J. Braz. Chem. Soc., Vol. 14, No. 3, 369-374, 2003.
Printed in Brazil - ©2003 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Article

SISTAX – An Intelligent Tool for Recovering Information on Natural Products Chemistry

Sandra A.V. Alvarenga[a], Jean P. Gastmans[a], Marcelo J. P. Ferreira[b], Gilberto V. Rodrigues[c], Antônio J. C. Brant[b] and Vicente P. Emerenciano*[,b]

[a] *Faculdade de Engenharia de Guaratinguetá, Universidade Estadual Paulista, 12516-410 Guaratinguetá - SP, Brazil*

[b] *Instituto de Química, Universidade de São Paulo, CP 26077, 05513-970 São Paulo - SP, Brazil*

[c] *Departamento de Química, ICEx, Universidade Federal de Minas Gerais, 30161-000 Belo Horizonte - MG, Brazil*

Este trabalho descreve o desenvolvimento de um novo programa para o sistema especialista SISTEMAT, denominado de SISTAX. Este programa permite aos interessados em quimiotaxonomia realizar uma "pesquisa inteligente" de substâncias orgânicas em bancos de dados através de estruturas químicas. Quando acoplado com um eficiente sistema de códigos, este programa reconhece tipos de esqueletos e pode encontrar quaisquer restrições subestruturais solicitadas pelo usuário. Um exemplo da aplicação do programa para diterpenos encontrados em plantas é descrito.

This work describes the development of a new program, named SISTAX, for the expert system SISTEMAT. This program allows anyone interested in chemotaxonomy to carry out an intelligent search for organic compounds in databases through chemical structures. When coupled with can efficient encoding system, the program recognizes skeletal types and can find any substructural constraints demanded by the user. An example of an application of the program to the diterpene class found in plants is described.

Keywords: expert system, chemotaxonomy, substructures search, diterpenes

Introduction

The identification of substructures, parts of structures, has several applications in organic chemistry. Two research fields that apply the concept of structures are computer-assisted structure elucidation and chemotaxonomy. In both fields the implementation of computer programs involves chemists, mathematicians, computer engineers, and the interdisciplinarity of the problems results in a great challenge.

Several works found in the chemical literature explain the recognition of substructures from spectra data.[1-11] Substructures, allied to other biochemical inferences, are the main tools for chemotaxonomy methodology, and may be useful to discrimimate genera, species *etc.*[12,13] The aim of this work is to demonstrate how a specialist system developed to assist the chemist in both fields described above can be used for chemotaxonomic purposes.

To accomplish the recognition of substructures for classification purposes in chemotaxonomy and evolution, a new program named SISTAX was developed, which permits realization of a search, at a determined botanical rank (family or genera) by chemical category, such as chemical class, carbon skeleton type and functional groups. This program is stored in a database, especially built for chemical data. We show applications of this program to natural products chemistry, due to the great diversity of compounds already recorded in this field of science as well as the great number of plants chemically studied in laboratories. This program will be integrated into the expert system SISTEMAT.[6-16]

Methods

The expert system SISTEMAT

The specialist system SISTEMAT is formed by a set of programs projected to be used primarily as an auxiliary tool in natural product determination processes, and secondly for chemotaxonomic studies. The former task has been thoroughly explored by our research group.[4-16] The latter is only beginning.[17] The system allows the analysis of the

* e-mail: vdpemere@quim.iq.usp.br

stored data due to an efficient method of structure encoding rendered by SISTEMAT. The database of SISTEMAT currently has about 23000 occurrences of compounds isolated from plants of several chemical classes.[6,9-11,18,19]

The database of SISTEMAT shows all the facilities of an associated database allied to the storage of compound structures in compacted vector forms, which are transformed into connectivity matrices during the search process.[14-16] This enables anyone to recover any chemical information contained within the substance encoding. This type of codification still allows that this information is obtained quickly and simply, which has been of fundamental importance in the development of the SISTAX program.

The SISTAX program: Definitions

Skeletons are different carbon arrangements exhibited by a determined chemical class.[20] Chemical classes are large groupings of natural products possessing a common biosynthetic origin, that is, a same chemical precursor. In Figure 1, chemical classes with their respective precursors and different carbon skeletons are shown. For the chemist dealing with natural products chemistry the concept of skeletal types is frequently used for taxonomic and structural determination purposes.

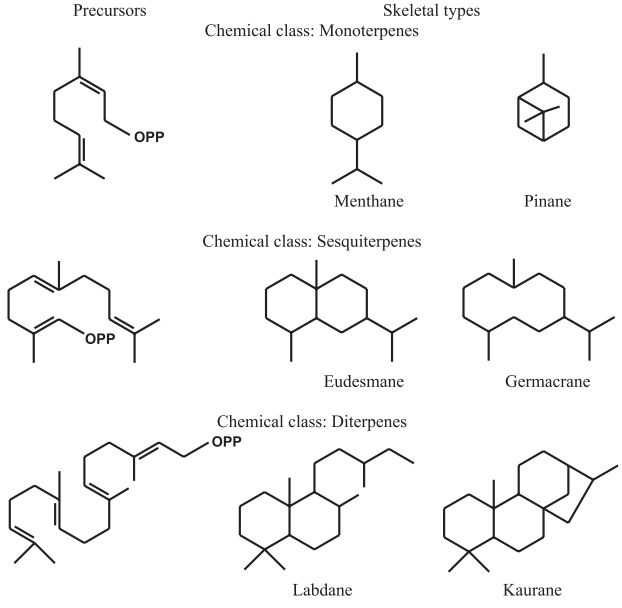


Figure 1. Some chemical classes of natural products, skeletal types and biosynthetic precursors.

The SISTAX program

The SISTAX program was developed to realize intelligent searches in SISTEMAT's database. At this moment the program has a version written in FORTRAN, with the facilities for controlling screens and data entries in PASCAL.

The first approach utilized by the chemists in this field is investigation of the distribution of the structural types (carbon skeletons or substructures) of one or several existing classes of natural products in a botanical taxon (family or genus).

The intelligent search processed by the program uses the method of encoding compounds from SISTEMAT. Through this encoding type it is possible to investigate within the connectivity matrices of a given compound information about its chemical structure. As the database containing chemical information is interlinked to others containing botanical data, one can therefore recover both types of data simultaneously. The searching processes in the database are performed quickly and simply by the user, who has only to answer the questions from the program through the encoding exhibited on the computer screen. With these answers, the researcher defines the structure types and the extent of the search relevant to his or her research. The research results are listed in tables that can be imported to statistical worksheets.

The structure types to be defined are: *Chemical class*: triterpene, diterpene, monoterpene *etc.*; *Carbon skeleton*: lupane, clerodane, menthane *etc.*; *Substructures (parts of structures)*: they can be functional groups such as hydroxyl or carbonyl and also sets of interlinked atoms such as an acetate, an aromatic or furanic ring among others.

The extent of the search can be represented by means of a flow chart (Figure 2), where the user can examine: the

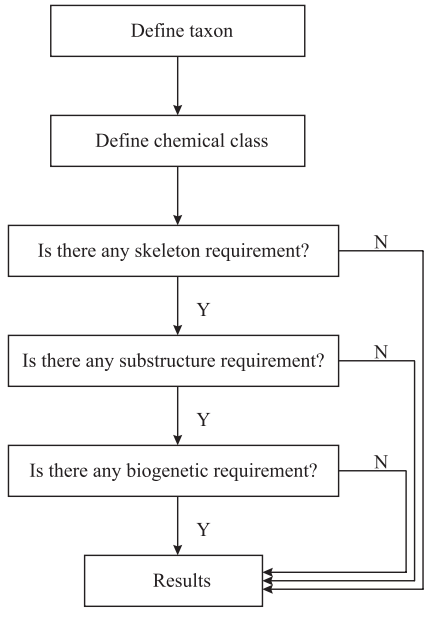


Figure 2. Flow chart of the SISTAX program.

occurrence of one or various chemical classes among the plant families and genera; the occurrence of a specific skeleton or various skeletons belonging to a chemical class; the occurrence of one or various substructures in one chemical class or on a specific skeleton belonging to a given chemical class.

The database

To evaluate the SISTAX program, SISTEMAT's database containing 2359 occurrences of diterpenes isolated from the Lamiaceae family was used. This database was built based on data published in the literature from journals indexed by Chemical Abstracts up to 1997.

Results

Verification of occurrence of a determined skeleton

In this test the occurrence of the clerodane skeleton (Figure 3) was verified among genera of the Lamiaceae family. In Figure 4 the information demanded by the SISTAX program from the user is exhibited, so that the analysis can be done. The results obtained are shown in Table 1. This

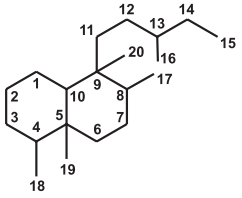


Figure 3. Biosynthetic numbering of clerodane diterpenes.

Type the database's name to record the results: **Test-4.1**
Type the code number where the biosynthetic vector is stored: **09**

Botanical constraints

The search will be made at the level of:
1. Families
2. Genera

Choose: **2**
Type the family's name: **Lamiaceae**

Structural constraints

1. Chemical classes
2. Skeletal types
3. One specific skeleton
4. Finish

Choose: **3**
Type the skeleton: **Clerodane**
Are there other constraints, Y/N ? **N**

END

Figure 4. Information requested by the SISTAX program for verification of occurrences of clerodane skeleton from Lamiaceae.

approach enables one to verify, for example, whether an accumulation of a preferential skeleton exists in some genera of a family. It is important to note that the skeleton in Figure 3 is numbered according to an arbitrary criterion adopted by the chemists, named as "biosynthetic numbering". From the computational point of view, the SISTEMAT program stores this numbering as a vector attached to the conectivity matrix of the compounds. By analyzing this encoding computationally, the biosynthetic vector permits a more precise search within the connectivity matrices, so that the user can discriminate, for instance, which, between C-6 and C-7, is methylenic (Figure 3).

Table 1. Ocurrence number of clerodane skeletal types in Lamiaceae's genera

Genus	Occurrence number
Teucrium	296
Scutellaria	97
Ajuga	78
Salvia	77
Stachys	10
Leonurus	6

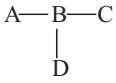
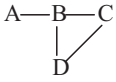
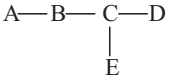
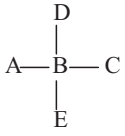
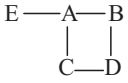
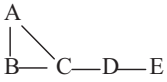
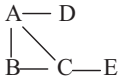
Another utility of the biosynthetic vector is searching for functional groups attached to specific positions of a carbon skeleton. Generally these groups are associated with some pharmacological properties or appear in compounds that are mainly isolated from characteristic genera of plants.[21]

Occurrence verification of a defined substructure

To carry out the search of a substructure on the SISTEMAT's data banks, a substructure code is needed, that is, it is necessary to define the size and type of existing atoms in that substructure, whose presence is to be searched in the connectivity matrices. The possible substructures are presented in Table 2 and the chemical groupings in Table 3, wherein it is feasible to select a substructure and the desired chemical groupings. As an example, we show in Table 4 the encoding for a furanic ring that may be present in clerodane diterpenes.

The aim of this test is to verify the occurrence of the furan ring in clerodanes from among the genera of the Lamiaceae family. As a demand, it was established that the furan ring should be located at carbons 13-16, according to biogenetic numbering, which is a numbering often used by natural product chemists for the clerodane skeleton (Figure 3). In Figure 5, the results of the search for the furan ring requested by the user through the SISTAX program are exhibited. Table 5 summarizes the results obtained through the analysis carried out by the program,

Table 2. The substructure sets used by SISTEMAT

Substructures	Label*	Substructures	Label*
A	01	A—B	02
A—B—C	03	A—B—C—D	05
	04		06
	07		08
A—B—C—D—E	09		10
	11		12
	13		14
	15		16
	17		

* - The labels are in agreement with the programming code.

that is, discriminating family, genera, the number of compounds from the clerodane skeleton having a furan ring at carbons 13-16.

Verification of oxidation in specifics

The SISTAX program permits to verify whether a determined position in a skeleton type, a taxon, shows oxidation more frequently than another position does. For example, one can search for the occurrence of CH_2 groups at C-6 and C-7 in clerodanes (Figure 6). The results are presented in Table 6, where one can see that in clerodanes from *Teucrium*, *Scutellaria* and *Ajuga*, C-6 is more frequently oxidized than C-7.

Conclusions

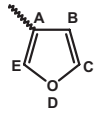
With the SISTAX program development, the expert system SISTEMAT acquires a new tool that allows the search for requirements such as chemical classes, carbon skeletons and substructures at a determined level in botanical classification. This program permits to correlate botanical information with chemical constraints. Thus, the results obtained can help forthcoming chemosystematic and evolutive studies. Since chemosystematics and evolution papers usually comprise studies on occurrences of compounds at several hierarchical levels[13], the SISTAX program may be seen as a powerful computer program at the basic step of chemotaxonomic tasks.

Table 3. Atomic groupings used by SISTEMAT

Atomic grouping	Label	Atomic grouping	Label	Atomic grouping	Label
$-CH_3$	01	=C=	12	-F	23
$-CH_2-$	02	=O	13	-Cl	24
$-\overset{\vert}{C}H-$	03	-OH	14	-Br	25
$-\overset{\vert}{\underset{\vert}{C}}-$	04	-O-	15	-I	26
$=CH_2$	05	$-NH_2$	16	-SH	27
=CH-	06	-NH-	17	-S-	28
$-\underset{\vert}{C}=$	07	$-\overset{\vert}{N}-$	18	=S	29
TCH-	08	=NH	19	$-\overset{\vert}{\underset{\vert}{S}}=$	30
TC-	09	=N-	20	$=\overset{\vert}{\underset{\vert}{S}}=$	31
HC*	10	TN	21	P	32
C*	11	N*	22		

T = represents a triple bond and * = represents the aromatic ring.

Table 4. Furan ring code

Substructures	Code
	Substructure code: 12 Atomic group codes: 07 06 06 15 06 Bond codes: 01 02 01 01 02

Codes for bonds: Simple = 01; Double = 02; Triple = 03.

Table 5. Occurrence number of clerodanes with a furan ring in Lamiaceae's genera

Genera	Occurrence number
Salvia	42
Teucrium	292

Table 6. Ocurrence number of CH_2 groups at C-6 and C-7 in clerodane skeletal types in Lamiaceae's genera

Genus	CH_2 at C-6	CH_2 at C-7
Teucrium	4	248
Scutellaria	0	86
Ajuga	0	78
Salvia	68	44
Stachys	10	4
Leonurus	6	6

Type the database's name to record the results: **Test-4.2**
Type the code number where the biosynthetic vector is stored: **09**

Botanical constraints

The search will be made at the level of:
1. Families
2. Genera

Choose: **2**
Type the family's name: **Lamiaceae**

Structural constraints

1. Chemical classes
2. Skeletal types
3. One specific skeleton
4. Finish

Choose: **3**
Type the skeleton: **Clerodane**
Are there other constraints, Y/N ? **Y**

First condition about the Clerodane skeleton:
Type the substructure code: **12**
Type the atomic group codes: **0706061506**
Type the bond codes: **0102010102**
Biosynhetic constraints: the X atom must have the number Y:
Type X and Y in order nI2, or <ENTER> **0113**
Are there other constraints, Y/N? **N**

Imposed constraint:
The atom 1 must have the number 13

END

Figure 5. Information requested by the SISTAX program for verification of occurrences of a furan ring in the clerodane skeleton from Lamiaceae.

Type the database's name to record the results: **Test-4.3**
Type the code number where the biosynthetic vector is stored: **09**

Botanical constraints

The search will be made at the level of:
1. Families
2. Genera

Choose: **2**
Type the family's name: **Lamiaceae**

Structural constraints

1. Chemical classes
2. Skeletal types
3. One specific skeleton
4. Finish

Choose: **3**
Type the skeleton: **Clerodane**
Are there other constraints, Y/N ? **Y**

First condition about the Clerodane skeleton:
Type the substructure code: **01**
Type atomic group codes: **02**
Type the bond codes: **01**
Biosynthetic constraints: the X atom must have the number Y:
Type X and Y in order nI2, or <ENTER> **0106**
Are there other constraints, Y/N? **N**

Imposed constraint:
The atom 1 must have the number 6

END

Figure 6. Information requested by the SISTAX program for verification of occurrence of an atomic grouping (CH_2) at a detremined position (C-6) in clerodanes from Lamiaceae.

At this time, correlations between hundreds of genera and, for example, dozens of chemical constraints are a task impossible to be carried out without a computer-assisted tool.

Acknowledgements

This work was supported by grants from FAPESP and CNPq.

References

1. Tomellini, S.A.; Hartwick, R.A.; Stevenson, J.M.; Woodruff, H.D.; *Anal. Chim. Acta* **1984**, *162*, 227.
2. Gray, N.A.B.; *Computer-Assisted Structure Elucidation*; Jonh Wiley Sons, Inc.: New York, 1986.
3. Funatsu, K.; Miyabayashi, N.; Sasaki, S.-I.; *J. Chem. Inf. Comp. Sci.* **1988**, *28*, 18.
4. Carabedian, M.; Dagane, I.; Dubois, J.E.; *Anal. Chem.* **1988**, *60*, 2186.
5. Munk, M.E.; Christie, B.D.; *Anal. Chim. Acta* **1989**, *216*, 57.
6. Emerenciano, V.P.; Bussolini, A.C.; Rodrigues, G.V.; *Spectroscopy* **1993**, *11*, 95.
7. Fromanteau, D.L.G.; Gastmans, J.P.; Vestri, S.A.; Emerenciano, V.P.; Borges, J.H.G.; *Comp. Chem.* **1993**, *17*, 369.
8. Emerenciano, V.P.; Rodrigues, G.V.; Macari, P.A.T.; Vestri, S.A.; Borges, J.H.G.; Gastmans, J.P.; Fromanteau, D.L.G.; *Spectroscopy* **1994**, *12*, 91.
9. Macari, P.A.T.; Gastmans, J.P.; Rodrigues, G.V.; Emerenciano, V.P.; *Spectroscopy* **1994/1995**, *12*, 139.
10. Emerenciano, V.P.; Melo, L.D.; Rodrigues, G.V.; Gastmans, J.P.; *Spectroscopy* **1997**, *13*, 181.
11. Alvarenga, S.A.V.; Gastmans, J.P.; Rodrigues, G.V.; Emerenciano, V.P.; *Spectroscopy* **1997**, *13*, 227.
12. Seaman, F.C.; Funk, V.A.; *Taxon* **1983**, *32*, 1.
13. Gottlieb, O.R.; *Micromolecular Evolution, Systematics and Ecology*; Springer-Verlag: Berlin, 1982.
14. Gastmans, J.P.; Furlan, M.; Lopes, M.N.; Borges, J.H.G.; Emerenciano, V.P.; *Quim. Nova* **1990,** *13*, 10.
15. Gastmans, J.P.; Furlan, M.; Lopes, M.N.; Borges, J.H.G.; Emerenciano, V.P.; *Quim. Nova* **1990**, *13*, 75.
16. Emerenciano, V.P.; Rodrigues, G.V.; Gastmans, J.P.; *Quim. Nova* **1993**, *16*, 431.
17. Alvarenga, S.A.V.; Gastmans, J.P.; Rodrigues, G.V.; Moreno, P.R.H.; Emerenciano, V.P.; *Phytochemistry* **2001**, *56*, 583.
18. Lins, A.P.; Furlan, M.; Gastmans, J.P.; Emerenciano, V.P.; *An. Acad. Bras. Cienc.* **1991**, *63*, 141.
19. Ferreira, M.J.P.; Emerenciano, V.P.; Linia, G.A.R.; Romoff, P.; Macari, P.A.T.; Rodrigues, G.V.; *Prog. Nucl. Magn. Reson. Spec.* **1998**, *33*, 153.
20. Emerenciano, V.P.; Ferreira, M.J.P.; Branco, M.D.; Dubois, J.E.; *Chemom. Intel. Lab. Syst.* **1998**, *40*, 83.
21. Alvarenga, S.A.V.; Rodrigues, G.V.; Gastmans, J.P.; Emerenciano, V.P.; *Nat. Prod. Lett.* **1995**, *7*, 133.
22. Magri, F.M.M.; Militão, J.S.L.; Ferreira, M.J.P.; Brant, A.J.C.; Emerenciano, V.P.; *Spectroscopy* **2001**, *15*, 99.
23. Ferreira, M.J.P.; Rodrigues, G.V.; Emerenciano, V.P. *Can. J. Chem.* **2001**, *79*, 1915.
24. Ferreira, M.J.P.; Costantin, M.B.; Sartorelli, P.; Rodrigues, G.V.; Limberger, R.; Henriques, A.T.; Kato, M.J.; Emerenciano, V.P. *Anal. Chim. Acta* **2001**, *447*, 125.
25. Ferreira, M.J.P.; Oliveira, F.C.; Alvarenga, S.A.V.; Macari, P.A.T.; Rodrigues, G.V.; Emerenciano, V.P.; *Comp. Chem.* **2002**, *26*, 601
26. Ferreira, M.J.P.; Alvarenga, S.A.V.; Macari, P.A.T.; Rodrigues, G.V.; Emerenciano, V.P.; *Biochem. Syst. Ecol.* **2003**, *31*, 25.

Received: December 17, 2001
Published on the web: February 26, 2003

FAPESP helped in meeting the publication costs of this article.