Universidade Estadual de Campinas
Instituto de Matemática Estatística e Computação Científica
Departamento de Matemática Aplicada

Algoritmos para Problemas de Geometria Molecular

Felipe Delfini Caetano Fidalgo
Mestrado em Matemática Aplicada

Orientador: Prof. Dr. Carlile Campos Lavor

Este trabalho contou com suporte financeiro do CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico.

Campinas-SP
Maio de 2011

Algoritmos para Problemas de Geometria Molecular

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por **Felipe Delfini Caetano Fidalgo** e aprovada pela comissão julgadora.

Campinas, 13 de maio de 2011.

Prof. Dr. Carlile Campos Lavor
Orientador

Banca examinadora:

Prof. Dr. Carlile Campos Lavor (UNICAMP)

Prof. Dr. Luiz Satoru Ochi (UFF)

Prof. Dra. Márcia Aparecida Gomes Ruggiero (UNICAMP)

Dissertação apresentada ao Instituto de Matemática Estatística e Computação Científica, UNICAMP, como requisito parcial para a obtenção do título de **MESTRE em Matemática Aplicada**.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Bibliotecária: Maria Fabiana Bezerra Müller – CRB8 / 6162

Fidalgo, Felipe Delfini Caetano

F448a Algoritmos para problemas de geometria molecular/Felipe Delfini Caetano Fidalgo – Campinas, [S.P. : s.n.], 2011.

Orientador: Carlile Campos Lavor.

Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1.Geometria molecular. 2.Algoritmos. I. Lavor, Carlile Campos. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título

Título em inglês: Algorithms for molecular geometry problems

Palavras-chave em inglês (Keywords): 1.Molecular Geometry. 2.Algorithms.

Área de concentração: Matemática da Computação.

Titulação: Mestre em Matemática Aplicada.

Banca examinadora: Prof. Dr. Carlile Campos Lavor (IMECC - UNICAMP)

Prof. Dr. Luiz Satoru Ochi (IC - UFF)

Prof. Dra. Márcia Aparecida Gomes Ruggiero (IMECC - UNICAMP)

Data da defesa: 13/05/2011

Programa de Pós-graduação: Mestrado em Matemática Aplicada

Dissertação de Mestrado defendida em 13 de maio de 2011 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Prof.(a). Dr(a). CARLILE CAMPOS LAVOR

Prof.(a). Dr(a). LUIZ SATORU OCHI

Prof.(a). Dr(a). MÁRCIA APARECIDA GOMES RUGGIERO

Para aqueles que me inspiram mais e mais a cada dia: minha família! ....

"Portanto, se alguém está em Cristo, é nova criação. As coisas antigas já passaram; eis que surgiram coisas novas."
I Coríntios 5:17.

Agradecimentos

Agradeço, primeiramente, ao autor e consumador de minha fé, Jesus Cristo, pela realização deste trabalho.

Agradeço aos meus pais, Luiz Carlos e Margaret, que me ensinaram, desde muito cedo, tudo o que aprendi de valioso, o valor do conhecimento, a respeitar e valorizar as outras pessoas e nunca desistir dos meus sonhos. Agradeço à minha querida irmã, Ana Karina, pelo apoio e pela cumplicidade de irmãos. Agradeço à minha avó, Anacyr, que sempre me orientou e sempre me ensinou a ter garra.

Sou grato ao meu orientador, Prof. Dr. Carlile Lavor, pelos valores que me ensinou. Por ter me mostrado que a grandeza de um cientista está em sua humildade.

Agradeço, também, à Prof. Dra. Márcia Aparecida Gomes Ruggiero pelas sugestões de grande valia para este trabalho, ao Prof. Dr. Luiz Satoru Ochi pelas dicas dispensadas à dissertação, professores que compuseram a banca de defesa, e ao Prof. Dr. Aurélio Ribeiro Leite de Oliveira, coordenador do Programa de Pós-Graduação em Matemática Aplicada do IMECC, por todo apoio, incentivo e pelas sugestões dadas na pré-defesa.

Não posso me esquecer de todos os mestres que se passaram pela minha vida, tanto os bons quanto os maus exemplos.

Agradeço aos grandes amigos Carlos Renato Medeiros e Bruno Amaro, amigos desde os tempos de graduação, que me acompanham até hoje com grande amizade, cumplicidade e apoio. Sou grato aos companheiros do MiLab - IMECC (Mathematical Imaging and Computational Intelligence Laboratory) pela amizade e pelas boas discussões matemáticas.

Agradeço aos queridos amigos andradinenses Daniel Barros Gomes, Hermes Dantas, Éder Cavalcante, Rodrigo Franco, José Ricardo Barros Gomes e Amanda Giacomette que sempre me suportaram com seu amor, carinho e dedicação. Pensando em amizade, não deixo de lembrar todos os amigos que passaram pela minha vida (e ainda passam) tanto em Andradina quanto em Campinas. A amizade é um bem conquistado que o tempo não apagar.

Enfim, agradeço a todos os que, direta ou indiretamente, contribuíram para a formação de mais um mestre o qual nunca quer perder a criatividade e o desejo por aprender mais-e-mais como um pequeno aluno a descobrir o que há de mais belo nesta vida.

Resumo

Neste trabalho, analisamos dois algoritmos da literatura para o "Molecular Distance Geometry Problem" (MDGP) e propomos um novo algoritmo que mantém a qualidade das soluções obtidas pelos dois anteriores e apresenta ganhos em termos de eficiência computacional. O MDGP consiste em determinar as posições dos átomos de uma molécula, no espaço tridimensional, a partir de um conjunto de distâncias entre eles. Quando todas as distâncias são conhecidas, o problema pode ser resolvido em tempo polinomial. Caso contrário, é um problema NP-difícil.

Abstract

In this work, we analyse two algorithms from the bibliography to solve the so-called "Molecular Distance Geometry Problem" (MDGP). Then, we propose a new algorithm that keeps the quality on the solutions obtained by both the previous ones and shows gains regarding computacional efficiency. The MDGP consists on the determination of positions of atoms in a molecule, on the tridimensional space, from a set containing distances among them. When all the distances are known, the problem might be solved in polynomial time. Otherwise, it is an NP-hard problem.

Sumário

Lista de Figuras

Introdução

Em 2001, cientistas do consórcio internacional *Human Genome Project* e da empresa americana *Celera*, empresas concorrentes, decifraram mais de 3,1 bilhões de bases químicas do DNA, mapeando, assim, o genoma humano. Apesar da extrema importância do DNA para as questões inerentes ao ser humano e sua saúde, decifrar o significado de cada nucleotídeo (bases do DNA) e suas funções é uma tarefa árdua a ser desenvolvida.

Segundo [12], sabemos que o DNA apresenta pouca mobilidade, restringindo-se apenas ao interior da célula, e que sua ação na determinação das características hereditárias é feita de um modo indireto. O DNA induz à formação do chamado RNA mensageiro o qual migra para o citoplasma e une-se a um ribossomo. Ambos iniciam o processo de ordenação e ligação dos aminoácidos, em cadeia, que formarão a proteína. Tal processo é chamado de tradução. Os aminoácidos são conectados por fortes ligações químicas e eles, juntamente com sua ordem na cadeia, são fixados para cada proteína. Eles são especificados pelo gene, que são sequências de moléculas de DNA.

As proteínas formam uma classe importante de moléculas. São codificadas nos genes e produzidas nas células pelo processo de tradução, descrito acima. Essas, depois de formadas, é que atuarão nas mais variadas funções de nosso organismo como: determinação de caracteristicas hereditárias, transporte de nutrientes e metabólitos, composição das células, entre outros.

Com a ajuda de proteínas, vírus são capazes de crescer, traduzir, integrar e replicar,

causando doenças. Algumas proteínas são tóxicas por si só (e até infecciosas), como as proteínas encontradas em plantas venenosas e na carne bovina, que são as que causam a doença da Vaca Louca [18].

A história das proteínas começa no século XVIII, a partir da descoberta de que alguns componentes do mundo vivo coagulam a altas temperaturas e em meio ácido, como a clara de ovo (albúmen), o sangue, o leite, dentre outros. Substâncias com tais características foram denominadas albuminóides. Já no século XIX, descobriram que os principais componentes das células eram albuminóides. Agora, o primeiro a usar o termo *proteína* (do grego, *proteios*, que significa primitivo) foi Gerardus Johannes Mulder (1802-1880), um químico holandês, em artigo de 1838, sugerido pelo sueco Jons Jacob Berzelius (1779-1848), já que este acreditava que os albuminóides eram *constituintes fundamentais dos seres vivos*([12], p. 5). No século XX, os cientistas chegaram à conclusão que as proteínas são formadas por aminoácidos encadeados, o que foi sugerido pelo químico alemão Franz Hofmeister (1850-1922). Em 1940, completou-se a identificação dos 20 aminoácidos que estão presentes naturalmente nas proteínas dos seres vivos.

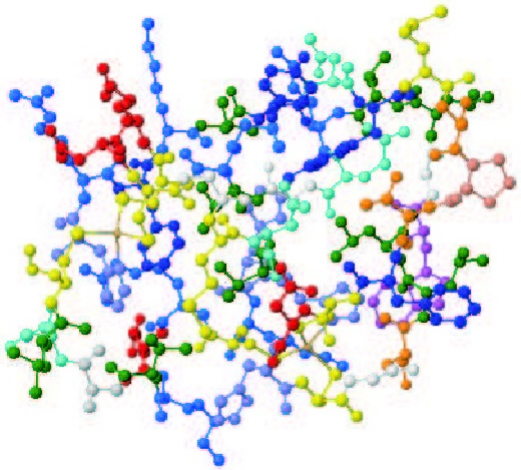

Figura 0.1: Proteína *kinase C* (1PTQ) formada por 402 átomos

Em relação à geometria protéica, a maioria das proteínas naturais têm estruturas tridimen-

sionais específicas que estão intimamente ligadas às suas funções biológicas. Essas estruturas apresentam pequenas variações, considerando condições de temperatura e configurações locais típicas. No início da década de 60, Christian B. Anfinsen, e alguns colaboradores, descobriram uma relação determinística entre a disposição tridimensional da proteína e a sequência de aminoácidos. Dessa forma, a determinação de sua estrutura tornou-se um passo importante para a compreensão das propriedades biológicas de todas as proteínas [2].

Neste trabalho, discutiremos um problema bem abrangente em relação à estrutura geométrica de moléculas para a determinação das mesmas, partindo de um conjunto de distâncias entre os átomos da molécula obtidas por meio de experimentos físicos e modelos teóricos. Tais problemas são chamados de *Problemas Moleculares de Geometria de Distância (PMGD)*. Vamos apresentar e analisar uma série de algoritmos a fim de resolver este problema com a maior eficácia possível. As proteínas são, apenas, motivação para estudar geometria de moléculas, já que sua estrutura diz muito sobre sua função, como vimos anteriormente.

No Capítulo 2, analisaremos o PMGD considerando seus dois casos. Primeiro, o caso particular, onde todas as distâncias entre quaisquer pares de átomos são conhecidas. Mostraremos dois algoritmos para resolver este caso: um, usando a decomposição SVD de uma matriz de distâncias e, o outro, executado em ordem de complexidade linear definido pela resolução de um sistema linear 3×3, ambos apresentados em [5]. Segundo, o caso geral, no qual dispomos de um conjunto arbitrário de distâncias inter-atômicas, podendo ser completo ou incompleto. Mostraremos um algoritmo que chamaremos de *Algoritmo Iterativo de Determinação Geométrica (AIDG)*, baseado na resolução iterativa de sistemas lineares 3×3, publicado em [6].

No Capítulo 3, mostraremos uma modificação do AIDG introduzida por Wu e Wu em [16], à qual chamaremos de *Algoritmo Iterativo de Determinação Geométrica Atualizado*, que visa diminuir a incidência da propagação de erros de arredondamento, que podem, eventualmente, acumular a cada iteração.

No Capítulo 4, apresentaremos nossa proposta para este trabalho, a ser comparada com as mostradas previamente, a fim de resolver o PMGD com conjunto arbitrário. Introduziremos um algoritmo chamado *Algoritmo T (AT)* que também consiste na resolução de um sistema linear, entretanto, de dimensão 4×4.

Por fim, no Capítulo 5, vamos comparar os algoritmos AIDG e AT, analisando tempo de execução e erros na precisão, em implementação feita em MATLAB.

Capítulo 1

Problema Molecular de Geometria de Distância

Sejam n o número de átomos em uma dada molécula e $x_1, ..., x_n$ os vetores de coordenadas de tais átomos, onde $x_i = (x_{i,1}, x_{i,2}, x_{i,3})$, com $x_{i,k} \in \mathbb{R}$ $(k = 1, 2, 3)$, indica a posição do átomo i em $\mathbb{R}^3$, para cada $i \in \{1, ..., n\}$. $x_{i,k}$ é a k-ésima coordenada do átomo i.

Suponhamos que as coordenadas $x_1, ..., x_n$, dos n átomos da molécula em questão, são conhecidas, ou seja, conhecemos as posições em $\mathbb{R}^3$ de seus n átomos. Então, as distâncias entre os átomos i e j, quaisquer, podem ser calculadas na forma

$$d_{i,j} = \|x_i - x_j\|,$$

onde $\|\cdot\|$ é a norma Euclidiana.

Entretanto, a situação que mais interessa é a inversa. Uma pergunta que surge de forma "natural" é a seguinte: se os únicos dados conhecidos sobre a molécula são as distâncias entre seus átomos, será que é possível descobrir as posições de cada um deles em $\mathbb{R}^3$?

Matematicamente, seja S um conjunto de pares (i, j), com $i, j \in \{1, .., n\}$, para os quais conhecemos a distância $d_{i,j} \in \mathbb{R}$ entre eles. Então, as coordenadas das posições $x_1, ..., x_n$ dos

n átomos definem o sistema não-linear de equações

$$\|x_i - x_j\| = d_{i,j}, \qquad (i,j) \in S. \tag{1.1}$$

Logo, o problema é determinar tais soluções de modo eficiente. Nem sempre pode-se encontrar todas as coordenadas, já que nem sempre se possui um número suficiente de valores de distâncias entre os átomos que nos permita definir um sistema viável.

O problema acima é chamado de *Problema Molecular de Geometria de Distância* (PMDG). Este problema é uma classe especial do chamado Problema de Geometria de Distância (PDG) que consiste em encontrar as coordenadas de certos pontos, dadas as distâncias como instâncias conhecidas.

Na prática, os valores das distâncias podem não ser exatos, isto é, eles podem conter erros. De forma mais geral, os dados conhecidos para o PMGD podem ser cotas superiores ($u_{i,j}$) e inferiores ($l_{i,j}$) para as distâncias $d_{i,j}$ de modo que

$$l_{i,j} \leq \|x_i - x_j\| \leq u_{i,j}, \qquad (i,j) \in S. \tag{1.2}$$

Neste trabalho, apenas os casos onde as distâncias são dadas exatamente serão considerados.

As distâncias entre muitos dos pares de átomos em uma molécula (especialmente proteínas) podem ser determinadas a partir de um certo conhecimento em Química usando, por exemplo, alguns tipos de comprimentos e ângulos de ligações e experimentos de Ressonância Magnética Nuclear (RMN) [2]. Se conseguirmos obter um conjunto de distâncias inter-atômicas suficientemente grande, então a estrutura de tal molécula pode ser determinada via resolução de um PMDG [3].

Na maioria dos casos, Problemas Moleculares de Geometria de Distâncias são computacionalmente intratáveis (NP-difícil), principalmente em sua forma mais geral [13].

Pode-se dividir o PMDG em duas classes: os que possuem um conjunto completo de dis-

tâncias inter-atômicas, ou seja, para cada par de átomos da molécula em questão, a distância é conhecida, e aqueles que possuem um conjunto arbitrário de distâncias inter-atômicas, isto é, nem todas as distâncias entre pares de átomos da molécula precisam estar disponíveis.

1.1 Conjunto Completo

A primeira classe do PMGD com distâncias exatas é aquela em que todas as distâncias $d_{i,j}$ entre todos os pares de átomos i e j (com vetor de coordenadas x_i e x_j, respectivamente) da molécula são conhecidas *exatamente*.

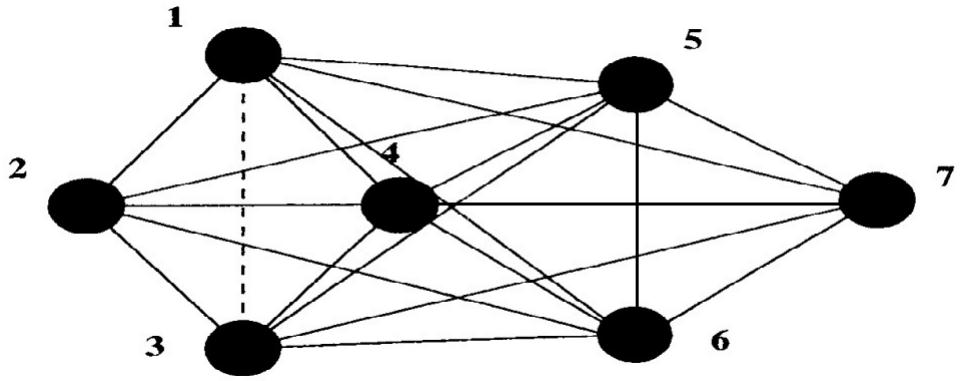

Figura 1.1: Molecula com sete átomos com conjunto completo de distâncias

Matematicamente, o problema consiste em encontrar os vetores coordenadas $x_1, ..., x_n$ tais que

$$\|x_i - x_j\| = d_{i,j}, \qquad \forall i, j = 1, ..., n, \tag{1.3}$$

onde todos os valores de $d_{i,j}$ são dados.

Duas abordagens serão apresentadas, nesta primeira seção, a fim de resolver tal problema: uma, via Decomposição em Valores Singulares de uma matriz que abrigará os dados de distâncias entre os átomos e, outra, usando um algoritmo em ordem de complexidade linear cujo cerne é a resolução de um sistema linear 3×3.

Ambas construções a serem discutidas foram publicadas no artigo [5] por Dong et al.

1.1.1 Um tratamento matemático via Decomposição em Valores Singulares

Considere as equações (1.3). Elevando suas equações ao quadrado e as expandindo, temos

$$\|x_i\|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \qquad i,j = 1,...,n. \tag{1.4}$$

Pode-se definir um sistema de coordenadas de modo que o último átomo de nossa lista se torne sua origem, isto é, de forma que $x_n = (0,0,0)^T$. Tal mudança é possível, sem perda de generalidade, já que a estrutura molecular de uma proteína é invariante sob movimentos rígidos (rotação, reflexão e translação). Logo,

$$\|x_i\|^2 = \left\|x_i - (0,0,0)^T\right\|^2 = \|x_i - x_n\|^2 = d_{i,n}^2, \tag{1.5}$$

para $i = 1,...,n-1$.

Das equações (1.4) e (1.5), obtém-se

$$d_{i,n}^2 - d_{i,j}^2 + d_{j,n}^2 = 2x_i^T x_j, \qquad i,j = 1,...,n-1. \tag{1.6}$$

Define-se, então, a matriz $D = (D_{i,j})$, onde

$$D_{i,j} = \frac{d_{i,n}^2 - d_{i,j}^2 + d_{j,n}^2}{2}, \tag{1.7}$$

para $i,j = 1,...,n-1$.

Segue, de (1.6) e (1.7), que

$$D_{i,j} = x_i^T x_j, \tag{1.8}$$

para $i,j = 1,...,n-1$.

Se $X \in \mathbb{R}^{(n-1)\times 3}$ é dada por

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{n-1}^T \end{bmatrix},$$

tem-se que

$$D = XX^T. \tag{1.9}$$

De fato, pelas definições de D e X, temos

$$D = \begin{bmatrix} x_1^T x_1 & \dots & x_1^T x_{n-1} \\ \vdots & \vdots & \vdots \\ x_{n-1}^T x_1 & \dots & x_{n-1}^T x_{n-1} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{n-1}^T \end{bmatrix} \cdot \begin{bmatrix} x_1 \dots x_{n-1} \end{bmatrix} = XX^T.$$

Como $X \in \mathbb{R}^{(n-1)\times 3}$, temos $posto(X) \leq 3$. Além disso, se $x \in Im(D)$, existe $y \in \mathbb{R}^{n-1}$ tal que $Dy = x$, isto é, tal que $XX^T y = x$. Associando, convenientemente os termos na equação anterior, temos que $X(X^T y) = x$, ou seja, $x \in Im(X)$. Portanto, $Im(D) \subseteq Im(X)$ e, logo, $posto(D) \leq posto(X) \leq 3$.

Desse modo, como D deve ter posto menor ou igual a 3 e, além disso, é simétrica (pois $D = XX^T = D^T$), é possível calcular sua decomposição em valores singulares (SVD) reduzida, cuja parte teórica é apresentada em [14],

$$D = \widehat{U}\widehat{\Sigma}\widehat{U}^T, \tag{1.10}$$

onde $\widehat{U} \in \mathbb{R}^{(n-1)\times 3}$ é uma isometria ($\widehat{U}^T\widehat{U} = I_3$) e $\widehat{\Sigma} \in \mathbb{R}^{3\times 3}$ é uma matriz diagonal cujos três elementos diagonais são $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$, que são os três valores singulares de D.

De (1.10), segue que a solução de (1.9) pode ser obtida tomando

$$X = \widehat{U}\widehat{\Sigma}^{1/2}, \tag{1.11}$$

pois

$$XX^T = \widehat{U}\widehat{\Sigma}^{1/2}\left(\widehat{U}\widehat{\Sigma}^{1/2}\right)^T = \widehat{U}\widehat{\Sigma}^{1/2}\left(\widehat{\Sigma}^{1/2}\right)^T\widehat{U}^T = \widehat{U}\widehat{\Sigma}^{1/2}\widehat{\Sigma}^{1/2}\widehat{U}^T = \widehat{U}\widehat{\Sigma}\widehat{U}^T = D.$$

Podemos resumir o demonstrado acima na seguinte proposição:

Proposição 1. Dada uma molécula com n átomos, sendo que o último é colocado como origem do sistema de coordenadas, sejam $D \in \mathbb{R}^{(n-1)\times(n-1)}$ a matriz definida pela regra (1.7) e

$$D = \widehat{U}\widehat{\Sigma}\widehat{U}^T$$

sua decomposição em valores singulares reduzida. Então, a matriz

$$X = \widehat{U}\widehat{\Sigma}^{1/2},$$

cujas linhas são as posições dos $n-1$ primeiros átomos da molécula, resolve a equação matricial

$$D = XX^T.$$

Por ([7], p. 254), são necessárias $14mn^2 - 2n^3$ operações aritméticas de ponto flutuante para calcular as matrizes $\widehat{\Sigma}$ e $\widehat{U}$ da decomposição em valores singulares reduzida de uma matriz em $\mathbb{R}^{m\times n}$, lançando mão do Algoritmo Golub-Reinsch. Assim, para calcular tais termos da SVD reduzida para nossa matriz $D \in \mathbb{R}^{(n-1)\times(n-1)}$, é preciso fazer $\mathcal{O}(12n^3)$ operações. Portanto, quando todas as distâncias entre os átomos da molécula são conhecidas, o PMGD involvido pode ser resolvido em ordem de complexidade polinomial (de grau três, como mencionado acima) pelo método descrito nesta seção.

1.1.2 Um algoritmo com ordem de complexidade linear

Analisaremos, agora, o segundo método que resolve o PMGD com um conjunto completo de distâncias exatas inter-atômicas para uma molécula com n átomos, descrito no início desta seção. A nova abordagem é bem diferente da anterior: nesta, utiliza-se iterativas resoluções de um sistema linear 3×3. Tal algoritmo resolve o problema com complexidade linear, isto é, usando $\mathcal{O}(n)$ operações de ponto flutuante, segundo Dong e Wu em [5].

Antes de discuti-lo, faz-se necessária a demonstração de dois resultados imprescindíveis para o funcionamento do mesmo, descrito em [16].

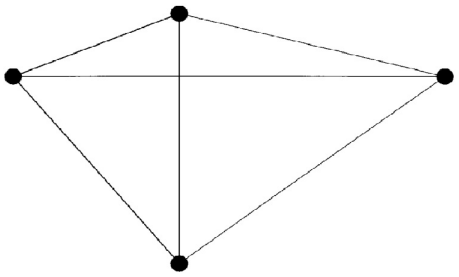

Figura 1.2: Quatro átomos da molécula com todas as distâncias conhecidas

Teorema 1. Em uma dada molécula, dadas as distâncias entre quatro átomos dois-a-dois, suas coordenadas podem ser determinadas, a menos de movimentos rígidos.

Demonstração. Sejam $d_{i,j}$ as distâncias conhecidas entre os átomos i e j, como na Figura 1.2, e $x_i = (u_i, v_i, w_i)^T$ os vetores de coordenadas de tais átomos, para $i, j = 1, ..., 4$. Tais vetores estão em um sistema de coordenadas tridimensional arbitrário. Realizando uma mudança de coordenadas por movimentos rígidos (rotação, translação e reflexão), as distâncias entre os átomos permanecem as mesmas no sistema novo em relação ao sistema anterior. Ou seja, as distâncias, neste caso, são invariantes a mudanças de coordenadas por movimentos rígidos da estrutura molecular.

Logo, sem perda de generalidade, considere a posição do primeiro átomo como a origem do sistema, a posição do segundo átomo como pertencente ao eixo das abscissas e, por fim, a

posição do terceiro átomo como pertencente ao plano abscissa-ordenada. A posição relativa ao quarto átomo pode se localizar em qualquer lugar no espaço. Neste sistema, os vetores de coordenadas dos três primeiros átomos são

$$x_1 = (0, 0, 0),$$
$$x_2 = (u_2, 0, 0),$$
$$x_3 = (u_3, v_3, 0),$$

com u_2, u_3 e v_3 a serem determinados. Já que os valores de $d_{i,j}$, para $i, j = 1, 2, 3$, são conhecidos, podemos definir, então, o sistema não-linear

$$\begin{cases} \|x_2 - x_1\|^2 = d_{2,1}^2 \\ \|x_3 - x_1\|^2 = d_{3,1}^2 \\ \|x_3 - x_2\|^2 = d_{3,2}^2 \end{cases}, \tag{1.12}$$

ou seja,

$$\begin{cases} u_2^2 = d_{2,1}^2 \\ u_3^2 + v_3^2 = d_{3,1}^2 \\ (u_3 - u_2)^2 + v_3^2 = d_{3,2}^2 \end{cases} \tag{1.13}$$

Resolvendo este sistema, encontramos os valores

$$u_2 = \pm d_{2,1}, \tag{1.14}$$

$$u_3 = \frac{d_{3,1}^2 - d_{3,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2} \tag{1.15}$$

e

$$v_3 = \pm (d_{3,1}^2 - u_3^2)^{1/2}, \tag{1.16}$$

podendo escolher tanto a parte positiva quanto a negativa para u_2 e para v_3 sem alterar as

distâncias entre os átomos.

Como o ponto x_4 está em algum lugar nosso sistema de coordenadas, com a única condição de respeitar as restrições de distâncias aos outros três átomos, então, escrevendo $x_4 = (u_4, v_4, w_4)^T$ e utilizando tais distâncias, definimos o seguinte sistema não-linear

$$\begin{cases} \|x_4 - x_1\|^2 = d_{4,1}^2 \\ \|x_4 - x_2\|^2 = d_{4,2}^2 \\ \|x_4 - x_3\|^2 = d_{4,3}^2 \end{cases} \tag{1.17}$$

Ou seja, temos o sistema

$$\begin{cases} u_4^2 + v_4^2 + w_4^2 = d_{4,1}^2 \\ (u_4 - u_2)^2 + v_4^2 + w_4^2 = d_{4,2}^2 \\ (u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2 = d_{4,3}^2 \end{cases} \tag{1.18}$$

Resolvendo (1.18), encontramos as coordenadas de x_4 em função, apenas, das distâncias inter-atômicas, isto é,

$$u_4 = \frac{d_{4,1}^2 - d_{4,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2}, \tag{1.19}$$

$$v_4 = \frac{d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2}{2v_3} + \frac{v_3}{2} \tag{1.20}$$

e

$$w_4 = \pm(d_{4,1}^2 - u_4^2 - v_4^2)^{1/2}. \tag{1.21}$$

podendo escolher em (1.21) tanto a parte positiva quanto negativa sem afetar as restrições de distâncias da hipótese do teorema.

Portanto, é possível determinar as coordenadas dos quatro átomos, a menos de movimentos rígidos, apenas em função das distâncias entre as posições dos átomos no $\mathbb{R}^3$, como queríamos.

□

Antes do próximo resultado, enunciaremos um lema técnico:

Lema 1. Se $\{x_1, x_2, x_3, x_4\} \subset \mathbb{R}^3$ são pontos não-coplanares, então o conjunto de vetores $B = \{(x_2 - x_1), (x_3 - x_1), (x_4 - x_1)\} \subset \mathbb{R}^3$ é linearmente independente.

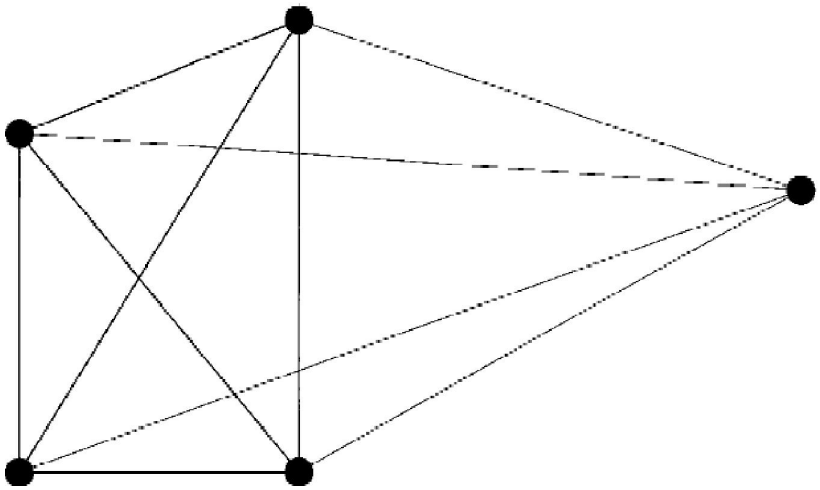

Figura 1.3: Conjunto com quatro átomos não-coplanares e um átomo indeterminado com todas as distâncias disponíveis

Teorema 2. Suponhamos que possuímos uma molécula com conjunto completo de distâncias exatas conhecido. Se temos quatro átomos fixos não-coplanares, então as coordenadas dos outros $n-4$ átomos de tal molécula podem ser determinadas unicamente através da resolução de sistemas lineares.

Demonstração. Por hipótese, temos quatro átomos já fixos, não-coplanares, cujas posições em $\mathbb{R}^3$ são x_1, x_2, x_3 e x_4 .

Chamaremos estes átomos de *átomos base* (eles serão definidos formalmente e serão mais amplamente usados na seção seguinte).

Seja, também, x_i as coordenadas desconhecidas de um quinto átomo, um dos $n-4$ átomos restantes (os átomos base e o átomo indeterminado estão representados na Figura 2.1-3).

Como conhecemos as distâncias entre o átomo i e os quatro átomos base, isto é, sabemos os valores das distâncias $d_{i,1}, d_{i,2}, d_{i,3}$ e $d_{i,4}$, temos o seguinte sistema de equações

$$\begin{cases} d_{i,1}^2 = \|x_i - x_1\|^2 = \|x_i\|^2 - 2x_i^T x_1 + \|x_1\|^2 \\ d_{i,2}^2 = \|x_i - x_2\|^2 = \|x_i\|^2 - 2x_i^T x_2 + \|x_2\|^2 \\ d_{i,3}^2 = \|x_i - x_3\|^2 = \|x_i\|^2 - 2x_i^T x_3 + \|x_3\|^2 \\ d_{i,4}^2 = \|x_i - x_4\|^2 = \|x_i\|^2 - 2x_i^T x_4 + \|x_4\|^2 \end{cases} \tag{1.22}$$

ou seja,

$$\begin{cases} \|x_i\|^2 - 2x_i^T x_1 + \|x_1\|^2 = d_{i,1}^2 \\ \|x_i\|^2 - 2x_i^T x_2 + \|x_2\|^2 = d_{i,2}^2 \\ \|x_i\|^2 - 2x_i^T x_3 + \|x_3\|^2 = d_{i,3}^2 \\ \|x_i\|^2 - 2x_i^T x_4 + \|x_4\|^2 = d_{i,4}^2 \end{cases} \tag{1.23}$$

Este sistema não-linear possui solução, já que os átomos existem e suas posições respeitam as restrições de distâncias.

Subtraindo a primeira equação do sistema não-linear (1.23) das outras três, temos o seguinte sistema linear

$$\begin{cases} -2x_i^T(x_2 - x_1) = (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ -2x_i^T(x_3 - x_1) = (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ -2x_i^T(x_4 - x_1) = (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{cases} . \tag{1.24}$$

Mas, como $v^T w = w^T v$, para $v, w \in \mathbb{R}^3$, então podemos reescrever o sistema (1.24) na forma

$$\begin{cases} -2(x_2 - x_1)^T x_i = (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ -2(x_3 - x_1)^T x_i = (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ -2(x_4 - x_1)^T x_i = (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{cases} . \tag{1.25}$$

Colocando o sistema (1.25) em forma matricial, temos

$$\begin{bmatrix} -2(x_2 - x_1)^T \\ -2(x_3 - x_1)^T \\ -2(x_4 - x_1)^T \end{bmatrix} \cdot x_i = \begin{bmatrix} (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{bmatrix}, \tag{1.26}$$

ou seja,

$$-2 \begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_1)^T \\ (x_4 - x_1)^T \end{bmatrix} \cdot x_i = \begin{bmatrix} (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{bmatrix}. \tag{1.27}$$

Logo, temos que

$$Ax_i = b_i, \tag{1.28}$$

onde

$$A = -2 \begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_1)^T \\ (x_4 - x_1)^T \end{bmatrix} \tag{1.29}$$

e

$$b_i = \begin{bmatrix} (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{bmatrix}. \tag{1.30}$$

Como os pontos x_1, x_2, x_3 e x_4 não são coplanares, por hipótese, segue do Lema 1 que $\{(x_2 - x_1), (x_3 - x_1), (x_4 - x_1)\}$ é um conjunto linearmente independente. Assim, a matriz

$$\begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_1)^T \\ (x_4 - x_1)^T \end{bmatrix}$$

tem posto completo, e, então, A é não-singular.

Portanto, temos que o sistema $Ax_i = b_i$ possui uma única solução x_i^*, isto é, conseguimos

encontrar uma única posição possível para o átomo em questão. Procedemos, iterativamente, como acima para todos os $n-4$ átomos ainda indeterminados, determinando, unicamente, a estrutura de nossa molécula. □

Observação 1. A hipótese do teorema anterior é a existência de um conjunto de quatro átomos não-coplanares e cujas posições são previamente determinadas. Se encontrarmos tal conjunto, mas não conhecermos suas coordenadas, podemos calculá-las com base no Teorema 1.

Estabelecidas as bases teóricas do método, segue um roteiro para tal algoritmo publicado em [16]:

Algoritmo de tempo linear para PMDGs
com conjunto completo de distâncias exatas

(1) Encontre quatro átomos não coplanares na molécula para serem átomos base. Determine suas coordenadas a partir das distâncias entre si;

(2) Para cada átomo remanescente, determine suas coordenadas através do sistema linear (1.28);

(Todos os átomos estão determinados.)

■

O próximo resultado também é discutido por Wu et al. em [16].

Teorema 3. Este algoritmo necessita de $\mathcal{O}(n)$ para determinar uma molécula com n átomos.

Demonstração. De fato. A cada iteração, resolvemos um sistema linear do tipo (1.28). Tal resolução pode ser feita em $\mathcal{O}(1)$ operações aritméticas de ponto flutuante. Como, precisamos resolver $n-4$ sistemas deste tipo, são necessárias $\mathcal{O}(n)$ operações de ponto flutuante.

Não são necessárias mais do que $\mathcal{O}(1)$ operações para calcular as posições dos átomos base. Além disso, no pior caso, serão necessárias $\mathcal{O}(n)$ operações para encontrar o terceiro átomo sem que este seja colinear com os dois primeiros e o quarto átomo que não seja coplanar com os outros três.

Portanto, o o problema pode ser resolvido através deste método em ordem de complexidade linear. □

Este método é mais eficiente do que o método anterior, que faz uso da decomposição SVD. Ambos são calculados em tempo polinomial, mas, o último, é calculado em ordem de complexidade linear, enquanto o primeiro é calculado em ordem de complexidade cúbica.

1.2 Conjunto Arbitrário

A segunda classe do PMGD a ser considerada neste trabalho é aquela cujo conjunto de distâncias exatas $d_{i,j}$ é arbitrário. Tanto podemos possuir um conjunto completo de distâncias quanto incompleto.

A primeira abordagem para resolver este problema é uma generalização do algoritmo cuja ordem de complexidade é linear, descrito na Seção 1.1.2. Entretanto, há um relaxamento nas condições e é possível considerar o problema com um conjunto incompleto de distâncias exatas.

Este novo algoritmo é chamado *Algoritmo Iterativo de Determinação Geométrica* (AIDG) e foi publicado no artigo [6] por Dong et al. Originalmente, o algoritmo foi denominado *Geometric Build-Up Algorithm*.

1.2.1 Algoritmo Iterativo de Determinação Geométrica

Para definir este algoritmo, usa-se, praticamente, as mesmas idéias do algoritmo com ordem de complexidade linear, comentado anteriormente. Algumas definições para a base teórica de tal método são necessárias, descritas em [20], baseadas em [1].

Definição 1. Chamamos de ***átomos posicionados*** os átomos que possuem as coordenadas em $\mathbb{R}^3$ já conhecidas em uma molécula. Os que não possuem posição conhecida são chamados de ***átomos não-posicionados***.

A idéia básica do AIDG é, essencialmente, adicionar um átomo à lista de átomos posicionados a cada iteração, partindo apenas das distâncias disponíveis entre eles.

No método com complexidade linear para o conjunto completo, precisa-se de quatro átomos não-coplanares para determinar todos os átomos ainda não-posicionados da molécula. Neste novo algoritmo, lança-se mão de quatro átomos não-coplanares para cada um dos átomos não-posicionados. Tal conjunto é definido a seguir.

Definição 2. (Base Métrica) Um conjunto de pontos B (cujas coordenadas são conhecidas) em um espaço (geralmente $\mathbb{R}^3$) é uma ***base métrica*** para um conjunto S de pontos se o sistema de coordenadas de qualquer ponto de S é unicamente determinado pelas distâncias conhecidas deste ponto aos pontos de B.

Em tal método, busca-se quatro átomos cujas posições formem uma base métrica para o conjunto que contém apenas a posição do átomo indeterminado de cada iteração.

Definição 3. (Átomos Base) Aos átomos cujas posições formem a base métrica chamaremos de ***átomos base***.

Definição 4. (Conjunto Independente) Um conjunto de quatro pontos em $\mathbb{R}^3$ é chamado ***independente*** se são não-coplanares.

Definição 5. (Ponto Vizinho) Um ponto P é chamado de ***ponto vizinho*** do ponto Q se a distância entre eles, $d(P, Q)$ é conhecida.

De posse dessas definições, temos o seguinte resultado:

Teorema 4. Quaisquer quatro átomos, cujos pontos são posicionados, independentes (em $\mathbb{R}^3$) e vizinhos do ponto de um átomo não-posicionado na molécula, formam uma base métrica para ele. A partir desta base métrica, o átomo não-posicionado pode ser determinado unicamente resolvendo um sistema linear.

Demonstração. Sejam x_i um átomo não-posicionado em nossa molécula e x_1, x_2, x_3 e x_4 átomos posicionados cujos pontos são independentes e vizinhos de x_i. Ou seja, os pontos x_1, x_2, x_3 e x_4 são posicionados, não-coplanares e as distâncias $d_{i,1}, d_{i,2}, d_{i,3}$ e $d_{i,4}$ são conhecidas.

Definimos as equações

$$\begin{cases} \|x_i - x_1\| = d_{i,1} \\ \|x_i - x_2\| = d_{i,2} \\ \|x_i - x_3\| = d_{i,3} \\ \|x_i - x_4\| = d_{i,4} \end{cases} \tag{1.31}$$

Elevando-as ao quadrado, temos

$$\begin{cases} d_{i,1}^2 = \|x_i - x_1\|^2 = \|x_i\|^2 - 2x_i^T x_1 + \|x_1\|^2 \\ d_{i,2}^2 = \|x_i - x_2\|^2 = \|x_i\|^2 - 2x_i^T x_2 + \|x_2\|^2 \\ d_{i,3}^2 = \|x_i - x_3\|^2 = \|x_i\|^2 - 2x_i^T x_3 + \|x_3\|^2 \\ d_{i,4}^2 = \|x_i - x_4\|^2 = \|x_i\|^2 - 2x_i^T x_4 + \|x_4\|^2 \end{cases} \tag{1.32}$$

Segue que

$$\begin{cases} \|x_i\|^2 - 2x_i^T x_1 + \|x_1\|^2 = d_{i,1}^2 \\ \|x_i\|^2 - 2x_i^T x_2 + \|x_2\|^2 = d_{i,2}^2 \\ \|x_i\|^2 - 2x_i^T x_3 + \|x_3\|^2 = d_{i,3}^2 \\ \|x_i\|^2 - 2x_i^T x_4 + \|x_4\|^2 = d_{i,4}^2 \end{cases} \tag{1.33}$$

Subtraindo a primeira equação do sistema não-linear (1.33) das outras três, temos o seguinte sistema linear

$$\begin{cases} -2x_i^T(x_2 - x_1) = (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ -2x_i^T(x_3 - x_1) = (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ -2x_i^T(x_4 - x_1) = (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{cases} . \tag{1.34}$$

Mas, como $v^T w = w^T v$, para $v, w \in \mathbb{R}^3$, então podemos reescrever o sistema (1.34) na forma

$$\begin{cases} -2(x_2 - x_1)^T x_i = (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ -2(x_3 - x_1)^T x_i = (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ -2(x_4 - x_1)^T x_i = (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{cases} . \tag{1.35}$$

Colocando o sistema (1.35) em forma matricial, temos

$$\begin{bmatrix} -2(x_2 - x_1)^T \\ -2(x_3 - x_1)^T \\ -2(x_4 - x_1)^T \end{bmatrix} \cdot x_i = \begin{bmatrix} (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{bmatrix}, \tag{1.36}$$

ou seja,

$$-2 \begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_1)^T \\ (x_4 - x_1)^T \end{bmatrix} \cdot x_i = \begin{bmatrix} (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{bmatrix}. \tag{1.37}$$

Logo, temos o sistema linear matricial

$$Ax_i = b_i, \tag{1.38}$$

onde

$$A = -2 \begin{bmatrix} (x_2 - x_1)^T \\ (x_3 - x_1)^T \\ (x_4 - x_1)^T \end{bmatrix} \tag{1.39}$$

e

$$b_i = \begin{bmatrix} (\|x_1\|^2 - \|x_2\|^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (\|x_1\|^2 - \|x_3\|^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (\|x_1\|^2 - \|x_4\|^2) - (d_{i,1}^2 - d_{i,4}^2) \end{bmatrix}. \tag{1.40}$$

Como x_1, x_2, x_3 e x_4 são não-coplanares, segue que os vetores $x_2 - x_1$, $x_3 - x_1, x_4 - x_1$ são linearmente independentes, pelo Lema 1. Assim, A é não-singular.

Portanto, x_i pode ser determinado unicamente como solução do sistema linear (1.38) e $\{x_1, x_2, x_3, x_4\}$ é base métrica para $\{x_i\}$. □

Segue um esboço do Algoritmo Iterativo de Determinação Geométrica para resolver o

Problema Molecular de Geometria de Distâncias com um conjunto arbitrário de distâncias exatas, como publicado em [6].

AIDG para PMGDs com conjunto arbitrário de distâncias

(1) Encontre quatro átomos não-coplanares e vizinhos entre si. Determine suas coordenadas a partir de suas distâncias segundo o Teorema 1;

(2) Repita:

Para cada átomo remanescente

(i) Encontre quatro átomos posicionados, vizinhos e independentes;

(ii) Resolva o sistema (1.38), usando a matriz dos coeficientes que possua o *maior determinante*, para encontrar a posição do átomo indeterminado.

Se nenhum átomo é determinado em todo o loop, pare.

Fim.

■

No primeiro passo do algoritmo acima, procura-se quatro átomos não-coplanares e vizinhos entre si. Procede-se como no caso completo e encontramos as posições dos quatro átomos com uma mudança de coordenadas, como descrito no Teorema 1.

No segundo passo, como a molécula possui n átomos e as coordenadas de quatro deles já foram encontradas, trabalha-se com os outros $n-4$ átomos.

Seja $x_i \in \mathbb{R}^3$ a posição a ser obtida de um dos $n-4$ átomos indeterminados. A partir disso, busca-se quatro átomos determinados, não-coplanares e vizinhos dele cujas posições sejam x_1, x_2, x_3 e x_4. Tais posições formam o conjunto $B_i = \{x_1, x_2, x_3, x_4\}$ que servirá de base métrica para o conjunto $\{x_i\}$, segundo o Teorema 4.

Dessa forma, é possível estabelecer as equações

$$\begin{cases} d_{i,1}^2 = \|x_i - x_1\|^2 = \|x_i\|^2 - 2x_i^T x_1 + \|x_1\|^2 \\ d_{i,2}^2 = \|x_i - x_2\|^2 = \|x_i\|^2 - 2x_i^T x_2 + \|x_2\|^2 \\ d_{i,3}^2 = \|x_i - x_3\|^2 = \|x_i\|^2 - 2x_i^T x_3 + \|x_3\|^2 \\ d_{i,4}^2 = \|x_i - x_4\|^2 = \|x_i\|^2 - 2x_i^T x_4 + \|x_4\|^2 \end{cases} \tag{1.41}$$

Consideremos a seguinte proposição.

Proposição 2. Vinte-e-quatro sistemas lineares diferentes podem ser derivados deste sistema não-linear de equações.

Demonstração. Inicialmente, escolhe-se uma das quatro equações a qual se chamará *equação-pivô*. Subtrai-se, então, a equação-pivô das outras três equações remanescentes encontrando um sistema linear nas incógnitas x_{i1}, x_{i2}, x_{i3}. Como é possível escolher quatro equações-pivôs diferentes, então encontramos quatro sistemas lineares diferentes nas incógnitas x_{i1}, x_{i2}, x_{i3}.

Além disso, para cada um desses quatro sistemas lineares, existem seis permutações de equações, as quais geram seis sistemas lineares distintos.

Portanto, é possível derivar vinte-e-quatro sistemas lineares nas incógnitas x_{i1}, x_{i2}, x_{i3} a partir do sistema não-linear (1.41). □

Por exemplo, fixamos a quarta equação de (1.41) como equação-pivô. Assim, a subtraimos das outras três obtendo o sistema

$$\begin{cases} -2x_i^T(x_1 - x_4) = (\|x_4\|^2 - \|x_1\|^2) - (d_{i,4}^2 - d_{i,1}^2) \\ -2x_i^T(x_2 - x_4) = (\|x_4\|^2 - \|x_2\|^2) - (d_{i,4}^2 - d_{i,2}^2) \\ -2x_i^T(x_3 - x_4) = (\|x_4\|^2 - \|x_3\|^2) - (d_{i,4}^2 - d_{i,3}^2) \end{cases} . \tag{1.42}$$

Podemos reescrever o sistema (1.42) na forma

$$\begin{cases} -2(x_1 - x_4)^T x_i = (\|x_4\|^2 - \|x_1\|^2) - (d_{i,4}^2 - d_{i,1}^2) \\ -2(x_2 - x_4)^T x_i = (\|x_4\|^2 - \|x_2\|^2) - (d_{i,4}^2 - d_{i,2}^2) \\ -2(x_3 - x_4)^T x_i = (\|x_4\|^2 - \|x_3\|^2) - (d_{i,4}^2 - d_{i,3}^2) \end{cases} . \tag{1.43}$$

Rearranjando este sistema em forma matricial, temos

$$Ax_i = b_i, \tag{1.44}$$

onde

$$A = -2 \begin{bmatrix} (x_1 - x_4)^T \\ (x_2 - x_4)^T \\ (x_3 - x_4)^T \end{bmatrix}$$

e

$$b_i = \begin{bmatrix} (\|x_4\|^2 - \|x_1\|^2) - (d_{i,4}^2 - d_{i,1}^2) \\ (\|x_4\|^2 - \|x_2\|^2) - (d_{i,4}^2 - d_{i,2}^2) \\ (\|x_4\|^2 - \|x_3\|^2) - (d_{i,4}^2 - d_{i,3}^2) \end{bmatrix} .$$

Assumimos, inicialmente, que x_1, x_2, x_3 e x_4 são átomos não-coplanares. Segue, do lema 1, que $\{(x_1 - x_4), (x_2 - x_4), (x_3 - x_4)\}$ é um conjunto de vetores linearmente independentes. Portanto, a matriz A é não-singular e, logo, o sistema $Ax_i = b_i$ possui uma única solução x_i^*.

A partir de qualquer um dos vinte-e-quatro sistemas lineares, que podemos derivar do sistema não-linear (1.41), é possível determinar as coordenadas do átomo não-posicionado em questão de modo único. Entretanto, alguns erros de arredondamento podem aparecer como, por exemplo, na resolução do sistema linear eleito.

Dong et al. propoem um critério em [6] para escolher o sistema linear adequado a ser derivado do sistema não-linear (1.41) a fim de minimizar tais erros de arredondamento. Para

que haja estabilidade numérica na resolução do sistema linear, é proposto que a matriz de coeficientes A esteja o mais distante possível de uma matriz singular. Logo, dentre as vinte-e-quatro possibilidades de extrair um sistema linear a partir do sistema não-linear citado acima, escolhe-se aquela cujo sistema tenha a matriz de coeficientes A com o maior valor absoluto do determinante.

Algumas observações se fazem necessárias sobre este critério.

Observação 2. Notemos que não é necessário escolher dentre as vinte-e-quatro matrizes aquela cujo módulo do determinante é o maior, mas, apenas, entre quatro delas. Isto segue imediatamente do fato de que se $P, A \in \mathbb{R}^{n\times n}$, de modo que P é uma matriz de permutação, então

$$|det(P \cdot A)| = |det(P) \cdot det(A)| = |det(P)| \cdot |det(A)| = |det(A)|, \tag{1.45}$$

já que P é uma matriz ortogonal e, logo, $det(P) = \pm 1$.

■

Considere o seguinte teorema, demonstrado em [4].

Teorema 5. Seja A não-singular. Então,

$$min\left\{\frac{\|\delta A\|_2}{\|A\|_2} : A + \delta A \ \ singular\right\} = \frac{1}{\mathcal{K}(A)}.$$

Logo, a distância para a matriz singular mais próxima é $d = \dfrac{1}{\mathcal{K}(A)}$.

Observação 3. Considere a matriz

$$A = \begin{bmatrix} 10^3 & 0 \\ 0 & 10^{-1} \end{bmatrix}.$$

Temos que

$$det(A) = 10^2 \text{ e } \mathcal{K}(A) = 10^4.$$

Como o número de condição de A é grande, então ela está bem próxima de uma matriz singular, pelo teorema enunciado anteriormente, e o valor absoluto de seu determinante é grande.

Logo, o critério proposto por Dong et al. não é tão efetivo já que podemos encontrar matrizes com determinantes grandes, mas mal-condicionadas. Ou seja, mesmo que tal critério seja satisfeito, podemos nos deparar com sistemas com matrizes próximas de singulares que introduzam grandes erros prejudicando a determinação da estrutura.

■

Neste processo iterativo para resolver o PMDG com dados arbitrários de distâncias, um erro foi constatado e está citado em [6]. Em algum momento na execução do algoritmo, quatro vetores determinados são escolhidos como átomos base. Estes foram determinados pelo mesmo processo em iterações anteriores e carregam consigo erros numéricos. Quando tais átomos forem utilizados na determinação de outro, estes erros numéricos serão passados adiante. Ou seja, é possível ter uma propagação de erros de grandes proporções de modo a não permitir que a precisão desejada na determinação de vários átomos seja alcançada. Tais erros numéricos podem ser oriundos de um eventual mal-condicionamento da matriz dos coeficiente do sistema linear utilizado na iteração. Utilizar o determinante para tentar corrigir este problema resolve alguns casos particulares testados no artigo, como descrito por Dong e Wu em [6], mas não é um artifício genericamente eficaz. A observação anterior cita um exemplo de matriz mal-condicionada com módulo do determinante bem grande.

A cada iteração, as coordenadas de um átomo não-posicionado são determinadas. Tal procedimento avança até que a estrutura de toda a molécula seja determinada ou, então, até que nenhum átomo seja determinado por falta de dados necessários para o loop. Não há garantias de que qualquer PMGD seja resolvido através deste método. Em cada loop, o algoritmo requer que, pelo menos, um dos átomos não-fixados possa ser determinado usando

os átomos fixados. De outra forma, o algoritmo vai parar e vai imprimir uma estrutura parcial composta das coordenadas dos átomos descobertas até então.

A mesma idéia do algoritmo para conjunto completo é utilizada aqui. Entretanto, agora, o conjunto de átomos base não é fixo durante todo o funcionamento do algoritmo e é reajustado a cada iteração. Melhor dito, a cada vez que requeremos um átomo não-determinado, o algoritmo procura quatro átomos não-coplanares cujas distâncias entre eles estejam disponiveis para funcionarem como átomos base e, assim, o algoritmo funciona como o outro.

Se o conjunto de distâncias for completo, o AIDG resolve o problema em ordem de complexidade linear. Agora, para o caso arbitrário, temos o seguinte teorema:

Teorema 6. Suponhamos que quaisquer quatro átomos iniciais possam nos conduzir para a determinação da estrutura molecular pelo AIDG. Então, no pior caso, serão necessárias $\mathcal{O}(n^3)$ operações aritméticas para tal determinação.

Demonstração. São necessárias $\mathcal{O}(n)$ operações aritméticas, no máximo, para encontrar quais átomos indeterminados podem ser posicionados através dos átomos previamente determinados. Para cada um destes, por sua vez, serão necessários $\mathcal{O}(n)$ operações para encontrar os quatro átomos que servirão de base métrica no pior caso. Cada sistema linear tem ordem de complexidade constante.

Além disso, temos n átomos a serem determinados, no máximo.

Portanto, em caso extremo, será preciso $\mathcal{O}(n^3)$ operações aritméticas de ponto flutuante para determinar toda a estrutura da molécula através do AIDG. □

Algumas linhas para a demonstração do teorema acima pode ser encontrada em [20], de Wu et al.

Capítulo 2

Algoritmo Iterativo de Determinação Geométrica Atualizado

Neste capítulo, vamos descrever e analisar o chamado *Algoritmo Iterativo de Determinação Geométrica Atualizado (AIDGA)* para a resolução de Problemas Moleculares de Geometria de Distâncias cujo conjunto de distâncias exatas entre os átomos da molécula é arbitrário. Este método foi proposto por Wu et al. em [16] com o nome de *Updated Geometric Build-up Algorithm*. É chamado *atualizado*, pois é uma versão modificada do AIDG geral, enunciado na Seção 1.2.1.

Para cada átomo não-posicionado, busca-se uma base métrica, cujos quatro átomos determinados e independentes sejam não-coplanares, de modo que tais átomos unidos com o átomo indeterminado formem um tetraedro T [20], assim como no AIDG. Como principal modificação em relação ao método anterior, realiza-se a mudança de coordenadas (translação, rotação, reflexão) no tetraedro T, descrita no Teorema 1. É possível, então, calcular as coordenadas do átomo não-posicionado de forma independente de cálculos anteriores, usando as distâncias originais como instâncias do problema.

Tal modificação tem como objetivo controlar a propagação de erros na descoberta de

novas coordenadas para os átomos indeterminados. Já que sempre é necessário recalcular as posições dos átomos da base métrica no início de cada iteração, o átomo a ser calculado não vai receber os erros numéricos provenientes de cálculos prévios.

Depois da mudança de coordenadas, o tetraedro T será recolocado em sua estrutura original, alinhando as novas coordenadas e as antigas da melhor forma de modo que a "distância" (da qual falaremos a seguir) entre elas seja mínima. Para tanto, translada-se os cinco átomos (de cada uma) para que as duas estruturas tenham o mesmo centro geométrico e, então, realiza-se uma rotação para encaixá-las de modo ótimo. Procedendo assim, os erros de arredondamento nas posições dos cinco átomos são minimizados simultaneamente. Tal procedimento é denominado *reinicialização* das coordenadas dos átomos do tetraedro.

Em resumo, o AIDGA possui dois passos principais. Primeiro, as posições dos quatro átomos base são recalculadas, baseado no Teorema 1. Suas novas posições são completamente independentes de suas posições antigas. Tal passo assegura que os quatro átomos base formem um tetraedro onde as distâncias inter-atômicas são as mais exatas possível. Segundo, o vetor de translação e a matriz de rotação para a reinicialização devem ser encontrados.

Vamos explicar o processo de reinicialização para o tetraedro quando todas as distâncias entre os quatro átomos estão disponíveis, assim como feito por Davis, Ernst e Wu em [20].

Sejam x_1, x_2, x_3 e x_4 os vetores de coordenadas dos quatro átomos base, previamente determinados, para o posicionamento do átomo indeterminado x_k, e $d_{i,j}$ a distância entre os átomos i e j, para $i, j = 1, 2, 3, 4$. Temos que $x_i = (x_{i1}, x_{i2}, x_{i3})^T$, $i = 1, 2, 3, 4$.

O primeiro passo é recalcular as coordenadas dos átomos da base métrica utilizando apenas as distâncias entre eles, assim como feito no Teorema 1. Para tanto, realiza-se uma mudança de coordenadas (translação, rotação, reflexão) colocando o primeiro átomo na origem do novo sistema de coordenadas, o segundo átomo no eixo das abscissas e, o terceiro, no plano abscissa-ordenada. Estas novas posições serão chamadas de y_1, y_2, y_3 e y_4, onde

- $y_1 = (0, 0, 0)^T$

- $y_2 = (u_2, 0, 0)^T$
- $y_3 = (u_3, v_3, 0)^T$
- $y_4 = (u_4, v_4, w_4)^T$.

A partir das distâncias entre eles, define-se os sistemas de equações

$$\begin{cases} \|y_2 - y_1\| = d_{2,1} \\ \|y_3 - y_1\| = d_{3,1} \\ \|y_3 - y_2\| = d_{3,2} \end{cases} \tag{2.1}$$

e

$$\begin{cases} \|y_4 - y_1\| = d_{4,1} \\ \|y_4 - y_2\| = d_{4,2} \\ \|y_4 - y_3\| = d_{4,3} \end{cases} \tag{2.2}$$

Elevando as equações de ambos os sistemas ao quadrado, temos

$$\begin{cases} d_{2,1}^2 = \|y_2 - y_1\|^2 = u_2^2 \\ d_{3,1}^2 = \|y_3 - y_1\|^2 = u_3^2 + v_3^2 \\ d_{3,2}^2 = \|y_3 - y_2\|^2 = (u_3 - u_2)^2 + v_3^2 \end{cases} \tag{2.3}$$

e

$$\begin{cases} d_{4,1}^2 = \|y_4 - y_1\|^2 = u_4^2 + v_4^2 + w_4^2 \\ d_{4,2}^2 = \|y_4 - y_2\|^2 = (u_4 - u_2)^2 + v_4^2 + w_4^2 \\ d_{4,3}^2 = \|y_4 - y_3\|^2 = (u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2 \end{cases} \tag{2.4}$$

Ou seja,

$$\begin{cases} u_2^2 = d_{2,1}^2 \\ u_3^2 + v_3^2 = d_{3,1}^2 \\ (u_3 - u_2)^2 + v_3^2 = d_{3,2}^2 \end{cases} \tag{2.5}$$

e

$$\begin{cases} u_4^2 + v_4^2 + w_4^2 = d_{4,1}^2 \\ (u_4 - u_2)^2 + v_4^2 + w_4^2 = d_{4,2}^2 \\ (u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2 = d_{4,3}^2 \end{cases} \tag{2.6}$$

Resolvendo o sistema não-linear (2.5), temos

$$u_2 = \pm d_{2,1}, \tag{2.7}$$

$$u_3 = \frac{d_{3,1}^2 - d_{3,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2} \tag{2.8}$$

e

$$v_3 = \pm\sqrt{d_{3,1}^2 - u_3^2}, \tag{2.9}$$

podendo optar tanto pela parte positiva quanto pela parte negativa para u_2 e v_3 sem alterar as restrições de distâncias.

A partir do sistema não-linear (2.6), podemos calcular as coordenadas de y_4

$$u_4 = \frac{d_{4,1}^2 - d_{4,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2}, \tag{2.10}$$

$$v_4 = \frac{d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2}{2v_3} + \frac{v_3}{2}, \tag{2.11}$$

e

$$w_4 = \pm\sqrt{d_{4,1}^2 - u_4^2 - v_4^2}, \tag{2.12}$$

podendo escolher entre as partes negativa e positiva de w_4, pois as distâncias serão preservadas.

Segue, do Teorema 4, que é possível calcular as coordenadas de x_k no novo sistema de

coordenadas de modo único.

Sejam X e $Y \in \mathbb{R}^{4\times 3}$ as matrizes de coordenadas correspondentes aos vetores de coordenadas x_i e y_i, para $i = 1, 2, 3, 4$, respectivamente. Isto é, tais matrizes são definidas por

$$\begin{cases} X(i,j) = x_i(j) \\ Y(i,j) = y_i(j) \end{cases} \tag{2.13}$$

onde $i = 1, 2, 3, 4$ e $j = 1, 2, 3$.

Os centros geométricos dessas duas estruturas, representadas pelas matrizes X e Y, são dados pelos vetores xc e yc, respectivamente, definidos por

$$xc(j) = \frac{1}{4}\sum_{i=1}^{4} X(i,j)$$

e

$$yc(j) = \frac{1}{4}\sum_{i=1}^{4} Y(i,j),$$

para $j = 1, 2, 3$.

Redefinindo as matrizes X e Y(chamando-as, agora, de $X^{(1)}$ e $Y^{(1)}$, respectivamente) por meio das translações, temos

$$X^{(1)} = X - v \cdot xc^T,$$

e

$$Y^{(1)} = Y - v \cdot yc^T,$$

onde $v = (1, 1, 1, 1)^T$. Ou seja, os elementos das matrizes $X^{(1)}$ e $Y^{(1)}$ são dados pelas expressões

$$X^{(1)}(i,j) = X(i,j) - xc(j),$$

e

$$Y^{(1)}(i,j) = Y(i,j) - yc(j),$$

para $j = 1, 2, 3$.

Segue, então, a proposição:

Proposição 3. As estruturas representadas pelas matrizes $X^{(1)}$ e $Y^{(1)}$ têm a origem como mesmo centro geométrico.

Demonstração. Temos que

$$X^{(1)}(i,j) = X(i,j) - \frac{1}{4}\sum_{k=1}^{4} X(k,j)$$

e

$$Y^{(1)}(i,j) = Y(i,j) - \frac{1}{4}\sum_{k=1}^{4} Y(k,j),$$

para $i = 1, ..., 4$ e $j = 1, .., 3$.

Calculando os centros geométricos $xc_{(1)}$ e $yc_{(1)}$, de $X^{(1)}$ e $Y^{(1)}$, respectivamente, temos que

$$xc_{(1)}(j) = \frac{1}{4}\sum_{i=1}^{4} X^{(1)}(i,j) = \frac{1}{4}\sum_{i=1}^{4} (X(i,j) - xc(j))$$

e

$$yc_{(1)}(j) = \frac{1}{4}\sum_{i=1}^{4} Y^{(1)}(i,j) = \frac{1}{4}\sum_{i=1}^{4} (Y(i,j) - yc(j)),$$

para cada $j = 1, 2, 3$.

Ou seja,

$$xc_{(1)}(j) = \frac{1}{4}\sum_{i=1}^{4} X^{(1)}(i,j) = \frac{1}{4}\sum_{i=1}^{4} X(i,j) - \frac{1}{4}\sum_{i=1}^{4} xc(j)$$

e

$$yc_{(1)}(j) = \frac{1}{4}\sum_{i=1}^{4} Y^{(1)}(i,j) = \frac{1}{4}\sum_{i=1}^{4} Y(i,j) - \frac{1}{4}\sum_{i=1}^{4} yc(j),$$

para $j = 1, 2, 3$.

Logo,

$$xc_{(1)}(j) = xc(j) - \frac{1}{4} \cdot 4xc(j) = xc(j) - xc(j) = 0$$

e

$$yc_{(1)}(j) = yc(j) - \frac{1}{4} \cdot 4yc(j) = yc(j) - yc(j) = 0,$$

para $j = 1, 2, 3$.

Portanto, $yc_{(1)} = xc_{(1)} = 0$, como queríamos. □

Por fim, fazendo $X = X^{(1)}$ e $Y = Y^{(1)}$, realiza-se translação em X e em Y de modo que fiquem sobrepostos e tenham o mesmo centro geométrico: a origem.

Para encontrar a matriz de rotação que "melhor alinha" as duas estruturas, já transladadas, uma nova ferramenta é introduzida [20].

RMSD (*Root-Mean-Square Deviation*) é uma medida da distância média entre as estruturas de duas moléculas sobrepostas. O modo mais comum de comparar duas estruturas biomoleculares é aplicar uma translação, para que os centros geométricos coincidam, seguida da rotação, de uma estrutura com relação à outra, de modo que o RMSD entre elas seja minimizado.

Definição 6 (RMSD). Sejam X e Y duas matrizes em $\mathbb{R}^{n\times 3}$ representando duas estruturas sobrepostas com mesmo centro geométrico. Definimos o RMSD entre X e Y por

$$\text{RMSD}(X, Y) = \frac{min_Q \|X - YQ\|_F}{\sqrt{n}},$$

onde Q é uma matriz de rotação e $\|\cdot\|_F$ é a norma de Frobënius.

A unidade de distância comumente usada para esse desvio é o Ângstrom (denotado por Å). Temos que $1\text{Å} = 10^{-10}$ metros.

Em nosso caso, $n = 4$ pois vamos reinicializar, primeiramente, os átomos base.

Após as translações descritas anteriormente, aplica-se uma rotação de Y em relação a X, descrita pela matriz de rotação Q, de modo a minimizar o valor do RMSD. Logo, é preciso encontrar encontrar uma matriz ortogonal de rotação Q_0 ($Q_0^T Q_0 = I$) tal que

$$\frac{\|X - YQ_0\|_F}{\sqrt{4}} = \frac{min_Q \|X - YQ\|_F}{\sqrt{4}},$$

onde Q é ortogonal.

A matriz que minimiza $\frac{\|X - YQ\|_F}{\sqrt{4}}$ é a matriz que minimiza $\|X - YQ\|_F$. Assim, basta resolver o problema

$$\begin{aligned} &\min && \|X - YQ\|_F \\ &\text{s.a. } && QQ^T = I \end{aligned}$$

a fim de encontrar a matriz de rotação necessária para a reinicialização dos átomos base.

Sabemos que

$$\|C\|_F^2 = tr(C^T C) = tr(CC^T),$$

para $C \in \mathbb{R}^{m \times n}$.

Logo, pela linearidade da função traço,

$$\begin{aligned} \|X - YQ\|_F^2 &= tr((X - YQ)^T (X - YQ)) = tr((X^T - Q^T Y^T)(X - YQ)) \\ &= tr(X^T X - X^T YQ - Q^T Y^T X + Q^T Y^T YQ) \\ &= tr(X^T X) + tr((YQ)^T YQ) - tr(X^T YQ) - tr(Q^T Y^T X) \\ &= tr((X^T X) + tr(YQ(YQ)^T) - tr(X^T YQ) - tr(Q^T Y^T X) \\ &= tr(X^T X) + tr(YQQ^T Y^T) - tr((Q^T Y^T X)^T) - tr(Q^T Y^T X) \\ &= tr(X^T X) + tr(YY^T) - tr(Q^T Y^T X) - tr(Q^T Y^T X) \\ &= tr(X^T X) + tr(Y^T Y) - 2tr(Q^T Y^T X), \end{aligned}$$

isto é,

$$\|X - YQ\|_F^2 = tr(X^T X) + tr(Y^T Y) - 2tr(Q^T Y^T X).$$

Assim, resolver o problema

$$\begin{aligned} &\min && \|X - YQ\|_F \\ &\text{s.a. } && QQ^T = I \end{aligned}$$

é o mesmo que resolver

$$\begin{aligned} \min \quad & -tr(Q^T Y^T X), \\ \text{s.a. } & QQ^T = I \end{aligned}$$

ou seja, resolver

$$\begin{aligned} \max \quad & tr(Q^T Y^T X), \\ \text{s.a. } & QQ^T = I \end{aligned}$$

já que X e Y são matrizes fixas.

Tomemos $C = Y^T X$. Seja $C = U\Sigma V^T$ sua decomposição em valores singulares, onde U, V são ortogonais e $\Sigma = diag(\sigma_1, \sigma_2, \sigma_3)$, com $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$. Definimos, assim, a matriz $Z = V^T Q^T U$ que é ortogonal, já que consiste em um produto de matrizes ortogonais.

Agora, temos que

$$tr(Q^T Y^T X) = tr(Q^T U\Sigma V^T) = tr((Q^T U\Sigma)V^T) = tr(V^T(Q^T U\Sigma)) = tr(V^T Q^T U\Sigma) = tr(Z\Sigma).$$

Além disso,

$$z_{i1}^2 + z_{i2}^2 + z_{i3}^2 = 1, \tag{2.14}$$

para $i = 1, 2, 3$.

Então,

$$z_{ii}^2 \leq 1, \tag{2.15}$$

ou seja,

$$-1 \leq z_{ii} \leq 1, \tag{2.16}$$

para $i = 1, 2, 3$.

Segue que

$$0 < z_{ii} \leq 1 \tag{2.17}$$

e, assim, que

$$z_{ii}\sigma_i \leq \sigma_i, \tag{2.18}$$

para $i = 1, 2, 3$.

Portanto,

$$tr(Z\Sigma) = \sum_{i=1}^{3} z_{ii}\sigma_i \leq \sum_{i=1}^{3} \sigma_i,$$

isto é,

$$tr(Q^T Y^T X) \leq \sum_{i=1}^{3} \sigma_i.$$

Agora, se $Q = UV^T$, então

$$Z = V^T Q^T U = V^T (UV^T)^T U = V^T V U^T U = I,$$

ou seja,

$$tr(Z\Sigma) = \sum_{i=1}^{3} \sigma_i.$$

Portanto, $Q = UV^T$ resolve o problema

$$\begin{aligned} &\min && \|X - YQ\|_F, \\ &\text{s.a.} && QQ^T = I \end{aligned}$$

já que é a matriz ortogonal que maximiza $tr(Q^T Y^T X)$.

Podemos, então, enunciar tal resultado como o seguinte teorema:

Teorema 7. Para $X, Y \in \mathbb{R}^{n\times 3}$, a matriz ortogonal $Q = UV^T$ resolve o problema

$$\begin{aligned} &\min && \|X - YQ\|_F, \\ &\text{s.a.} && QQ^T = I \end{aligned}$$

onde U e V são matrizes ortogonais da decomposição em valores singulares $Y^T X = U\Sigma V^T$.

Segue, do teorema 7, que

$$\mathrm{RMSD}(X,Y) = \frac{min_Q \|X - YQ\|_F}{\sqrt{4}} = \frac{\|X - YQ_0\|_F}{\sqrt{4}},$$

onde $Q_0 = UV^T$, sendo U e V as matrizes ortogonais relativas à decomposição SVD da matriz Y^TX.

Considerando a base métrica $B' = \{y_1, y_2, y_3, y_4\}$, deve-se determinar o vetor de coordenadas x_k que indica a posição do átomo indeterminado referente à base $B = \{x_1, x_2, x_3, x_4\}$. Seja $x \in \mathbb{R}^3$ tal que

$$\begin{cases} d_{k,1} = \|x - y_1\| \\ d_{k,2} = \|x - y_2\| \\ d_{k,3} = \|x - y_3\| \\ d_{k,4} = \|x - y_4\| \end{cases} \tag{2.19}$$

onde $d_{k,j}$ é a distância entre o átomo indeterminado e o j-ésimo átomo da base métrica B. Assim,

$$\begin{cases} d_{k,1}^2 = \|x - y_1\|^2 = \|x\|^2 - 2x^Ty_1 + \|y_1\|^2 \\ d_{k,2}^2 = \|x - y_2\|^2 = \|x\|^2 - 2x^Ty_2 + \|y_2\|^2 \\ d_{k,3}^2 = \|x - y_3\|^2 = \|x\|^2 - 2x^Ty_3 + \|y_3\|^2 \\ d_{k,4}^2 = \|x - y_4\|^2 = \|x\|^2 - 2x^Ty_4 + \|y_4\|^2 \end{cases} \tag{2.20}$$

A partir do sistema não-linear (2.20), da mesma forma como feito no AIDG, pode-se derivar vinte-e-quatro sistemas lineares. Outra modificação no AIDGA é proposta em [16], em relação ao que foi proposto no AIDG: examinar o número de condição em norma-2 da matriz, ao invés de analisar o determinante no momento de escolher qual sistema linear resolver a cada iteração. Quando o número de condição é grande, um conjunto de átomos base diferente é procurado de modo a evitar possíveis erros devidos a uma matriz mal-condicionada. Esta modificação deve-se a uma constatação: o fato do determinante ser grande, não exclui necessariamente a possibilidade de estarmos trabalhando com uma matriz mal-condicionada. Um exemplo disto foi dado no capítulo anterior.

Observação 4. Neste método, também deve-se avaliar o número de condição das matrizes de coeficientes de tão somente quatro sistemas lineares deduzidos do sistema não-linear que modela o problema. Ou seja, não é necessário avaliar as vinte-e-quatro possibilidades.

De fato: sejam $P, A \in \mathbb{R}^{n\times n}$ de modo que P é uma matriz de permutação. Então,

$$\mathcal{K}_2(P \cdot A) = \mathcal{K}_2(A), \tag{2.21}$$

já que P é uma matriz ortogonal.

■

Considerando estas modificações, encontra-se o sistema linear

$$-2\begin{bmatrix}(y_{j_2}-y_{j_1})^T\\(y_{j_3}-y_{j1})^T\\(y_{j_4}-y_{j_1})^T\end{bmatrix} x = \begin{bmatrix}(\|y_{j_1}\|^2-\|y_{j_2}\|^2)-(d_{k,j_1}^2-d_{k,j_2}^2)\\(\|y_{j_1}\|^2-\|y_{j_3}\|^2)-(d_{k,j_1}^2-d_{k,j_3}^2)\\(\|y_{j_1}\|^2-\|y_{j_4}\|^2)-(d_{k,j_1}^2-d_{k,j_4}^2)\end{bmatrix}, \tag{2.22}$$

onde $\{j_1, j_2, j_3, j_4\} = \sigma(\{1, 2, 3, 4\})$ e σ é uma permutação.

Após obter a solução do sistema (2.22), reinicializa-se o vetor v encontrando a posição de x_k de forma ótima segundo as técnicas da RMSD. Primeiramente, translada-se v para que esteja no mesmo sistema da matriz Y transladada, que possui centro geométrico na origem, fazendo

$$v = v - yc.$$

Depois, aplica-se a rotação Q^T ao vetor v de solução do sistema anterior já transladado

$$v = Q^T v.$$

Por fim, translada-se v para o sistema original de coordenadas segundo o vetor de translação xc, fazendo

$$x_k = v + xc.$$

Segue um roteiro do processo de reinicialização a ser feito a cada iteração do AIDGA.

Reinicialização

(1) Armazenar as coordenadas antigas da base métrica na matriz $X \in \mathbb{R}^{4\times 3}$ e as novas coordenadas na matriz $Y \in \mathbb{R}^{4\times 3}$.

(2) Determinar os vetores xc e yc correspondentes aos centros geométricos de X e Y, respectivamente.

(3) Transladar X e Y segundo xc e yc, respectivamente.

(4) Calcular a Decomposição em Valores Singulares $Y^T X = U\Sigma V^T$ e computar a matriz de rotação $Q = UV^T$.

(5) Fazer a translação $x = x - yc$, sendo x a solução do sistema linear (2.22) nas novas coordenadas da base métrica.

(6) Fazer a rotação $x = Q^T x$.

(7) Fazer a translação $x_k = x + xc$ para encontrar a posição do átomo indeterminado da iteração.

■

A busca por quatro átomos da base métrica, pode necessitar de $\mathcal{O}(n^4)$ passos, fazendo com que o algoritmo necessite de $\mathcal{O}(n^6)$ passos para completar sua execução ([19], p.176).

Desse modo, temos um resumo do AIDGA:

AIDGA para PMGDs com conjunto arbitrário de distâncias exatas

(1) Encontre quatro átomos determinados, vizinhos entre si e não-coplanares e determine suas coordenadas segundo o Teorema 1.

(2) Repita:

Para cada átomo indeterminado

(i) Encontre quatro átomos vizinhos a ele, independentes e determinados a servir de base métrica.

(ii) Resolva o sistema linear (2.22) com os átomos base transladados segundo o Teorema 1.

(iii) Reinicialize os cinco átomos e coloque-os na estrutura inicial.

Se nenhum átomo é determinado em todo o loop, pare.

Fim.

■

Capítulo 3

Algoritmo T

Neste capítulo, vamos apresentar uma outra abordagem para resolver um PMGD com conjunto arbitrário de distâncias exatas. Este algoritmo tem grande semelhanças com o AIDG descrito em [6], já que seu passo mais importante é resolver um sistema linear a partir do conhecimento de quatro átomos vizinhos e não-coplanares e suas distâncias ao átomo indeterminado. A principal diferença é que, ao invés de resolvermos um sistema 3×3, utilizamos um sistema 4×4. O chamamos de Algoritmo T.

Suponhamos, então, que temos uma molécula com n átomos e possuimos um conjunto de distâncias arbitrário entre pares de átomos. Queremos encontrar suas coordenadas $x_1, ..., x_n$ a partir dessas distâncias, $d_{i,j}$, entre os átomos i e j já conhecidas.

Para definirmos as bases teóricas deste método, considere os seguintes teoremas.

Teorema 8. Se $\{x_1, x_2, x_3, x_4\} \subset \mathbb{R}^3$ é um conjunto de pontos não-coplanares, então o conjunto de vetores $\{x_1 - x_4, x_2 - x_4, x_3 - x_4\}$ é linearmente independente.

Teorema 9. Se $\{x_1, x_2, x_3, x_4\} \subset \mathbb{R}^3$ é um conjunto de pontos não-coplanares, então a matriz

$$B = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ 1 & x_3^T \\ 1 & x_4^T \end{bmatrix}$$

é não-singular.

Demonstração: Considere os vetores $v_i = \begin{bmatrix} 1 & x_i^T \end{bmatrix}^T$, para $i = 1, 2, 3, 4$. Vamos mostrar que o conjunto $\{v_1, v_2, v_3, v_4\} \subset \mathbb{R}^4$ é linearmente independente.

De fato: seja

$$\alpha v_1 + \beta v_2 + \gamma v_3 + \theta v_4 = 0$$

uma combinação linear do vetor nulo de $\mathbb{R}^4$. Logo, temos as relações

$$\begin{bmatrix} \alpha + \beta + \gamma + \theta \\ \alpha x_1 + \beta x_2 + \gamma x_3 + \theta x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

ou seja, temos o sistema

$$\begin{cases} \alpha + \beta + \gamma + \theta = 0 \\ \alpha x_1 + \beta x_2 + \gamma x_3 + \theta x_4 = 0 \end{cases}. \tag{3.1}$$

Isto é,

$$\begin{cases} \alpha + \beta + \gamma = -\theta \\ \alpha x_1 + \beta x_2 + \gamma x_3 = -\theta x_4 \end{cases}. \tag{3.2}$$

Então

$$\alpha x_1 + \beta x_2 + \gamma x_3 = (\alpha + \beta + \gamma) x_4 = \alpha x_4 + \beta x_4 + \gamma x_4. \tag{3.3}$$

De (3.3), segue que

$$0 = (\alpha x_1 - \alpha x_4) + (\beta x_2 - \beta x_4) + (\gamma x_3 - \gamma x_4) = \alpha(x_1 - x_4) + \beta(x_2 - x_4) + \gamma(x_3 - x_4). \tag{3.4}$$

Por hipótese, x_1, x_2, x_3 e x_4 são não-coplanares. Assim, pelo Teorema 8, temos que $x_1 - x_4, x_2 - x_4$ e $x_3 - x_4$ são vetores linearmente independentes. Segue, em (3.4), que $\alpha = \beta = \gamma = 0$. Substituindo α, β e γ na primeira equação do sistema (3.1), temos que $\theta = 0$.

Então, $\{v_1, v_2, v_3, v_4\}$ é um conjunto de vetores linearmente independente e, portanto, a matriz

$$B = \begin{bmatrix} v_1^T \\ v_2^T \\ v_3^T \\ v_4^T \end{bmatrix} \in \mathbb{R}^{4\times 4}$$

tem posto-completo, como queríamos.

■

Teorema 10. Sejam $\{b_1, b_2, b_3, b_4\}$ e $\{y_1, y_2, y_3, y_4\}$ subconjuntos de $\mathbb{R}^3$ e $\mathbb{R}$, respectivamente. Se o sistema quadrático

$$\begin{cases} \|a - b_1\| = y_1 \\ \|a - b_2\| = y_2 \\ \|a - b_3\| = y_3 \\ \|a - b_4\| = y_4 \end{cases} \tag{3.5}$$

possui uma solução a^*, então o sistema

$$Ax = b, \tag{3.6}$$

onde

$$A = -2 \begin{bmatrix} 1 & b_1^T \\ 1 & b_2^T \\ 1 & b_3^T \\ 1 & b_4^T \end{bmatrix} \text{ e } b = \begin{bmatrix} y_1^2 - \|b_1\|^2 \\ y_2^2 - \|b_2\|^2 \\ y_3^2 - \|b_3\|^2 \\ y_4^2 - \|b_4\|^2 \end{bmatrix},$$

possui uma solução x^* em função de a^*.

Demonstração: Seja $a^* \in \mathbb{R}^3$ uma solução para o sistema (3.5). Assim, aplicando a^* às equações deste sistema e elevando-as ao quadrado, temos

$$\begin{cases} \|a^*\|^2 - 2b_1^T a^* + \|b_1\|^2 = y_1^2 \\ \|a^*\|^2 - 2b_2^T a^* + \|b_2\|^2 = y_2^2 \\ \|a^*\|^2 - 2b_3^T a^* + \|b_3\|^2 = y_3^2 \\ \|a^*\|^2 - 2b_4^T a^* + \|b_4\|^2 = y_4^2 \end{cases}, \tag{3.7}$$

ou seja,

$$\begin{cases} \|a^*\|^2 - 2b_1^T a^* = y_1^2 - \|b_1\|^2 \\ \|a^*\|^2 - 2b_2^T a^* = y_2^2 - \|b_2\|^2 \\ \|a^*\|^2 - 2b_3^T a^* = y_3^2 - \|b_3\|^2 \\ \|a^*\|^2 - 2b_4^T a^* = y_4^2 - \|b_4\|^2 \end{cases} \tag{3.8}$$

Fazendo $t = -\dfrac{\|a^*\|^2}{2}$ e substituindo-o em (3.8), temos

$$\begin{cases} -2t - 2b_1^T a^* = y_1^2 - \|b_1\|^2 \\ -2t - 2b_2^T a^* = y_2^2 - \|b_2\|^2 \\ -2t - 2b_3^T a^* = y_3^2 - \|b_3\|^2 \\ -2t - 2b_4^T a^* = y_4^2 - \|b_4\|^2 \end{cases} \tag{3.9}$$

Matricialmente, podemos escrever (3.9) como

$$\begin{bmatrix} -2 & -2b_1^T \\ -2 & -2b_2^T \\ -2 & -2b_3^T \\ -2 & -2b_4^T \end{bmatrix} \begin{bmatrix} t \\ a^* \end{bmatrix} = \begin{bmatrix} y_1^2 - \|b_1\|^2 \\ y_2^2 - \|b_2\|^2 \\ y_3^2 - \|b_3\|^2 \\ y_4^2 - \|b_4\|^2 \end{bmatrix}. \tag{3.10}$$

Segue, então, que

$$-2\begin{bmatrix} 1 & b_1^T \\ 1 & b_2^T \\ 1 & b_3^T \\ 1 & b_4^T \end{bmatrix}\begin{bmatrix} t \\ a^* \end{bmatrix} = \begin{bmatrix} y_1^2 - \|b_1\|^2 \\ y_2^2 - \|b_2\|^2 \\ y_3^2 - \|b_3\|^2 \\ y_4^2 - \|b_4\|^2 \end{bmatrix}. \tag{3.11}$$

Portanto, $x^* = \begin{bmatrix} t & a^{*T} \end{bmatrix}^T$ é uma solução para o sistema (3.6) em função da solução a^* do sistema (3.5), como queríamos.

■

Teorema 11. Sejam $B = \{b_1, b_2, b_3, b_4\}$ e $Y = \{y_1, y_2, y_3, y_4\}$ subconjuntos de $\mathbb{R}^3$ e $\mathbb{R}$, respectivamente, de modo que B é um conjunto de pontos não-coplanares. Então, o sistema

$$Ax = b, \tag{3.12}$$

onde

$$A = -2\begin{bmatrix} 1 & b_1^T \\ 1 & b_2^T \\ 1 & b_3^T \\ 1 & b_4^T \end{bmatrix} \text{ e } b = \begin{bmatrix} y_1^2 - \|b_1\|^2 \\ y_2^2 - \|b_2\|^2 \\ y_3^2 - \|b_3\|^2 \\ y_4^2 - \|b_4\|^2 \end{bmatrix},$$

possui solução única.

Demonstração: Como os pontos de B são não-coplanares, pelo Teorema 9, concluímos que a matriz A é não-singular. Portanto, tal sistema linear possui solução única.

■

Sejam x_j a posição de um átomo a ser determinado de nossa molécula e x_1, x_2, x_3 e x_4 as posições de seus átomos base, ou seja, átomos determinados, não-coplanares e vizinhos de

tal átomo. A partir das distâncias $d_{j,1}, d_{j,2}, d_{j,3}$ e $d_{j,4}$, entre x_j e os átomos x_1, x_2, x_3 e x_4, podemos considerar o sistema quadrático

$$\begin{cases} \|x_j - x_1\| = d_{j,1} \\ \|x_j - x_2\| = d_{j,2} \\ \|x_j - x_3\| = d_{j,3} \\ \|x_j - x_4\| = d_{j,4} \end{cases} \tag{3.13}$$

Este sistema modela nosso problema que é encontrar as coordenadas x_j.

Pelo Teorema 10, a solução x_j^* desse sistema integra a solução do sistema

$$Ax = b, \tag{3.14}$$

onde

$$A = -2 \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ 1 & x_3^T \\ 1 & x_4^T \end{bmatrix} \text{ e } b = \begin{bmatrix} d_{j,1}^2 - \|x_1\|^2 \\ d_{j,2}^2 - \|x_1\|^2 \\ d_{j,3}^2 - \|x_1\|^2 \\ d_{j,4}^2 - \|x_1\|^2 \end{bmatrix}.$$

Além disso, o sistema (3.14) tem solução e esta é única, pelo Teorema 11. Logo, uma pergunta vem naturalmente ao centro de nossa discussão: será que possuindo tal solução, então nosso problema de encontrar as coordenadas do átomo indeterminado pode ser resolvido partindo dela? Tal resposta vem através do seguinte teorema.

Teorema 12. Suponhamos que $\{x_1, x_2, x_3, x_4\}$ é um conjunto de pontos determinados não-coplanares vizinhos do ponto indeterminado x_j.

Se o sistema não-linear

$$\begin{cases} \|x_j - x_1\| = d_{j,1} \\ \|x_j - x_2\| = d_{j,2} \\ \|x_j - x_3\| = d_{j,3} \\ \|x_j - x_4\| = d_{j,4} \end{cases} \tag{3.15}$$

admite solução única x_j^*, então $x^* = \begin{bmatrix} t_j & x_j^{*T} \end{bmatrix}^T$, onde $t_j = -\dfrac{\|x_j^*\|^2}{2}$, é solução única do sistema

$$Ax = b, \tag{3.16}$$

com

$$A = -2 \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ 1 & x_3^T \\ 1 & x_4^T \end{bmatrix} \text{ e } b = \begin{bmatrix} d_{j,1}^2 - \|x_1\|^2 \\ d_{j,2}^2 - \|x_2\|^2 \\ d_{j,3}^2 - \|x_3\|^2 \\ d_{j,4}^2 - \|x_4\|^2 \end{bmatrix}.$$

Demonstração: Por hipótese, x_1, x_2, x_3 e x_4 são posições de átomos determinados, não-coplanares e vizinhos do átomo indeterminado x_j. Então, o sistema (3.15) tem solução única, segundo desenvolvido neste mesmo trabalho nas páginas 17 - 20 e baseado em [6]. Além disso, pelo Teorema 11, o sistema (3.16) possui solução única também.

Suponhamos que x_j^* é a única solução do sistema (3.15). Segue, do Teorema 10, que o vetor $x^* = \begin{bmatrix} t_j & x_j^{*T} \end{bmatrix}^T$ é a solução do sistema (3.16), com $t_j = -\dfrac{\|x_j^*\|^2}{2}$. Como queríamos demonstrar.

■

Este resultado nos dá uma direção a seguir a fim de resolver o Problema Molecular de Geometria de Distâncias com um conjunto de distâncias exatas arbitrário. Para encontrar as coordenadas dos átomos indeterminados, é preciso resolver o sistema (3.15), que é não-linear.

O Teorema 12 nos diz que uma forma equivalente de resolver tal problema é encontrarmos a solução de um sistema linear do tipo $Ax_j = b$, onde

$$A = -2 \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ 1 & x_3^T \\ 1 & x_4^T \end{bmatrix} \text{ e } b = \begin{bmatrix} d_{j,1}^2 \\ d_{j,2}^2 \\ d_{j,3}^2 \\ d_{j,4}^2 \end{bmatrix},$$

para cada átomo j ainda desconhecido da proteína.

Determinamos, assim, a posição do átomo não-posicionado fazendo $x_j(i) = x(i+1)$, para $i = 1, 2, 3$. Portanto, podemos responder à nossa pergunta feita acima: sim, é possível resolver um PMGD partindo do sistema (3.14) .

Segue um esboço do Algoritmo T (AT):

Algoritmo T para PMGDs com conjunto arbitrário de distâncias

(1) Encontre quatro átomos base não-coplanares; determine as coordenadas dos átomos base com as distâncias entre si;

(2) Repita:

   (i) Para cada átomo remanescente, encontre quatro átomos determinados, não-coplanares e vizinhos como átomos base.

   (ii) Encontra-se a solução x do sistema linear 4×4 (3.16);

   (iii) Temos, então, que as coordenadas do átomo indeterminado x_j da iteração são dadas por $x_j(i) = x(i+1)$, para $i = 1, 2, 3$.

   Se nenhum átomo é determinado em todo o loop, pare.

   Fim.

■

Observação 5. A priori, existem vinte e quatro permutações a serem feitas no sistema linear (3.16) que geram sistemas lineares equivalentes a ele. Ou seja, temos vinte e quatro sistemas lineares para resolvermos a fim de encontrar tal posição.

Um critério razoável para escolher o sistema a ser resolvido é analisar o número de condição de suas matrizes de coeficientes e escolher aquele que possui o menor. Queremos, assim, resolver um sistema que seja equivalente ao original e o mais estável possível.

Mas, como o número de condição de todas as matrizes de coeficientes dos vinte e quatro sistemas são iguais, pela Observação 4, então não há porque resolver mais de um sistema linear.

Ter apenas um sistema linear a ser resolvido é bem melhor do que ter de escolher entre quatro sistemas, como ocorre tanto no AIDG como no AIDGA.

■

A complexidade deste algoritmo é a mesma do AIDG. Precisamos de, no máximo, uma ordem de n operações para encontrar os átomos ainda não-posicionados que podem ser encontrados a partir do conjunto de átomos cujas coordenadas já são conhecidas. Para cada um desses átomos, serão necessárias $\mathcal{O}(n)$ operações para encontrar sua base métrica. A partir desta, resolvemos um sistema linear cuja ordem de complexidade é constante. Como temos n átomos para determinar, em um caso extremo, então segue o resultado.

Teorema 13. *São necessárias $\mathcal{O}(n^3)$ operações aritméticas de ponto flutuante para que o Algoritmo T resolva um Problema Molecular de Geometria de Distâncias com conjunto arbitrário de distâncias exatas.*

Capítulo 4

O Algoritmo T e comparações

Neste capítulo, vamos comparar o Algoritmo T com o Algoritmo Iterativo de Determinação Geométrica.

Para tal comparação, geramos instâncias artificiais com valores fixos de modo que a distância entre um átomo e o quarto subsequente a ele esteja abaixo de 6Å ao longo de toda a estrutura. Algumas definições se fazem necessárias antes de enunciar a regra de formação delas.

Definição 7 (Comprimento de Ligação). O comprimento da ligação entre dois átomos é a distância euclidiana entre as posições deles no espaço tridimensional.

Definição 8 (Ângulo de Torsão). Um Ângulo de Torsão, ou Ângulo Diedral, consiste no ângulo entre dois planos, isto é, o ângulo formado entre os vetores normais a estes planos.

Em nosso caso, o Ângulo de Torsão é o ângulo entre os planos que contém os átomos base, não-coplanares por definição.

Definição 9 (Ângulo de Ligação). Considerando as posições de três átomos da molécula como vértices de duas ligações, o ângulo de ligação consiste no ângulo plano formado entre os vetores que representam tridimensionalmente essas ligações.

A estrutura artificial gerada para fazer a comparação entre os algoritmos é uma cadeia de átomos estabelecidas em uma ordem de 1 a n, sendo n a quantidade de átomos presentes na estrutura. Seguem as regras:

(i) Os comprimentos de ligação entre os átomos consecutivos $i-1$ e i, denotados por $d_{i-1,i}$, foram fixos valendo 1.5Å cada, para $i \in \{2, 3, ..., n\}$.

(ii) Os ângulos planos das ligações entre os átomos $i-2, i-1$ e i, com o vértice em $i-1$, denotados por $\theta_{i-2,i}$, foram fixos em $\frac{2\pi}{3}$, para $i \in \{3, 4, ..., n\}$.

(iii) Os ângulos de torsão para cada conjunto de átomos $i-3, i-2, i-1$ e i, denotados por $\omega_{i-3,i}$, variam aleatoriamente entre os valores $\frac{\pi}{3}, \frac{\pi}{2}$ e $\frac{5\pi}{3}$, para $i \in \{4, ..., n\}$.

Tais instâncias foram motivadas por sua simplicidade de implementação e por representarem, minimamente, as características de um problema real [11].

Na figura seguinte, temos a representação de um conjunto com os átomos $i-3, i-2, i-1$ e i da estrutura artificial definida acima. O que sabemos, a priori, são as distâncias entre os átomos consecutivos, os ângulos planos e os ângulos de torsão, definidos acima.

Desse modo, vamos gerar as posições dos n átomos de nossa estrutura a partir do produto matricial, que está enunciada em [10], páginas 5 e 6,

$$\begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ 1 \end{bmatrix} = B_1 B_2 B_3 \dots B_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{4.1}$$

para $i \in \{4, ..., n\}$, onde (x_{i1}, x_{i2}, x_{i3}) é a posição do i-ésimo átomo, $B_1 = I_4$,

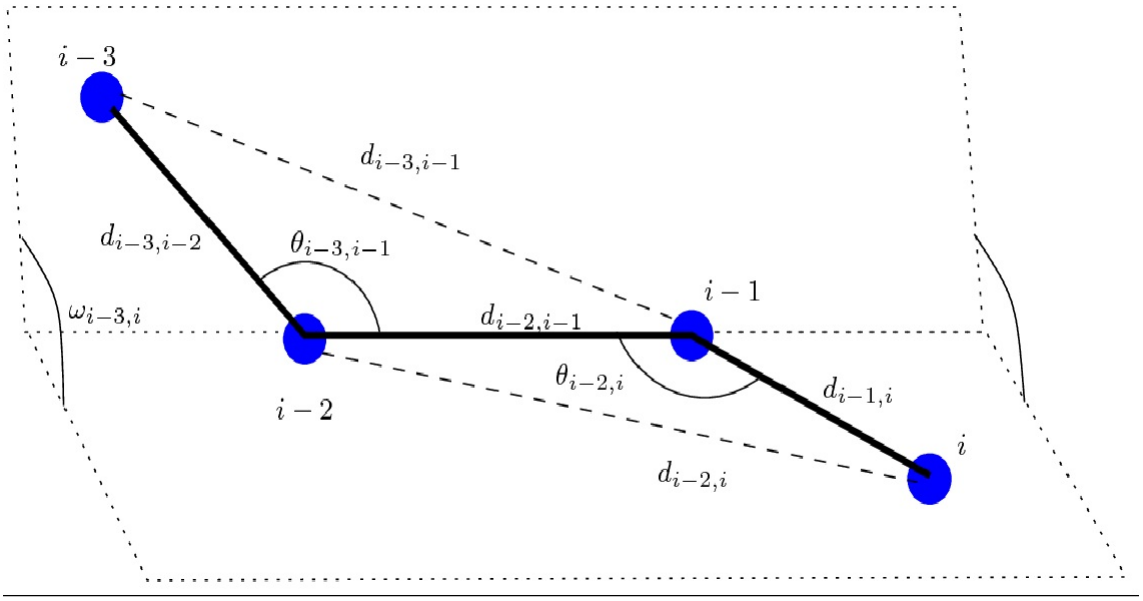

Figura 4.1: Representação de uma sequência de quatro átomos da estrutura artificial

$$B_2 = \begin{bmatrix} -1 & 0 & 0 & -1.5\mathring{\text{A}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; B_3 = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 & \frac{3}{4}\mathring{\text{A}} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 & \frac{3\sqrt{3}}{4}\mathring{\text{A}} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ e}$$

$$B_i = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 & \frac{3}{4}\mathring{\text{A}} \\ \frac{\sqrt{3}}{2}\cos\omega_{i-3,i} & \frac{1}{2}\cos\omega_{i-3,i} & -\operatorname{sen}\omega_{i-3,i} & \frac{3\sqrt{3}}{4}\mathring{\text{A}}\cos\omega_{i-3,i} \\ \frac{\sqrt{3}}{2}\operatorname{sen}\omega_{i-3,i} & \frac{1}{2}\operatorname{sen}\omega_{i-3,i} & \cos\omega_{i-3,i} & \frac{3\sqrt{3}}{4}\mathring{\text{A}}\operatorname{sen}\omega_{i-3,i} \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

para $i \in \{4, ..., n\}$.

Após terminar a geração da estrutura artificial, calculamos as distâncias entre os átomos e desprezamos aquelas que forem estritamente maiores do que $6\mathring{\text{A}}$. As distâncias que sobram são as instâncias que forneceremos como dados de entrada para ambos os algoritmos a fim de avaliar sua performance: tanto em tempo quanto em precisão.

4.1 AT versus AIDG

Fizemos dez experimentos computacionais gerados aleatóriamente. Testamos trinta e seis dimensões diferentes em cada um dos experimentos, com n variando de 5 a 40. Uma observação pertinente é a de que, para cada dimensão em cada experimento, testamos a mesma estrutura artificial tanto no AIDG quanto no AT.

A seguir apresentamos gráficos com os resultados dos dez experimentos. Cada figura possui dois gráficos onde as curvas azuis representam o Algoritmo T e, as vermelhas, o Algoritmo Iterativo de Determinação Geométrica.

Considerando todas as dimensões, os gráficos contidos nas primeiras dez figuras mostram:

(i) Os desempenhos da RMSD para os dois algoritmos, no primeiro gráfico.

(ii) Os tempos respectivos necessários para que os dois algoritmos determinem as estruturas, no segunqdo gráfico.

As outras dez figuras são semelhantes às dez primeiras. A diferença é que consideramos apenas até à quantidade de quinze átomos de cada experimento. Fizemos isto a fim de mostrar diferenças locais nas primeiras quantidades de átomos, levando em conta que a escala cresce consideravelmente até completar as quantidades. É interessante observar os crescimentos dos erros no início do processo, quando o mal-condicionamento das matrizes ainda é relativamente pequeno.

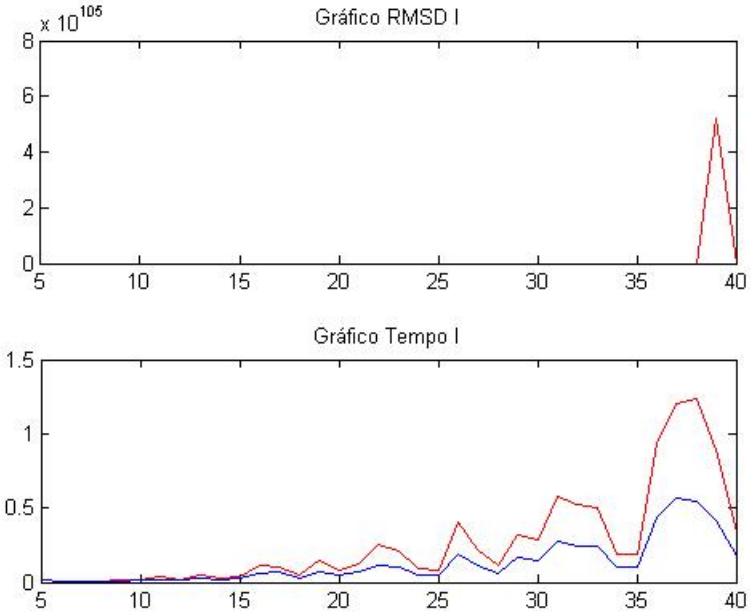

Figura 4.2: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o primeiro experimento.

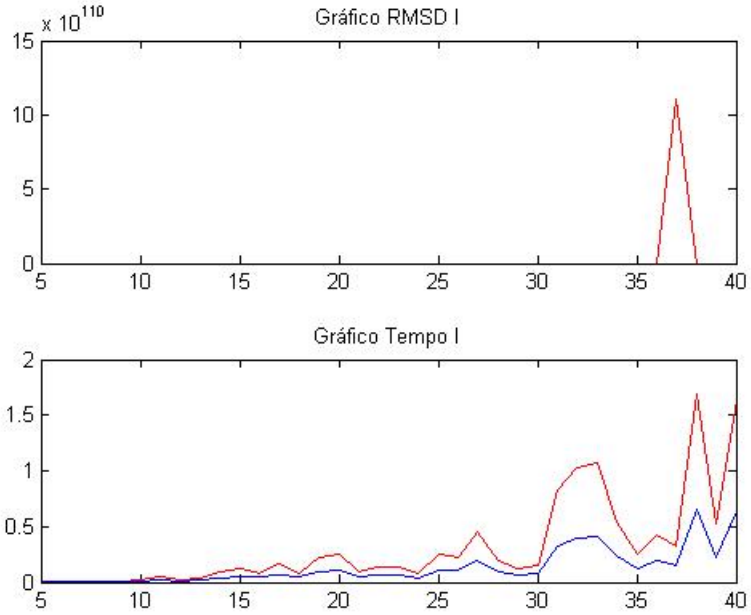

Figura 4.3: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o segundo experimento.

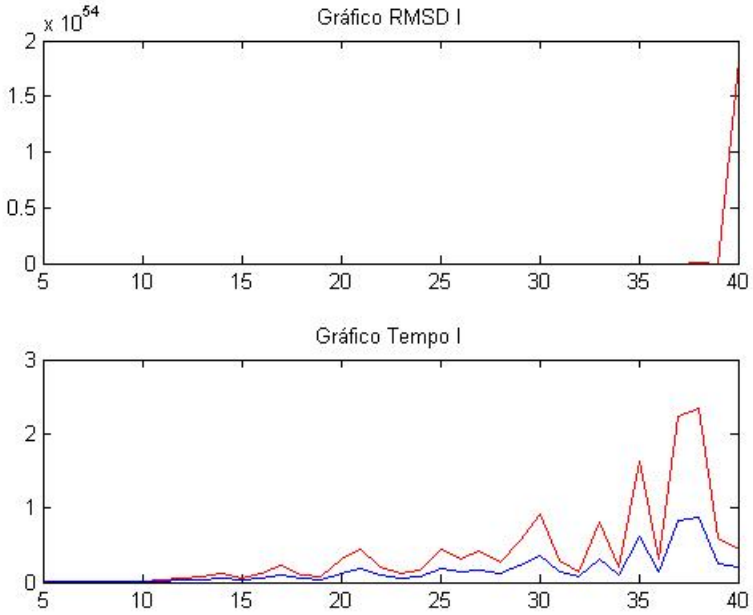

Figura 4.4: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o terceiro experimento.

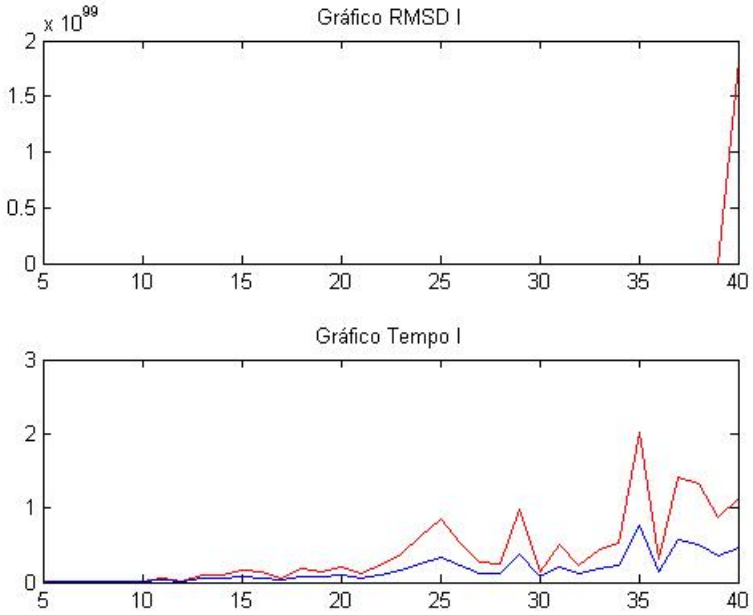

Figura 4.5: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o quarto experimento.

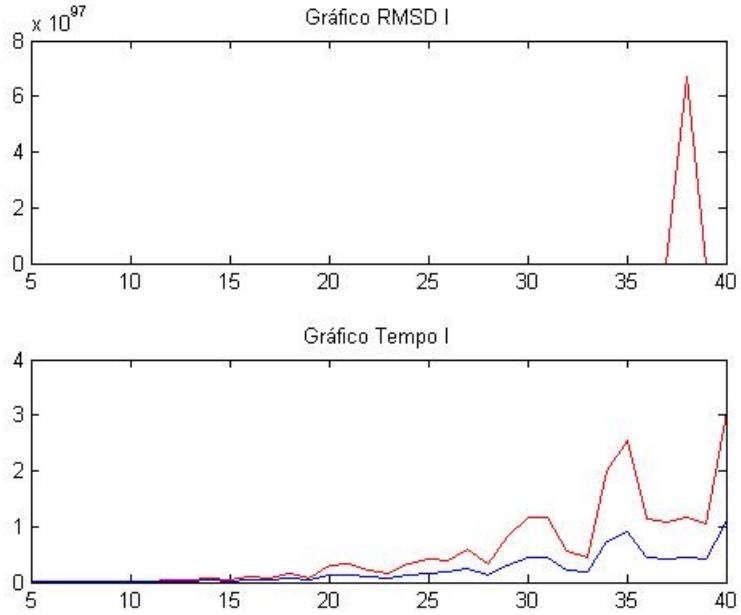

Figura 4.6: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o quinto experimento.

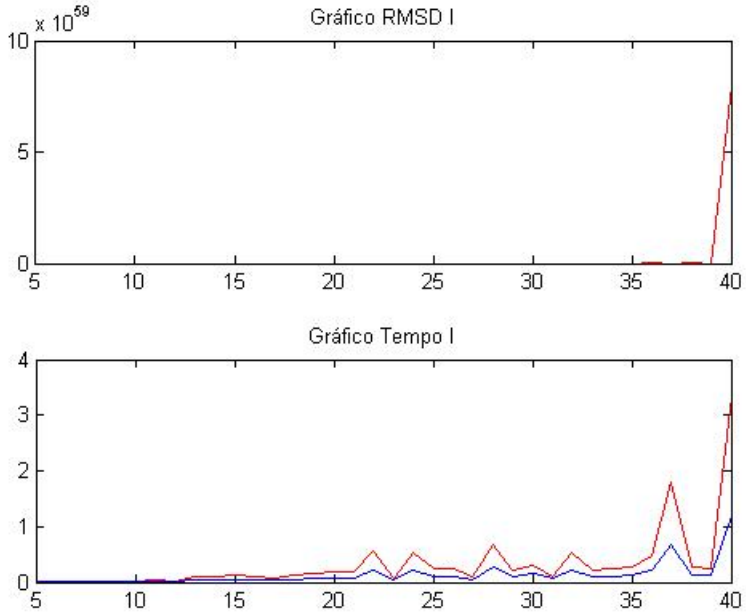

Figura 4.7: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o sexto experimento.

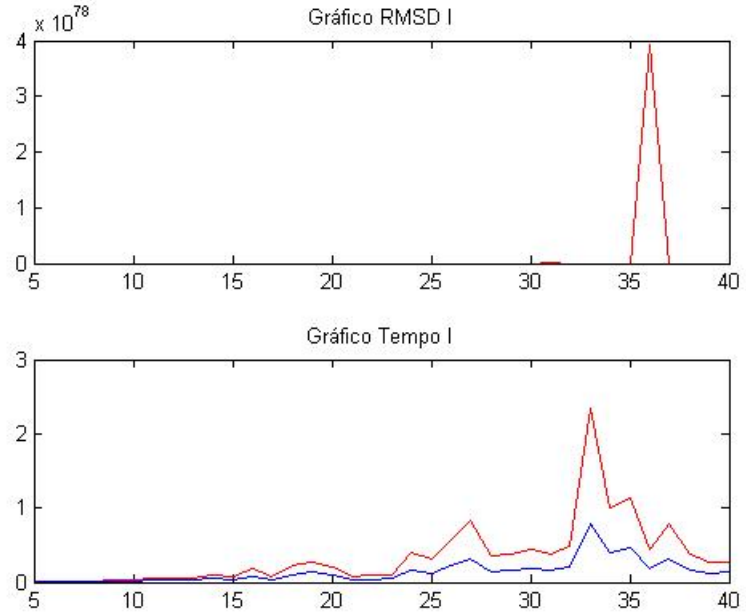

Figura 4.8: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o sétimo experimento.

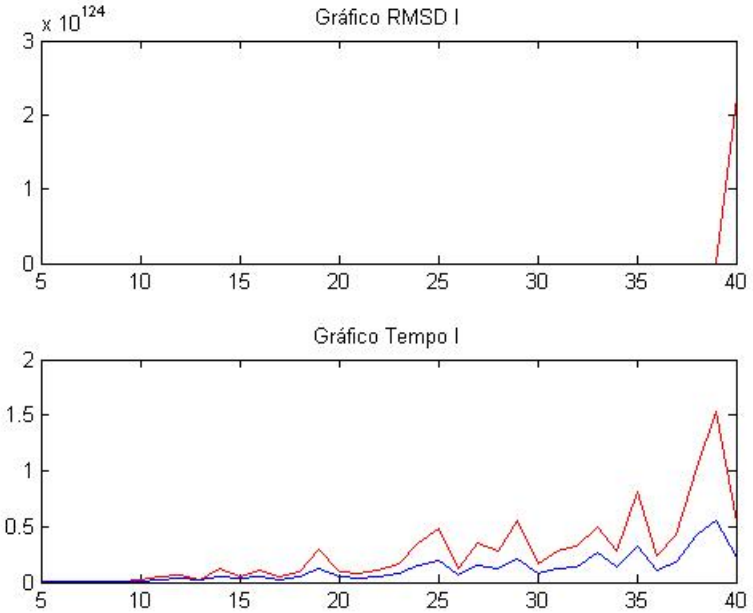

Figura 4.9: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o oitavo experimento.

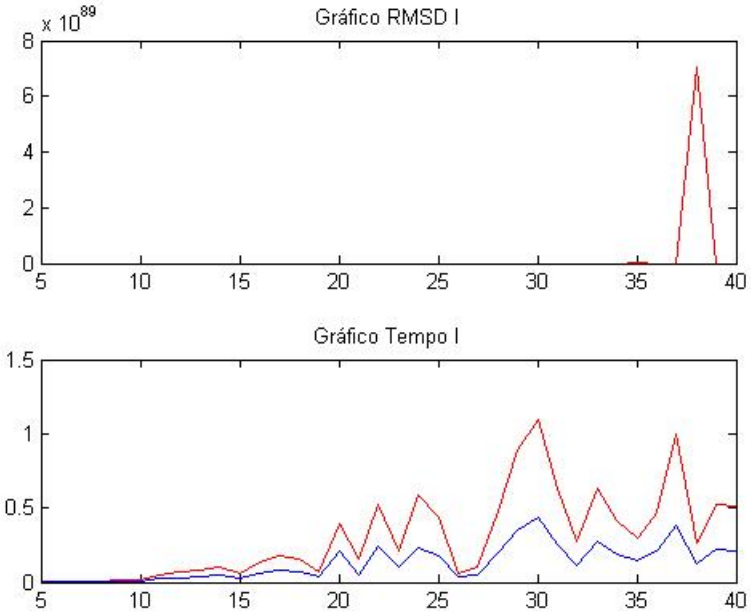

Figura 4.10: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o nono experimento.

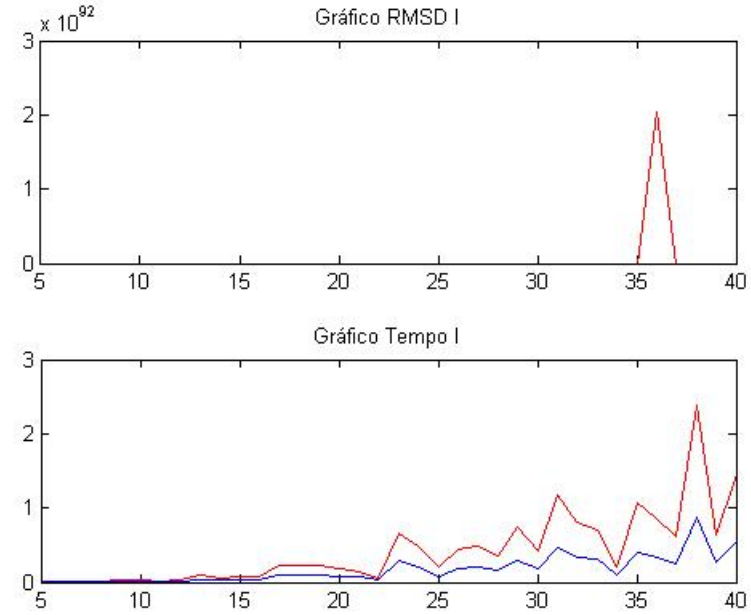

Figura 4.11: Gráficos da RMSD e do Tempo, em função de todas as quantidades de átomos, para o décimo experimento.

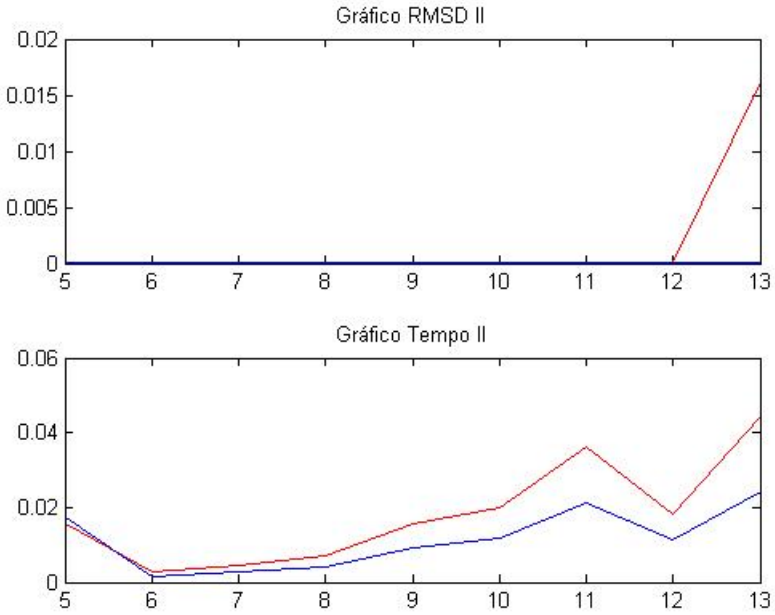

Figura 4.12: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o primeiro experimento.

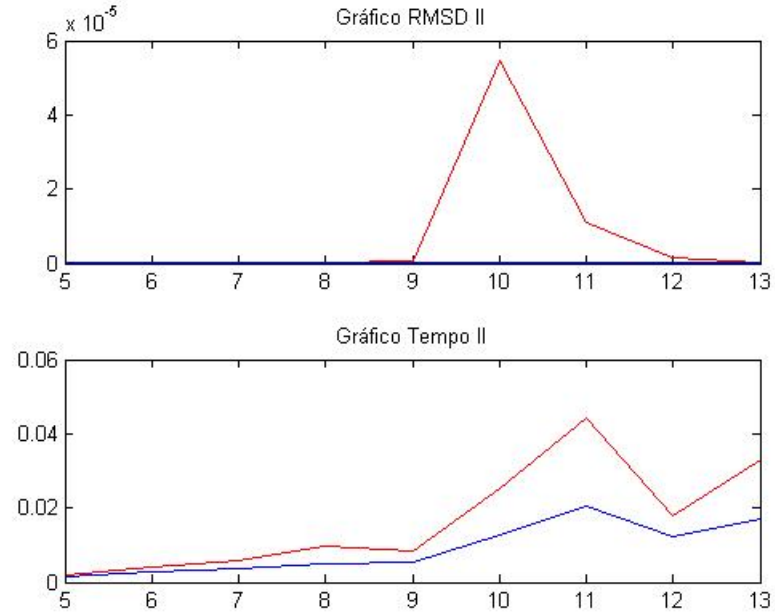

Figura 4.13: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o segundo experimento.

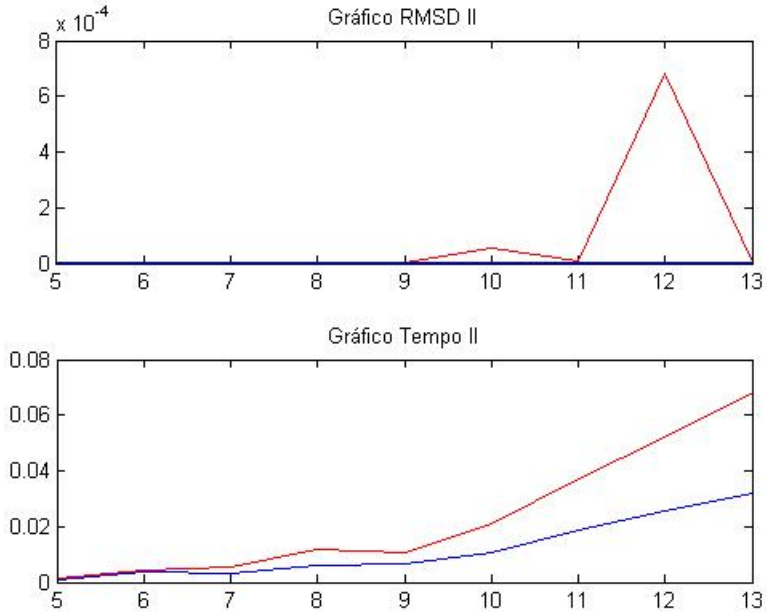

Figura 4.14: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o terceiro experimento.

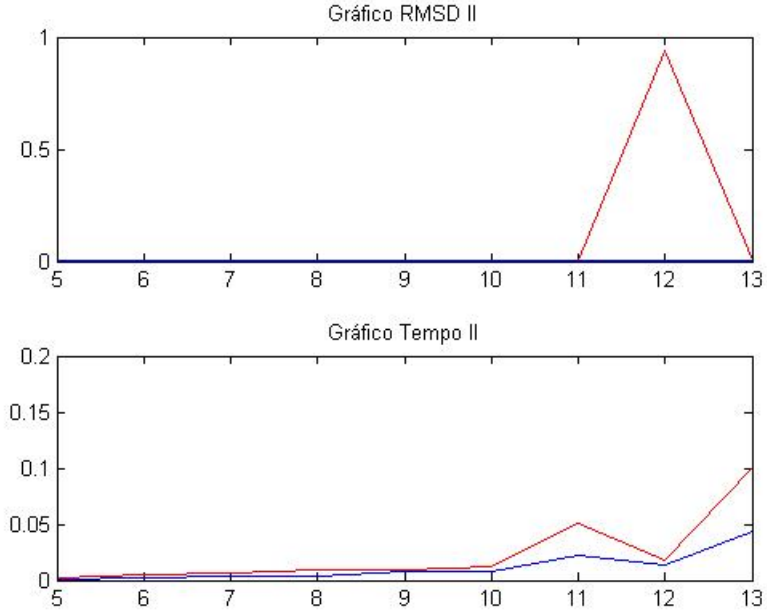

Figura 4.15: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o quarto experimento.

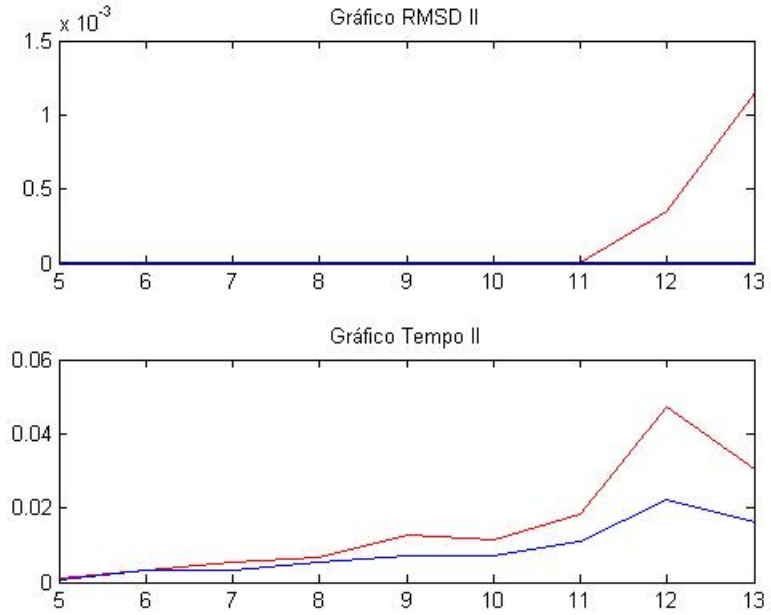

Figura 4.16: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o quinto experimento.

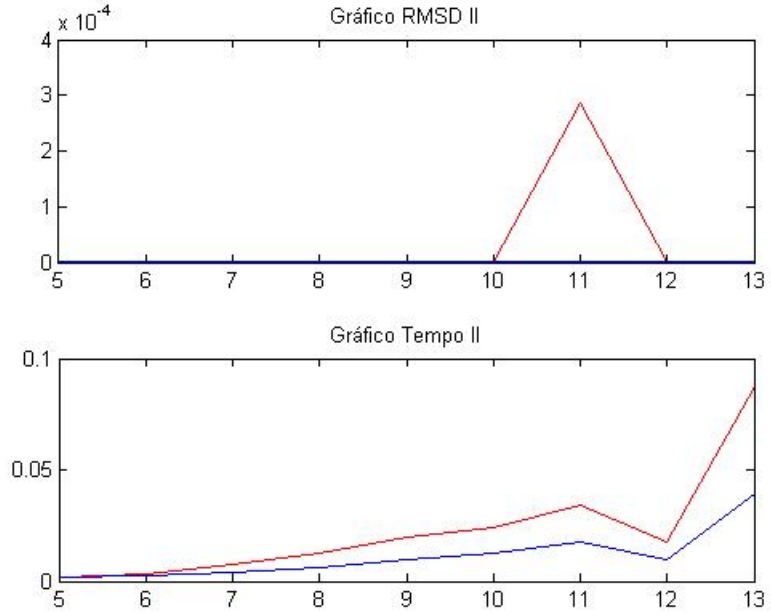

Figura 4.17: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o sexto experimento.

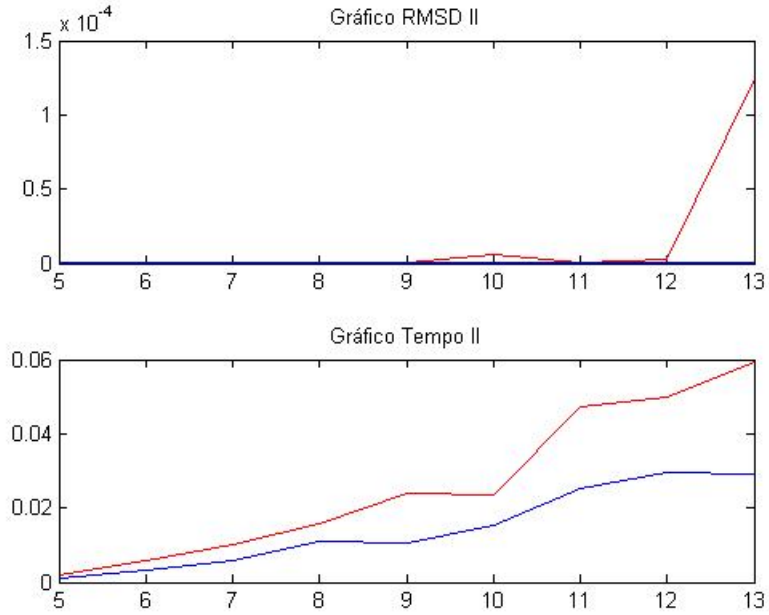

Figura 4.18: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o sétimo experimento.

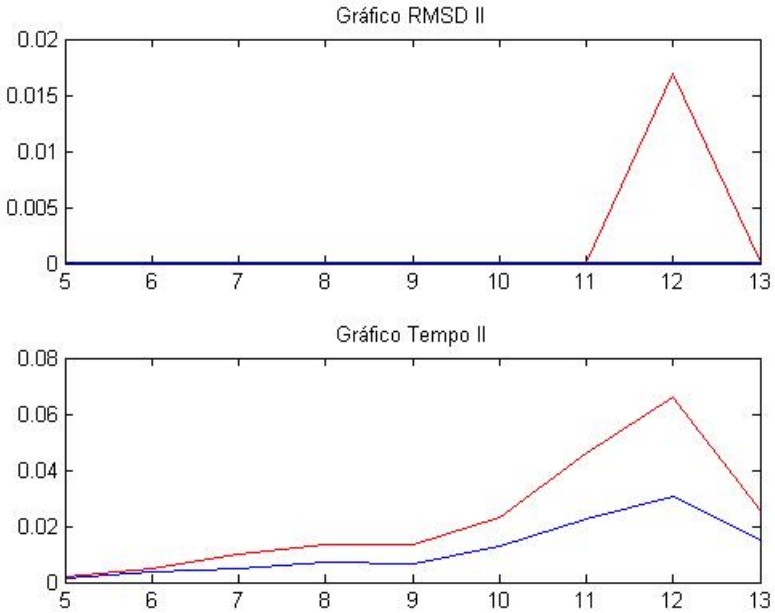

Figura 4.19: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o oitavo experimento.

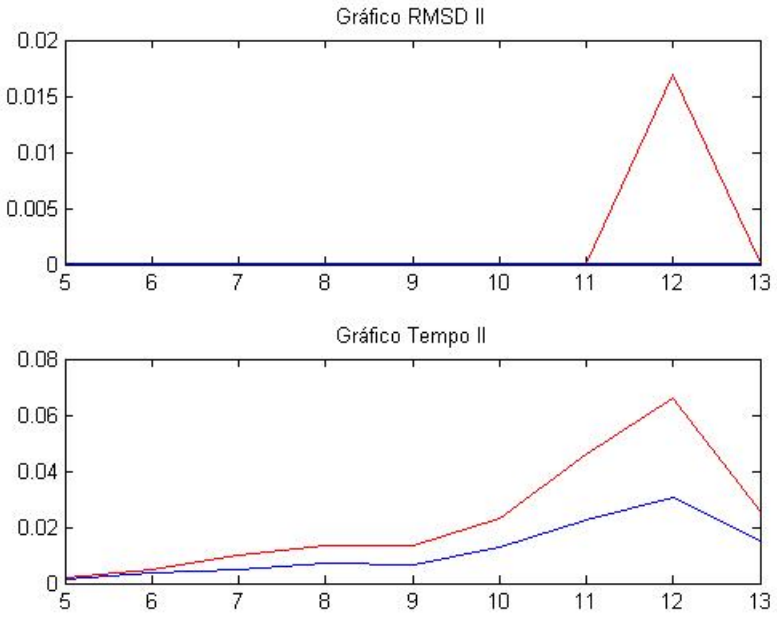

Figura 4.20: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o oitavo experimento.

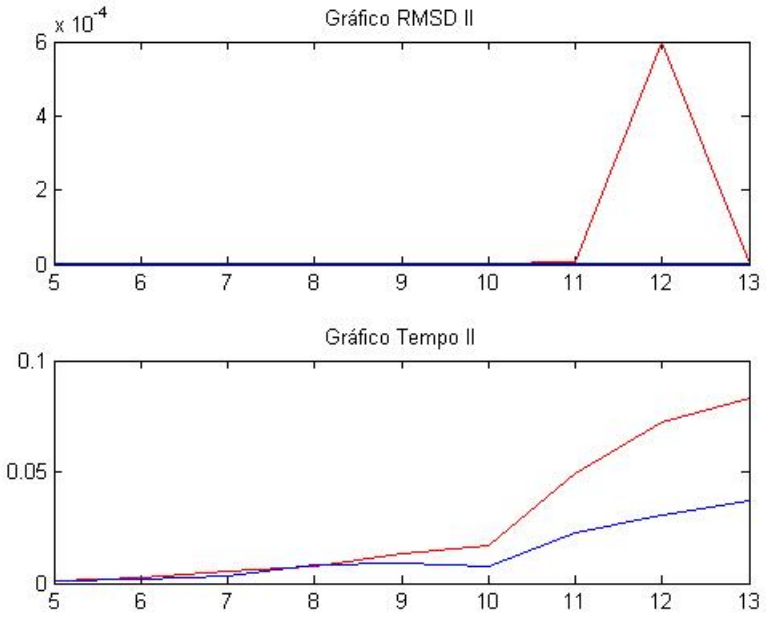

Figura 4.21: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o nono experimento.

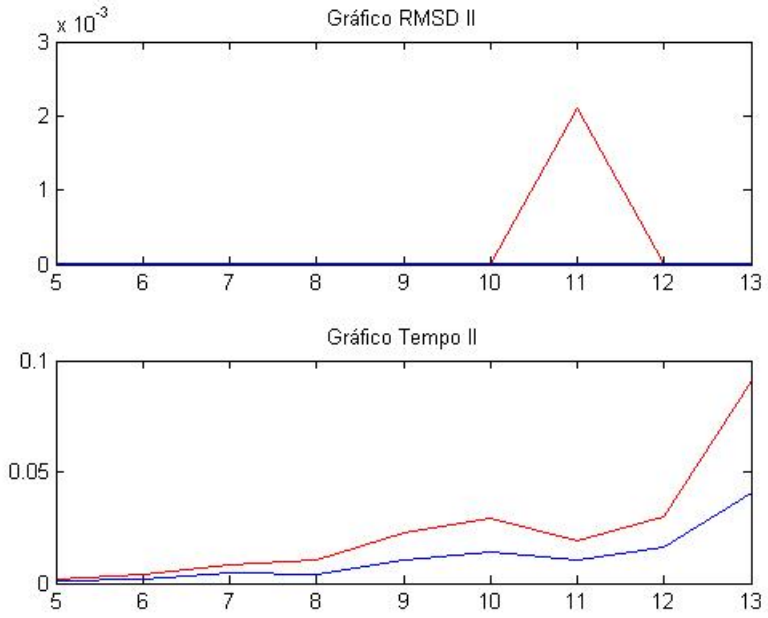

Figura 4.22: Gráficos da RMSD de do Tempo, em função de até quinze átomos, para o décimo experimento.

A partir desses dez experimentos, calculamos uma média entre eles, tanto da RMSD quanto do tempo para cada número de átomos para cada algoritmo. Os resultados médios estão mostrados na Tabela 5.1.

Nas figuras 5.1-22 e 5.1-23, apresentamos dois gráficos cada. A primeira figura possui um gráfico representando a RMSD média do resultados dos dois algoritmos para todas as quantidades de átomos e, o outro, o tempo médio da determinação dos dois algoritmos. A segunda figura apresenta também gráficos de RMSD média e tempo médio, mas apenas para as quantidades de átomos até $n = 15$. Como feito acima, as curvas referentes ao Algoritmo de Determinação Geométrica estão representadas em vermelho e as referentes ao Algoritmo T representadas em azul.

Em todos os resultados, podemos perceber que o Algoritmo T tem performance melhor em relação ao AIDG. O MatLab necessita de menos tempo de execução para determinar as estruturas através do AT do que através do AIDG. Além disso, o AT tem precisão melhor do

que o AIDG e esta precisão perdura por mais tempo no processo de determinação. Ou seja, através do AT, conseguimos reconstruir estruturas com mais átomos de modo a ter precisão mais efetiva do que através do AIDG.

A partir de certo ponto, as matrizes de ambos os sistemas se tornam muito mal-condicionadas, o que prejudica a determinação de estruturas maiores devido à acumulação de erros. Tal instabilidade é oriunda das instâncias artificiais utilizadas para a comparação.

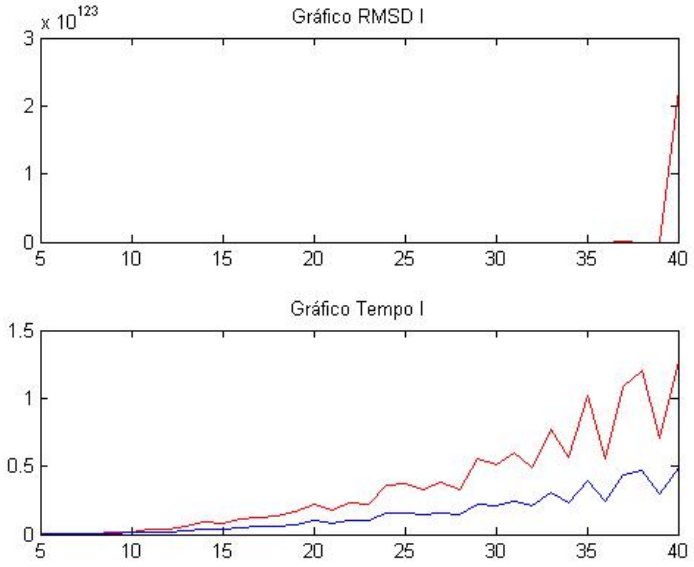

Figura 4.23: Gráficos da RMSD média de do Tempo médio, em função de todas as quantidades de átomos, dos dez experimentos.

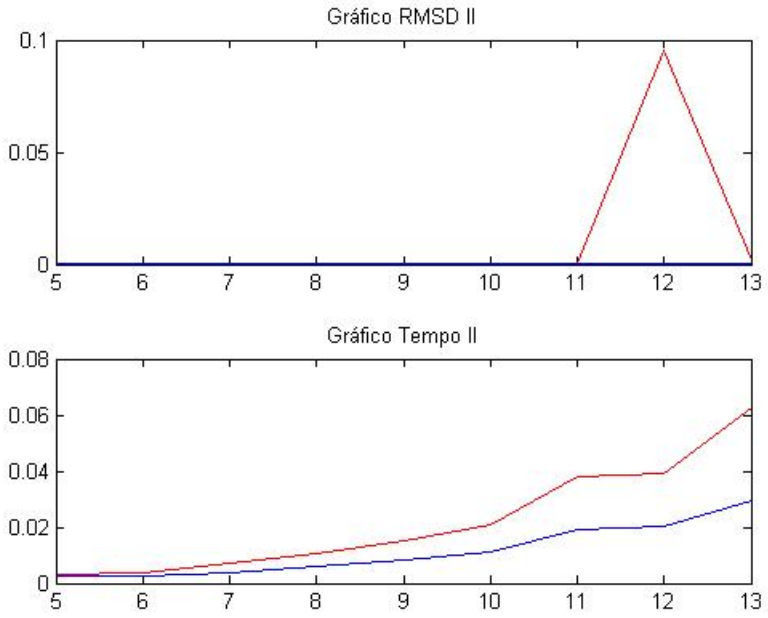

Figura 4.24: Gráficos da RMSD média de do Tempo médio, em função de até quinze átomos, dos dez experimentos.

Nº	RMSD AIDG	RMSD AT	Tempo AIDG	Tempo T
5	5.02044311992e-010	3.22526515958e-010	3.05872250849e-003	2.74885577565e-003
6	5.48629693326e-009	1.12125378813e-009	3.95689220243e-003	2.70284224151e-003
7	2.51680706912e-008	1.22373000609e-009	6.92345177591e-003	3.95526180161e-003
8	7.96796458494e-008	3.17362242677e-009	1.03870118648e-002	6.16916495399e-003
9	3.27872836912e-007	1.25318840727e-008	1.49864178554e-002	8.08180626672e-003
10	1.16474351631e-005	1.29775870036e-008	2.06640622523e-002	1.13256245567e-002
11	2.43226647158e-004	1.80408513011e-008	3.82293665983e-002	1.92013209869e-002
12	9.55141042216e-002	3.14085543219e-008	3.89526305158e-002	2.00902517429e-002
13	1.74946610568e-003	4.05432514162e-008	6.23031857126e-002	2.94232094802e-002
14	1.46541237858e+002	1.01441663599e-007	8.86035888745e-002	4.19613087770e-002
15	2.33872898220e-002	1.04785625856e-007	8.19213005526e-002	3.92819714807e-002
16	3.65852811041e+005	7.55767288913e-008	1.14965041489e-001	5.34879255233e-002
17	4.23935326774e+001	2.50695913290e-006	1.24229794125e-001	5.75491633779e-002
18	4.43637085387e+004	3.34561527210e-005	1.38573969926e-001	6.43761950611e-002
19	5.98171265547e+017	1.24123206876e-005	1.66384758288e-001	7.52890564779e-002
20	6.45770416613e+004	6.29541773903e-007	2.22773483885e-001	1.02363084827e-001
21	2.50012196565e+014	1.60092519547e-004	1.73891350089e-001	7.60468758350e-002
22	8.44745186216e+021	2.54488879922e-006	2.36008398375e-001	1.05498617329e-001
23	5.81280841857e+031	1.21590835824e-005	2.22625434433e-001	1.01973237876e-001
24	7.94645235963e+013	1.32901437771e-005	3.65388148720e-001	1.51797607477e-001
25	4.34432875067e+009	2.05405708380e-004	3.73247477634e-001	1.53890815680e-001
26	4.07910153096e+027	5.32872262330e-004	3.27491954424e-001	1.43982779344e-001
27	3.66509505196e+021	4.67810639431e-005	3.84352409324e-001	1.60635013944e-001
28	2.97377958160e+065	7.45321229635e-003	3.28669783147e-001	1.41617928249e-001
29	3.87880219527e+033	1.82024537654e-001	5.54455115518e-001	2.23202324589e-001
30	2.53034620422e+049	5.77526786689e-001	5.10642214881e-001	2.14594985792e-001
31	8.72637317428e+068	2.62506287973e-002	5.95134521653e-001	2.44847480532e-001
32	5.03540783224e+037	6.94872458228e-003	4.94934254087e-001	2.07718453329e-001
33	6.34762570183e+070	3.75377564749e-003	7.67021158073e-001	3.09192290378e-001
34	5.79460569726e+074	8.27630101685e-004	5.61976607371e-001	2.30406794061e-001
35	6.89511604068e+079	1.41923570019e+004	1.01804831058e+000	3.98563435725e-001
36	2.04413558683e+091	2.10362253131e-003	5.58491127449e-001	2.39506423326e-001
37	1.10281229115e+110	6.58011512547e-003	1.08906938532e+000	4.37144380144e-001
38	6.70756656466e+096	1.47717358770e+000	1.20706185469e+000	4.72767143438e-001
39	5.21593138486e+104	5.07860782788e+002	7.11115846318e-001	2.93687223907e-001
40	2.25866145737e+123	7.66586056470e+034	1.27742284354e+000	4.92092510753e-001

Tabela 4.1: *RMSDs e Tempos médios tanto para AIDG quanto para AT.*

Considerações Finais

Em resumo, apresentamos neste trabalho uma série de algoritmos para tratar Problemas Moleculares de Geometria de Distâncias.

Primeiramente, apresentamos dois algoritmos que tratam o caso em que o conjunto de distâncias do PMGD é completo, ou seja, no qual conhecemos as distâncias entre quaisquer pares de átomos da molécula. O primeiro deles utiliza como ferramenta principal a Decomposição SVD da matriz de distâncias. Já o segundo, que é o precursor da série de algoritmos iterativos de determinação geométrica, tem como cerne a resolução de um sistema linear 3×3, derivado de um sistema não-linear que modela o problema. Este tem ordem de complexidade linear.

Depois, apresentamos três algoritmos para tratar o caso mais geral do PMGD. Este é o caso que nos interessa neste trabalho, no qual o conjunto de distâncias é totalmente arbitrário. O primeiro método é uma generalização do algoritmo com ordem de complexidade linear citado acima e é chamado Algoritmo Iterativo de Determinação Geométrica (AIDG). O segundo método é uma atualização deste, visando diminuir a incidência de erros de arredondamento e é chamado de Algoritmo Iterativo de Determinação Geométrica Atualizado (AIDGA). Ambos são definidos a partir de resoluções de sistemas lineares 3×3. Por fim, o terceiro é um algoritmo inédito que tem como passo principal a resolução de um sistema linear 4×4. Este último é chamado de Algoritmo T e é a proposta que fizemos para este trabalho.

Nos testes realizados com as implementações que fizemos, pudemos constatar que nosso

Algoritmo T é mais eficiente do que o Algoritmo Iterativo de Determinação Geométrica tanto em precisão (medido pela RMSD) quanto em tempo que a máquina usou para determinar as estruturas.

As instâncias escolhidas para a realização dos testes geraram matrizes mal-condicionadas e, desse modo, sistemas lineares instáveis. Logo, a partir de certo ponto as soluções ficam imprecisas e não resolvem mais o problema. Ainda assim, o AT resolve o problema para uma quantidade maior de átomos do que o AIDG.

No trabalho de Doutorado a seguir, iremos estudar, primeiramente, o problema do mal-condicionamento a fim de resolver problemas maiores usando o Algoritmo T.

Referências Bibliográficas

[1] L. Blumenthal: *Theory and Applications of Distance Geometry*, 2nd. Edition. New York, Chelsea Publishing Company, 1970.

[2] T. Creighton: *Proteins: Structures and Molecular Properties*. 2nd. Edition. Freeman and Company, 1993.

[3] G. Crippen and T. Havel: *Distance Geometry and Molecular Conformation*. John Wiley and Sons, 1988.

[4] J. Demmel: *Applied Numerical Linear Algebra*. SIAM, 1996.

[5] Q. Dong and Z. Wu: *A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances*, Journal of Global Optimization, Volume 22, p. 365 - 375, 2002.

[6] Q. Dong and Z. Wu: *A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data*, Journal of Global Optimization, Volume 26, p. 321 - 333, 2003.

[7] G. Golub and C. Van Loan: *Matrix Computations*. The Johns Hopkins University Press, 3rd. Edition, 1996.

[8] B. Hendrickson: *The molecule problem: exploting structure in global optimization*, SIAM Journal on Optimization, Volume 5, 835-857, 1995.

[9] C. Lavor, A. Mucherino, L. Liberti and N. Maculan: *On the computation of protein backbones by using artificial backbones of hydrogens*, Journal of Global Optimization, aceito para publicação, 2010.

[10] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino: *The discretizable molecular distance geometry problem, Computational Optimization and Applications*, aceito para publicação.

[11] C. Lavor: *On generating instances for the molecular distance geometry problem*, Nonconvex Optimization and Its Applications, Volume 84, p.405-414, 2006.

[12] M. Souza: *Suavização Hiperbólica Aplicada à Otimização de Geometria Molecular*, Tese de Doutorado, COPPE-UFRJ, 2011.

[13] J. Saxe: *Embeddability of Weighted Graphs in k-space is Strongly NP-hard*, Proceedings of the 17th Allerton Conference on Communication, Control and Computing, University of Illinois Jose B. Cruz, Jr, p.480-489, Outubro, 1979.

[14] L. Trefethen and D. Bau III: *Numerical Linear Algebra*. SIAM, 1997.

[15] D. Watkins: *Fundamentals of Matrix Computations*, 2nd. Edition. New York, John Wiley and Sons, Inc., 2002.

[16] Z. Wu and D. Wu: *An updated geometric build-up algorithm for solving the molecular distance geometry problems with sparse distance data*, Journal of Global Optimization, Volume 37, p. 661 - 673, 2007.

[17] Z. Wu, D. Wu and Y. Yuan: *Rigid versus unique determination of protein structures with geometric buildup*, Optimization Letters, Volume 2, p.319-331, 2008.

[18] Z. Wu, D. Wu and Y. Yuan: *The solution of the distance geometry problem in protein modeling via geometric buildup*, Biophysical Reviews and Letters, Volume 3, p.43-75, 2008.

[19] D. Wu, R. Davis and C. Ernst: *An Efficient Geometric Build-up Algorithm for Protein Structure Determination with Sparse Exact Distance Data*, forthcoming in the Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, p. 173-180, 2009.

[20] D. Wu, R. Davis and C. Ernst: *Protein structure determination via an efficient geometric build-up algorithm*, Computational Structural Bioinformatics Workshop, Washington D.C., November 2009.