Apprentissage symbolique et interprétation de gels d'électrophorèse bidimensionnelle

Pierre Nugues, Robert Whalen, Jean-Paul Haton

INRIA

UNITÉ DE RECHERCHE
INRIA-LORRAINE

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 1128

Programme 8
Communication Homme-Machine

APPRENTISSAGE SYMBOLIQUE ET INTERPRETATION DE GELS D'ELECTROPHORESE BIDIMENSIONNELLE

Pierre NUGUES
Robert WHALEN
Jean-Paul HATON

Décembre 1989

RR_1128

APPRENTISSAGE SYMBOLIQUE ET INTERPRÉTATION DE GELS D'ÉLECTROPHORÈSE BIDIMENSIONNELLE

SYMBOLIC LEARNING AND INTERPRETATION OF 2-D ELECTROPHORESIS GELS

Pierre Nugues, INRIA-CRIN, Nancy et Cognitech, Paris,
Robert Whalen, Institut Pasteur, Paris,
Jean-Paul Haton, INRIA-CRIN, Nancy.

RÉSUMÉ

Nous présentons une application des techniques d'apprentissage symbolique. Cette application permet l'extraction automatique et la formalisation d'expertise biologique. Elle a pour support une méthode d'analyse de protéines : l'électrophorèse bidimensionnelle. Cette dernière produit des gels plans dont on extrait les paramètres numériques par traitement d'image. Grâce à la méthode que nous exposons, nous identifions, à partir des paramètres, des protéines qui peuvent jouer un rôle dans les transitions entre les différentes étapes de la maturation musculaire.

Mots-clés : électrophorèse bidimensionnelle, traitement d'image, apprentissage, systèmes experts, extraction de règles, classification symbolique.

An application of machine learning techniques is presented. Automated rule extraction and biological expertise formalizing is allowed through this application. A method to analyze proteins: two-dimesional electrophoresis, is used as a subject of investigation. This technique of electrophoresis yields two-dimensional images of which parameters are extracted by means of image processing. Using the method presented here, proteins which may play a role in transitions between different stages of muscle maturation are identified from these extracted parameters.

Key words: two-dimensional electrophoresis, image processing, machine learning, expert systems, rule extraction, conceptual clustering.

I. INTRODUCTION

Dans cet article nous considérons l'analyse et l'interprétation d'images de gels d'électrophorèse bidimensionnelle. Cette technique a été mise au point par O'Farrell en 1975 [ofarrell] et améliorée par Garrels [garrels]. Elle permet de détecter la plupart des protéines présentes dans une substance animale ou végétale : des plus

importantes aux espèces mineures. On situe l'identification à près de 80% de la masse totale du tissu.

On analyse un échantillon de matière en distribuant, suivant deux axes orthogonaux : l'un caractérise le point isoélectrique et l'autre le poids moléculaire, les différentes molécules de cet échantillon. Elles se focalisent à des endroits bien définis et cela se traduit par l'addition de courbes sur l'image du gel. La figure 1 représente une électrophorèse bidimensionnelle de plasma.

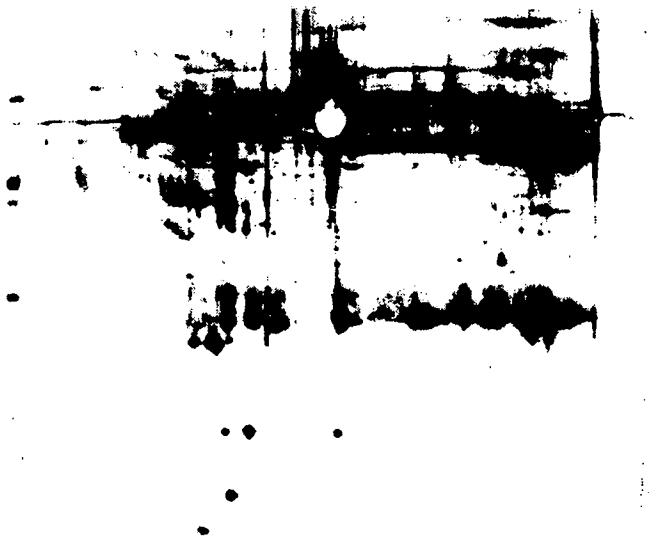

Figure 1 : Électrophorèse bidimensionnelle de sérum humain

Chaque amas de matière peut se modéliser par la somme de gaussiennes bidimensionnelles :

$$f(x, y) = A \exp(- \frac{(x - xo)^2}{2\sigma_x^2} - \frac{(y - yo)^2}{2\sigma_y^2})$$

où A représente l'amplitude de la gaussienne, (xo, yo), les coordonnées du sommet, (σ_x, σ_y), les écarts-type suivant les deux axes orthogonaux. On extrait généralement ces paramètres à l'aide d'une approximation par la méthode des moindres carrés.

Cette technique d'électrophorèse bidimensionnelle permet, par exemple, de distinguer plus de 500 points différents sur un gel de plasma. La contrepartie de cette richesse d'informations tient dans la complexité d'interprétation. À l'exception de certains tissus tels que le plasma, la plupart des protéines qui figurent sur les gels ont une fonction inconnue et ne possèdent même pas de nom. L'automatisation

du déchiffrement des gels est essentielle pour que l'électrophorèse bidimensionnelle puisse connaître une extension, car si les sources de gels sont nombreuses et touchent la plupart des tissus, il se heurte toujours à la quantité de points à considérer.

L'étude que nous présentons concerne une expérimentation sur des cellules musculaires, où il apparaît un grand nombre de protéines et où nous essayons de cerner celles qui peuvent être pertinentes dans l'évolution et les transitions des différents états musculaires afin d'extraire de nouvelles règles d'interprétation. L'essentiel de ce travail est ainsi relatif à des méthodes d'apprentissage symbolique inductif.

Dans l'avenir, ces techniques devraient permettre d'alimenter des bases de données d'images où on pourra localiser automatiquement sur un gel les protéines caractéristiques d'un état pathologique ou physiologique et d'interpréter la nature du tissu ou l'état du patient [nugues].

II. LES DONNÉES

Les gels que nous avons analysés proviennent de quatre lignées de cellules murines. Il s'agit en quelque sorte d'un examen temporel car ces cellules correspondent à quatre étapes du développement du muscle. Ce sont par ordre de maturité :

1. une lignée de souche primitive, avec des caractéristiques proches des cellules embryonnaires ;

2. une lignée d'une souche de cellules pouvant être le précurseur des cellules d'un mésoderme : muscle, adipocyte, chondroblastes...

3. une lignée de myoblastes. On ne trouve ces cellules que dans les tissus musculaires. Il s'agit d'une phase de maturation qui traduit déjà la spécialisation vers le muscle.

4. Une lignée de myotubes qui est caractéristique de la cellule musculaire. Son aspect est celui d'un cylindre. Chacun de ces cylindres résulte de l'agencement de cellules mononuclées pour former des fibres musculaires plurinuclées.

On a analysé les échantillons en triple pour chacune de ces étapes de la croissance du muscle ce qui a donc donné douze gels. L'électrophorèse a été réalisée selon des conditions standardisées par la firme Protein Databases, aux États-Unis. On a ensuite rentré les images de ces gels, sous forme numérique, dans le système PDQUEST. Du fait du traitement par des précurseurs radioactifs, les protéines ont pu être révélées par une radiographie sur film. Le système a extrait les caractéristiques de chacun des points par une approximation par les moindres carrés et apparié de manière semi-automatique les protéines des douze gels entre elles. Cet appariement est une opération délicate et sujette à certaines erreurs, notamment pour les plus petites taches.

Nous représentons chaque protéine par l'intégrale de sa quantité de matière sur les différents gels, ce que traduit par exemple la figure 2.

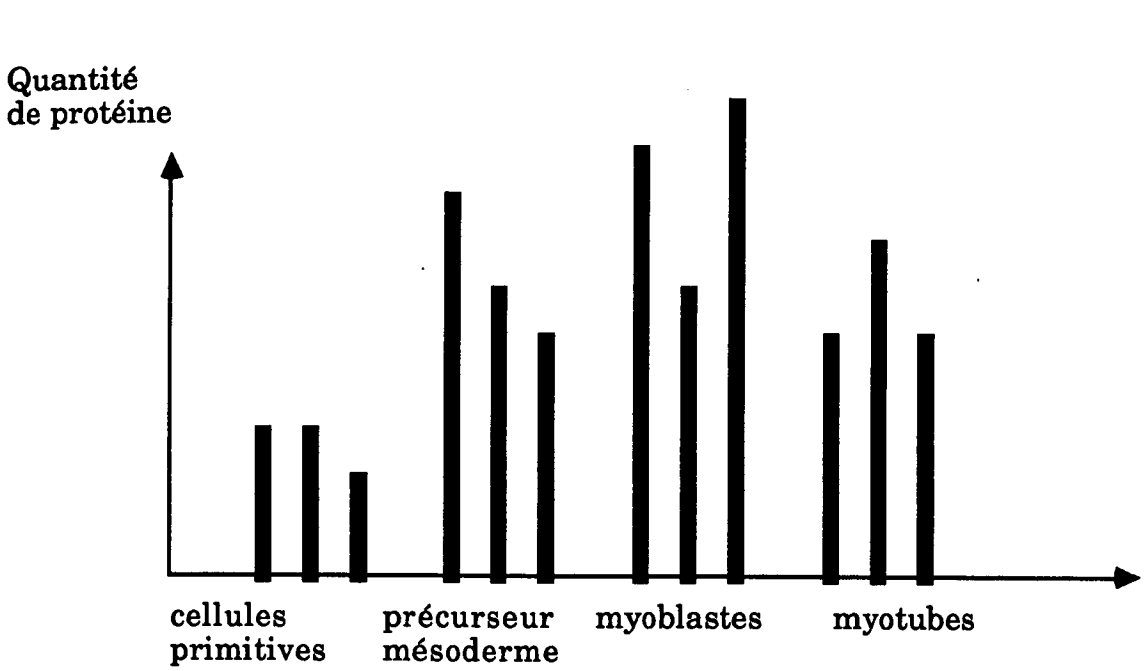

Figure 2 : Courbe de variation de la quantité de protéine suivant les différents états de la cellule.

On devrait retrouver une similitude de quantité par groupes de trois. Ceci n'est pas toujours le cas. Ces différences sont caractéristiques de la très grande variabilité biologique pour la plupart des protéines. Nous n'avons donc retenu dans une première analyse que les protéines présentant une certaine constance au travers des triplés. En effet, il est impossible de décrire la fonctionnalité d'une protéine à la seule connaissance de sa quantité en fonction de l'état biologique, si pour un même état celle-ci est variable. On considérera donc que ces variations à l'intérieur d'un même état ont d'autres explications auxquelles nous ne pouvons accéder dans l'état actuel de nos données et que nous laisserons de côté pour le moment.

Nous avons donc procédé à un filtrage qui s'est fait en majeure partie grâce au critère de la moyenne sur trois instants semblables divisée par l'écart-type de ces trois instant, soit $CV = \frac{x - m}{\sigma}$. Nous avons conservé au total 208 protéines pour notre analyse.

III. L'IMPLANTATION DE L'ALGORITHME

Considération sur les données

Le processus que nous avons développé est totalement expérimental et suit l'analyse biologique. En particulier, nous devons prendre garde au fait que nous ne savons pas si les quatre tranches temporelles sont suffisantes pour mettre en évidence l'action d'une protéine. Notre objectif est que, dans la mesure du possible, les résultats de l'analyse symbolique puissent guider l'orientation et la conception de manipulations à venir.

Nous avons repris les méthodes de classification conceptuelle définies par Michalski [michalski], afin de pouvoir offrir une description logique conjonctive de chaque classe et identifier les éléments ou la relation conjonctive entre les éléments, qui puis se déclencher ou jouer un rôle dans le déclenchement des passages d'un groupe à l'autre.

Cet algorithme, malheureusement, prend difficilement en compte les imprécisions ou les erreurs possibles dans les gels d'électrophorèse. Celles-ci peuvent provenir d'appariements défectueux ou simplement du caractère approximatif de la mesure de la quantité de matière de la protéine. On peut noter que la plupart des expériences réalisées en classification conceptuelle se font avec des jeux de données réduits ou sur des cas d'école, alors qu'ici nous traitons un exemple réel, de complexité nettement supérieure.

Les données correspondant à chaque protéines prennent la forme de 12 nombres caractérisant les quatre instants répétés trois fois[1]. Notre algorithme de classification ne traite que des domaines discrets, nous avons donc normalisé l'allure de chaque protéine par rapport à la quantité sur le gel où sa présence est la plus importante. Nous avons enfin discrétisé les valeurs dans un domaine de dix entiers. Ainsi, nous avons fait abstraction de la quantité absolue de matière pour privilégier le comportement. Sur ces données, deux types de classements sont possibles :

1. Un classement suivant les types de forme de courbes. Les éléments sont alors les protéines et leurs attributs sont leurs quantités au cours du temps. Ce type de classement permet de regrouper les protéines suivant leur allure au cours du temps (croissante, décroissante...) et d'émettre ainsi des hypothèses sur leur fonctionnalité.

2. Un classement suivant les types de gels. Les éléments sont alors les gels et leurs attributs sont les quantités de protéine à un instant donné.

Quelques particularités

[1]Ces quantités sont en fait les pourcentages de matière de la protéine par rapport à la quantité de matière du gel total multiplié par un certain coefficient.

Pour rendre l'algorithme de classification réalisable par un ordinateur de type *SUN-3*, nous avons dû y apporter un certain nombre d'altérations :

1. Dans la classification, telle que l'a décrite Michalski, on se trouve libre de prendre les noyaux au hasard et s'attendre à une convergence de l'algorithme. Ici nous ne pouvons procéder de cette façon, car les temps de calcul seraient beaucoup trop longs. Nous avons à chaque fois essayé de choisir un représentant typique d'une classe pressentie par l'expert en fonction de ses connaissances et du contexte expérimental.

2. Nous décrivons les classes par la disjonction des différences. Au sens strict, cela impose de créer un terme même si les deux valeurs d'attribut ne diffèrent que d'une unité. Ceci ne tient pas compte de la variabilité biologique et nous avons fixé une marge m_d de création de terme logique qui soit en rapport avec le coefficient de variation moyen pour des mesures concernant le même état temporel.

Ainsi, si les éléments sont décrits par une suite de variables x_i, pour un noyau e_i, dont les valeurs sont r_i^k, la description $G(e_1 \mid e_2)$ sera la disjonction des termes logiques $[x_k \neq r_2^k]$ tels que :

$$r_2^k \notin [r_2^k - m_d ; r_1^k + m_d].$$

De même, on considérera que la formule logique $[x_k \neq r_2^k]$ décrira les éléments qui se trouvent à l'extérieur d'un intervalle dont le centre est donné par la valeur r_2^k et la marge de filtrage m_f, soit :

$$x \notin [r_1^k - m_f ; r_1^k + m_f].$$

La description d'un noyau e par rapport aux autres s'écrit :

$$\wedge_i \, G(e \mid e_i).$$

L'évaluation des formules de partition

De telles formules logiques peuvent atteindre des tailles très importantes nécessitant d'effectuer des simplifications qui peuvent se révéler coûteuses.

À chaque étape de la production de conjonctions, nous élaguons la formule en évaluant chacune des conjonctions et en ne gardant que les p meilleures. Ceci empêche bien sûr toute optimalité de l'ensemble

final, mais on ne peut procéder autrement, car il y aurait au total une disjonction de plus d'un million de formules conjonctives.

Les méthodes d'évaluation telles que la simplicité des formules ou la discrétion ("sparseness") de l'ensemble décrit sont trop générales pour notre problème. C'est pourquoi nous avons conçu deux types de fonctions d'évaluation. Le premier type mesure la qualité des formules en cours de construction afin de limiter leur nombre et le second mesure la qualité de la partition.

1. Le premier type ne pourra porter que sur la formule en construction et éventuellement les classes précédemment calculées. Ici, nous avons jugé trop lourd de prendre en compte le contexte précédent et nous n'évaluons que la formule en cours. Pour cela nous avons bâti un indice composite qui est la somme pondérée de mesures sur l'ensemble décrit par chacune des conjonctions. Nous faisons en sorte que chacun de ces indices croisse avec la médiocrité de l'ensemble décrit (au besoin en prenant l'inverse de la fonction) et nous le divisons par son minimum sur toute la disjonction.

Si k est le nombre des indices et ω le facteur de pondération, ceci nous donne :

$$c = \sum_{i=1}^{k} \omega_i \cdot \frac{indice_i(conj_j)}{\min\ (indice_i(conj_j))}$$

Comme nous traitons des données numériques, nous avons repris le critère de la somme des écarts-type des valeurs prises par les éléments de l'ensemble décrit divisés par la moyenne pour chacune des variables :

$$\sum_{i=1}^{l} \frac{\sigma_i}{m_i}$$

l étant le nombre de variables

Nous avons ensuite défini deux indices différents selon qu'on traite les protéines ou les gels.

- Pour les protéines nous avons voulu privilégier de nouveau les formules qui agglomeraient les protéines les plus fiables, c'est à dire comportant le moins de variations par groupes de trois instants. Le deuxième critère est la moyenne de la somme des écarts-type de chacun des quatre groupes des trois valeurs des protéines définies par la formule logique.

- Pour les gels nous savions quel était le nombre des éléments de chaque classe et nous avons repris l'indice donné par [appel] :

$$\frac{n}{k} \cdot n_c$$

où n est le nombre de gels, k le nombre de classes et n_c le nombre de gels observés dans la classe.

2. La seconde fonction d'évaluation s'associe à l'algorithme éliminant les éléments inclus dans des intersections. À chaque germe, nous avons associé une disjonction formée des meilleures formules conjonctives. La partition est créée en prenant une de ces formules pour chaque germe et en déterminant les éléments décrits par ces formules. Nous considérons ensuite les cœurs de ces groupes, c'est à dire leurs éléments propres (qui ne sont inclus dans aucun autre groupe). Les éléments qui sont l'objet d'intersections sont rassemblés puis l'un après l'autre rattachés au groupe le plus proche, c'est à dire, ici, au groupe qui minimise le critère :

$$\sum_{j=1}^{k} \sum_{i=1}^{l} \frac{s_{ij}}{m_{ij}}$$

où l est le nombre de variables, k le nombre de classes et m_{ij} et σ_{ij} sont respectivement la moyenne et l'écart-type de la variable i dans la classe j.

IV. LES RÉSULTATS

Nous avons procédé à deux types de classements : le premier suivant les protéines, par forme de courbe et le second suivant les gels. Au préalable de tout classement, notre méthode nous impose de donner le nombre de classes. On pourrait imaginer qu'une protéine intervenant dans la maturation de la cellule sera croissante au cours du temps ou qu'elle décroîtra, si elle joue un rôle majeur à sa genèse. Lorsque nous classons les protéines par forme de courbes, on pourrait ainsi déterminer, au jugé, le nombre de classes fonctionnelles. En fait, tout n'est pas aussi clair, en particulier à cause de la grande variabilité biologique. Les essais auxquels nous avons procédés, avec différentes valeurs de nombres de classes, ne permettent pas de partitionner, dans l'état actuel de l'implantation de la méthode, avec suffisamment de fiabilité l'ensemble de départ. Certains groupes sont stables au cours des itérations, certains autres ne le sont pas. Dans l'état actuel de notre travail, les résultats obtenus n'ont pas donné entière satisfaction. Ceci est sans doute dû au fait que nous n'avions pas d'idée suffisante sur ce que nous devions trouver. De plus, il n'y avait pratiquement aucune

connaissance *a priori*, ce qui nous imposait presque une analyse combinatoire.

En ce qui concerne le classement par gel, nous connaissions les résultats auxquels nous devions parvenir : quatre classes de trois éléments et évaluer les capacités de l'algorithme. Il a pu retrouver sans erreur les classes ce qui très encourageant. La difficulté résidait dans la longueur des formules logiques et le temps nécessaire à leur traitement, car nous faisions face à un très grand nombre de combinaisons possibles, à chaque étape de la construction de la disjonction de conjonctions. Nous avons donc été sélectif lors de l'étape de filtrage, c'est à dire que nous avons pris une marge importante m_f, ce qui n'a laissé pratiquement que des termes de descripteurs avec les valeurs 0, 1, 2 ou 8, 9, 10.

Pour chaque gel nous avons obtenu une disjonction de complexes de longueur 1, 2 ou 3 dont voici un exemple pour la première catégorie de gels[1] (cellules embryonnaires) :

Formules de longueur 1

((prot8001 ≠ 0)) ∨ ((prot7612 ≠ 0)) ∨ ((prot7004 ≠ 0)) ∨ ((prot8108 ≠ 0))

Formules de longueur 2

((prot7409 ≠ 0) (prot8502 ≠ 0)) ∨ ((prot7409 ≠ 0)(prot8304 ≠ 0)) ∨
((prot7409 ≠ 0) (prot8005 ≠ 0)) ∨ ((prot7409 ≠ 0) (prot7706 ≠ 10)) ∨
((prot7409 ≠ 0)(prot7705 ≠ 10)) ∨ ((prot7409 ≠ 0) (prot7704 ≠ 0)) ∨
((prot7409 ≠ 0) (prot7621 ≠ 0)) ∨ ((prot7409 ≠ 0)(prot7503 ≠ 10)) ∨
((prot7408 ≠ 0) (prot7409 ≠ 0)) ∨ ((prot7404 ≠ 10) (prot7409 ≠ 0)) ∨
((prot7207 ≠ 10)(prot7409 ≠ 0)) ∨ ((prot7201 ≠ 0) (prot7409 ≠ 0)) ∨
((prot6004 ≠ 6) (prot7409 ≠ 0)) ∨ ((prot5305 ≠ 2)(prot7409 ≠ 0)) ∨
((prot7409 ≠ 0) (prot5211 ≠ 0)) ∨ ((prot7409 ≠ 0) (prot4605 ≠ 10)) ∨
((prot7409 ≠ 0) (prot4414 ≠ 10)) ∨ ((prot4309 ≠ 0) (prot7409 ≠ 0)) ∨
((prot303 ≠ 10) (prot306 ...

Au total 77 formules conjonctives.

Les termes de longueur 3 sont similaires aux précédents.

(prot-x ≠ 0) s'interprète comme : la protéine x est présente sur ce type de gel et (prot-x ≠ 10) signifie que la protéine x est absente de ce type de gel.

---

[1]Le système PDQUEST numérote automatiquement les protéines pour le traitement qu'il effectue. C'est une référence interne. Elle ne correspond pas à une nomenclature internationale. Nous avons repris ce numéro préfixé par prot.

V. CONCLUSION

L'algorithme que nous avons présenté permet de retrouver les étapes d'une série "temporelle" de gels d'électrophorèses bidimensionnelles et d'en donner des explications symboliques. Cette analyse s'est faite simplement sur des nombres avec un ensemble minimal de "connaissance ajoutée". Ceci empêche notamment le développement important d'une étape de généralisation.

Une des voie d'évolution sera d'introduire une partie de cette connaissance dans notre système. On peut en effet établir des classes entre les protéines. En particulier, les protéines de cellules musculaires peuvent se classer en groupes tels que : les protéines de l'appareil contractile, des enzymes musculaires, et même les protéines de choc thermique... Certaines de ces protéines sont identifiées ainsi qu'une partie de leurs propriétés, on pourra donc guider les partitions futures en fonction de ces protéines spécifiques, rassembler celles qui ont des comportements similaires ou au contraire étudier les comportements symétriques.

RÉFÉRENCES

[appel] R.D. Appel. *Melanie. Un système d'analyse et d'interprétation automatique d'images de gels d'électrophorèses bidimensionnelles.* Thèse de doctorat. Université de Genève, 1987.

[garrels] J.I.Garrels, J.T. Farrar et C.B. Burwell IV. *Two-Dimensional Gel Electrophoresis of Proteins.* Ch. 2, Academic Press, 1984.

[michalski] R.S. Michalski et R.E. Stepp. Learning from observation: conceptual clustering. R.S. Michalski, J.G. Carbonnell et T.M. Mitchell éd. *Machine Learning*, ch. 4, Springer Verlag, 1984.

[nugues] P. Nugues. *Interprétation de gels d'électrophorèses bidimensionnelles.* Thèse de l'université de Nancy I, 1989.

[ofarrell] P.H. O'Farrell. High resolution two-dimensional gel electrophoresis of proteins. *Journal of Biological Chemistry*. 250:4007-4021, 1975.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique