COPPE/UFRJ

SUAVIZAÇÃO HIPERBÓLICA APLICADA À OTIMIZAÇÃO DE GEOMETRIA MOLECULAR

Michael Ferreira de Souza

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientadores: Nelson Maculan Filho

Carlile Campos Lavor

Rio de Janeiro

Janeiro de 2010

SUAVIZAÇÃO HIPERBÓLICA APLICADA À OTIMIZAÇÃO DE GEOMETRIA MOLECULAR

Michael Ferreira de Souza

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Nelson Maculan Filho, D.Sc.

Prof. Carlile Campos Lavor, D.Sc.

Prof. Adilson Elias Xavier, D.Sc.

Prof[a]. Márcia Helena Costa Fampa, D.Sc.

Prof. Aurélio Ribeiro Leite de Oliveira, D.Sc.

Prof. Fábio Protti, D.Sc.

RIO DE JANEIRO – RJ, BRASIL

JANEIRO DE 2010

Souza, Michael Ferreira de

Suavização Hiperbólica Aplicada à Otimização de Geometria Molecular/ Michael Ferreira de Souza. – Rio de Janeiro: UFRJ/COPPE, 2010.

XXII, 86 p.: il.; 29,7 cm.

Orientadores: Nelson Maculan Filho

Carlile Campos Lavor

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2010.

Referencias Bibliográficas: p. 73-86.

1. Geometria molecular. 2. Suavização. 3. Otimização. I. Maculan Filho, Nelson et. al.. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Titulo.

À minha Carol, essa última década ao seu lado foi fantástica. Já anseio pelas próximas.

Agradecimentos

Agradecer é uma tarefa gratificante, mas de imenso risco. Ao agradecer, arriscamos esquecer das mais diversas contribuições, dos pequenos gestos. Sendo assim, agradeço inicialmente a todos os que eu deixarei de citar, quer por omissão, falta de espaço ou esquecimento.

Agradeço ao professor Nelson Maculan pela cândura e diligência. E pela grande facilidade de tornar as coisas simples. E estendendo o agradecimento, não poderia esquecer das suas fiéis escudeiras Maria de Fátima Cruz Marques e Josefina Solange Silva Santos que não me deixaram sucumbir na grandiosa burocracia que está ao nosso redor procurando a quem possa tragar.

Agradeço encarecidamente ao professor Adilson Elias Xavier pelas incontáveis horas de conversa e ensinamento sobre os mais diversos assuntos, indo do sagrado ao profano. Mesmo não ocupando formalmente o papel de orientador deste trabalho, foi um colaborador contínuo e sem ele, este trabalho não teria tomado a forma nem o conteúdo presentes. Obrigado pela amizade.

Agradeço aos amigos que tornaram o período do doutorado uma agradável lembrança. Em especial, aos amigos Jurair, Alberto, Jesus, Thiago, Francisco e Marcelo Dibb.

Agradeço ao professor Carlile Campos Lavor, que deste minha iniciação-científica ainda na UERJ, passando pelo mestrado, e, agora, no doutorado, tem sido uma presença constante e um grande parceiro. Sempre atencioso e correto nos mais diferentes cenários.

Agradeço a minha esposa Carol pela paciência e amor gratuito e, às vezes, imerecido, só ela sabe como eu posso ser chato!

Agradeço aos meus pais e irmãos pelas maiores e mais importantes lições que já tive e pelo suporte inconteste e irrestrito. Meu porto seguro.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

SUAVIZAÇÃO HIPERBÓLICA APLICADA À OTIMIZAÇÃO DE GEOMETRIA MOLECULAR

Michael Ferreira de Souza

Janeiro/2010

Orientadores: Nelson Maculan Filho
Carlile Campos Lavor
Programa: Engenharia de Sistemas e Computação

A determinação de estruturas tridimensionais de proteínas é um dos grandes desafios da biologia moderna. No presente trabalho, abordamos o problema da determinação de estruturas tridimensionais, a partir de algumas das distâncias entre pares de pontos que as compõem. Este problema está fortemente relacionado à determinação da conformação proteica via ressonância magnética nuclear, onde apenas um subconjunto das distâncias entre pares de átomos é conhecido. As usuais formulações utilizadas para esse problema são NP-difíceis, não-diferenciáveis e não-convexas, possuindo um elevado número de mínimos. A contribuição deste trabalho é um método especializado que combina suavização e penalização hiperbólicas, para obtenção de diferenciabilidade e convexificação, com uma estratégia de dividir-para-conquistar, para escalabilidade.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

HYPERBOLIC SMOOTHING APPLIED TO MOLECULAR GEOMETRY OPTIMIZATION

Michael Ferreira de Souza

January/2010

Advisors: Nelson Maculan Filho
Carlile Campos Lavor
Department: Systems Engineering and Computer Science

The determination of three-dimensional protein structures is a major challenge in modern biology. In the present work, we consider the problem of estimating relative positions of all points in a structure, given a subset of all the pair-wise distances between a set of its points. This problem is related to the protein folding determination via nuclear magnetic resonance, where only a subset of all pair-wise distance between atoms are available. The usual formulations to this problem are NP-hard, nonsmoothed and nonconvex, having a high number of local minima. The contribution of this work is a specialized method that combines hyperbolic smoothing and penalty in order to obtain differentiability and a specific divide-and-conquer strategy to get scalability.

Sumário

Lista de Figuras **x**

Lista de Tabelas **xiii**

1 Introdução **1**

2 Conceitos básicos sobre proteínas **6**

2.1 Estrutura química 7

2.2 Geometria local 12

2.3 Métodos experimentais para determinação da conformação protéica . 15

3 O problema geométrico da distância molecular **19**

3.1 Problemas relacionados 24

3.2 Comparação de estruturas via RMSD 25

3.3 MDGP via programação matemática 26

3.4 Estratégias de solução 27

3.4.1 EMBED 28

3.4.2 ABBIE 28

3.4.3 Perturbação estocástica 29

3.4.4 DGSOL 30

3.4.5 Otimização de diferenças de funções convexas 31

3.4.6 Algoritmo de construção geométrica 33

3.4.7 Algoritmo branch-and-prune 33

3.4.8 Programação semidefinida 35

3.4.9 GNOMAD 36

4 MDGP via suavização e penalização hiperbólicas 38

4.1 Proposta de suavização 39

4.2 Proposta de penalização 40

4.3 Convexificação 43

4.4 O algoritmo de suavização e penalização hiperbólicas (SPH) 47

4.5 Aliando o algoritmo SPH e a técnica de dividir-para-conquistar . . . 48

4.5.1 Combinando soluções 51

4.5.2 Dividindo problemas 53

4.5.3 O método *sphdc* 57

5 Experimentos computacionais 59

5.1 Validando o procedimento *sph* 60

5.1.1 Redução de mínimos de locais 61

5.1.2 Robustez 62

5.1.3 Coerência das soluções 63

5.2 Validando o procedimento *sphdc* 64

5.3 Experimentos mais complexos 66

6 Conclusão e propostas para trabalhos futuros 70

Referências Bibliográficas 73

Lista de Figuras

2.1 Estrutura química dos aminoácidos. 8

2.2 Processo de formação da ligação peptídica das proteínas. 9

2.3 Estrutura química dos aminoácidos. 9

2.4 Proteína *myohemerythrin* (2MHR) formada por 118 resíduos e quatro hélices. À direita, a visão simplificada e, à esquerda, a estrutura da cadeia principal. 10

2.5 Proteína *fibronectin* (1TTG) formada por 94 resíduos e quatro hélices. À direita, a visão simplificada e, à esquerda, a estrutura da cadeia principal. 10

2.6 Proteína *kinase C* (1PTQ) formada por 402 átomos. 11

2.7 Complexo *gvp-ssdna* (1GPV) formado por 1842 átomos e 182 resíduos e 4 cadeias (subunidades). As subunidades estão respresentadas com cores distintas. 11

2.8 Estrutura tridimensional local. 12

2.9 Ângulo diedral γ, ângulos das ligações $\theta_{ijk}, \theta_{jkl}$, vetores de ligação r_{ij}, r_{jk}, r_{kl} e vetores de posição x_i, x_j, x_k, x_l. 13

2.10 Ângulos diedrais da cadeia prinicipal de uma proteína. 13

2.11 Nestas configurações, todos os quatro átomos estão situados no mesmo plano. À esquerda, a forma *cis* ($\omega = 0^{\circ}$). À direita, a forma *trans* com a distância entre os átomos C_1^{α} e C_2^{α} sendo a maior possível. . . . 14

2.12 Ângulos das torções do resíduo lisina representados pelas variáveis χ_1 a χ_4. 14

2.13 À direita, vê-se o mapa 3D da densidade eletrônica em cada ponto do espaço. À esquerda, sendo conhecidas a sequênca e a forma dos aminoácidos que compõem a proteína, pode-se ajustar o modelo protéico a esse mapa. 16
2.14 Número de estruturas disponíveis na base de dados PDB entre 1972 e março de 2009. 18

3.1 Equivalência do particionamento de conjuntos e o MDGP unidimensional. 24
3.2 Definições dos comprimentos, ângulos das ligações e os ângulos de torção. 34

4.1 O gráfico da suavização hiperpólica $\theta_\tau(x)$ é uma hipérbole equilátera. 40
4.2 No gráfico da função de penalidade $\phi_{\lambda,\tau}$, o parâmetro λ controla a intensidade (inclinação) da penalização, e τ, a suavização. 41
4.3 À medida que aumentamos o valor do parâmetro τ, mais convexa torna-se a função objetivo. 46
4.4 (a) A matriz de conectividade da instância derivada da geometria conhecida da proteína 1PTQ. (b) Na região destacada, observa-se blocos com alta densidade sobre a diagonal da matriz de conectividade. 50
4.5 Os blocos destacados estão relacionados às restrições envolvendo átomos no mesmo resíduo. Estes blocos são os mais densos nas matrizes de conectividade tipicamente encontradas na literatura. 50
4.6 Na divisão binária, os grupos são divididos recursivamente até que existam apenas dois resíduos em cada grupo. 54
4.7 Na figura à esquerda, vemos a matriz de conectividade dos átomos. À direita, a matriz de conectividade dos resíduos. 55
4.8 Na figura à esquerda, vemos a matriz de conectividade dos átomos. À direita, a matriz de conectividade dos resíduos. 55

4.9 Na figura à esquerda, vemos o grafo obtido aplicando a árvore geradora máxima e, à direita, o grafo obtido pela divisão binária. A divisão com a árvore geradora máxima produz um conjunto de subproblemas (arestas) com maior acoplamento e, consequentemete, menor número de soluções. 57

4.10 Inicialmente, o problema original é dividido. Em seguida, a rotina *sph* é aplicada a cada um dos subproblemas. As soluções X e Y são combinadas em uma estrutura Z pela rotina *combinar*, utilizando a medida *rmsd* como critério de coerência da combinação das soluções. 58

5.1 Os valores de f nas soluções geradas pela rotina $va35$ e nas soluções geradas pela rotina *sph*. 61

5.2 Os erros máximos presentes nas soluções geradas pela rotina $va35$ e *sph*. Os erros associados às soluções obtidas pela rotina sph foram consideravelmente menores do que os obtidos pela rotina $va35$ aplicada individualmente. 62

5.3 À direita, a matriz de conectividade dos resíduos da instância 8DRH. Cada um dos quadrados em vermelho identificam os arcos associados às restrições envolvendo átomos átomos no mesmo resíduo. À esquerda, a matriz de conectividade dos resíduos, note que não há correlação (arcos) entre resíduos que não sejam vizinhos. 68

5.4 As matrizes de conectividade da instância 1PTQ com $C = 8$Å. As instâncias da forma dada na Eq. 5.8 são mais complexas que as idealizadas por Moré e Wu. 68

Lista de Tabelas

2.1 Nomes e siglas dos aminoácidos encontrados nas células vivas. Os nomes e siglas dos aminoácidos essenciais estão em negrito. 8

5.1 Frequência de obtenção de soluções pelos métodos *dgsol*, *va*35 e *sph*. 63

5.2 Desvio (RMSD) entre as coordenadas das soluções obtidas pelos métodos *dgsol* e *sph* e as coordenadas originais do fragmento com 100 átomos. 64

5.3 Desvio (RMSD) entre as coordenadas das soluções obtidas pelos métodos *dgsol* e *sph* e as coordenadas originais do fragmento com 200 átomos. 64

5.4 Performance dos métodos *sphdc* e *gcdca* com dados do PDB com $\varepsilon = 0,001$. O número de átomos e de resíduos são representados, respectivamente, por m e n e o tempo é dado em segundos. 66

5.5 Performance dos métodos *sphdc* e *gcdca* com dados do PDB com $\varepsilon = 0,08$. O número de átomos e de resíduos são representados, respectivamente, por m e n e o tempo é dado em segundos. 67

5.6 Performance dos método *sphdc* com dados do PDB em instâncias da forma dada na Eq. (5.8) com $C = 8$Å. 69

5.7 Performance dos método *sphdc* com dados do PDB em instâncias da forma dada na Eq. (5.8) com $C = 6$Å. 69

Capítulo 1

Introdução

> "A grande celebração da conclusão do mapa do genoma humano é um raro dia na história da ciência, um dia em que um evento de significância histórica é reconhecido não em restrospectiva, mas enquanto ele acontece ... Ainda que este dia mereça a atenção de toda a humanidade, não devemos confundir progresso com solução. Existe ainda muito trabalho a ser feito. Levará muitas décadas até que consigamos comprender totalmente a magnificência do edifício DNA construído sobre quatro bilhões de anos de evolução e escondido no núcleo de cada célula do corpo de cada organismo na Terra."
>
> David Baltimore, *The New York Times*, 25 de Junho de 2000.

Em 2001, dois grupos concorrentes, o consórcio internacional *Human Genome Project* e a empresa americana Celera anunciaram que conseguiram, pela primeira vez na história da humanidade, mapear o genoma humano e estabelecer sua sequência [96, 68]. O que os cientistas fizeram foi decifrar 3,1 bilhões de bases químicas (nucleotídeos) do DNA presentes no genoma humano [56].

Como a maioria dos aspectos da saúde humana, sejam eles positivos ou não, é influenciada/determinada pelas interações entre o DNA e os fatores ambientais, em tese, no futuro, será possível realizar considerável progresso no diagnóstico, tratamento e prevenção de doenças importantes com base no mapa do genoma humano. No entanto, a identificação das bases do DNA é apenas o primeiro passo. Resta ainda a tarefa muito mais complicada de decifrar o significado de cada base, sua função

e o que pode ser feito no caso de trazerem mensagens defeituosas, que resultem em doenças.

Sabe-se que o DNA apresenta pouca mobilidade, restrigindo-se ao interior do núcleo celular. Portanto, sua ação na determinação das características hereditárias é feita indiretamente. Em um processo denominado *transcrição*, o DNA presente no núcleo celular induz à formação do RNA mensageiro que migra para o citoplasma celular e se liga a um ribossomo. Juntos, RNA e ribossomo, iniciam a *tradução*, i.e., o processo de ordenação e ligação dos aminoácidos que formarão as proteínas. Serão as proteínas que atuarão diretamente não só na determinação das características hereditárias, mas também nas mais variadas funções nos organismos, desde o transporte de nutrientes e metabólitos, catálise de reações biológicas até a composição estrutural das células. Grosso modo, o genoma, o conjunto completo da informação genética, contém somente a receita para fabricação de proteínas, enquanto que as proteínas desempenham o papel de cimento e tijolos das células e realizam a maior parte do trabalho. Assim, a compreensão do real significado do mapeamento do genoma humano e suas possíveis aplicações estão profundamente ligados ao entendimento do papel desempenhado pelas proteínas.

Infelizmente, o proteoma, i.e., o conjunto de todas as proteínas produzidas por uma dada célula, tecido ou organismo, é muito mais complicado que o genoma [42]. O alfabeto do DNA é composto por quatro bases químicas conhecias por suas iniciais: adenina (A), citosina (C), guanina (G) e timina (T). As proteínas, no entanto, são formadas pela combinação de 20 blocos fundamentais denominados *aminoácidos*. Os genes especificam os aminoácidos que devem se combinar para formar uma dada proteína. Mas, mesmo quando a sequência de aminoácidos de uma proteína é conhecida, não se sabe ao certo determinar a função da proteína e a que outras proteínas ela pode se associar. Diferentemente dos genes, que são lineares, as proteínas assumem formas curvas que, em alguns casos, desafiam a predição e estão diretamente ligadas às funções desempenhadas pela proteína [16, 42, 27].

Além disso, as células normalmente modificam as proteínas pela adição de açúcar e gordura de uma forma que também pode ser difícil antecipar. Por isso, para produzir uma proteína codificada por um gene, não basta formar a sequência de aminoácidos ditada pelo gene, é também necessário realizar as corretas modificações

pelo acréscimo de açúcar e gordura. E para determinar o comportamento/funcionamento da proteína, é preciso ainda considerar o ambiente, água, óleo, etc, em que a proteína atua.

Um grande volume de recursos tem sido aplicado no estudo do mapa tridimensional das moléculas, mais especificamente, das proteínas [17, 97]. A criação de bases de dados de estruturas protéicas como, por exemplo, *Protein Data Bank* [10], fornece a possibilidade de detecção de homologias[1] entre diferentes proteínas que eventualmente não seriam percebidas simplesmente pela comparação das sequências de aminoácidos que as compõem. Ao catalogar as estruturas tridimensionais básicas das proteínas, cria-se a possibillidade de detecção de famílias de proteínas com características similares [11]. Essas estruturas são fundamentais na determinação dos mecanismos e funções das proteínas e podem ser utilizadas na redução dos custos de desenvolvimento e teste de medicamentos (Andrew Pollack, "Drug Testers Turn to 'Virtual Patients' as Guinea Pigs", 10 de novembro de 1998, *New York Times*).

Até 1984, informação estrutural em resolução atômica só poderia ser determinada via técnicas de difração de raio-X com unidades de proteínas cristalizadas [38]. A introdução da ressonância magnética nuclear (RMN) como uma técnica para a determinação da estrutura protéica tornou possível a obtenção de estruturas com elevada precisão em um ambiente (solução) muito mais próximo da situação natural de um organismo vivo do que os cristais utilizados na cristalografia [1, 107, 54, 89, 11].

Os experimentos de RMN baseiam-se no fato de que os núcleos de hidrogênio têm dois estados (spins) que podem ser alterados pelo fornecimento de energia em uma dada frequência. A informação estrutural vem do acoplamento spin-spin entre os núcleos de hidrogênio. Se dois núcleos estão espacialmente próximos, então seus spins interagem e a frequência necessária para alterar um spin é modificada. Os picos no espectro tornam-se ligeiramente alterados, o que torna possível a inferência não só de distâncias envolvendo pares de átomos de hidrogênio espacialmente próximos, com distância inferior a 5-6 Å (1 Å $= 10^{-10}$ m), mas também de ângulos entre átomos em uma dada proteína [54, 55, 113]. Para calcular a estrutura tridimensional da macromolécula, essas distâncias são usadas como restrições em combinação com diversas informações suplementares, tais como: a sequência de aminoácidos que

[1]Homologia: semelhança de origem e estrutura.

compõem a proteína, referências geométricas para o comprimento e os ângulos das ligações químicas existentes, entre outras. Consideráveis recursos computacionais são requeridos para analisar sistematicamente a informação produzida via RMN.

No presente trabalho, abordaremos um dos problemas relacionados à determinação da estrutura tridimensional das proteínas via RMN, mais especificamente, versaremos sobre o *problema geométrico da distância molecular* (MDGP, do inglês *molecular distance geometry problem*), onde o objetivo é determinar uma estrutura tridimensional que seja compatível com os dados (distâncias átomo-átomo) provenientes dos experimentos com RMN.

Na verdade, o estudo de formas de inferência de estruturas a partir de distâncias é um tema importante que vem aumentando seu número de aplicações, seja na predição de estruturas moleculares [11, 79, 30], estimação de posição em redes sem fio [12, 26, 90], visualização de informação [46, 84], tomografia da internet [23] ou reconstrução de mapas [33]. Mais recentemente, esta teoria tem sido aplicada no reconhecimento de face [62] e segmentação de imagem [102]. Segundo Biswas em [11], a questão essencial nesses problemas é, dado um conjunto incompleto e impreciso de distâncias euclideanas entre uma rede de pontos (em uma dada dimensão), podemos obter algoritmos robustos, eficazes e escaláveis para encontrar suas posições relativas?

A contribuição deste trabalho é um novo algoritmo que combina a técnica de suavização e penalização hiperbólicas e uma especializada estratégia de dividir-para-conquistar para solução do problema geométrico da distância molecular. As técnicas de suavização e penalização hiperbólicas, propostas por Xavier em [108], permitem a aplicação de métodos clássicos de otimização ao introduzirem diferenciabilidade na formulação do MDGP como um problema de programação matemática e, mais importante ainda, reduzem o número de mínimos locais pelo controle adequado dos parâmetros relacionados. Já a estratégia de dividir-para-conquistar permite que, ao invés de resolver um único e grande problema, possamos atacar uma sequência de problemas menores e, por isso, mais fáceis. Detalhes da implementação e resultados de experimentos numéricos com dados provenientes do *Protein Data Bank* são apresentados.

O capítulo 2 introduz conceitos básicos sobre estrutura química das proteínas,

suas representações geométricas e faz um apanhando dos principais métodos experimentais para determinação da estrutura proteica.

No capítulo 3, o problema geométrico da distância molecular (MDGP) é formalmente definido. Alguns aspectos históricos são explorados, nuances sobre a complexidade do MDGP são ressaltadas e diferentes abordagens encontradas na literatura são apresentadas.

Ao longo do capítulo 4, apresentamos a proposta de suavização e penalização hiperbólicas para solução dos problemas de mínimos quadrados relacionados e um algoritmo baseado na técnica de divididir-para-conquistar visando à resolução de instâncias do MDGP com elevado número de átomos.

O capítulo 5 é reservado aos experimentos computacionais realizados com instâncias geradas a partir de proteínas reais. Os resultados são comparados aos encontrados na literatura.

Finalmente, no capítulo 6, propomos caminhos para trabalhos futuros e sintetizamos as contribuições do presente trabalho.

Capítulo 2

Conceitos básicos sobre proteínas

> "As proteínas são as máquinas e tijolos das células. Se nós compararmos um organismo com o mundo, cada célula corresponderá a uma cidade, e as proteínas serão as casas, pontes, carros, guindastes, estradas, aeroportos, etc."
>
> Arnold Neumaier, em [82].

A história das proteínas começa no século XVIII, com a descoberta de que certos componentes do mundo vivo, como a clara de ovo (albúmen), o sangue, o leite, entre outros, coagulam em altas temperaturas e em meio ácido. Substâncias com esse tipo de comportamento foram denominadas albuminóides (semelhante ao albúmen).

No início do século XIX, descobriu-se que os principais constituintes das células vivas eram substâncias albuminóides. Em um artigo publicado em 1838, o químico holândes Gerardus Johannes Mulder (1802-1880) usou, pela primeira vez, o termo *proteína* (do grego *proteios*, primeiro, primitivo) para se referir às substâncias albuminóides. Na verdade, foi o sueco Jöns Jacob Berzelius (1779-1848), um dos mais importantes químicos da época, quem sugeriu o termo a Mulder, por acreditar que as substâncias albuminóides eram os constituintes fundamentais de todos os seres vivos.

Na virada para o século XX, o interesse pelas proteínas continuava a crescer. Os químicos passaram a analisar minuciosamente essas substâncias, descobrindo que a sua degradação liberava *aminoácidos*. Por volta de 1900, já haviam sido identificados 12 aminoácidos diferentes liberados pela degradação de proteínas. Face

a essa evidência, o químico alemão Franz Hofmeister (1850-1922) sugeriu, em 1902, que os proteínas seriam formadas por aminoácidos encadeados.

Em 1906, já haviam sido identificados 15 tipos de aminoácidos liberados pela degradação de proteínas; em 1935, esse número subiu para 18 e, em 1940, chegou a 20, completando a lista de aminoácidos que ocorrem naturalmente nas proteínas dos seres vivos [4].

A maioria das proteínas naturais adota estruturas tridimensionais específicas que estão associadas as suas atividades biológicas. Apesar de dinâmica, sobre condições térmicas e configurações locais típicas, a estrutura tridimensional de cada proteína apresenta pequenas variações. Uma das grandes descobertas sobre a estrutura biomolecular é a relação determinística entre a sequência de aminoácidos e a estrutura tridimensional da proteína [89]. Isto foi apontado pela primeira vez por Christian B. Anfinsen e colaboradores, no início da década de 1960 [7]. Anfinsen compartilhou o prêmio Nobel em química de 1972 com Stanford Moore e William H. Stein, por seus trabalhos sobre ribonuclease, conectando a sequência de aminoácidos à conformação biologicamente ativa.

Fundado em 1971 pelos doutores Edgar Meyer e Walter Hamilton, o banco de dados PDB (*Protein Data Bank*) é um repositório para estruturas tridimensionais de proteínas e aminoácidos [10]. Os dados encontrados no PDB são frutos de experimentos de RMN e Raio-X, ou de desenvolvimento teórico realizados por pesquisadores de diferentes partes do mundo e podem ser gratuitamente acessados. Ao longo dos anos, o PDB tem se transformado em uma importante fonte de dados para o avanço e divulgação do conhecimento sobre as proteínas.

2.1 Estrutura química

Do ponto de vista puramente químico, uma proteína é simplesmente uma longa cadeia de aminoácidos unidos por ligações peptídicas, daí as proteínas também serem denominadas polipeptídeos. Cada aminoácido (exceto a prolina) é formado por um carbono central, conhecido como carbono alfa (C^{α}), ao qual estão ligadas quatro unidades: um átomo de hidrogênio, um grupo amina (NH_3^+), um grupo carboxílico (COO^-) e uma característica cadeia lateral, ou grupo R (ver Figura 2.1).

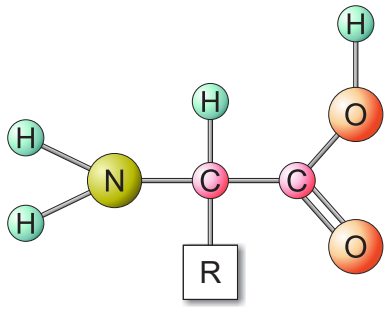

Figura 2.1: Estrutura química dos aminoácidos.

Embora sejam inúmeras, as proteínas das células vivas são formadas por um "alfabeto" de apenas 20 aminoácidos, que se repetem numa sequência específica para cada proteína. Nove dos vinte aminoácidos existentes não são sintetizados pelos seres humanos e, por isso, precisam ser incluídos em sua dieta, estes aminoácidos são denominados *aminoácidos essenciais*. Os grupos R (resíduos) são usualmente identificados pelas três letras iniciais dos nomes dos aminoácidos dos quais eles derivam (ver Tabela 2.1).

Ala	-	Alanina	Arg	-	Arginina
Asn	-	Asparagina	Asp	-	Aspartato
Cys	-	Cisteína	Gln	-	Glutamina
Glu	-	Glutamato	Gly	-	Glycine
His	-	**Histidina**	**Ile**	-	**Isoleucina**
Leu	-	**Leucina**	**Lys**	-	**Lisina**
Met	-	**Metionina**	**Phe**	-	**Fenilalanina**
Pro	-	Prolina	Ser	-	Serina
Thr	-	**Treonin**	**Trp**	-	**Tripofano**
Tyr	-	Tirosina	**Val**	-	**Valina**

Tabela 2.1: Nomes e siglas dos aminoácidos encontrados nas células vivas. Os nomes e siglas dos aminoácidos essenciais estão em negrito.

Durante a formação das proteínas, sobre a influência da informação genética contida no RNA, o grupo carboxílico de um aminoácido se une ao grupo amina de outro aminoácido formando uma ligação peptídica ($C-N$) e liberando uma molécula de água (ver Figura 2.2). A forma geral de uma proteína é a repetição da estrutura exibida na Figura 2.3.

A cadeia formada pela repetição da sequência "$-NC^{\alpha}C-$" é conhecida como *cadeia principal*. Apesar da forma linear sugerida pela Figura 2.3, forças interatômicas encurvam e torcem a estrutura protéica produzindo uma configuração tridimensional

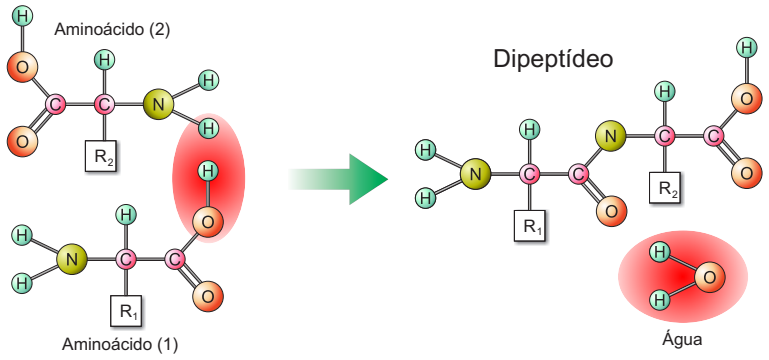

Figura 2.2: Processo de formação da ligação peptídica das proteínas.

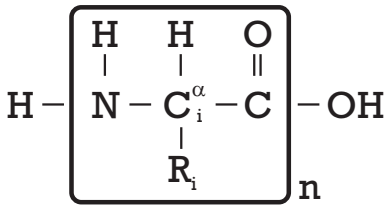

Figura 2.3: Estrutura química dos aminoácidos.

característica para cada proteína. Essa configuração e grupos quimicamente ativos na superfície das proteínas são os fatores que determinam as funções biológicas por elas desempenhadas.

Existem proteínas de diversos tamanhos, algumas realmente grandes, como a proteína muscular *titin*, com cerca de 27000 aminoácidos e massa de 3000 kDa[1], e outras pequenas, como a *trypsin* (inibidora da pancreatite bovina), formada por 58 aminoácidos. Mas, em média, as proteínas são formadas por algumas centenas de aminoácidos. Portanto, a quantidade de átomos envolvidos varia de poucas centenas a algumas centenas de milhares. O tamanho dos polipeptídeos pode ser determinado via experimentos com *gel electrophoresis*, pois a taxa de migração da molécula é inversamente proporcional ao logaritmo de seu comprimento. Assim, a massa de um polipetídeo ou proteína pode ser estimada pelas relações mobilidade-massa estabelecidas por proteínas de referência e pelas medições de espectometria de massa. Uma outra técnica para determinação de várias características macromoleculares, incluindo peso, baseada nas propriedades de transporte é a *equilibrium ultracentrifugation* [19].

Quatro níveis de abstração são utilizados para a descrição de estruturas protéicas:

[1]Da, ou Dalton, é uma unidade de medida de massa utilizada para expressar a massa de partículas atômicas. Ela é definida como 1/12 da massa de um átomo de carbono-12 em seu estado fundamental.

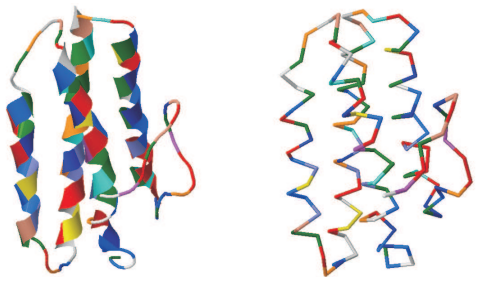

Figura 2.4: Proteína *myohemerythrin* (2MHR) formada por 118 resíduos e quatro hélices. À direita, a visão simplificada e, à esquerda, a estrutura da cadeia principal.

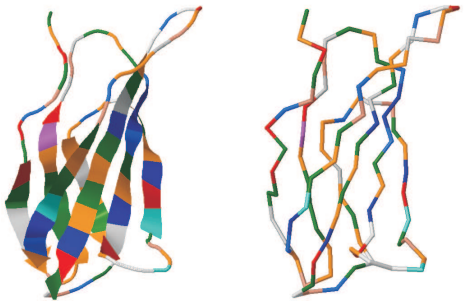

Figura 2.5: Proteína *fibronectin* (1TTG) formada por 94 resíduos e quatro hélices. À direita, a visão simplificada e, à esquerda, a estrutura da cadeia principal.

- estrutura primária: a sequência de resíduos na cadeia polipeptídica;
- estrutura secundária: padrões estruturais locais tais como α-hélices (ver Figura 2.4) e β-folhas (ver Figura 2.5), ou combinações destes padrões;
- estrutura terciária: o arranjo tridimensional de todos os átomos da cadeia polipeptídica (ver Figura 2.6);
- estrutura quaternária (utilizada para grandes proteínas com subunidades independentes): a rede tridimensional completa de interações entre as diferentes subunidades. A estrutura quaternária descreve a organização espacial das subunidades (ver Figura 2.7).

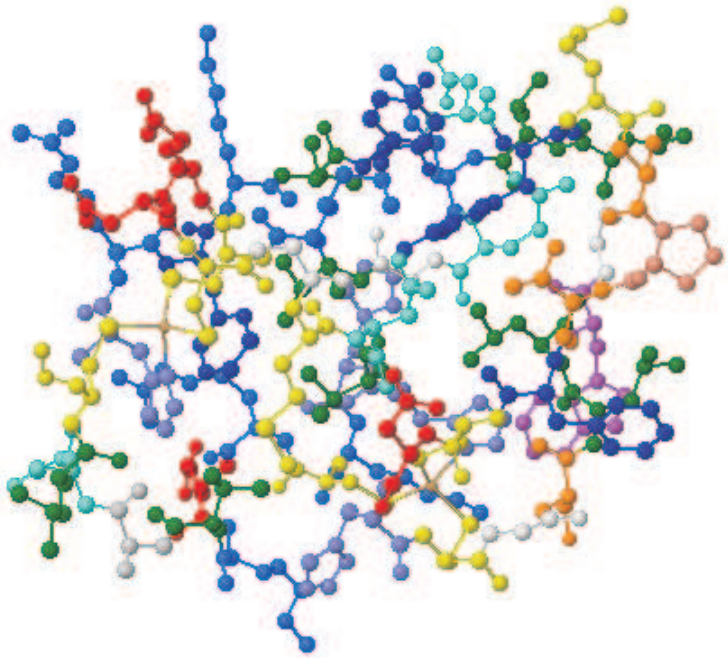

Figura 2.6: Proteína *kinase C* (1PTQ) formada por 402 átomos.

Figura 2.7: Complexo *gvp-ssdna* (1GPV) formado por 1842 átomos e 182 resíduos e 4 cadeias (subunidades). As subunidades estão respresentadas com cores distintas.

2.2 Geometria local

A geometria de uma proteína pode ser matematicamente representada atribuindo-se ao i-ésimo átomo que a compõe um vetor tridimensional

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix}, \tag{2.1}$$

que especifique a posição desse átomo no espaço. Podemos representar uma eventual ligação química entre os átomos i e j pelo vetor

$$r_{ij} = x_j - x_i, \tag{2.2}$$

e, neste caso, o comprimento da ligação é dado por

$$\|r\| = \sqrt{\langle r, r \rangle}, \tag{2.3}$$

onde

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3 \tag{2.4}$$

é o produto interno canônico em $\mathbb{R}^3$.

Assumindo que os átomos i, j e k estejam quimicamente unidos da forma representada na Figura 2.8, podemos definir os vetores

$$r_{ij} = x_j - x_i, \quad r_{jk} = x_k - x_j. \tag{2.5}$$

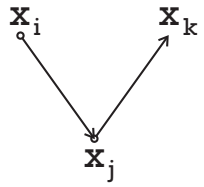

Figura 2.8: Estrutura tridimensional local.

A partir da definição dos vetores r_{ij}, r_{jk}, podemos calcular o ângulo θ_{ijk} pelas expressões

$$\cos\theta_{ijk} = \frac{\langle r_{ij}, r_{jk} \rangle}{\|r_{ij}\| \, \|r_{jk}\|}, \quad \sin\theta_{ijk} = \frac{\|r_{ij} \times r_{jk}\|}{\|r_{ij}\| \, \|r_{jk}\|}. \tag{2.6}$$

Finalmente, o ângulo diedral $\gamma_{ijkl} \in [-180°, 180°]$ é definido como o ângulo entre os vetores normais aos planos definidos pelos átomos i, j, k e j, k, l (ver Figura 2.9). O ângulo diedral γ pode ser calculado pelas fórmulas

$$\cos\gamma_{ijkl} = \frac{\langle r_{ij} \times r_{jk}, r_{jk} \times r_{kl} \rangle}{\|r_{ij} \times r_{jk}\| \, \|r_{jk} \times r_{kl}\|}, \quad \sin\gamma_{ijkl} = \frac{\langle r_{ij}, r_{jk} \times r_{kl} \rangle \, \|r_{jk}\|}{\|r_{ij} \times r_{jk}\| \, \|r_{jk} \times r_{kl}\|}. \tag{2.7}$$

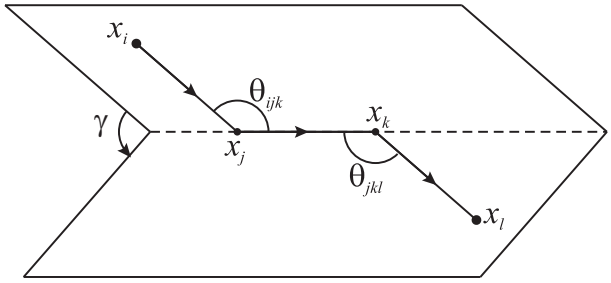

Figura 2.9: Ângulo diedral γ, ângulos das ligações $\theta_{ijk}, \theta_{jkl}$, vetores de ligação r_{ij}, r_{jk}, r_{kl} e vetores de posição x_i, x_j, x_k, x_l.

Um conjunto completo de vetores de ligação, ângulos de ligação e ângulos diedrais caracteriza completamente a geometria de uma molécula (na verdade, a superdetermina, i.e., fornece mais informação do que o mínimo necessário para a completa caracterização). Em uma proteína, os ângulos de ligação são frequentemente representados pela letra θ e os ângulos diedrais que descrevem torções ao redor das ligações $N - C^\alpha$, $C^\alpha - C$ e $C - N$ na cadeia principal, são representados pelas letras φ, ψ e ω respectivamente (ver Figura 2.10). Os ângulos diedrais nas cadeias laterais são descritos pela letra χ (ver Figura 2.12).

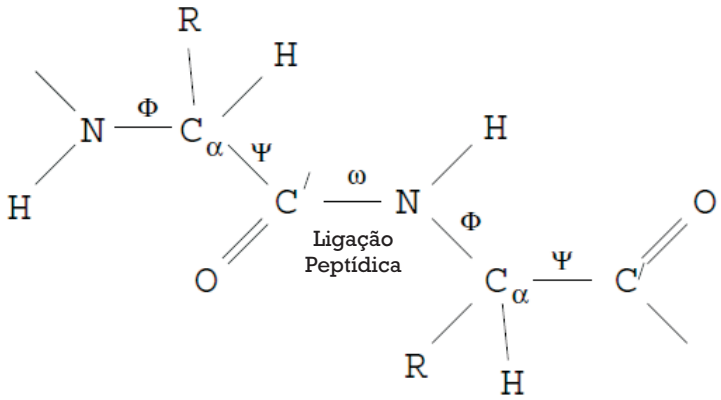

Figura 2.10: Ângulos diedrais da cadeia prinicipal de uma proteína.

Devido às propriedades químicas e sobre condições típicas, os ângulos diedrais ω são relativamente rígidos, assumindo a configuração denominada *trans* em que

$\omega = 180^o$, ou *cis* com $\omega = 0^o$ (ver Figura 2.11). A forma *trans* é mais frequente na maioria das ligações peptídicas (aproximadamente, 1000:1), a exceção são os grupos ligados aos resíduos proline onde a frequência de formas *trans* é bem menos expressiva (3:1). Os ângulos e vetores de ligação também são razoavelmente rígidos, com desvio padrão menor que 2^o, para os ângulos, e 0,2 Å, para os comprimentos [59, 39]. Já os ângulos φ e ψ são mais flexíveis, sendo responsáveis pelas principais características da geometria protéica.

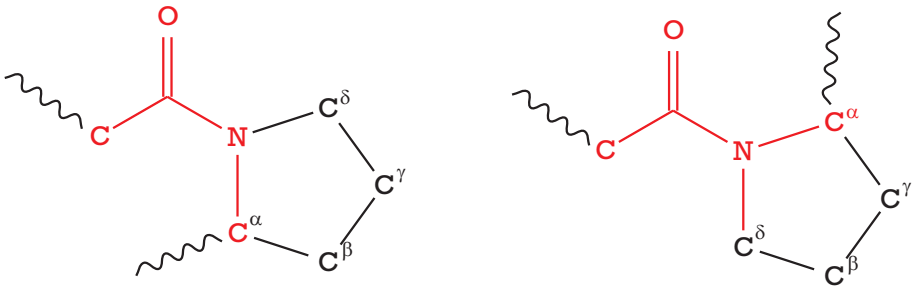

Figura 2.11: Nestas configurações, todos os quatro átomos estão situados no mesmo plano. À esquerda, a forma *cis* ($\omega = 0^o$). À direita, a forma *trans* com a distância entre os átomos C_1^α e C_2^α sendo a maior possível.

Cabe ressaltar que, além da flexibilidade $\{\varphi, \psi\}$ associada às ligações envolvendo os carbonos alfa, múltiplas conformações são possíveis para 18 dos 20 aminoácidos, as exceções são os aminoácidos glicina e alanina. Estruturas *rotameric* de proteínas são aquelas que possuem os mesmos ângulos $\{\varphi, \psi\}$, mas diferem nas configurações das cadeias laterais. Os ângulos diedrais utilizados para definir as rotações nas cadeias laterais são denotados pela letra χ, com subscritos sendo utilizados quando necessário (ver Figura 2.12).

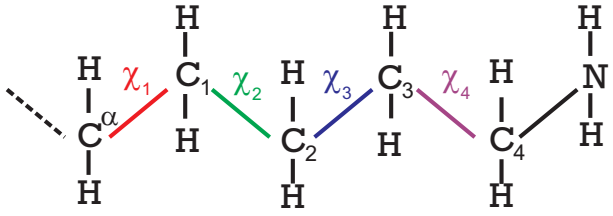

Figura 2.12: Ângulos das torções do resíduo lisina representados pelas variáveis χ_1 a χ_4.

Apesar de mais flexíveis, os ângulos $\{\varphi, \psi\}$ não assumem todos os valores possíveis devido às restrições impostas pelo tamanho das nuvens eletrônicas dos átomos de oxigênio e hidrogênio ao redor das ligações peptídicas. Por isso, somente certas

combinações são tipicamente observadas, com alguma dependência com respeito ao tamanho e forma dos resíduos. Na verdade, somente cerca de um décimo do espaço $\{\varphi, \psi\}$ é geralmente ocupado por proteínas e polipeptídeos [89]. Os primeiros a observar essa limitação foram G. N. Ramachandran e seus colaboradores em 1963. Por isso, o gráfico que mostra as regiões mais estáveis (de menor energia) para os pares de ângulos $\{\varphi, \psi\}$ é chamado gráfico de Ramachandram. Este gráfico é utilizado constantemente por pesquisadores durante o processo de construção de modelos de proteína para verificação da viabilidade dos ângulos $\{\varphi, \psi\}$ [91].

2.3 Métodos experimentais para determinação da conformação protéica

Um dos fundamentos da modelagem molecular é a noção de que a geometria molecular, a energia e várias propriedades moleculares podem ser calculadas a partir de modelos mecânicos sujeitos a forças físicas básicas. Uma molécula pode ser representada como um sistema mecânico no qual as párticulas (átomos) são conectados por molas (as ligações). Como uma resposta às forças inter- e intramamoleculares, a molécula então gira, vibra e se desloca para assumir uma conformação favorável (menor energia) no espaço.

As forças que atuam sobre a molécula são expressas como uma soma de termos harmônicos para o desvio com relação a valores de equilíbrio para o comprimento e os ângulos das ligações; termos para as torções que consideram rotações internas (rotações de subgrupos ao redor das ligações que as conecta); e potenciais eletroestáticos e de van der Waals [89]. O uso das mecânicas molecular e quântica tem sido uma rica fonte de modelos mais aderentes à dinâmica real da conformação protéica [99].

Apesar dos recentes avanços tanto computacionais quanto teóricos, o problema da conformação protéica, i.e., o problema da determinação da estrutura tridimensional a partir da estrutura primária e da descrição do ambiente, continua aberto, sem uma resposta definitiva [82, 89]. Por isso, os métodos experimentais para determinação da estrutura terciária das proteínas conservam seu status de ferramentas fundamentais.

Os dois principais métodos experimentais utilizados na determinação da estrutura tridimensional em alta resolução são a cristalografia de raio-X [16, 74] e a ressonância magnética nuclear [107, 54].

A técnica de cristalografia com raio-X envolve a análise dos padrões de difração produzidos quando um feixe de raios-X incide diretamente sobre um cristal bem-ordenado. Os padrões de difusão podem ser interpretados como reflexões da fonte primária do raio sobre o conjunto de planos paralelos no cristal. As marcas produzidas pela difração são gravadas sobre um detector (equipamento eletrônico ou filme de raio-X), escaneadas por um computador, e analisadas para determinar o mapa de densidade eletrônica (ver Figura 2.13). Os objetos (átomos) podem ser distinguidos se estiverem separados por uma distância superior ao valor da resolução do equipamento utilizado. Assim, menores resoluções estão associadas à representações estruturais mais detalhadas. Um dos maiores empecilhos à aplicação da cristalografia é a dificuldade de crescimento de cristais bem-ordenados de macromoléculas biológicas. Atualmente, é possível obter imagens das estruturas tridimensionais via cristalografia com resolução inferior a 2 Å [9].

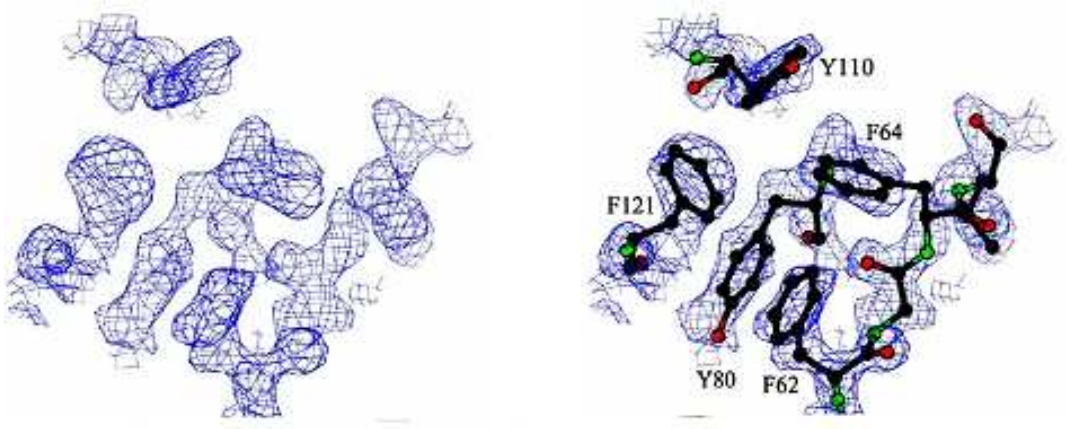

Figura 2.13: À direita, vê-se o mapa 3D da densidade eletrônica em cada ponto do espaço. À esquerda, sendo conhecidas a sequênca e a forma dos aminoácidos que compõem a proteína, pode-se ajustar o modelo protéico a esse mapa.

Com a habilidade de determinar estruturas de macromoléculas biológicas em resolução atômica em condições semifisiológicas, a espectroscopia de ressonância magnética nuclear (RMN) se tornou uma eminente ferramenta da biologia estrutural [98]. A informação tridimensional resultante dos experimentos com RMN não são tão detalhadas quanto as provenientes da cristalografia de raio-X, mas, em contra-

partida, a informação da RMN não é estática e incorpora efeitos devidos à variação térmica da solução.

Na RMN, poderosos campos magnéticos e ondas de radiação de alta frequência são aplicados na investigação do ambiente magnético dos núcleos atômicos. O ambiente local dos núcleos determina a frequência da absorção da ressonância. O espectro resultante da RMN contém informações sobre as interações e deslocamentos locais das moléculas que contêm os núcleos ressonantes. A frequência de absorção de diferentes grupos podem ser distinguidas quando equipamentos RMN de alta-frequência são utilizados, mas a necessidade de separar os diferentes sinais, com o intuito de produzir uma imagem clara, limita o tamanho (menos de 100 kDa) das macromoléculas que podem ser analisadas via RMN.

Cabe ressaltar que a cristalografia de raio-X e a resonância magnética nuclear não competem entre si, são, na verdade, técnicas que se complementam. Juntas fornecem uma imagem com detalhamento atômico da estrutura e dinâmica macromoleculares que pode ser utilizada para a melhor compreensão dos processos biológicos no nível molecular [18].

Mais recentemente, a *cryogenic electron microscopy* (cryo-EM) passou a ser utilizada no estudo de proteínas difíceis de cristalizar ou analisar via RMN [45, 95]. Esta técnica envolve o rápido congelamento de amostras de complexos moleculares que são então expostos às altas radiações dos microscópios eletrônicos, criando uma imagem tridimensional pela projeção das estruturas. A detecção adequada de partículas impõe um limite inferior de centenas de kDa para os complexos que podem ser analisados via cryo-EM. Apesar de não possuir a resolução superior a da cristalografia e RMN, com o advento de computadores mais eficientes e melhores algoritmos de reconstrução tridimensional, a cryo-EM desponta como uma boa promessa de contribuição no campo da biologia estrutural.

A Figura 2.14 mostra o crescimento do volume de dados do PDB entre 1972 e 2009. A cristalografia de raio-X é a mais profícua das técnicas experimentais com cerca de 48.618 contribuições, sendo seguida pela RMN com 7777 e cryo-EM com 230.

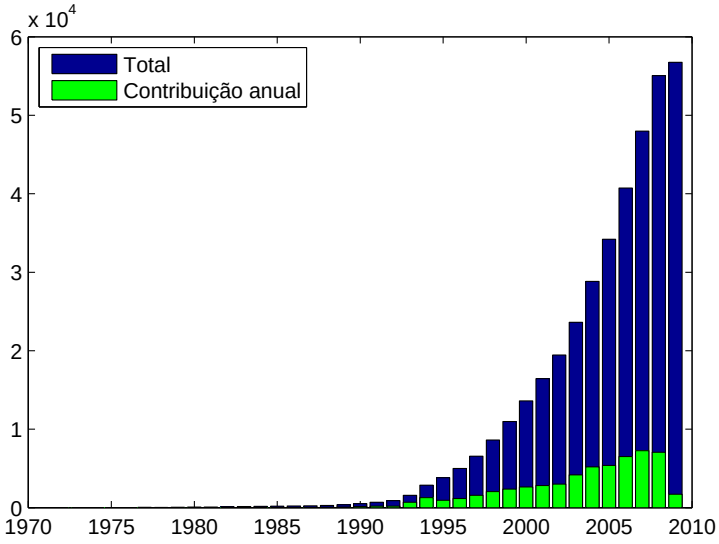

Figura 2.14: Número de estruturas disponíveis na base de dados PDB entre 1972 e março de 2009.

Capítulo 3

O problema geométrico da distância molecular

A espectroscopia de RMN fornece uma preciosa informação: uma rede de distâncias envolvendo pares de átomos de hidrogênio espacialmente próximos. As distâncias são derivadas de efeitos Overhauser nucleares (NOEs[1]) entre átomos de hidrogênio vizinhos (distantes a menos de 5-6Å).

Tendo o resultado dos experimentos com RMN, ou seja, uma rede de distâncias entre pares de átomos, o desafio passa a ser a obtenção de uma configuração válida para os átomos. Este problema pode ser colocado da seguinte forma: dado um conjunto de limites inferiores e superiores para um subconjunto esparso do conjunto de todas as distâncias interatômicas, determine uma configuração para a molécula (todos os átomos) que satisfaça as restrições de distância. Ou, de forma equivalente,

$$\begin{aligned} \textit{determinar} \quad & x_1, x_2, \ldots, x_m && (3.1)\\ s.a \quad & l_{ij} \leq \|x_i - x_j\| \leq u_{ij}, \forall (i,j) \in K \subset \{1,\ldots,m\}^2 && (3.2)\\ & x_i \in \mathbb{R}^3, \;\; i = 1,\ldots,m, && (3.3) \end{aligned}$$

onde K é um subconjunto do produto cartesiano $\{1,\ldots,m\}^2$ e identifica as distâncias cujos limites são conhecidos, e m é o número de átomos que se deseja posicionar.

Esse problema é conhecido como o problema da distância molecular (MDGP, do inglês *molecular distance geometry problem*) e, segundo Crippen e Havel [30],

[1] O efeito de Overhauser é o nome dado a trasferência de polarização de spins entre populações de átomos. Estas transferências geram alterações (picos) na espectroscopia de RMN.

foi definido por Cayley em 1841. No entanto, o problema só veio a ser sistematicamente estudado em 1928, quando Menger mostrou como a convexidade e muitas outras propriedades geométricas poderiam ser definidas e estudadas em termos das distâncias entre pares de pontos. Já em 1935, Schoenberg encontrou uma caracterização equivalente e percebeu a conexão do problema com as formas bilineares. Em 1953, Blumenthal [14] publicou uma monografia sobre o tema, onde foi enunciado o problema fundamental da distância geométrica como sendo

> "When we have given a set of distances between pairs of points, the distance geometry can give a clue to find a correct set of coordinates for the points in three-dimensional Euclidean space satisfying the given distance constraints."

Um caso particular do problema da distância molecular é obtido quando consideramos conhecidas as distâncias exatas entre todos os pares de átomos da proteína. Neste caso, o MDGP pode ser resolvido pela fatoração da matriz formada pelas distâncias conhecidas.

De fato, se a distância d_{ij} existente entre os átomos i e j é conhecida para todo par $(i,j) \in \{1,2,\ldots,m\}^2$, o problema passa a ser determinar uma configuração $\{x_1, x_2, \ldots, x_m\}$ viável, ou seja, tal que

$$\|x_i - x_j\| = d_{ij}, \quad i,j = 1,\ldots,m, \tag{3.4}$$

onde $x_k = (x_{k1}, x_{k2}, x_{k3})$ representa a posição no espaço tridimensional do k-ésimo átomo.

Podemos, sem perda de generalidade, posicionar o primeiro átomo na origem, ou seja, $x_1 = (0,0,0)$ pois a estrutura tridimensional da proteína é invariante com respeito à translação. Com isso, temos o seguinte sistema de equações não-lineares

$$d_{i1} = \|x_i - x_1\| = \|x_i\|, \tag{3.5}$$

$$d_{ij} = \|x_i - x_j\|, \quad i,j = 2,\ldots,m, \tag{3.6}$$

cujas variáveis são as posições $x_i = (x_{i1}, x_{i2}, x_{i3}), x_j = (x_{j1}, x_{j2}, x_{j3})$ ocupadas pelos átomos i e j respectivamente. Antes de determinarmos uma solução para esse

sistema, será conveniente reescrevermos a equação (3.6) na forma

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{3}(x_{ik} - x_{jk})^2 && (3.7)\\ &= \|x_i\|^2 + \|x_j\|^2 - 2x_i^t x_j && (3.8)\\ &= d_{i0} + d_{j0} - 2x_i^t x_j, \quad i,j = 1,\ldots,m. && (3.9) \end{aligned}$$

Ou, de forma equivalente,

$$x_i^t x_j = (d_{i1} + d_{j1} - d_{ij})/2, \quad i,j = 2,\ldots,m. \tag{3.10}$$

Definindo $X = [x_2,\ldots,x_m] \in \mathbb{R}^{(m-1)\times 3}$ como sendo a matriz cujas linhas são as posições dos átomos, podemos escrever a equação (3.10) na forma matricial

$$D = XX^t, \tag{3.11}$$

com $D_{ij} = (d_{i1} + d_{j1} - d_{ij})/2$. Se as distâncias são consistentes, i.e., se existe um conjunto de pontos viáveis no espaço tridimensional, então a matriz D deve necessariamente ser semidefinida positiva com posto menor ou igual a três (mesmo posto de X). Tomando a decomposição em valores singulares de D, obtemos

$$D = U\Sigma U^t,$$

onde U é uma matriz $n \times 3$ ortogonal e Σ uma matriz 3×3 diagonal com autovalores $\sigma_1, \sigma_2, \sigma_3$ positivos. Assim, uma solução para o sistema (3.11) pode ser obtida tomando

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

Uma vez que a decomposição em valores singulares pode ser feita em $\mathcal{O}(m^3)$ operações em ponto flutuante [51], então a solução para o MDGP com todas as distâncias exatas entre os pares de átomos sendo conhecidas pode ser obtida em tempo polinomial [14, 30, 36].

Na verdade, um algoritmo muito simples apresentado por Dong e Wu em [35] permite obter uma solução para o MDGP em $\mathcal{O}(m)$ operações, no caso em que todas as

distâncias são conhecidas. O algoritmo se baseia na propriedade geométrica elementar de que é possível determinar a posição de um ponto, a partir do conhecimento das distâncias entre este ponto e quatro pontos fixos e não-colineares conhecidos.

Com efeito, sejam $x_1, x_2, x_3, x_4 \in \mathbb{R}^3$ quatro pontos fixos conhecidos e não-colineares, e d_i as distâncias entre x_i, $i = 1, \ldots, 4$, e y, o ponto cuja posição desejamos determinar. Com estas hipóteses, podemos escrever o sistema não-linear

$$\begin{aligned} d_i^2 &= \|x_i - y\|^2 && (3.12)\\ &= \|x_i\|^2 + \|y\|^2 - 2x_i^t y, \quad i = 1, \ldots, 4. && (3.13)\\ & && (3.14) \end{aligned}$$

Isolando $\|y\|^2$ na equação associada ao ponto x_4, obtemos

$$\|y\|^2 = d_4^2 - \|x_4\|^2 + 2x_4^t y.$$

Substituindo $\|y\|^2$ nas equações associadas aos demais pontos ($i = 1, 2, 3$), obtemos o sitema linear

$$y^t(x_4 - x_i) = (d_i^2 - d_4^2 + \|x_4\|^2 - \|x_i\|^2)/2, \quad i = 1, 2, 3,$$

cuja única solução é o ponto procurado.

O primeiro passo do algoritmo proposto por Dong e Wu é determinar a posição de quatro átomos não-colineares. A partir desta base, os quatro átomos, tenta-se fixar todos os demais átomos. Como é baixa a probabilidade de que quatro átomos tomados aleatoriamente sejam colineares, a primeira base, muito provavelmente, será suficiente para a fixação dos outros átomos. Portanto, por esse algoritmo, seriam necessárias apenas $\mathcal{O}(m)$ operações em ponto flutuante.

Uma abordagem mais realista do MDGP é obtida quando, ao invés de considerar conhecidas todas as distâncias exatas entre os pontos (átomos), supõe-se disponível apenas um subconjunto esparso dessas distâncias. Essa abordagem ainda que simplista, pois considera distâncias exatas, transforma o MDGP em um problema muito mais complexo do ponto de vista computacional. Na verdade, neste caso, o MDGP se inclui na classe dos problemas NP-difícil. Em [87], Saxe mostrou que o MDGP

unidimensional é equivalente ao problema do particionamento de conjuntos, um problema conhecido da classe NP-difícil[2].

Em [113], encontra-se um exemplo de como o problema do particionamento de grafos pode ser reduzido ao MDGP unidimensional. Primeiro, suponha que se tenha um conjunto $S = \{1, 2, 2, 4, 3\}$ de números inteiros e que se deseja particioná-lo em dois subconjuntos cujas somas dos elementos sejam iguais. A partir desse problema, construa a seguinte instância unidimensional do MDGP: considere 6 pontos no espaço unidimensional e exija que a distância entre o primeiro e o segundo seja igual ao primeiro inteiro em S, que a distância entre o segundo e o terceiro seja igual ao segundo inteiro em S, que a distância entre o terceiro e o quarto seja igual ao terceiro inteiro em S, e assim por diante. Finalmente, exija que a distância entre o primeiro e o último ponto seja igual a zero.

Se uma solução para esse MDGP unidimensional for encontrada, então obtém-se automaticamente uma solução para o problema do particionamento de S, através da seguinte regra de associação: se o segundo ponto estiver à direita do primeiro ponto, o primeiro elemento de S pertencerá ao subconjunto S_d, caso contrário, pertencerá ao subconjunto S_e; se o terceiro ponto estiver à direita do segundo ponto, o segundo elemento de S pertencerá ao subconjunto S_d, caso contrário, pertencerá ao subconjunto S_e; e assim sucessivamente.

No fundo, o que se fez nesse exemplo foi associar cada um dos elementos de S a um segmento de reta, e a restrição de coincidência nas posições do primeiro e último pontos é motivada pelo fato de que as somas dos elementos de cada um dos subconjuntos que particionam S devem ser iguais. A Figura 3.1 mostra como o problema do particionamento pode ser reduzido ao MDGP unidimensional.

Na prática, os experimentos com RMN fornecem apenas um subconjunto das distâncias entre os átomos (distâncias menores que 5-6 Å[89]) e, como em todo experimento físico, os dados possuem um limite em sua precisão, portanto, são conhecidos apenas limites superiores e inferiores para as distâncias. Assim, na definição mais realista do MDGP, o objetivo é determinar as posições $x_1, x_2, \ldots, x_m \in \mathbb{R}^3$ tais que

$$l_{ij} \leq \|x_i - x_j\| \leq u_{ij}, \quad \forall (i,j) \in K \subsetneq \{1, \ldots, m\}^2, \tag{3.15}$$

[2]Saxe mostrou ainda que o MDGP em um espaço n-dimensional é NP-difícil para todo n maior ou igual a um.

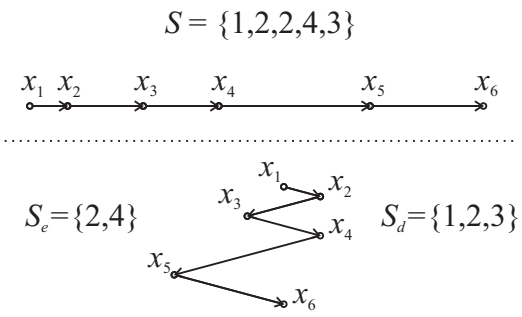

Figura 3.1: Equivalência do particionamento de conjuntos e o MDGP unidimensional.

onde l_{ij} e u_{ij} são, respectivamente, os limites inferior e superior para as restrições de distância.

A princípio, essa forma do MDGP parece ser mais fácil de solucionar, já que as restrições das distâncias são relaxadas e, portanto, mais fáceis de satisfazer. Contudo, na prática, os limites inferior e superior são próximos e, assim, o problema é ainda de difícil solução. Em [76], Moré mostrou que se os limites superior e inferior são próximos, o MDGP com as distâncias relaxadas também pertence a classe NP-difícil sendo, portanto, tão difícil de solucionar quanto o MDGP com distâncias exatas.

3.1 Problemas relacionados

Na literatura, encontram-se problemas intimamente relacionados ao MDGP, i.e., cujas soluções podem com frequência ser aplicadas ao MDGP. Um desses problemas é o *problema de localização em rede com informação de distância* [41, 11, 26, 90]. No problema de localização, o objetivo também é posicionar os pontos de um determinado conjunto de forma que as distâncias entre eles atendam às restrições impostas. A principal diferença é que no problema de localização são conhecidas não apenas as distâncias entre os pontos que se deseja posicionar, mas também as distâncias a alguns pontos fixos (âncoras).

Outro problema associado ao MDGP é o *problema do preenchimento da matriz de distância euclideana* [2, 69], onde, sendo conhecidas apenas algumas das entradas de uma matriz de distâncias, deseja-se determinar os demais elementos dessa matriz. Um método de solução do problema de preenchimento pode ser utilizado para determinar as distâncias desconhecidas de uma instância do MDGP exato, recaindo

assim na formulação polinomial do MDGP.

O *problema da aproximação da matriz de distância euclideana* também possui estreita ligação com o MDGP [73, 22, 47]. No problema de aproximação, uma matriz cujos elementos são valores aproximados das distâncias entre os átomos é dada *a priori* e o objetivo é determinar a matriz de distância euclideana mais próxima desta matriz. Em uma instância do MDGP em que são conhecidos os limites de todas as distâncias, a matriz formada pelos pontos médios de cada um dos intervalos pode ser tomada como entrada para um algoritmo de solução do problema de aproximação. É razoável supor que, se a matriz euclideana obtida na solução for suficientemente próxima da matriz de entrada, então as distâncias obtidas estarão no interior dos intervalos (restrições) e, novamente, recai-se na formulação polinomial do MDGP.

3.2 Comparação de estruturas via RMSD

Uma interessante questão teórica/experimental relacionada ao MDGP é a unicidade de solução. Note que, por ser o MDGP definida unicamente em função de distâncias, uma mesma solução apresentada em diferentes posições do espaço pode dar a impressão da existência de múltiplas soluções. Assim, é importante dispor de ferramentas analíticas/computacionais que permitam comparar estruturas tridimensionais independentemente das regiões do espaço que elas ocupem.

A $RMSD$ (*root mean square deviation*) é uma forma frequentemente empregada para quantificar discrepâncias entre estruturas tridimensionais [60, 51, 113]. Supondo que duas estruturas tridimensionais $X, Y \in \mathbb{R}^{m\times 3}$ possuam os mesmos centros de massa, define-se a RMSD entre as estruturas X e Y como sendo o mínimo da norma de Frobenius da diferença entre as duas estruturas, sujeita a uma apropriada rotação em X, i.e.,

$$RMSD(X,Y) = \min_{Q} \|Y - XQ\|_F, \quad Q^tQ = I, \tag{3.16}$$

onde Q é uma matriz 3×3 ortogonal.

Note que, por definição,

$$\|Y - XQ\|_F^2 = \mathrm{tr}\,(Y^tY) + \mathrm{tr}\,(X^tX) - 2\mathrm{tr}\,(Q^tX^tY). \tag{3.17}$$

Então, minimizar $\|Y - XQ\|_F$ equivale a maximizar $\mathrm{tr}\,(Q^t X^t Y)$.

Sejam $C = X^t Y \in \mathbb{R}^{3\times 3}$ e $C = U\Sigma V^t$ a decomposição em valores singulares de C. Segue que,

$$\mathrm{tr}\,(Q^t X^t Y) = \mathrm{tr}\,(Q^t C) = \mathrm{tr}\,(Q^t U\Sigma V^t) = \mathrm{tr}\,(V^t Q^t U\Sigma) \leq \mathrm{tr}\,(\Sigma), \tag{3.18}$$

pois os elementos da diagonal de Σ são todos não-negativos e $V^t Q^t U$ é uma matriz ortogonal e, por isso, possui traço. Assim, $\mathrm{tr}\,(Q^t X^t Y)$ é maximizado quando $Q = UV^t$.

Segue que a $RMSD(X, Y)$ pode ser computada pelo seguinte procedimento:

1. Faça $C = X^t Y$;
2. Obtenha a decomposição em valores singulares de C, i.e., $C = USV^t$;
3. Faça $Q = UV^t$, e então $RMSD(X, Y) = \|Y - XQ\|$.

Note que o procedimento acima fornece uma medida de similaridade que é invariante sobre rotações e translações das estruturas e também permite determinar a rotação Q que melhor sobrepõe as estruturas.

3.3 MDGP via programação matemática

O MDPG pode ser formulado como um problema global de mínimos quadrados. De fato, no caso mais geral, a determinação da estrutura protéica está associada ao problema de determinar posições $x_1, \ldots, x_n \in \mathbb{R}^3$ tais que

$$l_{ij} \leq \|x_i - x_j\| \leq u_{ij}, \quad (i, j) \in K \subset \{1, \ldots, m\}^2, \tag{3.19}$$

onde l_{ij} e u_{ij} são, respectivamente, os limites inferior e superior conhecidos para as distância e K um subconjunto dos pares de átomos. Assim, resolver o MDGP equivale a obter uma solução global do seguinte problema de mínimos quadrados:

Definição 1 (MDGP via programação matemática)

$$
\begin{aligned}
(FG) \quad \min \quad & f(x) = \sum_{(i,j)\in K} p_{ij}(d_{ij}) \\
s.a \quad & d_{ij} = d_{ij}(x) = \|x_i - x_j\|, \ \ \forall (i,j) \in K \subset \{1,\ldots,m\}^2 \\
& x_i \in \mathbb{R}^3 \quad \forall i \in \{1,\ldots,m\},
\end{aligned} \tag{3.20}
$$

onde p_{ij} é qualquer função de penalidade com as propriedades

(P1) *$p_{ij}(d_{ij}) = 0$, se $d_{ij} \in [l_{ij}, u_{ij}]$;*

(P2) *$p_{ij}(d_{ij}) > 0$, se $d_{ij} \notin [l_{ij}, u_{ij}]$.*

É fácil ver que $f(x) = f(x_1,\ldots,x_m) = 0$ se, e somente se, são satisfeitas todas as restrições da forma (3.19). Ou, em outras palavras, são equivalentes as sentenças "$x \in \mathbb{R}^{3\times m}$ é uma solução global de (FG)" e "x é uma configuração viável do MDGP".

Duas das infinitas possibilidades para as funções de penalidade p_{ij} são

$$p_{ij}(d_{ij}) = \max\{l_{ij} - d_{ij}, 0\} + \max\{d_{ij} - u_{ij}, 0\}; \tag{3.21}$$

$$p_{ij}(d_{ij}) = \max^2\left\{\frac{l_{ij}^2 - d_{ij}^2}{l_{ij}^2}, 0\right\} + \max^2\left\{\frac{d_{ij}^2 - u_{ij}^2}{u_{ij}^2}, 0\right\}. \tag{3.22}$$

Infelizmente, o problema da melhor formulação em programação matemática para o problema geométrico da distância molecular continua sem solução. Não existem resultados de unicidade (convexidade) e muitas formulações (como, por exemplo, as que utilizam as funções de penalidade definidas anteriormente) não são sequer diferenciáveis o que inviabiliza a aplicação direta dos métodos clássicos e mais robustos de otimização.

3.4 Estratégias de solução

Não encontramos na literatura uma formulação ideal (convexa) do problema (FG) para o caso mais realista, i.e., o caso em que K é um subconjunto próprio e esparso de $\{1,2,\ldots,m\}^2$ e $l_{ij} < u_{ij}$. Por isso, diferentes proposta vêm sendo apresentadas: suavização [79, 5], programação semidefinida [11], diferença de funções convexas [6],

etc. Segue abaixo uma lista incompleta das propostas para solução do MDGP, via programação matemática.

3.4.1 EMBED

Em [28, 30], Crippen e Havel apresentaram um algoritmo chamado EMBED para resolver o MDGP. O algoritmo EMBED lida com limites inferior e superior para todas as distâncias entre os átomos. Na prática, apenas algumas das distâncias têm seus limites inferior e superior conhecidos. No algoritmo EMBED esta escassez de limites para as distâncias é contornada através da desigualdade triangular. Mais especificamente, na ausência de limites, assumi-se um grande valor positivo como limite superior e zero como limite inferior. Como a desigualdade triangular deve ser satisfeita, para quaisquer átomos i, j, k tem-se $u_{ij} \leq u_{ik} + u_{jk}$ e, então, u_{ij} pode ser reduzido para a soma do lado direito. Este processo pode ser repetido até que todos os limites superiores alcancem seus valores mínimos. Algo similar pode ser feito com os limites inferiores também utilizando triplas de átomos.

O algoritmo EMBED segue sucessivamente três etapas. Na primeira delas, toma como entrada as distâncias obtidas experimentalmente (usualmente esparsas e imprecisas) e tenta estimar (via desigualdade triangular) um intervalo possível de valores para cada uma delas. Na segunda etapa, um conjunto de coordenadas para os átomos são calculadas de tal forma que as distâncias estejam nos intervalos obtidos na etapa anterior. A terceira etapa utiliza métodos de otimização numérica para minimizar uma função erro (mínimos quadrados) cuja magnitude mede os desvios das coordenadas com relação às restrições de distância fornecidas na entrada.

3.4.2 ABBIE

Em [57, 58], Hendrickson descreve uma estratégia para o MDGP exato que tenta evitar resolver grandes problemas de otimização. Esta estratégia pode ser vista como um algoritmo de dividir-para-conquistar. A idéia dos algoritmos baseados na estratégia de dividir-para-conquistar é dividir um problema grande em partes menores, essas partes são resolvidas separadamente, possivelmente de forma recursiva, e suas soluções são recombinadas em uma solução para o problema original.

A observação fundamental da proposta de Hendrickson é que existem no MDGP

subproblemas que podem ser resolvidos independentemente. Se for possível identificar um subgrafo que possua muitas arestas, então, considerando apenas as restrições das arestas desse subgrafo, pode ser possível determinar as posições relativas de seus vértices. Uma vez que o subproblema de determinar as posições dos vértices do subgrafo tenha sido resolvido, o subgrafo pode ser tratado como um corpo rígido.

A grande vantagem dessa proposta está relacionada ao número de variáveis dos problemas de otimização considerados. No espaço tridimensinal, um corpo rígido possui somente seis graus de liberdade, já cada vértice, se considerado independentemente, possui três graus de liberdade. Assim, ao tratar um conjunto de vértices como um corpo rígido, o número de variáveis consideradas pode ser drasticamente reduzido.

Esta proposta foi implementada em um código chamado ABBIE e testada com dados simulados provenientes da ribonuclease pancreática bovina A, uma pequena proteína formada por 124 aminoácidos, cuja estrutura tridimensional é conhecida [83]. O conjunto de dados utilizado era formado por todas as distâncias entre pares de átomos no mesmo aminoácido e 1167 distâncias adicionais correspondentes aos pares de átomos de hidrogênio com proximidade menor que 3,5 Å.

3.4.3 Perturbação estocástica

Em [114], Zou, Bird e Schnabel apresentaram um algoritmo de otimização global via pertubação estocástica para resolução do MDGP com restrições definidas tanto por distâncias exatas quanto por distâncias limitadas. Este algoritmo combina uma fase estocástica que identifica um conjunto inicial de minimizadores locais com uma segunda fase, mais determinística, que busca mínmimos locais mais profundos. Segundo os autores, uma das vantagens deste algoritmo é que, na segunda fase, ao incorporar a estrutura separável do problema, são resolvidos subproblemas de otimização global com dimensão muito menor que a do problema original.

Durante a primeira fase do algoritmo de Zou, gera-se um conjunto aleatório de pontos no espaço das variáveis do problema original através do posicionamento aleatório dos átomos no espaço tridimensional. As piores das configurações são descartadas, e tenta-se aprimorar as melhores configurações pela movimentação de um átomo ou um par de átomos, até que a função objetivo atinja um valor de corte.

Algumas das configurações aprimoradas são então utilizadas como pontos iniciais para um algoritmo de otimização local no espaço do problema original. Alguns dos minimizadores locais encontrados nesta fase são utilizados na segunda fase na tentativa de se obter resultados ainda melhores.

Na segunda fase, seleciona-se sucessivamente um minimizador local para a tentativa de aprimoramento. Um par de átomos é escolhido e um algoritmo estocástico de otimização global de pequena escala é aplicado à configuração tomando como variável apenas este par de átomos e mantendo os demais átomos fixos. Após esse passo, um algoritmo de otimização local é aplicado sobre a estrutura completa tomando como ponto inicial a melhor configuração obtida durante o passo de otimização global de pequena escala. As melhores configurações são inseridas numa lista de minimizadores locais e a segunda fase é iterada um definido número de vezes.

3.4.4 DGSOL

A exposição do algoritmo DGSOL será aqui feita com maior atenção/detalhamento devido à semelhança existente entre a abordagem desenvolvida por Moré e Wu e a proposta do presente trabalho.

Seguindo as idéias expostas em [106], Moré e Wu desenvolveram um algoritmo, chamado DGSOL [79], para resolver uma variação do $MDGP$. O algoritmo DGSOL é baseado em uma estratégia de otimização global por aproximação contínua. A idéia é transformar a função objetivo

$$f(x_1, \ldots, x_n) \equiv \sum_{(i,j)\in K} \min^2\{\|x_i - x_j\|^2 - l_{ij}^2, 0\} + \max^2\{\|x_i - x_j\|^2 - u_{ij}^2, 0\} \quad (3.23)$$

em uma função suave que possua um menor número de minimizadores através da técnica de suavização Gaussiana [64, 65, 66, 85]. Um algoritmo de otimização local é aplicado à função suavizada e técnicas de continuação são utilizadas para rastrear os minimizadores da função original, a partir dos minimizadores da função suavizada. Uma das vantagens do DGSOL é a possibilidade de utilização de um conjunto esparso de dados relativos às distâncias.

A transformada Gaussiana depende de um parâmetro λ que controla o grau de suavização. A função original f é obtida se $\lambda = 0$ e a intensidade da suavização

aumenta à medida que aumentam os valores de λ. A transformada Gaussiana $\langle f \rangle_\lambda$ de uma função $f : \mathbb{R}^n \to \mathbb{R}$ é definida como

$$\langle f \rangle_\lambda \equiv \frac{1}{\pi^{n/2}\lambda^n} \int_{\mathbb{R}^n} f(y) \exp\left(-\frac{\|y-x\|^2}{\lambda^2}\right) dy. \tag{3.24}$$

O valor $\langle f \rangle_\lambda(x)$ é uma média de f na vizinhança de x, com tamanho relativo desta vizinhança sendo controlado pelo parâmetro λ. A transformada Gaussiana $\langle f \rangle_\lambda$ também pode ser vista como a convolução de f com a função densidade Gaussiana.

Em casos particulares, por exemplo, quando os limites l_{ij} e u_{ij} são iguais (distâncias exatas), a transformada Gaussiana possui expressão analítica, mas, no caso geral, a integral da expressão (3.24) deve ser aproximada. Em [79], Moré utiliza a quadratura Gaussiana para realizar as aproximações das integrais.

O parâmetro λ é de extrema importância, pois tem a capacidade de convexificar a aproximação da função transformada. Mais especificamente, para valores λ acima de um valor de corte λ_c, a aproximação a $\langle f \rangle_{\lambda,q}$ da transformada Gaussiana $\langle f \rangle_\lambda$, calculada pela quadratura Gaussiana com q nós, é uma função convexa. Curiosamente, a convexificação da função aproximada é algo indesejável, pois a solução obtida é degenerada, todos os pontos ocupam a mesma posição. Então, a escolha do valor do parâmetro λ deve ser feita com parcimônia, buscando convexificar algumas das parcelas da expressão (3.23), mas evitando valores altos que conduzam a degeneração das soluções.

Esta estratégia para o MDGP foi implementada e testada em instâncias artificiais com tamanhos modestos (entre 100 e 200 átomos), geradas a partir de dados do PDB. Nos experimentos numéricos realizados, o algoritmo DGSOL foi capaz de determinar uma solução independentemente do ponto inicial escolhido. Portanto, a estratégia de aproximação permite determinar a solução global com menos esforço computacional que o requerido em estratégias de multistart.

3.4.5 Otimização de diferenças de funções convexas

Em [5, 6], An e Tao abordaram o MDGP sob a perspectiva dos algoritmos de otimização para diferenças de funções convexas (d.c. algorithms, do inglês *difference of convex functions*). Eles trabalharam no espaço $\mathcal{M}_{m,3}(\mathbb{R})$, o espaço das matrizes

reais de ordem $m \times 3$, onde para $X \in \mathcal{M}_{m,3}(\mathbb{R})$, X_i é sua i-ésima linha. Identificando um conjunto de posições $x_1, \ldots, x_m$ com a matriz X, $X_i^t = x_i$ para $i = 1, \ldots, m$, o MDGP pode ser expresso por

$$0 = \min \left\{ \sigma(X) := \frac{1}{2} \sum_{(i,j) \in K, i<j} w_{ij} \theta_{ij}(X) : X \in \mathcal{M}_{m,3}(\mathbb{R}) \right\}, \tag{3.25}$$

onde $w_{ij} > 0$ para $i \neq j$ e $w_{ij} = 0$ para todo i. O potencial de par (pairwise) $\theta_{ij} : \mathcal{M}_{m,3}(\mathbb{R}) \to \mathbb{R}$ é definido para as restrições de distâncias exatas como

$$\theta_{ij}(X) = (d_{ij}^2 - \left\| X_i^t - X_j^t \right\|^2)^2 \tag{3.26}$$

ou

$$\theta_{ij}(X) = (d_{ij} - \left\| X_i^t - X_j^t \right\|)^2, \tag{3.27}$$

e para as restrições de distâncias limitadas como

$$\theta_{ij}(X) = \overset{2}{\min} \left\{ \frac{\left\| X_i^t - X_j^t \right\|^2 - l_{ij}^2}{l_{ij}^2}, 0 \right\} + \overset{2}{\max} \left\{ \frac{\left\| X_i^t - X_j^t \right\|^2 - u_{ij}^2}{u_{ij}^2}, 0 \right\}. \tag{3.28}$$

Uma matriz X é uma solução do MDGP se, e somente se, for um minimizador global do problema (3.25) e $\sigma(X) = 0$.

An e Tao demonstraram que o algoritmo para diferenças de funções convexas pode ser adaptado para desenvolvimento de algoritmos eficientes para a solução de MDGP's exatos de grande porte. Eles propuseram várias versões do algoritmo d.c. baseadas em diferentes formulações do problema. Devido ao seu caráter local, a otimalidade global não pode ser garantida para um problema d.c. genérico. No entanto, o fato de que a otimalidade global pode ser obtida com pontos iniciais convenientes os motivou a investigar uma técnica para geração de bons pontos iniciais para os algoritmos d.c. na solução de (3.25), com θ_{ij} definido em (3.27).

An e Tao realizaram experimentos numéricos com três conjuntos de dados: os dados artificiais de Moré e Wu [78] (até 4096 átomos), 16 proteínas do PDB [10] (de 146 até 4189 átomos), e os dados de Hendrickson [58] (de 63 até 777 átomos). Utilizando esses dados, os algoritmos d.c. mostraram-se capazes de resolver eficientemente MDGP's exatos de grande porte.

3.4.6 Algoritmo de construção geométrica

Como já citado, Dong e Wu em [35] apresentaram um algoritmo com complexidade $O(m)$ para a solução da formulação exata do MDGP com todas as distâncias sendo conhecidas. O algoritmo baseia-se em simples relações geométricas entre as coordenadas dos átomos e as distâncias existentes entre eles. Em [36], Dong e Wu exibiram uma versão modificada do algoritmo de construção geométrica para um conjunto esparso de distâncias extas. Assumindo que seja possível fixar as coordenadas de pelo menos quatro átomos, propuseram que cada um dos átomos não fixados seja examinado a fim de se determinar se são conhecidas as distâncias entre ele e pelo menos quatro átomos fixados. Em caso afirmativo, as coordenadas desse átomo podem ser imediatamente determinadas. O algoritmo continua até que todos os átomos sejam fixados, mas não há garantias de que uma solução seja encontrada pois, em cada loop, o algoritmo requer que pelo menos um dos átomos não fixados possa ser determinado utilizando quatro dos átomos fixados. São requeridos $O(m^4)$ passos para a conclusão do método.

3.4.7 Algoritmo branch-and-prune

Em [71], Liberti, Lavor e Maculan propuseram um algoritmo, denominado branch-and-prune (BP), baseado em uma formulação discreta para o MDGP exato. Eles observaram que, com hipóteses adicionais, é possível formular o MDGP, aplicado à cadeia principal das proteínas, como um problema discreto de busca. Nesta abordagem, são consideradas as hipóteses adicionais de conhecimento dos ângulos e comprimentos das ligações covalentes e também conhecimento das distâncias entre átomos separados por três ligações consecutivas.

Para descrever a cadeia principal de uma proteína com m átomos, além dos comprimentos $d_{i-1,i}$, para $i = 2, \ldots, m$ e ângulos $\theta_{i-2,i}$ das ligações, para $i = 3, \ldots, m$, é necessário considerar os ângulos de torção $\omega_{i-3,i}$, para $i = 4, \ldots, m$, que são os ângulos entre as normais aos planos definidos pelas átomos $i-3, i-2, i-1$ e $i-2, i-1, i$ (ver Figura 3.2).

Dados todos os comprimentos $d_{1,2}, \ldots, d_{m-1,m}$ e os ângulos $\theta_{1,3}, \ldots, \theta_{m-2,m}$ das ligações e os ângulos de torção $\omega_{1,4}, \ldots, \omega_{m-3,m}$ de uma molécula com m átomos, as coordenadas Cartesianas (x_{i1}, x_{i2}, x_{i3}) para cada átomo i na molécula podem ser

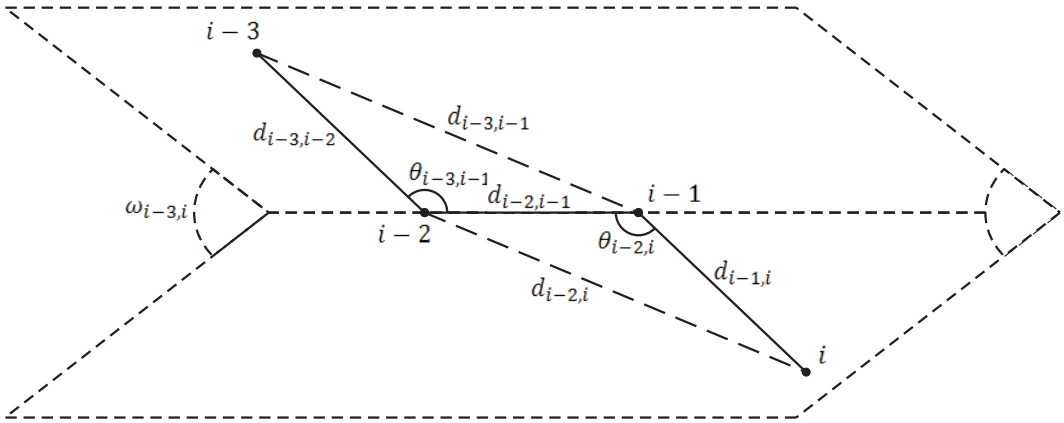

Figura 3.2: Definições dos comprimentos, ângulos das ligações e os ângulos de torção.

obtidos utilizando a seguinte fórmula:

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ 1 \end{pmatrix} = B_1 B_2 \dots B_i \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \forall i = 1, \dots, m, \tag{3.29}$$

onde

$$B_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, B_2 = \begin{pmatrix} -1 & 0 & 0 & -d_{1,2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.30}$$

$$B_3 = \begin{pmatrix} -\cos\theta_{1,3} & -\sin\theta_{1,3} & 0 & -d_{2,3}\cos\theta_{1,3} \\ \sin\theta_{1,3} & -\cos\theta_{1,3} & 0 & d_{2,3}\cos\theta_{1,3} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.31}$$

e

$$B_i = \begin{pmatrix} -\cos\theta_{i-2,i} & -\sin\theta_{i-2,i} & 0 & -d_{i-1,i}\cos\theta_{i-2,i} \\ \sin\theta_{i-2,i}\cos\omega_{i-3,i} & -\cos\theta_{i-2,i}\cos\omega_{i-3,i} & -\sin\omega_{i-3,i} & d_{i-1,i}\sin\theta_{i-2,i}\cos\omega_{i-3,i} \\ \sin\theta_{i-2,i}\sin\omega_{i-3,i} & -\cos\theta_{i-2,i}\sin\omega_{i-3,i} & \cos\omega_{i-3,i} & d_{i-1,i}\sin\theta_{i-2,i}\cos\omega_{i-3,i} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.32}$$

para $i = 4, \dots, m$.

Uma vez que os comprimentos e ângulos das ligações são, por hipótese, conhecidos, as coordenadas Cartesianas de todos os átomos de uma molécula podem ser determinadas usando os valores $\cos\omega_{i-3,i}$ e $\sin\omega_{i-3,i}$, para $i = 4, \ldots, m$. Na verdade, a posição do i-ésimo átomo pode ser expressa em termos das posições dos três átomos que o precedem, assim, existem 2^{m-3} conformações possíveis, o que caracteriza a viabilidade de discretização do problema.

Em termos gerais, no algoritmo BP, a cada passo, o i-ésimo átomo pode ser situado em duas posições. A busca é ramificada sobre todas as posições que são viáveis com respeito a todas as restrições, e, se uma posição não é viável, seu ramo é abandonado.

Em [71] e [31], são encontrados resultados numéricos obtidos com o algoritmo BP com dados artificiais propostos por Moré e Wu [78] e Lavor [70] e dados reais obtidos no PDB.

3.4.8 Programação semidefinida

Em [11, 13], Patrick Biswas e colaboradores propuseram um algoritmo para solucionar o MDGP via programação semidefinida. Neste algoritmo, o grafo cujos vértices são os índices dos átomos e as aresta, as distâncias (ou limites) conhecidas, é dividido em subgrafos usando um método de agrupamento. Então, uma estratégia de relaxação de programação semidefinida e o método de busca do gradiente são aplicados para determinar uma realização tridimensional[3] de cada subgrafo. Finalmente, um algoritmo é utilizado para combinar as soluções, determinando as posições no sistema de coordenadas global.

O problema de realização de grafo relacionado ao MDGP é

$$\begin{aligned} &Determinar && x_1, \ldots, x_m \\ &\quad s.a && l_{ij}^2 \leq \|x_i - x_j\| \leq u_{ij}^2, \ \ \forall (i,j) \in K \subset \{1, ..., m\}^2. \end{aligned} \tag{3.33}$$

Seja $X = [x_1 \ x_2 \ \ldots \ x_m]$ a matriz $3 \times m$ que se deseja determinar. O problema

[3]Uma realização de um grafo com peso em um espaço tridimensional é um conjunto de pontos cujas distâncias correspondem aos pesos das arestas que os ligam.

(3.33) pode ser reescrito na forma matricial abaixo

$$\begin{aligned} \textit{determinar} \quad & Y \\ s.a \quad & l_{ij}^2 \leq e_{ij}^t Y e_{ij} \leq u_{ij}^2, \ \ \forall (i,j) \in K \subset \{1,...,m\}^2 \\ & Y = X^t X, \end{aligned} \tag{3.34}$$

onde e_i é o i-ésimo vetor unitário em $\mathbb{R}^m$ e $e_{ij} = e_i - e_j$.

Assim, uma relaxação em programação semidefinida para o problema (3.34) é dada por

$$\begin{aligned} \textit{minimizar} \quad & \mathrm{tr}\,(Y) \\ s.a \quad & l_{ij}^2 \leq e_{ij}^t Y e_{ij} \leq u_{ij}^2, \ \ \forall (i,j) \in K \subset \{1,...,m\}^2 \\ & Y \succeq 0, \end{aligned} \tag{3.35}$$

onde $Y \succeq 0$ significa que Y é uma matriz semidefinida positiva.

A função objetivo do problema (3.35) equivale a $\sum_i \|x_i\|^2$ e, portanto, implica na minimização das normas. Uma vez que, sem perda de generalidade, os pontos x_i podem ter seu centro de gravidade fixado na origem, não há qualquer limitação nessa escolha dessa função objetivo.

O termo relaxação condiz com o fato de que pode não ser possível decompor a solução do problema (3.35), i.e., a matriz $Y \in \mathbb{R}^{n \times n}$ no produto $X^t X$ com $X \in \mathbb{R}^{3 \times m}$, ou de forma equivalente, a matriz Y pode ter posto maior que três. Se este for caso, então toma-se como solução os três autovetores associados aos maiores autovalores da decomposição em valores singulares de Y.

O método proposto por Patrick Biswas e colaboradores foi aplicado em instâncias derivadas do PDB, obtendo soluções com baixo RMSD (inferior a 1 Å) quando os limites l_{ij} e u_{ij} são próximos.

3.4.9 GNOMAD

Em [104], Williams e colaboradores propuseram um algoritmo de otimização global para o MDGP que, além das restrições de distância, utiliza as restrições de van der

Waals[4] para reduzir o espaço de busca das soluções.

Este método baseia-se na construção de esferas ao redor de cada ponto (átomo). E, a cada passo, busca-se reduzir o valor da função objetivo movendo-se um único átomo de tal forma que a nova posição esteja fora das esferas que envolvem os demais átomos e permita reduzir o valor da função objetivo. Estas esferas possuem raio proporcional à força de van der Waals existente entre o ponto que se move e os demais. Com o intuito de aumentar a taxa de convergência, Williams e os demais autores propõem que durante o deslocamento do ponto, evite-se também a esfera de centro na atual posição e raio proporcional ao erro associado ao ponto. Desta forma, pontos (átomos) que ocupem posições inviáveis movem-se mais rapidamente que os pontos próximos da viabilidade. As direções de decréscimo consideradas são as obtidas via iterações do método BFGS e os átomos vão sendo posicionados um a um, i.e., partindo de dois átomos, depois de alcançar a viabilidade, introduz-se um novo átomo, até que todos os átomos tenham sido posicionados corretamente (sejam viáveis).

[4]As forças de van der Waals são forças de curto alcance que exprimem a tendência dos átomos de se repelirem, quando estão muito próximos, e de se atraírem, quando se aproximam de uma distância internuclear ótima [89].

Capítulo 4

MDGP via suavização e penalização hiperbólicas

Os modelos (problemas de programação) matemáticos frequentemente utilizados para o MDGP possuem dois inconvenientes: a não-diferenciabilidade e a existência de muitos mínimos locais [79]. Juntos, eles impedem a aplicação direta dos métodos clássicos e mais robustos de otimização. O primeiro destes inconvenientes, a não-difenciabilidade, advém da presença da função norma Euclideana na definição das restrições de distância e do uso da função $\max\{\cdot, 0\}$ na caracterização das funções de penalidade. Já o elevado número de mínimos locais decorre do forte apelo combinatório do problema, o que se agrava à medida que se aumenta a quantidade de átomos considerados.

A contribuição deste trabalho é um novo algoritmo que combina as técnicas de suavização e penalização hiperbólicas e uma especializada estratégia de dividir-para-conquistar para a solução do problema geométrico da distância molecular. As técnicas de suavização e penalização hiperbólicas propostas originalmente por Xavier em [108] permitem a aplicação de métodos clássicos de otimização ao introduzir diferenciabilidade na formulação do MDGP como um problema de programação matemática e, mais importante ainda, reduzem o número de mínimos locais pelo controle adequado dos parâmetros relacionados. Já a estratégia de dividir-para-conquistar permite que, ao invés de resolver um único e grande problema, possamos atacar uma sequência de problemas menores e, portanto, mais fáceis do que o problema original.

A suavização e penalização hiperbólicas têm se mostrado extremamente eficazes na resolução dos mais diferentes problemas de otimização da forma Minimax envolvendo a norma Euclideana (Determinação de estruturas geométricas [37, 109], empacotamento [110], classificação [80, 81, 111], alocação de recursos [43], entre outros). Os bons resultados obtidos em [109] e [37], por Adilson Elias Xavier e Ana Flávia Macambira, apontaram a viabilidade de aplicação da suavização para determinação de estruturas tridimensionais de proteínas.

4.1 Proposta de suavização

Analisando o problema

$$(P)\ \textit{determinar} \quad x_1, \ldots, x_m \tag{4.1}$$

$$s.a \quad l_{ij} \leq \|x_i - x_j\| \leq u_{ij},\ \ \forall (i,j) \in K \subset \{1, \ldots, m\}^2 \tag{4.2}$$

$$x_i \in \mathbb{R}^3,\ \ \forall i = 1, \ldots, m, \tag{4.3}$$

vemos que a definiçãos das restrições envolve a função $\|\cdot\|$. Isto faz de (P) um modelo não-diferenciável, como a maioria das formulações para o MDGP. Em vista disto, o método que adotamos para resolver o problema (P) faz uso da estratégia de suavização hiperbólica [108]. Nesta abordagem, defini-se a função

$$\theta_\tau(y) = \sqrt{\tau^2 + \sum_{k=1}^{3} y_k^2},\ \ y = (y_1, y_2, y_3) \in \mathbb{R}^3,\ \tau > 0. \tag{4.4}$$

A função θ_τ apresenta as seguintes propriedades imediatas:

(T1) $\lim_{\tau \to 0} \theta_\tau(y) = \|y\|$;

(T2) θ_τ é uma função de classe C^∞;

(T3) $\theta_\tau(y) > \|y\|, \quad \forall \tau > 0$;

(T4) $\tau_1 > \tau_2 \Rightarrow \theta_{\tau_1}(y) > \theta_{\tau_2}(y), \quad \forall y \in \mathbb{R}^3$.

A propriedade $(T1)$ indica que a função θ_τ é uma boa aproximação da função $\|\cdot\|$. O parâmetro τ indica o nível da aproximação, pois à medida que τ tende a

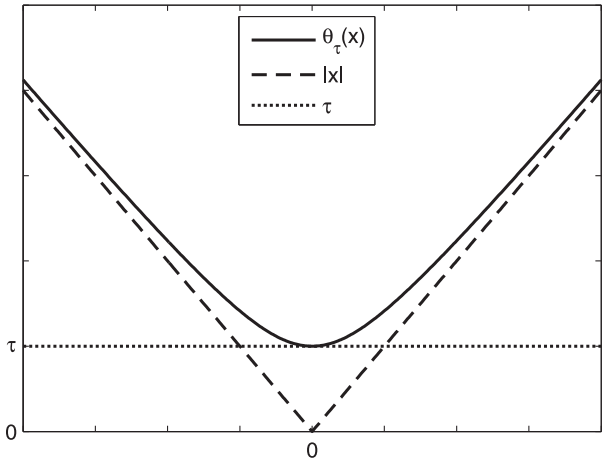

Figura 4.1: O gráfico da suavização hiperpólica $\theta_\tau(x)$ é uma hipérbole equilátera.

zero, a função suavizada θ_τ se aproxima da função original $\|\cdot\|$. Isto fica evidente na Figura 4.1, onde é possível ver com clareza que a distância máxima entre a função original e a função suavizada ocorre na origem e tem valor igual ao do parâmetro τ. O gráfico de $\theta_\tau(x)$, na Figura 4.1, é uma hipérbole equilátera, o que motiva a nomenclatura suavização hiperbólica.

Substituindo a função $\|\cdot\|$ por θ_τ no problema (P), obtemos o problema suavizado (diferenciável)

$$(P_\tau) \quad \textit{determinar} \qquad x_1, \ldots, x_m \tag{4.5}$$

$$s.a \qquad l_{ij} \leq \theta_\tau(x_i - x_j) \leq u_{ij}, \;\; \forall (i,j) \in K \subset \{1, \ldots, m\}^2 \tag{4.6}$$

$$x_i \in \mathbb{R}^3, \;\; \forall i = 1, \ldots, m. \tag{4.7}$$

4.2 Proposta de penalização

Nos métodos de penalização, um problema com restrições é substituído por uma série de problemas irrestritos cujas soluções convergem para a solução do problema original. No caso particular do problema (P), as funções de penalidade frequentemente utilizadas para as restrições de desigualdade envolvem a função $p(y) = \lambda \max\{y, 0\}$, onde λ é um parâmetro (peso) que indica a intensidade da penalização. Um dos inconvenientes no uso da função $\max\{\cdot, 0\}$ é a não-diferenciabilidade na origem. Como

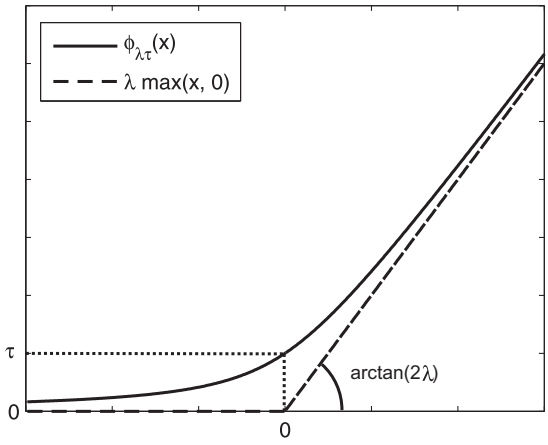

Figura 4.2: No gráfico da função de penalidade $\phi_{\lambda,\tau}$, o parâmetro λ controla a intensidade (inclinação) da penalização, e τ, a suavização.

alternativa, propomos o uso da função de penalidade

$$\phi_{\alpha,\tau}(y) = \left(\frac{1}{2}\tan(\alpha)\right) y + \sqrt{\left(\frac{1}{2}\tan(\alpha)y\right)^2 + \tau^2}, \tag{4.8}$$

onde $\alpha \in (0, \pi/2)$, $\tau > 0$ e $y \in \mathbb{R}$.

Note que

$$\lim_{\tau \to 0} \phi_{\pi/4,\tau}(x) = \max\{x, 0\},$$

ou seja, se $\alpha = \pi/4$, a função $\phi_{\alpha,\tau}$ é uma boa aproximação para $\max\{\cdot, 0\}$.

Substituindo $\tan(\alpha)/2$ por λ, obtemos a forma mais conveniente

$$\phi_{\lambda,\tau}(y) = \lambda y + \sqrt{\lambda^2 y^2 + \tau^2}, \tag{4.9}$$

para a função de penalização hiperbólica.

O gráfico de $\phi_{\lambda,\tau}$ é também uma hipérbole (ver Figura 4.2). O parâmetro τ controla o grau de suavização e λ a intensidade (peso) da penalidade.

A função $\phi_{\lambda,\tau}$ goza das seguintes propriedades:

(P1) $\phi_{\lambda,\tau}$ é uma função de classe C^∞;

(P2) $\phi_{\lambda,\tau}(y) > \max\{y, 0\} \quad \forall \lambda > 0, \tau > 0$;

(P3) $\tau_1 > \tau_2 \Rightarrow \phi_{\lambda,\tau_1}(y) > \phi_{\lambda,\tau_2}(y) \quad \forall y \in \mathbb{R}$;

(P4) $\lambda_1 > \lambda_2 \Rightarrow \phi_{\lambda_1,\tau}(y) > \phi_{\lambda_2,\tau}(y) \quad \forall y \in \mathbb{R}$.

Utilizando a função de penalização hiperbólica $\phi_{\lambda,\tau}$, podemos obter o problema irrestrito

$$(P_{\lambda,\tau}) \;\; minimizar \qquad f_{\lambda,\tau}(x) = \sum_{(i,j)\in K} \phi_{\lambda,\tau}(l_{ij} - \theta_\tau^{ij}) + \phi_{\lambda,\tau}(\theta_\tau^{ij} - u_{ij}), \quad (4.10)$$

$$\text{onde} \qquad \theta_\tau^{ij} = \theta_\tau^{ij}(x) = \theta_\tau(x_i - x_j), \quad (4.11)$$

$$K \subset \{1, \ldots, m\}^2, \quad (4.12)$$

$$x_k \in \mathbb{R}^3, \;\; \forall k = 1, \ldots, m. \quad (4.13)$$

O problema $(P_{\lambda,\tau})$ é infinitamente diferenciável com respeito a x e, portanto, permite a aplicação dos métodos clássicos de otimização. No entanto, deve ser observado que o problema $(P_{\lambda,\tau})$ não é exatamente igual ao problema (P). Assim, para se obter a solução do problema original, propõe-se que seja resolvida uma sequência infinita de problemas suavizados (P_{λ,τ_k}) parametrizados por uma sequência decrescente de parâmetros τ_k, $k = 1, 2, \ldots$, tendendo a zero, ou seja,

$$\tau^{k+1} < \tau_k, \quad (4.14)$$

$$\lim_{k\to+\infty} \tau_k = 0. \quad (4.15)$$

Através desse procedimento, a sequência de problemas suavizados se aproxima gradativamente do problema original. Note que na definição da sequência de problemas P_{λ,τ_k}, utilizamos o mesmo parâmetro λ em todas as parcelas da função objetivo (Isto não é obrigatório e traduz apenas uma escolha dos autores). Na verdade, poderíamos definir, para cada restrição, diferentes valores para os parâmetros λ e τ, o que permitiria controlar, em certa medida, a ordem de cumprimento das restrições. Numa aplicação com dados reais, pode ser de interesse satisfazer exatamente apenas as restrições cujos dados são confiáveis e de forma relaxada (imprecisa) as demais. Utilizando a suavização hiperbólica, podemos cumprir esse contrato apenas modificando os parâmetros λ e τ adequadamente.

4.3 Convexificação

O uso da suavização e penalização hiperbólicas não se justifica exclusivamente pela introdução de diferenciabilidade, já que, no caso de distâncias exatas, a diferenciabilidade poderia ser alcançada de forma mais simples, bastando para isso considerar o quadrado das distâncias e, conseqüentemente, da função norma. A grande vantagem no uso destas funções está relacionada à convexidade, como originalmente apontado por Xavier em [109]. Os resultados abaixo obtidos seguem a argumentação apresentada por Xavier e determinam uma cota superior para o valor mínimo do parâmetro de suavização que convexifica o problema $(P_{\lambda,\tau})$.

Lema 4.1 *A função $\theta_\tau^{ij} : \mathbb{R}^{m\times 3} \to \mathbb{R}$, dada por $\theta_\tau^{ij}(x) = \theta_\tau(x_i - x_j)$, é convexa para todo $\tau > 0$.*

Prova. Verifica-se diretamente que a matriz Hessiana de θ_τ^{ij} possui a seguinte estrutura:

$$\nabla^2\theta_\tau^{ij}(x) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & H & 0 & -H & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -H & 0 & H & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}_{3m\times 3m}, \tag{4.16}$$

onde

$$H = H_\tau^{ij}(x) = \frac{1}{\theta_\tau^{ij}(x)} I_{3\times 3} - \frac{1}{(\theta_\tau^{ij}(x))^3}(x_i - x_j)(x_i - x_j)^T. \tag{4.17}$$

Dado $y \neq 0$ em $\mathbb{R}^3$, tem-se que

$$\begin{aligned} y^T H y &= \frac{1}{\theta_\tau^{ij}(x)} \|y\|^2 - \frac{1}{(\theta_\tau^{ij}(x))^3} \langle y, (x_i - x_j)\rangle^2 && (4.18)\\ &\geq \frac{1}{\theta_\tau^{ij}(x)} \|y\|^2 - \frac{1}{(\theta_\tau^{ij}(x))^3} \|y\|^2 \|x_i - x_j\|^2 && (4.19)\\ &\geq \frac{1}{(\theta_\tau^{ij}(x))^3} \|y\|^2 \left((\theta_\tau^{ij}(x))^2 - \|x_i - x_j\|^2\right) && (4.20)\\ &= \frac{1}{(\theta_\tau^{ij}(x))^3} \|y\|^2 \tau^2 && (4.21)\\ &> 0, && (4.22) \end{aligned}$$

com a última desigualdade sendo estrita pois, por hipótese, $\tau > 0$. Ou seja, H é uma matriz definida positiva.

Finalmente, para qualquer $v = (v_1, v_2, \ldots, v_m) \neq 0$ em $\mathbb{R}^{m\times 3}$, vale

$$\begin{aligned} v^T \nabla^2 \theta_\tau^{ij}(x) v &= v_i^T H v_i - v_i^T H v_j - v_j^T H v_i + v_j^T H v_j && (4.23)\\ &= (v_i - v_j)^T H (v_i - v_j) && (4.24)\\ &\geq 0. && (4.25) \end{aligned}$$

Portanto, $\nabla^2 \theta_\tau^{ij}(x) \geq 0$ e, consequentemente, θ_τ^{ij} é um função convexa.

□

Proposição 4.1 *A função objetivo do problema $(P_{\lambda,\tau})$ é convexa para todo $\tau > \max\{u_{ij} : (i,j) \in K\}$.*

Prova. Por definição, cada uma das parcelas da função objetivo do problema $(P_{\lambda,\tau})$ é dada por

$$\begin{aligned} f_{\lambda,\tau}^{ij}(x) &= \phi_{\lambda,\tau}(l_{ij} - \theta_\tau^{ij}(x)) + \phi_{\lambda,\tau}(\theta_\tau^{ij}(x) - u_{ij}) && (4.26)\\ &= (l_{ij} - u_{ij}) + \sqrt{\lambda^2(\theta_\tau^{ij}(x) - l_{ij})^2 + \tau^2} && (4.27)\\ &+ \sqrt{\lambda^2(\theta_\tau^{ij}(x) - u_{ij})^2 + \tau^2}. && (4.28) \end{aligned}$$

É suficiente provar que $\nabla^2 f_{\lambda,\tau}^{ij}(x)$ é semidefinida positiva para todo $x \in \mathbb{R}^{m\times 3}$. De fato, provaremos que cada uma das parcelas $f_{\lambda,\tau}^{ij}$ é também a soma de funções convexas.

Pelo lema 4.1, $\nabla^2 \theta_\tau^{ij}(x)$ é semidefinida positiva, portanto, para qualquer constante $z \in \mathbb{R}$ tal que

$$z \leq \theta_\tau^{ij}(x), \quad \forall x \in \mathbb{R}^{m\times 3},$$

a função $h_{\lambda,\tau}^{ij}(x) = \sqrt{\lambda^2(\theta_\tau^{ij}(x) - z)^2 + \tau^2}$ será convexa, pois

$$\nabla^2 h_{\lambda,\tau}^{ij}(x) = \frac{\lambda^2\tau^2}{h_\tau^3(x)} \nabla\theta_\tau^{ij}(x) \nabla^t \theta_\tau^{ij}(x) + \frac{(\theta_\tau^{ij}(x) - z)}{h(x)} \nabla^2 \theta_\tau^{ij}(x). \tag{4.29}$$

Note que cada uma das parcelas de $f_{\lambda,\tau}^{ij}(x)$ é dada por

$$f_\tau^{ij}(x) = (l_{ij} - u_{ij}) + h_{\lambda,\tau}^{ij}(\theta_\tau^{ij}(x) - l_{ij}) + h_{\lambda,\tau}^{ij}(\theta_\tau^{ij}(x) - u_{ij}), \tag{4.30}$$

e como, por hipótese, $\theta_\tau^{ij}(x) \geq \tau > \max\{u_{ij} : (i,j) \in K\} \geq \max\{l_{ij} : (i,j) \in K\}$, segue que cada uma das parcelas $f_{\lambda,\tau}^{ij}(x)$ da função objetivo de $(P_{\lambda,\tau})$ é uma função convexa, concluindo a prova.

□

Na prática, a convexificação do problema $(P_{\lambda,\tau})$ obtida através da escolha de valores elevados para o parâmetro τ deve ser evitada, pois

$$\lim_{\tau \to \infty} f_{\lambda,\tau}(x)/\tau = 2\sqrt{\lambda^2 + 1}.$$

Ou seja, para valores excessivamente elevados, a função objetivo $f_{\lambda,\tau}$ se aproxima da constante $2\tau\sqrt{\lambda^2 + 1}$. E, portanto, não é possível estabelecer correlações entre seus mínimos e os mínimos do problema original.

Desta forma, o parâmetro τ deve expressar um compromisso harmônico entre a simplificação (convexificação) do problema e a coerência da solução. Assim, a escolha do valor inicial do parâmetro de suavização τ deve ser feita cautelosamente, buscando convexificar algumas das parcelas presentes na definição da função f_τ, mas evitando valores excessivamente altos que conduzam à degeneração do problema suavizado. A esperança é que mínimos locais com valores distantes do mínimo global sejam removidos pela escolha conveniente do parâmetro de suavização τ.

Na figura 4.3, vemos um exemplo da capacidade convexificadora da presente proposta e sua forte relação com o valor adotado para o parâmetro τ. Para a geração da imagem, consideramos a instância do MDGP definida pelo conjunto de restrições

$$\|x_1 - x_3\| = 1; \tag{4.31}$$

$$\|x_1 - x_6\| = 0; \tag{4.32}$$

$$\|x_2 - x_3\| = 2; \tag{4.33}$$

$$\|x_2 - x_4\| = 2; \tag{4.34}$$

$$\|x_2 - x_5\| = 2; \tag{4.35}$$

$$\|x_4 - x_5\| = 4; \tag{4.36}$$

$$\|x_4 - x_6\| = 3. \tag{4.37}$$

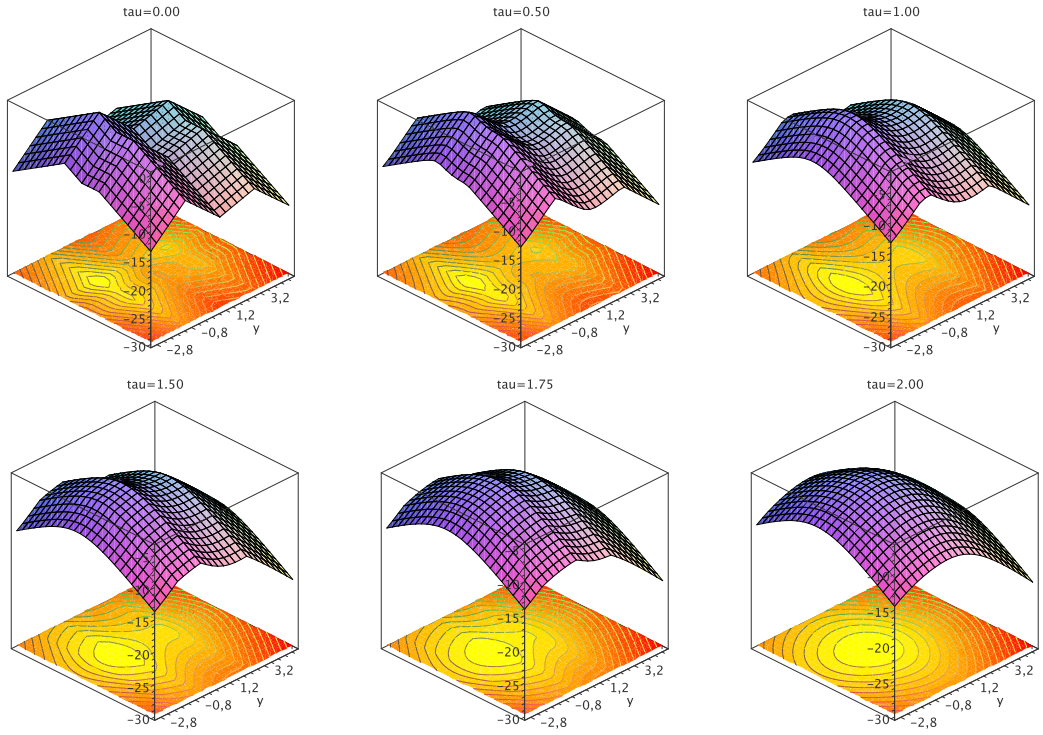

Figura 4.3: À medida que aumentamos o valor do parâmetro τ, mais convexa torna-se a função objetivo.

Além disso, fixamos as variáveis $(x_3, x_4, x_5, x_6) = (1, 1, -3, 0)$ e deixamos livres as variáveis x_1, x_2 com intuito de obter uma representação tridimensional da função objetivo. Para facilitar a visualização das curvas de nível e dos pontos críticos, plotamos o gráfico de $-f(x_1, x_2)$, ou seja, invertemos o sinal da função objetivo. Assim, os mínimos do problema são identificados como máximos nas superfícies representadas.

Na figura 4.3, fica clara a relação entre a convexificação do problema e o valor do parâmetro de suavização τ. À medida que aumentamos o valor de τ, mais convexo torna-se o problema graças à remoção gradativa dos mínimos locais menos atrativos. Também é possível perceber a existência de uma trajetória, com início no minimizador do problema convexo ($\tau = 2,00$) e fim no minimizador global do problema original ($\tau = 0,00$), passando pelos minimizadores globais dos problemas intermediários gerados pela redução do valor do parâmetro τ.

4.4 O algoritmo de suavização e penalização hiperbólicas (SPH)

As propriedades de diferenciabilidade e, principalmente, convexificação da suavização e penalização hiperbólicas são exploradas no Algoritmo 1 para resolução do problema (P). Seguindo a notação proposta em [79], denota-se por $\mathbf{locmin}(f, y, \mathcal{M})$ o minimizador obtido pelo método de minimização local $\mathcal{M}$ assumindo y como ponto inicial.

Algoritmo 1 Suavização e Penalização Hiperbólicas

função $sph(x,\ \lambda,\ \tau,\ \rho) : x \in \mathbb{R}^3,\ \rho \in (0,1),\ \tau > 0,\ \lambda > 0;$
  enquanto$(f_{\lambda,\tau}(x) > \epsilon_f)$
    $x = locmin(f_{\lambda,\tau},\ x,\ \mathcal{M});$
    $\tau = \rho\tau;$
  fim
retorna $(x,\ f_{\lambda,\tau}(x));$

No Algoritmo 1, a rotina de minimização local $\mathcal{M}$ utiliza como ponto inicial o minimizador obtido na iteração anterior. Por trás desta escolha, há a esperança de que uma conveniente escolha do parâmetro de atualização ρ exija um número reduzido de iterações internas da rotina de minimização local. A escolha do valor do parâmetro ρ deve ser feita com cautela, já que está diretamente relacionada à razão entre o número de iterações internas (realizadas pelo método de otimização local) e o número de iterações externas (atualizações do parâmetro τ). Uma escolha por valores baixos para ρ pode implicar não apenas em um elevado número de iterações internas a cada chamada do método $\mathcal{M}$, mas também na possibilidade de desperdício da trajetória engendrada pela convexificação dos problemas suavizados, tornando mais improvável a obtenção de minimizadores globais ao final do processo.

O ponto $x \in \mathbb{R}^{m\times 3}$ tomado como entrada do Algoritmo 1 pode ser gerado de maneira aleatória, mas uma escolha razoável pode reduzir o número de iterações requeridas para a obtenção de um minimizador global. Do ponto de vista teórico, o método de otimização local $\mathcal{M}$ pode ser escolhido com total liberdade já que os parâmetros λ e τ são positivos em todas as iterações e, por isso, os problemas de minimização relacionados são infinitamente diferenciáveis.

É clara a relação entre o método aqui proposto e o método *dgsol* apresentado

por Moré e Wu em [106]. Sendo assim, cabem comentários sobre as semelhanças e diferenças existentes entre essas propostas. Em ambas, a função objetivo do problema original é aproximada por uma seqüência de funções suavizadas com o intuito de obter tanto a diferenciabilidade quanto a redução do número de mínimos locais através da manipulação adequada dos parâmetros de suavização relacionados. Contudo, cabe ressaltar a simplicidade da presente proposta, já que a estratégia de suavização aqui empregada permite que a função suavizada $f_{\lambda,\tau}$ seja explicitamente avaliada sem que se requeira qualquer esforço computacional extra. O mesmo não pode ser dito a respeito da proposta de Moré e Wu, onde o cálculo do valor da função objetivo suavizada requer a aplicação de métodos de integração numérica.

4.5 Aliando o algoritmo SPH e a técnica de dividir-para-conquistar

Utilizando as técnicas de suavização e penalização hiperbólicas é possível reduzir consideravelmente o número de mínimos locais dos problemas associados. Contudo, em problemas que envolvam um elevado número de átomos, a quantidade remanescente de mínimos locais pode ser grande, o que torna pouco provável que uma rotina de minimização local consiga determinar um minimizador global para o problema partindo de um ponto longe de uma solução ótima. Em vista disto, propomos a rotina *sphdc* que combina os procedimento *sph* e a técnica de dividir-para-conquistar para controlar o tamanho dos problemas para os quais bons pontos iniciais não estejam disponíveis.

A técnica de dividir-para-conquistar ($D\&C$) é um importante modelo conceitual de algoritmo [63, 25, 32]. Tipicamente, um algoritmo que aplica a técnica $D\&C$ para solucionar um problema grande e/ou complexo divide repetidamente o problema original em subproblemas menores. Este procedimento de divisão é repetido até que os subproblemas gerados sejam suficientemente pequenos para serem resolvidos facilmente. Uma vez que as soluções dos subproblemas estejam disponíveis, inicia-se a fase de combinação destas soluções para se obter uma solução do problema original.

Os dois processos mais importantes para se definir um algoritmo que utilize a

técnica *D&C* são os métodos de divisão de problemas e combinação de soluções. O método de divisão especifica o momento e a forma pela qual os problemas são decompostos. Já o método de combinação determina o modo de agrupar/harmonizar as soluções dos subproblemas para obter uma solução para o problema original.

Na Figura 4.4a, vemos a matriz de conectividade de uma instância do MDGP derivada da geometria conhecida da proteína 1PTQ disponibilizada no *Protein Data Bank*. Um elemento (i, j) desta matriz tem valor diferente de zero quando existe uma restrição envolvendo a distância $\|x_i - x_j\|$, e zero, caso contrário. Ou seja, é uma matriz de conectividade derivada das restrições do problema.

Apesar de ser uma instância particular, alguns elementos desta matriz são tipicamente encontrados em todas as instâncias do MDGP consideradas na literatura. Uma destas características típicas é a presença de uma diagonal por blocos bastante densa. De fato, na Figura 4.4b, onde destacamos uma pequena região sobre a diagonal, podemos ver que existem blocos com elevado número de restrições (pontos). A alta densidade de restrições nestes blocos reduz o número de possíveis soluções para os problemas a eles restritos. Assim, parece natural considerá-los como blocos elementares (mínimos) na definição do método de decomposição da estratégia *D&C* aplicada ao MDGP. Uma questão natural que se impõe a esta proposta de decomposição é forma de caracterizar sistematicamente estes blocos elementares.

Como dito anteriormente (Ver capítulo 2), as proteínas são compostos químicos formados por uma cadeia linear de resíduos que pode ser determinada via *espectroscopia de massa*[1]. Observando os índices dos átomos nas instâncias tipicamente consideradas na literatura, percebemos que os blocos mais densos da matriz de conectividade são aqueles formados pelas restrições envolvendo átomos em um mesmo resíduo (Ver Figura 4.5, onde as restrições relacionadas aos resíduos são destacadas). Assim, propomos uma decomposição natural das instâncias do MDGP em blocos que envolvem restrições entre átomos em um mesmo resíduo.

Note que podemos representar cada um dos n resíduos de uma dada proteína R

[1]Um espectrômetro de massa é um instrumento que mede a razão massa/carga de moléculas eletricamente carregadas. Esta informação permite estabelecer o peso molecular e as estruturas dos compostos analisados [21].

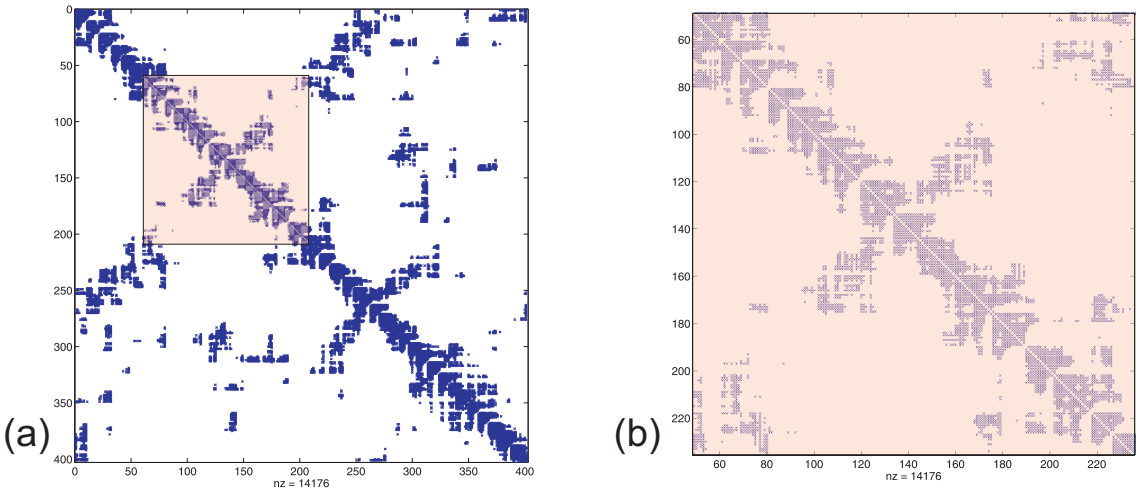

Figura 4.4: (a) A matriz de conectividade da instância derivada da geometria conhecida da proteína 1PTQ. (b) Na região destacada, observa-se blocos com alta densidade sobre a diagonal da matriz de conectividade.

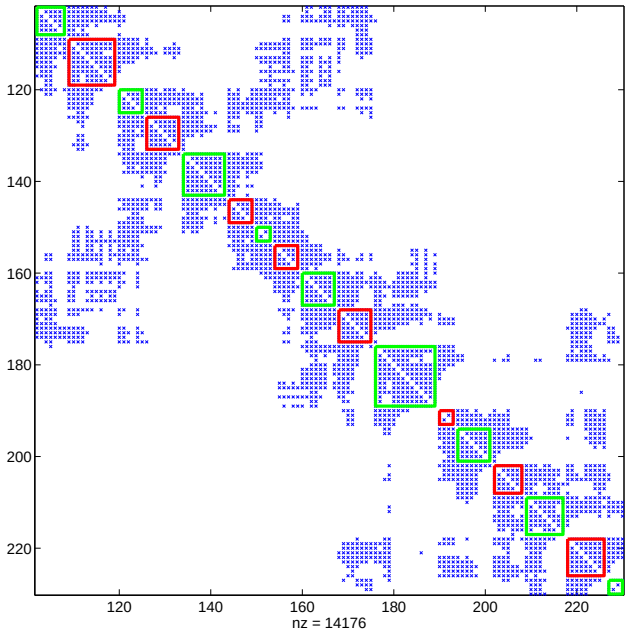

Figura 4.5: Os blocos destacados estão relacionados às restrições envolvendo átomos no mesmo resíduo. Estes blocos são os mais densos nas matrizes de conectividade tipicamente encontradas na literatura.

pelos conjuntos da forma

$$R(k) = \{x_i \in \mathbb{R}^3 : i \in \mathcal{I}(k)\} \text{ com } k = 1, \ldots, n, \tag{4.38}$$

onde $\mathcal{I}(k) \subset \{1, 2, \ldots, m\}$ é um subconjunto dos índices dos m átomos presentes na proteína R.

Uma vez que cada átomo pertence a um único resíduo, temos:

$$R = \bigcup_{k=1}^{n} R(k) \quad \text{e} \quad R(i) \cap R(j) = \emptyset, \forall i \neq j. \tag{4.39}$$

Com esta defingão, podemos decompor o problema original em subproblemas do seguinte tipo:

$$(P(\mathcal{K})) \quad \textit{determinar} \quad \bigcup_{k \in \mathcal{K}} R(k) \tag{4.40}$$

$$s.a \qquad l_{ij} \leq \|x_i - x_j\| \leq u_{ij}, \tag{4.41}$$

$$\forall (i,j) \in \bigcup_{k \in \mathcal{K}} \mathcal{I}(k). \tag{4.42}$$

Ou seja, em cada um dos subproblemas $P(\mathcal{K})$, o conjunto $\mathcal{K}$ indica quais resíduos $R(k)$ fazem parte do subproblema e as restrições consideradas são aquelas com átomos cujos índices pertencem ao conjunto $\bigcup_{k \in \mathcal{K}} \mathcal{I}(k)$. Com esta definição, no caso de uma proteína formada por n resíduos, o problema original é representado por $P(\{1, 2, ..., n\})$.

A fim de simplificar a notação, representaremos por $\mathcal{I}(\mathcal{K})$ a união de todos os conjuntos de índices $\mathcal{I}(k)$ com $k \in \mathcal{K}$, ou seja,

$$\mathcal{I}(\mathcal{K}) = \bigcup_{k \in \mathcal{K}} \mathcal{I}(k). \tag{4.43}$$

4.5.1 Combinando soluções

Na seção 3.2, apresentamos a medida *RMSD* para comparação de estruturas tridimensionais. Agora, apresentaremos um algoritmo que, utilizando a matriz ótima de rotação Σ obtida como acessório para o cálculo da medida *RMSD*, permite combinar eficientemente as soluções de subproblemas da forma (4.40), desde que estes possuam

alguma interseção. Com este procedimento, como veremos, é possível obter um bom ponto inicial para o problema combinado a partir das soluções de subproblemas.

De fato, se $X = \{x_i\}$ é uma solução de $P(\mathcal{K})$ e $X' = \{x'_i\}$, de $P(\mathcal{K}')$ com

$$\mathcal{K} \cap \mathcal{K}' \neq \emptyset, \tag{4.44}$$

então o conjunto de resíduos cujos índices pertencem a interseção $\mathcal{K} \cap \mathcal{K}'$ são representados tanto em X quanto em X'. Definindo, respectivamente, Y e Y' como sendo as representações das interseções nas soluções X e X', podemos obter uma estrutura combinada Z através do seguinte procedimento:

(A) aplicar em X e X' a translação que iguala à origem os centros de Y e Y', ou seja,

$$x_i = x_i - c_i, \forall x_i \in X \supset Y \tag{4.45}$$

$$x'_i = x'_i - c'_i, \forall x'_i \in X' \supset Y', \tag{4.46}$$

onde

$$c_i = \sum_{y \in Y} \frac{y}{|Y|}, \tag{4.47}$$

$$c'_i = \sum_{y' \in Y'} \frac{y'}{|Y'|}. \tag{4.48}$$

(B) rotacionar toda a estrutura transladada X' utilizando a matriz de rotação que define a medida $RMSD$ das substrutras Y e Y'. Explicitamente,

$$U\Sigma V^t = svd(Y^t Y); \tag{4.49}$$

$$X' = X'UV^t. \tag{4.50}$$

(C) a estrutura combinada $Z = \{z_i : i \in \mathcal{R}(\mathcal{K} \cup \mathcal{K}')\}$ resultante é dada por

$$z_i = \begin{cases} x_i, \text{ se } i \in \mathcal{I}(\mathcal{K} - \mathcal{K}') \\ x'_i, \text{ se } i \in \mathcal{I}(\mathcal{K}' - \mathcal{K}) \\ (x_i + x'_i)/2, c.c. \end{cases} \tag{4.51}$$

Note que os elementos da matriz diagonal Σ associada ao $RMSD$ indicam se a combinação das estruturas é coerente ou não: quanto menores forem os valores dos elementos da diagonal, menor é a discrepância entre as estruturas obtidas para os resíduo na interseção dos dois subproblemas.

No algoritmo *combinar* (Algoritmo 2), estas idéias são aplicadas para obter uma aproximação Z para a solução do problema $P(\mathcal{K} \cup \mathcal{K}')$, partindo das soluções X e X' dos subproblemas $P(\mathcal{K})$ e $P(\mathcal{K}')$. Observe que o problema combinado possui restrições que nenhum dos subproblemas considerados individualmente possui e, portanto, possui informações que podem ser aplicadas à estrutura Z com o intuito de aumentar sua qualidade. Por isso, na penúltima linha da rotina *combinar*, há uma chamada do método de otimização local $\mathcal{M}$ para o problema combinado tomando como ponto inicial a estrutura combinada Z (Isto é feito para refinar a solução).

Algoritmo 2 Combinação de estruturas

função $combinar(X,\ X',\ P(\mathcal{K} \cup \mathcal{K}'))$
  $Y = \{x_i \in X : i \in \mathcal{I}(\mathcal{K} \cap \mathcal{K}')\}$;
  $Y' = \{x_i \in X' : i \in \mathcal{I}(\mathcal{K} \cap \mathcal{K}')\}$;
  $w = \sum_{y \in Y} y/\|Y\|$; //centróide
  $w' = \sum_{y \in Y'} y/\|Y'\|$; //centróide
  $Y = Y - w$; //translação para origem
  $Y' = Y' - w'$; //translação para origem
  $U\Sigma V^t = svd(Y^t Y')$;
  $Q = UV^t$;
  $X = (X - w)Q$;
  $X' = X' - w'$;
  $Z = \{z_i :$ conforme eq. (4.51)$\}$;
  $[Z,\ f_Z] = locmin(P(\mathcal{K} \cup \mathcal{K}'),\ Z,\ \mathcal{M})$; //refinamento
retorna $(Z,\ \Sigma)$

4.5.2 Dividindo problemas

A estratégia de divisão do problema original em problemas menores é um elemento crítico na viabilidade de obtenção de subproblemas cujas soluções sejam compatíveis, i.e., soluções que possam ser combinadas. Nesta seção, apresentaremos duas diferentes estratégias de decomposição. A primeira delas, a divisão binária, tem caráter apenas didático e servirá como base para o aprofundamento da exposição.

Assumindo que os resíduos são dispostos sequencialmente na molécula de proteína, uma alternativa de decomposição do problema original é a divisão binária.

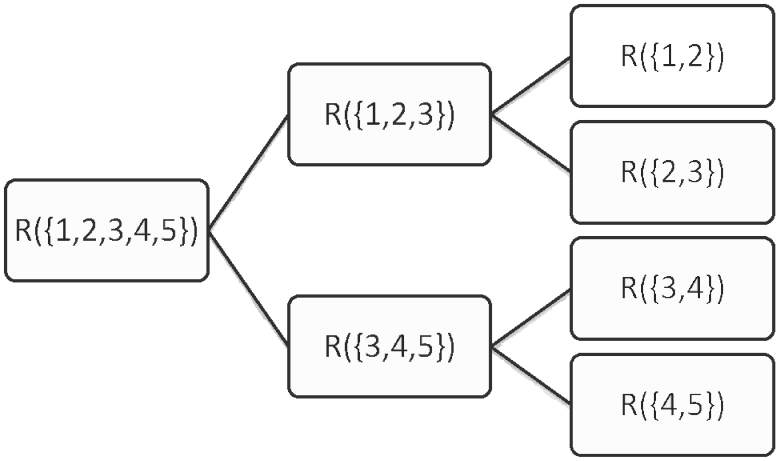

Figura 4.6: Na divisão binária, os grupos são divididos recursivamente até que existam apenas dois resíduos em cada grupo.

Neste caso, partindo do conjunto de todos os resíduos, cada conjunto é divido ao meio dando origem a dois novos grupos. E, para que suas soluções possam ser combinadas, os grupos em cada par gerado compartilham ao menos um resíduo. O processo é repetido até que cada um dos grupos possua um número mínimo (dois) de resíduos (Ver Figura 4.6).

Na Figura 4.7a, vemos a matriz de conectividade de uma instância do MDGP com 100 átomos e 10 resíduos obtida a partir de um fragmento da molécula 1GPV e todas as distâncias inferiores a 6Å. Um ponto (i, j) nesta matriz indica a existência de restrição envolvendo o i-ésimo e o j-ésimo átomos. E na figura 4.7b, vemos a matriz de conectividade dos resíduos. Um ponto (i, j) nesta matriz indica a existência de uma restrição envolvendo um átomo no resíduo $R(i)$ e um átomo no resíduo $R(j)$. Para esta instância em particular, a divisão binária parece ser uma boa alternativa, pois cada subproblema possui elevado número de restrições, o que sugere existir um pequeno conjunto de soluções possíveis. Naturalmente, quanto menor for o número de soluções possíveis para cada subproblema, maior será a chance de coerência na combinação de suas soluções.

Já na Figura 4.8a, vemos a matriz de conectividade de uma instância com 1003 átomos e 130 resíduos obtida a partir dos dados da proteína 1AX8. Esta instância aponta uma das fraquezas da decomposição binária: os resíduos consecutivos $R(22)$ e $R(23)$ não possuem conexão, ou seja, não existe restrição envolvendo um átomo em $R(22)$ e um átomo em $R(23)$. Como consequência, existirá uma infinidade de soluções possíveis para o subproblema $P(\{22, 23\})$. Isto decorre da indiferença da

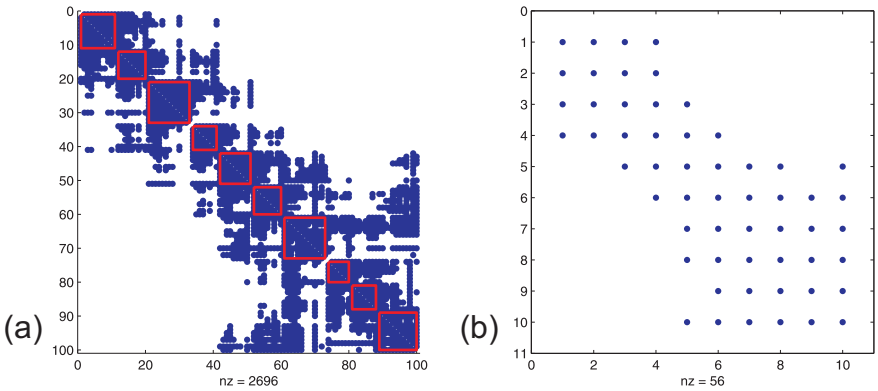

Figura 4.7: Na figura à esquerda, vemos a matriz de conectividade dos átomos. À direita, a matriz de conectividade dos resíduos.

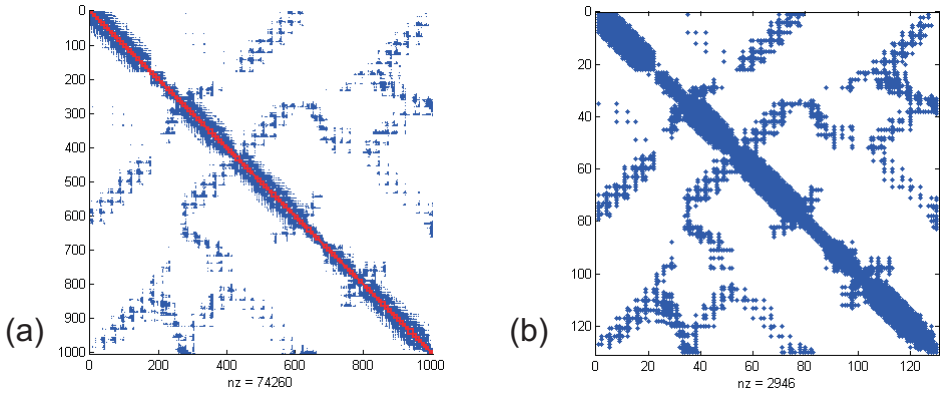

Figura 4.8: Na figura à esquerda, vemos a matriz de conectividade dos átomos. À direita, a matriz de conectividade dos resíduos.

decomposição binária com respeito ao número de soluções possíveis para cada subproblema, na medida em que ignora o número de restrições em cada um deles.

Essa fragilidade pode ser superada, considerando a matriz de conectividade dos resíduos no momento da formação dos pares que irão compor cada um dos subproblemas. O objetivo é definir subproblemas com o maior número possível de restrições. Com isto, reduzimos o número possível de soluções para cada um deles e, como consequência, aumentamos a probabilidade de obtenção de soluções compatíveis.

Seja G um grafo, onde cada vértice $g(i)$ está associado a um resíduo $R(i)$ e cada aresta $g(i, j)$ representa um subproblema a ser resolvido, considerando apenas as restrições envolvendo os átomos do resíduo $R(i)$ e os do resíduo $R(j)$. O peso de cada aresta de G é dado pelo número de restrições envolvendo os átomos em cada vértice.

A idéia por trás da estratégia que propomos é a construção de um subgrafo co-

nexo[2] $S \subset G$, que tenha peso máximo e número mínimo de arestas. A racionalidade da definição de S decorre dos seguintes argumentos:

(1) A imposição de conexidade é necessária, pois, do contrário, cada componente de S seria independente, podendo ser rotacionada e/ou transladada sem violar qualquer restrição, ou seja, haveria uma inifinidade de soluções para os subproblemas e nenhum critério para combiná-las.

(2) O peso de S é definido pelo número de restrições em cada subproblema (aresta), portanto quanto maior for, menor será a quantidade de minimizadores globais em cada subproblema.

(3) Cada aresta de S representa um subproblema a ser resolvido, portanto, ao reduzir o número de arestas, mantendo a conexidade, reduz-se o número de subproblemas redundantes. Por exemplo, com as soluções dos subproblemas (arestas) $g(1,2) = P(\{1,2\})$ e $g(2,3) = P(\{2,3\})$, é possível obter um bom ponto inicial para o problema $g(1,2,3) = P(\{1,2,3\})$, logo o subproblema $g(1,3) = P(\{1,3\})$ passa a ser redundante.

O subgrafo $S \subset G$ com as características citadas é uma *árvore geradora máxima*[3] de G. O problema de obtenção da árvore geradora máxima é um problema eficientemente resolvido [67, 86, 92].

Na Figura 4.9, vemos os grafos relacionados à decomposição binária e o obtido pela decomposição com a árvore geradora máxima (agm) para a instância com 100 átomos. É possível notar que o acoplamento (peso das arestas) nos subproblemas gerados pela estratégia com a agm são maiores do que os obtidos pela decomposição binária, mesmo nesta instância com apenas 10 resíduos. Na instância com 1003 átomos, o resultado á ainda mais interessante, já que os subproblemas obtidos formam um grafo conexo (uma única componente, algo imprescíndivel para obtenção de uma solução combinada).

[2]Um grafo não-vazio S é dito *conexo* se existe um caminho ligando qualquer par de seus vértices.

[3]Uma árvore geradora S de G é um subgrafo conexo e sem ciclos que contém todos os vértices de G. Uma árvore geradora máxima é uma árvore cuja soma dos pesos de suas arestas é a maior possível.

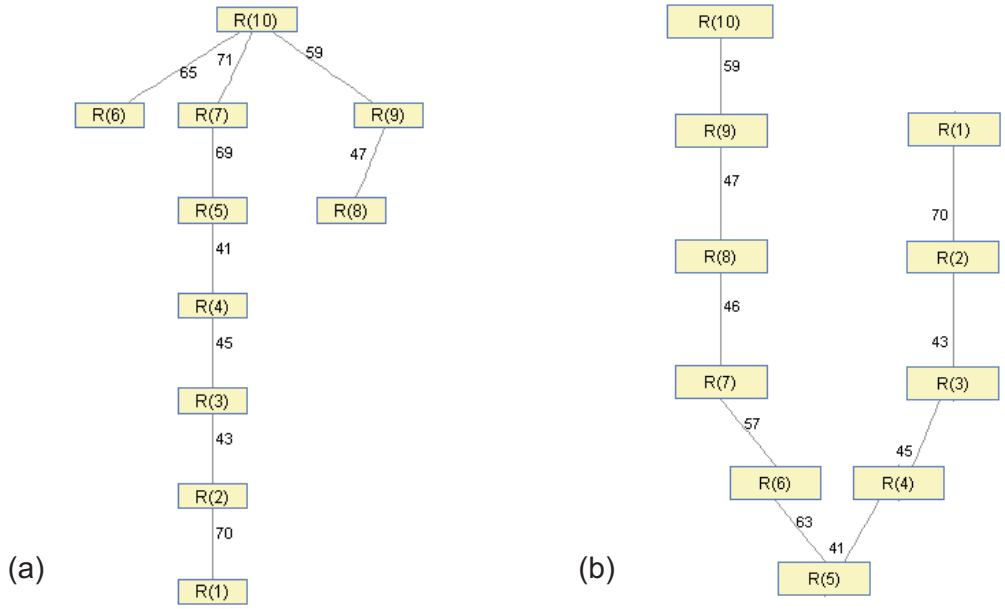

Figura 4.9: Na figura à esquerda, vemos o grafo obtido aplicando a árvore geradora máxima e, à direita, o grafo obtido pela divisão binária. A divisão com a árvore geradora máxima produz um conjunto de subproblemas (arestas) com maior acoplamento e, consequentemete, menor número de soluções.

4.5.3 O método *sphdc*

No algoritmo *sphdc* (Ver Algoritmo 3), incorporamos a suavização e penalização hiperbólicas, as estratégias de divisão de problemas e a combinação de soluções apresentadas. Inicialmente, aplicando a estratégia de divisão com a árvore geradora máxima (agm), dividimos o problema original em um grafo G de subproblemas $g(i,j)$, onde cada um deles envolve apenas dois resíduos. Em seguida, aplica-se a rotina *sph* em cada um dos subproblemas até que uma solução $X(i,j)$, com a precisão desejada, seja obtida. Ao final do procedimento, todas as soluções $X(i,j)$ geradas são combinadas em uma estrutura Z através da rotina *combinar* (Ver Figura 4.10 para uma instância com apenas três resíduos).

Algoritmo 3 Suavização e penalização hiperbólicas com $D\&C$

```
função sphdc(P(K)
    G = agm(P(K)
    para cada g(i, j) em G
        X(i,j) = sph(g(i,j))
    fim
    Z = combinar(X, P(K));
retorna Z
```

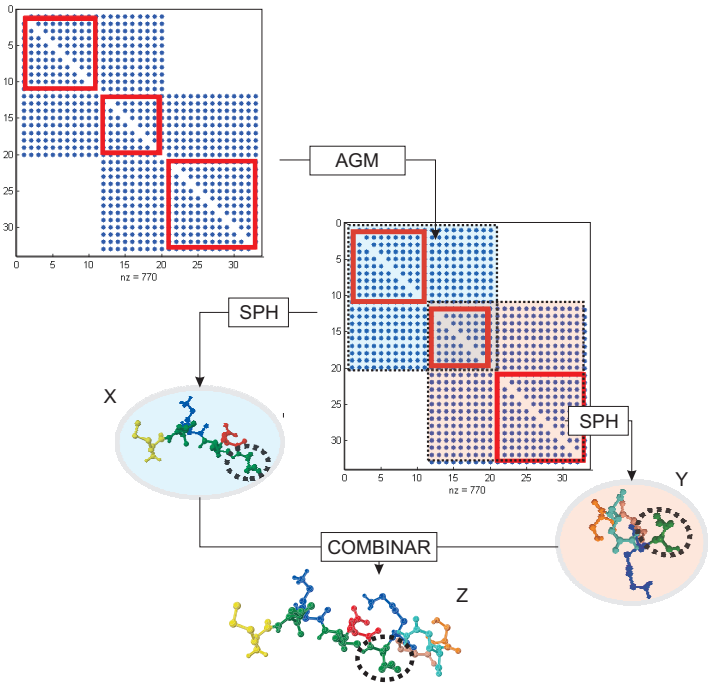

Figura 4.10: Inicialmente, o problema original é dividido. Em seguida, a rotina *sph* é aplicada a cada um dos subproblemas. As soluções X e Y são combinadas em uma estrutura Z pela rotina *combinar*, utilizando a medida *rmsd* como critério de coerência da combinação das soluções.

É importante salientar a separação existente entre a estratégia de dividir-para-conquistar aqui apresentada e as técnicas de suavização e penalização hiperbólicas. Da mesma forma que a suavização e penalização hiperbólicas podem ser combinadas com qualquer método de otimização local (BFGS, gradiente conjugado, método do gradiente, entre outros), a estratégia de dividir-para-conquistar aqui apresentada pode ser combinada com qualquer heurística ou método de otimização global.

Capítulo 5

Experimentos computacionais

Neste capítulo, apresentaremos os resultados dos experimentos numéricos realizados para validar os procedimentos *sph* e *sphdc*. Os experimentos foram divididos em dois grupos. No primeiro, testamos a capacidade de obtenção de mínimos globais da rotina *sph* nas mesmas instâncias com 100 e 200 átomos utilizadas por Moré e Wu em [79], para a validação do método *dgsol*. No segundo grupo de experimentos, procuramos validar o procedimento *sphdc*, para a resolução de instâncias do $MDGP$ envolvendo de 329 a 4200 átomos. Os resultados obtidos no segundo grupo de experimentos foram comparados aos obtidos por An com o algoritmo *cgdca* em [5] e aos obtidos por Di Wu e Zhijun Wu em [105]. Em todos os experimentos, foram utilizados dados de proteínas reais disponibilizados pelo PDB (*Protein data bank*) [10].

Os procedimentos *sph* e *sphdc* foram implementados em linguagem Matlab e C, à exceção da rotina de minimização local *va*35, codificada em linguagem FORTRAN e disponibilizada pela Harwell System Library (www.cse.scitech.ac.uk/nag/hsl/). A rotina *va*35 implementa o método BFGS com memória limitada [72]. A rotina utilizada na etapa de geração da árvore geradora mínima foi a *graphminspantree*, disponbilizada no Matlab e que implementa as idéias apontadas em [86]. Utilizamos um computador com processador Intel Core 2, CPU 6320 de 1,86 GHz, 4,096 GB de memória RAM e sistema operacional Linux-64bits em todos os experimentos.

5.1 Validando o procedimento *sph*

No primeiro grupo de experimentos, realizamos testes com dados derivados da estrutura tridimensional de fragmentos formados pelos 100 e 200 primeiros átomos da cadeia A da proteína 1GPV [53, 93]. Para cada fragmento, foi gerado um conjunto de restrições tomando apenas átomos no mesmo resíduo ou em resíduos vizinhos. Formalmente,

$$K = \{(i,j) : x_i \in R(k), x_j \in R(k) \cup R(k+1)\}, \tag{5.1}$$

onde $R(k)$ representa o k-ésimo resíduo.

Assim como Moré e Wu, consideramos como solução do MDGP qualquer conjunto de coordenadas $x \in \mathbb{R}^{m\times 3}$ tal que

$$(1-\tau_d)l_{ij} \leq \|x_i - x_j\| \leq (1+\tau_d)u_{ij}, \quad (i,j) \in K, \tag{5.2}$$

para uma tolerância $\tau_d = 10^{-2}$. Esta tolerância refletia a precisão disponível para o comprimento das ligações à época dos experimentos realizados por Moré e Wu [39, 79].

Os limites l_{ij} e u_{ij} foram gerados pelas equações

$$lij = (1-\varepsilon)\|x_i - x_j\|, \quad u_{ij} = (1+\varepsilon)\|x_i - x_j\|. \tag{5.3}$$

Ao todo, foram geradas 8 instâncias, 4 para cada fragmento, fazendo ε assumir os valores 0,04; 0,08; 0,12 e 0,16.

Em todos os experimentos, os parâmetros do método *sph* foram fixados em $\lambda = 1$, $\rho = 0.99$ e o valor do parâmetro de suavização τ foi definido em tempo de execução como sendo o terceiro quatil[1] da sequência gerada pela ordenação crescente do conjunto $\{(l_{ij}+u_{ij})/2 : (i,j) \in K\}$. Com essa escolha para o valor do parâmetro τ, pela Proposição 4.1, aproximadamente 75% das parcelas da função objetivo do problema $(P_{\lambda,\tau})$ serão convexas.

[1]Quando uma sequência crescente é divida em quatro partes iguais, os pontos da divisão são chamados de quartis. O terceiro quartil ou quartil superior possui valor maior que aproximadamente 75% dos elementos da sequência.

5.1.1 Redução de mínimos de locais

No primeiro experimento, comparamos a capacidade de obtenção de mínimos globais da rotina *sph* com a da rotina de otimização local *va*35 aplicada isoladamente. A instância utilizada foi gerada a partir do fragmento com os 100 primeiros átomos da proteína 1GPV, tomando $\varepsilon = 0,0$ na equação (5.3). Foram tomados como pontos iniciais 100 pontos selecionados aleatoriamente no interior do cubo de lado 100. O critério qualitativo utilizado para comparação das soluções foi o valor da função

$$f(x) = \sum_{(i,j)\in K} \max\{l_{ij} - \|x_i - x_j\|, 0\} + \max\{\|x_i - x_j\| - u_{ij}, 0\}. \qquad (5.4)$$

O uso correto da rotina *va*35 exige suficiente diferencibilidade da função a ser minimizada. Por isso, substituímos a função original não-diferenciável dada na eq. (5.4) pela aproximação diferenciável $f_{\lambda,\tau}$, conforme definido na eq. (4.13) com $\lambda = 1$ e $\tau = 10^{-6}$. A Figura 5.1 mostra os resultados obtidos pelas rotinas *sph* e *va*35 para cada um dos 100 diferentes pontos iniciais.

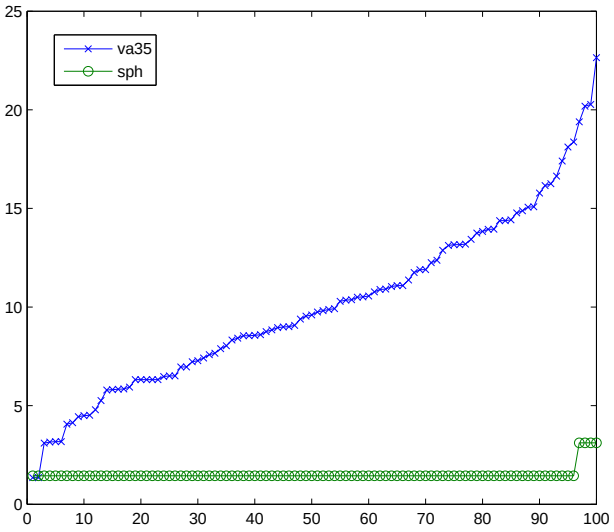

Figura 5.1: Os valores de f nas soluções geradas pela rotina *va*35 e nas soluções geradas pela rotina *sph*.

A rotina *va*35 que implementa o método de minimização local BFGS encontrou apenas uma solução e 87 diferentes[2] mínimos locais. Esta baixa frequência de

[2]Consideramos como sendo diferentes as soluções cuja diferença entre os valores da função objetivo nelas avaliadas foi maior que 0,001.

mínimos globais e elevado número de minimizadores locais distintos em uma pequena instância é um forte indicativo da dificuldade do problema. Já a suavização e penalização hiperbólicas implementadas na rotina *sph* permitiram evitar, em 96% das tentativas, a convergência para um minimizador local, o que aponta a capacidade convexificadora (redução de mínimos locais) da presente proposta. Na Figura 5.2, vemos que os valores dos erros máximos associados às soluções obtidas pela rotina *va*35 são consideravelmente maiores que os associados às soluções obtidas via *sph*.

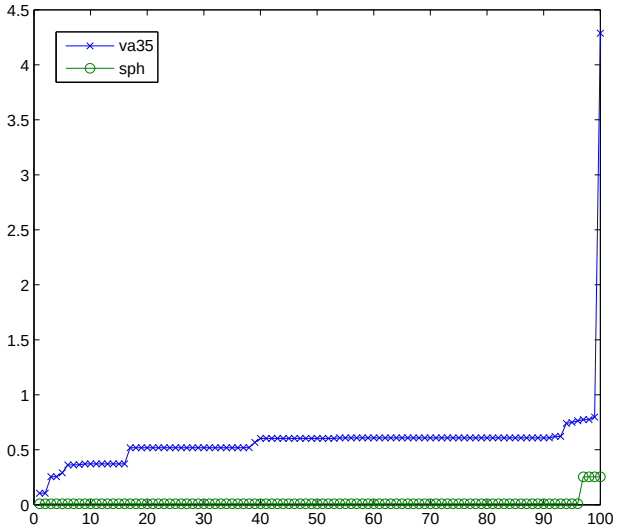

Figura 5.2: Os erros máximos presentes nas soluções geradas pela rotina *va*35 e *sph*. Os erros associados às soluções obtidas pela rotina sph foram consideravelmente menores do que os obtidos pela rotina *va*35 aplicada individualmente.

5.1.2 Robustez

No segundo experimento, comparamos a performance dos métodos *dgsol*, *sph* e *va*35 em problemas com 100 e 200 átomos. Fizemos o parâmetro ε de relaxamento das restrições assumir os valores $0,04; 0,08; 0,12$ e $0,16$. Em cada caso, testamos os métodos *sph* e *va*35 tomando novamente como pontos iniciais 100 pontos aleatoriamente distribuídos em um cubo de lado 100. Os resultados obtidos pelo método *dgsol* foram retirados de [79]. Assim como Moré e Wu em [79], consideramos como solução qualquer estrutura $x \in \mathbb{R}^{m\times 3}$ que satisfez a restrição (5.2) para $\tau_d = 10^{-2}$. Os resultados aparecem na Tabela 5.1.

100 átomos				200 átomos			
ε	$va35$	$dgsol$	sph	ε	$va35$	$dgsol$	sph
0,04	2	80	89	0,04	0	41	75
0,08	2	74	90	0,08	1	66	68
0,12	4	100	100	0,12	0	97	100
0,16	67	100	100	0,16	14	100	100

Tabela 5.1: Frequência de obtenção de soluções pelos métodos $dgsol$, $va35$ e sph.

A Tabela 5.1 mostra que, nos experimentos realizados, o método sph obteve maior frequência de mínimos globais do que os métodos $dgsol$ e $va35$. A elevada frequência de mínimos globais obtidos pelo método sph, comparada à baixa frequência do procedimento $va35$, indica a capacidade de convexificação (redução de mínimos locais) e uma certa independência com respeito à qualidade (profundidade) do ponto inicial. Outro ponto importante é a evidência de simplificação do problema à medida que ocorre o relaxamento das restrições com o aumento do parâmetro ε, já que todos os métodos apresentaram melhora substancial à medida que as restrições foram relaxadas.

5.1.3 Coerência das soluções

No terceiro experimento, avaliamos a coerência das soluções geradas pelos métodos $dgsol$ e sph medindo o desvio entre estas estruturas e a estrutura original obtida no PDB. A medida utilizada para comparação das estruturas foi o

$$RMSD(x,y) = \min\left\{\left(\frac{1}{m}\sum_{i=1}^{m}\|y_i - Qx_i\|^2\right)^{1/2} : Q \in \mathbb{R}^{3\times 3}\right\}, \qquad (5.5)$$

onde as estruturas x e y possuem m átomos e centro de massa na origem (Veja a seção 3.2 para maiores detalhes sobre o cálculo da medida $RMSD$). Os resultados obtidos são exibidos nas Tabelas 5.2 e 5.3.

As estruturas geradas pelos métodos $dgsol$ e sph foram bastante similares à estrutura original do fragmento com 100 átomos. Nas instâncias com 200 átomos, embora ambas as propostas tenham obtido soluções similares, o método $sphdc$ gerou, em média, estruturas mais próximas da original. Contudo, a proximidade entre as estruturas originais e as soluções obtidas pelos métodos parece ser mais um resultado do conjunto de restrições de cada instância do que propriamente uma propriedade

$RMSD$ - Fragmento com 100 átomos				
	dgsol		*sph*	
ε	*min*	*med*	*min*	*med*
0,04	0,063	0,067	0,055	0,057
0,08	0,11	0,12	0,104	0,107
0,12	0,27	0,60	0,155	0,311
0,16	0,37	1,0	0,224	0,325

Tabela 5.2: Desvio (RMSD) entre as coordenadas das soluções obtidas pelos métodos *dgsol* e *sph* e as coordenadas originais do fragmento com 100 átomos.

$RMSD$ - Fragmento com 200 átomos				
	dgsol		*sph*	
ε	*min*	*med*	*min*	*med*
0,04	1,5	1,7	1,5	1,6
0,08	1,5	1,9	0,9	1,6
0,12	1,4	2,2	0,9	2,1
0,16	0,7	2,9	1,0	2,2

Tabela 5.3: Desvio (RMSD) entre as coordenadas das soluções obtidas pelos métodos *dgsol* e *sph* e as coordenadas originais do fragmento com 200 átomos.

dos métodos, já que ambos lidam apenas com as restrições e ignoram qualquer informação adjacente.

5.2 Validando o procedimento *sphdc*

Apesar de reduzir consideravelmente o número de mínimos locais, como verificado nos experimentos da seção anterior, a suavização e penalização hiperbólicas implementadas no método *sph* possuem capacidade de redução limitada e não acompanham o crescimento exponencial dos mínimos locais decorrentes do crescimento das instâncias. De fato, com instâncias envolvendo mais de 1000 átomos, não fomos capazes de encontrar uma solução depois de cem tentativas. Portanto, fica claro que, à medida que cresce o número de átomos nas instâncias, mais improvável passa ser encontrar um minimizador global aplicando a rotina *sph* isoladamente. Daí a necessidade de controlar o tamanho dos problemas de otimização a serem resolvidos e, para este fim, utilizamos a rotina *sphdc*.

Nesta seção, procuramos validar o procedimento *sphdc* (*sph* + dividir-para-conquistar) para determinação de estruturas tridimensionais de proteínas com elevado número de átomos. Os resultados obtidos pelo procedimento *sphdc* foram

comparados aos resultados obtidos pelo método *cgdca*, divulgados por An em [5]. O método *cgdca*, assim como o método *dgsol*, utiliza a transformada Gaussiana para obter uma formulação diferenciável do MDGP cuja solução é obtida através da resolução de uma série de problemas que aproximam o problema original e guardam, assim, algum paralelo com a presente proposta. As instâncias utilizadas por An foram geradas a partir de dados do PDB seguindo o modelo proposto por Moré, ou seja, considerando apenas as distâncias entre pares de átomos no mesmo resíduo ou em resíduos vizinhos.

Nas Tabelas 5.4 e 5.5, podemos verificar os desempenhos dos métodos *cgdca* e *sphdc* em instâncias derivadas de proteínas compostas por 237 a 4189 átomos. O conjunto K considerado foi o definido pela equação (5.1), ou seja, todos os arcos envolvendo átomos no mesmo resíduo ou em resíduos vizinhos. Os limites l_{ij} e u_{ij} foram gerados pela equação (5.3), tomando $\varepsilon = 0,001$ e $\varepsilon = 0,08$. Nas tabelas 5.5 e 5.4, vemos os resultados obtidos pelas diferentes propostas.

As medidas de erro utilizadas foram: o erro médio, dado por

$$E_{med} = \frac{1}{|K|}\left(\sum_{(i,j)\in K} \max\left\{\frac{\|x_i - x_j\| - u_{ij}}{u_{ij}}, \frac{l_{ij} - \|x_i - x_j\|}{l_{ij}}, 0\right\}\right), \tag{5.6}$$

e o erro máximo, dado por

$$E_{max} = \max\left\{\max\left\{\frac{\|x_i - x_j\| - u_{ij}}{u_{ij}}, \frac{l_{ij} - \|x_i - x_j\|}{l_{ij}}, 0\right\}\right\}. \tag{5.7}$$

As soluções obtidas pelo método *sphdc* apresentaram erros menores que as apresentadas pelo método *gcdca* em todas as instâncias. Na verdade, em todas as instâncias, as soluções obtidas pelo método *sphdc* tiveram erro máximo igual à zero. Isto deve-se ao fato dos pontos de mínimo global da função suavizada pertencerem estritamente ao interior do conjunto das restrições, diferentemente do que acontece com as soluções obtidas pelo método *gcdca*, onde os pontos pertencem claramente à fronteira do conjunto viável.

Com respeito ao tempo, o método *sphdc* teve desempenho muito superior ao apresentado pelo algoritmo *gcdca*. Isto em parte se deve à diferença entre os hardwares[3], mas também à diferença do tamanho dos problemas de otimização resolvidos.

[3]Os resultados obtidos pelo algortimo *gcdca* foram produzidos em servidor multiprocessador

			cgdca			*sphdc*		
proteína	m	n	E_{med}	E_{max}	tempo	E_{med}	E_{max}	tempo
8DRH	329	16	1,01E-06	1,04E-03	67,76	0,00	0,00	4,27
1AMD	380	12	7,23E-05	9,99E-03	47,68	0,00	0,00	14,20
2MSJ	480	66	2,23E-06	9,32E-03	28,63	0,00	0,00	3,08
124D	508	16	2,03E-05	9,99E-03	153,20	0,00	0,00	10,18
141D	527	26	1,07E-05	2,17E-03	313,43	0,00	0,00	7,16
132D	750	24	5,42E-06	2,96E-03	433,80	0,00	0,00	15,69
1A84	758	24	1,62E-06	3,45E-03	382,64	0,00	0,00	15,27
104D	766	24	3,20E-05	9,99E-03	151,35	0,00	0,00	16,55
103D	772	24	2,87E-05	9,99E-03	263,19	0,00	0,00	15,14
2EQL	1023	129	2,34E-06	9,99E-03	232,47	0,00	0,00	8,02
6GAT	1853	92	4,52E-05	9,98E-03	541,07	0,00	0,00	31,75
7HSC	2482	159	1,31E-05	9,99E-03	387,53	0,00	0,00	35,27
2CLJ	4189	543	2,12E-05	1,03E-02	1079,59	0,00	0,00	57,92

Tabela 5.4: Performance dos métodos *sphdc* e *gcdca* com dados do PDB com $\varepsilon = 0,001$. O número de átomos e de resíduos são representados, respectivamente, por m e n e o tempo é dado em segundos.

Enquanto que, por exemplo, na instância com 4189 átomos o método *gcdca* resolve problemas envolvendo sempre cerca de 32400 variáveis, no método *sphdc* essa quantidade cresce gradativamente, pois apenas dois resíduos são inicialmente considerados em cada subproblema. De fato, nos subproblemas iniciais, onde bons pontos iniciais não estão disponíveis, o número de variáveis não passa de 78. À medida que o procedimento *sphdc* avança, as soluções dos subproblemas são combinadas para gerar bons pontos iniciais para os problemas que envolvem maior número de variáveis. Assim, no momento em que os problemas maiores são tratados, já se está consideravelmente próximo da solução, requerendo um número reduzido de iterações para obtenção da solução.

5.3 Experimentos mais complexos

As instâncias idealizadas por Moré e Wu, utilizadas também por An, são extremamente convenientes para a decomposição implementada no método *sphdc*, pois cada bloco (resíduo) possui elevado número de restrições e só há restrições envolvendo resíduos vizinhos. Como consequência, é possível determinar a estrutura tridimensional de cada resíduo sem que se requeira informação adicional. Isto fica evidente

SGI Origin 2000 com sistema IRIX.

			cgdca			*sphdc*		
proteína	m	n	E_{med}	E_{max}	tempo	E_{med}	E_{max}	tempo
8DRH	329	16	1,93E-05	9,98E-05	35,03	0,00	0,00	2,32
1AMD	380	12	1,25E-05	9,99E-03	102,85	0,00	0,00	3,96
2MSJ	480	66	1,00E-04	3,53E-02	32,69	0,00	0,00	2,08
124D	508	16	1,78E-05	9,99E-03	387,06	0,00	0,00	5,71
141D	527	26	1,22E-04	8,77E-02	241,23	0,00	0,00	3,82
132D	750	24	1,35E-05	9,99E-03	320,04	0,00	0,00	8,54
1A84	758	24	9,50E-06	9,99E-03	484,84	0,00	0,00	8,66
104D	766	24	5,63E-06	1,30E-02	371,74	0,00	0,00	9,12
103D	772	24	1,00E-05	9,99E-03	411,48	0,00	0,00	8,80
2EQL	1023	129	2,97E-05	9,99E-03	183,27	0,00	0,00	5,59
6GAT	1853	92	3,22E-05	9,99E-03	1229,52	0,00	0,00	19,49
7HSC	2482	159	8,19E-06	9,99E-03	1129,08	0,00	0,00	23,61
2CLJ	4189	543	5,00E-05	1,12E-01	914,04	0,00	0,00	51,72

Tabela 5.5: Performance dos métodos *sphdc* e *gcdca* com dados do PDB com $\varepsilon = 0,08$. O número de átomos e de resíduos são representados, respectivamente, por m e n e o tempo é dado em segundos.

na Figura 5.3, onde podemos ver a matriz de conectividade da instância 8DRH (329 átomos e 16 resíduos). Os quadrados em destaque na imagem, coloridos em vermelho, representam os arcos associados a um único resíduo. Esta simplicidade das instâncias propostas por Moré é o que nos motivou a realizar um terceiro grupo de experimentos numéricos com problemas gerados a partir de dados do PDB, considerando todas as distâncias menores que um valor de corte C, ou seja,

$$K = \{(i,j) : \|x_i - x_j\| \leq C\}. \tag{5.8}$$

As instâncias geradas desta forma possuem matrizes de conectividade muito mais complexas com restrições envolvendo resíduos com índices bem distantes, propiciando um teste mais realista para a presente proposta.

Geramos dois grupos de instâncias; o primeiro, tomando $C = 6$Å, e o segundo, fazendo $C = 8$Å. O primeiro valor de corte $C = 6$Å é condizente com o comprimento das ligações detectáveis pelos equipamentos de RMN [89]. O último valor, $C = 8$Å, foi utilizado para podermos comparar o desempenho da presente proposta com os resultados disponibilizados por Di Wu e Zhijun Wu em [105], que aplicaram o algoritmo de construção geométrica, aqui denominado *gbu*, baseado na resolução de uma série de sistemas lineares. Assim como Wu e Wu, consideramos restrições de

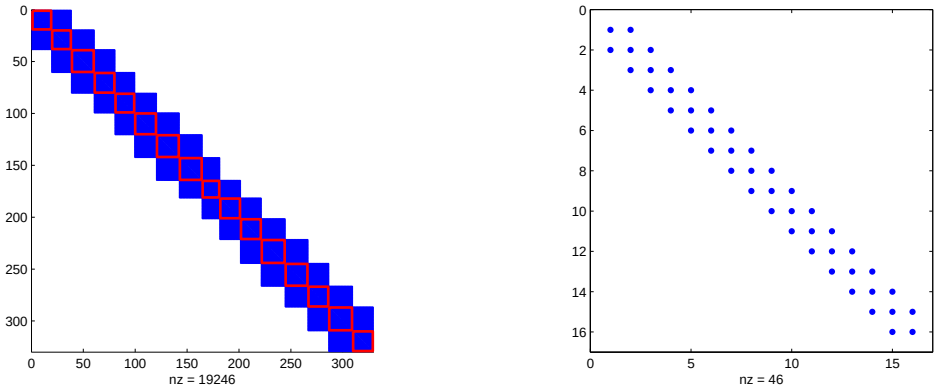

Figura 5.3: À direita, a matriz de conectividade dos resíduos da instância 8DRH. Cada um dos quadrados em vermelho identificam os arcos associados às restrições envolvendo átomos átomos no mesmo resíduo. À esquerda, a matriz de conectividade dos resíduos, note que não há correlação (arcos) entre resíduos que não sejam vizinhos.

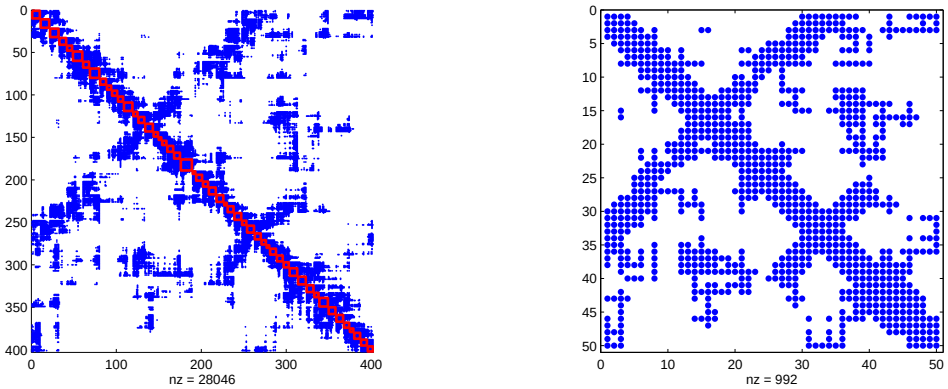

Figura 5.4: As matrizes de conectividade da instância 1PTQ com $C = 8$Å. As instâncias da forma dada na Eq. 5.8 são mais complexas que as idealizadas por Moré e Wu.

igualdade tomando $\varepsilon = 0$ na eq. (5.3).

Na Figura 5.4, vemos que a matriz de conectividade da instância gerada a partir da geometria da proteína 1PTQ (304 átomos e 50 resíduos), tomando $C = 8$Å, é muito mais complexa do que as geradas nas instâncias idealizadas por Moré e Wu.

A Tabela 5.6 mostra resultados obtidos pelo método *sphdc* e os resultados disponibilizados por Wu e Wu, em [105], com instâncias geradas tomando $C = 8$Å. Em todas as instâncias, os resultados obtidos pela proposta de Wu e Wu foram superiores (menor RMSD) aos obtidos pelo método *sphdc*. No entanto, os resultados obtidos pela presente proposta são em todos os sentidos satisfatórios. Na verdade, segundo Moré e Wu em [79], muitos pesquisadores consideram similares estruturas

		RMSD	
proteína	m	*gbu*	*sphdc*
1PTQ	404	3,5E-13	2,98E-04
1HOE	558	1,0E-11	3,27E-04
1LFB	641	3,9E-12	3,88E-04
1F39A	767	2,4E-12	3,68E-04
1PHT	811	1,8E-12	3,71E-04
1POA	914	1,7E-11	3,46E-04
1AX8	1003	3,5E-12	3,06E-04
1RGS	2015	1,1E-9	3,88E-04
1BPM	3671	3,2E-7	3,88E-04
1HMV	4200	2,5E-5	4,50E-04

Tabela 5.6: Performance dos método *sphdc* com dados do PDB em instâncias da forma dada na Eq. (5.8) com $C = 8$Å.

sphdc		
proteína	m	RMSD
1PTQ	404	2,95E-04
1HOE	558	2,63E-04
1LFB	641	4,03E-04
1F39A	767	3,46E-04
1PHT	811	3,59E-04
1POA	914	3,27E-04
1AX8	1003	2,94E-04
1RGS	2015	3,86E-04
1BPM	3671	3,95E-04
1HMV	4200	4,47E-04

Tabela 5.7: Performance dos método *sphdc* com dados do PDB em instâncias da forma dada na Eq. (5.8) com $C = 6$Å.

com RMSD entre um e dois angstrons.

A tabela 5.7 exibe os resultados dos experimentos numéricos mais significativos, onde as distâncias consideradas possuem comprimento compatível com a resolução dos experimentos de RMN, ou seja, $C = 6$Å. O desempenho do algoritmo *sphdc* foi plenamente satisfatório e semelhante ao obtido nas instâncias com $C = 8$Å.

Capítulo 6

Conclusão e propostas para trabalhos futuros

O problema da determinação de estruturas moleculares tem despertado grande interesse, graças à sua aplicação em áreas relevantes como medicina, farmácia, biologia, design de materiais e química [89, 103, 15, 112]. Uma das alternativas mais utilizadas nesta empreitada tem sido os experimentos envolvendo ressonância magnética nuclear (RMN) [107, 54], onde as distâncias entre alguns dos átomos que compõem a proteína podem ser estimadas. No problema geométrico de distâncias moleculares (MDGP), busca-se determinar a conformação tridimensional das estruturas proteicas, a partir das estimativas sobre algumas das distâncias entre átomos da proteína.

Com respeito à complexidade, se todas as distâncias entre os pares de átomos forem conhecidas, o MDGP pode ser resolvido em tempo polinomial [14, 30, 36]. No entanto, apenas um subconjunto das distâncias pode ser obtido via experimentos de RMN. Neste caso, o problema torna-se muito mais complexo. De fato, Saxe posicionou o MDGP, com apenas um subconjunto das distâncias sendo conhecido, na classe dos problemas NP-difíceis [87].

No presente trabalho, buscando maior aderência à realidade experimental, apresentamos um algoritmo para a determinação da estrutura proteica, a partir de um conjunto esparso de distâncias. Nosso algoritmo sphdc parte de uma formulação de mínimos quadrados e, através da aplicação de funções de suavização e penalização hiperbólicas, busca convexificar a função potencial envolvida. A racionalidade da proposta fundamenta-se na propriedade demonstrada de que valores adequados do

parâmetro de suavização são capazes de convexificar a função objetivo considerada e, assim, reduzir a probabilidade de convergência para mínimos locais pouco expressivos (profundos). A fim de manter escalabilidade, propomos ainda uma estratégia de decomposição e combinação de soluções baseada no modelo teórico de dividir-para-conquistar que permitem controlar o tamanho dos problemas de otimização enfrentados.

Realizamos experimentos computacionais com dados reais obtidos no PDB (*Protein Data Bank*). Os resultados, quando comparados aos demais encontrados na literatura, indicaram a viabilidade de aplicação do algoritmo *sphdc* na resolução do MDGP.

Como proposta de pesquisa futura, estamos interessados em aprimorar a função potencial considerada. Neste trabalho, consideramos apenas restrições de distâncias, mas inserindo restrições advindas de modelos teóricos de conformação proteica como, por exemplo, quiralidade, poderíamos, em tese, obter resultados ainda mais expressivos. Além disso, uma fase crucial de nosso algoritmo, a inicialização dos parâmetros, carece de justificação teórica que permita, por exemplo, automatizar a escolha dos mesmos a partir da magnitude e/ou precisão dos dados de entrada.

Ainda no campo das propostas de trabalhos futuros, apesar de possuirem alta performance, as implementações dos algoritmos utilizados nos experimentos numéricos realizados podem ser aperfeiçoadas utilizando, por exemplo, o paralelismo e uma linguagem não interpretada como é o caso do Matlab. O conjunto de rotinas nativas do Matlab aceleram consideravelmente o tempo de desenvolvimento, mas, infelizmente, comprometem sensivelmente a performance de rotinas com elevada comunicação de dados. Por exemplo, para a execução de instruções em paralelo, faz-se necessária, para cada *pool* de tarefas, uma instância do Matlab, o que produz grande *overhead* na comunicação de dados. Assim, a transcrição de o todo código-fonte para uma linguagem compilada pode aumentar a performance dos algoritmos implementados.

Outro tópico de interesse está relacionado à detecção de possíveis inconsistências dos dados de entrada como, por exemplo, a violação da desigualdade triangular. Consideramos que seria útil dispor de ferramentas que indicassem a inviabilidade dos dados e também apontassem que conjuntos de distâncias são incompatíveis entre

si.

Outro ponto que merece atenção é a natureza multidisciplinar do MDGP e a necessidade de obtenção/formulação de instâncias mais realistas, já que as atualmente consideradas simplificam consideravelmente o problema real. Esta simplificação aparece, por exemplo, na quantidade e qualidade dos dados disponíveis. Assim, a formação de parceria com laboratórios que produzam dados reais de experimentos de RMN seria algo realmente produtivo e permitiria o aprofundamento e comprovação em cenário real dos resultados aqui apresentados.

Portanto, concluímos que existe um vasto campo de possibilidades a ser explorado em trabalhos futuros, seja diretamente no MDGP ou aplicando a presente proposta em problemas correlatos.

Referências Bibliográficas

[1] A. Abragam. *Principles of nuclear magnetism*. Oxford: Clarendon Press, 1961.

[2] A. Y. Alfakih. On the uniqueness of euclidean distance matrix completions: the case of points in general position. *Linear Algebra and its Applications*, 397(1):265–277, 2005.

[3] A. Y. Alfakih, A. Khandani, and H. Wolkowicz. Solving euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, 12(1-3):13–30, 1999.

[4] J. M. Amabis and G. R. Martho. Do gene à proteína. *Atualidades Biológicas - Editora Moderna*, 6:1–8, 1997.

[5] L. T. H. An. Solving large scale molecular distance geometry problems by smoothing technique via the gaussian transform and d. c. programming. *Journal of Global Optimization*, 27:375–397, 2003.

[6] L. T. H. An and P. D. Tao. Large-scale molecular optimization from distance matrices by d. c. optimization approach. *SIAM Journal on Optimization*, 14:77–114, 2003.

[7] C. B. Anfinsen, E. Haber, M. Sela, and F. H. W. Jr. The kintics of formation of native ribonuclease durig oxidation of reduced polypeptide chain. *Proceedings of the National Academy of Sciencies of the USA*, 47:1309–1314, 1961.

[8] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5):93–96, October 2001.

[9] J. C. Beauchamp and N. W. Isaacs. Methods for x-ray diffraction analysis of macromolecular structures. *Current Opinion in Structural Biology*, 3:525–529, 1999.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[11] P. Biswas. *Semidefinite programming approaches to distance geometry problems*. PhD thesis, Stanford University, June 2007.

[12] P. Biswas, T. C. Liang, T. C. Wang, and Y. Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networs*, 2(2):188–200, 2006.

[13] P. Biswas, K.-C. Toh, and Y. Ye. A distributed sdp approach for large-scale noisy anchor-free graph realization with applications to molecular conformation. *SIAM Journal on Scientific Computing*, 30(3):1251–1277, 2008.

[14] L. M. Blumenthal. *Theory and applications of distance geometry*. Clarendon Press, 1953.

[15] D. B. Boyd. Rational drug design: Controlling the size of the haystack. *Modern Drug Discovery*, 1:41–47, 1998.

[16] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, 1991.

[17] S. E. Brenner. A tour of structural genomics. *Nature Reviews of Genetics*, 2(10):801–809, October 2001.

[18] A. T. Bruenger. X-ray crystallography and nmr reveal complementary views of structure and dynamics. *Nature Structural Biology*, Supl. 4(10):862–865, October 1997.

[19] C. R. Cantor and P. R. Schimmel. *Biophysical Chemistry*. W. H. Freeman and Company, San Francisco, 1980.

[20] L. Cayton and S. Dasgupta. Robust euclidean embedding. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 169–176, New York, NY, USA, 2006. ACM.

[21] C. M. Chiu and D. C. Muddiman. What is mass spectrometry? http://www.asms.org/whatisms/index.html, May 2008.

[22] M. T. Chu, R. E. Funderlic, and R. J. Plemmonsc. Structured low rank approximation. *Linear Algebra and its Applications*, 366(1):157–172, 2003.

[23] F. Chung, M. Garret, R. Graham, and D. Shallcross. Distance realization problems with applications to internet tomography. *Journal of Computer and System Sciences*, 63(17):432–448, November 2001.

[24] G. M. Clore and A. M. Gronenborn. New methods os structure refinement for macromolecular structure determination by nmr. *Proceedings of the National Academy of Sciencies of the USA*, 95:5891–5898, 1998.

[25] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2009.

[26] J. A. Costa, N. Patwari, and I. A. O. Hero. Distributed weighted-multidimensional scaling for node localization in sensor networks. *ACM Transactions on Sensor Networs*, 2(1):39–64, 2006.

[27] T. Creighton. *Proteins Structures and Molecular Properties*. New York: Freeman, 1993.

[28] G. M. Crippen. Distance geometry and conformational calculations. In *Chemometrics Research Studies Series*, volume 1. Research Studies Press (Wiley), New York, 1981.

[29] G. M. Crippen. Linearized embedding: a new metric matrix algorithm for calculating molecular conformations subject to geometric constraints. *Journal of Computational Chemistry*, 10(7):896–902, 1989.

[30] G. M. Crippen and T. F. Havel. Distance geometry and molecular conformation. In *Chemometrics Research Studies Series*. Research Studies Press (Wiley), New York, 1988.

[31] W. G. da Silva. Algoritmos para o cálculo de estruturas de proteínas. Master's thesis, Universidade Federal Fluminense, 2008.

[32] S. Dasgupta, C. Papadimitriou, and U. Vazirani. *Algorithms*. McGraw-Hill Science/Engineering/Math, 2006.

[33] J. Datorro. *Convex Optimization and Euclidean Distance Geometry*. Meboo Publishing, 2006.

[34] K. Davies. *Cracking the genome: inside the race to unlock human dna*. The Free Press (A Simon & Schuster Division), New York, NY, 2001.

[35] 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*, 22:365–375, 2002.

[36] Z. Dong, Qunfeng Wu. A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization*, 26(3):1573–2916, 2003.

[37] A. F. U. dos Santos Macambira. Determinação de estruturas de proteínas via suavização e penalização hiperbólica. Master's thesis, Universidade Federal do Rio de Janeiro, 2003.

[38] J. Drent. *Principles of protein x-ray crystallography*. New York: Springer, 1994.

[39] R. A. Engh and R. Huber. Accurate angles and bond parameters for x-ray protein structure refinement. *Acta Crystallographica A*, 47:392–400, 1991.

[40] T. Eren, D. Goldenberg, W. Whiteley, Y. R. Yang, A. S. Morse, B. D. O. Anderson, and P. N. Belhumeur. Rigidity, computation, and randomization in network localization. In *Proceedings of the International Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pages 2673–2684, March 2004.

[41] T. Eren, W. Whiteley, and P. N. Belhumeur. Further results on sensor network localization using rigidity. In *Proceedings of the Second European Workshop on Sensor Networks (EWSN)*, pages 405–409, January 2005.

[42] C. Ezzel. Proteins rule. *Scientific American*, 286(4):40–7, April 2002.

[43] F. D. Fagundez, A. E. Xavier, R. Medronho, J. L. D. Faco, and L. L. Xavier. A study on the universal access to vaccines in brazil. *XL SBPO*, 1, 2008.

[44] A. Fiaco and G. McCormick. *Nonlinear Programming, Sequential Unconstrained Minimization Techniques, Classics in Applied Mathematics 4*. SIAM, Philadelphia, 1990.

[45] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, San Diego, CA, 1996.

[46] D. Gleich, L. Zhukov, M. Rasmussen, and K. Lang. The world of music: Sdp layout of high dimensional data. In *Info Vis 2005*, 2005.

[47] W. Glunt and T. L. Hayden. The embedding problem for predistance matrices. *Bulletin of Mathematical Biology*, 53(5):769–796, 1991.

[48] W. Glunt, T. L. Hayden, S. Hong, and J. Wells. An alternating projection algorithm for computing the nearest euclidean distance matrix. *SIAM Journal on Matrix Analysis and Applications*, 11:589–600, 1990.

[49] W. Glunt, T. L. Hayden, and M. Raydan. Molecular conformations from distance matrices. *Journal of Computational Chemistry*, 14(1):114–120, 1993.

[50] W. Glunt, T. L. Hayden, and M. Raydan. Preconditioners for distance matrix algorithms. *Journal of Computational Chemistry*, 15:227–232, 1994.

[51] G. H. Golub. *Matrix Computations*. Johns Hopkins University Press, 1989.

[52] P. Green. Whole-genome disassembly. *Proceedings of the National Academy of Sciencies of the USA*, 99(7):4143–4144, Apr 2002.

[53] Y. Guan, H. Zhang, R. N. H. Konings, C. W. Hilbers, T. C. Terwilliger, and A. H. J. Wang. Crystal structure of y41h and y41f mutants of gene v suggest possible protein-protein interactions in the gvp-ssdna complex. *Biochemistry*, 33:7768, 1994.

[54] P. Guentert. Structure calculation of biological macromolecules from nmr data. *Quarterly Reviews of Biophysics*, 31(2):145–237, 1998.

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

[56] L. Hama. O mapa da vida. *Super Interessante*, 204A:10–10, 2004.

[57] B. A. Hendrickson. *The molecule problem: determining conformation from pairwise distances*. PhD thesis, Cornell University, 1990.

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

[59] W. A. Hendrickson. Stereochemically restrained refinement of macromolecular structures. *Methods in Enzymology*, 115:252–270, 1985.

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

[61] A. J. Kearsley, R. A. Tapia, and M. Trosset. The solution of the metric stress and sstress problems in multidimensional scaling by newton's method. *Computational Statistics*, 13:369–396, 1998.

[62] R. Kimmel. *Numerical Geometry of Images: Theory, Algorithms, and Applications*. Springer Verlag, 2003.

[63] D. E. Knuth. *Art of Computer Programming*, volume Volume 3: Sorting and Searching. Addison-Wesley Professional, 1998.

[64] J. Kostrowicki and L. Piela. Diffusion equation method of global minimization: performance for standard test functions. *Journal of Optimization Theory and Applications*, 69:269–284, 1991.

[65] J. Kostrowicki, L. Piela, B. J. Cherayil, and H. A. Scheraga. Performance of the diffusion equation method in searches for optimun structures of clusters of lennard-jones atoms. *The Journal of Physical Chemistry*, 95:4113–4119, 1991.

[66] J. Kostrowicki and H. A. Scheraga. Applications of diffusion equation method for global optimization oligopeptides. *The Journal of Physical Chemistry*, 96:7442–7449, 1992.

[67] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956.

[68] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump,

H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blï¿½cker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and I. H. G. S. Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[69] M. Laurent. Cuts, matrix completions and graph rigidity. *Mathematical Programming*, 79:255–283, 1997.

[70] C. Lavor. On generating instances for the molecular distance geometry problem. In L. Liberti and N. Maculan, editors, *Global optimization: from theory to implementation*. Springer, 2006.

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

[72] D. Liu and Nocedal. On the limited memory bfgs method for large scale optimization. Technical Report NA-03, Department of Electrical Engineering and Computer Science Northwestern University, 1988.

[73] R. Mathar. The best euclidean fit to a given distance matrix in prescribed dimensions. *Linear Algebra and its Applications*, 67:1–6, 1985.

[74] A. McPherson. Macromolecular crystals. *Scientific American*, 260:62–69, 1989.

[75] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides. *The Journal of Physical Chemistry*, 79:2361–2381, 1975.

[76] J. J. Moré and Z. Wu. ϵ-optimal solutions to distance geometry problems via global continuation. Technical report, Argonne: Mathematics and Computer Science Division, 1995.

[77] J. J. Moré and Z. Wu. Smoothing techniques for macromolecular global optimization. In G. D. Pillo and F. Giannessi, editors, *Nonlinear Optimization and Applications*, pages 297–312. Plenum Press, 1996.

[78] J. J. Moré and Z. Wu. Global continuation for distance geometry problems. *SIAM Journal on Optimization*, 7:814–836, 1997.

[79] J. J. Moré and Z. Wu. Distance geometry optimization for protein structures. *Journal of Global Optimization*, 15:219–234, 1999.

[80] A. O. Moreno. *Um novo algoritmo para resolução de problemas de classificação.* PhD thesis, Universidade Federal do Rio de Janeiro, 2008.

[81] A. O. Moreno, M. Souza, and A. E. Xavier. A new algorithm to solve classification problems. In *The XIIIth International Conference: Applied Stochastic Models and Data Analysis (ASMDA-2009)*, pages 11–13. Vilnius Gediminas Technical University Publishing House, Technika, 2009.

[82] A. Neumaier. Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Review*, 39:407–460, 1997.

[83] K. A. Palmer and H. A. Scheraga. Standard-geometry chains fitted to x-ray derived structures: validation of the rigid-geometry approximation. i. chain closure through a limited search of "loop" conformations. *Journal of Computational Chemistry*, 12(4):505–526, 1991.

[84] N. Patwari, I. A. O. Hero, and A. Pacholski. Manifold learning visualization of network traffic data. In *MiniNet '05: Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data*, pages 191–196, New York, NY, USA, 2005. ACM Press.

[85] L. Piela, J. Kostrowicki, and H. A. Scheraga. The multiple-minima problem in the conformational abalysis of molecules: deformation of the protein energy hypersurface by diffusion equation method. *The Journal of Physical Chemistry*, 93:3339–3346, 1989.

[86] R. C. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36:1389–1401, 1957.

[87] J. B. Saxe. Embeddability of weighted graphs in k-space is strongly np-hard. In *Proc. 17th Allerton Conf. in Communications, Control, and Computing*, pages 480–489, 1979.

[88] H. A. Scheraga, J. Lee, J. Pillardy, Y.-J. Ye, A. Liwo, and D. Ripoll. Surmounting the multiple-minima problem in protein folding. *Journal of Global Optimization*, 15(3):235–260, 1999.

[89] T. Schlick. *Molecular modeling and simulation: an interdisciplinary guide.* Springer, 2002.

[90] Y. Shang, W. Ruml, Y. Zhang, and M. P. J. Fromherz. Localization from mere connectivity. In *Proceedings of the Fourth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 201–212. ACM Press, 2003.

[91] S. S. Sheik, P. Sundararajan, A. S. Hussain, and K. Sekar. Ramachandran plot on the web. *Bioinformatics*, 18(11):1548–1549, 2002.

[92] J. G. Siek, L.-Q. Lee, and A. Lumsdaine. *The Boost Graph Library User Guide and Reference Manual*. Upper Saddle River, 2002.

[93] M. Skinner, H. Zhang, D. Leschnitzer, Y. Guan, H. Bellamy, R. Sweet, C. Gray, R. Konings, A. Wang, and T. Terwilliger. Structure of the gene v protein of bacteriophage f1 determined by multi-wavelength x-ray diffraction on the selenomethionyl protein. *Proceedings of the National Academy of Sciencies of the USA*, 91:2071, 1994.

[94] M. Trosset. Applications of multidimensional scaling to molecular conformation. *Computing Science and Statistics*, 29:148–152, 1998.

[95] V. M. Unger. Electron cryomicroscopy methods. *Current Opinion in Structural Biology*, 11:548–554, 2001.

[96] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin,

H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigï¿½, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[97] D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nature Structural Biology*, 8(6):559–566, June 2001.

[98] G. Wagner. An account of nmr in structural biology. *Nature Structural Biology*, Supl. 4(10):841–844, October 1997.

[99] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman. Biomolecular simula-

tions: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annual Review of Biophysics and Biomolecular Structure*, 30:211–243, 2001.

[100] R. H. Waterston, E. S. Lander, and J. E. Sulston. On the sequencing of the human genome. *Proceedings of the National Academy of Sciencies of the USA*, 99(6):3712–3716, Mar 2002.

[101] R. H. Waterston, E. S. Lander, and J. E. Sulston. More on the sequencing of the human genome. *Proceedings of the National Academy of Sciencies of the USA*, 100(6):3022–4; author reply 3025–6, Mar 2003.

[102] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[103] B. Werth. *The Billion-Dollar Molecule: One Company's Quest for the Perfect Drug*. Simon & Schuster, New York, NY, 1994.

[104] G. A. Williams, J. M. Dugan, and R. B. Altman. Constrained global optimization for estimating molecular structure from atomic distances. *Journal of Computational Biology*, 8(5):523–547, 2001.

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

[106] Z. Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM Journal on Optimization*, 6(3):748–768, 1996.

[107] K. Wuthrich. *NMR of proteins and nucleic acids*. New York: Wiley, 1986.

[108] A. E. Xavier. Solução de problemas de programação não-diferenciáveis via suavização. Technical report, Universidade Federal do Rio de Janeiro, 1993.

[109] A. E. Xavier. Convexificação do problema de distância geométrica através da técnica da suavização hiperbólica. In *Workshop em Biociências*. COPPE/UFRJ, Dezembro 2003.

[110] A. E. Xavier, J. L. D. Facó, and J. R. Paula Júnior. A novel contribution to the tammes probem by the hyperbolic smoothing method. In *EURO Mini Conference : EurOPT-2008*, Lituania, 2008. EurOPT-2008.

[111] A. E. Xavier, V. S. A. Menezes, and P. C. Rodrigues. The hyperbolic smoothing approach for solving a support vector machine problem: An outline. In *The XIIIth International Conference: Applied Stochastic Models and Data Analysis (ASMDA-2009)*, pages 527–530. Vilnius Gediminas Technical University Publishing House, Technika, 2009.

[112] B. I. Yakobson and R. E. Smalley. Fullerene nanotubes: $C_{1,000,000}$ and beyond. *American Scientist*, 85(4):324, 1997.

[113] J.-M. Yoon, Y. Gad, and Z. Wu. Mathematical modeling of protein structure using distance geometry. Technical report, Rice University, 2000.

[114] Z. Zou, R. H. Bird, and R. B. Schnabel. A stochastic/pertubation global optimization algorithm for distance geometry problems. *Journal of Global Optimization*, 11:91–105, 1997.