

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

PARTE I - BASES TEÓRICAS

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

^aFaculdade de Engenharia - UNESP - C.P. 205, CEP 12500 - Guaratinguetá (SP)

^bInstituto de Química - UNESP - Araraquara (SP)

^cInstituto de Química - USP - S. Paulo (SP)

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

ABSTRACT

The theoretical bases used to develop an expert system named SISTEMAT for Natural Products Chemistry is described. The main goals of this system are: 1) compactness: each compound needs only 84 bytes to store all the chemical informations that a chemist can discover from inspection of chemical formulae. 2) flexibility: the system may be applied in chemotaxonomy, spectroscopy and structural investigation.

A new randomness access method is proposed that combines the connectivity and polynomial method.

INTRODUÇÃO

A elaboração do banco de dados em química não é um fenômeno recente. Um dos primeiros e mais completos é provavelmente o banco de dados do CAS que manipula mais de 5 milhões de compostos¹. Atualmente devem existir várias dezenas de bancos de dados, a maioria comerciais, enfocando algumas propriedades, sejam espectrais²⁻¹² sintéticas¹³⁻¹⁷ ou farmacológicas das substâncias¹⁸⁻²⁰.

A grande maioria desses bancos são arqueturados segundo o DBASE, isto é, são bancos repetitivos que fornecem dados explicitamente neles contidos.

A introdução dos conceitos de inteligência artificial²¹ em bancos de dados químicos é relativamente recente. Os maiores sucessos alcançados foram nas áreas de espectrometria, principalmente de massas²²⁻²⁹ e de RMN³⁰⁻⁴⁰.

Baseados nesses conceitos e outros, anteriormente empregados, desenvolvemos o programa SISTEMAT, objeto do presente trabalho.

METAS

Ao contrário do DBASE, pretendemos desenvolver um sistema inteligente no seguinte sentido: da descrição de cada composto, o sistema deveria ser capaz de extrair todas as informações que um químico consegue visualizar a partir da inspeção da fórmula de uma substância, como por exemplo funções químicas, fórmula bruta, peso molecular, índice de oxidação, etc. Desta forma, o sistema não se limitaria aos dados explicitamente neles contidos mas poderia, potencialmente, criar seus próprios dados através de uma programação judiciosa.

Uma vez criado o banco, seria possível desenvolver aplica-

tivos aproveitando esta flexibilidade para resolver um determinado problema.

MÉTODO

Quimicamente falando, uma substância se define por duas características:

- a) o nome que caracteriza sua unicidade
- b) a fórmula que caracteriza suas propriedades

O primeiro passo seria, portanto, encontrar um meio de traduzir essas características em linguagem matemática, suscetível de ser entendida e manuseada pelo computador.

A. A Fórmula Química

1. A matriz topológica

Para representar matematicamente uma substância⁴¹, usaremos a matriz topológica que é o recurso matemático mais simples para representar um grafo. Um grafo é simplesmente a interligação, através de vértices (chamadas ligações em química), de nós (correspondentes aos átomos). Como em química existem vários tipos de ligações (simples, duplas, etc.) e vários tipos de átomos (C, O, N, ...), o grafo é chamado de ponderado. Como também as ligações não são orientadas, a fórmula de uma substância é um grafo ponderado não orientado que levaria a uma matriz topológica simétrica. Este fato será de grande importância no item seguinte.

Como foi mencionado, os nós e vértices são ponderados. Os pesos atribuídos podem ser puramente arbitrários, desde que sejam constantes.

Usamos os seguintes valores apresentados na tabela I.

TABELA I - Ponderação de nós e vértices

NÓS		VÉRTICES	
carbono	peso 1	ligação simples	peso 1
carbono aromático	peso 2	ligação aromática	peso 1
oxigênio	peso 3	ligação dupla	peso 2
nitrogênio	peso 4	ligação tripla	peso 3
nitrogênio aromático	peso 5		
flúor	peso 6		
cloro	peso 7		
bromo	peso 8		
iodo	peso 9		
enxofre	peso 10		
fósforo	peso 11		

se um átomo estiver ligado às posições *alfa* ou *beta* (ver adiante) todas as suas ligações, em relação aos átomos de numeração maior, serão consideradas *alfa* ou *beta*. A numeração não segue nenhuma outra norma uma vez que numerações diferentes levarão a matrizes isomorfas.

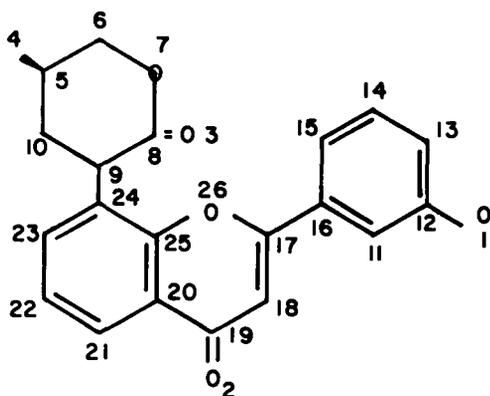


Figura 2

2. O vetor será formado pelos números iniciais e finais de cada cadeia:

- precedido por -1 relacionaremos, na ordem, os átomos aos quais os nós monoatômicos estão ligados, seguidos pelos bígrafos não incluídos nas cadeias;
- precedido por -2 indicaremos o número atômico dos heteroátomos e suas posições no grafo;
- precedido por -3 indicaremos os bígrafos correspondentes às ligações duplas;
- precedido por -4 indicaremos um átomo pertencente a cada ciclo aromático (o computador é capaz de encontrar os demais átomos);
- precedido por -5 indicaremos os nós das ligações triplas;
- precedido por -6-1 indicaremos os átomos de configuração relativa *beta*, por -2 os átomos das ligações *alfa*, por -3 os átomos com carga positiva, por -4 os átomos com carga negativa;
- precedido por -7 indicaremos os átomos das ligações transanulares de conformação *cis*;
- precedido por -8 indicaremos os tipos e posições dos macronós que possam existir;
- o sinal -9 serve para fechar este primeiro vetor.

O flavonóide da fig. 2 será portanto representado pelo vetor seguinte:

04101126-112190805100924111617262025-2080126020703-3171803080219-41321-6-104-9

Além desse vetor serão necessárias informações suplementares para que o computador possa transformar a matriz topológica em representação gráfica. Vários métodos encontram-se descritos na literatura, seja pelo "plate system"⁴²⁻⁴³, seja por métodos heurísticos⁴⁸. Nenhum deles, contudo, leva em conta a conformação e todos usam o modo gráfico, o que torna a impressão da fórmula excessivamente lenta. Por isso preferimos desenvolver nosso próprio método que usa exclusivamente caracteres alfanuméricos, e portanto é mais rápido. Se o resultado não é artisticamente perfeito (ver fig. 4) a fórmula desenhada é, contudo, suficientemente clara para ser entendida por um químico.

O método consiste em definir a posição inicial de cada cadeia seguida sequencialmente pelas orientações dos vértices

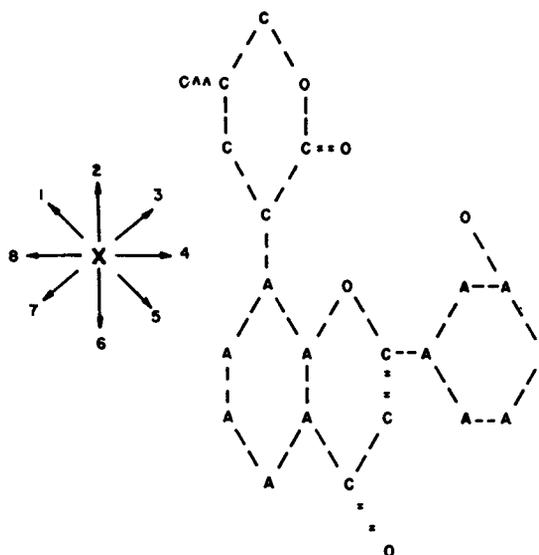


Figura 3

numa rede retilínea⁴³ (fig. 3) que dela fazem parte. Assim o vetor gráfico do flavonóide da fig. 2 seria:

12050403050607012126040507080108060701070102030503

Onde 1205 correspondem às coordenadas iniciais do átomo 04, início da cadeia 0410; 04 corresponde à orientação do bígrafo 0405; 03, do bígrafo 0506 e o processo se repete até a orientação 01 do bígrafo 0910.

O processo é repetido para cada cadeia. A representação gráfica deste flavonóide, expressa pelo computador, teria a forma indicada na fig. 3.

O vetor final será o resultado da concatenação do vetor reduzido e do vetor gráfico.

Como exemplo, forneceremos a seguir os vetores finais das substâncias apresentadas na fig. 4.

Composto 3

0422-1151211040811161520-2082205-3040806071920-6-110-2010203-91120030108069607050506070102010706050607

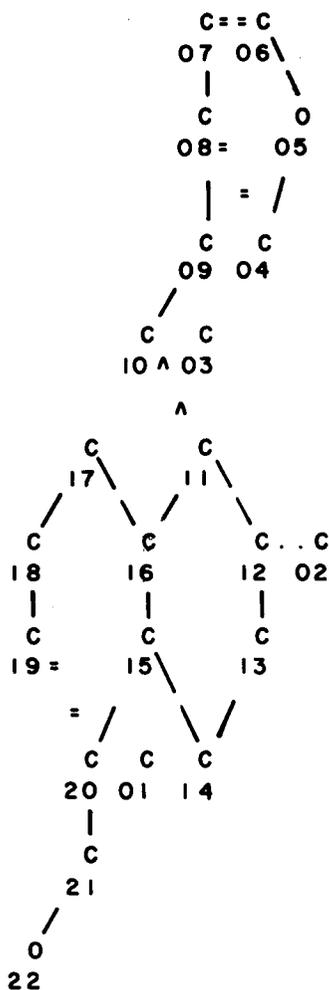
Composto 4

04050607081415161727-121201905220623072308121317141618-18272722-20806040102151014-31211080915260121-6-1071217-205020316-9291502290904081501030506060517240814150505060-7010701020305

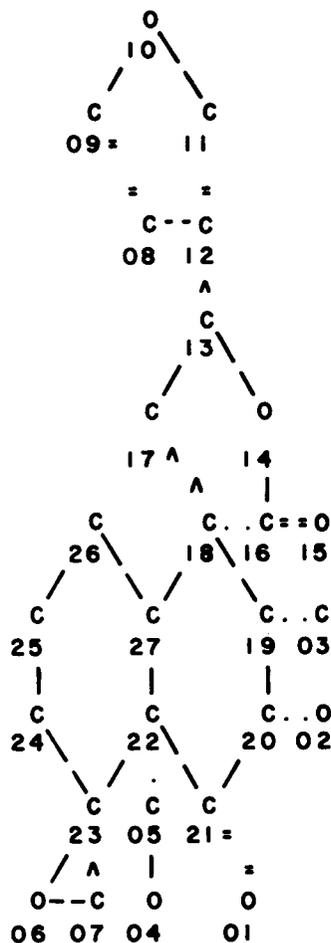
Composto 5

0616-1151511121308161216-2080107090405-31011-6-101070405-202-71216-9282201020302010708070503

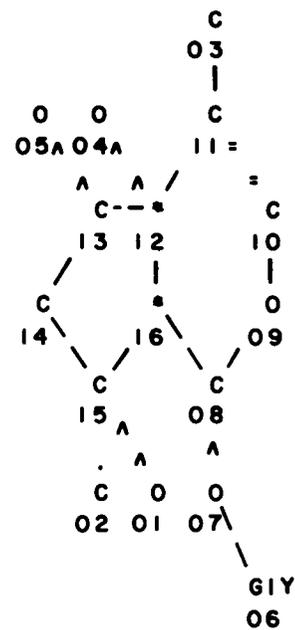
Esses dois vetores ainda podem ser compactados. Cada elemento do vetor reduzido e cada coordenada do nó que inicia uma cadeia no vetor gráfico podem ser transformados num termo ASCII, ocupando 1 byte. Cada par de orientação, da mesma forma, pode ser substituído por um termo ASCII, uma vez que o valor máximo de duas orientações consecutivas será 88. Dessa maneira, o flavonóide da fig. 2 ocuparia 51 bytes, e neste espaço de memória estariam contidas todas as informações químicas da substância, assim como a sua representação gráfica. As moléculas da fig. 4 ocupariam respectivamente 41 (composto 3), 81 (composto 4) e 44 bytes (composto 5). Pensamos que dificilmente conseguir-se-ia condensar tantas informações num espaço tão pequeno.



composto 3



composto 4



composto 5

Figura 4

B. O Nome

O nome serve para identificar inequivocamente um determinado composto. Devido a sua grande importância, a identificação de uma substância por computador foi objeto de muitos trabalhos. Discutiremos inicialmente os métodos mais conhecidos com suas vantagens e desvantagens antes de propormos o método desenvolvido.

1. O Método dos Grafos⁴⁵⁻⁴⁷

Este método consiste em descrever vetorialmente o grafo que representa a molécula. Para evitar o isomorfismo dos grafos, a numeração dos nós obedece a regras rígidas de tal forma que exista uma única maneira de escrever corretamente o vetor. A vantagem deste método é que cada composto é definido sem nenhum equívoco. A grande desvantagem é que para um sistema informatizado, ele não tem qualquer aplicabilidade. A elaboração do vetor tem que ser feita manualmente o que, sem dúvida, criará uma grande fonte de erros devido à complexidade das regras empregadas. Por outro lado, a partir do vetor não há como relacioná-lo a um número de registro: isto é, o acesso ao banco só pode ser seqüencial e não randômico. Esta limitação tornaria a procura excessivamente demorada.

2. O Método dos Nós Condensados⁴⁸⁻⁴⁹

Este método é semelhante ao anterior, os nós sucessivos e iguais sendo condensados num nó único. Evidentemente as mesmas vantagens e desvantagens permanecem.

3. O Método do Polinômio Característico⁵⁰⁻⁵² (ou determinante secular)

À matriz simétrica (M) corresponde um determinante secular (D) através da operação:

$$D = [M - \lambda I] = 0$$

onde I é a matriz de identidade, e λ são as raízes do determinante que pode ser caracterizado pelo seu polinômio ou pelas raízes.

Desta maneira, pensou-se que o isomorfismo dos grafos teria sido resolvido. Herndon⁵², contudo, mostrou que matrizes não isomorfas podiam apresentar o mesmo polinômio e que essas coincidências não eram tão raras. Na realidade, todo grafo que apresentar pontos isospectrais ou pontos de indução acusará este problema.

No entanto, o método é elegante pois permitiria um acesso

randômico aos bancos, apesar do volume respeitável de memória necessária para caracterizar uma substância.

4. O Método das Matrizes de Conectividade⁴⁴

Este método desenvolvido por MORGAN é provavelmente o mais usado. Aplicado no banco de dados do CAS, a proporção de erros não passou de 1 em 10.000 compostos. Ele consiste em definir cada átomo da substância não só em função dele mesmo, mas também dos seus primeiros, segundos, etc., vizinhos, cada um sendo considerado como um nó de peso arbitrário e diferente. É evidentemente o método mais eficaz. Contudo, a caracterização de cada substância requer um respeitável volume de memória.

MÉTODO PROPOSTO

Como foi mencionado, a nossa proposta de trabalho consiste na criação de bancos inteligentes, onde o computador é capaz de identificar as propriedades químicas das substâncias contidas nesses bancos. Por isso, o vetor reduzido deve forçosamente fazer parte do registro. Numa abordagem seqüencial do seu banco de dados, o computador conseguiria diferenciar as substâncias pelas matrizes de conectividade. O nosso único problema seria portanto conseguir um acesso randômico.

Desenvolvemos o seguinte método:

– Dada uma substância A, o computador deveria ser capaz de localizá-la através de seu vetor reduzido (V_A).

Numa primeira etapa, o computador elabora o determinante secular correspondente à matriz e calcula o valor da sua maior raiz. A matriz sendo simétrica, pode-se usar métodos iterativos bastante rápidos. Desta raiz, retém-se os 5 primeiros algarismos da sua parte fracionária e forma um número que chamamos de identificador. Evidentemente, várias substâncias terão o mesmo identificador. Este número serve para subdividir o conjunto total das substâncias em 100.000 subconjuntos (ver contudo observação adiante).

Dentro de cada subconjunto, numa segunda etapa, o computador identifica a substância através das matrizes de conectividade simples.

A matriz de conectividade de uma substância de n átomos é uma matriz (5xn), onde está relacionado em cada coluna o tipo de átomo e dos seus vizinhos. Para aumentar a selectividade do processo diferenciamos cada átomo em função do número de hidrogênios e do tipo de ligações nas quais está envolvido. Foram criados assim 35 tipos de átomos cujos códigos encon-

TABELA II - Código dos Agrupamentos Atômicos

-CH ₃	: 01	O=	: 13	-Br	: 25
-CH ₂ -	: 02	HO-	: 14	-I	: 26
-CH-	: 03	-O-	: 15	S=	: 27
-C-	: 04	-NH ₂	: 16	HS-	: 28
CH ₂ =	: 05	-NH-	: 17	-S-	: 29
-CH=	: 06	-N-	: 18	-S=	: 30
-C=	: 07	NH=	: 19	=S=	: 31
TCH	: 08	-N=	: 20	-PH ₂	: 32
TC	: 09	N=	: 21	-PH-	: 33
HC	: 10	N _{ar}	: 22	-P-	: 34
C _{ar}	: 11	-F	: 23	-P-	: 35
=C=	: 12	-Cl	: 24		

tram-se na Tabela II. Como exemplo transcrevemos a matriz de conectividade da substância da fig. 1 (Tabela III).

Nesta matriz, cada coluna corresponde a um átomo (1ª linha) e seus vizinhos. Para completar a matriz anula-se as posições não ocupadas. Assim a sexta coluna corresponde ao C₆ (CH – código 03 na primeira linha) e seus vizinhos (C₅ – código 02; C₁₁ – código 03 e C₇ – código 15), a quinta linha é anulada pois o quarto vizinho não existe.

TABELA III - Matriz de Conectividade da Substância Ia

21	09	07	07	02	03	15	02	02	17	03	02	01
09	21	07	07	07	02	03	15	02	02	17	03	07
00	07	02	01	03	03	02	02	17	03	03	17	00
00	00	09	02	00	15	00	00	00	00	02	00	00
00	00	00	00	00	00	00	00	00	00	00	00	00

De posse da matriz de conectividade da substância A(MC_A), o computador compara para todas as substâncias do mesmo subconjunto as suas matrizes de conectividade (MC_X) e verifica se cada coluna de MC_A tem o seu correspondente em MC_X. Se houver coincidência, trata-se da mesma substância, senão passa-se para a substância seguinte. O fluxograma da procura encontra-se na fig. 5.

O método proposto, que combina as técnicas das matrizes de conectividade e do determinante secular, apresenta uma grande vantagem pois reduz o espaço de memória necessário para identificar uma substância. Num sistema planejado para 50.000 compostos cada subconjunto é identificado num registro de 2 bytes e cada substância gasta 2 bytes para conter o registro de continuação.

A grande dúvida evidentemente é: “o método é eficiente ou não?”. Pelo fato de ser uma combinação de duas técnicas bem distintas acreditamos que a porcentagem de erros deveria ser aproximadamente o produto das porcentagens de cada uma, ou seja, algo em torno de um em cada 10⁶ substâncias. Realizamos testes com algumas centenas de compostos sem verificar nenhum erro. Estatisticamente, contudo, esta série de testes não tem significado conclusivo.

Finalmente, é necessário ter cuidado com a precisão do cálculo do identificador. Mesmo trabalhando em precisão dupla, o computador é incapaz de distinguir dois valores próximos. Por razões de velocidade de processamento, impusemos a precisão de cálculo até a nona casa decimal, isto é, o computador não vai distinguir entre 1,234569999 e 1,234570000. Os identificadores seriam diferentes para uma mesma substância. Para evitar esse possível erro, a procura se processa em 3 subconjuntos, o do identificador calculado e dos seus vizinhos.

CONCLUSÃO

O método descrito representa um conjunto de leis que permitem traduzir a química, pelo menos no seu aspecto descritivo, em linguagem programável. Usando essas leis, o papel do computador não se limita mais a um mero reservatório bibliográfico, mas ele pode criar os seus próprios dados ou analisar quimicamente uma substância. Tudo depende evidentemente da programação. SISTEMAT foi desenvolvido tendo como enfoque a química de Produtos Naturais, mas as mesmas leis poderiam ser usadas para desenvolver outros sistemas especialistas, como em Química de Síntese, Química de Coordenação, etc.

Uma outra vantagem do método é a sua compactação que o torna acessível a um simples PC equipado de um disco rígido. Esta compactação foi a nossa preocupação constante. Os PCs são largamente difundidos nos laboratórios, enquanto main

frame ou mesmo mini são menos acessíveis.

Se o químico assim o desejar, o seu PC pode se transformar numa ferramenta preciosa como outros equipamentos de uso corriqueiro, o que será mostrado na 2ª parte deste trabalho.

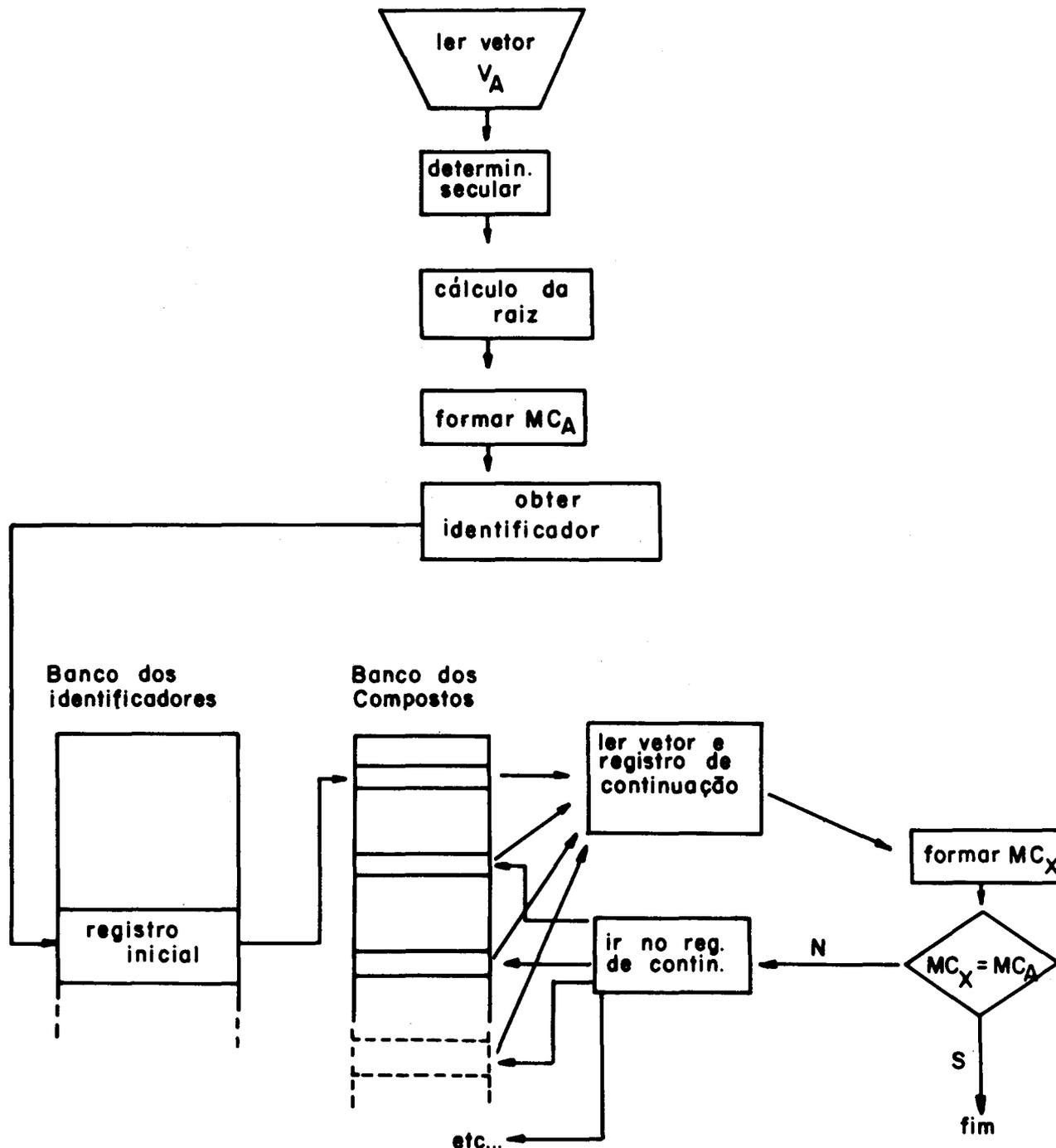


Figura 5

REFERÊNCIAS

1. Vanderstouw, G.G.; Gustafson, C.; Rulf, J.D.; Watson, C.E.; *J. Chem. Inf. Comput. SC.* (1976), *16*, 213.
2. Martinsen, D.P.; Song, B.H.; *Mas. Spectrom. Rev.* (1985), *1*, 173.
3. Ishida, Y.; Sasaki, S.; *Comput. Enhanced Spectrosc.* (1983), *1*, 173.
4. Bally, R.W.; Krimpton, D. Van; Klooster, H.A. Vant't; *Anal. Chim. Acta.* (1984), *157*(2), 227.
5. Tanabe, K.; Hiraishi, J.; Saeki, S.; Suzuki, I.; Tasuni, M.; *Comput. Enhanced Spectrosc.* (1984), *2*, 13.
6. Kimber, B.J.; *J. Magn. Res.* (1984), *57*, 140.
7. Zippel, M.; Mowitz, J.; Koehler, I.; Opferkuch, H.J.; *Anal. Chim. Acta.* (1982), *140*(1), 123.
8. Puthman, A.; Koontz, J.P.; Hinderliter, J.; Woodward, W.S.; Reilley, C.N.; *Anal. Chem.* (1982), *54*, 1772.
9. Saeki, S.; Yamamoto, O.; *Codata Bull.* (1981), *40*, 53.
10. Shaps, R.H.; Sprouse, J.F.; *Eur. Spectrosc. News.* (1980), *32*, 39.
11. Clerc, J.T.; Szekely, G.; *Tr. Anal. Chem.* (1983), *2*, 50.
12. Trulson, M.O.; Munk, M.E.; *Anal. Chem.* (1983), *55*, 2137.
13. Hippe, Z.; *Anal. Chim. Acta.* (1981), *133*(4), 677.
14. Cohen, B.J.; *J. Chem. Inf. Comput. SC.* (1982), *22*, 195.
15. Choplin, F.; Goudian, S.; Kaufmann, G.; *J. Chem. Inf. Comput. SC.* (1983), *23*, 26.
16. Chodosh, D.F.; Mendelson, W.L.; *Pharm. Technol.* (1983), *7*, 90.
17. Peter, A.J.; *Chem. Br.* (1985), *21*, 59, 62, 66.
18. Hodes, L.; *J. Chem. Inf. Comput. SC.* (1981), *21*, 128, 132.
19. Klopman, G.; *J. Am. Chem. Soc.* (1984), *106*, 7315.
20. Choplin, F.; *Pour La Science* (1985), *10*, 1.
21. Pierce, T. H.; Hohne, B. A.; "Artificial Intelligence Applications in Chemistry"; ACS Ed., Washington DC (1985).
22. Clark, H.A.; Jurs, P.C.; *Anal. Chim. Acta.* (1981), *132*, 75.
23. McLafferty, F.W.; Atwater, A.L.; Haraki, K.S.; Hosokawa, K.; Mun, I.K.; Venkataraghavan, R.; *Adv. Mass. Spectrosc.* (1980), *8B*, 1564.
24. Martinsen, D.P.; *Appl. Spectrosc.* (1981), *35*, 255.
25. Nakayama, T.; Fujiwara, Y.; *J. Chem. Inf. Comput. SC.* (1981), *21*, 142.
26. Gray, N.A.B.; Buchs, A.; Smith, D.H.; Djerassi, C.; *Helv. Chim. Acta.* (1981), *64*, 458.
27. Gray, N.A.B.; Buchs, A.; Smith, D.H.; *Anal. Chem. Symp. Ser.* (1981), *7*, 295.
28. Yasuhara, A.; Shindo, J.; Ito, H.; Misoguchi, T.; *Comp. Enhanced Spectrosc.* (1983), *1*, 117.
29. Shackelford, W.M.; Cline, D.M.; *Anal. Chim. Acta.* (1984), *164*, 251.
30. Robien, W.; *Monatsh. Chem.* (1983), *114*, 365.
31. Abe, H.; Fujiwara, I.; Nishimura, T.; Okuyama, T.; Kida, T.; Sasaki, S.; *Comput. Enhanced Spectrosc.* (1983), *1*, 55.
32. Dubois, J.E.; Carabedian, M.; Dagane, I.; *Anal. Chim. Acta.* (1984), *158*, 217.
33. Lindsay, R.K.; "Applications of Artificial Intelligence for Organic Chemistry"; McGraw-Hill Book Co., NY (1980).
34. Grandell, C.W.; Gray, N.A.B.; Smith, D.H.; *J. Chem. Inf. Comput. SC.* (1982), *22*, 48.
35. Djerassi, C.; Smith, D.H.; Grandell, C.W.; Gray, N.A.B.; Nourse, J.G.; Lindley, M.R.; *Pure Appl. Chem.* (1982), *54*, 2425.
36. Gray, N.A.B.; *Prog. Nucl. Magn. Reson. Spectros.* (1982), *15*, 201.
37. Kalchhauser, H.; Robien, W.; *J. Chem. Inf. Comput. SC.* (1985), *25*, 103.
38. Gastmans, J.P.; Zurita, J.C.; Sahão, J.; Emerenciano, V. de P.; *Anal. Chim. Acta.* (1989), *217*, 85.
39. Gastmans, J.P.; *QCPE Bull.* (1988), *8*, 181.
40. Gastmans, J.P.; Furlan, M.; Emerenciano, V. de P.; *Comput. & Chem.* (1988), *12*, 285.
41. Calusaru, A.; Volanshi, C.; *Rev. Chim.* (1984), *35*, 307.
42. Dittmar, P.G.; Mockus, J.; Couvreur, K.M.; *J. Chem. Inf. Comput. SC.* (1977), *17*, 186.
43. Carhart, R.E.; *J. Chem. Inf. Comput. SC.* (1976), *16*, 82.
44. Shelley, C.A.; *J. Chem. Inf. Comput. SC.* (1983), *23*, 61.
45. Lozac'h, N.; Goodson, A.L.; Powell, W.H.; *Ang. Chem. Int. Ed.* (1979), *18*, 887.
46. Lozac'h, N.; Goodson, A.L.; *Ang. Chem. Int. Ed.* (1984), *23*, 33.
47. Gottlieb, O.R.; Kaplan, A.C.; *J. Chem. Inf. Comput. SC.* (1986), *26*, 135.
48. Read, R.C.; *J. Chem. Inf. Comput. SC.* (1983), *23*, 135.
49. Read, R.C.; *J. Chem. Inf. Comput. SC.* (1985), *25*, 116.
50. Spialter, L.; *J. Am. Chem. Soc.* (1963), *85*, 2012.
51. Spialter, L.; *J. Chem. Doc.* (1964), *4*, 261, 269.
52. Kudo, Y.; Yamasaki, T.; Sasaki, S.; *J. Chem. Doc.* (1973), *13*, 225.
53. Herndon, W.C.; *J. Chem. Doc.* (1974), *14*, 150.
54. Morgan, H.L.; *J. Chem. Doc.* (1965), *5*, 107.