Description des structures moléculaires basée sur un classement ordonné de leurs atomes constitutifs. Application aux composés possédant un centre de chiralité (1re partie).

par Robert Luft,

(Laboratoire de chimie organique, Institut Polytechnique Méditerranéen, Université de Nice, 28, av. de Valrose, 06034 Nice Cedex)

Au milieu du siècle dernier les chimistes avaient déjà reconnu que l'une des caractéristiques les plus marquées des atomes de carbone correspondait à leur grande aptitude à s'unir entre-eux. Ce comportement distingue le carbone de ses voisins du tableau périodique et explique, en liaison avec son indice de « connectivité » élevé, la possibilité d'existence d'un nombre important d'isomères, dès que la condensation en atomes de carbone d'une molécule atteint un seuil relativement bas. C'est ainsi qu'à une « formule brute » $C_3H_6O_2$ peuvent déjà correspondre 25 isomères (diastéréoisomères inclus) et il est par conséquent aisé de comprendre que la représentation convenable et non ambiguë de toutes ces entités peut constituer un problème délicat.

Représentations conventionnelles des structures

Pour représenter les molécules sur un support matériel (feuille de papier, tableau noir), A.M. Boutlerov a proposé (1) dès 1863 l'utilisation de « formules structurales » conventionnelles dont toute considération d'ordre stéréochimique, c'est-à-dire se rapportant à la répartition spatiale des grains de matière constituant la molécule, était exclue (2). D'ailleurs, à cette époque la controverse autour de la notion d'atome (réalité physique ou concept métaphysique) était encore loin de s'apaiser.

Le maniement de ses formules permit à Boutlerov d'affirmer qu'aucune théorie structurale ne pouvait aboutir si elle n'était basée sur le principe d'une correspondance biunivoque entre un composé et sa représentation, cette dernière devant permettre de rendre compte des propriétés liées à la distribution des atomes dans la molécule (relations structurales internes) et aussi à la morphologie de cette dernière (caractères structuraux globaux). Dans une étude relative à la structure du benzène et aux « formules structurales » aptes à la représenter, Ladenburg écrivait (3) en 1869 « si, comme il est courant, des représentations graphiques sont utilisées pour visualiser la constitution d'une molécule, alors les relations géométriques déterminent les relations entre atomes. En aucune façon la figure n'est présumée indiquer la disposition spatiale des atomes ».

Telle était aussi la pensée de A. Kekulé, lorsqu'il introduisit, peu après 1860, l'usage de modèles matériels des « atomes » dans ses enseignements. Puisqu'à l'époque la notion d'atome, au sens où nous l'entendons aujourd'hui, était loin d'être admise par le plus grand nombre, les modèles de Kekulé n'avaient en aucune façon la prétention de représenter la disposition spatiale réelle, mais seulement la « valence » de chaque atome. Pour le carbone Kekulé avait retenu un modèle tétraédrique, car il avait observé (4) que ce dernier traduisait convenablement les différentes connectivités du carbone, c'est à dire que ce modèle permettait une bonne visualisation aussi bien des liaisons simples que des liaisons multiples.

Ces restrictions et mises en garde attachées aux représentations conventionnelles n'ont malheureusement pas toujours réussi à éviter la confusion des esprits, même des plus éminents. Dans ce contexte on peut citer l'erreur historique de A. von Baeyer. Ce chimiste de grande réputation, habitué à représenter conventionnellement les carbo-

cycles par des polygones réguliers, finit par confondre son modèle (imparfait) et la réalité pour dégager du premier sa célèbre théorie de la tension dans les cycles (5). Mais, tout autant que le sien, l'esprit de ses contemporains était obnubilé par la représentation plane des cycles, de sorte que les idées de Sachse (6, 7) sur la non-planéité du cyclohexane et sur les conformations « chaise » et « flexible » de ce composé n'arrivèrent pas à percer. La « simplicité et la clarté » de la théorie de Baeyer étaient telles que près de 40 ans après, ni les études théoriques de Mohr (8) sur les isomères Z et E de la décaline en particulier, ni leur vérification expérimentale par Hückel (9) n'ont dessillé les yeux des chimistes. Il a fallu attendre les travaux de Barton sur les conformations (10) pour que l'ensemble des chimistes concernés par des problèmes structuraux acceptent l'évidence. Mais cela n'empêche pas la théorie de la tension dans les cycles de hanter encore de nos jours l'esprit de bon nombre de scientifiques et de trouver sa place dans des ouvrages didactiques parmi les plus récents, tant elle a la vie dure.

Quoi qu'il en soit, le besoin d'une représentation spatiale des molécules s'est rapidement fait sentir, car les formules structurales conventionnelles ne permettaient pas de résoudre tous les problèmes liés à l'isomérie moléculaire, problèmes d'autant plus aigus que le nombre de molécules organiques synthétisées croissait très rapidement.

Principes généraux de description ordonnée des molécules

Le pas décisif vers une description complète de la topographie des molécules a été franchi en 1874 par J.H. Van't Hoff (11) et J.A. Le Bel (12), lorsque ces deux chimistes, partant de deux hypothèses nettement distinctes, ont abouti séparément à la conclusion que le modèle tétraédrique directionnel du carbone, utilisé par Kekulé pour indiquer les axes des liaisons, n'avait pas seulement la signification formelle qui lui était attachée jusqu'alors, mais traduisait convenablement les orientations réelles des substituants d'un atome de carbone.

Les premières tentatives de descriptions moléculaires ont été effectuées dès 1874 par A. Cayley (13) qui appliqua la théorie des graphes arborescents qu'il avait créée en 1857 (14) à la détermination du nombre des isomères caractérisés par une même « formule brute » (signalons cependant que l'étude exhaustive des isomères et stéréoisomères d'une population chimique donnée n'a été vraiment menée à bien (15) que vers 1930). Depuis cette époque les chimistes confrontés aux problèmes de description des molécules pratiquent implicitement ou explicitement les théories des graphes. En effet, à partir du moment où les traits qui, dans les représentations structurales, relient les atomes entre-eux se sont vus attribuer la signification de liens « réels et localisés dans l'espace », les représentations des enchaînements atomiques constituent des graphes au sens strict. Il n'entre pas dans notre propos de développer ici cette théorie aisément accessible au lecteur (8 à 10) qui est à la base de certaines méthodes modernes d'analyse structurale et dont les applications à des problèmes d'ordre chimique font l'objet d'un ouvrage récent (19).

Relevons simplement que par les procédés logiques qu'elle implique, la théorie des graphes a permis la création d'un métalangage cohérent (20) particulièrement adapté à la description topologique des molécules. Dans ce cadre la méthodologie DARC (21 à 23) par exemple constitue un outil très intéressant pour obtenir, à partir d'un diagramme structural moléculaire, un descripteur unilligne utilisable pour la codification, le stockage et la manipulation informatique des formules moléculaires.

Un tel descripteur est basé sur l'énumération de chacun des sites d'atomes rencontrés dans la molécule, sur leur localisation, ainsi que sur la caractérisation des espèces atomiques et des liens existant entre-elles. Ces opérations doivent être effectuées selon un ordre logique et aboutir à un descripteur univoque. A cet effet, il faut procéder au préalable à un classement ordonné de tous les atomes de la molécule. Plusieurs méthodes ont été proposées depuis la création des formules dualistiques par Berzelius. En nous limitant à celles qui ont été discutées depuis la seconde guerre mondiale, nous devons citer tout d'abord le travail de Dyson (24), puis celui de Wiswesser (25 et 26). Tous deux aboutissent à une description univoque des molécules, mais contiennent une bonne part d'arbitraire dans les choix indispensables pour arriver à une telle description. Le second (appelé « notation WLN ») a trouvé un certain nombre d'applications dans la mise au point de fichiers informatiques de substances chimiques (Index Chemicus Registry System, Excepta Medica Drugdoc, etc...), car son maniement semble simple. L'emploi d'un nombre restreint de symboles nécessite quelquefois le recours au même signe pour des significations différentes, le bon choix dans l'interprétation est l'affaire de l'analyste.

La logique sous-jacente à l'attribution des symboles n'apparaît pas toujours clairement, comme le montre l'extrait suivant de la liste des symboles WLN :

1, 2, 3, 4... *n*	chaîne alkyle linéaire à *n* chaînons
	1 représente $-CH_3$, $-CH_2-$, $-CH=$, etc, mais pas $>CH-$ ou $-\overset{\vert}{\underset{\vert}{C}}-$
	2 représente CH_3-CH_2-, $-CH_2-CH_2-$, $-CH_2-CH=$, etc, mais pas $-CH=CH-$, $-CH=C<$, $-CH_2-CH<$, etc.
B, C, D, E, F...	peuvent représenter une position de substituant sur un cycle
B, E, F, G, I	symbolisent le bore, le brome, le fluor, le chlore, l'iode
C	symbolise le carbone dans $-C\equiv N$, $=C=$ et des cas analogues
Y	caractérise un carbone siège de ramification, par exemple un carbone tertiaire $-CH<$, ou $>C=$
X	signale un carbone quaternaire
U	représente $>HC=CH<$
UU	représente $-C\equiv C-$

Pour l'azote et l'oxygène les symboles sont tout autant variés :

N	azote non porteur d'hydrogène, de connectivité 3
K	azote de connectivité 4, par exemple ammoniums $\overset{+}{N}R_4$
M	groupe $>NH$ ou $=NH$
Z	groupe NH_2
O	oxygène des éthers, lactones, etc., c'est-à-dire non porteur d'hydrogène (sauf cas symbolisés par V et W)
Q	groupe OH
V	groupe $>C=O$
W	deux oxygènes fixés sur un même atome (sauf esters, lactones)
Etc...	

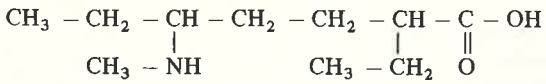
Les risques de confusions et d'erreurs d'interprétation nécessitent un bon contrôle de la codification, mais un tel contrôle est souhaitable pour tous les systèmes. Ce qui paraît plus gênant est le choix imposé du point origine de la description. En considérant la séquence

espace & – /Φ 1 2 3 ... A B C ... Z

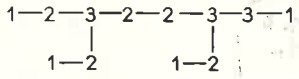
on doit commencer la description à partir du symbole de la formule linéaire placé le plus à droite dans cette liste. Comme la liste des symboles ne correspond pas à la séquence de l'alphabet, pas plus qu'à une autre séquence consacrée ou logique, l'origine de la description de plusieurs corps, ne se distinguant que par les substituants d'un même squelette carboné, changera au gré de l'ordre de la liste des symboles.

Le système de Chemical Abstracts Service (CAS)

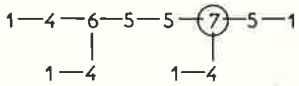
Chemical Abstracts Service a retenu un autre système univoque de description des molécules qui fait appel à un algorithme mis au point par Morgan (27), et est basé sur la préséance du graphe topologique de la molécule à analyser sur son graphe chromatique. Avec l'algorithme « Morgan-CAS » l'ordre de préséance des atomes d'une molécule est établi à partir des valeurs de connectivité des nœuds du graphe topologique de la molécule. L'étape de notation de chaque connectivité franchie, on procède, au niveau de chaque nœud, à la sommation des connectivités des nœuds voisins, puis renouvelle l'opération aussi souvent que nécessaire pour faire apparaître le nœud prioritaire et des priorités entre les nœuds du premier rang, reliés au nœud prioritaire. Le nœud prioritaire représente l'atome origine de la description à partir de laquelle on décrit successivement les atomes de chaque rang. Notons encore que les atomes d'hydrogène sont « transparents », c'est-à-dire qu'ils ne sont pas pris en compte dans le graphe. Voici quelques exemples de détermination des origines de description et de priorité des branches (voir ci-contre) :



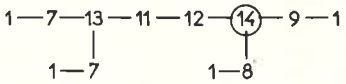
molécule à décrire



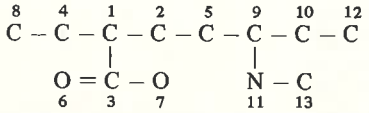
relevé des connectivités dans le graphe initial
(les atomes H sont « transparents », donc n'entrent pas en ligne de compte).



première sommation des connectivités
un atome prioritaire apparaît, il y a indétermination au niveau des atomes du premier rang (issus de l'atome prioritaire.

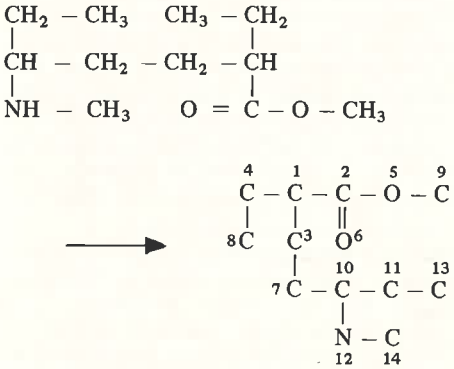


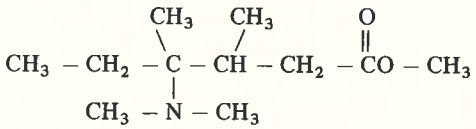
seconde sommation
les atomes du premier rang sont différenciés.

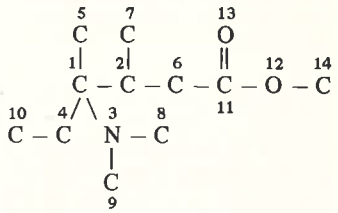


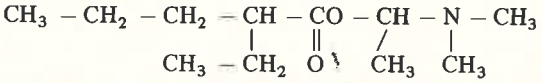
graphe final
avec indication de l'ordre de description des atomes, numérotés dans l'ordre des priorités décroissantes.

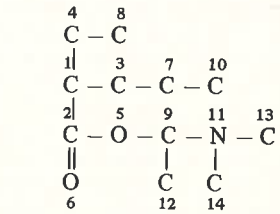
Fixons maintenant un méthyle sur le carboxyle, nous obtenons, à partir de la structure initiale, le graphe final (placé en dessous) :



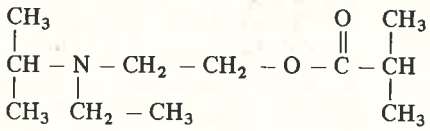


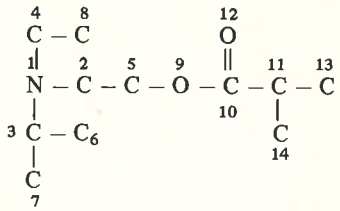




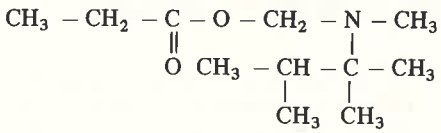


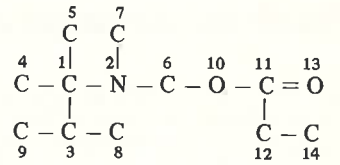
Si l'atome prioritaire n'a pas changé, il y a eu par contre mutation au niveau des directions de développement du graphe. Examinons maintenant quelques composés isomères du précédent, c'est-à-dire présentant tous la même formule brute $C_{11}H_{23}NO_2$, une fonction ester et une fonction amine, nous trouvons (voir ci-contre) :

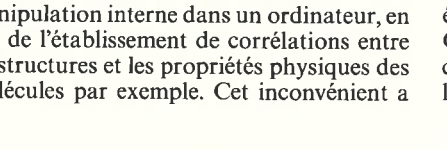


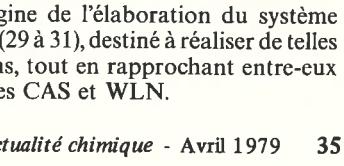


Ainsi, pour les cinq substances isomères précédentes le point origine de chacune des arborescences est imposé par l'ordre induit sur la structure moléculaire par application de l'algorithme « Morgan – CAS ». Comme dans le système WLN la hiérarchisation des différents sites de la molécule est imposée par des critères situés en dehors de ceux sur lesquels un chimiste pourrait peser en fonction d'aspects circonstanciels propres à sa recherche. D'autre part, comme l'ont relevé Campey, Hyde et Jackson (28), aucun des deux systèmes CAS ou WLN ne conduit à une restitution de données qui permette leur manipulation interne dans un ordinateur, en vue de l'établissement de corrélations entre les structures et les propriétés physiques des molécules par exemple. Cet inconvénient a été à l'origine de l'élaboration du système Crossbow (29 à 31), destiné à réaliser de telles corrélations, tout en rapprochant entre-eux les systèmes CAS et WLN.



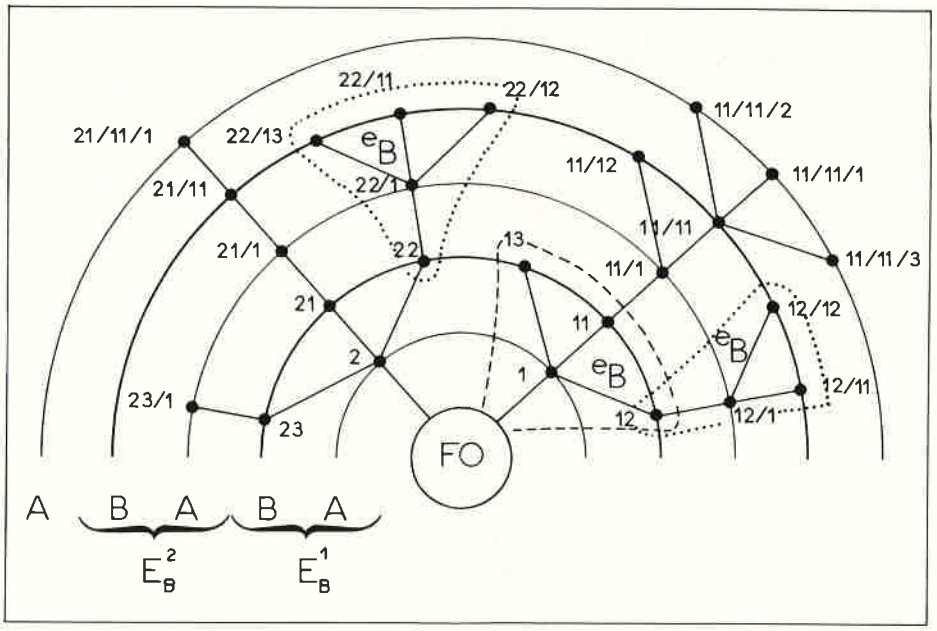






Le système DARC

Alors que l'algorithme « Morgan – CAS » attribue la préséance au graphe topologique, l'approche de la description des molécules est totalement différente avec l'algorithme DARC/DUPEL (32). Pour ce dernier l'ordonnance du graphe relève à la fois de la topologie et du chromatisme, les hiérarchisations successives apparaissent de façon concrète (20). Ce second algorithme conduit à un « maître-graphe » à partir duquel on établit entre autres le nom codé univoque de la molécule. Dans ce maître-graphe les sites se voient affecter un ordre canonique de référence ou « maître-ordre », mais l'algorithme DARC/DUPEL est compatible avec l'établissement d'ordres différents, ayant pour origine ou foyer « FO » un site quelconque du maître-graphe, « finalisés » en fonction de tel ou tel but propre. Ce foyer peut être constitué par un atome ou un ensemble d'atomes, par exemple une fonction chimique ou un fragment structural tel qu'un cycle, une chaîne ou, plus simplement, une partie du squelette de la molécule. Contrairement à ce qui a lieu pour les systèmes CAS et WLN, l'interversion des ordres finalisés entre-eux et avec le maître-ordre est ici aisément réalisable par transcodage au niveau de la mémoire de travail d'un ordinateur.



Examinons de plus près le système DARC/DUPEL. En vue de sa description, chaque composé chimique est assimilé à un graphe chromatique (20) dont les nœuds représentent les atomes et les arêtes les liaisons qui les joignent (par convention et pour alléger les graphes on convient habituellement de ne pas prendre en compte les atomes d'hydrogène). Le chromatisme des nœuds traduit la nature des atomes et leur connectivité maximale, tandis que le chromatisme des arêtes exprime la multiplicité des liaisons. La molécule et le graphe ordonné G_χ qui la représente sont homomorphes, c'est-à-dire qu'ils possèdent même topologie et même chromatisme ; en outre le graphe contient des informations supplémentaires se rapportant à un « ordre complexe ». Celui-ci englobe à la fois la notion de maître-ordre et la notion de distance topologique concentrique et discrète. A côté d'autres applications la notion d'ordre complexe permet de situer un composé parmi ses analogues.

Le graphe chromatique G_χ est élaboré selon une « séquence de génération ordonnée des atomes » qui comporte quatre étapes. Celles-ci se reflètent d'autre part dans la description DARC du composé dont on veut établir le graphe G_χ. Ces étapes sont les suivantes :

- la focalisation ou sélection d'un foyer FO à partir duquel la procédure de génération permet la découverte progressive de l'environnement E et son investigation.
- l'organisation de l'environnement E. A cet effet l'espace est découpé en couronnes centrées sur le foyer FO ; les nœuds du graphe sont affectés d'un rang par rapport au foyer. Dans un rang il y a, au plus, autant de nœuds qu'il existe de « directions de développement » à partir du foyer ou des nœuds du rang antérieur. Par convention on affecte chaque noeud de rang impair à une couronne notée A_i, et chaque nœud de rang pair à une couronne notée B_{ij}, le foyer correspondant au rang zéro. Par ailleurs on subdivise l'environnement total en zones d'« environnement limité E_B^i » contenant chacune deux couronnes successives, A_i et B_{ij}. Enfin, à l'intérieur de chaque E_B^i on regroupe dans un « segment d'environnement e_B » l'ensemble des positions qui dérivent d'un même nœud de l'environnement limité antérieur E_B^i.
- l'ordonnance du graphe qui consiste à associer un indice à chacune des positions d'un segment e_B, grâce à une fonction d'ordonnance, résultante de la composition d'une fonction d'ordonnance topologique et d'une fonction d'ordonnance chromatique. On procède successivement au niveau de chacun des segments e_B, en fonction des priorités décroissantes. Lorsqu'un e_B est totalement ordonné, on dit qu'il constitue un Environnement Limité Concentrique Ordonné ou ELCO. Le graphe est totalement ordonné lorsque la propagation de l'ELCO a englobé tout l'environnement du foyer qui constitue le nœud initial du graphe.
- la valuation chromatique se fait progressivement au niveau de chaque e_B.

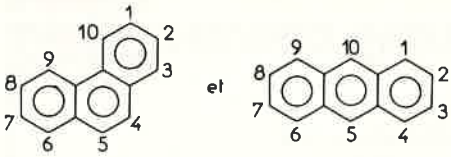
Dans la pratique, pour les noeuds d'un même rang, le classement se déduit de l'examen d'une série de critères hiérarchisés : *la connectivité* de chaque nœud, *la multiplicité* de la liaison entre le nœud considéré et le nœud antérieur, *le numéro atomique* de chacun des atomes figurés par les nœuds d'un rang ordonné.

La priorité la plus élevée est attribuée à la valeur numérique la plus élevée de chacun des critères. Lorsque l'application d'un critère ne permet pas de lever une indétermination, on passe au critère suivant.

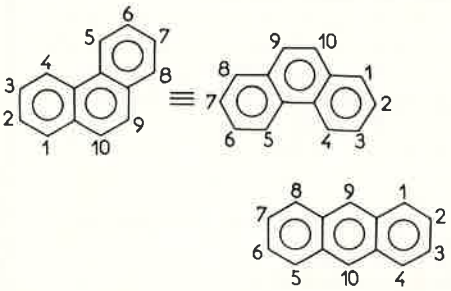
Intérêt des descripteurs topologiques moléculaires univoques

La méthode taxonomique que constitue la nomenclature chimique IUPAC (33) aurait pu permettre une description univoque des molécules si elle n'avait rencontré deux écueils qui bloquent toute vue prospective. Le premier réside dans la difficulté qu'il y a à cerner convenablement l'ensemble des classes de substances restant à découvrir et à maintenir ouverts en conséquence un certain nombre de créneaux dans les règles de nomenclature, pour pouvoir accueillir ces substances inconnues. Le second écueil est constitué par l'énorme quantité d'informations publiées dans les périodiques chimiques antérieurement à l'établissement de la nouvelle nomenclature. Bouleverser une seule règle peut alors revenir à créer une confusion énorme, lorsqu'on exploite simultanément des documents antérieurs et postérieurs au changement taxonomique. Pour cette raison les commissions de nomenclature ont été très prudentes dans leurs propositions de changement et ont préféré dévoyer dans certains cas la logique sous-tendant les principes taxonomiques de base en introduisant des exceptions ou des possibilités de choix dans le lot des règles établies. A titre d'exemple simple, indiquons que la règle A-22-1 de l'IUPAC impose pour le phénanthrène et l'anthracène les numérotations

suivantes, sur la base d'un numérotage systématique.



Mais, compte tenu de l'importance de la littérature relative à ces deux composés, parue avant la refonte de la nomenclature dans les règles actuelles, la règle A-22-5 de 1965 « recommande » (mais n'impose pas) les exceptions suivantes aux principes de numérotage :



Ces exceptions et possibilités de choix ne sont pas les seules, loin de là ; mais elles suffisent à démontrer le caractère non univoque du système de nomenclature.

De la même façon on serait amené à constater que les systèmes de répertoriation sur lesquels sont fondés « Beilstein's Handbuch der organischen Chemie », le « Ring Index » de Patterson, le « Merck Index », etc... ne sont pas basés, en partie ou en totalité, sur des aspects structuraux des substances décrites. Rappelons enfin que les idéographes que constituent les formules structurales conventionnelles, bien qu'ils fournissent d'un coup d'œil une information variée et dense, ne sont pas exempts d'inconvénients ; nous l'avons vu sur l'exemple de la représentation plane des cycles saturés.

Rappelons que les seules descriptions structurales univoques des molécules sont basées sur des classements ordonnés de leurs atomes constitutifs. Les descriptions obtenues avec l'algorithme DARC/DUPEL ont l'avantage de restituer complètement la topologie moléculaire, car elles situent tous les atomes de la molécule en précisant pour chacun ses relations de voisinage. Les descripteurs DARC sont biunivoques ; modulaires (les ELCO sont décrits successivement) et séquenciels (au niveau de la description juxtaposée du graphe d'existence ainsi que de la nature des liaisons et des atomes). Ces qualités leur confèrent un caractère « ouvert » qui permet l'adjonction de nouvelles informations sans remise en cause de celles acquises antérieurement.

Mais là ne s'arrêtent pas les avantages que présentent les descripteurs topologiques biunivoques sur les descripteurs taxonomiques classiques. En effet, ils sont particulièrement utiles dans la recherche ou la mise en évidence de sous-structures ou de fragments de structure communs à un lot de molécules. Dans le système DARC la génération de fragments est algorithmique, contrairement au cas du WLN, où les fragments sont prédéterminés. De ce fait la génération de fragments, dont certaines parties ne sont pas définies, restent indéterminées ou floues, est possible. Enfin, ces fragments DARC peuvent être jointifs ou de nature recouvrante, au gré de l'opérateur. Notons en particulier l'intérêt de la fragmentation recouvrante algorithmique DARC/FREL comme outil de restitution de structures, les fragments isolés (ou sous-structures) étant caractéristiques de la structure.

Ce sont ces qualités de restitution, jointes à la facilité de transcodage entre CAS et DARC qui ont amené récemment (34) le Chemical Abstracts Service à accorder une exclusivité pour l'exploitation de leur fichier en France au Centre National d'Informatique Chimique. Ce Centre, dans lequel sont représentés, comme on le sait, le CNRS, l'UIC, ainsi que les Ministères de la recherche, des Universités, de l'industrie et du commerce, est équipé pour le traitement des données par le système DARC.

Deuxième colloque audio-visuel (AVEMS)

Selon le désir exprimé par les participants au premier colloque audio-visuel qui s'est tenu à Poitiers les 23 et 24 mars 1978, le deuxième colloque aura lieu à Montpellier les 6 et 7 juin 1979, sur le thème bien précis suivant : *Vidéoscopie et formation des Maîtres scientifiques.*

Les deux journées de travail comprendront des matinées avec conférences plénières et discussions, et des après-midi avec présentation de documents réalisés sur ce sujet dans diverses Universités.

Quatre conférences sont prévues par : Mme A. Kornhauser (Yougoslavie), M. M. Champagne (Québec), M. J.-M. de Keteke (Belgique), M. Postic (France).

Les personnes désirant recevoir la seconde circulaire ainsi que celles qui désirent présenter un document sont priées de contacter, le plus rapidement possible, Mme Danièle Cross à l'adresse suivante : Laboratoire de chimie physique, U.S.T.L., Place E. Bataillon, 34060 Montpellier Cedex.

Recherches Coopératives en Didactique de la Chimie

Appel aux volontaires

Groupe de travail sur les « Comportements étudiants-enseignants (Chimie) ».

Après une année d'essais, le groupe de travail a mis au point une version adaptée du questionnaire Québécois « PERPE » (Perception par les Étudiants des Relations Professeurs-Étudiants). Ce questionnaire permet à l'enseignant de mesurer la satisfaction ou l'insatisfaction des étudiants.

Le groupe souhaite qu'une centaine d'enseignants de chimie des Universités françaises utilise ce questionnaire en 1978-9. La mise en commun des résultats permettra des progrès pédagogiques personnels et institutionnels.

Il est donc fait appel à tous les volontaires pour l'expérimentation de ce questionnaire. Pour tous renseignements sur le matériel nécessaire s'adresser d'abord à M. P. Boyer, Case Officielle 140, 54037 Nancy Cedex.

Les résultats fournis par ordinateur sont fort riches et l'interprétation peut en être délicate. Des collègues expérimentés peuvent aider à cette interprétation. S'adresser pour cela à M. G. Lepoutre, 13, rue de Toul, 59046 Lille Cedex.