Universidade Federal Fluminense

Pedro Costa Nucci

Métodos Computacionais para o Cálculo de Estruturas de Proteínas: Aproximando o Problema Molecular de Geometria de Distâncias de Dados de Ressonância Magnética Nuclear

Niterói

2010

Métodos Computacionais para o Cálculo de Estruturas de Proteínas: Aproximando o Problema Molecular de Geometria de Distâncias de Dados de Ressonância Magnética Nuclear

Pedro Costa Nucci

Dissertação de Mestrado submetida ao corpo docente do Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como parte dos requisitos necessários para obtenção do grau de Mestre em Computação.

Orientadores:

Loana Tito Nogueira

Carlile Campos Lavor (orientador externo pela UNICAMP)

Niterói

2010

Métodos Computacionais para o Cálculo de Estruturas de Proteínas: Aproximando o Problema Molecular de Geometria de Distâncias de Dados de Ressonância Magnética Nuclear

Pedro Costa Nucci

Dissertação de Mestrado submetida ao corpo docente do Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como parte dos requisitos necessários para obtenção do grau de Mestre em Computação.

Banca examinadora:

Prof. Alexandre Andreatta, D.Sc.
UNIRIO

Prof. Carlile Lavor, D.Sc.
(Orientador)
UNICAMP

Prof. Carlos Alberto Martinhon, D.Sc.
UFF

Prof. Fabio Almeida, D.Sc.
UFRJ

Profª. Loana Tito Nogueira, D.Sc.
(Orientadora)
UFF

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e Instituto de Computação da UFF

N962 Nucci, Pedro Costa.
Métodos computacionais para o cálculo de estruturas de proteínas : aproximando o problema molecular de geometria de distâncias de dados de ressonância magnética nuclear / Pedro Costa Nucci. – Niterói, RJ : [s.n.], 2010.
50 f.

Dissertação (Mestrado em Computação) - Universidade Federal Fluminense, 2010.
Orientadores: Loana Tito Nogueira, Carlile Campos Lavor.

1. Otimização combinatória (Computação). 2. Problema molecular de geometria de distância. 3. Estrutura molecular. 3. Proteína. I. Título.

CDD 004.0151

Resumo

O Problema Molecular de Geometria de Distâncias (PMGD) consiste em determinar as coordenadas cartesianas de cada átomo de uma molécula, a partir de algumas distâncias conhecidas entre seus átomos. Ao assumirmos que os dados de entrada têm algumas propriedades bastante compatíveis com dados experimentais obtidos por métodos como a espectroscopia por ressonância magnética nuclear de proteínas, o problema é classificado como NP-Completo, e tem espaço de busca discreto, sendo chamado de Problema Molecular de Geometria de Distâncias Discreto (PMGDD). Para este problema um algoritmo *Branch-And-Prune* (BP) é proposto na literatura.

Neste trabalho, lidamos com dois métodos baseados no BP que, visando a aplicabilidade prática, utilizam cadeias artificiais de átomos que não são ligados quimicamente, mas satisfazem as restrições do problema. Uma vez determinadas as estruturas dessas cadeias, determina-se a cadeia principal da proteína. Para um dos métodos, presente na literatura, estudamos o processo de obtenção da cadeia principal, apresentando um método alternativo que não requer resolução de sistemas lineares ou quadráticos, diferentemente do original. A outra aplicação do BP está sendo proposta neste trabalho, consistindo de uma nova cadeia artificial, com vantagens como a não utilização de átomos com distâncias imprecisas, menor custo de exploração do espaço de busca e maior facilidade de obtenção da cadeia principal.

Palavras-Chave:

Geometria de Distâncias, Otimização Combinatória, Estrutura Terciária de Proteínas, *Branch-And-Prune*.

Abstract

The Molecular Distance Geometry Problem (MDGP) consists in finding the cartesian coordinates of the atoms of a molecule, given some inter-atomic distances previously known. When the input data has some properties quitely agreeable with experimental data provided by methods as protein nuclear magnetic resonance spectroscopy, this problem is classified as NP-Complete, having discrete search space, and is known as Discrete Molecular Distance Geometry Problem (DMDGP). For this problem, a *Branch-And-Prune* (BP) algorithm is proposed under literature.

In this work, we deal with two methods based upon BP that, aiming practical applicability, use artificial atom chains not chemically bounded, but which satisfy problem restrictions. Once determined the structures for these chains, one determines the backbone of the protein. For one of these methods, present in the literature, we study the process of obtaining the backbone, presenting a new method, which does not require quadratic or linear systems solving, differently from the former. The other application of BP is proposed in this work, comprehended by a new artificial chain, with advantages such as no use of atoms with imprecise measures, smaller cost of exploring the search space and easier determining of the backbone.

Keywords:

Distance Geometry, Combinatorial Optimization, Tertiary Protein Structures, Branch-And-Prune.

Aos meus avós, meus verdadeiros mestres.

Agradecimentos

Agradeço ao meu orientadores.

Minha orientadora Loana é modelo, em diferentes aspectos, do que eu gostaria que os alunos de universidades públicas encontrassem em sala de aula e nas orientações. Nela pude encontrar eco para as minhas próprias idéias, bem como ver as coisas por outro ângulo (muitas vezes mais interessante) e essa ressonância é sempre produtiva.

Carlile, que desde os tempos de iniciação científica, tem sido uma ótima referência, fora os atributos usuais de um professor, na habilidade de enxergar beleza e simplicidade em alguns problemas e soluções matemáticas. É isso o que transforma uma lenta viagem de ônibus em uma oportunidade de se encontrar respostas (ou perguntas) interessantes.

Agradeço, em especial, a parte do corpo docente do IC-UFF, por me fazer olhar para fora da área acadêmica, algo que me trouxe um novo e interessante rumo.

Ao pessoal do Departamento de TI do Arsenal de Marinha do Rio de Janeiro: Ânderson, Aurélio, Cristiane, Déborah, Denise, Eni, Mônica Pardellas, Flavinha, Jorge, Luciene, Marcos, Mitiko, Thiago. Agradeço ao Aurélio mais uma vez por me ter deixado flexibilizar o horário em alguns dias, algo crucial para que eu continuasse no mestrado.

Meus colegas e amigos da UFF: além do pessoal da graduação (os que sobreviveram), a Cláudia, Leo Motta, Luciene Motta, Flávio. Teresa (especialmente), Viviane, Carlos e Rafael.

A banca examinadora contribuiu ativamente com a discussão de pontos importantes desse trabalho. Fabio Almeida, que já tinha me dado a oportunidade de estar no laboratório que coordena por algumas semanas, onde me senti num ambiente agradável e produtivo. Trouxe ótimas reflexões para o trabalho. Espero continuar com a colaboração. Alexandre Andreatta. Após assistir o curso de Heurísticas para Problemas Combinatórios lecionado por ele, saí convicto de que meu trabalho deveria ser avaliado por sua maneira de pensar inteligente e diferente. Carlos Martinhon, que após dois cursos (difíceis, mas muito bem lecionados) na pós-graduação e conversas informais dignas de quem tem interesse pela pesquisa, é para mim também um exemplo.

Sumário

Lista de Figuras

Lista de Algoritmos

Lista de Tabelas

Lista de Abreviações

PMGD	Problema Molecular de Geometria de Distâncias
PMGDD	Problema Molecular de Geometria de Distâncias Discreto
BP	*Branch-And-Prune*
RMN	Ressonância Magnética Nuclear
PDB	*Protein Data Bank*
pH	Potencial Hidrogeniônico

Capítulo 1

Introdução

1.1 Motivação

As macromoléculas têm sido cada vez mais um importante alvo de estudos em biologia, devido às importantes funções que desempenham nos sistemas biológicos. Em particular, a elucidação da estrutura tridimensional de macromoléculas é necessária para compreendermos a natureza de suas funções fisiológicas, bem como propriedades físico-químicas que nos permitam inferir possíveis interações com outras moléculas, como por exemplo, a interação entre um fármaco e uma proteína que tem função inibida por ele.

Dentre os tipos de macromoléculas presentes em organismos vivos, as proteínas certamente são as mais diversificadas e ricas quanto à estrutura tridimensional [6]. São sequências de resíduos de aminoácidos (frequentemente chamados apenas de resíduos), unidos por ligações peptídicas, que devido à diversidade de aminoácidos existentes e ao modo como são ligados, possuem estrutura tridimensional complexa e sensível a condições do meio, como temperatura e pH. Portanto é interessante conhecermos a estrutura de uma proteína em condições semelhantes ao seu meio natural, para que os estudos sejam condizentes com a realidade biológica.

1.2 Aminoácidos e Proteínas

Aminoácidos são moléculas orgânicas compostas fundamentalmente por átomos de carbono, nitrogênio, oxigênio e hidrogênio. Existem 20 principais tipos de aminoácidos que ocorrem em organismos vivos, classificáveis quanto a características tais como polaridade e carga (o Apêndice I apresenta uma relação deles). Para os aminoácidos que não os do tipo glicina, o carbono-α (carbono ligado ao grupo carbonila) está ligado a quatro diferentes grupos: grupo amina, grupo

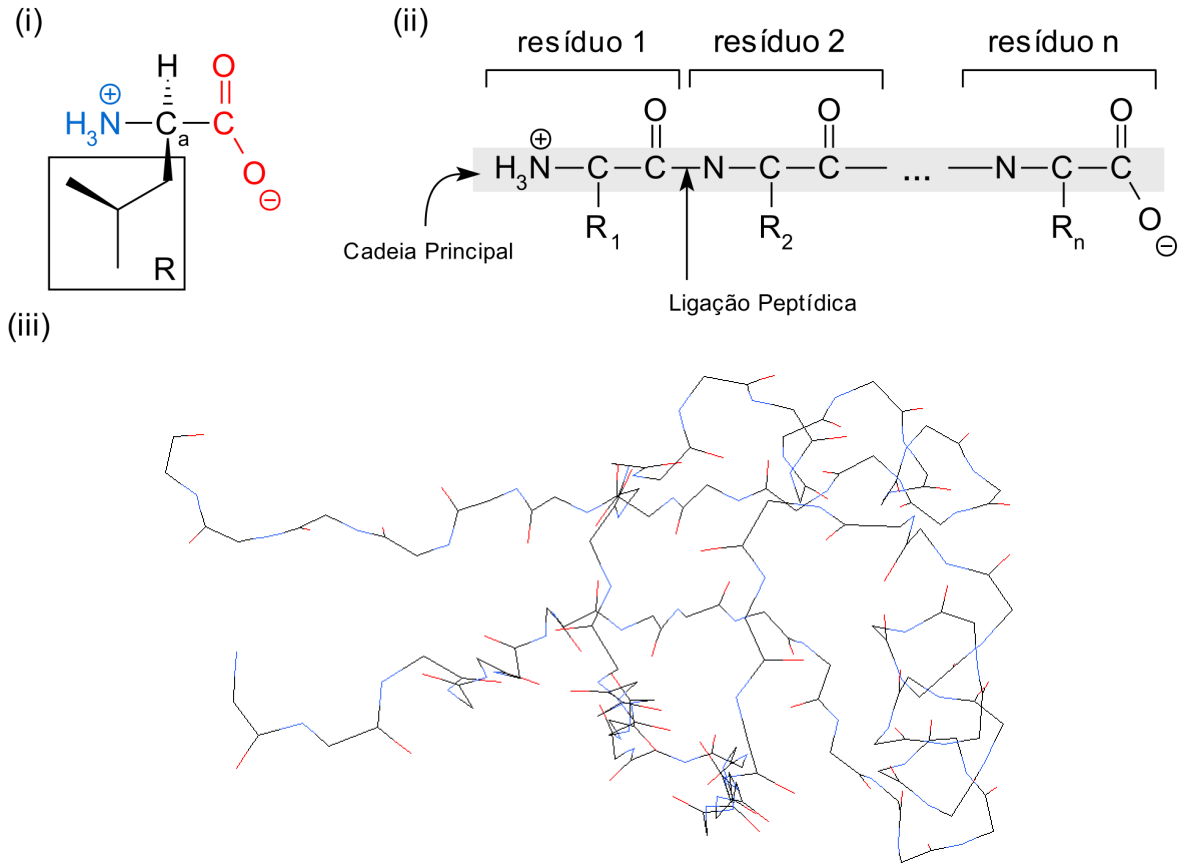


Figura 1.1: Aminoácidos e Proteínas

(i) Estrutura geral dos aminoácidos precursores dos resíduos das proteínas, apresentando um grupo amina (H_2N-), um grupo carboxila ($-COOH$) e um radical R (destacado), que varia com cada tipo de aminoácido (no exemplo está representada a leucina). (ii) Estrutura de uma cadeia polipeptídica, resultado da polimerização de vários aminoácidos. (iii) Conformação tridimensional adotada por uma proteína de 71 resíduos de aminoácidos (para facilitar a visualização, os radicais R foram ocultados).

carboxila, grupo R e um átomo de hidrogênio (ver figura 1.1). O carbono-α desses aminoácidos é chamado de centro quiral, pois, devido à geometria tetraédrica dos orbitais moleculares em torno dele, os quatro grupos diferentes podem assumir somente duas disposições espaciais. Essas duas disposições são imagens especulares umas das outras não sobreponíveis, sendo chamadas de enantiômeros. No caso da glicina, o grupo R é um outro átomo de hidrogênio, fazendo com que essa situação não se configure (pois há dois grupos iguais).

Para distinguir os enantiômeros de uma molécula, pode-se usar a convenção de Fischer, que toma como referência a molécula de gliceraldeído (um açúcar que tem um centro quiral ligado a quatro grupos distintos). Nela, é estabelecida uma associação entre os quatro grupos do gliceraldeído e os quatro grupos da molécula em questão. O enantiômero é então classificado como D ou L em função de sua correspondência ao D-gliceraldeído ou L-gliceraldeído, os dois enantiômeros do gliceraldeído (de propriedades óticas distintas). A figura 1.2 exemplifica a clas-

sificação dos dois enantiômeros do aminoácido alanina. Os resíduos de aminoácidos encontrados em proteínas são todos enantiômeros do tipo L.

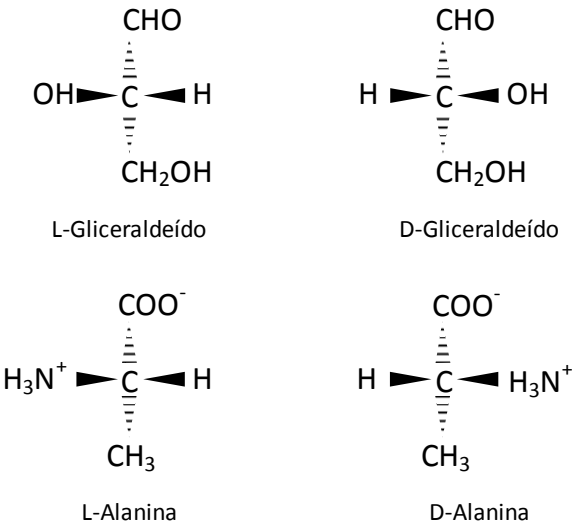


Figura 1.2: Classificação dos enantiômeros da alanina segundo a convenção de Fischer.

A sequência dos aminoácidos correspondentes aos resíduos de uma proteína é chamada de *estrutura primária*. Alguns resíduos próximos entre si assumem padrões estruturais localizados decorrentes de interações de curta distância (por exemplo, quando um conjunto de resíduos em sequência assume uma estrutura cilíndrica com as cadeias laterais viradas para fora, chamada alfa-hélice). Tal nível de organização é chamado de *estrutura secundária*. Um nível acima, está o arranjo tridimensional de toda a cadeia polipeptídica, a *estrutura terciária*, sendo mais diretamente responsável pela função biológica da molécula. Os dois extremos de uma proteína são chamados N-terminal e C-terminal, em função de qual é o grupo livre em cada extremidade, amina ou carboxila (na figura 1.1(ii), o N-terminal está à esquerda). Uma proteína pode ser monomérica, consistindo de somente uma cadeia, ou oligomérica, quando tem mais de uma cadeia (nesse caso, apesar de não estarem necessariamente unidas por ligações covalentes, as cadeias se mantêm unidas por outras interações químicas). A cadeia principal é considerada a "espinha dorsal" de uma proteína. Conhecer sua estrutura é um passo crucial para a determinação da estrutura terciária e portanto para determinação das funções e propriedades da molécula.

Uma técnica que vem sendo utilizada e aperfeiçoada para determinação de estruturas protéicas é a espectroscopia por Ressonância Magnética Nuclear (RMN)[2]. Dentre os dados experimentais que esta técnica fornece, podemos conhecer faixas de valores para distâncias inter-atômicas da molécula analisada que tenham comprimento de até um certo valor, sendo necessário um processamento computacional posterior, para se calcular as possíveis estruturas adotadas pela

molécula durante o experimento.

1.3 Problema Molecular de Geometria de Distâncias

O Problema Molecular de Geometria de Distâncias (PMGD) consiste em encontrar todas as estruturas tridimensionais possíveis para uma molécula, quando conhecemos algumas de suas distâncias inter-atômicas. Para os casos em que não são conhecidas todas as distâncias inter-atômicas, o PMGD é classificado como um problema NP-Completo [1] e requer uso de técnicas de otimização.

Diferentes métodos foram propostos na literatura para resolver o PMGD, como o algoritmo ABBIE [9], que utiliza uma abordagem baseada em grafos com paradigma dividir-e-conquistar, o algoritmo DGSOL [10], que utiliza técnicas de otimização global, o algoritmo de incremento geométrico ("Geometric build-up algorithm") [15, 17], que calcula a posição de um átomo a partir das posições de quatro outros átomos não-coplanares, utilizando sistemas lineares.

Ao assumirmos algumas propriedades para o conjunto de distâncias inter-atômicas (compatíveis com muitas proteínas reais), o problema passa a ter espaço de busca discreto, e é chamado de Problema Molecular de Geometria de Distâncias Discreto (PMGDD). Em [1], é proposto o algoritmo *Branch-And-Prune* para resolver o PMGDD.

1.4 Objetivos

Neste trabalho visamos aproximar o PMGDD de dados oriundos de experimentos de RMN. Em trabalhos anteriores na literatura, o problema havia sido aplicado a instâncias baseadas diretamente na cadeia principal de proteínas. No entanto, os dados de entrada referentes à cadeia principal não podem ser obtidos em laboratório. Para contornar a situação, trabalhamos apenas com átomos cujas distâncias podem ser medidas nesses experimentos, através de cadeias artificiais de átomos, que, apesar de não serem ligados quimicamente, suas distâncias podem ser obtidas na prática. Uma vez solucionadas, tais cadeias artificiais podem ser usadas para obtermos a cadeia principal das proteínas. Apresentamos uma alternativa para uma parte de um método proposto na literatura. Propomos também um outro método, consistindo de uma nova cadeia artificial, cujas vantagens em relação ao método existente podem ser úteis para futuras aplicações.

Capítulo 2

Problema Molecular de Geometria de Distâncias

2.1 Conceitos Básicos e Definições

2.1.1 Formulação Contínua

O Problema Molecular de Geometria de Distâncias (PMGD) consiste em determinar a estrutura tridimensional de uma sequência linear de átomos, tendo como entrada algumas distâncias entre esses átomos. Na prática, tal sequência pode ser composta por átomos de uma grande molécula, como a mostrada na Figura 1.1, mas o PMGD pode ter outras aplicações, como o desenho de grafos tridimensionais [8] e o projeto de redes [11].

Formalmente, o PMGD é definido como um problema de otimização contínua, cujo objetivo é encontrar as coordenadas cartesianas $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n} \in \mathbb{R}^3$ para os n átomos (não necessariamente ligados quimicamente) de uma molécula que minimizem a função

$$f(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}) = \sum_{(i,j)\in S} \left(\|\mathbf{x_i} - \mathbf{x_j}\|^2 - d_{i,j}^2 \right)^2, \tag{2.1}$$

onde S é o conjunto de pares de átomos (i, j) cuja distância $d_{i,j}$ é conhecida. Uma instância do PMGD pode ter muitas soluções, pois podem existir diversos posicionamentos $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n} \in \mathbb{R}^3$ de seus átomos obedecendo as distâncias de entrada. Na prática, em experimentos de RMN, os pares de átomos que têm distância inferior a 5Å[1] são conhecidos, independentemente dos átomos serem elementos próximos na sequência.

[1] Å = Ångström, unidade do SI para distâncias de ordem atômica. 1Å=10^{-10}m.

Definimos o **intervalo** $[i, j]$ de uma molécula de n átomos como a subsequência de seus átomos $\{i, ..., j\}$, onde $1 \leq i \leq j \leq n$. As relações de pertinência entre átomos e intervalos, bem como as relações de igualdade e continência entre intervalos são análogas àquelas que envolvem elementos e conjuntos, conhecidas em Teoria dos Conjuntos. O tamanho de um intervalo $[i, j]$ é $j - i$. Uma **realização** R para o intervalo $[a, b]$ é uma função $R : [a, b] \mapsto \mathbb{R}^3$ que associa os átomos $k \in [a, b]$ a coordenadas cartesianas de um mesmo sistema de coordenadas, dadas pelo vetor $\mathbf{R}(\mathbf{k}) \in \mathbb{R}^3$.

2.1.2 Formulação Discreta

Como proposto em [7], o PMGD ganha uma formulação discreta, Problema Molecular de Geometria de Distâncias Discreto (PMGDD), se assumirmos as seguintes hipóteses:

1. Todos os pares de átomos (i, j), onde $1 \leq j - i \leq 3$, têm suas distâncias d_{ij} conhecidas.

2. Os ângulos entre as ligações de átomos consecutivos, ou seja, os ângulos formados pelos vetores $\mathbf{x_{i+2}} - \mathbf{x_{i+1}}$ e $\mathbf{x_{i+1}} - \mathbf{x_i}$, onde $1 \leq i \leq n - 3$, nunca serão múltiplos de π.

Ambas as hipóteses podem ser assumidas como verdadeiras ao aplicarmos o problema a proteínas, seja diretamente pelo cálculo da cadeia principal ([1, 3]), uma vez que os ângulos entre ligações e distâncias entre átomos consecutivos da cadeia principal ocorrem com valores já conhecidos que validam as duas hipóteses, seja através de aplicações mais realistas, das quais trataremos nos próximos capítulos.

O conjunto S de pares de átomos com distâncias conhecidas pode então ser particionado em dois subconjuntos:

- Conjunto E, que compreende todos os pares de átomos (i, j), onde $1 \leq j - i \leq 3$, decorrência da hipótese 1.

- Conjunto F dos pares de átomos (i, j), onde $j - i > 3$, cujas distâncias são conhecidas.

A figura 2.1 ilustra os dois tipos de pares de átomos através de um exemplo.

Chamaremos as distâncias $d_{i,j}$ que estiverem relacionadas aos pares $(i, j) \in E$ de **distâncias tipo** E, e as que estiverem relacionadas aos pares $(i, j) \in F$ de **distâncias tipo** F.

O conhecimento das distâncias tipo E já é suficiente para determinarmos os cossenos dos ângulos de ligação θ_i para os átomos $i = 3, ..., n$, e os cossenos dos ângulos de torção ω_i para os átomos $i = 4, ..., n$ (ver figura 2.2), através da lei dos cossenos

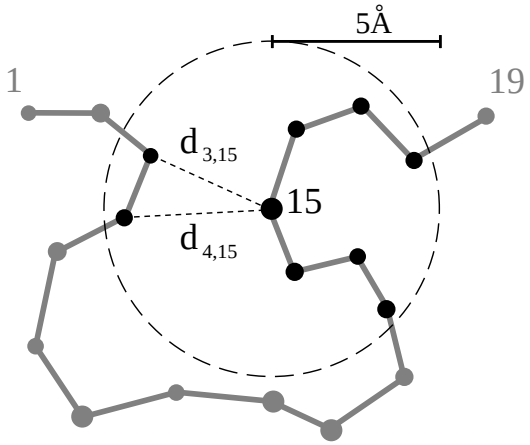


Figura 2.1: Conjuntos de Distâncias E e F

No exemplo acima, a circunferência delimita o raio de corte de 5Å para o átomo 15. Pela definição do PMGDD, os átomos separados do átomo 15 por até 3 ligações têm suas distâncias conhecidas, estando dentro do raio de corte. Logo os pares $(12, 15), (13, 15), ..., (15, 18)$ pertencerão ao conjunto E. Além disso, nesse caso, as distâncias para os átomos 3 e 4 também são menores que 5Å, e portanto os pares $(3, 15)$ e $(4, 15)$ pertencerão ao conjunto F.

$$\cos\theta_i = \frac{d_{i-1,i}^2 + d_{i-2,i-1}^2 - d_{i-2,i}^2}{2d_{i-1,i}d_{i-2,i-1}}$$

e de sua adaptação para ângulos de torção, apresentada em [12]. A aplicação no PMGDD da lei dos cossenos para ângulos de torção foi mostrada primeiramente em [7]. Recentemente, retificamos e melhor desenvolvemos a equação em [19], da seguinte forma:

$$\cos\omega_i = \frac{\cos\gamma - \cos\alpha\cos\beta}{\sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta}}, \tag{2.2}$$

onde

$$\begin{aligned}
\cos\alpha &= \frac{d_{i-2,i}^2 + d_{i-2,i-1}^2 - d_{i-1,i}^2}{2d_{i-2,i}d_{i-2,i-1}}, \\
\cos\beta &= \frac{d_{i-3,i-2}^2 + d_{i-2,i-1}^2 - d_{i-3,i-1}^2}{2d_{i-3,i-2}d_{i-2,i-1}} \quad \text{e} \\
\cos\gamma &= \frac{d_{i-3,i-2}^2 + d_{i-2,i}^2 - d_{i-3,i}^2}{2d_{i-3,i-2}d_{i-2,i}}.
\end{aligned}$$

Com os dados de distâncias entre átomos consecutivos ($d_{i-1,i}$), cossenos de ângulos entre ligações ($\cos\theta_i$) e cossenos de ângulos de torção ($\cos\omega_i$), podemos determinar a coordenada

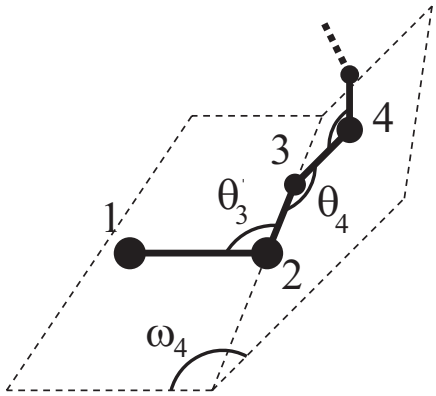


Figura 2.2: Ângulos de Ligação e Ângulos de Torção

O ângulo de ligação θ_i é formado entre a ligação dos átomos $(i-2)$ e $(i-1)$ e a ligação dos átomos $(i-1)$ e (i). O ângulo de torção ω_i é formado entre o plano determinado por $(i-3)$, $(i-2)$ e $(i-1)$ e o plano determinado por $(i-2)$, $(i-1)$ e (i).

cartesiana $x_i = (x_{i_x}, x_{i_y}, x_{i_z})$ do átomo i, usando a seguinte fórmula [13]:

$$\begin{bmatrix} x_{i_x} \\ x_{i_y} \\ x_{i_z} \\ 1 \end{bmatrix} = B_1 B_2 ... B_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{2.3}$$

onde

$$B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} -1 & 0 & 0 & -d_{1,2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$B_3 = \begin{bmatrix} -\cos\theta_3 & -\sin\theta_3 & 0 & -d_{2,3}\cos\theta_3 \\ \sin\theta_3 & -\cos\theta_3 & 0 & d_{2,3}\cos\theta_3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.4}$$

e

$$B_i = \begin{bmatrix} -\cos\theta_i & -\sin\theta_i & 0 & -d_{i-1,i}\cos\theta_i \\ \sin\theta_i\cos\omega_i & -\cos\theta_i\cos\omega_i & -\sin\omega_i & d_{i-1,i}\sin\theta_i\cos\omega_i \\ \sin\theta_i\sin\omega_i & -\cos\theta_i\sin\omega_i & \cos\omega_i & d_{i-1,i}\sin\theta_i\sin\omega_i \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ 4 \leq i \leq n \tag{2.5}$$

As matrizes B_i são chamadas de *matrizes de torção*. O primeiro átomo da molécula pode ser fixado em uma origem arbitrária, o segundo átomo em um eixo arbitrário, a distância $d_{1,2}$ da origem, e o terceiro em um plano arbitrário, a distâncias $d_{1,3}$ e $d_{2,3}$ dos outros dois átomos já fixados. As matrizes B_1, B_2 e B_3 são usadas para isso.

Uma vez fixados os três primeiros átomos, utilizando-se as matrizes de torção mostradas acima, os átomos seguintes terão suas matrizes de torção determinadas pela equação (2.5). Entretanto, para cada átomo $i \geq 4$, existem duas raízes possíveis para $\sin\omega_i$, que é calculado a partir de $\cos\omega_i$:

$$\sin\omega_i = \pm\sqrt{1-\cos^2\omega_i}, \tag{2.6}$$

gerando duas matrizes de torção possíveis, B_i^1 e B_i^2 (uma para cada raiz de $\sin\omega_i$). Com isso, uma instância de n átomos terá 2^{n-3} soluções possíveis, resultado das combinações de duas matrizes de torção de cada átomo $i = 4, 5, ..., n$. Portanto, o espaço de busca do problema é finito e pode ser explorado através das diferentes combinações de matrizes de torção de cada átomo na obtenção de posições através da fórmula (2.3).

2.2 Algoritmo *Branch-And-Prune*

Para resolver o PMGDD, Lavor, Liberti e Maculan propuseram em [1] o algoritmo *Branch-And-Prune* (BP), cuja idéia principal consiste em explorar o espaço de busca da seguinte forma: a cada passo i, cada matriz de torção para o átomo i (B_i^1 e B_i^2) é conjugada às matrizes de torção dos átomos anteriores na fórmula (2.3), fornecendo dois posicionamentos para o átomo i (*branching*). Os posicionamentos que não atenderem às distâncias dos pares do conjunto F são descartados (*prunning*, ou poda). Com essa abordagem, a exploração do espaço de busca se dá em forma de árvore, onde cada nível da árvore corresponde a um átomo da molécula.

Inicialmente, como podemos fixar arbitrariamente os três primeiros átomos da molécula em posições que respeitam as distâncias tipo E, inicializamos uma árvore com três nós ligados em série. Em seguida, o algoritmo explora as combinações de matrizes de torção obtendo, em cada nó de nível i, as posições para o átomo i que obedecem às distâncias tipo F. Cada nó da árvore em nível i contém as seguintes informações:

- uma coordenada $x_i \in \mathbb{R}^3$ para o átomo i;
- uma *matriz de torção acumulada*, que é dada por $C_i = \prod_{j=1}^{i} B_i$;

Algoritmo 2.1 Branch-and-Prune

Inicialização:
Criar nós n_1, n_2 e n_3 para os três primeiros átomos da molécula,
    contendo suas respectivas matrizes de torção
Conectar os nós n_1, n_2 e n_3 em série
BranchAndPrune(n_3)

BranchAndPrune(no)
    $j \leftarrow Nivel(no)$
    se $j \leq n$ **então**
        obter matriz de torção acumulada C_{j-1} do pai de no
        calcular matrizes de torção B_j^1 e B_j^2, pela fórmula (2.5)
        $C_j \leftarrow C_{j-1}B_j^1$; $x_j \leftarrow C_j[0,0,0,1]^T$;
        $C_j^{'} \leftarrow C_{j-1}B_j^2$; $x_j^{'} \leftarrow C_j^{'}[0,0,0,1]^T$;
        //testar distâncias tipo F relacionadas a j com uma tolerância ϵ
        $valido \leftarrow true$; $valido' \leftarrow true$
        para cada par $(i,j) \in F$ **faça**
            se $|\|x_j - x_i\| - d_{ij}| > \epsilon$ **então** $valido \leftarrow false$
            se $\left|\left\|x_j^{'} - x_i\right\| - d_{ij}\right| > \epsilon$ **então** $valido' \leftarrow false$
        fim para
        //criar nós com posições validadas
        se $valido$ **então**
            criar um nó z, contendo C_j e x_j
            marcar z como filho de no
            marcar no como pai de z
            BranchAndPrune(z)
        fim se
        se $valido'$ **então**
            criar um nó z', contendo $C_j^{'}$ e $x_j^{'}$
            marcar z' como filho de no
            marcar no como pai de z'
            BranchAndPrune(z')
        fim se
    senão
        imprimir solução formada pelas posições dos nós ancestrais até a raiz
    fim se

Capítulo 3

Sequência de átomos

$$H_\alpha^r - H_N^r - H_{AUX}^r - H_N^{r+1} - H_\alpha^r$$

3.1 Visão Geral

No trabalho de Lavor, Mucherino, Liberti e Maculan [14], foi proposta uma maneira de se obter as possíveis estruturas da cadeia principal de uma proteína, a partir de distâncias apenas entre átomos de hidrogênio, característica que contribui com a aplicabilidade prática do método, como será discutido mais adiante. De modo geral, dada uma proteína monomérica qualquer, considera-se uma sequência específica de átomos para cada um de seus resíduos r: $H_\alpha^r - H_N^r - H_{AUX}^r - H_N^{r+1} - H_\alpha^r$. A figura 3.1 mostra como os átomos dessa sequência são associados aos átomos dos resíduos. Os resíduos dos tipos glicina e prolina têm diferenças em relação aos demais tipos que são significativas na atribuição dos átomos da sequência aos átomos da molécula. A glicina é um tipo de aminoácido sem cadeia lateral e portanto não possui C_β, o que obriga H_{AUX}^r a ter significado diferente dos demais tipos. Como seu C_α é ligado a dois átomos de hidrogênio (diferentemente dos demais aminoácidos cujo C_α é ligado a apenas um), podemos, para a glicina, associar H_{AUX}^r a um dos átomos de hidrogênio ligados ao C_α, que não o átomo associado a H_α^r. A prolina se diferencia dos demais por não ter um átomo de hidrogênio ligado a seu átomo N, já que uma de suas valências é usada para formar um anel com a cadeia lateral. Em [14], tratou-se apenas o caso da glicina.

A sequência $H_\alpha^r - H_N^r - H_{AUX}^r - H_N^{r+1} - H_\alpha^r$, repetida ao longo dos resíduos $r = 1, ..., m$, onde m é a quantidade de resíduos na proteína, com seus dois últimos elementos do último resíduo sendo substituídos pelo hidrogênio da hidroxila C-terminal $H^{m'}$ e sem o elemento H_{AUX}^1, ou

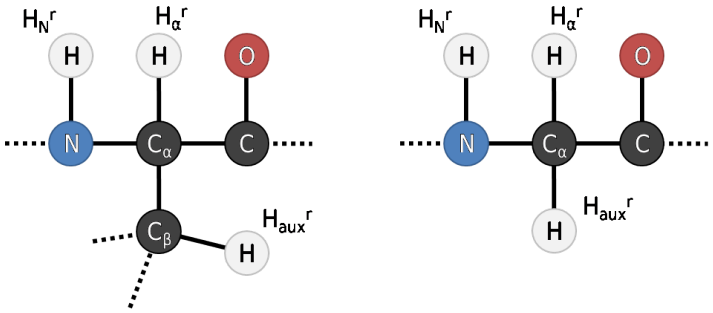


Figura 3.1: Associação entre átomos da sequência $H_\alpha^r - H_N^r - H_{AUX}^r - H_N^{r+1} - H_\alpha^r$ e átomos dos resíduos.
À direita, é ilustrada a associação dos átomos para resíduos do tipo glicina. À esquerda, para os demais tipos (que apresentam C_β).

seja,

$$H_\alpha^1, H_N^1, H_N^2, H_\alpha^1, H_\alpha^2, H_N^2, H_{AUX}^2, H_N^3, H_\alpha^2, ..., H_\alpha^m, H_N^m, H_{AUX}^m, H^{m'},$$

compõe uma cadeia de átomos, cujas posições queremos conhecer, que acompanha a cadeia principal ao longo da proteína. O valor dessa sequência está no fato de que as distâncias interatômicas que devem ser conhecidas para que se obtenha uma instância do PMGDD (atendendo às duas premissas estabelecidas na seção 3.3) são mensuráveis experimentalmente, como mostrado em [14], podendo ter suas coordenadas encontradas através do algoritmo BP.

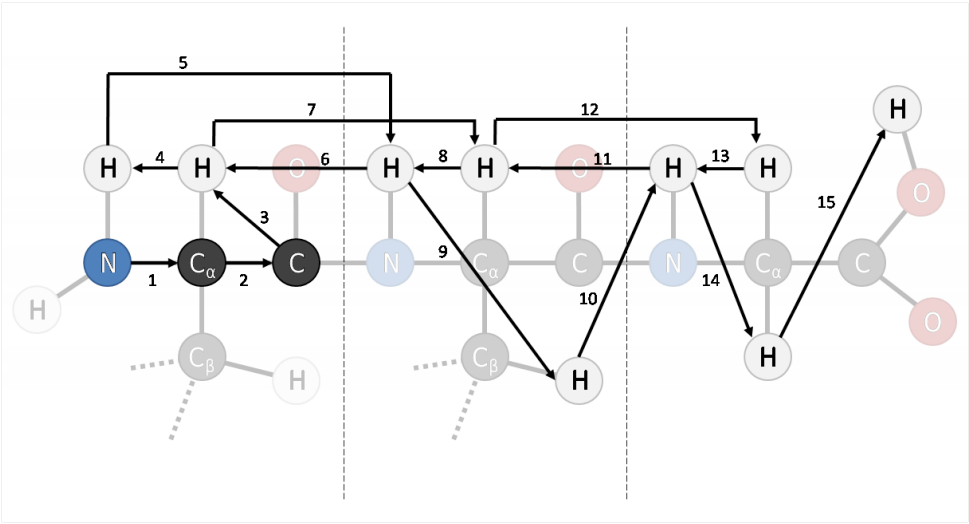


Figura 3.2: Cadeia artificial induzida pela sequência $H_\alpha^r - H_N^r - H_{AUX}^r - H_N^{r+1} - H_\alpha^r$.
Está exemplificado um peptídeo de três resíduos (sendo o terceiro do tipo glicina). As numeração das setas apresenta a ordem dos átomos da cadeia artificial.

Ainda em [14], é mostrado que o BP também pode ser utilizado com a seguinte modificação

no início da sequência anterior:

$$N^1, C^1_\alpha, C^1, H^1_\alpha, H^1_N, H^2_N, H^1_\alpha, H^2_\alpha, H^2_N, H^2_{AUX}, H^3_N, H^2_\alpha, ..., H^m_\alpha, H^m_N, H^m_{AUX}, H^{m'}.$$

A cadeia de átomos formada por essa sequência - à qual nos referiremos doravante como **cadeia artificial** - tem a vantagem de estar no mesmo sistema de coordenadas da cadeia principal, pois ambas as cadeias compartilham átomos (ver figura 3.2).

Uma vez tendo em mãos a instância do PMGDD correspondente à cadeia artificial, pode-se utilizar o algoritmo BP para se obter as estruturas que solucionam essa cadeia. No entanto, tais estruturas terão apenas as coordenadas dos átomos da cadeia artificial, e não da cadeia principal, que é composta pela sequência $N^r - C^r_\alpha - C^r$, para todo $1 \leq r \leq m$. Utilizando o conhecimento de algumas distâncias entre átomos das duas cadeias que variam insignificantemente e são conhecidas *a priori*, é possível chegar às estruturas correspondentes da cadeia principal. A figura 3.3 sintetiza o método descrito nesta seção.

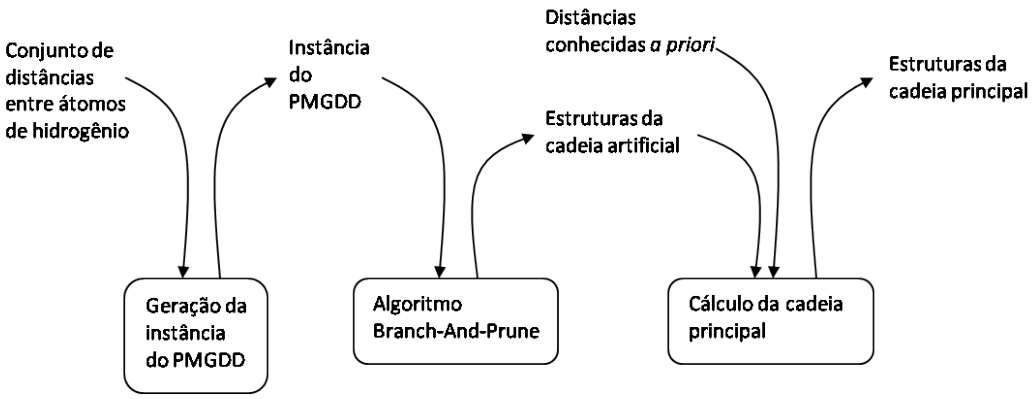


Figura 3.3: Fluxograma: visão geral do método.

3.2 Obtenção da cadeia principal a partir de uma estrutura da cadeia artificial

Abordaremos aqui com mais detalhes o processo de obtenção da cadeia principal a partir de uma estrutura da cadeia artificial proposto em [14], já que na seção 3.3 vamos apresentar uma alternativa a ele.

Após utilizarmos o algoritmo BP para resolver uma instância do PMGDD formada com uma cadeia artificial, obteremos estruturas possíveis para a cadeia artificial. Seja x a posição

que queremos determinar de um átomo da cadeia principal, e b_1, b_2, b_3, b_4 posições conhecidas de outros quatro átomos, com respectivas distâncias até x conhecidas. A posição x pode ser determinada unicamente caso b_1, b_2, b_3 e b_4 não sejam coplanares [16, 17].

Seja d_{x,b_i} a distância euclidiana entre x e b_i, para todo $i \in \{1, 2, 3, 4\}$. Para encontrarmos as coordenadas de x, o seguinte sistema deve ser resolvido:

$$\begin{cases} \|x - b_1\| &= d_{x,b_1} \\ \|x - b_2\| &= d_{x,b_2} \\ \|x - b_3\| &= d_{x,b_3} \\ \|x - b_4\| &= d_{x,b_4} \end{cases}. \tag{3.1}$$

Em [14] mostra-se que se o sistema (3.1) tem uma solução, então ela também é solução do sistema linear de 4 variáveis

$$-2 \begin{bmatrix} 1 & b_1^T \\ 1 & b_2^T \\ 1 & b_3^T \\ 1 & b_4^T \end{bmatrix} \begin{pmatrix} t \\ x \end{pmatrix} = \begin{bmatrix} d_{x,b_1}^2 - \|b_1\|^2 \\ d_{x,b_2}^2 - \|b_2\|^2 \\ d_{x,b_3}^2 - \|b_3\|^2 \\ d_{x,b_4}^2 - \|b_4\|^2 \end{bmatrix}. \tag{3.2}$$

Se, além disso, $t \approx \dfrac{\|x\|^2}{2}$ (dentro de uma certa tolerância), então x pode ser considerada uma boa estimativa da posição que queremos determinar.

Com o que foi disposto acima, é possível determinar as posições de todos os átomos da cadeia principal resolvendo-se sistemas lineares como em (3.2). A tabela 3.1 mostra os diferentes casos de uso do sistema (3.2), para se determinar as posições dos átomos da cadeia principal N^r, C_α^r e C^r, para $i > 1$, uma vez que N^1, C_α^1 e C^1 já têm suas posições conhecidas na própria cadeia artificial.

Átomo a ser determinado	Átomos utilizados			
N^r	C_α^{r-1}	C^{r-1}	H_N^r	H_α^r
C_α^r	C^{r-1}	N^r	H_N^r	H_α^r
C^r	N^r	C_α^r	H_α^r	H_N^{r+1}

Tabela 3.1: Átomos utilizados para determinar átomos da cadeia principal.

3.3 Uma nova forma de se obter a cadeia principal a partir de uma estrutura da cadeia artificial

Examinemos novamente as condições apresentadas na seção 3.2: seja x a posição que queremos determinar de um átomo da cadeia principal, e b_1, b_2, b_3, b_4 posições conhecidas de outros quatro átomos, com respectivas distâncias até x conhecidas.

Seja $B = \{b_1, b_2, b_3, b_4\}$ o conjunto das posições conhecidas. Como consequência imediata das condições acima, conhecemos todas as distâncias euclidianas $d_{i,j}$ e $d_{i,x}$, onde $i, j \in B$. Consideremos um conjunto $A = \{x\} \cup C$ de quatro elementos, onde C é um subconjunto qualquer de B contendo três elementos. Podemos formular uma pequena instância I do PMGDD de quatro átomos correspondentes às posições de A. As distâncias de I são dadas por $d_{f(i),f(j)} = d_{i,j}$, onde $i, j \in A$ e $f : A \mapsto \{1, 2, 3, 4\}$ é uma bijeção. Por exemplo, podemos escolher $C = \{b_1, b_3, b_4\}$ e $f = \{(b_1, 1), (b_3, 2), (b_4, 3), (x, 4)\}$.

Ao solucionar essa instância, o algoritmo BP encontra as duas realizações possíveis R_1 e R_2 para o intervalo $[1, 4]$. Já que na instância foram usadas as distâncias das posições A, sabemos que uma das realizações é sobreponível à estrutura das posições A (cujo elemento x ainda queremos conhecer). No entanto, as soluções do BP são estruturas com uma orientação específica (devido à fixação dos três primeiros átomos na origem, eixo OX e plano OXY respectivamente) que não necessariamente é a mesma orientação das posições A. Para resolver isso, podemos alinhar R_1 e R_2 às posições $C \subset A$ através de rotações e translações, produzindo duas novas estruturas $R_1^{'}$ e $R_2^{'}$ tais que $R_1^{'}(f(i)) = R_2^{'}(f(i)) = i$, $i \in C$, obtendo as duas posições possíveis de x na orientação correta. Vale observar que se R_1 e R_2 são distintas, suas transformações euclidianas $R_1^{'}$ e $R_2^{'}$ também o serão. Nesse caso, como três de seus átomos já têm mesma posição, obrigatoriamente $R_1^{'}$ e $R_2^{'}$ se diferenciam quanto à posição x, ou seja, $R_1^{'}(f(x)) \neq R_2^{'}(f(x))$.

Agora, a posição $b^{'} \in B - C$ que tinha ficado fora da instância I pode ser usada para determinar qual das duas realizações alinhadas é a correta, uma vez que conhecemos $d_{b',x}$. A realização correta $R_k^{'}$, portanto, é aquela que satisfaz

$$\left\| R_k^{'}(f(x)) - b^{'} \right\| = d_{b',x} \tag{3.3}$$

e dela obtemos a posição que queríamos encontrar, $x = R_k^{'}(f(x))$. A situação final é ilustrada na figura 3.4.

Com vistas à aplicação prática do método acima, devemos considerar a importância de alguns

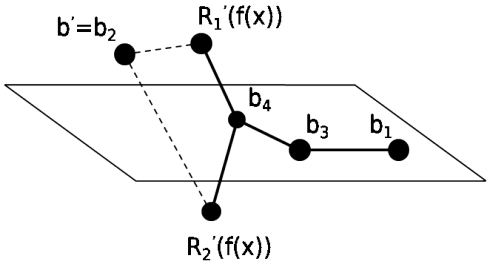


Figura 3.4: Inferência de uma posição a partir de quatro outras conhecidas.

fatores na busca da realização correta $R_k^{'}$. Na prática, por questões de precisão numérica e/ou utilização de parâmetros (valores de distância $d_{i,j}$ e $d_{i,x}$, $i,j \in B$) aproximados, é bastante plausível que a igualdade da equação (3.3) não seja atingida para nenhuma das duas realizações. Nesse cenário, $R_k^{'}$ pode ser vista como aquela que mais se aproxima do caso ideal em (3.3), ou seja,

$$R_k^{'}(f(x)) = \min_{R_i^{'} \in \left\{R_1^{'}, R_2^{'}\right\}} \left| \left\| R_i^{'}(f(x)) - b^{'} \right\| - d_{b^{'},x} \right|. \tag{3.4}$$

Uma característica importante desse método é que ele não requer a resolução de sistemas lineares ou quadráticos. Além disso, ambas as soluções para a posição x foram determinadas apenas pelos três átomos do conjunto C. O átomo $b^{'}$ só serviu para escolhermos uma dentre essas duas opções (através do teste com $d_{b^{'},x}$). No método apresentado na seção 3.2, x é sempre influenciada pelas quatro posições e distâncias conhecidas. Acredita-se então que no método atual há menor propagação de erros para a posição x. Pode ser feito um estudo mais formal sobre as incertezas envolvidas no cálculo da equação (3.4), mas como os resultados práticos foram satisfatórios, tal assunto será abordado em trabalhos futuros.

O algoritmo 3.1 resume o método descrito.

Algoritmo 3.1 InferePos

Entrada: $b^{'}$ e C tal que as distâncias $d_{i,j}$ e $d_{i,x}$, $i,j \in \left(C \cup \left\{b^{'}\right\}\right)$ sejam conhecidas.
Saída: a posição x.
InferePos($b^{'}$,C)
  seja $f : A \mapsto \{1,2,3,4\}$ uma bijeção qualquer;
  seja I a instância do PMGDD tal que $n = 4$, $S = \{(f(i), f(j))\}$ e $d_{f(i),f(j)} = d_{i,j}$, $i,j \in A$;
  $\{R_1, R_2\} \leftarrow$ soluções encontradas pelo BranchAndPrune para a instância I;
  $R_1^{'} \leftarrow$ alinhamento de R_1 com base nas posições de C;
  $R_2^{'} \leftarrow$ alinhamento de R_2 com base nas posições de C;
  retorna $\min_{R_i^{'} \in \left\{R_1^{'}, R_2^{'}\right\}} \left| \left\| R_i^{'}(f(x)) - b^{'} \right\| - d_{b^{'},x} \right|$;

3.4 Experimentos Computacionais

Validamos o método proposto nesse capítulo, gerando instâncias do PMGDD com base na cadeia artificial descrita na seção 3, para estruturas protéicas de uma base de testes.

Para construir a base de testes, selecionamos um conjunto de estruturas protéicas monoméricas já resolvidas, com similaridade abaixo de 30%, obtidas por experimentos de RMN e que foram acrescentadas ao Protein Data Bank entre o ano de 2008 e 2010. Deste conjunto, selecionamos ao acaso algumas proteínas para os testes. Para atender à atual formulação do PMGDD, as distâncias tipo E levadas em conta foram distâncias inter-atômicas exatas obtidas a partir de cada estrutura conhecida (o que não é dado pelos experimentos de RMN, que, em vez de valores exatos, fornecem faixas para tais distâncias). As distâncias conhecidas *a priori* para geração das cadeias peptídicas a partir de soluções foram obtidas usando-se toda base de testes (valores médios de distâncias inter-atômicas).

Para podermos trabalhar com estruturas de proteínas reais, o que não havia sido feito em [14], tivemos que tratar resíduos do tipo prolina. Portanto, consideramos que H_N^r, no caso da prolina, se refere à posição do próprio átomo N do resíduo (como se ele tivesse suas distâncias detectadas em laboratório). Consequentemente, para esses resíduos, a posição do nitrogênio da cadeia principal ja é dada de antemão na cadeia artificial, por H_N^r.

Implementamos o algoritmo BP para solucionar as instâncias, obtendo estruturas tridimensionais da cadeia artifical. Para cada solução, aplicamos o algoritmo 3.1 para calcular a cadeia principal correspondente. As implementações foram feitas em linguagem Java e executadas em ambiente Intel Core 2 Duo 2.2 GHz, 2GB RAM, Sistema Operacional Linux.

Nesse experimento, estivemos interessados em medir a similaridade entre a cadeia principal calculada e a cadeia principal da estrutura original, excluindo-se os efeitos causados por seus estados de rotação e translação. Para isso, medimos o LRMSD (least root mean square deviation) das cadeias principais obtidas, R, em relação à cadeia principal da estrutura original, R^*. O LRMSD é o RMSD (root mean square deviation) mínimo entre duas realizações, dentre todas as rotações e translações possíveis que elas podem assumir (para obtermos o alinhamento espacial ótimo, implementamos o método proposto em [18]). O RMSD entre duas realizações A e B, com mesma quantidade de átomos n, é dado pela seguinte expressão:

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|A(i) - B(i)\|}. \tag{3.5}$$

A tabela 3.2 mostra os resultados dos experimentos. O algoritmo BP foi executado sobre as instâncias com tolerância $\epsilon = 10^{-5}$ Å, por um tempo máximo de duas horas, guardando-se sempre o melhor LRMSD, E_{min}, dentre todas as soluções encontradas.

Instância	Res	Soluções	E_{min}(Å)
2K1I	32	1274386	0,4002
2WH9	36	1130582	0,9694
2KB8	37	1109953	0,4044
2VY5	37	1010233	4,9210
2KGH	39	1146169	0,4027
2K9D	44	905086	0,3966
2JZM	53	799015	0,4023
2KKR	59	736034	0,3598
2WNM	59	632249	0,5045
2KEC	68	620378	0,3373
2KN9	81	552936	0,2958
2VXE	88	426019	3,1961
2VXF	88	420314	0,6360
2K22	90	491412	0,5313
2KC1	91	485518	0,2226

Instância	Res	Soluções	E_{min}(Å)
2KUO	91	431848	0,2204
2ROK	100	425119	0,3454
2ROP	142	260925	16,4201
2K3A	155	264731	2,9535
2WCY	155	238348	0,6131
2KTU	164	260159	0,1805
2RR6	168	219780	0,2641
2K7N	203	213998	0,1662
2K0L	210	0	-
2KRG	216	0	-
2RON	242	150352	0,1757
2K6H	248	166283	0,2373
2KDO	252	169870	0,1761
2K4T	285	145565	0,1750
2ROQ	343	109101	0,1649

Tabela 3.2: Resultados computacionais de comparação das soluções do BP com a cadeia original.

3.5 Discussão

Fizemos os primeiros experimentos da sequência $H^r_\alpha - H^r_N - H^r_{AUX} - H^{r+1}_N - H^r_\alpha$ utilizando proteínas reais. Tal sequência gera instâncias com 5 átomos por resíduo da cadeia peptídica original. Utilizamos a maneira alternativa de se calcular a cadeia principal a partir da cadeia artificial, proposta na seção 3.3. Nenhuma instância teve seu espaço de busca completamente explorado. Pôde-se perceber que as instâncias testadas tiveram um número elevado de soluções (muitas vezes chegando a números da ordem de 10^6). Testes feitos separadamente mostraram que a maior parte dessas soluções é geometricamente idêntica, apesar de serem resultados de caminhos distintos na árvore (a quantidade de soluções duplicatas por estrutura encontrada pelo algoritmo chegou a números da ordem de 10^5, para algumas instâncias). Por conta disso, a árvore percorrida pelo algoritmo para cada instância era muito densa e o tempo de duas horas para cada instância não foi suficiente para que todos os nós da árvore (o espaço de busca) tivessem sido explorados. Aproximadamente metade das instâncias tiveram soluções encontradas nesse período com scores razoavelmente satisfatórios, com E_{min} abaixo de 0.4 Å.

Esses resultados podem ser úteis para aproximar o PMGDD da aplicabilidade prática. A formulação atual do PMGDD ainda não nos permite trabalhar com faixas de distâncias de pares

do conjunto E, mas a facilidade com que o BP encontrou soluções similares à cadeia original (mesmo com a dificuldade de se terminar a exploração do espaço de busca) pode nos indicar que as distâncias tipo F da cadeia artificial apresentada nesse capítulo restringem o espaço de busca razoavelmente bem, de modo a nos direcionar a boas soluções. Por outro lado, a existência de muitas soluções geometricamente idênticas é um alvo em potencial para a busca de melhores estratégias de exploração do espaço de busca. O método descrito no próximo capítulo, por exemplo, não apresentou esse problema nos testes realizados.

Capítulo 4

Sequência de átomos $H_N^r - H_\alpha^r - C^r$

4.1 Motivação

A sequência de átomos contemplada no capítulo 3 utiliza átomos H_β (exceto para resíduos do tipo glicina). Isso implica que algumas distâncias tipo E sejam relacionadas a esses átomos. Por causa da maior liberdade conformacional do H_β (que, para a maioria dos resíduos, pode girar livremente em torno do C_β), as medidas em experimentos de RMN sobre ele costumam resultar em faixas de distância maiores, o que tornaria menos realistas futuras aplicações práticas baseadas no PMGDD e no BP, já que a formulação atual do problema exige que as distâncias tipo E sejam exatas. Nesse capítulo, proporemos uma alternativa a essa cadeia (uma outra cadeia artificial) que também nos permite obter a cadeia principal, mas que não incorpora medidas sobre átomos H_β às distâncias tipo E.

4.2 Uma nova cadeia artificial

Novamente, estaremos tratando de proteínas monoméricas. Consideremos a seguinte sequência de átomos que ocorre em seus resíduos r: $H_N^r - H_\alpha^r - C^r$. Apesar de não conter apenas átomos de hidrogênio, essa sequência compõe uma cadeia de átomos que pode ser resolvida utilizando-se apenas distâncias entre átomos de hidrogênio e distâncias conhecidas *a priori*, mostrada na figura 4.1. O princípio utilizado baseia-se na estabilidade da geometria dos átomos de carbono e nitrogênio da cadeia principal. Em termos gerais, rearranjamos a equação (2.2) a fim de obtermos distâncias relacionadas aos C^r a partir de distâncias já conhecidas.

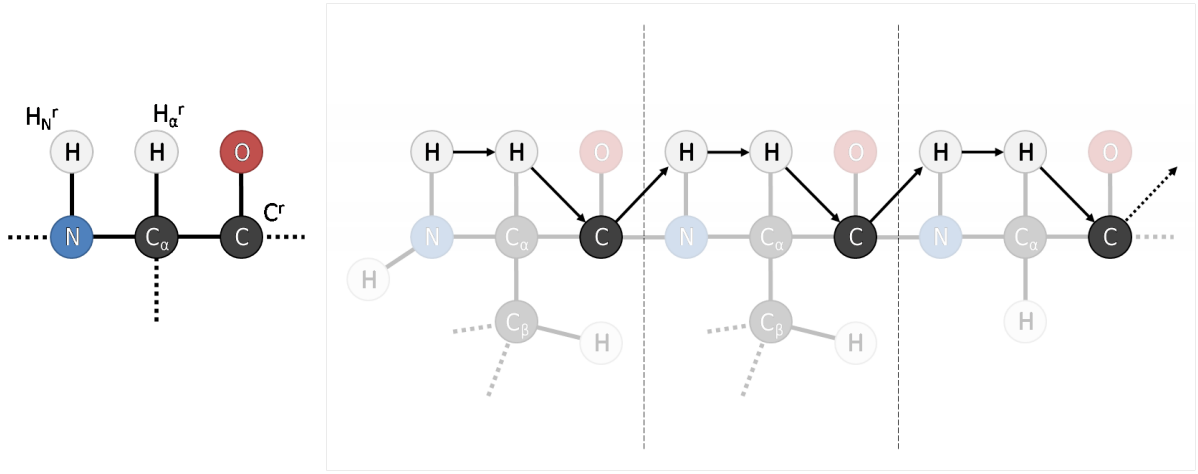


Figura 4.1: Cadeia induzida por $H_N^r - H_\alpha^r - C^r$.

4.3 Distâncias associadas aos átomos da cadeia

Para que possa ser resolvida pelo BP, a cadeia induzida por $H_N^r - H_\alpha^r - C^r$ precisaria ter os seguintes pares $(i, j) \in E$ com distância conhecida (ver figura 4.1):

- (H_N^r, H_α^r), (H_N^r, C^r), $\left(H_N^r, H_N^{r+1}\right)$
- (H_α^r, C^r), $\left(H_\alpha^r, H_N^{r+1}\right)$, $\left(H_\alpha^r, H_\alpha^{r+1}\right)$
- $\left(C^r, H_N^{r+1}\right)$, $\left(C^r, H_\alpha^{r+1}\right)$, $\left(C^r, C^{r+1}\right)$

Dos pares relacionados acima, aqueles que contêm apenas átomos de hidrogênio podem ter suas distâncias obtidas a partir de experimentos de RMN, então as consideraremos conhecidas. Os pares (H_α^r, C^r) e $\left(C^r, H_N^{r+1}\right)$ têm distâncias conhecidas *a priori*. No caso dos demais pares, (H_N^r, C^r), $\left(C^r, H_\alpha^{r+1}\right)$ e $\left(C^r, C^{r+1}\right)$, vamos utilizar o resultado descrito a seguir.

Sejam $i-3$, $i-2$, $i-1$, i átomos com

1. distâncias euclidianas $d_{j,k}$ conhecidas, onde $j \geq i-3, k \leq i$ e $k-j \leq 2$,
2. cosseno do ângulo de torção entre os quatro átomos, $\cos\omega_i$, conhecido,

cuja distância $d_{i-3,i}$ queiramos descobrir. Podemos obter $d_{i-3,i}$ usando a lei dos cossenos para ângulos de torção de maneira alternativa:

$$\begin{aligned} \cos\omega_i &= \frac{\cos\gamma - \cos\alpha\cos\beta}{\sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta}} \Rightarrow \\ \cos\gamma &= (\cos\omega_i)\sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta} + \cos\alpha\cos\beta \Rightarrow \end{aligned}$$

$$\frac{d^2_{i-3,i-2} + d^2_{i-2,i} - d^2_{i-3,i}}{2d_{i-3,i-2}d_{i-2,i}} = (\cos\omega_i)\sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta} + \cos\alpha\cos\beta \Rightarrow$$

$$d^2_{i-3,i-2} + d^2_{i-2,i} - d^2_{i-3,i} = 2d_{i-3,i-2}d_{i-2,i}\left[(\cos\omega_i)\sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta} + \cos\alpha\cos\beta\right] \Rightarrow$$

$$d_{i-3,i} = \sqrt{d^2_{i-3,i-2} + d^2_{i-2,i} - 2d_{i-3,i-2}d_{i-2,i}\left[(\cos\omega_i)\sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta} + \cos\alpha\cos\beta\right]}. \tag{4.1}$$

Os pares (H^r_N, C^r), $\left(C^r, H^{r+1}_\alpha\right)$ e $\left(C^r, C^{r+1}\right)$ têm a primeira condição atendida ao serem conjugados a átomos da cadeia principal, formando diferentes quádruplas de átomos $i-3$, $i-2$, $i-1$, i cujas distâncias euclidianas $d_{j,k}$, onde $j \geq i-3, k \leq i$ e $k-j \leq 2$, são conhecidas *a priori* (por serem comprimentos de ligações covalentes ou distâncias entre átomos com ângulos de ligação fixos), como mostra a figura 4.2.

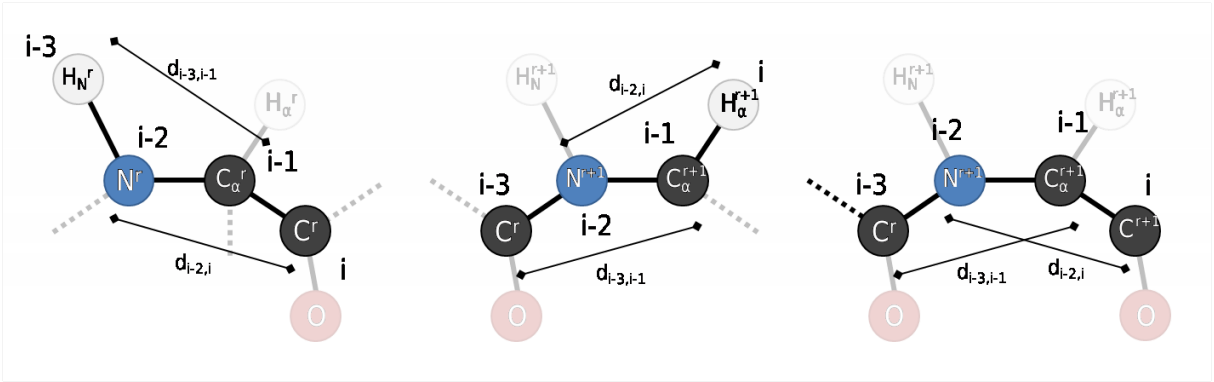


Figura 4.2: Torções dos pares envolvendo átomos de carbono e suas distâncias conhecidas *a priori*.

Podemos nos valer da estabilidade geométrica dos átomos de carbono e nitrogênio da cadeia principal para descobrir ω_i, tendo como base o ângulo de torção ω' de uma quádrupla de átomos auxiliar q'. Com isso, a condição 2 será atendida.

Por exemplo, vamos contemplar o caso do par (H^r_N, C^r). Considere $q = (H^r_N, N^r, C^r_\alpha, C^r)$ e $q' = (H^r_N, N^r, C^r_\alpha, H^r_\alpha)$. O ângulo de torção de q, ω_i, é definido pelos vetores p_1 (perpendicular ao plano formado por H^r_N, N^r, C^r_α) e p_2 (perpendicular ao plano formado por N^r, C^r_α, C^r). Seu valor é desconhecido a princípio. O ângulo de torção de q', ω', é definido pelos vetores p_1 e p'_2 (perpendicular ao plano formado por N^r, C^r_α, H^r_α). Temos dois valores possíveis para ω' (já que todas as distâncias inter-atômicas de q' são conhecidas e pela equação (2.2) obtemos o cosseno desse ângulo).

Como ambas as quádruplas compartilham do mesmo vetor p_1, a diferença entre seus ângulos de torção $\triangle\omega$ é determinada somente pelos vetores p_2 e $p_2^{'}$. O átomo C_α^r não apresenta variação significativa de seus comprimentos e ângulos de ligação, nem tem a ordem de seus ligantes trocada (pois conhecemos *a priori* a quiralidade do resíduo que estamos calculando), para diferentes proteínas e resíduos onde q e $q^{'}$ forem situadas, fazendo com que o ângulo entre p_2 e $p_2^{'}$ esteja dentro de uma faixa de valores bastante restrita. Consequentemente, $\triangle\omega$ também é restrito (fato que pode ser verificado intuitivamente e empiricamente). Isso significa que quando conhecemos o ângulo de uma dessas torções, podemos deduzir o ângulo da outra:

$$\omega_i = \omega^{'} + \triangle\omega. \tag{4.2}$$

O princípio descrito pode ser aplicado analogamente aos outros pares, $\left(C^r, H_\alpha^{r+1}\right)$ e $\left(C^r, C^{r+1}\right)$. A tabela 4.1 relaciona as quádruplas utilizadas para cada um e as medições de $\triangle\omega$ sobre a base de testes. Dessa forma, a partir dos dois valores possíveis de $\omega^{'}$, obtemos dois valores para ω_i e consequentemente dois valores para $\cos\omega_i$, atendendo à condição 2. Ao utilizá-los na equação (4.1), obtemos dois valores para $d_{i-3,i}$ (sendo um deles o correto), o que não acompanha a definição do PMGDD (que prevê apenas um valor para cada distância $d_{i,j}$ associada a um par $(i,j) \in E$). Deveremos então reformular o problema e o algoritmo que o resolve para tratar instâncias desse tipo.

	Par (H_N^r, C^r)	Par $\left(C^r, H_\alpha^{r+1}\right)$	Par $\left(C^r, C^{r+1}\right)$
Quádrupla q	$(H_N^r, N^r, C_\alpha^r, C^r)$	$(C^r, N^{r+1}, C_\alpha^{r+1}, H_\alpha^{r+1})$	$(C^r, N^{r+1}, C_\alpha^{r+1}, C^{r+1})$
Quádrupla $q^{'}$	$(H_N^r, N^r, C_\alpha^r, H_\alpha^r)$	$(H_N^{r+1}, N^{r+1}, C_\alpha^{r+1}, H_\alpha^{r+1})$	$(H_N^{r+1}, N^{r+1}, C_\alpha^{r+1}, H_\alpha^{r+1})$
$\overline{\triangle\omega}$ (Prolina)	-2,1413	3,1278	0,9865
$\sigma\left(\triangle\omega\right)$ (Prolina)	0,0121	0,0171	0,0113
n (Prolina)	3754	3754	3754
$\overline{\triangle\omega}$ (Glicina)	-2,1056	3,1268	1,0212
$\sigma\left(\triangle\omega\right)$ (Glicina)	0,0102	0,0193	0,0211
n (Glicina)	6656	6463	6463
$\overline{\triangle\omega}$ (Outros)	-2,0770	3,1213	1,0443
$\sigma\left(\triangle\omega\right)$ (Outros)	0,0205	0,0412	0,0511
n (Outros)	68204	67809	67809

Tabela 4.1: Torções usadas para inferir as distâncias associadas a (H_N^r, C^r), $\left(C^r, H_\alpha^{r+1}\right)$ e $\left(C^r, C^{r+1}\right)$.

4.4 Reformulação do PMGDD

Segundo o que foi apresentado na seção 4.3, cada par (H_N^r, C^r), $\left(C^r, H_\alpha^{r+1}\right)$ e $\left(C^r, C^{r+1}\right)$ terá duas distâncias associadas possíveis. Para incorporarmos essa característica ao problema, vamos redefinir o PMGDD. Uma vez que a ordem dos átomos da cadeia artificial é $H_N^r - H_\alpha^r - C^r$, o conjunto dos pares de átomos que terão duas distâncias $d'_{i,j}$ e $d''_{i,j}$ associadas é definido por $E' = E_{\left(H_N^r, C^r\right)} \cup E_{\left(C^r, H_\alpha^{r+1}\right)} \cup E_{\left(C^r, C^{r+1}\right)}$, onde

$$\begin{aligned} E_{\left(H_N^r, C^r\right)} &= \{(x, x+2) \in E | x = 3k+1, k \in \mathbb{Z}\}, \\ E_{\left(C^r, H_\alpha^{r+1}\right)} &= \{(x, x+2) \in E | x = 3k, k \in \mathbb{Z}\}, \\ E_{\left(C^r, C^{r+1}\right)} &= \{(x, x+3) \in E | x = 3k, k \in \mathbb{Z}\}. \end{aligned}$$

No que segue, redefinimos o PMGDD da seguinte forma:

Encontrar as coordenadas cartesianas $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n} \in \mathbb{R}^3$ para os n átomos de uma molécula que minimizem a função

$$f(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}) = \sum_{(i,j) \in S - E'} \left(\|\mathbf{x_i} - \mathbf{x_j}\|^2 - d_{i,j}^2 \right)^2 + \sum_{(i,j) \in E'} \left[\min_{d_{i,j} \in \left\{ d'_{i,j}, d''_{i,j} \right\}} \left(\|\mathbf{x_i} - \mathbf{x_j}\|^2 - d_{i,j}^2 \right)^2 \right], \tag{4.3}$$

onde $S - E'$ é o conjunto de pares de átomos (i, j) cuja distância tem valor conhecido $d_{i,j}$ e E' é conjunto de pares de átomos (i, j) cuja distância é um dos valores conhecidos $d'_{i,j}$ ou $d''_{i,j}$.

As matrizes de torção apresentadas na seção 2.1.2 continuam sendo úteis para encontrarmos as soluções do problema, mas a existência de diferentes valores para algumas distâncias resultará na existência de diferentes matrizes de torção, pois tais distâncias participam das equações (2.4) e (2.5). Assim, uma escolha dos valores dessas distâncias configura uma situação semelhante ao PMGDD original, quando cada distância $(i, j) \in E$ está associada a somente um valor.

A equação (2.4) contém distâncias dos pares $(1, 2)$, $(1, 3)$ e $(2, 3)$. Desses, o par $(1, 3)$ corresponde aos átomos $\left(H_N^1, C^1\right)$, que têm dois valores possíveis de distância. Assim são possíveis duas matrizes de torção, B'_3 e B''_3, uma para cada valor de $d_{1,3} \in \left\{ d'_{1,3}, d''_{1,3} \right\}$.

Para a equação (2.5), que se aplica aos demais átomos i, $4 \leq i \leq n$, temos três cenários: o i-ésimo átomo pode ser H_N^r, H_α^r ou C^r. A figura 4.3 ilustra esses casos, detalhados a seguir. No caso do H_N^r, o único par envolvido na equação com dois valores de distância é $(i-3, i-1)$, que corresponde aos átomos $\left(H_N^{r-1}, C^{r-1}\right)$. Sendo assim, com $d'_{i-3,i-1}$ obtemos as matrizes

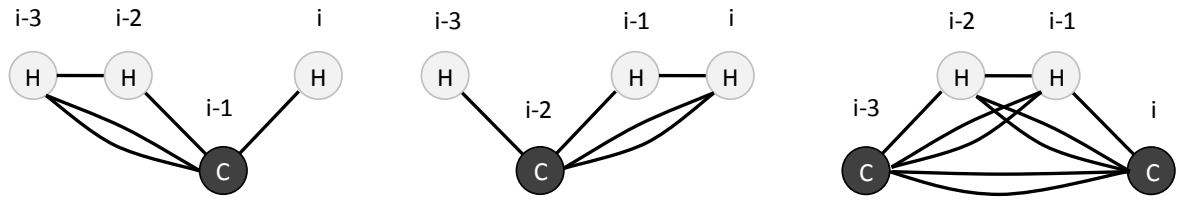


Figura 4.3: Os três tipos de quádruplas aplicáveis à equação (2.5) e suas distâncias.

de torção $B_i^{1'}$ e $B_i^{2'}$, enquanto que com $d''_{i-3,i-1}$ obtemos as matrizes $B_i^{1''}$ e $B_i^{2''}$. No caso do H_α^r, similarmente, há somente um par na equação com dois valores de distância, $(i-2, i)$, correspondendo aos átomos $\left(C^{r-1}, H_\alpha^r\right)$, cujas alternativas nos dão também quatro matrizes de torção $B_i^{1'}$, $B_i^{2'}$, $B_i^{1''}$ e $B_i^{2''}$.

O caso do C^r é mais complexo, pois nele vários pares têm distância com mais de um valor possível: $(i-3, i-1)$, $(i-3, i)$ e $(i-2, i)$, que correspondem respectivamente a $\left(C^{r-1}, H_\alpha^r\right)$, $\left(C^{r-1}, C^r\right)$ e (H_N^r, C^r). À primeira vista, deveríamos combinar as diferentes alternativas de distância de cada par, obtendo, feitas as contas, 16 matrizes de torção. Entretanto, de acordo com a tabela 4.1 (e corrigidos os índices dos átomos), todos os três pares utilizam uma mesma quádrupla auxiliar $q' = (H_N^r, N^r, C_\alpha^r, H_\alpha^r)$. Se os dois valores possíveis para seu ângulo de torção ω' são iguais, teremos apenas um conjunto de valores para $(i-3, i-1)$, $(i-3, i)$ e $(i-2, i)$. Caso contrário, digamos que os valores possíveis para ω' sejam a e b. Seria contraditório, por exemplo, combinar a distância $d'_{i-3,i-1}$, calculada com base em $\omega' = a$, com $d''_{i-2,i}$, calculada com base em $\omega' = b$, uma vez que ω' assume apenas um dos valores, a ou b. A matriz de torção gerada por essa combinação não tem significado útil, pois os pares $(i-3, i)$ e $(i-2, i)$ não podem ter esses valores de distância simultaneamente. Então, novamente, precisamos considerar apenas quatro matrizes: as matrizes $B_i^{1'}$ e $B_i^{2'}$, geradas com as distâncias $d'_{i-3,i-1}$, $d'_{i-3,i}$, $d'_{i-2,i}$, decorrentes de $\omega' = a$, e as matrizes $B_i^{1''}$ e $B_i^{2''}$, geradas com as distâncias $d''_{i-3,i-1}$, $d''_{i-3,i}$, $d''_{i-2,i}$, decorrentes de $\omega' = b$. Em resumo, seja qual for o átomo i, $4 \leq i \leq n$, temos quatro matrizes de torção a serem consideradas.

4.5 Reformulação do Algoritmo BP

Para solucionar o problema redefinido, vamos modificar o algoritmo Branch-and-Prune para que ele seja capaz de lidar com a maior quantidade de matrizes de torção a cada átomo i. Antes de começarmos a percorrer o espaço de busca através de árvores quaternárias, de maneira cegamente análoga ao BP original, podemos nos atentar para uma outra propriedade da cadeia

artificial proposta nesse capítulo, a fim de obtermos um algoritmo mais eficiente.

Para cada resíduo r, as matrizes de torção de átomos H^r_α, C^r e H^{r+1}_N são calculadas com distâncias baseadas no mesmo ângulo de torção $\omega^{'}$ da quádrupla $q^{'} = (H^r_N, N^r, C^r_\alpha, H^r_\alpha)$, que denotaremos como ω^r. Logo, também não faria sentido combinar matrizes desses átomos que fossem baseadas em diferentes valores de ω^r. Isso significa que, assumido um valor para ω^r, precisamos considerar apenas duas matrizes de torção para cada um dos três átomos H^r_α, C^r e H^{r+1}_N.

Podemos então percorrer o espaço de busca da seguinte forma: para cada resíduo r, antes de uma tripla de átomos H^r_α, C^r e H^{r+1}_N, ramificamos dois nós, um para cada valor de ω^r, os quais chamaremos de **nós de torção**. Em cada ramo de ω^r, as matrizes de H^r_α, C^r e H^{r+1}_N calculadas com distâncias decorrentes do valor respectivo de ω^r são combinadas, gerando nós que chamaremos de **nós de átomo**. Esses nós são ramificados e podados como feito no BP original, diferentemente dos nós de torção, que nunca sofrem poda (pois sempre queremos considerar ambos os valores de ω^r). O algoritmo encontra uma solução quando um nó de átomo correspondente ao último átomo da cadeia artificial, C^m, é aberto.

Dessa forma, a exploração do espaço de busca se dá em forma de árvore binária (exemplificada na figura 4.4), tal que cada resíduo r acrescenta quatro níveis à árvore: um nível para o nó de torção do resíduo e três níveis para nós de átomo. O método aqui descrito está sintetizado no algoritmo 4.1.

Cada *nó de átomo* em nível i possui:

- uma coordenada $x_i \in \mathbb{R}^3$ para o átomo i;
- uma matriz de torção acumulada C_i;
- ponteiros para os filhos e pai.

Cada *nó de torção* em nível i possui:

- um valor binário t^r que representa uma escolha para o ângulo de torção ω^r, onde $r = \frac{(i-2)}{4} + 1$;
- ponteiros para os filhos e pai.

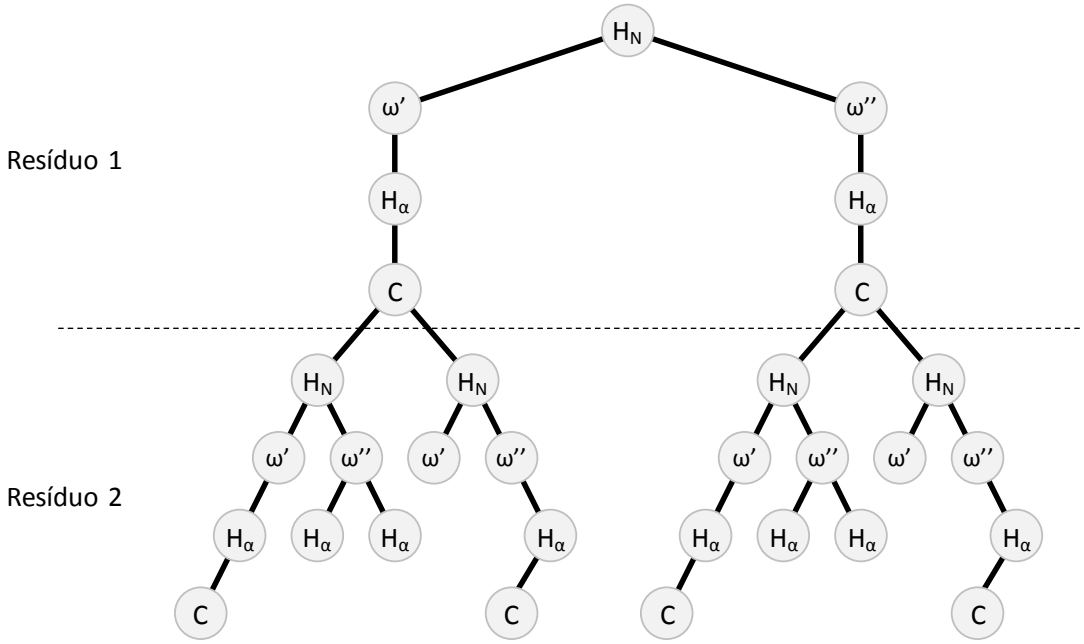


Figura 4.4: Árvore binária percorrida pelo BP modificado.
Ilustração da exploração do espaço de busca de uma instância de uma proteína com dois resíduos. Os nós de átomo do primeiro resíduo não são ramificados, pois, analogamente ao BP original, são os três primeiros átomos da realização, podendo ser fixados na origem, em um eixo e em um plano.

4.6 Obtenção da cadeia principal a partir de uma estrutura da cadeia artificial

Diferentemente do caso da cadeia artificial apresentada no capítulo 3, é fácil usar as realizações da cadeia induzida por $H_N^r - H_\alpha^r - C^r$, obtidas pelo BP modificado, para chegarmos à cadeia principal. Como os átomos C^r já têm suas posições conhecidas, falta descobrir as posições de N^r e C_α^r.

Os átomos H_N^r, H_α^r e C^r, além de já estarem fixados em alguma posição encontrada pelo algoritmo, têm distâncias em relação a C_α^r e N^r conhecidas *a priori*. Escolhendo-se arbitrariamente calcular primeiro C_α^r, há informação suficiente para serem calculadas duas posições desse átomo no $\mathbb{R}^3$, a partir das conhecidas. Em seguida, para cada posição calculada de C_α^r, teremos uma situação que já foi discutida anteriormente, com quatro posições conhecidas a fim de se determinar uma quinta, no caso, N^r. Pelo procedimento descrito em 3.2 ou pelo que foi proposto em 3.3, chegaremos a uma única posição de N^r para cada posição de C_α^r, totalizando duas realizações da cadeia principal do resíduo r.

As duas realizações encontradas para o resíduo r são simétricas em relação ao plano definido por H_N^r, H_α^r e C^r, uma vez que as duas posições de C_α^r são simétricas em relação a esse plano.

Algoritmo 4.1 Branch-And-Prune Modificado

Inicialização:
criar um nó H_N^1 com matriz de torção correspondente à origem
BPModificado(H_N^1)

BPModificado(*pai*)
se ($pai \neq C^m$) **então**
se *pai* é do tipo H_N^r **então**
criar nós n_1 e n_2, cada um contendo um valor possível de ω^r
$filhos \leftarrow \{n_1, n_2\}$
senão
obter o produto P de matrizes de torção dos nós do caminho da raiz até *pai*
computar as matrizes de torção B^1 e B^2 para o próximo átomo da cadeia,
usando distâncias derivadas do nó ancestral ω^r
criar nós n_1 e n_2, um para cada produto PB^1 e PB^2
$filhos \leftarrow \{n_1, n_2\}$
podar qualquer nó em $filhos$ caso sua posição não satisfaça
alguma *distância tipo F*, segundo um valor de tolerância ϵ
fim se
para cada nó b de $filhos$ **faça**
conectar b com *pai*
BPModificado(b)
fim para
senão
imprimir a solução formada pelas posições de cada nó do caminho de *pai* até a raiz
fim se

Isso implica que uma realização seja a imagem especular da outra. Portanto, se uma delas não for um L-aminoácido, a outra o será e vice-versa, o que torna apenas uma delas interessante para nós. Podemos então calcular ambas e verificar qual delas tem a quiralidade que desejamos. A característica mais importante dessa abordagem é que não precisamos utilizar nenhuma informação dos resíduos anteriores. Como vantagem, não temos propagação de erros entre resíduos no cálculo da cadeia principal.

4.7 Experimentos Computacionais

Testamos o método proposto nesse capítulo de maneira análoga à seção 3.4, gerando instâncias do PMGDD modificado a partir das mesmas estruturas protéicas (mesma base de testes).

Para podermos trabalhar com estruturas de proteínas reais, tivemos que tratar resíduos do tipo prolina. Dessa vez, o fizemos diferentemente da seção 3.4 (que considerava H_N^r como a posição do próprio átomo N do resíduo), pois, para que os cálculos de distâncias dos pares (H_N^r, C^r), $\left(C^r, H_\alpha^{r+1}\right)$ e $\left(C^r, C^{r+1}\right)$ funcionem, é necessário que N e H_N^r sejam diferentes (vide

figura 4.2). Sendo assim, consideramos H_N^r como a posição do C_δ (carbono delta, o carbono do anel da prolina ligado ao nitrogênio, que não o C_α). Outras maneiras de se contornar essa situação podem ser procuradas e tentadas em trabalhos futuros (por exemplo, utilizar a cadeia do capítulo 3 especialmente para os resíduos prolina). As distâncias conhecidas *a priori* para geração das cadeias peptídicas a partir de soluções foram obtidas usando-se toda base de testes (valores médios de distâncias inter-atômicas), com a exceção dos ângulos $\triangle\omega$, cujos valores exatos originais foram utilizados. Isso foi necessário pois o PMGDD modificado requer valores exatos de distâncias e se usássemos valores médios de $\triangle\omega$, as distâncias dos pares (H_N^r, C^r), $\left(C^r, H_\alpha^{r+1}\right)$ e $\left(C^r, C^{r+1}\right)$ seriam inconsistentes com as demais. Esperamos que aplicações futuras do algoritmo possam trabalhar com faixas de distâncias e assim os valores originais não precisarão ser conhecidos.

Implementamos o algoritmo BP modificado para solucionar as instâncias, obtendo estruturas tridimensionais da cadeia artifical $H_N^r - H_\alpha^r - C^r$. Para cada solução, procedemos de acordo com a seção 4.6 para calcular a cadeia principal correspondente. As implementações foram feitas em linguagem Java e executadas em ambiente Intel Core 2 Duo 2.2 GHz, 2GB RAM, Sistema Operacional Linux.

Como anteriormente, medimos a similaridade entre a cadeia principal calculada e a cadeia principal da estrutura original através da equação 3.5. O algoritmo BP modificado foi executado sobre as instâncias com tolerância $\epsilon = 10^{-5}$, guardando-se sempre a solução de melhor LRMSD, E_{min}. A tabela 4.2 resume os resultados dos experimentos. O tempo de execução T está expresso em minutos.

A figura 4.5 compara o desempenho da cadeia do capítulo 3 (na figura como Cadeia Artificial A), com o da cadeia proposta (na figura como Cadeia Artificial B).

4.8 Discussão

De maneira semelhante ao método do capítulo 3, os testes feitos com a implementação do BP modificado resultaram em soluções de boa qualidade, geometricamente próximas das moléculas originais. Entretanto, as instâncias do PMGDD modificado que foram geradas (a partir das mesmas proteínas utilizadas nos experimentos do capítulo 3) tiveram uma menor quantidade de soluções. Consequentemente, a quantidade de instâncias cujo espaço de busca pôde ser completamente explorado no tempo de duas horas foi muito maior (apenas 7 instâncias não

Instância	Res	Sols	T	E_{min}
2K1I	32	16	0	0,1966
2WH9	36	128	0	0,3617
2KB8	37	4	0	0,1249
2VY5	37	4	0	0,1172
2KGH	39	4	0	0,1472
2K9D	44	4	0	0,0230
2JZM	53	4	0	0,1469
2KKR	59	16	0	0,1991
2WNM	59	4	0	0,1580
2KEC	68	256	0,05	0,1322
2KN9	81	262144	55,4	0,0345
2VXE*	88	278328	120	1,4089
2VXF	88	64	0,3333	0,2406
2K22	90	4	0	0,0222
2KC1	91	64	0	0,0314

Instância	Res	Sols	T	E_{min}
2KUO	91	1024	0,1	0,0564
2ROK*	100	2	120	2,2883
2ROP	142	128	63,75	0,1869
2K3A*	155	4130	120	12,8616
2WCY	155	1024	16,6666	0,2448
2KTU	164	4	0,5333	0,0801
2RR6	168	4096	0,4833	0,1552
2K7N	203	4	0,2666	0,0975
2K0L*	210	0	120	-
2KRG*	216	0	120	-
2RON	242	16	0,15	0,0229
2K6H	248	4	4,0333	0,1622
2KDO	252	16	4,5166	0,0552
2K4T*	285	0	120	-
2ROQ	343	256	0,1166	0,0214

Tabela 4.2: Resultados computacionais de comparação das soluções do BP modificado com a cadeia original.
* Instâncias cuja exploração do espaço de busca não foi concluída. O algoritmo foi automaticamente interrompido após 2 horas de execução.

foram completamente resolvidas, ficando 4 sem solução e 3 com solução não ótima).

É importante ainda ressaltar que somente foram utilizadas para podas as distâncias dos átomos da cadeia artificial $H^r_N - H^r_\alpha - C^r$. Poderíamos, por exemplo, em tempo de execução do BP modificado, gerar a estrutura protéica da solução e utilizar as distâncias relacionadas a átomos H_β para efetuar mais podas. Ou seja, o método proposto nesse capítulo utilizou menos informações de distância que o método do capítulo 3 e obteve resultados com melhor qualidade e mais rapidamente para a maioria das instâncias. No geral, a qualidade das soluções de ambos os métodos foi satisfatória. A figura 4.6 nos permite comparar visualmente a solução encontrada por ambos os métodos BP com a cadeia original de uma das instâncias.

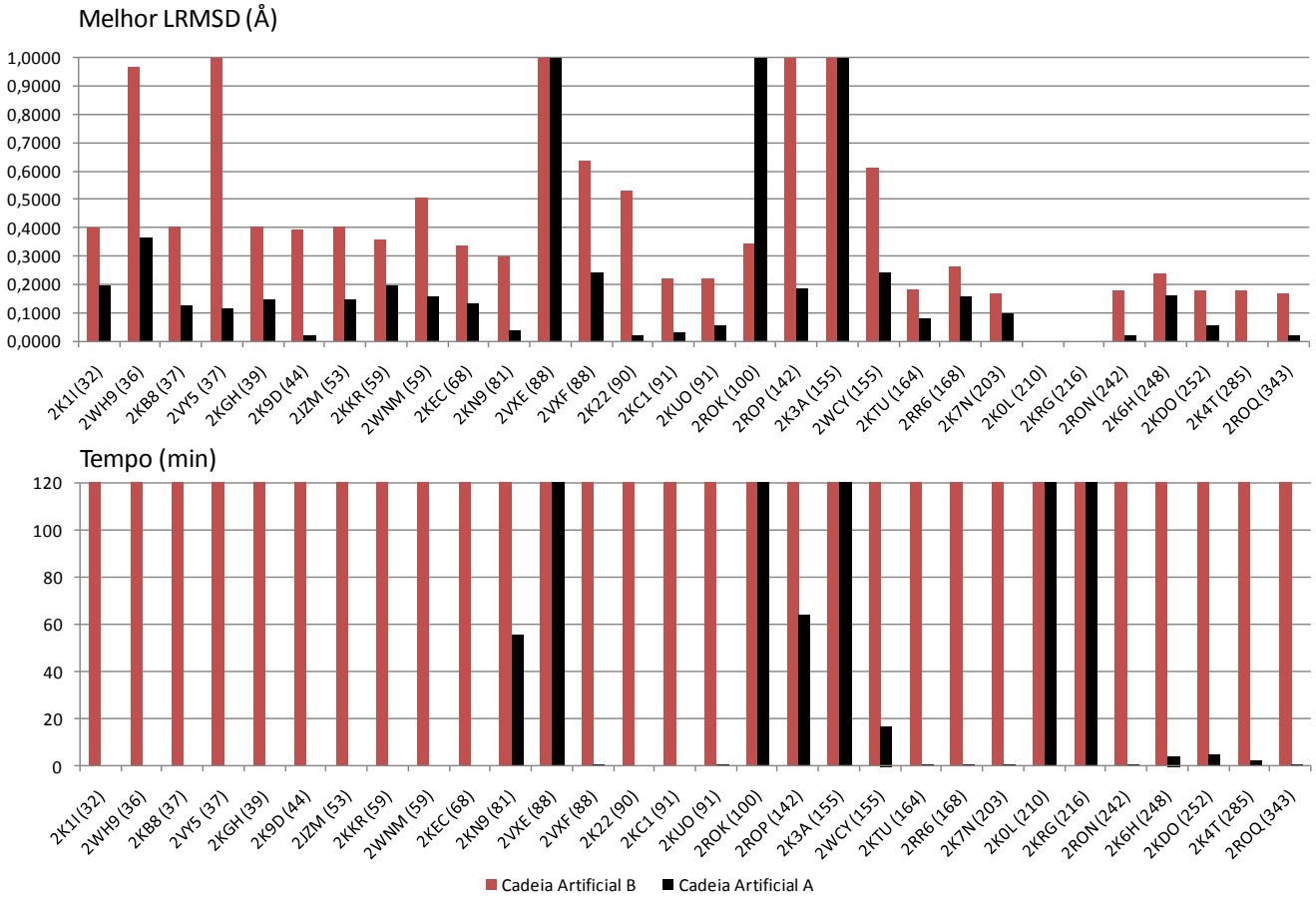


Figura 4.5: Resultados computacionais para os métodos estudados.

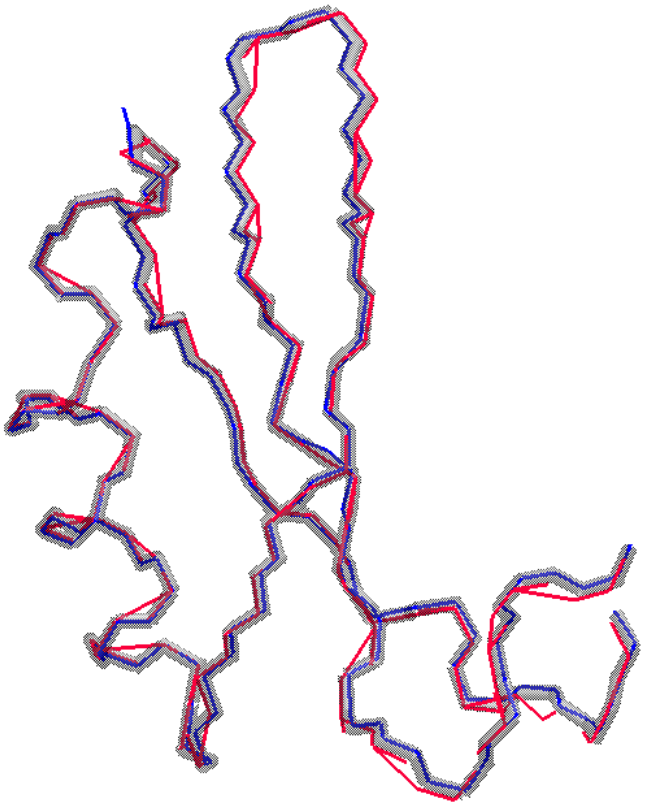


Figura 4.6: Estruturas encontradas pelos dois métodos comparadas à estrutura original.
A estrutura original da instância 2WNM (cinza), comparada com as estruturas obtidas pelo algoritmo BP (azul) e pelo algoritmo BP modificado (vermelho).

Capítulo 5

Conclusões e Trabalhos Futuros

Foram estudadas duas abordagens que usam o PMGDD e o BP envolvendo cadeias artificiais de átomos. Na primeira (capítulo 3), presente na literatura, estudamos com mais detalhes o processo de obtenção da cadeia principal a partir da artificial, apresentando um método alternativo que não requer resolução de sistemas lineares ou quadráticos, diferentemente do original. Além da sua facilidade de implementação, acreditamos que essa abordagem pode ter vantagens relacionadas à propagação de erros e assim merece ser investigada posteriormente. A segunda abordagem (capítulo 4) foi proposta neste trabalho, consistindo de uma nova cadeia artificial.

No aspecto teórico, podemos destacar como vantagens da cadeia proposta a não utilização de medidas relacionadas a átomos H_β (menos precisas). Além disso, a cadeia proposta acrescenta 4 níveis por resíduo à árvore de exploração do espaço de busca percorrida pelo BP Modificado, enquanto que a cadeia do capítulo 3 acrescenta 5 níveis por resíduo à árvore do BP. Finalmente, com a cadeia proposta temos mais facilidade de obtenção da cadeia principal (devido à fácil implementação e à não propagação de erros entre os resíduos).

No aspecto prático, a cadeia proposta teve desempenho inferior para a instância 2K4T, por não ter encontrado soluções no tempo de duas horas, enquanto que a outra cadeia encontrou boas soluções. De resto, a cadeia proposta teve desempenho superior: mesmo usando menos informações que a cadeia do capítulo 3 (já que poderíamos utilizar as distâncias relacionadas a átomos H_β como distâncias tipo F, para que houvesse mais podas), ela teve soluções com melhor qualidade (menor LRMSD) para 25 das 27 instâncias que ambos os algoritmos conseguiram solucionar (tendo desempenho inferior apenas para 2ROK e 2K3A). Uma diferença importante entre os resultados para os dois métodos é que a cadeia proposta teve uma menor quantidade de nós e soluções explorados, permitindo uma inspeção completa do espaço de busca para a maioria

das instâncias, o que não aconteceu com a outra cadeia.

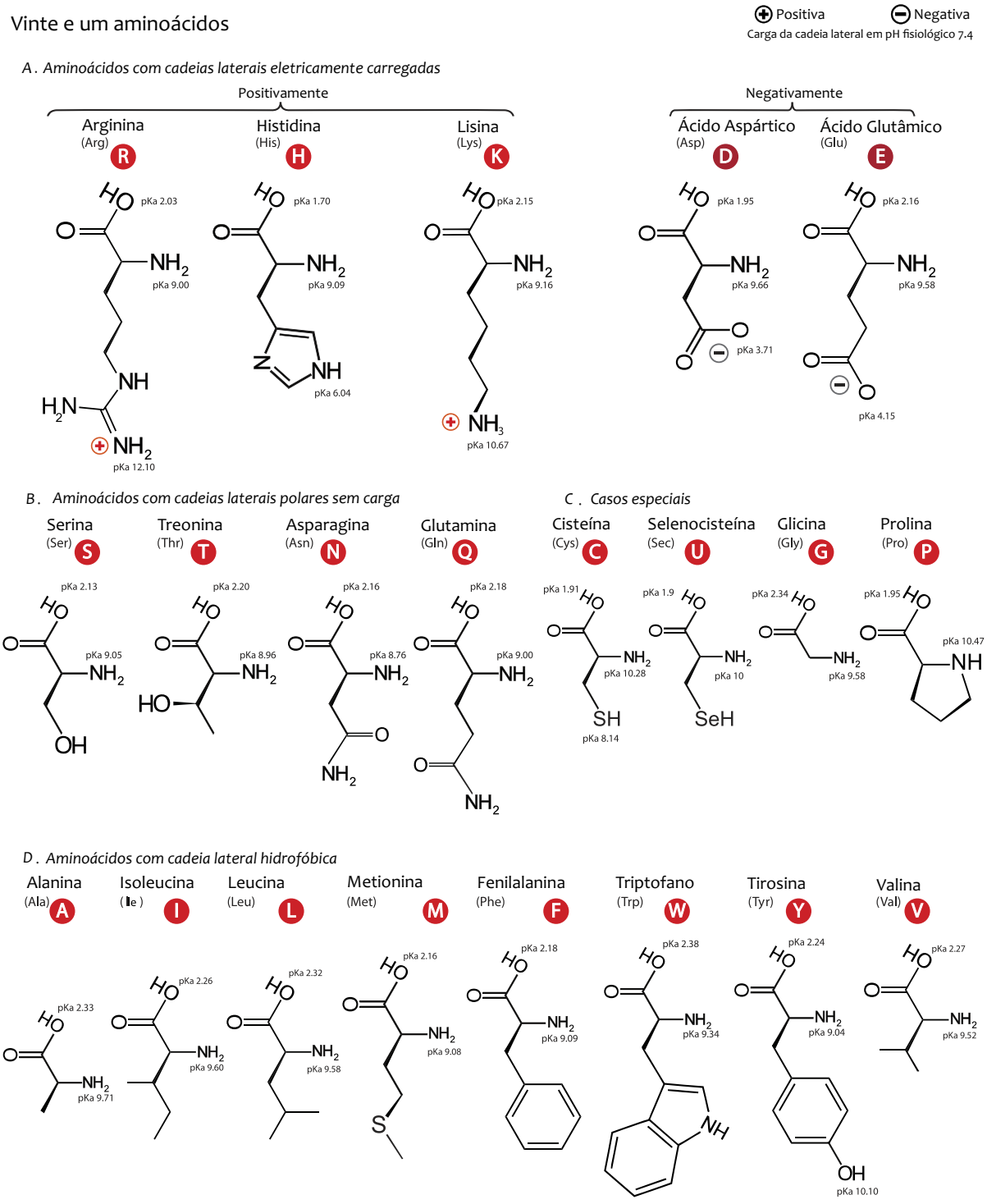
Recentemente, fizemos uma investigação teórica [5] sobre a influência de características das instâncias do PMGDD sobre a sua quantidade de soluções. Abordagens nesse sentido podem ser úteis para esclarecer por que as instâncias do capítulo 3 tiveram tantas soluções. Fizemos outro trabalho [4] mostrando que o algoritmo BP pode ser utilizado de maneira segmentada, em diferentes partes da molécula que posteriormente seriam unidas. Isso pode ser útil particularmente na determinação das cadeias laterais dos resíduos de uma proteína, ou mesmo na resolução de partes da molécula separadas por resíduos sem informações experimentais suficientes para que se gere uma instância contínua (seja por problemas técnicos em laboratório, seja pela ausência de átomos da cadeia artificial no resíduo, como é o caso da prolina para a sequência $H_N^r - H_\alpha^r - C^r$).

Ambos os métodos estudados nesse trabalho contribuem no sentido de aproximar o PMGDD dos experimentos de RMN, uma vez que a maioria dos valores de entrada considerados são plausíveis de se obter experimentalmente. Os próximos passos mais importantes nesse sentido estariam, sem dúvida, na reformulação do problema e do algoritmo para que trabalhem com faixas de valores de distância, já que os experimentos não fornecem valores exatos. Os métodos aqui estudados (e a maneira como foram estudados), dentre outras finalidades, serviram para esboçar como pode ser a complexidade do espaço de busca quando considerarmos faixas em torno dos valores utilizados.

Abre-se a possibilidade para trabalhos futuros que podem melhorar a aplicabilidade prática do BP, bem como aumentar o seu desempenho e a precisão sobre o cálculo de estruturas proteicas, a saber:

- Utilização de parâmetros biofísicos (por exemplo, restrições conhecidas decorrentes da estrutura secundária) como critério de poda durante a execução do BP;
- Melhor tratamento computacional de resíduos do tipo Prolina para a cadeia $H_N^r - H_\alpha^r - C^r$;
- Demonstração formal da relação (4.2);
- Estudo sobre as incertezas envolvidas no cálculo da equação (3.4);
- Adaptação do PMGDD e do BP para distâncias tipo E não exatas, com faixas obtidas através de experimentos.

Apêndice I - Tabela de Aminoácidos



pKa Data: CRC Handbook of Chemistry, v.2010

Traduzido e Adaptado de: Dan Cojocari, Department of Medical Biophysics, **University of Toronto**, 2010

Referências Bibliográficas

[1] Liberti, L., Lavor, C., Maculan, N. A branch-and-prune algorithm for the molecular distance geometry problem. International Transactions in Operational Research 15, 1–17, 2008.

[2] Gordon S. Rule, T. Kevin Hitchens. Fundamentals of Protein NMR Spectroscopy, Springer, 2006.

[3] Nucci, P. Lavor, C. Nogueira, L. Heurísticas para o Problema Molecular de Geometria de Distâncias Aplicado a Proteínas. Monografia de conclusão do curso de Bacharelado em Ciência da Computação. Universidade Federal Fluminense. 2008.

[4] Nucci, P., Nogueira, L., Lavor, C. Múltiplas Árvores de Realização no Problema de Geometria de Distâncias Aplicado a Moléculas. In: Simpósio Brasileiro de Pesquisa Operacional, 2009, Porto Seguro. 41o. Simpósio Brasileiro de Pesquisa Operacional, 2009.

[5] Nucci, P., Nogueira, L., Lavor, C. Influência de Distâncias Inter-atômicas no Algoritmo Branch-and-prune Aplicado a Moléculas. In: Simpósio Brasileiro de Pesquisa Operacional, 2009, Porto Seguro. 41o. Simpósio Brasileiro de Pesquisa Operacional, 2009.

[6] Berg, J., Tymoczko, J., Stryer, L.. Biochemistry. 5th ed. New York, Freeman and Comp, 2002.

[7] Lavor, C., Liberti, L., Maculan, N. The discretizable molecular distance geometry problem, arXiv:q-bio/0608012v1, 2006.

[8] Cruz, I.F. e Twarog, J.P. 3D Graph Drawing with Simulated Annealing, in Brandenburg, F.-J. (ed.), Graph Drawing, Symposium on Graph Drawing, GD '95, Passau, Germany, LNCS 1027:162-165, Springer, Berlin, 1996.

[9] Hendrickson, B.A. The molecule problem: exploiting structure in global optimization, SIAM Journal on Optimization 5 (1995), 835-857.

[10] Wu, Z., The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation, SIAM Journal on Optimization 6 (1996), 748-768.

[11] Eren, T., Goldenberg, D.K., Whiteley, W., Yang, Y.R., Morse, A.S., Anderson, B.D.O. e Belhumeur, P.N. Rigidity, Computation and Randomization in Network Localization, in IEEE InfoCom 2004 Proceedings, 2673-2684, 2004.

[12] Pogorelov, A. Geometry, Mir Publishers, Moscow, 1987.

[13] Phillips, A.T., Rosen, J.B., e Walke, V.H. Molecular structure determination by convex underestimation of local energy minima, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 23:181-198, American Mathematical Society, Providence, 1996.

[14] (Artigo aceito). C. Lavor, A. Mucherino, L. Liberti, N. Maculan. On the Computation of Protein Backbones by using Artificial Backbones of Hydrogens. Journal of Global Optimization. 2010.

[15] Dong, Q., e Wu, Z., A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data, Journal of Global Optimization 26 (2003) 321–333.

[16] Q. Dong, Z. Wu. A Linear-Time Algorithm for Solving the Molecular Distance Geometry Problem with Exact Inter-Atomic Distances, Journal of Global Optimization 22, 365–375, 2002.

[17] D. Wu e Z. Wu. An Updated Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse Distance Data, Journal of Global Optimization 37, 661–673, 2007.

[18] B. Horn. Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America. A, vol. 4, no. 4, pp. 629-642, 1987.

[19] (Artigo submetido) Nucci, P., Nogueira, L., Lavor, C., Liberti, L., Maculan, N. Comment on: "A branch-and-prune algorithm for the molecular distance geometry problem". International Transactions in Operational Research. 2010.