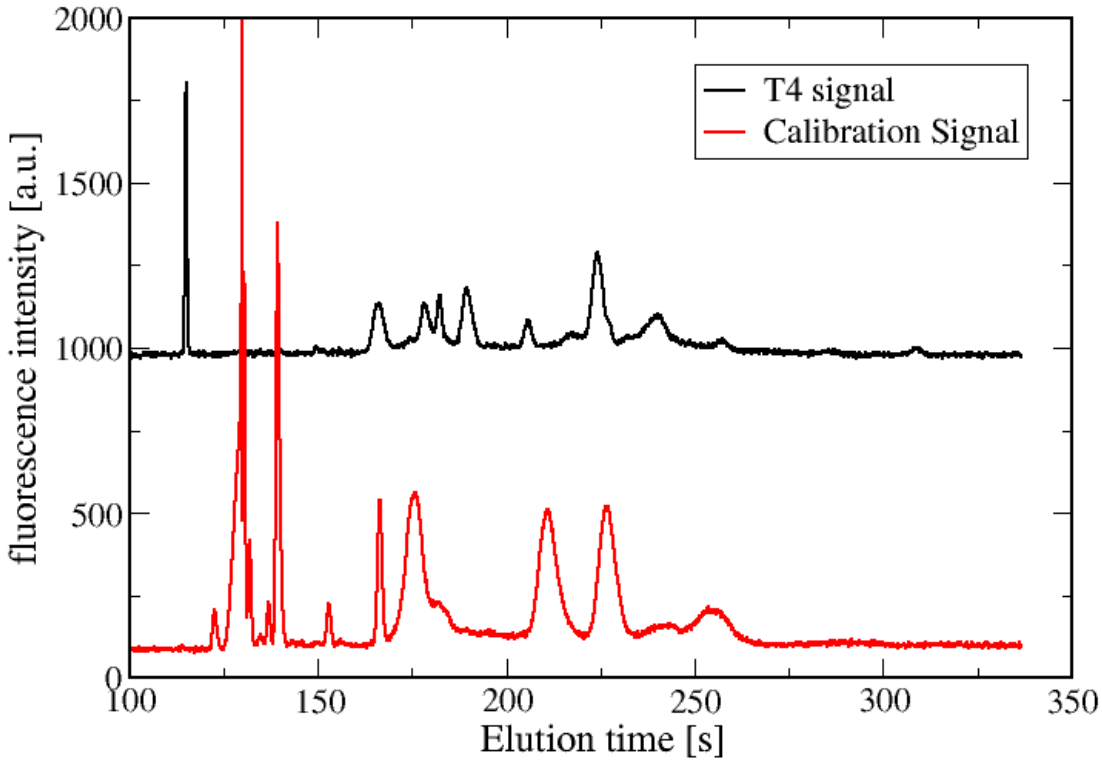
SAND2007-1628P

Bayesian Data Analysis for μChemLab Detection

The challenge



- Infer analyte identity from CGE separation
- Calibration signal supplies time correction

Analyzed Data Sets:

- "February Data Set": data taken in February '05 (or before)
  - MS2 (28 runs)
  - RSV (14 runs)
  - Vaccinia (29 runs)
- "March Data Set": data taken in March '05
  - Epstein-Barr (13 runs)
  - MS2 (13 runs)
  - T2 (11 runs)
  - T4 (18 runs)
- Total of 126 runs

Overview of analysis methodology

1. Preprocess calibration signal
2. Identify standards in calibration signal
3. Infer mapping to reference signal
4. Preprocess analyte signal
5. Classify analyte

Analyte signal preprocessing

1. Filter out spikes
2. Smooth with Savitzky-Golay filter
3. Remove baseline
4. Detect peaks
5. Take n largest peaks
6. Apply time shift obtained from calibration channel

Analyte classification

- Uses maximum likelihood over all agents in the data base

$$\underbrace{\mathrm{P}\left(C_i \middle| A_1, \ldots, A_N\right)}_{\text{Probability of substance } C_i, \text{ given the attributes } A_j \text{ in the signal}} \propto \underbrace{\mathrm{P}\left(A_1, \ldots, A_N \middle| C_i\right)}_{\text{Likelihood of attributes } A_j, \text{ given substance } C_i}$$

$$\approx \mathrm{P}\left(A_1 \middle| C_i\right)\mathrm{P}\left(A_2 \middle| C_i\right)\cdots\mathrm{P}\left(A_N \middle| C_i\right)$$

- Analyte signals shifted based on manually identified standards peaks in calibration signal
- Location of 5 largest peaks chosen as attributes

Hints:

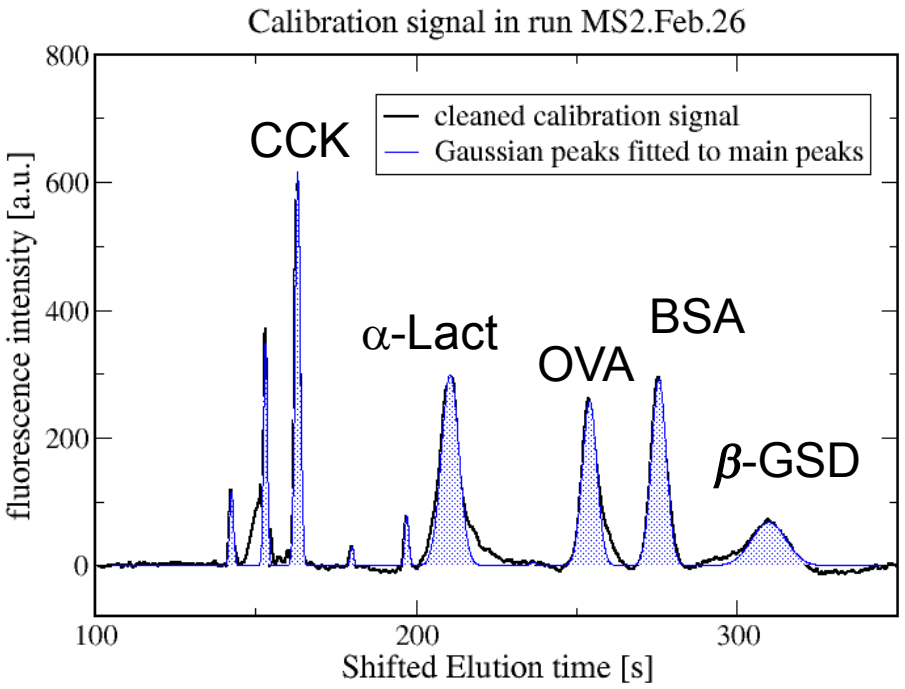
Full Bayesian approach means:

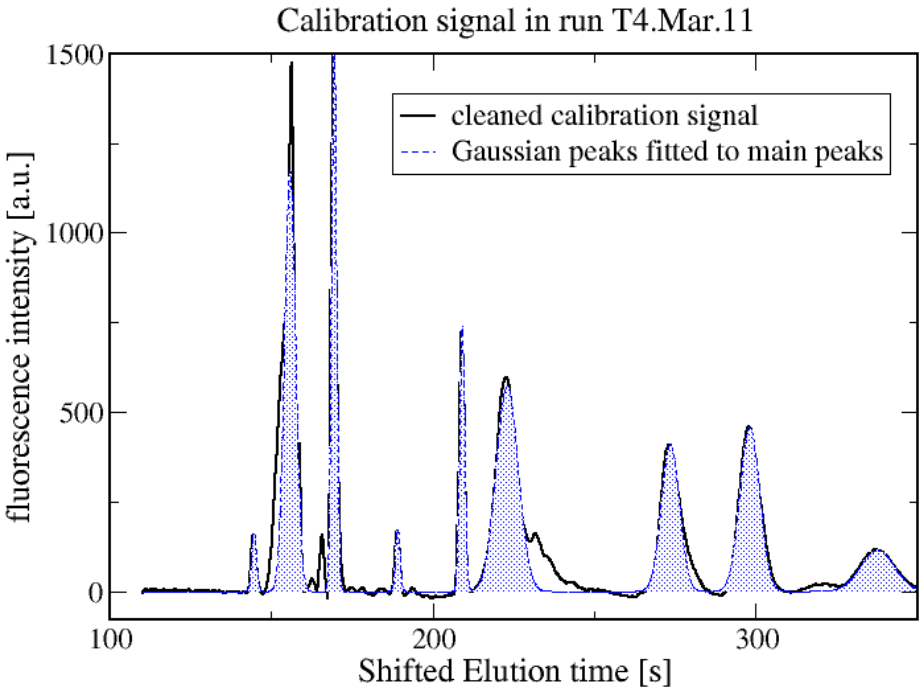
- Explicit use of prior knowledge
  - Selection of discriminating attributes
- Modeling stochastic variability using PDFs
  - Time shift
  - Elution time variability
  - Spurious peaks

Internal calibration makes detection more robust, but is challenged by variability in attributes

Peak location is a strong predictor

Decay products from standards add new peaks



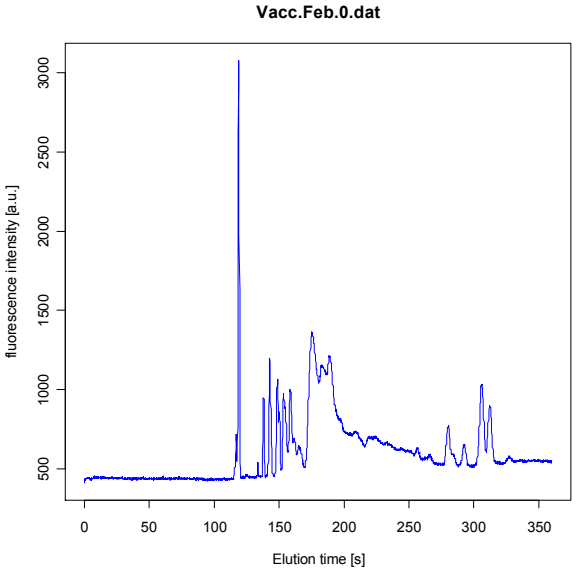


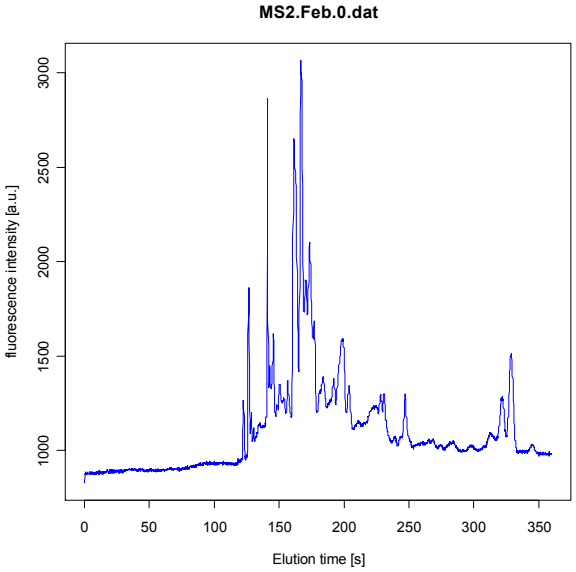
- A new peaks appears next to α-Lact, most likely degradation products of α-Lact
- Peak height similar to α-Lact
- Distorts signal pre-shifting
- Include as part of standards signature or differentiate based on peak width

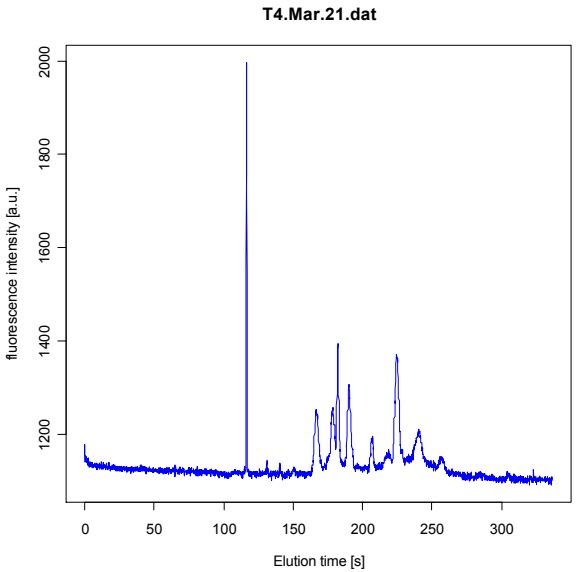
Implementing Full Bayesian Analysis

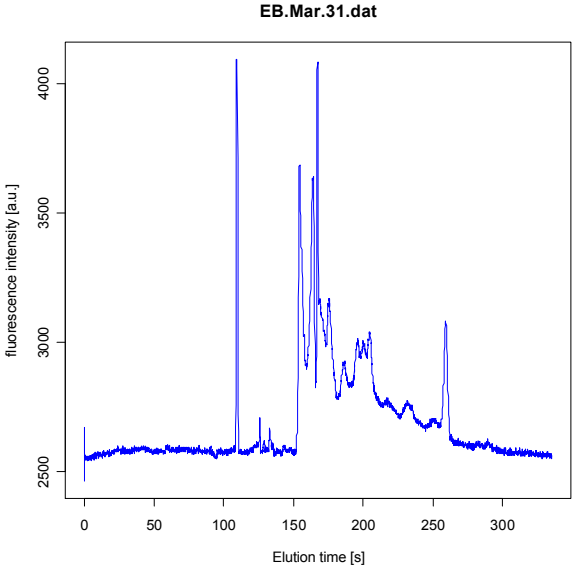
- Learning from data
- Explicit formulation of statistical model
  - Selection of attributes
  - Use PDFs to characterize unknown quantities (time shift, peak location…)
  - Prior probabilities for different classes (the agents)

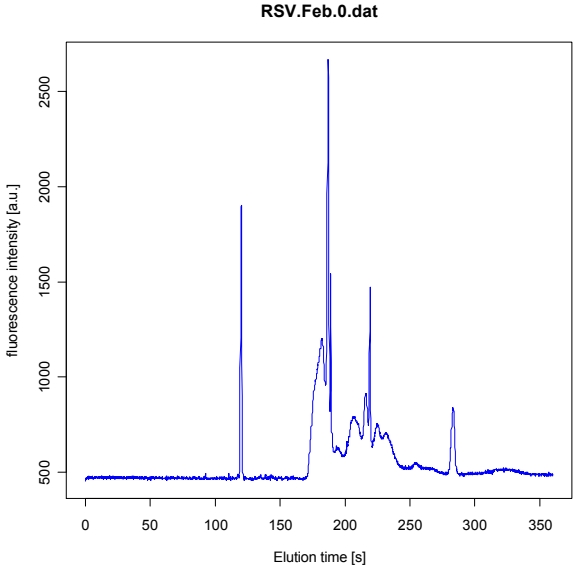
CGE spectra for different agents

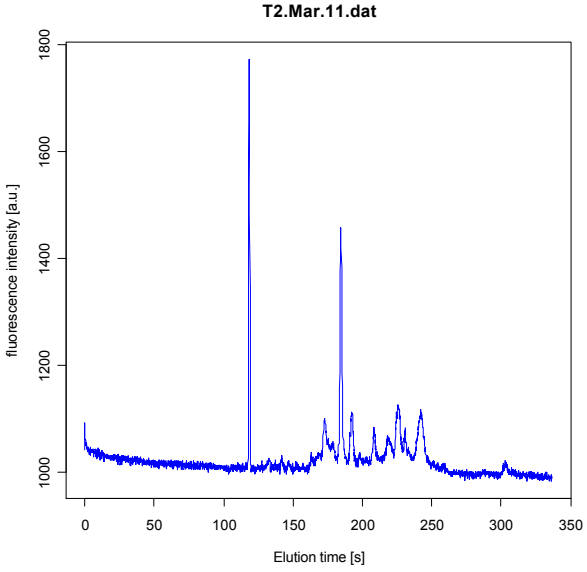






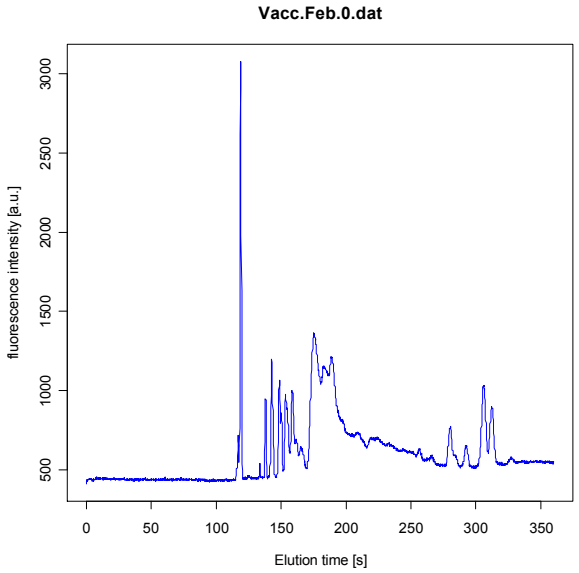




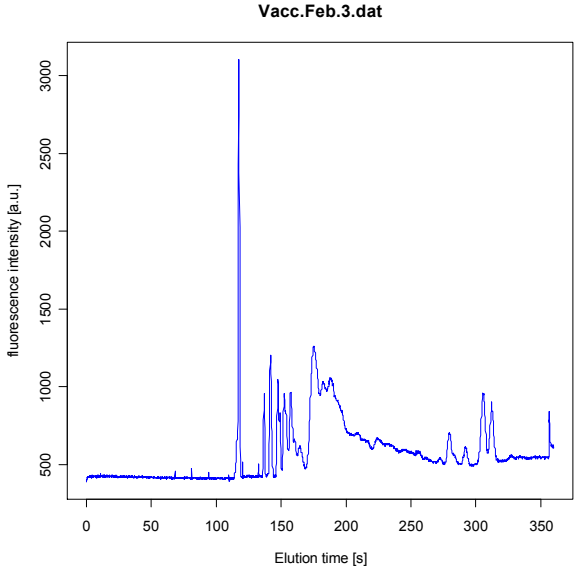


Analyte signal preprocessing

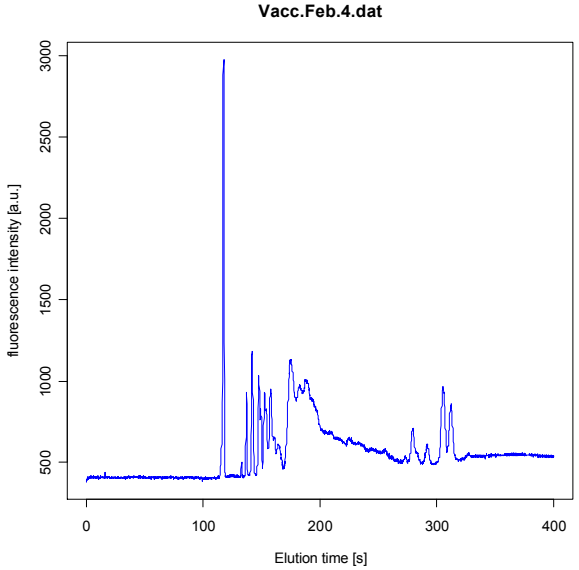
spike removal, smoothing, baseline correction, detection and selection of most significant peaks



time	amplitude
118.90	2501.633
138.00	454.538
142.90	711.156
148.70	566.627
153.50	485.319
158.60	506.787
175.50	850.016
188.80	692.541
305.90	481.619
312.20	344.298



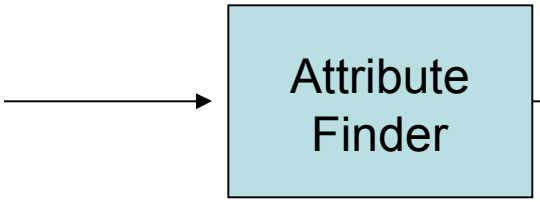
time	amplitude
117.50	2578.768
137.00	473.291
141.80	729.185
147.60	574.582
152.40	478.509
157.50	489.451
175.00	769.757
305.60	415.470
312.20	318.156
356.80	256.960



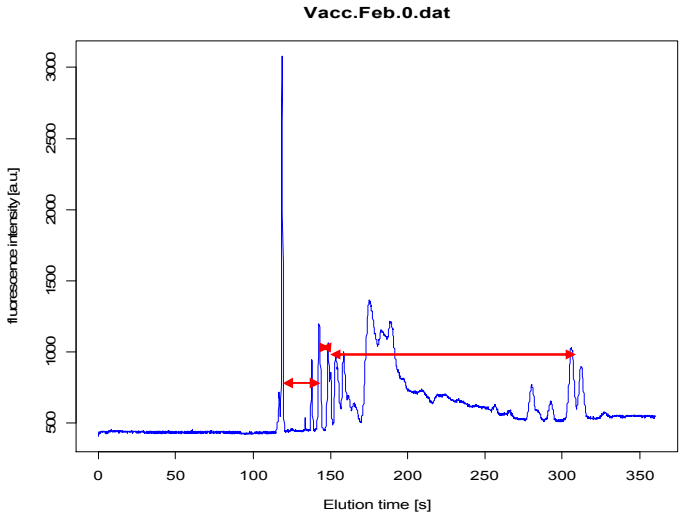
time	amplitude
117.30	2454.998
137.00	467.677
141.90	723.768
147.70	573.881
152.50	477.670
157.60	489.051
175.00	657.384
279.40	181.800
305.30	430.558
312.00	317.412

Training: learning from data

Raw Data
Vacc.Feb.0.dat



Elution Times / Amplitudes
of most significant peaks
Vacc.Feb.0.scattributes.dat



Time	Amplitude
118.9	2501.633
138.0	454.5384
142.9	711.1556
148.7	566.6274
153.5	485.3187
158.6	506.7869
175.5	850.0159
188.8	692.5406
305.9	481.619
312.2	344.298

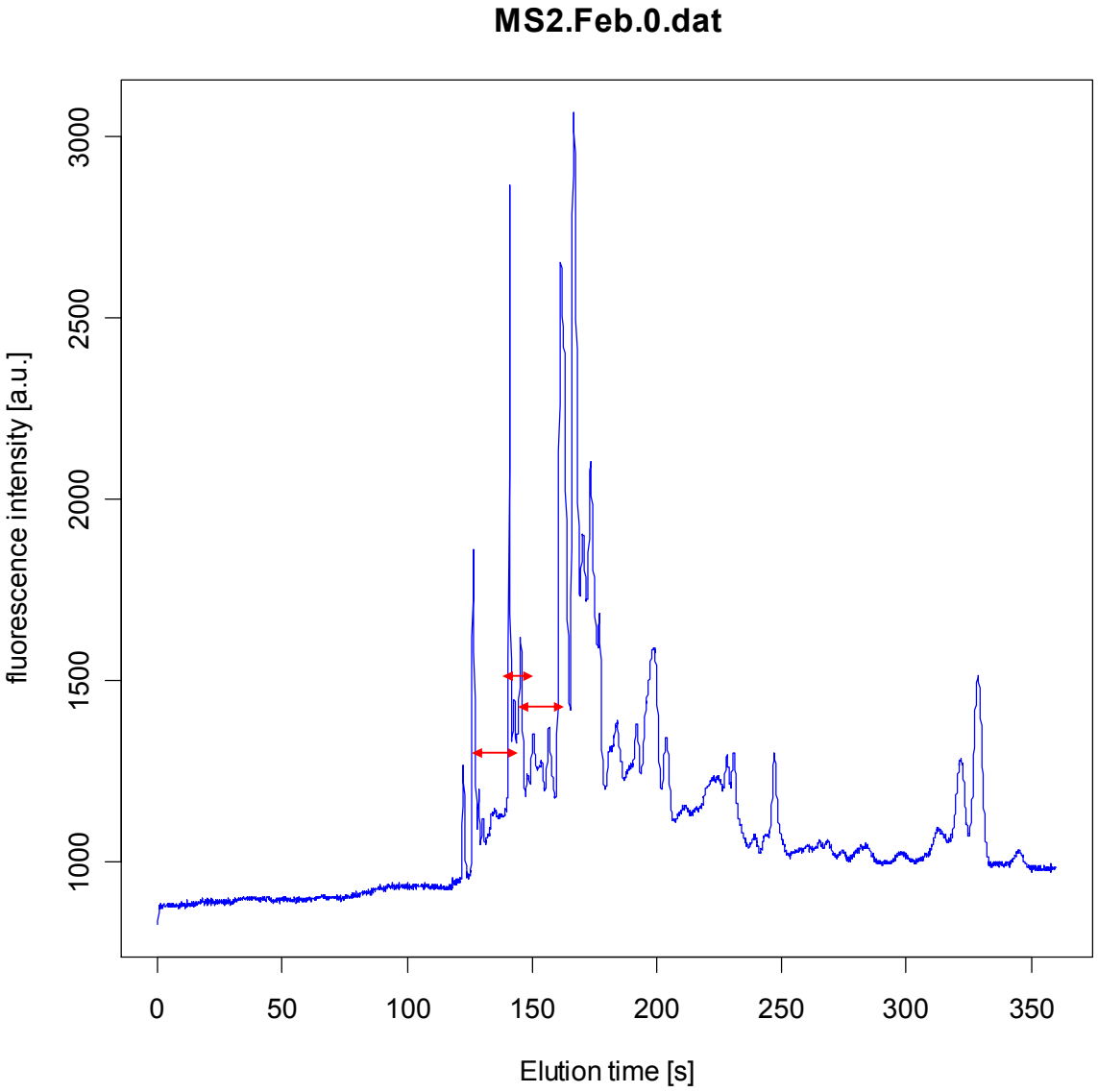
Probabilistic representation of Vaccinia pattern:

d_{12} ~ N (mean = 23.70, var = 0.188)

d_{23} ~ N (mean = 5.80, var = 0.004)

d_{34} ~ N (mean = 158.66, var = 3.538)

Training: attribute characterization



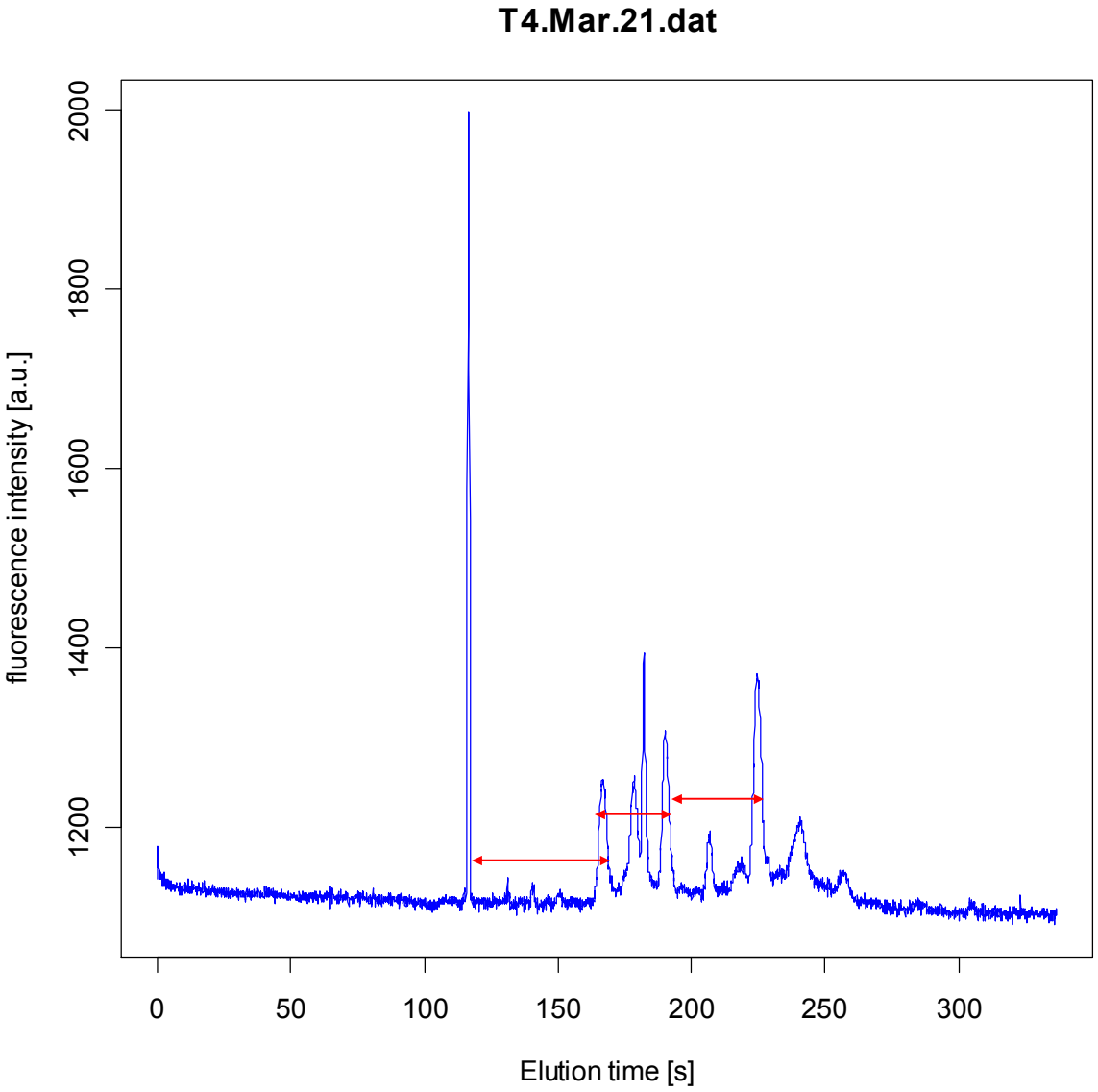
MS2 :

d_{12} ~ N (mean = 13.87, var = 0.344)

d_{23} ~ N (mean = 19.90, var = 1.207)

d_{34} ~ N (mean = 5.23, var = 0.104)

Training: attribute characterization



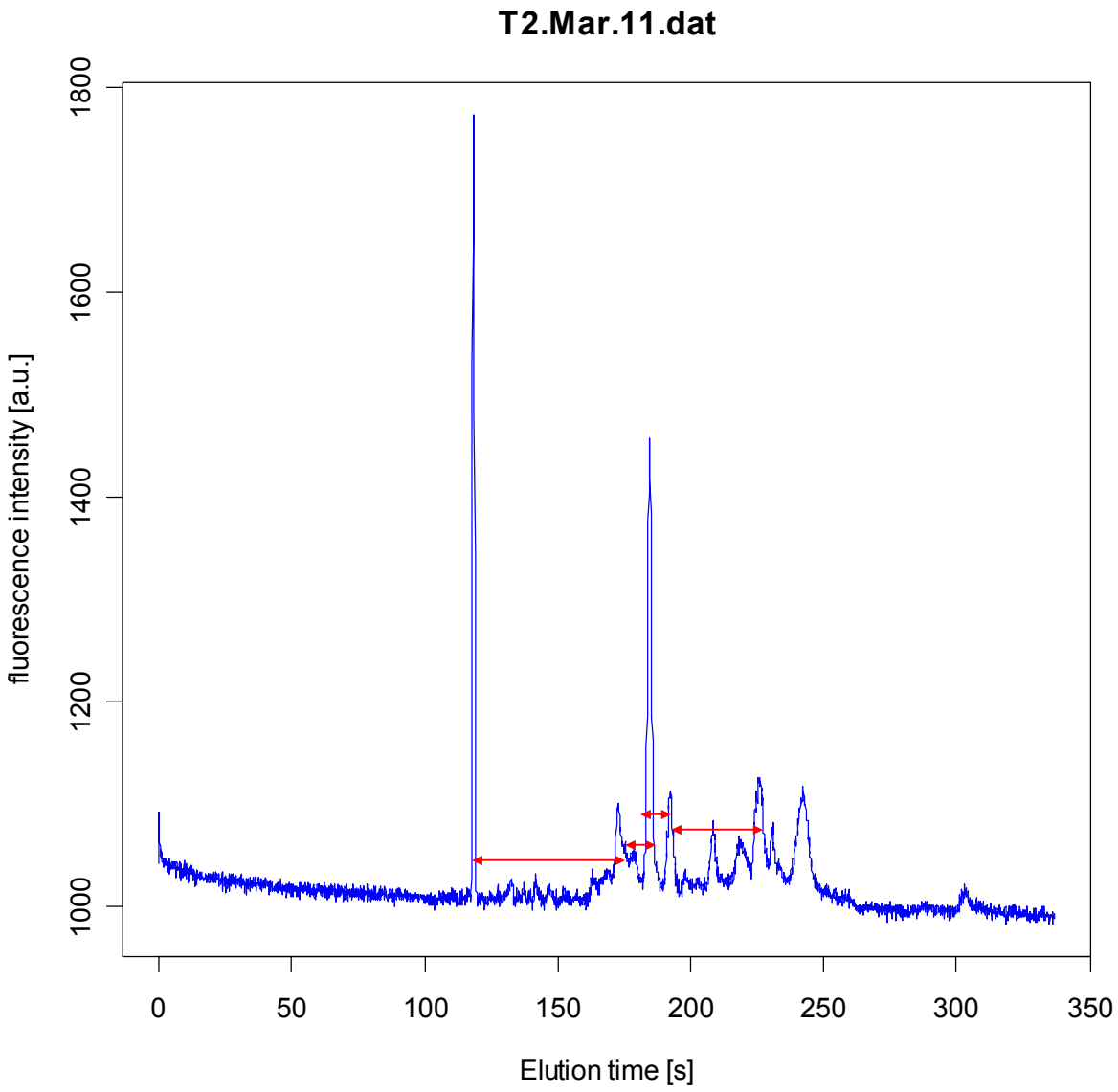
T4 :

d_{12} ~ N (mean = 50.62, var = 0.188)

d_{23} ~ N (mean = 23.52, var = 0.004)

d_{34} ~ N (mean = 34.39, var = 3.538)

Training: attribute characterization



T2 :

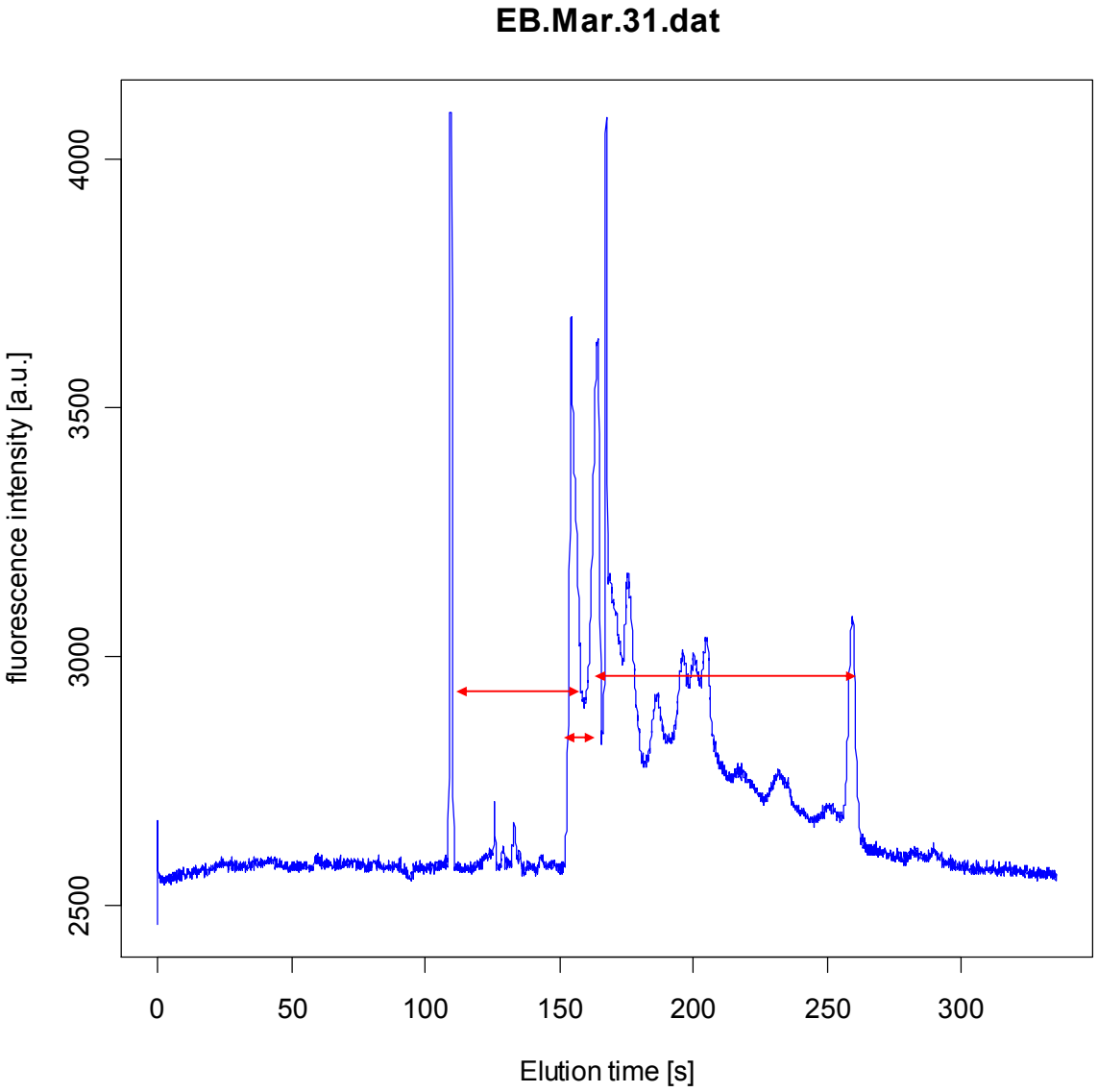
d_{12} ~ N (mean = 59.25, var = 0.188)

d_{23} ~ N (mean = 6.48, var = 0.004)

d_{34} ~ N (mean = 8.05, var = 3.538)

d_{45} ~ N (mean = 33.16, var = 3.538)

Training: attribute characterization



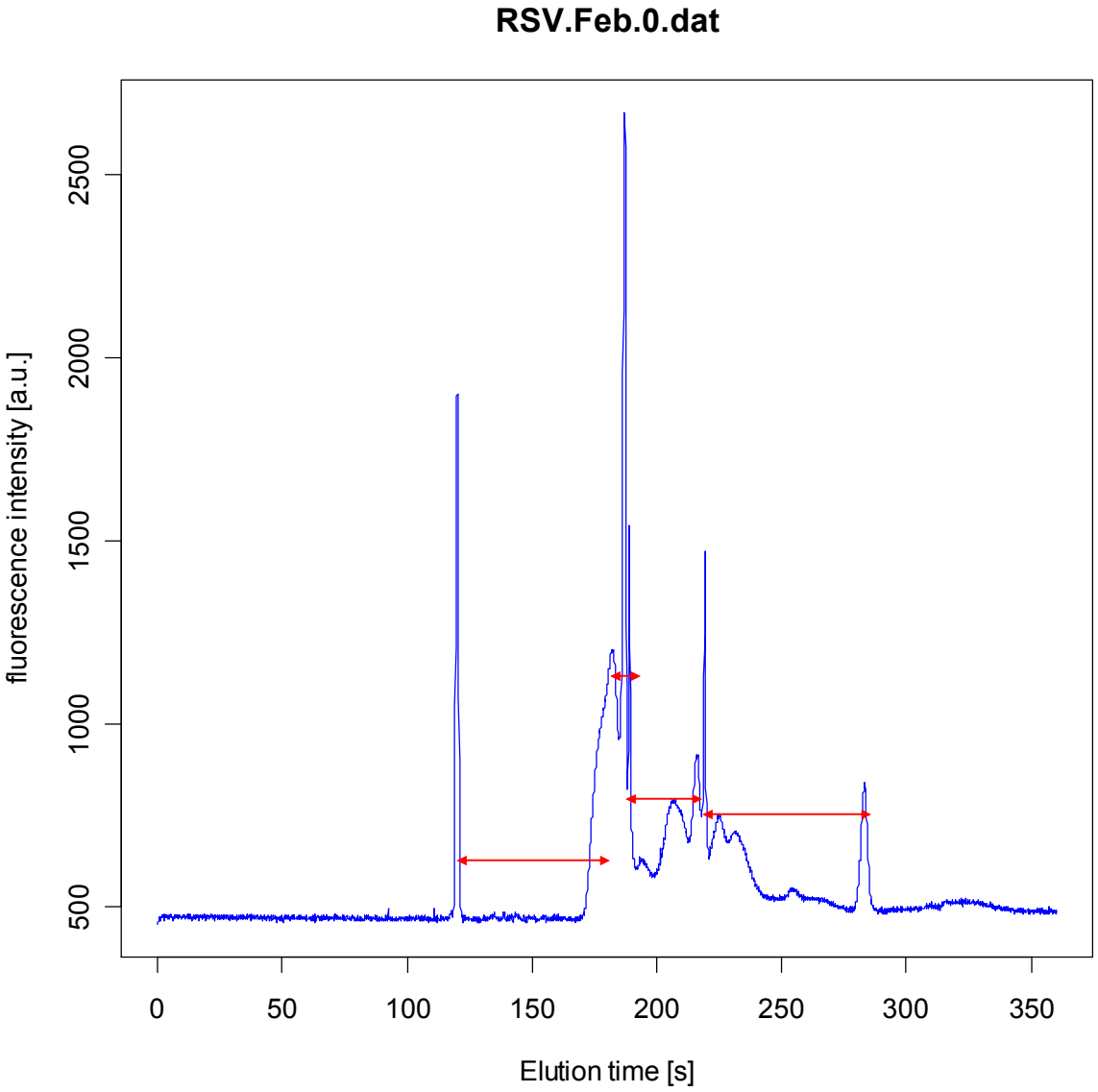
EB :

d_{12} ~ N (mean = 45.11, var = 0.188)

d_{23} ~ N (mean = 10.16, var = 0.004)

d_{34} ~ N (mean = 94.96, var = 3.538)

Training: attribute characterization



RSV :

d_{12} ~ N (mean = 59.85, var = 3.202)

d_{23} ~ N (mean = 5.70, var = 0.939)

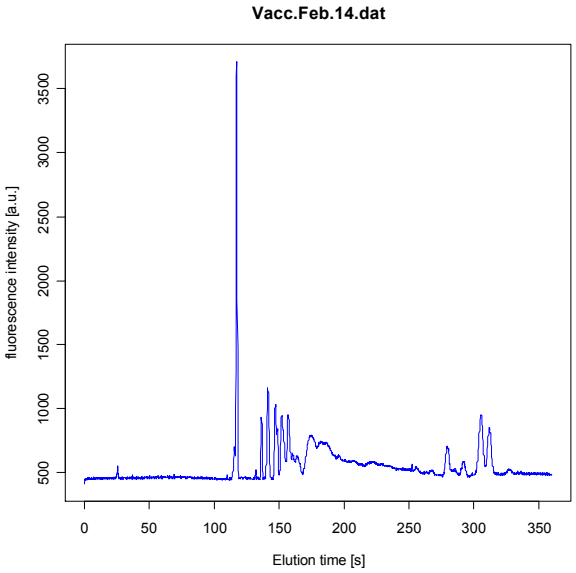
d_{34} ~ N (mean = 28.52, var = 1.201)

d_{45} ~ N (mean = 69.87, var = 1.187)

Classification

Finding the attributes in an unknown sample. SA optimization

Vacc.Feb.14.dat



Time	Amplitude
117.3	3140.424
136.5	432.514
141.4	675.7318
147.2	538.1851
152	450.2342
157.1	467.5786
279.6	189.5629
305.4	446.1932
311.9	346.7245

24.1

5.8

159.9

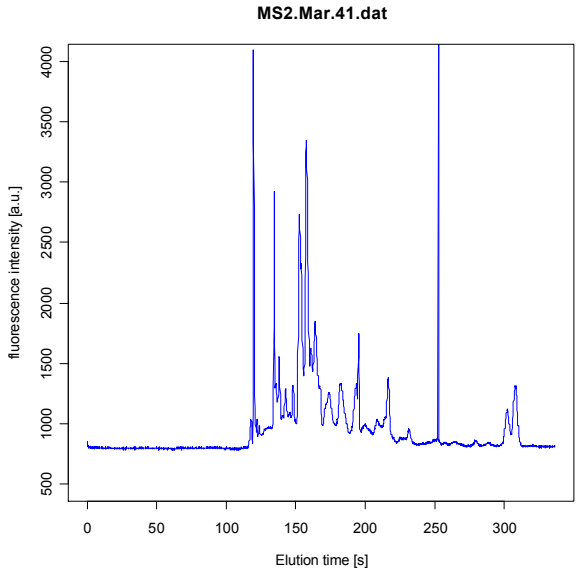
Vaccinia pattern:

d_{12} ~ N (23.70, 0.188)

d_{23} ~ N (5.80, 0.004)

d_{34} ~ N (158.66, 3.538)

MS2.Mar.41.dat



Time	Amplitude
119.8	2633.082
134.5	1107.37
147.8	421.6494
152.6	1861.521
157.6	2446.918
163.9	963.0131
182.3	445.8168
195.2	726.4242
216.5	500.4109
307.9	500.0817

23.1

6.3

144.0

Bayesian rule for classification

- Posterior Probabilities of agent presence:

$$p\left(Vaccinia \mid data\right) = \frac{p\left(data \mid Vaccinia\right) p\left(Vaccinia\right)}{p(data)}$$

$$p\left(MS2 \mid data\right) = \frac{p\left(data \mid MS2\right) p\left(MS2\right)}{p(data)}$$

$$data = \{attributes, other\}$$

- Ratio of posterior probabilities (Bayes Factor)

$$\frac{p\left(Vaccinia \mid data\right)}{p\left(MS2 \mid data\right)} = \frac{p\left(data \mid Vaccinia\right) p\left(Vaccinia\right)}{p\left(data \mid MS2\right) p\left(MS2\right)}$$

Bayesian rule for classification

Likelihood function (predictive probability of data):

observations are assumed to be independent

$$p(data \mid agent) = p(attributes, other \mid agent) = p(attributes \mid agent) \cdot p(other \mid agent)$$

Likelihood of the data for each hypothesis

$$p(attributes_{vacc} \mid Vacc) = \prod_{i=1}^{nattrib\_vacc} p(y_i \mid Vacc) \qquad p(attributes_{MS2} \mid MS2) = \prod_{i=1}^{nattrib\_MS2} p(y_i \mid MS2)$$

$$p(other_{vacc} \mid Vacc) = \prod_{j=1}^{nother\_vacc} U(t_0, t_f) \qquad p(other_{MS2} \mid MS2) = \prod_{j=1}^{nother\_MS2} U(t_0, t_f)$$

$$data = \{attributes_{vacc}, other_{vacc}\} = \{attributes_{MS2}, other_{MS2}\}$$

Bayesian rule for classification

Bayes Factor:

$$B_{Vacc,MS2} = \frac{p(Vaccinia \mid data)}{p(MS2 \mid data)} = \frac{\prod_{i=1}^{n\_vacc} p(y_i \mid Vacc) \prod_{j=1}^{other\_vacc} U(t_0, t_f)}{\prod_{i=1}^{n\_ms2} p(y_i \mid MS2) \prod_{j=1}^{other\_ms2} U(t_0, t_f)} \frac{p(Vacc)}{p(MS2)}$$

More convenient representation using log transform

$$\log\left(\frac{p(Vaccinia \mid data)}{p(MS2 \mid data)}\right) = \log\left(p(Vaccinia \mid data)\right) - \log\left(p(MS2 \mid data)\right)$$

Interpretation

$B_{Vacc,MS2} > 1$

$\log\left(B_{Vacc,MS2}\right) > 0$

Odds in favor of Vaccinia

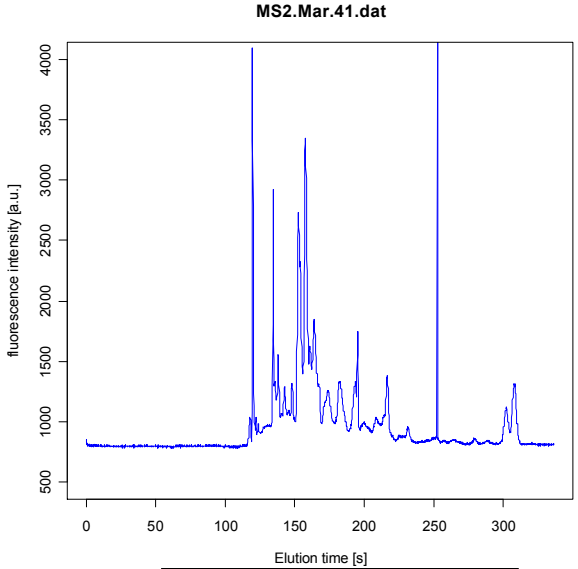
$B_{Vacc,MS2} < 1$

$\log\left(B_{Vacc,MS2}\right) < 0$

Odds in favor of MS2

Ratio of posterior probabilities

MS2.Mar.41.dat



$$\log\left(\frac{p(Vaccinia \mid data)}{p(MS2 \mid data)}\right) = \log\left(\frac{p(data \mid Vacc)}{p(data \mid MS2)}\right) + \log\left(\frac{p(Vacc)}{p(MS2)}\right)$$

	Time	Amplitude	
	119.8	2633.082	
14.7	134.5	1107.37	23.1
18.1	147.8	421.6494	
	152.6	1861.521	
5.0	157.6	2446.918	
	163.9	963.0131	6.3
	182.3	445.8168	
	195.2	726.4242	144.0
	216.5	500.4109	
	307.9	500.0817	

Vaccinia pattern:
d12 ~ N (23.70, 0.188)
d23 ~ N (5.80, 0.004)
d34 ~ N (158.66, 3.538)

Log Likelihood = -91.492

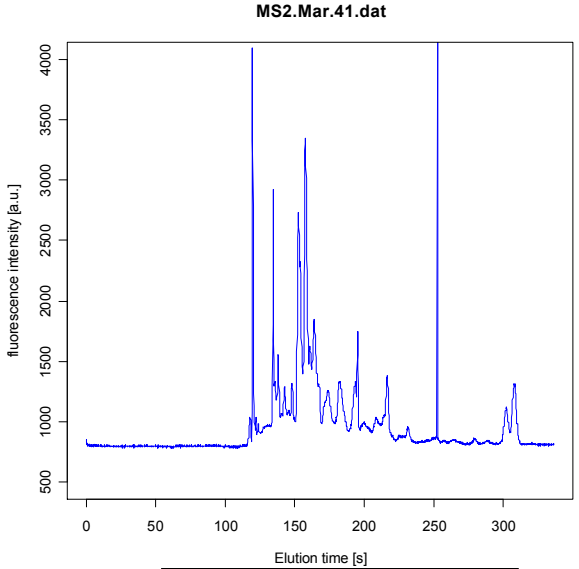
MS2 pattern:
d12 ~ N (13.87, 0.347)
d23 ~ N (19.90, 1.207)
d34 ~ N (5.23, 0.104)

Log Likelihood = -32.620

log p(data | Vacc) – log p(data | MS2) < 0
data strongly favors MS2 presence hypothesis

Vaccinia or something else?

MS2.Mar.41.dat



Time	Amplitude
119.8	2633.082
134.5	1107.37
147.8	421.6494
152.6	1861.521
157.6	2446.918
163.9	963.0131
182.3	445.8168
195.2	726.4242
216.5	500.4109
307.9	500.0817

23.1

6.3

144.0

$$\log\left(\frac{p(Vaccinia \mid data)}{p(S.E. \mid data)}\right) = \log\left(\frac{p(data \mid Vacc)}{p(data \mid S.E.)}\right) + \log\left(\frac{p(Vacc)}{p(S.E.)}\right)$$

Vaccinia pattern:
d12 ~ N (23.70, 0.188)
d23 ~ N (5.80, 0.004)
d34 ~ N (158.66, 3.538)

Log Likelihood = -91.492

"Something else" statistical model:
$peak_i \sim U(t_0, t_f)$

Log Likelihood = -51.914

log p(data | Vacc) – log p(data | s.e.) < 0
data strongly favors Vaccine absence hypothesis

Classification Results for Experimental Data February – March 2005

- Training Data
  - Vacc 12 / 29 { Feb.0, Feb.1, Feb.3, …, Feb.12}
  - EB 8 / 13 { Mar.21, Mar.31,…, Mar.91}
  - RSV 5 / 14 { Feb.0, Feb.1, Feb.8, Feb.9, Feb.10}
  - T4 9 / 18 { Mar.11, Mar.21,…, Mar.91}
  - T2 6 / 11 { Mar.11, Mar.21,…, Mar.61}
  - MS2 17 / 41 { Feb.0, Feb.1,…, Feb.18}

    Total 63 / 126

- Test Data: all (but only the test on skipped data is really meaningful).

Classification Results for Experimental Data February – March 2005

- % Successful Classification with Training data:

100 % (this was easy!)

- % Successful Classification with All Data:

123 / 126 → 97.62 %

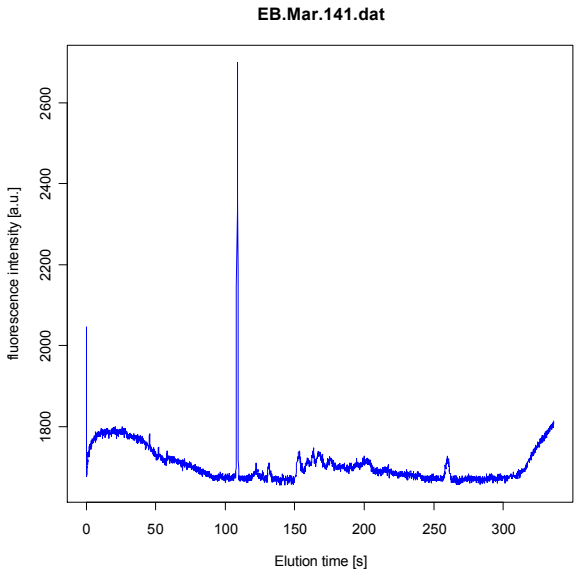
- % Successful Classification with Non-Training Data

60 / 63 → 95.24 %

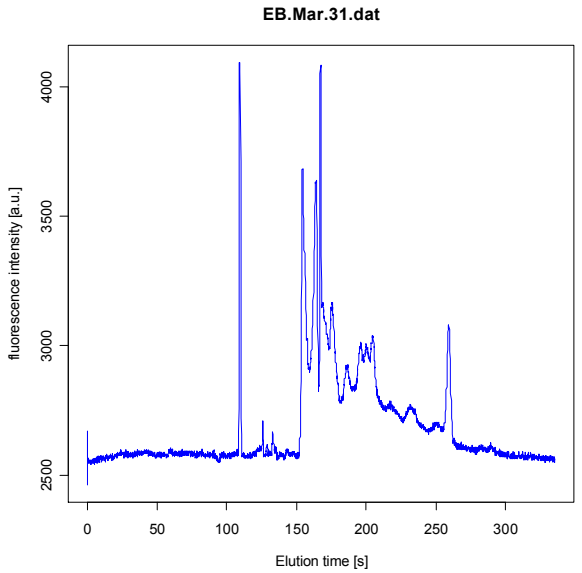
Classification Results for Experimental Data February – March 2005

3 Misclassified samples (1/3):

Identified as None



should be EB

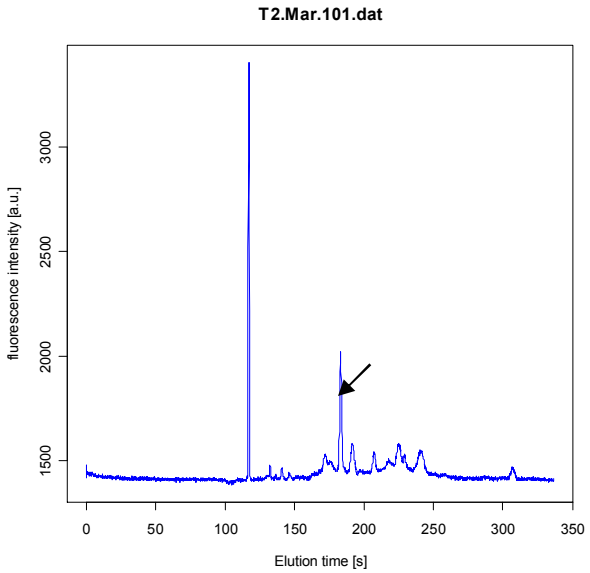


Reason: very weak signal

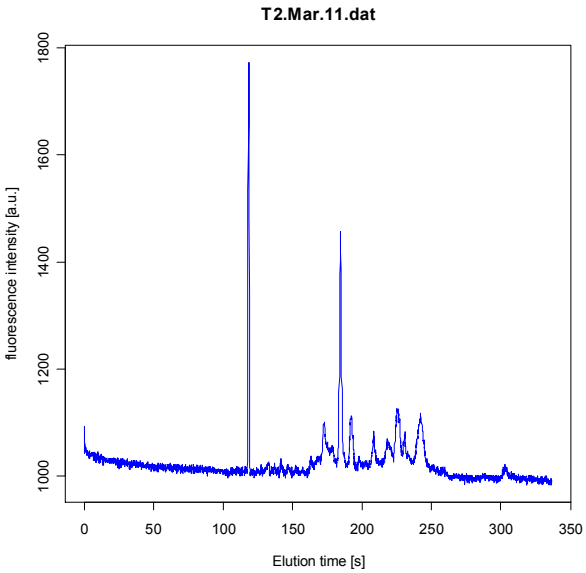
Classification Results for Experimental Data February – March 2005

3 Misclassified samples (2/3):

Identified as None



should be T2

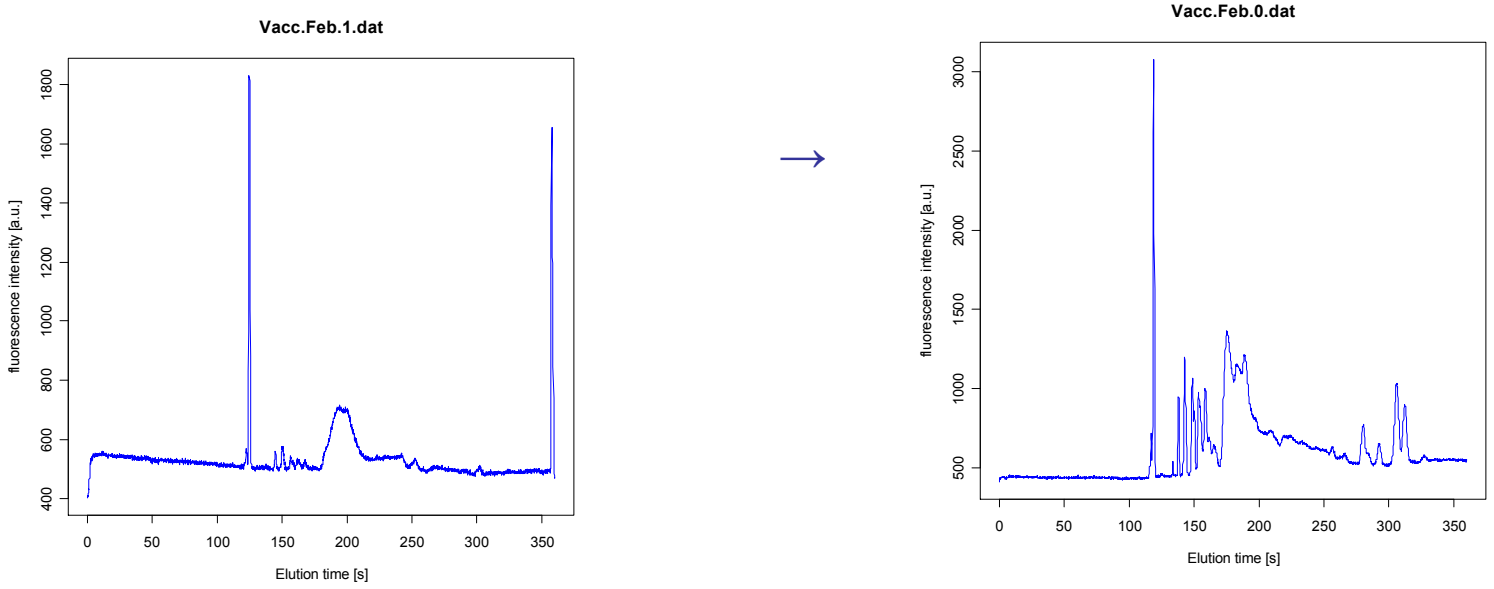


Reason: attribute finder missed significant peak

Classification Results for Experimental Data February – March 2005

3 Misclassified samples (3/3):

Identified as None → should be Vacc



Reason: not the expected Vaccinia pattern

Conclusions

Selective attribute picking improves classification

- seek discriminating peaks / patterns
- elution times are strong predictors

Bayesian approach allows for modeling flexibility

- null hypothesis
- presence of spurious peaks