

Objet et évolution méthodologique de l'analyse protéomique

Sandrine Thébault, Nadine Machour, Frédérique Perrot, Thierry Jouenne, Catherine Lange, Marie Hubert, Marc Fontaine, François Tron, Roland Charlionet

Le protéome désigne l'ensemble des produits des gènes exprimés par un organisme. C'est une entité variable, car les systèmes biologiques sont dynamiques et présentent des états différents selon leurs stades de différenciation, leurs réponses aux stimulus et leur transformation pathologique. L'analyse de l'ensemble des produits des gènes fait appel à deux approches : l'une consiste à mesurer le niveau quantitatif de chacun des ARNm du transcriptome d'une cellule soit par hybridation, soit par séquençage à haut débit ; l'autre, l'analyse protéomique, s'adresse directement aux protéines. Leur stratégie commune est d'étudier le contenu en protéines ou

en ARNm d'un organisme à un instant et dans un environnement donnés ; elle peut conduire à des analyses comparatives de plusieurs états. Ces deux approches sont complémentaires : la durée de vie des protéines et des ARNm n'étant pas identique, il n'y a pas de corrélation automatique entre leurs concentrations ; les modifications post-traductionnelles et les interactions protéiques ne sont accessibles que par l'analyse protéomique. Les techniques et les évolutions méthodologiques de l'analyse protéomique sont présentées ici, et nous engageons le débat sur ses domaines d'application.

Le protéome a été défini par Wilkins [1] comme « la totalité de la partie protéique exprimée dans une cellule, un tissu ou un organisme donnés ». Son étude, au niveau tant quantitatif que qualitatif, est complémentaire de celle du transcriptome [2]. Il s'agit en effet d'approches concernant des entités moléculaires, les protéines et les ARNm, certes liées dans le fonctionnement cellulaire, mais possédant des propriétés physico-chimiques totalement différentes, des durées de vie non corrélées [3] et jouant des rôles physiologiques tout à fait distincts. Les supports technologiques et méthodologiques qui fondent ces deux approches, sont également très différents. L'analyse protéomique, que nous présentons dans cette revue, comporte trois étapes [4, 5]. De manière classique, l'étape initiale

consiste à séparer les protéines contenues dans l'échantillon biologique étudié. Cette étape fait le plus souvent appel à l'électrophorèse bidimensionnelle (E-2D), technique séparative de très haut pouvoir de résolution, généralement préférée aux autres méthodes de séparation (électrophorèses mono dimensionnelles en gel ou en capillaire, chromatographies...). La deuxième étape est le traitement et la mise en image de la séparation protéique qui permet d'établir après E-2D une véritable carte protéique comme le montre la figure 1. Les techniques de traitement d'image permettent de déterminer les différences qualitatives et quantitatives des protéines quand deux échantillons biologiques sont comparés. Elles comportent éventuellement les méthodes de statistique multifactorielle qui permettent de prendre en compte et d'ana-

lyser les variations de très nombreuses taches protéiques. La troisième et dernière étape est l'identification et la caractérisation des protéines. Elle est le plus souvent réalisée grâce aux techniques de spectrométrie de masse. Les techniques utilisées au cours de ces trois étapes ont connu des progrès considérables sans lesquels les objectifs visés par l'analyse protéomique n'auraient pu être abordés avec succès. Cette évolution des technologies est loin d'être terminée. Aussi ferons-nous ici un état des lieux, en gardant à l'esprit que les méthodologies et les stratégies d'analyse connaîtront des modifications dans un avenir proche.

La séparation initiale

L'E-2D combine successivement deux séparations des protéines réalisées selon des critères indépendants,

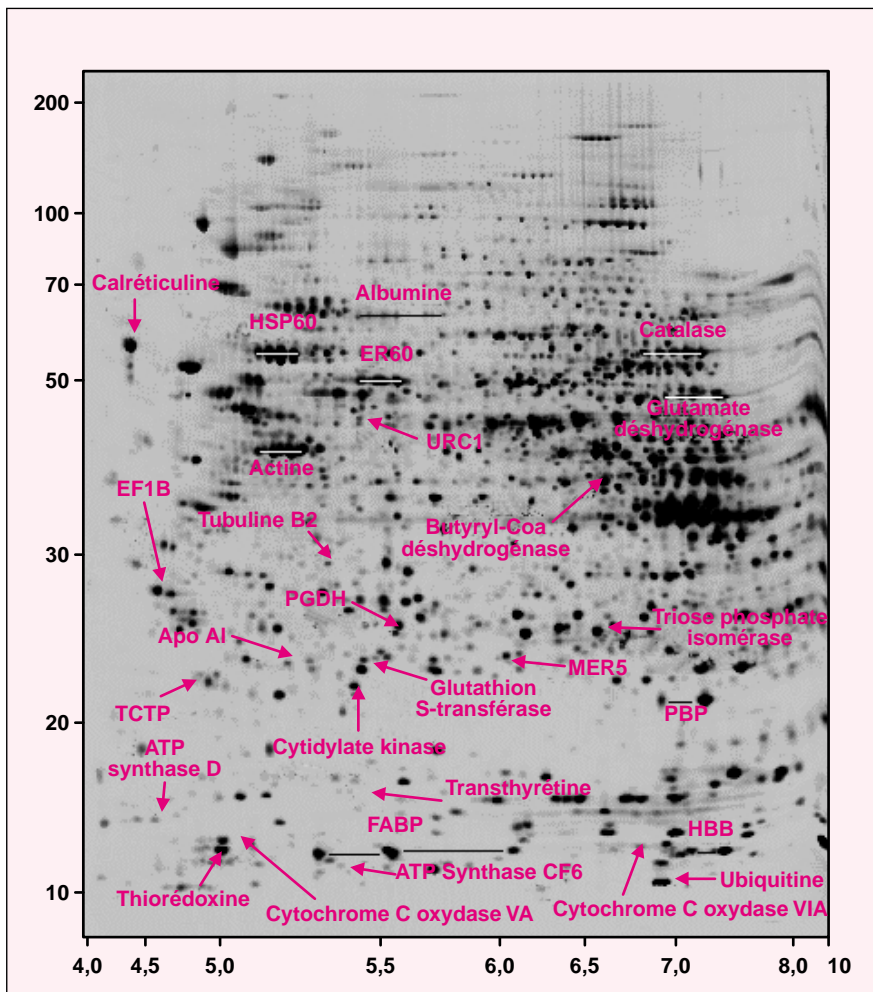


Figure 1. Carte protéique de la lignée cellulaire HepG2 obtenue par électrophorèse bidimensionnelle et colorée au nitrate d'argent. Les protéines sont séparées selon leur pI en abscisse et leur poids moléculaire en ordonnée. Cette carte protéique peut être trouvée sur la base de données SWISS-2DPAGE (<http://www.expasy.ch/ch2d/>). (copyright, Swiss Institute of Bioinformatics, Genève, Suisse).

le point isoélectrique (pI) puis la masse, ce qui lui confère un pouvoir de résolution très élevé. Elle nécessite au préalable une extraction et une solubilisation des protéines compatibles avec les conditions des expériences électrophorétiques. C'est une méthode de séparation fiable et reproductible, notamment depuis la commercialisation de gradients de pH préformés (*immobilized pH gradients*, IPG) [6] et de gradients de porosité [7]; ceux-ci sont respectivement utilisés pour la première dimension, c'est-à-dire la focalisation iso-

électrique (*isoelectric focusing*, IEF) et, pour la deuxième dimension, l'électrophorèse en gel de polyacrylamide en présence de sodium dodécyle sulfate (*sodium dodecyl sulfate-polyacrylamide gel electrophoresis*, SDS-PAGE). Des outils et des protocoles précis ont été développés à tous les stades de l'E-2D, qu'il s'agisse de la préparation et du dépôt de l'échantillon protéique, du passage de la première à la deuxième dimension, des milieux pour l'IEF et la SDS-PAGE, des gels de polyacrylamide présentant des gradients de pH préformés étroits ou

larges [8] et couvrant tous les domaines de pH (jusqu'à pH 12), des nouveaux appareils automatisés, etc. Les traitements subis par les protéines durant cette étape initiale de séparation sont compatibles avec la plupart des techniques de mise en image, d'identification et de caractérisation des protéines. Ainsi, l'E-2D peut-elle être considérée à la fois comme une méthode de séparation et de présentation des protéines. La spécificité de la présentation offerte par l'E-2D est double: d'une part, les concentrations relatives de chaque protéine séparée reflètent approximativement ce qu'elles étaient dans les cellules ou les tissus étudiés (nous serons amenés à préciser les limites de cette affirmation, voir plus loin); d'autre part, les protéines sont classées dans l'ordre croissant de leur point isoélectrique et de leur masse moléculaire (respectivement de gauche à droite et de bas en haut sur la carte protéique de la figure 1), ce qui fournit une première caractérisation des isoformes des protéines, complémentaire des renseignements structuraux, généralement plus précis, que l'on peut obtenir par spectrométrie de masse. Il faut souligner que cette présentation ordonnée sans *a priori* des protéines permet de détecter des modifications post-translationnelles qui ne sont pas soupçonnées au départ de l'analyse.

Révélation et analyse d'image

Les méthodes de mise en image sont maintenant bien définies, qu'il s'agisse de la coloration spécifique des protéines (bleu de Coomassie colloïdal, argent, etc.), de détection par fluorescence (Sypro orange, Sypro rubis, etc.) et par chimioluminescence, de révélation immunochimique (*Western blot*) ou de marquage radio-isotopique. Ces méthodes peuvent atteindre des sensibilités très importantes, en particulier lorsqu'elles sont couplées à certaines technologies de détection et de digitalisation (impression de plaques photographiques, Imager, caméra CCD ou densitomètre laser). Des progrès considérables ont été réalisés dans le traitement de l'image. Des logiciels performants permettent de

détecter les taches protéiques, de caractériser leur forme, leur surface et de mesurer leur densité optique, de préciser les coordonnées de point isoélectrique et de poids moléculaire, de comparer les cartes protéiques entre elles, de réaliser la synthèse de plusieurs cartes protéiques afin d'établir des cartes de référence, de calculer des histogrammes de densité optique pour chaque protéine ou groupe de protéines, ou encore de mettre en œuvre des protocoles d'analyse multifactorielle. Cependant, il faut remarquer que les méthodes de quantification sont généralement peu fiables. En effet, les marquages utilisés, quels qu'ils soient, ne traitent pas de manière égale toutes les protéines et, en outre, les seuils de saturation sont trop facilement atteints pour la plupart des appareils de détection, à l'exception des Imagers. Nous décrivons plus loin les nouveaux moyens mis en œuvre pour mesurer de manière précise les variations quantitatives des protéines.

Limites des deux premières étapes de l'analyse protéomique

Les deux premières étapes de l'analyse protéomique répondent-elles à l'objectif initial de description de l'ensemble protéique d'un organisme? Tout d'abord, le pouvoir de résolution de l'électrophorèse bidimensionnelle est-il suffisant pour séparer les dizaines de milliers de protéines qui composent le protéome d'un organisme? Théoriquement, la réponse est affirmative car on peut jouer sur l'étape de la focalisation isoélectrique de l'E-2D pour améliorer la séparation des protéines. Le pouvoir de résolution de la focalisation isoélectrique en gradient de pH préformé est en effet favorisé par un champ électrique élevé et des gradients de pH étroits [9]. En outre, on trouve dans le commerce d'une part, pour la première dimension, des appareils permettant d'atteindre des différences de potentiel de 10 000 volts et, d'autre part, toute une série de bandelettes de polyacrylamide présentant des gradients de pH préformés étroits et se chevau-

chant afin de couvrir toute la gamme de pH utile. Il en résulte que pour présenter, séparées les unes des autres, l'ensemble des protéines d'un protéome, il est nécessaire de réaliser non pas une, mais plusieurs cartes protéiques regroupant les protéines par domaine de points isoélectriques. Le caractère hydrophobe et insoluble de certaines protéines posent également question. Certes, l'E-2D est la méthode séparative la plus performante mais il reste vrai que, pour les protéines, la focalisation au point isoélectrique et le franchissement de plusieurs interfaces liquide/gel ou gel/liquide sont des événements qui favorisent leur précipitation. Pour y remédier, on fait appel à l'action synergique d'agents chaotropes (urée, thiourée), réducteurs (dithiothréitol, tributylphosphine) et de détergents neutres (CHAPS, sulfobetaine) [10]. Néanmoins, de nombreuses protéines, notamment membranaires, restent insolubles ou faiblement solubles, ce qui rend leur présentation par E-2D aléatoire voire impossible. Une extraction différentielle des protéines est alors nécessaire et il faut étudier séparément les protéines solubles, les protéines moyennement hydrophobes et les protéines très hydrophobes [11].

Une autre difficulté rencontrée est la grande variabilité d'expression des protéines [12]: certaines sont présentes en grande quantité dans les cellules ou les tissus, alors que d'autres, qui peuvent jouer un rôle clé dans certains processus, ne seront que très faiblement exprimées. Augmenter la concentration des protéines au-delà d'une certaine limite entraîne généralement des précipitations locales de protéines et des profils de cartes protéiques brouillés par des traînées, et ne constitue donc pas une solution. L'utilisation de gradients de pH étroits résout en partie ce problème car les protéines dont le pI se situe en dehors du gradient de pH sont automatiquement exclues du gel. Ces gradients de pH étroits sont beaucoup plus résolutifs, comme nous l'avons dits plus hauts, et autorisent l'emploi de concentrations protéiques beaucoup plus importantes, ce qui permet de faire apparaître des protéines faiblement

exprimées. Cette stratégie est encore plus efficace en réalisant un fractionnement préalable des protéines à l'aide d'un électrolyseur à compartiments multiples, séparés par des membranes isoélectriques (Isoprime) [13], qui permet d'utiliser des concentrations énormes de protéines sélectionnées dans des domaines étroits de pI. Il convient de signaler également que l'E-2D n'est pas bien adaptée aux protéines de poids moléculaire supérieur à 200 000 Da ou inférieur à 7 000 Da. Une autre limite de la technique séparative est liée précisément aux conditions dénaturantes, nécessaires à la solubilisation et à la bonne séparation des protéines, qui font obstacle à leur caractérisation fonctionnelle ultérieure.

Enfin, nous devons signaler que la plus grande difficulté de l'analyse protéomique réside actuellement dans l'analyse et le traitement informatisé des cartes protéiques. Malgré l'apparition de logiciels performants, la comparaison des cartes protéiques est difficilement réalisable lorsqu'on prend en compte un trop grand nombre de protéines [14]. Les problèmes de solubilité que nous venons d'évoquer, notamment, mais également les problèmes de distorsion que l'on observe au moment des séparations électrophorétiques, peuvent rendre aléatoire la détection de certaines protéines et entraînent de grandes difficultés pour mener à bien les opérations de paramétrage et d'appariements des taches protéiques sur les cartes que l'on doit comparer. Une étude récente a particulièrement bien mis en évidence cette difficulté [14]. Cependant de nouveaux traitements informatiques sont actuellement à l'étude [15]; ils comprennent des processus d'apprentissage qui intègrent les différents niveaux d'information: les spectres de masse pour l'identification des protéines (que nous abordons plus loin), les différents gels bidimensionnels à comparer pour la caractérisation et l'appariement des taches protéiques, la nature des échantillons protéiques, les principaux contaminants, etc. Cela permettra sans aucun doute dans un avenir proche de mener à bien l'analyse

d'un nombre plus élevé de protéines d'un organisme.

Différents travaux actuels n'étudient qu'un nombre limité (quelques centaines au plus) de protéines à la fois [12]. On peut trier les protéines d'intérêt après leur séparation électrophorétique et ne les révéler (parfois après transfert sur membrane) qu'au moyen de réactifs spécifiques (anticorps, ligand, marquage aux radioéléments, etc.). On peut également mettre en place un tri sélectif préalablement à leur séparation. Dans la plupart des cas, cela s'effectue selon des critères biologiques : préparation des protéines à partir d'organites cellulaires isolés (membrane plasmique, appareil de Golgi, noyau, mitochondrie, etc.) ; sélection des protéines interagissant avec un ligand (protéines associées, immunoréactives, etc.). Ailleurs, ce sont des critères physiques qui sont utilisés pour sélectionner les protéines : principalement la solubilité ou le point isoélectrique des protéines grâce aux gradients de pH modulables utilisés pour la focalisation isoélectrique (comme cela a été décrit plus haut). Lorsque ce tri préalable – éventuellement très restrictif – est effectué, il est possible alors de réaliser la séparation et la présentation des protéines avec une technique séparative de pouvoir de résolution plus faible que l'E-2D (par exemple, SDS-PAGE, électrophorèse non-dénaturante, électrophorèse capillaire, chromatographie capillaire, etc.).

Identification et caractérisation des protéines

On ne peut parler d'analyse protéomique sans évoquer les procédés d'identification des protéines. En effet, identifier une tache protéique permet de l'associer à une famille de protéines, de connaître sa localisation cellulaire et son lieu de synthèse ou de l'inclure dans une chaîne réactionnelle ou un mécanisme physiopathologique et donc déjà de lui attribuer une (ou des) fonction(s). C'est cette étape qui permet de donner une interprétation biologique des modifications observées dans un échantillon biologique par comparai-

son des cartes protéiques. De même, l'analyse protéomique doit comporter des étapes de caractérisation structurale des protéines. En effet, la plupart des protéines subissent après leur synthèse des modifications au niveau de leur structure (protéolyse, conformation, action sur divers substrats, etc.) ou au niveau de leurs acides aminés. Ces modifications peuvent être permanentes ; elles confèrent alors aux protéines une structure tridimensionnelle stable ou leur permettent d'être dirigées vers un compartiment cellulaire distinct. Elles peuvent être transitoires et caractériser ainsi différents stades de l'activité ou de la fonctionnalité des protéines. Bon nombre de ces modifications sont connues : formation de ponts disulfures, désamination des asparagines ou glycines, méthylation, acétylation, hydroxylation, oxydation des méthionines, formylation, phosphorylation, sulfatation, glycosylation, farnesylation, myristoylation, etc. Une autre source de variabilité des protéines concerne les mutations ponctuelles, c'est-à-dire les échanges d'acides aminés liés au polymorphisme des protéines ou introduits artificiellement dans les protéines recombinantes. Un certain nombre des modifications post-traductionnelles ou des mutations ponctuelles peuvent être décelées sur les cartes protéiques après E-2D puisqu'elles entraînent des changements de la charge et/ou de la masse des protéines. Mais l'identification des protéines ainsi que la caractérisation fine de leurs modifications post-traductionnelles sont devenues possibles à grande échelle grâce à la précision, la sensibilité et le pouvoir de résolution des nouvelles techniques de spectrométrie de masse.

Application de la spectrométrie de masse aux protéines

Jusqu'à ces dernières années, l'étape de l'identification des protéines était difficile et faisait appel à des techniques d'application limitée (immunodétection, co-migration avec des protéines isolées) ou lourdes à mettre en œuvre (composition en acides aminés, microséquençage). Aujourd'hui,

la mise au point de techniques d'identification à grande échelle a transformé cette étape. Deux avancées sont à la base de cette transformation. La première est le séquençage du génome, déjà réalisé pour certains organismes mono- ou multicellulaires, dont la généralisation à l'ensemble du monde vivant permet le développement concomitant des bases de données protéiques et des outils informatiques nécessaires à leur exploitation. La deuxième avancée est l'adaptation de la spectrométrie de masse à l'analyse de polymères biologiques (protéines, acides nucléiques).

La spectrométrie de masse est capable de déterminer la masse des peptides ou des protéines de manière très précise, jusqu'au dixième de Da, grâce notamment au développement de techniques d'ionisation douce et à l'ensemble des progrès techniques intéressant tous les compartiments de l'instrumentation : source d'ionisation, analyseur, cellule de détection, système de pompe à vide (la *figure 2* donne le schéma de fonctionnement d'un spectromètre de masse). Les spectromètres de masse, qui analysent les protéines à partir des cartes protéiques en gel ou après transfert sur membrane, fonctionnent grâce au système d'ionisation MALDI (*matrix assisted-laser desorption/ionisation*) [16] : la protéine, co-cristallisée avec une matrice absorbant les photons (par exemple, l'acide sinapinique ou l'acide α -cyano-4-hydroxycinnamique), est bombardée avec un rayon laser ; la matrice transforme l'énergie transportée par le laser en énergie d'excitation pour l'échantillon et ce processus s'accompagne notamment d'un mécanisme de transfert du proton ; la protéine donne ainsi naissance à un ion gazeux monochargé $(M+H)^+$ ou $(M-H)^-$ mais il peut apparaître également, en faible quantité, des ions dimériques ou doublement chargés ou des adduits dus aux contaminations affectant l'échantillon. La spectrométrie de masse peut être également couplée aux séparations capillaires (HPLC ou électrophorèse capillaire), notamment en utilisant les ionisations électrospray (ESI) [17] ou nano-électrospray : l'échantillon protéique en

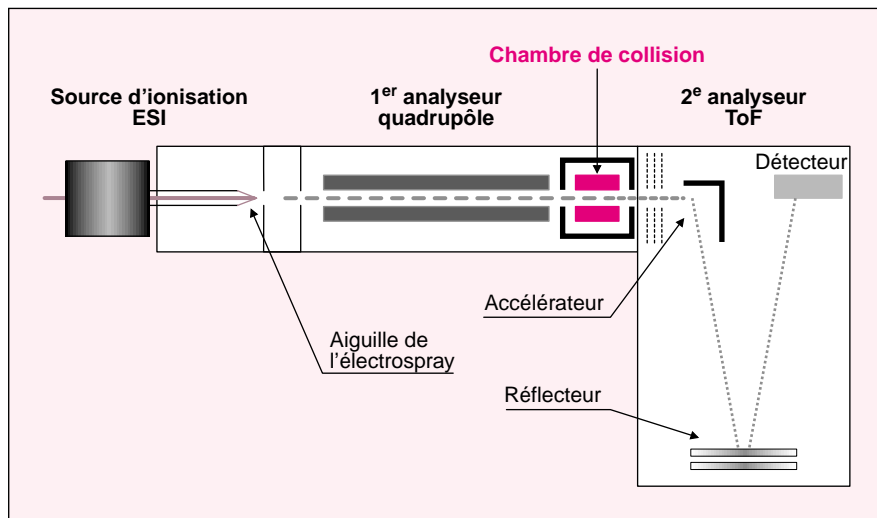


Figure 2. **Spectromètre de masse en tandem SM/SM (Q-ToF).** On distingue la source d'ionisation ESI (electrospray ionisation) où les peptides sont transformés en ion gazeux, le premier analyseur quadripôle qui permet de les séparer selon la masse et de sélectionner l'ion à fragmenter, la chambre de collision dans laquelle le peptide sélectionné est fragmenté et le deuxième analyseur Temps de Vol (ToF) qui donne accès à sa séquence en acides aminés.

solution qui arrive à l'embout du capillaire est soumise à un champ électrique intense qui entraîne sa pulvérisation en gouttelettes de plus en plus fines jusqu'à l'obtention d'ions protéiques multichargés. Ces méthodes d'ionisation douce ne fragmentent pas les échantillons protéiques et permettent donc de les étudier dans leur intégralité.

On appelle *analyseur* la partie du spectromètre de masse qui est responsable de la séparation des ions gazeux formés dans la source d'ionisation. Les analyseurs sont le plus souvent de type temps-de-vol (*time-of-flight*, ToF), quadripôle, secteur-magnétique, ou trappe à ion. Chacun de ces types sépare les ions sur la base de leur rapport masse sur charge (m/z). Les spectromètres de masse peuvent avoir également plusieurs analyseurs séparés par une chambre de fragmentation (*collision-induced dissociation*, CID) : il s'agit alors de combinaison en tandem d'analyseurs (*SM/SM*) qui peuvent être de types différents (figure 2). Lorsque les ions peptidiques sélectionnés par le premier analyseur arrivent dans la chambre de fragmentation, ils entrent en collision avec des molécules de gaz rares et se fragmentent principalement au niveau des

liaisons peptidiques (mais il peut y avoir accessoirement d'autres sites de coupures) ; les ions engendrés sont alors étudiés dans le deuxième analyseur. Les spectromètres de masse trappe à ion permettent de réaliser plusieurs cycles de sélection d'ions peptidiques/fragmentation/analyse des ions fragmentés et il est alors possible d'accéder à des expériences SM^2 , SM^3 ou SM^4 qui correspondent respectivement à des combinaisons SM/SM , $SM/SM/SM$ ou $SM/SM/SM/SM$. Les appareils les plus souvent utilisés sont le MALDI-ToF pour l'identification des protéines, les tandem SM/SM avec notamment la *quadrupole-orthogonal time of flight* (Q-ToF) et la trappe à ion pour le séquençage des peptides et leur caractérisation structurale.

La stratégie générale d'identification des taches protéiques est décrite dans la figure 3. En résumé, elle consiste à analyser par spectrométrie de masse les fragments peptidiques obtenus après excision d'une tache protéique et sa digestion par une enzyme protéolytique (généralement la trypsine). Le spectre de masse obtenu par MALDI-ToF est considéré comme l'empreinte digitale de la protéine. Il est comparé, en utilisant les outils informatiques disponibles

sur internet (par exemple MS-FIT), aux spectres de masse virtuels des protéines dont le gène est caractérisé et qui sont rassemblées dans les banques de données spécialisées (par exemple, SWISS-PROT) [18]. Cette comparaison est associée à un score fondé sur le nombre des fragments protéolytiques obtenus dont les masses peuvent correspondre à celle des fragments protéolytiques des banques de données, et par la précision avec laquelle les masses sont déterminées. Les meilleurs scores correspondent aux comparaisons les plus fiables. Pour être univoque, l'identification doit parfois comporter une recherche complémentaire soit portant sur le spectre lui-même (par exemple, recherche de doublets arginine ou lysine – dans le cas de digestion trypsique, etc.) soit en réalisant le séquençage d'un des peptides par SM/SM [19]. Quel que soit le mode de spectromètre de masse utilisé, les modifications post-traductionnelles doivent être prises en compte, mais pour détailler la nature précise et la localisation de ces modifications il faut entreprendre une étude avec les tandem SM/SM .

Place de la spectrométrie de masse dans l'évolution prévisible des techniques de l'analyse protéomique

L'identification des protéines par les techniques de spectrométrie de masse et notamment par les appareils en tandem SM/SM est devenue une opération rapide et, avec l'informatique adéquate, automatisée et applicable sur une grande échelle. Cependant comme l'efficacité d'ionisation, quel que soit le mode MALDI ou ESI choisi, est totalement variable pour les protéines et leurs fragments peptidiques, la spectrométrie de masse ne peut pas fournir directement des informations quantitatives absolues. Toutefois, l'expression différentielle des protéines peut être quantifiée à condition d'inclure dans les échantillons protéiques des isotopes lourds stables comme étalon interne (2H , ^{13}C , ^{15}N , etc.). Parmi les nombreuses tentatives récentes dans ce sens, nous décrivons (figure 4) celle de l'équipe

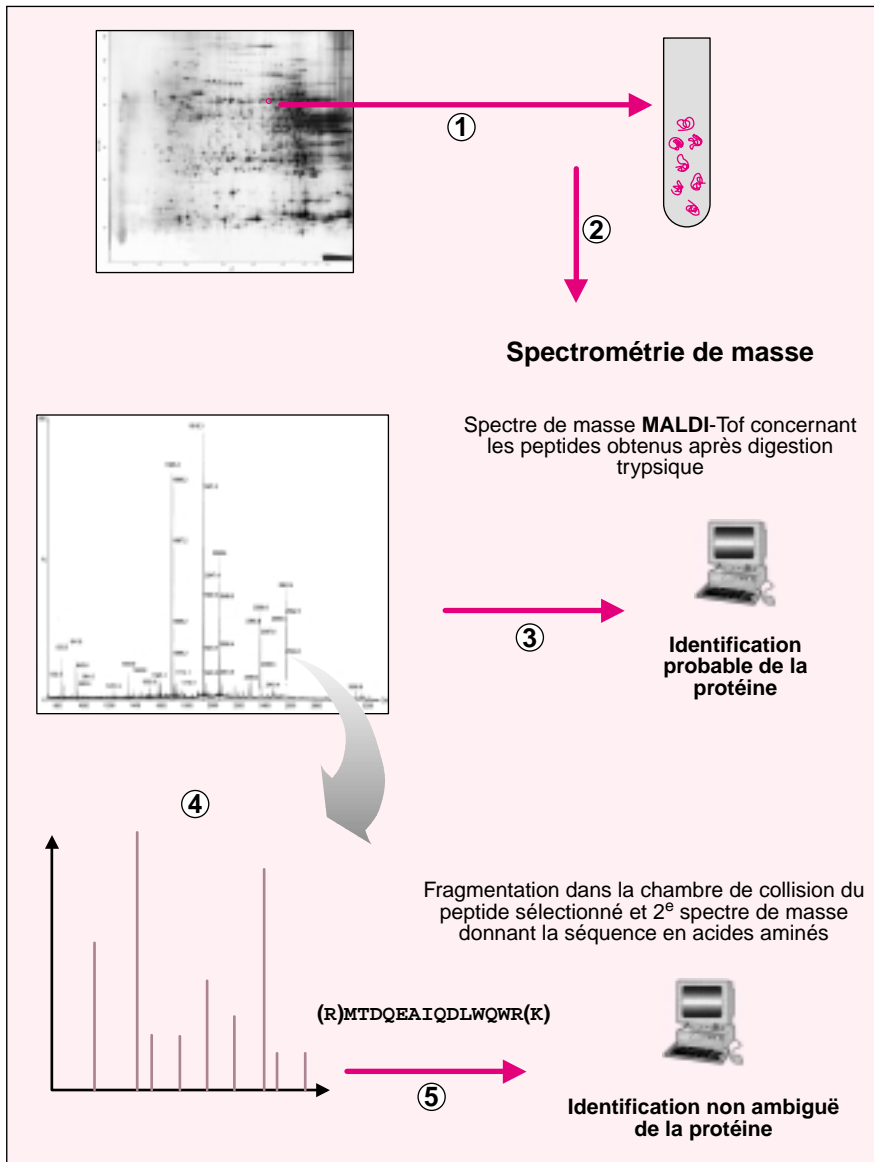


Figure 3. **Présentation schématique des principales étapes de l'identification par spectrométrie de masse des protéines séparées après électrophorèse bidimensionnelle (analyse protéomique classique).** Des extraits cellulaires sont préparés, les protéines sont solubilisées et séparées par E-2D. 1. Découpage de la tache protéique, digestion trypsique dans le gel, extraction des peptides tryptiques, dessalage. 2. Analyse par un premier spectromètre de masse. 3. L'identification se fait par comparaison du spectre de masse réel obtenu avec les spectres virtuels de l'ensemble des protéines trypsinisées des bases de données. Les protéines sont classées dans l'ordre des meilleures correspondances. 4. Pour une identification non ambiguë un peptide est sélectionné puis fragmenté dans une chambre de collision et les fragments sont analysés dans un deuxième spectromètre de masse donnant accès au séquençage en acides aminés de ce peptide. 5. La comparaison de cette séquence avec celles des fragments tryptiques de protéines répertoriées dans les bases de données identifie de manière univoque la protéine.

d'Aebersold [20]. Un composé ICAT (*isotope-coded affinity tag*) a été synthétisé qui comporte trois éléments: un

pôle réactif se liant spécifiquement aux cystéines et qui va permettre d'accrocher l'ICAT sur les protéines;

une étiquette biotine permettant par chromatographie d'affinité de récupérer les fragments protéiques ayant fixé l'ICAT; un support qui peut incorporer huit isotopes stables (hydrogène ou deutérium). Les formes deutérée et hydrogénée de l'ICAT ont une différence de masse de 8 Da, parfaitement détectable par spectrométrie de masse. Les protéines de deux états cellulaires différents sont recueillies, dénaturées, réduites et marquées respectivement avec l'ICAT léger et lourd; les deux échantillons sont alors mélangés et trypsinisés; les peptides possédant une cystéine (liée à un ICAT) sont isolés par chromatographie d'affinité puis analysés par HPLC couplée en ligne à un spectromètre de masse tandem SM/SM. Les peptides sont donc détectés par paires, chaque couple de pics ayant une différence de masse caractéristique des deux isotopes stables. Le rapport d'intensité des ions appariés quantifie l'abondance relative des protéines dont ils sont issus. En même temps, le spectre de masse SM/SM révèle la séquence du peptide et identifie sans ambiguïté la protéine. Ainsi, cette nouvelle stratégie appliquée à l'analyse protéomique, permet-elle à la fois d'identifier les protéines et de quantifier leur expression différentielle: elle est désignée sous le nom d'*analyse protéomique quantitative (quantitative proteome analysis-QPA)*. Il faut signaler qu'une autre méthode d'analyse protéomique quantitative permettant également de quantifier comparativement les protéines et de les identifier a été développée [21]: mais elle se réalise à partir des cartes protéiques bidimensionnelles et elle utilise un double marquage radioisotopique lié à l'analyse de la composition en acides aminés.

L'analyse protéomique quantitative, telle que nous venons de la décrire, présente des différences notables avec l'analyse protéomique classique dont nous avons rappelé les principales étapes plus haut. Notamment les étapes de séparation et présentation des protéines suivies de la mise en image et de l'analyse des cartes protéiques, qui sont celles consommant le plus de temps et présentant le plus de limitations par rapport à la

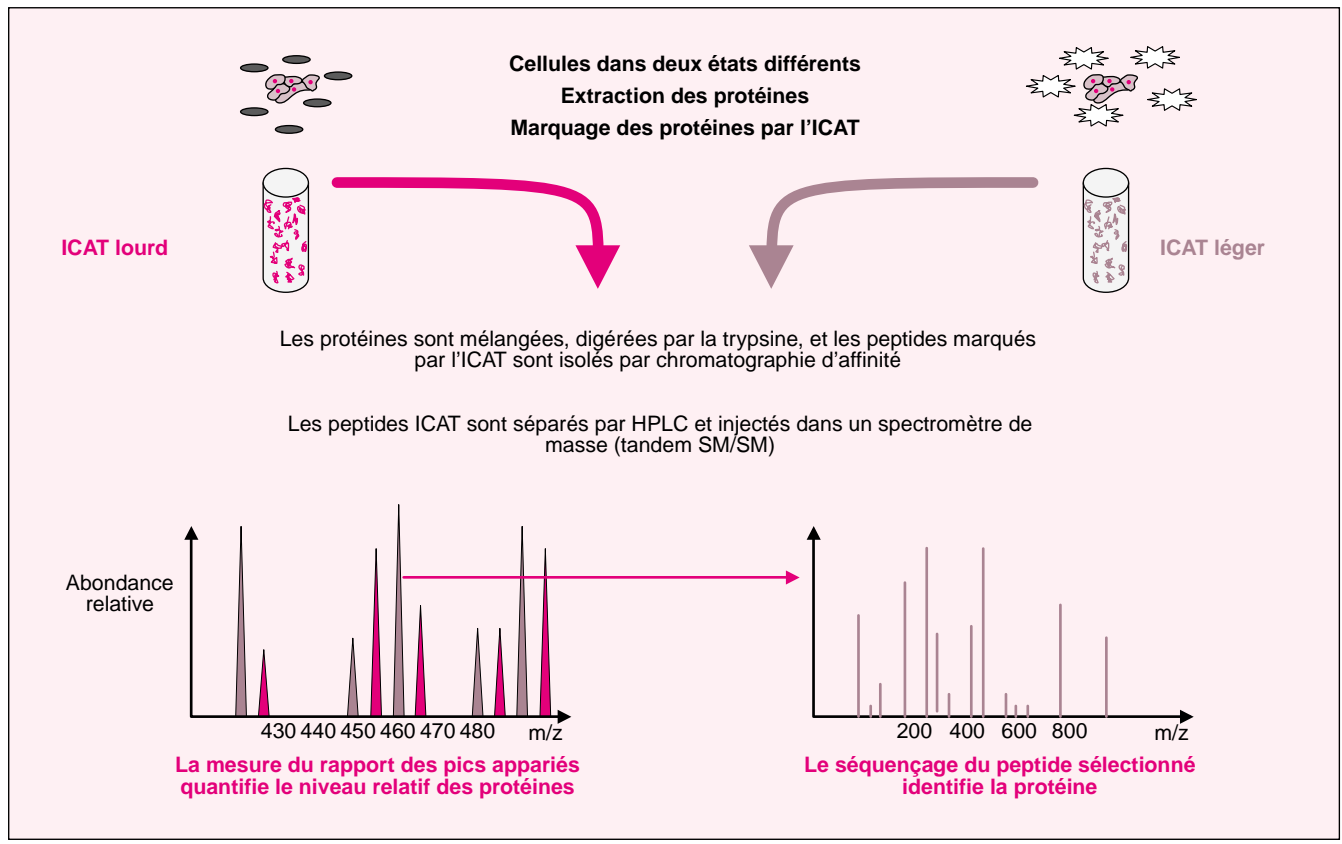


Figure 4. **Présentation schématique des principales étapes de l'analyse protéomique quantitative.** L'ICAT est un composé polyréactif pouvant se lier à la cystéine et à l'avidine, et fixant 8 Hydrogène ou 8 Deutérium. Les protéines de deux extraits cellulaires à comparer sont marquées respectivement par l'ICAT lourd et léger, puis mélangées et protéolysées par la trypsine. Les peptides marqués par l'ICAT, c'est-à-dire possédant une cystéine, sont isolés par chromatographie d'affinité (système avidine/biotine). L'ensemble des peptides ICAT subissent une séparation bidimensionnelle (HPLC et premier analyseur d'un spectromètre de masse tandem SM/SM). L'intensité relative de 2 pics appariés (séparés par 8 Da) indique le niveau relatif de la protéine d'où ils sont issus dans les deux extraits cellulaires. Le peptide sélectionné est alors fragmenté dans une chambre de collision et les fragments, analysés dans le deuxième analyseur du tandem SM/SM, donnent la séquence du peptide et permettent donc d'identifier la protéine (selon Aebersold [23]).

globalité de l'analyse, sont supprimées ou en tout cas ne sont plus obligatoires pour la QPA. Subsistent, en commun aux deux stratégies, la préparation des protéines (avec en plus, pour la QPA, le marquage des protéines par l'ICAT) et la digestion protéolytique. La préparation des peptides pour l'analyse en spectrométrie de masse est simplifiée et plus rapide dans le cas de la QPA puisqu'elle ne fait appel qu'à une opération de chromatographie d'affinité. La QPA n'accède à la première étape classique qu'au niveau des peptides protéolytiques où une double séparation en ligne selon deux critères indépendants est réalisée : l'hydrophobicité avec l'HPLC, technique qui autorise

des concentrations importantes, et la masse avec le premier spectromètre de masse du tandem SM/SM. Puis le deuxième spectre de masse est utilisé pour le séquençage des peptides. Les étapes d'identification des protéines et de la quantification de leur expression différentielle sont automatisées et très rapides. Néanmoins, il reste pour la QPA deux principales limites. Les protéines n'étant représentées dans la phase de présentation que par leurs peptides protéolytiques contenant une ou plusieurs cystéines, il n'est pas possible de mettre en évidence et donc de caractériser les modifications post-traductionnelles comme on peut le faire à partir des cartes protéiques d'E-2D. En cela, la

QPA se rapproche de l'analyse des transcriptomes. Le problème de l'analyse de la globalité des gènes exprimés reste quand même posé, bien que de manière moins aiguë que pour l'analyse protéomique classique, car la capacité de la méthode à mesurer les différences quantitatives au niveau des peptides marqués avec l'ICAT dépend à la fois de l'intensité des pics correspondant aux peptides appariés et de la complexité du mélange à analyser. Autrement dit, pour atteindre l'expression différentielle des protéines les moins abondantes, il est nécessaire d'avoir des mélanges moins complexes à analyser : il faut donc fractionner le mélange protéique initial.

Lorsque il s'agit de mettre en évidence les modifications post-traductionnelles des protéines – et on peut rappeler que pour un organisme complexe, il y a une moyenne de 7 à 10 protéines différentes par gène exprimé ! – il est obligatoire de passer par une étape adéquate de séparation et présentation des protéines. L'E-2D reste, malgré tous ses inconvénients, la méthode de choix. Cependant des recherches sont entreprises pour la remplacer et les deux principales voies qui s'ouvrent, offrent également une place centrale à la spectrométrie de masse. Il s'agit de substituer à la SDS-PAGE peu résolutive de la deuxième dimension, une technique de spectrométrie de masse afin d'obtenir une carte protéique virtuelle ayant une capacité de pics encore plus importante et permettant encore mieux d'obtenir une présentation ordonnée sans *a priori* des protéines. Dans un premier cas, il s'agit de faire, directement à partir du gel, le spectre de masse MALDI-ToF des protéines séparées par focalisation isoélectrique [22]; dans le deuxième, la focalisation isoélectrique est remplacée par une technique séparative de type HPLC ou électrophorèse capillaire qui est couplée *en ligne* à un nanospray SM/SM [23].

Enfin, il convient de souligner le développement considérable des méthodologies bio-informatiques qui, à partir des renseignements fournis par les banques de données, permettent maintenant de prévoir la structure des protéines et leur fonction. Bien entendu, ces prévisions théoriques doivent être confirmées expérimentalement. La caractérisation fonctionnelle réelle des protéines séparées actuellement par E-2D est souvent difficile à réaliser en raison des conditions dénaturantes signalées plus haut. De nouveaux développements méthodologiques sont à attendre également dans ce domaine.

Objet et applications de l'analyse protéomique

Dans un article récent [24], Klose, l'un des pères de l'électrophorèse bidimensionnelle, éprouve le besoin de préciser à nouveau le sens de termes couramment employés : l'ana-

lyse protéomique structurale devrait s'appliquer à l'isolement et au séquençage de toutes les protéines primaires codées dans le génome d'un organisme et l'analyse protéomique fonctionnelle désignerait la détermination des caractéristiques chimiques, biochimiques et biologiques de toutes les protéines primaires, c'est-à-dire non seulement la séquence en acides aminés des protéines mais aussi l'ensemble des modifications structurales et les changements quantitatifs auxquels les protéines sont soumises dans les différents tissus, organites cellulaires et étapes de développement. Comme nous l'avons signalé, la détermination quantitative, l'identification et la caractérisation de l'ensemble des protéines d'un protéome forment un objectif qui est encore loin d'être atteint. Dans le développement technique classique, avec présentation des protéines séparées par E-2D, analyse d'image comparative et identification/caractérisation des protéines, le nombre d'éléments protéiques qui peuvent être étudiés en une seule fois est fortement limité par rapport au nombre de protéines réellement en jeu. Cependant, de nombreuses études remarquables d'analyse protéomique *stricto sensu* ont été entreprises. L'équipe danoise de Celis a créé le premier et, probablement encore, le plus complet des sites d'analyse protéomique sur Internet en établissant notamment les premières cartes protéiques de kératocytes cultivés en présence de stimulus variés. Actuellement, cette équipe utilise les profils d'expression protéiques comme empreinte digitale des tumeurs de la vessie et comme point de départ de la recherche de marqueurs diagnostiques et pronostiques [25]. Nous pouvons citer aussi, à titre d'exemple également, la mise en évidence par l'équipe de Boucherie [21] des changements globaux de l'expression protéique qui sont à l'origine du stress adaptatif de la levure en réponse à l'action de peroxydants.

En fait, la caractéristique des analyses protéomiques est double. En premier lieu, il s'agit toujours d'études *in vivo* : elles doivent refléter au niveau protéique l'état physiologique de l'organisme étudié à un moment donné dans des conditions environ-

nementales données. Ensuite, elles doivent revêtir un caractère global, mais il apparaît nettement que ce n'est que la somme d'études partielles et le développement de l'informatique, qui peuvent donner accès à cette vue globale. L'avènement récent de l'analyse protéomique quantitative de Aebersold [20] permet sûrement d'identifier et de comparer quantitativement un très grand nombre de protéines primaires exprimées par un génome, mais alors le pan considérable de la diversité protéique due aux modifications post-traductionnelles et aux interactions protéines/ligands, c'est-à-dire tout ce qui donne lieu à la cascade des événements dynamiques survenant entre la lecture d'un génome et les fonctions cellulaires d'un organisme vivant, échappe à l'analyse ! Actuellement, dans la pratique courante de l'analyse protéomique, des objectifs partiels raisonnables sont la plupart du temps visés. Godovac-Zimmermann [26] faisait remarquer dans une publication récente que pour étudier un système de transduction du signal, il n'est sans doute pas pertinent d'identifier les milliers de protéines présentes sur un gel bidimensionnel préalablement à l'étude du phénomène cellulaire, et qu'une stratégie adaptée de stimulation ou de détection sélectives – dans le cas cité, révélation des gels bidimensionnels par des anticorps anti-phosphotyrosine et anti-phosphosérine – se révèle beaucoup plus efficace. L'auteur propose d'utiliser le terme d'analyse protéomique fonctionnelle dans un autre sens que celui suggéré par Klose et de le réserver à la dénomination de stratégies d'études partielles fondées sur la stimulation sélective et l'observation au niveau cellulaire de réseaux de protéines interagissantes. Dans le même ordre d'idée, Pasquali [27] a forgé le terme d'analyse protéomique ciblée (*target-oriented proteomics*, TOP) pour désigner les stratégies qui permettent d'analyser comment les produits des gènes interagissent les uns avec les autres en milieu physiologique. Il illustre son propos en identifiant l'ensemble des substrats des protéines tyrosine phosphatases par la combinaison des techniques de frac-

tionnement subcellulaire avec des séparations électrophorétiques bidimensionnelles révélées par une méthode permettant de mettre en évidence les interactions protéines/protéines (*far-Western immunoblotting*). Des banques de données protéomiques de plus en plus complètes se construisent sur les sites spécialisés Internet. Elles concernent les protéomes d'organismes simples ou complexes (virus, procaryote, eucaryote, plante, animal, homme), les différents tissus biologiques etc. Ces cartes protéiques sont utilisées pour des études biologiques variées portant aussi bien sur différents effets environnementaux (stress, effecteurs hormonaux, virus etc.) que sur des phénomènes biologiques complexes (cycle cellulaire, apoptose, etc.). En biomédecine sont étudiés de nombreux tissus et fluides biologiques sains ou malades, en particulier des échantillons solides de tumeurs d'origines variées. Les études des protéomes de lignées cellulaires de différents types de cancer sont en cours. L'identification et la caractérisation des partenaires de protéines ou de ligands sont réalisées. Le but de ces études est non seulement de comprendre les mécanismes physiopathologiques mis en œuvre, mais également de repérer des marqueurs spécifiques de la pathologie à forte valeur diagnostique ou pronostique. En pharmacologie, l'analyse protéomique permet d'évaluer l'action des médicaments en comparant les profils protéiques d'un état normal ou malade avec ou sans traitement. Il s'agit d'identifier les cibles protéiques des médicaments, d'évaluer les relations structure/activité des analogues de médicaments déjà connus, de donner des indications sur les effets secondaires ou toxiques des composés testés. L'appréhension globale de la complexité des systèmes vivants est à l'ordre du jour [28] et la combinaison des approches protéomique et génomique constitue actuellement le meilleur moyen d'atteindre cet objectif [29]. Dans le cadre des applications biomédicales, cela va se traduire par l'amélioration de nos outils diagnostiques et pronostiques et par le progrès de la recherche pharmacologique ■

S. Thébault
N. Machour
R. Charlionet
M. Fontaine
F. Tron

Inserm U. 519, Faculté de médecine pharmacie, 76183 Rouen Cedex, France.

F. Perrot
T. Jouenne

Cnrs UMR 6522, Faculté des sciences, 76821 Mont-Saint-Aignan Cedex, France.

M. Hubert
C. Lange

Cnrs UPRESA 6014, Faculté des sciences, 76821 Mont-Saint-Aignan Cedex, France.

RÉFÉRENCES

1. Wilkins MR, Sanchez JC, Gooley AA, *et al*. Progress in proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev* 1995; 13: 19-50.
2. Chee M, Yang R, Hubbell E, *et al*. Accessing genetic information with High-density DNA arrays. *Science* 1996; 274: 610-4.
3. Gygi ST, Rochon Y, Franz BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999; 19: 1720-30.
4. Rabilloud T. *Proteome research: two-dimensional gel electrophoresis and identification methods*. Berlin: Springer Verlag; 2000.
5. Wilkins MR, Williams KL, Appel RD, Hochstrasser DF. *Proteome research: new frontiers in functional genomics*. Berlin: Springer verlag; 1997.
6. Bjellqvist B, Ek K, Righetti PG, *et al*. IEF in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods* 1982; 6: 317-39.
7. Lambin P. Reliability of molecular weight determination of proteins by polyacrylamide gradient gel electrophoresis in the presence of SDS. *Anal Biochem* 1978; 85: 114-25.
8. Görg A, Obermaier G, Boguth A, *et al*. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* 2000; 21: 1037-53.
9. Rilbe H. Historical and theoretical aspect of isoelectric focusing. *Ann NY Acad Sci* 1973; 209: 11-22.
10. Santoni V, Molloy M, Rabilloud T. Membrane proteins and proteomics: un amour impossible? *Electrophoresis* 2000; 21: 1054-70.

11. Herbert B. Advances in protein solubilisation for two-dimensional electrophoresis. *Electrophoresis* 1999; 20: 660-3.
12. Corthals GL, Wasinger VC, Hochstrasser DF, Sanchez JC. The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* 2000; 21: 1104-15.
13. Righetti PG, Barzaghi B, Luzzana M, Manfredi G, Faupel M. A horizontal apparatus for isoelectric protein purification in a segmented immobilized pH gradient. *J Biochem Biophys Methods* 1987; 15: 189-98.
14. Voss T, Haberl P. Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: consequences for comprehensive data analysis. *Electrophoresis* 2000; 21: 3345-51.
15. Gras R, Müller M, Gasteiger E, *et al*. Improving protein identification from peptide mass fingerprinting through a parametered multi-level scoring algorithm and optimized peak detection. *Electrophoresis* 1999; 20: 3535-50.
16. Karas M, Hillenkamp F. Laser desorption ionisation of proteins with molecular masses exceeding 10 000 daltons. *Anal Chem* 1988; 60: 2299-301.
17. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionisation for mass spectrometry of large biomolecule. *Science* 1989; 246: 64-71.
18. Mann M, Hojrup P, Roepstorff P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* 1993; 22: 338-45.
19. Clauser KR, Hall SC, Smith DM, *et al*. Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. *Proc Natl Acad Sci USA* 1995; 92: 5072-6.
20. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotech* 1999; 17: 994-9.
21. Godon C, Lagniel G, Lee J, *et al*. The H₂O₂ stimulin in *Saccharomyces cerevisiae*. *J Biol Chem* 1998; 273: 22480-9.
22. Ogorzalek Loo RR, Mitchell C, Stevenson TI, *et al*. Sensitivity and mass accuracy for proteins analyzed directly from polyacrylamide gels: Implication for proteome mapping. *Electrophoresis* 1997; 18: 382-90.
23. Valascovic GA, Kelleher NL, McLafferty FW. Attomole protein characterization by capillary electrophoresis -mass spectrometry. *Science* 1996; 273: 1199-202.
24. Klose J. Genotypes and phenotypes. *Electrophoresis* 1999; 20: 643-52.
25. Celis JE, Ostergaard M, Rasmussen HH, *et al*. A comprehensive protein resource for the study of bladder cancer: <http://bio-base.dk/cgi-bin/celis>. *Electrophoresis* 1999; 20: 300-9.
26. Godovac-Zimmermann J, Soskic V, Poznanovic S, Brianza F. Functional proteomics of signal transduction by membrane receptors. *Electrophoresis* 1999; 20: 952-61.

RÉFÉRENCES

27. Pasquali C, Vilbois F, Curchod ML, van Huijsduijnen RH, Arigoni F. Mapping and identification of proteins-protein interactions by two-dimensional far-Western immunoblotting. *Electrophoresis* 2000 ; 21 : 3357-68.
28. Harry JL, Wilkins MR, Herbert BR, Packer NH, Gooley AA, Williams KL. Proteomics : capacity versus utility. *Electrophoresis* 2000 ; 21 : 1071-81.
29. Nelson PS, Hon D, Rochon Y, *et al.* Comprehensive analysis of prostate gene expression : convergence of expressed sequence tag databases, transcript profiling and proteomics. *Electrophoresis* 2000 ; 21 : 1823-31.

Summary

Object and methodological development of proteomics

The long-term goal of proteomics is to identify, quantify and characterize the whole set of proteins from biological medium, in such a way that we can understand, at the molecular level, how living cells and organisms function, organize and respond to environmental changes. Unlike genomics, which involves the study of simple and replicable polymers, proteomics involves the study of very complex and non-replicable heteropolymers having a wide range of concentration, size, physical properties, secondary and tertiary structures, solubility and functions. Hence proteomics requires the synergistic support of different techniques, classically two-dimensional gel electrophoresis and mass spectrometry. In this review we describe the technological foundations of proteomics and the methodological evolutions which are beginning to take shape for this new analytical science.

TIRÉS À PART

R. Charlionet.