A INTELIGÊNCIA ARTIFICIAL APLICADA À QUÍMICA DE PRODUTOS NATURAIS. O PROGRAMA SISTEMAT

PARTE II - ORGANIZAÇÃO DO PROGRAMA E APLICATIVOS

Jean P. Gastmans[a], Maysa Furlan[b], Márcia N. Lopes[b], João H. G. Borges[b], Vicente de P. Emerenciano[c]

[a]*Faculdade de Engenharia - UNESP - C.P. 205, CEP 12500 - Guaratinguetá (SP)*
[b]*Instituto de Química - UNESP - Araraquara - (SP)*
[c]*Instituto de Química - USP - S. Paulo (SP)*

Recebido em 22/02/89; cópia revisada em 18/10/89

ABSTRACT

The fluxogram and framework of an expert system for Natural Products Chemistry (SISTEMAT) and three programs based on this expert system are described.

The first (PICKUPS) searches atomic subgraphs and analyses its ^{13}C spectra. The second (SISBOTA) analyses botanic family or genus and calculates the histograms family or genus towards class, skeleton or oxidation index. The last (SISSTRUC) searches atomic substructures from NMR ^{13}C spectra.

INTRODUÇÃO

Baseados nos princípios teóricos enunciados na parte I[1], desenvolvemos um sistema especialista em Química dos Produtos Naturais que chamamos de SISTEMAT.

Ao contrário dos demais sistemas já existentes (ver Parte I), tentamos impor a maior flexibilidade possível ao SISTEMAT de tal forma que ele possa, potencialmente, abordar qualquer problema de interesse do químico trabalhando em Produtos Naturais, tanto na parte de determinação estrutural quanto espectroscópica ou de quimiossistemática.

Devido à complexidade do assunto, resolvemos deliberadamente dividir o problema em duas partes. A primeira cuida da criação dos bancos de dados e das suas interligações.

A segunda trata de aplicativos que são conjuntos de programas que permitem ao computador responder a perguntas específicas.

CRIAÇÃO DOS BANCOS - SISTEMAT

Ao invés de desenvolver um único programa para a criação dos bancos preferimos subdividir esta tarefa em 5 programas distintos pelas seguintes razões:

1. Possibilidade de verificação e correção dos dados a medida que são fornecidos;

2. separar as operações automáticas, que o computador pode realizar sozinho, das outras nas quais a presença do operador é indispensável;

3. criar bancos fontes a partir dos quais pode-se modificar dados e rearquiteturar todos os bancos, sem ter que digitá-los.

Resumidamente, os programas são os seguintes:

- VERISIS: verifica se a substância já existe no banco, neste caso os vetores reduzidos e gráficos não precisam ser elaborados, agilizando assim o processo. Ele inicia também os bancos fonte.

- FONTESIS: recolhe os vetores e origens botânicas (se tiver). Ele verifica se os dados fornecidos são corretos e calcula o índice de oxidação. Este processo é semi automático. Antes de iniciar o cálculo do índice de oxidação, o computador retira da substância em estudo todos os macronós e os grupos metila das funções metoxiladas, contudo ele é ainda incapaz de retirar os demais carbonos existentes e que não pertencem ao esqueleto (p. ex. grupos prenila...). Esses átomos têm que ser assinalados pelo operador. Estamos trabalhando atualmente numa segunda versão na qual o cálculo do índice de oxidação será automático.

- NUMSIS: calcula o identificador. O processamento é automático.

- ARQUISIS: arquitetura os bancos. Este programa é automático.

- DTSIS: recolhe os dados físico-químicos ou outros que o computador não pode deduzir por si próprio. Ele elabora também os bancos fonte. Os bancos fonte são bancos seqüenciais não arquiteturados que contêm todos os dados fornecidos.

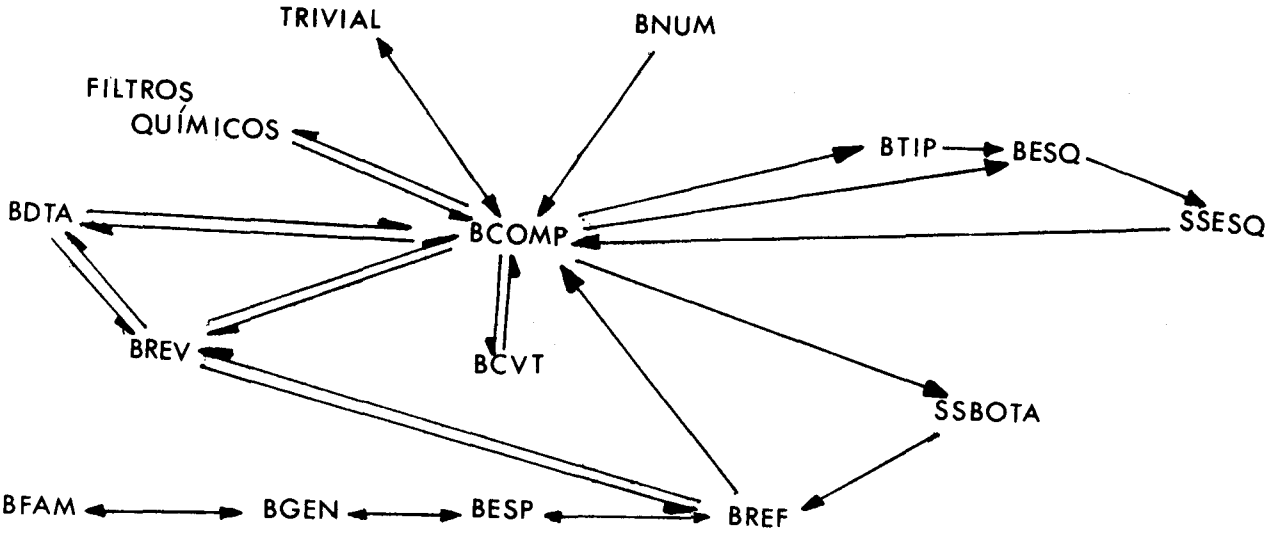
A esses cinco programas, acrescentamos um sexto, "MISTAKE" que serve para rearquiteturar todos os bancos a partir dos bancos fonte. Este programa é automático. Ele se tornou necessário porque freqüentemente as propostas estruturais são modificadas.

A arquitetura e interligações dos bancos que compõem o SISTEMAT estão representadas na fig. 1.

O sistema foi desenvolvido levando em conta os seguintes fatores:

1. a compactação. Isso foi conseguido usando uma aritmética de base 222, aproveitando os termos ASCII de 33 em diante e evitando ao máximo a repetição dos mesmos dados. Isso explica o grande número de bancos interligados existentes.

Para se ter uma idéia, o banco teste sobre o qual trabalhamos é previsto para 3.000 substâncias e ocupa menos que 1 Mb.



BCOMP é o banco dos compostos. Cada registro de 84 bytes contém:

- 70 bytes para vetor
- 02 bytes para a seta do registro de BCVT onde o resto de vetor é armazenado, se este for maior que 70 bytes
- 01 byte para o registro de BTIP onde a classe do composto é armazenado
- 02 bytes para o registro de BESQ onde o nome do esqueleto é armazenado
- 02 bytes para o registro do composto seguinte pertencente ao mesmo identificador
- 03 bytes para o registro de BDTA onde os dados físico-químicos são armazenados
- 02 bytes para o registro do banco TRIVIAL onde o nome comum do composto é armazenado
- 02 bytes para o índice de oxidação calculado durante o programa FONTESIS

BNUM é um banco de 100.000 registros de 2 bytes cada um, onde se encontra o registro do primeiro composto pertencente a um identificador.

BFAM, BGEN, BESP são os bancos das famílias, gêneros e espécies botânicas com setas e retrosetas.

BREF é o banco das referências botânicas. Cada registro de 12 bytes contém:

- 03 bytes para a retroseta em direção à BESP
- 01 byte para o registro de BREV, onde o título de revista é armazenado
- 01 byte para o ano
- 02 bytes para a página
- 03 bytes para a seta do registro de continuação
- 02 bytes para a seta do registro de BCOMP

SSEQ e SSBOTA são bancos de procura randômica. A sua existência não é essencial, mas agiliza sobremaneira procuras dentro de BCOMP e de BREF. Sem eles a procura teria que ser seqüencial.

Figura 1 Arquitetura do SISTEMAT

2. a velocidade de execução. Isso foi conseguido criando bancos de acesso direto. Uma vez definido o problema a ser resolvido, o computador vai diretamente para o(s) registro(s) onde a(s) reposta(s) se encontra(m).

3. a flexibilidade do sistema. Potencialmente toda pergunta para a qual o enunciado e a resposta são interligados por setas ou retrosetas pode ser respondida pelo computador. Por exemplo, "encontrar os espectros de ^{13}C dos clerodanos, com pelo menos 2 grupos carbonila, das substâncias isoladas da Família Compositae", é uma tarefa realizável (ver programa SISBOTA adiante).

APLICATIVOS

Desenvolvemos até agora 3 aplicativos que apresentaremos a seguir:

A - PICKUPS

Este aplicativo é uma versão melhorada do PICKUP[2] desenvolvido recentemente. Ele se destina principalmente ao espectroscopista e realiza a seguinte tarefa: encontrar as faixas de sinais de ^{13}C de agrupamentos atômicos e verificar se esses sinais são realmente característicos desses agrupamentos. Ele é subdividido em 3 programas:

– O primeiro retira do SISTEMAT, os vetores, os sinais e as referências bibliográficas de todas as substâncias, ou daquelas pertencentes a uma classe ou esqueleto, se assim desejar. A rigor esta primeira fase não seria necessária mas observamos que quando o PICKUPS é usado com uma certa freqüência obtém-se um sensível ganho de tempo procedendo desta forma.

– O segundo é o PICKUPS propriamente dito.

– Impõem-se em primeiro lugar os pré-requisitos químicos. Esses requisitos são definidos pelos subgrafos (Tabela I), pelos pesos dos nós e vértices (Tabela II) e pelo número de vezes que cada requisito tem que ser obedecido.

Pode-se impor também, e isso é o melhoramento introduzido, que os subgrafos estejam interligados. Por exemplo, impondo-se como requisito o ângulo químico (2a) da fig. 2, pode-se supor 2 requisitos (2b e 2c) e exigir que os átomos 5 e 2 do subgrafo 2b ligados através de ligações simples respectivamente aos átomos 1 e 2 do subgrafo 2c.

TABELA I

Subgrafos do PICKUPS

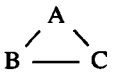
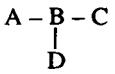
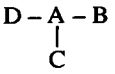
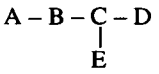
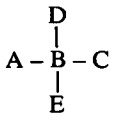
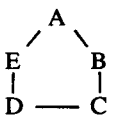
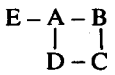
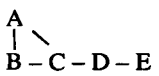
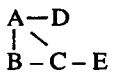
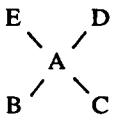
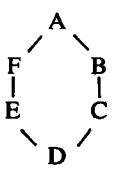
GRUPO	CÓDIGO
A	01
A – B	02
A – B – C	03
	04
A – B – C – D	05
	06
	07
	08
A – B – C – D – E	09
	10
	11
	12
	13
	14
	15
	16
	17

TABELA II

Códigos dos nós e vértices

CÓDIGOS DOS NÓS

$-CH_3$	:	01	–NH–	:	17
$-CH_2-$	:	02		:	18
–CH–	:	03	=NH	:	19
	:	04	=N–	:	20
$=CH_2$	:	05	=N	:	21
=CH–	:	06	N_{ar}	:	22
	:	07	–F	:	23
≡CH	:	08	–Cl	:	24
≡C–	:	09	–Br	:	25
HC_{ar}	:	10	–I	:	26
C_{ar}	:	11	–SH	:	27
=C=	:	12	–S–	:	28
=O	:	13	=S	:	29
–OH	:	14		:	30
–O–	:	15		:	31
$-NH_2$	:	16	$P_{(qualquer)}$	:	32

CÓDIGO DOS VÉRTICES

ligação simples ou aromática : 01
ligação dupla : 02
ligação tripla : 03

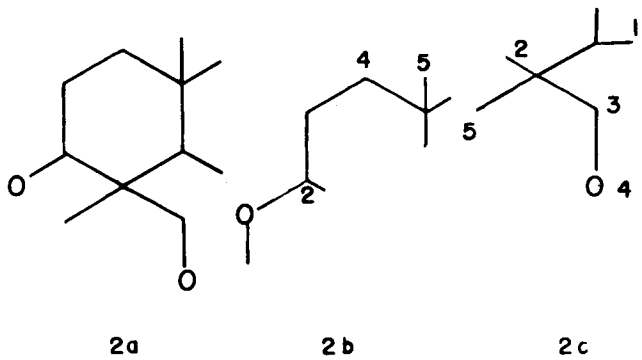
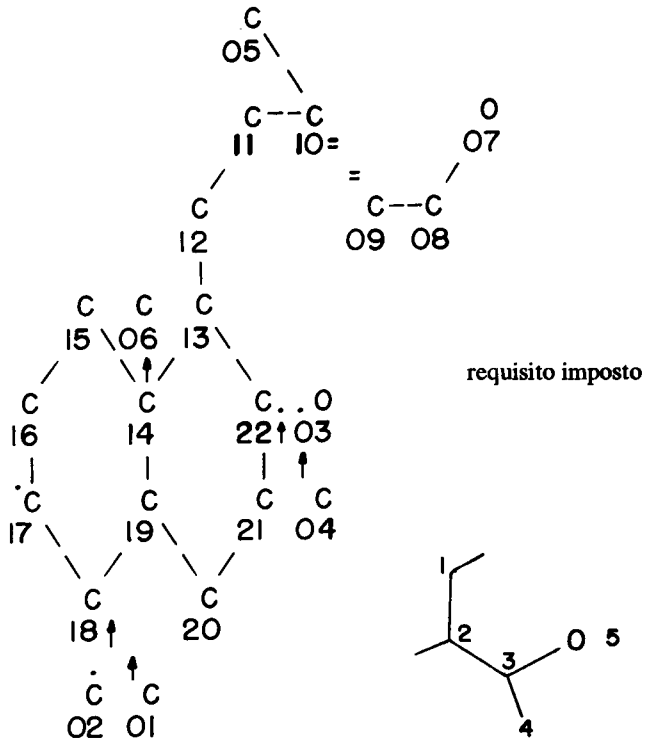


Figura 2. *Interligações de subgrafos*

Conditions remalies par
13E–2ABD–13–ENE–8a, 15–DIOL
registro número 68 1985 p. 2991
referência Phytochem.



requisito imposto

Condition número 1

carbone 12 absorption = 23.7 trip.
carbone 13 absorption = 61.2 dubl.
carbone 22 absorption = 74.1 sing.
carbone 4 absorption = 24.0 quad.
atome 3 0

TRAITEMENT

número de composes = 7
condition número 1
atome 1 trip.
moyenne = 23.5 max.= 25.1 min.= 20.5
deviation = .55 cas = 7

atome 2 dubl.
moyenne = 61.1 max.= 61.9 min.= 59.5
deviation = .28 cas = 7

atome 3 sing.
moyenne = 74.7 max.= 77.0 min.= 73.0
deviation = .42 cas = 7

atome 4 quad.
moyenne = 23.4 max.= 25.2 min.= 19.3
deviation = .66 cas = 7

atome 5 0

Fig. 3. Estrutura encontrada a partir do requisito imposto

Pode-se impor até 9 requisitos, cada um podendo ser repetido até 9 vezes.

De posse dos requisitos impostos, o computador elabora a matriz topológica e verifica se os requisitos são obedecidos. Se forem, os dados são armazenados e após a análise de todas as substâncias, ele as desenha e faz o tratamento estatístico dos seus sinais. Dependendo do tipo de micro usado, a análise de uma substância leva quase 2 segundos, num XT, 1.5 segundos num PC com processador aritmético, e menos de 1 segundo num AT, o que é ainda aceitável.

Transcrevemos uma parte da listagem operacional na fig. 3, junto com o requisito imposto.

O terceiro faz o trabalho inverso do PICKUPS; isto é, a partir dos valores máximos, mínimos e da multiplicidade dos sinais ele desenha e lista as substâncias que obedecem a esses requisitos espectrais. Pode-se assim determinar se um conjunto de sinais é realmente característico de um tipo de agrupamento atômico ou não. Existem duas variantes deste programa: a variante chamada de rígida (PICKRVR) e flexível (PICKRVF). As duas vão listar os mesmos compostos mas a atribuição dos sinais difere segundo os preceitos lógicos usados. A versão rígida obedece aos seguintes preceitos:

a. Se um único sinal for atribuído a um carbono os demais não podem lhe ser atribuídos,
b. se um sinal é atribuído a um único carbono, este sinal não pode ser atribuído aos demais.

A versão flexível utiliza exclusivamente o primeiro preceito.

B - SISBOTA

Trata-se de um programa de procura dentro do SISTEMAT, a partir da definição botânica e através de vários filtros. Após selecionar as substâncias, o computador pode fornecer também os histogramas família ou gênero (ou todos os gêneros da família) contra índice de oxidação e/ou esqueleto e/ou classe.

Após definir a família, o gênero (ou todos) e a espécie (ou todas), definem-se os filtros desejados. Esses filtros são:

1. a classe
2. o esqueleto
3. o peso molecular (superior, igual ou inferior a um número dado)
4. a fórmula bruta (nº de carbonos, hidrogênios, etc...)
5. o índice de oxidação
6. os requisitos químicos (os mesmos anteriormente definidos no PICKUPS).

Finalmente estabelecem-se os dados que se quer do computador e os histogramas desejados.

Uma parte da listagem operacional encontra-se na fig. 4.

Pode-se ver que este programa mobiliza praticamente a totalidade dos bancos; apesar disto o processamento é muito rápido, sendo limitado pela velocidade da impressora.

Este programa que se destina principalmente ao quimiossistemático depende evidentemente da riqueza do banco de dados. A listagem acima, feita a partir de um banco de teste, só tem valor demonstrativo da operação do sistema.

C - SISSTRUC

Este programa destina-se especificamente ao químico trabalhando em determinação estrutural. Ele propõe e desenha

Tecle a família 25A COMPOSITAE
Tecle o gênero
Se forem todos tecle ALL ALL
Tecle a espécie
Se forem todas tecle ALL ALL

REQUISITOS

1. classe-esqueleto
2. peso molecular
3. índice de oxidação
4. fórmula bruta
5. requisitos químicos

Se terminou os requisitos tecle (ENTER)
Escolha-II 1

Tecle a classe – 20A DITERPENE
Tecle o esqueleto
Se forem todos tecle ALL LABDANE

HISTOGRAMA COMPOSITAE
CONTRA ÍNDICE DE OXIDAÇÃO

A unidade padrão de gradiente é 0.001
Quantas unidades você vai usar? 50

De	−1.6000 até	−1.5500	1 caso
De	−1.5500 até	−1.5000	2 casos
De	−1.5000 até	−1.4500	0 casos
De	−1.4500 até	−1.4000	0 casos
De	−1.4000 até	−1.3500	0 casos
De	−1.3500 até	−1.3000	0 casos
De	−1.3000 até	−1.2500	3 casos
De	−1.2500 até	−1.2000	6 casos
De	−1.2000 até	−1.1500	1 caso
De	−1.1500 até	−1.1000	0 casos
De	−1.1000 até	−1.0500	0 casos
De	−1.0500 até	−1.0000	2 casos

Quer mudar de gradiente S/N? S
A unidade padrão de gradiente é 0.001
Quantas unidades você vai usar? 25

De	−1.6000 até	−1.5750	1 caso
De	−1.5750 até	−1.5500	0 casos
De	−1.5500 até	−1.5250	0 casos
De	−1.5250 até	−1.5000	0 casos
De	−1.5000 até	−1.4750	0 casos
De	−1.4750 até	−1.4500	0 casos
De	−1.4250 até	−1.4000	3 casos
De	−1.4000 até	−1.3750	6 casos
De	−1.3750 até	−1.3500	1 caso
De	−1.3500 até	−1.3250	0 casos
De	−1.3250 até	−1.3000	0 casos
De	−1.3000 até	−1.2750	3 casos
De	−1.2750 até	−1.2500	0 casos
De	−1.2500 até	−1.2250	0 casos
De	−1.2250 até	−1.2000	6 casos
De	−1.2000 até	−1.1750	0 casos
De	−1.1750 até	−1.1500	1 caso
De	−1.1500 até	−1.1250	0 casos
De	−1.1250 até	−1.1000	0 casos
De	−1.0000 até	−1.0750	0 casos
De	−1.0750 até	−1.0500	0 casos
De	−1.0500 até	−1.0250	0 casos
De	−1.0250 até	−1.0000	2 casos

Figura 4

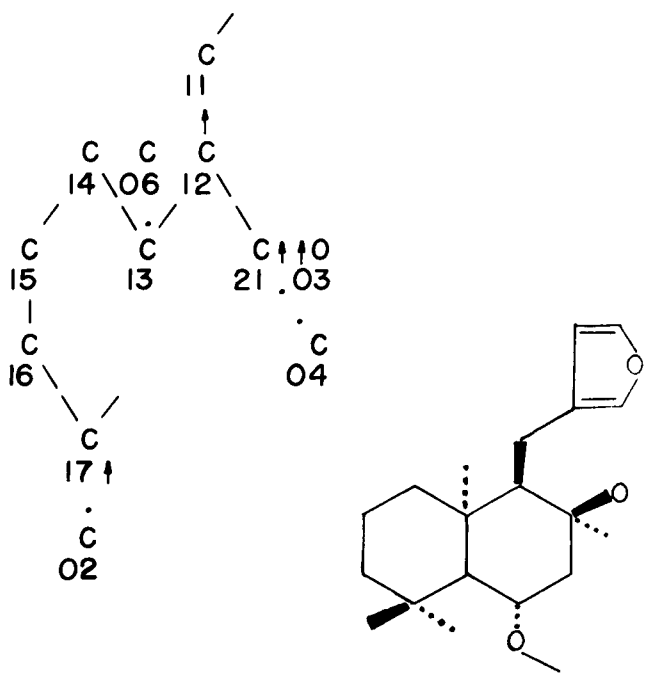
Multiplicidade? 2
Sinal 20 valor? 167.3
Multiplicidade? 1
Sinal 21 valor? 50.7
Multiplicidade? 4
O número padrão de maching é: 11
Quer mudar S/N – S
Tecle o número minimum de maching 12 09
A procura se fará
1. sobre todos os compostos
2. sobre os compostos de uma classe
3. sobre os compostos de um esqueleto

escolha 1

Procura terminada
Selecione 11 estruturas
O maior número de átomos foi: 12

Quantos átomos deseja?
Se terminou tecle [ENTER] 12

Ref. Phytochem. (1985) p. 1789



Carbono 2 experimental : 21.500
Carbono 17 experimental : 33.200
Carbono 16 experimental : 42.000
Carbono 15 experimental : 18.500
Carbono 14 experimental : 39.700
Carbono 13 experimental : 39.200
Carbono 6 experimental : 15.500
Carbono 12 experimental : 61.200
Carbono 21 experimental : 74.300
Carbono 11 experimental : 24.700
Carbono 4 experimental : 24.100

Figura 5

subestruturas suscetíveis de serem responsáveis pelos sinais de ^{13}C de uma substância desconhecida.

Trata-se de um melhoramento sensível com respeito aos sistemas de Bremser e Clerc.

Numa primeira fase, o sistema extrai do SISTEMAT um banco formado pelo vetor, os sinais e as multiplicidades calculadas. De novo, esta fase não seria, a rigor, necessária, mas como esses dados vão ser usados muitas vezes, obtém-se um ganho de tempo apreciável.

Como nos sistemas de Bremser[3] e Clerc[4], ele compara os sinais da substância desconhecida com aqueles de cada composto contido no banco. Contudo, se desejar, ele pode restringir sua procura a uma determinada classe ou esqueleto, o que torna o processo bem mais rápido.

A diferença consiste no seguinte; após determinar os sinais comuns, o computador procura quais são os carbonos responsáveis por esses sinais, os combina entre si e verifica se eles estão interligados criando assim subestruturas completas. Uma parte da listagem operacional encontra-se na fig. 5,junto com a substância cujo espectro foi fornecido e que fez o papel de substância desconhecida.

CONCLUSÕES

Acreditamos que esses poucos exemplos demonstram a flexibilidade e a abrangência do SISTEMAT. Outros aplicativos estão sendo desenvolvidos e estamos melhorando aqueles apresentados neste trabalho. Por exemplo, no SISSTRUC o computador ainda não é capaz de reconhecer subestruturas iguais, o que torna a listagem repetitiva. Recentemente publicou-se um algorítmo[5] que poderia ser aproveitado com ligeiras modificações para sanar esta falha.

Resta contudo, o problema do banco de dados. Em inteligência artificial, o computador pode errar. Esses erros não provêm da programação, mas sim do conhecimento do problema que ele tem, e que pode ser falho se o banco de dados não for suficientemente completo. Para as aplicações em quimiossistemática, este problema é crucial.

REFERÊNCIAS

1. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Borges, J. H. G.; Emerenciano, V. de P.; *Química Nova*, (1990), **13**, 10.
2. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Emerenciano, V. de P.; *Comput & Chem.* (no prelo).
3. Bremser, W.; Wagner, H.; Francke, B.; *Org. Magn. Res.* (1981),**15**, 178.
4. Clerc, J. T.; Sommerauer, H.; *Anal. Chim. Acta*, (1977), **95**, 33.
5. Barone, R.; Arbelot, M.; Chanon, M.; *Tetrahedron Comput. Meth.*, (1988), **1**, 1.