

L'étude du protéome constitue un moyen d'aborder la dynamique d'expression des produits du génome en permettant de repérer les produits des gènes qui sont exprimés à un moment donné dans un contexte donné.

Dans le cadre récent des études à grande échelle, le terme *protéomique* correspond à l'ensemble des techniques et outils permettant d'étudier le contenu en protéines d'un échantillon biologique ou de caractériser leurs interactions. Les informations fournies permettent d'appréhender la réponse cellulaire dans sa globalité et non plus de façon partielle⁶.

La grande diversité des techniques de la protéomique est liée à la poursuite d'objectifs différents comme de déterminer la quantité de protéines, la présence de modifications post-traductionnelles, la localisation subcellulaire, la structure tridimensionnelle, ainsi que les interactions protéine-protéine ou protéine-biomolécule. Le défi majeur de l'analyse protéomique est de fournir les outils capables de prendre en compte l'ensemble de ces informations afin de définir avec précision l'état d'un système biologique à un instant donné (figure 2.1).

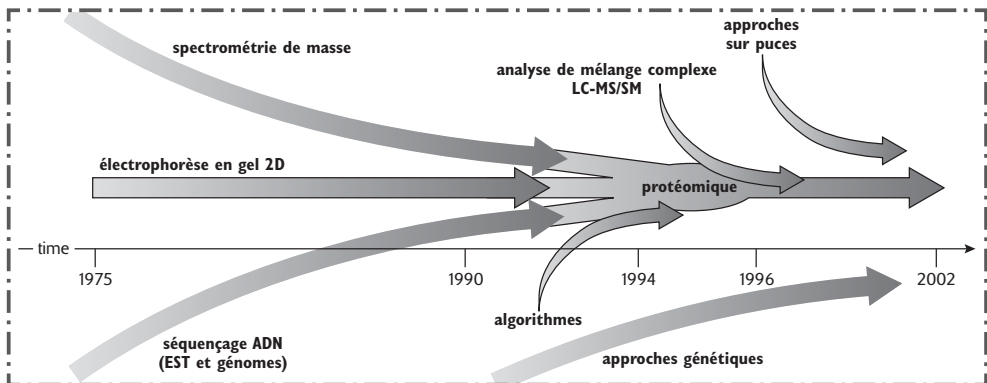
6. « Analyse protéomique, exploration cellulaire et annotation », J. Garin, *Biofutur*, 2002, hors-série 4, p. 28-38.

➤ Les outils de la protéomique : une grande diversité de techniques et de méthodes

▼ Figure 2.2

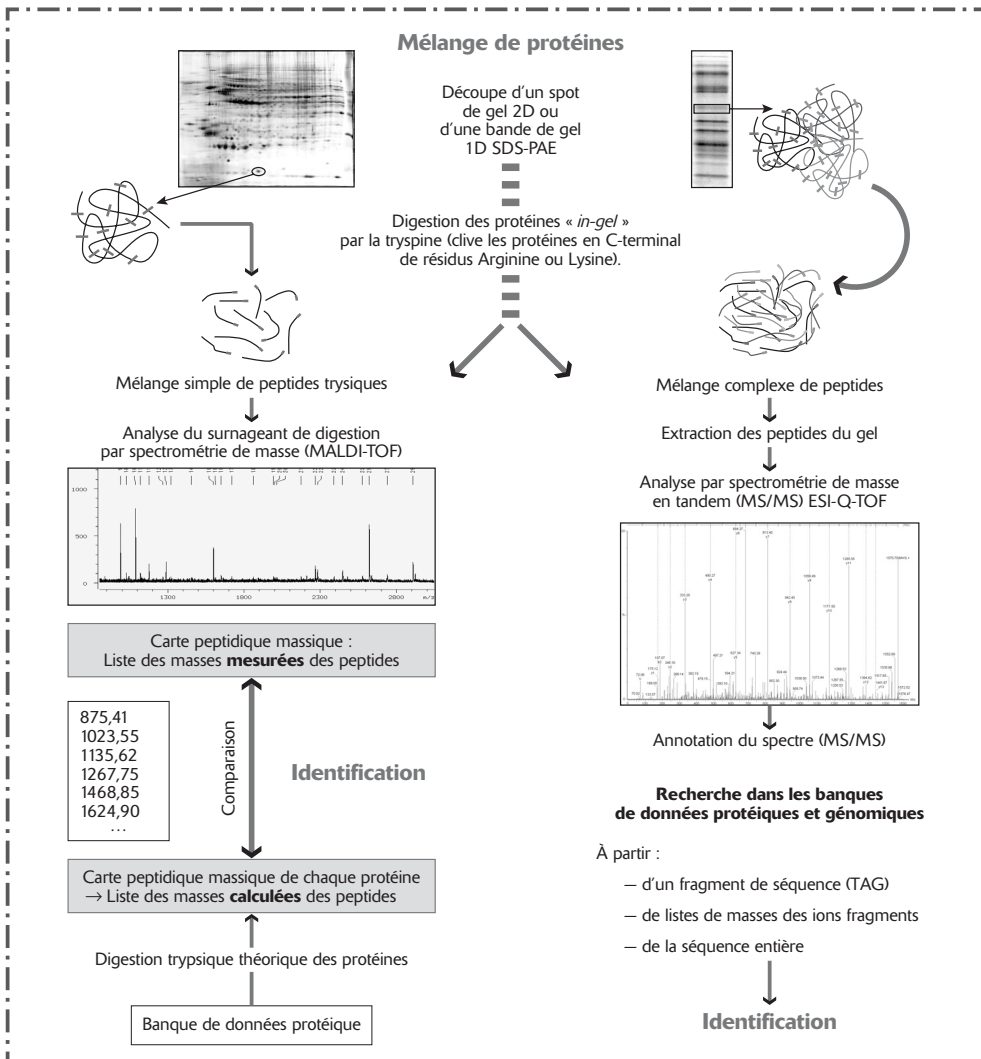
Illustration de la place de la protéomique au sein des différentes technologies et de leur complémentarité. D'après « Proteomics: evolution of the technology », S. D. Patterson, Biotechniques, septembre 2003, 35 (3), p. 440-444. (CL : chromatographie liquide; SM : spectrométrie de masse.)

Les défis méthodologiques et technologiques posés par la protéomique à haut-débit tiennent à la fois au très grand nombre de protéines à analyser, à l'hétérogénéité de leurs propriétés physico-chimiques, et au fait que de très nombreuses protéines sont faiblement exprimées. Comme présenté sur la figure 2.2, c'est la complémentarité des différentes technologies développées ces dernières décennies qui apporte une nouvelle dimension à la protéomique.



En effet, selon la question scientifique posée, par exemple, « Quelles protéines sont présentes dans un échantillon biologique? », « Quelles protéines sont différemment exprimées entre deux conditions données? », ou encore « Quelles protéines interagissent avec quelles autres? », les stratégies d'analyse vont être différentes. Le plus souvent, c'est la combinaison de plusieurs techniques, utilisées de façon séquentielle, qui va permettre de répondre à ces questions. En figure 2.3 est

Figure 2.3 ▾
 Déroulement d'une analyse protéomique classique telle qu'elle a été décrite par Shevchenko et al. en 1996. Les abréviations usuelles (en anglais) des différentes méthodologies sont indiquées entre parenthèses. D'après C. Ramus.



présentée une chaîne « classique » d'analyse protéomique à des fins d'identification des protéines contenues dans un échantillon biologique complexe. Partant d'un mélange de protéines, cette chaîne met en œuvre à la fois des méthodes séparatives (gels d'électrophorèse à une ou deux dimensions), un ou plusieurs spectromètres de masse permettant de mesurer finement la masse des composés et des outils et ressources bioinformatiques pour procéder à la comparaison des données de masse avec les données de séquences protéiques publiques. Cette première série d'approches vise à l'analyse des protéines et de leurs constituants, les peptides. En revanche, si l'on cherche à déterminer quelle protéine interagit avec quelle autre, on se tournera alors vers des techniques comme celle dite du « double-hybride » qui combine à la fois des outils de la biologie moléculaire et des outils bioinformatiques⁷.

7. « The continued evolution of two-hybrid screening approaches in yeast: how to outwit different preys with different baits », S.J. Fashena, *Gene*, 2000, 250 (1-2), p. 1-14.

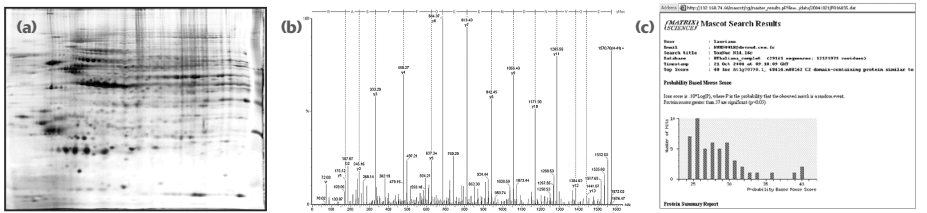
› Les enjeux et objectifs de la standardisation en protéomique

» Des données dispersées et hétérogènes

Les données associées aux articles scientifiques sont publiées sous différents formats, distribuées par différents moyens et hétérogènes

Données et métadonnées

Une distinction est opérée sur la nature de l'information produite par les expérimentateurs. Les données se définissent comme le résultat d'un dispositif expérimental d'acquisition et/ou calculatoire. À titre d'exemples : les images de gels 2D (a), les données brutes ou calculées issues d'un spectromètre de masse (b) ou encore le résultat d'une identification protéique par un logiciel de recherche dans les banques de séquences protéiques (c).



Exemple de type de données produites en analyse protéomique

Les métadonnées se définissent comme « les données au sujet des données », c'est-à-dire les informations associées à la production de ces données ; elles renvoient au contexte qui a permis de générer les données. À titre d'exemple, on citera les protocoles utilisés, les valeurs de paramétrage des appareillages ou d'un logiciel de calcul.

La déclaration du bureau d'édition de la revue *Molecular and Cellular Proteomics* (MCP)

"In the absence of accepted standards and widely available tools that operate on such standards, there are guidelines that the journal can formulate that would help ensure the publication of high-quality information and to assist readers in being able to make their own assessment of the validity of the assignments in manuscripts... Achieving the ultimate goal of publishing only high-quality datasets with small and known false-positive rates will require new data analysis tools and methods. As such tools become available, published results will be subject to reanalysis and interpretation. This is a healthy situation for the field, which MCP will assist by moving in the future to require that authors submit their data (in a suitable format) as a condition for acceptance of their manuscript."⁸

« En l'absence de standards acceptés et d'outils les manipulant largement disponibles, il y a des directives que le journal peut énoncer qui garantiraient la publication d'informations de haute qualité et autorisent aux lecteurs la capacité de valider en propre les informations consignées dans les manuscrits... Atteindre l'objectif ultime de ne publier uniquement que des données de haute qualité avec des taux d'erreurs infimes et connus nécessitera de nouvelles méthodes et outils d'analyse. Au fur et à mesure de la mise à disposition de tels outils, les résultats publiés pourront être soumis à de nouvelles analyses et à leur interprétation. Ceci constitue une situation saine pour le domaine dans lequel MCP participera dans le futur en exigeant que les auteurs soumettent leurs données (dans un format exploitable) comme condition pour l'acceptation de leur manuscrit. »⁸

en nature. Ceci peut aller de la publication des données sur le site web des auteurs jusqu'à des tables au format PDF comme matériel supplémentaire et qui ne sont pas systématiquement déposées dans des dépôts publics. Enfin, certains types de données sont produits sous des formats « propriétaires » (format exclusif utilisé par un logiciel, le plus souvent en format binaire) et ne sont exploitables qu'avec les logiciels fournis par les constructeurs d'instruments; c'est le cas des spectromètres de masse.

Concernant les bases de données de protéomique, on constate la même hétérogénéité qui rend la comparaison de données interbase très délicate. Ainsi, la diversité des formats à laquelle s'ajoute la difficulté de disposer de rapports détaillés, souvent incomplets, sur les données et les métadonnées limitent non seulement la validation d'une publication mais aussi la possibilité de regrouper et de procéder à des analyses comparatives d'expériences.

Alors qu'il est aujourd'hui aisé de procéder à la comparaison de tous les génomes publiés, il est encore très difficile de parcourir tous les protéomes humains publiés en fonction du tissu ou encore de l'état physiologique d'un tissu. Par exemple, à la proposition en apparence

8. « The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data », S. Carr et al., *Mol. Cell Proteomics*, 2004, 3 (6), p. 531-533.