Discovery of Specific Protein Biomarkers for the Differentiation of Pathogenic Bacteria

Steven M. Musser, Ph.D.

Center for Food Safety and Applied Nutrition
U.S. Food and Drug Administration
College Park, Maryland U.S.A.

FDA
CFSAN

Pathogenic Microorganisms

- Millions of cases of human illness each year
- High incidence in foods
- Hospital or clinical setting

Problem

- Emergence of new pathogens
- Acquired resistance to drug therapy
- Disease diagnosis/treatment
- Differentiation of pathogenic/nonpathogenic species

MS based proteomics tools

- Identification of “trait” characteristic proteins
- Rational design of molecular probes
- Quantitative measure of expression rates

MS based approaches

- Enzymatic digest of protein extract
- 2D gels – digest spots
- MALDI of whole cell
- LC/MS of protein extract

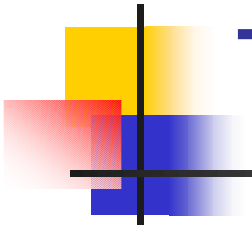
Whole digest approach

- Long run times (2-3 hours)
- Thousands of peptides
- Quantitative measure of expression

Too much non-specific data?

>sp|P02144|MYG_HUMAN Myoglobin - Homo sapiens (Human).
Trypsin:/K-\P /R-\P

Frag#	Res#	Sequence	Theor(Bo)	[M+H]	[M+2H]	[M+3H]
T7-8	48-56	(K)HLKSEDEMK(A)	1115.53	1116.54	558.77	372.85
T3-4	32-42	(R)LFKGHPETLEK(F)	1297.70	1298.71	649.86	433.58
T4-5	35-45	(K)GHPETLEKFDK(F)	1299.65	1300.65	650.83	434.22
T11	64-77	(K)HGATVLTALGGILK(K)	1349.80	1350.81	675.91	450.94
T8-9	51-62	(K)SEDEMKASEDLK(K)	1380.61	1381.62	691.31	461.21
T21-22	141-153	(K)DMASNYKELGFQG(-)	1458.64	1459.65	730.33	487.22
T11-12	64-78	(K)HGATVLTALGGILKK(K)	1477.90	1478.91	739.96	493.64
T10-11	63-77	(K)KHGATVLTALGGILK(K)	1477.90	1478.91	739.96	493.64
T18	119-133	(K)HPGDFGADAQGAMNK(A)	1514.66	1515.67	758.34	505.89
T2	17-31	(K)VEADIPGHGQEVLIR(L)	1631.86	1632.87	816.94	544.96
T1	1-16	(-)GLSDGEWQLVLNVWGK(V)	1799.92	1800.93	900.97	600.98
T14	80-96	(K)GHHEAEIKPLAQSHATK(H)	1852.95	1853.96	927.49	618.66
T17	103-118	(K)YLEFISECIIQVLQSK(H)	1912.00	1913.01	957.01	638.34
T13-14	79-96	(K)KGHHEAEIKPLAQSHATK(H)	1981.05	1982.06	991.53	661.36
T2-3	17-34	(K)VEADIPGHGQEVLIRLFK(G)	2020.11	2021.12	1011.06	674.38
T14-15	80-98	(K)GHHEAEIKPLAQSHATKHK(I)	2118.11	2119.12	1060.06	707.04
T18-19	119-139	(K)HPGDFGADAQGAMNKALELFR(K)	2244.07	2245.08	1123.05	749.03
T16-17	99-118	(K)IPVKYLEFISECIIQVLQSK(H)	2349.30	2350.31	1175.66	784.11
T17-18	103-133	(K)YLEFISECIIQVLQSKHPGDFGADAQGAMNK(A)	3408.65	3409.66	1705.33	1137.22
T1-2	1-31	(-)GLSDGEWQLVLNVWGKVEADIPGHGQEVLIR(L)	3413.77	3414.78	1707.89	1138.93



Too much data?

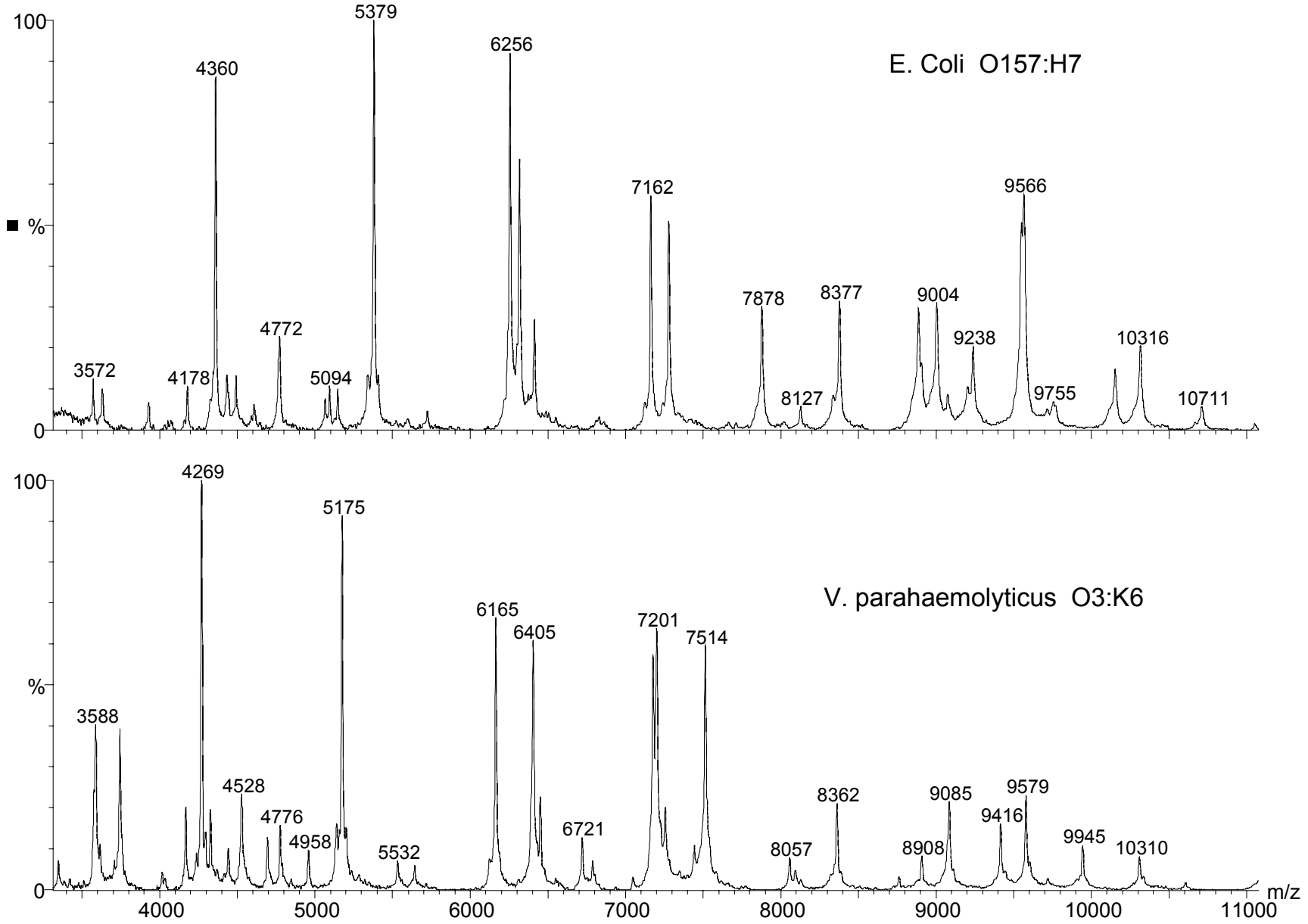
- Example – Myoglobin
  - 40 peptides from human trypsin (1 missed)
  - Compared to Gorilla, only 4 different peptides
  - Whole protein 17052 avg. (human) vs. 17042 avg. gorilla

2D Gels

- Long run times
- Low resolution
- Repeatability
- Poor recovery from gel spot

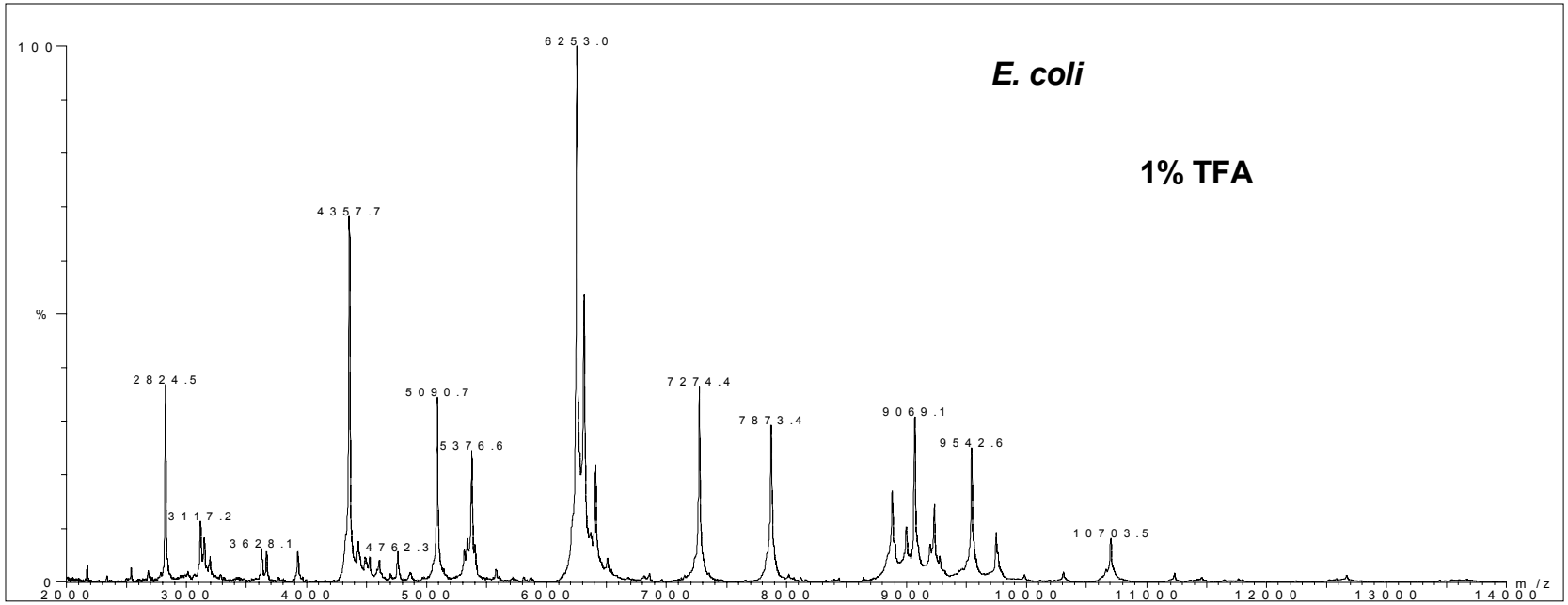
Intact protein approach MALDI

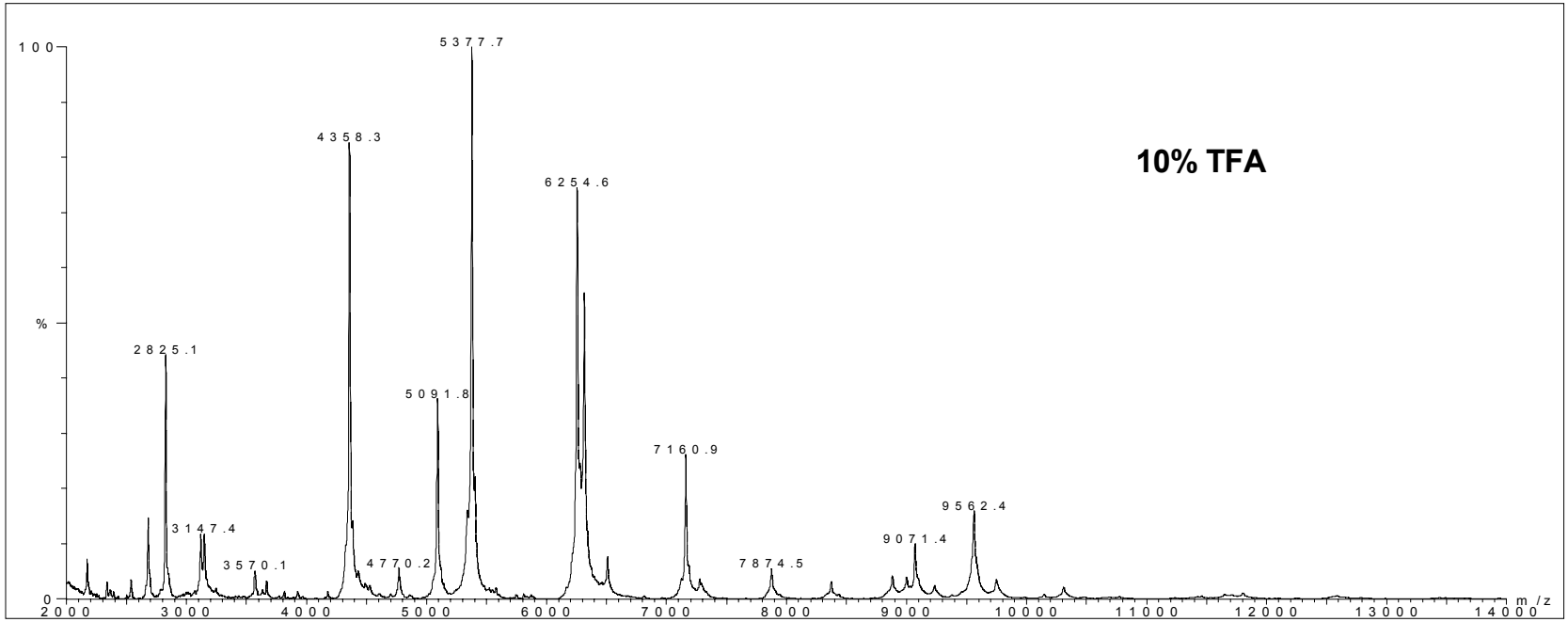
- Easy sample prep.
  - Sterilize cell in ethanol
  - Mix cell pellet with appropriate matrix
- Rapid analysis

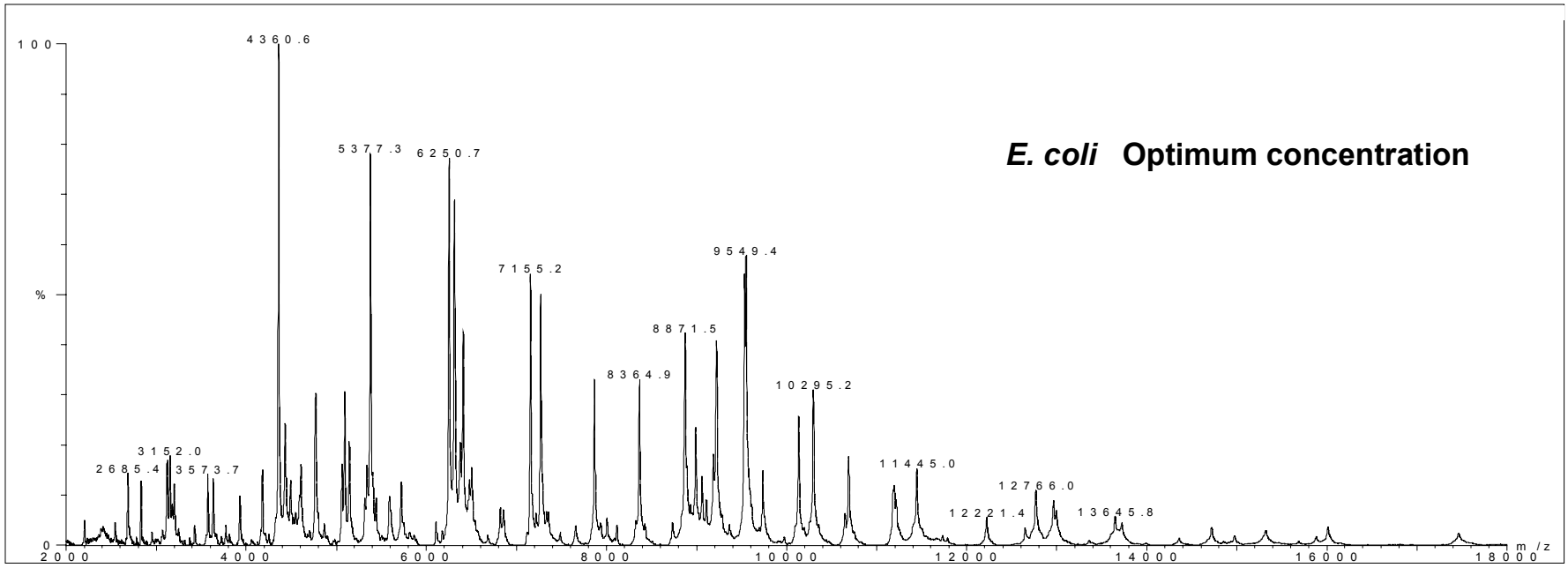

E. Coli O157:H7
100
%
0
5379
6256
4360
7162
9566
7878
8377
9004
9238
4772
10316
3572
4178
5094
8127
9755
10711
V. parahaemolyticus O3:K6
100
%
0
4269
5175
6165
6405
7201
7514
3588
4528
4776
4958
5532
6721
8362
9085
9579
9416
8057
8908
9945
10310
4000
5000
6000
7000
8000
9000
10000
11000
m/z

Experimental variables

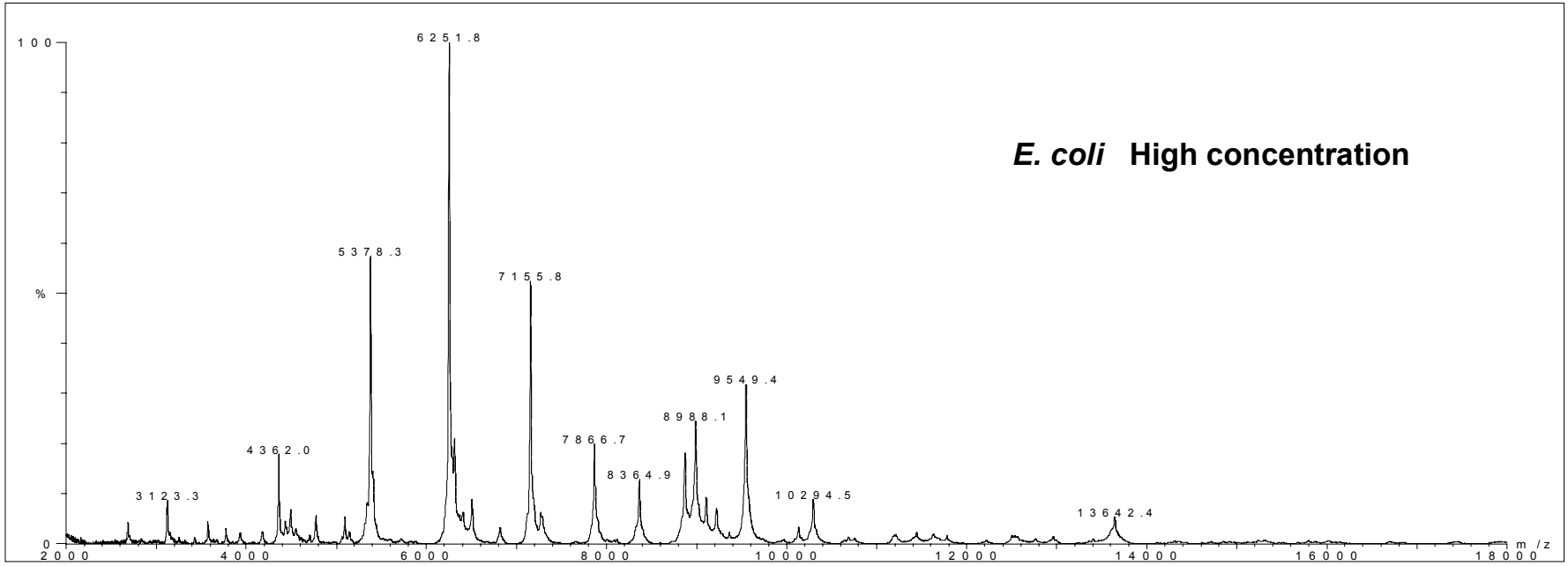
- Accelerating Voltage
- Matrix
- Type of laser
- Acid concentration and type
- Method of sample preparation
- Concentration of bacteria


E. coli
1% TFA
100
%
0
6253.0
4357.7
2824.5
5090.7
5376.6
3117.2
3628.1
4762.3
7274.4
7873.4
9069.1
9542.6
10703.5
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
m/z


10% TFA
100
%
0
5377.7
4358.3
6254.6
2825.1
5091.8
7160.9
3147.4
3570.1
4770.2
7874.5
9071.4
9562.4
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
m/z



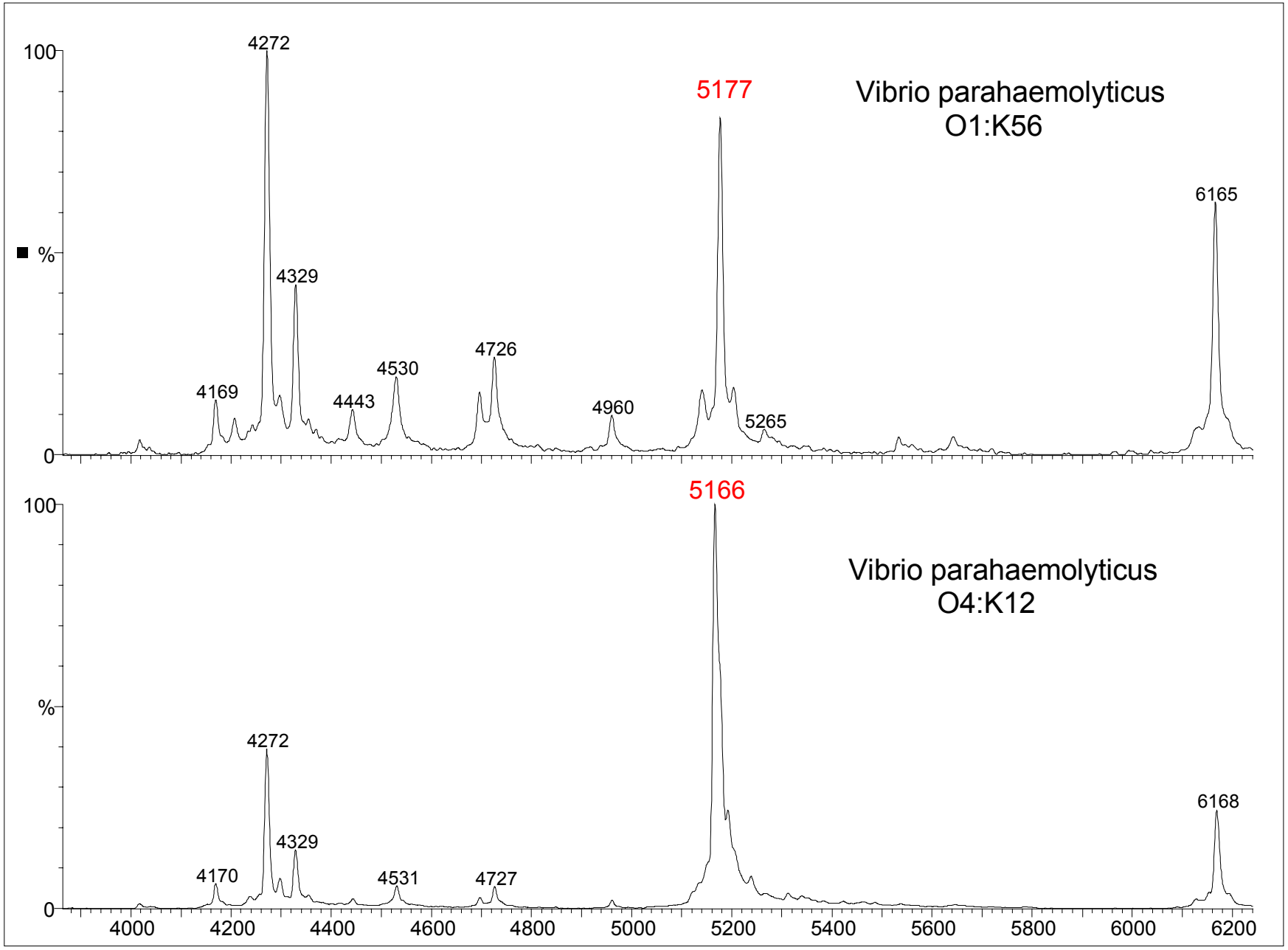
E. coli **Optimum concentration**



E. coli **High concentration**

Optimum conditions

- Cultures grown for 24 hrs
- Bacteria in 70% ethanol
- Matrix (α-cyano-4-hydroxycinnamic acid)
- Nonpolar solvent for matrix (acetonitrile)
- 2% TFA
- Three concentrations of sample
- Nd:YAG laser


100
%
0
4272
5177
6165
4329
4726
4530
4169
4443
4960
5265
Vibrio parahaemolyticus
O1:K56
100
%
0
5166
6168
4272
4329
4170
4531
4727
Vibrio parahaemolyticus
O4:K12
4000
4200
4400
4600
4800
5000
5200
5400
5600
5800
6000
6200

Problems with MALDI

- Limited to species level identification
- Multiple experimental variables
- Poor mass accuracy at higher masses (>10K)
- Biased towards low mass proteins
- Have to “find” protein in mixture for further analysis

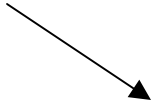
New approach – ES/LC/MS

- Better mass accuracy than MALDI
- Can be combined with LC
- LC/MS/MS of Protein Digest
- Multiply charged ions

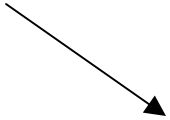
The plan

Extract cells

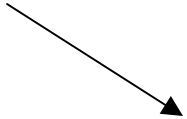
Fractionate cell extract by HPLC



Protease digestion of selected protein

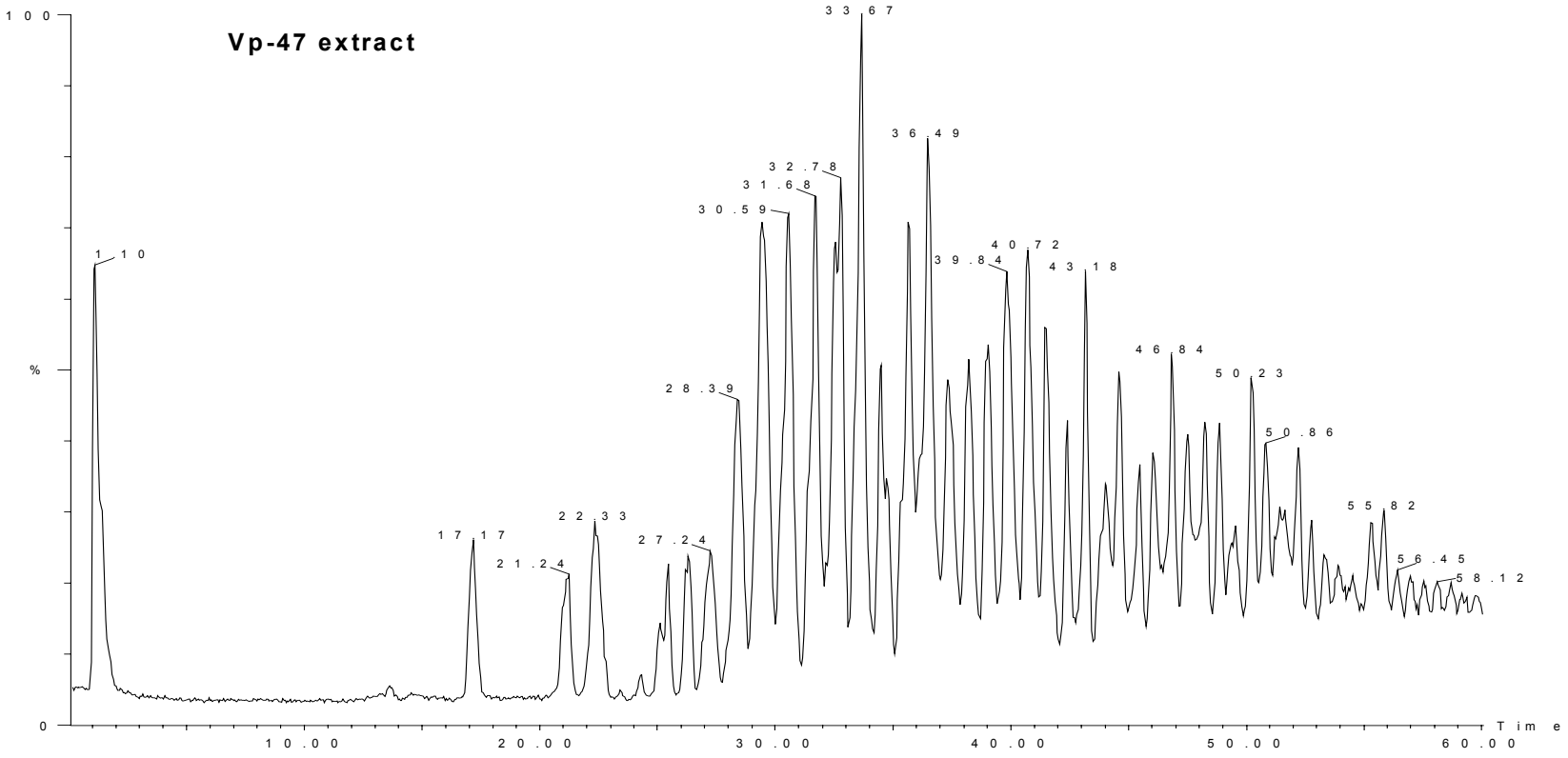


LC/MS/MS analysis of protein digest

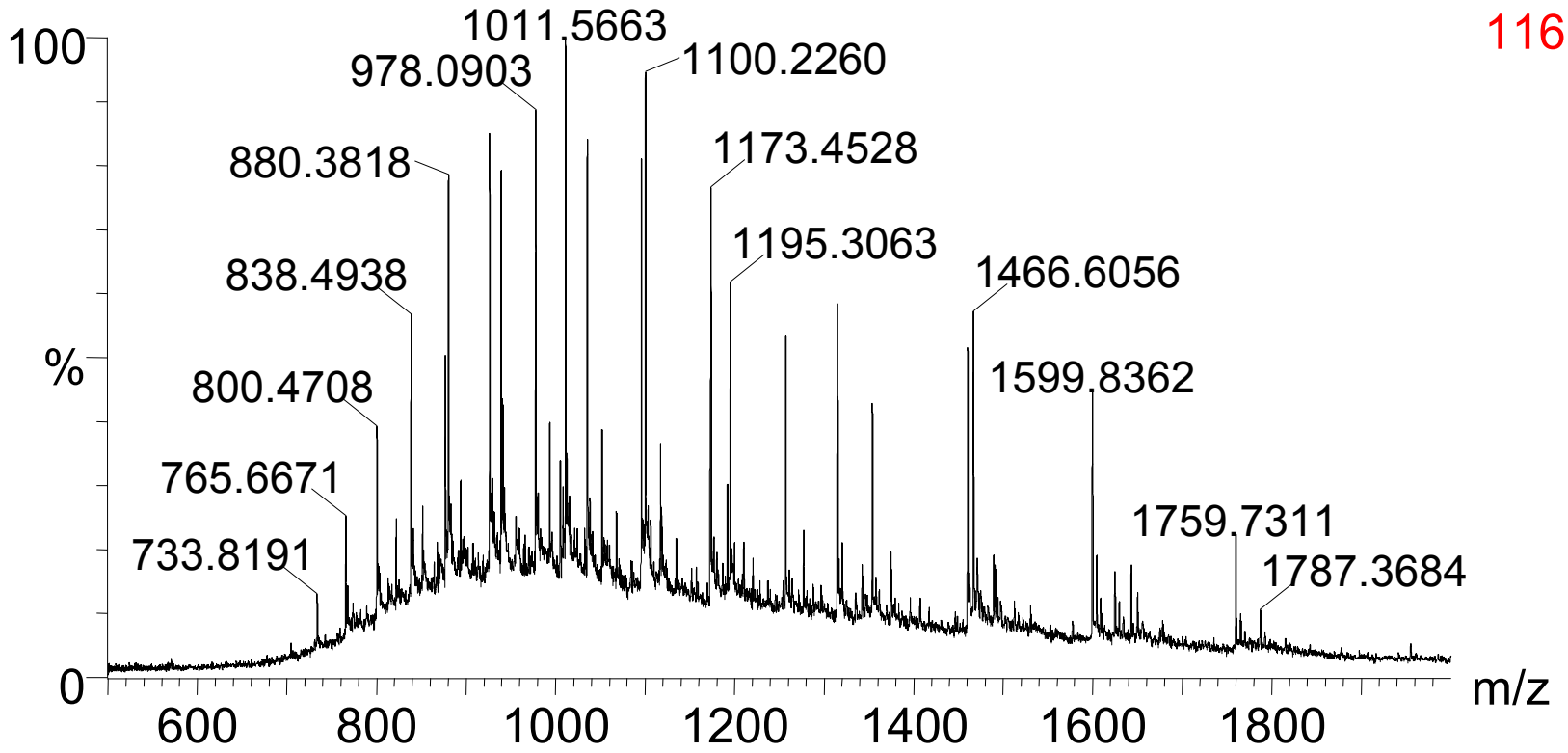


Identify protein via database search

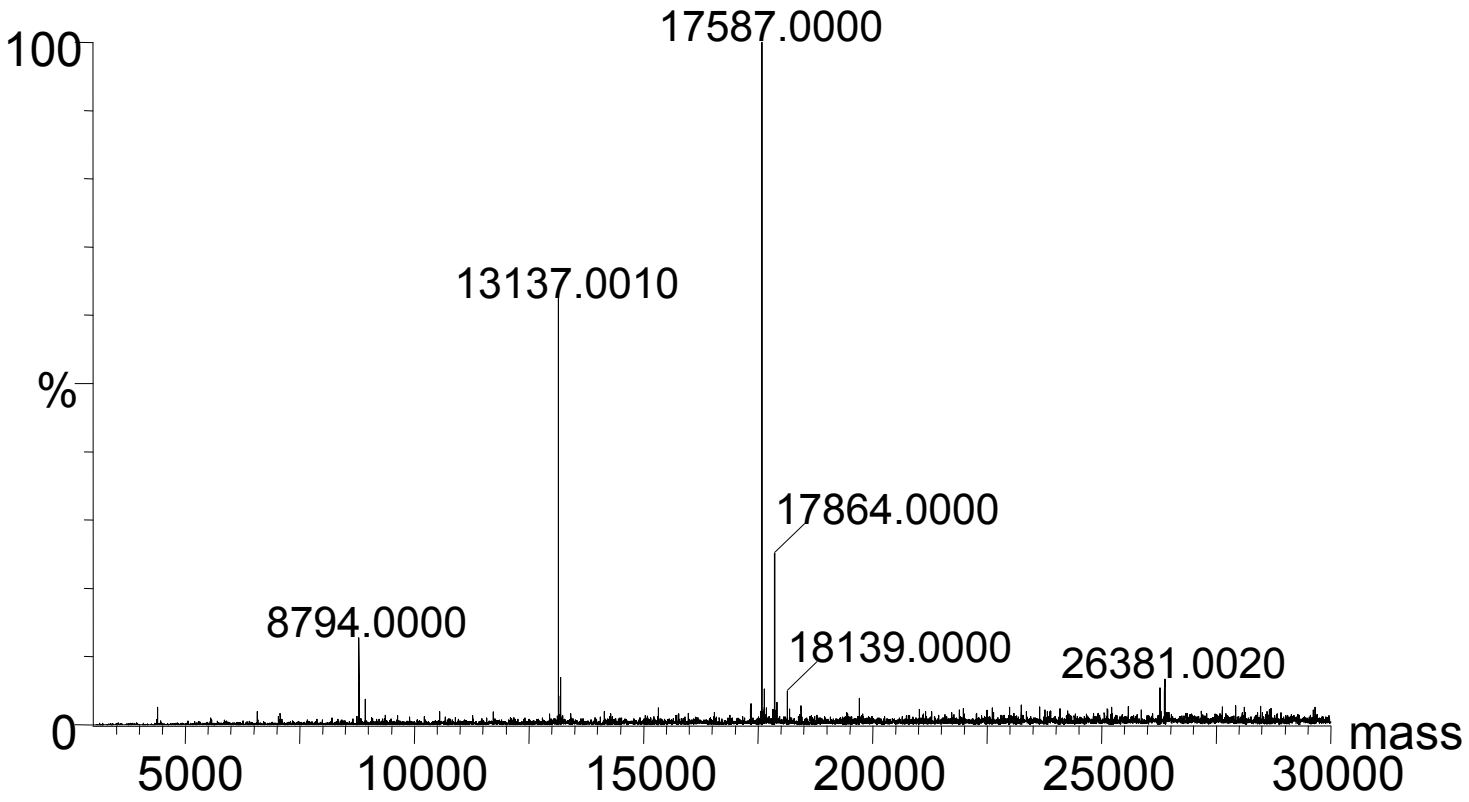
The problem


Vp-47 extract
100
%
0
1.10
17.17
21.24
22.33
27.24
28.39
30.59
31.68
32.78
33.67
36.49
39.84
40.72
43.18
46.84
50.23
50.86
55.82
56.45
58.12
10.00
20.00
30.00
40.00
50.00
60.00
Time

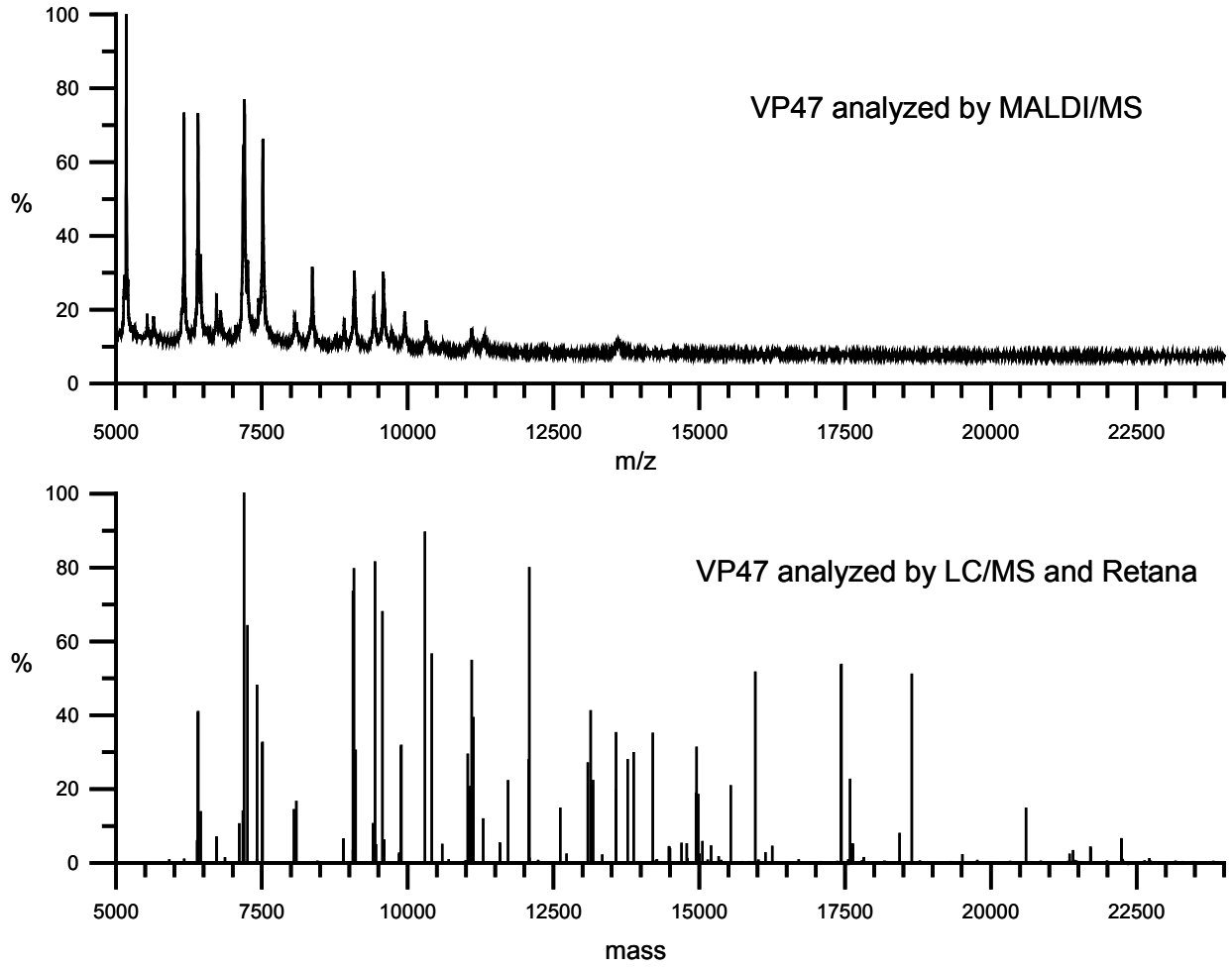
...and the problem continues



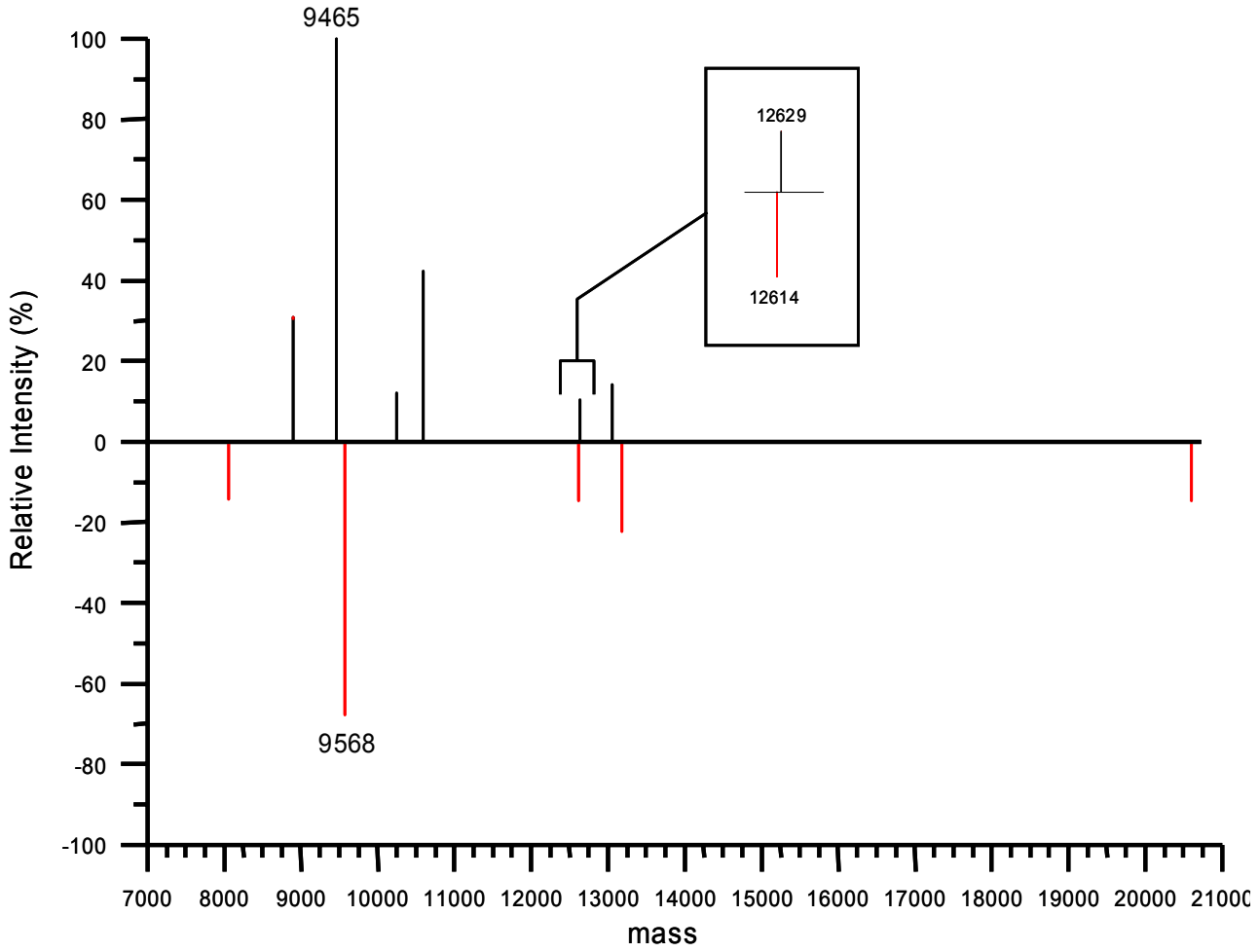
Enter software “solutions”



Automated protein profile from ES/LC/MS data



Accurate difference spectra



Sequencing information

Vibrio proteolyticus - DNA-binding protein HU-alpha (9449 Daltons)

Vibrio paraheamolyticus
Protein from nonclinical strain = 9468 Daltons
Protein from clinical strain = 9568 Daltons

AALEATLE GVTGALK
MNKTQLIDFI AEKADLSKAQ AKAALEATLD GVTDALKEGD

QVQLIGFGTF KVNHRAARTG RNPKTGAEIQ IAAANVPAFV

AGKALKDAVK

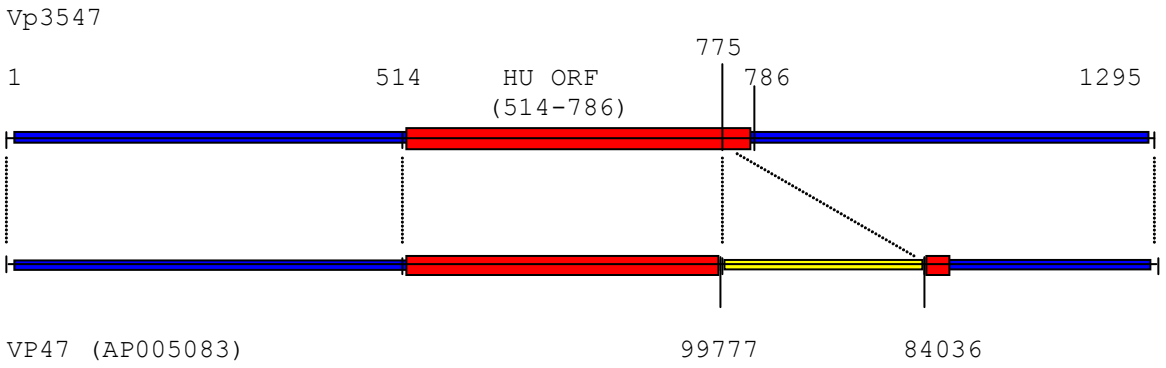
More challenges

	1	6	11	16	21	26	31	36	41
V. proteolyticus[a]	MNKTQ	LIDFI	AEKAD	LSKAQ	AKAAL	EATLD	GVTDA	LKEGD	QVQLI
V. cholerae[b]	MNKTQ	LIDFI	AEKAD	LTKVQ	AKAAL	EATLG	AVEGA	LKDGD	QVQLI
V. vulnificus[c]	MNKTQ	LIDFI	AEKAD	LSKAQ	AKAAL	EATLN	GVTDA	LKEGD	QVQLI
V. parahaemolyticus (O3:K6)[d]	MNKTQ	LIDFI	AEKAD	LSKAQ	AKAAL	EATLE	GVTGA	LKEGD	QVQLI
V. parahaemolyticus (O4:K55)	MNKTQ	LIDFI	AEKAD	LSKAQ	AKAAL	EATLE	GVTGA	LKEGD	QVQLI

	46	51	56	61	66	71	76	81	86	91
V. proteolyticus	GFGTF	KVNHR	AARTG	RNPKT	GAEIQ	IAAAN	VPAFV	AGKAL	KDAVK	
V. cholerae	GFGTF	KVNHR	SARTG	RNPKT	GEEIK	IAAAN	VPAFV	AGKAL	KDAIK	
V. vulnificus	GFGTF	KVNHR	AARTG	RNPKT	GDEIQ	IAAAN	VPAFV	AGKAL	KESVN	
V. parahaemolyticus (O3:K6)	GFGTF	KVNHR	AARTG	RNPKT	GDEIQ	IAAAN	VPAFV	AGKAL	KEACN	D
V. parahaemolyticus (O4:K55)	GFGTF	KVNHR	AARTG	RNPKT	GDEIQ	IAAAN	VPAFV	AGKAL	KESVN	

[a] Swissprot Accession #P28080 [11]
[b] NCBI Accession #NP_229929 [12]
[c] NCBI Accession #NP_760155 [13]
[d] NCBI Accession #NP_799290 [14]

The solution



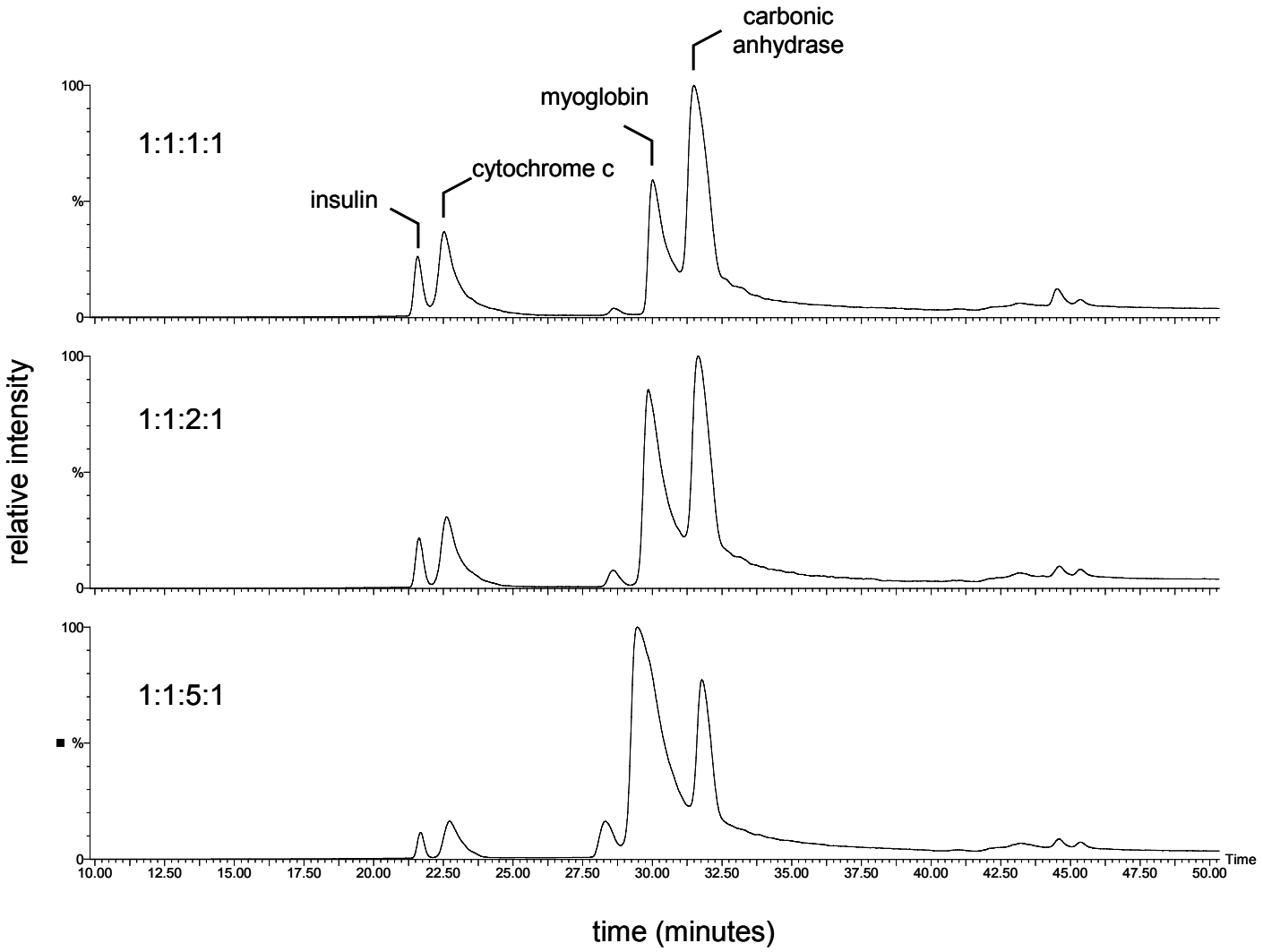
HU open reading frame

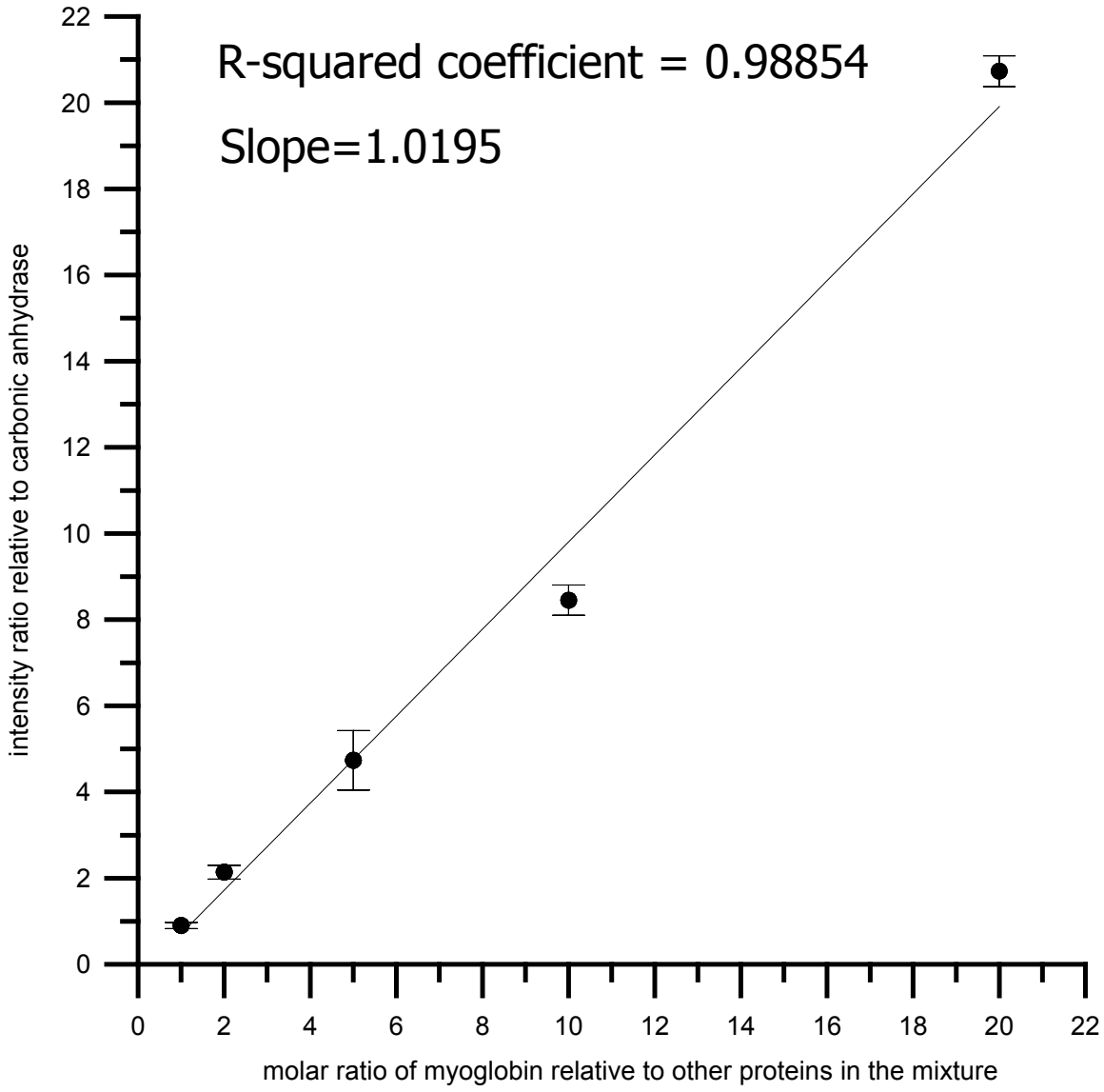
Vp 3547 sequence data

Vp 47 gene sequence (AP005083)

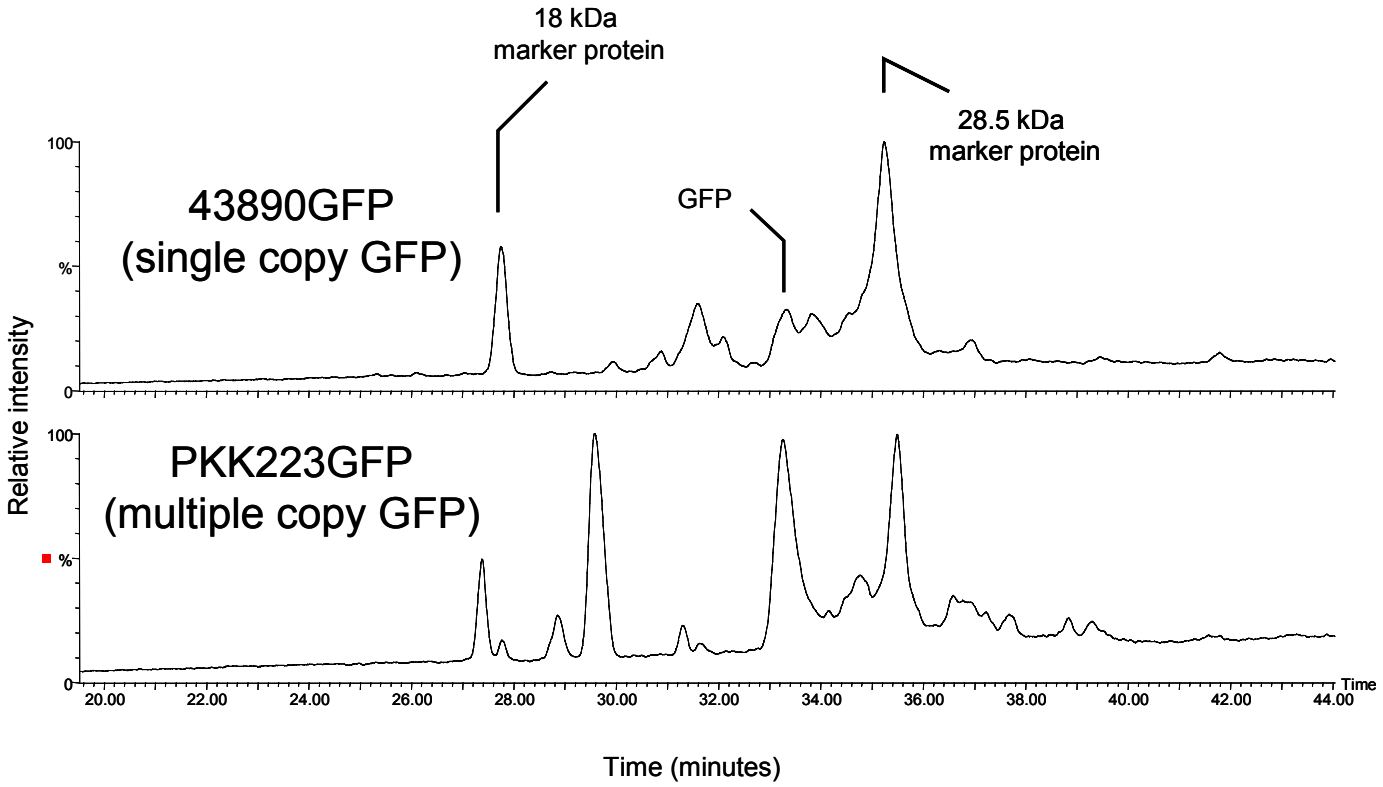
Quantitation?

- Deconvoluted profile “spectra” appear to be highly repeatable.
- Can housekeeping proteins be used as internal standards?
- New genetic material will not change the expression of all proteins.

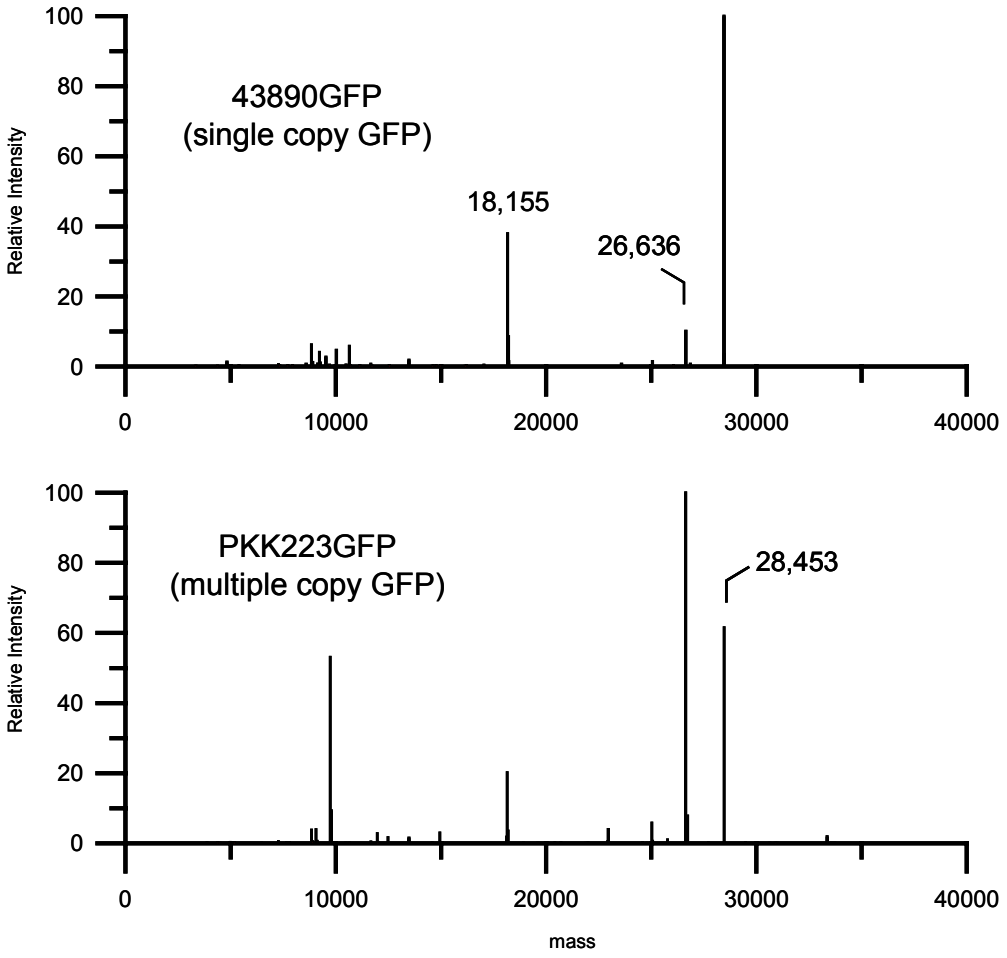

carbonic
anhydrase
myoglobin
cytochrome c
insulin
1:1:1:1
1:1:2:1
1:1:5:1
100
%
0
relative intensity
Time
10.00
12.50
15.00
17.50
20.00
22.50
25.00
27.50
30.00
32.50
35.00
37.50
40.00
42.50
45.00
47.50
50.00
time (minutes)


R-squared coefficient = 0.98854
Slope=1.0195
intensity ratio relative to carbonic anhydrase
molar ratio of myoglobin relative to other proteins in the mixture
0
2
4
6
8
10
12
14
16
18
20
22

Does it work in bacterial extracts?



Does it work in bacterial extracts? Yes



Conclusions

- Rapid screening of microorganisms at the species level
- Identification of new “targets” for detection of pathogenic strains
- Quantitative measure of protein expression
- Development of computer tools in informatics and chemometrics are necessary

Acknowledgements

- Tracie Williams — CFSAN, College Park
- John Callahan — CFSAN, College Park

- Andy Depaola — CFSAN, Dauphin Island, AL
- Steve Monday — CFSAN, College Park, MD
- Peter Feng — CFSAN, College Park, MD

- Peter Leopold — Bioanalyte, Inc., ME