59th ASMS Conference on Mass Spectrometry and Allied Topics

Denver (USA), June 5 – 9, 2011

Bayesian hierarchical reconstruction of protein profiles including a digestion model

Pierre GRANGEAT[1], Pascal SZACHERSKI[12], Laurent GERFAULT[1], Jean-François GIOVANNELLI[2]

[1] CEA, LETI, MINATEC Campus, DTBS,
17 rue des Martyrs, F-38054 Grenoble cedex 9, France.
[2] Université de Bordeaux 1 – CNRS - IPB, IMS,
351 Cours de la Libération, F-33405, Talence cedex, France.

E-mail: pierre.grangeat@cea.fr, pascal.szacherski@cea.fr, laurent.gerfault@cea.fr, Giova@IMS-Bordeaux.fr

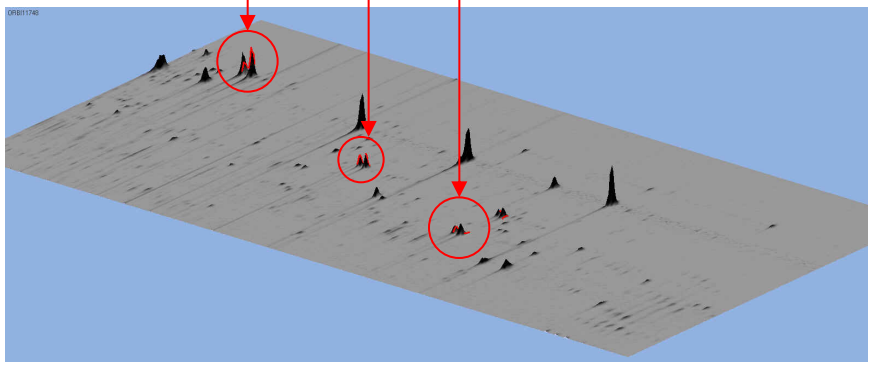
Ref: DRT/LETI/DTBS/STD/LE2S 11-156

Outline

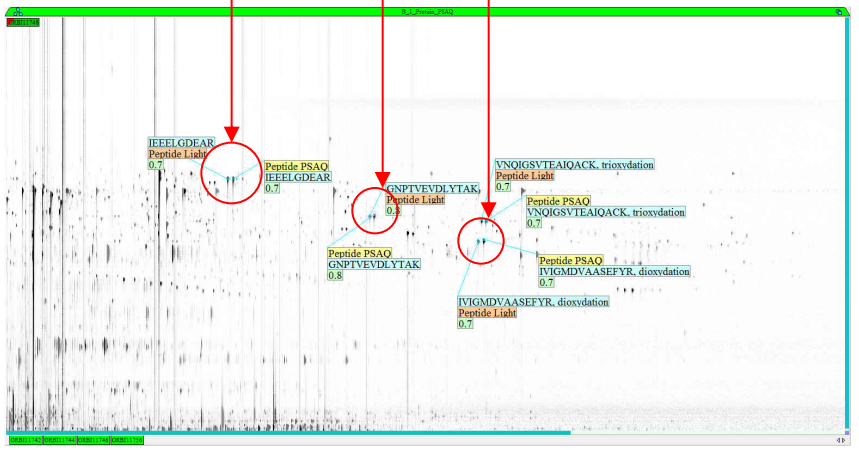
1. Introduction
2. Hierarchical model of the LC-MS analytical chain
3. Digestion model
4. Bayesian hierarchical reconstruction
5. Results on simulated data
6. Conclusion

The challenge of molecular profile reconstruction

Targeted peaks



Targeted peaks



3D view (zoom) 2D view

LC-MS chromato-spectrogram for the analysis of the NSE protein in serum:
3 peptides in light and heavy forms (PSAQ) [14]

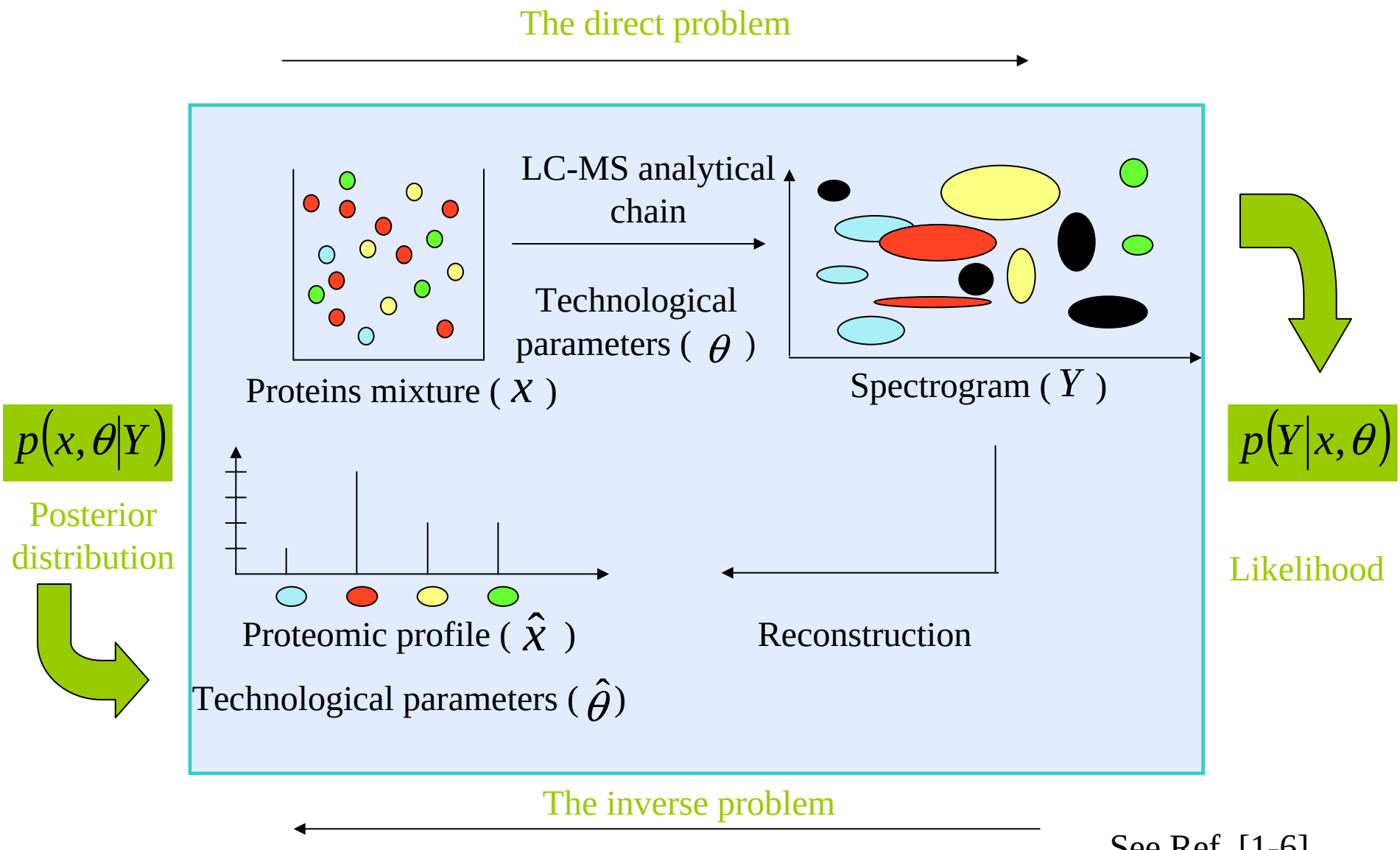
Visualisation using MSight software developed by SIB

- **The molecular profile reconstruction challenge:**

How to compute the quantity of targeted proteins from a molecular signature embedded in a complex measurement?

- **Technological variability: a major issue for quantification**

The Bayesian inverse problem approach



See Ref. [1-6]

Outline

1. Introduction
2. **Hierarchical model of the LC-MS analytical chain**
3. Digestion model
4. Bayesian hierarchical reconstruction
5. Results on simulated data
6. Conclusion

Graph structure of the LC-MS analytical chain

Protein content X_p $p = 1,\ldots, P$

Peptide content κ_i $i = 1,.., I$

Elementary Ion LC-MS spectrum Y_i $i = 1,.., I$

Total LC-MS spectrum $Y = \sum_{i=1}^{I} Y_i$

The hierarchical mixture model

Protein content x_p $p = 1,\ldots,P$

Digestion factor: d_{ip}

$$\kappa_i = \sum_{p=1}^{P} d_{ip} x_p$$

Peptide content κ_i $i = 1,..,I$

Ionisation gain: ξ_i

LC response function: $C_i(t_i)$

MS response function: S_i

$$Y_i = \xi_i \kappa_i S_i C_i^T$$

Elementary Ion LC-MS spectrum Y_i $i = 1,..,I$

Measurement noise: $N(\gamma_n)$

Total LC-MS spectrum:

$$Y = \sum_{i=1}^{I} \xi_i \kappa_i S_i C_i^T(t_i) + N(\gamma_n)$$

Graph structure of the LC-MS analytical chain including the PSAQ standard for calibration

	Native protein	Isotope labeled protein [7-8]	
Protein content	X_p	X_p^*	$p = 1,\ldots, P$
	↓	↓	
Peptide content	κ_i	κ_i^*	$i = 1,.., I$
	↓	↓	
Elementary Ion LC-MS spectrum	Y_i	Y_i^*	$i = 1,.., I$

Total LC-MS spectrum

$$Y = \sum_{i=1}^{I} \left(Y_i + Y_i^* \right) + N(\gamma_n)$$

Outline

1. Introduction
2. Hierarchical model of the LC-MS analytical chain
3. **Digestion model**
4. Bayesian hierarchical reconstruction
5. Results on simulated data
6. Conclusion

The digestion process

- Cause of variability on digestion:
  - temperature , pH, process time length [9], trypsin parameters.
- Statistical peptide content distribution:
  - Gaussian distribution associated with a Gaussian digestion noise $N(\gamma_\kappa)$ of inverse variance γ_κ
- Digestion variability model:
  - Bernoulli random process associated with a cleavage probability α_{ip} controlled by the digestion kinetic law.
- Digestion parameters:
  - digestion factor d_{ip} linking a protein p content to a peptide i content for a complete digestion
  - Bernoulli parameter $\alpha_{i_c p}^{cleaved}$ for the cleaved peptide and parameter $\alpha_{i_m p}^{miscleaved} = 1 - \alpha_{i_c p}^{cleaved}$ for the miscleaved molecule.
- In this presentation, we suppose each coefficient d_{ip}, α_{ip} , γ_κ to be known

Graph structure of the LC-MS analytical chain including a digestion model

Protein content $X_p \quad X_p^* \qquad p = 1,\ldots,P$

Digestion factor: d_{ip}

Digestion yield: α_{ip}

Conservation principle: $\alpha_{i_c p}^{cleaved} + \alpha_{i_m p}^{miscleaved} = 1$

Digestion noise: $N(\gamma_\kappa)$

$$\begin{cases} H_i(x) = \sum_{p=1}^{P} \alpha_{ip} d_{ip} x_p \\ \kappa_i = H_i(x) + N(\gamma_\kappa) \\ \kappa_i^* = H_i(x^*) + N(\gamma_\kappa) \end{cases}$$

Peptide content $\kappa_i \quad \kappa_i^* \qquad i = \underbrace{1,\ldots,I}_{\text{Cleaved form}}, \underbrace{I+1,\ldots,I'}_{\text{Miscleaved form}}$

Elementary Ion LC-MS spectrum $Y_i \quad Y_i^* \qquad i = 1,\ldots,I, I+1,\ldots,I'$

Total LC-MS spectrum $$Y = \sum_{i=1}^{I'} \left(Y_i + Y_i^*\right) + N(\gamma_n)$$

The full hierarchical mixture model

Protein content

Digestion factor: d_{ip}

Digestion yield: α_{ip}

Conservation principle: $\alpha_{i_c p}^{cleaved} + \alpha_{i_m p}^{miscleaved} = 1$

Digestion noise: $N(\gamma_\kappa)$

$X_p \quad X_p^* \qquad p = 1,\ldots,P$

$$\begin{cases} H_i(x) = \sum_{p=1}^{P} \alpha_{ip} d_{ip} x_p \\ \kappa_i = H_i(x) + N(\gamma_\kappa) \\ \kappa_i^* = H_i(x^*) + N(\gamma_\kappa) \end{cases}$$

Peptide content

Ionisation gain: ξ_i

LC response function: $C_i(t_i)$

MS response function: S_i

$\kappa_i \quad \kappa_i^* \qquad i = 1,..,I'$

$$\begin{cases} Y_i = \xi_i \kappa_i S_i \, C_i^T(t_i) \\ Y_i^* = \xi_i \kappa_{i\,i}^* S_i^* C_i^T(t_i) \end{cases}$$

Elementary Ion LC-MS spectrum

$Y_i \quad Y_i^* \qquad i = 1,..,I'$

Measurement noise: $N(\gamma_n)$

Total LC-MS spectrum:

$$\begin{cases} G(\kappa,\xi,t) = \sum_{i=1}^{I'} \xi_i \left(\kappa_i S_i + \kappa_i^* S_i^*\right) C_i^T(t_i) \\ Y = G(\kappa,\xi,t) + N(\gamma_n) \end{cases}$$

Outline

1. Introduction
2. Hierarchical model of the LC-MS analytical chain
3. Digestion model
4. **Bayesian hierarchical reconstruction**
5. Results on simulated data
6. Conclusion

Bayesian approach for profile reconstruction

Statistical framework [10-13]:

	Associated distribution	Notation
Direct model + noise model	*likelihood*	$p\left(Y \middle\| x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa\right)$
Modeling the *a priori* information on the parameters	*prior*	$p\left(x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa\right)$
Combined model	*posterior*	$p\left(x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa \middle\| Y\right)$

Bayes rule:

$$p\left(x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa \middle| Y\right) = \frac{p\left(Y \middle| x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa\right) p\left(x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa\right)}{\int p\left(\mathbf{Y} \middle| x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa\right) p\left(x, \boldsymbol{\xi}, \mathbf{t}, \gamma_n, \kappa\right) \mathrm{d}x \, \mathrm{d}\boldsymbol{\xi} \, \mathrm{d}\mathbf{t} \, \mathrm{d}\gamma_n \, \mathrm{d}\kappa}$$

The a posteriori distribution

$$p\left(x,\boldsymbol{\xi},\mathbf{t},\gamma_n,\boldsymbol{\kappa}\middle|\mathbf{Y}\right)\propto \exp\left(-\frac{1}{2}\gamma_n\left\|Y-G\left(\boldsymbol{\kappa},\boldsymbol{\xi},\mathbf{t}\right)\right\|^2\right)$$

Noise: normal distribution

Peptide level prior

$$\times\prod_{i=1}^{I}\exp\left(-\frac{1}{2}\gamma_\xi^i\left(\xi_i-\bar{\xi}_i\right)^2\right)$$

System gain: normal distribution

$$\times\frac{\gamma_n^{\alpha_n-1}}{\beta_n^{\alpha_n}\Gamma\left(\alpha_n\right)}\exp\left(-\frac{\gamma_n}{\beta_n}\right)$$

Noise inverse variance: gamma distribution

$$\times\prod_{i=1}^{I}U\left(t_i;t_i^m,t_i^M\right)$$

Retention time: uniform distribution

$$\times\prod_{i=1}^{I}\exp\left(-\frac{1}{2}\gamma_\kappa\left(\kappa_i-H_i\left(x\right)\right)^2\right)$$

Peptide content: normal distribution

Protein level prior

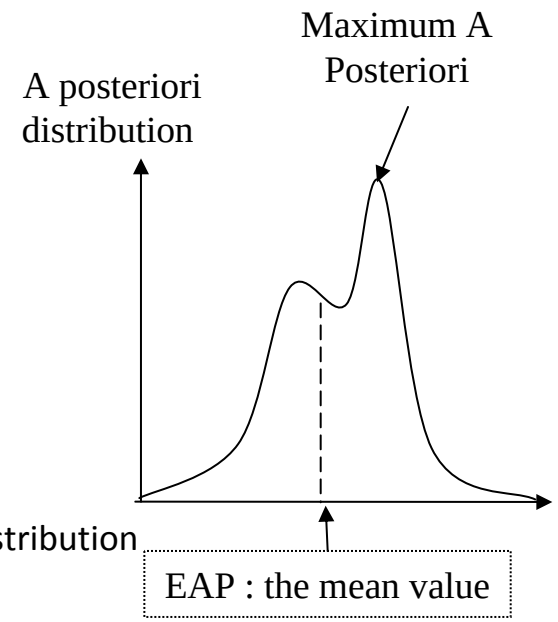
$$\times\prod_{p=1}^{P}\exp\left(-\frac{1}{2}\gamma_x^p\left(x_p-\bar{x}_p\right)^2\right)$$

Protein content: normal distribution

Bayesian reconstruction based on the EAP

- **Bayesian reconstruction [10-13]:**
  - Expectation of the A Posteriori distribution (EAP)

$$\mathrm{EAP}\left([x,\xi,\mathbf{t},\gamma_n,\kappa]\right)=\int [x,\xi,\mathbf{t},\gamma_n,\kappa]\, p\left(x,\xi,\mathbf{t},\gamma_n,\kappa|Y\right)\mathrm{d}x\,\mathrm{d}\xi\,\mathrm{d}\mathbf{t}\,\mathrm{d}\gamma_n\mathrm{d}\kappa$$



- **EAP computation :**

MCMC (Monte Carlo methods based on Markov Chain) algorithm

  - random sampling of vector $[x \quad \xi \quad \mathrm{t} \quad \gamma_{\mathrm{n}} \quad \kappa]$ under the a posteriori distribution
  - computation of the empirical mean of the samples

$$\left[\hat{x} \quad \hat{\xi} \quad \hat{\mathbf{t}} \quad \hat{\gamma}_n \quad \hat{\kappa}\right]=\frac{1}{K}\sum_{k=K_0}^{K+K_0-1}\left[x^{(k)} \quad \xi^{(k)} \quad \mathbf{t}^{(k)} \quad \gamma_n^{(k)} \quad \kappa^{(k)}\right]$$

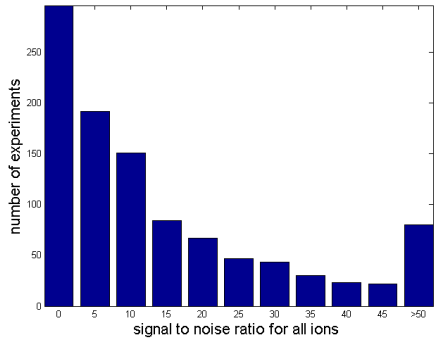
- **Hierarchical Gibbs sampling algorithm [10-13]:**
  - The sampling of a complex multivariate distribution is implemented as a combination of sampling of simpler monovariate distributions
  - Hierarchical decomposition of the Gibbs algorithm

Outline

1. Introduction
2. Hierarchical model of the LC-MS analytical chain
3. Digestion model
4. Bayesian hierarchical reconstruction
5. **Results on simulated data**
6. Conclusion

A simple simulated data set

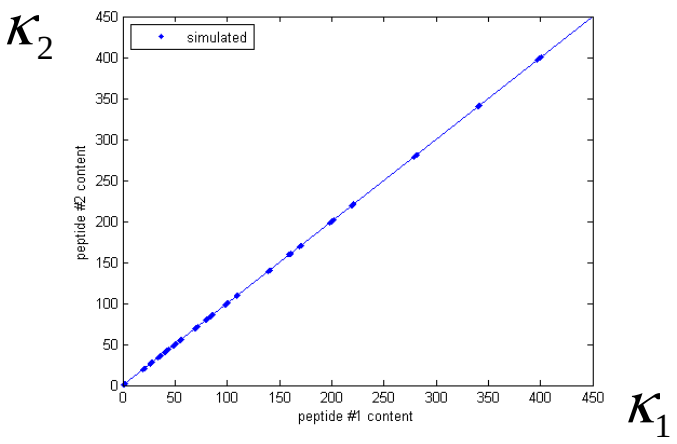
- Reference experimental framework [14]:
  - human NSE
  - combining immunoenrichment, trypsin digestion and LC-MS analysis on an Orbitrap
  - PSAQ quantification is performed spiking an isotopically-labeled version of NSE [7-8]
- 5 statistical protein content distributions of NSE within 1 ml serum samples :
  - Distribution mean: 1, 50, 100, 200, 400 ng/ml
- Digestion products of the protein:
  - Cleaved peptides: 2 peptides pep1 and pep2
  - Miscleaved molecule combining pep1 and pep2: pep3
- Technological variability:
  - 5 digestion yield parameter α: 0.4, 0.55, 0.7, 0.85, 1
  - variation of 10% on LC retention time
  - variation of 30% on MS ionization gain
  - measurement noise: see the SNR histogram
- Two data processing strategies:
  - case 1: protein content estimated from the 2 cleaved peptides pep1 and pep2
  - case 2: protein content estimated from the 2 cleaved peptides pep1 and pep2 and the miscleaved molecule pep3
- 345 spectrograms simulated

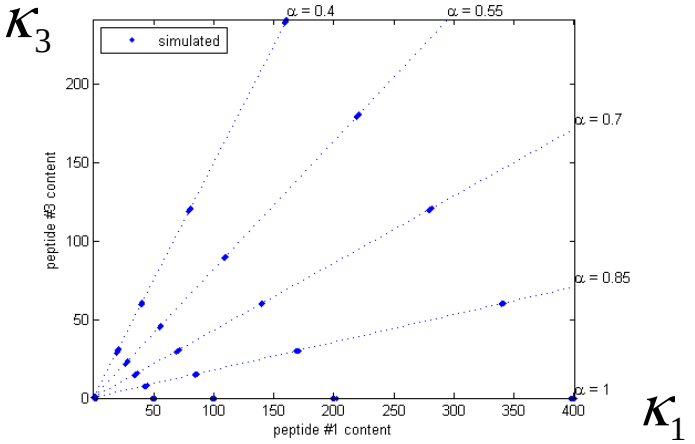


SNR histogram for all ions on simulated LC-MS data set

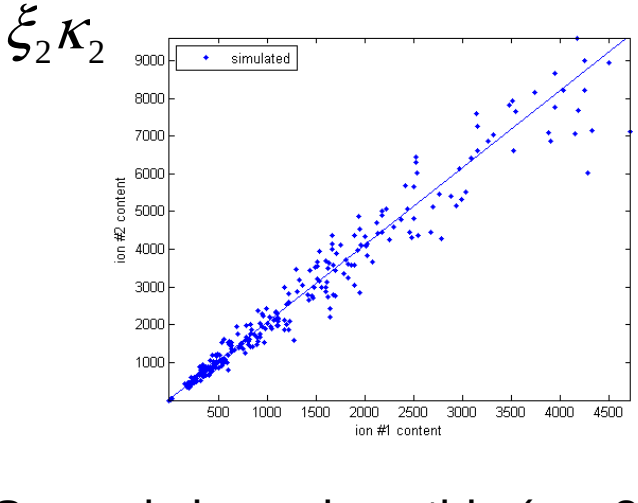
Technological variability on simulated data

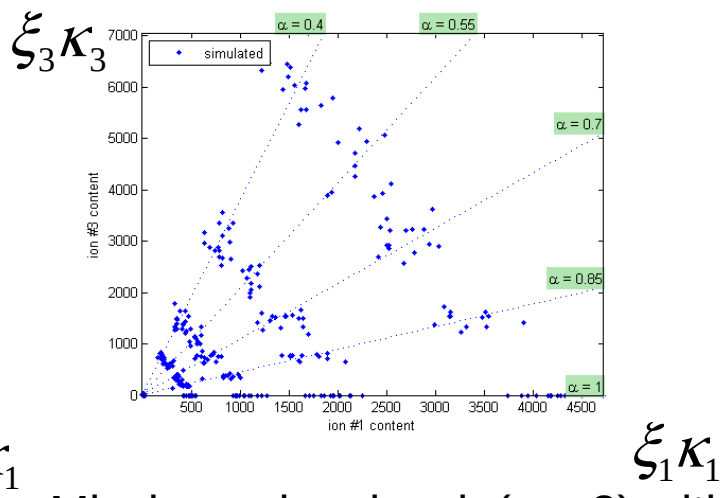
Peptide content





Ion content



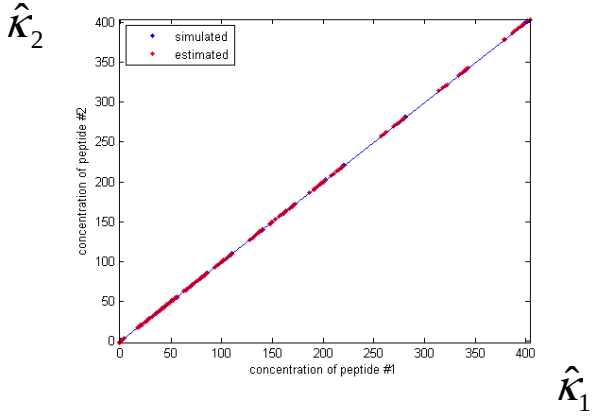


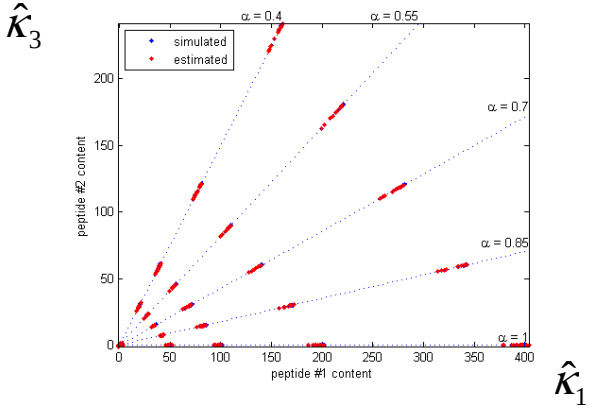
Second cleaved peptide (pep2) with respect to the first cleaved peptide (pep1)

Miscleaved molecule(pep3) with respect to the first cleaved peptide (pep1)

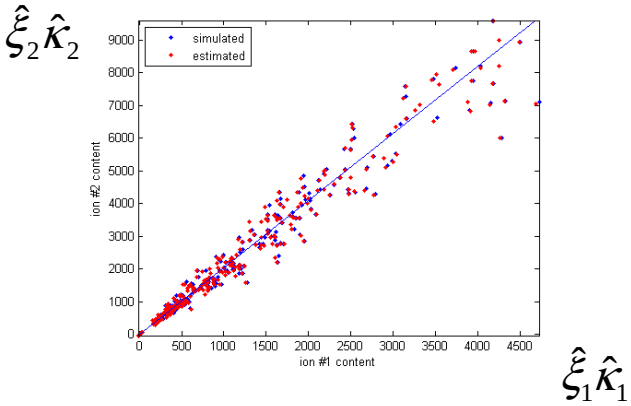
Estimated peptide and ion content after Bayesian profile reconstruction

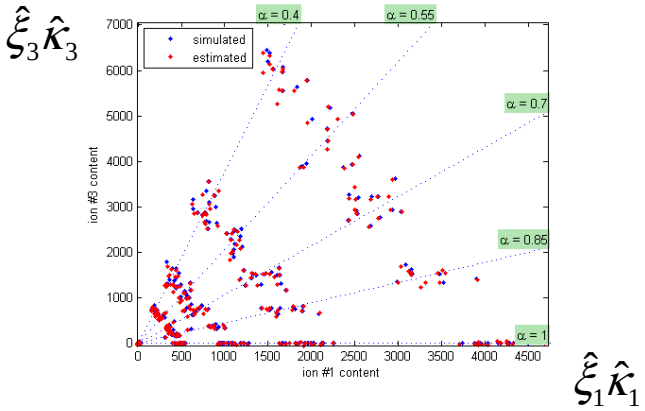
Peptide content





Ion content



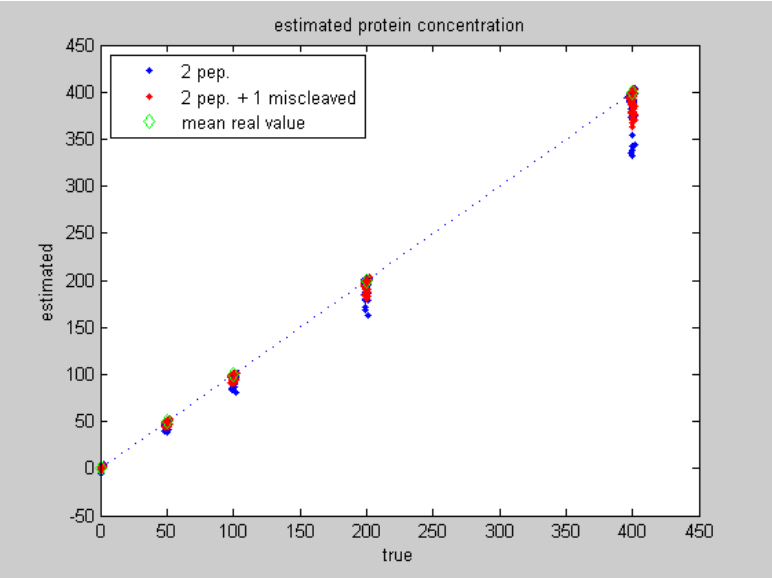


Second cleaved peptide (pep2) with respect to the first cleaved peptide (pep1)

Miscleaved molecule(pep3) with respect to the first cleaved peptide (pep1)

Estimation of protein content after Bayesian profile reconstruction

$\hat{X}$



X

mean value of true protein content (ng/ml)	1	50	100	200	400
CV2 = CV (%) of protein content estimated with 2 peptides	167.6	6.1	5.1	4.4	4.4
CV3 = CV (%) of protein content estimated with 3 peptides	151.6	4	3.4	2.8	2.9
CV3/CV2	0.9	0.7	0.7	0.6	0.7

Estimated versus real protein content

- with 2 peptides (pep1, pep2) (in blue)
- with 3 peptides (pep1, pep2, pep3) (in red)

Relative coefficient of variation on estimated protein content

Outline

1. Introduction
2. Hierarchical model of the LC-MS analytical chain
3. Digestion model
4. Bayesian hierarchical reconstruction
5. Results on simulated data
6. Conclusion

Conclusion

- Hierarchical parametric probabilistic model of the LC-MS analytical chain:
  - Hierarchical mixture model
  - Graph structure of the protein decomposition
  - Technological variability
  - Digestion model
- Bayesian hierarchical reconstruction:
  - Joint estimation of technological parameters and protein content
  - Recovery of peptide and protein content according to the hierarchical graph structure
  - Robust protein quantification
- Perspectives:
  - Joint Bayesian hierarchical inversion-classification [13].
  - BHI-PRO project

References

- [1] Gelman A. et al. (2003), *Bayesian Data Analysis, Second Edition (Texts in Statistical Science)*, Chapman & Hall/CRC..
- [2] Robert C. (2007), *The Bayesian choice : from decision-theoretic foundations to computational implementation*, Springer, New York, NY..
- [3] Do K.-H. et al. (2006), *Bayesian Inference for Gene Expression and Proteomics*, Cambridge University Press, New York, USA.
- [4] Idier J., Ed. (2008), *Bayesian Approach to Inverse Problems*, ISTE Ltd and John Wiley & Sons Inc., London, 2008.
- [5] Grangeat P. (Ed) (2009), *Tomography*, ISTE Ltd, London, UK and John Wiley & Sons Inc., Hoboken, USA.
- [6] Grangeat P. (2009), "Data processing", in. M. Lahmani, P. Boisseau, P. Houdy: *Nanobiotechnology and Nanobiology*, chap. 13, 775-802, Springer.
- [7] Brun V. et al. (2007), " Isotope-labeled Protein Standards "**,** Mol. and Cell. Proteomics 6.12, 2139-2149.
- [8] Brun V. et al. [2009), " Isotope dilution strategies for absolute quantitative proteomics ", *Journal of Proteomics*, vol. 72, no. 5, pp. 740–749, 2009.
- [9] Finehout et al. (2005), " Kinetic characterization of sequencing grade modified trypsin "**,** Proteomics, 5, 2319-2321.
- [10] Strubel G. et al. (2007), "Bayesian estimation for molecular profile reconstruction in proteomics based on liquid chromatography and mass spectrometry" , 29th IEEE EMBS Conference, Lyon, France, 5979-5982.
- [11] Strubel G. et al. (2008), "Robust protein quantification in mass spectrometry", HUPO 7th annual world congress, 16-20 August 2008, Amsterdam, The Netherlands.
- [12] Grangeat P. et al. (2009), "Robust statistical reconstruction of protein profiles in mass spectrometry", 57th ASMS Conference on Mass Spectrometry, Philadelphia, Pennsylvania, USA.
- [13] Szacherski P. et al. (2011), "Joint Bayesian hierarchical inversion-classification and application in proteomics", SSP2011, 2011 IEEE Workshop on Statistical Signal Processing, Nice, France.
- [14] Grangeat et al. (2010), "First demonstration on NSE biomarker of a computational environment dedicated to lab-on-chip based cancer diagnosis“, 58th ASMS Conference on Mass Spectrometry, Salt Lake City, USA.

LABORATOIRE D'ÉLECTRONIQUE
ET DE TECHNOLOGIES
DE L'INFORMATION

CEA-Leti
MINATEC Campus, 17 rue des Martyrs
38054 GRENOBLE Cedex 9
Tel. +33 4 38 78 36 25

www.leti.fr

Thank you for your attention

Contact: pierre.grangeat@cea.fr