Iêdo Luiz Martinovsky

RECONHECIMENTO DE PADRÕES EM MOLÉCULAS ORGÂNICAS DE ORIGEM VEGETAL EM UM BANCO DE ESTRUTURAS TRIDIMENSIONAIS

Florianópolis – SC
2000

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

IÊDO LUIZ MARTINOVSKY

RECONHECIMENTO DE PADRÕES EM MOLÉCULAS ORGÂNICAS DE ORIGEM VEGETAL EM UM BANCO DE ESTRUTURAS TRIDIMENSIONAIS

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Orientador:
PROF. DR. JÚLIO SANCHO LINHARES TEIXEIRA MILITÃO

Florianópolis, outubro de 2000

RECONHECIMENTO DE PADRÕES EM MOLÉCULAS ORGÂNICAS DE ORIGEM VEGETAL EM UM BANCO DE ESTRUTURAS TRIDIMENSIONAIS

Iêdo Luiz Martinovsky

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, Área de Concentração Sistemas de Conhecimento, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Júlio Sancho Linhares Teixeira Militão, Dr.

_____

Fernando Ostuni Gauthier, Dr.

Banca Examinadora:

_____

Júlio Sancho Linhares Teixeira Militão, Dr.

_____

Prof. Rogério Cid Bastos, Dr.

_____

Luiz Fernando Jacintho Maia, Dr.

_____

Anita Maria da Rocha Fernandes, Dra.

DEDICATÓRIA

À Elena, eterna namorada e companheira, pelo estímulo, carinho, paciência e colaboração, sem o qual este trabalho não teria sido possível.

AGRADECIMENTOS

Muitas são as pessoas com as quais estou em débito pelo que auxiliaram durante o curso e a elaboração desta dissertação. Em especial, desejo agradecer ao Prof. Júlio Sancho Linhares Teixeira Militão que, mais do que um orientador, foi um inspirador, um verdadeiro amigo, quase um irmão. Meu reconhecimento,também, à equipe do SISTEMAT, do Instituto de Química d Universidade de São Paulo, sobretudo ao Prof. Vicente Emerenciano, por gentilmente me franquearem o acesso aos seus bancos de dados e códigos fontes dos programas. Ao Prof. Rogério Cid Bastos, coordenador do curso de mestrado, que com sua seriedade e cordialidade serviu de exemplo a estímulo, meus sinceros agradecimentos. Aos professores, colegas e a todos que de uma forma ou outra contribuíram para que este momento se tornasse realidade, minha gratidão mais sincera.

Tabelas

Tabela 2.1: Algumas definições de IA - Fonte: Wooldrige, 1995, p. 5 23
Tabela 2.2: Exemplo de conjunto de regras para sistema de produção 27
Tabela 4.1: Códigos de agrupamentos atômicos 57
Tabela 4.2: Matriz de conectividade da substância da fig. 4.1 58
Tabela 4.3: Matriz de conectividade da subestrutura da fig 4.2 59
Tabela 4.4: Número de ligantes diferentes de hidrogênio, por tipo de átomo 60
Tabela 4.5: Correspondência entre os átomos da subestrutura da molécula 62
Tabela 4.6: Correspondência final entre a subestrutura e a molécula 63

Figuras

Fig. 1.1: Espectro de RMN ^{13}C do ácido betulínico (Militão, 1996) 2

Fig. 1.2: Exemplo de arquivo .HIN 8

Fig. 1.3: Sintaxe de arquivo tipo .HIN 8

Fig. 1.4: Tela do Sistemat (sub-programa DATASIS) 9

Fig. 1.5: Matriz topológica representativa de estrutura 16

Fig. 1.6 Estrutura dos arquivos de armazenamento de moléculas e subestruturas 18

Fig. 1.7: Diagrama de abordagem de decisão teórica (Fu, 1976) 20

Fig. 2.1: Representação das funções h1, h2 e h3 (alturas próximas de 2 metros) 38

Fig. 2.2: Conjunto nebuloso “jovem” 39

Fig. 2.3: Conjunto nebuloso “meia idade” 40

Fig. 2.4: Conjunto nebuloso “velho” 40

Fig. 2.5: Conjunto nebuloso complementar de “jovem” 41

Fig. 2.6: Conjunto nebuloso complementar de “meia idade” 41

Fig. 2.7: Conjunto nebuloso complementar de “velho” 42

Fig. 2.8: Intersecção dos conjuntos nebulosos “jovem” e “meia idade” 42

Fig. 2.9: Intersecção dos conjuntos nebulosos “jovem” e “velho” 43

Fig. 2.10: Intersecção dos conjuntos nebulosos “meia idade” e “velho” 43

Fig. 2.11: Reunião dos conjuntos nebulosos “jovem” e “meia idade” 44

Fig. 2.12: Reunião dos conjuntos nebulosos “jovem” e “velho” 44

Fig. 2.13: Reunião dos conjuntos nebulosos “meia idade” e “velho” 45

Fig. 2.14: Representação gráfica de “livro caro” 46

Fig. 2.15: Estrutura aninhada de cortes 47

Fig. 3.1: Fórmula estrutural do ácido acético 51

Fig. 3.2: Modelo de Le Bel e Van´t Hoff (tetraedro) 52

Fig. 3.3: 2-cloro-propano no modelo bow-tie 52

Fig. 3.4 n-Butano representado no modelo de Stuart 53

Fig. 3.5: Modelo de pau e bola representando o diclorometano 53

Fig. 3.6: O diclorometano e a projeção de seus átomos no plano α 54

Fig. 3.7: Isomeria plana 55

Fig. 3.8: Isomeria espacial 55

Fig. 4.1: Molécula de exemplo 57

Fig. 4.2: Subestrutura 59

Fig. 4.3 Primeira fase da busca 61

Fig. 4.4: Continuação da busca 62

Fig. 4.5: Subestrutura linear em forma de “grade” 64

Fig. 4.6: Exemplos de subestruturas 66

Fig. 5.1: Figuras geométricas com o mesmo número de vértices 68

SUMÁRIO

Capítulo I – INTRODUÇÃO 1
1.1. Histórico 1
1.2. Determinação estrutural de um composto 3
1.3. Sistemas informatizados 4
1.3.1. DENDRAL 4
1.3.2. ACCESS 4
1.3.3. DARC/EPIOS 4
1.3.4. CASE 5
1.3.5. Outros Sistemas 5
1.3.6. Sistema especialista SISTEMAT 6
1.4. Objetivo e justificativa 10
1.5. Método de trabalho 13
1.6. Linguagem utilizada 15
1.7. Bancos de dados 15
1.8. Arquivos xBase 16

Capítulo II - FUNDAMENTAÇÃO TEÓRICA 19
2.1. Reconhecimento de padrões (RP) 19
2.1.1. Conceito 19
2.1.2. Técnicas de RP 20
2.2. Inteligência artificial (IA) 22
2.2.1. O que é inteligência artificial? 22
2.2.2. Paradigmas da inteligência artificial 24
2.2.3. Regras de produção 25
2.2.4. Busca em amplitude 27
2.2.5. Busca em Profundidade 31
2.2.6. Subida de Encosta 33
2.3. A incerteza 35
2.3.1. Conjuntos nebulosos 36
2.3.2. Operações com conjuntos nebulosos 38

2.3.3. Cortes α de conjuntos nebulosos 45
2.3.4. Lógica nebulosa 48

Capítulo III - A TRIDIMENSIONALIDADE DA MOLÉCULA 51
3.1. Representação da molécula 51
3.2. Isomeria 54

Capítulo IV - LOCALIZAÇÃO PLANA 56
4.1. Matriz de Conectividade 56
4.2. Reconhecimento plano 58
4.2.1. Decisão 59
4.2.2. Localização 61

Capítulo V - RECONHECIMENTO ESPACIAL 67
5.1. Minimização das distâncias médias 67
5.2. Minimização por translação 69
5.3. Minimização por rotação 71
5.4. Aplicação do método para a localização de substruturas 75
5.5. Critério de interrupção 76

Capítulo VI – CONCLUSÃO 78

Referências Bibliográficas 79

Anexos 82

RESUMO

Através da técnica de Morgan-Gastmans, combinada com a de Sippl-Stegbuchner, fazendo uso de recursos de IA e de lógica difusa, desenvolveu-se um método para localização de subestruturas de micromoléculas vegetais em uma base de dados de substâncias isoladas de plantas, com geometria otimizada por um programa modelador (HyperChem®). O método consiste no reconhecimento de padrões de moléculas planas, através do uso de matrizes de conectividade e na sobreposição tridimensional de moléculas pela técnica de minimização das distâncias médias entre os átomos correspondentes. Após a implementação, o método deverá integrar-se a um sistema que, futuramente, aproveitará uma base de dados para estudos em quimiosistemática e evolução química vegetal, além de servir de base para a elaboração de um programa gerador estrutural automático.

ABSTRACT

Through the technique of Morgan-Gastmans, combined with Sippl-Stegbuchner's method, using resources of Artificial Intelligence and of fuzzy logic, was developed a new method for vegetable micro molecules substructures localization in a database of isolated plant's substances, with geometry optimized for a modelating program (HyperChem®). The method consist in recognizing patterns of plane molecules, through the use of connectivity arrays, and in the superposing of three-dimensional molecules by technique of minimization of the medium distances among the corresponding atoms. After the implementation, the method should integrate into a system that will take advantage of a database for studies in chemiosistematics, and vegetable chemical evolution studies, as well as serving as a base for the elaboration of an automatic structural generating program.

CAPÍTULO I

INTRODUÇÃO

1.1. Histórico

A identificação de compostos orgânicos é um dos problemas com que se defrontam os que trabalham com a química de produtos naturais, que isolam muitos compostos. Isto é especialmente importante no Brasil, com uma flora rica, cujas espécies, em sua maioria, são ainda pouco conhecidas, embora muito contribuam, especialmente na indústria farmacêutica, para o desenvolvimento de novos produtos.

Desde épocas muito remotas se reconhece o valor alimentício e terapêutico das plantas. Posteriormente, constatou-se que alguns extratos mais concentrados poderiam produzir efeitos mais potentes e/ou rápidos. Esses extratos, entretanto, eram compostos por diversas substâncias e era provável que apenas uma ou algumas delas seriam importantes para o efeito que produziam. Passou-se, assim, a buscar formas de isolar essas substâncias. Diversas técnicas foram sendo desenvolvidas para tal fim. O passo seguinte foi determinar qual das substâncias isoladas produzia os efeitos esperados. Especialmente no que tange aos efeitos terapêuticos, é muito importante determinar também a composição molecular da substância isolada a fim de tentar sintetizá-la em laboratório. (Rodrigues, 1996)

As primeiras técnicas utilizadas, quando comparadas aos métodos atuais, eram rudimentares, consistindo especialmente em reações de degradação e observação de aspectos físicos como pontos de solidificação, ebulição e fusão, por exemplo, além da solubilidade e outras características facilmente observáveis no seu aspecto ou comportamento. Podia-se levar anos estudando uma substância e chegar a propostas apenas parcialmente corretas. Por isso, para o reconhecimento da composição molecular de uma substância, diversas técnicas independentes da análise por degradação foram sendo desenvolvidas, tais como a espectrometria na região do infravermelho (IV), a

espectrometria de massas (EM) e a ressonância magnética nuclear (RMN). Uma das mais usadas pelos químicos orgânicos é a Ressonância Magnética Nuclear de Carbono 13 (RMN ^{13}C), que submete a amostra a ser analisada a uma forte radiação eletromagnética, que é absorvida pelos núcleos do isótopo de carbono com massa 13. O resultado dessa análise aparece em um gráfico do tipo mostrado na figura 1.1, cujos picos indicam o nível de absorção de átomos de carbono com o mesmo ambiente químico e, conforme este, originam deslocamentos relativos a um padrão TMS (Tetrametil-silano).

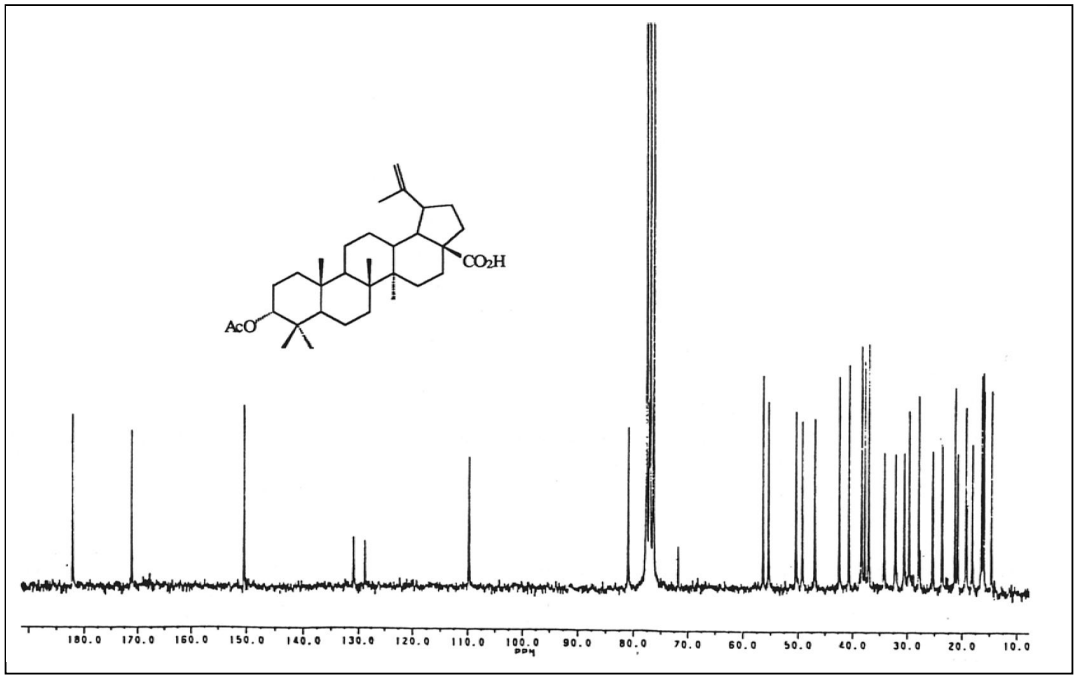


Fig. 1.1: Espectro de RMN ^{13}C do ácido betulínico (Militão, 1996)

Átomos com ambientes químicos semelhantes absorvem em regiões semelhantes, dentro do espectro de RMN ^{13}C. Essa técnica passou a ser utilizada a partir do final da década de 50, mas os primeiros trabalhos sobre ela surgiram apenas a partir de 1973. Através dela o analista estrutural pode determinar a fórmula estrutural e o grau de oxidação do composto. Isto lhe permite uma visão básica de sua classe biossintética, com o que é possível buscar na literatura padrões semelhantes. (Militão, 1996)

A RMN ^{13}C, como já foi dito, é apenas uma das técnicas possíveis para a, análise estrutural de um composto. Entretanto, pelo seu custo relativamente baixo e pela

importância das informações que fornece, essa modalidade de espectrometria é a mais utilizada pelos químicos de produtos naturais.

1.2. Determinação estrutural de um composto

A literatura dispõe de milhares de trabalhos relatando dados de RMN ^{13}C de compostos orgânicos de origem natural, especialmente vegetal, correlacionando os dados individuais de cada átomo de carbono à sua estrutura espacial. A identificação se dá, em primeiro lugar, pela determinação do tipo de átomo e de seu ambiente químico e, em seguida, pela comparação com os dados da literatura. Estabelece-se, então um esqueleto básico e, juntando-se o conhecimento do especialista sobre o assunto e condições lógicas inerentes ao tipo de substância, busca-se determinar a sua estrutura molecular. Para compostos inéditos, às vezes é possível propor estrutura somente a partes da molécula. (Borges, 1998)

Observa-se que as espécies vegetais produzem substâncias obedecendo a determinadas ordens de seu metabolismo biossintético, fazendo com que as substâncias ocorram como variações de determinados esqueletos carbônicos, cujo número é estimado em apenas algumas centenas. Uma vez de posse dos dados espectrométricos e do tipo de esqueleto, a proposição de subestruturas é possível. A busca dessas subestruturas em compostos conhecidos é fundamental para a proposição da estrutura final. (Borges, 1990)

Do exposto, duas conclusões imediatas sobressaem:

1. A determinação estrutural, além de demorada, depende de um conhecimento especializado sobre a questão, e até de uma certa intuição do químico, o que nos leva a concluir que poucos são os profissionais realmente especialistas nesse campo;

2. Os profissionais da área química, em geral, são capazes de utilizar as diversas técnicas de determinação estrutural. Entretanto, pela natureza de sua linha de pesquisa ou pela orientação do trabalho que desenvolvem, na maioria das vezes, não conseguem dedicar a essa área o tempo que seria necessário para interpretar os resultados com eficiência.

1.3. Sistemas informatizados

Para auxiliar os químicos, especialistas ou não, em questões estruturais, diversos trabalhos foram desenvolvidos utilizando sistemas computadorizados nessa área. Dentre os que adquiriram maior notoriedade, temos:

1.3.1. DENDRAL

Seu desenvolvimento iniciou-se em 1960 na Universidade de Standford. É um dos mais conhecidos e resultou em dezenas publicações que foram reunidas em dois livros publicados por participantes do projeto (Linsey *et al*, 1980 e Gray, 1986). Foi desenvolvido para uso em computadores de grande porte, utilizando dados espectrométricos diversos, especialmente de espectrometria de massa e de RMN ^{13}C, aliados a informações decorrentes da experiência do químico, o que exige grande interação com o usuário.

1.3.2. ACCESS

Também de abordagem multi-espectral (RMN 13C, RMN 1H, IV e EM), com predominância para os dados de RMN ^{13}C (mais de 60%) e tem mais de 150.000 espectros em seus bancos de dados. Desenvolvido por Bremser e colaboradores para a BASF (Bremser *et al*, 1975 e Bremser, 1988), usa ainda a interação com o usuário a fim de limitar o número de propostas geradas. Utiliza um banco de subestruturas associadas a deslocamentos químicos de RMN ^{13}C para auxiliar na geração estrutural.

1.3.3. DARC/EPIOS

Serve-se apenas de dados RMN ^{13}C e pode gerar estruturas tanto de forma automática como com interação com o usuário. Como o ACCESS, possui um banco de subestruturas, aqui chamadas de ELCOs (*Environment that is Limited, Concentric and*

Ordered - Ambiente limitado centrado e ordenado), obtidos a partir de milhares de espectros constantes da literatura. Seu desenvolvimento foi iniciado na França a partir dos anos 60 (Attias, 1983 e Carabedian *et al*, 1988)

1.3.4. CASE

Utiliza abordagem multi-espectral, mas se fundamenta especialmente em dados de RMN ^{13}C, podendo servir-se de RMN ^{1}H e IV. Este sistema também possui um banco de dados de fragmentos (Shelley et *al*, 1978 e Christie & Munk, 1991). Esses fragmentos são denominados ACFs (*Atoms-Centered Fragments* - fragmentos centrados em átomos). Também se serve de interação com o usuário que pode fornecer restrições a fim de reduzir o tempo de geração de estruturas e o número de propostas finais.

1.3.5. Outros Sistemas

Pela importância de que se reveste o assunto, muitos outros sistemas foram desenvolvidos. Alguns deles são:

Chemics: desenvolvido no Japão a partir da década de 70 por Sasaki e colaboradores, pode fazer uso de dados de diversos tipos (Sasaki *et al*, 1978).

SpecInfo: também faz uso de dados de diversos tipos como espectros de massa, infravermelho e RMN. Tem um banco de dados com cerca de 100.000 espectros, cujo tamanho se justifica pelo fato de se destinar a grandes indústrias químicas. Geralmente propõe vários candidatos, cabendo ao espectroscopista a tarefa de selecionar os melhores (Neudert *et al*, 1996).

Há ainda várias pesquisas sobre o assunto nas universidades brasileiras. Dentro da química de produtos naturais pode-se citar, por exemplo, o aplicativo **RMN-TRIT**, objeto de Tese de Doutoramento, que faz uso dos dados de RMN ^{13}C para identificação de compostos triterpênicos (Militão, 1996).

1.3.6. Sistema especialista SISTEMAT

O **Sistemat** começou a ser desenvolvido em 1988 no Instituto de Química da USP (Emerenciano *et al*, 1990). Este sistema é totalmente orientado à química dos produtos naturais, podendo rodar até em um computador XT, o que não acontece com os anteriormente citados, Apresenta especial importância ao presente projeto, razão pela qual merecerá aqui um destaque maior.

O **Sistemat**, na verdade, não é constituído de um único programa, mas de diversos aplicativos que acessam o mesmo banco de dados, subdivididos em dois pacotes (Rodrigues, 1996):

O primeiro destina-se à criação e manutenção do banco de dados, era composto inicialmente por 6 programas isolados: INICSIS, VERISIS, FONTESIS, NUMSIS, ARQSIS e DTSIS. Hoje, todos eles estão integrados em um programa único chamado DATASIS.

O segundo pacote, com programas aplicativos, utiliza os dados organizados pelo DATASIS. Os principais programas desde pacote são:

- **SISCONT***:* Utiliza apenas dados de RMN ^{13}C, confronta-os com os existentes no banco de dados e apresenta como resultado os esqueletos mais prováveis.
- **SISPICK** e **SISPICK2***:* Também usam os dados de RMN ^{13}C. Permitem que o usuário especifique requisitos químicos e/ou numeração biossintética para a identificação de compostos.
- **SISBOTA***:* Além de efetuar pesquisa bibliográfica, possibilita ao usuário estabelecer diversas limitações, tais como família, gênero, espécie, classe, esqueleto, massa molecular, índice de oxidação, e outros.
- **C13MATCH** e **MASMATCH***:* São dois programas praticamente iguais, diferindo apenas no fato de que o primeiro usa dados de RMN ^{13}C e o segundo de espectrometria de massas, confrontando os dados do espectro-problema com os do banco de dados e verificando as semelhanças existentes com os dados já cadastrados.

- **PREVSIS**, **DITREGRA** e **TRITREGRA*:*** Com abordagem multi-espectral aliada a dados botânicos, busca inferir a estrutura mais provável para um composto. São voltados especialmente para as classes dos diterpenos e triterpenos.

- **MACRONO*:*** Fazendo uso de RMN ^{13}C, busca identificar múltiplas ligações ou cadeias carbônicas como hidroxilas, halogênios e outras ligadas à cadeia principal através de um heteroátomo (normalmente o oxigênio) que aumentam o número de sinais no espectro, dificultando a identificação do composto ou conduzindo a estruturas erradas

- **SISHYPER*:*** Permite criar, para um composto cadastrado no SISTEMAT, um arquivo de texto, de extensão .HIN, que pode ser lido pelo aplicativo HyperChem®, marca registrada de HyperCube Inc. e Autodesk Inc., o qual permite realizar cálculos de energia para uma estrutura química. Esses arquivos de texto, podem ser abertos pelo HyperChem® e regravados por este após a otimização de energia, o que dá origem a um arquivo do tipo mostrado pela figura 1.2, cujos dados, segundo o próprio manual HyperChem®, tem seus significados parcialmente especificados na figura 1.3. Nesta última figura, podemos ver que, na linha iniciada pela expressão ***atom***, existem os parâmetros <x> <y> <z>, identificadores das coordenadas do átomo num sistema de coordenadas cartesianas com três eixos ortogonais, o que permite uma visão tridimensional da molécula de um composto. Observando a sintaxe da figura 1.3, nota-se que, após as coordenadas cartesianas, seguem-se o número de ligações do átomo com os demais e a quais átomos e com que tipo de ligação cada um está conectado. Por exemplo, na linha da figura 1.2 iniciada por ***atom 3***, dentre outros dados, podemos ver a seqüência "*-0.406352 1.35485 0 3 2 a 4 a 9 s*" que indica: *x*=***-0.406352***, *y*=***1.35485***, *z*=***0***, havendo ***3*** ligações com átomos vizinhos, das quais duas são ***a***romáticas, com os átomos ***2*** e ***4*** (carbonos) e uma ***s***imples com o átomo ***9*** (hidrogênio).

```
;BENZENE.HIN
forcefield amber
sys 0
view 40 0.302419 40 10 1 0 0 0 1 0 0 0 1 0.406355 0.0451501 -40
mol 1 0
atom 1 - C CA - 0 -1.61879 -0.74515 0 3 2 a 6 a 7 s
atom 2 - C CA - 0 -1.61879 0.65485 0 3 1 a 3 a 8 s
atom 3 - C CA - 0 -0.406352 1.35485 0 3 2 a 4 a 9 s
atom 4 - C CA - 0 0.806083 0.654849 0 3 3 a 5 a 10 s
atom 5 - C CA - 0 0.806083 -0.74515 0 3 4 a 6 a 11 s
atom 6 - C CA - 0 -0.406352 -1.44515 0 3 5 a 1 a 12 s
atom 7 - H HC - 0 -2.5541 1.28515 0 1 1 s
atom 8 - H HC - 0 -2.5541 1.19485 0 1 2 s
atom 9 - H HC - 0 -0.406352 2.43485 0 1 3 s
atom 10 - H HC - 0 1.74139 1.19485 0 1 4 s
```

Fig. 1.2: Exemplo de arquivo .HIN
Fonte: Manual do HyperChem

```
Hyperchem .HIN files:
This is a pretty complex format with lots of optional parameters. The
description is about 10 pages in a manual which I'd rather not retype, so here
are the basics and a couple of sample files:

; <comment>
forcefield <force-field-name>
sys <temperature>
        for moleculars dynamics
view <distance> <scale> <slab dist> <slav thickness> <viewing x-form> ....
mol <mol #> <mol name>
        starts a molecule
atom <#> <name> <element> <tipe (depends on forcefield)> <flags> <atomic
charge> <x> <y> <z> <cn> <nbor> <type> etc ...
        cn=#covalent bonds to this atom
        nbor is atom # of bonded atom
```

Fig. 1.3: Sintaxe de arquivo tipo .HIN

- **SISGER***:* Objeto de tese de doutorado (Borges, 1998), é um pacote de 5 programas (INIGER, SQUEL, INPUT, ELCOGR e SISGER), é totalmente destinado à química de produtos naturais e tem como objetivo "*propor uma estrutura orgânica a partir de dados de RMN ^{13}C, fórmula molecular e dados estruturais (esqueleto carbônico ou uma subestrutura definida pelo usuário)*" Borges, 1998). Utiliza basicamente os bancos de dados do SISTEMAT,

melhora a codificação molecular e acrescenta alguns arquivos ao banco de dados existente. A seguir, utilizando a filosofia dos "ELCOs" do sistema DARC/EPIOS (item 1.3.3), procura gerar a estrutura do composto.

O SISTEMAT vem sendo desenvolvido por diversos pesquisadores da USP há mais de 12 anos. Foi proposto para rodar nos microcomputadores do final da década de 80, os antigos XT, enquanto todos os outros sistemas voltados para elucidação estrutural rodavam apenas em computadores de grande porte. Por essa razão, o uso da memória disponível era crítico e os arquivos de dados tinham que ser compactados. Embora, hoje, isto não mais se justifique, a situação ainda permanece, praticamente igual, dificultando o uso dos dados.

O sistema operacional utilizado foi o DOS e uma tela típica do sistema se parece com o mostrado na figura 1.4. A equipe do SISTEMAT apontou a necessidade de migrar os aplicativos para uma linguagem que facilitasse a estruturação dos dados, tendo escolhido o Pascal como a plataforma de desenvolvimento futuro. (Rodrigues, 1996)

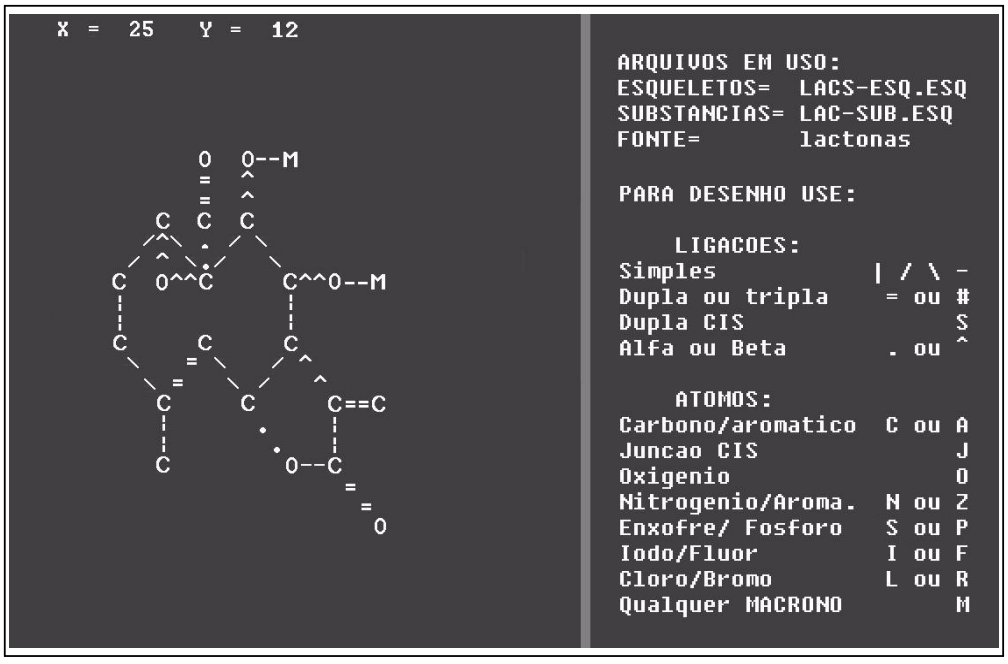


Fig. 1.4: Tela do ***Sistemat*** (sub-programa DATASIS)
Fonte: Rodrigues, 1996

A linguagem utilizada inicialmente foi o FORTRAN, seguido pelo PASCAL. Houve também algumas rotinas auxiliares desenvolvidas em dBase III Plus (Rodrigues, 1996). Os programas até hoje não estão integrados em um sistema único, mas as rotinas desenvolvidas em FORTRAN foram ou estão sendo gradativamente transformadas para o PASCAL que, hoje, é a linguagem padrão do SISTEMAT, havendo planos para portá-lo para a plataforma Windows (Borges, 1996)

As linguagens usadas não facilitam o gerenciamento do banco de dados, especialmente no que se refere à indexação, razão pela qual, após os dados de uma substância serem gravados em um arquivo mestre, é utilizada uma rotina que os distribui em diversos arquivos com um sistema de vetores (ponteiros) num relacionamento um para um (Rodrigues, 1996), o que não seria necessário se tivesse sido utilizada uma linguagem que facilitasse mais o gerenciamento. Dessa forma, o que se ganha em espaço de armazenamento é perdido com a redundância dos dados, além do processamento mais lento que isso acarreta pois, além das contínuas descompactações, o banco de dados muitas vezes deve ser percorrido inteiramente a cada busca (observado nos arquivos fonte dos programas).

1.4. Objetivo e justificativa

O presente trabalho tem como objetivo a criação de um sistema auxiliar de identificação tridimensional de micromoléculas orgânicas através da combinação de técnicas de sobreposição espacial e de conectividade molecular.

De forma mais específica, pretende-se:

- Reconhecer subestruturas em um composto: Para tanto, ter-se-á sempre a preocupação com a tridimensionalidade da molécula, conforme delineado no capítulo III deste trabalho.
- Contribuir para a determinação estrutural de compostos orgânicos de origem vegetal. Isto implica na utilização de técnicas e uso de um banco de dados (ver item 1.7, a seguir) com estrutura adequada para que o reconhecimento de subestruturas possa, em seqüência, ser utilizado na elucidação de estruturas moleculares complexas.

Isto se justifica pelo fato de que, conforme foi dito, a identificação de um composto pode consumir um tempo considerável do químico. O primeiro passo consiste em reunir o maior número de informações possível, estado físico, pontos de solidificação, ebulição e fusão, presença ou ausência de propriedades ácidas ou básicas, características de solubilidade, índice de refração, etc.

As técnicas espectrais no infravermelho ou ultravioleta, a ressonância magnética nuclear (Carbono 13 e Hidrogênio), a espectrometria de massa e outras podem fornecer informações valiosas sobre grupos funcionais, o ambiente dos hidrogênios, pesos moleculares, saturação, quantidade de carbonos, etc.

De posse de todos esses dados, a segunda fase é a pesquisa na literatura. Se todas as propriedades físicas e químicas de uma amostra forem iguais às de uma substância registrada, elas corresponderão a um mesmo composto químico. Assim, comparando-se os dados obtidos com os da literatura, pode-se identificar o composto sob análise, se ele já tiver sido estudado. A pesquisa na literatura, entretanto, se feita "manualmente", é demorada, em razão do grande número de compostos já identificados. Somente com métodos informatizados é possível fazer essa verificação com certa agilidade. Quando se tratar de composto novo, o químico defrontar-se-á com novo desafio: caracterizá-lo, elucidar a sua estrutura e publicar o resultado.

E esse é um desafio gigantesco, pois o número de compostos existentes é muito grande. Basta notar, por exemplo, que:

- Uma simples molécula de *Escherichia coli*, invisível a olho nu, tem cerca de 5.000 compostos diferentes. Somente de proteínas (moléculas complicadas, formadas por aminoácidos, com pesos moleculares variando de alguns milhares a milhões) o ser humano possui cerca de cinco milhões e elas são diferentes das encontradas em outros seres vivos. Estima-se que existam na terra cerca de 1.200.000 espécies, o que daria, somente de proteínas, 10^{12} tipos participando dos processos vitais em nosso planeta. Até hoje se conhece a estrutura completa de menos de mil dessas proteínas. E elas são apenas uma parte dos tipos de compostos orgânicos de carbono existentes (Allinger, 1978).
- O carbono, elemento básico dos compostos orgânicos, possui quatro elétrons na sua camada externa e pode compartilhar esses elétrons com elementos como nitrogênio, oxigênio, hidrogênio e cloro, e outros. Além disso, pode partilhar elétrons com

outros átomos do próprio carbono, permitindo a formação de estruturas lineares, ramificadas, cíclicas ou em forma de gaiolas. Tentar determinar todas as estruturas possíveis para determinado número de sinais de RMN ^{13}C entraria num processo de análise combinatória, gerando um número tão grande de estruturas que ultrapassaria a capacidade dos meios de armazenamento de dados. Um exemplo disso é o programa EPIOS, que não consegue gerar propostas para espectros com mias de 15 sinais de RMN ^{13}C (Borges, 1998). Por outro lado, mesmo com um número de quatro ou cinco sinais, por exemplo, a quantidade de estruturas geradas seria de tal ordem que o químico não teria como se decidir por uma delas e, conseqüentemente, o trabalho seria inócuo.

Como se soluciona, ao menos parcialmente, um problema desses? Já que a "força bruta" não consegue resolver, é necessário lançar mão de técnicas de "inteligência artificial", buscando métodos "heurísticos" que consigam reduzir o número final de estruturas geradas. É esta a razão pela qual quase todos os sistemas informatizados voltados para essa área permitem a interação com o usuário, que pode introduzir condições ou limitações baseadas em sua experiência pessoal sobre o assunto, em geral tipos de subestruturas que obrigatoriamente precisam estar presentes ou que definitivamente não devem ser consideradas. Por esta razão, esses sistemas são chamados de sistemas especialistas e os que existem estão em constante evolução pelo acréscimo de rotinas decorrentes de novos conhecimentos auferidos pelos químicos ou de novos recursos proporcionados pelo desenvolvimento da informática.

Os programas que conseguem geração automática de estruturas o fazem com base na experiência acumulada pelos técnicos, o que lhes proporciona uma "lógica" de trabalho que não apenas conduz a um número reduzido de estruturas, mas ainda permite que a geração delas seja feito num tempo relativamente curto.

É aqui que se revela a importância do **reconhecimento de subestruturas** dentro da estrutura geral de um composto. A experiência do químico pode mostrar, por exemplo, que determinados fragmentos tem a tendência de ligar-se desta ou daquela maneira a outros fragmentos. Na química orgânica, em especial, a existência de determinado grupo de subestruturas pode colocar a substância numa classe conhecida e, verificando determinadas ordens de metabolismo biossintético em que as substancias vegetais são produzidas, pode-se a eliminar muitos ramos da árvore de possibilidades.

Isso facilita a geração da estrutura de um composto inédito, além de possibilitar o estabelecimento de um índice de possibilidades, que indique quais das estruturas geradas seriam as mais ou menos prováveis.

Como objetivo secundário, pretende-se fornecer uma contribuição ao projeto SISTEMAT descrito no item 1.3.6, tanto no reconhecimento de subestruturas, como na reformulação do banco de dados e na possibilidade de migração da plataforma DOS para a plataforma Windows. Esta é a razão pela qual, anteriormente, as características deste sistema foram detalhadas em maior profundidade que as dos demais.

1.5. Método de trabalho

A fim de concretizar os objetivos acima, fundamentalmente, optou-se pela utilização de técnicas ligadas à inteligência artificial, aliadas às dos conjuntos difusos, ambas com sua fundamentação teórica detalhada no próximo capítulo. O problema de reconhecimento foi dividido em algoritmos para solucionar problemas de decisão, localização e otimização.

A **decisão** é uma simples questão de lógica boolena, ou seja, uma resposta SIM ou NÃO à questão sobre se existem ou não propriedades básicas da subestrutura pertencentes à estrutura considerada. Por exemplo, foram excluídos da busca, todos os compostos cuja quantidade de carbonos seja inferior à da subestrutura a ser localizada. Há ainda outros critérios de exclusão que serão discutidos quando for tratada a localização plana (Capítulo IV).

Havendo uma resposta SIM para a decisão, passa-se à tentativa de **localização** da subestrutura dentro da estrutura em questão. Aqui utiliza-se algoritmos com técnicas de IA, em especial a busca em amplitude, a busca em profundidade e a subida de encosta. A dificuldade básica é que, sendo grande o número de átomos da estrutura maior dentro da qual se deseja localizar a subestrutura, o número de iterações da busca pode ser elevado. O problema se agrava à medida que aumenta também o número de átomos da subestrutura a ser localizada. Trata-se, entretanto, apenas de um problema de tempo de processamento. Entretanto, como a estrutura maior tem um número finito de átomos de tipos perfeitamente determinados, geralmente menor que 50, para micromoléculas vegetais, esse tempo de processamento não é muito significativo. São

usadas, ainda, técnicas heurísticas que reduzem o tempo de busca com a utilização da técnica de IA chamada de subida de encosta, o que permite eliminar rapidamente vários ramos da árvore de busca. Para isso, o conceito de ELCO, utilizado inicialmente pelo sistema DARC/EPIOS e adotado com certa variação também pelo SISTEMAT (Borges, 1998), foi de grande valia.

Por fim, a **otimização** consiste em verificar se, encontrada uma subestrutura com os mesmos átomos e ligações, sua conformação espacial é a mesma. Três questões, nesse ponto, são cruciais:

- A otimização de energia de uma molécula baseia-se na física quântica, levando em conta os diversos tipos de átomos envolvidos e os tipos de ligações entre eles, com suas forças de atração e repulsão determinando as distâncias e ângulos relativos aos átomos. Não obstante todas as estruturas sejam otimizadas através da mesma técnica (no caso o HyperChem®), na estrutura maior os átomos da subestrutura nela presente podem sofrer desvios em razão da proximidade com os átomos restantes de seu entorno, podendo o seu posicionamento relativo não ser exatamente igual ao da subestrutura procurada.

- Os valores das coordenadas cartesianas que definem a posição de cada átomo no espaço podem ser diferentes para um mesmo composto quando se desenha sua estrutura numa escala diferente. Dessa forma, não é possível fazer coincidir, por sobreposição, as estruturas comparadas, o que também pode mascarar o resultado, fazendo parecer distintas duas estruturas que na realidade não o são.

- Em razão da tridimensionalidade da molécula (Capítulo III), a mesma figura, vista sob ângulos diferentes, pode parecer diferente.

Este último caso será solucionado pela sobreposição espacial da subestrutura à estrutura maior, utilizando método desenvolvido por Sippl & Stegbuchner, 1991.

Quanto ás duas primeiras questões (otimização de energia e escala de desenho), entretanto, faz-se necessário o estabelecimento das condições em que duas formas a rigor distintas possam ser consideradas congruentes. E isto envolve avaliações que fogem à lógica tradicional, exigindo a adoção de critérios *fuzzy* (Klir, 1997). Estes aspectos serão discutidos nos capítulos IV e V.

1.6. Linguagem utilizada

Para o desenvolvimento do sistema foi utilizado o ambiente Delphi. Esta escolha não se deve apenas ao fato de ser essa plataforma de desenvolvimento uma das mais utilizadas hoje, com excelentes recursos de gerenciamento de bancos de dados, mas especialmente por ser o PASCAL sua linguagem subjacente. Dessa maneira, uma vez que PASCAL é também uma das linguagens básicas do SISTEMAT, códigos escritos nessa linguagem quase não precisarão ser reescritos ou traduzidos, possibilitando sua utilização praticamente integral em qualquer futura transformação para a plataforma Windows.

1.7. Bancos de dados

Foram utilizados bancos de dados no formato xBase, que permitem rápida consulta, para armazenar os dados relativos à configuração espacial da molécula. Haverá dois modelos desses arquivos, um com os dados das moléculas e outro com os das subestruturas. Foram também utilizados dois tipos de arquivos texto, um para armazenar as **estruturas** e outro para as **subestruturas**, no formato do HyperChem® (figura 1.2). O programa identifica se alguma das subestruturas constantes no segundo arquivo está presente no primeiro (compostos cujas estruturas já tenham sido elucidadas).

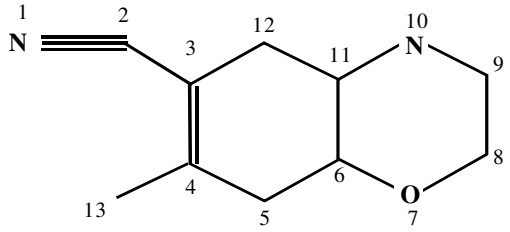
Qualquer nova estrutura ou subestrutura incluída nos arquivos deverá ser submetida à otimização de energia através do HyperChem®.

A forma de armazenamento dos dados servirá a três propósitos:

- Identificar de forma inequívoca as ligações existentes entre os diversos átomos, mostrando tanto os átomos adjacentes aos quais cada um está ligado, como o tipo de cada ligação (simples, dupla, tripla ou aromática).
- Estabelecer a posição relativa de cada átomo em relação aos demais, não somente permitindo uma visão tridimensional da molécula como eliminando os problemas decorrentes da isomeria espacial.

- Representar graficamente os compostos, facilitando a visualização na tela de um sistema computacional e possibilitando a visão sob diversos ângulos através da rotação da figura em torno de qualquer dos eixos.

Para a finalidade aqui proposta, há necessidade de uma representação dos dados adequada a propósitos de manipulações computacionais. A própria representação da uma molécula através de sua fórmula estrutural (modelo de Kekulé, figura 3.1) tem a forma clássica de um grafo, razão pela qual as representações relativas a grafos podem ser adotadas para as buscas. Para isso, normalmente são usadas tanto as matrizes quanto as listas. O SISTEMAT faz uso de matrizes, com uma codificação que permite identificar tanto os elementos envolvidos como os tipos de ligações entre eles, à semelhança de sua fórmula estrutural, conforme se vê na figura 1.5, a seguir:



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	4	3	0	0	0	0	0	0	0	0	0	0	0
2	3	1	1	0	0	0	0	0	0	0	0	0	0
3	0	1	1	2	0	0	0	0	0	0	0	1	0
4	0	0	2	1	1	0	0	0	0	0	0	0	1
5	0	0	0	1	1	1	0	0	0	0	0	0	0
6	0	0	0	0	1	1	1	0	0	0	1	0	0
7	0	0	0	0	0	1	3	1	0	0	0	0	0
8	0	0	0	0	0	0	1	1	1	0	0	0	0
9	0	0	0	0	0	0	0	1	1	1	0	0	0
10	0	0	0	0	0	0	0	0	1	4	1	0	0
11	0	0	0	0	0	1	0	0	0	1	1	1	0
12	0	0	1	0	0	0	0	0	0	0	1	1	0
13	0	0	0	1	0	0	0	0	0	0	0	0	1

Fig. 1.5: Matriz topológica representativa de estrutura

Contudo, a "espacialidade" da molécula não é contemplada aqui, o que pode ser suprido pelas coordenadas cartesianas constantes nos arquivos do tipo HyperChem®.

Note-se que a matriz que representa um composto vai além daquelas utilizadas para representar grafos. Nas matrizes de grafos os elementos são apenas 1 ou 0, indicando se dois nós estão ou não conectados. Aqui, os elementos da matriz não apenas indicam a existência da conexão, mas ainda o tipo de ligação (simples, dupla, tripla, aromática) e o "tipo" dos átomos envolvidos (Rodrigues, 1996).

A figura 1.5 mostra a fórmula estrutural e a matriz que a representa. Na diagonal principal dessa figura (sombreado claro) temos o "tipo" de átomo. Segundo a codificação adotada, vemos nessa diagonal que o átomos 1 e 10 são do tipo 4 (nitrogênio), o átomo 7 é do tipo 1 (oxigênio) e todos os demais são do tipo 1 (carbono).

Os elementos fora da diagonal principal representam o tipo de ligações. Por exemplo: o átomo 2 (linha 2) tem os elemento 3 e 1 (desprezando-se aqui o elemento da diagonal principal), respectivamente nas colunas 1 e 3; isto indica que possui uma ligação tripla com o átomo 1 e uma ligação simples com o átomo 3.

Pode-se notar, portanto, que, com exceção das informações relativas ao posicionamento espacial, todas os demais elementos da estrutura do composto (tipos de átomos, ligações e tipos de ligações) estão presentes nessa matriz. Por simplificação e para economia de memória, estão aqui omitidas as informações relativas aos átomos de hidrogênio, que podem ser facilmente deduzidas a partir dos dados da matriz, sabendo-se que a valência do carbono é 4. Por exemplo: notando-se que o carbono de número 3 possui uma ligação dupla e duas simples com os átomos vizinhos, já completou sua valência (3+1+1), não podendo haver mais nenhum átomo de hidrogênio ligado a ele. Por outro lado, o carbono de número 9 possui apenas duas ligações simples com os átomos vizinhos. Assim, para completar sua valência, é necessário que esteja ligado a mais dois átomos de hidrogênio que, aqui, não estão representados.

Uma matriz desse tipo, que inclua ou não os dados referentes aos átomos de hidrogênio, pode ser obtida facilmente, em tempo de execução, a partir dos dados constantes de um arquivo tipo HyperChem® (figura 1.2).

1.8. Arquivos xBase

Durante os trabalhos para esta dissertação, para armazenar os dados estruturais das moléculas e das subestruturas depois de otimizadas, foram utilizados arquivos com a estrutura da figura 1.6, a seguir:

Nr Átomo	El Quím	Ligações	Lig 1	Lig 2	Lig 3	Lig 4	Tipo Átomo	Coord X	Coord Y	Coord z
1	C	3	2s	10d	27s		6	-2,082822	1,346342	2,254128
2	C	4	1s	3s	18s	28s	3	-0,9476988	0,335979	2,331074
3	C	4	2s	4s	29s	30s	2	0,1621841	0,9209145	3,275352
4	C	3	3s	5d	15s		7	0,4842592	2,369217	2,921885
5	C	3	4d	6s	31s		6	-0,3335669	3,321353	3,427809
6	C	4	5s	7s	16s	32s	3	-0,5303969	4,722461	2,900488
7	C	4	6s	8s	11s	33s	3	-0,9405544	4,773518	1,384718
8	C	4	7s	9s	19s	34s	3	-2,470678	4,71648	1,073364
9	C	4	8s	10s	35s	36s	2	-3,204014	3,397624	1,489018
10	C	3	1d	9s	14s		7	-2,413493	2,127068	1,199601
11	C	3	7s	12s	13d		7	-0,2491829	6,05196	0,9643525
12	C	3	11s	16s	17d		7	0,6864418	6,435999	2,018589
13	C	3	11d	37s	38s		5	-0,4072232	6,700567	-0,2002764
14	C	4	10s	39s	40s	41s	1	-1,963723	1,861854	-0,2211566
15	C	4	4s	42s	43s	44s	1	1,556034	2,623389	1,884338
16	O	2	6s	12s			15	0,5945235	5,550527	3,040894

Fig. 1.6: Estrutura dos arquivos de armazenamento de moléculas e subestruturas

É nos dados gravados nesse arquivo que se busca identificar as subestruturas. Por isso, merece um pequeno exame.

Entre outros, possui os seguintes campos:

- ***Número do átomo***: identifica o número de ordem do átomo dentro da molécula, a fim de se poder estabelecer as ligações existentes;
- ***Ligação 1*** até ***ligação 4***: identifica os números dos átomos aos quais está ligado e o tipo de ligação (***s***imples, ***d***upla, ***t***ripla ou ***a***romática);
- ***Coordenadas de X, Y e Z***: guardam as coordenadas relativas aos três eixos;
- ***Tipo do átomo***: identifica o tipo de átomo quanto à conectividade.

CAPÍTULO II
FUNDAMENTAÇÃO TEÓRICA

2.1. Reconhecimento de padrões (RP)

2.1.1. Conceito

O objetivo do presente trabalho é o **reconhecimento de subestruturas de moléculas orgânicas de origem vegetal, em um banco de estruturas tridimensionais**. Para tanto, será elaborado um software que possibilite essa identificação de subestruturas.

E isso envolve técnicas de ***reconhecimento de padrões***, que pode ser definido como "a busca de estrutura nos dados" (Bezdek, 1981). Várias outras definições podem ser encontradas na literatura, dependendo do enfoque utilizado pelo autor e da caracterização das expressões "padrão" e "reconhecimento de padrões".

Para Uhr, sensação e percepção são fundamentais porque é através delas que as pessoas interagem com o meio ambiente e se tornam capazes de associar um determinado **padrão** ao seu significado dentro de um certo contexto, sendo o "re-conhecimento" a finalidade da sensação-percepção. (Uhr, 1973)

Para Beale e Jackson, contudo, a prioridade está no entrelaçamento do padrão com a informação, ressaltando que a maior parte das informações que absorvemos se apresenta na forma de padrões. A tarefa primordial do reconhecimento de padrões seria, então, a classificação. (Beale, 1991)

Para Fu, RP relaciona-se à descrição e à classificação de medidas obtidas em processos físicos ou mentais. Para ele, também, a classificação desempenha o principal papel no reconhecimento de padrões. (Fu, 1976)

2.1.2. Técnicas de RP

Dependendo das conceituações adotadas pelos autores, as técnicas de reconhecimento de padrões também variam. Várias abordagens são propostas, dentre as quais pode-se destacar desde a simples técnica observação visual ("*eyeball" technique* - técnica do globo ocular (Bezdek, 1981), até outras mais rigorosas em termos algorítmicos, como as abordagens estatística, da decisão teórica ou discriminante, sintática ou estrutural, redes neurais etc.

Para os propósitos deste trabalho será utilizada a técnica proposta por Bezdek, baseada numa abordagem estatística aliada à lógica difusa (Bezdek, 1981). Para Fu, essa abordagem, assim como a de redes neurais, enquadrar-se-ia no modelo de decisão teórica ou discriminante que, segundo ele, pode ser descrita através do diagrama da fig. 1.7.

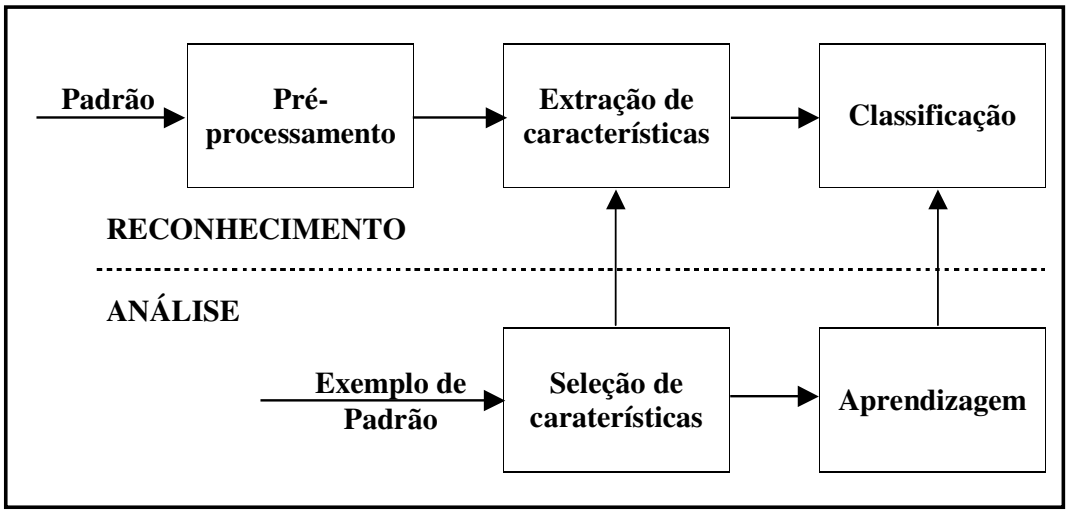


Fig. 1.7: Diagrama da abordagem de decisão teórica (Fu, 1976)

Andrews tem uma concepção similar, quando descreve o reconhecimento de padrões num mapeamento $\mathbf{P} \rightarrow \mathbf{F} \rightarrow \mathbf{C}$ (Andrews, 1972), onde:

1. **P** representa o **universo do padrão**, ou seja, um conjunto de atributos identificáveis em um objeto (padrão). Para uma pessoa, esses atributos poderiam ser nome, estado civil, idade, sexo, altura, peso, nacionalidade, naturalidade, número da cédula de identidade, número do CPF, formação e experiência profissionais etc. Esse conjunto de atributos é representado por

um vetor $X = (x_1, x_2, x_3, ..., x_k, ..., x_n)$, onde cada x_k representa um valor particular relacionado à *k*-ésima dimensão do vetor e *X* é um ponto no espaço de dimensão *n* do padrão.

2. **F** é o **universo de características**, conjunto representado por um vetor de ordem $r < n$, com a forma $Y = (y_1, y_2, y_3, ..., y_k, ..., y_r)$, cujos elementos simbolizam os atributos mais marcantes do universo *P*, que servirão de base para algum tipo classificação. Por exemplo, se se deseja classificar candidatos a um emprego, características como idade, formação e experiência profissional deverão certamente integrar o universo de características.
3. **C** é o **universo de classificação**, constituído por um conjunto de regras de decisão que se pode representar por $C = (\omega_1, \omega_2, \omega_3, ..., \omega_k, ..., \omega_c)$, um vetor de características relacionadas aos dados que se deseja classificar.

Abordagem semelhante pode-se notar em Jain e Mao para quem o reconhecimento de padrões envolve, basicamente, três etapas: aquisição de dados, representação/extração de características e tomada de decisão ou clusterização (Jain, 1994).

Em resumo, conforme Fu, 1976, pode-se dizer que, segundo a abordagem discriminante, o reconhecimento de padrões segue os três passos seguintes:

a) **Pré-processamento**: Aqui se faz a representação e codificação dos dados. Podem, ainda ser efetuadas operações destinadas a filtrar, restaurar ou realçar um padrão degradado para melhorar sua qualidade. É freqüente, aqui, o uso das transformadas de Fourier ou das transformadas de Hadamard. Assim, pode-se extrair as características do padrão, obtendo-se o vetor *X* que caracteriza o universo do padrão *P*.

b) **Seleção de características**: É a fase em que se analisa o problema e definem-se os aspectos mais marcantes do padrão a fim de formar o vetor *Y*, que representa o universo de características.

c) **Classificação**: É o passo final do reconhecimento de padrões cujas técnicas podem ser de dois tipos: numéricas e não numéricas (Beale, 1991). As primeiras envolvem medidas estatísticas e redes neurais, enquanto as últimas usam os conjuntos difusos. Trata-se de associar um vetor de

características $Y = (y_1, y_2, y_3, ..., y_k, ..., y_r)$ de dimensão *r* a uma das a uma das *c* características 2 do espaço de classificação $C = (\omega_1, \omega_2, \omega_3, ..., \omega_k, ..., \omega_c)$.

2.2. Inteligência artificial (IA)

"Teorias sobre argumentação e aprendizagem emergiram de mais de 2000 anos de tradição em filosofia, junto com o ponto de vista que o pensamento é constituído pela operação de um sistema físico. De mais de 400 anos de desenvolvimento matemático, originaram-se teorias formais de lógica, probabilidade, tomada de decisão e computação. Da psicologia obtivemos as ferramentas para investigação da mente humana e uma terminologia científica com a qual podemos expressar as teorias resultantes. Da lingüística conseguimos teorias sobre a estrutura e o significado de um idioma. Finalmente, da informática nos vieram as ferramentas para fazer da IA uma realidade." (Wooldridge, 1995, p. 8)

2.2.1. O que é inteligência artificial?

Com a introdução acima, Wooldridge estabelece os fundamentos da IA, cuja conceituação passa pelos filtros de diversas ciências. Assim, cada autor, oferece a sua, segundo determinado ponto de vista. Por isso, "uma definição capaz de separar seres inteligentes de não inteligentes, naturais e artificiais, parece escapar como uma nuvem que se tenta pegar e tem limites nebulosos. Esta nebulosidade que freqüentemente tem um caráter pejorativo, é no entanto intrínseca ao comportamento inteligente, desde sua definição até suas mais sofisticadas aplicações." (Bezerra, 1997, p. 3)

Wooldridge observa que as definições de IA variam segundo duas dimensões: algumas repousam sobre processos mentais e raciocínio, enquanto outras se apoiam nos conceitos de comportamento. Apresenta ele uma tabela com oito definições de diversos autores, reunidas duas a duas em quatro categorias, segundo a similaridade das

respectivas abordagens (pensamento humano, pensamento racional, atitude humana, atitude racional). Essas definições são transcritas a seguir (tabela 2.1).

Tabela 2.1: Algumas definições de IA - **Fonte:** Wooldrige, 1995, p. 5

"O excitante esforço novo para fazer o computador pensar... *máquinas com mentes,* no sentido completo e literal" (Haugeland, 1985)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak and McDermott, 1985)
"[Automação] de atividades associadas ao pensamento humano, como tomada de decisão, resolução de problemas, aprendizado..." (Bellman, 1978)	"O estudo da computação que torna possível perceber, raciocinar e agir." (Winston, 1992)
"Arte de criar máquinas para executar funções que requeiram inteligência quando executadas por pessoas." (Kurzweil, 1990)	"Campo de estudo que busca explicar e emular comportamento inteligente em processos computacionais." (Schalkoff, 1990)
"Estudo de como fazer com que computadores façam algo que, no momento, as pessoas fazem melhor." (Rich and Knight, 1991)	"Ramo da informática que se preocupa com a automatização do comportamento inteligente." (Luger and Stubblefield, 1993)

Assim, segundo Wooldridge, da esquerda para a direita e de cima para baixo, vê-se nessa tabela que sistemas inteligentes podem:

- **Pensar de forma humana**, o que envolve, por exemplo, ser capaz de resolver problemas "planejando um longo trajeto, considerando as várias rotas possíveis, comparando-as e escolhendo a melhor, tudo antes de iniciar a jornada". Ou seja:

  > *Se o organismo possui dentro de si um "modelo em pequena escala" da realidade externa e de suas próprias possíveis ações, é capaz de experimentar várias alternativas, decidir qual a melhor, reagir a situações futuras antes que surjam, utilizar o conhecimento anterior para lidar com o presente e o futuro e, em todos os sentidos, reagir de uma maneira muito mais completa, mais segura e mais competenteàs emergências que encontrar. (Craik, 1943, apud Wooldridge, 1995, p. 13)*

- **Pensar racionalmente**, ou seja, de forma **lógica**, deduzindo fatos novos a partir de outros conhecidos e de regras que permitam obter esses novos fatos a partir dos antigos. Isto exige uma linguagem formal com a qual o

conhecimento possa ser expresso e regras de como raciocinar a partir dela. Estes dois elementos constituem uma l**ógica**. O componente central é a **base de conhecimento**, um conjunto de representações de fatos acerca da realidade. Cada representação individual é uma **sentença**, que é expressa numa **linguagem de representação do conhecimento**.

- **Agir de forma humana** sendo capaz de: i) processar uma linguagem natural humana e comunicar-se através dela; ii) armazenar informações fornecidas durante o processamento; iii) usar essas informações para responder perguntas e extrair conclusões; iv) adaptar-se a novas circunstâncias, descobrindo e extrapolando padrões.
- **Agir racionalmente**, atingindo os objetivos a partir de premissas. Isto significa perceber e agir. Sob esse enfoque, IA é vista como o estudo e a construção de agentes racionais. Neste sentido, o sistema é capaz de sentir o ambiente e agir sobre ele de forma adequada.

2.2.2. Paradigmas da inteligência artificial

Segundo Barreto, 1997, o encontro no *Darthmouth College*, em l956, constituiu-se no primeiro esforço conjunto para estudar a IA. Estiveram presentes quatro dos grandes pesquisadores de IA Estados Unidos, Minsky, Simon, Newel e McCarthy. A este último atribui-se a criação da expressão "inteligência artificial", cunhada nesse encontro. Logo em seguida, Shannon e McCarthy publicaram o livro *Automata Studies*, onde pela primeira vez as redes neurais foram tratadas como um paradigma computacional. Das conclusões do encontro e do livro nasceram os dois paradigmas da IA: ***simbólico*** e ***conexionista***.

A inteligência artificial simbólica (IAS) lida com problemas bem definidos, cujo modo de solução seja conhecido. Aqui tudo é previsto, como raciocínio impreciso ou por falta, generalizações etc. A principal ferramenta de trabalho é a lógica, tanto a clássica quanto a difusa.

Já da inteligência artificial conexionista (IAC) espera-se um melhor desempenho em problemas mal definidos onde o conhecimento de como realizar uma tarefa não pode ser bem estabelecido. As redes neurais são bastante utilizadas e uma característica muito forte deste paradigma é a generalização, que permite ao sistema aplicar conhecimentos já aprendidos ou conhecidos na solução de outros problemas similares. O raciocínio impreciso é uma característica muito forte, pelo que, em geral, a lógica nebulosa é uma companheira assídua da IAC. A ferramenta básica da IAC é um complexo de circuitos semelhante a uma rede de neurônios cerebrais humanos. (Barreto, 1997)

Para os propósitos deste trabalho será usado o paradigma simbólico. Assim, doravante, a expressão IA isto terá o significado de "inteligência artificial simbólica".

2.2.3. Regras de produção

"Pode-se dizer que IA serve para resolver problemas, imitando de uma certa forma a inteligência dos seres vivos (geralmente seres humanos)," definindo-se por problema "o objeto matemático $P = \{ D, R, q \}$, consistindo de dois conjuntos não vazios, D os dados e R os resultados possíveis de uma relação binária $q \subset D \times R$, a condição que caracteriza uma solução satisfatória." (Barreto, 1997)

O conceito acima serve para testar se um elemento é solução, mas nada diz sobre a maneira de resolver o problema, o que é definido por Wooldridge como **percepção** de um ambiente e **ação** sobre ele, o que também nada diz sobre a forma de fazê-lo.

Um dos meios para a resolução de um problema é um mecanismo genérico chamado de **sistema de produção**, que transforma o problema em um grafo de estados em que é preciso ter um estado inicial e um forma de identificar se um estado final foi atingido. Resolver o problema é, utilizando determinadas regras, caminhar sobre esse grafo, partindo do estado inicial até atingir um estado final, se houver algum. Essa é a abordagem de Rich, 1994, que servirá de base para a seqüência do desenvolvimento deste tópico.

Um **sistema de produção** é definido como uma tupla $SP = (R,E,e_0,F)$, onde R é um conjunto de regras, E é um conjunto de estados, $e_0 \in E$ é o estado inicial e F é o conjunto de estados finais.

As regras são constituídas de um lado esquerdo (um padrão) que determina a que estados a regra pode ser aplicada, e um lado direito, que descreve a transformação a ser aplicada aos estados que se encaixam no padrão, originando novos estados. Ou seja, uma **regra de produção** é constituída por um par (p,f), onde $p:E \rightarrow \{V,F\}$ e $f:E \rightarrow E$. O elemento p é o **padrão** da regra e f constitui a **operação**.

O padrão p consiste de um predicado que mapeia o conjunto de estados do problema em valores booleanos (verdadeiro ou falso). O padrão define como verdadeiros os estados aos quais a regra é aplicável. A aplicação da regra consiste em aplicar a operação p a um destes estados, gerando um novo estado.

A descrição de um modelo formal através de um sistema de produção envolve os seguintes passos:

- Definição de um espaço de estados com todas as configurações relevantes.
- Especificação dos **estados iniciais** por onde a resolução poderá começar.
- Definição dos **estados finais** aceitáveis como solução do problema
- Especificação de um conjunto de regras descritoras das ações (operadores), compostas por um **padrão** que define os estados que podem sofrer a aplicação da regra e de uma **ação** que estabelece a maneira de construir novos estados a partir dos que pertencem a determinado padrão.

Exemplificando: um problema clássico da área de resolução de problemas, que aparece em várias obras é: "Há dois baldes sem graduação com capacidades de quatro e três litros. Fazendo uso de uma torneira e desses baldes, colocar exatamente dois litros de água no balde de quatro litros."

O espaço de estados para este problema pode ser modelado como o conjunto de pares ordenados de números naturais (x,y) tal que $x \in \{0, 1, 2, 3, 4\}$ e $y \in \{0, 1, 2, 3\}$, onde x representa a quantidade de água no balde de 4 litros, e y representa a quantidade de água no balde de 3 litros.

O estado inicial do problema é o estado no qual ambos os baldes estão vazios, (0,0), e o conjunto de estados finais é constituído por todos os estados onde a quantidade de água no primeiro balde é 2, ou seja, $(2,n)$, $n \in \{0, 1, 2, 3\}$.

Um possível conjunto de regras para este problema seria:

Tabela 2.2: Exemplo de conjunto de regras para sistema de produção

r_1	$(x,y \mid x<4) \rightarrow (4,y)$	Encher o balde de 4 litros
r_2	$(x,y \mid y<3) \rightarrow (x,3)$	Encher o balde de 3 litros
r_3	$(x,y \mid x>0) \rightarrow (0,y)$	Jogar fora toda a água do balde de 4 litros
r_4	$(x,y \mid y>0) \rightarrow (x,0)$	Jogar fora toda a água do balde de 3 litros
r_5	$(x,y \mid x+y>4) \rightarrow (4,y-(4-x))$	Despejar água do balde de 3 litros no balde de 4 litros até enchê-lo
r_6	$(x,y \mid x+y>3) \rightarrow (x-(3-y),3)$	Despejar água do balde de 4 litros no balde de 3 litros até enchê-lo
r_7	$(x,y \mid x+y\leq 4$ e $y>0) \rightarrow (x+y,0)$	Despejar toda a água do balde de 3 litros no balde de 4 litros
r_8	$(x,y \mid x+y\leq 3$ e $x>0) \rightarrow (0,x+y)$	Despejar toda a água do balde de 4 litros no balde de três litros

Uma solução possível para o problema seria aplicar em seqüência as regras r_2, r_7, r_2, r_7, r_5, r_3 e r_7.

A modelagem de um problema como um sistema de produção consiste apenas em definir o espaço de estados e as regras. Isto, entretanto, não mostra como chegar à solução, isto é, como encontrar um estado final em tempo razoável. Par tanto, são usadas técnicas de busca em grafos, algumas das quais são mostradas a seguir:

2.2.4. Busca em amplitude

A busca em amplitude consiste em construir uma árvore de estados a partir do inicial, aplicando, a cada momento, todas as regras possíveis aos estados do nível mais

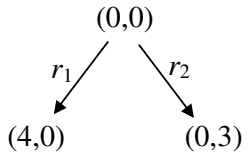
baixo da árvore, gerando todos os sucessores de cada estado. Assim, a busca por um estado final se dá por níveis, na árvore do problema.

Na notação apresentada para o algoritmo, são usadas funções de acesso a lista conforme a definição da linguagem LISP. Assim, a função CAR retorna o primeiro elemento de uma lista, enquanto CDR retorna o resto da lista, isto é, a lista sem o primeiro elemento. Nenhuma destas funções altera a lista original. Por exemplo, se x := [1,2,3] e y := CDR(x), então y = [2,3], mas x continua sendo igual a [1,2,3]. A função *append* tem dois parâmetros, uma lista e um elemento. Esta função retorna uma nova lista com o elemento acrescido ao final da lista, e também não altera a lista original. Para obter o efeito de alteração da lista original, pode-se usar a atribuição. Por exemplo, para eliminar o primeiro elemento de uma lista pode-se fazer: x := CDR(x). A seguir é apresentado o algoritmo da busca em amplitude.

algoritmo BuscaEmAmplitude (R, E, e_0, F)

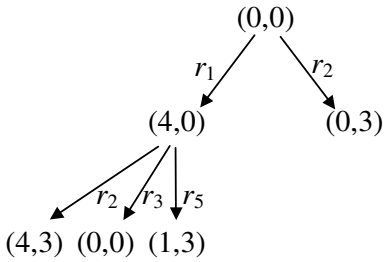
1. *ListaDeNós* := [e_0];
2. enquanto não Vazia(*ListaDeNós*) faça:
   - 2.1 e := CAR(*ListaDeNós*);
   - 2.2 *listaDeNós* := CDR(*ListaDeNós*);
   - 2.3. Para toda $r_i \in R$ tal que $r_i(E) \in E$ faça:
     - 2.3.1. e' := $r_i(e)$;
     - 2.3.2. Se $e' \in F$ então retorne e';
     - 2.3.3. *ListaDeNós* := append(*ListaDeNós*,e');

Desenvolvendo o algoritmo para o exemplo das duas jarras, atribui-se o estado inicial (0,0) à lista de nós pela instrução 1. Em seguida, gera-se os sucessores desse estado (instrução 2.3). Observando-se a tabela 2.2, nota-se que apenas as regras 1 e 2 podem ser aplicadas ao estado inicial. Assim, tem-se:



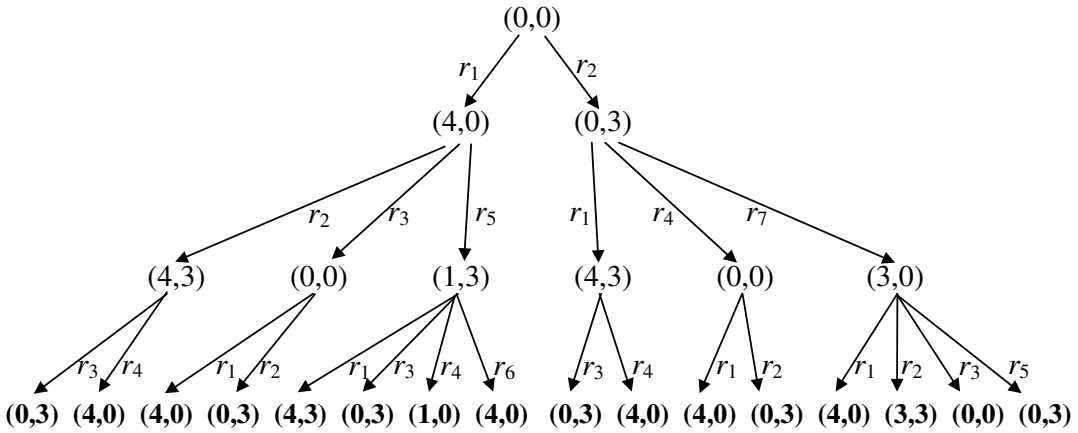
A instrução 2.3.3 acrescentou os dois últimos estados a *ListaDeNós* e, como a instrução 2.2 havia retirado o estado inicial, a nova *ListaDeNós* é igual a [(4,0), (0,3)], ou seja, os estados que ainda não foram expandidos.

Na segunda aplicação da instrução 2.3, são gerados inicialmente os sucessores de (4,0):



Neste momento, *ListaDeNós* = [(0,3), (4,3), (0,0), (1,3)], já que (0,3), que já estava na lista ainda não foi explorado e os sucessores de (4,0) foram adicionados no final da lista pela instrução *append*. Então, o próximo estado a ter seus sucessores gerados por busca em extensão será (0,3), que será retirado da *ListaDeNós*.

Na seqüência do algoritmo, são gerados os sucessores de (4,3), (0,0), (1,3), (4,3), (0,0) e (3,0), nesta ordem, resultando na seguinte árvore:



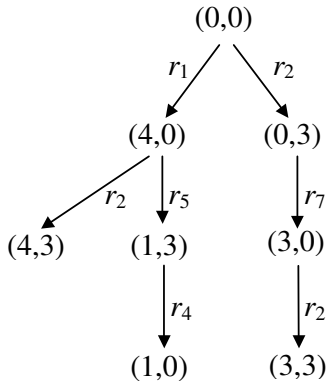
Neste momento, *ListaDeNós* = [(0,3), (4,0), (4,0), (0,3),(4,3),(0,3), (1,0),(4,0), (0,3), (4,0), (4,0), (0,3),(4,0),(3,3), (0,0),(0,3)],

Um problema com este algoritmo é que nós já visitados na busca podem ser visitados novamente se forem sucessores de outros estados. Esta repetição é desnecessária, porque, por exemplo, as sub-árvores abaixo de (0,0) no segundo e terceiro níveis serão exatamente iguais à sub-árvore de (0,0) no primeiro nível, isto é, elas terão exatamente os mesmos sucessores. Para evitar essa repetição, o algoritmo pode ser reescrito eliminando a busca em estados já visitados, usando-se a busca em amplitude sem repetição de estados a seguir:

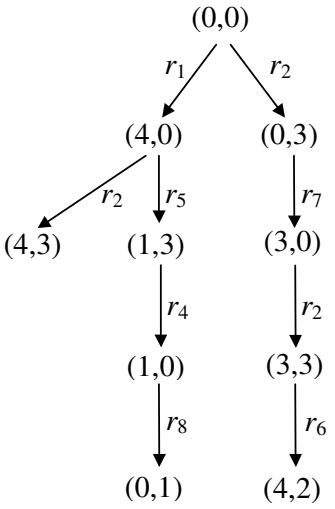
algoritmo BuscaEmAmplitudeSemRepetiçãoDeEstados (R,E,e_0,F)

1. *listaDeNós* := [e_0];
2 *visitados* :=∅;
3. enquanto não Vazia(*listaDeNós*) faça:
    3.1. e:=CAR(*listaDeNós*);
    3.2. *listaDeNós*:=CDR(*listaDeNós*);
    3.3. *visitados* := *visitados*∪{e};
    3.4. Para toda $r_i \in R$ tal que $r_i(e) \in E$ faça:
        3.4.1. e':=$r_i(e)$;
        3.4.2. Se $e' \in F$ então retorne e';
        3.4.3 Se $e' \notin$ *visitados* então *listaDeNós* :=append(*listaDeNós*,e');

Com a utilização deste algoritmo, a última árvore do exemplo anterior ficaria:



É mais fácil, então continuar a expansão desta árvore. Na interação seguinte tem-se:



E assim por diante. Agora, o algoritmo caminha rapidamente para a solução. Basta aplicar sobre o ramo da esquerda as regras r_1 e r_6 para chegar a (2,1) que é uma solução, ou as regras r_3 e r_7 sobre os ramos da esquerda para obter (2,0), que também é um dos estados finais.

2.2.5. Busca em Profundidade

Enquanto a busca em amplitude procura expandir todos os nós em um determinado nível da árvore antes de partir para o nível seguinte, a busca em profundidade procura explorar completamente cada ramo da árvore antes de tentar o ramo vizinho, como segue:

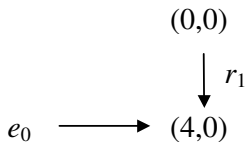
Algoritmo BuscaEmProfundidade(R,E,e_0,F,*visitados*):

1. Se $e_0 \in F$ então retorne (e_0,*sucesso*);
2. Para toda $r_i \in R$ tal que $r_i(e_0) \in E$ e $r_i(e_0) \notin$ *visitados* faça:
   2.1. (e,x) := *BuscaEmProfundidade*(R,E,$r_i(e_0)$,F,*visitados*$\cup\{r_i(e_0)\}$)
   2.2. Se x = *sucesso* então retorne(*sucesso* e,x)
   2.3. Caso contrário, retorne (*nada*,*fracasso*)

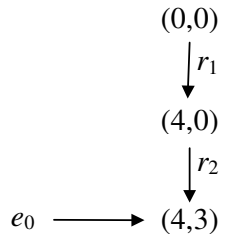
O algoritmo deve ser invocado na sua primeira chamada com o sistema de produção e mais a lista de nós inicializada como uma lista vazia. Ele verifica se cada uma das sub-árvores do nós e_0 possui a solução do problema. A verificação em cada sub-árvore consiste em chamar recursivamente *BuscaEmProfundidade* passando o nó filho como nó inicial.

Para exemplificar a utilização de busca em profundidade, é apresentado a seguir o desenvolvimento do algoritmo para o mesmo sistema de produção que serviu de exemplo à busca em extensão.

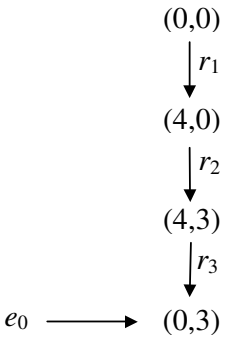
O algoritmo inicia com o estado (0,0). A primeira regra aplicável é r_1, gerando o estado (4,0) . Então o algoritmo é chamado recursivamente com o estado (4,0) como estado inicial:



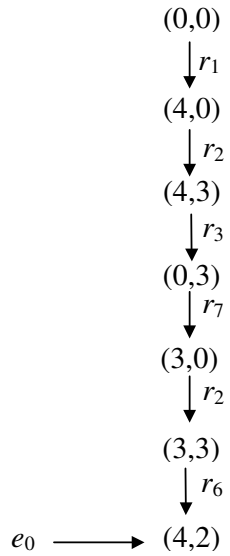
Como (4,0) não é o estado final, então o algoritmo é chamado recursivamente outra vez. A primeira regra não é aplicável a este estado, mas r_2 produz o estado (4,3). Chega-se, assim, à seguinte situação, após uma nova chamada recursiva ao algoritmo:



Para este estado, as regras r_1 e r_2 não se aplicam, mas a regra r_3 produz:



No estado (0,3), as regras r_1 e r_4 produzem estados nós já visitados e, por isso, não são aplicadas. As regras r_2 e r_3 não são aplicáveis. As regras r_5 e r_6 não são aplicáveis. Já a regra r_7 pode ser aplicada e leva a (3 ,0). A seguir, a única regra aplicável que não conduz a um estado já visitado é r_2 que produz o estado (3,3) e, em seguida, r_6 leva ao estado (4,2). Depois disso o algoritmo chegará rapidamente a um estado final.



Quando nenhuma regra puder ser aplicada ocorre o processo que se chama *backtracking*, ou seja, o algoritmo volta atrás e tenta outro caminho.

A principal vantagem do algoritmo de **busca em extensão** é encontrar o menor caminho do nó inicial até o nó final mais próximo. Já o algoritmo de **busca em profundidade** não encontra necessariamente a solução mais próxima, mas pode ser mais eficiente se o problema possui um número grande de soluções ou se a maioria dos caminhos pode levar a uma solução.

2.2.6. Subida de Encosta

Outros métodos mais eficientes do que busca em profundidade e busca em extensão podem ser empregados se o espaço de estados do problema puder ser organizado em uma estrutura de ordem.

Empregando uma ordenação total ou parcial no conjunto de estados, é possível dizer se um estado sucessor leva para mais perto ou para mais longe da solução. Assim, o algoritmo de busca pode preferir explorar em primeiro lugar os estados que levam para mais perto da solução, realizando, assim, uma espécie e subida de encosta, na qual o algoritmo só se move na direção de estados melhores.

A subida de encosta é uma variante da busca em profundidade. A diferença básica é que ao invés de simplesmente selecionar a regra a aplicar arbitrariamente do conjunto de regras existentes, o algoritmo dá preferência àquelas regras que levam a estados melhores.

Para que se possa comparar dois estados e decidir se um é melhor do que o outro, é necessário que o conjunto de estados possua uma relação de ordem. Assim, nos exemplos seguintes, o conjunto de estados E será substituído por uma estrutura de ordem total irreflexiva $(E,<)$, onde $<$ é uma relação transitiva, anti-simétrica e irreflexiva.

Por definição, de $e_i<e_j$ então se diz que o estado e_j está mais próximo da solução do que o estado e_i.

Há duas variações do método de subida de encosta: a *subida de encosta simples* e a *subida de encosta pela trilha mais íngreme*. A diferença entre os dois métodos está no fato de que a subida de encosta pela trilha mais íngreme examina todos os sucessores do estado e escolhe entre estes sucessores qual é o que está mais próximo da solução, e então segue a expansão por este estado. A subida de encosta simples não examina este fato, seguindo simplesmente para o primeiro estado encontrado que seja melhor do que o atual. Este procedimento pode ser mais eficiente, principalmente se o número de regras for muito grande. Todavia, em muitos casos, a subida de encosta pela trilha mais íngreme poderá ser mais efetivo por se aproximar mais rapidamente da solução.

A seguir, apresenta-se um modelo de cada algoritmo:

Algoritmo SubidaDeEncosta (R,(E,<),e_0,F,*visitados*):

1. Se $e_0 \in F$ então retorne (e_0,*sucesso*);
2. Para toda $r_i \in R$ tal que $r_i(e_0) \in E$ e $r_i(e_0) \notin$ *visitados* e $e_0 < r_i(e_0)$ faça:
   2.1 (e,x):=*SubidaDeEncosta*(R,E,$r_i(e_0)$,F,*visitados*$\cup\{r_i(e_0)\}$)
   2.2 Se x = *sucesso* então retorne(*sucesso* e,x)
   2.3 Caso contrário, retorne (*nada*,*fracasso*)

Algoritmo SubidaDeEncostaÍngreme (R,(E,<),e_0,F,*visitados*):

1. Se $e_0 \in F$ então retorne (e_0,*sucesso*);
2. *sucessores* := {e|para toda $r_i \in R$, $e=r_i(e_0)$ e $e \notin$ *visitados*}
3. Enquanto $x \neq \varnothing$ faça:
   3.1 e':= $max_<$(*sucessores*);
   3.2 *sucessores*:=*sucessores*-{e'};
   3.3 (e,x):=*SubidaDeEncostaÍngreme*(R,(E,<),e',F,*visitados*$\cup\{e'\}$)
   3.4 Se x=*sucesso* então retorne(e,x)
4. Retorne (*nada*,*fracasso*)

Este último tipo de algoritmo será utilizado na localização plana de uma subestrutura (ver capítulo IV).

2.3. A incerteza

> *Nebuloso, vago, impreciso e qualitativo são termos que para muitos tem uma conotação pejorativa. É comum se ter respeito pelo que é preciso, claro, e que se olhe com desdém para o que é desprovido da precisão que se costuma atribuir à matemática. Mas apesar disto, quanto mais se aprende sobre cognição mais se percebe que a habilidade de manipular conceitos mal definidos é uma das maiores qualidades, e não defeito, e que isto constitui a chave da diferença entre inteligência humana e a da máquina. (Barreto, 1997)*

Bezdek, 1981, afirma que "um dos mais divertidos paradoxos da ciência moderna é a tentativa de capturar *precisamente* o valor e os efeitos da incerteza nos modelos matemáticos." Segundo ele, as principais causas da incerteza são: falta de precisão nas medidas, ocorrências aleatórias e descrições vagas. Essas três manifestações de incerteza são ditas determinísticas, probabilísticas e nebulosas, respectivamente.

Num processo determinístico o resultado pode ser previsto com absoluta certeza repetindo-se as circunstâncias que lhe deram origem, o que nem sempre é possível fazer exatamente, como por exemplo, na determinação da quantidade de bactérias presentes numa amostra num determinado momento. Por outro lado, mesmo sendo o processo repetido exatamente, medidas imprecisas podem ser causa de incerteza.

A incerteza probabilística deriva de fenômenos que ocorrem aleatoriamente, como no lançamento de uma moeda. Modelos de processos com esse tipo de incerteza são chamados de estocásticos ou probabilísticos.

No primeiro caso a incerteza não decorre do processo, mas da incapacidade de executá-lo ou monitorizá-lo. No segundo, o próprio resultado é incerto.

Considere-se, agora, a pergunta "A altura da pessoa x está próxima de dois metros?" Ainda segundo Bezdek, 1981, respondê-la significa tentar associar-lhe uma resposta **sim** ou **não**, o que pode suscitar certa dúvida, que pode decorrer de erro de medida, de uma ocorrência aleatória e até do próprio significado da expressão 'próxima'.

Em outros termos: seja X um conjunto de pessoas e A_1 o subconjunto de X para o qual a altura $h(x)$ é exatamente dois metros, ou seja: $A_1 = \{x \in X \mid = 2$. Supondo que

'x próximo de 2 metros' se e somente se $x \in A_1$, o conjunto A_1 será muito pequeno, inclusive porque é bastante difícil medir $h(x)$ exatamente. Para superar isto, poder-se-ia considerar os 'x próximos de 2 metros' pertencentes ao conjunto $A_2 = \{x \in X \mid x = 2 \pm 0{,}005\}$. Agora o conjunto A_2 possui um número maior de elementos que A_1. Pode-se, então, concluir que a expressão 'próximo', neste contexto, está ligada à imprecisão de medida e o modelo seria determinístico.

Seja agora X um subconjunto de um conjunto de n pessoas de S, a população de um país. Seja ainda $x \in X$ o evento elementar de uma "ocorrência de x" e $h: X \rightarrow (0, \infty)$ a variável aleatória que expressa a "altura de x". Definindo o conjunto das pessoas com altura 'próxima de 2' como $A_2 = \{ x \in X \mid 1{,}995 \leq h(x) \leq 2{,}005 \}$, tem-se uma outra representação para o mesmo conjunto A_2 acima. Uma série adequada de observações seguida de inferência estatística, pode atribuir a cada x a sua probabilidade de pertencer a A_2, de forma que se possa estimar a probabilidade $Pr(1{,}995 \leq h(x) \leq 2{,}005) = Pr(x \in A_3)$. Supondo que se chegue a $Pr(x \in A_3) = 0{,}95$, isto significa ser de 95% a possibilidade de que $h(x)$ esteja no intervalo [0,995; 2,005]. Neste caso, a incerteza estaria associada a um modelo probabilístico.

Há ainda uma terceira abordagem sobre essa incerteza, decorrente do próprio significado da palavra próximo, conforme se verá a seguir.

2.3.1. Conjuntos nebulosos

A idéia dos conjuntos nebulosos nasceu com L. A. Zadeh e a publicação do seu artigo "*Fuzzy Sets*" (Zadeh, 1965) criou o nome e formalizou a teoria. A expressão "nebuloso" foi cunhada quando Jean Paul Jacob, da IBM, responsável pelo financiamento das pesquisas de Zadeh, proferiu uma palestra no IME - Instituto Militar de Engenharia, em 1971. Em 1974, após uma palestra ocorrida no INPE - Instituto Nacional de Pesquisas Espaciais, um oficial sugeriu o termo "difuso" e esta, talvez, seja a origem do termo também usado no Brasil. "Esta terminologia é no entanto imprópria pois difusão se refere a um fenômeno físico bem definido e que implica em um movimento. Por exemplo, difusão de neurotransmissores no espaço sináptico, difusão de moléculas de uma gota de tinta em um copo d'água. Fala-se também de luz difusa ao se referir a uma iluminação que vem de todas as direções, etc. Nenhum desses conceitos

exprime que uma afirmação possa ser mais pertinente que outra nem que uma afirmação possa ter graus de pertinência, dependendo do caso; o conceito de nuvem em que se entra mais ou menos dentro dela exprime perfeitamente o caso." (Barreto, 1997)

No item anterior discutiu-se sobre a incerteza ligada à pergunte "A altura da pessoa x está próxima de dois metros?" Em cada um dos casos, a resposta seria **sim** somente se a pessoa pertencesse a um conjunto (A_1 ou A_2) das pessoas cuja altura estivesse próxima de dois metros, segundo modelos determinísticos ou probabilísticos, respectivamente (Bezdek, 1981). Em outras palavras, atribuindo-se valor 1 à resposta **sim** e valor 0 à reposta **não**, a função característica, $h_n: X \rightarrow \{0, 1\}$, dessa pertinência ao conjunto A_1 ou ao conjunto A_2, é do tipo:

$$h_1(x) = \left\{ \begin{array}{ll} 1; & x \in A_1 \\ 0; & \text{caso contrário} \end{array} \right\} \quad \text{ou}$$

$$h_2(x) = \left\{ \begin{array}{ll} 1; & x \in A_2 \\ 0; & \text{caso contrário} \end{array} \right\} .$$

Nota-se que a pertinência ao conjunto A_n é a chave para qualquer decisão e o raciocínio se encaixa na lógica booleana ou bi-valorada (sim/não, verdadeiro/falso, 0/1).

Os conjuntos nebulosos introduzidos por Zadeh, entretanto, levam a uma terceira possibilidade, segundo a qual, pode-se definir um conjunto de forma semelhante a:

$$A_3 = \{\, x \in X \mid x \text{ tem altura próxima de 2 metros} \,\},$$

cuja função de pertinência pode ser definida como:

$$h_3(x) = \left\{ \begin{array}{lll} 1; & 1{,}995 \le h(x) \le 2{,}005 & \\ 0{,}95; & 1{,}990 \le h(x) < 1{,}995 \text{ ou } & 2{,}005 < h(x) \le 2{,}010 \\ \vdots & & \vdots \\ 0{,}05; & \cdots & \cdots \\ \vdots & & \vdots \end{array} \right\}$$

Aqui, também, a decisão depende de x pertencer ou não ao conjunto A_3. Entretanto, essa pertinência, agora, se faz **segundo determinado grau**, o que permite julgamentos qualitativos sobre alturas relativas. Assim, sabendo-se que $h_3(y) = 0{,}65$ e $u_3(x) = 0{,}95$, percebe-se que x está mais próximo de dois metros que y, embora tanto x quanto y pertençam a A_3. O grau de pertinência pode variar no intervalo de 0 a 1, ou

seja, $h_3: X \rightarrow [0, 1]$. Como se pode perceber, a função $h_n: X \rightarrow \{0, 1\}$, acima citada, é um caso particular de h_3 (Bezdek, 1981)

A representação gráfica das funções acima pode ser feita na forma a seguir (figura 2.1):

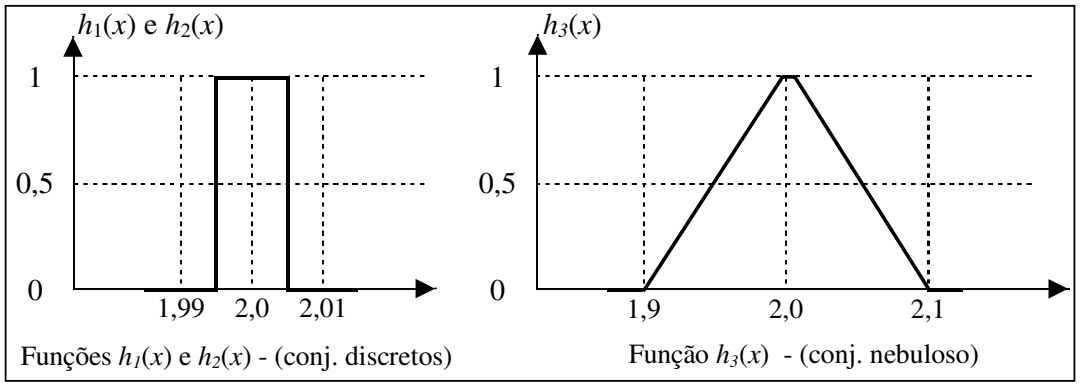


Fig. 2.1: Representação gráfica das funções h_1, h_2 e h_3 (alturas próximas de 2 metros)

2.3.2. Operações com conjuntos nebulosos

As operações básicas com conjuntos nebulosos, segundo Klir, 1997, São definidas da seguinte forma:

- **Complementação**: Dado um conjunto nebuloso A, definido no conjunto universo X, seu complemento $\overline{A}$ é outro conjunto nebuloso definido sobre X de forma que, enquanto para todo $x \in X$, $A(x)$ expressa o grau segundo o qual x **pertence** A, $\overline{A}(x)$ expressa o grau segundo o qual x **não pertence** A. Isto pode ser expresso pela fórmula:

$$\overline{A}(x) = 1 - A(x)$$

- **Reunião:** Considerando-se dois conjuntos A e B definidos sobre um conjunto universo X, a reunião $A \cup B$ é definida com base em seus graus de pertinência pela fórmula:

$$(A \cup B)(x) = \max[A(x), B(x)]$$

- **Intersecção:** Sejam também se dois conjuntos A e B definidos sobre um conjunto universo X. Sua intersecção $A \cap B$ é definida por:

$$(A \cap B)(x) = \min[A(x), B(x)]$$

Exemplificando:

Sejam *A*, *B* e *C* os conjuntos nebulosos *jovem*, *meia-idade* e *velho*, definidos, respectivamente, pelas seguintes funções de pertinência no intervalo [0, 100]:

$$A(x)=\begin{cases} 1 & \text{se } x<20 \\ \dfrac{35-x}{15} & \text{se } 20\le x\le 35 \\ 0 & \text{caso contrário} \end{cases}$$

$$B(x)=\begin{cases} \dfrac{x-20}{15} & \text{se } 20\le x<35 \\ 1 & \text{se } 35\le x\le 50 \\ \dfrac{65-x}{15} & \text{se } 50<x\le 65 \\ 0 & \text{caso contrário} \end{cases}$$

$$C(x)=\begin{cases} 0 & \text{se } x<50 \\ \dfrac{x-50}{15} & \text{se } 50\le x\le 65 \\ 1 & \text{caso contrário} \end{cases}$$

Sua representação gráfica é a das figuras 2.2, 2.3 e 2.4, a seguir:

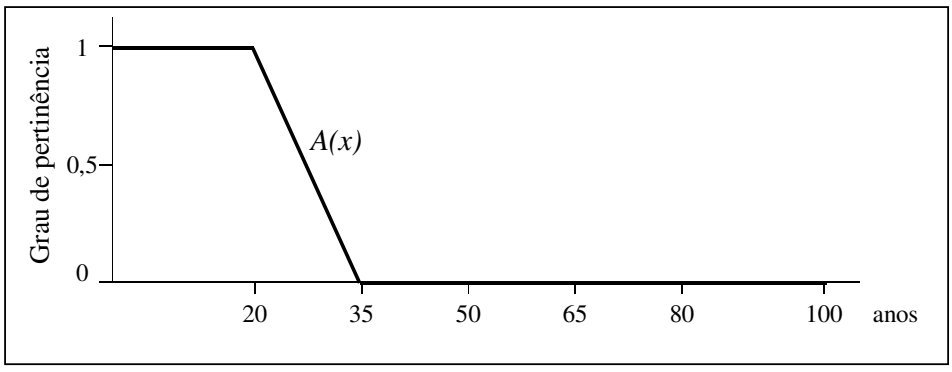


Fig. 2.2: Conjunto nebuloso "jovem"

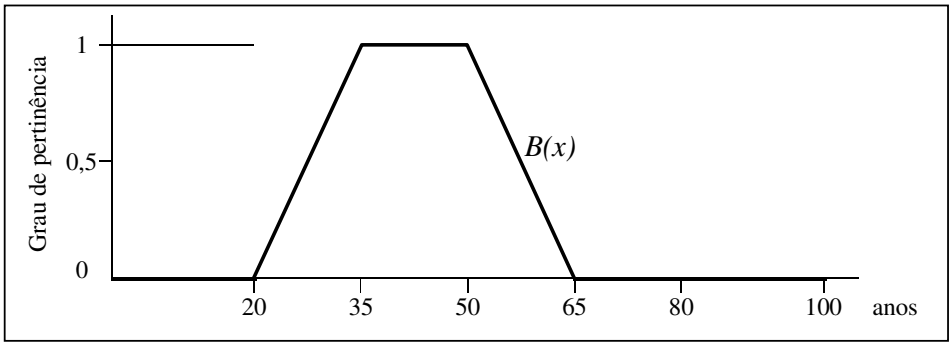


Fig. 2.3: Conjunto nebuloso "meia idade"

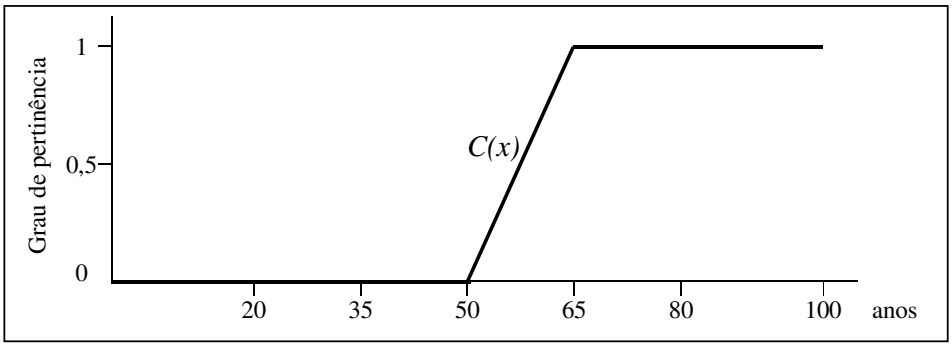


Fig. 2.4: Conjunto nebuloso "velho"

Os **conjuntos nebulosos complementares** de *A*, *B* e *C* são dados pelas funções e pelos gráficos seguintes:

Complemento do conjunto difuso “meia-idade”:

$$\overline{A}(x) = 1 - A(x)$$

$$\overline{A}(x) = \begin{cases} 0 & \text{se } x < 20 \\ \dfrac{x - 20}{15} & \text{se } 20 \leq x \leq 35 \\ 1 & \text{caso contrário} \end{cases}$$

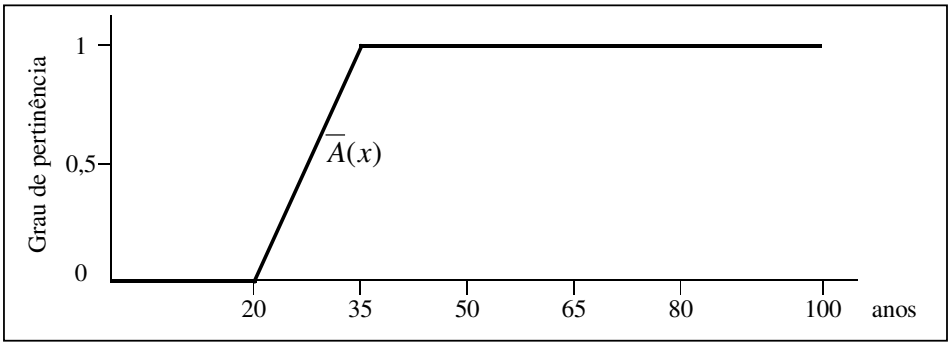


Fig. 2.5: Conjunto nebuloso complementar de "jovem"

Complemento do conjunto difuso “meia-idade”:

$$\overline{B}(x) = 1 - A(x)$$

$$\overline{B}(x) = \begin{cases} \dfrac{35 - x}{15} & \text{se } 20 \le x \le 35 \\ 0 & \text{se } 35 \le x \le 50 \\ \dfrac{x - 50}{15} & \text{se } 50 \le x \le 65 \\ 1 & \text{caso contrário} \end{cases}$$

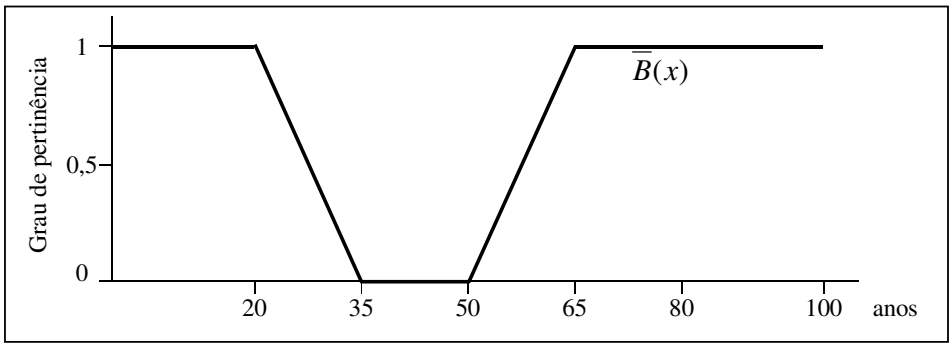


Fig. 2.6: Conjunto nebuloso complementar de "meia idade"

Complemento do conjunto difuso “velho”:

$$\overline{C}(x) = 1 - C(x)$$

$$\overline{C}(x) = \begin{cases} 1 & \text{se } x < 50 \\ \dfrac{65 - x}{15} & \text{se } 50 \le x \le 65 \\ 0 & \text{caso contrário} \end{cases}$$

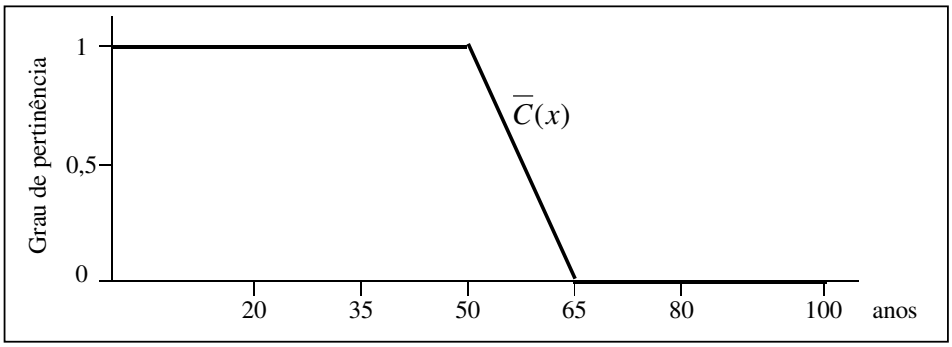


Fig. 2.7 - Complemento do conjunto nebuloso "velho"

A **interseção dos conjuntos nebulosos** *A*, *B* e *C* é calculadas através da fórmula $(X \cap Y)(x) = \min[X(x), Y(x)]$ e dada pelas funções e pelos gráficos seguintes:

Interseção dos conjuntos "jovem" e "meia-idade":

$$(A \cap B)(x) = \min[A(x), B(x)] = \begin{cases} 1 - \dfrac{|27{,}5 - x|}{15} & \text{se } 20 \le x \le 35 \\ 0 & \text{caso contrário} \end{cases}$$

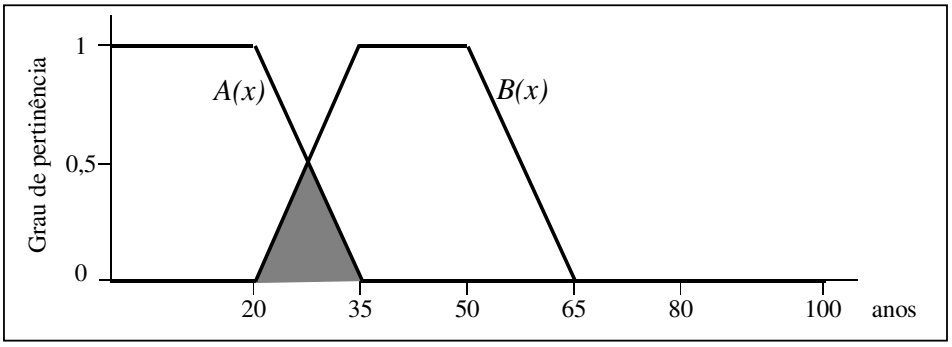


Fig. 2.8 - Intersecção dos conjuntos nebulosos "jovem" e "meia-idade"

A interseção dos conjuntos "jovem" e "velho" é um conjunto vazio ($A \cap C = \varnothing$), ou seja:

$$(A \cap B)(x) = 0, \quad \forall x \in \mathrm{I} \mid 0 \le x \le 100$$

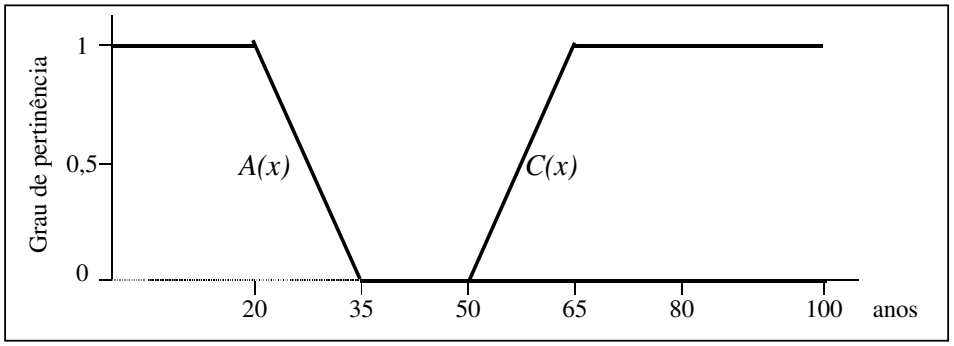


Fig. 2.9 - Intersecção dos conjuntos nebulosos "jovem" e "velho"

A interseção dos conjuntos “meia-idade” e “velho” é dada pela função de pertinência e pelo gráfico a seguir:

$$(B \cap C)(x) = \begin{cases} 1 - \dfrac{|57{,}5 - x|}{15} & \text{se } 50 \le x \le 65 \\ 0 & \text{caso contrário} \end{cases}$$

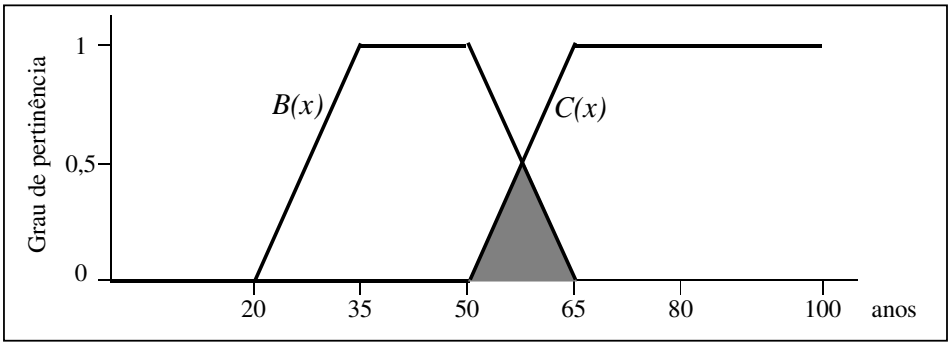


Fig. 2.10 - Interseção dos conjuntos nebulosos "meia-idade" e "velho"

A **reunião dos conjuntos nebulosos** *A*, *B* e *C* é calculadas através da fórmula $(X \cup Y)(x) = \max[X(x), Y(x)]$ e definida pelas funções de pertinência e pelos gráficos a seguir:

Reunião dos conjuntos nebulosos “jovem” e “meia-idade”:

$$(A \cup B)(x) = \begin{cases} 0{,}5 + \dfrac{|x - 27{,}5}{15} & \text{se } 20 \le x \le 35 \\ \dfrac{65 - x}{15} & \text{se } 50 \le x \le 65 \\ 0 & \text{se } x > 65 \\ 1 & \text{caso contrário} \end{cases}$$

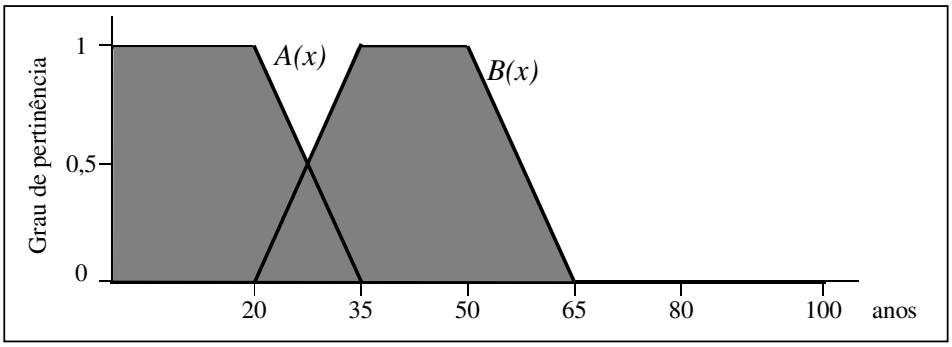


Fig. 2.11 - Reunião dos conjuntos nebulosos “jovem” e “meia-idade”

Reunião dos conjuntos nebulosos “jovem” e “velho”:

$$(A \cup B)(x) = \begin{cases} 0 & \text{se } 35 \le x \le 50 \\ \dfrac{35 - x}{15} & \text{se } 20 < x < 35 \\ \dfrac{x - 50}{15} & \text{se } 50 < x < 65 \\ 1 & \text{caso contrário} \end{cases}$$

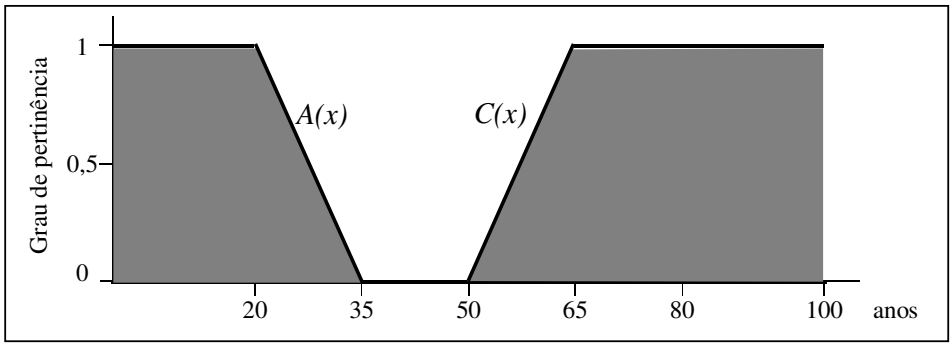


Fig. 2.12 - Reunião dos conjuntos nebulosos “jovem” e “velho”

A reunião dos conjuntos nebulosos “meia-idade” e “velho” é dada por:

$$(B \cup C)(x) = \begin{cases} 0 & \text{se } x < 20 \\ \dfrac{x-20}{15} & \text{se } x \le 20 \le 35 \\ \dfrac{0{,}5+|57{,}5|}{15} & \text{se } 50 \le x \le 65 \\ 1 & \text{caso contrário} \end{cases}$$

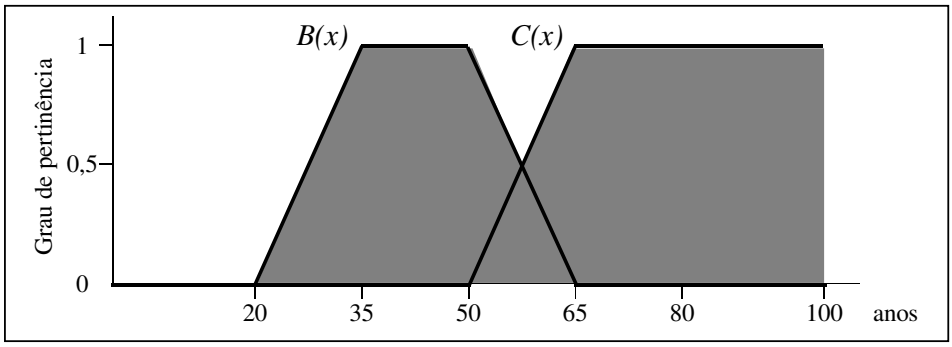


Fig. 2.13: Reunião dos conjuntos nebulosos e "meia-idade" "velho"

2.3.3. Cortes α de conjuntos nebulosos

Os conjuntos nebulosos podem ter várias representações mas, em cada uma delas, a cada elemento *x* do conjunto universo *X* é sempre atribuído um único grau de satisfação *A(x)*. Dessa forma, as diversas representações (gráficos, tabelas, listas, fórmulas matemáticas, ou coordenadas em um cubo de *n* dimensões) podem ser vistas como representações do mesmo tipo. (Klir, 1997)

Pode-se, ainda, representar conjuntos através de relações numéricas no intervalo [0,1] com conjuntos **abruptos**. Considere-se um conjunto nebuloso *E* que expresse o conceito de "livro caro" num dado contexto (figura 2.14).

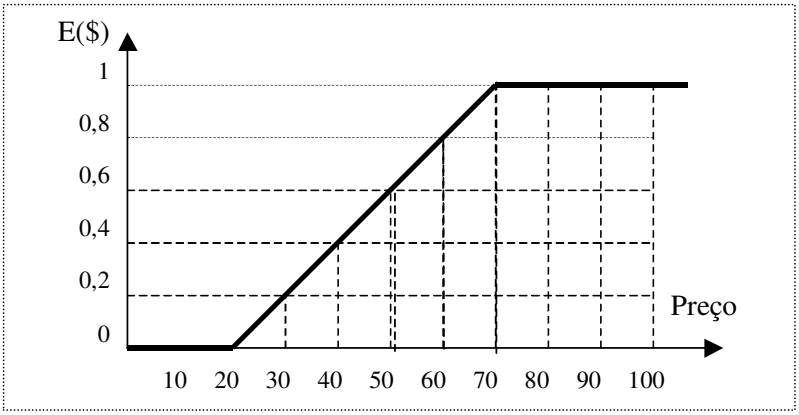


Fig. 2.14: Representação gráfica de "livro caro"

Tome-se uma faixa de valores entre \$0 (livros grátis) e \$100. Pode-se também considerar uma versão discreta da função de pertinência, na qual:

$X = \{0, 10, 20, ..., 100\}$ e

$E = \{(0;0), (0;20), (0.2;30), (0.4;40), (0.6;50), (0.8;60), (1;70), (1;80), (1;90), (1;100)\}$,

onde E representa o conjunto dos pares ordenados cujo primeiro elemento é o grau de pertinência do preço representado pelo segundo elemento do par. Este conjunto pode ser representado na forma:

$E = 0/0 + 0/20 + 0.2/30 + 0.4/40 + 0.6/50 + 0.8/60 + 1/70 + 1/80 + 1/90 + 1/100.$

Uma maneira particularmente importante de restringir os graus de satisfação é considerar apenas os elementos cuja pertinência é maior ou igual a certo valor escolhido α do intervalo [0,1]. Quando esta restrição é aplicada a um conjunto nebuloso A obtem-se um subconjunto **abrupto** ${}^{\alpha}A$ do conjunto universo X chamado **corte α** (alpha-cut) de A, cuja notação é ${}^{\alpha}A = \{x \in X \mid A(x) \geq \alpha\}$, para qualquer $\alpha \in [0,1]$. Esta equação mostra que o corte α do um conjunto nebuloso A é um conjunto abrupto ${}^{\alpha}A$ que contém todos os elementos do conjunto universo X cujos graus de satisfação em A são maiores ou iguais ao valor especificado para α.

Para o conjunto nebuloso E (figura 2.14), alguns exemplos de cortes α são: ${}^{0}E$ = [0,100], ${}^{0.2}E$ =[30,100], ${}^{0,5}E$ =[45,100], ${}^{0,9}E$ =[65,100], ${}^{1}E$ =[45,100]. Para a versão discreta de E, os exemplos correspondentes são: ${}^{0}E$ = {0,10,...100}, ${}^{0.2}E$ = {30,40,...,100}, ${}^{0,5}E$ = {50,60,...,100}, ${}^{0,9}E$ = {70,80...,100}, ${}^{1}E$ = {45,100}.

Note-se que aumentando o valor de α o valor do corte α não aumenta. Ele tanto pode diminuir como permanecer o mesmo. Esta é uma propriedade geral dos cortes α. Se considerarmos os diversos valores de α em ordem crescente, como 0,1, 0,2, 0,3, e assim por diante, os cortes α serão associados por inclusão de conjuntos. Assim, se $\alpha_1 < \alpha_2$, então ${}^{\alpha 1}A \supseteq {}^{\alpha 2}A$, o que significa que: ${}^{\alpha 1}A \cap {}^{\alpha 2}A = {}^{\alpha 2}A$ e ${}^{\alpha 1}A \cup {}^{\alpha 2}A = {}^{\alpha 1}A$.

É de se notar que os diversos cortes α vão formando uma família de conjuntos abruptos aninhados, conforme figura 2.15, a seguir:

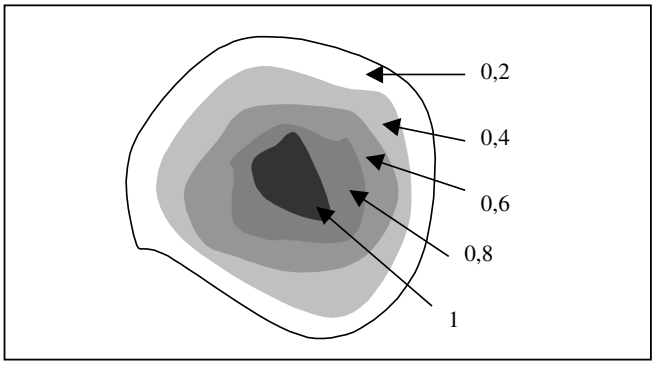


Fig. 2.15: Estrutura aninhada de cortes α.

Considerando, agora, o conjunto abrupto que contém todos os elementos do conjunto universo cujos graus de satisfação em dado conjunto são maiores que α, denotando-o por α+A,. Ele é chamado de corte α forte e definido como:

$$^{\alpha+}A = \{ x \in X \mid A(x) > \alpha \}$$

Valem para este corte todas as observações feitas para o corte α, com a única diferença de que contempla apenas a faixa do gráfico superior à cota α pelo que, por exemplo, ${}^{0+}E$ = [20,100], onde não aparece o intervalo [0,20) do corte α ${}^{0}E$ = [0,100], em razão de sua definição.

2.3.4. Lógica nebulosa

> *Lógica nebulosa fornece uma forma de inferência que permite ao raciocínio humano aproximado ser aplicado a sistemas baseados em conhecimento. A teoria da lógica nebulosa fornece meios matemáticos para manipular incertezas associadas aos processos cognitivos humanos, como pensar e argumentar. As abordagens convencionais para representação de conhecimento não permitem representar os conceitos nebulosos. Como conseqüência, as abordagens baseadas em lógica de primeira ordem e teoria de probabilidade clássica não fornecem uma base conceitual apropriada para lidar com a representação do conhecimento de senso comum, pois tal conhecimento é, por natureza, tanto lexicalmente impreciso como indeterminado. (Fuller, 1999)*

A lógica preocupa-se com os princípios formais que suportam as leis do raciocínio e serve da base para várias áreas da engenharia e da computação. Um sistema de lógica bi-valorada lida com proposições cujos valores podem ser verdadeiros ou falsos. Em sistemas lógicos multivalorados as proposições podem ser verdadeiras ou falsas e ainda assumir valores de verdade intermediários. Por seu lado, a lógica nebulosa ultrapassa os limites da lógica multivalorada, admitindo valores relativos a conjuntos difusos do intervalo unitário. A verdade pode ser vista como um valor lingüístico. Assim, a lógica nebulosa preocupa-se com os princípios que regem o raciocínio aproximado. (Pedricz, 1998)

Segundo Fuller, 1999, o desenvolvimento da lógica nebulosa foi incentivado em grande parte pela necessidade de um sistema conceitual que pudesse lidar com temas como incerteza e imprecisão léxica. Algumas das suas características básicas são:

- Na lógica nebulosa, raciocínio exato é visto como um caso particular do raciocínio aproximado.
- Na lógica nebulosa tudo é uma questão de grau.
- Na lógica nebulosa o conhecimento é visto como uma coleção flexível, ou seja, conceituação nebulosa sobre uma coleção de variáveis.
- A inferência é vista como um processo de propagação de conceitos flexíveis.
- Qualquer sistema lógico pode ser encarado como nebuloso.

Há duas características importantes dos sistemas nebulosos que lhes permitem melhor desempenho em aplicações específicas:

- Sistemas nebulosos são adequados para raciocínio incerto ou aproximado, especialmente em sistemas com modelos matemáticos de difícil definição.
- Lógica nebulosa permite tomar decisões a partir de valores determinados com base em informações incompletas ou incertas.

Na lógica clássica, o valor de uma proposição é verdadeiro ou falso, não podendo ter um terceiro valor (princípio do meio excluído) nem ser verdadeira ou falsa ao mesmo temo (princípio da não contradição). Há, contudo, proposições cujo valor verdade é indeterminado ou vago e é necessário considerar estruturas lógicas que tratem com a incerteza de verdade, a fim de interpretar valores compreendidos entre verdadeiro (1) e falso (0). A partir da década de 1930 foram desenvolvidas várias lógicas capazes de expressar a veracidade não apenas através dos valores 1 e 0, mas através de *n* valores. A mais comum é a lógica **trivalorada** que, além dos valores 0 e 1, admite um valor intermediário (1/2), para representar a indeterminação. Isto derruba os dois princípios básicos da lógica clássica (meio excluído e não contradição). (Klir, 1997)

Os estudiosos das lógicas trivaloradas, dentre eles Bochvar, Lukaziewicz e Kleene, têm pontos de vista diferentes sobre o valor-verdade de sentenças compostas com valor-verdade indeterminado e não concordam inteiramente sobre o resultado derivado do uso dos conectivos lógicos, sendo mais estritos ou mais liberais quando o valor ½ ocorre em algumas proposições. Quer concordem ou não sobre certos pontos, o certo é que estabeleceram formas de raciocinar sobre verdades parciais, segundo certas intuições básicas ou objetivos específicos. Aqui reside, em especial, a necessidade da lógica trivalorada, como ponto de partida para as lógicas n-valoradas e para as lógicas contínuas, mais adequadas à captura do valor de verdade de proposições com componentes difusos. (Klir, 1997)

Como, intuitivamente, é mais simples raciocinar com quantidades menores de elementos, na vida diária é comum a utilização de proposições cujo valor verdade pode ser estabelecido com o uso da lógica trivalorada, como, por exemplo:

- O arrazoado da defesa estava repleto de ***meias***-verdades.

- Dirigir embriagado é uma atitude ***meio*** perigosa.
- Hoje estou ***meio*** cansado.
- O diretor, na próxima semana, andará ***meio*** ocupado.

Certamente, com mais freqüência, usamos as expressões bastante, pouco, um tanto, ligeiramente, muito, etc., mais relacionadas com lógicas multivaloradas ou contínuas. Entretanto, a simplificação para uma lógica trivalorada é a mais usual.

Segundo Klir, 1997, a lógica nebulosa pode ser vista como "um sistema para raciocinar de modo aproximado, em vez de exato." E ainda: a lógica nebulosa "usa conceitos, princípios e métodos desenvolvidos na teoria dos conjuntos nebulosos para formular as várias maneiras de emitir argumentos aproximados. Para utilizar os recursos da teoria dos conjuntos nebulosos para raciocínio aproximado, é necessário estabelecer uma correlação entre os graus de pertinência nos conjuntos nebulosos e graus de verdade das proposições nebulosas".

Seja A um conjunto nebuloso. O grau de pertinência $A(x)$ para um determinado elemento x do conjunto universo X pode ser interpretado como o grau de verdade da proposição nebulosa "x é um elemento de A". Reciprocamente, dada uma proposição "x é F", onde x pertence ao conjunto X e F é uma expressão lingüística nebulosa (como baixo, alto, muito distante, extremamente lento etc.), seu grau de verdade pode ser interpretado como o grau de pertinência $A(x)$ pelo qual um conjunto nebuloso A, caracterizado pela expressão F é definido num determinado contexto. Dessa forma, operações de negação, conjunção e disjunção entre proposições nebulosas são definidas exatamente como as operações de complementação, interseção e reunião de conjuntos nebulosos. Esta correspondência permitirá o desenvolvimento de conceitos adicionais necessários à lógica nebulosa, tais como qualificadores nebulosos, quantificadores nebulosos e probabilidades nebulosas.

CAPÍTULO III

A TRIDIMENSIONALIDADE DA MOLÉCULA

Todo o conteúdo deste capítulo, texto e ilustrações, acha-se baseado no Cap. III do vol. 4 da obra "Química orgânica" (Feltre, 1991)

3.1. Representação da molécula

A molécula de um composto pode ser representada através de fórmulas e de modelos. Entre as fórmulas mais usadas tem-se:

- **Fórmula molecular:** fornece o nome e o número efetivo de cada átomo presente na molécula. Assim, para representar o ácido acético, por exemplo, teríamos a fórmula molecular $C_2H_4O_2$, indicando que é composto de dois átomos de carbono, quatro de hidrogênio e 2 de oxigênio. Essa fórmula, entretanto, não nos diz nada sobre a maneira como cada átomos se liga aos outros.

- **Fórmula estrutural:** Este tipo de fórmula foi **estabelecido** ainda em 1859, pelo químico alemão August Kekulé que propôs ter o átomo de carbono 4 ligações, representadas por quatro traços partindo do símbolo do carbono ("C"), situados num mesmo plano. Esse tipo de representação, com algumas adaptações, é utilizado até hoje e mostra as ligações químicas existentes entre os átomos. O mesmo acido acético acima citado, através deste tipo de fórmula, seria representado como se vê na figura 3.1, a seguir.

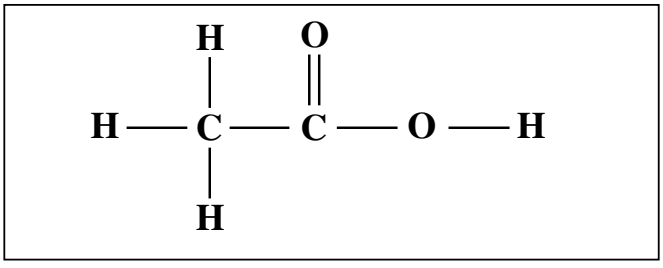


Fig. 3.1 : Fórmula estrutural do ácido acético

- **Modelo de Le Bel e Van't Hoff:** O cientista holandês, Jacobus H. Van't Hoff e o francês Joseph A. Le Bel, em 1974, de forma independente, procuraram compatibilizar a representação de Kekulé com fatos reais, propondo que as quatro valências do carbono possuíam uma disposição espacial em que o carbono ocupa o centro de um tetraedro regular, com suas quatro valências dirigindo-se para os vértices, pois só assim se explicavam certas constatações experimentais (figura 3.2).

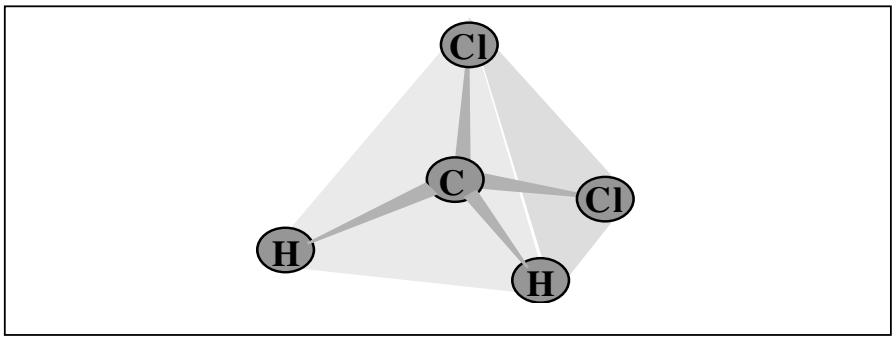


Fig. 3.2: Modelo de Le Bel e Van't Hoff (tetraedro)
Estrutura espacial do diclorometano

Há diversos outros modelos que representam a disposição espacial dos átomos de um composto: Os mais importantes são:

- **Modelo bow-tie:** É semelhante à formula estrutural da figura 3.1 e é mostrado na figura 3.3. Aqui já se tem algumas informações de estereoquímica, ou seja, sobre a disposição espacial dos átomos. Esse modelo é amplamente usado e se interpreta da seguinte forma:

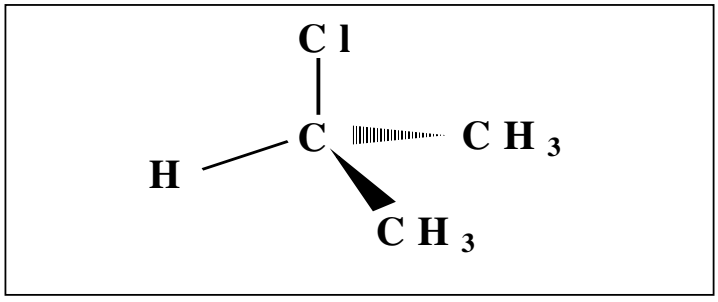


Fig. 3.3: 2-cloro-propano no modelo *bow-tie*

a) Os traços finos indicam ligações que estão no mesmo plano (plano do papel), perpendicular ao raio de visão do observador.

b) As cunhas em negrito indicam átomos que estão acima do plano (mais próximos do observador). São chamadas de ligações α (alfa).

c) Finalmente, as cunhas tracejadas, chamadas de ligações β (beta), indicam átomos que estão abaixo do plano.

Evidentemente, dependendo da posição do observador em relação à molécula, uma ligação α pode se transformar em uma β, por exemplo. Mais precisamente, qualquer um dos três tipos de ligações pode se transformar em qualquer dos outros.

- **Modelo de Stuart:** Aqui, a representação se faz através de esferas adjacentes, representando cada uma delas um átomo do composto. A figura 3.4 mostra a representação do C_4H_{10} (n-Butano) no modelo de Stuart. Este é o modelo que mais se aproxima da realidade.

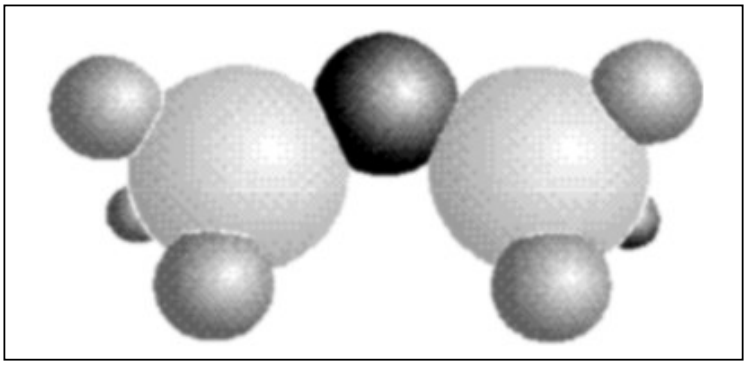


Fig. 3.4: n-Butano representado no modelo de Stuart

- **Modelo pau e bola:** É um modelo didático, que nos permite notar com bastante clareza a natureza tridimensional das moléculas. A figura 3.5 mostra um exemplo deste modelo, onde aparece o diclorometano (CH_2Cl_2).

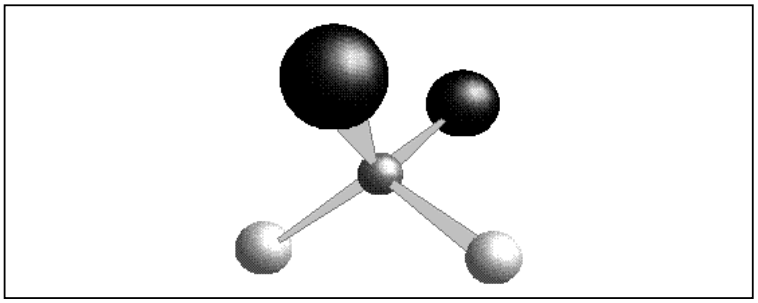


Fig. 3.5: Modelo pau e bola representando o diclorometano

Na figura 3.6 vê-se o mesmo diclorometano acima e a projeção de seus átomos sobre o plano α. Pode-se notar que essa projeção nada mais é do que a fórmula estrutural de Kekulé.

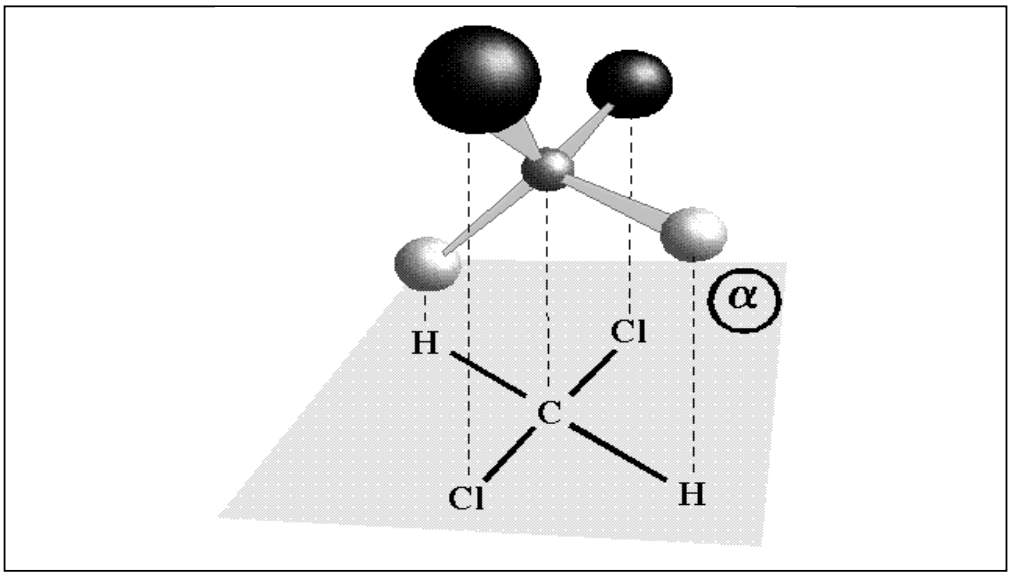


Fig. 3.6: O diclorometano e a projeção de seus átomos no plano α

3.2. Isomeria

Alguns compostos possuem ***mesma fórmula molecular*** (mesmo número de átomos de cada elemento) e, no entanto, tem ***disposições espaciais diferentes***, o que lhes dá propriedades diferentes. Assim, são compostos diferentes, embora seus componentes sejam os mesmos. Compostos com essa característica são chamados de *isômeros*.

Há duas espécies de isomeria: a isomeria plana e a isomeria espacial.

- **Isomeria plana:** As fórmulas estruturais, como vimos acima, podem ser consideradas como a "projeção" dos elementos de um composto no plano. Quando dois compostos, embora com os mesmos componentes, apresentam fórmulas estruturais diferentes, podemos notar no próprio plano a sua diferença estrutural. Nesse caso temos um caso de isomeria plana. Por

exemplo: O etanol e o éter dimetílico tem a mesma fórmula molecular, ou seja, C_2H_6O. Entretanto, como podemos ver na figura 3.7, no etanol o átomo de oxigênio está ligado a ambos os átomos de carbono, enquanto no éter dimetílico ele está ligado apenas a um dos átomos de carbono.

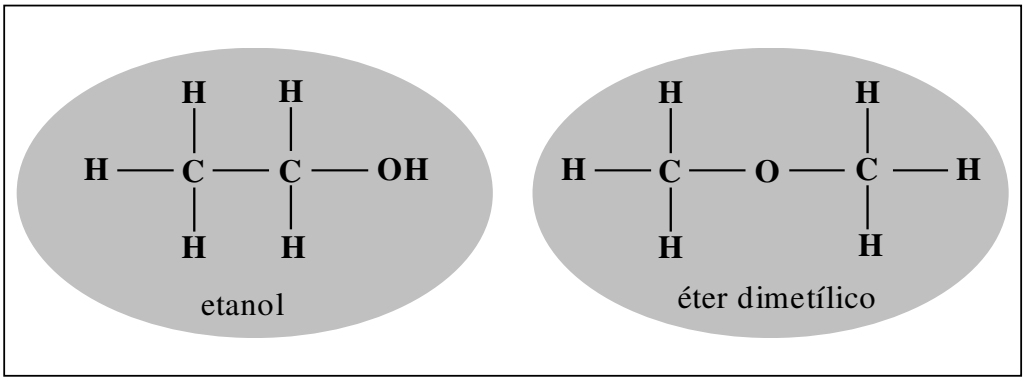


Fig. 3.7: Isomeria plana

- **Isomeria espacial:** Quando dois compostos com os mesmos componentes, apresentam fórmulas estruturais iguais, mas estrutura espacial diferente, a sua diferença estrutural só pode ser notada no espaço. Nesse caso temos um caso de **isomeria espacial**. Isto é o que acontece, por exemplo, com o 1,2,dicloro-ciclobutano (figura 3.8), onde os dois átomos de cloro, em relação ao plano que contém os átomos de carbono, podem estar no mesmo semi-espaço ou em semi-espaços opostos. No primeiro caso dizemos que temos uma forma **cis** e, no segundo, uma forma **trans**.

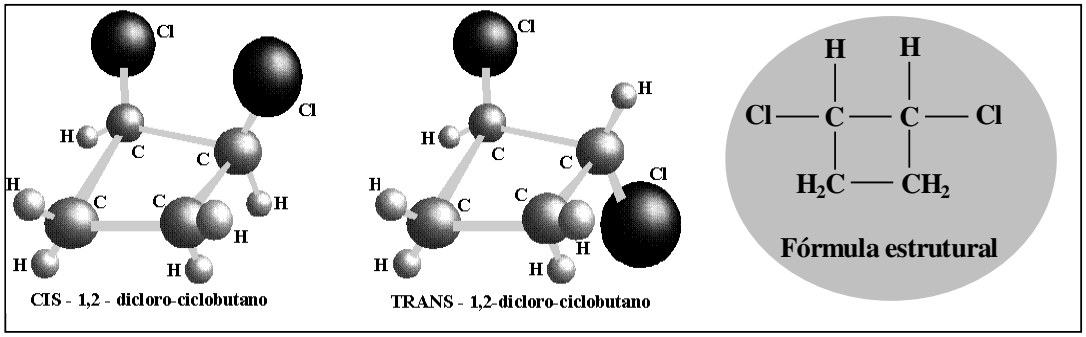


Fig. 3.8: Isomeria espacial

CAPÍTULO IV

LOCALIZAÇÃO PLANA

4.1. Matriz de Conectividade

Dentre os campos do arquivo de moléculas há um que define o **tipo de átomo**. Sobre ele é necessário tecer algumas considerações, em razão da sua importância na heurística que será adotada para reduzir o trabalho computacional na identificação da subestrutura dentro da molécula.

Além da matriz topológica descrita no item 1.7 (figura 1.5), relativa à substância exemplificada (figura 4.1), é usada na química uma outra, que será aqui denominada **matriz de conectividade**, cujo método de construção se deve a Morgan (*apud* Gastmans, 1990). Para sua elaboração, é preciso que se defina um código de agrupamentos atômicos. O SISTEMAT, por tratar apenas de substâncias orgânicas naturais, usa uma tabela de códigos com 35 tipos, contemplando apenas os átomos encontrados nesse tipo de substâncias (carbono, hidrogênio, oxigênio, nitrogênio, flúor, cloro, bromo, iodo, enxofre e fósforo) (Gastmans, 1990). Esses códigos aparecem na tabela 4.1 a seguir e se baseiam no elemento químico e nos tipos de ligações que ele apresenta na molécula.

Tabela 4.1: Códigos de agrupamentos atômicos

Agrupamento	Código	Agrupamento	Código	Agrupamento	Código
– CH_3	01	O =	13	– Br	25
– CH_2 –	02	HO –	14	– I	26
$\rangle$CH –	03	– O –	15	S ≡	27
$\rangle$C$\langle$	04	– NH_2	16	HS –	28
= CH_2	05	$\rangle$NH	17	– S –	29
$\rangle\!\!\rangle$CH	06	$\rangle$N –	18	– S =	30
$\rangle$C =	07	= NH	19	= S =	31
≡CH	08	– N =	20	– PH2	32
≡C	09	N ≡	21	$\rangle$PH	33
HC	10	N_{ar}	22	$\rangle$P –	34
C_{ar}	11	– F	23	– P –	35
= C =	12	– Cl	24		

A matriz de conectividade tem 5 linhas e tantas colunas quantos sejam os átomos da molécula. Exemplificando: na figura 4.1 a matriz de conectividade é a que aparece na tabela 4.2.

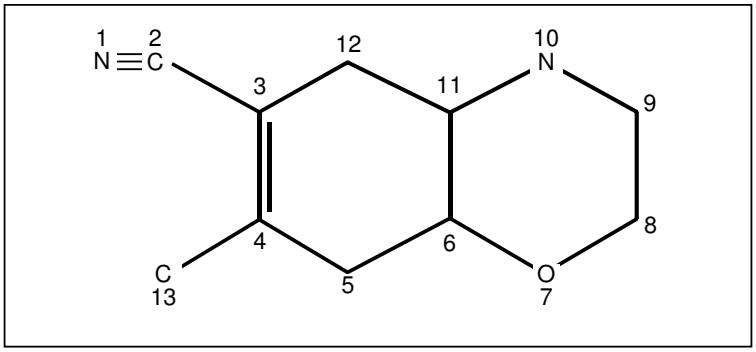


Fig. 4.1 – Molécula de exemplo

Tabela 4.2: Matriz de conectividade da substância da fig. 4.1

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

No topo de cada coluna está o tipo de átomo, abaixo do qual aparecem os tipos dos átomos a ele ligados, de acordo com os códigos da tabela 4.1. Preenche-se com zeros as ligações que não existem. Dessa forma, a sexta coluna corresponde ao $\mathbf{C_6}$ (código 03, na primeira linha) e seus vizinhos ($\mathbf{C_5}$: cód. 02, $\mathbf{C_{11}}$: cód. 03 e $\mathbf{O_7}$: cód. 15).

Esse tipo de representação facilita a identificação da subestrutura dentro da molécula, pois não se trabalha com os elementos químicos e suas ligações, mas com um tipo que, por si, já esclarece tanto o elemento químico quanto a que átomos e com quais tipos de ligações está conectado. Assim, não se buscará, por exemplo, um carbono com estes ou aqueles ligantes, mas um átomo do tipo 01 a 12. A busca, dessa forma, fica muito mais seletiva. Note-se, por exemplo, que, na molécula da figura 4.1 existem 10 átomos de carbono. Entretanto, há apenas duas colunas da tabela 4.2 com os mesmos dados, que são as de número 6 e 12, relativas aos carbonos C_5 e C_{12}, respectivamente.

Durante os trabalhos para elaboração desta dissertação foi feito um programa para conversão dos dados existentes no arquivo HyperChem®, que grava os dados convertidos na tabela estruturada cujo layout aparece na figura 3.8, determinando, automaticamente, os tipos de átomos, conforme especificado na tabela 4.1. O fluxograma para essa determinação de tipo é mostrado no Anexo I.

4.2. Reconhecimento plano

A presente fase é chamada de reconhecimento plano porque aqui os objetos são considerados como se não tivessem conformação espacial, levando-se em conta apenas as conexões que existem entre seus componentes. Conforme definido no item 1.5, o problema de reconhecimento é dividido em algoritmos que solucionem problemas de decisão, localização e otimização.

4.2.1. Decisão

Em fase de decisão, monta-se a matriz de conectividade tanto da molécula quanto da subestrutura a ser localizada, conforme o método descrito no item 4.1. A molécula deve possuir todas as colunas existentes na subestrutura, em número igual ou superior. Caso alguma das colunas da subestrutura não esteja presente na molécula, conclui-se que aquela não está presente nesta e encerra-se a busca, caso contrário, prossegue-se com a fase de localização.

Considere-se, por exemplo, a subestrutura da figura 4.2, a seguir.

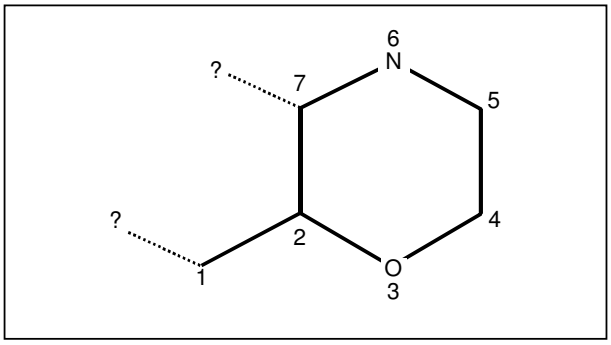


Fig. 4.2: Subestrutura

Elaborando-se sua matriz de conectividade, chegamos aos dados expressos na tabela 4.3, a seguir:

Tabela 4.3: Matriz de conectividade da subestrutura da fig. 4.2

02	03	15	02	02	17	03
03	15	03	15	17	03	17
-1	03	02	02	02	02	02
00	02	00	00	00	00	-1
00	00	00	00	00	00	00

Aqui, os átomos estão sendo representados por um conjunto de 5 elementos, constituído pelo tipo do átomo e dos seus vizinhos. É possível ver que cada uma das colunas está presente na matriz de conectividade da tabela 4.2 estando, inclusive, na mesma ordem (colunas 5 a 11).

A primeira e a última colunas da tabela 4.3 contém um identificador "***–1***", que indica não estarem completas as ligações dos átomos correspondentes. Por exemplo, o átomo da última coluna é do tipo 3 e, conforme se vê na tabela 4.1, ele deve ter três ligações diferentes de hidrogênio. Na figura 4.2, entretanto, só se pode ver duas. A ligação que falta está sendo indicada por linhas pontilhadas com um ponto de interrogação.

O identificador "**–1**", aqui funciona como um "coringa", de forma que ao verificar-se a correspondência com os átomos da molécula, pode-se considerar como correspondente qualquer coluna que tenha o mesmo cabeçalho (linha 1) e os números não negativos. Inicialmente havia-se pensado em um identificador "99", pois o número de tipos de átomos só chega a 35. Resolveu-se, entretanto, que, se fosse necessário ampliar o número de tipos no futuro, com um identificador negativo nada precisaria ser alterado nas rotinas até então implementadas.

Como as matrizes de conectividade são montadas apenas na memória do computador, em tempo de execução, deve-se estabelecer uma forma de reconhecer as ligações incompletas. Isto pode ser feito utilizando os dados da tabela 4.4, a seguir:

Tabela 4.4: Número de ligantes diferentes de hidrogênio, por tipo de átomo.

Código	Ligantes	Código	Ligantes	Código	Ligantes
01	1	13	1	25	1
02	2	14	1	26	1
03	3	15	2	27	1
04	4	16	1	28	1
05	1	17	2	29	2
06	2	18	3	30	2
07	3	19	1	31	2
08	1	20	2	32	1
09	2	21	1	33	2
10	1	22		34	3
11		23	1	35	2
12	2	24	1		

Os tipos 11 e 22 não estão sendo considerados, pois se trata de elementos aromáticos e, para o presente propósito, esse tipo de ligação pode ser considerado como uma ligação simples.

Até aqui, o que se fez foi uma busca em amplitude, verificando uma condição essencial para o prosseguimento da busca.

4.2.2. Localização

A fase da localização é um pouco mais complexa. É preciso verificar se os átomos estão ligados na mesma ordem, ou seja, saber quais átomos da molécula correspondem aos átomos da subestrutura, o que será necessário para a fase seguinte (otimização), que fará o reconhecimento espacial. Na fase atual estão envolvidas técnicas de inteligência artificial, especialmente a subida de encosta.

Começa-se montando a árvore de conectividade da subestrutura, iniciando por uma coluna que, se possível, tenha um representante único na estrutura. Desejável, ainda, é que o átomo dessa coluna tenha o maior número de ligantes (números que aparecem da segunda à quinta linha). Observe-se que os átomos de hidrogênio não são considerados, pelas razões expostas ao final do item 3.3.

No segundo nível da árvore, coloca-se os átomos que estão ligados ao primeiro e busca-se uma correspondência na molécula. Exemplificando graficamente (figura 4.3), pode-se iniciar pelo átomo número 2 da subestrutura, por ter o maior número de ligantes conhecidos (3 neste exemplo) e por existir na molécula apenas um átomo com todos os elementos da coluna iguais ao do átomo 2 da subestrutura, especificamente, o de número 6 (sexta coluna). Assim, pode-se montar com certeza uma parte da árvore dos dois objetos, até o segundo nível, verificando as correspondências entre os átomos.

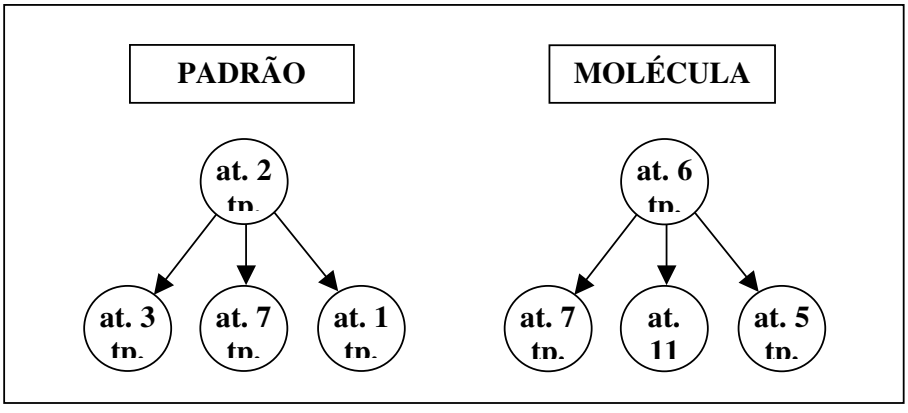


Fig. 4.3: Primeira fase da busca

Ao átomo 2 da subestrutura só pode corresponder ao átomo 6 da molécula, cujos ligantes (15, 03 e 02) pode-se ver na coluna 6 da matriz de conectividade da molécula. Entretanto, como se pode ter certeza de que o átomo de tipo 03 colocado na árvore é mesmo o átomo de número 11 e não o de número 6, que também é do tipo 3? Simplesmente verificando o arquivo de moléculas, onde constam as conectividades de todos os átomos (ver figura 4.4).

Assim, para esta busca, não bastam as informações das matrizes de conectividade, mas é preciso trabalhar paralelamente com os dados dos arquivos de moléculas e de subestruturas.

Nessa primeira fase, portanto, já se pode inferir que, até este ponto, existe a seguinte correspondência entre os átomos:

Tabela 4.5: Correspondências entre os átomos da subestrutura e da molécula

Subestrutura	Molécula
2	6
3	7
7	11
1	5

Para evitar pesquisa circular, os átomos cuja correlação já ficou definida são excluídos da busca. Prossegue-se verificando na subestrutura as ligações dos átomos do segundo nível, buscando-se, ao mesmo tempo, sua correlação na molécula (figura 4.4).

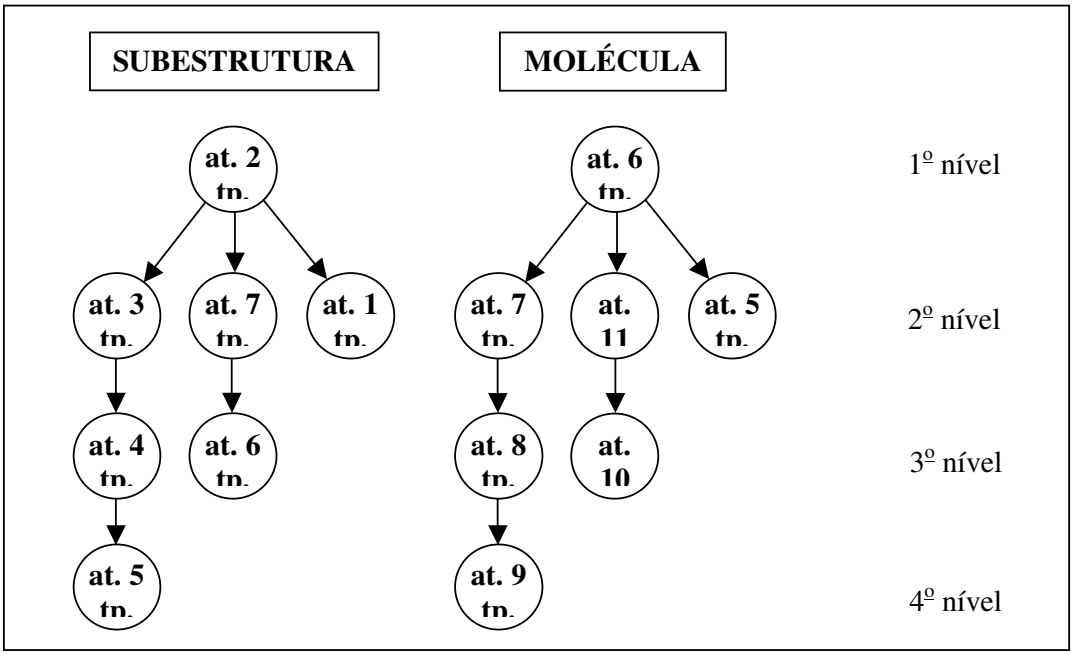


Fig. 4.4: Continuação da busca

A subestrutura a ser localizada possui uma estrutura bem definida e sua árvore pode ser construída com facilidade. O trabalho, basicamente, consiste em verificar na estrutura, com base nas conexões que cada átomo tem com os outros, quais os elementos correspondentes ao da subestrutura.

Neste caso, a busca rapidamente chega ao fim e a subestrutura foi encontrada dentro da molécula. A tabela de correspondência entre os átomos é a seguinte:

Tabela 4.6: Correspondência final entre a subestrutura e molécula

Átomos da subestrutura	Átomos da molécula
2	6
3	7
7	11
1	5
4	8
6	10
5	9

Se qualquer ramo da árvore da estrutura não puder ter continuação, por não haver entre seus ligantes um átomo do mesmo tipo que o átomo correspondente da subestrutura, toda a árvore fica prejudicada e, a menos que se possa iniciar nova árvore com outro átomo, a subestrutura não estará presente na molécula.

A altura da árvore será, no máximo, igual ao número de átomos da subestrutura, o que ocorre quando este tiver uma conformação "linear", ou seja, cada átomo está ligado a outros 2, no máximo.

Em alguns casos, a busca pode se tornar bem mais complexa, especialmente se o formato da subestrutura for "linear" e o da estrutura for um tipo de "grade" homogênea. Nesses casos, podem surgir muitas alternativas, com as conseqüentes ramificações possíveis na árvore da estrutura, as quais, depois, terão que ser combinadas. Isto pode levar a um número muito grande de verificações espaciais posteriores, sem contar que a "pilha" necessária para manter as diversas ramificações pode ser tornar muito grande e difícil de ser manuseada.

Exemplificando (figura 4.5):

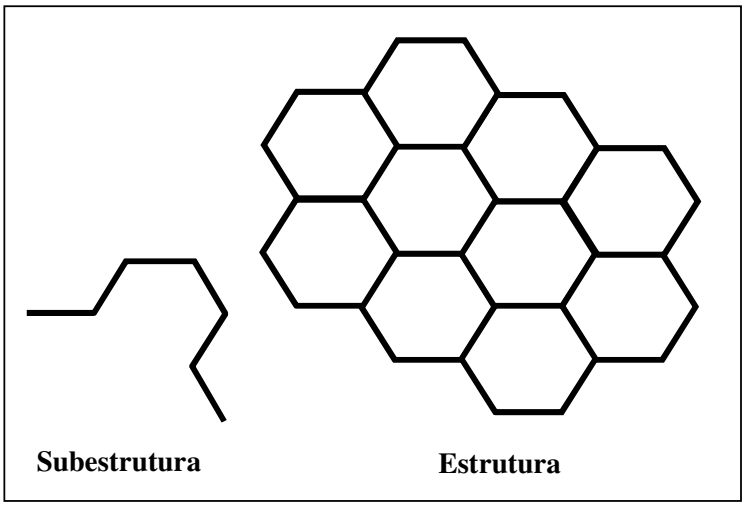


Fig 4.5: Subestrutura linear em estrutura tipo "grade"

A figura fala por si mesma. É fácil ver que é grande o número de posições diferentes em a subestrutura pode se encontrar dentro da estrutura. Isto, entretanto, não traz grandes problemas, pois o objetivo é saber se a subestrutura se acha contido na estrutura e, uma vez encontrada uma correlação, o trabalho estará terminado. O problema mais grave é que, como não se está considerando a estereoquímica, só se leva em conta os tipos de átomos e de ligações e, por isso, não se cogita sobre a "direção" em que procurar. Assim, neste exemplo, depois de estabelecida uma correlação inicial, só se sabe que o átomo seguinte é de um determinado tipo e, como quase todos os átomos são do mesmo tipo, o seguinte poderá ser quase que cada um dos vizinhos. Mesmo se considerando subestruturas de pequeno porte como a presente, o número de possibilidades se tornará muito elevado.

Se houver vários candidatos no primeiro nível, inicia-se por um deles e passa-se para o segundo nível. Se houver vários candidatos no segundo nível, pode-se estar numa situação como a descrita no parágrafo anterior. Assim, para evitar a explosão computacional e a difícil manutenção de "pilhas", é melhor ter certeza da "direção" correta, para não tomar um caminho errado que poderia levar a outros. E isto só se consegue utilizando uma pesquisa espacial. Então, considera-se correto o candidato do primeiro nível e o primeiro do segundo, montando com eles uma subestrutura que será submetida a um reconhecimento espacial (ver capítulo seguinte). Se não houver

coincidência espacial, descarta-se o último candidato e passa-se ao seguinte, até encontrar o correto ou chegar ao último, cuja verificação espacial não se fará, pois, por eliminação, deve ser o correto, nesse nível. Embora isto pareça aumentar a carga computacional, evitará árvores cheias de ramificações, e verificações espaciais posteriores cujo número seria a combinação de todas as alternativas possíveis para as diversas árvores. Prossegue-se na busca e, não sendo encontrada configuração plana igual, retorna-se ao primeiro nível e tenta-se com o candidato seguinte, até encontrar uma coincidência ou não haver mais candidatos. Repete-se este procedimento sempre que, em dois níveis consecutivos, aparecerem vários candidatos para a seqüência.

Diversas experiências feitas mostram que não é comum aparecerem muitas alternativas, em razão da heurística adotada (método de Morgan/Gastmans – item 6.2). De fato, quando se define um átomo através de sua coluna na matriz de conectividade, não se está atingindo apenas o átomo em questão e seus vizinhos imediatos, mas até o segundo átomo ao seu redor. Isto se deve ao fato de que no tipo do vizinho está implícito o tipo de substância que constitui o "vizinho do vizinho".

Encontra-se maiores problemas, como foi dito acima, quando a subestrutura é "linear". Isto, na prática, não deverá ser comum, pois a maior utilidade do presente reconhecimento de padrões deverá ocorrer na busca de esqueletos químicos dentro das moléculas. Estes esqueletos (p. ex.: germacrolídeo, melampolídeo, anéis lactônicos etc.), são subestruturas comuns a muitas substâncias, as quais variam em decorrência dos substituintes agregados ao esqueleto básico. Podem ser também conjuntos de moléculas que constituem substituintes comumente agregados aos esqueletos básicos (glicose, acetato, angelato etc.). E estes padrões, com raras exceções, não são lineares (figura 4.6).

O fluxograma pra a localização plana pode ser visto no anexo II.

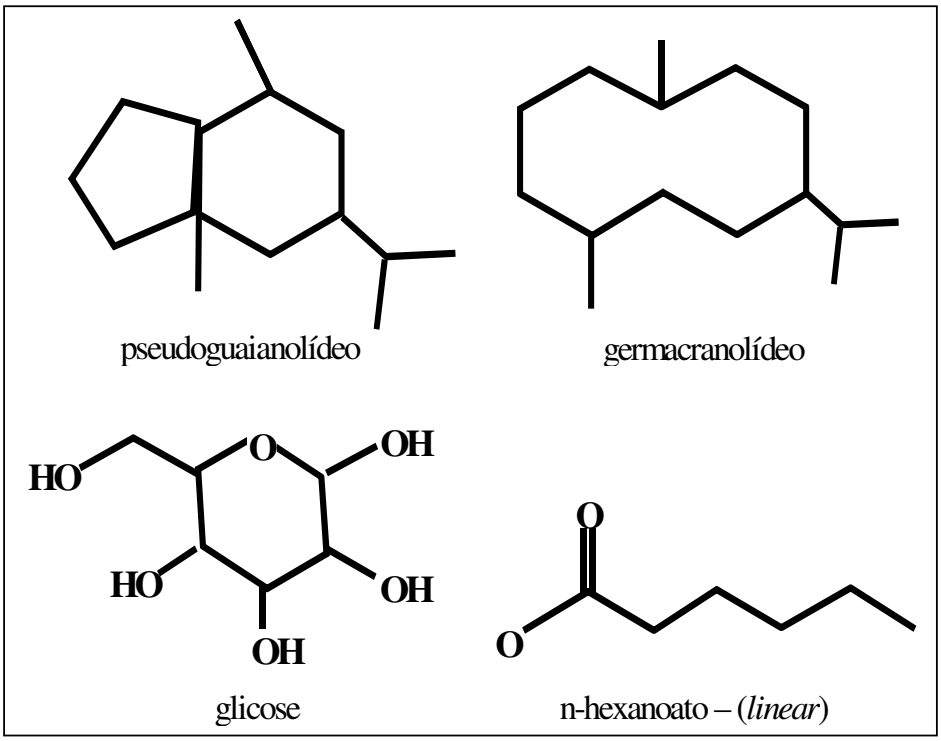


Fig. 4.6: Exemplos de subestruturas

CAPÍTULO V

RECONHECIMENTO ESPACIAL

5.1 Minimização das distâncias médias

O método que será apresentado a seguir, com base nas coordenadas cartesianas de cada átomo de duas moléculas, fundamenta-se, com algumas adaptações, no trabalho de Sippl & Stegbuchner, 1991. Um de seus aspectos interessantes é que, embora robusto e computacionalmente rápido, opera sem necessidade de recursos de álgebra matricial superior. Não requerendo programas matemáticos especiais, torna-se facilmente portável para qualquer computador pessoal.

Fundamenta-se na sobreposição de duas moléculas (ou fragmentos de moléculas), mediante deslocamento espacial de uma delas para sobre a outra, buscando a minimização das distâncias médias entre os átomos correspondentes, com uso de movimentos de translação e rotação relativos aos três eixos coordenados, como segue.

Sejam x e y dois pontos no espaço, cujas coordenadas são (x_1, x_2, x_3) e (y_1, y_2, y_3), respectivamente.

A distância d entre esses pontos é dada por:

$$R(x, y) = d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}\ , \qquad (1)$$

que pode ser expressa como

$$d^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 \qquad (2)$$

Para vários pontos, a distância média entre eles seria calculada pelo somatório das distâncias entre os pontos correspondentes, dividido pelo número de pontos, como segue:

$$d_m = \frac{d_1 + d_2 + ... + d_n}{N}$$

ou ainda,

$$N \times d_m = d_1 + d_2 + \ldots + d_N \qquad (3)$$

Isto se aplica a qualquer objeto que possa ser representado por um conjunto de coordenadas espaciais, como figuras geométricas tridimensionais ou, como no presente caso, moléculas de substâncias químicas. Como exemplo, os pontos poderiam ser os vértices ABC e A'B'C' dos triângulos X e Y, respectivamente, todos relativos a um mesmo sistema de coordenadas, conforme figura 5.1 a seguir. Os objetos não têm a mesma posição, nem a mesma orientação. Quanto à dimensão nada se pode afirmar, pois, são figuras espaciais e sua representação plana pode estar distorcida. A relação entre suas dimensões só poderá ser determinada pela sobreposição dos objetos, o que pode ser conseguido por meio de movimentos de translação e/ou rotação, buscando a ***minimização da distância média*** entre os vértices correspondentes.

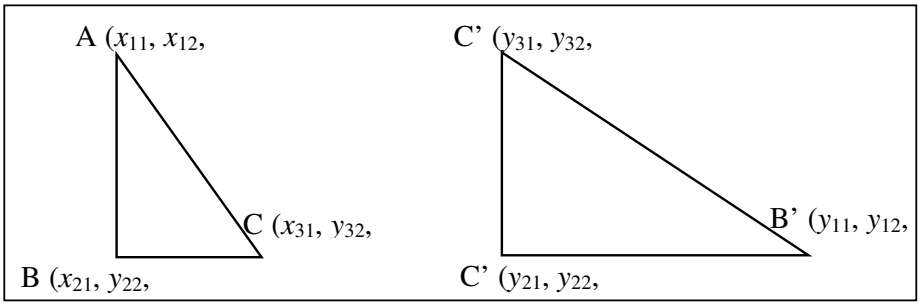


Fig. 5.1: Figuras geométricas com mesmo número de vértices

Pelas sentenças (2) e (3), a distância média entre os pontos correspondentes é:

$$N \times d_m^2 = (x_{11} - y_{11})^2 + (x_{12} - y_{12})^2 + (x_{13} - y_{23})^2 +$$
$$(x_{21} - y_{21})^2 + (x_{22} - y_{22})^2 + (x_{23} - y_{23})^2 +$$
$$\ldots +$$
$$(x_{n1} - y_{n1})^2 + (x_{n2} - y_{n2})^2 + (x_{n3} - y_{n3})^2 \qquad (4)$$

A expressão (4) pode ser representada da seguinte forma:

$$F(X, Y) = N \times d_m^2 = \sum_{i=1}^{N} \sum_{j=1}^{3} \left(x_{ij} - y_{ij}\right)^2 \qquad (5)$$

As coordenadas dos conjuntos de pontos de duas figuras *X* e *Y* quaisquer, na notação matricial, são assim representadas:

$$X = \begin{Bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} \end{Bmatrix} \qquad \text{e} \qquad Y = \begin{Bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \vdots & \vdots & \vdots \\ y_{N1} & y_{N2} & y_{N3} \end{Bmatrix} \tag{6}$$

Usando vetores, os pontos de ordem *i* de *X* ou de *Y*, tem a seguinte notação:

$$x_i = (x_{i1}, x_{i2}, x_{i3}) \quad \text{e} \quad y_i = (y_{i1}, x_{i2}, x_{i3}) \tag{7}$$

Assim, pode-se reescrever a equação 5 como:

$$N \times d_m^2 = F(X, Y) = \sum (x_i - y_i)^2 \tag{8}$$

e, fazendo

$$(X - Y)^2 = \sum (x_i - y_i)^2 \tag{9}$$

tem-se, de (8) e (9):

$$F(X, Y) = (X - Y)^2 \tag{10}$$

onde *X* e *Y* são dois conjuntos de vetores cuja expressão geral é denotada em (7), com extremidades nos pontos definidos nas matrizes descritas em (6). A expressão $(X - Y)^2$ é uma forma reduzida da dupla somatória que aparece em (5), sendo equivalente a $N \times d_m$, como se pode ver em (8).

O valor de $F(X, Y)$ depende tanto da posição quanto da orientação relativas de *X* e *Y*. Entretanto, a similaridade dos objetos representados por *X* e *Y* é independente dessas características.

Para aquilatar o grau de similaridade de *X* e *Y* é necessário a minimização do valor de $F(X, Y)$ o que corresponde, como foi dito, à tentativa de sobreposição dos dois objetos seja por movimentos de translação (minimização da diferença de posição), seja de rotação (minimização da diferença de orientação).

5.2 Minimização por translação

Para translacionar um ponto em relação a um eixo, soma-se um determinado valor à componente relativa a esse eixo. Em termos vetoriais, considerando-se a posição de um ponto definida pelo vetor y_{ij}, a nova posição y'_{ij} será dada por:

$$y'_{i1} = y_{i1} + r_1$$

$$y'_{i2} = y_{i2} + r_2 \quad \text{e}$$

$$y'_{i3} = y_{i3} + r_3 \qquad (11)$$

onde r_1, r_2 e r_3 representam as componentes, em relação aos três eixos, do vetor que define uma determinada translação para os ponto de um dado objeto.

Para minimizar a diferença de posição entre os objetos mantêm-se fixo um deles, por exemplo, *X*, e translaciona-se o outro, *Y*. Considerando-se que todos os pontos de *Y* sofrerão a mesma translação, é preciso somar ao vetor representativo de cada um dos pontos de *Y* um vetor constante r_k (*k* variando de 1 a 3), conforme definido em (11). Assim, a expressão (5) pode ser reescrita como:

$$F(X, Y) = \sum_{i=1}^{N}\sum_{j=1}^{3}\left(x_{ij} - (y_{ij} + r_j)\right)^2 \qquad (12)$$

O menor valor de *F* (*X*, *Y*) ocorre quando sua derivada é igual a zero. Assim, inicialmente, deriva-se *F* (*X*, *Y*) em relação a r_k. Observando que *k* varia de 1 a 3, elimina-se o segundo somatório e acha-se a derivada em relação a cada uma das componentes de *k*.

Usando a regra de cadeia tem-se:

$$\frac{\delta F}{\delta r_k} = \sum_{1=1}^{N} 2\left[x_{ik} - \left(y_{ik} + r_k\right)\right] . 1 = 2\left[\sum_{i=1}^{N} x_{ik} - \sum_{i=1}^{N} y_{ik} - \sum_{i=1}^{N} r_k\right]$$

Assim, a derivada de *F (X, Y)* pode ser expressa como:

$$\frac{\delta F}{\delta r_k} = -2Nr_k + \sum_{i=1}^{N} x_{ik} - \sum_{i=1}^{N} y_{ik} \qquad (13)$$

Fazendo essa derivada igual a zero (minimização), tem-se:

$$-2Nr_k + \sum_{i=1}^{N} x_{ik} - \sum_{i=1}^{N} y_{ik} = 0$$

$$r_k = \frac{1}{N} . \sum_{i=1}^{N} x_{ik} - \frac{1}{N} . \sum_{i=1}^{N} y_{ik} \text{ , para } k = 1, 2, 3 \qquad (14)$$

Observando-se a igualdade expressa em (13), pode-se notar que o segundo membro é a diferença entre as expressões $\frac{1}{N} \cdot \sum_{i=1}^{N} x_{ik}$ e $\frac{1}{N} \cdot \sum_{i=1}^{N} y_{ik}$, com k variando de 1 a 3.

Como essas expressões representam, respectivamente, os centros de gravidade de X e Y, conclui-se que a minimização da média das distâncias entre os pontos correspondentes de X e Y, por translação, consiste exatamente em mover o centro de gravidade de um dos objetos para a posição onde se acha o centro de gravidade do outro.

5.3 Minimização por rotação

Resta minimizar a distância média entre os pontos correspondentes às duas moléculas pela melhor rotação de um dos objetos sobre cada um dos eixos. Note-se que, na translação, cada vez que se desliza um objeto, paralelamente a um dos eixos, a distância média entre todos pontos correspondentes vai sucessivamente diminuindo até atingir a menor distância média que é possível atingir com um movimento deste tipo. Diz-se, então, que os movimentos de translação são independentes entre si.

Os movimentos de rotação sobre os eixos, entretanto, não são independentes, pois um deles pode afastar dois ou mais pontos que haviam sido aproximados pelo movimento anterior. Isto ocorre pelo fato de que uma rotação sobre o eixo Y, por exemplo, altera os ângulos que os vetores representativos dos pontos formam com os outros dois eixos. Entretanto, no processo que será visto a seguir, a distância ***média*** será sempre ***estritamente decrescente*** a cada movimento rotação sobre um dos eixos.

Assim, o problema será resolvido iterativamente, contando-se um ciclo do processo a cada conjunto de três rotações, uma sobre cada eixo. O processo termina quando se atinge determinado grau de precisão preestabelecido, o que será discutido ao final deste desenvolvimento.

Observe-se que a minimização da função F (X, Y), agora, não depende da posição dos objetos (cujos centros de gravidade estão sobrepostos), mas de sua

orientação. Assim, a sobreposição dos pontos correspondentes, se possível, será feita mediante rotação de um deles. Para que os centros de gravidade, já sobrepostos, não mudem de lugar quando for feita a rotação, basta fazer uma translação dos três eixos de forma que sua origem coincida com o centro de gravidade comum.

Considerando se um ponto *y* qualquer, suas coordenadas podem ser definidas pela matriz linear $\{y_1, y_2, y_3\}$. Após um movimento de rotação sobre um dos eixos, a nova posição de *y* será definida através do produto da matriz de rotação relativa a esse eixo pela matriz transposta de *y* (y^T).

Seja γ o ângulo de rotação sobre o eixo *Z*. A matriz associada à sua rotação é:

$$\mathrm{M}(\gamma) = \begin{Bmatrix} \cos\gamma & -\mathrm{sen}\,\gamma & 0 \\ \mathrm{sen}\,\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{Bmatrix}. \qquad (15)$$

Assim, depois de determinada rotação γ, a matriz que define a nova posição do ponto *y* é definida por:

$$y' = \mathrm{M}(\gamma) \times \mathrm{y}^{\mathrm{T}} \qquad (16)$$

ou ainda:

$$y' = \begin{Bmatrix} \cos\gamma & -\mathrm{sen}\,\gamma & 0 \\ \mathrm{sen}\,\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{Bmatrix} \times \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} \qquad (17)$$

De forma geral, usando (16) e a notação estabelecida em (9) e (10), pode-se definir a distância média, depois de determinada rotação, da seguinte forma:

$$N \times d_m^2 = F(\alpha,\beta,\gamma) = \sum \left[x_i - y_i^{'}(\alpha,\beta,\gamma) \right]^2$$

$$= (X - \mathrm{M}Y^{\mathrm{T}})^2 \qquad (18)$$

Para um ponto de ordem *i* de *Y*, segundo (16), as novas coordenadas, após determinada rotação sobre *Z* serão:

$$y_i^{'}(\gamma) = \mathrm{M}(\gamma)\ y_i^T, \qquad (19)$$

que, desenvolvido dá:

$$y'_{i1}(\gamma) = y_{i1}\cos\gamma - y_{12}\,\text{sen}\,\gamma$$

$$y'_{i2}(\gamma) = y_{i1}\cos\gamma - y_{12}\,\text{sen}\,\gamma$$

$$y'_{i3}(\gamma) = y_{i3} \qquad (20)$$

Para os quadrados

$$(x_i - y'_1)^2 = (x_{i1} - y'_{i1})^2 + (x_{i2} - y'_{i2})^2 + (x_{i3} - y'_{i3})^2$$

obtém-se, após substituição dos valores expressos em (19),

$$\begin{aligned}(x_i - y'_1)^2 = {} & x_{i1}^2 + y_{i1}^2\cos^2\gamma + y_{i2}^2\,\text{sen}^2\gamma - 2\,x_{i1}\,y_{i1}\cos\gamma + 2\,x_{i1}\,y_{i2}\,\text{sen}\,\gamma \\ & + x_{i2}^2 + y_{i1}^2\,\text{sen}^2\gamma + y_{i2}^2\cos^2\gamma - 2\,x_{i2}\,y_{i1}\,\text{sen}\,\gamma - 2\,x_{i2}\,y_{i2}\cos\gamma \\ & + x_{i3}^2 + y_{i3}^2 - 2\,x_{i3}\,y_{i3} \qquad (21)\end{aligned}$$

o que leva à expressão

$$\begin{aligned}(x_i - y'_1)^2 = {} & x_{i1}^2 + x_{i2}^2 + x_{i3}^2 + y_{i1}^2 + y_{i2}^2 + y_{i3}^2 - 2\,x_{i3}\,y_{i3} \\ & - 2(x_{i1}\,y_{i1} + x_{i2}\,y_{i2})\cos\gamma + 2(x_{i1}\,y_{i2} - x_{i2}\,y_{i2})\,\text{sen}\,\gamma \qquad (22)\end{aligned}$$

Substituindo esses valores em (17) e diferenciando em relação a γ, obtem-se:

$$\frac{\delta F(\gamma)}{\delta\gamma} = 2\,\text{sen}\,\gamma\sum_{i+1}^{N}(x_{i1}\,y_{i1} + x_{i2}\,y_{i2}) + 2\cos\gamma\sum_{i+1}^{N}(x_{i1} - x_{i2}\,y_{i2}) \qquad (23)$$

O valor mínimo da expressão (21), portanto, ocorre quando a diferencial que aparece em (23) for nula, o que dá, sucessivamente:

$$2\,\text{sen}\,\gamma\sum_{i=1}^{N}(x_{i1}\,y_{i1} + x_{i2}\,y_{i2}) = -2\cos\gamma\sum_{i=1}^{N}(x_{i1}\,y_{i2} - x_{i2}\,y_{i2})$$

$$\text{tg}\,\gamma\sum_{i+1}^{N}(x_{i1}\,y_{i1} + x_{i2}\,y_{i2}) = -\sum_{i+1}^{N}(x_{i1}\,y_{i2} - x_{i2}\,y_{i2})$$

$$\gamma = \arctan\frac{\sum_{i=1}^{N}x_{i2}y_{i1} - \sum x_{i1}\,y_{i2}}{\sum_{i=1}^{N}x_{i1}\,y_{i1} + \sum_{i=1}^{N}x_{i2}\,y_{i2}} \qquad (24)$$

A equação (24) fornece o ângulo que minimiza as distâncias médias entre os pontos correspondentes de *X* e *Y* mediante a rotação sobre o eixo *Z* (γ_m).

Processos semelhantes levam aos ângulos que minimizam as distâncias médias entre os pontos correspondentes de *X* e *Y* mediante a rotação sobre os eixo *X* e *Y* (α_m e β_m , respectivamente).

Os ângulos a seguir definem as melhores rotações sobre os eixos:

$$\alpha_m = \arctan \frac{\sum_{i=1}^{N} x_{i3} y_{i2} - \sum x_{i2}\, y_{i3}}{\sum_{i=1}^{N} x_{i2}\, y_{i2} + \sum_{i=1}^{N} x_{i3}\, y_{i3}} \qquad (25)$$

$$\beta_m = \arctan \frac{\sum_{i=1}^{N} x_{i3} y_{i1} - \sum x_{i1}\, y_{i3}}{\sum_{i=1}^{N} x_{i1}\, y_{i1} + \sum_{i=1}^{N} x_{i3}\, y_{i3}} \qquad (26)$$

$$\gamma_m = \arctan \frac{\sum_{i=1}^{N} x_{i2} y_{i1} - \sum x_{i1}\, y_{i2}}{\sum_{i=1}^{N} x_{i1}\, y_{i1} + \sum_{i=1}^{N} x_{i2}\, y_{i2}} \qquad (27)$$

As três matrizes associadas são:

$$M(\alpha_m) = \begin{Bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_m & -\operatorname{sen}\alpha_m \\ 0 & \operatorname{sen}\alpha_m & \cos\alpha_m \end{Bmatrix} \qquad (28)$$

$$M(\beta_m) = \begin{Bmatrix} \cos\beta_m & 0 & \operatorname{sen}\beta_m \\ 0 & 1 & 0 \\ -\operatorname{sen}\beta_m & 0 & \cos\beta_m \end{Bmatrix} \qquad (29)$$

$$M(\gamma_m) = \begin{Bmatrix} \cos\gamma_m & -\operatorname{sen}\gamma_m & 0 \\ \operatorname{sen}\gamma_m & \cos\gamma_m & 0 \\ 0 & 0 & 1 \end{Bmatrix} \qquad (30)$$

5.4 Aplicação do método para a localização de substruturas

Preliminarmente, destaca-se da molécula a parte onde foi localizada a subestrutura, seguindo a relação da tabela 4.6. Tem-se, então, dois objetos (pedaço de molécula e subestrutura) com o mesmo número de pontos (átomos) e o mesmo padrão de conectividade.

A seguir, aplica-se o método seguindo os passos:

1. Determina-se o centro de gravidade dos dois objetos:

   a. $c_x = \frac{1}{N} . \sum_{i=1}^{N} x_{ij}$

   b. $c_y = \frac{1}{N} . \sum_{i=1}^{N} y_{ij}$

   c. $c_z = \frac{1}{N} . \sum_{i=1}^{N} z_{ij}$

2. Move-se ambos os objetos para a origem dos eixos coordenados:

   a. $x' = x_i - c_x$, com i variando de 1 a N

   b. $y' = y_i - c_y$, com i variando de 1 a N

   c. $z' = z_i - c_z$, com i variando de 1 a N

3. Verifica-se o critério de parada. Se atingida a condição preestabelecida, estará terminada a comparação, caso contrário, executa-se os passos a seguir.

4. Faz-se a rotação sobre o eixo *X:*

   a. Determina-se o ângulo ótimo para a rotação sobre o eixo *X*, utilizando a fórmula de (25).

   b. Determina-se as novas coordenadas dos pontos do objeto *Y*, segundo (16), utilizando a matriz expressa em (28), ou seja:

$$Y' = M (\alpha_m) \times Y^T$$

5. Faz-se a rotação sobre o eixo *Y:*

   a. Determina-se o ângulo ótimo para a rotação sobre o eixo *Y*, utilizando a fórmula de (26).

   b. Determina-se as novas coordenadas dos pontos do objeto *Y*:

$$Y' = \mathrm{M}\,(\,\beta_m\,) \times Y^{\mathrm{T}}$$

6. Faz-se a rotação sobre o eixo *Z:*

   a. Determina-se o ângulo ótimo para a rotação sobre o eixo *Z*, utilizando a fórmula de (27).

   b. Determina-se as novas coordenadas dos pontos do objeto *Y*:

$$Y' = \mathrm{M}\,(\,\gamma_m\,) \times Y^{\mathrm{T}}$$

7. Tendo completado um ciclo, verifica-se o critério de parada. Se atingida a condição de parada (item 7.5), estará terminada a comparação, caso contrário, retorna-se ao item 4.

5.5. Critério de interrupção

Considerando que o presente método aproxima os pontos correspondentes de dois objetos sem garantir uma verdadeira sobreposição, é preciso estabelecer um critério para interromper a iteração, ou seja, uma condição a partir da qual se pode estabelecer se as moléculas são ou não congruentes. Utilizando a terminologia "fuzzy", é preciso estabelecer a *função de pertinência* com relação à congruência dos objetos.

Como o processo diminui progressivamente a distância média entre os pontos correspondentes, uma condição a ser considerada é a diferença dessas distâncias médias entre duas iterações consecutivas.

Em outros termos, o decremento da distância média *R,* após uma iteração *n*, teria que ser menor que um valor ε. Ou seja:

$$\Delta R_n = R_{n-1} - R_n < \varepsilon$$

Isto implica no cálculo da distância média *R* após cada iteração, segundo a fórmula: ε

$$R = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{3}\left(x_{ij} - y_{ij}\right)^2}$$

Esse cálculo demandaria considerável esforço computacional, especialmente se o número de átomos de cada molécula fosse elevado.

Uma vez que os ângulos ótimos de rotação (α_m, β_m, γ_m) já são calculados durante o processo, e sua soma é progressivamente decrescente, uma condição mais racional seria

$$\phi = |\alpha_m| + |\beta_m| + |\gamma_m| < \delta \,.\, \varepsilon$$

Usando esse critério, a iteração seria interrompida a partir do momento em que o movimento de rotação fosse menor que um valor δ predeterminado, o que não demandaria qualquer computação adicional além da soma acima.

Uma vez que o método ora proposto se presta a diversos propósitos, os conceitos "idêntico" ou "sobreposto" não são de imediato muito claros e dependem da interpretação do usuário. Assim, as constantes δ e ε são números difusos, cujo grau de inclusão no conjunto "sobreposto" deve ser definido pelo químico, com base na sua realidade, na sua interpretação dos dados, na precisão das medidas ou outras circunstâncias.

Os valores adotados para as finalidades deste trabalho serão de $\varepsilon=10^{-1}$ (decremento da distância média de *R*) e de $\delta=10^{-2}$ ((somas dos valores absolutos dos ângulos de uma iteração). Além disso, considerar-se-á sobrepostas e, portanto, idênticas, duas moléculas cuja distância média entre seus átomos for menor que 5.10^{-1}, por ser esse valor correspondente a cerca de um terço da distância geralmente existente entre dois átomos consecutivos de uma molécula de origem vegetal.

O Anexo III mostra o fluxograma do reconhecimento espacial e o Anexo IV apresenta tabelas comparativas de verificações feitas, a fim de que se possa avaliar a eficiência do método e a relação entre ΔR e ϕ.

CAPÍTULO VI

CONCLUSÃO

Diversos sistemas computacionais foram elaborados para auxílio à determinação estrutural de mão de "heurísticas" diversas, tais como fornecer o maior número de informações possível, tanto relacionadas com as características imediatamente observáveis (densidade, pontos de fusão, de solidificação etc.), como resultantes de métodos físicos de análise (EM, IV, UV, RMN). São, ainda, acrescentadas regras decorrentes da experiência dos especialistas, mormente aquelas sobre condições que as estruturas devam ou não satisfazer.

Uma maneira rápida e prática de reduzir o tempo gasto para a identificação de moléculas é a determinação de características estruturais de partes de uma molécula por meio de dados experimentais, obtendo-se subestruturas. Especialmente na química dos produtos naturais, o conhecimento de determinadas subestruturas pode facilitar a comparação com compostos já elucidados, podendo sugerir a pertinência da substância em análise a determinadas classes conhecidas, o que reduziria mais ainda os ramos da árvore de possibilidades.

Utilizando-se técnicas de reconhecimento de padrões propôs-se um método para a identificação estérica de subestruturas de moléculas orgânicas de micromoléculas vegetais em uma base de dados de substâncias isoladas de plantas, com geometria otimizada por um programa modelador (HyperChem®). O método é uma combinação das técnicas de Morgan-Gastmans e de Sippl-Stegbucnher, fazendo uso de recursos de IA e de lógica difusa.

O método pode ainda aproveitar a base de dados de sistemas existentes para estudos de quimiosistemática e evolução química vegetal, alem de servir de base para a elaboração de um programa gerador estrutural automático.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ALLINGER, N. L. et alii. *Química orgânica* 2.ed. Rio de Janeiro: Guanabara, 1978

2. ANDREWS, H. C. *Introduction to Mathematical Techniques in Pattern Recognition*, John Willy & Sons, 1972

3. ATTIAS, R. *J. Chem Inf. Comp. Scii,* **23**: 102, 1993

4. BARRETO, J. M., *Inteligência artificial no limiar do século XXI*, Florianópolis: J. M. Barreto, 1997

5. BEALE, R. and JACKSON, T. *Neural Computing: An Introduction,* Philadelphia: Adam Higler, 1991

6. BEZDEK, J. *Pattern Recognition With Fuzzy Objetctive Function Algorithms*, New York: Plenum Press, 1981

7. BORGES, J H G. *Design e Implementação de um gerador estrutural automático para o projeto Sistemat*, Tese de doutoramento, USP, 1998

8. BREMSER, W. *Angew. Chem. Int. Ed. Engl.,* **27**: 247, 1988

9. BREMSER, W; KLIER, M and MEYER, E. *Org. Magn. Res.*, **7**: 97, 1975

10. CARABEDIAN, M; DAGANE, I and DUBOIS, J E. *Anal. Chem.,* **60**: 2186, 1988

11. CHRISTIE, B D and MUNK, M E, *J. Am. Chem. Soc.*, **113**: 3570, 1991

12. EMERENCIANO, V P. *Química nova,* **16**(6), 551-559, 1993

13. FELTRE, R. *Química*, Vol. 1 e 2, 3. ed. São Paulo: Moderna, 1991

14. FU, K. S. *Syntactic Pattern Recognition and Applications*, Englewoods Cliffs: Prentice-Hall, 1982

15. FULLER, R., *Introducing to Neuro-Fuzzy Systems*,: Berlin: Springer-Verlag, 1999

16. GASTMANS, J. P., FURLAN M. et al. *A inteligência artificial aplicada à química de produtos naturais,* revista Química Nova, 13(1), pág. 10 a 15, 1990

17. GRAY, N. A. B. *Computer Assisted Structure Elucidation*, New York: John Wiley & Sons, 1986

18. KLIR, G. J.; ST.CLAIR, U. and YUAN, B. *Fuzzy Set Theory – Foundations and applications*, Upper Saddle River: Prentice-Hall, 1997

19. JAIN, A. e MAO, J. *Neural Networks and Pattern Recognition*, p. 194-212 in *Computational Intelligence - Imitating Life*, Piscataway, New Jersey/USA, IEEE Press, 1994

20. LINDSAY, R. K. et alii. *Applications of Artificial Intelligence in Organic Chemistry: The Dendral Project*, New York: McGraw-Hill, 1980

21. MILITÃO, J. S. L. T. *Identificação de triterpenos com auxílio de computador*, Tese de Doutoramento, Universidade Federal do Ceará, 1996

22. NEUDERT, E. and PENK, M. *J. Chem. Inf. Comput. Sci.*, **36**: 224, 1996

23. PEDRICZ, W. and GOMIDE, F., *Na Introducing to Fuzzy Sets Analysis and Design*, Cambridge: MIT Press: 1998

24. RODRIGUES, G. V. *Um sistema especialista para determinação estrutural de sesquiterpenos lactonizados - Um enfoque multi-espectral aliado a dados botânicos*, Tese de doutoramento, USP, 1996

25. RICH, E. and KNIGHT, K. *Inteligência artificial*. 2. ed. São Paulo: Makron Books, 1994

26. SASAKI, S. I. and KUDO, Y. *J. Chem. Inf. Comput. Sci.* **16**(1): 43-49, 1975

27. SHELLEY, C.; HAYS, T.; MUNK, M. E. and RAMAN, R. *Anal. Chim. Acta,* **103**: 121, 1978

28. SIPPL, M. J. and STEGBUCHNER, H. *Superposition of three-dimensional objects: a fast and numerically stable algorithm for the calculation of the matrix of optimal rotation*. Computer Chem. Vol. 15, pág 73-78, 1991

29. UHR, L. *Pattern Recognition, Learning and Thought*, Englewoods Cliffs: Prentice-Hall, 1973

30. WOOLDRIDGE, M., MULLER, J. P. and TAMBE, M., *Intelligent Agents II - Agent Theories, Architectures and Langauages,* Berlin: Springer Verlag, 1995

31. ZADEH, L. A., *Fuzzi Sets*, Information and Control, vol. 8 , p. 338-353, 1965

A N E X O S

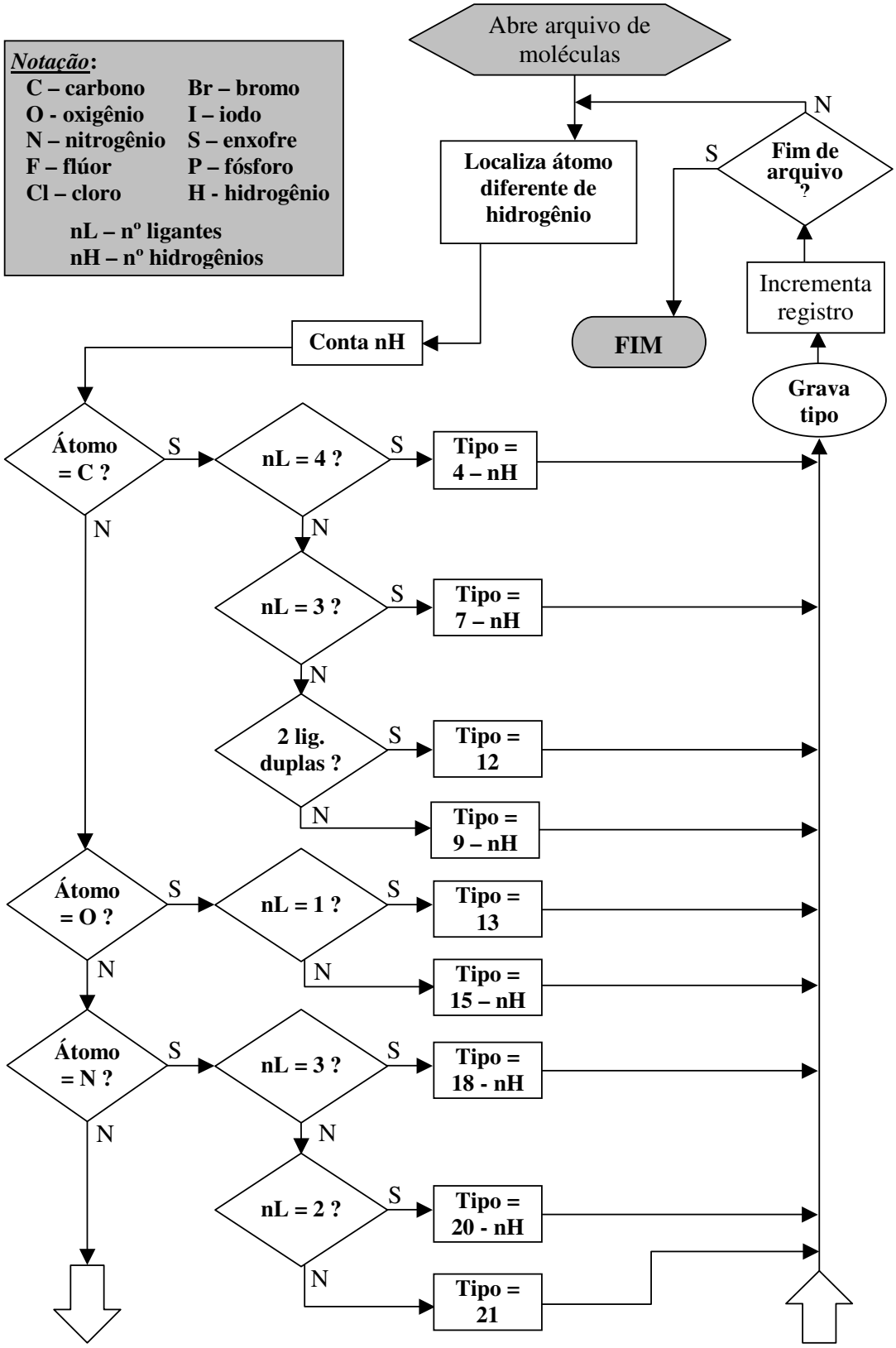
Anexo I: Fluxograma de cálculo do tipo para tabela de conectividade

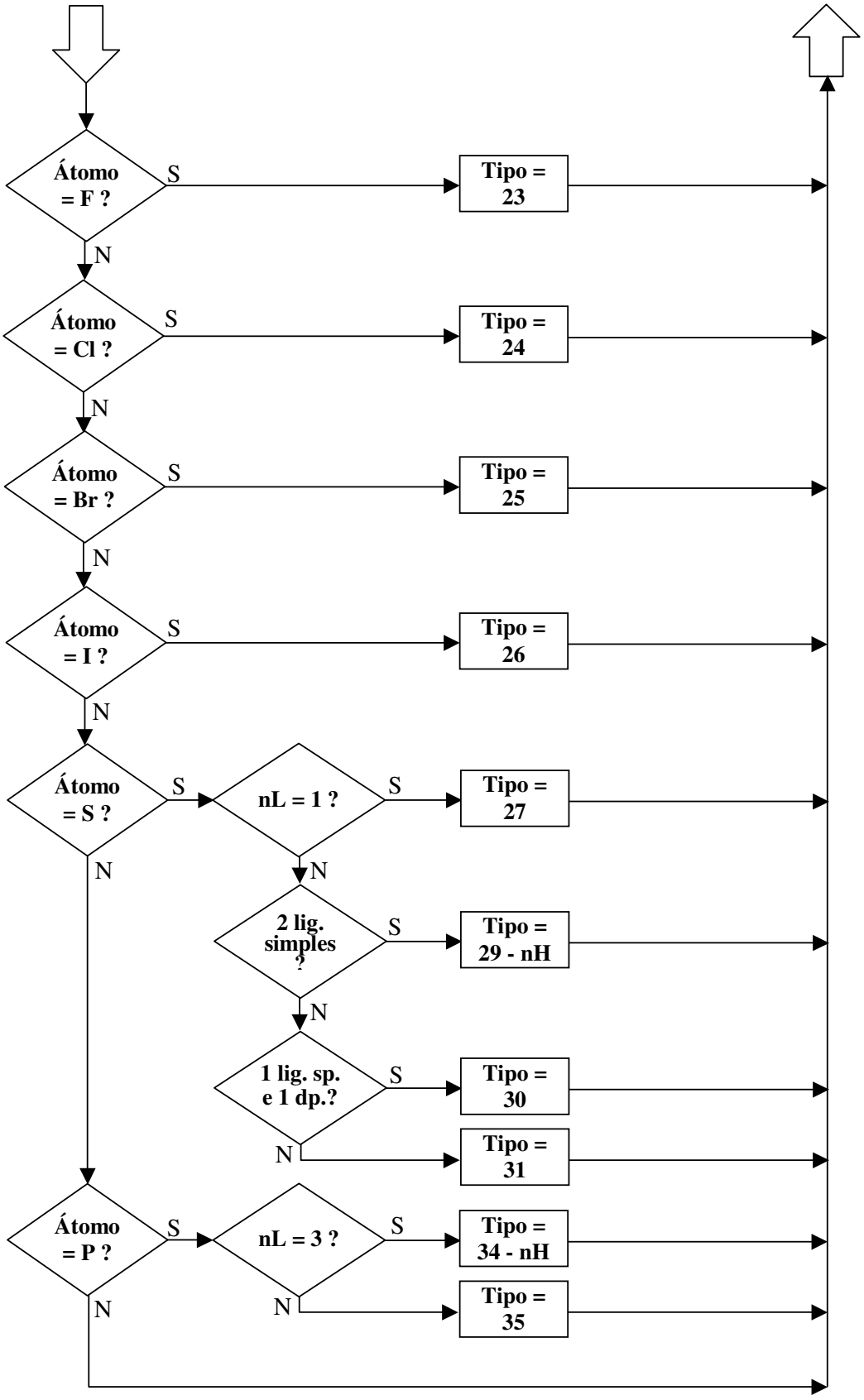
Anexo II: Fluxograma de localização plana

Anexo III Fluxograma de reconhecimento espacial

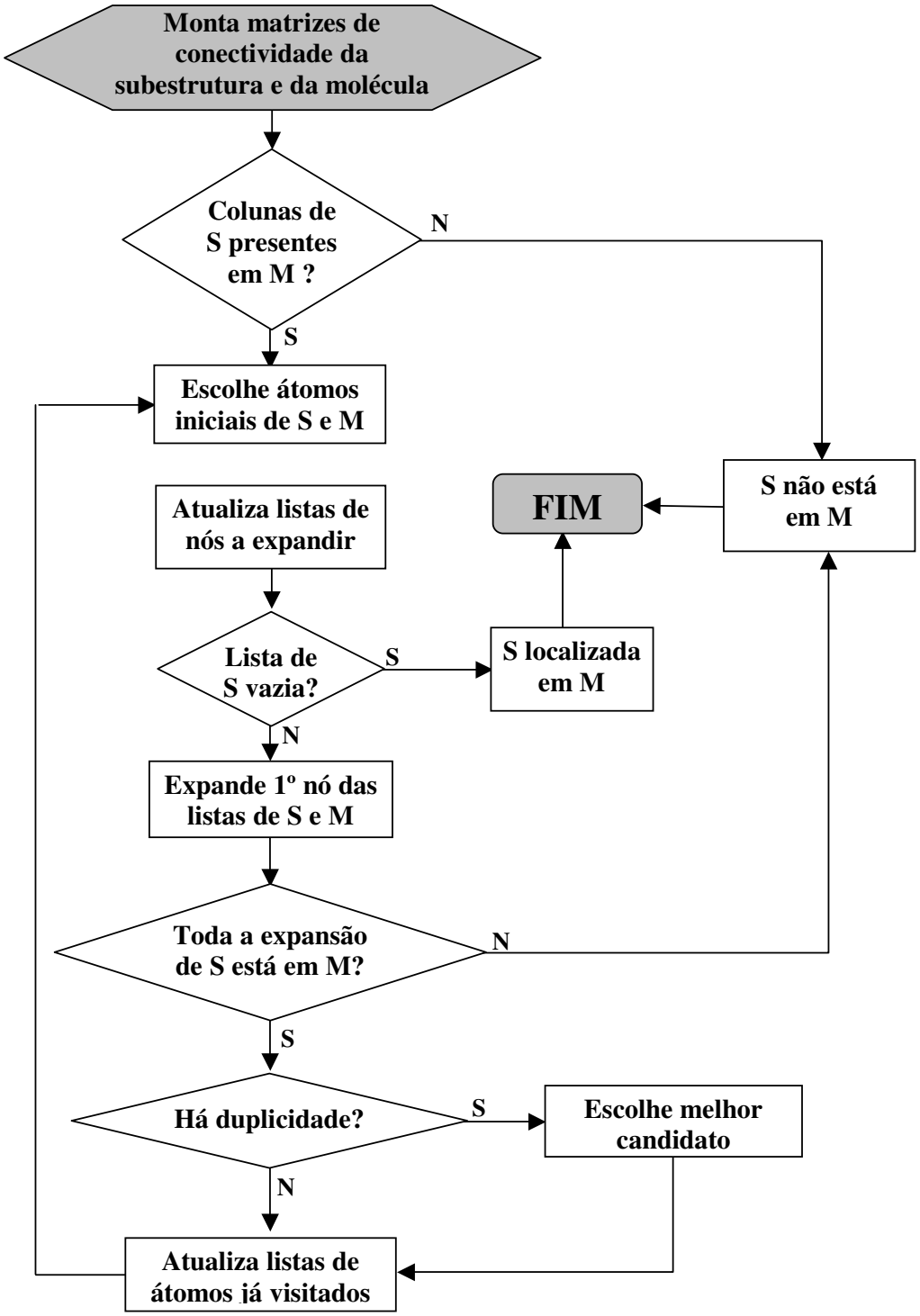
Anexo IV: Tabelas comparativas entre incrementos de ângulos e incrementos de distâncias.

ANEXO I - Fluxograma de cálculo do tipo para tabela de conectividade

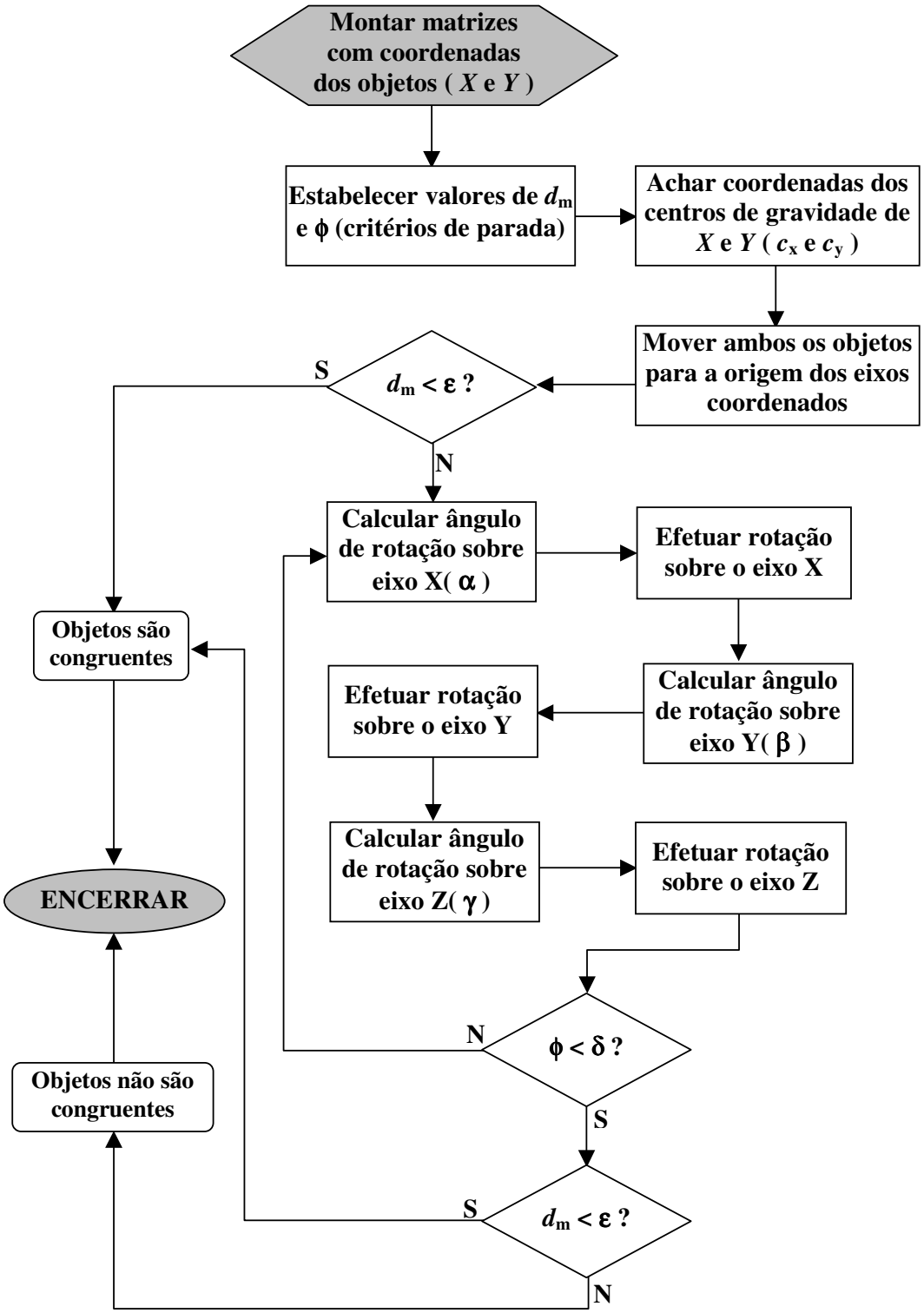



Átomo = F ?
S
Tipo = 23
N
Átomo = Cl ?
S
Tipo = 24
N
Átomo = Br ?
S
Tipo = 25
N
Átomo = I ?
S
Tipo = 26
N
Átomo = S ?
S
nL = 1 ?
S
Tipo = 27
N
N
2 lig. simples ?
S
Tipo = 29 - nH
N
1 lig. sp. e 1 dp.?
S
Tipo = 30
N
Tipo = 31
Átomo = P ?
S
nL = 3 ?
S
Tipo = 34 - nH
N
N
Tipo = 35

ANEXO II - Fluxograma da Localização plana



ANEXO III – Fluxograma do reconhecimento espacial



ANEXO IV – Tabelas comparativas entre incrementos de ângulos e incrementos de distâncias

Para ilustrar e avaliar a eficiência do método de reconhecimento espacial, utilizaram-se as moléculas de triterpenos cujos gráficos aparecem na figura IV.1. Na parte superior vê-se as moléculas com sua maior face voltada para o observador, aparentando ter formatos idênticos. Na parte inferior é mostrada cada molécula "de perfil" e pode-se notar que, nas moléculas A e B os átomos 1 e 2 acham-se dispostos em lados opostos, enquanto na molécula C ambos os átomos se acham do mesmo lado da molécula. Trata-se, portanto, de uma isomeria espacial, que não se pode verificar no plano, pois suas projeções planas podem parecer idênticas.

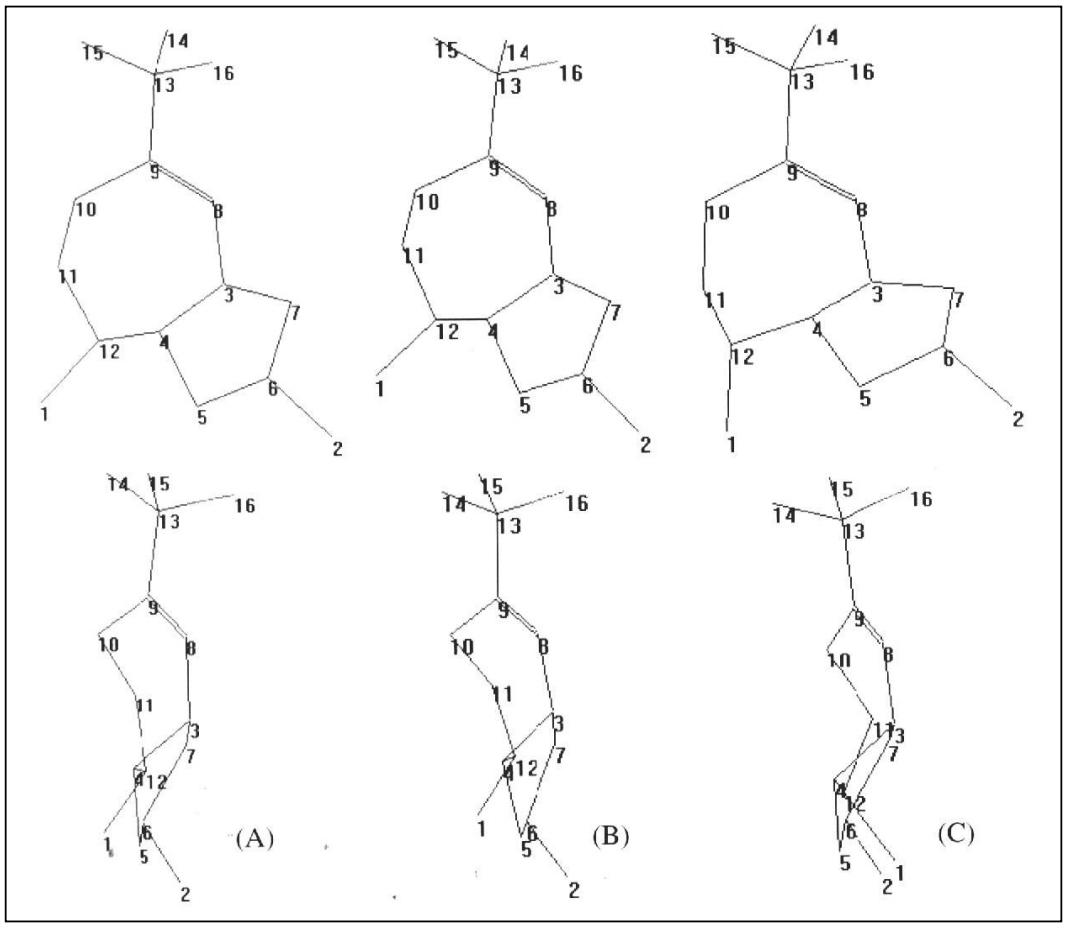


Fig IV.1 – Moléculas de triterpenos utilizadas para exemplo

A tabela IV.1, a seguir, apresenta as coordenadas em relação aos eixos cartesianos X, Y e Z. É a partir dessas coordenadas que se aplicará o método de reconhecimento espacial entre os pares de moléculas A-B e A-C, com seis iterações (ciclos) para cada par de moléculas. As tabelas IV.2 e IV.3 mostram os resultados obtidos em cada verificação.

Tabela IV.1 – Coordenadas cartesianas dos 16 átomos das moléculas da fig IV.1

Molécula A			**Molécula B**			**Molécula C**		
x_1	x_2	x_3	y_1	y_2	y_3	z_1	z_2	z_3
2,428284	4,250217	11,25272	-7,515808	14,33306	0,8913323	-11,51111	12,20547	4,116331
7,582174	6,003406	12,61314	-2,362386	16,08547	2,256574	-6,70893	14,1131	5,116726
4,110606	7,341539	13,01213	-5,833404	17,42234	2,654915	-10,28162	15,24223	5,345923
3,851242	6,310585	11,8845	-6,093065	16,39196	1,526958	-10,322	14,54621	3,960117
5,159136	5,478489	11,90329	-4,785036	15,5599	1,545946	-8,980255	13,76517	3,930583
6,260557	6,561929	12,03257	-3,683759	16,64348	1,674982	-7,963408	14,78108	4,502519
5,622437	7,686541	12,89564	-4,322377	17,76865	2,53699	-8,776301	15,60725	5,536069
3,245057	8,577883	13,06231	-6,701668	18,65918	2,704412	-11,11388	16,49978	5,484125
2,06123	8,786441	12,43829	-7,881506	18,86721	2,082447	-12,29285	16,78544	4,898228
1,491127	7,801477	11,43346	-8,453971	17,88225	1,076526	-13,01566	15,8329	3,95858
1,270035	6,390421	12,03843	-8,674099	16,47141	1,682165	-12,85709	14,33491	4,312163
2,549898	5,504972	12,16094	-7,394207	15,58574	1,80241	-11,59429	13,69381	3,675887
1,323631	10,10137	12,74462	-8,620596	20,18303	2,386784	-12,83837	18,20671	5,133443
1,869657	11,21718	11,80937	-8,07427	21,29739	1,450099	-11,99502	19,20118	4,281791
-0,225423	9,989369	12,6056	-10,16947	20,07133	2,248705	-14,34668	18,38227	4,787442
1,594351	10,45404	14,11984	-8,348841	20,53702	3,761564	-12,67812	18,51044	6,536659

IV.1 – Comparação entre as moléculas A e B.

A distância média original entre os átomos das duas moléculas é de 17,54445245. Após a translação que fez a junção dos centros de gravidade, essa distância ficou em 0,001398471. De acordo com o critério de interrupção adotado (R<5.10^{-1}), essa distância média já indica serem as moléculas idênticas. Entretanto, para efeito comparativo, realizaram-se seis ciclos de rotações. A soma dos ângulos (ϕ) em cada ciclo, como se pode ver na tabela IV.2, foi sempre decrescente e menor que 10^{-4}. Como as moléculas já estavam sobrepostas, as rotações ocasionaram até um pequeno aumento na distância média entre os átomos, decorrentes de erros de aproximações nos cálculos e menores que um centésimo de milésimo, portanto desprezíveis, o que reforça a certeza de que se trata da mesma substância química.

Tabela IV.2 – Resultados das rotações das moléculas A e B

Ciclo	ϕ	R_n	ΔR
1	$8{,}155.10^{-5}$	0,001406255	$-7{,}784.10^{-6}$
2	$2{,}571.10^{-5}$	0,001411709	$-5{,}454.10^{-6}$
3	$5{,}620.10^{-6}$	0,001412962	$-1{,}253.10^{-6}$
4	$1{,}203.10^{-6}$	0,001413232	$-2{,}699.10^{-7}$
5	$2{,}574.10^{-7}$	0,001413290	$-5{,}784.10^{-8}$
6	$5{,}509.10^{-8}$	0,001413302	$-1{,}238.10^{-8}$

IV.2 – Comparação entre as moléculas A e C

A distância média original entre os átomos das duas moléculas é de 18,05610282. A Superposição dos centros de gravidade por translação fez a distância média diminuir para 0,312265265. Esse valor está bem acima do valor de $5{,}10^{-1}$ exigido como média máxima. Torna-se, portanto necessária a tentativa de sobreposição por rotação, cujo resultado se acha na tabela IV.3.

Tabela IV.3 – Resultados das rotações das moléculas A e C

Ciclo	ϕ	R_n	ΔR
1	$5{,}838.10^{-2}$	0,295220474	$1{,}704.10^{-2}$
2	$2{,}368.10^{-2}$	0,290741075	$4{,}479.10^{-3}$
3	$5{,}096.10^{-3}$	0,290287089	$4{,}540.10^{-4}$
4	$1{,}116.10^{-3}$	0,290209632	$7{,}746.10^{-5}$
5	$2{,}445.10^{-4}$	0,290193659	$1{,}597.10^{-5}$
6	$5{,}361.10^{-5}$	0,290190206	$3{,}454.10^{-6}$

Note-se que os três últimos ciclos não seriam necessários, pois o valor de ϕ ao final da terceira iteração (5,096E-03) já é menor que δ=10^{-2} estabelecido como critério de parada. A partir desse ponto, também o ΔR (incremento da distância média) já é menor que o ε=10^{-1} adotado.

Dessa forma, como a distância média entre os átomos ficou próxima de 0,2902, de acordo com os critérios adotados, pode-se concluir que não se trata da mesma substância.