: different
sample preparation
sample complexity
quality of spectra

easy comparison
quality control,
known error rates



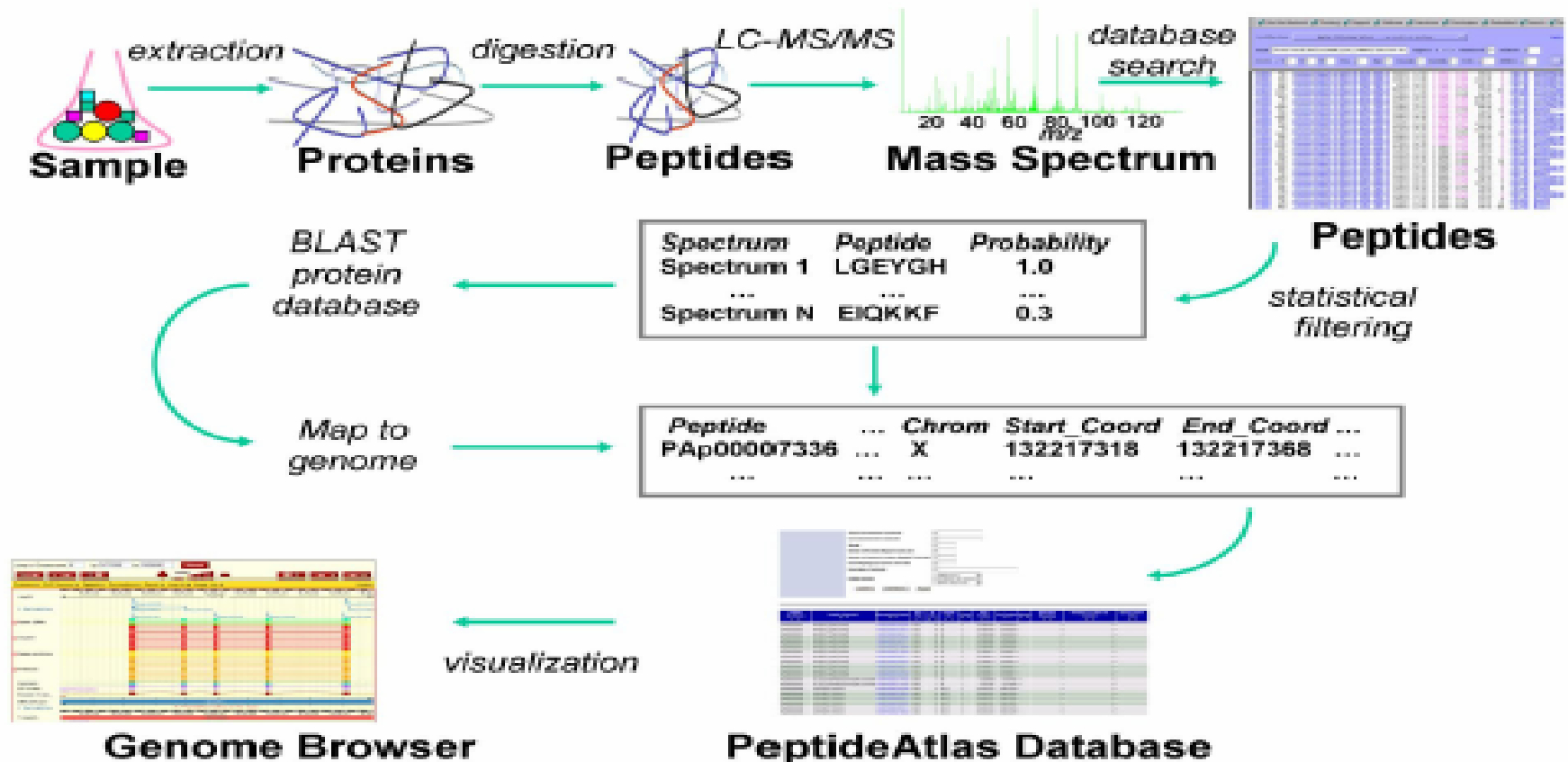
PeptideProphet

**peptide identifications
with accurate probabilities**

download from www.proteomecenter.org

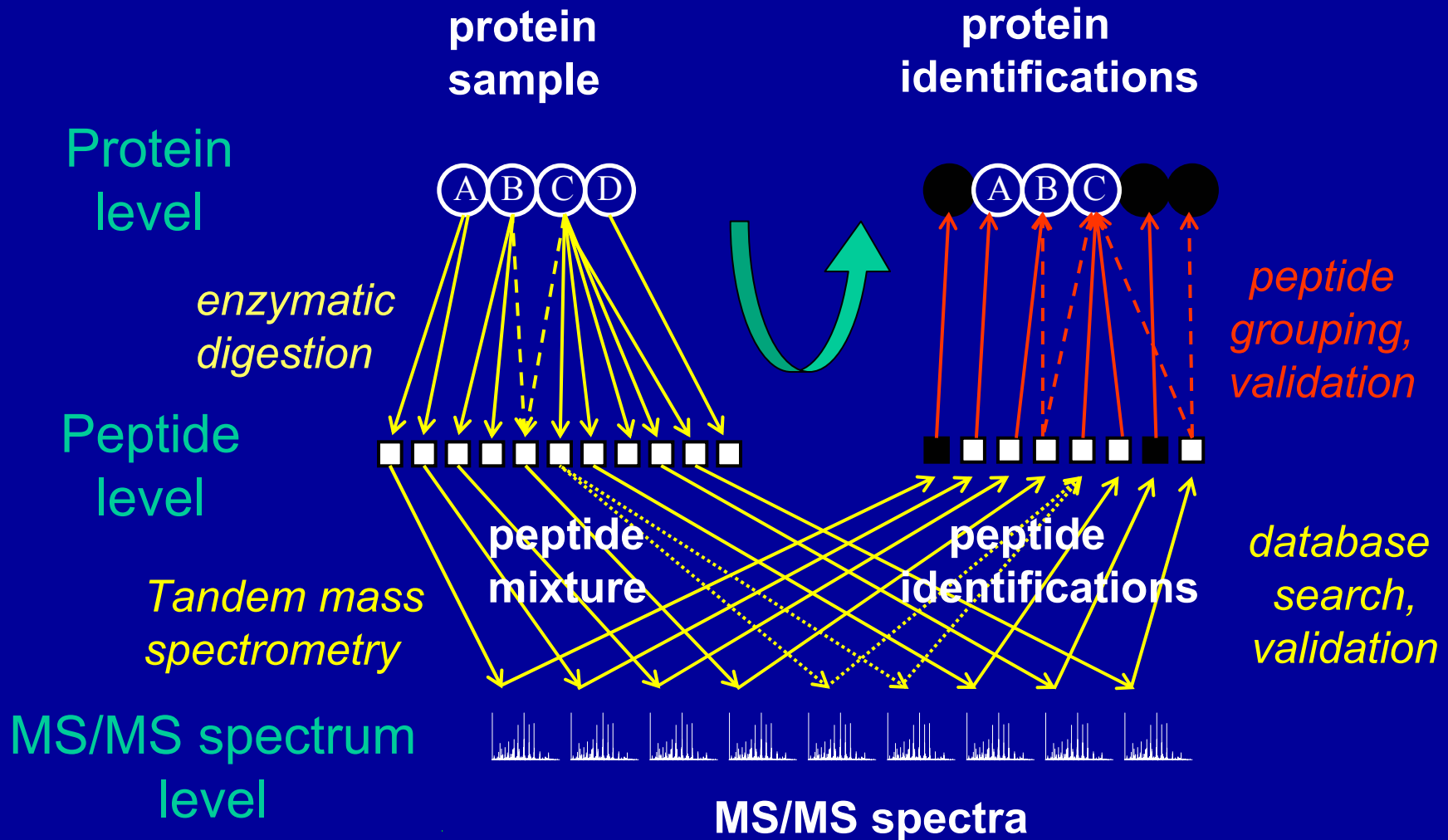
PeptideAtlas: High Throughput Pipeline

From Peptides to Genome Annotation



Protein Identification

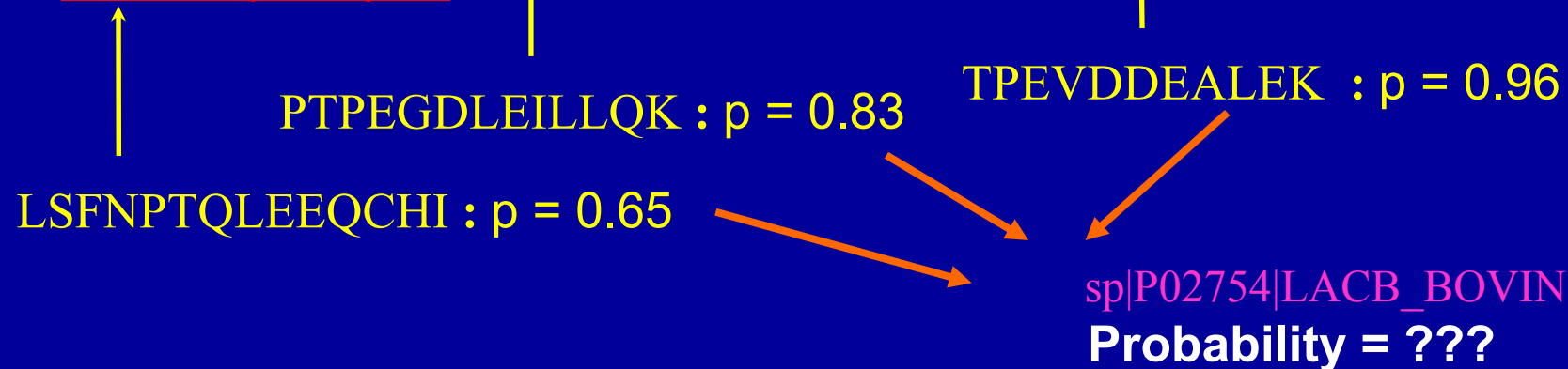
Shotgun Protein Identification



Protein Identification

>sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG)
(ALLERGEN BOS D 5) - Bos taurus (Bovine).

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQKWENGECQAQKKIIAEKTKIPAVFKIDALNENKL
VLDTDYKKYLLFCMENS~~AEPEQSLACQCLVR~~TPEVDDEALEKFDKALKALPMH
IRLSFNPTQLEEQCHI



ProteinProphet combines probabilities of peptides assigned to MS/MS spectra to compute *accurate* probabilities that corresponding proteins are present?

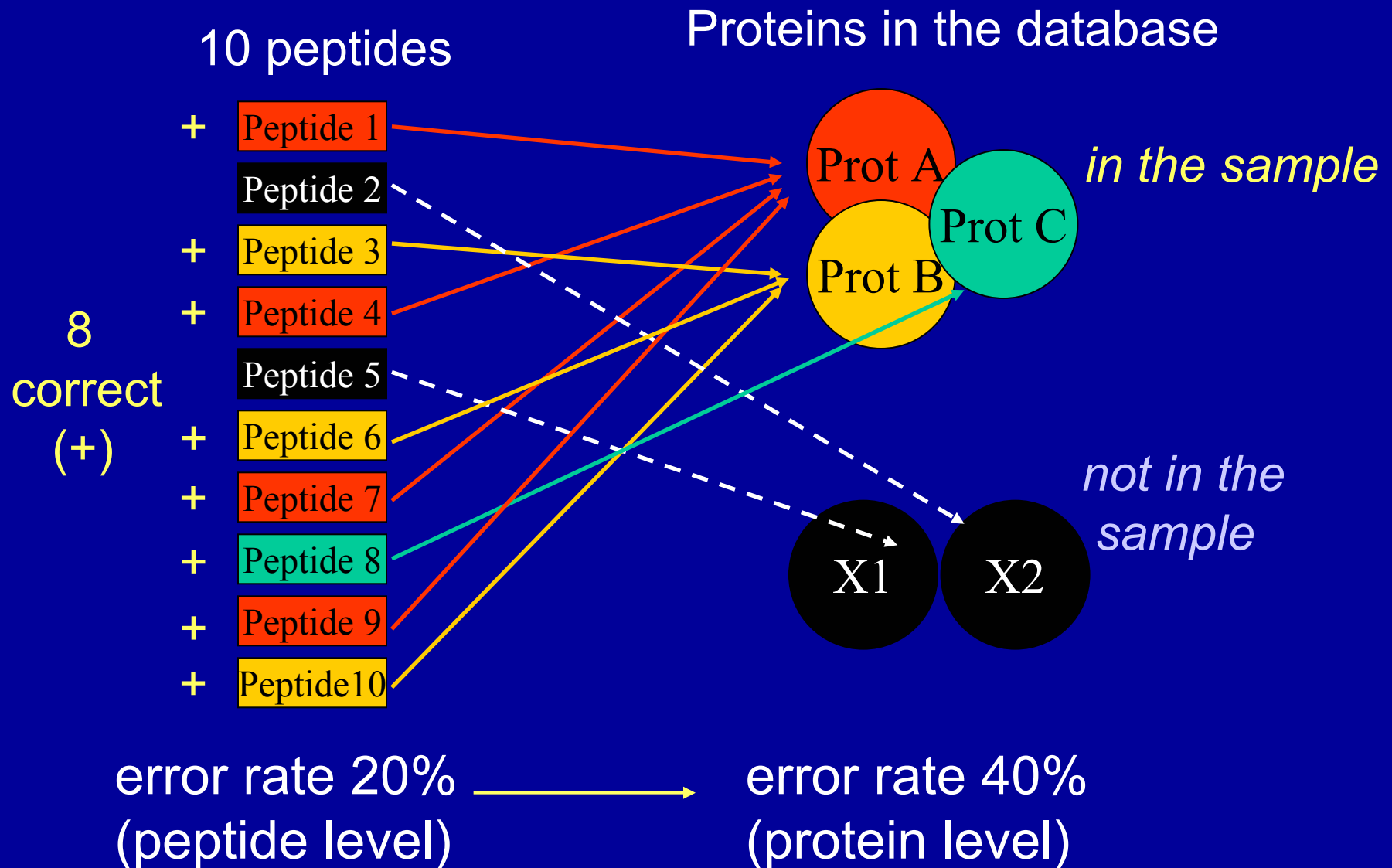
Issues for Protein Identification

- Peptides corresponding to ‘single-hit’ proteins are less likely to be correct than those corresponding to ‘multi-hit’ proteins
- Many peptides are present in more than a single protein (isoforms, homologous proteins etc.)

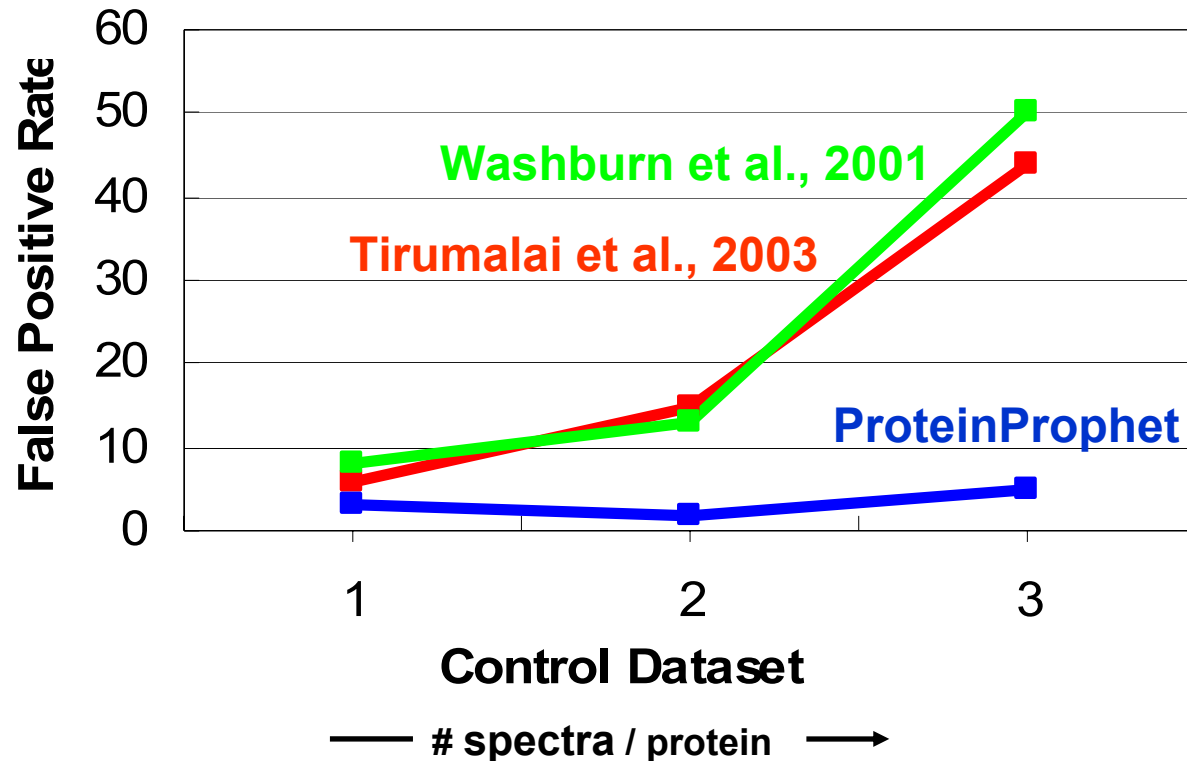
A. I. Nesvizhskii *et al.*, *Anal. Chem.* **75**, 4646-4658 (2003)

A. I. Nesvizhskii & R. Aebersold *Drug Discov Today* **9**,
173-81 (2004)

Amplification of False Positive Error Rates from Peptide to Protein Level



Protein ID False Positive Rates: ProteinProphet vs. Score Thresholds



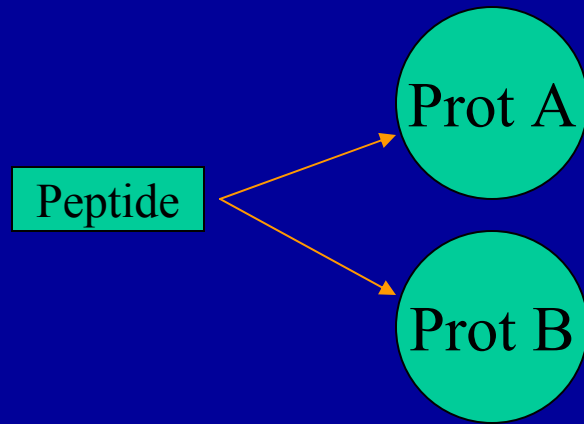
Keep acquiring
data and you will
“identify”
everything in the
database!

Control Datasets:

- 1 Halobacterium vs. Halo+Human (4 runs)
- 2 Halobacterium vs. Halo+Human (45 runs)
- 3 18 purified proteins vs. 18+Human (22 runs)

Protein Inference Problem

Degenerate peptides: correspond to more than a single entry in protein database



protein A or protein B ??
Or both?

In shotgun proteomics the connectivity between peptides and proteins is lost

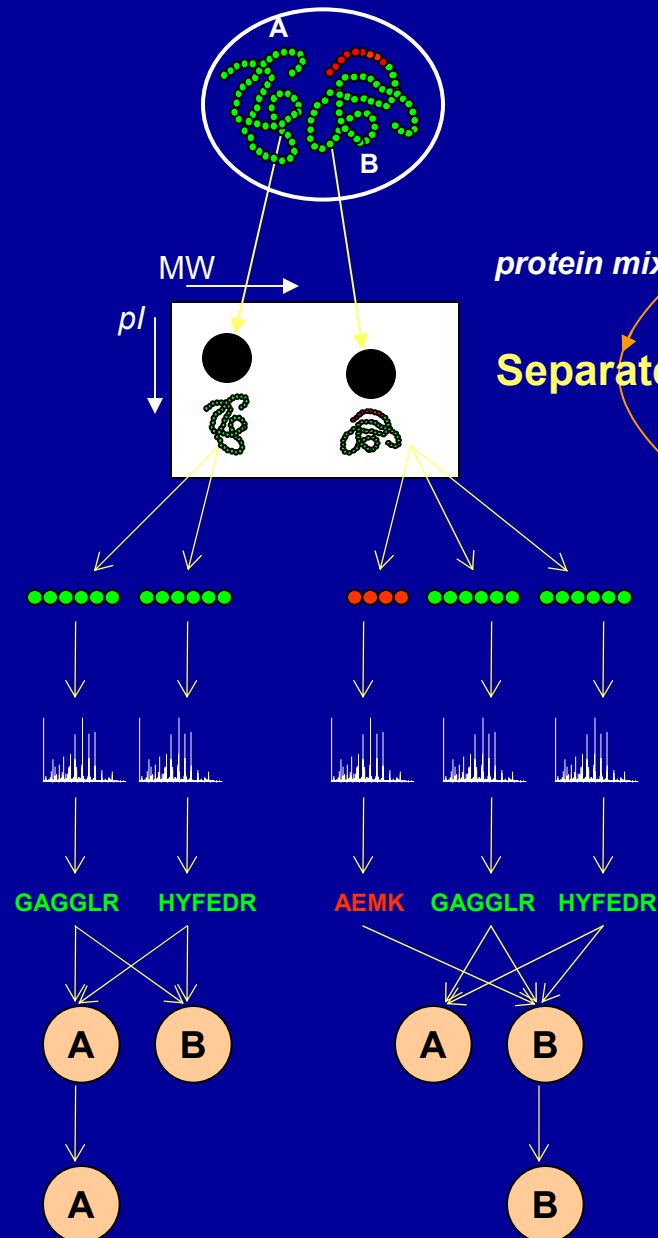
Degenerate peptides are more prevalent with databases of higher eukaryotes due to the presence of:

- related protein family members
- alternative splice forms
- partial sequences

2D Gel-based Approach

vs.

Shotgun Approach



Protein mixture

protein mixture separation

Separated proteins

protein digestion

Peptide mixture

peptide mixture separation
MS/MS sequencing

MS/MS spectra

peptide identification

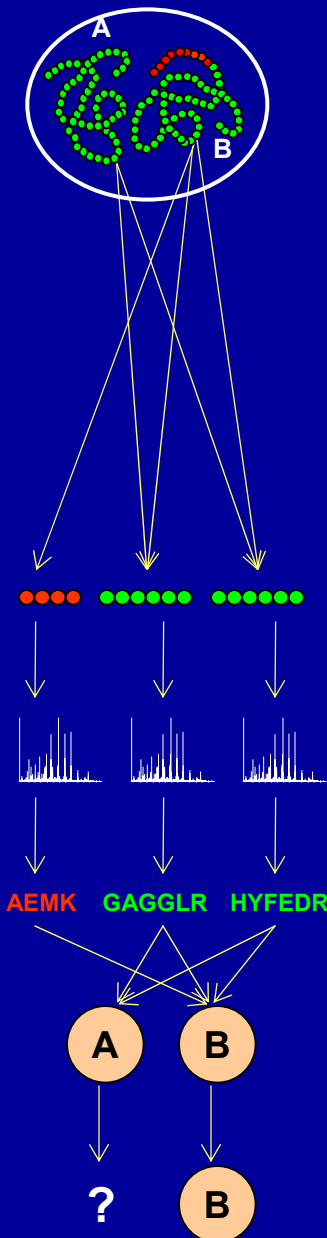
Peptide sequences

peptide grouping

Implicated database proteins

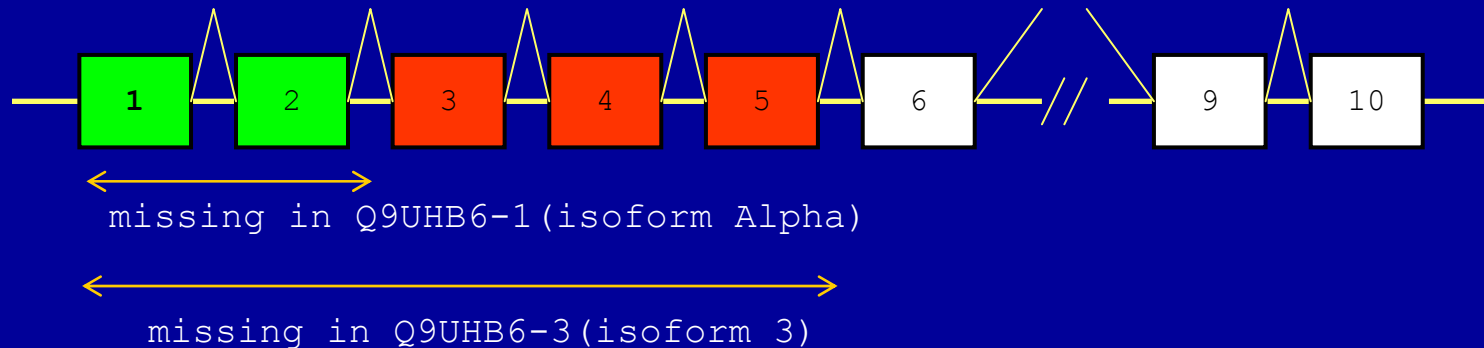
protein inference

Identified proteins



Isoforms

Gene: **EPLIN** location: **12q13** Q9UHB6-1 (isoform Beta)



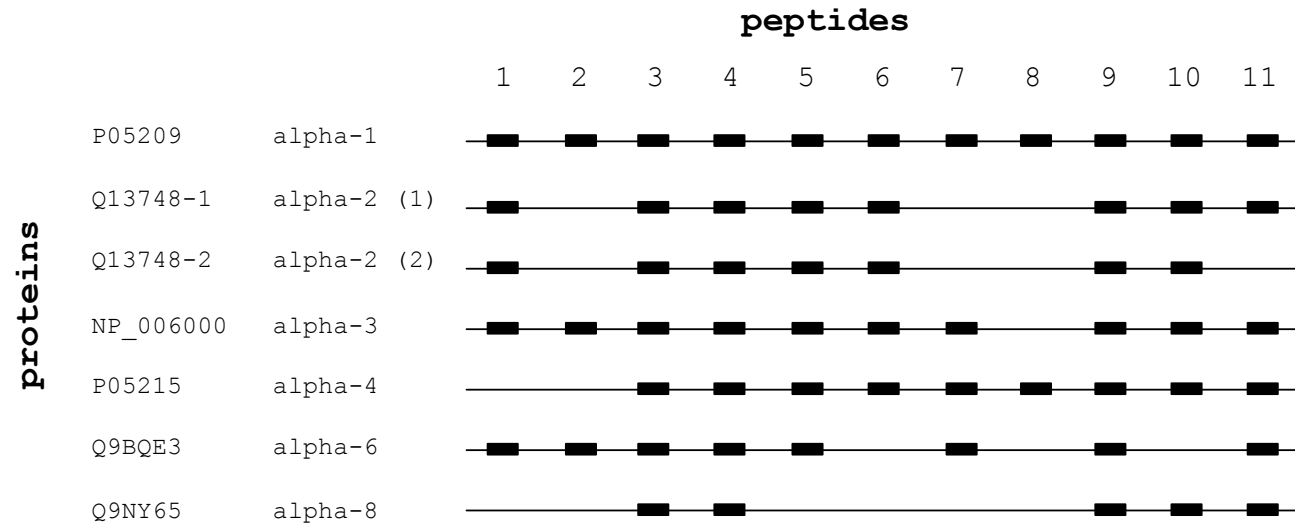
>IPI:IPI00220465 Splice isoform Alpha of Q9UHB6 Epithelial protein lost in neoplasm

MESSPFNRRQWTSLSLRVTAKELSLVNKNKSSAIVEIFSKYQKAAEETNMEKKRSNTENLSQHFRKGTTLTVLKKKWENP
GLGAESHTDSLNRNSSTEIRHRADHPPAEVTSHAASGAKADQEEQIHPRSRLRSPPEALVQGRYPHIKDGEDLKDHSTES
KKMENCLGESRHEVEKSEISENTDASGKIEKYNVPLNRLKMMFEKGEPTQTKILRAQSRASGRKISENSYSLDDLEIG
PGQLSSSTFDSEKNESRRNLELPRLSETSIKDRMAKYQAAVSKQSSSTNYTNELKASGGEIKIHKMEQKENVPPGPPEVC
ITHQEGEKISANENSLAVRSTPAEDDSRDSQVKSEVQQPVHPKPLSPDSRASSLSESSPPKAMKKFQAPARETCVECQK
TVYPMERLLANQQVFHISCFRCSYCNKLSLGTYASLHGRIYCKPHFNQLFKSKGNYDEGFGHRPHKDLWASKNENEEI
LERPAQLANARETPHSPGVEDAPIAKVGVLAASMEAKASSQOEKEDKPAETKKLRIAWPPPTELGSSSGSALEEGIKMSK
PKWPPDEDEISKPEVPEDVDLDLKKLRRSSSLKERSRPFTVAASFQSTSVKSPKTVSPPIRKGWSMSEQSEESVGGRAVE
RKQVENAKASKKNGNVGKTTWQNKESKGETGKRSKEGHSLEMENENLVENGADSDDDNSFLKQQSPQEPKSLNWSSFVD
NTFAEEFTTQNQKSQDVELWEGEVVKELSVVEEQIKRNRYYDEDEDEE

Gene families (paralogues)

C Tubulin-alpha protein family

1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISA EK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		



>IPI00387144 IPI:IPI00387144.2|SWISS-PROT:P05209 Tubulin alpha-1 chain

MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSPDK**TIGGGDDSFNTFFSETGAGK**HVPRAVFVDLEPTVIDEVRTGTYR**QLFHPEQ**
LITGKEDAANNYARGHYTIGKEIIDLVLDRIKRLADQCTGLQGFLVFHSFGGGTSGSFTSLLMERLSVDYGGKSKLEFSIYPAPQVSTAVV
 EPYNSILTHTTTLEHSDCAFMDNEAIYDICRR**NLDIERPTYTNLNR**LISQIVSSITASLRFDGALNVDLTFEQTNLVPYPRI**IHFPLATYA**
PVISA EKAYHEQLSVAEITNACFEPANQMVKCDPRHGK**YMACCLLYR**GDVVPKDVNAAIATIKTKR**SIQFVDWCPTGFK**VGINYQPPTVVP
 GGD**LAKVQRAVCMLSNTTAIAEAWARLDHKFDLMYAK**RAVFVHWYVGEGMEEGEFSEAREDMAALEKDYE EVGVDSVEGEGEGEEGEEY

Mature Protein Forms

A

>IPI00019906.1|UniProt/Swiss-Prot:P35613 Basigin precursor

signal peptide

MAAALFVLLGFALLGTHGASG**AAGTVFTTVEDLGSK**ILLTCSLNDSEVTGHR**WLKGGVVLKEDALPGQK**TEFKVDSDDQWGEYSCVFL
PEPMGTANIQLHGPPRVKAVK**SSEHINEGETAMLVCK**SESVPPVTDWAWYKITDSEDKALMNGSESRRFFVSSSQGR**SELHIENLNMEADP**
GQYRCNGTSSKGSDQAIITLRVRSHLAALWPFLGIVAEVLVLVTIIFIYEKR**RKPEDVLDDDDAGSAPLKSSGQHQNDKGKNVRQRN**SS

signal peptide removed

Protein Inference using ProteinProphet

- Peptides corresponding to ‘single-hit’ proteins are less likely to be correct than those corresponding to ‘multi-hit’ proteins

ProteinProphet learn from the data itself by how much peptide probabilities should be adjusted to reflect this protein grouping information

- Many peptides are present in more than a single protein (isoforms, homologous proteins etc.)

ProteinProphet resolves ambiguous cases (when possible) and presents the list of protein identifications in a convenient “biologist-friendly” format

A. I. Nesvizhskii et al., *Anal. Chem.* **75**, 4646-4658 (2003)

Publishing Large-Scale Datasets: Example

Technology

The Application of New Software Tools to Quantitative Protein Profiling Via Isotope-coded Affinity Tag (ICAT) and Tandem Mass Spectrometry

II. EVALUATION OF TANDEM MASS SPECTROMETRY METHODOLOGIES FOR LARGE-SCALE PROTEIN ANALYSIS, AND THE APPLICATION OF STATISTICAL TOOLS FOR DATA ANALYSIS AND INTERPRETATION*

Priska D. von Haller‡, Eugene Yi, Samuel Donohoe, Kelly Vaughn, Andrew Keller, Alexey I. Nesvizhskii, Jimmy Eng, Xiao-jun Li, David R. Goodlett, Ruedi Aebersold, and Julian D. Watts§

Proteomic approaches to biological research that will prove the most useful and productive require robust, sensitive, and reproducible technologies for both the qualitative and quantitative analysis of complex protein mixtures. Here we applied the isotope-coded affinity tag (ICAT) approach to quantitative protein profiling, in this case proteins that copurified with lipid raft plasma membrane domains isolated from control and stimulated Jurkat human T cells. With the ICAT approach, cysteine residues of the two related protein isolates were covalently labeled with isotopically normal and heavy versions of the same reagent, respectively. Following proteolytic cleavage of combined labeled proteins, peptides were fractionated by multidimensional chromatography and subsequently analyzed via automated tandem mass spectrometry. Individual tandem mass spectrometry spectra were searched

against a human sequence database, and a variety of

ing the rapid evaluation of large proteomic datasets possible. Finally, by repeating the experiment, information relating to the general reproducibility and validity of this approach to large-scale proteomic analyses was also obtained. *Molecular & Cellular Proteomics* 2:428–442, 2003.

A main objective of proteomics research is the systematic identification and quantification of the proteins expressed in a cell, or contained within a cell compartment or other protein complex. The common approach to quantitative protein analysis to date has been the combination of protein separation, most commonly high-resolution two-dimensional polyacrylamide gel electrophoresis (2DE)¹ and tandem mass spectrometry (MS/MS). For this approach, protein identification is accomplished by individual spot excision, in-gel digestion

P. Von Haller *et al.* *Molecular & Cellular Proteomics* (2003)

Conclusions

- consistent and transparent analysis;
reduce if not eliminate need for manual validation

need to make tools available (and published)!

- known error rates → comparison between different experiments becomes possible
- facilitates publishing large-scale datasets
 - publish entire (minimally filtered) datasets of protein and peptide identifications together with statistically computed confidence measures
 - provide access to raw data (e.g., MS/MS spectra)

Publishing Large-scale Datasets

Editorial

The Need for Guidelines in Publication of Peptide and Protein Identification Data

WORKING GROUP ON PUBLICATION GUIDELINES FOR PEPTIDE AND PROTEIN IDENTIFICATION DATA*

Steven Carr^{‡§}, Ruedi Aebersold[¶], Michael Baldwin^{||}, Al Burlingame^{||}, Karl Clauser^{**},
and Alexey Nesvizhskii[¶]

Over the past few years, the number and size of proteomic datasets composed of mass spectrometry-derived protein identifications reported in the literature have grown dramatically. This is a direct result of the widespread availability of instruments, methods, and easy-to-use software for collecting large amounts of data and for converting the observed peptide and fragment-ion masses to peptide and then protein identities. In particular, the analysis of samples containing large numbers of proteins by multidimensional liquid chromatography (LC/LC)¹ coupled on-line with tandem mass spectrometry (MS/MS) is now a common component of many biological projects. Clearly it is in the interest of the scientific community to make such data readily available. However, the

the most probable peptide assignment to the top of the "hit" list. In addition, new filtering criteria are being developed that, when layered onto the results from the above algorithms, help to eliminate a certain additional percentage of false positives (3, 4). It is very important that the users of these tools, our authors, have at least a working understanding of how the algorithm they use works. However, even the judicious use of scoring, threshold parameters, and additional filtering criteria for search engines, while serving the very important purpose of reducing the number of misassigned peptides and proteins, does not eliminate the problem. It is almost always possible to match a MS/MS spectrum to a peptide in the database; the difficult part is validating that the match is

Molecular & Cellular Proteomics 3, 532-534 (2004)

Acknowledgements

Ruedi Aebersold

Andrew Keller Jimmy Eng

David Goodlett Eugene Yi Sam Donahue

Julian Watts Jeff Ranish

Dan Martin Victor Ng

Serg Stolyar E. Kolker

Funding: NHLBI Proteomics Center at ISB

www.proteomecenter.org