

# Applications d'intelligence artificielle dans la chimie organique. XVII. Nouveaux programmes du projet SISTEMAT

Vincente de Paulo Emerenciano<sup>1</sup>, Gilberto do V. Rodrigues<sup>1</sup>, Patricia A. T. Macari<sup>1</sup>,  
Sandra A. Vesti<sup>1</sup>, Joao H. G. Borges<sup>1</sup>, J. P. Gastmans<sup>2</sup> et Denise L. G. Fromanteau<sup>2</sup>

1. Instituto de Quimica, USP, C. P. 20780, Sao Paulo, Brasil. 2. Faculdade de  
Engenharia, UNESP, 12500, Guaratingueta, S. P. Brasil

**Abstract.** This work describes the new improvements of the SISTEMAT project, one system for structural elucidation mainly in the field of Natural Products Chemistry. Some examples of the resolution of problems using <sup>13</sup>C Nuclear Magnetic Resonance and Mass Spectroscopy are given. Programs to discover new heuristic rules for structure generation are discussed. The data base contains about 10000 <sup>13</sup>C NMR spectra.

## 1. Introduction

Le SISTEMAT est un système expert qui utilise des techniques d'Intelligence Artificielle développées pour les chimistes qui travaillent avec les substances naturelles. Il pourra aussi dans le future être utilisé dans d'autres domaines tels que la pharmacologie.

Dans ce cas le Système est surtout utilisé pour la détermination de structures à partir de spectres RMN au <sup>13</sup>C, spectres de masse et données taxonomiques. Récemment nous avons publié les bases théoriques sur le système [1], et les premiers programmes qui montrent les applications qui ont été développées [2].

La plus grande différence entre notre système et les autres est la possibilité de rendre compacte des spectres et des données (botaniques, bibliographiques). La méthode pour faire la codification des formules chimiques est une autre différence, et elle est très efficace. Les programmes dits d'applications peuvent reconnaître quelques sous-structures contenues dans les formules et les utiliser dans le processus de détermination des structures ou dans les travaux taxonomiques.

Tous les programmes qui constituent le système ont été écrits employant PASCAL et FORTRAN pour les ordinateurs IBM (PC/XT/AT) qui contiennent 512 Kb de mémoire vive et un disque dur. Les programmes pourront être instaurés dans les stations de travail ou dans les ordinateurs de haut niveau qui détiennent le langage PASCAL ou FORTRAN.

Les premiers résultats dans le domaine de détermination de structures ont été publiés [1-16].

Les domaines d'intérêts du groupe sont actuellement:

- 1) Reconnaissance de modèle employant la RMN au  $^{13}\text{C}$ .
- 2) Reconnaissance de modèle employant la spectrométrie de masse.
- 3) Prédiction spectrale.
- 4) Reconnaissance de squelettes de substances naturelles employant heuristiques obtenus à partir de spectres de RMN au  $^{13}\text{C}$ .
- 5) Création automatique de structures chimiques employant des données spectrales.
- 6) Programmes qui imitent l'apprentissage humain.
- 7) Utilisation des données botaniques comme heuristique pendant les processus de détermination des structures.
- 8) Études sur la chimiotaxonomie.

## 2. Expérimental

La technique pour la création des banques de données et le développement des programmes d'applications dans le projet SISTEMAT ont été publiées [1, 2]. Par conséquent nous présenterons ici exclusivement un résumé des méthodes de création des banques de données.

Le Système travaille avec deux groupes de banques de données. Le premier est dit "banque-source". Il regroupe toutes les informations relatives aux substances publiées dans la littérature, c'est-à-dire, le code qui fait ses représentations graphiques, la bibliographie, les données qui seront indispensables à l'étude spécifique du groupe de recherche comme le lieu où la plante a été cueillie, la partie utilisée, l'activité pharmacologique, etc.

Le second groupe de banque des données est fait par le groupe système. Il est dit interlié. Il favorise une meilleure liaison parmi les données qui peuvent être employées pour les programmes d'applications comme nous le verront plus loin. En cas d'erreurs dans les banques-sources tout le système pourra être reformulé.

Une des différences clés entre notre système et les autres en cours [20, 21, 22] est la présence de l'information sur la classe chimique et sur le squelette de chaque substance. Ces informations seront de grande portée dans le processus de détermination structurale et dans les études chimiotaxonomiques et pharmacologiques.

La Codification des substances était faite manuellement, résultant en un lent processus pour la croissance des banques de données. Postérieurement nous avons développé l'autre méthode semi-automatique [14] et finalement nous sommes arrivés à un programme qui codifie automatiquement les substances par le biais d'une représentation graphique par ordinateur.

La technique de codification de toute la structure moléculaire avec les données spectrales permet, dans le cas des spectres de RMN au  $^{13}\text{C}$ , d'utiliser ces données dans un programme qui fait la confrontation spectrale d'une part et d'autre part de créer des banques de données avec des sous-structures comme le DENDRAL [20] et le DARC [21]. Dans notre système les sous-structures ressemblent aux ELCOs (Environnements limités concentriques et ordonnés) du système DARC, cependant elles contiennent aussi les représentations stéréochimiques. Ces sous-structures faciliteront la création des structures agencées de façon différente du système DARC. Au même moment, le système dans le but de maintenir les structures complètes, ne perd pas la connaissance sur le squelette et le générateur aura implicitement cette grande contrainte pour l'aider pendant le processus de création. Un fluxogramme du système est décrit dans la Figure 1.

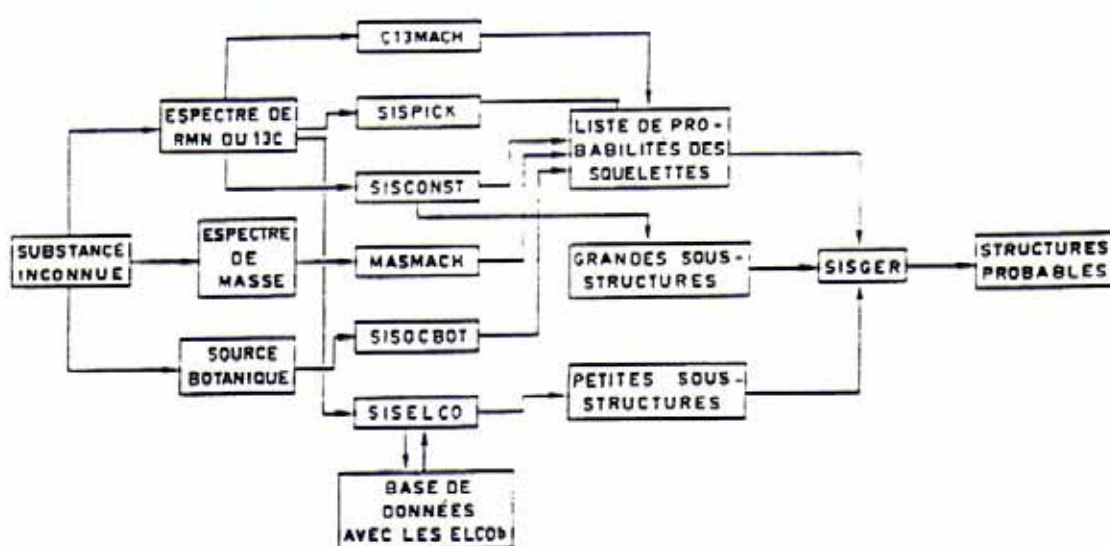


Figure 1. Fluxogramme du Système Expert SISTEMAT.

### 3. Résultats

#### *Le programme SISPICK - La reconnaissance de modèles pour l'identification des squelettes.*

Pour démontrer l'utilisation de plusieurs programmes nous avons vérifié le système avec les spectres de RMN  $^{13}\text{C}$  et de masse d'un triterpène isolé de la plante *Picrolemma granatensis* [24].

Nous avons développé [10, 19] le système PICK-UP qui détermine les règles ou modèles de RMN du  $^{13}\text{C}$  pour l'identification des squelettes. Ce système a été amélioré et est dénommé SISPICK. La nouvelle version a un algorithme pour chercher des sous-structures assujetties à des contraintes stéréochimiques et biogénétiques. Le SISTEMAT a outre la codification créée automatiquement, une autre codification identique à celle que les chimistes sont habitués d'utiliser (numération biogénétique). Cette numération donne la possibilité d'inclure des contraintes biogénétiques pendant les processus de recherche de modèles.

L'autre programme dénommé SANTIPO a été inclus dans le système SISPICK et il permet au chimiste de chercher dans un ensemble de substances qui appartiennent à un type de squelette, les carbones qui n'altèrent pas beaucoup la donation chimique ou l'hybridation (dans les cas des carbones aliphatiques). En général ce sont les carbones dont les signaux caractérisent les squelettes. En utilisant le système SISPICK nous pouvons chercher l'intervalle de déplacement chimique qui caractérise les squelettes. Ces intervalles sont utilisés comme règles heuristiques pour le SISTEMAT quand il tente de classifier un spectre. Le résultat (P4) du programme SISPICK avec le spectre de RMN au  $^{13}\text{C}$  du triterpène utilisé pour le test figure dans le Tableau 1, colonne 5.

Tableau 1

Résultats du Systeme Expert SISTEMAT apres l'analyse des données botaniques et spectrales du triterpene décrit dans la Figure 2. De P1 à P5 sont les probabilités discutées dans le texte.

ESQUELETTE	P1	P2	P3	P4	P5	Pf
QUASSINOID	99.6	15.1	50.0	100.0	1.8	55.2
22ETILQUASSINOID			50.0			8.7
LANOSTANE		19.5				5.6
URSANE		13.3				3.8
OTHERS					28.3	13.7
SECODAMARANE		9.5				2.8
OLEANE		9.5				2.7
CICLOARTANE		7.7			1.1	2.4
24METHYLSECODARANE		3.9			1.9	1.3
TIRUCALANE		2.0			1.0	0.7
16NOROUASSINOIDE		2.4				0.7
HEXANORLANOSTANE		2.1				0.6
FERNANE		2.0				0.6
27NORLANOSTANE		2.0				0.6
18NORQUASSINOID		1.9				0.6

### *Le programme SISCONST*

Ce programme a été élaboré dans le but d'indiquer les grandes sous-structures et les squelettes employant un processus très sophistiqué à partir du moment où la confrontation des spectres de RMN au  $^{13}\text{C}$  est réalisée. L'importance de ce programme pour le projet SISTEMAT réside dans le fait que comme il a été démontré [16], que le manque des sous-structures est le principal problème des générateurs du type DENDRAL [20] ou DARC [21] qui n'évitent pas le problème de l'explosion combinatoire. Nous montrerons par la suite quelques exemples d'application du SISCONST avant de l'utiliser avec le générateur en finition.

Le programme a été vérifié avec les données de RMN au  $^{13}\text{C}$  d'une substance donnée décrite à la Figure 2. Ce triterpène a le squelette quassinoid et a été judicieusement identifié (Tableau 1, colonne 1).

Le programme a aussi indiqué trois grandes sous-structures qui lorsqu'elles ont été superposées nous ont permis d'arriver à une meilleure approximation de la structure (Figure 3).

### *Le programme C13MACH*

Ce programme utilise la méthode de confrontation spectrale de BREMSER [25]. Le programme de BREMSER et la nôtre calculent les indices de ressemblance entre une substance inconnue et d'autres contenues dans les banques de données. Comme le SISTEMAT connaît les squelettes, le programme peut calculer la probabilité de la substance d'appartenir à n'importe quel type de squelette. Nous avons fait des études avec 50 spectres qui ont donné un pourcentage de réussite d'environ 70%. Dans le Tableau 1 (seconde

colonne) nous avons le résultat (P2) du programme avec le spectre de la substance décrit dans la Figure 2. Le programme détermine le squelette quassinoid de la même façon que le programme SISCONST.

#### *Le programme MASMACH*

Les spectres de masse de 800 triterpènes obtenus dans la littérature figurent dans la banque de données du SISTEMAT. Ces spectres ont été faits à partir de différents appareils qui résultent des différentes conditions de fragmentation, bien qu'il existe quelques modèles typiques qui offrent la possibilité de regroupement des squelettes. Nous avons travaillé avec l'hypothèse qu'un spectre d'une substance inconnue doit être un degré de ressemblance plus notoire avec substances d'un même squelette. Le programme a un algorithme pour faire la confrontation spectrale d'une façon presque identique à la méthode de BREMSER [25] qui confronte les spectres de RMN au  $^{13}\text{C}$ . Les résultats (P5) du programme avec le spectre de masse du triterpène utilisé pour le test figure dans le Tableau 1, colonne 5.

#### *Le programme SISOCBOT*

L'origine botanique, c'est-à-dire, la famille ou le genre, utilisés pour isoler une substance, est une heuristique importante pour identifier le type de squelette. Le programme SISOCBOT peut calculer la probabilité que la plante a un type de squelette employant les fréquences des occurrences de ces squelettes dans la famille ou dans le genre. L'information botanique n'est pas aussi importante comme les autres spectrales, et il est nécessaire que les banques de données botaniques du système soient très complètes. Dans le cas du triterpène utilisé pour le test du programme, les probabilités (P3) obtenues figurent dans le Tableau 1, colonne 3.

#### *L'union des différentes méthodes pour identifier les squelettes.*

Le SISTEMAT utilise les cinq informations obtenues à partir des programmes cités ci-dessus pour calculer une probabilité finale (Pf) qu'une substance appartienne à un squelette.

$$P_f = P_i + ((P_i + 1) \times (1 - P_i))$$

De P1 à 5 sont les cinq probabilités du Tableau 1.

Nous avons fait 30 essais avec de nouveaux triterpènes isolés récemment et par la suite nous avons attribué un poids pour chaque programme avant de calculer la probabilité finale. Pour chaque programme nous avons trouvé des poids différents: SISCONST (0.7), C13MACH (0.5), SISOCBOT (0.3), SISPICK (0.8) et MASMACH (0.2). Dans ce cas le système a indiqué la probabilité correcte au squelette quasinoid (Figure 2 et Tableau 1).

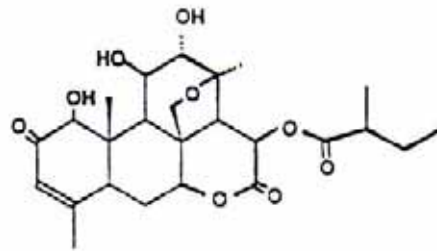


Figure 2. Un triterpène du type quasinoïde isolé de *Picroiema granatensis* (Simaroubaceae).

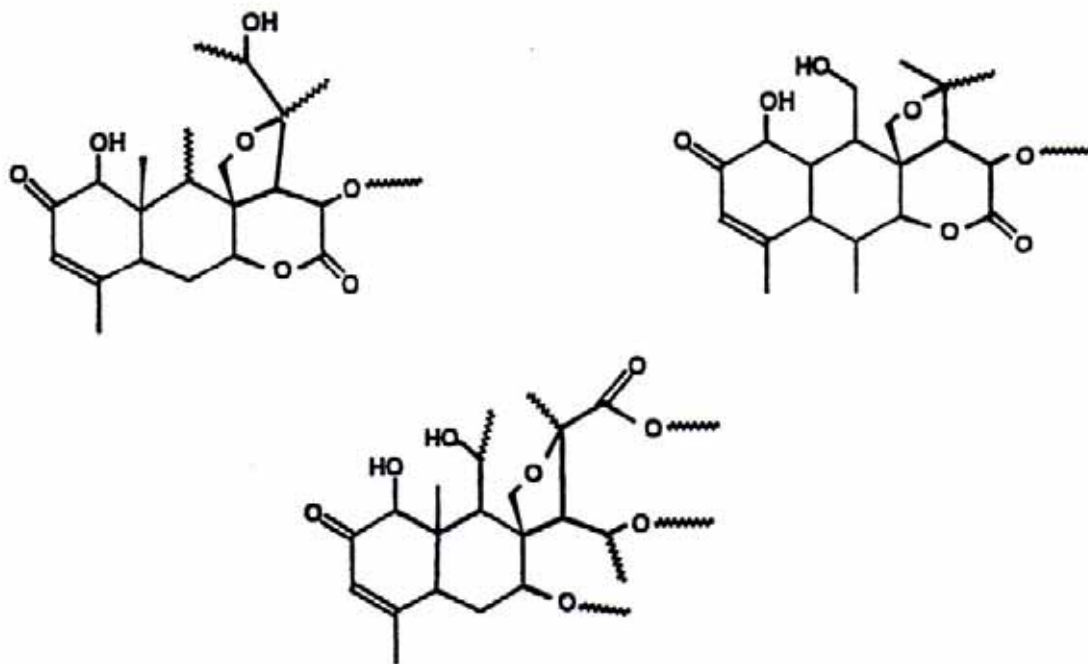


Figure 3. Trois sous-structures fournies pour le programme SISCONST après l'analyse d'un spectre de RMN au  $^{13}\text{C}$  [24].

### Le programme SISELCO

Les banques de données interliées qui contiennent les représentations topologiques des substances et les données spectrales ont été transformées en sous-structures du type ELCOb [21]. Actuellement le SISTEMAT a 10000 spectres de RMN du  $^{13}\text{C}$  qui fournissent 5000 ELCOb. Ces sous-structures contiennent aussi les descriptions stéréochimiques des carbones autour d'un foyer. Le générateur (SISGER), qui est en phase de construction, devra utiliser trois contraintes. La première, sera le squelette carbonique que le système lui-même peut indiquer. La seconde, sera les grandes sous-structures fournies par le programme SISCONST, et la troisième sera le nouveau type de ELCOb.

#### 4. Conclusion

Tout au long de ce travail nous avons décrit le stage actuel du système expert SISTEMAT. Nous avons brièvement présenté chaque programme car ils seront décrit plus en détail dans des manuscrits futures. Nous avons développé les programmes inexistantes dans les autres systèmes [20-23], c'est-à-dire, les algorithmes et les heuristiques qui peuvent chercher leurs propres contraintes, comme les squelettes et les grandes sous-structures. Avec ces mécanismes le générateur du système pourra tester les problèmes de détermination structurale de substances qui atteignent 30 atomes de carbone. A l'avenir d'autres données pourront être utilisées par le système comme la RMN du  $^1\text{H}$  et de l'Infra-Fouge (IR). Ces types de données sont utiles pour annuler des sous-structures inexacts qui proviennent d'un programme comme le SISCONST.

La simple superposition manuelle des sous-structures fournies pour le SISCONST a été suffisante pour identifier quelques molécules isolées dans notre laboratoire.

Actuellement la croissance des banques de données est faite sur la base d'agencement des spectres des classes chimiques et le système a surtout les spectres de terpènes (environ 5000) avec 15, 20 et 30 atomes de carbone. Les autres classes sont aussi bien représentées comme les flavonoides, les esteroides, les iridoïdes et les lignanes.

La méthode faisant utilisation de comprimer tous les types de données permet l'utilisation des mini-ordinateurs d'une part, et d'autre part en vue de sa simplicité, le système pourra être acheminé à d'autres universités.

L'inclusion de données botaniques dans le système est une nouveauté et elle donne plus une heuristique aux processus de détermination structurale, et éventuellement ces données pourront être utilisées pour les études taxonomiques.

#### 6. Remerciements

Nous tenons à remercier la FAPESP et CNPq pour le financement de ces travaux.

#### 7. Références

- [1] J. P. Gastmans, M. Furlan, M. N. Lopes, J. H. G. Borges et V. P. Emerenciano, *Química Nova*, **13** (1990) 10.
- [2] J. P. Gastmans, J. C. Zurita, J. S. Junior et V. P. Emerenciano, *Anal. Chim. Acta*, **217** (1989) 85.
- [3] J. P. Gastmans, V. P. Emerenciano, et M. Furlan, *Computer and Chemistry*, **12** (1988) 285.
- [4] J. P. Gastmans, M. Furlan, V. P. Emerenciano, N. F. Roque et A. C. Bussolini, *Química Nova*, **12** (1989) 25.
- [5] J. P. Gastmans, M. Furlan et V. P. Emerenciano, *Computer and Chemistry*, **14** (1990) 75.
- [6] M. N. Lopes, J. H. G. Borges, J. P. Gastmans, et V. P. Emerenciano, *Eclética Química*, **14** (1989) 69.
- [7] V. P. Emerenciano, N. F. Roque, M. Furlan, et L. M. B. Torres, *Anal. Chim. Acta*, **236** (1990) 501.
- [8] C. M. B. Maia, R. Braz Filho, et V. P. Emerenciano, *An. Acad. Bras. Ciências*, **62** (1990) 119.
- [9] P. A. T. Macari, V. P. Emerenciano, et Z. S. Ferreira, *Química Nova*, **13** (1990) 260.
- [10] A. P. Lins, M. Furlan, J. P. Gastmans, et V. P. Emerenciano, *An. Acad. Bras. Ciências*, **63** (1991) 41.
- [11] V. P. Emerenciano, A. C. Bussoline, M. Furlan, G. V. Rodrigues, et D. L. G. Fromanteau, *Spectroscopy*, **11** (1993) 95.
- [12] V. P. Emerenciano, *Química Nova*, **16** (1993) 551.
- [13] V. P. Emerenciano, G. V. Rodrigues et J. P. Gastmans, *Química Nova*, **16** (1992) 431.
- [14] V. P. Emerenciano, M. Furlan, L. Lopes, J. P. Gastmans, et L. D. Melo, *Spectroscopy*, **10** (1992) 113.

- [15] D. L. G. Fromanteau, J. P. Gastmans, S. A. Vestri, V. P. Emerenciano, et J. H. G. Borges, *Compuer and Chemistry*, **17** (1993) 369.
- [16] S. A. V. Alvaranga, *Thèse de doctorat*, University of Sao Paulo (1993).
- [17] P. A. T. Macari, *Thèse de doctorat*, University of Sao Paulo (1994).
- [18] J. H. G. Borges, résultats non publiés.
- [19] C. Djerassi, D. H. Smith, C. W. Crandell, N. A. B. Gray, J. G. Nourse et M. R. Lindley, *Pure and Applied Chemistry*, **54** (1982) 2445.
- [20] M. Carabedian, I. Gagane et J. E. Dubois, *Anal. Chem.*, **60** (1991) 2186.
- [21] B. D. Christie et M. E. Munk, *J. Amer. Chem. Soc.*, **113** (1991) 3570.
- [22] W. Bremser, *Angew. Chem. Int. Ed. Engl.*, **27** (1988) 247.
- [23] E. Rodrigues, J. B. Fernandes, P. C. Vieira, M. F. Das, et G. F. Da Silva, *Phytochemistry*, **33** (1993) 891.
- [24] W. Bremser, M. Klier, et E. Meyer, *Org. Magn. Res.*, **7** (1975) 97.