JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 2, 2009
© Mary Ann Liebert, Inc.
Pp. 395–406
DOI: 10.1089/cmb.2008.22TT

Modeling Exopeptidase Activity from LC-MS Data

BOGUSŁAW KLUGE,[1] ANNA GAMBIN,[1] and WOJCIECH NIEMIRO[2]

ABSTRACT

Recent studies demonstrate that the peptides in the serum of cancer patients that are generated (*ex vivo*) as a result of tumor protease activity can be used for the detection and classification of cancer. In this paper, we propose the first formal approach to modeling exopeptidase activity from liquid chromatography–mass spectrometry (LC-MS) samples. We design a statistical model of peptidome degradation and a Metropolis-Hastings algorithm for Bayesian inference of model parameters. The model is successfully validated on a real LC-MS dataset. Our findings support the hypotheses about disease-specific exopeptidase activity, which can lead to new diagnostic approach in clinical proteomics.

Key words: exopeptidase activity, liquid chromatography mass spectrometry, Markov chain Monte Carlo, proteomics, stochastic modeling.

1. INTRODUCTION

With the development of proteomic analytic technologies, especially mass spectrometry (MS), great hopes for early diagnostics of cancer were expressed (Petricoin et al., 2002). However, the initial optimism has encountered strong criticism. The criticism was addressed not against the idea of using protein profiles as a diagnostic tool but against poor quality of data obtained from SELDI type detectors and non-reproducibility of experimental conditions (Diamandis, 2003, 2004).

Moreover, despite years of intensive MS analysis, only a small number of proteins have been validated as cancer biomarkers. Also, the MS samples where characterized as highly unstable, mainly because of *ex vivo* proteolytic processing (Marshall et al., 2003; Verrills, 2006). Changes in protein profiles can be generated simply by the amount of time between sample draw and analysis. Surprisingly, this obstacle gives rise to a completely new approach enthusiastically described as “spinning biological trash into diagnostic gold” (Liotta and Petricoin, 2006).

1.1. Research objective

In Diamandis (2006), the advantages and limitations of clinical peptidomics were summarized. The authors proposed to characterize the proteolytic activity, as it could lead to better patient discrimination. Therefore, our research objective was to build a mathematical model of exopeptidase activity and to check whether the model exhibits differences between samples from healthy donors and diseased patients.

[1] Institute of Informatics, University of Warsaw, Warsaw, Poland.
[2] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland.

1.2. *Related research*

In a typical liquid chromatography–mass spectrometry (LC-MS) experiment, a complex mixture of peptides is separated using liquid chromatography coupled on-line with electrospray mass spectrometer. After appropriate preprocessing (Gambin et al., 2007), each detected peptide is characterized by two coordinates—its molecular mass to charge ratio and retention time value.

Much work has already been invested into detection of molecular mass biomarkers for various pathologies and diagnostic procedures have been suggested (Adam et al., 2002; Geurts et al., 2005; Jacobs and Menon, 2004; Li et al., 2002; Lilien et al., 2003; Tibshirani et al., 2004; Wu et al., 2003; Yu et al., 2005). Unfortunately, it is extremely hard to obtain stable MS results reproducible over time and across different laboratories (Hu et al., 2005). Often the differences in sample collection or sample handling protocol affect the proteome to a degree that can dominate biological changes. Also, the *ex vivo* peptide degradation process was regarded as a serious obstacle in MS analysis.

Recently, a novel way of diagnosing cancer was suggested in Villanueva et al. (2006a, 2006b). The authors postulate that the diagnostic peptides originate after *ex vivo* exoproteolytic processing of high abundance protein fragments. Paradoxically, these findings indicate that inhibition of proteolysis in *ex vivo* samples could limit biomarker discovery. See also Koomen et al. (2005) for the information on the peptidome degradation process analyzed with the use of mass spectrometry technology.

Using peptide degradation pattern for the diagnostic purposes seems biologically sound as the amount of peptides in the circulation changes dynamically according to the physiological or pathological state of an individual. Moreover, it was reported that the degradation enzymes (especially exopeptidases) affect the dynamics of signaling pathways (Reznik and Fricker, 2001).

Even though there exists a large body of research concerning modeling enzymatic reaction systems with differential equations (Ciliberto et al., 2007), to the best of our knowledge this work is the first attempt to build a model specifically with exopeptidase activity in mind.

1.3. *Our results*

We propose a comprehensive statistical and computational framework for analysis of peptide degradation patterns in LC-MS samples. In our approach, the exopeptidase activity is modeled as a continuous time Markov process. The stationary distribution of this process is proved to be a product of Poisson laws. A Metropolis-Hastings sampler is implemented to estimate the parameters of the model. These correspond to the rates of cleavage for different amino acids. The model is tested on simulated data and validated on a colorectal cancer dataset. Parameter estimates for diseased patients and healthy donors differ significantly and allow for accurate classification. Moreover, the estimated differences in activity of proteolytic enzymes in cancer and healthy samples correlates with experimentally verified activity of metallopeptidases in colorectal cancer development (Leeman et al., 2003; Masaki et al., 2001). The scheme of data processing and analysis workflow is depicted in Figure 1.

1.4. *Availability*

The source code (R with C) of our estimation procedure is freely available at *http://bioputer.mimuw.edu.pl/papers/exopep*. The site also contains additional figures and peptide sequences generating the cleavage graph.

2. RESULTS AND DISCUSSION

Our model has two main components: the first one describes the cleavage (peptide degradation) process itself, while the second accounts for imperfections at the data acquisition stage.

2.1. *Model for the cleavage process*

Peptide sequences whose proteolysis we wish to model give rise to a graph $(\mathcal{V}, \mathcal{E})$, which we will call the *cleavage graph*. Nodes $\mathcal{V}$ of this graph correspond to all peptide subsequences of length at least 2. A directed edge from node i to j is placed if subsequence j can be obtained from subsequence i by

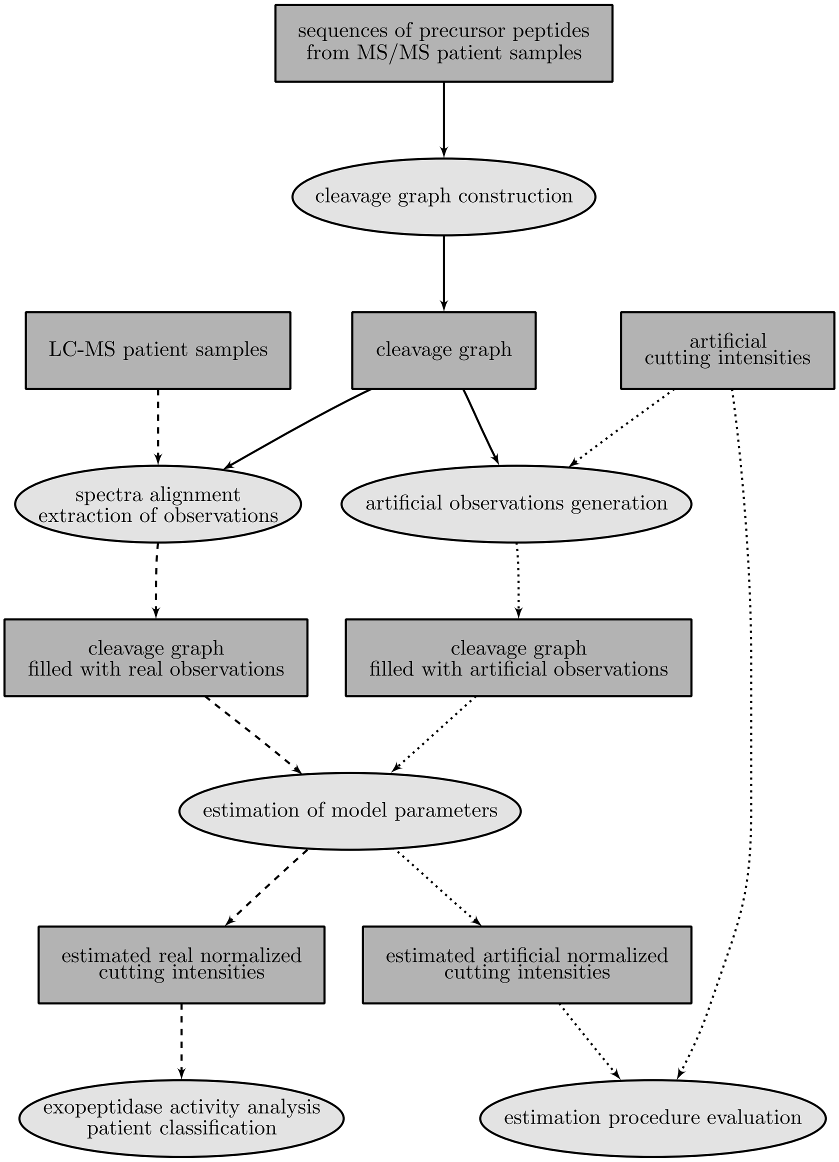
sequences of precursor peptides from MS/MS patient samples
cleavage graph construction
LC-MS patient samples
cleavage graph
artificial cutting intensities
spectra alignment extraction of observations
artificial observations generation
cleavage graph filled with real observations
cleavage graph filled with artificial observations
estimation of model parameters
estimated real normalized cutting intensities
estimated artificial normalized cutting intensities
exopeptidase activity analysis patient classification
estimation procedure evaluation

FIG. 1. Data processing and analysis workflow.

cutting off a single amino acid from the N-terminus or the C-terminus. Each edge is labeled with the amino acid being cut off and the terminus it is being cut off from, thus the set $\mathcal{R}$ of possible labels has 20×2 elements. The label for edge $i \to j$ is denoted by $r(i, j)$. We assume that the labeling and structure of the cleavage graph is known. An exemplary cleavage graph is presented in Figure 2.

It is helpful to think of the peptide subsequences as particles placed at nodes of the cleavage graph and moving along its edges. Then the probabilistic dynamics of the cleavage process is described by the following intensities of transition:

- particles are created at node i with intensity $a_{\star i}$,
- every particle placed at i can move to j with intensity $a_{r(i,j)}$ independently of all other particles, provided that there exists an edge $i \to j$,
- every particle placed at i can be annihilated with intensity $a_{i\dagger}$ independently of all other particles.

We refer to the $(a_r)_{r\in\mathcal{R}}$ parameters as the *cutting intensities.*

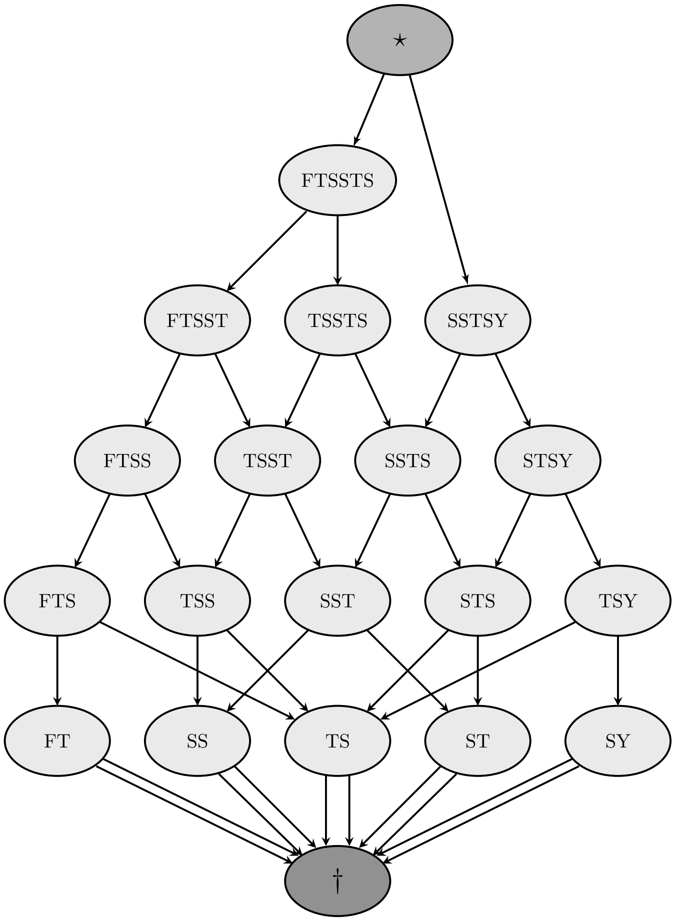

FIG. 2. The cleavage graph for two precursor peptides, FTSSTS and SSTSY, with source and sink nodes added.

More formally, let random variable $X_i(t)$ denote the number of particles at node $i \in \mathcal{V}$ at time t and write $X(t) = (X_i(t))_{i \in \mathcal{V}}$. We regard $(X(t), t \geq 0)$ as a homogeneous Markov process in the space of configurations $x = (x_i)_{i \in \mathcal{V}}$, $x_i \in \{0, 1, \dots\}$. We use the standard notation for restricted configurations, writing e.g., $x_{-i} = (x_k)_{k \in \mathcal{V}:\, k \neq i}$. The process has the following intensities of transition ($x \neq x'$):

$$Q(x, x') = \begin{cases} a_{\star i} & \text{if } x'_i = x_i + 1,\ x'_{-i} = x_{-i} \text{ for some } i, \\ a_{r(i,j)} x_i & \text{if } x'_j = x_j + 1,\ x'_i = x_i - 1, \\ & \text{and } x'_{-i-j} = x_{-i-j} \text{ for some } i \to j, \\ a_{i\dagger} x_i & \text{if } x'_i = x_i - 1,\ x'_{-i} = x_{-i} \text{ for some } i. \end{cases}$$

We assume that the process reached the equilibrium state. At each node, we are interested in the distribution of the number of particles. Perhaps surprisingly, we can prove that those numbers are independent and each one follows a Poisson distribution.

Proposition 1 (Equilibrium distribution). *The process $(X(t))$ has the equilibrium (stationary) distribution π given by:*

$$\pi(x) = \prod_{i \in \mathcal{V}} \mathrm{e}^{\lambda_i} \frac{\lambda_i^{x_i}}{x_i!},$$

where the configuration of intensities $(\lambda_i)_{i \in \mathcal{V}}$ *is the unique solution to the following system of "balance" equations:*

$$\sum_{k \to i} \lambda_k a_{r(k,i)} + a_{\star i} = \lambda_i \left(\sum_{i \to j} a_{r(i,j)} + a_{i\dagger} \right) \qquad \textit{for every } i \in \mathcal{V}.$$

Note that it is easy to solve the system of "balance" equations recursively starting from the nodes without parents. The proposition can be proved by simply checking the global balance condition (i.e., that for every configuration x the equality $\sum_{x' \neq x} \pi(x) Q(x, x') = \sum_{x' \neq x} \pi(x') Q(x', x)$ holds).

The above description of the cleavage process is valid for any directed acyclic graph. Since we are concerned with exopeptidase activity modeling, we impose some restrictions. Let $\mathcal{V}_{\text{in}}$ be the set of nodes that have no parents. We set $a_{\star i}$ to 0 for $i \in \mathcal{V} \setminus \mathcal{V}_{\text{in}}$ and $a_{i\dagger}$ to 0 if node i has children. If i has no children then $\alpha_{i\dagger}$ is expressed as a sum of two elements from $\{a_r \mid r \in \mathcal{R}\}$, corresponding to the amino acids on both ends of subsequence i.

In order to go further with the description of the model, we need to change the parameterization a little bit. Write $b_{\star i} = s_1^{-1} a_{\star i}$ for $i \in \mathcal{V}$, where $s_1 = \sum_{i \in \mathcal{V}} a_{\star i}$ forcing $\sum_{i \in \mathcal{V}} b_{\star i} = 1$ and similarly $b_r = s_2^{-1} a_r$ for $r \in \mathcal{R}$, where $s_2 = \sum_{r \in \mathcal{R}} a_r$ forcing $\sum_{r \in \mathcal{R}} b_r = 1$. Now we can express λ_i as $s\mu_i$, $s = \frac{s_2}{s_1}$ for $i \in \mathcal{V}$ where μ_i depend only on $(b_r)_{r \in \mathcal{R}}$ and $(b_{\star k})_{k \in \mathcal{V}_{\text{in}}}$. We place a Gamma prior with parameters $(S_{\text{shape}}, S_{\text{rate}})$ on s and a Dirichlet prior with parameters $(B_r)_{r \in \mathcal{R}}$ on $(b_r)_{r \in \mathcal{R}}$ and $(B_{\star i})_{i \in \mathcal{V}_{\text{in}}}$ on $(b_{\star i})_{i \in \mathcal{V}_{\text{in}}}$. Since we are interested in relative intensities only, our goal is to estimate $(b_r)_{r \in \mathcal{R}}$, which we will call the *normalized cutting intensities*.

2.2. *Model for data acquisition*

Ideally, after the data preprocessing step one would get an exact reading on the numbers of particles corresponding to every possible subsequence present in the cleavage graph. In reality, we must deal with many kinds of experimental errors.

First of all, many readings are missing. We can see which readings are missing and which are not. A vector of binary variables $(\epsilon_i)_{i \in \mathcal{V}}$ indicates the non-missing readings.

Some of the non-missing readings may be incorrect, meaning that they are taken from the wrong peaks from the LC-MS spectra, and have little to do with the peptides mentioned in the cleavage graph. This information is hidden and modeled by the δ variables coming from a Bernoulli process with success probability q.

Moreover, assuming that each correct reading is a sample from a Poisson distribution would imply low relative errors for readings from high peaks. This is clearly not realistic in case of the LC-MS data. Therefore, we assume that correct readings y_i for i such that $\delta_i = 1$ come from independent log-normal distributions with parameters $\ln x_i$ and τ (see Equation (1)), where x is the hidden realization of the cleavage process. Incorrect readings y_i for i such that $\delta_i = 0$ come independently from a background distribution with density bg. This density is estimated from the data (all mono-isotopic peak intensities in an LC-MS sample).

Note that from now on we define δ_i, x_i and y_i only for $i \in \mathcal{V}$ such that $\epsilon_i = 1$. When we write $i : \delta_i = 1$ we mean only those indices i, for which δ_i is defined. When we write x, we mean $(x_i)_{i\,:\,\epsilon_i = 1}$, for example.

2.3. *Posterior distribution*

The dependence structure of the variables in the hierarchical Bayesian model is shown in Figure 3. The posterior distribution can be written as:

$$f(s, b_\star, b, \delta, x \mid y) \propto f(y \mid s, b_\star, b, \delta, x) f(s, b_\star, b, \delta, x)$$

$$= f(y \mid \delta, x) f(\delta) f(x \mid s, b_\star, b) f(s) f(b_\star) f(b),$$

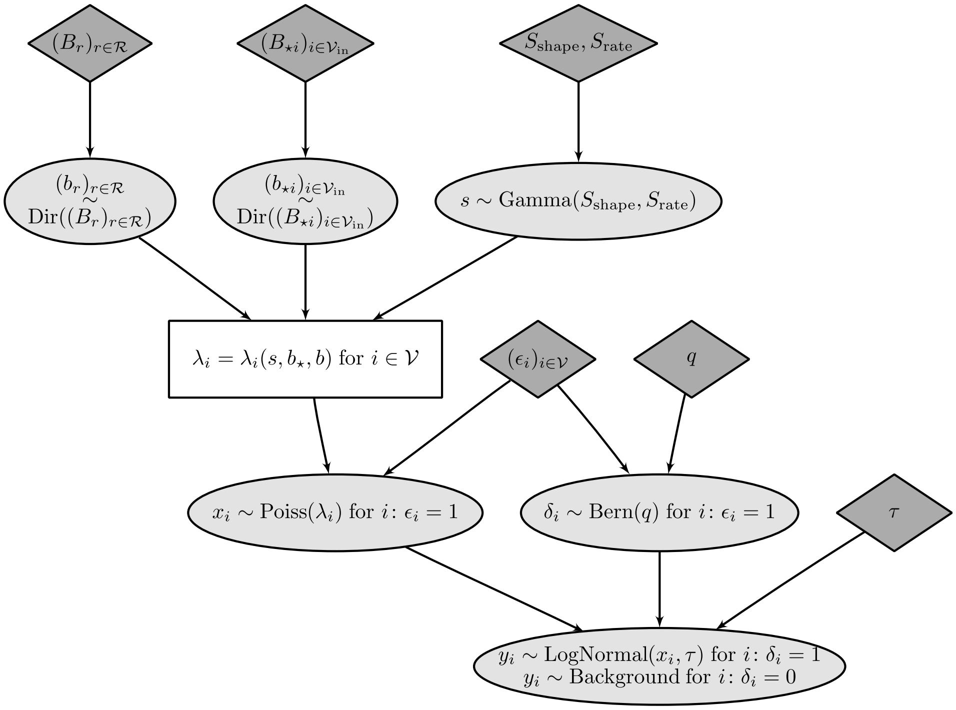
$(B_r)_{r\in\mathcal{R}}$
$(B_{\star i})_{i\in\mathcal{V}_{\text{in}}}$
$S_{\text{shape}}, S_{\text{rate}}$
$(b_r)_{r\in\mathcal{R}} \sim \text{Dir}((B_r)_{r\in\mathcal{R}})$
$(b_{\star i})_{i\in\mathcal{V}_{\text{in}}} \sim \text{Dir}((B_{\star i})_{i\in\mathcal{V}_{\text{in}}})$
$s \sim \text{Gamma}(S_{\text{shape}}, S_{\text{rate}})$
$\lambda_i = \lambda_i(s, b_\star, b)$ for $i \in \mathcal{V}$
$(\epsilon_i)_{i\in\mathcal{V}}$
q
$x_i \sim \text{Poiss}(\lambda_i)$ for $i\colon \epsilon_i = 1$
$\delta_i \sim \text{Bern}(q)$ for $i\colon \epsilon_i = 1$
τ
$y_i \sim \text{LogNormal}(x_i, \tau)$ for $i\colon \delta_i = 1$
$y_i \sim$ Background for $i\colon \delta_i = 0$

FIG. 3. The hierarchical Bayesian model of cleavage activity and data acquisition.

where:

$$f(y \mid \delta, x) = \prod_{i\,:\,\delta_i=0} \text{bg}(y_i) \prod_{i\,:\,\delta_i=1} \frac{1}{y_i \tau \sqrt{2\pi}} \mathrm{e}^{-\frac{(\ln y_i - \ln x_i)^2}{2\tau^2}}, \tag{1}$$

$$f(\delta) = q^{|\{i|\delta_i=1\}|}(1-q)^{|\{i|\delta_i=0\}|},$$

$$f(x \mid s, b_\star, b) = s^{\sum_i x_i} \prod_i \frac{\mu_i^{x_i}}{x_i!} \mathrm{e}^{-s\mu_i},$$

$$f(s) = s^{S_{\text{shape}}-1} \frac{S_{\text{rate}}^{S_{\text{shape}}} \mathrm{e}^{-S_{\text{rate}}\, s}}{\Gamma(S_{\text{shape}})},$$

$$f(b) = \frac{\Gamma\left(\sum_{r\in\mathcal{R}} B_r\right)}{\prod_{r\in\mathcal{R}} \Gamma(B_r)} \prod_{r\in\mathcal{R}} b_r^{B_r-1},$$

$$f(b_\star) = \frac{\Gamma\left(\sum_{i\in\mathcal{V}_{\text{in}}} B_{\star i}\right)}{\prod_{i\in\mathcal{V}_{\text{in}}} \Gamma(B_{\star i})} \prod_{i\in\mathcal{V}_{\text{in}}} b_{\star i}^{B_{\star i}-1}.$$

Integrating out s and δ gives:

$$f(b_\star, b, x \mid y) \propto f(y \mid x) f(x \mid b_\star, b) f(b) f(b_\star),$$

where:

$$f(y \mid x) = \prod_i \left((1-q)\mathrm{bg}(y_i) + \frac{q}{y_i \tau \sqrt{2\pi}} \mathrm{e}^{-\frac{(\ln y_i - \ln x_i)^2}{2\tau^2}} \right),$$

$$f(x \mid b_\star, b) = \frac{S_{\mathrm{rate}}^{S_{\mathrm{shape}}}}{\Gamma(S_{\mathrm{shape}})} \frac{\Gamma\left(S_{\mathrm{shape}} + \sum_i x_i\right)}{\left(S_{\mathrm{rate}} + \sum_i \mu_i\right)^{S_{\mathrm{shape}} + \sum_i x_i}} \prod_i \frac{\mu_i^{x_i}}{x_i!}.$$

2.4. Estimation procedure

We wish to estimate the $(b_r)_{r \in \mathcal{R}}$ parameters. The closed form of the expression $f(b_\star, b, x \mid y)$ was derived in the previous section. Since we are unable to integrate out $b_\star$ and x, we use the Metropolis–Hastings algorithm with the standard acceptance rule to sample $(b_\star, b, x)$ from the posterior distribution.

Transition proposal is generated by selecting with equal probability one of the three following rules:

1. Changing $b_\star$:
   - generate $i, j \in \mathcal{V}_{\mathrm{in}}$, $i \neq j$ uniformly,
   - generate

     $$(b'_{\star i}, b'_{\star j}) \sim (b_{\star i} + b_{\star j})\mathrm{Dir}\left(c\frac{b_{\star i}}{b_{\star i} + b_{\star j}} + 1, c\frac{b_{\star j}}{b_{\star i} + b_{\star j}} + 1 \right),$$

     where c is a parameter of the procedure,
   - set $b'_{\star k}$ to $b_{\star k}$ for $k \notin \{i, j\}$,
   - propose transition $(b_\star, b, x) \mapsto (b'_\star, b, x)$,
2. changing b (analogously to changing $b_\star$),
3. changing x:
   - generate i such that $\epsilon_i = 1$ uniformly,
   - generate $x'_i \sim \mathrm{LogNormal}(\ln x_i, d)$, where d is a parameter of the procedure,
   - set x'_k to x_k for $k \neq i$,
   - propose transition $(b_\star, b, x) \mapsto (b_\star, b, x')$.

2.5. Model testing

Three datasets (readings for the nodes of the cleavage graph) were generated according to the model with parameters $B_{\star i} = 2$ for $i \in \mathcal{V}_{\mathrm{in}}$, $B_r = 2$ for $r \in \mathcal{R}$, $s = 10^6$, $\tau = 0.2$. Based on each of these datasets, another dataset was derived by selecting nodes with correct readings with $q = 0.7$. Readings at all other nodes were resampled as being incorrect, each with λ parameter selected uniformly from $(\lambda_i)_{i \in \mathcal{V}}$. Thus, we have two version of each of the three datasets—with and without incorrect readings.

The Metropolis-Hastings algorithm was run with parameters $c = 80$ (for changing $b_\star$ and b) and $d = 0.05$ (for changing x) for 3×10^6 iterations to recover the $(b_r)_{r \in \mathcal{R}}$ parameters.

Initial $b_\star$ and b were sampled from the Dirichlet priors. Initial x_i parameters were sampled from the log-normal distributions with parameters $\ln y_i, \tau$.

During the algorithm run the same Dirichlet priors and τ as during data generation were used. The S_{shape}, S_{rate} parameters were set to 0 and the q parameter was set appropriately to 1 or 0.7. On each dataset, eight independent algorithm runs were conducted (therefore in total there were $3 \times 2 \times 8$ algorithm runs), each time with randomly selected 90% of the readings hidden as missing data. This was motivated by the fact that on real data only about 10% of the nodes of the cleavage graph had readings.

Results for one dataset with and without incorrect readings are presented in Figure 4. Results for other datasets look similarly and are available as supplementary materials (see online Supplementary Material at *www.liebertonline.com*). Clearly the estimates for data without incorrect readings are very accurate. Since the estimates for data with incorrect readings are visually less appealing, we computed the Aitchison distance (see Section 3.1) between the averaged (over eight runs using the Aitchison mean) estimated

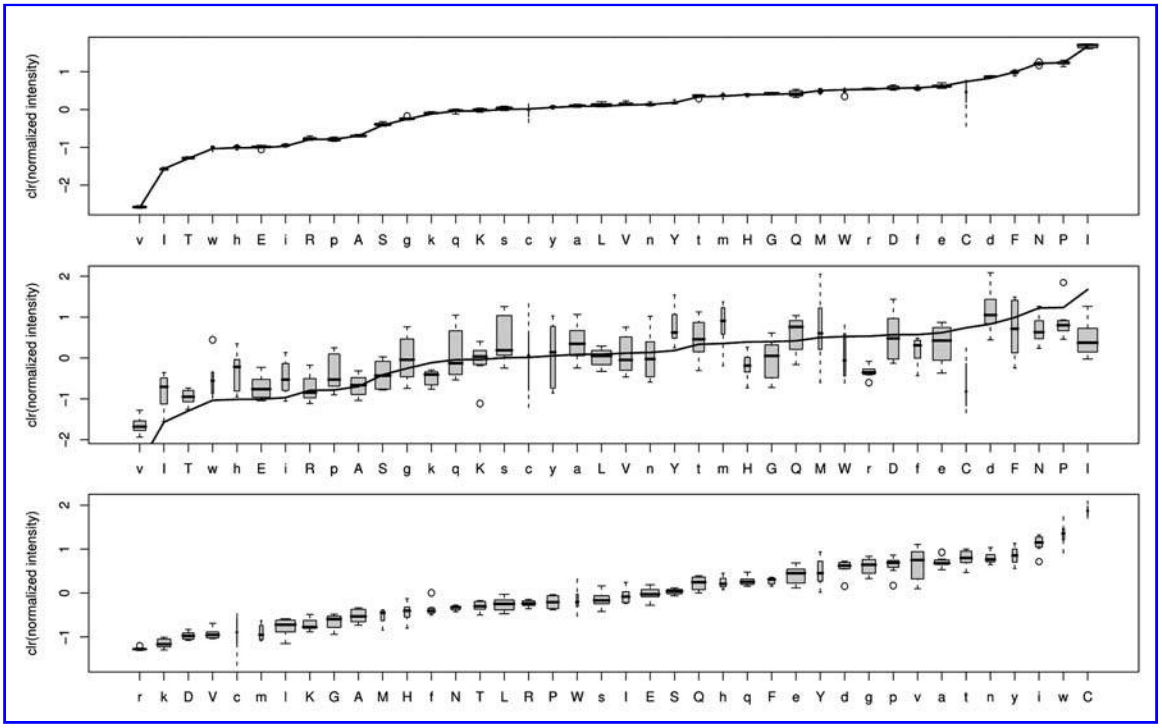

FIG. 4. Clr-transformed normalized amino acid cutting intensities estimated from an artificial dataset without (top) and with (middle) incorrect readings and from a real dataset (bottom) based on eight runs of the algorithm. The horizontal axis is sorted by the true intensities marked with the thick black line (top, middle) or by the Aitchison mean (bottom). Uppercase letters denote cutting from the N-terminus; lowercase letters denote cutting from the C-terminus. The width of the bars is proportional to the number of appearances of the corresponding amino acid in the maximal vertices of the cleavage graph (i.e., the precursor peptides).

intensities and the true intensities. The results were 3.12 (dataset in Fig. 4), 2.64, and 3.23. To give those numbers some meaning, if we take two points from the Dirichlet prior, then with probability greater than 0.999 the Aitchison distance is greater than 4.5.

2.6. *Validation on LC-MS data*

In order to illustrate the applicability of the model to real data, we analyzed the colorectal cancer dataset. We show that our model can be used to discriminate between diseased patients and healthy donors. In each of the 29 samples, about 90% readings in the nodes of the cleavage graph were missing. Since there were only about 0.7% nodes with readings from all samples, it is not straightforward to bypass the model parameters estimation and perform classification directly on the data (one would have to deal somehow with the missing data). Therefore, we leave comparison with other classification methods as a topic for further research, but we stress that our model can provide insights into the peptide degradation process.

The estimation algorithm was run eight times on each sample with the q parameter set to 0.7 and all other parameters as described in the previous section. Figure 4 shows that the results are quite consistent (additional figures can be found in the supplementary materials, see online Supplementary Material at *www.liebertonline.com*). Obtained intensities were averaged over these eight runs using the Aitchison mean.

Figure 5 shows the data projected on the first three principal components (accounting for almost 75% of total variance). The first and third components are heavily influenced by Cysteine cutting intensity (see loadings on Fig. 5) and do not discriminate samples well. The second component is the only one significant in this respect (Bonferroni corrected p-value from Kolmogorov-Smirnov test below 0.01). We tried to confirm whether it carries the information about the altered pattern of exopeptidase activity.

To determine the hypothetical enzymes involved, we scanned MEROPS database (Rawlings and Barrett, 2000) for peptidases cleaving the specific bond (e.g., Threonine, Valine, and Histidine from C-terminus,

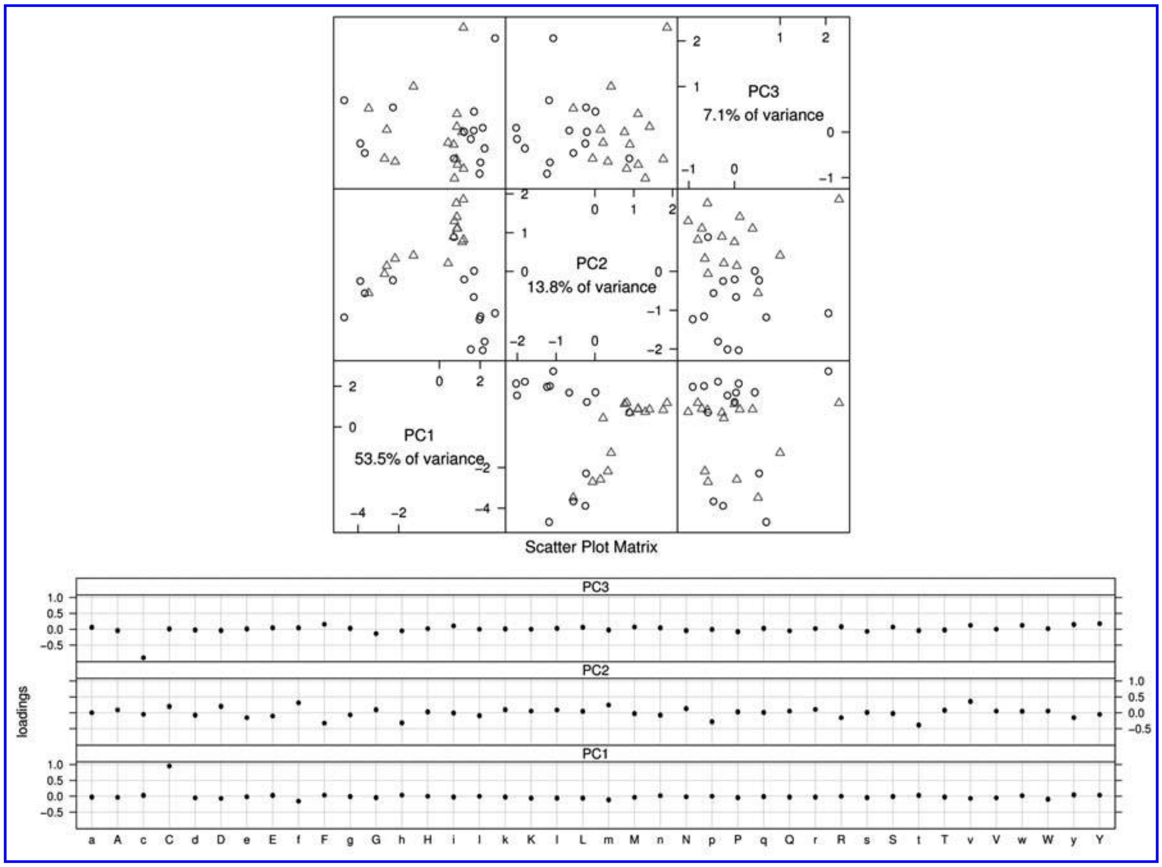

FIG. 5. Scatter plot of the colorectal cancer dataset for the first three principal components and the corresponding loadings. Healthy donors are marked with black circles. Diseased patients are marked with gray triangles. Uppercase letters denote cutting from the N-terminus; lowercase letters denote cutting from the C-terminus.

Phenylalanine and Aspartic acid from N-terminus; Fig. 5). Resulting list contains many metallopeptidases, which are experimentally verified as crucial for colorectal cancer development (Leeman et al., 2003; Masaki et al., 2001).

We also investigated whether patient classification based on clr-transformed normalized cutting intensities can be performed. Using the SVM classifier (Schölkopf and Smola, 2002) with linear kernel, the .632+ bootstrap (Efron and Tibshirani, 1997) error estimate based on 1000 bootstrap replicates was 12.4%. We repeated the whole procedure 1000 times with class labels permuted randomly. The average .632+ bootstrap error estimate was 54.5% with standard deviation 12.5%, suggesting that cutting intensities indeed contain information about patient state (and perhaps that the .632+ estimator is too pessimistic).

2.7. Conclusions

This work models the protein degradation process from LC-MS data. We described a mathematical framework allowing for adequate statistical modeling. The model was extensively tested on suitably chosen artificial datasets, as well as on real LC-MS samples.

The outcome of computational experiments is very promising. The estimation procedure yielded robust results even when dealing with errors and missing values in the input data. Moreover, the accurate classification results for colorectal cancer patients suggest diagnostic potential of the model.

On the other hand, we are aware of the problems with reproducibility of the LC-MS experiments. Two more datasets were at our disposal. They were collected at different times and processed with different HPLC columns. After preliminary analyzes we decided however to base the presentation of our model only on one dataset, because the results were hard to compare between different datasets. We believe that better LC-MS spectra alignment procedures would decrease the variability between the datasets.

Recently, new diagnostic test has been proposed to compare proteolytic activities within individual proteome of two groups of biological samples (Villanueva et al., 2008). It tracks degradation of artificial substrates under strictly controlled conditions. We plan to adopt our model to this setting and infer the hypotheses on the activity of postulated but as yet unidentified exopeptidases.

Finally, some concerns may be raised regarding the cleavage process stationarity assumption, especially as it is hard to strictly control the time between sample collection and MS analysis. We consider modifying our model to remove this assumption.

3. METHODS

3.1. Compositional data analysis

The normalized cutting intensities lie on a simplex. The theory for analysis of such data (termed *compositional data analysis*) is summarized in Aitchison and Egozcue (2005). In short, it consists of interpreting the simplex as a Euclidean linear vector space and then applying standard analysis techniques. We use the following concepts:

- the *centered log ratio* transform of a point $(z_i)_{i=1,\dots,n}$ on a simplex is defined as:

$$\mathrm{clr}(z) = \left(\ln \frac{z_i}{\mathrm{g}(z)}\right)_{i=1,\dots,n},$$

  where g denotes the geometric mean,
- the *Aitchison distance* is the Euclidean distance between clr-transformed points,
- the analogue of the expected value of a variable $(Z_i)_{i=1,\dots,n}$ on a simplex is the *Aitchison mean* defined as:

$$\mathcal{C}((\exp \mathrm{E}[\ln Z_i])_{i=1,\dots,n}),$$

  where $\mathcal{C}$ denotes rescaling of the components so that their sum is 1,
- the analogue of principal component analysis is well defined (it amounts to performing PCA on clr-transformed data).

3.2. Colorectal cancer dataset

LC-MS and MS-MS data was provided by the Mass Spectrometry Laboratory from the Institute of Biochemistry and Biophysics of the Polish Academy of Sciences. The mass spectrometer used in the experiments was an ElectroSpray Ionization Fourier Transform Ion Cyclotron Resonance (ESI-FTICR) coupled with an HPLC retention column.

The dataset comprised mass spectra acquired from serum samples for colorectal cancer patients. Apart from the patient data, control samples were also collected from healthy donors and analyzed with the mass spectrometer. The colorectal cancer dataset consisted of 29 spectra, 15 samples corresponding to diseased patients and 14 to healthy donors.

3.3. Cleavage graph construction

For the construction of the cleavage graph we start with the information about successfully sequenced peptides from a LC-MS/MS experiment. This information covers a little over 1000 peptides and is comprised of peptide mass, mass to charge ratio, retention time, charge, amino acid sequence, protein of origin, and optionally the information about the oxidation for each peptide. For simplicity, the oxidated molecules are omitted. The set of vertices of the cleavage graph is defined using all other maximal sequences (i.e., the precursor peptides; they can be found in the supplementary materials, see online Supplementary Material at *www.liebertonline.com*) and their subsequences. It has 39,544 elements, including 243 maximal vertices. This graph is fixed in all our experiments (both on artificial and real data).

3.4. Data preprocessing

For each spectrum, we used the *mz2m* program (Gambin et al., 2007) to obtain a list of mono-isotopic peak coordinates (m/z values and retention times) together with their charges and intensities.

For each cleavage graph vertex encoded by amino acid sequence, we need to find a corresponding signal in the spectrum. One can easily compute mass of the sequence and consider mass to charge ratios for charge $z \in \{1, \dots, 8\}$ (greater charges do not occur in the data). There is a problem however with the retention time. As we mentioned in Section 3.3, the retention time is readily available for some sequences. We use this information to train the Random Forest regression algorithm (Breiman, 2001) to predict the retention time from the amino acid composition.

The lists of mono-isotopic peaks from spectra together with the list describing the nodes of the graph were aligned by applying to each list a linear transformation along the retention time axis.

Assuming we know the retention time and several possible mass to charge ratios for a given sequence, we find peaks nearest to those locations on the LC-MS spectrum. The Euclidean metric is used with the retention time scaled by 10^{-2}. Signals which are further than 0.05 or with charge mismatch are discarded. Intensities of the rest are summed and returned as the observed value at a suitable node of the cleavage graph.

We are aware that there are many factors influencing the intensity measurements (Mallick et al., 2007), for instance, the isoelectric point. We leave integrating this knowledge into the model for the future, especially as we already have a component that accounts for inexact readings (generating y from x, cf. Fig. 3).

ACKNOWLEDGMENTS

The research described in this paper was partially supported by the Ministry of Science and Higher Education (grant PBZ-MNiI-2/1/2005). The computational resources were provided by the CoE BioExploratorium (project WKP 1/1.4.3/1/2004/44/44/115). Grateful acknowledgment is due to J. Ostrowski (Department of Gastroenterology, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology) and M. Dadlez (Mass Spectrometry Laboratory, Institute of Biochemistry and Biophysics, Polish Academy of Sciences) for providing the LC-MS/MS data.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Adam, B.L., Qu, Y., Davis, J.W., et al. 2002. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* 62, 3609–3614.

Aitchison, J., and Egozcue, J.J. 2005. Compositional data analysis: where are we and where should we be heading? *Math. Geol.* 37, 829–850.

Breiman, L. 2001. Random forests. *Mach. Learn.* 45, 5–32.

Ciliberto, A., Capuani, F., and Tyson, J.J. 2007. Modeling networks of coupled enzymatic reactions using the total quasi-steady state approximation. *PLoS Comput. Biol.* 3, e45.

Diamandis, E.P. 2003. Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin. Chem.* 49, 1272–1275.

Diamandis, E.P. 2004. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell. Proteomics* 3, 367–378.

Diamandis, E.P. 2006. Peptidomics for cancer diagnosis: present and future. *J. Proteome Res.* 5, 2079–2082.

Efron, B., and Tibshirani, R. 1997. Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.* 92, 548–560.

Gambin, A., Dutkowski, J., Karczmarski, J., et al. 2007. Automated reduction and interpretation of multidimensional mass spectra for analysis of complex peptide mixtures. *Int. J. Mass Spectrom.* 260, 20–30.

Geurts, P., Fillet, M., de Seny, D., et al. 2005. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics* 21, 3138–3145.

Hu, J., Coombes, K.R., Morris, J.S., et al. 2005. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief. Funct. Genomic Proteomic* 3, 322–331.

Jacobs, I.J., and Menon, U. 2004. Progress and challenges in screening for early detection of ovarian cancer. *Mol. Cell. Proteomics* 3, 355–366.

Koomen, J.M., Li, D., Xiao, L.-C., et al. 2005. Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. *J. Proteome Res.* 4, 972–981.

Leeman, M.F., Curran, S., and Murray, G.I. 2003. New insights into the roles of matrix metalloproteinases in colorectal cancer development and progression. *J. Pathol.* 201, 528–534.

Li, J., Zhang, Z., Rosenzweig, J., et al. 2002. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin. Chem.* 48, 1296–1304.

Lilien, R.H., Farid, H., and Donald, B.R. 2003. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J. Comput. Biol.* 10, 925–946.

Liotta, L.A., and Petricoin, E.F. 2006. Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. *J. Clin. Invest.* 116, 26–30.

Mallick, P., Schirle, M., Chen, S.S., et al. 2007. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* 25, 125–131.

Marshall, J., Kupchak, P., Zhu, W., et al. 2003. Processing of serum proteins underlies the mass spectral fingerprinting of myocardial infarction. *J. Proteome Res.* 2, 361–372.

Masaki, T., Matsuoka, H., Sugiyama, M., et al. 2001. Matrilysin (mmp-7) as a significant determinant of malignant potential of early invasive colorectal carcinomas. *Br. J. Cancer* 84, 1317–1321.

Petricoin, E.F., Ardekani, A.M., Hitt, B.A., et al. 2002. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359, 572–577.

Rawlings, N.D., and Barrett, A.J. 2000. Merops: the peptidase database. *Nucleic Acids Res.* 28, 323–325.

Reznik, S.E., and Fricker, L.D. 2001. Carboxypeptidases from a to z: implications in embryonic development and wnt binding. *Cell. Mol. Life Sci.* 58, 1790–1804.

Schölkopf, B., and Smola, A.J. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Tibshirani, R., Hastie, T., Narasimhan, B., et al. 2004. Sample classification from protein mass spectrometry, by "peak probability contrasts." *Bioinformatics* 20, 3034–3044.

Verrills, N.M. 2006. Clinical proteomics: present and future prospects. *Clin. Biochem. Rev.* 27, 99–116.

Villanueva, J., Martorella, A., Lawlor, K., et al. 2006a. Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age. *Mol. Cell. Proteomics* 5, 1840–1852.

Villanueva, J., Nazarian, A., Lawlor, K., et al. 2008. A sequence-specific exopeptidase activity test (sseat) for "functional" biomarker discovery. *Mol. Cell. Proteomics* 7, 509–518.

Villanueva, J., Shaffer, D.R., Philip, J., et al. 2006b. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Invest.* 116, 271–284.

Wu, B., Abbott, T., Fishman, D., et al. 2003. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19, 1636–1643.

Yu, J.S., Ongarello, S., Fiedler, R., et al. 2005. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics* 21, 2200–2209.

Address reprint requests to:
Bogusław Kluge
Institute of Informatics
University of Warsaw
02-097 Warsaw, Poland

E-mail: bogklug@mimuw.edu.pl

This article has been cited by:

1. 2009. Current literature in mass spectrometry. *Journal of Mass Spectrometry* **44**:8, 1262-1273. [CrossRef]