Metody obliczeniowe i modele stochastyczne w proteomice

streszczenie rozprawy doktorskiej

Bogusław Kluge

marzec 2011

Według (RNCOS, 2010) rynek bioinformatyki będzie rósł w latach 2011–2013 o 24% rocznie. Raport wskazuje, że proteomika wniesie znaczący wkład w ten wzrost, ze względu na coraz większe zainteresowanie leczeniem zindywidualizowanym.

Jedną z technologii, która umożliwia zgłębianie proteomiki jest *spektrometria mas* (MS). Dzięki niej możemy w pojedynczym eksperymencie wykonywać wyczerpujące analizy złożonych mieszanin zawierających tysiące molekuł. Typowy spektrometr mas jonizuje cząsteczki analizowanej mieszaniny, rozdziela je według stosunku masy do ładunku przy pomocy pola elektromagnetycznego, by w końcu zmierzyć ich liczbę specjalnym detektorem. Spektrometry mas różnią się implementacją powyższych trzech etapów (Fenn et al., 1989; March, 2000; Marshall et al., 1998). W wyniku takich eksperymentów powstają ogromne ilości danych, których nie można zinterpretować bez użycia specjalizowanych algorytmów wykonywanych na komputerze.

W kolejnych rozdziałach rozprawy poruszam zagadnienia związane z następującymi po sobie etapami analizy danych proteomicznych. Poczynając od algorytmów automatycznej interpretacji surowych danych spektrometrycznych, poprzez problem uliniawiania (ang. *alignment*) próbek, którego rozwiązanie umożliwia porównywanie danych z różnych eksperymentów, a kończąc na modelach stochastycznych opisujących proces trawienia białek w mieszaninach analizowanych z wykorzystaniem spektrometrii mas. Modele te po-

zwalają na sformułowanie interesujących z punktu widzenia biologii hipotez i wniosków.

Rozprawa zawiera wyniki dwojakiego rodzaju:

- Opracowane zostały metody obliczeniowe dla określonych problemów związanych ze spektrometrią mas. Metody te bazują w dużej mierze na schematach *Expectation Maximization*, *algorytmu Metropolisa-Hastingsa* oraz *programowania dynamicznego*.

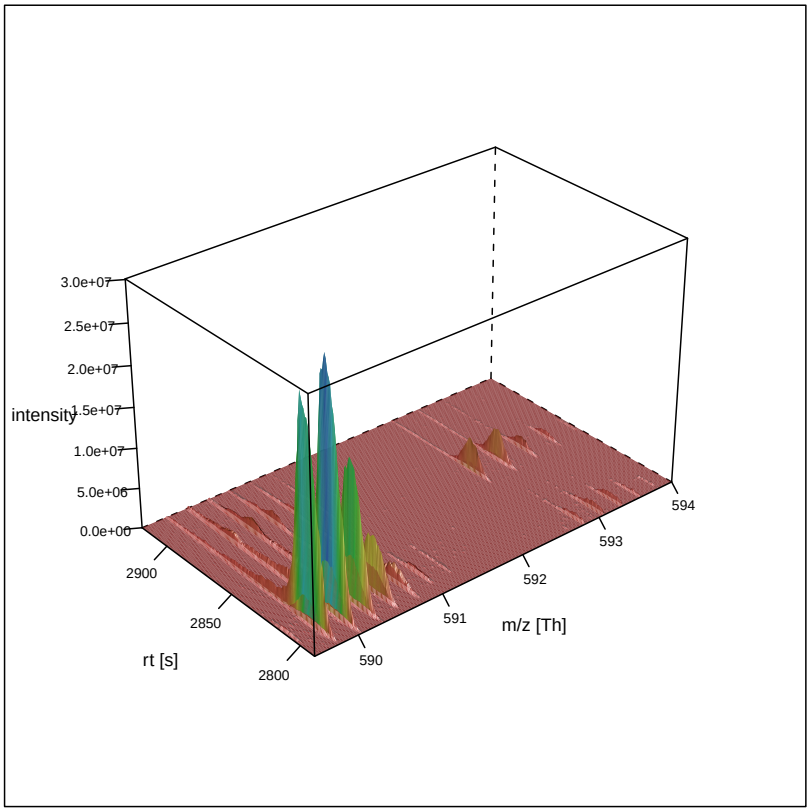
- Zaproponowany został stochastyczny model procesu proteolizy (trawienia peptydów), który opiera się na wiedzy z biologicznych baz danych i jest sformalizowany jako chemiczne równanie główne (ang. *chemical master equation*). Dane ze spektrometru mas służą jako dane wejściowe dla tego modelu.

Dodatkowo udało się uzyskać ciekawy rezultat dotyczący obliczania funkcji wykładniczej macierzy trójkątnych, który może być bezpośrednio przełożony na algorytm programowania dynamicznego.

Chciałbym podkreślić, że tego rodzaju interdyscyplinarne badania wymagają współpracy wielu specjalistów, od lekarzy, (bio)informatyków i (bio)statystyków mających doświadczenie z przetwarzaniem i analizą danych, po techników wykonujących „mokre" eksperymenty w laboratoriach. Podczas studiów doktoranckich miałem okazję współpracować z wiodącymi laboratoriami eksperymentalnymi — z grupą prof. Michała Dadleza z Instytutu Biochemii i Biofizyki Polskiej Akademii Nauk oraz z grupą grupa prof. Jerzego Ostrowskiego z Centrum Onkologii – Instytutu im. Marii Skłodowskiej-Curie w Warszawie.

Wstępne przetwarzanie danych spektrometrycznych

Pierwsza część rozprawy dotyczy wstępnego przetwarzania danych spektrometrycznych, na które składa się redukcja szumów oraz eliminacja redundancji. O surowym spektrum można myśleć jak o funkcji $\mathrm{R} \to \mathrm{R}$ lub $\mathrm{R}^2 \to \mathrm{R}$



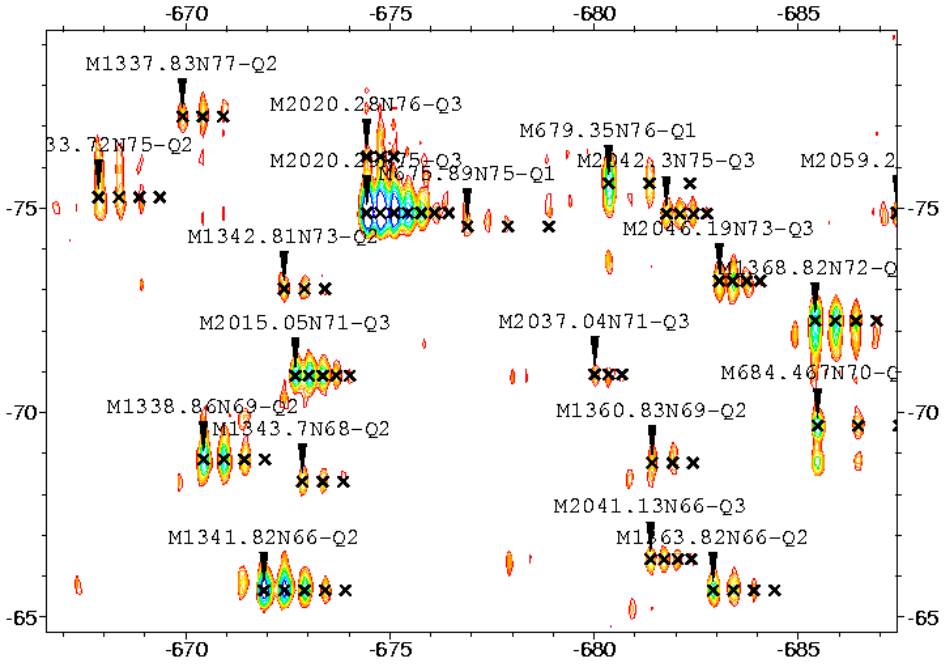
Rysunek 1: Niewielki fragment dwuwymiarowego spektrum LC-MS. Widoczne jest ok. 0.007% dziedziny. Wyraźne serie szczytów odpowiadają cząsteczkom o takim samym składzie chemicznym i ładunku, ale różnych wersjach izotopowych.

(Rys. 1), użyteczną informacją są jednak jedynie pozycje i wysokości wyraźnych lokalnych maksimów tej funkcji, zwanych *szczytami* (ang. *peak*). W rozprawie prezentuję wyniki dla dwuwymiarowych danych spetrometrycznych ($R^2 \to R$), czyli pochodzących ze spektrometru wyposażonego dodatkowo w kolumnę chromatografii cieczowej (ang. *liquid chromatography mass spectrometry* (LC-MS)). Taka kolumna umożliwia rozdzielenie skomplikowanej mieszaniny peptydów — są one stopniowo wypłukiwane odpowiednim rozpuszczalnikiem, co powoduje, że analizowane widmo zyskuje drugi wymiar (tzw. *czas retencji*).

Pierwszym krokiem przetwarzania jest utworzenie listy szczytów (ang. *peak picking*) i grupowanie szczytów odpowiadających cząsteczkom tej samej substancji (np. określonego peptydu), różniącym się ładunkiem lub wersją izotopową. Rozdział 2, oparty na pracy (Gambin et al., 2007), prezentuje procedury służące do rozwiązywania tych problemów. Rozpoczyna się wprowadzeniem do danych LC-MS, następnie opisuje wykorzystanie pakietu oprogramowania NMRPipe (Delaglio et al., 1995) stworzonego do przetwarzania danych NMR (ang. *Nuclear Magnetic Resonance*) do wykrywania szczytów. W kolejnej sekcji przedstawiam nasze podejście do problemu wykrywania obwiedni izotopowych (czyli grupowania szczytów z różnych wersji izotopowych takich samych cząsteczek), którego nowatorstwo polega na spojrzeniu na dwuwymiarowe spektrum (LC-MS) jako na całość. Jest ono istotnym uogólnieniem metody THRASH (Horn et al., 2000) zaprojektowanej dla danych jednowymiarowych. Oprócz prac przy projektowaniu schematu działania naszego rozwiązania (*metoda zamiatania*) mój wkład w głównej mierze polegał na stworzeniu algorytmu programowania dynamicznego, który proponuje składy aminokwasowe peptydów o zadanej masie. Warto podkreślić, że opisane w rozprawie podejście zostało zaimplementowane jako program o nazwie *mz2m* dostępny pod adresem `http://mz2m.sourceforge.net` i było wykorzystywane w Instytucie Biochemii i Biofizyki Polskiej Akademi Nauk. Rysunek 2 przedstawia fragment widma LC-MS zinterpretowany naszym algorytmem.

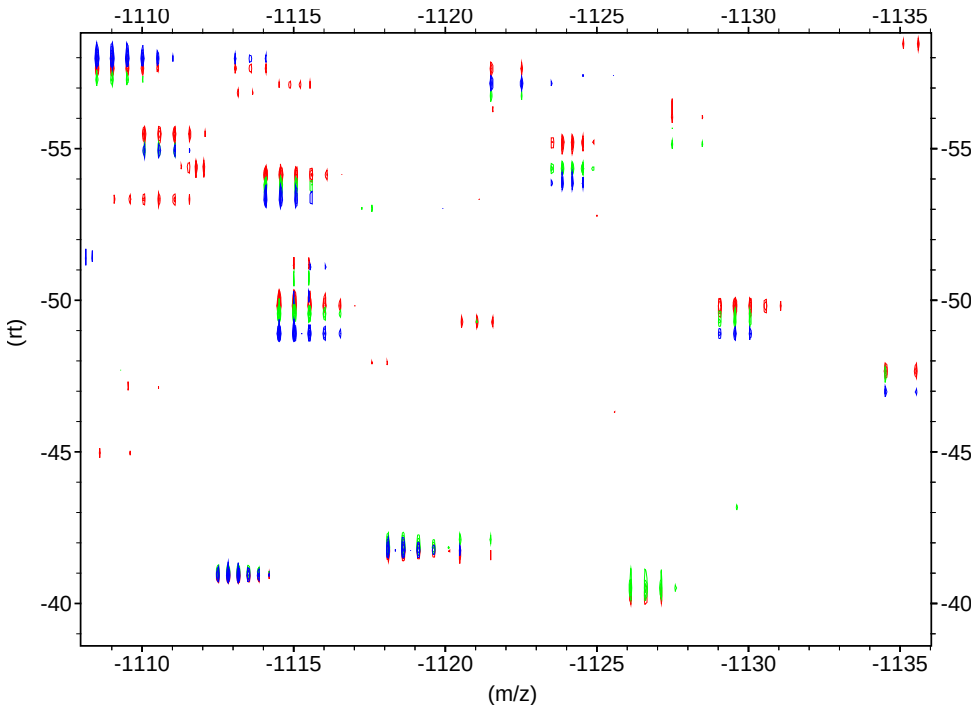
Uliniawianie danych spektrometrycznych

W kolejnym rozdziale rozprawy poruszam kluczowe dla zastosowań diagnostyki medycznej zagadnienie uliniawiania widm spektrometrycznych. Zbiory danych otrzymane w wyniku wstępnej interpretacji widm spektrometrycznych krwi pacjentów wykazują wiele różnic. Niektóre z nich są istotne, ponieważ wynikają np. z patologicznych procesów pozostawiających ślady w badanej mieszaninie peptydów, a inne wynikają z niedoskonałości technologicznych. Z tego powodu kolejnym etapem analizy zbiorów widm spektrometrycznych jest zidentyfikowanie sygnałów odpowiadających tym samym peptydom



Rysunek 2: Fragment spektrum z naniesionym wynikiem działania programu *mz2m*. Szczyty oznaczone są za pomocą czarnych krzyżyków. Małe strzałki wskazują szczyty monoizotopowe (odpowiadające najlżejszej wersji izotopowej związku). Przy każdej obwiedni widoczna jest masa monoizotopowa (M) oraz ładunek (Q).

w różnych próbkach (Rys. 3). W pracy (Lange et al., 2008) można znaleźć przegląd i porównanie takich procedur. Niektóre z nich omijają fazę przetwarzania wstępnego i działają bezpośrednio na surowych danych spektrometrycznych, np. próbując wyznaczyć transformację czasów retencji dzięki której dwa spektra dobrze do siebie pasują (Bylunda et al., 2002; Prince and Marcotte, 2006), większość jednak próbuje grupować dyskretne obiekty, takie jak szczyty (Lange et al., 2007; Li et al., 2005; Smith et al., 2006). Źródłem największych problemów przy uliniawianiu jest duża zmienność pomiędzy eksperymentami w wymiarze czasu retencji (związana z chromatografią cieczową). Prezentuję dwa podejścia do problemu uliniawiania widm spektrometrycznych: afiniczne transformacje czasów retencji oraz grupowanie



Rysunek 3: Kolorami oznaczone są trzy spektra powstałe w wyniku analizy tej samej mieszaniny peptydowej w różnych eksperymentach spektrometrycznych. Widoczne są grupy szczytów odpowiadające tym samym peptydom w różnych eksperymentach. Zaburzenia pomiędzy eksperymentami wzdłuż osi pionowej (czas retencji) są dużo bardziej znaczące niż wzdłuż osi poziomej (stosunek masy do ładunku).

z zastosowaniem mieszanek gaussowskich.

W pierwszej metodzie rozważam poprawianie czasów retencji za pomocą funkcji afinicznych $f_k(t) = a_k t + b_k$ indeksowanych uliniawianymi spektrami. Definiuję kryterium, które jest optymalizowane poprzez dobór wektorów parametrów a i b. Kryterium faworyzuje te przekształcenia, które sprawiają, że w bliskim sąsiedztwie szczytów znajdują się szczyty z innych spektr. Jako procedurę optymalizacyjną wykorzystuję algorytm Metropolisa-Hastingsa (Hastings, 1970). Metodę tą stworzyłem na potrzeby pracy (Kluge et al., 2009), ale jej szczegóły nie były nigdzie opublikowane.

Druga metoda uliniawiania wykorzystuje grupowanie oparte o model (ang. *model based clustering*). Takie podejście zostało po raz pierwszy zastosowa-

ne przez nas do danych proteomicznych (Gambin et al., 2006; Łuksza et al., 2009), co z powodu dużego rozmiaru zadania obliczeniowego było problemem istotnie trudniejszym niż spotykane w literaturze analizy danych transkryptomicznych (Yeung et al., 2001). Zastosowaliśmy dwa poziomy grupowania sygnałów. W pierwszej fazie algorytmem DBSCAN (Ester et al., 1996) grupujemy gęste obszary sygnałów na sumie widm z analizowanych eksperymentów. Następnie zgodnie z maksymą *dziel i zwyciężaj* rozwiązujemy mniejsze zadanie dla każdej grupy sygnałów osobno. W naszym podejściu zakładamy, że pozycje szczytów x analizowanych widm pochodzą z mieszaniny dwuwymiarowych rozkładów normalnych (każdy odpowiada populacji cząsteczek jednego rodzaju). Funkcja wiarygodności ma więc postać:

$$f(x \mid \tau, \mu, \Sigma) = \prod_{p \in P} \sum_{r=1}^{R} \tau_r \frac{1}{2\pi \left|\Sigma_r\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(x_p - \mu_r\right)^{\mathrm{T}} \Sigma_r^{-1} \left(x_p - \mu_r\right)\right),$$

gdzie P jest zbiorem wszystkich szczytów, zbiór R indeksuje komponenty mieszaniny. Za pomocą algorytmu Expectation-Maximization (Dempster et al., 1977; Minka, 1998) estymujemy[1] wartości oczekiwane komponentów μ, ich macierze kowariancji Σ oraz proporcje mieszaniny τ. Porównujemy 9 wariantów różniących się parametryzacją macierzy kowariancji oraz metodę z pakietu XCMS (Smith et al., 2006). Mój wkład polegał na implementacji niektórych z tych wariantów i pracach koncepcyjnych nad modelem.

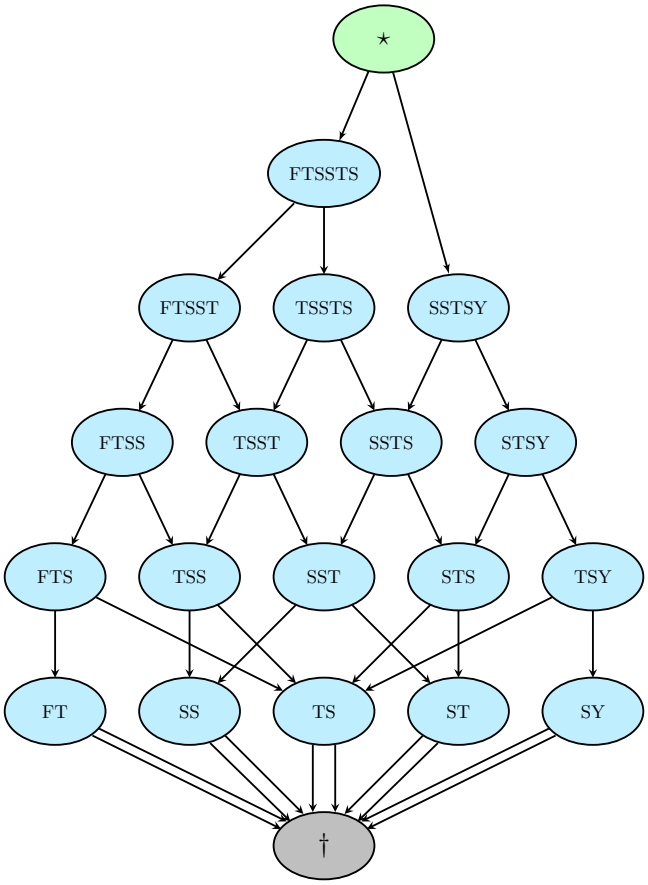
Ocena jakości uliniowień jest trudna, ponieważ nie dysponujemy wzorcowym, najlepszym rozwiązaniem. Oprócz porównania wizualnego stosujemy metodę *False Discovery Rates* (Benjamini and Hochberg, 1995) do oceny przydatności różnych uliniowień przy klasyfikacji pacjentów.

Modelowanie procesu proteolizy na podstawie danych spektrometrycznych

Ostatni rozdział skupia się na modelowaniu aktywności proteolitycznej. Jest to bardzo ważna tematyka m. in. ze względu na rolę procesu degradacji peptydów w chorobach nowotworowych (Villanueva et al., 2006). Diagnostyka

[1] estymator *maximum a posteriori* wynikający z zadania rozkładów a priori

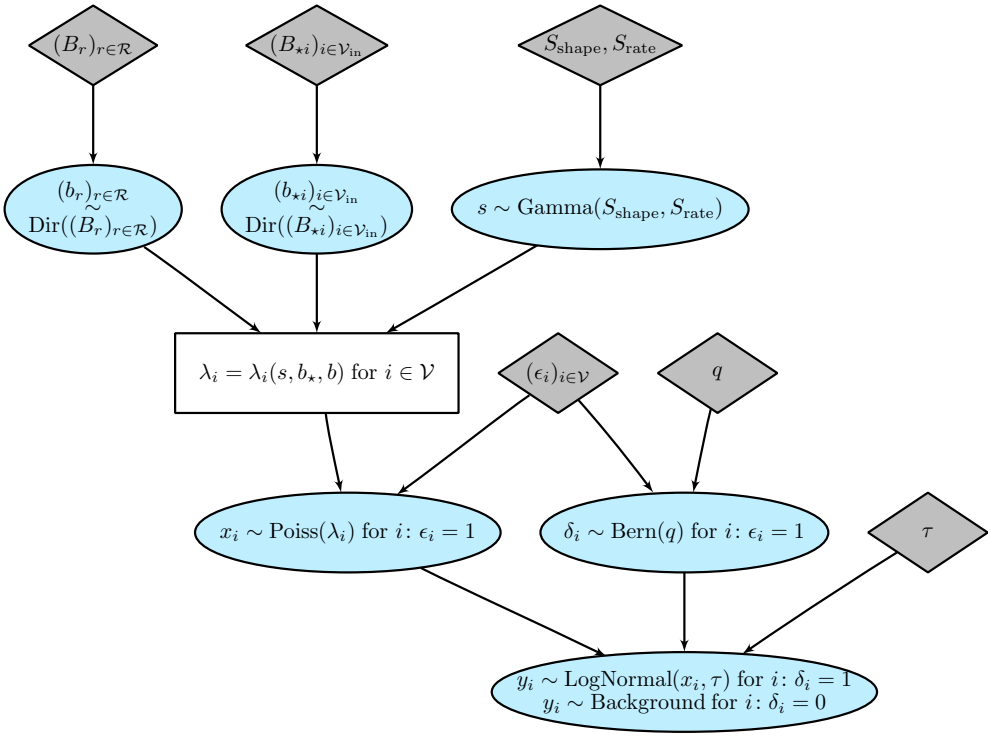
medyczna na podstawie obrazów spektrometrycznych krwi jest trudna, ponieważ obserwuje się dużą ich zmienność, związaną z proteolizą zachodzącą *ex-vivo*. Paradoksalnie, do badania stanu zdrowia można jednak wykorzystać różnice w samym procesie proteolizy. W tym celu zbudowaliśmy bayesowski model (Kluge et al., 2009) składający się z dwóch niezależnych komponentów, który formalizuje ten proces. Pierwszy komponent opisuje degradację łańcu-



Rysunek 4: Graf cięć dla przykładowych sekwencji aminokwasowych FTSSTS and SSTSY. Ponieważ ograniczamy się do działania egzopeptydaz, przedstawione są jedynie cięcia pojedynczych amiokwasów z końców sekwencji. Wierzchołek ⋆ zapewnia dopływ nowych cząsteczek do układu, natomiast wierzchołek † symbolizuje ostateczny produkt degradacji, czyli jedno-aminokwasowe sekwencje.

chów aminokwasowych (Rys. 4) na podstawie ich stężeń, drugi natomiast

pozyskiwanie informacji o stężeniach z danych spektrometrii mas. Struktura



Rysunek 5: Struktura modelu proteolizy i zbierania danych. Zmienne b są intensywnościami cięcia łańcuchów aminokwasowych. Z nich można wyznaczyć parametry λ rozkładów Poissona opisujących liczby cząsteczek poszczególnych peptydów. Przyjmujemy, że wielkości, które ostatecznie obserwujemy na wyjściu spektrometru (zmienne y) są zaszumionymi prawdziwymi liczbami cząsteczek lub pochodzą z rozkładu tła i nie mają z nimi nic wspólnego.

zależności zmiennych tego modelu jest pokazana na Rys. 5. Uzasadnieniem opisywania liczb cząsteczek poprzez rozkłady Poissona jest następujące twierdzenie.

Twierdzenie (Rozkład stacjonarny procesu cięcia)**.** *Rozkład stacjonarny* π *procesu* $(X(t))$ *określony jest wzorem:*

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

gdzie intensywności $(\lambda_i)_{i\in\mathcal{V}}$ *są jedynym rozwiązaniem następującego układu równań*[2]*:*

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

W celu wyestymowania parametrów modelu został zaprojektowany i zaimplementowany algorytm Metropolisa-Hastingsa próbkujący z rozkładu a posteriori. Po pomyślnym przejściu testów na sztucznie skonstruowanych zbiorach danych zastosowaliśmy go do danych rzeczywistych dotyczących raka jelita grubego.

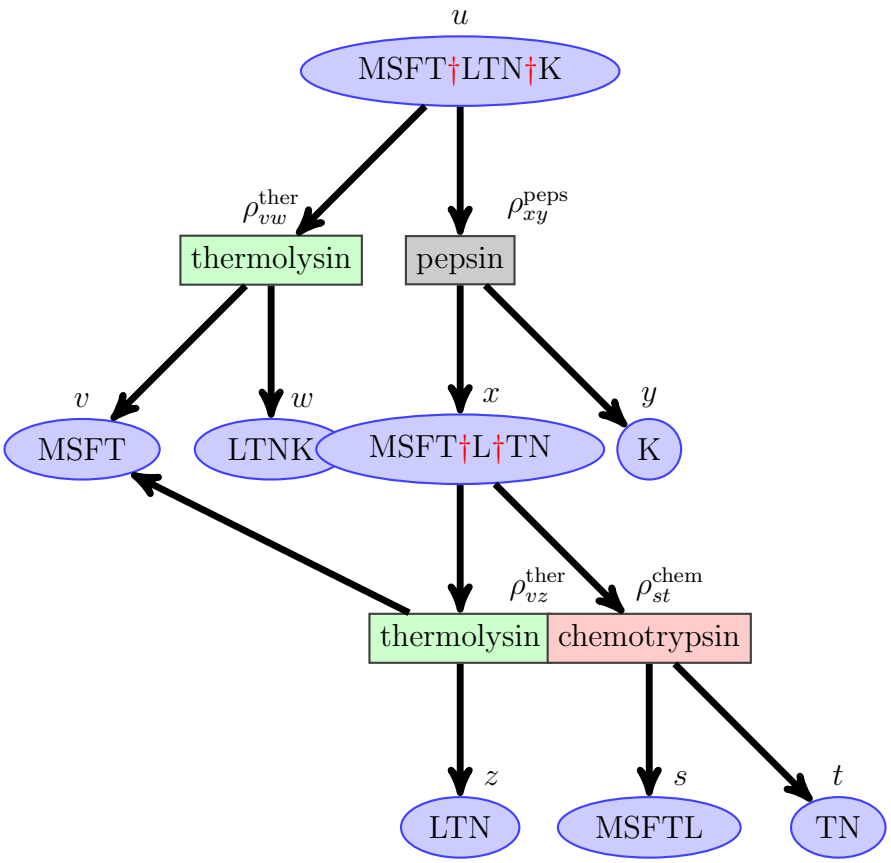
Druga część ostatniego rozdziału bazuje na pracy (Gambin and Kluge, 2010) i opisuje nową wersję modelu proteolizy, w której rozważamy cięcia w dowolnych miejscach łańcucha aminokwasowego, czyli zarówno cięcia egzopeptydazami jak i endopeptydazami (Rys. 6). Tym razem nie ograniczamy się do rozkładu stacjonarnego procesu proteolizy i opisujemy jego zachowanie w dowolnym punkcie czasowym (tym samym nie musimy zakładać stałego dopływu cząsteczek do układu, żeby rozważania asymptotyczne miały sens). Dodatkowo wykorzystujemy wzorce cięcia peptydaz z bazy danych MEROPS (Rawlings and Barrett, 2000). Uproszczeniu uległa natomiast procedura estymacji i model błędów — za pomocą algorytmu L-BFGS-B (Lu et al., 1994) rozwiązujemy problem nieliniowych najmniejszych kwadratów.

Niezbędnym składnikiem naszej procedury jest obliczanie funkcji wykładniczej macierzy trójkątnej w celu rozwiązywania liniowych układów równań różniczkowych. Podajemy rekurencyjne wzory na współczynniki macierzy wynikowej, które można bezpośrednio przełożyć na algorytm programowania dynamicznego. Mógłby on być szczególnie przydatny do obliczeń symbolicznych[3]. Nie spotkaliśmy w literaturze takiej charakteryzacji funkcji wykładniczej.

Miałem wkład we wszystkie etapy prac nad modelami proteolizy (uczestniczyłem w ich projektowaniu, implementowałem je i wykonywałem eksperymenty obliczeniowe). Podkreślam, że zgodnie z naszą wiedzą jest to pierwsza

[2]Parametry a są nieznormalizowanymi wersjami intensywności cięcia b z Rys. 5.

[3]W naszej implementacji używamy gotowej funkcji z jednej z bibliotek języka R (Team, 2009) wykorzystującej metody numeryczne.



Rysunek 6: Fragment grafu cięcia w rozszerzonym modelu, dla sekwencji aminokwasowej MSFTLTNK. Rozważamy cięcia w dowolnym miejscu sekwencji wykonywane przez różne peptydazy. Skłonności peptydaz do cięcia w określonych miejscach (parametry ρ) pochodzą z bazy danych MEROPS.

próba sformalizowania procesu proteolizy w postaci modelu przystosowanego do danych pochodzących ze spektrometru mas i że ma on potencjalne zastosowania w diagnostyce medycznej.

Literatura

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

D. Bylunda, R. Danielssona, G. Malmquistb, and K. E. Markides. Chromatographic alignment by warping and dynamic programming as a preprocessing tool for parafac modelling of liquid chromatography–mass spectrometry data. *Journal of Chromatography A*, 961(2):237–244, 2002.

F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRpipe: A multidimensional spectral processing system based on Unix pipes. *Journal of Biomolecular NMR*, 6:277–293, 1995.

A. P. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statististical Society*, Series B(39):1–38, 1977.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

A. Gambin and B. Kluge. Modeling proteolysis from mass spectrometry proteomic data. *Fundamenta Informaticae*, 103:89–104, 2010.

A. Gambin, M. Łuksza, B. Kluge, J. Ostrowski, and J. Karczmarski. Efficient model-based clustering for lc-ms data. *Workshop on Algorithms in Bioinformatics, LNBI*, 4175:32–43, 2006.

A. Gambin, J. Dutkowski, J. Karczmarski, B. Kluge, K. Kowalczyk, J. Ostrowski, J. Poznański, J. Tiuryn, M. Bakun, and M. Dadlez. Automated reduction and interpretation of multidimensional mass spectra for analysis of complex peptide mixtures. *International Journal of Mass Spectrometry*, 260:20–30, 2007.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

D. M. Horn, R. A. Zubarev, and F. W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11(4): 320–332, 2000.

B. Kluge, A. Gambin, and W. Niemiro. Modeling exopeptidase activity from lc-ms data. *Journal of Computational Biology*, 16(2):395–406, 2009.

E. Lange, C. Gröpl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber, and K. Reinert. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics*, 23(13):i273–i281, 2007.

E. Lange, R. Tautenhahn, S. Neumann, and C. Gröpl. Critical assessment of alignment procedures for lc-ms proteomics and metabolomics measurements. *BMC Bioinformatics*, 9:375, 2008.

X. J. Li, E. C. Kemp, H. Zhang, and R. Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular & Cellular Proteomics*, 4(9):1328–1340, 2005.

P. Lu, J. Nocedal, C. Zhu, and R. H. Byrd. A limited-memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, 1994.

M. Łuksza, B. Kluge, J. Ostrowski, J. Karczmarski, and A. Gambin. Two-stage model-based clustering for liquid chromatography mass spectrometry data analysis. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 15, 2009.

R. E. March. Quadrupole ion trap mass spectrometry: a view at the turn of the century. *International Journal of Mass Spectrometry*, 200(1–3):285–312, 2000.

A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews*, 17(1):1–35, 1998.

T. Minka. Expectation-maximization as lower bound maximization, 1998. URL http://research.microsoft.com/en-us/um/people/minka/papers/em.html.

J. T. Prince and E. M. Marcotte. Chromatographic alignment of esi-lc-ms proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry*, 78(17):6140–6152, 2006.

N. D. Rawlings and A. J. Barrett. Merops: the peptidase database. *Nucleic Acids Research*, 28(1):323–325, 2000.

RNCOS. Global bioinformatics market outlook. Research Report, 2010. URL http://www.rncos.com/Report/IM554.htm.

C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3): 779–787, 2006.

R. D. C. Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org.

J. Villanueva, D. R. Shaffer, J. Philip, C. A. Chaparro, H. Erdjument-Bromage, A. B. Olshen, M. Fleisher, H. Lilja, E. Brogi, J. Boyd, M. Sanchez-Carbayo, E. C. Holland, C. Cordon-Cardo, H. I. Scher, and P. Tempst. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *Journal of Clinical Investigation*, 116(1):271–284, 2006.

K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.