

MassTodon

Oprogramowanie Służące Interpretacji Widm Spektroskopu Masowego z
Wykorzystaniem Techniki Dysocjacji Transferem Elektronów

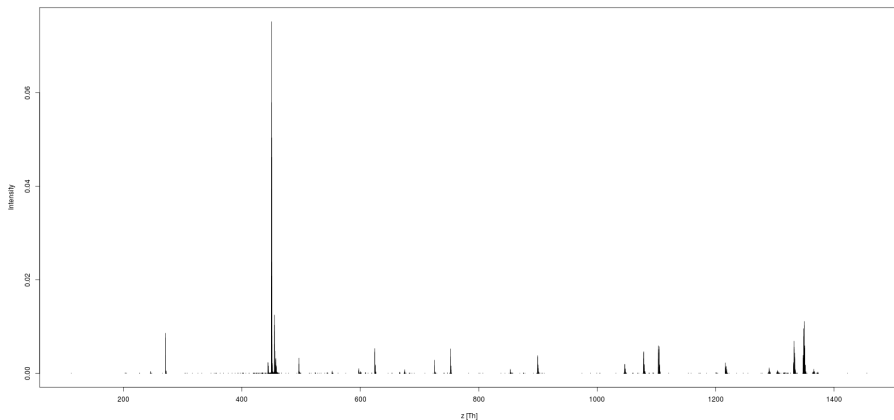
Mateusz Łącki

Uniwersytet Warszawski

9 Stycznia 2014



Substancja P okiem Spektrometru Masowego



Projekt MassTodon

- Spektrometr Masowy

- Bada skład molekularny próbek
- Wynik pomiaru:

$$\star \left\{ \frac{\text{Masa}_j}{\text{\u0142adunek}_j}, \text{Intensywno\u015b\u0107}_j \right\}_j^J$$

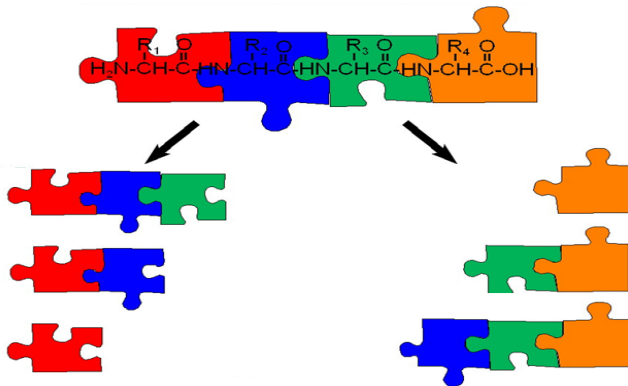
- Technologia MS/MS

- Uszeregowienie dwu spektroskop\u00f3w
- Filtracja moleku\u0142 o ustalonej masie do \u0142adunku
- Wykorzystanie ETD

- Motywacja

- Wyja\u015bnienie struktury spektrum posiekanego elektronami polimeru

Przydatna Analogia

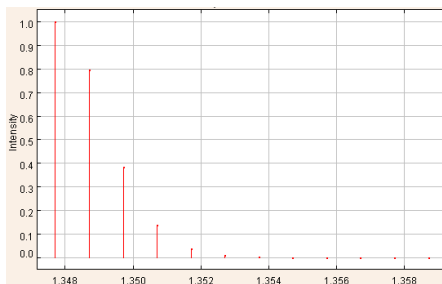


Dzisiejszy Program

- 1 Studium Przypadku
- 2 Statistical Modelling
- 3 Procedury Dopasowujące
- 4 What remains to be done?

Produkt Modelów Wielomianowych

- Czemu służy?
 - Rozkład Liczby Dodatkowych Neutronów
 - Rozkład Masy Związku Chemicznego
 - Odchył od obserwowanej masy monoizotopowej



Model Wielomianowy

- Rozkład Liczby Dodatkowych Neutronów w *przyrodzie*

- $\mathcal{P}(^{16}\text{O}) = 99.757$
- $\mathcal{P}(^{17}\text{O}) = 0.038$
- $\mathcal{P}(^{18}\text{O}) = 0.205$

źródło: International Union of Pure and Applied Chemistry 1997

- Molekuła = $\text{C}_c\text{H}_h\text{O}_o\text{N}_n\text{S}_s$

- Założenia

- Dużo niezależności: izotop danego atomu $\text{C}_c\text{H}_h\text{O}_o\text{N}_n\text{S}_s$ (np. C) nie zależy od izotopów pozostałych atomów
- np. molekula trój-atomowa wody

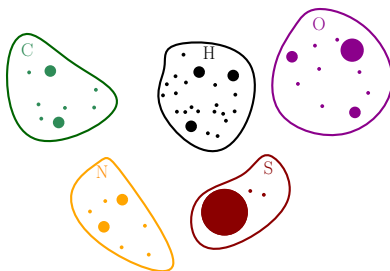
$$\mathcal{P}(^1\text{H}^2\text{H}^{17}\text{O}) = \mathcal{P}(^1\text{H}) \times \mathcal{P}(^2\text{H}) \times \mathcal{P}(^{17}\text{O})$$

- Nie rozróżniamy izomerów

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

Produkt Modelów Wielomianowych

- Związek Chemiczny = lista zbiorów atomów



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

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

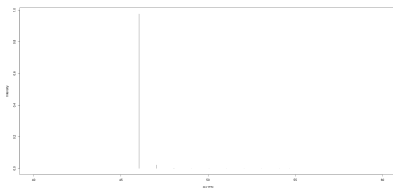
Produkt Modelów Wielomianowych a Masa Molekuł

- m_W masa molekuły $W = (c_{12}, c_{13}, \dots, s_{36})$
 - ★ $m_W = c_{12}m_{^{12}\text{C}} + \dots + s_{36}m_{^{36}\text{S}}$
 - ★ $\mathcal{P}(\text{Masa Molekuły} = m) = \sum_{W: m_W = m} \mathcal{P}(c_{12}^{^{12}\text{C}}, \dots, s_{36}^{^{36}\text{S}})$
 - s.t. $m_W = c_{12}m_{^{12}\text{C}} + \dots + s_{36}m_{^{36}\text{S}}$
 - :) Wzór $W \rightarrow$ miara $p_W \rightarrow \rightarrow$
 - :(Ilość pików rośnie.
- Fakt
 - Dodatkowe neutrony zmieniają masę pierwiastków różnorako
 - Różnice są możliwe do zaobserwowania w spektroskopie
- ★ Algorytm BRAIN zaniedbuje powyższą właściwość materii

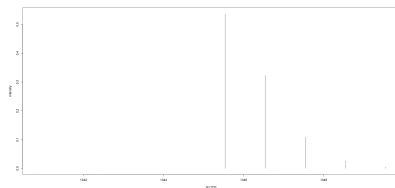
Wizualizacja Modelu Wielomianowego

- Oprogramowanie BRAIN - Piotr Dittwald[®]
- Założenie:
 - ★ Neutrony mają masę 1 Daltona dla wszystkich pierwiastków.

Etanol C_2H_5OH

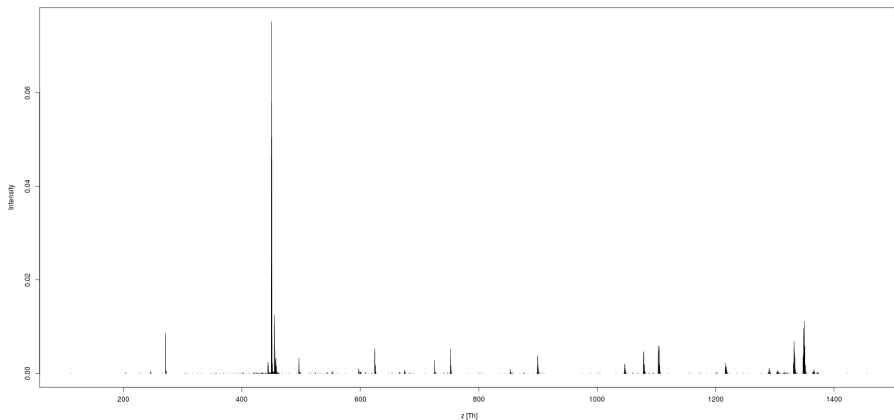


Angiotensyna II $C_{50}H_{71}N_{13}O_{12}$

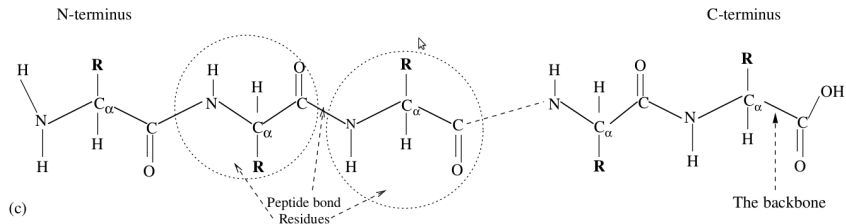


- Hola hola! Spektrum z początkowego slajdu było wielomodalne!
 - fragmentacja (ETD)
 - szafowanie ładunkami (ETD, ETnoD, PTR)

Substancja P okiem Spektrometru Masowego



Polimery jako sekwencje aminokwasów



- Dodatkowa struktura pomiędzy wzorem sumarycznym a poszczególnymi pierwiastkami

$$\text{C}_c\text{H}_h\text{O}_o\text{N}_n\text{S}_s = A_1A_2 \dots A_k$$

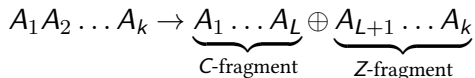
$$A_i \in \{ \text{Alanina, Cysteina, Kwas Asparaginowy, Kwas Glutaminowy, ...} \}$$

Dodatkowa struktura - sekwencje aminokwasów



Electron Transfer Dissociation

- Wynik: losowe cięcie białka na dwa podśłowa



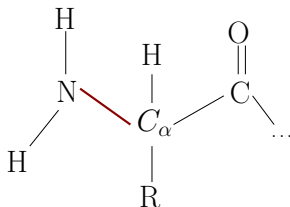
- L = numer przeciętego wiązania peptydowego (od N do C)

- Założenia

- L nie zależy od składu izotopowego

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

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



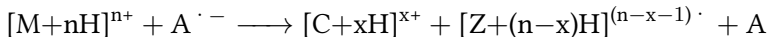
- Drobna komplikacja: dodatkowe wiązania peptydowe na A_1 - wprowadzamy stan $L = 0$

Możliwe reakcje okiem zaprzyjaźnionego chemika Friderika

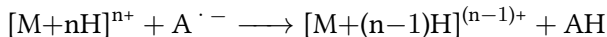
Hipoteza:

★ Spektrum Empiryczne = Wynik Kilku Reakcji

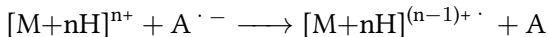
- ETD



- PTR



- ETnoD



- To nie jest poprawny zapis reguł

e.g. Błędy w konkatencjach reakcji: $ETD \rightarrow PTR \rightarrow PTR \rightarrow ETnoD$

Poprawne Zasady

- Dodatowy opis molekuly o wzorze $C_cH_hO_oN_nS_s$ - para (p, q)
 - p - *protonizacja*
 - liczba dodatkowych protonów bez sparowanych elektronów
 - to ładunek molekuly
 - to coś waży
 - q - *zneutralizowana protonizacja*
 - liczba dodatkowych protonów ze sparowanymi elektronami
 - to coś waży ale jest elektrycznie obojętne
- Wsad algorytmu: $(A_1A_2 \dots A_k, p, q)$

Możliwe Reakcje: o potrzebie rachunkowości w przyrodzie

- Wsad: $(A_1 A_2 \dots A_k, p, q)$

- Reakcje Częstkowe

- ♣ ETD

- $\rightarrow (A_1 \dots A_L, p_1, q_1)$

- $\rightarrow (A_{L+1} \dots A_k, p_2, q_2)$

- że $p_1 + p_2 = p - 1$ oraz $q_1 + q_2 = q$ i $q_2 \geq 0$

- ◇ PTR

- $\rightarrow (A_1 \dots A_k, p - 1, q)$

- ♥ ETnoD

- $\rightarrow (A_1 \dots A_k, p - 1, q + 1)$

- ♠ HTR

- $\rightarrow (A_1 \dots A_L, p_1, q_1),$

- $\rightarrow (A_{L+1} \dots A_k, p_2, q_2),$

- że $p_1 + p_2 = p$ oraz $q_1 + q_2 = q + 1$ i $q_2 \geq 1$

Algorytm - generowanie wzorów chemicznych

- $S = \left[M = A_1 \dots A_k, p = \text{Maksymalne Naładowanie}, q = 0 \right]$

- Reakcje Częstkowe =

$$= \{ \heartsuit, \clubsuit_{L,p_1,q_1}^C, \clubsuit_{L,p_2,q_2}^Z, \diamond, \heartsuit, \spadesuit_{L,\tilde{p}_1,\tilde{q}_1}^C, \spadesuit_{L,\tilde{p}_2,\tilde{q}_2}^Z \}$$

-

Reakcje = $\{ r_1 r_2 \dots r_k : r_i \text{ to reakcja cząstkowa i jest OK} \}$

np. $\left(\clubsuit_{L,2,1}^C \heartsuit \heartsuit \diamond \right)$ może być OK jeśli

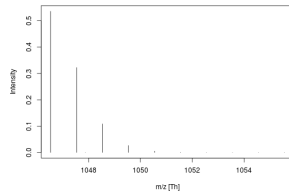
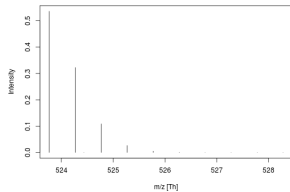
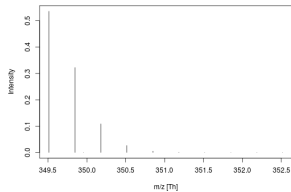
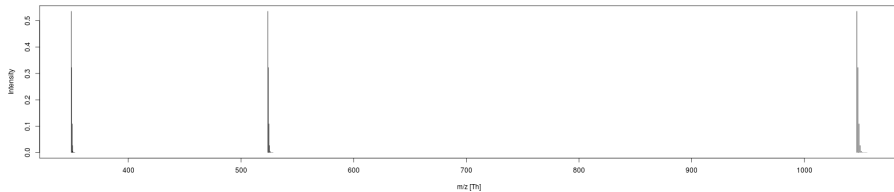
- S ma wystarczająco ładunków
- M jest wystarczająco długa
- q protonów zostało *zneutralizowanych* przed ET

- Wynik pełny: lista trójek $\left\{ [W, p_W, q_W] \right\}_{W \in \text{Reakcje}}$

Rozkłady indukowane przez reakcje

- Rozkład na osi $\frac{m}{z}$:
 - Algorytm dostarcza trójkę $R = [W, p, q]$
 - $W \rightarrow$ Model Wielomianowy \rightarrow Rozkład masy $m_W \rightarrow$

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



Konsekwencje zaniedbania czasu

- Niektóre reakcje dają te same wyniki!

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

- Model abstrahuje od czasu przeprowadzania reakcji
- Rozróżnialne Reakcje \subset Reakcje
- Reakcje $:=$ Klasy Równoważności w zbiorze Reakcji

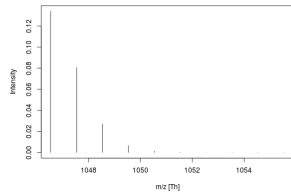
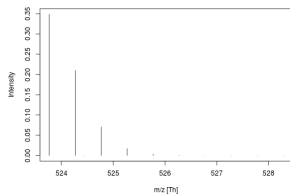
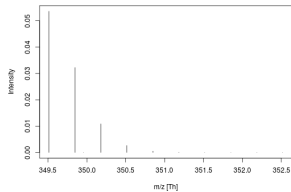
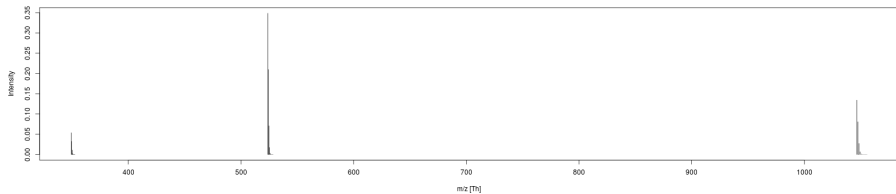
Doprecyzowanie Hipotezy Badawczej

- Spektrum Empiryczne: $y = \left\{ \left(\frac{m_j}{z_j}, l_j \right) \right\}_{j=1}^J$

Hipoteza

$$\begin{aligned} &\star l_j = \sum_{R \in \text{Reakcje}} \alpha_R \mathcal{P}\left(\frac{m_R}{z_R}\right) + \text{Błąd} \\ &\text{że } \alpha_R \geq 0, \\ &\text{i } \sum_R \alpha_R \leq 1 \end{aligned}$$

- Potrzeba oprogramowania które
 - znajduje zbiór nierozróżnialnych reakcji R
 - estymuje α_R minimalizując błąd



massTodon

- Prototyp istnieje

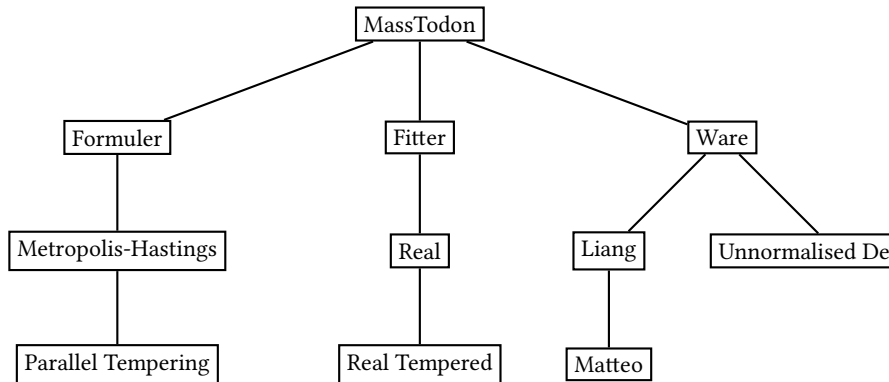


Figure : Current operational entity-relations diagram.

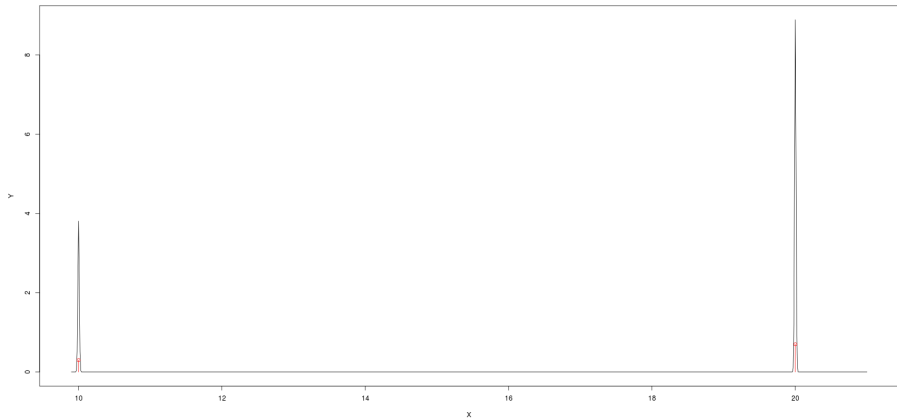
Kernelizacja

- $y = \left\{ \left(\frac{m_j}{z_j}, l_j \right) \right\}_{j=1}^J$
- Ale może lepiej wygładzić wszystko?

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

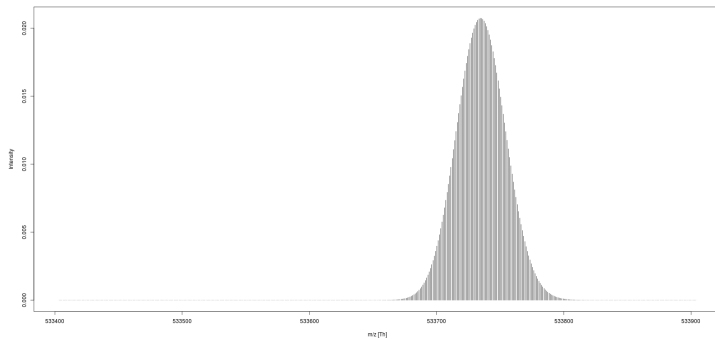
$$\text{s.t. } y_j(z) = \frac{1}{\sqrt{\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{\sigma^2}}$$

$$\mu = \frac{m_j}{z_j}$$



Gaussian Approximation to Multinomial Model

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



Gaussian Approximation to Multinomial Model

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

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

- Usual least squares humbug

$$\|\epsilon\|^2 = \left\| y - \sum_R \alpha_R f_k \right\|^2 \rightarrow \min$$

$$\text{s.t. } \alpha_R \geq 0$$

$$\text{s.t. } \sum_R \alpha_R \leq 1$$

$$\text{where } \|y\|^2 = \langle y|y \rangle = \int_{\mathbb{R}} y(x)^2 dx.$$

Gaussian Approximation to Multinomial Model

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

where $\mathbf{m}^t = [\dots, \langle y | f_R \rangle, \dots]$

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

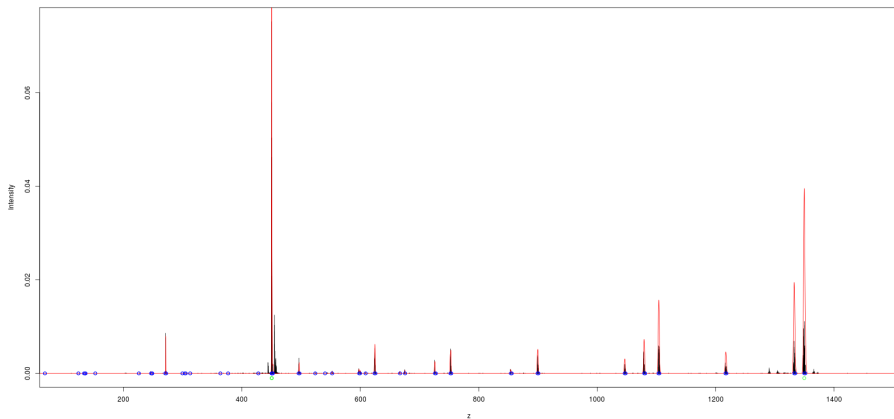
- Easy to evaluate scalar product

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

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

- 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.