Supplementary materials[†]

Piotr Dittwald,[*,‡] J. Ostrowski,[¶,§] J. Karczmarski,[¶] and Anna Gambin[‡]

Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, Department of Oncological Genetics, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, 02-781 Warsaw, Poland, and Department of Gastroenterology, Medical Center for Postgraduate Education, 01-813 Warsaw

E-mail: piotr.dittwald@mimuw.edu.pl

Abstract

The supplementary materials for paper *Inferring serum proteolytic activity from LC-MS/MS data*. The electronic version of this document as well as some additional resources can be found on Web Supplement (`http://bioputer.mimuw.edu.pl/papers/proteolysis/`).

In this document we will use some notation introduced in the main paper.

Frequency and specificity matrices

First of all we build the matrix M^p with rows corresponding to amino acids, including empty position (set $\mathscr{I}$) and columns corresponding to loci around the cleavage point (denoted by $\mathscr{J}$). The value m^p_{ij} is the amount of amino acid i on position j summed for all cleavages made by peptidase p.

[†]Inferring serum proteolytic activity from LC-MS/MS data

[*]To whom correspondence should be addressed

[‡]Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

[¶]Department of Oncological Genetics, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, 02-781 Warsaw, Poland

[§]Department of Gastroenterology, Medical Center for Postgraduate Education, 01-813 Warsaw

The frequency matrix F^p (for $p \in \mathscr{P}$) is then defined by the formula:

$$f_{ij}^p = \frac{m_{ij}^p}{\sum_{k \in \mathscr{I}} m_{kj}^p}$$

More examples of frequency matrices are available on our Web Supplement.

Then we construct specificity matrix $S^p = (s^p)_{ij}$, where $(i, j) \in \mathscr{I} \times \mathscr{J}, p \in \mathscr{P}$ such that

$$s_{ij}^p = f_{ij}^p(\log_2(20) + \sum_{k \in I} f_{kj}^p \log_2 f_{kj}^p)$$

.

We used the concept of information theory described in.[1]

Pattern matrix

Let us construct the table $\mathscr{T}[0..2] = \{0.3, 0.1, 0.001\}$ with values of thresholds. Then we build the pattern matrix $Q^p = (q^p)_j$, where $j \in \mathscr{J}, p \in \mathscr{P}, q_j^p \subseteq \mathscr{I}$. For given $p \in \mathscr{P}$, $j \in \mathscr{J}$ and $i \in \mathscr{I}$ we assume that $i \in q_j^p$ iff the following condition is satisfied:

$$\exists_{0 \leq k \leq 2} s_{ij}^p \geq \mathscr{T}[k] \wedge \forall_{0 \leq k' < k} \forall_{i' \in \mathscr{I}} s_{i'j}^p < \mathscr{T}[k']$$

So the values of $\mathscr{T}$ are used to separate different classes of amino acid specificity on the given locus.

Affinity coefficients

Let us consider cleavage $v \dagger w$ made by peptidase $p \in \mathscr{P}$. Assume, that the cleavage point is surrounded by the sequence of amino acids (possibly with some empty positions at the ends) $a_{P4} \ldots a_{P1} a_{P1'} \ldots a_{P4'}$ and the cleavage point is between a_{P1} and $a_{P1'}$. If this sequence contains more than one empty place on any end then additional $-$'s are truncated, hence we obtain the subsequence $a_{Pk} \ldots a_{P1} a_{P1'} \ldots a_{Pl'}$ where $1 \leq k, l \leq 4$.

$$\bar{\rho}^p_{vw} = (\prod_{j \in \mathscr{J}'} f^p_{ja_j})^{\frac{8}{k+l}} \tag{1}$$

We use this kind of formula and not just the simple product in order to have similar sizes of coefficients' order for cleavages made by exo- and endopeptidases. The affinity coefficients ρ^p_{vw} are obtained from $\bar{\rho}^p_{vw}$ by normalization.

Cleavage detection

For the given $p \in \mathscr{P}$ and the following subsequence of polypeptide chain $a_{P4} \ldots a_{P1} a_{P1'} \ldots a_{P4'}$ s.t. $\forall_{j \in \mathscr{J}}\, a_j \in \mathscr{I}$ we assume that there is a cleavage between amino acids a_{P1} and $a_{P1'}$ iff

$$\forall_{j \in \mathscr{J}}\, a_j \in q^p_j \;\; \& \;\; \bar{\rho}^p_{vw} > 5 \times 10^{-10}$$

MS data analysis

The program `mz2m` [2] was run with the appropriate value of parameter *cut-param* (set to reduce the size of input data about 10 times).

The search is processed for each sequence charged by each of eight charges (values from 1 to 8) used by MS machine. The mass-to-charge ratio of the searched sequence is computed by the use of the following formula:

$$r(s) = \frac{m(s) + cm_\pi}{c} = \frac{m(s)}{c} + m_\pi \tag{2}$$

where

$$m(s) = m_{H2O} + \sum_a \#(a,s) m_a \tag{3}$$

Variables are described below:

- $r(s)$ - mass-to-charge ratio for sequence s,
- $\#(a,s)$ - how many times the amino acid a occurs in the sequence s,

- m_a - monoisotopic mass of the amino acid a,
- m_{H2O} - monoisotopic water mass,
- c - charge from set $\{1,\ldots,8\}$,
- m_π - monoisotopic proton mass that equals 1.0078250321 (1 Dalton); LC-MS machine charge sequences by adding proton.

The retention time is accessible only for some sequences. Thus we use this data to train linear regression model classifier[3] to predict retention time from amino acid composition assuming the following:

- the retention time depends on amino acids occurring in the polypeptide chain,
- only the amounts of amino acids are important, not their order in the sequence (i.e. we use the multiset of amino acids).

Graph pruning

We minimize our graph by the pruning process with the following steps:

1. we cut (recursively) roots, that were not identified in data set (and we also deleted elements of $\mathscr{P}$),
2. we cut (recursively) leaves, that were not identified in data set (with unused peptidases),
3. we delete roots that are also leaves (connected components that consist of one vertex).

Parameters normalization

By LMA we obtained vector $\hat{x} = ((\hat{\varphi}^\star_u)_{u\in\mathscr{S}}, (\hat{\varphi}^\perp_w)_{w\in\mathscr{L}}, (\hat{\lambda}_p)_{p\in\mathscr{P}}) = (\hat{x}_i)_{i=1\ldots m}$. We can see that for normalized vector x defined for $1 \leq i \leq m$ by the formula

$$x_i = \frac{\hat{x}_i}{\sum_{1 \leq i \leq m} \hat{x}_i}$$

the value of *rse* is the same as before normalization. Moreover $\sum_{1 \leq i \leq m} x_i = 1$, so vector x lies on simplex.

References

(1) Schneider, T. D.; Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **1990**, *18*, 6097–6100.

(2) Gambin, A.; Dutkowski, J.; Karczmarski, J.; Kluge, B.; Kowalczyk, K.; Ostrowski, J.; Poznański, J.; Tiuryn, J.; Bakun, M.; Dadlez, M. Automated reduction and interpretation of multidimensional mass spectra for analysis of complex peptide mixtures. *Int. J. Mass Spectrom.* **2007**, *260*, 20–30.

(3) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning*; Springer Verlag, 2001.