La protéomique

Une approche nécessaire en physiologie et en génétique des plantes

H. Thiellement[1]

Introduction

Des avancées considérables en robotique, en informatique et en nanotechnologie ont permis l'émergence de différentes approches globales en biologie.

Au niveau de l'ADN, la séquence de génomes entiers de micro-organismes a été réalisée et le premier génome végétal sera intégralement connu en 2000 (plus de 93 % du génome de 130 Mb d'*Arabidopsis thaliana* est déjà séquencé fin juin 2000).

Au niveau de l'ARN messager, plus de 2 millions d'étiquettes d'(EST) (Expressed Sequence Tag) des ADN complémentaires ont été séquencés chez l'Homme. En juin 2000, plus de 85 000 ADNc sont répertoriés chez le soja, 72 000 chez la tomate, 61 000 chez le maïs, 53 000 chez le riz et 51 000 chez *Arabidopsis*, soit, même compte tenu des redondances, presque tous les gènes (dont le nombre est estimé entre 10 000 et 20 000).

On réalise, avec ces données de séquence, des filtres à haute densité ou des « puces à ADN » (DNA chips, microarrays). Sur une très

[1] Station de Génétique végétale, Inra, ferme du Moulon, 91190 Gif-sur-Yvette, France.

petite surface est fixée toute la collection d'ADNc, ou des fragments de leurs séquences, que l'on hybride avec les messagers présents à un stade ou dans un organe donné. On peut alors en déduire quels gènes, parmi des milliers, sont transcrits et en quelles proportions (Marshall et Hodgson, 1998).

On dispose donc aujourd'hui d'outils extrêmement puissants pour étudier la transcription et sa régulation.
Au niveau des protéines, la méthode utilisée est encore aujourd'hui l'électrophorèse bidimensionnelle (O'Farrell 1975 ; Klose 1975). Cette méthode sépare les protéines dénaturées selon 2 critères indépendants : i) le point isoélectrique dans une première dimension (isoélectrofocalisation), ii) la masse moléculaire dans la seconde (SDS PAGE). On obtient ainsi des gels en plaque où se répartissent de plusieurs centaines à plusieurs milliers de spots protéiques (c.-a.-d. de produits de gènes) (fig. 1).

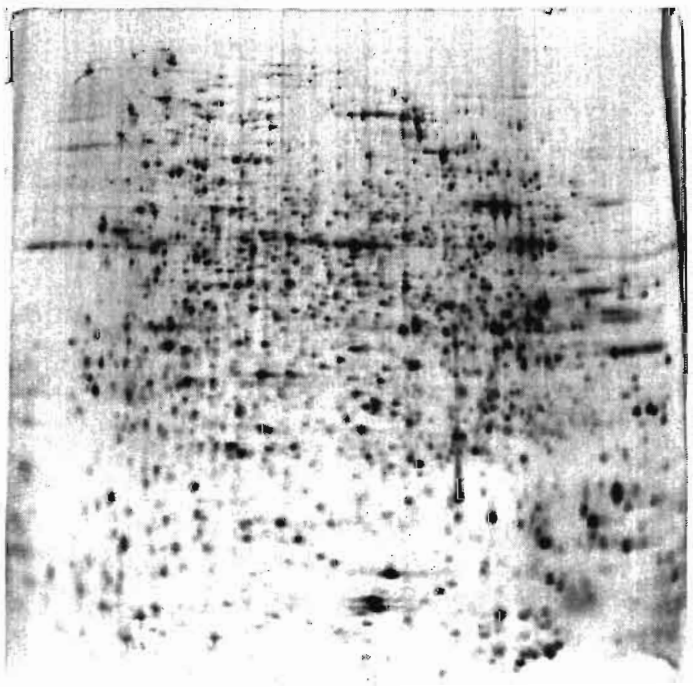

Figure 1
Gel bidimensionnel obtenu avec les protéines totales dénaturées de la partie aérienne de germinations étiolées d'*Arabidopsis thaliana*, écotype Landsberg *erecta*.

Ces protéines peuvent être caractérisées par différentes méthodes. La caractérisation immunologique peut être utilisée quand on soupçonne *a priori* tel ou tel spot ou que l'on recherche une protéine connue.

Le séquençage, par les méthodes classiques d'Edman, des 10 à 15 premiers acides aminés de la partie N-terminale, ou de séquences internes, est très fiable mais coûteuse. Lente (1spot/jour), elle exige de grandes quantités de protéines.

Cependant, il vaut mieux combiner les différents critères suivants :

– la masse moléçulaire et le point isoélectrique déduits directement de la position du spot sur le gel ;

– la composition globale en acides aminés, qui consiste, après hydrolyse acide, à estimer les pourcentages des différents acides aminés. Peu coûteuse et rapide (plus de 20 spots/jour), elle se révèle très utile en combinaison avec d'autres critères ;

– le séquençage de 3 à 4 acides aminés consécutifs (« sequence tag »), méthode rapide et qui permet d'utiliser le même spot pour d'autres analyses.

La combinaison des données obtenues sur la composition en acides aminés, une « sequence tag », le point isoélectrique et la masse moléculaire suffit en général à une caractérisation sans ambiguïté si la protéine, ou son gène, se trouve déjà archivé dans les banques de données.

Le nombre d'échantillons traités est accru et les coûts considérablement réduits par rapport au séquençage classique selon Edman.

Mais la caractérisation la plus performante est aujourd'hui réalisée par spectrométrie de masse. Les spectromètres de type MALDI-TOF mesurent la masse des peptides obtenus après digestion trypsique de la protéine. Appareils à haut débit, ils permettent l'identification rapide de nombreuses protéines quand une information génomique riche est disponible pour l'espèce étudiée. En effet, les données recueillies (« peptide mass fingerprints ») sont comparées à celles des banques de données de type Swissprot ou TrEMBL. Les spectromètres de masse de type ESI-MS/MS permettent quant à eux d'aller, par fragmentations successives, jusqu'à la séquence en acides aminés des peptides obtenus. Comme avec le séquençage d'Edman, ils autorisent alors la comparaison avec les séquences d'autres espèces et sont donc plus adaptés quand le génome de l'espèce étudiée est peu caractérisé.

Intérêts de la protéomique

On assiste au développement de 3 nouvelles disciplines : au niveau de l'ADN, du génome, la génomique, au niveau de l'ARNm (de l'ADNc), des transcriptomes, la transcriptomique et au niveau de la protéine, des protéomes, la protéomique, qui représentent une nouvelle façon de faire de la biologie avec l'analyse et l'étude simultanée d'un très grand nombre de données.

Au niveau génomique, malgré le développement d'algorithmes toujours plus astucieux pour les études *in silico* des séquences, on se demande toujours où sont les gènes et surtout quelles sont leurs fonctions. La grande question est donc celle de l'analyse fonctionnelle. En effet, plus de la moitié des gènes séquencés sont encore « orphelins » (de fonction).

Au niveau transcriptomique, toutes les régulations ne sont pas accessibles : la corrélation est médiocre (0,5) entre abondance des messagers et quantité des protéines (Anderson et Seilhammer, 1997), et de très nombreuses régulations sont post-transcriptionnelles (modifications post-traductionnelles, turn over des protéines...).

Le niveau protéomique est donc indispensable :

– pour savoir si, où, quand et en quelle quantité un messager sera traduit ;

– pour étudier les modifications post-traductionnelles (des dizaines ont été décrites : coupure du peptide d'adressage, phosphorylation, glycosylation, etc.) ;

– pour décrire les différentes formes d'une protéine (le paradigme un gène-une protéine n'est plus d'actualité) ;

– pour étudier les interactions fonctionnelles entre protéines et entre les protéines et les molécules qu'elles synthétisent (sucres, lipides, etc.) ;

– pour connaître la localisation subcellulaire des protéines.

La protéomique permet d'étudier :

– les effets de molécules à cibles multiples comme les drogues, les médicaments ou les hormones ;

– les effets multiples de stress biotiques (parasites, maladies) ou abiotiques (froid, sécheresse, pollution...) (Hausmann *et al.*, 2000 ; Leonardi, cet ouvrage) ;

– les effets pléïotropes d'une mutation, par exemple de repérer les cibles d'un facteur de transcription (Damerval et Le Guilloux, 1998) ;

mais aussi de :

– compléter les cartes génétiques avec des gènes exprimés (les marqueurs protéiques ne sont pas codés par des séquences anonymes ni par des pseudo gènes) (de Vienne *et al.*, 1996) ;

– de localiser sur les cartes génétiques des « PQL » pour Protein Quantity Loci (Damerval *et al.*, 1994). À partir de la quantification des spots, une analyse de type QTL sur un dispositif adéquat permettra de localiser les régions du génome impliquées dans la régulation en quantité de telle ou telle protéine. On pourra alors rechercher les co-localisations des QTL pour un caractère d'intérêt et des PQL, permettant de faire des hypothèses sur des « protéines candidates » dont la fonction pourrait être liée à l'expression de ce caractère ;

– calculer des distances génétiques. La comparaison des protéomes de plusieurs génotypes, à un même stade d'un même organe, permet de compter pour chaque couple de génotypes les spots communs et spécifiques et de calculer des indices de similarité ou de distance à partir de ces nombres (Thiellement *et al.*, 1989 ; Zivy *et al.*, 1995, Marquès *et al.*, 2000). La construction d'arbres phénétiques à partir de ces matrices de distances permet alors d'estimer les relations génétiques entre ces génotypes (variétés, espèces ou genres) (fig. 2).

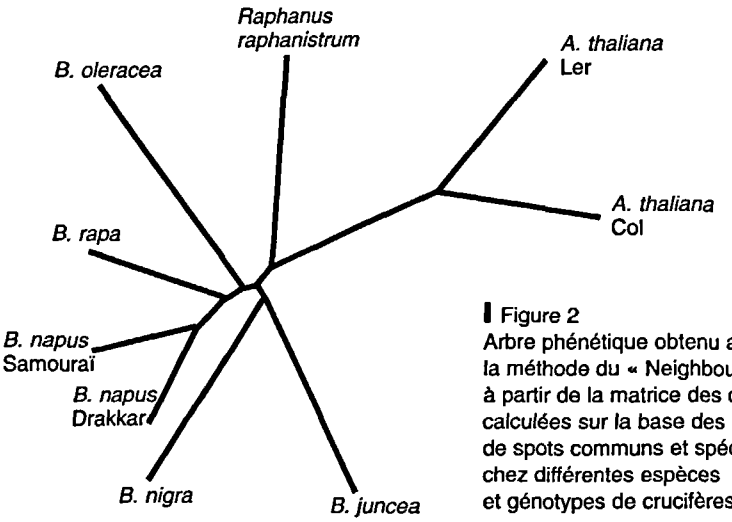

Figure 2
Arbre phénétique obtenu avec la méthode du « Neighbour Joining » à partir de la matrice des distances calculées sur la base des nombres de spots communs et spécifiques chez différentes espèces et génotypes de crucifères.

On peut également, avec cette approche, étudier les relations d'homéologie entre les génomes des polyploïdes (Bahrman et Thiellement, 1987). La comparaison de génotypes tétraploïdes (par ex. AACC et AABB) et de chacun des diploïdes (AA, BB et CC) porteurs des mêmes génomes a permis de comparer, chez les crucifères, le niveau d'expression d'un même génome à différents niveaux de ploïdie et vis-à-vis de différents homéologues (Marquès *et al.*, 2000).

Quelques perspectives

Malgré les avancées technologiques extraordinaires de la génomique et plus récemment de la transcriptomique, la protéomique apparaît comme une discipline complémentaire et indispensable pour l'analyse fonctionnelle des gènes (Humphery-Smith *et al.*, 1997, Wilkins *et al.*, 1997). Ses applications potentielles en diagnostic médical et en pharmacologie ont poussé à son développement technologique à différents niveaux :

– au niveau de la biochimie des protéines (extraction de protéines hydrophobes, de pI extrêmes, standardisation, automatisation et reproductibilité des protocoles) ;

– au niveau de la caractérisation des protéines par la spectrométrie de masse (vers un « laser moléculaire » qui balayera le gel et caractérisera les spots au fur et à mesure) ;

– au niveau bioinformatique : analyse d'images, quantification des protéines, comparaison multiple de gels, interconnection entre les bases de données protéiques, génétiques, physiologiques et bibliographiques (Appel *et al.*, 1997).

Des cartes protéiques sont actuellement développées, accessibles sur internet, où on peut cliquer sur un spot et accéder aux informations le concernant. La figure 3 montre les informations recueillies récemment sur les protéines de feuille d'*Arabidopsis* (Sarazin *et al.*, 2000).

Les applications de la protéomique à la taxonomie ont encore été peu exploitées et une protéomique comparative permettra aussi :

– de court-circuiter les délais de passage des espèces modèles aux espèces cibles et inversement ;

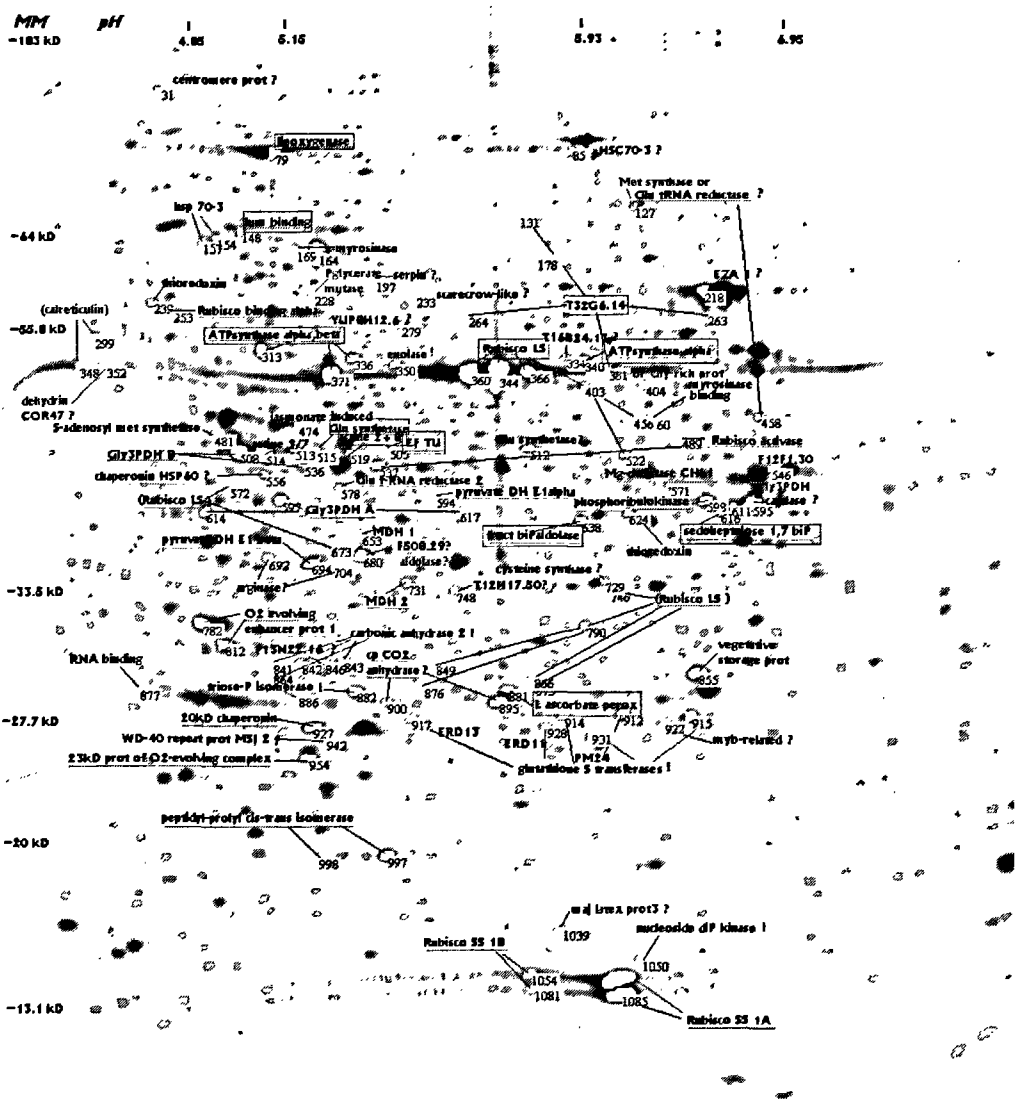

Figure 3
Carte protéique selon le standard « Swiss 2D PAGE »
du protéome de feuille verte d'*Arabidopsis*, écotype Columbia.
Les points d'interrogation indiquent que la masse moléculaire observée
sur le gel ne correspond pas à la description de la protéine selon Swissprot.
Quand il s'agit de fragments protéiques avérés, ils sont indiqués entre parenthèses.
Les protéines trouvées communes à toutes les crucifères étudiées
(des genres *Brassica, Raphanus* et *Arabidopsis*) sont encadrées.

– de caractériser ces protéines invariantes (spots communs) et donc ces gènes « contraints » à évolution plus lente (gènes codant pour des protéines « de ménage » ou impliquées dans des complexes, ou gènes d'une autre nature ?).

Conclusion

La protéomique s'avère donc aujourd'hui incontournable pour répondre à un grand nombre de questions fondamentales et appliquées. Ses applications à la physiologie et à l'amélioration des plantes ont fait l'objet de revues récentes (Thiellement *et al.*, 1999, Thiellement *et al.*, 2001). Elle permet aussi d'imaginer et de répondre à de nouvelles questions.

Comme les autres approches globales en biologie, elle débouche sur la définition d'ensembles de protéines ou de gènes co-régulés. Elle montre la nécessité et l'urgence de développer de nouveaux concepts et de nouveaux modèles pour gérer d'énormes quantités de données afin de pouvoir rendre compte, et prédire, la complexité des systèmes biologiques.

Bibliographie

Anderson L., Seilhammer J 1997 — A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18 : 533-537.

Appel RD, Palagi PM, Walther D, Vargas JR, Sanchez J-C, Ravier F, Pasquali C, Hochstrasser DF 1997 — Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: 1. Features and user interface. Electrophoresis 18 : 2724-2734.

Bahrman N, Thiellement H 1987 — Parental genome expression in synthetic wheats (*Triticum turgidum* sp. x *T. tauschii* sp.) revealed by 2D electrophoresis of seedling proteins. Theor Appl Genet 74 : 246-251

Damerval C, Le Guilloux M 1998 — Characterization of novel proteins affected by the *o2* mutation and expressed during maize endosperm development. Mol Gen Gene. 257 : 354-361.

Damerval C, Maurice A, Josse JM, de Vienne D 1994 — *Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression.* Genetics 137 : 289-301.

Hausman JF, Evers D, Thiellement H, Jouve L 2000 — Compared responses of poplar cuttings and *in vitro* raised shoots to short-term chilling treatments. Plant Cell Rep 19 : 954-960.

Humphery-Smith I, Cordwell SJ, Blackstock WP 1997 — Proteome research: Complementarity and limitations with respect to the DNA and RNA words. Electrophoresis 18 : 1217-1242.

Klose J 1975 — Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. Humangenetik 26 : 231-243.

Marques K, Sarazin B, Chane-Favre L, Thiellement H 2000 — Recent developments and new prospects in *Cruciferae* proteomics. *In* From genome to proteome: Knowledge acquisition and representation. DF Hochstrasser, V Pallini, L Bini eds. 4th Siena 2D Electrophoresis Meeting : 227-228.

Marshall A, Hodgson J 1998 — DNA chips: An array of possibilities. Nature Biotechnology 16 : 27-31.

O'Farrell PH 1975 — High resolution two-dimensional electrophoresis of proteins. J Biol Chem. 250 : 4007-4021.

Sarazin B, Tonella L, Marques K, Paesano S, Chane-Favre L, Heller M, Sanchez JC, Hochstrasser DF, Thiellement H 2000 — The Swiss-2DPage reference map of the leaf proteins of *Arabidopsis thaliana* ecotype Columbia. *In* From genome to proteome: Knowledge acquisition and representation. DF Hochstrasser, V Pallini, L Bini eds. 4th Siena 2D Electrophoresis Meeting (Abst.) (in press).

Thiellement H, Seguin M, Bahrman N, Zivy M 1989 — Homoeology and phylogeny of the A, S and D genomes of the *Triticinae.* J Mol Evol 29 : 89-94.

Thiellement H, Bahrman N, Damerval C, Plomion C, Rossignol M, Santoni V, De Vienne D, Zivy M 1999 — Proteomics for genetic and physiological studies in plants. Electrophoresis 20 : 2013-2026.

Thiellement H, Plomion C, Zivy M 2001 — Proteomics as a tool for plant genetics and breeding. *In* Proteomics from protein sequence to function. M. Dunn, S Pennington eds. Bios, UK : 289-309.

Vienne D de, Burstin J, Gerber S, Leonardi A, Le Guilloux M, Murigneux A, Beckert M, Bahrman N, Damerval C, Zivy M 1996 — Two-dimensional electrophoresis of proteins as a source of monogenic and codominant markers for population genetics and mapping the expressed genome. Heredity 76 : 166-177

Wilkins MR, Williams KL, Appel RD, Hochstrasser DF 1997 — Proteome Research: New Frontiers in Functional Genomics. Springer-Verlag, Berlin. 243 p.

Zivy M, El Madidi S, Thiellement H 1995 — Distance indices in a comparison between the A, D, and R genomes of the *Triticeae* tribe. Electrophoresis 16 : 1295-1300.