L'Inférence de Protéines à travers le Modèle Peptide Quantity Assignment

Emile Benoist, Guillaume Fertin, Géraldine Jean

L'Inférence de Protéines à travers le Modèle Peptide Quantity Assignment

Emile Benoist, Guillaume Fertin, Géraldine Jean

Université de Nantes, LS2N (UMR 6004), CNRS, Nantes, France
{emile.benoist,guillaume.fertin,geraldine.jean}@univ-nantes.fr

Mots-clés : *Spectrométrie de masse, peptides, protéines, modélisation, complexité algorithmique.*

1 Introduction

La *protéomique* est la discipline scientifique visant à identifier, caractériser, quantifier et comparer des protéines au sein d'échantillons. Ces protéines, qui jouent un rôle clé au sein de tout organisme, font de la protéomique un domaine de recherche très important.

Pour analyser un échantillon de protéines, la technique la plus utilisée est la *spectrométrie de masse*. Celle-ci a pour but de mesurer la masse de petites molécules. Dans un premier temps, les protéines d'un échantillon à analyser sont digérées afin de réduire leur taille. Une protéine est une longue molécule formée d'une séquence d'acides aminés et celle-ci, en étant digérée, se transforme en plusieurs petites chaînes d'acides aminés appelées *peptides*. Le résultat obtenu après digestion est donc un échantillon de peptides. Ces peptides peuvent ensuite être fournis au spectromètre de masse, dont les spectres qu'il fournit en sortie peuvent permettre d'identifier les peptides d'entrée, c'est-à-dire retrouver leurs séquences. Cette étape est en réalité bien plus complexe qu'elle ne l'est présentée ici et en général, seule une petite fraction des peptides réellement présents dans l'échantillon est correctement identifiée.

Vient ensuite la dernière étape : l'*inférence de protéines*. Cette étape consiste à retrouver, grâce aux peptides identifiés, quelles protéines étaient présentes dans l'échantillon d'origine. Ces protéines sont sélectionnées à partir d'une banque de données contenant un grand nombre de protéines, ainsi que les peptides que celles-ci engendreraient théoriquement suite à leur digestion. Notre travail vise à modéliser ce problème, et notamment à en étudier ses aspects algorithmiques (P, NP-difficile, FPT, etc.). Pour plus de détails sur la protéomique et la spectrométrie de masse dans leur ensemble, se référer à Ingvar et al. [2].

2 Le modèle Peptide Quantity Assignment et ses variantes

Nous introduisons ici notre première modélisation du problème d'inférence de protéines, que nous appelons Peptide Quantity Assignment (PQA). Celui-ci prend en compte chaque peptide identifié ainsi que le nombre de fois qu'il a été identifié. On appelle m le nombre de peptides identifiés et p_j, $\forall j \in [\![1, m]\!]$, le j-ème peptide identifié. Pour un peptide p_j donné, on appelle q_j le nombre de fois qu'il a été identifié (q_j correspond alors à la *quantité* de peptide p_j dans l'échantillon digéré). On nomme n le nombre de protéines candidates à être identifiées. Une protéine candidate est une protéine de la banque de données, composée d'au moins un peptide identifié. On nomme P_i, $\forall i \in [\![1, n]\!]$, la i-ème protéine candidate.

Le but est de déterminer, pour chaque peptide p_j, de quelle manière est distribuée sa quantité q_j à travers les différentes protéines possédant le peptide p_j (par exemple, un peptide présent dans les protéines P_1, P_2 et P_3 et ayant quantité 7 pourra être expliqué 3 fois par P_1, 4 fois par P_2 et 0 fois par P_3). On introduit également une contrainte imposant que la distribution des

Protéines \ Peptides	p_1	p_2	p_3	p_4	p_5
P_1	•	•		•	•
P_2		•	•		•
P_3	•	•	•	•	
P_4		•	•	•	
P_5	•			•	•
Quantités $\mathcal{Q}$	3	10	4	6	8

Protéines \ Peptides	p_1	p_2	p_3	p_4	p_5
P_1	3	3		3	0
P_2		0	0		8
P_3	0	4	4	0	
P_4		3	0	3	
P_5	0			0	0
Quantités $\mathcal{Q}$	3	10	4	6	8

TAB. 1 – (Gauche) Exemple d'instance de PQA, où les points noirs représentent l'appartenance d'un peptide à une protéine, et où la quantité de chaque peptide est fournie à la dernière ligne. (Droite) Une solution possible pour cette instance, dans laquelle les protéines P_1, P_2, P_3 et P_4 ont été identifiées (P_5 n'est pas utilisée).

quantités de chaque peptide au sein d'une même protéine doit être la même, ou bien égale à 0 (par exemple, une protéine possédant les peptides p_1, p_2, p_3 et p_4 pourra expliquer la présence de 5 peptides p_1, 5 peptides p_2, 0 peptide p_3, mais pas 6 peptides p_4 car $6 \neq 5$). Une fois qu'une distribution valide a été trouvée, chaque protéine expliquant la présence d'au moins un peptide est considérée comme identifiée, c'est-à-dire présente dans l'échantillon d'origine.

Nous considérons ensuite une variante de ce modèle en y ajoutant une contrainte forçant chaque peptide à n'être expliqué que par une seule protéine. Dans la Table 1, la solution ne respecte pas cette contrainte ; en effet, les peptides p_2 et p_4 sont chacun distribués respectivement sur 3 et 2 protéines, et non pas une seule. Nous appelons cette variante Exact Peptide Quantity Assignment (EPQA). Enfin, nous proposons une version "optimisation" du problème EPQA, que l'on appelle Max-EPQA, et dans laquelle certains peptides peuvent ne pas être expliqués, mais où l'objectif est d'en expliquer le plus possible.

	NP-*difficile*	λ-*approximable*	FPT
PQA	•	?	en n / en m
EPQA	•	?	en n / en m
Max-EPQA	•	$\lambda = 2$	en m

TAB. 2 – Résultats algorithmiques obtenus pour les 3 modèles présentés. Ici, m est le nombre de peptides identifiés et n le nombre de protéines candidates.

Pour chacun des trois problèmes PQA, EPQA et Max-EPQA, nous proposons une modélisation sous la forme d'un programme linéaire en nombre entiers (ILP). Par ailleurs, nous avons démontré un certain nombre d'autres résultats de complexité algorithmique concernant ces problèmes, qui sont résumés dans la Table 2.

Notons que d'autres modèles liés à l'inférence de protéines existent dans la littérature (voir par exemple [1]). Ils ont pour inconvénient de ne pas considérer leurs aspects algorithmiques tels que nous les présentons ici, mais leur avantage est d'être plus réalistes, car ils prennent plus de paramètres en considération. Notre but, suite à cette première modélisation, est de nous rapprocher de la réalité biologique pour obtenir des modèles plus précis, tout en accordant une grande importance à leur complexité. A terme, des tests sur des données (simulées et réelles) seront réalisés.

Références

[1] T. Huang, J. Wang, W. Yu, and Z. He. Protein inference : a review. *Oxford University Press*, 2012.

[2] E. Ingvar, F. Kristian, M. Lennart, and M. Svein-Ole. *Computational Methods for Mass Spectrometry Proteomics*. Wiley, 2008.