Faculty research interests seminar

George C. Tseng

11/19/2004

1999-2000	Ph.D. program, Department of Statistics, UCLA
2000-2003	Ph.D., Department of Biostatistics, Harvard University
August, 2003	Joined the department as an assistant professor.

Course: BIOST 2070 Special Topics 4

Statistical methods and data mining in microarray analysis

Time: WEDNESDAY 9:00 – 10:55 AM

Spring Term 05-2

Room A425 Crabtree Hall

Research interests:

1. Unsupervised machine learning (clustering):
   - **Tight clustering**: systematically extract stable and tight patterns in large complex data through resampling approach.
   - **Penalized and weighted K-means**: a class of loss function extended from K-means that allows a noise set not being clustered and incorporation of prior knowledge.

2. Supervised machine learning (classification):
   - **Psi learning**: utilize a modified penalty term in SVM to achieve better error rate performance. (joint work with Xiaotong Shen and Wing Wong)

3. Microarray data analysis and related statistical issues:
   - Quality filtering, normalization, gene selection and multiple comparison, Bayesian hierarchical model

4. Data mining and graphical visualization for genomic and proteomic data

5. Proteomics

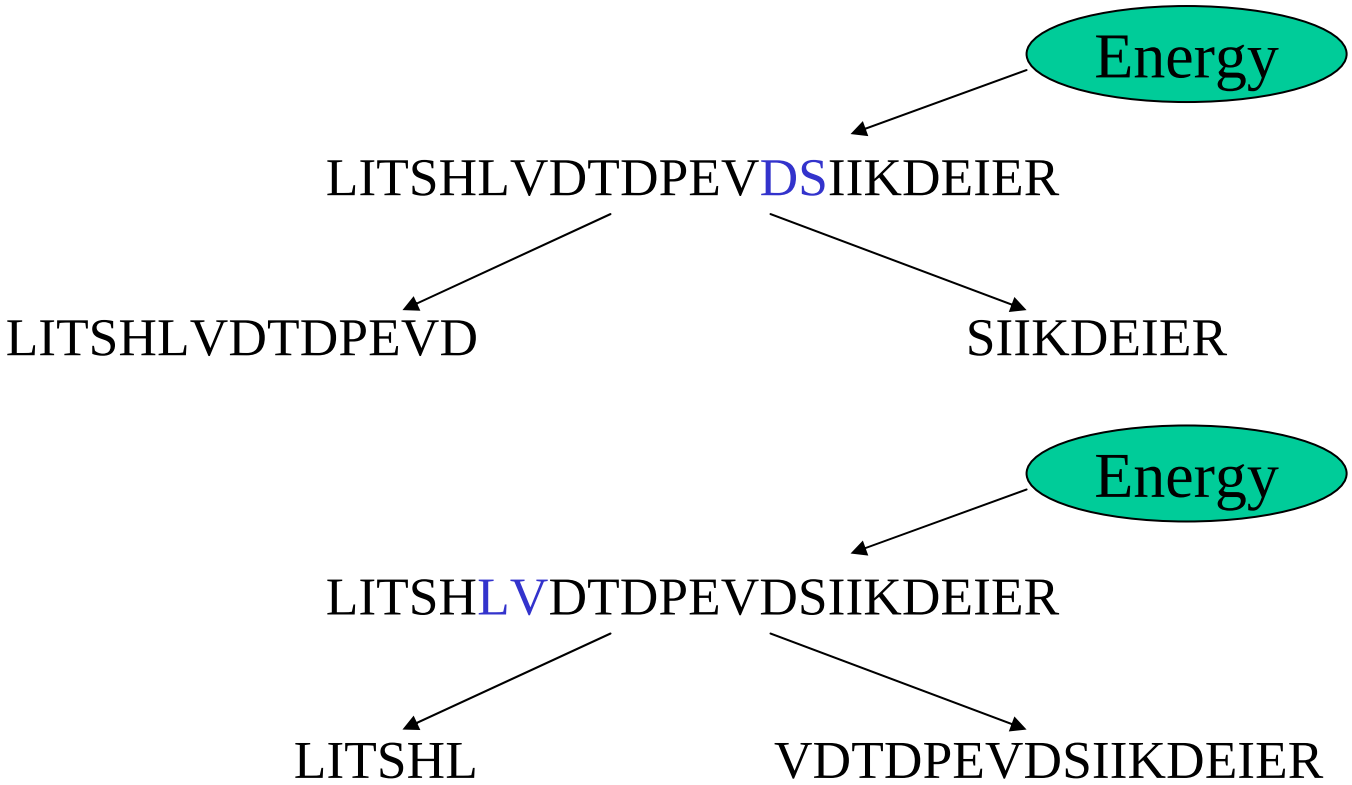
Some biological terms and facts:

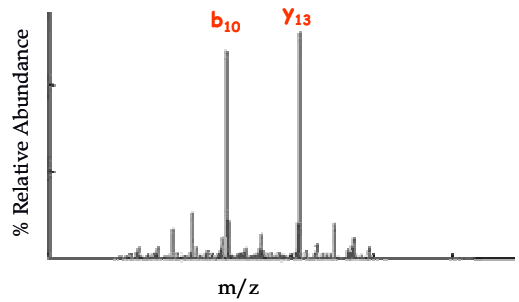
1. Proteins play important roles in most of the molecular activities in our body.

2. Proteins are composed of 20 amino acids (usually abbreviated as letters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) as a sequence.

3. A peptide is defined as a sequence of amino acid combinations, usually a subsequence of a real protein.
   e.g. "LITSHLVD"

Low energy Collision-Induced Dissociation (CID)





The abundance of such cleavages are recorded as intensities.

AAAMDAQAEAK

A	C	D	E	F	G	H	I	K	L	M	N	P	Q
0	1	2	3	4	5	6	7	8	9	10	11	12	13
1st	2nd	count	intensities										
A	A	2	0	0.381									
A	C	0											
A	D	0											
A	E	1	0.031										
A	F	0											
A	G	0											
A	H	0											
A	I	0											
A	K	0											
A	L	0											
A	M	1	0.514										
A	N	0											
A	P	0											
A	Q	1	0.096										
A	R	0											
A	S	0											
A	T	0											
A	V	0											
A	W	0											
A	Y	0											
C	A	0											

AA

AE

AM

AQ

For a specific set of peptides:

A	C	D	E	F	G	H	I	K	L	M	N	P	Q
0	1	2	3	4	5	6	7	8	9	10	11	12	13
1st	2nd	count	intensities										
A	A	188	0	0	0	0	0	0	0	0	0	0	0
A	C	1	0.105										
A	D	91	0	0	0	0	0	0	0	0	0	0	0
A	E	94	0	0	0	0	0	0	0	0	0	0	0
A	F	41	0	0	0	0	0	0	0	0	0.008	0.008	0.01
A	G	129	0	0	0	0	0	0	0	0	0	0	0
A	H	0											
A	I	69	0	0	0	0	0	0	0	0	0	0	0.019
A	K	5	0	0.01	0.013	0.122	0.139						
A	L	137	0	0	0	0	0	0	0	0	0	0	0
A	M	28	0.004	0.02	0.023	0.034	0.034	0.06	0.062	0.068	0.093	0.101	0.122
A	N	52	0	0	0	0	0.006	0.022	0.025	0.025	0.033	0.039	0.046
A	P	152	0	0	0	0	0.012	0.029	0.041	0.042	0.049	0.059	0.068
A	Q	58	0	0	0	0	0	0	0	0	0	0	0
A	R	0											
A	S	59	0	0	0	0	0	0.01	0.021	0.031	0.038	0.04	0.049
A	T	78	0	0	0	0	0	0	0	0	0	0.013	0.018
A	V	88	0	0	0	0	0	0	0	0	0	0	0
A	W	11	0	0	0	0	0.029	0.149	0.205	0.321	0.428	0.454	1
A	Y	26	0	0	0	0	0.008	0.011	0.047	0.063	0.065	0.078	0.079
C	A	0											

20×20=400 independent distributions

Each with 0 or multiple (up to hundreds) observations

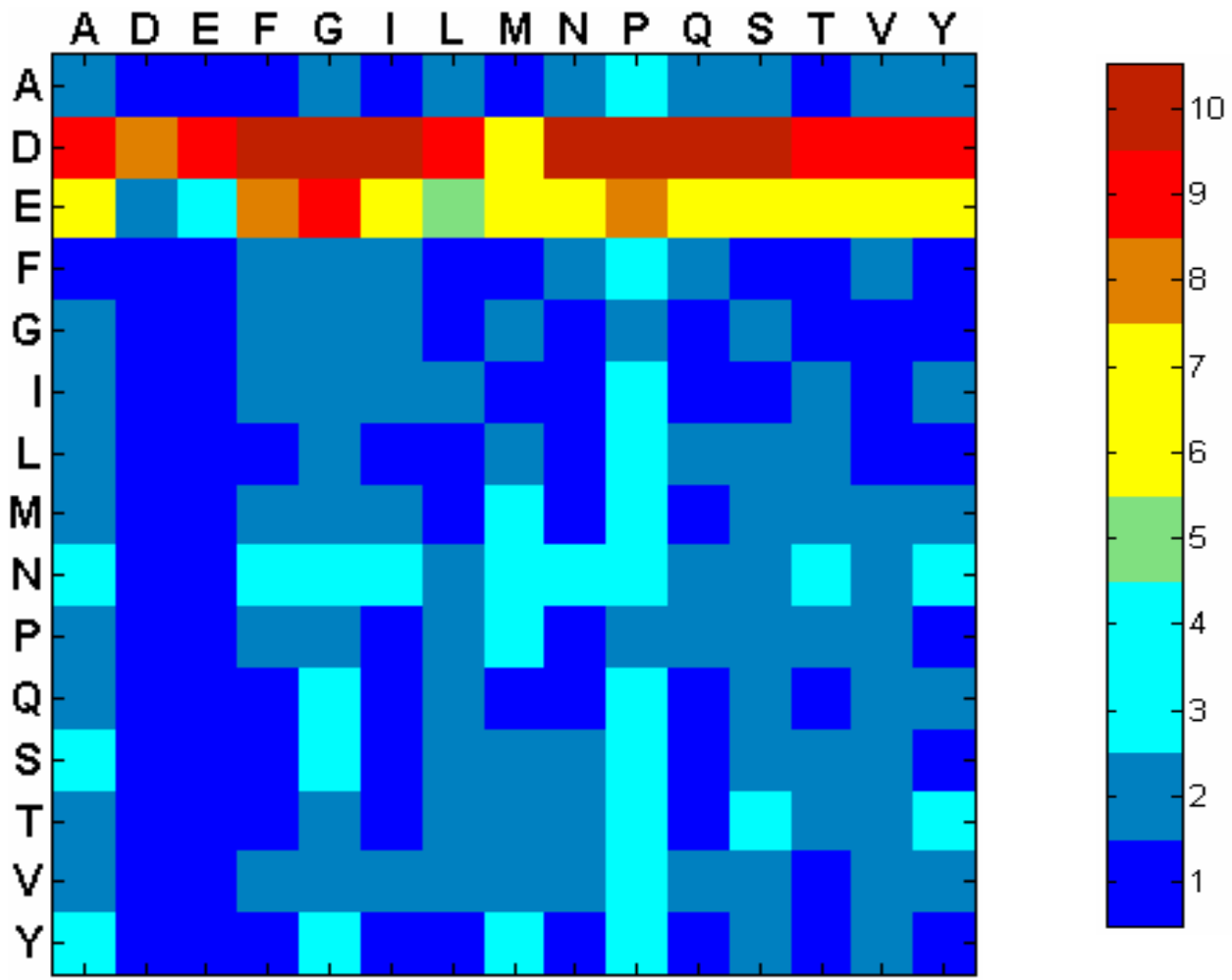
A	A	188	0	0	0	0	0	0	0	0	0	0	0
A	C	1	0.105										
A	D	91	0	0	0	0	0	0	0	0	0	0	0
A	E	94	0	0	0	0	0	0	0	0	0	0	0
A	F	41	0	0	0	0	0	0	0	0	0.008	0.008	0.01
A	G	129	0	0	0	0	0	0	0	0	0	0	0
A	H	0											
A	I	69	0	0	0	0	0	0	0	0	0	0	0.019
A	K	5	0	0.01	0.013	0.122	0.139						
A	L	137	0	0	0	0	0	0	0	0	0	0	0
A	M	28	0.004	0.02	0.023	0.034	0.034	0.06	0.062	0.068	0.093	0.101	0.122
A	N	52	0	0	0	0	0.006	0.022	0.025	0.025	0.033	0.039	0.046
A	P	152	0	0	0	0	0.012	0.029	0.041	0.042	0.049	0.059	0.068
A	Q	58	0	0	0	0	0	0	0	0	0	0	0
A	R	0											
A	S	59	0	0	0	0	0	0.01	0.021	0.031	0.038	0.04	0.049
A	T	78	0	0	0	0	0	0	0	0	0	0.013	0.018
A	V	88	0	0	0	0	0	0	0	0	0	0	0
A	W	11	0	0	0	0	0.029	0.149	0.205	0.321	0.428	0.454	1
A	Y	26	0	0	0	0	0.008	0.011	0.047	0.063	0.065	0.078	0.079

AX: low

D	A	73	0.018	0.026	0.032	0.046	0.046	0.093	0.108	0.146	0.155	0.173	0.193
D	C	1	0										
D	D	38	0	0.012	0.056	0.097	0.097	0.118	0.135	0.142	0.156	0.182	0.197
D	E	53	0	0	0	0.021	0.026	0.035	0.05	0.073	0.08	0.085	0.095
D	F	38	0	0.003	0.044	0.1	0.119	0.214	0.232	0.283	0.41	0.468	0.507
D	G	44	0.024	0.128	0.128	0.226	0.239	0.247	0.383	0.395	0.491	0.529	0.693
D	H	0											
D	I	38	0.029	0.054	0.063	0.128	0.15	0.173	0.229	0.233	0.257	0.268	0.284
D	K	0											
D	L	76	0	0	0	0	0	0.01	0.057	0.064	0.113	0.114	0.126
D	M	15	0	0.147	0.212	0.376	0.419	0.709	0.806	0.841	0.885	0.887	0.947
D	N	18	0.047	0.108	0.232	0.442	0.458	0.506	0.508	0.575	0.585	0.844	0.904
D	P	63	0.122	0.444	0.481	0.61	0.631	0.675	0.753	0.882	0.883	1	1
D	Q	14	0	0.098	0.163	0.228	0.262	0.297	0.421	0.459	0.463	0.55	0.667
D	R	0											
D	S	42	0	0.03	0.049	0.075	0.093	0.144	0.147	0.217	0.255	0.362	0.403
D	T	34	0	0	0.064	0.145	0.149	0.223	0.235	0.24	0.243	0.297	0.313
D	V	58	0	0	0.031	0.053	0.07	0.076	0.088	0.098	0.135	0.148	0.173
D	W	5	0.204	0.222	0.225	0.376	1						
D	Y	29	0.021	0.033	0.037	0.039	0.056	0.056	0.07	0.098	0.099	0.099	0.101

DX: high

Take median for each of the 400 distributions

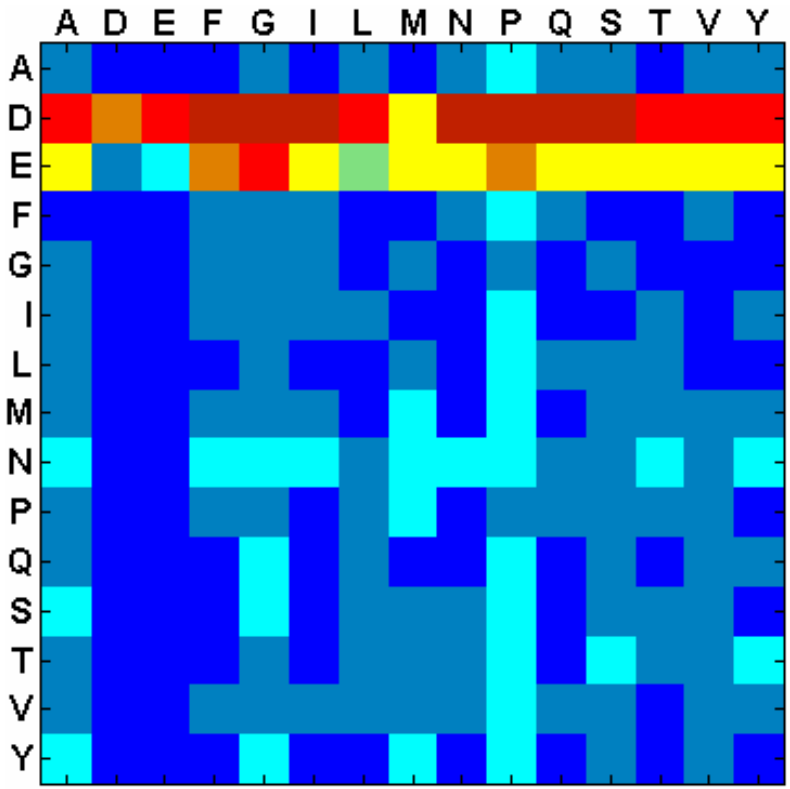


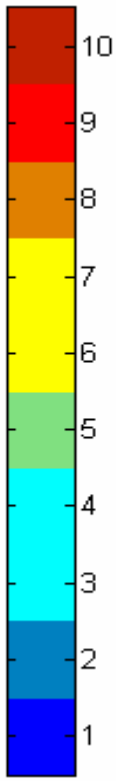
Some peptide sequence motif can contribute to different fragmentation pattern

spectra set 1

y ion, with P, no H, end in R

e.g. aaaaeldiplyr



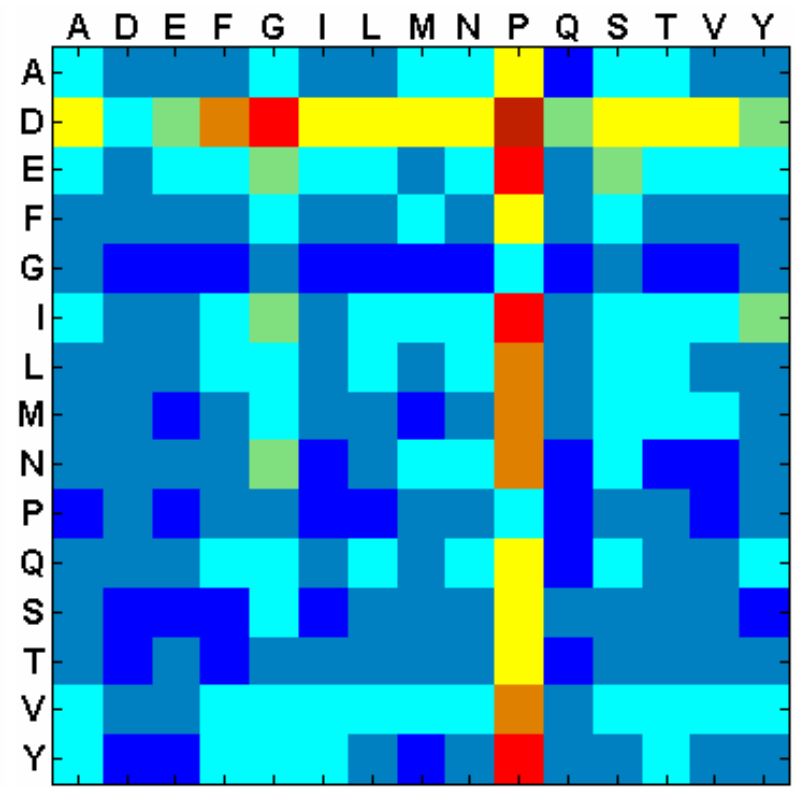


DX and EX

spectra set 2

y ion, with P, no H, end in K

e.g. aaainiiptstgaak



DX and XP

X: means any

Part I: New Visualization Tool

What's the problem?

1. "Median" is an effective data reduction from the whole distribution but also lose lots of information

2. The discrete color bar can be improved.

3. Each cell now contribute equally to visualization despite that each cell contain different degree of information.

   e.g. information contained in the two distributions (0.41, 0.39, 0.45, 0.44, 0.39) & (0.01, 0.79) are very different.

What alternatives do we have so far?

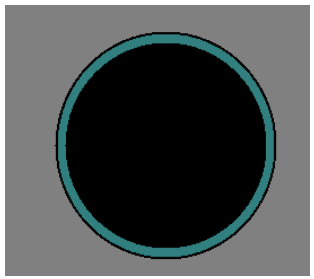
Descriptive statistics:
-- statistics:
mean, median, mode, trimmed mean, inter-quartile mean….
variance, higher-order moments….

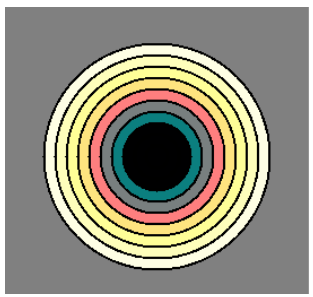
-- visualization (plots):
Histogram, density plot, box plot……

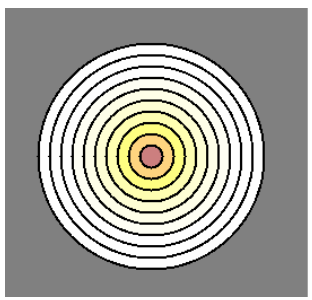
What about visualizing 400 distributions?!

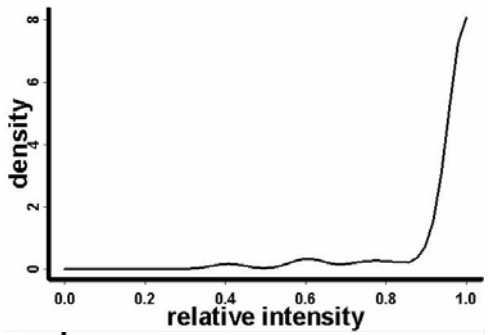
New proposal:

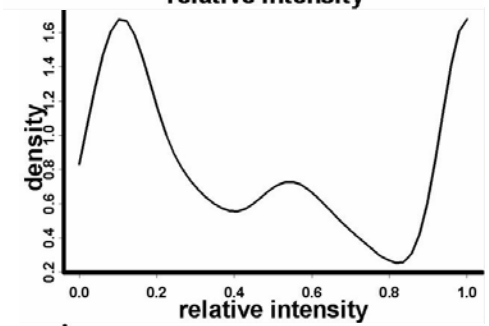
1. Gradient color to represent intensities
2. 10 concentric doughnuts representing 5%, 15%,…,95% quantiles (from outside inward).

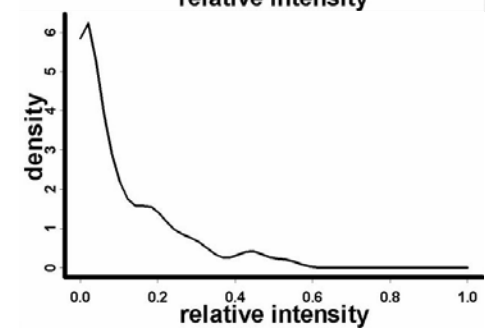












3. Width: proportional to the information contained in the distribution.

Count (number) of the observations??

Fisher Information!!

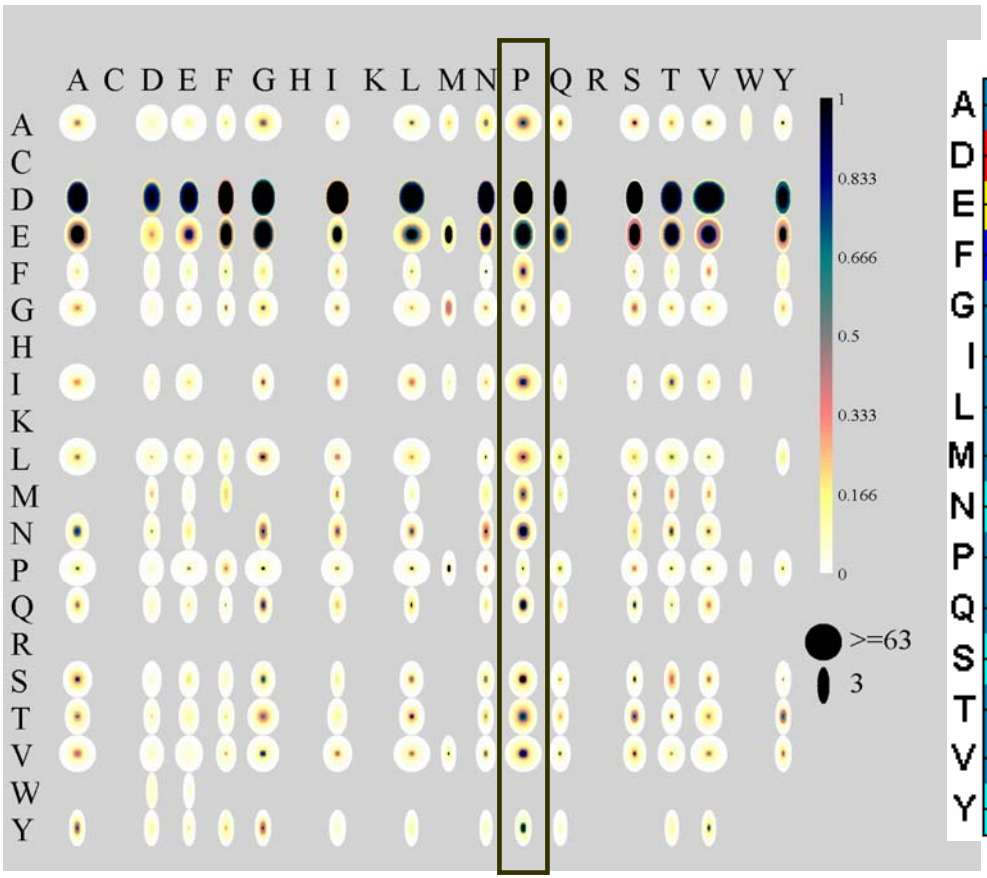
$X_1, X_2,\ldots,X_n \sim N(\mu, \sigma^2)$

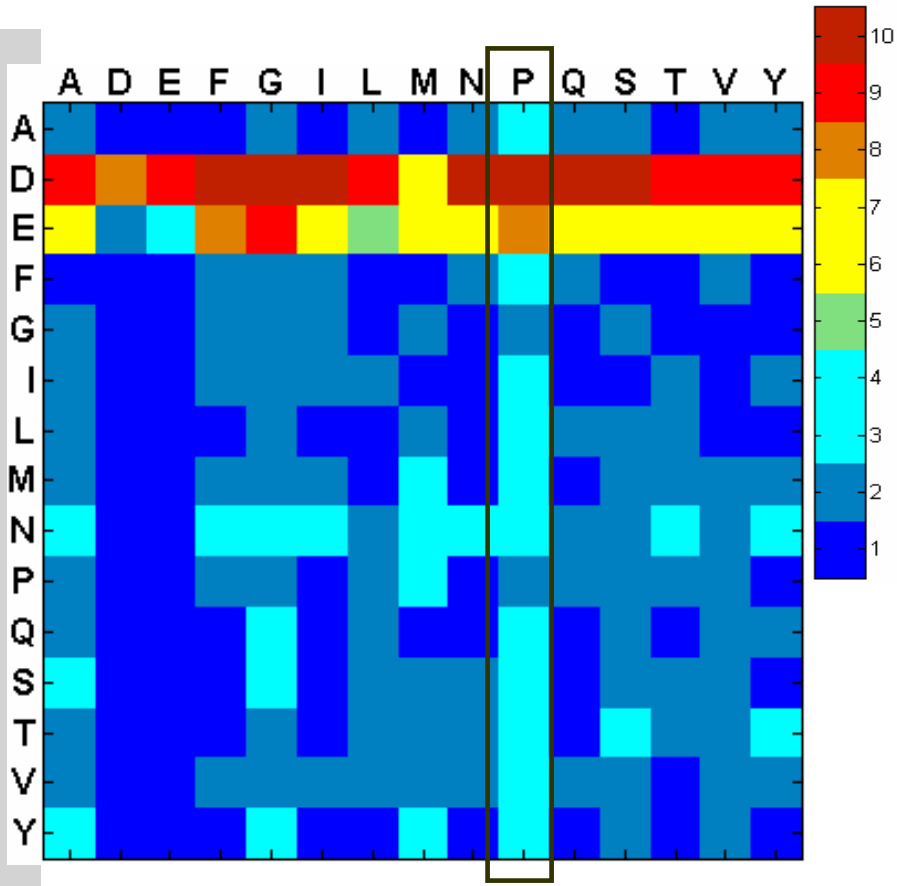
$$\mathrm{I}(\mu) = E\left(\left[\frac{\partial}{\partial \mu}\log f(X;\mu)\right]^2\right) = -E\left[\frac{\partial^2}{\partial \mu^2}\log f(X;\mu)\right] = \frac{n}{\sigma^2}$$

$$\hat{\mathrm{I}}(\mu) = \frac{n}{\hat{\sigma}^2} = \frac{n\cdot(n-1)}{\sum\left(X_i - \overline{X}\right)^2}$$

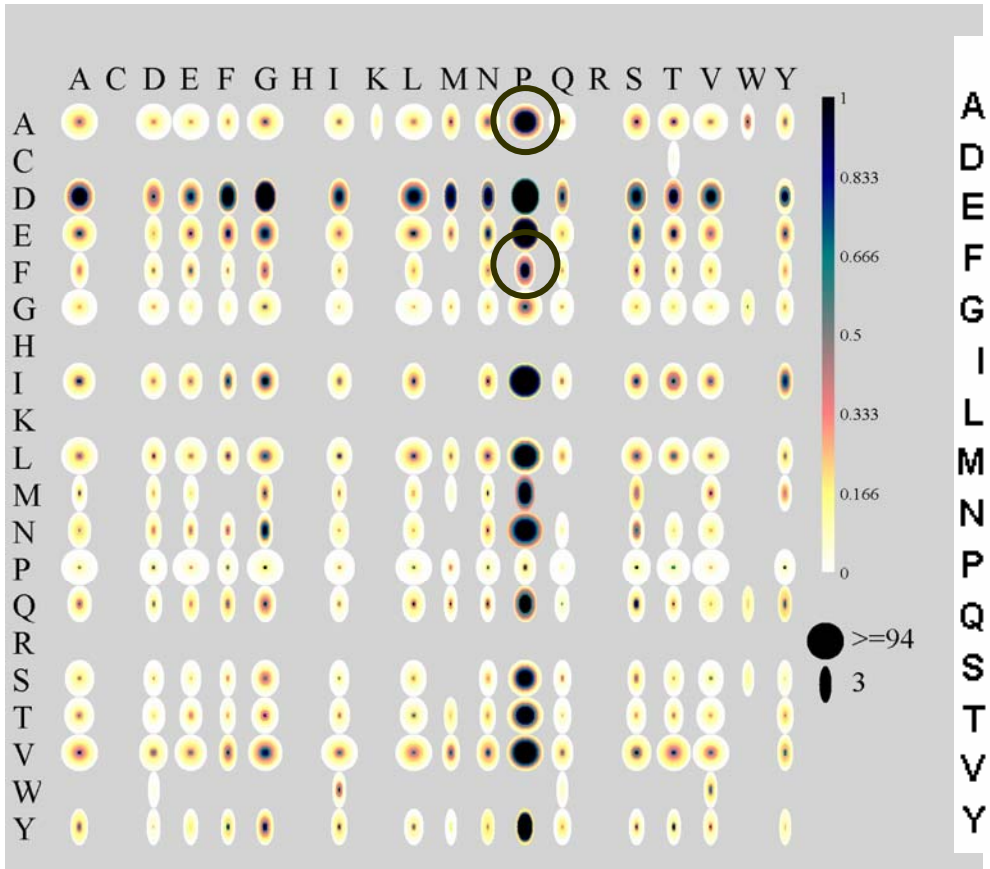
The larger number of observations and the smaller variance, the more information data contains

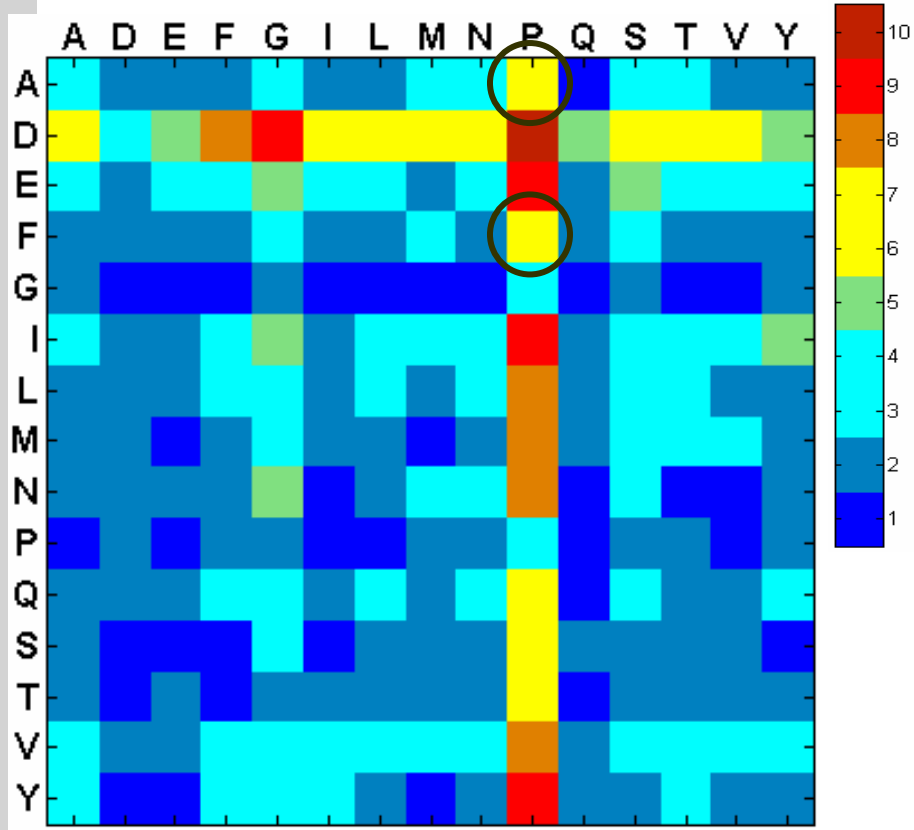
y ion, with P, no H, end in R





y ion, with P, no H, end in K





Some remaining technical issues:

(0,0) => $I=\infty$

(0,0,0,0,0) => $I=\infty$

(0.45, 0.51, 0.47, 0.48, 0.45, 0.49, 0.47, 0.49, 0.46, 0.50)
=> I=23622.05

(0, 0.003) => I=444444.4

Model through Empirical Bayes!!

Part II: Statistical Testing of Pattern Differences

- Wilcoxon (Rank-sum) test for single cell

$$Z_{ij} = \frac{\left(\sum_{k=1}^{n_{ij}} R(X_{ijk}) \right) - \frac{n_{ij}\left(n_{ij} + m_{ij} + 1\right)}{2}}{\sqrt{\frac{n_{ij} m_{ij}\left(n_{ij} + m_{ij} + 1\right)}{12}}}$$

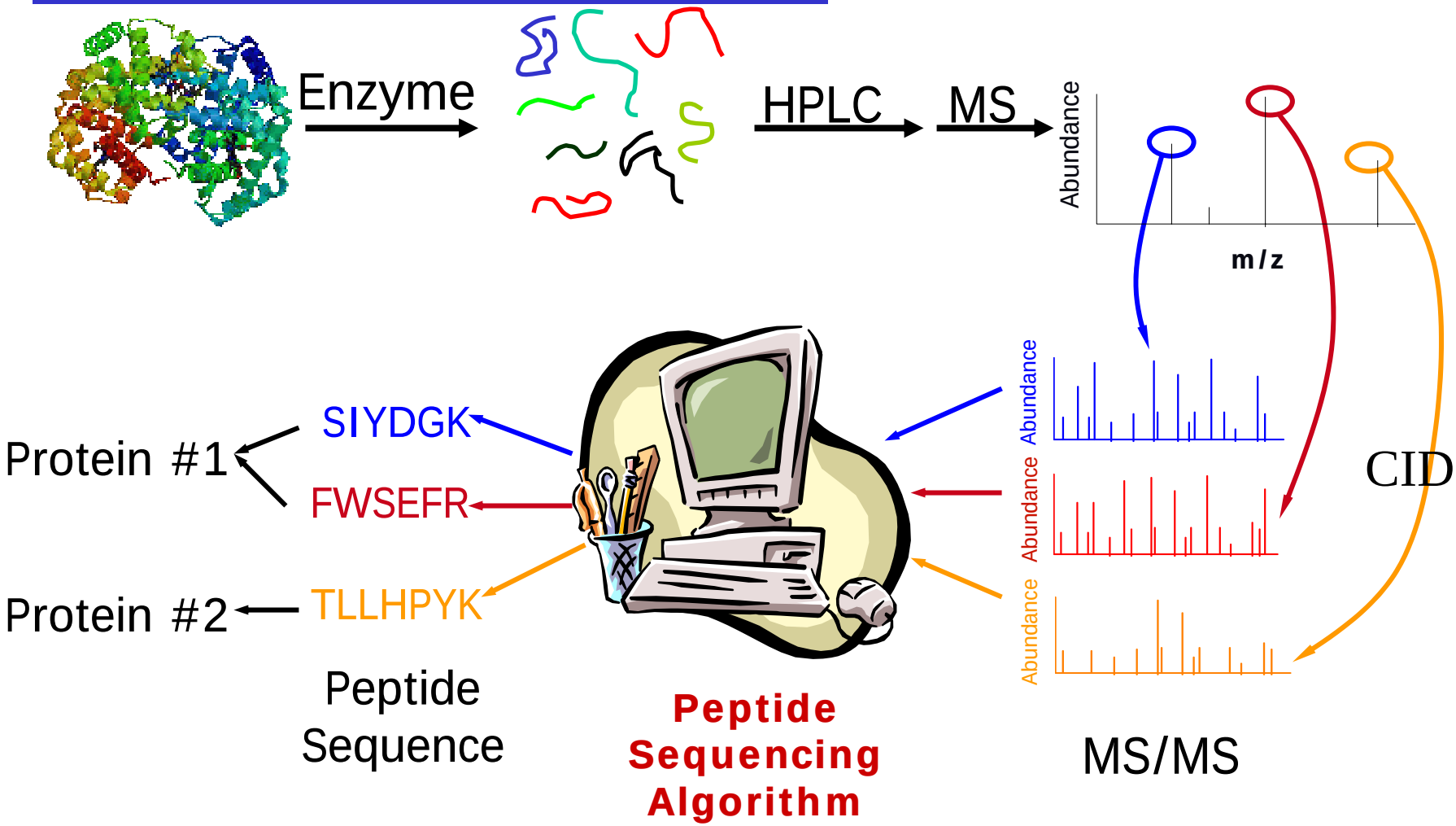
- Compare two tables

$$W = \sum_{(i,j): n_{ij} > 10 \text{ and } m_{ij} > 10} \left| Z_{ij} \right|^d$$

$$When \quad d = 2, \quad W \sim \chi^2_n \quad under \quad null.$$

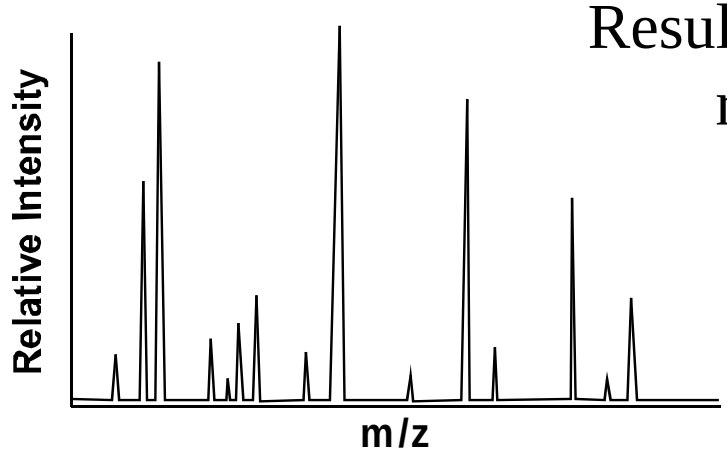
Part III:
a data mining scheme for extracting fragmentation patterns and learning chemical knowledge

Protein Identification by MS/MS (Tandem Mass Spectrometry)



Why Current Algorithms Fail ?

- Based on overly-simplified fragmentation model
  - Cleavage happens uniformly throughout the peptide sequence
  - 20 AA residues fragment the same way
- Use only m/z information, ignore the intensity information



Result in poor accuracy rate (<50%)!!

Hypothesis

- Algorithms can be improved by using a better peptide dissociation model
- Overall fragmentation behavior can be obtained by analyzing a large number of spectra

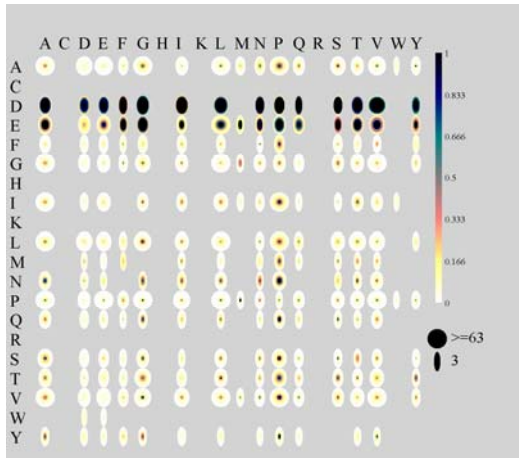
Through such analyses:

- Cluster the fragmentation patterns
- Extract features behind these patterns
- Confirm the factors through experiments
- Incorporate into sequencing algorithm

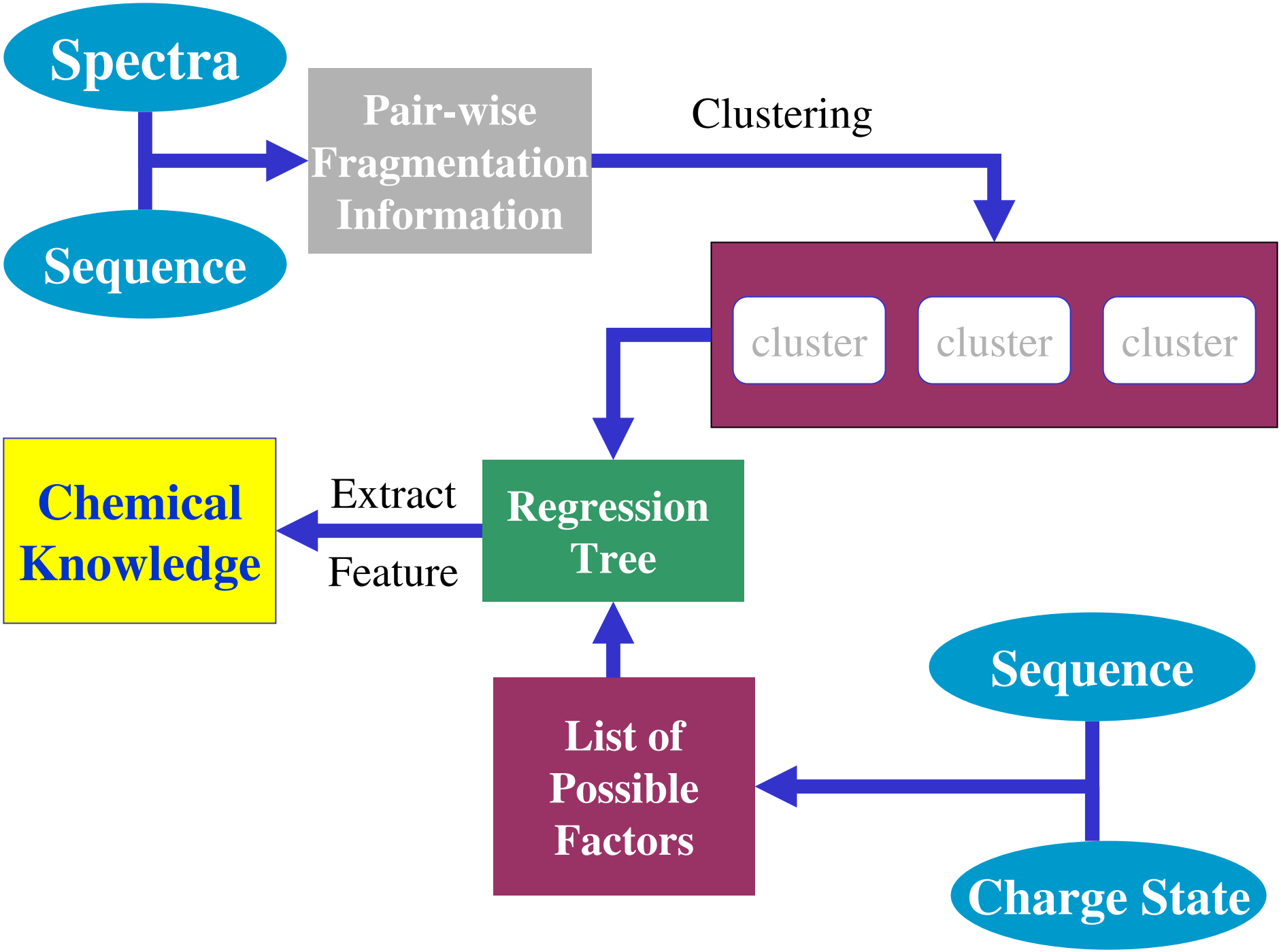
Known feature (sequence motifs)

Plot fragmentation pattern

y ion, with P, no H, end in R

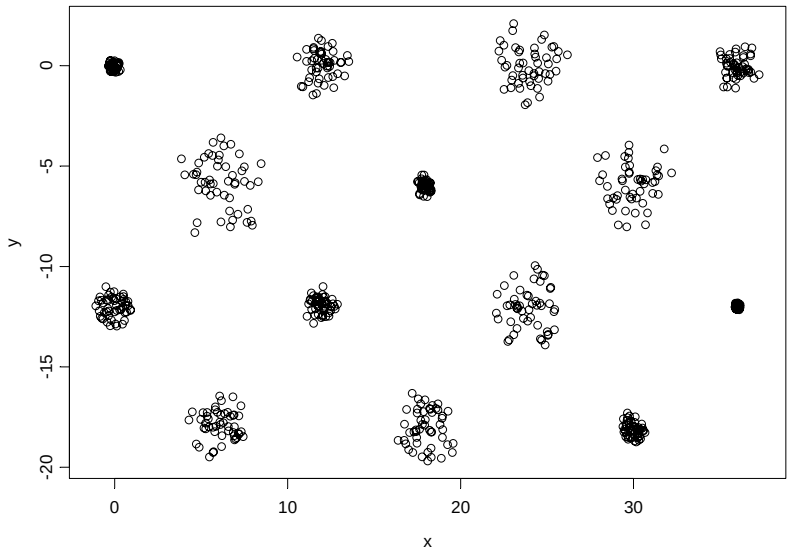


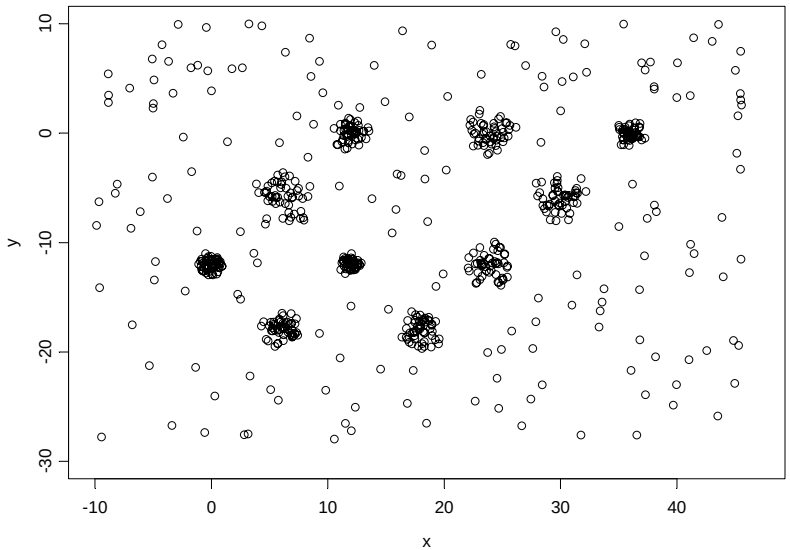
- Motifs known to affect fragmentation patterns are very few.
- We want to learn unknown motifs through a large scale data mining approach.


Spectra
Sequence
Pair-wise Fragmentation Information
Clustering
cluster
cluster
cluster
Chemical Knowledge
Extract
Feature
Regression Tree
List of Possible Factors
Sequence
Charge State

Clustering

Potential Problem for Kmeans





Works well

Noise confuses algorithm

Clustering

A class of penalized and weighted K-means:

$$W(C;k,\lambda) = \sum_{j=1}^{k} \sum_{x_i \in C_j} w(x_i) \cdot d(x_i, C_j) + \lambda |S|$$

$X = (\cup_j C_j) \cup S$ S: the set of scattered genes

d(x, C): with-in cluster dispersion of point *x*

w(x): weight for preferred or prohibited patterns

Proposition 1. If $\lambda_1 > \lambda_2$, then $|S^*(k, \lambda_1)| \leq |S^*(k, \lambda_2)|$.

Clustering

Penalized K-means

$$W_p(C;k,\lambda_0) = \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \bar{x}^{(j)}\|^2 + \lambda_0 |S| \cdot \left(\frac{H}{\sqrt[p]{k}} \right)^2$$

Equivalent to the mixture classification likelihood

$$f(x|C,\theta) = \prod_{j=1}^{k} \prod_{i \in C_j} f(x_i|\mu_j,\Sigma_j) \prod_{i \in S} \frac{1}{|V|}$$

Feature Extraction

Now we have the clusters, then:

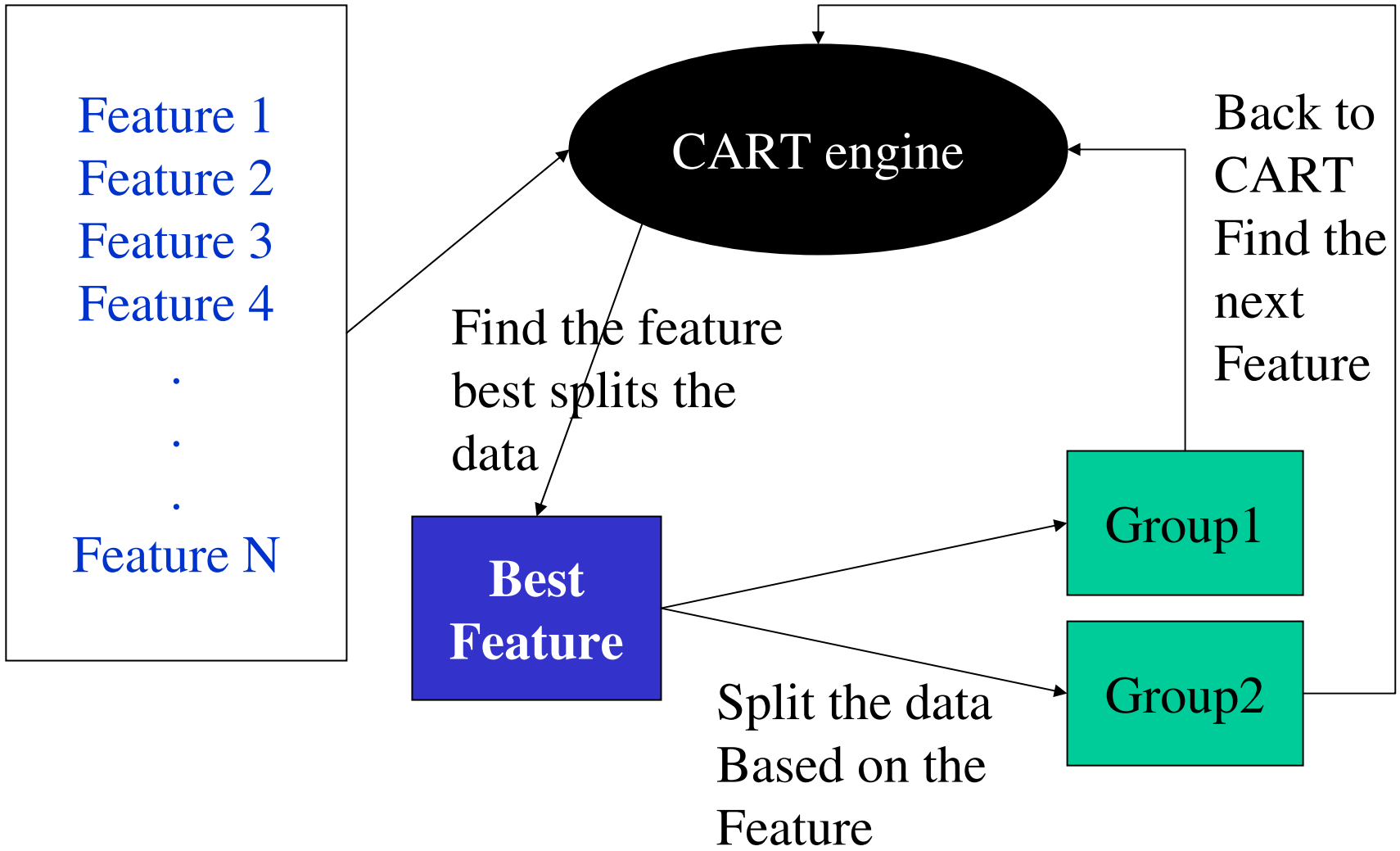
"**what are the underlying chemical factors (sequence motif)?**"

Feature Extraction

CART (Classification And Regression Tree)

- We decided to use CART to find the factors which correspond to the clusters formed by the similarity of the intensity spectra.
- In our application of CART method, the loss function is the **“misclassification rates”** by the selected feature.

Feature Extraction



Features Extraction

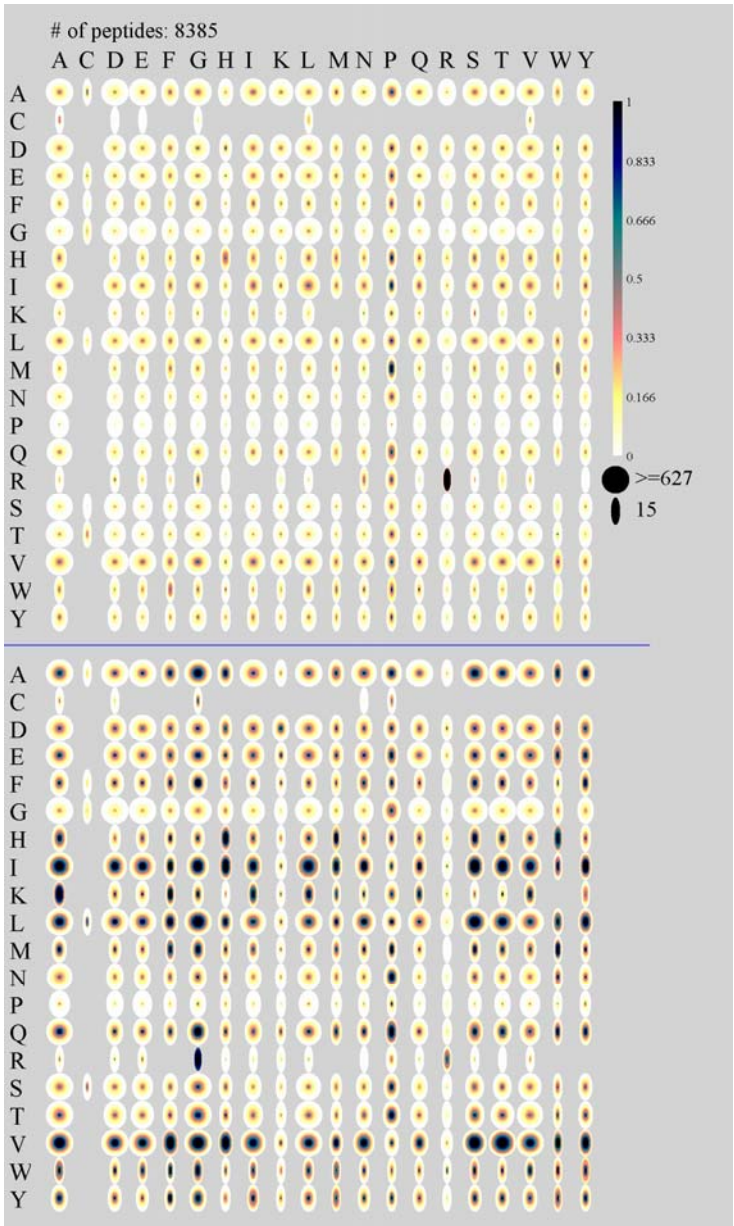
-- The Feature Space:

These features are determined mostly based on prior chemical knowledge.

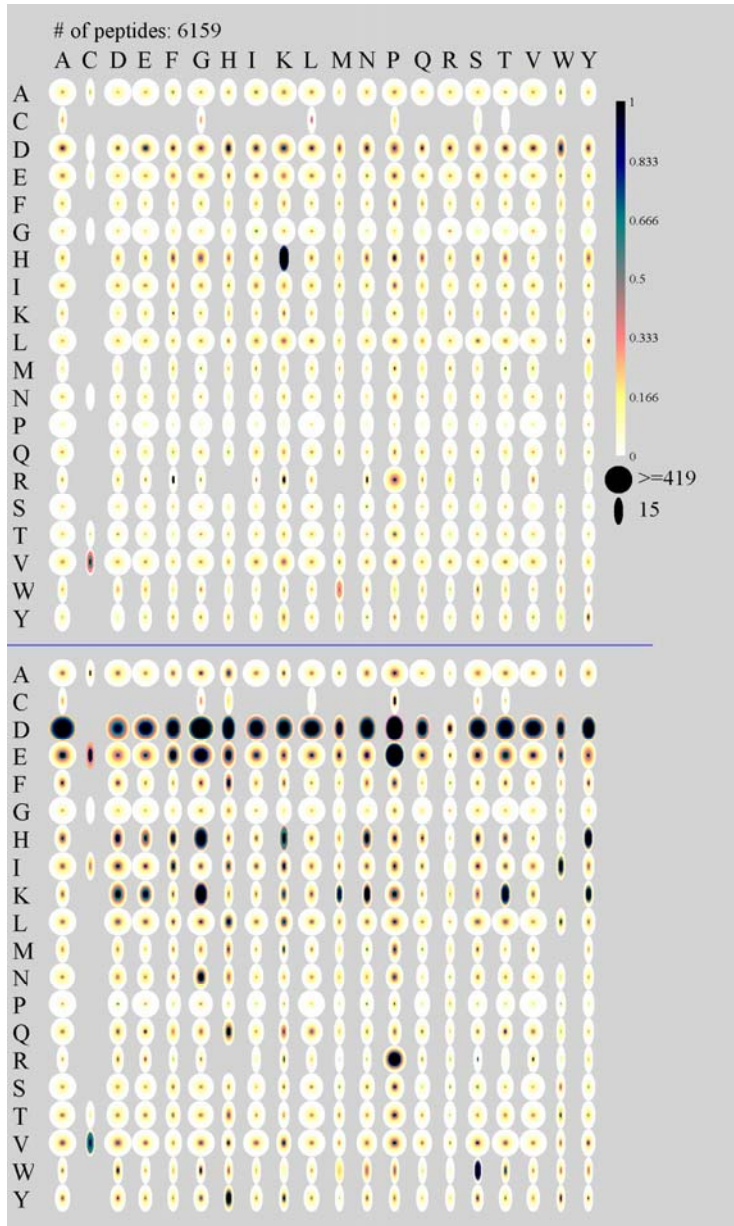
- m/z ratio
- Residue content
- Absolute position and relative position of a residue
- Relative basicity of peptide

Clustering Result (28330 spectra):

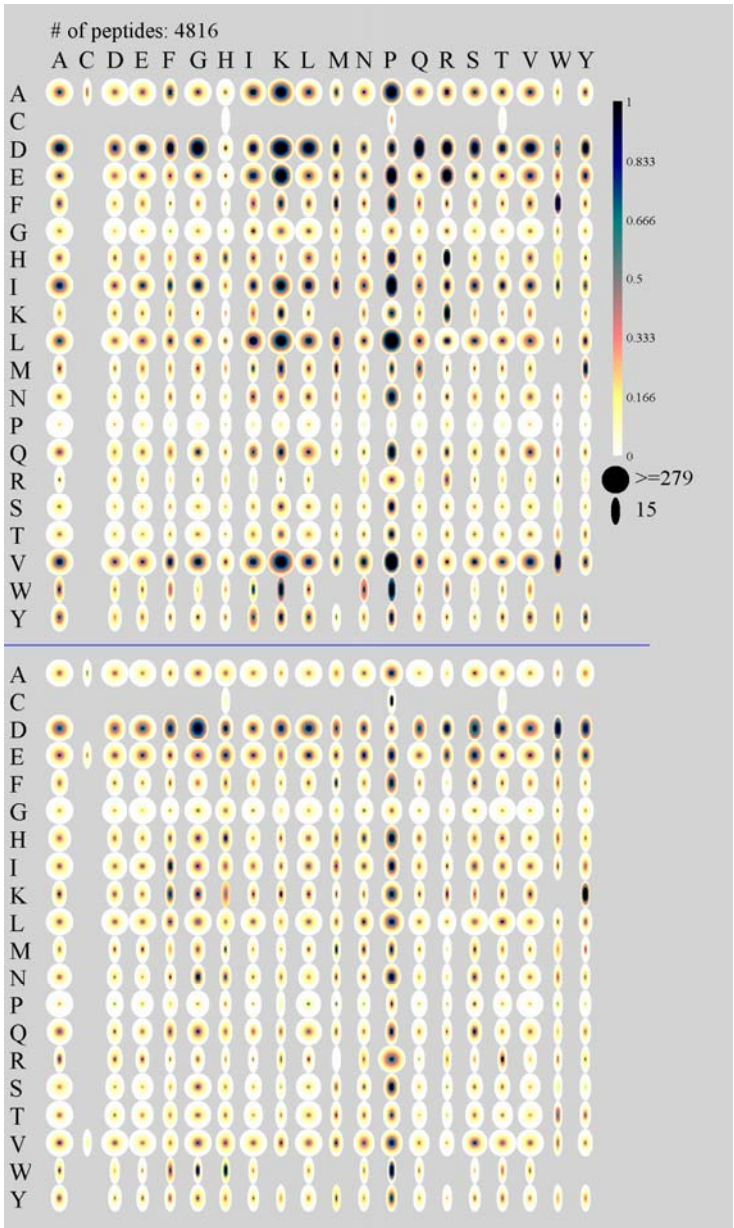
Cluster 1:



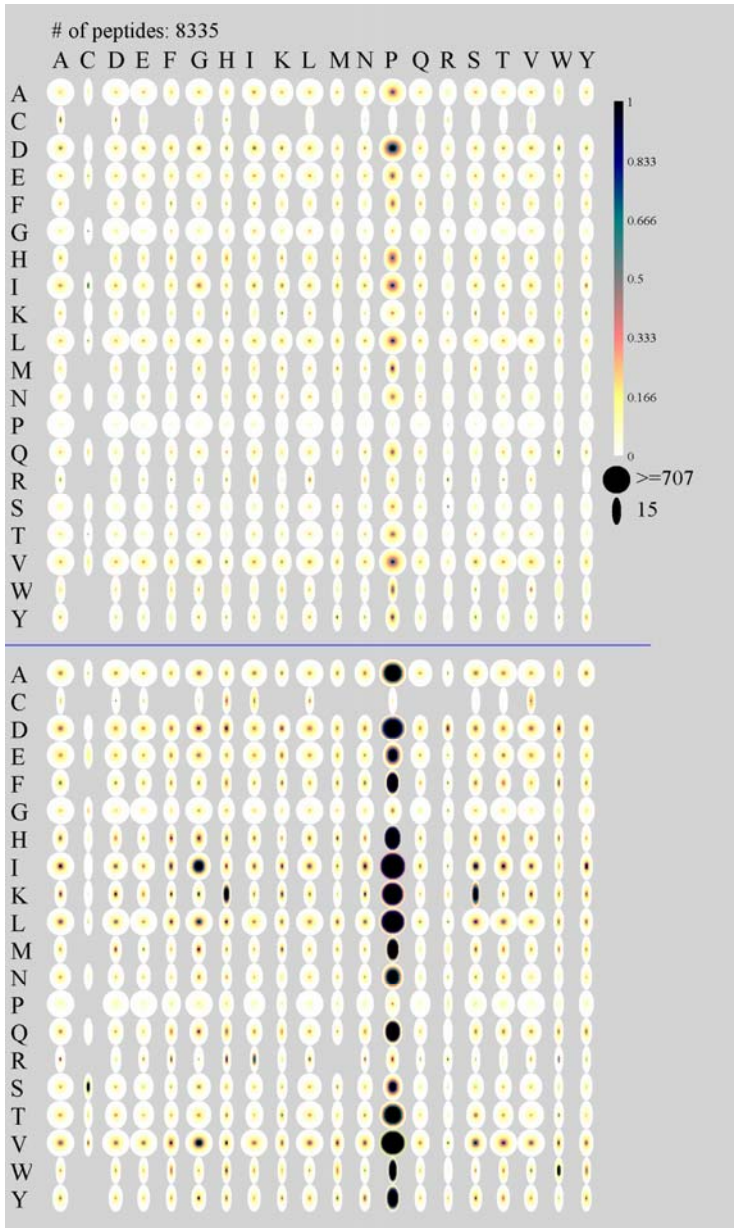
Cluster 2:



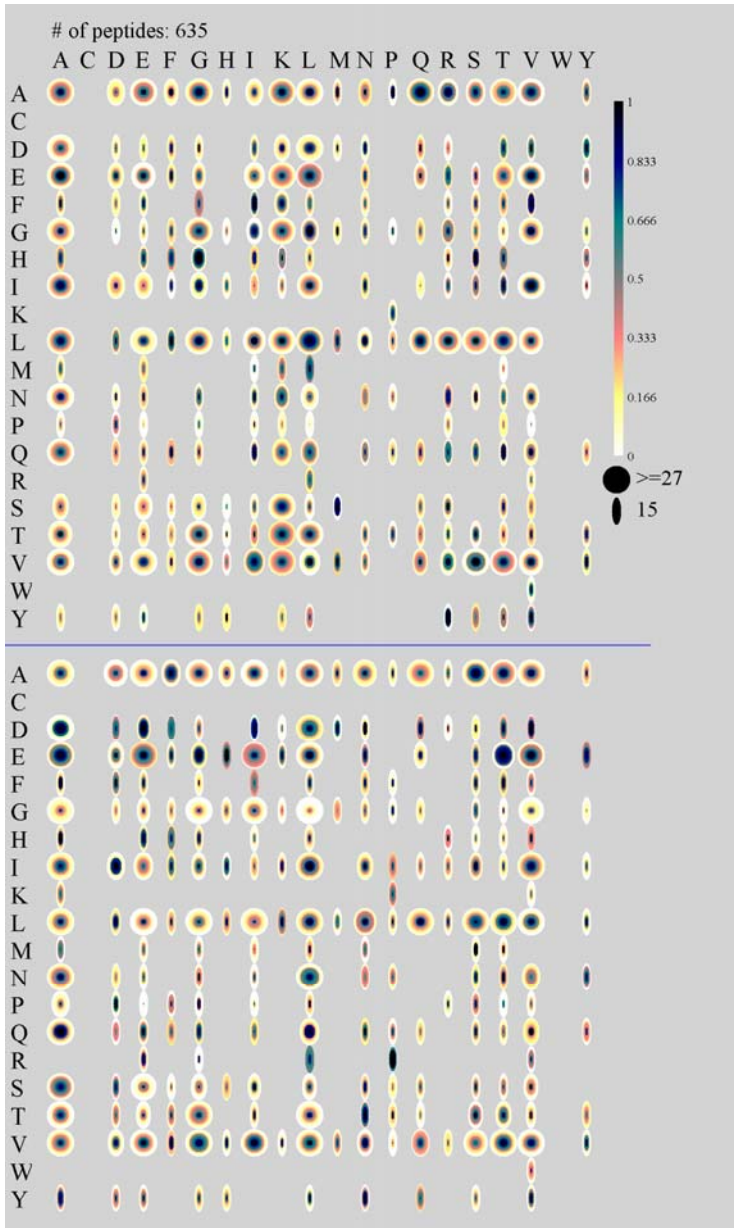
Cluster 3:

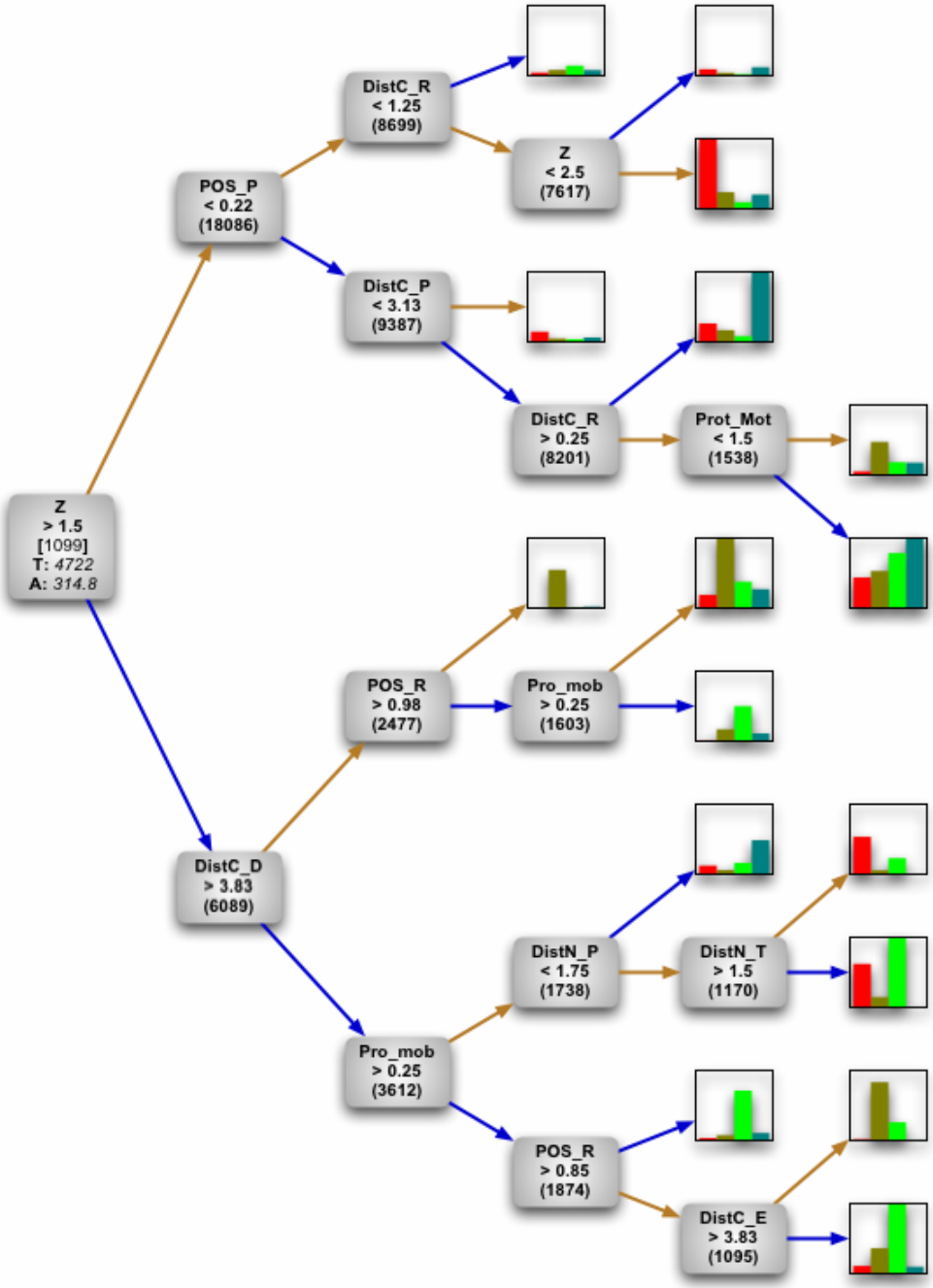


Cluster 4:



Noise set :


# of peptides: 635
A C D E F G H I K L M N P Q R S T V W Y
1
0.833
0.666
0.5
0.333
0.166
0
>=27
15


Z
> 1.5
[1099]
T: 4722
A: 314.8
POS_P
< 0.22
(18086)
DistC_D
> 3.83
(6089)
DistC_R
< 1.25
(8699)
Z
< 2.5
(7617)
DistC_P
< 3.13
(9387)
DistC_R
> 0.25
(8201)
Prot_Mot
< 1.5
(1538)
POS_R
> 0.98
(2477)
Pro_mob
> 0.25
(1603)
Pro_mob
> 0.25
(3612)
DistN_P
< 1.75
(1738)
DistN_T
> 1.5
(1170)
POS_R
> 0.85
(1874)
DistC_E
> 3.83
(1095)

Related working papers

- Yingying Huang, Joseph M. Triscari, **George C. Tseng**, Ljiljana Pasa-Tolic, Mary S. Lipton, Richard D. Smith, Vicki H. Wysocki. *Statistical Characterization of Charge State and Residue Dependence of Low Energy CID Peptide Dissociation Patterns.*

- Yingying Huang, **George C. Tseng**, Shinsheng Yuan, Ljiljana Pasa-Tolic, Mary S. Lipton, Richard D. Smith, Vicki H. Wysocki. *A data mining scheme for identifying peptide structural motifs behind different MS/MS fragmentation intensity patterns.*

- **George C. Tseng.** *A class of penalized and weighted K-means method for clustering.*

- **George C. Tseng** and Yingying Huang. *Pattern visualization of multiple independent distributions.*