ProteinProphet: Statistical Validation of Protein Identifications

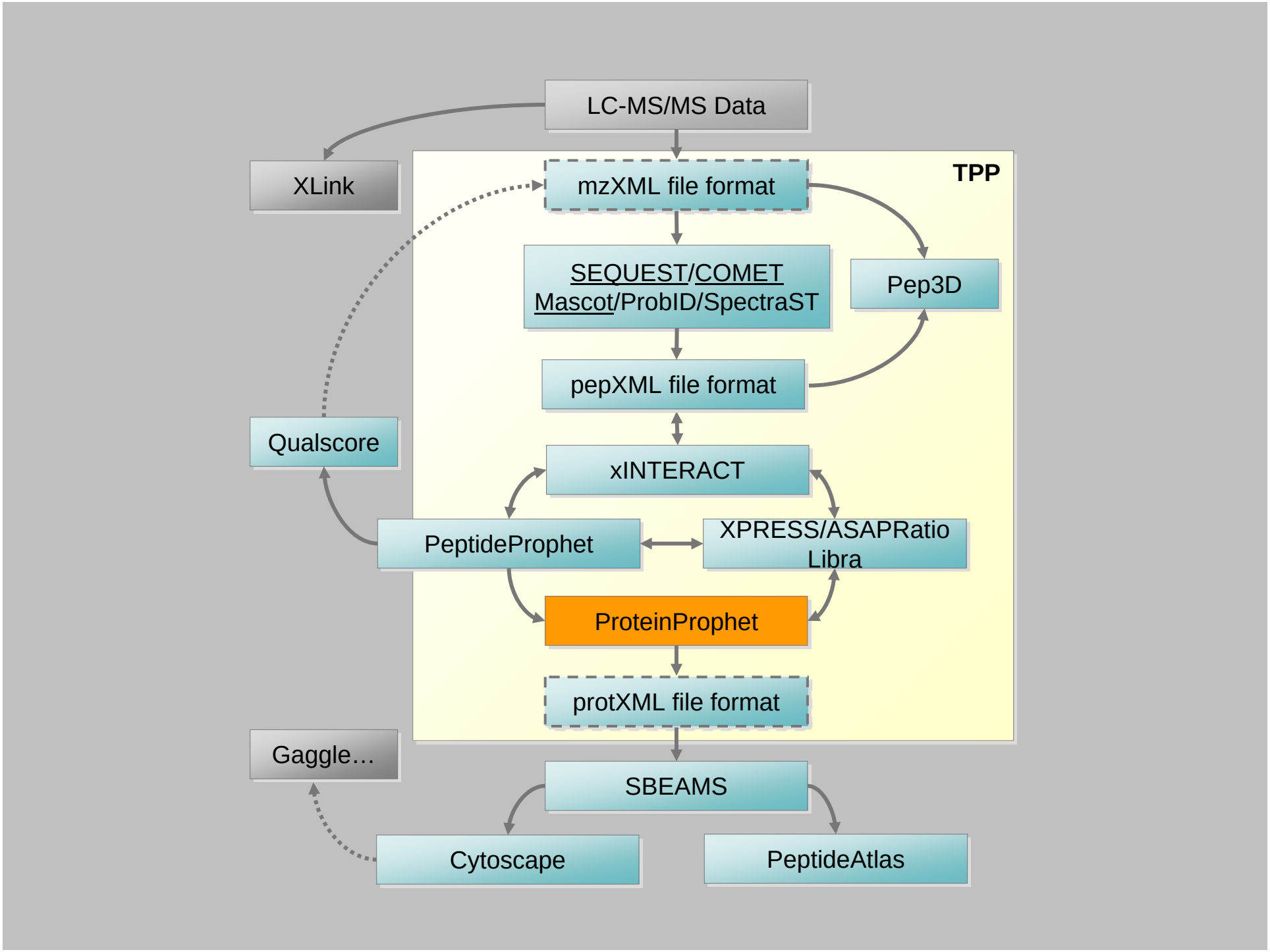
Alexey Nesvizhskii

Day 3

October 18, 2006

Alex Nesvizhskii
University of Michigan

INSTITUTE FOR Systems Biology


LC-MS/MS Data
XLink
TPP
mzXML file format
SEQUEST/COMET
Mascot/ProbID/SpectraST
Pep3D
pepXML file format
Qualscore
xINTERACT
PeptideProphet
XPRESS/ASAPRatio
Libra
ProteinProphet
protXML file format
Gaggle…
SBEAMS
Cytoscape
PeptideAtlas

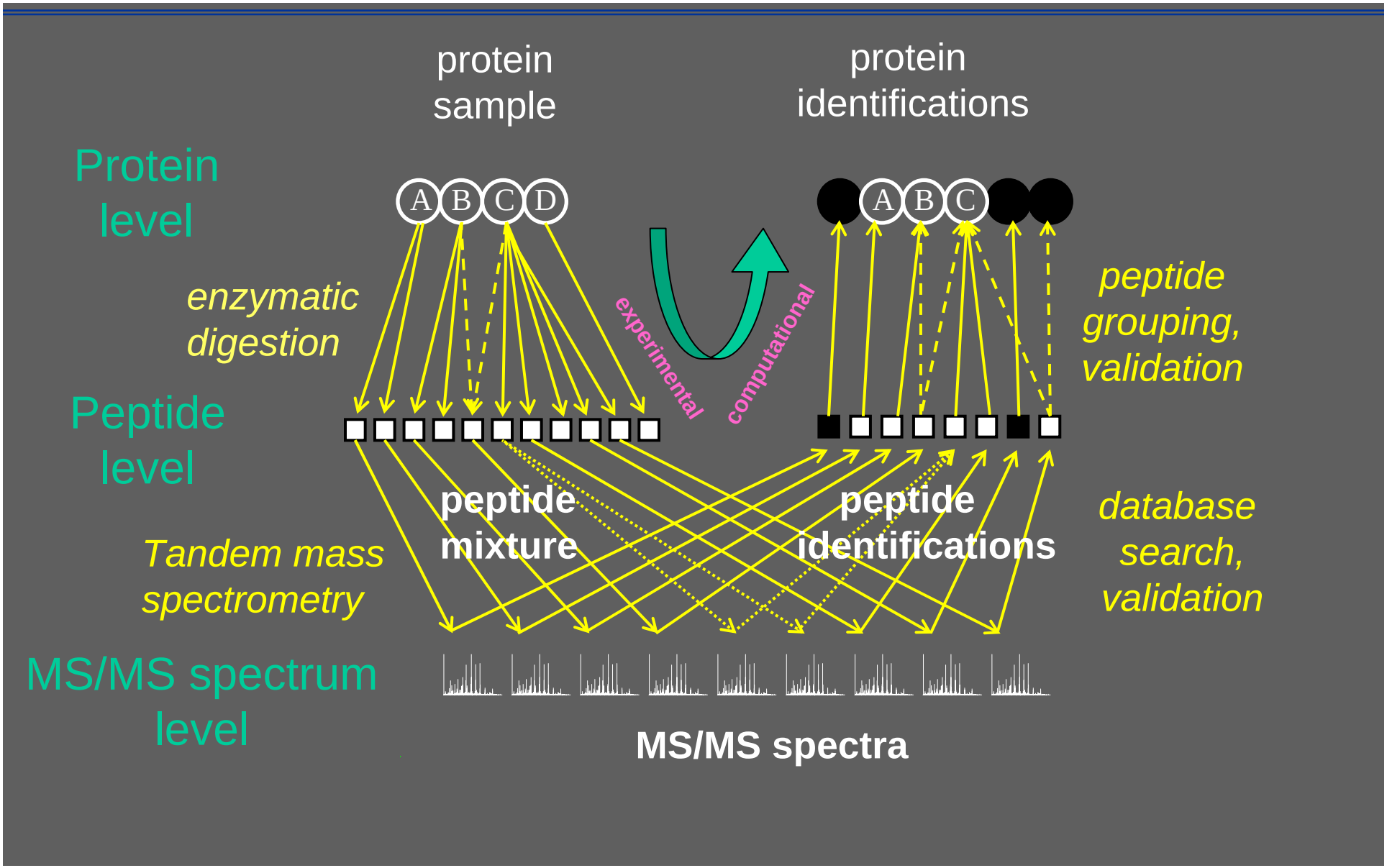
Outline

- ProteinProphet:
    1. peptide grouping
    2. adjusting peptide probabilities for protein grouping information
    3. protein inference problem
- Interpretation of shotgun proteomics data
- Data publication guidelines

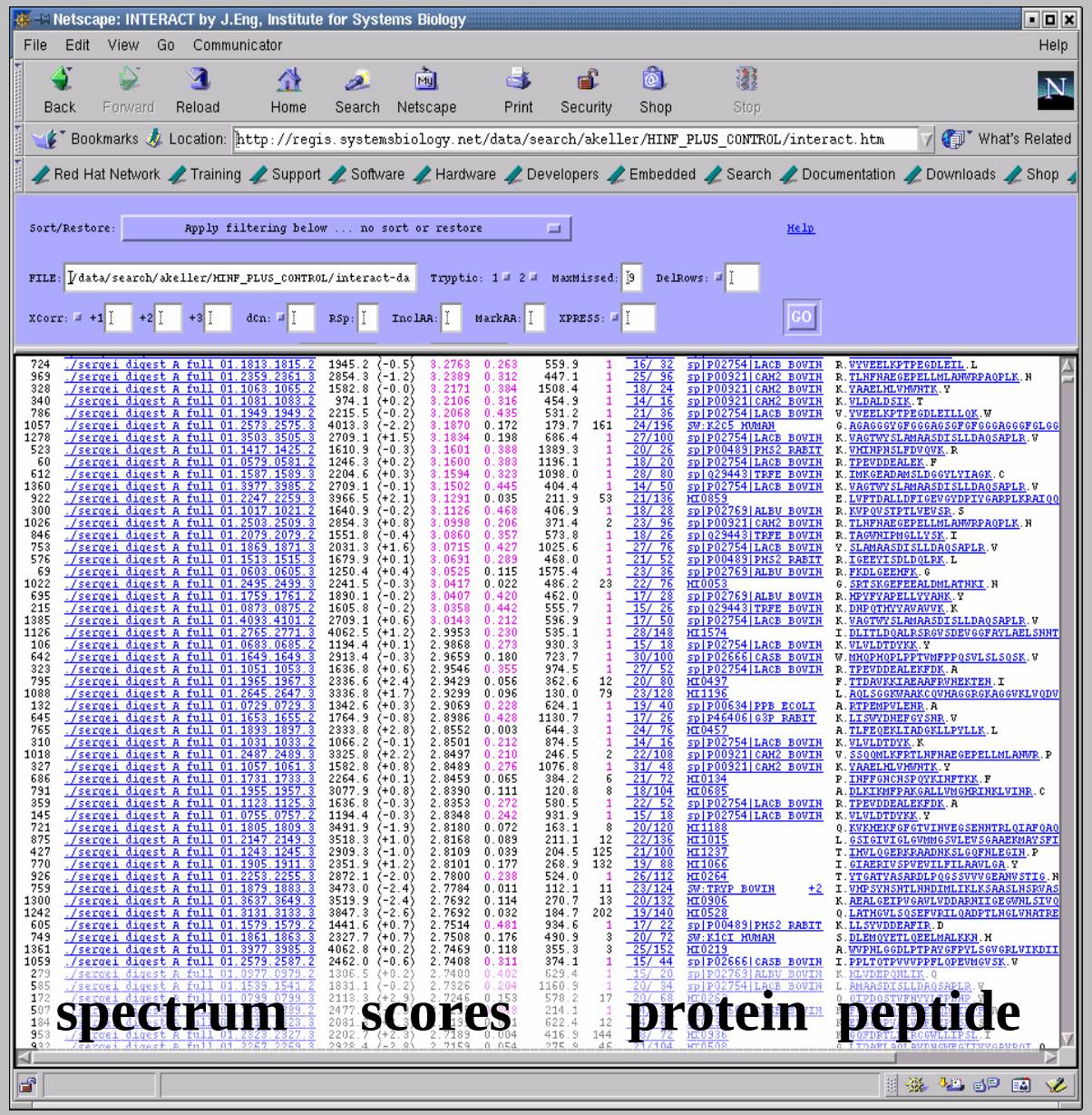
References: A statistical model for identifying proteins by tandem mass spectrometry, A. I. Nesvizhskii *et al*., *Anal. Chem.* 75, 4646-4658 (2003)

Interpretation of shotgun proteomic data: The Protein Inference Problem, A. I. Nesvizhskii, R. Aebersold, *Mol. Cell. Proteomics* 4, 1419-1440 (2005)

Shotgun Protein Identification



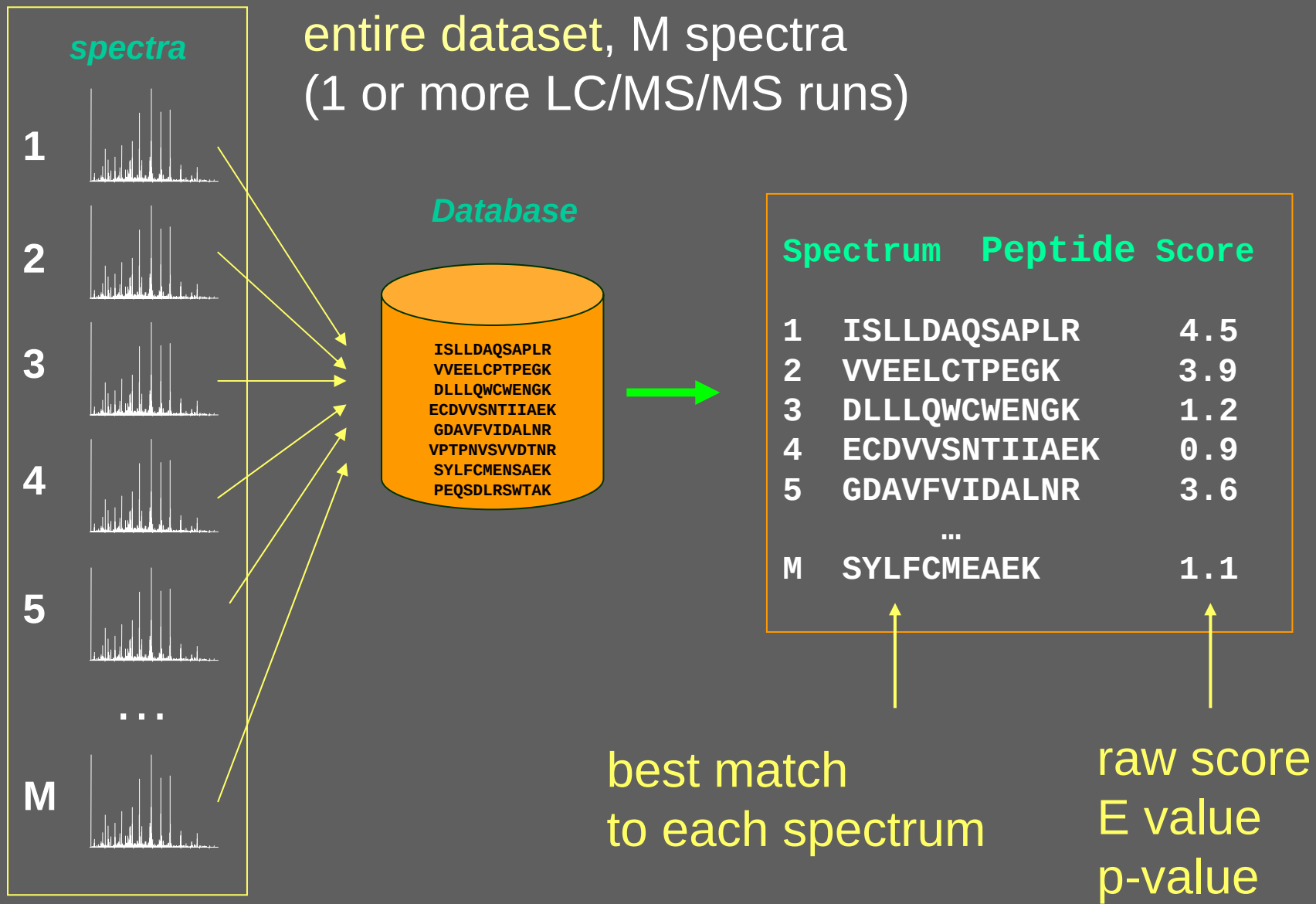
Peptide validation using PeptideProphet



MS/MS database search algorithms: SEQUEST, Mascot, Sonar, etc.

- thousands of spectra
- which peptide identification are correct?
- how to filter data?

Statistical Model



Bayesian Classification

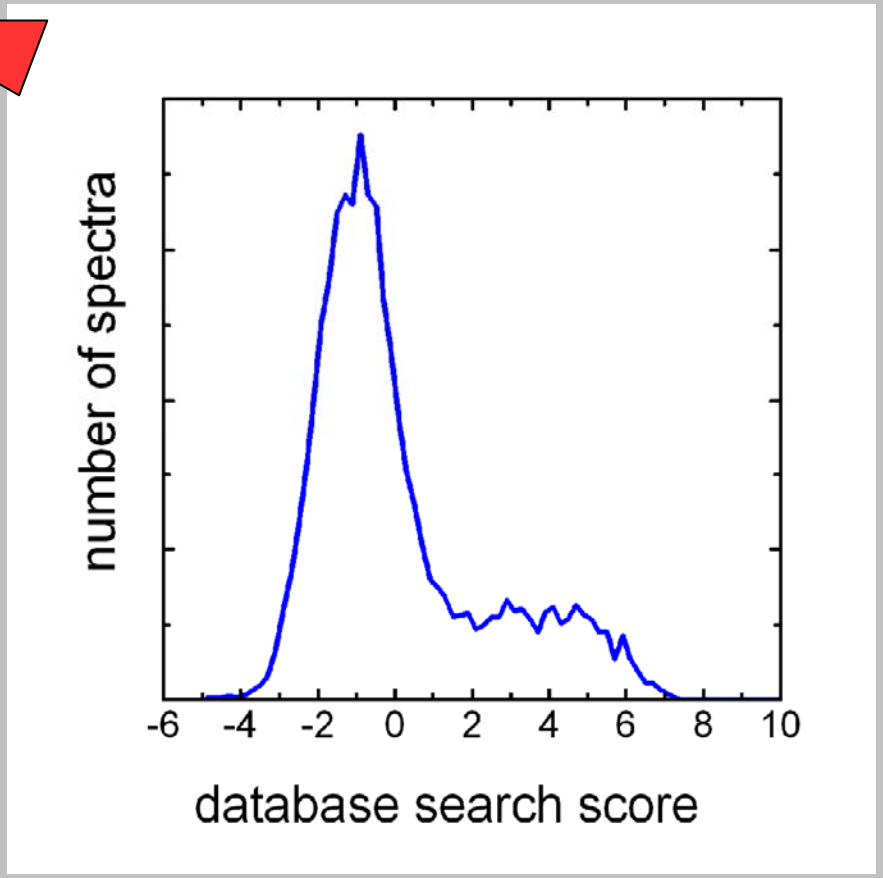
entire dataset:

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

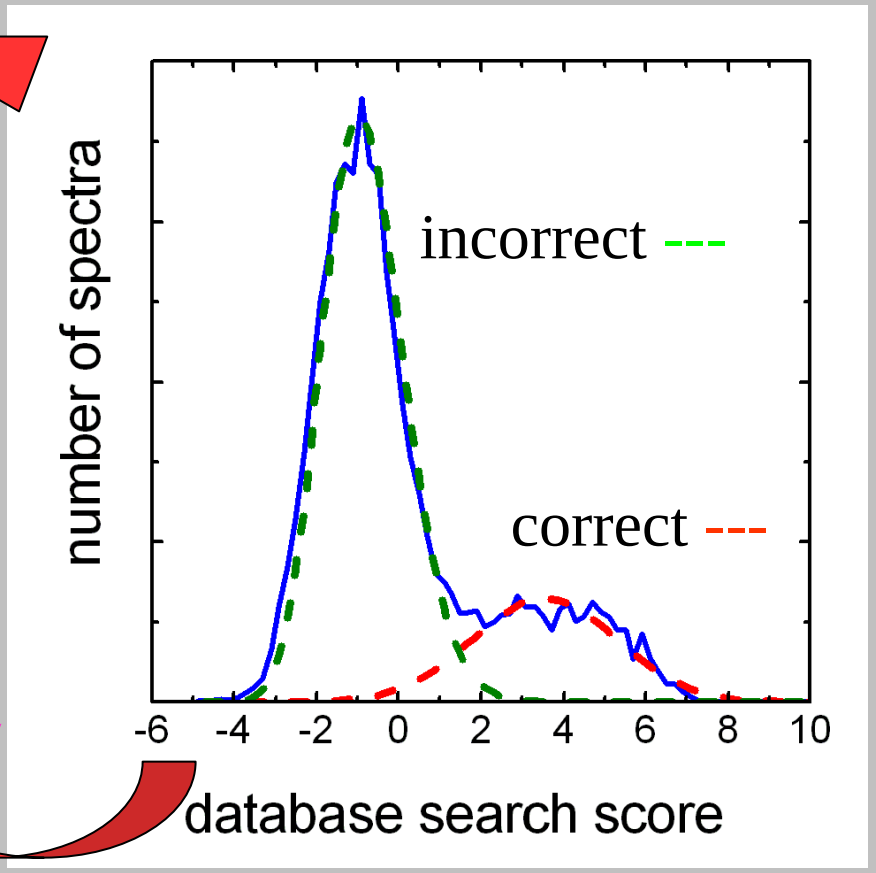


A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold *Anal. Chem*. 74, 5383 (2002)

Statistical Model for Computing Peptide Probabilities (PeptideProphet)

entire dataset:

Spectrum	Peptide	Score	probability
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



'unsupervised clustering'

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

PeptideProphet Software

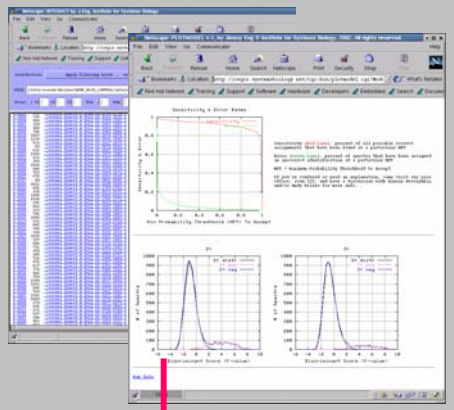


LCQ TOF-TOF Q-TOF

SEQUEST MASCOT

hard to compare datasets, unknown error rates

Also: different sample preparation sample complexity quality of spectra



Statistical model

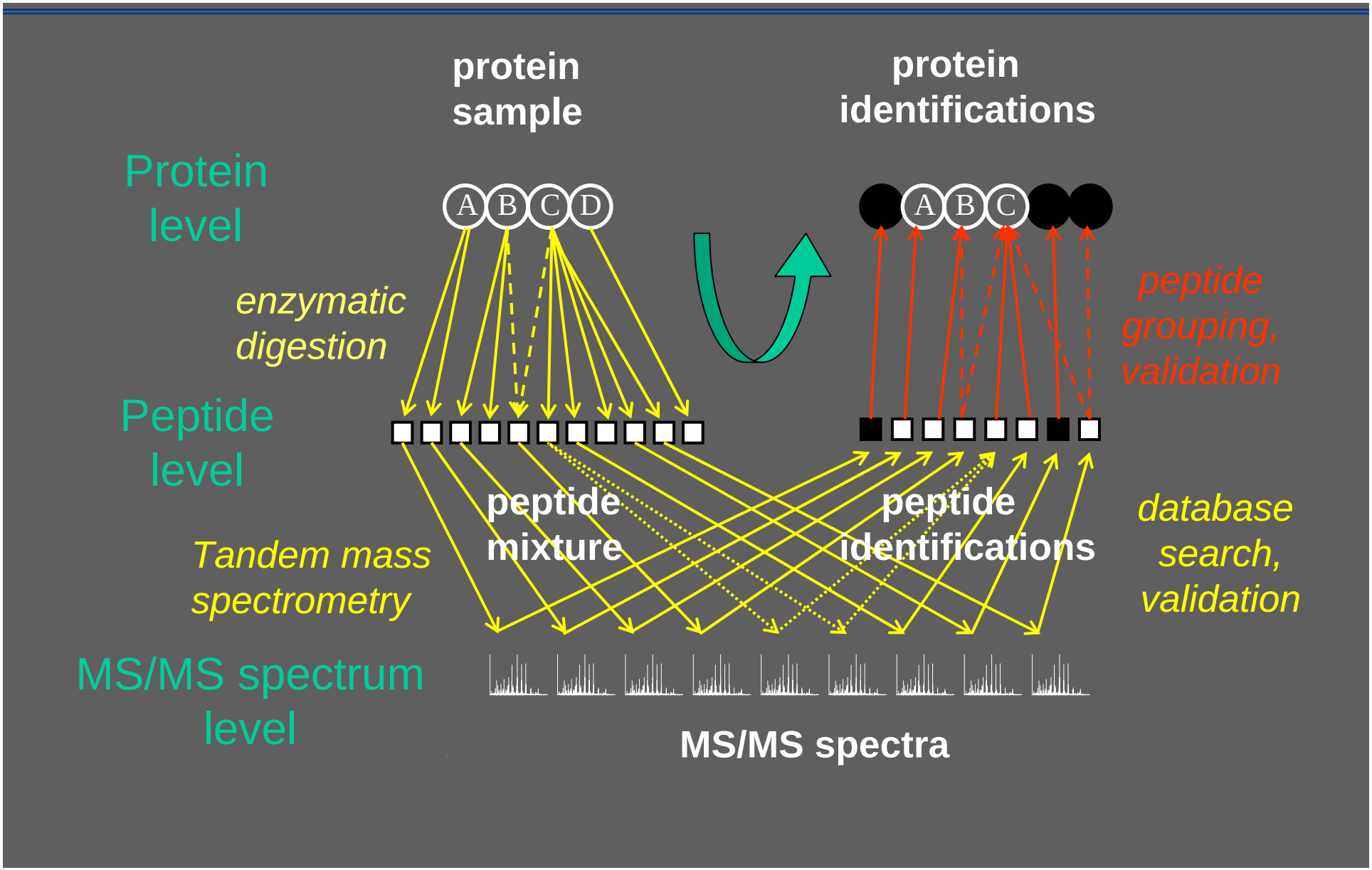
easy comparison quality control, known error rates

peptide identifications with accurate probabilities

Nice, but …

Want to know what proteins are present in the sample ?

Shotgun Protein Identification



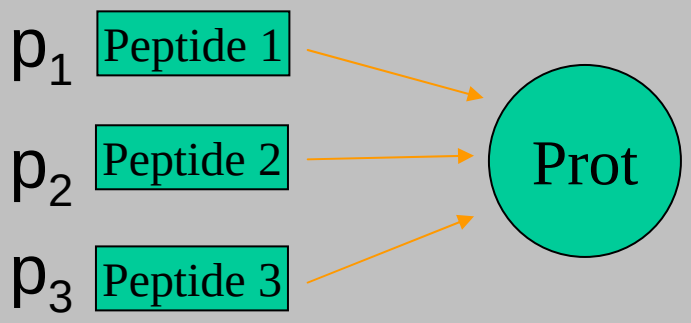
Statistical Model for Identifying Proteins (simple model)

- Group peptides by protein
- Compute a probability that a protein is present in the sample based upon the evidence of corresponding peptides in the dataset

Computing Protein Probabilities

Compute probability that at least one corresponding peptide is a correct identification

$$P(\text{protein}) = 1 - \Pi_i (1 - p_i)$$



$$P_{\text{incorrect}} = (1 - p_1)(1 - p_2)(1 - p_3)$$

$$P = 1 - (1 - p_1)(1 - p_2)(1 - p_3)$$

Computing Protein Probabilities

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

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKL
VLDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMH
IRLSFNPTQLEEQCHI

PTPEGDLEILLQK : p = 0.81

TPEVDDEALEK : p = 0.96

LSFNPTQLEEQCHI : p = 0.65

P = 1 – (1-0.81)(1-0.91)(1-0.65) = 0.99

Repeated Sequencing Events

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

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKL
VLDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMH
IRLSFNPTQLEEQCHI

PTPEGDLEILLQK : p = 0.81

PTPEGDLEILLQK : p = 0.62

PTPEGDLEILLQK : p = 0.95

same peptide sequenced multiple times

P = ??

Repeated Sequencing: Example

4 MS/MS spectra

input file /data2/search/nesvi/NEW_HUM_PLUS_CONTROL/TRYPTIC/interact-data.htm:

1	0.4946	618	./sergei_digest_A_full_02.1531.1535.2	1830.1 (+1.3)	2.3264	0.134	543.4	1	13/ 32	PIR2:T34540	+3	R.EAMDPIAELLSQLSGVR.R
2	0.8710	599	./sergei_digest_A_full_12.1565.1567.2	1830.1 (+1.1)	2.7030	0.221	616.9	1	15/ 32	PIR2:T34540	+3	R.EAMDPIAELLSQLSGVR.R
3	0.9654	622	./sergei_digest_A_full_13.1566.1568.2	1830.1 (+0.2)	2.4338	0.257	715.4	1	15/ 32	PIR2:T34540	+3	R.EAMDPIAELLSQLSGVR.R
4	0.5612	619	./sergei_digest_A_full_14.1564.1566.2	1830.1 (+0.1)	2.1465	0.062	529.1	1	13/ 32	PIR2:T34540	+3	R.EAMDPIAELLSQLSGVR.R

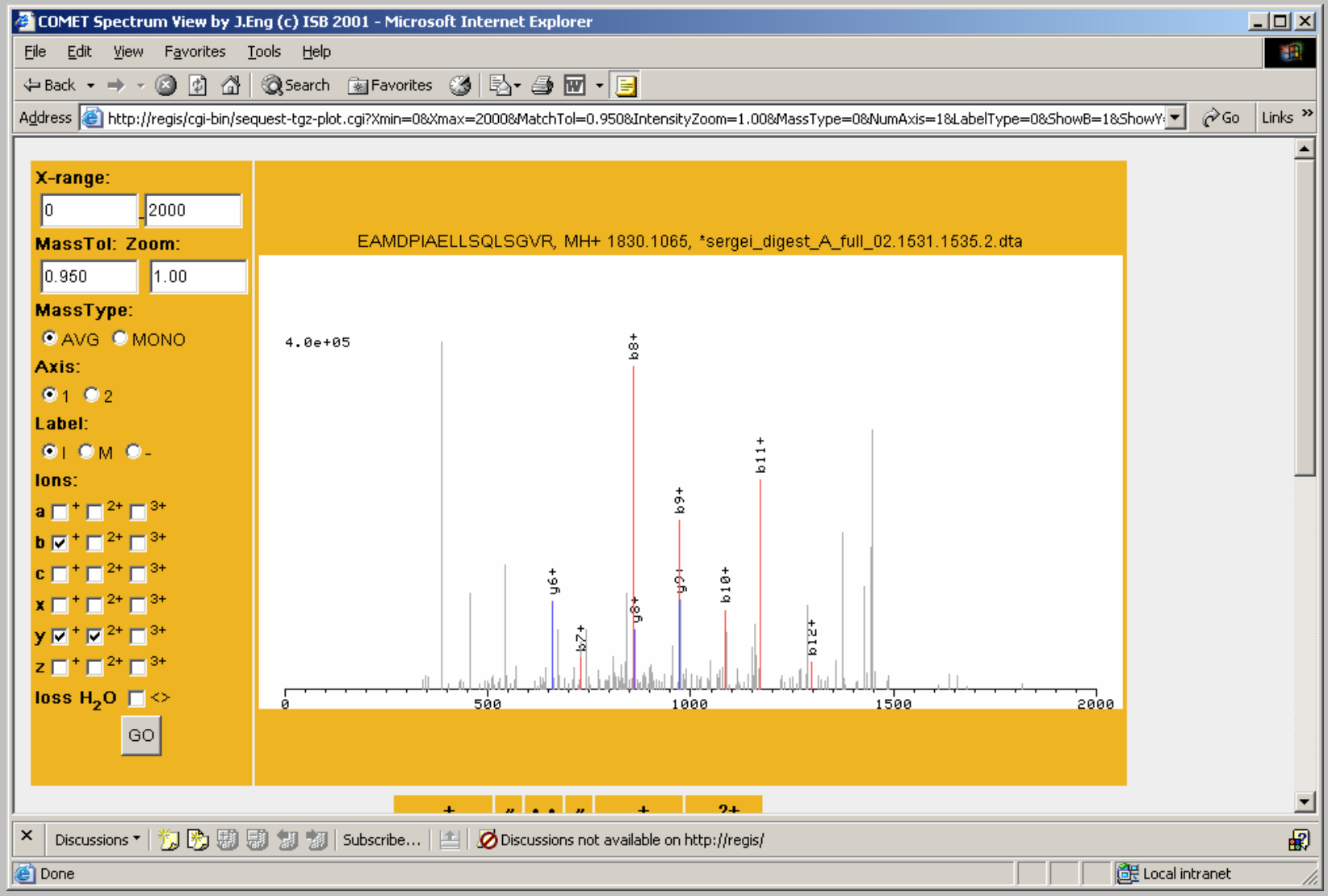
high probability peptide assignments

consistent retention time as indicated by the scan numbers

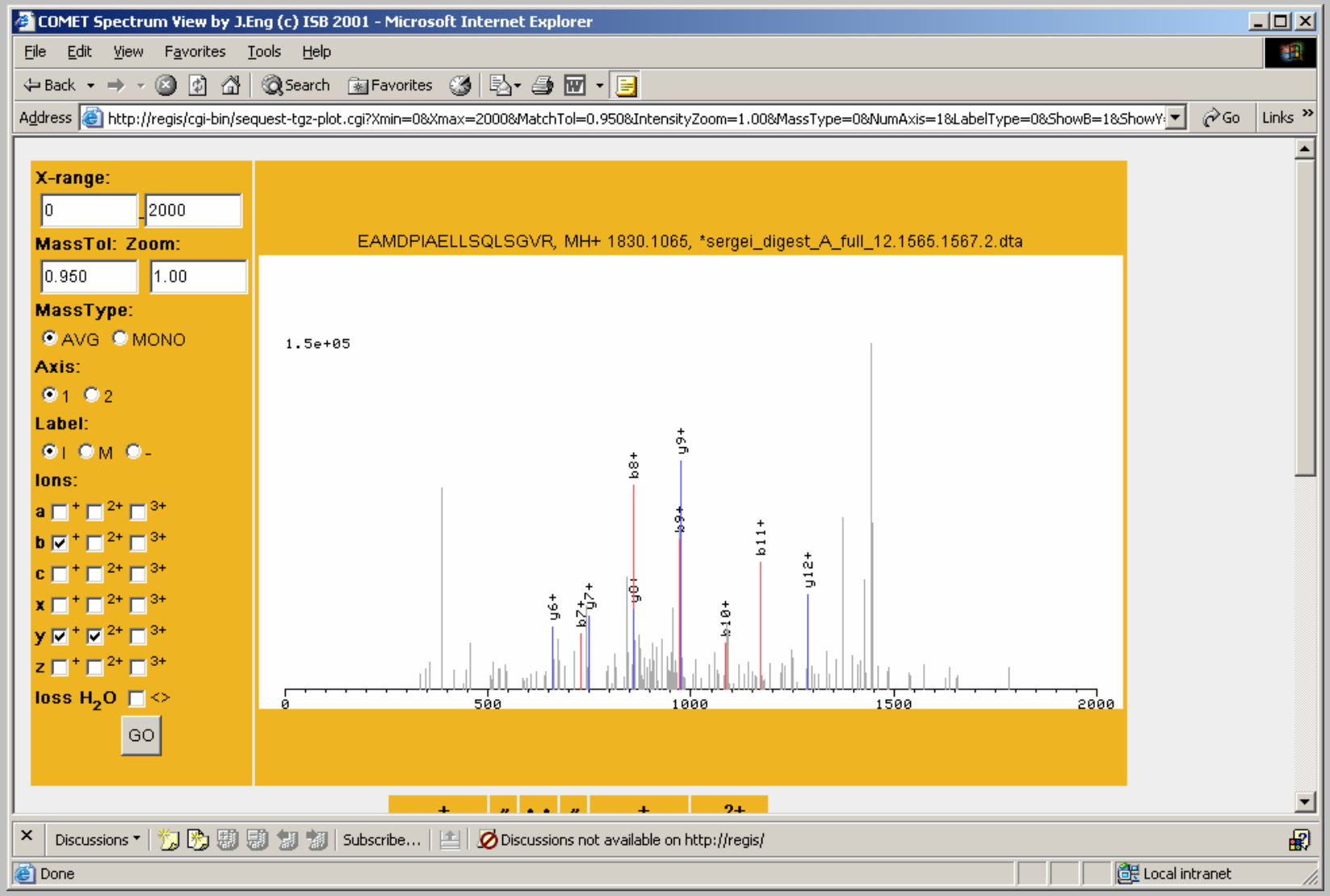
match to the same database peptide

note: spectra searched using SEQUEST with trypsin specified as the digestion enzyme

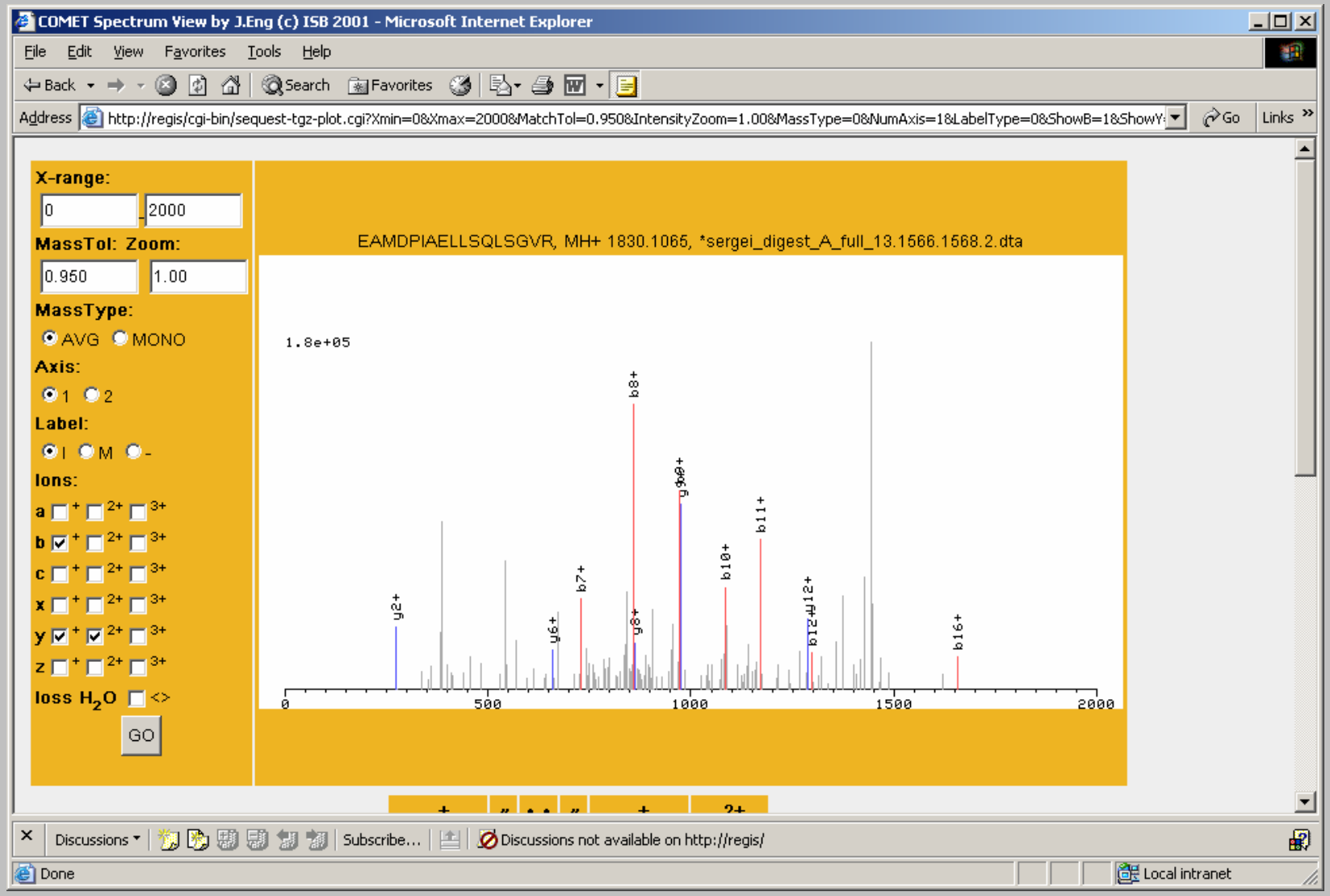
Spectrum 1


COMET Spectrum View by J.Eng (c) ISB 2001 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://regis/cgi-bin/sequest-tgz-plot.cgi?Xmin=0&Xmax=2000&MatchTol=0.950&IntensityZoom=1.00&MassType=0&NumAxis=1&LabelType=0&ShowB=1&ShowY
Go Links
X-range:
0
2000
MassTol: Zoom:
0.950
1.00
MassType:
AVG MONO
Axis:
1 2
Label:
I M -
Ions:
a + 2+ 3+
b + 2+ 3+
c + 2+ 3+
x + 2+ 3+
y + 2+ 3+
z + 2+ 3+
loss H2O <>
GO
EAMDPIAELLSQLSGVR, MH+ 1830.1065, *sergei_digest_A_full_02.1531.1535.2.dta
4.0e+05
b8+
b11+
b9+
y9+
y6+
y8+
b7+
b10+
b12+
0
500
1000
1500
2000
Discussions
Subscribe...
Discussions not available on http://regis/
Done
Local intranet

Spectrum 2

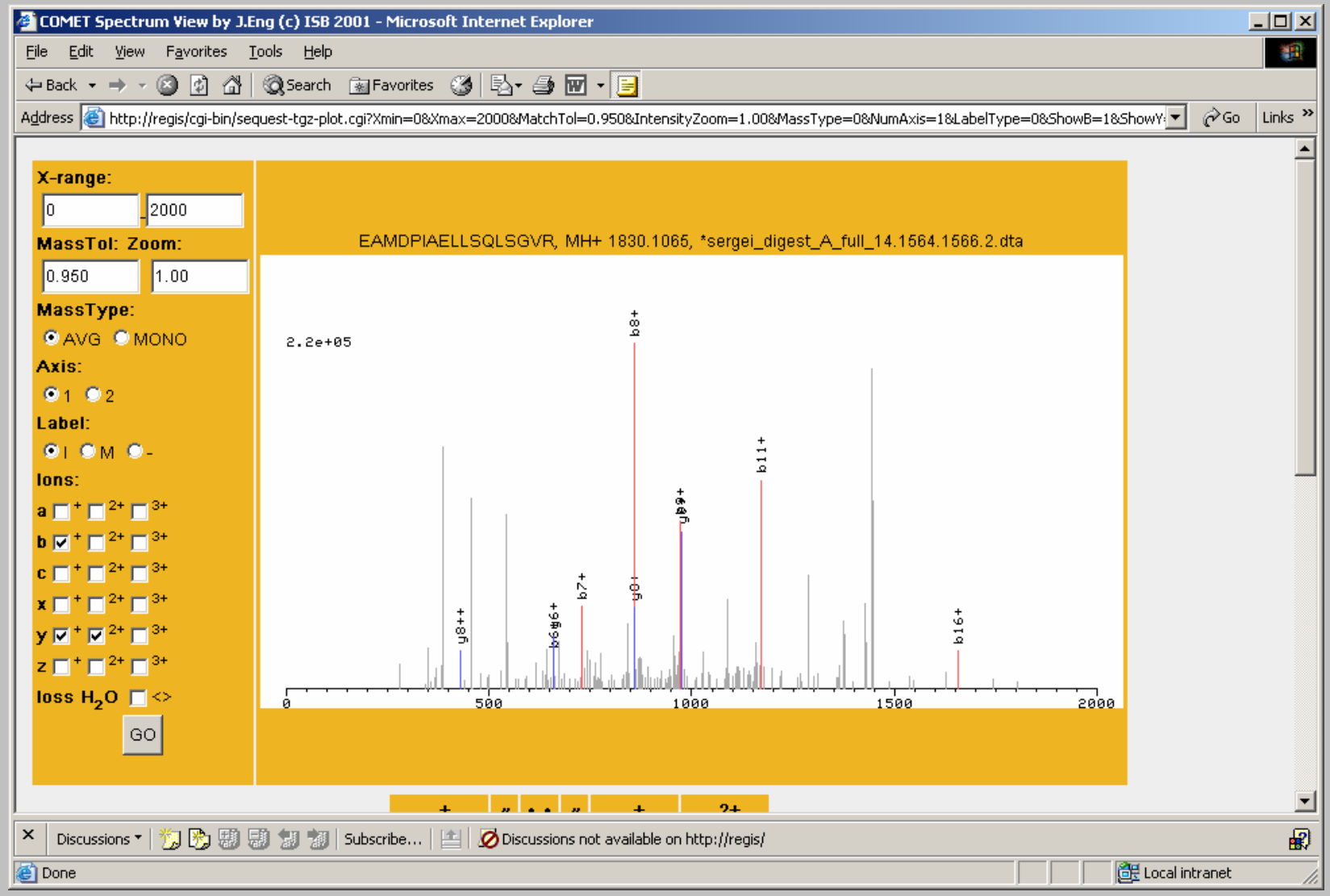

COMET Spectrum View by J.Eng (c) ISB 2001 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://regis/cgi-bin/sequest-tgz-plot.cgi?Xmin=0&Xmax=2000&MatchTol=0.950&IntensityZoom=1.00&MassType=0&NumAxis=1&LabelType=0&ShowB=1&ShowY
Go Links
X-range:
0
2000
MassTol: Zoom:
0.950
1.00
MassType:
AVG MONO
Axis:
1 2
Label:
I M -
Ions:
a + 2+ 3+
b + 2+ 3+
c + 2+ 3+
x + 2+ 3+
y + 2+ 3+
z + 2+ 3+
loss H2O <>
GO
EAMDPIAELLSQLSGVR, MH+ 1830.1065, *sergei_digest_A_full_12.1565.1567.2.dta
1.5e+05
b8+
y9+
b9+
b11+
y12+
y6+
b7+
y7+
y8+
b10+
0
500
1000
1500
2000
Discussions
Subscribe...
Discussions not available on http://regis/
Done
Local intranet

Spectrum 3



COMET Spectrum View by J.Eng (c) ISB 2001 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://regis/cgi-bin/sequest-tgz-plot.cgi?Xmin=0&Xmax=2000&MatchTol=0.950&IntensityZoom=1.00&MassType=0&NumAxis=1&LabelType=0&ShowB=1&ShowY
Go Links
X-range:
0
2000
MassTol: Zoom:
0.950
1.00
MassType:
AVG MONO
Axis:
1 2
Label:
I M -
Ions:
a + 2+ 3+
b + 2+ 3+
c + 2+ 3+
x + 2+ 3+
y + 2+ 3+
z + 2+ 3+
loss H2O <>
GO
EAMDPIAELLSQLSGVR, MH+ 1830.1065, *sergei_digest_A_full_13.1566.1568.2.dta
1.8e+05
b8+
b11+
b10+
b7+
y2+
y6+
y8+
b16+
0
500
1000
1500
2000
Discussions
Subscribe...
Discussions not available on http://regis/
Done
Local intranet

Spectrum 4


COMET Spectrum View by J.Eng (c) ISB 2001 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://regis/cgi-bin/sequest-tgz-plot.cgi?Xmin=0&Xmax=2000&MatchTol=0.950&IntensityZoom=1.00&MassType=0&NumAxis=1&LabelType=0&ShowB=1&ShowY=
Go Links
X-range:
0
2000
MassTol: Zoom:
0.950
1.00
MassType:
AVG MONO
Axis:
1 2
Label:
I M -
Ions:
a + 2+ 3+
b + 2+ 3+
c + 2+ 3+
x + 2+ 3+
y + 2+ 3+
z + 2+ 3+
loss H2O <>
GO
EAMDPIAELLSQLSGVR, MH+ 1830.1065, *sergei_digest_A_full_14.1564.1566.2.dta
2.2e+05
b8+
b11+
y8++
b7+
y8+
b16+
0
500
1000
1500
2000
Discussions
Subscribe...
Discussions not available on http://regis/
Done
Local intranet

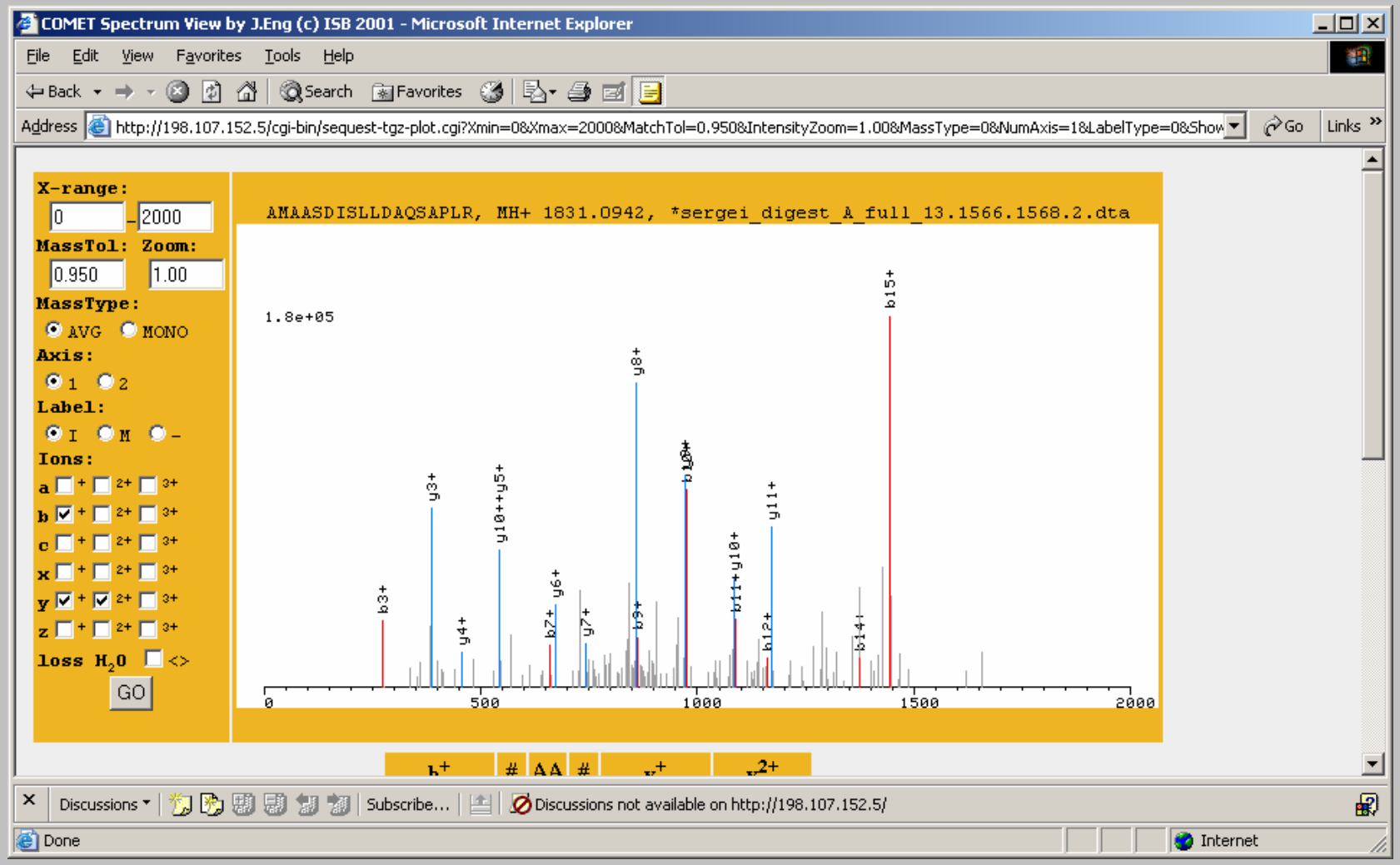
All 4 peptide assignments shown above are incorrect!

Search without enzyme specificity reveals the true peptide (partially tryptic):

Rank/Sp	(M+H)+	deltCn	XCorr	Sp	Ions	Reference		Peptide
1 / 1	1831.0942	0.0000	3.6893	1636.0	20/34	sp|P02754|LACB_BOVIN		L.AMAASDISLLDAQSAPLR.V
2 / 25	1830.1065	0.3403	2.4338	715.4	15/32	PIR2:T34540	+3	R.EAMDPIAELLSQLSGVR.R
3 / 10	1830.1035	0.3483	2.4042	827.1	16/34	SW:GBAZ_HUMAN	+1	L.TGPAESKGEITPELLGVM.R
4 /126	1830.0453	0.3855	2.2671	554.0	15/32	GP:AB032958_1		T.KDNGDEVVSTLKLKPAD.R
5 /344	1830.1956	0.4158	2.1553	451.5	13/34	GP:U69881_1		R.TPSVASKGVVLLMHGLYG.S
6 / 13	1828.0752	0.4214	2.1345	786.5	15/34	GP:BC007653_1		S.TPLASFKILALESADGHG.G
7 / 74	1830.0079	0.4277	2.1114	610.4	13/30	SW:CG1_HUMAN	+1	I.RSLTPTSNLLSQQQQQ.Q
8 /188	1830.0483	0.4357	2.0820	508.2	13/30	GP:AF263742_1	+1	Q.VENNQALSLLSKEQKQ.R
9 / 57	1830.0068	0.4363	2.0798	640.8	13/28	SWN:EPPL_HUMAN		G.QRVSAWELINSEYFS.E
10 / 49	1829.8743	0.4379	2.0736	653.2	15/32	GP:AF055994_1	+1	S.SSSGHSQSTLFDSDVFQ.T

sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG) (ALLERGEN BOS D 5) - Bos taurus (Bovine).
PIR2:T34540 hypothetical protein DKFZp434L1021.1 - human (fragment) [MASS=30460]

Correct Assignment

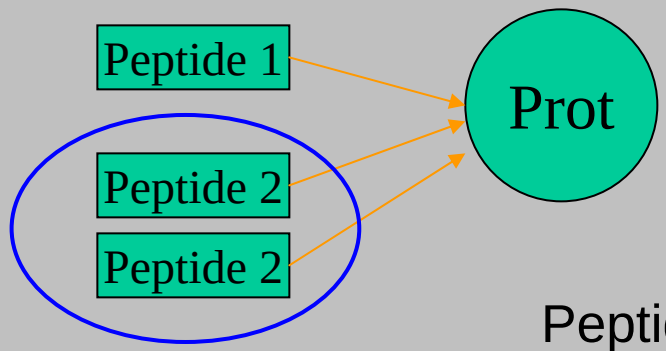


Assignments of Multiple Spectra to the Same Peptide

Assignments to the ***same*** peptide are not independent

conservative approach:

$$P(\text{protein}) = 1 - \Pi_i \, (1 - max_j \, p_i^j)$$



$$P = 1 - \; (1 - p_1) \, (1 - max \, (p_{2A}, \, p_{2B}))$$

Peptide assignment with highest probability is kept, other instances of the same peptide do not contribute to the protein probability.

Treatment of Repeated Sequencing Events

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

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKL
VLDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMH
IRLSFNPTQLEEQCHI

PTPEGDLEILLQK : p = 0.81

PTPEGDLEILLQK : p = 0.62

PTPEGDLEILLQK : p = 0.95

same peptide sequenced multiple times

protein P = 0.95

Note: identifications of the same peptide sequence from MS/MS spectra of different charge state are still considered ‘independent’ events. Same in the case of modifications (light/heavy isotope tags, oxidation etc.)

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)

A. I. Nesvizhskii, R. Aebersold, *MCP* 4, 1419-1440 (2005)

Problem # 1

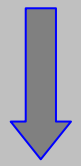
Non-random Grouping of Peptides according to Their Corresponding Protein

(single-hit protein identification problem)

Non-random Grouping of Peptides

Correct peptide assignments tend to correspond to "multihit" proteins, those to which other correctly assigned peptides correspond

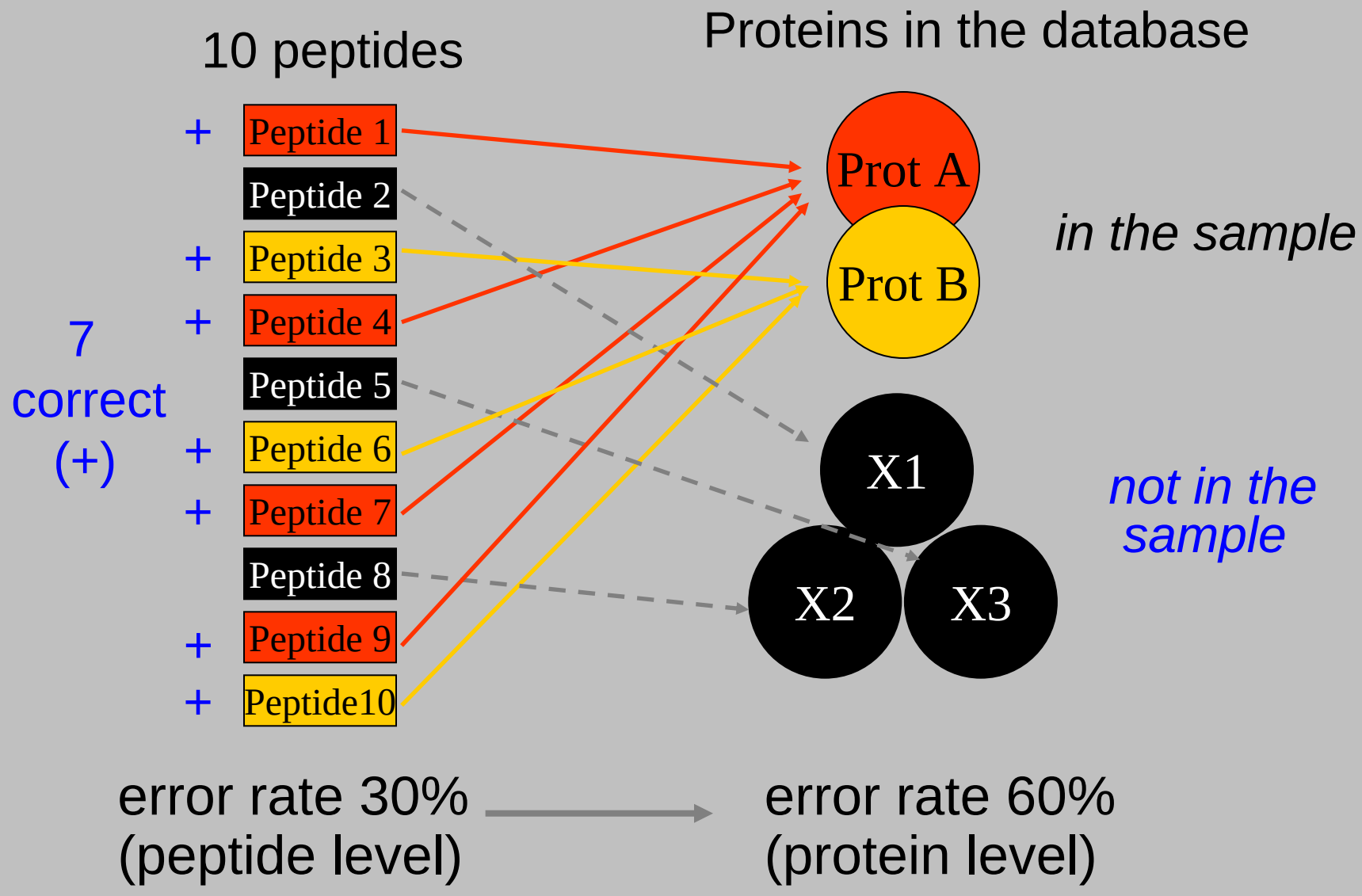
Incorrect peptide assignments tend to correspond to "single-hit" proteins to which no other correctly assigned peptide corresponds



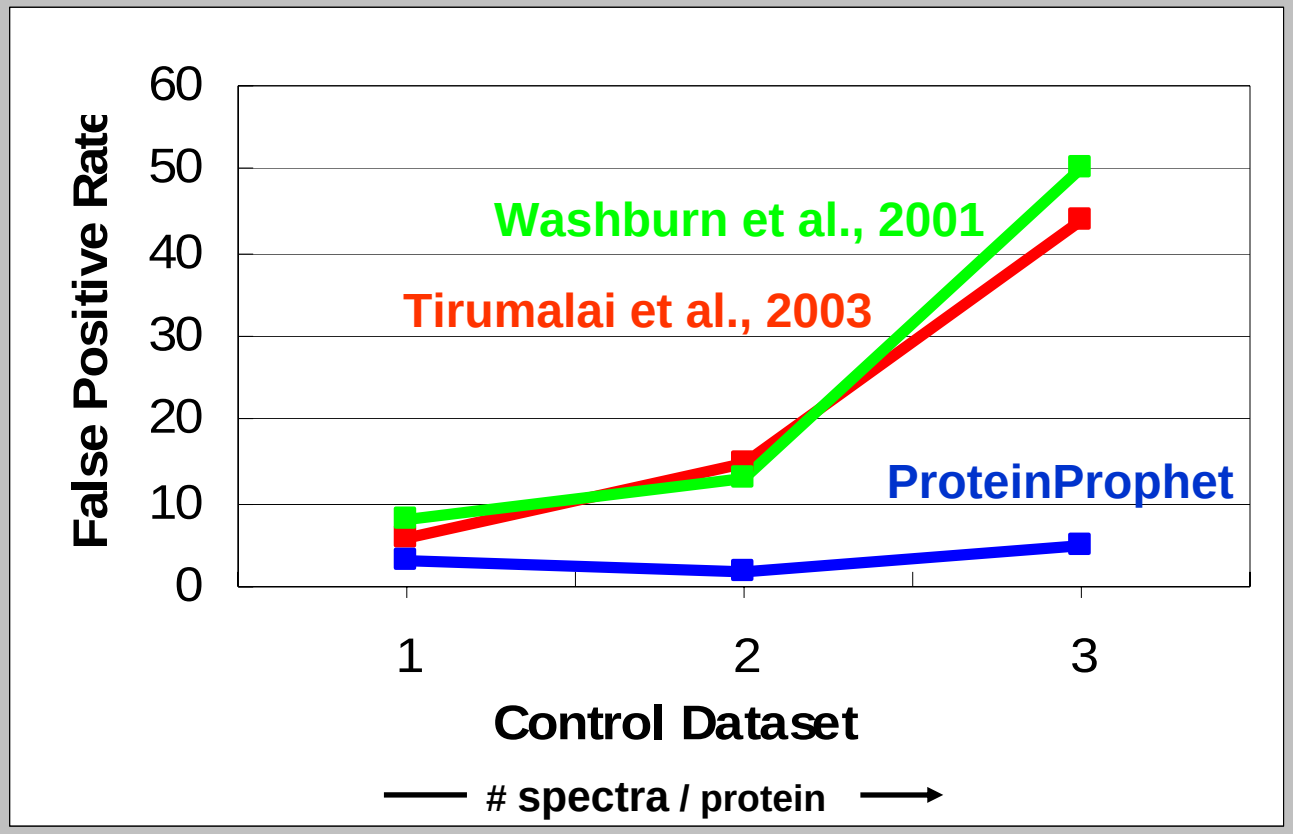
False positive identification error rate on the protein level is higher than on the peptide level

Hard to distinguish single-hit correct proteins form the incorrect ones

Non-random Grouping of Peptides according to Corresponding Protein



Protein ID False Positive Rates: ProteinProphet vs. Score Thresholds



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

Control Datasets:

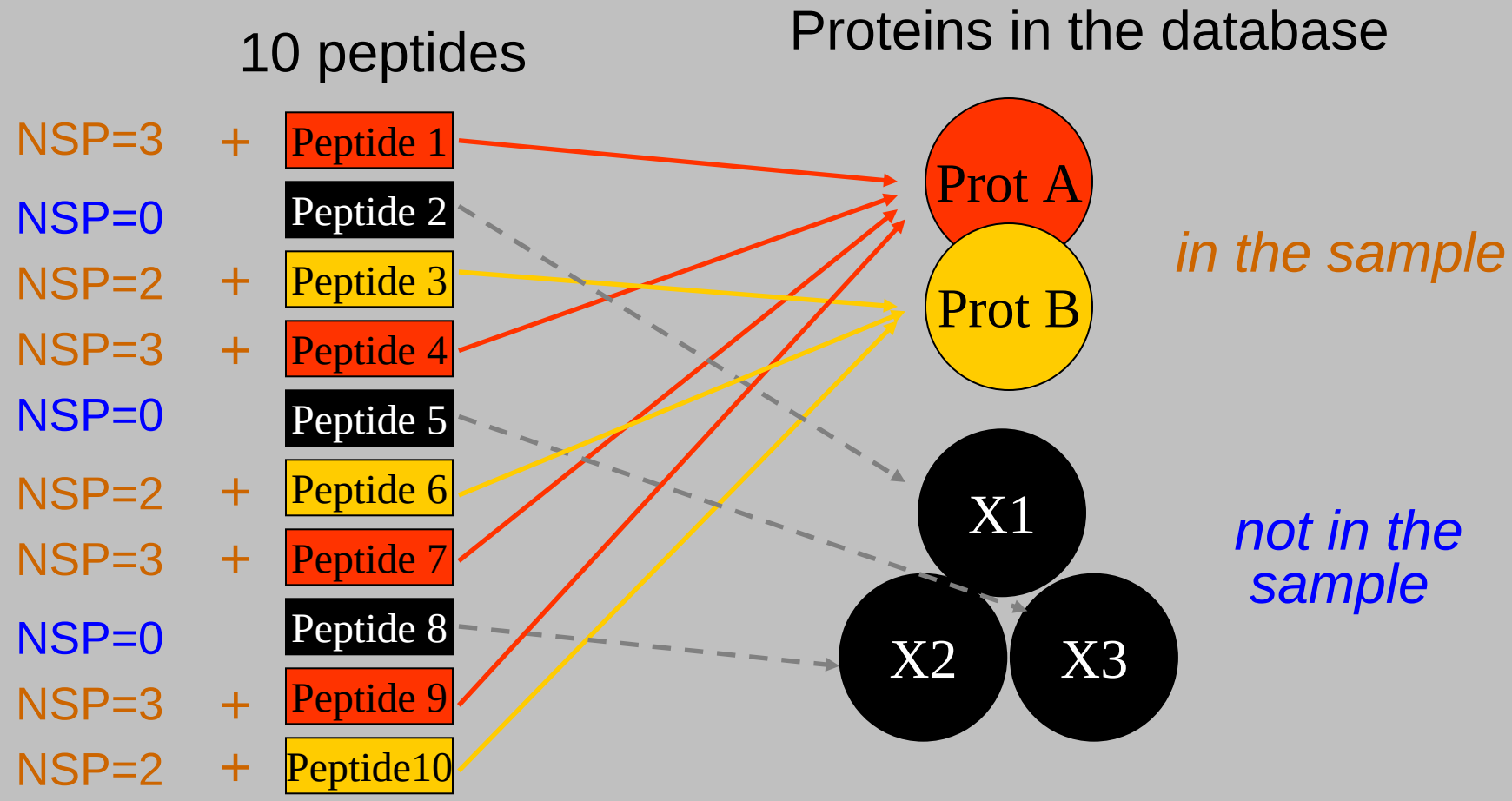
1 Halobacterium vs. Halo+Human (4 runs)
2 Halobacterium vs. Halo+Human (45 runs)
3 18 purified proteins vs. 18+Human (22 runs)

Serum Protein Identifications from Large-scale (~375 runs) Experiment

Data Filter	*# ids*	*# non-single hits*	*# single-hits*
Publ. Threshold mode l#1	2257	359	1898
Publ. Threshold model #2	2742	441	2301
ProteinProphet, p≥ 0.5 (*predicted error rate: 7%*)	713	511	202

Reference: *H. Zhang et al., in prep*

Number of Sibling Peptides (NSP) as Measure of Protein Grouping



Number of sibling peptides (NSP): measure of how many other peptides in dataset correspond to same protein

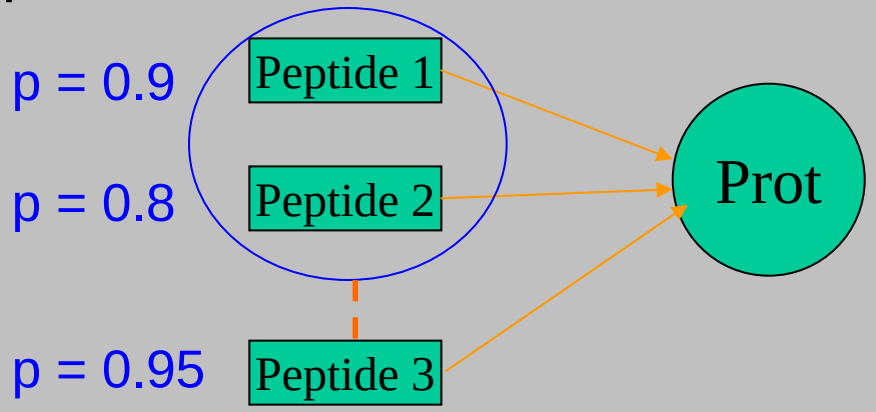
Number of Sibling Peptides (NSP)

peptide i ∈ protein P; sum over all other j ∈ protein P:

$$NSP_i = \sum_{j \neq i} p(+ \mid D_j)$$

peptide probability
computed by PeptideProphet

example:



peptide 3 has 2 siblings, peptide 1 and peptide 2
It has NSP = 1.7 (calculated as 0.9 + 0.8)

Adjusting Peptide Probabilities for NSP

$$p(+\mid D, NSP) = \frac{p(+\mid D)\,p(NSP\mid +)}{p(+\mid D)\,p(NSP\mid +) + p(-\mid D)\,p(NSP\mid -)}$$

adjusted probability

Amount of adjustment depends on $\dfrac{p(NSP\mid +)}{p(NSP\mid -)}$

- Learn NSP distributions from each dataset (EM algorithm)
- Adjust peptide probabilities to include NSP information

  peptides with high *NSP* – more confidence $\rightarrow$ p $\uparrow$
  peptides with low *NSP* – less confidence $\rightarrow$ p $\downarrow$

The appropriate amount of adjustment for NSP is determined by the model from the data

NSP Distributions

Why do we need to determine the appropriate amount of adjustment for NSP (e.g., by how much we should penalize single-hits) from the data?

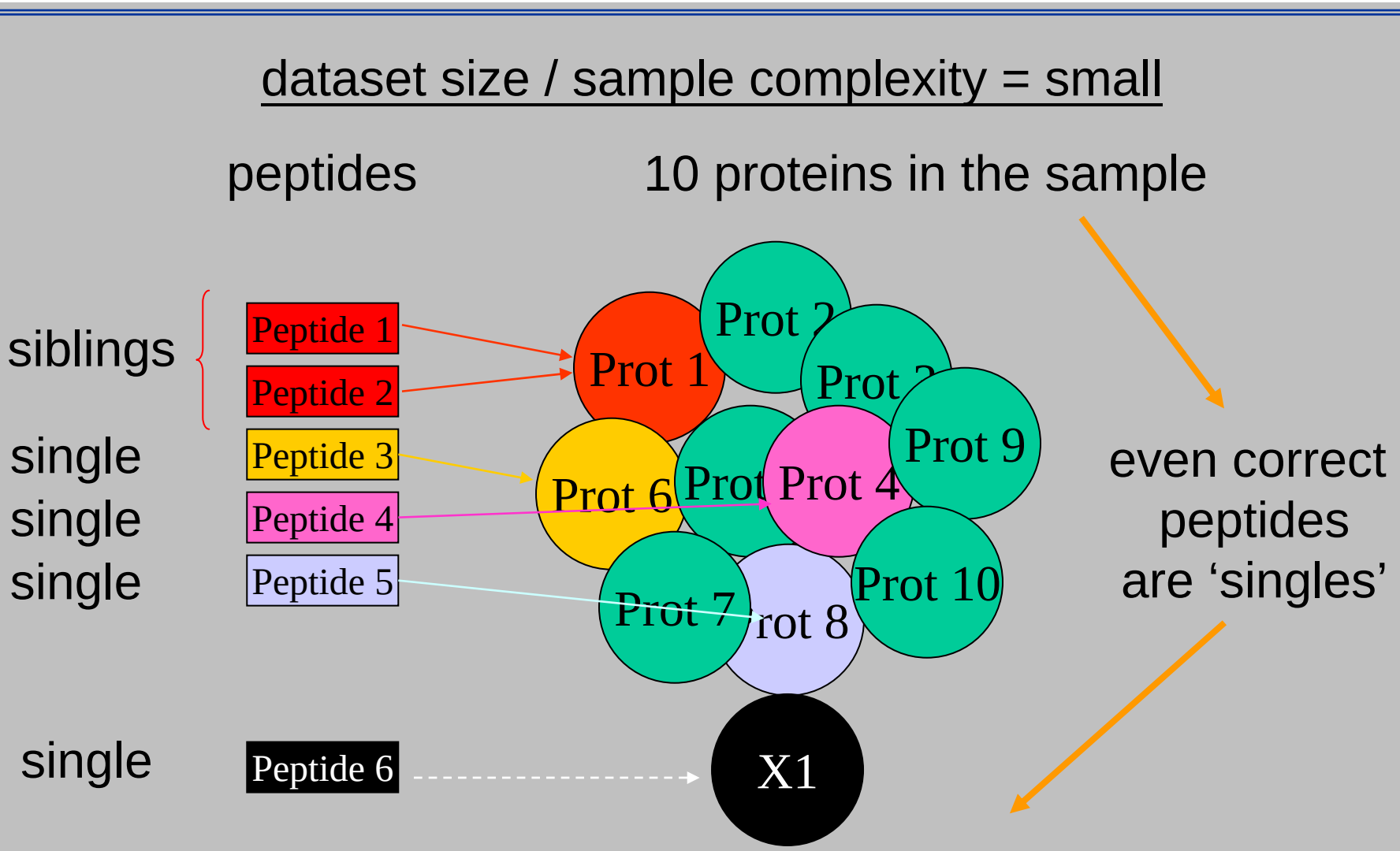
NSP distributions among correct and incorrect peptide assignments, p(NSP|+) and p(NSP|-), vary from dataset to dataset. They strongly depend on sample "coverage".

coverage = dataset size / sample complexity

sample complexity:	number of proteins
dataset size:	number of peptide assignments (number of acquired MS/MS spectra)

Low Coverage Dataset

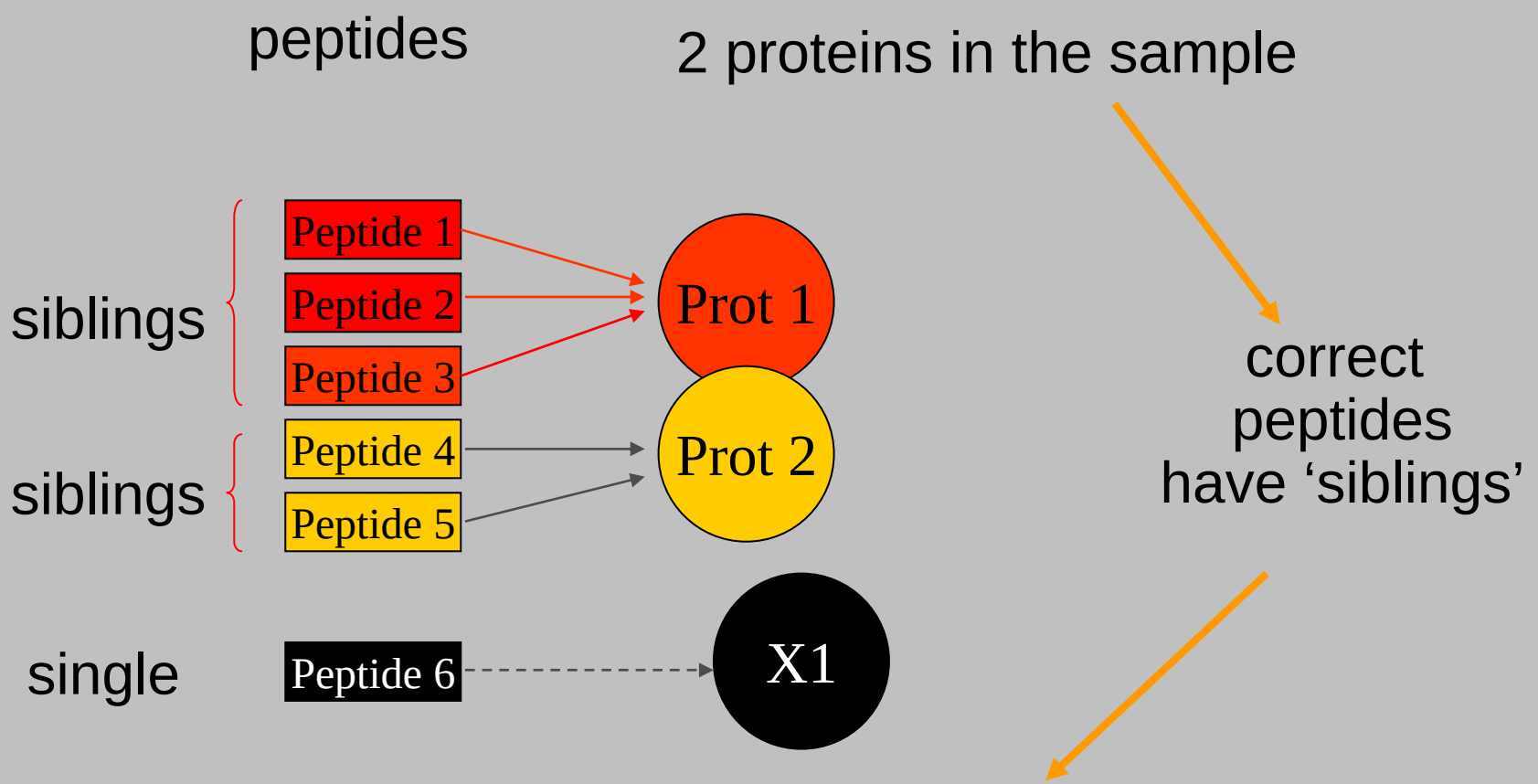
dataset size / sample complexity = small



small difference between p(NSP|+) and p(NSP|-)

High Coverage Dataset

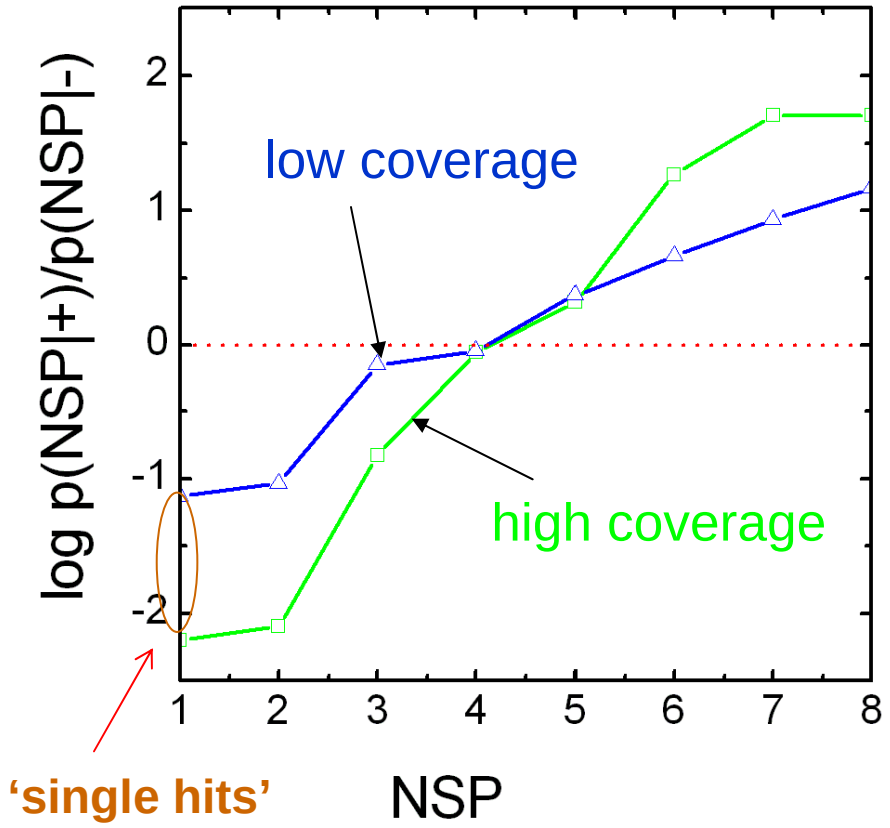
dataset size / sample complexity = large



large difference between p(NSP|+) and p(NSP|-)

NSP distributions: dependence on sample coverage

before adjusting for *NSP*



Similar sample size (# of sequencing attempts)

Different sample complexity:

- complex *H. Influenzae* sample (low coverage)
- 18 protein sample (high coverage)

The higher the coverage, the bigger the adjustment for NSP

Adjusting Peptide Probabilities: Example 1

Given the initial peptide probabilities, the model "learns" that:

If a peptide has "1 sibling" (*NSP=0.5+0.5=1*), then *p(NSP|+) = 0.4* and *P(NSP|-) = 0.1* (that is, a peptide with 2 siblings is 4 times more likely to be correct assignment than incorrect one

10 peptides with p=0.5

- Peptide 1
- Peptide 2
- Peptide 3
- Peptide 4
- Peptide 5
- Peptide 6
- Peptide 7
- Peptide 8
- Peptide 9
- Peptide10

$$p' = \frac{p \cdot p(NSP \mid +)}{p \cdot p(NSP \mid +) + (1-p) \cdot p(NSP \mid -)}$$

$p(NSP|+) = 0.4 \quad p(NSP|-) = 0.1$

$$p' = \frac{0.5 \cdot 0.4}{0.5 \cdot 0.4 + (1-0.5) \cdot 0.1} = 0.8$$

Peptides 1, 4, and 7: adjusted p = 0.8

Adjusting Peptide Probabilities: Example 2

Given the initial peptide probabilities, the model "learns" that:

If a peptide has 0 sibling (*NSP=0*) , then *p(NSP|+) = 0.1* and *P(NSP|-) = 0.4* (that is, a peptide with no siblings is 4 times less likely to be correct assignment than incorrect one

10 peptides with p=0.5

- Peptide 1
- Peptide 2
- Peptide 3
- Peptide 4
- Peptide 5
- Peptide 6
- Peptide 7
- Peptide 8
- Peptide 9
- Peptide10

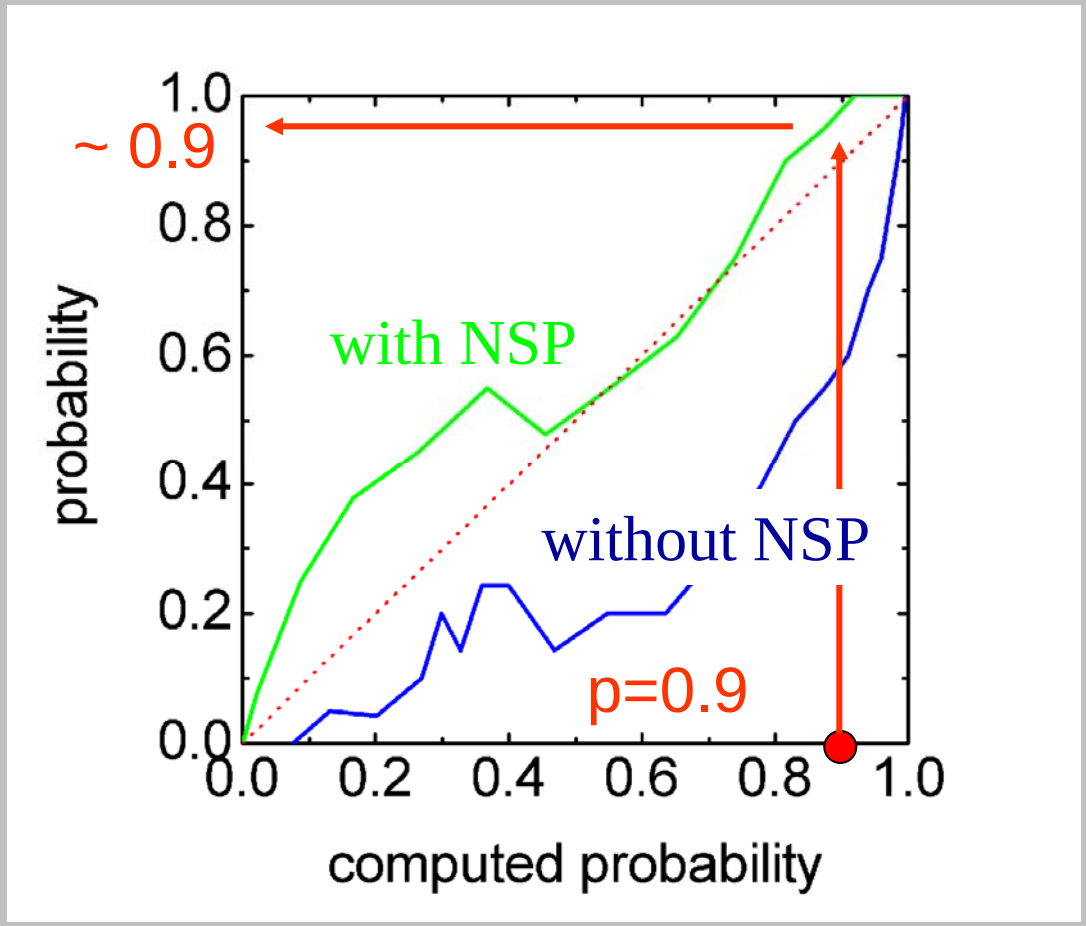
$$p' = \frac{p \cdot p(NSP \mid +)}{p \cdot p(NSP \mid +) + (1-p) \cdot p(NSP \mid -)}$$

$p(NSP|+) = 0.1 \quad p(NSP|-) = 0.4$

$$p' = \frac{0.5 \cdot 0.1}{0.5 \cdot 0.1 + (1-0.5) \cdot 0.4} = 0.2$$

Peptides 2, 5, 6, 8, and 9: adjusted p = 0.2

Accuracy of Computed Protein Probabilities



100 proteins with computed p ~ 0.9

90% of them (90) should be correct

observed probability is around 0.9

Statistical model is more accurate with adjustment for protein grouping information (NSP)

ProteinProphet (with adjustment for NSP)

- Group peptides by protein
- Adjust peptide probabilities for grouping information (penalize ‘single hits’)
- Compute a probability that a protein is present in the sample based upon the evidence of corresponding peptides in the dataset (using adjusted peptide probabilities)

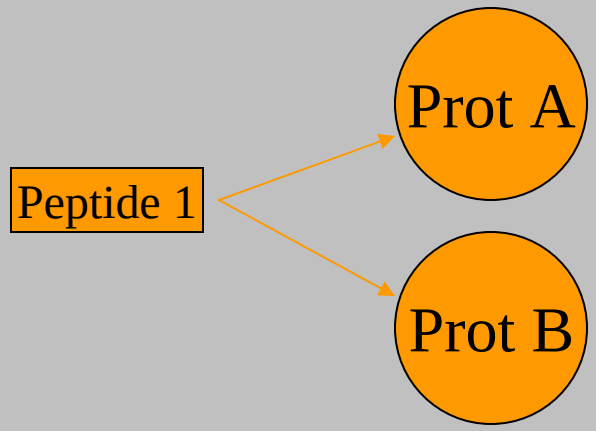
Problem # 2

Protein Inference Problem

(“degenerate“ or “shared” peptides)

Protein Inference Problem

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



protein A or protein B ??

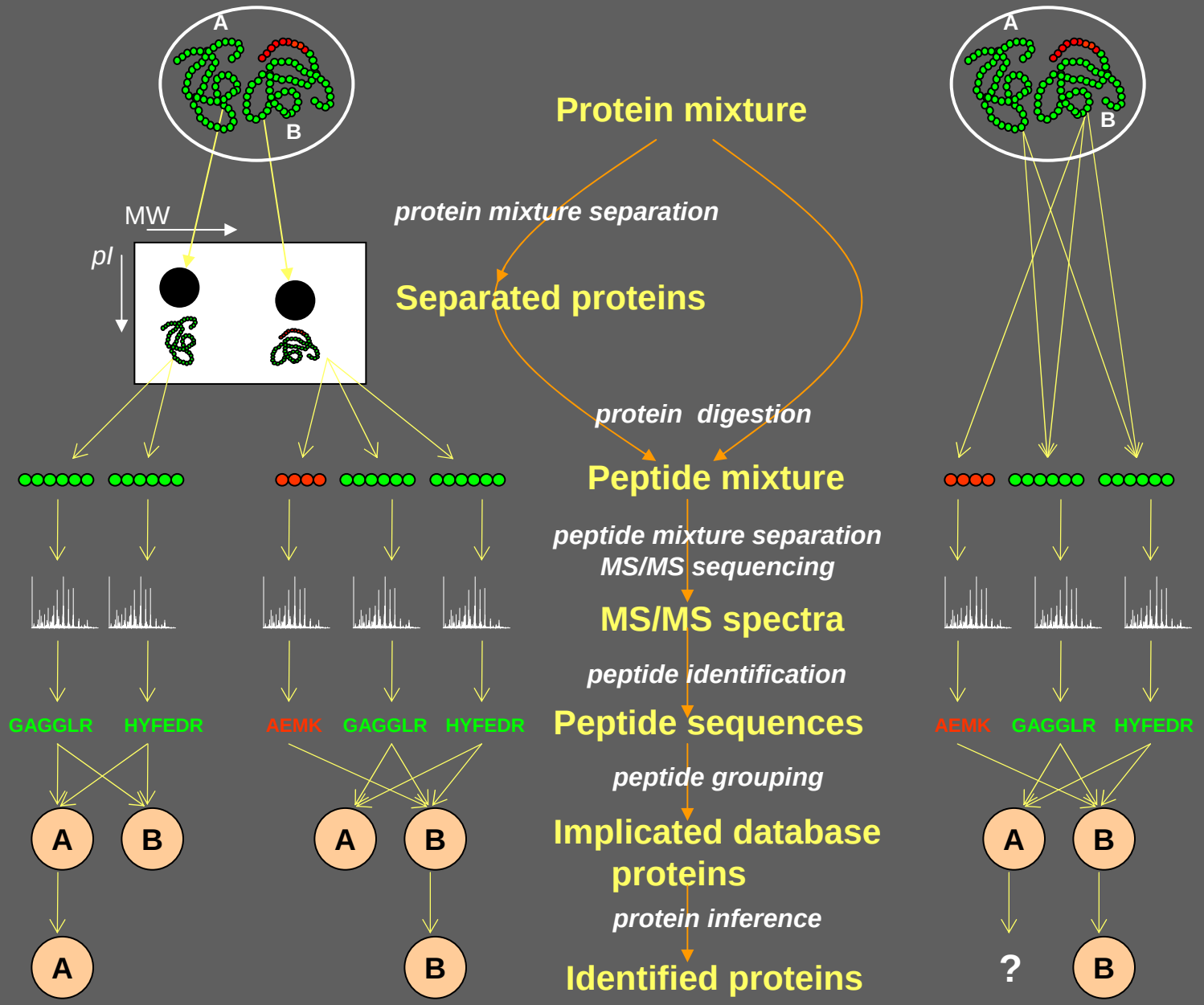
Or both?

In shotgun proteomics, the connectivity between peptides and proteins is lost

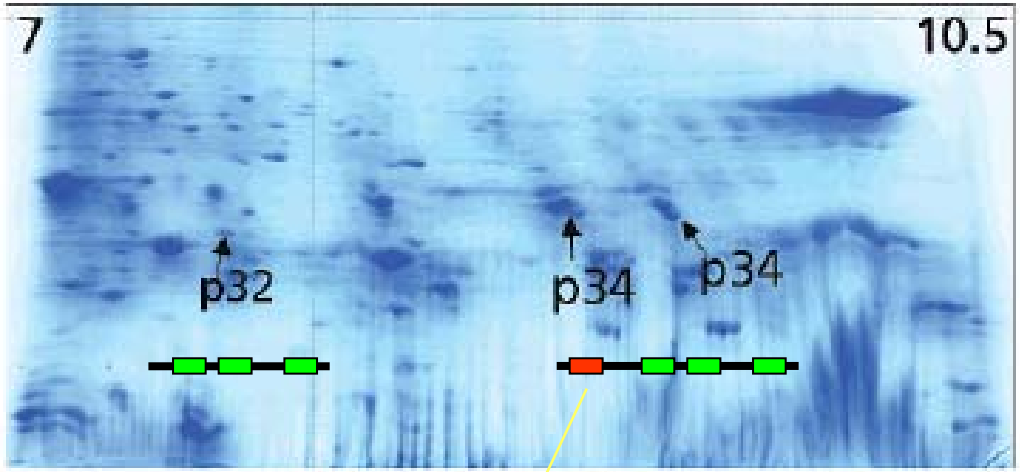
Shared peptides are more prevalent with databases of higher eukaryotes due to the presence of:

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

2D Gel-based Approach vs. Shotgun Approach



Protein Isoforms and 2D-gels



S.K. Pedersen *et al.* J. Proteome Research 2, 303-311 (2003)

NADH-cytochrome b5 reductase

p32: soluble form, Inter-membrane space

p34: membrane bound, outer membrane of the mitochondrion

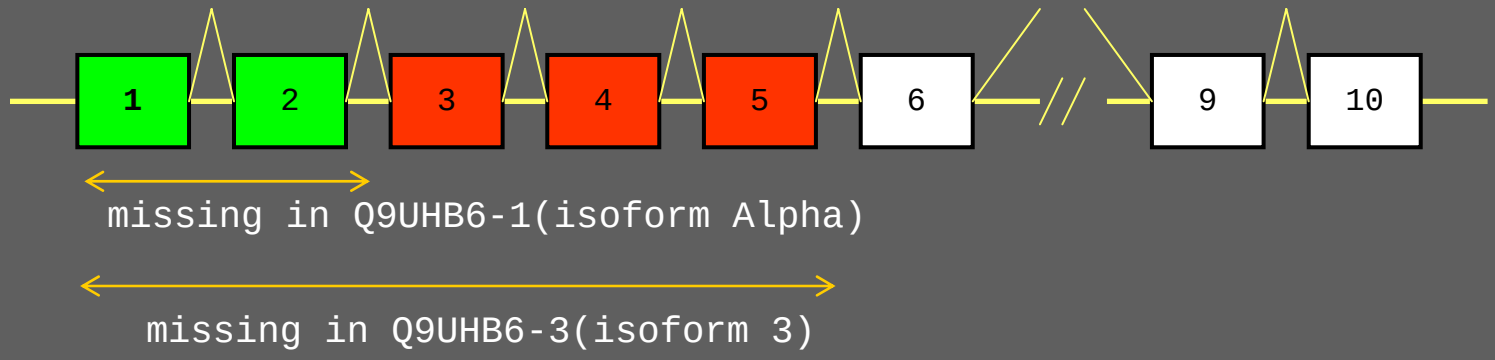
p32/p34 cleavage site

MFSRLSRSHSKALPIALGTVAIAAATAFYFANRNQHSFVFNESNKVFKGDDKWIDLPISK
IEEESHDTRRFTFKLPTEDSEMGLVLASALFAKFVTPKGSNVVRPYTPVSDLSQKGHFQL
VVKHYEGGKMTSHLFGLKPNDTVSFKGPIMKWKWQPNQFKSITLLGAGTGINPLYQLAHH
IVENPNDKTKVNLLYGNKTPQDILLRKELDALKEKYPDKFNVTYFVDDKQDDQDFDGEIS
FISKDFIQEHVPGPKESTHLFVCGPPPFMNAYSGEKKSPKDQGELIGILNNLGYSKDQVF
KF

When purified proteins are digested, the connectivity between peptides and proteins is usually apparent

Challenge for Shotgun Proteomics: Alternative Splice Forms

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



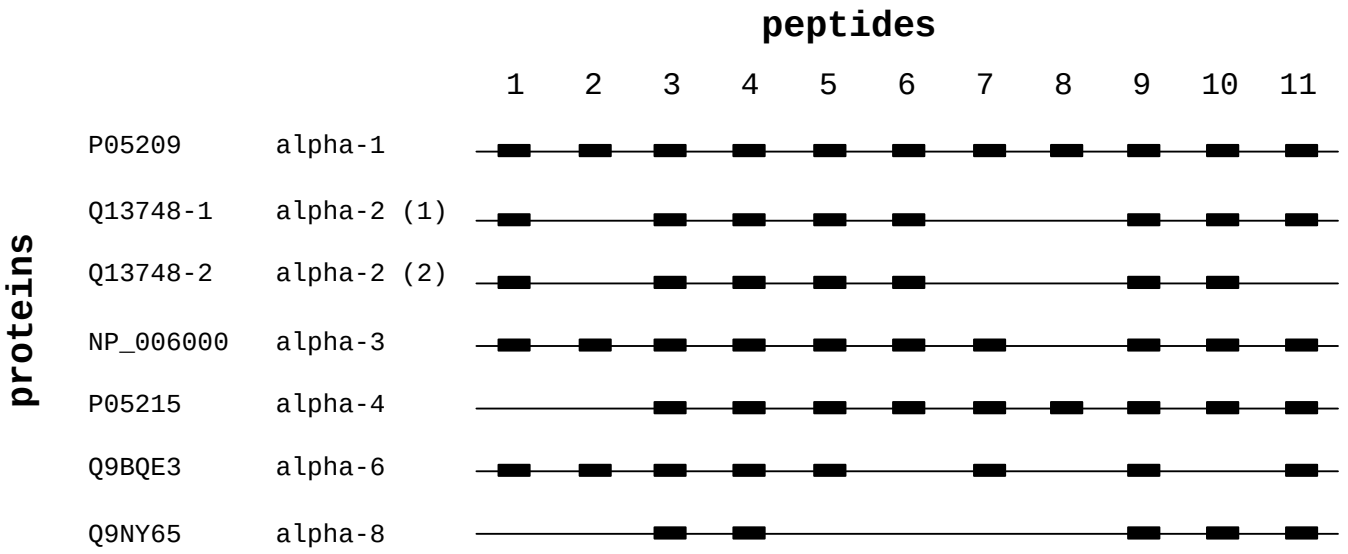
>IPI:IPI00220465 Splice isoform Alpha of Q9UHB6 Epithelial protein lost in neoplasm

MESSPFNRRQWTSLSLRVTAKELSLVNKNKSSAIVEIFSKYQKAAEETNMEKKRSNTENLSQHFRKGTLTVLKKKWENP
GLGAESHTDSLRNSSTEIRHRADHPPAEVTSHAASGAKADQEEQIHPRSRLRSPPEALVQGRYPHIKDGEDLKDHSTES
KKMENCLGESRHEVEK**SEISENTDASGKIEK**KYNVPLNRLKMMFEKGEPTQTKILRAQSRSASGRKISENSYSLDDLEIG
PGQLSSSTFDSEKNESRRNLELPRLSETSIKDRMAKYQAAVSKQSSSTNYTNELKASGGEIKIHKMEQKENVPPGPEVC
ITHQEGEKISANENSLAVRSTPAEDDSRDSQVKSEVQQPVHPKPLSPDSRASSLSESSPPKAMKKFQAPARETCVECQK
TVYPMERLLANQQVFHISCFRCSYCNNKLSLGTYASLHGRIYCKPHFNQLFKSKGNYDEGFGHRPHK**DLWASKNENEEI**
LERPAQLANARETPHSPGVEDAPIAKVGV**LAASMEAKASSQQEK**EDKPAETKKLRIAWPPPTELGSSGSALEEGIKMSK
PKWPPEDEISKPEVPEDVDLDLKKLRRSSSLKER**SRPFTVAASFQSTSVK**SPKTVSPPIRKGWSMSEQSEESVGGRVAE
RKQVENAKASKKNGNVGKTTWQNKESKGETGKRSKEGHSLEMENENLVENGADSDEDDNSFLKQQSPQEPKSLNWSSFVD
NTFAEEFTTQNQKSQDVELWEGEVVKELSVEEQIKR**NRYYDEDEDE**E

Challenge for Shotgun Proteomics: Gene Families (Paralogues)

C **Tubulin-alpha protein family**

1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISAEK	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

MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDK**TIGGGDDSFNTFFSETGAGK**HVPR**AVFVDLEPTVIDEVR**TGTYR**QLFHPEQ**
LITGKEDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVV
EPYNSILTTHTTLEHSDCAFMVDNEAIYDICRR**NLDIERPTYTNLNR**LISQIVSSITASLRFDGALNVDLTEFQTNLVPYPR**IHFPLATYA**
PVISAEKAYHEQLSVAEITNACFEPANQMVKCDPRHGK**YMACCLLYR**GDVVPKDVNAAIATIKTKR**SIQFVDWCPTGFKVGINYQPPTVVP**
GGDLAKVQR**AVCMLSNTTAIAEAWARLDHKFDLMYAK**RAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGVDSVEGEGEEEGEEY

Complication: Database Redundancies

Entrez Protein **Stress-70 protein, mitochondrial precursor**

Sequence entry 1:

gi|21264428|sp|P38646|GR75_HUMAN Stress-70 protein, mitochondrial precursor (75 kDa glucose regulated protein) (GRP 75) (Peptide-binding protein 74) (PBP74) (Mortalin) (MOT)

gi|24234688|ref|NP_004125.3| heat shock 70kDa protein 9B precursor; heat shock 70kD protein 9;
gi|477763|pir||B48127 dnaK-type molecular chaperone precursor, mitochondrial - human

Sequence entry 2: *G → R at position 540*

gi|292059|gb|AAA67526.1| MTHSP75
Human mRNA L15189

Sequence entry 3: *Q → R at position 74; H → R at position 184*

gi|12653415|gb|AAH00478.1| Heat shock 70kDa protein 9B, precursor [Homo sapiens].
Human mRNA BC000478
Human mRNA BC024034

gi|18645123|gb|AAH24034.1| Heat shock 70kDa protein 9B, precursor [Homo sapiens].

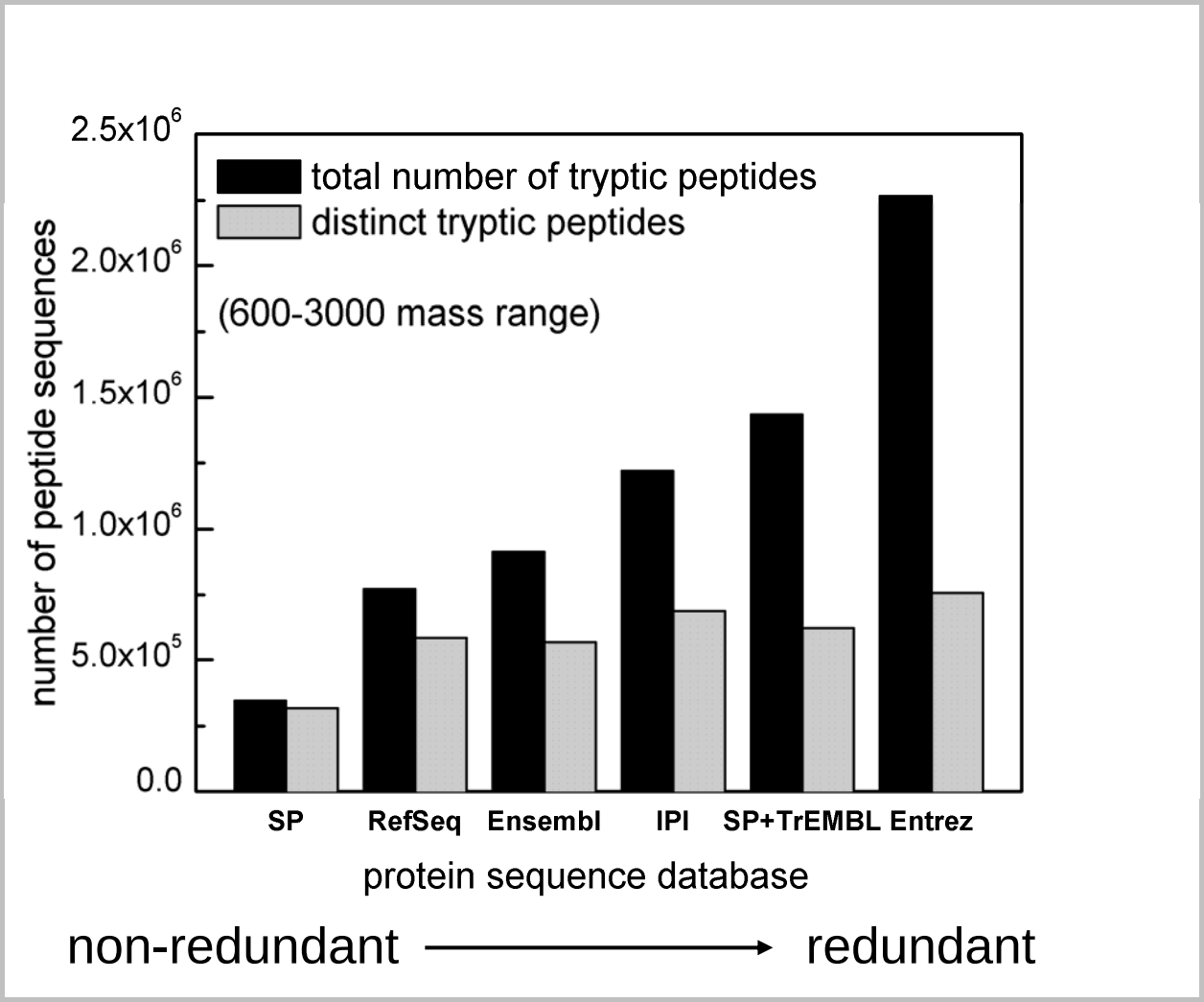
Sequence entry 4: *M → V at N terminus*

gi|21040386|gb|AAH30634.1| heat shock 70kD protein 9B (mortalin-2) [Homo sapiens].

Human mRNA BC030634

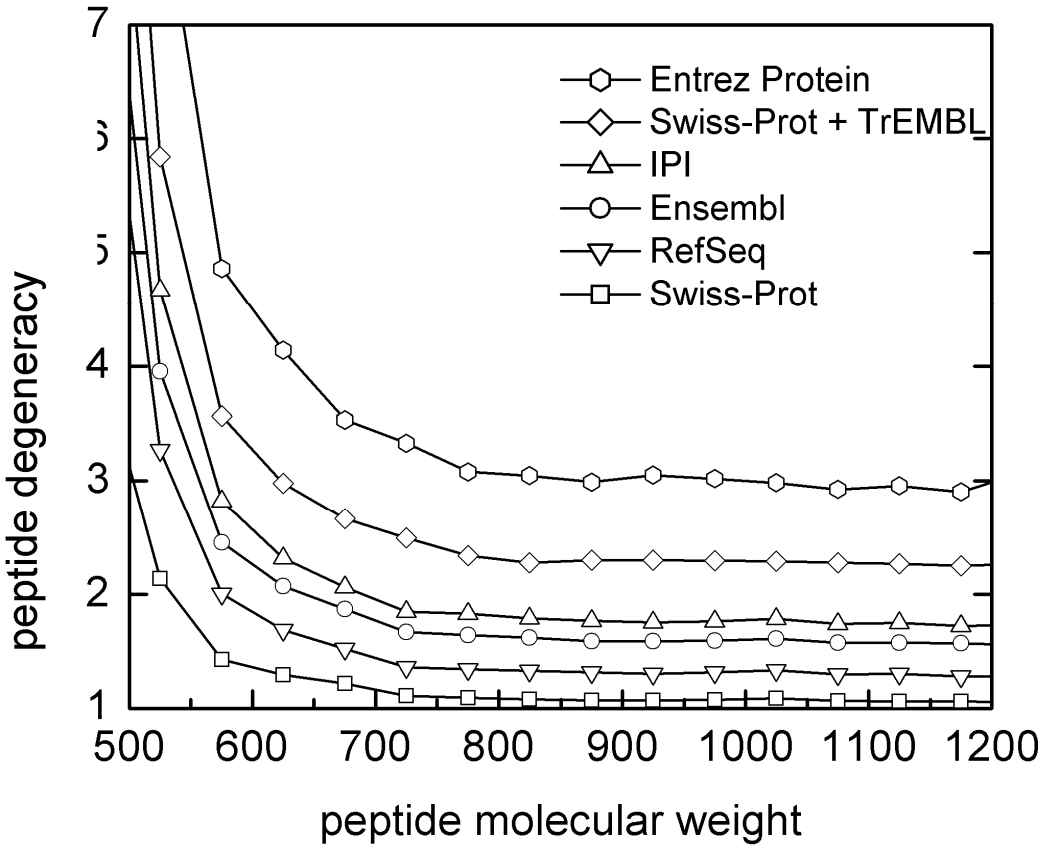
redundant database entries

Protein Sequence Databases



Protein Sequence Databases

Average number of database entries containing each peptide



choice of the database depends on the goals of the experiment

e.g. is ability to identify polymorphisms Important?

Mature Protein Forms

A

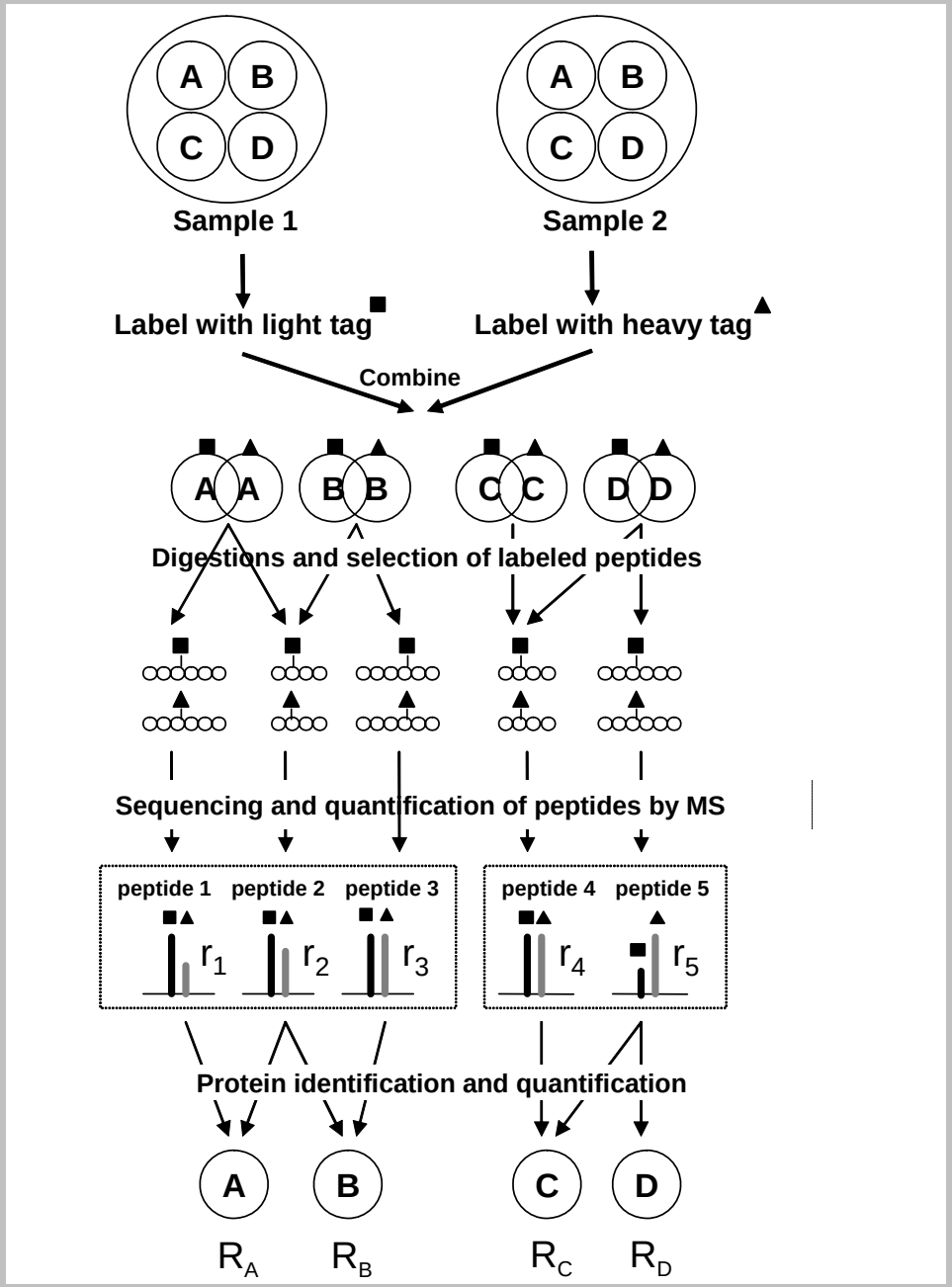
>IPI00019906.1|UniProt/Swiss-Prot:P35613 Basigin precursor

signal peptide **N-terminal peptide**

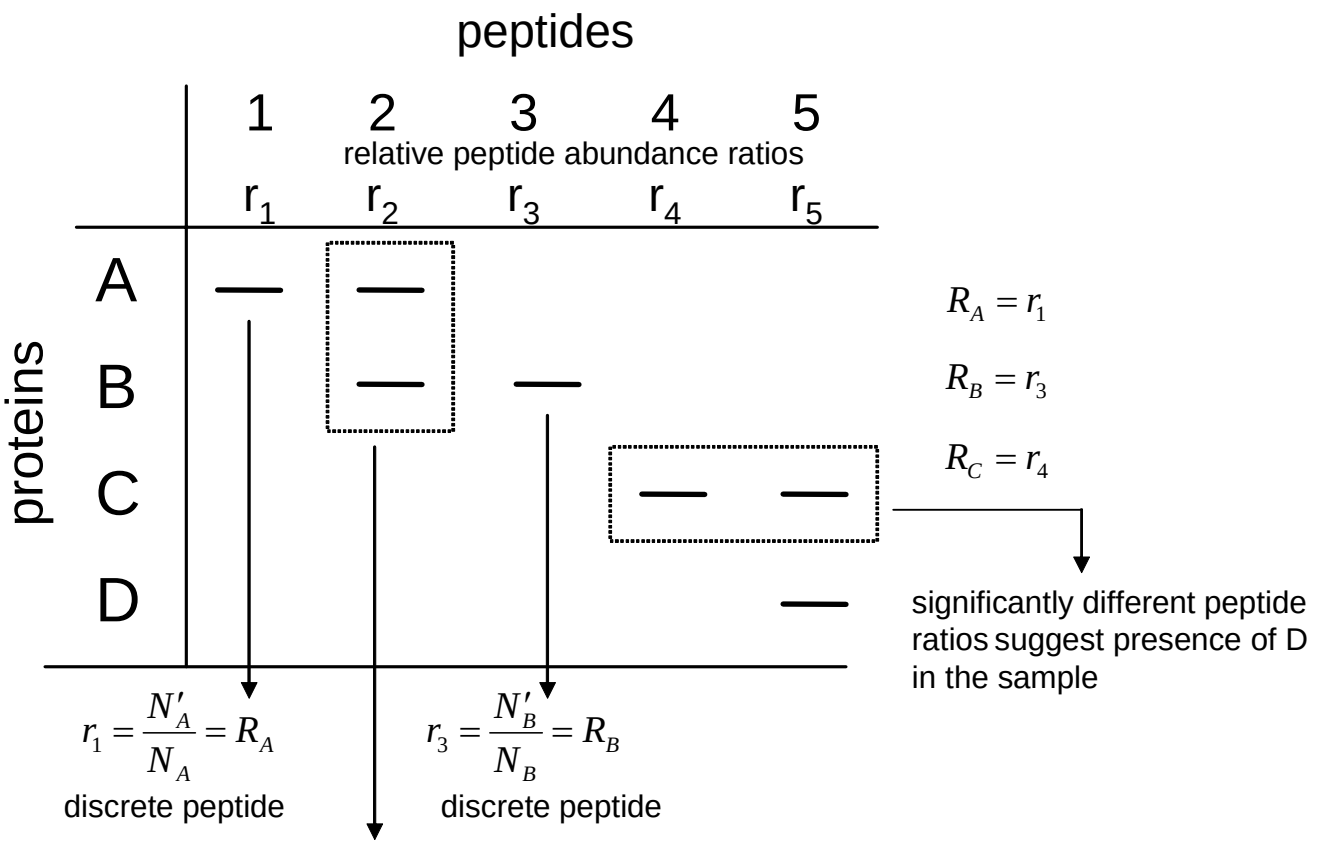
MAAALFVLLGFALLGTHGASG**AAGTVFTTVEDLGSK**ILLTCSLNDSATEVTGHR**WLKGGVVLKEDALPGQK**TEFKVDSDDQWGEYSCVFL
PEPMGTANIQLHGPPRVKAVK**SSEHINEGETAMLVCK**SESVPPVTDWAWYKITDSEDKALMNGSESRFFVSSSQGR**SELHIENLNMEADP**
GQYRCNGTSSKGSDQAIITLRVRSHLAALWPFLGIVAEVLVLVTIIFIYEKR**RKPEDVLDDDDAGSAPLK**SSGQHQNDKGKNVRQRNSS

signal peptide removed

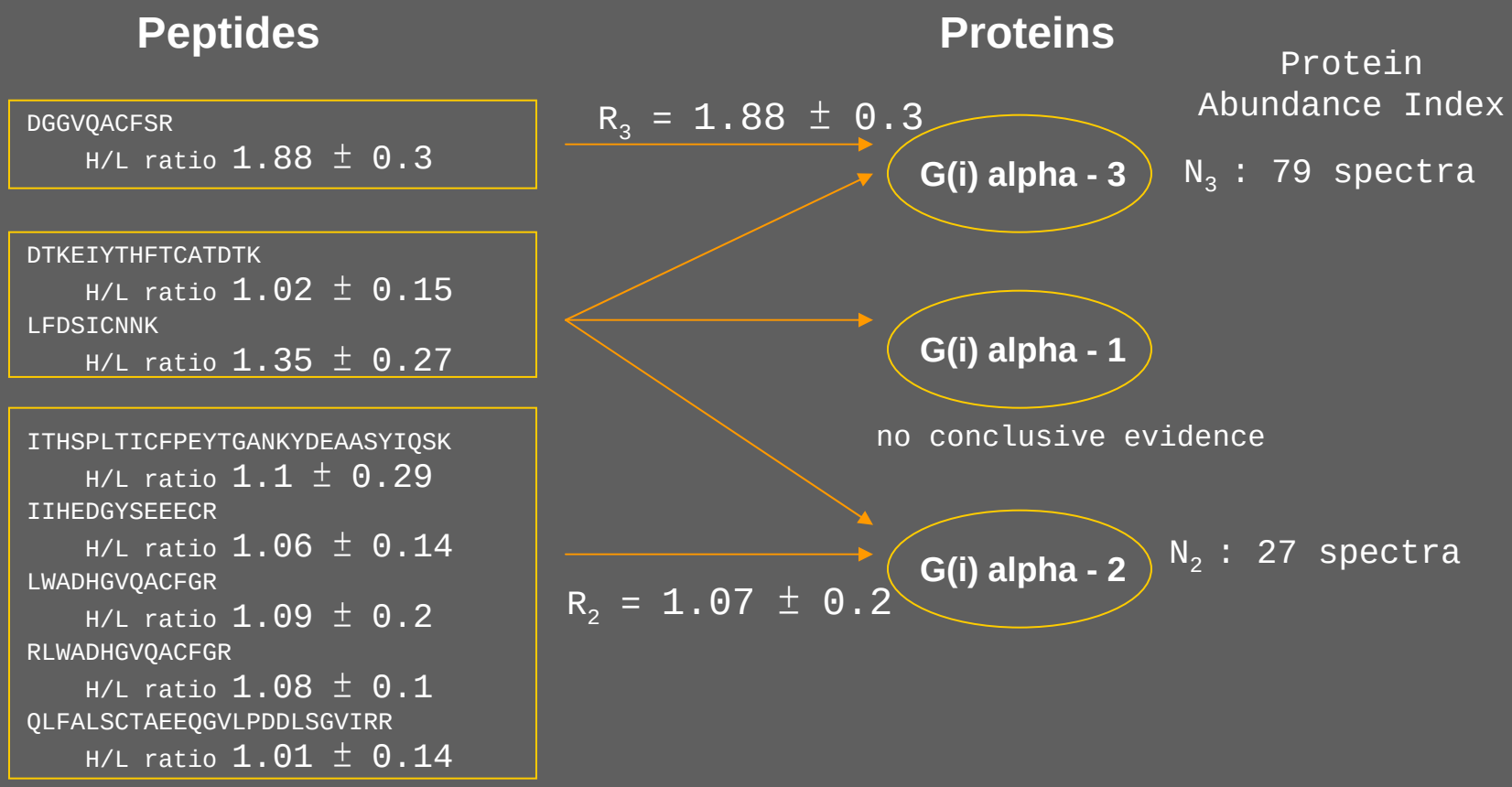
Quantitative Proteomics



Quantitative Proteomics

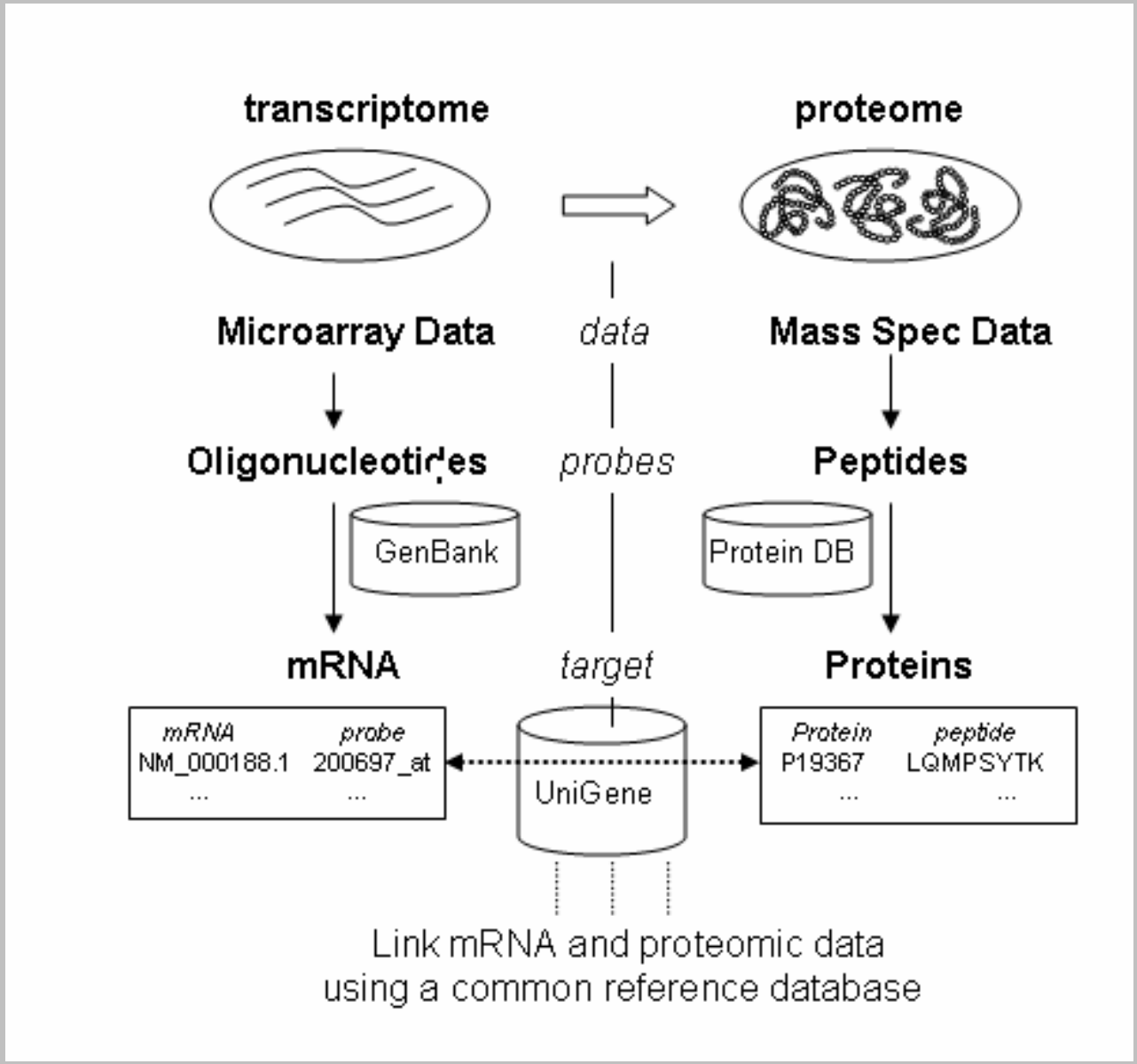


Quantitative Proteomics

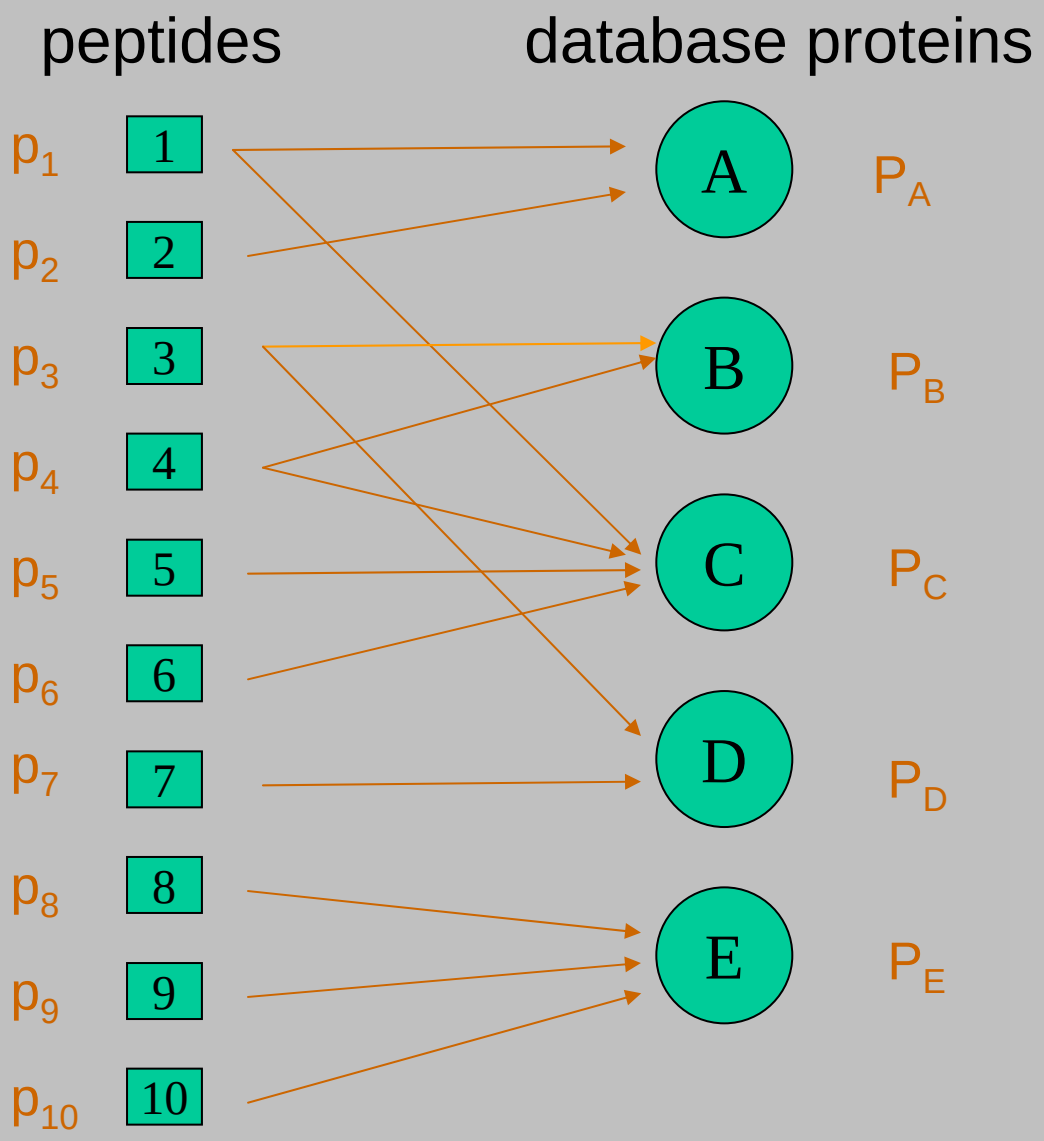


Quantification and identification are interdependent problems

Connecting Protein and mRNA Measurements



Protein Inference with ProteinProphet

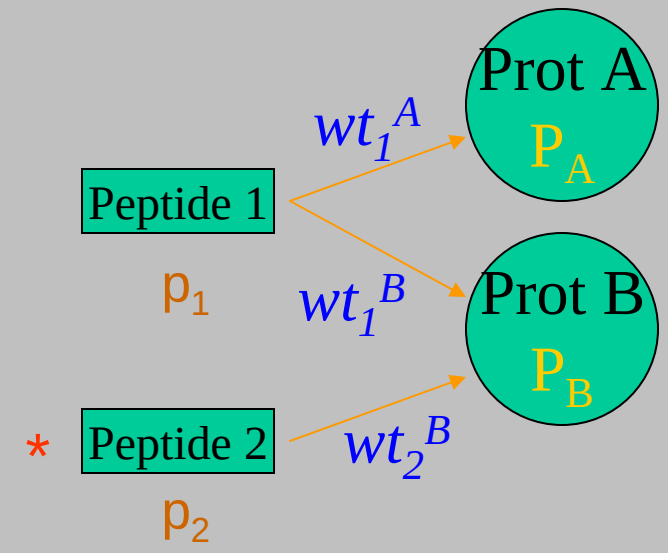


PeptideProphet computational tool:

Probability-based model aims at deriving minimum list of proteins that can explain all observed peptides

Resolves ambiguous cases (when possible) and presents results in a convenient “biologist-friendly”format

Apportionment of Shared Peptides



* independent evidence

EM

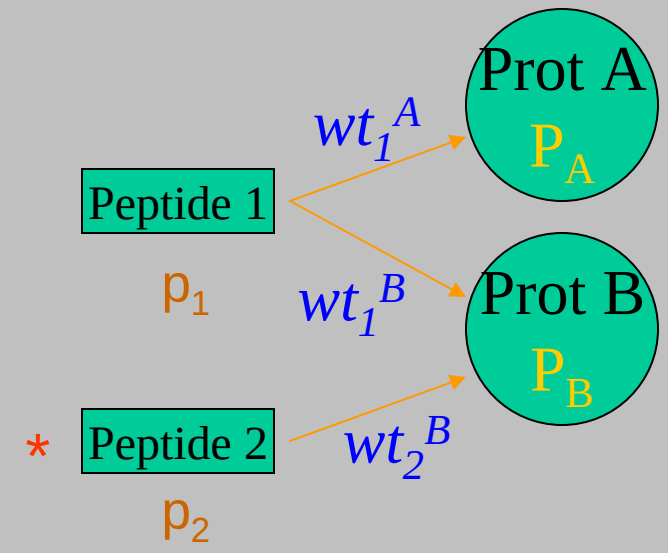
$$wt_i^n = \frac{P_n}{\sum_{j=1\ldots N} P_j}$$

$$P_n = 1 - \prod_i (1 - wt_i^n p_i)$$

For each peptide, weights sum to 1

Model aims at deriving the simplest list of proteins sufficient to explain the observed peptides (Occam's razor)

Apportionment of Shared Peptides



* independent evidence

Initialize:

$P_A = P_B$

$wt_1^A = wt_1^B = 0.5$

$wt_2^B = 1$

Run EM:

$wt_1^A = 0$

$P_A = 0$

$wt_1^A = 1$

$P_B = 1 - (1-0.8)(1-0.7) = 0.94$

Protein B is present in the sample

Protein Group: Example 1

proteins	peptides 1	2	3	4
A	-	-		
B			-	-

distinct proteins

ProteinProphet output

#1 Prot A P=1
*wt=1 peptide 1
*wt=1 peptide 2

#2 Prot B P=1
*wt=1 peptide 3
*wt=1 peptide 4

*: no other sequence database entry has this peptide (wt=1)

Protein Group: Example 2

proteins	peptides 1	2	3
A	-	-	-
B	-	-	-

indistinguishable proteins

ProteinProphet output

#1	Prot A	Prot B	P=1
	*wt=1	peptide 1	
	*wt=1	peptide 2	
	*wt=1	peptide 3	

*: no other sequence database entry has this peptide (wt=1)

Indistinguishable proteins

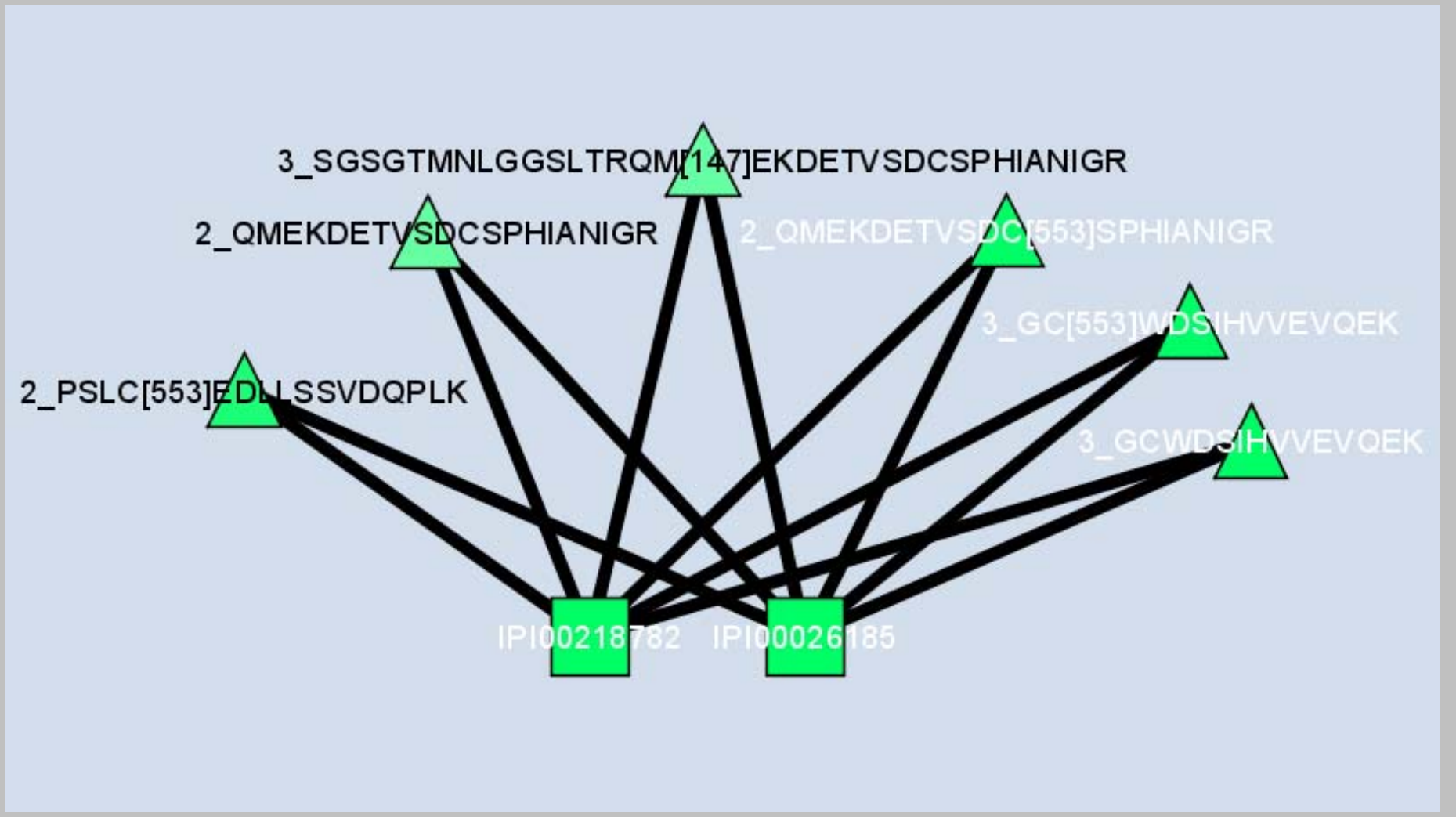


Illustration created using Cytoscape (Luis Mendoza)

Protein Group: Example 3

protein	peptides 1	2	3
A	-	-	-
B		-	-

B is a subset protein

Occam's razor: Prot A is present in the sample. No conclusive evidence for the presence of Prot B.

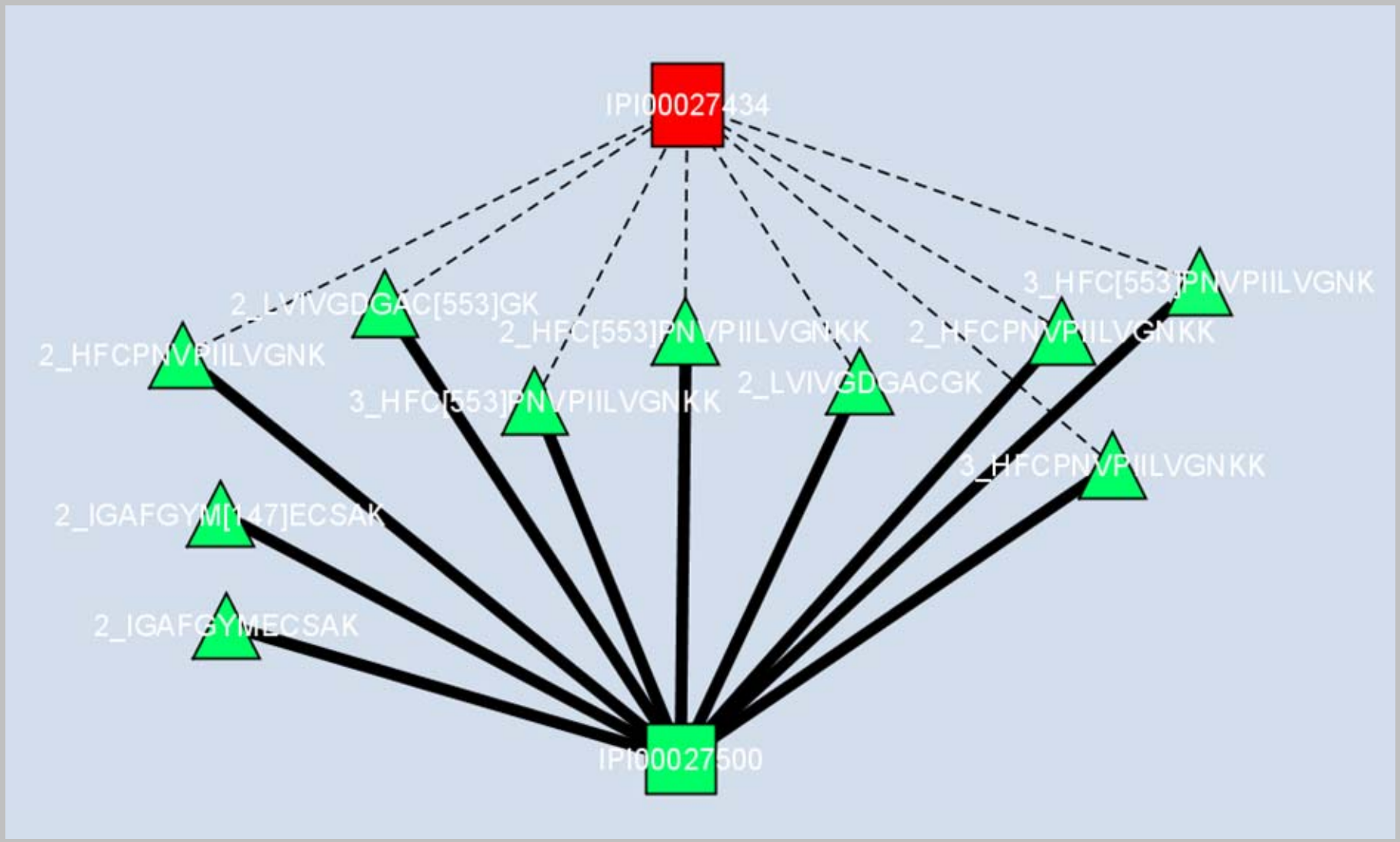
ProteinProphet output

#1 Prot A P=1
*wt=1 peptide1
wt=1 peptide 2
wt=1 peptide 3

(end of the list)

#2 Prot B P=0
wt=0 peptide 2
wt=0 peptide 3

Subset proteins



Protein Group: Example 4

proteins	peptides 1	2	3	4
A	-	-	-	
B		-	-	-

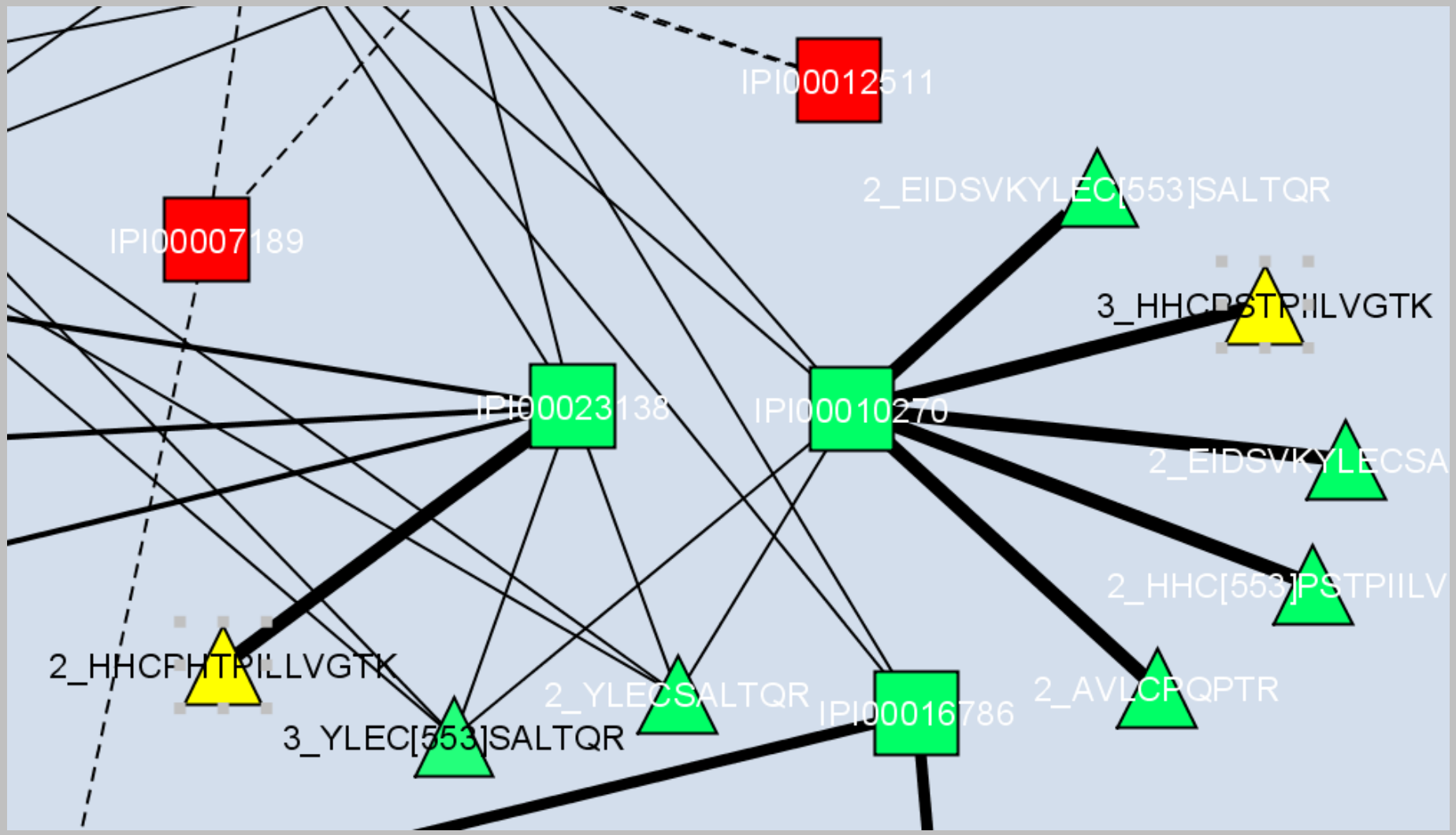
differentiable proteins

ProteinProphet output

#1 Prot A P=1
- *wt=1 peptide 1
- wt=0.5 peptide 2
- wt=0.5 peptide 3

#2 Prot B P=1
- wt=0.5 peptide 2
- wt=0.5 peptide 3
- *wt=1 peptide 4

Differentiable proteins



Protein Group: Example 5

proteins	peptides 1	2	3	4
A	-	-		
B		-	-	
C			-	-

B: subsumable protein

ProteinProphet output

#1 Prot A P=1
*wt=1 peptide 1
wt=1 peptide 2

#2 Prot C P=1
wt=1 peptide 3
*wt=1 peptide 4

(end of the list)

#3 Prot B P=0
wt=0 peptide 2
wt=0 peptide 3

Protein Group: Example 6

proteins	peptides 1	2	3
A	-	-	-
B	-		-
C		-	-

special case:

A has all three peptides (most likely protein). However, B and C combined can account for all the observed peptides as well

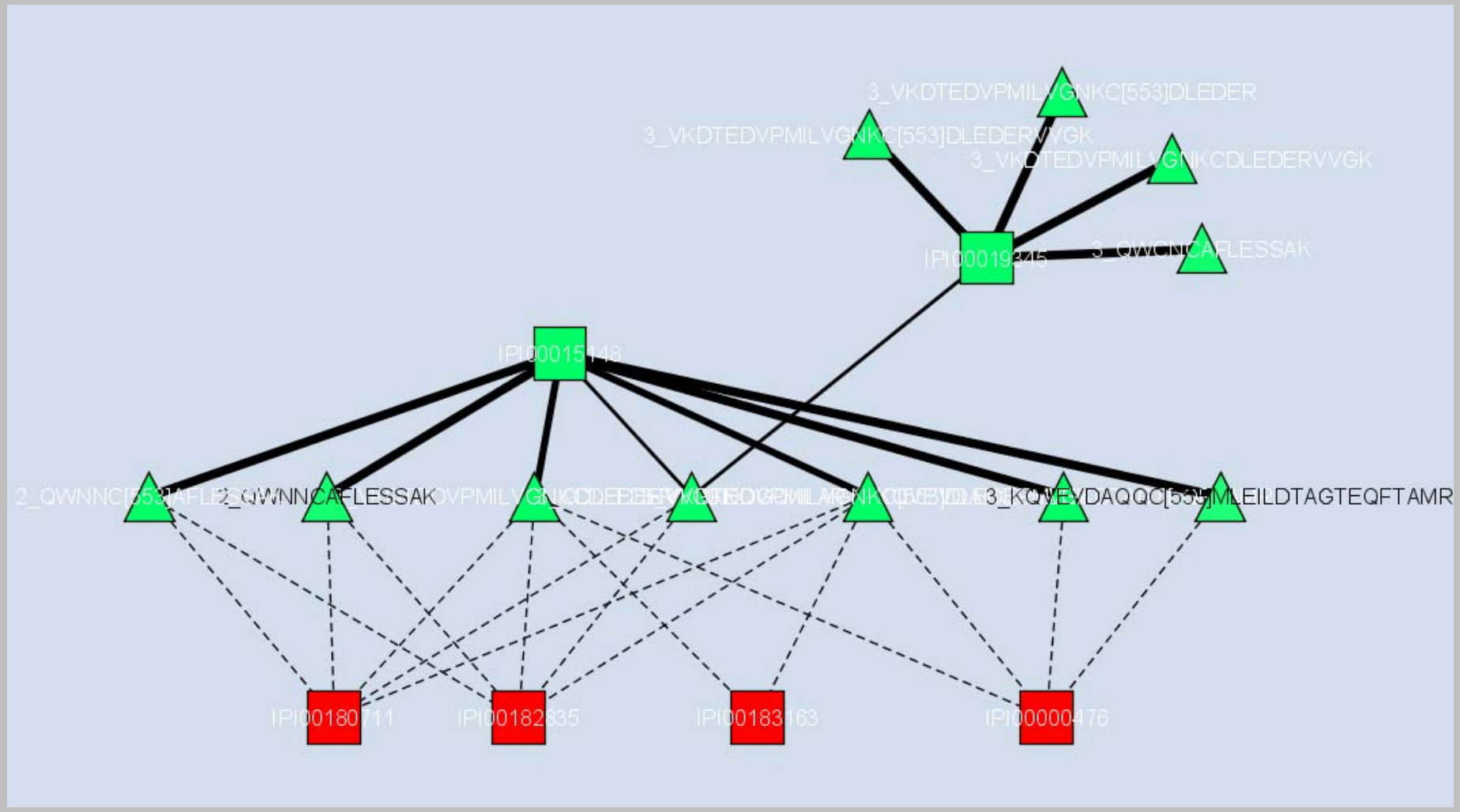
ProteinProphet output

#1 Protein group P=1
a Prot A *P=1*
wt=1 peptide 1
wt=1 peptide 2
wt=1 peptide 3

b Prot B *P=0*
wt=0 peptide 1
wt=0 peptide 3

c Prot C *P=0*
wt=0 peptide 2
wt=0 peptide 3

Protein group



ProteinProphet (full analysis)

- Retrieves ***all*** proteins corresponding to each assigned peptide from search database
- Groups peptides by protein and computes a probability that a protein is present based upon the evidence of corresponding peptides in the dataset
- Adjusts peptide probabilities for grouping information (penalizes ‘single hits’)
- Apportions shared peptides, those corresponding to more than a single protein in the database, among ***all*** corresponding proteins
- Collapses redundant and indistinguishable protein database entries into one identification

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

Publishing Large-Scale Datasets

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

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

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

Publishing Large-Scale Datasets

eters used to date (18, 23). Fig. 5*A* shows one such MS/MS scan, along with the search results for the given peptide sequence from both the ICAT 1 and ICAT 2 datasets. While the peptide sequence was only partially tryptic, the biology of MIF was in keeping with the biology of the experiment performed and made acceptance of the identification tempting. However, when PeptideProphet™ interpreted the data, it reported low values for p_{comp}, indicating that MIF was not identified. However, as shown in Fig. 5*B*, when the same data were searched against a nonredundant database, the same MS/MS scan (better) matched a peptide sequence for the bovine heterotrimeric Gγ2 protein. The human homologue of this gene, though known, was not in the human database searched for some reason, resulting in the errant MIF identification. Because the human and bovine Gγ2 amino acid sequences are conserved for the region spanning the assigned peptide, when the human Gγ2 sequence was added to the human database and the data re-searched, the p_{comp} values for this new match were now very high (*i.e.* most likely correct). Indeed as discussed elsewhere,[2] heterotrimeric G proteins are highly abundant in lipid rafts (the source of our initial sample), thus this result was also in keeping with the biology.

Remarkably, as can be seen in Fig. 5, many of the MS/MS fragment ion peaks matched potential fragment ions from both peptide sequences (peaks in *bold*), eight of which were common to both sequences (indicated with *asterisks*). Thus, using SEQUEST™ alone, even an experienced user could be forgiven for assuming MIF to be the correct identification, whereas PeptideProphet™ yielded an unequivocal result. Indeed, immunoblotting confirmed that MIF was not present in the original samples (data not shown). This being said, even if statistical tools such as PeptideProphet™ are used, it is still likely that some incorrect identifications will result from searching incomplete databases.

Statistical Analysis and Validation of Protein Identifications—When interpreting proteomic MS/MS data, there are two related but entirely separate steps to the process of identifying the protein(s) in the original sample. The first, as has been discussed above, is assigning individual MS/MS spectra to peptide sequences in a database. For this purpose, PeptideProphet™ was developed to calculate a level of con-

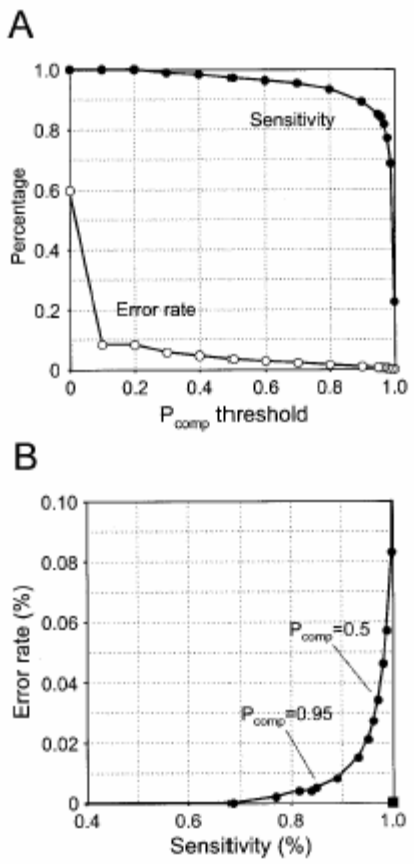
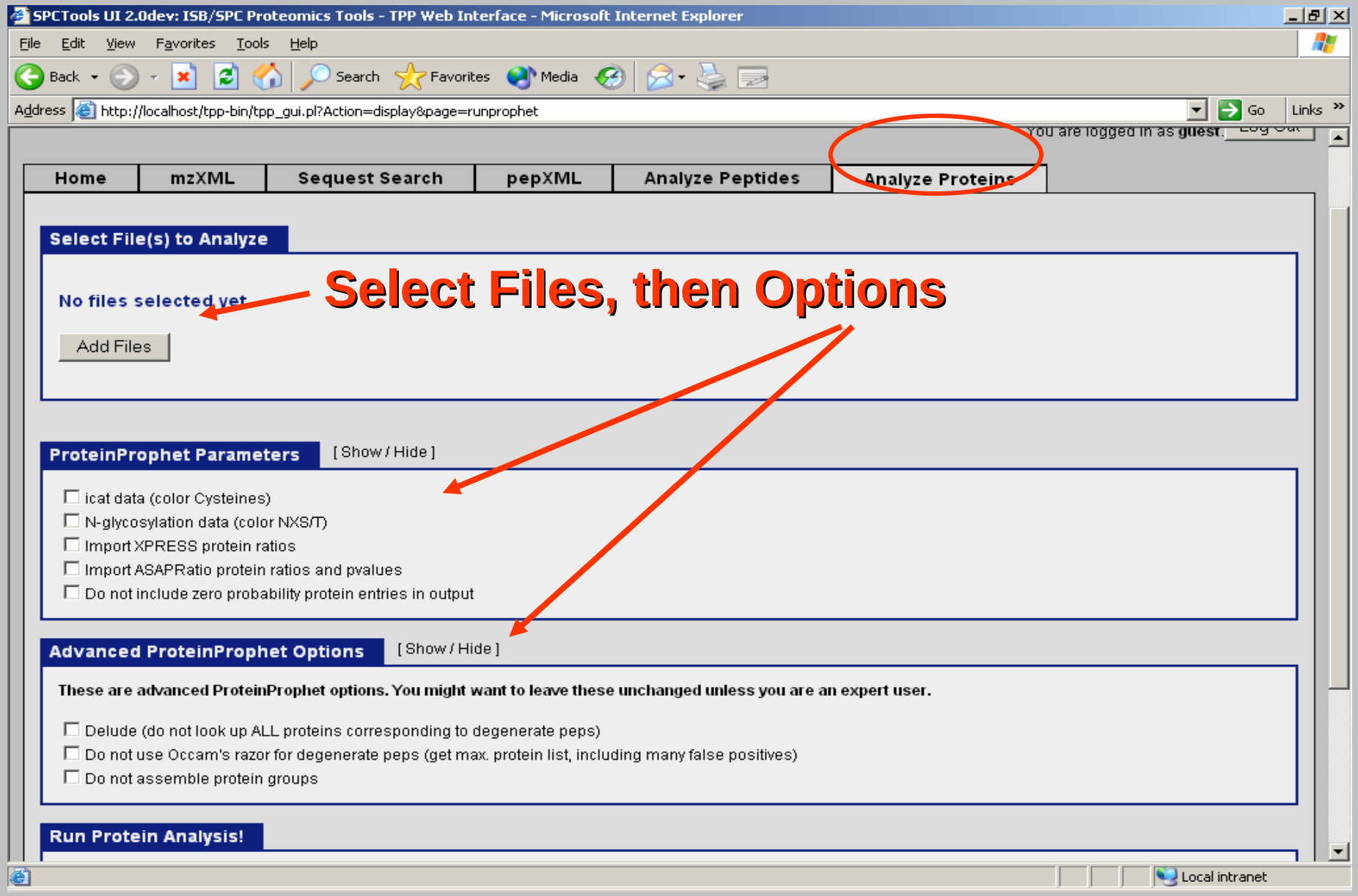
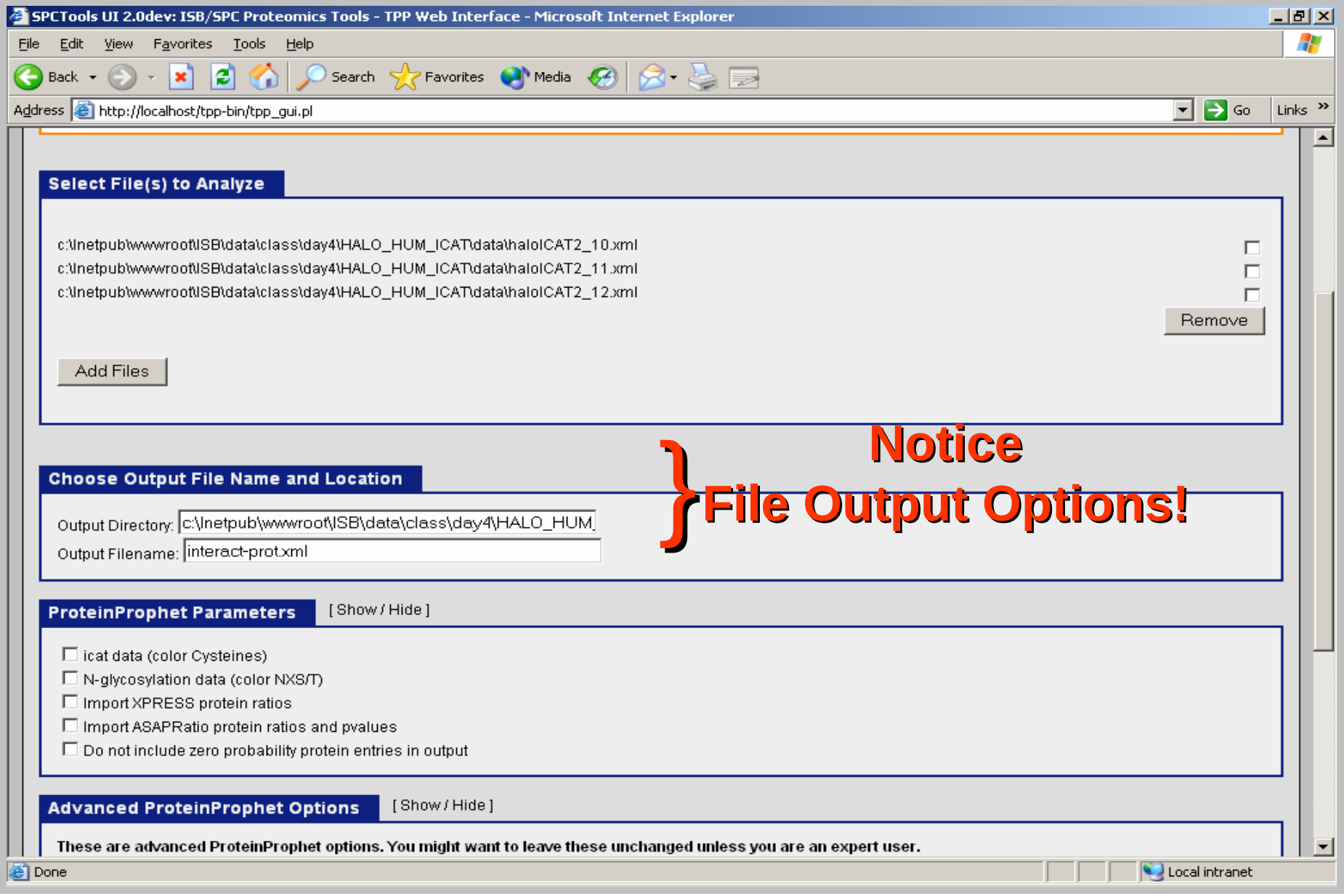


FIG. 6. Protein identification error and sensitivity rates determined by ProteinProphet™. *A*, Error rates (*open*, percentage of total identifications which are false) and sensitivity (*filled*, percentage of total correct identifications remaining after restricting data at a given P_{comp} threshold) as a function of minimum protein P_{comp} threshold set to restrict the data (a compilation of the ICAT 1, ICAT 2, and Flow-through 1 datasets). *B*, Error rate *versus* sensitivity plot for the same

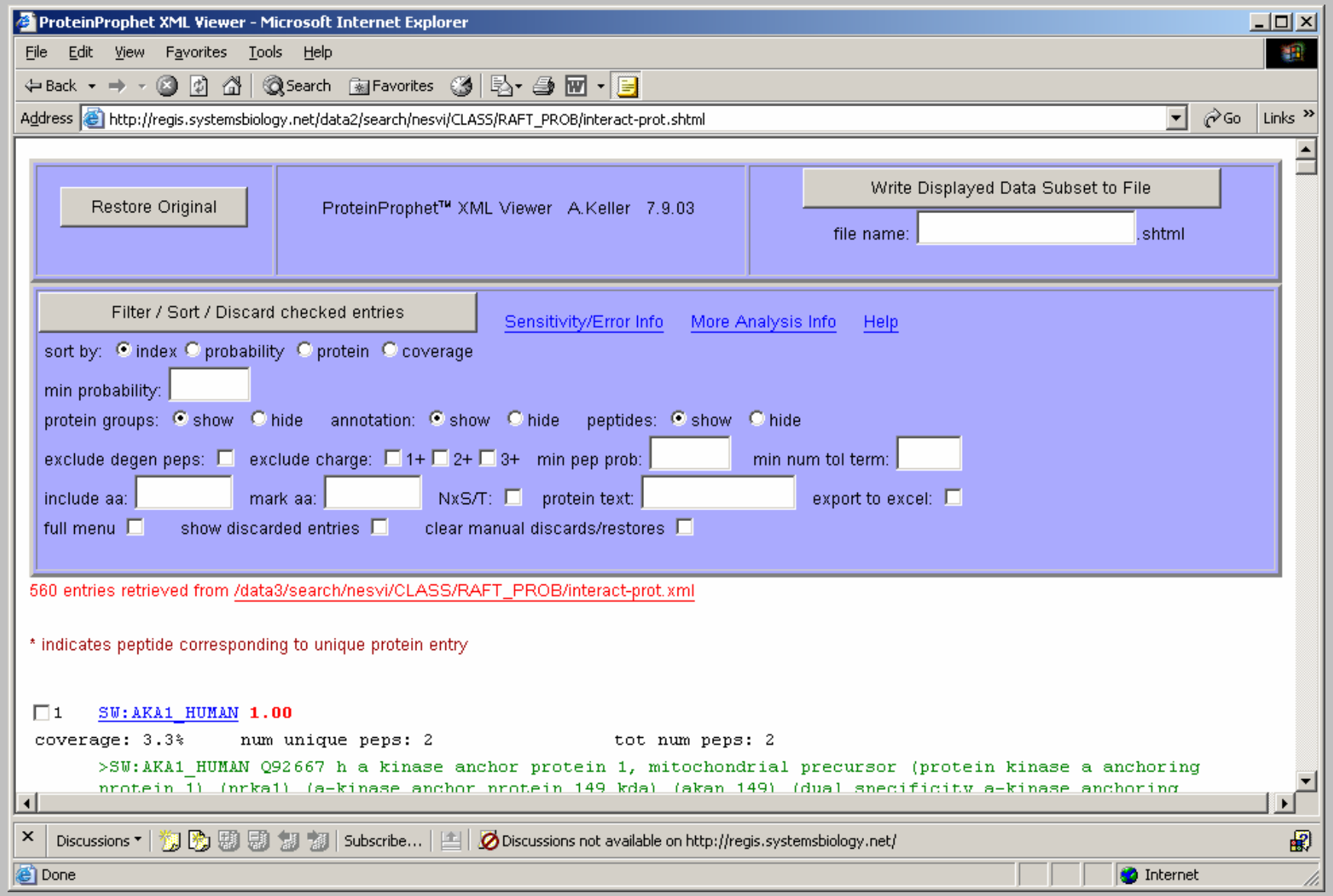
Running ProteinProphet



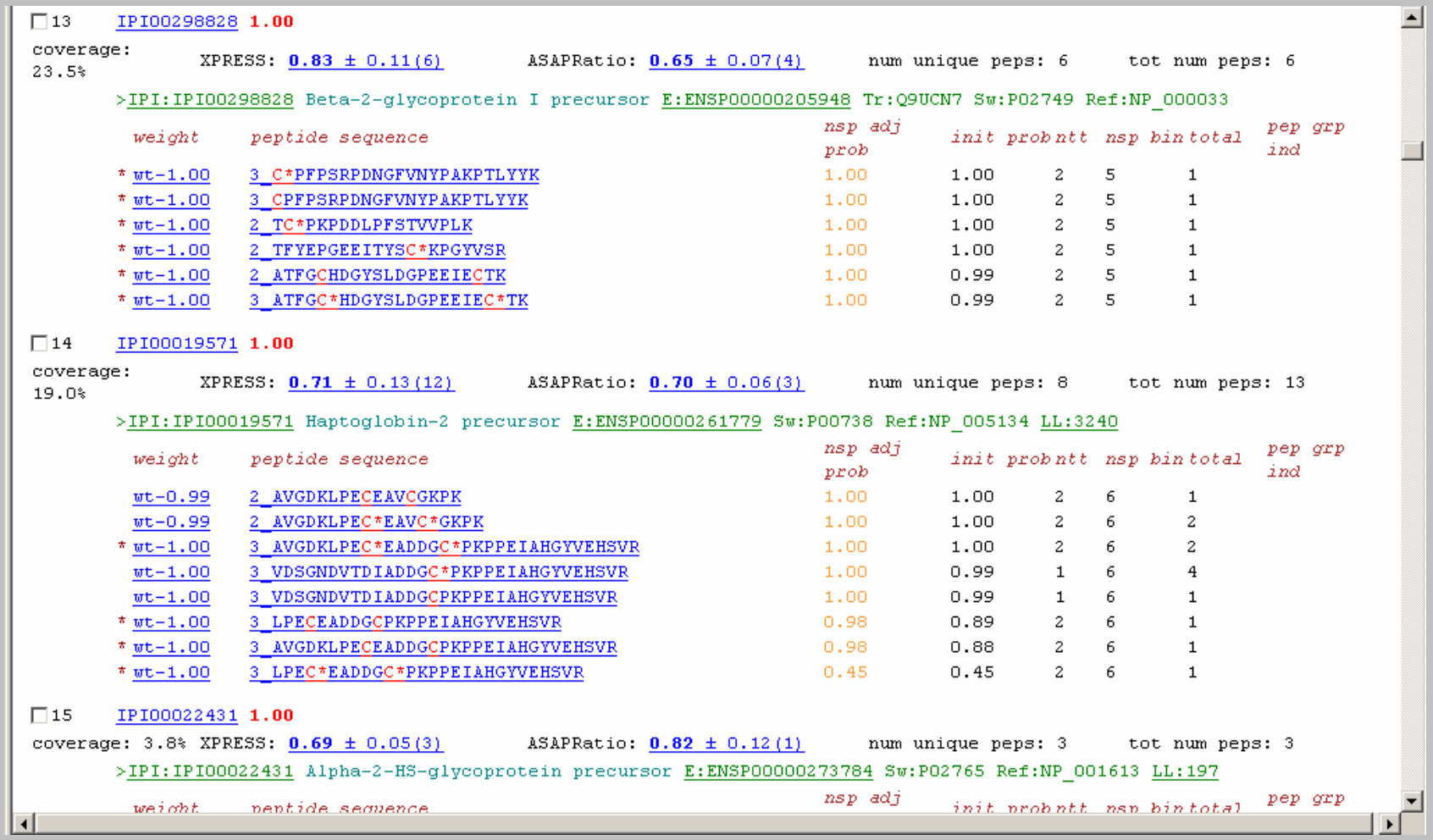
Running ProteinProphet (after file select)



ProteinProphet Output



ProteinProphet Output



13 IPI00298828 1.00

coverage: 23.5% XPRESS: 0.83 ± 0.11(6) ASAPRatio: 0.65 ± 0.07(4) num unique peps: 6 tot num peps: 6

>IPI:IPI00298828 Beta-2-glycoprotein I precursor E:ENSP00000205948 Tr:Q9UCN7 Sw:P02749 Ref:NP_000033

	weight	peptide sequence	nsp adj prob	init prob	ntt	nsp bin	total	pep grp ind
*	wt-1.00	3_C*PFPSRPDNGFVNYPAKPTLYYK	1.00	1.00	2	5	1	
*	wt-1.00	3_CPFPSRPDNGFVNYPAKPTLYYK	1.00	1.00	2	5	1	
*	wt-1.00	2_TC*PKPDDLPFSTVVPLK	1.00	1.00	2	5	1	
*	wt-1.00	2_TFYEPGEEITYSC*KPGYVSR	1.00	1.00	2	5	1	
*	wt-1.00	2_ATFGCHDGYSLDGPEEIECTK	1.00	0.99	2	5	1	
*	wt-1.00	3_ATFGC*HDGYSLDGPEEIEC*TK	1.00	0.99	2	5	1	

14 IPI00019571 1.00

coverage: 19.0% XPRESS: 0.71 ± 0.13(12) ASAPRatio: 0.70 ± 0.06(3) num unique peps: 8 tot num peps: 13

>IPI:IPI00019571 Haptoglobin-2 precursor E:ENSP00000261779 Sw:P00738 Ref:NP_005134 LL:3240

	weight	peptide sequence	nsp adj prob	init prob	ntt	nsp bin	total	pep grp ind
	wt-0.99	2_AVGDKLPECEAVCGKPK	1.00	1.00	2	6	1	
	wt-0.99	2_AVGDKLPEC*EAVC*GKPK	1.00	1.00	2	6	2	
*	wt-1.00	3_AVGDKLPEC*EADDGC*PKPPEIAHGYVEHSVR	1.00	1.00	2	6	2	
	wt-1.00	3_VDSGNDVTDIADDGC*PKPPEIAHGYVEHSVR	1.00	0.99	1	6	4	
	wt-1.00	3_VDSGNDVTDIADDGCPKPPEIAHGYVEHSVR	1.00	0.99	1	6	1	
*	wt-1.00	3_LPECEADDGCPKPPEIAHGYVEHSVR	0.98	0.89	2	6	1	
*	wt-1.00	3_AVGDKLPECEADDGCPKPPEIAHGYVEHSVR	0.98	0.88	2	6	1	
*	wt-1.00	3_LPEC*EADDGC*PKPPEIAHGYVEHSVR	0.45	0.45	2	6	1	

15 IPI00022431 1.00

coverage: 3.8% XPRESS: 0.69 ± 0.05(3) ASAPRatio: 0.82 ± 0.12(1) num unique peps: 3 tot num peps: 3

>IPI:IPI00022431 Alpha-2-HS-glycoprotein precursor E:ENSP00000273784 Sw:P02765 Ref:NP_001613 LL:197

weight	peptide sequence	nsp adj	init prob	ntt	nsp bin	total	pep grp

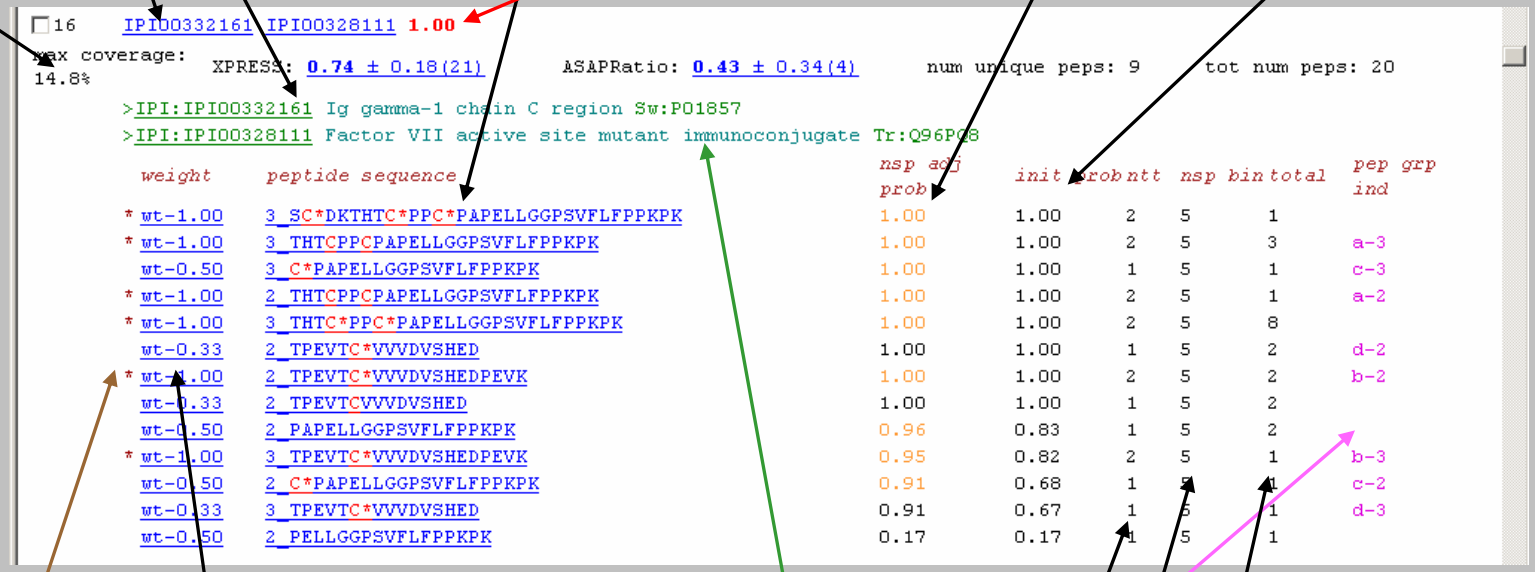
ProteinProphet Output

protein name(s)

protein probability

charge and peptide

adj pep probs/orig pep probs



16 IPI00332161 IPI00328111 **1.00**

max coverage: 14.8% XPRESS: **0.74** ± 0.18(21) ASAPRatio: **0.43** ± 0.34(4) num unique peps: 9 tot num peps: 20

>IPI:IPI00332161 Ig gamma-1 chain C region Sw:P01857

>IPI:IPI00328111 Factor VII active site mutant immunoconjugate Tr:Q96PQ8

	weight	peptide sequence	nsp adj prob	init prob	ntt	nsp bin	total	pep grp ind
*	wt-1.00	3_SC*DKTHTC*PPC*PAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	1	
*	wt-1.00	3_THTCPPCPAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	3	a-3
	wt-0.50	3_C*PAPELLGGPSVFLFPPKPK	1.00	1.00	1	5	1	c-3
*	wt-1.00	2_THTCPPCPAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	1	a-2
*	wt-1.00	3_THTC*PPC*PAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	8	
	wt-0.33	2_TPEVTC*VVVDVSHED	1.00	1.00	1	5	2	d-2
*	wt-1.00	2_TPEVTC*VVVDVSHEDPEVK	1.00	1.00	2	5	2	b-2
	wt-0.33	2_TPEVTCVVVDVSHED	1.00	1.00	1	5	2	
	wt-0.50	2_PAPELLGGPSVFLFPPKPK	0.96	0.83	1	5	2	
*	wt-1.00	3_TPEVTC*VVVDVSHEDPEVK	0.95	0.82	2	5	1	b-3
	wt-0.50	2_C*PAPELLGGPSVFLFPPKPK	0.91	0.68	1	5	1	c-2
	wt-0.33	3_TPEVTC*VVVDVSHED	0.91	0.67	1	5	1	d-3
	wt-0.50	2_PELLGGPSVFLFPPKPK	0.17	0.17	1	5	1	

shared peptide weight

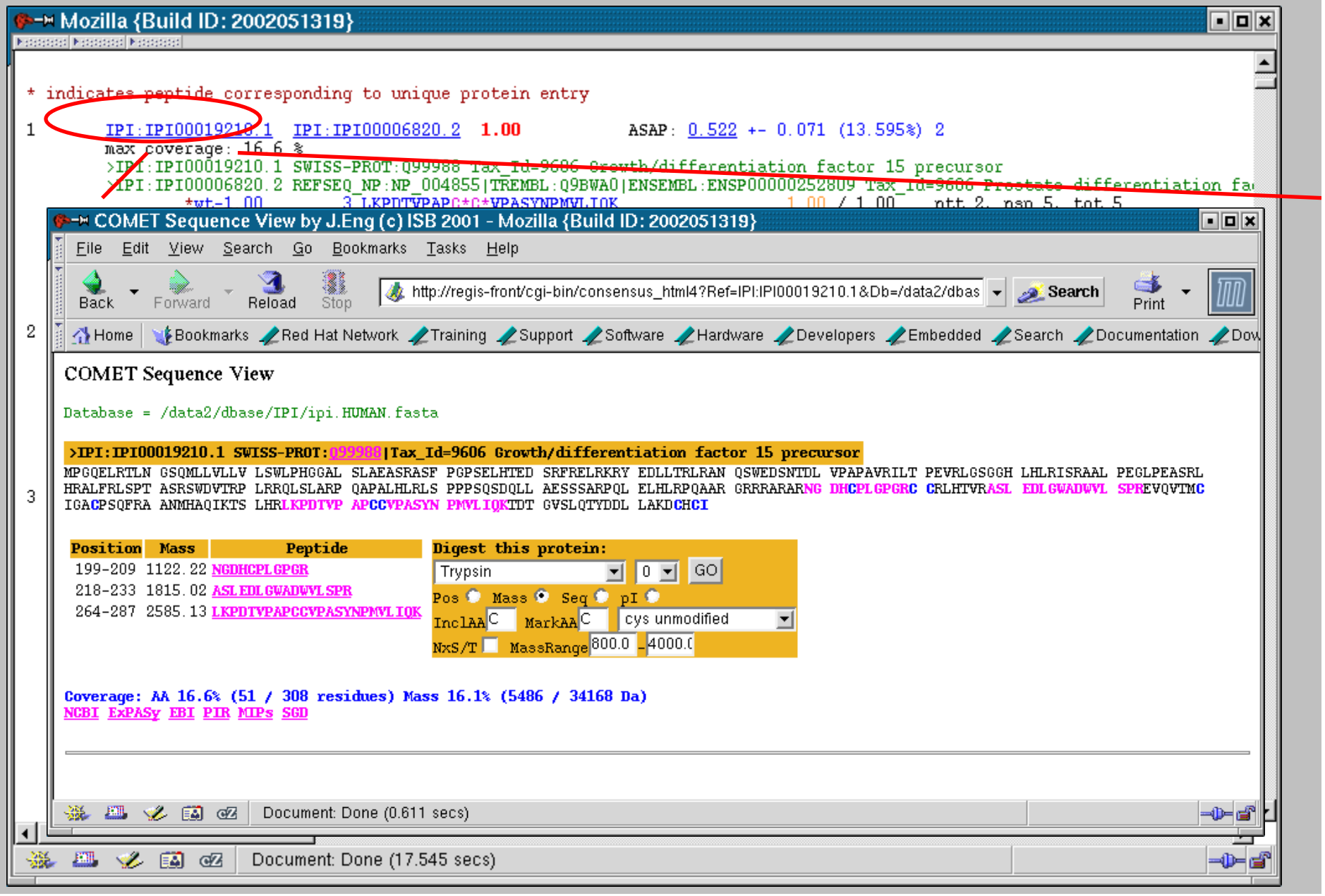
independent evidence

annotation

number tryptic termini,
NSP,
number peptide instances

1+/2+/3+ same peptide

Protein Entry Link



Peptide Link

16 IPI00332161 IPI00328111 **1.00**

max coverage: 14.8% XPRESS: **0.74** ± 0.18(21) ASAPRatio: **0.43** ± 0.34(4) num unique peps: 9 tot num peps: 20

>IPI:IPI00332161 Ig gamma-1 chain C region Sw:P01857

>IPI:IPI00328111 Factor VII active site mutant immunoconjugate Tr:Q96PQ8

	weight	peptide sequence	nsp adj prob	init prob	ntt	nsp bin	total	pep grp ind
*	wt-1.00	3_SC*DKTHTC*PPC*PAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	1	
*	wt-1.00	3_THTCPPCPAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	3	a-3
	wt-0.50	3_C*PAPELLGGPSVFLFPPKPK	1.00	1.00	1	5	1	c-3
*	wt-1.00	2_THTCPPCPAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	1	a-2
*	wt-1.00	3_THTC*PPC*PAPELLGGPSVFLFPPKPK	1.00	1.00	2	5	8	
	wt-0.33	2_TPEVTC*VVVDVSHED	1.00	1.00	1	5	2	d-2
*	wt-1.00	2_TPEVTC*VVVDVSHEDPEVK	1.00	1.00	2	5	2	b-2
	wt-0.33	2_TPEVTCVVVDVSHED	1.00	1.00	1	5	2	
	wt-0.50	2_PAPELLGGPSVFLFPPKPK	0.96	0.83	1	5	2	
*	wt-1.00	3_TPEVTC*VVVDVSHEDPEVK	0.95	0.82	2	5	1	b-3
	wt-0.50	2_C*PAPELLGGPSVFLFPPKPK	0.91	0.68	1	5	1	c-2
	wt-0.33	3_TPEVTC*VVVDVSHED	0.91	0.67	1	5	1	d-3
	wt-0.50	2_PELLGGPSVFLFPPKPK	0.17	0.17	1	5	1	

http://regis.systemsbiology.net/cgi-bin/akeller/peptide_html.pl?Ref=3_THTC*PPC*PAPELLGGPSVFLFPP - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

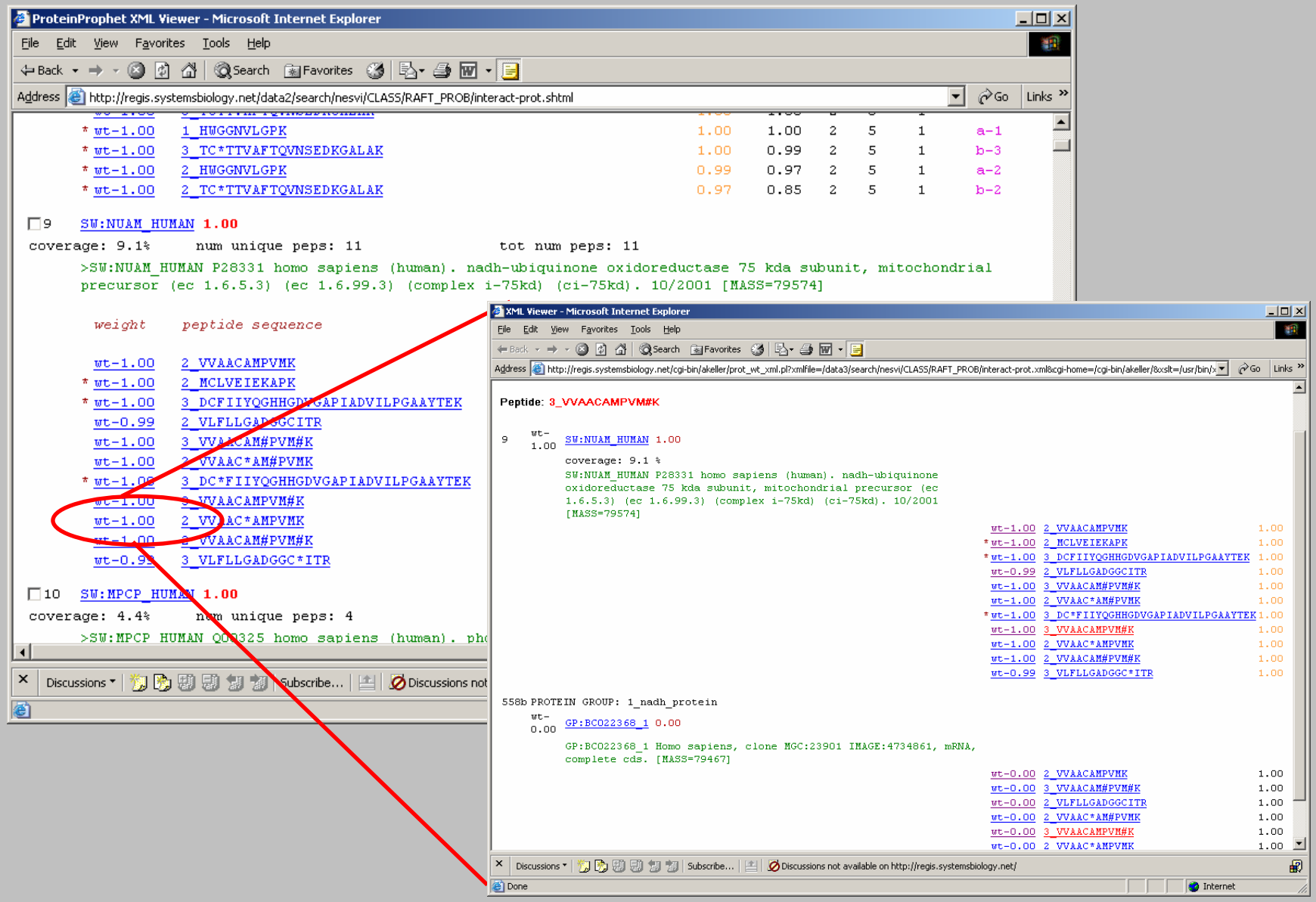
Address http://regis.systemsbiology.net/cgi-bin/akeller/peptide_html.pl?Ref=3_THTC*PPC*PAPELLGGPSVFLFPPKPK&Infile=/data2/search/guest/catherine/20031119/interac Go Links

input file /data2/search/guest/catherine/20031119/interact-data.htm:

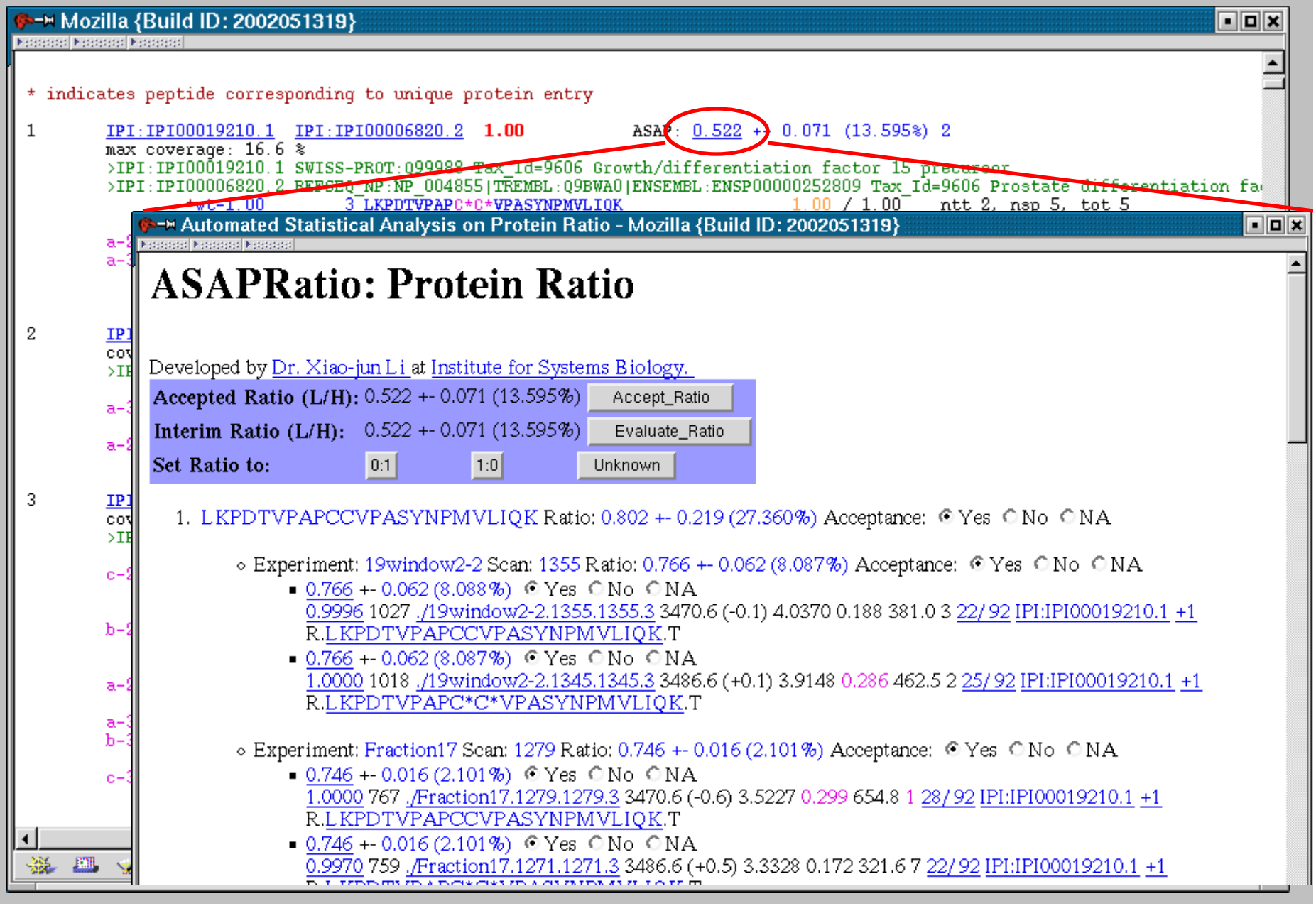
1	0.9054	859	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_05.2723.2732.3	3204.7 (+2.8)	5.1691	0.392*	1202.6	1	31/100	IPI00332161	+1	K.THT
2	0.8693	835	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_05.2678.2678.3	3204.7 (+2.1)	3.9854	0.369*	569.9	1	27/100	IPI00332161	+1	K.THT
3	0.9798	850	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_05.2705.2705.3	3204.7 (+2.0)	3.8541	0.094	595.3	1	25/100	IPI00332161	+1	K.THT
4	0.9853	807	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_05.2645.2645.3	3204.7 (+2.4)	3.3801	0.205	485.0	12	23/100	IPI00332161	+1	K.THT
5	0.8854	799	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_05.2619.2624.3	3204.7 (-0.3)	2.8926	0.073	240.2	38	19/100	IPI00332161	+1	K.THT
6	0.9996	726	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_06.2654.2658.3	3204.7 (+0.4)	4.9003	0.151	1033.2	1	29/100	IPI00332161	+1	K.THT
7	0.8684	782	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_06.2768.2771.3	3204.7 (+0.8)	4.6267	0.227*	464.4	23	22/100	IPI00332161	+1	K.THT
8	0.7754	416	./CSF_OD_AD_ICAT_1_5P_Repeat_111303_07.2636.2640.3	3204.7 (+0.1)	2.8502	0.001	605.3	5	26/100	IPI00332161	+1	K.THT

Generate Pep3D image | Save as | Pep3D.htm

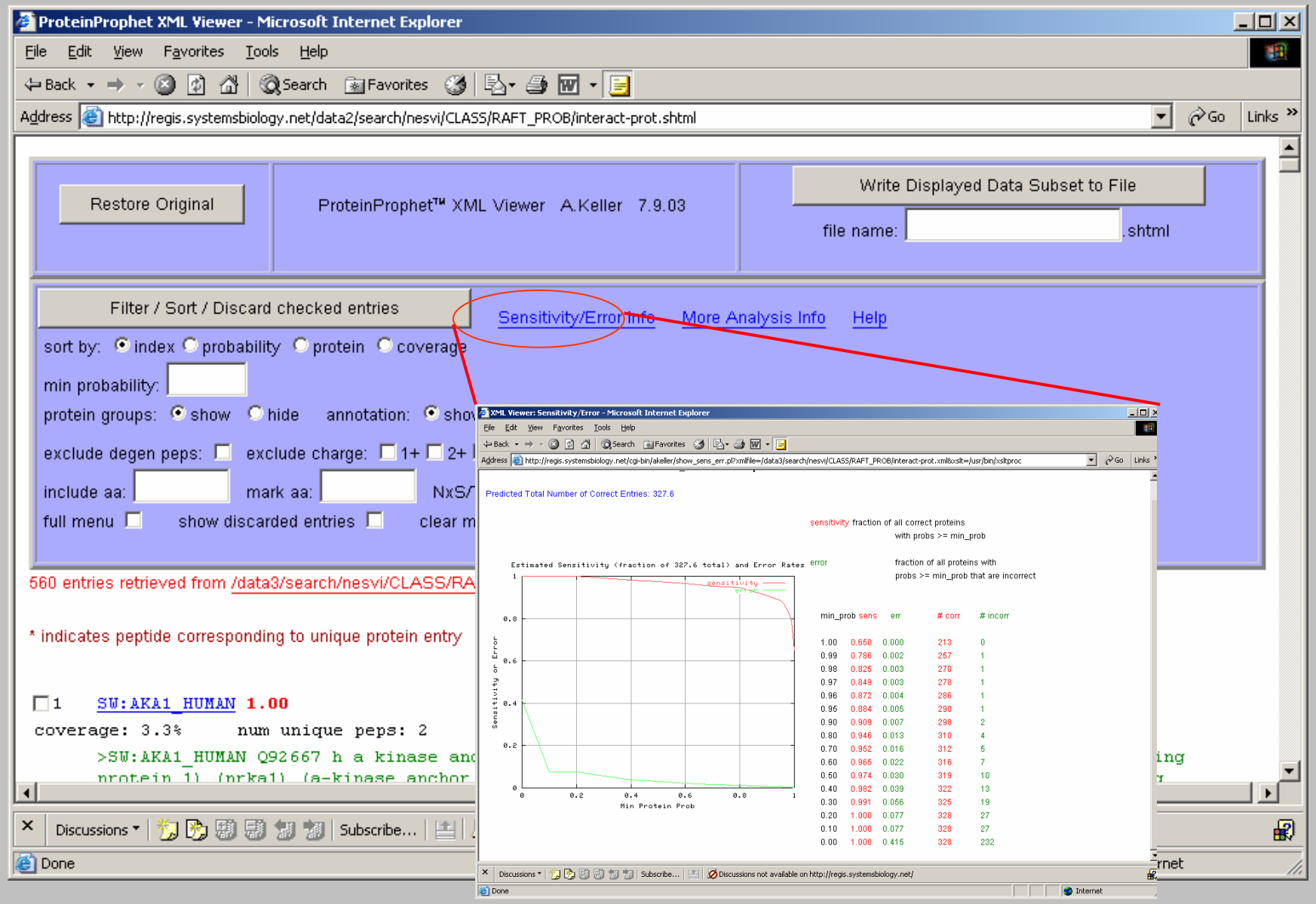
Peptide Weight Link



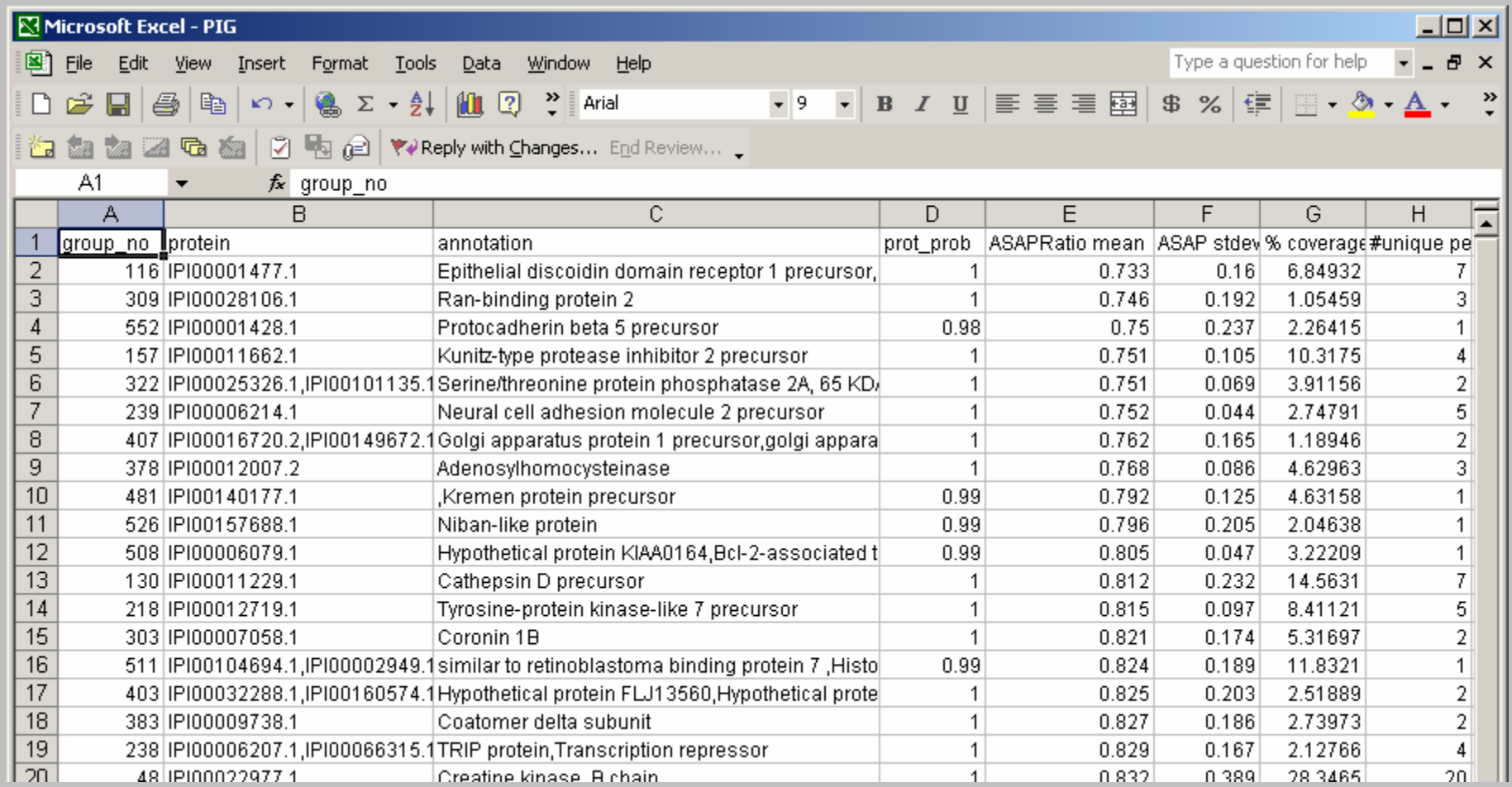
ASAPRatio Link (ICAT experiment)



Sensitivity/Error Rate Plot



Tab Delimited File Output



Microsoft Excel - PIG

File Edit View Insert Format Tools Data Window Help

Type a question for help

Arial 9

Reply with Changes... End Review...

A1 group_no

	A	B	C	D	E	F	G	H
1	group_no	protein	annotation	prot_prob	ASAPRatio mean	ASAP stdev	% coverage	#unique pe
2	116	IPI00001477.1	Epithelial discoidin domain receptor 1 precursor,	1	0.733	0.16	6.84932	7
3	309	IPI00028106.1	Ran-binding protein 2	1	0.746	0.192	1.05459	3
4	552	IPI00001428.1	Protocadherin beta 5 precursor	0.98	0.75	0.237	2.26415	1
5	157	IPI00011662.1	Kunitz-type protease inhibitor 2 precursor	1	0.751	0.105	10.3175	4
6	322	IPI00025326.1,IPI00101135.1	Serine/threonine protein phosphatase 2A, 65 KD/	1	0.751	0.069	3.91156	2
7	239	IPI00006214.1	Neural cell adhesion molecule 2 precursor	1	0.752	0.044	2.74791	5
8	407	IPI00016720.2,IPI00149672.1	Golgi apparatus protein 1 precursor,golgi appara	1	0.762	0.165	1.18946	2
9	378	IPI00012007.2	Adenosylhomocysteinase	1	0.768	0.086	4.62963	3
10	481	IPI00140177.1	,Kremen protein precursor	0.99	0.792	0.125	4.63158	1
11	526	IPI00157688.1	Niban-like protein	0.99	0.796	0.205	2.04638	1
12	508	IPI00006079.1	Hypothetical protein KIAA0164,Bcl-2-associated t	0.99	0.805	0.047	3.22209	1
13	130	IPI00011229.1	Cathepsin D precursor	1	0.812	0.232	14.5631	7
14	218	IPI00012719.1	Tyrosine-protein kinase-like 7 precursor	1	0.815	0.097	8.41121	5
15	303	IPI00007058.1	Coronin 1B	1	0.821	0.174	5.31697	2
16	511	IPI00104694.1,IPI00002949.1	similar to retinoblastoma binding protein 7 ,Histo	0.99	0.824	0.189	11.8321	1
17	403	IPI00032288.1,IPI00160574.1	Hypothetical protein FLJ13560,Hypothetical prote	1	0.825	0.203	2.51889	2
18	383	IPI00009738.1	Coatomer delta subunit	1	0.827	0.186	2.73973	2
19	238	IPI00006207.1,IPI00066315.1	TRIP protein,Transcription repressor	1	0.829	0.167	2.12766	4
20	48	IPI00022977.1	Creatine kinase, B chain	1	0.832	0.389	28.3465	20