Bibliométrie et chimie. Exemple sur les acides gras et phospholipides

L Quoniam*, H Dou*, P Hassanaly*, G Mille **

Des essais d'analyse systématique d'informations scientifiques ont déjà été entrepris [1-5]. Ces analyses sont intéressantes pour mettre en évidence des méthodologies par analyse statistique des présences de mots au sein d'une interrogation de banques de données. Ces méthodes présentent l'avantage de dégager automatiquement et systématiquement les corrélations (positives ou négatives) entre mots (donc sujets) abordés dans les références. Cependant, plusieurs problèmes sont posés par cette approche.

La procédure de saisie d'une interrogation de banques de données est automatique. Par contre, il n'en est pas de même pour le codage du contenu des réponses afin de réaliser une étude statistique des résultats de l'interrogation. Or ce type de traitement ne présente d'intérêt que si tout le processus est automatique.

Nous travaillons au laboratoire à la résolution d'une partie de ces obstacles. En effet, il serait intéressant à plusieurs titres de pouvoir recourir de façon plus systématique à ce type d'analyse :

– les coûts d'interrogation étant élevés, il apparaît indispensable de pouvoir utiliser et cerner davantage les résultats d'une interrogation, tant sur son contenu strict que sur les perspectives mises en évidence;

– les interrogations apparaissent souvent comme une liste plus ou moins homogène de références. Certains travaux ont un niveau informatif plus important que d'autres.

Or, par le type d'analyse que nous proposons, il est possible de mettre en évidence les hétérogénéités afin d'aider le chercheur à analyser avec le plus de pertinence possible le sujet tout en le replaçant dans un contexte plus global.

Le CRRM (Centre de recherche rétrospective de Marseille) développe un logiciel de mise au format (programme Textrans) [6] et d'extraction de champs des références obtenues lors de l'interrogation de la banque de données (programme Datalink) [7, 8].

De plus, un logiciel de visualisation graphique appelé Pole [9] permet de résumer les informations de certains champs en quelques graphes.

À ces programmes, conçus et élaborés au sein du laboratoire, nous avons adjoint un logiciel de traitement statistique (Statitcf) élaboré par l'Institut des techniques des céréales et des fourrages. Cependant, certaines adaptations sont nécessaires pour le rendre opérationnel et compatible avec le matériel et les programmes existants.

Structure d'une interrogation

Chaque référence obtenue par interrogation est découpée en rubriques ou champs. Le nombre et l'intitulé des champs désirés sont spécifiés à l'interrogation et structurent les réponses. Un exemple est présenté à la *fig 1*.

L'ensemble des exemples utilisés est extrait d'une interrogation de 500 titres concernant le dosage des acides gras et phospholipides par chromatographie dont l'historique est donné *fig 1*.

Une des fonctions de Datalink est de réaliser une mise au format d'un champ choisi, puis un codage (sous forme de *hash-coding*) des mots avec mémorisation des occurrences [8,9]. Ce *hash-coding* est en fait une conversion de mots en chiffres destinée à accélérer tris et recherches. Tous les éléments d'un champ, fonction du demandeur, peuvent être codés, exploités.

Dans un premier temps, nous nous sommes intéressés au champ section. Ce champ indique le classement des références effectué par le producteur, dans les différentes rubriques concernées (dans le cas du CA, 80 sections identifiées par un numéro). L'étude de ce champ va permettre de cerner le sujet sur lequel a porté l'interrogation, et aussi de le replacer dans un contexte élargi.

Cette étude a été entreprise de deux façons dans notre laboratoire.

Dans le champ code des CA *(fig 2)*, la disposition des sections à l'intérieur du champ est codifiée. La première section

```
SS 1:  ALL CHROMATOG AND ( ALL PHOSPHOLIPID#.OR ALL LIPID# OR ALL FATTY (W)
ACID# )  (1240)
SS 2:  1 AND BIOCHEM/FS  (972)
USER:
PRT 1-500 AN TI AU OS SO CC ST

-1-
AN  - CA05-183117(20)
TI  - Large-scale separation of lipids from organochlorine pesticides and
      polychlorinated biphenyls using a polymeric high-performance liquid
      chromatographic column
AU  - Seymour, Mark P.; Jefferies, Terry M.; Notarianni, Lidia J.
OS  - Univ. Bath, Sch. Pharm. Pharmacol., Bath, UK, BA2 7AY
SO  - Analyst (London) (ANALAO), V 111 (10), p. 1203-5, 1986, ISSN 00032654
CC  - SEC80-4; SEC5; SEC48
ST  - organochlorine; pesticide; HPLC; HPLC; pesticide; chlorinated;
      biphenyl; lipid; liq; chromatog; pesticide; PCB; lipid; residue;
      analysis; cleanup; HPLC; polymeric; column; HPLC

où:
-2-   :est le n° du tiré à part dans l'interrogation.
AN  - :n° de registre du C.A.
TI  - :titre de l'article.
AU  - :les auteurs.
OS  - :adresse des auteurs.
SO  - :références de l'ouvrage.
CC  - :sections du C.A.
ST  - :mots supplémentaires attribués par un analyste du C.A. afin de mieux
       définir l'article.
```

Fig 1 - Historique et structure du résultat de l'interrogation

* Centre de recherche rétrospective de Marseille

** Centre de spectroscopie moléculaire, faculté des sciences de Saint-Jérôme, avenue Escadrille-Normandie-Niemen, 13397 Marseille Cedex 13. Tél : (16) 91 02 90 94. Fax : (16) 91 28 80 30.

Tableau I - Formalisme du tableau de départ

	Mot X			Mot Y				...
	État 1	État 2	État 3	État 1	État 2	État 3	État 4	
Tiré à Part 1								
Tiré à Part 2								
...								

Tableau II - Formalisme de la matrice obtenue

	s10	s17	s30	...
Tiré à Part 1	1	1	1	
Tiré à Part 2	1	2	3	
Tiré à Part 3	1	2	1	
...				

est appelée section principale (notée : PSECX), elle peut être suivie par une sous-section (le chifffre seul), et par une section secondaire (notée : Secy). Toutes les sections secondaires sont au même niveau. Le nombre des sous-sections est variable en fonction de la section de rattachement.

Ce champ extrait sera traité soit par les logiciels Datalink et Pole (développés au paragraphe C), soit par le logiciel STATITCF (développé au paragraphe D). Le travail accompli par les deux premiers programmes a été repris succinctement ici pour comparaison, mais a déjà été longuement développé.

```
CC    PSEC8Ø 4  SEC5  SEC48
CC    PSEC3Ø 2Ø  SEC17
CC    PSEC2Ø 4  SEC17
CC    PSEC17 1
CC    PSEC17 1
```

Fig 2 - Exemple de chap extrait par Datalink

Étude des fréquences de sections et leurs liaisons

Après extraction du champ section, trois traitements sont réalisés.

Les fréquences d'apparition des sections peuvent être éditées sous forme de liste par Datalink avec le choix d'une différenciation entre section principale ou secondaire. Mais une formulation plus explicite des fréquences de sections principales est proposée par Pole, soit sous forme bidimentionnelle, soit sous forme tridimentionnelle *(figs 3 et 4)*.

Après codage du champ extrait (réalisé pour augmenter la rapidité), le programme Datalink permet d'obtenir la liste de toutes les paires, avec en regard la fréquence d'apparition.

Le programme Pole élimine automatiquement les sections non liées. Un graphe est tracé donnant les liaisons des différentes sections entre elles. Le graphe général *(fig 5)* donne la totalité des liaisons.

La section placée au centre est toujours celle à la plus haute fréquence (pôle principal). Les sections périphériques sont représentées selon deux symboliques :

– désigne une section apparaissant comme section principale (ou principale et secondaire).

– désigne une section n'apparaissant que comme section secondaire.

Ces liaisons peuvent aider à cerner les interactions entre sujets de recherche.

Une idée de quantification de cette interaction peut être obtenue en interprétant simultanément ce graphe et la liste des paires : ainsi nous rendons-nous compte que la liaison section 9 section 10 (Biochemical methods Microbial biochemistry) a une fréquence de 42/500, elle sera donc plus importante que celle de la section 80 section 46 (Organic analytical chemistry Surface-active agents and detergents) qui n'a qu'une fréquence de 1/500.

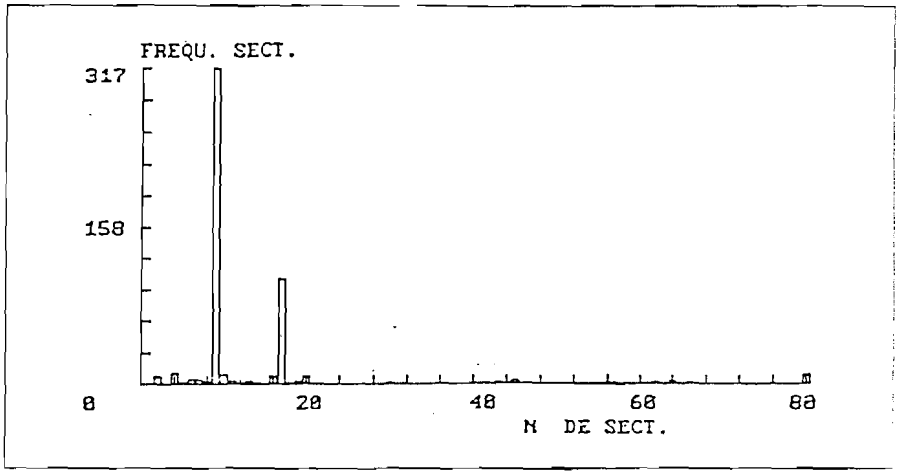

Fig 3 - Sections CAS de 500 références sur le thème chromatographie des acides gras et phospholipides

Une fois le graphe général édité, d'autres graphes peuvent être réalisés : par exemple *(fig 6)* l'environnement du pôle 17 (Food and feed chemistry). Nous voyons ainsi les sections principales et secondaires auxquelles est rattachée la section 17. La quantification de cette liaison se fera, comme précédemment, en regardant la fréquence d'apparition d'une paire.

Ce système de graphes, créé par Pole, puissant par la simplicité de son interprétation et la rapidité de sa mise en œuvre, nous a incités à développer davantage ce type d'interprétations, en tenant compte des inconvénients suivants :

– la non-prise en compte des sections non appariées. Ce traitement, nous l'avons vu, élimine les sections non appariées; donc, perte de l'importance des sections indépendantes dans l'interrogation;

– la perte de la trace des références. En effet, il est intéressant de pouvoir toujours retrouver la ou les références qui présentent une particularité mise en évidence par l'analyse;

– le non respect des distances entre sections dans l'édition des graphes. Ces graphes renseignent sur les liaisons entre sections, mais difficilement sur l'intensité de ces liaisons, leur proximité relative ou le nombre de paires formées par ces sections.

Recours à des méthodes statistiques

Nous avons réalisé un programme de codage de mots ajouté en option à Datalink capable de construire automatiquement une matrice de répartition en classes *(tableau II)*. Cette matrice est

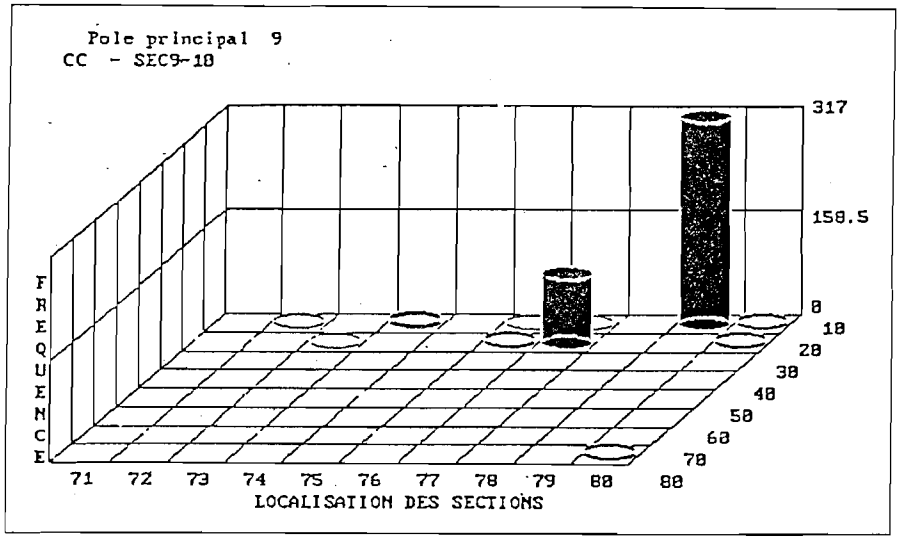

Fig 4 - Sections CAS de 500 références sur le thème chromatographie des acides gras et phospholipides

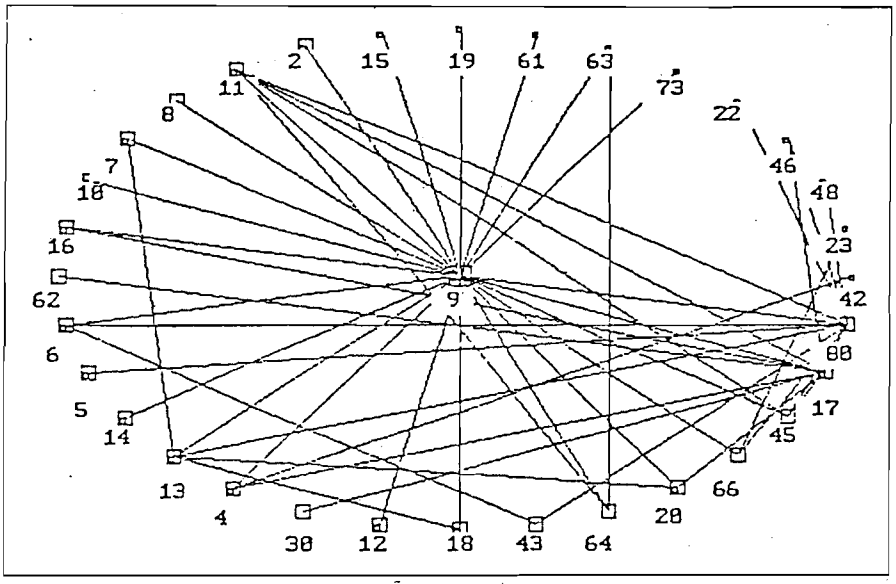

Fig 5 - Graphe général des liaisons entre sections

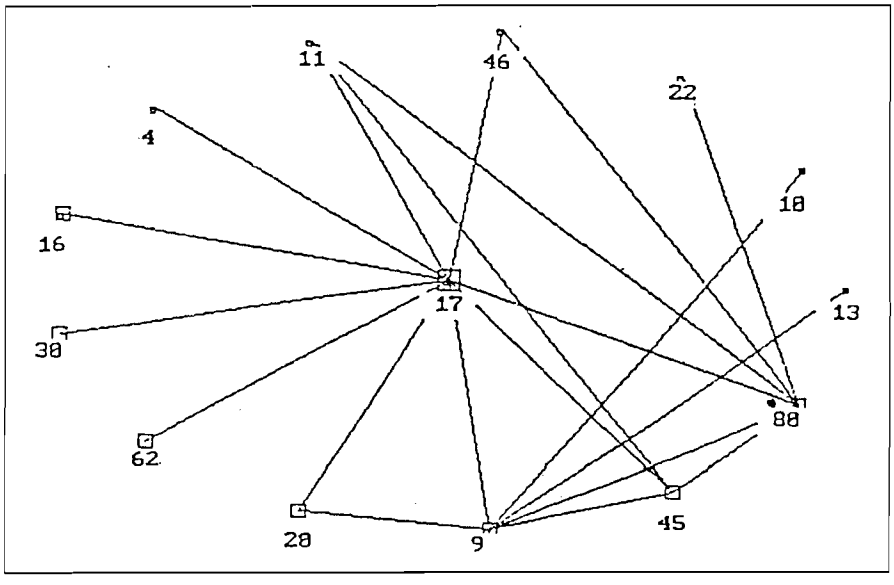

Fig 6 - Graphe des liens de la section CAS 17

élaborée à partir d'un champ codé par Datalink et revient au remplissage d'un tableau d'une variable à plusieurs états *(tableau I)*.

Ce programme travaille sur un champ extrait et codé par le programme Datalink. Il permet de choisir entre trois formes de codage différentes. La façon dont le codage (présence/absence) des mots est entrepris va avoir une grande incidence sur l'analyse statistique ultérieure. En effet, un codage sous forme 1 ou 0 peut générer des erreurs (division par 0 impossible); un codage sous forme 1 ou 2 pallie ce problème, mais reste sommaire, car n'offrant qu'une variation entre 1 et 2 donc une variance faible et un pouvoir discriminant faible; un codage allant de 1 à *n* est la solution idéale mais est difficile à réaliser dès que des répétitions apparaissent.

Une fois choisi le type de codage désiré apparaissent une série de questions quant à l'analyse envisagée. Puis le programme demande de rentrer la liste des mots dont les occurrences sont désirées ainsi que le numéro de code souhaité. Le programme établit la matrice directement transcrite dans des fichiers au format du logiciel de traitement des données STATITCF élaboré par l'Institut technique des céréales et des fourrages.

Cela nous a permis dans un premier temps une application sur le champ section où :

Mot X = section ...
État 1 de X= absence de la section ...
État 2 de X= présence de la section... secondaire
État 3 de X= présence de la section... principale

Le traitement proposé pour l'édition des listes de paires et des liaisons entre sections aide grandement pour le choix du codage à réaliser et de l'analyse à entreprendre. En fait, ces deux analyses sont complémentaires l'une de l'autre. La première, simple et rapide mais incomplète par certains aspects, aide à l'élaboration de la seconde, qui, si elle est assez difficile à mettre en œuvre, apporte le maximum d'informations.

Une fois le traitement effectué par Datalink, des analyses statistiques sont réalisables.

Obtention des paramètres statistiques de base

Les moyennes, écart-types, variances... peuvent fournir des indications sur la fréquence des états d'un mot dans une interrogation, ainsi que sur les différentes dispersions de ceux-ci.

Analyses factorielles

L'analyse factorielle des correspondances multiples (AFCM), mettant en évidence les très faibles fréquences, n'est pas adaptée en cas de grandes différences de fréquence d'apparition. Non présentée ici, une AFCM nous a

permis de voir qu'elle représentait essentiellement les références qui pouvaient être considérées comme « hors sujet » par rapport à l'interrogation.

Pour l'application choisie, l'analyse multivariée retenue est l'analyse factorielle des correspondances (AFC). Cette analyse est adaptée aux variables qualitatives et tolère de grandes différences de fréquences en expliquant essentiellement les fortes fréquences [10].

La contribution des valeurs propres à l'inertie totale (pourcentages expliqués par les axes principaux) permet de constater qu'avec cette analyse nous représentons :
41,3% + 8,5% + 7,1% + 5,0% + 4,7% = 66,6% de l'information contenue dans le champ section de l'interrogation de 500 titres. De plus, il faut remarquer la grande importance de l'axe 1 qui représente à lui seul 41,3% de l'information.

Les coordonnées des observations (références) et leurs contributions ne seront pas mentionnées ici, car, bien que le logiciel les fournisse, le listing est long (500 lignes). Il en est de même pour les graphiques. Nous avons choisi de présenter les plans 1-3 et 4-5 qui représentent respectivement 48,4 et 9,7%. Le deuxième plan, pauvre en quantité d'informations, est riche par sa qualité, alors que l'axe 2 qui représente 8,5% n'apporte pas d'information intéressante.

Représentation du plan factoriel 1-3. Ce plan nous permet d'observer 48,4% de l'information et nous montre que la contribution des sections 9 (Biochemical methods) et 17 (Food and feed chemistry) est très forte pour l'axe 1 (respectivement 39,8 et 59 %) alors que l'axe 3 est influencé en plus de façon non négligeable par les sections 10 (74,8%) [Microbial biochemistry], 80 (7,6%) [Organic analytical chemistry].

Au niveau de l'examen des points cachés, il apparaît que la grosse majorité des références de cette interrogation se répartit en deux paquets d'inégales grosseurs : un sous l'influence de la section 9 (Biochemical methods), l'autre sous l'influence de la section 17 (Food and feed chemistry). Une étude détaillée des numéros de référence permettrait de connaître les sections constituantes, ceci grâce à l'isomorphisme des espaces duaux dû à l'utilisation de la métrique du $\chi 2$. Autrement dit, des numéros des références situés au barycentre de sections contiennent celles-ci.

Représentation du plan factoriel 4-5. Ce plan représente 9,7% de l'information et nous permet de découvrir le rôle d'autres sections. Ainsi l'axe 4 est essentiellement influencé par les sections 80 (63,8%) [Organic analytical chemistry], 13 (19,3%) [Mammalian biochemistry], 4 (13,1%) [Toxicology]; alors que l'axe 5 est commandé par les sections 13 (58,8%) et 4 (24,1%). L'apport de ce plan est donc essentiellement représenté par les sections 80, 13, 4.

Classifications automatiques. Des classifications automatiques sont également envisageables permettant l'élaboration de dendrogrammes afin de pouvoir analyser le degré de similitude entre les différentes sections. Ces études, possibles avec les programmes dont nous disposons, sont réalisables avec les données de départ ou avec les coordonnées des variables sur les axes factoriels. Cependant elles ont l'inconvénient de faire perdre la trace des références de départ. De plus, elles vont indiquer des similitudes de « comportement » plus que des similitudes réelles. Par exemple, les sections 09 et 17 qui apparaissent opposées dans l'analyse factorielle sont proches en analyse hiérarchique ascendante.

Cette étude nous amène à développer deux types de discussions.

Interprétation de l'interrogation

Ces analyses nous ont permis, sur 500 titres, de cerner les thèmes des recherches s'intéressant à la chromatographie des acides gras et des phospholipides.

Pour comprendre le phénomène qui amène une section à émerger de l'analyse, il faut différencier la façon dont interviennent les sections dans l'ensemble des références bibliographiques sélectionnées lors de l'interrogation :
- 1 un nombre de liaisons importantes;
- 2 un nombre de liaisons faibles;
- 3 une forte fréquence;
- 4 une faible fréquence;
- 5 principale essentiellement;
- 6 secondaire essentiellement;
- 7 fréquence trop faible.

Cela nous permet de montrer le rôle de quelques sections dans l'ensemble des références sélectionnées.

Catégorie Sec	1	2	3	4	5	6	7
s10		x	x			x	
s11	x		x			x	
s12		x		x		x	
s13	x	x			x		
s14		x	x			x	
s16		x		x	x		
s17	x		x		x		
s18		x		x		x	
s02				x			x
s20		x		x			x
s04	x			x	x		
s06		x		x			x
s80	x		x		x		
s09	x		x		x		

Le fait qui met en évidence les sections dans l'AFC est un compromis entre la fréquence, la quantité de liaison, et la nature (principale ou secondaire) de la section.

Parmi les deux sections à plus forte fréquence 09 (Biochemical methods) et 17 (Food & feed chemistry), la section 17 dans le premier axe de l'AFC est plus importante, car, en proportion, elle est beaucoup plus liée que la section 09 (qui est pourtant à plus forte fréquence). Elles permettent d'avoir une première idée : ce sujet est encore dominé par la mise au point des méthodes (confirmé par la présence des sections 22 : Physical organic chemistry, 64 : Pharmaceutical analysis, 72 : Electrochemistry, 73 : Optical; electron; and mass spectroscopy and other related properties, 06 : General biochemistry, 08 : Radiation biochemistry, qui sont toutes des sections de méthodologie chimique).

L'apparition des sections 10 (Microbial biochemistry), 80 (Organic analytical chemistry), 13 (Mammalian biochemistry) ne fait que confirmer le premier point avec un nouveau centre d'intérêt : la biochimie microbienne et des mammifères. Le rôle de la section 80 dans l'AFC est important et s'explique par la grande quantité de liaisons de cette section qui a une relativement faible fréquence et est essentiellement secondaire.

La présence avec un rôle non négligeable de la section 04 (Toxicology) est à relier à la présence de la section 14 (Mammalian pathological biochemistry) qui a une forte fréquence mais un faible rôle vu son peu de liaisons avec le reste et le fait qu'elle est toujours secondaire.

Les sections 11 : Plant biochemistry, 12 : Nonmammalian biochemistry, 15 : Immunochemistry, 2 : Mammalian hormones, 5 : Agrochemical bioregulators, 6 : General biochemistry, même si elles ne sont remarquables ni par leurs fréquences ni par leurs liaisons, montrent que de nombreux secteurs s'intéressent à ces techniques.

En voyant les sections représentant les secteurs industriels (42 : coating; inks and related products, 43 : cellulose; lignin; paper; and other wood products, 45 : industrials organic chemicals; leathers; facts; and waxes, 46 : surface-active agents and detergents, 62 : essential oils and cosmetics, 66 : surface chemistry and colloids, 16 : fermentation and bioindustrial), nous pouvons penser que les applications de ces techniques peuvent être nombreuses; mais leurs très faibles fréquences nous confirment le fait que ce sujet en est encore à la mise au point et soulève suffisamment d'intérêt pour que de nombreuses techniques soient essayées sans qu'une méthode émerge encore. Les principaux secteurs actuels d'applications potentiels sont: la chimie de l'alimentation avec ses implications biochimiques, la microbiologie avec surtout ses applications pathologiques, la biochimie végétale en relation avec l'alimentation.

L'étude d'ouvrages de synthèse et revues, obtenues à partir de l'interrogation [11-16], a été entreprise parallèlement pour chercher une confirmation *a posteriori*.

Il y apparaît que la séparation de mélanges d'acides gras ou phospholipides par des méthodes chromatographiques n'est pas évidente :
– vu la multiplicité des produits possibles, une méthode ne peut tout séparer. Si

l'HPLC n'est pas destructrice, la détection pose des problèmes. Si la détection par GC-MS est quasi universelle, la chaleur requise rend difficile le dosage des phospholipides ou de certains acides gras insaturés [17];

– la mise au point de méthodes se pratique sur les échantillons les plus riches en ces substances et les moins chers, à savoir les extraits de soja, de bactéries et d'œufs;

– la constitution graisseuse du système nerveux et le rôle des lipides saturés dans les maladies cardiovasculaires expliquent l'intérêt que portent les pathologistes et les bromatologistes à ces méthodes [15];

– les techniques de la chromatographie couche mince, si elles sont faciles à mettre en œuvre, n'ont pas le même pouvoir séparateur que la chromatographie en phase gazeuse capillaire ou la chromatographie liquide haute performance avec cependant des résultats intéressants à cause de la possibilité de détection par marquage radioactif [13, 14, 16].

Perspective de cette méthodologie

Cette application nous permet de dégager des axes de développement. En effet, nous signalions au début la possibilité de travailler sur un autre champ. Ainsi, il serait possible de la même façon d'étudier les auteurs ou les revues ou les *suplementary terms* ou d'étudier une évolution temporelle en incluant l'année dans un champ traité.

Nous pensons que l'intérêt de ce type d'analyse est maintenant clair. Une interrogation de banques de données, exploitée de la sorte, est un gain de temps non négligeable pour le demandeur de l'interrogation, quelle que soit l'utilisation qu'il veut en faire.

La multiplicité et la facilité d'obtention des informations scientifiques implique que parallèlement se développent des méthodes de résumé et d'extraction automatique de l'information ainsi que des représentations schématisées.

Si ces méthodes sont (ou seront) disponibles au niveau des serveurs, nous pensons qu'il est intéressant de le développer au niveau local. Le micro-ordinateur constituera ainsi une station de travail intégrée permettant à la fois les traitements scientifiques, bureautiques et de l'information scientifique et technique.

Note : nous tenons à remercier Orbirt Search Service pour l'aide apportée lors de la réalisation de ce travail. Les figures correspondant aux graphes factoriels peuvent être obtenues auprès du laboratoire.

Bibliographie

[1] Deroulède A, Dutheuil C *Informations Chimie* (1990) 315, 240

[2] Leydesdorff L, Zaal R *First International Conference on Bibliometrics and Theorical Aspects of Information Retrieval* (1987) 24-28/08, 29

[3] Dore JC, Gilbert J, Miquel JF, Deroulede A, Dutheuil C *Revue française de bibliométrie appliquée* (1987) 1, 14

[4] Dou H, Hassanaly P, La Tela A *Congrès de la société française de bibliométrie appliquée* Ile Rousse (1987) 23-25/10, 19

[5] Gilbert J, Dore JC, Miquel JF, Dutheuil C, Seite B et al *L'actualité chimique* (1986) 04, 21

[6] Hassanaly P, Pullino J, Dou H *Congrès de la société française de bibliométrie appliquée* Ile Rousse (1987) 23-25/10, 189

[7] La Tela A Thèse de doctorat Marseille (1987)

[8] La Tela A *Congrès de la société française de bibliométrie appliquée* Ile Rousse (1987) 23-25/10

[9] Dou H, Hassanaly P, Messiane E *Congrès de la société française de bibliométrie appliquée* Ile Rousse (1987) 23-25/10, 93

[10] Legendre L, Legendre P *Écologie numérique*, Masson, 2e édition, tome I et II, 260 pp et 335 pp

[11] *Aitzetmuller K Fette Seifen Anstrichm* (1984) 8, 318

[12] Christie W W Z Lebensm Unsters Forsch (1981) 181 (3), 171

[13] Kates M Laboratory *Techniques in biochemistry and molecular biology* (Work J J, Ed) North-Holland (1972)

[14] Kuksis A Myher JJ *J Chromatog* (1986) 379, 57

[15] Macrae R *Food Science and Technology* (Stewart GF, Schweigert B S, Hauwthorn J Ed) Academic Press (1982)

[16] McCluer RH, Ullman M D, Jungalwala FB *Adv Chromatog* (1986) 25, 309

[17] Mallet G communication personnelle