Douglas Silva Maioli

Estabilidade de Sistemas Lineares em Problemas de Geometria Molecular

CAMPINAS

2013

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

Douglas Silva Maioli

Estabilidade de Sistemas Lineares em Problemas de Geometria Molecular

Orientador: Prof. Dr. Eduardo Cardoso de Abreu

Coorientador: Prof. Dr. Carlile Campos Lavor

Dissertação de mestrado apresentada ao Instituto de Matemática, Estatística e Computação Científica da Unicamp para obtenção do título de **Mestre em Matemática Aplicada**

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO DOUGLAS SILVA MAIOLI, E ORIENTADA PELO PROF. DR. EDUARDO CARDOSO DE ABREU.

Assinatura do Orientador

Eduardo Cardoso de Abreu.

Assinatura do Coorientador

Carlile Campos Lavor

CAMPINAS

2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

M285e

Maioli, Douglas Silva, 1987-
Estabilidade de sistemas lineares em problemas de geometria molecular / Douglas Silva Maioli. – Campinas, SP : [s.n.], 2013.

Orientador: Eduardo Cardoso de Abreu.
Coorientador: Carlile Campos Lavor.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Geometria molecular. 2. Sistemas lineares. 3. Proteínas - Estrutura. 4. Simulação (Computadores). I. Abreu, Eduardo Cardoso de,1974-. II. Lavor, Carlile Campos,1968-. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações para Biblioteca Digital

Título em inglês: Stability of linear systems in molecular geometry problems
Palavras-chave em inglês:
Molecular geometry
Linear systems
Proteins - Structure
Computer simulation
Área de concentração: Matemática Aplicada
Titulação: Mestre em Matemática Aplicada
Banca examinadora:
Eduardo Cardoso de Abreu [Orientador]
Maicon Ribeiro Correa
Tiberius de Oliveira e Bonates
Data de defesa: 04-03-2013
Programa de Pós-Graduação: Matemática Aplicada

Dissertação de Mestrado defendida em 04 de março de 2013 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Prof.(a). Dr(a). EDUARDO CARDOSO DE ABREU

Prof.(a). Dr(a). MAICON RIBEIRO CORREA

Prof.(a). Dr(a). TIBERIUS DE OLIVEIRA E BONATES

À minha esposa Fernanda, com amor, admiração e gratidão por sua compreensão, carinho, presença e incansável apoio ao longo do período de elaboração deste trabalho.

Agradecimento

Em primeiro lugar quero agradecer a Deus pelo dom da vida, por iluminar meus pensamentos e principalmente por me dar forças e coragem nos momentos mais difíceis.

Agradeço a minha família por todo apoio. Ao meu pai Wilson por tudo que me ensinou. À minha mãe Márcia por todo esforço que sempre fez por mim. À minha vó Lourdes por todo amor que me deste. Às minhas irmãs Gláucia e Gabriela que sempre trouxeram um toque a mais de beleza em minha vida. Sem eles tudo seria mais difícil.

Agradeço à minha esposa Fernanda pelo apoio, carinho, incentivo e por permanecer ao meu lado em todos os momentos.

Agradeço aos amigos, que perto ou longe, me ajudam a ter força para seguir em frente. À Thais, Bruno, Rinaldo e Luciana que conheci em Campinas e que desde então passamos a trilhar juntos nesse árduo caminho. À juventude da igreja Assembleia de Deus de Barão Geraldo, onde fiz muitos amigos e de onde consigo forças diariamente. Aos amigos da UNESP, onde cresci e aprendi muito, quero agradecer em especial: João Vitor, Luiz Fernando, Lílian, Larissa, Anieli, Juliana, Marcel, Guemael, Fernando, Divane, Máira, Ulcilea, Nathália, Robson, Flávio, Silmara, Jéssica e Viviane. Não poderia também esquecer os grandes amigos que desde criança sei que posso contar e que terei sempre ao meu lado Douglinhas, Fernando, Cássio e a mãe dos dois últimos, a dona Lia.

Um agradecimento especial aos meus orientadores Eduardo e Carlile pelos sábios conselhos, pela dedicação, pela paciência e pela oportunidade de crescimento profissional e pessoal. E ao professor Jaime que me orienta e ajuda desde o início de minha graduação.

Agradeço também aos professores Aurélio, Maicon e Tibérius pela participação na banca e pelas valiosas sugestões que enriqueceram muito este trabalho.

Agradeço a Fundação de Amparo a Pesquisa do Estado de São Paulo - FAPESP - pelo suporte ao projeto 2011/11897-6, e a UNICAMP/FAEPEX (Fundo de Apoio ao Ensino, à Pesquisa e Extensão) pelo suporte ao projeto 519.292-785/11, ambos sob responsabilidade do orientador Prof. Eduardo Cardoso de Abreu. Estes projetos viabilizaram em parte a realização do estudo proposto neste projeto de mestrado. Agradeço também à CAPES pelo apoio financeiro neste trabalho.

Agradeço a todos que de uma maneira ou de outra me auxiliaram a transformar o sonho em uma realidade e que não foram nomeados aqui. Sou eternamente grato à todos que tornaram meu caminho mais leves e mais alegre.

"O coração sábio buscará o conhecimento."

Provérbios $15,14^{a}$

Resumo

No presente trabalho é abordado um Problema de Geometria de Distâncias Moleculares (PGDM) que consiste na determinação de estruturas tridimensionais de moléculas a partir de distâncias entre pares de seus átomos. Inicialmente, apresentamos métodos da literatura utilizados para tentar resolver tal problema, como o Updated Geometric Build-Up (UGB) de Wu e Wu (2007) e o Algoritmo T (AT) de Fidalgo (2011). O novo método introduzido nesta dissertação de mestrado é baseado no AT e foi denominado de Algoritmo T Atualizado (ATA). Esta nova proposta utiliza a mesma estratégia desenvolvida no UGB, que busca obter uma maior estabilidade, com respeito ao número de condição, dos sistemas lineares resolvidos na execução do ATA. Por fim, um estudo baseado em experimentos numéricos foi feito para a verificação da qualidade das soluções obtidas pelo ATA, levando em conta o custo computacional, e em comparação com o método UGB.

Abstract:

The present work approaches the Molecular Distance Geometry Problem (MDGP) which consists on determining three-dimensional molecular structures from distance values between pairs of its atoms. Initially, we present methods from the literature which have been used in order to solve this problem, such as the Updated Geometric Build-Up (UGB) algorithm, from Wu and Wu (2007), and the T Algorithm (TA), from Fidalgo (2011). The new method, introduced in this master dissertation, is based on the TA and was named Updated T Algorithm (UTA). This new approach uses the same strategy developed in the UGB, which looks for obtaining a better numerical stability, with respect to the condition number of the coefficient matrices of the linear systems which are solved in UTA. Finally, an study based on numerical experiments was done for verifying the quality of the solutions obtained from UTA, considering the computational cost and comparing with the UGB.

Lista de Siglas

Å	Ângstron
DNA	*Deoxyribonucleic Acid*
MDGP	*Molecular Distance Geometry Problem*
PDB	*Protein Data Bank*
PGDM	Problema de Geometria de Distâncias Moleculares
SVD	*Singular Value Decomposition*
RMSD	*Root-Mean-Square Deviation*
RNA	*Ribonucleic Acid*

Métodos:

AT	Algoritmo T
ATA	Algoritmo T Atualizado
GB	*Geometric Build-Up*
UGB	*Updated Geometric Build-Up*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Neste trabalho iremos apresentar um novo método, denominado Algoritmo T Atualizado (ATA), utilizado na resolução do Problema de Geometria de Distâncias Moleculares (PGDM), que consiste no problema de determinar a estrutura de certas moléculas conhecendo apenas um conjunto de distâncias entre pares de seus átomos. Nesta dissertação será considerado apenas moléculas de proteínas, que são os produtos finais da expressão genética, responsáveis diretas pela maior parte das funções celulares, já que o DNA, onde esta armazenado toda nossa informação genética, fica restrito ao núcleo nas células eucariontes, e portanto, não pode atuar diretamente na maioria dos processos celulares. Dessa forma, por meio de um processo chamado transcrição, o DNA induz à formação do RNA, transcrevendo as informações necessárias para a produção das proteínas. Observa-se, ainda, que essas informações contém a sequência de aminoácidos que determinam a estrutura da proteína, e portanto sua função, já que existe uma relação determinística entre a estrutura e a função das proteínas [1]. Por isso, é de grande importância a determinação das estruturas das proteínas.

Com auxílio do método ATA, que é uma atualização do Algoritmo T (AT) de F. Fidalgo [21], buscamos resolver o PGDM com distâncias exatas. E uma de nossas perspectivas de trabalhos futuros é estender tal algoritmo para trabalhar também com distâncias imprecisas. Ambos os algoritmos (AT e ATA) foram baseados na família de métodos *"Geometric Build-Up"* de Wu et al [17, 18, 45], que também apresentaremos neste trabalho. Porém, antes de discutirmos estes métodos, iremos, no capítulo 1, discorrer um pouco sobre a história das proteínas, mostrando brevemente como elas são formadas, e comentar sobre os experimentos de Ressonância Magnética Nuclear (RMN), que são utilizados para colher certos tipos de dados da proteína, como distâncias de pares de átomos desta. Podemos usar estas distâncias como os dados de entrada para execução de nossos algoritmos. Na seção 2 deste capítulo, iremos introduzir uma modelagem do problema, que nos ajudará no estudo dos métodos, e por fim, apresentamos algumas ferramentas e conceitos fundamentais com o objetivo de facilitar a leitura deste trabalho.

Dedicamos o capítulo 2 para apresentar os métodos da família "Geometric Build-Up", formado por todos os métodos que buscam resolver o PGDM utilizando a mesma estratégia que o *"Geometric Build-Up Algorithm"* (GB). Além do GB, outros métodos que fazem parte dessa família e que apresentaremos no capítulo 2, serão o *Updated Geometric Build-Up Algorithm* (UGB) e um método de complexidade linear que resolve o PGDM com um conjunto completo de distâncias exatas, este último método citado foi uma alternativa apresentada por Dong e Wu em [17] a outro método que utiliza a fatoração em fatores valores singulares (SVD), que tem ordem de complexidade $O(n^3)$, e que também será discutido na primeira seção deste capítulo.

No capítulo 3 iremos apresentar a "família T" de algoritmos, que é composta pelo AT, que assim como o GB utiliza uma estratégia de transformar sistemas quadráticos em sistemas lineares para resolvê-los, no qual implementamos um conveniente procedimento, a fim de diminuir o acúmulo de erros provenientes do mal condicionamento das matrizes de coeficientes, em cada

sistema linear baseada na atualização feita no UGB, obtendo uma maior estabilidade numérica. Esta atualização do AT, juntamente com a utilização do método LU com pivoteamento parcial, é nossa proposta neste trabalho, e o denominamos de Algoritmo T Atualizado.

No capítulo 4 realizaremos uma comparação numérica entre os algoritmos UGB e ATA por meio de experimentos computacionais. Para isso, utilizaremos dois tipos diferentes de instâncias: (1) geradas artificialmente conforme descrito em [32], escolhidas pela simplicidade de implementação, e por simular estruturas moleculares reais e (2) retiradas do banco de dados *Protein Data Bank* (PDB) [5], que são estruturas moleculares mais realísticas. Para fazer tal comparação entre os métodos, iremos levar em consideração duas variáveis, o tempo em que cada algoritmo leva para calcular as coordenadas dos átomos e o erro entre a estrutura molecular original e a molécula reportada pelos métodos. Para obter uma quantificação do erro cometido nesta aproximação será utilizado o *Root-Mean-Square Deviation* (RMSD), que é uma forma de medir a distância entre as estruturas de duas moléculas sobrepostas [19, 41]. Nestes testes foi possível verificar que apesar do erro ter sido praticamente igual quanto aos dois métodos, tivemos que o tempo que o ATA precisou para determinar as estruturas foi em média 45% menor que o tempo do UGB. No capítulo 4 também é apresentado os testes referentes a estabilidade numérica obtida com a atualização do AT.

Finalmente, no capítulo 5 apresentamos nossas conclusões referentes ao trabalho e aos testes computacionais, apresentando uma lista de perspectivas futuras e de questões que no nosso entendimento requerem um estudo mais detalhado e um maior entendimento na sequência deste trabalho.

Capítulo 2

Um Problema de Geometria de Distância Molecular

Neste capítulo iremos descrever o Problema de Geometria de Distância Molecular (PGDM), cuja sigla em inglês é MDGP de *"Molecular Distance Geometry Problem"*. Na primeira seção apresentaremos alguns apectos da origem do problema, para isso, discorreremos um pouco sobre a história das proteínas, tais como sua formação e como sua estrutura influencia na função que irá desempenhar. Até o momento não existem meios físicos diretos que observam a estrutura de uma proteína em uma resolução suficiente para determinar sua função. Dessa forma, citaremos dois experimentos, (1) Difração de raio-X da cristalização da proteína e (2) Ressonância Magnética Nuclear (RMN), utilizados para colher dados da proteína para que assim a sua estrutura possa ser determinada através de métodos indiretos, como por exemplo, o algoritmo T Atualizado. Os dados de entrada de tal algoritmo podem ser obtidos através de experimento de RMN, que podem medir, entre outras coisas, distâncias inter-atômicas entre átomos da molécula de proteína. Portanto, utilizando propriedades geométricas e de resolução de equações algébricas advindas da modelagem matemática é possível determinar a estrutura da molécula desejada.

Na segunda seção, realizaremos uma das possíveis modelagens do PGDM, que tem como objetivo calcular a estrutura tridimensional de uma molécula de proteína conhecendo apenas as distâncias entre pares de seus átomos, ou em muitos casos, quando não é possível determinar a distância exata, conhecendo apenas intervalos onde estas distâncias podem estar contidas. Por fim, na terceira seção, apresentaremos algumas ferramentas e conceitos fundamentais que iremos utilizar nos métodos descritos neste trabalho.

2.1 Uma Breve Discussão sobre Proteínas no Contexto de Geometria Molecular

Em fevereiro de 2001, dois grupos concorrentes, o consórcio público internacional *Human Genome Sequencing* [28] e a empresa americana *Celera* [44], anunciaram que conseguiram, pela primeira vez na história da humanidade, mapear o genoma humano e estabelecer sua seqüência, ao decifrar $3,1$ bilhões de bases químicas (nucleotídeos) do DNA, que constituem cerca de 40000 genes diferentes, presentes no genoma humano [27]. Genoma é toda a informação hereditária única, de cada espécie de organismo, que está codificada em seu DNA. Para os seres humanos, o genoma é essencialmente equivalente ao conteúdo de informação genética que está presente em um conjunto completo de cromossomos humanos [31].

Com a conclusão do mapeamento do genoma humano, os estudos sobre as proteínas, os produtos finais da expressão genética, tornaram-se cada vez mais importantes para a interpretação dos genes e das suas implicações para a vida. As proteínas formam uma classe importante de

moléculas, no entanto, para compreender as proteínas e suas funções, é necessário conhecer as suas estruturas tridimensionais, as quais, devido a várias razões técnicas, são muito difíceis de determinar [4, 10]. A maioria das proteínas naturais são dotadas de estruturas tridimensionais específicas que estão associadas às suas atividades biológicas. Apesar de dinâmica, sobre condições térmicas e configurações locais típicas, a estrutura tridimensional de cada proteína apresenta pequenas variações. Uma das grandes descobertas sobre essa estrutura biomolecular é a relação determinística entre a sequência de aminoácidos e sua estrutura [40]. Isso foi apontado pela primeira vez, no início da década de 1960, por Christian Boehmer Anfinsen $(1916-1995)$ e colaboradores [1, 43].

Iniciaremos a discussão sobre a formação das proteínas onde se inicia todo processo de sua formação: no DNA (ou ácido desoxirribonucléico). O DNA é o principal responsável pelo armazenamento de nossas características hereditárias, coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus, e suas interações com os fatores ambientais determinam a maioria dos aspectos da saúde humana. Entretanto, o DNA apresenta pouca mobilidade, restringindo-se apenas ao interior do núcleo celular, por isso, sua atuação na maioria das vezes é feita de forma indireta. Por meio de um processo chamado transcrição, o DNA induz à formação do RNA, copiando ou transcrevendo as informações genéticas contidas em si para moléculas de RNA. As sequências de nucleotídeos, que são blocos que constituem a molécula de DNA, contém o "código" para a ordenação específica de aminoácidos, que formarão as proteínas. O RNA, por sua vez, junto com o ribossomo, iniciam um processo denominado tradução, processo no qual ocorre a ordenação e ligação dos aminoácidos, determinada pelo "código" passado pelo DNA, que formam um tipo de proteína [43, 31].

Existem três principais classes de RNA (ou ácido ribonucléico): o RNA mensageiro (mRNA), o RNA de transferência (tRNA) e o ribossômico (rRNA), sendo que os três estão envolvidos na síntese protéica. O mRNA contém a informação genética, transmitida pelo DNA no processo de transcrição, para a sequência de aminoácidos, o tRNA identifica e transporta as moléculas de aminoácidos até o ribossomo, e o rRNA representa 50% da massa dos ribossomos, organelas que fornecem um suporte molecular para as reações químicas da montagem de uma proteína [15, 36].

Dessa forma, podemos dizer que o DNA gera o RNA através da transcrição e passando a este as informações necessárias para a formação de algum tipo de proteína, e o RNA com estas informações gera as proteínas através da tradução. Assim, a proteína será a molécula responsável por agir diretamente em quase todas as funções celulares. Estima-se que as células de um mamífero típico tenham cerca de 100.000 diferentes proteínas, com estruturas e funções diversas [31]. Em outras palavras, o genoma, o conjunto completo da informação genética, contém somente a receita para a fabricação de proteínas, enquanto que as proteínas desempenham o papel de cimento e tijolos das células e realizam a maior parte do trabalho. Por isso, a compreensão do real significado do mapeamento do genoma humano e suas possíveis aplicações estão profundamente ligadas ao entendimento do papel desempenhado pelas proteínas.

Infelizmente, o proteoma, isto é, o conjunto de todas as proteínas produzidas por uma dada célula, tecido ou organismo, é muito mais complicado que o genoma [20]. O alfabeto do DNA é composto por quatro bases químicas conhecidas por suas iniciais: adenina (A), citosina (C), guanina (G) e timina (T). As proteínas, no entanto, são formadas pela combinação de 20 blocos fundamentais denominados aminoácidos. Porém, mesmo quando a sequência de aminoácidos de uma proteína é conhecida, é possível que não se consiga determinar a função da proteína, ou a que outras proteínas ela pode se associar, pois, além da sequência de aminoácidos, outro fator importante para se determinar a função de uma proteína é conhecer seus ângulos de torção, que em alguns casos, desafiam a predição e estão diretamente ligadas às funções desempenhadas pela proteína [10, 20, 7, 43].

A história das proteínas se inicia no século XVIII, quando foi descoberto que, assim como o sangue e o leite, o albúmen (clara de ovo) também coagulava em altas temperaturas e em

meio ácido. E por isso, todas as substâncias que tinham essa característica foram denominadas albuminoides. Posteriormente, no século XIX, foi descoberto que os principais elementos das células vivas eram substâncias albuminoides.

O primeiro a usar o termo proteína (do grego proteitos, que significa primeiro, primitivo) foi o químico Holandês Gerardus Johannes Murder $(1802-1880)$ em um artigo publicado em 1838 para se referir às substâncias albuminoides. A partir de então, o interesse pelo estudo das proteínas só vem crescendo. Químicos descobriram que a degradação liberava aminoácidos. Baseado nisso, em 1902, o alemão Franz Hofmeister $(1850-1922)$ propôs que as proteínas seriam formadas por aminoácidos encadeados, o que hoje se sabe que é verdade. [2]

Segundo a literatura indicada anteriormente, existe 20 diferentes tipos de aminoácidos que compõem as proteínas, estando presentes nestas em diversas proporções, unidos por ligações peptídicas, o motivo das proteínas também poderem ser chamadas de polipeptídeos. A ordem em que estes 20 aminoácidos podem se unir dá origem a um número astronômico de combinações em diferentes moléculas de proteínas, e que determinam não só sua especificidade, mas também sua atividade biológica. Afinal, a função de cada proteína depende de sua estrutura tridimensional, por isso, a importância de se estudar sua estrutura.

Como não existem meios físicos diretos que observam a estrutura de uma proteína em uma resolução suficiente para que se possa determinar sua função, várias abordagens experimentais têm sido utilizadas para obter alguns dados estruturais indiretos de tal forma que suas estruturas pudessem ser deduzidas. Por exemplo, os dados de difração para um cristal de proteína podem ser obtidos pela cristalografia de raios-X e utilizados para achar a distribuição de densidade dos elétrons e, portanto, a estrutura da proteína [18]. Até 1984, este era o único método para determinar uma estrutura protéica em resolução atômica, porém, a cristalografia de raios-X requer cristalização da proteína, o que é demorado e muitas vezes falha. Foi então, que iniciou-se a utilização da ressonância magnética nuclear (RMN), que tornou possível a determinação de estruturas de proteínas com uma elevada precisão em um ambiente (solução) muito mais próximo da situação natural de um organismo vivo do que os cristais utilizados na cristalografia.

Os experimentos de RMN baseiam-se no fato de que os núcleos de hidrogênio têm dois estados (spins) que podem ser alterados pelo fornecimento de energia em uma dada frequência. A informação estrutural vem do acoplamento spin-spin entre os núcleos de hidrogênio. Se dois núcleos estão espacialmente próximos, então seus spins interagem e a frequência necessária para alterar um spin é modificada. Os picos no espectro tornam-se ligeiramente alterados, o que torna possível a inferência não só de distâncias envolvendo pares de átomos de hidrogênio espacialmente próximos, com distância inferior a $5-6$ Å (1 Å $= 10^{-10}$ m), mas também de ângulos entre átomos em uma dada proteína [25, 26, 48]. Para calcular a estrutura tridimensional da macromolécula, essas distâncias são usadas como restrições em combinações com diversas informações suplementares, tais como: a sequência de aminoácidos que compõem a proteína, referências geométricas para o comprimento e os ângulos das ligações químicas existentes, entre outras [43].

No nosso caso, os dados que necessitaremos dos experimentos de RMN são as distâncias entre pares de átomos da molécula de proteína, porém, como nem sempre os experimentos de RMN são capazes de calcular as distâncias exatas entre todos pares, mas apenas intervalos onde estas distâncias estão contidas, então em trabalhos futuros pretendemos estudar sobre distâncias imprecisas, que é um caso mais geral do que é estudado nesta dissertação.

O Protein Data Bank (PDB) é um banco de dados para estruturas tridimensionais de proteínas e aminoácidos, fundado em 1971 por Edgar Meyer e Walter Hamilton [5]. Os dados contidos no PDB são frutos de experimentos de RMN, cristalografia de raio-X ou de desenvolvimento teórico realizados por pesquisadores de diferentes partes do mundo. Hoje, estima-se que 80% das estruturas do PDB foram determinadas por cristalografia de raios X, 15% através da RMN, e 5% em outras abordagens. Destas estruturas, cerca de algumas dezenas de milhares no total, contêm uma elevada porcentagem de repetições (estruturas para a mesma proteína determinadas

com diferentes técnicas ou sob diferentes condições). Existem, pelo menos, várias centenas de milhares de diferentes proteínas no corpo humano por si só, no entanto, a maioria de suas estruturas ainda é desconhecida [9, 38, 42].

Um exemplo de molécula de proteína é a membrana associada ao componente essencial do complexo de replicação do vírus da hepatite C, denominada de proteína não estrutural 5A (ou NS5A) representada na figura 2.1, cujo nome no PDB é 1R7C e que é constituída por 532 átomos. No caso dessa figura, cada cor representa um tipo de aminoácido diferente.

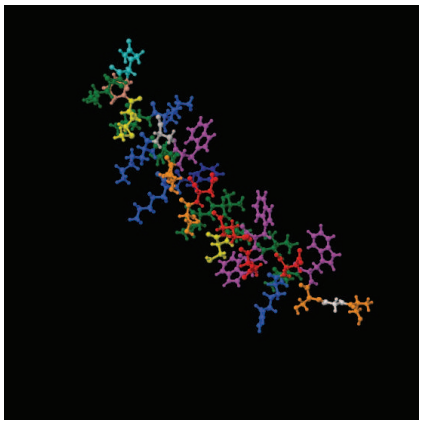


Figura 2.1: Proteína NS5A (1R7C) formada por 532 átomos [5]

O problema que apresentaremos a seguir, denominado Problema de Geometria de Distâncias Moleculares, tem como obejtivo calcular a estrutura tridimensional de uma dada molécula de proteína, determinando as coordenadas de seus átomos em um espaço tridimensional, utilizando apenas as distâncias entre os átomos da molécula como instâncias iniciais, distâncias estas que podem ser medidas através de experimentos de RMN. As moléculas que utilizaremos em nossos testes terão duas procedências, usaremos proteínas do banco de dado PDB e estruturas geradas artificialmente que simulam uma molécula de proteína real.

2.2 Uma Modelagem do Problema de Geometria de Distâncias Moleculares para Proteínas

Como já foi discutido, nosso objetivo é determinar a estrutura tridimensional de uma dada molécula de proteína, conhecendo inicialmente apenas um conjunto de distâncias entre pares de átomos desta molécula, que podem ser calculadas através de experimentos de Ressonância Magnética Nuclear (RMN). Este conjunto de distâncias pode ser completo, sendo de nosso conhecimento as distâncias entre todos os pares de átomos dessa molécula de proteína, ou pode ser um conjunto arbitrário, o que geralmente ocorre nos experimentos de RMN que nem sempre consegue estimar todas as distâncias entre os pares de átomos. Além disso, muitas vezes também não é possível obter os valores exatos das distâncias, podendo ser apenas obtido um intervalo para algumas das mesmas.

Como estamos interessados na estrutura tridimensional de uma proteína, vamos considerar que essa molécula está em um espaço tridimensional. Neste trabalho utilizaremos o $\mathbb{R}^3$ como sendo o espaço em que a estrutura está contida. Dessa forma, cada átomo dessa proteína será um ponto deste espaço. Assim, se a molécula tiver n átomos, poderemos enumerar cada átomo de 1 até n, e a posição de cada molécula i em $\mathbb{R}^3$ será $x_i = (x_{i,1}, x_{i,2}, x_{i,3})$ onde $x_{i,1}, x_{i,2}, x_{i,3} \in \mathbb{R}$ e $i = 1, ..., n$. Por simplificação de notação, por algumas vezes, iremos chamar a molécula i de x_i.

Se as coordenadas dos átomos de uma molécula forem conhecidas, podemos determinar a distância entre eles. Por exemplo, para calcularmos a distância, $d_{i,j}$, entre os átomos x_i e x_j, basta fazermos:

$$\|x_i - x_j\| = d_{i,j}, \tag{2.1}$$

onde $\|.\|$ é a norma euclidiana. Porém, o que queremos é fazer o inverso, isto é, calcular as posições x_i, $i = 1,...,n$ dos átomos da molécula de proteína sabendo apenas as distâncias $d_{i,j}$ entre eles, ou em muitos casos, apenas um subconjunto do conjunto destas distâncias. Como na figura 2.2 onde temos como instâncias iniciais apenas as distâncias conhecidas que são as que estão descritas na figura, como por exemplo, entre os átomos 1 e 3. Já outras distâncias entre pares de átomos, como $d_{1,4}$ ou $d_{3,6}$ não são dadas. O PGDM é determinar as coordenadas de cada átomo, ainda desconhecidas.

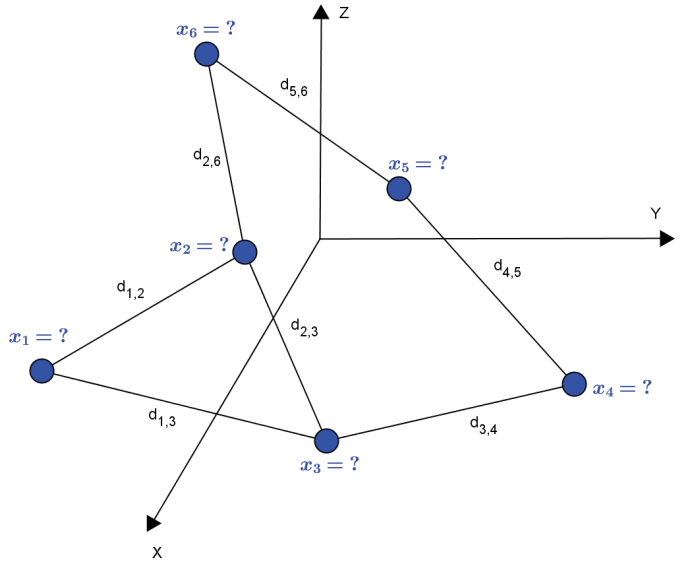


Figura 2.2: Estrutura molecular com coordenadas dos átomos desconhecidas

Como já foi mencionado, haverá casos, em que os esperimentos de RMN não conseguirão fornecer todas distâncias exatas, mas apenas intervalos numéricos onde estarão parte destas distâncias. Assim, haverá apenas uma cota inferior $l_{i,j}$ e uma cota superior $u_{i,j}$ para alguma distância $d_{i,j}$. Dessa forma, teremos o seguinte:

$$l_{i,j} \leq d_{i,j} \leq u_{i,j}, \tag{2.2}$$

onde $d_{i,j}$ é a distância entre os átomos i e j. Esse problema da determinação da estrutura tridimensional de uma molécula, onde é conhecido apenas distâncias entre seus átomos, é nomeado de *Problema de Geometria de Distâncias Moleculares* (PGDM), que segundo Crippen e Havel [12] foi definido em 1841 por Arthur Cayley $(1821-1895)$, e que é uma particularização de um outro problema, que é chamado de *Problema de Geometria de Distâncias* (PGD), este último, também baseado no cálculo das coordenadas de certos pontos, tendo como instâncias apenas as distâncias de alguns desses pontos, porém abrangendo uma quantidade maior de aplicações, não apenas em problemas moleculares. Podemos, assim, definir o PGDM da seguinte forma:

Problema de Geometria de Distâncias Moleculares (PGDM): Considere um molécula formada por uma sequência de n átomos, da qual é conhecido um subconjunto das distâncias entre pares deles. É possível obter uma configuração tridimensional para tal molécula, compatível com as distâncias dadas, isto é, determinar um conjunto de coordenadas $\{x_1,....,x_n\}$ no $\mathbb{R}^3$ para seus

átomos de modo que as distâncias euclidianas entre essas posições sejam iguais às distâncias conhecidas?

Dessa forma, uma possível modelagem matemática do PGDM seria a seguinte: "Como representar a estrutura molecular tridimensional no espaço $\mathbb{R}^3$ de uma proteína composta por n átomos, onde teremos que calcular as coordenadas $x_1, x_2, ..., x_n$ dos átomos, representando suas posições $x_i = (x_{i,1}, x_{i,2}, x_{i,3})$ no $\mathbb{R}^3$ para $i = 1, 2 \ldots n$, conhecendo um subconjunto do conjunto D das distâncias dos pares de átomos molécula."

Tendo como dados iniciais: (1) no caso completo, as cotas inferiores, $l_{i,j}$, e superiores, $u_{i,j}$, das distâncias, $d_{i,j}$, entre os átomos i e j, para todo $i, j \in \{1, 2, ..., n\}$, e (2) no caso arbitrário, um subconjunto destas cotas. Teremos, dessa forma, que calcular as coordenadas de todos átomos (ou o máximo que é possível), de modo que as distâncias entre eles estejam entre as cotas dadas no início.

No caso de distâncias exatas, temos que $l_{i,j} = d_{i,j} = u_{i,j}, \forall i, j \in \{1, 2, ..., n\}$, então, as instâncias do problema serão as distâncias entre os átomos, e da mesma forma que relatado anteriormente, teremos que obter $x_i = \{x_{i,1}, x_{i,2}, x_{i,3}\}, \forall i \in \{1, 2, ..., n\}$, de tal forma, que as distâncias destas coordenadas encontradas satisfaçam as distâncias dadas.

Em outras palavras, conhecendo as distâncias $d_{i,j}$, onde $(i, j) \in K$ e K é um subconjunto de $\{1, 2, ..., n\}^2$, queremos encontrar os pontos $x_1, x_2, ..., x_n$ de $\mathbb{R}^3$, tais que satisfazem o seguinte:

$$\|x_i - x_j\| = d_{i,j}, \forall (i, j) \in K. \tag{2.3}$$

Podemos perceber que este é um sistema quadrático, e as coordenadas dos átomos serão as soluções deste sistema, e isto será o cerne dos métodos que apresentaremos, a tranformação linear de sistemas de equações quadráticas e a resolução destes sistemas lineares.

Geralmente não é fácil calcular as coordenadas de todos os átomos da molécula estudada, e nem sempre é possível determinar alguma estrutura que satisfaça todas as condições iniciais dadas. Além disso, em muitos casos, principalmente no problema de distâncias imprecisas, onde é dado apenas o intervalo em que estão as distâncias [12], pode existir um número infindável de possíveis estruturas que satisfaçam todas as condições iniciais, e, nestes casos, é importante não apenas encontrar uma destas estruturas possíveis, mas todo o conjunto de estruturas, porque os desvios das estruturas umas das outras no conjunto fornecem informações importantes sobre a forma como a estrutura da proteína pode flutuar dinamicamente em torno de seu estado de equilíbrio. Esta propriedade dinâmica é muitas vezes tão crítica quanto a estrutura em si para a compreensão da função da proteína [12, 47].

Nosso trabalho se iniciará com o estudo dos métodos de resolução do PGDM com um conjunto completo de distâncias exatas, passando em seguida a estudar os métodos que resolvem este problema quando se tem um conjunto arbitrário de distâncias exatas.

Por fim traremos algumas possibilidades de estudos futuros para trabalhar com as distâncias calculadas via RMN que não são exatas, ou seja, como já explicamos, aquelas cujas distâncias estão em algum intervalo. Além disso, o novo algoritmo ATA também terá o objetivo de corrigir alguns erros causados por aproximação e o número de condição de algumas matrizes.

2.3 Alguns Conceitos Fundamentais de Modelagem Computacional de Proteínas

Antes de iniciarmos a discussão sobre a resolução do PGDM, iremos enumerar algumas definições e conceitos preliminares muito recorrentes nas discussões dos métodos que citaremos. Essas primeiras quatro definições irão nos mostrar os conceitos de átomo determinado, átomos base, átomos vizinhos e independência entre átomos.

Definição 1. *É chamado de* ***átomo posicionado*** *(ou* ***átomo determinado****), todo átomo de uma molécula, cujas coordenadas são conhecidas em* $\mathbb{R}^3$*. Se não conhecemos as coordenadas de um átomo, ele é dito* ***átomo não-determinado*** *(ou* ***átomo indeterminado****)*

Iremos chamar de átomos remanescentes a todos átomos que nos resta determinar, ou seja, todos átomos não determinados.

Definição 2. *Diremos que um conjunto de átomos B são* ***átomos base*** *de um outro átomo i, se for possível determinar unicamente as coordenadas de i, por meio das distâncias conhecidas deste átomo i aos átomos do conjunto B.*

Definição 3. *Dizemos que dois átomos i e j são* ***átomos vizinhos****, se conhecemos a distância entre eles, isto é,* $d_{i,j}$ *é conhecida.*

Definição 4. *Um conjunto de quatro átomos será chamado* ***independente****, se suas coordenadas no* $\mathbb{R}^3$ *forem independentes, ou seja, os pontos que os representam no plano são não coplanares.*

Algumas ferramentas, descritas com maiores detalhes em [24], necessárias para o estudo dos métodos que apresentaremos são:

1) Determinante
O determinante de uma matriz A é definido da seguinte forma:

Definição 5. *Se* $A = (a) \in \mathbb{R}^{1\times 1}$*, então seu determinante é dado por* $det(A) = a$*. De modo mais geral, o determinante de* $A \in \mathbb{R}^{n\times n}$ *é definido por*

$$det(A) = \sum_{j=1}^{n} (-1)^{j+1} a_{1j} det(A_{1j}), \tag{2.4}$$

onde A_{1j} *é uma matriz* $(n-1)\times(n-1)$ *obtida através da exclusão da primeira linha e da j-ésima coluna de A.*

Algumas propriedades dos determinantes de matrizes são as seguintes: $\forall A, B \in \mathbb{R}^{n\times n}$ e $c \in \mathbb{R}$, temos que
$det(AB) = det(A).det(B)$,
$det(A^T) = det(A)$,
$det(cA) = c^n det(A)$ e
$det(A) \neq 0 \Leftrightarrow$ A é não singular

2) Traço de uma matriz
Em álgebra linear o traço de uma matriz quadrada A é a função matricial que associa a matriz à soma dos elementos da sua diagonal principal e é denotado por $tr(A)$, dessa forma se $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ então temos que

$$tr(A) = a_{11} + a_{22} + \ldots + a_{nn}. \tag{2.5}$$

Algumas importantes propriedades desta função devem ser destacadas: $\forall A, B \in \mathbb{R}^{n\times n}$ e $\lambda \in \mathbb{R}$ temos que
$tr(A+B) = tr(A) + tr(B)$,
$tr(\lambda A) = \lambda tr(A)$,
$tr(A) = tr(A^T)$,
$tr(AB) = tr(BA)$ e
$tr(A) = \sigma_1 + \ldots \sigma_p$, onde $\sigma_1, \ldots, \sigma_p$ são os autovalores de A.

3) Normas de Matrizes
Uma função $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ é uma norma de matrizes se, e somente se para todo $A, B \in \mathbb{R}^{m\times n}$ e $\alpha \in \mathbb{R}$ são satisfeitas as seguintes propriedades:

$f(A) \geq 0$ ($f(A) = 0 \Leftrightarrow A = 0$),
$f(A+B) \leq f(A) + f(B)$ e
$f(\alpha A) = |\alpha| f(A)$.

As normas de matrizes são denotadas por $||A|| = f(A)$. E uma das normas de matrizes mais usada é a Norma de Frobenius definidida por:

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}^2|}. \tag{2.6}$$

Esta norma satisfaz a propriedade sub-multiplicativa, isto é,

$$||AB||_F \leq ||A||_F ||B||_F.$$

Além disso, podemos calcular a Norma de Frobenius utilizando a função traço:

$$||A||_F = \sqrt{tr(AA^T)}. \tag{2.7}$$

4)Número de Condição

Na análise numérica, o número de condicionamento ou número de condição da matriz de coeficientes de um sistema linear é uma medida indicando se o sistema tem "boas condições" para ser tratado numericamente. Um problema com um número de condição pequeno é chamado de bem condicionado, enquanto os problemas que possuem um número de condição elevado são denominados mal condicionados. O número de condição de uma matriz $A \in \mathbb{R}^{m\times n}$ é denotado por $\mathscr{K}(A)$ e definido por

$$\mathscr{K}(A) = ||A|| \cdot ||A^{-1}||, \tag{2.8}$$

onde $||.||$ é uma norma de matriz. Portanto, o número de condição depende da norma escolhida, por exemplo, no caso da Norma de Frobenius o número de condição é dado por $\mathscr{K}_F(A) = ||A||_F \cdot ||A^{-1}||_F$.

5) Centro Geométrico

Em geometria , o centro geométrico (ou centróide) de uma figura (ou conjunto de pontos) M é a intersecção de todas as retas que dividem M em duas partes de iguais momentos. Informalmente, é a "média" (média aritmética) de todos os pontos de M. A definição se estende a qualquer objeto N em um espaço n-dimensional, tal que o seu centróide é a intersecção de todos os hiperplanos que dividem N em duas partes de momentos iguais.

Portanto, o centro geométrico de um conjunto finito de k pontos, digamos $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$ é dado por

$$n_c = \frac{v_1 + v_2 + \ldots + v_k}{k}. \tag{2.9}$$

Portanto, se temos uma estrutura molecular X com n átomos, $x_1, x_2, \ldots, x_n$, onde suas posições estão em $\mathbb{R}^3$, então o centro geométrico dessa estrutura é

$$x_c = \frac{v_1 + v_2 + \ldots + v_n}{n},$$

ou seja,

$$x_c(j) = \frac{1}{k}\sum_{i=1}^{n} v_i(j) \tag{2.10}$$

para $j = 1, 2, 3$.

6) LU

É muito comum nas aplicações da matemática problemas que necessitam da resolução de sistemas lineares e por muitas vezes, a estratégia mais recorrida é fatorar a matriz de coeficientes do sistema em outras que são mais fáceis de serem resolvidas ou que requerem menos tempo computacional. Uma das fatorações de matriz mais conhecida e utilizada é a fatoração LU, que consiste em decompor uma matriz em outras duas, uma triangular inferior e outra triangular superior, motivo do nome da fatoração: L de *lower* (inferior em inglês) e U de *upper* (superior em inglês).

A fatoração LU se aproveita da facilidade de resolver sistemas lineares triangulares e é baseada no seguinte teorema:

Teorema 2.3.1. *[Golub, 1996] Seja $A \in \mathbb{R}^{n\times n}$ uma matriz quadrada de ordem n, e A_k o menor principal k, isto é, a matriz $k \times k$ formada pelas primeiras k colunas e k linhas de A . Se $det(A_k) \neq 0$, para $k = 1,2,...,n-1$, então existe uma única matriz triangular inferior com diagonal unitária $L \in \mathbb{R}^{n\times n}$, e uma única matriz triangular superior $U \in \mathbb{R}^{n\times n}$, tal que LU = A. Além disso, $det(LU) = u_{11}u_{22}\dots u_{nn}$, onde u_{ii} são os elementos da diagonal principal de U.*

Na verdade, o fato da matriz A ser não singular, ou seja, se $det(A) \neq 0$, já é suficiente para a existência da fatoração LU, para alguma permutação de A. Portanto, se é preciso resolver um sistema linear $Ax = b$ com a matriz A sendo quadrada de ordem n e não singular, então poderemos fatorar a matriz em $A = LU$, e assim, a solução x^* deste sistema pode ser encontrado através da resolução dos seguintes sistemas lineares:

$$Ly = b \qquad \text{e} \qquad Ux = y$$

que requerem menos tempo computacional para serem resolvidos, pois L e U são matriz triangulares. O algoritmo da fatoração LU requer no máximo $\frac{2}{3}n^3$ operações [24] para a decomposição de uma matriz A de ordem n.

7) *Singular Value Decomposition* (SVD)

Outro método de decomposição de matrizes é o *Singular Value Decomposition* (SVD), mas ao contrário da fatoração LU também trabalha com matrizes retangulares e/ou singulares, na verdade, toda matriz admite uma fatoração SVD, porém com um custo computacional maior que a LU. A decomposição SVD é baseada no seguinte teorema.

Teorema 2.3.2. *[Golub, 1996] Se $A \in \mathbb{R}^{m\times n}$ é uma matriz real, então existem ortogonais $U \in \mathbb{R}^{m\times m}$ e $V \in \mathbb{R}^{n\times n}$ tal que*

$$A = U\Sigma V^T$$

onde $\Sigma = (\sigma_1,\dots,\sigma_p)$ com $p \leq min\{m,n\}$ é uma matriz diagonal e $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$ são os valores singulares não nulos de A, isto é, são os autovalores das matrizes AA^T e A^TA.

Temos também, que as colunas de V formam uma base para os autovetores de A^TA, enquanto que as colunas de U formam uma base para os autovetores de AA^T. E o algoritmo da decomposição SVD requer $\frac{8}{3}n^3$ operações [24] para fatorar uma matriz $A \in \mathbb{R}^{n\times n}$.

Existem outras ferramentas matemáticas que iremos utilizar neste trabalho como o *Root-Mean-Square Deviation* (RMSD), que é uma forma de medir a distância entre duas estruturas sobrepostas, e tem sido amplamente utilizada na modelagem de proteínas, para comparar e validar as estruturas de proteínas [19, 41], porém iremos defini-las em um outro momento mais oportuno.

Com estas definições e ferramentas preliminares, podemos iniciar a apresentação dos métodos utilizados para resolver o PGDM com distâncias exatas. Dividiremos a apresentação dos métodos em dois capítulos, um dedicado à "família Geometric Build-Up" e outro para a "família T".

Capítulo 3

Revisão da Literatura

Neste capítulo iremos apresentar os métodos da família "Geometric Build-Up", formada por todos os métodos que buscam resolver o PGDM utilizando a transformação linear do teorema 3.1.5, que será apresentado na seção 3.1.2. Além do GB, fazem parte dessa família o *"Updated Geometric Build-Up Algorithm"* (UGB) e um método com complexidade linear, que resolve o PGDM com um conjunto completo de distâncias exatas, também apresentado neste capítulo. Porém antes de introduzirmos os métodos dessa família, iremos apresentar um outro algoritmo que se utiliza da fatoração em valores singulares (SVD, do inglês *Singular Value Decomposition*) da matriz de distâncias para resolver o PGDM com um conjunto completo de distâncias exatas. O primeiro algoritmo da "família Geometric Build-Up" a ser apresentado é um método de complexidade linear $O(n)$, que foi uma alternativa apresentada por Dong e Wu em [17] ao método que utiliza a fatoração SVD e, portanto tem ordem de complexidade $O(n^3)$.

Posteriormente apresentaremos dois métodos utilzados para a resolução do PGDM com um conjunto arbitrário de distâncias, que são o GB [18] e sua atualização, o UGB [45], de Wu *et al*. A principal diferença entre os dois é que no segundo, as coordenadas da base são recomputadas a cada passo, a fim de tentar conter o acúmulo de erros numéricos. Foi com essa reinicialização das coordenadas dos átomos base que realizamos a mesma atualização no Algoritmo T [21].

3.1 Métodos para Resolução do Problema de Geometria de Distâncias Moleculares com Conjunto Completo de Distâncias Exatas

Iniciaremos a discussão sobre a resolução do Problema de geometria de distâncias moleculares com conjunto completo de distâncias exatas com dois algoritmos usados para resolver tal problema. O primeiro apresenta a estratégia de decomposição da matriz de distância através da fatoração SVD, porém, o custo computacional deste algoritmo é muito alto, já que a sua ordem de complexidade é $O(n^3)$. Assim, Q. Dong e Z. Wu [17] desenvolveram uma alternativa a esse primeiro método, que utiliza a estratégia de aplicar uma transformação linear em sistemas não-lineares transformando-os em sistemas lineares, este foi o primeiro passo para o nascimento da "família Geometric Build-Up".

3.1.1 Método usando a Decomposição em Valores Singulares da Matriz de Distâncias Moleculares

Iremos, agora, tratar sobre um algoritmo [6, 12, 17] que se utiliza da Decomposição em Valores Singulares (SVD) da matriz de distâncias para resolver o PGDM com um conjunto completo de

distâncias exatas. A matriz B de distâncias é formada pelas distâncias inter-atômicas dadas como instâncias iniciais do algoritmo, definida do seguinte modo

$$B_{i,j} = \frac{d_{1,i}^2 - d_{i,j}^2 + d_{1,j}^2}{2} \qquad \forall i,j = 1,2,\ldots,n.$$

Vejamos os dois próximos resultados antes de continuar a explanação sobre o método.

Proposição 3.1.1. *[Dong e Wu, 2002] Sendo $X = [x_1^T \ldots x_n^T]^T \in \mathbb{R}^{n\times 3}$ a matriz contendo as coordenadas de todos os n átomos de uma molécula, e sendo B a matriz de distâncias como definido acima, temos que $B = XX^T$.*

Demonstração. Sejam $x_1, x_2, \ldots, x_n$ as coordenadas dos átomos da molécula, se $d_{i,j}$ é a distância entre os átomos i e j da molécula, então temos as seguintes equações:

$$\|x_i - x_j\| = d_{i,j}, \qquad \forall i,j = 1,2,...,n.$$

Elevando ambos lados ao quadrado, teremos

$$\|x_i\|^2 - 2x_i^T x_j + \|x_j\|^2 = d_{i,j}^2, \qquad \forall i,j = 1,2,...,n, \tag{3.1}$$

como a estrutura molecular de uma proteína é invariante sob movimentos rígidos, isto é, de rotação, reflexão ou translação, então, podemos definir o sistema de coordenadas de tal forma que o primeiro átomo esteja na origem desse sistema, ou seja, $x_1 = (0,0,0)$. Assim,

$$\|x_i\|^2 = \|(0,0,0)^T - x_i\|^2 = \|x_1 - x_i\|^2 = d_{1,i}^2, \qquad \forall i = 1,2,3,...,n.$$

Substituindo $\|x_i\|^2 = d_{1,i}^2$ na equação 3.1, obetmos

$$d_{1,i}^2 - 2x_i^T x_j + d_{1,j}^2 = d_{i,j}^2, \qquad \forall i,j = 1,2,3,...,n.$$

E, então, agrupando as distâncias em um lado da igualdade, resulta em

$$\frac{d_{1,i}^2 - d_{i,j}^2 + d_{1,j}^2}{2} = x_i^T x_j, \qquad \forall i,j = 1,2,3,...,n.$$

Perceba agora que no lado esquerdo da igualdade temos $B_{i,j}$, assim,

$$B_{i,j} = x_i^T x_j, \qquad \forall i,j = 1,2,3,...,n.$$

Então, teremos que

$$B = \begin{bmatrix} x_1^T x_1 & \ldots & x_1^T x_n \\ \vdots & \vdots & \vdots \\ x_n^T x_1 & \ldots & x_n^T x_n \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \begin{bmatrix} x_1 \cdots x_n \end{bmatrix} = XX^T$$

□

Lema 3.1.2. *[Dong e Wu, 2002] Sendo B e X como definidos acima, então, temos que Im(B) $\subseteq$ Im(X).*

Demonstração. Seja $x \in \mathbb{R}^n$, tal que $x \in$ Im (B), então, existe $y \in \mathbb{R}^n$, onde $By = x$, assim, substituindo $B = X.X^T$, temos

$X.X^T y = x \;\Rightarrow\; X(X^T y) = x \;\Rightarrow\; x \in$ Im (X).

Portanto, Im(B) $\subseteq$ Im(X).

□

Por meio deste último resultado que acabamos de demonstrar, podemos ver que $Im(B) \subseteq Im(X)$, então $posto(B) \leq posto(X)$, e como sabemos que $posto(X) \leq 3$, já que X tem apenas três colunas, então $posto(B) \leq posto(X) \leq 3$. Assim, como B é simétrica, pois, $B^T = (XX^T)^T = (X^T)^T X^T = XX^T = B$ e como $posto(B) \leq 3$, então, a decomposição SVD reduzida de B é a seguinte:

$$B = U\Sigma U^T, \tag{3.2}$$

onde $U \in \mathbb{R}^{nx3}$ é uma matriz ortogonal e $\Sigma \in \mathbb{R}^{3x3}$ é uma matriz diagonal, cujos elementos são os valores singulares de B, $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$.

Nosso objetivo é obter as coordenadas dos átomos da molécula, para isso podemos calcular a solução de $B = XX^T$, mas como já conhecemos a fatoração de B, tal solução é dada por:

$$X = U\Sigma^{\frac{1}{2}}, \tag{3.3}$$

onde $\Sigma^{\frac{1}{2}}(i,i) = \sigma_i^{\frac{1}{2}}$ para $i = 1,2,3$.

Portanto, se tivermos um conjunto completo de distâncias exatas entre os átomo de uma determinada molécula, podemos calcular sua estrutura tridimensional da seguinte forma: (1) Calcular a matriz de distâncias B como foi definido anteriormente, (2) Fatorar B em valores singulares, e finalmente, (3) calcular as coordenadas dos átomos em nosso sistema de coordenadas, através da equação (3.3).

Em [24] é possível observar que a decomposição em SVD pode ser feita em no máximo $\frac{8}{3}n^3$ operações, então sua ordem de complexidade é $O(n^3)$. Portanto, a solução para o PGDM, se dado todas as distâncias exatas entre pares de átomos, através deste método será obtido em tempo polinomial, onde a ordem de complexidade também será $O(n^3)$.

3.1.2 Método com ordem de complexidade linear

Em [17], os autores Q. Dong e Z. Wu apresentaram um método com ordem de complexidade linear para resolver o PGDM com um conjunto completo de distâncias exatas, que poderia ser usado como alternativa a este último apresentado, já que sua ordem de complexidade é linear, isto é, com o custo computacional de $O(n)$. As duas idéias centrais do método nasceram de observações de propriedades geométricas no $\mathbb{R}^2$.

Em um espaço de duas dimensões, podemos fazer duas observações: (1) Se sabemos as distâncias entre três pontos, dois a dois, então podemos calcular as coordenadas desses três pontos, a menos de movimentos rígidos, por meio da resolução de um sistema linear de equações algébricas, e (2) se estes três pontos não estiverem na mesma reta e soubermos as distâncias entre eles e um quarto ponto, então podemos calcular, de forma única, a posição deste quarto ponto, por meio novamente da resolução de um sistema de equações algébricas [17]. Foram essas duas idéias que ajudaram Q. Dong e Z. Wu a formular os próximos dois resultados, que verificam a validade dessas ideias no $\mathbb{R}^3$, e que formam o fundamento no desenvolvimento de seu algoritmo.

Teorema 3.1.3. *[Dong e Wu, 2002] Se conhecermos todas as distâncias entre quatro átomos de uma molécula, então podemos determinar de modo único, a menos de movimentos rígidos, as coordenadas destes quatro átomos preservando a estrutura da molécula.*

Demonstração. Sejam x_1, x_2, x_3 e x_4 os quatro átomos que queremos determinar, e $d_{i,j}$, onde $i, j = 1,2,3,4$ e $i \neq j$, as distâncias entre eles. Como estes átomos estão em um espaço de três dimensões, temos que

$$x_1 = (u_1, v_1, w_1), \quad x_2 = (u_2, v_2, w_2), \quad x_3 = (u_3, v_3, w_3), \quad x_4 = (u_4, v_4, w_4).$$

Sem perda da estrutura da molécula, podemos via de movimentos rígidos realizar uma mudança de coordenadas de modo que, o primeiro átomo esteja na origem do espaço, isto é, $x_1 = (0,0,0)$, o segundo átomo esteja sobre o eixo das abscissas (eixo-x), assim, suas coordenadas serão da forma $x_2 = (u_2,0,0)$ e, por fim, o terceiro átomo esteja no plano abscissa-ordenada (plano xy), ou seja, $x_3 = (u_3,v_3,0)$. De posse dessas hipóteses, podemos definir o seguinte sistema não-linear

$$\begin{array}{l} \|x_2 - x_1\|_2 = d_{2,1} \\ \|x_3 - x_1\|_2 = d_{3,1} \\ \|x_3 - x_2\|_2 = d_{3,2} \end{array} , \quad \text{ou seja,} \quad \begin{array}{r} u_2^2 = d_{2,1}^2 \\ u_3^2 + v_3^2 = d_{3,1}^2 \\ (u_3 - u_2)^2 + v_3^2 = d_{3,2}^2 \end{array} . \tag{3.4}$$

Portanto, se resolvermos este sistema, obtemos os valores

$$u_2 = \pm d_{2,1}, \quad u_3 = \frac{d_{3,1}^2 - d_{3,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2} \quad \text{e} \quad v_3 = \pm (d_{3,1}^2 - u_3^2)^{1/2}. \tag{3.5}$$

Podemos tanto escolher u_2 e v_3 positivos, quanto ambos negativos, pois as distâncias não serão alteradas, apenas o que ocorre de uma escolha para outra é uma reflexão em torno do eixo y, mas mantendo a mesma estrutura rígida. No nosso caso, escolhemos ambos positivos.

Depois de termos determinado a posição dos três primeiros pontos podemos determinar a posição de $x_4 = (u_4, v_4, w_4)$, utilizando as distâncias conhecidas aos outros átomos podemos definir suas coordenadas através do seguinte sistema:

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

Que é equivalente ao seguinte sistema quadrático:

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

E, se resolvermos o sistema acima, obtemos os seguintes resultados:

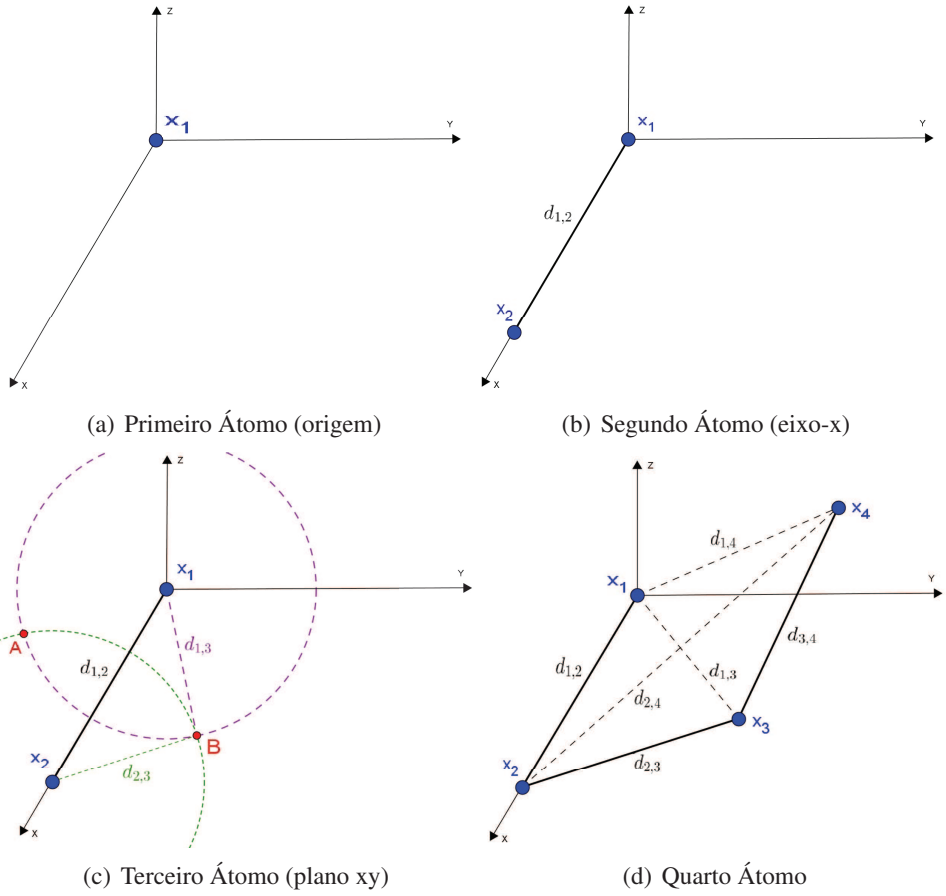
$$u_4 = \frac{d_{4,1}^2 - d_{4,2}^2}{2u_2} + \frac{u_2}{2}, \; v_4 = \frac{d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2}{2v_3} + \frac{v_3}{2} \text{ e } w_4 = \pm (d_{4,1}^2 - u_4^2 - v_4^2)^{1/2}$$

Novamente, podendo escolher w_4 positivo ou negativo, pois isso não afetará a estrutura, já que a diferença de uma escolha para a outra é uma reflexão.

□

Na figura 3.1 é possível visualizar como ocorre a determinação dos quatro primeiros átomos através do teorema 3.1.3, no item (a) da figura foi posicionado o primeiro átomo na origem, no item (b) foi determinada a posição do segundo átomo no eixo-x através da distância $d_{1,2}$, posteriormente foi calculada em (c) a posição do terceiro átomo no plano xy utilizando as coordenadas dos dois átomos já calculados e as distâncias $d_{1,3}$ e $d_{2,3}$. É possível visualizar que existem duas possibilidades para a posição do terceiro átomo, assim como diz o teorema. Por fim, a partir das coordenadas destes três átomos e das distâncias deles ao quarto átomo, determinamos na figura 3.1(d) a posição x_4 deste átomo.

O cálculo de quatro átomos como foi mostrado na demonstração deste teorema será o primeiro passo não apenas deste algoritmo, como também de todos os outros que apresentaremos adiante. O seguinte lema nos auxiliará no próximo teorema:



(a) Primeiro Átomo (origem) (b) Segundo Átomo (eixo-x)

(c) Terceiro Átomo (plano xy) (d) Quarto Átomo

Figura 3.1: Construção dos quatro primeiros átomos pelo Teorema 3.1.3

Lema 3.1.4. *[Dong e Wu, 2003] Se* $\{x_1, x_2, x_3, x_4\} \subset \mathbb{R}^3$ *são pontos não coplanares, então o conjunto de vetores* $B = \{(x_1 - x_2), (x_1 - x_3), (x_1 - x_4)\} \subset \mathbb{R}^3$ *é linearmente independente.*

Já o próximo resultado será o cerne deste e de todos algoritmos da família *Geometric Build-Up*, sendo utilizado em todos passos subsequentes ao cálculo das coordenadas dos 4 primeiros átomos através do teorema 3.1.3. Pois em cada um destes $n-4$ passos poderemos calcular a posição de um átomo não determinado por meio da resolução de um sistema quadrático, que por este teorema pode ser linearizado a um sistema que tem a matriz de coeficentes não-singular, portanto, com solução única.

Teorema 3.1.5. *[Dong e Wu, 2002] Sejam* 4 *átomos determinados e não coplanares, cujas posições em* $\mathbb{R}^3$ *são* x_1, x_2, x_3 *e* x_4*. Então se um quinto átomo não determinado for vizinho destes* 4 *átomos, temos que é possível achar as coordenadas deste átomo,* x_5*, que será a solução do seguinte sistema linear:*

$$Ax = c, \qquad com \tag{3.6}$$

$$A = 2 \begin{bmatrix} (x_1 - x_2)^T \\ (x_1 - x_3)^T \\ (x_1 - x_4)^T \end{bmatrix} \quad e \quad c = \begin{bmatrix} d_2^2 - d_1^2 + ||x_1||^2 - ||x_2||^2 \\ d_3^2 - d_1^2 + ||x_1||^2 - ||x_3||^2 \\ d_4^2 - d_1^2 + ||x_1||^2 - ||x_4||^2 \end{bmatrix}.$$

onde $A \in \mathbb{R}^{3x3}$ *é não-singular e* $c \in \mathbb{R}^3$.

Demonstração. Uma das formas de determinar a posição do átomo não determinado x_i sabendo apenas a distância entre ele e outros quatro átomos determinados é resolver o seguinte sistema quadrático:

$$\begin{array}{l} ||x_i - x_1|| = d_{i,1} \\ ||x_i - x_2|| = d_{i,2} \\ ||x_i - x_3|| = d_{i,3} \\ ||x_i - x_4|| = d_{i,4} \end{array} . \tag{3.7}$$

Podemos ver no apêndice A, que pelo teorema A.2 este sistema quadrático possui, a menos de erro de medidas, apenas uma solução. Mas pela dificuldade de resolver tal sistema, iremos aplicar uma transformação linear em tal sistema para que seja reduzido a um sistema linear. Supomos que $a^* \in \mathbb{R}^3$ seja a solução deste sistema.

Primeiramente, iremos elevar ambos lados das equações ao quadrado

$$\begin{array}{l} ||a^* - x_1||^2 = d_{i,1}^2 \\ ||a^* - x_2||^2 = d_{i,2}^2 \\ ||a^* - x_3||^2 = d_{i,3}^2 \\ ||a^* - x_4||^2 = d_{i,4}^2 \end{array} .$$

Como sabemos que $||u-v||^2 = ||u||^2 - 2u^T v + ||v||^2$, então

$$\begin{array}{l} ||a^*||^2 - 2x_1^T a^* + ||x_1||^2 = d_{i,1}^2 \\ ||a^*||^2 - 2x_2^T a^* + ||x_2||^2 = d_{i,2}^2 \\ ||a^*||^2 - 2x_3^T a^* + ||x_3||^2 = d_{i,3}^2 \\ ||a^*||^2 - 2x_4^T a^* + ||x_4||^2 = d_{i,4}^2 \end{array} .$$

Substituindo $||a^*||^2 = 2x_1^T a^* - ||x_1||^2 + d_1^2$ da primeira equação, nas outras três, obtemos

$$\begin{array}{l} 2x_1^T a^* - ||x_1||^2 + d_{i,1}^2 - 2x_2^T a^* + ||x_2||^2 = d_{i,2}^2 \\ 2x_1^T a^* - ||x_1||^2 + d_{i,1}^2 - 2x_3^T a^* + ||x_3||^2 = d_{i,3}^2 \\ 2x_1^T a^* - ||x_1||^2 + d_{i,1}^2 - 2x_4^T a^* + ||x_4||^2 = d_{i,4}^2 \end{array} ,$$

ou seja,

$$\begin{array}{l} 2(x_1 - x_2)^T a^* = ||x_1||^2 - ||x_2||^2 + d_{i,2}^2 - d_{i,1}^2 \\ 2(x_1 - x_3)^T a^* = ||x_1||^2 - ||x_3||^2 + d_{i,3}^2 - d_{i,1}^2 \\ 2(x_1 - x_4)^T a^* = ||x_1||^2 - ||x_4||^2 + d_{i,4}^2 - d_{i,1}^2 \end{array} .$$

Assim, sendo

$$A = 2 \begin{bmatrix} (x_1 - x_2)^T \\ (x_1 - x_3)^T \\ (x_1 - x_4)^T \end{bmatrix} \quad e \quad c = \begin{bmatrix} d_{i,2}^2 - d_{i,1}^2 + ||x_1||^2 - ||x_2||^2 \\ d_{i,3}^2 - d_{i,1}^2 + ||x_1||^2 - ||x_3||^2 \\ d_{i,4}^2 - d_{i,1}^2 + ||x_1||^2 - ||x_4||^2 \end{bmatrix},$$

onde $A \in \mathbb{R}^{3x3}$ e $c \in \mathbb{R}^3$, temos que $Aa^* = c$.

Portanto, se a^* é solução do sistema quadrático (3.7), então, também será solução do sistema linear $Ax = c$, com A e c definidos acima.

Já vimos que o sistema (3.7) tem apenas uma solução e que ela também será solução do sistema linear (3.6). Agora, nos resta provar que o sistema (3.6) tem apenas uma solução, ou seja, que A é não singular. Como x_1, x_2, x_3 e x_4 são não-coplanares, então pelo lema (3.1.4) temos que $x_1 - x_2, x_1 - x_3$ e $x_1 - x_4$ são linearmente independentes, logo, A é não singular.

□

Tendo estes dois resultados em mãos, Q. Dong e Z. Wu [17] mostraram que para resolver o PGDM com um conjunto completo de distâncias, basta escolhermos 4 pontos não coplanares e então determinamos suas coordenadas x_1, x_2, x_3 e x_4 através do teorema 3.1.3 e os fixamos como átomos base do restante dos átomos, já que eles satisfazem as condições do teorema 3.1.5, pois as distâncias entre estes 4 átomos e todos os outros não determinados são conhecidas, então para cada átomos remanescente i, ou seja, para cada átomo não determinado, resolvemos o sistema (3.6), já que o teorema 3.1.5 nos garante que a solução deste sistema será as coordenadas do átomo indeterminado x_i que estamos buscando.

Apresentamos abaixo o algoritmo deste método que acabamos de apresentar.

Algoritmo 1: Método com complexidade linear para conjunto completo

Entrada: O conjunto D de distâncias euclidianas entre todos os pares de átomos de uma molécula com n átomos.
Saída: Estrutura completa da molécula de proteína.

1. Encontre quatro átomos não-coplanares.
2. Determine suas coordenadas x_1, x_2, x_3 e x_4, segundo as fórmulas dadas pelo Teorema 3.1.3.
3. Para cada átomo não-determinado j, usando as coordenadas de x_1, x_2, x_3 e x_4, podemos calcular a posição de x_j através do sistema linear 3.6.

O próximo método que apresentaremos, chamado *Geometric Build-Up Algorithm*, pode ser utilizado para resolver o PGDM com um conjunto arbitrário de distâncias, e é uma generalização deste último método, pois quando se trata do problema com um conjunto completo de distâncias temos que ambos são iguais. A principal diferença é que como podemos não ter todas distâncias conhecidas, então em cada passo, teremos que buscar 4 átomos que possam ser base para o átomo não determinado que está sendo buscado.

3.2 Métodos para Resolução do Problema de Geometria de Distâncias Moleculares com Conjunto Arbitrário de Distâncias Exatas

Os dados de distâncias moleculares derivados da interação atômica medidos por experimentos de RMN podem ser utilizados para calcular a estrutura tridimensional de uma proteína. Se as distâncias entre todos os pares de átomos estão disponíveis, então uma estrutura de proteína única pode ser calculada em tempo polinomial, conforme o algoritmo que vimos na seção anterior, o qual foi baseado em uma simples relação geométrica entre as coordenadas e as distâncias. No entanto, geralmente, os experimentos de RMN só podem fornecer um conjunto arbitrário de distâncias entre pares de átomos de uma determinada proteína. Mesmo podendo combinar os dados de RMN com o nosso conhecimento sobre certos vínculos entre comprimentos e ângulos, nós ainda não podemos obter todas as distâncias [18]. Algoritmos heurísticos, tais como o algoritmo EMBED [12], têm sido desenvolvidos para estimar as distâncias "em falta", isto é, as distâncias que não puderam ser calculadas via RMN, e assim, obter um conjunto completo de distâncias que possam ser utilizadas para calcular a estrutura da proteína. Infelizmente, tais estimativas são geralmente caras e propensas a introduzir erros em algumas dessas distâncias "em falta". Em relação às distâncias maiores, pode ser muito difícil obtê-las a partir das distâncias disponíveis. O algoritmo EMBED usa uma técnica chamada *"Bound Smoothing"*, que em geral, possui um custo computacional de $O(n^4)$ para estimar as distâncias que não foram possíveis de calcular pelo RMN [11, 12]. Esta técnica depende de conceitos geométricos como a desigualdade triangular,

que na melhor das hipóteses só pode estimar um limite superior e inferior para as distâncias. Em seguida, as distâncias encontradas são provavelmente usadas para estimar as outras distâncias "em falta", havendo assim, uma propagação de erros e fazendo da estimativa não confiável. Uma melhor abordagem seria resolver a estrutura diretamente usando o conjunto arbitrário das distâncias. Nesse caso, em vez de estimar cada distância "em falta", nós podemos apenas usar as distâncias mais confiáveis obtidas a partir de experimentos de RMN.

Neste trabalho, todos os métodos descritos na resolução do PGDM com um conjunto arbitrário de distâncias utilizam essa estratégia de trabalhar apenas com as distâncias conhecidas. Inicialmente apresentaremos os algoritmos *Geometric Build-Up Algorithm* (GB) e *Updated Geometric Build-Up Algorithm* (UGB). O algoritmo GB foi baseado no algoritmo de tempo linear relatado anteriormente, porém ao invés de usar uma mesma base para encontrar toda a estrutura, no GB a base pode ser alterada para cada átomo remanescente. É possível que alguns átomos não determinados não possuam vizinhos o suficiente para determiná-lo, assim a estrutura da proteína poderá não ter solução.

Os passos dos algoritmos GB e UGB consiste em escolher inicialmente quatro átomos independentes e determinar suas posições, através do teorema 3.1.3, depois escolher um átomo não-determinado que seja vizinho aos quatro primeiros átomos e determinar sua posição resolvendo o sistema linear (3.7) do teorema 3.1.5. Por fim, para cada átomo não-determinado restante, buscar 4 átomos posicionados, independentes e vizinhos deste átomo a determinar, e assim, calcular sua posição resolvendo o mesmo sistema linear, porém atualizado com os novos átomos base. Já no algoritmo UGB nestes últimos passos antes de resolver o sistema linear (3.7), são recalculadas através do teorema 3.1.3 as coordenadas dos átomos que foram escolhidos para ser base do átomo a ser determinado. Após feito esse cálculo, resolvemos o sistema linear (3.7) encontrando a posição do átomo indeterminado em relação às coordenadas da base recalculada. Verificamos as transformações que diminuem a diferença entre a estrutura dos átomos recalculados e de suas posições anteriores, e assim, aplicamos essas transformações no átomo que acabamos de determinar para verificar suas coordenadas em relação a estrutura principal.

3.2.1 *Geometric Build-Up Algorithm* (GB)

Wu apresentou o *"Geometric Build-up Algorithm"* (GB) em 2003 [18] como uma alternativa aos outros métodos existentes para resolver o PGDM com um conjunto arbitrário de distâncias, já que na época da sua publicação as abordagens precisavam estimar todas as outras distâncias que não puderam ser determinadas via RMN, para assim, construir um conjunto completo de distâncias, e assim calcular a estrutura da proteína. No entanto, os passos para a estimação eram caros e sujeitos a muitos erros. O GB foi inspirado pelo método que acabamos de mostrar na seção anterior, de autoria dos mesmos autores e que resolve o PGDM com um conjunto completo de distâncias em tempo polinomial, porém ele não necessita de todas as distâncias entre os átomos e não precisa estimá-las também. Tanto o GB quanto sua atualização, o UGB, foram o que nos deram base para que desenvolvessemos o Algoritmo T e sua atualização, o Algoritmo T Atualizado.

Primeiramente, escolhemos quatro átomos não coplanares. As coordenadas destes átomos podem ser facilmente calculadas usando as distâncias entre eles, como vimos no Teorema 3.1.3. Em seguida, para cada átomo remanescente procuramos quatro átomos que possam servir de base para esse átomo, ou seja, quatro átomos não coplanares e vizinhos com o átomo não determinado. Posteriormente, o algoritmo resolve um sistema linear pequeno (4×4) de equações algébricas. Como este algoritmo irá repetir no máximo n iterações, então a quantidade de operações é proporcional ao número de átomos na molécula.

A principal diferença entre o algoritmo GB e o algoritmo da seção anterior com complexidade linear que resolve o PGDM com um conjunto completo de distâncias é que neste algoritmo,

os primeiros quatro átomos determinados pelo Teorema 3.1.3 podem servir de base para todos os outros $n-4$ átomos remanescentes, dado que sabemos as distâncias entre todos os átomos da molécula de proteína. Porém, no caso do conjunto arbitrário, podemos não saber todas as distâncias, e portanto, no algoritmo GB devemos para cada átomo remanescente, procurar quatro átomos já determinados que podem servir de base para este átomo. Porém, nem sempre podemos encontrar quatro átomos base para todos átomos, e nesse caso, o GB pode informar somente uma estrutura parcial.

Assim, para que o método GB consiga reportar toda a estrutura, é preciso que para todo átomo i não determinado existam pelo menos outros quatro átomos, k_1, k_2, k_3 e k_4, determinados que possam servir de base para o átomo i, ou seja, que sejam não coplanares e vizinhos do átomo requerido.

Seja x_i as coordenadas do átomo que queremos determinar e $x_{k_1}, x_{k_2}, x_{k_3}$ e x_{k_4} as coordenadas dos átomos base para o átomo i. Então, para calcular a posição do átomo i devemos resolver o seguinte sistema quadrático:

$$\begin{array}{l} ||x_i - x_{k_1}|| = d_{i,k_1} \\ ||x_i - x_{k_2}|| = d_{i,k_2} \\ ||x_i - x_{k_3}|| = d_{i,k_3} \\ ||x_i - x_{k_4}|| = d_{i,k_4} \end{array} . \tag{3.8}$$

A transformação linear que realizaremos neste sistema é similar ao do método anterior, mostrado no Teorema 3.1.5, onde primeiramente são elevadas as equações ao quadrado, e posteriormente usando uma equação como pivô, primeiro é isolado $||x||^2$ nesta equação e depois o substituímos nas outras equações. Por exemplo, se utilizarmos a primeira equação como pivô temos o seguinte:

Substituindo $||x||^2 = 2x_{k_1}^T x_i - ||x_{k_1}||^2 + d_1^2$ da primeira equação, nas outras três, obtemos

$$\begin{array}{l} 2a_1^T x_i - ||a_1||^2 + d_1^2 - 2a_2^T x_i + ||a_2||^2 = d_2^2 \\ 2a_1^T x_i - ||a_1||^2 + d_1^2 - 2a_3^T x_i + ||a_3||^2 = d_3^2 \\ 2a_1^T x_i - ||a_1||^2 + d_1^2 - 2a_4^T x_i + ||a_4||^2 = d_4^2 \end{array} ,$$

ou seja,

$$\begin{array}{l} 2(x_{k_1} - x_{k_2})^T x_i = ||x_{k_1}||^2 - ||x_{k_2}||^2 + d_2^2 - d_1^2 \\ 2(x_{k_1} - x_{k_3})^T x_i = ||x_{k_1}||^2 - ||x_{k_3}||^2 + d_3^2 - d_1^2 \\ 2(x_{k_1} - x_{k_4})^T x_i = ||x_{k_1}||^2 - ||x_{k_4}||^2 + d_4^2 - d_1^2 \end{array} .$$

Assim, obtemos uma equação linear, da forma $Ax = c$, com

$$A = 2 \begin{bmatrix} (x_{k_1} - x_{k_2})^T \\ (x_{k_1} - x_{k_3})^T \\ (x_{k_1} - x_{k_4})^T \end{bmatrix} \quad e \quad c = \begin{bmatrix} d_2^2 - d_1^2 + ||x_{k_1}||^2 - ||x_{k_2}||^2 \\ d_3^2 - d_1^2 + ||x_{k_1}||^2 - ||x_{k_3}||^2 \\ d_4^2 - d_1^2 + ||x_{k_1}||^2 - ||x_{k_4}||^2 \end{bmatrix} , \tag{3.9}$$

onde $A \in \mathbb{R}^{3x3}$ e $c \in \mathbb{R}^3$.

Porém, podemos escolher qualquer uma das quatro equações do sistema (3.8) como equação pivô e em cada uma dessas escolhas obtemos um sistema linear diferente com três equações e três incógnitas, $x_{i,1}, x_{i,2}$ e $x_{i,3}$, e portanto, cada um desses sistemas terá $3! = 6$ permutações diferentes. Assim, existem $6 \cdot 4 = 24$ diferentes sistemas lineares que podem ser derivados do sistema não linear (3.8). Entretanto, pelo Teorema 3.1.5, temos que todos esses 24 sistemas têm a propriedade de que a matriz de coeficientes é não-singular, e portanto, os determinantes dessas matrizes serão diferente de zero. Além disso, Dong e Wu afirmam que quanto mais próximo o determinante estiver de zero, mais próxima a matriz A estará de ser uma matriz singular [18],

e por esta razão, no método GB, dos 24 sistemas lineares possíveis, é escolhido aquele que tem o maior valor absoluto do determinante da matriz de coeficientes, para resolvê-lo e assim determinarmos as coordenadas do átomo remanescente. Como veremos adiante, no artigo da atualização do GB, os próprios autores reconhecem que o determinante não é uma boa medida para verificar o condicionamento de uma matriz e começam a utilizar o número de condição para tal verificação.

Porém, não precisamos calcular os determinantes das vinte e quatro matrizes de coeficientes que determinam os sistemas, pois, se uma matriz B é uma permutação de A, então o determinante das duas matrizes são iguais, basta ver que $|det(B)| = |det(P \cdot A)| = |det(P)||det(A)| = 1.|det(A)|$, pois $|det(P)| = 1$. Assim, precisamos calcular o determinante de apenas quatro dos vinte e quatro sistemas lineares possíveis.

O algoritmo do método GB é como se segue:

Algoritmo 2: Geometric Build-Up (GB)

Entrada: Um conjunto D de distâncias euclidianas entre pares de átomos de uma molécula com n átomos.
Saída: Estrutura completa (ou parcial) da molécula de proteína.

1. Encontre quatro átomos não-coplanares com distâncias conhecidas entre si.
2. Determine suas coordenadas x_1, x_2, x_3 e x_4, segundo a demonstração do Teorema 3.1.3.
3. Para cada átomo não-determinado j, se possível, encontre quatro átomos determinados, x_{j1}, x_{j2}, x_{j3} e x_{j4}, não-coplanares, com distâncias conhecidas entre si e com o átomo indeterminado.
4. A partir desses quatro átomos, resolva o sistema $Ax = b$ do Teorema 3.1.5 que tem a matriz de coeficientes com o maior determinante.
5. Se não for possível encontrar 4 átomos base para os átomos remanescentes, pare e reporte a estrutura parcial. Senão, volte ao passo 3.

Observação 1. *Como foi dito acima, na determinação de cada átomo remanescente teremos quatro possíveis sistemas linares que poderemos resolver, pois são derivados do sistema quadrático (3.8). Tais sistemas são $A_k x = c_k$, com $k = 1,2,3,4$ e onde*

$$A_1 = 2\begin{bmatrix} (x_{k_1} - x_{k_2})^T \\ (x_{k_1} - x_{k_3})^T \\ (x_{k_1} - x_{k_4})^T \end{bmatrix} \quad e \quad c_1 = \begin{bmatrix} d_{i,2}^2 - d_{i,1}^2 + ||x_{k_1}||^2 - ||x_{k_2}||^2 \\ d_{i,3}^2 - d_{i,1}^2 + ||x_{k_1}||^2 - ||x_{k_3}||^2 \\ d_{i,4}^2 - d_{i,1}^2 + ||x_{k_1}||^2 - ||x_{k_4}||^2 \end{bmatrix}, \tag{3.10}$$

$$A_2 = 2\begin{bmatrix} (x_{k_2} - x_{k_1})^T \\ (x_{k_2} - x_{k_3})^T \\ (x_{k_2} - x_{k_4})^T \end{bmatrix} \quad e \quad c_2 = \begin{bmatrix} d_{i,2}^2 - d_{i,1}^2 + ||x_{k_2}||^2 - ||x_{k_1}||^2 \\ d_{i,3}^2 - d_{i,1}^2 + ||x_{k_2}||^2 - ||x_{k_3}||^2 \\ d_{i,4}^2 - d_{i,1}^2 + ||x_{k_2}||^2 - ||x_{k_4}||^2 \end{bmatrix}, \tag{3.11}$$

$$A_3 = 2\begin{bmatrix} (x_{k_3} - x_{k_1})^T \\ (x_{k_3} - x_{k_2})^T \\ (x_{k_3} - x_{k_4})^T \end{bmatrix} \quad e \quad c_3 = \begin{bmatrix} d_{i,2}^2 - d_{i,1}^2 + ||x_{k_3}||^2 - ||x_{k_1}||^2 \\ d_{i,3}^2 - d_{i,1}^2 + ||x_{k_3}||^2 - ||x_{k_2}||^2 \\ d_{i,4}^2 - d_{i,1}^2 + ||x_{k_3}||^2 - ||x_{k_4}||^2 \end{bmatrix}, \tag{3.12}$$

$$A_4 = 2\begin{bmatrix} (x_{k_4} - x_{k_1})^T \\ (x_{k_4} - x_{k_2})^T \\ (x_{k_4} - x_{k_3})^T \end{bmatrix} \quad e \quad c_4 = \begin{bmatrix} d_{i,2}^2 - d_{i,1}^2 + ||x_{k_4}||^2 - ||x_{k_1}||^2 \\ d_{i,3}^2 - d_{i,1}^2 + ||x_{k_4}||^2 - ||x_{k_2}||^2 \\ d_{i,4}^2 - d_{i,1}^2 + ||x_{k_4}||^2 - ||x_{k_3}||^2 \end{bmatrix}. \tag{3.13}$$

Destes, escolheremos, para resolver, o sistema i, onde $det(A_i) \geq det(A_k) \forall k = 1,2,3,4$.

Apresentaremos a seguir, uma atualização do algoritmo GB, que mantém as suas ideias centrais, porém em cada etapa, onde são determinadas as coordenadas de um átomo remanescente, o algoritmo atualizado realiza um passo a mais. Ele recalcula a posição dos quatro átomos base em um novo sistema de coordenadas, conforme o Teorema 3.1.3, depois calcula a posição do átomo não determinado conforme o Teorema 3.1.5, fazendo assim que ele se torne o quinto átomo nesta estrutura no novo sistema de coordenadas. Por fim são verificadas as tranformações que são necessárias fazer para que os quatro átomos bases recalculados voltem a sua posição original, e é aplicado no quinto átomo, encontrando enfim, as coordenadas do átomo remanescente no sistema de coordenandas principal onde está sendo construida a estrutura da molécula.

3.2.2 *Updated Geometric Build-up Algorithm* (UGB)

Nesta seção iremos discutir sobre o "Updated Geometric Build-up Algorithm" (UGB) que foi apresentado por D. Wu e Z. Wu em [45]. Este algoritmo é uma versão atualizada do GB para o PGDM com um conjunto arbitrário de distâncias exatas, e a principal diferença é que no algoritmo atualizado foram implementadas duas estratégias para minimizar os erros introduzidos nos cálculos das coordenadas dos átomos. A primeira mudança é que ao invés de calcular o determinante da matriz de coeficientes, será o número de condição que será examinado para verificar qual sistema linear será resolvido em cada passo do algoritmo. O que é melhor do que avaliar o determinante, uma vez que a matriz pode ser mal condicionada mesmo se o seu determinante for grande. A segunda modificação é que as coordenadas dos quatro átomos base são recalculadas (ou reinicializadas). Assim, o acúmulo de erros introduzidos na resolução dos sistema lineares, para a determinação das coordenadas dos átomos, são controlados nas etapas intermediárias. Essa segunda estratégia também foi implementada na atualização do Algoritmo T.

O UGB, se utiliza das mesmas transformações realizadas do Algoritmo GB, descritas no Teorema 3.1.5, porém, no início de cada passo, quando vamos calcular o i-ésimo átomo, nós fazemos uma reinicialização, recalculando as coordenadas, a partir da origem, dos átomos base do átomo a ser determinado, através do Teorema 3.1.3 gerando assim uma nova estrutura com 4 átomos. Depois calculamos a posição do átomo i nessa nova estrutura através da resolução do sistema da Observação 1 com o menor número de condição, depois verificamos as transformações necessárias para que os 4 átomos base voltem à suas posições originais, por fim aplicamos essa transformação no quinto átomo da nova estrutura, e o acrescentamos na estrutura principal.

Os próximos dois resultados mostram que o cálculo do determinante pode não ser tão confiável para verificar se uma matriz está bem ou mal condicionada.

Teorema 3.2.1 (Demmel, 1996)**.** *Seja a matriz A não-singular, então*

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

onde $\mathscr{K}_2(A)$ *é o número de condição na norma-2 da matriz A.*

A demonstração deste teorema se encontra em [16] e mostra que a distância de A para a matriz singular mais próxima dela é de $\frac{1}{\mathscr{K}_2(A)}$. O próximo exemplo, mostra bem que mesmo que o determinante de uma matriz seja grande, ela pode ser mal condicionada.

Observação 2. *Considere a matriz*

$$A = \begin{bmatrix} 10^4 & 0 \\ 0 & 10^{-1} \end{bmatrix}$$

Temos que $det(A) = 10^3$ *e* $\mathscr{K}_2(A) = 10^5$. *Como o número de condição de A é grande, então pelo Teorema anterior, temos que A está bem próxima de uma matriz singular, e portanto é uma matriz mal condicionada. Por outro lado, o determinante de A também é grande, o que mostra que o determinante não é uma boa medida para determinar se uma matriz é bem ou mal condicionada.*

Portanto, o número de condição é uma forma mais confiável que o determinante para verificar o condicionamento de uma matriz. E como temos que $\mathscr{K}_2(AP) = \mathscr{K}_2(PA) = \mathscr{K}_2(A)$ onde P é uma matriz de permutação, então na determinação de cada átomo remanescente, os sistemas lineares da Observação 1 continuam sendo os 4 que precisam ser escolhidos para resolver, porém, escolhendo o sistema que tem a matriz de coeficientes com o menor número de condição.

O algoritmo UGB se inicia da mesma forma que o seu antecessor. Primeiramente buscamos por 4 átomos não-coplanares que sejam vizinhos entre si, e calculamos suas coordenadas segundo as fórmulas do Teorema 3.1.3. Posteriormente para cada átomo remanescente i escolhemos quatro átomos determinados $x_{k_1}, x_{k_2}, x_{k_3}$ e x_{k_4} que possam servir de base para o cálculo deste átomo. Para isso, pelo Teorema 3.1.5, além de precisarem ser vizinhos do átomo não determinado, é necessário que os quatro sejam vizinhos entre si e não coplanares.

Após encontrar 4 átomos que satisfaçam essas restrições, nós recalculamos a posição destes em um outro sistema de coordenadas, segundo as fórmulas usadas na demonstração do Teorema 3.1.3, obtendo os pontos $y_{k_1}, y_{k_2}, y_{k_3}$ e y_{k_4} neste novo sistema. Posteriormente calculamos o número de condição das matrizes A_k, descritas na Observação 1, e resolvemos o sistema linear que possuir a matriz de coeficientes com o menor número de condição. A solução y_i de tal sistema será as coordenadas do átomo i no novo sistema de coordenadas, e para acrescentarmos o átomo i na estrutura principal, aplicamos as transformações necessárias em y_i para que os 4 átomos reinicializados voltem às suas posições originais. Esse passo adicional foi desenvolvido a fim de obter uma maior estabilidade nos sistemas lineares que são resolvidos ao longo do algoritmo.

Vejamos então quais são estas transformações citadas para que os átomos voltem a estrutura principal. Primeiro é realizada uma translação para que a nova estrutura tenha o mesmo centro geométrico que os quatro átomos base na estrutura principal e depois é aplicado uma rotação para que os 4 átomos voltem às suas posições originais. O cálculo da matriz de rotação é feito através do uso do *Root-Mean-Square Deviation* (RMSD), que é uma estimativa de erro entre duas estruturas, no caso, a matriz de rotação será a matriz Q que minimiza a função do RMSD.

Seja $X \in \mathbb{R}^{4,3}$ a matriz que contém as coordenadas dos átomos base, e $Y \in \mathbb{R}^{4,3}$ a matriz contendo as coordenadas dos átomos base reinicializados, ou seja, temos que

$$\begin{aligned} X(i,j) &= x_{k_i}(j) \\ Y(i,j) &= y_{k_i}(j), \end{aligned}$$

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

Os centros geométricos das estruturas X e Y, denotados por x_c e y_c são definidos por

$$x_c(j) = \tfrac{1}{4}\sum_{i=1}^{4} X(i,j) \quad \text{e} \quad y_c(j) = \tfrac{1}{4}\sum_{i=1}^{4} Y(i,j), \tag{3.14}$$

para $j = 1,2,3$. Logo, se aplicarmos a seguinte translação em Y:

$$\begin{aligned} Y^{(1)}(i,1) &= Y(i,1) - [y_c(1) - x_c(1)] \\ Y^{(1)}(i,2) &= Y(i,2) - [y_c(2) - x_c(2)], \\ Y^{(1)}(i,3) &= Y(i,3) - [y_c(3) - x_c(3)] \end{aligned} \tag{3.15}$$

com $i = 1,2,3,4$, então teremos que a matriz Y terá o mesmo centro geométrico de X, basta observar o seguinte Teorema:

Teorema 3.2.2. *As estrutura X e $Y^{(1)}$ como definidas acima têm o mesmo centro geométrico.*

Demonstração. Iremos calcular o centro geométrico de $Y^{(1)}$ e mostra que é igual ao de X. Temos inicialmente pelo sistema 3.15 que

$$Y^{(1)}(i,j) = Y(i,j) - y_c(j) + x_c(j)$$

Então, o centro geométrico de $Y^{(1)}$ é o seguinte:

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

para $j = 1,2,3$. E, como o centro geométrico de Y é $y_c(j) = \frac{1}{4}\sum_{i=1}^{4} Y(i,j)$ com $j = 1,2,3$, então

$$y_c^{(j)} = y_c(j) - \frac{1}{4}(4y_c(j)) + \frac{1}{4}(4x_c(j)) = y_c(j) - y_c(j) + x_c(j) = x_c(j)$$

Portanto, o centro geométrico da estrutura $Y^{(1)}$ é igual ao centro geométrico de X. □

Como as estruturas X e $Y^{(1)}$ têm o mesmo centro geométrico, então podemos utilizar o RMSD para verificar qual a matriz de rotação Q que minimiza a diferença entre ambas estruturas, e assim aplicando essa rotação Q em $Y^{(1)}$ teremos que os átomos da estrutura Y voltarão às suas posições originais na estrutura X, isto é, na estrutura principal. A definição do RMSD é dada a seguir:

Definição 6 (RMSD)**.** *[3, 19, 29, 30, 41] Sejam X e Y duas matrizes em $\mathbb{R}^{n\times 3}$ representando duas estruturas sobrepostas, isto é, com o mesmo centro geométrico. Definimos o RMSD entre X e Y por*

$$RMSD(X,Y) = min_Q \frac{\|X - YQ\|_F}{\sqrt{n}}, \tag{3.16}$$

onde Q é uma matriz 3×3 de rotação e $\|.\|_F$ é a norma de Frobënius.

No próximo teorema, verificamos qual é a matriz de rotação que minimiza a função do RMSD.

Teorema 3.2.3. *[Wu, 2008] Seja $X, Y \in \mathbb{R}^{nx3}$ duas matrizes com o mesmo centro geométrico, temos que a matriz de rotação que minimiza a função do RMSD é dada por*

$$Q = UV^T, \tag{3.17}$$

onde $C = U\Sigma V^T$ é a decomposição em valores singulares da matriz $C = Y^T X$.

Demonstração. Seja X e Y duas matrizes de ordem $nx3$ que representam duas estruturas sobrepostas. Queremos obter a matriz de rotação Q que minimiza a função do RMSD. Porém, a matriz que minimiza $\frac{\|X-YQ\|_F}{\sqrt{n}}$ será a mesma matriz que minimiza $\|X - YQ\|_F$. Assim, para encontrar Q basta resolvermos o problema

$$min_Q \|X - YQ\|_F. \tag{3.18}$$

Utilizando as propriedades da norma de Frobenius e da função traço, podemos fazer os seguintes cálculos:

$\|X-YQ\|_F^2 = \text{tr}((X-YQ)^T(X-YQ)) = \text{tr}(X^TX - X^TYQ - Q^TY^TX + Q^TY^TYQ) = \text{tr}(X^TX) + \text{tr}(Q^TY^TYQ) - 2\text{tr}(Q^TY^TX) = \text{tr}(X^TX) + \text{tr}(Y^TY) - 2\text{tr}(Q^TY^TX)$, pois $\text{tr}(X^TYQ) = \text{tr}(X^TYQ)^t = \text{tr}(Q^TY^TX)$ e $\text{tr}(Q^TY^TYQ) = \text{tr}(QQ^TY^TY) = \text{tr}(Y^TY)$.

Assim, chegamos que $\|X - YQ\|_F^2 = \text{tr}(X^TX) + \text{tr}(Y^TY) - 2\text{tr}(Q^TY^TX)$, e portanto, o problema de minimizar $\|X - YQ\|_F$ é equivalente ao problema de maximizar $\text{tr}(Q^TY^TX)$.

Calculando a fatoração SVD da matriz $C = Y^TX$, obtemos $C = U\Sigma V^T$, onde U e V são matrizes ortogonais e Σ é uma matriz diagonal, portanto

$$\text{tr}(Q^TY^TX) = \text{tr}(Q^TU\Sigma V^T) = \text{tr}(V^TQ^TU\Sigma).$$

E como U, V e Q são matrizes ortogonais, segue que $\text{tr}(V^TQ^TU\Sigma) \leq \text{tr}(\Sigma)$. Porém, se fizermos $Q = UV^T$, ou seja, $Q^T = VU^T$, chegamos que $\text{tr}(V^TQ^TU\Sigma) = \text{tr}(V^T(VU^T)U\Sigma) = \text{tr}(\Sigma)$, e portanto, $Q = UV^T$ é a matriz ortogonal de rotação que maximiza $\text{tr}(Q^TY^TX)$ e consequentemente que minimiza $\frac{\|X-YQ\|_F}{\sqrt{n}}$.

□

Portanto, depois de calcularmos a posição y_i do átomo i, então basta aplicar a translação 3.15 seguida da rotação 3.17, mostrada no teorema anterior, em y_i que encontraremos as coordenadas x_i do átomo i na estrutura principal. E, assim, repetimos o mesmo procedimento para o próximo átomo remanescente. Dessa forma, o algoritmo do UGB é como se segue:

Algoritmo 3: Updated Geometric Build-Up (UGB)

Entrada: Um conjunto D de distâncias euclidianas entre pares de átomos de uma molécula com n átomos.
Saída: Estrutura completa (ou parcial) da molécula de proteína.

1. Encontre quatro átomos não-coplanares com distâncias conhecidas entre si.
2. Determine suas coordenadas x_1, x_2, x_3 e x_4, segundo a demonstração do Teorema 3.1.3.
3. Para cada átomo não-determinado j, se possível, encontre quatro átomos determinados, x_{j1}, x_{j2}, x_{j3} e x_{j4}, não-coplanares, com distâncias conhecidas entre si e com o átomo indeterminado.
4. Utilizando apenas as distâncias disponíveis, reinicialize os quatro átomos, ou seja, encontre as coordenadas y_{j1}, y_{j2}, y_{j3} e y_{j4}, segundo a demonstração do Teorema 3.1.3.
5. Determine a translação 3.15 e a rotação 3.17.
6. A partir desses quatro átomos (com suas novas coordenadas), calcule as coordenadas y_j do átomo j na nova estrutura, resolvendo o sistema $Ax = b$ do Teorema 3.1.5 que tem a matriz de coeficientes com o menor número de condição.
7. Aplique em y_j a translação e a rotação encontrada no passo 5, calculando dessa forma as coordenadas x_j do átomo j na estrutura principal.
8. Se não for possível encontrar 4 átomos base para os átomos remanescentes, pare e reporte a estrutura parcial. Senão, volte ao passo 3.

Capítulo 4

Descrição da Nova Família de Métodos para o Problema de Geometria de Distâncias Moleculares

Neste capítulo iremos apresentar a nossa proposta de algoritmo que nomeamos de Algoritmo T Atualizado (ATA), que consiste em utilizar no Algoritmo T (AT) a mesma atualização que foi feita no UGB, em relação ao GB, a fim de tentar evitar erros numéricos, como por exemplo, instabilidade, arredondamento, etc. Apesar do algoritmo T ter sido apresentado em [21] por F. Fidalgo, o colocamos separado dos outros métodos da revisão da literatura, ficando no mesmo capítulo da descrição do novo método que estamos propondo, pois além de que a idéia do ATA ter dependido do AT, ainda quisemos separar o capítulo 3 para os algoritmos da "família Geometric Build-Up" e o capítulo 4 para os da "família T", do qual ambos métodos fazem parte.

Os dois primeiros passos do AT e ATA são iguais, inicialmente escolhemos quatro átomos não coplanares e vizinhos entre si, determinamos suas coordenadas através do Teorema 3.1.3, buscamos um átomo que seja vizinho destes quatro primeiros átomos e determinamos suas coordenadas através da resolução de um sistema linear. A partir desse momento os dois métodos começam a se diferenciar. Para cada átomo remanescente, o AT busca quatro átomos determinados que possam servir de base para esse átomo requerido, e assim, através da resolução de um sistema linear pode obter as coordenadas deste átomo até então não determinado, enquanto que o ATA faz uma reinicialização das coordenadas dos átomos base antes de resolver o sistema.

4.1 Algoritmo T (AT)

O Algoritmo T foi apresentado por F. Fidalgo em [21], e é bem parecido com o algoritmo GB, diferindo na transformação aplicada no sistema não-linear (3.8). O primeiro passo do AT é determinar as coordenadas de quatro átomos não coplanares e vizinhos entre si através das fórmulas na demonstração do Teorema 3.1.3, depois para cada átomo não determinado i, buscar quatro átomos que possam servir de base para o mesmo, e resolver um sistema linear, calculando dessa forma a posição deste átomo.

Os próximos resultados são os fundamentos do AT e mostram entre outras coisas a transformação que será aplicada no do sistema (3.8).

Lema 4.1.1. *[Fidalgo, 2011] Se* $\{x_1, x_2, x_3, x_4\} \subset \mathbb{R}^3$ *é um conjunto de pontos não-coplanares, então a matriz*

$$B = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \end{bmatrix}^T$$

é não-singular, onde $v_i = \begin{bmatrix} 1 & x_i^T \end{bmatrix}^T$, $i = 1, \ldots, 4$.

Demonstração. Verificaremos para quais $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ temos que

$$\alpha v_1 + \beta v_2 + \gamma v_3 + \delta v_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T, \tag{4.1}$$

ou seja,

$$\alpha \begin{bmatrix} 1 & x_1^T \end{bmatrix}^T + \beta \begin{bmatrix} 1 & x_2^T \end{bmatrix}^T + \gamma \begin{bmatrix} 1 & x_3^T \end{bmatrix}^T + \delta \begin{bmatrix} 1 & x_4^T \end{bmatrix}^T = 0.$$

E, portanto, teremos o seguinte sistema linear

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

que é equivalente ao sistema

$$\begin{cases} \alpha + \beta + \gamma = -\delta \\ \alpha x_1 + \beta x_2 + \gamma x_3 = -\delta x_4, \end{cases} \tag{4.2}$$

Logo, podemos observar que

$$\alpha x_1 + \beta x_2 + \gamma x_3 = -\delta x_4 = (\alpha + \beta + \gamma) x_4 = \alpha x_4 + \beta x_4 + \gamma x_4. \tag{4.3}$$

Da equação (4.3), temos

$$0 = (\alpha x_1 - \alpha x_4) + (\beta x_2 - \beta x_4) + (\gamma x_3 - \gamma x_4) = \alpha(x_1 - x_4) + \beta(x_2 - x_4) + \gamma(x_3 - x_4)$$

Como x_1, x_2, x_3 e x_4 são não coplanares, então pelo Lema 3.1.4 temos que $(x_1 - x_4), (x_2 - x_4)$ e $(x_3 - x_4)$ são vetores linearmente independentes, e portanto, temos que $\alpha = \beta = \gamma = 0$. Substituindo estes valores no sistema 4.2 temos que $\delta = 0$, então o sistema 4.1 tem apenas uma solução: $\alpha = \beta = \gamma = \delta = 0$.

Dessa forma, podemos afirmar que a matriz

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

é não-singular.

□

A seguir, apresentamos o resultado central do algoritmo T, que faz uso do lema anterior.

Teorema 4.1.2. *[Fidalgo, 2011] Sejam $\{b_1, b_2, b_3, b_4\}$ e $\{y_1, y_2, y_3, y_4\}$, respectivamente, subconjuntos de $\mathbb{R}^3$ e $\mathbb{R}$. Se o sistema quadrático*

$$\begin{aligned} \|a - b_1\|_2 &= y_1 \\ \|a - b_2\|_2 &= y_2 \\ \|a - b_3\|_2 &= y_3 \\ \|a - b_4\|_2 &= y_4 \end{aligned} \tag{4.4}$$

possui uma solução a^*, então o sistema $Ax = b$,

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

possui uma única solução x^*, em função de a^*, da forma $x^* = \begin{bmatrix} t & a^{*T} \end{bmatrix}^T$, onde $t = -\dfrac{\|a^*\|^2}{2}$.

Demonstração. Seja $a^* \in \mathbb{R}^3$ uma solução para o sistema 4.4. Assim, substituindo a^* neste sistema, elevando suas equações ao quadrado e fazendo operações com elas, temos

$$\begin{array}{l} \|a^*\|^2 - 2b_1^T a^* = y_1^2 - \|b_1\|^2 \\ \|a^*\|^2 - 2b_2^T a^* = y_2^2 - \|b_2\|^2 \\ \|a^*\|^2 - 2b_3^T a^* = y_3^2 - \|b_3\|^2 \\ \|a^*\|^2 - 2b_4^T a^* = y_4^2 - \|b_4\|^2 \end{array},$$

ou seja,

$$\begin{array}{l} -2t - 2b_1^T a^* = y_1^2 - \|b_1\|^2 \\ -2t - 2b_2^T a^* = y_2^2 - \|b_2\|^2 \\ -2t - 2b_3^T a^* = y_3^2 - \|b_3\|^2 \\ -2t - 2b_4^T a^* = y_4^2 - \|b_4\|^2 \end{array}, \tag{4.5}$$

tomando $t = -\|a^*\|^2/2$. Matricialmente, podemos escrever (4.5) como

$$-2 \begin{bmatrix} 1 & b_1^T \\ 1 & b_2^T \\ 1 & b_3^T \\ 1 & b_4^T \end{bmatrix} \begin{bmatrix} t \\ a^* \end{bmatrix} = \begin{bmatrix} y_1^2 - \|b_1\|^2 \\ y_2^2 - \|b_2\|^2 \\ y_3^2 - \|b_3\|^2 \\ y_4^2 - \|b_4\|^2 \end{bmatrix}. \tag{4.6}$$

Portanto, $x^* = \begin{bmatrix} t & a^{*T} \end{bmatrix}^T$ é solução para $Ax = b$ em função de a^*. □

Em cada passo do AT, deseja-se determinar a posição x_j de um átomo indeterminado da molécula. Sejam x_{k_1}, x_{k_2}, x_{k_3} e x_{k_4} as posições de quatro átomos determinados, não-coplanares, com distâncias entre si e com x_j todas conhecidas. A partir das distâncias $d_{j,k_1}, d_{j,k_2}, d_{j,k_3}$ e d_{j,k_4}, entre x_j e $x_{k_1}, x_{k_2}, x_{k_3}$ e x_{k_4}, respectivamente, podemos considerar o seguinte sistema quadrático:

$$\begin{array}{l} \left\|x_j - x_{k_1}\right\|_2 = d_{j,1} \\ \left\|x_j - x_{k_2}\right\|_2 = d_{j,2} \\ \left\|x_j - x_{k_3}\right\|_2 = d_{j,3} \\ \left\|x_j - x_{k_4}\right\|_2 = d_{j,4} \end{array}. \tag{4.7}$$

Apresentaremos, a seguir, o principal resultado que embasa este método, fazendo uso do Teorema 4.1.2 e do Lema 4.1.1.

Corolário 4.1.3. *[Fidalgo, 2011] Sejam $x_{k_1}, x_{k_2}, x_{k_3}$ e x_{k_4} as coordenadas de quatro átomos determinados, não-coplanares, com distâncias com um átomo indeterminado x_j conhecidas. Se conhecermos as distância $d_{j,k_1}, d_{j,k_2}, d_{j,k_3}, d_{j,k_4}$ e se $x^* = \begin{bmatrix} t_j & {x_j}^{*T} \end{bmatrix}^T$, onde $t_j = -\left\|{x_j}^*\right\|^2/2$, é solução de*

$$Ax = b, \text{ com} \tag{4.8}$$

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

onde $A \in \mathbb{R}^{4x4}$ e $b \in \mathbb{R}^{4x3}$, então temos que as coordenadas do átomo j serão $x_j = x_j^*$.

Demonstração. Como é de nosso conhecimento as coordenadas de quatro átomos determinados e as distâncias entre estes e um outro átomo não determinado, então para calcularmos a posição deste átomo basta resolver o seguinte sistema quadrático

$$\begin{aligned} \left\|x_j - x_{k_1}\right\|_2 &= d_{j,k_1} \\ \left\|x_j - x_{k_2}\right\|_2 &= d_{j,k_2} \\ \left\|x_j - x_{k_3}\right\|_2 &= d_{j,k_3} \\ \left\|x_j - x_{k_4}\right\|_2 &= d_{j,k_4} \end{aligned} . \tag{4.9}$$

E como podemos ver no apêndice A, no teorema A.2 este sistema tem no máximo 1 solução, logo a menos de erros nas medidas, temos que tal sistema tem exatamente 1 solução, que será as coordenadas do átomo não determinado, então os átomos $x_{k_1}, x_{k_2}, x_{k_3}$ e x_{k_4} podem servir como base para o átomo x_j.

Pelo Teorema 4.1.2, como a solução do sistema 4.9 é x_j, então a solução do sistema linear $Ax = b$ será $x^* = \begin{bmatrix} t_j & x_j{}^T \end{bmatrix}^T$, onde $t_j = -\left\|x_j\right\|^2/2$.

Mas como A é não-singular, pelo Lema 4.1.1, então temos que $Ax = b$ também possui solução única. E portanto, basta resolver tal sistema linear para encontrar as coordenadas do átomo j.

□

Assim, podemos dizer que o Algoritmo T funciona da seguinte forma: Considere uma molécula com n átomos da qual conhecemos um conjunto arbitrário de distâncias entre pares de seus átomos. Denotaremos as coordenadas desses átomos, que desejamos calcular, por $x_1, x_2, \ldots, x_n \in \mathbb{R}^3$.

Assim, como nos métodos anteriores, calculamos a posição dos quatro primeiros átomos através das fórmulas descritas na demonstração do Teorema 3.1.3. Posteriormente, para cada átomo j remanescente, buscamos por quatro átomos determinados que sejam não coplanares, e vizinhos do átomo não determinado. Segundo o Corolário 4.1.3, estes quatro átomos servirão como base para o cálculo do átomo remanescente, bastando calcular a solução $x^* = (t_j, u_j, v_j, w_j)$ do sistema linear (4.8), que teremos que as coordenadas deste átomo requerido será $x_j = (u_j, v_j, w_j)$. Além disso, a primeira coordenada da solução de $Ax = b$ talvez possa ser utilizada como um teste para a qualidade da solução encontrada, já que a mesma tem de estar próxima de $t_j = -\left\|x_j{}^*\right\|^2/2$. Essa questão ainda é alvo de investigação.

Quanto à estrutura de saída deste algoritmo, há que se fazer uma observação.

Observação 3. *Há outros métodos que mostram que é possível encontrar mais de uma solução para PGDM [34], ou seja, determinar mais de uma configuração tridimensional para seus átomos, respeitando as restrições de distâncias. Estas soluções podem ser semelhantes (i.e., podem ser sobrepostas a partir da aplicação de movimentos rígidos, como translação e/ou rotação) ou não. Tanto o ATA quanto o UGB fornecem, apenas, uma delas.*

Assim como acontece com o algoritmo GB e UGB, caso não seja possível determinar quatro átomos base para algum dos átomos remanescentes ou resolver um dos sistema linear no decorrer do método, o AT para e reporta uma estrutura molecular parcial.

Dessa forma, apresentamos o método proposto por F. Fidalgo em [21], chamado Algoritmo T, e que continua sendo alvo de nossos estudos. Utilizando os resultados demonstrados acima, segue o algoritmo do AT.

Algoritmo 3: Algoritmo T (AT)

Entrada: Um conjunto D de distâncias euclidianas entre pares de átomos de uma molécula com n átomos.
Saída: Estrutura completa (ou parcial) da molécula de proteína.

1. Encontre quatro átomos não-coplanares com distâncias conhecidas entre si.

2. Determine suas coordenadas x_1, x_2, x_3 e x_4, segundo a demonstração do Teorema 3.1.3.

3. Para cada átomo não-determinado j, se possível, encontre quatro átomos determinados, x_{j1}, x_{j2}, x_{j3} e x_{j4}, não-coplanares, com distâncias conhecidas entre si e com o átomo indeterminado.

4. A partir desses quatro átomos, resolva o sistema $Ax = b$ do Corolário 4.1.3.

5. Faça $x_j(i) = x(i+1)$, para $i = 1,2,3$. x_j é a posição determinada para o átomo j.

6. Se não for possível encontrar 4 átomos base para os átomos remanescentes, pare e reporte a estrutura parcial. Senão, volte ao passo 3.

4.2 Algoritmo T Atualizado (ATA)

Verificando que a reinicialização feita em cada passo no Algoritmo UGB foi bem sucedida e que diminuía em muito o erro entre a estrutura da molécula original e a encontrada no pelo método, decidimos então, implementar essa reinicialização no AT, que se mostrou tão eficiente, em relação ao erro calculado, quanto o UGB, porém, com um menor custo computacional. Essa modificação no AT originou o Algoritmo T Atualizado (ATA), que já foi apresentado no XVI *Congresso Latino-Iberoamericano de Investigación Operativa* (CLAIO) [22], sendo incluído como trabalho completo em seus pré-anais, e que também é tema do artigo [23] em construção.

Nesta atualização, em que pretendemos evitar a propagação e o acúmulo de erros numéricos e garantir mais estabilidade, há uma mudança do sistema de coordenadas em cada passo k: passamos os quatro átomos base $x_{k_1}, x_{k_2}, x_{k_3}$ e x_{k_4} do sistema original para outro, de modo que as posições destes quatro átomos no novo sistema $y_{k_1}, y_{k_2}, y_{k_3} y$ e y_{k_4} dependam, apenas, das distâncias entre eles para serem determinadas e independam de cálculos previamente realizados. Depois, calcula-se a posição do átomo indeterminado no novo sistema através do sistema linear (4.8) do Teorema 4.1.3 e, em seguida, colocamos este novo átomo no sistema de coordenadas original, observando as transformações necessárias para que os 4 átomos base voltem a sua posição original. O ATA se mostrou mais estável do que o AT, garantindo maior qualidade das soluções, quanto a precisão numérica.

Iremos detalhar melhor a sequência de passos do Algotitmo T Atualizado. Primeiramente, buscamos 4 átomos que sejam não coplanares e vizinhos entre si, e determinamos suas coordenadas através do Teorema 3.1.3, ou seja, conforme as seguintes fórmulas:

$$x_1 = (0,0,0),\ \ x_2 = (x_{2,1},0,0),\ \ x_3 = (x_{3,1},x_{3,2},0)\ \ x_4 = (x_{4,1},x_{4,2},x_{4,3})$$

onde,

$$
\begin{aligned}
x_{2,1} &= \pm d_{2,1} \\
x_{3,1} &= \frac{d_{3,1}^2 - d_{3,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2} \\
x_{3,2} &= \pm (d_{3,1}^2 - x_{3,1}^2)^{1/2} \\
x_{4,1} &= \frac{d_{4,1}^2 - d_{4,2}^2}{2x_{2,1}} + \frac{x_{2,1}}{2} \\
x_{4,2} &= \frac{d_{4,2}^2 - d_{4,3}^2 - (x_{4,1} - x_{2,1})^2 + (x_{4,1} - x_{3,1})^2}{2x_{3,2}} + \frac{x_{3,2}}{2} \\
x_{4,3} &= \pm (d_{4,1}^2 - x_{4,1}^2 - x_{4,2}^2)^{1/2}
\end{aligned}
\quad .
$$

Então procedemos igual ao AT para obter o quinto átomo determinado. Depois para cada átomo x_j não determinado, buscamos quatro átomos determinados, não-coplanares, com distâncias conhecidas entre si e com x_j, onde suas coordenadas x_{k_1}, x_{k_2}, x_{k_3} e $x_{k_4} \in \mathbb{R}^3$, e que armazenaremos como linhas de uma matriz $X \in \mathbb{R}^{4\times 3}$. Recalculamos as suas coordenadas conforme as fórmulas descritas acima, obtendo suas novas posições y_{k_1}, y_{k_2}, y_{k_3} e y_{k_4}, as quais também são armazenadas como linhas de uma matriz $Y \in \mathbb{R}^{4\times 3}$. Este processo é chamado reinicialização dos quatro pontos. Observe que os centros geométricos das estruturas tridimensionais representadas pelas matrizes X e Y são calculados, respectivamente, da forma

$$
x_c(k) = \tfrac{1}{4}\sum_{i=1}^{4} X(i,k) \quad \text{e} \quad y_c(k) = \tfrac{1}{4}\sum_{i=1}^{4} Y(i,k), \tag{4.10}
$$

para $k = 1,2,3$. Logo, para que a matriz Y tenha o mesmo centro geométrico de X, realizamos a seguinte translação em Y:

$$
\begin{aligned}
Y(i,1) &= Y(i,1) - [y_c(1) - x_c(1)] \\
Y(i,2) &= Y(i,2) - [y_c(2) - x_c(2)] \\
Y(i,3) &= Y(i,3) - [y_c(3) - x_c(3)]
\end{aligned}
\quad , \tag{4.11}
$$

para $i = 1,2,3,4$. Desse modo as estruturas representadas por X e Y ficaram sobrepostas e teram o mesmo centro geométrico [21]. Para finalizar, resta-nos rotacionar Y para que os quatro átomos, recalculados, voltem o mais próximo de suas posições originais. Para tanto, usamos a RMSD

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

onde X e Y são como definimos acima, a matriz Q é a matriz de rotação que melhor alinha X e Y, estando sobrepostos sobre o mesmo centro geométrico, e $\|\cdot\|_F$ é a norma de Frobenius. Como vimos na seção 3.2.2, neste procedimento, a matriz Q que minimiza a função RMSD é dada por $Q = UV^T$, onde U e V são as componentes ortogonais da decomposição em valores singulares da matriz $C = Y^T X$, isto é, $C = U\Sigma V^T$ [21]. Porém, antes de aplicarmos tais transformações na matriz Y, usamos os átomos y_{k_1}, y_{k_2}, y_{k_3} e y_{k_4} para calcular, pelo Teorema 4.1.3, a posição do átomo não determinado j no novo sistema de coordenadas resolvendo o sistema linear $Ay = b$ com

$$
A = -2\begin{bmatrix} 1 & y_{k_1}^T \\ 1 & y_{k_2}^T \\ 1 & y_{k_3}^T \\ 1 & y_{k_4}^T \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} d_{k_1,1}^2 - \|y_{k_1}\|^2 \\ d_{k_1,2}^2 - \|y_{k_2}\|^2 \\ d_{k_1,3}^2 - \|y_{k_3}\|^2 \\ d_{k_1,4}^2 - \|y_{k_4}\|^2 \end{bmatrix}.
$$

Sendo $y^* = (y^*(1), y^*(2), y^*(3), y^*(4))$ a solução do sistema acima, teremos um quinto átomo nessa nova estrutura, tal que suas coordenadas serão $y_j = (y^*(2), y^*(3), y^*(4))$. Após este cálculo, acrescentamos o átomo j no sistema de coordenadas original, aplicando a translação (4.11) e a rotação

$$Q = UV^T \tag{4.12}$$

em y_j, obtendo a posição x_j do átomo indeterminado. Este procedimento é repetido até que encontremos a estrutura completa da molécula de proteína, reportando toda sua estrutura, ou até que não seja possível determinar as posições dos átomos restantes da molécula, podendo ser pelo mal condicionamento do sistema linear que gera uma instabilidade numérica ou pela impossibilidade de se encontrar os quatro átomos base para algum átomo remanescente. Desse modo, o método se encerra, reportando uma estrutura parcial da molécula.

A figura 4.2 mostra um exemplo do uso dessa estratégia implementada no método ATA a fim de obter uma melhor estabilidade. Nesse exemplo, já foram determinados 6 átomos, cujas coordenadas são x_1, x_2, x_3, x_4, x_5 e x_6. Além disso, busca-se calcular as coordenadas do átomo 7 ainda indeterminado. Os átomos que estão ligados por semi-retas são aqueles para os quais conhecemos as distâncias entre eles, então, pela figura 4.1(a) é possível observar que os átomos $2, 3, 4$ e 5 tem todas distâncias entre eles conhecidas, assim como também as distâncias entre eles e o átomo remanescente. Supomos neste caso, que estes átomos sejam não coplanares, portanto, podem servir de átomos base para a determinação do átomo 7. Assim, no item (a) é realizado um novo calculo das coordenadas destes 4 átomos através do Teorema 3.1.3, levando-os assim a um novo sistema de coordenadas, e obtendo dessa forma as coordenadas y_1, y_2, y_3 e y_4. Na figura (b) é calculada a posição y_7 do átomo 7 neste novo sistema de coordenadas a partir da resolução do sistema $Ax = b$ do Corolário 4.1.3. Por fim, identificamos a translação (4.11) e a rotação (4.12) que levariam as novas posições dos átomos base nas suas posições originais e na figura 4.1(c) aplicamos estas duas transformações lineares em y_7, encontrando dessa forma as coordenadas x_7 do átomo 7 na estrutura que o algoritmo está construindo.

Segue o algoritmo do ATA.

Algoritmo 3: Algoritmo T Atualizado (ATA)

Entrada: Um conjunto D de distâncias euclidianas entre pares de átomos de uma molécula com n átomos.

Saída: Estrutura completa (ou parcial) da molécula de proteína.

1. Encontre quatro átomos não-coplanares com distâncias conhecidas entre si.

2. Determine suas coordenadas x_1, x_2, x_3 e x_4, segundo a demonstração do Teorema 3.1.3.

3. Para cada átomo não-determinado j, se possível, encontre quatro átomos determinados, x_{j1}, x_{j2}, x_{j3} e x_{j4}, não-coplanares, com distâncias conhecidas entre si e com o átomo indeterminado.

4. Utilizando apenas as distâncias disponíveis, reinicialize os quatro átomos, ou seja, encontre as coordenadas y_{j1}, y_{j2}, y_{j3} e y_{j4}, segundo a demonstração do Teorema 3.1.3.

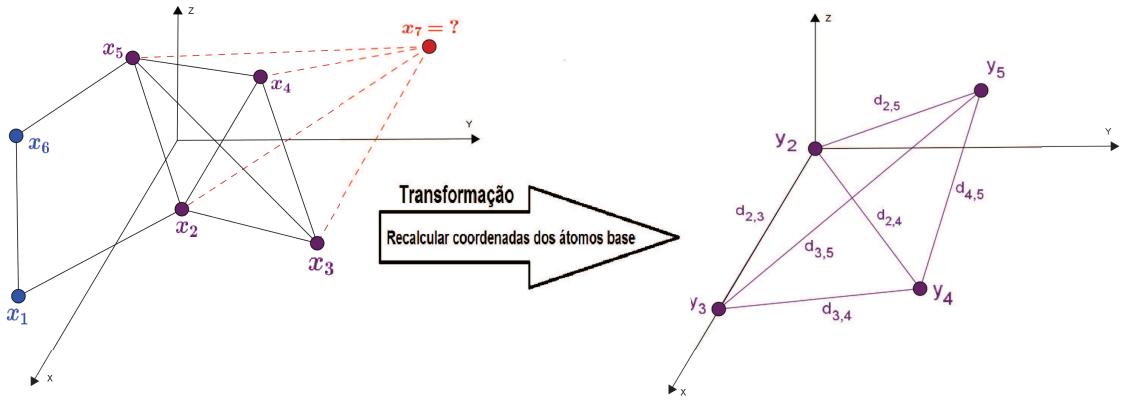
5. Determine a translação 4.11 e a rotação 4.12.

6. Resolva o sistema $Ax = b$, do Corolário 4.1.3, utilizando as novas coordenadas dos átomos base y_{j1}, y_{j2}, y_{j3} e y_{j4}, obtendo y_j^*.

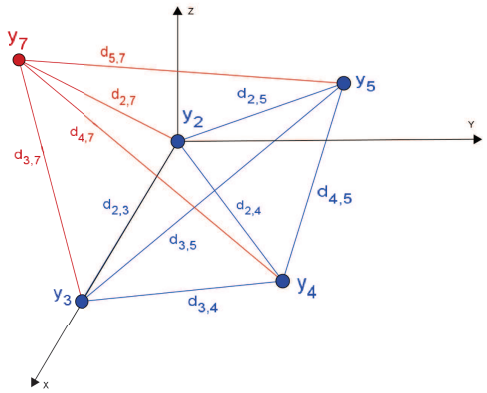
7. Calcule a posição na nova estrutura do átomo indeterminado $y_j(i) = y_j^*(i+1)$, para $i = 1, 2, 3$.

8. Aplique em y_j a translação e a rotação determinada no passo 5, encontrando as coordenadas, x_j, do átomo j na estrutura principal.

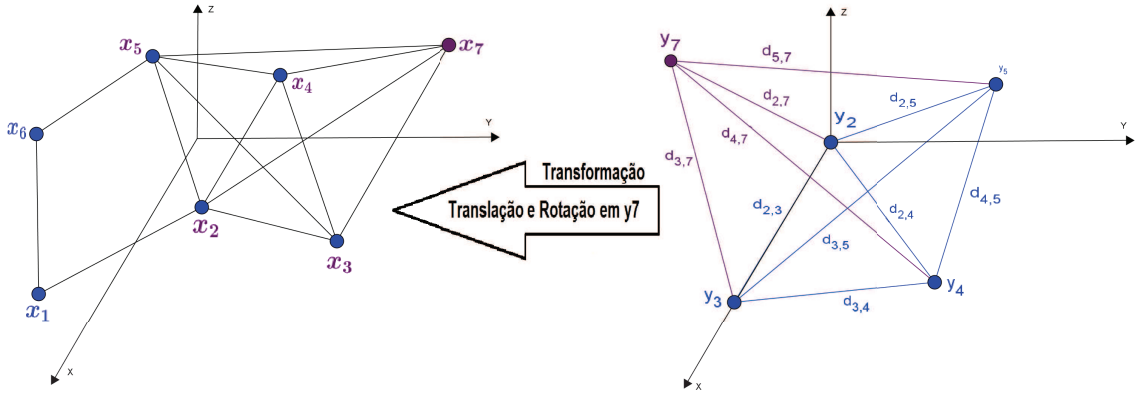
9. Se não for possível encontrar 4 átomos base para os átomos remanescentes, pare e reporte a estrutura parcial. Senão, volte ao passo 3.



(a) Primeiro Passo: recalcular coordenadas dos átomos base através do Teorema 3.1.3.



(b) Segundo Passo: Calcular a posição do átomo indeterminado no novo sistema de coordenadas, através do Corolário 4.1.3.



(c) Terceiro Passo: Aplicar a translação (4.11) e a rotação (4.12) neste átomo para conhecer suas coordenadas na estrutura principal.

Figura 4.1: Estratégia utilizada para melhorar estabilidade dos Sistemas Lineares no algoritmo ATA

Essa reinicialização feita nos átomos base é realizada com o objetivo de melhorar o condicionamento das matrizes de coeficientes e com isso obter mais estabilidade nos sistema lineares que serão resolvidos no decorrer do método ATA. Na seção 5.1.1 são apresentados os testes que foram realizados para verificar se o condicionamento das matrizes realmente eram melhorados depois da utilização dessa estratégia desenvolvida por Wu [45]. Nestes testes o algoritmo ATA mostrou que de fato é mais estável do que o algoritmo T sem essa atualização, e que por causa

dessa estabilidade, o erro, calculado via RMSD, também se comporta melhor no Algoritmo T Atualizado.

Poderíamos inclusive optar por outras estratégias de pré-condicionamento que melhorassem o condicionamento das matrizes de coeficientes, e assim, conseguir uma melhor estabilidade quanto a estes sistema lineares, porém estratégias tais como fazer um reescalonamento utilizando autovalores ou aplicar uma transformação linear para que a base ficasse ortonormal poderiam aumentar o custo computacional, como por exemplo, o tempo de processamento do algoritmo, e um aumento no tempo do algoritmo poderia se tornar uma grande desvantagem em um estudo do tipo Monte Carlo no caso de distâncias imprecisas, que é o caso mais frequente no estudo do PGDM com proteínas. Essa preocupação se explica pela **Lei dos Grandes Números**, também chamada de Primeiro Teorema Fundamental da Probabilidade, que é definido da seguinte forma: "Se um evento de probabilidade p é observado repetidamente em ocasiões independentes, a proporção da freqüência observada deste evento em relação ao total de número de repetições converge em direção a p à medida que o número de repetições se torna arbitrariamente grande."[8]. E pelo **Teorema Central do Limite**, também denominado de Segundo Teorema Fundamental da Probabilidade, um importante resultado estatístico em aplicações práticas, que garante que mesmo que os dados não sejam distribuídos conforme uma distribuição normal, a média dos dados converge para a distribuição normal conforme o número de dados aumenta [37, 13]. Portanto, no estudo de caso de distâncias imprecisas, precisaremos fazer um estudo de quantificação de incertezas e poderemos necessitar destes teoremas citados anteriormente, e assim, o algoritmo escolhido precisará ser utilizado um grande número de vezes, e portanto, uma pequena diferença no tempo do caso de distâncias exatas, poderá se tornar uma grande diferença no caso mais geral.

Por esse motivo acreditamos que o tempo de processamento do ATA é uma de suas vantagens, já que nos testes realizados, o ATA se mostrou um pouco mais rápido que o UGB e essa pequena diferença no tempo para o caso de distâncias exatas será bem maior no caso de distâncias imprecisas, e portanto, o algoritmo ATA poderá ter uma melhor vantagem no estudo deste caso.

Como já foi comentado, os algoritmos ATA e UGB são semelhantes estruturalmente, diferindo principalmente pela dimensão do sistema linear resolvido. No caso do ATA, a dimensão é igual a 4, enquanto que no algoritmo UGB são quatro possíveis escolhas de sistemas lineares de dimensão 3. Portanto, enquanto no UGB é preciso calcular os números de condição de quatro matrizes, para depois resolver um sistema linear 3×3, no ATA não há busca a ser feita, isto é, resolvemos diretamente um sistema linear 4×4, resultando por isso, em um menor custo computacional.

No próximo capítulo iremos fazer algumas simulações computacionais nos métodos ATA e UGB, onde compararemos o tempo gasto em cada algoritmo e o erro entre a estrutura utilizada e a estrutura reportada pelo método. Tais estruturas podem ser artificiais ou reais. E a unidade de medida usada para a distância entre os átomos, e consequentemente usada na posição destes átomos no plano tridimensional, é o Ângstron, que é denotada por Å, onde 1 Å $= 10^{-10}$ metros. O erro, calculado via RMSD, que citaremos no próximo capítulo também é dado em Ângstron.

Capítulo 5

Experimentos Computacionais

Neste capítulo, iremos realizar uma comparação numérica entre os algoritmos UGB e ATA por meio de experimentos computacionais. Para isso, utilizaremos dois tipos diferentes de instâncias: geradas artificialmente e retiradas do banco de dados PDB. Quanto às instâncias artificiais, iremos gerá-las conforme descrito em [32], escolhidas pela simplicidade de implementação, e por simular estruturas moleculares reais. Além disso, serve como uma etapa preliminar da verificação da qualidade do algoritmo ATA, pois sabemos a solução. Essas estruturas são estabelecidas por uma cadeia de átomos enumerados de 1 a n, onde a distância entre dois átomos consecutivos é igual a 1.5 Å. Então, calculamos as distâncias entre todos os átomos, mas levamos em consideração, apenas as menores que 6 Å, formando o conjunto de dados iniciais dea entrada de nossos dois algoritmos. Além disso, observa-se que estas estruturas artificiais representam algumas características qualitativas de um problema real [32]. Em relação às moléculas reais que utilizaremos do banco de dados PDB [5] iremos medir todas as distâncias, mas utilizaremos como dados de entrada nos algoritmos apenas aquelas que forem menores que 8 Å.

Para fazer tal comparação entre os métodos, iremos levar em consideração duas variáveis, o tempo que cada algoritmo leva para calcular as coordenadas dos átomos e o erro entre a molécula original e a molécula reportada pelos métodos. Esse erro será calculado pelo RMSD. Como já discutimos anteriormente, o RMSD é uma forma de medir a distância média entre as estruturas de duas moléculas sobrepostas.

A forma que utilizamos o RMSD para calcular esse erro é similar com que fizemos na seção do ATA, sendo X a matriz com todas as distâncias da molécula original, seja ela artificial [32] ou real do PDB [5], e Y a matriz com todas as distâncias reportada por um dos métodos. Primeiro aplicamos a translação descrita em 4.11, resultando em Y'. Depois fatoramos a matriz $C = Y^T X$ atraves da SVD, obtendo $C = U\Sigma V^T$, e assim, calculamos a matriz de rotação $Q = UV^T$, finalmente poderemos obter o valor do RMSD, que será o erro, através do seguinte cálculo $\frac{\|X - YQ\|_F}{\sqrt{n}}$.

5.1 Geração de Instâncias Artificiais

Para realizarmos os experimentos, iremos utilizar como instâncias dois tipos de moléculas, moléculas reais extraídas do banco de dados PDB, e também moléculas artificiais, criadas através da descrição feita em [32]. Neste artigo Lavor mostra como "construir" moléculas artificiais que possam ser condizentes com as reais.

Antes, de introduzirmos a construção das coordenadas destas moléculas artificiais, precisamos enunciar algumas definições:

Definição 7. *Chamamos de **Comprimento de Ligação** o comprimento da ligação entre dois átomos, ou seja, a distância euclidiana entre as posições deles no espaço tridimensional.*

Definição 8. *O* ***Ângulo de Torção****, ou* ***Ângulo Diedral****, é o ângulo entre dois planos, ou seja, o ângulo entre os vetores normais a estes planos.*

Definição 9. *Sejam A, B e C três átomos da molécula em um espaço tridimensional, que são os vértices de duas ligações, então o ângulo entre os vetores,* $\vec{AB}$ *e* $\vec{BC}$*, que representam essas ligações é chamado de* ***Ângulo de Ligação****.*

Podemos ver melhor estas definições na figura 5.1, onde por exemplo, temos uma ligação entre os átomos i e $i+1$, então o Comprimento de ligação dos átomos i e $i+1$ é a distância entre eles. Temos que o ângulo $\Theta_{i,i+2}$ é o Ângulo de ligação entre os átomos $i, i+1$ e $i+2$, e finalmente o ângulo $\omega_{i,i+3}$ é o ângulo de torsão entre o plano definido pelos átomos $i, i+1$ e $i+2$ e o plano definido pelos átomos $i+1, i+2$ e $i+3$. Nesse caso os 4 átomos são não coplanares, da mesma forma que os átomos base nos métodos.

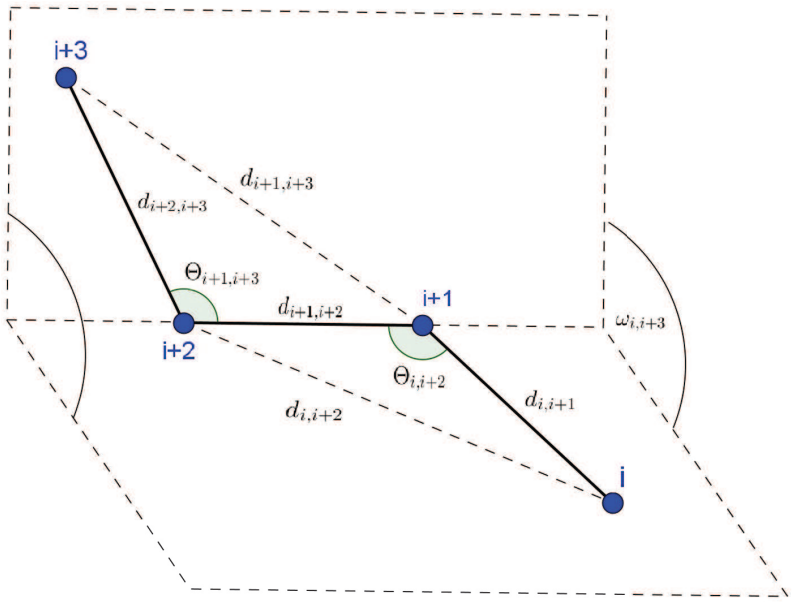


Figura 5.1: Representação de 4 átomos em uma molécula

Portanto, no caso das moléculas, temos que os comprimentos de ligação serão as distâncias moleculares entre dois átomos. E como em todas iterações dos métodos, os átomos base serão não coplanares, então eles estarão dispostos em dois planos, e assim, o ângulo entre estes dois planos será o ângulo de torsão da base. Estas estruturas artificiais foram motivadas pela sua simplicidade de implementação e por representarem algumas das características de uma molécula real de proteínas [33].

Sendo n a quantidade de átomos da molécula artifical, que queremos gerar, então, a estrutura artificial será uma cadeia com n átomos, isto é, uma sequência de n pontos, enumeradas de 1 a n. Conforme C. Lavor descreveu em [32], para gerar as coordenadas da estrutura artificial, devemos, para cada átomo i realizar as seguintes multiplicações:

$$\begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ 1 \end{bmatrix} = B_1 B_2 ... B_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{5.1}$$

onde $i \in \{1, 2, ..., n\}$ para encontrar suas coordenadas $x_i = (x_{i1}, x_{i2}, x_{i3})$. Sendo que $B_1 = I_4$,

$$B_2 = \begin{bmatrix} -1 & 0 & 0 & -d_{1,2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_3 = \begin{bmatrix} -\cos(\theta_{1,3}) & -\sin(\theta_{1,3}) & 0 & -d_{2,3}\cos(\theta_{1,3}) \\ \sin(\theta_{1,3}) & -\cos(\theta_{1,3}) & 0 & d_{2,3}\sin(\theta_{1,3}) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ e}$$

$$B_i = \begin{bmatrix} -\cos(\theta_{i-2,i}) & -\sin(\theta_{i-2,i}) & 0 & -d_{i-1,i}\cos(\theta_{i-2,i}) \\ \sin(\theta_{i-2,i})\cos(\omega_{i-3,i}) & -\cos(\theta_{i-2,i})\cos(\omega_{i-3,i}) & -\sin(\omega_{i-3,i}) & d_{i-1,i}\sin(\theta_{i-2,i})\cos(\omega_{i-3,i}) \\ \sin(\theta_{i-2,i})\sin(\omega_{i-3,i}) & -\cos(\theta_{i-2,i})\sin(\omega_{i-3,i}) & \cos(\omega_{i-3,i}) & d_{i-1,i}\cos(\theta_{i-2,i})\sin(\omega_{i-3,i}) \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

onde $i \in \{4, \ldots, n\}$.

A estrutura artificial que iremos gerar para fazer a comparação computacional entre os métodos será uma particularização desta estrutura que acabamos de apresentar. Iremos fixar os comprimentos de ligação e os ângulos de ligação entre átomos consecutivos, e apenas o ângulo de torsão irá variar, ou seja, nossas moléculas artificiais seguirão as seguintes regras [32, 21]:

1. Os comprimentos de ligação entre dois átomos consecutivos foram fixados em 1.5 Å, ou seja, $d_{i,i+1} = 1.5$ Å, onde $i \in \{1, \ldots, n-1\}$.

2. Os ângulos de ligação entre três átomos consecutivos foram fixados em $\frac{2\pi}{3}$, isto é, o ângulo entre os átomos $i, i+1$ e $i+2$ com vétice em $i+1$, denotado por $\theta_{i,i+2}$ é igual a $\frac{2\pi}{3}$ para todo $i \in \{1, \ldots, n-2\}$.

3. Os ângulos de torsão para cada conjunto de quatro átomos consecutivos $i, i+1, i+2$ e $i+3$, que denotaremos por $\omega_{i,i+3}$, são selecionados aleatoriamente do conjunto $\{\frac{\pi}{3}, \frac{\pi}{2}$ e $\frac{5\pi}{3}\}$, onde $i \in \{1, \ldots, n-3\}$.

Portanto, para gerar uma estrutura artificial com essas restrições, teremos que fazer a multiplicação (5.1), para cada átomo i, onde as matrizes B_i são:

$$B_1 = I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} -1 & 0 & 0 & -1,5 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_3 = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 & \frac{3}{4} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 & \frac{3\sqrt{3}}{4} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} e$$

$$B_i = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 & \frac{3}{4} \\ \frac{\sqrt{3}}{2}\cos(\omega_{i-3,i}) & \frac{1}{2}\cos(\omega_{i-3,i}) & -\sin(\omega_{i-3,i}) & \frac{3\sqrt{3}}{4}\cos(\omega_{i-3,i}) \\ \frac{\sqrt{3}}{2}\sin(\omega_{i-3,i}) & \frac{1}{2}\sin(\omega_{i-3,i}) & \cos(\omega_{i-3,i}) & \frac{3\sqrt{3}}{4}\sin(\omega_{i-3,i}) \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

onde $i \in \{4, \ldots, n\}$.

Após ter concluído a construção da estrutura artificial, iremos calcular a distância entre todos os átomos, desprezando aquelas que forem maiores que 6 Å. As distâncias restantes, aquelas que forem menores ou iguais a 6, serão utilizadas como os dados de entrada dos algoritmos, os quais irão calcular uma estrutura artificial conforme estas distâncias dadas, e ao final iremos comparar as estruturas que os métodos ATA e UGB calcularam com a estrutura gerada artificialmente através das fórmulas que acabamos de introduzir.

5.1.1 Experimentos Preliminares: Estabilidade nos Sistemas Lineares

Nesta seção serão apresentados os testes realizados para a verificação da estabilidade numérica obtida pela estratégia de reinicialização dos átomos base em relação ao algoritmo T e sua atualização. Foram utilizados três algoritmos nestes testes, o algoritmo T utilizando a fatoração SVD, a fatoração LU com estratégia de pivoteamento parcial e o algoritmo ATA. E como dados de entrada dos algoritmos, as distâncias obtidas de uma molécula artificial com 1500 átomos. O teste foi realizado para cada um dos métodos da seguinda forma: no calculo de cada átomo j remanescente nos algoritmos, foi computado o número de condição da matriz de coeficientes do sistema linear que seria resolvido para obter a posição deste átomo, após resolver tal sistema e ter encontrado a posição do átomo, comparavamos a estrutura parcial com j átomos encontrada até o momento com as posições do j primeiros átomos da estrutura original. Por exemplo, no algoritmo ATA, o primeiro passo é encontrar a posição dos 4 primeiros átomos, portanto, o cálculo da posição do primeiro átomo remanescente ocorre para o átomo 5. Nesse passo é preciso resolver um sistema linear para encontrar a posição deste átomo e pela tabela 5.1 vemos que o número de condição da matriz de coeficientes deste sistema foi de $1,50E+01$. Após calcular a posição deste quinto átomo, havia sido determinado até o momento uma estrutura parcial com 5 átomos, e finalmente comparamos, via RMSD, esta estrutura parcial com a estrutura dos 5 primeiros átomos da molécula original, resultando pela tabela 5.2 em um erro de $6,37E-16$.

Na figura 5.2 é possível verificar que a partir do centésimo átomo calculado no Algoritmo T, o número de condição das matrizes de coeficientes dos sistemas começa a aumentar drasticamente, independente da fatoração utilizada, enquanto que no caso do Algoritmo T Atualizado, estes números de condição permanecem com a mesma ordem de grandeza, gerando dessa forma uma maior estabilidade na resolução dos sistemas lineares. Esta estabilidade pode ser percebida na figura 5.4 que mostra que à medida que o número de condição cresce, o erro calculado pelo RMSD também cresce. Foi acrescentado um zoom nos 180 primeiros átomos para cada uma desta duas figuras, a fim de facilitar a visualização de que desde as menores moléculas o ATA já tem uma melhor estabilidade em relação ao AT, afinal todos os gráficos apresentados neste trabalho estão em escala logarítmica.

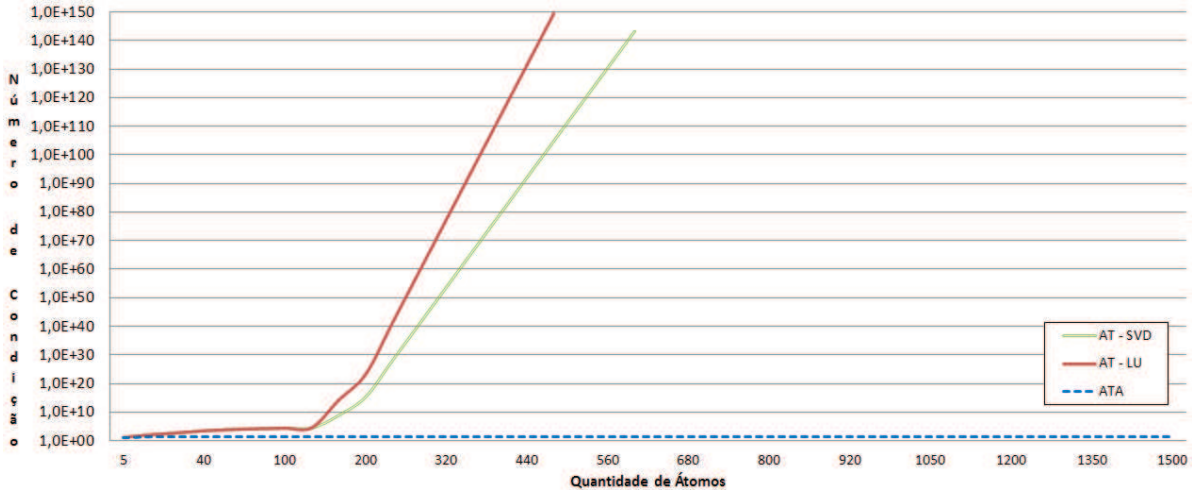


Figura 5.2: Gráfico com os Números de condição das matrizes de coeficientes dos Sistema lineares

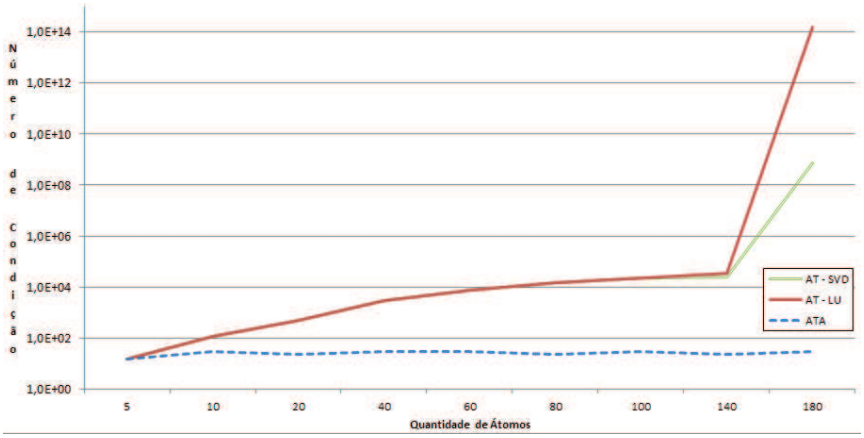


Figura 5.3: Gráfico com os Números de condição das matrizes de coeficientes dos Sistema lineares (ZOOM da Figura 5.2)

Átomo	Nº Cond. AT (LU)	Nº Cond. AT (SVD)	Nº Cond. ATA
5	1,50E+01	1,50E+01	1,50E+01
10	1,20E+02	1,20E+02	3,07E+01
15	3,06E+02	3,06E+02	3,07E+01
20	5,21E+02	5,21E+02	2,43E+01
25	9,68E+02	9,68E+02	3,07E+01
30	2,04E+03	2,04E+03	3,07E+01
35	1,67E+03	1,67E+03	2,43E+01
40	2,90E+03	2,90E+03	3,07E+01
45	4,43E+03	4,43E+03	3,07E+01
50	5,56E+03	5,56E+03	2,43E+01
60	7,86E+03	7,86E+03	3,07E+01
70	1,10E+04	1,10E+04	3,07E+01
80	1,50E+04	1,50E+04	2,43E+01
90	2,19E+04	2,19E+04	3,07E+01
100	2,38E+04	2,38E+04	3,07E+01
120	3,49E+04	3,63E+04	3,07E+01
140	3,53E+04	2,47E+04	2,43E+01
160	1,19E+05	2,52E+04	3,07E+01
180	1,46E+14	7,10E+08	3,07E+01
200	1,53E+23	1,78E+15	2,43E+01
240	1,69E+41	1,12E+28	3,07E+01
280	1,87E+59	7,04E+40	3,07E+01
300	1,96E+68	1,77E+47	3,07E+01
340	2,17E+86	1,11E+60	3,07E+01
380	2,40E+104	6,99E+72	2,43E+01
400	2,52E+113	1,75E+79	3,07E+01
440	2,78E+131	1,10E+92	2,43E+01
480	3,07E+149	6,94E+104	3,07E+01
500		1,74E+111	2,43E+01
540		1,09E+124	3,07E+01
580		6,88E+136	3,07E+01
600		1,73E+143	3,07E+01
640			3,07E+01
680			2,43E+01
720			3,07E+01
760			3,07E+01
800			2,43E+01
840			3,07E+01
920			2,43E+01
1000			3,07E+01
1100			2,43E+01
1200			3,07E+01
1300			3,07E+01
1400			2,43E+01
1500			3,07E+01

Tabela 5.1: Tabela com os Números de Condição das matrizes de coeficientes dos Sistema lineares

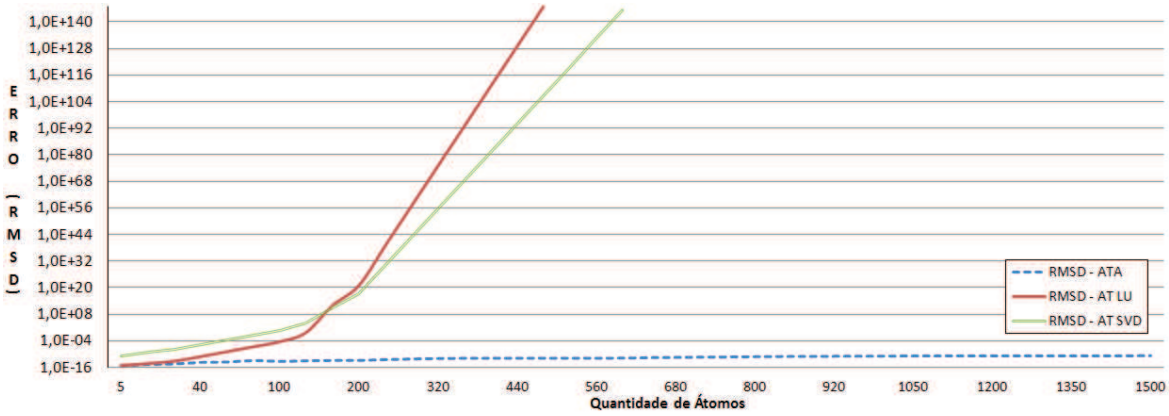


Figura 5.4: Gráfico relacionado ao erro (RMSD) no teste com os algoritmo ATA e AT (utilizando a fatoração LU e SVD)

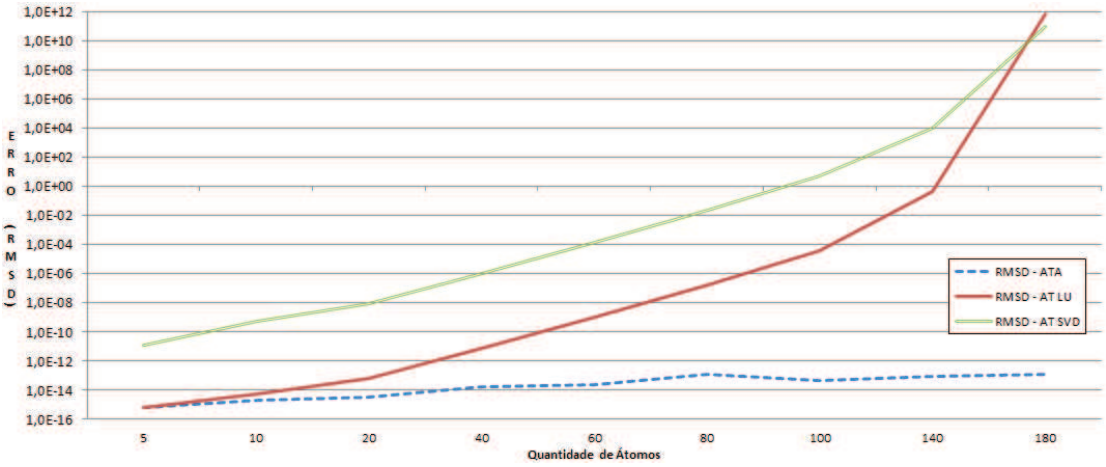


Figura 5.5: Gráfico relacionado ao erro (RMSD) no teste com os algoritmo ATA e AT (utilizando a fatoração LU e SVD) (ZOOM da Figura 5.4)

5.1.2 Comparação Numérica entre os Métodos ATA e UGB Utilizando Estruturas Artificiais

Para efeito de verificar a qualidade do algoritmo proposto, é realizada uma comparação entre a estrutura original, gerada de forma artificial, e aquela reconstruída pelos algoritmos UGB e ATA, por meio da RMSD, definida anteriormente. Quanto maior o valor da RMSD, maior será a distância entre a estrutura gerada pelo método e a gerada artificialmente. O objetivo desta medida é avaliar o quanto ambas estruturas tridimensionais se parecem geometricamente, ou seja, o quanto podem estar bem alinhadas no espaço tridimensional. É possível que o método gere uma estrutura que respeite as distâncias, mas não tenha forma semelhante à da instância testada (veja a Observação 3). Entretanto, nossa intenção é buscar estruturas tridimensionais parecidas geometricamente com a instância testada, além de que suas distâncias sejam compatíveis com as restrições. Logo, escolhemos a RMSD de modo a testar se a solução obtida satisfaz tais condições, visando recuperar a estrutura original. Assim, grandes valores para a RMSD mostram que o algoritmo não recuperou a estrutura como da instância testada.

Tal estimativa nos diz, qualitativamente, o quanto o algoritmo falhou em determinar a estrutura das moléculas. Porém, em alguns casos, esse valor alto do RMSD pode também nos indicar que as distâncias utilizadas na definição do problema não definem uma solução única.

# Átomos	RMSD AT (LU) (Å)	RMSD AT (SVD) (Å)	RMSD ATA (Å)
5	6,43E-16	1,33E-11	6,37E-16
10	5,02E-15	5,39E-10	2,08E-15
15	1,70E-14	3,15E-09	2,63E-15
20	5,98E-14	8,26E-09	3,21E-15
25	1,54E-13	2,56E-08	6,82E-15
30	4,83E-13	5,72E-08	9,80E-15
35	1,65E-12	2,17E-07	1,34E-14
40	7,03E-12	1,04E-06	1,70E-14
45	2,29E-11	3,23E-06	2,03E-14
50	7,86E-11	1,09E-05	2,24E-14
60	1,10E-09	1,53E-04	2,55E-14
70	1,58E-08	2,21E-03	3,48E-14
80	1,82E-07	2,53E-02	1,15E-13
90	2,57E-06	3,58E-01	5,53E-14
100	3,74E-05	5,21E+00	4,79E-14
120	6,24E-03	8,93E+02	6,50E-14
140	4,68E-01	1,05E+04	8,99E-14
160	7,30E+02	5,13E+04	1,16E-13
180	7,31E+11	9,98E+10	1,27E-13
200	7,29E+20	2,37E+17	1,37E-13
240	7,36E+38	1,36E+30	3,09E-13
280	7,53E+56	7,95E+42	6,23E-13
300	7,65E+65	1,93E+49	9,77E-13
340	7,94E+83	1,14E+62	1,07E-12
380	8,30E+101	6,77E+74	1,23E-12
400	8,51E+110	1,66E+81	1,27E-12
440	8,97E+128	9,93E+93	1,35E-12
480	9,49E+146	5,98E+106	1,37E-12
500	erro	1,47E+113	1,36E-12
540		8,90E+125	1,33E-12
580		5,40E+138	1,38E-12
600		1,33E+145	1,53E-12
640		erro	2,24E-12
680			3,10E-12
720			4,02E-12
760			5,08E-12
800			6,31E-12
840			7,78E-12
920			1,00E-11
1000			1,20E-11
1100			1,44E-11
1200			1,45E-11
1300			1,41E-11
1400			1,47E-11
1500			1,73E-11

Tabela 5.2: Tabela com erro (RMSD) no teste com os algoritmo ATA e AT (utilizando a fatoração LU e SVD)

No método UGB, para cada átomo não determinado, são construídos quatro sistemas lineares de ordem 3. Dentre estes, apenas o que possui menor número de condição de sua matriz de coeficientes é resolvido. Esta medida visa garantir mais estabilidade numérica evitando, assim, uma excessiva propagação de erros nos cálculos subsequentes [45]. Essa decisão demanda mais tempo de processamento. Já no ATA, é preciso construir um único sistema linear de ordem 4, para cada átomo remanescente. Como não é necessário fazer buscas prévias de sistemas lineares como no UGB, essa estratégia representa uma vantagem. Para resolver esses sistemas lineares, usamos fatoração LU com estratégia de pivoteamento parcial, uma vez que a qualidade dos resultados sugere que até o momento, uma estratégia de pivoteamento total não se faz necessária, tendo em vista a limitação do número de operações. Além disso no método ATA, há um fator de comparação, ainda em estudo, para decidir se a solução para a posição requerida é aceitável ou não, a saber, t_j definido no método. Este termo é estimado pelo próprio algoritmo e, também, é calculado em função da solução obtida, podendo funcionar como critério de parada alternativo [22].

Ambos os métodos foram implementados em linguagem de programação Matlab, em uma máquina com processador Intel Core $i7-2600$ CPU, 3.4 GHz e com memória RAM de 4 GB.

Foi utilizado o mesmo método de busca dos átomos base em todos algoritmos testados.

Visando comparar estruturas de diversas naturezas, realizamos testes com moléculas de pequeno porte com cinco átomos, o mínimo possível para o funcionamento dos métodos, até estruturas de grande porte com dezesseis mil átomos. Foram realizados 7 testes com estruturas artificiais de moléculas de proteínas. Em cada teste foram simuladas 66 diferentes moléculas, com diferentes tamanhos, isto é, quantidades diferentes de átomos. Para cada teste mostramos abaixo uma tabela resumida, com os resultados de 17 das 66 moléculas testadas, onde a primeira coluna é a quantidade de átomos da molécula, a segunda é o tempo (em segundos) que o ATA precisou para determinar a estrutura, a terceira coluna é o erro entre a estrutura original e a estrutura encontrada pelo ATA, calculado pelo RMSD, a quarta e a quinta coluna são o tempo de processamento e o erro referente ao algoritmo UGB, e finalmente a sexta coluna é a razão entre o tempo de processamente do ATA e o tempo do UGB, mostrando, então, o quanto um algoritmo foi mais rápido que o outro. Na última linha de cada tabela, temos a média dessa razão para as 66 moléculas do teste em questão. Apresentamos também, para cada teste, um gráfico para uma melhor comparação do tempo dos algoritmos e outro para a comparação do erro, levando em consideração as 66 moléculas.

Por exemplo, os resultados do primeiro teste podem ser observados na tabela 5.3 e na figura 5.6. Das 66 moléculas artificiais utilizadas no teste 1, os resultados de 17 são apresentados na tabela 5.3. É possível verificar nessa tabela que praticamente em todas moléculas descritas o tempo de processamento do algoritmo ATA ficou entre 50% e 60% do tempo do UGB, e que na média das 66 moléculas temos que o tempo do ATA foi $54,58\%$ do tempo do UGB, ou seja, em relação ao custo computacional o algoritmo ATA foi aproximadamente $45,42\%$ melhor que o algoritmo UGB. Quando verificamos o erro, que foi calculado via RMSD, entre a estrutura original e as estruturas reportadas pelos métodos, é possível visualizar que nas moléculas com $5, 40, 300, 400, 1000, 2000, 6000, 8000, 10000$ e 16000 átomos o erro do algoritmo ATA é menor que do algoritmo UGB, enquanto que nas 7 restantes o erro do UGB é menor, ou seja, não é possível dizer que um algoritmo tenha um erro consistentemente maior que outro, o que se confirma no gráfico da figura 5.6(a), que mostra claramente, que em alguns momentos o erro do algoritmo ATA é maior que o UGB, como nas moléculas entre 15 e 20 átomos e entre 50 e 100 átomos. Já em outros momentos, o erro do algoritmo UGB é maior, como por exemplo, nas moléculas com uma quantidade entre 2000 e 3500 átomos. O gráfico da figura 5.6(b) é referente ao tempo de cada algoritmo. É possível verificar que em todas moléculas o tempo do algoritmo ATA é menor que o UGB. Em alguns momentos o gráfico do tempo do ATA se "aproxima" do gráfico do tempo do UGB, ou seja, o tempo de ambos ficam próximos, e em outros momentos se "afastam", mostrando que o tempo dos algoritmos se distanciaram, porém, na maioria das moléculas a diferença dos tempos é aproximadamente igual.

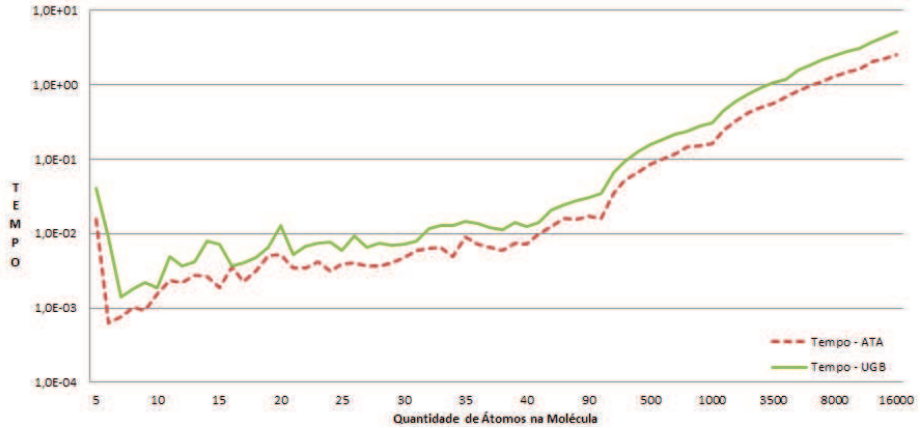
Quanto ao segundo teste, foram verificadas outras 66 moléculas diferentes do teste 1, porém com as mesmas quantidades de átomos, e podemos fazer observações parecidas com as anteriores, verificando as informações de seus resultados contidos na tabela 5.4 e nos gráficos da figura 5.7. Uma pequena diferença, foi que no teste 2 tivemos que o tempo do UGB foi menor que o tempo do ATA em uma molécula, no gráfico da figura 5.7 vemos que na molécula com 32 átomos o gráfico do tempo do algoritmo UGB fica abaixo do gráfico do tempo do ATA, mostrando que para aquela molécula o algoritmo UGB foi mais rápido. Porém, casos como esse foram exceções, e nós entendemos que essa diferença ocorreu por fatores inerentes ao computador e não ao algoritmo. Já em relação ao terceiro teste, há uma informação relevante que precisamos comentar: de todos testes realizados, esse foi o que o tempo do UGB foi menor que do ATA em mais moléculas, no caso 3 moléculas ($12000, 14000$ e 16000). Esse fato pode ser visto na tabela 5.4 e no gráfico da figura 5.7(b), por isso, resolvemos refazer o teste 3 e o denominamos de teste 4. Este teste por sua vez, foi similar aos demais, ou seja, um erro comparável entre os métodos , porém, com o algoritmo ATA em média $46,78\%$ mais rápido que o algoritmo UGB. Essas informações

podem ser verificadas na tabela 5.6 e na figura 5.9. Os outros 3 testes foram similares a estes já comentados e podemos verificar que na última coluna das tabelas, com exceção do teste 3, o tempo de processamento do algoritmo ATA está entre 50 % e 60 % do tempo do UGB.

# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
5	1,54E-02	1,47E-15	4,03E-02	1,54E-15	38,14%
40	7,18E-03	7,08E-15	1,25E-02	1,23E-14	57,25%
80	1,56E-02	4,17E-14	2,74E-02	2,56E-14	56,96%
100	1,59E-02	2,83E-14	3,44E-02	1,44E-14	46,24%
200	3,41E-02	1,03E-13	6,51E-02	8,28E-14	52,40%
300	5,26E-02	8,53E-14	9,49E-02	1,86E-13	55,46%
400	6,52E-02	1,00E-13	1,24E-01	1,12E-13	52,49%
500	8,45E-02	3,67E-13	1,56E-01	2,19E-13	54,33%
1000	1,60E-01	5,31E-13	3,06E-01	5,57E-13	52,39%
2000	3,26E-01	1,62E-12	6,09E-01	2,66E-12	53,52%
4000	6,73E-01	7,60E-12	1,19E+00	6,10E-12	56,49%
6000	9,83E-01	5,01E-12	1,82E+00	5,86E-12	54,13%
8000	1,31E+00	7,71E-12	2,46E+00	2,52E-11	53,18%
10000	1,61E+00	1,38E-11	3,09E+00	1,99E-11	51,95%
12000	2,00E+00	3,14E-11	3,69E+00	2,98E-11	54,21%
14000	2,25E+00	2,05E-11	4,33E+00	1,57E-11	51,81%
16000	2,54E+00	2,86E-11	5,07E+00	4,36E-11	50,13%
Média (66)					54,58%

Tabela 5.3: Tabela - Teste 1

(a) Gráfico RMSD



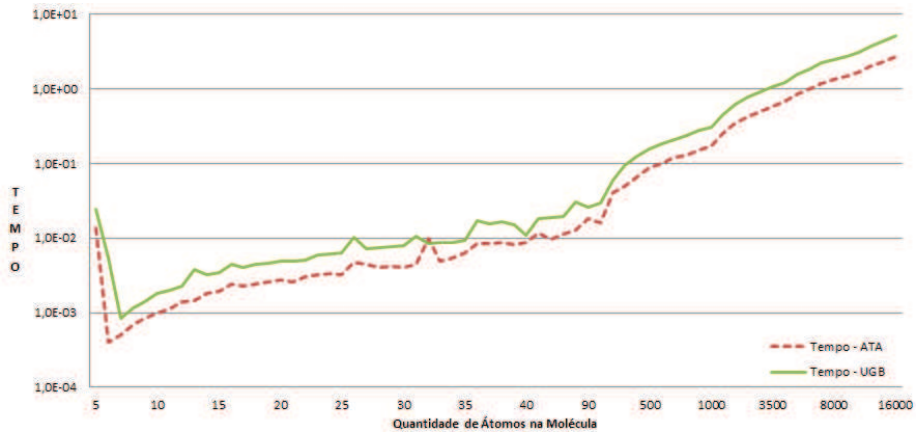
(b) Gráfico Tempo

Figura 5.6: Gráficos relacionados ao teste 1

# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
5	1,38E-02	9,68E-16	2,41E-02	9,97E-16	57,24%
40	8,83E-03	7,65E-15	1,11E-02	1,16E-14	79,81%
80	1,27E-02	1,97E-14	3,03E-02	4,02E-14	41,83%
100	1,61E-02	2,21E-14	2,94E-02	2,52E-14	54,88%
200	4,04E-02	6,12E-14	6,01E-02	1,84E-13	67,34%
300	4,93E-02	2,52E-13	9,28E-02	2,31E-13	53,07%
400	6,59E-02	2,85E-13	1,24E-01	2,04E-13	53,33%
500	8,77E-02	1,36E-13	1,57E-01	1,32E-13	55,94%
1000	1,69E-01	1,19E-12	3,01E-01	1,27E-12	56,19%
2000	3,43E-01	2,46E-12	6,16E-01	3,24E-12	55,65%
4000	6,91E-01	2,29E-12	1,23E+00	7,37E-12	56,35%
6000	1,01E+00	9,48E-12	1,85E+00	1,08E-11	54,28%
8000	1,33E+00	1,30E-11	2,42E+00	2,04E-11	54,82%
10000	1,67E+00	4,14E-11	3,02E+00	1,93E-11	55,25%
12000	2,00E+00	2,12E-11	3,69E+00	2,26E-11	54,26%
14000	2,32E+00	4,96E-11	4,39E+00	3,05E-11	52,88%
16000	2,68E+00	2,07E-11	5,07E+00	2,16E-11	52,84%
Média (66)					55,83%

Tabela 5.4: Tabela - Teste 2

(a) Gráfico RMSD



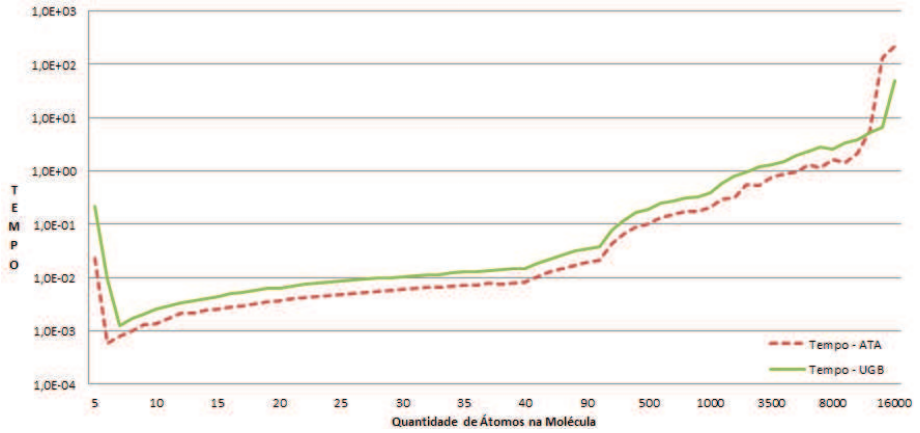
(b) Gráfico Tempo

Figura 5.7: Gráficos relacionados ao teste 2

# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
5	2,31E-02	1,11E-15	2,16E-01	8,64E-16	10,72%
40	8,34E-03	1,23E-14	1,47E-02	7,89E-15	56,60%
80	1,73E-02	2,92E-14	3,12E-02	3,30E-14	55,44%
100	2,12E-02	5,82E-14	3,80E-02	9,34E-14	55,92%
200	4,36E-02	9,37E-14	7,87E-02	7,50E-14	55,46%
300	6,56E-02	3,08E-13	1,18E-01	4,08E-13	55,37%
400	8,77E-02	3,02E-13	1,71E-01	3,53E-13	51,41%
500	1,04E-01	3,18E-13	1,88E-01	2,19E-13	55,34%
1000	2,08E-01	1,02E-12	3,94E-01	7,36E-13	52,86%
2000	3,17E-01	2,71E-12	8,00E-01	3,27E-12	39,61%
4000	8,83E-01	2,29E-12	1,52E+00	3,67E-12	57,98%
6000	1,29E+00	7,50E-12	2,40E+00	1,58E-11	53,78%
8000	1,65E+00	5,70E-12	2,64E+00	6,85E-12	62,38%
10000	2,18E+00	1,73E-11	3,80E+00	1,61E-11	57,43%
12000	5,69E+00	1,88E-11	5,29E+00	6,60E-11	107,69%
14000	1,32E+02	1,75E-11	6,57E+00	2,11E-11	2002,92%
16000	2,17E+02	2,26E-11	5,09E+01	2,78E-11	426,88%
Média (66)					90,09%

Tabela 5.5: Tabela - Teste 3

(a) Gráfico RMSD



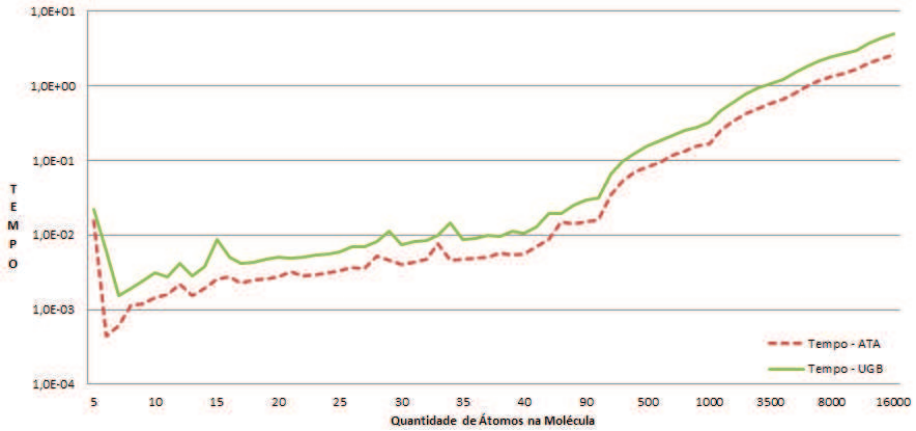
(b) Gráfico Tempo

Figura 5.8: Gráficos relacionados ao teste 3

# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
5	1,57E-02	1,11E-15	2,20E-02	8,64E-16	71,44%
40	5,56E-03	1,23E-14	1,07E-02	7,89E-15	51,96%
80	1,40E-02	2,92E-14	2,49E-02	3,30E-14	56,19%
100	1,60E-02	5,82E-14	3,11E-02	9,34E-14	51,34%
200	3,47E-02	9,37E-14	6,52E-02	7,50E-14	53,21%
300	5,26E-02	3,08E-13	9,55E-02	4,08E-13	55,11%
400	6,90E-02	3,02E-13	1,23E-01	3,53E-13	55,90%
500	8,27E-02	3,18E-13	1,55E-01	2,19E-13	53,43%
1000	1,69E-01	1,02E-12	3,24E-01	7,36E-13	52,07%
2000	3,36E-01	2,71E-12	6,05E-01	3,27E-12	55,48%
4000	6,66E-01	2,29E-12	1,22E+00	3,67E-12	54,68%
6000	9,99E-01	7,50E-12	1,82E+00	1,58E-11	54,93%
8000	1,33E+00	5,70E-12	2,48E+00	6,85E-12	53,46%
10000	1,67E+00	1,73E-11	3,01E+00	1,61E-11	55,38%
12000	2,01E+00	1,88E-11	3,70E+00	6,60E-11	54,29%
14000	2,30E+00	1,75E-11	4,44E+00	2,11E-11	51,75%
16000	2,59E+00	2,26E-11	4,95E+00	2,78E-11	52,30%
Média (66)					53,22%

Tabela 5.6: Tabela - Teste 4

(a) Gráfico RMSD

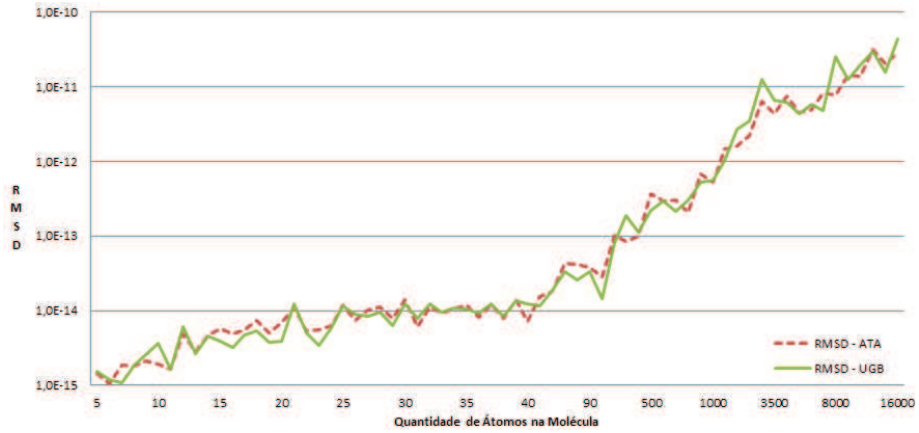


(b) Gráfico Tempo

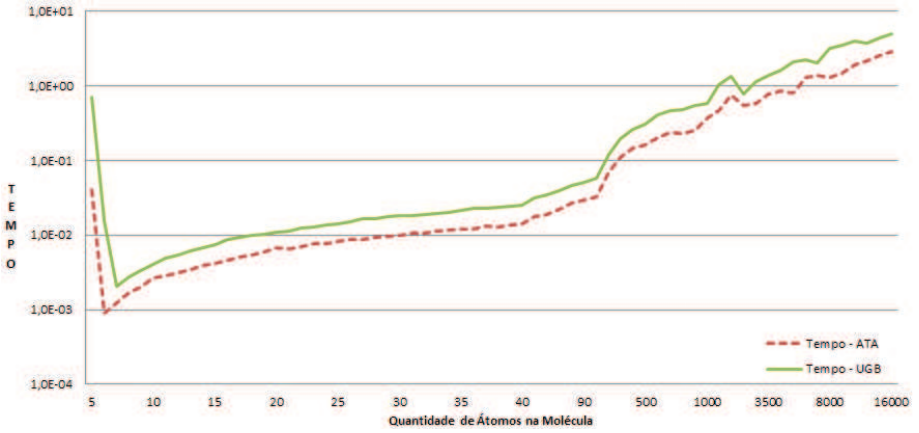
Figura 5.9: Gráficos relacionados ao teste 4

# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
5	4,09E-02	1,47E-15	6,93E-01	1,54E-15	5,90%
40	1,40E-02	7,08E-15	2,49E-02	1,23E-14	56,10%
80	2,64E-02	4,17E-14	4,65E-02	2,56E-14	56,71%
100	3,28E-02	2,83E-14	5,79E-02	1,44E-14	56,66%
200	7,08E-02	1,03E-13	1,19E-01	8,28E-14	59,54%
300	1,11E-01	8,53E-14	1,98E-01	1,86E-13	55,92%
400	1,48E-01	1,00E-13	2,63E-01	1,12E-13	56,48%
500	1,63E-01	3,67E-13	3,07E-01	2,19E-13	53,06%
1000	3,72E-01	5,31E-13	5,85E-01	5,57E-13	63,65%
2000	7,48E-01	1,62E-12	1,35E+00	2,66E-12	55,29%
4000	8,58E-01	7,60E-12	1,62E+00	6,10E-12	52,93%
6000	1,30E+00	5,01E-12	2,21E+00	5,86E-12	58,74%
8000	1,28E+00	7,71E-12	3,16E+00	2,52E-11	40,60%
10000	1,87E+00	1,38E-11	3,91E+00	1,99E-11	47,81%
12000	2,15E+00	3,14E-11	3,72E+00	2,98E-11	57,82%
14000	2,53E+00	2,05E-11	4,40E+00	1,57E-11	57,41%
16000	2,88E+00	2,86E-11	5,00E+00	4,36E-11	57,60%
Média (66)					54,24%

Tabela 5.7: Tabela - Teste 5



(a) Gráfico RMSD

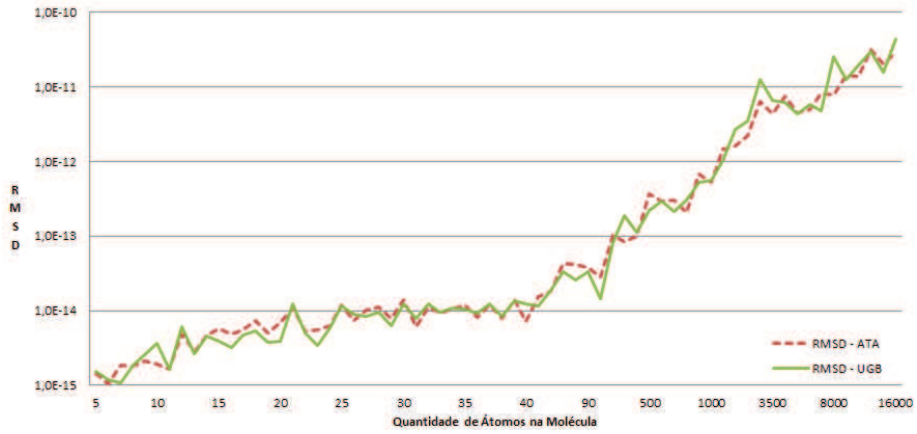


(b) Gráfico Tempo

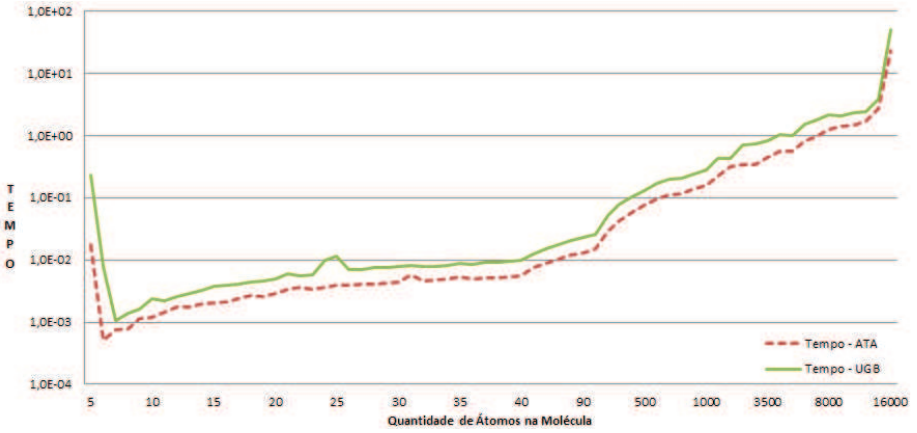
Figura 5.10: Gráficos relacionados ao teste 5

# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
5	1,78E-02	1,47E-15	2,28E-01	1,54E-15	7,80%
40	5,57E-03	7,08E-15	9,72E-03	1,23E-14	57,27%
80	1,18E-02	4,17E-14	2,05E-02	2,56E-14	57,69%
100	1,50E-02	2,83E-14	2,57E-02	1,44E-14	58,26%
200	2,89E-02	1,03E-13	5,25E-02	8,28E-14	55,09%
300	4,34E-02	8,53E-14	7,88E-02	1,86E-13	55,07%
400	5,84E-02	1,00E-13	1,04E-01	1,12E-13	56,05%
500	7,46E-02	3,67E-13	1,31E-01	2,19E-13	56,89%
1000	1,57E-01	5,31E-13	2,82E-01	5,57E-13	55,74%
2000	3,09E-01	1,62E-12	4,32E-01	2,66E-12	71,49%
4000	5,51E-01	7,60E-12	1,02E+00	6,10E-12	54,00%
6000	8,32E-01	5,01E-12	1,51E+00	5,86E-12	55,10%
8000	1,24E+00	7,71E-12	2,11E+00	2,52E-11	58,56%
10000	1,49E+00	1,38E-11	2,30E+00	1,99E-11	64,94%
12000	1,73E+00	3,14E-11	2,42E+00	2,98E-11	71,43%
14000	2,67E+00	2,05E-11	3,85E+00	1,57E-11	69,17%
16000	2,33E+01	2,86E-11	5,02E+01	4,36E-11	46,32%
Média (66)					56,26%

Tabela 5.8: Tabela - Teste 6



(a) Gráfico RMSD



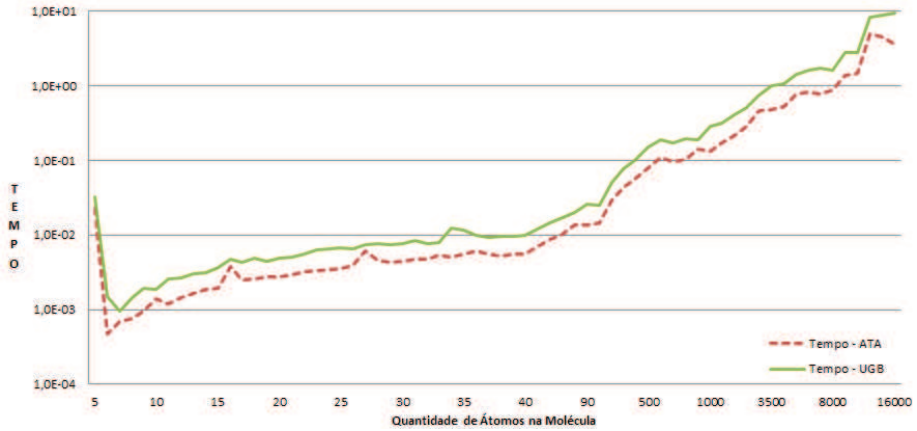
(b) Gráfico Tempo

Figura 5.11: Gráficos relacionados ao teste 6

# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
5	2,26E-02	1,47E-15	3,24E-02	1,54E-15	69,75%
40	5,52E-03	7,08E-15	9,87E-03	1,23E-14	55,94%
80	1,39E-02	4,17E-14	2,01E-02	2,56E-14	69,09%
100	1,48E-02	2,83E-14	2,53E-02	1,44E-14	58,52%
200	2,97E-02	1,03E-13	5,16E-02	8,28E-14	57,64%
300	4,36E-02	8,53E-14	7,82E-02	1,86E-13	55,74%
400	5,82E-02	1,00E-13	1,02E-01	1,12E-13	56,85%
500	7,95E-02	3,67E-13	1,52E-01	2,19E-13	52,24%
1000	1,33E-01	5,31E-13	2,87E-01	5,57E-13	46,36%
2000	2,17E-01	1,62E-12	4,10E-01	2,66E-12	52,86%
4000	5,34E-01	7,60E-12	1,08E+00	6,10E-12	49,53%
6000	8,14E-01	5,01E-12	1,62E+00	5,86E-12	50,31%
8000	8,87E-01	7,71E-12	1,60E+00	2,52E-11	55,58%
10000	1,45E+00	1,38E-11	2,75E+00	1,99E-11	52,86%
12000	4,93E+00	3,14E-11	8,29E+00	2,98E-11	59,44%
14000	4,48E+00	2,05E-11	8,77E+00	1,57E-11	51,10%
16000	3,63E+00	2,86E-11	9,55E+00	4,36E-11	38,01%
Média (66)					56,43%

Tabela 5.9: Tabela - Teste 7

(a) Gráfico RMSD



(b) Gráfico Tempo

Figura 5.12: Gráficos relacionados ao teste 7

5.2 Estruturas Moleculares do *Protein Data Bank* (PDB)

As moléculas reais foram retiradas do banco de dados *"Protein Data Bank"* [5], que como já foi dito é um banco de dados para estruturas tridimensionais de proteínas e aminoácidos, fundado em 1971 por Edgar Meyer e Walter Hamilton, e em outras técnicas, os dados contidos no PDB são frutos de métodos que utilizam instâncias dos experimentos de RMN. A escolha das proteínas no PDB foi baseada na escolha de Dong e Wu em [17, 18, 45].

5.2.1 Comparação Numérica entre os Métodos ATA e UGB Utilizando Moléculas do PDB

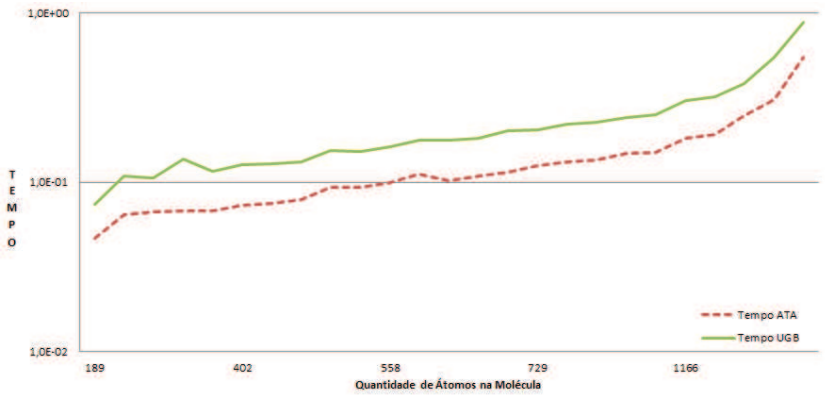
Foram realizados dois testes com proteínas reais retiradas do banco de dados PDB, em ambos foi calculado as distâncias entre todos pares de átomos da molécula escolhida, porém no primeiro teste, apresentado na tabela 5.10 e na figura 5.13, foram utilizados como instâncias iniciais dos algoritmos apenas as distâncias que fossem menores que 8 Å. Já no segundo teste, apresentado na tabela 5.11 e na figura 5.14, foram utilizados como entrada dos algoritmos apenas as distâncias que fossem menores que 6 Å.

Por exemplo, em relação ao primeiro teste, a tabela 5.10 apresenta os resultados numéricos realizados com as instâncias geradas à partir das moléculas proteicas do PDB com um corte de 8 Å. Na primeira e segunda coluna, temos o nome da molécula no PDB e o número de átomos presentes em sua estrutura. Na segunda e terceira colunas, apresentamos os valores de tempo de CPU, em segundos, e os valores da RMSD para o ATA. A quarta e quinta colunas apresentam os mesmos dados, mas para o UGB. Por fim, a última coluna da tabela apresentada mostra o valor relativo (ATA/UGB) entre os tempos de CPU. Em comparação com o UGB, os resultados do ATA são satisfatórios, com respeito à precisão e, para os resultados obtidos em tempo de CPU, o ATA exibiu um melhor desempenho. É possível ver isto tanto pelo gráfico da Figura 5.13(a) quanto pela última coluna da tabela 5.10 que apresenta o tempo relativo entre os dois métodos e que em média o algoritmo ATA foi $42,28\%$ mais rápido que o UGB. Quanto ao erro, que foi calculado pela RMSD entre a estrutura real e as estruturas reportadas pelos métodos, no gráfico da Figura 5.13(b) é possível verificar que em ambos os métodos foram praticamente iguais, e a tabela 5.10 mostra o quanto os erros foram parecidos, alguns se diferenciando apenas na terceira ou quarta casa decimal. Portanto, o método ATA, aqui proposto, é tão robusto e preciso quanto o UGB, porém obteve resultados computacionais satisfatórios e promissores, se mostrando mais rápido que o algoritmo UGB em relação ao tempo de processamento, o que acreditamos ser um bom resultado para a continuação de nossos estudos no caso de distâncias imprecisas, por causa dos Teoremas Fundamentais da Probabilidade.

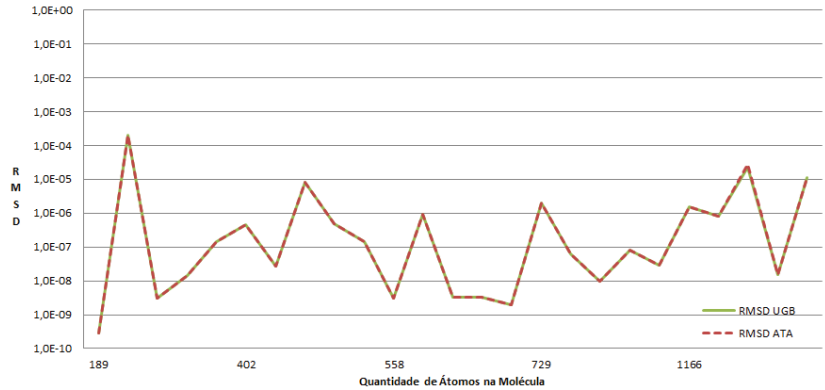
Em relação ao segundo teste, no qual realizamos um corte de 6 Å, isto é, utilizamos como entrada dos algoritmo as distâncias que fossem menores que 6 Å, obtemos resultados similares ao do primeiro teste, conforme pode ser observado na tabela 5.11 e na figura 5.14.

PDB Name	# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
1ID7	189	4,67E-02	2,9121E-10	7,47E-02	2,8377E-10	62,52%
1B5N	332	6,49E-02	2,0046E-04	1,09E-01	2,0046E-04	59,82%
1FW5	332	6,70E-02	3,1029E-09	1,07E-01	3,1027E-09	62,68%
1SOL	353	6,82E-02	1,4518E-08	1,37E-01	1,4513E-08	49,78%
1JAV	360	6,81E-02	1,4328E-07	1,16E-01	1,4327E-07	58,76%
1PTQ	402	7,36E-02	4,4987E-07	1,27E-01	4,4985E-07	58,00%
1MEQ	405	7,56E-02	2,7598E-08	1,28E-01	2,7581E-08	58,88%
1AMB	438	7,90E-02	8,2328E-06	1,32E-01	8,2329E-06	59,85%
1R7C	532	9,32E-02	4,8412E-07	1,54E-01	4,8315E-07	60,40%
1HLL	540	9,39E-02	1,4397E-07	1,52E-01	1,4398E-07	61,61%
1HOE	558	9,99E-02	3,0329E-09	1,62E-01	3,0810E-09	61,70%
1VII	596	1,12E-01	9,3210E-07	1,77E-01	9,3214E-07	63,65%
1HIP	617	1,02E-01	3,3136E-09	1,77E-01	3,3065E-09	57,72%
1LFB	641	1,09E-01	3,1951E-09	1,82E-01	3,1951E-09	60,13%
1URL	677	1,15E-01	2,0006E-09	2,03E-01	2,0025E-09	56,45%
1AIK	729	1,26E-01	1,9971E-06	2,04E-01	1,9971E-06	61,54%
1PHT	811	1,32E-01	6,3880E-08	2,20E-01	6,3802E-08	60,15%
1CEU	854	1,36E-01	9,7657E-09	2,27E-01	9,7687E-09	60,10%
1POA	914	1,48E-01	8,2019E-08	2,43E-01	8,2043E-08	60,91%
1KVX	954	1,51E-01	2,9637E-08	2,51E-01	2,9603E-08	60,01%
1VMP	1166	1,83E-01	1,5584E-06	3,03E-01	1,5584E-06	60,40%
1HSM	1251	1,93E-01	8,1260E-07	3,20E-01	8,1265E-07	60,34%
1HAA	1310	2,47E-01	2,7982E-05	3,83E-01	2,2997E-05	64,54%
1RGS	2015	3,06E-01	1,5418E-08	5,46E-01	1,5405E-08	56,05%
1BPM	3671	5,50E-01	1,1548E-05	8,85E-01	1,1547E-05	62,16%
Média						57,62%

Tabela 5.10: Tabela com moléculas do PDB - Teste com 8 Å



(a) Gráfico Tempo

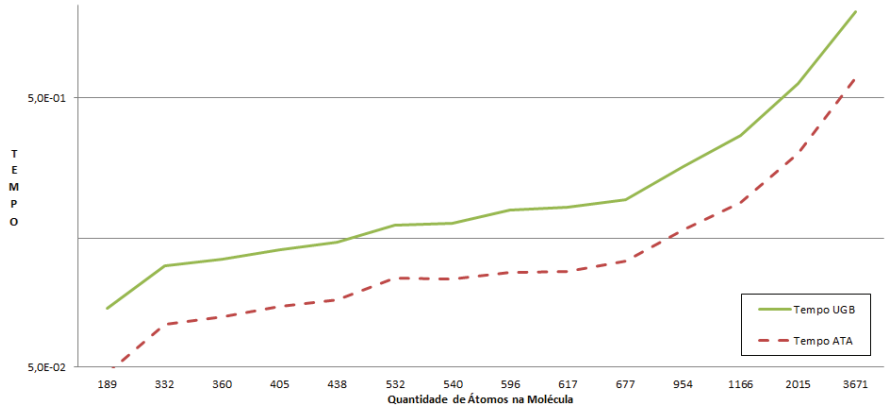


(b) Gráfico RMSD

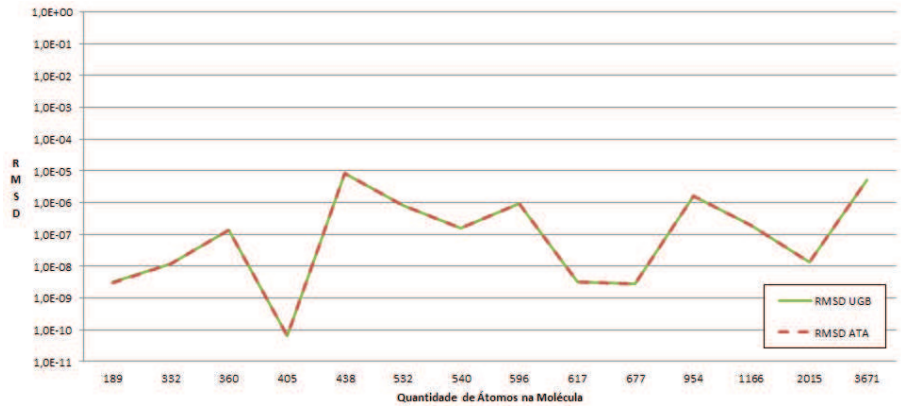
Figura 5.13: Gráficos relacionados ao teste das moléculas do PDB (8Å)

PDB Name	# Átomos	Tempo ATA (*s*)	RMSD ATA (Å)	Tempo UGB (*s*)	RMSD UGB (Å)	Tempo(ATA/UGB)
1ID7	189	4,74E-02	3,12E-09	8,27E-02	3,12E-09	57,32%
1FW5	332	7,19E-02	1,18E-08	1,19E-01	1,18E-08	60,42%
1JAV	360	7,66E-02	1,38E-07	1,26E-01	1,38E-07	60,70%
1MEQ	405	8,39E-02	6,39E-11	1,36E-01	6,42E-11	61,51%
1AMB	438	8,86E-02	8,22E-06	1,45E-01	8,22E-06	61,15%
1R7C	532	1,07E-01	8,39E-07	1,68E-01	8,38E-07	63,83%
1HLL	540	1,06E-01	1,59E-07	1,71E-01	1,59E-07	61,96%
1VII	596	1,12E-01	9,19E-07	1,92E-01	9,19E-07	58,55%
1HIP	617	1,13E-01	3,15E-09	1,95E-01	3,14E-09	57,98%
1ULR	677	1,24E-01	2,82E-09	2,09E-01	2,81E-09	59,20%
1KVX	954	1,61E-01	1,65E-06	2,77E-01	1,65E-06	58,19%
1VMP	1166	2,04E-01	1,97E-07	3,63E-01	1,97E-07	56,25%
1RGS	2015	3,11E-01	1,37E-08	5,63E-01	1,37E-08	55,23%
1BPM	3671	5,94E-01	5,08E-06	1,04E+00	5,08E-06	57,11%

Tabela 5.11: Tabela com moléculas do PDB - Teste com 6 Å



(a) Gráfico Tempo



(b) Gráfico RMSD

Figura 5.14: Gráficos relacionados ao teste das moléculas do PDB (6Å)

Capítulo 6

Conclusões e Perspectivas Futuras

Neste trabalho, apresentamos um novo método para resolver Problemas de Geometria de Distâncias Moleculares com Conjuntos Arbitrários de Distâncias, chamado Algoritmo T Atualizado (ATA), que fora uma atualização do Algoritmo T introduzido em [21]. Ambas famílias de métodos que citamos, "família *Geometric Bulid-Up*" e "família T", têm como cerne a resolução de sistemas lineares de pequeno porte, e inicialmente mal-condicionados, porém através de uma reinicialização nos átomos base, introduzida em [45], há uma melhora quanto ao condicionamento das matrizes de coeficientes dos Sistemas Lineares que serão resolvidos na determinação de cada átomo remanescente, obtendo portanto, uma maior estabilidade nestes Sistemas, pois como comentado anteriormente, os átomos base, nesta estratégia, são tratados de modo a diminuir o acúmulo de erros para não afetar o cálculo das posições dos átomos que seguem. Os algoritmos AT e ATA podem ser vistos como uma extensão do método GB introduzido em [18], com a possibilidade de tratar incertezas/imprecisões inerentes no processo de medição de problemas de geometria de distâncias moleculares, já que estes algoritmos contam com a variável $t_j = -\left\|x_j^*\right\|^2/2$, descrita no capítulo 3, que pode ser usada como uma medida, a cada passo do ATA, da imprecisão dos dados. Esse importante aspecto continuará sendo explorado para compreender em detalhes todo seu potencial para essa classe de problemas, bem como suas limitações [22].

Para trabalhos futuros, a nossa principal perspectiva é a aplicação de método numérico proposto ATA para o tratamento de distâncias com ruídos nas medições ou distâncias inexatas. Para tanto, uma alternativa científica é introduzir uma nova abordagem via uma modelagem estocástica do problema de Geometria de Distância Molecular em conjunto com técnicas do tipo de Monte Carlo.

Os resultados obtidos com o ATA em relação ao UGB são promissores, pois mostram eficiência no cálculo das estruturas e em tempo relativamente menor, controlando, assim, a propagação de erros numéricos. Através dos gráficos da figura 5.13(b) é possível visualizar que o erro, calculado via RMSD, entre a estrutura original (real) e a calculada pelos métodos ATA e UGB são praticamente iguais, enquanto pelos gráficos da figuras 5.13(a) podemos verificar que o tempo que o algoritmo ATA precisa para reportar a estrutura molecular é menor que o do algoritmo UGB. Essa diferença no tempo entre os algoritmos foi pequena nos nossos testes (alguns décimos de segundos), e se nosso objetivo fosse apenas determinar estruturas moleculares onde se conhecem apenas distâncias exatas entre pares de átomos, poderíamos inclusive implementá-los com outras estratégias para melhorar o condicionamento da matriz de coeficientes, afinal, apesar dessas implementações deixarem os algoritmos mais caros computacionalmente, esse tempo "a mais" não faria grande diferença, pois estamos tratando de segundos. Porém, nosso objetivo é trabalhar com a determinação de estruturas moleculares onde para algumas das distâncias são dados apenas intervalos onde estas podem estar contidas, que é o caso mais comum. Mas para trabalhar com esse problema de distâncias imprecisas precisamos fazer um estudo de quantifica-

ção de incertezas, e para isso precisaremos utilizar ferramentas da probabilidade e da estatística, como a Lei dos Grandes Números, também chamado de Teorema Fundamental da Probabilidade. Por exemplo, em um estudo de Monte Carlo precisaríamos de um número muito grande de iterações dos algoritmos para cada átomo não determinado, e essa pequena diferença do tempo em nossos testes que o algoritmo ATA tem de vantagem pode se transformar em uma grande diferença quando formos tratar de distâncias imprecisas.

Abaixo temos uma lista com algumas perspectivas futuras de questões que pretendemos dar continuidade em nossos estudos:

- Estender o método para tratar dos problemas onde somente cotas inferiores e superiores para os valores de distâncias são fornecidos pela RMN (Distâncias imprecisas);
- Aplicar nosso método numérico no caso de distâncias imprecisas, através de uma modelagem estocástica do Problema de geometria de distância molecular (PGDM), por meio de uma abordagem de Monte Carlo;
- Realizar um estudo sobre a complexidade dos algoritmos;
- Melhorar o condicionamento das matrizes dos sistemas lineares resolvidos na busca de cada átomo remanescente, sem afetar o custo computacional;
- Verificar as vantagens e limitações da variável t_j quanto ao PGDM com distâncias exatas;
- Verificar a possibilidade da utilização da variável t_j no caso de distâncias imprecisas;
- Considerar estruturas de maior porte do *Protein Data Bank*;
- Fazer um estudo sobre quantificação de imprecisão/incerteza;
- Realizar um estudo sobre as possíveis visualizações em 3D da estruturas moleculares;
- Aplicar o método em estruturas proteicas com estrutura ainda desconhecida.

Referências Bibliográficas

[1] **Anfinsen, C. B., Haber, E., Sela, M., White, F. H. Jr.** (1961), *The kinetics of formation of native ribonuclease during oxidation of reduced polypeptide chain*, Proceedings of the national academy of sciencies of USA, 47 : 1309 – 1314.

[2] **Amabis, J.M., Martho, G.R.** (1997), *Do gene à proteína*, Atualidades Biológicas, Editora Moderna.

[3] **Arun, K.S., Huang, T.S., Blostein, S.D.** (1987), *Least-squares fitting of two* $3-D$ *point sets*, IEEE Transactions on Pattern Analysis Machine Intelligence, 9 : 698 – 700.

[4] **Berg, J. M., Tymoczko, J. L., Stryer, L.** (2006), *Biochemistry*, W. H. Freeman.

[5] **Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I., Bourne, P.** (2000), *The protein data bank*, Nucleid Acids Research.

[6] **Blumenthal, L.M.** (1953), *Theory and Applications of Distance Geometry*, Clarendon Press, Oxford.

[7] **Branden, C., Tooze, J.** (1991), *Introduction to Protein Structure*, Garland Publishing.

[8] **Casella, G., Berger, R. L.** (2010), *Inferência estatística*, Cengage Learning, tradução da 2ª edição, São Paulo-SP.

[9] **Chandonia, J. M., Brenner, S. E.** (2006), *The impact of structural genomics: expectations and outcomes*, Science 311, 347 – 351.

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

[11] **Crippen, G. M.** (1981), *Distance Geometry and Conformational Calculations*, In Chemometrics Research Studies Series, volume 1. Research Studies Press (Wiley), New York.

[12] **Crippen, G.M., Havel, T.F.** (1988), *Distance Geometry and Molecular Conformation*, John Wiley e Sons, New York.

[13] **Dantas, C. A. B.** (1997), *Probabilidade: Um Curso Introdutório*, Edusp, 3ª Edição, São Paulo-SP.

[14] **Davis, R. T., Ernest, C. e Wu, D.** (2009): *Protein structure determination via an efficient geometric build-up algorithm*, Computational Structural Bioinformatics Workshop, Washington, DC, USA;

[15] **De Roberts Jr., E.M.F.** (1985), *Bases da Biologia Celular e Molecular*, Editora Guanabara.

[16] **Demmel, J.** (1996), *Applied Numerical Linear Algebra*, SIAM.

[17] **Dong, Q. e Wu, Z.** (2002), *A linear-time algorithm for solving the Molecular Distance Geometry Problem with exact inter-atomic distances*, Journal of Global Optimization 22, pp. 365-375.

[18] **Dong, Q. e Wu, Z.** (2003), *A Geometric Build-Up Algorithm for solving the Molecular Distance Geometry Problem with sparse distance data*, Journal of Global Optimization 26, pp. 321-333.

47

[19] **Eggert, D.W., Lorusso, A., and Fisher. R.B.** (1997), *Estimating* $3-D$ *rigid body transformations: a comparison of four major algorithms*, Machine Vision and Applications, $9 : 272-290$.

[20] **Ezzel, C.** (2002), *Proteins rule*, Scientific American, $286(4) : 40-7$.

[21] **Fidalgo, F.** (2011), *Algoritmos para Problemas de Geometria Molecular*, Dissertação de Mestrado, IMECC - UNICAMP, Campinas.

[22] **Fidalgo, F., Maioli, D., Abreu, E. e Lavor, C.** (2012), *Uma formulação numérica para resolução de problemas de geometria de distâncias moleculares*, XVI CLAIO, ALIO, Rio de Janeiro-RJ. http://www2.claiosbpo2012.iltc.br/pdf/102269.pdf

[23] **Fidalgo, F., Maioli, D., Abreu, E. e Lavor, C.** (2012), *A Numerical Formulation For Solving The Molecular Distance Geometry Problem*, em preparação.

[24] **Golub, G. e Van Loan, C.** (1996), *Matrix Computations*, The Johns Hopkins University Press, 3ª Edição.

[25] **Guentert, P.** (1998), *Structure calculation of biological macromolecules from NMR data*, Quarterly Reviews of Biophysics, $31(2) : 145-237$..

[26] **Gunther, H.** (1995), *NMR Spectroscopy: basic principles, concepts and applications in chemistry*, John Wiley & Sons.

[27] **Hama, L.** (2004), *O mapa da vida*, Super Interessante.

[28] **Lander, E. S., et al.** (2001), **International Human Genome Sequencing Consortium**, *Initial sequencing and analysis of the human genome*, Nature, $409(6822) : 860-921$.

[29] **Kabsch, W.** (1976), *A solution for the best rotation to relate two sets of vectors*, Acta Crystallographica A, 32:922-923.

[30] **Kabsch, W.** (1978), *A discussion of the solution for the best rotation to relate two sets of vectors*, Acta Crystallographica A, 34:827-828.

[31] **Karp, G.** (2005), *Biologia Celular e Molecular - Conceitos e Experimentos*, Editora Manole, Barueri.

[32] **Lavor, C.** (2006), *On generating instances for the molecular distance geometry problem*, Nonconvex Optimization and Its Applications 84, pp. 405-414.

[33] **Lavor, C., Liberti, L, Maculan, N., Mucherino, A.** (2012), *The discretizable molecular distance geometry problem*, Computational Optimization and Application, Vol. 52, 1:115-146.

[34] **Liberti, L, Lavor, C. e Maculan, N.** (2008), *A Branch-And-Prune algorithm for the Molecular Distance Geometry Problem*, International Transactions in Operational Research 15, pp.1-17.

[35] **Liberti, L., Lavor, C., Mucherino, A. e Maculan, N.** (2010), *Molecular Distance Geometry methods: from continuous to discrete*, International Transactions in Operational Research 18, pp. 33-51.

[36] **Lemos, M.** (2000), *Gerenciamento de Memória para Comparação de Biossequências*, Dissertação de Mestrado, PUC, Rio de Janeiro.

[37] **Magalhaes, M. N.** (2006), *Probabilidade e variáveis aleatórias*, Edusp, São Paulo-SP.

[38] **Russell, R. B., Eggleston, D. S.** (2000), *New roles for structure in biology and drug discovery*, Nature Structural Biology 7 ,928 – 930, 2000.

[39] **Saxe, J.** (1979), *Embeddability of weighted graphs in k-space is strongly NP-hard*, Proceedings of 17th Allerton Conference in Communications, Control and Computing, Monticello, IL, pp. 480-489.

[40] **Schlick, T.** (2002), *Molecular modeling and simulation: an interdisciplinary guide*, Springer, New York.

[41] **Shibuya, T.** (2007), *Efficient Substructure RMSD Query Algorithms*, Journal of Computacional Biology, Vol. 14, 9:1201-1207.

[42] **Sit, A.** (2010), *Solving distance geometry problems for protein structure determination*, Tese de Doutorado, Iowa State University, Ames, Iowa.

[43] **Souza, M. F.** (2010), *Suavização Hiperbólica Aplicada à Otimização de Geometria Molecular*, Tese de Doutorado, UFRJ, Rio de Janeiro.

[44] **Venter, J. C., et al.** (2001), *The sequence of the human genome*, Science, 291(5507) : 1304 – 1351.

[45] **Wu, D. e Wu, Z.** (2007), *An Updated Geometric Build-Up Algorithm for solving the Molecular Distance Geometry Problem with sparse distance data*, Journal of Global Optimization 37, pp. 661-673.

[46] **Wu, Z.** (2008), *Lecture Notes on Computational Structural Biology*, World Scientific Publishing Company.

[47] **Wüthrich, K.** (1995), *NMR in Structural Biology*, World Scientific Publishing Company.

[48] **Yoon, J. M., Gad, Y., Wu, Z.** (2000), *Mathematical modeling of protein structure using distance geometry*, Technical report, Rice University.

Apêndice A

Demonstração da Consistência dos Métodos

Uma de nossas preocupações era verificar e provar que os métodos baseados no teorema 3.1.5 e no corolário 4.1.3 são consistentes, isto é, deveríamos provar que a solução do sistema quadrático

$$\begin{aligned} ||x-a_1|| &= d_1 \\ ||x-a_2|| &= d_2 \\ ||x-a_3|| &= d_3 \\ ||x-a_4|| &= d_4 \end{aligned} \tag{A.1}$$

nestes teoremas além de existir, deveria ser única, e assim a transformação linear aplicada nos Sistemas não-lineares iriam manter a mesma solução. Chegamos através de contra-exemplos que matematicamente é possível que não tenha nenhuma solução, porém como o problema é físico, sabemos que a molécula existe, e que as distâncias foram calculadas dela, então, a menos de erro na medição, a solução real do problema também será uma solução do sistema, e portanto, o sistema quadrático terá pelo menos uma solução, que é a solução real. Assim, nos resta provar que o sistema quadrático em questão não poder ter mais de uma solução, para podermos afirmar que a solução é única. Vejamos então, um lema auxiliar para o próximo teorema.

Lema A.1. *Se* $\{a_1,a_2,a_3,a_4\} \subset \mathbb{R}^3$ *são pontos não coplanares, então o conjunto de vetores* $B = \{(a_2-a_1),(a_3-a_1),(a_4-a_1)\} \subset \mathbb{R}^3$ *é linearmente independente.*

Por meio da contrapositiva deste lema, temos que se $B = \{(a_2-a_1),(a_3-a_1),(a_4-a_1)\} \subset \mathbb{R}^3$ é linearmente dependente, então $\{a_1,a_2,a_3,a_4\} \subset \mathbb{R}^3$ são coplanares.

Resolvemos elaborar o teorema abaixo e demonstrá-lo ao observar que em nenhum dos trabalhos que contém os teoremas mencionados citam que o sistema não linear A.1 tem apenas uma solução, e percebemos que para que os métodos fossem consistentes deveríamos provar que tal sistema deveria ter uma única solução dado que não houve erros nas medições.

Teorema A.2. *Se tivermos que* $a_1,a_2,a_3,a_4 \in \mathbb{R}^3$ *são não coplanares, então o sistema quadrático*

$$\begin{aligned} ||x-a_1|| &= d_1 \\ ||x-a_2|| &= d_2 \\ ||x-a_3|| &= d_3 \\ ||x-a_4|| &= d_4 \end{aligned} \tag{A.2}$$

tem no máximo 1 solução.

Demonstração. Supomos inicialmente, que o sistema (2.11) tenha duas soluções distintas, ou seja, que b e $\bar{b}$ sejam soluções, onde $b \neq \bar{b}$, então

$$\begin{array}{lll} ||b-a_1|| = d_1 & & ||\bar{b}-a_1|| = d_1 \\ ||b-a_2|| = d_2 & e & ||\bar{b}-a_2|| = d_2. \\ ||b-a_3|| = d_3 & & ||\bar{b}-a_3|| = d_3 \\ ||b-a_4|| = d_4 & & ||\bar{b}-a_4|| = d_4 \end{array}$$

Elevando os dois lados ao quadrado de todas as 8 equações obtemos,

$$\begin{array}{ll} ||b||^2 - 2a_1^T b + ||a_1||^2 = d_1^2 & ||\bar{b}||^2 - 2a_1^T \bar{b} + ||a_1||^2 = d_1^2 \\ ||b||^2 - 2a_2^T b + ||a_2||^2 = d_2^2 & ||\bar{b}||^2 - 2a_2^T \bar{b} + ||a_2||^2 = d_2^2, \\ ||b||^2 - 2a_3^T b + ||a_3||^2 = d_3^2 \text{ e} & ||\bar{b}||^2 - 2a_3^T \bar{b} + ||a_3||^2 = d_3^2 \\ ||b||^2 - 2a_4^T b + ||a_4||^2 = d_4^2 & ||\bar{b}||^2 - 2a_4^T \bar{b} + ||a_4||^2 = d_4^2 \end{array}$$

Substituindo $d_i^2 = ||b||^2 - 2a_i^T b + ||a_i||^2$ do sistema (2.14), onde $i = 1,2,3,4$, no sistema (2.15), teremos

$$\begin{array}{l} ||\bar{b}||^2 - 2a_1^T \bar{b} + ||a_1||^2 = ||b||^2 - 2a_1^T b + ||a_1||^2 \\ ||\bar{b}||^2 - 2a_2^T \bar{b} + ||a_2||^2 = ||b||^2 - 2a_2^T b + ||a_2||^2 \\ ||\bar{b}||^2 - 2a_3^T \bar{b} + ||a_3||^2 = ||b||^2 - 2a_3^T b + ||a_3||^2 \\ ||\bar{b}||^2 - 2a_4^T \bar{b} + ||a_4||^2 = ||b||^2 - 2a_4^T b + ||a_4||^2 \end{array}$$

Como $b^T a_i = a_i^T b$, já que $b, a_i \in \mathbb{R}^3 \ \ \forall i = 1,2,3,4$, então fazendo essa substituição,

$$\begin{array}{l} ||\bar{b}||^2 - ||b||^2 = 2\bar{b}^T a_1 - 2b^T a_1 \\ ||\bar{b}||^2 - ||b||^2 = 2\bar{b}^T a_2 - 2b^T a_2 \\ ||\bar{b}||^2 - ||b||^2 = 2\bar{b}^T a_3 - 2b^T a_3 \\ ||\bar{b}||^2 - ||b||^2 = 2\bar{b}^T a_4 - 2b^T a_4 \end{array}$$

Obtemos, assim, o seguinte sistema,

$$\begin{array}{l} 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2} a_1 = 1 \\ 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2} a_2 = 1 \\ 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2} a_3 = 1 \\ 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2} a_4 = 1 \end{array}$$

Se subtraírmos 1 dos dois lados das 3 últimas equações e substituirmos $1 = 2\frac{(\bar{b}^T - 2b^T)}{||\bar{b}||^2 - ||b||^2} a_1$, referente a primeira equação, teremos

$$\begin{array}{l} 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2} (a_2 - a_1) = 0 \\ 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2} (a_3 - a_1) = 0 \\ 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2} (a_4 - a_1) = 0 \end{array}$$

Como $0+0+0=0$, então

$$2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2}(a_2 - a_1) + 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2}(a_3 - a_1) + 2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2}(a_4 - a_1) = 0$$

Colocando $2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2}$ em evidência,

$$2\frac{(\bar{b}^T - b^T)}{||\bar{b}||^2 - ||b||^2}((a_2 - a_1) + (a_3 - a_1) + (a_4 - a_1)) = 0$$

Como supomos inicialmente, que $b \neq \bar{b}$, então,

$$(a_2 - a_1) + (a_3 - a_1) + (a_4 - a_1) = 0$$

O que implica que $(a_2 - a_1)$, $(a_3 - a_1)$ e $(a_4 - a_1)$ são Linearmente Dependentes, portanto, pelo lema 1, a_1, a_2, a_3, a_4 são coplanares. Absurdo, pois supomos que eles são não coplanares. Logo, o sistema (1), tem no máximo 1 solução.

□

Assim, supondo que não houve erro no cálculo da distância entre as moléculas, temos que a solução do sistema quadrático existe e é única. Sabemos também que a solução do sistema linear de ambos teoremas existem e são únicas, pois as matrizes de coeficientes dos dois sistemas tem posto completo (não-singular). Então, podemos afirmar que a solução do sistema A.1 também será a solução do sistema linear do Teorema 3.1.5 e as três últimas coordenadas da solução do Corolário 4.1.3.