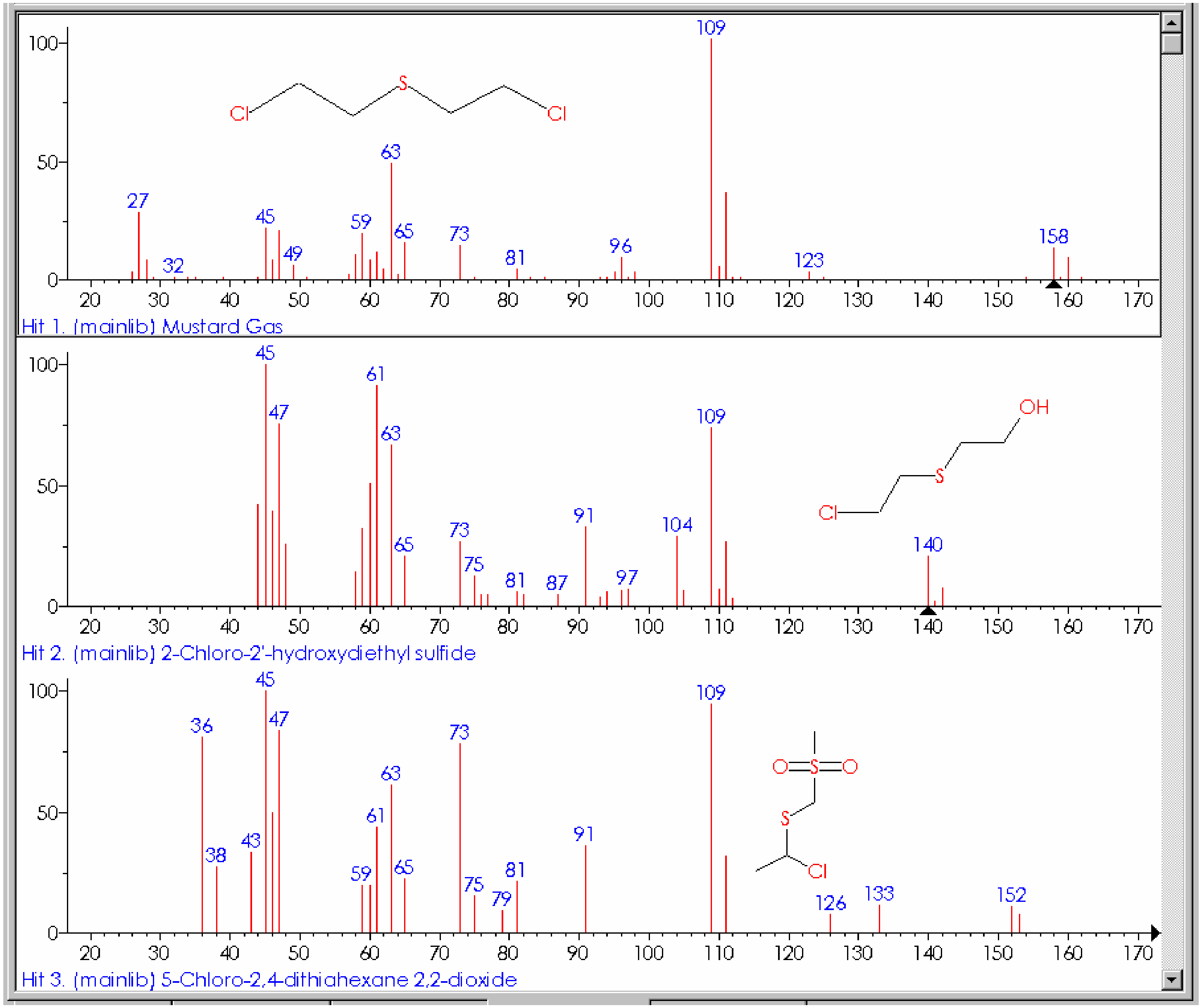
How We Handle Mass Spectra

NIST Mass Spectrometry Data Center

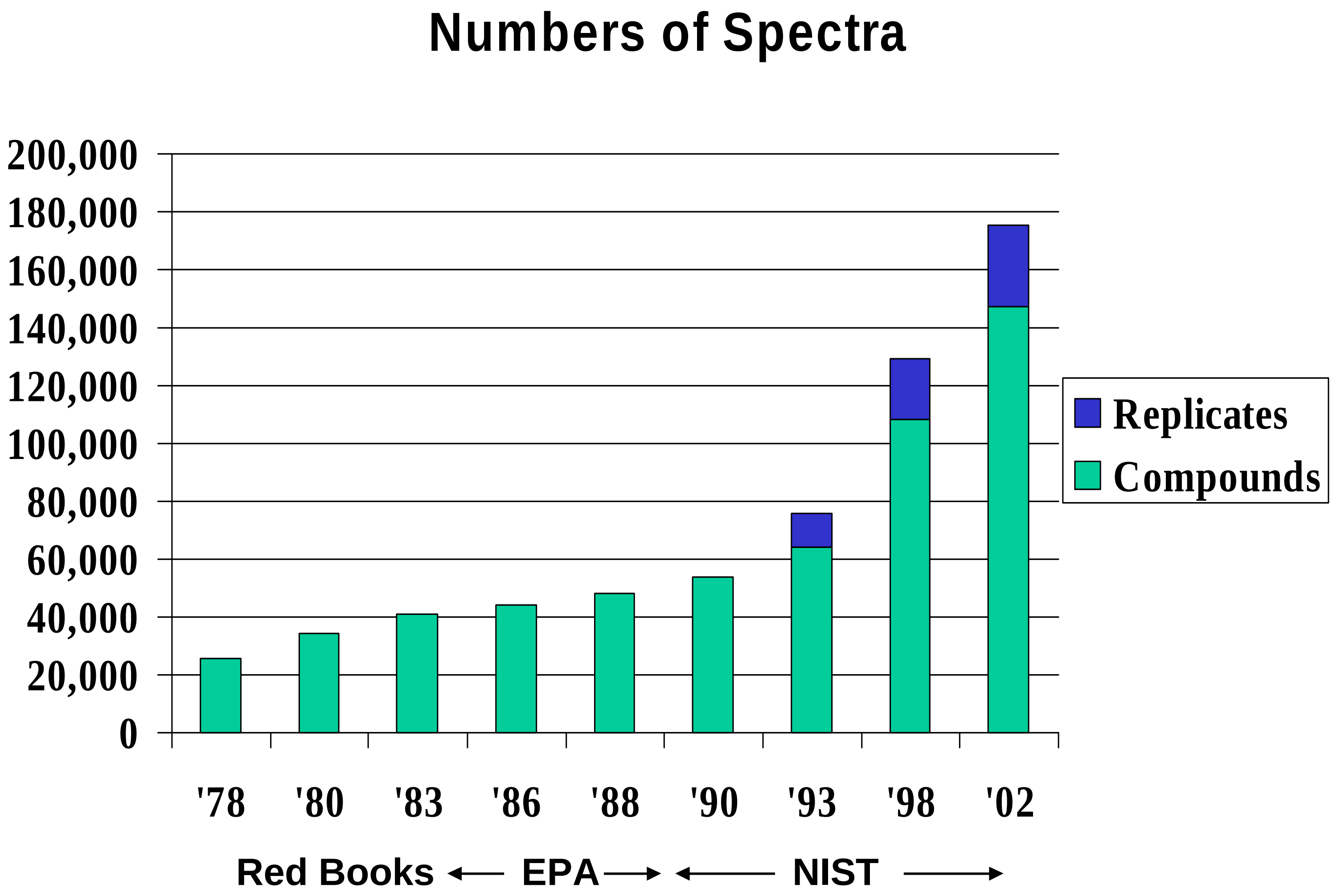


Hit 1. (mainlib) Mustard Gas

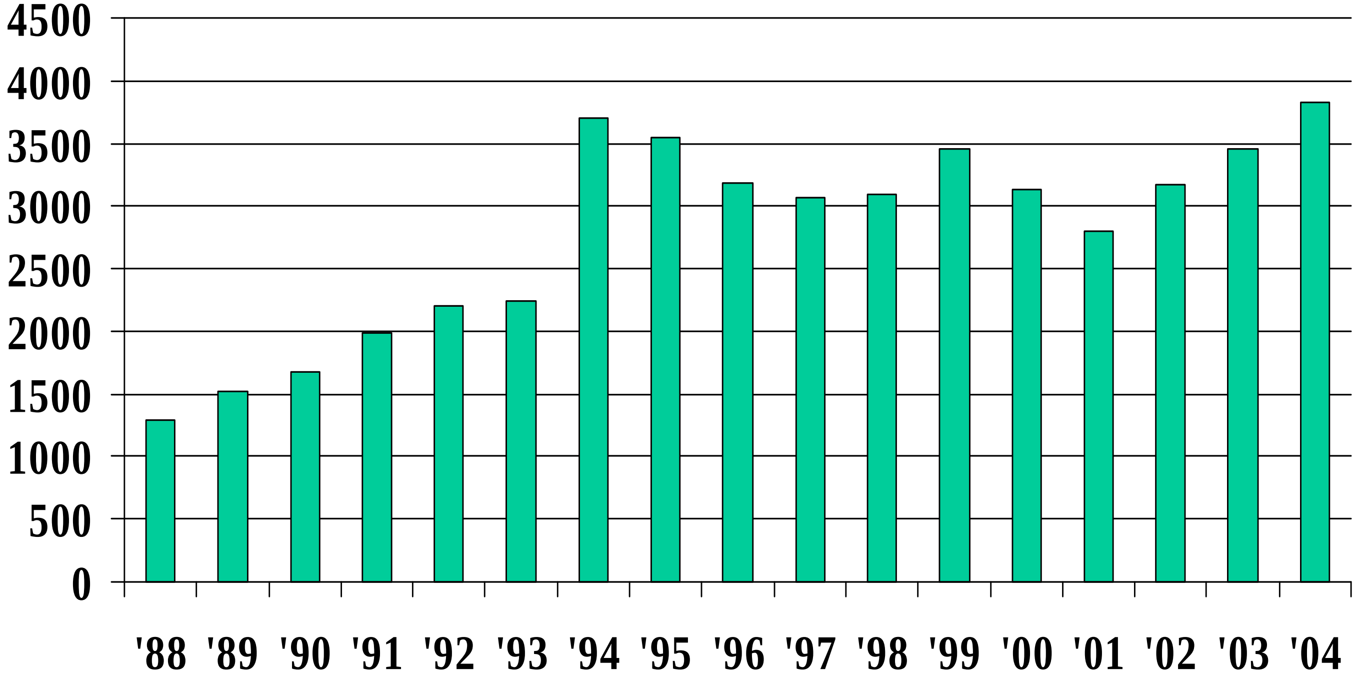
Hit 2. (mainlib) 2-Chloro-2'-hydroxydiethyl sulfide

Hit 3. (mainlib) 5-Chloro-2,4-dithiahexane 2,2-dioxide

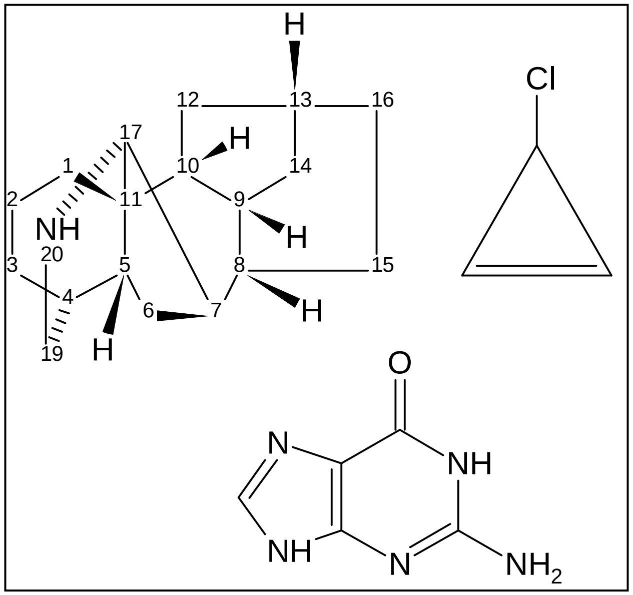
NIST/EPA/NIH Mass Spectral Library



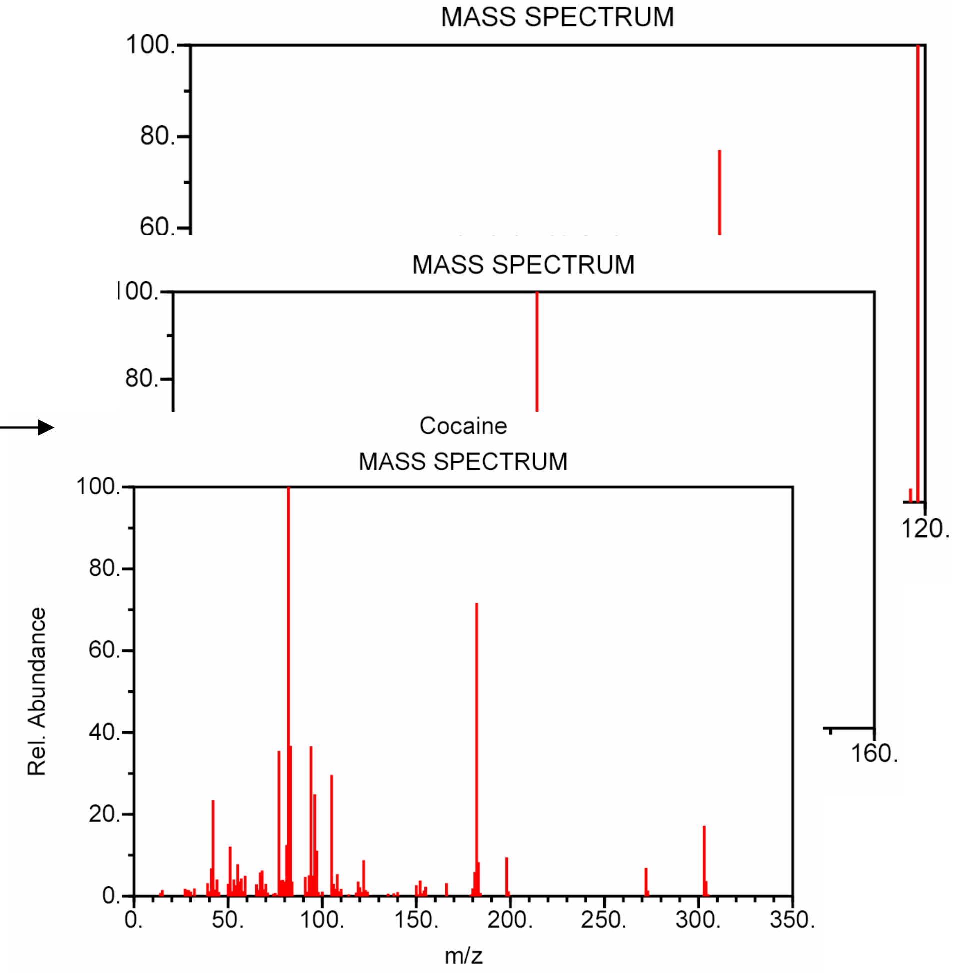
Libraries Distributed/Year



The Data

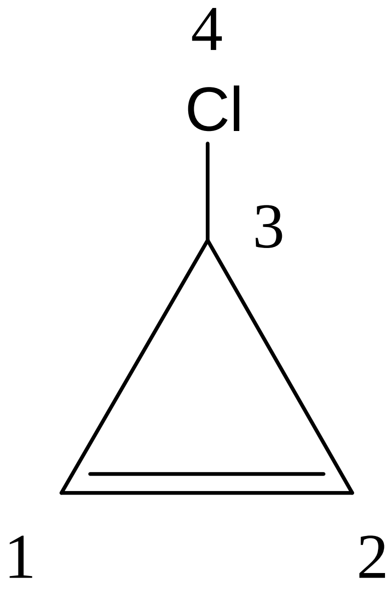


H
12
13
16
17
H
1
10
14
2
11
9
NH
H
20
8
3
5
15
4
6
7
H
19
H
Cl
O
N
NH
NH
N
NH_2



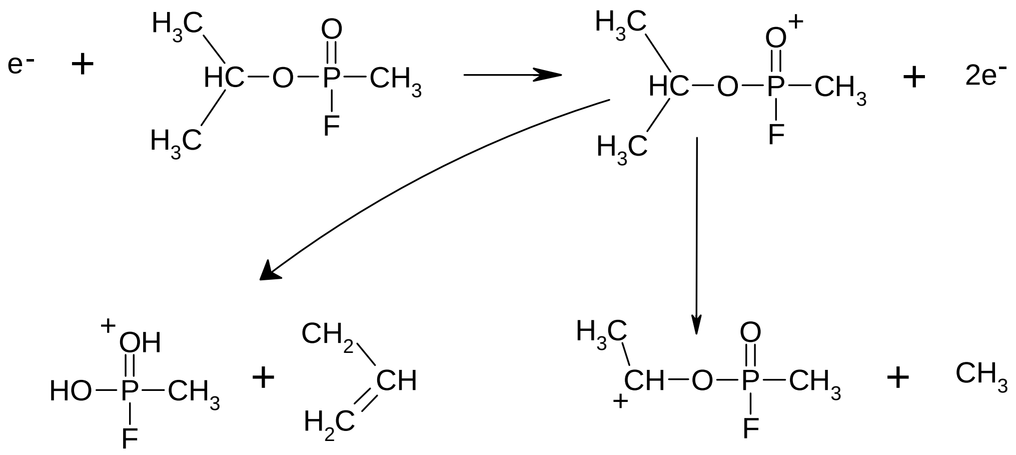
MASS SPECTRUM
100.
80.
60.
120.
MASS SPECTRUM
100.
80.
160.
Cocaine
MASS SPECTRUM
Rel. Abundance
100.
80.
60.
40.
20.
0.
0.
50.
100.
150.
200.
250.
300.
350.
m/z

Connection Table



	1	2	3	4
1		D	S	
2	D		S	
3	S	S		S
4				S

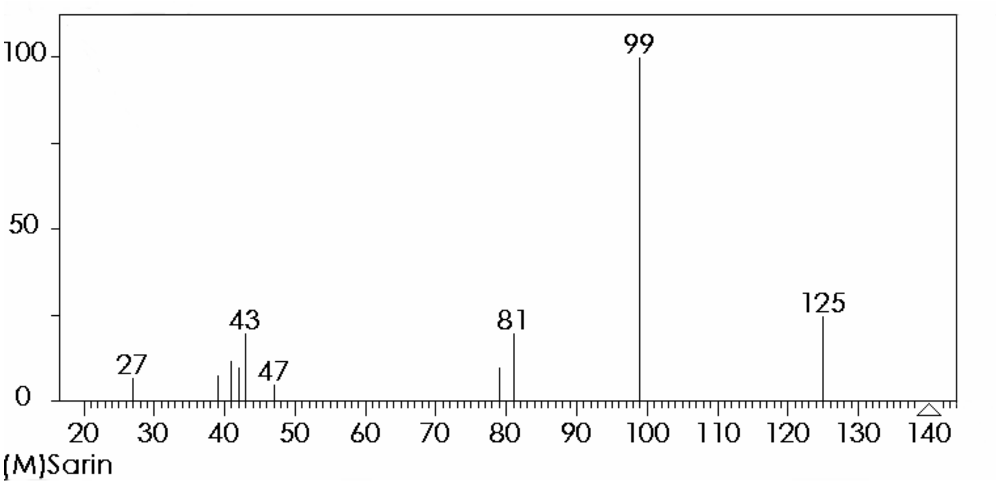
From Structure to Spectrum: A Mass “Fragmentogram”



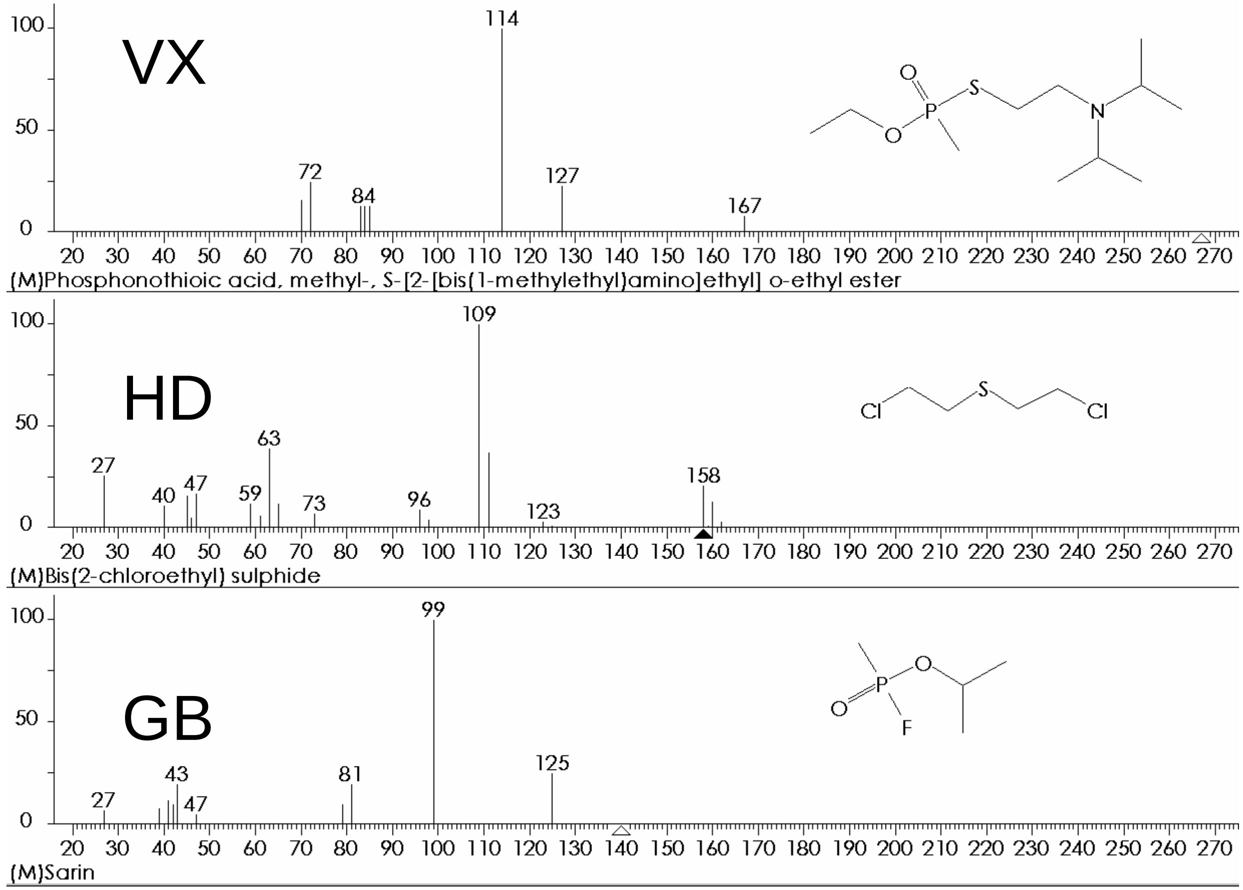
mass = 140 u

mass = 99 u

mass = 125 u



Molecular Fingerprints

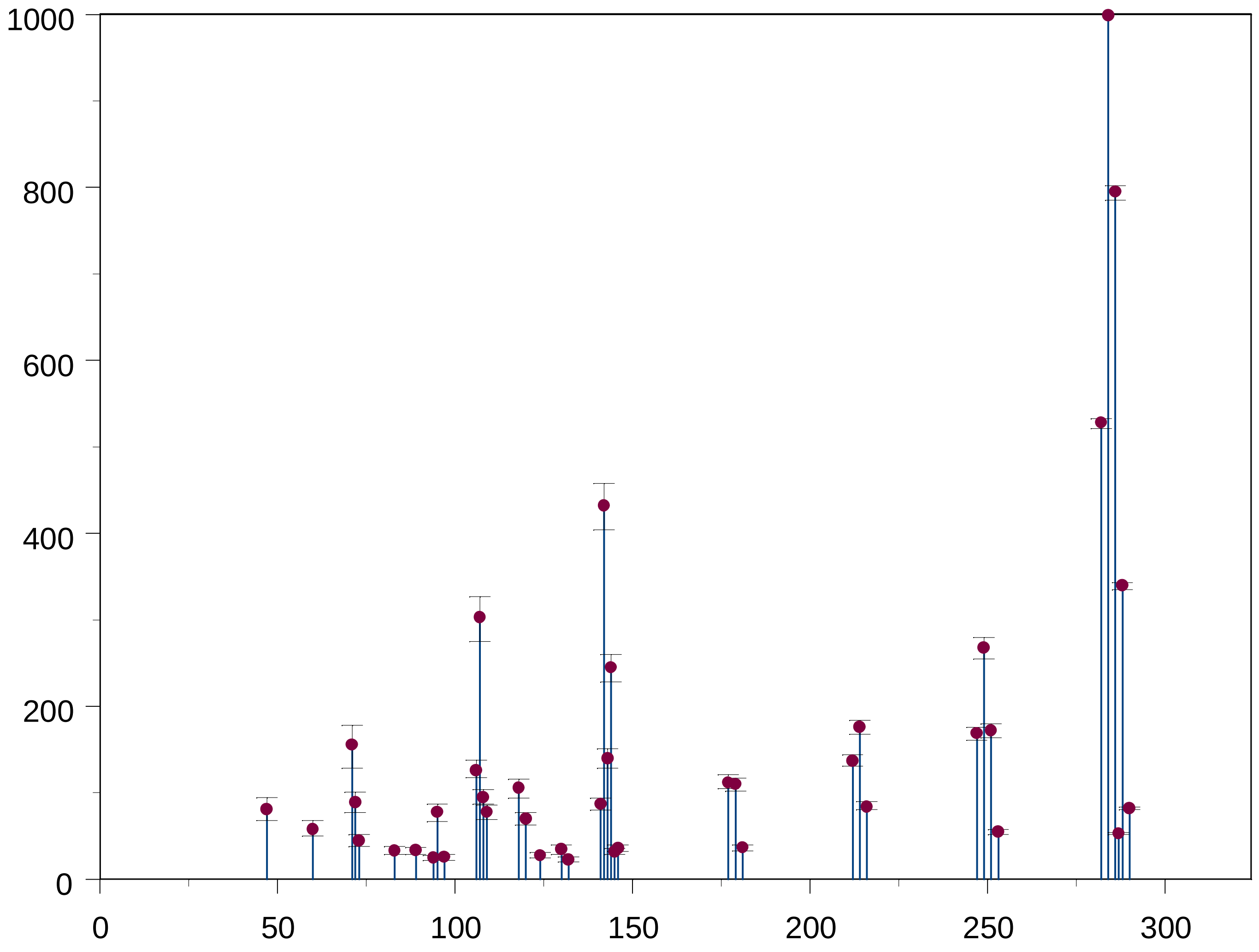


I will discuss

- Library Searching
  - Full and Partial Spectra
- Spectrum Purification
- Chemical Structure Representation
- Peptide Spectra Libraries

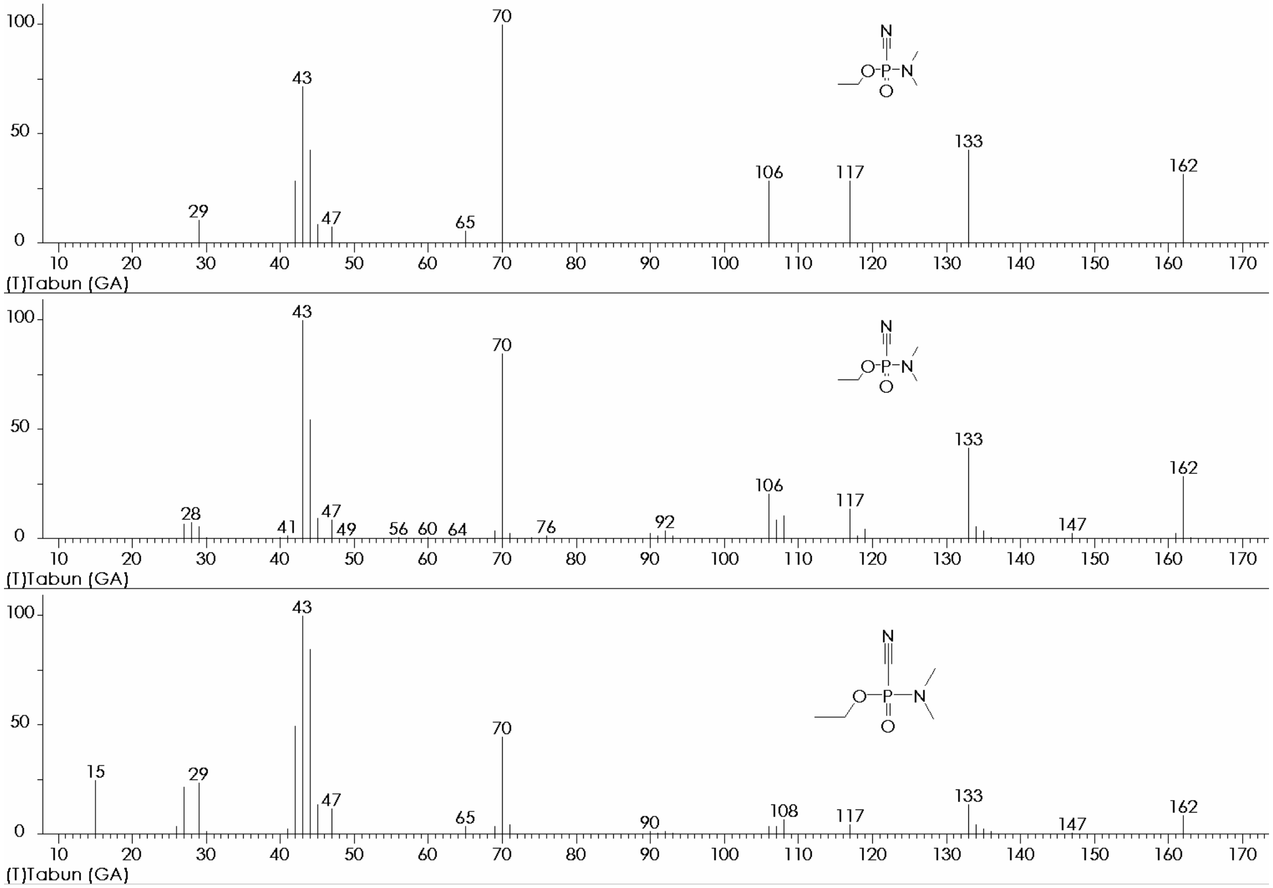
Instrument ‘Noise Signature’

250 Hexachlorobenzene Spectra
same instrument, calibration mix

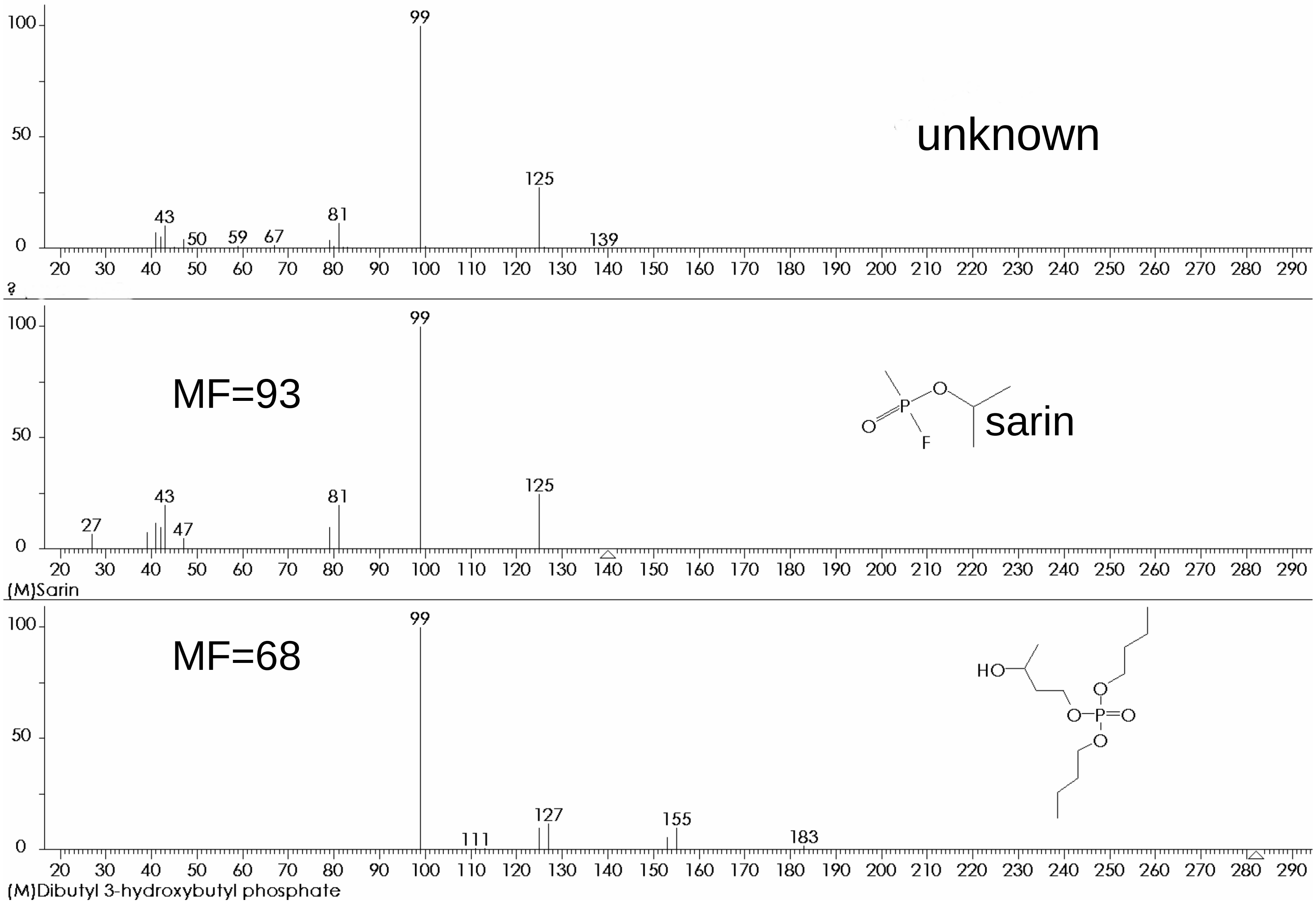


Bars show quartiles

Instrument Effects



Library Search



Spectral Similarity

$$\frac{\sum \sqrt{MR}}{\sqrt{\sum M \sum R}}$$

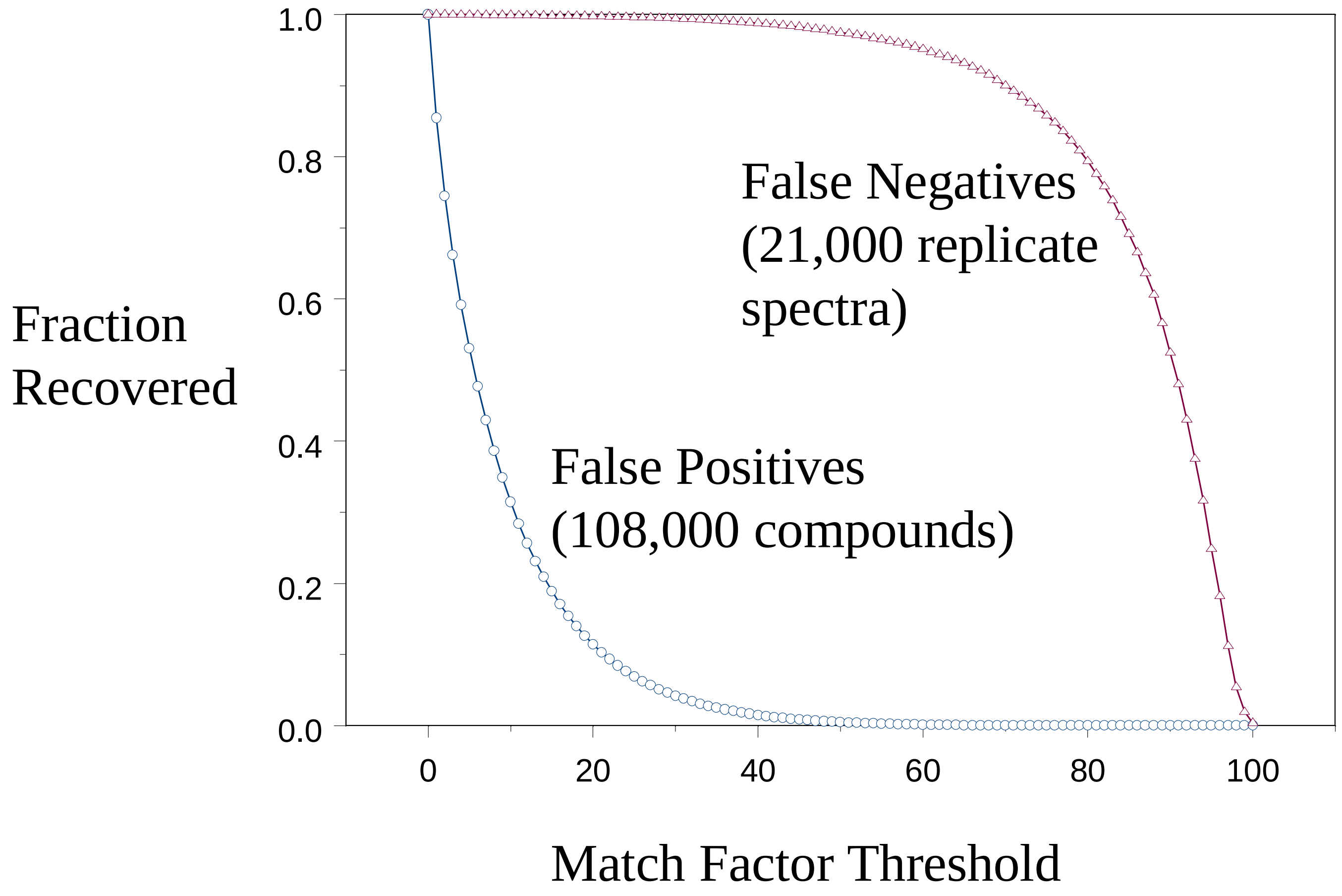
- M = ***f***(Abundance) Peak in Measured Spectrum
- R = ***f***(Abundance) Peak in Reference Spectrum
- Sum over all peaks
- ***f***(Abundance)
  - Abundance
  - Abundance * m/z
  - Certainty

Algorithm Performance

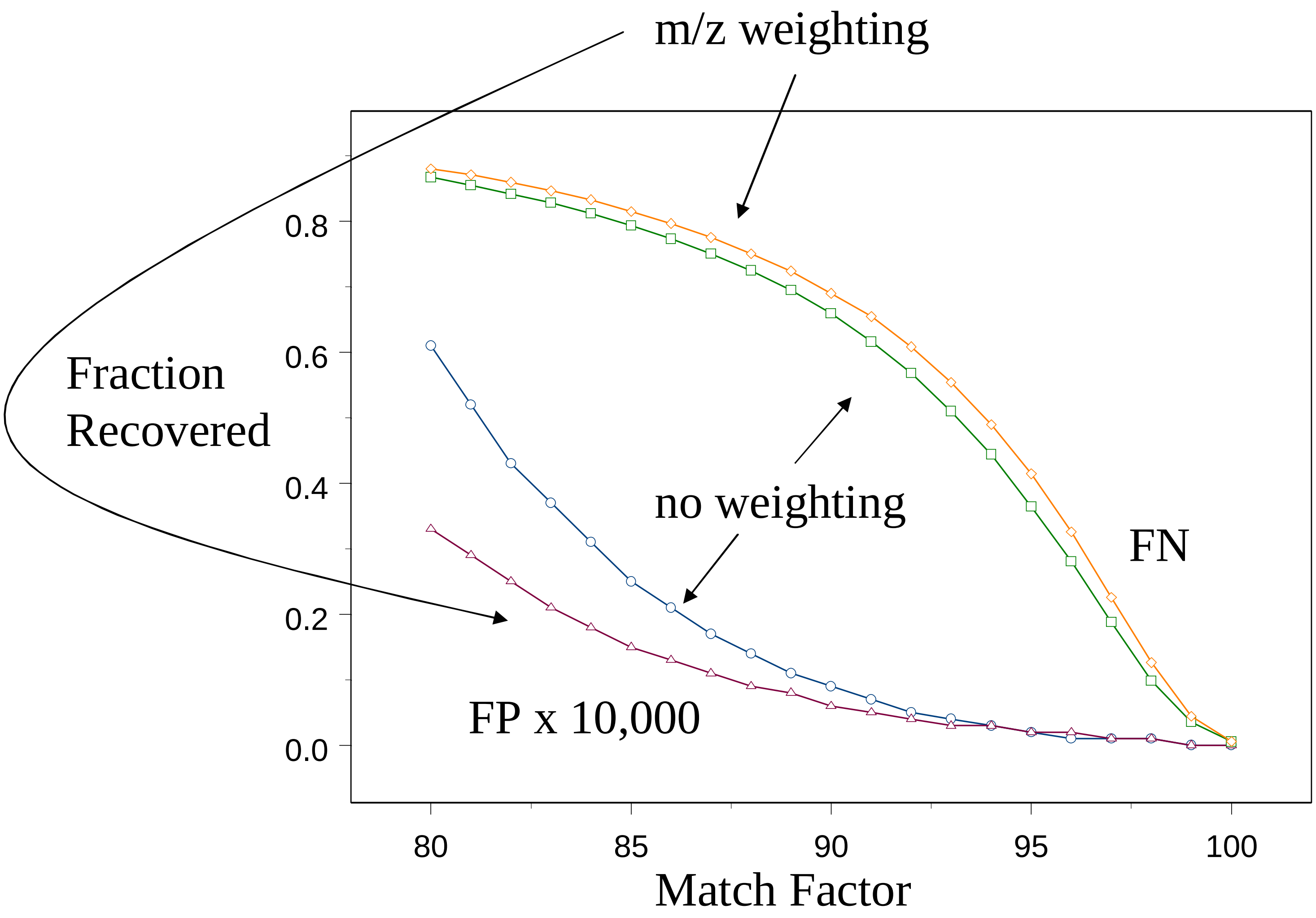
12,592 Replicate Spectra against NIST Library

Model	**Percent Correct**		
	Top Hit	**Top 2 Hits**	**Top 3 Hits**
Correlation – Weighted	74.9	86.9	91.7
Correlation	72.9	85.9	90.8
Euclidean Distance	71.9	83.9	88.9
Absolute Distance	67.9	80.3	85.5
PBM - Published	64.7	78.4	84.8
Hites/Hertz/Biemann	64.4	77.2	83.2

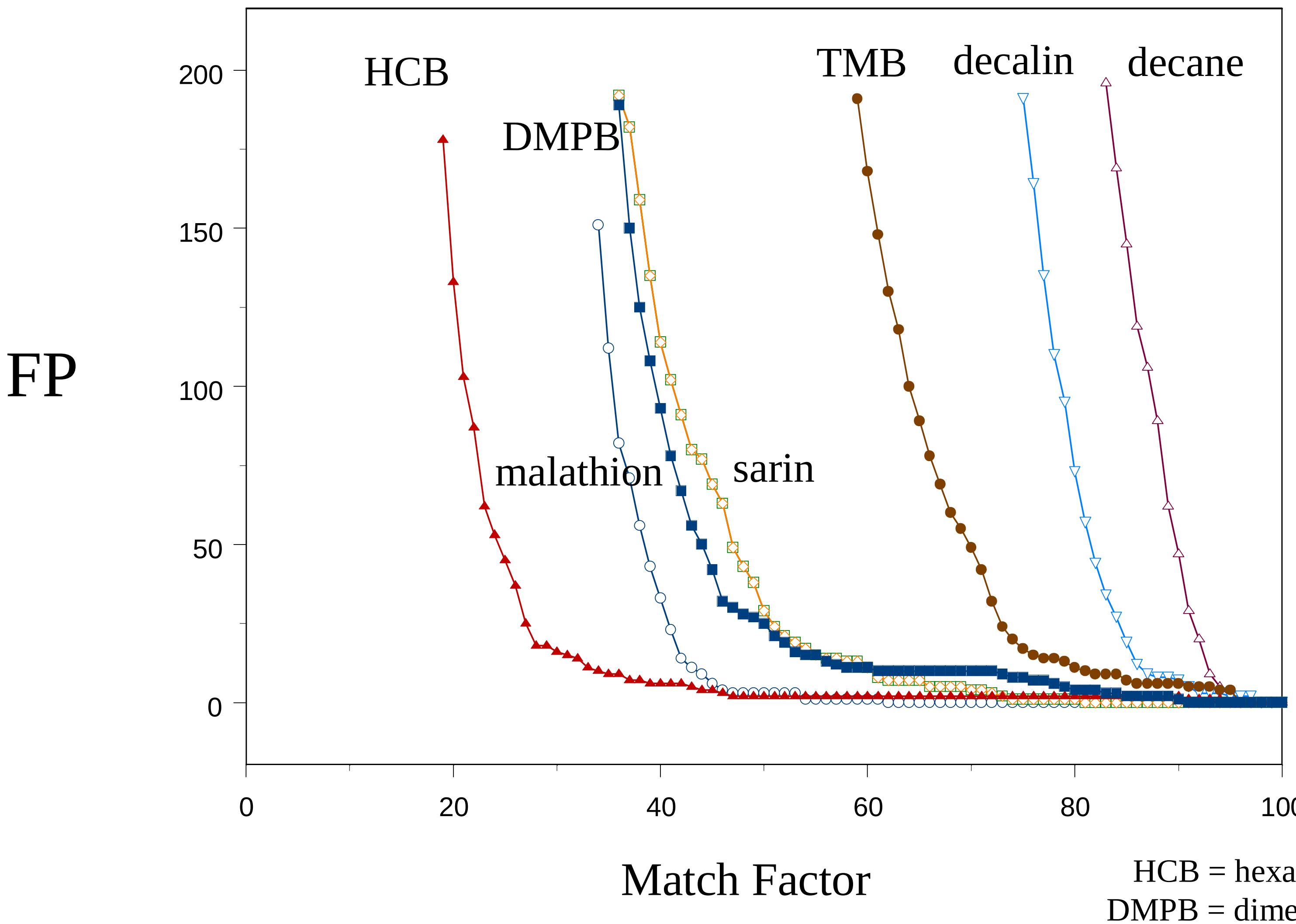
FP/FP Above Given Match Factor for NIST Library Spectra



FP/FN
Expanded View



FP Depends on Spectrum Uniqueness



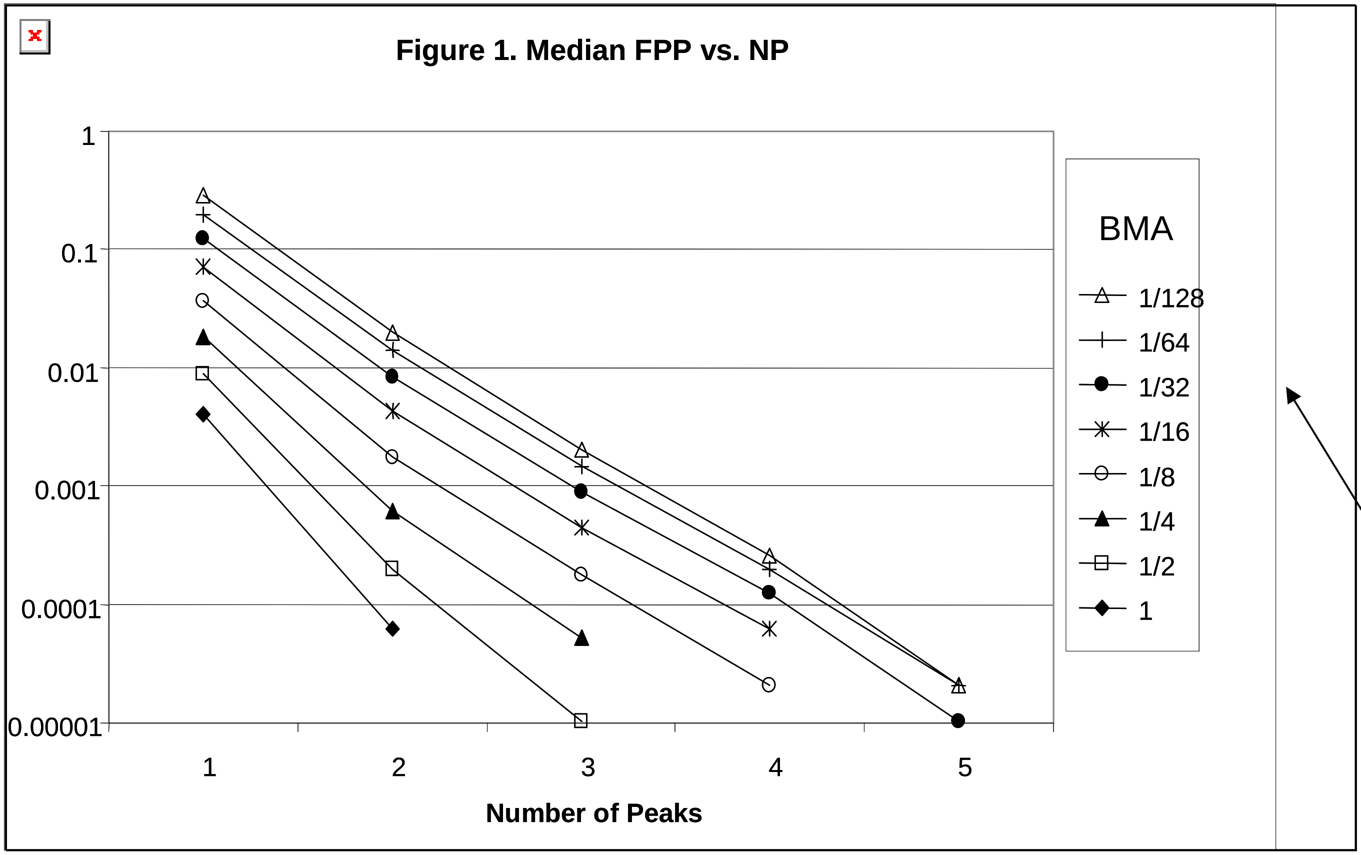
HCB = hexachlorobenzene
DMPB = dimethylpenobarbital
TMB = 1,2,3-trimethylbenzene

Multiple Ion Monitoring

- What is is?
  - Use 2-5 Major Peaks in Spectrum of Target
    - 10 – 100 more sensitive
- What's the problem?
  - Can match major Target peaks with Minor Sample Peaks
- What we have done:
  - Examine risk using library as source of potential false positive IDs

False Positive Risk vs Number of Peaks Used

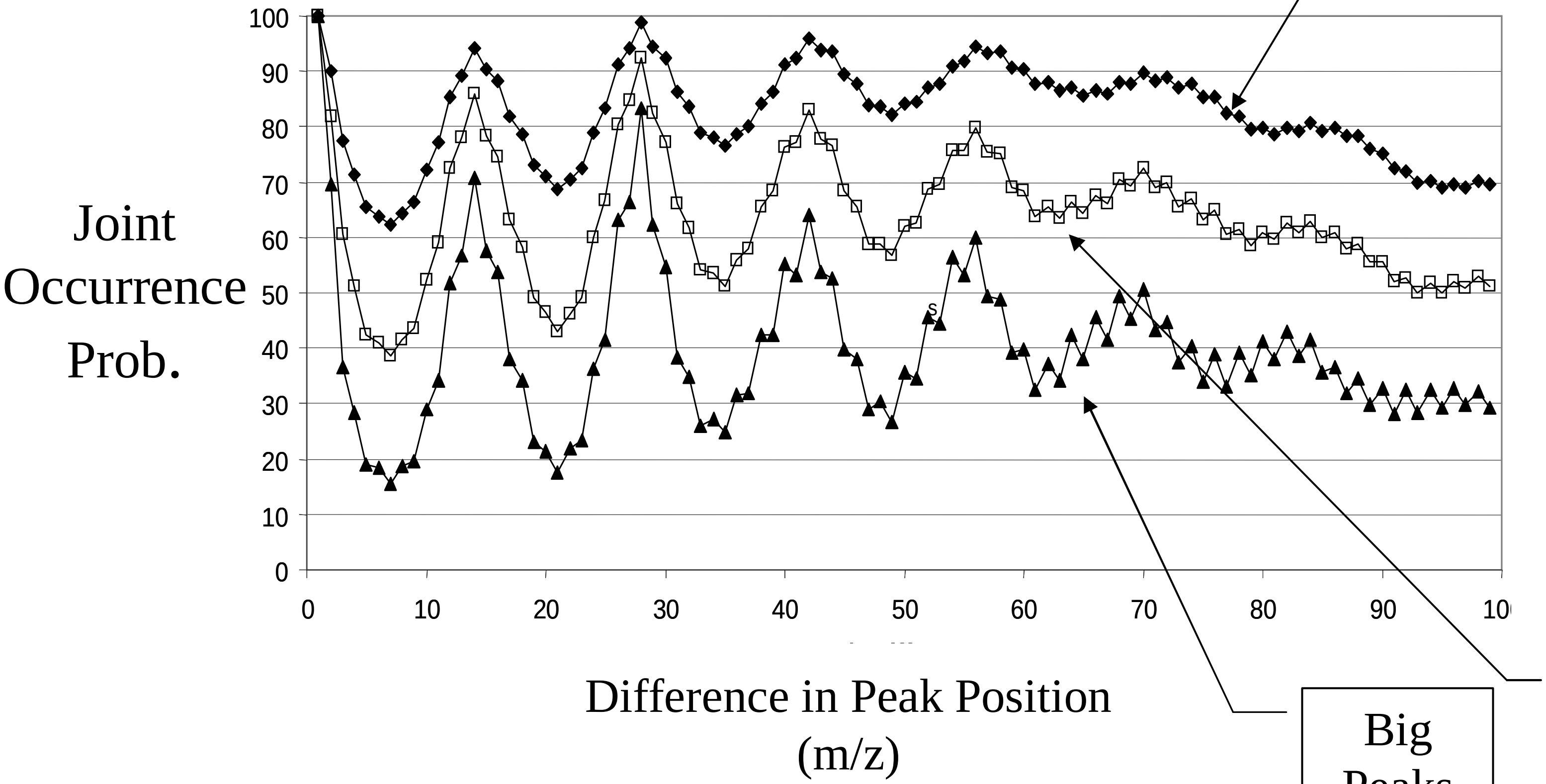
FP/
spectrum
(median)



Number of Peaks

Abundance Ratio:
Biggest Search Peak/
Matching Peak in FP

Mass Spectral Peak Occurrences are Correlated



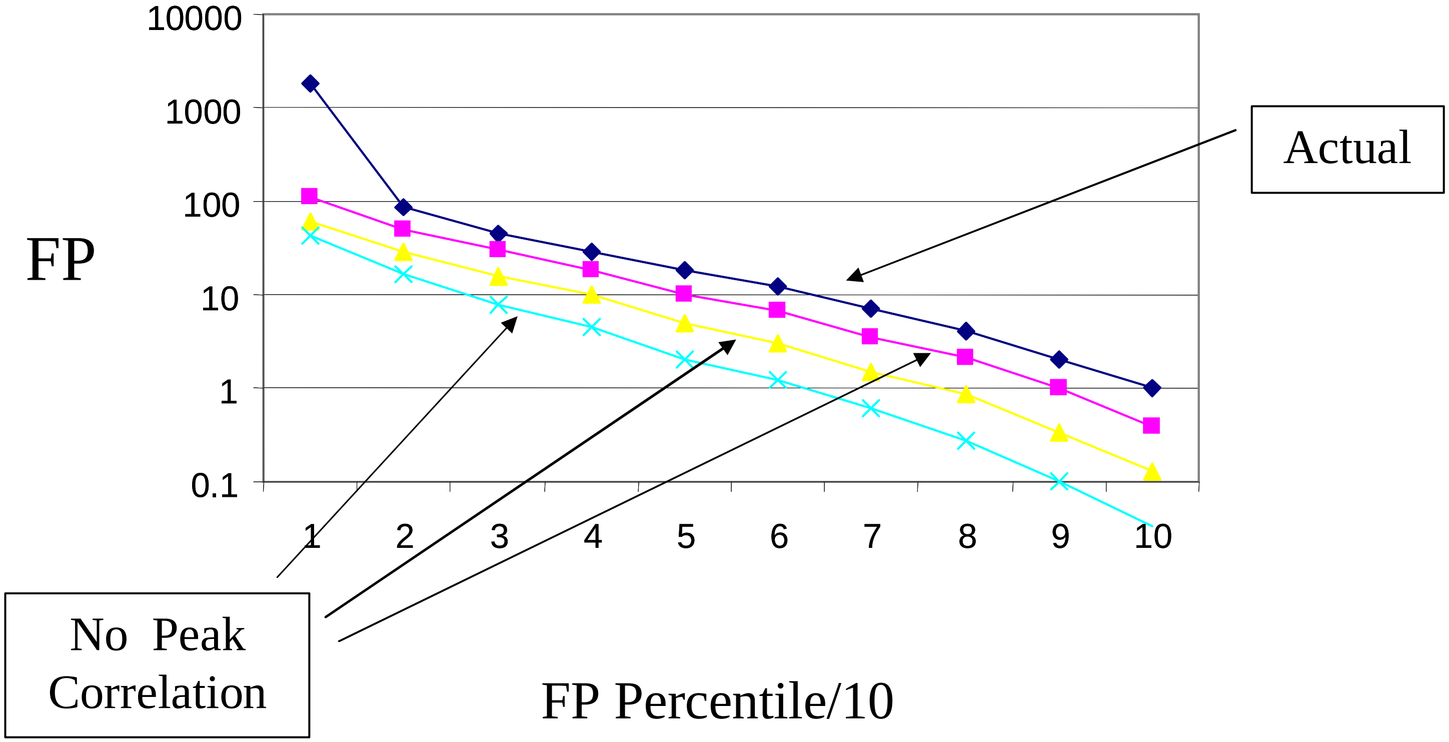
Small Peaks

Medium Peaks

Big Peaks

FP Observed and Computed

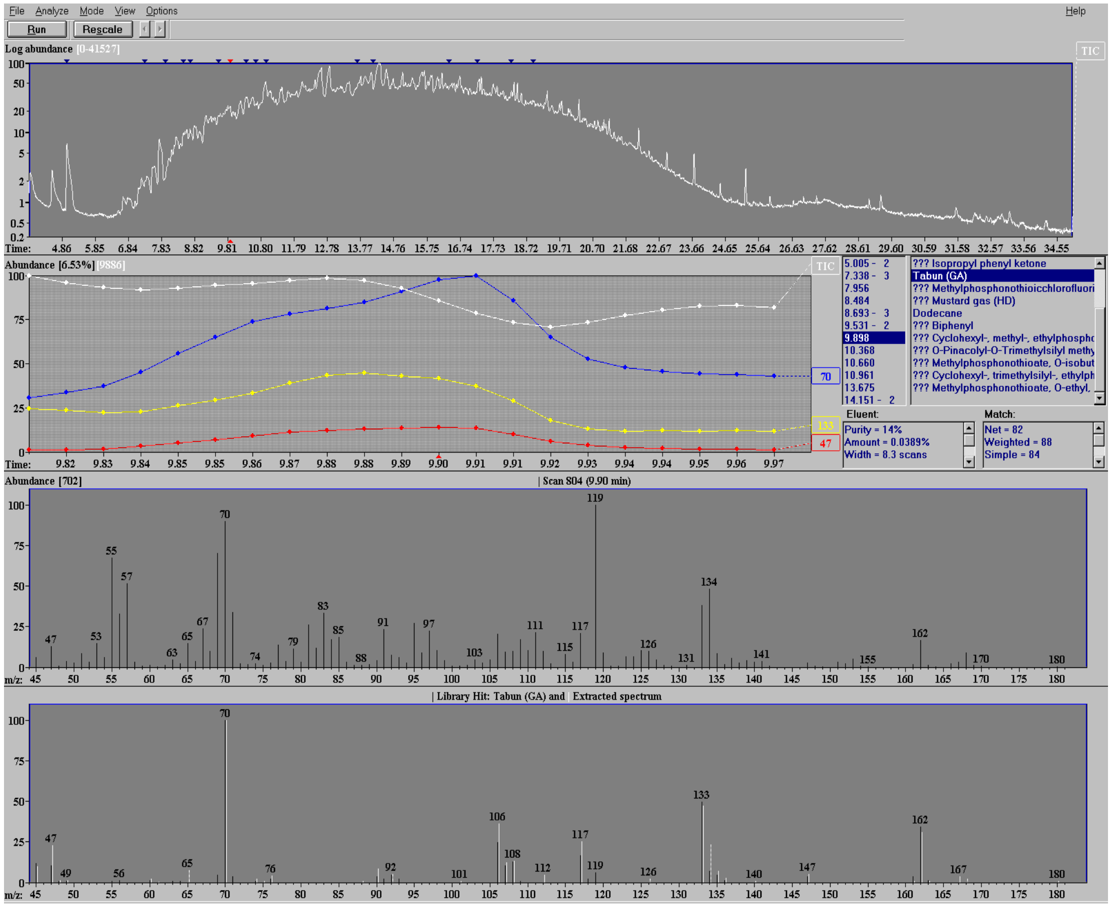
(from individual peak probabilities)



Actual

No Peak Correlation

Search Results Depend on Search Spectrum Quality


File Analyze Mode View Options
Help
Run
Rescale
Log abundance [0-41527]
TIC
100
50
20
10
5
2
1
0.5
0.2
Time: 4.86 5.85 6.84 7.83 8.82 9.81 10.80 11.79 12.78 13.77 14.76 15.75 16.74 17.73 18.72 19.71 20.70 21.68 22.67 23.66 24.65 25.64 26.63 27.62 28.61 29.60 30.59 31.58 32.57 33.56 34.55
Abundance [6.53%] [9886]
TIC
100
75
50
25
0
70
133
47
Time: 9.82 9.83 9.84 9.85 9.85 9.86 9.87 9.88 9.88 9.89 9.90 9.91 9.91 9.92 9.93 9.94 9.94 9.95 9.96 9.97
Abundance [702]
Scan 804 (9.90 min)
100
75
50
25
0
119
70
55
57
134
83
67
85
91
97
111
117
47
53
65
162
79
126
63
74
88
103
115
141
131
155
170
180
m/z: 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180
Library Hit: Tabun (GA) and Extracted spectrum
100
75
50
25
0
70
133
106
162
117
108
47
65
119
147
112
92
167
76
56
49
101
126
140
180
m/z: 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180

5.005 - 2
7.338 - 3
7.956
8.484
8.693 - 3
9.531 - 2
9.898
10.368
10.660
10.961
13.675
14.151 - 2

??? Isopropyl phenyl ketone
Tabun (GA)
??? Methylphosphonothioicchlorofluori
??? Mustard gas (HD)
Dodecane
??? Biphenyl
??? Cyclohexyl-, methyl-, ethylphospho
??? O-Pinacolyl-O-Trimethylsilyl methy
??? Methylphosphonothioate, O-isobut
??? Cyclohexyl-, trimethylsilyl-, ethylph
??? Methylphosphonothioate, O-ethyl,

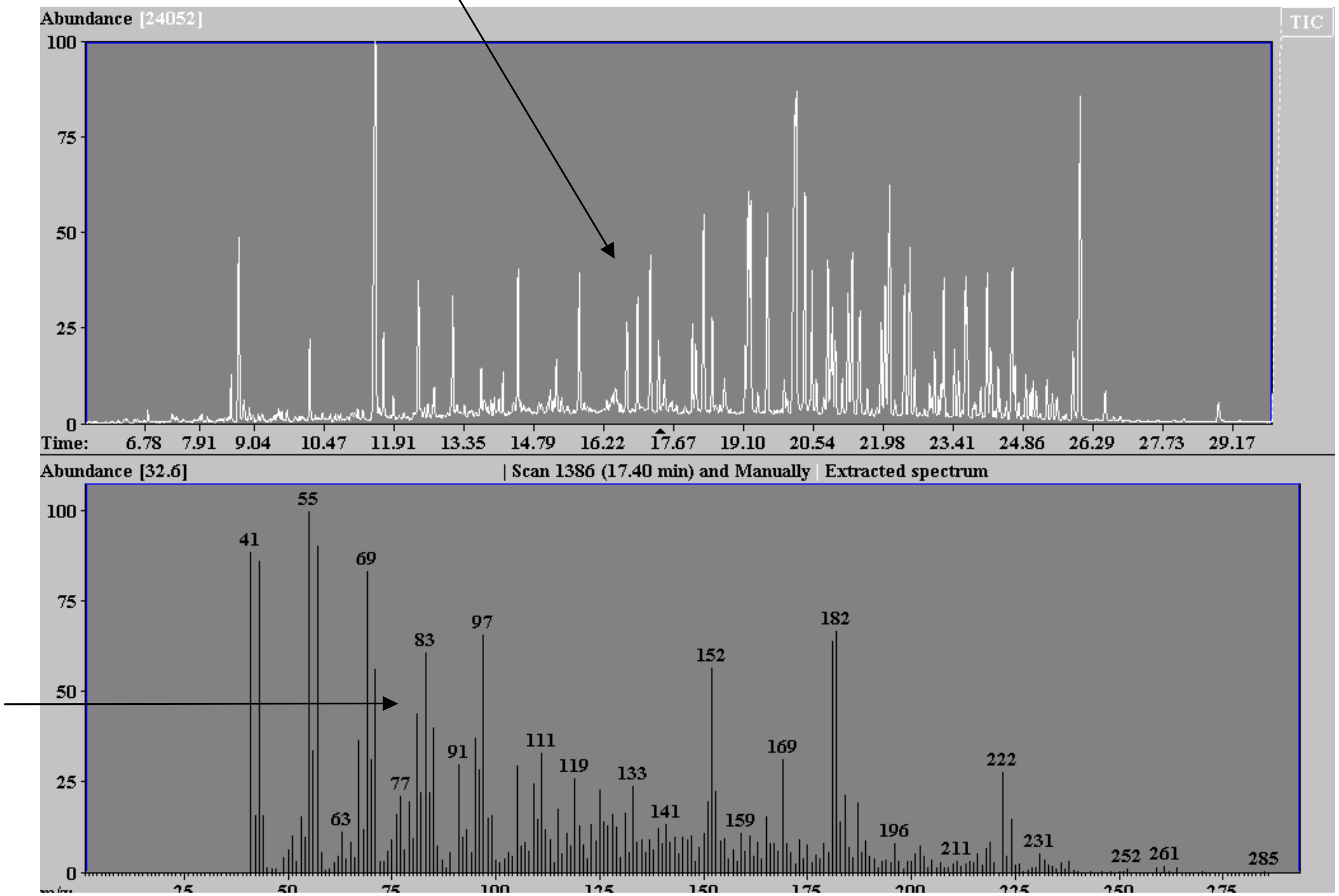
Eluent:
Purity = 14%
Amount = 0.0389%
Width = 8.3 scans

Match:
Net = 82
Weighted = 88
Simple = 84

AMDIS:
http://chemdata.nist.gov

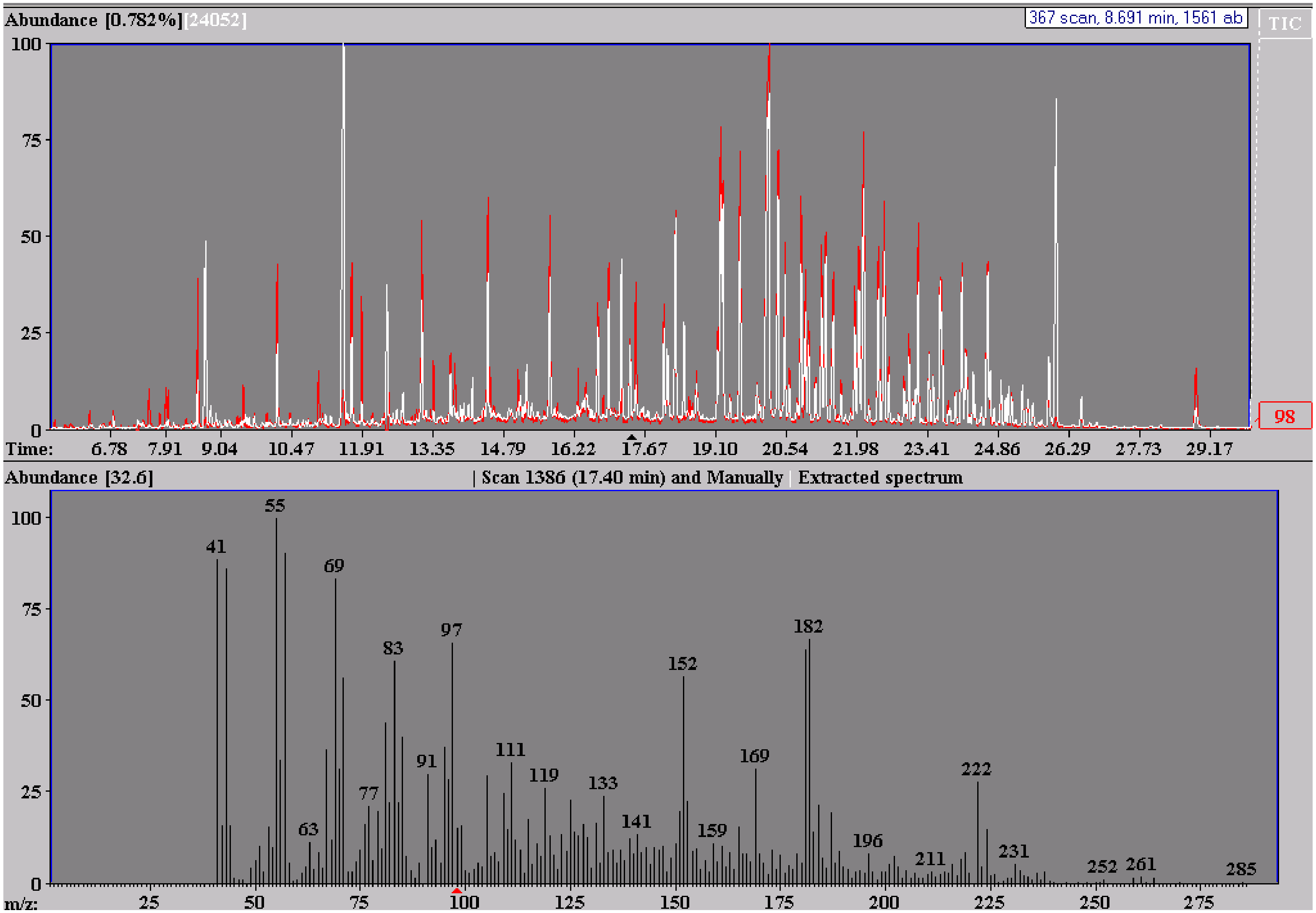
Real Data

Total ion chromatogram

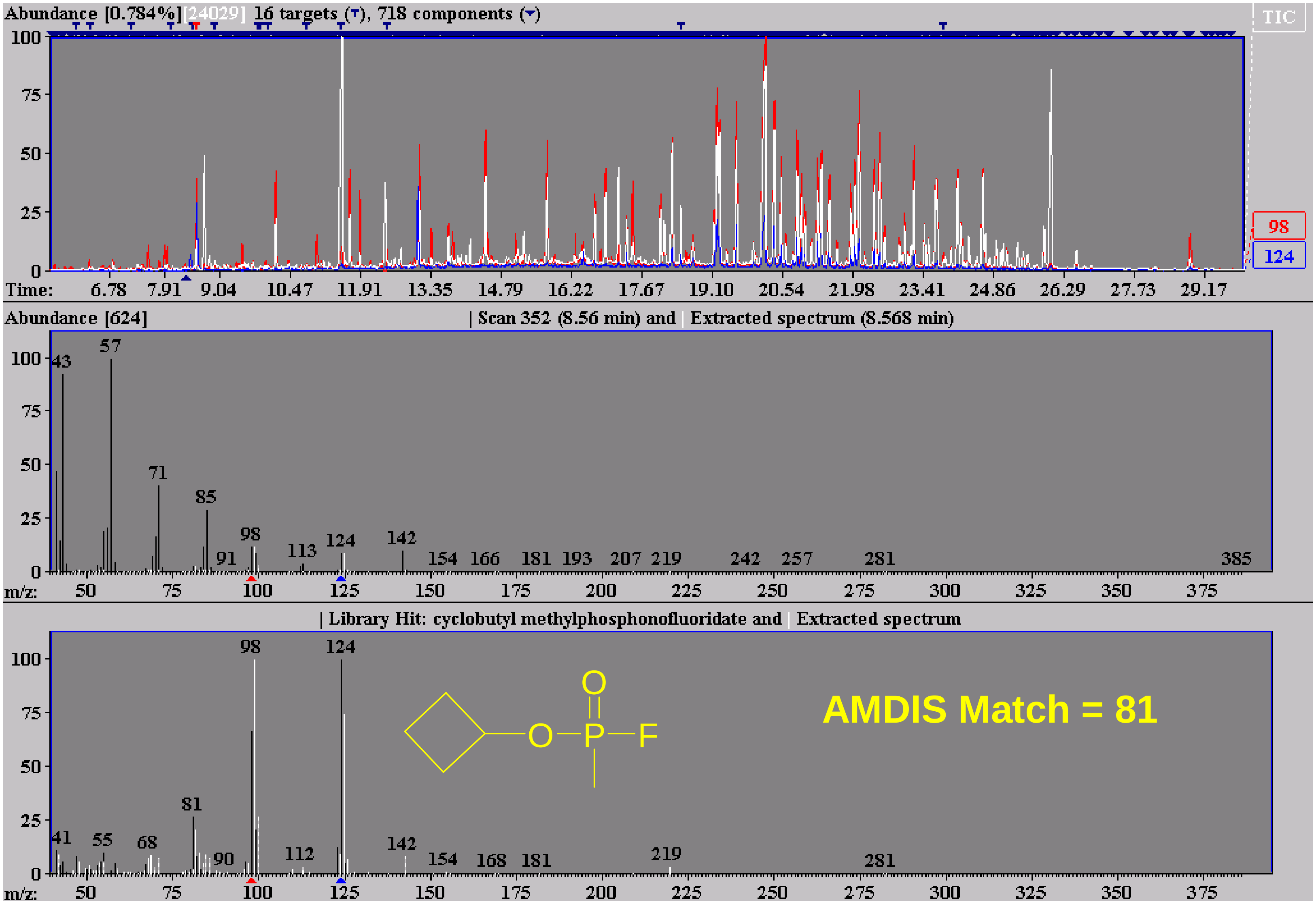


A mass spectrum (scan)

Chromatogram with single ion



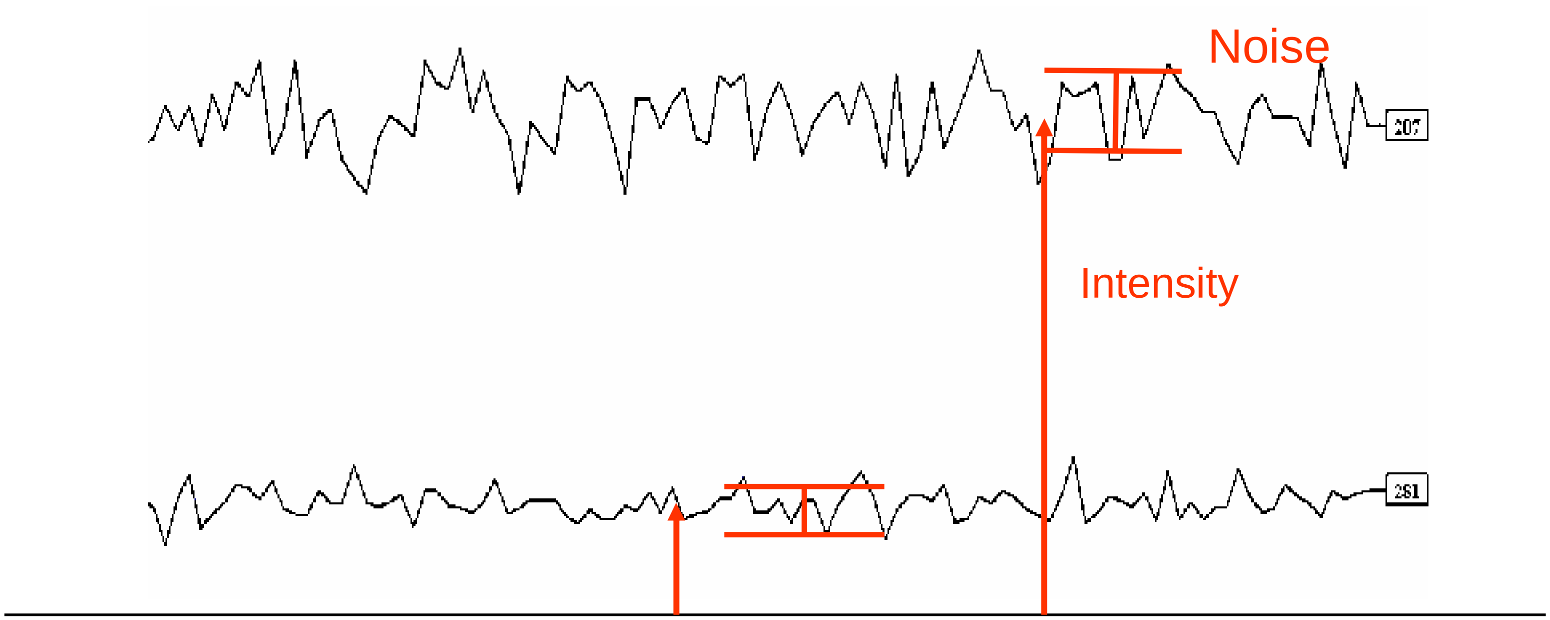
AMDIS Analysis of Data


Abundance [0.784%] [24029] 16 targets (▼), 718 components (▼)
TIC
100
75
50
25
0
Time: 6.78 7.91 9.04 10.47 11.91 13.35 14.79 16.22 17.67 19.10 20.54 21.98 23.41 24.86 26.29 27.73 29.17
98
124
Abundance [624]
| Scan 352 (8.56 min) and | Extracted spectrum (8.568 min)
100
75
50
25
0
43
57
71
85
91
98
113
124
142
154 166 181 193 207 219 242 257 281 385
m/z: 50 75 100 125 150 175 200 225 250 275 300 325 350 375
| Library Hit: cyclobutyl methylphosphonofluoridate and | Extracted spectrum
100
75
50
25
0
98
124
41 55 68 81 90 112 142 154 168 181 219 281
m/z: 50 75 100 125 150 175 200 225 250 275 300 325 350 375
O
O
P
F
AMDIS Match = 81

Order of Analysis

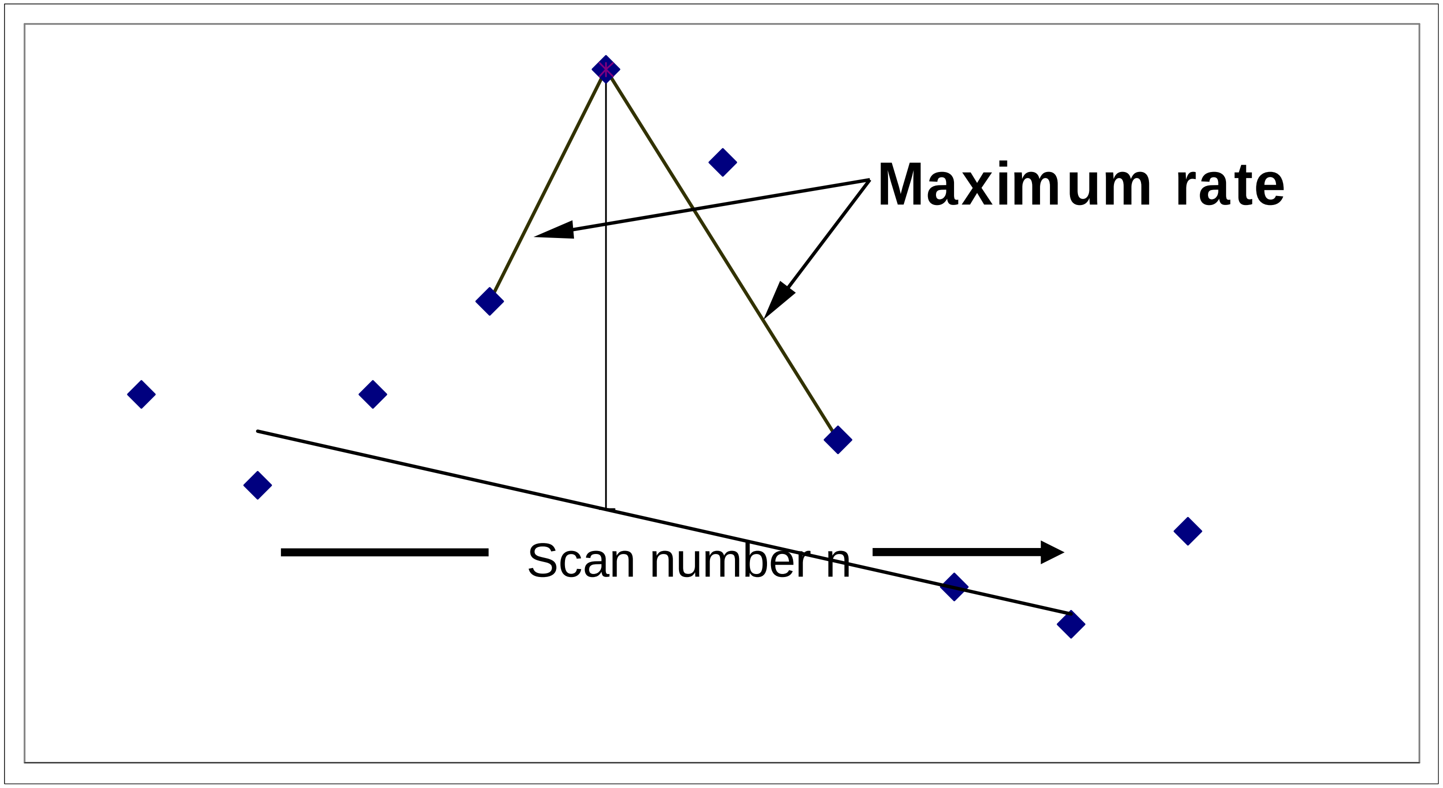
- Noise Analysis – find ‘Noise Factor’
- Find and quantify maximizing ions
- Combine to create ‘Model Peak’
- Use Model Peak shape (intensity vs time) to purify spectra
- Find best matching library spectrum

Derive Noise Factor

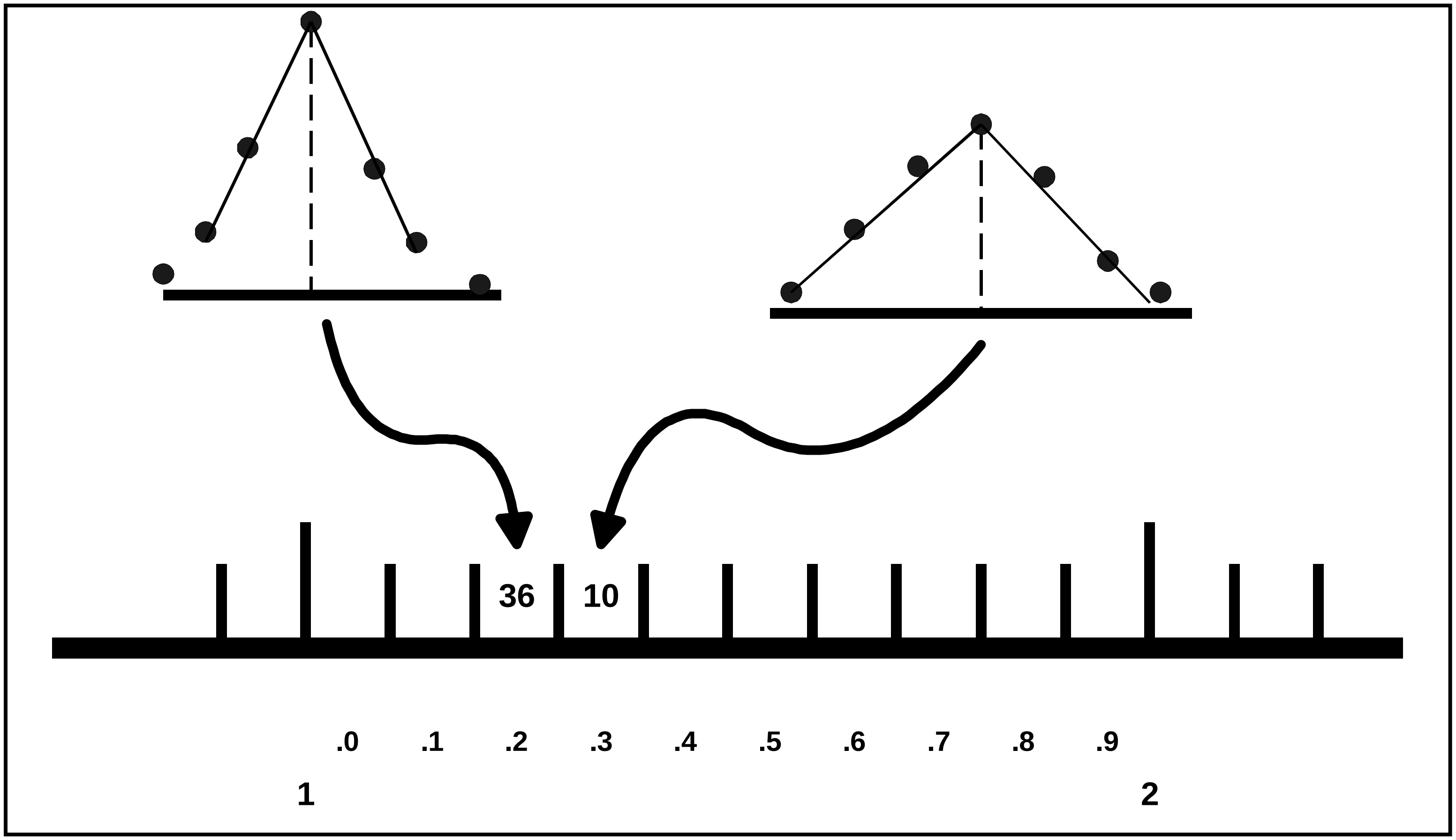


$$Noise = K_{noise}\sqrt{Intensity}$$

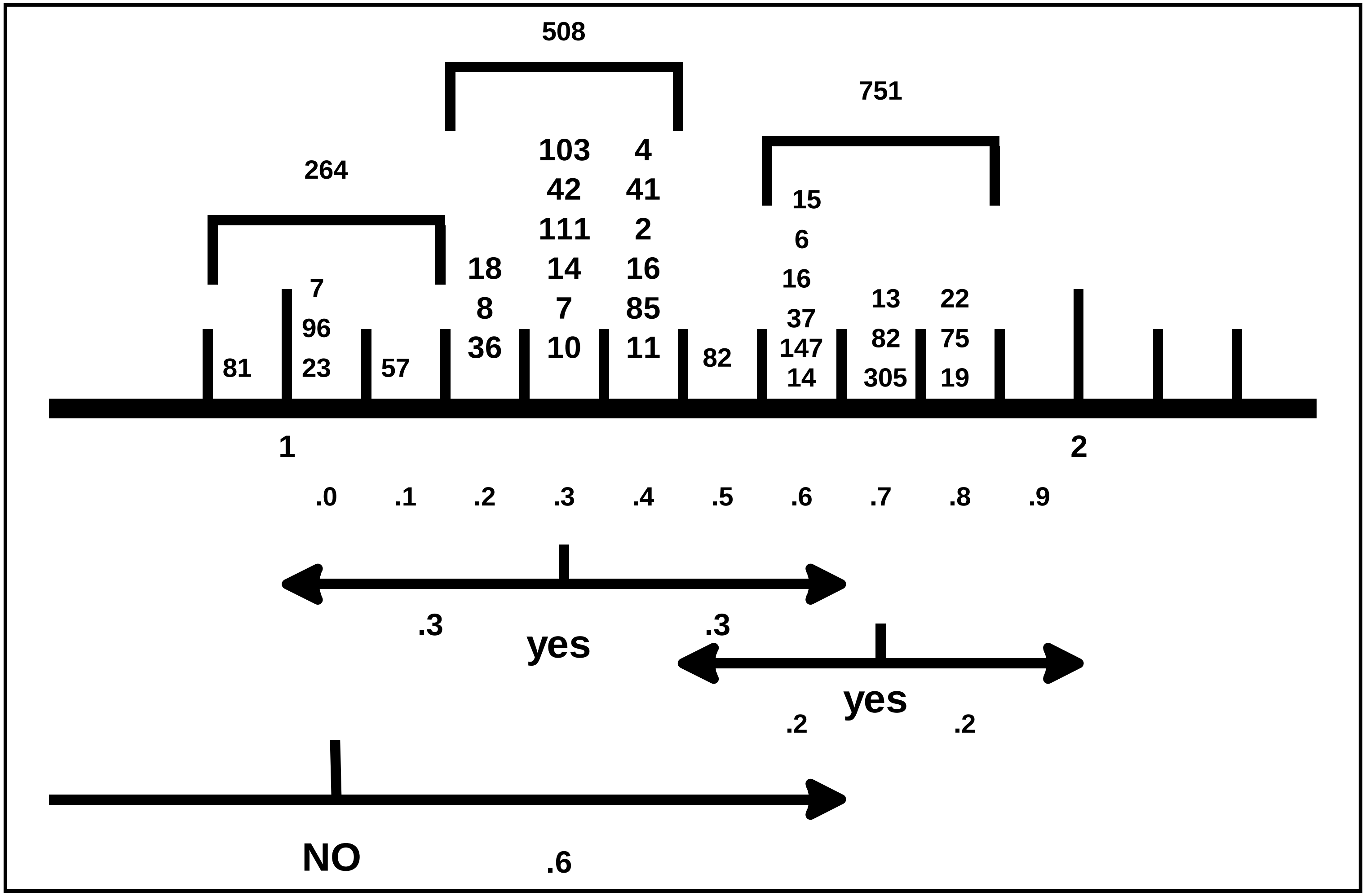
Finding Possible Peaks for Each m/z



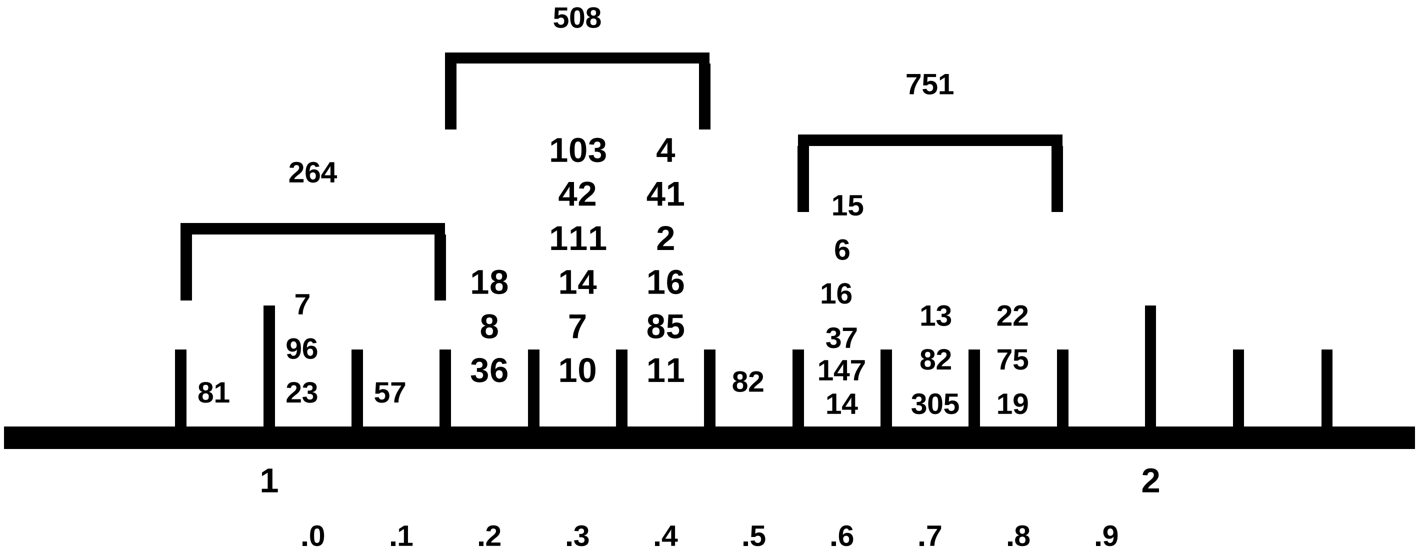
Find Possible Compounds:
Do Ions Maximize at Same Time?



Separate the Components

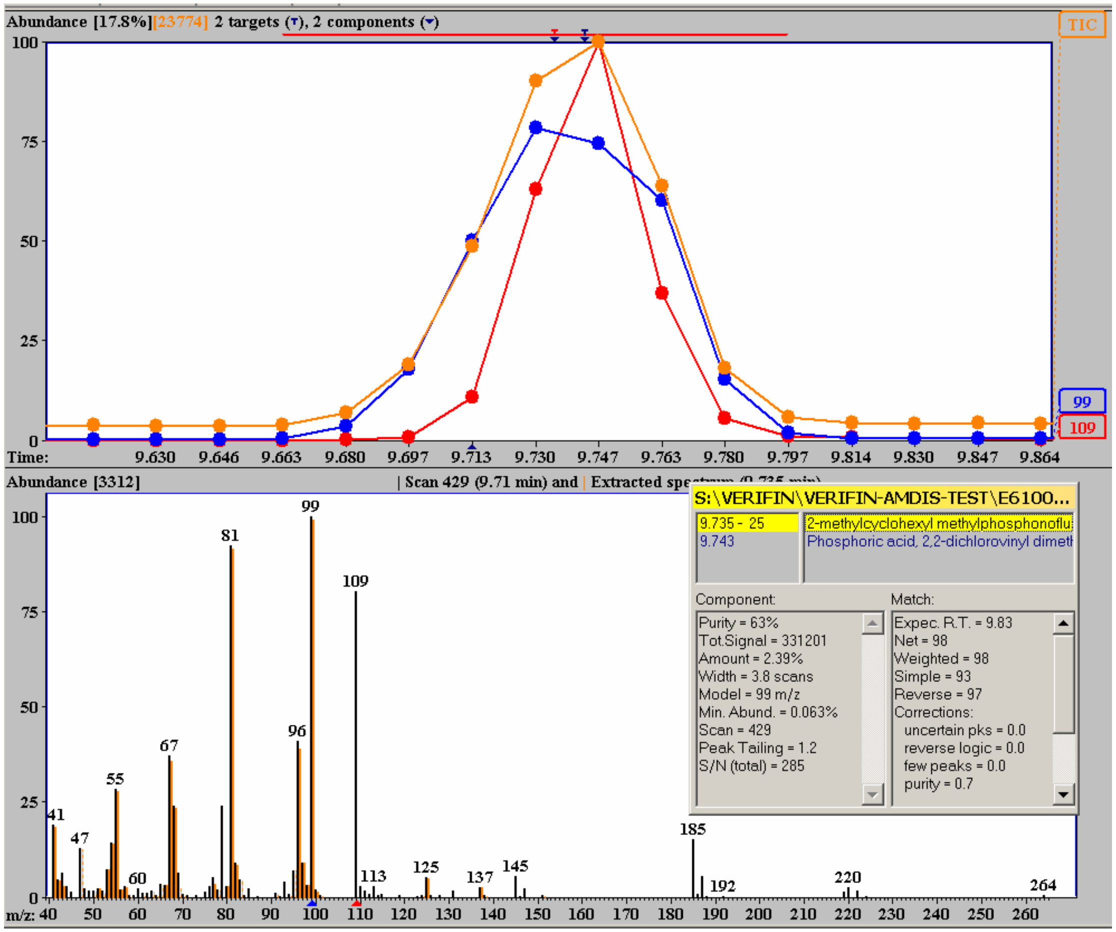


A ‘Model Peak’ Provides Shape



The model shape is defined as the sum of all of the ion chromatograms that maximize within the range and have a sharpness value within 75% of the maximum.

AMDIS Testing – Closely Eluting Components



S:\VERIFIN\VERIFIN-AMDIS-TEST\E6100...

9.735 - 25	2-methylcyclohexyl methylphosphonoflu
9.743	Phosphoric acid, 2,2-dichlorovinyl dimetl

Component:

Purity = 63%
Tot.Signal = 331201
Amount = 2.39%
Width = 3.8 scans
Model = 99 m/z
Min. Abund. = 0.063%
Scan = 429
Peak Tailing = 1.2
S/N (total) = 285

Match:

Expec. R.T. = 9.83
Net = 98
Weighted = 98
Simple = 93
Reverse = 97
Corrections:
uncertain pks = 0.0
reverse logic = 0.0
few peaks = 0.0
purity = 0.7

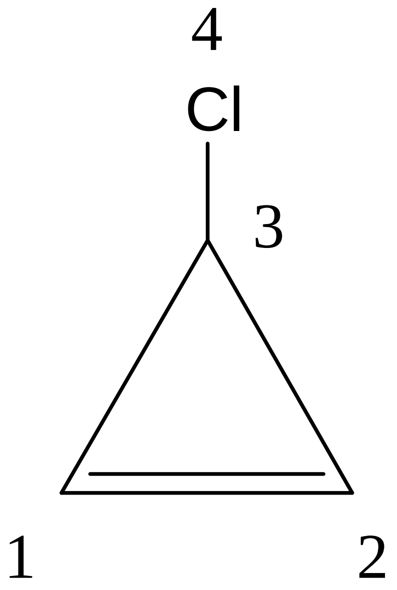
Representing Chemical Identity

- Visual: 2D Structure
- Text: IUPAC Name
- Digital: No Accepted, Open Method

- Solution:

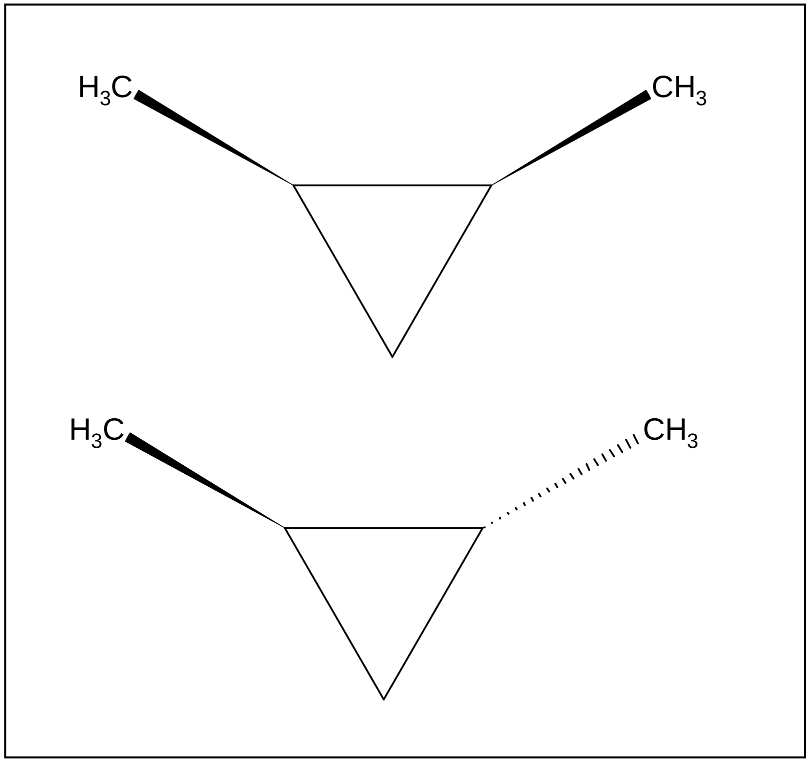
The IUPAC/NIST Chemical Identifier

Connection Table



	1	2	3	4
1		D	S	
2	D		S	
3	S	S		S
4				S

Chemical Identity Problems



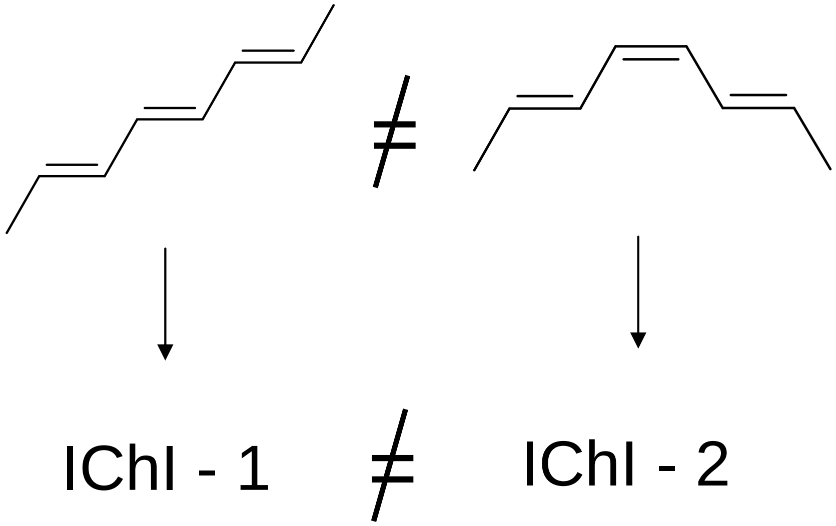
Registry Number possible for each exact form, mixture, unknown, unspecified

Experts required

Expensive, ambiguous and error prone

Requirements

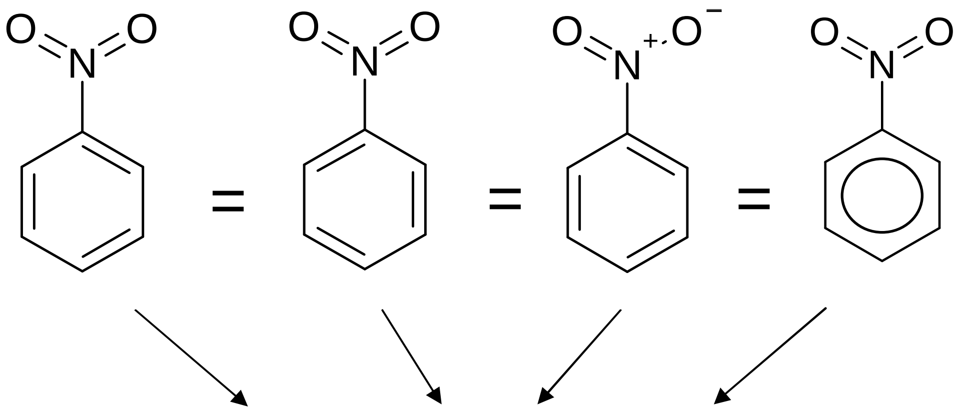
- Different compounds have different identifiers
  - Keep all distinguishing structural information



IChI - 1 ≠ IChI - 2

Requirements

- One compound has only one identifier
  - Omit unnecessary information



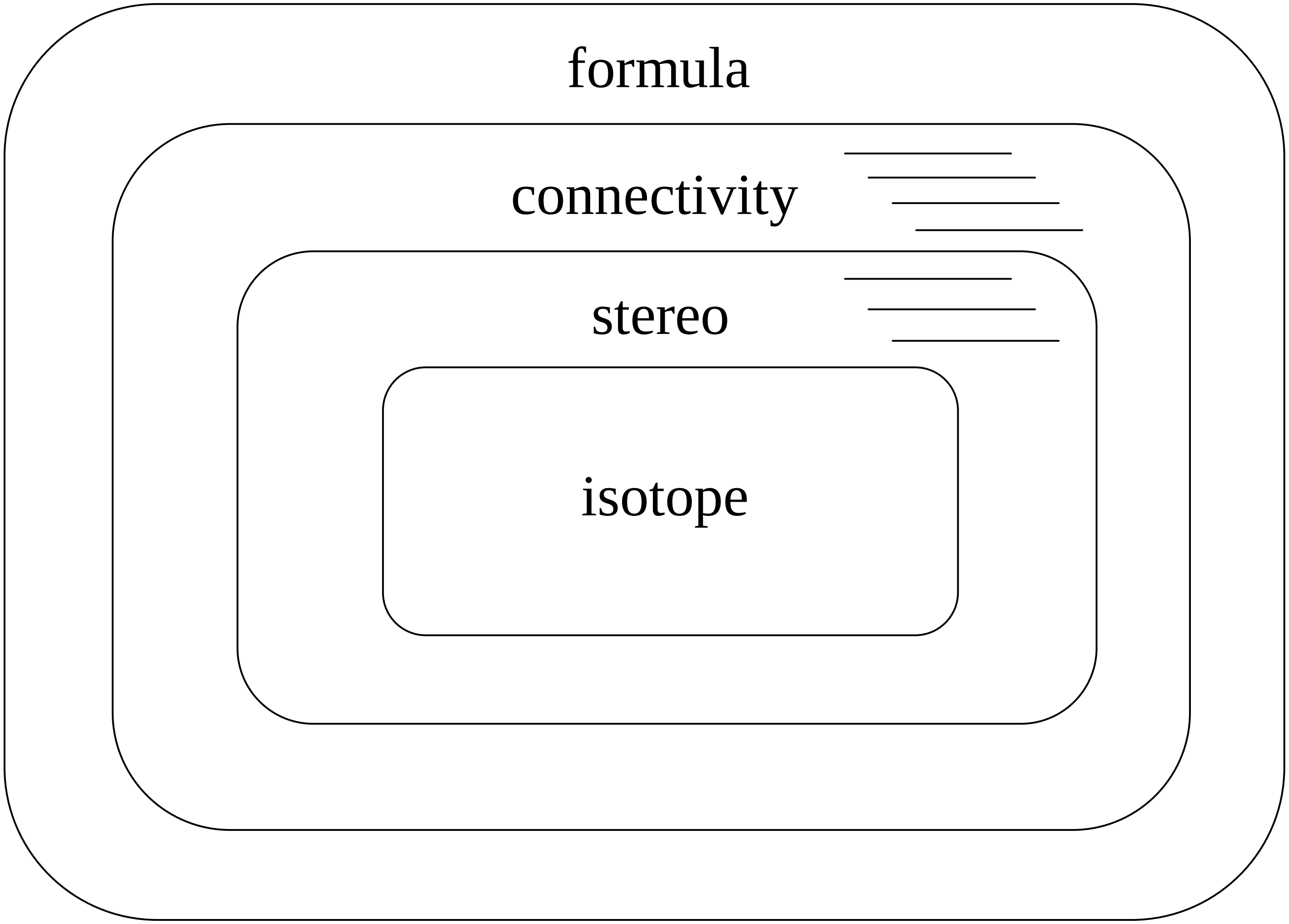
Same INChI

3 Steps to INChI

- Chemistry
  - ‘Normalize’ Input Structure
    - Implement chemical rules
- Math
  - ‘Canonicalize’ (label the atoms)
    - Equivalent atoms get the same label
- Format
  - ‘Serialize’ Labeled Structure
    - Output as character string (‘name’)

“Layers”

Chemical Substances



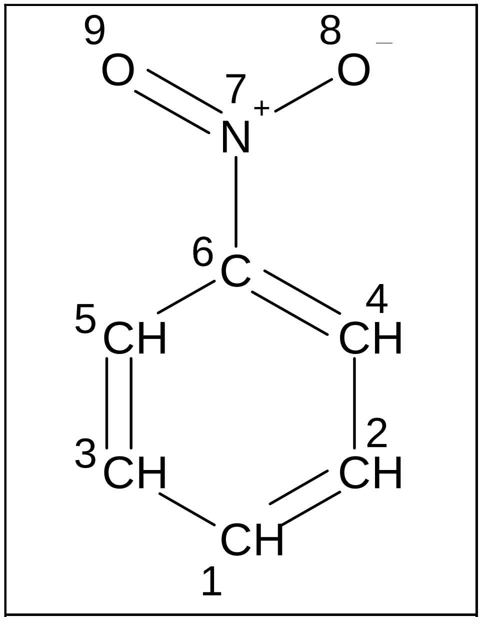
formula

connectivity

stereo

isotope

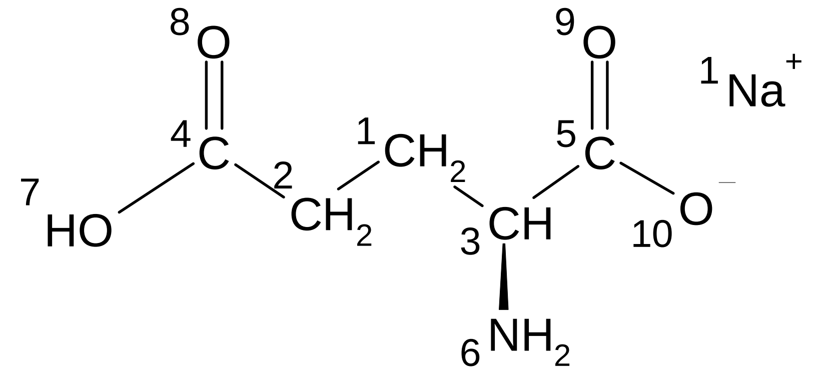
Nitrobenzene



Canonical numbering

Description	Layers
formula	C6H5NO2
connectivity	8-7(9)6-4-2-1-3-5-6
H-atoms	1-5H
charges	

MSG



Canonical numbering

Description	Layers
formula	**C5H8NO4.Na**
connectivity	**6-3(5(9)10)1-2-4(7)8;**
H-atoms	**1-2H2,3H,6H2(H-,7,8,9,10);**
stereo sp^3	**3-;**
charges	**-1;+1**

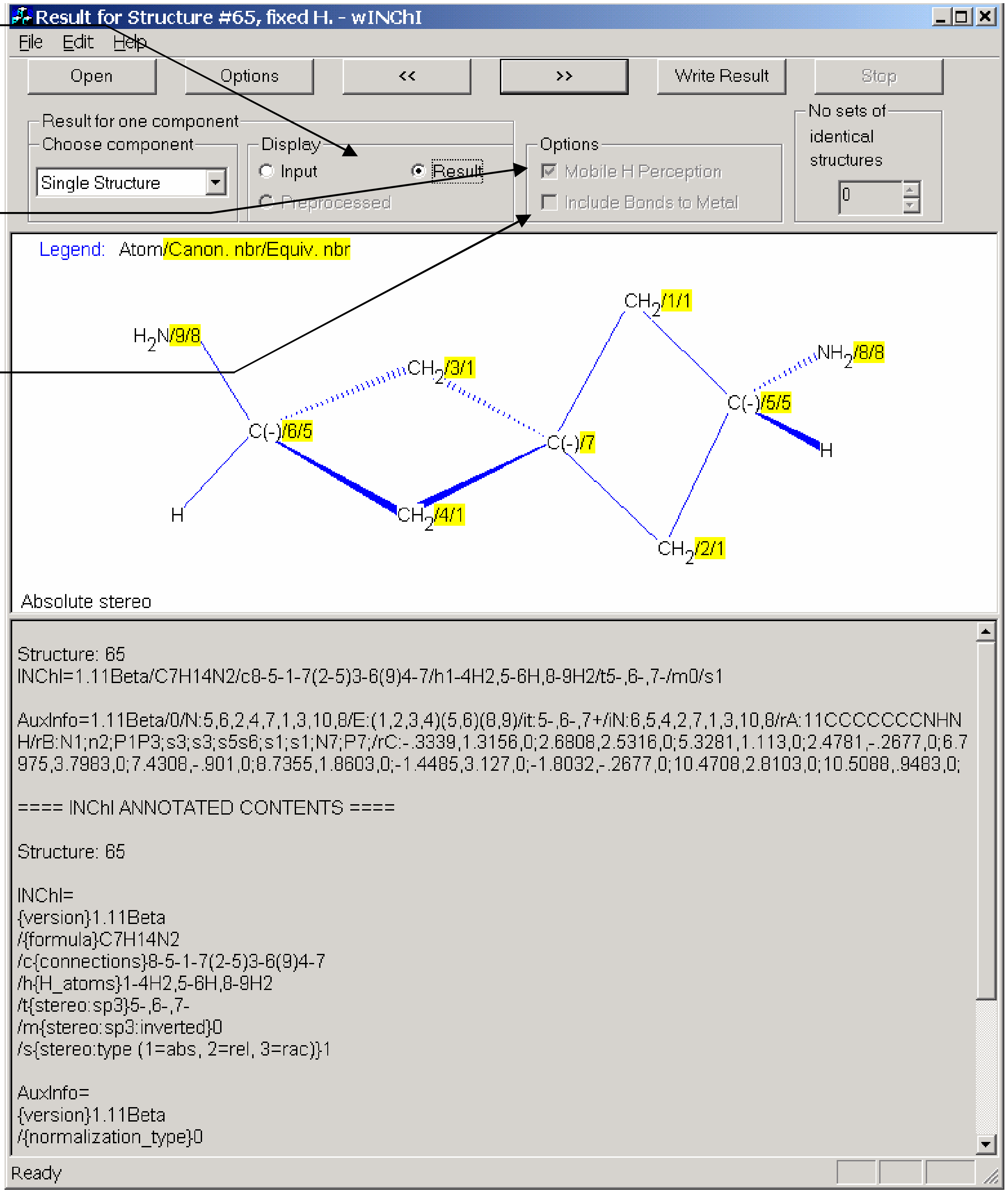
C5H9NO4.Na/c6-3(5(9)10)1-2-4(7)8;/h1-2H2,3H,6H2,(H,7,8)(H,9,10);/q;+1/p-1/t3-;/m1./s1

INChI Test Version

Input/ Result

Mobile H On/Off

Include Org-Metal Bonds

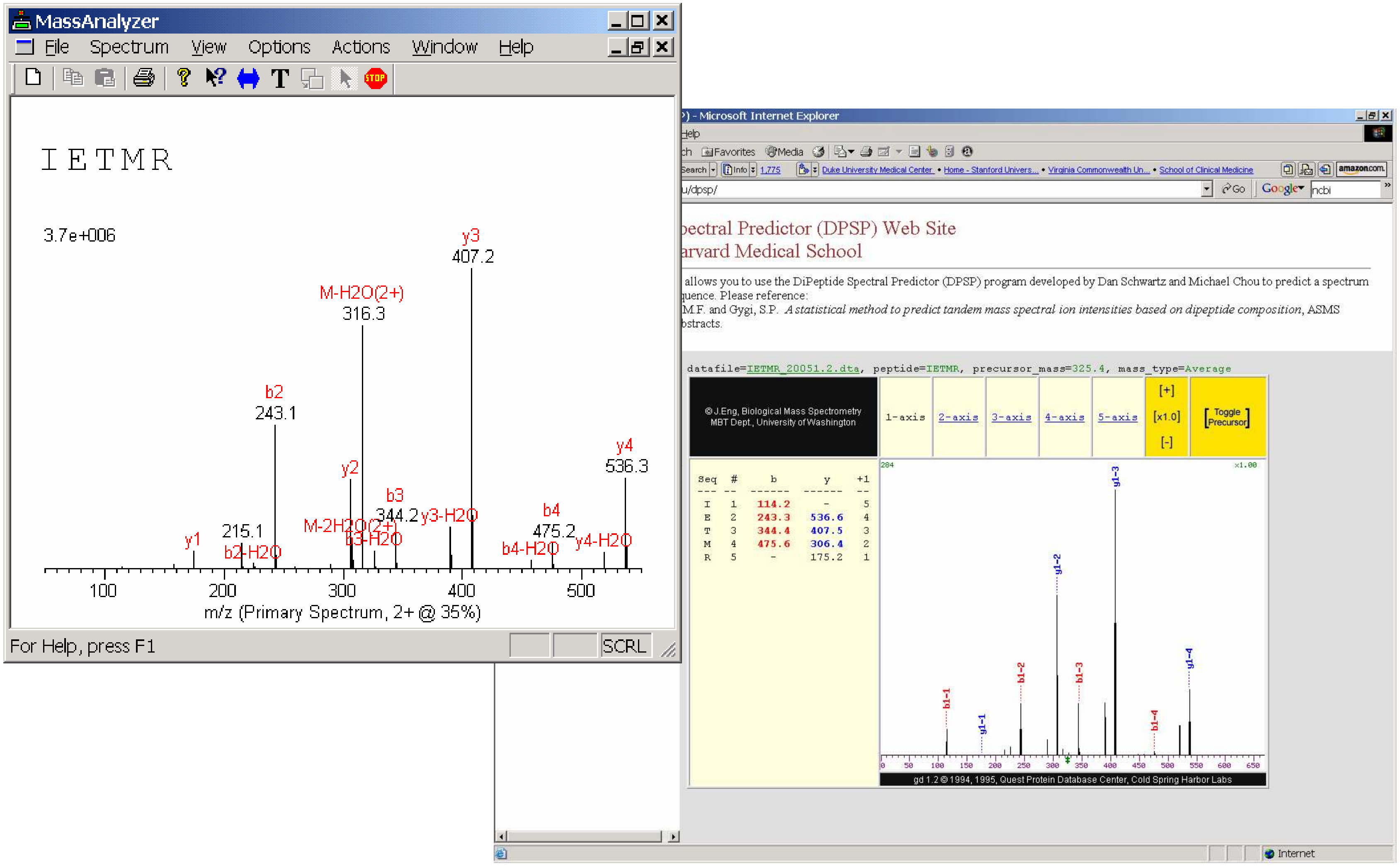


```
Structure: 65
INChI=1.11Beta/C7H14N2/c8-5-1-7(2-5)3-6(9)4-7/h1-4H2,5-6H,8-9H2/t5-,6-,7-/m0/s1

AuxInfo=1.11Beta/0/N:5,6,2,4,7,1,3,10,8/E:(1,2,3,4)(5,6)(8,9)/it:5-,6-,7+/iN:6,5,4,2,7,1,3,10,8/rA:11CCCCCCCNHN
H/rB:N1;n2;P1P3;s3;s3;s5s6;s1;s1;N7;P7;/rC:-.3339,1.3156,0;2.6808,2.5316,0;5.3281,1.113,0;2.4781,-.2677,0;6.7
975,3.7983,0;7.4308,-.901,0;8.7355,1.8603,0;-1.4485,3.127,0;-1.8032,-.2677,0;10.4708,2.8103,0;10.5088,.9483,0;

==== INChI ANNOTATED CONTENTS ====

Structure: 65

INChI=
{version}1.11Beta
/{formula}C7H14N2
/c{connections}8-5-1-7(2-5)3-6(9)4-7
/h{H_atoms}1-4H2,5-6H,8-9H2
/t{stereo:sp3}5-,6-,7-
/m{stereo:sp3:inverted}0
/s{stereo:type (1=abs, 2=rel, 3=rac)}1

AuxInfo=
{version}1.11Beta
/{normalization_type}0
```

Peptide Mass Spectra: Libraries for Organisms

- Proteins are linear sequences of amino acids
  - characteristic of Genome (organism)
- Peptides are ‘digested’ fragments of proteins
- MS ‘sequences’ peptides to reveal source Protein
- Peptides fragmentation spectra are not quite predictable
- Peptide fragmentation spectra for a ‘genome’ can be contained in one Library.

Spectrum Prediction Programs



MassAnalyzer
File Spectrum View Options Actions Window Help
IETMR
3.7e+006
y3
407.2
M-H2O(2+)
316.3
b2
243.1
y2
b3
344.2
y3-H2O
M-2H2O(2+)
b3-H2O
y1
215.1
b2-H2O
b4
475.2
b4-H2O
y4
536.3
y4-H2O
100
200
300
400
500
m/z (Primary Spectrum, 2+ @ 35%)
For Help, press F1
SCRL
) - Microsoft Internet Explorer
Help
ch Favorites Media
Search Info 1,775 Duke University Medical Center • Home - Stanford Univers... • Virginia Commonwealth Un... • School of Clinical Medicine
u/dpsp/
Go
Google
ncbi
Internet

pectral Predictor (DPSP) Web Site
arvard Medical School

allows you to use the DiPeptide Spectral Predictor (DPSP) program developed by Dan Schwartz and Michael Chou to predict a spectrum
quence. Please reference:
M.F. and Gygi, S.P. *A statistical method to predict tandem mass spectral ion intensities based on dipeptide composition*, ASMS
bstracts.

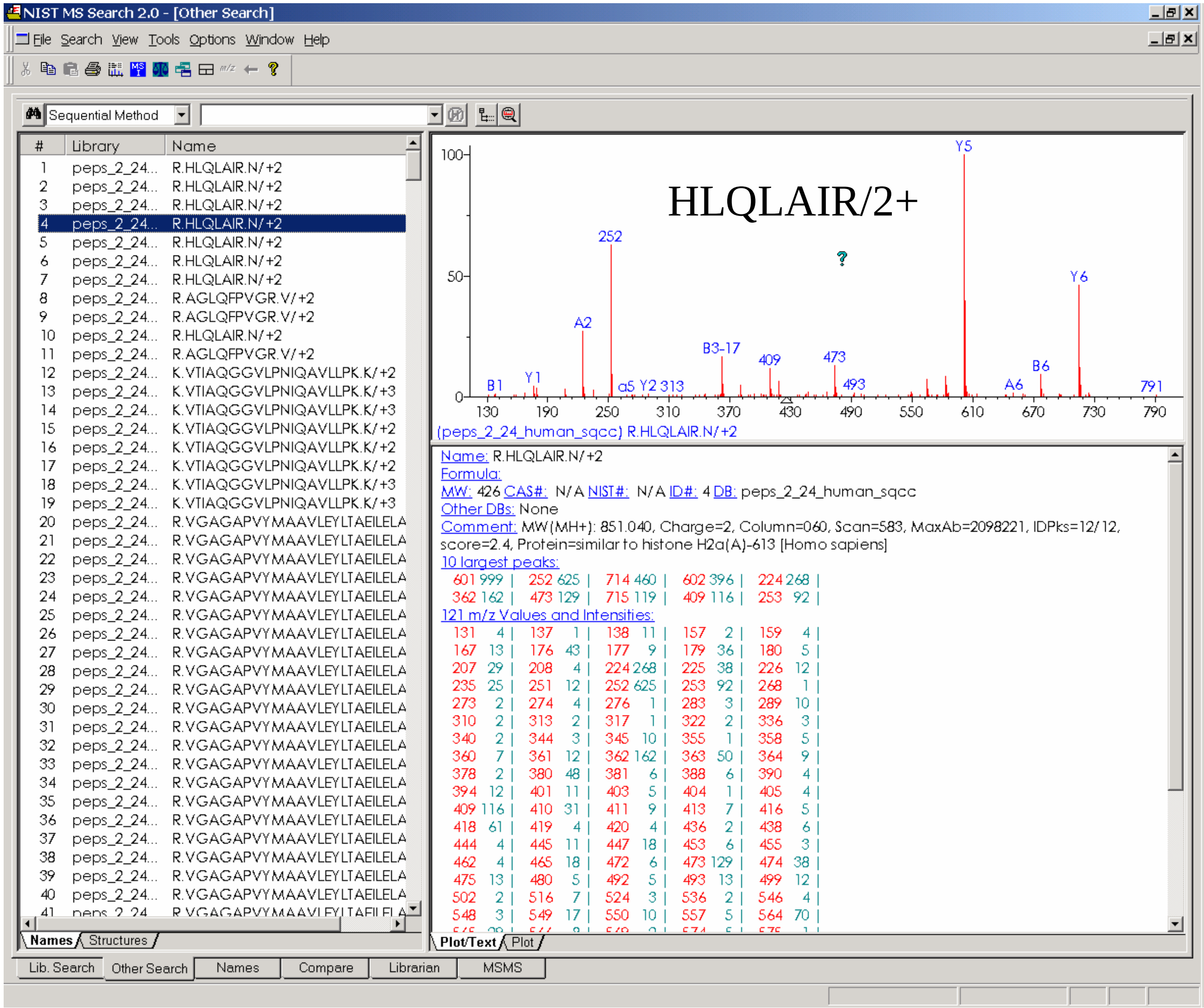
```
datafile=IETMR_20051.2.dta, peptide=IETMR, precursor_mass=325.4, mass_type=Average
```

© J.Eng, Biological Mass Spectrometry
MBT Dept., University of Washington
1-axis
2-axis
3-axis
4-axis
5-axis
[+]
[x1.0]
[-]
Toggle Precursor

```
Seq  #     b       y     +1
---  --  ------  ------  --
 I   1   114.2     -      5
 E   2   243.3   536.6    4
 T   3   344.4   407.5    3
 M   4   475.6   306.4    2
 R   5     -     175.2    1
```

Peptide Spectra Reference Library

(multiple measurements each of 10,000 peptides)



#	Library	Name
1	peps_2_24...	R.HLQLAIR.N/+2
2	peps_2_24...	R.HLQLAIR.N/+2
3	peps_2_24...	R.HLQLAIR.N/+2
4	peps_2_24...	R.HLQLAIR.N/+2
5	peps_2_24...	R.HLQLAIR.N/+2
6	peps_2_24...	R.HLQLAIR.N/+2
7	peps_2_24...	R.HLQLAIR.N/+2
8	peps_2_24...	R.AGLQFPVGR.V/+2
9	peps_2_24...	R.AGLQFPVGR.V/+2
10	peps_2_24...	R.HLQLAIR.N/+2
11	peps_2_24...	R.AGLQFPVGR.V/+2
12	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+2
13	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+3
14	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+3
15	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+2
16	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+2
17	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+2
18	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+3
19	peps_2_24...	K.VTIAQGGVLPNIQAVLLPK.K/+3
20	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
21	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
22	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
23	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
24	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
25	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
26	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
27	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
28	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
29	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
30	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
31	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
32	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
33	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
34	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
35	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
36	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
37	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
38	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
39	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
40	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA
41	peps_2_24...	R.VGAGAPVYMAAVLEYLTAEILELA

Name: R.HLQLAIR.N/+2
Formula:
MW: 426 CAS#: N/A NIST#: N/A ID#: 4 DB: peps_2_24_human_sqcc
Other DBs: None
Comment: MW(MH+): 851.040, Charge=2, Column=060, Scan=583, MaxAb=2098221, IDPks=12/12, score=2.4, Protein=similar to histone H2a(A)-613 [Homo sapiens]

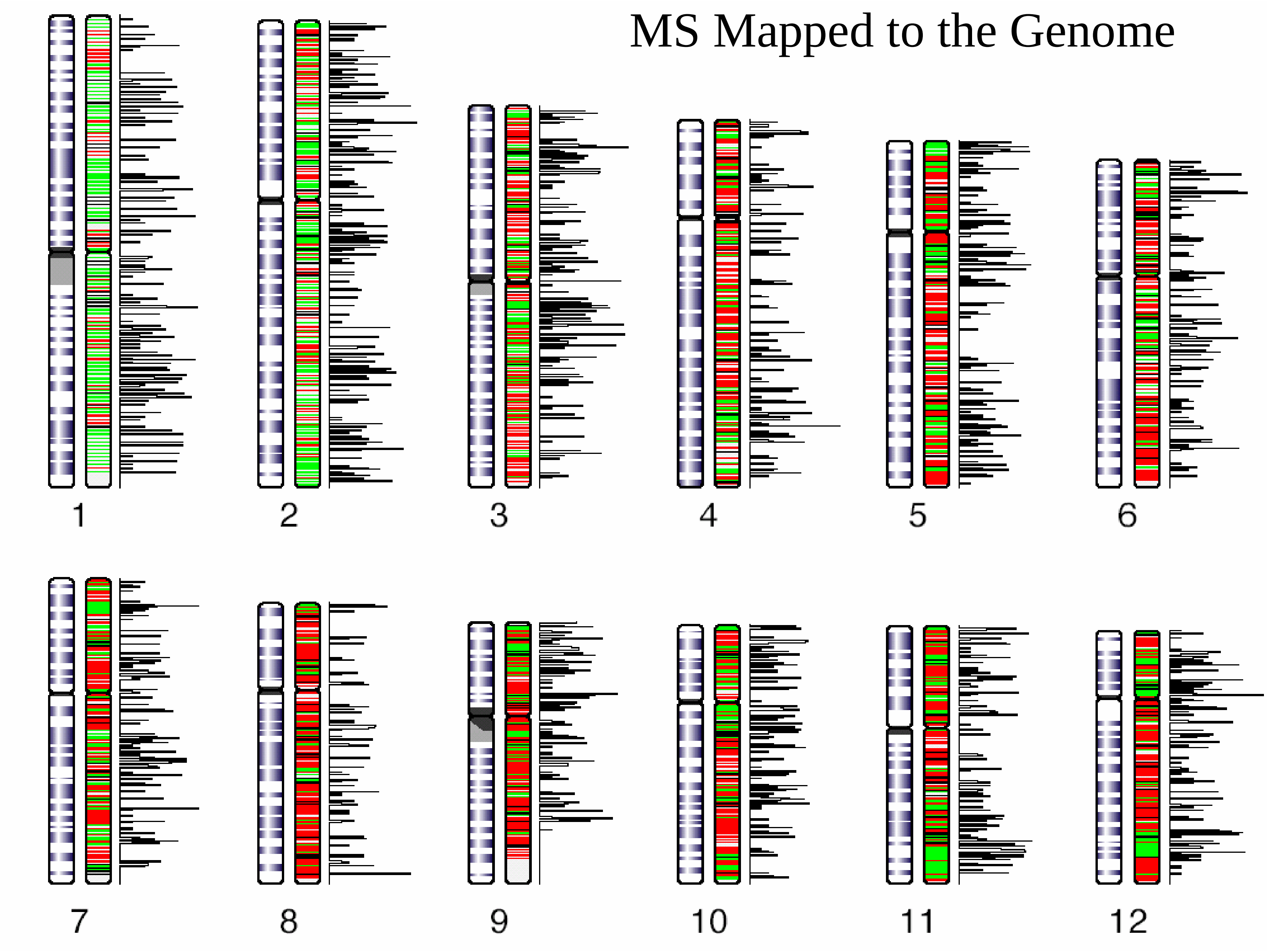
10 largest peaks:

```
 601 999 |  252 625 |  714 460 |  602 396 |  224 268 |
 362 162 |  473 129 |  715 119 |  409 116 |  253  92 |
```

121 m/z Values and Intensities:

```
 131   4 |  137   1 |  138  11 |  157   2 |  159   4 |
 167  13 |  176  43 |  177   9 |  179  36 |  180   5 |
 207  29 |  208   4 |  224 268 |  225  38 |  226  12 |
 235  25 |  251  12 |  252 625 |  253  92 |  268   1 |
 273   2 |  274   4 |  276   1 |  283   3 |  289  10 |
 310   2 |  313   2 |  317   1 |  322   2 |  336   3 |
 340   2 |  344   3 |  345  10 |  355   1 |  358   5 |
 360   7 |  361  12 |  362 162 |  363  50 |  364   9 |
 378   2 |  380  48 |  381   6 |  388   6 |  390   4 |
 394  12 |  401  11 |  403   5 |  404   1 |  405   4 |
 409 116 |  410  31 |  411   9 |  413   7 |  416   5 |
 418  61 |  419   4 |  420   4 |  436   2 |  438   6 |
 444   4 |  445  11 |  447  18 |  453   6 |  455   3 |
 462   4 |  465  18 |  472   6 |  473 129 |  474  38 |
 475  13 |  480   5 |  492   5 |  493  13 |  499  12 |
 502   2 |  516   7 |  524   3 |  536   2 |  546   4 |
 548   3 |  549  17 |  550  10 |  557   5 |  564  70 |
```

MS Mapped to the Genome



From Eric Deutsch, ISB, 6/2004