A Model for Enzymatically ^{18}O-Labeled Mass Spectra

Tomasz Burzykowski, Qi Zhu, Dirk Valkenborg

Interuniversity Institute for Biostatistics and statistical Bioinformatics

Hasselt University, Belgium

Outline

Mass spectrometry

- Allows to separate (peptide) molecules by their atomic mass
- Allows to quantify the abundance of the molecules in a sample

Mass spectrometry

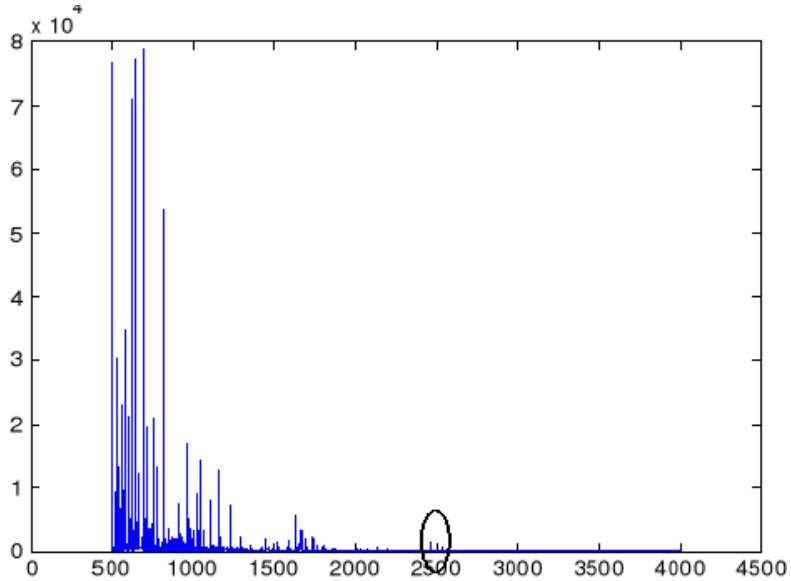
- Allows to separate (peptide) molecules by their atomic mass
- Allows to quantify the abundance of the molecules in a sample

MASS	INTENSITY
499.866376	26.1438
499.879673	24.9673
499.892970	20.6536
499.906267	21.5686
499.919565	24.5752
499.932863	27.4510
499.946160	28.7582
499.959458	27.8431
499.972757	25.8824
499.986055	30.5882
499.999353	37.3856
500.012652	32.2876
500.025951	27.4510
500.039250	32.4183
500.052549	38.8235
. . .	. . .

Mass spectrometry

- Allows to separate (peptide) molecules by their atomic mass
- Allows to quantify the abundance of the molecules in a sample

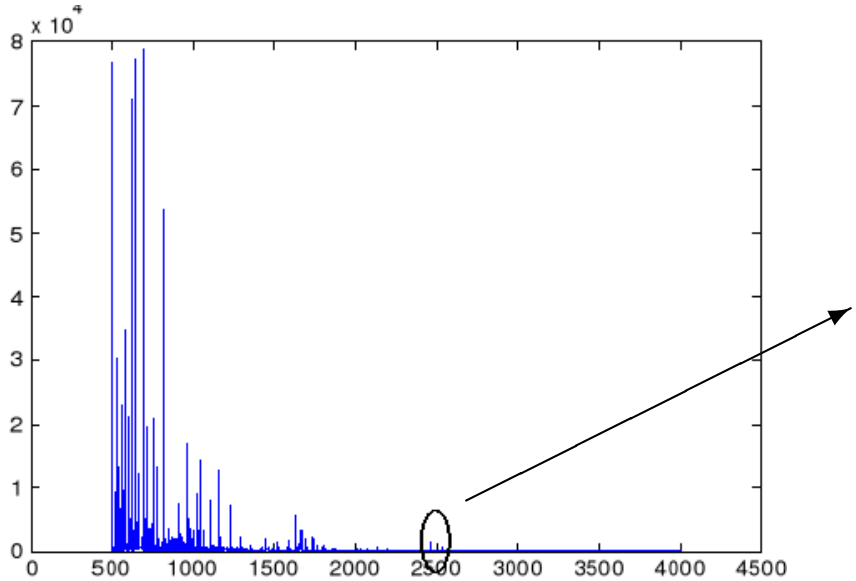
MASS	INTENSITY
499.866376	26.1438
499.879673	24.9673
499.892970	20.6536
499.906267	21.5686
499.919565	24.5752
499.932863	27.4510
499.946160	28.7582
499.959458	27.8431
499.972757	25.8824
499.986055	30.5882
499.999353	37.3856
500.012652	32.2876
500.025951	27.4510
500.039250	32.4183
500.052549	38.8235
. . .	. . .



Mass spectrometry

- Allows to separate (peptide) molecules by their atomic mass
- Allows to quantify the abundance of the molecules in a sample

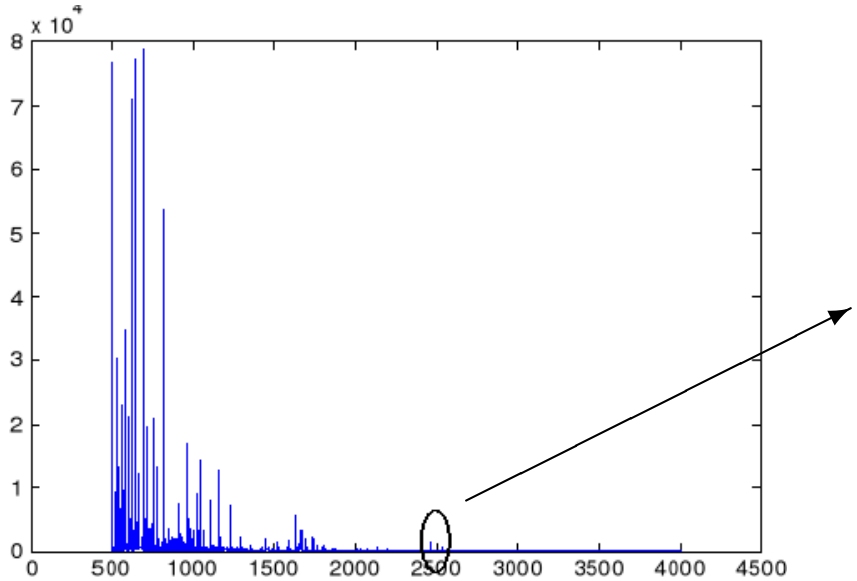
MASS	INTENSITY
499.866376	26.1438
499.879673	24.9673
499.892970	20.6536
499.906267	21.5686
499.919565	24.5752
499.932863	27.4510
499.946160	28.7582
499.959458	27.8431
499.972757	25.8824
499.986055	30.5882
499.999353	37.3856
500.012652	32.2876
500.025951	27.4510
500.039250	32.4183
500.052549	38.8235
. . .	. . .

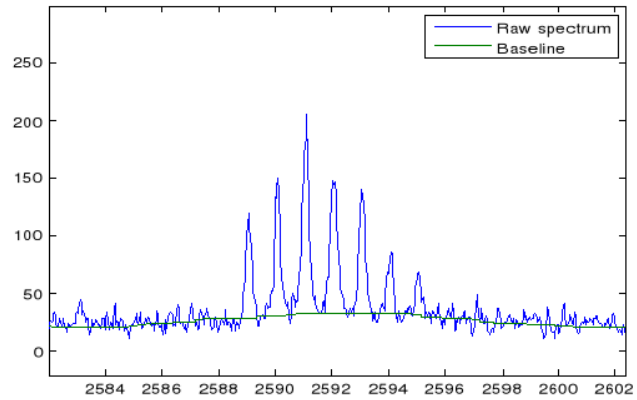


Mass spectrometry

- Allows to separate (peptide) molecules by their atomic mass
- Allows to quantify the abundance of the molecules in a sample

MASS	INTENSITY
499.866376	26.1438
499.879673	24.9673
499.892970	20.6536
499.906267	21.5686
499.919565	24.5752
499.932863	27.4510
499.946160	28.7582
499.959458	27.8431
499.972757	25.8824
499.986055	30.5882
499.999353	37.3856
500.012652	32.2876
500.025951	27.4510
500.039250	32.4183
500.052549	38.8235
. . .	. . .





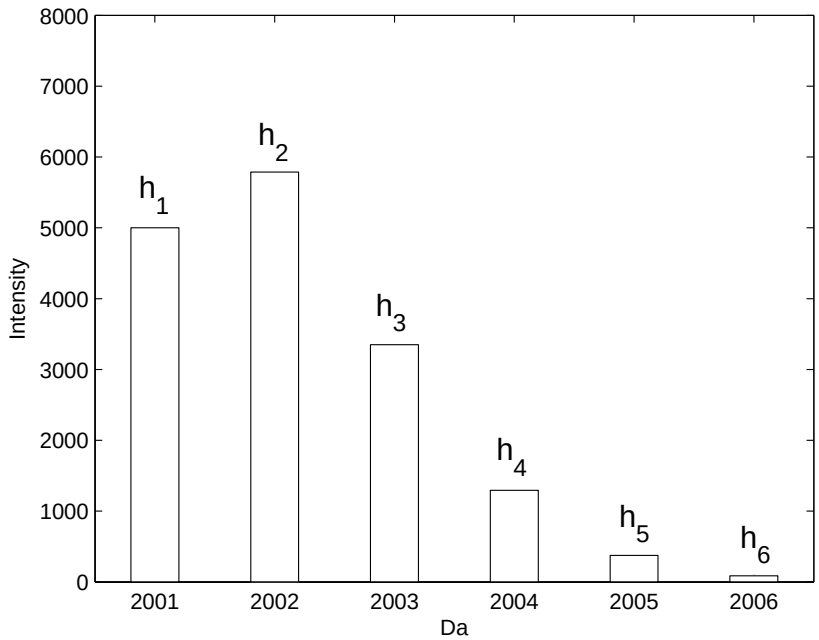
Isotopic distribution

- Molecules of a peptide contain different isotopes of chemical elements:
  $(^{1}H,^{2}H)$ $(^{16}O,^{17}O,^{18}O)$ $(^{12}C,^{13}C)$ $(^{14}N,^{15}N)$ $(^{32}S,^{33}S,^{34}S,^{36}S)$
- Lead to isotopic variants of a molecule, with masses differing by ≈ 1 Da

- Distribution represented by *isotopic ratios*:

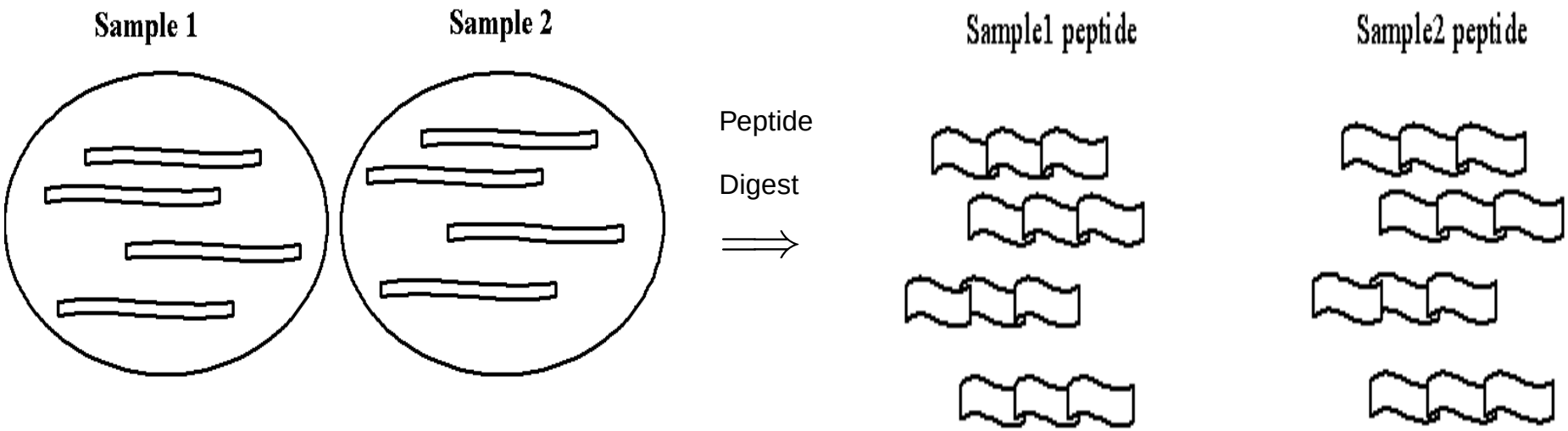
  $R_1 = h_1/h_1,$
  $R_2 = h_2/h_1,$
  $R_3 = h_3/h_1,$

  $\ldots$



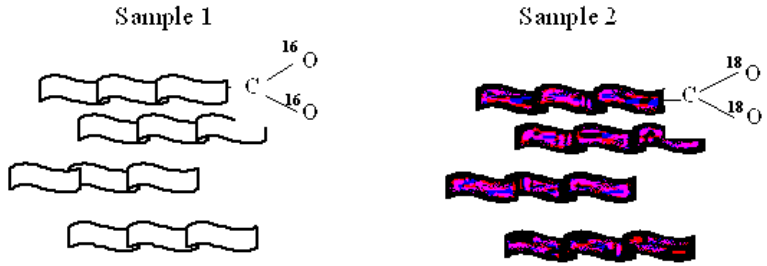
^{18}O-labeling strategy

- Reduces the beween-spectra variability by comparing samples in the same spectrum
- Idea similar to double-channel cDNA microarrays



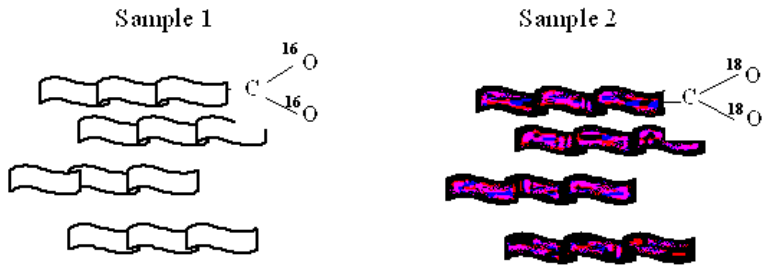
^{18}O-labeling strategy

- Samples get labeled ($^{16}O \longrightarrow {}^{18}O$)

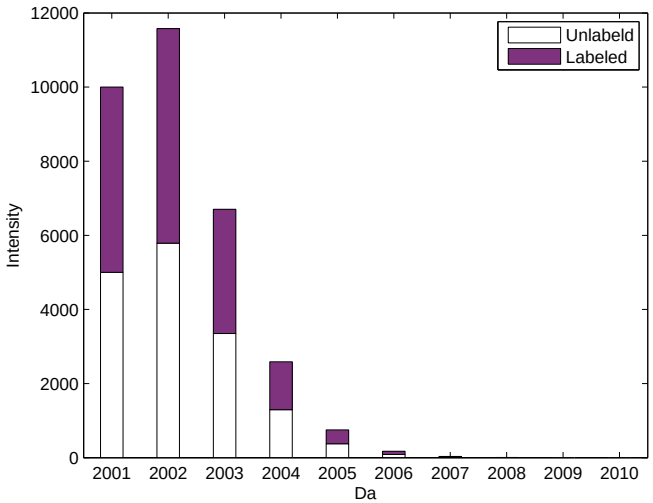


^{18}O-labeling strategy

- Samples get labeled ($^{16}O \longrightarrow {}^{18}O$)

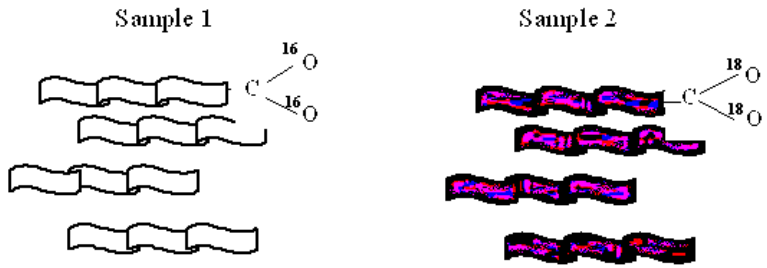


- Labeled samples are processed simultaneously

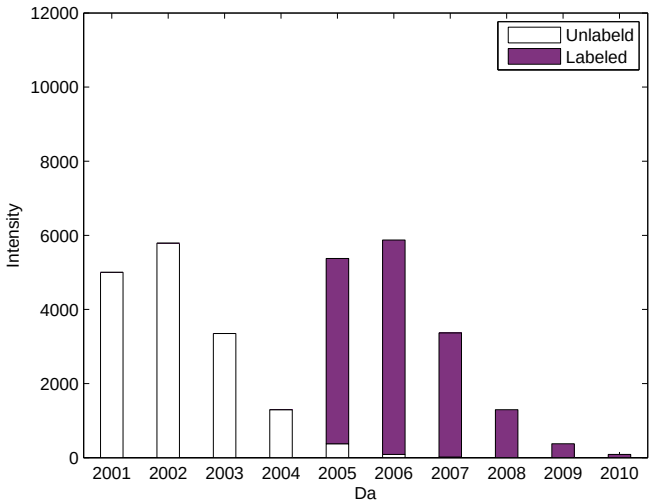


^{18}O-labeling strategy

- Samples get labeled ($^{16}O \longrightarrow {}^{18}O$)



- Labeled samples are processed simultaneously



- Ideally, the labeled sample shifted by 4 Da ($2 \times {}^{18}O$)
- Parameter of interest: the relative abundance Q of the peptides

Incomplete labeling

Not all oxygens in the carboxyl terminus are replaced by ^{18}O, due to

- Water impurities – presence of ^{16}O & ^{17}O in the heavy oxygen water

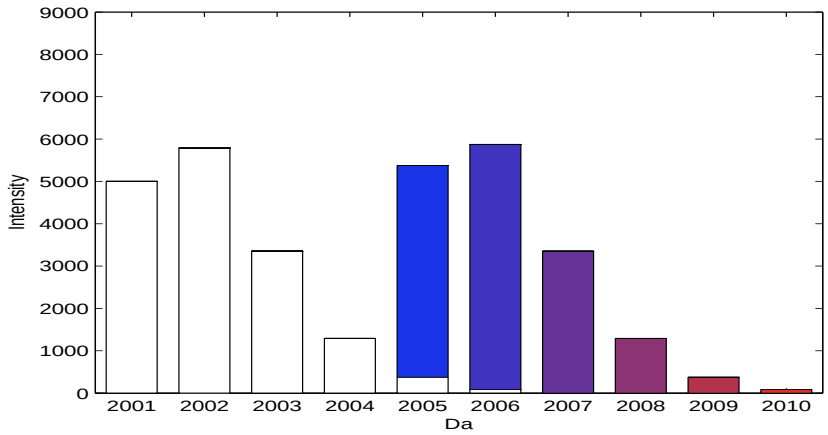
Factors influencing the completeness of labeling

- Speed of atom exchange, quantified by *incorporation rate* λ
- Time τ available for the exchange
- Prevalence of water impurities (π_{16}, π_{17})

Consequences of incomplete labeling

- The isotopic distribution of the labeled peptide shifts by 0, . . ., 4 Da
- It overlaps with the distribution of the unlabeled peptide
- Shift probabilites

$P_0 = P(^{16}O, ^{16}O)$
$P_1 = 2P(^{16}O, ^{17}O)$
$P_2 = 2P(^{16}O, ^{18}O) + P(^{17}O, ^{17}O)$
$P_3 = 2P(^{17}O, ^{18}O)$
$P_4 = P(^{18}O, ^{18}O)$

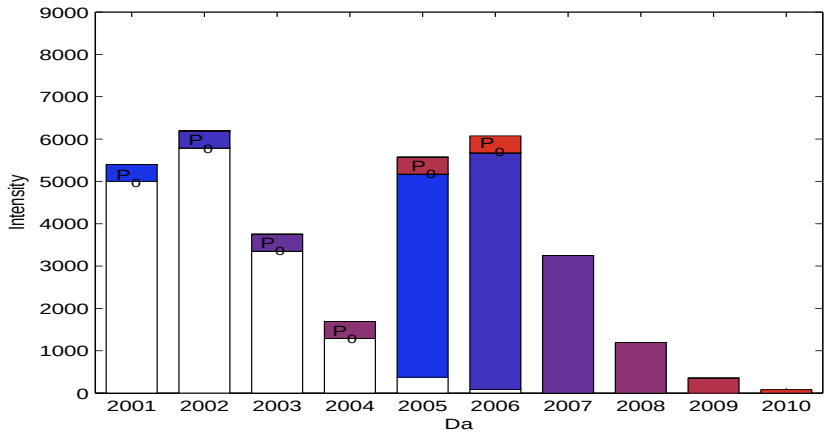


Ignoring the incomplete labeling leads to a biased estimate of Q

Consequences of incomplete labeling

- The isotopic distribution of the labeled peptide shifts by 0, . . ., 4 Da
- It overlaps with the distribution of the unlabeled peptide
- Shift probabilites

$P_0 = P(^{16}O, ^{16}O)$
$P_1 = 2P(^{16}O, ^{17}O)$
$P_2 = 2P(^{16}O, ^{18}O) + P(^{17}O, ^{17}O)$
$P_3 = 2P(^{17}O, ^{18}O)$
$P_4 = P(^{18}O, ^{18}O)$

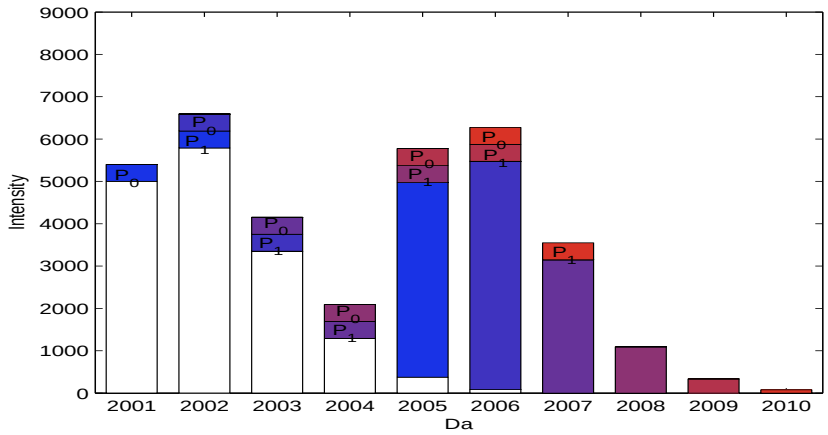


Ignoring the incomplete labeling leads to a biased estimate of Q

Consequences of incomplete labeling

- The isotopic distribution of the labeled peptide shifts by 0, . . ., 4 Da
- It overlaps with the distribution of the unlabeled peptide
- Shift probabilites

$P_0 = P(^{16}O, ^{16}O)$
$P_1 = 2P(^{16}O, ^{17}O)$
$P_2 = 2P(^{16}O, ^{18}O) + P(^{17}O, ^{17}O)$
$P_3 = 2P(^{17}O, ^{18}O)$
$P_4 = P(^{18}O, ^{18}O)$

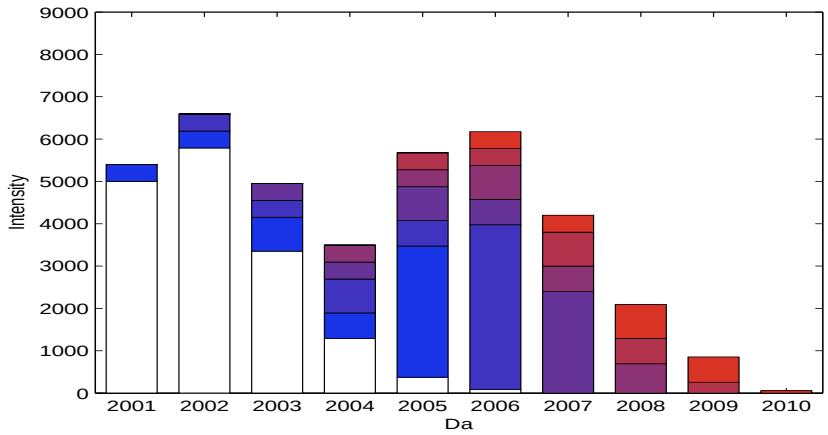


Ignoring the incomplete labeling leads to a biased estimate of Q

Consequences of incomplete labeling

- The isotopic distribution of the labeled peptide shifts by 0, . . ., 4 Da
- It overlaps with the distribution of the unlabeled peptide
- Shift probabilites

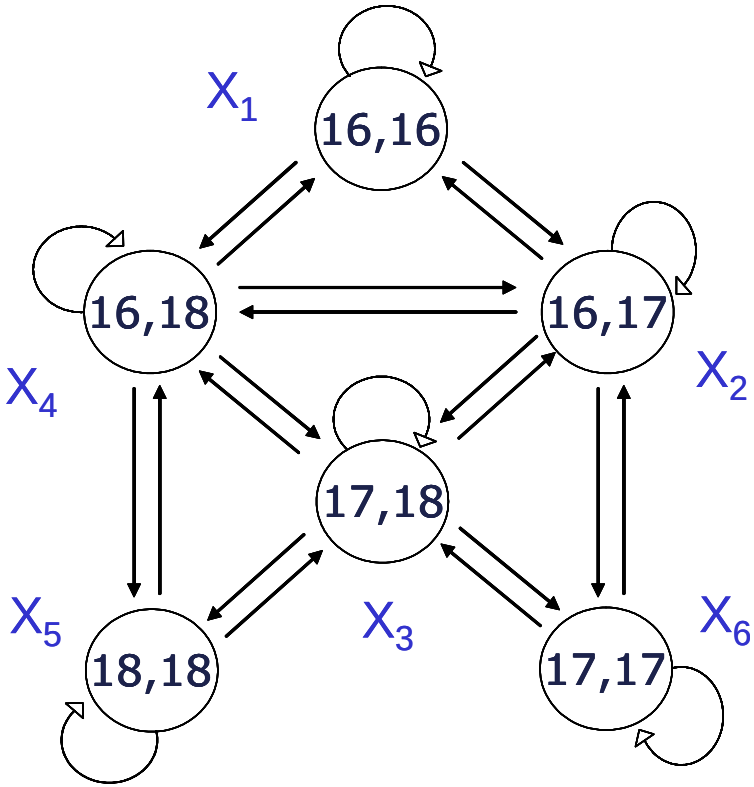
$P_0 = P(^{16}O, ^{16}O)$
$P_1 = 2P(^{16}O, ^{17}O)$
$P_2 = 2P(^{16}O, ^{18}O) + P(^{17}O, ^{17}O)$
$P_3 = 2P(^{17}O, ^{18}O)$
$P_4 = P(^{18}O, ^{18}O)$



Ignoring the incomplete labeling leads to a biased estimate of Q

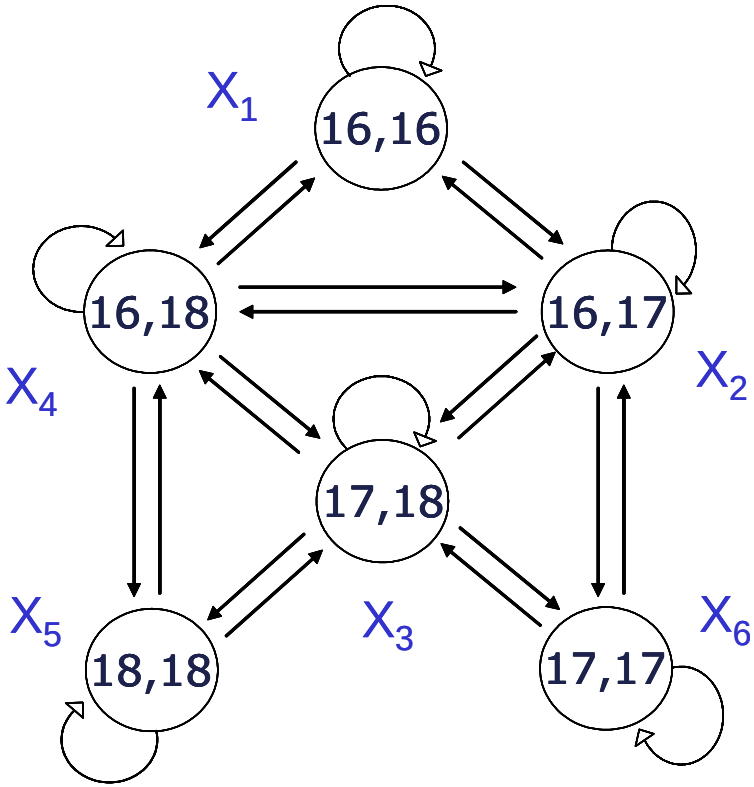
A solution to the incomplete labeling issue

Valkenborg and Burzykowski (2009): a discrete-time Markov-chain model to model shift probabilities P_0-P_4



A solution to the incomplete labeling issue

Valkenborg and Burzykowski (2009): a discrete-time Markov-chain model to model shift probabilities P_0-P_4



- Transition matrix $\boldsymbol{T}(\pi_{16}, \pi_{17}, \pi_{18})$
- State probability vector

$$\boldsymbol{S}(\tau) = (P(X_1;\tau), P(X_2;\tau), \ldots, P(X_6;\tau))$$

- Initial state $\boldsymbol{S}(0) \equiv \boldsymbol{S}_0 = (1, 0, \ldots, 0)$
- $\boldsymbol{S}(\tau) = \boldsymbol{S}_0 e^{-\lambda\tau} e^{\boldsymbol{T}\lambda\tau}$
- Resulting shift probabilities:

$$P_0(\tau) = P(X_1;\tau),\ P_1(\tau) = P(X_2;\tau)$$
$$P_2(\tau) = P(X_4;\tau) + P(X_6;\tau)$$
$$P_3(\tau) = P(X_3;\tau),\ P_4(\tau) = P(X_5;\tau)$$

Model for the peak intensity

Observed intensity for the jth peak in the ith spectrum (peptide with l isotopic variants):

$$y_{ij} = \mu_{ij} + \varepsilon_{ij},$$

$$\mu_{ij} \equiv \mathrm{E}(y_{ij}) = \begin{cases} H_i R_j + Q H_i \sum_{k=0}^{min(4,j-1)} P_k R_{j-k} & \text{if } 1 \leq j \leq l \\ Q H_i \sum_{k=j-1}^{4} P_k R_{j-k} & \text{if } l+1 \leq j \leq l+4 \end{cases}$$

- Q: relative abundance of the peptide
- H_i: intensity scale for the ith spectrum
- R_j: isotopic ratio
- P_k: shift probability (depends on λ, π_{16}, π_{17}, π_{18}, τ)
- Homoscedasticity: $\varepsilon_{ij} \sim N(0, \sigma^2)$

Proposed extensions

- Heteroscedasticity: $\text{Var}(\varepsilon_{ij}) = \sigma^2 \mu_{ij}^{2\theta}$
- Mixed-effects modeling:

  to capture the technical variability: $H_i \sim N(\mu_H, \sigma_H^2)$

  to capture the biological variability: $Q \sim N(\mu_Q, \sigma_Q^2)$
- Combination (heteroscedastic mixed-effects model)
- Bayesian approach

Bayesian model implementation

- Straightforward implementation of random effects
- Parametrization and prior distributions:
  - non-informative priors for all the parameters
  - Q, H_i, R_j constrained to be positive – logarithmic transformation
  - λ constrained to $(0,20/\tau)$ – Box-Cox transformation
- Practical implementation:
  - WinBUGS through WBDiff: an interface for differentiation equations
  - JAGS: internal matrix exponential function *mexp* in the *msm* module

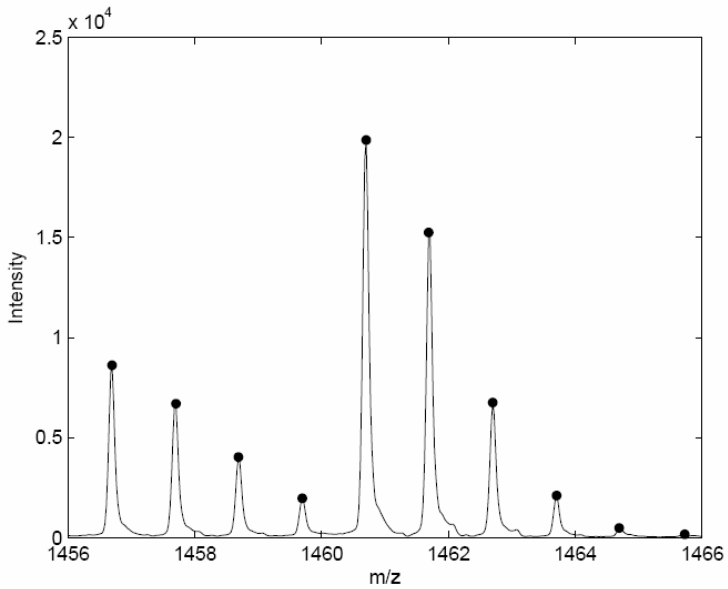
The data

- Mixture of peptides of bovine Cytochrome C (17 protein fragments)
- One part labeled with a stable ^{18}O-isotope, the other unlabeled
- Three unlabeled units mixed with one labeled; the relative abundance of $1/3$...
- ... and *vice versa* – the relative abundance of $3/1$
- Six spectra for each of the relative abundances

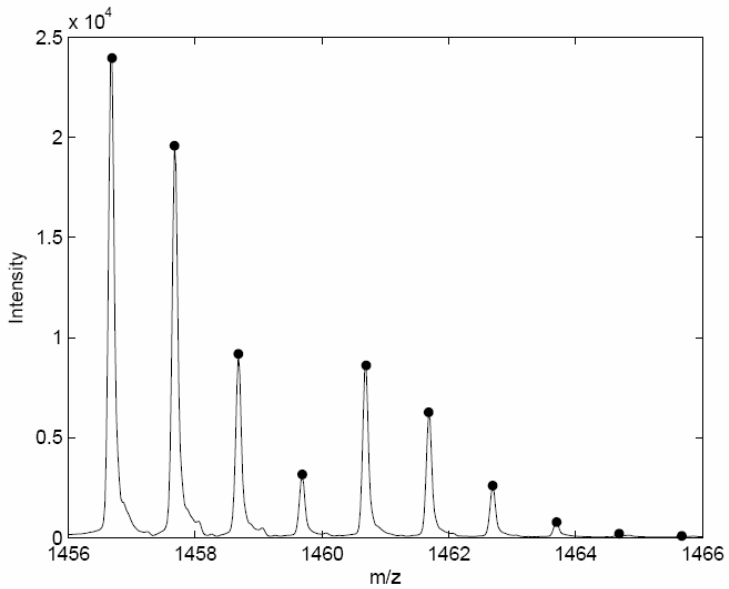
The data

- Peptide with the monoisotopic mass = 1456.66 Da

$Q = 1/3$



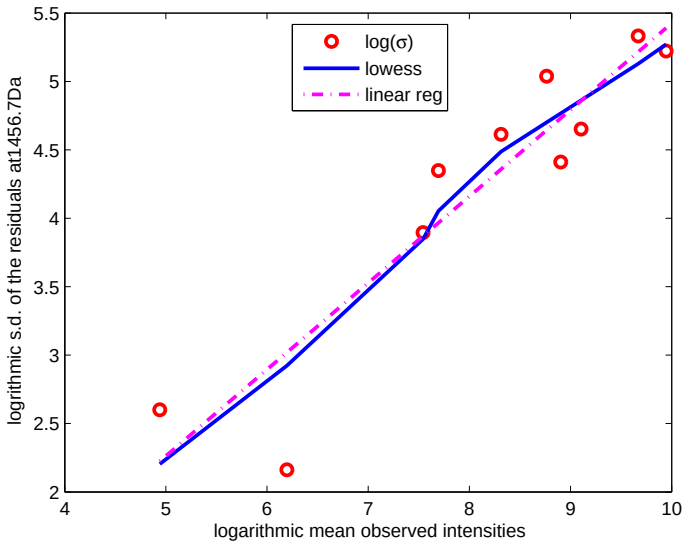
$Q = 3/1$



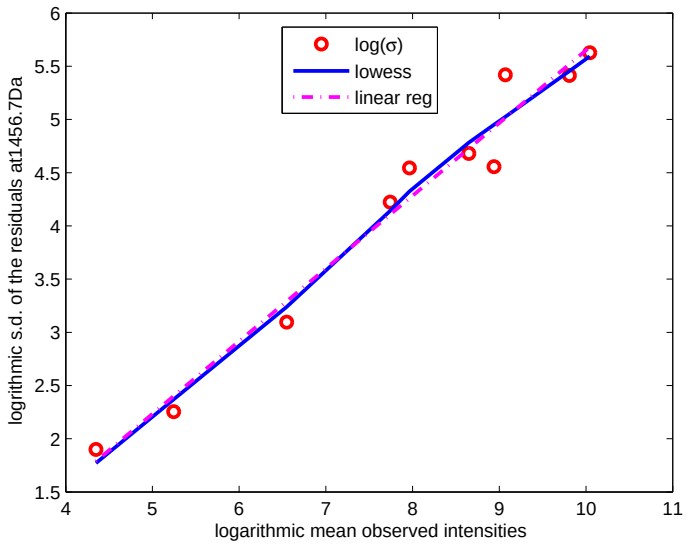
Mean-dependent variance

- Peptide with the monoisotopic mass = 1456.66 Da

$Q = 1/3$



$Q = 3/1$



- $0.5 \log\{\mathrm{Var}(\varepsilon_{ij})\} \approx \theta_0 + \theta \log(\mu_{ij}) \rightarrow \mathrm{Var}(\varepsilon_{ij}) \approx \sigma^2 \mu_{ij}^{2\theta}$

Fixed-effects, heteroscedastic model

- Peptide with the monoisotopic mass = 1456.66 Da

Parameter	Ratio 1/3			Ratio 3/1		
	True	Median	95% c.i.	True	Median	95% c.i.
H_1	–	24852.0	(24360.1, 25457.3)	–	8720.3	(8568.3, 8870.5)
H_2	–	22871.3	(22334.2, 23476.7)	–	8819.8	(8670.9, 8973.7)
H_3	–	22667.0	(22179.8, 23214.7)	–	7846.5	(7710.6, 7993.0)
H_4	–	24828.4	(24266.6, 25408.4)	–	10232.0	(10064.8, 10407.4)
H_5	–	19716.8	(19247.8, 20178.4)	–	9812.8	(9645.8, 9989.7)
H_6	–	25179.1	(24635.0, 25806.4)	–	8742.4	(8596.3, 8909.4)
Q	0.33	0.33	(0.32, 0.34)	(2.4)	2.4	(2.3, 2.4)
$\lambda\tau$	–	8.8	(5.5, 11.2)	–	11.5	(10.3, 13.2)
σ	–	0.21	(0.10, 0.51)	–	0.43	(0.17, 1.23)
θ	–	0.77	(0.66, 0.86)	–	0.66	(0.53, 0.77)
R_1	0.7933	0.7770	(0.7611, 0.7926)	0.7933	0.7763	(0.7651, 0.7868)
R_2	0.3567	0.3416	(0.3335, 0.3490)	0.3567	0.3430	(0.3368, 0.3493)
R_3	0.1166	0.1030	(0.1002, 0.1059)	0.1166	0.1015	(0.0987, 0.1045)
R_4	0.0306	0.0275	(0.0261, 0.0292)	0.0306	0.0245	(0.0235, 0.0256)
R_5	0.0068	0.01115	(0.0104, 0.0119)	0.0068	0.0070	(0.0066, 0.0076)

Homoscedastic mixed-effects model (random H & Q)

- Peptide with the monoisotopic mass = 1456.66 Da

Parameter	Ratio 1/3			Ratio 3/1		
	True	Median	95% c.i.	True	Median	95% c.i.
μ_H	–	22978.2	(20713.2, 24835.0)	–	8962.2	(8225.2, 9617.3)
σ^2_H	–	4538005.5	(1515783.0, 21798794.0)	–	884115.12	(302543.88, 3915923.5)
μ_Q	0.33	0.33	(0.27, 0.44)	(2.4)	2.4	(2.3, 2.4)
σ^2_Q	–	0.0052	(0.0017, 0.0254)	–	0.0056	(0.0019, 0.0245)
$\lambda\tau$	–	7.4	(6.8, 8.0)	–	10.4	(9.4, 11.8)
σ	–	158.6	(131.0, 199.1)	–	128.7	(106.0, 160.3)
R_1	0.7933	0.7905	(0.7831, 0.7977)	0.7933	0.7761	(0.7695, 0.7826)
R_2	0.3567	0.3494	(0.3436, 0.3555)	0.3567	0.3396	(0.3341, 0.3448)
R_3	0.1166	0.1062	(0.1004, 0.1114)	0.1166	0.1004	(0.0950, 0.1052)
R_4	0.0306	0.0392	(0.0313, 0.0535)	0.0306	0.0246	(0.0198, 0.0295)
R_5	0.0068	0.0090	(0.0040, 0.0188)	0.0068	0.0062	(0.0040, 0.0114)

Concluding Remarks

- Frequentist estimation works well (results not shown)
- Bayesian approach is feasible
- Nest step: heteroscedasticity AND random effects

Concluding Remarks

- Frequentist estimation works well (results not shown)
- Bayesian approach is feasible
- Nest step: heteroscedasticity AND random effects

Concluding Remarks

- Frequentist estimation works well (results not shown)
- Bayesian approach is feasible
- Nest step: heteroscedasticity AND random effects