UMA FORMULAÇÃO NUMÉRICA PARA RESOLUÇÃO DE PROBLEMAS DE GEOMETRIA DE DISTÂNCIAS MOLECULARES

Felipe Fidalgo
felipefidalgo@ime.unicamp.br

Douglas Maioli
douglasmaioli@bol.com.br

Eduardo Abreu
eabreu@ime.unicamp.br

Carlile Lavor
clavor@ime.unicamp.br

Departamento de Matemática Aplicada
Instituto de Matemática, Estatística e Computação Científica - IMECC
Universidade Estadual de Campinas - UNICAMP
Rua Sérgio Buarque de Holanda, 651 - Barão Geraldo - Campinas/SP

RESUMO

Dada uma molécula com n átomos, o Problema de Geometria de Distâncias Moleculares (PGDM) consiste em encontrar as posições $x_1, \ldots, x_n$ destes átomos em $\mathbb{R}^3$, dadas algumas distâncias entre eles. Esse problema é relevante, por exemplo, na determinação de estruturas de proteínas. Neste trabalho, é proposto um novo método numérico para resolver o PGDM, chamado Algoritmo T (AT), o qual demonstrou boa performance em tempo computacional, quando comparado com resultados disponíveis na literatura. Além disso, será apresentada uma modificação deste método, chamado Algoritmo T Atualizado (ATA), com o objetivo de evitar a propagação e o acúmulo de erros numéricos referentes ao mal-condicionamento dos sistemas lineares provenientes da modelagem do problema. Para controlar esse acúmulo, são adotadas duas estratégias: *(i)* resolver os sistemas lineares usando fatoração LU com estratégia de pivoteamento parcial e *(ii)* a reinicialização, na qual as coordenadas dos átomos base são recalculadas a partir das distâncias entre eles. Resultados numéricos preliminares, utilizando estruturas artificiais de moléculas publicadas na literatura, são apresentados para validação da proposta.

Palavras Chave: Geometria de Distâncias, Estruturas Moleculares, Fatoração LU.
Área principal: Programação Matemática, Otimização Combinatória.

ABSTRACT

Given a molecule with n atoms, the Molecular Distance Geometry Problem (MDGP) is defined as the problem of finding the positions $x_1, \ldots, x_n$ of these atoms in $\mathbb{R}^3$, given some distances among these points. This problem is significant, for example, in protein structure determination. In this paper, it is proposed a new numerical method to solve the MDGP, called T Algorithm (TA), which has shown good performance in computational time, when compared with results available in the literature. In additon, it will be shown a modification of this method, called the Updated T Algorithm (UTA), in order to avoid numerical error accumulation and propagation related to the ill-conditioning of the linear systems from the modeling problem at hand. For controlling this accumulation, two strategies are adopted: *(i)* solving the linear systems using LU factorization with partial pivoting and *(ii)* the re-initialization, where the coordinates of the base atoms are recalculated from the distances among them. Preliminary numerical results using artificial structures of molecules published in the literature are also reproduced in order to validate our approach.

Keywords: Distance Geometry, Molecular Structures, LU Factorization.
Main area: Mathematical Programming, Combinatorial Optimization.

1 Introdução

A partir de experimentos de Ressonância Magnética Nuclear (RMN) [7], é possível obter distâncias entre pares de átomos de uma molécula que estejam próximos (5 a 6 angstrons) [12]. De posse desses dados, é possível definir um problema, baseado em uma modelagem matemática de geometria de distâncias [1], para determinar estruturas de moléculas e, em especial, de proteínas [3, 6, 13]. O PGDM pode ser formalizado como segue.

Problema de Geometria de Distâncias Moleculares (PMGD)**.** *Considere uma molécula formada por uma sequência de n átomos, da qual é conhecido um subconjunto de distâncias entre pares deles. É possível obter uma configuração tridimensional para tal molécula, compatível com as distâncias conhecidas, isto é, determinar um conjunto de vetores de coordenadas* $\{x_1,\ldots,x_n\}$ *em* $\mathbb{R}^3$ *para seus átomos de modo que as distâncias euclideanas entre essas posições sejam iguais às distâncias conhecidas?*

Usualmente, o PGDM é formulado como um problema de otimização global contínua sem restrições [10] do seguinte modo:

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} \sum_{(u,v)\in E} (\|x_u - x_v\|_2^2 - d_{uv}^2)^2,$$

onde $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $E \subseteq \{1,2,\ldots,n\}^2$ e d_{uv} é a distância dada entre as posições dos átomos u e v. Sabe-se que $\mathbf{x}$ é solução para o PGDM se, e somente se, $f(\mathbf{x}) = 0$. Encontrar este minimizador global é um problema NP-difícil [11].

O PGDM pode ser classificado de duas formas: *(i)* Se todas as distâncias entre quaisquer pares de átomos são conhecidas, dizemos que este é um PGDM com *Conjunto Completo de Distâncias*. *(ii)* Caso contrário, temos um PGDM com *Conjunto Arbitrário de Distâncias*. Dong e Wu [2] apresentaram um algoritmo em tempo polinomial para os problemas da primeira classe baseado na resolução de uma sequência de sistemas lineares. Para os problemas da segunda classe, eles apresentaram um método, chamado *Geometric Build-Up Algorithm (GB)*, envolvendo, também, resoluções de sistemas lineares [3]. Entretanto, este não se mostrou tão eficiente e robusto, pois tinha grande acúmulo de erros de arredondamento, principalmente para grandes moléculas. Posteriormente, Dong e Wu reformularam o GB, criando o *Updated Geometric Build-Up Algorithm (UGB)*, que faz uma atualização a cada iteração para descobrir a posição do átomo em questão. Esta abordagem diminui drasticamente a propagação de erros de arredondamento nos cálculos, principalmente, ao lidar com moléculas de grande porte [13].

Neste trabalho, vamos apresentar um novo algoritmo que resolve o PGDM com Conjunto Arbitrário de Distâncias, que tem como uma de suas vantagens, em relação ao UGB, um menor custo computacional e a possibilidade de tratar instâncias do problema que apresentam erros nas distâncias dadas.

2 Algoritmo T

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

Para a demonstração do Teorema 1, faremos uso do seguinte lema de demonstração trivial.

Lema 1. *Se* $\{x_1,x_2,x_3,x_4\} \subset \mathbb{R}^3$ *é um conjunto de pontos não-coplanares, então a matriz*

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

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

A seguir, apresentamos o primeiro resultado para a definição do novo método.

Teorema 1. *Sejam* $\{b_1,b_2,b_3,b_4\}$ *e* $\{y_1,y_2,y_3,y_4\}$*, respectivamente, subconjuntos de* $\mathbb{R}^3$ *e* $\mathbb{R}$*. Se o sistema quadrático*

$$\begin{aligned} \|a-b_1\|_2 &= y_1 \\ \|a-b_2\|_2 &= y_2 \\ \|a-b_3\|_2 &= y_3 \\ \|a-b_4\|_2 &= y_4 \end{aligned} \tag{1}$$

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

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

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

Demonstração. Seja $a^* \in \mathbb{R}^3$ uma solução para o sistema (1). Assim, substituindo a^* neste sistema, elevando suas equações ao quadrado e fazendo operações com elas, temos

$$\begin{aligned} \|a^*\|^2 - 2b_1^T a^* &= y_1^2 - \|b_1\|^2 \\ \|a^*\|^2 - 2b_2^T a^* &= y_2^2 - \|b_2\|^2 \\ \|a^*\|^2 - 2b_3^T a^* &= y_3^2 - \|b_3\|^2 \\ \|a^*\|^2 - 2b_4^T a^* &= y_4^2 - \|b_4\|^2 \end{aligned} \quad , \quad \text{ou seja,} \quad \begin{aligned} -2t - 2b_1^T a^* &= y_1^2 - \|b_1\|^2 \\ -2t - 2b_2^T a^* &= y_2^2 - \|b_2\|^2 \\ -2t - 2b_3^T a^* &= y_3^2 - \|b_3\|^2 \\ -2t - 2b_4^T a^* &= y_4^2 - \|b_4\|^2 \end{aligned} \quad , \tag{2}$$

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

$$-2\begin{bmatrix} 1 & b_1^T \\ 1 & b_2^T \\ 1 & b_3^T \\ 1 & b_4^T \end{bmatrix} \begin{bmatrix} t \\ a^* \end{bmatrix} = \begin{bmatrix} y_1^2-\|b_1\|^2 \\ y_2^2-\|b_2\|^2 \\ y_3^2-\|b_3\|^2 \\ y_4^2-\|b_4\|^2 \end{bmatrix}. \tag{3}$$

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

A cada iteração do AT, deseja-se determinar a posição x_j de um átomo indeterminado da molécula. Sejam x_1, x_2, x_3 e x_4 as posições de quatro átomos determinados, não-coplanares, com distâncias entre si e com x_j todas conhecidas. A partir das distâncias $d_{j,1}, d_{j,2}, d_{j,3}$ e $d_{j,4}$, entre x_j e x_1, x_2, x_3 e x_4, respectivamente, podemos considerar o seguinte sistema quadrático:

$$\begin{aligned} \|x_j - x_1\|_2 &= d_{j,1} \\ \|x_j - x_2\|_2 &= d_{j,2} \\ \|x_j - x_3\|_2 &= d_{j,3} \\ \|x_j - x_4\|_2 &= d_{j,4} \end{aligned} . \tag{4}$$

Apresentaremos, a seguir, o principal resultado que embasa o novo método, fazendo uso do Teorema 1 e do Lema 1.

Corolário 1. *Suponhamos que* $\{x_1, x_2, x_3, x_4\}$ *é um conjunto de pontos determinados, não-coplanares, com distâncias conhecidas entre si e com o ponto indeterminado* x_j*. Se o sistema não-linear*

$$\begin{aligned} \left\|x_j - x_1\right\|_2 &= d_{j,1} \\ \left\|x_j - x_2\right\|_2 &= d_{j,2} \\ \left\|x_j - x_3\right\|_2 &= d_{j,3} \\ \left\|x_j - x_4\right\|_2 &= d_{j,4} \end{aligned} \tag{5}$$

admite solução ${x_j}^*$*, então* $x^* = \begin{bmatrix} t_j & {x_j}^{*T} \end{bmatrix}^T$*, onde* $t_j = -\left\|{x_j}^*\right\|^2/2$*, é solução única de* $Ax = b$ *com*

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

Demonstração. O sistema não-linear (5) possui solução única [3, 4]. Como A é não-singular, pelo Lema 1, então $Ax = b$ também possui solução única. Seja ${x_j}^*$ solução de (5). Portanto, pelo Teorema 1, o vetor $x^* = \begin{bmatrix} t_j & {x_j}^{*T} \end{bmatrix}^T$ é solução única de $Ax = b$, onde $t_j = -\left\|{x_j}^*\right\|^2/2$.

□

Neste momento, observe que o problema (4) satisfaz as condições do Teorema 1. Logo, partindo da solução $x^* = \begin{bmatrix} t_j & {x_j}^{*T} \end{bmatrix}^T$, de $Ax = b$, temos que x_j^* é a posição que queremos encontrar para o átomo j. Além disso, a primeira coordenada da solução de $Ax = b$ pode ser utilizada como um teste para a qualidade da solução encontrada, já que a mesma tem de estar próxima de $t_j = -\left\|{x_j}^*\right\|^2/2$. Deste modo, pequenas imprecisões nos dados de distâncias podem ser toleradas neste método, i.e., mesmo que este algoritmo não determine a solução exatamente, ele é capaz de determinar uma aproximação da mesma. Dessa forma, apresentamos o método proposto por este trabalho, chamado Algoritmo T (AT): para cada átomo a ser determinado, mesmo não encontrando uma solução exata, podemos ter uma solução aproximada de acordo com uma tolerância conveniente, ajustada à natureza de cada problema. Se não for possível determinar os quatro átomos base das iterações ou resolver o sistema linear do método, o AT pára, reportando uma estrutura molecular parcial.

Antes da definição do AT, apresentamos o seguinte resultado que foi demonstrado por Wu e Wu [13]. Vamos reproduzir esta demonstração, pois os cálculos empregados serão usados para a construção do primeiro passo do AT.

Teorema 2. *Dadas as distâncias entre quatro átomos de uma molécula, dois-a-dois, suas coordenadas cartesianas podem ser determinadas, usando translação, rotação e reflexão.*

Demonstração. Sejam $d_{i,j}$ a distância entre os átomos i e j, com $i, j = 1, \ldots, 4$. Sem perda de generalidade, realizamos uma mudança de coordenadas, isto é, translação, rotação e reflexão de modo que: o primeiro átomo está na origem de nosso sistema, isto é, $x_1 = (0,0,0)$, o segundo átomo está sobre o eixo das abscissas, ou seja, suas coordenadas são da forma $x_2 = (x_{21}, 0, 0)$ e, por fim, o terceiro átomo está no plano abscissa-ordenada, isto é, $x_3 = (x_{31}, x_{32}, 0)$.

De posse dessas hipóteses, podemos definir o sistema não-linear

$$\begin{aligned} \|x_2 - x_1\|_2 &= d_{2,1} \\ \|x_3 - x_1\|_2 &= d_{3,1} \\ \|x_3 - x_2\|_2 &= d_{3,2} \end{aligned} \quad , \quad \text{ou seja,} \quad \begin{aligned} x_{21}^2 &= d_{2,1}^2 \\ x_{31}^2 + x_{32}^2 &= d_{3,1}^2 \\ (x_{31} - x_{21})^2 + x_{32}^2 &= d_{3,2}^2 \end{aligned} . \tag{6}$$

Resolvendo o sistema (6), obtemos os valores

$$x_{21} = \pm d_{2,1}, \quad x_{31} = \frac{d_{3,1}^2 - d_{3,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2} \quad \text{e} \quad x_{32} = \pm(d_{3,1}^2 - x_{31}^2)^{1/2}. \tag{7}$$

Nestas coordenadas, partindo do sistema

$$\begin{array}{l} \|x_4 - x_1\| = d_{4,1} \\ \|x_4 - x_2\| = d_{4,2} \\ \|x_4 - x_3\| = d_{4,3} \end{array}, \quad \text{isto é,} \quad \begin{array}{r} x_{41}^2 + x_{42}^2 + x_{43}^2 = d_{4,1}^2 \\ (x_{41} - x_{21})^2 + x_{42}^2 + x_{43}^2 = d_{4,2}^2 \\ (x_{41} - x_{31})^2 + (x_{42} - x_{32})^2 + x_{43}^2 = d_{4,3}^2 \end{array}, \tag{8}$$

determinamos $x_4 = (x_{41}, x_{42}, x_{43})^T$. De fato, por substituição, obtemos explicitamente

$$x_{41} = \frac{d_{4,1}^2 - d_{4,2}^2}{2d_{2,1}} + \frac{d_{2,1}}{2}, \quad x_{42} = \frac{d_{4,2}^2 - d_{4,3}^2 - (x_{41} - x_{21})^2 + (x_{41} - x_{31})^2}{2x_{32}} + \frac{x_{32}}{2} \tag{9}$$

$$\text{e} \quad x_{43} = \pm(d_{4,1}^2 - x_{41}^2 - x_{42}^2)^{1/2}. \tag{10}$$

Podemos tomar tanto a parte negativa quanto positiva para os valores x_{21}, x_{32} e x_{43}. □

Utilizando os resultados demonstrados acima, segue o pseudo-algoritmo do AT.

Algoritmo (AT)**.** *Dado um conjunto D de distâncias euclideanas entre pares dentre os n átomos de uma molécula:*

1. *Encontre quatro átomos não-coplanares com distâncias conhecidas entre si.*
2. *Determine suas coordenadas x_1, x_2, x_3 e x_4, segundo a demonstração do Teorema 2.*
3. *Faça:*
   - *(i) Para cada átomo não-determinado j, se possível, encontre quatro átomos determinados, não-coplanares, com distâncias conhecidas entre si e com o átomo indeterminado, x_{j1}, x_{j2}, x_{j3} e x_{j4}. Caso não encontre, pare.*
   - *(ii) A partir desses quatro átomos, resolva o sistema $Ax = b$ do Teorema 1.*
   - *(iii) Faça $x_j(i) = x(i+1)$, para $i = 1,2,3$. x_j é a posição determinada para o átomo j.*
4. *Se nenhum átomo é determinado em todo o loop, pare e reporte a estrutura parcial. Senão, volte ao passo (3).*

Esta é a proposta deste trabalho. Os algoritmos AT e GB são parecidos em sua estrutura, entretanto o cerne do AT se difere do GB: o sistema linear resolvido a cada iteração no AT é de dimensão 4, enquanto que no GB, o sistema linear é de dimensão 3.

3 Algoritmo T Atualizado

Para evitar a propagação e o acúmulo de erros numéricos e garantir mais estabilidade, Wu e Wu [13] realizam uma atualização a cada iteração do GB, gerando o UGB. Nesta atualização, há uma mudança do sistema de coordenadas: passamos do sistema original para outro, de modo que as posições dos quatro átomos x_1, x_2, x_3 e x_4 dependam, apenas, das distâncias entre eles para serem determinadas e independam de cálculos previamente realizados. Depois, calcula-se a posição do átomo indeterminado no novo sistema e, em seguida, voltamos os cinco átomos para o sistema de

coordenadas original. O UGB se mostra mais robusto do que o GB, garantindo maior qualidade das soluções [13]. Portanto, fazendo o mesmo processo de atualização no AT, obtemos o ATA.

Segundo esta idéia, calcula-se as posições dos quatro primeiros átomos da molécula, de acordo com a demonstração do Teorema 2, e procede-se a primeira iteração do AT, obtendo cinco átomos determinados. Para cada átomo x_j seguinte, busca-se quatro átomos determinados, não-coplanares, com distâncias conhecidas entre si e com x_j e cujas coordenadas são x_{j1}, x_{j2}, x_{j3} e x_{j4} em $\mathbb{R}^3$, as quais serão armazenadas como linhas de uma matriz X. Recalculamos as coordenadas dos átomos x_{j1}, x_{j2}, x_{j3} e x_{j4}, como na demonstração do Teorema 2, obtendo os vetores y_{j1}, y_{j2}, y_{j3} e y_{j4}, os quais são armazenados como linhas de uma matriz Y. Este processo é chamado *reinicialização* dos quatro vetores. Observe que os centros geométricos das estruturas tridimensionais representadas pelas matrizes X e Y são calculados, respectivamente, da forma

$$x_c(k) = \frac{1}{4}\sum_{i=1}^{4} X(i,k) \quad \text{e} \quad y_c(k) = \frac{1}{4}\sum_{i=1}^{4} Y(i,k), \tag{11}$$

para $k = 1,2,3$. Logo, realiza-se a translação de Y,

$$\begin{array}{l} Y(i,1) = Y(i,1) - [y_c(1) - x_c(1)] \\ Y(i,2) = Y(i,2) - [y_c(2) - x_c(2)] \\ Y(i,3) = Y(i,3) - [y_c(3) - x_c(3)] \end{array}, \tag{12}$$

de modo que as estruturas representadas por X e Y fiquem sobrepostas e tenham o mesmo centro geométrico [4]. Para finalizar, resta-nos rotacionar Y para que os quatro átomos, recalculados, voltem o mais próximo de suas posições originais. Para tanto, lançamos mão da RMSD (Root-Mean-Square Deviation), definida por

$$RMSD(X,Y) = min_Q \frac{\|X - YQ\|_F}{\sqrt{n}}, \tag{13}$$

onde X contém as coordenadas originais dos átomos da molécula, Y contém as coordenadas determinadas pelo método, a matriz Q é a matriz de rotação que melhor alinha X e Y, estando sobrepostos sobre o mesmo centro geométrico, e $\|\cdot\|_F$ é a norma de Frobenius. Neste procedimento, a matriz Q que minimiza a função RMSD é dada por $Q = UV^t$, onde U e V são as componentes ortogonais da decomposição em valores singulares da matriz $C = Y^tX$, isto é, $C = U\Sigma V^t$ [4]. Usando os vetores y_{j1}, y_{j2}, y_{j3} e y_{j4}, aplicamos o passo principal do AT, ou seja, resolvemos o sistema linear do Teorema 1, para descobrir a posição y_j do átomo indeterminado no novo sistema de coordenadas. Após esta descoberta, voltamos o vetor y_j para o sistema de coordenadas original, obtendo a posição x_j do átomo indeterminado da iteração. Este procedimento é repetido até que não seja possível determinar as posições dos átomos restantes da molécula, ou pelo mal condicionamento do sistema linear que gera uma instabilidade numérica ou pela impossibilidade de se encontrar os quatro átomos base da iteração. Desse modo, o método se encerra, reportando uma estrutura parcial da molécula. Segue o pseudo-algoritmo do ATA.

Algoritmo (ATA)**.** *Dado um conjunto D de distâncias euclideanas entre pares dentre os n átomos de uma molécula:*

1. *Encontre quatro átomos não-coplanares e com distâncias conhecidas entre si.*
2. *Determine suas coordenadas x_1, x_2, x_3 e x_4, segundo a demonstração do Teorema 2.*
3. *Faça:*
   - *(i) Para cada átomo não-determinado j, se possível encontre quatro átomos determinados,*

não-coplanares, com distâncias conhecidas entre si e com o átomo j, x_{j1}, x_{j2}, x_{j3} e x_{j4}. Caso não encontre, pare.

(ii) Utilizando apenas as distâncias disponíveis, reinicialize os quatro átomos, ou seja, encontre as coordenadas y_{j1}, y_{j2}, y_{j3} e y_{j4}, segundo a demonstração do Teorema 2.

(iii) Resolva o sistema $Ax = b$, do Teorema 1, usando y_{j1}, y_{j2}, y_{j3} e y_{j4}, obtendo y_j^*.

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

(v) Coloque os átomos de volta à estrutura original.

4. *Se nenhum átomo é determinado em todo o loop, pare e reporte a estrutura parcial. Senão, volte ao passo (3).*

Assim como comentado na seção anterior, os algoritmos ATA e UGB são semelhantes estruturalmente, diferindo pela dimensão do sistema linear resolvido. No caso do ATA, a dimensão é igual a 4.

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

Observação 1. *Há outros métodos que mostram que é possível encontrar mais de uma solução para PGDM [9], ou seja, determinar mais de uma configuração tridimensional para seus átomos, respeitando as restrições de distâncias. Estas soluções podem ser semelhantes (i.e., podem ser sobrepostas a partir da aplicação de movimentos rígidos, como translação e/ou rotação) ou não. Tanto o ATA quanto o UGB fornecem, apenas, uma delas.*

4 Resultados Numéricos

Nesta seção, iremos comparar os algoritmos UGB e ATA. Para tanto, geramos instâncias artificiais, conforme descrito em [8], escolhidas pela simplicidade de implementação, a fim de simular estruturas moleculares. Essas estruturas são estabelecidas por uma cadeia de átomos enumerados de 1 a n, onde a distância entre dois átomos consecutivos é igual a $1.5\mathring{A}$. Então, calculamos as distâncias entre todos os átomos, mas levamos em consideração, apenas, as que são menores que $6\mathring{A}$, para simular dados de RMN [8]. Essas, que restaram, formam um conjunto de distâncias considerado como a instância artificial de nossos dois algoritmos. Além disso, observa-se que estas estruturas artificiais representam algumas características qualitativas de um problema real [8].

Para efeito de verificar a qualidade do algoritmo proposto, é realizada uma comparação entre a estrutura original, gerada de forma artificial, e aquela reconstruída pelos algoritmos UGB e ATA, por meio da RMSD, definida anteriormente. Quanto maior o valor da RMSD, maior será a distância entre a estrutura gerada pelo método e a gerada artificialmente. O objetivo desta medida é avaliar o quanto ambas estruturas tridimensionais se parecem geometricamente, ou seja, o quanto podem estar bem alinhadas no espaço tridimensional. É possível que o método gere uma estrutura que respeite as distâncias, mas não tenha forma semelhante à da instância testada (veja a Observação 1). Entretanto, nossa intenção é buscar estruturas tridimensionais parecidas geometricamente com a instância testada, além de que suas distâncias sejam compatíveis com as restrições. Logo, escolhemos a RMSD de modo a testar se a solução obtida satisfaz tais condições, visando recuperar a estrutura original. Assim, grandes valores para a RMSD mostram que o algoritmo não recuperou a estrutura como da instância testada.

No método UGB, a cada iteração, são construídos quatro sistemas lineares de ordem 3. Dentre estes, apenas o que possui menor número de condição de sua matriz de coeficientes é resolvido. Esta medida visa garantir mais estabilidade numérica evitando, assim, uma excessiva propagação de erros nos cálculos subsequentes [13]. Essa decisão demanda mais tempo de processamento. Já no ATA, é preciso construir um único sistema linear de ordem 4, por iteração. Como não é necessário

fazer buscas prévias, essa estratégia representa uma vantagem. Para resolver esses sistemas lineares, usamos fatoração LU com estratégia de pivoteamento parcial. Além disso, a cada iteração no ATA, há um fator de comparação para decidir se a solução para a posição requerida é aceitável ou não, a saber, t_j definido no método. Este termo é estimado pelo próprio algoritmo e, também, é calculado em função da solução obtida, funcionando como critério de parada alternativo.

# Átomos	Tempo ATA (s)	RMSD ATA	Tempo UGB (s)	RMSD UGB	Tempo(ATA/UGB)
5	2,3300E-02	6,3672E-16	3,6700E-02	6,9688E-16	63,49%
20	5,5000E-03	3,2133E-15	1,2500E-02	3,3246E-15	44,00%
40	9,2632E-03	1,6979E-14	1,1630E-02	1,2573E-14	79,65%
60	1,2187E-02	2,5530E-14	1,4811E-02	2,9629E-14	82,28%
80	1,4455E-02	1,1453E-13	2,1918E-02	1,1859E-13	65,95%
100	1,7803E-02	4,7948E-14	2,7703E-02	8,0261E-14	64,26%
200	2,9270E-02	1,3678E-13	5,4015E-02	2,4121E-13	54,19%
300	5,8600E-02	9,7663E-13	9,6600E-02	8,3107E-13	60,66%
500	1,0610E-01	1,3605E-12	1,5170E-01	1,3755E-12	69,94%
1000	1,4398E-01	1,2040E-11	2,7506E-01	7,9073E-12	52,34%
2000	2,7278E-01	3,1801E-11	4,8366E-01	8,5692E-11	56,40%
4000	5,7585E-01	1,7805E-10	9,7256E-01	4,0795E-10	59,21%
6000	8,6527E-01	1,0882E-09	1,4000E+00	7,8574E-10	61,81%
8000	1,1181E+00	2,6738E-09	1,8639E+00	2,3254E-09	59,99%
10000	1,4461E+00	4,5426E-09	2,6326E+00	4,0318E-09	54,93%
12000	3,5009E+00	6,0816E-09	4,3294E+00	5,3836E-09	80,86%
15000	5,6546E+01	1,0190E-08	6,8244E+01	7,6604E-09	82,86%

Tabela 1: Tabela com valores de RMSD e tempo de CPU para o ATA e o UGB relativo às estruturas artificiais geradas.

Na Tabela 1 são apresentados resultados numéricos preliminares realizados com as instâncias artificiais geradas [4, 8]. Na primeira coluna, temos o número de átomos presentes na estrutura artificial testada. Na segunda e terceira colunas, apresentamos os valores de tempo de CPU, em segundos, e os valores da RMSD para o ATA. A quarta e quinta colunas apresentam os mesmos dados, mas para o UGB. Por fim, a última coluna da tabela apresentada mostra o valor relativo (ATA/UGB) entre os tempos de CPU. Visando comparar estruturas de diversas naturezas, realizamos testes com instâncias artificiais de pequeno porte com cinco átomos, o mínimo possível para o funcionamento do método, até estruturas de grande porte com quinze mil átomos. Em comparação com o UGB, os resultados são satisfatórios, com respeito à precisão e, para os resultados obtidos em tempo de CPU, o ATA exibiu um melhor desempenho. É possível ver isto tanto pelo gráfico da Figura 2 quanto pela última coluna da tabela mostrada neste trabalho, que apresenta o tempo relativo entre os dois métodos. Nos testes realizados, com ambos os algoritmos, UGB e ATA, os valores da RMSD são qualitativamente da mesma ordem de grandeza, mesmo para estruturas de grande porte, como no caso de 15000 átomos. Pode-se ver este fato pelo gráfico da Figura 1. Portanto, o método ATA, aqui proposto, é tão robusto e preciso quanto o UGB, e obteve resultados computacionais satisfatórios e promissores.

Ambos os métodos foram implementados em linguagem de programação Matlab, em uma máquina com processador Intel Core $i5-2410M$, 2.3 GHz e com memória RAM de 4 GB.

5 Conclusão

Neste trabalho, apresentamos um novo método para resolver Problemas de Geometria de Distâncias Moleculares com Conjuntos Arbitrários de Distâncias, chamado Algoritmo T (AT), que fora inicialmente introduzido em [4]. Ambos, GB e AT, têm como cerne a resolução de sistemas lineares de

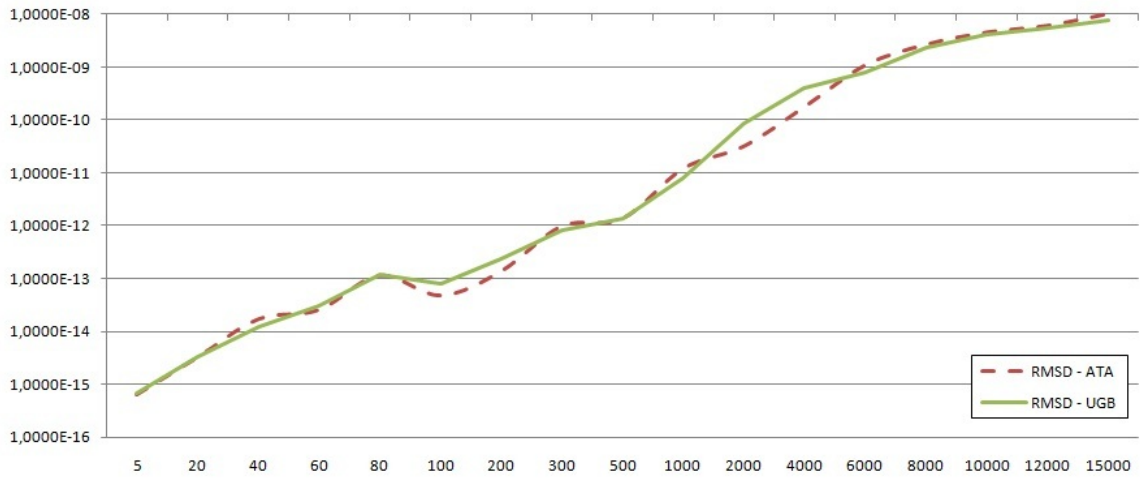

Figura 1: Gráfico com os valores da RMSD para UGB e ATA.

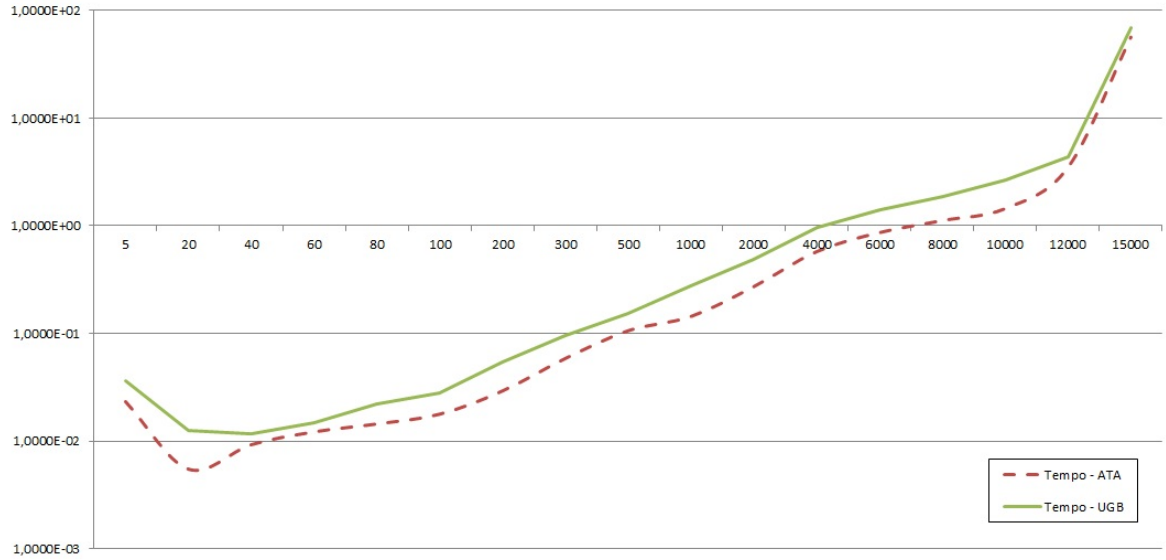

Figura 2: Gráfico com os tempos de CPU para UGB e ATA.

pequeno porte e mal-condicionados. Em [13], há uma melhora no GB, o *Updated Geometric Algorithm (UGB)*. Como comentado anteriormente, os átomos base são tratados a cada iteração de modo a diminuir o acúmulo de erros para não afetar o cálculo das posições dos átomos que seguem. No presente trabalho, a mesma atualização dos átomos base para a determinação do átomo não-posicionado é também realizada, em cada iteração. Com esta modificação, obtemos o Algoritmo T Atualizado (ATA). Os algoritmos AT e ATA podem ser vistos como uma extensão do método GB introduzido em [3], com a possibilidade de tratar incertezas/imprecisões inerentes no processo de medição de problemas de geometria de distâncias moleculares, já que estes algoritmos contam com a variável $t_j = -\left\|x_j^*\right\|^2/2$, descrita na seção 2, que pode ser usada como uma medida, a cada iteração, da imprecisão dos dados. Esse importante aspecto está sendo explorado [5] para compreender em detalhes todo seu potencial para essa classe de problemas, bem como suas limitações.

Os resultados obtidos com o ATA em relação ao UGB são promissores, pois mostram eficiência no cálculo das estruturas e em tempo relativamente menor, controlando, assim, a propagação de

erros numéricos. Temos algumas outras linhas de trabalho em andamento [5]:

- Melhorar o condicionamento das matrizes dos sistemas lineares resolvidos a cada iteração.
- Estender o método para tratar os problemas onde somente cotas inferiores e superiores para os valores de distâncias são fornecidos pela RMN.
- Considerar instâncias do PGDM provenientes de estruturas do *Protein Data Bank*.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq - e à Fundação de Amparo à Pesquisa do Estado de Sâo Paulo - FAPESP, pelo suporte financeiro, e à Universidade Estadual de Campinas. F. Fidalgo agradece ao CNPq por fomento aos projetos 133311/2009-0 e 142379/2011-5. E.Abreu agradece à FAPESP pelo suporte ao projeto 2011/11897-6 e à UNICAMP/FAEPEX (Fundo de Apoio ao Ensino, Pesquisa e Extensão) pelo apoio ao projeto 519.292-785/11. Os autores agradecem, também, aos revisores anônimos pelas sugestões e críticas pertinentes e que contribuíram com o formato final do artigo.

Referências

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

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

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

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

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

[6] **Glunt, W., Hayden, T., e Raydan, M.** (1993), *Molecular conformations from distance matrices*, Journal of Computational Chemistry 14, pp. 114-120.

[7] **Havel, T.** (1995), *Distance Geometry*, Encyclopedia of Nuclear Magnetic Resonance, John Wiley and Sons, pp. 1702-1710.

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

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

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

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

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

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