

O Algoritmo de Branch and Prune para obter formato de moléculas

Francisco Lenilson de Souza do Nascimento

Resumo

Conhecer o formato de moléculas e proteínas assim como sua estrutura em três dimensões têm desempenhado um papel de grande importância na indústria farmacêutica no que diz respeito ao desenvolvimento de novos remédios entre outras aplicações. Este artigo tem como objetivo estudar o código de branch and prune que se propõe fornecer posições de moléculas através de dados que podem ser obtidos por meio de uma ressonância nuclear magnética. O código usa o conceito de geometria de distância para obter posições dos átomos no plano tridimensional. Foi abordada ainda o índice de participação de artigos brasileiros assim como os principais nomes do país a estudar o tema, e pôde-se constatar que esta é uma área praticamente inexplorada em nosso país.

Palavras-chave: estrutura molecular; geometria de distância; *Branch and Prune*

1 INTRODUÇÃO

Desde 1953, sabe-se da correlação entre o formato da molécula e a função desta, conforme foi comprovado pelo bioquímico James Watson e o biofísico Francis Crick que, através da imagem de uma molécula de ácido desoxirribonucleico, propuseram um modelo tridimensional que explicava todos os dados disponíveis em estudos realizados sobre DNA (LAVOR et al, 2017).

Obter o formato das moléculas é de interesse de várias áreas de pesquisa, como a farmacêutica, robótica e redes neurais (LIBERTI et al, 2014). No entanto, os átomos são invisíveis, uma vez que são menores que o menor comprimento de onda visível.

Precisa-se, portanto, de outras maneiras de se "enxergá-las". Uma delas é por meio de uma máquina de ressonância magnética nuclear. Esta máquina não indica o formato da molécula, mas mede a distância entre os átomos. O problema então é, a partir das distâncias dadas, obter as posições dos átomos. A área da matemática que estuda este problema é a Geometria das Distâncias.

Este trabalho tem dois objetivos: fazer um levantamento bibliográfico sobre os estudos recentes de geometria molecular, em especial com a aplicação de determinar o formato de moléculas e estudar uma das técnicas de se determinar o formato de molécula, o algoritmo de *Branch and Prune*.

Ao final deste trabalho, faremos ainda uma implementação computacional do método e faremos experiências com dados aleatórios.

2 REVISÃO BIBLIOGRÁFICA

Na primeira análise, buscamos estudar a correlação entre a área matemática "Geometria de Distâncias" com sua principal aplicação, a determinação de formato de moléculas. Buscando-se as palavras chave "geometria de distâncias" para trabalhos publicados de 2016 a 2020, a plataforma Google Acadêmico retornou 24 resultados. Ao se acrescentar a palavra chave "molécula", retornou 16 resultados.

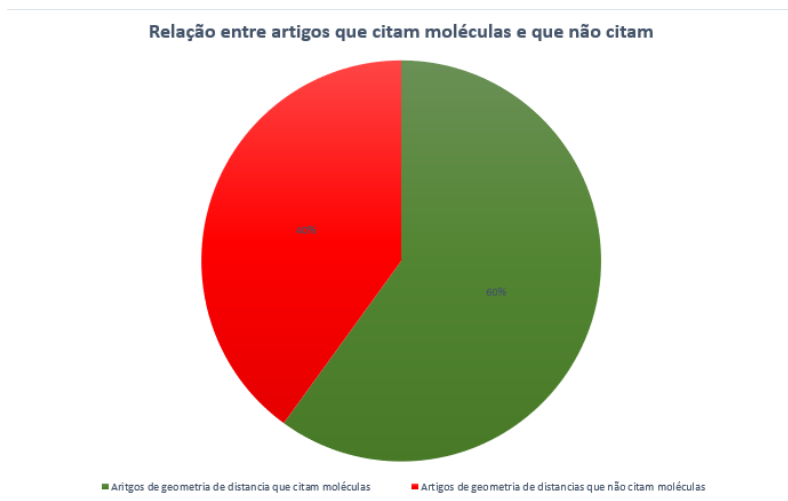


Figura 1 – A maioria dos artigos que tratam de geometria de distâncias relacionam o tema à cálculo estrutural de moléculas

Ao se procurar "geometria de distância", no singular, surge um novo artigo (CHP Rodrigues 2016), que trata de um assunto diverso ao estudado (trabalho sobre química com aplicação forense)

Uma segunda análise sobre os estudos de geometria de distância seria identificar o principal autor brasileiro na área. Inicialmente pôde-se notar a presença massiva do professor Carlile Campos Lavor, professor da Universidade Estadual de Campinas (Unicamp) e presidente da Sociedade Brasileira de Matemática Aplicada (SBMAC). Dentre os 24 artigos brasileiros encontrados, pôde-se observar que 20 mencionam (ou foram escritos por) Lavor.



Figura 2 – Dos artigos brasileiros publicados a maior parte tem alguma influência de Lavor.

A última análise é sobre a presença de pesquisadores brasileiros no campo de estudo. Ao se buscar na plataforma Google Acadêmico "distance geometry", retornou-se

muitos resultados. Por outro lado, uma análise superficial mostrou que poucos tratavam da geometria de distâncias conforme foi definido neste trabalho. Por isso, filtramos os resultados com as palavras chave "distance geometry molecules". Desta forma, obtivemos aproximadamente 65100 trabalhos.

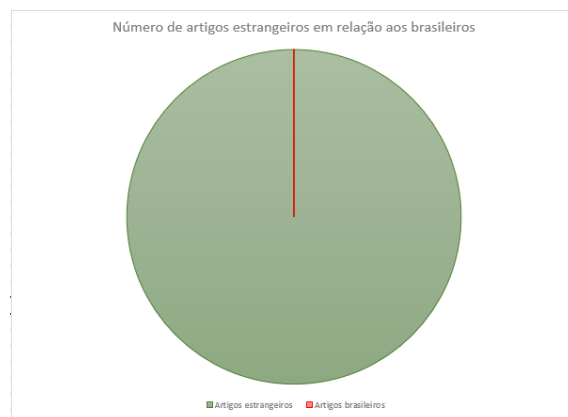


Figura 3 – Quase nenhum trabalho sobre o tema foi escrito por brasileiros.

Isso mostra que a produção brasileira corresponde a apenas 0,004% da produção no mundo inteiro, um indicativo que o assunto ainda é inexplorado no Brasil.

3 DESENVOLVIMENTO

Segundo o glossário de termos usados no planejamento de fármacos, geometria de distâncias é definida como

“um método matemático usado para construir modelos moleculares tridimensionais (3D) a partir de um conjunto de distâncias interatômicas aproximadas (por exemplo, experimentos de efeito nuclear Overhauser (NOE, do inglês nuclear Overhauser effect) em ressonância magnética nuclear (RMN) sugerem apenas faixas de distâncias)”(SANTANA,2002).

Já segundo Liberti et al (2014), a geometria de distância é o estudo da geometria euclidiana no qual os dados de entrada consistem em um conjunto incompleto de distâncias e a saída é um conjunto de pontos no espaço euclidiano. Embasada principalmente pela computação e a matemática, em 1928 surge a ideia de geometria de distâncias GD a partir de vários conceitos geométricos usando a ideia de distância. Entretanto apenas em 1953 veio a se tornar uma área do conhecimento a partir dos resultados apresentados por Blumenthal intitulado *The theory and Applications of the Distance Geometry* chamada geometria de distância (LIMA; CAMARGO, 2016).

O método clássico de se resolver sistemas não-lineares é o algoritmo de newton-raphson (NR). Este método é utilizado com o intuito de obter aproximações decimais de raízes de funções deriváveis, sob condições apropriadas, podendo ser estendido a funções complexas holomorfas bem como as funções reais de várias variáveis. Basicamente esse método consiste em linearizar a função, buscar a raiz dessa linearização e repetir esse processo até obter-se a precisão desejada. No entanto, este método não é o ideal para se resolver este problema pois não temos como garantir que a aproximação inicial estará na bacia de convergência.

O algoritmo de *Branch and Prune* (BP) foi desenvolvido em 1960 por A.H Land e A.G Doig e foi proposto como um método de resolver variáveis discretas não negativas. Entretanto em 1999 foi resgatado por Jens Clausen com o intuito de resolver o problemas de rota de veículos, agendamento de equipe e planejamento de produção, todos problemas NP-difíceis.

Dizer que um algoritmo é NP-difícil, diz respeito a sua classificação quanto ao seu grau de complexidade computacional. Um problema é NP-difícil se um algoritmo para resolvê-lo puder ser traduzido em um algoritmo para resolver qualquer problema NP-problema (tempo polinomial não-determinístico). NP-difícil, portanto, significa "pelo menos tão difícil quanto qualquer problema NP", embora possa, de fato, ser mais difícil (WEISSTEIN, ERIC W., 2020). Como a quantidade de cálculos deste algoritmo cresce exponencialmente, é um problema NP-difícil.

Neste trabalho nos baseamos nos códigos escritos por Lavor para descrever o algoritmo de branch and prune e para realizar as experiências.

Os códigos foram escritos na linguagem Wolfram, presente no software Mathematica. Esta linguagem foi desenvolvida em 1988, por Stephen Wolfram, com o objetivo de ser o mais abrangente possível, com ênfase em computação simbólica, programação funcional e programação baseada em regras. É construída de modo a representar estruturas e dados arbitrários. É ainda muitas vezes erroneamente chamada de "Mathematica", que é o nome do principal aplicativo que trabalha com esta linguagem. (WOLFRAM, 2013)

3.1 O problema e principais definições

O problema de se determinar a posição cartesiana dos átomos pode ser matematicamente descrito pela busca da raiz da função

$$f : \mathbb{R}^{3n} \rightarrow \mathbb{R}$$

$$f(x_1, \dots, x_n) = \sum_{(i,j) \in E} (\|x_i - x_j\| - d_{ij})^2,$$

onde n é a quantidade de átomos da molécula, x_i é a posição cartesiana do átomo i ($x_i \in \mathbb{R}^3$), E é o conjunto de arestas e d_{ij} é a distância entre os átomos i e j ($d_{ij} \in \mathbb{R}$). As distâncias d_{ij} são obtidas de um aparelho de ressonância magnética. No entanto, a máquina não é capaz de medir algumas distâncias; a saber, para quando os átomos estão a uma distância maior que 5 angstroms, ou (10^{-10} metros).

Para se organizar os dados do problema costuma-se utilizar uma área da matemática denominada de grafos, nesse caso grafos ponderados. Um grafo G é um par de conjuntos (V, E) , tal que V é o conjunto dos vértices e E é o conjunto formado de pares não ordenados de elementos de V , a qual chamamos de arestas. Por simplicidade de notação, uma aresta entre os vértices x e y será representada por xy .

Um grafo diz-se simples se não existem arestas paralelas (mais do que uma aresta entre os mesmos dois vértices) nem lacetes (arestas com ambos os extremos no mesmo vértice) (CARDOSO, 2011). Um grafo é ponderado se para cada aresta em E associamos um número real.

Os vértices V representam os átomos, as arestas E representam as distâncias conhecidas e a função $d : E \rightarrow \mathbb{R}$ armazena as distâncias dadas pela máquina de ressonância magnética.

	i	j	d _{ij}
1	1	2	1.526
2	1	3	2.49139
3	1	4	2.89393
4	1	6	5.1867
5	2	3	1.526
6	2	4	2.49139
7	2	5	3.83445
8	3	4	1.526
9	3	5	2.49139
10	3	6	2.92201
11	4	5	1.526
12	4	6	2.49139
13	4	7	3.8389
14	5	6	1.526
15	5	7	2.49139
16	6	7	1.526

Figura 4 – Exemplo de dados de entrada do problema: as distâncias entre cinco átomos. Note que não sabemos a distância entre alguns átomos, como por exemplo a distância do átomo 1 ao 5.

Fonte: Autoria própria.

Abaixo temos um exemplo onde podemos que observar que algumas distâncias são conhecidas, mas outras não conhecemos.

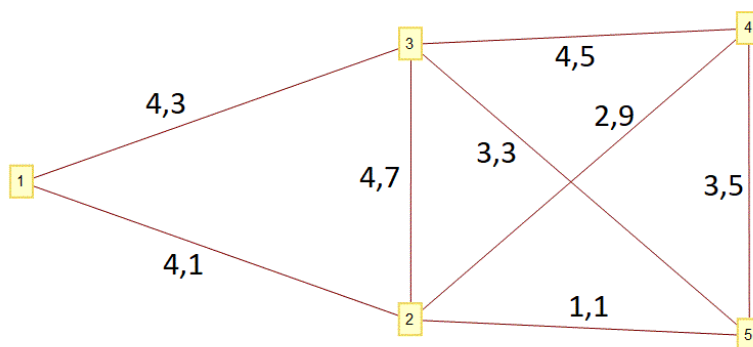


Figura 5 – Exemplo dos dados de entrada do problema. Este grafo representa que a distância entre o átomo 1 e 3 são conhecidas (4,3 angstroms), mas a distância do átomo 1 para o 4 não conhecemos.

3.2 Abordagem ingênua: método NR

A tentativa usual seria utilizar o método de NR para encontrar essas distâncias. Porém, dependendo das posições dos átomos, o método pode não convergir. Por exemplo,

considere o caso de uma molécula de 5 átomos, no qual a solução é;

x	y	z
0.732761	0.169653	0.656641
0.246863	0.660452	0.742898
0.426759	0.729058	0.897434
0.100975	0.67008	0.241258
0.204643	0.236955	0.655968

temos o sistema que associa as distâncias às coordenadas cartesianas,

$$\begin{aligned}
(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 &= 0.484421 \\
(x_{11} - x_{31})^2 + (x_{12} - x_{32})^2 + (x_{13} - x_{33})^2 &= 0.464552 \\
(x_{11} - x_{41})^2 + (x_{12} - x_{42})^2 + (x_{13} - x_{43})^2 &= 0.822124 \\
(x_{11} - x_{51})^2 + (x_{12} - x_{52})^2 + (x_{13} - x_{53})^2 &= 0.283439 \\
(x_{21} - x_{31})^2 + (x_{22} - x_{32})^2 + (x_{23} - x_{33})^2 &= 0.0609507 \\
(x_{21} - x_{41})^2 + (x_{22} - x_{42})^2 + (x_{23} - x_{43})^2 &= 0.273019 \\
(x_{21} - x_{51})^2 + (x_{22} - x_{52})^2 + (x_{23} - x_{53})^2 &= 0.188689 \\
(x_{31} - x_{41})^2 + (x_{32} - x_{42})^2 + (x_{33} - x_{43})^2 &= 0.540181 \\
(x_{31} - x_{51})^2 + (x_{32} - x_{52})^2 + (x_{33} - x_{53})^2 &= 0.349807 \\
(x_{41} - x_{51})^2 + (x_{42} - x_{52})^2 + (x_{43} - x_{53})^2 &= 0.370329 \\
x_{11} &= 0 \\
x_{12} &= 0 \\
x_{13} &= 0 \\
x_{21} &= 0.696003 \\
x_{22} &= 0 \\
x_{23} &= 0
\end{aligned}$$

;

Ao se aplicar o método de NR neste sistema, por meio do comando NSolve do Mathematica, o método não convergiu para a solução. Segundo Yang et al (2006), essa dificuldade se dá pelo fato da função 3.1 ser não convexa em torno da raiz. Dizemos que uma função é convexa se, dados x, y no domínio de f e $t \in [0, 1]$, temos que

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

(BERTSEKAS, 2006).

3.3 O método BP

O algoritmo de Branch and Prune ("galho e poda", em inglês) é uma estratégia de busca. Basicamente, o algoritmo observa a distância de três átomos até um quarto átomo e lista as possíveis posições para este quarto átomo.

É, portanto, necessário definir uma boa ordenação. Neste contexto, um grafo é bem ordenado se, para todo i , existirem as arestas $x_{i,i+3}$, $x_{i+1,i+3}$ e $x_{i+2,i+3}$ em E . Intuitivamente, dizemos que o grafo é ordenado se para cada vértice conhecemos as distâncias até seus três vértices anteriores.

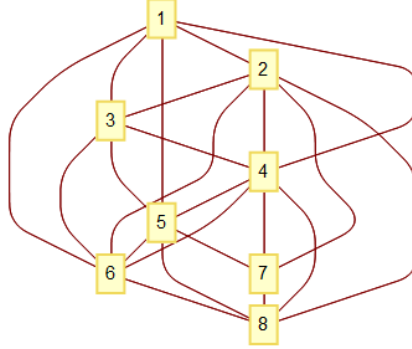


Figura 6 – Ordenação ruim. Veja como o vértice 4 não possui uma aresta ligando com o vértice 1

Seja $S(x_i, d_{ij})$ a esfera de centro x_i e raio d_{ij} . O algoritmo de branch and prune funciona da seguinte maneira:

Entradas:

- Um grafo ponderado $G(V, E, d)$.
- Uma ordenação boa do grafo.

1. Fixa-se as posições dos três primeiros átomos desta maneira:
 - a) o primeiro é colocado na origem do sistema, $(0,0,0)$.
 - b) O segundo é colocado em $(-d_{12}, 0, 0)$
 - c) o terceiro é colocado em uma das duas intersecções entre o plano xy e os círculos de raio d_{13} e raio d_{23} com centros x_1 e x_2 , respectivamente.
2. A posição do quarto átomo é dada fazendo-se a intersecção em $S(x_1, d_{14})$, $S(x_2, d_{24})$ e $S(x_3, d_{34})$. Seja D o conjunto dos pontos dessa intersecção. Essa intersecção pode ter um, dois ou nenhum ponto, ou seja, $|D| = 0, 1$ ou 2 .
3. Caso $|D| = 2$
 - a) Denominamos os elementos de D como x_4^0 e x_4^1 .
 - b) Supomos que a posição x_4^0 é válida. Repetimos então o processo a partir do passo 2 para encontrar a posição da molécula 5, mas utilizando as esferas $S(x_2, d_{25})$, $S(x_3, d_{35})$ e $S(x_4^0, d_{45})$.
 - c) Fazemos a intersecção deste novo D também com as esferas $S(x_i, d_{i5})$, caso d_{i5} exista. Desta maneira, $|D| = 0$ ou 1 .
 - i. Caso $|D| = 0$, concluímos que a escolha de x_4^0 não foi válida. Vamos "podar" este "galho".

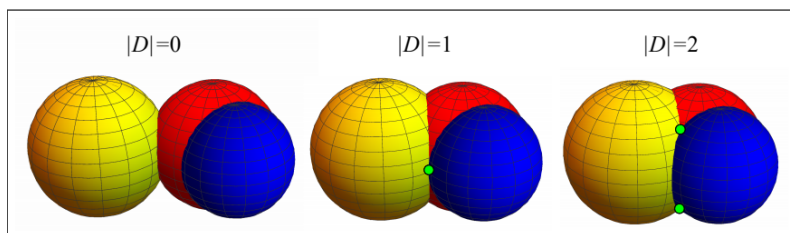


Figura 7 – Formas de intersecções possíveis (pontos verdes) ao tentar se obter as posições viáveis para a molécula 4

Fonte: LAVOR et al, 2017;

- ii. Caso $|D| = 1$, seguimos o algoritmo considerando a posição do átomo 4 como x_4^0 .
- d) Caso $|D| = 1$, a posição do átomo 4 é bem definida. Seguimos o algoritmo para encontrar a posição da molécula 5, utilizando as esferas $S(x_2, d_{25})$, $S(x_3, d_{35})$ e $S(x_4, d_{45})$.
- 4. Caso $|D| = 0$, temos o indício de que os dados foram colhidos com erro. Não é possível determinar a posição da molécula 4.

Como se pode observar, a complexidade da análise pode crescer exponencialmente.

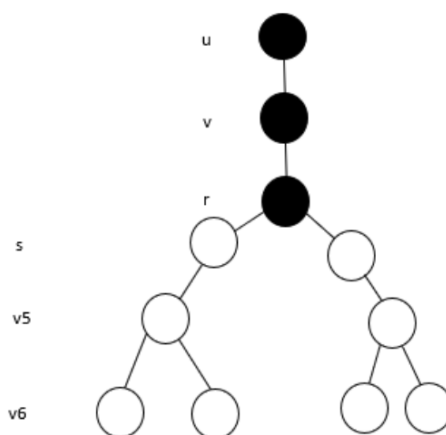


Figura 8 – Árvore de escolha das possíveis posições do primeiro átomo (u), segundo átomo (v), Observe que os três primeiros átomos têm posições fixadas.

Este algoritmo foi programado por Lavor em linguagem Wolfram. Implementamos este código no *Mathematica* e fizemos simulações.

3.4 Exemplo

Considerando as distâncias dadas na figura a seguir, o algoritmo encontra quatro soluções.

	i	j	dij
1	1	2	1.526
2	1	3	2.49139
3	1	4	2.91852
4	1	6	5.16605
5	2	3	1.526
6	2	4	2.49139
7	2	5	3.83955
8	3	4	1.526
9	3	5	2.49139
10	3	6	3.83463
11	4	5	1.526
12	4	6	2.49139
13	4	7	2.93228
14	5	6	1.526
15	5	7	2.49139
16	6	7	1.526

As soluções (posições dos átomos) são

$$\left\{ \begin{pmatrix} 0 & 0 & 0 \\ -1.526 & 0 & 0 \\ -2.03376 & 1.43905 & 0 \\ -1.52012 & 2.15871 & -1.24376 \\ -2.04705 & 3.59081 & -1.2557 \\ -1.38214 & 4.36706 & -2.38884 \\ -1.72462 & 3.71071 & -3.72323 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ -1.526 & 0 & 0 \\ -2.03376 & 1.43905 & 0 \\ -1.52012 & 2.15871 & -1.24376 \\ -2.04705 & 3.59081 & -1.2557 \\ -1.38214 & 4.36706 & -2.38884 \\ 0.126381 & 4.40478 & -2.16162 \end{pmatrix}, \right.$$

$$\left. \begin{pmatrix} 0 & 0 & 0 \\ -1.526 & 0 & 0 \\ -2.03376 & 1.43905 & 0 \\ -1.52012 & 2.15871 & 1.24376 \\ -2.04705 & 3.59081 & 1.2557 \\ -1.38214 & 4.36706 & 2.38884 \\ 0.126381 & 4.40478 & 2.16162 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ -1.526 & 0 & 0 \\ -2.03376 & 1.43905 & 0 \\ -1.52012 & 2.15871 & 1.24376 \\ -2.04705 & 3.59081 & 1.2557 \\ -1.38214 & 4.36706 & 2.38884 \\ -1.72462 & 3.71071 & 3.72323 \end{pmatrix} \right\}$$

4 CONCLUSÕES

Tendo em vista trabalhos revisados, pôde-se concluir que o algoritmo se mostra eficaz apresentando boas soluções, se mostrando mais eficaz que o método de Newton-Raphson que é o método usual para resolver inequações. No entanto, o algoritmo ainda está na categoria NP-difícil, uma vez que a complexidade da busca aumenta exponencialmente. Como estudos futuros, recomenda-se buscar formas de diminuir o espaço de busca. Talvez analisar outros dados moleculares, como ângulos de átomos ou possíveis simetrias. Percebeu-se ainda que o assunto é um tanto quanto novo em nosso país visto que apenas 0,004% dos artigos encontrados sobre o assunto são de escritores brasileiros.

Referências

- BERTSEKAS, Dimitri. *Convex Analysis and Optimization*. Athena Scientific, 2003.
 CARDOSO, Domingos Moreira. *Teoria dos grafos e aplicações*. 2011.
 LAVOR, C. et al. *Álgebra e Geometria no Cálculo de Estrutura Molecular*. 2017

LIBERTI, Leo et al. Euclidean distance geometry and applications. *Siam Review*, v. 56, n. 1, p. 3-69, 2014.

LIMA, Ana Maria Lemes; CAMARGO, Valter Soares. Cálculo Estrutural de Proteínas. In: II *Encontro Anual de Iniciação Científica da Unespar*. 2016.

SANT'ANNA, Carlos Mauricio R.. *Glossário de termos usados no planejamento de fármacos (recomendações da IUPAC para 1997)*. Quím. Nova, São Paulo , v. 25, n. 3, p. 505-512, May 2002 . Disponível em <<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422002000300027&lng=en&nrm=iso>>. acesso em 16 Jan. 2020. <<http://dx.doi.org/10.1590/S0100-40422002000300027>>.

WOLFRAM.S Celebrate Mathematica's First Quarter Century. 2013.

WEISSTEIN, ERIC W. *NP-Hard Problem*. From MathWorld—A Wolfram Web Resource.