Mass Spectrometry Analysis of Proteins Using Electron Transfer Dissociation

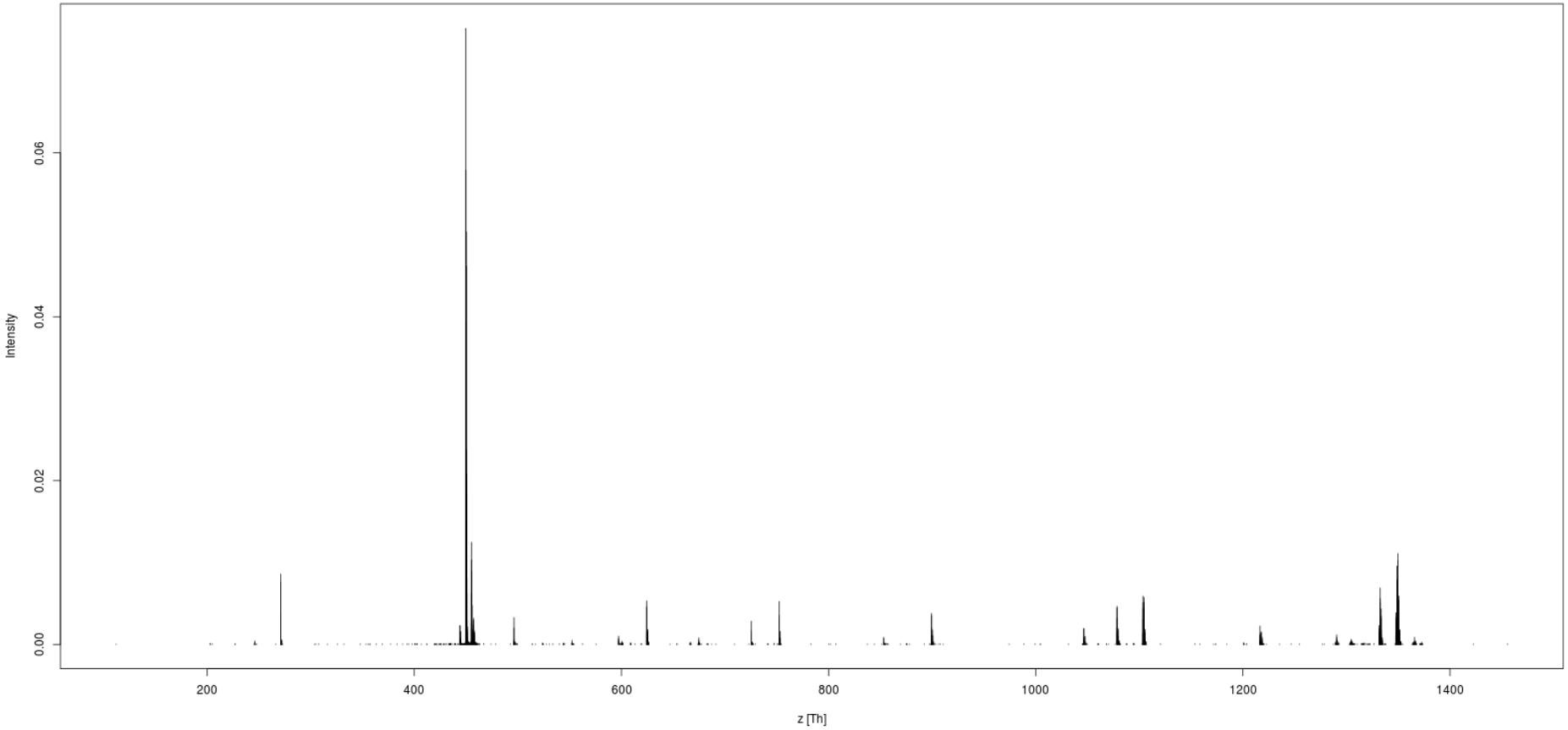
Statistical Model

Mateusz Łącki

Uniwersytet Warszawski

17 October 2013

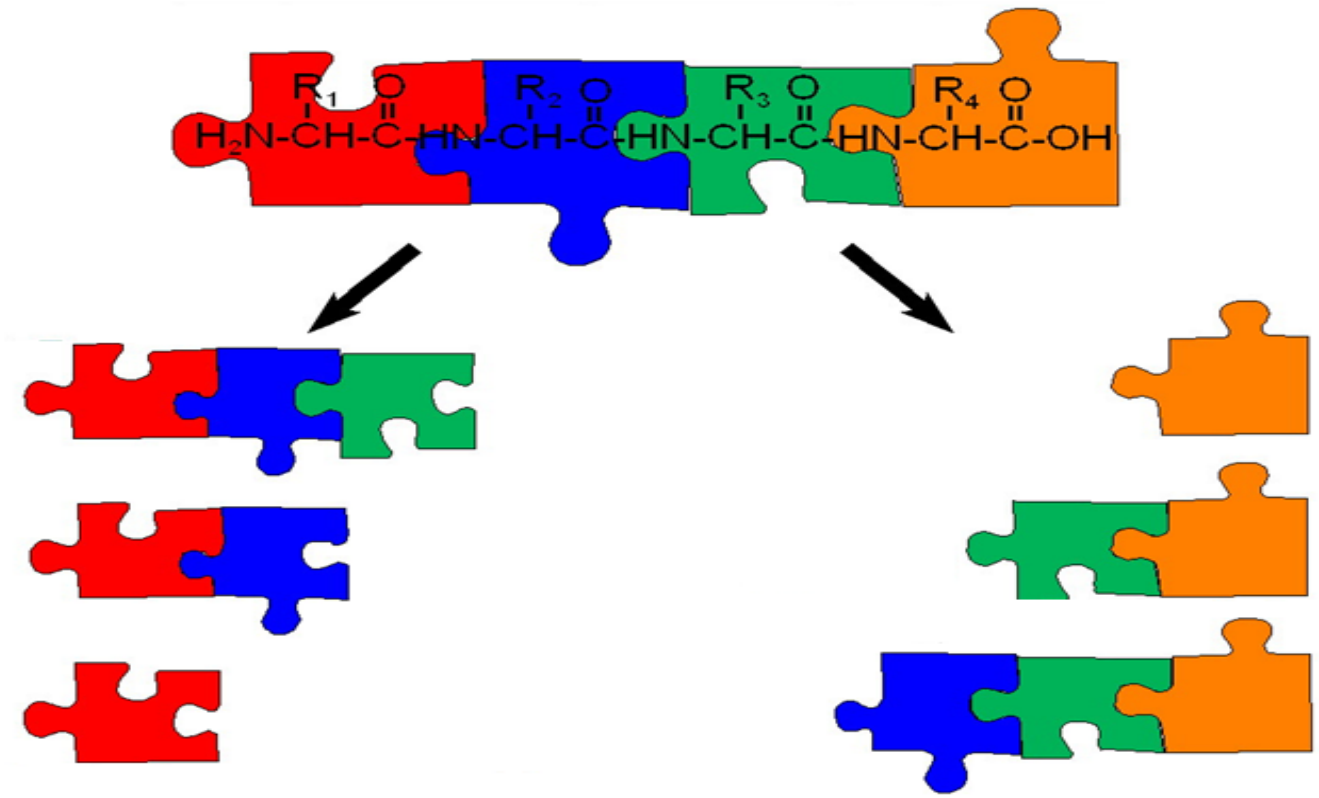
Data from a Mass Spectrometer for a given substance.



Project massTodon: Debriefing

- Mass Spectrometer
  - Evaluation of chemical composition of molecules
  - Measurements
    - $\star$ $\left\{ \frac{\text{Mass}_j}{\text{Charge}_j}, \text{Intensity}_j \right\}_j^J$
- We use MS/MS instrument
  - Coupling two mass specs
  - Filtering specific mass to charge
  - Use of the ETD instrument
- Why all that?
  - Study structure of peptides by inducing cleavages

Inducing Cleavages

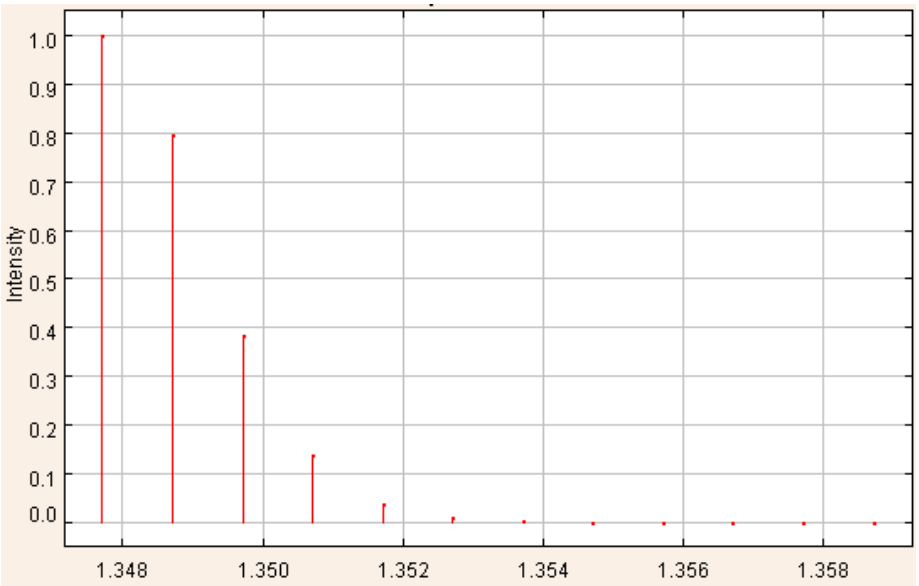


Today's Agenda

1. What is our motivation?
2. Statistical Modelling
3. Fitting Procedures
4. What remains to be done?

Multinomial Model

- What is being explained?
  - Distributions of masses
  - Deviations from monoisotopic peaks



Multinomial Model

- Modelling isotope distributions
  - $\mathcal{P}(^{16}\mathrm{O}) = 99.757$
  - $\mathcal{P}(^{17}\mathrm{O}) = 0.038$
  - $\mathcal{P}(^{18}\mathrm{O}) = 0.205$
- Molecule = $\mathrm{C_c H_h O_o N_n S_s}$
- Assumptions
  - Isotope variant of a single atom from $\mathrm{C_c H_h O_o N_n S_s}$(e.g. C) independent of isotope variants of other atoms

    i.e. for a molecule with 2 atoms

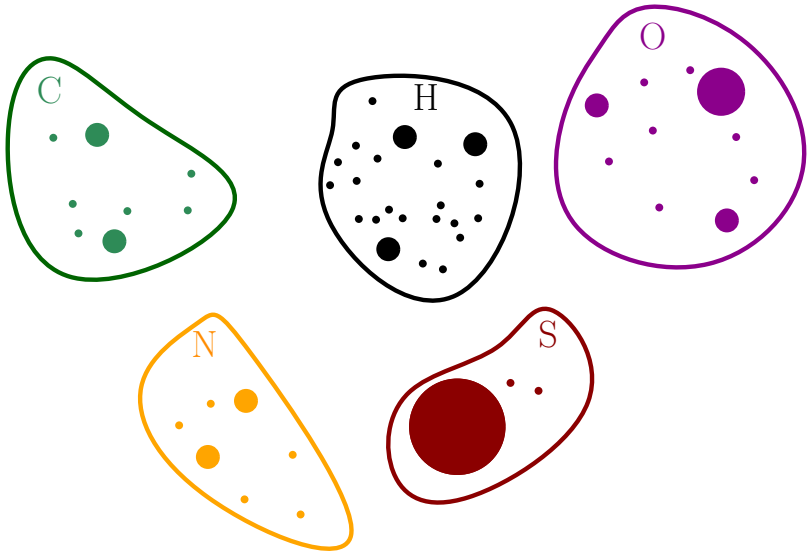
$$\mathcal{P}(^{13}\mathrm{C}^{17}\mathrm{O}) = \mathcal{P}(^{13}\mathrm{C}) \times \mathcal{P}(^{17}\mathrm{O})$$

  - We cannot discern among isomers

$$^{13}\mathrm{C}^{17}\mathrm{O}^{17}\mathrm{O} \simeq {}^{17}\mathrm{O}^{13}\mathrm{C}^{17}\mathrm{O} \simeq {}^{17}\mathrm{O}^{17}\mathrm{O}^{13}\mathrm{C}$$

Multinomial Model

- Chemical compound = list of sets of atoms



$$\mathcal{P}(\underbrace{c_{12}\,{}^{12}C,\ c_{13}\,{}^{13}C}_{c_{12}+c_{13}=c},\ \underbrace{h_1\,{}^{1}H,\ h_2\,{}^{2}H}_{h_1+h_2=h},\ \ldots,\ s_{36}\,{}^{36}S) =$$

$$\binom{c}{c_{12}, c_{13}}\mathcal{P}({}^{12}C)^{c_{12}}\mathcal{P}({}^{13}C)^{c_{13}}\binom{h}{h_1, h_2}\mathcal{P}({}^{1}H)^{h_1}\mathcal{P}({}^{2}H)^{h_2}\ldots$$

Multinomial Model and Molecular Mass

- Atomic mass m of a molecule $R = (\mathrm{c}_{12}, \mathrm{c}_{13}, \ldots \mathrm{s}_{36})$
  - ⋆ $m_R = \mathrm{c}_{12} m_{^{12}\mathrm{C}} + \cdots + \mathrm{s}_{36} m_{^{36}\mathrm{S}}$
  - ⋆ $\mathcal{P}(m_R = m) = \sum \mathcal{P}(\mathrm{c}_{12}\,{}^{12}\mathrm{C}, \ldots, \mathrm{s}_{36}\,{}^{36}\mathrm{S})$

  s.t. $m_R = \mathrm{c}_{12} m_{^{12}\mathrm{C}} + \cdots + \mathrm{s}_{36} m_{^{36}\mathrm{S}}$

  :) Good news: theory operates on chemical formulas
    - No need to solve these equations!
    - given a formula F, we derive it's mass probability function, $p_F(m)$.

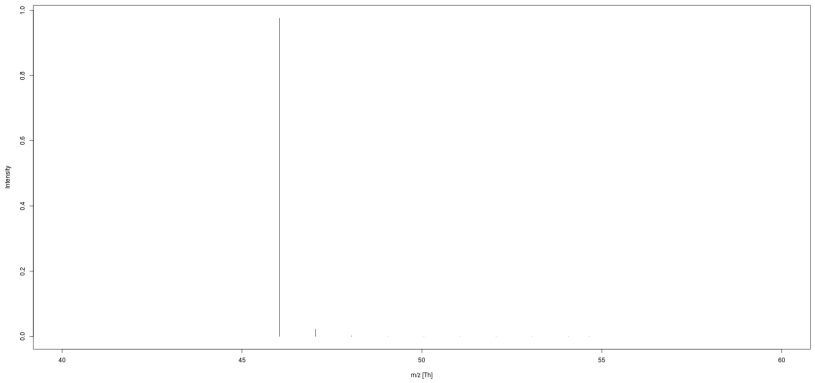
  :(Bad news
    - The number of peaks grow's quite big this way
- Cool thing
  - Masses of neutrons for different elements are not equal!
  - Modern mass specs can already discern them

★ Neglecting that phenomenon gives rise to BRAIN algorithm

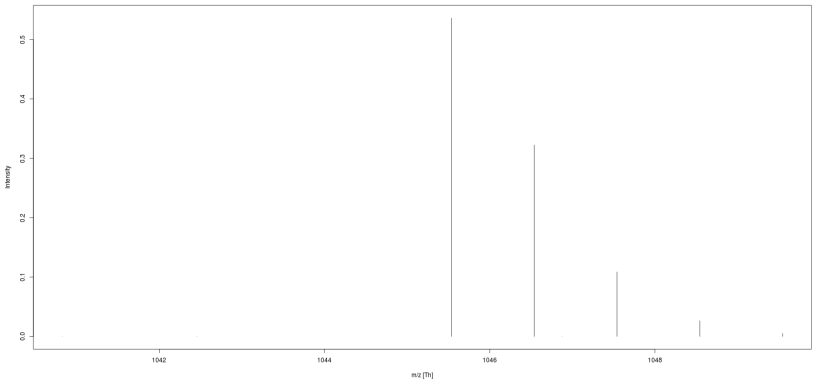
Visualising Multinomial Model

- BRAIN software - Piotr Dittwald®
- Assumption:
  - ⋆ All neutrons have equal masses and cannot be discerned.

Ethanol C_2H_5OH

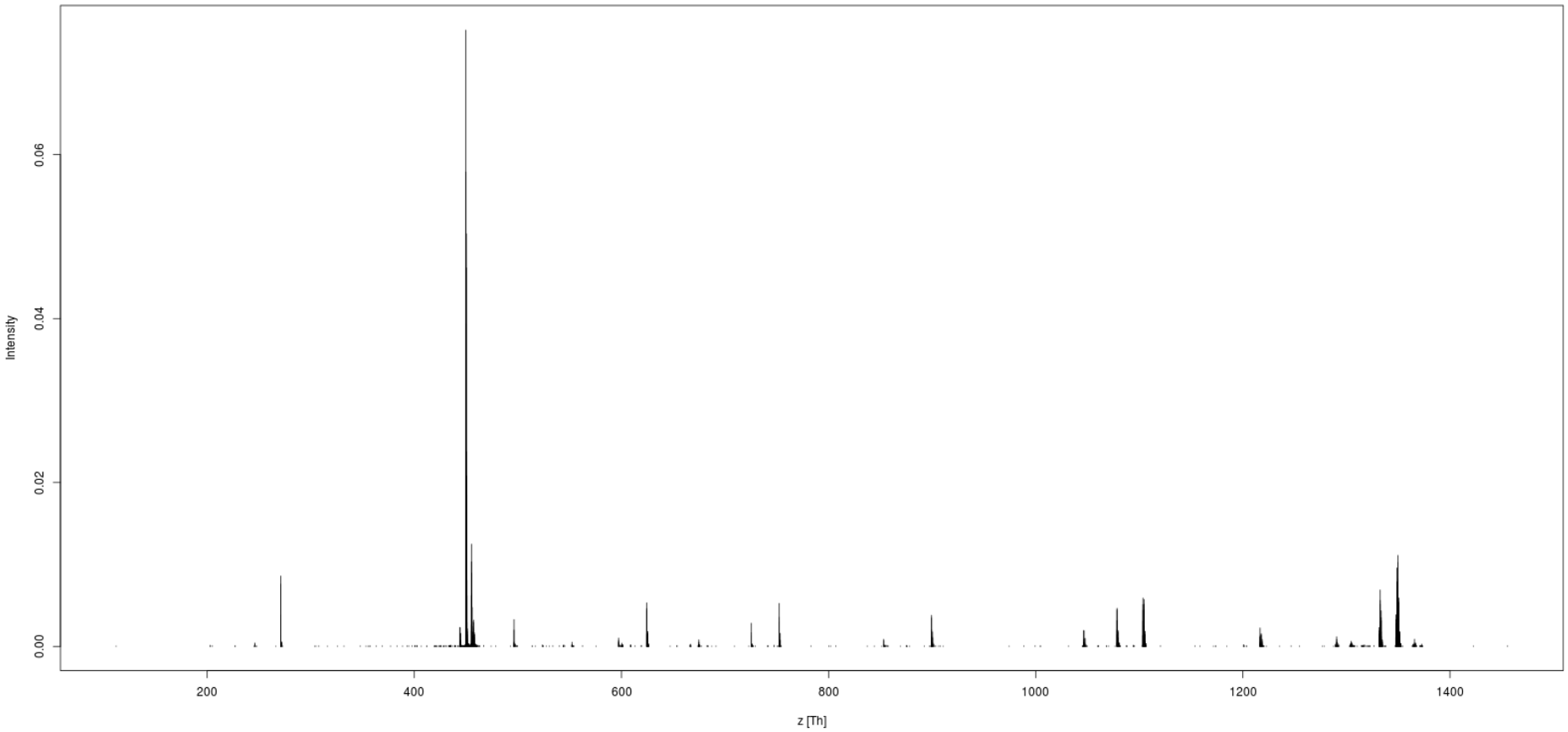


Angiotensine II $C_{50}H_{71}N_{13}O_{12}$

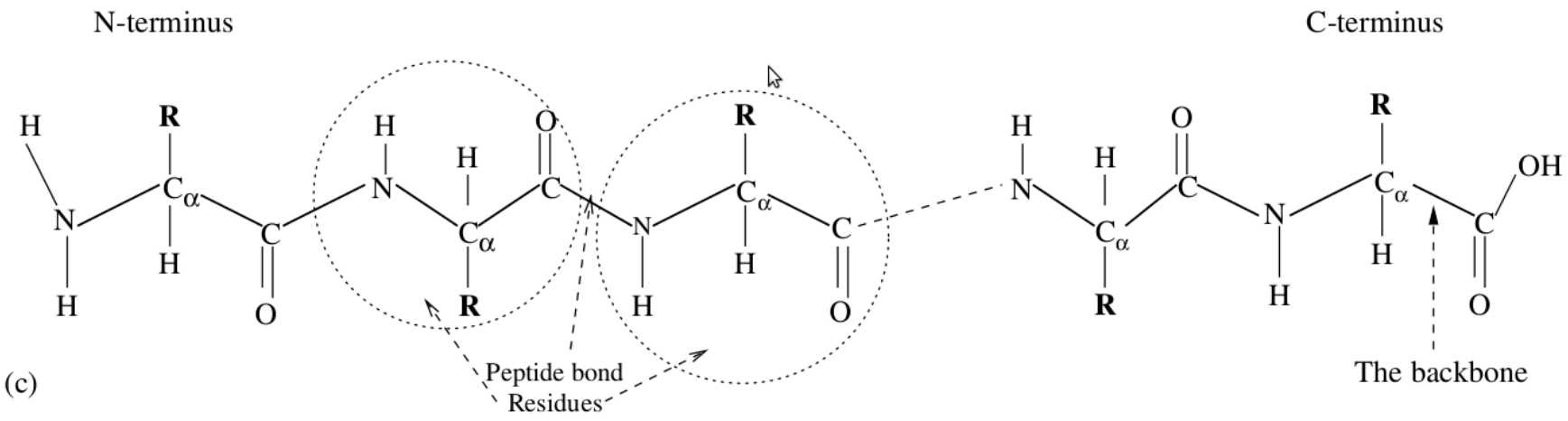


- Alas! Real spectra are multimodial!
  - fragmentation? (ETD)
  - charge reductions? (ETD, ETnoD, PTR)

Mass Spec results for substance P



Polymer as a sequence of Amino Acids

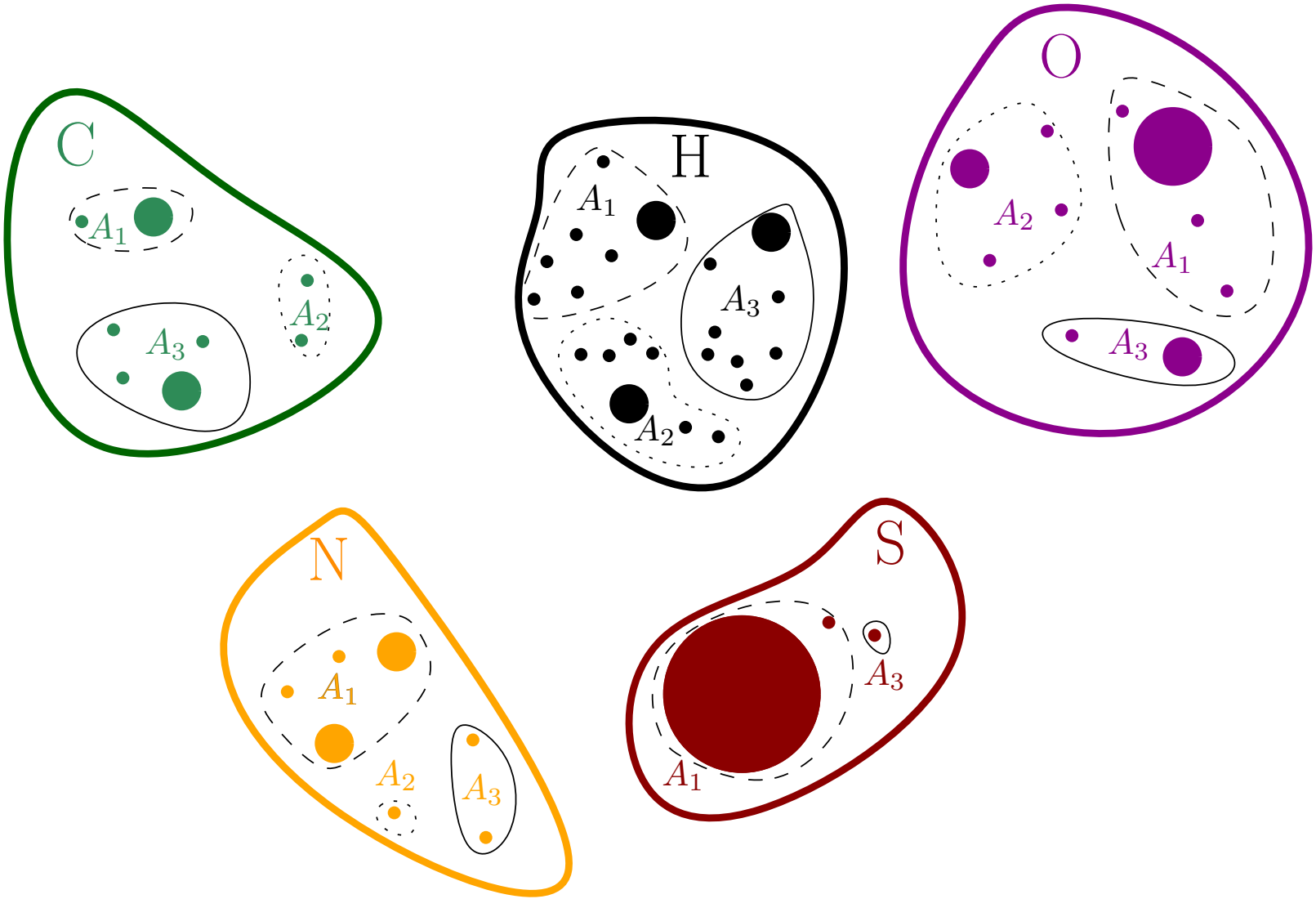


- Extra structure in our model must be added

  $C_cH_hO_oN_nS_s = A_1 A_2 \ldots A_k$

  $A_i \in \{$ Alanine, Cysteine, Aspartic Acid, Glutamic Acid, ...$\}$

Molecule Subdivided into Amino Acids



Electron Transfer Disociation

- Result: random cleavage of the peptide in two subsequences

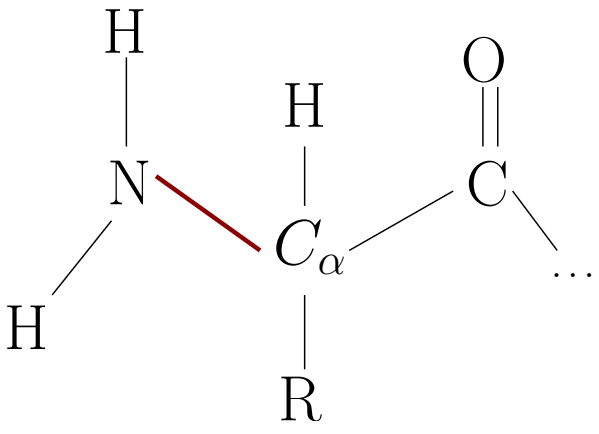
$$A_1 A_2 \ldots A_k \rightarrow \underbrace{A_1 \ldots A_L}_{C\text{ fragment}} \oplus \underbrace{A_{L+1} \ldots A_k}_{Z\text{ fragment}}$$

- L = index of the cleaved peptide bond (N-terminus to C-terminus)

- Assumption
  - Cleavage independent of isotope composition

$$\mathcal{P}(A_1 \ldots A_L = a_1 \ldots a_L | L = l) =$$

$$\mathcal{P}(A_1 \ldots A_l = a_1 \ldots a_l)$$



- Minor complication
  - Cleavage solely on A_1

Reactions considered by Frederik, our fellow chemist

Hypothesis:

- ⋆ Empirical spectrum = Result of Several Reactions

- ETD

$$[M+nH]^{n+} + A^{\cdot -} \longrightarrow [C+xH]^{x+} + [Z+(n-x)H]^{(n-x-1)\cdot} + A$$

- PTR

$$[M+nH]^{n+} + A^{\cdot -} \longrightarrow [M+(n-1)H]^{(n-1)+} + AH$$

- ETnoD

$$[M+nH]^{n+} + A^{\cdot -} \longrightarrow [M+nH]^{(n-1)+\cdot} + A$$

- Not good description of rules

  e.g. Concatenating reactions: $ETD \rightarrow PTR \rightarrow PTR \rightarrow ETnoD$

Correct Rules

- Additional Description of $C_cH_hO_oN_nS_s$- (p, q)
  - p - protonisation
    - n° of extra protons
    - charge state of a given molecule
    - adds weight to molecule
  - q - neutralised protonisation
    - n° of extra protons paired with electrons
    - only adds weight
- Problem Input $(A_1A_2 \ldots A_k, p, q)$

Reactions revised: post-doc in accountancy

- Problem Input $(A_1 A_2 \ldots A_k,\ p,\ q)$
- Some Partial Reactions
  - ♣ ETD

    $\rightarrow\ (A_1 \ldots A_L,\ p_1,\ q_1)$
    $\rightarrow\ (A_{L+1} \ldots A_k,\ p_2,\ q_2)$
    s.t. $p_1 + p_2 = p - 1$ and $q_1 + q_2 = q$ and $q_2 \geq 0$
  - ♢ PTR

    $\rightarrow\ (A_1 \ldots A_k,\ p - 1,\ q)$
  - ♡ ETnoD

    $\rightarrow\ (A_1 \ldots A_k,\ p - 1,\ q + 1)$
  - ♠ *HTR*

    $\rightarrow\ (A_1 \ldots A_L,\ p_1,\ q_1)$
    $\rightarrow\ (A_{L+1} \ldots A_k,\ p_2,\ q_2)$
    s.t. $p_1 + p_2 = p$ and $q_1 + q_2 = q + 1$ and $q_2 \geq 1$

Algorithm?

- Inputs:
  - $S = \left[M = A_1 \dots A_k,\ p = \text{Maximal Charge}, q = 0\right]$
  - Partial Reactions =

$$= \{\mathfrak{Id},\ \clubsuit^{C}_{L,p_1,q_1},\ \clubsuit^{Z}_{L,p_2,q_2},\ \diamondsuit,\ \heartsuit,\ \spadesuit^{C}_{L,\tilde{p}_1,\tilde{q}_1},\ \spadesuit^{Z}_{L,\tilde{p}_2,\tilde{q}_2}\}$$

  - Reactions =

$$= \{r_1 r_2 \dots r_k : r_i \text{ is a Partial Reactions and is OK}\}$$

e.g. $\clubsuit^{C}_{L,2,1}\heartsuit$, $\heartsuit\diamondsuit$ might be valid reactions if
  - S has enough charges
  - M long enough
  - there were q reactions before ET resulting in proton neutralisation

  - Empirical Spectrum, $y = \left\{\left(\frac{m_j}{z_j}, I_j\right)\right\}_{j=1}^{J}$

Deriving many probability functions

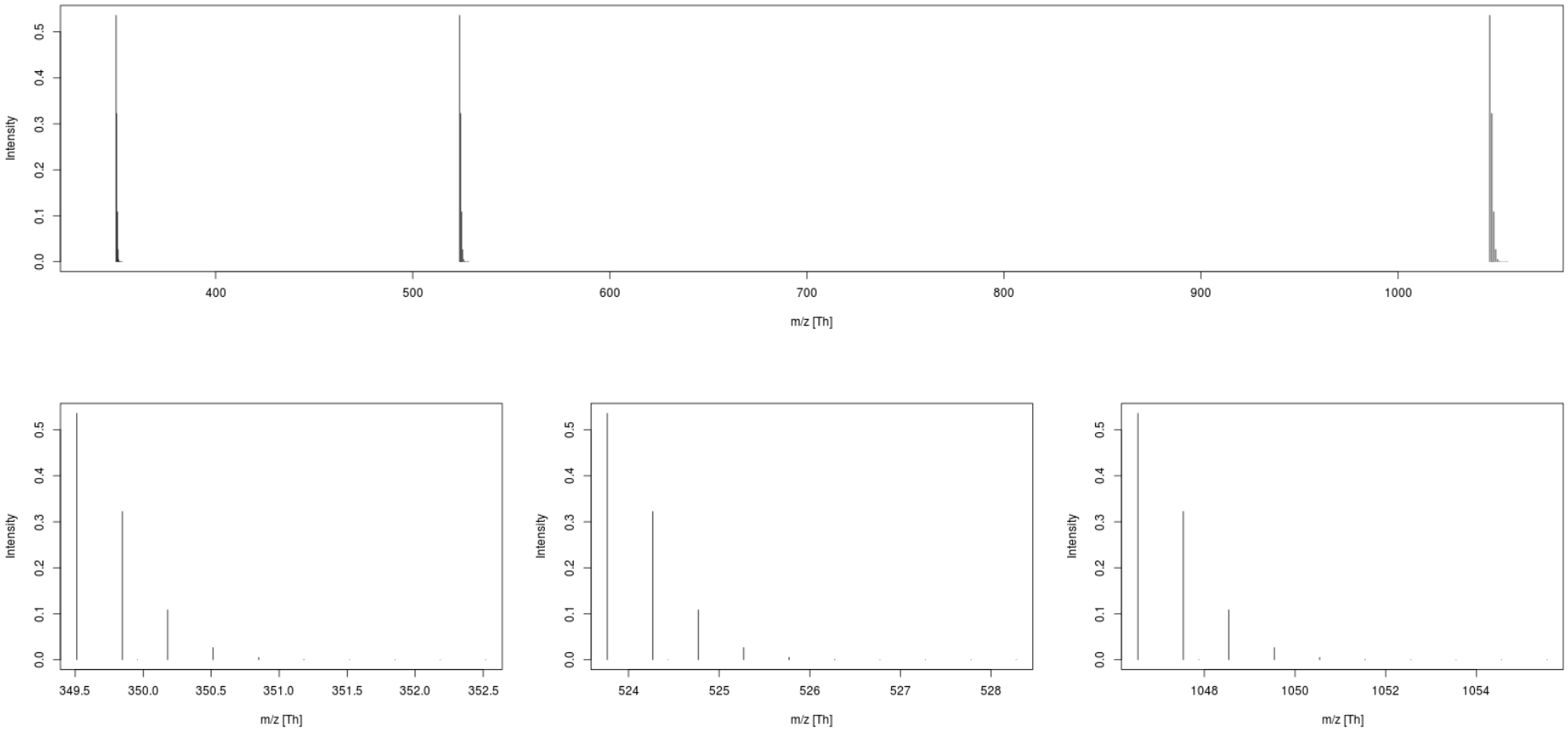
- Observations:
  - Each reaction is also a triplet $R = [F, p, q]$
  - $F \sim$ Multinomial Distribution
  - m_F - corresponding mass distribution

$$\mathcal{P}\left(\frac{m_R}{z_R}\right) = \mathcal{P}\left(\frac{m_F + p + q}{p}\right) = p_F(m_F)$$

  - Some reactions give the same results

$$\heartsuit\diamondsuit = \diamondsuit\heartsuit$$

    - The problem is static: we do not model time explicitly.
    - Discernible Reactions $\subset$ Reactions
    - Reactions := Equivalence Classes within Reactions


Intensity
0.0
0.1
0.2
0.3
0.4
0.5
400
500
600
700
800
900
1000
m/z [Th]
349.5
350.0
350.5
351.0
351.5
352.0
352.5
524
525
526
527
528
1048
1050
1052
1054

Precising Hypothesis

- Empirical Spectrum, $y = \left\{ \left(\frac{m_j}{z_j}, I_j \right) \right\}_{j=1}^{J}$

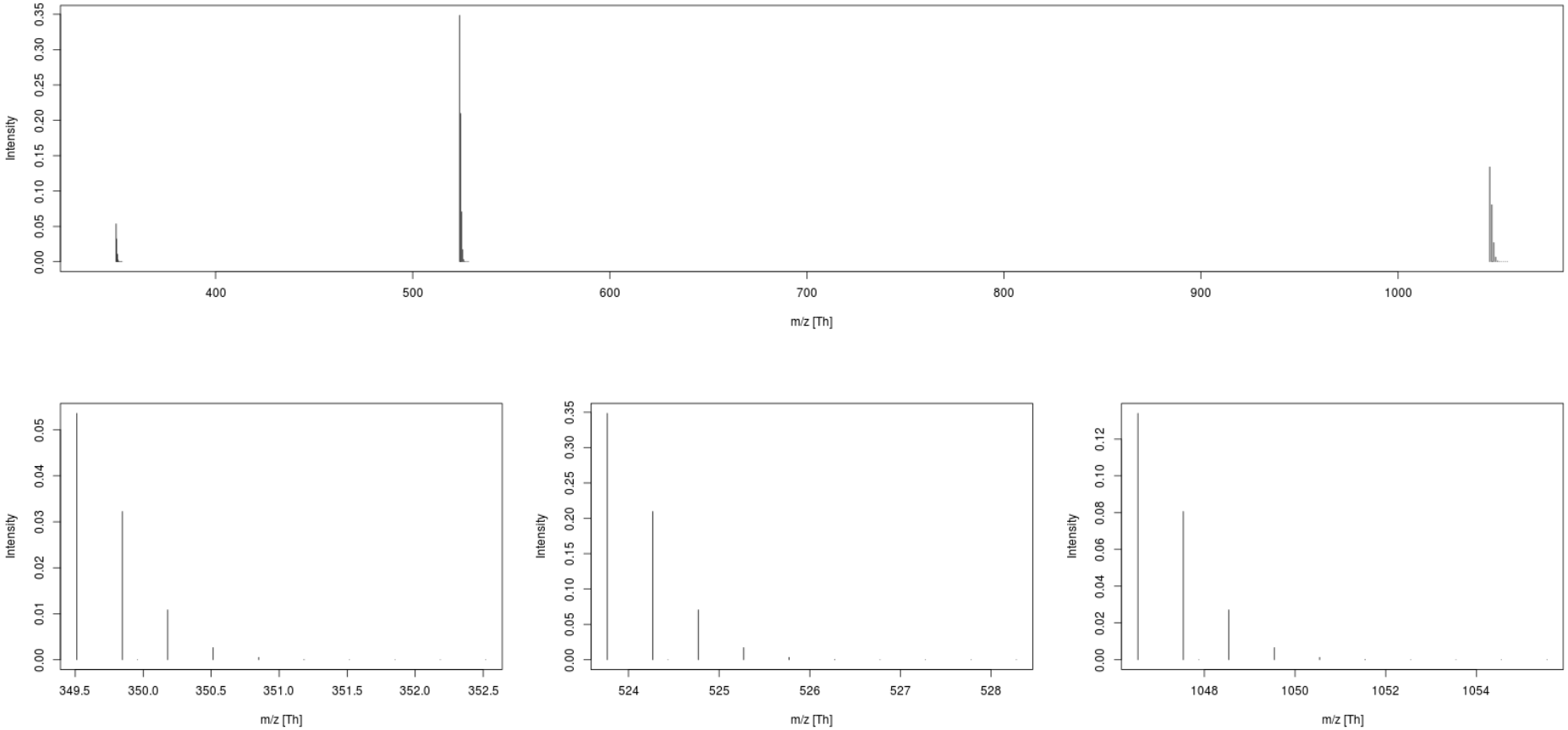
  Hypothesis

  $\star$ $I_j = \sum_{R \in \text{Reactions}} \alpha_R \mathcal{P}\left(\frac{m_R}{z_R}\right) + \text{Error}$

  s.t. $\alpha_R \geq 0,$

  $\sum_R \alpha_R \leq 1$

- Problem: need software that
  - finds Reactions
  - estimates α_R so that error is smallest possible


Intensity
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
400
500
600
700
800
900
1000
m/z [Th]
Intensity
0.00
0.01
0.02
0.03
0.04
0.05
349.5
350.0
350.5
351.0
351.5
352.0
352.5
m/z [Th]
Intensity
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
524
525
526
527
528
m/z [Th]
Intensity
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1048
1050
1052
1054
m/z [Th]

massTodon

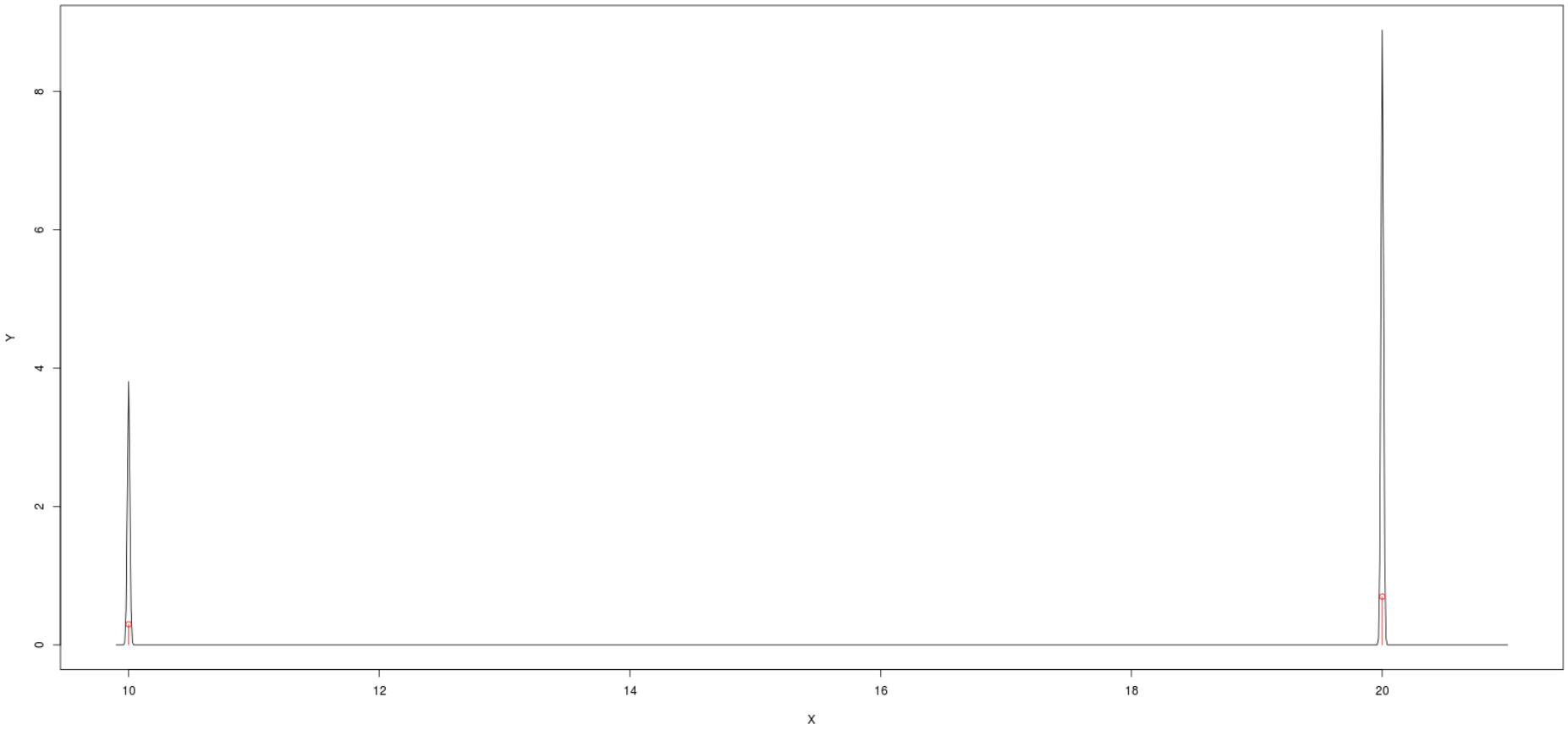
- The prototype of the software already exists
- Codename - massTodon
- Stupid fitting procedure: does not use BRAIN

Kernelisation

- $y = \left\{ \left(\frac{m_j}{z_j}, I_j \right) \right\}_{j=1}^{J}$
- Consider a function

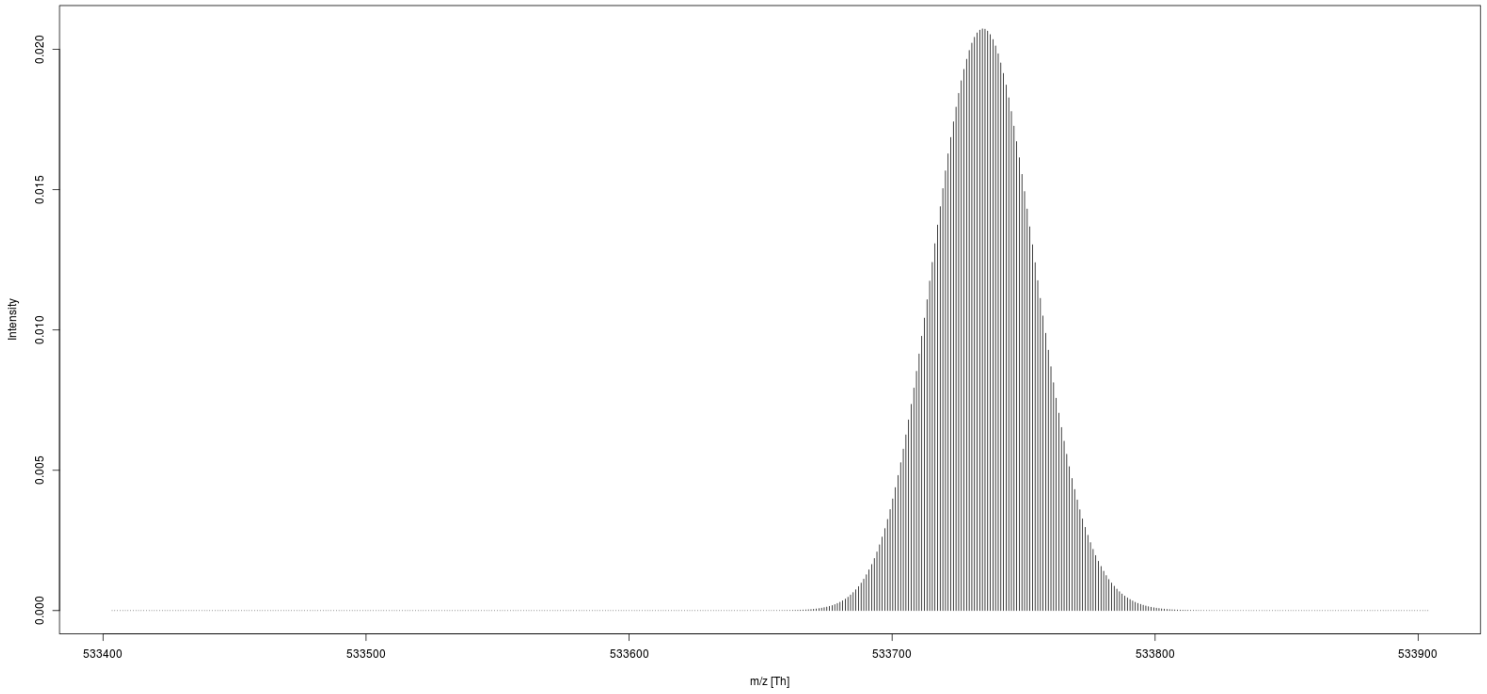
$$y(z) = \sum_{j=1}^{J} I_j \times y_j(z)$$

s.t. $y_j(z) = \frac{1}{\sqrt{\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{\sigma^2}}$
$\mu = \frac{m_j}{z_j}$


Y
0
2
4
6
8
10
12
14
16
18
20
X

Gaussian Approximation to Multinomial Model

Human Dynein: $C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$



Gaussian Approximation to Multinomial Model

- Theoretical Spectra: $\left\{ f_R(z) \right\}_{R \in \text{Reactions}}$
- Empirical Spectrum: $y(z)$
- Linear model:

$$y(z) = \sum_R \alpha_R f_R(z) + \epsilon(z)$$

- Usual least squares humbug

$$||\epsilon||^2 = \left|\left| y - \sum_R \alpha_R f_k \right|\right|^2 \to \text{min}$$

s.t. $\alpha_R \geq 0$
s.t. $\sum_R \alpha_R \leq 1$
where $||y||^2 = \langle y|y \rangle = \int_{\mathbb{R}} y(x)^2 \mathrm{d}\, x.$

Gaussian Approximation to Multinomial Model

$$\left|\left|y - \sum_R \alpha_R f_k\right|\right|^2 = ||y||^2 - 2\alpha^{\mathrm{t}}\mathfrak{m} + \alpha^{\mathrm{t}}\mathfrak{H}\alpha$$

where $\mathfrak{m}^{\mathrm{t}} = [\ldots, \langle y|f_R\rangle, \ldots]$

and $\mathfrak{H} = [\langle f_R|f_P\rangle]_{R,P\in\mathrm{Reactions}}$

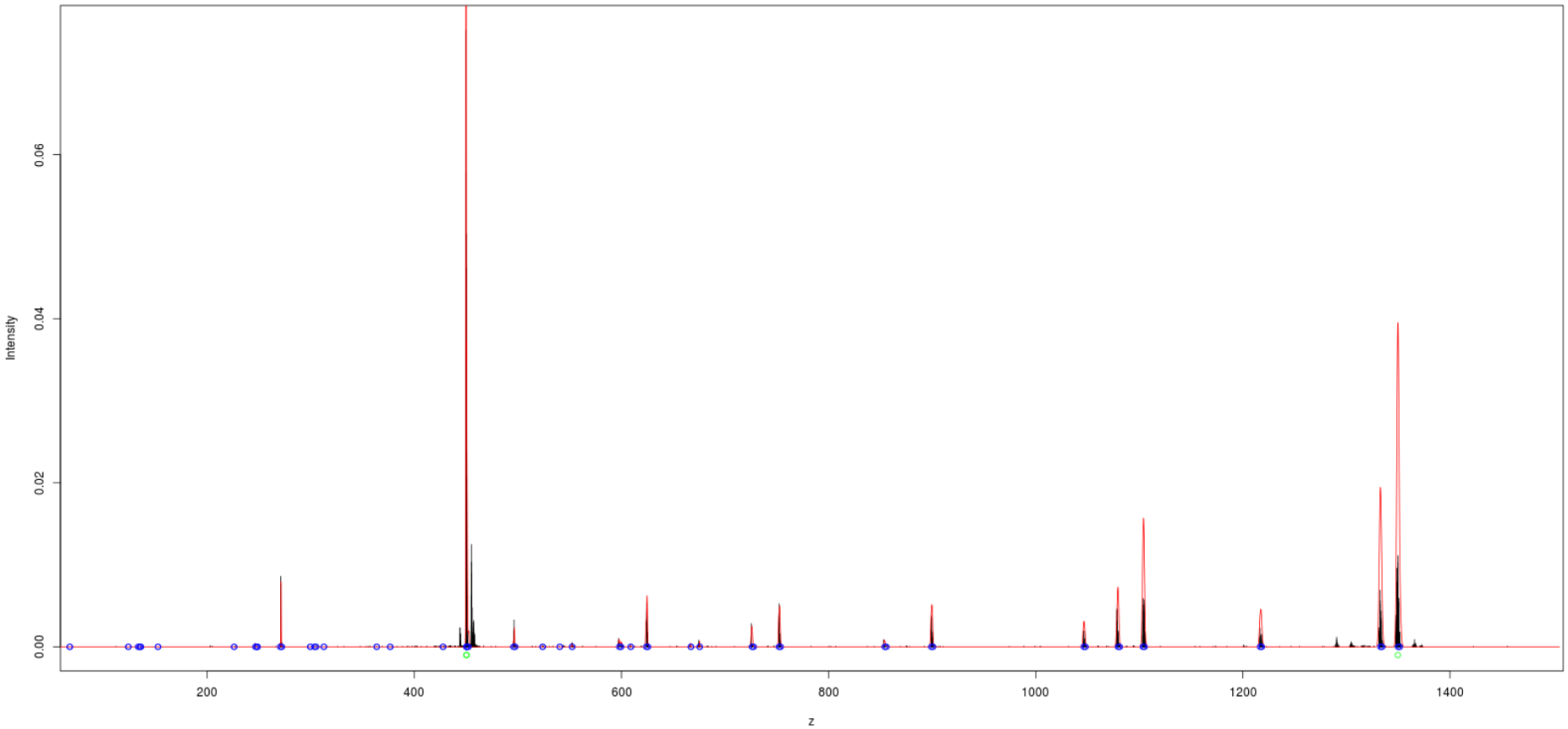
- Easy to evaluate scalar product

$$m(z) = \frac{1}{\sqrt{\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{\sigma^2}} \text{ and } n(z) = \frac{1}{\sqrt{\pi\eta^2}} e^{\frac{(z-\eta)^2}{\nu^2}}$$

$$\langle m|n\rangle = \int_{\mathbb{R}} m(z)n(z)\mathrm{d}\,z = \frac{1}{\sqrt{\pi(\sigma^2+\nu^2)}} e^{-\frac{(\mu-\eta)^2}{\sigma^2+\nu^2}}$$

- This problem is numerically very nice
  - The Gramm Matrix is diagonally dominant, well conditioned.

Theoretically Explained Spectrum



Whole lotta of things to do

- Algorithmically
  - Optimise generation of Reactions:
    - $\heartsuit\diamondsuit = \diamondsuit\heartsuit$
  - Calibrate fitting procedures:
    - stick spectra more jazzy among chemists
    - no-one uses gaussian approximations
  - Derive quick procedures for stick spectra generation: modify BRAIN.
- Experimentally
  - More substances to analyze
  - Mixtures of substances
- Pragmatically
  - What if the substance is not known?

massTodon potential

- Understanding ETD statics
- Next step : understand dynamics
- Quantitative approach : potential characterisation of peptides through the use of MASS SPEC.
    - But not certain yet how to do it yet

- Ingvar Eidhammer, Kristian Flikka, Lennart Martens, Svein-Ole Mikalsen, *Computational Methods for Mass Spectrometry Proteomics*. Wiley-Interscience, 2007.
- Igor Kaltashov, Stephen J. Eyles *Mass Spectrometry in Biophysics: Conformation and Dynamics of Biomolecules*. Wiley-Interscience, 2005.