Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 2, N. 1, 2014.
Trabalho apresentado no CMAC-Sul, Curitiba-PR, 2014.

MÉTODO DE INTERSEÇÃO DE ESFERAS APLICADO AO CÁLCULO DE ESTRUTURAS DE PROTEÍNAS

Inajara da Silva Freitas
Programa de Pós Graduação em Métodos Numéricos em Engenharia - UFPR
E-mail: inajara@ufpr.br.

Luiz Carlos Matioli
Departamento de Matemática - UFPR
R. Cel. Francisco Heráclito dos Santos, 210, CEP 81531-970
Jardim das Américas, Curitiba - PR
E-mail: matioli@ufpr.br.

Resumo: *Neste trabalho discutiremos sobre o problema de determinar a estrutura de uma proteína quando um conjunto de distâncias entre seus átomos são conhecidas. Este problema também é conhecido como problema geométrico de distância molecular. Serão abordadas duas visões diferentes. Primeiro, formularemos o problema para que ele possa ser resolvido de forma linear, utilizando técnicas básicas de álgebra linear. Em seguida, falaremos sobre um método quadrático que é formulado utilizando o cálculo dos pontos de interseção entre esferas. Para encontrar os pontos de interseção entre esferas no $\mathbb{R}^n$, dado um conjunto de n equações não-lineares, desejamos encontrar sua solução resolvendo um sistema de equações quadráticas. Os métodos foram implementados e testados no MATLAB e em seguida comparados utilizando o Root-Mean-Square-Deviation (RMSD), que serve para medir o erro acumulado.*

Palavras-chave: *Interseção de Esferas, Estruturas de Proteínas, Root-Mean-Square-Deviation.*

1 INTRODUÇÃO

O Protein Data Bank [6] é um repositório que contém informações de coordenadas atômicas e outras informações que servem para descrever proteínas e ácidos nucleicos importantes, tais como vírus e proteínas.

Para descobrir a estrutura dessas proteínas são utilizados métodos como difração de raios X, ressonância magnética nuclear (RNM) e crio-microscopia de elétrons, e com isso determinar a localização de cada átomo em relação ao outro na molécula. Essa informação, é disponibilizada livremente pelo wwPDB.

Ao utilizar o método RNM temos como resultado a distância entre os átomos do organismo observado. Para formular a estrutura de uma proteína precisamos descobrir qual é a menor distância entre cada um de seus átomos, com isso temos o "problema geométrico de distância molecular" mais conhecido como MDGP (molecular distance geometry problem), que pode ser representado da seguinte forma:

Dado um conjunto C de pares de átomos (i, j) de um conjunto de n átomos com suas distâncias d_{ij} definida sobre C, ache as posições $x_1, ..., x_m \in \mathbb{R}^3$ de átomos na estrutura da molécula de modo que

$$\|x_i - x_j\| = d_{ij} \quad \forall (i, j) \in C \tag{1}$$

DOI: 10.5540/03.2014.002.01.0114

2 Estruturas de Proteínas

Inicialmente iremos apresentar o método usado em [4], onde foi utilizado a hipótese de que todas as distâncias entre um número finito de átomos são conhecidas, fixaremos este número como n átomos. Em um espaço tridimensional se conhecermos a localização de 4 destes átomos e eles não estão contidos em um mesmo plano, podemos formar um sistema para encontrar todos os átomos restantes de maneira única. Iremos denotar as coordenadas destes 4 pontos por:

$$x_j = (u_j, v_j, w_j)^T,\ para\ j = 1, ..., 4. \tag{2}$$

Queremos determinar as coordenadas $x_i = (u_i, v_i, w_i)^T$ de um certo átomo. Como sabemos as distâncias de todas as possíveis combinações de pares de átomos, denotamos a distância entre o átomo i e j por $d_{i,j}$ onde $j = 1, ..., 4$ o que nos possibilita obter as seguintes equação:

$$||x_i - x_j|| = d_{i,j},\ para\ j = 1, ..., 4. \tag{3}$$

Elevando ao quadrado os dois lados da igualdade e substituindo $x_i^T x_j$, por $(u_i, v_i, w_i)(u_j, v_j, w_j)^T$ onde $j = 1, ..., 4$:

$$||x_i - x_j||^2 = ||x_i||^2 - 2u_iu_j - 2v_iv_j - 2w_iw_j + ||x_j||^2 = d_{i,j}^2 \tag{4}$$

Subtraindo a primeira equação das outras e representando de forma matricial, temos $Ax_i = b_i$, onde:

$$A = 2\begin{pmatrix} u_1 - u_2 & v_1 - v_2 & w_1 - w_2 \\ u_1 - u_3 & v_1 - v_3 & w_1 - w_3 \\ u_1 - u_4 & v_1 - v_4 & w_1 - w_4 \end{pmatrix}\ \ e\ \ b_i = \begin{pmatrix} (||x_1||^2 - ||x_2||^2) - (d_{i,1}^2 - d_{i,2}^2) \\ (||x_1||^2 - ||x_3||^2) - (d_{i,1}^2 - d_{i,3}^2) \\ (||x_1||^2 - ||x_4||^2) - (d_{i,1}^2 - d_{i,4}^2) \end{pmatrix} \tag{5}$$

Este sistema pode ser resolvido em tempo linear. Após o cálculo inicial dos quatro primeiros átomos são executadas no máximo n operações, onde n é o número de átomos em cada molécula. Com isso é possível determinar a estrutura da molécula, onde a distância exata entre cada átomo é dada.

Resultados Computacionais

O algoritmo foi implementado em MATLAB, chamaremos este algoritmo de "Método Linear", como dado de entrada precisamos apenas definir o nome da proteína escolhida. Com esse dado o MATLAB buscará a estrutura da proteína no *site* http://www.wwpdb.org. Este *site* contém uma base de dados de estruturas biológicas macromoleculares. No MATLAB podemos extrair esta matriz de todos os pontos que formam a estrutura da proteína, o que nos permite criar uma matriz de distâncias exatas e fazendo operações simples podemos calcular a distância entre todos os pontos.

Seja $x_i = (u_i, v_i, w_i)^T$, para fixarmos o primeiro átomo na origem basta fixar $u_1 = 0$, $v_1 = 0$ e $w_1 = 0$, em seguida fixamos o segundo átomo em um dos eixos, denotando $u_2 = d_{2,1}$, $v_2 = 0$ e $w_2 = 0$, onde $d_{2,1}$ é a distância entre os átomos 1 e 2. O terceiro átomo deve ser fixado de modo que não esteja no mesmo plano de x_1 e x_2, para que isto aconteça temos $w_3 = 0$. As outras duas coordenadas podem ser encontradas fazendo uma relação entre suas distâncias $d_{3,1}$ e $d_{3,2}$.

O quarto átomo deve ser escolhido de modo que não esteja no plano formado pelos 3 primeiros átomos. Podemos obter os valores de x_4 usando o mesmo método do átomo anterior. Com isto temos as coordenadas dos quatro primeiros átomos, já podemos determinar suas coordenadas restantes de forma única utilizando o algoritmo da seção anterior.

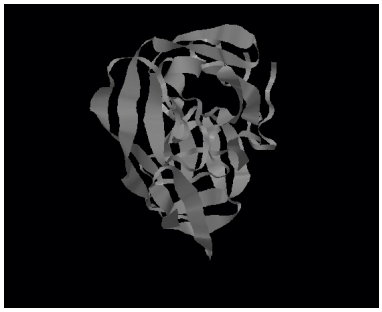
Para determinar a matriz de distâncias entre os átomos, tomamos os dados de todos os pontos de uma proteína e com estes pontos é possível fazer o cálculo estimado da matriz de distâncias. Lembrando que as distâncias não são exatas, devido acumulo de erros gerados pelo computador ao calcular operações necessárias para achar as distâncias originais. Abaixo segue uma tabela com os resultados dos testes:

DOI: 10.5540/03.2014.002.01.0114

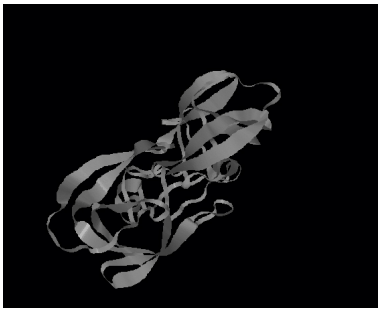
Proteína	Tempo/segundos	Nº de átomos
1HAA	0.018870	1310
1AJV	0.010216	1516
1PTQ	0.007361	402
1HOE	0.007767	558
1HIV	0.010275	1502

Tabela 1: Resultado dos testes - Método Linear

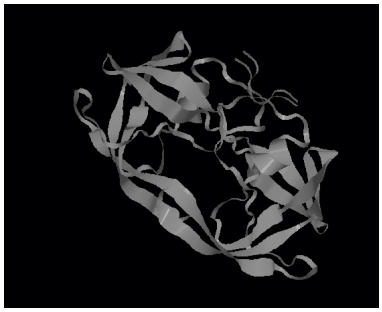
O MATLAB nos possibilita mostrar a estrutura de cada proteína. Como vemos abaixo para a proteína 1AJV. Podemos verificar que as estruturas parecem ser diferentes, mas ao girá-las é possível encontrar a posição na qual são semelhantes. Desta forma as figuras 2(a) e 2(b) foram giradas para obter 2(c) e 2(d).



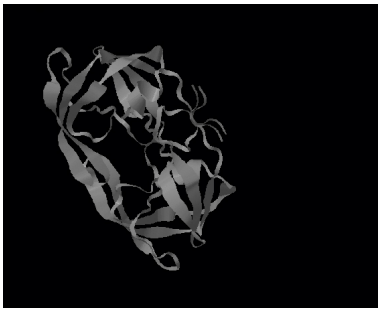
(a) Figura gerada com os dados PDB(Protein Data Bank)



(b) Figura gerada com o método linear apresentado



(c) Figura gerada com os dados PDB(Protein Data Bank)



(d) Figura gerada com o método linear apresentado

Figura 1: Figuras geradas pelo MATLAB

Como definimos, os valores de w_k para $k = 1, ..., 4$, a figura pode ficar de forma espelhada em relação a figura original. Para obtermos a figura original, basta usar w_k para $k = 1, ..., 4$ com valores negativos.

3 Determinação de Pontos de Interseção Entre Esferas

De acordo com [1], dados $x_i \in \mathbb{R}^n$ e $d_i \in \mathbb{R}^+$, $i = 1, ..., n$, onde x_i é representado como o centro de uma esfera E_i e d_i o raio correspondente.

O problema de interseção de esferas é encontrar $x \in \mathbb{R}^n$ tal que $\|x - x_i\|^2 = d_i^2,\ i = 1, ..., n$. Onde d_i é a distância entre os pontos x e x_i, ou seja o raio da esfera E_i.

DOI: 10.5540/03.2014.002.01.0114

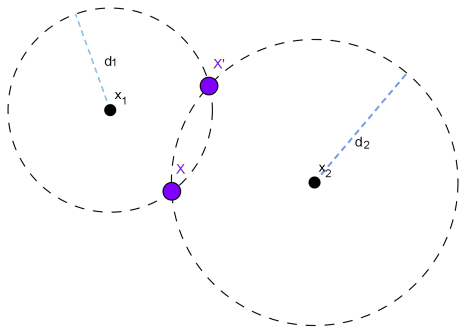


Figura 2: Pontos de interseção entre duas esferas no $\mathbb{R}^2$

Seja $A \in \mathbb{R}^{n\times n}$ uma matriz cujas colunas são x_i, $i = 1, ..., n$, onde $x_1, x_2, ..., x_n$ são linearmente independentes, desta forma a matriz A será não singular e portanto inversível. Como visto acima temos

$$\|x - x_i\|^2 = d_i^2 \quad i = 1, ..., n \iff x^T x - 2x^T x_i + x_i^T x_i = d_i^2,\ i = 1, ..., n \tag{6}$$

isolando $x^T x_i$ e substituindo $r = x^T x$ e $b_i = x_i^T x_i - d_i^2$ a equação ficará da forma

$$A^T x = \frac{re + b}{2} \tag{7}$$

onde $e = [1, 1, 1, ..., 1, 1]^T \in \mathbb{R}^n$ e $b = [b_1, ..., b_n]^T$. Como definimos A invertível, existe A^{-1} tal que

$$A^{-T}A^T x = A^{-T}\frac{(re + b)}{2} \iff x = \frac{A^{-T}re + A^{-T}b}{2}. \tag{8}$$

Considere $u = A^{-T}e$ e $v = A^{-T}b$, como $r = x^T x$, temos

$$r = \frac{2 - u^T v \pm \sqrt{(u^T v - 2)^2 - (u^T u)(v^T v)}}{u^T u} \tag{9}$$

4 Aplicação do Método de Interseção de Esferas

Nesta seção iremos analisar e implementar um método baseado em interseção de esferas para resolver o problema de estruturas de proteínas. Para encontrar a distância entre dois pontos, partimos do seguinte problema:

$$\|x_i - x_j\| = d_{i,j},\ \ para\ j = 1,\ 2,\ 3,\ 4\ e\ i = 5 : n \tag{10}$$

onde o ponto x_i é o átomo procurado e d_{ij} uma distância conhecida. Através de operações básicas concluimos que

$$2x_i(x_1 - x_j) = \|x_1\|^2 - \|x_j\|^2 - (d_{i,1}^2 - d_{i,j}^2) \tag{11}$$

para $j = 2,\ 3,\ 4$. Podemos observar que esta equação não é linear, portanto não é possível aplicar o método como foi descrito na seção anterior. Como estamos em $\mathbb{R}^3$ precisamos apenas de 3 vetores linearmente independentes, iremos usar os pontos $x_2,\ x_3,\ x_4$ que foram construídos de maneira que sejam linearmente independentes e manteremos x_1 como ponto inicial para que a estrutura comece na origem. A matriz A é dada por:

$$A = [x_2,\ x_3,\ x_4] = \begin{pmatrix} u_2 & u_3 & u_4 \\ 0 & v_3 & v_4 \\ 0 & 0 & w_4 \end{pmatrix} \tag{12}$$

DOI: 10.5540/03.2014.002.01.0114

As distâncias d_{ij} para $j = 2,\ 3,\ 4$ serão os raios das esferas de centros x_j.

Ao utilizar o Método de Interseção de esferas podemos obter:

- 1 ponto x_i - neste caso x_i será o ponto procurado.
- 2 pontos x_{i1} e x_{i2} - podemos descobrir qual é o ponto correto verificando qual deles está mais próximo da distância d_{i1}.
- 0 pontos - este caso não deverá ocorrer visto que todas as distâncias existem. Para garantir este resultado basta colocar um $\varepsilon > 0$ e verificar $(u^T v - 2)^2 - (u^T u)(v^T v) <= \varepsilon$ ao invés da condição $(u^T v - 2)^2 - (u^T u)(v^T v) = 0$.

Resultados Computacionais

O algoritmo foi implementado no MATLAB, com base no algoritmo de interseção de esferas, chamaremos de "Método Quadrático".

Abaixo segue a tabela com os resultados dos testes:

Proteína	Tempo/segundos	Nº de átomos
1HAA	0.146791	1310
1AJV	0.168065	1516
1PTQ	0.065965	402
1HOE	0.078101	558
1HIV	0.165347	1502

Tabela 2: Resultado dos testes - Interseção de esferas

Neste caso também pode ocorrer das figuras estarem espelhadas o que pode ser resolvido escolhendo x_i no caso em que existem dois pontos de interseção de esferas, como sendo a maior distancia até d_{i1}.

5 Comparação dos Métodos

Usaremos [2], [3] e [5] para comparar os dois métodos. Seja X a matriz dos pontos originais do PDB e Y a matriz de pontos recalculados, primeiro precisamos colocar as duas estruturas no mesmo centro geométrico.

Podemos obter este resultado calculando

$$xc(j) = \frac{1}{n}\sum_{i=1}^{n} X(i,j) \quad yc(j) = \frac{1}{n}\sum_{i=1}^{n} Y(i,j) \tag{13}$$

para $j = 1,\ 2,\ 3$.

Os valores de xc e yc serão usados para atualizar os valores de X e Y, fazendo uma translação:

$$X_1(:,j) = X(:,j) - xc(j) \quad Y_1(:,j) = Y(:,j) - yc(j) \tag{14}$$

Agora as estruturas X_1 e Y_1 estão transladadas para origem. Para rotacioná-las utilizaremos o cálculo de Root-Mean-Square Deviation (RMSD), ele serve para medir o grau de semelhança de duas estruturas X_1 e Y_1 que possuem o mesmo centro geométrico, definimos:

$$RMSD(X_1, Y_1) = \frac{min_Q \|X_1 - Y_1 Q\|_F}{\sqrt{n}} \tag{15}$$

Onde Q_{3x3} é a matriz de rotação e $QQ^T = I$.

A tabela abaixo mostra os resultados para os métodos linear e quadrático (Interseção de esferas).

Proteina	Nº de átomos	Linear	Quadrático
1HAA	1310	4.971775530784199e-04	4.971775530784199e-04
1AJV	1516	4.991662587665603e-04	4.991662587665603e-04
1PTQ	402	5.024860013613150e-04	5.024860013613150e-04
1HOE	558	4.899717950387293e-04	4.899717950387293e-04
1HIV	1502	5.037966418678951e-04	5.037966418678951e-04

Tabela 3: Resultado dos testes - RMSD

Conclusões

Neste trabalho analisamos dois métodos para determinar a estrutura molecular de proteínas. Um método linear, que podemos utilizar para analisar um sistema linear do tipo $Ax = b$ e um método quadrático, onde precisamos determinar um sistema equações quadráticas para achar a solução do sistema. Tanto o método linear quanto o quadrático, usam matrizes fixas de dados iniciais para o cálculo dos pontos seguintes, nos dois casos são utilizados 4 pontos. No caso linear os pontos formam a matriz inicial, no caso quadrático utilizamos 3 pontos iniciais e um quarto ponto é necessário para fixar uma das soluções obtidas.

Como estamos usando apenas um subconjunto das distâncias, estes pontos podem ser praticamente coplanares o que causa uma inconsistência nos dados. Uma maneira de solucionar este problema seria verificar se a matriz é consistente antes de calcular o próximo ponto, caso não seja, teríamos que calcular uma nova matriz entre os pontos obtidos, mas adicionar esse passo ao algoritmo resultaria em um processo muito lento.

Apesar dos métodos apresentados fornecerem os mesmos resultados, o método linear tem uma vantagem sobre o método quadrático pois ele executa menos operações, o que o torna mais rápido.

Além da distância atômica obtida pelo método RNM não ser exata, pois este método utiliza ondas de radio para obter uma lista de núcleos atômicos que estão perto um do outro, também impossibilita encontrar as distâncias entre todos os átomos. Os dados não obtidos são substituídos por zeros, o que acaba criando uma matriz esparsa. Existem algoritmos capazes de estimar algumas dessas distâncias, mas acabam produzindo um erro muito grande na estrutura final, isso impossibilita o uso dos métodos apresentados para calcular uma estrutura com dados reais.

Referências

[1] Coope I.D., Reliable computation of the points of intersection of n spheres in $\mathbb{R}^n$.*Journal of Global Optimization*, pg 365 - 375, 2002.

[2] Davis, R. T.; Ernst, C.; Wu, D., Protein Structure Determination Via An Efficient Geometric Build-Up Algorithm. *BMC Structural Biology*, 2010.

[3] Di Wu, Distance-Based Protein Structure Modeling. *Tese de Doutorado, Iowa State University*, 2006.

[4] Dong e Wu, A Linear-Time Algorithm For Solving The Molecular Distance Geometry Problem With Exact Inter-Atomic Distances. *Journal of Global Optimization*, pg 365 - 375, 2002.

[5] Souza Michael F., Suavização Hiperbólica Aplicada À Otimização De Geometria Molecular. *Tese de Doutorado*, UFRJ, 2010.

[6] RCSB PDB, RCSB *Protein Data Bank: www.rcsb.org/.*

