UM MÉTODO DE CODIFICAÇÃO SEMI-AUTOMÁTICO ASSISTIDO POR COMPUTADOR PARA O *SISTEMAT*

Vicente de Paulo Emerenciano, Gilberto do Vale Rodrigues
Instituto de Química - USP - São Paulo - CP 20.780 - SP
Jean Pierre Gastmans
Faculdade de Engenharia - UNESP - Guaratinguetá - CP 205 - SP

Recebido em 15/12/92; cópia revisada em 5/5/93

A new technique to improve the speed of data codification to be used on SISTEMAT system, an expert program for structure elucidation of Natural Products. The method is based on drawing of the structure on the screen using only ASCII characters. The draw is used by the program to build the Coded Vector that describes the chemical properties of compounds. Tests carried on our laboratory have shown that the method here proposed is ten to one faster than the old one.

Keywords: computer techniques; codification; artificial intelligence.

INTRODUÇÃO

O SISTEMAT é um Sistema Especialista[1], em desenvolvimento, que tem por objetivo auxiliar o químico de Produtos Naturais, no processo de determinação estrutural. O seu funcionamento se baseia em técnicas de Inteligência Artificial, cujas bases teóricas foram publicadas nesta revista[2]. Resumidamente, podemos dividir os programas que compõem o sistema em dois grupos: programas utilizados para a montagem e correção dos bancos de dados, e aplicativos, que usam os bancos de dados para o processo de determinação estrutural. Atualmente dispomos de vários programas aplicativos que utilizam principalmente dados da espectrometria de ressonância magnética de ^{13}C e de massas, além de dados botânicos. A utilização destes programas foi também publicada recentemente[3].

Para atingir um grande potencial de aplicação, o sistema necessita de dados. Quanto mais número de dados o sistema tiver sobre uma classe de substâncias, maior será a confiabilidade das propostas estruturais fornecidas pelos aplicativos. Estes dados devem ser fornecidos ao computador na forma codificada. A codificação se faz necessária devido ao grande volume de memória requerida para armazenar os dados de uma substância: informações estruturais, dados físico-químicos e dados botânicos. Os principais Sistemas Especialistas usados em determinação estrutural foram revistos e comparados com o SISTEMAT[4]. Todos estes sistemas se baseiam em algum tipo de codificação[5,6] cujas características ideais são: facilidade de elaboração, versatilidade, compactação acessíveis a programas escritos em linguagens de alto nível[7,8].

A fase crítica para o crescimento dos bancos de dados é a obtenção do vetor codificado, que representa a estrutura da substância. Originalmente, esta codificação era feita manualmente usando um determinado conjunto de regras. O processo era muito moroso e freqüentemente levava o codificador a cometer erros que só muito mais tarde eram percebidos. Objetivando aumentar o volume e a confiabilidade dos dados, idealizamos um novo processo de codificação semi-automático assistido pelo computador. Nosso sistema dispensa o uso de mesas digitalizadoras, como no sistema DARC[9].

LINGUAGEM DE PROGRAMAÇÃO

O SISTEMAT foi originalmente programado em FORTRAN[7]. Esta linguagem apresenta muitos aspectos positivos no que diz respeito à manipulação de operações matemáticas, mas é muito pobre em recursos de manipulação de periféricos, como por exemplo a tela do computador.

Para melhorar a interface com o usuário, e em consequência aumentar a produtividade, passamos a usar a linguagem de programação Pascal[8]. Esta linguagem apresenta vários recursos sofisticados entre os quais podemos destacar:

- programação estruturada, que possibilita escrever códigos fontes mais claros e de fácil manutenção por outros programadores;
- controle de tela, que permite escrever e ler informações em qualquer parte da tela do computador;
- geração de código executável menor e mais rápido.

O PROGRAMA CODISIS

Existem dois tipos de bancos de dados no SISTEMAT: os bancos estruturados e os bancos fontes. Os dados são primeiramente armazenados juntos nos bancos fontes, ou seja, vetor codificado da estrutura, dados botânicos e dados físico-químicos, para uma determinada substância, são armazenados em um único arquivo. Posteriormente, cada informação é colocada em um arquivo específico. Assim temos os bancos de vetores codificados, bancos de nomes triviais, bancos de dados botânicos e assim por diante. Os bancos específicos são denominados bancos estruturados.

Anteriormente era necessário seguir o seguinte processo para armazenar dados no SISTEMAT:

- executar um programa chamado INICSIS, que cria os bancos estruturados;
- executar o programa VERISIS, que cria um banco fonte, armazena os nomes triviais de todas as substâncias que o mesmo vai conter e verifica a existência dos mesmos nomes já armazenados anteriormente;
- executar FONTESIS, que recolhe os dados preliminares da substância: dados botânicos, classe química, esqueleto, vetor codificado e referência bibliográfica.

Resolvemos juntar as funções dos três programas citados anteriormente em um único programa denominado CODISIS, que passa a ser responsável por todo o processo de armazenamento dos dados de uma substância no computador, além de permitir corrigir ou mesmo excluir um ou todos os dados.

Não se trata de uma simples junção de programas, mas alteração de procedimentos que tornam muito mais eficiente a

construção de bancos de dados no SISTEMAT. O CODISIS é todo controlado por "menus" que permitem um fácil manuseio, mesmo por um usuário iniciante, com pouco conhecimento em computação.

O ponto crítico do armazenamento é a codificação da estrutura. Para acelerar esta codificação o programa trabalha com o desenho feito na tela do computador. A fim de melhorar o desempenho do usuário, os desenhos podem ser gravados em arquivos e lidos posteriormente para evitar redesenha-los. A gravação pode ser feita em dois arquivos, que contém esqueletos e estruturas prontas.

O "menu" principal apresenta oito opções:

Opção 0 - sai do programa voltando ao sistema operacional;

Opção 1 - define os nomes dos arquivos de esqueletos e substâncias;

Opção 2 - define o nome do arquivo de banco fonte que vai ser montado;

Opção 3 - permite desenhar estruturas na tela do computador e grava-las em disco;

Opção 4 - permite a obtenção do vetor codificado para uma estrutura. Inicialmente, é ativada automaticamente a opção 3 para que se possa desenhar a estrutura ou fazer a leitura da mesma se esta estiver gravada em disco. Com a estrutura pronta na tela, é iniciada a codificação que será explicada posteriormente;

Opção 5 - permite armazenar todos os dados para uma substância. É solicitado o nome trivial e o número da ficha da substância. Em seguida é feita uma verificação para determinar se a mesma substância já foi arquivada anteriormente. Em caso afirmativo, são solicitadas apenas as informações referentes à nova ocorrência botânica e à referência bibliográfica. Se a mesma for nova no sistema, será ativada automaticamente a opção 3 para desenhar a estrutura ou recupera-la a partir do disco. Em seguida será ativada a opção 4 para a obtenção do vetor codificado. Finalmente, será mostrada uma tela onde serão recolhidos os demais dados constantes da Tabela 1;

Tabela 1. Dados recolhidos pelo programa CODISIS, incluindo um exemplo da literatura[11]. Obs.: 51 em nosso código representa o periódico Phytochemistry.

Item	Exemplo
Classe	Sesqui-lactona
Esqueleto	Guaiano
Código da Revista	51
Ano	88
Página	1771
Família	Asteraceae
Gênero	Liabum
Espécie	Floribundum
Índice de Evolução Morfológica	72

Opção 6 - substitui o programa INICSIS na criação dos bancos estruturados;

Opção 7 - permite a correção dos dados armazenados para uma substância. É mostrada uma tela contendo todos os dados da substância. A substituição na tela de qualquer um dos dados seguido da tecla ENTER, corrige no disco o mesmo dado. Nesta opção, também é possível alterar o "status" de uma substância para OUT, que faz com que o registro de dados da mesma seja ignorado pelos demais programas que compõem o SISTEMAT;

Opção 8 - permite eliminar definitivamente uma substância marcada com "status" OUT do arquivo em disco.

DESENHO DA ESTRUTURA

A primeira fase da codificação consiste em desenhar a estrutura da substância na tela do computador.

A tela do computador é dividida em duas partes: a parte esquerda é reservada para o desenho da estrutura e é composta por 50 colunas. As 30 últimas colunas da tela são usadas para comunicação com o usuário. É mostrado um "menu" com todos os elementos de desenho permitidos. Podem ser vistos os nomes dos arquivos abertos para uso, além de instruções para gravação e recuperação de estruturas em disco. As teclas de movimentação do cursor, permitem desloca-lo por toda a área de desenho. Uma vez atingida a posição adequada, basta apertar uma tecla de ligação ou átomo. Deve-se seguir a convenção de colocar dois elementos de ligação entre dois átomos ligados.

A área de desenho pode ser visualizada como uma matriz de 50x50 elementos. Cada elemento da matriz pode conter apenas um dos elementos de desenho possíveis. Estes são classificados em dois grupos: átomos e ligações. A Tabela 2 mostra os elementos de desenho utilizados pelo CODISIS. Os elementos de ligação incluem a capacidade de representar ligações com informações estereoquímicas. A Figura 1B mostra um exemplo usando os elementos de desenho.

Tabela 2. Elementos de desenho utilizados pelo programa CODISIS.

Ligação	Tecla	Átomo	Tecla
Simples	I / \ -	Carbono	C
Dupla	=	Carbono Aromático	A
Dupla *cis*	$	Junção *cis* de Anel	J
Tripla	#	Oxigênio	O
Alfa	.	Nitrogênio	N
Beta	^	Nitrogênio Aromático	Z
		Enxofre	S
		Fósforo	P
		Iodo	I
		Flúor	F
		Cloro	L
		Bromo	R
		Qualquer Macronó	M

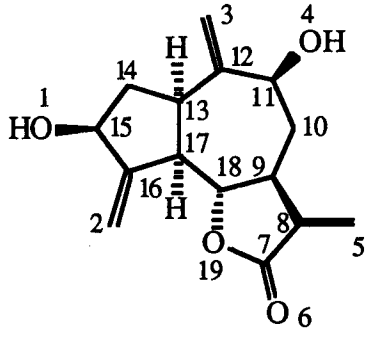
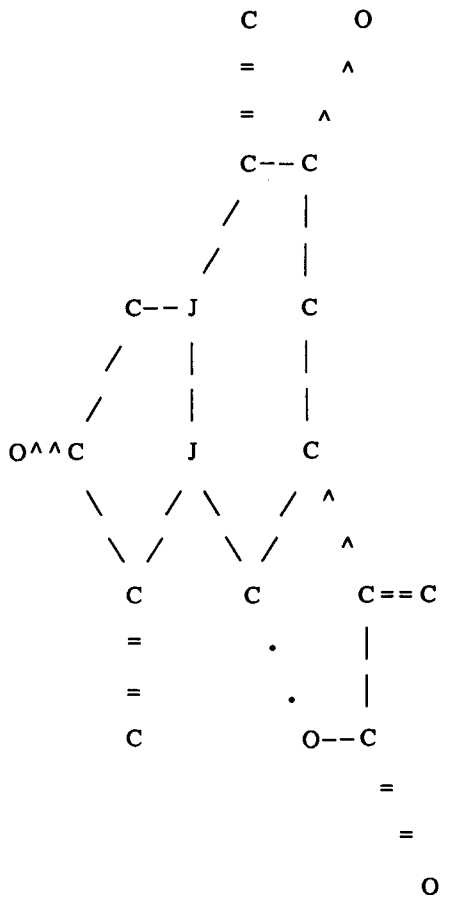


Figura 1A. Estrutura de um sesquiterpeno lactonizado com a numeração dos átomos a ser usada na codificação.



Arquivos em uso:
Esqueletos = LAC-SESQ. ESQ
Substâncias = LAC-SUBS. ESQ
Fonte = GIL008
0-Limpa a tela
Esqueleto 1-LE; 3-GRAVA
Substância 2-LE; 4-GRAVA
5-Redefine arquivos
Para desenho use:
Ligações:

Simples	| / \ -		
Dupla ou Tripla	=	ou	#
Dupla CIS			$
Alfa ou Beta	.	ou	^
Átomos:			
Carbono/Aromático	C	ou	A
Junção CIS			J
Oxigênio			O
Nitrogênio/Aroma.	N	ou	Z
Enxofre/ Fósforo	S	ou	P
Iodo/Flúor	I	ou	F
Cloro/Bromo	L	ou	R
Qualquer MACRONO			M

Figura 1B. *Tela de desenho do programa CODISIS, mostrando como exemplo a estrutura de uma substância (Figura 1A) da classe dos sesquiterpenos lactonizados com esqueleto guaiano.*

A possibilidade de gravar uma estrutura em disco agiliza bastante o processo. Em nosso caso particular, inicialmente montamos um banco com todos os esqueletos da classe em questão. Para desenhar a estrutura de uma substância, recuperamos o desenho de seu esqueleto e apenas acrescentamos ou modificamos os grupos funcionais. Pode-se gravar a estrutura já funcionalizada em um arquivo separado dos esqueletos, pois freqüentemente as estruturas podem diferir apenas na estereoquímica de uma ligação, e nesse caso ganha-se tempo em relação; à opção de se começar pelo esqueleto.

CODIFICAÇÃO DA ESTRUTURA

A partir do desenho na tela do computador, pode-se obter o vetor codificado de acordo com as regras estabelecidas pelo SISTEMAT[2]. O processo é semi-automático, pois ainda requer a interação com o usuário em alguns aspectos. As 30 últimas colunas da tela, que antes eram usadas para instruções de desenho, são nesta fase utilizadas para permitir ao usuário acompanhar o funcionamento do programa durante a codificação.

Cada átomo no desenho é denominado nó. O processo de codificação segue os seguintes passos:

1) O programa identifica automaticamente os nós monoatômicos (um nó é monoatômico se está ligado a um único nó que por sua vez está ligado a pelo menos dois outros nós);

2) O usuário marca os demais nós, formando o menor número possível de cadeias. Quanto maior o número de cadeias usadas, maior será o espaço de memória necessário para armazenar o vetor codificado. Após compactação, os dados deverão ocupar o máximo de 100 bytes. A divisão em várias cadeias não influi na qualidade dos dados armazenados. A marcação é feita posicionando o cursor sobre o átomo e apertando a tecla ENTER. Para iniciar uma nova cadeia deve ser teclado N sobre o primeiro átomo da nova cadeia. Ao final da marcação o programa codifica automaticamente, o início e final de todas as cadeias, os nós monoatômicos e ligações não incluídas nas cadeias;

3) O programa identifica e codifica automaticamente todos os tipos e posições dos heteroátomos presentes no desenho e seus números atômicos;

4) Freqüentemente é necessário saber para um dado átomo no desenho, quantos e de que tipo são os demais átomos ligados a ele. É necessário também saber o tipo de ligação entre os átomos. Para responder a estas perguntas foi construído um procedimento para obter as respostas. Este procedimento é usado para obter o restante da codificação. Deste modo, são automaticamente identificados e codificados todos os átomos conectados por ligações duplas e triplas;

5) O usuário deve marcar um átomo de cada anel aromático, que não seja comum a dois anéis. A marcação é feita como de costume, posicionando o cursor sobre o átomo e apertando a tecla ENTER;

6) Usando o procedimento descrito acima, são automaticamente identificados e codificados os átomos conectados por ligações com informações estereoquímicas: ligações *alfa* e *beta*, além de ligações transanulares de configuração *cis*;

7) A presença da letra M no desenho da estrutura leva o programa a requisitar do usuário a definição do tipo de macronó presente. Um macronó é qualquer substituinte ligado ao esqueleto da substância, como por exemplo, o grupo acetato. Caso o macronó não tenha sido previamente codificado, a codificação pode ser feita com o auxílio do próprio CODISIS;

8) A última etapa da codificação consiste na definição da orientação das ligações entre os átomos, o que é denominado de vetor passo. Usando as regras estabelecidas para o SISTEMAT o computador é capaz de obter este vetor automaticamente a partir do desenho da tela.

Completada a obtenção do vetor, é feito o cálculo do número de oxidação da substância. Este é um parâmetro usado em estudos quimiotaxonômicos[10] que o SISTEMAT calcula automaticamente para cada molécula armazenada. Devido à aplicações taxonômicas, pode ser necessário retirar alguns átomos, especialmente macronós não definidos como tal, antes do cálculo do NOX. Usando o desenho ainda na tela do computador, o usuário pode marcar, como já mostrado anteriormente, os átomos a serem retirados do cálculo.

Depois de calculado o NOX, o mesmo é mostrado na área de interface com o usuário, juntamente com o vetor codificado e a estatística do número de elementos de codificação usados e o número de bytes a ser ocupado pela descrição da estrutura após a etapa de compactação.

O exemplo da Figura 1A submetido ao CODISIS para codificação mostrou os seguintes resultados:

Vetor Codificado: 0719-11516121108071317180919O7-
20801040619-31602031208050706-6-
1010408-218-71317-9213202010202
0807080705030505
Vetor = 55 dados
Compactação = 49 bytes
NOX = -10

DISCUSSÕES E CONCLUSÕES

O programa CODISIS torna o processo de codificação de estruturas no SISTEMAT muito mais rápido e fácil de ser aprendido. Isto se deve em grande parte ao uso da linguagem de programação Pascal. O programa fonte possui aproximadamente 2200 linhas e gera um código executável de apenas 72.736 bytes, bem menor se comparado ao espaço ocupado pelos três programas que ele substitui, 112.950 bytes.

Outro aspecto importante é o grau de confiabilidade do código obtido, que está livre de erros humanos do usuário. Este fato, torna o processo de montagem dos bancos de dados do SISTEMAT uma tarefa possível de ser realizada por alunos de iniciação científica, que normalmente não tem um conhecimento profundo de computação e mesmo de química.

Em nosso grupo de trabalho os bancos tem crescido a uma taxa 10 vezes maior se comparado com o processo antigo de codificação. Testes realizados em nosso laboratório mostraram ser possível armazenar 100 substâncias por dia de trabalho. As estruturas dos esqueletos já haviam sido desenhadas anteriormente. O desenho foi obtido a partir do esqueleto. Em seguida foi feita a codificação e a entrada dos dados complementares: classe química, esqueleto, referência bibliográfica e dados botânicos. Estima-se que existam hoje em dia aproximadamente 3000 diterpenos descritos na literatura. De posse deste levantamento bibliográfico, em menos de dois meses poderíamos armazenar todas as estruturas no computador.

O exemplo mostrado no item **Codificação da Estrutura** permite verificar o grau de dificuldade do processo de entrada de dados do método antigo. Todos os 120 algarismos que compõem o vetor codificado, teriam que, primeiramente serem obtidos usando as regras de codificação e depois digitados no computador. A possibilidade de erro era muito grande, além da demora para realizar ambas as tarefas. Usando o CODISIS, basta desenhar a estrutura da Figura 1, marcar as cadeias e o processo de codificação está concluído.

OBTENÇÃO DOS PROGRAMAS

O programa CODISIS, assim como os demais programas que compõem o sistema SISTEMAT podem ser obtidos junto aos autores. Os programas contam com pequenos manuais que ajudam a compreender sua utilização. Os interessados podem entrar em contato com os autores no Instituto de Química - USP - São Paulo - SP.

AGRADECIMENTOS

Os autores agradecem o auxílio financeiro concedido pela Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP e Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq.

REFERÊNCIAS

1. Weiss, S. M. & Kulikoski, C. A.; Guia Prático para Projetar Sistemas Especialistas, 1988, Livros Técnicos e Científicos, Rio de Janeiro.
2. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Borges, J. H. G. & Emerenciano, V. P.; *Quím. Nova,* (1990), **13**, 10.
3. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Borges, J. H. G. & Emerenciano, V. P.; *Quím. Nova,* (1990), **13**, 75.
4. Emerenciano, V. P.; *Quím. Nova,* (submetido).
5. Lindsay, R.K.; Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project, McGraw-Hill, New York, 1980.
6. Gray, N. B. A.; Computer Assisted Structure Elucidation, Wiley, New York, 1986.
7. Microsoft FORTRAN[77], versão 3.31, Manual do Usuário, Microsoft Corporation, 1985.
8. Turbo Pascal, versão 6.00, Manual do Usuário, Borland International, 1990.
9. Dubois, J. E., Carabedian, M. & Dagane, I.; *Anal. Chim. Acta,* (1984), **158**, 217.
10. Gottlieb, O. R.; Micromolecular Evolution, Systematics and Ecology, Springer-Verlag, Berlin, 1982.
11. Jakupovic, J.; Schuster, A.; Bohlmann, F. & Dillon, M. O.; *Phytochemistry,* (1988), **27**, 1771.

Publicação financiada pela FAPESP