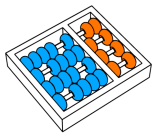
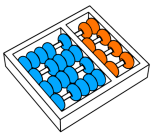


Ivan da Silva Sendin

"Tratamento de Incertezas no Cálculo de Estruturas de Proteínas"

CAMPINAS
2012



Universidade Estadual de Campinas
Instituto de Computação

Ivan da Silva Sendin

"Tratamento de Incertezas no Cálculo de Estruturas de Proteínas"

Orientador(a): **Prof. Dr. Siome Klein Goldenstein**

Co-Orientador(a): **Prof. Dr. Carlile Lavor**
IMECC - Unicamp

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Doutor em Ciência da Computação.

Este exemplar corresponde à versão final da Tese defendida por Ivan da Silva Sendin, sob orientação de Prof. Dr. Siome Klein Goldenstein.

Assinatura do Orientador(a)

CAMPINAS
2012

FICHA CATALOGRÁFICA ELABORADA POR
MARIA FABIANA BEZERRA MULLER - CRB8/6162
BIBLIOTECA DO INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E
COMPUTAÇÃO CIENTÍFICA - UNICAMP

Sendin, Ivan da Silva, 1975-
Se55t Tratamento de incertezas no cálculo de estruturas de proteínas / Ivan da Silva Sendin. – Campinas, SP : [s.n.], 2012.

Orientador: Siome Klein Goldenstein.
Coorientador: Carlile Campos Lavor.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Incerteza (Teoria da informação). 2. Proteínas - Estrutura. I. Goldenstein, Siome Klein,1972-. II. Lavor, Carlile Campos,1968-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em inglês: Uncertainty propagation in protein structure determination
Palavras-chave em inglês:
Measure of uncertainty (Information theory)
Proteins - Structure
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Siome Klein Goldenstein [Orientador]
Guilherme Pimentel Telles
Aurelio Ribeiro Leite de Oliveira
Nelson Maculan Filho
Luiz Satoru Ochi
Data de defesa: 10-12-2012
Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

Tese Defendida e Aprovada em 10 de dezembro de 2012, pela Banca examinadora composta pelos Professores Doutores:

Prof. Dr. Guilherme Pimentel Telles
IC / UNICAMP

Prof. Dr. Aurelio Ribeiro Leite de Oliveira
IMECC / UNICAMP

Prof. Dr. Nelson Maculan Filho
COPPE / UFRJ

Prof. Dr. Luiz Satoru Ochi
IC / UFF

Prof. Dr. Siome Klein Goldenstein
IC / UNICAMP

Instituto de Computação
Universidade Estadual de Campinas

Tratamento de Incertezas no Cálculo de Estruturas de Proteínas

Ivan da Silva Sendin[1]

10 de Dezembro de 2012

Banca Examinadora:

- Prof. Dr. Siome Klein Goldenstein (Orientador)
- Prof. Dr. Guilherme Pimentel Telles
  IC - Unicamp
- Prof. Dr. Aurelio Ribeiro Leite de Oliveira
  IMECC - Unicamp
- Prof. Dr. Nelson Maculan Filho
  COPPE - UFRJ
- Prof. Dr. Luiz Satoru Ochi
  IC - UFF

[1]**Parcialmente apoiado pela CAPES, Bolsa Número 1079591**

Resumo

A determinação da estrutura de uma proteína usando dados de Ressonância Magnética Nuclear precisa lidar com incertezas provenientes do experimento laboratorial. Neste trabalho, apresentamos um método híbrido utilizando aritmética afim e propagação de incerteza por partículas para o tratamento e o controle de incertezas. Aplicado no cálculo da estrutura de proteínas, o método proposto é capaz de gerar estruturas de proteínas que atendem satisfatoriamente as restrições do problema.

Abstract

The protein structure determination using Nuclear Magnetic Resonance data uses imprecise information from laboratorial experiments. In this work we introduce a new hybrid method that combines affine arithmetic and particles to uncertainty propagation and control. Applied to protein structure determination this new method was able to determine protein structures that satisfy most of problem constraints.

Agradecimentos

Aos orientadores, Siome e Carlile, por me guiarem neste caminho, pelas inúmeras e preciosas orientações.

Aos professores e funcionários do Instituto de Computação.

Aos colegas de Catalão.

A minha família, pelo suporte e pela paciência infinita e por darem um sentido ao termo incerteza...

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A Ciência nunca "gostou" de incertezas... O sonho de Laplace manifestava-se em um mundo determinístico, onde com o conhecimento dos estados do sistema, talvez do universo, um algoritmo poderia prever o futuro por completo.

O início do Século 20 foi marcado por diversas descobertas científicas que revolucionaram a Ciência: o princípio da incerteza e a física quântica descrevem um mundo onde a certeza não existe.

O teorema da incompletude de Gödel também introduz incerteza na Matemática. Na Computação, a teoria da NP-completude mostra que existem problemas que dificilmente serão resolvidos de forma exata por computadores e a Teoria de Computabilidade de Turing mostrou que existem problemas que não podem ser resolvidos por computadores.

Neste trabalho, apresentamos um método novo de propagação de incertezas que combina duas formas já existentes e conhecidas pela comunidade científica: a propagação usando aritmética afim e por partículas. O método proposto é testado em um problema prático conhecido: a determinação de estrutura de proteínas usando informações de experimentos de Ressonância Nuclear Magnética. Até onde foi possível investigar, não existe uma abordagem semelhante a nossa: um método construtivo que usa explicitamente a incerteza.

No Capítulo 2, são apresentados os métodos de propagação de incertezas que serão usados como base para o nosso método. Neste capítulo, expomos as suas limitações quando usados separadamente.

No Capítulo 3, descrevemos o método híbrido desenvolvido e, no Capítulo 4, o problema abordado: a determinação de estruturas de proteínas usando dados incertos.

Os resultados obtidos são expostos e analisados no Capítulo 5. O trabalho é finalizado no Capítulo 6, quando apresentamos as conclusões obtidas e trabalhos futuros.

Capítulo 2

Incertezas

Modelagem, representação e propagação de incertezas são problemas abordados pela Teoria da Informação e pela Modelagem Estocástica. No Capítulo 1 de [31], o autor apresenta a definições de vários tipos de incertezas. Neste trabalho, o termo incerteza está ligado à deficiência de informação. Dizemos que uma grandeza é incerta se o seu valor é parcialmente conhecido e podemos limitar esta grandeza a um intervalo, isto é, o valor exato $\dot{x}$ é desconhecido, mas conhecemos min e max, tal que:

$$min \leq \dot{x} \leq max,$$

ou ainda

$$\mathcal{P}(min \leq \dot{x} \leq max) \geq k,$$

para um valor conhecido k, onde $\mathcal{P}(x)$ é a probabilidade do evento x ocorrer.

Praticamente qualquer informação do mundo real está sujeita a erros. Este erro pode ter várias origens:

Medições Imprecisas Seja por restrições tecnológicas, financeiras ou operacionais, as medições feitas no mundo real estão sujeitas a imprecisões. Mesmo dados discretos estão sujeitos a incertezas [21].

Modelo Aproximado O modelo matemático utilizado para representar o mundo real do sistema pode estar incompleto ou não representar fielmente a realidade [21, 22, 17].

Computação Aproximada As grandezas representadas em um computador precisam ser armazenadas em espaços finitos de memória. O sistema de ponto flutuante usado nos computadores pode adicionar erros nos cálculos [20, 29, 35, 16].

2.1 Métodos de representação e propagação de incertezas

Os modelos tradicionais de representação e propagação de incertezas podem ser paramétricos, onde a incerteza é representada por parâmetros, ou não paramétricos, onde a incerteza é representada por amostras.

2.1.1 Estatística Paramétrica

A representação paramétrica é uma forma natural de representação de incerteza [47]. Utilizando poucos parâmetros e distribuições estatísticas, muitas formas de incerteza podem ser bem representadas. Por exemplo, inúmeros fenômenos naturais podem ser descritos por uma distribuição gaussiana: $x \sim (\mu, \sigma)$, com μ sendo a média e σ o desvio padrão da distribuição. Utilizando estes dois parâmetros, podemos gerar amostras que representam adequadamente a incerteza.

Um fator limitante para o uso desta representação é a dificuldade da propagação de incerteza através de operações matemáticas, pois as distribuições, em geral, não são fechadas para operações matemáticas. Na Figura 2.1, observamos que a distribuição de x^2, para $x \sim (0, 1)$, não pode ser representada adequadamente por uma distribuição gaussiana.

2.1.2 Representação por Faixa

Na representação por faixa, um valor parcialmente conhecido é representado por uma faixa que contenha este valor [27]. Uma operação genérica binária $\otimes_F$ em valores incertos representados por faixas é definida por:

$$\tilde{a} \otimes_F \tilde{b} := \{a \otimes b | a \in \tilde{a}, b \in \tilde{b}\}, \tag{2.1}$$

onde $\tilde{a}$ e $\tilde{b}$ são faixas e $\otimes$ é uma operação definida entre a e b.

Esta definição garante que os valores corretos contidos nas faixas sejam propagados durante a cadeia de cálculos, pois se

$$\dot{a} \in \tilde{a} \text{ e } \dot{b} \in \tilde{b},$$

então

$$\dot{a} \otimes \dot{b} \in (\tilde{a} \otimes_F \tilde{b}).$$

Esta garantia de propagação do valor correto faz com que os intervalos cresçam demasiadamente durante uma cadeia de cálculos, pois a faixa representa valores improváveis.

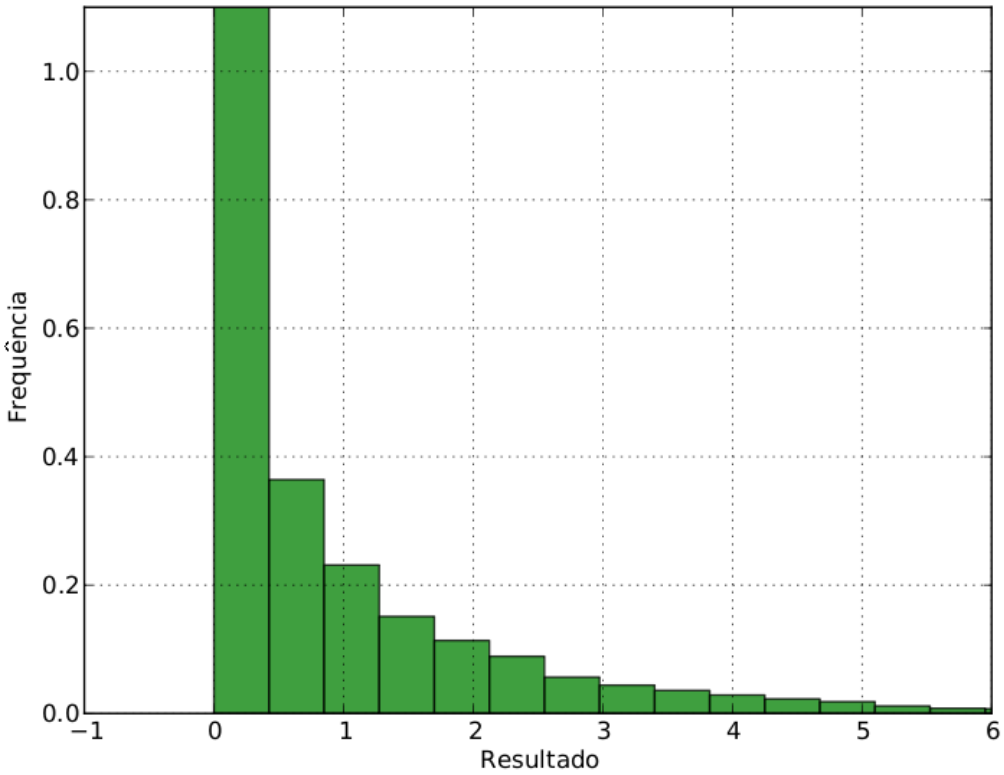


Figura 2.1: Distribuição de x^2, com $x \sim (0, 1)$.

Por exemplo, uma soma de n faixas $\tilde{v}_i$, cada uma limitada inferiormente por $v_i.min$ e superiormente por $v_i.max$, resulta em uma faixa com limite inferior

$$\sum_{i=1}^{n} v_i.min$$

e limite superior

$$\sum_{i=1}^{n} v_i.max.$$

Pelo Teorema Central do Limite [47], sabemos que para faixas independentes e n suficientemente grande, temos uma distribuição que se aproxima de uma distribuição gaussiana, e portanto, os valores extremos são menos prováveis que os próximos da média.

Qualquer faixa que contenha o valor correto $\dot{x}$ estará correta, mesmo que a faixa seja arbitrariamente grande, como $[-\infty; \infty]$. Ou seja, uma faixa correta pode não

ter informação útil, por exemplo, dizer que a altura de uma pessoa esta no intervalo $[-0,5; 6,0]$ metros.

A seguir apresentamos duas formas de propagação de incetezas usando faixas: **Aritmética Intevalar** e **Aritmética Afim**

Aritmética Intervalar

Uma forma de se implementar a representação por faixas é a utilização de intervalos, desenvolvida por Moore [36]. A Aritmética Intervalar (**AI**) define um conjunto de operações aritméticas sobre valores intervalares. Um valor parcialmente conhecido $\tilde{x}$ limitado por min e max é representado por um intervalo $\bar{x} = [min; max]$.

Na Aritmética Intervalar (**AI**) um valor parcialmente conhecido x limitado por min e max é representado por valor intervalar $\bar{x} = [min; max]$. A AI define um conjunto de operações aritméticas sobre valores intervalares que respeitam a restrição definida em 2.1.

Exemplo 1 *O valor de π pode ser representado pelo intervalo* $[3,14; 3,15]$ *, trabalhando com duas casas decimais de precisão. Para trabalhar apenas com inteiros, teríamos π representado por* $[3; 4]$ *, implicando que a área de um círculo de raio 10 seria de* $[300; 400]$ *.*

Cada operação aritmética precisa ser redefinida para ser aplicada sobre intervalos. Por exemplo a adição de intervalos $\bar{a} = [a.min; a.max]$ e $\bar{b} = [b.min; b.max]$ é definida da seguinte forma:

$$\bar{a} + \bar{b} := [a.min + b.min; a.max + b.max], \tag{2.2}$$

e a subtração por:

$$\bar{a} - \bar{b} := [a.min - b.max; a.max - b.min]. \tag{2.3}$$

A soma com um valor real k é definida por:

$$k + \bar{a} := [a.min + k; a.max + k], \tag{2.4}$$

e a subtração:

$$k - \bar{a} := [k - a.max; k - a.min], \tag{2.5}$$

que provoca um deslocamento do intervalo sem alterar o seu diâmetro.

A faixa de um intervalo é dada por:

$$faixa(\bar{a}) := a.max - a.min. \tag{2.6}$$

Como a AI não mantém informação sobre a origem das incertezas ou correlação entre elas, cancelamentos aritméticos não podem ser executados e a incerteza pode aumentar desnecessariamente durante uma cadeia de cálculos. A correlação entre intervalos também não existe. Por exemplo, o gráfico de uma incerteza em $\mathcal{R}^2$ é retangular (ver Figura 2.2).

Exemplo 2 *Suponha que* $\bar{x} = [1; 9]$ *seja o resultado de uma medida experimental. Agora fazemos:*

$$\begin{aligned} \bar{x} + (1 - \bar{x}) &= && (2.7)\\ [1; 9] + [-8; 0] &= && (2.8)\\ [-7; 9] &\quad . && (2.9) \end{aligned}$$

Se substituirmos as operações aritméticas intervalares pelo cancelamento algébrico, obteremos 1, valor também correto e mais preciso.

Aritmética Afim

Na Aritmética Afim (**AA**) [10, 9], um valor parcialmente conhecido $\tilde{x}$ é representado por uma **forma afim**, composta por um valor central e um somatório de parcelas com os valores desconhecidos:

$$\hat{x} = x_0 + x_1\epsilon_1 + x_2\epsilon_2 + \cdots + x_n\epsilon_n, \tag{2.10}$$

com $x_i \in \mathcal{R}$ representando a porção conhecida e $\epsilon_i \in [-1;\ 1]$ a porção **desconhecida**. A **faixa** de uma forma afim $\hat{x}$ é obtida usando:

$$faixa(\hat{x}) := \sum_{i=1}^{n} |x_i| \tag{2.11}$$

e os valores mínimo e máximo são dados por

$$\begin{aligned} min(\hat{x}) &:= x_0 - faixa(\hat{x}) && (2.12)\\ max(\hat{x}) &:= x_0 + faixa(\hat{x}). && (2.13) \end{aligned}$$

Em algumas situações, é preferível trabalhar com o diâmetro da forma afim:

$$diametro(\hat{x}) := 2\sum_{i=1}^{n} |x_i|. \tag{2.14}$$

Uma forma afim pode gerar um intervalo:

$$intervalo(\hat{x}) := [min(\hat{x}); max(\hat{x})]. \tag{2.15}$$

A faixa é uma forma de medida de incerteza da forma afim. Quando trabalhamos com mais de uma forma afim, outra maneira de medir a incerteza é considerar o volume produzido. Geometricamente, uma forma afim sempre forma um zonotopo, com uma simetria em relação ao centro. Na Figura 2.2, vemos a incerteza combinada das formas afins:

$$\begin{cases} \hat{x} = 5 + 2\epsilon_1 + 1\epsilon_2 \\ \hat{y} = 5 - 1\epsilon_1 + 1\epsilon_3. \end{cases}$$

Considerando o problema intervalar, temos uma área de 24 unidades, onde a área afim tem apenas 18 unidades.

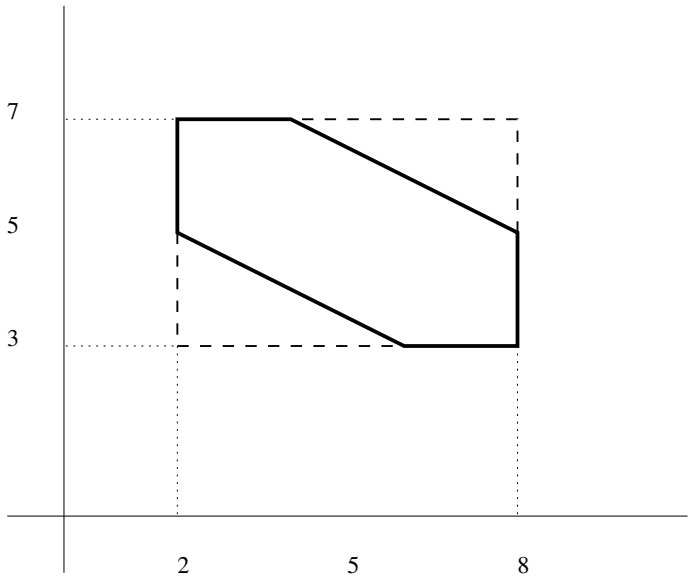


Figura 2.2: Zonotopo formado por formas afins e o retângulo formado pelos intervalos.

Na aritmética afim, cada valor desconhecido é identificado pelos seus respectivos ϵ_i's. Nas operações de soma, subtração, multiplicação e divisão por números reais, e soma ou subtração por outra forma afim, usa-se a aritmética simples aplicada às formas afins.

Por exemplo, as operações de soma ou subtração entre formas afins com n desconhecidos, são definidas por:

$$\hat{x} \pm \hat{y} := x_0 \pm y_0 + (\sum_{i=1}^{n} (x_i \pm y_i)\epsilon_i),$$

e a multiplicação por um real k:

$$k\hat{x} := kx_0 + k\sum_{i=1}^{n} x_i\epsilon_i.$$

Exemplo 3 *Usando a mesma incerteza do Exemplo 2, criamos a forma afim* $\hat{x} = 5+4\epsilon_1$*, onde obtemos:*

$$\begin{aligned} \hat{x} + (1 - \hat{x}) &= && (2.16)\\ 5 + 4\epsilon_1 + (-4 - 4\epsilon_1) &= && (2.17)\\ 1 &. && (2.18) \end{aligned}$$

2.1.3 Funções Não-Afins

Como nem todas as operações nas formas afins são fechadas, aproximações precisam ser desenvolvidas.

Exemplo 4 *Considere a operação de multiplicação*

$$\begin{aligned} (2 + 2\epsilon_1 + 3\epsilon_2) \times (3 + 4\epsilon_2) &= && (2.19)\\ 6 + 8\epsilon_2 + 6\epsilon_1 + 8\epsilon_1\epsilon_2 + 9\epsilon_2 + 12\epsilon_2^2 && (2.20) \end{aligned}$$

com um termo quadrático e um produto entre desconhecidos que não podem ser representados adequadamente na forma afim. Para representar este resultado usando uma forma afim, é necessário utilizar uma aproximação dos termos não afins com a criação de um novo termo desconhecido. A expressão não-afim $8\epsilon_1\epsilon_2 + 12\epsilon_2^2$*, que tem como mínimo o valor -1,33, para* $(\epsilon_1; \epsilon_2) = (1; -0,33)$ *ou* $(\epsilon_1; \epsilon_2) = (-1; 0,33)$ *e máximo o valor 20, para* $(\epsilon_1; \epsilon_2) = (1; 1)$ *é subistituída pela forma afim* $20\epsilon_3$*. A nova forma afim* $6+6\epsilon_1+17\epsilon_2+20\epsilon_3$ *terá uma faixa maior que a expressão com termos não-afins e parte da informação de correlação é perdida.*

No caso geral, a multiplicação afim é definida da seguinte forma:

$$\hat{x}.\hat{y} := x_0y_0 + \sum_{i=1}^{n}(x_0y_i + y_0x_i)\epsilon_i + faixa(\hat{x})faixa(\hat{y})\epsilon_{n+1}, \quad (2.21)$$

onde ϵ_{n+1} é um novo desconhecido.

No Algoritmo 1, é mostrado como calcular os valores de α, ζ e δ que são usados para calcular o inverso de uma forma afim. Com estes valores, o inverso de $\hat{x}$ é dado por:

$$\zeta + \alpha\hat{x} + \delta\epsilon_{n+1}. \tag{2.22}$$

Uma vez que sabemos calcular o inverso de $\hat{x}$, a divisão pode ser feita da seguinte forma:

$$\frac{\hat{y}}{\hat{x}} := \frac{1}{\hat{x}}\hat{y}. \tag{2.23}$$

Portanto, a operação de divisão é calculada usando duas operações não-afins.

Em [9], são apresentadas funções de aproximações para operações $\hat{x}^2$, $\sqrt{\hat{x}}$, $|\hat{x}|$, seno($\hat{x}$) e cosseno($\hat{x}$), e também um método geral de aproximação usando aproximação de Chebyshev.

Entrada: $\hat{x}$
Saída: α, ζ e δ
$mn \leftarrow min(\hat{x})$
$mx \leftarrow max(\hat{x})$
`/*` *a* `recebe o menor de` $|mn|$ `e` $|mx|$ `*/`
$a \leftarrow minimo(|mn|, |mx|)$
$b \leftarrow maximo(|mn|, |mx|)$
$\alpha \leftarrow -(1/b^2)$
$d_{max} \leftarrow 1/(a - \alpha * a)$
$d_{min} \leftarrow 1/(b - \alpha * b)$
$\zeta \leftarrow d_{min} + (d_{max} - d_{min})/2$
se $mn < 0$ **então**
  $\zeta \leftarrow -\zeta$
fim
$\delta \leftarrow (d_{max} - d_{min})/2$
retorna α, ζ, δ

Algoritmo 1: Calcula α, ζ e δ para o inverso de $\hat{x}$.

2.2 Evitando explosões

Um dos objetivos nas operações sobre formas afins é produzir resultados com faixas menores. Em geral, podemos dizer que a representação usando forma afim produz faixas menores que a representação intervalar [44]. Mesmo assim, a faixa cresce durante a computação e, apesar de correta, pode perder o seu significado ou utilidade.

2.2.1 Variáveis Temporárias

A criação de novos desconhecidos, para cada operação não-afim, cria situações indesejadas, tais como:

$$\hat{x}^2 - \hat{x}^2 \neq 0. \tag{2.24}$$

Isto ocorre porque cada uma das operações de potência cria um novo desconhecido, cada um com um identificador diferente. Operações não-afins que são repetidas devem ser efetuadas apenas uma vez e seus resultados armazenados em variáveis temporárias. Por exemplo:

$$\hat{t} = \hat{x}^2 \tag{2.25}$$

e

$$\hat{t} - \hat{t} = 0. \tag{2.26}$$

2.2.2 Ordem das Operações

Em geral, podemos dizer que:

$$\frac{(\hat{x}\hat{y})}{\hat{z}} \neq \frac{\hat{x}}{\hat{z}}\hat{y} \neq \frac{\hat{y}}{\hat{z}}\hat{x},$$

onde cada um dos resultados terá faixas diferentes. Quando possível, uma análise pré-implementação deve ser feita para produzir um resultado final mais preciso.

2.2.3 Cancelamentos Algébricos e Número de Operações

Como para cada operação não-afim é criada uma aproximação, é importante minimizar o número total de operações não-afins. Por exemplo, a expressão

$$(\hat{a} + \hat{b})^2$$

é calculada com duas operações não-afins e é preferível do que a expressão

$$\hat{a}^2 + 2\hat{a}\hat{b} + \hat{b}^2$$

com três operações não-afins.

2.3 Variações da Forma Afim

Em [33], são apresentadas variações sobre a forma afim tradicional introduzindo mais duas classes de desconhecidos: uma para os termos positivos ($\epsilon^+ \in [0;\ 1]$) e outra para os termos negativos ($\epsilon^- \in [-1;\ 0]$). Usando estas classes, podemos, por exemplo, modelar melhor os termos quadráticos produzidos pela multiplicação e consequentemente produzir faixas menores. Em [34], é apresentado um método mais preciso para a operação de divisão afim.

Nenhuma destas variações chega a apresentar uma melhoria significativa no controle do crescimento da faixa. Em geral, produzem aproximações menores. Porém, muitas vezes, a incerteza inerente do problema é suficiente para criar faixas grandes e fazer com que as formas afins não tenham informação relevante.

2.4 Implementações

Para realizar os testes deste trabalho, uma biblioteca de aritmética afim foi implementada em Python, de acordo com [9]. Um detalhe que não foi implementado foram os cuidados com os problemas de erros de arredondamento em aproximações por ponto-flutuante, que causam aumento da complexidade e a diminuição da velocidade. As garantias obtidas por estes cuidados não serão úteis, como veremos no próximo capítulo, usarmos as partículas para alterar as formas afins.

2.4.1 Coeficientes Nulos

Espera-se que existam muitos coeficientes iguais a zero nas formas afins. Portanto, uma biblioteca de matriz esparsa foi usada na construção das formas afins.

2.4.2 Agrupamento de Desconhecidos

Nos cálculos para determinação de uma forma afim, podem ser usadas várias operações não-afins, cada uma delas criando um novo desconhecido. Estes novos desconhecidos, vistos separadamente, não acrescentam informação à nova forma afim, pois não estão presentes nas outras formas afins já existentes no sistema. A manutenção destes torna os cálculos mais caros, o que torna interessante agrupar estes desconhecidos em um único termo. No Algoritmo 2 mostramos como podemos acumular os desconhecidos novos e, portanto, não compartilhados com outras formas afins.

Coeficientes pequenos, em relação à faixa total de incerteza representada por uma forma afim, também podem se tornar um custo computacional sem benefícios evidentes.

Entrada: $\hat{x}$: forma afim recém-criada; i: índice do último desconhecido utilizado antes de $\hat{x}$ ser criado; u: índice do último desconhecido
Saída: $\hat{x}$ e u alterados
$c \leftarrow 0$
para $i < k \leq u$ **faça**
  $c \leftarrow c + |\hat{x}_k|$
  $\hat{x}_k \leftarrow 0$
fim
$\hat{x}_{i+1} = c$
$u \leftarrow i + 1$

Algoritmo 2: Forma Afim Agrupada.

Assim, é interessante eliminar estes desconhecidos. No Algoritmo 3, é apresentado um método para agrupar todos os coeficientes menores que um determinado limite em um desconhecido escolhido. Em geral, é possível acumular as incertezas no último desconhecido, pois, vindo de uma aproximação, este desconhecido é exclusivo em uma forma afim recém criada.

Entrada: $\hat{x}$: forma afim; i: índice do desconhecido destino; u: índice do último desconhecido; l: limite do desconhecido
Saída: $\hat{x}$ alterada
$c \leftarrow 0$
para $1 \leq k \leq u$ **faça**
  se $|\hat{x}_k| < l$ **então**
    $c \leftarrow c + |\hat{x}_k|$
    $\hat{x}_k \leftarrow 0$
  fim
fim
$\hat{x}_i \leftarrow c$

Algoritmo 3: Eliminação de Coeficientes Pequenos.

2.5 Partículas

Partículas é um método não paramétrico de representação de incertezas [48]. Um valor incerto é representado por uma coleção de valores exatos chamados de partículas. Cada uma das partículas é um representante de sua "região". Como cada uma das partículas é uma instância simples do problema, as operações matemáticas são fáceis de serem implementadas, tão como uma função de pontuação específica ao problema. No Algoritmo

4, mostramos como construir uma operação $\otimes_P$ sobre partículas, dada uma operação $\otimes$ sobre tipos básicos das partículas.

2.5.1 Propagação da Incerteza

A propagação da incerteza leva em conta a importância das partículas. A dinâmica do processo consiste em atribuir importâncias diferentes para as partículas, de acordo com a probabilidade desta ser uma partícula viável. Esta atribuição é feita usando alguma função de pontuação que pode ser dependente do problema. Uma vez que cada partícula tem sua importância determinada, as partículas são amostradas. Assim, partículas com maior probabilidade de estarem corretas têm maior probabilidade de serem propagadas. Como cada partícula é um representante de sua região, também é usual criar partículas *próximas* às partículas amostradas e usá-las no processo de propagação.

Entrada: Conjuntos A e B de partículas; tamanho da saída n
Saída: $A \otimes_p B$
$r = \emptyset$
para $i = 0; i < n; i++$ **faça**
  $a \leftarrow$ amostra A
  $b \leftarrow$ amostra B
  $r \leftarrow \cup\{a \otimes b\}$
fim

Algoritmo 4: Operação $\otimes_p$ em partículas usando a operação $\otimes$.

A forma como as partículas são propagadas naturalmente aplica as propriedades estatísticas que em outras situações teriam que ser explicitamente identificadas e calculadas.

Um problema que ocorre nas partículas é a possibilidade de operações com partículas incompatíveis, já que nem todas as combinações de partículas são "corretas".

2.6 Comparações Entre os Métodos

Faremos comparações entre operações simples sobre formas afins e partículas para ilustrar os conceitos expostos e que serão relevantes na construção do método de propagação de incerteza. Para as operações sobre partículas serão feitas duas formas de amostragem:

- Sem correlação: cada operando é amostrado de forma independente.
- Com correlação: cada desconhecido da forma afim é amostrado e o valor é substituído nas formas afins.

2.6.1 Operação de Soma

O primeiro conjunto de experimentos será para a operação de soma de duas incertezas.

Sem correlação

A primeira comparação feita será com dois valores desconhecidos e independentes:

$$\begin{cases} \hat{a} = 1 + 1\epsilon_1 \\ \hat{b} = 1 + 1\epsilon_2 \end{cases}$$

O resultado afim da soma de $\hat{a}$ e $\hat{b}$ é $2 + 1\epsilon_1 + 1\epsilon_2$, e pode assumir como menor valor 0,0 (zero) e maior valor 4,0 (quatro). Amostrando 10.000 partículas, temos um resultado muito próximo: um valor mínimo de 0,024 e máximo de 3,98, com média de 2,005. Na Figura 2.3 é apresentada a distribuição dos resultados, e na Figura 2.4 é mostrada a porcentagem das partículas que estão dentro de um raio variável. Nesta figura, vemos que 75% das partículas estão limitadas pelo raio de um, correspondendo aproximadamente à faixa de um a três.

Com correlação

Neste caso, temos os mesmos limites para a forma afim, mas fazendo a amostragem dos desconhecidos, temos uma distribuição uniforme (Figura 2.5).

2.6.2 Operação de Subtração

Com correlação

Também iremos considerar:

$$\begin{cases} \hat{a} = 1 + 1\epsilon_1 \\ \hat{b} = 1 + 1\epsilon_1. \end{cases}$$

Na subtração de formas afins com correlação, ocorre um cancelamento da incerteza e o resultado afim é zero (o mesmo ocorre quando fazemos a amostragem dos desconhecidos). Usando partículas, para amostragens independentes, temos um mínimo de -1,98 e máximo de 1,99. Este é um exemplo claro de partículas incompatíveis sendo usadas. Aqui é fácil identificar o erro e usar apenas uma amostragem, mas em casos mais complexos isso não é trivial. A distribuição obtida é mostrada na Figura 2.6.

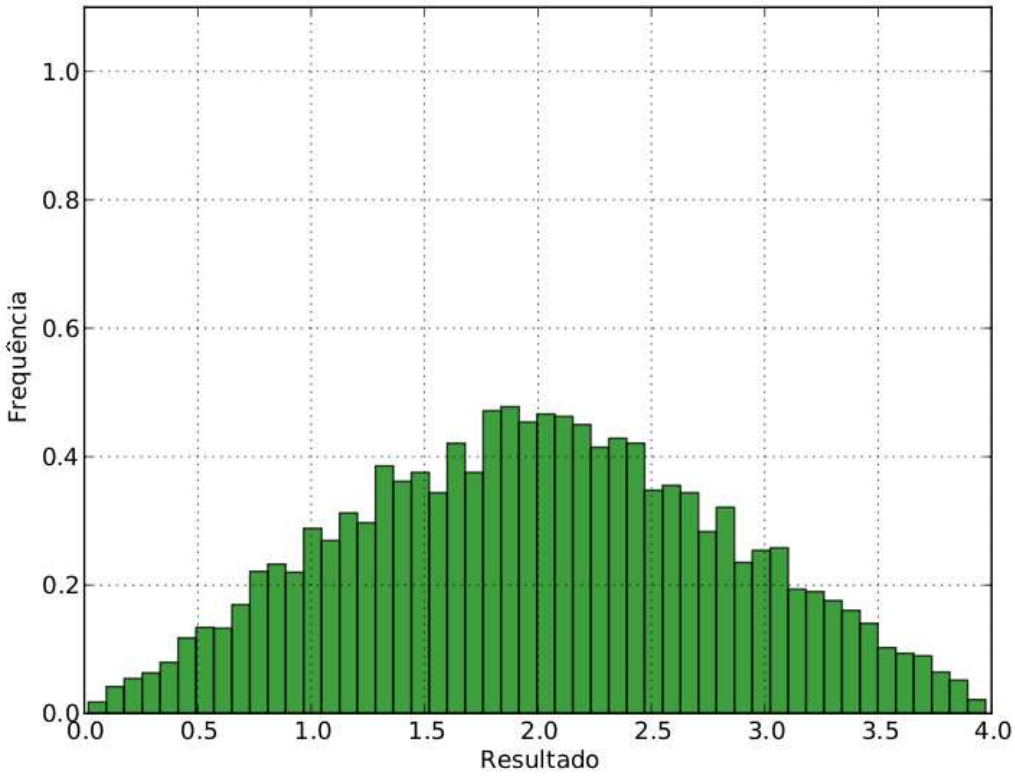


Figura 2.3: Distribuição dos resultados da soma com amostragens independentes.

2.6.3 Operação de Divisão

Neste exemplo, vamos comparar os resultados para a seguinte operação:

$$\frac{1+0,5\epsilon_1}{1+0,5\epsilon_1}$$

Analiticamente, concluímos que o valor correto da operação é 1. Na operação afim, obtemos o seguinte resultado:

$$1,333+0,444\epsilon_1+0,444\epsilon_2+0,333\epsilon_3,$$

com mínimo de 0,111 e máximo de 2,55.

Já com as partículas, com amostras independentes para cada um dos operandos, obtemos um valor médio de 1,09, com mínimo de 0,33 e máximo de 2,98, para uma amostra de 10.000 partículas. A distribuição dos resultados é mostrada na Figura 2.7. Na Figura 2.8,

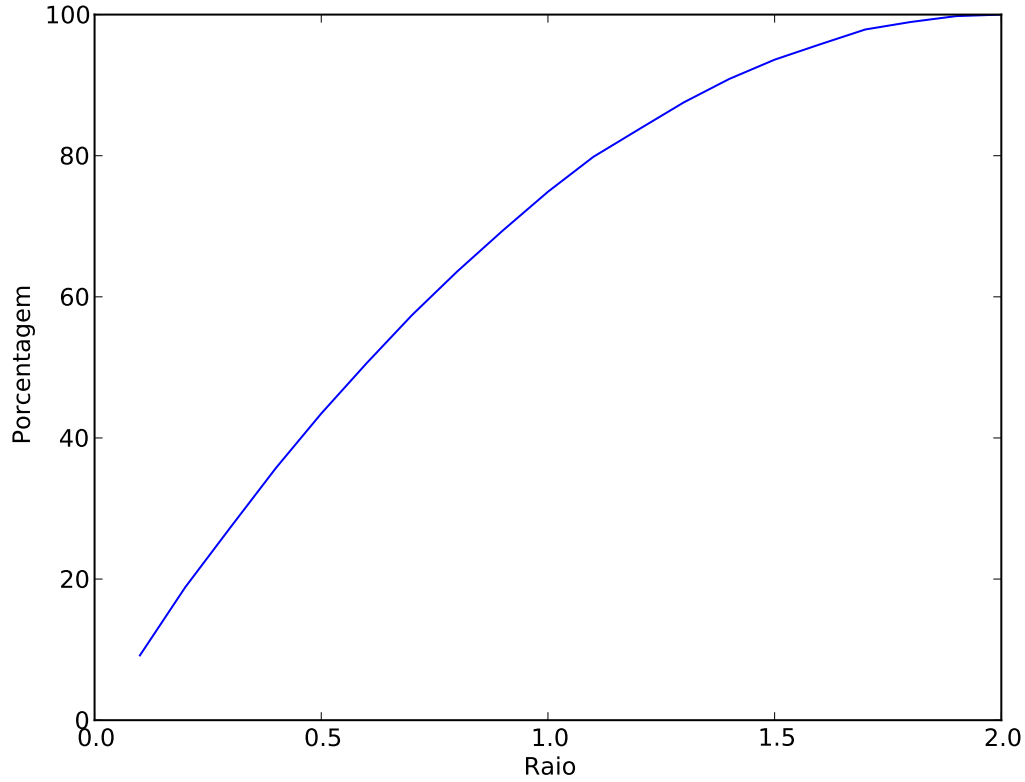


Figura 2.4: Evolução da porcentagem de partículas limitadas pelo centróide ± raio.

mostramos a porcentagem de partículas que estão dentro da faixa limitada pela média ± raio. Utilizando o mínimo das amostras e o raio de 1,0, vemos que, na faixa de 0,33 a 2,0, temos aproximadamente 95% da partículas Comparando com a forma afim, vemos que o limite inferior das partículas é melhor, mas a forma afim tem um limite superior melhor.

Fazendo uma amostragem para cada desconhecido do sistema, temos sempre o resultado correto: 1.

Agora, vamos repetir a operação com o coeficiente do desconhecido do divisor negativo:

$$\frac{1 + 0,5\epsilon_1}{1 - 0,5\epsilon_1}$$

O resultado afim obtido é $1,33 + 0,888\epsilon_1 + 0,444\epsilon_2 + 0,333\epsilon_3$, com mínimo e máximo de -0,33 e 3,0 respectivamente. As partículas tiveram resultados similares, tanto para a amostragem independente como na amostragem do desconhecido: 0,341 e 2,972, para as partículas, e 0,333 e 2,999, para a amostragem nos desconhecidos.

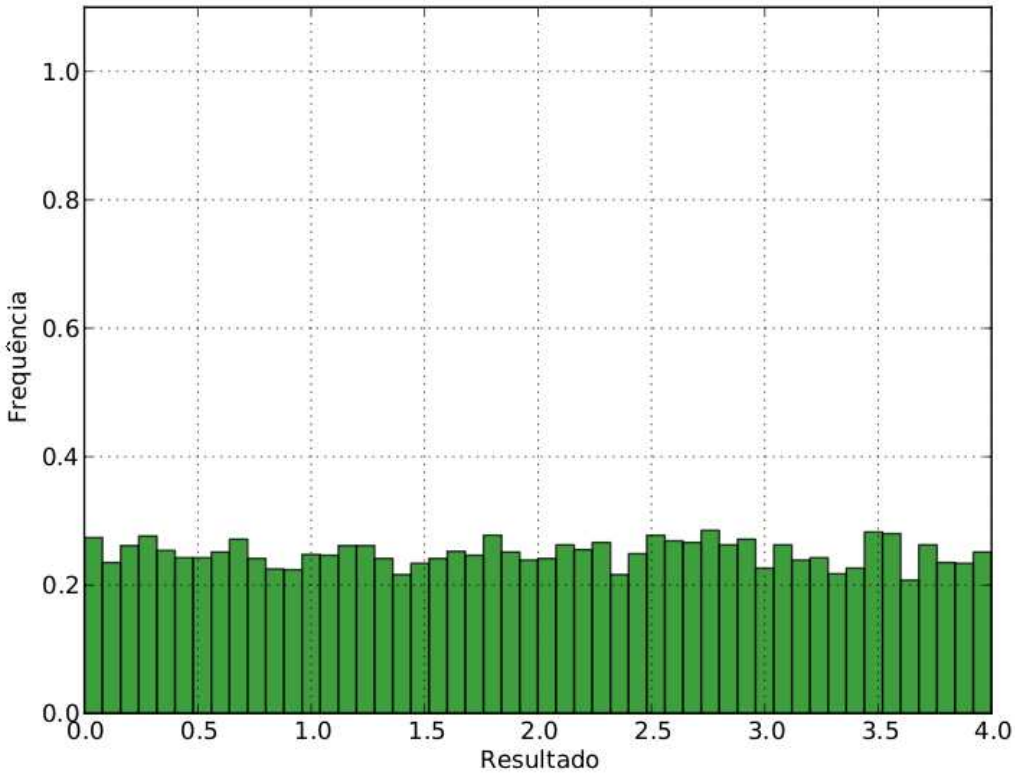


Figura 2.5: Distribuição dos resultados da soma com amostragem com alta correlação.

Por último, consideraremos o caso

$$\frac{1+0,5\epsilon_1}{1+0,5\epsilon_2}$$

O resultado da forma afim é $1,33+0,66\epsilon_1-0,22\epsilon_2+0,44\epsilon_3+0,33\epsilon_4$, com mínimo -0,33 e máximo 3,0, e com as partículas amostradas, temos 0,34 e 2,946, para o mínimo e máximo, respectivamente.

2.6.4 Resumo dos Resultados Obtidos

Na Tabela 2.1 apresentamos o resumo dos resultados obtidos. Em geral, as partículas apresentam resultados com faixas menores, onde ainda é possível escolher a porcentagem de partículas. Porém, a forma afim pode ser melhor em situações quando a correlação pode ser aproveitada. O sinal '+' indica correlação positiva, o sinal '-' indica correlação

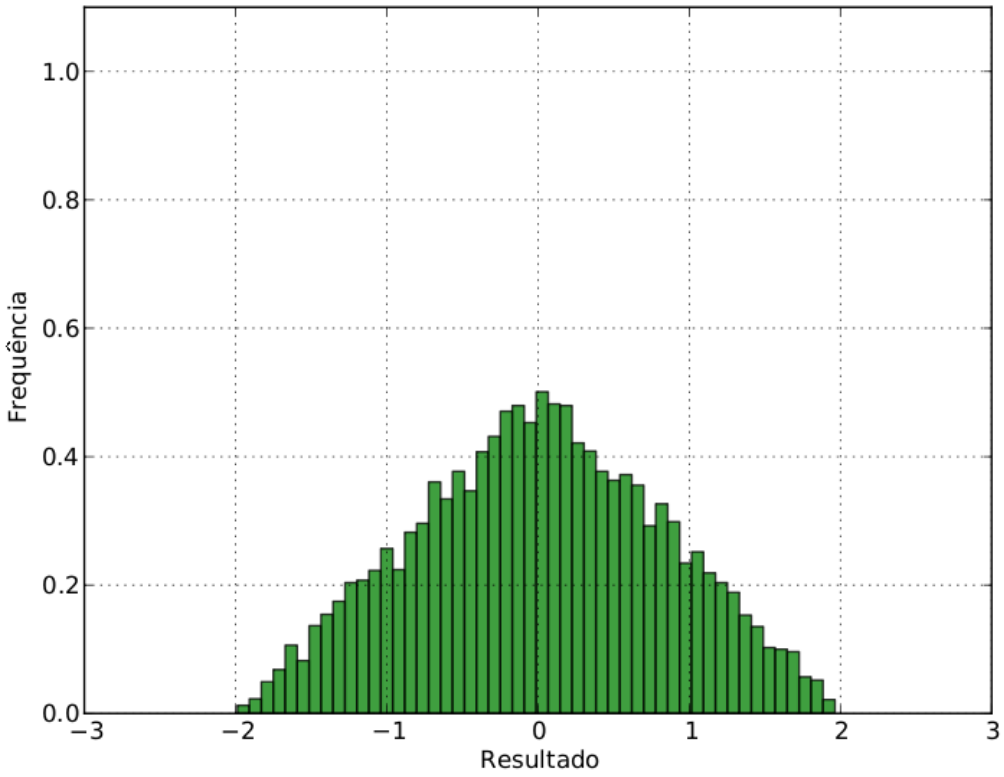


Figura 2.6: Distribuição dos resultados da subtração com amostragem independentes.

negativa e, sem o sinal, indica independência das incertezas. Os números entre parênteses nas colunas das partículas indicam o diâmetro que limita 75% das partículas.

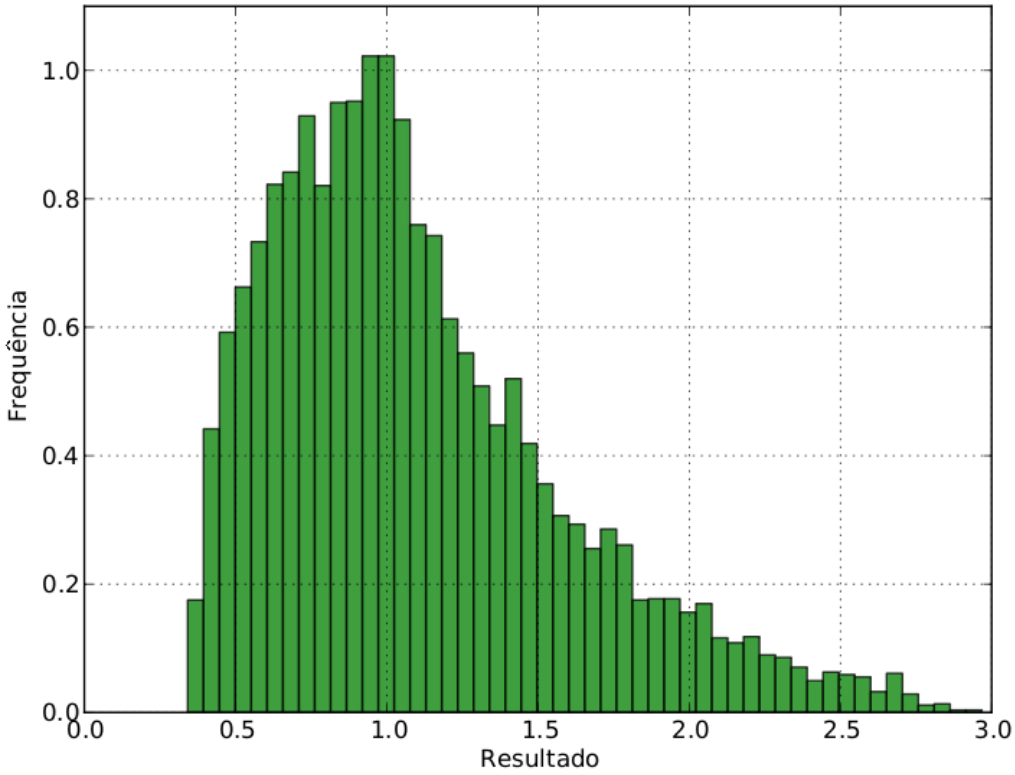


Figura 2.7: Distribuição dos resultados da divisão com amostragem do desconhecido.

Experimento	**Afim**	**Partículas Independentes**	**Partículas Desconhecidos**
Soma	4,0	3,75 (2,0)	3,74 (2,0)
Soma +	4,0	3,74 (3,0)	3,74 (3,0)
Subtração +	0,0	3,99 (2,0)	0
Divisão +	2,44	2,65 (1,1)	0
Divisão -	3,3	2,63 (1,1)	2,63(1,55)
Divisão	3,3	2,6 (1,1)	2,6(1,1)

Tabela 2.1: Resultados dos testes para operações simples sobre formas afins e partículas.

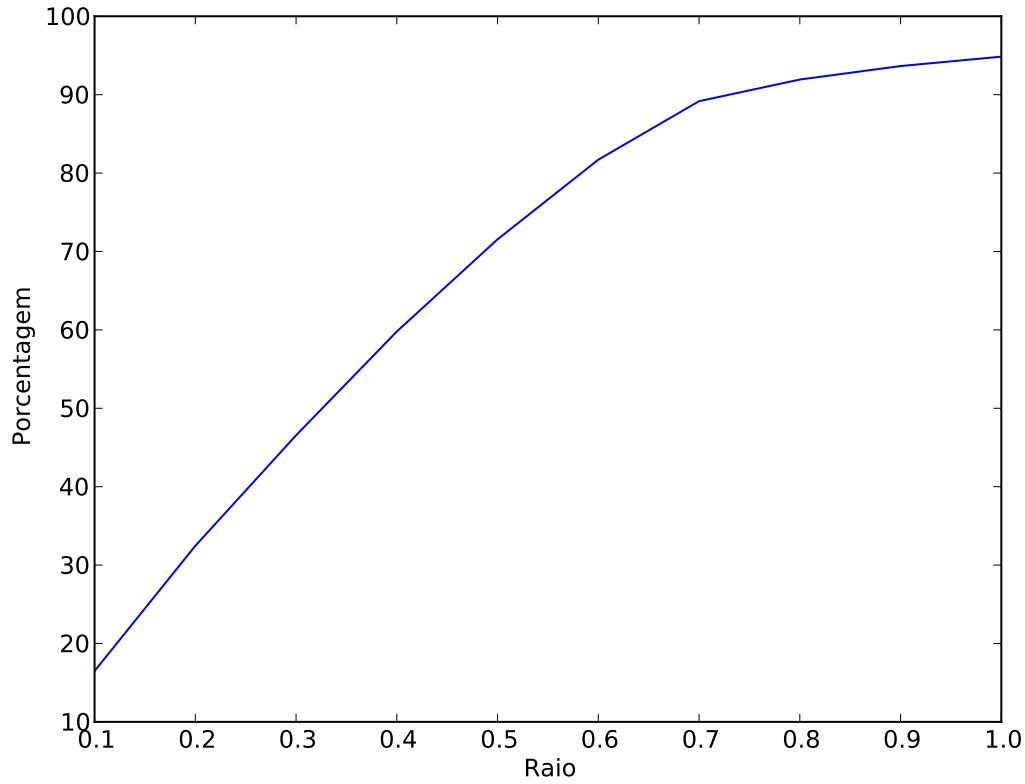


Figura 2.8: Evolução da porcentagem de partículas limitadas pelo centróide ± raio.

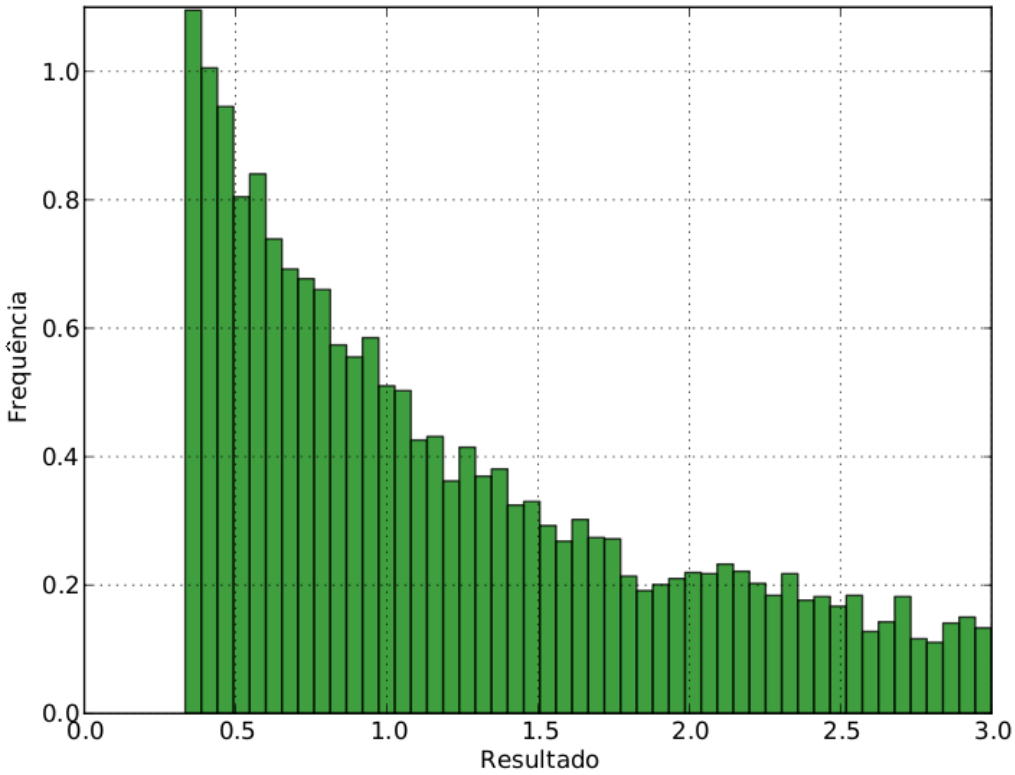


Figura 2.9: Distribuição dos resultados da divisão com correlação negativa.

Capítulo 3

Método Híbrido de Propagação e Controle de Incertezas

Neste capítulo, apresentamos um novo método de propagação e controle de incertezas. O método proposto combina a representação usando partículas e a representação afim, em uma tentativa de obter o melhor de cada uma dessas representações:

Correlação Explícita A forma afim, com o compartilhamento de desconhecidos, permite o cancelamento da incerteza e amostragens consistentes. Mesmo com intervalos grandes, a incerteza representada pela forma afim pode ser mais controlada que nas partículas por causa da correlação mantida com outros estados incertos do sistema.

Informação Estatística As partículas obedecem às leis estatísticas e fornecem informações de densidade, probabilidade e, em geral, limites mais estreitos que a forma afim. A natureza exata da representação por partículas permite o seu controle com o uso de métodos já existentes. Na seção 3.3 serão apresentadas como as partículas podem ser controladas.

3.1 Descrição do Método

Para cada estado incerto do sistema, são calculadas paralelamente duas configurações: uma usando aritmética afim e outra usando partículas. As partículas são selecionadas e usadas para controlar a forma afim, e a forma afim é usada para gerar partículas em futuras amostragens. O método é descrito na Figura 3.1 e no Algoritmo 5. O primeiro passo do método é criar uma solução afim para o novo estado usando os estados anteriores e dados incertos do problema. Em seguida a solução usando partículas é criada, as partículas são seçecionadas e usadas para controlar a forma afim. Veremos a seguir como as partículas podem ser selecionadas e, também, usadas para controlar a forma afim.

Entrada: Estados anteriores: $\hat{x}[1], \hat{x}[2], \ldots, \hat{x}[n-1]$; dados incertos: I; np: número de partículas usadas no processo
Saída: $\hat{x}[n]$
$\hat{x}[n] \leftarrow f_a(\hat{x}[1], \hat{x}[2], \ldots, \hat{x}[n-1], I)$
$u \leftarrow$ índice do último desconhecido do sistema
$P \leftarrow \emptyset$
enquanto $|P| < np$ **faça**
  $d \leftarrow u$ amostras uniformes entre -1 e 1
  `/* Cria amostras das incertezas novas */`
  $a \leftarrow$ amostra(I)
  `/*` $\hat{x} \times d$ `cria um valor exato usando a amostra` $d \ldots$ `*/`
  $t \leftarrow f(\hat{x}[1] \times d, \hat{x}[2] \times d, \ldots, \hat{x}[n-1] \times d, a)$
  $P \leftarrow P \cup t$
fim
Seleciona partículas de P
Controla $\hat{x}[n]$ usando P
retorna $\hat{x}[n]$

Algoritmo 5: Cálculo do estado $\hat{x}[n]$ usando o método híbrido.

O método proposto é similar à **transformada unscented** [28]. Entretanto a grande diferença é que a transformada unscented alterna a representação por partículas com a representação paramétrica estatística e estas transições são feitas com métodos bem estabelecidos. No nosso método as transições são feitas entre formas afins e partículas e a transição de partículas para formas afins não é trivial.

3.2 Controle da forma afim

Um valor incerto representado por uma forma afim, obtida por uma longa cadeia de cálculos usando aritmética afim, pode ser dividido em três regiões:

Correta e Provável É a região onde o valor correto deve estar.

Correta e Improvável É a região onde o valor correto pode estar, mas com probabilidade baixa (a baixa probabilidade é uma consequência direta do Teorema Central do Limite, ou da combinação da informação vinda de várias fontes imprecisas).

Incorreta Algumas partes da forma afim devem estar incorretas, que podem ser consequência de aproximações e podem ser determinadas usando outras informações disponíveis.

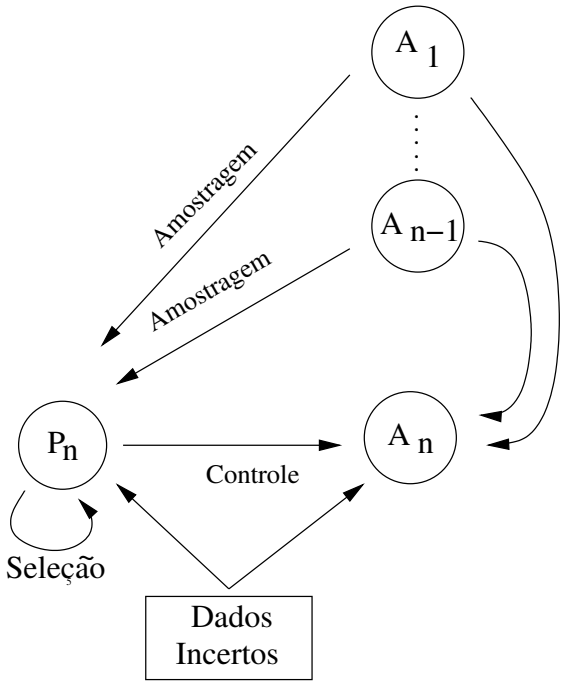


Figura 3.1: Visão geral do método híbrido.

Usando apenas a informação da forma afim, é difícil delimitar as três regiões descritas acima pois as distorções criadas pelas aproximações da aritmética afim inviabilizam a sua execução. Por exemplo, uma restrição de distância entre dois pontos representados por uma forma afim terá muito ruído (veremos que no cálculo da distância em $\mathcal{R}^3$ serão usadas quatro aproximações).

Uma vez selecionadas, as partículas são usadas para controlar a forma afim, passando para a forma afim os controles aplicados às partículas. A nova forma afim será mais compacta, formada pela região mais provável de conter os valores corretos e ainda mantendo a correlação existente nas formas afins.

Para controlar a forma afim, dividimos os desconhecidos que constituem a faixa de incerteza em dois grupos, de acordo com a sua origem:

1. Desconhecidos originais do problema.

2. Desconhecidos criados pelas aproximações de operações não afim.

Como os desconhecidos do segundo grupo podem crescer rapidamente e dominar a faixa de incerteza, precisamos diferenciá-los dos desconhecidos originais do problema.

O processo de controle da forma afim, usando partículas é descrito no Algoritmo 6: a faixa ocupada pelas partículas é calculada duas vezas, para porcentagens de propagações

diferentes (definidas pelos parâmetros $L1$ e $L2$ no algoritmo). A forma afim é então modificada, os desconhecidos originais ocupam a faixa definida por $L1$ e os demais desconhecidos ocupam a faixa definida por $L2 - L1$.

Input: $\hat{a}$: forma afim; P: partículas usadas no controle; $L1$: limite para os desonhecidos originais; $L2$: limite para os demais desconhecido; O: conjunto de desconhecidos originais; nO: conjunto dos demais desconhecidos
Output: $\hat{a}$ modificada
$c \leftarrow controide(P)$
$f1 \leftarrow faixa(P, L1)$
$f2 \leftarrow faixa(P, L2)$
$fo \leftarrow 0$
`/* Calculo da faixa dos originais */`
para cada $i \in O$ **faça**
  $fo \leftarrow fo + |\hat{a}_i|$
fim
$fno \leftarrow 0$
`/* Calculo da faixa dos não originais */`
para cada $i \in nO$ **faça**
  $fno \leftarrow fno + |\hat{a}_i|$
fim
`/* Altera o centro da forma afim */`
$\hat{a}_0 = c$
`/* Ajuste dos desconhecidos originais usando f1 */`
para cada $i \in O$ **faça**
  $\hat{a}_i \leftarrow (\hat{a}_i / fo) * f1$
fim
`/* Ajuste dos desconhecidos originais usando a diferença de f2 e f1 */`
para cada $i \in nO$ **faça**
  $\hat{a}_i \leftarrow (\hat{a}_i / fno) * (f2 - f1)$
fim

Algoritmo 6: Controle de uma forma afim usando partículas.

3.3 Controle das partículas

Como a forma afim contém regiões improváveis, espera-se que para um determinado estado, a faixa das partículas geradas seja menor que a faixa da forma afim. Mesmo assim, podemos melhorar a representação da incerteza feita pelas partículas. A seguir,

apresentamos duas formas de melhorar a representação.

3.3.1 Reamostragem com Importância

A reamostragem com importância (*Sample Importance Resample* -SIR) de partículas é um método usado para criar um conjunto de partículas melhores a partir de um conjunto inicial [2].

Descrito no Algoritmo 7, o SIR inicia com uma amostragem e com a pontuação do conjunto inicial. Também pontua-se uma versão com ruídos das partículas amostradas. A função de pontuação e de ruído são dependentes da aplicação. A importância é então usada para gerar a probabilidade de cada uma das partículas a ser escolhida na reamostragem, que é feita com reposição. Finalmente, um conjunto final de partículas é gerado através da amostragem. Espera-se que este conjunto tenha uma pontuação total melhor que a inicial.

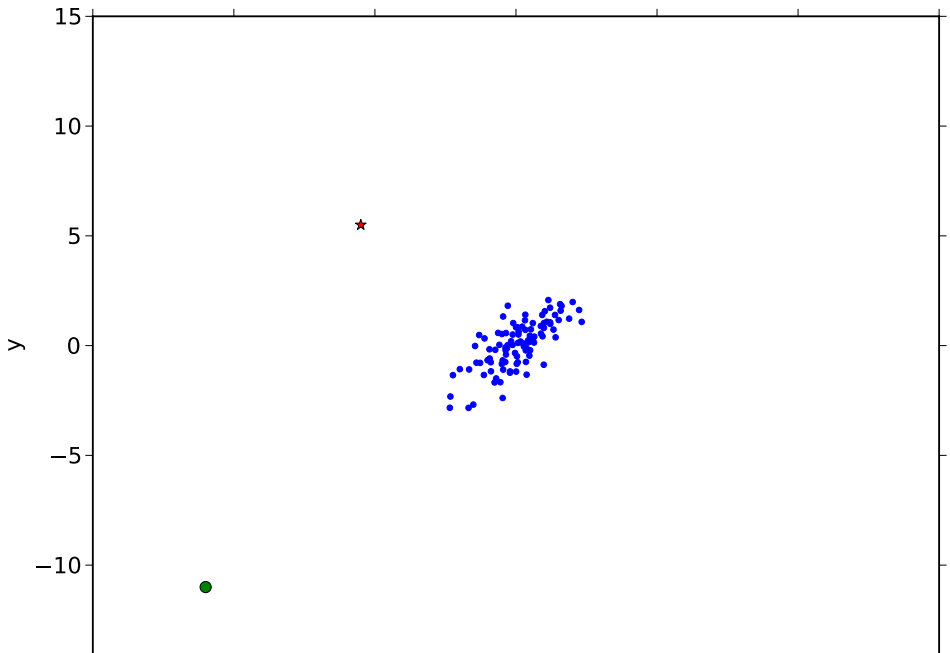
3.3.2 Seleção de Partículas por Distâncias

Também é possível implementar nas partículas um processo de seleção para a propagação de uma fração da incerteza representada pela partícula.

Quando queremos calcular a distância de uma determinada partícula em relação às demais partículas, não podemos usar a distância euclidiana. Na Figura 3.2, o centróide das partículas está na origem. O ponto em vermelho está na posição (-5,5;5,5) e a uma distância de 7,7 da origem. Já o ponto verde está na posição (-11;-11) e a uma distância de 15,5 da origem. Estas medidas não consideram a dispersão das partículas.

Entrada: P: conjunto de partículas; t: tamanho da saída; k: o número de partículas criadas para cada partícula existente.
Saída: Conjunto de t partículas selecionadas.
$T \leftarrow \emptyset$
enquanto $|T|$ *não é suficiente* **faça**
  $p \leftarrow amostra(P)$
  $T \leftarrow T \cup (p, pontuacao(p))$
  para $0 \leq i < k$ **faça**
    `/* A função` *perturba* `acrescenta um erro em na entrada */`
    $pp \leftarrow pertuba(p)$
    $T \leftarrow T \cup (pp, pontuacao(pp))$
  fim
fim
$S \leftarrow 0$
`/* Acumula em` S `a pontuação,`$T_{i,1}$ `contém a pontuação de` $T_{i,0}$ `*/`
para $0 \leq i < |T|$ **faça**
  $S \leftarrow S + T_{i,1}$
fim
`/* ...e cria as probabilidades */`
para $0 \leq i < |T|$ **faça**
  $T_{i,1} \leftarrow T_{i,1}/S$
fim
$R \leftarrow \emptyset$
para $0 \leq i < t$ **faça**
  $temp \leftarrow$ amostra de T de acordo com as probabilidades em $T_{i,1}$
  $R \leftarrow R \cup \{temp\}$
fim
retorna R

Algoritmo 7: Reamostragem com Importância.



Uma forma de calcular a distância de uma partícula em relação as demais é obter a sua verossimilhança, usando a Distância de Mahalanobis [41],

$$DM(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}, \tag{3.1}$$

onde $x \in \mathcal{R}^n$, S^{-1} é o inverso da matriz de covariância das partículas e μ é a sua média. O uso da covariância nos permite criar elipses com a mesma verossimilhança, e assim, sabermos qual partícula está mais distante. Na Figura 3.3, vemos que a partícula em vermelho tem uma distância de Mahalanobis de 14,24, enquanto a partícula verde tem distância 11,4.

No Algoritmo 8, implementamos a seleção de partículas usando a Distância de Mahalanobis, a distância de cada partícula é calculada e as partículas com menor distância são selecionadas.

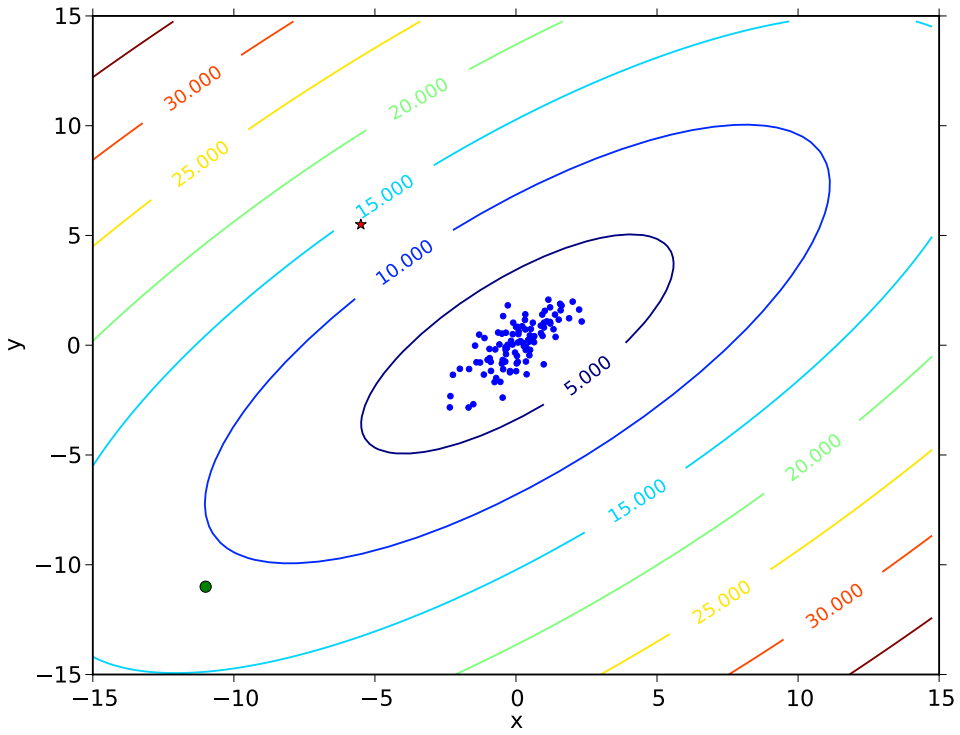


Figura 3.3: Distância de Mahalanobis aplicada nas partículas. A partícula em vermelho tem distância euclidiana até o centróide de 7,7 e distância de Mahalanobis de 14,24. A partícula em verde tem distância euclidiana de 15,5 e distância de Mahalanobis de 11,4.

3.4 Criação de partículas

Em diversos momentos, precisamos da representação exata de uma solução incerta. Para um único estado, uma amostragem das partículas consegue resolver o problema, mas para um sistema multi-estado, uma simples amostragem pode gerar um conjunto inconsistente.

Usando a representação afim, conseguimos gerar amostras mais coerentes dos vários estados amostrando valores para os desconhecidos. Como a correlação entre as formas afins se dá pelo compartilhamento dos desconhecidos, quando usamos os mesmos valores de desconhecidos para várias formas afins, a correlação entre os estados é mantida, a não ser pelas aproximações.

Entrada: P: conjunto de partículas; k: limite para seleção das partículas($0 < k \leq 1$)
Saída: Conjunto de Partículas selecionadas
$T \leftarrow \emptyset$
para $0 \leq i < |P|$ **faça**
    `/* Armazena o par partícula,distância */`
    $T_i \leftarrow (P_i, distanciaMahalobis(P_i, P))$
fim
ordena T em ordem crescente pela distância
$R \leftarrow \emptyset$
para $0 \leq i < |P| * k$ **faça**
    `/* Armazena a partícula */`
    $R \leftarrow R \cup \{T_{i,0}\}$
fim
retorna R

Algoritmo 8: Seleção de Partículas pela Distância de Mahalanobis.

Capítulo 4

Determinação da Estrutura de Proteínas

O método de propagação de incertezas descrito no Capítulo 3 será aplicado na determinação da estrutura de proteínas usando dados de Ressonância Magnética Nuclear. Neste Capítulo, descrevemos as proteínas e como a sua estrutura pode ser obtida.

4.1 Introdução

As proteínas são essenciais para os seres vivos. Elas executam inúmeras funções dentro dos organismos: estruturais, hormonais, defesa entre outras. Quimicamente, elas são polímeros formados por uma sequência de aminoácidos. Existem 20 aminoácidos naturais que podem fazer parte de uma proteína. Na Figura 4.1, mostramos uma esquematização da composição química de uma proteína, onde R é chamada de cadeia lateral e é o que distingue um aninoácido de outro. O carbono que se liga a R é chamado de carbono alfa. A sequência H-(N-C_α-C)$_n$-OH, onde n é o número de aminoácidos, forma o *backbone* de uma proteína. As principais características estruturais são definidas pelo *backbone* e, em geral, o seu conhecimento é suficiente para determinar a estrutura de uma proteína [3].

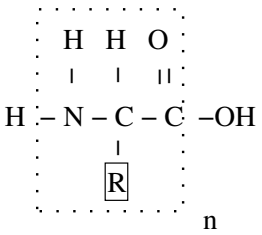


Figura 4.1: Proteína esquematizada

A função executada por cada proteína está diretamente ligada à sua forma. Portanto, o conhecimento da estrutura de uma proteína é essencial para a execução de diversas tarefas, como classificação, comparação, inferência de função ou *docking* [39].

Existem dois métodos laboratoriais principais para a determinação da estrutura: Cristalografia por Raio-x e Ressonância Magnética Nuclear. Existe um banco público de proteínas, o *Protein Data Bank* (`www.pdb.org`), que centraliza informações sobre a estrutura de proteínas e também define um padrão para a publicação da estrutura de proteínas. As proteínas publicadas nem sempre respeitam as informações obtidas pelos experimentos laboratoriais ou mesmo a geometria que toda molécula deve ter [14, 46, 42, 38].

4.2 Geometria Molecular

A **Geometria Molecular** estuda os padrões geométricos existentes nas moléculas. Estes padrões podem ser usados para auxiliar o processo de determinação da estrutura de proteína.

As **ligações covalentes** são caracterizadas por pares de elétrons compartilhados por dois átomos. Como as propriedades geométricas das ligações covalentes são muito rígidas, no processo de determinação da estrutura de proteína, podemos considerar como conhecidas as distâncias entre átomos separados por uma a duas ligações covalentes [18].

As distâncias entre C_α consecutivos de uma proteína têm variação muito pequena e também são consideradas constantes na maioria dos experimentos [5].

4.3 Métodos Laboratoriais

A cristalografia por raio-x é o método mais preciso usado para determinar a estrutura de proteínas [15]. Inicialmente, a proteína é cristalizada. Em seguida, usando a densidade dos elétrons, a estrutura da proteína é calculada. Este método apresenta duas principais limitações [49, 24]:

- A proteína, que tem certa flexibilidade natural, tem sua estrutura alterada pelo processo de cristalização;
- Algumas proteínas não podem ser cristalizadas.

A Ressonância Magnética Nuclear (RMN) é um método menos preciso que a cristalografia, mas não apresenta as suas desvantagens, pois os experimentos de RMN ocorrem em meio aquoso, mais natural para as proteínas, permitindo que elas sejam estudadas na sua conformação natural.

A RMN mede a interação entre os *spins* de elétrons de átomos, principalmente hidrogênios. A interação medida é classificada em fraca, média ou forte, e então a distância entre estes átomos pode ser inferida. Na Tabela 4.1, são apresentadas as distâncias usadas nos processos de determinação de estrutura [24]. As distâncias inter-atômicas, juntamente com outras informações geométricas, são usadas para determinar a estrutura da proteína.

Tipo de Iteração	**Distância**
Forte	Até 2,7Å
Média	2,7Å a 3,3 Å
Fraca	3,3Å a 5Å

Tabela 4.1: Distâncias obtidas por RMN.

4.4 Problema da Geometria de Distâncias Moleculares

O Problema da Geometria de Distâncias Moleculares (**PGDM**) está relacionado com a determinação da estrutura de proteínas usando as informações da RMN e pode ser definido da seguinte forma:

Dados um conjunto S de pares de átomos (i, j) sobre um conjunto de m átomos e distâncias d_{ij} definidas sobre S, determine as posições $x_1, \ldots, x_m \in \mathcal{R}^3$ dos átomos da molécula tal que

$$||x_i - x_j|| = d_{ij} \ \forall (i, j) \ \in S \tag{4.1}$$

Quando as distâncias entre todos os pares de átomos da molécula são conhecidas, uma estrutura única pode ser determinda em tempo linear [12]. Entretanto, devido a erros experimentais, a solução pode não existir ou pode não ser única, tornando o PGDM um problema difícil no caso geral. O problema também é chamado de **Problema de Imersão de Grafos**, Saxe [40] mostrou que o PGDM é NP-Completo mesmo no caso unidimensional. A prova consiste em reduzir o **Problema da Partição de Conjuntos** ao problema de imersão com $k = 1$. O mecanismo da redução é bem direto: para cada elemento do conjunto que queremos particionar é criado uma aresta com o mesmo valor. Se existe uma imersão na reta, então os pontos à direita do seu sucessor formam uma partição e os demais pontos outra partição. O processo é descrito no Algoritmo 9).

O PGDM pode ser naturalmente formulado como um problema de otimização global

Entrada: Conjunto S de inteiros
Saída: S_e,S_d se houver partição
para *i de 1 a* $|S|$ **faça**
  $d_{i,i\ (\mathrm{mod}\ S)+1} = S_i$;
fim
$V = PGD(d)$;
para *i de 1 a* $|V|$ **faça**
  se $V_i < V_{i\ (\mathrm{mod}\ |V|)+1}$ **então**
    $S_e = S_e \cup \{V_i\}$;
  else
    $S_d = S_d \cup \{V_i\}$;
  end
fim
retorna S_e, S_d;

Algoritmo 9: Redução do Problema da Partição para PGDM

não linear, cuja função objetivo é dada por

$$f(x_1, \ldots, f_m) = \sum_{(i,j)\in S} (||x_i - x_j||^2 - d_{ij}^2)^2. \tag{4.2}$$

Esta função é infinitamente diferenciável em todo o domínio e tem um número exponencial de mínimos locais. Assumindo que todas as distâncias estão corretas , $x \in \mathcal{R}^{3m}$ é uma solução para o problema se, e somente se, $f(x) = 0$.

Como erros experimentais podem ocorrer, podemos também considerar uma solução ε-ótima, isto é uma solução $x_1, \ldots, x_m$ que satisfaça

$$|||x_i - x_j|| - d_{ij}| \leq \varepsilon \ \forall\ (i,j) \in S. \tag{4.3}$$

Em [37] é mostrado que obter uma solução ε-ótima é NP-Difícil para ε pequeno o suficiente.

Na prática, os experimentos de RMN obtêm os limites superiores e inferiores das distâncias, assim uma definição mais realista do PGDM é determinar $x_1, \ldots, x_m \in \mathcal{R}^3$ tal que

$$l_{ij} \leq ||x_i - x_j|| \leq u_{ij}\ \forall (i,j)\ \in S, \tag{4.4}$$

onde l_{ij} e u_{ij} são respectivamente os limites inferior e superior da distância d_{ij}.

O PGDM é um caso particular do Problema da Geometria de Distâncias [6, 32] que está relacionado com o problema de completar a matriz de distâncias euclidianas [4, 26].

4.5 Geometric Build-Up

Em [12], é apresentado um algoritmo em tempo polinomial para determinar a estrutura de uma proteína dado um grafo suficientemente denso, com distâncias exatas. O Algoritmo iterativamente constrói a proteína e é chamado de **Geometric Build-Up**.

Usando o fato de que, para um átomo qualquer, se forem conhecidas as distâncias para quatro átomos cujas posições são conhecidas, pode-se determinar sua posição (Algoritmo 11).

Assim, com uma clique de tamanho 4, o algoritmo constrói uma base de quatro átomos (Algoritmo 10) e iterativamente a proteína é construída (Algoritmo 12).

É razoável acreditar que este algoritmo determine a estrutura da maioria das proteínas, visto que existem as informações geométricas e do experimento de RMN.

Existem algumas variações para o algoritmo, visando melhorar os problemas causados pela instabilidade numérica [13], ampliar o conjunto de estruturas que são determinadas [11], aumentar sua eficiência [8], mas sempre considerando distâncias exatas.

Entrada: Distâncias $d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}$
Saída: Base geométrica (x_i, y_i, z_i) para $i \in \{1, 2, 3, 4\}$
$(x_1, y_1, z_1) = (0, 0, 0)$
$(x_2, y_2, z_2) = (d_{12}, 0, 0)$
$x_3 = (d_{13}^2 * d_{23}^2)/(d_{12} * 2) + (d_{12}/2)$
$y_3 = \sqrt{d_{13}^2 - x_3^2}$
$z_3 = 0$
$x_4 = (d_{14}^2 - d_{24}^2)/(d_{12} * 2) + d_{12}/2$
$y_4 = d_{24}^2 - d_{34}^2 - (x_4 - d_{12})^2 + (x_4 - x_3)^2$
$y_4 = y_4/(y_3 * 2) + y_3/2$
$z_4 = \sqrt{d_{14}^2 - x_4^2 - y_4^2}$
retorna (x_i, y_i, z_i) para $i \in \{1, 2, 3, 4\}$

Algoritmo 10: Cria uma base geométrica.

Entrada: Pontos $p_i = (x_i, y_i, z_i)$ e distâncias d_i para $i \in \{1, 2, 3, 4\}$
Saída: $p = (x, y, z)$

$$A = 2 \begin{bmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \\ x_1 - x_3 & y_1 - y_3 & z_1 - z_3 \\ x_1 - x_4 & y_1 - y_4 & z_1 - z_4 \end{bmatrix}$$

$$b = \begin{bmatrix} ||p_1||^2 - ||p_2||^2 - (d_1^2 - d_2^2) \\ ||p_1||^2 - ||p_3||^2 - (d_1^2 - d_3^2) \\ ||p_1||^2 - ||p_4||^2 - (d_1^2 - d_4^2) \end{bmatrix}$$

resolve $Ap = b$
retorna p

Algoritmo 11: Calcula a posição de um ponto dada uma base.

Entrada: $G(V, E)$ com distâncias exatas nas arestas
Saída: $G(V, E)$ com as posições nos vértices
D = findClique4(G)
Cria uma base usando o Algoritmo 10
enquanto $D \neq G(V)$ **faça**
    Encontre um vértice v em $G(V) - D$ com ao menos 4 arestas em D
    Determine a posição de v usando 4 vizinhos em D e o Algoritmo 11
    $D = D \cup \{v\}$
fim

Algoritmo 12: Geometric Build-Up com distâncias exatas.

Capítulo 5

Determinação da Estrutura de Proteínas com incertezas

5.1 Algoritmos Desenvolvidos

Os algoritmos desenvolvidos neste capítulo são extensões dos algoritmos descritos no Capítulo 4. Nos Algoritmos 11 e 10, as operações de ponto flutuante foram substituídas por operações em partículas e por operações afins.

O Algoritmo 13 foi desenvolvido modificando o Algoritmo 12, onde além das operações aritméticas sobre formas afins e partículas, foram inseridos a seleção de partículas e o controle da forma afim pelas partículas.

A implementação dos algoritmos descritos foi feita em Python, com amplo uso das bibliotecas SciPy[1] e NumPy[2]. O *parsing* dos arquivos de proteínas foi feito pela biblioteca BioPython[3].

[1] `http://www.scipy.org/`
[2] `http://www.numpy.org/`
[3] `http://www.biopython.org`

Entrada: $G(V, E)$: grafo com distâncias intervalares
Saída: $G(V, E)$ com as posições dos vértices determinadas
D = findClique4(G)
Cria uma base para os vértices usando partículas
Cria uma base para os vértices usando aritmética afim
while $D \neq G(V)$ **do**
    Encontre um vértice v em $G(V) - D$ com ao menos 4 arestas em D
    Encontre 4 vértices u_i em D, vizinhos de v
    para $1 \leq i \leq 4$ **faça**
        `/* Novos vizinhos podem mudar` u_i `*/`
        Aplica a seleção de partículas SIR em u_i
        Controla a forma afim de u_i
    fim
    Calcule a posição afim de v usando u_i e o Algoritmo 11
    Calcule a posição com partículas de v usando o algoritmo 11 e com partículas amostradas da forma afim
    Aplica a seleção de partículas SIR em v
    Controla a forma afim de v usando as partículas
    $D = D \cup \{v\}$
end

Algoritmo 13: Geometric Build-Up com controle híbrido da incerteza.

5.2 Conjunto de Testes

Para testar o método, foram obtidas as estruturas de proteínas originadas por RMN disponíveis no *Protein Data Bank*[4] em oito de outubro de 2012, totalizando 8366 proteínas. Destas, 367 foram selecionadas para os testes pois não apresentaram erros: o *parser* do Biopython conseguiu fazer a leitura sem erros e as distâncias inter-atômicas respeitavam a geometria molecular.

5.3 Construção das restrições

Nos experimentos de construção de proteínas, usamos as proteínas existentes para criar restrições que imitam as restrições disponíveis nos experimentos de determinação de estrutura usando dados de RMN. Serão construídos apenas os átomos do *backbone* da proteína,

[4] `www.pdb.org`

pois estes átomos são suficientes para definir a maior parte das suas características estruturais. Assim, o conjunto de restrições é criado da seguinte forma:

- Para átomos separados por uma ou duas ligações covalentes, cria-se uma restrição com o valor exato da distância obtida na estrutura original.
- Para C_α consecutivos, também usamos a distância exata disponível.
- Os demais átomos são analisados par a par, e são criadas restrições de distâncias intervalares com limites de 2Å a 3Å, 3Å a 4Å ou 4Å a 5Å, caso a distância original pertença a um destes intervalos.

5.4 Pré e pós-processamento

Antes de iniciar a montagem da proteína, as restrições passam pelo processo de *bound smoothing* [43], que usa a desigualdade triangular para diminuir a faixa de incerteza das restrições. Para distâncias intervalares, um conjunto de três pontos i, j e k com as distâncias conhecidas deve obedecer as seguintes inequações:

$$u_{ij} \leq u_{ik} + u_{kj} \tag{5.1}$$

$$l_{ij} \leq l_{ik} + l_{kj} \tag{5.2}$$

onde l_{ij} e u_{ij} são as distâncias mínima e máxima entre os pontos i e j. Iterativamente, estas inequações são usadas para gerar distâncias com faixas menores (Algoritmo 14).

Entrada: $G(V, E)$: grafo com as distâncias nas arestas; m : mínimo de melhoria para cada passo
Saída: Grafo G alterado
repita
  $\delta \leftarrow 0$
  para cada *clique* (i, j, k) *de* G **faça**
    aplica a desigualdade em (i, j, k)
    acumula em δ a alteração do intervalo
  fim
até $\delta < m$

Algoritmo 14: Desigualdade triangular.

O pós-processamento melhora a qualidade de uma proteína exata com uma adaptação para restrições intervalares da função de *update* do algoritmo proposto em [1]. O Algoritmo 15 altera a posição de um determinado ponto caso a distância entre ele e um

segundo ponto, usado como referência, não seja respeitada. Cada alteração faz com que a distância dos pontos fique mais próxima da distância usada como restrição, o fator de aproximação é controlado pela variável λ. O Algoritmo 16 escolhe aleatoriamente o par de pontos que será usado na atualização e controla o fator λ.

Entrada: $x_i[], x_j[]$: posições em $\mathcal{R}^3$; l: limite inferior da distância; u: limite superior da distância; λ: fator de aprendizado
Saída: $x_j[]$ alterado se necessário
$d \leftarrow ||x_i - x_j||$
se $d < l$ **então**
  $o \leftarrow l - d$
  $sa \leftarrow o/d$
  para $dim \in [1, 2, 3]$ **faça**
    $des \leftarrow \lambda * (sa) * (x_i[dim] - x_j[dim])$
    $x_j[dim] \leftarrow x_j[dim] - des$
  fim
  retorna
fim
se $u > d$ **então**
  $o \leftarrow d - u$
  $sa \leftarrow o/d$
  para $dim \in [1, 2, 3]$ **faça**
    $des \leftarrow \lambda * (sa) * (x_i[dim] - x_j[dim])$
    $x_j[dim] \leftarrow x_j[dim] + des$
  fim
  retorna
fim

Algoritmo 15: Função *update* para distâncias imprecisas.

Entrada: S: estrutura com erros; D: conjunto de distâncias intervalares
Saída: Estrutura alterada
para λ *de* 10000 *a* 1000 **faça**
  Escolhe uma d_{ij} de D
  update$(S_i, S_j, d_{ij}.min, d_{ij}.max, \lambda/10000)$
fim

Algoritmo 16: Executa a iteração para a função *update*.

5.5 Métricas

Usualmente, as proteínas são comparadas usando o *Root-mean-square deviation* (RMSD) [24]:

$$RMSD(a,b) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||a_i - b_i||^2}, \tag{5.3}$$

onde a e b são as proteínas sendo comparadas. Antes do RMSD ser calculado as proteínas preciam ser superimpostas [24].

Como usamos dados imprecisos, mesmo com as restrições sendo satisfeitas, o RMSD obtido pode ser alto. Não existe um padrão para determinar a qualidade de uma estrutura em função do RMSD. Em [19], é afirmado que um RMSD de 6 Å é suficiente para determinar a função da proteína e [45] considera de 4 a 6 Å significativo. Já em [7], temos a estrutura da mesma proteína, determinada por experimentos diferentes com RMSD próximo de 15Å.

Para analisar os resultados, também serão usadas uma métrica para verificar cada restrição em relação a posição do átomo descrita por partículas e outra métrica para analisar uma proteína exata em relação às restrições existentes.

5.5.1 Restrições Aplicadas às Partículas

Para determinar se dois átomos (cujas posições não são exatas e estão representados por partículas) obedecem a uma determinada restrição intervalar de distância, são calculadas a média μ e o desvio padrão σ das distância entre as partículas de cada átomo. Uma restrição [min,max] é k-satisfeita se:

$$\mu - k\sigma \leq min \leq \mu + k\sigma$$

ou

$$\mu - k\sigma \leq max \leq \mu + k\sigma.$$

5.5.2 Restrições da estrutura

Para analisarmos quais restrições uma estrutura exata obedece, primeiramente, precisamos determinar uma estrutura a partir da estrutura incerta, definida por partículas ou por formas afins. Dois métodos foram desenvolvidos:

1. Centróide das partículas. O centróide de cada conjunto de partículas é calculado e a estrutura exata é determinada.

2. Amostragem dos desconhecidos. Com as posições dos átomos representadas por formas afins, amostras dos desconhecidos são feitas e uma estrutura exata é criada subistituindo os desconhecidos amostrados nas formas afins.

5.6 Exemplo

Antes de apresentar os resultados obtidos, mostraremos um exemplo de como o método híbrido funciona. Neste exemplo usaremos a proteína com o identificador **1b0q** e determinaremos os seus sete primeiros átomos. Na Tabela 5.1 mostramos as distâncias que serão usadas.

	1	2	3	4	5	6	7
1	-	1,33	2,49	[3; 4]	[3; 4]	[4; 5]	[4; 5]
2	1,33	-	1,45	2,48	[3; 4]	[4; 5]	Desconhecida
3	2,49	1,45	-	1,53	2,42	3,83	[4; 5]
4	[3; 4]	2,48	1,53	-	1,33	2,49	[3; 4]
5	[3; 4]	[3; 4]	2,42	1,33	-	1,45	2,42
6	[4; 5]	[4; 5]	3,83	2,49	1,45	-	1,53
7	[4; 5]	Desconhecida	[4; 5]	[3; 4]	2,42	1,53	-

Tabela 5.1: Distâncias inter-atômicas dos sete primeiros átomos.

Os quatro primeiros pontos são determinados usando o Algoritmo 10. Formando uma clique com distâncias exatas nas arestas, os átomos 1,2 e 3 são determinados de forma precisa:

- Átomo 1 posição $(0;\ 0;\ 0)$
- Átomo 2 posição $(1,33;\ 0;\ 0)$
- Átomo 3 posição $(2,19;\ 1,18; 0)$

Como a distância entre o quarto e o primeiro átomo é imprecisa, a sua localização também é imprecisa. Na Figura 5.1 temos os pontos em verde, a representação em partículas do quarto átomo e em vermelho a estrutura original.

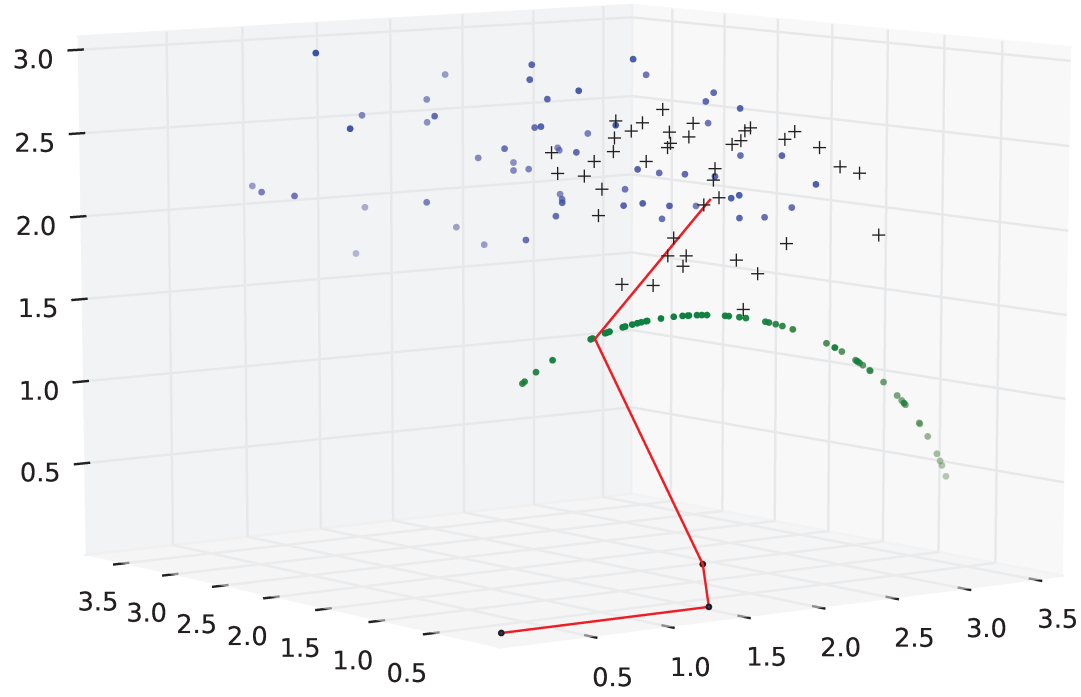


Figura 5.1: Cinco átomos da proteína **1b0q**. Em vermelho a estrutura original. Os pontos em verde indicam a posiçõa imprecisa do quarto átomo. Os pontos em azul a localização imprecisa do quinto átomo antes do processo de seleção e os sinais de '+', em preto, as posições do quinto átomo após a seleção.

Nos cálculos do quinto átomo, é obtido um conjunto inicial de partículas (pontos em azul da Figura 5.1). Este conjunto é melhorado usando o processo de SIR (sinais '+' em preto, na mesma figura). As incertezas das formas afins para este ponto são inicialmente muito grandes com diâmetros de 5,4, 7,2 e 10,2 para os eixos x, y e z, respectivamente. Após serem controladas pelas partículas usando o Algoritmo 6, os diâmetros das formas afins ficam os valores 0,65, 1,2 e 0,72.

Na Figuras 5.2 e 5.3, mostramos o processo para o sexto e o sétimo átomos, respectivamente. Em azul, as partículas iniciais e em verde, com o sinal '+', as partículas após o SIR.

O processo se repete até a proteína ser montada completamente ou que não seja possível determinar a sua estrutura por falta de distâncias conhecidas.

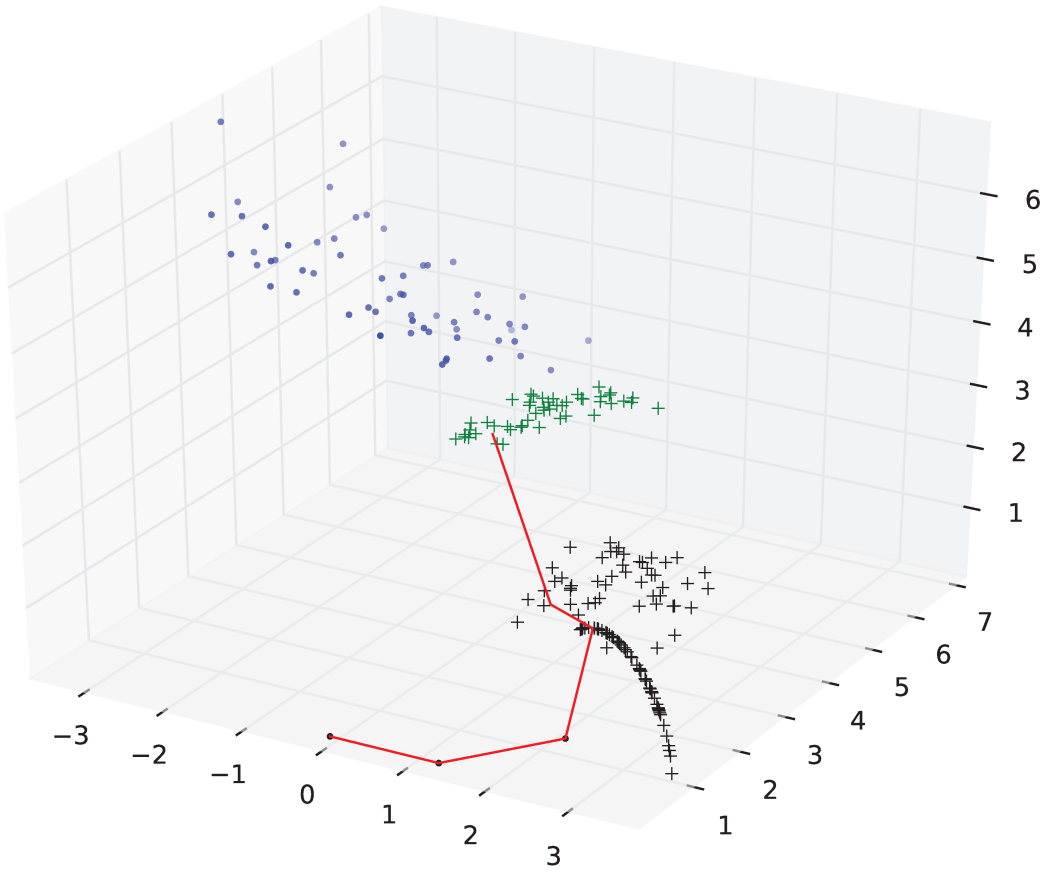


Figura 5.2: Seis átomos da proteína **1b0q**.Os pontos em azul, a localização imprecisa do sexto átomo antes da seleção e os sinais de '+', em preto, as posições do sexto átomo após a seleção.

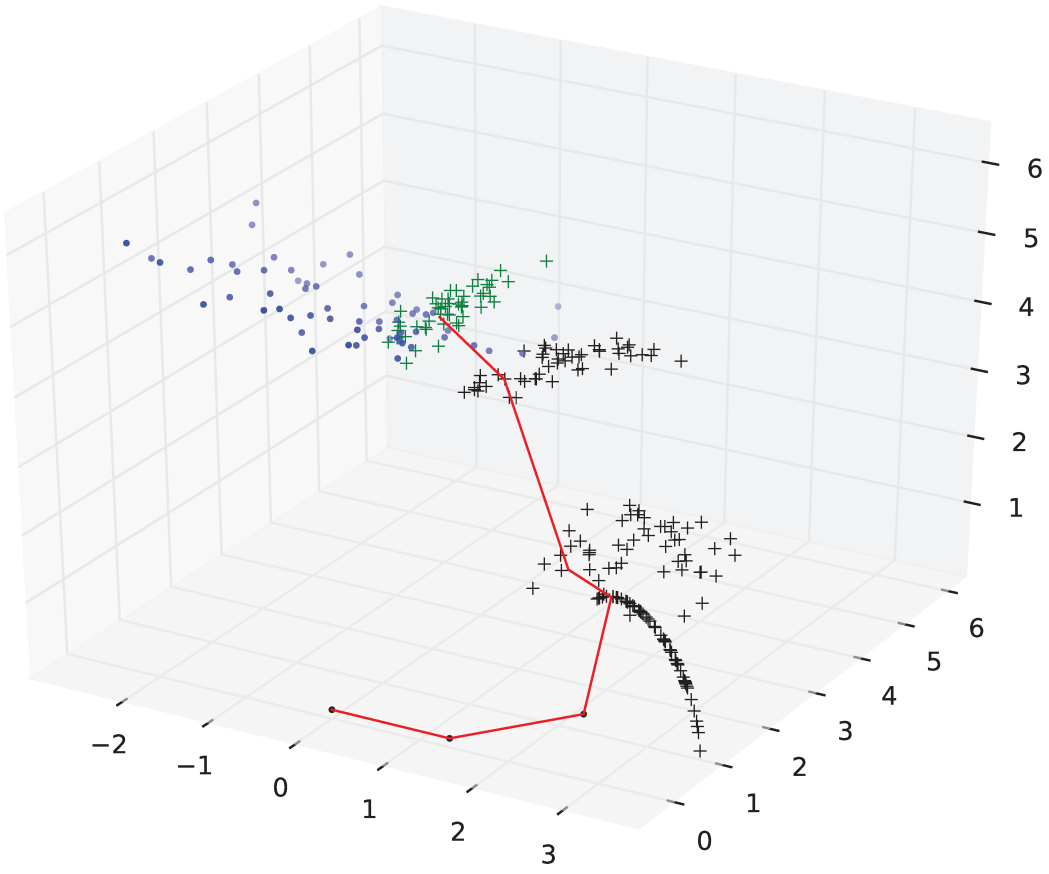


Figura 5.3: Sete átomos da proteína **1b0q**.Os pontos em azul, a localização imprecisa do sétimo átomo antes da seleção e os sinais de '+', em preto, as posições do sétimo átomo após a seleção.

5.7 Resultados Obtidos

Nas tentativas de montagem usando apenas a aritmética afim, sem o controle das partículas, não foi possível concluir a determinação da estrutura de nenhuma das proteínas. O crescimento da incerteza é muito rápido e as operações matemáticas não podem ser efetuadas.

Usando apenas partículas, com 60 partículas por átomo e propagando 85% da incerteza, 261 experimentos foram finalizados com sucesso. Usando o método híbrido com as mesmas configurações, 274 dos experimentos obtiveram êxito.

Nas Figuras 5.4, 5.5 e 5.6 mostramos a evolução das restrições aplicadas às partículas atendidas conforme a variação de k.

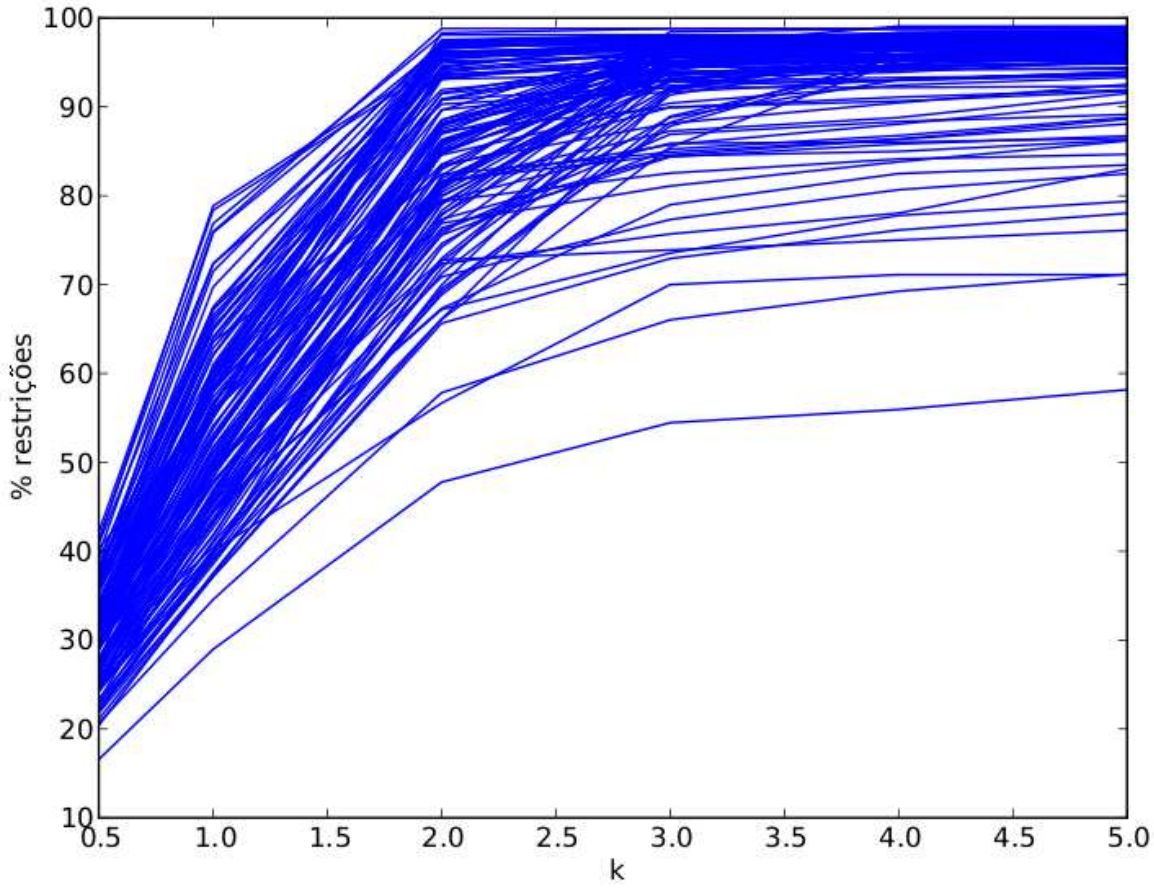


Figura 5.4: Evolução da restrições atendidas com a variação de k para proteínas de até 100 átomos.

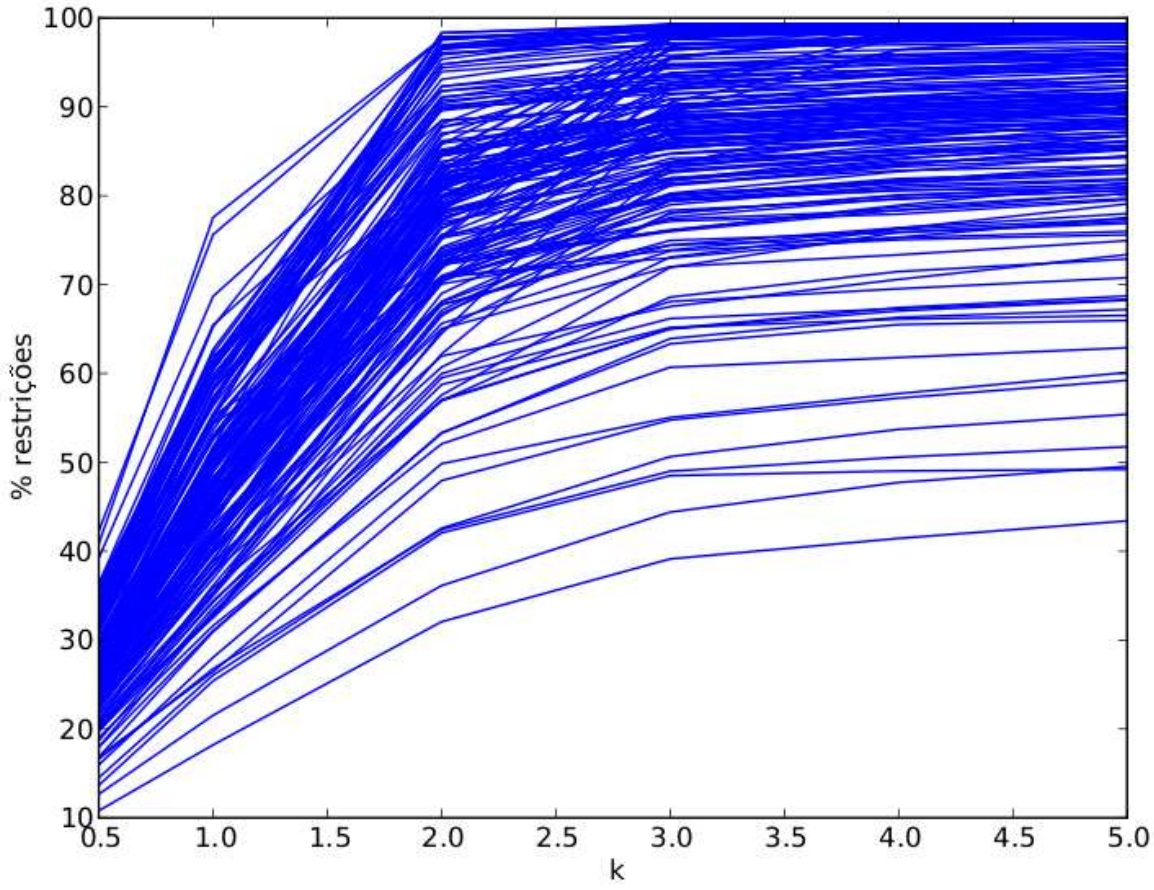


Figura 5.5: Evolução da restrições atendidas com a variação de k para proteínas de 100 até 200 átomos.

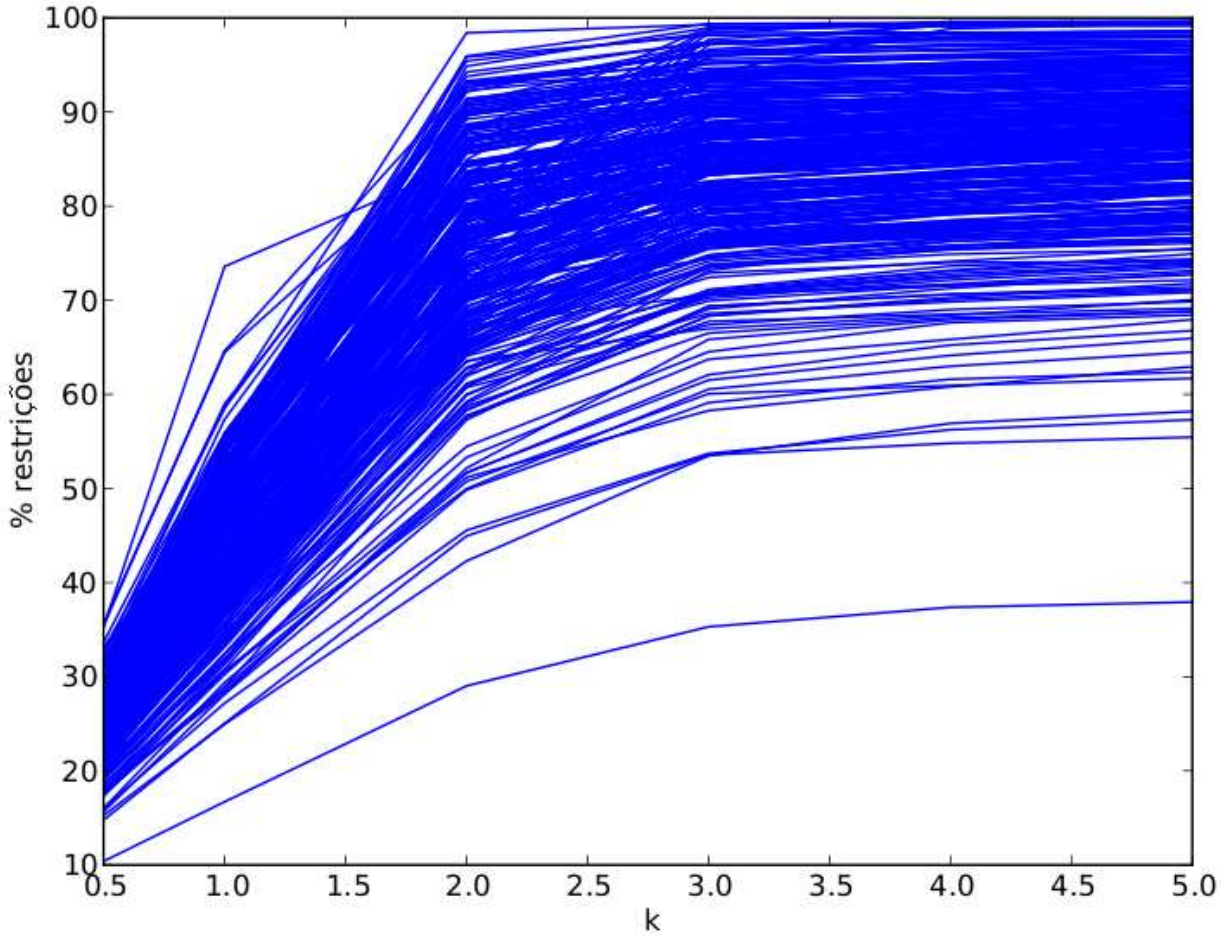


Figura 5.6: Evolução da restrições atendidas com a variação de k para proteínas com mais 200 átomos.

Na Tabela 5.2, são mostradas as porcentagens de restrições respeitadas pelas proteínas geradas pelo experimento com partículas e pelo método híbrido.

Experimento/Tamanho	**50**	**100**	**150**	**200**	**>200**
Partículas	65,9	68,4	63,0	64,5	63,2
Híbrido	73,2	73,8	62,5	63,7	64,2

Tabela 5.2: Porcentagem das restrições atendidas.

Na Tabela 5.3 mostramos a média dos RMSD's obtidos, agrupados pelo tamanho da proteína. Nas Tabelas 5.4 e 5.5, mostramos a porcentagem dos alinhamentos que ficaram com RMSD abaixo de 6 Å e 10Å, respectivamente.

RMSD/Tamanho	**50**	**100**	**150**	**200**	**>200**
Partículas	2,8	4,6	7,2	8,4	10,5
Híbrido	3,2	4,6	6,7	8,1	9,9

Tabela 5.3: RMSD em relação a proteína original.

RMSD/Tamanho	**50**	**100**	**150**	**200**	**>200**
Partículas	100	77,4	23,8	10,1	3,8
Híbrido	100	74,1	47,6	10,0	2,8

Tabela 5.4: Porcentagem dos experimentos com RMSD menor que 6Å.

RMSD/Tamanho	**50**	**100**	**150**	**200**	**>200**
Partículas	100	100	95,0	80,0	54,2
Híbrido	100	100	90,0	86,6	55,2

Tabela 5.5: Porcentagem dos experimentos com RMSD menor que 10Å.

Na Figura 5.7, mostramos o resultado do alinhamento dos 186 átomos do *backbone* da proteína **1era**, com RMSD de 4,6 Å.

Nas Figuras 5.8 e 5.9, temos os resultados para as proteínas **2gp8**, com 120 átomos e RMSD de 4,3 Å, e **1t2y**, com 75 átomos e RMSD de 1,4 Å.

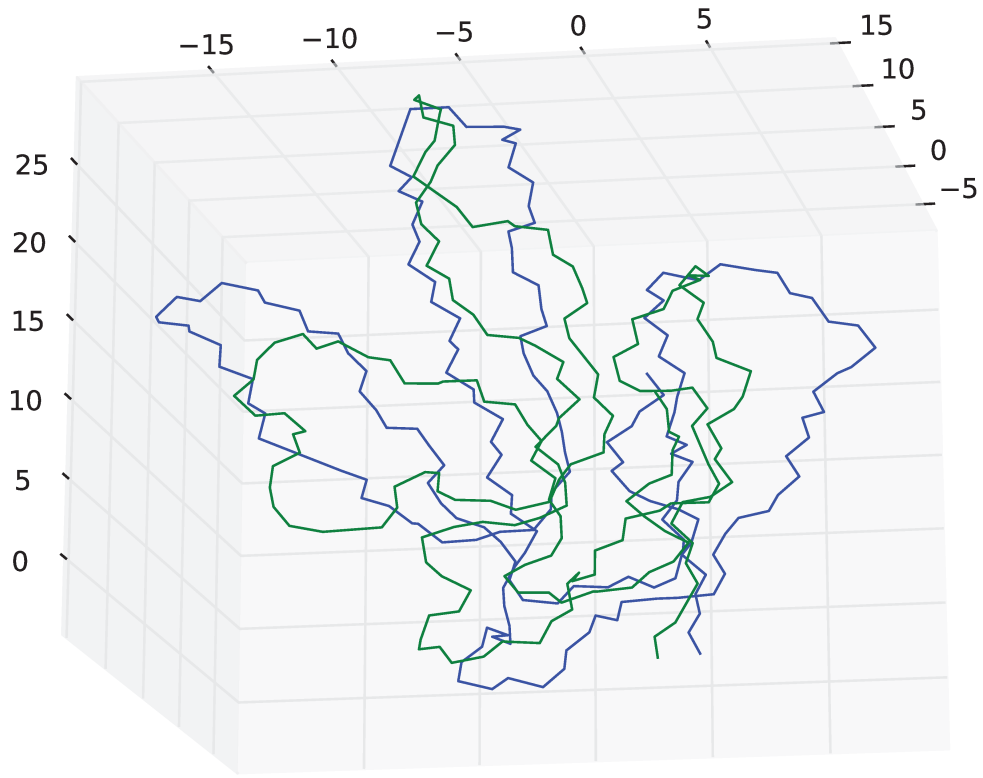


Figura 5.7: Alinhamento da proteína **1era** reconstruída (em azul) com a proteína original (em verde).

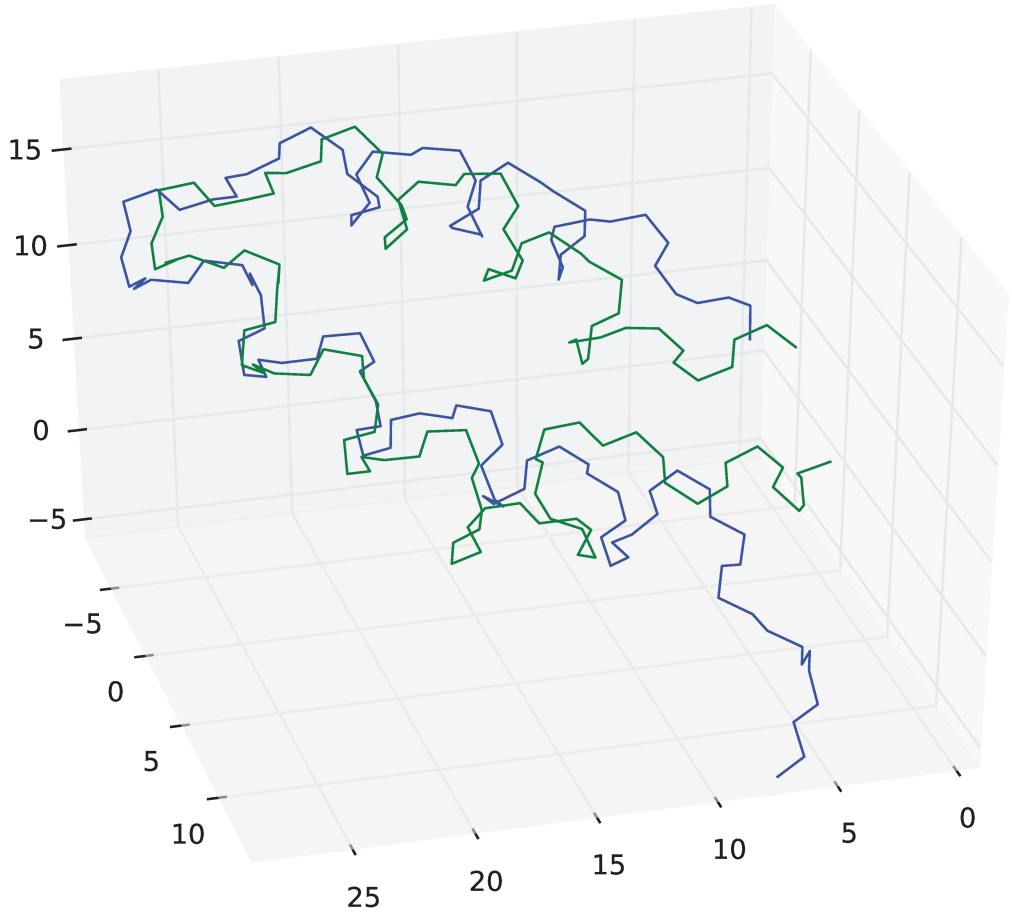


Figura 5.8: Alinhamento da proteína **2gp8** reconstruída (em azul) com a proteína original (em verde).

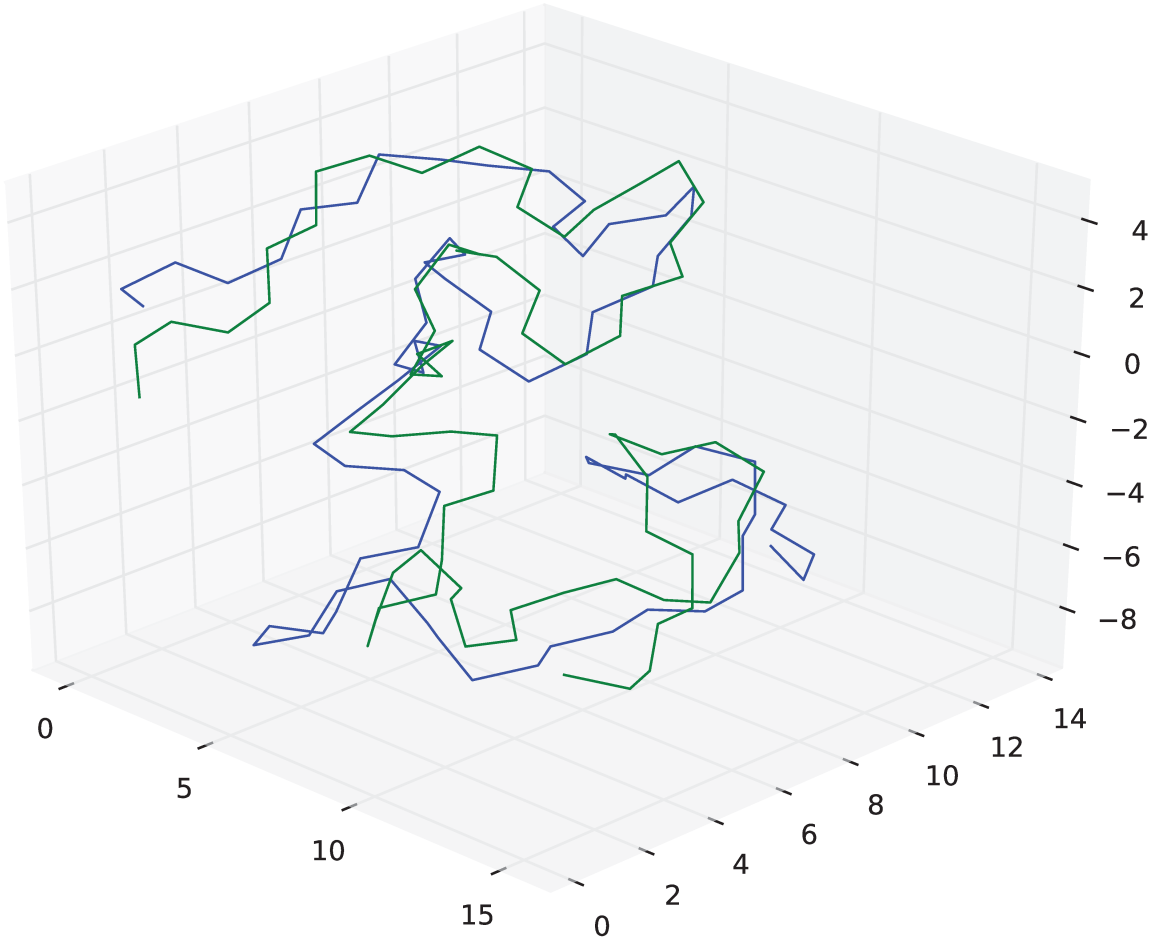


Figura 5.9: Alinhamento da proteína **1t2y** reconstruída (em azul) com a proteína original (em verde).

Capítulo 6

Conclusão e Trabalhos Futuros

Neste trabalho foram desenvolvidos três formas de determinação da estrutura de proteínas usando dados incertos:

Formas Afins Isoladamente, as formas afins não obtiveram êxito no controle da propagação da incerteza durante a construção das proteínas. Apesar de ter mais informações que as partículas, a propagação da incerteza de uma forma determinística aumenta a faixa de incerteza.

Partículas As partículas foram eficazes no controle da propagação de incerteza, o controles a elas aplicados limitaram o seu crescimento, permitindo a construção da maioria das proteínas testadas.

Método Híbrido O método híbrido permitiu que o controle aplicado às partículas fosse também utilizado na propagação da incerteza usando formas afins, assim as formas afins também foram capazes de construir as proteínas testadas com resultados similares aos resultados obtidos com as partículas.

A comparação dos resultados obtidos com outros métodos desenvolvidos não foi possível pois não existem outros métodos que usem dados incertos com a mesma abordagem que usamos.

Verificamos que o método híbrido, desenvolvido na tese, apresentou bons resultados para instâncias com até 200 átomos, contendo incertezas nos dados de entrada. Ou seja, a nova metodologia proposta pode ser incorporada a métodos que particionam a proteína em grupos menores, identificam suas estruturas, e fazem depois a junção dessas estruturas para identificar a estrutura global. Podemos citar, por exemplo, o método desenvolvido em [25].

6.1 Trabalhos Futuros

Melhorar a construção da proteína, incorporando no algoritmo outras informações estruturais, como com as restrições de ângulos de Ramachandran [23, 19] e minimização de energia [30], podem tornar os métodos desenvolvidos mais eficientes. Uma vez que estas informações sejam incorporadas aos algoritmos, também será viável testá-los com dados reais de RMN, que devem conter menos informações: o grafo deve ser menos denso e os dados podem conter erros experimentais.

Outra idéia seria melhorar o método de controle das formas afins usando o covariância ou otimização no espaço dos desconhecidos para alterar os desconhecidos de forma mais precisa.

Seria importante também criar uma biblioteca que disponibilize os métodos desenvolvidos, oferecendo transparência no uso. Com a implementação atual, as operações aritméticas e os controles precisam ser executadas explicitamente.

Referências Bibliográficas

[1] Dimitris K. Agrafiotis. Stochastic proximity embedding. *Journal of Computational Chemistry*, 24(10):1215–1221, 2003.

[2] Jim Albert. *Bayesian computation with R*. Springer, 2009.

[3] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*, volume 10. Garland Science Taylor & Francis Group, 2002.

[4] A.Y. Alfakih, A. Khandani, and H. Wolkowicz. Solving euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, 12:13–30, 1999.

[5] A. Aszodi, M.J. Gradwell, and W.R. Taylor. Global fold determination from a small number of distance restraints. *Journal of molecular biology*, 251:308–326, 1995.

[6] G.M. Crippen and T.F. Havel. *Distance Geometry and Molecular Conformation*. Wiley, 1988.

[7] Heather A. Damm, Kelly L.; Carlson. Gaussian-weighted RMSD superposition of proteins: A structural comparison for flexible proteins and predicted protein structures. *Biophysical journal*, 90:4558–4573, 2006.

[8] Robert T. Davis, Claus Ernst, and Di Wu. Protein structure determination via an efficient geometric build-up algorithm. *BMC structural biology*, 10 Suppl 1(Suppl 1):S7, January 2010.

[9] Luiz Henrique de Figueiredo and Jorge Stolfi. *Self-Validated Numerical Methods and Applications*. Brazilian Mathematics Colloquium monographs. IMPA/CNPq, Rio de Janeiro, Brazil, 1997.

[10] Luiz Henrique de Figueiredo and Jorge Stolfi. Affine Arithmetic: Concepts and Applications. *Numerical Algorithms*, 37(1-4):147–158, 2004.

[11] Z.W. Di Wu and Yaxiang Yuan. Rigid vs. Unique Determination of Protein Structures with Geometric Buildup. *Optimization Letters*, 2(3):319–331, 2007.

[12] Q. Dong and Z. Wu. A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. *Journal of Global Optimization*, 22(1):365–375, 2002.

[13] Qunfeng Dong and Zhijun Wu. An updated geometric build-up algorithm for solving the molecular distance geometry problems with sparse distance data. *Journal of Global Optimization*, 37(4):661–673, 2006.

[14] Jurgen F Doreleijers, Johan A C Rullmann, and Robert Kaptein. Quality Assessment of NMR Structures : a Statistical Survey. *J. Mol. Biology*, 281:149–164, 1998.

[15] J. Drenth. *Principles of Protein X-ray Crystallography*. Springer, 2007.

[16] Claire Fang Fang, Tsuhan Chen, and Rob A. Rutenbar. Floating-point error analysis based on affine arithmetic. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 561–564, 2003.

[17] Richard Feynman and Peter W. Shor. Simulating physics with computers. *International Journal of Theoretical Physics*, 21:467–486, 1982.

[18] A.V. Finkelstein and O.B. Ptitsyn. *Protein physics: a course of lectures*. Academic Pres, 2002.

[19] R.A. Friesner, I. Prigogine, and S.A. Rice. *Computational methods for protein folding*, volume 120 of *Advances in Chemical Physics*. Wiley-Interscience, 2002.

[20] David E Goldberg. What Every Computer Scientist Should Know About Floating-Point Arithmetic. *Computing Surveys*, 1991.

[21] Barry Gower. *Scientific Method: An historical and philosophical introduction*. Routledge, 1996.

[22] David Griffiths. *Introduction to Quantum Mechanics*. Addison-Wesley, 2004.

[23] K. Gunasekaran, C. Ramakrishnan, and P. Balaram. Disallowed Ramachandran conformations of amino acid residues in protein structures. *Journal of molecular biology*, 264:191–198, 1996.

[24] P. Guntert. Structure calculation of biological macromolecules from NMR data. *Quarterly reviews of biophysics*, 31(2):145–237, 1998.

[25] B. Hendrickson. The molecule problem: exploiting structure in global optimization. *SIAM Journal on Optimization*, 5:835–857, 1995.

[26] H.X. Huang, Z.A. Liang, and P.M. Pardalos. Some properties for the euclidean distance matrix and positive semidefinite matrix completion problems. *Journal of Global Optimization*, 25:3–21, 2003.

[27] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis*. Springer, 2001.

[28] Simon J Julier and Jeffrey K Uhlmann. Unscented Filtering and Nonlinear Estimation. *Computer Engineering*, 92(3), 2004.

[29] W. Kahan and E Darcy, J.D. How Java's Floating-Point Hurts Everyone Everywhere. In *ACM 1998 workshop on java for high-performance network computing*, pages 1–81, 1998.

[30] John L. Klepeis, Christodoulos A. Floudas, Dimitrios Morikis, and John D. Lambris. Predicting peptide structures using nmr data and deterministic global optimization. *Journal of Computational Chemistry*, 20(13):1354–1370, 1999.

[31] G.J. Klir. *Uncertainty and information: foundations of generalized information theory*. Wiley-IEEE Press, 2006.

[32] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. Recent advances on the discretizable molecular distance geometry problem. *European Journal of Operational Research*, 219:698–706, 2012.

[33] Frederic Messine. Extensions of Affine Arithmetic : Application to Unconstrained Global Optimization. *Journal of Universal Computer Science*, 8(11):992–1015, 2002.

[34] Shinya Miyajima and Masahide Kashiwagi. A dividing method utilizing the best multiplication in affine arithmetic. *IEICE Electronics Express*, 1(7):176–181, 2004.

[35] David Monniaux. The pitfalls of verifying floating-point computations. *ACM Transactions on Programming Languages and*, pages 1–40, 2008.

[36] Ramon E Moore. *Interval Analysis*. Prentice-Hall, 1966.

[37] J.J. Moré and Z. Wu. ε-optimal solutions to distance geometry problems via global e continuation. pages 151–168, 1996.

[38] Aart J. Nederveen, Jurgen F. Doreleijers, Wim Vranken, Zachary Miller, Chris A. E. M. Spronk, Sander B. Nabuurs, Peter Gu, Miron Livny, John L. Markley, Michael Nilges, Eldon L. Ulrich, Robert Kaptein, and Alexandre M. J. J. Bonvin. RECOORD : A Recalculated Coordinate Database of 500+ Proteins from the PDB Using Restraints from the BioMagResBank. *Structure*, 672(October 2004):662– 672, 2005.

[39] Daniel John. Rigden. *From protein structure to function with bioinformatics*. Springer, 2008.

[40] J.B. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. *Proceedings of the 17th Allerton Conference on Communication, Control, and Computing*, pages 480–489, 1979.

[41] Richard Szeliski. *Computer vision: Algorithms and applications*. Springer, 2010.

[42] M. Tagari, J. Tate, G. J. Swaminathan, R. Newman, A. Naim, W. Vranken, A. Kapopoulou, A. Hussain, J. Fillon, K. Henrick, and S. Velankar. E-MSD : improving data deposition and structure quality. *Database*, 34:287–290, 2006.

[43] Quincy Teng. Structural biology. Practical NMR applications. *Science (New York, N.Y.)*, 334(6054):320–1, October 2011.

[44] Ahmed Touhami. A general reliable quadratic form: An extension of affine arithmetic. *Reliable computing*, pages 171–192, 2006.

[45] Ron Unger. The Genetic Algorithm Approach to Protein Structure Prediction. *Structure and Bonding*, 110:153–175, 2004.

[46] Wim Vranken. A global analysis of NMR distance constraints from the PDB. *Journal of Biomolecular NMR*, pages 303–314, 2007.

[47] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics)*. Springer, December 2003.

[48] Larry Wasserman. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag, 2006.

[49] D.M. Webster. *Protein Structure Prediction: Methods and Protocols*. Methods in Molecular Biology. Humana Press, 2000.