Quim. Nova, Vol. 30, No. 5, 1347-1356, 2007

Divulgação

ELUCIDAÇÃO ESTRUTURAL DE SUBSTÂNCIAS ORGÂNICAS COM AUXÍLIO DE COMPUTADOR: EVOLUÇÕES RECENTES

Ricardo Stefani e Paulo Gustavo Barboni Dantas Nascimento
Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes, 3900, 14040-901 Ribeirão Preto – SP, Brasil
Fernando Batista Da Costa*
Departamento de Ciências Farmacêuticas, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Av. do Café, s/n, 14040-903 Ribeirão Preto – SP, Brasil

Recebido em 4/4/06; aceito em 1/2/07; publicado na web em 30/07/07

COMPUTER-AIDED STRUCTURE ELUCIDATION OF ORGANIC COMPOUNDS: RECENT ADVANCES. The development of new tools for chemoinformatics, allied to the use of different algorithms and computer programmes for structure elucidation of organic compounds, is growing fast worldwide. Massive efforts in research and development are currently being pursued both by academia and the so-called chemistry software development companies. The demystification of this environment provoked by the availability of software packages and a vast array of publications exert a positive impact on chemistry. In this work, an overview concerning the more classical approaches as well as new strategies on computer-based tools for structure elucidation of organic compounds is presented. Historical background is also taken into account since these techniques began to develop around four decades ago. Attention will be paid to companies which develop, distribute or commercialize software as well as web-based and open access tools which are currently available to chemists.

Keywords: artificial intelligence; computer programmes; organic compounds.

INTRODUÇÃO

A elucidação estrutural de compostos orgânicos é um ramo tão antigo da Química Orgânica quanto ela própria. Durante muito tempo o processo de elucidação estrutural era empírico, baseado em observações e experimentos simples, sendo que basicamente se empregavam processos degradativos e obtenção de derivados, o que muitas vezes induzia a erros. Esta situação começou a mudar a partir da segunda metade do século XX, quando os métodos espectrométricos (espectrometria de massas – EM, espectroscopia no infravermelho – IV, no ultravioleta – UV e ressonância magnética nuclear – RMN) se sedimentaram, permitindo maior precisão e confiabilidade na elucidação de estruturas. Com o passar do tempo, os métodos de separação e purificação de substâncias, as metodologias analíticas e a tecnologia relacionada à espectrometria evoluíram consideravelmente. Como conseqüência, houve aumento considerável da precisão e confiabilidade dos dados obtidos através de análises instrumentais. Isso teve como resultado o crescimento expressivo de dados espectrométricos disponíveis para diversas substâncias orgânicas, os quais passaram a ser organizados e catalogados como qualquer outra propriedade da molécula. Caso um químico tivesse uma coleção de dados espectrométricos disponível, poderia compará-las com os dados experimentais obtidos para uma substância com estrutura desconhecida, evitando que todo o processo de interpretação dos espectros fosse feito a partir da etapa inicial. Assim, surgiram as primeiras coleções – "handbooks" – de dados espectrométricos. A comparação dos dados espectrométricos obtidos de uma amostra desconhecida com aqueles disponíveis na literatura tornou-se o *modus operandi* mais comum em elucidação estrutural e os "handbooks" passaram a ser um excelente auxílio nesta tarefa. Esta abordagem, a mais convencional de todas, é também empregada em ferramentas computacionais de elucidação estrutural, como será discutido posteriormente. Entretanto, os "handbooks" tornaram-se cada vez mais volumosos e complexos; realizar uma simples busca em um deles sem possuir nenhum conhecimento prévio sobre a classe de substância à qual a amostra desconhecida pertencia, muitas vezes, era o mesmo que procurar uma agulha no palheiro.

Alguns grupos de pesquisa perceberam essas limitações e aproveitaram o desenvolvimento da informática[1] para criar as primeiras formas de tratamento de informações químicas através de computador. Estes grupos pesquisaram maneiras de se representar uma estrutura química no computador[2], sendo que outros, como o do pioneiro projeto DENDRAL[3], pesquisaram como se poderia automatizar o processo de elucidação estrutural de substâncias utilizando-se computadores. A esta abordagem iremos nos referir como "elucidação estrutural auxiliada por computador", uma tradução do famoso jargão inglês CASE ("Computer-Assisted Structure Elucidation"). Após os projetos pioneiros, vários outros surgiram nos últimos 40 anos e alguns desenvolvem-se até os dias atuais.

Após os resultados bem sucedidos de alguns destes projetos e com o rápido desenvolvimento de diferentes técnicas computacionais, foram criadas condições para a evolução e o aperfeiçoamento das técnicas que pudessem auxiliar o químico na elucidação estrutural de substâncias orgânicas. Técnicas de Inteligência Artificial (IA) no seu sentido mais amplo passaram a ser empregadas e forneceram excelentes resultados. Uma vez que a demanda pela elucidação estrutural de substâncias orgânicas cresce a cada instante, busca-se aumentar a produtividade, pois torna-se evidente que a etapa limitante do processo de elucidação de uma substância não é mais a geração de dados, mas sim como interpretá-los. Rapidez também é essencial, sendo que já existem programas que realizam tarefas complexas de elucidação estrutural em apenas alguns segundos. Logo, é evidente a necessidade do emprego de técnicas computacionais que possam auxiliar o químico no processo de elucidação estrutural. Desta

*e-mail: febcosta@fcfrp.usp.br

forma, tanto a academia quanto as empresas de pesquisa e desenvolvimento de "software" dedicados à química investiram maciçamente no setor, gerando bons resultados. Tal fato pode ser constatado observando-se o aumento da quantidade de publicações de artigos científicos inerentes ao tema e nos produtos oferecidos pelas empresas. A facilidade ao acesso a programas de computador e a massificação da internet deram o impulso que faltava neste campo. Atualmente, excelentes ferramentas computacionais – comerciais ou não – voltadas para a elucidação estrutural de substâncias orgânicas estão à disposição dos usuários.

Histórico

O projeto pioneiro no ramo de elucidação estrutural automatizada e o mais clássico de todos foi o DENDRAL[2]. Basicamente o programa funcionava com as estratégias clássicas denominadas "planejar-montar-testar". Durante a realização do projeto, foram desenvolvidos diversos algoritmos[4] e técnicas que se tornaram clássicas e são atuais até os dias de hoje. O DENDRAL possuía um banco de dados contendo vários fragmentos de estruturas químicas com seus respectivos deslocamentos químicos. Todos esses fragmentos eram pequenos, com no máximo quatro ou cinco átomos, incluindo heteroátomos. O sistema iniciava confrontando os dados de RMN ^{13}C do espectro-problema com os dados disponíveis em seu banco de dados e então era obtida uma lista de fragmentos compatíveis com os dados do espectro e com a fórmula molecular oriunda de EM. A partir desta fase, o DENDRAL não era mais totalmente automatizado, pois o químico deveria informar ao programa quais grupamentos funcionais ou subestruturas deveriam estar presentes na solução final do problema ("goodlist") e quais deveriam estar ausentes ("badlist"). Após esta etapa o programa elaborava as propostas estruturais. Um exemplo completo de elucidação estrutural utilizando-se o DENDRAL está disponível na literatura[5].

Um dos primeiros programas que foi desenvolvido na década de 60 a partir do DENDRAL era um gerador de estruturas, ou seja, um programa que, partindo da fórmula molecular, podia gerar todos os seus possíveis isômeros. Este gerador foi a base para os sistemas CONGEN ("CONectivity GENerator")[6] e uma extensão deste, o GENOA ("GENeration with Overlapping Atoms")[7]. Um outro programa pioneiro, porém mais recente e avançado que o DENDRAL, foi o DARC/EPIOS ("Direct Access Radar Channel/Elucidation by Progressive Intersection of Ordered Substructures")[8]. A começar pelo banco de dados, este programa utilizava um sistema diferente daquele utilizado pelo DENDRAL. O DARC/EPIOS também possuía um banco de fragmentos estruturais, entretanto, estes eram baseados em um átomo de carbono central ligado a seus respectivos vizinhos α e β juntamente com a descrição dos deslocamentos químicos dos mesmos. Desta forma, haviam subestruturas conhecidas como ELCOs ("Environment Limited and Concentric Ordered")[9], capazes de descrever diversos ambientes químicos. O algoritmo utilizado era capaz de selecionar os ELCOs cujos deslocamentos químicos do átomo central fossem compatíveis com os do espectro-problema. A partir dos ELCOs, a estrutura era gerada. O DARC/EPIOS utilizava um algoritmo mais eficiente que o do DENDRAL, necessitando de um mínimo de interferência do químico.

Um dos programas desenvolvidos por Munk e colaboradores[10] armazenava em seu banco de dados estruturas completas e seus respectivos deslocamentos químicos de RMN ^{13}C. O químico deveria informar os dados experimentais de RMN ^{13}C e o número mínimo de sinais que as soluções deveriam apresentar. O programa fazia uma busca dos dados fornecidos pelo usuário em todo o banco e comparava os dados espectrométricos de cada estrutura do banco com os dados experimentais. Em seguida, o programa filtrava as estruturas que continham dados espectrométricos compatíveis com os experimentais. Aquelas com o mínimo de sinais requeridos eram novamente filtradas, sendo que estruturas duplicadas eram descartadas. Assim, tinha-se uma lista de *n* estruturas compatíveis com os dados experimentais do espectro-problema. Outros sistemas desenvolvidos pelo grupo de Munk incluem o ASSEMBLE[11], o SESAMI[12] e o HOUDINI[13]. O ASSEMBLE se diferencia-se dos outros sistemas por não possuir banco de dados e tampouco trabalhar com dados espectrométricos. Tal sistema lida apenas com a fórmula molecular e todos os isômeros possíveis são gerados a partir desta fórmula molecular. Entretanto, apenas gerar todos os isômeros possíveis para uma dada fórmula molecular é algo inútil, pois o número de isômeros cresce exponencialmente de acordo com o número de átomos presentes na molécula, juntamente com o tempo de computação necessário para gerar tais isômeros. Desse modo, os projetistas do ASSEMBLE adicionaram opções para o usuário poder indicar ao programa fragmentos que deveriam estar presentes ou ausentes na solução final, o que em quimioinformática se denomina restrições ("constraints"). Assim, a partir da interpretação dos dados espectrométricos, o usuário poderia saber se por exemplo uma molécula possuía ou não uma função epóxido ou um sistema α,β-insaturado ligado a uma carbonila e informar ao gerador. O programa ASSEMBLE evoluiu desde então e hoje está disponível em sua versão 2.0, porém paga. O SESAMI ("Systematic Elucidation of Structure Applying Machine Intelligence")[12], também do mesmo grupo de pesquisa, possui um gerador que trabalha primeiro procurando todos os centros quirais possíveis na molécula, em conjunto com as restrições impostas pelo usuário. Seu algoritmo é baseado em tabelas que correlacionam estruturas com características de um espectro. A interpretação de espectros pelo SESAMI inicia-se pela fórmula molecular e a extração de fragmentos compatíveis com dados espectrométricos da substância desconhecida. Esses dados são utilizados como uma segunda lista, que a partir daí é utilizada para gerar as estruturas compatíveis com a fórmula molecular.

Recentemente, Munk e colaboradores ainda desenvolveram o HOUDINI[13], a partir do SESAMI, que é um sistema completo que utiliza dados de RMN mono e bidimensionais e mais a fórmula molecular da substância desconhecida. O HOUDINI possui uma abordagem totalmente diferente dos sistemas anteriores, pois o seu algoritmo envolve primeiro a criação de uma hiper-estrutura com todas as ligações possíveis entre seus átomos. A partir deste ponto, as ligações excedentes vão sendo removidas de acordo com as valências atômicas e correlações bidimensionais, até haver apenas a presença de poucas estruturas compatíveis com os dados experimentais.

O sistema CSEARCH[14] ("Carbon-13 SEARCH"), desenvolvido pelo grupo de Robien, na Universidade de Viena, é baseado no sistema de procura desenvolvido por Munk. Entretanto, possui alguns melhoramentos, tais como predição de deslocamentos químicos baseada em código HOSE[15] ("Hierarchically Ordered Spherical Description of Environment"), procura por grupos funcionais e por similaridade de espectros, sendo que o sistema ainda é capaz de quebrar as estruturas encontradas em fragmentos de até três átomos e combinar estes fragmentos com novas estruturas "on the fly"[16]. Isso significa que o CSEARCH possui um gerador estrutural, ainda que primitivo.

O sistema CHEMICS[17], desenvolvido por um grupo de pesquisadores japoneses, também utiliza algoritmos semelhantes aos usados pelo DENDRAL e DARC/EPIOS e é capaz tratar dados de RMN bidimensionais para descartar fragmentos inválidos durante a geração estrutural. Já o ACCESS[18] é outro sistema que combina um gerador estrutural com busca em uma biblioteca de espectros.

Participação do Brasil – O SISTEMAT

O desenvolvimento do SISTEMAT[19] foi iniciado nos anos 80 e é o único dos sistemas especialistas utilizando IA que foi desenvolvido para múltiplas aplicações além de elucidação estrutural. Uma das aplicações que merece destaque é a utilização de informações sobre ocorrências botânicas das substâncias naturais de origem vegetal existentes no banco, o que permite seu uso para estudos quimiotaxonômicos.

O SISTEMAT é um sistema modular, formado por diversos pequenos programas que executam tarefas específicas, tais como inserção de dados no banco (DATASIS[20]), análise e extração de dados botânicos (SISBOTA[21] e SISTAX[22]) e busca de dados espectrométricos (REGRAS[23] e SISCONST[24]). O sistema pode ser executado sob as plataformas DOS ou Microsoft® Windows e está em contínuo desenvolvimento, sendo que atualmente se estuda incorporar Redes Neurais (RN) artificiais em seus módulos[25]. Dos diferentes programas disponíveis no SISTEMAT para a tarefa de elucidação estrutural, o mais útil é, sem dúvida, o SISCONST. Ele trabalha exclusivamente com dados de RMN ^{13}C e utiliza-os para procurar subestruturas compatíveis com o espectro-problema em todo o banco do SISTEMAT. Porém o usuário é responsável por reunir os fragmentos e montar as propostas estruturais, visto que ainda falta um gerador de estruturas.

ESTRATÉGIAS PRINCIPAIS NA ELUCIDAÇÃO ESTRUTURAL AUTOMATIZADA

No contexto da elucidação estrutural, que envolve a situação onde todas as informações obtidas a partir dos dados espectrais são insuficientes para se propor uma estrutura para uma substância desconhecida, as principais estratégias para se realizá-la de forma automatizada são: *planejamento*, quando os dados disponíveis são confrontados com dados de uma biblioteca e subestruturas são obtidas a partir destes dados; *geração de estruturas*, quando as estruturas químicas são geradas; *validação*, quando se verifica se as estruturas geradas são compatíveis com os dados apresentados, incluindo-se aqui o processo de predição e comparação de espectros. Essa abordagem é conhecida como "planejar-gerar-testar" (Figura 1) e está presente nos sistemas completos para elucidação estrutural automatizada. Contudo, a maioria dos programas é desenvolvida especificamente para apenas uma destas três etapas, funcionando como módulos simples, porém de grande utilidade. O químico ou o espectroscopista são responsáveis por utilizar outros programas para realizar as demais etapas do processo, ou devem realizá-las usando apenas seu próprio raciocínio.

Planejamento

Nesta etapa, os dados espectrais obtidos (IV, RMN, EM) são confrontados com os dados existentes em um banco e aqueles que forem compatíveis com os dados experimentais são selecionados e apresentados ao usuário. Em seguida, os dados seguem para a próxima etapa, a geração de estruturas (Figura 1). Os dados presentes em tais bancos usualmente são estruturas químicas ou fragmentos destas, que geralmente são armazenados como cadeias no formato SMILES[26] ("Simplified Molecular Input Line Entry System")[27], MDL/MOL (ou SDF)[28] ou ainda em algum formato próprio do sistema. Juntamente com esses fragmentos são armazenados os respectivos dados de espectrais. Alguns sistemas de busca de dados de RMN limitam-se a esta única etapa, como o SISCONST do SISTEMAT e o sistema on-line NMRShiftDB.

Geração de estruturas

A etapa seguinte do processo é ao mesmo tempo a etapa limitante e consiste em combinar os fragmentos compatíveis (e seus respectivos dados espectrais) com a fórmula molecular em novas estruturas químicas (Figura 1).

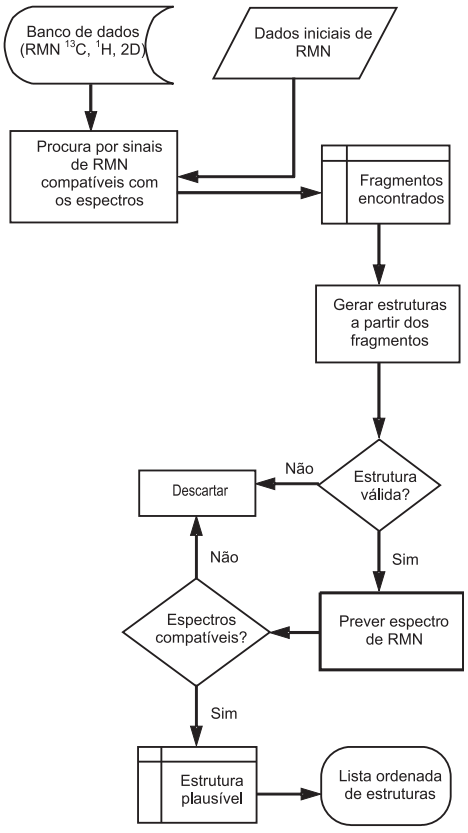

Figura 1. *Diagrama contendo as estratégias principais da elucidação estrutural automatizada (planejar-gerar-testar), tendo como exemplo dados de RMN*

Estas metodologias podem ser agrupadas em dois grandes grupos. O primeiro é do tipo *determinístico* que, através de um algoritmo, vai testar todas as combinações possíveis de acordo com os dados presentes e combinações impostas e, por isso, também são conhecidas por exaustivas. Dentre estas metodologias, estão a montagem[29], a redução[30], ou a combinação das duas, utilizando-se técnicas bidimensionais[31]. Entre os sistemas determinísticos tem-se o DENDRAL, DARC/EPIOS, ASSEMBLE, MOLGEN ("MOLecular GENerator")[32], CHEMICS, SESAMI, HOUDINI e ACD/StrucEluc. O segundo grupo de metodologia é do tipo *estocástico*, fazendo uma geração aleatória, de acordo com um determinado conjunto de dados e combinações possíveis, gerando então estruturas aleatórias. No entanto, tais estruturas são quimicamente corretas e compatíveis com os dados fornecidos, porém não são geradas todas as estruturas possíveis. Dentre os métodos estocásticos, pode-se citar o método de Faulon[33] e Algoritmos Genéticos (AG). O desenvolvimento de metodologias estocásticas para geração estrutural é recente e, até o momento, apenas o SENECA (Faulon + AG) e o GENIUS[34] (AG) as utilizam.

Validação

Nesta etapa, verifica-se se a estrutura gerada é ou não compatível com os dados fornecidos (Figura 1). Pode ser realizada após todas as estruturas terem sido geradas ou em paralelo com a geração estrutural, onde logo após a sua geração é verificado se a estrutura é compatível com os dados do espectro real. Na primeira fase

da validação, são checadas todas as valências e a ordem das ligações para se verificar se são válidas ou não. Isso evita absurdos como, por exemplo, a presença de átomos de carbono tri ou pentavalentes na molécula. Se tudo estiver correto com as ligações e valências, a estrutura é válida (Figura 1). Contudo, deve-se saber se a estrutura gerada possui dados espectrométricos compatíveis com os dados experimentais. Para isso, o mais comum é predizer os dados de RMN ^{1}H e de ^{13}C e depois compará-los com os dados do espectro real. Para essa etapa, podem ser utilizadas metodologias empíricas, tais como regras de adição ("ChemNMR", da CambridgeSoft), procura em bancos de dados utilizando-se código HOSE ("NMRPredict", da Modgraph e os produtos da ACD/Labs, como "ACD/HNMR" ou "ACD/CNMR Predictor") ou ainda IA, como nos sistemas especialistas, com emprego de RN ("SPINUS-WEB", "GENIUS" e "SpecSolv"[35]).

O código HOSE é um método de descrever a vizinhança de um átomo central, sendo muito utilizado para descrever o ambiente químico deste átomo (Figura 2) e, a partir daí, predizer o seu deslocamento químico. O método foi descrito por Bremser[15], em 1978, e consiste em codificar a vizinhança de um átomo central de uma até *n* esferas, onde cada esfera representa átomos de uma até *n* ligações distantes do átomo central. Por exemplo, um código HOSE de duas esferas é capaz de descrever as vizinhanças α e β de um átomo central X, sendo que um de três esferas descreve as vizinhanças α, β e γ do átomo X (Figura 2) e assim por diante. Quanto maior for o número de esferas do código HOSE, mais confiável será a qualidade da predição de deslocamento químico do átomo X.

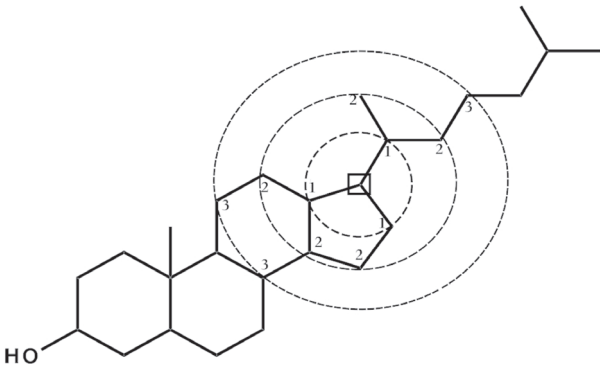

Figura 2. *Vizinhança descrita por um código HOSE de três esferas (vizinhanças α, β e γ) para a estrutura de um esterol. O átomo central está indicado com um quadrado*

As redes neurais artificiais – ou redes neuronais – compreendem uma outra metodologia muito utilizada para a predição de deslocamentos químicos ou de espectros, como IV, RMN ou EM. Por exemplo, as RN têm a finalidade de validar uma determinada proposta estrutural, o que será discutido adiante com maiores detalhes. É uma das técnicas que utilizam IA e que atualmente competem em igualdade com o clássico código HOSE.

PRINCIPAIS METODOLOGIAS

Além das regras peculiares de alguns dos sistemas já descritos – muitas delas envolvendo IA – diferentes metodologias estão sendo amplamente utilizadas, tais como mecânica quântica, redes neurais (RN) e algoritmos genéticos (AG), ou até mesmo combinações destas, sendo que algumas serão discutidas a seguir.

Mecânica quântica

Atualmente é possível utilizar os conceitos da mecânica quântica para o cálculo teórico de grandezas relacionadas à espectrometria de RMN, sobretudo para o átomo de ^{13}C, no auxílio à elucidação estrutural de substâncias orgânicas.

Com a utilização de métodos de estrutura eletrônica, baseados no formalismo de Hartree-Fock-Roothan[36], em conjunto com métodos de inserção da correlação eletrônica, é possível obter valores muito próximos dos experimentais para um dado confôrmero molecular.

Até recentemente, a utilização de técnicas de química quântica estava restrita a moléculas com peso molecular muito baixo, devido ao alto custo computacional para a criação de modelos suficientemente complexos da estrutura eletrônica de moléculas orgânicas que pudessem ser utilizados na predição das suscetibilidades magnéticas. Com funções de base pequenas, a descrição do ambiente molecular não é suficiente e pode levar a erros. Por isso, o cálculo da suscetibilidade magnética deve ser realizado com funções de base extensas[37-39], contendo funções difusas e de polarização, de maneira a obter dados confiáveis.

A técnica GIAO[40] ("Gauge Independent Atomic Orbital") é a maneira mais utilizada para a obtenção das suscetibilidades magnéticas de átomos leves, definindo para cada átomo uma origem do potencial vetorial do campo magnético externo. Há ainda a técnica CSGT[41] ("Continuous Set of Gauge Transformations") que utiliza uma origem única para o campo vetorial magnético externo. Existem ainda outras técnicas, as quais utilizam orbitais localizados IGLO[42] ("Individual Gauge Localized Orbital") e LORG[43] ("Localized Orbital/Local Origin"), porém são menos indicadas por sua maior dependência com a função de base utilizada.

Todas estas técnicas e métodos encontram-se implementados e disponíveis em vários programas comerciais, como por ex. o Gaussian 03.

Os valores obtidos para o deslocamento químico são qualitativamente semelhantes aos experimentais e em muitos casos quantitativamente também. Enfatiza-se que os valores são obtidos para apenas uma conformação, no vácuo, e ajustes paramétricos dos valores teóricos podem compensar pelas diferenças do modelo, buscando menores erros estatísticos.

Um experimento que auxilia a determinação estrutural de substâncias orgânicas é verificar a concordância linear entre dados experimentais de ^{13}C e dados teóricos, pois grandes desvios e pontos fora da reta podem indicar alguma troca na atribuição dos dados experimentais. Uma das vantagens do cálculo é a certeza de qual deslocamento químico corresponde a cada átomo. Para deslocamentos químicos de ^{1}H esta análise é mais complexa, pois se deve considerar a maior suscetibilidade dos prótons, os efeitos de solvente e a conformação da molécula.

Inteligência Artificial (IA) - algoritmos e metodologias

A IA é um dos ramos da computação que pesquisa metodologias para tentar simular o raciocínio humano através de computador, ou pelo menos, no mínimo, tais metodologias tentam reduzir o tempo que o computador gasta para realizar tarefas em que o cérebro humano é melhor que um computador. A IA não é um ramo novo, mas apenas recentemente surgiram as condições consideradas ideais para a proliferação desta técnica, tais como computadores mais velozes, linguagens de programação e algoritmos mais eficientes. Algumas das mais importantes metodologias aplicadas em IA são a heurística, as redes neurais (RN) e os algoritmos genéticos (AG).

A *heurística* é uma das primeiras técnicas em IA e consiste em um conjunto de regras de tomada de decisão. Algumas delas são inseridas pelos especialistas durante o projeto do sistema e outras são inferidas pelo sistema conforme novos casos são apresentados

a este, sendo que as regras vão sendo armazenadas em um banco. Assim, quando aparece um problema semelhante, o sistema é capaz de "julgar" qual é o melhor caminho para a resolução deste. Dentre os sistemas heurísticos, encontram-se o DENDRAL, SISTEMAT, DARC/EPIOS, SESAMI, HOUDINI e ACD/StrucEluc.

As *redes neurais* são um método computacional que simula o funcionamento do cérebro humano e têm a capacidade de aprender a partir de exemplos. Podem ser consideradas como uma "caixa preta" que recebe uma série de estímulos de entrada ("input") e, a partir destes, produz um ou mais dados de saída ("output") (Figura 3). Por ex., recebem dados médicos de um paciente e realizam predições sobre o tipo de doença que ele possui, ou a partir de um espectro de uma substância podem prever sua estrutura. As RN consistem de um conjunto de neurônios e um conjunto de sinapses artificiais, onde um neurônio artificial recebe estímulos e envia sinais para o neurônio seguinte, assim como os neurônios biológicos (Figura 3). Detalhes sobre o funcionamento das RN e suas aplicações em química, bem como em elucidação estrutural de substâncias (RMN, IV e EM) foram anteriormente publicados[44,45].

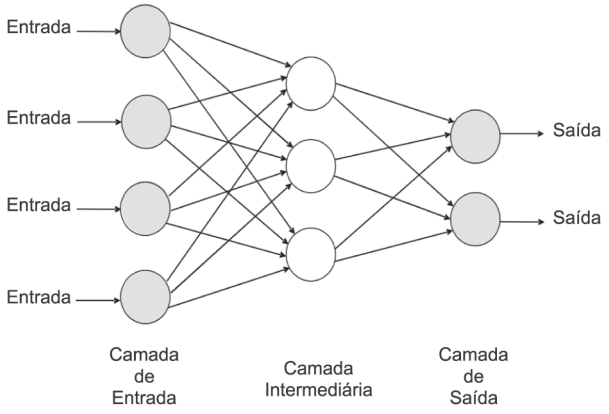

Figura 3. *Esquema de uma rede neural artificial do tipo Back Propagation, destacando-se a entrada, a saída e os neurônios das camadas de entrada, escondida (intermediária) e de saída, contendo todas as sinapses. Cada neurônio é simbolizado por um círculo*

As RN têm exercido atração aos químicos, pois em vários casos pode-se resolver problemas de interpretação de espectros e elucidação estrutural, uma vez que elas conseguem trabalhar com as complexas relações entre propriedades moleculares e dados espectrais. Para citar um exemplo, como entrada podem-se utilizar estruturas químicas e como saída, seus respectivos deslocamentos químicos ou vice-versa. Uma vez devidamente treinada, as RN são capazes de receber exemplos desconhecidos e realizar predições. Dentre as metodologias mais comuns em elucidação estrutural de substâncias utilizando-se RN estão os métodos supervisionados como CPG[46] ("CounterPropaGation"), BP ("BackPropagation", Figura 3) e ASNN[47] ("ASsociative Neural Networks"). Programas como o SPINUS-WEB, GENIUS e SpecSolv possuem RN em sua arquitetura.

Os *algoritmos genéticos* são também chamados de "computação evolucionária" e baseiam-se em uma analogia com os sistemas biológicos, tendo também várias aplicações em química[48]. Na abordagem de algoritmos genéticos, cada solução do problema é chamada de cromossomo. Os cromossomos consistem de genes e cada característica da solução, como por exemplo um grupo funcional de uma molécula, é chamada de gene. Nesta metodologia, o algoritmo realiza mutações e combinações para encontrar a melhor solução para o problema. A rotina que realiza tal trabalho é chamada de função de adequação/adaptação. O algoritmo segue os passos demonstrados na Figura 4. Primeiramente, é gerado um número aleatório de soluções possíveis que são passadas para a função de adaptação, por exemplo a predição de espectros. Tal função pontua cada solução de acordo com os resultados e as que obtiverem maior pontuação sobrevivem (neste exemplo, o espectro previsto mais próximo do experimental). As soluções sobreviventes sofrem mutação – como por exemplo oxidação, redução, mudança da ordem de uma ligação etc. – ou então são recombinadas ("crossover") e geram descendentes, ou seja, uma nova população. Todo o processo continua com os descendentes até as respectivas pontuações convergirem, isto é, até não ser mais possível melhorar a qualidade das soluções. Os AG estão implementados, por exemplo, no programa SENECA.

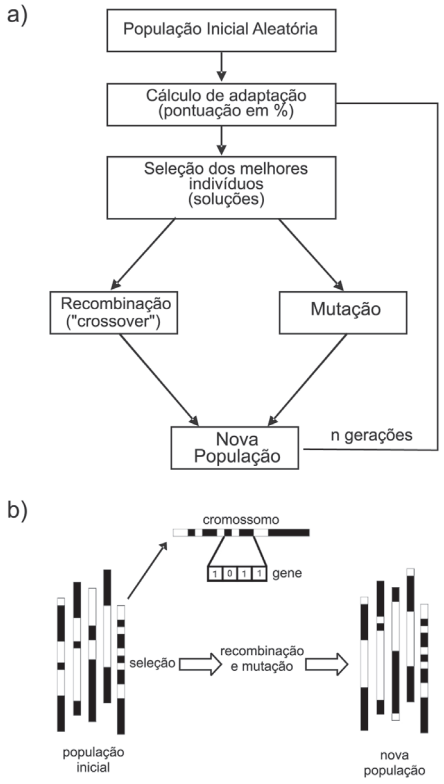

Figura 4. *Diagrama das etapas envolvidas no algoritmo genético (a); representação esquemática destacando a população inicial e a nova, um cromossomo e seu gene codificado com cadeia binária (b)*

O ACESSO DO USUÁRIO AOS PROGRAMAS DE ELUCIDAÇÃO ESTRUTURAL

Atualmente existem diversos programas disponíveis para auxiliar o usuário – químico ou espectroscopista – na elucidação estrutural de substâncias. Esses programas têm sua arquitetura baseada em diferentes metodologias e operam de acordo com as diferentes estratégias de elucidação estrutural descritas anteriormente. Um resumo contendo os principais programas disponíveis que são discutidos neste trabalho e suas principais características encontra-se na Tabela 1. Existem inúmeras formas de se ter acesso a tais programas, sejam eles aplicativos ou ferramentas: pode ser realizada a compra de sua licença de uma empresa; utilizar material de acesso livre, seja da academia ou de código aberto; pode-se ainda fazer uso de programas disponíveis em páginas da internet, como os serviços on-line, gratuitos ou não; finalmente, o acesso pode ser feito mediante solicitação ao(s) seu(s) criador(es) ou responsável(is), com envio posterior ao usuário.

Empresas de desenvolvimento de software para química

A "Advanced Chemistry Development" é bem conhecida pelos químicos por ser a firma que disponibiliza o ChemSketch, um programa livre muito utilizado para desenho e edição de estruturas químicas. Possui em sua linha de produtos comerciais (pagos) uma variedade enorme de programas e pacotes para diversas finalidades. Na linha de elucidação estrutural, destacam-se o ACD/HNMR Predictor 1D e 2D e de constantes de acoplamento, módulos para diferentes núcleos (^{13}C, ^{31}P, ^{15}N, ^{19}F), além de um outro programa para a predição de fragmentos de espectros de massas. Possui serviço on-line em sua página da internet, onde após registro, o usuário pode realizar gratuitamente tanto a avaliação de produtos pagos como também efetuar testes por um período de tempo determinado. Esta jovem empresa canadense de tecnologia é uma das que mais rapidamente se desenvolveu e inovou no setor, contando com vários doutores em sua equipe que pesquisam continuamente novos métodos e desenvolvem novas ferramentas computacionais.

A "CambridgeSoftware Corporation" é a empresa que comercializa o ChemOffice e o ChemDraw, recomendado por alguns periódicos de química como editor padrão de estruturas, o que tem causado desconforto por parte dos que são simpatizantes do software livre. Juntamente com o ChemOffice Pro a empresa comercializa o simulador ChemNMR embutido. O ChemNMR é um programa que utiliza regras empíricas para calcular os deslocamentos químicos de RMN. O ChemNMR possui um sistema de predição menos sofisticado que o da ACD/Labs, mas ainda confiável. Pesquisadores que desenvolvem novas metodologias ou novos programas para a predição de deslocamentos químicos de RMN geralmente comparam seus resultados com aqueles originados pelos programas da ACD/Labs e CambridgeSoftware.

Existem várias outras empresas, todas de menor porte, a maioria delas localizada na Europa, as quais se dedicam à pesquisa e ao desenvolvimento de software para a química e ferramentas de quimioinformática. Muitas delas empregam químicos, dando preferência a doutores da área de quimioinformática, sendo que algumas serão citadas a seguir.

O acesso livre ou público e os gratuitos ("freeware")

O acesso livre ou não a tais ferramentas gera basicamente as mesmas discussões que ocorrem quando se discute a comercialização de software e o monopólio da Microsoft – por exemplo com o Windows® e o Office® – e as empresas que produzem software de código livre para o sistema Linux – como a Conectiva+Mandrake (hoje Mandriva), Red Hat (Fedora), Suse Linux, etc. – e o OpenOffice. Esse tipo de discussão muitas vezes chega a ser tão fervorosa quanto discussões políticas, futebolísticas ou religiosas, quando cada lado defende cegamente seu ponto de vista. Isto é feito sem levar em conta que ambos os modelos de distribuição de software têm vantagens e desvantagens, tanto para desenvolvedores como para usuários. Estas discussões também levam à criação e perpetuação de muitos mitos sobre o software livre, dos quais dois valem a pena ser esclarecidos. O primeiro é o mito de que "software livre e de código aberto é grátis e nunca deve ser cobrado", ou seja, é comum confundir software livre com software grátis ou "freeware"[49]. Tal confusão é devida ao termo inglês[50], que levou muitos a confundir o sentido de "livre" – que no movimento do código aberto ("opensource") quer dizer que o usuário tem a liberdade para distribuir, modificar e adaptar o programa às suas necessidades e redistribuí-lo se quiser ou até mesmo criar um trabalho derivado totalmente novo – com o sentido de "livre", que muitas vezes em relações comercias quer dizer que um produto é dado como brinde ou vendido por um preço simbólico. Se fosse verdade que todo software livre e de código aberto deveria ser grátis, não haveria tantas empresas e pessoas tirando deste modelo de distribuição o sustento de suas vidas. Na prática, o que ocorre é que as empresas que vendem software livre cobram pelo serviço de empacotamento, gravação de mídia, distribuição, documentação e suporte, e não pelo software em si, fazendo com que o custo de aquisição dos produtos seja, em média, um décimo ou um centésimo do custo de aquisição de um software distribuído pelo modelo tradicional. O segundo mito é que "software livre é de domínio público", algo tão equivocado quanto o mito anterior, acreditando-se que o software livre é de domínio público e qualquer um pode fazer um trabalho derivado e vender como se fosse seu trabalho original. Pelo contrário, a grande maioria dos software livres possuem licença de distribuição e direitos autorais. As licenças de software livre visam proteger os direitos autorais dos desenvolvedores e garantir o direito dos usuários de compartilhar o software, sendo que cada licença protege os direitos e estabelece os deveres de ambas as partes de maneira diferente. Dentre as licenças mais comuns estão a GPL ("General Public License"), LGPL ("Lesser General Public License"), BSD ("Berkeley-Software Development License"), "Academic License", "Apache License", "Artistic License", SPL ("Sun Public License"), MPL ("Mozilla Public License") e mais recentemente a MSL ("Microsoft Shared License"), sendo que cada licença tem suas próprias características. No entanto, o que todas estas licenças têm em comum é a exigência de que o devido direito autoral seja mantido e respeitado em todos os trabalhos derivados. O desrespeito a essa diretriz, além ser anti-ético, pode acarretar ao infrator sanções jurídicas.

Um caso clássico de desrespeito aos direitos autorais de software livre em quimioinformática envolve o RasMol e o conhecido miniaplicativo para visualização de moléculas 3D MDL®/Chime, o qual gerou até uma publicação a respeito[51]. O RasMol é um visualizador de moléculas em 3D de código aberto produzido por Roger Sayle, do departamento de pesquisa e desenvolvimento da GlaxoWellcome e liberado sob licença GPL, a qual não permite o uso do código-fonte em projeto proprietários e de código-fonte fechado, caso do Chime. No passado, programadores da MDL apropriaram-se indevidamente do código do RasMol para desenvolver o Chime, sem darem o devido crédito a Sayle. Quando o fato foi descoberto, a GlaxoWellcome entrou com uma ação judicial contra a MDL e esta foi forçada a reconhecer publicamente que tinha utilizado indevidamente partes do RasMol, tendo de pagar uma indenização a Sayle e a liberar o Chime, anteriormente pago, gratuitamente na rede. A MDL também oferece gratuitamente o MDL/IsisDraw para desenho e edição de estruturas químicas.

Acesso on-line

Existem pesquisadores que implementaram poderosas plataformas de livre acesso e de código aberto. Tais ferramentas são bibliotecas para desenvolvimento de novos programas para quimio- e bioinformática, tais como OpenBabel, com licença GPL em C++; Chemistry Development Kit[52], com licença LGPL em Java; JOELib, com licença GPL, uma biblioteca para quimioinformática e cálculo de descritores em Java. De todas essas bibliotecas, as mais maduras e que se autocomplementam são a CDK e a JOELib. Dentre os sistemas de código livre e com acesso livre em rede para elucidação estrutural estão o SENECA[53], com licença "Artistic", e o NMRShiftDB[54], com licença GPL, ambos oriundos do mesmo grupo de pesquisa. Alguns sistemas são projetados para o acesso livre e on-line, e mesmo não sendo de código aberto, são ferramentas de grande auxílio para o químico. Dentre esses sistemas, pode-se citar o SPINUS-WEB para predição de deslocamentos químicos

e espectros de RMN ^{1}H, o TeleSpec para predição de espectros na região do IV, o SpecInfo para procura de dados de RMN e elucidação estrutural, dentre outros.

PROGRAMAS PARA AUXÍLIO NA ELUCIDAÇÃO ESTRUTURAL

Conforme foi discutido, existem disponíveis aos usuários vários pacotes, aplicativos, ferramentas e bancos de dados, comerciais ou não, os quais podem ser obtidos de diferentes fontes (Tabela 1). As metodologias implementadas em alguns destes programas e suas características principais, bem como as respectivas fontes, serão descritos a seguir.

Programas comerciais

ASSEMBLE

Desenvolvido pelo grupo de pesquisas de Munk, possui duas linhas para a abordagem da elucidação: geração de estruturas e redução de estruturas. Originalmente um puro gerador de estruturas, recentemente propagandeado como um módulo independente, está na versão 2.0[55]. Não realiza interpretação de espectros, baseando-se puramente nas informações fornecidas pelo usuário, que deve realizar a sua interpretação. Ele também gera subestruturas. As informações fornecidas são restrições e devem envolver, por ex., contagem do número máximo e mínimo de ligações duplas e triplas, número de átomos de carbono nas moléculas, número esperado de anéis, contagem de átomos de hidrogênio, tipo de hibridação para metais pesados etc. Com base nestas informações, o programa gera e fornece listas de subestruturas ao usuário, que deve observá-las e reanqueá-las de acordo com a concordância dos dados espectrométricos experimentais anteriormente obtidos para a estrutura desconhecida. O usuário é quem realiza a interpretação dos espectros, sendo que o programa apenas lista estruturas com base nos fragmentos. Uma versão de demonstração que pode trabalhar com estruturas de até 15 átomos que não sejam de hidrogênio pode ser baixada gratuitamente na página da empresa suíça Upstream Solutions.

Tabela 1. Exemplos de software utilizados para elucidação estrutural auxiliada por computador e suas principais características

Nome	Disponibilidade	Licença	Estratégia	Metodologia	Facilidade de uso	URL
SPINUS-WEB	on-line, livre	N/D	predição	RN	fácil (WEB)	http://www.dq.fct.unl.pt/spinus/
ASSEMBLE	comercial, demo disponível	proprietária	geração	heurística	fácil (GUI[1])	http://www.upstream.ch/products/assemble.html
CSEARCH	on-line (apenas predição)	N/D	busca/ predição	RN, HOSE, banco de dados	fácil (WEB, e-mail)	http://homepage.univie.ac.at/wolfgang.robien/csearch_server_info.html
DENDRAL	academia	-	busca/geração	banco de dados, heurística	-	-
CONGEN	código aberto, livre	domínio público	geração	heurística	difícil (linhas de comando; conhecimento de programação necessário)	http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/reasonng/chem/congen/
GENOA	academia	-	geração	heurística	-	-
HOUDINI	academia	-	geração	heurística	-	-
SESAMI	academia, sob requisição	-	geração/ predição	heurística	fácil (GUI)	http://chemistry.asu.edu/faculty/M_munk.asp
DARC/EPIOS	academia	-	geração/predição	heurística	-	-
SPECINFO	on-line, comercial	proprietária	busca/predição	banco de dados, HOSE	fácil (WEB)	http://specinfo.wiley.com/specsurf/welcome.html
SENECA	código aberto, livre	*Artistic*	busca/geração	GA, heurística	médio (o programa vem apenas como código-fonte e deve ser compilado; possui GUI)	http://almost.cubic.uni-koeln.de/cdk/jrg/software/seneca
NMRShiftDB	código aberto, livre	GPL	busca, predição	banco de dados/ HOSE	fácil (WEB)	http://www.nmrshiftdb.org
ACD/HNMR/ CNMR	comercial	proprietária	predição	banco de dados/ HOSE	fácil (GUI)	http://www.acdlabs.com/products/spec_lab/predict_nmr/
StrucEluc	comercial	proprietária	busca/geração/ predição	banco de dados/ HOSE/heurística	médio (GUI complicada)	http://www.acdlabs.com/products/spec_lab/complex_tasks/str_elucidator/
MOLGEN	comercial, demo	proprietária	geração	heurística	médio (com GUI, mas mal documentado)	http://www.mathe2.uni-bayreuth.de/molgen4/
GENIUS	academia, sob requisição	N/D	geração/ predição	GA, RN	fácil (GUI)	http://www.jens-meiler.de/index_soft.html
SISTEMAT	academia, sob requisição	N/D	busca	banco de dados	médio (linha de comando, modo texto)	-
NMRBenefit	comercial	proprietária	predição	HOSE, RN	fácil (GUI)	http://www.modgraph.co.uk/product_nmr_benefit.htm
LSD	academia, código aberto	GPL	geração	heurística	médio (linha de comando)	http://www.univ-reims.fr/Labos/UMR6013

[1]"Graphical User Interface" (interface gráfica para o usuário)

Advanced Chemistry Development

Conforme mencionado, os programas para elucidação estrutural desenvolvidos por esta empresa são pagos. No entanto, é possível cadastrar-se na página da internet para obter uma autorização válida por 15 dias para teste ilimitado de alguns aplicativos e bancos de dados que a empresa oferece, bastando o pesquisador interessado se cadastrar em http://ilab.acdlabs.com.

ACD/HNMR Predictor 1D e 2D

A metodologia empregada neste programa é de bancos de dados relacionais. De acordo com a empresa, existem armazenados dados de RMN ^{1}H de mais de 175.000 estruturas diferentes, com cerca de 1.440.000 deslocamentos químicos atribuídos. O método funciona com base em uma tabela de correlação de fragmentos de estruturas com seus respectivos deslocamentos químicos, o que torna possível um desempenho melhor que dos sistemas baseados em regras. Este sistema possui alto desempenho durante a predição de deslocamentos químicos e foi usado como parâmetro de comparação para várias metodologias[56,57]. O algoritmo é inteligente e heurístico, pois pode achar os fragmentos da molécula que estão presentes em seus bancos de dados e calcular as interações spin-spin presentes na molécula para ajustar os valores de deslocamento químico. Ainda reconhece diferenças no espectro dos seguintes tipos de estruturas isoméricas: isômeros *cis-trans* e isômeros cíclicos endo-exo.

ACD/StrucEluc (Structure Elucidator)

Este programa, que está na versão 8.0[58], foi desenvolvido para a elucidação estrutural automatizada de estruturas químicas. O programa possui uma moderna interface com o usuário e é necessária pouca interferência do químico. Os algoritmos do sistema baseiam-se na fórmula molecular (um espectro de massas é necessário) e correlações de espectros 1D/2D para realizar o processo de elucidação estrutural. Se dados de NOESY ("Nuclear Overhauser Effect SpectroscopY") estiverem disponíveis, o sistema também é capaz de determinar a estereoquímica relativa automaticamente[59]. Ele ainda utiliza as correlações entre os espectros 2D para criar uma lista de fragmentos compatíveis com as correlações apresentadas. Tais fragmentos são enviados ao gerador para combinação e geração das soluções, que são validadas pelo ACD/NMR Predictor.

Além da predição de espectros de ^{1}H e de ^{13}C, a empresa ainda possui programas para predição de espectros para outros núcleos, como ^{19}F, ^{15}N e ^{31}P, além de fragmentos de espectros de massas (ACD/MS Fragmenter).

SPECINFO

Este programa foi originalmente desenvolvido pela BASF. Em seguida, sua licença passou para a Chemical Concepts GmbH, da Alemanha, em 1998, e desde 2004 é comercializado pela Wiley Interscience, que fornece acesso on-line e a atualização para o Specinfo XS Client, utilizado para acesso ao servidor SPECINFO, versão 4.0. O SPECINFO, uma das maiores coleções do mundo com mais de 420.000 espectros, é um sistema de gerenciamento de banco de dados projetado para armazenar, buscar e manipular espectros de IV, RMN (^{1}H, ^{13}C, ^{31}P e ^{15}N) e EM de substâncias orgânicas. Possui ainda plataforma integrada para visualização, predição e busca de espectros. O programa foi projetado para ser uma ferramenta auxiliar no processo de elucidação estrutural, utilizando um algoritmo de busca de espectro/subespectro no banco para encontrar uma estrutura/subestrutura compatível com os dados do espectro-problema. Possui também um módulo de predição de espectros. Pode ser utilizado por químicos que desejam ter uma idéia para onde direcionar a elucidação estrutural ou na confirmação de uma proposta para uma determinada estrutura. Uma versão on-line da ferramenta SpecSurf XS para a busca de espectros e predição de deslocamentos químicos, juntamente com um guia ilustrado para o usuário, pode ser acessada gratuitamente em http://cds.dl.ac.uk/cds/datasets/spec/specinfo/specinfo.html (apenas para membros de universidades britânicas) mediante registro na página da CDS ("Chemical Database Service", http://cds.dl.ac.uk/), da Inglaterra.

Sistemas de código aberto e livres

SENECA

Trata-se de um pacote de programas para elucidação estrutural auxiliada por computador. Esta ferramenta utiliza o método estocástico e AG para geração de estruturas[60], sendo capaz de buscar espaços constitucionais de moléculas que sejam mais amplos que os algoritmos determinísticos. Utilizando este procedimento, o programa tenta encontrar a constituição de uma molécula desconhecida a partir de evidências de seus dados espectrométricos experimentais. No processo de elucidação estrutural são utilizados basicamente dados de RMN, porém a fórmula molecular obtida por EM é recomendável, sendo que quaisquer dados espectrométricos são aceitáveis. Porém, antes de iniciar o processo, o usuário deve providenciar os dados de entrada, os quais são retirados dos espectros obtidos experimentalmente. Uma característica importante do SENECA é a utilização de dados de RMN monodimensionais de ^{13}C como DEPT ("Distortionless Enhancement by Polarization Transfer") 90 e 135°. Entretanto, dados bidimensionais também podem ser submetidos, como por exemplo ^{1}H-^{1}H COSY ("COrrelated SpectroscopY"), dados de correlação $^{1}J_{CH}$ a curta distância como HMQC ("Heteronuclear Multiple Quantum Coherence") e HSQC ("Heteronuclear Single Quantum Coherence") e, ainda, dados de correlação a longa distância, como HMBC ("Heteronuclear Multiple Bond Correlation"), tanto para C-H como N-H. O programa tem inclusive a capacidade de importar arquivos do programa Win-NMR da Bruker (editor/processador de espectros de RMN), além de trabalhar com o formato XML ("eXtensible Markup Language"). O pacote, disponível apenas em inglês, foi escrito na linguagem Java e roda nas interfaces Cocoa (MacOS X), Gnome, KDE e Win32 (MS Windows). Possui licença "Artistic", podendo ser baixado na própria página que o descreve, a partir do atalho http://www.sf.net/projects/seneca. O sistema é distribuído e caso o acesso à internet esteja disponível, é capaz de distribuir a tarefa para outros computadores executando o SENECA, o que pode tornar o processo mais rápido. Embora os autores não afirmem, o SENECA pode ser uma ótima alternativa livre ao ACD/StrucEluc, tendo inclusive chegado aos mesmos resultados deste último software na elucidação do triterpeno policapol[61,62].

Sistemas on-line livres com código fechado

SPINUS-WEB

É uma ferramenta on-line destinada à verificação e validação de estruturas orgânicas através da predição de deslocamentos químicos de RMN ^{1}H[63,64]. Roda na plataforma Java e foi desenvolvido por J. A. de Sousa, pesquisador da Universidade Nova de Lisboa, Portugal, estando disponível na página desta universidade (http://www.dq.fct.unl.pt/spinus) e com um espelho ("mirror") na Universidade Erlangen-Nuremberg (http://www2.chemie.uni-erlangen.de/services/spinus), na Alemanha. Possui em sua interface o editor de estruturas Marvin Applet a fim de que o usuário possa desenhar a estrutura desejada (2D) antes do cálculo de seus deslocamentos químicos. Entretanto, o aplicativo ainda suporta a importação de estruturas em outros formatos individuais como o

SMILES e MDL/MOL (com tabelas de conectividade[65]), além de grupos de estruturas em formato MDL/SD. Para rodar o programa, além da instalação do software gratuito Java (Sun Microsystems, http://java.sun.com/), o usuário ainda necessita instalar o "plug-in" Chime para visualizar estruturas 3D. O MDL/Chime está disponível gratuitamente para "download" na página da MDL.

A entrada de cada estrutura (em 2D) é feita após sua inserção ou desenho no editor da tela principal. Porém, antes da predição de seus deslocamentos químicos, o SPINUS-WEB gera automaticamente as coordenadas em 3D através do software CORINA, além de calcular várias propriedades físico-químicas para cada tipo de hidrogênio presente na estrutura. O CORINA[66], um software comercializado pela empresa alemã Molecular Networks GmbH (http://www.mol-net.de/) e também disponível gratuitamente para uso on-line na página da Universidade Erlangen-Nuremberg (http://www2.chemie.uni-erlangen.de/software/corina/index.html), é um gerador de estruturas em 3D que foi incorporado ao SPINUS-WEB. A metodologia empregada para a predição dos deslocamentos químicos que foi incorporada ao SPINUS-WEB baseia-se em conjuntos de várias RN artificiais com o algoritmo do tipo BP (ou FFNN). Além da predição de deslocamentos químicos, o programa fornece a simulação do espectro de RMN ^{1}H da substância em questão. Todo este processo é muito rápido, sendo que a predição para uma molécula relativamente pequena (< 600 Da) realizada em computador com processador Pentium IV ou equivalente conectado a uma rede DSL dura menos de 15 s. Existe também uma versão paga do SPINUS, com itens adicionais, comercializada pela empresa Molecular Networks GmbH.

CSEARCH (http://homepage.univie.ac.at/wolfgang.robien/csearch_main.html). Trata-se de um banco de dados com mais de 230.000 espectros de RMN ^{13}C e mais de 2.700.000 deslocamentos químicos de ^{13}C atribuídos. É um programa baseado em código HOSE para predição de espectros que recentemente incorporou RN artificiais. Existe uma versão on-line para a predição de espectros de RMN ^{13}C onde os dados de entrada são estruturas 2D no formato MDL/MOL, podendo ou não conter informações sobre a estereoquímica. Após o registro do usuário no servidor, estas informações devem ser submetidas através de e-mail, sendo que o espectro previsto é posteriormente enviado ao usuário. Em 2005 o CSEARCH foi incorporado a um programa denominado NMR Predict que se encontra na versão 2.0 e é atualmente comercializado pela empresa britânica Modgraph Consultants LTD (http://www.modgraph.co.uk/). Esta versão realiza predições de RMN tanto para ^{13}C (com a incorporação do CSEARCH) como para ^{1}H (com a incorporação de um programa denominado CHARGE Proton NMR Prediction), além de oferecer algumas melhorias ao usuário.

Sistemas on-line livres com código aberto

NMRShiftDB

Trata-se de um banco de dados de código aberto ("opensource") de moléculas orgânicas e seus dados de RMN. Atualmente, seu banco de dados possui cerca de 19.000 espectros, número que aumenta a cada dia, pois qualquer pesquisador pode se cadastrar na página da internet e enviar os dados de RMN de qualquer substância química. Para garantir a integridade e veracidade dos dados enviados, estes são revistos e confirmados por dois pesquisadores voluntários. Se tais pesquisadores encontrarem algum erro, aquele que enviou os dados é contactado por e-mail e deve corrigi-los em 48h. Caso a correção não seja feita, os dados são eliminados do banco. No caso de dados originais, eles são eliminados caso não sejam publicados em periódico dentro de um prazo de 120 dias. Esse processo lembra o processo de revisão por pares ("peer review"), comum para avaliar manuscritos submetidos a periódicos. O sistema também faz predição de dados de RMN, baseando-se em códigos HOSE de até seis esferas. Para utilizar o sistema, o usuário entra com os dados de RMN e o sistema retorna uma lista de estruturas contendo uma porcentagem de similaridade entre o espectro da substância do banco e o espectro informado pelo usuário ou então, pode-se entrar com uma estrutura e obter-se o espectro previsto. Esse sistema mostra-se como uma alternativa livre, ainda que com um banco de dados menor que o do SpecInfo. A utilização deste sistema para a elucidação estrutural de um cromeno foi descrita na literatura[67].

CONCLUSÕES E PERSPECTIVAS

Os diferentes programas para elucidação estrutural auxiliada por computador estão provocando uma revolução no setor, seja na academia ou em empresas que desenvolvem software para a química. Uma tendência é integrar tais programas de computador em pesquisas para estudos de bioprospecção de vegetais e microrganismos ou de produtos de reações orgânicas em larga escala, com o intuito de se aumentar a produtividade. Como exemplo, recentemente surgiu o termo "High Throughput Structure Elucidation" (HiTSE), que consiste em minimizar ao máximo o tempo necessário para a elucidação estrutural de uma determinada substância química. O princípio baseia-se na comparação de dados espectrométricos, físico-químicos e cromatográficos com os dados presentes em uma biblioteca de substâncias puras[68]. A procura em princípio é baseada em tempos de retenção relativos e pesos moleculares obtidos por EM, sendo que após esta fase são selecionadas substâncias com dados cromatográficos muito semelhantes entre si, para em seguida poder compará-los aos dados de espectros de RMN. Desta forma, as estruturas são identificadas com maior acuidade e precisão, além de maior rapidez. É válido lembrar que todo este processo é realizado de forma automatizada e para uma grande quantidade de substâncias. Caso alguma substância não possa ser identificada através deste procedimento, como por exemplo aquelas novas na literatura ou ausentes no banco de dados, é realizada a elucidação estrutural parcial, com base em subestruturas. A elucidação da estrutura é concluída com o auxílio de técnicas de RMN bidimensionais (HSQC, HMBC, ^{1}H-^{1}H COSY etc.). O HiTSE ainda pode ser utilizado ou adaptado para uso com outras técnicas, tais como HPLC-EM[69].

Constata-se que os programas de computador já estão completamente consolidados na área de elucidação estrutural de substâncias orgânicas. Vários indicadores atestam tal afirmação: o aumento significativo de publicações em periódicos especializados a cada ano; o crescimento e investimento em pesquisa, desenvolvimento e na distribuição por empresas da área de química que desenvolvem e comercializam software, como a ACD/Labs, CambridgeSoft, de outras menores e da academia; surgimento dos programas de código aberto e as ferramentas on-line na internet, havendo ainda espaço para o software livre e os que funcionam baseados em "collaborative development". Cada um destes produtos tem vantagens e desvantagens, tais como suporte, preço, validação, plataformas, metodologias, acuidade, além da capacidade da interface ser amigável ou não ao usuário. A escolha por um produto ou outro depende do problema em questão, sendo que a precisão varia para diferentes classes de substâncias. Uma das limitações a ser vencida é o efeito do solvente nos deslocamentos químicos, em especial em RMN ^{1}H, pois os dados utilizados são obtidos apenas em deuteroclorofórmio. Torna-se óbvio que ainda não existe um produto que seja capaz de operar muito bem em todas as situações e o

conselho é que o usuário teste diferentes programas com diferentes moléculas, ainda mais se pretende comprar algum, pois o preço não costuma ser baixo. Embora alguns almejem, dificilmente o software irá substituir a inteligência do especialista no processo de elucidação estrutural de substâncias. A elucidação estrutural auxiliada por computador ainda não é um processo totalmente automatizado e exige a presença do especialista, tanto no seu desenvolvimento quanto em seu uso.

MATERIAL SUPLEMENTAR

Nesta seção encontra-se breve descrição do material suplementar (figuras) que está disponível gratuitamente em http://quimicanova.sbq.org.br, na forma de arquivo PDF.

As Figuras 1Sa e 1Sb são referentes a páginas da web do programa *SPINUS-WEB*, discutido na seção sistemas on-line livres com código fechado. A Figura 1Sa mostra a tela inicial com uma estrutura 2D usada como dado de entrada para as predições; a Figura 1Sb mostra a mesma estrutura, 2D e também 3D, com os deslocamentos químicos previstos de seus hidrogênios (tabela à direita) e a simulação do respectivo espectro de RMN ^{1}H (abaixo). A página de entrada do sistema *CSEARCH* para predição de espectros de RMN ^{13}C é mostrada na Figura 2S. A Figura 3S mostra a página da web do banco de dados *NMRShiftDB*, discutido na seção sistemas on-line livres com código aberto; pode-se observar o processo de auxílio da elucidação estrutural de um produto natural através de seus dados de RMN ^{13}C que foram utilizados como dados de entrada.

REFERÊNCIAS

1. Ciência que trata da organização, busca e extração de informação de forma automatizada através do uso de um computador; não é sinônimo de computação.
2. Morgan, H. L.; *J. Chem. Doc*. **1965***, 5*, 107.
3. Lindsay, R.; Buchanan, B. G.; Feigenbaum, E. A.; Ledberg, J.; *Applications of Artificial Intelligence in Organic Chemistry: The Dendral Project*, McGraw-Hill: Nova York, 1980.
4. É a descrição, de forma lógica, de um conjunto finito de passos a serem executados no cumprimento de determinada tarefa; uma receita para um pocesso computacional.
5. Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C.; *Org. Mag. Reson.* **1981**, *15*, 375.
6. Masinter, L. M.; Sridharan, N. S.; Ledeberg, J.; Smith, D. H.; *J. Am. Chem. Soc.* **1974**, *96*, 7702.
7. Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C.; *J. Org. Chem.* **1981**, *46*, 1708.
8. Dubois, J.-E.; Sobel, Y.; *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 326.
9. Dubois, J.-E.; Carabedian, M.; Dagane, I.; *Anal. Chim. Acta* **1984**, *158*, 217.
10. Shelley, C. A.; Munk, M. E.; *Anal. Chem.* **1982**, *54*, 516.
11. Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V.; *Anal. Chim. Acta* **1978**, *103*, 121.
12. Madison, M. S.; Schulz, K. P.; Korytko, A. A.; Munk, M. E.; *Internet J. Chem*. **1998**, *1*, 34.
13. Korytko, A.; Schulz, K. P.; Madison, M. S.; Munk, M. E.; *J. Chem. Inf. Comput. Sci.* **2003**, *32*, 1434.
14. Kalchhauser, H.; Robien, W.; *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103.
15. Bremser, W.; *Anal. Chim. Acta* **1978**, *103*, 355.
16. Jargão de informática; significa que o sistema é capaz de inferir ou calcular novos dados durante a execução do programa a partir de dados pré-existentes em um banco.
17. Kudo, Y.; Sasaki, S.; *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 43.
18. Bremser, W.; Fachinger, W.*; Magn. Reson. Chem.* **1985**, *23*, 1056.
19. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Borges, J. H. G.; Emerenciano, V. P.; *Quim. Nova* **1990**, *13*, 10.
20. Gastmans, J. P.; Furlan, M.; Lopes, M. N.; Borges, J. H. G.; Emerenciano, V. P.; Q*uim. Nova* **1990**, *13*, 75.
21. Alvarenga, S. A. V.; Rodrigues, G. V.; Gastmans, J. P.; Emerenciano, V. P.; *Nat. Prod. Lett.* **1995**, *7*, 133.
22. Alvarenga, S. A. V.; Gastmans, J. P.; Rodrigues, G. V.; Brandt, A. J. C.; Emerenciano, V. P.; *J. Braz. Chem. Soc.* **2003**, *14*, 369.
23. Ferreira, M. J. P.; Brandt, A. J. C.; Rodrigues, G. V.; Emerenciano, V. P.; *Anal. Chim. Acta* **2001**, *429*, 151.
24. Fromanteau, D. L. G.; Gastmans, J. P.; Vestri, S. A.; Emerenciano, V. P.; Borges, J. H. G.; *Comput. Chem.* **1993**, *17*, 369.
25. Rufino, A. R.; Brandt, A. J. C.; Santos, J. B. O.; Ferreira, M. J. P.; Emerenciano, V. P.; *J. Chem. Inf. Model.* **2005**, *45*, 645.
26. Weiniger, D.; *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
27. Nomenclatura química para a representação de estruturas, mais especificamente um modelo de valência altamente simplificado e compactado; descreve uma estrutura química como uma notação de linha.
28. Ctfile Formats. MDL Information Systems, http://www.mdli.com, San Leandro, 2002.
29. Masinter, L. M.; Shriodharan, N. S.; Lederberg, J.; Smith, D. H.; *J. Am. Chem. Soc*. **1974**, *96*, 7702.
30. Bradley, D. C.; Munk, M. E.; *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87.
31. Fontana, P.; Pretsch, E.; *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 614.
32. Kerber, A.; Laue, R.; Grüner, T.; Meringer, M.; *Match* **1998**, *37*, 205.
33. Faulon, J. L.; *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1204.
34. Meiler, J.; Will, M.; *J. Am. Chem. Soc.* **2002**, *124*, 1868.
35. Will, M.; Fachinger, W.; Richert, J. R.; *J. Chem. Inf. Comput. Sci.* 1996, *36*, 221,
36. Pople, J. A.; Nesbet, R. K.; *J. Chem. Phys.* **1954**, *22*, 571.
37. Chesnut, D. B.; Phung, C. G.; *Chem. Phys.* **1990**, *147*, 91.
38. Chesnut, D. B.; *Ann. Rep. NMR Spectrosc.* **1994**, *29*, 71.
39. Fileti, E. E.; Canuto, S.; *Int. J. Quantum Chem.* **2005**, 102, 554.
40. Wolinski, K.; Hilton, J. F.; Pulay, P.; *J. Am. Chem. Soc.* **1990**, *112*, 8251.
41. Keith, T. A.; Bader, R. F. W.; *Chem. Phys. Lett.* **1992**, *194*, 1.
42. Schindler, M.; Kutzelnigg, W.; *Mol. Phys.* **1983**, *48*, 781.
43. Hansen, A. E.; Bouman, T. D.; *J . Chem. Phys.* **1989**, *91*, 3552.
44. Zupan, J.; Gasteiger, J.; *Neural Networks in Chemistry and Drug Design*, Winheim, Wiley-VCH, 2nd ed., 1999.
45. Munk, M. E.; Madison, M. S.; *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231.
46. Gasteiger, J.; *Chem. Intell. Lab. Syst*., no prelo.
47. Tetko, I. V.; Vsevold, Y. T.; *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136.
48. Filho, P. A. C.; Poppi, R. J.; *Quim. Nova* **1999**, *22*, 405.
49. Termo em inglês para designar um *software* disponível gratuitamente.
50. Em inglês: *free software*; a palavra *free* em inglês tem duplo sentido e depende muito do contexto; pode significar tanto "livre" no sentido de liberdade ou "livre" no sentido de obter-se algo de forma gratuita, ou seja, grátis; por isso, o termo vem sendo rapidamente substituido por *opensource*.
51. Hodgson, J.; *Nat. Biotechnol.* **1996**, *14*, 690.
52. Steinbeck, C.; Youngquan, H.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighangem, E.; *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493.
53. Steinbeck, C.; *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1500.
54. Steinbeck, C.; Kuhn, S.; Krause, S.; *J. Chem. Inf. Comput Sci.* **2003**, *45*, 1733.
55. Baderstcher, M.; Korytko, A.; Schulz, K. P.; Madison, M.; Munk, M. E.; Portmann, P.; Junghans, M.; Fontana, P.; Pretsch, E.; *Chem. Intell. Lab. Syst.* **2000**, *51*, 73.
56. Magri, F. M. M.; Militão, J. S. L.; Ferreira, M. J. P.; Brandt, A. J. C.; Emerenciano, V. P.; *Spectroscopy* **2001**, *15*, 99.
57. Meiler, J.; Maier, W.; Will, M.; Meusinger, R.; *J. Magn. Reson*. **2002**, *157*, 242.
58. Elyashberg, M. E.; Blinov, K. A.; Willians, A. J.; Molodtsov, S. G.; Martin, G. E.; Martirosian, E. R.; *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 771.
59. Sumurnyy, Y. D.; Elyashberg, M. E.; Blinov, K. A.; Lefbrvre, B. A.; Martin, G. E.; Williams, A. J.; *Tetrahedron* **2005**, *61*, 9980.
60. Han, Y.; Steinbeck, C.; *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 489.
61. Elyashber, M. E.; Blinov, K. A.; Williams, A. J.; Martinrosian, E. R.; Molodtsov, S. G.; *J. Nat. Prod.* **2002**, *65*, 693.
62. Steinbeck, C.; *Nat. Prod. Rep.* **2004**, *21*, 512.
63. Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J.; *Anal. Chem.* **2002**, *74*, 80.
64. Binev, Y.; Aires-de-Sousa, J.; *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 940.
65. Do inglês *Connection Table* (CT): dentre várias outras, é a forma predominante de representação de estruturas químicas em programas de computador, baseada em uma matriz com uma lista de átomos e outra de ligações químicas que fornece as conexões entre os átomos.
66. Sadowski, J.; Gasteiger, J.; *Chem. Rev*. **1993**, *93*, 2567.
67. Steinbeck, C.; Kuhn, S.; *Phytochemistry* **2004**, *65*, 2711.
68. Bindseil, K. U.; Jakupovic, J.; Wolf, D.; Lavayre, J.; Leboul, J.; Pyl, D.; *Drug Disc. Today* **2001**, *16,* 840.
69. Wolf, C.; Villalobos, C. N.; Cummings, P. G.; Kennedy-Gabbs, S.; Olsen, M. A.; Trescher, G.; *J. Am. Soc. Mass. Spectrom*. **2005**, *16*, 553.