N° d'ordre:

THÈSE

présentée à

L'UNIVERSITÉ DE NANCY I, FACULTÉ DES SCIENCES

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITÉ DE NANCY I EN INFORMATIQUE

par

Pierre NUGUES

Sujet de la thèse :

INTERPRÉTATION DE GELS D'ÉLECTROPHORÈSES BIDIMENSIONNELLES

Soutenue le 21 Mars 1989 devant le jury composé de :

MM.	**R. MOHR**	Président
	Jean-Laurent MALLET	Rapporteurs
	Christian PELLEGRINI	
	Jean-Paul HATON	Examinateurs
	Michel CLERGET	
Mlle	**Marie-Madeleine GALTEAU**	
M.	**Philippe GARDERET**	

N° d'ordre:

THÈSE

présentée à

L'UNIVERSITÉ DE NANCY I, FACULTÉ DES SCIENCES

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITÉ DE NANCY I EN INFORMATIQUE

par

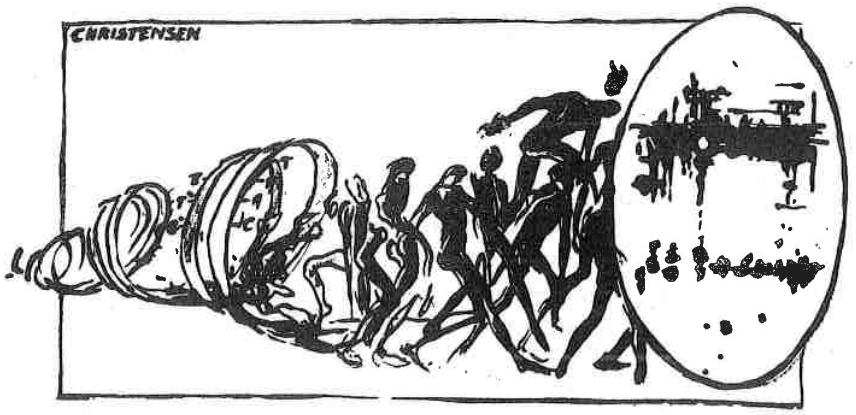
Pierre NUGUES

Sujet de la thèse :

INTERPRÉTATION DE GELS D'ÉLECTROPHORÈSES BIDIMENSIONNELLES

Soutenue le 21 Mars 1989 devant le jury composé de :

MM.	R. MOHR	Président
	Jean-Laurent MALLET	Rapporteurs
	Christian PELLEGRINI	
	Jean-Paul HATON	Examinateurs
	Michel CLERGET	
Mlle	Marie-Madeleine GALTEAU	
M.	Philippe GARDERET	



In memoriam Anita-Marie Galoppin
Tout une vie de peines et de prières
Par respect pour son souvenir.

Cette thèse est bien sûr le fruit d'une collaboration et elle n'aurait pu voir le jour sans l'aide de ceux qui m'ont dirigé et financé :

Jean-Paul Haton, Professeur à l'université de Nancy I, qui m'a accueilli dans son équipe, encadré et donné les moyens matériels pour effectuer cette thèse.

Liliane de Lassus, Président de Cognitech et Michel Clerget, Directeur des études de Cognitech, qui ont bien voulu me faire confiance et financer ce travail.

Roger Mohr, Professeur à l'université de Grenoble, pour ses conseils et l'amitié qu'il me fait de présider ce jury.

Marie-Madeleine Galteau, Professeur à l'université de Nancy I, pour l'appui donné à cette entreprise.

Jean-Laurent Mallet, Professeur à l'INP de Lorraine, et Christian Pellegrini, Professeur à l'université de Genève, qui ont accepté de juger mon travail.

Josiane Steinmetz, chercheur au Centre de médecine préventive, dont l'expertise a permis la mise en œuvre du projet.

Robert Whalen, chercheur à l'Institut Pasteur qui m'a confié les premières données que j'ai pu analyser.

Karl Tombre et Gérald Masini, qui m'ont fait partager leurs connaissances et leur entousiasme pour les langages C et Lisp.

Enfin cette liste de remerciements ne serait pas complète sans la mention de Philippe Garderet, qui, grâce à ses qualités de cœur et d'esprit, a guidé mes premiers pas dans le domaine de l'électrophorèse bidimensionnelle, et qui a accepté de participer à ce jury.

À mes camarades "thèsards", aussi, pour toutes les discussions, Yves Laprie, Marie-Odile Berger, Philippe Hoggan, Noëlle Carbonell en particulier, ainsi que Gong YiFan et Quan Long qui ont éclairé mon esprit sans subtilité d'occidental.

Résumé

Cette thèse a porté sur l'étude et la réalisation d'un système d'interprétation d'images et d'apprentissage symbolique.

Le système d'interprétation d'images identifie automatiquement, sur un gel d'électrophorèse, des protéines, isolées ou à l'intérieur de constellations, en reproduisant les méthodes d'experts biologistes. Il se fonde sur une architecture modulaire comprenant des procédures de traitement d'images conduisant à l'extraction des paramètres et d'un processus de raisonnement ascendant et descendant. Ce processus fait d'abord correspondre les paramètres extraits aux modèles géométriques potentiels des protéines puis revient sur l'image pour déterminer les éléments éventuellement manquants. Il est précédé d'une focalisation de l'attention. Ce système a été appliqué avec succès aux apoliporotéines du plasma.

Le système d'apprentissage symbolique a pour objectif de fournir une interprétation de séries d'expériences et s'inspire de Cluster-2 de R. Michalski. Sa description est précédée d'une étude et d'une comparaison de méthodes symboliques et numériques ainsi que de l'exposé de deux types d'amélioration concernant la généralisation et l'évaluation. Le système d'apprentissage permet de classer les gels correspondant aux étapes de la croissance de cellules musculaires.

Mots clés : électrophorèse bidimensionnelle, interprétation d'image, traitement d'image, apprentissage symbolique, classification conceptuelle, intelligence artificielle.

Summary

This dissertation studies and realizes a system for image interpretation and conceptual clustering.

The image interpretation system automatically identifies proteins on an electrophoretic gel by reproducing methods of biological experts whether isolated or inside constellations. This system is based on a modular architecture featuring image processing procedures which allows extraction of parameters and a top-down and bottom-up reasoning process. First the process matches extracted parameters to possible geometric models of proteins, it then returns to the

image to determine possibly missing elements on the gel. This system was successfully applied to plasma apolipoproteins.

The conceptual clustering system is aimed at supplying an interpretation of a series of experiments and is inspired by Cluster-2 of R. Michalski. Its description is preceded by a study and a comparison of symbolic and numerical methods and by the presentation of two types of improvements which concern generalization and evaluation. This learning system allows the classification of gels corresponding to stages in the growth of muscular cells.

Table des matières

1
Introduction

Notre travail porte sur *l'interprétation* des images d'électrophorèses bidimensionnelles. Ce terme recouvre, dans le vocabulaire scientifique, un concept multiforme, car il peut s'appliquer aux images elles-mêmes, aux états physiologiques ou pathologiques qui les sous-tendent ou bien encore aux conclusions qu'on peut tirer d'un classement de ces images.

Ce mémoire débute par une description succincte du contexte biologique, car c'est lui qui a motivé notre travail et notre orientation. Il se situe, c'est notre *credo*, dans l'explosion du savoir sur la vie et nous espérons que la *vulgarisation* des méthodes que nous détaillons puisse contribuer à l'amélioration des connaissances sur le sujet.

Plus précisément, nous replaçons dans ce chapitre les techniques d'électrophorèse dans un environnement biologique et nous donnons au lecteur une perspective d'ensemble qui lui permettra, nous le souhaitons, de juger de l'utilité de notre démarche. Il va sans dire que c'est la complexité des interprétations diverses qui justifie l'exploitation informatique et que même si, pour parvenir aux objectifs que nous nous sommes fixés, nous devons mettre en œuvre de nouvelles techniques de calcul, elles ne se destinent qu'à être *appliquées*.

La première partie de notre étude prend sa source effective dans un stage que nous avons effectué au Commissariat à l'énergie atomique à Grenoble en 1984, dans le cadre d'un projet français qui avait pour but de déterminer les potentialités de l'électrophorèse bidimensionnelle et de défricher son exploitation automatique. À l'époque, les seules réalisations venaient des États-Unis où était née cette technique et concernaient le traitement d'image. On échafaudait des constructions grandioses de bases de données destinées à contenir la description de toutes les protéines humaines, voire animales ou végétales. Au cours de ce stage, nous avons étudié l'adaptation du logiciel américain TYCHO qui nécessitait un processeur vectoriel, à un matériel sans ce type d'extension.

L'exploration de cette voie pourrait sembler désuète aujourd'hui du fait de l'extraordinaire banalisation des capacités de calcul informatique. Elle nous a cependant permis d'entrer de plain-pied dans le domaine du traitement d'image qui est fondamental en l'électrophorèse bidimensionnelle. Les bases de données sur les protéines n'ont pas encore eu l'extension qu'on leur prédisait, sans doute à cause de la grande complexité du sujet mais plusieurs équipes poursuivent ce travail et aident ainsi à alimenter ces bases de données.

Dans cette partie, nous reprenons notre travail sous un angle original, grâce à des méthodes d'intelligence artificielle et des systèmes à base de connaissances. Nous décrivons une utilisation plus rationnelle de ces techniques de traitement que de nombreux auteurs ont publiées dans la littérature scientifique et nous donnons les lignes directrices de ce que pourrait être l'association du traitement d'image à la mise en œuvre de bases de données, de connaissances et de l'expertise humaine pré-existante.

De manière plus précise, nous évaluons d'abord les procédures de "bas niveau", executées sur l'image : morphologiques, fréquentielles ou autres. Nous traitons ensuite des difficultés et des erreurs qui interviennent dans la mise en correspondance directe d'images de gels. Ensuite, nous présentons une nouvelle méthodologie pour tirer parti de l'expertise des laboratoires biologiques en ce qui concerne l'identification des protéines sur les gels ainsi que les conclusions qui en résultent sur l'état des patients. Cette partie propose une démarche originale de contruction de bases de données. Nous donnons un exemple d'application de cette méthode pour identifier les apolipoprotéines du plasma. En conclusion de cette partie, nous associons deux types d'interprétation :

- l'une se bornant à l'identification d'éléments particuliers sur une image,
- l'autre, à partir de l'identification des éléments visibles, déduit le contexte qui les régit.

La seconde partie est venue de la nécessité d'accroître automatiquement la connaissance au-delà des limites des découvertes humaines. Elle fournit plusieurs méthodes d'interprétation électrophorétique fondées sur une description automatisée (sous forme de règles) des classes résultant d'une description symbolique.

Nous décrivons d'abord des techniques numériques classiques d'analyse des données. Nous reconsidérons ensuite la représentation et le traitement de ces données sous l'angle de l'intelligence artificielle ce qui nous conduit à :

- présenter diverses méthodes d'apprentissage symbolique et de classification conceptuelle existantes. Ces méthodes se fondent sur des techniques encore en pleine investigation et dont les résultats, en terme d'implantation, ne sont pas, à ce jour, totalement probants ;

- puis décrire un certain nombre d'améliorations relatives aux techniques de généralisation et d'évaluation ;
- appliquer les outils d'apprentissage élaborés à des courbes de variation de la production de protéines au cours de la croissance musculaire ;
- enfin, proposer une méthode de dépouillement d'enquêtes épidémiologiques.

2
Qu'est ce que l'électrophoèse

2.1 Les fondements biologiques : le protéines

Dans ce chapitre, nous n'avons pas l'intention de présenter de manière exhaustive les théories de la biologie cellulaire, mais de situer notre travail dans le cadre plus global qui l'a inspiré. Afin que cette thèse n'apparaisse pas seulement comme l'agencement de diverses techniques informatiques réalisé avec plus ou moins d'habileté, nous avons tenté d'exposer les grandes lignes qui permettent de comprendre la signification des "buvards tachés" que sont les gels d'électrophorèses bidimensionnelles au premier coup d'œil. Ceci représente l'unique ambition des paragraphes qui suivent. Le lecteur soucieux de détails pourra se reporter à [Lewin 87].

Les protéines sont les composants essentiels de tout être vivant. Chaque protéine est formée de la liaison d'une ou plusieurs chaînes polypeptidiques qui elles mêmes sont constituées de la succession parfaitement définie de dizaines ou centaines d'acides aminés. Ces protéines peuvent aussi s'associer à d'autres composants, tels, par exemple qu'un atome de fer dans le cas de l'hémoglobine.

Leur poids moléculaire varie entre 6.000 et 1.500.000 daltons. Elles ont une structure tridimensionnelle bien définie, qui dépend en partie des propriétés chimiques des acides aminés qui la constituent et de modifications post-traductionnelles : glycosylation, méthylation... Ainsi à l'état naturel, une protéine présente, sous des conditions données, des propriétés caractéristiques. En solution, les protéines se présentent sous des formes globulaires, l'enchaînement des acides aminés étant replié de nombreuses fois. De plus, des liaisons peuvent se former entre des chaînes polypeptidiques différentes à l'intérieur d'une chaîne peptidique et maintiennent ainsi l'édifice dans une (ou des) conformation(s) spécifique(s) et stable(s).

Une des propriétés importantes des protéines, pour ce qui va suivre, est le point isoélectrique. Chaque acide aminé possède à ses extrémités un groupe COO^- et un groupe NH_3^+. Il agit comme un tampon à la fois dans le domaine acide et dans le domaine basique. Dans une solution acide, l'acide aminé capte un proton (un ion H^+), en saturant son groupement COO^- il gagne ainsi des charges positives grâce au groupement NH_3^+ et en présence d'un

champ électrique, il migrera donc vers la cathode. Dans une solution basique, il cèdera un proton aux ions OH^- environnants à partir de son groupement NH_3^+, qui deviendra NH_2, l'acide aminé sera alors globalement négatif, à cause de son groupe COO^- et il migrera vers l'anode dans un champ électrique. Il existe un pH caractéristique de chaque acide aminé pour lequel celui-ci reste neutre, c'est à dire ne capte, ni ne cède de proton et qu'on appelle le point isoélectrique. Dans un champ électrique, l'acide aminé ne migre pas. Cette constante est par exemple de 6,1 pour la glycine. Les protéines, qui sont un agencement d'acides aminés, possèdent des caractéristiques du même type qui découlent de ce raisonnement, mais seuls les acides animés périphériques interviennent dans la définition du point isoélectrique des protéines.

On estime qu'il existe plus d'un million de protéines dont certaines sont très bien connues et d'autres sont encore à découvrir. On connaît la composition de moins de 1% des protéines, cependant certains liquides biologiques tels que le plasma sont mieux élucidés et leurs protéines mieux caractérisées.

Parmi les méthodes qui permettent d'analyser et éventuellement d'identifier les protéines dans un mélange complexe, l'électrophorèse bidimensionnelle occupe une place de choix puisqu'elle permet, en une seule manipulation, d'avoir une vue d'ensemble de la quasi-totalité des protéines qui composent un tissu ou un liquide biologique. Elle interviendra sans doute dans le futur de manière complémentaire avec un autre procédé qui est celui de l'analyse directe de l'ADN [Landegren 88]. C'est, semble-t-il *la* technique d'avenir dans de nombreux domaines, car elle offre notamment une signature directe de l'expression du génome pour n'importe quelle matière vivante. Cependant, son avenir dépend largement de son coût, qui est très élevé actuellement, et des possibilités de son automatisation.

Enfin, il existe des tentatives de constitution de banques de données sur les protéines et les tissus qui enregistent pour chaque protéine qu'un laboratoire a analysé son nom, ses caractéristiques physico-chimiques, telles que son poids moléculaire, son pH, son point isoélectrique et sa composition en acide aminés. Un exemple d'une telle banque est le *Human Protein Index* [Anderson 82].

2.2 L'électrophorèse monodimensionnelle

2.2.1 Présentation générale

L'électrophorèse est une technique de séparation, sous l'effet d'un champ électrique, de particules en solution qui se fonde principalement sur les propriétés d'ionisation, donc des charges électriques, des groupes qui composent ces particules. Cette technique est assez ancienne et [Righetti 83] la fait remonter à 1912 au Japon pour la production du glutamate de sodium. Plus couramment, on indique la date de 1941 où Tiselius réalisa une électrophorèse

grâce à une solution d'ampholytes. Cette technique permettait de séparer les protéines d'un mélange selon leurs charges électriques.

Il existe plusieurs types d'électrophorèses. Nous ne considérerons que les deux principales méthodes : l'électrophorèse suivant le point isoélectrique et l'électrophorèse suivant la masse moléculaire. Pour des détails concernant les autres procédés électrophorétiques, on consultera [Righetti 83,Hames 81].

Généralement, on dégrade les protéines des tissus en polypeptides par chauffage et par addition d'urée et de détergent. Même lorsque les protéines sont en solution, dans le plasma par exemple, il faut dissocier les protéines en leurs composants polypeptidiques.

2.2.2 La séparation suivant le point isoélectrique

L'électrofocalisation est une des techniques électrophorétique les plus utilisées. Elle sépare chaque protéine suivant son point isoélectrique et elle s'effectue selon les principes suivants [Righetti 83] :

1. le support sur lequel on opère la manipulation est variable. C'est en général un gel d'acétate, d'agarose ou de polyacrylamide. Le support en polyacrylamide est très courant, on l'obtient par polymérisation sous différentes formes, soit une bande plane, soit un cylindre [Hames 81] ;
2. on ajoute à ce gel des immobilines ou des polyampholytes de transport dont les pH s'étagent régulièrement entre deux bornes ;
3. on applique un champ électrique qui produit un gradient linéaire de pH le long du support ;
4. on introduit l'échantillon de tissu solubilisé, et grâce au champ électrique, les protéines migrent jusqu'au pH qui définit leur point isoélectrique.

2.2.3 La séparation suivant la masse

La séparation suivant la masse s'opère sur un support pouvant constituer un filtre. En principe ce support est un gel de polyacrylamide [Allen 84,Righetti 83,Hames 81]. On forme ainsi un milieu de texture plus ou moins compacte qui variera en fonction de la taille de ses pores. Lors de la migration dans un champ électrique, le milieu opère un frottement et la course de migration pourra s'écrire comme une fonction de la charge et du coefficient de frottement.

L'électrophorèse en présence de *dodécyl sulfate de sodium*[1] sur un gel de polyacrylamide s'effectue selon les principes suivants :

1. qui s'abrège en *SDS* en anglais

1. comme pour la séparation isoélectrique, le support est plan ou cylindrique :
2. on introduit du dodécyl sulfate de sodium dans la solution contenant les protéines puis on la chauffe ;
3. Le SDS dénature les protéines ;
4. il se fixe de manière homogène et régulière le long de la chaîne peptidique. À chaque molécule "accrochée", correspondant une charge négative ;
5. on place l'échantillon à l'extrémité du gel et on applique un champ électrique ;
6. la vitesse de migration est une fonction du nombre de molécules de SDS et donc du nombre d'acides aminés de la protéine et du coefficient de frottement du gel de polyacrylamide ; on admet en effet généralement que la masse est proportionnelle à la longueur de la chaîne et par là au nombre d'acides aminés, ainsi la vitesse de migration se définit comme une fonction de la masse moléculaire et du coefficient de frottement ;
7. on observe une séparation des molécules selon une distance proportionnelle au logarithme de leur masse moléculaire valable dans un large domaine de masses.

2.3 L'électrophorèse bidimensionnelle

2.3.1 La méthode de séparation

L'électrophorèse bidimensionnelle combine deux techniques de séparation électrophorétique, suivant deux directions orthogonales. [Righetti 83] en cite un très grand nombre, mais une seule est réellement répandue. Elle correspond à la double séparation des protéines solubilisées d'un tissu suivant le point isoélectrique et la massse moléculaire. C'est en 1975 que O'Farrell a déterminé les conditions de cette manipulation [OFarrell 75]. Il a combiné ainsi deux des techniques les plus résolutives.

Les principes de cette manipulation restent aujourd'hui les mêmes. On utilise donc ces deux mêmes méthodes de séparation, cependant de nombreuses variantes ont vu le jour. Parmi ses améliorations les plus connues, on peut citer [Garrels 84,Anderson 77]

La première séparation s'effectue par électrofocalisation suivant la même méthode que pour une électrophorèse monodimensionnelle, avec un tube d'un vingtaine de cm de long ou sur une bande de la même longueur. On expulse ensuite le gel cylindrique résultant du tube capillaire en présence d'une solution de SDS ou on découpe la bande et on place le polymère sur le bord supérieur d'un gel plan de polyacrylamide d'environ 1 mm d'épaisseur, moulé entre deux plaques de 20 cm de large sur 20 cm de long environ. On maintient ces deux gels fixés l'un contre l'autre par une substance conductrice telle que l'agarose et on opère à la séparation suivant la seconde dimension en appliquant un champ électrique orthogonal au

premier gel. La migration s'effectue de manière totalemant indépendante de la précédente. On obtient ainsi un gel plan dont les dimensions sont d'environ 20 cm × 20 cm, transparent où les protéines se distribuent suivant leur pI et leur poids moléculaire.

Pour ces deux migrations, il est possible d'établir une échelle en introduisant des protéines standards de caractéristiques connues.

2.3.2 Les colorations

Après la séparation, les protéines ne sont pas visibles à l'œil nu et on doit donc les révéler par l'intermédiaire d'une substance chimique ou radio-active. On dispose pour ceci de plusieurs méthodes.

1. La première méthode est la coloration au *bleu de Coomassie*. Sa sensibilité est assez faible. Il faut donc que la protéine soit en assez grande quantité dans le tissu traité pour qu'elle devienne visible. L'intérêt de cette coloration est d'être linéaire avec la quantité de protéine et relativement stable pour toutes les familles de protéines.

2. Une seconde méthode de coloration est la révélation au nitrate d'argent. Elle est beaucoup plus sensible que la précédente (100 fois plus environ) pour la plupart des protéines, mais l'intensité de la coloration est plus délicate à interpréter, car elle dépend à la fois de la quantité de protéine à révéler et des propriétés intrinsèques de cette protéine à fixer l'argent. Ces propriétés sont mal connues. Certaines protéines sont réfractaires à cette coloration, c'est le cas de l'orosomucoïde sur les gels de plasma sanguin, alors qu'elle se révèle bien au bleu de Coomassie.

3. l'autoradiographie s'utilise pour les protéines tissulaires. On marque les protéines par l'intermédiaire d'éléments radio-actifs, tels que le soufre 35 ou le carbone 14, qui se substituent aux mêmes éléments – aux isotopes près – des protéines. La révélation se fait par l'exposition d'une plaque sensible aux rayons émis par le gel. L'intensité du rayonnement est proportionnel aux éléments radio-actifs insérés dans la protéine. On postule que le nombre d'éléments radio-actifs substitués est à peu près proportionnel au nombre total d'atomes pour chaque protéine. La quantité de radio-activité émise correspond donc à la quantité de chaque protéine. Cette dernière méthode est très sensible mais ne peut s'appliquer au plasma, seuls les tissus peuvent être rendus radio-actifs.

4. Enfin, il existe des colorations spécifiques de protéines, pour les glycoprotéines par exemple. [Righetti 83] donne une revue des colorations spécifiques.

Figure 2.1. Aspect d'une électrophorèse bidimensionnelle

2.3.3 L'aspect d'une électrophorèse bidimensionnelle

Une fois révélé, le gel d'électrophorèse se présente sous l'aspect d'un buvard dont les taches sont d'intensité et d'étendue plus on moins grandes. Nous en donnons un exemple sur la figure 2.1 page 14.

De manière classique, les axes sont les suivants : l'abscisse représente l'échelle des points isoélectriques croissants et l'ordonnée représente l'échelle des poids moléculaires croissants. Le domaine d'étude est : pI $\in$ [3 ; 10] et PM $\in$ [10 ; 220.000].

On distingue des points isolés et bien résolus, dont la représentation en niveaux de gris prend la forme d'un dôme, alors que certaines autres protéines se chevauchent. On remarque aussi des traînées horizontales ou verticales le long des axes de migration, au dessus desquelles émergent des pics. La présence de ces traînées est encore mal expliquée. Certaines taches sont nettement plus importantes que les autres et "noient" les taches plus petites ce qui

rend difficile le dépouillement de ces régions. On découvre ainsi les limites de détection de l'électrophorèse bidimensionnelle qui viennent d'une part de la révélation argentique et d'autre part de la présence d'ampholytes (focalisation).

2.3.4 La modélisation

Les taches protéiques sont le fruit d'une migration de molécules et elles se distribuent de manière statistique autour de leur point caractéristique, ici le point isoélectrique et le poids moléculaire. La coloration sur l'image peut se modéliser en terme de probabilités. Au delà de l'intérêt théorique, cette modélisation permet de connaître le volume d'une tache correspondant à une protéine simple, ce qui n'est pas un grand progrès, car il suffit pour cela d'intégrer la tache sur l'image, mais aussi d'une tache de plusieurs protéines se chevauchant, en supposant que les modèles individuels s'additionnent. Par l'intermédiaire de la modélisation, nous pouvons accéder alors au volume de chaque tache particulière.

On a proposé plusieurs modèles. Le modèle le plus couramment utilisé est celui de la gaussienne bidimensionnelle [Taylor 79], dont l'équation est la suivante :

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

où $f(x,y)$ représente l'intensité de gris au point (x,y) ; A, l'amplitude de la gaussienne ; (x_o, y_o), les coordonnées du sommet ; (σ_x, σ_y), les écarts-type suivant les deux axes orthogonaux.

Plusieurs auteurs ont noté une asymétrie des gaussiennes et ont proposé des modèles plus complexes, ainsi [Lutin 79] modélise les taches par des "gaussiennes" munient d'un *skewness* et d'un *kutosis*, c'est à dire dont les termes carrés sont remplacés par des polynômes d'ordre 4. [Vincens 87a] propose un modèle à 6 paramètres qui est une gaussienne bidimensionnelle, dont l'axe principal suit une orientation quelconque. (Dans le cas classique, la gaussienne bidimensionnelle est le produit d'une gaussienne suivant l'axe des x par une gaussienne suivant l'axe des y).

Certains auteurs s'affranchissent de modèles, ainsi [Miller 82] et le système qu'il a inspiré [Funk 87]. La détermination du volume des taches se chevauchant devient alors nettement plus heuristique.

2.3.5 La reproductibilité

L'électrophorèse bidimensionnelle est un technique nouvelle dont les conditions optimales de manipulation sont encore mal définies et de ce fait certains points sont mal maîtrisés. C'est le cas de la reproductibilité des gels, entre les différentes expériences d'un même laboratoire et entre les différents labratoires. Le caractère de reproductibilité est très important, car sur

un gel il peut apparaître jusqu'à 6.000 points identifiables dans des conditions favorables et une mauvaise procédure peut le rendre complétement illisible. Cette reproductibilité a pour limite la variabilité naturelle due aux différences entre les individus et dont l'étude est un des objectifs de l'électrophorèse.

La reproductibilité dépend étroitement du pouvoir de séparation des méthodes [Taylor 83], et on peut s'attendre à des améliorations pour cette partie [Hanash 87], notamment grâce à l'emploi des gradients de pH immobilisés.

Pour la majeure partie des cas, le problème de la reproductibilité reste ouvert. On constate que certains laboratoires obtiennent des gels pratiquement superposables, C'est le cas de J. Garrels à Cold Spring Harbor, aux États-Unis, particulièrement avec des méthodes de révelation par autoradiographie. D'autres sont plus variables, notamment avec les colorations argentiques. Cette dernière coloration s'obtient de manière assez empirique et on s'efforce autant que possible de normaliser les manipulations. Enfin le plasma, milieu sur lequel sont effectuées les électrophorèses bidimensionnelles que nous avons étudiées, constitue le milieu le plus difficile en électrophorèse bidimensionnelle car il contient une protéine majeure, l'albumine, qui représente plus de 50% des protéines totales. L'albumine, par sa présence, perturbe sensiblement la migration des protéines situées dans la même région de pI et de masse.

2.3.6 La quantification

La quantification des taches s'opère par l'intermédiaire de la densité optique de la coloration car c'est la seule donnée qui nous soit accessible sur l'image. À partir de cette quantité, on peut retrouver avec plus ou moins de précision, la quantité de protéine dans le tissu.

Après avoir modélisé les taches, on peut accèder à leur volume. Dans le cas d'une tache simple, il suffit de l'intégrer sur l'image. Dans le cas de protéines qui se recouvrent, on doit extraire les paramètres de chacune – en émettant l'hypothèse d'un modèle linéaire – avant de connaître leur volume.

La quantification en masse dépend étroitement de la méthode de révélation.

1. Avec le *bleu de Coomassie*, il est assez facile de quantifier approximativement les taches. En effet, la fixation de colorant par les protéines est à peu près proportionnelle á la quantité de cette protéine. Il suffit de rapporter l'intégrale des densités optiques à une constante et pour connaître le pourcentage d'une protéine particulière, de la rapporter à l'intégrale des densités optiques sur tout le gel.

2. Avec *l'autoradiographie*, dans le cas des protéines tissulaires, l'émission de rayonnement est aussi proportionnelle aux quantités de matière[2]. On accède donc à la quantité grâce à un rapport avec l'intégrale de la densité optique des points étudiés. On procède donc comme avec le bleu de Coomassie. Cette technique étant beaucoup plus sensible, il apparaît des saturations sur la pellicule qui reçoit les rayons radio-actifs. Pour pallier cet inconvénient, on doit procéder à des expositions multiples, où les protéines les plus grosses apparaissent d'abord sur les premiers films, à ce moment, on modélise leur forme, qu'on interpolera pour simuler le résultat final et ainsi de suite pour les protéines en quantité moyenne et faible, jusqu'aux quantités qui ne sont perceptibles qu'après une exposition prolongée de la pellicule au gel. On créera finalement une image de synthèse où on restituera les gaussiennes sans saturations[3].

3. Avec la *coloration argentique* les problèmes sont plus délicats, car la fixation de l'argent n'est pas proportionnelle à la quantité de matière. On peut seulement effectuer un rapport entre protéines similaires. On doit alors introduire dans l'échantillon à analyser des quantités connues des protéines de calibration qui ont les mêmes propriétés de fixation de l'argent que les protéines qui intéressent le chercheur. Dans le cas de la coloration argentique, on ne peut honnêtement parler que de quantification très approximative.

La quantification est un point délicat et pourtant essentiel pour l'exploitation future de l'électrophorèse bidimensionnelle. Elle est imprécise sur ce point, plus ou moins selon les procédés, et c'est ce qui la fait encore concurencer par d'autres techniques. Pour l'instant on ne peut réellement employer que le terme de *semi-quantification* et nous espérons qu'à l'avenir des progrès notables dans les processus d'obtention interviendront.

2.4 Les applications de l'électrophorèse bidimensionnelle

L'électrophorèse bidimensionnelle a ouvert un champ considérable d'investigations et d'applications dans des domaines aussi différents que la biologie clinique humaine, la pharmacologie, la biologie végétale, l'embryologie... Sur une électrophorèse bidimensionnelle de plasma on peut distinguer plus de 500 taches. Ceci représente un accroissement considérable des moyens d'analyse par rapport à l'électrophorèse monodimensionnelle classiquement effectuée en biologie clinique. Le plasma n'est pas le seul tissu à pouvoir en bénéficier et nous pouvons citer aussi les électrophorèses de foie [Appel 87] qui comptent environ 3.000 taches.

2.4.1 Les applications cliniques

Ce sont certainement les applications cliniques de l'électrophorèse bidimensionnelle qui sont actuellement les plus développées [Tracy 84,Allen 84]. Elle touche presque tous les liquides biologiques et on peut imaginer d'analyser n'importe quel tissu.

2. Communication orale avec J. Klose. Institut für Toxikologie und Embryonal-Pharmakologie, Université libre, Berlin

3. C'est ainsi que procède la firme Protein Databases. Huntington Station. États-Unis

Parmi les liquides biologiques, le plus étudié est le sang, ou ses dérivés : le plasma et le sérum sanguin, car il est d'obtention très facile. On a identifié de nombreuses protéines et on a pu dresser une carte assez précise de ce tissu [Anderson 84]. On a ainsi la liste des taches, leurs caractéristiques sur l'image et leurs principales variantes. On a pu lier certaines de ces protéines ou de ces variantes à divers phénomènes physiologiques ou pathologiques.

[Vincens 87a] a procédé à l'examen de liquides amniotiques de fœtus atteints de deux maladies graves : la mucoviscidose et le spina-bifida. Il pu mettre en évidence des modifications pathologiques bien précises de protéines particulières et les relier à la condition du fœtus.

Par cette technique, on peut déceler en une seule analyse, une série d'états fœtaux, car on pourra sans doute reprendre la méthodologie de P. Vincens pour d'autres exemples, grâce à la vue d'ensemble des protéines sur le gel. Actuellement, on doit avoir recours à un ensemble d'analyses, souvent pénibles, spécifiques de chaque pathologie.

Les tissus solides font l'objet de moins d'études que les liquides, on en trouve pourtant de multiples exemples dans la littérature :

[Appel 87] a déterminé par l'analyse de tissus de foie sain et cirrhotique quelles étaient les protéines qui étaient spécifiques de cette maladie.

[Endler 86] a découvert des protéines jouant un rôle dans le cancer du poumon,

[Wiederkehr 86] établit un diagnostic de la sclérose multiple à l'aide d'électrophorèses bidimensionnelles.

Ces exemples d'applications cliniques montre que les domaines d'action de cette technique sont extrêmenent variés et qu'ils offrent une généralité que ne possède aucune autre analyse biochimique.

2.4.2 L'étude des mutations et de l'évolution

Par la comparaison de gels provenant de différentes lignées de cellules, on peut étudier les mutations génétiques. L'électrophorèse bidimensionnelle permet de détecter deux types de mutations [Neel 84] :

- les substitutions de bases nucléotidiques qui entraînent une synthèse d'acides aminés différents et modifient le poids ou la charge et donc la position géométrique de la protéine.
- Les substitutions qui empêchent l'élaboration d'une protéine.

Elle présente l'avantage de pouvoir remonter à un niveau plus fondamental que la simple observation clinique, car on peut accéder à l'expression génétique et éventuellement à la

structure de l'ADN. On pourra évaluer puis quantifier les risques induits par les environnements toxiques des industries, par les divers traitements médicamenteux, ainsi que par les utilisations d'armes chimiques ou nucléaires.

[Skolnick 86a] a mis au point une série d'algorithmes spécialement destinés à révéler les protéines potentiellement mutantes. Elle implique la comparaison des gels provenant du père, de la mère et d'un des enfants, et recherche les protéines présentes chez l'enfant, qui sont absentes chez les parents.

[Vincens 87a] compare les gels d'espèces de tortues différentes, afin de déterminer les mécanismes de différenciation sexuelles en fonction de la température. En effet, si vous ne saviez pas, le sexe des tortues dépend de leur température d'incubation. On a pu déterminer, le point est d'importance, grâce au gel des gonades de ces animaux, les protéines jouant un rôle dans cette différenciation et leur variation au cours de l'évolution.

2.4.3 Les applications industrielles

Là encore, les applications sont multiples. Dans le domaine de la nutrition, on peut contrôler les apports protéiques des différents aliments en effectuant une électrophorèse bidimensionnelle. Ainsi Marcia Goldfarb a étiqueté toutes les protéines qui apparaissent sur les gels de lait maternel[4]. On pourra dans l'avenir détecter les principales carences protéiques dont souffrent les nourrissons. Ces études pour la nutrition peuvent déboucher vers d'autres applications telles que l'amélioration des espèces végétales comestibles. On peut d'ores et déjà comparer les espèces de blés, de pommes de terre. On peut aussi déterminer l'origine des viandes animales, car les aspects des gels bidimensionnels sont différents, qu'ils s'agissent de chèvres, de bœufs, de moutons, etc. L'électrophorèse bidimensionnelle permet de visualiser les différences entre lignées cellulaires et de ce fait de comprendre le rôle potentiel d'effecteurs chimiques ou physiques, (choc thermique, molécules,...) sur des cellules spécifiques. [Vincens 87a] l'utilise pour déterminer les effets sur une lignée de cellule de rat de la farnesyl acétone. On met en évidence en une seule manipulation les protéines atteintes dans le tissu considéré, dans le même ouvrage P. Vincens, détermine une carte fonctionnelle de cellule à l'aide de divers effecteurs et de techniques d'analyse de données qu'il applique aux gels bidimensionnels.

On peut envisager son application à l'étude de la toxicité de produits chimiques, des cosmétiques..., des effets des médicaments, des drogues,...

L'emploi de l'électrophorèse bidimensionnelle est aussi possible dans le contrôle de qualité et la purification de protéines. L'examen d'un échantillon permet de déterminer exactement quel sont les protéines présentes, de les quantifier, de calculer les variations par rapport aux

4. Présentation sur panneaux. Sixth Meeting of The International Electrophoresis Society, juillet 1988. Copenhague.

normes voulues, et de rejeter les échantillons qui contiendraient des substances indésirables.

PARTIE I

LE TRAITEMENT D'IMAGE D'ÉLECTROPHORÈSE

Dans cette partie, nous abordons le traitement et l'analyse d'images appliqués au domaine de l'électrophorèse bidimensionnelle. Ce traitement, que l'image soit numérique ou analogique[5] fait appel à un nombre considérable de théories mathématiques ou de techniques calculatoires. Ceci provient d'une part de la très grande diversité des sources d'images : médicales, géographiques, spatiales, industrielles... auquelles on doit appliquer des méthodes différentes, et d'autre part d'un certain manque de formalisme unificateur. Ainsi parmi les domaines des mathématiques qu'on se doit de connaître, figurent : l'informatique au premier rang, car c'est elle qui a donné naissance au traitement d'image, mais aussi l'algèbre vectorielle, la théorie des distributions, des systèmes linéaires, des probabilités, de la morphologie mathématique...

Nous tentons de présenter les techniques, voir les théories qui nous permettent de "lire" un gel bidimensionnel d'électrophorèse, sans avoir la prétention d'être exhaustif, ces techniques n'ayant comme fonction que d'être au service d'un but. Comme ouvrage général sur le traitement de l'image numérique, on pourra consulter [Pratt 78].

Le but qui détermine l'emploi des procédures de traitement d'image est de pouvoir identifier de la même manière que le ferait un biologiste ou un médecin les protéines qui peuvent présenter un intérêt quelconque sur un gel. Nous avons d'abord indiqué les défaut des techniques actuelles et nous avons proposé une méthode, que nous espérons être originale, pour résoudre ce problème ou au moins l'envisager sous un autre angle. Pour la mettre en œuvre, nous avons dû faire appel à des procédés d'intelligence artificielle. Nous ne définirons pas ce dernier terme, car il recouvre encore plus de concepts que celui de traitement d'image et si peu de gens sont d'accord pour dire où il commence, tous s'entendent pour lui touver des extensions illimitées. Nous décrivons simplement dans cette partie les techniques utilisées pour mettre en œuvre notre méthode. Nous donnons un exemple d'application avec une famille de protéines du plasma : les apolipoprotéines.

La méthode que nous présentons ouvre de nouvelles perspectives dans l'exploitation future des gels bidimensionnels et notamment dans la formation et la manière d'organiser les bases de données. Pour terminer cette partie, nous exposons ces idées bien que celles-ci n'aient pas, pour le moment, pris la forme d'une réalisation.

5. c'est à dire résultant d'un phénomène continu, et dont l'amplitude en chaque point n'est accessible que par l'intermédiaire d'un dispositif de transformation physique, alors que l'image numérique est toujours discrète, bien qu'on puisse imaginer un système numérique continu

1
Saisie et traitement de l'image

Le traitement informatique de l'image commence par une saisie de données par l'ordinateur. Ces données proviennent d'un phénomène physique, qui est mesuré par un capteur à certains intervalles spatiaux ou temporels et qui sont converties en nombres.[1] Ce capteur peut mesurer plusieurs types d'informations. En général, c'est l'intensité d'un rayonnement sur une surface à des longueurs d'onde variables, ça peut être aussi un déphasage pour une télémétrie, ou encore une mesure thermique. Avant d'atteindre ce capteur, l'image peut subir un certain nombre de transformations optiques. Ces transformations sont souvent similaires à celles qui se produisent dans l'œil et on parle alors de vision monoculaire ou binoculaire[2]. Elles peuvent aussi ne pas exister comme par exemple lorsqu'on acquière une image plane avec un densitomètre, ou bien quand l'ordinateur se charge de reconstruire les images à partir de mesures physiques, comme c'est le cas par exemple dans la tomographie à rayons X.

Nous restituons cette acquisition sous la forme de matrices bidimensionnelles, en considérerant l'amplitude comme une somme de trois fonctions représentants les trois couleurs fondamentales. La restitution s'effectuant dans trois domaines spectraux différents. Nous avons alors :

$$I(x,y) = I_{rouge}(x,y) + I_{bleu}(x,y) + I_{vert}(x,y)$$

Nous n'abordons pas ce type de modélisation dans notre thèse et nous considérons les images à saisir comme des matrices, sans profondeur de champ, qui s'écriront sous la forme :

$$I(x,y)$$

cette quantité désignant la densité optique telle qu'on peut la mesurer au point (x,y).

1.1 La numérisation

La saisie de l'image d'un gel bidimensionnel s'effectue par l'intermédiaire :

1. On ne peut traiter en général l'image par ordinateur qu'après l'avoir rendue sous une forme numérique, quoiqu'il existe encore des formes de traitements analogiques dans certains domaines bien spécifiques tels que celui des signaux vidéos de haute fréquence.
2. Pour les problèmes de formation des images, de perception et de vision, on pourra consulter [Marr 82]

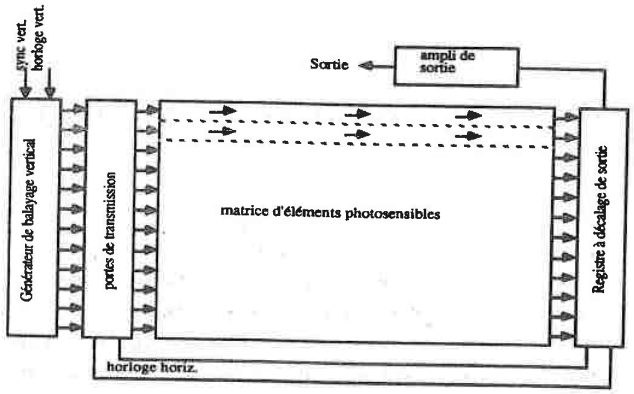


Figure 1.1. Principes des caméras CCD

- d'une caméra optique ou
- d'un densitomètre.

Les seules caméras qu'on utilise désormais, fonctionnent avec des cellules *CCD*[3]. Succintement, le principe de la caméra est le suivant : l'objet vient se former sur le foyer du système optique. À l'emplacement de ce foyer se trouve une plaque sensible composée de cellules CCD de tailles identiques. Chacunes de ces cellules intègre la quantité de lumière qu'elle reçoit à sa surface. Un dispositif électronique numérise chacune de ces quantités et les rentre en mémoire de l'ordinateur. Les dimensions actuelles maximales des matrices CCD sont de 1024 × 1024 cellules. Elles permettent une acquisition d'images à la vitesse du cinéma. Certaines caméras ne disposent que d'une barrette linéaire de cellules CCD, qui balaye le foyer optique de la caméra. On ne les utilise que pour des poses fixes[4], car le balayage est mécanique et dure quelques secondes. Elles permettent, en revanche, des matrices de bien plus grandes dimensions, de l'ordre de 3000 × 3000 cellules. On donne une vue schématique sur la figure 1.1 page 26

3. Abréviation de *Charge-Coupled Device*
4. C'est le cas pour l'électrophorèse

Les densitomètres mesurent point par point à intervalles définis la densité optique du gel. Ils transmettent la valeur en la convertissant par un convertisseur analogique-numérique. Ils nécessitent un dispositif mécanique assez élaboré. La mesure peut se faire à plat et le capteur se déplace au dessus du gel à chaque endroit où il doit prendre une mesure ou en le fixant sur un tambour, le capteur se déplace alors le long d'un axe fixe pendant la rotation du tambour. Dans ce dernier cas, il est préférable de reproduire le gel en le photographiant ou bien en le posant sur une plaque photographique et en l'illuminant pendant un temps donné. On numérise alors le négatif photo.

Dans tous les cas de saisie, on remarque que l'un des facteurs déterminant de qualité est l'éclairage. Deux types sont possibles, l'éclairage par réflexion, dans ce cas, on place la source de lumière et le capteur au-dessus du gel et l'éclairage par transmission où on place le gel entre la source de lumière et le capteur.

À partir de ces dispositifs d'acquisition, on peut procéder à la numérisation des images. Cette numérisation se définit essentiellement par deux paramètres :

- le pas de discrétisation spatiale selon les abscisses et les ordonnées ;
- le nombre maximal de niveaux de numérisation pour chaque échantillon.

Le pas d'échantillonnage spatial permet de définir la valeur minimale d'information fréquentielle qu'on pourra restituer en discrétisant l'image. Le critère de Nyquist établit cette fréquence minimale d'échantillonage au double de celle de la borne du spectre fréquentiel du phénomène continu. Il est nécessaire d'avoir une idée de cette valeur pour procéder à l'acquisition. L'interprétation de la transformée de Fourier d'une image d'électrophorèse prise au hasard d'une expérience est délicate car elle contient de nombreux artéfacts. On peut, par contre, modéliser son information théorique et déterminer quelle est la valeur optimale d'échantillonnage pour conserver cette information. Le modèle que nous avons pris étant le produit de deux gaussiennes suivant les abscisses et les ordonnées, il nous suffit de raisonner dans le cas monodimensionnel.

L'équation d'une gaussienne s'écrit :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{x^2}{2\sigma^2}$$

Sa transformée de Fourier est :

$$F(\nu) = \sqrt{2\pi\sigma^2} \exp -2\pi\sigma^2\nu^2$$

En utilisant cette transformée, on peut calculer l'énergie perdue à partir d'une fréquence de coupure. Il vient :

$$1 - \int_{-\infty}^{fc} F$$

P. Vincens [Vincens 87a, page 35 du document] a calculé les fréquences spatiales d'échantillonnage pour diverses tailles des taches. Pour détecter les taches dont l'écart-type est de 0,2 mm, le pas d'échantillonnage doit être au moins de 10 points par mm, l'énergie perdue est alors de moins de 1%.

On prend les pas d'échantillonnage en général égaux suivant les abscisses et les ordonnées en supposant que les taches protéiques prennent une forme relativement symétrique statistiquement selon les axes.

Une fois déterminée la fréquence spatiale de discrétisation, nous devons quantifier chaque échantillon de l'image. Cette quantification s'opère suivant une loi de codage et avec un nombre de niveaux déterminés.

1. La loi de codage définit comment se transforment les niveaux analogiques à la sortie du capteur en chiffres. La plus commune est la loi linéaire. On utilise aussi d'autres lois, telle que les lois logarithmiques en télécommunications [Bellanger 84]. Elles améliorent notablement le rapport signal sur bruit. Ici, nous supposons que l'influence du bruit est négligeable comparée aux autres perturbations telles que les dérives de coloration, etc. le seul usage qu'on pourrait faire d'une loi de codage non linéaire serait de modifier la lisibilité, ceci est inutile car on peut très facilement effectuer ces opérations une fois l'image dans l'ordinateur par une transformation d'histogramme. D'autre part ce codage a pour effet de modifier la quantification des taches de protéines. Comme nous faisons déjà face à une assez grande imprécision de mesure, il n'est sans doute pas approprié d'introduire des manipulations supplémentaires.

2. Le nombre de niveaux est, en général, de 256. Plus ce nombre est élevé, plus la numérisation est précise, car il limite le bruit d'arrondi. Cette pratique de codage sur un octet est sutout due aux structures internes des ordinateurs et aux convertisseurs analogique-numérique qui existent sur le marché, plus que pour des raisons théoriques. Ce nombre est largement suffisant pour la lecture et [Pratt 78] donne 64 niveaux comme valeur nécessaire en se fondant sur des expériences psychologiques. [Vincens 87a,Vincens 86b] justifie ce choix en calculant l'entropie d'une image d'électrophorèse numérisée avec différentes valeurs de niveaux. Il observe une dégradation en dessous de 64 niveaux.

Jusqu'à présent, on a étudié l'optimalité de la numérisation en considérant la discrétisation puis la quantification des échantillons de manière séparée. En fait, ces notions ne sont pas indépendantes et pour restituer une même qualité de lecture avec une quantification binaire, noir ou blanc, il faut une fréquence d'échantillonnage beaucoup plus importante qu'avec une quantification sur 256 niveaux de gris. [Lee 87,Pavlidis 82] donnent des méthodes de calcul pour optimiser, au sens d'une erreur, l'occupation en mémoire. Ils déterminent ainsi une fréquence de discrétisation et une plage de numérisation. En fait, cette démarche, de même si elle améliore celle de Nyquist ne tient compte que de quelques facteurs alors que la qualité

de la perception des informations par l'œil répond sans doute à d'autres facteurs tels que la netteté des contours.

Pour les réalisations de notre travail, nous avons pris des valeurs au-dessus des seuils optimaux, aussi bien pour le pas de discrétisation spatiale que pour le nombre de niveaux de codage.

Une fois l'image rentrée dans l'ordinateur, on la restitue sur un dispositif de visualisation, tel par exemple qu'un écran. Il existe plusieurs façon de procéder à cet affichage.

1. Si on dispose d'un sytème capable d'afficher les niveaux d'acquisition, le cas est trivial et on a pas besoin de procéder à un codage de sortie ;

2. si le système d'affichage ne possède que deux niveaux de sortie, généralement noir et blanc, comme par exemple pour les écrans bit-map et les imprimantes à encre noire, on doit simuler les niveaux par un codage approprié : un tramage. Le principe général en est le suivant, on applique une trame sur l'espace où s'affichera l'image de sortie de mêmes dimensions que l'image originale en niveaux à restituer. Chacun des intervalles de la trame est constitué d'une matrice de points élementaires qui prennent la valeur noir ou blanc. Le tramage consiste dans le remplissage de ces carreaux de trame – plus ou moins selon le niveau d'origine – suivant un algorithme approprié. Ces algorithmes sont très nombreux. On pourra consulter [Ulichney 88]. Nous avons choisi de reproduire un algorithme défini par Knuth [Knuth 87] toutes les fois que l'image s'affiche sur un écran bit-map.

3. Bien que la saisie des gels d'électrophorèse se fasse de manière monochromatique, on peut la restituer en couleurs. Il s'agit alors de pseudo-couleurs. On associe à un niveau de gris, une couleur particulière choisie sur une palette adéquate. Ce choix s'effectue suivant des critères psychologiques et subjectifs et on peut s'interroger sur la pertinence de cette manière de faire. Dans notre système, nous avons introduit cette possibilité de codage. Nous avons fondé notre choix de couleurs sur une analogie avec les cartes topographiques. En effet, on pourrait comparer les pics de protéines à un relief où les pseudo-couleurs se dégradent du bleu, pour les niveaux bas, au rouge, pour les niveaux hauts, en passant par le vert et l'orange.

1.2 L'amélioration de l'image

Ce sujet a été très abordé dans la littérature, aussi bien générale [Wang 83,Hahn 84, Pratt 78] que concernant l'électrophorèse bidimensionnelle [Skolnick 82b,Vincens 86a]. Cette amélioration résulte de la séparation du signal utile : ici les taches protéiques, des signaux parasites. Ces derniers nous sont connus par l'intermédiaire de leurs propriétés statistiques

comme pour le bruit ou par une certaine modélisation. On distingue en général quatre types de perturbation sur les gels :

1. les pertubations dues aux défauts dans les processus d'obtention ou de manipulation du gel, telles que la mauvaises migrations le long des axes : en biais ou incomplètes, les bulles d'air, les fêlures ou les écornures qui sont la conséquence de la grande fragilité des gels bidimensionnels.
2. L'inégalité du niveau du fond. En effet celui-ci subit des dérives qui le réhausse ou l'abaisse par endroit.
3. Les traînées horizontales ou verticales qui suivent les directions de la migration.
4. les bruits de haute fréquence sans doute dûs aux dipositifs électroniques de mesures.

Il n'est pas raisonnable de tenter de restaurer les gels abimés par le premier type de perturbations. Celles-ci ne sont pas inhérentes aux gels mais seulement dues à des fautes dans les procédures de développement biologique. On doit simplement chercher à les éviter. Pour les autres séries de perturbations, nous décrivons et nous critiquons un ensemble de méthodes.

1.2.1 La morphologie mathématique

La morphologie mathémathique [Serra 82,Serra 86] est la méthode la plus utilisée pour l'amélioration d'images d'électrophorèses [Skolnick 82b,Skolnick 86a,Vincens 87a,Vincens 86a]. Elle s'applique à des images binaires et en niveaux de gris. Elle met en œuvre deux opérations principales simples : l'érosion et la dilatation [Serra 86]. Les opérandes sont d'une part la forme à traiter et d'autre part la forme de référence du filtre : *l'élément structurant.* Dans le cas des images binaires, on définit la forme à traiter par des 1, le fond par des 0 et l'élément structurant par des 1.

Soit X la forme et B_x l'élément structurant de centre x. Les opérations se décrivent de la manière suivante :

- l'érosion de X est l'ensemble de points :

$$X \ominus B = \{x : B_x \subset X\}$$

  Nous donnons un exemple d'érosion figure 1.2 page 31

- la dilatation de X est l'ensemble des points :

$$X \oplus B = \{x : B_x \cap X \neq \emptyset\}$$

  Nous donnons un exemple de dilatation figure 1.3 page 31

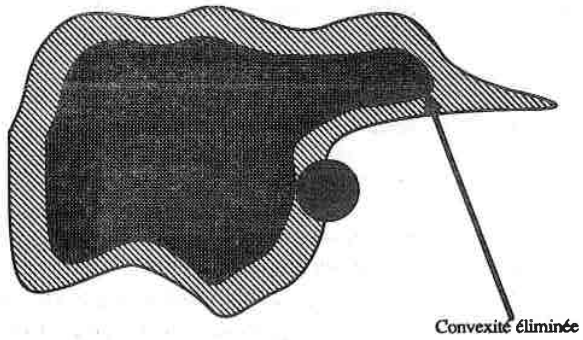


Figure 1.2. Érosion

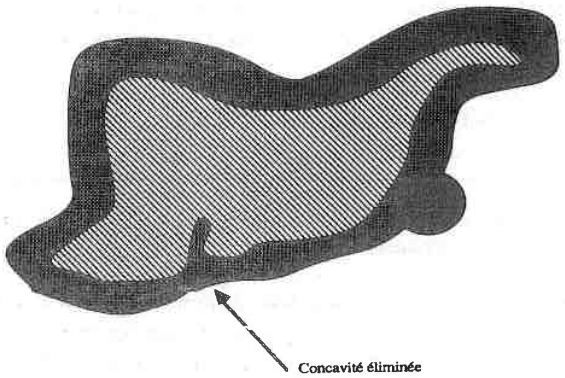


Figure 1.3. Dilatation

S. Sternberg [Sternberg 86] a étendu ces deux opérations au domaine des images à niveaux de gris. Les définitions pour une image décrite par la matrice $a(x,y)$ et pour un élément structurant $B(x,y)$ centré en $(0,0)$ deviennent :

- la dilatation $d(x,y)$ de l'image $a(x,y)$ s'écrit :

$$d(x,y) = \max_{i,j}[a(x-i, y-j) + b(i,j)]$$

- l'érosion $e(x,y)$ de l'image $a(x,y)$ s'écrit :

$$e(x,y) = \min_{i,j}[a(x-i, y-j) - b(-i,-j)]$$

S. Sternberg a par ailleurs définit des algorithmes pour créer itérativement, à partir de matrices 3×3, un certain nombre d'éléments structurants.

Les deux opérations d'érosion et de dilatation se combinent aussi bien dans le domaine des images binaires que des images en niveau de gris pour réaliser les ouvertures et les fermetures de l'image.

- L'ouverture de l'ensemble X par l'élément structurant B se définit par :

$$X_B = [X \ominus B] \oplus B$$

  Cette opération se traduit concrètement par l'élimination dans l'image des parties convexes des formes qui ne peuvent pas contenir l'élément structurant. En niveaux de gris, l'image peut se considérer comme une nappe, sous laquelle on introduirait l'élément structurant et dont on garderait tous les parcours possibles. On obtient un image lissée morphologiquement.

- La fermeture de l'ensemble X par l'élément structurant B se définit par :

$$X^B = [X \oplus B] \ominus B$$

  Cette opération se traduit par le "remplissage" des parties concaves des formes[5] qui ne peuvent pas contenir l'élément structurant.

L'utilisation des ouvertures et des fermetures morphologiques permet d'effectuer des filtrages de signaux parasites sur les images. L'esprit de cette opération est de manipuler l'image avec un élément qui puisse s'apparier, en une fois ou par l'intermédiaire d'itérations, à la forme recherchée. Cette utilisation permet de retrouver des formes géométriques particulières, sans lissage, ce qui serait impossible par une autre méthode. En revanche, cette méthode nécessite une réglage très précis des éléments structurants. [Sternberg 86, page 347 du document] donne un exemple avec un cubeoctahèdre des possibilités et des résultats des mauvaises adaptations éventuelles.

5. ou convexes du fond

Dans le domaine de l'électrophorèse bidimensionnelle, la morphologie mathématique s'utilise essentiellement pour l'élimination des inégalités du fond et des traînées verticales et horizontales. Pour ceci on tâche de définir les éléments dont les itérations peuvent le mieux ressembler aux formes à éliminer ou aux taches d'électrophorèse.

- On élimine le fond par une ouverture avec un plateau ellipsoïdal ou avec une sphère. Les dimensions de ces éléments doivent être légèrement supérieures à celles des plus grosses taches.

- On élimine les traînées par des ouvertures avec une barre horizontale ou verticale, dont les dimensions sont supérieures aux élongations des taches de protéines.

Ces manipulations appellent un certain nombre de critiques. D'une part elles ne reposent sur aucune modélisation biologique des perturbations qu'elles sont sensées amoindrir. La réalisation de l'objectif tient seulement à l'amélioration de la lecture visuelle. Ensuite, une fois effectuées, elles faussent notablement les résultats des procédures suivantes telles que la quantification ou la modélisation éventuelle et, du fait de leur non réversibilité, on ne pourra plus accéder aux valeurs originelles. Enfin, les résultats finaux semblent assez peu convaincants. Nous avons procédé à une série d'essais sur des images de plasma. Ces images sont particulièrement intéressantes, car elles ne présentent pas d'homogénéité concernant les formes des taches protéiques. En effet, la partie centrale est "inondée" par l'albumine et prend l'aspect d'une pieuvre. Ces images forment ainsi une des meilleures mise à l'épreuve possible pour des opérateurs morphorlogiques. Nous avons essayé par les moyens d'ouvertures par des éléments structurants plats ou sphériques de tailles différentes[6] d'isoler le fond et nous n'y sommes jamais parvenu. Nous dégradions toujours une partie des taches, notamment autour de l'albumine. En comparant attentivement les images d'origine page 309 et les résultats page 313 publiés par [Skolnick 86a] qui correspondent à des électrophorèses relativement simples, nons constatons que la perte d'information est considérable, même si celle qui reste est plus lisible. Ceci provient du fait que la morphologie se destine au filtrage de forme très précises incompatibles avec la diversité de celles qu'on rencontre sur la plupart des électrophorèses bidimensionnelles. Nous avons abandonné cette méthode pour des utilisations générales.

Le seul point qu'on peut porter à l'actif de la morphologie mathématique, dans le cadre de l'électrophorèse bidimensionnelle, est l'élimination des bruits de hautes-fréquences qu'on peut réaliser par une ouverture avec un disque de diamètre 1 ou 2.

6. avec pratiquement tous les rayons possibles

1.2.2 Les autres méthodes

Les autres méthodes ont eu moins de fortune dans les publications. Elles se fondent essentiellement sur l'analyse des fréquences.

Les méthodes fréquentielles [Pratt 78] étudient les répartitions spectrales dans l'image et éliminent certains domaines de fréquences à cause de la quantité importante de bruit ou de parasites qu'ils comportent.

Dans le cas des électrophorèses, on ne peut les utiliser que pour éliminer le fond et les bruits de hautes fréquences :

- dans le cas de la dérive du fond qu'on pourrait modéliser comme un phénomène de très basse fréquence, on se heurte de nouveau à la diversité des formes possibles sur un gel : les gels de tissus offrent des taches relativement petites et homogènes alors que les gels de plasma offrent la masse compacte de l'albumine. De plus le filtrage fréquentiel repose sur un principe de linéarité des signaux qui n'a jamais été démontré dans le cas de l'électrophorèse.

- On peut par contre tout à fait éliminer les bruits de hautes fréquences par – c'est une tautologie – une méthode fréquentielle. Dans notre réalisation, nous avons implanté une méthode morphologique en pratiquant une ouverture par un disque simplement car la réalisation en est plus simple. Le filtrage a cependant des effets mieux calculables et mieux maîtrisables.

On pourrait aussi mettre en œuvre des méthodes *spatiales*, telles que le filtrage de Kalman ou les techniques ARMA. Ces techniques n'ont, pour l'instant, à notre connaissance, jamais été utilisées dans le domaine de l'électrophorèse bidimensionnelle.

Le filtrage de Kalman [Woods 77], se fonde sur une modélisation paramétrique du signal utile. Il met en œuvre un mécanisme de prédiction-vérification des paramétres en minimisant une erreur quadratique. Il demande des moyens de calcul assez lourds. On pourrait peut être l'utiliser pour éliminer le bruit de fond mais l'outil serait démesuré par rapport aux résultats à obtenir.

Les méthodes ARMA permettent d'éliminer ou de synthétiser des signaux présentant une certaine régularité. On pourrait envisager leur application à l'extraction des traînées si celles-ci sont prédictibles. Il ne s'agit ici que d'une conjecture qui demanderait des études supplémentaires.

De toute manière et en conclusion, l'amélioration des images par analyse du signal se fonde sur un principe de linéarité des phénomènes qui composent l'image. Par analogie avec les phénomènes acoustiques, on suppose que les divers signaux qui forment l'image s'additionnent et qu'on peut alors soustraire les perturbations. Il faut bien prendre garde

que ceci n'a jamais été réellement prouvé dans le cas de l'électrophorèse bidimensionnelle et que c'est manifestement faux dans certaines situations. Les signaux ne subissent pas toujours des modifications linéaires, ainsi on observe souvent des écrêtages de pics qu'on ne peut peut pas réparer par ces méthodes[7].

1.2.3 Alors quoi faire ?

Il est assez délicat de répondre à cette question. Nous avons développé un point de vue différent de celui qu'on présente habituellement et qui est de constater l'inefficacité de la plupart des méthodes. Cependant, il est parfois très difficile de travailler en présence d'un fond important ou de traînées qui rendent une chaîne de points illisibles, et il est donc admissible de chercher à les amoindrir par une technique ou par une autre. Toutes ces techniques sont néanmoins imparfaites et on ne doit s'en servir qu'avec la connaissance de leurs effets et pour isoler certains points particuliers. Elle ne permettent pas la découverte d'autres points, comme pourrait le faire une déconvolution par exemple, et on doit donc les manier avec une extrême précaution. Dans le cas d'un usage empirique, qui est le plus fréquent, on doit présenter clairement les choses qui sont de limiter l'information disponible à celle la plus visible, les détails se trouvant éliminés.

Ce qui précède ne répond toujours pas à la question que nous nous sommes posée. En fait, la réponse vient des applications. Nous avons choisi la démarche d'identifier des points particuliers de gels – nous l'exposerons dans les chapitres qui suivent – et dans les cas que nous avons eu à traiter les protéines se détachaient correctement, nous avons donc réduit les traitements au minimum qui est l'élimination du bruit de fond. Si pour un point particulier, l'étude impose l'élimination de l'environnement de ce point, alors il est parfaitement possible et justifié de concevoir un *ensemble* de filtres morphologiques destinés à agir sur cet environnement. L'utilisation des procédures d'amélioration d'image repose alors, en tout état de cause, sur une analyse puis une application contextuelle.

1.3 La détection des pics

La détection des taches se fait essentiellement par l'intermédiaire de la détection des maxima locaux. Ceci signifie qu'on ne détectera pas deux protéines trop proches qui fusionneraient en un seul sommet, comme sur le dessin du bas de la figure 1.4, page 36. La condition de séparation, par exemple, pour deux gaussiennes monodimensionnelles de même amplitude et de même écart-type est que les centres soient séparés d'au moins deux écarts-type.

Nous verrons dans le paragraphe suivant que dans une certaine mesure on peut déconvoluer les signaux et retrouver l'existence de protéines qui ne montreraient pas d'extremum

7. ni par aucune autre à notre connaissance

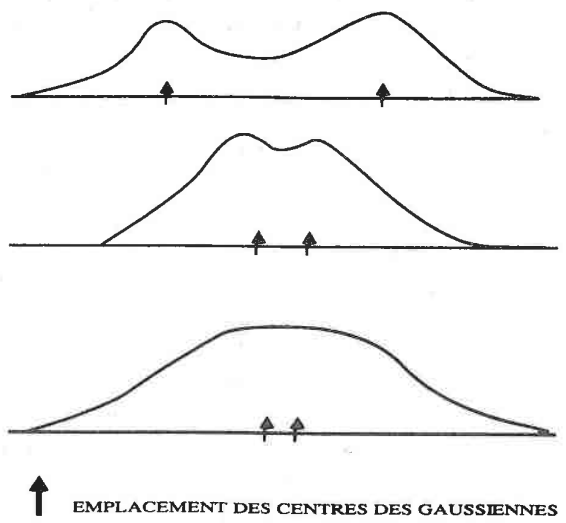


Figure 1.4. Conditions de détections

local.

On trouve un certain nombre de méthodes de détection de pics dans la littérature. Elles se fondent sur des esprits différents, mais elles nécessitent toutes au départ une image avec peu de bruit de haute fréquence. Si celui-ci est trop important, il conduira à une sur-détection notable, qui entraînera des erreurs dans les étapes ultérieures d'identification. Avant de procéder, il faudra donc éliminer ce bruit, par les moyens d'un lissage léger ou d'une ouverture avec un petit élément structurant.

1.3.1 L'analyse directe

L'analyse directe [Vincens 87a,Vincens 86a,Garrels 84] se fonde sur le balayage des lignes ou de l'image avec un élément géométrique, par exemple, un segment de droite, un rectangle ou une ellipse et un examen à l'intérieur de ces régions des maxima, puis une reconstruction de maxima locaux grâce à la comparaison avec les résultats trouvés dans les régions voisines. Cet examen est donc localisé à l'intérieur de formes bien précises. Il perd ainsi en généralité,

et impose un réglage de la taille des éléments. De plus, les auteurs de ces méthodes sont amenés à ajouter un certaine nombre "d'heuristiques" pour éviter les détections erronées. Elles font appel à un grand nombre de conditions, notamment lorsque l'examen est linéaire, pour la reconstruction des maxima locaux à partir des maxima des lignes adjacentes.

1.3.2 Les méthodes dérivatives

Les méthodes dérivatives [Lemkin 81,Funk 87,Ridder 84] se fondent sur les propriétés analytiques des surfaces continues. Étudions les dérivées secondes d'une courbe gaussienne $f(x)$, centrée en zéro, les calculs s'étendant de manière triviale aux courbes bidimensionnelles :

$$f(x) = A \exp -\frac{x^2}{2\sigma^2}$$

sa dérivée est :

$$f'(x) = -\frac{x}{\sigma^2} f(x)$$

et sa dérivée seconde :

$$f''(x) = \frac{1}{\sigma^2}(\frac{x^2}{\sigma^2} - 1) f(x)$$

On remarque que les dérivées secondes des courbes sont négatives dans une région centrée au sommet des pics et délimitée par les écarts-type. On suppose que ces écarts-type ne dépassent pas quelques unités. La méthode consiste dans le calcul des dérivées secondes de l'image et la recherche des zones ou ces deux dérivées sont négatives. On détermine les sommets en considérant leur barycentre.

Cette technique est séduisante du point de vue théorique pour les taches bien séparées, mais elle ne propose pas de solution claire et facile à mettre en œuvre dans le cas des recouvrements de protéines. Ceux-ci modifient sensiblement le comportement des opérateurs aux jonctions entre deux taches et on ne peut s'en servir dans tous les cas avec tranquilité d'esprit.

1.3.3 L'analyse de convexités

L'analyse des convexités [Miller 82,Prehm 87] se fonde sur une corrélation de l'image avec un masque dont les dimensions sont déterminées empiriquement. C'est un type de comparaison de formes ("template matching"), où on convolue l'image avec un prototype de fonction convexe. Comme dans le cas précédent cette méthode est très dépendante du modèle gaussien, bien que ceci ne soit pas toujours clairement indiqué par les auteurs. On peut imaginer un certain nombre d'exceptions qui prendraient en défaut cette technique. En particulier, comme la méthode précédente, elle s'accommode mal des zones de recouvrement et on doit encore faire appel à un certain nombre de règles heuristiques pour pouvoir les traiter. En résumé, l'efficacité de cette technique dépend beaucoup des détails d'implantations.

1.3.4 Les méthodes morphologiques

Contrairement aux méthodes précédentes, les méthodes morphologiques [Skolnick 86a] ne dépendent pas du modèle de courbe. Elles correspondent à la suite d'opérations suivantes :

1. on considère chaque point de la surface comme un sommet éventuel ;
2. on examine pour chacun de ces points si il possède un voisin[8] d'amplitude strictement supérieure et dans ce cas on le marque ;
3. on examine chacun de points marqués et on marque de la même manière ses voisins d'amplitude égale ;
4. on itère l'opération précédente jusqu'à la stabilité de l'image ;
5. on regroupe les points connexes restants qui correspondent aux maxima.

Par analogie, on peut comparer cet algorithme à de l'eau qui s'écoulerait de chaque point de l'image et dont on extraierait tous les points qui restent au sec.

Cette méthode présente l'avantage d'être universelle, c'est à dire indépendants des formes étudiées et de traiter dans le cas général les recouvrements de taches. C'est celle que nous avons implanté dans notre système.

1.4 L'extraction des paramètres du modèle des taches

Nous avons choisi un modèle paramétrique gaussien (voir le chapitre : *La technique électrophorétique*) pour nos taches protéiques afin de pouvoir les quantifier de manière précise et sans faire appel à trop de manipulations culinaires.

On modélise une tache par :

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

où $f(x,y)$ représente l'intensité de gris au point (x,y) ; A, l'amplitude de la gaussienne ; (x_o, y_o), les coordonnées du sommet ; (σ_x, σ_y), les écarts-type suivant les deux axes orthogonaux.

Ceci s'étend facilement aux taches qui se recouvrent[9] en supposant le modèle additif.

Avec N taches :

$$f(x,y) = \sum_{i=1}^{N} A_i \exp -\frac{(x-x_{oi})^2}{2\sigma_{xi}^2} - \frac{(y-y_{oi})^2}{2\sigma_{yi}^2}$$

8. au sens de l'octoconnexité ou de la quadriconnexité
9. c'est à dire les taches connexes

Une telle fonction se détermine par $n = 5 \times N$ paramètres qu'on peut approximer par la méthode des moindres carrés. L'esprit de cette approximation est de minimiser la différence quadratique entre l'image $I(x,y)$ et la fonction des paramètres. L'algorithme est un peu différent de celui qu'on expose en général du fait de la non linéarité des paramètres. On doit procéder par itérations.

Nous supposons que nous connaissons le nombre de fonctions gaussiennes : N. La surface s'écrit :

$$F(x,y,p_1,p_2,...,p_n) = \sum_{k=1}^{n} F_k(x,y)$$

où $p_0, p_1, \ldots$, représentent les $5 \times N$ paramètres : amplitudes, centres et écarts-type des N gaussiennes.

À une itération donnée du processus d'ajustement, nous avons les estimations, $p'_1, p'_2, p'_3, \ldots$ la fonction s'écrit alors :

$$F(x,y,p_1,p_2,...,p_n) \approx F(x,y,p'_1,p'_2,...,p'_n) + \sum_{i=1}^{n} \frac{\partial F(p'_i)}{\partial p_i}(p_i - p'_i)$$

Le reste entre la paramétrisation optimale et l'image s'écrit :

$$r(x,y) = F(x,y,p_1,p_2,...,p_n) - I(x,y)$$

Par définition de la méthode, si Ω est le domaine d'intégration, cette paramétrisation est telle que :

$$Q = \int\int_{\Omega} r^2(x,y)dxdy$$

soit minimal

On l'approxime le reste par :

$$r(x,y) \approx F(x,y,p'_1,p'_2,...,p'_n) - I(x,y) + \sum_{i=1}^{n} \frac{\partial F}{\partial p_i}\delta p_i$$

Soit $R(x,y)$, la différence calculée sur l'image :

$$R(x,y) = F(x,y,p'_1,p'_2,...,p'_n) - I(x,y)$$

Il vient :

$$Q \approx \int\int_{\Omega} (R(x,y) + \sum_{i=1}^{n} \frac{\partial F}{\partial p_i}\delta p_i)^2 dxdy$$

On cherche à minimiser Q qui est fonction des δp_i. Ceci est donné par la condition sur les dérivées partielles :

$$\frac{\partial Q}{\partial \delta p_k} = 0$$

ce qui se traduit par :

$$\frac{\partial Q}{\partial \delta p_k} = 2 \int \int_\Omega (R(x,y) + \sum_{i=1}^{n} \frac{\partial F}{\partial p_i} \delta p_i) \frac{\partial F}{\partial p_k} dxdy = 0$$

On peut représenter cette équation par le produit de matrices :

$$B\Delta = D$$

où

$$b_{ij} = \int \int_\Omega \frac{\partial F}{\partial p_i} \frac{\partial F}{\partial p_j} dxdy$$

et

$$d_i = \int \int_\Omega (I(x,y) - F(x,y)) \frac{\partial F}{\partial p_i} dxdy$$

Les incréments des paramètres à chaque l'itération sont donnés par la formule :

$$\Delta = B^{-1} D$$

où

$$\delta_i = p_i - p_i'$$

L'exposé théorique de l'ajustement par les moindres carrés est assez simple, par contre sa mise en œuvre pose un certain nombre de problèmes de réglages et d'implantation.

Le premier de ces problèmes est lié à la puisssance de calcul requise. En effet l'ajustement d'une gaussienne simple impose plusieurs inversions d'une matrice 5×5 et par exemple celui d'un groupe de 5 protéines connexes, l'inversion de matrices 25×25. À ceci s'ajoute un grand nombre d'opérations sur des réels pour calculer les coefficients. En fait, ce sont surtout les protéines connexes qui demandent un temps de calcul important, car les gaussiennes simples sont très vite ajustées grâce à la connaissance des coordonnées de leur sommet et de leur amplitude. Par ailleurs, on peut accéder directement à leur quantité de matière par une simple intégration.

La complexité de l'analyse dépend donc des amas de taches. Certains gels en contiennent plus que d'autres, certaines régions sur les gels sont plus ou moins compactes et certaines procédures de migrations sont plus ou moins séparatives. Il est donc délicat de déterminer la puissance de calcul exactement nécessaire. L'analyse d'une image de gel 512×512 est impossible en un temps raisonnable pour une machine de type VAX-11 sans processeur spécialisé et pour obtenir des résultats, on a dû simplifier les algorithmes d'ajustement [Nugues 84]. Aujourd'hui les processeurs arithmétiques et vectoriels étant largement répandus, on peut analyser une image en quelques minutes sur n'importe quel station de travail ou micro-ordinateur équipé de manière adéquate. Les procédures de simplifications n'ont plus lieu d'être.

Le domaine Ω sur lequel on réalise les opérations est en théorie infini ou comprend toute l'image. Ceci n'est pas envisageable dans l'état actuel de la technique. On limite donc Ω aux domaines de connexité de l'image seuillée, dans nos programme à 1/10 de l'amplitude maximale soit 25.

Lorsque l'approximation se fait avec des protéines possédant un maximum local, on initialise l'amplitude et les coordonnées des sommets avec les valeurs données par les procédures de détection des pics. Les écarts-type prennent eux des valeurs par défaut.

La procédure d'approximation peut aussi ajuster une tache en imposant le nombre de protéines. Cela se produit notamment losqu'un module du système détecte une incompatibilité entre le nombre d'extrema locaux et le nombre de protéines possibles dans cette zone : qu'il s'agisse d'une sur-détection ou d'une sous-détection. Nous pouvons qualifier cette manipulation de déconvolution en considérant le procéssus de migration comme une fonction de transfert. Dans ce cas, nous initialisons le processus d'approximation avec les valeurs potentielles des paramètres des protéines considérées.

Lors de l'ajustement, il peut se produire des oscillations d'assez grande amplitude du vecteur d'incrémentation Δ. On doit donc mettre en place un "modérateur" pour limiter ces oscillations qui autrement pourraient causer des dépassements de la borne supérieure de codage des réels en machine. Nous avons implanté une méthode très simple : nous analysons chaque composante composante du vecteur Δ avant de l'additionner aux paramètres et nous remplaçons cette composante par sa racine carrée si elle dépasse un seuil donné.

On arrête le processus lorsque $\| \Delta \| < \epsilon$. La vitesse de convergence pour les cas que nous avons eu à traiter dépend du nombre de taches qu'on ajuste à la fois. Il faut moins d'une dizaine d'itérations pour une gaussienne simple et jusqu'à 30 pour une tache de cinq gaussiennes.

La quantification de chaque protéine se fait en calculant l'intégrale :

$$I = \int \int A \exp - \frac{(x - x_o)^2}{2\sigma_x^2} - \frac{(y - y_o)^2}{2\sigma_y^2}$$

dont la valeur est :

$$2A\Pi\sigma_x\sigma_y$$

2
Mise en correspondance d'images

2.1 Généralités sur la mise en correspondance

La mise en correspondance d'images est un problème qui a fait l'objet de très nombreuses études. Le domaine est vaste et les méthodes de traitement dépendent en grande partie du sujet traité.

Dans le cas général, on distingue plusieurs types de mise en correspondance [Ballard 82]. Les variations tiennent surtout à la connaissance du contexte sous-jacent.

Par ordre de connaissance croissante sur les images à apparier, il vient d'abord l'appariement *iconique*. Il caractérise la mise en correspondance directe entre images sans aucune référence à un modèle. Il peut être la source de nombreuses erreurs et on ne le pratique que faute de mieux. Ces appariements font en général appel à des techniques corrélatives.

Les cas le plus courants de mise en correspondance se réfèrent à l'appariement entre une image et un modèle. Sous ce terme on décrit plusieurs types de méthodes :

- l'approximation à un modèle paramétrique. Ceci englobe, par exemple, la transformée de Radon[1] pour les modèles simples, l'ajustement par la méthode des moindres-carrés comme nous l'avons exposé au chapitre précédent ainsi que les techniques issues du filtrage de Kalman, telle que la prédiction-vérification [Ayache 86] qui se fondent sur une minimisation récursive de l'écart quadratique entre les éléments d'une scène et un modèle paramétrique.

- La mise en correspondance la plus générale et la plus élevée du point de vue conceptuel s'opère par isomorphisme de graphes. On dresse les relations géométriques ou structurelles qui existent entre les objets ou leurs différentes parties, par l'intermédiaire d'un réseau sémantique par exemple, et on tâche d'établir un isomorphisme entre l'objet et le modèle ou bien entre des parties de l'objet et des sous graphes du modèle (recherche de cliques) ou encore en établissant des isomorphismes "inexacts" [Shapiro 81] entre l'objet et le graphe du modèle. On peut procéder par étapes et d'abord rechercher à

1. ou de Hough, qui l'a redécouvert 60 ans plus tard

grouper certains éléments, tels que les segments colinéaires, avant de procéder à la mise en correspondance [Mohr 88].

L'électrophorèse bidimensionnelle est une technique relativement récente pour laquelle il n'existe pas encore de modèle général d'image. Les seules données disponibles pratiquement étant l'aspect des gels. Son application ayant permis de découvrir une multitude de nouvelles protéines, les premiers appariements n'avaient pour but que de pouvoir comparer, tant bien que mal, d'un gel à un autre quelles étaient les taches différentes : celles qui apparaissaient, celles qui disparaissaient, les changement notables de quantité... L'extension de la connaissance *a priori* sur les gels permettra sans doute de passer à un autre type de procédé, c'est entre autres l'un des objectifs de notre travail, mais les méthodes utilisées couramment, à ce jour, ne relèvent encore que de l'appariement "iconique".

2.2 Appariement de gels dans le cadre d'une expérience

Les aspects des gels bidimensionnels d'électrophorèse sont d'une très grande diversité. Il reflètent les particularités des tissus et leur quantité d'informations. Du fait du peu de connaissance sur la plupart des protéines révélées, il est illusoire, à présent, de comparer des gels de natures différentes. Les méthodes d'appariement concernent des gels issus en général d'un même laboratoire et d'une même expérience. Dans l'avenir, c'est sans doute par l'intermédiaire de bases de données qu'on réalisera les comparaisons de niveaux supérieurs.

Toutes les méthodes que nous présentons fonctionnent en prenant comme point de départ les grandes similitudes – c'est à dire les protéines les plus importantes – entre les images et à partir de ces ressemblances marquantes, elles étendent la connaissance en détectant les protéines communes ou manquantes. Elle ne font pas de miracles et elles ne sont efficaces que lorsqu'on peut déjà à l'œil effectuer une comparaison. D'autre part, sans une bonne reproductibilité et un pouvoir de résolution correcte du procédé chimique, leur résultats seront incohérents.

2.2.1 L'appariement par déformation des gels

La première méthode se fonde sur la déformation dynamique de la surface des gels [Garrels 84] ou *stretching*. On considère un gel comme l'image de référence et on opère une anamorphose sur le gel à apparier.

L'algorithme ne fonctionne pas de manière totalement automatique. On doit l'initialiser en appariant manuellement un certain nombre de points remarquables ou *"landmarks"*. L'algorithme apparie les protéines en partant des plus proches de ces points remarquables et étendant au fur et à mesure des itérations, la zone d'appariement. Le programme procède en plusieurs passes, en considérant dans une zone d'appariement donnée, la protéines

les plus importantes, puis toutes les protéines et enfin en reliant entre elles toutes les régions environnant les landmarks. À chacune de ces passes, le programme réapparie toutes les protéines.

La correspondance entre les coordonnées du gel de référence et celles du gel à apparier est locale à chaque région entourant les landmarks. Elle est linéaire et on l'obtient en calculant la droite de régression qui lie les abscisses des points appariés des deux régions à l'itération précédente, ainsi que celle qui lie les ordonnées.

Cette méthode fait souvent appel à un opérateur en cas de risque d'échec et elle demande l'appariement manuel d'un grand nombre de "landmarks". Dautre part le raccordement des régions qu'ils définissent fait appel à un certain nombre d'heuristiques. Enfin, elle est assez sensible à la reproductibilité et à la qualité des gels.

2.2.2 L'appariement par création de relations

Mark Miller a tenté de réduire [Miller 82] puis d'éliminer la phase d'appariement manuelle. Le programme considère d'abord les protéines les plus importantes des gels et les voisins proches de ces protéines. Pour chacun de ces groupes, il opère une transformation en coordonnées polaires avec comme origine le centre de gravité des taches. Il apparie deux groupes sur deux gels différents si ils possèdent des descriptions géométriques similaires. Il détermine ensuite la probabilité de l'appariement et si celle-ci est suffisante, il apparie les groupements connexes au groupement traité précédemment utilisant la même méthode. [Appel 87] a repris cet algorithme dans le système MÉLANIE.

2.2.3 L'appariement par maillage géométrique

Les méthodes précédentes présentent l'inconvénient d'opérer de manière locale. [Skolnick 82a, Skolnik 85,Skolnick 86b] tente de pallier ceci en projetant un graphe de Gabriel [Toussaint 80] sur la surface du gel. On considère les points d'un gel au-dessus d'un seuil fixé et on constuit le graphe en reliant deux protéines par un arc si le cercle dont la diamètre est défini par cet arc exclu toutes les autres protéines. Cette construction est relativement stable. On apparie deux protéines en comparant leurs arcs respectifs. Autour de ces points de confiance, on tente d'apparier les protéines restantes en examinant les points au-dessous du seuil en utilisant leurs coordonnées relatives.

2.2.4 L'appariement par création d'automate

Pierre Vincens [Vincens 87a,Vincens 87b] utilise une démarche dont l'inspiration est différente des précédentes. Elle consiste dans la description en termes d'automates des relations géométriques des groupes de protéines. Un groupe de protéines se décrit par une grammaire formelle. La relation qui lie deux protéines se code par trois mots :

1. l'angle entre les deux protéines,
2. la distance entre les deux protéines,
3. leur différence d'intensité.

Les valeurs que prennent ces trois paramètres étant discrètes.

On effectue l'appariement en comparant successivement les phrases d'une grammaire provenant d'un groupe de protéines sur un gel aux phrases issues d'un groupe sur l'autre gel. On crée au départ les relations pour des groupes d'environ dix protéines autour des protéines les plus importantes, puis on les étend aux groupes voisins. Étant donné que l'égalité est rare entre les groupes à cause des distances ou orientations légérement différentes, on pondère chaque appariement de phrase par un coefficient de vraisemblance allant de -1 à $+1$. Ces coefficients sont définis de manière empirique et sont propagés par combinaison arithmétique au fur et à mesure des déductions.

2.2.5 La prise en compte de plusieurs gels

Les méthodes que nous venons d'exposer permettent la comparaison de gels deux à deux. Or le but d'une expérience est, en général, la comparaison d'une série de gels. Cette comparaison multiple est fondamentale et difficile. Elle est traitée de manière assez empirique [Vincens 87b,Miller 88] et entraîne des erreurs inévitables. Deux stratégies sont possibles :

1. la première consiste dans la construction d'un gel maître : le "master". La démarche est la suivante. On commence par apparier deux gels de la série et on fusionne toutes les taches, qu'elles soient communes aux deux gels ou propres à l'un des deux, en un gel qui constituerra l'embryon du gel maître. On apparie les gels suivants avec le gel maître et lorqu'une protéine nouvelle apparaît on l'ajoute au gel maître. Ce gel maître contitue donc une carte virtuelle de toutes les protéines qui peuvent apparaître sur un gel. Cette technique est séduisante mais elle est sujette à des erreurs. Elle dépend notamment de l'ordre d'introduction des gels lors de la construction du gel maître. Un moyen biologique de vérification serait de procéder à une co-électrophorèse où on mélangerait tous les échantillons à analyser en un seul qu'on ferait migrer. Toutes les protéines étant présentes, on devrait les retouver sur le gel bidimensionnel.
2. La seconde stratégie est de prendre un gel comme référence et de l'apparier à tous les autres. À partir de ces paires, on fusionne les gels et on dresse une table des protéines présentes sur un ou plusieurs gels de la série. Une fois cette table établie, on doit vérifier la cohérence globale en appariant chaque gel de la série avec le modèle obtenu.

En l'absence de modèle, ces méthodes d'appariement multiples sont les seules disponibles. Elles conduisent à un certain nombre d'erreurs. Par un calcul purement statistique, lorsqu'on

apparie 40 gels avec 1% d'erreur à chaque appariement, ce qui est très faible, on obtient 30 % d'erreur globale : $(1-0,99^{40})$. Cette méthode de calcul est approximative, car il existe des zones confuses où les taux d'erreurs sont très forts et d'autres où il n'y en a pas mais ceci est tout de même difficilement acceptable pour la plupart des expériences biologiques et cette imprécision a sans doute beaucoup retardé l'extension des applications de l'électrophorèse bidimensionnelle.

2.3 Relation avec une base de données

La très grande quantité d'informations provenant des électrophorèses bidimensionnelles nécessite l'utilisation de bases de données. Ces bases ont deux objectifs :

- le premier devra permettre la comparaison de lignées de cellules et aider au dépouillement des gels [Anderson 88,Miller 88,Vincens 87c].
- Le second est la construction d'une base généralisée sur les protéines [Anderson 82] dont l'électrophorèse bidimensionnelle sera le source majeure d'information. Avant cette étape lointaine, beaucoup d'obstacles – et notamment la maîtrise du premier point – restent à surmonter.

[Garrels 84] distingue quatre niveaux de complexité de construction de bases de données.

1. Le premier niveau résulte de la construction d'une base à partir des gels d'une expérience sur un même type de tissu ou de cellule dans un même laboratoire.
2. Le second niveau traite les données du niveau précédent mais provenant d'expériences effectuées à des périodes différentes.
3. le troisième niveau de base de données contient les informations tirées de manipulation faites sur lignées de cellules ou des tissus voisins.
4. Le quatrième niveau définit la construction de bases à partir de données provenant de lignées de cellules et de laboratoires différents.

Le cœur de ces bases est constitué par les taches des images qu'on doit mettre en relation avec d'une part les informations sur les images, telles que les conditions d'obtention, le tissu, l'opérateur, etc. et d'autre part les connaissances externes sur ces taches, telles que les protéines qu'elles représentent, les phénomènes auquelles elles sont associées, etc.

Cet exposé est assez théorique, car dans la réalité, la construction de ces bases de données se heurte à un grand nombre de difficultés. Nous faisons notamment face à une absence de nomenclature universelle de codage des données et en général beaucoup d'incertitude en ce qui concerne l'identification des points visibles à la surface des gels.

À partir de l'instant où on peut clairement identifier une protéine sur un gel, comme c'est le cas de l'albumine pour la plupart des tissus par exemple, il est toujours possible de lui associer d'autres données, qui sont dans le cas de l'albumine assez complètes. Pour le reste, et c'est la majorité de points, il semble difficile de rentrer autre chose que ce qu'on peut directement tirer du gel[2] et des procédures d'obtention. On pallie le manque d'identification des points par la mise en relation avec leurs homologues sur les images issues d'une même expérience. On doit alors être très prudent car nous avons vu qu'il est délicat d'établir la liste exacte des polypeptides appariées issus d'une série d'expériences ; il faudra sans doute franchir de nombreuses étapes pour accéder à la généralisation de la mise en relation de taches issues de multiples expériences.

La classification de Garrels nous semble assez artificielle bien qu'elle fasse autorité dans le monde de l'électrophorèse bidimensionnelle. L'étape déterminante à franchir est plutôt la possibilité d'identification des taches sur une image :

- d'abord à l'intérieur d'une expérience,
- puis à l'intérieur d'autres expériences.

Que cette identification se fasse par l'intermédiaire d'un nom ou d'une propriété telle que : inductible par l'hormone juvénile, par exemple. Dans le cadre d'une expérience, et grâce à des appariements iconiques, [Vincens 87c] a réalisé ceci en l'implantant sous une forme relationnelle [Codd 70]. On peut questionner la base en faisant intervenir des propriétés et visualiser les protéines correspondantes.

Dans l'état actuel, faute d'un langage de description géométrique suffisamment souple et sûr des protéines des gels, il est pratiquement impossible de construire une base qui puisse s'étendre au dehors d'une manipulation avec la généralité nécessaire.

2.4 Que conclure ?

L'étape de mise en correspondance des gels est la partie la plus délicate du processus actuel d'analyse. Faute d'une connaissance suffisante du contexte, c'est la seule qu'on puisse mettre en œuvre et elle permet, dans une certaine mesure, d'assister le biologiste pour retrouver des protéines communes aux images d'une même expérience.

Il est difficile de juger la qualité des algorithmes, car ils s'appliquent à des situations différentes. Certains s'implantent facilement et conviennent bien à des gels reproductibles, ou bien échouent à apparier, d'autres tolèrent plus de variations de position mais sont plus complexes à mettre en place et induisent eventuellement plus de faux appariements. C'est,

2. l'abscisse et l'ordonnée

en définitive, surtout à travers l'optique de ce compromis qu'il faut considérer la mise en correspondance de gels.

L'amélioration des résultats de cette étape, pour l'essentiel, n'est pas de notre ressort, mais plutôt de celui des méthodes d'obtentions biologiques qui puissent assurer la reproductibilité la plus exacte possible. Nous n'avons donc pas cherché à perfectionner une autre méthode (une de plus ?) et nous avons préféré consacrer nos efforts à implanter un procédé d'identification qui tire parti de l'expertise des biologistes dans la localisation des protéines connues. Nous le décrivons dans le chapitre suivant.

3
Identification par expertise

3.1 Pourquoi une expertise

Au chapitre précédent, nous avons présenté et critiqué les méthodes d'appariement de gels entre eux. Ces méthodes entraînent des erreurs inévitables, car elles impliquent la création du modèle au cours de l'analyse sans en connaître la structure sous-jacente. Ce modèle ne peut être que géométrique et très pauvre. Il ne tient compte que des relations entre le point qu'on considère et ses abords immédiats. Pour obtenir un appariement parfait, il faudrait que les manipulations biologiques soient exactement reproductibles, ce qui n'est jamais le cas. On obtient, cependant des résultats acceptables lorsque le nombre d'appariements est limité.

Lorsqu'on ne dispose d'aucune autre connaissance sur la structure des gels à examiner, c'est malheureusement la seule manière de procéder. En revanche quand celle-ci est diponible, elle peut conduire à des résultats bien supérieurs. Pour illustrer l'avantage qu'on peut tirer de ces *points de repère*, considérons l'exemple suivant : supposons que deux polypeptides de même masse moléculaire et de pI voisins n'apparaissent jamais ensembles, comme sur la figure 3.1 page 52. Sans connaissance *a priori*, un analyste humain confondra sans doute ces deux polypeptides. Si on admet qu'on dispose de cette connaissance et qu'on puisse différencier les deux protéines, grâce par exemple à une autre dont le pI est connu, l'appariement – ou dans ce cas l'identification – sera possible, figure 3.2 page 52.

C'est ce type d'expertise à laquelle la machine ne peut pas accéder seule[1], que nous allons tenter de décrire et de communiquer à la machine.

3.2 L'implantation d'une expertise dans un système

Pour certains gels, un expert peut identifier sans fautes certaines protéines présentes et éventuellement en déduire, si les connaissances sont suffisantes, des conséquences sur l'état de la personne ayant subi cette analyse.

1. dans l'état des possibilités techniques actuelles

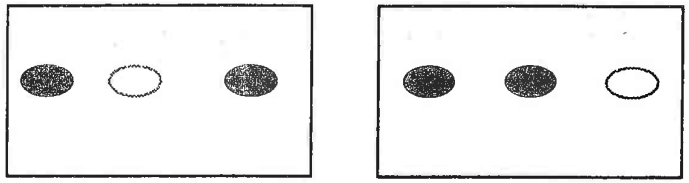


Figure 3.1. Protéines mutuellement exclusives

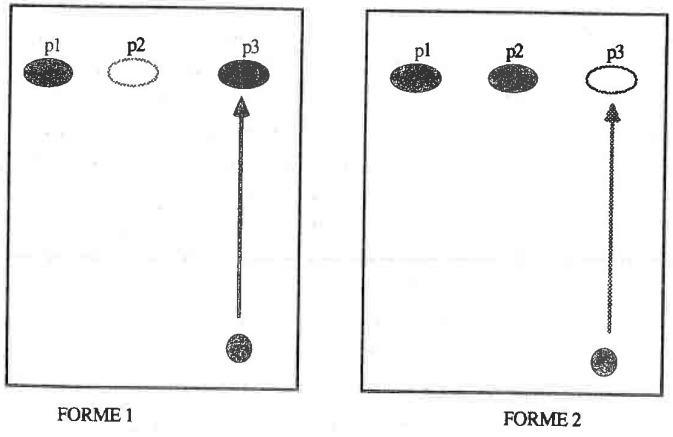


Figure 3.2. Identification de protéines

Cette connaissance repose essentiellement sur les liens entre les aspects géométriques des gels et les propriétés physico-chimiques des protéines ou des groupes de protéines à identifier. On sait par exemple que telle protéine apparaît sous une forme sialylée, c'est à dire qu'elle aura fixé une molécule de sucre et sera plus lourde ou bien que dans tel phénotype on aura une chaîne de trois taches.

L'introduction d'expertise dans le domaine de l'électrophorèse bidimensionnelle consiste à pouvoir faire le lien entre ces affirmations et l'image contenue dans l'ordinateur. Nous avons examiné dans les chapitres précédents les méthodes qui pouvaient nous servir pour extraire les informations des images. Nous allons maintenant considérer la façon de les agencer afin de tenter de reproduire le raisonnement d'un expert humain.

Cette connaissance, bien sûr, ne porte pas sur la totalité du gel, car comme nous l'avons indiqué précédemment, la plupart des taches représentant des protéines ne font l'objet d'aucune étude. Elle permet simplement l'extraction de protéines ou de formes que l'analyste a l'habitude d'identifier.

3.2.1 La représentation des connaissances et les raisonnements

Logique et prédicats

L'une des propriétés les plus marquantes des systèmes à bases de connaissances en intelligence artificielle est la séparation très claire entre les données et les mécanismes qui les manipulent : les connaissances sont en général déclaratives.

Cette déclaration intervient notamment sous une forme logique de propositions ou de prédicats reliés par les opérateurs :

- de négation $\neg$,
- de conjonction $\wedge$,
- de disjonction $\vee$,
- d'implication $\Longrightarrow$,
- d'équivalence $\Longleftrightarrow$,
- de quantification universelle $\forall$ et
- de quantification existentielle $\exists$.

La combinaison de ces opérateurs logiques fondamentaux, des propositions ou des prédicats fournissent les axiomes du problème. À partir de ces axiomes, il existe des algorithmes de résolution en logique d'ordre 0[2] ou d'unification et de résolution en logique d'ordre 1[3]

2. la logique des propositions
3. la logique des prédicats

[Rich 83,Charniak 85].

Les prédicats représentent des affirmations telles que le point A de coordonnées (100, 100) est une haptoglobine ou que la préalbumine est plus légère que l'albumine. On transcrit généralement ceci par :

Haptoglobine(A)
PlusLégère(préalbumine, albumine)

L'unification correspond à la liste des substitutions possibles qui permettent d'obtenir deux formules égales. Ainsi en ayant les prédicats :

PlusLégère(préalbumine, albumine)
PlusLégère(préalbumine, x)

x pourra s'unifier à albumine.

On peut rarement exprimer les problèmes ainsi dans leur totalité et les résoudre grâce à des axiomes. Dans la réalité, le raisonnement est souvent incertain et on est obligé d'introduire des coefficients de vraisemblance dans l'enchaînement des règles ou de les moduler par une logique floue [Sowa 84]. D'autre part, on ne peut pas appliquer un algorithme général de résolution sur toutes les données par manque d'efficacité.

Objets et structures

Après avoir défini les opérateurs logiques, la seconde étape consiste à donner une stucture aux entités à manipuler. La représentation "naturelle" consiste à prendre des objets munis d'attributs [Cointe 86]. Ainsi nous définissons une tache protéique quelconque sur un gel par son amplitude, son abscisse, son ordonnée et l'intégrale de sa densité.

(protéine (amplitude) (abscisse) (ordonnée) (masse))

Les *frames* de Marvin Minsky permettent une extension du rôle des attributs comme par exemple, la prise d'une valeur par défaut ou l'attachement procédural que nous n'avons pas implanté.

Ces objets se structurent en classes et en sous-classes. On obtient une organisation hiérarchique de type *est-un*. Ainsi une protéine contient la sous-classe des apolipoprotéines qui contient la sous-classe de apolipoprotéines E. Les classes font partie de la connaissance *a priori* et on ne peut pas les modifier dynamiquement. Nous donnons un exemple sur la figure 3.3, page 55.

On crée un objet en l'instanciant par moulage de la classe créatrice, de la même manière qu'en *Smalltalk*[4]. Le nouvel objet possède donc les attributs de sa classe que ceux-ci aient

4. par opposition à l'instanciation par duplication

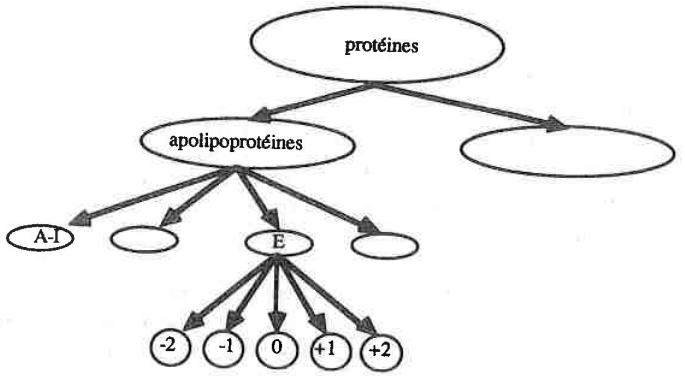


Figure 3.3. Organisation hiérarchique

une valeur ou non.

Parmi une série d'objets crées, on peut en filtrer[5] un ou plusieurs satisfaisant à un ensemble de contraintes, tel que, par exemple, trouver une protéine dont l'amplitude soit comprise entre 100 et 200 et dont l'ordonnée soit inférieure à 10 ou supérieure à 300 [Winston 84].

On définit pour chacun de ces objets un certain nombre d'opérations admissibles sur les instances d'une classe particulière, en plus des manipulations fondamentales, telles que la création, la destruction, la modification de la valeur des attributs. Ainsi par exemple, on définit une classe **zone-d-analyse** qui admet les opérations :

1. **translater** avec en paramètre un vecteur de translation,
2. **agrandir** avec en paramètre un coefficient d'agrandissement,
3. **réduire** avec en paramètre un coefficient de réduction.

Le déclenchement de ces actions sur les objets se fait par l'intermédiaire de règles de production du type :

SI *condition* **ALORS** *action*

5. par *pattern-matching*

La partie condition est rendue valide par filtrage sur la base des objets. Elle est constituée d'une conjonction de "motifs" qui doivent s'unifier, en tenant compte de contraintes, à des objets d'une classe.

Méthodes de raisonnement

Les méthodes de raisonnement se classent en deux groupes [Nagao 84] :

- les raisonnements ascendants qui agissent à partir des données brutes extraites par les procédures de traitement d'image, les assemblent et tentent de les faire coïncider avec les modèles d'organisation spatiale des protéines.
- Les raisonnements descendants qui partent de ces modèles d'organisation géométriques, les décomposent en sous-modèles, émettent les hypothèses nécessaires et tentent d'extraire, par l'intermédiaire des procédures de traitement d'image, les données correspondant à ces hypothèses.

L'application de ces types de raisonnements n'intervient pas dans les mêmes conditions. Le flux du raisonnement se dirige dans l'un des deux sens : des faits potentiels vers les ensembles à construire ou bien le contraire. Le "bon sens" suggère de prendre la méthode la plus économique. Si le nombre de faits est très important, combiné avec plusieurs règles, il créera un grand nombre de conclusions possibles à comparer aux modèles, ceci est d'autant plus génant que l'extraction de ces faits potentiels peut être coûteuse. Réciproquement, si il y peu de faits, il est facile de les combiner pour constituer des ensembles. Dans le cas d'une image, les faits potentiels sont considérables, il est donc absolument nécessaire de les limiter. Nous devrions alors opter pour le choix de méthodes descendantes, car les modèles ne sont pas si nombreux en général. Ça n'est pas toujours possible et on ne peut pas toujours, par exemple, avec une certitude absolue trouver le pixel qui correspond aux modèles. Nous fondons donc nos raisonnements sur l'image sur un modèle descendant faisant appel à des démarches ascendantes. Un certain nombre de systèmes combinent les deux méthodes avec succès [Matsuyama 85,Matsuyama 87,Nagao 88].

Les procédures ascendantes fonctionnent essentiellement de deux manières. La première est la focalisation de l'attention[6]. La seconde est la prédiction des parties manquantes du modèle.

La focalisation de l'attention a pour objectif de réduire le nombre de données à extraire. C'est un point essentiel de l'analyse des images du fait de leur très grande quantité d'informations potentielles. Ainsi, il est inutile d'extraire tous les sommets d'une image si on sait que la protéine recherchée se trouve dans son quart inférieur. La focalisation fonctionne

6. traduction de l'anglais : *focus of attention*

en pilotant les procédures d'analyse d'image dans des zones présentant des caractéristiques bien determinées :

1. localisation topologique,
2. moyenne du niveau de gris environnant,
3. homogénéité fréquentielle, etc.

C'est une première application des modèles. En général, on ne classe pas la focalisation d'attention comme un processus descendant, mais simplement comme une façon de commander l'application des processus ultérieurs à l'intérieur des zones déterminées, que ces processus soient ascendants – bottom-up – ou descendants – top-down –. On l'assimile donc au "contrôle" d'un système à base de connaissance. En fait ce contrôle est largement indépendant de la focalisation d'attention, c'est pourquoi, nous l'avons répertorié différemment.

L'application du modèle dans sa totalité intervient lors de la vérification de l'instanciation des configurations trouvées par les procédures ascendantes, qu'elles soient complètes ou partielles. En effet, lors du processus de migration des protéines, il est fréquent que des variations se produisent par rapport à un modèle donné, qui ne permettent pas la construction parfaite d'une instance par voie ascendante. Par exemple un point se trouve en-dessous d'un seuil fixé ou alors il est presque masqué par une protéine plus importante, etc. La stratégie que nous avons utilisée comprend donc l'application du modèle à l'image grâce à la connaissance de son instanciation partielle. Elle permet la recherche plus précise, par l'application des procédures de traitement de bas niveau adéquates, des parties manquantes.

Les méthodes ascendantes agissent par analyse syntaxique des faits issus des procédures de bas niveau, essentiellement les extrema locaux et leur caractéristiques topologiques. Cette analyse de structures spatialement proches se fonde sur la correspondance aux modèles partiels possibles, afin de construire des entités de plus en plus élaborées. Elle s'exprime sous la formes de règles dont la partie gauche doit s'unifier avec des objets en respectant les contraintes des sous-modèles. On cherche à déclencher les règles à partir de points dont les propriétés géométriques sont particulièrement stables au cours des migrations protéiques. Il se forme alors des îlots de confiance, c'est à dire des ébauches "sûres" d'instanciations partielles. Ces ébauches permettent l'émission de nouvelles hypothèses pour pouvoir procéder à l'instanciation complète ou partielle minimale du modèle. L'instanciation partielle minimale correspond au seuil au delà duquel l'analyse descendante peut opérer.

3.2.2 Les architectures

L'architecture réalise l'implantation des points que nous avons décrits précédemment. Elle permet la gestion effective de la connaissance. Nous exposons les grandes lignes des architectures courantes puis la solution que nous avons retenue.

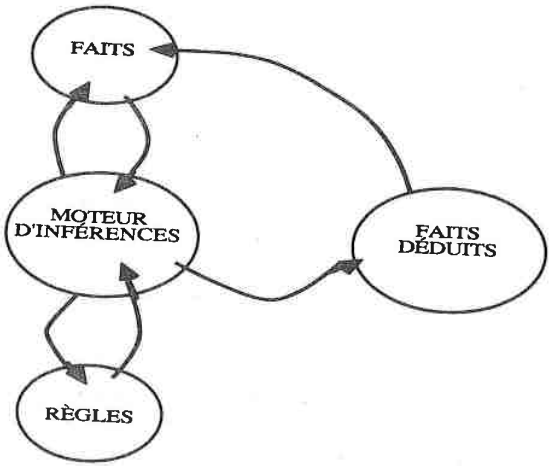


Figure 3.4. Structure d'un système expert

Les systèmes experts

Les systèmes experts ordinaires forment la première des organisations pour gérer la connaissance. Leur caractéristique est d'accumuler une expertise, s'écrivant sous forme de *règles*[7], sans grande structure. Le système prend connaissance d'un certain nombre de données : *les faits*. Un *moteur d'inférence* associe ces faits aux règles pour inférer des hypothèses afin d'en tirer des conclusions. On peut attribuer des coefficients de vraisemblance aux règles pour moduler ces conclusions. Nous donnons un schéma simplifié d'un système expert sur la figure 3.4, page 58.

Il existe deux mécanismes d'enchaînement de ces règles : le chaînage avant chaînage arrière on peut faire une analogie au niveau microscopique avec les méthodes descendantes et ascendantes. Un certain nombre de générateurs de systèmes experts possèdent la possibilité d'effectuer ces deux chaînages. Cette classification est cependant un peu arbitraire car on peut presque toujours transformer un mécanisme de chaînage. Ainsi le système *ART* qui

7. d'après la définition classique

permet l'écriture de règles en chaînage avant et arrière procède toujours par chaînage avant en récrivant les règles arrières du type :

C si A et B

en :

si but(C) **et** A **et** B **alors** C **et** non-but(C).

Les systèmes experts considèrent la connaissance de manière globale, sans structuration particulière. Lorsque elle est importante cela pose de nombreux problèmes, notamment dans le choix du déclenchement des règles où la résolution des conflits se fait en général par heuristiques simples : précédence dans l'entrée dans le système, etc.

Ce manque de distinction à l'intérieur des types de connaissances à conduit à l'évolution des systèmes expert vers d'autres types d'architectures.

l'introduction

Les *blackboards*

L'architecture des blackboards est une tentative de structuration de la connaissance. Elle permet l'inclusion de types de connaissances diverses gérés selon un mode qui leur est propre. Ceci évite d'imposer un schéma de structure en règles qu'on trouve dans les systèmes experts. La figure 3.5 page 60 représente un exemple de blackboard.

Cette organisation très souple a permis l'extension de son utilisation à des domaines très variés, que ce soit dans la reconnaissance de la parole d'où est né le premier blackboard [Lesser 75] ou de l'image [Matsuyama 87,Shafer 86,Thorpe 88].

À l'intérieur d'un blackboard, la division de la connaissance se fait suivant son type d'obtention et de manipulation. On construit ainsi des sources de connaissances possédant leur propre mécanisme d'inférence, qui écrivent et lisent des informations dans une base de donnée commune : le *blackboard*. Les objets du blackboards sont structurés hiérarchiquement.

La construction de la solution peut s'opérer par les faits ou les modèles, par des procédures algorithmiques qui fournissent des données sur l'image par exemple ou par émissions d'hypothèses par les sources de connaissances de niveaux supérieurs. Le schéma de raisonnement est essentiellement opportuniste en fonction des solutions partielles en cours de construction et dans la mesure où une source de connaissance peut compléter ces solutions partielles. Chaque source de connaissance est maintenue distincte et connaît les conditions auxquelles elle peut se déclencher pour contribuer à l'élaboration du résultat.

La structure réelle de "contrôle" dépend néanmoins pour sa plus grande part de l'implantation. Elle règle l'entité sur laquelle va se focaliser l'attention à un instant donné notamment. On

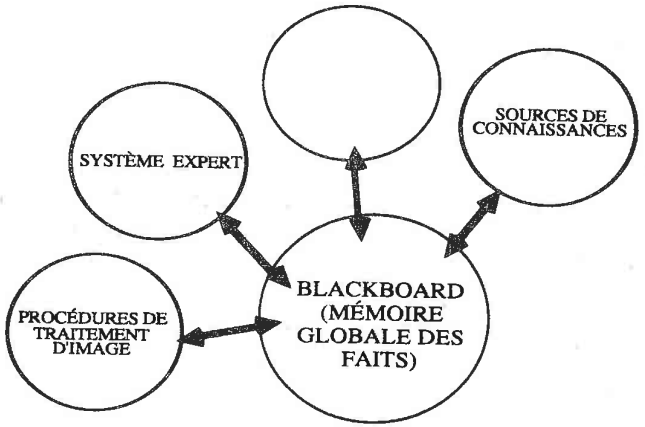


Figure 3.5. Structure d'un système blackboard

peut imaginer ce contrôle à l'intérieur du blackboard, des sources de connaissances, dans un module séparé ou comme une combinaison de ces trois possibilités [Engelmore 88b].

De fait, le blackboard ne correspond plus à une entité informatique, il est devenu un concept, au même titre que les systèmes d'exploitation, par lequel on désigne un gestionnaire de sources de connaissances. Des réalisations récentes incluent le raisonnement hypothétique et temporel [Laasri 88].

Les émules

L'architecture blackboard a connu un certain succès et a donné naissance à un grand nombre de successeurs. Sur ce sujet, on pourra consulter [Engelmore 88a]. Parmi toutes les réalisations, ce sont [Gong 88a,Gong 88b,Terry 88] qui ont le plus inspiré notre travail.

Le travail de [Gong 88a] a porté sur la construction d'une société de spécialistes pour l'interprétation du chinois parlé. Ce système est constitué de plusieurs associations permettant de traiter un niveau d'abstraction de l'interprétation. Ces associations sont formées de spécialistes indépendants partageant une base de données grâce à laquelle elles échangent

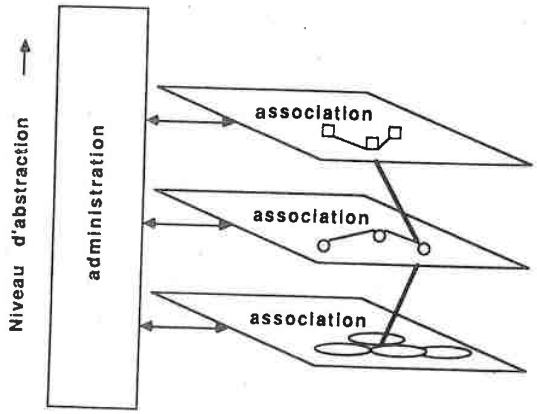


Figure 3.6. Structure d'une société de spécialistes

leurs informations et peuvent construire une solution.

Une administration actionne ces associations et transmet, par l'intermédiaire de messages, à chacune des autres associations, les solutions partielles déjà construites. La structure générale du système est comparable à des blackboards simplifiés mis en parallèle. La figure 3.6, page 61 représente le schéma de cette structure.

Pour l'interprétation de la parole continue, l'auteur distingue quatre niveaux :

1. le niveau acoustique,
2. le domaine phonétique,
3. le niveau phonologico-lexical et
4. le niveau syntaxico-sémantique.

Le niveau syntaxico-sémantique comprend notamment les spécialistes suivants :

- un mécanisme de chaînage-avant,
- un mécanisme de chaînage-arrière,
- un mécanisme de recherche en faisceau,

- un mécanisme de fusion d'arbre d'instanciation partielle...

L'objectif du système CRYSALIS de [Terry 88] est de retrouver la structure spatiale de protéines dont on connaît la composition en terme d'acides aminés ainsi que des données d'électro-densité. Le système procède essentiellement par focalisation de l'attention pour réduire le nombre d'inférences possibles. Toutes ses sources de connaissances prennent la forme de règles du type :

SI *condition* ALORS *action*

Elles s'organisent autour d'un contrôle hiérarchique qui est une source de connaissance particulière. L'aspect est rendu très homogène, ce qui conduit à une très grande simplicité d'écriture des connaissances. CRYSALIS est une tentative de construction de système hiérarchique à règles de production.

Notre solution

Le système que nous avons construit s'inspire des architectures blackboard[8] [Nugues 88b]. Le trait principal des blackboards est la possibilité de manipuler des procédures et des données de nature très différente. Notre idée de départ était d'homogénéiser la structure pour la rendre plus claire. En effet, si la possibilité d'inclure des méthodes qui tiennent de plusieurs modes de raisonnement peut être un avantage, elle présente aussi une grande difficulté de réalisation pratique, notamment pour ce qui concerne les conflits ainsi que le désordre qui peut régner dans l'exécution des procédures et qui entraînent l'imprévisibilité relative du fonctionnement.

Les sources de connaissances de notre système ont toutes le même aspect extérieur : elle prennent la forme d'un système de production hiérarchique, dont les règles incluent des procédures dans leur partie action. Une source de connaissance particulière effectue le contrôle : le *pilote*. La figure 3.7 page 63 représente le schéma du système.

Pour résoudre notre problème – l'identification de constellations de protéines – nous avons défini quatre modules en dessous du pilote :

- l'accès à l'image, la localisation de zones pertinentes et le traitement de bas niveau de l'image,
- l'appariement structurel partiel,
- la qualification, l'évaluation et l'interprétation,
- l'approximation numérique.

8. Il est en fait facile d'utiliser ce terme car il recouvre un grand nombre d'organisations possibles

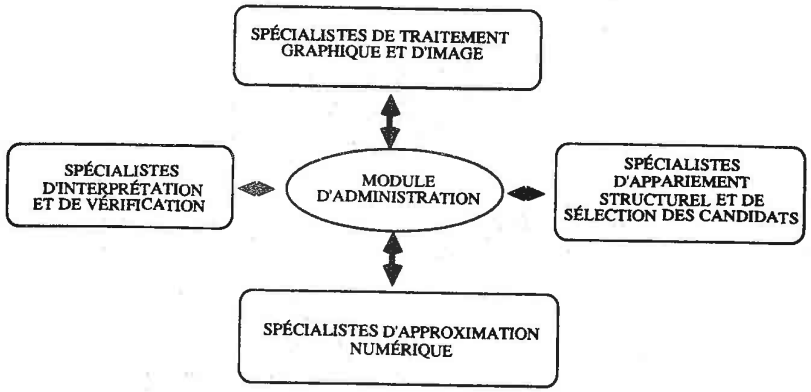


Figure 3.7. Structure du système

Ces modules sont activables séquentiellement par le pilote. C'est lui qui peut définir l'ordre d'enchaînement en fonction de ses buts : identifier une constellation, et des données qui lui parviennent des modules inférieurs. Un module particulier ne peut manipuler, modifier ou détruire que les instances d'un sous-ensemble des classes qui correspond à son domaine de compétence.

La stratégie d'un module particulier est donnée par un agenda qui définit l'ensemble des règles applicables à l'intérieur de ce module par ordre de priorité. Un élément de l'agenda peut définir plusieurs règles. Cet agenda est dynamique. On peut le créer ou le modifier de trois manières :

1. Avant l'activation du module, lorsque c'est possible, le module de pilotage définit l'agenda en fonction des données courantes.
2. Il existe un certain nombre de cas qui peuvent ne pas correspondre à une situation prévue et chaque module possède un agenda par défaut.
3. L'agenda est modifié dynamiquement par la partie action des règles en fonction des événements.

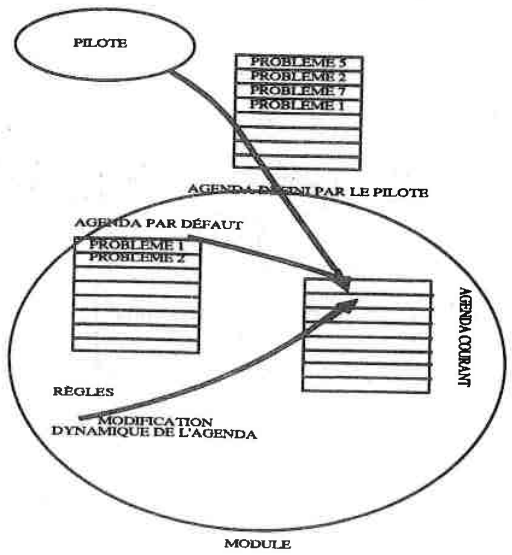


Figure 3.8. Structure des agendas

La figure 3.8 page 64 résumé cette description.

En nous référant à la terminologie utilisée dans les systèmes blackboards, la stratégie du système est une combinaison de mode de pilotage par les événements et de pilotage par les règles.

Le pilotage par les événements correspondrait à la tentative d'unifier un événement particulier à la partie gauche de toutes les règles du module et de les déclencher. Le pilotage par les modèles consiterait en un filtrage d'un événement dans une ou plusieurs classes par l'ensemble des règles.

L'utilisation des agendas permet de restreindre la portée du chaînage-avant de notre système et de définir de manière souple les objectifs de chacun des modules. La direction générale à partir d'un module de pilotage évite la plupart des conflits entre les modules subalternes. Elle permet l'enchaînement des niveaux de traitements appropriés, c'est à dire ici des modules, et grâce à la gestion adéquates des agendas, par exemple, on opère facilement

les retours en arrière en cas d'échec. Une amélioration possible du système serait d'implanter une recherche en faisceau ou bien la possibilité de raisonnement hypothétique [deKleer 86] au lieu de procéder par retour en arrière.

La communication entre les modules de traitement et le module de pilotage se fait par des messages qui prennent la forme d'une classe particulière d'objets. Ces messages reprennent pour l'essentiel ceux décrits par [Gong 88a]

Le système a été implanté grâce au générateur Iroise@CEWB [Cog].

3.3 L'implantation des règles d'identification des protéines

Pour identifier les protéines ou les constellations de protéines, nous opérons à partir des sommets détectés. Nous connaissons alors leur amplitude, leur abscisse et ordonnée et l'intégrale de la tache connexe à l'intérieur de laquelle ils se trouvent. Pour réduire le nombre de sommets, on a procédé à une focalisation de l'attention dans une zone où les points recherchés se situent vraisemblablement. Si ce n'était pas le cas, l'identification échouera et on opérera un retour en arrière en agrandissant ou en translatant la zone d'analyse.

À partir de ces points qu'on rassemble dans une classe banalisée : **$sommet**, on filtre une configuration potentielle. Cette dernière correpond à la forme géométrique complète ou seulement partielle des sommets. On est souvent obligé de rechercher ces dernières, car la totalité des points théoriques est rarement présente. Par exemple, lorsque une petite tache se trouve souvent masquée par une plus grosse, ou bien apparaît, mais en dessous du seuil de détection. Cette recherche s'effectue avec une stratégie de retour en arrière. On tente d'intancier la constellation la plus complète possible puis des constellations partielles jusqu'à une forme minimale tolérable.

Il existe souvent plusieurs configurations possibles pour un même problème, notamment lorsqu'il se présente un polymorphisme. La stratégie est aussi le retour en arrière. Le choix de l'ordre des règles est déterminé par l'agenda et il correspond par valeurs décroissantes, aux probabilités d'apparition des configurations. Ces probabilités sont estimées par l'expert.

Le filtrage proprement dit d'une configuration s'opère par l'unification d'une conjonction de conditions avec des contraintes géométriques sur la classe **$sommet**. On cherche d'abord les points remarquables les plus hauts et les plus constants, présents dans cette configuration et en dernier lieu, les points plus instables. Comme on ne peut pas toujours détecter ces derniers dans des conditions normales, on tâche d'écrire une forme de filtrage minimale correspondant à une configuration partielle. Cette configuration sera évaluée par le module de qualification. Il déterminera les points manquants et prédira leur position. À ce moment, le système pourra opérer de trois manières :

- il demandera une nouvelle activation du module de localisation avec un seuil de détec-

tion plus faible pour les points potentiels ;

- il prédira la positions des points manquants et les insérera de manière artificielle au milieu des points réels. Ceci est s'effectue par exemple lorsque une petite protéine est souvent mal résolue dans son environnement ;
- il modifiera les dimensions du cadre d'analyse et relancera le processus.

Nous donnons une règle de détection simplifiée de la configuration de la figure 3.9 page 67. Elle s'écrira :

($sommet ?s1 (amplitude ?a1) **et** (abscisse ?x1) **et** (ordonnée ?y1))
(pas de
($sommet ?s2
(amplitude ?a2 **(contrainte** (> ?a1))) **et**
(abscisse ?x2 **(contrainte** (> (?x1 − valx1)) **et** (< (?x1 + valx1)))) **et**
(ordonnée ?y2 **(contrainte** (> (?y1 − valy1)) **et** (< (?y1 + valy1))))))
($sommet ?s3 (amplitude ?a3) **et**
(abscisse ?x3 **(contrainte** (> (?x1 + valx2)) **et** (< (?x1 + valx3)))) **et**
(ordonnée ?y3 **(contrainte** (> (?y1 + valy2)) **et** (< (?y1 + valy3)))))
(pas de
($sommet ?s4
(amplitude ?a4 **(contrainte** (> ?a3))) **et**
(abscisse ?x4 **(contrainte** (> (?x1 + valx2)) **et** (< (?x1 + valx3)))) **et**
(ordonnée ?y4 **(contrainte** (> (?y1 + valy2)) **et** (< (?y1 + valy3))))))

3.4 L'interprétation

L'interprétation est liée – pour l'instant – à la détection de protéines particulières ou de constellations spécifiques. Sur ce point, on pourra citer les travaux de Vincens [Vincens 87a, page 171] sur la mucoviscidose où un certain nombre de protéines présentent des intensités différentes sur les gels normaux et les gels de fœtus atteints. Notre architecture possède des avantages évidents par rapport aux analyseurs classiques puisque elle permet l'identification des protéines qu'on aura reconnues comme responsables de tel ou tel état physique et donc l'interprétation des pathologies. Malheureusement la connaissance sur la plupart des protéines des gels reste, à présent, embryonnaire et l'implantation de capacités interprétatives demeure rudimentaire, du type :

SI constellation présente **ALORS** état pathologique

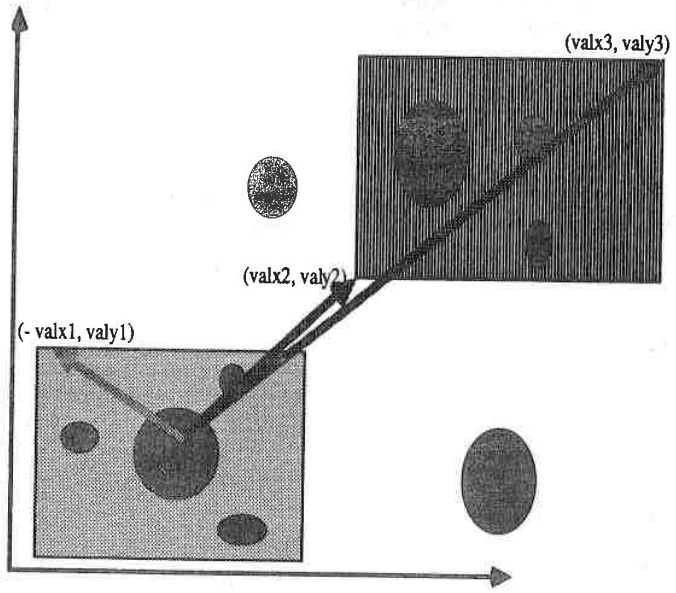


Figure 3.9. Exemple théorique de configuration

Lorsque celle-ci se développera, on pourra facilement l'inclure.

3.5 Pour une nouvelle façon de concevoir les bases de données d'électrophorèses bidimensionnelles

La technique d'électophorèse bidimensionnelle permet l'extraction potentielle d'une quantité considérable d'informations sur les protéines et les tissus. Toutes ces informations n'ont pu jusqu'à présent être exploitées, d'une part à cause du très grand nombre d'investigations à mener et d'autre part à cause de la difficulté relative d'obtention des gels qu'entraîne cette technique. Il est peu vraisemblable qu'on puisse reproduire, pour chaque protéine à étudier,

au moment voulu, toutes les manipulations nécessaires pour caractériser ses fonctionnalités et ses conditions d'expression. Par contre, avec les disques optiques de grande capacité, on peut conserver les images de ces expériences et les rappeler au besoin pour identifier les fonctions des protéines visibles sur les gels.

Envisager ces perspectives impose de coordonner quatre points.

1. Il faut d'abord pouvoir disposer de bases de données d'images où on conservera les gels de divers laboratoires,
2. il faut pouvoir associer à chacun de ces gels, des informations annexes, sur les patients, les conditions de l'expérimentation..., les plus complètes,
3. on doit pouvoir interroger la base de données de manière élaborée sur le contenu de l'image
4. et enfin on doit pouvoir relier les protéines des gels à d'autres données telles que la composition en acides aminés, le classement de lignées de cellules...

Ce dernier point dépasse le cadre de notre système. D'autres auteurs ont tenté de le clarifier [Vincens 87c,Garrels 84,Anderson 82]. Une tentative d'implantation de bases de données a été effectuée dans le système PD-QUEST de la firme Protein Databases mais le contenu de l'image doit être codé manuellement par un opérateur.

3.5.1 Les bases de données d'images

On peut considérer une base de données d'images de plusieurs manières. [Tamura 84] définit trois façons de les concevoir :

- comme une extension des bases de données classiques avec un type particulier supplémentaire qui serait l'image. La recherche des données ne concerne alors que les attributs de chaque entrée. Pour ce qui touche directement l'image, il ne peut s'agir que de son nom, ou des données complémentaires qu'on jugera utiles, comme les dimensions, la scène représentée..., et qu'un opérateur devra rentrer en même temps que l'image elle-même. La seule manipulation réelle permise est la visualisuation.
- Comme le champ potentiel d'opérations de traitement d'image. La base de données est alors un système de fichiers, muni de procédures pouvant extraire des informations telles que la moyenne, l'histogramme, le nombre d'extrema locaux ou d'autres paramètres plus élaborés. L'image n'est plus une entité figée de la base de données mais une source d'informations dynamiques.
- Comme un domaine d'étude où l'image joue un rôle. C'est le cas de la télédetection, de la cartographie, de l'imagerie médicale...Chaque métier ou activité trouvera alors

les fonctionnalités à inclure dans une base de données d'images pour servir ses besoins propres : la mise en relation des images prises dans des longueurs d'ondes différentes pour la télédétection, la coloration des niveaux topographiques pour la cartographie, la transmission des images à différents services pour l'imagerie médicale...Ce dernier point peut combiner, en fait, les deux premiers.

Notre démarche est essentiellement liée au domaine médical. Elle aborde les images suivant le point de vue du traitement du signal et de la reconnaissance des formes. L'objectif est de pouvoir transformer les systèmes existants en véritable bases de données dont les champs seraient dynamiques et modifiables par l'image elle-même.

3.5.2 Les perspectives pour les électrophorèses bidimensionnelles

Les bases de données sur les gels d'électrophorèse bidimensionnelles peuvent comprendre plusieurs types d'informations qu'on peut ranger par ordre de connaissance croissante. [Anderson 88], notamment, distingue trois niveaux de données sur les protéines :

1. le niveau quantitatif, qui se borne à considérer un groupe de gels appariés et à indiquer la quantité de matière pour chaque protéine suivant les gels, quelques tests statistiques tels que les écart-types...
2. le niveau "connecté", où on associe à une protéine, chaque fois que c'est possible, une information ou une citation extraite de la littérature. Cette construction permet d'associer directement une protéine à un ensemble de références bibliographiques mais n'offre aucune structure.
3. Enfin, le niveau "annoté", qui est une tentative de structuration par attributs. Parmi eux l'attribut *propriété* où on pourra trouver cytoplasmique, mitochondrial etc,...

Ce dernier niveau est le seul souhaitable car c'est le seul qui puisse se prêter à des interrogations faciles par les systèmes de gestion de bases de données actuels, tel que les systèmes relationnels [Codd 70]. Ceci d'autant plus qu'on peut facilement l'associer aux systèmes graphiques et par exemple demander l'affichage des protéines mitochondriales d'un gel donné ou d'une série de gels appariées.

Nous proposons [Nugues 88a] de compléter ce modèle par une possibilité de recherche et d'identification automatique de protéines sur les images de la base. Le modèle précédent de bases de données exige en effet une intervention manuelle constante lors de l'introduction d'un nouveau gel, surtout lorsque l'ensemble des gels est quelque peu hétérogène, ce qui est inévitable dans une grosse base. L'identification de toutes les protéines du nouveau gel par comparaison visuelle avec un gel de référence devient alors obligatoire.

Dans notre modèle, il suffit de décrire les protéines clairement identifiées par leurs moyens d'accès en se fondant sur le formalisme que nous avons développé au cours de ce chapitre. L'identification des constellations sur le gel peut alors être automatique. Les techniques d'appariements iconiques deviennent accessoires dans le dépouillement des gels. Elles interviennent uniquement pour l'investigation des taches inconnues, dont le nombre, inéluctablement se réduira. Le contenu d'une image n'est plus figé et il devient la source potentielle d'une quantité considérable d'informations.

On peut concevoir l'organisation des bases de deux manières :

- soit en considérant l'organisation classique, c'est à dire comme une *collection d'images*, de préférence de type "annoté", auquelles on ajoute un langage d'accès aux constellations qui puisse permettre d'identifier les protéines intéressantes ou les phénotypes.

- Soit comme une *collection de moyens d'accès* à différentes constellations. Le champ d'expérience est alors une seule image, provenant, par exemple, d'un patient dont on aurait à déterminer la pathologie et sur laquelle on rechercherait divers types de protéines ou de constellations potentielles.

L'extension d'une telle méthodologie permet de multiples applications telles que mise à jour d'une base de données sur un tissu lors de l'identification, ou du nommage d'une nouvelle protéine. Il suffira alors de la décrire en termes de relations pour la retrouver sur les images de la base de données et lui affecter son nom sur chaque gel. La sélection automatique d'un lot d'images de la base, et donc de patients, par une requête adéquate sur des formes de protéines précises. On peut parfaitement combiner ces interrogations par les moyens d'une algèbre relationnelle et rechercher des conjonctions de constellations ou des configurations s'exprimant par des formules plus complexes. Un laboratoire peut, par ailleurs, transmettre facilement son expertise à des laboratoires différents et leur permettre aussi d'identifier sans risques d'erreurs les protéines qu'il aura pu lui-même identifier.

Dans le cadre des applications cliniques, on peut, en face d'un gel reproduire à peu près n'importe quelle interprétation du moment qu'on peut la lier avec une forme précise sur l'image. Il existe une certain nombre d'états pathologiques qu'on peut déterminer par une électrophorèse bidimensionnelle. Il suffira de leur associer la méthode d'accès correspondante.

En conclusion, les méthodes d'identification que nous venons de proposer permettent d'associer et d'intégrer pleinement l'image d'électrophorèse aux bases de données. Ce ne sera plus un simple support visuel, mais une source d'informations manipulable automatiquement.

- Elles permettent l'exploitation rationnelle de l'expertise actuelle en évitant de réapparier les protéines connues.

- Au-delà des simples protéines, elles permettent l'identification des constellations plus complexes et ouvrent la voie à l'interprétation médicale liée à l'électrophorèse bidimensionnelle.

- Elles permettent une mise à jour souple de la connaissance contenue dans les bases de données.

- Elles autorisent un véritable partage de l'expertise de différents laboratoires.

4

Application : l'identification des apolipoprotéines sur un gel de plasma

La construction de notre système s'est faite en collaboration avec le Centre de médecine préventive de Vandœuvre dont l'un des sujets d'étude concerne la famille des apolipoprotéines. Ce sujet fait partie d'un cadre plus vaste qui est celui de la recherche sur les maladies cardio-vasculaires. La perspective du Centre de médecine préventive est celle de l'épidémiologie. Elle implique le dépouillement et l'analyse fastidieux d'enquêtes. L'ampleur de cette dernière est d'une centaine de familles ce qui correpond à environ 400 personnes. Pour chacun des patients examinés, on devra réaliser une séparation des protéines plasmatiques par électrophorèse bidimensionnelle puis repérer et quantifier les protéines qui nous intéressent... Notre système trouve son application dans l'automatisation de ces deux derniers points qui permettent la "lecture" des apolipoprotéines.

4.1 La famille des apolipoprotéines

Les apolipoprotéines [Sprecher 84] sont des protéines qui jouent un rôle important dans le métabolisme des lipides, notamment en se fixant sur les lipides ce qui permet leur ciculation dans le sang.

Les apolipoprotéines se divisent en familles et sont visibles sur les gels bidimensionnels ordinaires de plasma. On les classe en apoA-I, apoA-II, apoA-IV, apoC-2, apoC-3, apoD, apoE et apoH. La figure 4.1, page 74 représente une image de plasma où sont indiqués quelques apolipoprotéines.

L'objectif du Centre de médecine préventive est d'étudier les principales, c'est à dire les apo A-I, A-II, A-IV, E et C. Nous n'avons pas implanté, cependant, dans notre système l'identification des apoA-IV.

Chacune de ces familles apparaît sous la forme d'une constellation de points assez proches

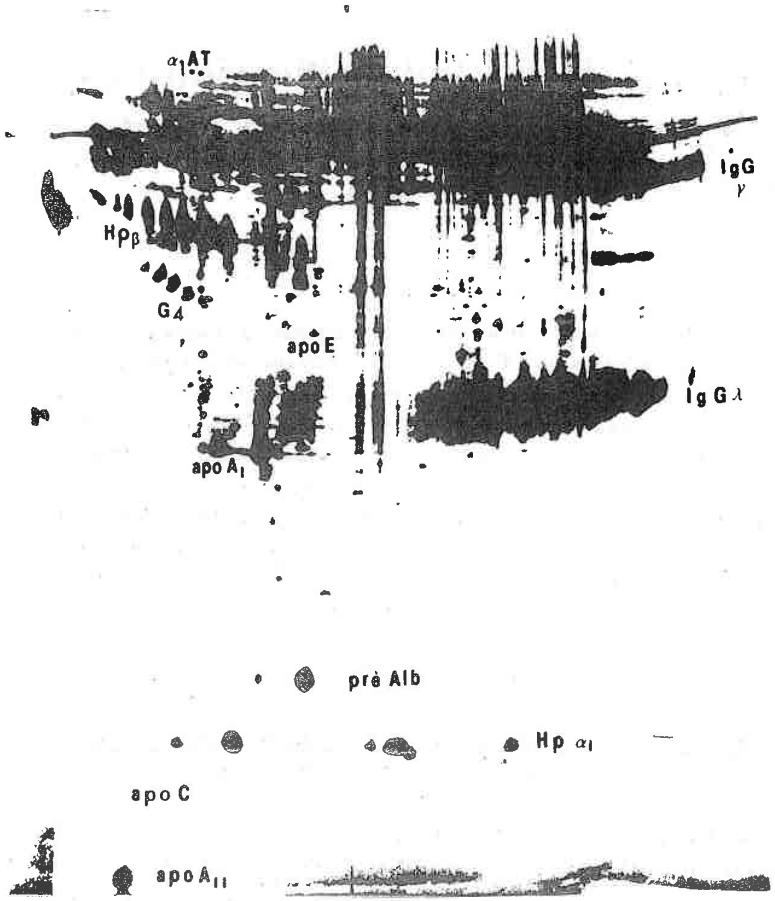


Figure 4.1. Électrophorèse bidimensionnelle de plasma

Figure 4.2. ApoA-I normale

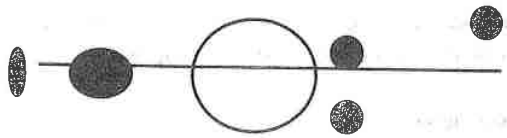


Figure 4.3. ApoA-I chez un patient atteint de la maladie de Tangier

qui prennent des emplacements constants. À chacun de ces emplacements la protéine peut être plus ou moins importante. Ainsi, l'apoA-I est une chaîne d'au plus six points dont la forme la plus fréquente est représentée sur la figure 4.2, page 75.

Cette forme n'est pas constante suivant les individus, elle présente un polymorphisme. Certaines de ces formes sont liées au phénotype génétique et peuvent ne pas avoir de conséquences pathologiques, d'autres sont liées à des maladies ainsi l'apoA-I apparaît sous une forme différente dans le cas de la maladie de Tangier, qui se caractérise par un affaiblissement de la tache majeure centrale. (figure 4.3, page 75)

Les variantes les plus fréquentes de ce polymorphisme sont connues et répertoriées et on a pu associer certaines d'entre elles à des maladies – comme la maladie de Tangier – ou à des variations des lipides plasmatiques, notamment pour ce qui concerne les apoE.

4.2 L'expertise

Un expert peut identifier avec certitude les familles d'apolipoprotéines sur un gel de plasma et déterminer les différentes isoformes. C'est cette extraction d'expertise qui a été notre premier travail. Nous avons répertorié les démarches de détermination des formes que prennent les apoliporotéines que rencontrent les analystes du Centre de médecine préventive.

Elles se fondent sur quelques points de confiance à partir desquels l'expert étaye son raisonnement. On procède par sous-buts successifs tels que :

1. repérer un ensemble de points faciles,
2. à l'intérieur de cet ensemble, localiser un point remarquable,
3. ce point remarquable permet d'en localiser un autre
4. qui permet d'identifier une constellation d'apolipoprotéines.

La collecte de cette expertise a représenté une aide essentielle pour la construction et la formalisation de notre pensée. En particulier en ce qui concerne la focalisation de l'attention et la décomposition des problèmes. C'est grâce à ces entretiens que nous avons pris conscience de l'impossibilité d'une solution par une méthode classique d'appariement.

4.3 La saisie des gels

Les gels à analyser ont une taille de $20cm \times 20cm$ environ. Nous avons effectué leur saisie de manière à pouvoir distinguer des points de $1mm$ de diamètre à mi-hauteur, ceci nous contraignait à une résolution de 5 points par mm.

La saisie a été faite avec un densitomètre à balayage de très grande résolution (jusqu'à 30 μ) de l'université de Strasbourg. Nous l'avons réalisé en tirant d'abord un négatif photographique à partir du gel original. On impressionne le négatif en l'appliquant à la surface du gel puis en l'illuminant pendant un temps donné. Nous avons ensuite numérisé les gels par transparence. (figure 4.4, page 77)

4.4 L'organisation informatique et un exemple de module

L'organisation respecte celle que nous avons décrite au chapitre précédent. Nous l'avons cependant adapté à cause de la similarité de détection des différentes familles. Nous avons reproduit pour chacune d'elles les quatre modules suivants :

- l'accès à l'image, la localisation de zones pertinentes et le traitement de bas niveau de l'image,

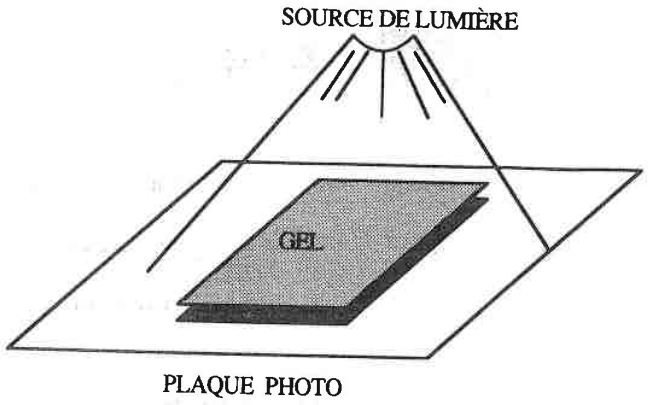


Figure 4.4. Dispositif de prise de vue

- l'appariement structurel partiel,
- la qualification, l'évaluation et l'interprétation,
- l'approximation numérique.

Donnons l'exemple du module de localisation. Il se compose des problèmes suivants :

- le *pilote interne*. Il détermine en fonction du message venant du pilote le type de cadrage et si il doit ou non effectuer un lissage de la zone ;
- les cadrages possibles qui sont au nombre de trois :
  1. le *cadrage automatique* extrait la zone la plus probable en fonction d'heuristiques ou de résultats précédents ;
  2. le *recadrage* opère en cas d'échec en agrandissant ou en translatant la zone ;
  3. le *cadrage quelconque* s'effectue avec des données provenant du module de pilotage ou d'un autre module ;

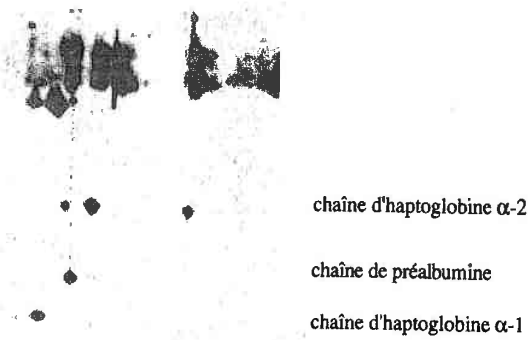


Figure 4.5. Haptoglobines et préalbumine

- la *détection des sommets* qui s'opère uniquement à l'intérieur du cadre déterminé précédemment ;
- le *lissage* des bruits de hautes fréquences qu'on effectue lorsque on a détecté trop de maxima pour une zone donnée ;
- la *sortie* qui rend la main au module de pilotage et lui adresse le message adéquat.

4.5 Le déroulement

Le déroulement de l'analyse se fonde sur le repérage de points de confiance, qui sont constitués ici des chaînes d'haptoglobines et de préalbumine pour les deux raisons suivantes :

- elles se localisent dans la partie inférieure gauche du gel qui est l'une des moins chargée du gel.
- Le point isoélectrique de la tache majeure de la chaîne de préalbumine permet le cadrage du gel et l'identification de l'abscisse d'un des points remarquables de la chaîne d'apolipoprotéines A-I : la proapolipoprotéine A-I.

Il est donc essentiel d'identifier d'abord et sans erreurs ces taches et en particulier la tache majeure de préalbumine. (figure 4.5, page 78)

Ces chaînes sont horizontales et normalement au nombre de trois, mais du fait des dégradations qui peuvent intervenir lors de la migration ou bien de variations entre les individus, parfois deux seulement sont visibles : les deux chaînes inférieures ou les deux chaînes supérieures. En tout état de cause, la chaîne de préalbumine reste présente.

On lance l'analyse dans le cadre où leur présence est la plus probable et on détecte tous les sommets au dessus d'un certain seuil afin d'éviter les éventuelles disparités du niveau du fond. On procède ensuite par instanciations partielles des modèles jusqu'à un niveau suffisant de confiance, qui est l'identification de la tache majeure de préalbumine. Si l'instanciation partielle des trois modèles échoue, le système redéclenchera la détection des sommets, soit avec un seuil plus faible dans le cas où on juge l'instanciation partielle possible, mais insuffisante, soit avec un cadre plus grand lors d'un échec total. Dans le cas d'un échec concernant une famille à identifier par la suite, le système peut revenir en arrière sur cette première zone.

À partir de là deux processus parallèles identifient :

1. les apolipoprotéines de plus grand poids moléculaire, les apoA-I, en commençant par la proapolipoprotéine A-I, ainsi que leur phénotype puis les apoE. Nous donnons un schéma de l'identification de la chaîne d'apoA-I sur la figure 4.6, page 80.

   En ce qui concerne les apoA-I, l'interprétation médicale est limitée à la découverte de la maladie de Tangier. Pour les apoE, on n'effectue la recherche que sur les présences potentielles des trois protéines centrales[1]. On détermine leur phénotype qui permet d'associer un risque cadio-vasculaire, ainsi que la présence où non de formes syalilées. Sur le figure 4.7 de la page 81 nous présentons les quatres étapes de cette analyse. Le cadre inférieur correspond à la zone des haptoglobines. Il vient ensuite le cadre des apolipoprotéines A-I dont on détermine la position en fonction des coordonnées des chaînes du cadre précedent. On analyse les apolipoprotéines E (dont le cadre est inversé et qui n'est pas visible sur la reproduction) et une fois identifiées on recherche juste au-dessus d'elles les syalilations éventuelles dans le petit cadre supérieur.

2. les apoliporotéines de poids plus faible, c'est à dire les apoC-2, apoC-3 et apoA-II, lorsqu'elles sont présentes sur les gels, en tenant compte des formes de leur polymorphisme les plus courantes.

À toutes ces étapes, on considère le meilleur candidat en revenant en arrière en cas d'échec. Une amélioration pourrait être l'implantation d'une recherche en faisceau ou d'un raisonnement hypothétique [deKleer 86] sur les différentes configurations potentielles d'haptoglobines par exemple.

1. il en existe cinq, en fait, mais la présence des deux autres est très rare

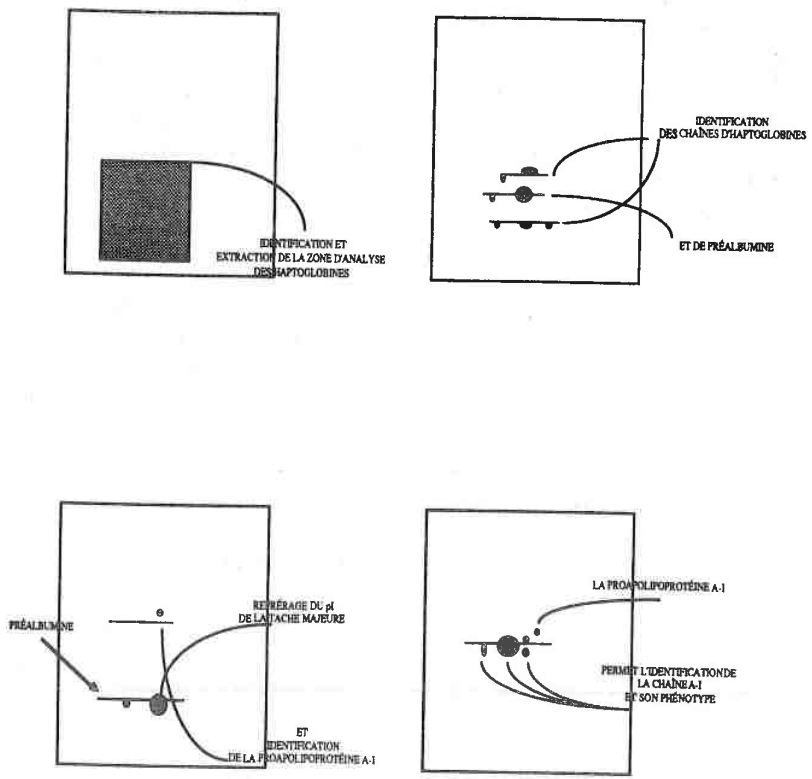


Figure 4.6. Schéma simplifié de l'identification de la chaîne d'apoA-I

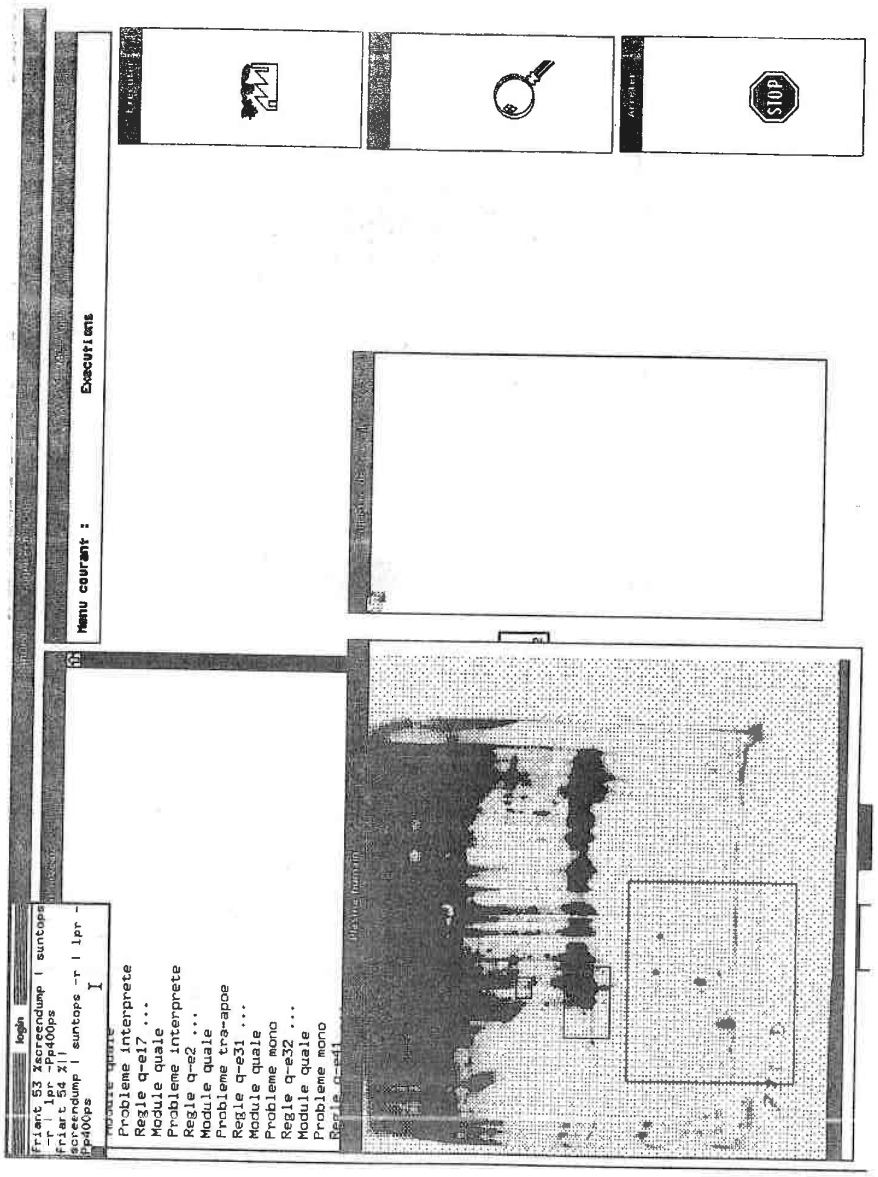


Figure 4.7. Les cadres d'analyse des haptoglobines, des apo A-I et des formes syalilées des apo E

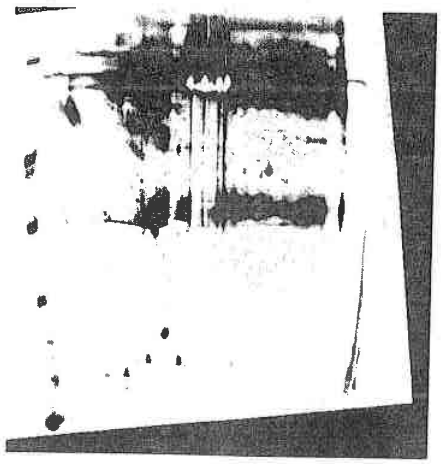


Figure 4.8. Gel 1

4.6 Les résultats

Nous avons évalué notre système sur cinq gels fournis par le Centre de médecine préventive, choisis au hasard, que nous reproduisons sur les pages 82, 83 et84. Nous n'avons pas activé le module de quantification qui n'avait pas de sens ici du fait de l'absence de protéines de calibration mais simplement considéré les capactités d'identification.

Nous avons construit le système à partir de l'expertise sur papier et des deux premiers gels qui nous ont servi de validation intermédiaire au cours de l'implantation des différents modules. Nous l'avons ensuite fait fonctionner, une fois écrit, sur les trois derniers gels.

Ces gels se présentent sous des aspects divers et sont très dégradés par rapport aux meilleures obtentions des laboratoires. Ce sont pourtant des exemplaires semblables auquels font face les analystes chargés de donner une interprétation clinique. On remarque notamment une grande disparité dans les niveaux de gris moyens, les deux premiers gels sont beaucoup plus clairs que les suivants, l'existence d'artéfacts tels que les numérotations ou les déchirements...

Toutes les protéines et les phénotypes ont été identifiés avec succès, sur tous les gels, aussi bien pour les haptoglobines, les apo A-I, E, A-II, C-2 et C-3 à l'exception d'une seule confusion sur le gel 3, pour l'apo A-I, où le système a désigné les deux points horizontaux

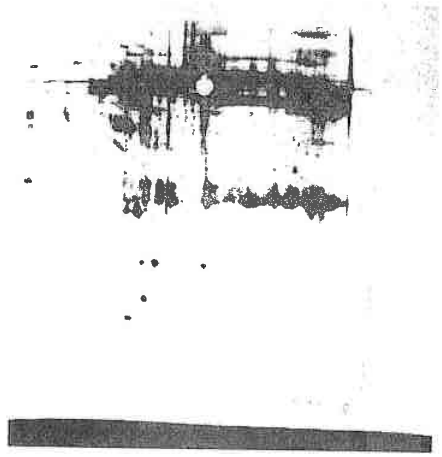


Figure 4.9. Gel 2

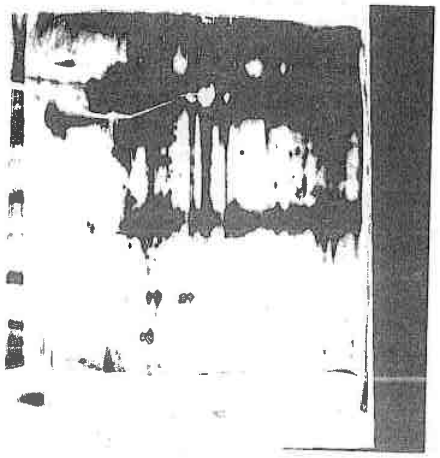


Figure 4.10. Gel 3

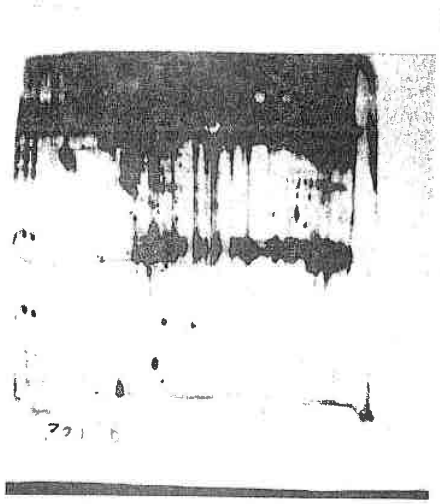


Figure 4.11. Gel 4

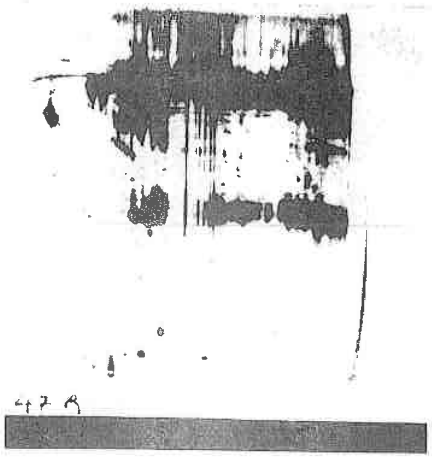


Figure 4.12. Gel 5

juste en-dessous comme les protéines en question.

C'est un résultat qu'aucune méthode d'appariement iconique ne pourrait sans doute obtenir.

De cette manipulation nous pouvons tirer les conclusions que l'emploi d'une expertise, lorsqu'elle existe, est sans conteste supérieure. De plus, dans le système présent les procédures de traitement d'image sont réduites au minimum, il n'y a pas de nettoyage ou de normalisation de niveau de gris...ces opérations préliminaires qui pourraient agir avant les opérations d'identification diminuraient sans doute encore les possibilités d'erreurs.

PARTIE II

VERS L'INTERPRÉTATION D'EXPÉRIENCES BIOLOGIQUES

Avant d'entrer de plain pied dans cette partie, nous allons préciser le sens du mot interprétation, car il recouvre définitions diverses suivant les sujets abordés. Cette confusion de vocabulaire se retrouve aussi parfois dans les esprits.

Dans le domaine de l'image informatique, le terme interprétation désigne la reconnaissance de scènes simples telles que la présence d'objets dans une pièce : un bureau, une table, une chaise, etc. qui puisse, par exemple, permettre à un robot de se déplacer [Thirion 88] dans cette scène ou d'étiqueter les différentes parties d'une scène extérieure en ciel,arbres, voitures, etc. [Ohta 85]. D'une certaine manière c'est une*interprétation perceptuelle*.

Ce terme d'interprétation, quand on aborde lamédecine, paraît abusif. En effet, il ne suffit pas de détecter des pustules rouges sur un visage ou même d'apercevoir un tache sombre sur une radiographie pulmonaire pour prétendre à une interprétation médicale. [Appel 87, Funk 87] proposent un résumé de l'interprétation médicale citant les travaux de [Fieschi 84, Newell 72,Elstein 78].

On y distingue deux étapes principales :

1. le médecin considère les signes les plus marquants d'un patient. Il ne recherchera plus la solution parmi le répertoire de toutes les maladies possibles, mais seulement parmi celles qui peuvent donner ces signes les plus visibles.

2. À partir des signes, le médecin émet des hypothèses qu'il essaie de confirmer avec d'autres signes ou données cliniques. Il élimine alors les hypothèses contradictoires. Au besoin, il répète ce point jusqu'à ce qu'il puisse établir un diagnostic.

Ce type de raisonnement à deux limitations :

1. il y a d'abord l'incertitude dans la connaissance des signes, des thérapeutiques ou des données ainsi que l'impossibilité matérielle d'accéder à toutes les informations.

2. Il y ensuite la limitation humaine à prendre an compte la totalité des informations disponibles. La prise de décision s'effectue toujours en privilégiant certaines données. C'est la rationnalité limitée.

L'interprétation ici est intimement liée au raisonnement avec des données incomplètes et à l'inférence statistique. On pourrait la qualifier *d'interprétation déductive*.

Il existe enfin un troisième niveau d'interprétation qui touche à l'acquisition de la connaissance. On emploie ce terme par exemple dans le cadre de dépouillements d'enquêtes statistiques – l'interprétation statistique – où on doit extraire les traits caractéristiques d'une masse grossière, éventuellement à partir d'un petit échantillon pour les inférer à un plus grand. Cette interprétation reprend l'aspect descriptif de l'interprétation d'images et l'aspect déductif de l'interprétation de signes médicaux. Dans la dernière partie de cette thèse, nous avons cherché

1
Les méthodes numériques

Dans l'histoire de *l'interprétation* des données par ordinateur, ce sont les méthodes numériques qui viennent en premier. Leur développement est parallèle à celui de l'informatique, bien que leur origine soit beaucoup plus ancienne. Pour cette raison d'antériorité nous les avons placé devant les méthodes symboliques, qui elles sont issues directement de l'intelligence artificielle et qui ont constitué le champ principal de notre étude.

1.1 La démarche statistique

Le support de cette analyse est formé des protéines ou d'un sous-ensemble des protéines d'une suite de gels. Les protéines que nous considérons dans cette suite de gels peuvent être :

1. identifiées, manuellement ou grâce à la reproduction automatique d'une expertise par le système que nous avons décrit précédemment,
2. ou bien simplement appariées, sur des critères de positions équivalentes, à l'intérieur de cette série de gels.

À chacune de ces protéines, on associe un ensemble de paramètres, variables suivant les gels, correspondant aux caractéristiques numériques des taches, c'est à dire, dans le cadre de notre analyse :

- l'amplitude A,
- les coordonnées (x, y),
- les écarts-type (σ_x, σ_y),
- l'intégrale de la densité de la tache calculée directement sur l'image ou à partir des paramètres précédents.

Le seul paramètre que nous avons réellement examiné est en fait l'intégrale densitométrique.

Les méthodes purement statistiques permettent trois types principaux d'analyses [Vincens 87a, pages 144–148] :

- l'étude des distributions de ces analyses. Elle permet de mettre en évidence la reproductibilité relative des gels d'électrophorèses bidimensionnelles. P. Vincens montre qu'il existe deux populations de taches, les unes très reproductibles, les autres moins.
- L'analyse de la variance de ces distributions,
- la normalisation des paramètres dans une série de gels. Pour que l'examen de ces paramètres soit pertinent, il faut qu'ils soient en rapport les uns avec les autres, or les densités optiques qu'on mesure sur l'image sont la conséquence, à la fois de la quantité de matière et du temps d'exposition nécessaire pour la révélation du gel. Qu'il s'agisse de sels d'argent or d'éléments radio-actifs. Il est essentiel, lorsqu'on désire comparer la quantité de matière d'une même protéine dans une série de gels, de pouvoir corriger les différences de temps d'exposition. On tâche autant que possible de les réduire en normalisant les protocoles de développement, mais on observe toujours des différences de teintes moyennes[1]. Ceci de toute manière ne vaut que pour les gels tirés par un même laboratoire, or l'ambition est de pouvoir interroger des bases de données sur l'expression des protéines et donc d'accéder aux quantités de matière de façon relativement universelle.

Le dernier point est fondamental pour l'exploitation généralisée de l'électrophorèse bidimensionnelle. Précédemment nous avons proposé une solution qui est d'introduire des quantités connues de protéines de calibration et de faire le rapport des intégrales des densités optiques des protéines à analyser, à ces protéines de calibration. La plupart des laboratoires ne pratiquent pas ceci de manière courante et des auteurs ont proposé d'autres méthodes reposant sur l'analyse statistique.

À partir d'un ensemble de gels provenant d'une même expérience, on choisit un sous-ensemble de protéines. On note $Y_{i,j}$, l'intégrale densitométrique de la $j^{ième}$ protéine du $i^{ième}$ gel.

[Kuick 87] propose trois types de normalisations dont l'une, fondée sur un calcul de droite de régression, donne des résultats nettement meilleurs que les autres. On suppose que la relation de correction entre les gels est linéaire, c'est à dire de la forme :

$$Y'_{i,j} = a + b.Y_{i,j}$$

Où $Y'_{i,j}$ est la valeur corrigée pour le $i^{ième}$ gel.

On choisit un gel parmi les autres qui ait une teinte de gris moyenne. On note S_i l'intégrale densitométrique du point i de ce gel et on calcule pour chaque gel j les paramètres a et b de

1. Communication orale au Centre de médecine préventive. Vandœuvre.

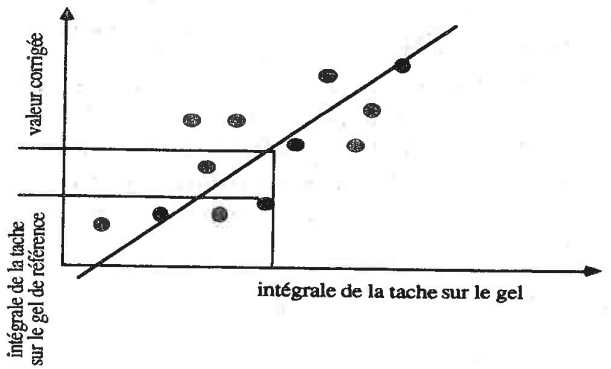


Figure 1.1. Correction d'intensité

la droite de régression entre les valeurs S_i du gel de référence et $Y_{i,j}$:

$$S_i = a + b.Y_{i,j}$$

(figure 1.1, page 93)

Cette méthode est relativement robuste car ses résultats ne s'améliorent pas de façon sensible en itérant l'opération sur un sous-ensemble de taches ayant les meilleurs coefficients de variation[2].

Une autre méthode, celle-ci très simple, est utilisée par le système PD-QUEST pour normaliser les intensités. Elle consiste à diviser l'intégrale de chaque tache par l'intégrale de toute l'image. Nous avons utilisé des données traitées par cette méthode qui semble produire des résultats corrects avec des protocoles d'obtention des gels très stricts.

1.2 L'analyse des données multidimensionnelles

Nous avons considéré deux types d'analyse de données multidimensionnelles, *l'analyse en composantes principales* et *l'analyse factorielle des correspondances* [Lebart 82].

2. Défini comme l'écart-type sur la moyenne

Ces deux techniques permettent de réduire un tableau de données de dimension n à une dimension inférieure : 2, 3 ou 4, par la projection du nuage de points dans les conditions d'obtention originelles, sur les axes principaux d'inertie qui portent le maximum d'information. Elles se fondent sur l'hypothèse – *a priori* contestable – qu'on peut expliquer un individu dont on connaît beaucoup de paramètres par un nombre nettement moins grand de paramètres et éventuellement présenter simultanément les individus et les variables sur un plan de dimension 2. C'est cette possibilité graphique et visuelle qui a sans doute été responsable de la plus grande part du succès des méthodes factorielles.

Les deux méthodes partent d'un tableau contenant n variables et p individus. On suppose que les individus sont en ligne et les variables en colonne et on note $r_{i,j}$, la valeur de la variable j pour l'individu i.

1.2.1 L'analyse en composantes principales

Cette analyse s'effectue en cinq étapes :

1. on remplace chacun des éléments $r_{i,j}$ par leur valeur centrée : $(r_{i,j} - \bar{r}_j)/\sqrt{p}$ ou normée : $(r_{i,j} - \bar{r}_j)/\sigma_j\sqrt{p}$.
2. On calcule la matrice des corrélations : $C = {}^tX\,X$.
3. On diagonalise cette matrice dont les valeurs propres sont toujours positives.
4. On ordonne les valeurs propres dont les vecteurs propres constituent les nouvelles variables.
5. On obtient la réduction en projetant les données sur l'espace défini par les deux, trois ou quatre vecteurs propres dont les valeurs propres sont les plus importantes.

1.2.2 L'analyse factorielle des correspondance

La procédure de cette méthode est la même que la précédente à l'exception de la distance utilisée qui est le χ^2 [Benzecri 76]. On note :

$$k = \sum_i \sum_j r_{i,j}$$

la somme totale du tableau,

$$f_{i,j} = r_{i,j}/k$$

les fréquences relatives du tableau,

$$f_{i.} = \sum_j f_{i,j}$$

les fréquences selon les lignes,

$$f_{.j} = \sum_i f_{i,j}$$

les fréquences selon les colonnes.

La méthode devient :

1. on remplace chacun des éléments $r_{i,j}$ par la valeur $f_{i,j}/\sqrt{f_{i.}f_{.j}}$.
2. On calcule la matrice des corrélations : $C = {}^tX\,X$.
3. On diagonalise cette matrice et on obtient les valeurs propres dont on écarte la valeur 1.
4. On ordonne les valeurs propres dont les vecteurs propres constituent les nouvelles variables.
5. On obtient la réduction en projetant les données sur l'espace défini par les deux, trois ou quatre vecteurs propres dont les valeurs propres sont les plus importantes.

1.2.3 L'utilisation de ces méthodes dans le cadre de l'électrophorèse bidimensionnelle

Dans le cadre de ces méthodes, nous ne considérons que les intégrales densitométriques des protéines supposées normalisées.

Les tableaux de données prennent la forme suivante[3,4] :

numéro de la protéine	numéro du gel									
	1	2	3	4	5	6	7	8	9	10
1	700	500	na	100	1000	600	400	na	400	400
2	na	na	na	na	na	na	na	100	200	na
3										
4										

Ces méthodes permettent un très grand nombre d'applications :

- en supposant que les groupes de protéines correspondant aux mêmes fonctions cellulaires variaient de façon similaire au cours de leur développement, on a pu établir une carte fonctionnelle des protéines, à partir de données temporelles de tissus [Rabilloux 85].
- Pierre Vincens [Vincens 87a,Tarroux 87] les a utilisées pour comparer des liquides amniotiques sains et pathologiques, des lignées cellulaires et pour constituer des cartes fonctionnelles en soumettant les cellules à l'action de divers effecteurs chimiques ou

3. na est l'abréviation de non apparié
4. Les chiffres correspondant aux protéines sont fictifs

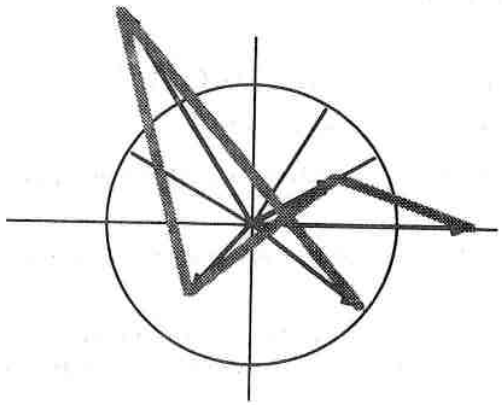


Figure 1.2. Étoile de projection

thermique. La représentation graphique ne peut cependant se faire sur un plan car les deux axes principaux ne portent pas assez d'information et on doit comparer les regroupements d'individus dans un espace de dimension 6. Ceci n'est pas aisé à tracer sur du papier et on a décrit graphiquement les coordonnées de chaque protéine par une "étoile" dont les branches sont séparées par des secteurs d'angles égaux et dont les coordonnées sur chaque axe de l'espace se traduisent par la longueur des branches. (Figure 1.2 page 96.)

- [Pun 88] a utilisé l'analyse des correspondances pour déterminer les protéines caractéristiques de gels de plasma pathologiques et sains. En effet, après avoir calculé les valeurs propres λ_k de la matrice de covariance, on peut connaître les individus qui contribuent le plus à leur formation par la formule :

$$\frac{f_i . S_{k,i}^2}{\lambda_k}$$

où $S_{k,i}$ représente la projection de la protéine i sur l'axe défini par la valeur propre λ_k. En triant les protéines ayant les plus fortes contributions pour la plus grande valeur propre, on parvient à déterminer des protéines discriminantes.

1.3 La taxinomie numérique

La taxinomie numérique définit les classifications qu'on peut obtenir d'individus décrits par des paramètres numériques tels que leur taille, leur poids,..., ou éventuellement par une matrice de similarité ou de dissimilarité. Elles ont donné naissance à une littérature très importante que nous ne pourrons examiner ici dans tout son détail. On pourra consulter sur ce point [Jambu 78]. Dans ce paragraphe nous présentons les grandes lignes des principales méthodes taxinomiques pour l'esprit qu'elle ont contribué à faire naître. L'utilisation intensive de l'ordinateur a largement permis leur extension et la création de nouveaux concepts qui dépassent le domaine des nombres. Telles qu'elles sont actuellement appliquées, leur défaut essentiel est d'être liée à la notion de distance numérique, sans inclure aucune autre connaissance. On pourrait cependant imaginer des les étendre aux données symboliques car ces méthodes ont pour support – dans la plupart des cas – des procédures algorithmiques itérables où intervient une distance entre des individus ou des groupes d'individus et non pas une équation tenant en une ligne qui ne lierait que des chiffres. Nous tenterons d'exposer dans le prochain chapitre dans quelle mesure on peut reprendre une partie de ces algorithmes.

1.3.1 La classification ascendante hiérarchique

Les concepts de cette méthode sont simples ; il s'agit de regrouper itérativement les éléments à classer en réduisant d'une unité à chaque étape de l'itération leur nombre. Pour cela, on fusionne les deux éléments les plus proches en un nouveau. La condition d'application est de disposer d'une distance entre éléments et entre groupes d'eléments.

Les principales distances entre deux éléments notés $(x_1, \ldots, x_n)$ et $(x_1, \ldots, y_n)$ sont :

- La *distance des blocs* [5] qui se définit par

$$\sum_{i=1}^{n} | x_i - y_j |$$

- La *distance euclidienne* qui se définit par

$$\sqrt{\sum_{i=1}^{n} (x_i - y_j)^2}$$

La distance entre un élément particulier et un groupe d'éléments définissent, en fait, les critères d'agrégation. Parmi les plus utilisées, il y a :

- la *distance moyenne* entre un élément x et un ensemble E qui se définit par :

$$\sum_{y \in E} d(x, y)/Card(E).$$

5. Ou distance d_1 dans certains manuels

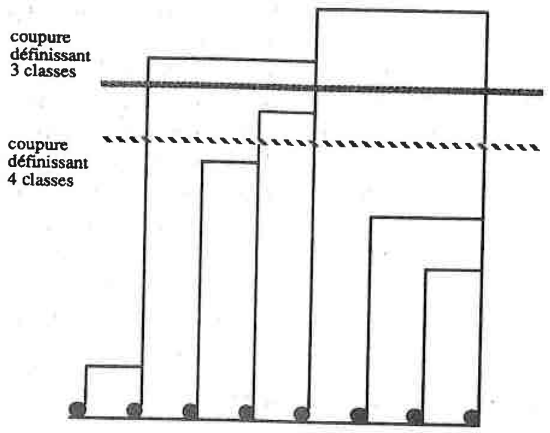


Figure 1.3. Dendogramme

- Le *saut minimal*, qui définit la distance entre un élément x et un ensemble E par :

$$\min_{y \in E} d(x, y).$$

- La *distance du diamètre*, qui définit la distance entre un élément x et un ensemble E par :

$$\max_{y \in E} d(x, y).$$

- La *variance minimale* qui définit la fusion de deux classes sur le critère de la perte minimale d'inertie interclasse au cours du regroupement. L'inertie interclasse étant la somme des inerties de toutes les classes par rapport au centre de gravité de l'ensemble total.

Le résultat final peut alors se mettre sous la forme graphique d'un dendogramme. (Figure 1.3 page 98.)

L'avantage de cette méthode réside dans la facilité de partition dans un nombre variable de classes. Il suffit de couper horizontalement le dendogramme à la hauteur désirée pour les

obtenir. Leur désavantage, dans le cadre de l'électrophorèse, est de ne pas avoir produit de résultats totalement probants.

[Westerbrink 84] a tenté de séparer des gels issus de sujets sains et atteints de plusieurs types de cancers des ganglions lymphatiques : les lymphomes centroblastiques et centrocytiques d'une part et les lymphogranulomatoses d'autre part. Il définit la distance entre un gel A et un gel B par la formule :

$$\frac{m(A) + m(B) - m(A \cap B)}{m(A \cap B)}$$

où $m(A)$ est le nombre de taches de A, $m(B)$ est le nombre de taches de B et $m(A \cap B)$ le nombre de taches appariées entre A et B.

Il effectue la classification successivement selon les critères de saut minimal et de distance du diamètre. Les dendogrammes reproduits[6] permettent de séparer les sujets sains des autres, mais échouent à différencier les types de lymphomes. Cet échec est complet, car il ne s'agit pas simplement d'un individu égaré dans un mauvais groupe, mais d'un mélange total entre les deux classes de lymphomes.

Cette manipulation ne met pas fondamentalement en cause l'utilisation des méthodes hiérarchiques pour l'électrophorèse bidimensionnelle car on pourra objecter que la mesure de distance proposée par K. Westerbrink est assez simpliste.

[Appel 87,Appel 88] a expérimenté une méthode de classification ascendante hiérarchique sur des cellules de foie de rats sains et atteints de cirrhose. Il définit chaque gel comme un vecteur où les paramètres sont les protéines identifiées de la même manière pour tous les gels. Les valeurs des protéines correspondent à leur intégrale densitométrique. R. Appel a réalisé sa manipulation avec la distance euclidienne et la distance des blocs entre les éléments et la distance moyenne entre les classes.

Les résultats obtenus sont les mêmes quelque soit les distances et ne parviennent pas à effectuer une partition correcte, cependant l'examen des gels mals classés montre qu'ils sont particulièrement bruités ou bien qu'ils comportent une forte quantité de faux appariements. Cette expérience ne rejette donc pas totalement la méthode de classification hiérarchique mais elle prouve sa sensibilité au bruit. Elle comporte aussi le désavantage de ne pas fournir de description finale des classes trouvées en donnant par exemples les protéines caractéristiques.

[Pun 88] a réalisé une manipulation similaire à celle de R. Appel avec des tissus différents et en réduisant les données à un seul axe par l'intermédiaire d'une analyse factorielle. Il a finalement obtenu une classification correcte.

6. pages 81 et 82 de sa thèse

1.3.2 Les méthodes dynamiques

Les méthodes dynamiques tiennent d'un esprit très différent des méthodes hiérarchiques. Elles partent d'un nombre connu de classes et doivent parvenir à partitionner l'ensemble à traiter en maximisant un critère. C'est uniquement pour des raisons historiques que nous les présentons comme une taxinomie numérique car en fait la notion de distance apparaît peu et on les utilise aussi comme un "moteur" d'itération pour des méthodes symboliques.

Le principe est assez simple. Soit k le nombre de classes à former. On choisit au départ k centres $c_1, \ldots, c_k$ au hasard ou selon une heuristique parmi les éléments de l'ensemble et on agrège les éléments restants autour du centre le plus proche. On remplace les centres initiaux $c_1, \ldots, c_k$ par les centres de gravité des groupes formés et on itère la méthode. L'arrêt intervient lorsque les centres de gravité sont relativement stables. On montre que dans le cadre numérique, l'algorithme converge toujours.

L'algorithme des *nuées dynamiques* [Diday 83] utilise le même principe d'itération mais remplace les centres mobiles par des noyaux mobiles qui sont formés d'un nombre fixé d'éléments (n par exemple). On regroupe les classes autour des noyaux initiaux et on procéde à l'itération en les remplaçant par les n éléments des classes ainsi formées, les plus proches des nouveaux centres de gravité.

Pour ces deux méthodes, les partitions finales dépendent du choix du noyau initial et on devra donc effectuer plusieurs traitements en modifiant ce choix et comparer les résultats finaux obtenus.

1.4 En conclusion

On peut appliquer les méthodes numériques classiques avec succès à l'électrophorèse bidimensionnelle. Elles ont prouvé leur fiabilité dans ce domaine et elles restent un instrument fondamental du dépouillement traditionnel de série de gels, que ça soit, entre autres, pour la réduction et la visualisation des données par des méthodes multidimensionnelles ou pour la classification. Elle présentent néanmoins le désavantage d'avoir un langage de description de connaissances très pauvre et de ne pas se prêter agréablement à une interprétation fine. P. Vincens [Vincens 87a, page 143 du document] déclare notamment qu'elles peuvent mal :

> *prendre en compte des données externes, ou [...] localiser certains événements sur le gel*

puis que :

> *les méthodes dérivées de l'intelligence artificielle [lui] paraissent adéquates pour l'étude de ces problèmes.*

C'est ce domaine sur lequel nous avons fait porter nos efforts de recherche et que nous détaillons dans les chapitres suivants.

2
Les méthodes symboliques

Dans le chapitre précédent, nous avons abordé le traitement des résultats des électrophorèses bidimensionnelles par des méthodes numériques. On représente les objets à traiter et à interpréter par des vecteurs de chiffres. Ces méthodes se sont développées grâce à l'extension considérable des ordinateurs et la facilité de codage, de représentation et de manipulation des nombres par ces machines. Pourtant sur deux points essentiels, les méthodes numériques sont supplantées par les méthodes symboliques.

Le premier point tient à la restriction de la représentation. Pouvons-nous décrire le monde qui nous entoure par une suite de nombres ? La réponse va de soi et bien qu'on ne puisse prétendre à la vision objective de la nature, on peut admettre qu'il existe des manières plus "humaines" et plus directes que des chiffres pour formuler des problèmes. Nous savons représenter grâce aux techniques issues de l'intelligence artificielle des attributs tels que des couleurs ou des formes ou bien des relations telles que : *les animaux respirent* ou SI *oiseau* ALORS *vole*, ou bien des structures de classification déjà existantes, telles que des arbres taxinomiques... il serait ridicule de ne pas chercher à les exploiter afin d'enrichir de connaissances les domaines de classification.

Le second point est l'interprétation automatique qui s'effectue, quand on utilise des méthodes symboliques, par le ré-assemblage des concepts et des relations qui décrivent les objets en entrée. Lors de la classification qui intervient à la suite d'une expérience, on pilote toujours les opérations par ces concepts, on les retrouvent donc toujours à la fin de manière à ce qu'ils puissent justifier les formations des classes.

Enfin, il n'est pas question de mettre en procès les traitements classiques. Les méthodes symboliques n'éliminent pas les méthodes numériques, elles élargissent simplement leur cadre d'action, et il est parfaitement concevable d'inclure des algorithmes numériques à l'intérieur d'un ensemble décrit de manière symbolique.

En résumé, les méthodes symboliques étendent les capacités de classification des objets et des scènes au-delà des nombres.

2.1 L'apprentissage

La classification symbolique ou conceptuelle fait partie du cadre plus vaste qu'est l'apprentissage. Ce mot représente un domaine très large dont la définition, comme pour beaucoup de termes abstraits d'usage courant, est relativement imprécise et ambiguë. On en distingue de nombreuses variétés de telle sorte qu'il recouvre des concepts différents.

On peut réaliser une classification de l'apprentissage selon trois critères [Carbonell 84] :

- les stratégies sous-jacentes ;
- la représentation de la connaissance ;
- les domaines d'applications.

Ici, le domaine d'application est clair ; la représentation des connaissance est celle utilisée de manière classique en intelligence artificielle, nous allons examiner quels peuvent être les les stratégies d'apprentissage sous-jacentes.

La première stratégie – et aussi la plus primitive – est l'apprentissage par mémorisation directe[1]. Elle agit en retenant toutes les configurations qui se présentent pendant la phase d'apprentissage et les reproduit lors de la phase d'action. C'est une méthode longue et fastidieuse, qui ne pourrait s'appliquer à l'électrophorèse bidimensionnelle.

L'apprentissage par approximation fait appel aux techniques numériques et à l'imitation possible de mécanismes humains par la machine. L'apprentissage bayésien en est un exemple, bien qu'il ne s'agisse pas d'une méthode symbolique. On doit grâce à une série d'exemples ajuster les paramètres d'un modèle. Une autre forme de ces techniques est née des idées visionnaires du *perceptron* [Rosenblatt 58]. Cette tentative s'inscrit dans la lignée des copies plus ou moins conformes des réseaux nerveux des êtres vivants. En dépit de certains succès, son application semble réservée au domaine de la perception de concepts simples, tels que la lecture de caractères.

L'apprentissage à partir d'exemples est l'une des formes les plus fécondes de l'apprentissage. Elle correspond à une induction et une détermination de concepts qui permettent de créer les taxinomies [Winston 84, chapitre 11], [Michalski 84a]. À partir d'exemples, positifs et négatifs, d'éléments appartenant à une classe, on donne pour but aux algorithmes d'extraire les structures ou les caractéristiques qui rendrons possible la reconnaissance et la classification automatique. En la comparant à la première méthode, où on tâche de donner une description en *extension*, on remarque que le saut est considérable, puisqu'il s'agit ici de déterminer la définition en *compréhension*. Les idées fondamentales qui régissent ces algorithmes sont :

1. la généralisation et

1. En anglais *rote learning*

2. la spécialisation.

La généralisation se réfère à l'agrégation des données, c'est à dire à la détermination d'ensembles, puis à leur abstraction. La **spécialisation** intervient comme le mécanisme correcteur de la généralisation. Elle agit par la séparation de plus en plus fine des concepts, en se déplaçant, par exemple, vers les feuilles d'un arbre de classification ou en ajoutant des prédicats par une conjonction : $rouge(x) \wedge grand(x)$, jusqu'à l'instanciation ou l'individuation[2].

C'est cette classe de stratégies que nous avons implantée et adaptée et que nous détaillerons par la suite. Elles représentent un point de vue "réalisable" dans les conditions techniques actuelles et répondent en partie aux problèmes auxquels nous sommes confrontés.

Il existe enfin une dernière classe de stratégies, qui est l'apprentissage par découvertes [Langley 86] brillament illustrée par [Lenat 84]. Elle permet la définition de nouveaux théorèmes ou concepts. Cette vue est la plus ambitieuse mais pose de très nombreux problèmes d'implantations et elle nous a paru hors de portée quant à la validation par des manipulations biologiques des concepts trouvés.

2.2 La classification conceptuelle

L'interprétation d'expériences d'électrophorèse entre dans le champ de l'apprentissage par induction et plus précisément, pour l'objectif que nous nous sommes fixés, dans celui de l'apprentissage à partir d'exemples. Cette interprétation fait appel à des données numériques telles que celles extraites des gels, par notre système ou par d'autres, mais aussi à toute une connaissance sur la structure du vivant qui ne s'exprime que difficilement par des chiffres. La classification conceptuelle intervient comme un enrichissement des méthodes traditionnelles. [Fisher 86] – et plus généralement [Gale 86] – en donnent un certain nombre d'exemples.

Avant de définir les implantations des opérations d'apprentissage telles que la généralisation et la spécialisation, nous allons définir les modes de représentations spécifiques.

Il vient d'abord les domaines des variables manipulées. On en distingue en général trois types [Michalski 84a] :

- le domaine *nominal*, qui caractérise les ensembles sans ordre, ni structure, tels que les noms de personnes[3], les prédicats tels que la couleur, etc. Par exemple, les gels résultants d'une manipulation appartiennent au domaine nominal :

$$[\text{manipulation} = gel_1 \vee gel_2 \vee gel_3]$$

2. Ces mécanismes de généralisations et de spécialisations sont primordiaux dans le raisonnement humain, ainsi Polya les décrit comme les deux composantes de l'analogie. (Cité par Michel Sintzoff lors de l'anniversaire de Centre de recherche en informatique de Nancy)

3. en ignorant l'ordre alphabétique

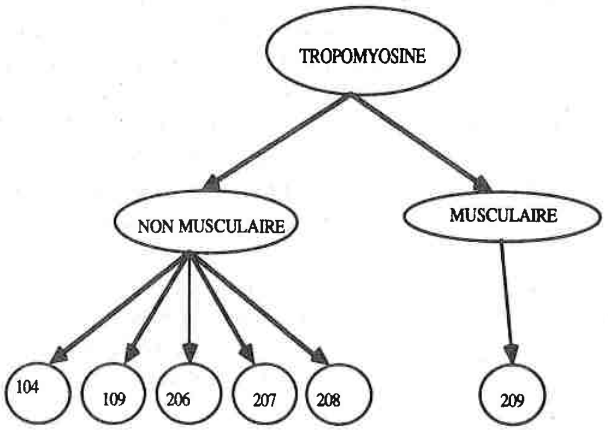


Figure 2.1. Un domaine structuré

- le domaine *linéaire* ou ordonné qui caractérise par exemple les nombres ou les échelles diverses ; l'intégrale densitométrique d'une protéine sur un gel appartient à un domaine ordonné. Supposons que cette quantité puisse varier de 0 à 100, on aura par exemple sur le gel_1 :

  [quantité albumine = 55]

- enfin le domaine *structuré* qui détermine les arbres de classification, comme la classification de Linné. Par exemple, les tropomyosines de la figure 2.1, page 106, sont des protéines qui possèdent des formes musculaires et non musculaires.

À ces domaines, nous ajoutons le produits cartésiens de domaines ordonnés afin de permettre une meilleure intégration des systèmes numériques. En effet, ceux-ci raisonnent sur des vecteurs et il est indispensable de pouvoir faire le produit de deux ou plusieurs domaines pour les appliquer.

Une fois les types définis, il nous reste à donner la description des classes. Ce langage de description est déjà une interprétation[4] du "monde" [Rendell 86]. Bien que nous ne propo-

4. *bias* dans la littérature anglo-saxonne

sions aucune solution pour nous en affranchir, nous nous devons d'en être conscient.

Nous avons essentiellement deux possibilités de présentation :

1. le caractérisation par prédicats logiques [Michalski 84b] qui peut s'exprimer par
   - une disjonction de termes ou
   - une conjonction ;
2. un arbre de décision [Quinlan 84].

La description par disjonction de termes permet d'exprimer les classes en extension directement à partir des exemples : si les individus e_1, e_2 et e_3, se traduisent par une conjonction d'attributs et sont membres de la classe C, sans autres connaissances, on ne peut qu'écrire $C = e_1 \vee e_2 \vee e_3$ en les réduisant éventuellement grâce aux règles de la logique booléenne par factorisation des termes communs. Ceci peut s'exprimer sous la forme d'un graphe **OU-ET**

La représentation des classes par conjonction de prédicats est la plus utilisée. C'est celle que nous avons adopté, elle correspond à une extension de la description des individus par attributs[5] et peut s'exprimer sous la forme d'un graphe **ET-OU**. On peut la rendre très élaborée et [Diday 87] propose une structure générale de données pour définir les objets et les règles à partir de ces conjonctions.

Par exemple, on pourra exprimer la classe des sujets atteints de la rougeole par :

$$[\hat{a}ge = 6..12] \wedge [température = 39..40] \wedge [marques sur le visage = boutons \vee pustules \vee croûtes]$$

Les deux premiers descripteurs sont de type ordonné, le dernier peut être nominal ou éventuellement structuré. (Figure 2.2, page 108.)

Pour passer de la description des individus à celle d'une classe, on doit actionner un mécanisme que nous détaillerons par la suite.

La description par arbre de décision est une méthode plus ancienne et dont l'étendue d'application dépasse le cadre de l'apprentissage tel que nous le définissons ici. Pour chaque objet, on définit un arbre comprenant autant de niveaux que d'attributs et à chaque niveau autant des branches que l'attribut peut prendre de valeurs. Les exemples viennent se situer au bas de l'arbre. La difficulté principale dans leur leur construction réside dans l'arrangement de l'ordre des attributs de manière à ce que les décisions soient les plus rapides possibles.

2.2.1 Le raisonnement par généralisation

La généralisation est le processus le plus important mis en jeu lors de l'apprentissage. Il permet d'extraire les concepts d'une série d'événements. C'est le mécanisme fondamental

5. ou des variables suivant la terminologie

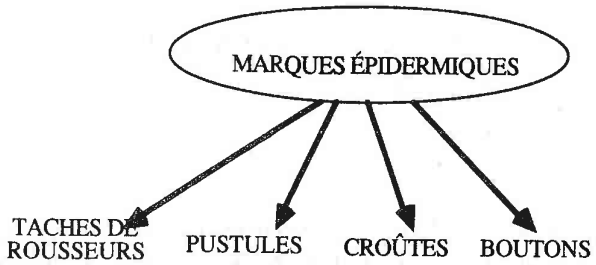


Figure 2.2. La rougeole

de l'induction. Il s'applique à un ensemble d'exemples ou à des formules permettant des descriptions et on dit qu'une formule S_1 est plus générale qu'une formule S_2 si l'ensemble de description de S_2 est inclus dans S_1.

À partir d'une ensemble d'exemples ou de formules, on peut pratiquer plusieurs types de généralisations qui dépendent à la fois du mode de représentation et des domaines des attributs. Ce sont ces règles que nous allons détailler.

La généralisation minimale

Cette généralisation se définit pour un ensemble d'objets ou d'ensembles dans le cadre d'une représentation des connaissances, c'est à dire ici avec une description des classes sous la forme d'une conjonction d'attributs.

Soit E, l'ensemble des individus ou des descriptions d'individus qu'on désire généraliser. La généralisation minimale de E se caractérise pour chaque attribut commun à tout les éléments de E par l'union de toutes les valeurs prises par ces attributs.

Par exemple, la généralisation minimale de :

$$[couleur = blanc \vee rouge \vee vert] \wedge [taille = petit] \wedge [forme = rond]$$

et

$$[couleur = blanc \vee noir] \wedge [forme = carré]$$

se traduit par :

$$[couleur = blanc \vee rouge \vee vert \vee noir] \wedge [forme = rond \vee carré]$$

L'antiunification et la transformation de constantes en variables

L'antiunification est un algorithme très puissant de généralisation [Pair 88] des exemples en formules de description. Elle définit de manière formelle la transformation des constantes en variables.

Soit S_1 et S_2, deux exemples de scènes, on peut antiunifier ces deux termes si il existe un appariement direct à un renommage près des constantes transformables en variables. Ceci s'exprime par l'existence d'une substitution biunivoque telle que :

$$substit(t, \sigma_1) = S_1$$

et

$$substit(t, \sigma_2) = S_2$$

où t est la variable et σ_1, σ_2 sont les constantes.

Cet algorithme peut s'appliquer dans les cas où il n'y a pas d'ambiguïté dans les descriptions de scènes, ce qui est rarement le cas dans le monde réel, car l'algorithme procède par comparaison directe des termes. De plus, il ne tient pas compte des exceptions qui peuvent intervenir, ni des règles de transformations qu'on peut attacher aux différents contextes.

[Kodratoff 86] reprend l'antiunification, en la rebaptisant : "appariement structurel", et propose une amélioration de cet algorithme de manière à pallier éventuellement la non uniformité des formules de description. Elle repose sur la liaison adéquate des constantes aux variables et sur la récriture des prédicats grâce à la connaissance sémantique du contexte et des théorèmes.

L'abandon de termes conjonctifs

Lorsqu'on choisit une description par conjonction de termes, cette généralisation est très facile. En effet, si une formule caractérisant un gel s'écrit :

$$(\bigwedge_N C_i) \wedge [albumine > 250]$$

la formule :

$$\bigwedge_N C_i$$

est plus générale, car aucune condition n'est posée sur la quantité d'albumine.

Cette méthode peut être d'utilisation assez dangereuse car elle simplifie de manière imprévisible l'ensemble d'exemples. On ne doit l'utiliser qu'en dernier ressort.

L'ajout de termes disjonctifs

Cette généralisation est facile dans les descriptions constituées de disjonction. Ainsi la formule :

$$\bigvee_N C_i$$

se généralise en :

$$(\bigvee_N C_i) \vee [albumine > 250]$$

Lorsque la description est donnée sous la forme d'une conjonction, on doit ajouter les éléments disjonctifs au niveau de la partie *ou*, c'est à dire à l'intérieur des termes de la conjonction, ainsi la formule :

$$(\bigwedge_N C_i) \wedge [albumine = 250..300]$$

peut se généraliser en :

$$(\bigwedge_N C_i) \wedge [albumine = 250..300 \vee 100..150]$$

Cette forme peut s'étendre de la manière suivante :

$$(\bigwedge_N C_i) \wedge [A = D_1]$$

se généralise en :

$$(\bigwedge_N C_i) \wedge [A = D_2]$$

lorsque $D_1 \in D_2$

La généralisation d'ensembles ordonnés

On opère la généralisation d'ensembles ordonnés par la fermeture d'intervalles. Ainsi, le terme :

$$[val = a \vee b]$$

se généralise en

$$[val = a..b]$$

où $a..b$ désigne l'intervalle $[a, b]$.

On fixe en général un seuil au-dessus duquel on ferme l'intervalle. On peut comparer cette méthode au lissage de segments[6] en traitement d'images [Wong 82]. La règle de transformation devient alors :

$$[val = a \vee b] \; et \; (b - a) < \epsilon$$

6. Run-Length Smoothing

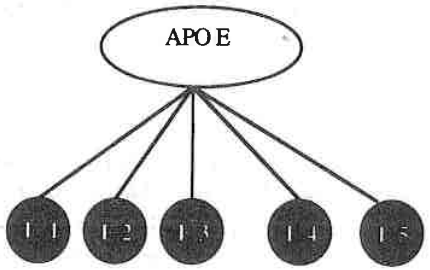


Figure 2.3. Les apolipoprotéines E

se généralise en

$$[val = a..b]$$

On obtient des généralisations du type :

$$[val = a \vee b \vee c \vee d]$$

en

$$[val = a \vee b..c \vee d]$$

La généralisation d'ensembles structurés

Cette généralisation tire parti des connaissances *a priori* du domaine qui constitue le champ d'investigation. Lorsqu'on peut hiérarchiser ce domaine en un arbre de concepts de plus en plus généraux, on procède à la généralisation par une remontée au concept qui inclut toutes les valeurs que peut prendre l'attribut. Ainsi, en considérant l'arbre de la figure 2.3, page 111.

On généralise le terme :

$$[protéines = apolipoprotéine E1 \vee apolipoprotéine E5]$$

en

$$[protéines = apolipoprotéine E]$$

2.2.2 Le raisonnement par spécialisation

La spécialisation intervient comme le mécanisme inverse de la généralisation dans les opérations d'apprentissage inductif. C'est un processus plus délicat à manier et auquel on fait moins référence dans la littérature. Nous examinons deux types de mise en œuvre.

Les arbres de décisions

La technique des arbres de décisions est plutôt delaissée dans le domaine de la classification conceptuelle au profit de méthodes plus en vogue. Elle offre pourtant un moyen simple et efficace de déterminer une partition entre exemples et contre-exemples [Quinlan 84, Quinlan 86]. On pourrait exprimer la description des classes sous la forme d'une disjonction de conjonctions, mais ceci n'est pas la forme de présentation habituelle.

Pour construire l'arbre, on place l'attribut le plus discriminant à la racine de l'arbre, qu'on choisit, suivant un critère d'entropie. Les individus de l'ensemble d'apprentissage se divisent suivant les valeurs de cet attribut et viennent se répartir aux nœuds du premier niveau.

On poursuit ce processus en ajoutant à chaque nœud dont les éléments ne sont pas tous de la même classe, un attribut choisi comme le plus discriminant parmi ceux qui restent. Pour un même niveau, les attributs peuvent être, bien sûr, différents.

L'algorithme rend les conditions de plus en plus spécifiques, – ramifie l'arbre ou le spécialise –, jusqu'à obtenir des feuilles qui soient toutes de classe homogène.

[Nordhausen 86] généralise ces arbres de décision à un nombre indéfini de classes, c'est à dire au-delà des exemples et contre exemples. Il ajoute la possibilité de recombiner les prédicats pour former de nouveaux attributs.

L'introduction de contre exemples

Lorsqu'on crée une description à partir d'un ensemble d'exemples et qu'on obtient une formule trop générale, on peut réduire sa portée en introduisant des contre exemples. C'est la démarche suivie par [Vere 80].

L'algorithme suit un processus de raffinement récursif de la description. À partir des exemples, il crée une formule par la transformation de constantes en variables et d'abandon de conditions. Celle-ci est rapide et souvent trop générale. Il considère alors les contre exemples et introduit les exceptions nécessaires et ainsi de suite jusqu'à la description complète.

Cette méthode, cependant, ne semble pas adaptée à la description de classes multiples.

2.2.3 L'introduction de probabilités

Jusqu'ici, nous nous sommes placés dans un cadre idéal de classification où on pouvait déterminer sans faute l'appartenance des individus par la valeur de certains attributs. Ceci n'est pratiquement jamais le cas dans la réalité. Il existe toujours des éléments atypiques qui ne remettront pas en cause la généralité des concepts mais qui pourront détruire beaucoup de constructions universitaires.

Les descriptions conceptuelles sont nécessairement inexactes car on ne peut prétendre cerner une réalité par une formule. On peut cependant mesurer leur pertinence d'approximation et donc d'application par l'utilisation de probabilités. Il nous semble inutile sinon impossible[7] de chercher à décrire logiquement les individus qui ne rentrerons pas dans les formules d'une classe, car ils se présentent plutôt comme un bruit par rapport à un phénomène déterministe. L'ambition réaliste que peut avoir la classification conceptuelle est de fournir une définition symbolique de classes et d'en évaluer statistiquement la vraisemblance.

Bien que ce sujet soit tout à fait essentiel il n'a pas joui d'une totale reconnaissance. Nous allons examiner deux cas de traitement statistique.

[Gascuel 88] fixe comme objectif à son système *PLAGE* de déterminer les différences entre les signaux peptidiques[8] des E. coli et ceux des êtres humains. Pour ceci, il se donne un certain nombre de descripteurs portant sur les propriétés des chaînes des caractères, tel que :

- le nombre d'occurences d'un caractère dans une chaîne ;
- la présence d'une sous-chaîne dans une chaîne ;
- le nombre de ces sous-chaînes...

et il les combine selon une grammaire munie de règles de production qui peuvent, par exemple, engendrer des sous-chaînes de plus en plus spécifiques.

L'évaluation d'un descripteur se fait grâce au critère du χ^2. On suppose que l'ensemble d'apprentissage comprend n classes. Pour un descripteur, on dresse la table de contingence des éléments bien et mal classés :

	C1	C2	...	Cn
vrai	v1	v2	...	vn
faux	f1	f2	...	fn

Soit k_{ij} les valeurs du tableau ci-dessus. Avec les notations :

$$k = \sum_i \sum_j k_{ij}; \quad f_{ij} = k_{ij}/k; \quad f_{i.} = \sum_j f_{ij}\ ; \quad f_{.j} = \sum_i f_{ij},$$

7. À l'opposé de Vere
8. Ces sont les chaines de départ des protéines. Elles sont bien sûr constituées d'acides aminés.

on calcule le χ^2 par la formule :

$$k \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$$

et on retient les descripteurs dont les probabilités de dépendances sont supérieures à 99%.

Une autre méthode tient de l'apprentissage incrémental [Fisher 87,Schlimmer 86]. Elle tâche d'évaluer chaque descripteur suivant une démarche bayésienne déterminant les probabilités *a priori* du type : $P(A_i = V_{ij} \mid C_k)$ ou $P(C_k \mid A_i = V_{ij})$, où $A_i = V_{ij}$ désigne la $j^{ième}$ paire attribut-valeur du $i^{ième}$ attribut et C_k, la $k^{ième}$ classe. On construit ensuite un arbre de classification suivant les meilleures "chances" d'appartenance. À chaque nœud de cet arbre, correspondant à un concept probabilisé, on ajoute la valeur $P(A_i = V_{ij} \mid C_k)$.

L'acquisition des connaissances est incrémentale. Dans le système *COBWEB* de D. Fisher, on peut :

1. classer un nouvel individu dans une catégorie existante. On cherchera alors la classe qui maximise une fonction d'utilité de type bayésien ;
2. créer une nouvelle classe en y plaçant l'objet, lorsque la fonction d'utilité est meilleure que lors de l'insertion dans une classe déjà existante ;
3. fusionner des classes ;
4. partager des classes.

On n'obtient plus une description avec des termes conjonctifs, mais simplement une série de descripteurs probabilisés. Nous considérons cette dernière technique plutôt comme une extension de la classification bayésienne que comme un introduction de probabilités dans la classification conceptuelle.

2.2.4 Comment créer des classes

Lorsqu'on considère un ensemble d'individus dont les classes ne sont pas connues ou alors de manière imparfaite, on doit pouvoir disposer d'un mécanisme de création de classes ou de correction de celles données en exemple.

Ce mécanisme dépend étroitement du mode de représentation choisi. Avec une construction incrémentale telle que celle de *COBWEB*, la création des classes est implicite.

Avec la description sous forme conjonctive, l'utilisation de l'algorithme des nuées dynamiques nous semble la plus appropriées [Diday 83].

[Michalski 80] l'implante dans son système *CLUSTER/2* et [Appel 88] l'utilise pour reclasser avec succès des gels de cellules de foies sains et atteints de cirrhose. Il opère par un

déplacement dynamique des formules de description des classes potentielles, en activant pour chacunes de ces formules un mécanisme de généralisation et de spécialisation.

Soit N le nombre d'individus à classer, on doit disposer au départ du nombre n de classes à former.

1. Au départ on choisit n éléments parmi les N individus, soit de manière totalement aléatoire, soit suivant une heuristique si on a une idée des classes à former ce qui est presque toujours le cas. Ces n individus vont former les noyaux des classes.

2. Pour donner une description logique des classes, on utilise les noyaux en les redéfinissant par leurs différences les uns par rapport aux autres[9].

   Les éléments étant décrits par une suite de variables x_k. Soit e_i, un noyau, dont les valeurs sont r_i^k, alors $G(e_j \mid e_i)$ correspondant à la description du noyau e_j par rapport à e_i sera la disjonction des termes logiques $[e_j^k \neq r_i^k]$ tel que $r_j^k \neq r_i^k$. Soit :

   $$G(e_j \mid e_i) = \bigvee_k [e_j^k \neq r_i^k]$$

   Ceci correspond à la fonction la *plus générale possible* pouvant décrire e_j en excluant e_i.

   La description d'un noyau par rapport à tous les autres correspond à la conjonction des termes précédents. Soit E l'ensemble des noyaux restants, il vient :

   $$G(e_j \mid E) = \bigwedge_{j=1, j \neq i}^{n} (\vee_k [e_j^k \neq r_i^k])$$

   On obtient ainsi une formule de généralité maximale pour chaque noyau.

3. On développe les termes et on les réduit de manière à obtenir une disjonction de conjonctions. On appelle *complexe*, chacune de ces conjonctions. La réduction se fait par subsomption selon les règles suivantes :

   (a) $a \vee 1 \rightarrow 1$

   (b) $a \wedge 1 \rightarrow a$

   (c) $a \vee (a \wedge b) \rightarrow a$

   (d) $a \wedge (a \vee b) \rightarrow a$

   Dans la pratique, on doit limiter le nombre des complexes de la disjonction pour des raisons de capacité de mémoire. Pour cela, on évalue chaque complexe grâce à un critère approprié et on conserve les meilleurs.

9. Ceci correspond à une remarque de bon sens, car dans une population on assimile souvent la classes aux signes distinctifs. Il est bien entendu que la population est formée d'individus relativement comparables

4. Les points précédents correspondent à des créations de formules de généralisation. On opère maintenant une spécialisation en resteignant ces formules à l'ensemble qu'elles recouvrent, Pour cela, on détermine en extension l'ensemble des éléments décrits par chaque complexe et on opère une *généralisation minimale* de cet ensemble pour le rendre compatible avec le modèle de représentation des données. Il en résulte une nouvelle disjonction de complexes, issue de chaque noyau, qu'on peut éventuellement généraliser dans une certaine mesure. Chacun d'eux peut représenter une description de la classe.

5. On réalise une partition de l'ensemble en considérant un complexe par disjonction et en disjoingnant ces complexes, c'est à dire en éliminant les termes communs à deux ou plusieurs descriptions, suivant une procédure spécifique.

6. On évalue la meilleure partition selon un critère dépendant de l'application.

7. À cette étape, la partition peut être satisfaisante, et dans ce cas on arrête. Si elle ne l'est pas, on considère un nouvel ensemble de noyaux choisis à l'intérieur des classes et on revient au point 2.

2.2.5 La complexité de la création de classes

On peut évaluer la complexité de l'algorithme *CLUSTER-2* pour les étapes de création et de réduction des formules. Nous reprenons les notations du paragraphe précédent :

- soit n le nombre de classes à former,
- N le nombre d'individus à classer et
- K le nombre d'attributs.

Soit e_1, ..., e_n les n noyaux, la description de e_j par rapport à e_i est donné par la disjonction :

$$G(e_j \mid e_i) = \bigvee_k [e_j^k \neq r_i^k]$$

Elle revient à effectuer K comparaisons et crée au pire K termes logiques. La description de e_j par rapport à tous les autres noyaux est donc de complexité $(n-1)K$.

La description de la classe déterminée par e_j est donnée par la conjonction :

$$G(e_j \mid E) = \bigwedge_{j=1, j\neq i}^{n} (\vee_k [e_j^k \neq r_i^k])$$

qui se simplifie à chaque itération de sa formation, c'est à dire à chaque conjonction par une formule $G(e_j \mid e_i)$.

À la première itération, la constitution de la formule résultante nécessite au pire K^2 multiplications, puis K^2 comparaisons pour fusionner les termes ayant les mêmes variables et enfin $\sum_1^{K^2-1}$ (soit $\frac{(K^2)(K^2-1)}{2}$) opérations de subsomption. Chaque subsomption demande 4 comparaisons au pire. La seconde itération demande au pire K^3 multiplications, K^3 comparaisons de fusion et $\sum_1^{K^3-1}$ subsomptions qui elles-mêmes requièrent 9 comparaisons au pire. Par récurrence, la dernière itérartion (étape n) demande K^n multiplications, puis K^n comparaisons et $\sum_1^{K^n-1}$ subsomptions avec n^2 comparaisons.

Les opérations les plus coûteuses sont les subsomptions et nous approximons la complexité de l'algorithme à ces seules opérations qui portent au pire sur $\sum_2^n \frac{(iK^i)^2}{2}$ opérations. La complexité est bornée par le terme :

$$n^3 K^{2n}$$

Ce calcul est sans grande signification en pratique car on limite, au cours du développement, le nombre des formules à une valeur fixée. Soit m cette valeur, le nombre d'opérations à l'itération i est alors de $K.m$ multiplications, de $K.m$ comparaisons et de $\sum_1^{K.m-1}$ subsomptions comprenant i^2 comparaisons. Ces dernières opérations déterminent la complexité qui devient :

$$O(n^2(K.m)^2)$$

La complexité est en fait augmentée par l'évaluation des formules qui permet de limiter leur nombre mais que nous ne pouvons déterminer dans le cas général. La constitution de la partition pour les n classes est donc de complexité minimale :

$$O(n^3(K.m)^2)$$

2.3 Ce qui reste en suspens

L'algorithme de R. Michalski, associé aux techniques de généralisation, est le noyau de la plupart des recherches sur la classification conceptuelle. La présentation que nous en avons donnée est cependant très théorique et on doit l'adapter aux cas réels.

Tel quel, il effectue une création d'un nombre insurmontable, pour le matériel actuel, de formules pour la description des classes et donc il est indispensable d'évaluer ces formules par des fonctions appropriées et selon une méthodologie qui s'écarte un peu de l'exposé que nous venons de faire.

D'autre part cet algorithme tient peu compte des bruits ou des exceptions – qui sont inévitables dans la plupart des cas – pouvant perturber son déroulement et ceci conduit notamment à sur-généraliser les descriptions des classes ce qui peut en donner une caractérisation insipide ou rentrant en intersection avec celle d'autres classes.

Ces faiblesses, et peut être d'autres, empêchent cet algorithme de trouver, pour l'instant,

des véritables applications industrielles. Dans le chapitre suivant, nous tentons de proposer des solutions qui puisse permettre son extension.

3
Une intégration des méthodes de classification

Dans ce chapitre, nous décrivons les modifications théoriques qu'on pourrait apporter au processus de généralisation pour lui faire prendre en compte la nature toujours "exceptionnelle" des cas réels.

Nous établissons notre filiation intellectuelle par les thèses développées par R. Michalski dont nous avons cité dans le chapitre précédent les articles fondamentaux. Nous nous situerons donc dans son cadre de description des données – par conjonction de descripteurs logiques – avec des domaines de définition précis sur lesquels on peut effectuer certaines opérations, ainsi que des règles de transformation qui traduisent la connaissance du sujet traité.

C'est aussi dans le cadre des algorithmes dynamiques, pouvant redessiner les frontières de concepts mal établis[1] que nous avons travaillé. Il y a au départ le concept extrêmement riche de E. Diday que nous n'avons pas réussi à remettre en cause.

Le formalisme que nous proposons concernent deux points particuliers de ces méthodes. Le premier porte sur la généralisation et le second, sur les méthodes d'évaluation des partitions dynamiques.

3.1 L'antiunification modérée

L'antiunification représente une des méthodes les plus puissantes et les mieux formalisées de généralisation. Il s'agit d'effectuer un appariement littéral entre les formules à généraliser et de substituer les termes différents entre ces formules par des variables existentielles. Nous supposons ici qu'il n'y a pas d'ambiguïté lors de l'appariement structurel.

L'antiunification est aussi une technique assez rigide, car elle n'admet pas de degré dans la variabilisation : dès qu'au cours des formules, deux termes sont différents, on opérera

1. Il est pratiquement impossible de trouver une partition pertinente totalement nouvelle, car on fait presque toujours face à un choix combinatoire immense, on ne peut que préciser ou améliorer ce qui existe déjà

la substitution, ce qui revient dans le cas d'une description par termes conjonctifs à un abandon du terme antiunifié. Nous proposons deux types d'améliorations qui permettront la "modération" de l'antiunification en établissant un palier entre la généralisation minimale d'un ensemble, qui vise à lui donner une forme de réprésentation en termes d'attributs et l'application brutale de la variabilisation.

3.1.1 L'antiunification avec contraintes

La généralisation respectant la représentation des données à généraliser remplace en fait les constantes des attributs par un ensemble de données. Cet ensemble peut varier d'un minimum, défini comme l'union de toute les valeurs prises par les exemples, il s'agit alors d'une *généralisation minimale*, à tout le domaine de définition de l'attribut en question dans le cas de l'antiunification.

Par exemple, considérons les trois exemples suivants, avec l'attribut polymorphisme, prenant ses valeurs dans le domaine D, avec $D = \{forme_1, forme_2, forme_3, forme_4\}$

1. $E_1 = [polymorphisme = forme_1]$
2. $E_2 = [polymorphisme = forme_2]$
3. $E_3 = [polymorphisme = forme_1]$
4. $E_4 = [polymorphisme = forme_1]$

La généralisation minimale rend : $[polymorphisme = forme_1 \vee forme_2]$

L'antiunification rend : $[polymorphisme = x]$, $x \in D$

Il est clair que si le domaine est de grande cardinalité par rapport à la cardinalité de l'ensemble décrit par une généralisation minimale, une antiunification est un gaspillage. Nous proposons donc de modérer l'antiunification en tâchant de faire intervenir des règles de généralisation moins strictes.

Ces règles dépendent des domaines sur lesquels sont définis les attributs. On ne peut les appliquer tel quel et on doit définir des contraintes au-delà desquelles, l'antiunification est obligatoire.

- Lorque le domaine de l'attribut est structuré, on opère la généralisation par remontée dans l'arbre des structures. La contrainte est le *degré de remontée* dans l'arbre de manière à ce que le nœud puisse contenir les valeurs des exemples. Si la formule ne peut contenir les exemples en remontant de ce degré maximal, on procède à l'antiunification. Ce degré peut être défini différemment suivant les branches de l'arbre, il suffit alors de les étiqueter.

- Lorsque le domaine de l'attribut est ordonné, on généralise par la fermeture des intervalles des valeurs des exemples. On définit la contrainte comme *la longeur des segments* nécessaire pour opérer la fermeture. Si il faut une longueur supérieure pour contenir les exemples, on effectue l'antiunification.

- Lorque le domaine de l'attribut est nominal, il n'existe pas d'autre généralisation que la réunion des valeurs des exemples. On définit alors la contrainte comme *la cardinalité maximale* au-delà de laquelle on opère à l'antiunification.

3.1.2 La prise en compte du bruit

Malgré les contraintes définies précédemment, les exceptions peuvent dégrader notablement les résultats de l'antiunification. Il existe, en effet, toujours des individus atypiques d'une classe ou encore les frontières de cette classe sont parfois mal définies ce qui entraîne des fautes dans la répartition des exemples. Dans le domaine ordonné par exemple, une population, à partir de l'instant où elle est suffisament importante, se distribue souvent suivant une courbe gaussienne. On peut pourtant en donner une description approximative en indiquant sa valeur moyenne et une plage centrée autour de cette moyenne qui réunit la plus grande part des individus[2]. Le "bruit" inhérent aux cas reéls et "mal" choisis est une des principales causes du défaut d'application des méthodes symboliques. Dans l'état des connaissances actuelles, on ne peut pas modéliser ce bruit autrement que par des critères statistiques[3]. On postule donc que les exemples se laissent décrire conceptuellement et qu'il existe des exceptions dont nous ne cherchons pas à extraire les traits pertinents mais que nous pouvons quantifier de différentes manières.

L'introduction des probabilités se fait par la fixation d'un nombre d'ensembles devant rassembler un certain pourcentage des exemples. Pour déterminer les ensembles, on utilise, par exemple, l'algorithme des nuées dynamiques et on fixe les tailles limites admissibles des plages dans lesquelles doivent se trouver le pourcentage minimal d'individus. On peut associer des tailles de plages différentes suivant les ensembles.

Ceci se traduit, pour le domaine ordonné par le nombre d'ensembles, les plages et le seuil d'individus.

Par exemple en considérant la figure 3.1, page 122, de la répartition des tailles dans la population, on constate que la forme est bimodale. On peut la décrire par les moyens de la méthode précédente en fixant les contraintes d'antiunification à deux segments – ou deux noyaux – de longueur fixée dans un domaine ordonné avec un pourcentage de 90%

Pour les domaines structurés, la découverte des noyaux prend une forme légèrement différente, puisque la généralisation se fait par montée en hiérarchie. En remontant du maximum

2. L'approximation statistique est dans ce cas plus précise, mais moins compréhensible.
3. À moins d'en savoir beaucoup plus sur les applications.

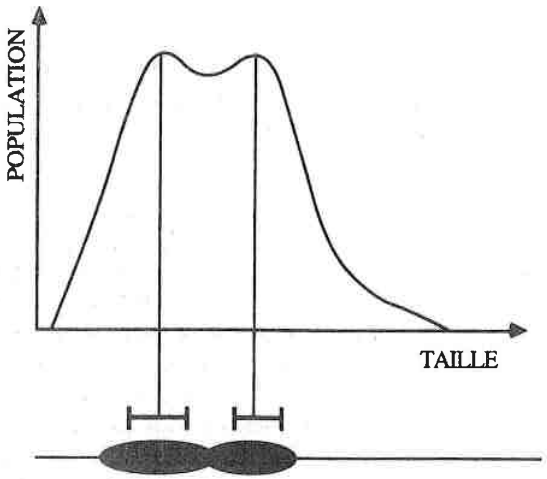


Figure 3.1. La répartition bimodale des tailles

autorisé dans la hiérarchie, on cherche donc les sous-ensembles contenant les plus grandes quantité d'individus.

Pour les domaines nominaux, la cas est encore plus simple, puisqu'il s'agit de déterminer par ordre d'importance, les noms qui rassemblent les pourcentages fixés à condition que leur nombre ne dépasse pas le maximum permis.

Tout ce que nous venons de décrire peut se généraliser au produit cartésien d'attributs de même domaines.

3.1.3 La combinaison de l'antiunification modérée et des statistiques

L'introduction des statistiques dans le mécanisme d'antiunification permet une description différente de celles que nous avons examinées jusqu'à présent, puisque elle ne décrit pas, à proprement parler, les traits d'individus, mais de populations. On cherche alors les valeurs dominantes de chacun des attributs. Elle pourrait fournir des descriptions telles que celles

que peuvent donner les médecins de populations sujettes à des risques cardiaques :

$$[sexe = masculin] \; et \; [taille = grande] \; et \; [poids = important] \; et \; [fumeur = 10..50]$$

Les maladies cardio-vasculaires sont une cause de mortalité très importante et pourtant il est bien difficile de trouver beaucoup d'individus aussi typiques ?

Ici l'algorithme produit une description de "bon sens" plutôt qu'un système de calibrage.

3.2 Les fonctions d'évaluation

Les fonctions d'évaluation interviennent de deux manières différentes dans la classification conceptuelle :

1. d'une manière théorique, pour déterminer la cohérence ou la compatibilité des ensembles construits. En effet, les classes potentielles se présentent comme des disjonctions de descriptions. Nous faisons donc face à un choix combinatoire qu'il faut effectuer en optimisant une formule et éventuellement en incluant une recherche. Deux points sont à évaluer – qui ne sont pas complétement indépendants –,
   - d'une part la qualité des classes prises l'une après l'autre ;
   - et d'autre part la qualité de la partition globale.
2. d'une manière plus pratique, pour limiter l'importance des calculs. La création des descriptions des classes potentielles est extrêmement coûteuse, notamment lors de la transformation du graphe ET/OU en graphe OU/ET et de sa simplification. En effet, si on considère des individus ayant k attributs et qu'on désire former n classes, la formule aura une taille de l'ordre de k^{n-1}, ce qui donne avec 50 attributs et 10 classes – hypothèses d'un cas réel simple –, 10^{17} descripteurs élémentaires. Il faut donc limiter dynamiquement – au cours de la création – le nombre de formules élémentaires.

En fait, c'est l'implantation qui prend le pas sur le reste et on effectue l'évaluation au cours du calcul pour éliminer au plus vite les complexes inutiles. On perd ainsi toute espérance d'optimalité de classification.

Pour limiter le choix des partitions, R. Michalski [Michalski 84b] propose de borner $G(e_i \mid E)$ à m disjonctions et évalue dynamiquement la partition qu'entraîne $G(e_{i+1} \mid E)$ en fonction du contexte précédent. On détermine ce contexte en prenant un complexe parmi les m retenus de $G(e_1 \mid E)$, puis de $G(e_2 \mid E)$, ...jusqu'à $G(e_i \mid E)$. On obtient ainsi l'arbre de recherche de la figure 3.2 : pour chaque noyau e_i, on crée et on ordonne par le mérite les m complexes différemment à chaque nœud. page 124. R. Michalski inclut donc dans le même processus, l'évaluation de la partition et des parties.

Nous proposons deux améliorations à cet algorithme :

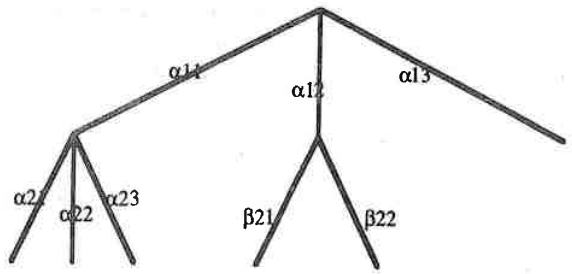


Figure 3.2. Partition dynamique

1. Dans son algorithme, R. Michalski ne distingue pas les fonctions d'évaluation qui s'appliquent à la limitation des disjonctions de $G(e_i \mid E)$ en cours de développement et les fonctions d'évaluation de la partition en construction. Il utilise simplement une "évaluation contextuelle" des complexes dans tous les cas. Ces fonctions sont pourtant de nature très différentes et on ne peut pas les appliquer de la même manière. De plus le premières posent un important problème de coût car à chaque nœud, on est amené à développer une nouvelle disjonction de complexes.

   Nous jugeons nécessaire d'établir de manière distincte les deux types de fonctions d'évaluation. Le premier, lié au développement des formules doit se faire de manière indépendante du contexte, ce qui :

   - limite les calculs, car pour chaque noyau et pour un ordre donné de développement de $G(e_i \mid E)$, il n'existe qu'une seule disjonction de m complexes ;
   - réduit l'influence de l'ordre des valeurs de i, sur le traitement de $G(e_i \mid E)$.

   Le second, pour l'évaluation globale de la partition, peut être de nature contextuelle. Il est alors sensible à l'ordre des classes et il revient à une recherche en faisceau sur les meilleurs complexes de chaque disjonction. Dans ce cas il réduit la quantité de traitement.

2. La limitation du nombre de disjonctions, lors des transformations ET/OU en OU/ET à chaque étape du développement, est très délicate du point de vue théorique, car on doit évaluer des formules en cours de construction. Nous avons choisi une limitation variable du nombre de formules suivant le stade de construction de $G(e_i \mid E)$. En effet, à la première étape, après les simplifications, les complexes conjonctifs sont de taille deux au maximum, de taille trois à la deuxième étape et ainsi de suite. Ils sont donc beaucoup moins spécifiques au départ et plus difficiles à évaluer. À chaque étape l de la construction de $G(e_i \mid e_j)$, on définit le nombre r_l de complexes à garder. Cette suite sera décroissante jusqu'à m.

L'évaluation proprement dite peut prendre des formes diverses suivant la nature de la connaissance *a priori* du problème.

1. On doit profiter autant que possible des connaissances sur le lot de données à traiter. C'est à notre sens la principale des évaluations possibles. L'algorithme cité suppose la connaissance du nombre de classes. C'est un pur cas d'école que de considérer acquise la connaissance de ce nombre sans aucune idée des classes à former. Aussi bien lors de la construction des complexes, que dans le choix de la partition et dans l'application des nuées dynamiques, on doit avoir le plus large recouvrement des "éléments sûrs", établis au départ par un expert, par les formules logiques et éliminer celles qui échouent à caractériser les classes dont on peut avoir une idée. On peut aussi privilégier un descripteur particulier et forcer la conservation des complexes qui le contiennent.

   [Stepp 86] définit les critères de classification par un but et des sous-buts à atteindre, par exemple en établissant un réseau de dépendance où *survivre* serait le but final qui entraînerait les actions *manger* et *boire* et favoriserait les propriétés telles que *comestibles* ou *potable*. C'est une vision séduisante et semblable dans son esprit à celle que nous avons exposée mais qui est difficilement réalisable dans toute sa généralité à partir de l'instant où on crée de très nombreuses formules logiques.

2. Si on ne dispose pas de toute cette connaissance *a priori* ou qu'on ne veut pas en faire usage, on peut évaluer la qualité d'une partition en fonction de la non-intersection des données filtrées par les descriptions des classes. Ce second type d'évaluation peut se définir de manière stricte, en pénalisant les intersections selon le nombre des éléments qu'elles recouvrent ou de manière floue, en privilégiant les descriptions de classes contenant le plus grand nombre d'éléments se distinguant des autres classes. On devra alors définir un indice de dissimilarité pour les attributs symboliques. C'est la voie choisie par [Appel 87, page 183 du document] dont le meilleur critère est de : maximiser le nombre de taches protéiques dont les valeurs qu'elles peuvent prendre dans une classe sont nettement différentes de celles qu'elles peuvent prendre dans les autres classes.

3. Enfin un troisième type est de considérer uniquement les attributs numériques ou ordonnés pour lesquels il existe déjà dans la littérature un très grand nombre de distances et de critères. En effet, on peut supposer que dans beaucoup d'exemples, le comportement de ces attributs n'est pas totalement indépendant de celui des autres. Cette méthode peut s'appliquer à l'évaluation des complexes d'une disjonction, en effectuant sur eux un test de cohésion au sens d'une distance numérique, ou bien d'une partition, en maximisant les distances entre les classes.

4
Application : l'analyse de la croissance des cellules musculaires

Nous présentons une application des techniques d'apprentissage symbolique concernant des paramètres numériques et permettant leur validation. Grâce à la méthode que nous exposons, nous identifions des protéines qui peuvent jouer un rôle dans les transitions entre les différentes étapes de la maturation musculaire.

Cette étude s'est faite en collaboration avec Robert Whalen du laboratoire de biochimie de l'Institut Pasteur. Nous traitons ici des données issues de gels dont le dépouillement, ainsi que l'appariement, n'a pas été réalisé par notre système.

4.1 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

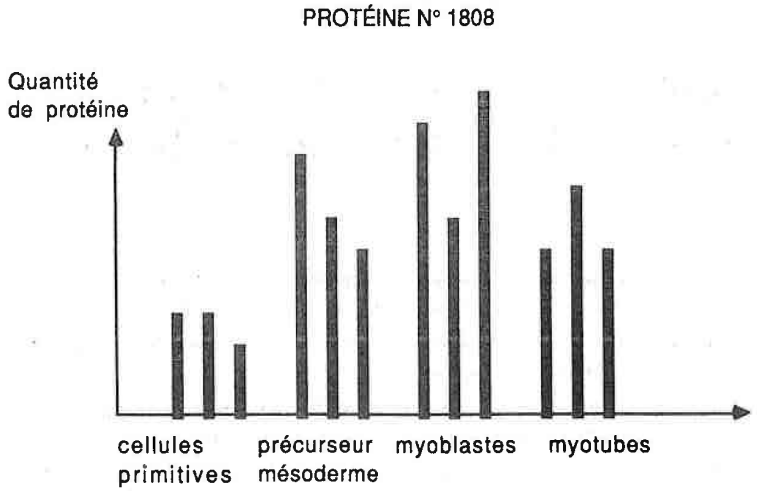


Figure 4.1. Courbe de variation de la quantité de protéine suivant les différents états de la cellule.

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 un film. Le système a extrait les caractéristiques de chacun des points par une approximation par les moindres carrés et apparié les douze gels entre eux.

L'appariement est effectué par l'expert biologiste en reliant manuellement des protéines remarquables d'aspect géométrique identique sur les douze gels, puis en ajoutant automatiquement et itérativement les protéines nouvellement appariées à l'ensemble d'appariement jusqu'à ce qu'il n'en reste plus. La procédure automatique consiste en une transformation des régions des protéines précédemment appariées par l'intermédiaire d'une fonction quadratique. On détermine les nouvelles coordonnées pour les protéines considérées et on compare les gels deux à deux (méthode de "stretching" [Garrels 84]).

Enfin nous représentons chaque protéine par l'intégrale de sa quantité de matière sur les différents gels, ce que traduit par exemple la figure 4.1 page 128

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 divisé par l'écart-type de ces trois instant, soit $CV = \frac{x - m}{\sigma}$. Nous avons conservé au total 208 protéines pour notre analyse.

4.2 L'implantation de l'algorithme

4.2.1 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 classifications conceptuelle définies par Michalski [Michalski 84b], afin de pouvoir offrir une description logique conjonctive de chaque classe et identifier les éléments ou la relation conjonctive entre les éléments, qui puisse 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'électrophorèses. 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 correpondant à chaque protéines prennent la forme de 12 nombres caractérisant les quatre instants répétés trois fois[1]. Notre algoritme 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

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.

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 fonctionalité.

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.

4.2.2 Quelques particularités

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

- 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.

- 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 $[x_k \neq r_2^k]$ tel que $r_2^k \notin [r_1^k - m_d; r_1^k + m_d]$.

  De même, on considèrera 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 : $\bigwedge_i G(e \mid e_i)$.

4.2.3 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. Nons 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 :

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

   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 agglomèraient 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 87] :

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

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

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 :

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

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.

4.3 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 génèse. 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 auquels 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 sélectifs 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[2] (cellules embryonnaires) :

Longueur 1

$$((prot8001 \neq 0)) \vee ((prot7612 \neq 0)) \vee ((prot7004 \neq 0)) \vee ((prot8108 \neq 0))$$

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)) ∨
((prot7409 ≠ 0) (prot3612 ≠ 2)) ∨ ((prot7409 ≠ 0) (prot3506 ≠ 10)) ∨
((prot7409 ≠ 0) (prot3310 ≠ 10)) ∨ ((prot3301 ≠ 0) (prot7409 ≠ 0)) ∨
((prot7409 ≠ 0) (prot2611 ≠ 10)) ∨ ((prot7409 ≠ 0) (prot2610 ≠ 9)) ∨
((prot7409 ≠ 0) (prot2608 ≠ 10)) ∨ ((prot7409 ≠ 0) (prot2110 ≠ 0)) ∨
((prot2104 ≠ 10) (prot7409 ≠ 0)) ∨ ((prot7409 ≠ 0) (prot1510 ≠ 10)) ∨
((prot801 ≠ 10) (prot7409 ≠ 0)) ∨ ((prot501 ≠ 10) (prot7409 ≠ 0)) ∨
((prot306 ≠ 7) (prot7409 ≠ 0)) ∨ ((prot209 ≠ 10) (prot7409 ≠ 0)) ∨
((prot303 ≠ 10) (prot306

2. 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

Au total 77 formules conjonctives.

Les termes de longueur 3 sont similaires aux précédents.

4.4 Les améliorations possibles

Jusqu'ici nous n'avons pas tiré le parti maximal des techniques symboliques. Néanmoins, nous avons pu obtenir des résultats pertinents avec un ensemble d'entrée minimal en "connaissance ajoutée".

Cette connaissance minimale *a priori*, empêche notamment le développement important d'une étape de généralisation. Une des voies d'évolution est 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. On cherchera à mettre en évidence les couplages possibles entre différentes protéines.

5

Application : le dépouillement d'une enquête épidémiologique

Nous décrivons cette application bien qu'elle ne soit à présent qu'à l'état d'étude. Elle fait partie du projet ambitieux d'examen de 400 familles pour lequel nous avons d'abord conçu un identificateur automatique d'apolipoprotéines. Nous détaillons ici les données à analyser afin qu'un autre chercheur – nous l'espérons – ait plus de facilités pour l'implanter.

5.1 La problématique des apolipoprotéines

L'objectif du dépouillement de cette enquête est de relier le polymorphisme des apolipoprotéines et quelques traits déterminants physiologiques, aux risques cardio-vasculaires. L'influence de ce polymorphisme a été démontrée, notamment en ce qui concerne les apoE [Galteau 88].

Cette étude implique un certains nombre de relations de structure entre les différentes familles d'apolipoprotéines et d'autres données qu'on aurait peut être pu traiter par une méthode classique mais qu'il semble bien plus intéressant de considérer avec des méthodes symboliques.

Les données extraites des gels concernent d'une part la détermination du polymorphisme des apolipoprotéines A-I, A-II, C-2, C-3 et E et d'autre part la semi-quantification des taches grâce à des protéines de référence.

5.2 Les autres données et leur structure

En dehors de apolipoprotéines, le Centre de médecine préventive a répertorié environ 500 autres protéines dont l'investigation pourrait se révéler intéressante. Du fait que notre système de lecture ne les détecte pas, elles ne seront pas prises en compte dans le système de classification.

Les patients examinés doivent repondre à un questionnaire sur leur conditions de vie et de santé.

Les conditions de vie portent sur les conditions sociales et professionnelles, l'environnement familial, les loisirs, les quantités d'absorption d'alcool et de tabac, les nuisances chimiques toxiques... Ce sont surtout ces trois derniers points qui sont à prendre en compte.

Les conditions de santé sont rassemblées dans un questionnaire et portent sur un domaine très large. En fait, seul un nombre limité de questions est intéressant à considérer, au moins dans une première analyse. Ce sont notamment les questions sur le système circulatoire :

- hypertension artérielle,
- maladie cardiaque éventuelle et son type,
- appréciation de l'état cardiaque général.

On interrogera les patients sur les médicaments qu'ils peuvent prendre. L'effet de tous les médicaments présente un interêt et parmi eux on en définit une liste liée au métabolisme des apolipoprotéines : le lipathyl, le lurselle, le mediator, le ticlid, le atherolip, le befiziol, le lipavlon, le lipur et le questran.

Le bilan biologique complémentaire portera sur le dosage de :

- l'urée et la créatinine,
- le glucose,
- le cholestérol et les triglicérides,
- le taux d'albumine et de préalbumine,
- le calcium et les phosphates,
- des enzymes hépatiques : bilirubine, T.G.P., T.G.O. et G.G.T.

5.3 Le jugement *a priori*

L'étude des patients de l'enquête se fonde sur un pré-classement établi par l'expert entre les individus à risques et ceux sans risques.

On raffinera cette distinction en désignant parmi les individus jugés à risques, ceux dont le risque est avéré et ceux dont le risque est incertain. On procédera de même avec les individus jugés sans risques. Ceci de manière à pouvoir construire une fonction de mérite qui évaluera les capacités de recouvrement en pondérant différemment les individus suivant la gravité du risque. (Figure 5.1, page 137.)

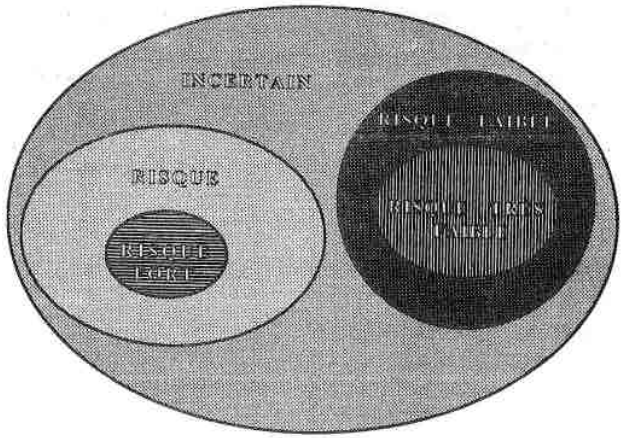


Figure 5.1. Le jugement *a priori* du risque

Dans la plupart des cas, le médecin peut fonder son jugement sur certains critères qui sont fonctions de conditions de vie ou de résultats d'analyses biologiques. Cependant, du simple fait des lacunes de la science humaine, il reste des facteurs de risque qui lui sont inconnus. D'autre part ce jugement ne peut être d'une sûreté absolue car il porte sur des concepts flous.

L'apport des méthodes de classification conceptuelle sera :

1. de retrouver les attributs et leurs valeurs sur lesquels se fonde le médecin pour porter son jugement et de faire intervenir tous les autres attributs de manière automatique afin de pouvoir les lier aux risques ;
2. de redessiner les frontières du risque de manière plus précise en tenant compte de tous ces facteurs.

5.4 Les algorithmes à mettre en œuvre

La disponibilité d'un classement préalable permet d'utiliser une méthode de généralisation. Ce classement résultant d'une connaissance floue, on ne peut pas utiliser directement

des procédures traditionnelles car elles entraîneraient trop d'imprécision. Cependant, l'anti-unification modérée que nous avons définie dans un chapitre précédent peut s'accommoder de flou et de bruit, et elle produit, comme résultat, une caractérisation des groupes en termes d'attributs et de valeurs majoritaires sans pour autant cerner des individus particuliers. Ceci permet, dans le cadre du projet que nous avons décrit, de fournir les caractéritiques dominantes des différents groupes de risques.

On pourra de plus introduire cette méthode au sein d'un algorithme dynamique qui redessinera les contours des groupes de risques définis au départ par l'expert. À partir de cette classification *a priori*, on déplacera les frontières des groupes en choisissant comme nouveaux groupes à risques l'ensemble des éléments qui seront les mieux caractérisés par les formules de description produites à chaque étape de l'itération.

On pourra ainsi déterminer, nous l'espérons, une nouvelle série de définitions plus précises des risques ou de l'absence de risques cardio-vasculaires.

6
Conclusion

Le travail que nous avons fourni sur l'interprétation des images d'électrophorèses bidimensionnelles a consisté, dans un premier temps à réaliser un système capable d'identifier automatiquement des protéines particulières, isolées ou à l'intérieur d'une constellation de polypeptides, sur des images de gels, en reproduisant les méthodes de localisation des experts, médecins ou biologistes.

Ce système possède une architecture modulaire, inspirée de celle des *blackboards*. Parmi les procédures de traitement d'image qui conduisent à l'extraction des paramètres, nous avons détaillé les techniques existantes et choisi celles qui suffisaient et convenaient le mieux pour atteindre nos objectifs. Nous avons utilisé un processus mixte de raisonnement : ascendant-descendant qui fait d'abord correspondre les paramètres extraits des images aux modèles géométriques potentiels des constellations de protéines, puis qui revient sur l'image pour déterminer les éléments éventuellement manquants. Nous avons fait précéder ce processus d'une focalisation de l'attention afin de réduire fortement la recherche des éléments pertinents. La partie ascendante est fondée sur un appariement des caractériques des taches avec des contraintes, à la fois sur les taches elles-mêmes et sur la disposition géométrique de l'environnement. Cet appariement peut se faire de manière incomplète pour tenir compte d'imperfections sur les images. La cohérence des instances trouvées est vérifiée et elle est complétée au besoin par la partie descendante qui active les procédures de traitement d'image avec des seuils de détection plus bas.

Nous avons validé cette architecture sur la localisation des apolipoprotéines du plasma et la détermination de leur phénotype.

Pour permettre le dépouillement des résultats nombreux provenant de gels d'électrophorèse, nous avons étudié et implanté des méthodes d'analyse de données numériques et symboliques. Nous avons particulièrement considéré les procédures symboliques en indiquant leurs avantages, notamment en ce qui concerne leur puissance de représentation, et leurs limites, qui tiennent à leurs exigences en puissance de calcul.

Nous avons ensuite proposé deux types d'améliorations, concernant la généralisation et

l'évaluation, pour rendre les techniques symboliques applicables à des cas réels

Les améliorations portant sur l'évaluation, nous ont permis d'adaptater l'algorithme *CLUSTER-2* de R. Michalski à la classification de gels caractéristiques d'étapes de la croissance des cellules musculaires. Notre implantation a pu retrouver les bonnes classes et pour chacune d'elles proposer des descriptions sous la forme de conjonctions d'attributs.

Nous avons enfin décrit une méthode pour dépouiller une enquête épidémiologique qui s'appuie sur estimation humaine du risque cardio-vasculaire des patients en vue de préciser les liens entre ces risques et les familles d'apoliprotéines du plasma.

Deux des intérêts de notre travail résident, nous semle-t-il dans :

- l'application originale de techniques d'intelligence artificielle au domaine de l'électrophorèse bidimensionnelle ;
- la mise en œuvre des méthodes mises au point sur des applications réelles et non de simples cas d'école.

Néanmoins, les méthodes que nous avons proposées devraient être améliorées et complétées pour leur conférer une plus grande généralité. En particulier :

- Ces méthodes pourraient tirer davantage parti des techniques de traitement d'image qu'utilisent les analyseurs actuels par :
  1. l'intégration de prétraitements supplémentaires au sein d'une architecture qui permette de les activer d'une manière contextuelle.
  2. l'introduction d'une méthode d'appariement iconique pour seconder les processus d'identification et traiter les protéines inconnues qui ne ne peuvent l'être par nos méthodes car elles mettent en défaut l'expertise humaine.
- Il serait envisageable de renforcer l'aspect de système à base de connaissances par :
  1. une amélioration des méthodes de raisonnement en introduisant le raisonnement hypothétique.
  2. une mise à jour de l'architecture qui autorise des interrogations sur des grandes bases de données d'images de gels.
- Pour l'analyse des données, on pourra effectuer :
  1. une validation des méthodes de généralisation que nous avons élaborée (antiunification modérée) ;
  2. une association des méthodes symboliques à des méthode traditionnelles de classification et d'analyse des données.

Liste des Figures

Bibliographie

[Allen 84] R. C. Allen, C. A. Saravis, et H. R. Maurer. *Gel Electrophoresis and Isoelectric Focusing of Proteins.* Walter de Gruyter, 1984.

[Anderson 77] N. L. Anderson et N. G. Anderson. High resolution two-dimensional electrophoresis of human plasma proteins. *Proc Natl. Acad. Sci.*, 74, 1977.

[Anderson 82] Norman Anderson et Leigh Anderson. The human protein index. *Clinical Chemistry*, 28(4):739–748, 1982.

[Anderson 84] N. Leigh Anderson, Russell P. Tracy, et Norman G. Anderson. High-resolution two-dimensional electrophoretic mapping of plasma proteins. F.W. Putman, éditeur, *The Plasma Proteins, Vol. IV*, Academic Press, 1984.

[Anderson 88] Leigh Anderson. New architectures appropriate for large 2-d gel databases. *Electrophoresis '88*, pages 313–321, 1988.

[Appel 87] Ron David Appel. *Melanie. Un système d'analyse et d'interprétation automatique d'image de gels d'électrophorèse bidimentionnelle. Systèmes-experts et apprentissage automatique.* Thèse de Doctorat, Université de Genève. Faculté des sciences. Département d'informatique., 1987.

[Appel 88] Ron Appel, Denis Hochstrasser, Christian Roch, Mattieu Funk, Alex F. Muller, et Christian Pellegrini. Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis*, 9(3):136–142, March 1988.

[Ayache 86] Nicholas Ayache et Olivier Faugeras. Hyper: a new approach for the recognition and positioning of 2d objects. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 8(1):44–54, 1986.

[Ballard 82] Dana H. Ballard et Christopher M. Brown. *Computer Vision.* Prentice-Hall, 1982.

[Bellanger 84] M. Bellanger. *Traitement numérique du signal.* Masson, 1984.

[Benzecri 76] J.-P. Benzecri. *L'analyse des données,* tome 2 : *l'analyse des correspondances.* Dunod, 1976.

[Carbonell 84] J. Carbonell, R. Michalski, et T. Mitchell. An overview of machine learning. Ryszard S. Michalski, Jaime G. Carbonnell, et Tom M. Mitchell, éditeurs, *Machine Learning: An Artificial Intelligence Approach*, pages 3–23, Springer Verlag, 1984.

[Charniak 85] Eugene Charniak et Drew Mc Dermott. *Introduction to Artificial Intelligence.* Addison-Wesley, 1985.

[Codd 70] E.F. Codd. A relational model of data for large data banks. *Communications of the ACM*, 1970.

[Cog] *Iroise@CEWB, manuel de référence.* Cognitech.

[Cointe 86] Pierre Cointe. Une introduction á la programmation par objets. *Giens 86: deuxièmes journées bases de données avancées*, 1986.

[deKleer 86] J. deKleer. An assumption-based tms. *Artificial Intelligence*, 28(1), 1986.

[Diday 83] E. Diday, J. Lemaire, J. Pouget, et F. Testu. *Élements d'analyse des données.* Dunod, 1983.

[Diday 87] E. Diday. Introduction à l'approche symbolique en analyse des données. *JOURNÉES SYMBOLIQUE-NUMÉRIQUE*, Université de Paris-Dauphine, 1987.

[Elstein 78] A.S. Elstein, L.S. Shulman, et S.A. Sprafka. *Medical Problem Solving: An Analysis of Clinical Reasoning.* Harvard University Press, 1978.

[Endler 86] A.T. Endler et D. S. Young. Two-dimensional electrophoresis of proteins in tumors of the lung. *Journal of Clinical Chemitry and Clinical Biochemistry*, 24(12):981–992, 1986.

[Engelmore 88a] Robert Engelmore et Tony Morgan, éditeurs. *Blackboard Systems.* Addison-Wesley, 1988.

[Engelmore 88b] R.S. Engelmore, A.J. Morgan, et H.P. Nii. Introduction. *Blackboard Systems*, pages 1–22, Addison-Wesley, 1988.

[Fieschi 84] M. Fieschi. *Intelligence artificielle en médecine.* Masson, 1984.

[Fisher 86] Douglas Fisher et Pat Langley. Conceptual clustering and its relation to numerical taxonomy. William A. Gale, éditeur, *Artificial Intelligence & Statistics*, pages 77–116, Addison-Wesley, 1986.

[Fisher 87] D.H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[Funk 87] Matthieu Funk. *Melanie. Un système d'analyse et d'interprétation automatique d'image de gels d'électrophorèse bidimentionnelle. Traitement de l'image et système-expert.* Thèse de Doctorat, Université de Genève. Faculté des sciences. Département d'informatique., 1987.

[Gale 86] William A. Gale, éditeur. *Artificial Intelligence & Statistics.* Addison-Wesley, 1986.

[Galteau 88] M.-M. Galteau, S. Visvikis, J. Steinmetz, et G. Siest. Relations between apoliporoteins structure, hyperlipidemia and response to hypolipidemic treatments. T. Endler et S. Hanash, éditeurs, *Two Dimensional Electrophoresis*, VCH, Weinheim, Allemagne, 1988.

[Garrels 84] James I. Garrels, John T. Farrar, et Carter B. Burwell IV. The QUEST system for computer-analyzed two-dimensional electrophoresis of proteins. J. E. Celis et R. Bravo, éditeurs, *Two-Dimensional Gel Electrophoresis of Proteins*, chapitre 2, pages 37–91, Academic Press, 1984.

[Gascuel 88] Olivier Gascuel et Antoine Danchin. Data analysis using a learning program, a case study: an application of plage to a biological sequence analysis. *Proceedings of the European Conference on Artificial Intelligence*, pages 390–395, August 1988.

[Gong 88a] Yifan Gong. *Contribution à l'interprétation automatique des signaux en présence d'incertitude.* Thèse de Doctorat, Université de Nancy I, 1988.

[Gong 88b] Yifan Gong et Jean-Paul Haton. A specialist society for continuous speech understanding. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1988*, New York City, April 1988.

[Hahn 84] Saúl Hahn et Eugenio E. Mendoza. Simple enhancement techniques in digital image processing. *Computer Vision, Graphics, and Image Processing*, 26:233–241, 1984.

[Hames 81] B. D. Hames et D. Rickwood, éditeurs. *Gel Electrophoresis of Proteins: a practical approach.* IRL Press, 1981.

[Hanash 87] Samir M. Hanash, John R. Strahler, Luke Somerlot, Wilhelm Postel, et Angelika Görg. Two-dimensional electrophoresis with immobilized pH gradients in the first dimension: Protein focusing as a function of time. *Electrophoresis*, 8(5):229–234, May 1987.

[Jambu 78] M. Jambu et M.O. Lebeaux. *Classification automatique pour l'analyse des doneées.* Dunod, 1978.

[Knuth 87] Donald E. Knuth. Digital halftones by dot diffusion. *ACM Transactions on Graphics*, 6(4):245–273, 1987.

[Kodratoff 86] Y. Kodratoff et J.-G. Ganascia. Improving the generalization step in learning. Ryszard S. Michalski, Jaime G. Carbonnell, et Tom M. Mitchell, éditeurs, *Machine Learning: An Artificial Intelligence Approach, Volume II*, pages 215–243, Morgan Kaufmann, 1986.

[Kuick 87] Rork D. Kuick, Samir M. Hanash, Ernest H. Y. Chu, et John R. Strahler. A comparison of some adjustment techniques for use with quantitative spot data from two-dimensional gels. *Electrophoresis*, 8(4):199–204, April 1987.

[Laasri 88] Hassan Laasri, Brigitte Maître, Thierry Mondot, François Charpillet, et Jean-Paul Haton. *Atome: A blackboard architecture with temporal and hypothetical reasoning.* Rapport no. 855, INRIA, juin 1988.

[Landegren 88] U. Landegren, R. Kaiser, C.T. Caskey, et L. Hood. Dna diagnostics–molecular techniques and automation. *Science*, 242:229–237, 1988.

[Langley 86] P. Langley et R.S. Michalski. Editorial: machine learning and discovery. *Machine Learning*, 363–366, 1986.

[Lebart 82] Ludovic Lebart, Alain Morineau, et Jean-Pierre Fénelon. *Traitement des données statistiques.* Dunod, Paris, 2ème édition, 1982.

[Lee 87] D. Lee, T. Pavlidis, et G.W. Wasilkowski. A note on the trade-off between sampling and quantization in signal processing. *Journal of Complexity*, 3(4):359–371, 1987.

[Lemkin 81] P.F. Lemkin et L.E. Lipkin. Gellab: a computer system for 2d gel electrophoresis analysis i. segmentation of spots and system preliminaries. *Computers Biomed. Res.*, 14:272–297, 1981.

[Lenat 84] D. Lenat. The role of heuristics in learning by discovery: three case studies. Ryszard S. Michalski, Jaime G. Carbonnell, et Tom M. Mitchell, éditeurs, *Machine Learning: An Artificial Intelligence Approach*, pages 243–306, Springer Verlag, 1984.

[Lesser 75] V.R. Lesser, R.D. Fennell, L.D. Erman, et D.R. Reddy. Organisation of the Hearsay-II speech understanding system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23:11–23, 1975.

[Lutin 79] W.A. Lutin, C.F. Kyle, et J.A. Freeman. Quantitation of brain proteins by computer-analysed two-dimensional electrophoresis. *Electrophoresis'78*, pages 93–106, Elsevier, 1979.

[Marr 82] David Marr. *Vision.* Freeman, 1982.

[Matsuyama 85] Takashi Matsuyama et Vincent Hwang. Sigma: a framework for image understanding - integration of bottom-up and top-down analyses-. *IJCAI*, pages 908–915, 1985.

[Matsuyama 87] Takashi Matsuyama. Knowledge-based aerial image understanding systems and expert systems for image processing. *IEEE Transaction on Geoscience and Remote Sensing*, 25(3):305–316, May 1987.

[Michalski 80] Ryszard S. Michalski. Pattern recognition as rule-guided inductive inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4):349–361, July 1980.

[Michalski 84a] Ryszard S. Michalski. A theory and methodology of inductive learning. Ryszard S. Michalski, Jaime G. Carbonnell, et Tom M. Mitchell, éditeurs, *Machine Learning: An Artificial Intelligence Approach*, chapitre 4, pages 83–134, Springer Verlag, 1984.

[Michalski 84b] Ryszard S. Michalski et Robert E. Stepp. Learning from observation: conceptual clustering. Ryszard S. Michalski, Jaime G. Carbonnell, et Tom M. Mitchell, éditeurs, *Machine Learning: An Artificial Intelligence Approach*, chapitre 11, pages 331–363, Springer Verlag, 1984.

[Miller 82] Mark J. Miller, Phong K. Vo, Chris Nielsen, E. Peter Geiduschek, et Nguyen huu Xuong. Computer analysis of two-dimensional gels: semi-automatic matching. *Clinical Chemistry*, 28(4):867–875, 1982.

[Miller 88] Mark M. Miller. Analysis of sets of two-dimensional gel electrophoretograms: building a data base, correction of errors, analysis of data. Claus Schafer-Nielsen, éditeur, *Electrophoresis '88*, pages 322–335, The International Electrophoresis Society, The Protein Laboratory, University of Copenhagen, July 1988.

[Mohr 88] Roger Mohr, Long Quan, et Éric Thirion. Feature grouping: a way to deterministic matching. *Proceedings of the Workshop on Syntactical and Structural Pattern Recognition*, IAPR, 1988.

[Nagao 84] Makoto Nagao. Control strategies in pattern analysis. *Pattern Recognition*, 17(1), 1984.

[Nagao 88] Makoto Nagao, Takashi Matsuyama, et Hisayuki Mori. Structural analysis of complex aerial photographs. *Blackboard Systems*, pages 219–230, Addison-Wesley, 1988.

[Neel 84] J.V. Neel, B.B. Rosenblum, C.F. Sing, M.M. Skolnick, S.M. Hanash, et S. Sternberg. Adapting two-dimensional electrophoresis to the study of human germ-line mutation rates. Celis et Bravo, éditeurs, *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, 1984.

[Newell 72] A. Newell et H.A. Simon. *Human problem solving.* Prentice-Hall, 1972.

[Nordhausen 86] B. Nordhausen. Conceptual clustering using relational information. *Fifth National Conference on Artificial Intelligence*, AIII, 1986.

[Nugues 84] Pierre Nugues. *Algorithme d'ajustement de paramètres de spots d'électrophorèse bidimensionnelle.* Rapport de maîtrise, ENST, 1984.

[Nugues 88a] Pierre Nugues, Josiane Steinmetz, Jean-Paul Haton, Marie-Madeleine Galteau, et Gérard Siest. Using artificial intelligence in 2-d electrophoresis experiments. Claus Schafer-Nielsen, éditeur, *Electrophoresis '88*, pages 336–343, The International Electrophoresis Society, The Protein Laboratory, University of Copenhagen, July 1988.

[Nugues 88b] Pierre Nugues, Josiane Steinmetz, Jean-Paul Haton, Marie-Madeleine Galteau, et Gérard Siest. New perspectives in data bases query using artificial intelligence. T. Endler et S. Hanash, éditeurs, *Two Dimensional Electrophoresis*, VCH, Weinheim, Allemagne, 1988.

[OFarrell 75] P. H. O'Farrell. High resolution two-dimentional gel electrophoresis of proteins. *Journal of Biological Chemistry*, 250:4007–4021, 1975.

[Ohta 85] Yuichi Ohta. *Knowledge-based Interpretation of Outdoor Natural Color Scenes.* Pitman, 1985.

[Pair 88] C. Pair, R. Mohr, et R. Schott. *Construire les algorithmes.* Dunod, 1988.

[Pavlidis 82] Theo Pavlidis. *Algorithms for Graphics and Image Processing.* Springer-Verlag, 1982.

[Pratt 78] William K. Pratt. *Digital Image Processing.* Wiley-Interscience, 1978.

[Prehm 87] Joachim Prehm, Peter Jungblut, et Joachim Klose. Analysis of two-dimensional electrophoretic protein patterns using a video camera and a computer II. adaptation of automatic spot detection to visual evaluation. *Electrophoresis*, 8(12):562–572, December 1987.

[Pun 88] Thierry Pun, Denis Hochstrasser, et Christian Pellegrini. Correspondence analysis and hierarchical classification of complex images: application to two-dimensional gel electrophoretograms. *Signal Processing IV*, EURASIP, 1988.

[Quinlan 84] J.R. Quinlan. Learning efficient classification procedures and their application to chess end games. Ryszard S. Michalski, Jaime G. Carbonnell, et Tom M. Mitchell, éditeurs, *Machine Learning: An Artificial Intelligence Approach*, pages 463–482, Springer Verlag, 1984.

[Quinlan 86] J.R. Quinlan. The effect on noise on concept learning. Ryszard S. Michalski, Jaime G. Carbonnell, et Tom M. Mitchell, éditeurs, *Machine Learning: An Artificial Intelligence Approach, Volume II*, pages 149–166, Morgan Kaufmann, 1986.

[Rabilloux 85] T. Rabilloux, P. Vincens, et P. Tarroux. A new tool to study genetic expression using 2-d electrophoresis data: the functional map concept. *FEBS Letters*, 189(2), 1985.

[Rendell 86] Larry Rendell. A general framework for induction and a study of selective induction. *Machine Learning*, 1(2):177–226, 1986.

[Rich 83] Elaine Rich. *Artificial Intelligence.* McGraw-Hill, 1983.

[Ridder 84] Gregg Ridder, Ed VonBargen, David Burgard, Harvey Pickrum, et Emma Williams. Quantitative analysis and pattern recognition of two-dimensional electrophoresis gels. *Clinical Chemistry*, 30(12):1919–1914, 1984.

[Righetti 83] Pier Giorgio Righetti. *Isoelectric Focusing: Theory, Methodology and Applications.* Volume 11, série *Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Biomedical, 1983.

[Rosenblatt 58] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 58:386–407, 1958.

[Schlimmer 86] J.C. Schlimmer et R.H. Granger Jr. Incremental learning from noisy data. *Machine Learning*, 1:317–354, 1986.

[Serra 82] Jean Serra. *Image Analysis and Mathematical Morphology.* Academic Press, 1982.

[Serra 86] Jean Serra. Introduction to mathematical morphology. *Computer Vision, Graphics, and Image Processing*, 35(3):283–305, September 1986.

[Shafer 86] Steven A. Shafer, Anthony Stentz, et Charles E. Thorpe. An architecture for sensor fusion in a mobile robot. *Proc. IEEE Int. Conf. Robotics and Automation*, 1986.

[Shapiro 81] L.G. Shapiro et R.M. Haralick. Structural description and inexact matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-3(5), 1981.

[Skolnick 82a] Michael M. Skolnick. An approach to completely automatic comparison of two-dimensional gels. *Clincal Chemistry*, 28(4):978–986, 1982.

[Skolnick 82b] Michael M. Skolnick, Stanley R. Sternberg, et James V. Neel. Computer programs for adapting two-dimensional gels to the study of mutation. *Clinical Chemistry*, 28(4):969–978, 1982.

[Skolnick 86a] Michael M. Skolnick. Application of morphological transformations to the analysis of two-dimensional electrophoretic gels of biological materials. *Computer Vision, Graphics, and Image Processing*, 35(3):306–332, September 1986.

[Skolnick 86b] Michael M. Skolnick et James V. Neel. An algorithm for comparing two-dimentional electrophoretic gels, with particular reference to the study of mutation. Harry Harris et Kurt Hirschorn, éditeurs, *Advances in Human Genetics*, chapitre 2, pages 55–160, Plenum Publishing Corporation, 1986.

[Skolnik 85] Michael M. Skolnik. Automatic comparison of 2-d electrophoretic gels. *IEEE Proceeding on Computer Vision and Pattern Recognition*, IEEE, 1985.

[Sowa 84] J. F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine. The Systems Programming Series*, Addison-Wesley, 1984.

[Sprecher 84] Dennis L. Sprecher, Lila Taam, et H. Bryan Brewer Jr. Two-dimensional electrophoresis of human plasma apolipoproteins. *Clinical Chemistry*, 30(12):2084–2092, 1984.

[Stepp 86] Robert E. Stepp et Ryszard S. Michalski. Conceptual clustering of structured objects: a goal-oriented approach. *Artificial Intelligence*, 28(1):43–69, 1986.

[Sternberg 86] Stanley R. Sternberg. Grayscale morphology. *Computer Vision, Graphics, and Image Processing*, 35(3):333–355, September 1986.

[Tamura 84] Hideyuki Tamura et Naokazu Yokoya. Image database systems: a survey. *Pattern Recognition*, 17(1):29–43, 1984.

[Tarroux 87] Philippe Tarroux, Pierre Vincens, et Thierry Rabilloud. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels Part V: Data analysis. *Electrophoresis*, 8(4):187–199, April 1987.

[Taylor 79] J. Taylor, N. L. Anderson, et B. P. Coultern. Estimation of 2-dimentional electrophoretic spot intensities and positions by modelling. B. J. Radola, éditeur, *Electrophoresis'79*, pages 329–339, De Gruyter, 1979.

[Taylor 83] John Taylor, N. Leigh Anderon, et Norman G. Anderson. Numerical measures of two-dimensional gel resolution and positional reproducibility. *Electrophoresis*, 4(5):338–346, October 1983.

[Terry 88] Allan Terry. Using explicit strategic knowledge to control expert systems. *Blackboard Systems*, pages 159–188, Addison-Wesley, 1988.

[Thirion 88] É. Thirion et R. Mohr. Matching 3-d images without backtracking through feature grouping. *Proceedings of the 8th European Conference on Artificial Intelligence*, pages 678–682, 1988.

[Thorpe 88] Charles Thorpe, Martial Hebert, Takeo Kanade, et Steven Shafer. Vision and navigation for the carnegie-mellon navlab. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10(3), 1988.

[Toussaint 80] Godfried T. Toussaint. Pattern recognition and geometrical complexity. *5th International Conference on Pattern Recognition*, pages 1324–1347, IAPR, IEEE, December 1980.

[Tracy 84] Russel Tracy et Donald Young. Clinical applications of two-dimensional gel electrophoresis. Celis et Bravo, éditeurs, *Two-Dimensional Gel Electrophoresis of Proteins*, Academic Press, 1984.

[Ulichney 88] Robert A. Ulichney. Dithering with blue noise. *Proceedings of the IEEE*, 76(1), 1988.

[Vere 80] S.A. Vere. Multilevel counterfactuals for generalizations of relational concepts and productions. *Artificial Intelligence*, 14:139–164, 1980.

[Vincens 86a] Pierre Vincens. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels Part II: Spot detection and integration. *Electrophoresis*, 7(8):357–367, August 1986.

[Vincens 86b] Pierre Vincens, Nicolas Paris, Jean-Luc Pujol, Christine Gaboriaud, Thierry Rabilloud, Jean-Louis Pennetier, Philippe Matherat, et Philippe Tarroux. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels Part I: Data acquisition. *Electrophoresis*, 7(8):347–356, August 1986.

[Vincens 87a] Pierre Vincens. *Analyse informatisée des images de gels d'électrophorèse bidimentionnelle.* Volume 32, série *Publications du laboratoire de zoologie*, École normale supérieure, E. N. S. laboratoire de zoologie, 46, rue d'Ulm Paris, 1987.

[Vincens 87b] Pierre Vincens et Philippe Tarroux. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels Part IV: Data base organization and management. *Electrophoresis*, 7(4):173–186, April 1987.

[Vincens 87c] Pierre Vincens et Philippe Tarroux. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels Part III: Spot list matching. *Electrophoresis*, 8:100–107, 1987.

[Wang 83] David C.C. Wang, Anthony H. Vagnucci, et C.C. Li. Digital image enhancement: a survey. *Computer Vision, Graphics, and Image Processing*, 24:363–381, 1983.

[Westerbrink 84] Klaas Westerbrink. *High Resolution Protein Mapping, Technique and applications, with special reference to the characterization of human lymphomas.* Thèse de Doctorat, Rijsuniversitet te Groningen, Groningen, Pays-Bas, 1984.

[Wiederkehr 86] F. Wiederkehr, H. Imfeld, et D.J. Vonderschmitt. Two-dimensional gel electrophoresis, isoelectric focusing and agarose gel electrophoresis in the diagnosis of multiple sclerosis. *Journal of Clinical Chemitry and Clinical Biochemistry*, 24(12):1017–1021, 1986.

[Winston 84] Patrick Winston et Klaus Horn. *Lisp.* Addison Wesley, 1984.

[Wong 82] K.Y. Wong, R.G. Casey, et F.M. Wahl. Document analysis system. *IBM Research and Development Journal*, 26:647–655, 1982.

[Woods 77] J.W Woods et C.H. Radewan. Kalman filtering in two dimensions. *IEEE Transactions on Information Theory*, IT-23(4), 1977.

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

NOM DE L'ETUDIANT : Monsieur NUGUES Pierre

NATURE DE LA THESE : DOCTORAT DE L'UNIVERSITE DE NANCY I
EN INFORMATIQUE

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 3 Mars 1989 n° 462

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

M. Boulangé

Résumé

Cette thèse a porté sur l'étude et la réalisation d'un système d'interprétation d'images et d'apprentissage symbolique.

Le système d'interprétation d'images identifie automatiquement, sur un gel d'électrophorèse, des protéines, isolées ou à l'intérieur de constellations, en reproduisant les méthodes d'experts biologistes. Il se fonde sur une architecture modulaire comprenant des procédures de traitement d'images conduisant à l'extraction des paramètres et d'un processus de raisonnement ascendant et descendant. Ce processus fait d'abord correspondre les paramètres extraits aux modèles géométriques potentiels des protéines puis revient sur l'image pour déterminer les éléments éventuellement manquants. Il est précédé d'une focalisation de l'attention. Ce système a été appliqué avec succès aux apoliporotéines du plasma.

Le système d'apprentissage symbolique a pour objectif de fournir une interprétation de séries d'expériences et s'inspire de Cluster-2 de R. Michalski. Sa description est précédée d'une étude et d'une comparaison de méthodes symboliques et numériques ainsi que de l'exposé de deux types d'amélioration concernant la généralisation et l'évaluation. Le système d'apprentissage permet de classer les gels correspondant aux étapes de la croissance de cellules musculaires.

Mots clés : électrophorèse bidimensionnelle, interprétation d'image, traitement d'image, apprentissage symbolique, classification conceptuelle, intelligence artificielle.

I S B N - 2 - 7261 - 0588 - 2