Chapitre 15

Forme des données et recherche de profils biologiques en pathologie hépato-biliaire

0 - FICHE TECHNIQUE

A - Tableau T_0 de mesures biologiques numériques ; variables numériques × Individus de dimension 16 × 228, avec données manquantes et sous tableau T_1 de dimension 16 × 117 où aucun résultat de mesure ne manque. Après avoir constaté, à un niveau global, la stabilité des résultats lorsqu'on passe de T_0 à T_1 , on travaille surtout par rapport à T_1 . En construisant une subdivision "statistiquement significative" de l'intervalle de variation d'un même paramètre biologique, on associe à T_1 un tableau T_2 de même dimension 16 × 228 où les variables sont discrètes totalement préordinales (i.e. caractères descriptifs aux ensembles respectifs des modalités totalement ordonnés). En attachant à chaque modalité un attribut de description, on définit un tableau d'incidence T_3 de dimension 61 × 117. (cf. chap. 2, § II.1, 2, 2').

B - Classification hiérarchique de l'ensemble des variables et de l'ensemble des sujets pour chacun des codages définis par T_1 , T_2 et T_3 . Définition des profils biologiques des classes de malades, relativement au codage initial matérialisé par T_1 .

C - Nous allons, pour chacun des codages supportés par les différents tableaux T_0 , T_1 , T_2 et T_3 , préciser l'indice ou les indices de proximité considérés entre variables descriptives et entre individus.

Tableau T_0

Soit (v, w) le couple de variables sur lequel il y a lieu de déterminer la valeur du coefficient d'association. Désignons par E_v de cardinal n_v (resp. E_w de cardinal n_w) l'ensemble des individus pour lesquels la mesure de v (resp. w) est présente. Un premier indice est défini, à un coefficient multiplicatif près, par le coefficient de corrélation entre v et w calculé sur $E_v \cap E_w$. Un deuxième indice peut être défini par la formule :

$$\sum_{x \in E_v \cap E_w} (v(x)-\bar{v})(w(x)-\bar{w})/\sigma(v)\sigma(w),$$

à un coefficient multiplicatif près, où $\bar{v}$ et $\sigma^2(v)$ (resp. $\bar{w}$ et $\sigma^2(w)$) sont la moyenne et la variance de la variable v sur E_v (resp. de w sur E_w), (cf. [3] chap. IV). Pour l'un ou l'autre de ces deux indices, le résultat de la classification des variables ayant été quasi-identique à celui, dans le cadre du tableau T_1 où les données sont complètes ; on s'est limité à l'analyse de ce dernier tableau T_1 et de ses dérivés.

Tableau T_1

L'indice de proximité entre les deux variables v et w est défini par $\sqrt{(n-1)}\ \rho(v,w)$ où n est le cardinal de l'ensemble des individus (ici 117) et $\rho(v,w)$ est le coefficient de corrélation entre v et w (cf. chap. 2, § IV.2). Pour la classification de l'ensemble des sujets (matériel clinique), deux indices de proximité ont été testés ; le premier est, au facteur multiplicatif $\sqrt{m-1}$ près où m est le cardinal de l'ensemble des variables, le coefficient de corrélation entre individus, formellement identique à celui entre variables après transposition du tableau des données (cf. formule (1), chap. 2,§VI.5). Le deuxième indice (cf. formule (2), chap. 2, § VI.5) est le cosinus de l'angle des deux vecteurs représentant les deux individus où les composantes d'un même vecteur sont les valeurs de la suite des variables de description sur l'un des individus. Rappelons l'expression de chacun des deux indices.

$$\rho(x,y)=\frac{\sum_{v\in V}[v(x)-m(x)][v(y)-m(y)]}{\{\sum_{v\in V}[v(x)-m(x)]^2 \sum_{v\in V}[v(y)-m(y)]^2\}^{1/2}}, \qquad (1)$$

où $m(x)=\frac{1}{m}\sum_{v\in V} v(x)$.

$$\gamma_o(x,y)=\frac{\sum_{v\in V} v(x)v(y)}{\{\sum_{v\in V} v(x)^2 \sum_{v\in V} v(y)^2\}^{1/2}}, \qquad (2)$$

Tableau T_2.

L'indice de proximité entre variables discrètres totalement préordinales (i.e. caractères descriptifs aux ensembles respectifs des modalités totalement ordonnés) est celui (L) déjà evoqué dans la fiche technique du chapitre 12 précédent (cf. chapitre 2, § IV. 4.2.3.).

L'indice de proximité entre individus est celui développé au chapitre 2 paragragraphe VI.3. L'expression précise est trop lourde pour être reproduite ici. Toutefois, on peut la retrouver sans difficulté à partir des formules (3) et (4) (chap. §VI.3) et son calcul sur ordinateur s'organise clairement.

Tableau T_3.

Il s'agit d'un tableau d'incidence ; l'indice de proximité Q(a,b) adopté entre attributs de description a et b est, avec des notations entendues (cf. chap. 2, §IV. 1) :

$$Q(a,b)=[\,\mathrm{card}(E_a \cap E_b)- \frac{n_a n_b}{n}] \Big/ \sqrt{\frac{n_a n_b}{n}} .$$

Deux indices de proximité entre individus ont ici été utilisés. Le premier R(x,y) est, à une transposition du tableau des données T_3 près, formellement identique à Q(a,b) ; soit

$$R(x,y) =[\,\mathrm{card}(A_x \cap A_y)- \frac{m_x m_y}{m}] \Big/ \sqrt{\frac{m_x m_y}{m}}$$

où m est le cardinal de l'ensemble A des attributs et où $m_x=\mathrm{card}(A_x)$ (resp. $m_y=\mathrm{card}(A_y)$) est le cardinal de l'ensemble des attributs possédés par l'individu x (resp. y), pour tout x et tout y appartenant à E.

Le deuxième indice S(x,y), où on tient compte de la fréquence relative des différents attributs descriptifs sur E, est défini par la formule

$$S(x,y)=[\,\mathrm{card}(A_x \cap A_y)-(\sum_{1\leq j\leq m} p_j^2))] \Big/ \sqrt{\sum_{1\leq j\leq m} p_j^2}$$

où p_j est la fréquence relative du j-ème attribut ; pour tout x et tout y de E.

La réduction globale des similarités a été à chaque fois effectuée en centrant et en réduisant par rapport à la distribution de l'indice sur l'ensemble des paires de l'ensemble à organiser ; ensemble des variables ou bien, ensemble des individus (cf. chap.2. fin § IV.5).

D - Algorithme de la Vraisemblance des Liens (A.V.L.). Utilisation de la chaîne de programmes PROX-ORDON-POLON-ARBRE pour la classification de l'ensemble des variables dans le cas du codage supporté par le tableau T_2 et de la chaîne ORDON-POLON-ARBRE dans les autres cas où l'indice "centré réduit" a été calculé dans un programme séparé lorsque ce calcul n'est pas prévu dans ORDON où alors, on rentre dans la chaîne par une option spéciale.

E - Analyse de la variation des résultats de la classification des variables et de celle des objets lorsqu'on subtitue aux échelles des variables descriptives numériques des échelles discrètes plus pauvres obtenues à partir d'un algorithme de découpage qui respecte "au mieux" la densité de la distribution d'une même variable sur l'échantillon étudié. Test de nouveaux indices de proximité entre individus décrits par des variables discrètes ; notamment des caractères aux ensembles respectifs des modalités totalement ordonnés.

F - J.Y.LAFAYE

G - Dr HITA DE NERCY, M. KERBAOL et Pr. LENOIR.

La suite des titres des paragraphes est

1 - Introduction

2 - Présentation du fichier brut.

3 - Exposé et Interprétation des résultats.

1 - INTRODUCTION

La présente étude a vu sa genèse en la rencontre de deux préoccupations majeures :

- d'une part celle de répondre à un certain nombre de questions posées par le spécialiste au sujet de la pathologie hépato-bilaire,

- d'autre part celle de tester dans un domaine nouveau, les algorithmes de classification hiérarchique de I.C. Lerman dont l'efficacité était déjà apparue sur d'autres types de données ; notamment celles en Psycho-Sociologique et en Psycho-Pédagogie (cf. Chapitres précédents).

a - LE PROBLEME ET LE MEDECIN

1 - LE CADRE

Le problème tel qu'il se posait au praticien peut schématiquement être exposé comme suit (cf. [1]).

La pathologie hépato-biliaire présente un ensemble important de manifestations diverses. Le diagnostic peut en conséquence s'effectuer sur la base d'arguments de différents ordres :

- évolutif,
- laparoscopique,
- chirurgical,
- histologique...
- biologique.

Précisément, la biologie, et plus particulièrement le diagnostic enzymologique tend actuellement à occuper une place privilégiée parmi les différentes méthodes d'investigation de la pathologie hépato-biliaire.

2 - LES QUESTIONS SOULEVEES

Dans ces conditions, peut-on préciser les limites en diagnostic

purement enzymologique, par rapport à un diagnostic mettant en jeu l'ensemble des arguments énoncés ci-dessus ?

Le diagnostic enzymologique et le bilan fonctionnel hépatique font appel à un certain nombre de mesures "Traditionnelles".

Il s'agit alors, dans la mesure du possible, de se prononcer sur le rôle spécifique de chacune de ces mesures dans le dépistage, le diagnostic et la surveillance des affections hépato-biliaires.

Le rôle de chaque enzyme peut être étudié sous l'angle qualitatif ; on met alors en lumière les différents déséquilibres qu'il indique. Il peut également être abordé sous l'angle quantitatif, on s'intéresse alors tout particulièrement au degré de sensibilité de l'enzyme considéré.

Enfin, le rôle d'un enzyme peut être étudié de manière comparative, on s'attachera alors à observer son caractère plus ou moins discrimant, plus ou moins pertinent, ou à l'opposé redondant, relativement aux autres enzymes dont on dispose.

On voit, se basant sur d'éventuelles conclusions à ce sujet, l'intérêt qu'il y aurait à réduire, à efficacité diagnostique égale, le nombre d'examens systématiquement effectués.

Cela tant au niveau des nuisances subies par le patient qu'à celui des coûts supportés par les unités médicales.

Parallèlement, on prône la mesure et la prise en compte de nouveaux enzymes. Ceux-ci mériteraient, selon la nature de l'information apportée, ou bien de s'ajouter au bilan fonctionnel hépatique classique, ou bien d'y prendre la place d'un test ancien, moins pertinent ou plus coûteux.

Qu'en est-il ? Doit-on effectivement introduire ces nouveaux enzymes dans le bilan de routine ?

Si l'on élève la réflexion à un plan plus général, on est amené à s'interroger sur la nature de la démarche diagnostique elle-même.

Le problème du diagnostic biologique des affections hapato-biliaires réside grossièrement dans trois questions fondamentales :

- Comment caractériser de façon objective, et pour chaque paramètre biologique envisagé, une valeur théorique "normale" servant de référence ?

- Comment juger, en terme de pathologie, la grandeur d'un écart à une telle valeur ?

- De quelle manière effectuer une synthèse des informations apportées par l'ensemble des paramètres biologiques.

Pour illustrer la difficulté d'une telle synthèse nous citerons deux exemples précis :

- l'interprétation de l'élévation du taux de fer sérique sera différente selon la présence ou l'absence d'hémorragie chez le sujet observé.

- de même une élévation de la "Gamma glutamyl transpeptidase" apportera des renseignements différents selon qu'elle se situe dans un cadre d'insuffisance hepato-cellulaire ou bien dans un cadre cholestatique.

La simplicité de ces deux exemples typiques et bien connus, ne doit pas faire douter de la complexité et du nombre des nuances des interprétations des résultats d'un bilan fonctionnel hépatique.

D'une manière générale, la spécificité d'un paramètre biologique, ne peut s'interpréter valablement qu'à l'intérieur d'un "contexte".

Cette notion de contexte correspond à l'ensemble de toutes les manifestations concernant, directement ou non, l'état pathologique du patient.

Dans les exemples précédents, le contexte était artificiellement réduit aux notions de syndrômes cholestatique ou d'insuffisance hépato-cellulaire, à la présence ou à l'absence d'hémorragie, mais il s'agit, de façon idéale, d'un concept beaucoup plus large.

Or, si l'on se limite à la biologie, ce contexte peut être abordé uniquement par la prise en compte des informations apportées par les autres paramètres biologiques mesurés.

Schématiquement on peut représenter ainsi les liens de dépendance entre les trois questions fondamentales précédentes.

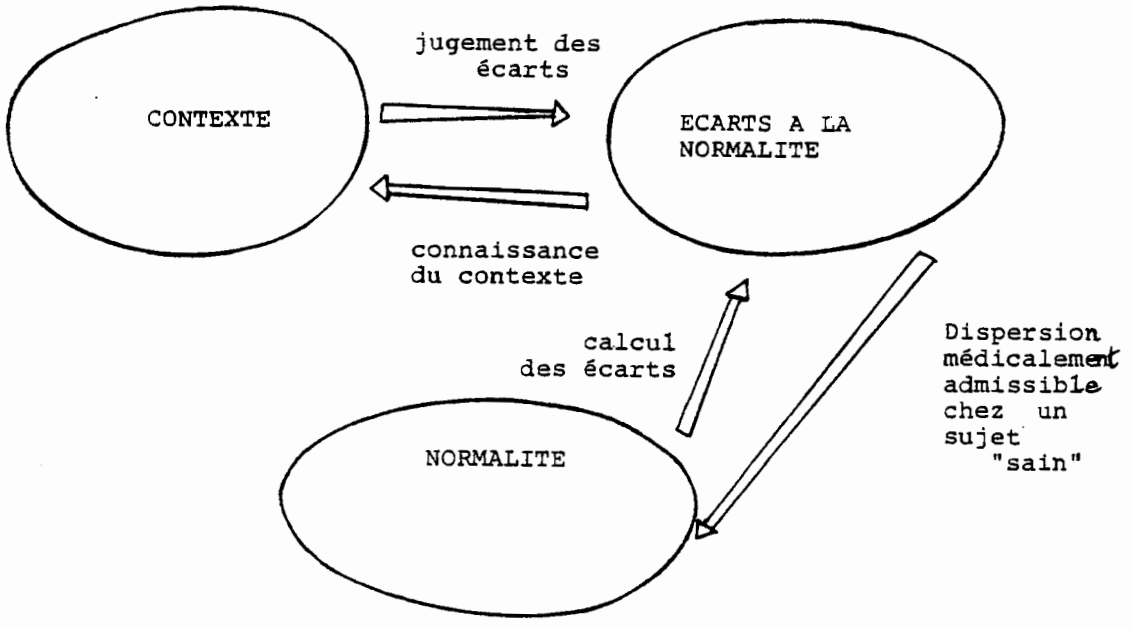


Figure 1

Le rôle important du contexte dans la formulation d'une réponse aux interrogations du praticien, le caractère très intime des liaisons existant entre les problèmes soulevés par les trois questions fondamentales relatives au diagnostic chacune renvoyant finalement plus ou moins directement aux deux autres, justifie l'approche statistique globale qui sera la nôtre.

b - LE PROBLEME ET LE STATISTICIEN

1 - LES APPROCHES TRADITIONNELLES

Jusqu'ici, la majorité des travaux conduits dans le domaine de l'investigation biologique de la pathologie hépato-biliaire, consistait en l'analyse d'un nombre limité d'enzymes étudiés le plus souvent séparément.

Ces études étaient principalement à caractère qualitatif, les résultats quantifiés se bornant à des analyses de distributions statistiques unidimensionnelles ou, au mieux de correlations simples.

Une telle approche du problème rejetait donc entièrement sur le médecin, la formulation personnelle d'une synthèse. Celle-ci, dans ces conditions, était le fruit de la compilation, difficile et fastidieuse d'une bibliographie pléthorique.

2 - APPORT DES ANALYSES MULTIDIMENSIONNELLES

La généralisation des méthodes de traitement de données multidimensionnelles, aussi bien géométriques que combinatoires, a permis dans l'ensemble de la médecine une approche statistique globale, élégante et féconde.

Plus précisément, différentes études ont déjà apporté des résultats positifs dans le domaine de la pathologie hépato-biliaire, et cela en utilisant l'ensemble des techniques d'analyse multidimensionnelles classiques.

A titre d'illustration :

- Des conclusions basées sur l'analyse discriminante apportent des éléments de réponse intéressants (bien que médicalement peu réconfortants) sur la nécessité du recours à la chirurgie pour se prononcer sur la nature intra ou extra hépatique des cholestases (cf. [4]).

- L'utilisation de méthodes dérivées des nuées dynamiques apporte de riches enseignements sur la modification des protéines sériques en pathologie clinique (cf. [5]).

- Enfin, un traitement basé sur l'analyse des correspondances a été mené dans le domaine précis qui nous préoccupe de l'approche globale de la "géographie" de la pathologie hépato-biliaire ainsi que de l'organisation de l'ensemble des paramètres biologiques qui permettent de la caractériser (cf. [1]).

c - UNE METHODE D'ANALYSE FACE A DES APPLICATIONS NOUVELLES

Outre les goûts et les habitudes personnels qui conduisent à préférer telle ou telle méthode d'analyse de données et telle ou telle démarche de traitement préalable d'un échantillon, la meilleure solution qui s'offre au statisticien pour guider son choix, est la comparaison des résultats obtenus dans différents cas sur des données semblables à celles qu'il se propose d'analyser.

Cependant, dans le cas d'un domaine d'application nouveau pour une technique donnée, l'expérience faisant gravement défaut, il est difficile et très peu naturel, de transposer des résultats obtenus dans des domaines autres. Nous avons, en l'espèce, déjà montré le caractère tout particulier des données médicales.

La question première qui se pose dans cette situation peut s'exprimer en terme de stratégie :

Il s'agit, face à des données appréhendées sous une forme brute d'opérer une suite de transformations. Les données présentées sous forme élaborée étant alors directement soumises à l'interprétation.

La perplexité du statisticien est en proportion du nombre important de façons variées d'opérer ces transformations et du manque de critère pour juger, a priori, du bien fondé de celles-ci.

Une fois le choix d'une méthode et d'un algorithme de traitement, il s'agira dans notre cas de la classification hiérarchique par l'A.V.L. l'adoption d'une conduite fructueuse face aux données est le résultat d'une stratégie qui doit se développer principalement dans deux directions :

- choix d'un codage

On se trouve alors directement confronté avec le problème de la reconnaissance de la nature et de la part d'information effectivement utilisable pour la conduite du diagnostic.

La complexité -dans la plupart des cas- du processus menant au diagnostic biologique des maladies hépato-biliaire fait que le médecin peut difficilement analyser et traduire nettement sa démarche, il ne nous apporte que peu de renseignements au sujet de l'information utilisée, et encore moins sur l'information utilisable. Un des objets du présent travail sera d'ailleurs de préciser ce point.

En tout état de cause, le codage se doit de respecter cette part d'information pertinente indispensable à la clarté des résultats et au diagnostic automatique ou non. Gardant cette pré-occupation ; nous adopterons différents codages passant des échelles descriptives les plus "riches" à celles, les plus "pauvres".

- choix d'un indice de similarité

Ce choix est certes intimement lié à celui précédent du codage.

D'autre part, on restera dans le cadre des indices de proximités qui ont été développés au chapitre 2 et dont certains, concernant notamment la comparaison des individus dans le cas de variables discrètes, ont été élaboré à l'occasion de cette étude. Toutefois, précisément surtout dans ce dernier cas, différentes formes de l'hypothèse d'absence de lien par rapport à laquelle l'indice est défini sous la forme d'une vraisemblance, peuvent a priori être considérées. Principalement deux formes ; celle où on ne tient pas compte de la distribution observée sur l'échantillon des variables discrètes de description et celle où on tient compte.

Dans le cas, comme il se présente ici, où les variables sont au départ numériques, la définition d'un indice de similarité entre individus doit procéder de considération géométriques. Deux indices seront principalement expérimentés : le coefficient de corrélation et le cosinus de l'angle des deux vecteurs représentant les deux objets où les composantes d'un même vecteur représentent les valeurs de la suite des variables de description sur l'un des objets.

Dans le cadre où nous nous sommes placés, le choix d'une stratégie correspond donc au choix d'un codage et d'un indice de proximité associé ; entre variables ou bien entre objets selon le cas analysé.

Ceci dit, il reste que, à l'intérieur de ce cadre, des stratégies différentes mais également fécondes amènent des résultats variés. Chacun ayant sa signification, son optique et ses nuances propres. Il est naturel que l'on puisse souhaiter des résultats à un niveau très global correspondant à une organisation grossière mais simple, ou bien au contraire à une analyse locale à un niveau de détail et plus complexe.

On peut de la même façon souhaiter privilégier par exemple les liaisons tenues au détriment des liens trop grossier qui les masquent, etc...

Notre ambition serait alors, au terme de notre étude, de pouvoir évaluer l'intérêt particulier des différentes stratégies envisageables, et de contribuer à caractériser, dans un cadre général, leurs spécificités. Entendons par là, contribuer à prédire que tel type de stratégie conduit à telle forme de résultats.

L'objectif final étant alors, outre le traitement d'un fichier particulier, d'apporter une expérience généralisable concernant les possibilités d'une méthode et les différents formes de résultats accessibles, dans le domaine médical ou dans des domaines structurellement voisins.

2 - PRESENTATION DU FICHIER BRUT

Le fichier résumant les données à analyser nous a été fourni par l'INSERM (U49 - Rennes), qui achevait d'en effectuer le traitement par analyse des correspondances (cf. [1]).

Il s'agissait des valeurs de 16 paramètres biologiques mesurés sur 117 malades, atteints principalement d'affections hépatiques. Chaque malade avait effectivement subi la totalité des 16 mesures.

A cela s'ajoutait un fichier complémentaire présentant de nombreuses données manquantes, concernant toutefois les 16 mêmes paramètres biologiques mesurés sur 111 autres malades. Le traitement de ce fichier par des méthodes métriques nécessitait la reconstitution des données absentes et n'avait pas été entrepris par l'INSERM.

Les mesures des 16 paramètres sur les 228 malades sont représentées par un tableau 16 × 228 de nombres réels positifs.

Nous exposerons successivement et d'une manière détaillée la composition du matériel biologique (les 16 paramètres) et celle du matériel clinique (les malades). Nous terminerons par quelques remarques d'ordre général concernant à la fois l'un et l'autre.

a - LE MATERIEL BIOLOGIQUE

1 - CONDITIONS DE MESURE

Les 16 paramètres biologiques ont été mesurés sur les 228 malades dans des conditions identiques et suivant les mêmes méthodes de dosage. Les observations ont été recueillies avant tout traitement, dès l'entrée dans les services hospitaliers. Ces remarques ont leur importance particulière étant donné la sensibilité des mesures effectuées aux facteurs : temps, matériel de mesure, méthode de dosage... La liste des observations est fournie dans [3].

Il convenait de ne pas ajouter artificiellement à l'hétérogénéité intrinsèque que nous mettrons par la suite en évidence dans l'échantillon.

2 - NATURE DES PARAMETRES BIOLOGIQUES

Parmi les 16 paramètres biologiques, 12 constituent le bilan hépatique de routine et leur mesure est effectuée systématiquement. Les quatre autres sont les nouveaux enzymes dont on désire étudier l'intérêt. (Elles sont marquées "Nouv" dans la liste qui suit).

Afin de justifier certaines remarques ultérieures, nous donnerons rapidement un aperçu de la méthode de dosage utilisée pour la mesure de chaque paramètre, ainsi qu'un tableau récapitulant les noms, les observations et les unités de mesure.

Liste des différentes variables, unités, symboles

Dénomination	Symbole	Unité
- sérum glutamo pyruvique transaminase	T G P	unité Frankel
- sérum glutamo oxalacétique transaminase	T G O	unité Frankel

- lactico-deshydrogénase totale	L D H T	m u i
- sidérémie totale	F E R	gramme par litre de sang
- protidémie totale	PROTIDES	gramme par litre de sang
- albuminémie	ALBU	%
- prothrombinémie	PROTHR	%
- bilirubinémie totale	BILIRU	milligramme par litre de sang
- phosphatase alcaline	PH.ALC	unité Bodanski
- 5'nucléotidase	5'NUCL	u i par litre de sérum
- leucine-amino peptidase	L A P	m u i
- gamma glutamil transpeptidase	G G T P	m u i par millilitre
- alpha 2 globulinémie	ALPHA	%
- gamma globulinémie	GAMMA	%
- vitesse de sédimentation	S V S	
- K 1 de la BSP	Kl BSP	%

u i = unité internationale
m u i = milli unité internationale

DIFFERENTES METHODES DE MESURE DES PARAMETRES BIOLOGIQUES

(mesures anciennes : ANC, nouvelles : NOUV)

* Sérum GLUTAMO PYRUVIQUE (ANC)
* Sérum GLUTAMO OXALACETIQUE (ANC)

Ces deux transaminases interviennent comme catalyseurs des réactions de transamination. Le facteur retenu pour évaluer les deux variables qui nous intéressent sera l'activité enzymatique mesurée par une vitesse d'oxydation intervenant dans la transamination.

* LACTICO-DESHYDROGENASE TOTALE (NOUV)

Cet enzyme catalyse une réaction de transfert d'hydrogène. Son activité sera quantifiée par la variation, par minute d'extraction d'hydrogène.

* PHOSPHATASE ALCALINE (ANC)
* 5'NUCLEOTIDASE (NOUV)
* LEUCINE AMINO PEPTIDASE (NOUV)
* GAMMA GLUTAMYL TRANSPEPTIDASE (NOUV)

On détermine traditionnellement l'activité de ces enzymes grâce à leurs propriétés d'hydrolyse, mesurée par l'intermédiaire de la quantité de substrat libéré lors de la réaction.

* SIDEREMIE TOTALE (ANC)
* PROTIDEMIE TOTALE (ANC)
* BILIRUBINEMIE (ANC)

Mesurent respectivement le poids de fer, de protides, de bilirubine par unité de sérum analysé.

* ALBUMINEMIE (ANC)

* 2 GLOBULINEMIE (ANC)
* GLOBULINEMIE (ANC)

Les mesures de ces trois variables résultent de l'électrophorèse des protides.

* TAUX DE PROTHROMBINE (ANC)
* VITESSE DE SEDIMENTATION (ANC)
* K1 DE LA B S P (ANC)

Cette dernière mesure globale résume la facilité de captation et d'excrétion du colorant (B S P) au niveau du système réticulo-endothilique.

Il a été procédé en accord avec le médecin, à un écrètement de certaines variables dont la valeur anormalement élevée était sans enseignement supplémentaire par rapport à la valeur limite choisie pour l'ecrètement.

On a ainsi :

TGP, TGO	écrété à	999 UF
5'N	"	99,9 UI
ggtp	"	999 mUI

b - LE MATERIEL CLINIQUE

Les 116 malades ayant subi la totalité des mesures -et à fortiori les 228 de l'échantillon total- recouvrent dans la mesure du possible l'ensemble de la pathologie hépato-biliaire.

Ils correspondent à la totalité des malades venus consulter les services d'hépatologie à l'hôpital Pontchaillou (35 RENNES) durant une durée d'un an et demi.

1 - LES DIAGNOSTICS

Précisons dès maintenant que les diagnostics qui nous serviront de références lors de la caractéristation de chaque individu, tout au long de l'étude, ont été établis à partir de l'ensemble des arguments disponibles et jamais uniquement à partir d'arguments biologiques.

Cela est nécessaire, si l'on désire effectivement étudier -comme nous nous le sommes proposés- les limites du diagnostics biologique. Ces limites ne peuvent être appréhendées et définies que par rapport à un cadre plus large. Tout argument de décision, autre que biologique contribue à la fois à élargir ce cadre et à confirmer le bien fondé du diagnostic finalement établi.

2 - STRUCTURE DU MATERIEL CLINIQUE

Nous fournirons très maintenant un aperçu de la constitution de l'échantillon en précisant les classes pathologiques présentes ainsi que leur fréquence d'apparition. La liste détaillée des malades et des

diagnostics qui leur correspondent pourra être trouvée dans [3].

Les classes pathologiques principales et leur fréquence dans l'échantillon

CIRRHOSES	: 30	
CIRRHOSES suspectes de cancer	: 8	38
HEPATITES aïgues	: 12	
HEPATITES chroniques cirrhogènes	: 8	24
HEPATITES prolongées	: 4	
HEMOCHROMATOSES primitives	: 17	
MEMOCHROMATOSES secondaires	: 7	24
CANCERS	:	10
HEPATOPATHIES ALCOOLIQUES	:	7
LITHIASES	:	5
SAINS	:	5
DIVERS	:	4
		117

La séduisante simplicité de ce tableau ne doit pas faire douter de la complexité de l'organisation du matériel clinique. Il n'est donné ici qu'à titre indicatif à seule fin de juger de la proportion relative des classes pathologiques.

Chaque malade est affecté à une classe pathologique particulière d'une façon plus ou moins nette. Certains peuvent présenter à la fois deux affections, par exemple cirrhose et cancer. Enfin, ils peuvent souffrir d'affections autres que hépatiques. Précisément, les individus qualifiés de "sains" le sont par rapport à une "ponction biopsie hépatique" normale, mais sont des sujets atteints de maladies générales. Ils ne peuvent donc servir de "témoins" d'une normalité idéale.

Enfin, certains malades figurent plusieurs fois dans l'échantillon, ayant été retenus à différents stades de leur évolution.

Cela pourraît être effectué de façon systématique. On observe alors des évolutions diverses, selon les cas, décompensation ou retour vers la "normalité".

c - REMARQUES SUR LES MATERIELS BIOLOGIQUE ET CLINIQUE

1 - NATURE ET STRUCTURE

- Matériel biologique :

L'étude, même superficielle de l'ensemble des 16 paramètres biologiques fait apparaître :

- son caractère particulièrement hétérogène, celui-ci regroupant des mesures de poids, vitesse, %, etc...

- les étendues et dispersions très diverses. (Celles-ci découlent d'ailleurs en partie du point précédent). On notera des intervalles de variation allant, après écrètement de [0,30] à [0,1000] !

- Matériel clinique :

L'échantillon "choisi" peut choquer par sa taille très limitée, a priori nuisible à la stabilité et à l'interprétabilité des résultats.

Il faut alors bien se pénétrer de la spécificité du domaine médical.

Le recueil des données telles qu'elles se présentent, a nécessité dix huit mois ininterrompus.

Si cela est hélas grandement suffisant pour observer un nombre statistiquement satisfaisant de cas de cirrhoses et d'hépatites, il en va autrement pour des affections plus rares telles que granulomatoses ou hemochromatoses (la Bretagne présentait curieusement une fréquence élevée d'hemochromatoses par rapport au niveau national).

En médecine, l'échantillon peut rarement être construit suivant l'orthodoxie statistique. Il résulte plus généralement de contraintes dues à la rareté de certains groupes, aux difficultés d'observation, et à la déontologie ; la raison scientifique ne pouvant justifier une pléthore de mesures ou d'interventions.

En conséquence, l'échantillon étudié représente assez fidèlement la fréquence de chaque classe pathologique dans la population bretonne.

2 - TRAITEMENT

Les inconvénients inhérents au domaine étudié ne peuvent être réduits au niveau de la collecte des données.

Ce sont donc eux qui serviront de base à toute réflexion sur l'élaboration d'une stratégie pour leur traitement.

- Codages et indices, devront donc être envisagés par rapport à des données hétérogènes dans leur nature et dans leur échelle.

- Des pondérations éventuelles devront privilégier des entités rares mais pertinentes.

- Les algorithmes devront être efficaces dans la constitution et la reconnaissance de classes de faibles effectifs, et éviter en particulier des effets de chaînage abusifs dont le risque est ici important.

3 - EXPOSE ET INTERPRETATION DES RESULTATS

Nous exposerons ici successivement les résultats des traitements suivants :

3.1. Classifications du matériel biologique et du matériel clinique dans le cas du codage initial numérique.

3.2. Analyse des relations entre syndrômes et classes pathologiques ; reconnaissance de profils biologiques.

3.3. Classifications des matériels biologique et clinique après discrétisation des variables numériques, pour différents codages.

3.4. Conclusion.

Compte tenu de la multiplicité des traitements effectués ; nous ferons précéder chacune des analyses par ses paramètres qui définiront la nature et les dimensions du tableau des données, le matériel organisé et l'indice utilisé ; la formation de l'arbre des classifications étant à chaque fois assurée au moyen de A.V.L. Il s'agira en quelque sorte de fiches techniques locales.

Précisons qu'on se contentera ci-dessous de figurer uniquement les arbres condensés.

3.1. Classifications des matériels biologiques et clinique dans le cas du codage numérique

a - ORGANISATION DU MATERIEL BIOLOGIQUE.

ANALYSE [VNA] de paramètres

- Tableau de données variables numériques × individus, de dimension 16 × 117.
- Classification hiérarchique de l'ensemble des variables.
- Coefficient de corrélation entre variables numériques.
- Figure 2.

La classification des 16 paramètres biologiques initiaux montre une structure bien hiérarchisée, le niveau 4, où la statistique globale atteint son maximum correspond à une partition en 6 classes. Si l'on excepte les classes réduites à un seul élément, la statistique locale indique une cohérence significative pour chacun des noeuds où ces classes atteignent leur complétion.

L'ensemble traité apparaît donc particulièrement bien classifiable.

Plus précisément, la partition du niveau 10, met en évidence les quatre syndrômes principaux de la pathologie hépato-biliaire :

* la cytolyse indiquée par la réunion de la bilirubine à la classe très cohérente des deux transaminases TGP et TGO.

* le syndrôme inflammatoire, constitué par la réunion de la protidémie, variable neutre, avec deux variables plus pertinentes : gamma-globulinémie et vitesse de sédimentation.

* la cholestase, classe de forte cohésion correspondant à

quatre paramètres bien significatifs au sens de la statistique de neutralité : LAP, GGTP, 5'NUCL, PH.ALC.

* l'insuffisance hépato-cellulaire attachée à la perturbation des taux de ALBU, PROTHR, Kl BSP ; la variable ALPHA très neutre semble en conséquence très peu représentative de cette classe.

Parallèlement, l'isolement de LDHT confirme son rôle peu discriminant, voir, son manque d'intérêt dans le bilan fonctionnel hépatique. Ceci est d'ailleurs corroboré par la valeur de la statistique de neutralité calculée pour LDHT, elle atteint là sa valeur minimum.

Au contraire, l'isolement de FER doit être rattaché non pas à un caractère peu discriminant, mais, au comportement tout particulier de la sidérémie dans les hémochromatoses.

* Extension en cas de données manquantes :

Nous nous limiterons à signaler la stabilité des résultats lorsque l'analyse s'effectue sur le fichier (16 × 228) en utilisant les indices généralises en cas de données manquantes (cf. [3]).

Si les quelques variables les plus neutres peuvent éventuellement permuter, les noyaux de forte cohésion relatifs aux quatre syndrômes et à l'hypersidérémie restent inchangés.

* Intérêt des nouveaux enzymes :

Rappelons que quatre parmi les seize paramètres biologiques envisagés, ne font pas partie, à proprement parler du bilan de routine. Il s'agit de LDHT, LAP, 5'NUCL, GGTP.

Nous avons déjà conclu en ce qui concerne LDHT. Par contre, si LAP, 5'NUCL et GGTP semblent avoir un rôle redondant du côté de LAP, comme témoins cholestatiques, elles apparaissent dans notre analyse comme plus discriminantes que cette dernière. Leur intérêt n'est donc pas négligeable. Enfin, nous aurons l'occasion de mettre par la suite en évidence, le rôle à la fois prépondérant et nuancé de GGTP dans le dépistage et la reconnaissance d'affections hépato-biliaires variées.

b - ORGANISATION DU MATERIEL CLINIQUE.

1 - ANALYSE [MNA] de paramètres

- Tableau de données variables numériques × individus, de dimension 16 × 117.
- Classification hiérarchique de l'ensemble des individus.
- Coefficient de corrélation entre individus.
- Figure 3.

Une fois encore, les résultats de la statistique globale permettent de distinguer clairement les niveaux les plus significatifs d'une arborescence nettement hiérarchisée.

En allant du plus général au plus détaillé, on distingue, au niveau le plus élevé une partition en quatre classes correspondant à quatre grandes classes pathologiques :

(A) - CIRRHOSES
(B) - CHOLESTASES
(C) - HEMOCHROMATOSES ET NORMACITE
(D) - HEPATITES

Les niveaux significatifs directement inférieurs permettent une analyse plus fine et en particulier, distinguent :

- un gradient de gravité à quatre degrés (A1, A2, A3, A4) à l'intérieur des cirrhoses, allant des cirrhoses avec décompensation ascitique jusqu'aux hépatopathies cirrhogènes bénignes et granulomatoses. Les cirrhoses peu évoluées sont assimilées aux hépatopathies alcooliques, que leur caractère cholestatique rapproche des cirrhoses suspectes de cancer.

- les hémochromatoses primitives (C2), des hémochromatoses secondaires (C1) et des individus sains (C3) à l'intérieur de la classe de quasi-normalité.

- les hépatites aïgues et prolongées des hépatites chroniques qui, de par leur caractère cirrhogène se retrouvent proches des cirrhoses.

On notera par contre, la non discrimination des cholestases intra ou extra hépatiques. Les cancers du foie étant à ce point intimement mélangés avec les lithiases biliaires (B).

La complétion de chacune des sous-classes analysées précédemment correspond de façon précise aux maxima locaux de la statistique locale.

Dans le schéma suivant (figure 3) condensé de l'arbre des classifications, chaque branche correspond à plusieurs sujets.

Nous avons effectué une analyse (MNB) de mêmes paramètres que l'analyse (MNA) à cela près que les mesures ont été au préalable centrées réduites par rapport aux distributions respectives des différentes variables sur l'échantillon étudié.

A des différences de détail près, les résultats de l'analyse précédente se trouvent confirmés. Il semble donc que la disparité d'échelle et de variance des différentes variables ne soit pas ici un inconvénient majeur pour la distinction des classes ; il peut l'être pour certains types de données.

2 - ANALYSE (MNC) de paramètres :

- Tableau de données variables numériques × individus, de dimension 16 × 117.
- Classification hiérarchique de l'ensemble des sujets.
- Cosinus de l'angle des deux vecteurs représentant les deux individus à comparer.
- Figure 4.

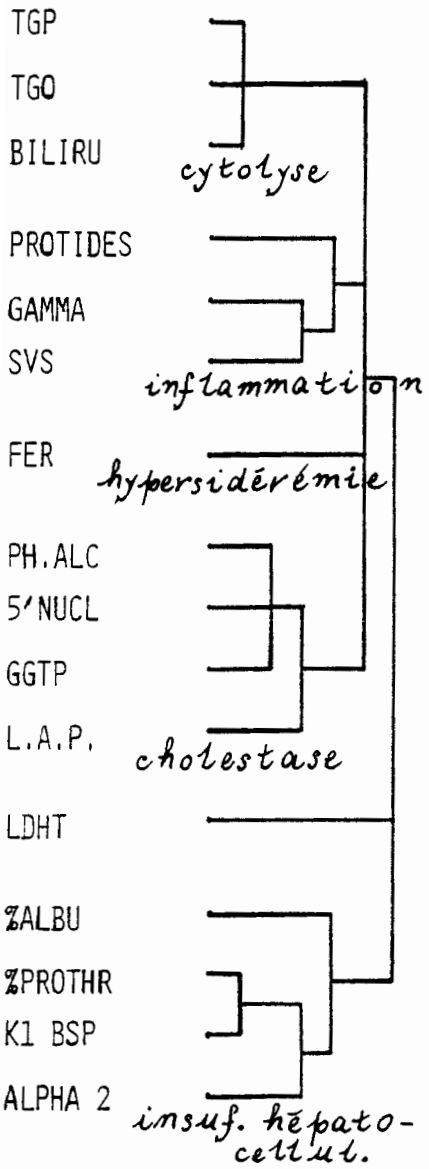


Figure 2.
analyse (VNA)

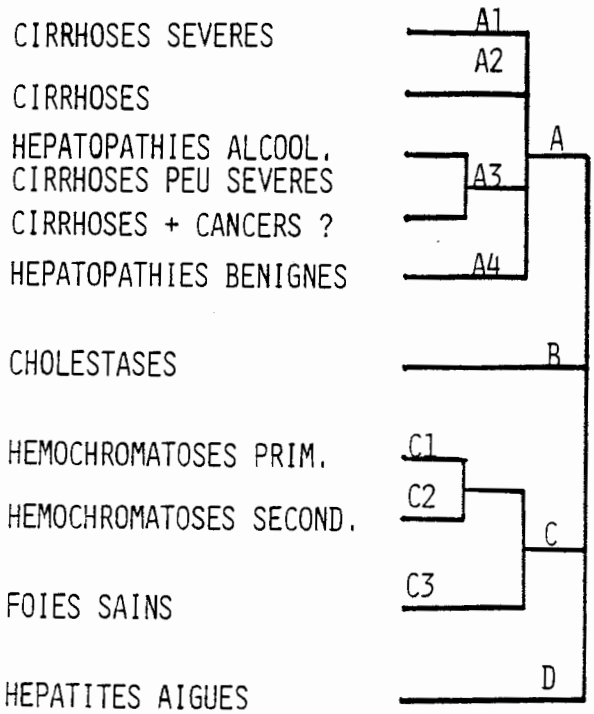


Figure 3.
analyse (MNA)

L'utilisation du cosinus de deux vecteurs au lieu du coëfficient de corrélation usuel pour mesurer la similarité, si elle laisse inchangée la structure générale du matériel clinique, modifie notablement l'optique à travers laquelle celle-ci est perçue.

Une telle notion de similarité conduit à un arbre condensé nettement moins hiérarchisé que les précédents.

Au niveau le plus significatif au sens de la statistique globale, on reconnaît toutefois les classes pathologiques habituelles, et, si l'analyse semble avoir perdu en globalité, elle paraît avoir gagné en finesse, correspondant à une optique classificatoire plus locale.

Ainsi, maintenant, reconnaît-on :

- les hémochromatoses en tant que classe pathologique distincte des individus sains.

- les hépatites aïgues comme séparées des hépatites prolongées.

- les cirrhoses avec ictères et ascite distincts des cirrhoses non décompensées.

Mais aussi, et ce qui est particulièrement nouveau et intéressant, les deux types de cholestases : intra ou extra-hépatiques jusque là non discriminés, se trouvent ici nettement reconnus.

D'une manière générale, il semble que dans cet exemple, le caractère cholestatique joue un rôle central autour duquel s'organise les autres aspects de la pathologie.

Plus précisément, on peut conclure que l'aspect cholestatique de certaines hémochromatoses a prévalu sur l'hypersidérémie rapprochant celle-ci de la classe des cancers secondaires en début d'évolution.

De la même façon, les cirrhoses ayant une cholestase associée se sont vus éloignées des autres cirrhoses du fait de cette teinte cholestatique.

Enfin, l'accent est mis sur une typologie des cholestases d'une façon analogue à celle qui faisait apparaître dans l'analyse [MNA] un gradient des cirrhoses.

Un mode d'interprétation de ce phénomène pourrait être le suivant :

dans l'analyse [MNA], du fait du caractère fortement cirrhotique de l'échantillon, tout centrage s'effectue par rapport à un profil moyen cirrhotique. Ainsi, les cirrhoses apparaîssent-elles comme plus centrales et les différents syndromes associés sont appréhendés comme des nuances du schéma cirrhotique.

A l'inverse, si aucun centrage n'est réalisé, la cirrhose apparaît comme fortement pathologique et les différents syndrômes associés s'effacent devant cette caractéristique principale.

Les nuances apparaissent alors dans les autres classes moins "excentriques", telles, ici, les différentes formes de cholestases.

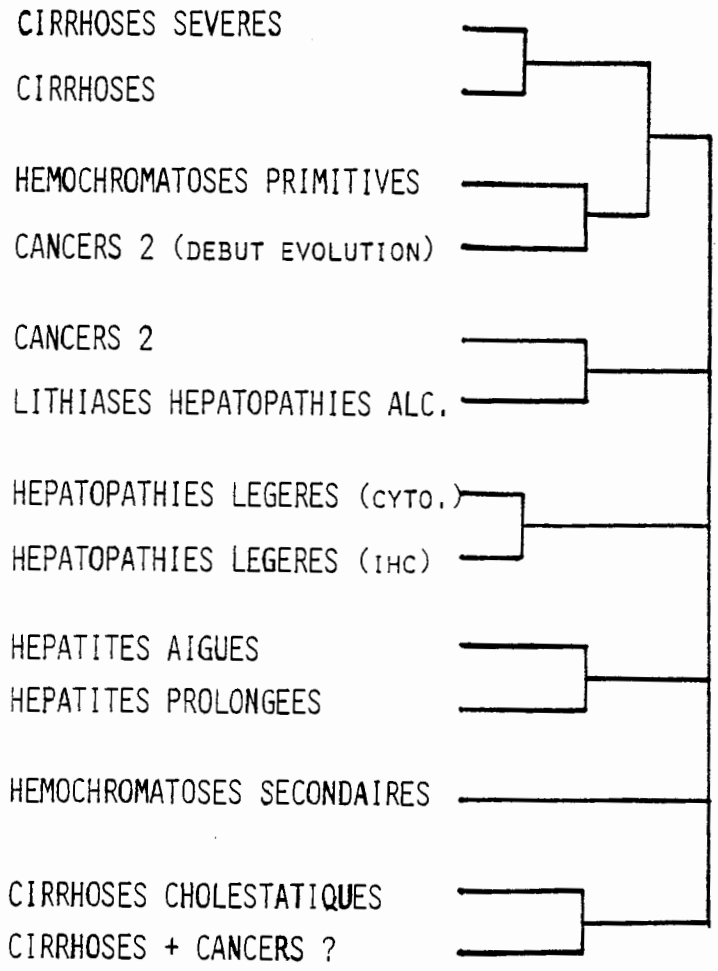


Figure 4
analyse (MNG)

c - ANALYSE FACTORIELLE DES CORRESPONDANCES

Une analyse factorielle des correspondances a été menée sur le tableau initial complet de dimension 16 × 117 par l'INSERM (U49-RENNES) dans le cadre d'une thèse de doctorat en médecine (cf.[1]).

Les résultats obtenus alors, sont dans leur ensemble en accord avec ceux que nous venons d'exposer. Plutôt que de représenter ici les différents plans factoriels, nous renvoyons au texte de la thèse citée en référence.

3.2. Liaisons entre syndromes et classes pathologiques ; profils biologiques

Jusqu'ici, l'interprétation des différents arbres s'est bornée à reconnaître dans le matériel biologique d'une part et dans le matériel clinique d'autre part, une organisation cohérente avec la toxinomie clinique expérimentale.

Le but du présent paragraphe est de chercher à apprendre, de la seule considération des données, les rapports qui existent entre les différents syndrômes et les différentes classes pathologiques mis en évidence lors des analyses séparées de l'ensemble des variables et de celui des malades.

Notre démarche sera la suivante. Pour chacune des classes pathologiques pertinentes exhibées dans l'analyse [MNA] nous calculerons un profil moyen. C'est-à-dire, la suite des valeurs moyennes des différents paramètres biologiques sur la classe considérée :

Soit "C" une partie de l'ensemble E des individus, le profil $\mathcal{P}_c$ de la classe C est la suite $\mathcal{P}_c = \{ \mu_v(C) / v \in V \}$ où v décrit l'ensemble des variables et où $\mu_v(C)$ représente la quantité :

$$\mu_v(C) = \frac{1}{\text{card}(C)} \sum_{x \in C} v(x)$$

Le profil ainsi obtenu peut être comparé à un profil type "sain" fourni par le médecin $\mathcal{P}_o = \{N(v) / v \in V\}$ où N (v) est la valeur théorique "normale" de la variable v de V chez un sujet non pathologique. Notons que ce profil $\mathcal{P}_o$ s'écarte assez fortement du profil moyen $\mathcal{P}_E$, ce qui tient à la nature de l'échantillon étudié.

En fait, pour pallier aux disparités d'échelles nous préférerons utiliser comme argument final de jugement un profil normalisé, constitué par la suite des rapports

$\{\mu_v (C) / N (v) \Big/ v \in V \}$ ainsi une valeur observée sera d'autant plus pathologique qu'elle s'écarte de la valeur "1" qui constitue alors la norme pour une quelconque variable.

Il reste néanmoins que, l'on devra conserver présent à l'esprit, lors de l'étude des profils normalisés, que chaque variable possède une sensibilité qui lui est propre et qu'il s'agit ainsi de comparer avec discernement les grandeurs des différents écarts observés.

Dans les tableaux suivants nous accompagnons la valeur $p(C)=\mu_v(C)/N(v)$ d'un intervalle $[p - \sigma_p, p + \sigma_p]$ que nous dirons de "confiance" où σ_p^2 et la variance de p estimée au niveau de la classe C.

PROFIL DE CIRRHOSES (Classe A)

La considération du profil normalisé relatif à la classe des cirrhoses montre une perturbation quasi générale de tous les témoins de l'activité enzymatique.

Si l'on raisonne en termes de syndrômes tels qu'ils ont été reconnus dans l'analyse [VNA], on note donc :

- une cytolyse moyenne compte tenu de la grande sensibilité des deux transaminases.

- un syndrôme inflammatoire discret. L'élévation de GAMMA pouvant en outre être liée à la présence de cirrhoses disglobulinémiques.

- une légère cholestase liée à l'augmentation de LAP et PH. ALC plus que de 5'NUCL.

- une imprégnation alcoolique (GGTP).

- une grande insuffisance hépato cellulaire, à laquelle les autres syndrômes sont associés.

Profil des sous classes :

A1 : ici, l'insuffisance hépato-cellulaire est très accentuée, indice de cirrhoses fortement décompensées.

Les syndrômes associés sont eux aussi naturellement aggravés. Notons que la 5'NUCL reste cependant quasi normale par rapport à la hausse de LAP, PH. ALC, et surtout GGTP.

A2 : on trouve ici, à un degré moindre, la persistance de l'I.H.C., tandis que les paramètres témoignant des autres syndrômes voisinent la normalité. Ceci confirme la notion de gradient associée aux niveaux inférieurs de l'arbre.

L'inflation de GGTP dans un contexte peu cholestatique confirme sa sensibilité à l'imprégnation alcoolique.

A3 : on reconnaît un profil propre à l'I.H.C., dans un contexte cholestatique et inflammatoire présent dans les cirrhoses suspectes de cancer.

A4 : le profil atteint ici un degré très peu pathologique où une très légère IHC est associée à une cytolyse. Cette sous classe correspond à des hépatopathies cirrhogènes bénignes ou à des hépatites chroniques.

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$
(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	57.03	1.58	2.21
TGO	40	82.36	1.86	2.25
BILIRU	10	19.69	1.65	2.28
PROTIDES	75	70.90	0.93	0.95
GAMMA	18	35.29	1.86	2.05
SVS	85	246.78	2.75	3.05
FER	130	100.09	0.72	0.81
PH. ALC	5	9.77	1.75	2.15
5'NUCL	15	17.72	0.98	1.37
GGTP	23	133.85	5.13	6.50
L.A.P.	22	27.48	1.17	1.32
LDHT	195	237.14	1.14	1.29
% ALBU	55	35.63	0.62	0.66
% PROTHR	100	63.90	0.61	0.67
K1 BSP	14	5.85	0.38	0.45
ALPHA	9	10.25	1.08	1.19

Profil des cirrhoses classe (A)

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$

(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	47.58	1.34	1.83
TGO	40	100.50	2.03	2.99
BILIRU	10	28.75	2.16	3.59
PROTIDES	75	72.58	0.95	0.98
GAMMA	18	37.94	1.91	2.31
SVS	85	231.75	2.44	3.01
FER	130	105.00	0.69	0.93
PH.ALC	5	16.43	2.65	3.93
5'NUCL	15	17.37	0.86	1.46
GGTP	23	261.41	9.80	12.93
L.A.P.	22	32.58	1.29	1.67
LDHT	195	215.50	1.00	1.21
% ALBU	55	35.16	0.59	0.69
% PROTHR	100	53.66	0.49	0.58
K1 BSP	14	4.71	0.28	0.39
ALPHA	9	10.14	1.04	1.21

Profil des cirrhoses décompensées (classe A1)

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$
(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	39.00	1.10	1.50
TGO	40	63.14	1.38	1.77
BILIRU	10	11.42	0.69	1.59
PROTIDES	75	73.42	0.95	1.00
GAMMA	18	33.62	1.67	2.07
SVS	85	341.42	3.78	4.25
FER	130	99.64	0.67	0.86
PH. ALC	5	6.75	1.16	1.54
5'NUCL	15	16.28	0.73	1.43
GGTP	23	119.00	4.28	6.06
L.A.P.	22	27.57	1.12	1.39
LDHT	195	182.14	0.87	1.00
% ALBU	55	34.92	0.59	0.67
% PROTHR	100	66.92	0.62	0.71
K1 BSP	14	6.75	0.41	0.56
ALPHA	9	10.89	1.06	1.36

Profil des cirrhoses sévères sans syndromes associés (classe A2)

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$

(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	36.85	1.03	1.42
TGO	40	69.25	1.46	2.00
BILIRU	10	17.15	1.22	2.21
PROTIDES	75	68.55	0.89	0.94
GAMMA	18	36.71	1.91	2.17
SVS	85	216.55	2.34	2.76
FER	130	94.75	0.66	0.80
PH.ALC	5	8.59	1.49	1.95
5'NUCL	15	19.14	0.86	1.69
GGTP	23	78.60	2.64	4.20
L.A.P.	22	24.63	0.99	1.25
LDHT	195	308.55	1.45	1.71
% ALBU	55	34.45	0.60	0.65
% PROTHR	100	65.25	0.60	0.70
K1 BSP	14	5.62	0.34	0.46
ALPHA	9	9.49	0.98	1.12

Profil des cirrhoses suspectes de cancer (classe A3)

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$
(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	183.83	4.16	8.09
TGO	40	87.33	1.57	2.79
BILIRU	10	14.83	0.85	2.10
PROTIDES	75	68.83	0.91	0.93
GAMMA	18	23.11	1.18	1.39
SVS	85	173.50	1.77	2.31
FER	130	96.66	0.69	0.80
PH. ALC	5	8.33	1.50	1.83
5'NUCL	15	17.08	0.83	1.45
GGTP	23	128.16	3.49	7.66
L.A.P.	22	26.66	1.02	1.40
LDHT	195	213.50	0.82	1.37
% ALBU	55	45.33	0.74	0.91
% PROTHR	100	86.66	0.83	0.90
K1 BSP	14	8.10	0.52	0.64
ALPHA	9	11.36	1.14	1.35

Profil des hépatites cirrhogènes et granulomatoses (classe A4)

PROFIL DES CHOLESTASES (classe B)

Cette classe réunissant des cancers du foie et des lithiases est nettement caractérisée comme son nom le laisse supposer par l'inflation des paramètres relatifs au syndrôme cholestatique.

La 5'NUCL y participe cette fois au même titre que LAP ou PH.ALC. Ceci laisse à penser que 5'NUCL est le témoin des cholestases pures plutôt que des cholestases associées.

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$
(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	119.00	2.60	5.34
TGO	40	139.92	2.67	4.33
BILIRU	10	20.07	1.39	2.63
PROTIDES	75	79.64	0.93	0.96
GAMMA	18	24.67	1.24	1.50
SVS	85	172.35	1.76	2.29
FER	130	146.42	0.98	1.27
PH. ALC	5	19.46	3.28	4.50
5'NUCL	15	42.27	2.25	3.38
GGTP	23	651.13	25.80	30.81
L.A.P.	22	46.44	1.76	2.45
LDHT	195	203.28	0.90	1.18
% ALBU	55	41.64	0.71	0.80
% PROTHR	100	75.50	0.71	0.80
K1 BSP	14	7.84	0.48	0.64
ALPHA	9	10.54	1.10	1.23

Profil des cholestases classe (B)

PROFIL DES HEMOCHROMATOSES (Classe C)

Profil très peu pathologique dans cette classe qui réunit hémochromatoses et individus sains.

- Profil des sous classes

C3 : le caractère non perturbé de l'ensemble des 16 paramètres témoigne d'individus biologiquement sains.

C2 : le même profil se retrouve, à l'exception de la sidérémie élevée. On reconnaît alors le profil caractéritisque des hémochromatoses primitives, par distinction avec les hémochromatoses secondaires réunis dans la classe suivante.

C1 : toujours la présence de l'hypersidérémie, jointe cette fois à une IHC notable et à une inflation de GGTP. L'hémochromatose est alors secondaire, sur une cirrhose ou une hépatopathie alcoolique.

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$
(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	50.47	1.50	1.86
TGO	40	67.69	1.54	1.85
BILIRU	10	9.91	0.77	1.20
PROTIDES	75	69.58	0.91	0.94
GAMMA	18	26.42	1.40	1.54
SVS	85	83.66	0.88	1.09
FER	130	175.00	1.27	1.44
PH. ALC	5	5.87	1.05	1.29
5'NUCL	15	8.70	0.51	0.65
GGTP	23	77.87	2.75	4.02
L.A.P.	22	21.58	0.91	1.05
LDHT	195	218.08	1.02	1.21
% ALBU	55	40.75	0.72	0.76
% PROTHR	100	74.41	0.71	0.78
K1 BSP	14	9.51	0.62	0.74
ALPHA	9	5.87	0.59	0.72

Profil sain ou hémochromatique classe (C)

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$

(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	57.64	1.54	2.30
TGO	40	60.92	1.27	1.78
BILIRU	10	7.50	0.58	0.91
PROTIDES	75	69.42	0.90	0.95
GAMMA	18	24.43	1.27	1.44
SVS	85	51.21	0.52	0.68
FER	130	126.78	0.83	1.12
PH.ALC	5	3.92	0.68	0.89
5'NUCL	15	6.92	0.40	0.52
GGTP	23	36.57	1.25	1.93
L.A.P.	22	21.14	0.84	1.08
LDHT	195	271.78	1.23	1.55
% ALBU	55	41.00	0.70	0.79
% PROTHR	100	80.50	0.75	0.86
K1 BSP	14	10.75	0.67	0.86
ALPHA	9	10.67	1.09	1.28

Profil sain ou hépatopathies bénignes classe (C1)

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$
(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	42.12	1.27	1.54
TGO	40	57.93	1.80	2.03
BILIRU	10	10.43	0.65	1.43
PROTIDES	75	70.06	0.92	0.95
GAMMA	18	26.19	1.34	1.57
SVS	85	121.93	1.29	1.59
FER	130	206.56	1.52	1.66
PH. ALC	5	6.79	1.18	1.53
5'NUCL	15	8.56	0.47	0.67
GGTP	23	55.03	2.06	2.73
L.A.P.	22	23.95	0.99	1.19
LDHT	195	153.75	0.71	0.87
% ALBU	55	41.06	0.71	0.78
% PROTHR	100	70.06	0.65	0.75
K1 BSP	14	9.77	0.60	0.80
ALPHA	9	10.48	1.08	1.24

Profil des hémochromatoses primitives (Classe C2)

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$

(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	56.00	1.45	2.28
TGO	40	109.50	2.22	3.25
BILIRU	10	14.16	0.80	2.02
PROTIDES	75	68.66	0.88	0.95
GAMMA	18	31.70	1.63	1.89
SVS	85	57.33	0.45	0.90
FER	130	208.33	1.50	1.71
PH.ALC	5	7.96	1.21	1.97
5'NUCL	15	13.25	0.66	1.10
GGTP	23	235.16	8.35	12.10
L.A.P.	22	16.33	0.65	0.84
LDHT	195	264.33	1.13	1.58
% ALBU	55	39.33	0.66	0.77
% PROTHR	100	71.83	0.65	0.79
K1 BSP	14	5.91	0.33	0.52
ALPHA	9	8.29	0.82	1.02

Classe des hémochromatoses secondaires (Classe C3)

PROFIL DES HEPATITES (Classe D)

On mesure ici la sensibilité des deux transaminases. L'élévation de TGP excède ici celle de TGO. Ce phénomène est caractéristique des hépatites aïgues, ainsi que l'est également la cholestase associée fréquemment à ce genre d'affection.

Un léger syndrôme inflammatoire indique la présence d'hépatites aïgues infectieuses.

PROFILS NORMALISES : $p = \mu_v(C)/N(v)$
(La Normalité correspond à la valeur "1")

Variable	Norm.	Moy.Obs	Intervalle de confiance	
			$p - \sigma_p$	$p + \sigma_p$
TGP	30	397.50	12.06	14.43
TGO	40	370.25	3.09	4.67
BILIRU	10	38.83	3.09	4.67
PROTIDES	75	72.08	0.93	0.99
GAMMA	18	29.78	1.53	1.78
SVS	85	141.00	1.46	2.07
FER	130	168.75	1.19	1.40
PH. ALC	5	12.52	2.05	2.96
5'NUCL	15	30.41	1.62	2.43
GGTP	23	130.75	4.81	6.56
L.A.P.	22	50.02	1.92	2.62
LDHT	195	222.58	0.96	1.32
% ALBU	55	36.91	0.64	0.70
% PROTHR	100	69.66	0.63	0.77
K1 BSP	14	4.46	0.24	0.39
ALPHA	9	9.42	0.97	1.12

Profil des hépatites (Classe D)

BILAN DE L'ANALYSE DU FICHIER NUMERIQUE

On reconnaît 4 syndrômes :

Le Syndrôme cytolytique lié à l'inflation des transaminases, TGP-TGO et éventuellement de la bilirubinémie dans certains ictères.

TGO et surtout TGP sont des témoins très sensibles de la cytolyse.

Dans les hépatites aïgues ou prolongées on observe TGP > TGO, au contraire dans les atteintes cirrhogènes (hépatites chroniques ou cytolyse associée à une grave insuffisance hépato-cellulaire), l'inégalité s'inverse : TGO > TGP.

Le syndrôme cholestatique nettement indiqué par une hausse de 4 enzymes PH.ALC, 5'NUCL, L.A.P., GGTP.

5'NUCL et LAP sont les plus sensibles après GGTP, les effets d'une imprégnation alcoolique peuvent appuyer ceux d'une cholestase pour contribuer à augmenter de façon très importante cet enzyme déjà excessivement sensible.

La cholestase peut être intra hépatique et l'on observe alors des syndrômes inflammatoires et cytolytiques associés dans les cancers du foie.

Elle peut aussi être extra hépatique (lithiase).

Le syndrôme d'insuffisance hépato-cellulaire apparaît dans les atteintes cirrhotiques et entraîne dans les cas les plus graves un disfonctionnement général et l'apparition de tous les autres syndrômes.

Les deux paramètres ALBU, PROTHR, indiquent à un même titre ce syndrôme ; KI BSP, moins sensible, ne semble pas cependant devoir être négligé.

Enfin, nous noterons l'inflation des gamma-globulines qui témoigne de la disglobulinémie des cirrhoses décompensées.

Les hépatopathies alcooliques sont des atteintes cirrhogènes présentant l'IHC, mais une fréquente cholestase associée les rapproche de certaines cholestases ou des cirrhoses suspectes de cancer.

Le syndrôme inflammatoire apparaît principalement comme syndrôme associé dans de nombreux cancers ainsi que dans les graves IHC. Il peut être cependant le résultat d'une attaque infectieuse sans aucun rapport avec la biologie hépatique.

ΣVS est un témoin très sensible ; joint à GAMMA il permet de différentier le syndrôme inflammatoire d'une disglobulinémie. L'anormalité de la protidémie totale peut confirmer ce jugement.

L'hypersidérémie apparaît liée sans ambiguité aux hémochromatoses. La considération de syndrômes associés permet de distinguer les atteintes primitives des atteintes secondaires.

Dans les atteintes primitives en particulier le profil moyen est, si l'on excepte la sidérémie, non pathologique.

Les variables LDHT et ALPHA sont très neutres et de peu d'intérêt.

3.3. Analyse des données discrétisées

a) DISCRETISATION DES VARIABLES NUMERIQUES

La transformation des variables numériques en variables

discrètes passe nécessairement par un algorithme de découpage de l'intervalle de variation d'un même paramètre en sous intervalles.

I.C. Lerman nous a suggéré d'analyser un type d'algorithme où on balayerait l'axe, support de la distribution de la variable sur l'échantillon étudié, d'un petit intervalle (fenêtre mobile) ; afin de déterminer l'évolution d'une sorte de densité discrète correspondante au nombre de points dans la fenêtre. Le repérage des minima locaux d'une telle distribution permettrait de déterminer les points de coupure.

Cette analyse qu'on verra développée dans [3] comprend une présentation formelle qui permet d'étudier ses propriété algébriques, la situation de cette démarche par rapport aux différentes méthodes de séparation d'un mélange de lois de probabilité à support unidimensionnelle, la définition d'une stratégie de détermination de quelques amplitudes de la fenêtre afin de repérer les minima locaux stables, l'application à des données d'une certaine dimension dans [2] et la programmation qu'une récente technique d'évaluation de la "profondeur" des minima locaux permet de rendre le procédé complètement automatique.

L'application de la méthode de la fenêtre mobile conduit à adopter le système de codage décrit dans le tableau suivant.

Celui-ci est constitué par 16 caractères à modalités exclusives. Le nombre de modalité par caractère variant de deux à cinq. Le cardinal de l'ensemble des modalités est de 61.

Bien que cela corresponde à une abérration du point de vue médical, nous supposerons, dans un but de simplification de l'écriture, que chaque paramètre varie a priori entre "0" et "+ ∞" ; ainsi le découpage correspond à une subdivision de l'intervalle semi-ouvert $[0,+\infty[$

Le découpage retenu, tel qu'il va être présenté par la suite s'avère, de l'avis du médecin, respecter de façon satisfaisante la notion de valeur "normale", "perturbée", "très perturbée" et être en conséquence en accord avec la structure de la population étudiée.

La prise en compte d'un tel système d'intervalles garantit donc une perte minime d'information par rapport aux données initiales.

Ainsi, à une même variable numérique on peut associer soit un caractère à l'ensemble totalement ordonné des modalités, où une même modalité est associée à un même intervalle de la subdivision ; soit un ensemble d'attributs respectivement associés aux différentes modalités exclusives et exhaustives du caractère.

Résultat du découpage :

Codage des variables numériques par des caractères à modalités exclusives

MODALITES	1	2	3	4	5	
CARACTERES	←→	←→	←→	←→	←→	
TGP	0	29,50	55,00	78,00	+ ∞	
TGO	0	80,00	113,00	146,00	+ ∞	
LDHT	0	143,00	204,00	376,00	+ ∞	
FER	0	46,00	121,00	196,00	+ ∞	
PROTIDE	0	67,50	+ ∞			
% ALBU	0	32,50	45,50	+ ∞		
% PROTHR	0	41,00	53,00	81,00	+ ∞	
BILIRU	0	21,00	46,00	+ ∞		
PH. ALC	0	6,40	11,70	+ ∞		
S'NUCL	0	14,50	47,00	+ ∞		
L.A.P.	0	13,50	22,50	28,50	+ ∞	
GGTP	0	39,00	122,00	225,00	325,00	+ ∞
ALPHA	0	6,10	9,10	11,10	12,10	+ ∞
GAMMA	0	18,50	30,10	41,00	+ ∞	
Σ VS	0	25,00	69,00	175,00	241,00	+ ∞
KI BSP	0	2,60	7,20	12,00	+ ∞	

NB : Les bornes "0" et " ∞ " n'ont évidemment aucune signification médicale mais remplacent les extréma cliniquement observables et compatibles avec la vie.

: Les différentes modalités résultant du découpage seront codées de façon naturelle : "1TGO, 2TGO, 3TGO, 1TGP, 2TGP, etc...

b - ORGANISATION MATERIEL BIOLOGIQUE

1 -ANALYSE [VDA] de paramètres :

- tableau de données variables préordinales × individus, de dimension 16 × 117
- classification hiérarchique de l'ensemble des variables
- indice de proximité entre préordres totaux, mentionné dans la fiche technique
- figure 5.

La structure dégagée sur le matériel biologique représenté par les 16 paramètres initiaux, se trouve en tout point confirmée par les analyses menées à partir de la représentation par 16 variables préordinales.

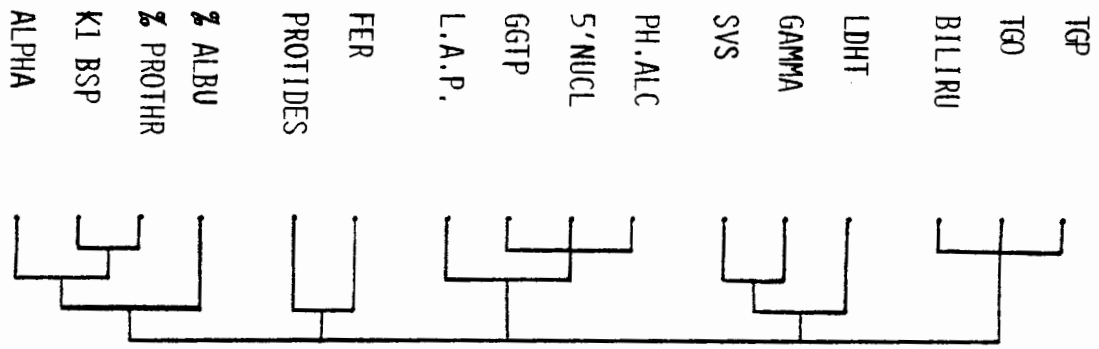
Ce résultat obtenu présente en outre deux intérêts majeurs:

D'une part, l'identité des deux arborescences obtenues permet de valider la segmentation des variables numériques.

L'affaiblissement de la richesse descriptive initiale n'a pas entamé de façon notable la part d'information réellement signifiante quant au diagnostic et aux relations mutuelles entre syndrômes.

Le système de modalités retenu pour chaque caractère est donc en accord avec les différentes modalités de réponse propres aux variables mesurées, devant les divers degrés et formes de la pathologie.

Parallèlement, se trouve ici confirmé le fait que, à juste titre la démarche diagnostique, néglige la précision effective de la mesure pour n'en considérer que l'ordre de grandeur.



Classification du matériel biologique
analyse : (VDA)
Figure 5

2 - ANALYSE [V D B] de paramètres

- Tableau d'incidence des données (Attributs ou Modalités) × Individus, de dimension 61 × 117
- Classification hiérarchique de l'ensemble des attributs
- Indice de proximité entre attributs de description mentionné dans la fiche technique
- Figure 6.

Le découpage des variables numériques et la considération de chaque modalité résultante comme un attribut descriptif en faisant abstraction de toute relation d'ordre, permet d'aborder l'analyse de la structure du matériel biologique sous un angle tout à fait neuf.

L'optique la plus globale, liée à l'étude de la partition relative au niveau significatif le plus élevé dans la hiérarchie, reconnaît un certain nombre de classes qui peuvent être analysées comme suit :

- une classe homogène correspondant aux modalités témoignant de la normalité de l'activité enzymatique.

- un ensemble bien hiérarchisé d'attributs fortement pathologiques. Cette organisation hiérarchique témoigne du lien intime existant entre les différents syndrômes dans les atteintes hépatiques sévères.

On reconnaît donc ici une classe de syndrômes dits "associés", dans les grandes insuffisances hépato cellulaires où les fortes cytolyses etc...

- par opposition, la dernière partie de l'arbre, correspond à une juxtaposition de formes moyennes de la pathologie. Cette représentation non hiérarchique témoigne de la distinction entre les quatre syndrômes initiaux sous leur forme "pure" : cholestase pure, cytolyse typique etc...

D'une façon plus détaillée, portant l'analyse à un degré moins élevé de l'arbre, on se doit de noter principalement :

- une typologie de la normalité où se trouvent réunis, à l'intérieur d'une classe à dominante non pathologique les différentes formes, classées selon le caractère non pathologique auquel elles s'opposent :

. non cytolytique,
. non cholestatique,
. non I.H.C.
. sidérémie normale

Enfin, l'hypersidérémie, en raison du caractère très particulier des hémochromatoses, se trouve encore une fois attachée à la normalité.

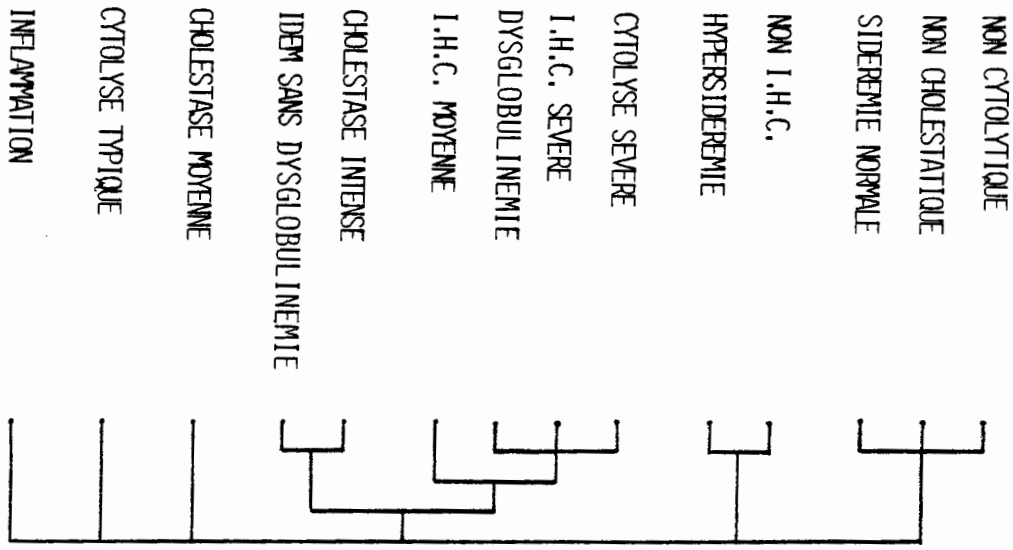
Cette typologie de la normalité, occultée par le caractère pathologique de l'échantillon dans les analyses précédentes peut, avec le présent codage, se révéler.

- Le syndrôme d'Insuffisance hépato-cellulaire est absent de l'ensemble des syndrômes isolés sous leur forme pure.

En fait, il sert de centre à l'organisation de la classe des syndrômes associés. Ceci s'explique naturellement à la fois par la composition de l'échantillon et par les graves désordres de l'ensemble de la fonction hépatique dans les cirrhoses décompensées.

Dans le dessin suivant de l'arbre chaque branche correspond à un paquet d'attributs de description.

Les résultats précédents se trouvent conservés quelle que soit l'hypothèse de lien choisie (modèle Hypergéométrique - modèle Binomial) (cf. chap.2 § IV.0 et 1).



Classification du matériel biologique
analyse : (VDB)
Figure 6

c - ORGANISATION DU MATERIEL CLINIQUE

Dans la totalité des analyses précédentes la considération de l'allure de la statistique globale ainsi que de sa cohérence avec la statistique locale, a permis de juger sans ambiguïté de la qualité des résultats et de la classificabilité de l'ensemble.

Une situation satisfaisante correspondant à une statistique globale présentant une suite régulière de maxima locaux bien individualisés le dernier étant attaché à un niveau élevé de l'arbre. Chaque maxima local, définit un niveau ou le noeud correspondant est significatif au sens de la statistique locale.

1 - ANALYSE [MDA] de paramètres :

- Tableau d'incidence Attributs × Individus, de dimension 61 × 117

- Classification hiérarchique de l'ensemble des individus
- Indices de proximité entre individus pour une description par des attributs, mentionnés dans la fiche technique.

Les résultats ont été ici tout à fait insatisfaisants pour un niveau assez global de synthèse ; d'ailleurs le comportement de la statistique globale le montre de façon notoire puisqu'elle atteint son maximum général à un niveau relativement bas de l'arbre.

En introduisant, pour l'établissement d'un indice de proximité entre individus, une hypothèse d'absence de lien qui tient compte de la fréquence relative des différents attributs sur l'échantillon étudié, on "repousse" à un niveau plus élevé le maximum général de la statistique globale ; ce qui "améliore" l'interprétation des résultats. Cependant, les classes restent d'effectif très réduit et si les liaisons y sont interprétables, elles se situent à un niveau trop bas de l'arbre pour permettre d'aboutir à une synthèse qui se comprend à partir de la notion de diagnostic.

On a ici beaucoup trop appauvri la richesse de la description pour pouvoir encore, relativement à la classification de l'ensemble des individus, obtenir une vue synthétique du matériel clinique. On peut certes reconnaître la constitution cohérente de groupes de très petits effectifs formés d'individus très voisins ; mais les liens plus faibles sont perdus.

Nous n'insisterons donc pas sur ce traitement pour nous attacher à la classification sur l'ensemble des malades obtenue à partir d'une description des sujets par des caractères aux modalités totalement ordonnées (i.e. variables qualitatives ordinales).

2 - ANALYSE [MDB] de paramètres

- tableau de données Variables préordinales × Individus, de dimension 16 × 117
- Classification hiérarchique de l'ensemble des individus.

- Indice de proximité entre individus décrits au moyen de variables préordinales ou caractères aux modalités totalement ordonnées ; indice mentionné dans la fiche technique.
- figure 7.

Ici l'appauvrissement de l'information prise en compte lors du codage des individus n'altère aucunement la qualité de la représentation des données.

L'organisation générale corrobore les analyses précédentes et l'on retrouve les quatre principales classes pathologiques :

- cirrhoses
- normalité/hémochromatoses
- lithiases/cancers
- hépatites

Plutôt que d'insister sur ce fait, nous allons examiner les acquis nouveaux apportés par cette forme nouvelle de codage et de similarité.

- La classe des cirrhoses se trouve décomposée en deux sous-classes selon la présence ou non d'un syndrôme cytolytique associé. Cette cytolyse témoignée par une élévation particulièrement sensible de TGO plutôt que TGP éclaire le rapprochement de cette classe de nette IHC avec la classe des hépatites, au niveau supérieur.

Ces hépatites en général chroniques se trouvent en outre distinguées des hépatites aïgues et prolongées elles-mêmes regroupées ailleurs. Elles présentent en outre une hypersidérémie absente chez ces dernières.

- La classe de normalité regroupe avec quelques sujets effectivement sains différents types d'hémochromatoses?

La distinction se fait suivant deux axes : présence ou non de disglobulinémie d'une part, d'une légère cholestase de l'autre.

- Le groupe cholestatique s'organise suivant un gradient de gravité ; on retrouve ainsi successivement :

. des cancers du foie non discriminés des lithiases de la voie biliaire

. des atteintes de même type qui, bien que moins évoluées présentent un net syndrôme inflammatoire.

. diverses atteintes cholestatiques telles que des hépatopathies alcooliques ayant une forte élévation de GGTP

. des individus sains du point de vue hépatique mais présentant tous un syndrôme inflammatoire biologique, ce qui explique leur distinction d'avec les individus réellement sains et proches des hémochromatoses.

- la classe des hépatites aïgues ou prolongées présentant un syndrôme cytolytique pur sans IHC et sans hypersidérémie ce qui les distingue des autres hépatites rencontrées plus haut.

Dans le dessin schématisé suivant de l'arbre, chaque branche soustend un groupe de sujets.

3.4. Conclusion

Médical

La série d'analyses dont nous avons ici détaillé certaines, a permis de répondre de façon précise aux questions initiales du médecin et ce sur les points suivants :

- organisation du matériel biologique
- organisation du matériel clinique
- rôle des différents paramètres biologiques
- intérêt des nouveaux enzymes
- validité de la notion de profils biologiques
- limites au diagnostic biologique.

Mesure de la similarité, interprétabilité

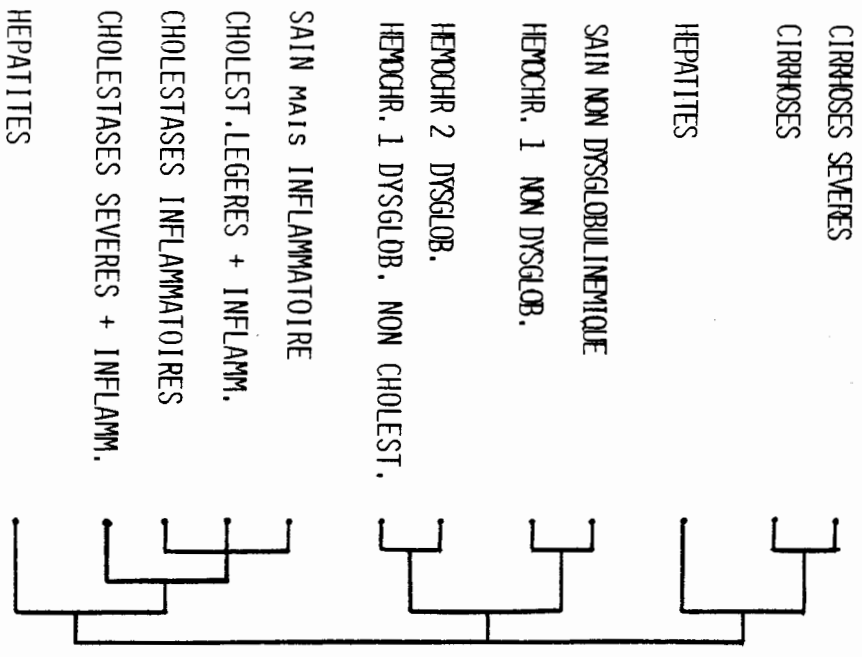
L'étude permet de conclure à l'efficacité de l'approche classificatoire selon I.C.Lerman dans le domaine médical et en particulier lors de la recherche d'une structure sur l'ensemble des individus.

Il s'est avéré que dans tous les cas, l'allure de la statistique globale de pertinence des niveaux de la hiérarchie de partitions, est un indice fidèle de pertinence et d'interprétabilité des résultats.

Nous avons été amenés à adapter et à construire différents indices de similarité, mesurés sous divers hypothèses d'absence de lien.

Les hypothèses d'absence de lien faisant référence à la répartition observée à travers l'échantillon analysé ont toujours été préférables à celles stipulant l'absence de lien comme une répartition uniforme.

L'analyse du fichier des variables numériques conduit à des résultats globaux. Une optique de détail, en particulier pour l'étude spécifique de certains paramètres nécessite l'étude du fichier des variables discrétisées.



Classification du matériel clinique
analyse : (MDB)
Figure 7

BIBLIOGRAPHIE

[1] Y.HITA DE NERCY (1975)
"Exploration fonctionnelle hépatique, questions soulevées par l'analyse des données"
Thèse de doctorat d'état, Université de Rennes.

[2] A.M. KERJAN née SALMON (1978)
"Tentative d'établissement de 100 typologie d'examens biologiques. Contribution à l'établissement du système "A.D.M."
Thèse de doctorat d'état, Université de Rennes 1.

[3] J.Y. LAFAYE (1978)
"Les différentes formes de l'appréhension des données dans l'exploration fonctionnelle hépatique ; discrétisation de variables numériques. Recherche de profils biologiques par une méthode de classification hiérarchique".
Rapport de D.E.A. (1977) et Thèse de doctorat 3ème cycle.
Université de Rennes 1 (1978).

[4] B. CYFFERS (1965)
"Analyse discriminatoire"
Revue Statistique Appliquée , Volume XIII n° 2 pages 29-46.

[5] Y. LECHEVALLIER (1974)
"Optimisation de quelques critères en classification automatique et application à l'étude des modifications des protéines sériques en pathologie clinique".
Thèse de doctorat de 3ème cycle. Université Paris VI.