DETERMINAÇÃO ESTRUTURAL COMPUTADORIZADA. PRINCÍPIOS, MÉTODOS, PERSPECTIVAS E MITOS

Vicente de Paulo Emerenciano
Instituto de Química - USP - CP.20780 - CEP.01498-970 - São Paulo - SP

Recebido em 7/12/92; cópia revisada em 5/5/93

The main Expert Systems used for structural elucidation of organic compounds are reviewed. The perspectives of future works in these field are discussed under the Natural Product Chemist point of view.

Keywords: expert systems; principles; methods; perspectives.

1. INTRODUÇÃO

A determinação estrutural de moléculas orgânicas tem sido um campo de trabalho extremamente interessante. A introdução dos métodos físicos de análise nos anos 60, começou a substituir as laboriosas tarefas de determinação de estruturas usando processos degradativos. Vários trabalhos pioneiros usando as duas técnicas (RMN protônica e degradações) apareceram neste período[1].

À medida que as técnicas avançavam, surgiam problemas cada vez mais complexos que dependiam do talento do químico para resolver verdadeiros quebra-cabeças. Estes problemas geralmente não têm um ponto certo por onde começar, nem têm um procedimento adequado a ser seguido rigorosamente e que garantam uma solução correta.

Com o advento dos computadores começou-se a pensar em arquivar os dados espectrais disponíveis para ajudar os químicos a resolverem seus problemas. Pensou-se em começar pela Espectrometria de Massas (EM) pois existiam vários espectrômetros em funcionamento e o espectro de massas pode ser reduzido facilmente a dois vetores (m/z e intensidades relativas) usando várias técnicas matemáticas já conhecidas.

A facilidade de obtenção de espectros de massas e a disponibilidade de técnicas matemáticas para sua confrontação resultou na criação de grandes bancos de dados cuja utilidade prática será discutida a seguir.

A proliferação de aparelhos de Ressonância Magnética (RMN), inicialmente focalizando ^{1}H e posteriormente ^{13}C , sua utilização rotineira por químicos orgânicos forneceu um número cada vez maior de espectros registrados na literatura e alguns grupos de pesquisa começaram a utilizar estes dados tentando criar os chamados Sistemas Especialistas (SE) que pudessem ajudar os químicos que se dedicam à elucidação estrutural.

Hoje, vários sistemas estão em fase operacional e outros em construção. Pretendemos abordar neste artigo a sua eficácia na resolução de problemas complexos, as principais metodologias empregadas e o futuro desta linha de pesquisa.

2. PRINCÍPIOS

2.1 Os Sistemas Especialistas

As técnicas de determinação estrutural fazem parte de um conjunto de estratégias utilizadas por uma disciplina chamada de Inteligência Artificial (IA). Este nome nasceu durante um Congresso[2] onde especialistas em várias áreas se reuniram para discutir os avanços no campo naquela época.

A IA, em um sentido amplo, tenta resolver problemas, usando técnicas computacionais, simulando o raciocínio humano. Seria preferível chamar este campo, ou os programas que dele emanam de Sistemas Especialistas, porém o termo IA acabou aceito e usado indiscriminadamente.

Várias subdivisões são conhecidas atualmente para a IA. Entre elas destacam-se a interpretação da linguagem natural[3], a robótica[4], a tradução assistida por computador[3] e aplicações em ciências básicas como a Matemática[5]. A terminologia em IA é repleta de termos técnicos específicos. O mais comum é a "Heurística". Segundo Polya[6], Heurística ou "ars inviniendi" é um ramo de estudo não bem delimitado, pertencente à Lógica, à Filosofia ou à Psicologia, cujo objetivo é o estudo dos métodos de descoberta e da invenção. A Heurística Moderna procura compreender os processos mentais solucionadores de problemas. Na prática, as heurísticas encurtam os caminhos para se chegar a uma solução de um problema mas nem sempre garantem que seja encontrada a solução perfeita. Os projetos computacionais em IA aplicados à Química Orgânica são chamados de heurísticos, porque geralmente simulam algum processo mental do químico quando resolve problemas, porém existe alguma controvérsia se estes projetos são verdadeiramente heurísticos.

A literatura apresenta atualmente vários compêndios sobre IA[7].Alguns livros abrangentes e que podem ser compreendidos por pesquisadores de diversas áreas são bastante úteis[3,4]. As técnicas bem gerais de construção de SE são bem descritas[8], e os aspectos filosóficos dos impactos da IA no fim do século XX são abordados de maneira bastante clara em dois artigos com opiniões divergentes[9,10].

As principais fontes de informação para quem trabalha no campo ou queira nele ingressar são:

1. Journal of Chemical Information and Computer Sciences.
2. Analytical Chemistry
3. Analytica Chimica Acta,
4. Chemomometrics and Intelligent Laboratory Systems
5. Journal of Organic Chemistry
6. Journal of American Chemical Society
7. Angewandte Chemie

Algumas aplicações de sistemas bem conhecidos apareceram recentemente nas três últimas revistas citadas acima.

Nos anos 70 a preocupação dos pesquisadores era com a resolução automática de teoremas, a compreensão da linguagem natural e programas para jogos. As técnicas de resolução de problemas empregadas em vários campos da pesquisa começaram a ser empregadas na Química, principalmente em Síntese Orgânica onde foram criados vários sistemas para o planejamento de reações orgânicas no computador[11] e em de-

terminação estrutural. O grande projeto realizado nesta área, na Universidade de Stanford, é conhecido como DENDRAL e é amplamente citado em livros textos básicos de IA[3] como o primeiro sistema especialista completo usado rotineiramente pelos químicos orgânicos. Outros sistemas apareceram deste então. Alguns podem ser considerados comerciais e são amplamente divulgados como o DARC[12].

Como o projeto DENDRAL teve sua origem ainda nos anos 60, o objetivo teórico inicial era criar um gerador de isômeros para substâncias orgânicas[13]. No início, havia o interesse em gerar apenas representações topológicas das estruturas. A geração de estereoisômeros apareceu no projeto bem recentemente[14].

Processos de determinação estrutural sejam eles realizados por químicos ou quando se tenta usar técnicas computacionais, utilizam várias informações. Os idealizadores do projeto DENDRAL estavam atentos a este fato e tentaram direcioná-lo para um enfoque multi-espectral, ou seja, explorar dados vindos de todas as fontes espectrométricas disponíveis.

Depois do pioneirismo do DENDRAL quase todos os outros SE seguiram seus passos e podem ser enquadrados no esquema da Figura 1.

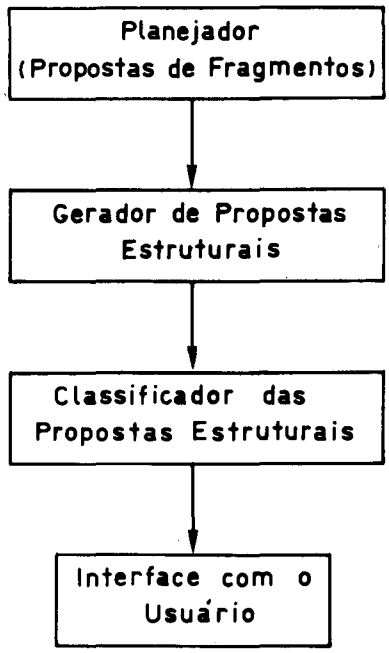


Figura 1. Esquema geral de um sistema especialista em determinação estrutural.

2.2 A geração de propostas estruturais

O gerador é basicamente um programa[15] capaz de agrupar um número de pedaços ou fragmentos e ligá-los respeitando as regras de valência. O procedimento seqüencial que realiza este processo deve ser exaustivo, isto é, deve garantir que todas as combinações serão feitas. Na maior parte dos casos o gerador cria um número muito grande de soluções quando são fornecidos muitos fragmentos e às vezes o número de propostas a serem testadas não é viável.

2.3 A previsão espectral

A previsão espectral é um processo dependente de regras empíricas[11] ou de bancos de dados[16]. Em EM vários grupos têm trabalhado bastante com regras empíricas[17] e com classes de substâncias específicas[17].

A previsão de RMN ^{13}C baseada em bancos de dados tem como fundamento a seguinte regra: se um átomo de carbono entra em ressonância a uma dada freqüência devido a certo ambiente químico (conseqüentemente magnético), então em uma situação análoga ele entrará em ressonância em campo parecido. Por situação análoga entende-se por vizinhanças iguais a distâncias compreendendo quatro ligações. Ou seja, o que se conhece como nível "delta". Necessita-se então de um banco de dados contendo átomos de carbono, seus respectivos deslocamentos químicos e a descrição de todos os seus vizinhos até este nível para que seja realizada uma previsão de um espectro de RMN ^{13}C. Estes bancos são criados por programas, que por sua vez, são alimentados com espectros inteiros, e com uma descrição(codificação) da topologia da molécula. Obviamente, os efeitos conformacionais que afetam o ambiente químico e magnético são desprezados mas os espectros teóricos obtidos são muito próximos dos reais[16].

Ainda seguindo o caminho do DENDRAL a previsão de espectros de EM não evoluiu muito. Muito se fez em análise de EM computadorizada, mas basicamente as técnicas confiáveis se limitam à confrontação espectral e ao reconhecimento de padrões[17]. Dentro do projeto DENDRAL tentou-se criar regras para propor um espectro teórico de EM[18] porém a relação entre os espectros previstos e o espectro real foi muito pobre. Ainda assim os espectros teóricos obtidos servem para eliminar estruturas extremamente aberrantes criadas pelo gerador quando este não conhece o esqueleto da substância problema.

A previsão de RMN ^{1}H e de espectros no infravermelho (IV) não tem sentido para a maioria dos produtos naturais com estruturas complicadas. Para RMN ^{1}H a sobreposição de sinais em campo alto, principalmente em terpenóides, desencorajou bastante a pesquisa neste campo. Alguns trabalhos de previsão de RMN ^{1}H usando bancos de dados são conhecidos[19].

2.4 Tendências Atuais da IA Aplicada à Química

Recentemente, a pesquisa em IA tem se voltado mais para modelos baseados no conhecimento do que em modelos dependentes de bancos de dados. Com esta nova tendência pretende-se criar sistemas cujo fluxograma geral está descrito resumidamente na Figura 2. Nesta nova etapa da pesquisa em

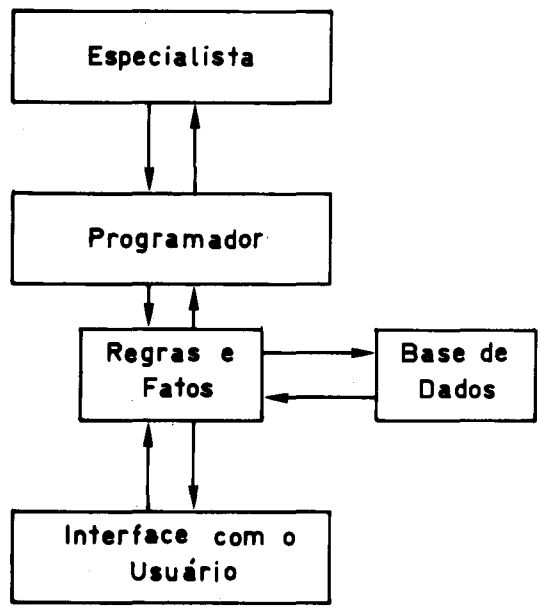


Figura 2. Esquema geral de um sistema especialista contendo um módulo de regras e fatos.

IA pretende-se passar o conhecimento diretamente do especialista para um programa de computador. Esta tendência é exemplificada fora do âmbito da Química pelo Sistema MYCIN, para diagnóstico de infecções bacterianas[20].

Quando não se conhece um especialista em determinado campo muito específico pode-se então obter as regras heurísticas através de programas que vasculham dentro de banco de dados. Isto foi feito no DENDRAL para EM e para algumas classes de substâncias o sistema conseguiu aprender e explicar algumas regras de fragmentação. Nosso grupo trabalha no mesmo sentido com RMN ^{13}C[21,22,23] porém direcionando a procura de regras para esqueletos representativos de substâncias naturais.

Sistemas recentemente montados segundo o esquema geral descrito na Figura 2 também foram construídos usando dados de IV. O mais conhecido é o PAIRS (Program for Analysis of IR Spectra). Sua novidade[11] é associar a presença de um sinal com a possibilidade de se encontrar um determinado grupo funcional. O interpretador de IV contido no PAIRS utiliza basicamente tabelas de absorções e fornece probabilidades de se encontrar grupos funcionais específicos.Para isto lança mão de outra técnica muito usada em IA chamada de "Raciocínio probabilístico"[3]. As subrotinas contidas no PAIRS são basicamente sequências "if...then...else" criadas em uma linguagem especial denominada CONCISE idealizada especialmente para este sistema. Obviamente não se pretende chegar a propostas estruturais com base em só em análise de IV. A utilidade do interpretador de IV é a capacidade de afirmar a existência de grupos funcionais ou de eliminá-los. No segundo caso é muito útil para eliminar fragmentos vindos da análise de RMN ^{13}C. O sistema CASE[24], em sua nova versão, denominada SESAMIN[25] incorpora um interpretador de IV com este objetivo.

3. ANÁLISE DO DESEMPENHO DE ALGUNS SISTEMAS ESPECIALISTAS

3.1 O DENDRAL

O projeto DENDRAL resultou na publicação de uma série de cerca de 50 artigos resumidos em uma publicação de 1980[13] e em outra de 1986[11]. Recentemente, Gray publicou dois artigos sobre os "mitos e realidades" do projeto"[26,27]. Segundo Gray, existe muita diferença entre o trabalho realizado por estudantes em laboratórios voltados à Química Orgânica de Produtos Naturais e os procedimentos envolvidos nos processos de elucidação empregando o DENDRAL. Geralmente o Químico trabalha pensando em um esqueleto básico e em uma classe de substâncias. O DENDRAL não incorpora um módulo automático de identificação dos esqueletos na fase de planejamento e tem que utilizar o conhecimento auxiliar do Químico.

3.1.1 Análise do Warburganal

Em 1982 os autores publicaram um longo artigo[28] mostrando a elucidação estrutural desta substância pelo sistema. O Warburganal (Fig.4) é um sesquiterpeno com esqueleto drimano, com fórmula molecular $C_{15}H_{22}O_3$, e que já tinha sido isolado da planta *Warburgia ugandensis*[29]. Após a análise de todos os dados espectrais (feita pelo Químico) foram fornecidos ao programa gerador de estruturas sete fragmentos e outros três foram definidos como impossíveis de existir na molécula (Figura 3). O processo de construção de prováveis estruturas tem que ser interativo, ou seja, precisa da ajuda do usuário em várias etapas, como veremos abaixo. Um detalhe importante dos geradores do tipo do DENDRAL é a necessidade do fornecimento de restrições(constraints). No exemplo acima, um experimento usando dupla radiação em RMN protônica garantiu que o CH2 isolado (C-6) está vizinho à

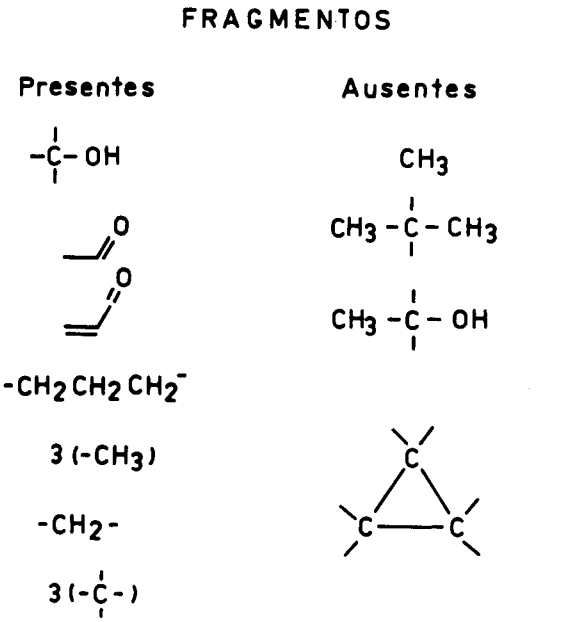


Figura 3. Fragmentos fornecidos ao DENDRAL para gerar a estrutura do Warburganal[28].

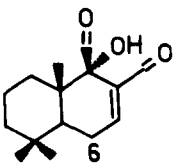


Figura 4. Estrutura proposta para o sesquiterpeno Warburganal e confirmada pelo DENDRAL.

ligação dupla, obrigando que todos as propostas estruturais geradas tenham a seqüência CH2-CH=CH-CHO. Durante o processo de construção o programa usa alguns fragmentos e gera estruturas parciais denominadas casos (cases). Estes são mostrados ao usuário que os elimina quando são incompatíveis ou os deixa para o processo final de criação. Esta também é outra técnica usada para diminuir a explosão combinatória durante o processo de construção. Para o Warburganal, o programa gerou 42 estruturas compatíveis com os fragmentos fornecidos. O passo seguinte foi a avaliação das propostas feitas. Neste ponto o programa consulta uma biblioteca de esqueletos[30] para saber se algumas das estruturas criadas incorporam algum esqueleto já isolado. Os autores admitem que lançam mão deste procedimento porque existe grande probabilidade que um sesquiterpeno bicíclico seja de um esqueleto já conhecido. Este exemplo demonstra a dependência do conceito de esqueleto como a grande restrição para que os resultados sejam satisfatórios. A etapa seguinte consiste em prever os espectros de RMN ^{13}C e EM para algumas propostas estruturais que "parecem" substâncias naturais e eliminar novamente aquelas improváveis. Dos 42 candidatos, somente uma das estruturas incorporava um esqueleto completo e era a estrutura proposta no trabalho original[29]. Esta teve a melhor relação entre o espectro de RMN ^{13}C previsto e o espectro real. O mesmo foi feito para EM e também foi observada a melhor relação entre os dois espectros. O DENDRAL tem ainda outro programa capaz de gerar todos os estereoisômeros a partir de uma representação plana da molécula.

3.1.2 Análise do Palustrol

O programa gerador de estruturas, nesta época ainda chamado de CONGEN, foi utilizado na elucidação da estrutura de um outro sesquiterpeno chamado palustrol[31]. A análise espectral forneceu os fragmentos descritos na Figura 5. Com estes fragmentos o CONGEN construiu 272 estruturas. Experimentos com desacoplamento homonuclear em RMN ^1H a 360 MHz revelaram a subestrutura denominada HEP (Fig.6) e com esta restrição o número de estruturas geradas caiu para 88. A desidratação da amostra e a análise do ambiente ao redor da ligação dupla dos produtos obtidos, permitiu inferir a subestrutura denominada PEN (Fig.6) e com esta o CONGEN criou 22 candidatos, destes alguns têm o grupo hidroxila em cabeça de ponte e não seriam coerentes com os resultados da desidratação, logo foram descartados e chegou-se a 7 candidatos.

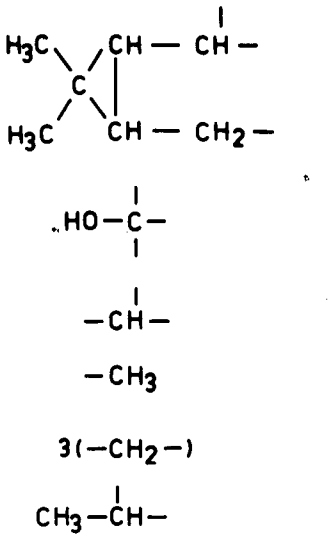


Figura 5. Fragmentos fornecidos ao DENDRAL para gerar a estrutura do Palustrol[31].

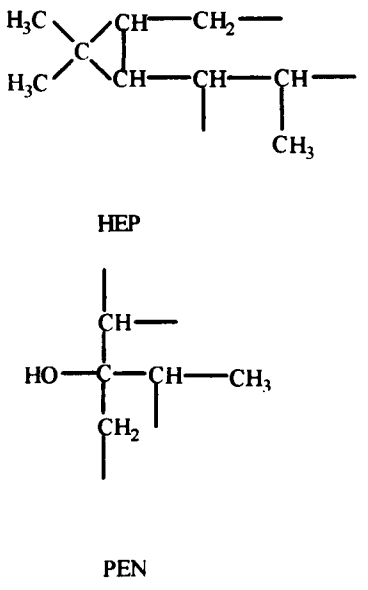


Figura 6. Fragmentos fornecidos ao DENDRAL, após experimentos de desacoplamento homonuclear em RMN 1H, para gerar a estrutura do Palustrol[28].

Quando estes foram analisados para ver se obedeciam a regra do isopreno, somente duas substâncias incorporaram esqueletos de produtos naturais (Figura 7) Somente a substância A tinha um esqueleto conhecido (aromadendrano) e a estrutura proposta correspondia a uma substância já conhecida (palustrol).

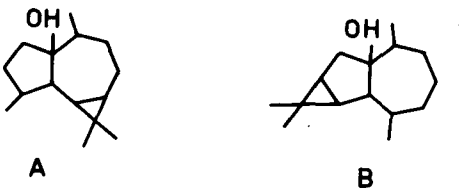


Figura 7. Estruturas propostas para o Palustrol[28], que incorporam esqueletos de sesquiterpenos.

3.1.3 Um diterpeno isolado de *Roylea cacylicina* (Roxb) Briq

Este experimento demonstra bem a dependência de grandes restrições estruturais (como a definição de um esqueleto) para que o gerador funcione bem com moléculas com número de carbono igual ou maior a 20. Os autores[16] trabalharam com os dados de RMN ^1H e ^{13}C, IV e EM de uma substância de fórmula molecular $C_{22}H_{32}O_5$. A dificuldade do problema obrigou os autores a impor, logo de início, uma grande restrição: admitir que estariam se defrontando com um esqueleto labdano(I) ou clerodano (II) (Figura 8). A análise detalhada dos espectros forneceu ainda as outras subestruturas vistas na Figura 8. Com estas restrições o gerador criou 112 substâncias. Após a previsão dos espectros de RMN ^{13}C foi eliminada a maior parte e ficou-se com os 34 restantes. Estes foram verificados para ver se eram compatíveis com o espectro de RMN protônica e foram eliminados mais 21 propostas estruturais. As 13 restantes estão na Figura 9. Neste ponto o DENDRAL não tem ferramentas capazes de distinguir um dos 13 melhores candidatos, ou mesmo de reduzir a lista para um número menor. Segundo os autores, dois argumentos reforçam a hipótese de a substância incorporar um esqueleto labdânico. Primeiro porque só labdanos tinham sido isolados da planta. Segundo porque o produto da desidratação apresenta um grupo metilvinílico incompatível com o uma desidratação de uma substância que incorporasse o esqueleto clerodânico. As estruturas 1 e 2 são, contudo, as mais prováveis.

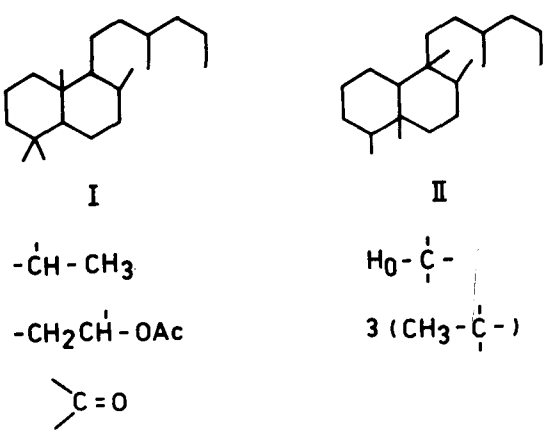
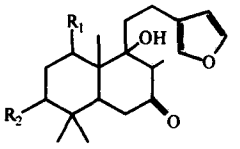
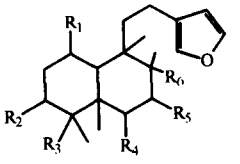


Figura 8. Dois esqueletos de diterpenóides (I e II) e alguns fragmentos estruturais fornecidos ao DENDRAL para evitar a explosão combinatória na elucidação estrutural de um problema cuja fórmula molecular é $C_{22}H_{32}O_5$.



	R_1	R_2
1	OAc	H
2	H	OAc



	R_1	R_2	R_3	R_4	R_5	R_6
3	OAc	H	H	=O	H	OH
4	H	OAc	H	=O	H	OH
5	OAc	H	OH	=O	H	H
6	OAc	H	H	H	=O	OH
7	H	OAc	H	H	=O	OH
8	H	OAc	OH	H	=O	H
9	OAc	H	OH	H	=O	H
10	=O	H	H	OAc	H	OH
11	H	=O	H	H	OAc	OH
12	H	=O	H	OAc	H	OH
13	H	=O	OH	H	OAc	H

Figura 9. *Estruturas prováveis para $C_{22}H_{32}O_5$ propostas pelo DENDRAL, obedecendo requisitos biogenéticos.*

3.1.4 Elucidação da Diasina

A Diasina é um diterpeno com esqueleto labdano modificado devido à migração de um grupo metila (substância I, Fig. 11), com fórmula molecular $C_{21}H_{24}O_7$, isolado da planta *Croton diasii* Pires[32] Vários requisitos estruturais foram passados ao programa gerador e são descritos com detalhes no artigo[33]. O programa forneceu 145 estruturas com base nestes requisitos. Novamente os pesquisadores do DENDRAL tentam imitar o raciocínio do Químico que desvendou a estrutura, usando também propostas biogenéticas, caso contrário, ele não conseguiria reduzir o número de candidatos gerados. Neste caso foram colocados como restrições as subestruturas descritas na Figura 10, que são partes de um esqueleto labdânico. Trabalhando com estas subestruturas o gerador reduziu o número de propostas estruturais para 2 (Figura 11). Novamente, vê-se a dependência do raciocínio baseado em biogênese para que o gerador funcione. Neste caso é interessante notar que ambas as propostas estruturais descritas na Figura 11 são compatíveis com as informações espectrais. Como a substância em questão tinha um esqueleto novo, verifica-se a utilidade do gerador ser exaustivo.

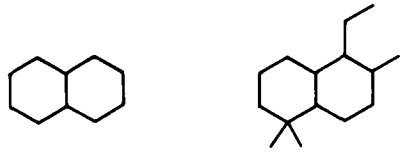
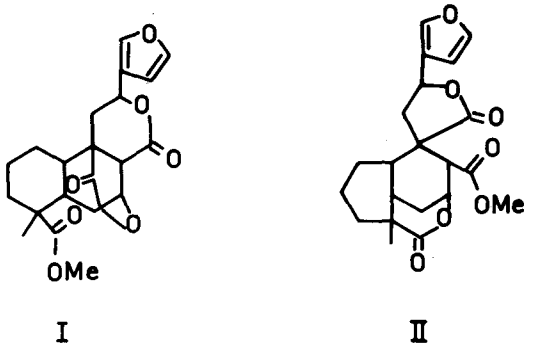


Figura 10. *Fragmentos de esqueletos, fornecidos ao DENDRAL, como restrições impostas para elucidar a estrutura da Diasina[32]*



I II

Figura 11. *Estruturas propostas para a Diasina pelo DENDRAL[33].*

3.1.5 Introdução de Requisitos Biogenéticos no Gerador

Os pesquisadores do DENDRAL introduziram uma modificação no gerador para tentar criar possíveis isômeros de esqueletos de mono e sesquiterpenóides[34]. Os autores propuseram um método com base em considerações mecanísticas para as possíveis ciclizações de duas ou três unidades isoprênicas. O programa é capaz de propor esqueletos já conhecidos ou outros ainda não isolados usando reações de ciclização, migrações de hidrogênio e de grupos alquílicos e juntando as unidades da maneira "cabeça-cauda". O número de esqueletos previstos mesmo usando este limitado conhecimento biogenético é muito maior do que os esqueletos já isolados de plantas. Os autores concordam que o número de esqueletos naturais é muito pequeno em relação aos esqueletos prováveis criados pelo gerador. Eles também sugerem que este reduzido número é uma das razões para o sucesso da análise estrutural feita através de analogias com terpenóides já conhecidos. A abordagem desta modificação do gerador poderá ainda ser fonte de investigação se novas fontes de conhecimento forem introduzidas.

3.2 O Sistema DARC/EPIOS

O sistema DARC/EPIOS foi construído na França[12] e seu mecanismo de elucidação estrutural é bem diferente do DENDRAL. Ele se baseia na metodologia conhecida como "elucidação por intersecção progressiva de subestruturas ordenadas". Resumidamente, podemos entender o DARC/EPIOS como um sistema especialista com um banco de dados contendo milhares de subestruturas com a descrição de carbonos ressonantes em diversos ambientes. Essas subestruturas são conhecidas como ELCOs (ambientes que são limitados, concêntricos e ordenados) originados de 11000 espectros retirados da literatura. No processo de elucidação estrutural do DARC/EPIOS tenta-se criar estruturas compatíveis com o espectro de RMN 13C de uma substância realizando simultaneamente os processos de interpretação e atribuição espectral. A geração de estruturas é feita progressivamente e alguns detalhes deste gerador foram publicados[35]. A diferença para o DENDRAL é que o próprio interpretador fornece os fragmentos para o gerador, e até certo ponto não é necessária a ajuda do químico. Porém não existem exemplos na literatura mostrando a utilização do sistema em elucidação de moléculas complexas. Um estudo feito pelo nosso grupo mostrou que com moléculas com 20 átomos de carbono o número de soluções geradas pelo DARC/EPIOS sem a ajuda de um Químico seria enorme e impossível de análise[36]. Os autores do DARC/EPIOS publicaram a utilização do sistema na elucidação das substâncias descritas na Figura 12[35].

Para a substância I (Figura 12) 14 estruturas foram geradas devido à ambigüidade dos sinais a 24.4t ou q e 27.6q ou t,

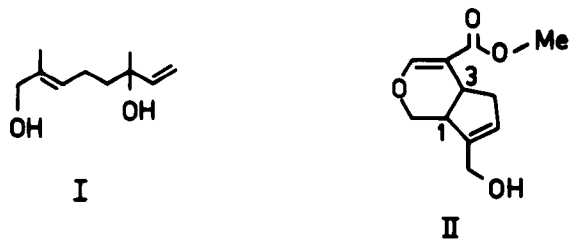


Figura 12. Exemplo de duas moléculas cujas estruturas foram propostas pelo sistema DARC/EPIOS[35].

quando esta ambigüidade foi resolvida, isto é, foi atribuída a multiplicidade correta a cada sinal, a única estrutura proposta foi a estrutura correta. Para a segunda molécula a solução foi encontrada em 7.65 segundos usando um computador VAX 11/780. Quando foi introduzida a informação sobre a obrigatoriedade da conexão entre C-1 e C-3 a solução foi encontrada em 1.27s.

3.3 O Sistema ACCESS

Este sistema foi desenvolvido na BASF por Bremser e colaboradores. Sua codificação é muito semelhante àquela do sistema DARC/EPIOS, porém tem um banco de dados muito maior, com cerca de 100.000 espectros de RMN 13C. Dois artigos ilustram bem algumas aplicações do sistema. No primeiro[37] os autores demonstram como um dos procedimentos mais simples usados em bancos de dados podem ajudar na determinação estrutural de moléculas com número de carbono em torno de 30. Usando técnicas de confrontação espectral, e aplicando um índice de semelhança que correlaciona uma amostra com os espectros contidos em um banco de dados, os autores demonstram que é possível inferir grandes partes de uma estrutura. No caso do ACCESS, o banco é extremamente grande e diversificado, e bons resultados foram conseguidos. É interessante notar neste trabalho, que mesmo os processos de confrontação com grandes bancos de dados podem ser auxiliados pela intuição ou pela experiência do Químico, usando programas interativos. Em um artigo mais recente[38] Bremser descreve um protótipo de um laboratório de análise bastante informatizado e que começa a ser robotizado.

Além dos processos de confrontação, o ACCESS tem um mecanismo interpretativo também semelhante ao DARC/EPIOS. Bremser publicou a elucidação do monoterpeno genipina (Figura 12, substância II) que já tinha sido simulado pelo DARC/EPIOS. O ACCESS levou 79 segundos de processamento em um computador de grande porte para realizar o mesmo trabalho e chegou à uma proposta estrutural correta[39].

O índice de semelhança criado por Bremser também foi extremamente útil na identificação de esqueletos de triterpenos pentacíclicos[40].

3.4 O Sistemat

O sistema de inteligência artificial, denominado Sistemat, que estamos desenvolvendo no Brasil, foi projetado para múltiplas aplicações[41,42]. A primeira , e que está sendo explorada no momento, visa basicamente o armazenamento de dados espectrais e a criação de programas para determinação estrutural. Um interesse posterior será explorar as informações botânicas contidas no bancos de dados para fins quimiotaxonômicos.

As principais diferenças do Sistemat com os outros sistemas descritos anteriormente são: a)utilização de micro computadores em ambiente DOS ou WINDOWS, permitindo que a construção dos bancos de dados seja feita em qualquer microcomputador. Estes bancos podem ser agrupados em único computador desde que haja capacidade de armazenamento em disco rígido. b) possui um sistema de codificação e compactação de dados espectrais, e de referências bibliográficas extremamente eficiente justamente para permitir a utilização de computadores de pequeno porte. c) permite a codificação das substâncias sem a necessidade de mesas digitalizadoras. Estas são acessórios especiais, pouco comuns, compostos por uma caneta óptica que permite a digitalização de uma imagem para o computador. Esta imagem é depois utilizada por um programa que a transforma em código. d) O armazenamento de dados espectrais de substâncias de origem natural juntamente com as informações botânicas (família, gênero, espécie) e com a classe de produto natural (terpenóide, alcalóide, etc.), permite que o próprio sistema encontre regras de classificação de esqueletos com base em RMN 13C.

O Sistemat foi projetado inicialmente usando a nossa experiência com outros programas gerenciadores de bancos de dados como o dBASE[43]. Este é muito útil para armazenar informações botânicas, a classe das substâncias e algum tipo de codificação linear das mesmas, porém o uso deste tipo de codificação para fins espectrais é limitado. A entrada dos dados no Sistemat era até 1991 a etapa limitante para o crescimento dos bancos de dados. Recentemente esta codificação passou a ser feita de maneira semi-automática e a velocidade de crescimento dos bancos de dados cresceu de um fator de 10 para 1. Apesar de todos os sistemas incorporarem os módulos descritos na figura 1, não nos preocupamos em criar programas para simulação espectral. Nossa técnica de previsão foi abandonada temporariamente porque o modelo por nós desenvolvido, apesar de ter um bom desempenho, era muito lento para criarmos um banco de dados com milhares de subestruturas[44]. A convivência dentro de um grupo de pesquisa em Química de Produtos Naturais mais a análise dos sistemas que vimos nos itens anteriores nos levou a trabalhar de uma maneira inversa, ou seja, direcionar o trabalho para a criação de técnicas de identificação automática de esqueletos, primeiro porque este é o ponto onde o químico se apoia para tentar elucidar uma estrutura e segundo porque os programas dos outros grupos de pesquisa não têm capacidade de tratar problemas complexos em Química de Produtos Naturais sem as informações auxiliares sobre os esqueletos ou parte dele. Vários programas foram criados pelo nosso grupo com este objetivo, entre eles está o SISPICK que é uma nova versão do programa PICKUP, descrito anteriormente[21], que cria regras de identificação de esqueletos usando RMN 13C. As regras fornecidas pelo SISPICK são passadas à um outro programa denominado SKELPRED que pode usa-las para prever o esqueleto de uma nova substância.

Usando um procedimento de confrontação foi também criado o programa SISCONST[36] que propõe um esqueleto a partir da confrontação do espectro de RMN 13C de uma amostra, com um banco de dados, usando uma faixa de tolerância bem ampla. Depois propõe subestruturas com o maior número possível de carbonos. Este programa já foi utilizado na comprovação de um sesquiterpeno (substância II, Fig. 13) inédito isolado em nosso laboratório[45] e deverá ter seu mecanismo de interpretação de dados de RMN 13C melhorado nos próximos anos.

Outras substâncias publicadas na literatura foram usadas para testar o SISCONST (substâncias I e III, Fig. 13). Com a sobreposição das subestruturas fornecidas pelo programa chegou-se facilmente à mesma proposta estrutural dos artigos onde tinham sido publicados. Recentemente o mesmo programa auxiliou na elaboração da estrutura de um triterpeno isolado na família Melastomataceae (substância IV, Fig. 13). O Sistemat tem atualmente um programa chamado SISSKE[46] que consegue reconhecer usando o sistema de codificação, qual o

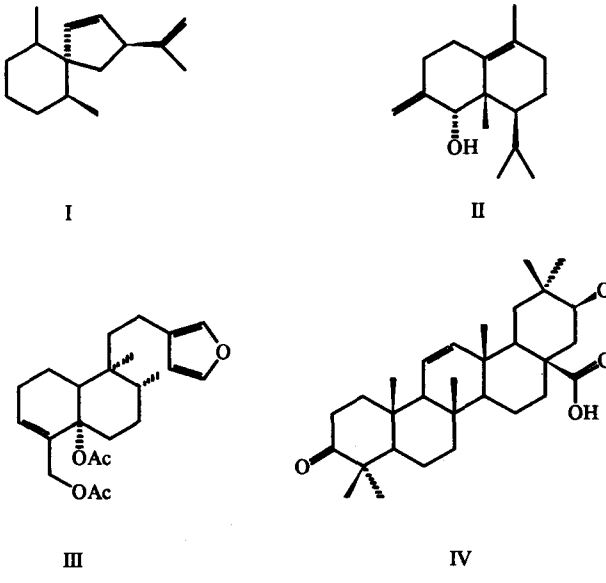


Figura 13. Exemplo de quatro moléculas usadas para testar o sistema especialista SISTEMAT[42].

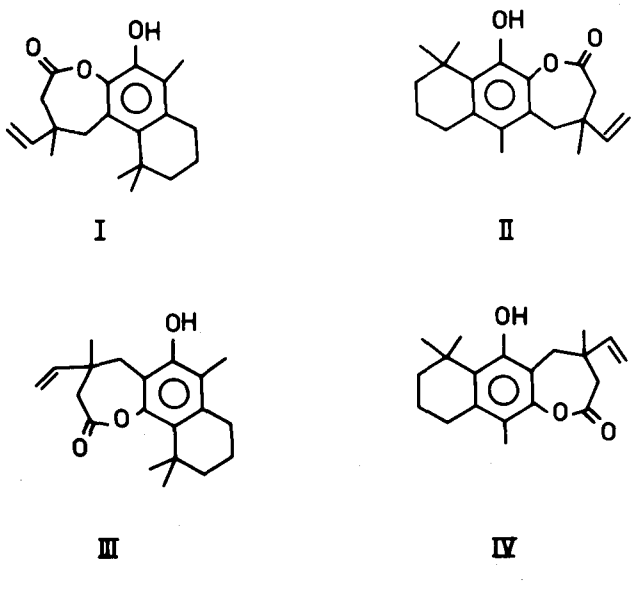


Figura 14. Propostas estruturais para o Velloziolídeo[47], usando o sistema especialista CASE[24].

esqueleto a que pertence determinada molécula. Este programa tem um mecanismo de aprendizado interessante pois os esqueletos vão sendo definidos um a um para ele, com as respectivas numerações biogenética. Quando o programa se defronta com uma nova molécula ele tenta reconhecer o esqueleto e grava um vetor que faz a transformação entre a codificação que obedece as regras do sistema e a outra, biogenética. O Sistemat tem no momento cerca de 5000 espectros de RMN 13C de terpenóides em geral. Após o banco de dados ser considerado representativo para substâncias predominantemente alifáticas outras classes deverão ser acrescentadas.

3.5 O Sistema CASE

Este sistema[24,25] tem outro gerador de estruturas bastante eficiente, porém, necessita de restrições durante o processo de construção de propostas estruturais. Um exemplo bastante interessante da utilização do CASE foi a elucidação de estrutura do Velloziolídeo[47] utilizando técnicas de IA e informações auxiliares. A novidade do CASE em relação aos outros sistemas é que ele possui mecanismos para eliminar estruturas que não são plausíveis com a fórmula molecular e com dados de RMN protônica. Após a análise do espectro de RMN 13C as subestruturas encontradas, chamadas de ACFs (atom-centered fragments) são analisadas para verificar sua consistência com os dados de RMN protônica. Para isto o sistema tem dois módulos interpretadores (HNMRPRUNE e 2DNMRPRUNE). Os ACFs podem também ser eliminados por um interpretador de IV. A intenção dos construtores do sistema é chegar a uma lista de ACFs pequena e consistente com todos os dados espectrais disponíveis chamada SHORTLIST muito parecida com aquela denominada GOODLIST do DENDRAL[13]. O sistema CASE também está apto para receber informações sobre a conectividade entre carbonos oriundas de experimentos como INADEQUATE. No exemplo do Velloziolídeo o número de combinações possíveis do gerador poderia levar a 2×10^{12} se o sistema não tivesse um mecanismo para eliminar os ACFs inválidos, antes da fase de criação. As quatro estruturas propostas para o Velloziolídeo estão na Figura 14 sendo que a estrutura descrita pelos autores no artigo original[47] é a de número 2.

4. EQUIPAMENTOS E LINGUAGENS

A evolução de equipamentos disponíveis na área de informática tem se dado tão rapidamente que é difícil definir o que é moderno hoje, pois neste momento novas invenções certamente estarão sendo lançadas no mercado. Em termos de tendências passamos rapidamente de computadores de grande porte tipo VAX, DEC-10, Burroughs para pequenos PCs (Personal Computers) e estes estão evoluindo e sendo cada vez mais utilizados em laboratórios de pesquisa. No meio, entre os PCs e os computadores de grande porte estão as estações de trabalho vendidas a um preço razoável. A tendência atual é passagem de grandes bancos de dados existentes em computadores de grande porte para as estações de trabalho com grande capacidade de armazenamento em discos rígidos. A construção dos SE requer o conhecimento de uma linguagem de computação para que seja feita a transferência de dados para o computador e para a construção dos programas.

Teoricamente, um algoritmo pode ser passado para qualquer linguagem, porém algumas delas são mais úteis para determinadas aplicações. Em geral trabalhou-se com o FORTRAN[48] para aplicações rigorosamente matemáticas, e com o LISP[49] para outras mais simbólicas. As primeiras versões do DENDRAL foram escritas em LISP e em outra linguagem denominada ALGOL[50]. O DARC foi basicamente escrito em PASCAL[51] e o CASE[24], na sua versão mais recente[25], tem partes em FORTRAN e em PASCAL. O Sistemat foi originalmente escrito em FORTRAN porém está sendo reescrito em PASCAL. Esta linguagem tem boas vantagens para o desenvolvimento de softwares pois tem grande potencial de manipulação de dados numéricos, fácil acesso a banco de dados, gera um código fonte estruturado e permite o amplo controle de periféricos.

Uma linguagem mais recente que tem sido divulgada como a ferramenta do futuro em IA é linguagem PROLOG[52,53]. As implementações atualmente existentes da linguagem, principalmente aquelas disponíveis para computadores do tipo IBM/PC não possuem recursos suficientes para se desenvolver uma grande sistema especialista só com base em PROLOG. Poucas aplicações são descritas na literatura exemplificando a utilização de PROLOG no âmbito da Química. Alguns métodos de codificação de substâncias orgânicas foram elaborados com esta linguagem[54] porém não foram usadas na criação de grandes bancos de dados.

5. O Futuro dos Sistemas Especialistas em Determinação Estrutural

5.1 Os mitos

Quando lemos em um livro de IA sobre o impacto dos SE como o DENDRAL neste campo, temos a impressão de que o gerador de estruturas é o módulo mais inteligente do sistema. Os exemplos escolhidos neste artigo não comprovam isto. O gerador foi considerado por um dos principais integrantes do grupo[26,27] como um algoritmo combinatorial para agrupar subunidades moleculares que satisfaçam um determinado peso molecular. O modelo utilizado no DENDRAL, conhecido em IA como "Planejar-Gerar-Testar", não tem nenhum componente de IA na fase de planejamento. O gerador não possui filtros heurísticos para reduzir, sem a ajuda do Químico, o número de propostas estruturais, ou eliminar estruturas contendo partes que seriam heuristicamente descartadas pelo Químico. Em outras palavras, não há "conhecimento" embutido no gerador. As pesquisas no campo da IA com aplicações usadas em Química tendem a ser direcionadas para programas que gerem eles próprios suas heurísticas, numa forma de aprendizado bem diferente do humano, mas criando sistemas cada vez mais independentes do especialista.

5.2 Perspectivas

Várias demonstrações da utilidade dos SE residem em exemplos oriundos da Química de Produtos Naturais porque é onde aparecem problemas intrigantes para serem resolvidos. Uma forma de conhecimento adjacente (meta-conhecimento) que deveria ser acoplado à estes SE poderia ser o conhecimento biogenético. Vimos por exemplo, no caso da Diasina como o usuário tem que fornecer informações biogenéticas ainda durante a fase de construção. Como conhecimento biogenético entendemos as sequências de formação de esqueletos, mesmo que hipotéticas, e seus principais processos de degradação e rearranjos. Para resolver problemas práticos em Química de Produtos Naturais em um tempo útil será também necessário o que denominamos de filtro botânico, ou seja, o conhecimento da origem do material de onde as substâncias foram isoladas. Neste filtro botânico poderão ser incluídas informações sobre o parentesco entre os taxons para que se façam inferências sobre a possibilidade de uma determinada planta possuir certas classes e/ou esqueletos específicos, desde que uma espécie ou gênero mais próxima já tenha apresentado estas classes ou esqueletos.

5.3 Métodos de Aquisição do Conhecimento

Os teóricos da IA consideram que o passo limitante para o crescimento de um SE é a integração do especialista com os construtores dos módulos, ou seja, com os programadores propriamente ditos. Para realizar a transferência de dados para a máquina é necessário um "especialista em especialistas", conhecido dentro do campo da IA como Engenheiro do conhecimento. Este deve ter alguma competência em informática, saber estabelecer um bom diálogo com os especialistas e, ao mesmo tempo, deve saber vencer suas resistências em transmitir seus conhecimentos armazenados através de anos[20].

5.3.1 A Integração com Químicos de Produtos Naturais

Obviamente, a resolução de problemas relacionados à Química de Produtos Naturais no computador pode incorporar heurísticas vindas do especialista, porém requer também técnicas matemáticas para resolução de problemas de rotina, que têm às vezes 30 a 40 átomos de carbono. Para restringir o problema (o que em IA se chama diminuir o espaço de busca) o Químico usa algum conhecimento prévio como sinais característicos de ligações duplas, de carbonos carbinólicos etc. A partir daí, segue em direção de uma proposta estrutural com base em algum esqueleto conhecido. Neste ponto, ele realiza dezenas de operações mentais do tipo "se...então".

Este método leva a sequências de raciocínio como a descrita a seguir: SE existem evidências do espectro estar relacionado a um triterpeno E não existem muitas evidências para um grande número de funções oxigenadas E existe um número de sinais de carbonos olefínicos absorvendo em uma região característica Então começa-se a pensar em uma proposta estrutural parcial como aquela descrita na Figura 15. Após verificadas as consistências com a fórmula molecular pode-se recomeçar o processo pensando em novo esqueleto. É difícil colocar este procedimento no computador pois a ordem do passos não pode ser determinada precisamente.

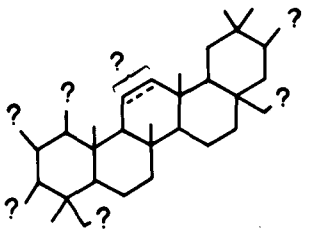


Figura 15. Esqueleto básico de um triterpeno, mostrando as principais posições funcionalizadas.

6. CONCLUSÕES

A determinação estrutural de substâncias orgânicas pode ser vista como um desafio típico para os interessados em IA. Todos os problemas onde atualmente os especialistas são melhores que os programas merecem ser investigados e alguma atenção deve ser dada a maneira pela qual os especialistas resolvem seus problemas. Já foi visto também em outras áreas que o número de soluções aumenta exponencialmente com o número de variáveis. Em determinação estrutural podemos considerar como variáveis o número de átomos (n). Geralmente o tempo necessário para resolver estes problemas é exponencial (x^n). Um dos interesses da IA é reduzir este tempo a uma função no mínimo polinomial, com o auxílio de heurísticas. Em Química Orgânica existe ainda o agravante da complexidade dos arranjos possíveis para uma determinada fórmula molecular. Um sistema para o futuro deverá saber agrupar métodos matemáticos com métodos heurísticos, e interagir com os usuários de modo a adquirir mais dados e mais conhecimento sem a necessidade de ser reprogramado.

REFERÊNCIAS

1. Lavie, D. e Gottlieb, O. R.; *Chem. & Ind.*, (1960), 929.
2. Feigenbaum, E. A. e Feldmann, J. A.; (Eds.) Computers and Thought, McGraw-Hill, Nova York, 1963.
3. Rich, E.; Inteligência Artificial, McGraw-Hill, São Paulo, 1988.
4. Winston, P. H.; Inteligência Artificial, Livros Técnicos e Científicos, Rio de Janeiro, 1987.
5. Davis, R. e Lenat, D.; Knowledge-Based Systems in Artificial Intelligence, McGraw-Hill, 1982.
6. Polya, G.; How to Solve It, Princeton University Press, 1957.
7. Barr, A. e Feigenbaum, E. A.; The Handbook of Artificial Intelligence, Kaufmann, Los Altos, 1981.

8. Weiss, S. M. e Kulikowski, C. A.; Guia Prático Para Projetar Sistemas Especialistas, Livros Técnicos e Científicos, Rio de Janeiro, 1988.
9. Searle, J. R.; *Scient. Amer.*, (1990), **262**,19.
10. Churchland, P. M. e Churchland, P. S.; *Scient. Amer.*, (1990), **262**,26.
11. Gray, N. A. B.; Computer Assisted Structure Elucidation, Wiley, Nova York, 1986.
12. Attias, R.; *J. Chem. Inf. Comp. Sc.*, (1983), **23**,102.
13. Lindsay, R. K.; Buchanan, B. G.; Feigenbaum. E. A. e Ledberg, J.; Applications of Artificial Intelligence in Organic Chemistry: The Dendral Project, McGraw-Hill, Nova York, 1980.
14. Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H. e Djerassi, C.; *Org. Magn. Reson.*, (1981), **15**,375.
15. Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G. e Djerassi, C.; *J. Org. Chem.*, (1981),**46**,1708.
16. Lindley, M. R.; Gray, N. A. B.; Smith, D. H. e Djerassi, C.; *J. Org. Chem.*, (1982), **47**, 1027
17. Bachiri, M. e Mouvier, G., *Org. Mass Spec.*, (1976), **11**, 1272.
18. Gray, N. A. B.; Carhart, R. E.; Lavanchy, A.; Smith, D. H.; Varkony, T.; Buchanan, B. G.; White, W. C. e Cleary, L.; *Anal. Chem.*, (1980), **52**, 1095.
19. Egli, H.; Smith, D. H. e Djerassi, C.; *Helv. Chim. Acta*, (1980), **65**, 1898.
20. Buchanan, B. G. e Shortliffe, E. H.; Rule-Based Expert Systems. The MYCIN Experiments of the Heuristic Programming Project. Addison-Wesley, Reading, 1984.
21. Gastmans, J. P.; Furlan, M. e Emerenciano, V. de P.; *Computer and Chemistry*, (1990), **14**, 75.
22. Gastmans, J. P.; Furlan, M. e Emerenciano, V. de P.; *Quantum Chemistry Program Exchange*, (1989), **9**, 134.
23. Lins, A. P.; Furlan, M.; Gastans, J.P. e Emerenciano, V. de P.; *An. Acad. Bras. Ciências*, (1991), **63**, 141.
24. Shelley, C.; Hays, T.; Munk, M. E. e Raman, R.; *Anal. Chim. Acta.*, (1978), **103**, 121.
25. Christie, B. D. e Munk, M.e.; *J. Am. Chem. Soc.*, (1991), **113**, 3570.
26. Gray, N. A. B.; *Chemom. and. Intel. Lab. Syst.*, (1988), **5**, 11.
27. Gray, N. A. B.; *Chemom. and. Intel. Lab. Syst.*, (1988), **5**, 37.
28. Djerasssi, C.; Smith, D. H.; Crandell, C. w.; Gray, N. A. B.; Nourse, J. G. e Lindley, M.r.; *Pure and appl. Chem.*, (1982), **54**, 2425.
29. Kubo, I.; Lee, W. Y.; Petter, M.; Pilkiewicz, F. e Nakanishi, K.; *J. Chem. Soc. Chem. Commun.*, (1976), 1013.
30. Devon, T. K. e Scott, A. I.; Handbook of Naturally Occuring Compounds, Vol. II, Terpenes. Academic Press, Nova York, 1972.
31. Cheer, C. J.; Smith, D. H.; Djerasssi, C.; Tursch, B.; Braekman, J. C. e Daloze, D.; *Tetrahedron*, (1976), **32**, 1807.
32. Alvarenga, M. A.; Gottlieb, H. E.; Gottlieb, O. R.; Magalhães, M. T. e DaSilva, V.O.; *Phytochemistry*, (1978), **17**, 1773.
33. Smith, D. H.; Gray, N. A. B.; Nourse, J. G. e Crandell, C. W.; *Anal. CHim. Acta*, (1981), **133**, 471.
34. Smith, D. H. e Carhart, R. E.; *Tetrahedron*, (1976), **32**, 2513.
35. Carabedian, M.; Dagane, I. e Dubois, J. E.; *Anal. Chem.*, (1988), **60**, 2186.
36. Fromanteau, D. L. G.; Gastmans, J. P.; Vestri, S. A.; Emerenciano, V. de P. e Borges, J. H. G.; Computer and Chemistry (submetido).
37. Bremser, W.; Klier, M. e Meyer, E.; *Org. Mag. Reson.*, (1975), **7**, 97.
38. Bremser, W.; *Angew. Chem. Int. Ed. Engl.*, (1988), **27**, 247.
39. Bremser, W. e Fachinger, W.; *Magn. Reson. Chem.*, (1985), **23**, 1056.
40. Maia, C. M. B. de F.; Braz Filho, R. e Emerenciano, V. de P.; *An. Acad. Bras. Ciências*, (1990), **62**, 119.
41. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Borges, J. H. G. e Emerenciano, V. de P.; *Química Nova*, (1990), **13**, 10.
42. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Borges, J. H. G. e Emerenciano, V. de P.; *Química Nova*, (1990), **13**, 75.
43. Jones, E. dBASE III - Guia do Usuário, McGraw-Hill, São Paulo, 1987.
44. Gastmans, J. P.; Zurita, J. C.; Sahão Junior, J. e Emerenciano, V. de P.; *Anal. Chim. Acta*, (1989), **213**, 85.
45. Lordello, A. L. L.; Análise Química de Três Variedade de Ocotea pretiosa (Mez) Ness - Dissertação de Mestrado, IQUSP, 1992.
46. Borges, J. H. G. SKELETON - Um Sistema Especialista na Determinação Automática de Esqueletos Carbônicos de Produtos Naturais - Dissertação de Mestrado, DQ-UFSCAR,1991.
47. Pinto, A. C.; Gonçalves, M. L. A.; Braz Filho, R.; Neszmelyi, A. e Luckacs, G.; *J. Chem. Soc. Chem. Commun.*, (1982), 293.
48. Hell, M. E.; Linguagem de Programação Estruturada Fortran 77, McGraw-Hill, São Paulo, 1986.
49. Brooks, R. A.; Programming in Common LISP, John Wiley & Sons, Nova York, 1985.
50. Segre, L. M.; Linguagem de Programação Algol, Editora Campus, Rio de Janeiro, 1981.
51. Turbo Pascal. Programmer's Guide, Borland International, 1983.
52. Bratko, I.; PROLOG, Programming for Artificial Intelligence, Addison-Wesley, Reading, 1990.
53. Coelho, H. e Cotta, J. C.; PROLOG by Examples - How to Learn, Teach and Use it, Springer-Verlag, Berlin, 1988.
54. Armstrong, J. L.; *J. Chem. Inf. Comput. Sci.*, (1989), **29**, 51.

Publicação financiada pela FAPESP