Jacques-Émile Dubois, le système DARC, et son influence sur la pensée, la représentation et la manipulation de l'information chimique

Jacques-Émile Dubois, the DARC system, and his influence on conceptualizing, representing and handling chemical information

Michel Petitjean

Dans les années 50, la nécessité d'une notation linéaire et compacte des formules chimiques est apparue avec les premières banques de données informatisées, car les systèmes de nomenclature chimiques étaient – et sont toujours – très incomplets. Hélas, les premiers codes linéaires dédiés aux applications informatisées, tels le WLN (Wiswesser Line Notation), souffraient d'un défaut commun aux systèmes de nomenclature : ils étaient pensés suivant le schéma squelette-substituants. Dans les années 60, Jacques-Émile Dubois a pensé la représentation des molécules d'une manière fondamentalement différente, basée sur une représentation matricielle des couches concentriques autour d'un foyer, les atomes d'une couche étant ordonnés [1-4]. La représentation matricielle est elle-même linéarisée, donnant lieu à un code linéaire particulièrement simple lorsque seules deux couches sont considérées, sous l'hypothèse qu'aucun atome n'a plus de quatre voisins. Cette représentation fournit un accès immédiat à l'environnement local d'un atome, et sa première application concerna la synthèse de familles de cétones [5-7].

Le système DARC

Ainsi, le système DARC est né en 1966, mais c'était bien plus qu'un code linéaire des formules structurales. Les lettres D,A,R,C indiquent les principales utilisations du système DARC :

- **D**escription : représentation des structures (formules structurales), sous-structures et hyperstructures.
- **A**cquisition : saisie par ordinateur des structures et sous-structures.
- **R**estitution : retrouver des sous-structures dans des bases de données chimiques.
- **C**orrélations : applications de type QSAR (Quantitative Structure-Activity Relationships).

Pour la recherche de sous-structures dans une base de données chimiques organisée suivant le système DARC, la saisie des sous-structures se faisait en dessinant les formules avec un périphérique graphique, ce qui permettait aux chimistes d'opérer sans avoir à maîtriser le code DARC. L'interrogation de la banque se faisait ensuite en deux étapes : « RE » (Recherche Écran), puis « AA » (Atome par Atome). La commande « RE » était un filtre sophistiqué [8-11] basé sur les FRELs (Foyer Réduit à un Environnement Limité). Un FREL n'est rien d'autre qu'un atome avec ses deux premières couches concentriques de voisins [12], l'ensemble étant ordonné suivant les règles DARC, prenant en compte le sous-graphe local sans les hydrogènes, mais considérant la nature des atomes et des liaisons, plus quelques informations supplémentaires telles les isotopes, les charges, etc. Le système de filtrage complet avait une structure hiérarchisée optimisée pour la chimie, mais il ne doit en aucun cas être confondu avec les filtres de type bitscreen habituellement rencontrés en informatique. La commande « AA » était basée sur des tables de connectivité ordinaires, et aurait été trop lente si elle avait été effectuée sur des bases de données volumineuses sans passer d'abord par la commande « RE ». Le système DARC complet de recherches par sous-structures s'est révélé hautement efficace en 1980, sur la version française de la base de données CAS, Chemical Abstracts Service (appelée EURECAS par la suite), contenant plusieurs millions de molécules, formatée suivant les besoins du système DARC. Dès 1981, le système DARC fut commercialisé par la division Questrel de Telesystemes, une société anonyme française, filiale de la DGT (Direction Générale des Télécommunications).

Le système DARC a été étendu aux formules de Markush afin de traiter les réactions chimiques et les brevets, ces derniers étant souvent basés sur des formules génériques. L'interrogation de bases de formules de Markush par des sous-structures de Markush est toujours effectuée chez Questrel-Orbit, qui est aujourd'hui une société indépendante spécialisée dans les bases de brevets (Derwent, INPI (Institut National de la Propriété Industrielle)). Le système DARC est également utilisé pour retrouver des informations chimiques dans d'autres bases de données chimiques, telles que des bases de spectres de masse ou de RMN. Dans les années 70, il été adopté comme système national d'information chimique [13].

Relations structure-activité (QSAR)

Rechercher des corrélations entre formules structurales et propriétés physico-chimiques ou activités biologiques est un moyen de prédire de nouvelles molécules actives. Malheureusement, la formule structurale s'apparente à un graphe, alors que les outils mathématiques actuels de calcul de corrélations sont inadaptés aux graphes, et travaillent presque tous sur des valeurs numériques. Aussi est-il d'usage de convertir le graphe en une série de descripteurs numériques avant de calculer les corrélations. Or, les descripteurs basés sur les formules structurales sont principalement pensés suivant le schéma squelette-substituants, schéma utile dans un contexte d'enseignement, mais peu pertinent dans bien des situations. En 1967, J.-E. Dubois a proposé de corréler les propriétés physico-chimiques aux descripteurs basés sur l'environnement concentrique local des atomes, comme dans le système DARC [14], rompant ainsi avec les approches traditionnelles, encore en vigueur aujourd'hui. Ce concept s'est avéré prolifique, et J.-E. Dubois en a dérivé plusieurs théories donnant lieu à de nombreuses publications [15].

De la stéréochimie au QSAR-3D

Le système DARC a également été pensé en fonction de la stéréochimie [16-17], et l'apport de J.-E. Dubois dans ce domaine

a été reconnu [18-19]. Toutefois, l'information stéréochimique est insuffisante pour de nombreuses applications QSAR, qui nécessitent une information 3D complète. Ce point était évidemment reconnu par J.-E. Dubois [20]. La relation entre l'information 2D (formule structurale) et l'information 3D (géométrie) a été mesurée [21], par comparaison entre l'indice de forme topologique et l'indice de forme géométrique, les deux ayant une expression mathématique commune [22]. Mais il apparaît que les deux coefficients de forme sont peu corrélés.
L'information 3D est d'un maniement complexe, en particulier dans la mesure où elle doit prendre en compte la flexibilité des molécules. Ce défi encore non résolu est l'un des thèmes de recherche d'actualité dans le domaine du QSAR-3D.

Conclusion

Dès les années 50-60, J.-E. Dubois se montra un chimiste visionnaire, concevant la représentation de l'information chimique en fonction des applications sur ordinateur. Il faut souligner ici que les programmes informatiques opérant sur des bases de données chimiques doivent être capables de lire n'importe quelle formule structurale, et non uniquement celles connues actuellement de la communauté scientifique.
Le progrès apporté par les théories de J.-E. Dubois peut se comparer à celui apporté par les règles de CIP (Cahn, Ingold, Prelog) aux anciennes manières de coder la stéréochimie, mais c'est le domaine de la chemoinformatique (autrefois informatique chimique) dans son ensemble qui a été profondément influencé par les travaux de Jacques-Émile Dubois.

Références

[1] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sci.*, **1966**, Paris, Série C, *263*, p. 764.
[2] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sci.*, **1996**, Paris, Série C, *263*, p. 1245.
[3] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sci.*, **1970**, Paris, Série A, *270*, p. 228.
[4] Dubois J.-E., Bonnet J.C., *C. R. Acad. Sci.*, **1970**, Paris, Série A, *270*, p. 1002.
[5] Dubois J.-E., Hennequin F., Chastrette M., *Bull. Soc. Chim. France*, **1966**, *11*, p. 3568.
[6] Dubois J.-E., Hennequin F., *Bull. Soc. Chim. France*, **1966**, *11*, p. 3572.
[7] Dubois J.-E., Schutz G., Normant J.-M., *Bull. Soc. Chim. France*, **1966**, *11*, p. 3578.
[8] Attias R., *J. Chem. Inf. Comput. Sci.*, **1983**, *23*, p. 102.
[9] Attias R., EURECAS/DARC : Thèse de Doctorat d'État, Université Paris 7, 14 mai **1992**.
[10] Attias R., *Encyclopedia of Library and Information Science*, Allan Kent (ed), Marcel Dekker, New York, **1992**, *50*, suppl. 13, p. 308.
[11] Attias R., *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, p. 415.
[12] Attias R., Petitjean M., *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, p. 649.
[13] Dubois J.-E., *J. Chem. Doc.*, **1973**, *13*, p. 8.
[14] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sci.*, **1967**, Paris, Série C, *264*, p. 1019.
[15] Dubois J.-E., Information Today Inc., Medford, New Jersey, **2004**, p. 149. www.chemheritage.org/pubs/asist2002/13-dubois.pdf
[16] Dubois J.-E., Alliot M.J., Viellard H., *C. R. Acad. Sci.*, **1970**, Paris, Série C, *271*, p. 1412.
[17] Luft R., *L'Act. Chim.*, **1979**, *5*, p. 37.
[18] Petitjean M., *Symmetry: Culture and Science*, **2005**, *16*, p. 309.
[19] Mezey P.G., *Symmetry: Culture and Science*, **2005**, *16*, p. 311.
[20] Dubois J.-E., Doucet J.P., Panaye A., *Bull. Soc. Chim. Belg.*, **1989**, *98*, p. 31-44.
[21] Bath P.A., Poirrette A.R., Willett P., Allen F.H., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, p. 141.
[22] Petitjean M., *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, p. 331.

Michel Petitjean
est chargé de recherche à l'ITODYS (Université Paris 7)*.

* ITODYS, Université Paris 7,
1 rue Guy de la Brosse, 75005 Paris.
Courriel : michel.petitjean@cea.fr