UNIVERSIDADE FEDERAL DO PARANÁ

INAJARA DA SILVA FREITAS

COMPARAÇÃO E ANÁLISE DE ALGORITMOS PARA O CÁLCULO DE ESTRUTURAS DE PROTEINAS

CURITIBA
2015

INAJARA DA SILVA FREITAS

COMPARAÇÃO E ANÁLISE DE ALGORITMOS PARA O CÁLCULO DE ESTRUTURAS DE PROTEÍNAS

Dissertação apresentada ao Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, do Departamento de Matemática, Setor de Ciências Exatas e do Departamento de Construção Civil, Setor de Tecnologia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Ciências.

Orientador: Prof. Dr. Luiz Carlos Matioli

CURITIBA
2015

F866c Freitas, Inajara da Silva
Comparação e análise de algoritmos para o cálculo de estruturas de proteínas/ Inajara da Silva Freitas. – Curitiba, 2015.
127 f. : il. color. ; 30 cm.

Dissertação - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-graduação em Métodos Numéricos em Engenharia, 2015.

Orientador: Luiz Carlos Matioli .
Bibliografia: p. 126-127.

1. Proteínas - Análise. 2. Geometria molecular. 3. Algoritmo - Otimização. I. Universidade Federal do Paraná. II.Matioli, Luiz Carlos. III. Título.

CDD: 539.6

TERMO DE APROVAÇÃO

INAJARA DA SILVA FREITAS

COMPARAÇÃO E ANÁLISE DE ALGORITMOS PARA O CÁLCULO DE ESTRUTURAS DE PROTEÍNAS

Dissertação aprovada como requisito parcial para obtenção do grau de mestre no Programa de Pós-Graduação em Métodos Numéricos em Engenharia, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof. Dr. Luiz Carlos Matioli
Orientador - Membro do PPGMNE/UFPR.

Prof.ª Dr.ª Diane Rizzotto Rossetto
Membro da UTFPR – Curitiba/PR

Prof.ª Dr.ª Solange Regina dos Santos
Membro da UNESPAR – Campo Mourão/PR

Prof. Dr. Carlos Henrique dos Santos
Membro da UFPR – Curitiba/PR

Curitiba, 29 de maio 2015

Dedico este trabalho a minha família e amigos.

AGRADECIMENTOS

Agradeço a todos que contribuiram direta e indiretamente a este trabalho.

RESUMO

Nesta dissertação serão abordados métodos para resolver o problema de determinar a estrutura de uma proteína quando apenas a distância entre os átomos é conhecida, que também é chamado de problema geométrico de distância molecular. Serão abordadas várias situações diferentes, referentes tanto ao conjunto de distâncias conhecidas, quanto ao erro que pode existir nestas distâncias. Também será mostrado como diferentes implementações do mesmo problema podem resultar em um erro maior ao determinar estas estruturas. A primeira formulação do problema será resolvida de forma linear e também não-linear. Esta variação da linearidade será obtida através de pequenas modificações na organização do problema de distâncias. Também será analisada a importância do número de átomos iniciais necessários para resolver este problema e a disposição de suas coordenadas no espaço para cada implementação. Por último será implementada uma versão para o problema de distâncias que pode ser resolvido com métodos de otimização, cujo método escolhido para solucionar este problema será o de Newton. Todos os algoritmos implementados serão comparados utilizando o Root-Mean-Square-Deviation, que é uma metodologia utilizada para calcular o erro gerado entre a estrutura original já conhecida e a estrutura obtida por cada método. Salientando que serão feitas algumas considerações especiais referentes a comparação de vários algoritmos, ao cálculo do erro gerado para quando as distâncias consideradas no problema molecular não forem exatas e pela implementação de um novo algoritmo iterativo.

Palavras-chave: Implementação, Distância Molecular, Otimização.

ABSTRACT

In this dissertation several methods are addressed to solve the problem of determining the structure of a protein when only the distance between these atoms is known, which can also be defined as the geometric molecular distance problem. It will be addressed different situations related to the set of known distances and also the error on these distances. It will be shown as well as different implementations of the same problem may result in a larger error when determining these structures. The first formulation of the problem will be solved linearly and also non-linear. This variation in linearity can be achieved by small changes in the organization of the problem. It will also be analyzed the importance of the number of initial atoms necessary to solve the problem and the position of its coordinates in space for each implementation. Finally it will be implemented a version for the distance problem that can be solved with optimization methods, the method chosen to solve this problem will be the Newton's. All implemented algorithms will be compared using Root-Mean-Square Deviation, which is a methodology used to calculate the error generated between the original structure already known and the obtained structure by each method. Pointing out that special considerations are made concerning the comparison of several algorithms for calculating the error generated when the distances considered in the molecular problem are not exact and the implementation of a new algorithm.

Key-words: Implementation, Molecular Distance, Optimization

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

RMSD	Root-Mean-Square Deviation
ML	Método Linear
MLA	Método Linear Atualizado
MLRV2	Método Linear Rígido - Versão 2
MLMQL	Método Linear com Mínimos Quadrados Linear
MLMQNL	Método Linear com Mínimos Quadrados Não-Linear
MIE	Método de Intercessão de Esferas

SUMÁRIO

1 INTRODUÇÃO

As proteínas possuem habilidades para desempenhar funções que qualquer organismo vivo execute. Elas são capazes de alterar o sistema biológico, tanto de forma benéfica através de ajustes no organismo, em funções como o crescimento de unhas, cabelos e mantendo o batimento cardíaco, como também prejudicando por meio de vírus ou doenças. Em geral, o funcionamento dessas proteínas depende totalmente do formato que elas possuem, sendo elas capazes de se encaixar umas nas outras ou em diferentes organismos executando papéis que podem ser alterados de acordo com o organismo no qual esta proteína se encaixa.

Para que uma proteína execute essa função de se encaixar em outra proteína, ela precisa ser compatível no encaixe, quase como um quebra cabeça. Portanto, pode-se dizer que é de grande importância conhecer essas estruturas, pois conhecendo o formato dessas proteínas, cientistas são capazes de criar vacinas e tratar ou evitar doenças. Salientando que ao identificar a estrutura da proteína que executa certa função, é possível manipular uma outra estrutura para que copie a mesma funcionalidade ou até mesmo reproduzir a mesma estrutura.

Neste trabalho serão apresentados algoritmos que recaem no caso de determinar a estrutura de uma proteína quando apenas distâncias a uma certa proximidade de cada átomo é conhecida, mas as coordenadas do átomo em si são desconhecidas. Assim, é gerada uma matriz esparsa de distâncias muito pequenas, onde, em vários casos, são conhecidas apenas distâncias de três ou quatro átomos iniciais, para a partir deles determinar cada átomo restante. Este problema também pode ser denominado como problema geométrico de distâncias.

Serão discutidos diferentes algoritmos gerados a partir da manipulação do problema geométrico de distâncias, que poderão resultar na resolução de um simples

sistema linear ou na resolução do problema por mínimos quadrados linear ou não-linear, como também formulando um problema de minimização que possa ser resolvido com o uso de um método iterativo. Para testar estes algoritmos serão utilizadas diferentes condições para as distâncias entre os átomos, considerando inicialmente que todas as distâncias são conhecidas, para então generalizar para casos em que as distâncias são inexatas, ou que poucas distâncias sejam conhecidas.

Todos estes casos serão apresentados e analisados separadamente, para descobrir se existe algum algoritmo que poderia ser considerado superior aos demais, levando em consideração o número de átomos inicialmente conhecidos para cada caso.

Vale salientar que este trabalho apresenta três contribuições especiais, a primeira é em relação a comparação de vários algoritmos, visto que em geral quando um novo algoritmo é implementado, ele é comparado apenas a um algoritmo já existente na literatura. A segunda é relacionada ao cálculo de uma aproximação para o erro gerado pelo Método Linear no caso onde todas as distâncias são conhecidas mas existe erro nessas distâncias. A última contribuição é a apresentação de um novo algoritmo que pode definir a estrutura da proteína utilizando métodos de otimização, sendo que este algoritmo ainda não foi implementado para o caso específico de estruturas de proteínas.

Todos os algoritmos implementados foram executados no MATLAB, versão R2014a e testados em um processador Core I7, com 8 Gigabites de memória. É possível obter qualquer um dos algoritmos implementados neste trabalho via e-mail a autora: *inajarafreitas@gmail.com*.

1.1 ESTRUTURA DO TRABALHO

O capítulo 2 é dedicado a apresentar os conceitos básicos necessários para a compreensão deste trabalho, focando mais profundamente na definição de proteínas e qual a importância de se conhecer estas estruturas, explicando também quais mé-

todos são usados para determiná-las e como fazer o cálculo do erro para que os métodos apresentados nos próximos capítulos possam ser validados.

No capítulo 3 é apresentado um método linear que resolve o problema geométrico de distâncias de forma linear, sendo necessário determinar apenas os quatro átomos iniciais para o seu funcionamento. Os resultados do método serão analisados inicialmente considerando que a matriz formada por todas as distâncias obtidas não é esparsa. Em seguida é feita uma análise dos resultados supondo que existe erro na matriz de distâncias, para então generalizar estes resultados para casos em que a matriz de distâncias é esparsa e também inexata.

No capítulo 4 são apresentados três algoritmos baseados no método linear para o caso em que a matriz de distâncias é esparsa e que, com algumas modificações no algoritmo linear, se tornam mais eficientes. O primeiro algoritmo é o método linear atualizado e nele é possível verificar que a cada iteração do método, os átomos utilizados para determinar o átomo procurado são reinicializados garantindo uma melhor precisão do método. Em seguida são apresentados dois algoritmos denominados método linear rígido e método linear rígido - versão 2. Estes algoritmos tem o diferencial de utilizar apenas três átomos iniciais ao invés de quatro, gerando múltiplas soluções para a estrutura final.

O capítulo 5 apresenta dois algoritmos para o mesmo problema de distâncias, que podem ser resolvidos por mínimos quadrados. O primeiro é chamado de método linear com o uso de mínimos quadrados e o segundo é chamado de método linear com mínimos quadrados não-linear e, como o nome já diz, são resolvidos com diferentes variações de mínimos quadrados.

No capítulo 6 é introduzido um novo algoritmo para o cálculo de estruturas de proteínas que é baseado no método de interseção de esferas. Este algoritmo será transformado em um problema de otimização, que por fim será resolvido pelo método de Newton. Para este algoritmo, só será apresentada a versão onde todas as distâncias são conhecidas.

Por fim, no capítulo 7 é feita a comparação de todos os algoritmos apresentados, considerando diferentes casos para a matriz de distâncias molecular, analisando qual método apresenta melhor desempenho em cada caso.

2 CONCEITOS BÁSICOS

Neste capítulo serão mostrados exemplos da funcionalidade de proteínas nos seres vivos e todas as proteínas usadas poderão ser encontradas em (Berman *et al.* 2000). Em seguida será calculado o erro usado para validar os métodos mostrados nos capítulos posteriores.

2.1 O QUE SÃO PROTEÍNAS?

De acordo com (Berman *et al.* 2000), as proteínas podem assumir diversos papéis na biologia, elas formam todas as estruturas vivas conhecidas, executando papéis essenciais para o funcionamento dos seres vivos. Estas proteínas são formadas por apenas vinte aminoácidos que, combinados das mais diversas maneiras, são capazes de criá-las, os quais são apresentados na Tabela 1.

TABELA 1: Aminoácidos encontrados em organismos vivos

Alanina	Cisteina	Glutamina	Lisina	Tirosina
Arginina	Fenilalanina	Histidina	Metionina	Treonina
Asparagina	Glicina	Isoleucina	Prolina	Triptofano
Aspartato	Glutamato	Leucina	Serina	Valina

Cada aminoácido é formado por átomos de carbono, oxigênio, nitrogênio e hidrogênio, sendo que apenas alguns deles podem também conter átomos de enxofre. Os aminoácidos apresentam características distintas, eles podem ser hidrofílicos que atraem água ou hidrofóbicos que repelem água, também podem ser ionizados apresentando cargas positivas ou negativas. O fato de um aminoácido ser hidrofílico ou hidrofóbico determina o quanto ele irá interagir com outras estruturas, assim como a polaridade os obriga a interagir com aminoácidos carregados opostamente.

Alguns aminoácidos são chamados de aminoácidos essenciais, eles são neces-

sários para a sobrevivência dos seres vivos e não podem ser sintetizados pelo organismo, mas podem ser obtidos a partir de alimentos, a Metionina é um desses aminoácidos, Figura 1. A Metionina não é sintetizada pelo corpo humano e nem por outros animais, porém, é possível encontrá-la em alimentos como ovos, peixes, castanhas e também cereais.

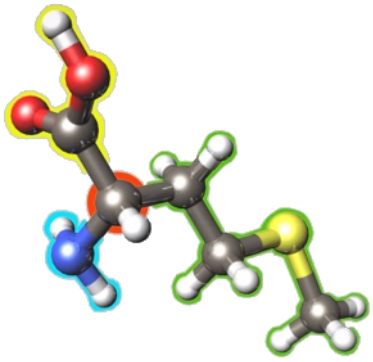
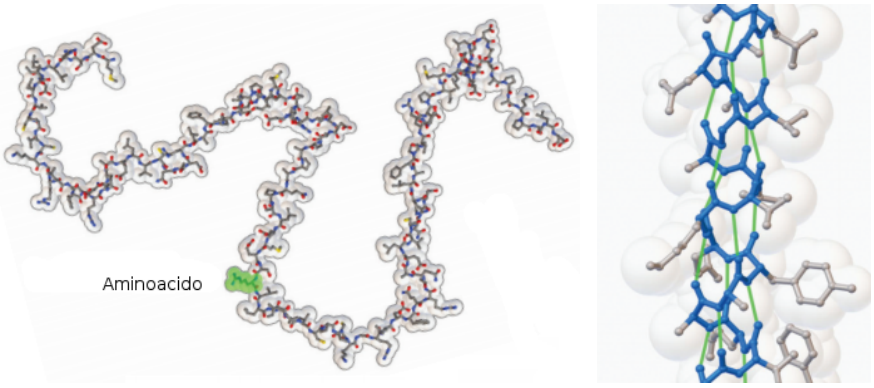


FIGURA 1: Metionina
Fonte: (Berman *et al.* 2000)

Um dos efeitos causados aos seres humanos pela falta da Metionina no organismo é o envelhecimento do cabelo, portanto, ingerindo alimentos que contenham Meotinina é possível recuperar a cor do cabelo.

Os aminoácidos possuem a capacidade de se unir, formando uma cadeia de proteínas que é codificada pela sequência de DNA, Figura 2(a) e a cada união entre aminoácidos uma molécula de água é produzida em torno deles, Figura 2(b).

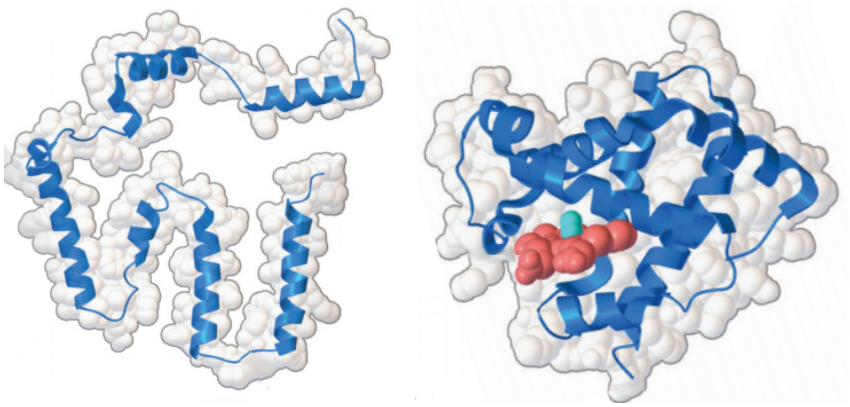


(a) Sequencia de DNA

(b) Aminoácido envolto por água

FIGURA 2: Formação da proteína - 1
Fonte: (Berman *et al.* 2000)

As ligações provocadas entre aminoácidos ocasionam dobras nessa cadeia de proteínas, Figura 3(a), criando a estrutura da proteína. A cadeia de DNA se dobra de uma maneira compacta em torno de uma pequena molécula de oxigênio que contém um átomo de ferro em seu centro, Figura 3(b).



(a) Cadeia de Proteínas

(b) Proteína Estruturada em torno de um átomo de ferro

FIGURA 3: Formação da proteína - 2
Fonte: (Berman *et al.* 2000)

Qualquer mudança diferente da usual executada entre os aminoácidos, pode resultar na junção das proteínas de forma incorreta, ocasionando uma anemia falciforme, ou seja, provocando uma má formação destas proteínas. As proteínas também podem se juntar formando uma única molécula com várias subunidades, Figura 4.

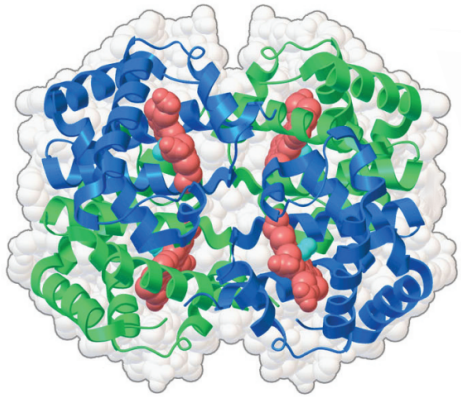


FIGURA 4: Hemoglobina
Fonte: (Berman *et al.* 2000)

Um exemplo deste acontecimento ocorre com a hemoglobina, ela possui quatro subunidades. A hemoglobina executa o importante papel de transportar através dos glóbulos vermelhos oxigênio do pulmão ao sistema circulatório.

2.2 FERRAMENTA BENÉFICA

Nessa seção serão dados exemplos do importante papel que as proteínas assumem em seres vivos e qual é o benefício que se pode obter ao conhecer estas estruturas.

2.2.1 SANGUE ARTIFICIAL

Os problemas encontrados ao armazenar sangue provido de doações é que ele não possui a capacidade de resistir por muito tempo fora do corpo humano e também que muitas vezes não é encontrado o tipo sanguíneo necessário para a transfusão. Por estes fatos é que muitas vezes não é possível encontrar o sangue necessário em bancos de sangue.

Através do conhecimento do funcionamento e da estrutura da hemoglobina, é possível criar um sangue artificial. Este sangue artificial, é uma solução pura de hemoglobina, que pode ser usada em transfusões. No entanto, sem a hemácia para proteger a estrutura da hemoglobina, Figura 5, sua cadeia de aminoácidos rapidamente se desfaz.

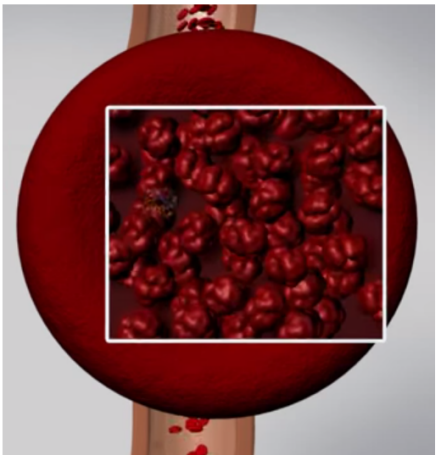


FIGURA 5: Hemoglobinas dentro da Hemácia
Fonte: (Berman *et al.* 2000)

Para evitar que a hemoglobina se desfaça, são criadas novas moléculas em laboratório e, estas novas hemoglobinas são interligadas duas a duas, evitando assim sua

separação.

2.2.2 APLICAÇÕES INDUSTRIAIS

A insulina é uma pequena proteína produzida por células do pâncreas, ela viaja pelo sangue, regulando o nível de glicose. A glicose que é absorvida através da comida pelo corpo humano, viaja na corrente sanguínea até chegar às células do pâncreas. Ao reconhecer a glicose, o pâncreas ativa a liberação de insulina. A insulina viaja pela corrente sanguínea em direção aos músculos, a gordura e ao fígado ativando-os para que armazenem glicose, esta glicose é então usada como fonte de energia, Figura 6.

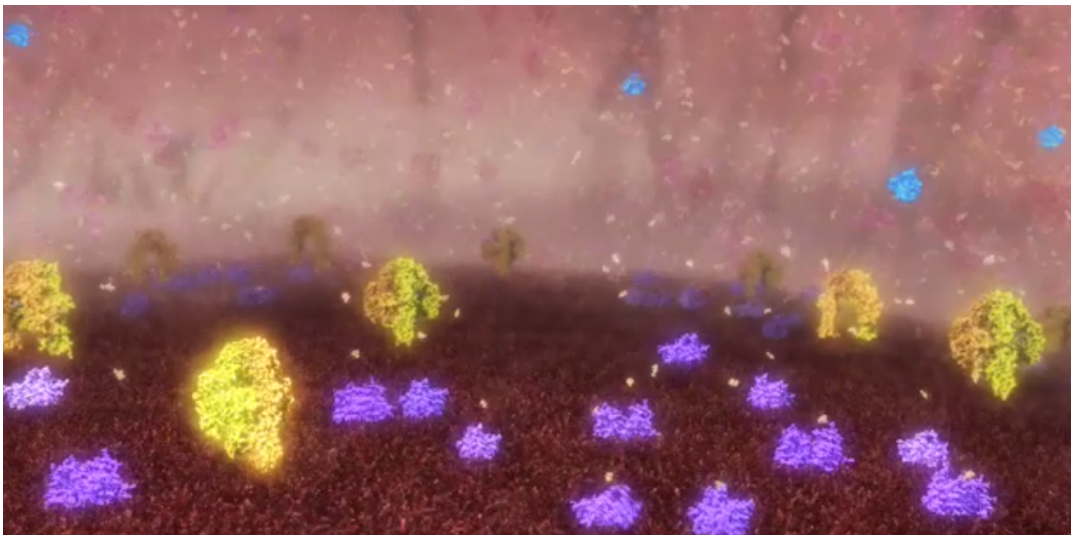


FIGURA 6: Processo de armazenamento da glicose
Fonte: (Berman *et al.* 2000)

Em pessoas com diabetes do tipo 1, o sistema imunológico ataca as células que produzem insulina tornando-as menores, o pâncreas então passa a não produzir insulina o suficiente para regular a glicose no sangue. Esse excesso de glicose acaba prejudicando tecidos de todo o corpo, tais como, a visão, rins, nervos, coração, estomago e a pele.

Uma das maneiras de verificar os níveis de glicose no sangue, para que seja possível controlar o diabetes com injeções de insulina, é através da enzima glicose oxidase, Figura 7.

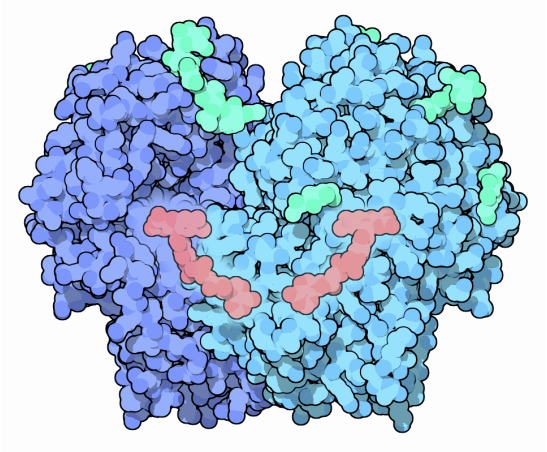


FIGURA 7: Glicose oxidase
Fonte: (Berman *et al.* 2000)

A glicose oxidase transforma a glicose em glucolactona, o qual contém um elemento chamado peróxido de hidrogênio. O peróxido de hidrogênio pode ser medido facilmente, tornando possível mensurar o nível de glicose no sangue. Se este nível de glicose estiver acima do normal, a pessoa deverá tomar insulina para controlar a diabete.

Atualmente existem três maneiras conhecidas de se conseguir insulina, a primeira delas é a insulina produzida pelos porcos, ela difere da insulina humana por apenas um aminoácido possuindo a Treonina ao invés de Alanina, Figura 8.

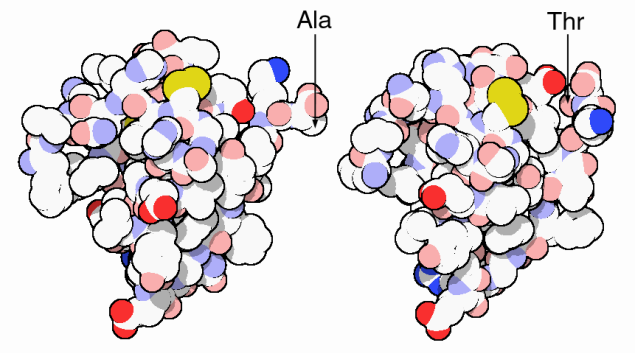


FIGURA 8: Insulina - 4HIN porcos (esq), 2HIN humanos (dir)
Fonte: (Berman *et al.* 2000)

A insulina das vacas também é muito parecida com a humana, difere por três aminoácidos, portanto, ambas são reconhecidas pelo corpo humano e podem ser usadas no tratamento da diabetes tipo 1. Hoje em dia a insulina humana é criada

com o uso de biotecnologia, ela é produzida por bactérias e possui formato idêntico a insulina humana.

2.2.3 REMÉDIOS

Drogas em geral podem ser usadas para ativar ou desativar proteínas, também são capazes de modificar ou melhorar ações que essas proteínas executam no corpo. Elas são capazes de matar bactérias e até mesmo células cancerígenas. A seguir serão descritos dois exemplos dessas drogas.

2.2.3.1 Cicloxigenase

A cicloxigenase, também conhecida como aspirina, é um remédio usado para bloquear sinais de dor enviados ao corpo. Ela é uma das drogas mais utilizadas nos dias de hoje, tem a função de diminuir febres, inflamações e até mesmo prevenir AVC e ataques do coração. Também existem algumas evidências de que a cicloxigenase serve como uma aliada na luta contra o câncer.

A cicloxigenase age bloqueando a produção de prostaglandinas, um hormônio que envia mensagens locais para o corpo, tais como a dor ou ativando o acúmulo de plaquetas durante a coagulação do sangue em machucados, Figura 9.

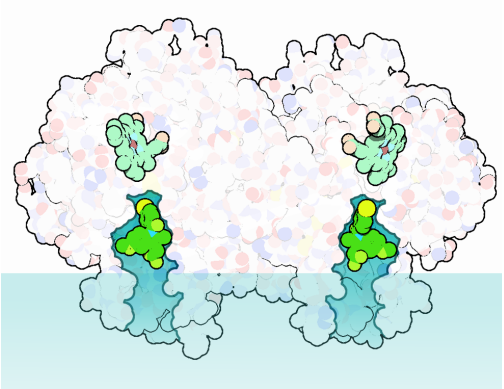


FIGURA 9: Cicloxigenase - Aspirina
Fonte: (Berman *et al.* 2000)

Existem dois tipos de prostaglandinas no corpo, COX-2 - sinaliza dor e inflamação no corpo e COX-1 - envia mensagens de organização por todo o corpo. A cicloxige-

nase bloqueia ambas o que pode causar efeitos colaterais, pois a falta de COX-1 pode causar sangramento do estômago e outros efeitos desconhecidos.

Hoje em dia já existem remédios no mercado que bloqueiam apenas a produção da prostaglandina COX-1.

2.2.3.2 HIV1-Protease

Antibióticos e Antivirais são considerados remédios para o corpo humano, no entanto, agem como veneno para bactérias e vírus. Eles atuam sem envenenar o paciente, atacando proteínas cruciais para a multiplicação e sobrevivência desses organismos, impedindo que doenças se desenvolvam.

A Aids ainda é uma doença sem cura, não existindo uma vacina eficaz para impedir seu desenvolvimento. Hoje em dia, pacientes infectados com HIV são capazes de manter uma vida saudável por vários anos. A saúde destes pacientes é garantida com o uso de drogas efetivas no tratamento da Aids, entre elas está o HIV1-Protease, Figura 10.

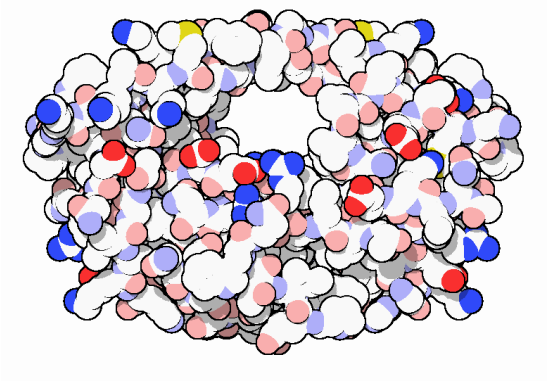


FIGURA 10: HIV1-Protease
Fonte: (Berman *et al.* 2000)

O HIV1-Protease age como um antibiótico contra o vírus HIV, sendo capaz de ligar-se ao vírus ainda em sua fase inicial, enquanto não existem sintomas da doença. Essa ligação impede que o vírus se multiplique, desta maneira o vírus HIV acaba ficando estável apesar de ainda presente no organismo, mas impossibilitado de amadurecer e estabilizando a doença, Figura 11.

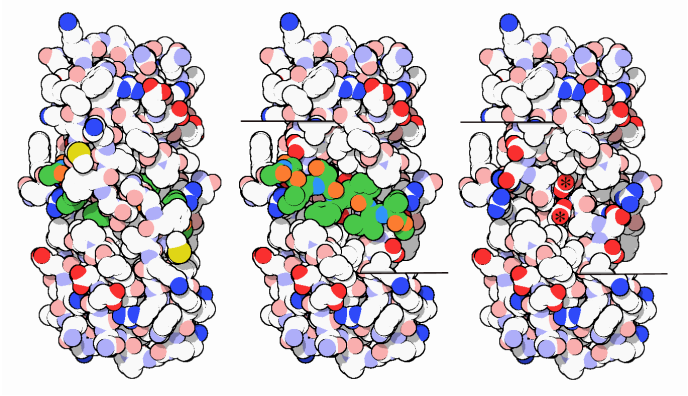


FIGURA 11: conexão do HIV1-Protease ao vírus HIV
Fonte: (Berman *et al.* 2000)

Para que seja possível o tratamento, é importante fazer exames sempre que houver a probabilidade de infecção pelo vírus HIV, para descobrir a doença antes que o corpo apresente sinais e não seja possível tratamento eficaz.

2.2.4 VÍRUS

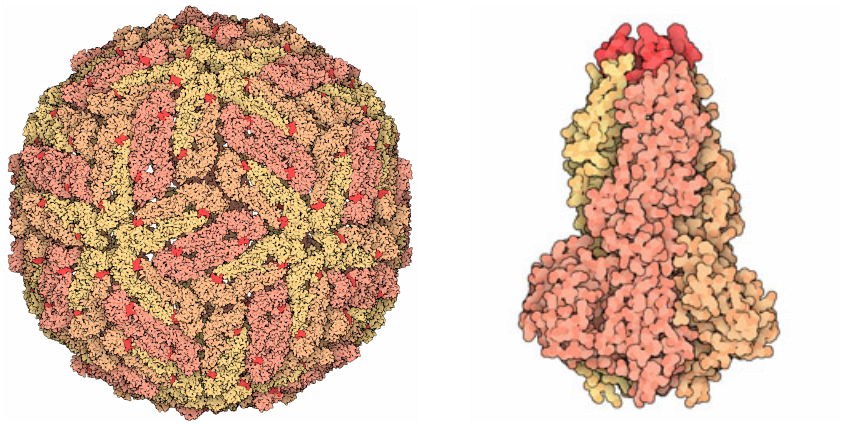
Vírus atacam células forçando-as a reproduzirem vírus, muitas vezes matando as células incubadoras no processo. Eles injetam um genoma virótico na célula, destruindo suas defesas. Os vírus podem ter formatos de um organismo simples ou assumir formas complexas que até hoje não são inteiramente conhecidas.

2.2.4.1 Vírus da Dengue

Criar uma vacina contra o vírus da dengue tem se mostrado difícil. O vírus age de maneira diferente em uma pessoa infectada por um dos quatro subtipos do vírus que entra em contato com um novo subtipo, os anticorpos existentes auxiliam o vírus ao invés de proteger o corpo, provocando a dengue hemorrágica. Portanto uma vacina contra o vírus deve gerar anticorpos contra todos os subtipos simultaneamente.

A fêmea do mosquito *Aedes aegypti* transmite o vírus da dengue, sendo que o mesmo é pequeno e formado por apenas dez moléculas de proteínas. Apenas três destas moléculas formam sua estrutura e as outras sete tem o papel de infectar a célula, Figura 12(a).

Ao invadir a célula, o vírus muda de formato se fundindo a organismos da célula para então liberar seu genoma viral no organismo, Figura 12(b).



(a) Vírus antes de invadir a célula (b) Vírus dentro da célula

FIGURA 12: 1K4R - Vírus da dengue
Fonte: (Berman *et al.* 2000)

2.2.4.2 Vírus Ebola

O vírus ebola se forma a partir da membrana de uma célula roubada, essa célula é invadida por glicoproteínas Figura 13 que contém moléculas do vírus. Estas glicoproteínas se ligam a membrana da célula formando uma cápsula capaz de armazenar várias proteínas, compondo assim a nova estrutura do vírus.

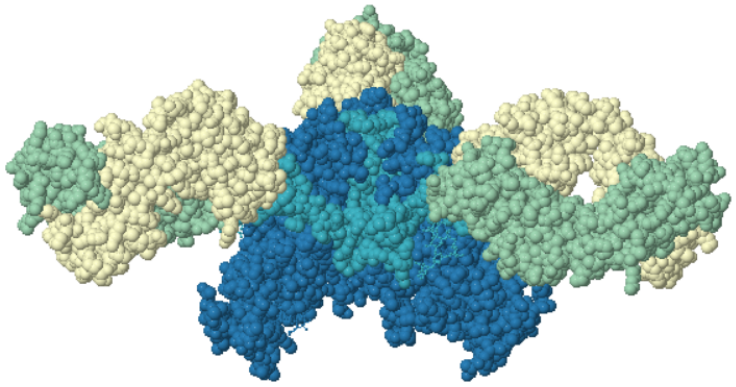


FIGURA 13: Glicoproteína
Fonte: (Berman *et al.* 2000)

A glicoproteína é o principal alvo na busca de uma vacina, ela está na superfície do vírus Figura 14 e é a mais acessível a anticorpos.

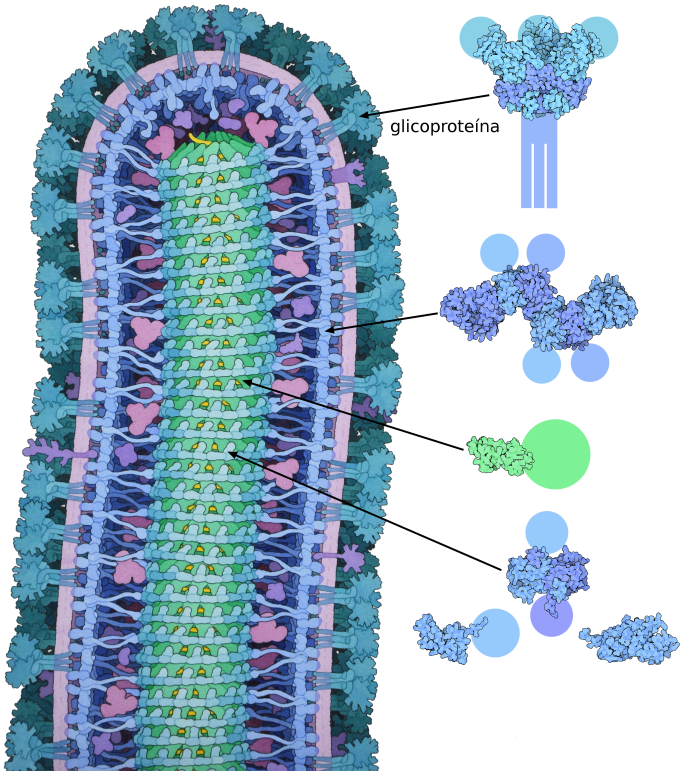


FIGURA 14: Vírus Ebola
Fonte: (Berman *et al.* 2000)

A Figura 13 mostra a glicoproteína com anticorpos de uma pessoa que sobreviveu ao vírus ebola. A parte mais escura no centro é a glicoproteína ligada a três anticorpos. É possível produzir uma vacina capaz de desabilitar o vírus, para isso basta criar um composto contendo a glicoproteína ligada a anticorpos. Ela é capaz de induzir o surgimento de novos anticorpos no organismo, protegendo os pacientes de uma infecção causada pelo vírus.

2.3 TÉCNICAS UTILIZADAS NA DETERMINAÇÃO DE PROTEÍNAS

De acordo com (Berman *et al.* 2000), é possível determinar a estrutura de uma proteína de três diferentes maneiras.

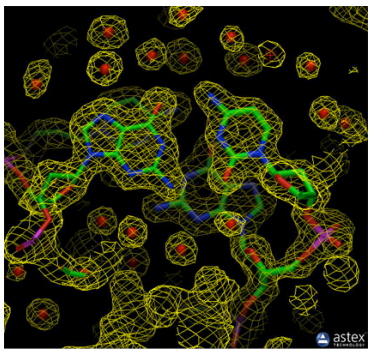
Cristalografia de raios X: A cristalografia de raios X Figura 15(a), é utilizada para determinar a maioria das proteínas conhecidas. Primeiramente a proteína é purificada e cristalizada, em seguida são feitos vários raios-x da proteína em diferentes sentidos formando feixes. Como resultado são obtidas apenas manchas de partes da estrutura,

mas com a ajuda de outros métodos é possível determinar os átomos contidos dentro dessas manchas, sendo assim possível determinar a estrutura da proteína. Esta técnica é útil no caso de proteínas que possuam um formato repetitivo, estrutura rígida e com cristais ordenados.

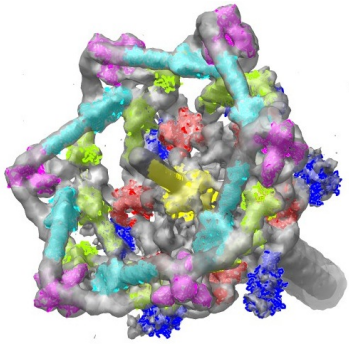
Microscopia Eletrônica: Na Microscopia Eletrônica Figura 15(b), são usados raios de elétrons sobre a proteína para formar sua imagem, no entanto esta imagem é pouco nítida e mostra apenas a aparência visual da molécula. Para obter a estrutura final da proteína são combinadas várias imagens, inclusive aliando-se a outras técnicas para obter resultados mais seguros.

Ressonância Magnética Nuclear (RMN): A proteína é purificada, colocada em um campo magnético e então atingida por várias ondas magnéticas, obtendo assim, uma lista de núcleos atômicos que se encontram próximos uns dos outros, também chamados de distância molecular. A distância molecular é medida em Angstrom ($\mathring{A}$), onde $1\mathring{A} = 10^{-10}$ metros, essa distância pode variar de $0\mathring{A}$ a $100\mathring{A}$ na estrutura da proteína.

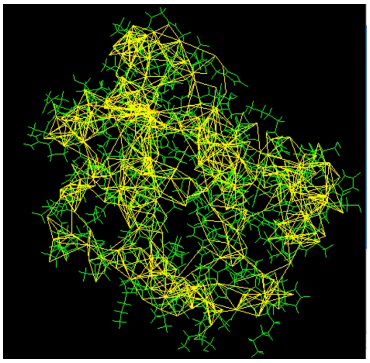
Com o uso da Ressonância Magnética Nuclear Figura 15(c), são obtidas distâncias moleculares em uma proximidade de $5\mathring{A}$, formando então uma matriz esparsa de distâncias moleculares. Devido ao fato do método RMN não ser exato, a matriz de distâncias também apresentará um erro, ou seja, estas distâncias serão valores aproximados dos valores reais, este erro é em torno de $1e-05\mathring{A}$.



(a) Cristalografia de raios X



(b) Microscopia Eletrônica



(c) Ressonância Magnética Nuclear (RMN)

FIGURA 15: Métodos para a determinação de Proteínas
Fonte: (Berman *et al.* 2000)

Encontrar o modelo da molécula através das distâncias geradas pelo RMN, recai em um problema do cálculo de distâncias:

Determinar $X = (x_1, ..., x_n)^T$, para $x_1, ..., x_n$ átomos em $\mathbb{R}^3$, tal que as distâncias entre os átomos i e j são iguais a $d_{i,j}$ e $d_{i,j}$ uma distância molecular conhecida. Matematicamente este problema pode ser escrito como,

$$\|x_i - x_j\| = d_{i,j}, \quad i, j = 1, ..., n, \tag{2.1}$$

X é denota a matriz das coordenadas das proteínas.

A vantagem ao utilizar a técnica (RMN) é que ela torna possível estudar proteínas flexíveis. Neste trabalho serão estudados apenas métodos para calcular essas coordenadas considerando o uso de (RMN), ou seja, serão tratados algoritmos para resolver o problema (2.1).

2.4 ROOT-MEAN-SQUARE DEVIATION (RMSD)

Para validar os métodos futuramente abordados neste trabalho, serão feitos testes com átomos obtidos no *website The Protein Data Bank Nucleic Acids Research* (Berman *et al.* 2000). Este *website* possui um banco de dados de proteínas já conhecidas, sendo possível acessar os dados de cada uma delas através de arquivos de extensão .pdb.

Neste arquivo .pdb estão contidas as coordenadas macromoleculares da estrutura no espaço tridimensional, que formam a matriz de coordenadas X. Para transformar as coordenadas de X em uma matriz de distâncias, calcula-se a distância entre cada átomo x_i e $x_j \in X$ para $i, j = 1, ..., n$.

Seja n o número total de átomos da estrutura, a matriz $D = [d_{ij}]$ será denominada a matriz de distâncias atômicas, onde d_{ij} é a distância entre os átomos x_i e x_j para $i, j = 1, ..., n$, tal que,

$$d_{ij} = ||x_i - x_j||. \tag{2.2}$$

Com o uso da matriz D de distâncias atômicas, é possível calcular as novas coordenadas da proteína. Estas coordenadas serão obtidas através de métodos exibidos posteriormente, para assim comprovar a eficácia dos mesmos. A matriz de coordenadas recalculadas será chamada de Y.

Com as matrizes de coordenadas X e Y Figura 16(a), é possível calcular o erro gerado ao definir a matriz de coordenadas Y. Para definir este erro primeiro é necessário posicionar as duas estruturas no mesmo centro geométrico. Desta maneira, X e Y terão seus centros na origem. De acordo com (Souza 2010), (Silva 2008) e (Davis, Ernst e Wu 2010), é possível transladar as estruturas X e Y para o mesmo centro geométrico calculando a média de todos os átomos contidos na estrutura:

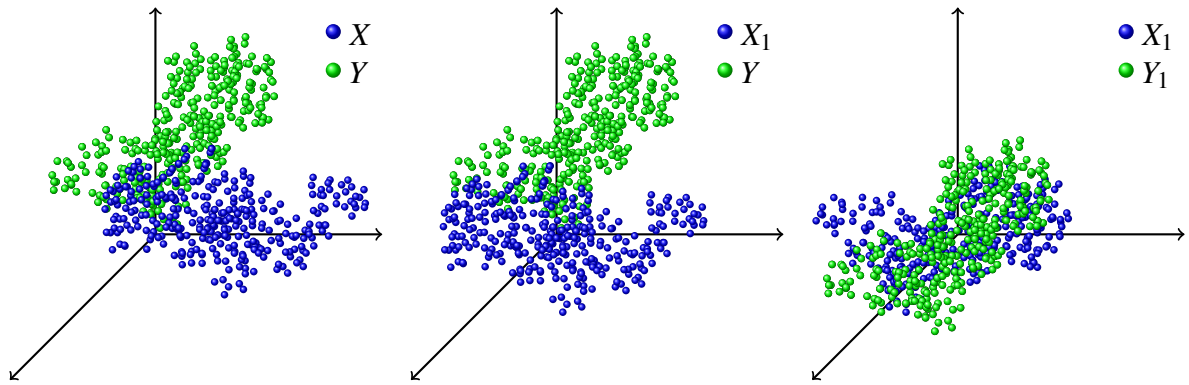
$$xc(j) = \frac{1}{n}\sum_{i=1}^{n} X(i,j), \quad yc(j) = \frac{1}{n}\sum_{i=1}^{n} Y(i,j), \tag{2.3}$$

para $j = 1,2,3$ e n o número de átomos da estrutura.

Os valores xc e yc serão usados para atualizar X e Y Figuras 16(b) e 16(c):

$$\begin{aligned} X_1(i,1) &= X(i,1) - xc(1), & Y_1(i,1) &= Y(i,1) - yc(1), \\ X_1(i,2) &= X(i,2) - xc(2), & Y_1(i,2) &= Y(i,2) - yc(2), \\ X_1(i,3) &= X(i,3) - xc(3), & Y_1(i,3) &= Y(i,3) - yc(3), \end{aligned} \tag{2.4}$$

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



(a) Coordenadas de X e Y (b) Translação de X (c) Translação de Y

FIGURA 16: Translação de X e Y para a origem

O próximo passo é rotacionar Y_1 para que seja possível fazer o cálculo do Root-

Mean-Square Deviation - RMSD. O RMSD é usado para medir o grau de semelhança dessas duas estruturas X_1 e Y_1, equivalentemente calculando qual o erro coordenada a coordenada ao rotacionar $Y1$, para que as estruturas se sobreponham da melhor maneira possível, a fórmula matemática usada é a seguinte:

$$RMSD(X_1, Y_1) = \frac{min_Q \|X_1 - Y_1 Q\|_F}{\sqrt{n}} \tag{2.5}$$

onde $Q \in \mathbb{R}^3$ é uma matriz ortogonal, $QQ^T = I$.

Para calcular este RMSD, é necessário fazer a construção da matriz Q de rotação. Note que,

$$\begin{aligned}
\|X_1 - Y_1 Q\|_F^2 =& Tr((X_1 - Y_1 Q)(X_1 - Y_1 Q)^T), \\
=& Tr(X_1 X_1^T - X_1 (Y_1 Q)^T - Y_1 Q X_1^T + Y_1 Q (Y_1 Q)^T), \\
=& Tr(X_1 X_1^T) - Tr(X_1 (Y_1 Q)^T) - Tr(Y_1 Q X_1^T) + Tr(Y_1 Q Q^T Y_1^T).
\end{aligned}$$

Pelas propriedades de traço de uma matriz, sabe-se que $Tr(A) = Tr(A^T)$, então

$$\begin{aligned}
Tr(X_1 (Y_1 Q)^T) =& Tr((X_1 (Y_1 Q)^T)^T), \\
=& Tr(Y_1 Q X_1^T),
\end{aligned}$$

e como $Tr(AB) = Tr(BA)$,

$$\begin{aligned}
Tr(Y_1 Q X_1^T) =& Tr((Y_1 Q) X_1^T). \\
=& Tr(X_1^T Y_1 Q),
\end{aligned}$$

concluindo que,

$$\begin{aligned}
\|X_1 - Y_1 Q\|_F^2 =& Tr(X_1 X_1^T) - Tr(X_1 (Y_1 Q)^T) - Tr(Y_1 Q X_1^T) + Tr(Y_1 Y_1^T), \\
=& Tr(X_1 X_1^T) - Tr(Y_1 Q X_1^T) - Tr(Y_1 Q X_1^T) + Tr(Y_1 Y_1^T), \\
=& Tr(X_1 X_1^T) - 2Tr(Y_1 Q X_1^T) + Tr(Y_1 Y_1^T), \\
=& Tr(X_1 X_1^T) - 2Tr(X_1^T Y_1 Q) + Tr(Y_1 Y_1^T).
\end{aligned}$$

Portanto, para obter $min_Q \|X_1 - Y_1 Q\|_F$, é necessário maximizar o termo $Tr(X_1^T Y_1 Q)$, onde Q é a variável a ser determinada.

Supondo que $C = X_1^T Y_1$ e usando decomposição por valores singulares (SVD), tem-se que $C = U\Sigma V^T$, portanto

$$Tr(X_1^T Y_1 Q) = Tr(CQ) = Tr(U\Sigma V^T Q) = Tr((U\Sigma)(V^T Q)) = Tr((V^T Q)(U\Sigma)) = Tr((V^T QU)\Sigma),$$

Σ é a matriz dos autovalores de C, ou seja, Σ é uma matriz diagonal com valores reais não negativos portanto, de acordo com (Harvey 2011), vale o seguinte.

Se $\Sigma \succeq 0$ então

$$Tr((V^T QU)\Sigma) \leq ||V^T QU||_2 Tr(\Sigma).$$

Sabe-se também que,

$$(V^T QU)(V^T QU)^T = V^T QUU^T (V^T Q)^T = V^T QQ^T V = V^T V = I,$$

$$||V^T QU||_2 = \max_{||x||\neq 0} \frac{||(V^T QU)x||}{||x||} = \max_{||x||\neq 0} \frac{\sqrt{x^T (V^T QU)^T (V^T QU)x}}{||x||} = \max_{||x||\neq 0} \frac{||x||}{||x||} = 1.$$

Concluindo, $Tr(V^T QU\Sigma) \leq Tr(\Sigma)$, ou seja, $Tr(V^T QU\Sigma)$ é máximo quando $Q = UV^T$, pois neste caso $Tr(V^T QU\Sigma) = Tr(\Sigma)$.

2.4.1 ALGORITMO - ROOT-MEAN-SQUARE DEVIATION (RMSD)

Com estas informações é possível descrever um algoritmo para calcular o erro entre duas estruturas:

Dados X e Y, calcule: $xc,\ yc,\ X1,\ Y1$ como em (2.3) e (2.4)

Faça $C = X_1^T Y_1$

Faça a decomposição em valores singulares de $C = U\Sigma V^T$

Faça $Q = UV^T$

$RMSD(X_1, Y_1) = \frac{\|X_1 - Y_1 Q\|_F}{\sqrt{n}}$

Fim

3 ESTRUTURAS DE PROTEÍNAS

Neste capítulo será abordado um método linear para a solução do problema (2.1), considerando diferentes casos referentes a matriz de distâncias moleculares.

3.1 DISTÂNCIAS EXATAS

De acordo com (Dong e Wu 2002), suponha que todas as distâncias entre um número finito de átomos são conhecidas. Portanto, existe $D = [d_{i,j}]$, onde $d_{i,j}$ é a distância entre os átomos x_i e x_j para $i, j = 1, ..., n$, tal que n é o número total de átomos da estrutura.

Considere primeiramente, o caso particular no plano, onde são conhecidas as coordenadas de apenas três átomos. Por exemplo, x_1, x_2 e x_3 são conhecidos e procura-se determinar as coordenadas de um átomo que ainda não foi determinado, neste caso x_4, conhecendo todas as distâncias entre estes átomos, conforme ilustrado na Figura 17.

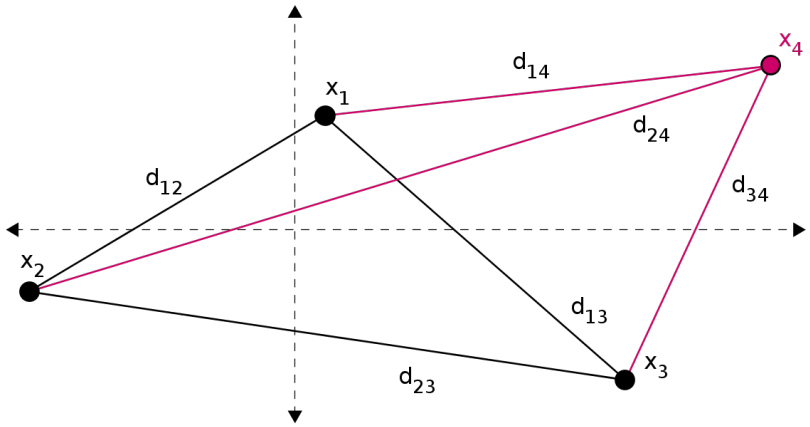


FIGURA 17: Exemplo 2D: Geração do 4^o átomo

De acordo com a Figura 17, as distâncias d_{12}, d_{13}, d_{14}, d_{23}, d_{24} e d_{34} são conhecidas; os átomos x_1, x_2 e x_3 estão fixados e x_4 é o átomo que pode ser determinado de

forma única. Este processo pode ser repetido para determinar os outros átomos da estrutura.

Agora, considere conhecidas as posições de quatro átomos no $\mathbb{R}^3$ e suponha que os mesmos não sejam coplanares. Se forem conhecidas as distâncias destes átomos em relação aos demais que ainda não foram determinados, será possível fixar as coordenadas destes átomos não determinados, Figura 18.

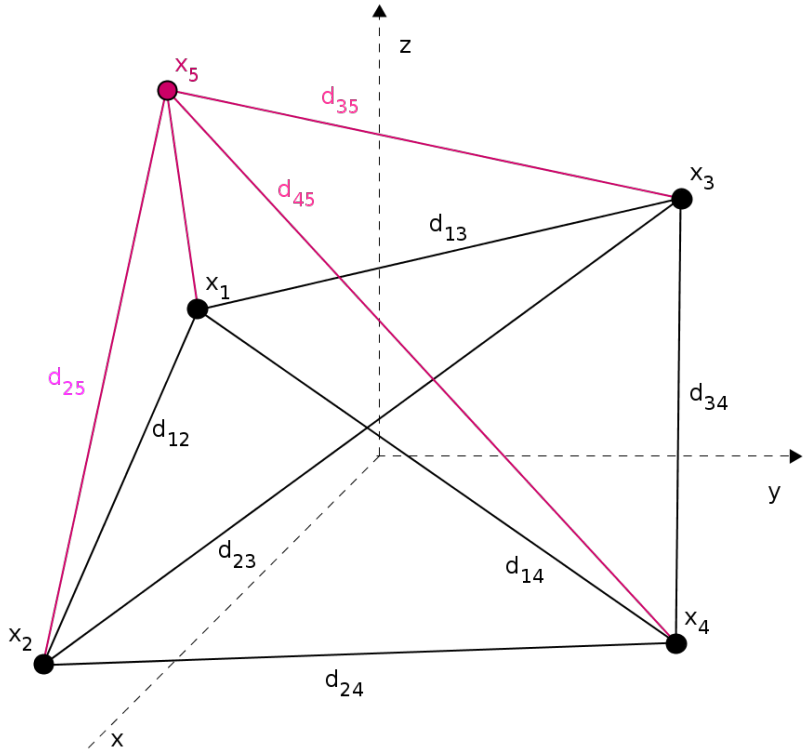


FIGURA 18: Exemplo 3D: geração 5^o átomo

De acordo com a Figura 18, as distâncias d_{12}, d_{13}, d_{14}, d_{15}, d_{23}, d_{24}, d_{25}, d_{34}, d_{35} e d_{45} são conhecidas. Os átomos x_1, x_2, x_3 e x_4 estão fixados e x_5 é o átomo encontrado de forma única.

3.1.1 MÉTODO LINEAR

Para determinar as coordenadas dos demais átomos da estrutura de uma proteína no espaço tridimensional, é necessário encontrar as coordenadas de cada átomo x_i, para $i = 5, ..., n$. Primeiro será necessário fixar as coordenadas de quatro átomos iniciais no espaço tridimensional. A partir desses quatro átomos é possível determinar

as coordenadas dos demais átomos. Suponha que as coordenadas dos átomos x_j, $j = 1,...,4$ são conhecidas, ou seja:

$$x_1 = (u_1, v_1, w_1)^T, \qquad x_3 = (u_3, v_3, w_3)^T,$$

$$x_2 = (u_2, v_2, w_2)^T, \qquad x_4 = (u_4, v_4, w_4)^T.$$

Para determinar as coordenadas dos átomos $x_i = (u_i, v_i, w_i)^T$ onde $i = 5,...,n$ e supondo que as distâncias de todas as possíveis combinações de pares de átomos, ou seja, $d_{i,j}$ para $j = 1,...,4$ são conhecidas. O que torna possível obter o seguinte sistema:

$$||x_i - x_1|| = d_{i,1}, \qquad ||x_i - x_3|| = d_{i,3},$$

$$||x_i - x_2|| = d_{i,2}, \qquad ||x_i - x_4|| = d_{i,4}.$$

Elevando ao quadrado dos dois lados de cada igualdade:

$$||x_i||^2 - 2x_i^T x_1 + ||x_1||^2 = d_{i,1}^2,$$

$$||x_i||^2 - 2x_i^T x_2 + ||x_2||^2 = d_{i,2}^2,$$

$$||x_i||^2 - 2x_i^T x_3 + ||x_3||^2 = d_{i,3}^2,$$

$$||x_i||^2 - 2x_i^T x_4 + ||x_4||^2 = d_{i,4}^2.$$

Substituindo as coordenadas de x_i e x_j por $x_i = (u_i, v_i, w_i)^T$ e $x_j = (u_i, v_j, w_j)^T$ em $x_i^T x_j$, para $j = 1,...,4$:

$$||x_i||^2 - 2u_i u_1 - 2v_i v_1 - 2w_i w_1 + ||x_1||^2 = d_{i,1}^2,$$

$$||x_i||^2 - 2u_i u_2 - 2v_i v_2 - 2w_i w_2 + ||x_2||^2 = d_{i,2}^2,$$

$$||x_i||^2 - 2u_i u_3 - 2v_i v_3 - 2w_i w_3 + ||x_3||^2 = d_{i,3}^2,$$

$$||x_i||^2 - 2u_i u_4 - 2v_i v_4 - 2w_i w_4 + ||x_4||^2 = d_{i,4}^2.$$

Subtraindo a primeira equação das demais:

$$2u_i(u_1 - u_2) + 2v_i(v_1 - v_2) + 2w_i(w_1 - w_2) = (||x_1||^2 - ||x_2||^2) - (d_{i,1}^2 - d_{i,2}^2),$$

$$2u_i(u_1-u_3)+2v_i(v_1-v_3)+2w_i(w_1-w_3)=(||x_1||^2-||x_3||^2)-(d_{i,1}^2-d_{i,3}^2),$$

$$2u_i(u_1-u_4)+2v_i(v_1-v_4)+2w_i(w_1-w_4)=(||x_1||^2-||x_4||^2)-(d_{i,1}^2-d_{i,4}^2),$$

O sistema acima pode ser representado da seguinte forma:

Seja $x_i=(u_i,v_i,w_i)^T$, para $i=5,...,n$, onde cada i assume a posição de um átomo ainda não determinado na estrutura da molécula.

$$Ax_i=b_i, \tag{3.1}$$

onde

$$A=2\begin{pmatrix}(u_1-u_2) & (v_1-v_2) & (w_1-w_2)\\ (u_1-u_3) & (v_1-v_3) & (w_1-w_3)\\ (u_1-u_4) & (v_1-v_4) & (w_1-w_4)\end{pmatrix}$$

e

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

Assim, a partir de quatro átomos fixados é possível determinar toda a estrutura da molécula.

3.1.2 PONTOS INICIAIS

Para que seja possível utilizar o método linear apresentado na seção anterior, é necessário conhecer quatro átomos iniciais. No entanto, a única informação conhecida do problema original gerado pelo RMN é a matriz de distâncias $D=[d_{i,j}]$, onde $i,j=1,...,n$ e os átomos iniciais precisam ser gerados a partir desta matriz.

O átomo $x_1=(u_1,v_1,w_1)$ será fixado na origem, ou seja, $u_1=0$, $v_1=0$ e $w_1=0$. O átomo $x_2=(u_2,v_2,w_2)$ será fixado em um dos eixos, o eixo x será escolhido, portanto $u_2=d_{2,1}$, $v_2=0$ e $w_2=0$, onde $d_{2,1}$ é a distância entre os átomos x_1 e x_2. O átomo

$x_3 = (u_3, v_3, w_3)$ é fixado de modo que não esteja na mesma reta de x_1 e x_2. Para que isto aconteça basta definir $u_3 \neq 0$, $v_3 \neq 0$ e $w_3 = 0$.

As coordenadas u_3 e v_3 podem ser encontradas fazendo uma relação entre as distâncias $d_{3,1}$ e $d_{3,2}$.

Distância $d_{3,1}$:

$$\begin{aligned} d_{3,1}^2 &= ||x_3 - x_1||^2, \\ d_{3,1}^2 &= ||x_3||^2, \\ d_{3,1}^2 &= u_3^2 + v_3^2 + w_3^2, \\ d_{3,1}^2 &= u_3^2 + v_3^2. \end{aligned}$$

Distância $d_{3,2}$:

$$\begin{aligned} d_{3,2}^2 &= ||x_3 - x_2||^2, \\ d_{3,2}^2 &= ||(u_3 - u_2), (v_3 - v_2), (w_3 - w_2)||^2, \\ d_{3,2}^2 &= (u_3 - u_2)^2 + (v_3 - v_2)^2 + (w_3 - w_2)^2, \\ d_{3,2}^2 &= (u_3 - u_2)^2 + v_3^2. \end{aligned}$$

Desta maneira, tem-se o seguinte sistema:

$$d_{3,1}^2 = u_3^2 + v_3^2, \tag{3.2}$$

$$d_{3,2}^2 = (u_3 - u_2)^2 + v_3^2. \tag{3.3}$$

Isolando v_3 na equação (3.2):

$$v_3 =_{-}^{+} (d_{3,1}^2 - u_3^2)^{1/2}. \tag{3.4}$$

Para determinar a coordenada u_3, basta substituir (3.4) na equação (3.3)

$$\begin{aligned} d_{3,2}^2 &= (u_3 - u_2)^2 + ((d_{3,1}^2 - u_3^2)^{1/2})^2, \\ d_{3,2}^2 - d_{3,1}^2 &= u_3^2 - 2u_3u_2 + u_2^2 - u_3^2, \\ 2u_3u_2 &= d_{3,1}^2 - d_{3,2}^2 + u_2^2, \\ u_3 &= (d_{3,1}^2 - d_{3,2}^2)/(2u_2) + u_2/2. \end{aligned}$$

Pode-se concluir que x_3 possui as seguintes coordenadas:

$$\begin{aligned} u_3 &= (d_{3,1}^2 - d_{3,2}^2)/(2u_2) + u_2/2, \\ v_3 &= \pm (d_{3,1}^2 - u_3^2)^{1/2}, \\ w_3 &= 0. \end{aligned}$$

É possível notar que a coordenada v_3 pode ser positiva ou negativa. A diferença de sinal pode resultar em um átomo simétrico ao procurado. Ao escolher a coordenada positiva o átomo será simétrico ao átomo encontrado utilizando a coordenada negativa, ou seja, resultaria em uma estrutura espelhada, Figura 19.

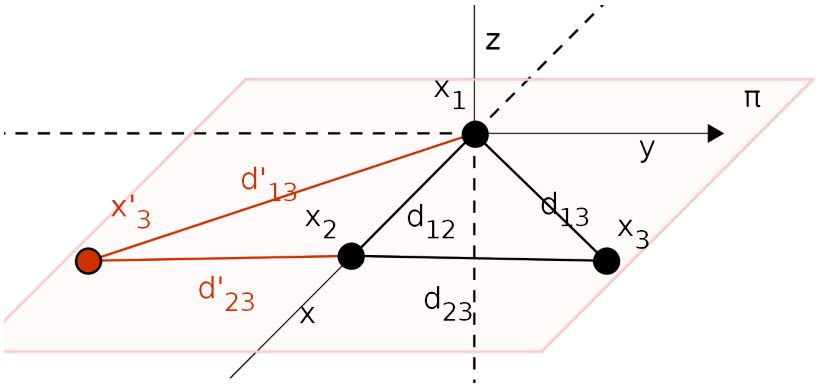


FIGURA 19: Átomo x_3 para $\pm v_3$

Na Figura 19, seja Π o plano $(x, y, 0)$, onde $d_{1,3} = d'_{1,3}$ e $d_{2,3} = d'_{2,3}$, existem dois átomos que satisfazem a condição $||x_3 - x_j|| = d_{j,3}$ para $j = 1, 2$, ou seja, o átomo x'_3 é simétrico a x_3.

Pode-se determinar a posição do átomo $x_4 = (u_4, v_4, w_4)$, onde u_4, v_4 e $w_4 \neq 0$, através do mesmo cálculo usado para x_3. Desta maneira x_4 não estará no plano formado pelos três primeiros átomos.

Distância $d_{4,1}$:

$$
\begin{aligned}
d_{4,1}^2 &= ||x_4 - x_1||^2, \\
d_{4,1}^2 &= ||(u_4 - u_1), (v_4 - v_1), (w_4 - w_1)||^2, \\
d_{4,1}^2 &= (u_4 - u_1)^2 + (v_4 - v_1)^2 + (w_4 - w_1)^2, \\
d_{4,1}^2 &= u_4^2 + v_4^2 + w_4^2.
\end{aligned}
$$

Distância $d_{4,2}$:

$$
\begin{aligned}
d_{4,2}^2 &= ||x_4 - x_2||^2, \\
d_{4,2}^2 &= ||(u_4 - u_2), (v_4 - v_2), (w_4 - w_2)||^2, \\
d_{4,2}^2 &= (u_4 - u_2)^2 + (v_4 - v_2)^2 + (w_4 - w_2)^2, \\
d_{4,2}^2 &= (u_4 - u_2)^2 + v_4^2 + w_4^2.
\end{aligned}
$$

Distância $d_{4,3}$:

$$
\begin{aligned}
d_{4,3}^2 &= ||x_4 - x_3||^2, \\
d_{4,3}^2 &= ||(u_4 - u_3), (v_4 - v_3), (w_4 - w_3)||^2, \\
d_{4,3}^2 &= (u_4 - u_3)^2 + (v_4 - v_3)^2 + (w_4 - w_3)^2, \\
d_{4,3}^2 &= (u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2.
\end{aligned}
$$

Com esses dados, pode-se obter o seguinte sistema:

$$d_{4,1}^2 = u_4^2 + v_4^2 + w_4^2, \tag{3.5}$$

$$d_{4,2}^2 = (u_4 - u_2)^2 + v_4^2 + w_4^2, \tag{3.6}$$

$$d_{4,3}^2 = (u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2. \tag{3.7}$$

Para encontrar u_4 basta subtrair (3.6) de (3.5) e isolar u_4

$$
\begin{aligned}
d_{4,1}^2 - d_{4,2}^2 &= u_4^2 + v_4^2 + w_4^2 - (u_4 - u_2)^2 - v_4^2 - w_4^2, \\
d_{4,1}^2 - d_{4,2}^2 &= u_4^2 - u_4^2 + 2u_4u_2 - u_2^2, \\
u_4 &= \frac{d_{4,1}^2 - d_{4,2}^2}{2u_2} + \frac{u_2}{2}.
\end{aligned}
$$

Para encontrar v_4 basta subtrair (3.7) de (3.6) e isolar v_4

$$\begin{aligned}
d_{4,2}^2 - d_{4,2}^2 &= (u_4-u_2)^2 + v_4^2 + w_4^2 - (u_4-u_3)^2 - (v_4-v_3)^2 - w_4^2, \\
d_{4,2}^2 - d_{4,2}^2 &= (u_4-u_2)^2 + v_4^2 - (u_4-u_3)^2 - v_4^2 + 2v_4v_3 - v_3^2, \\
d_{4,2}^2 - d_{4,2}^2 &= (u_4-u_2)^2 - (u_4-u_3)^2 + 2v_4v_3 - v_3^2, \\
v_4 &= \frac{d_{4,2}^2 - d_{4,3}^2 - (u_4-u_2)^2 + (u_4-u_3)^2}{2v_3} + \frac{v_3}{2}.
\end{aligned}$$

Em seguida, basta isolar w_4 na equação (3.5):

$$w_4 = \pm (d_{4,1}^2 - u_4^2 - v_4^2)^{1/2}. \tag{3.8}$$

É possível encontrar as seguintes coordenadas para x_4

$$\begin{aligned}
u_4 &= \frac{d_{4,1}^2 - d_{4,2}^2}{2u_2} + \frac{u_2}{2}, \\
v_4 &= \frac{d_{4,2}^2 - d_{4,3}^2 - (u_4-u_2)^2 + (u_4-u_3)^2}{2v_3} + \frac{v_3}{2}, \\
w_4 &= \pm (d_{4,1}^2 - u_4^2 - v_4^2)^{1/2}.
\end{aligned} \tag{3.9}$$

Novamente pode-se notar que a coordenada w_4 pode ser positiva ou negativa e a diferença entre escolher o sinal positivo ou negativo resultaria em uma estrutura espelhada. É possível ver um exemplo na Figura 20.

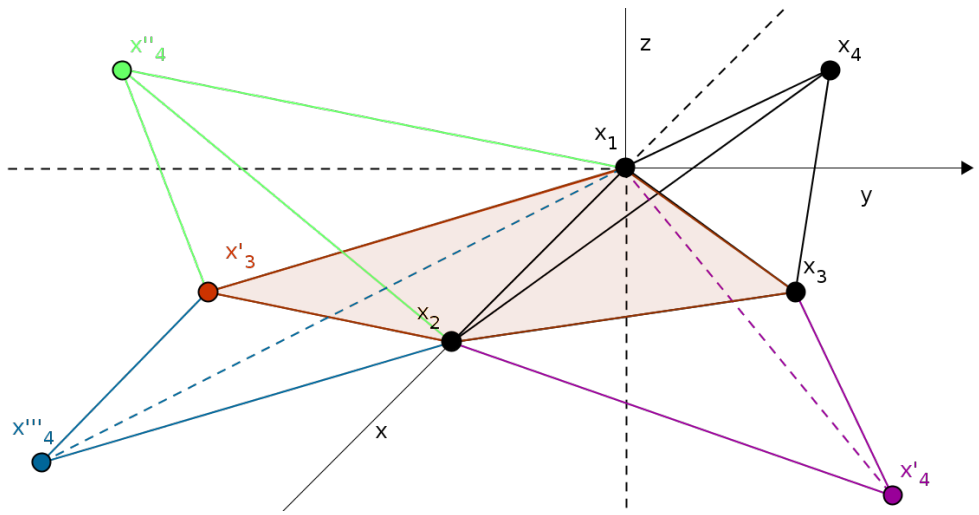


FIGURA 20: Átomo x_3 para $\pm v_3$ e átomo x_4 para $\pm w_4$

Na Figura 20, pode-se notar que definir v_3 e w_4 como positivos ou negativos resul-

tam em quatro pontos diferentes. Esse processo acumulado para determinada opção de sinais durante varias iterações provoca o espelhamento da figura:

- x_4 é obtido com v_3 positivo e w_4 positivo.
- x'_4 é obtido com v_3 positivo e w_4 negativo.
- x''_4 é obtido com v_3 negativo e w_4 positivo.
- x'''_4 é obtido com v_3 negativo e w_4 negativo.

Abaixo é possível notar o que acontece com a estrutura ao modificar estes sinais.

FIGURA 21: 1FW5

(a) 1FW5: v_3 positivo e w_4 positivo

(b) 1FW5: v_3 positivo e w_4 negativo

(c) 1FW5: v_3 negativo e w_4 positivo

(d) 1FW5: v_3 negativo e w_4 negativo

FIGURA 22: Variação dos sinais de v_3 e w_4

O exemplo mostra a mudança causada na proteína 1FW5 ao trocar os sinais de x_3 e x_4. Note que as Figuras 22(a) e 22(d) estão espelhadas em relação a estrutura original Figura 21, enquanto as Figuras 22(b) e 22(c) estão praticamente idênticas a estrutura original, Figura 21.

Em cada estrutura a determinação dos átomos para que fiquem ou não espelhados em relação a figura original, acontecem de forma aleatória. Por exemplo, ao determinar v_3 positivo e w_4 positivo, para o caso da Figura 21, foi gerada uma estrutura espelhada em relação a estrutura original da Figura 21.

Utilizando esta regra para os sinais originará um resultado diferente para outras estruturas, ou seja, não existe uma definição padrão das coordenadas v_3 e w_4 que origine a estrutura correta em todos os casos. Portanto, para que seja mais fácil utilizar o método, v_3 e w_4 serão positivos para todos os casos.

As coordenadas dos quatro primeiros átomos foram determinadas, com isso todos os coeficientes em (3.1) são conhecidos, de forma que A é uma matriz triangular inferior. Utilizando estes quatro átomos juntamente com o método definido na seção anterior é possível calcular os átomos restantes, ou seja, x_i para $i = 5, ..., n$.

3.1.3 ALGORITMO - MÉTODO LINEAR

Utilizando as informações apresentadas até o momento, é possível descrever o algoritmo para encontrar os pontos x_i, para $i = 5, ..., n$.

Este algoritmo será denominado "Método linear- ML".

Dados de entrada: $D = [d_{i,j}]$, para $i, j = 1, ..., n$

Calcule A, como em (3.1)

Para $i = 5, ..., n$

Calcule b_i, como em (3.1)

Resolva $Ax_i = b_i$

fim

O ponto x_i é calculado da seguinte forma:

$$x_i = \begin{pmatrix} \frac{b_{i,1}}{A_{1,1}} \\ \frac{b_{i,2} - A_{2,1}x_i(1)}{A_{2,2}} \\ \frac{b_{i,3} - A_{3,1}x_i(1) - A_{3,2}x_i(2)}{A_{3,3}} \end{pmatrix}. \tag{3.10}$$

3.1.4 RESULTADOS COMPUTACIONAIS - CASO EXATO

O algoritmo foi implementado em MATLAB, abaixo seguem os resultados dos testes do método linear para algumas proteínas:

Para cada proteína são mostrados os seguintes fatores: Proteína - Nome da proteína que consta no banco de dados, Átomos - número de átomos contidos na proteína, Tempo/s - Tempo em segundos que o método linear demorou para determinar todos os átomos, RMSD - Erro acumulado durante o processo calculado utilizando Root Mean Square Deviation (RMSD).

TABELA 2: Método Linear - Exato

Proteína	Átomos	Tempo/s	RMSD
103D	772	0.00094	1.87250e-13
104D	766	0.00096	7.64665e-14
124D	508	0.00076	6.51431e-14
132D	750	0.00093	2.40680e-13
141D	527	0.00071	1.34497e-13
1A1D	146	0.00046	3.75974e-14
1A23	2952	0.00275	3.22424e-13
1A84	758	0.00100	1.93925e-13
1AIK	729	0.00092	2.19168e-13
1AMB	438	0.00063	1.31720e-13
1AMD	380	0.00058	6.78567e-14
1AQR	524	0.00072	8.03857e-14
1AX8	1003	0.00119	1.79517e-13
1B5N	332	0.00055	5.50003e-14
1BOM	700	0.00086	6.12665e-14
1BQX	1166	0.00127	1.10360e-13
1CEU	854	0.00101	2.27163e-13

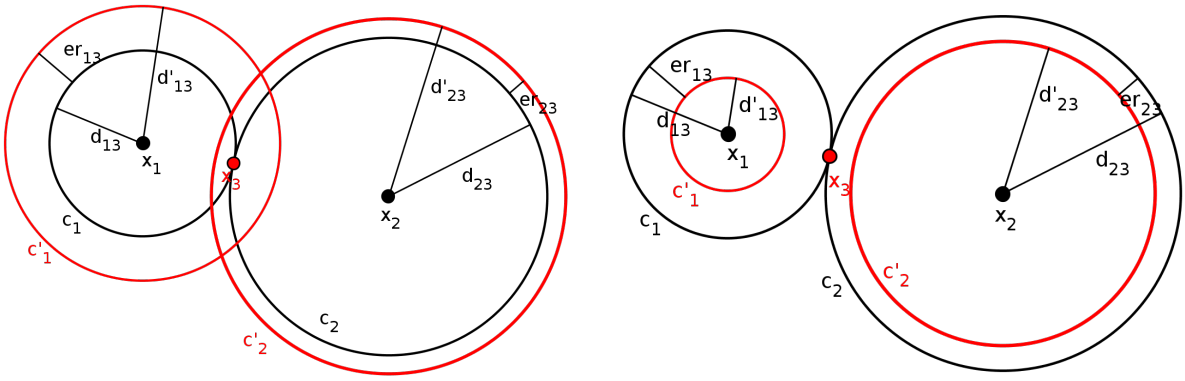
Proteína	Átomos	Tempo/s	RMSD
1D8V	4208	0.00380	2.38325e-13
1DKE	4380	0.35354	6.81280e-13
1F39	1534	0.00157	1.96630e-13
1FS3	951	0.00118	1.87145e-13
1FW5	332	0.00051	7.52774e-14
1HAA	1310	0.00142	6.64178e-12
1HIP	617	0.00086	1.44031e-13
1HLL	540	0.00072	3.24263e-13
1HMV	29592	0.50132	3.54499e-12
1HOE	558	0.01112	7.94426e-14
1HSM	1251	0.00137	1.74248e-13
1ID7	189	0.00039	1.29096e-14
1ITH	2126	0.00206	3.86159e-13
1JAV	360	0.00057	3.62559e-13
1JK2	1229	0.00138	1.40918e-13
1KVX	954	0.00107	2.17865e-13
1LFB	641	0.00084	4.44424e-13
1M40	5712	0.00504	6.42926e-12
1MBN	1216	0.00136	1.40550e-13
1MEQ	405	0.00064	2.28152e-14
1MQQ	5681	0.00489	5.96916e-13
1N4W	8616	0.00738	9.20071e-13
1PHT	811	0.00099	1.48231e-13
1POA	914	0.00107	1.35637e-13
1PTQ	402	0.00057	1.04122e-13
1QS5	1295	0.00137	2.58632e-13
1QSB	1293	0.00137	2.32342e-13
1R7C	532	0.00074	1.70274e-12
1RGS	2015	0.00202	2.81481e-13
1RWH	5646	0.00505	7.09502e-13
1SOL	353	0.00061	4.79160e-14
1ULR	677	0.00085	5.45141e-12
1VII	596	0.00078	5.83050e-14
1VMP	1166	0.00128	2.62561e-13
2CLJ	4189	0.00374	1.05899e-11
2E7Z	7633	0.00653	4.16725e-13
2EQL	1023	0.00129	1.00295e-12
2MSJ	480	0.00072	8.89519e-14
304D	159	0.00033	2.43078e-13
3B34	7479	0.00643	3.58936e-13
4MBA	1083	0.00124	2.72505e-11
8DRH	329	0.00053	2.71267e-13

Analisando os resultados na tabela 2 pode-se constatar que o método linear é realmente muito rápido no caso de distâncias exatas. Também é possível notar que o erro gerado durante este processo foi muito pequeno. Isto acontece devido ao fato de x_i possuir uma fórmula exata para ser calculado. Em grande parte este erro foi gerado pelo erro de máquina e pelos átomos iniciais predeterminados

Vamos verificar o que acontece quando essas distâncias não são exatas.

3.2 DISTÂNCIAS INEXATAS

Como foi discutido anteriormente, o processo de (RMN) Ressonância Magnética Nuclear de proteínas não é exato. As distâncias obtidas por esse processo contém erros e estes erros podem dificultar a determinação desses átomos. Na Figura 23 a seguir, é possível visualizar como seriam essas distâncias obtidas pelo processo de (RMN).



(a) caso 1: $er > 0$ (b) caso 2: $er < 0$

FIGURA 23: Variação do erro causado pelo (RMN)

A Figura 23 mostra as seguintes circunferências, c_1 de raio $||x_3 - x_1|| = d_{1,3}$ e c_2 de raio $||x_3 - x_2|| = d_{2,3}$. Nos dois casos x_3 é o átomo procurado. Devido ao erro gerado pelo (RMN) c_1' e c_2' serão as novas circunferências geradas com o erro, $d'_{1,3} = d_{1,3} + er_{1,3}$, $d'_{2,3} = d_{2,3} + er_{2,3}$ seus respectivos raios e $er_{1,3}$ e $er_{2,3}$ representam os erros gerados pelo RMN, este erro pode ser positivo Figura 23(a) ou negativo Figura 23(b).

Este problema pode ser generalizado da seguinte maneira:

Dada a esfera de raio x_j e raio $d'_{i,j}$, para $j = 1,...,4$ e $i = 5,...,n$, tal que $||x_i - x_j|| = d'_{i,j}$ e $d'_{i,j} = d_{i,j} + er_{i,j}$, para $er_{i,j} \in \mathbb{R}$.

3.2.1 APROXIMAÇÃO DO RMSD - MÉTODO LINEAR PARA DISTÂNCIAS INEXATAS

De acordo com a seção anterior, para o problema exato vale a seguinte definição:

$$Ax_i = b_i, \tag{3.11}$$

para $i = 5,...,n$, onde

$$A = 2\begin{pmatrix} (u_1 - u_2) & (v_1 - v_2) & (w_1 - w_2) \\ (u_1 - u_3) & (v_1 - v_3) & (w_1 - w_3) \\ (u_1 - u_4) & (v_1 - v_4) & (w_1 - w_4) \end{pmatrix}$$

e

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

Considere o problema inexato da seguinte maneira:

$$A\tilde{x}_i = b'_i, \tag{3.12}$$

com $\tilde{x}_i \neq x_i$, $i = 5,...,n$ onde $b'_i = b_i + er_i$.

Observando que b'_i é calculado utilizando a matriz de distâncias com um certo erro er_i gerado pelo processo (RMN). Desta forma, o objetivo é estimar o erro causado ao utilizar o método linear para resolver um problema inexato.

Considerando

$$b'_i = \begin{pmatrix} (||x_1||^2 - ||x_2||^2) - (d_{i,1}'^2 - d_{i,2}'^2) \\ (||x_1||^2 - ||x_3||^2) - (d_{i,1}'^2 - d_{i,3}'^2) \\ (||x_1||^2 - ||x_4||^2) - (d_{i,1}'^2 - d_{i,4}'^2) \end{pmatrix},$$

onde $d'_{i,j}$ é uma distância inexata, ou seja, $d'_{i,j} = d_{i,j} + er_{i,j}$, ver Figura 23, onde $j = 1,...,4$ e $i = 5,...,n$.

O próximo passo é substituir $d'_{i,j}$ em b'_i:

$$b'_i = \begin{pmatrix} (||x_1||^2 - ||x_2||^2) - ((d_{i,1} + er_{i,1})^2 - (d_{i,2} + er_{i,2})^2) \\ (||x_1||^2 - ||x_3||^2) - ((d_{i,1} + er_{i,1})^2 - (d_{i,3} + er_{i,3})^2) \\ (||x_1||^2 - ||x_4||^2) - ((d_{i,1} + er_{i,1})^2 - (d_{i,4} + er_{i,4})^2) \end{pmatrix},$$

ou

$$b'_i = \begin{pmatrix} (||x_1||^2 - ||x_2||^2) - ((d_{i,1}^2 + 2d_{i,1}er_{i,1} + er_{i,1}^2) - (d_{i,2}^2 + 2d_{i,2}er_{i,2} + er_{i,2}^2)) \\ (||x_1||^2 - ||x_3||^2) - ((d_{i,1}^2 + 2d_{i,1}er_{i,1} + er_{i,1}^2) - (d_{i,3}^2 + 2d_{i,3}er_{i,3} + er_{i,3}^2)) \\ (||x_1||^2 - ||x_4||^2) - ((d_{i,1}^2 + 2d_{i,1}er_{i,1} + er_{i,1}^2) - (d_{i,4}^2 + 2d_{i,4}er_{i,4} + er_{i,4}^2)) \end{pmatrix},$$

equivalentemente pode-se considerar

$$b'_i = \begin{pmatrix} (||x_1||^2 - ||x_2||^2) - (d_{i,1}^2 - d_{i,2}^2) - 2d_{i,1}er_{i,1} - er_{i,1}^2 + 2d_{i,2}er_{i,2} + er_{i,2}^2 \\ (||x_1||^2 - ||x_3||^2) - (d_{i,1}^2 - d_{i,3}^2) - 2d_{i,1}er_{i,1} - er_{i,1}^2 + 2d_{i,3}er_{i,3} + er_{i,3}^2 \\ (||x_1||^2 - ||x_4||^2) - (d_{i,1}^2 - d_{i,4}^2) - 2d_{i,1}er_{i,1} - er_{i,1}^2 + 2d_{i,4}er_{i,4} + er_{i,4}^2 \end{pmatrix}.$$

Como previsto, é possivel separar $b'_i = b_i + er_i$, onde

$$er_i = \begin{pmatrix} 2d_{i,2}er_{i,2} + er_{i,2}^2 - 2d_{i,1}er_{i,1} - er_{i,1}^2 \\ 2d_{i,3}er_{i,3} + er_{i,3}^2 - 2d_{i,1}er_{i,1} - er_{i,1}^2 \\ 2d_{i,4}er_{i,4} + er_{i,4}^2 - 2d_{i,1}er_{i,1} - er_{i,1}^2 \end{pmatrix}.$$

Como $d_{i,j}$, para $j = 1,...,4$ e $i = 5,...,n$, não são conhecidos no caso inexato, será feita uma substituição por $d_{i,j} = d'_{i,j} - er_{i,j}$

$$er_i = \begin{pmatrix} 2(d'_{i,2} - er_{i,2})er_{i,2} + er_{i,2}^2 - 2(d'_{i,1} - er_{i,1})er_{i,1} - er_{i,1}^2 \\ 2(d'_{i,3} - er_{i,2})er_{i,3} + er_{i,3}^2 - 2(d'_{i,1} - er_{i,1})er_{i,1} - er_{i,1}^2 \\ 2(d'_{i,4} - er_{i,4})er_{i,4} + er_{i,4}^2 - 2(d'_{i,1} - er_{i,1})er_{i,1} - er_{i,1}^2 \end{pmatrix},$$

o erro er_i procurado pode ser escrito da seguinte forma:

$$er_i = \begin{pmatrix} 2d'_{i,2}er_{i,2} - er_{i,2}^2 - 2d'_{i,1}er_{i,1} + er_{i,1}^2 \\ 2d'_{i,3}er_{i,3} - er_{i,3}^2 - 2d'_{i,1}er_{i,1} + er_{i,1}^2 \\ 2d'_{i,4}er_{i,4} - er_{i,4}^2 - 2d'_{i,1}er_{i,1} + er_{i,1}^2 \end{pmatrix}. \tag{3.13}$$

É necessário gerar o erro $er_{i,j}$ para executar os testes numéricos. Esse erro será gerado da mesma maneira descrita em (Sit, Wu e Yuan 2008):

$$er_{i,j} = 2RE(0,5 - rand)d_{i,j}, \tag{3.14}$$

para $j = 1,...,4$ e $i = 5,...,n$, onde $rand$ é uma constante que varia no intervalo aberto de 0 a 1 e $RE = 1e-08, 1e-06, 1e-04, 1e-02$.

Substituindo $d_{i,j} = d'_{i,j} - er_{i,j}$ em (3.14), e isolando $er_{i,j}$ obtém-se

$$er_{i,j} = \frac{2RE(0,5 - rand)d'_{i,j}}{1 + 2RE(0,5 - rand)}. \tag{3.15}$$

É possível notar que a variação $2RE(0,5 - rand)$ é limitada da seguinte forma

$$-RE < 2RE(0,5 - rand) < RE. \tag{3.16}$$

Somando 1 a todos os lados da desigualdade encontra-se o dividendo de (3.15).

$$1 - RE < 1 + 2RE(0,5 - rand) < 1 + RE,$$

$$\frac{1}{1-RE} > \frac{1}{1 + 2RE(0,5 - rand)} > \frac{1}{1+RE}.$$

Como $d'_{i,j} \geq 0$, é possível multiplicá-lo pela desigualdade sem que a mesma fique alterada,

$$\frac{d'_{i,j}}{1-RE} \geq \frac{d'_{i,j}}{1 + 2RE(0,5 - rand)} \geq \frac{d'_{i,j}}{1+RE}. \tag{3.17}$$

Como definido em (3.16), $2RE(0,5 - rand)$ pode assumir valores positivos e negativos, portanto para multiplicar este valor por (3.17) será necessário estudar dois casos

separados.

- Para $2RE(0,5-rand) < 0$

$$\frac{2RE(0,5-rand)d'_{i,j}}{1-RE} \geq \frac{2RE(0,5-rand)d'_{i,j}}{1+2RE(0,5-rand)} \geq \frac{2RE(0,5-rand)d'_{i,j}}{1+RE},$$

por (3.15) e (3.16)

$$\frac{RE d'_{i,j}}{1-RE} \geq er_{i,j} \geq \frac{-RE d'_{i,j}}{1+RE}.$$

- Para $2RE(0,5-rand) > 0$

$$\frac{2RE(0,5-rand)d'_{i,j}}{1-RE} \leq \frac{2RE(0,5-rand)d'_{i,j}}{1+2RE(0,5-rand)} \leq \frac{2RE(0,5-rand)d'_{i,j}}{1+RE},$$

por (3.15) e (3.16)

$$\frac{-RE d'_{i,j}}{1-RE} \leq er_{i,j} \leq \frac{RE d'_{i,j}}{1+RE}.$$

Sabe-se que $1+RE > 1-RE$, portanto vale a seguinte desigualdade:

$$\frac{1}{1+RE} < \frac{1}{1-RE},$$

e como RE e $d'_{i,j} \geq 0$, pode-se afirmar:

$$\frac{RE d'_{i,j}}{1+RE} \leq \frac{RE d'_{i,j}}{1-RE} \quad e \quad \frac{-RE d'_{i,j}}{1+RE} \geq \frac{-RE d'_{i,j}}{1-RE},$$

o que resulta na seguinte desigualdade

$$\frac{-RE d'_{i,j}}{1-RE} \leq er_{i,j} \leq \frac{RE d'_{i,j}}{1-RE},$$

ou seja

$$|er_{i,j}| \leq \frac{RE d'_{i,j}}{1-RE}.$$

Agora, usando (3.13) e $|h|$ representando o módulo componente a componente do

vetor $h \in \mathbb{R}^3$, tem-se:

$$|er_i| = \begin{pmatrix} |2d'_{i,2}er_{i,2} - er_{i,2}^2 - 2d'_{i,1}er_{i,1} + er_{i,1}^2| \\ |2d'_{i,3}er_{i,3} - er_{i,3}^2 - 2d'_{i,1}er_{i,1} + er_{i,1}^2| \\ |2d'_{i,4}er_{i,4} - er_{i,4}^2 - 2d'_{i,1}er_{i,1} + er_{i,1}^2| \end{pmatrix}.$$

Usando a desigualdade triangular do módulo.

$$|er_i| \leq \begin{pmatrix} 2d'_{i,2}|er_{i,2}| + |er_{i,2}|^2 + 2d'_{i,1}|er_{i,1}| + |er_{i,1}|^2 \\ 2d'_{i,3}|er_{i,3}| + |er_{i,3}|^2 + 2d'_{i,1}|er_{i,1}| + |er_{i,1}|^2 \\ 2d'_{i,4}|er_{i,4}| + |er_{i,4}|^2 + 2d'_{i,1}|er_{i,1}| + |er_{i,1}|^2 \end{pmatrix},$$

$$|er_i| \leq \begin{pmatrix} \frac{2REd'^2_{i,2}}{1-RE} + \frac{RE^2d'^2_{i,2}}{(1-RE)^2} + \frac{2REd'^2_{i,1}}{1-RE} + \frac{RE^2d'^2_{i,1}}{(1-RE)^2} \\ \frac{2REd'^2_{i,3}}{1-RE} + \frac{RE^2d'^2_{i,3}}{(1-RE)^2} + \frac{2REd'^2_{i,1}}{1-RE} + \frac{RE^2d'^2_{i,1}}{(1-RE)^2} \\ \frac{2REd'^2_{i,4}}{1-RE} + \frac{RE^2d'^2_{i,4}}{(1-RE)^2} + \frac{2REd'^2_{i,1}}{1-RE} + \frac{RE^2d'^2_{i,1}}{(1-RE)^2} \end{pmatrix},$$

$$= \begin{pmatrix} \frac{2REd'^2_{i,2}(1-RE)+RE^2d'^2_{i,2}+2REd'^2_{i,1}(1-RE)+RE^2d'^2_{i,1}}{(1-RE)^2} \\ \frac{2REd'^2_{i,3}(1-RE)+RE^2d'^2_{i,3}+2REd'^2_{i,1}(1-RE)+RE^2d'^2_{i,1}}{(1-RE)^2} \\ \frac{2REd'^2_{i,4}(1-RE)+RE^2d'^2_{i,4}+2REd'^2_{i,1}(1-RE)+RE^2d'^2_{i,1}}{(1-RE)^2} \end{pmatrix},$$

ou seja,

$$|er_i| \leq \begin{pmatrix} \frac{RE(2-RE)(d'^2_{i,2}+d'^2_{i,1})}{(1-RE)^2} \\ \frac{RE(2-RE)(d'^2_{i,3}+d'^2_{i,1})}{(1-RE)^2} \\ \frac{RE(2-RE)(d'^2_{i,4}+d'^2_{i,1})}{(1-RE)^2} \end{pmatrix}. \tag{3.18}$$

É possível concluir que, ao utilizar o método linear para o caso inexato, pode-se afirmar que o erro de cálculo acumulado em cada átomo é limitado superiormente por (3.18) e à partir deste erro é possível calcular uma aproximação do RMSD para o método linear inexato. Com base na fórmula do RMSD (2.5), a fórmula para calcular o erro aproximado pode ser escrita da seguinte forma:

$$RMSD_{APR}(er) = \frac{\|er\|_F}{\sqrt{n}}, \tag{3.19}$$

onde $er = (er_1, ..., er_n)^T$.

3.2.2 RESULTADOS COMPUTACIONAIS - CASO INEXATO

Será utilizada a expressão (3.14) para gerar erros na matriz de distâncias, simulando, desta forma, os erros cometidos pelo processo RMN.

Abaixo seguem resultados do cálculo do RMSD aproximado na subseção 3.2.1 e o RMSD original na seção 2.4. Os testes foram feitos para algumas das proteínas testadas no método linear, as mesmas serão utilizadas para comparar os próximos métodos.

Nas tabelas apresentadas a seguir, para cada proteína são exibidos os seguintes fatores: Proteína - Nome da proteína que consta no banco de dados, Átomos - número de átomos contidos na proteína, Tempo/s - Tempo em segundos que o método linear demorou para encontrar todos os átomos, RMSD-Apr - Erro acumulado durante o processo, calculado pelo (RMSD) utilizando a fórmula (3.19), RMSD - Erro acumulado durante o processo, calculado pelo (RMSD) comparando a proteína original e a calculada pelo método linear inexato.

TABELA 3: Método Linear - Re = 1e-08

Proteína	Átomos	Tempo/s	RMSD-Apr	RMSD
1A1D	146	0,005319	5,17E-05	2,84E-06
1AQR	524	0,000706	2,22E-05	5,75E-06
1CEU	854	0,000975	7,59E-05	1,72E-05
1D8V	4208	0,00356	4,68E-05	1,54E-05
1F39	1534	0,001479	9,88E-05	1,54E-05
1KVX	954	0,001048	1,21E-04	1,84E-05
1LFB	641	0,000811	1,50E-04	4,57E-05
1MBN	1216	0,00132	1,10E-04	1,04E-05
1N4W	8616	0,006916	1,21E-04	7,90E-05
1RGS	2015	0,001902	2,59E-04	2,05E-05
1RWH	5646	0,004701	2,23E-04	5,08E-05
2MSJ	480	0,000713	7,63E-04	6,58E-06
3B34	7479	0,00593	1,28E-04	2,87E-05
8DRH	329	0,000566	6,11E-04	2,25E-05

TABELA 4: Método Linear - Re = 1e-06

Proteína	Átomos	Tempo/s	RMSD-Apr	RMSD
1A1D	146	0,00634	5,17E-03	2,75E-04
1AQR	524	0,000509	2,22E-03	5,34E-04
1CEU	854	0,000751	7,59E-03	1,76E-03
1D8V	4208	0,003289	4,68E-03	1,55E-03
1F39	1534	0,001273	9,88E-03	1,57E-03
1KVX	954	0,000842	1,21E-02	1,81E-03
1LFB	641	0,000605	1,50E-02	3,53E-03
1MBN	1216	0,001025	1,10E-02	1,05E-03
1N4W	8616	0,006538	1,21E-02	7,99E-03
1RGS	2015	0,001936	2,59E-02	2,07E-03
1RWH	5646	0,004365	2,23E-02	5,20E-03
2MSJ	480	0,000487	7,63E-02	7,11E-04
3B34	7479	0,00596	1,28E-02	2,89E-03
8DRH	329	0,000359	6,11E-02	1,76E-03

TABELA 5: Método Linear - Re = 1e-04

Proteína	Átomos	Tempo/s	RMSD-Apr	RMSD
1A1D	146	0,006389	5,17E-01	2,89E-02
1AQR	524	0,000506	2,22E-01	5,38E-02
1CEU	854	0,000747	7,59E-01	1,74E-01
1D8V	4208	0,003326	4,68E-01	1,50E-01
1F39	1534	0,001242	9,88E-01	1,55E-01
1KVX	954	0,000825	1,21E+00	1,88E-01
1LFB	641	0,000597	1,50E+00	3,57E-01
1MBN	1216	0,001065	1,10E+00	1,11E-01
1N4W	8616	0,006519	1,21E+00	7,94E-01
1RGS	2015	0,001587	2,59E+00	2,08E-01
1RWH	5646	0,004272	2,23E+00	5,12E-01
2MSJ	480	0,000479	7,63E+00	6,83E-02
3B34	7479	0,00565	1,28E+00	2,92E-01
8DRH	329	0,000409	6,11E+00	2,22E-01

TABELA 6: Método Linear - Re = 1e-02

Proteína	Átomos	Tempo/s	RMSD-Apr	RMSD
1A1D	146	0,006445	5,25E+01	2,68E+00
1AQR	524	0,000518	2,26E+01	5,89E+00
1CEU	854	0,000892	7,70E+01	1,68E+01
1D8V	4208	0,003391	4,75E+01	1,62E+01
1F39	1534	0,001382	1,00E+02	1,54E+01
1KVX	954	0,02667	1,23E+02	1,72E+01
1LFB	641	0,005061	1,52E+02	3,40E+01

Proteína	Átomos	Tempo/s	RMSD-Apr	RMSD
1MBN	1216	0,002144	1,12E+02	1,09E+01
1N4W	8616	0,214704	1,23E+02	1,05E+02
1RGS	2015	0,002441	2,63E+02	2,03E+01
1RWH	5646	0,006509	2,26E+02	5,10E+01
2MSJ	480	0,000751	7,74E+02	7,23E+00
3B34	7479	0,008478	1,30E+02	2,66E+01
8DRH	329	0,000582	6,20E+02	1,37E+01

Analisando os resultados, é possível verificar que o erro gerado pelo erro aproximado na fórmula (3.19) é maior do que o erro calculado pelo RMSD, pois o erro aproximado é formulado para que haja como limitante superior do erro original.

Ao utilizar o método linear para distâncias inexatas, é possível mensurar o erro sem conhecer as coordenadas originais, assim como definido em (3.19). Isto é possível contanto que se tenha uma estimativa do erro na matriz de distâncias, o valor de RE, ou seja, o erro gerado no processo de RMN.

3.3 CONJUNTO DE DISTÂNCIAS ESPARSAS

O problema geométrico tem distâncias esparsas, quando a matriz de distâncias é esparsa. Isto ocorre devido ao fato do processo de RMN não encontrar todas as distâncias. Com este processo só é possível encontrar distâncias próximas de um ponto específico, portanto grande parte destas distâncias é desconhecida.

Lembrando que a distância molecular é medida em Angstrom ($\mathring{A}$), onde $1\mathring{A} = 10^{-10}$ metros e que esta distância pode variar de $0\mathring{A}$ a $100\mathring{A}$ na matriz de distâncias moleculares. Para gerar a matriz de distâncias esparsas, basta definir o valor máximo que estas distâncias poderão assumir.

Por exemplo, suponha que o corte nas distâncias é da ordem de $10\mathring{A}$, o que significa que a matriz de distâncias só possui valores menores ou iguais a $10\mathring{A}$, ou seja, à partir de um átomo x_i somente são conhecidas as distâncias de átomos na proximidade de no máximo $10\mathring{A}$ de x_i.

Na Figura 24 é mostrada a matriz de distâncias D para distâncias menores ou iguais a $3\mathring{A}$ da proteína 1HOE:

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1.4657	2.5290	0	2.4934	0	0	0	0	0	0	0
2	1.4657	0	1.5252	2.4219	1.5405	2.4682	0	0	2.3872	0	0	0
3	2.5290	1.5252	0	1.2543	2.5282	0	0	0	1.3052	2.4528	0	0
4	0	2.4219	1.2543	0	2.9470	0	0	0	2.2461	2.8018	0	0
5	2.4934	1.5405	2.5282	2.9470	0	1.5339	2.4151	2.4265	0	0	0	0
6	0	2.4682	0	0	1.5339	0	1.2633	1.2635	0	0	0	0
7	0	0	0	0	2.4151	1.2633	0	2.1798	0	0	0	0
8	0	0	0	0	2.4265	1.2635	2.1798	0	0	0	0	0
9	0	2.3872	1.3052	2.2461	0	0	0	0	0	1.4819	2.5115	2.9115
10	0	0	2.4528	2.8018	0	0	0	0	1.4819	0	1.5315	2.4083
11	0	0	0	0	0	0	0	0	2.5115	1.5315	0	1.2551
12	0	0	0	0	0	0	0	0	2.9115	2.4083	1.2551	0
13	0	0	0	0	0	0	0	0	2.4739	1.5454	2.5053	0
14	0	0	0	0	0	0	0	0	2.6605	2.4473	0	0
15	0	0	0	0	0	0	0	0	0	2.5445	0	0

FIGURA 24: Matriz de distâncias - $D = [d_{i,j}]$ para $i, j = 1, ..., n$

De acordo com a Figura 24, a distância $d_{1,1}$ é equivalente a $0\mathring{A}$ e $d_{2,1}$ equivale a $1.4657\mathring{A}$. Os demais valores iguais a $0\mathring{A}$ mas diferentes de $d_{i,i}$, para $i = 1, ..., n$, são distâncias com valores desconhecidos.

Utilizando a matriz de distâncias esparsas, aparece um novo problema na determinação dos átomos de uma proteína, ou seja, como determinar os átomos uma vez que as distâncias não são todas conhecidas? A Figura 25 faz uma representação desta esparsidade.

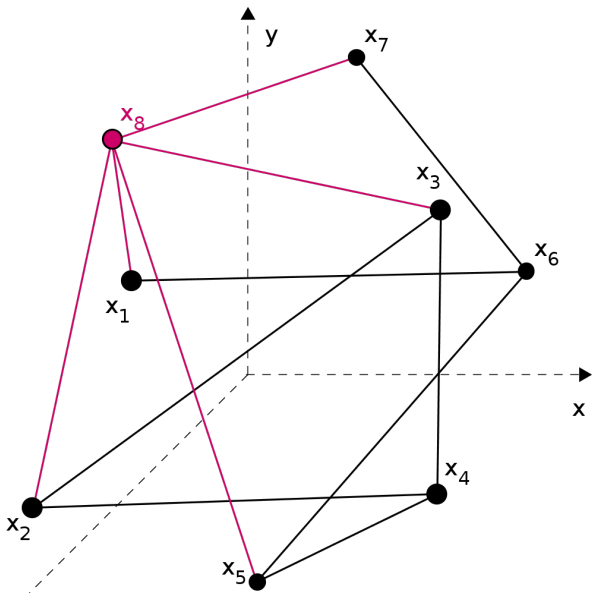


FIGURA 25: Esparsidade no $\mathbb{R}^3$

Analisando a Figura 25, considerando x_8 em relação aos outros átomos, é possível ver que nem todas as distâncias são conhecidas. No entanto, ainda pode-se determinar x_8 utilizando apenas as distâncias que são conhecidas, visto que existem pelo menos quatro delas conhecidas.

Em geral, para estruturas de proteínas, a matriz de distâncias está definida em torno de 5$\mathring{A}$ a 8$\mathring{A}$.

3.3.1 MÉTODO LINEAR PARA A MATRIZ DE DISTÂNCIAS ESPARSAS

Inicialmente o algoritmo utilizado para o caso de distâncias esparsas será o mesmo algoritmo apresentado na seção 3.1, com algumas modificações. Os quatro átomos iniciais são calculados da mesma maneira, no entanto, estes átomos não podem ser usados para fixar todos os átomos restantes da molécula, pois nem todas as distâncias são conhecidas.

Segundo (Dong e Wu 2003), para fixar estes átomos que não foram determinados, pode-se escolher novos átomos que já estejam determinados como novos átomos iniciais. Para tanto, é necessário que os novos átomos possuam distâncias com relação ao átomo procurado e à partir desses novos átomos utilizar o método linear.

A única restrição para esses novos átomos iniciais é que não sejam coplanares, desta maneira a matriz A em (3.1) será inversível. Para fazer a verificação de que estes átomos não são coplanares, serão geradas doze matrizes A diferentes, utilizando os átomos conhecidos que possuem distâncias em relação ao átomo procurado. A partir dessas matrizes será verificado qual delas tem o maior módulo do determinante. Desta maneira a matriz A estará o mais longe possível de ser singular. Isto aumentará a chance do ponto estar correto.

Também é possível verificar que como os átomos iniciais não serão fixados a priori e sim escolhidos de modo que não sejam coplanares, a nova matriz inicial A não será triangular inferior. Portanto, o procedimento utilizado no algoritmo da subseção 3.1.3, para resolver o sistema $Ax_i = b_i$, para A, b_i definidos em (3.1) e $x_i = (u_i, v_i, w_i)$, terá de

ser modificado.

Este sistema também pode ser escrito da seguinte maneira:

$$A_{1,1}u_i + A_{1,2}v_i + A_{1,3}w_i = b_{i,1}, \tag{3.20}$$

$$A_{2,1}u_i + A_{2,2}v_i + A_{2,3}w_i = b_{i,2}, \tag{3.21}$$

$$A_{3,1}u_i + A_{3,2}v_i + A_{3,3}w_i = b_{i,3}. \tag{3.22}$$

Se $A_{1,1}$ for diferente de zero, é possível eliminar u_i do sistema com as seguintes operações:

$$(3.20)\frac{A_{2,1}}{A_{1,1}} - (3.21) \quad \underbrace{(\frac{A_{1,2}A_{2,1}}{A_{1,1}} - A_{2,2})}_{B_1} v_i + \underbrace{(\frac{A_{1,3}A_{2,1}}{A_{1,1}} - A_{2,3})}_{C_1} w_i = \underbrace{\frac{b_{i,1}A_{2,1}}{A_{1,1}} - b_{i,2}}_{D_1}, \tag{3.23}$$

$$\Leftrightarrow$$

$$(3.20)\frac{A_{3,1}}{A_{1,1}} - (3.22) \quad \underbrace{(\frac{A_{1,2}A_{3,1}}{A_{1,1}} - A_{3,2})}_{B_2} v_i + \underbrace{(\frac{A_{1,3}A_{3,1}}{A_{1,1}} - A_{3,3})}_{C_2} w_i = \underbrace{\frac{b_{i,1}A_{3,1}}{A_{1,1}} - b_{i,3}}_{D_2}. \tag{3.24}$$

Usando as equações (3.23) e (3.24) pode-se formular um sistema mais compacto:

$$B_1 v_i + C_1 w_i = D_1, \tag{3.25}$$

$$B_2 v_i + C_2 w_i = D_2. \tag{3.26}$$

Para encontrar u_i e v_i, basta utilizar a regra de cramer em (Anton e Busby 2006), para resolver (3.25) e (3.26), supondo que $(B_1C_2 - C_1B_2)$ seja não nulo, tem-se

$$v_i = \frac{C_2D_1 - C_1D_2}{B_1C_2 - C_1B_2}, \tag{3.27}$$

$$w_i = \frac{D_2B_1 - D_1B_2}{B_1C_2 - C_1B_2}. \tag{3.28}$$

Substituindo (3.27) e (3.28) em qualquer uma das equações (3.20), (3.21) ou (3.22), é possível encontrar u_i. Neste trabalho a coordenada u_i será definida utilizando (3.20).

$$u_i = \frac{b_{i,1} - A_{1,2}v_i - A_{1,3}w_i}{A_{1,1}}. \tag{3.29}$$

Com estas informações é possível escrever o algoritmo para o método linear no caso esparso.

3.3.2 ALGORITMO - MÉTODO LINEAR PARA DISTÂNCIAS ESPARSAS

Dados de entrada: $D = [d_{i,j}]$ onde $i, j = 1, ..., n$

Operações:

$F = \{x_l \mid l = 1, ..., 4\}$, átomos fixados na subseção 3.1.2

$U = \{n-4$ átomos não fixados$\}$

Enquanto $U \neq \emptyset$ faça

Para $x_i \in U$ onde $i = 5, ..., n$.

Encontre x_1, x_2, x_3, $x_4 \in$ F, que contenham distâncias em relação x_i

Calcule A, como em (3.1)

Calcule b_i em (3.1)

Resolva $Ax_i = b_i$

Mova x_i de U para F

Fim

Fim

O ponto x_i é calculado utilizando:

B_1, C_1 e D_1 em (3.23) e B_2, C_2 e D_2 em (3.24). Então usando (3.27), (3.28) e (3.29),

para obter:

$$x_i = \begin{pmatrix} \frac{b_{i,1} - A_{1,2}x_{i,2} - A_{1,3}x_{i,3}}{A_{1,1}} \\ \frac{C_1D_2 - C_2D_1}{C_1B_2 - C_2B_1} \\ \frac{D_1B_2 - D_2B_1}{C_1B_2 - C_2B_1} \end{pmatrix} \tag{3.30}$$

3.3.3 RESULTADOS COMPUTACIONAIS - CASO ESPARSO

A partir da matriz de distâncias exatas, é calculada a matriz de distancias esparsas e, então, aplicado o método linear para distâncias esparsas apresentado na seção anterior. Como já foi mostrado na seção 3.2 este algoritmo pode ser usado para o caso inexato.

Abaixo seguem resultados do cálculo do erro absoluto e o tempo que cada proteína demorou para calcular as estruturas das proteínas, com as seguintes esparsidades na matriz de distâncias: 16$\mathring{A}$, 14$\mathring{A}$, 12$\mathring{A}$, 10$\mathring{A}$, 8$\mathring{A}$.

Nas Tabelas 7 a 16, “- ”significa que não foi possível determinar a estrutura da proteína.

TABELA 7: RMSD - Método Linear - Esparso e Exato

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	1,53E-09	4,26E-07	1,00E+07	-	-
1AQR	524	1,29E-11	5,22E-10	4,94E-10	7,99E-09	4,11E-06
1CEU	854	6,52E-10	1,44E-10	1,75E-09	1,52E-07	5,29E-07
1D8V	4208	4,71E-08	3,45E-06	1,47E+02	2,89E+19	1,90E+93
1F39	1534	3,32E-08	1,57E-03	8,09E-05	-	-
1KVX	954	9,88E-10	9,67E-10	4,10E-09	6,86E-08	7,07E-03
1LFB	641	5,46E-09	2,54E-10	1,97E-09	2,56E-07	6,33E-04
1MBN	1216	1,91E-09	2,54E-07	3,65E-07	8,11E+51	2,32E+28
1N4W	8616	2,49E-02	6,96E-02	-	-	-
1RGS	2015	4,84E-06	6,70E-03	2,96E-02	4,78E+16	1,91E+29
1RWH	5646	1,48E-01	1,98E-02	4,14E+101	-	-
2MSJ	480	2,11E-10	2,76E-11	3,39E-09	1,83E-08	7,20E-06
3B34	7479	-	-	-	-	-
8DRH	329	4,45E-10	1,73E-10	1,94E-10	-	-

TABELA 8: Tempo - Método Linear - Esparso e Exato

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
11A1D	146	0.09913	0.09610	0,003831	-	-
1AQR	524	0.43871	0.44347	0,459099	0,423149	0,412305
1CEU	854	0.74888	0.77890	0,76653	0,75375	0,772818
1D8V	4208	6.04534	5.78940	6,016496	5,436904	6,203874
1F39	1534	1.51065	1.45443	1,411285	-	-
1KVX	954	0.86776	0.81736	1,261225	0,994988	1,143398
1LFB	641	0.54090	0.51472	0,563153	0,985969	0,65356
1MBN	1216	1.20309	1.12706	1,095604	1,331054	1,494619
1N4W	8616	18.39623	16.98993	-	-	-
1RGS	2015	2.04988	2.00526	2,21765	2,371333	1,955006
1RWH	5646	8.16408	8.03048	8,446336	-	-
2MSJ	480	0.38785	0.38199	0,393038	0,449407	0,362921
3B34	7479	-	-	-	-	-
8DRH	329	0.24907	0.28296	0,240924	-	-

TABELA 9: RMSD - Método Linear - Esparso Re = 1e-08

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	1,34E-02	1,46E+00	-	-	-
1AQR	524	1,08E-03	2,68E-04	3,54E-03	1,33E-02	1,31E+07
1CEU	854	2,17E-02	5,80E-03	2,06E-02	9,94E+00	3,75E+05
1D8V	4208	1,09E+04	1,66E+61	7,84E+93	-	-
1F39	1534	2,58E-01	1,61E+09	1,23E+12	-	-
1KVX	954	1,33E-01	2,30E-03	1,45E+00	4,76E+00	8,22E+10
1LFB	641	5,04E-03	3,07E-02	3,43E-02	6,32E+10	1,15E+04
1MBN	1216	1,06E-01	2,56E-02	1,04E-01	3,02E+25	2,38E+17
1N4W	8616	1,33E+00	-	-	-	-
1RGS	2015	7,77E+14	1,47E+31	9,56E+30	5,68E+67	4,03E+51
1RWH	5646	-	-	-	-	-
2MSJ	480	7,72E-05	1,80E-03	5,59E-04	1,12E-01	1,11E-01
3B34	7479	-	-	-	-	-
8DRH	329	3,34E-04	1,17E-03	1,58E-03	-	-

TABELA 10: Tempo - Método Linear - Esparso Re = 1e-08

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	0,795898	0,604198	-	-	-
1AQR	524	0,588557	0,437329	0,456935	0,532734	0,522588
1CEU	854	0,897118	0,719866	0,737136	0,707906	0,680759
1D8V	4208	6,291598	5,696436	5,392499	-	-
1F39	1534	1,64473	1,608501	1,367503	-	-
1KVX	954	1,007609	0,88313	0,788641	0,820559	0,889716
1LFB	641	0,687144	0,548082	0,540032	0,506233	0,490199

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1MBN	1216	1,334172	1,110287	1,049303	1,041269	1,002401
1N4W	8616	1,334172	-	-	-	-
1RGS	2015	2,243418	2,63735	2,086572	1,993246	1,990058
1RWH	5646	-	-	-	-	-
2MSJ	480	0,594758	0,52672	0,516418	0,509115	0,496862
3B34	7479	-	-	-	-	-
8DRH	329	0,388539	0,383989	0,391384	-	-

TABELA 11: RMSD - Método Linear - Esparso Re = 1e-06

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	6,49E-02	1,93E+02	-	-	-
1AQR	524	1,10E-01	5,54E-02	1,28E-01	5,60E+01	7,13E+02
1CEU	854	3,35E-01	1,31E-01	2,44E+02	2,28E+07	2,09E+20
1D8V	4208	1,36E+62	4,49E+84	-	-	-
1F39	1534	1,22E+29	1,95E+35	1,08E+46	-	-
1KVX	954	1,04E+04	5,97E+18	3,55E+13	1,39E+24	2,24E+21
1LFB	641	1,34E-01	2,36E+12	3,62E+00	1,91E+18	7,41E+26
1MBN	1216	5,25E+10	2,65E+27	8,92E+16	1,67E+28	3,41E+09
1N4W	8616	-	-	-	-	-
1RGS	2015	2,23E+66	1,99E+59	3,44E+53	2,03E+00	9,89E+51
1RWH	5646	-	-	-	-	-
2MSJ	480	5,76E-02	1,42E-01	1,48E-01	1,86E-01	1,95E+07
3B34	7479	-	-	-	-	-
8DRH	329	4,56E-02	9,16E-02	5,78E-01	-	-

TABELA 12: Tempo - Método Linear - Esparso Re = 1e-06

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	0,230201	0,238272	-	-	-
1AQR	524	0,463051	0,447956	0,460224	0,594795	0,55677
1CEU	854	0,768024	0,747511	0,753366	0,722606	0,697831
1D8V	4208	6,102112	5,739064	-	-	-
1F39	1534	1,512271	1,437516	1,630447	-	-
1KVX	954	0,878901	0,962031	0,948559	0,838827	0,897775
1LFB	641	0,55475	0,53475	0,521652	0,499113	0,493614
1MBN	1216	1,157109	1,114816	1,073664	1,032587	1,000055
1N4W	8616	-	-	-	-	-
1RGS	2015	2,627558	2,152694	2,091098	1,31E+063	1,982104
1RWH	5646	-	-	-	-	-
2MSJ	480	0,953828	0,519375	0,553067	0,507158	0,502224
3B34	7479	-	-	-	-	-
8DRH	329	0,390405	0,390575	0,396223	-	-

TABELA 13: RMSD - Método Linear - Esparso Re = 1e-04

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	2,44E+01	2,43E+05	-	-	-
1AQR	524	8,14E+03	2,82E+12	3,20E+09	1,50E+11	2,97E+02
1CEU	854	1,43E+05	1,93E+07	3,71E+17	3,79E+20	1,65E+18
1D8V	4208	1,01E+82	5,12E+91	-	-	-
1F39	1534	1,42E+41	5,69E+38	2,65E+55	-	-
1KVX	954	3,77E+21	1,74E+22	1,64E+34	1,35E+23	4,26E+19
1LFB	641	8,32E+20	9,90E+16	2,40E+16	2,42E+14	3,40E+13
1MBN	1216	1,53E+31	2,36E+42	2,35E+43	1,01E+26	2,33E+30
1N4W	8616	-	-	-	-	-
1RGS	2015	2,30E+54	1,10E+63	4,13E+67	5,29E+66	1,64E+49
1RWH	5646	-	-	-	-	-
2MSJ	480	6,80E+01	1,54E+00	6,67E+07	1,86E+11	1,29E+09
3B34	7479	-	-	-	-	-
8DRH	329	1,22E+01	1,59E+02	3,41E+01	-	-

TABELA 14: Tempo - Método Linear - Esparso Re = 1e-04

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	0,231036	0,259283	-	-	-
1AQR	524	0,462597	0,449411	0,536118	0,539938	0,528177
1CEU	854	7,58E-001	0,753982	0,722599	0,711015	0,696083
1D8V	4208	6,193822	5,747442	-	-	-
1F39	1534	1,515493	1,607632	1,558847	-	-
1KVX	954	0,865053	0,965483	0,841824	0,832856	0,911168
1LFB	641	0,539403	0,682477	0,51795	0,507416	0,498779
1MBN	1216	1,157351	1,12527	1,078442	1,038289	1,007359
1N4W	8616	-	-	-	-	-
1RGS	2015	2,246021	2,225707	2,086793	2,035245	2,164628
1RWH	5646	-	-	-	-	-
2MSJ	480	0,530607	0,527902	0,507493	0,497862	0,488008
3B34	7479	-	-	-	-	-
8DRH	329	0,401338	0,384049	0,388642	-	-

TABELA 15: RMSD - Método Linear - Esparso Re = 1e-02

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	1,45E+01	1,86E+01	-	-	-
1AQR	524	7,61E+07	1,14E+01	6,17E+06	1,10E+16	1,31E+13
1CEU	854	3,27E+03	7,84E+08	7,56E+15	2,81E+16	7,34E+20
1D8V	4208	3,79E+68	3,50E+92	-	-	-
1F39	1534	5,12E+48	1,40E+56	7,03E+52	-	-
1KVX	954	1,82E+21	2,01E+32	3,04E+20	8,86E+36	1,57E+25
1LFB	641	2,26E+19	9,03E+24	1,46E+07	5,79E+24	1,84E+14

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1MBN	1216	2,43E+20	2,92E+39	3,09E+43	1,91E+32	1,09E+34
1N4W	8616	-	-	-	-	-
1RGS	2015	3,91E+49	1,77E+45	4,17E+57	2,55E+55	2,09E+61
1RWH	5646	-	-	-	-	-
2MSJ	480	5,83E-01	1,04E+00	6,33E+13	1,41E+05	1,44E+06
3B34	7479	-	-	-	-	-
8DRH	329	9,67E+00	8,74E-01	6,06E+00	-	-

TABELA 16: Tempo - Método Linear - Esparso Re = 1e-02

Proteína	Átomos	16 $\mathring{A}$	14 $\mathring{A}$	12 $\mathring{A}$	10 $\mathring{A}$	8 $\mathring{A}$
1A1D	146	0,237367	0,230034	-	-	-
1AQR	524	0,464753	0,454676	0,563644	0,53869	0,540227
1CEU	854	7,67E-001	0,748797	0,799063	0,711232	0,709456
1D8V	4208	6,413405	5,9635	-	-	-
1F39	1534	1,50292	1,60413	1,548886	-	-
1KVX	954	0,860335	0,857615	0,842928	0,829598	0,974873
1LFB	641	0,543725	0,524543	0,516871	0,498347	0,517203
1MBN	1216	1,147813	1,112056	1,077272	1,039237	1,03432
1N4W	8616	-	-	-	-	-
1RGS	2015	2,503484	2,215669	2,07319	2,032641	2,032624
1RWH	5646	-	-	-	-	-
2MSJ	480	0,64035	0,633243	0,55772	0,495365	0,509718
3B34	7479	-	-	-	-	-
8DRH	329	0,390139	0,388795	0,383215	-	-

De acordo com os resultados, tanto para o caso exato como inexato, proteínas que possuem muitos átomos demoram muito tempo para serem encontradas. Na Tabela 8, para a proteína 1N4W com $16\mathring{A}$ de esparsidade já se pode notar demora na resolução do algoritmo, cerca de 18 segundos. Conforme a esparsidade aumenta, fica mais difícil determinar a matriz A de átomos iniciais, fazendo com que o método seja mais demorado ou até mesmo não consiga determinar alguns átomos, impossibilitando determinar sua estrutura, como no caso das esparsidades $12\mathring{A}$, $10\mathring{A}$ e $8\mathring{A}$ para a mesma proteína.

Já no caso do erro, mesmo que a proteína seja determinada rapidamente, pode ocorrer desses resultados serem insatisfatórios, como na Tabela 7 para a proteína 1MBN com $10\mathring{A}$ de esparsidade. Mesmo apresentando uma resolução rápida, em

torno de 1 segundo, o erro é de $8,11E+51$ que é muito grande se comparado ao tamanho da proteína.

Mesmo considerando estas discrepâncias, o caso exato e esparso apresenta ótimos resultados, como no caso $16\mathring{A}$, alcançando um erro em torno de $1E-11$ a $1E-02$. O caso inexato e esparso tem resultados pouco satisfatórios, considerando que para $RE = 1e-08$ o erro já varia de $1E-05$ a $1E+14$.

O aumento da esparsidade também pode implicar na impossibilidade de encontrar algumas estruturas, devido ao fato de poucas distâncias serem conhecidas. Isso acontece pois não existem átomos com distâncias a x_i o suficiente para gerar a matriz A. Em geral, pode-se dizer que este método mostra-se eficaz para o caso de proteínas pequenas, com distâncias exatas e com pouca esparsidade na matriz de distâncias.

Os resultados podem apresentar certa diferença sempre que forem calculados, pois cada iteração depende fortemente dos pontos base encontrados, que são definidos ao calcular o determinante de A.

4 VARIAÇÕES DO MÉTODO LINEAR

Neste capítulo serão abordados métodos baseados no método linear, mas com algumas modificações que os tornam mais eficientes. Estes métodos são eficazes apenas nos casos em que o corte na matriz de distâncias esparsas é menor ou igual a $8\mathring{A}$, pois dependem da busca por novas distâncias a cada passo.

4.1 MÉTODO LINEAR ATUALIZADO

O Método Linear Atualizado - MLA descrito em (Wu e Wu 2007 - a), é baseado no método linear para distâncias esparsas, descrito na subseção 3.3.1. Ele foi remodelado para que o erro acumulado em cada iteração possa ser controlado.

Ao invés de serem utilizados átomos calculados em iterações anteriores para calcular x_i, estes átomos servirão apenas como base para determinar novos átomos que possuam as mesmas distâncias entre si. Então estes novos átomos substituirão os anteriores na estrutura original.

Ao calcular o determinante da matriz A, não existe garantia de que ela seja bem condicionada, sendo assim será verificado seu condicionamento e então escolhida a matriz que tiver o menor número de condicionamento. Pois, mesmo que o determinante da matriz seja grande, não há como garantir que a matriz seja bem condicionada. Isso acaba acarretando em uma resposta ruim para o problema.

O número condição de uma matriz A é dado por, $\kappa(A) = ||A^{-1}|| \cdot ||A||$, sendo que, uma matriz é considerada bem condicionada quando $\kappa(A) \simeq 1$.

Os átomos iniciais são determinados da maneira mostrada na subseção 3.1.2, portanto falta apenas calcular $n-4$ átomos. Estes $n-4$ átomos são calculados da seguinte maneira: para cada iteração são encontradas distâncias entre o átomo x_i

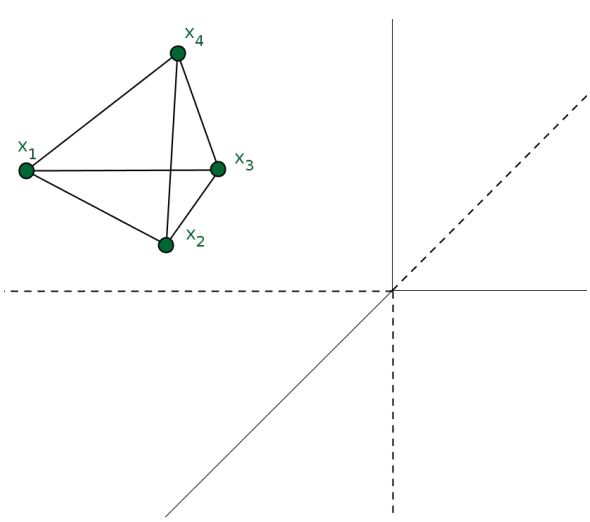
procurado e quatro outros átomos que já tenham suas coordenadas determinadas.

O próximo passo é verificar se estes quatro átomos utilizados possuem distâncias entre si. Se estas distâncias existirem, as coordenadas destes átomos serão atualizadas da forma descrita na subseção 3.1.2, de modo que uma nova base de átomos x'_j, para $j = 1,...,4$, será gerada e então com estas novas coordenadas o ponto x'_i será determinado.

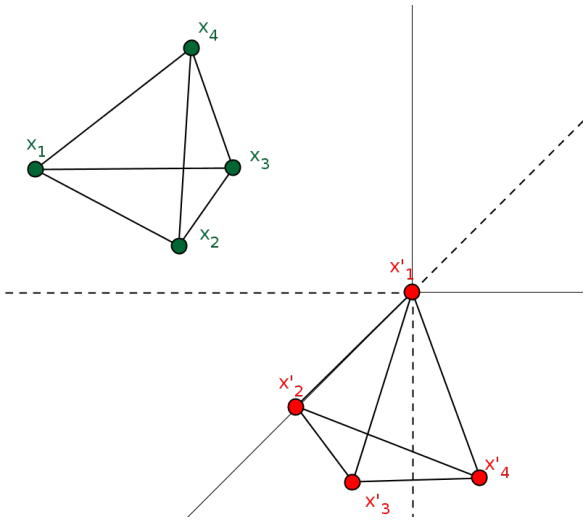
Através deste passo é possível eliminar o erro gerado quando estes átomos foram determinados. Visto que este fato ocorreu utilizando apenas átomos que já haviam sido determinados anteriormente por átomos com um certo erro. Agora é possível colocar esta base de átomos x'_j, para $j = 1,...,4$, e o novo ponto x'_i recalculado em sua posição correta e substituindo-os na proteína pelos átomos originais.

Este reposicionamento é possível através da translação e rotação dessas coordenadas encontradas para a base original dos átomos utilizados, ou seja, utilizando alguns passos do RMSD descritos na seção 2.4.

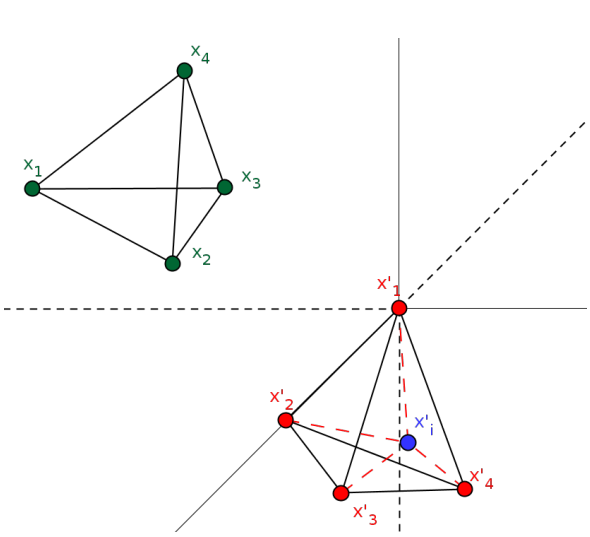
A Figura 26 mostra os passos executados pelo método linear atualizado para calcular o átomo x_i de uma proteína qualquer.



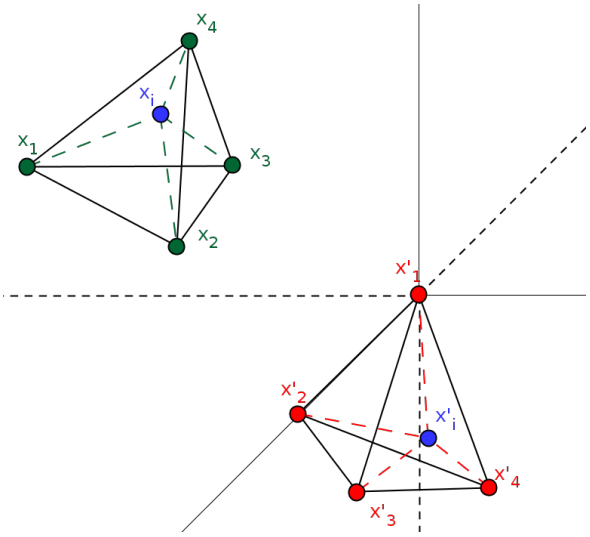
(a) Os átomos x_1, x_2, x_3 e x_4 quaisquer possuem distâncias conhecidas entre si mesmos e a x_i



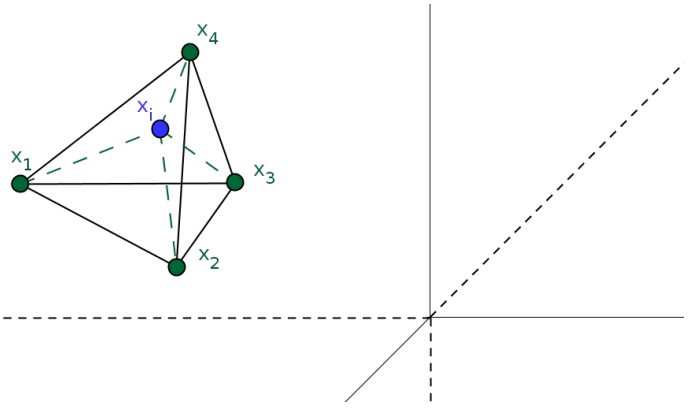
(b) Reinicialização dos átomos x_1, x_2, x_3 e x_4 para que se posicionem na origem



(c) Cálculo de x'_i com o uso dos átomos reposicionados x'_1, x'_2, x'_3 e x'_4



(d) Uso do RMSD para o reposicionamento x'_1, x'_2, x'_3, x'_4 e x'_i em x_1, x_2, x_3, x_4 e x_i



(e) As coordenadas auxiliares x'_1, x'_2, x'_3, x'_4 e x'_i são excluídas, restando apenas o sistema original

FIGURA 26: Passos executados por uma iteração do MLA

Seguindo os passos mostrados na Figura 26, o átomo x_i não carregará o erro que os átomos x_1, x_2, x_3 e x_4 acumularam ao serem calculados. Restando apenas o erro atual que cada distância possui, acarretando em um erro final bem menor.

Para reposicionar as novas coordenadas, considera-se X as coordenadas da nova base de pontos recalculada x'_j, para $j = 1, ..., 4$, e x_i e Y as coordenadas dos átomos escolhidos para serem substituídos. Agora é necessário transladar e rotacionar X para que suas coordenadas estejam de acordo com a base original como na seção 2.4.

Calcula-se a média de todos os átomos, como em (2.3) para encontrar xc e yc, onde $n = 4$, e então calcula-se a translação para X e para Y, como em (2.4).

De acordo com (2.5), pode-se utilizar $Y_1 Q$ para encontrar as coordenadas melhores alinhadas às antigas. Neste caso X será a coordenada a ser modificada.

Portanto, serão necessárias as seguintes operações: $C = X'Y$ e o cálculo da decomposição em valores singulares de C para encontrar U e V. De acordo com (Wu 2006), calcula-se $Q = UV'$ para então aplicar o valor de Q a X e x'_i.

$$\begin{aligned} X &= XQ + yc, \\ x'_i &= (x'_i - xc)Q + yc. \end{aligned} \tag{4.1}$$

Com essas novas coordenadas é possível substituir as coordenadas da base antiga de átomos por suas novas coordenadas X e x'_i.

4.1.1 ALGORITMO - MÉTODO LINEAR ATUALIZADO

Dados de entrada: $D = [d_{i,j}]$ onde $i, j = 1, ..., n$

$F = \{x_i \mid i = 1, ..., 4\}$, átomos fixados na subseção 3.1.2

$U = \{n - 4$ átomos não fixados$\}$

Enquanto $U \neq \emptyset$

Para $x_i \in U$ onde $i = 5, ..., n$.

Encontre x_1, x_2, x_3, $x_4 \in F$, que contenham distâncias em relação x_i

Se x_1, x_2, x_3, x_4 possuem distâncias entre si.

Reinicialize x'_1, x'_2, x'_3 e x'_4, como descrito na subseção 3.1.2

Substitua x_1, x_2, x_3, x_4 por X e $x_i = x'_i$, X e x'_i estão definidos na seção 4.1

Fim

Fim

O átomo x_i é calculado como na equação (3.10) utilizando x'_1, x'_2, x'_3 e x'_4 como

coordenadas da base.

4.1.2 RESULTADOS COMPUTACIONAIS - MÉTODO LINEAR ATUALIZADO

Abaixo seguem resultados do cálculo do erro absoluto e o tempo que cada proteína demorou para ser calculada com o uso do método linear atualizado.

Os testes foram realizados para distâncias esparsas exatas e inexatas, com as seguintes esparsidades na matriz de distâncias D: 8$\mathring{A}$, 7$\mathring{A}$, 6$\mathring{A}$, 5$\mathring{A}$ e 4$\mathring{A}$.

Para esparsidades menores que $4\mathring{A}$ não foi possível realizar testes devido ao fato de não haverem distâncias o suficiente relativas ao átomo x_i procurado.

TABELA 17: RMSD - MLA - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	1,71E-14	2,32E-14	2,88E-14	2,36E-14	1,59E-08
1CEU	854	1,77E-14	2,22E-14	2,74E-14	4,65E-14	5,98E-10
1D8V	4208	1,41E-13	1,42E-13	2,11E-12	3,52E-11	1,61E-06
1F39	1534	1,00E-13	8,38E-14	3,42E-13	6,87E-11	-
1KVX	954	2,42E-14	2,64E-14	4,87E-14	4,72E-11	-
1LFB	641	2,83E-14	2,83E-14	7,30E-14	6,04E-14	-
1MBN	1216	2,89E-14	2,73E-14	1,15E-13	-	-
1N4W	8616	1,03E-13	1,38E-13	1,60E-12	4,45E-13	5,52E-09
1RGS	2015	2,17E-13	3,36E-13	4,93E-12	-	-
1RWH	5646	3,22E-13	1,03E-12	2,30E-12	-	-
2MSJ	480	3,57E-14	3,67E-14	4,18E-14	9,41E-13	-
3B34	7479	1,76E-13	1,05E-11	2,05E-12	-	-
8DRH	329	9,69E-14	1,41E-13	1,79E-13	1,79E-12	5,95E-11

TABELA 18: Tempo - MLA - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,9548	1,1093	1,0356	0,9245	21,4876
1CEU	854	1,6515	1,4049	1,4455	2,3668	23,7313
1D8V	4208	9,8570	9,3019	9,1890	9,3276	133,6374
1F39	1534	2,8478	2,7480	2,9391	5,2337	-
1KVX	954	1,7339	1,7936	1,9437	1,7505	-

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1LFB	641	1,1005	1,1074	1,1335	1,1041	-
1MBN	1216	2,2151	2,1942	2,2125	-	-
1N4W	8616	24,5450	23,0617	22,9828	23,3882	312,9148
1RGS	2015	4,0591	3,9671	4,7871	-	-
1RWH	5646	15,7390	12,1425	13,9742	-	-
2MSJ	480	1,0151	1,1056	0,8882	1,1755	-
3B34	7479	17,6878	16,7901	17,2185	-	-
8DRH	329	0,5660	0,8553	1,2529	1,2782	29,7027

TABELA 19: RMSD - MLA - Esparso Re = 1e-08

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	3,46E-07	3,28E-07	9,81E-07	3,03E-06	6,72E+00
1CEU	854	4,52E-07	3,87E-07	6,35E-07	1,60E-06	8,66E-02
1D8V	4208	1,86E-06	1,16E-06	8,16E-04	4,57E-03	1,49E+01
1F39	1534	2,83E-06	2,66E-06	1,41E-04	2,05E-03	-
1KVX	954	4,77E-07	8,51E-07	2,30E-06	9,25E-03	-
1LFB	641	3,11E-07	4,13E-07	1,08E-06	3,65E-06	-
1MBN	1216	1,08E-06	1,18E-06	3,70E-06	-	-
1N4W	8616	2,99E-06	4,43E-06	1,63E-04	2,53E-04	2,06E+01
1RGS	2015	2,73E-06	5,56E-06	9,43E-06	-	-
1RWH	5646	5,12E-06	1,24E-05	1,47E-04	-	-
2MSJ	480	5,07E-07	6,66E-07	6,97E-07	6,13E-05	-
3B34	7479	9,23E-06	5,59E-04	8,29E-04	-	-
8DRH	329	4,33E-05	4,93E-06	2,39E-05	1,64E-04	3,91E-03

TABELA 20: Tempo - MLA - Esparso Re = 1e-08

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,8941	1,0362	1,0239	1,7548	3,7242
1CEU	854	1,3460	1,3997	1,4643	2,2306	4,5257
1D8V	4208	9,9905	9,6909	9,7478	11,3192	20,8390
1F39	1534	3,1448	3,1037	3,1794	3,3914	-
1KVX	954	1,7638	1,7791	1,7855	1,8536	-
1LFB	641	1,1221	1,1180	1,1235	1,1495	-
1MBN	1216	2,2356	2,2268	2,2383	-	-
1N4W	8616	25,7833	24,1605	26,6272	23,8257	48,0589
1RGS	2015	4,1357	4,0277	4,0149	-	-

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1RWH	5646	13,7860	12,8527	13,0072	-	-
2MSJ	480	0,8643	0,8733	0,8894	0,9246	-
3B34	7479	20,4012	17,6726	17,7894	-	-
8DRH	329	0,5459	0,5355	0,6485	0,6648	3,0161

TABELA 21: RMSD - MLA - Esparso Re = 1e-06

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	6,69E-05	4,19E-05	3,72E-05	2,92E-04	6,55E+00
1CEU	854	4,61E-05	5,79E-05	7,82E-05	3,45E-04	2,36E+00
1D8V	4208	1,04E-04	1,26E-04	6,98E-02	3,59E-02	1,67E+01
1F39	1534	3,78E-03	4,70E-04	1,22E-03	8,82E-01	-
1KVX	954	4,31E-05	6,55E-05	2,86E-04	1,35E+00	-
1LFB	641	3,49E-05	4,27E-05	8,58E-05	6,58E-04	-
1MBN	1216	5,89E-05	6,50E-05	2,81E-04	-	-
1N4W	8616	3,91E-04	1,94E-04	3,42E-04	1,20E-02	1,56E+01
1RGS	2015	6,46E-04	8,73E-04	5,12E-03	-	-
1RWH	5646	1,47E-04	5,91E-04	1,07E-02	-	-
2MSJ	480	4,04E-05	7,75E-05	6,05E-05	7,10E-03	-
3B34	7479	9,18E-04	8,59E-04	2,93E-01	-	-
8DRH	329	2,32E-04	2,60E-03	2,25E-03	6,59E-03	2,00E+00

TABELA 22: Tempo - MLA - Esparso Re = 1e-06

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,9301	1,0504	1,7186	1,0208	3,5926
1CEU	854	1,3788	1,4113	1,3305	1,4795	4,3598
1D8V	4208	10,1263	15,1964	8,9466	9,3500	18,9945
1F39	1534	3,1755	3,1139	2,9203	3,1597	-
1KVX	954	1,7674	1,7543	1,6581	1,7653	-
1LFB	641	1,1306	1,1139	1,0447	1,1032	-
1MBN	1216	2,2758	2,2338	2,0361	-	-
1N4W	8616	25,7820	24,3609	21,3830	22,1501	43,8787
1RGS	2015	4,1838	4,0247	3,6921	-	-
1RWH	5646	13,2928	13,0387	11,8175	-	-
2MSJ	480	0,8688	0,8852	0,8193	0,8767	-
3B34	7479	18,4370	17,6364	15,9327	-	-
8DRH	329	0,5409	0,5423	0,5909	0,6053	2,8166

TABELA 23: RMSD - MLA - Esparso Re = 1e-04

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	7,49E-03	5,91E-03	1,40E-02	1,07E-02	5,54E+00
1CEU	854	4,98E-03	4,67E-03	7,58E-03	8,62E-03	1,10E+01
1D8V	4208	1,61E-02	3,23E-02	5,17E-02	2,95E-01	1,49E+01
1F39	1534	7,03E-02	1,51E-02	1,23E+01	1,49E+01	-
1KVX	954	5,16E-03	6,18E-03	1,55E-02	7,69E+00	-
1LFB	641	5,11E-03	4,93E-03	6,15E-03	1,28E-01	-
1MBN	1216	5,61E-03	1,02E-02	3,45E-02	-	-
1N4W	8616	4,04E-02	5,05E-02	5,06E-02	3,19E+00	1,99E+01
1RGS	2015	1,85E-02	5,23E-02	4,22E+00	-	-
1RWH	5646	5,50E-02	4,47E-01	6,65E+00	-	-
2MSJ	480	8,00E-03	3,97E-03	6,52E-03	4,99E-01	-
3B34	7479	1,09E-01	3,29E-01	2,74E+00	-	-
8DRH	329	2,01E-01	5,70E-01	1,73E-01	6,27E-01	8,34E+00

TABELA 24: Tempo - MLA - Esparso Re = 1e-04

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,8891	0,9797	0,9702	0,9841	3,5407
1CEU	854	1,2908	1,3352	1,3691	1,4785	4,3234
1D8V	4208	9,4860	9,2333	9,2014	9,4310	19,0488
1F39	1534	2,9828	2,9493	2,9968	3,1724	-
1KVX	954	1,6754	1,6757	1,6860	1,7712	-
1LFB	641	1,0681	1,0569	1,0605	1,0944	-
1MBN	1216	2,1602	2,0970	2,1126	-	-
1N4W	8616	24,1458	22,8124	22,0525	22,2101	43,9626
1RGS	2015	4,0104	3,9246	3,8396	-	-
1RWH	5646	12,4583	12,1310	12,1580	-	-
2MSJ	480	0,8205	0,8313	0,8457	0,8805	-
3B34	7479	17,2070	16,6702	16,4553	-	-
8DRH	329	0,5023	0,5096	0,5932	0,6221	2,8234

TABELA 25: RMSD - MLA - Esparso Re = 1e-02

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	3,02E+00	5,18E-01	9,31E-01	7,36E+00	6,57E+00
1CEU	854	5,49E-01	4,99E-01	8,77E-01	7,29E-01	1,01E+01
1D8V	4208	9,43E+00	1,28E+01	9,68E+00	1,10E+01	1,73E+01
1F39	1534	9,92E+00	7,96E+00	1,92E+00	1,01E+01	-
1KVX	954	9,63E-01	3,85E+00	9,21E-01	1,29E+01	-

1LFB	641	3,38E-01	4,75E-01	1,05E+00	1,85E+00	-
1MBN	1216	1,58E+00	2,16E+00	8,34E+00	-	-
Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1N4W	8616	1,93E+01	8,10E+00	5,08E+00	1,76E+01	1,78E+01
1RGS	2015	1,80E+01	1,89E+01	1,62E+01	-	-
1RWH	5646	1,94E+01	2,22E+01	2,06E+01	-	-
2MSJ	480	7,58E+00	1,54E+00	7,28E-01	9,29E+00	-
3B34	7479	2,30E+01	2,38E+01	2,28E+01	-	-
8DRH	329	8,03E+00	5,47E+00	4,82E+00	3,54E+00	7,06E+00

TABELA 26: Tempo - MLA - Esparso Re = 1e-02

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,9318	0,9755	0,9821	1,0056	3,5155
1CEU	854	1,3035	1,4194	1,3834	1,5710	4,1196
1D8V	4208	9,5876	9,2403	9,2502	9,6346	22,7344
1F39	1534	3,0103	2,9635	3,0366	3,2400	-
1KVX	954	1,7326	1,7134	1,7256	1,8200	-
1LFB	641	1,0981	1,0829	1,0708	1,1177	-
1MBN	1216	2,1861	2,1223	2,1150	-	-
1N4W	8616	24,3223	23,2740	22,1576	22,9525	43,6256
1RGS	2015	4,0339	3,9092	3,8629	-	-
1RWH	5646	12,6110	12,2000	12,1770	-	-
2MSJ	480	0,8243	0,8338	0,8422	0,8926	-
3B34	7479	17,4032	16,6895	16,5749	-	-
8DRH	329	0,5053	0,5140	0,6078	0,7052	2,9931

Casos em que não foi possível determinar a estrutura da proteína ocorreram devido ao tamanho da esparsidade da matriz de distâncias. Esta esparsidade impossibilitou encontrar, em alguns casos, átomos o suficiente já determinados com distâncias a x_i'. Não houve nenhum caso de moléculas indeterminadas devido ao tempo de processamento.

Pode-se notar que o método linear atualizado obteve resultados muito melhores que o método linear esparso. Isso ocorreu como previsto, pois no método atualizado os átomos base são recalculados para então serem recolocados na estrutura da proteína. Este passo minimiza o erro a cada iteração.

Lembrando que como a base de cada iteração é escolhida de forma aleatória,

através da análise da condição da matriz de doze possíveis bases, o resultado pode ter uma pequena variação em cada iteração.

4.2 MÉTODO LINEAR RIGIDO

Em (Wu, Wu e Yuan 2007 - b) é apresentada uma variação do método linear atualizado. A diferença é que caso não existam quatro átomos para serem utilizados como base, o método do algoritmo rígido executa um algoritmo auxiliar com apenas três átomos iniciais. O algoritmo é chamado de rígido se nenhuma parte da estrutura pode ser modificada sem violar as condições de distâncias iniciais.

Como isso acontece na prática? O uso de apenas três átomos iniciais, torna possível encontrar dois átomos em cada passo do método, como mostado na Figura 19. Isto ocasiona a possibilidade de encontrar múltiplas estruturas geradas por reflexões a cada passo do método. Todas essas estruturas devem ser armazenadas, para mais tarde excluir estruturas que não satisfaçam restrições de distâncias conhecidas.

Este método pode oferecer certa desvantagem, pois é necessário armazenar todos os dados de possíveis estruturas. Em vários casos não é possível determinar ao fim do método uma única estrutura.

Este método é construído com base no método linear atualizado. Caso não exista solução pelo método linear atualizado, o ponto é buscado através do algoritmo rígido. O algoritmo rígido pode ser encontrado, usando o mesmo raciocínio do método linear na seção 3.1. Considerando que apenas três átomos são conhecidos ao invés de quatro.

$$||x_i - x_1|| = d_{i,1},$$

$$||x_i - x_2|| = d_{i,2},$$

$$||x_i - x_3|| = d_{i,3}.$$

Elevando ao quadrado dos dois lados de cada igualdade:

$$||x_i||^2 - 2x_i^T x_1 + ||x_1||^2 = d_{i,1}^2,$$
$$||x_i||^2 - 2x_i^T x_2 + ||x_2||^2 = d_{i,2}^2,$$
$$||x_i||^2 - 2x_i^T x_3 + ||x_3||^2 = d_{i,3}^2.$$

Substituindo as coordenadas de $x_i = (u_i, v_i, w_i)$ e $x_j = (u_j, v_j, w_j)$ em $x_i^T x_j$, para $j = 1,2,3$:

$$\begin{aligned} ||x_i||^2 - 2u_i u_1 - 2v_i v_1 - 2w_i w_1 + ||x_1||^2 &= d_{i,1}^2, \\ ||x_i||^2 - 2u_i u_2 - 2v_i v_2 - 2w_i w_2 + ||x_2||^2 &= d_{i,2}^2, \\ ||x_i||^2 - 2u_i u_3 - 2v_i v_3 - 2w_i w_3 + ||x_3||^2 &= d_{i,3}^2. \end{aligned} \tag{4.2}$$

Subtraindo a primeira equação das demais:

$$2u_i(u_1 - u_2) + 2v_i(v_1 - v_2) + 2w_i(w_1 - w_2) = (||x_1||^2 - ||x_2||^2) - (d_{i,1}^2 - d_{i,2}^2),$$
$$2u_i(u_1 - u_3) + 2v_i(v_1 - v_3) + 2w_i(w_1 - w_3) = (||x_1||^2 - ||x_3||^2) - (d_{i,1}^2 - d_{i,3}^2).$$

O resultado será o seguinte:

$$A = 2 \begin{pmatrix} (u_1 - u_2) & (v_1 - v_2) & (w_1 - w_2) \\ (u_1 - u_3) & (v_1 - v_3) & (w_1 - w_3) \end{pmatrix} \tag{4.3}$$

e

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

Também pode ser escrito como um sistema linear:

$$Ax_i = b, \tag{4.4}$$

onde $x_i = (u_i, v_i, w_i)$ é o ponto procurado.

Este sistema não fornece uma solução exata para x_i visto que a matriz A tem 2 linhas e 3 colunas. Portanto é necessário fazer algumas modificações para que seja

possível encontrar os dois valores procurados para x_i.

Seja $x_i = A^T y_i$ onde $y_i = (y_{i,1}, y_{i,2})$, substituindo x_i em (4.4)

$$Ax_i = AA^T y_i = b.$$

Se $x^T AA^T x = 0$, então $||A^T y||_2^2 = 0$, portanto $A^T y = 0$, ou seja, $y = 0$. Isso significa que AA^T tem dimensão 2 e é definida positiva, consequentemente AA^T é não singular. Logo, é possível resolver o sistema linear $AA^T y_i = b$ e obter uma única solução para $y_i = (y_{i,1}, y_{i,2})$.

Lembrando que os pontos x_1, x_2 e x_3, devem ser escolhidos de modo que não sejam colineares. Em (Anton e Rorres 2012), o posto de uma matriz A_{mxn} é denotado por

$$posto(A) \leq min(m,n),$$

onde m é o número de linhas e n o numero de colunas linearmente independentes de A.

Neste caso, como as linhas de A_{2x3} são definidas como linearmente independentes, o resultado será:

$$posto(A) = 2.$$

Agora considere $A' = A_{1:2,1:2}$, então com $x'_{iaux} = A' y_i$ é possível calcular $x'_{iaux} = (u_i, v_i)$. Porém, ainda falta calcular w_i para que $x_i = (u_i, v_i, w_i)$ seja determinado.

Para encontrar w_i basta substituir u_i e v_i em uma das equações de (4.2)

$$||x_i||^2 - 2u_i u_j - 2v_i v_j - 2w_i w_j + ||x_j||^2 = d_{i,j}^2$$

para $j = 1$, 2 ou 3. Substituindo x_i nesta última equação, resulta em

$$w_i^2 - 2w_j w_i - d_{i,j}^2 + u_i^2 + v_i^2 - 2u_i u_j - 2v_i v_j + ||x_j||^2 = 0. \tag{4.5}$$

Substituindo x'_{iaux} na equação (4.5) é possível obter uma equação do segundo grau

em relação a w_i, que como esperado resultará em dois átomos diferentes para x_i. No entanto, como solução final da iteração do algoritmo, é possível obter dois átomos, um átomo, ou dois átomos imaginários. Caso não seja encontrada como solução a opção de dois átomos, a ordem dos átomos base deve ser alterada, pois não é possível determinar uma estrutura válida no $\mathbb{R}^3$ com os resultados de um átomo, ou resultados imaginários, visto que o algoritmo proposto obtém como resultado duas soluções.

4.2.1 ALGORITMO - MÉTODO LINEAR RÍGIDO

Dados de entrada: $D = [d_{i,j}]$ onde $i, j = 1, ..., n$

$F = \{x_i \mid i = 1, ..., 4\}$, átomos fixados na subseção 3.1.2

$U = \{n-4$ átomos não fixados$\}$

Enquanto $U \neq \emptyset$ faça

Para $x_i \in U$ onde $i = 5, ..., n$.

Encontre x_1, x_2, x_3, $x_4 \in F$, que contenham distâncias em relação x_i

Se x_1, x_2, x_3, x_4 possuem distâncias entre si.

Reinicialize x'_1, x'_2, x'_3 e x'_4, utilizando a seção 3.1.2

Substitua x_1, x_2, x_3, x_4 por X e $x_i = x'_i$, X e x'_i estão definidos na subseção 4.1.

Se x_1, x_2 e x_3 possuem distâncias entre si.

Calcule x_i utilizando a seção 4.2.

Fim

Fim

Se existirem x_1, x_2, $x_3 \in$ F, que contenham distâncias em relação x_i, então x_i será calculado da maneira descrita na seção 4.2:

Calcule $AA^T y_i = b$

$x'_i = A'y_i$, onde $A' = A(1:2, 1:2)$,

$$w_i^2 - 2w_j w_i - (d_{i,j}^2 - u_i^2 - v_i^2 + 2u_i u_j + 2v_i v_j - ||x_j||^2) = 0$$

onde $j = 1$ ou 2 ou 3, resolvendo a equação do segundo grau em w_i é possível determinar x_i .

4.2.2 RESULTADOS COMPUTACIONAIS - MÉTODO LINEAR RÍGIDO

Como descrito no método rígido são necessários os átomos x_1, x_2 e x_3 para então encontrar x_i, onde estes três átomos não devem ser colineares, ou seja, a condição da matriz formada por estes átomos deve ser pequena. Para este algoritmo, o numero condição utilizado é $\kappa(A) \leq 20$, isso significa que para o sistema $Ax_i = b_i$, se houver uma perturbação em b_i, a perturbação permitida em A será até 20 vezes maior.

No entanto, a condição da matriz A é sempre muito grande $\kappa(A) \geq 20$, ocasionando resultados não desejados (um átomo ou dois átomos imaginários). Portanto, não serão feitos testes com este algoritmo.

4.3 MÉTODO LINEAR RÍGIDO - VERSÃO 2

O método descrito em (Wu 2006) também é baseado no método linear atualizado descrito na seção 4.1. Este método também segue o modelo descrito em (Wu, Wu e Yuan 2007 - b).

Primeiramente, é possível iniciar o método se houverem apenas três pontos com distâncias conhecidas entre si. Outro fato importante é que as coordenadas dos três átomos a serem calculados também serão reinicializadas e a partir delas calculado o novo ponto.

Para recalcular os quatro átomos iniciais será utilizado o mesmo cálculo realizado na subseção 3.1.2, então a partir deles será calculado o ponto x_i como no método atualizado na seção 4.1. Se houverem apenas três átomos eles também devem ser

calculados como na subseção 3.1.2, mas então o átomo x_i deve ser calculado como na fórmula (3.9). Trocando a distância $d_{4,j}$, para $j = 1,2,3$ por $d_{i,j}$, para $j = 1,2,3$.

$$\begin{aligned} u_i &= (d_{i,1}^2 - d_{i,2}^2)/(2u_2) + u_2/2 \\ v_i &= (d_{i,2}^2 - d_{i,3}^2 - (u_i - u_2)^2 + (u_i - u_3)^2)/2v_3 + v_3/2 \\ w_i &= \pm^{+}_{-} (d_{i,1}^2 - u_i^2 - v_i^2)^{1/2}. \end{aligned} \tag{4.6}$$

Como discutido na subseção 3.1.2, serão obtidos dois átomos. Este método também gera múltiplas estruturas, que devem ser eliminadas da mesma maneira discutida em (Wu, Wu e Yuan 2007 - b). A cada novo átomo encontrado, devem ser verificadas as distâncias de cada estrutura, para encontrar quais átomos não condizem com a distância procurada.

4.3.1 ALGORITMO - MÉTODO LINEAR RÍGIDO - VERSÃO 2

Dados de entrada: $D = [d_{i,j}]$ onde $i, j = 1, ..., n$

$F = \{x_i \mid i = 1, ..., 4\}$, átomos fixados na subseção 3.1.2

$U = \{n-4$ átomos não fixados$\}$

Enquanto $U \neq \emptyset$ faça

Para $x_i \in U$ onde $i = 5, ..., n$.

Encontre x_1, x_2, x_3, $x_4 \in F$, que contenham distâncias em relação x_i

Se x_1, x_2, x_3, x_4 possuem distâncias entre si.

Reinicialize x_1', x_2', x_3' e x_4', utilizando a subseção 3.1.2

Substitua x_1, x_2, x_3, x_4 por X e $x_i = x_i'$, X e x_i' estão definidos na seção 4.1.

Se x_1, x_2 e x_3 possuem distâncias entre si.

Reinicialize x_1', x_2' e x_3' como na subseção 3.1.2 e x_i' como na formula (4.6)

Substitua x_1, x_2 e x_3 por X da subseção 4.1 e x_i por x_i' com a formula (4.6).

Fim

Fim

4.3.2 RESULTADOS COMPUTACIONAIS - MÉTODO LINEAR RÍGIDO - VERSÃO 2

Os testes foram realizados para distâncias esparsas exatas, com as seguintes esparsidades na matriz de distâncias D: 8$\mathring{A}$, 7$\mathring{A}$, 6$\mathring{A}$ e 5$\mathring{A}$. Para a esparsidade $4\mathring{A}$ não foi feita uma análise utilizando as proteínas habituais, pois os resultados apresentados foram insatisfatórios e demorados. Ao invés disso os testes foram feitos para proteínas discutidas em (Wu 2006).

Devido a semelhança nos resultados, para o caso exato e inexato, os testes do algoritmo rígido V2 inexato não foram calculados. No entanto o algoritmo foi implementado para que sejam possíveis obter estes resultados se necessários.

Abaixo seguem os resultados do cálculo do erro absoluto e o tempo que cada proteína demorou para ser calculada com o uso do método linear rígido v2 - MLRV2.

TABELA 27: RMSD - MLRV2 - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$
1A1D	146	-	-	-	-
1AQR	524	1,99E-14	1,90E-14	1,74E-14	3,04E-14
1CEU	854	2,16E-14	1,91E-14	2,60E-14	4,18E-14
1D8V	4208	1,44E-13	1,46E-13	1,50E-11	6,35E-11
1F39	1534	3,56E-13	1,25E-13	8,26E-12	4,19E-12
1KVX	954	1,96E-14	2,69E-14	6,39E-14	1,68E-10
1LFB	641	3,62E-14	2,02E-14	4,99E-14	6,89E-14
1MBN	1216	3,29E-14	4,65E-13	8,75E-14	-
1N4W	8616	1,06E-13	9,32E-14	1,14E-12	3,17E-01
1RGS	2015	2,11E-13	2,22E-13	5,04E-13	-
1RWH	5646	3,26E-13	4,36E-13	1,05E-10	-
2MSJ	480	3,41E-14	3,93E-14	4,94E-14	7,39E-13
3B34	7479	1,64E-13	2,42E-12	2,50E-12	4,34E-03
8DRH	329	1,33E-13	7,26E-14	7,42E-14	1,65E-12

TABELA 28: Tempo - MLRV2 - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$
1A1D	146	-	-	-	-
1AQR	524	0,8056	0,7980	0,7793	0,8073
1CEU	854	1,1854	1,2028	1,2502	1,4303
1D8V	4208	8,8097	8,4873	8,6039	9,0094
1F39	1534	3,1673	3,0626	2,8068	3,2950
1KVX	954	1,5686	1,5307	1,5923	1,7074
1LFB	641	0,9624	0,9590	0,9968	1,0612
1MBN	1216	1,9860	1,9307	2,0136	-
1N4W	8616	22,5181	21,7297	21,3746	23,6486
1RGS	2015	3,6676	3,6229	3,6011	-
1RWH	5646	11,4421	11,4685	11,4801	-
2MSJ	480	0,7554	0,7819	0,7948	0,8545
3B34	7479	15,7973	15,6715	15,6023	53,9568
8DRH	329	0,5584	0,5085	3,6279	1,2980

Comparando os resultados obtidos pelo método linear rígido V2 com o método linear atualizado na subseção 4.1.2, pode-se notar que ambos obtiveram resultados satisfatórios. Não se pode afirmar que um resultado foi melhor que o outro, considerando o cálculo do erro, devido ao fato de ambos possuírem átomos que podem variar a cada iteração. No entanto o algoritmo rígido apresentou uma pequena vantagem no tempo de cálculo para as proteínas testadas.

Houve apenas um caso em que o algoritmo rígido v2 foi capaz de calcular a proteína e o método atualizado não conseguiu. Este caso ocorreu para a proteína 3B34 com distância 5$\mathring{A}$, mas o erro obtido foi grande comparado com os demais erros para distâncias 5$\mathring{A}$.

Para distâncias menores que $5\mathring{A}$, em vários casos não foi possível determinar a proteína e nos casos em que foi possível determinar a estrutura o resultado foi demorado (como previsto pelo método), mas o erro encontrado foi muito maior do que o esperado.

Para haver uma base de comparação, os próximos resultados foram calculados para as mesmas proteínas discutidas em (Wu 2006), para os algoritmos atualizado e rígido v2, visto que os casos em que é possível utilizar o algoritmo MLRV2 são poucos

e também difíceis de serem encontrados.

TABELA 29: RMSD - MLA - Esparso e Exato

Proteína	Átomos	8.5 $\mathring{A}$	7.5 $\mathring{A}$	7 $\mathring{A}$
1BKR	887	4,05E-14	4,25E-14	4,47E-14
1EJG	843	1,68E-14	1,73E-14	1,86E-14
1IO0	1297	6,42E-14	7,51E-14	8,00E-14
1LIT	1068	3,18E-14	4,00E-14	1,98E-13
1WRI	716	7,18E-14	3,31E-14	4,17E-14

TABELA 30: Tempo - MLA - Esparso e Exato

Proteína	Átomos	8.5 $\mathring{A}$	7.5 $\mathring{A}$	7 $\mathring{A}$
1BKR	887	2,7714	1,760877	1,846688
1EJG	843	1,4505	1,449669	1,473128
1IO0	1297	2,2476	2,244849	2,514185
1LIT	1068	2,0366	2,071206	2,014985
1WRI	716	1,3447	1,309612	1,319805

TABELA 31: RMSD - MLRV2 - Esparso e Exato

Proteína	Átomos	8.5 $\mathring{A}$	7.5 $\mathring{A}$	7 $\mathring{A}$
1BKR	887	4,84E-14	3,92E-14	4,47E-14
1EJG	843	1,17E-14	1,40E-14	1,40E-14
1IO0	1297	5,97E-14	8,59E-14	6,25E-14
1LIT	1068	2,99E-14	2,36E-14	4,34E-14
1WRI	716	3,04E-14	2,63E-14	8,30E-14

TABELA 32: Tempo - MLRV2 - Esparso e Exato

Proteína	Átomos	8.5 $\mathring{A}$	7.5 $\mathring{A}$	7 $\mathring{A}$
1BKR	887	1,54035	1,520315	1,522547
1EJG	843	1,361286	1,376549	1,367245
1IO0	1297	2,120638	2,085244	2,096999
1LIT	1068	1,963802	1,874665	1,858003
1WRI	716	1,292582	1,233094	1,235666

Diferente dos resultados mostrados em (Wu 2006), o método linear atualizado resolve todos os casos apresentados para distâncias $8.5\mathring{A}$, $7.5\mathring{A}$ e $7\mathring{A}$, assim como o método linear rígido v2. A única melhora pode ser notada no tempo de processamento para o algoritmo rígido v2.

TABELA 33: RMSD - MLA - Esparso e Exato

Proteína	Átomos	5 $\mathring{A}$	4 $\mathring{A}$
1ABA	728	5,92E-12	-
1BKR	887	-	-
1EJG	843	1,81E-14	1,58E-11
1HYP	656	3,32E-10	-

TABELA 34: Tempo - MLA - Esparso e Exato

Proteína	Átomos	5 $\mathring{A}$	4 $\mathring{A}$
1ABA	728	1,3973	-
1BKR	887	-	-
1EJG	843	1,5173	3,48402
1HYP	656	1,0778	-

TABELA 35: RMSD - MLRV2 - Esparso e Exato

Proteína	Átomos	5 $\mathring{A}$	4 $\mathring{A}$
1ABA	728	1,67E-12	-
1BKR	887	4,18E-13	-
1EJG	843	1,59E-14	5,11E-12
1HYP	656	4,23E-13	

TABELA 36: Tempo - MLRV2 - Esparso e Exato

Proteína	Átomos	5 $\mathring{A}$	4 $\mathring{A}$
1ABA	728	1,309864	-
1BKR	887	1,721799	-
1EJG	843	1,427729	6,709034
1HYP	656	1,829084	

Para a esparsidade $5\mathring{A}$ o método linear rígido v2 foi capaz de encontrar a estrutura 1BKR e neste caso o resultado foi satisfatório. No caso da proteína 1HYP o resultado também foi melhor, mas o tempo de processamento foi maior. Isto ocorreu porque apesar de 1HYP ter distâncias necessárias para ser calculado pelo método atualizado, a matriz A não estava bem definida. O que não interfere para o bom resultado no método linear rígido. No entanto, o uso deste método ocasionou o aumento do tempo de cálculo da estrutura.

No caso da esparsidade $4\mathring{A}$, o método linear rígido v2 não foi satisfatório. Pois o erro calculado foi aproximado e, no entanto, o tempo de processamento dobrou. Analisando estes resultados, é possível concluir que o método linear rígido v2 é eficaz para esparsidades entre $8.5\mathring{A}$ e $5\mathring{A}$, contanto que não seja necessário de fato encontrar x_i utilizando apenas 3 átomos.

5 MÉTODO LINEAR COM O USO DE MÍNIMOS QUADRADOS PARA DISTANCIAS INEXATAS

Neste capítulo serão discutidos dois métodos diferentes, que também são baseados no método linear. Para determinar o átomo x_i procurado, os algoritmos a serem abordados utilizam o método dos mínimos quadrados ao invés de serem determinados com uma fórmula exata. Estas duas abordagens para o uso do método linear podem ser encontradas em (Wu, Wu e Yuan 2007) e (Sit, Wu e Yuan 2009).

5.1 MÉTODO LINEAR COM MÍNIMOS QUADRADOS LINEAR

Quando a matriz de distâncias D possui erros, estes erros podem variar entre $]-RE, RE[$, como visto em (3.14). Porém, estas distâncias não são consistentes ao serem comparadas com todas as distâncias conhecidas em relação a um certo átomo.

Para fixar a coordenada de um átomo através do método linear, são utilizados apenas quatro átomos que já tenham sido determinados. Ao utilizar apenas quatro átomos o resultado obtido será totalmente baseado neles, ou seja, o erro gerado pode ser maior se comparado a média de todos os erros de distâncias conhecidas.

Para contornar essa situação, nesta versão do método linear ao invés de utilizar apenas quatro átomos, serão usados todos os átomos já determinados que possuírem distâncias a x_i, onde x_i é o átomo procurado. Seja l o número de distâncias conhecidas para determinar o átomo x_i, ou seja, $d_{i,j}$, para $j = 1,...,l$ são conhecidas e $x_j = (u_j, v_j, w_j)^T$, para $j = 1,...,l$, já foram previamente determinados.

Para satisfazer o problema de distâncias proposto em (2.1), tem-se:

$$||x_i - x_j||^2 = d_{i,j}^2, \ para \ j = 1,...,l.$$

Expandindo o problema, obtém-se um resultado análogo:

$$||x_i||^2 - 2x_i^T x_j + ||x_j||^2 = d_{i,j}^2, \ para \ j = 1, ..., l.$$

Substituindo $x_j = (u_j, v_j, w_j)^T$ e $x_i = (u_i, v_i, w_i)^T$ no termo $2x_i^T x_j$:

$$||x_i||^2 - 2u_i u_j - 2v_i v_j - 2w_i w_j + ||x_j||^2 = d_{i,j}^2, \ para \ j = 1, ..., l.$$

Ao invés de subtrair a primeira equação das demais, a equação j será subtraída da equação $j+1$, eliminando assim o termo quadrático de x_i. Desta maneira, o problema pode ser resolvido de forma linear.

$$2u_i(u_j - u_{j+1}) + 2v_i(v_j - v_{j+1}) + 2w_i(w_j - w_{j+1}) = (||x_j||^2 - ||x_{j+1}||^2) - (d_{i,j}^2 - d_{i,j+1}^2),$$

para $j = 1, ..., l-1$.

O sistema acima pode ser representado da seguinte forma:

$$Ax_i = b_i \tag{5.1}$$

onde $x_i = (u_i, v_i, w_i)^T$, para $i = 5, ..., n$, e cada i assume a posição de um átomo ainda não determinado na estrutura da molécula, em que

$$A = 2 \begin{pmatrix} (u_1 - u_2) & (v_1 - v_2) & (w_1 - w_2) \\ & \dots & \\ (u_{l-1} - u_l) & (v_{l-1} - v_l) & (w_{l-1} - w_l) \end{pmatrix}$$

e

$$b_i = \begin{pmatrix} (||x_1||^2 - ||x_2||^2) - (d_{i,1}^2 - d_{i,2}^2) \\ \dots \\ (||x_{l-1}||^2 - ||x_l||^2) - (d_{i,l-1}^2 - d_{i,l}^2) \end{pmatrix},$$

onde l é o número de coordenadas já determinadas que possuem distâncias a x_i.

Para que o sistema final obtido possa ser resolvido, é necessário que a matriz A possua pelo menos 3 linhas linearmente independentes, ou seja, $posto(A) = 3$. O

sistema final será da forma $A_{l-1,3}x_{i3,1} = b_{l,1}$, onde $l \geq 4$.

É possível encontrar a solução para o sistema $Ax = b$ de Mínimos quadrados com o uso da pseudo inversa, como descrito em (Kincaid e Cheney 2009).

$$A^{+} = (A^{T}A)^{-1}A^{T}$$

Portanto, multiplicando o sistema $Ax = b$ por A^{T}, tem-se:

$$A^{T}Ax = A^{T}b.$$

Como A foi definida com posto 3, então $A^{T}A$ tem posto cheio, ou seja, $A^{T}A$ é inversível.

$$x_i = (A^{T}A)^{-1}A^{T}b_i \tag{5.2}$$

Este método será chamado de Método linear com Mínimos Quadrados Linear - MLMQL.

5.1.1 ALGORITMO - MÉTODO LINEAR COM MÍNIMOS QUADRADOS LINEAR

Dados de entrada: $D = [d_{i,j}]$ onde $i, j = 1, ..., n$

$F = \{x_i \mid i = 1, ..., 4\}$, átomos fixados na subseção 3.1.2

$U = \{n - 4$ átomos não fixados$\}$

Enquanto $U \neq \emptyset$

Para $x_i \in U$ onde $i = 5, ..., n$.

Encontre $x_j \in$ F, onde $l \geq 4$ e $j = 1, ..., l$, com distâncias em relação x_i

Calcule A e b_i, utilizando (5.1)

Calcule x_i, utilizando (5.2)

Fim

5.1.2 RESULTADOS COMPUTACIONAIS - MÉTODO LINEAR COM MÍNIMOS QUADRADOS LINEAR

Ao invés de utilizar a pseudo inversa para calcular o sistema $Ax = b$, foi utilizado o comando do MATLAB "barra invertida ($A\backslash b$)". Este comando resolve o sistema linear não quadrático com o uso do método dos mínimos quadrados.

Abaixo seguem resultados do cálculo do erro absoluto e o tempo que cada proteína demorou para ser calculada através do método linear com mínimos quadrados linear.

Os testes foram realizados para distâncias esparsas exatas e inexatas, com as seguintes esparsidades na matriz de distâncias D: 8$\mathring{A}$, 7$\mathring{A}$, 6$\mathring{A}$, 5$\mathring{A}$ e 4$\mathring{A}$.

Para esparsidades menores que $4\mathring{A}$ não foi possível realizar testes devido ao fato de não haverem distâncias o suficiente relativas ao átomo x_i procurado.

TABELA 37: RMSD - MLMQL - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	3,05E-14	4,94E-14	8,15E-14	2,89E-09	7,95E+15
1CEU	854	4,78E-14	5,69E-14	1,35E-12	2,97E-08	1,55E+03
1D8V	4208	1,55E-13	2,53E-13	9,59E-09	3,10E+01	1,72E+11
1F39	1534	1,06E-11	6,44E-12	1,74E-08	5,35E+10	-
1KVX	954	1,28E-13	4,37E-13	5,36E-11	4,97E+00	-
1LFB	641	8,87E-14	6,39E-13	1,69E-10	3,11E-02	-
1MBN	1216	2,04E-13	3,28E-12	3,24E-09	-	-
1N4W	8616	3,54E-13	8,40E-13	8,72E-08	1,06E+01	6,77E+46
1RGS	2015	2,27E-12	2,80E-10	3,11E-02	-	-
1RWH	5646	4,04E-11	7,61E-08	1,30E+01	-	-
2MSJ	480	8,15E-14	1,22E-13	4,76E-13	9,24E-04	-
3B34	7479	2,97E-10	1,54E-06	1,45E+01	-	-
8DRH	329	3,59E-12	1,21E-11	2,87E-08	3,58E-04	4,84E+01

TABELA 38: Tempo - MLMQL - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,883356	0,333138	0,301687	0,286862	0,384967
1CEU	854	0,405099	0,390317	0,360018	0,335053	0,421712
1D8V	4208	3,36354	2,497951	2,248039	2,061954	2,263001

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1F39	1534	0,789055	0,719183	0,68397	0,676384	-
1KVX	954	0,470016	0,407111	0,373883	0,349793	-
1LFB	641	0,341372	0,262424	0,240153	0,225886	-
1MBN	1216	0,561978	0,51189	0,479227	-	-
1N4W	8616	7,518945	6,885352	6,086803	5,598397	6,399959
1RGS	2015	0,948628	0,971172	0,881718	-	-
1RWH	5646	3,423973	3,205227	2,987798	-	-
2MSJ	480	0,198158	0,184658	0,176975	0,165099	-
3B34	7479	4,993525	4,675276	4,368146	-	-
8DRH	329	0,149255	0,12852	0,122204	0,13177	0,228248

TABELA 39: RMSD - MLMQL - Esparso Re = 1e-08

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	1,67E-07	2,09E-07	6,49E-07	7,75E-01	3,55E+08
1CEU	854	1,93E-07	2,10E-07	1,44E-06	1,31E+00	1,50E+09
1D8V	4208	4,38E-07	4,48E-07	5,27E-01	1,66E+01	1,88E+41
1F39	1534	1,53E-04	7,61E-05	1,48E-01	3,73E+02	-
1KVX	954	7,88E-07	4,44E-06	4,67E-03	1,06E+01	-
1LFB	641	2,64E-06	6,29E-06	1,25E-02	1,28E+01	-
1MBN	1216	2,35E-06	9,08E-06	7,98E-03	-	-
1N4W	8616	1,08E-06	1,85E-06	1,72E-01	2,89E+02	2,07E+58
1RGS	2015	1,04E-05	1,30E-03	9,71E+00	-	-
1RWH	5646	1,62E-04	6,76E-02	1,99E+01	-	-
2MSJ	480	1,16E-06	1,87E-06	9,12E-06	8,46E+00	-
3B34	7479	3,03E-03	3,44E+00	2,44E+01	-	-
8DRH	329	1,04E-05	4,50E-05	3,36E-01	1,17E+01	8,16E+01

TABELA 40: Tempo - MLMQL - Esparso Re = 1e-08

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,244965	0,342427	0,296511	0,28328	0,382977
1CEU	854	0,412647	0,393611	0,361107	0,334906	0,419599
1D8V	4208	2,730342	2,466341	2,256412	2,066004	2,261704
1F39	1534	0,738112	0,695398	0,678838	0,673991	-
1KVX	954	0,426478	0,398712	0,373162	0,351434	-
1LFB	641	0,272119	0,255591	0,240853	0,226749	-

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1MBN	1216	0,547986	0,509842	0,478004	-	-
1N4W	8616	7,339331	6,654907	6,071414	5,617249	6,518732
1RGS	2015	0,988709	0,929969	0,878625	-	-
1RWH	5646	3,344701	3,144561	2,96457	-	-
2MSJ	480	0,2	0,189732	0,175774	0,165803	-
3B34	7479	4,916706	4,626036	4,381025	-	-
8DRH	329	0,133515	0,127355	0,123167	0,131956	0,228595

TABELA 41: RMSD - MLMQL - Esparso Re = 1e-06

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	2,28E-05	3,19E-05	5,68E-05	5,60E+00	1,89E+16
1CEU	854	1,41E-05	2,15E-05	3,70E-04	6,82E+00	9,59E+20
1D8V	4208	2,95E-05	3,65E-05	3,11E+00	3,29E+02	8,37E+24
1F39	1534	9,70E-02	1,50E-02	5,80E+00	7,52E+04	-
1KVX	954	7,33E-05	3,60E-04	1,35E-01	1,08E+01	-
1LFB	641	2,73E-04	5,63E-03	7,22E-02	1,24E+01	-
1MBN	1216	1,11E-04	1,16E-03	5,76E+00	-	-
1N4W	8616	1,23E-04	3,15E-04	9,64E+00	3,75E+01	1,38E+56
1RGS	2015	1,98E-03	1,93E+00	1,17E+01	-	-
1RWH	5646	8,64E-03	3,02E+00	2,23E+01	-	-
2MSJ	480	6,93E-05	1,84E-04	4,30E-04	8,62E+00	-
3B34	7479	2,10E-01	1,36E+01	1,98E+01	-	-
8DRH	329	1,33E-03	1,63E-02	4,72E+00	1,80E+01	2,20E+02

TABELA 42: Tempo - MLMQL - Esparso Re = 1e-06

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,244853	0,30951	0,29381	0,278196	0,374516
1CEU	854	0,412012	0,387273	0,358313	0,327384	0,419654
1D8V	4208	2,721765	2,460123	2,240982	2,065484	2,261021
1F39	1534	0,744069	0,695415	0,680592	0,670719	-
1KVX	954	0,427508	0,398669	0,373867	0,34481	-
1LFB	641	0,271641	0,255573	0,240739	0,225172	-
1MBN	1216	0,550025	0,513675	0,482743	-	-
1N4W	8616	7,321675	6,627854	6,072962	5,630076	6,383238
1RGS	2015	0,986938	0,925357	0,875631	-	-
1RWH	5646	3,341004	3,136827	2,956699	-	-
2MSJ	480	0,200421	0,187786	0,176255	0,165625	-
3B34	7479	4,908649	4,621291	4,364508	-	-
8DRH	329	0,132205	0,125868	0,122663	0,132312	0,229062

TABELA 43: RMSD - MLMQL - Esparso Re = 1e-04

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	1,97E-03	3,32E-03	5,19E-03	1,61E+01	2,58E+07
1CEU	854	1,54E-03	2,51E-03	1,29E-02	1,04E+01	1,45E+11
1D8V	4208	3,47E-03	4,02E-03	4,32E+00	1,47E+01	5,70E+17
1F39	1534	7,43E+00	2,10E+00	6,81E+00	1,02E+09	-
1KVX	954	1,47E-02	2,33E-01	4,80E+00	1,09E+01	-
1LFB	641	1,94E-02	1,34E+00	1,10E+01	3,87E+03	-
1MBN	1216	7,48E-03	3,81E-01	1,48E+01	-	-
1N4W	8616	1,21E-02	1,80E-02	1,43E+01	1,61E+02	3,92E+36
1RGS	2015	7,82E-02	1,10E+01	1,33E+01	-	-
1RWH	5646	1,17E+00	1,29E+01	2,31E+01	-	-
2MSJ	480	4,86E-03	5,62E-03	8,47E-02	9,75E+00	-
3B34	7479	7,75E+00	1,78E+01	2,42E+01	-	-
8DRH	329	6,69E-01	3,33E+00	7,77E+00	8,14E+01	9,09E+03

TABELA 44: Tempo - MLMQL - Esparso Re = 1e-04

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,240238	0,325368	0,290254	1,142029	0,387293
1CEU	854	0,412323	0,381595	0,379057	0,329154	0,41937
1D8V	4208	2,760234	2,472592	2,238018	2,052212	2,285033
1F39	1534	0,740325	0,691019	0,680801	0,656382	-
1KVX	954	0,419062	0,391442	0,365322	0,352703	-
1LFB	641	0,268849	0,253455	0,237983	0,225532	-
1MBN	1216	0,549166	0,512366	0,481539	-	-
1N4W	8616	7,308276	6,635103	6,046417	5,567752	6,456951
1RGS	2015	0,931359	0,928929	0,870857	-	-
1RWH	5646	3,330463	3,138106	2,953891	-	-
2MSJ	480	0,199551	0,188164	0,176097	0,16592	-
3B34	7479	4,886025	4,630085	4,368968	-	-
8DRH	329	0,132205	0,126664	0,122	0,132837	0,232376

TABELA 45: RMSD - MLMQL - Esparso Re = 1e-02

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	2,75E-01	2,29E-01	7,24E-01	7,09E+00	3,58E+18
1CEU	854	1,79E-01	2,57E-01	8,26E-01	1,60E+01	7,58E+10
1D8V	4208	2,79E-01	6,92E-01	9,62E+00	4,30E+01	-
1F39	1534	8,26E+00	7,71E+00	1,06E+01	3,81E+12	-
1KVX	954	3,73E+00	1,00E+01	5,24E+01	1,10E+01	-
1LFB	641	1,15E+00	5,94E+00	9,10E+00	1,15E+01	-

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1MBN	1216	6,99E-01	5,35E+00	1,15E+01	-	-
1N4W	8616	9,05E-01	1,88E+00	1,45E+01	3,90E+01	4,47E+38
1RGS	2015	6,41E+00	1,37E+01	1,78E+01	-	-
1RWH	5646	2,08E+01	2,40E+01	9,47E+03	-	-
2MSJ	480	6,29E-01	1,28E+00	6,95E+00	1,06E+01	-
3B34	7479	2,32E+01	2,46E+01	5,90E+02	-	-
8DRH	329	4,80E+00	7,46E+00	7,08E+00	1,07E+01	6,50E+51

TABELA 46: Tempo - MLMQL - Esparso Re = 1e-02

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	0,250242	0,312002	0,301376	0,282749	0,388941
1CEU	854	0,416235	0,389154	0,365364	0,341719	0,431764
1D8V	4208	2,800949	2,477207	2,258095	2,084062	-
1F39	1534	0,751451	0,704191	0,682999	0,654494	-
1KVX	954	0,437048	0,400416	0,485198	0,355597	-
1LFB	641	0,273749	0,263087	0,255525	0,229898	-
1MBN	1216	0,553194	0,519482	0,498859	-	-
1N4W	8616	7,319417	6,666252	6,25743	5,66109	6,376675
1RGS	2015	0,99436	0,940244	0,874831	-	-
1RWH	5646	3,342447	3,156388	3,065631	-	-
2MSJ	480	0,200938	0,188482	0,18421	0,168609	-
3B34	7479	4,902934	4,641546	4,504641	-	-
8DRH	329	0,132811	0,137892	0,128811	0,134627	0,306849

Para analisar os resultados do MLMQL, seria mais justo compará-lo ao Método Linear para a Matriz de Distâncias Esparsas - ML na seção 3.3.1, pois ambos os métodos resolvem um sistema linear sem atualizar resultados anteriores a cada passo.

Uma desvantagem ao se utilizar o MLMQL é que a cada iteração o erro ao gerar cada átomo será maior, o que torna o método eficaz apenas para pequenas estruturas de proteínas.

5.2 MÉTODO LINEAR COM MÍNIMOS QUADRADOS NÃO-LINEAR

O primeiro problema encontrado ao tentar modificar o algoritmo MLMQL para que ele seja atualizado de modo a recalcular as coordenadas dos átomos utilizados, para que seja eficaz comparado com o método linear atualizado, é que seria necessário

conhecer todas as distâncias entre os l átomos. Estas distâncias dificilmente serão conhecidas considerando que a matriz de distâncias é esparsa e que em alguns casos o l encontrado pode ser bem grande.

Para evitar este problema, considerando que estes l átomos já foram determinados, é possível calcular a distância entre eles. Tornando necessário apenas ser dado a distância $d_{k,j}$, para $k, j = 1, ..., l$ átomos determinados.

Seguindo o mesmo modelo de problema da seção 5.1, mas neste caso com l átomos, é possível satisfazer o problema de distâncias proposto em (2.1), ou seja:

$$||x_k - x_j||^2 = d_{k,j}^2, \ para\ k, j = 1, ..., l.$$

Expandindo o problema, obtém-se um resultado análogo:

$$||x_k||^2 - 2x_k^T x_j + ||x_j||^2 = d_{k,j}^2, \ para\ k, j = 1, ..., l. \tag{5.3}$$

De acordo com o modelo utilizado para o método linear atualizado na subseção 4.1.1, sabe-se que independentemente da localização das coordenadas do átomo no espaço, a estrutura geral da proteína será a mesma ao calcular o erro pelo RMSD. Portanto, se o átomo procurado x_i' for colocado na origem $x_i' = (0,0,0)^T$, criando um novo sistema de referências para encontrar x_i, as seguintes igualdades serão válidas.

$$\begin{aligned} ||x_k||^2 &= ||x_k - x_i'||^2 = d_{k,i}^2, \\ ||x_j||^2 &= ||x_j - x_i'||^2 = d_{j,i}^2, \end{aligned} \tag{5.4}$$

onde, $d_{k,i}$ e $d_{j,i}$ são distâncias conhecidas dos átomos k e j ao átomo procurado.

Substituindo (5.4) em (5.3):

$$d_{k,i}^2 - 2x_k^T x_j + d_{j,i}^2 = d_{k,j}^2, \ para\ k, j = 1, ..., l,$$

que é equivalente a

$$x_k^T x_j = \frac{d_{k,i}^2 - d_{k,j}^2 + d_{j,i}^2}{2}, \ para\ k, j = 1, ..., l. \tag{5.5}$$

Desta forma é possível definir os seguintes conjuntos,

$$X = \{x_j,\ para\ j = 1,...,l\},$$

$$D = \left\{ \frac{d_{k,i}^2 - d_{k,j}^2 + d_{j,i}^2}{2}\ para\ k, j = 1,...l \right\}. \tag{5.6}$$

Tornando possível escrever o sistema (5.5) da seguinte forma:

$$XX^T = D.$$

De acordo com (Souto 2000), a matriz $X_{l,3}$ de posto 3, pode ser decomposta em valores singulares, tal que $X = U\Sigma V^T$, onde Σ é uma matriz diagonal de autovalores e U e V matrizes de autovetores são bases ortonormais, portanto,

$$D = XX^T = U\Sigma V^T(U\Sigma V^T)^T = U\Sigma V^T V\Sigma^T U^T = U(\Sigma\Sigma^T)U^T = (U\Sigma)(U\Sigma)^T,$$

onde

$$\Sigma\Sigma^T = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \sigma_2^2 & & & & \\ & & \sigma_3^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}.$$

Considerando que a partir de D deve-se definir X, é possível encontra-lo facilmente calculando a decomposição em valores singulares de D.

A matriz de novas coordenadas é definida da seguinte forma:

$$X = U(k, j)\Sigma(j, j)^{\frac{1}{2}}, \tag{5.7}$$

onde $k = 1,...,l$ e $j = 1,2$ e 3, pois como as distâncias originais possuem certo erro, a matriz D pode não ter posto 3, o que não afeta muito o resultado final, mas pode resultar em mais de três autovalores e autovetores.

5.2.1 ALGORITMO - MÉTODO LINEAR COM MÍNIMOS QUADRADOS NÃO-LINEAR

Dados de entrada: $D = [d_{k,j}]$ onde $k, j = 1, ..., n$

$F = \{x_j \mid j = 1, ..., 4\}$, átomos fixados na subseção 3.1.2

$U = \{n-4$ átomos não fixados$\}$

Enquanto $U \neq \emptyset$

Para $x_i \in U$ onde $i = 5, ..., n$.

Encontre todos os átomos $x_j \in F$, onde $j = 1, ..., l$, com distâncias a x_i

Calcule as distâncias desconhecidas $||x_k - x_j|| \in F$, onde $k, j = 1, ..., l$

Calcule D, como em (5.6)

Calcule X como em (5.7)

Reposicione X em $x_j \in F$, onde $j = 1, ..., l, i$, como na seção 4.1

Fim

5.2.2 RESULTADOS COMPUTACIONAIS - MÉTODO LINEAR COM MÍNIMOS QUADRADOS NÃO-LINEAR

Abaixo seguem resultados do cálculo do erro absoluto e o tempo que cada proteína demorou para ser calculada através do Método linear com Mínimos Quadrados Não-Linear - MLMQNL.

Os testes foram realizados para distâncias esparsas exatas e inexatas, com as seguintes esparsidades na matriz de distâncias D: 8$\mathring{A}$, 7$\mathring{A}$, 6$\mathring{A}$, 5$\mathring{A}$ e 4$\mathring{A}$.

Para esparsidades menores $4\mathring{A}$ não foi possível realizar testes devido ao fato de não haverem distâncias o suficiente relativas ao átomo x_i procurado.

TABELA 47: RMSD - MLMQNL - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	2,17E-09	1,76E-10	2,97E-13	4,84E-14	4,64E-13
1CEU	854	4,44E-02	2,43E-07	4,35E-11	1,63E-13	1,47E-12
1D8V	4208	-	-	-	5,45E-10	6,02E-04
1F39	1534	4,47E+28	8,95E-08	1,51E-13	5,40E-13	-
1KVX	954	9,61E-04	2,46E-10	2,75E-13	1,20E-13	-
1LFB	641	4,43E-09	5,61E-13	1,13E-13	9,28E-14	-
1MBN	1216	3,98E+03	2,76E-10	1,45E-13	-	-
1N4W	8616	-	-	-	1,31E+150	2,61E-07
1RGS	2015	7,01E+36	5,05E-07	4,43E-12	-	-
1RWH	5646	-	-	1,38E-05	-	-
2MSJ	480	1,11E-09	1,87E-12	6,96E-14	7,56E-14	-
3B34	7479	-	-	2,68E+22	-	-
8DRH	329	7,62E-12	1,17E-12	6,25E-13	1,20E-13	1,34E-12

TABELA 48: Tempo - MLMQNL - Esparso e Exato

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	1,9944	1,1861	1,5270	0,9144	1,2258
1CEU	854	6,6421	2,6733	2,7769	2,1136	2,3678
1D8V	4208	-	-	-	170,5385	161,2988
1F39	1534	11,8141	8,853	8,2848	7,9146	-
1KVX	954	3,4866	2,9634	2,9853	2,3738	-
1LFB	641	2,1353	1,3891	1,5621	1,0630	-
1MBN	1216	9,3183	4,931	5,2954	-	-
1N4W	8616	-	-	-	1430,9261	1325,0007
1RGS	2015	17,1221	17,3680	15,5134	-	-
1RWH	5646	-	-	401,1939	-	-
2MSJ	480	0,9645	0,8304	0,7428	0,6514	-
3B34	7479	-	-	913,7067	-	-
8DRH	329	0,7377	0,7350	0,4409	0,4199	0,5607

TABELA 49: RMSD - MLMQNL - Esparso Re = 1e-08

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	5,82E-01	1,07E-04	2,00E-06	1,81E-07	5,25E-05
1CEU	854	3,10E+39	1,98E+23	5,38E-05	6,91E-07	5,36E-05
1D8V	4208	-	-	-	1,69E+15	9,41E+02
1F39	1534	5,17E+46	7,47E+02	3,67E-06	3,55E-05	-
1KVX	954	9,32E+18	5,77E-04	1,90E-06	5,88E-06	-
1LFB	641	2,39E-03	2,88E-06	3,19E-07	6,80E-07	-

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1MBN	1216	9,83E+28	2,68E-04	9,17E-07	-	-
1N4W	8616	-	-	-	-	3,69E+06
1RGS	2015	1,87E+63	5,04E+16	1,07E-05	-	-
1RWH	5646	-	-	5,45E+33	-	-
2MSJ	480	1,48E-02	5,25E-06	9,70E-07	1,59E-06	-
3B34	7479	-	-	2,20E+108	-	-
8DRH	329	1,45E-05	1,16E-06	4,04E-05	5,68E-06	1,48E-04

TABELA 50: Tempo - MLMQNL - Esparso Re = 1e-08

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	2,1638	-	-	-	-
1AQR	524	1,4012	2,3405	1,2218	1,1854	1,1819
1CEU	854	3,4099	3,3110	2,6809	2,6893	2,4564
1D8V	4208	83,4939	-	-	179,4080	153,275
1F39	1534	8,4747	10,6568	10,1280	9,6180	-
1KVX	954	3,3756	3,7108	2,3150	2,0436	-
1LFB	641	1,5703	0,9589	1,0281	0,8504	-
1MBN	1216	5,6829	6,2267	6,8447	-	-
1N4W	8616	334,0162	-	-	-	1305,8002
1RGS	2015	16,4993	34,1193	24,3210	-	-
1RWH	5646	233,1268	-	388,4762	-	-
2MSJ	480	1,4218	0,7455	0,4923	0,3953	-
3B34	7479	352,1601	-	860,5417	-	-
8DRH	329	0,6399	0,4015	0,3483	0,3352	0,4948

TABELA 51: RMSD - MLMQNL - Esparso Re = 1e-06

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	2,40E+14	-	-	-	-
1AQR	524	1,49E+44	1,96E-02	5,43E-04	3,38E-05	5,00E-01
1CEU	854	-	6,55E+31	1,67E-02	3,74E-05	4,87E-03
1D8V	4208	6,69E+55	-	-	1,56E+67	1,02E+03
1F39	1534	1,03E+21	1,05E+04	2,77E-04	1,47E-03	-
1KVX	954	2,83E+03	5,99E-02	7,86E-04	1,85E-03	-
1LFB	641	4,57E+35	3,22E-04	3,60E-05	7,28E-05	-
1MBN	1216	-	5,62E-01	1,26E-04	-	-
1N4W	8616	8,63E+74	-	-	-	4,82E+07
1RGS	2015	-	1,11E+27	9,69E-04	-	-
1RWH	5646	1,31E-01	-	3,59E+48	-	-
2MSJ	480	-	4,43E-04	3,90E-05	6,26E-05	-
3B34	7479	1,71E-02	-	6,42E+134	-	-
8DRH	329	6,50E-01	7,07E-04	5,38E-03	4,60E-04	8,00E-03

TABELA 52: Tempo - MLMQNL - Esparso Re = 1e-06

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	2,3598	-	-	-	-
1AQR	524	1,4756	1,4923	1,4110	1,0747	1,1337
1CEU	854	-	3,9502	3,1066	2,3877	2,4811
1D8V	4208	69,8130	-	-	177,1481	154,2588
1F39	1534	17,7405	17,0848	13,4613	9,0410	-
1KVX	954	2,6977	2,3940	2,0011	1,9169	-
1LFB	641	1,2226	1,0313	0,8178	0,7700	-
1MBN	1216	-	9,3098	8,2248	-	-
1N4W	8616	334,2067	-	-	-	1308,3001
1RGS	2015	-	34,2669	33,7840	-	-
1RWH	5646	191,3567	-	399,4525	-	-
2MSJ	480	-	0,7417	0,4809	0,4023	-
3B34	7479	333,1213	-	866,7538	-	-
8DRH	329	0,5069	0,4221	0,3520	0,3383	0,4890

TABELA 53: RMSD - MLMQNL - Esparso Re = 1e-04

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	1,71E+19	4,07E+02	1,34E-01	5,58E-03	6,49E+00
1CEU	854	2,60E+53	1,61E+44	7,16E+29	6,72E-03	1,97E+01
1D8V	4208	-	-	-	1,13E+66	1,23E+04
1F39	1534	2,25E+63	9,40E+18	2,19E-02	2,17E-02	-
1KVX	954	5,48E+26	4,19E+07	1,76E-02	3,63E-02	-
1LFB	641	2,48E+05	1,57E-02	5,94E-03	4,77E-02	-
1MBN	1216	2,99E+51	7,39E+32	9,97E-03	-	-
1N4W	8616	-	-	-	-	1,44E+08
1RGS	2015	7,62E+87	9,95E+25	1,38E+08	-	-
1RWH	5646	-	-	9,83E+85	-	-
2MSJ	480	2,19E+15	3,14E-02	2,68E-03	3,74E-03	-
3B34	7479	-	-	6,60E+150	-	-
8DRH	329	1,19E+02	1,02E-02	6,34E-02	3,93E-02	4,95E+00

TABELA 54: Tempo - MLMQNL - Esparso Re = 1e-04

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	1,4808	1,6180	1,4538	1,0920	1,4170
1CEU	854	3,3270	4,6613	3,2870	2,3633	2,6560
1D8V	4208	-	-	-	164,3655	166,6646
1F39	1534	10,6862	18,3500	10,7611	8,3420	-
1KVX	954	2,6160	2,8384	2,2957	1,8664	-

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1LFB	641	1,1413	1,2435	1,0132	0,7462	-
1MBN	1216	7,7379	9,5476	7,4287	-	-
1N4W	8616	-	-	-	-	1359,2279
1RGS	2015	37,6623	35,2856	33,532	-	-
1RWH	5646	-	-	388,4785	-	-
2MSJ	480	0,7207	0,5706	0,4772	0,4838	-
3B34	7479	-	-	849,1880	-	-
8DRH	329	0,5341	0,4119	0,3541	0,3342	0,4928

TABELA 55: RMSD - MLMQNL - Esparso Re = 1e-02

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	4,42E+27	1,63E+18	1,75E+10	5,07E+01	6,22E+00
1CEU	854	1,81E+59	3,65E+45	1,03E+28	3,63E+10	1,13E+01
1D8V	4208	-	-	-	1,18E+82	6,24E+03
1F39	1534	1,15E+74	1,45E+48	3,31E+15	1,67E+02	-
1KVX	954	1,10E+42	7,80E+05	1,02E+10	1,15E+02	-
1LFB	641	4,07E+12	8,38E+07	5,18E+03	4,43E+01	-
1MBN	1216	8,03E+68	5,46E+34	2,52E+16	-	-
1N4W	8616	-	-	-	-	1,97E+08
1RGS	2015	1,77E+99	3,08E+60	1,16E+21	-	-
1RWH	5646	-	-	2,59E+101	-	-
2MSJ	480	6,04E+26	6,61E+13	1,67E+05	5,83E+01	-
3B34	7479	-	-	-	-	-
8DRH	329	1,73E+09	1,22E+01	2,88E+02	3,02E+01	7,02E+00

TABELA 56: Tempo - MLMQNL - Esparso Re = 1e-02

Proteína	Átomos	8 $\mathring{A}$	7 $\mathring{A}$	6 $\mathring{A}$	5 $\mathring{A}$	4 $\mathring{A}$
1A1D	146	-	-	-	-	-
1AQR	524	1,4893	2,1223	1,3798	1,1788	1,1677
1CEU	854	3,3261	3,6943	3,6162	2,6909	2,5876
1D8V	4208	-	-	-	167,9607	156,2637
1F39	1534	11,1503	7,2919	10,2013	8,4122	-
1KVX	954	2,6140	30,3216	2,1655	1,7644	-
1LFB	641	11,0625	1,2598	0,8681	0,6878	-
1MBN	1216	9,1895	5,1845	6,8187	-	-
1N4W	8616	-	-	-	-	1294,0432
1RGS	2015	40,0898	38,5282	29,4735	-	-
1RWH	5646	-	-	399,2192	-	-
2MSJ	480	0,8622	0,8008	0,4892	0,3886	-
3B34	7479	-	-	-	-	-
8DRH	329	0,5996	1,5946	0,3513	0,3223	0,4836

Comparando os resultados do MLMQL ao MLMQNL, pode-se notar que o MLMQNL obteve resultados melhores que o MLMQL, tanto no caso exato como inexato para esparsidades menores ou iguais a $6\mathring{A}$ (em parte das proteínas), $5\mathring{A}$ e $4\mathring{A}$. No entanto, em alguns desses casos o algoritmo MLMQNL não conseguiu calcular a proteína e ao invés de obter um resultado aproximado o erro foi gigantesco. Em relação ao tempo o MLMQNL perdeu em todos os casos, visto que seu algoritmo é mais elaborado e precisa executar a decomposição em valores singulares de uma matriz, que pode ter um posto bem grande, dificultando o cálculo.

6 OTIMIZAÇÃO NO CÁLCULO DE ESTRUTURAS DE PROTEÍNAS

Neste capítulo será apresentado um método iterativo, que torna possível resolver o problema geométrico de distâncias através de minimização.

6.1 FORMULAÇÃO DO PROBLEMA DE OTIMIZAÇÃO PELO MÉTODO DOS MÍNIMOS QUADRADOS

No artigo (Coope 2000), é discutida uma maneira de calcular o problema geométrico de distâncias, que também é descrito como um problema de intercessão de esferas em $\mathbb{R}^n$. Neste artigo, são discutidas diversas maneiras para resolver o problema de intercessão de esferas com apenas n pontos, no entanto não existe na literatura uma aplicação do problema ao caso do problema geométrico.

Neste trabalho será feita uma modelagem, à partir de (Coope 2000), para que o problema geométrico de distâncias recaia no caso do cálculo de distâncias de proteínas, ou seja, permitindo que seja possível encontrar a estrutura de uma proteína no $\mathbb{R}^3$ com apenas três átomos.

Partindo do mesmo problema geométrico de distâncias em (2.1), é possível fazer uma formulação de (Coope 2000) por mínimos quadrados não-linear, de modo a minimizar o erro ao determinar cada átomo de acordo com (Gander, Golub e Strebel 1996):

$$min\ f(x_i) = \sum_{j=1}^{3} ||S_i||^2,\ \ para\ \ S_i = \{x_i \in \mathbb{R}^3, \|x_i - x_j\| = d_{i,j}\},\ x_i \in \mathbb{R}^3, j = 1,2,3.$$

onde x_j é o centro de cada esfera de raio $d_{i,j}$.

A função a ser minimizada, também pode ser escrita como,

$$min\ f(x_i) = \sum_{j=1}^{3} |\|x_i - x_j\| - d_{i,j}|^2,\ \ para\ \ j = 1,2,3. \tag{6.1}$$

Como $\|x_i - x_j\|$ é uma constante, em (6.1) é equivalente afirmar

$$min\ f(x_i) = \sum_{j=1}^{3} (\|x_i - x_j\| - d_{i,j})^2, \quad para \quad j = 1,2,3. \tag{6.2}$$

Também pode-se dizer que (6.2) é equivalente a encontrar

$$\tilde{x} \in Argmin \sum_{j=1}^{3} d^2(x, S_i),$$

onde $d^2(x,S_i)$, é a distância de x ao conjunto S_i, ou seja, $d^2(x,S_i) = \inf_{y \in S_i} |x-y|$. Isto significa que a função $f(x_i)$ minimiza o erro, sendo x_i o átomo que minimiza as distâncias mesmo que não exista intercessão entre as esferas.

A função $f(x_i)$ será minimizada com a ajuda do método de Newton descrito em (Ribeiro e Karas 2014) e para minimizar $f(x_i)$ será necessário que $\nabla^2 f(x_i)$ seja definida positiva. Como $x_i \in \mathbb{R}^3$, então $f(x_i)$ será expandida para calcular o gradiente com a ajuda das variáveis auxiliares β_i e α_i.

$$f(x_i) = \sum_{j=1}^{3} \underbrace{((\underbrace{(x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2 + (x_{i,3} - x_{j,3})^2}_{\alpha_i})^{\frac{1}{2}} - d_{i,j})^2}_{\beta_i}.$$

melhor dizendo

$$f(x_i) = \sum_{j=1}^{3} \beta_i^2,$$

com $\beta_i = \alpha_i^{\frac{1}{2}} - d_{i,j}$ e $\alpha_i^{\frac{1}{2}} = ((x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2 + (x_{i,3} - x_{j,3})^2)^{\frac{1}{2}} = ||x_i - x_j||$.

Portanto o gradiente de $f(x_i)$ pode ser escrito utilizando a regra da cadeia descrita em (Guidorizzi 2001):

$$\nabla f(x_i) = \sum_{j=1}^{3} 2\beta_i \frac{\partial \beta_i}{\partial x_{i,k}},$$

para $k = 1,2,3$, onde

$$\frac{\partial \beta_i}{\partial x_{i,k}} = \frac{1}{2\alpha_i^{\frac{1}{2}}} \frac{\partial \alpha_i}{\partial x_{i,k}} \; e$$
$$\frac{\partial \alpha_i}{\partial x_{i,k}} = 2((x_{i,1} - x_{j,1}),(x_{i,2} - x_{j,2}),(x_{i,3} - x_{j,3})) = 2(x_i - x_j).$$

Substituindo $\frac{\partial \beta_i}{\partial x_{i,k}}$ e $\frac{\partial \alpha_i}{\partial x_{i,k}}$ em $\nabla f(x_i)$, obtém-se o seguinte resultado:

$$\nabla f(x_i) = \sum_{j=1}^{3} 2\beta_i \frac{2(x_i - x_j)}{2\alpha_i^{\frac{1}{2}}}.$$

Agora basta retornar para as variáveis originais

$$\begin{aligned} \nabla f(x_i) &= \sum_{j=1}^{3} 2(\alpha_i^{\frac{1}{2}} - d_{i,j}) \frac{(x_i - x_j)}{\alpha_i^{\frac{1}{2}}}, \\ \nabla f(x_i) &= \sum_{j=1}^{3} 2(||x_i - x_j|| - d_{i,j}) \frac{(x_i - x_j)}{||x_i - x_j||}. \end{aligned} \tag{6.3}$$

A equação (6.3) também pode ser simplificada da seguinte maneira:

$$\nabla f(x_i) = 2 \sum_{j=1}^{3} (1 - \frac{d_{i,j}}{||x_i - x_j||})(x_i - x_j). \tag{6.4}$$

Para calcular a Hessiana também serão utilizadas variáveis auxiliares,

$$\nabla f(x_i) = 2 \sum_{j=1}^{3} \underbrace{(1 - \frac{d_{i,j}}{||x_i - x_j||})}_{\gamma_i} \underbrace{(x_i - x_j)}_{\delta_i},$$

onde $\delta_i = (x_i - x_j)$ e $\gamma_i = (1 - \frac{d_{i,j}}{||x_i - x_j||}) = (1 - d_{i,j}\alpha_i^{\frac{-1}{2}})$.

Neste caso será usada a regra do produto, também conhecida como lei de Leibniz

$$\nabla^2 f(x_i) = 2 \sum_{j=1}^{3} \left[\frac{\partial \gamma_i}{\partial x_{i,k}} \delta_i + \gamma_i \frac{\partial \delta_i}{\partial x_{i,k}} \right],$$

onde $k = 1,2,3$, para as seguintes derivadas parciais

$$\frac{\partial \delta_i}{\partial x_{i,k}} = I,$$
$$\frac{\partial \gamma_i}{\partial x_{i,k}} = (\frac{1}{2} d_{i,j} \alpha_i^{\frac{-3}{2}}) \frac{\partial \alpha_i}{\partial x_{i,k}} = 2(\frac{1}{2} d_{i,j} \alpha_i^{\frac{-3}{2}})(x_i - x_j) = \frac{d_{i,j}(x_i - x_j)}{||x_i - x_j||^3},$$

concluindo que

$$\nabla^2 f(x_i) = 2\sum_{j=1}^{3}\left[\frac{d_{i,j}(x_i - x_j)(x_i - x_j)^T}{||x_i - x_j||^3} + (1 - \frac{d_{i,j}}{||x_i - x_j||})I\right]. \tag{6.5}$$

A hessiana dada em (6.5) pode não ser definida positiva para valores de j tal que $||x_i - x_j|| < d_{i,j}$. Existem duas opções para contornar esta situação.

A primeira opção é forçar o algoritmo a escolher novos átomos sempre que $||x_i - x_j|| < d_{i,j}$ ocorrer, o que tornaria o algoritmo final mais lento e poderia ocasionar na impossibilidade de encontrar a solução dependendo da esparsidade na matriz de distâncias.

A outra opção, que será usada neste trabalho, é fazer uma pequena modificação na matriz Hessiana $\nabla^2 f(x_i)$ para uma nova matriz B_i. Excluindo a parte negativa da equação (6.5), $I(1 - \frac{d_{i,j}}{||x_i - x_j||})$ quando a mesma atingir valores negativos, implicando no uso de um método quase-Newton, visto em (Fausett 1999). Esta exclusão não terá grandes efeitos no resultado final, visto que o algoritmo alcança a melhor solução quando $||x_i - x_j|| = d_{i,j}$.

Portanto, a matriz B_i será formulada da seguinte maneira:

Para cada x_i procurado, define-se o conjunto $L = \{l \in \{1,2,3\} | (1 - \frac{d_{i,l}}{||x_i - x_l||}) > 0\}$,

$$B_i = 2\left[\sum_{j=1}^{3}\frac{d_{i,j}(x_i - x_j)(x_i - x_j)^T}{||x_i - x_j||^3} + \sum_{l\in L}(1 - \frac{d_{i,l}}{||x_i - x_l||})I\right]. \tag{6.6}$$

É preciso verificar primeiramente que esta nova matriz B_i é definida positiva, para que $f(x_i)$ seja convexa.

Seja $v \in \mathbb{R}^3 \backslash \{0\}$

$$v^T B_i v = 2\left[\sum_{j=1}^{3}\frac{d_{i,j}v^T(x_i - x_j)(x_i - x_j)^T v}{||x_i - x_j||^3} + \sum_{l\in L}(1 - \frac{d_{i,l}}{||x_i - x_l||})v^t I v\right]. \tag{6.7}$$

Na primeira parcela da equação, tem-se

$$\frac{d_{i,j}}{||x_i - x_j||^3} > 0 \; e \; v^T(x_i - x_j)(x_i - x_j)^T v = (v^T(x_i - x_j))^2 \geq 0,$$

sendo que $(v^T(x_i - x_j))^2 = 0$ somente quando v é ortogonal a $(x_i - x_j)$, ou seja, a primeira parcela da equação B_i, é sempre positiva. Na segunda parcela da equação $(1 - \frac{d_{i,l}}{||x_i - x_l||}) > 0$ por definição de B_i e $v^t I v = v^t v = ||v||^2$ que é uma constante positiva por definição de produto interno em (Anton e Rorres 2012). Pode-se afirmar que $v^T B_i v > 0$, ou seja, B_i é definida positiva.

É importante observar, que $(x_i - x_j)(x_i - x_j)^T$ é uma matriz simétrica, tornando B_i uma matriz simétrica.

6.1.1 MÉTODO QUASE-NEWTON

Neste algoritmo será aceita uma tolerância para a convergência do gradiente, visto que em casos reais é difícil alcançar um resultado exato e também será definido um número de iterações máximas. O algoritmo terá o tamanho do passo fixo para executar o mínimo possível de operações. Além disso, como mostrado em parágrafos precedentes, o problema a ser resolvido é convexo.

Para cada átomo x_i:

Dado $x_i^0 \in \mathbb{R}^3$, $tol = 1e-16$, $iti_{max} = 1000$ e $k = 0$.

Enquanto $((||\nabla f(x_i^k)|| > tol) \;\; e \;\; (k < iti_{max}))$

Defina $B_i^k d^k = -\nabla f(x_i^k)$

Faça $x_i^{k+1} = x_i^k + d^k$

$k = k + 1$

Fim

6.2 CONVERGÊNCIA

É possível mostrar a convergência global do algoritmo da seção 6.1, partindo da aproximação de Taylor de primeira ordem

$$\begin{aligned} f(x_i^{k+1}) &= f(x_i^k) + \nabla f(x_i^k)^t d_k + O(||d_k||^2), \\ &= f(x_i^k) - \nabla f(x_i^k)^t (B_i^k)^{-1} \nabla f(x_i^k) + O(||d_k||^2). \end{aligned} \tag{6.8}$$

Considerando que o termo $O(||d_k||^2)$ é pequeno comparado com o termo de primeira ordem, a cada passo do método de quase-Newton é diminuindo o seguinte valor da função objetivo,

$$f(x_i^{k+1}) - f(x_i^k) \cong -\nabla f(x_i^k)^t (B_i^k)^{-1} \nabla f(x_i^k). \tag{6.9}$$

Agora é necessário limitar $(B_i^k)^{-1}$ inferiormente. Seja $v \in \mathbb{R}^3$ e B_i^k uma matriz simétrica e positiva definida com λ_1 e λ_n menor e maior autovalores de B_i^k, então por (Nocedal e Wright 2006) vale o seguinte:

$$\frac{||v||^2}{\lambda_1} \geq v^t (B_i^k)^{-1} v \geq \frac{||v||^2}{\lambda_n}.$$

Como a intenção é limitar inferiormente e considerando de λ_n é um escalar positivo, existe $M > 0$ tal que $M \geq \lambda_n$. Concluindo,

$$v^t (B_i^k)^{-1} v \geq \frac{||v||^2}{\lambda_n} \geq \frac{||v||^2}{M}.$$

Substituindo este resultado em (6.9), com $\nabla f(x_i^k)$ no lugar de v

$$f(x_i^{k+1}) - f(x_i^k) \leq -\frac{||\nabla f(x_i^k)||^2}{M}. \tag{6.10}$$

Agora é possível concluir que a cada passo o decrescimento da função objetivo é proporcional a $||\nabla f(x_i^k)||^2$ e que a sequência x_i^k só chegará a um ponto de acumulação x_i se $\nabla f(x_i^k) = 0$.

Para encontrar este limitante M para B_i^k que não dependa de k será usada a hi-

pótese de que existe uma constante $\delta > 0$ tal que $||x_i^k - x_j|| > \delta$, para todo centro de esfera x_j e para todo k. Desta maneira se pode garantir que nenhuma subsequencia $\{x_i^k\}$ seja um centro x_j, evitando que eles se tornem pontos de acumulação.

Na prática esta hipótese não atrapalha o algoritmo, pois na solução final nenhum dos átomos se sobrepõem apesar de estarem bem próximos. O único desafio seria escolher um átomo inicial x_i^0 para o método.

Partindo da equação (6.7), para x_i^k tem-se

$$v^T B_i^k v = 2\left[\sum_{j=1}^{3} \frac{d_{i,j}(v^T(x_i^k - x_j))^2}{||x_i^k - x_j||^3} + \sum_{l\in L}(1 - \frac{d_{i,l}}{||x_i^k - x_l||})v^t I v\right].$$

Pela desigualdade de Cauchy-Schwarz, pode-se dizer que $|v^T(x_i^k - x_j)| \leq ||v||||x_i^k - x_j||$, portanto

$$\begin{aligned} v^T B_i^k v &\leq 2\left[\sum_{j=1}^{3} \frac{d_{i,j}||x_i^k - x_j||^2}{||x_i^k - x_j||^3}||v||^2 + \sum_{l\in L}(1 - \frac{d_{i,l}}{||x_i^k - x_l||})||v||^2\right], \\ &= 2\left[\sum_{j=1}^{3} \frac{d_{i,j}}{||x_i^k - x_j||} + \sum_{l\in L}(1 - \frac{d_{i,l}}{||x_i^k - x_l||})\right]||v||^2, \end{aligned}$$

considerando $|L|$ o número de elementos do conjunto L,

$$v^T B_i^k v \leq 2\sum_{l\notin L}(\frac{d_{i,l}}{||x_i^k - x_l||} + |L|)||v||^2. \tag{6.11}$$

Quando $l \notin L$,

$$1 - \frac{d_{i,l}}{||x_i^k - x_l||} \leq 0,$$

o que implica que

$$1 \leq \frac{d_{i,l}}{||x_i^k - x_l||}.$$

Como $||x_i^k - x_l|| > \delta$, pode-se concluir que

$$\frac{d_{i,l}}{||x_i^k - x_l||} \leq \frac{d_{i,l}}{\delta}. \tag{6.12}$$

Substituindo (6.12) em (6.11)

$$v^T B_i^k v \leq 2\sum_{l\notin L}(\frac{d_{i,l}}{\delta}+|L|)||v||^2 \leq 2(3-|L|)(\max_{l\notin L}\frac{\{d_{i,l}\}}{\delta}+|J|)||v||^2.$$

Escolhendo $M = 2(3-|L|)(\max_{l\notin L}\frac{\{d_{i,l}\}}{\delta}+|J|)$, é possível concluir em (6.10) que $v^T B_i^k v \leq \frac{||v||^2}{M}$, onde M não depende da iteração k, como desejado. Encontrando assim, uma taxa de convergência proporcional a $||\nabla f(x_i^k)||^2$.

6.2.1 ALGORITMO - MÉTODO DE INTERCESSÃO DE ESFERAS POR MÍNIMOS QUADRADOS

Para implementação deste algoritmo serão usados os átomos iniciais descritos na subseção 3.1.2.

Lembrando que no $\mathbb{R}^3$ quando as distâncias são exatas e com apenas 3 esferas, serão encontrados:

- Dois átomos - o primeiro átomo $x_i = (u_i, v_i, w_i)$ e o outro átomo simétrico a x_i em relação ao eixo w, $x_i' = (u_i, v_i, -w_i)$, seção 4.2 e 4.3;
- Um átomo - quando $x_i = x_i'$.

Para o caso descrito neste capítulo considera-se que as distâncias podem ser inexatas, portanto existe a possibilidade de não haver interseção entre as esferas, como visto na seção 3.2. Caso este fato aconteça, o átomo será calculado de maneira aproximada, como calculado na seção 6.1. Para determinar qual destes átomos é o correto será utilizado um átomo auxiliar.

Este algoritmo será denominado "Método de Intercessão de Esferas - MIE".

Dados de entrada: $D = [d_{i,j}]$, para $i, j = 1, ..., n$

Calcule x_1, x_2, x_3 e x_4 como na subseção 3.1.2

Para $i = 5, ..., n$

Defina x_j para $j = 1, 2, 3$

Encontre x_i pelo método de quase-Newton na subseção 6.1.1

Defina x_i', o simétrico de x_i

Se $|||x_i' - x_4|| - d_{i,4}| < |||x_i - x_4|| - d_{i,4}|$, então $x_i = x_i'$

Fim

6.3 RESULTADOS COMPUTACIONAIS

Abaixo seguem resultados do cálculo do erro absoluto e o tempo que cada proteína demorou para ser calculada através do Método de Intercessão de Esferas. Os testes foram realizados para distâncias exatas e inexatas, sem esparsidade.

TABELA 57: Intercessão de Esferas - Exato

Proteína	Átomos	Tempo/s	RMSD
103D	772	8,794156	4,57E-012
104D	766	7,131478	1,27E-012
124D	508	5,176324	1,09E-012
132D	750	6,221582	2,46E-013
141D	527	5,308486	1,50E-012
1A1D	146	1,328488	7,07E-013
1A23	2952	26,968942	4,56E-012
1A84	758	6,299246	2,65E-013
1AIK	729	6,610109	1,15E-012
1AMB	438	3,888149	1,87E-013
1AMD	380	3,123115	1,06E-013
1AQR	524	4,581673	6,42E-014
1AX8	1003	9,5345	9,10E-013
1B5N	332	2,99068	1,04E-013
1BOM	700	6,626209	2,28E-012
1BQX	1166	10,827479	3,51E-013
1CEU	854	8,440836	8,57E-012
1D8V	4208	39,54798	1,49E-012
1F39	1534	14,052799	2,22E-012
1FS3	951	8,494328	1,52E-011
1FW5	332	3,291347	2,64E-012
1HAA	1310	11,478982	4,44E-013
1HIP	617	5,493331	8,27E-013
1HLL	540	4,715865	2,56E-013
1HMV	29592	331,462919	1,68E-02

Proteína	Átomos	Tempo/s	RMSD
1HOE	558	5,472115	3,91E-13
1HSM	1251	12,947444	6,04E-12
1ID7	189	1,72875	2,13E-14
1ITH	2126	18,069281	2,80E-13
1JAV	360	3,330265	2,08E-12
1JK2	1229	12,272117	3,37E-11
1KVX	954	8,614939	8,67E-13
1LFB	641	5,561595	1,63E-12
1M40	5712	56,793138	1,32E-10
1MBN	1216	10,972044	3,71E-13
1MEQ	405	3,445488	6,73E-14
1MQQ	5681	51,156905	5,20E-11
1N4W	8616	81,794051	1,78E-11
1PHT	811	7,143892	8,84E-13
1POA	914	8,100768	4,67E-13
1PTQ	402	3,727238	5,78E-12
1QS5	1295	11,799191	9,59E-13
1QSB	1293	12,587642	3,80E-12
1R7C	532	4,715299	1,02E-11
1RGS	2015	17,489943	7,58E-13
1RWH	5646	53,48187	7,15E-13
1SOL	353	3,101127	8,40E-14
1ULR	677	5,986311	1,63E-12
1VII	596	5,339511	1,81E-13
1VMP	1166	11,067138	8,30E-13
2CLJ	4189	36,820862	2,44E-12
2E7Z	7633	71,338099	4,12E-11
2EQL	1023	8,995791	7,48E-12
2MSJ	480	4,144325	1,42E-12
304D	159	1,363198	9,30E-14
3B34	7479	68,35644	1,41E-11
4MBA	1083	9,972793	2,10E-13
8DRH	329	3,114104	1,84E-13

TABELA 58: Intercessão de Esferas - Re = 1e-08

Proteína	Átomos	Tempo/s	RMSD
1A1D	146	1,362057	3,89E-05
1AQR	524	4,571065	3,42E-06
1CEU	854	8,300881	5,64E-04
1D8V	4208	38,802656	5,73E-05
1F39	1534	14,042896	8,34E-05
1KVX	954	9,030252	6,91E-05
1LFB	641	5,880904	6,03E-05

Proteína	Átomos	Tempo/s	RMSD
1MBN	1216	10,249549	1,61E-05
1N4W	8616	81,206669	7,97E-04
1RGS	2015	17,107179	6,09E-05
1RWH	5646	54,772791	3,55E-05
2MSJ	480	4,069279	1,25E-04
3B34	7479	68,702147	4,24E-04
8DRH	329	2,959446	1,14E-05

TABELA 59: Intercessão de Esferas - Re = 1e-06

Proteína	N^o átomos	Tempo/s	RMSD
1A1D	146	1,109018	1,38E-02
1AQR	524	4,696985	3,32E-04
1CEU	854	8,812656	2,77E-02
1D8V	4208	38,327498	4,07E-03
1F39	1534	13,682919	7,76E-03
1KVX	954	8,46216	3,93E-03
1LFB	641	5,73765	3,88E-03
1MBN	1216	11,212505	1,48E-03
1N4W	8616	82,169154	1,49E-02
1RGS	2015	17,090178	2,90E-03
1RWH	5646	52,945701	3,49E-03
2MSJ	480	4,3479	8,96E-03
3B34	7479	67,982374	1,76E-02
8DRH	329	3,013629	1,03E-03

TABELA 60: Intercessão de Esferas - Re = 1e-04

Proteína	Átomos	Tempo/s	RMSD
1A1D	146	1,314579	2,00E-01
1AQR	524	4,65488	3,23E-02
1CEU	854	9,409247	7,30E-01
1D8V	4208	38,2917	1,73E-01
1F39	1534	13,871326	2,71E-01
1KVX	954	9,245265	1,35E-01
1LFB	641	6,330713	1,72E-01
1MBN	1216	11,625352	9,75E-02
1N4W	8616	83,644292	5,64E-01
1RGS	2015	17,57554	2,22E-01
1RWH	5646	52,757441	3,47E-01
2MSJ	480	4,24172	2,47E-01
3B34	7479	69,167306	5,17E-01
8DRH	329	3,023981	1,15E-01

TABELA 61: Intercessão de Esferas - Re = 1e-02

Proteína	Átomos	Tempo/s	RMSD
1A1D	146	1,625986	3,89E+00
1AQR	524	4,759954	3,13E+00
1CEU	854	12,513915	1,22E+01
1D8V	4208	45,776241	1,16E+01
1F39	1534	18,516878	1,43E+01
1KVX	954	10,52655	9,50E+00
1LFB	641	6,81088	1,31E+01
1MBN	1216	12,917846	1,03E+01
1N4W	8616	112,277641	2,44E+01
1RGS	2015	21,747004	1,84E+01
1RWH	5646	68,376081	3,83E+01
2MSJ	480	5,102544	7,53E+00
3B34	7479	89,47193	2,43E+01
8DRH	329	3,811804	1,21E+01

Analisando estes resultados, pode-se notar que o MIE mostrou-se eficaz tanto para o caso exato como também o inexato. Conforme a tolerância é diminuída o tempo de processamento diminui, mas o erro obtido é proporcionalmente maior.

O resultado para o erro pode ser melhorado ao exigir uma tolerância mais alta para o método de Quase Newton, no entanto o tempo de processamento será mais alto. A vantagem de se exigir uma tolerância tão alta se da ao fato de que vários átomos são encontrados rapidamente garantindo $\nabla f(x_i^k) = 0$, mas alguns átomos não convergem e quanto mais próximos chegarem do átomo procurado menos afetam o resultado final.

7 COMPARAÇÃO DOS ALGORITMOS

Neste capítulo serão feitas comparações entre os algoritmos mostrados em capítulos anteriores para melhor análise dos resultados.

7.1 DISTÂNCIAS EXATAS - DE

Considerando a matriz de distâncias moleculares exatas, foram testados dois métodos diferentes, o método linear e o Método de Interseção de Esferas. Nestes testes foram definidas estruturas para proteínas de 103 a 29.592 átomos.

Na Tabela 62 são apresentados o menor e maior valor para o RMSD e tempo obtidos em cada método.

TABELA 62: Métodos testados para DE

Método	RMSD	Tempo/s
Método Linear	1E-14 - 1E-11	0,00005 - 0,5
Método de Interceção de esferas	1E-14 - 1E-02	1-331

Comparando os resultados da Tabela 62, pode-se concluir que o ML é mais eficaz que o MIE tanto no tempo de processamento quanto nos resultados, apresentando erros mais estáveis. Mesmo considerando que o MIE é menos eficaz, ele apresenta vantagem por ser capaz de controlar o tamanho do erro para cada x_i, apresentando em alguns casos melhor precisão que o método linear. A vantagem ao utilizar o MIE é que a cada passo do método o erro pode ser controlado, sem prejudicar o cálculo dos demais átomos. Já o ML usa uma fórmula fixa, impossibilitando a redução do erro a cada iteração.

7.2 DISTÂNCIAS INEXATAS - DI

Para o caso em que as distâncias são inexatas, os algoritmos testados são os mesmos do caso anterior, o Método Linear e o Método de Interseção de Esferas. As estruturas definidas possuem de 103 a 8.616 átomos.

Nas Tabelas a seguir são apresentados o menor e maior valor para o RMSD, RMSD aproximado para o método linear e tempo obtidos em cada método.

TABELA 63: Método Linear - DI

RE	RMSD	RMSD-Apr	Tempo/s
1E-08	1E-06 - 1E-05	1E-05 - 1E-04	0,0005 - 0,005
1E-06	1E-04 - 1E-03	1E-03 - 1E-02	0,0003 - 0,006
1E-04	1E-02 - 1E-01	1E-01 - 1E+00	0,0004 - 0,006
1E-02	1E+00 - 1E+02	1E+01 - 1E+02	0,0005 - 0,2

TABELA 64: Interseção de esferas - DI

RE	RMSD	Tempo/s
1E-08	1E-06 - 1E-04	1 - 81
1E-06	1E-04 - 1E-02	1 - 82
1E-04	1E-02 - 1E-01	1 - 83
1E-02	1E+00 - 1E+01	1 - 112

Considerando primeiramente método linear pode-se notar que o RMSD aproximado cumpre o papel de atuar como limite superior para o RMSD original apresentando em todos os casos uma variação em torno de 1E+01. A vantagem ao se utilizar o RMSD aproximado é que não é necessário conhecer a proteína previamente para estimar o erro.

Comparando o método linear ao método de interseção de esferas, novamente pode-se notar que os algoritmos apresentam os mesmos resultados que o caso exato, com exceção do caso em que $RE = 1E-02$. Para $RE = 1E-02$ o método de interseção de esferas apresentou um erro menor que o método linear, o que significa que conforme o RE aumenta o método de interseção de esferas é mais eficaz.

7.3 DISTÂNCIAS EXATAS E ESPARSAS - DEE

Para o caso em que as distâncias são exatas, com esparsidade na matriz de distâncias, os seguintes algoritmos foram testados: Método Linear, Método Linear Atualizado, Método Linear Revisado Versão 2, Método Linear com Mínimos Quadrados Linear e Método Linear com Mínimos Quadrados não-Linear. As estruturas definidas possuem de 103 a 8.616 átomos.

Nas Tabelas a seguir são apresentados o menor e maior valor para o RMSD e tempo obtidos em cada método.

TABELA 65: Método Linear - DEE

ML	16$\mathring{A}$	14$\mathring{A}$	12$\mathring{A}$	10$\mathring{A}$	8$\mathring{A}$
RMSD	1E-11 - 1E-01	1E-11 - 1E-2	1E-10 - 1E+101	1E-09 - 1E+51	1E-07 - 1E+93
Tempo/s	0,09 - 18,3	0,09 - 16,9	0,003 - 8,4	0,4 - 5,4	0,3 - 6,2

TABELA 66: Método Linear Atualizado - DEE

MLA	8$\mathring{A}$	7$\mathring{A}$	6$\mathring{A}$	5$\mathring{A}$	4$\mathring{A}$
RMSD	1E-14 - 1E-13	1E-14 - 1E-11	1E-14 - 1E-12	1E-14 - 1E-11	1E-11 - 1E-06
Tempo/s	0,5 - 24,5	0,8 - 23,1	0,8 - 22,8	0,9 - 23,3	21,4 - 312,91

TABELA 67: Método Linear Revisado Versão 2 - DEE

MLRV2	8$\mathring{A}$	7$\mathring{A}$	6$\mathring{A}$	5$\mathring{A}$
RMSD	1E-14 - 1E-13	1E-14 - 1E-12	1E-14 - 1E-10	1E-14 - 1E-01
Tempo/s	0,5 - 22,5	0,5 - 21,7	0,7 - 21,3	0,8 - 23,6

TABELA 68: Método Linear com Mínimos Quadrados Linear - DEE

MLMQL	8$\mathring{A}$	7$\mathring{A}$	6$\mathring{A}$	5$\mathring{A}$	4$\mathring{A}$
RMSD	1E-14 - 1E-10	1E-14 - 1E-06	1E-14 - 1E+01	1E-09 - 1E+10	1E+01 - 1E+46
Tempo/s	0,1 - 7,5	0,1 - 6,8	0,1 - 6,0	0,1 - 5,5	0,2 - 6,3

TABELA 69: Método Linear com Mínimos Quadrados Não-Linear - DEE

MLMQNL	8$\mathring{A}$	7$\mathring{A}$	6$\mathring{A}$	5$\mathring{A}$	4$\mathring{A}$
RMSD	1E-12 - 1E+36	1E-13 - 1E-07	1E-13 - 1E+22	1E-14 - 1E+150	1E-13 - 1E-04
Tempo/s	0,7 - 17,1	0,7 - 17,3	0,4 - 913,7	0,4 - 1430,9	0,5 - 1325,0

Considerando primeiramente o método linear, nota-se que o algoritmo apresenta erros comportados para 16$\mathring{A}$ e 14$\mathring{A}$, conforme a esparsidade na matriz de distâncias moleculares aumenta estes erros se tornam muito instáveis, como o caso de 12$\mathring{A}$. O tempo de processamento se mostrou eficaz, visto que o algoritmo tem uma fórmula fixa para o problema sendo necessário apenas determinar a matriz A em cada iteração, mas em alguns casos não foi possível determinar a matriz A inversível, impossibilitando a determinação da proteína.

O método linear atualizado já se apresenta mais estável, obtendo pouca variação no erro mesmo para esparsidades muito grandes como 4$\mathring{A}$, mas também com erros menores que o método linear e portanto apresentando-se mais eficaz. O tempo de processamento é um pouco maior se comparado ao método linear, visto que a cada iteração é necessário recalcular os átomos iniciais para então encontrar o átomo desejado.

O método linear revisado versão 2 apresenta melhores resultados que os métodos anteriores, mostrando certa instabilidade para esparsidades 6$\mathring{A}$ e 5$\mathring{A}$. O tempo de processamento deste método é superior ao método linear atualizado mas não chega a ser tão rápido quanto o método linear, sendo que essa demora se da ao fato que a cada iteração são criadas múltiplas estruturas que devem ser comparadas.

O método linear com mínimos quadrados linear apresenta resultados superiores em relação ao erro se comparado ao método linear, sendo capaz de definir estruturas considerando maiores esparsidades na matriz de distâncias moleculares, mas com resultados inferiores aos demais métodos. Com relação ao tempo, os resultados são satisfatórios, alcançando o melhor tempo em relação a esparsidade desejada.

O método linear com mínimos quadrados não-linear apresentou melhores resultados para as esparsidades 5$\mathring{A}$ e 4$\mathring{A}$ em relação aos demais algoritmos considerando erro em relação a esparsidade, porém também apresentou os resultados mais instáveis. Mostrando que o algoritmo não é confiável ao analisar a confiabilidade em determinar a estrutura corretamente, sendo que o tempo de processamento cresceu

exponencialmente conforme a esparsidade aumentou.

7.4 DISTÂNCIAS INEXATAS E ESPARSAS - DIE

Para o caso em que as distâncias são inexatas com esparsidade na matriz de distâncias os seguintes algoritmos foram testados: Método Linear, Método Linear Atualizado, Método Linear com Mínimos Quadrados Linear e Método Linear com Mínimos Quadrados não-Linear. As estruturas definidas possuem de 103 a 8.616 átomos.

Nas Tabelas a seguir são apresentados o menor e maior valor para o RMSD e tempo obtidos em cada método.

TABELA 70: RMSD Método Linear - DIE

RE/ Esp	16Å	14Å	12Å	10Å	8Å
1E-08	1E-05 - 1E+14	1E-04 - 1E+61	1E-04 - 1E+93	1E-02 - 1E+67	1E-01 - 1E+51
1E-06	1E-02 - 1E+66	1E-02 - 1E+84	1E-01 - 1E+53	1E-01 - 1E+28	1E+02 - 1E+51
1E-04	1E+01 - 1E+82	1E+00 - 1E+91	1E+01 - 1E+67	1E+11 - 1E+66	1E+02 - 1E+49
1E-02	1E-01 - 1E+68	1E-01 - 1E+92	1E+00 - 1E+57	1E+05 - 1E+36	1E+06 - 1E+61

TABELA 71: Tempo Método Linear - DIE

RE/ Esp	16Å	14Å	12Å	10Å	8Å
1E-08	0,3 - 6,2	0,3 - 5,6	0,3 - 5,3	0,5 - 1,9	0,4 - 1,9
1E-06	0,2 - 6,1	0,2 - 5,7	0,3 - 1,6	0,4 - 1,3	0,5 - 1,9
1E-04	0,2 - 6,1	0,2 - 5,7	0,3 - 1,5	0,4 - 2,0	0,4 - 2,1
1E-02	0,2 - 6,4	0,2 - 5,9	0,3 - 2,0	0,4 - 2,0	0,5 - 2,0

TABELA 72: RMSD Método linear Atualizado - DIE

RE/ Esp	8Å	7Å	6Å	5Å	4Å
1E-08	1E-07 - 1E-05	1E-07 - 1E-04	1E-07 - 1E-04	1E-06 - 1E-03	1E-03 - 1E+01
1E-06	1E-05 - 1E-03	1E-05 - 1E-03	1E-05 - 1E-01	1E-04 - 1E+00	1E+00 - 1E+01
1E-04	1E-03 - 1E-01	1E-03 - 1E-01	1E-03 - 1E+00	1E-03 - 1E+01	1E+00 - 1E+01
1E-02	1E-01 - 1E+01	1E-01 - 1E+01	1E-01 - 1E+01	1E-01 - 1E+01	1E+00 - 1E+01

TABELA 73: Tempo Método linear Atualizado - DIE

RE/ Esp	8Å	7Å	6Å	5Å	4Å
1E-08	0,5 - 25,7	0,5 - 24,1	0,6 - 26,6	0,6 - 23,8	3,0 - 48,0
1E-06	0,5 - 25,7	0,5 - 24,3	0,6 - 21,3	0,6 - 22,1	2,0 - 43,8
1E-04	0,5 - 24,1	0,5 - 22,8	0,5 - 22,0	0,6 - 22,2	2,8 - 43,9
1E-02	0,5 - 24,3	0,5 - 23,2	0,6 - 22,1	0,7 - 22,9	2,9 - 43,6

TABELA 74: RMSD Método Linear com Mínimos Quadrados Linear - DIE

RE/ Esp	8Å	7Å	6Å	5Å	4Å
1E-08	1E-07 - 1E-03	1E-07 - 1E+00	1E-07 - 1E+01	1E-01 - 1E+02	1E+01 - 1E+58
1E-06	1E-05 - 1E-01	1E-05 - 1E+01	1E-05 - 1E+01	1E+0 - 1E+04	1E+02 - 1E+56
1E-04	1E-03 - 1E+00	1E-03 - 1E+01	1E-03 - 1E+01	1E+00 - 1E+09	1E+03 - 1E+36
1E-02	1E-01 - 1E+01	1E-01 - 1E+01	1E-01 - 1E+02	1E+00 - 1E+12	1E+10 - 1E+51

TABELA 75: Tempo Método Linear com Mínimos Quadrados Linear - DIE

RE/ Esp	8Å	7Å	6Å	5Å	4Å
1E-08	0,1 - 7,3	0,1 - 6,6	0,1 - 6,0	0,1 - 5,6	0,2 - 6,5
1E-06	0,1 - 7,3	0,1 - 6,2	0,1 - 6,0	0,1 - 5,6	0,2 - 6,3
1E-04	0,1 - 7,3	0,1 - 6,6	0,1 - 6,0	0,1 - 5,5	0,2 - 6,4
1E-02	0,1 - 7,3	0,1 - 6,6	0,1 - 6,2	0,1 - 5,6	0,2 - 6,3

TABELA 76: RMSD Método Linear com Mínimos Quadrados não-Linear - DIE

RE/ Esp	8Å	7Å	6Å	5Å	4Å
1e-08	1e-05 - 1e+63	1e-06 - 1e+16	1e-07 - 1e+108	1e-07 - 1e+15	1e-05 - 1e+06
1e-06	1e-02 - 1e+74	1e-04 - 1e+31	1e-05 - 1e+134	1e-05 - 1e+67	1e-03 - 1e+07
1e-04	1e+02 - 1e+87	1e-02 - 1e+44	1e-03 - 1e+150	1e-03 - 1e+66	1e+00 - 1e+08
1e-02	1e+09 - 1e+99	1e+01 - 1e+60	1e+02 - 1e+101	1e+01 - 1e+82	1e+00 - 1e+08

TABELA 77: Tempo Método Linear com Mínimos Quadrados não-Linear - DIE

RE/ Esp	8Å	7Å	6Å	5Å	4Å
1E-08	0,6 - 352,1	0,4 - 34,1	0,3 - 860,5	0,3 - 179,4	0,4 - 1305,8
1E-06	0,5 - 334,2	0,4 - 34,2	0,3 - 866,7	0,3 - 177,1	0,4 - 1308,3
1E-04	0,5 - 37,6	0,4 - 35,2	0,3 - 849,1	0,3 - 164,3	0,4 - 1359,2
1E-02	0,5 - 40,0	0,8 - 38,5	0,3 - 399,2	0,3 - 167,9	0,4 - 1294,0

Considerando primeiramente o método linear, pode-se notar que o algoritmo tem um decaimento do erro proporcional ao decaimento de RE, porém o método linear apresenta-se instável ao considerar que o erro de máquina é de 1E-16 e portanto o

resultado desejado seria próximo de RE. O método linear é capaz de determinar a proteína rapidamente obtendo uma melhora considerável no tempo conforme a esparsidade aumenta.

O método linear atualizado apresenta melhores resultados que o método linear para todas as esparsidades, obtendo resultados comportados, com poucas variações e mostrando-se confiável. Em relação ao tempo, o método linear também não deixa a desejar pois o algoritmo apresenta uma melhora no tempo conforme a esparsidade aumenta, com exceção de $4\mathring{A}$ que é o caso em que podem não haver distâncias ao átomo a ser determinado dificultando o andamento do algoritmo. Vale salientar que o fato de RE aumentar não prejudica o tempo do algoritmo.

Considerando o método linear com mínimos quadrados linear, pode-se notar que o método linear ganha em relação ao tempo, pois ele apenas resolve um sistema linear 3x3, enquanto MLMQL resolve um sistema com l-1 equações. No entanto o MLMQL ganha do ML em relação ao erro assim como esperado, visto que ao invés de utilizar apenas quatro átomos como base para encontrar o átomo x_i ele usa l átomos com distâncias conhecidas a x_i, e deste modo faz uma média entre o erro gerado em cada distância. Comparado ao método linear atualizado, o método linear com mínimos quadrados linear é mais rápido e alcança os mesmos resultados de erros com exceção das esparsidades $5\mathring{A}$ e $4\mathring{A}$ porém, apresentando maior variação nestes erros.

Ao comparar os resultados do método linear com mínimos quadrados não-linear ao método linear atualizado, os resultados obtidos foram similares em relação ao erro para os casos exato e inexato, exceto para $RE = 1e-02$, em esparsidades menores ou iguais a $6\mathring{A}$, $5\mathring{A}$ e $4\mathring{A}$. O algoritmo do método linear com mínimos quadrados não-linear alcançou resultados melhores em alguns casos, obtendo novamente resultados gigantescos como $1e+108$ e assim demora muito tempo para rodar o algoritmo para proteínas grandes. Já nos casos em que o algoritmo do método linear com mínimos quadrados não-linear determinou a proteína nestas mesmas esparsidades o método foi mais rápido. Para $Re = 1e-02$, o algoritmo perdeu tanto no tempo quanto no cálculo

do erro.

Mesmo apresentando resultados satisfatórios apenas para os casos com distâncias menores que $6\mathring{A}$ o método linear com mínimos quadrados não-linear pode ser considerado um algoritmo eficaz, visto que em casos reais o erro na matriz de distâncias é menor ou igual a $5\mathring{A}$. Porém o algoritmo é instável e deve ser usado em conjunto com outros algoritmos para garantir sua eficácia.

8 CONCLUSÃO

Através da análise de diversos algoritmos propostos neste trabalho, pode-se notar que cada algoritmo apresenta suas vantagens ou desvantagens. Considerando a motivação desse estudo, que é o problema do cálculo de estruturas de proteínas, onde a esparsidade na matriz de distâncias é em torno de $5\mathring{A}$ com um erro em torno de $1e-05$. Em nossa opinião o algoritmo que conseguiu alcançar melhores resultados, mas não em todos os casos, foi o método linear com mínimos quadrados não-linear. Mas neste caso, supondo que a estrutura real não é conhecida, para comparar se o resultado obtido foi eficaz, não se tem garantia de que o erro final é pequeno, visto que em vários casos o algoritmo alcançou inesperadamente erros gigantescos, em torno de $1e+30$, ocasionando uma estrutura completamente diferente da esperada.

Já em relação ao tempo de processamento o método linear rígido versão 2 apresentou melhores resultados, mas assim como o método linear com mínimos quadrados não-linear, o método linear rígido versão 2 também apresenta uma falha, pois não há garantia de que uma única estrutura será encontrada. No entanto, quando mais de uma estrutura for encontrada não haverá um resultado final, sabendo portanto que as estruturas encontradas não são confiáveis. Procurando por um algoritmo que obtenha o melhor tempo e melhores resultados e com a garantia de que estes resultados são confiáveis, o método linear atualizado é o melhor candidato para resolver o problema do cálculo de estruturas de proteínas.

Em relação as contribuições desta dissertação, quase sempre que novo algoritmo é proposto na literatura os testes para validar os resultados deste novo algoritmo são comparados ao melhor algoritmo existente, sem fazer um panorama geral de qual método realmente poderia ser mais eficaz. Um exemplo deste fato pode ser visto no método linear, pois o algoritmo foi proposto apenas para o caso exato e sem esparsi-

dade na matriz de distâncias, mas de acordo com os resultados mostrados, é possível verificar que o método linear também é eficaz para distâncias inexatas, alcançando inclusive melhores resultados com sua variação no método linear atualizado para distâncias esparsas.

Em relação ao algoritmo do método de interseção de esferas, até o momento, nenhuma das escolhas de átomos inicias discutidas neste trabalho se apresentou eficaz. Portanto, ainda não foi determinada uma escolha eficiente dos átomos iniciais escolhidos a cada passo do método. Mas já é possível notar para o caso sem esparsidade na matriz de distâncias que o método de interseção de esferas garante mais flexibilidade tanto em relação ao tempo como na tolerância do erro.

REFERÊNCIAS

ANTON, H.; BUSBY, R. C. **Algebra linear Contemporânea**. [S.l.]: Bookman, 2006.

ANTON, H.; RORRES, C. **Algebra linear com Aplicações**. [S.l.]: Bookman, 2012.

BERMAN, H. *et al.* **The Protein Data Bank Nucleic Acids Research**. 2000. Disponível em: <http://www.wwpdb.org>.

COOPE, I. D. A reliable computation of the points of intersection of n spheres in $\mathbb{R}^n$. Anziam J., v. 42 (E) ppC461-C477, 2000.

DAVIS, R. T.; ERNST, C.; WU, D. Protein structure determination via an efficient geometric build-up algorithm. **BMC Structural Biology**, v. 10, 2010.

DONG, Q.; WU, Z. A linear time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. **Journal of Global Optimization**, v. 22, p. 365 – 375, 2002.

DONG, Q.; WU, Z. A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. **Journal of Global Optimization**, v. 26, p. 321 – 333, 2003.

FAUSETT, L. V. **Appied Numerical Analysis Using Matlab**. [S.l.]: Prentice Hall, 1999.

GANDER, W.; GOLUB, G. H.; STREBEL, R. **Least-Squares Fitting of Circles and Ellipses**. [S.l.]: Bulletin of the Belgian Mathematical Society, 1996.

GUIDORIZZI, H. L. **Um Curso de Cálculo**. [S.l.]: LTC, 2001.

HARVEY, N. **Notes on Symmetric Matrices**. 2011. Disponível em: <http://www.cs.ubc.ca/ nickhar/W12/NotesMatrices.pdf>.

KINCAID, D.; CHENEY, W. **Numerical Analysis: Mathematics of Scientific Computing**. [S.l.]: American Mathematical Society, 2009.

NOCEDAL, J.; WRIGHT, S. J. **Numerical Optimization**. second edition. [S.l.]: Springer, 2006.

RIBEIRO, A. A.; KARAS, E. W. **Otimização Continua - Aspectos Teóricos e Computacionais**. [S.l.]: Cengage Learning, 2014.

SILVA, W. G. da. **Algoritmos para o Cálculo de Estruturas de Proteínas**. Niterói: Universidade Federal Fluminense, 2008.

SIT, A.; WU, Z.; YUAN, Y. A stable geometric buildup algorithm for the solution of the distance geometry problem using least-squares. **Institute for Mathematics and its applications**, IMA Preprint Series, v. 2206, 2008.

SIT, A.; WU, Z.; YUAN, Y. A geometric buildup algorithm for the solution of the distance geometry problem using least-squares approximation. Bulletin of Mathematical Biology, v. 71: 1914 - 1933, 2009.

SOUTO, G. **Decomposição em Valores Singulares**. FLORIANÓPOLIS - SC: TRABALHO DE CONCLUSÃO DE CURSO - UNIVERSIDADE FEDERAL DE SANTA CATARINA, 2000.

SOUZA, M. F. **Suavização Hiperbólica Aplicada à Otimização de Geometria Molecular**. Rio de Janeiro: Tese de Doutorado, UFRJ, 2010.

WU, D. **Distance-based protein structure modeling**. Ames, Iowa: Iowa State University, 2006.

WU, D.; WU, Z. An updated geometric build-up algorithm for solving the molecular distance geometry problems with sparse distance data. **J Glob Optim**, Springer, DOI 10.1007/s10898-006-9080-6, 2007 – a.

WU, D.; WU, Z.; YUAN, Y. The solution of the distance geometry problem in protein modeling via geometric buildup. **Institute for Mathematics and its applications**, IMA Preprint Series, v. 2184, 2007.

WU, D.; WU, Z.; YUAN, Y. Rigid versus unique determination of protein structures with geometric buildup. **Optimization Letters**, Springer, DOI 10.1007/s11590-007-0060-7, 2007 – b.