SIMO 2006 – systèmes d'information, modélisation, optimisation et commande en Génie des Procédés
Le virtuel dans la réalité quotidienne.
11-12 octobre 2006 - Toulouse (France)

APPROCHE MULTI-CLASSES DE REPRÉSENTATION DES MOLÉCULES POUR LA CONCEPTION DES PRODUITS-PROCÉDÉS ASSISTÉE PAR ORDINATEUR

Mourad KORICHI[1,2], Vincent GERBAUD[1], Xavier JOULIA[1], Pascal FLOQUET[1]

- (1) Laboratoire de Génie Chimique
  5 rue Paulin Talabot
  F-3110 Toulouse, FRANCE
  e-mail : Vincent.Gerbaud@ensiacet.fr ; France

- (2) Laboratoire de Valorisation et Promotion des Ressources Sahariennes
  Université de Ouargla, BP 511
  30000, Ouargla, ALGERIE
  e-mail : M.Korichi@arn.dz ; Algérie

Abstract: Computer Aided Product Design (CAPD) is widely used in process system engineering (PSE) as a powerful tool for searching novel chemicals. The crucial steps in CAPD are the generation of candidate molecules and the estimation of properties, especially when complex molecular structures like flavors are sought. In this paper, we present a multiclass molecular knowledge framework which is based on chemical graph theory and chemical knowledge. Three kinds of functional groups are defined: elementary, basic and composed groups. These serve to generate four classes of knowledge that can be useful for property estimation and molecular design (CAMD). An Input/output structure basing on XML language is defined to favor the interoperability between softwares.

Keywords: Molecular graph; Group contribution; Property prediction; CAMD

Résumé : La Conception de Produits Assistée par Ordinateur (CPAO) est largement utilisée dans le domaine « Process System Engineering » (PSE), comme un outil puissant pour la recherche de nouveaux produits chimiques. Les étapes cruciales de la CPAO sont la génération des molécules et l'estimation des propriétés, particulièrement quand les structures moléculaires complexes comme les arômes sont recherchées. Dans cet article, nous présentons une approche multi-classes de représentation des molécules basée sur les graphes moléculaires et la connaissance chimique. Trois catégories de groupes fonctionnels sont proposées : groupes élémentaires, groupes de base et groupes composés. Ces derniers servent à générer quatre classes de représentation qui peuvent être utiles pour la prédiction des propriétés et dans le design des molécules (CAMD). Une structure d'informations entrée-sortie basée sur le langage XML est définie, pour favoriser l'interopérabilité entre les logiciels.

MOTS-CLÉS : Graphe moléculaire ; Contribution de groupes ; Prédiction des propriétés ; CPAO.

INTRODUCTION

La Conception des Produits Assistée par Ordinateur (CPAO) est la technique inverse de la prédiction des propriétés basée sur le concept de contribution de groupes (Harper and Gani, 2000 ; Achenie and Sinha, 2003). Elle consiste à fixer un ensemble de propriétés dites cibles, puis à chercher/trouver par combinaison de groupes, les molécules satisfaisant ces propriétés. L'estimation des propriétés et la génération des molécules sont le plus fréquemment basées sur le modèle de contribution de groupes. Ici la connaissance de la structure moléculaire à différents niveaux est fondamentale. De plus, la disponibilité des valeurs de propriétés des corps purs et de mélanges est cruciale dans la conception des procédés chimiques. Les mesures expérimentales sont toujours la source primaire de ces données mais, cette voie n'étant pas toujours accessible, en raison des dépenses et des difficultés de réalisation, l'utilisation des techniques de prédiction peut lui être préférée. Le rôle de la molécule et éventuellement sa structure détaillée (formule brute, formule détaillée, groupes, coordonnées atomiques, …) est également important pour la conception des procédés.

Dans ce travail, nous présentons une approche multi-classes de représentation des molécules pour la conception des produits – procédés, basée sur la théorie des graphes. L'approche proposée propose quatre classes. Les relations entre ces diverses classes sont développées afin de générer l'ensemble des informations concernant la molécule. Cette dernière est utilisée en phase d'estimation des propriétés et dans la CPAO. Les problèmes à résoudre dans ce travail sont les suivants :

- Quel modèle choisir pour représenter les molécules ?

- Quelles structures de données choisir pour coder les graphes moléculaires ?
- Comment coder les différentes catégories de groupes fonctionnels ?
- Quel algorithme utiliser pour la décomposition des molécules en groupes fonctionnels ?
- Comment communiquent le système avec l'utilisateur, les bibliothèques de calcul des propriétés et la CPAO ?

REPRÉSENTATION DES MOLÉCULES

Une molécule est représentée par plusieurs informations, à savoir sa représentation 1D (par exemple formule brute) ou 2D (par exemple formule développée). Une telle description doit tout d'abord faciliter la manipulation et/ou la représentation de la structure ; même pour des structures moléculaires de grosse taille et complexe. Le respect des règles de base de chimie (valence neutre de la molécule, …) doit également être assumé. En général, deux familles de représentation sont utilisées en chimie :

- Les méthodes basées sur la ligne de notation. Celles-ci représentent la molécule sous forme d'une liste de caractères et symboles spécifiques pour identifier les atomes et les liaisons. Plusieurs types de ligne de notation sont disponibles dans la littérature, notamment, SMILES '*Simplified Molecular Input Line Specification*' (Weininger, 1988 ; Weininger *et al.*, 1989) et WLN '*Wiswesser Line Notation*' (Smith, 1968).
- D'autres méthodes plus générales, basées sur les concepts de graphe ou d'objet, décrivent les atomes et les liaisons de façon plus détaillée.

Le concept de graphe est ici utilisé comme une alternative pour symboliser la structure des molécules. Quelques modifications sont apportées afin de prendre en charge les différentes possibilités de représentation. Par définition un graphe moléculaire non orienté MG = (V, E) est constitué d'un ensemble V d'atomes et d'un ensemble E $\subseteq$ VxV de liaisons. Le graphe est considéré comme un graphe simple (Pogliani, 2000). Le MG peut être représenté par une variété de matrices telles que la matrice de connectivité entre atomes, la matrice de connectivité entre liaisons, la matrice d'incidence et la matrice de distance. La matrice de connectivité est la plus utilisée pour stocker la structure moléculaire. Le graphe moléculaire (MG) proposé dans ce travail est une matrice de taille N x N, où N est le nombre d'atomes dans la molécule (équation 1). Les éléments diagonaux représentent la structure de données des groupes élémentaires présents dans la molécule (voir section suivante). Les autres éléments de la matrice indiquent la connexion entre les atomes.

$$A = (a_{ij})$$

$$a_{ij} = \begin{cases} i \neq j \; a_{ij} \begin{cases} 1 & \text{si l'atome i est connecté à l'atome j} \\ 0 & \text{sinon} \end{cases} \\ i = j \; a_{ii} \quad \text{informations relative à l'atome i} \end{cases} \qquad (1)$$

DÉCOMPOSITION AUTOMATIQUE DE MOLÉCULE

La décomposition automatique de molécules en groupes fonctionnels est une étape fondamentale dans la prédiction des propriétés physiques et thermodynamiques à partir de la structure moléculaire, à savoir les techniques basées sur le concept de descripteurs moléculaires comme les méthodes de contribution de groupes (Joback et Reid, 1987 ; Constantinou et Gani, 1994) et les méthodes QSPR. La décomposition de la molécule est essentielle lorsque la molécule n'est pas disponible dans la base de données des simulateurs, ce qui peut être une alternative pour intégrer ce type de méthodes dans la simulation des procédés (Bünz et al., 1998).

Qu, Su Muraki et Hayakawa (1992 a & b) présentent une technique de décomposition de la molécule basée sur la notation WLN. Une modification a été apportée à cette notation pour supporter la génération des groupes. Récemment, Jeremy Rowley, Oscarson, Rowley et Winding (2001) ont développé une approche de génération des groupes fonctionnels à partir de la structure moléculaire en se basant sur la notation SMILES. La figure 1 présente les trois catégories de groupes fonctionnels définies dans ce travail: groupes élémentaires, groupes de bases et groupes composés. La méthodologie proposée est décomposée en quatre classes selon le mode de génération de l'information.

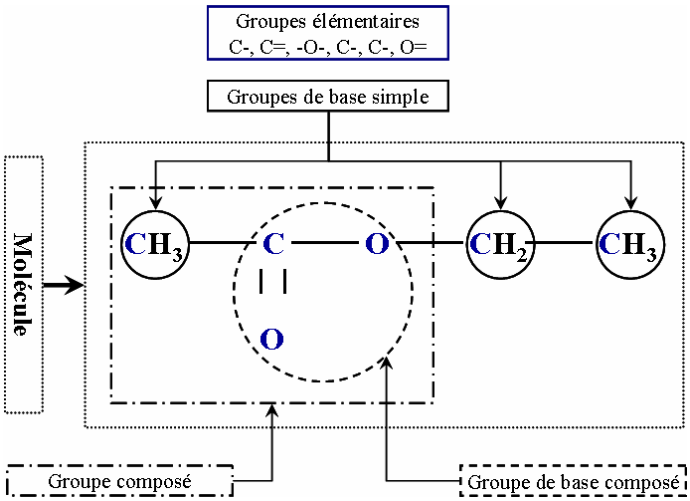

Figure (1) : Molécule et définition des groupes pour l'éthyle acétate.

Classe 01 : groupes élémentaires

Dans la première classe, on représente la molécule par un graphe moléculaire simple défini par l'équation 1. La structure d'information sur les atomes ainsi que leur voisinage sont décrits par un vecteur de données qui identifie le groupe élémentaire associé. Les éléments de ce vecteur sont présentés comme suit :

P_1	P_2	P_3	P_4	P_5	……	P_{N-1}	P_N

1. Identificateur d'atome (P_1) : il existe un nombre fini d'atomes dans les tables des méthodes de contribution de groupes. Ces atomes sont : carbone, oxygène, azote et halogène. Dans certains cas, on peut ajouter le soufre et le phosphore. Chaque atome dans la molécule est identifié par son nombre atomique. L'atome d'hydrogène est exclu de ce classement.
2. Identificateur de liaison (P_2) : Seules les liaisons carbone - carbone, carbone - non carbone et non carbone – non carbone sont codifiées. Les liaisons simple, double, triple et quadruple sont codées par 1, 2, 3 et 4 respectivement. La liaison avec l'atome d'hydrogène n'est pas attribuée.
3. Paramètre caractéristique (P_3) : Il permet de respecter la règle d'octet qui assume la neutralité de la valence de la molécule. Ce paramètre est employé dans l'équation 2, pour calculer le nombre d'atome d'hydrogène attaché à l'atome central dans le groupe élémentaire.
4. Identificateur de cycle (P_4) : Cet identificateur à pour objectif la codification des différents types de cycle dans la molécule. Il prend la valeur 0 si l'atome est non cyclique, 1 pour un cycle non aromatique et 2 pour un cycle aromatique. D'autres valeurs sont employées pour codifier certains cycles spécifiques (hétérocycle, par exemple).
5. Autres identificateurs (P_5, …, P_n) : On peut ajouter toute sorte d'information complémentaire importante pour identifier certaines spécificités des groupes élémentaires, par exemple, la position de OH dans un site primaire, secondaire ou tertiaire de la structure.

Classe 02 : groupes de base

La structure de données présentée dans la classe précédente est utilisée pour générer le vecteur des groupes de base présents dans la molécule. Ces derniers sont subdivisés en deux catégories : groupes de base simples et groupe de base composés.

1. Un groupe de base simple est définie comme un groupe élémentaire avec tous les atomes d'hydrogène attachés à l'atome central du groupe. Exemple de ce type de groupe : CH_3, CH_2, CH et OH. Un groupe de base simple est représenté par un couple (X, Y), où X caractérise la codification des groupes élémentaires définie dans la classe précédente et Y le nombre d'atomes d'hydrogène attachés à l'atome central de X calculé par l'équation 2, où NH_i est le nombre d'atomes d'hydrogène attachés à l'atome principal i, V_i^{STD} est la valence standard de l'atome principal i, a_{ij} la connexion entre l'atome i et l'atome j. Les constantes $C_i^{(P2)}$ et $C_i^{(P3)}$

correspondent à l'identificateur de liaisons et au paramètre caractéristique respectivement. L'acétate d'éthyle (voir figure 1 & tableau 1) contient ainsi trois groupes de base simple : deux CH_3 et un CH_2 décrits par (61000, 3) et (61000, 2) et la structure de données de ces groupes sont deux [1, (61000, 3)] et un [1, (61000, 2)].

$$NH_i = V_i^{STD} - \sum_{j=1, j\neq i}^{Natoms} a_{ij} - C_i^{(P2)} + C_i^{(P3)} + 1 \quad (2)$$

2. Un groupe de base composé est constitué de plusieurs groupes élémentaires associés à toutes les informations relatives à la liaison avec les atomes d'hydrogène. Ces groupes sont en nombre limité et liés à la fonctionnalité chimique de la molécule : HCOO, COOH, C=O, HC=O, C≡N et NO2. La structure de données est décrite selon la relation 3, où Index_EG indique le nombre de groupes élémentaires et $(X, Y)_i$ identifié la nature du $i^{ème}$ groupe élémentaire présent dans le groupe de base composé. La molécule d'acétate d'éthyle possède ainsi un groupe de base composé COO, lui-même composé de trois groupes élémentaires C =, O = et -O-. La structure de données devient alors [3,(62000,0),(82100,0),(82000,0)] (voir figure 1 & tableau 1).

$$\left[Index\_EG, (X,Y)_1, (X,Y)_2, ..., (X,Y)_{Index\_EG}\right] \quad (3)$$

Classe 03 : groupes composés

Dans la troisième classe, nous présentons la structure de données pour le groupe composé. Elle est basée sur les informations générées dans les classes 1 & 2, à l'aide du principe de profil de groupe et de règles heuristiques. Les groupes composés sont employés dans les méthodes de contribution de groupes multi-ordre et les modèles UNIFAC (par exemple UNIFAC Dortmund). Il n'y a pas une liste unique de ce type de groupe, mais ceux-ci varient selon la nature de la méthode d'estimation. Par exemple, nous prenons le groupe CH_3COO du modèle UNIFAC (Dortmund) qui est constitué de deux groupes : un groupe de base simple CH_3 et un groupe de base composé COO. La connexion entre ces deux groupes de base doit être assurée pour décrire le groupe composé. La structure de données relative à cette classe est liée à la relation 4, où BG_1, BG_2, …, BG_N sont des informations dérivées de la relation 3 (voir figure 1 & tableau 1).

$$\left[Index\_BG, BG_1, BG_2, ..., BG_{Index\_BG}\right] \quad (4)$$

Table 1. Structure de données pour les trios catégories de groupes

Groupe composé	Groupes de base	Groupes élémentaires	codifications
CH3COO-	CH3	C-	[1, (61000,3)]
	-COO-	C=	**[3, (62040,0),(82000,0),(82100,0)]**
		O=	
		-O-	
	{2, [1, (61000,3)]**, [3, (62040,0), (82000,0), (82100,0)]}**		

Classe 04 : structures détaillées

Enfin, et pour des applications spécifiques de conception Produits – Procédés, notamment pour l'industrie des arômes, les isomères doivent être différenciés par simulation moléculaire. L'information sur la structure moléculaire générée dans les classes précédentes (1,2 et 3) peut être raffinée pour représenter la structure tridimensionnelle en utilisant les coordonnées atomiques.

FORMALISME ENTREE – SORTIE BASÉ SUR XML

XML (e***X****tensible* ***M****arkup* ***L****anguage*) est une méthode universelle de représentation textuelle de données structurées selon une syntaxe normalisée. XML facilite la réalisation, l'échange et le stockage de fichiers qui ne soient pas ambigus et qui évitent les pièges courants, tels que la non extensibilité et la dépendance par rapport à certaines plate-formes. Dans ce travail, un formalisme entrée-sortie est construit, se basant sur deux langages XML standards connus : CML (*Chemical Markup Language*) pour la représentation de la structure moléculaire (Murray-Rouille et Rzepa, 2001)

et le ThermoML pour les propriétés expérimentales (Frenkel et al., 2003). Dans ce dernier, les données dérivées des méthodes d'estimation et de corrélations (contribution de groupes, état de correspondant, …) aussi bien que les méthodes de calcul théorique ne sont pas supportées (Frenkel et al., 2003). Celles-ci peuvent être une motivation supplémentaire pour proposer un formalisme d'entrée-sortie XML.

```
<?xml version="1.0" encoding="UTF-8"?>
<XmlFormalism>
        <Compound id="Ethyl Acetate">
          <ElementaryGroupArray>
            <elementaryGroup id="a1" atomType="6" bondType="1" hydrogenCount="0" ring="0" others="0"/>
            <elementaryGroup id="a2" atomType="6" bondType="2" hydrogenCount="0" ring="0" others="0"/>
            <elementaryGroup id="a3" atomType="8" bondType="2" hydrogenCount="1" ring="0" others="0"/>
            <elementaryGroup id="a4" atomType="6" bondType="1" hydrogenCount="0" ring="0" others="0"/>
            <elementaryGroup id="a5" atomType="6" bondType="1" hydrogenCount="0" ring="0" others="0"/>
            <elementaryGroup id="a6" atomType="8" bondType="2" hydrogenCount="0" ring="0" others="0"/>
          </ElementaryGroupArray>
          <BondArray>
                <bond id="b1" elementaryGroupRefs="a1 a2" order="1"/>
                <bond id="b2" elementaryGroupRefs="a2 a3" order="1"/>
                <bond id="b3" elementaryGroupRefs="a3 a4" order="1"/>
                <bond id="b4" elementaryGroupRefs="a4 a5" order="1"/>
                <bond id="b5" elementaryGroupRefs="a2 a6" order="1"/>
          </BondArray>
          <BasicgroupsArray>
          </BasicgroupsArray>
          <ComposedgroupsArray>
          </ComposedgroupsArray>
        </Compound>
        <PureOrMixtureProperty>
                <propertyType id="Pure"/>
                <propertyName id="BoilingTemperature"/>
                <propertyClass id="groupContribution" MethodName="GC.Joback.Ried.1987"/>
        </PureOrMixtureProperty>
</XmlFormalism>
```

Figure (2) : Fichier XML pour l' acétate d'éthyle.

Le formalisme entrée – sortie consiste en deux blocs majeurs, un bloc ″**Compound**″ et un bloc ″**PureOrMixtureProperty**″. Le premier est utilisé pour décrire la structure de données de la molécule, de la première classe (groupes élémentaires) à la dernière (structure détaillée). Le second bloc consiste en un ensemble d'éléments qui représente la structure de l'information relative aux propriétés et à leurs méthodes d'estimation. Plusieurs noms d'éléments sont proposés pour les deux blocs, à savoir, ″ElementaryGroupArray", ″BasicGroupArray″, ″ComposedGroupArray″, ″BondArray″, ″PropertyType″, ″PropertyName″ et ″PropertyClass″. D'autres éléments sont utilisés comme ″ElementaryGroup″, ″BasicGroup″, ″Bond″. Des attributs spécifiques aux noms d'éléments sont définis pour assurer la bonne description des éléments dans la structure XML, notamment ″AtomType″, ″BondType″, ″HydrogenCount″, ″Ring″, ″Order″, ″ElementaryGroupRefs″, ″HydrogenNumber″ et ″MethodName″. La figure 2 illustre la forme générale du fichier XML pour l'acétate d'éthyle ; la matrice de graphe moléculaire est uniquement représentée dans le bloc ″**Compound**″. De plus, dans le bloc ″**PureOrMixtureProperty**″, la température normale d'ébullition et la méthode de contribution de groupes de Joback et Ried (1987) est décrite. La figure 3 montre brièvement l'architecture générale de la méthodologie proposée. La structure peut être lue par différentes voies : fichier smiles, fichiers mol ou fichier utilisateur. Le fichier de données est

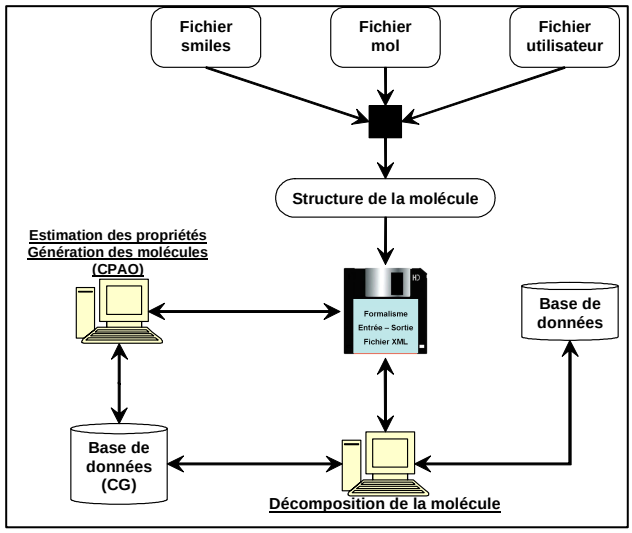

Figure (3) : Représentation schématique de la méthodologie de décomposition des molécules.

traduit au format XML. Un algorithme de parcours de graphe est utilisé pour décomposer les molécules. Néanmoins, pour générer la structure détaillée et les groupes composés, des données supplémentaires sont nécessaires.

EXEMPLE D'ILLUSTRATION

La molécule d'acétate d'éthyle est prise comme exemple d'illustration. Dans la figure 4, On schématise les différentes étapes de décomposition de l'acétate d'éthyle en groupes fonctionnels. A chaque étape, une analyse de la nature des groupes élémentaires ainsi que leur connexion est effectuée pour spécifier le point de coupure. La première étape est la recherche des groupes de base simples. Dans cet exemple le vecteur des groupes de base simples généré est {CH_3, CH_3, CH_2}. La deuxième étape correspond à l'analyse du sous groupe restant, pour identifier la structure de données du groupe de base composé. Ici une analyse des connexions est importante pour l'identification. Le groupe restant (il faut remarquer qu'il y a des connexions entre les trois groupes élémentaires restants) est COO. Dans la troisième étape, une vérification de connexions entre différents groupes de base est réalisée ; de plus la méthode d'estimation des propriétés doit être spécifiée pour identifier ce groupe. Dans ce cas il n'y a pas de groupe composé. La méthode de Joback et Ried (1987) ne possède pas ce type de groupe.

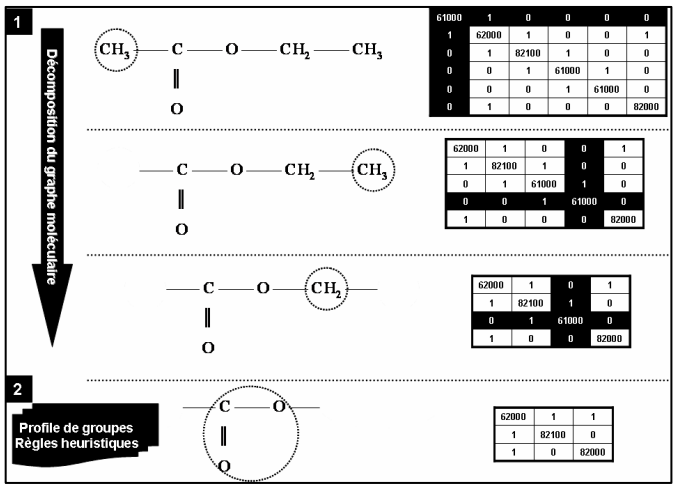

Figure (4) : schéma illustratif de la décomposition de l'éthyle acétate.

CONCLUSIONS & PERSPECTIVES

Une méthodologie multi-classes de représentation des molécules pour la conception produits – procédés est présentée. L'approche est décomposée en quatre classes. Pour chaque classe, un ensemble de méthodes et techniques de prédiction des propriétés est définie. La méthodologie est utile pour intégrer des méthodes de contribution de groupes dans les simulateurs où certaines molécules ne sont pas référencées. Un format d'entrée-sortie XML est brièvement décrit pour faciliter le stockage et l'échange des données. Cette méthodologie peut être aussi utile pour le développement de méthodes de contribution de groupes se basant sur une décomposition automatique. Dans les perspectives, un formalisme CAPE-OPEN est envisagé.

REFERENCES

Achenie, L. E. K. and Sinha, M., ***Advances in Environmental Research***, 8 (2003) 213-227.
Bünz, A. P., Braun, B. and Janowsky, R., ***Ind. Eng. Chem. Res.***, 37(1998), 3043-3051.
Constantinou, L. and Gani, R., ***AIChE Journal***, 40 (1994) 1697-1710.
Frenkel, M., Chirico, R. D., Diky, V. V., Dong, Q., Frenkel, S., Franchois, P. R., Embry, D. L., Teague, T. L., Marsh, K. N. and Wilhoit, R. C., ***J. of Chem. Eng. Data***, 48 (2003), 2-13.
Harper, P. M. and Gani, R., ***Computers and Chemical Engineering***, 24 (2000) 677-683.
Jeremy Rowley, R.., Oscarson, J. L., Rowley, R. L. and Wilding, W. V., ***J. Chem. Eng. Data***, 46 (2001), 1110-1113.
Joback, K. G.; Reid, R. C., ***Chemical Engineering Communications***, 57 (1987) 233-243.
Murray-Rust P. and Rzepa, H. S., ***J. Chem. Inf. Comput. Sci.***, 41(2001), 1113-1123.
Pogliani, L., *Chem. Rev.*, 100 (2000), 3827-3858.
Qu,D., Su, J., Muraki,M. and Hayakawat,T., ***J. Chem. Inf. Comput. Sci***, 32 (1992, a), 443-447.
Qu,D., Su, J., Muraki,M. and Hayakawat,T.***J. Chem. Inf. Comput. Sci***, 32 (1992, b), 448-452.
Smith, E.G., *Wiswesser Line-Formula Chemical Notation Method*, McGraw-Hill, NY, 1968.
Weininger, D., ***J. Chem. Inf. Comput. Sci.***, 28 (1988), 31-36.
Weininger, D., Weininger, A., Weininger, J. L., ***J. Chem. Inf. Comput. Sci.***, 29 (1989), 97-101.