[2] Teixeira, V.L.; Tomassini, T.; Fleury, B.G.; Kelecom. A.; *J. Nat. Prod.* (1986) *49*, 570.
[3] Gonzalez, A.G.; Martin, J.D.; Norte, M.; Rivera, P.; Perales, A.; Fayos, J.; *Tetrahedron* (1983) *39*, 3355.
[4] Crews, P.; Klein, T.E.; Hogue, E.R.; Myers, B.L.; *J. Org. Chem.* (1982) *47*, 811.

ARTIGO

APLICAÇÕES DA INTELIGÊNCIA ARTIFICIAL EM QUÍMICA ORGÂNICA O PROJETO PRONAT

Jean Pierre Gastmans e Maysa Furlan

Instituto de Química – UNESP; 14800 – Araraquara (SP).

Vicente de Paulo Emerenciano, Nídia Franca Roque e Ana Cristina Bussolini

Instituto de Química – USP; 01498 – São Paulo (SP)

(Recebido em 25/5/88; cópia revisada em 14/12/88)

ABSTRACT

We describe some expert systems based on ^{13}C NMR. They were developed in order to give assistance to organic chemists working, on structural determination. Some didatic and theoretical applications are described.

INTRODUÇÃO

Grandes esforços têm sido realizados nos últimos anos para a criação de sistemas especialistas abrangendo vários aspectos da química.

Como a Química Orgânica não é uma ciência matemática propriamente dita, os sistemas desenvolvidos por vários pesquisadores usaram os conceitos da inteligência artificial[1,2]. Vários programas foram elaborados enfocando, genericamente, uma das três áreas seguintes:

– proposta de rotas sintéticas[3,4];

– determinação estrutural de produtos naturais[5 a 12];

– atividade farmacológica de compostos orgânicos[13,14,15]

Alguns desses sistemas são comercializados[5,12] ou usados em grandes empresas[3,14,15].

O acesso a esses programas e seus bancos de dados geralmente é caro. Por esta razão pensamos em construir, aqui no Brasil, algum tipo de sistema especialista que pudesse auxiliar os químicos orgânicos no seu trabalho de determinação estrutural de produtos naturais.

Do esforço das nossas universidades, nasceu o projeto PRONAT (PROdutos NATurais). A realidade computacional do Brasil sendo bem diferente daquela existente no exterior, o sistema devia ser compatível com o hardware facilmente disponível no país; ou seja um microcomputador do tipo IBM-PC ou XT equipado com um disco rígido de 10 Mb sem extensão de memória.

Sem perder sua eficiência, o projeto dependia, portanto, do desenvolvimento de um sistema altamente especializado e automático de compactação dos dados.

Quando outros sistemas p.ex.[11,12] gastavam K$\not b$ para armazenar um dado, tínhamos que conseguir o mesmo com algumas dezenas de bytes. O sistema de compactação foi desenvolvido e encontra-se embutido em todos os programas codificando ou decodificando a compactação.

A memória limitada que impusemos nos levou a escolher os deslocamentos de ^{13}C como fonte de dados experimentais e não dados de espectrometria de massa. De fato cada subestrutura é definida spectralmente por três dados, o valor máximo do deslocamento, o mínimo e o valor médio, ou seja, 3$\not b$ enquanto que os sistemas baseados em espectroscopia de massas necessitam de 2$\not b$ para cada pico experimental.

SISTEMAS EM OPERAÇÃO

A. Sistema ^{13}C

Esse sistema, plenamente operacional nas nossas unidades, encontra-se em via de publicação[16,17] e foi registrado no Q.C.E.P. (Quantum Chemistry Exchange Program).

Trata-se de um programa de simulação espectral de ^{13}C que gera espectros teóricos para quaisquer substâncias, que podem ser comparadas com o espectro real, eliminando assim as propostas estruturais inadequadas e "orientando" o químico em direção a uma solução de seu problema.

A eficiência de um programa deste tipo depende essencialmente do sistema de codificação das subestruturas. Cada átomo vizinho que influencia o deslocamento químico deve

estar presente no código, a sua influência relativa deve ser destacada e os vizinhos que não influenciam no valor do deslocamento ou cuja influência é pequena não devem ser relacionados no código, pois criar-se-iam subestruturas equivalentes gastando, desnecessariamente, espaço de armazenamento.

Os métodos de codificação existentes[5,8,18,27,28] não nos pareciam obedecer a esses requisitos. Por isso resolvemos desenvolver o nosso próprio método de codificação[16]. Este se baseia nas distâncias interatômicas no tipo de átomos vizinhos e na orientação relativa dos vizinhos.

As distâncias interatômicas são obtidas da seguinte maneira:

Após montar o modelo molecular (Prentice – Hall Inc. N.Y.) com distâncias de ligação padronizadas na sua conformação mais estável, projeta-se os átomos nos planos xy e yz. Se ocorrer erros de medida, eles serão automaticamente corrigidos pelo computador no decorrer do programa portanto que eles não ultrapassam 0,2 Å.

O sistema é composto de dois programas (BANCO e C 13).

O primeiro recolhe os deslocamentos experimentais, a descrição do composto, as coordenadas de todos os átomos, exceto os hidrogênios e cria o banco de dados chamado DTSET. Para cada carbono, ele elabora a sua subestrutura, verifica se ela já existe no banco; neste caso ele modifica o registro da substância para incorporar este novo deslocamento. Se a substrutura não existe, ele cria um novo registro para incluí-la no banco DTSET.

A medida que novos dados experimentais são fornecidos, o sistema vai literalmente "aprendendo" a espectrometria de RMN de ^{13}C e tornando-se mais "inteligente". Atualmente o banco DTSET conta com 4.000 subestruturas provenientes de mais de 11.000 deslocamentos experimentais. Graças ao sistema de compactação já mencionado, o banco DTSET não ocupa mais que 450 K$\not b$.

Todo trabalho de construção de bancos tem-se restringido até o momento em terpenóides pois pretendemos tratar as substâncias essencialmente aromáticas (flavonóides, lignóides, etc.) e maneira a parte, uma vez que os deslocamentos químicos dependem das matrizes de densidade eletrônica[19].

O segundo programa (C 13) gera os espectros teóricos a partir do banco DTSET. Para a previsão espectral o sistema necessita de:

1. O vetor reduzido correspondente à matriz topológica do grafo molecular. As regras utilizadas para elaborar este vetor encontram-se descritas no próprio programa;

2. As coordenadas espaciais de todos os átomos, exceto os hidrogênios.

Um exemplo de previsão de espectro de RMN de ^{13}C de esteróide da Fig. 1 está mostrado na Tabela I.

A procura dentro do banco realiza-se em 4 níveis a saber:

– o nível O corresponde ao próprio carbono objeto da procura com os seus hidrogênios e ligações.
– o nível 1 corresponde ao nível O mais todos os átomos e ligações existentes numa esfera de raio igual a 1,7 Å centrada no carbono do nível O.

Este nível corresponde exatamente ao nível alfa do espectroscopista.

– o nível 2 corresponde ao nível 1 acrescentado de todos os átomos e ligações existentes numa esfera de raio igual a 2,9 Å.
– o nível 3 engloba todos os átomos e ligações existentes num raio de 4.2 Å.

Um nível é considerado preenchido, e portanto a procura prossegue no nível seguinte, se existe coincidência de todos os átomos (inclusive hidrogênio e de todas as distâncias interatômicas entre o carbono do nível O e os vizinhos englobados no nível objeto da procura. Evidentemente, os níveis 2 e 3 não correspondem aos níveis beta e gama do espectroscopista. Assim por exemplo, os carbonos 18 e 19 pertencem ao segundo nível do carbono 11 apesar de serem gama. Este tipo de codificação contudo, é mais coerente com as equações teóricas que regem a RMN de ^{13}C.

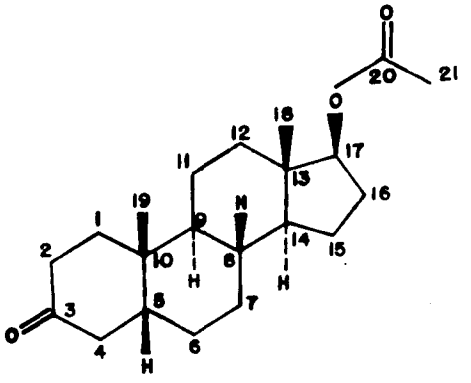


FIG. 1

Após a previsão, o computador propõe uma tentativa de atribuição dos deslocamentos experimentais se estes são conhecidos. A atribuição feita pelo computador no caso do esteróide da Fig. 1, está descrita na penúltima coluna da Tabela I. O método de codificação que desenvolvemos, por ser diferente dos demais, nos permitiu incluir várias opções que são impraticáveis nos outros programas e que já fazem parte do sistema.

Opção: "previsão fina"

Esta opção permite estudar o deslocamento de cada carbono incluindo na sua estrutura cada átomo vizinho um por um na ordem *das distâncias atômicas crescentes.* Por exemplo a procura do deslocamento do átomo 17 parou no primeiro nível com uma faixa de tolerância muito grande (de 61.2 até 88.3). Graças a esta opção, o computador incluiu sucessivamente os átomos 20, 14, 18 e 15. A previsão final foi de 79.4 com uma faixa de tolerância de 77.6 até 82.6 δ.

Tabela I

Carbono	Faixa de Tolerância		Previsão Estatística	Número de casos	Nível	Atribuído aos atomos	Exp.
	min.	max					
1	31.0	37.0	34.0	2	2	1,2,16,12	37.1
2	37.0	38.2	37.7	2	2	1,2,16,12	37.0
3	212.6	212.6	212.6	1	3	3	212.5
4	42.2	42.2	42.2	1	3	4	42.1
5	44.2	44.2	44.2	1	3	5	44.2
6	24.7	24.7	24.7	1	2	1,6,7,15,16	25.4
7	24.4	24.4	24.4	1	2	1,2,6,7,16	26.5
8	35.0	36.2	35.5	8	3	8	35.4
9	41.0	41.0	41.0	1	2	9	40.8
10	35.0	36.7	35.7	5	3	10	35.0
11	20.5	20.8	20.6	3	3	11	20.7
12	25.8	41.4	37.7	46	2	1,2,12,16	37.0
13	42.5	42.5	42.5	2	3	13	42.7
14	50.7	50.7	50.7	2	3	14	50.8
15	23.5	23.5	23.5	2	3	6,7,15,16	23.5
16	27.6	27.6	27.6	2	2	1,2,6,7,12,16	27.6
17	61.2	88.3	77.6	44	1	17	82.5
18	11.1	24.2	14.8	15	2	18,19	12.1
19	24.7	24.8	24.8	2	2	18,19,21	22.6
20	170.8	170.8	170.8	1	3	20	170.8
21	20.7	20.7	20.7	1	3	19,21	21.1

Obs.: Nível 1, 2 e 3 significa a inclusão dos vizinhos até 1.7 Å, 2.9 Å e 4.2 Å respectivamente.

Opção: "degradação"

Esta opção permite retirar da molécula de um a cinco átomos e realizar novas previsões para todos os átomos fornecendo novos valores de δ. Este tipo de degradação foi realizada com o esteróide da Fig. 1 e os resultados obtidos estão mostrados na Tabela II.

Estudando uma população estatisticamente significativa de compostos pode-se, inclusive, estabelecer relações lineares que regem a influência de uma determinada função química sobre os deslocamentos. Duas tentativas neste sentido foram feitas. Procurou-se equacionar a influência da carbonila[20] e da hidroxila[21] sobre o deslocamento químico dos átomos vizinhos (γ e δ) e os $\Delta\delta$ observados foram submetidos a uma regressão linear múltipla[22]. Obtivemos graus de correlação satisfatórios, superiores a 0.85.

B. Sistema Check-Onde

A eficiência de um sistema de inteligência artificial como o C 13, depende crucialmente da confiabilidade que podemos depositar no banco de dados.

Bremser[18], analisando o seu banco de dados avaliava que aproximadamente 10% das atribuições não estavam corretas e que 2% das propostas estruturais não eram confiáveis.

Cercamo-nos dos maiores cuidados na escolha dos dados fornecidos ao computador, escolhendo sempre que possível estruturas oriundas de trabalhos confirmados por síntese, raios-X, etc. Da mesma forma a coerência de cada deslocamento era testada frente ao banco existente antes de ser introduzida.

Para alguns tipos de esqueletos, este procedimento não é suficiente. Os primeiros a assinalar este problema foram os pesquisadores do grupo de Stanford[23]. Ao tentar colocar 96 lactonas sesquiterpênicas no banco de dados, esses pesquisadores descobriram que os erros de atribuição eram tão frequentes que isto ameaçaria a integridade do banco e um trabalho de correção foi proposto.

Desenvolvemos dois programas que permitem detectar prováveis erros de atribuições e apontá-los.

O primeiro (CKECK) compara entre si todas as subestruturas de uma população de compostos e aponta as discrepâncias que porventura existam. O segundo (ONDE) lista os compostos onde uma mesma subestrutura foi detectada e o deslocamento que lhe foi atribuído.

O sistema está sendo aplicado a uma população de 360 lactonas sesquiterpênicas e será utilizado futuramente com outras classes de substâncias.

Tabela II

Carbono	Faixa de Tolerância		Previsão Estatística	dist. (Å)	Δδ
	min.	max.			
13	40.2	41.2	40.8	2.32	6.9
16	20.5	20.8	20.6	2.41	7.0
17	40.2	40.5	40.4	1.42	42.1

C. Sistema Pickup[24]

Podemos qualificar este sistema como uma DBASE químico.

Usando a teoria dos grafos e das matrizes topológicas, desenvolvemos um sistema no qual o computador pode reconhecer por si só todas as funções químicas ou qualquer grupo de átomos ligados entre si, como o químico o faz pela inspeção da fórmula do composto. Ao contrário o DBASE onde os dados tem que ser explicitamente contidos no banco, o sistema PICKUP cria os seus próprios dados que servirão como critério de seleção.

É usado atualmente para extrair de uma dada população, as substâncias que apresentam o mesmo padrão de substituição, obedecendo a vários requisitos químicos (até 9). Ele fornece a faixa dos deslocamentos e inversamente lista as substâncias que apresentam deslocamentos dentro desta faixa com a multiplicidade adequada.

Pretendemos com esse sistema fornecer ao espectroscopista uma ferramenta capaz de detectar o(s) deslocamento(s) característico(s) de cada esqueleto, determinar estatisticamente os principais padrões de oxidação inerentes a ele e facilitar a análise dos Δδ ocasionados pelos átomos vizinhos.

D. O Sistema MATCH[30]

Este sistema foi desenvolvido inicialmente para resolver problemas simples em RMN de ^{13}C tais como a confrontação de um espectro de uma substância desconhecida com um banco de dados formado por substâncias da mesma classe. Este tipo de procedimento evita a procura demorada em catálogos e ajuda o químico fornecendo várias subestruturas que agrupadas podem levá-lo a uma proposta estrutural. O sistema foi ampliado posteriormente para auxiliar na análise de misturas. Este novo programa chamado MCAR13 permite ao químico propor uma mistura com até 10 triterpenos e o computador propõe um espectro de RMN de ^{13}C para esta mistura hipotética. O espectro proposto pode então ser comparado com o espectro experimental e o programa fornece a probabilidade de termos feito uma proposta coerente. Os procedimentos de confrontação espectral desenvolvidos durante a construção do sistema MATCH são extremamente úteis pois, aliados a um índice de semelhanças desenvolvido por Bremser[28], permitem indicar em amostras de substâmcias terpenóidicas desconhecidas, o esqueleto a que pertencem.

E. O Sistema TERPTRI

Outra área do interesse dentro do projeto PRONAT é a aplicação de técnicas heurísticas[1] em RMN de ^{13}C. Estas técnicas consistem em fornecer ao computador regras para a resolução de problemas que melhoram a eficiência dos processos de busca e confrontação, sacrificando às vezes as idéias de perfeição. Este tipo de procedimento pode ser empregado, por exemplo, para ensinar ao computador detalhes específicos sobre o espectro de RMN de uma substância terpenóidica, com os quais possa identificar a que esqueleto pertence esta substância. Detalhes deste tipo poderiam ser o número de CH_3, CH_2, CH e C de cada esqueleto e a maneira de chegar a estes números pela análise do espectro seria ensiná-lo a desfuncionalizar teoricamente a substância transformando C = O em CH_2, CO_2H em CH_3, etc... o que é feito automaticamente pelo sistema. O programa, denominado TERPTRI, já está em uso em nossos centros de pesquisa e consegue identificar esqueletos de triterpenos pentacíclicos através desta metodologia[29].

SISTEMAS EM DESENVOLVIMENTO

Estamos atualmente trabalhando no desenvolvimento de dois sistemas que estão em fase de teste. O primeiro, chamado de STRUCTURE realiza a tarefa inversa do C 13, ou seja a partir do espectro, pretendemos que ele forneça as subestruturas que potencialmente poderiam ser a origem dos deslocamentos observados. O segundo, mais ambicioso, chamado de SISTEMA (quimiosSISTEMAtica) pretende oferecer uma ferramenta auxiliar aos químicos que trabalham nesta área.

STRUCTURE

Existem na literatura dois sistemas que permitem em alguns casos a proposta de estruturas via computador. O primeiro é o sistema GENOA do projeto DENDRAL[5], o

segundo é o CHEMICS[6,25,26] desenvolvido no Japão. O sistema GENOA para ser eficiente na prática precisa ter em memória todas as subestruturas correspondentes a cada carbono do composto estudado. Comparando as informações extraídas do banco o sistema consegue propor um número aceitável de propostas estruturais (algumas dezenas).

É claro que se uma subestrutura não existir no banco, o sistema falhará.

O CHEMICS ao contrário, trabalha com um número reduzido de subestruturas (171 atualmente) contendo cada uma poucos átomos. Pelo mesmo processo de comparação, ele elimina as estruturas incoerentes, mas como existem poucos átomos em cada subestrutura, o número de subestrutura possíveis é grande e a combinação entre elas leva a milhares, e as vezes dezenas de milhares de propostas estruturais.

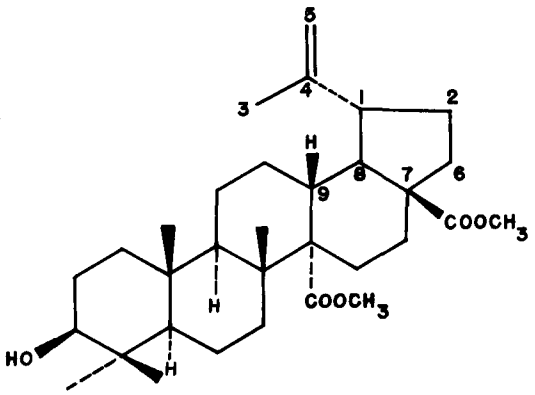
Sem ser tão ambicioso, o nosso projeto se limitará a propor ao químico quais as subestruturas possíveis oriundas de deslocamentos químicos.

A nossa proposta se baseia na observação de que a grande maioria (mais de 80%) das previsões feitas pelo sistema C13 atingem o segundo nível. Quando isso se verificar, podemos impor a cada subestrutura condições de coerência espectrais que vão limitar o número de subestruturas possíveis associadas a um par de deslocamentos. Vamos exemplificar com o triterpeno da figura abaixo (Fig. 2).

Esta substância não existe no banco mas as subestruturas dos átomos 1 e 2 existem até o segundo nível. Este nível engloba os átomos numerados.

Limitando-se a impor como condições, as faixas de tolerância espectral associadas aos carbonos 1 e 2, teremos que considerar que centenas de subestruturas podem ser relacionadas a esses dois deslocamentos.

Contudo, dentro dessas centenas de possibilidades, a substância estudada apresentará realmente a subestrutura da Fig. 2 se satisfizer também os requisitos espectrais dos outros carbonos[3 a 9] limitados ao volume da sub-estrutura (p.ex. para C_9 o nível adequado será o nível 0, para C_8 o



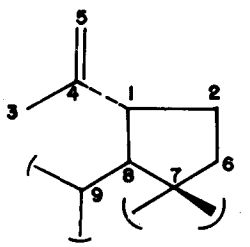


FIG. 2

Tabela III

Carbono	Multiplicidade	Requisitos de 1,2		Nível
		Faixa de Tolerância		
		máx.	min.	
1	2	39	51	2
2	3	27	39	2
Requisitos Complementares dos Outros Carbonos				
3	4	18	21	3
4	1	142	150	3
5	3	107	111	3
6	3	19	45	1
7	1	17	109	0
8	2	31	69	1
9	2	21	110	0

nível 1 etc...) Esses requisitos, obtidos, pelo computador estão descritos na Tabela III.

O conjunto desses 8 requisitos é evidentemente mais rigoroso que as simples faixas de tolerância dos carbonos 1 e 2.

O sistema STRUCTURE está atualmente em fase de elaboração e deverá ser operacional dentro de alguns meses.

SISTEMA

O SISTEMA ainda está em fase de programação e optimização. Ele se propõe em ajudar o trabalho do químico que atua em quimioSISTEMÁtica). O sistema deverá poder listar as substâncias por origem botânica, esqueleto, tipo, funções químicas, peso molecular, índice de oxidação, metoxilação, etc.

Devido às limitações impostas pelo hardware, o sistema será "inteligente", no sentido que ele poderá reencontrar todos esses dados (menos a origem botânica evidentemente) por si só sem que eles sejam explicitamente incluídos no banco. Desta maneira conseguimos dimensionar o sistema para 15.000 compostos e 50.000 origens botânicas em 3.2 Mb, o que o torna acessível para qualquer PC. A descrição completa do sistema será objeto de uma próxima publicação.

CONCLUSÕES

Pretendemos com este trabalho, apresentar uma revisão geral do software desenvolvido e/ou em desenvolvimento nas nossas instituições e colocá-lo à disposição dos pesquisadores brasileiros interessados.

Os sistemas descritos são auto explicativos e não requerem nenhum conhecimento específico em computação. Tentamos, na medida do possível, usar uma linguagem química na relação usuário-máquina de tal forma que o seu uso seja o mais fácil possível.

REFERÊNCIAS

[1] Rich, E.; *Inteligência Artificial;* Mc Graw Hill, São Paulo (1988).

[2] A.C.S. Symposium Series (1985) vol. 306.

[3] Martin, Y.; *Pharmaco-chem. Libr.* (1982) *4*, 269.

[4] Rozenblit, A.B.; *ibid* (1982) *4*, 287.

[5] Djerassi, C.; Smith, O.H.; Cranoell, C.W.; Gray, M.A.B.; Mourse, J.G.; Lindley, M.R.; *Pure Appl. Chem.* (1982) *54*, 2425.

[6] Abe, H.; Fujiwara, I.; Mishimiura, T.; Okuyama, T.; Kida, T.; Sasaki, S.; *Comput. Enh. Spectrosc.* (1983) *1*, 55.

[7] Adler, B.; Binshuba, A.; Herrmann, I.; *Z. Chem.* (1986) *26*, 157.

[8] Dubois, J.E.; Carabedian, M.; Dagane, I.; *Anal. Chim. Acta* (1984) *158*, 217.

[9] Shelley, C.A.; *Anal. Chem.* (1982) *54*, 516.

[10] Kalchhhauser, H.; Wolfgang, R.J.; *J. Chem. Inf. Comput. Sci.* (1985) *25*, 103.

[11] Wolfgang, R.; *Monatsh. Chem.* (1983) *114*, 365.

[12] Bremser, W.; *Magn. Reson. Chem.* (1985) *23*, 1056.

[13] Perun, T.J.; *W. Biotec. Rep.* (1985) *2*, 313.

[14] Boyd, D.B.; *Druy Inf. J.* (1983) *17*, 171.

[15] Chopin, F.; *Pour da Science* (1985) nov.

[16] Gastmans, J.P.; Silva, J.C.Z.; Sahão, J.; Emerenciano, V. de P.; *Anal. Chim. Acta.* (1988) (no prelo).

[17] Gastmans, J.P.; Emerenciano, V. de P.; Furlan, M.; *Comput. and Chem.* (1988) (no prelo).

[18] Bremser, W.; *Anal. Chim. Acta* (1978) *103*, 355.

[19] Pople, J.A.; *Proc. Roy. Soc.* (1957) *A239*, 550.

[20] Furlan, M.; Gastmans, J.P.; Emerenciano, V. de P.; *Magn. Res. Chem.* (submetido).

[21] Furlan M. (resultados não publicados).

[22] Morusis M.J.; SPSS/PC+ Statistical Package, Chicago U.S.A. (1986).

[23] Lindley, M.R.; Gray, M.A.B.; Surith, D.H.; Djerassi C.; *J. Org. Chem.* (1982) *47*, 1027.

[24] Gastmans, J.P.; Furlan, M.; Nasser, M.; Emerenciano, V. de P. *Comput and Chem.* (aceito).

[25] Fujiwara, I.; Okuyama, T.; Yamasaki, T.; Abe H.; Sasaki, S.; *Anal. Chim. Acta,* (1981) *133*, 527.

[26] Sasaki S.; Abe H.; *Anal. Chem. Symp. Ser.* (1983) *15*, 185.

[27] Gray, M.A.B.; Mourse, J.G.; Grandell C.W.; Smith D.H.; Djerassi, C.; *Org. Magn. Res.* (1981) *15*, 375.

[28] Bremser, W.; Klier, M.; Meyer, E.; *Org. Magn. Res.* (1975) *7*, 97.

[29] Maia, C.; Dissertação de Mestrado, Universidade Federal Rural do Rio de Janeiro, 1988.

[30] Emerenciano, V. de P.; Roque, M.F.; Furlan, M.; Torres, L. M.B.; *Anal. Chim. Acta* (1988) no prelo.