Algoritmos para Dobramento do RNA

Luiz Carlos da Silva Rozante[1]*, **José Augusto Ramos Soares**[1]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo
Rua do Matão, 1010 – 055085-090 – Cidade Universitária – São Paulo, SP

rozante@ime.usp.br, jose@ime.usp.br

Abstract

Similarly to the proteins, the RNA molecules assume a three-dimensional conformation (structure) that has an important role in determining their function. These molecules can develop catalytic activities or be a structural or a regulator element. The experimental methods for deducing RNA structures are costly. The RNA secondary structure gives information about the function of the molecule and serves as an important step in determining its tertiary structure. So, it is important develop fast and accurate computer methods on prediction of secondary structure from primary structure. The two most significant strategies for solving the problem are based on thermodynamic stability criteria (of minimum free-energy) and in the search of the common foldings among homologous molecules. In the first case, the most important algorithms are based on techniques of dynamic programming. In the second, the most important algorithms known are based on models of covariance and perform on a set of aligned sequences. Being situated on genomics and biocomputing areas, this work presents the models proposed for the problem and describes formally the several existing techniques and methods involved in the solution of this problem. We also developed efficient implementations of the most expressive algorithms based on free energy minimization.

Keywords: RNA folding, RNA Secondary Structure, Dynamic Programming, Free Energy Minimization.

Resumo

De forma similar à que ocorre com as proteínas, as moléculas de RNA assumem uma conformação espacial (estrutura) que desempenha um importante papel na definição de sua função. Além de influenciar os processos de transcrição, tradução e replicação, as moléculas de RNA podem desenvolver atividades catalíticas. Podem ainda desempenhar papel regulador ou estrutural. Os métodos laboratoriais para determinação da estrutura do RNA são onerosos. A estrutura secundária do RNA, além de fornecer informações acerca da função da molécula, serve também como importante etapa na definição de sua estrutura terciária. Daí a importância em se desenvolver métodos computacionais, rápidos e precisos, de predição da estrutura secundária a partir da estrutura primária. As duas mais importantes estratégias de resolução do problema estão baseadas em critérios de estabilidade termodinâmica (de energia livre mínima) e na identificação dos dobramentos comuns entre moléculas homólogas. No primeiro caso, os algoritmos mais importantes são baseados em técnicas de programação dinâmica. No segundo, os mais importantes algoritmos conhecidos são baseados em modelos de covariância e operam sobre um conjunto de seqüências homólogas alinhadas. Situado no contexto da genômica estrutural e da bioinformática, o trabalho apresenta os modelos propostos para o problema, além de descrever formalmente as várias técnicas e métodos envolvidos na sua resolução. Desenvolvemos também implementações eficientes dos algoritmos mais expressivos baseados em cálculo de energia livre mínima, tanto do ponto de vista da complexidade computacional (de tempo e espaço), como da representatividade do modelo termodinâmico.

Keywords: Dobramento de RNA, Estrutura Secundária, Programação Dinâmica, Minimização de Energia Livre.

*Financiado pelo CNPq.

1. Introdução

Uma molécula de RNA consiste em uma cadeia de nucleotídeos conectados por ligações covalentes. Cada nucleotídeo contém um grupo fosfato, um açúcar (ribose) e uma base. Essa molécula de RNA é um polímero e é formado pela ligação de grupos fosfato. Somente as bases diferem e elas são quatro: Adenina (A), Citosina (C), Guanina (G) e Uracil (U).

Sob condições naturais, uma cadeia de RNA dobra-se sobre si mesma, através da formação de pontes de hidrogênio entre bases complementares (A com U e C com G) e entre bases *wobble* (U com G). As bases complementares formam pares de bases estáveis através da criação de pontes de hidrogênio entre elas, que são ditos pares de bases de Watson-Crick. Além disso, é possível considerar também (e geralmente o é) o par G-U, cuja ligação é mais fraca e que é denominado par de base *oscilante* ou *instável* (em inglês *wobble*). Os pares de bases de Watson-Crick, juntamente com os oscilantes, são denominados pares de bases *canônicos*.

A *estrutura secundária* de uma molécula de RNA é o conjunto de pares de bases canônicos — ou simplificadamente pares de bases — que ocorrem na "dobradura" natural da molécula.

1.1. Representação e Conceituação Matemática

De um modo mais formal, uma molécula de RNA é representada como uma seqüência de n caracteres $R = r_1, r_2, \ldots, r_n$, onde $r_i \in \{A, U, C, G\}$ representa o i-ésimo nucleotídeo. Uma *estrutura secundária* da molécula — cuja noção topológica está ilustrada na Figura 1 — é um conjunto S de pares de inteiros tal que cada par $(i, j) \in S$, com $1 \leq i < j \leq n$, satisfaz as seguintes restrições:

Restrição 1 r_i e r_j é um par de bases canônico;
Restrição 2 $j - i > t$, onde tipicamente $t = 4$ ou $t = 3$;
Restrição 3 se $i < i'$ e $(i', j') \in S$, então somente um dos casos ocorre:
Caso 2 $i < j < i' < j'$;
Caso 3 $i < i' < j' < j$.

Se $(i, j) \in S$ dizemos que r_i e r_j são bases *pareadas*. A Restrição 2 modela um fato da realidade biológica, observado experimentalmente, que consiste na impossibilidade de uma molécula dobrar-se sobre si mesma — em alguma parte — de forma pontiaguda. A Figura 2 ilustra um exemplo onde temos $t = 4$.

Os Casos 2 e 3 excluem uma configuração natural chamada *pseudo-nó*. Dizemos que ocorre um *pseudo-nó* quando existem pares (r_i, r_j), $(r_{i'}, r_{j'}) \in S$ com $i < i' < j < j'$. Sua exclusão simplifica o problema. Alguns tipos de pseudo-nós foram tratados no algoritmo de [18], cujas complexidades de tempo e espaço são, respectivamente, $O(n^6)$ e $O(n^4)$ para uma molécula com n bases. Entretanto, [14] mostraram recentemente que o problema geral é NP-difícil.

Poderíamos ser levados a conceber um algoritmo trivial que enumerasse todas as possíveis candidatas a estruturas e depois simplesmente escolhesse, entre aquelas que podem ser estruturas secundárias, aquela que correspondesse à estrutura mais estável (baseada num critério termodinâmico, por exemplo) . No entanto, o número possível de candidatas a estruturas é de pelo menos 2^n, para seqüências de n nucleotídeos. Isto, evidentemente, torna tal algoritmo inviável para seqüências de tamanho razoável.

2. Usando Cálculo de Energia Livre Mínima

A estratégia mais difundida para se predizer a estrutura secundária de uma molécula de RNA é baseada no cálculo da estrutura secundária de energia livre mínima. Tal estratégia utiliza-se da idéia de se atribuir uma energia a cada um de seus pares de bases, ou a seus elementos formadores estruturais, como laços internos, barrigas, arcos, hélices e multilaços [27].

2.1. Algoritmo Básico

Um modelo simplificado do problema, cuja idéia inicial foi proposta por Nussinov [17], supõe que as energias de cada um dos pares de bases são independentes entre si, de maneira que a energia total da estrutura S pode ser escrita como

$$Energia(S) = \sum_{(i,j) \in S} \alpha(r_i, r_j), \text{ onde } \alpha(r_i, r_j) < 0. \quad (1)$$

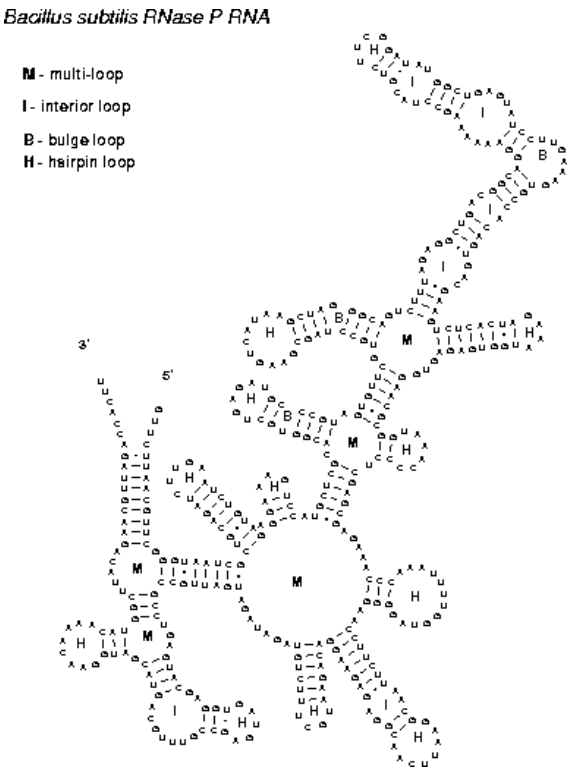


Figure 1: Exemplo de Estrutura Secundária do RNA na sua representação Normal. Obtida em <http://www.ibc.wustl.edu/∼zuker/Bio-5495/RNAfold-html>.

Ou seja, pressupõe-se existir uma função α tal que $\alpha(r_i, r_j)$ é definida como a energia de ligação do par de bases (r_i, r_j).

Esta idéia permite-nos modelar o problema como segue. Seja a seqüência $R = r_1, r_2, \ldots, r_n$ para a qual desejamos encontrar uma estrutura secundária S de energia livre mínima. Definimos

$$E(R) = \min_S \{Energia(S)\},$$

onde S varia em todas as estruturas secundárias de R.

Tomando a subseqüência $R_{i,j} = r_i, r_{i+1}, \ldots, r_j$, $1 \leq i < j \leq n$, para a qual desejamos encontrar a estrutura secundária $S_{i,j}$ correspondente de energia livre mínima, há quatro possibilidades a serem tratadas:

1. se r_i não é base pareada em nenhuma estrutura de energia mínima, então $E(R_{i,j}) = E(R_{i+1,j})$;
2. se r_j não é base pareada em nenhuma estrutura de energia mínima, então $E(R_{i,j}) = E(R_{i,j-1})$;
3. se em alguma estrutura de energia mínima, r_i e r_j são bases pareadas, mas não entre si, então $E(R_{i,j}) = \min_k\{E(R_{i,k}) + E(R_{k+1,j})\}$, para $i + t < k < j - t$;
4. se r_j é pareada com r_i em alguma estrutura de energia mínima, então $E(R_{i,j}) = E(R_{i+1,j-1}) + \alpha(r_i, r_j)$.

De modo mais formal, reescrevemos então as situações (modelo) acima como

$$E(R_{i,j}) = \left\{ \begin{array}{ll} 0, & se\ \ j - i \leq t \\ \min \left\{ \begin{array}{l} E(R_{i+1,j}), \\ E(R_{i,j-1}), \\ \min_{i+t<k<j-t}\{E(R_{i,k}) + E(R_{k+1,j})\}, \\ \alpha(r_i, r_j) + E(R_{i+1,j-1}) \end{array} \right\}, & \text{caso contrário.} \end{array} \right. \quad (2)$$

A Expressão 2 é resolvida por programação dinâmica. Resolvemos esta recorrência através de um algoritmo iterativo o qual preenche uma matriz de energias E, onde cada célula $E[i][j]$ armazena $E(R_{i,j})$, $1 \leq i < j \leq n$.

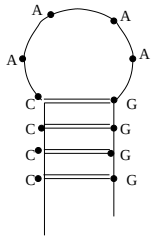


Figure 2: Um exemplo de "arco" formado com quatro ($t = 4$) bases do tipo 'A'.

Atribuímos $E[i][j] \leftarrow 0$ para valores iniciais $j - i \leq t$. Lembremos que o parâmetro t relaciona-se à impossibilidade da molécula dobrar-se, sobre si mesma, de forma demasiado pontiaguda (Figura 2 da Seção 1.1.). Este algoritmo é de complexidade $O(n^3)$.

Uma vez calculado $E[1][n]$, a computação (identificação) do dobramento de energia mínima $S_{1,n}$ é feita através de um algoritmo rastreador (do inglês *traceback*).

2.2. Incorporação de Laços

Infelizmente, a abordagem acima é insuficiente para capturar e representar algumas situações que concretamente ocorrem na definição das estruturas secundárias, pois não leva em consideração a influência que a energia de um par de bases exerce sobre outro, notadamente os pares adjacentes; tampouco contabiliza as energias associadas a estruturas denominadas laços, definidos a seguir e ilustrados na Figura 3. Esta formulação está baseada nos trabalhos de [27], [20], [23] e [28].

Seja $(i, j) \in S$ e sejam i',v e j' posições tais que $i < i' < v < j' < j$. Então dizemos que:

1. v é *acessível* a (i, j) se $(i', j') \notin S$ para todo i' e j';
2. (i', j') é *acessível* a (i, j) se $(i', j') \in S$ e i' e j' são acessíveis a (i, j);
3. o conjunto formado pelas bases dos pares de bases acessíveis a (i, j) e pelas bases não pareadas — também acessíveis a (i, j) — é o *laço fechado* por (i, j), ou simplesmente *laço*.
4. o laço formado por k pares de bases (o par de fechamento (i, j) juntamente com $(k - 1)$ pares de bases acessíveis a (i, j)) e por k' bases não pareadas é chamado *k-laço* (ou *k-ciclo*) de tamanho k' fechado por (i, j).
5. uma base não pareada não pertencente a nenhum laço é uma *base externa*; um par de bases pareadas não pertencente a nenhum laço é denominado *par externo*. A coleção formada pelas bases externas e pares externos é denominado *laço externo*.

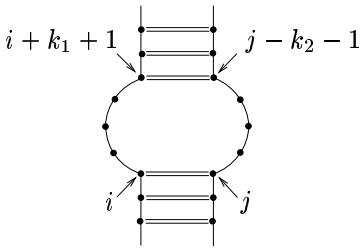
Uma estrutura secundária S induz uma decomposição de R em uma coleção de laços disjuntos $\mathit{Laço}_1, \mathit{Laço}_2, \ldots, \mathit{Laço}_m$, onde $m > 0$, se e somente se, $S \neq \emptyset$. Energias são atribuídas aos k-laços e a energia da estrutura S passa a ser escrita como

$$Energia(S) = \sum_{i=1}^{m} \varepsilon(\mathit{Laço}_i), \tag{3}$$

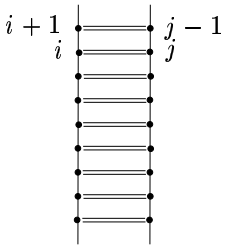
onde ε é uma função que fornece a energia de um k-laço $\mathit{Laço}_i$.

Para atribuir energias aos seis tipos de laços, são definidas as seguintes funções:

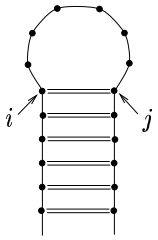
- $\varepsilon h(i, j)$ é a energia do laço arco fechado pelo par (i, j);
- $\varepsilon i(i, j)$ é a energia mínima de um laço interno fechado por (i, j);
- $\varepsilon bi(i, j)$ é a energia mínima de um laço barriga em i fechado por (i, j);
- $\varepsilon bj(i, j)$ é a energia mínima de um laço barriga em j fechado por (i, j);
- $\varepsilon s(i, j)$ é a energia de empilhamento de dois pares de bases adjacentes (i, j) e $(i + 1, j - 1)$;
- $\varepsilon m(i, j)$ é a energia mínima de um k-laço de tamanho k', com $k > 2$ (multi-laço), fechado por (i, j).



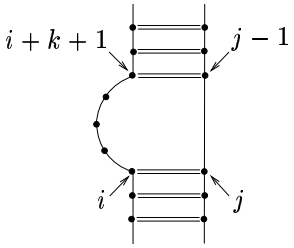
A: Laço Interno



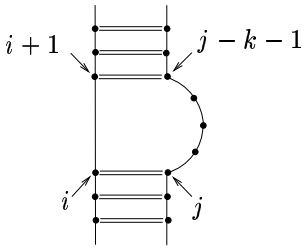
B: Laço Hélice (ou Empilhamento)



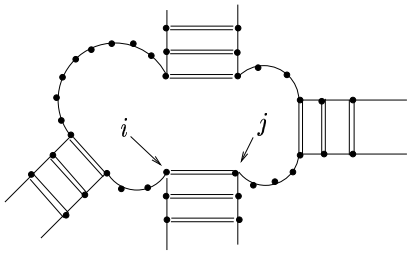
C: Laço Arco



D: Laço Barriga em i



E: Laço Barriga em j



F: Multi-laço (4-laço)

Figure 3: Vários Tipos de Laços. As linhas simples representam as ligações covalentes e as duplas as pontes de hidrogênio. Obtida de [21].

Novamente usamos a estratégia de programação dinâmica para resolver o problema. Seja a seqüência $R = r_1, r_2, \ldots, r_n$ para a qual desejamos encontrar a estrutura secundária $S_{1,n}$ de energia livre mínima. Consideremos a subseqüência $R_{i,j} = r_i, r_{i+1}, \ldots, r_j$, $1 \leq i < j \leq n$, para a qual desejamos encontrar a estrutura secundária $S_{i,j}$ de energia livre mínima. Como na modelagem anterior, há quatro possibilidades a serem tratadas, diferenciando-se apenas na situação 4:

$$E(R_{i,j}) = \min \left\{ \begin{array}{ll} 0, & se \;\; j - i \leq t \\ \min \left\{ \begin{array}{l} E(R_{i+1,j}), \\ E(R_{i,j-1}), \\ \min_{i<k<j}\{E(R_{i,k}) + E(R_{k+1,j})\}, \\ L(R_{i,j}) \end{array} \right\}, & \text{caso contrário,} \end{array} \right. \tag{4}$$

onde

$$L(R_{i,j}) = \min \{ \;\; \varepsilon h(i,j), \varepsilon i(i,j), \varepsilon bi(i,j), \varepsilon bj(i,j), \varepsilon s(i,j), \varepsilon m(i,j) \}. \tag{5}$$

Podemos então caracterizar os cálculos associados a $L(R_{i,j})$ para cada uma das configurações ilustradas na Figura 3, cuja opção correspondente pode ser vista na Expressão 5. Assim sendo,

- se $L_{i,j}$ é um laço arco, então

$$L[i][j] \leftarrow L(R_{i,j}) = \varepsilon h(i,j) = \zeta(j - i - 1).$$

- se $L_{i,j}$ é uma região empilhada (ou hélice), então

$$L[i][j] \leftarrow L(R_{i,j}) = \varepsilon s(i,j) = \eta + L[i+1][j-1].$$

- se $L_{i,j}$ é uma barriga em i, então

$$L[i][j] \leftarrow L(R_{i,j}) = \varepsilon bi(i,j) = \min_{k \geq 1}\{\beta(k) + L[i+k+1][j-1]\}.$$

- se $L_{i,j}$ é uma barriga em j, então

$$L[i][j] \leftarrow L(R_{i,j}) = \varepsilon bj(i,j) = \min_{k \geq 1}\{\beta(k) + L[i+1][j-k-1]\}.$$

- se $L_{i,j}$ é um laço interno, então

$$L[i][j] \leftarrow L(R_{i,j}) = \varepsilon i(i,j) = \min_{k_1,k_2 \geq 1}\{\gamma(k_1 + k_2) + L[i+1+k_1][j-1-k_2]\}.$$

- se $L_{i,j}$ é um multilaço, então

$$L[i][j] \leftarrow L(R_{i,j}) = \varepsilon m(i,j) = \min_{i<k<j-1}\{G[i+1][k] + G[k+1][j-1] + a\},$$

onde

$$G[i][j] = \min \left\{ \begin{array}{l} L[i][j] + b \\ \min_{i<h<j} \min \left\{ \begin{array}{l} G[i][h] + (j-h) \times c \\ G[i][h] + G[h+1][j] \\ (h-i+1) \times c + G[h+1][j], \end{array} \right. \end{array} \right. \tag{6}$$

sendo que a, b e c são constantes onde a representa a contribuição do par de fechamento (i,j) do multi-laço, b a contribuição de cada par acessível a (i,j) e c a contribuição de cada base não pareada acessível a (i,j).

As funções ζ, η, β e γ nas expressões acima são determinadas experimentalmente. A complexidade deste algoritmo é $O(n^4)$ e a computação (identificação) do dobramento de energia mínima $S_{1,n}$ é feita através de algoritmos recursivos.

2.3. Melhoria da Eficiência em Laços Internos

Os laços internos são as estruturas que dominam assintoticamente no algoritmo geral apresentado na seção anterior. A partir deste fato, esforços foram empreendidos a fim de melhorar a complexidade do algoritmo como um todo por meio da melhoria da complexidade dos laços internos.

Assumindo que a estabilidade de um laço interno depende apenas do seu tamanho, [23] mostraram como reduzir o tempo de computação do algoritmo geral para $O(n^3)$.

No entanto, o cálculo da energia associada ao laço interno fechado por (i, j) e (i', j') é determinado por quatro fatores:

- contribuição entrópica, que depende do tamanho do laço;
- contribuição referente ao par terminal não pareado adjacente a (i, j);
- contribuição referente ao par terminal não pareado adjacente a (i', j');
- penalidade associada a assimetria do laço.

Lyngs *et al* [15] propuseram um algoritmo que executa o algoritmo geral em tempo $O(n^3)$ — otimizando o cálculo de laços internos para $O(n)$ — mas com a vantagem de incorporar todos os fatores acima.

2.4. Geração de soluções sub-ótimas

Uma consideração importante que deve ser feita com relação às modelagens acima e aos respectivos algoritmos, é que elas nos fornecem uma única solução, que pode não ser necessariamente a estrutura verdadeira. É desejável, então, que se tenha um conjunto de soluções, onde algumas delas representem valores sub-ótimos, no que se refere à energia livre. Zuker [26] descreve um algoritmo que oferece soluções sub-ótimas.

Wuchty *et al* [25] apresentaram um algoritmo que gera todas as estruturas secundárias sub-ótimas dentro de um intervalo energético definido pela energia livre mínima da estrutura ótima e um limite superior arbitrário.

2.5. Melhoria da Eficiência para Classes Especiais de Funções

Com o intuito de melhorar o desempenho do algoritmo geral, alguns esforços foram empreendidos a partir de suposições acerca do comportamento das funções de desestabilização dos laços. Alguns algoritmos melhoram sensivelmente a complexidade de tempo supondo linearidade, convexidade ou concavidade destas funções. Infelizmente estas suposições não refletem de maneira confiável a realidade biológica; ou seja, estas funções na prática não são lineares, convexas ou côncavas. De qualquer modo, estes algoritmos representam avanços no campo estrito da computação.

Caso façamos a suposição de que as funções de desestabilização de laços sejam lineares no tamanho do laço, então é possível reduzir o tempo de computação de barrigas e laços interiores para uma constante. Isto leva a um algoritmo geral de complexidade de tempo $O(n^2)$ [23], caso não consideremos os multilaços.

Eppstein *et al* [4] propuseram um algoritmo com tempo de execução $O(n^2 \log^2 n)$. Neste algoritmo faz-se a suposição de que não ocorrem multi-laços na estrutura, bem como a função que fornece o custo (contribuição energética) de um laço é convexa.

Lamore e Schieber [10] melhoraram o tempo de execução para $O(n^2)$ no caso de funções côncavas. Apresentaram ainda um algoritmo cujo crescimento assintótico é expresso por $O(n^2\alpha(n))$ no caso de funções convexas, onde $\alpha(n)$ é uma função com crescimento extremamente lento (inversa da função de Ackermann).

3. Usando Análise Comparativa entre Homólogos

Uma outra forma de calcular a estrutura secundária de moléculas de RNAs é através da comparação de um dado conjunto de seqüências homólogas, entre as quais supõe-se existir um grau de similaridade que permita que as mesmas sejam alinhadas de modo confiável. Duas ou mais seqüências de RNA's são ditas *homólogas* se derivam de uma mesma seqüência; ou seja, se têm — do ponto de vista da história evolutiva — um ancestral comum do qual descendem. Esta estratégia está baseada no princípio da biologia molecular que diz que, no processo evolucionário, a estrutura tende a ser mais conservada do que a seqüência de nucleotídeos correspondente.

Ou seja, como o que realmente determina a função de uma molécula é sua estrutura, supõe-se que na eventualidade de uma alteração em uma dada base da seqüência que implique na modificação de sua estrutura (devido, por exemplo, a fatores mutacionais), esta vai ser acompanhada de uma outra alteração que neutralize aquela, de modo que a estrutura e função são conservadas. Este tipo de alteração é chamada de *alteração compensatória*.

Os métodos que se utilizam deste princípio geralmente fazem uso de técnicas de análise de covariância para inferir a estrutura da molécula. Eles também geralmente envolvem, de alguma forma, algoritmos de alinhamento múltiplo, pois ou recebem um alinhamento das seqüências homólogas como entrada ou produzem um alinhamento para estas seqüências. Alguns métodos utilizam programas de alinhamento prontos, como por exemplo o CLUSTAL W [22, 8].

Vários métodos foram formulados baseados nestas idéias. Existem também esforços no sentido de incorporar em um único método características oriundas do cálculo de energia livre mínima, bem como da análise comparativa entre homólogos. Há ainda métodos que não se utilizam diretamente das técnicas de análise de covariância, mas exploram o fato de disporem de um conjunto de seqüências homólogas e as propriedades que daí advém.

Han e Kim [7] propuseram uma heurística baseada em técnicas de análise de covariância com complexidade $O(mn^2 + n^3)$, onde m é o número de seqüências homólogas e n o número de bases em cada uma das moléculas alinhadas. Este método é descrito adiante.

Winker *et al* [24] também propuseram um algoritmo que se utiliza de técnicas de covariância para determinação de estrutura secundária a partir de um alinhamento de seqüências homólogas.

Lück *et al* [11] elaboraram um método que combina elementos de análise comparativa com distribuição de probabilidade de estruturas. Eles se utilizam do algoritmo para cálculo da função Partição Equilíbrio e de probabilidade de pares de [16].

Alguns métodos recentes baseados em gramáticas estocásticas independentes de contexto, que utilizam a dependência estatística entre as colunas de um alinhamento de homólogos, também foram propostos para identificar o dobramento comum entre as seqüências: [12], [13], [3], [19] e [5].

3.1. Análise de Covariância

Como já dissemos, os métodos que se utilizam do princípio enunciado acima, geralmente, fazem uso de técnicas de análise de covariância para inferir a estrutura da molécula. A fim de melhor descrever este tipo de técnica, suponhamos um alinhamento múltiplo de um conjunto de seqüências homólogas $R^1, R^2, \ldots, R^m$ todas de tamanho n, como ilustrado abaixo:

$$\begin{array}{rccccc} R^1 = & r_1^1 & r_2^1 & r_3^1 & \cdots & r_n^1 \\ R^2 = & r_1^2 & r_2^2 & r_3^2 & \cdots & r_n^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ R^m = & r_1^m & r_2^m & r_3^m & \cdots & r_n^m \end{array}$$

Uma medida quantitativa da interdependência entre pares de colunas i e j — denotadas por $M_{i,j}$ — que vem da teoria da informação [1, 6, 2], é denominada *índice de informação mútua*.

Para melhor entender esta medida, façamos antes algumas definições e considerações:

Uma *entropia* é uma medida da incerteza de um resultado. Dado uma variável aleatória X com probabilidades $P(x_i)$ para o conjunto discreto de k eventos $x_1, x_2, \cdots, x_k$, a entropia de Shannon é definida por

$$H(X) = -\sum_{i=1}^{k} P(x_i) \log_2 P(x_i).$$

Esta entropia é maximizada quando todos os termos da soma são iguais, isto é, $P(x_i) = 1/k$. Deste modo, temos

$$H(X) = -\sum_{i=1}^{k} P(x_i) \log_2 P(x_i) = -\sum_{i=1}^{k} \frac{1}{k} \log_2 \frac{1}{k} = log_2 k.$$

Se estamos certos do resultado de um evento x_i, isto é, se temos que $P(x_i) = 1$ ou $P(x_i) = 0$, então a entropia é zero.

Índice de informação, aqui denotada por I, é uma medida da redução da incerteza. Por exemplo, no caso de nossa variável aleatória X, se a você é informado o resultado de um evento x_i, a incerteza é reduzida de $H(X)$ para zero, devido ao fato de você ter obtido esta informação. De um modo mais geral, índice de informação é a diferença das entropias antes e depois da "informação":

$$I(X) = H_{antes}(X) - H_{depois}(X).$$

Índice de informação pode ser usado para mensurar o grau de conservação de bases em um alinhamento de RNAs. Por exemplo, suponhamos que a expectativa para uma base de RNA é aleatória, ou seja, $P(r) = 1/4$ e $H_{antes} = 2$, onde $r \in \{A, C, G, U\}$. Suponhamos também que observamos, para uma posição particular, sempre um 'A' ou um 'G' com $P(A) = 0.7$ e $P(G) = 0.3$. Desta forma, temos $H_{depois} = -0.7 \log_2 0.7 - 0.3 \log_2 0.3 = 0.88$. Portanto, o índice de informação para esta posição é $2 - 0.88 = 1.12$. Entretanto, notemos que o índice de informação pode assumir valor negativo se a distribuição observada tem uma entropia maior do que a esperada.

Por conta disto, é então melhor mensurar a diferença entre duas distribuições (frequentemente nos referimos a entropia da distribuição de probabilidade P, como $H(P)$) por meio da entropia relativa. Para duas distribuições P e Q, a *entropia relativa* é definida como

$$H(P \parallel Q) = \sum_{i=1}^{k} P(x_i) \ln \frac{P(x_i)}{Q(x_i)}.$$

Entropia relativa apresenta a propriedade de ser sempre maior ou igual a zero, onde a igualdade é verificada se $P(x_i) = Q(x_i)$.

É útil encarar a entropia relativa $H(P \parallel Q)$ como a "distância" entre as distribuições de probabilidade P e Q. Entretanto, devemos observar que elas não são simétricas, ou seja, $H(P \parallel Q) \neq H(Q \parallel P)$.

Duas variáveis aleatórias X e Y são independentes se $P(X, Y) = P(X)P(Y)$. Nos interessa saber quão independentes elas são. Isto pode ser mensurado por meio da entropia relativa entre as distribuições $P(X, Y)$ e $P(X)P(Y)$, denominada *informação mútua* e aqui denotada por M. Assim,

$$M(X, Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)P(y_j)},$$

onde $M(X, Y)$ é uma medida da interdependência entre X e Y e os valores possíveis de X e Y são $\{x_i\}$ e $\{y_j\}$.

Em teoria da informação, é frequente se assumir que a distribuição de probabilidade é conhecida. No entanto, em muitas aplicações não se conhece a distribuição verdadeira. Então, as entropias são calculadas através das freqüências relativas dos eventos ao invés da distribuição de probabilidade. Para uma coluna i, seja $f_i(r)$ a freqüência relativa com que a base r ocorre na i-ésima coluna, onde $r \in \{A, C, G, U\}$. Vamos considerar uma outra coluna j e tomar $f_{i,j}(r, s)$ como sendo a freqüência relativa conjunta de ocorrência dos dois nucleotídeos r e s nas colunas i e j. Em outras palavras, seja $f_{i,j}(r, s)$ a freqüência relativa com que ocorre um dos dezesseis possíveis pares de bases nas colunas i e j.

Na hipótese onde quaisquer duas colunas são independentes, devemos esperar que $f_{i,j}(r, s)$ seja aproximadamente igual a $f_i(r) f_j(s)$. Desta forma, esperamos

$$\log \frac{f_{i,j}(r, s)}{f_i(r) f_j(s)} \approx 0.$$

Isto leva à definição do *índice de informação mútua* entre duas diferentes colunas i e j, denotado por $M_{i,j}$, como

$$M_{i,j} = \sum_{r,s \in \{A,C,G,U\}} f_{i,j}(r, s) \log_2 \frac{f_{i,j}(r, s)}{f_i(r) f_j(s)}. \tag{7}$$

Utilizamos o logaritmo na base 2, de tal modo que o resultado de $M_{i,j}$ — para as quatro letras do alfabeto do RNA — varia entre 0 e 2. Quando i e j covariam perfeitamente, devemos esperar que

$$f_{i,j}(r,s) = f_i(r) = f_j(s)$$

para todos os pares de nucleotídeos. Esta quantidade deve ser 0 quando r e s não formam par.

Desta forma, esperamos que $M_{i,j}$ seja máximo (isto é, $M_{i,j} = 2$) quando i e j correlacionam perfeitamente e mínimo (isto é, $M_{i,j} = 0$) quando i e j não estão correlacionados. Ou seja, $M_{i,j}$ é máximo se i e j, individualmente, aparecem completamente aleatórios ($f_i = f_j = 0.25$) com i e j perfeitamente correlacionados (formando par).

4. Implementação e Conclusões

Desenvolvemos implementações eficientes dos algoritmos mais expressivos baseados em cálculo de energia livre mínima, tanto do ponto de vista da complexidade computacional (de tempo e espaço), como da representatividade do modelo termodinâmico: [17], [28], [23] e [15].

Estas implementações tiveram como propósito possibilitar o entendimento, preciso e detalhado, destes métodos e modelos. O código C gerado, os arquivos de dados e demais arquivos necessários à execução dos programas estão disponíveis e podem ser acessados no endereço $http://www.ime.usp.br/dcc/posgrad/teses/rozante/$.

Neste endereço encontra-se também disponível um arquivo no formato dvi/ps contendo um documento com mais detalhes a respeito dos métodos, modelos e implementação em questão. Este documento foi construído sob a filosofia *Literate Programming* (ambiente CWEB [9]), de modo que inserimos, de forma "diluída", código C ao longo do texto principal. Ou seja, à medida que os conceitos e métodos são apresentados, incluímos, no ponto da apresentação, o código correspondente ao conceito ou método.

A precisão dos resultados fornecidos pelos algoritmos baseados em minimização de energia tendem a melhorar na medida em que os parâmetros termodinâmicos se tornam mais precisos. Neste contexto, os parâmetros de energia tendem a incorporar cada vez mais casos especiais que surgem à medida em que avança o conhecimento sobre as propriedades físico-químicas dos ácidos nucléicos.

Em relação às soluções fornecidas pelos algoritmos baseados em cálculo de energia livre mínima, podemos dizer que essas soluções podem não descrever adequadamente a situação real. Em outras palavras, o modelo adotado para descrever as interações termodinâmicas pode não capturar a totalidade das situações que efetivamente ocorrem na natureza. Isto ocorre por dois motivos.

Primeiro, os parâmetros de energia com os quais os algoritmos trabalham são inevitavelmente imprecisos. Logo, a estrutura de energia livre mínima pode ser sub-ótima em relação aos parâmetros usados. O mesmo pode ocorrer em função do não conhecimento (ou não tratamento) de alguma restrição biológica que pode alterar as energias relativas, tornando/levando a uma outra estrutura sub-ótima entre as mais favoráveis. Estes fatos justificam o desenvolvimento de algoritmos que forneçam várias soluções.

Os algoritmos de Eppstein e Larmore representam importantes avanços no campo estrito da computação para o problema. No entanto, eles não representam um avanço importante do ponto de vista da contribuição biológica, pois na natureza as funções de desestabilização de laços não são convexas nem côncavas. Além disto a suposição de não existência de multilaços não é razoável.

Como a suposição de linearidade é ainda mais restritiva do que a convexidade e a concavidade, não acreditamos que algoritmos baseados nesta suposição representem contribuições significativas para o problema.

Talvez o leitor espere por considerações que apontem para uma relação superioridade/inferioridade entre as duas estratégias estudadas aqui. Concluímos não ser possível fazer afirmações categóricas neste sentido, tendo em vista a grande diferença de princípios que caracterizam uma e outra estratégia. No entanto, é possível estabelecer vantagens/desvantagens entre elas.

Algoritmos baseados em minimização energética conseguem operar sobre uma única seqüência e, em relação à qual, não é necessário conhecer qualquer informação filogenética. Isto representa uma vantagem desta estratégia em

relação àquela baseada em análise comparativa, já esta última exige que se disponha de um conjunto de moléculas homólogas como entrada, o que, nem sempre é possível.

No entanto, parece razoável supor que, com o tempo, o acúmulo de informação nas bases de dados de bioseqüências pode levar a um estado onde raramente, para um dada seqüência, não se disponha de um conjunto de homólogos.

Os métodos baseados em análise comparativa conseguem detectar interações terciárias (pseudo-nós) na estrutura. Isto representa uma vantagem desta estratégia em relação àquela baseada no cálculo de energia livre mínima. Esses métodos, geralmente, são insensíveis a pequenas variações na seqüência de nucleotídeos, enquanto nos métodos baseados em cálculo de energia livre mínima, pode-se chegar a estruturas muito diferentes a partir de seqüências que variam em poucas bases.

References

[1] D. K. Y. Chiu and T. Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *Computer Applications in the Biosciences*, (7):347–352, 1991.

[2] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis / Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[3] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.

[4] D. Eppstein, Z. Galil, and R. Giancarlo. Speeding up dynamic programming. In *28th Symposium on the Foundations of Computer Science*, pages 488–495, 1988.

[5] L. Grate. Automatic RNA secondary structure determination with stochastic context-free grammars. In *Intelligent Systems for Molecular Biology, 136–144*, 1995.

[6] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and applications of comparative sequence analysis methods. *Nucleic Acids Research*, (20):5785–5795, 1992.

[7] K. Han and H. J. Kim. Prediction of common folding structures of homologous RNAs. *Nucleic Acids Res.*, (21):1251–1257, 1993.

[8] D. G. Higgins and P. M. Sharp. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, (73):237–244, 1988.

[9] D. E. Knuth and S. Levy. *The CWEB System of Structured Documentation*. Reading, Massachusetts: Addison-Wesley, 1993.

[10] L. Lamore and B. Schieber. On-line dynamic programming with applications to the prediction of RNA secondary structure. In *First ACM-SIAM Symposium on Discrete Algorithms*, pages 503–512, 1990.

[11] R. Lück, G. Steger, and D. Riesner. Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *J. Mol. Biol.*, (258):813–826, 1996.

[12] F. Lefebvre. An optimized parsing algorithm well-suited to RNA folding. In *Proc. of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 222–230. AAAI Press, 1995.

[13] F. Lefebvre. A grammar-based unification of several alignment and folding algorithms. In *Proc. of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 143–154. AAAI Press, 1996.

[14] R. B. Lyngs and C. N. S. Pedersen. Pseudoknots in RNA secondary structure. In *Proc. 4rd Int. Conf. Computational Molecular Biology (RECOMB'00)*. ACM, Apr 2000.

[15] R. B. Lyngs, M. Zuker, and C. N. S. Pedersen. Internal loops in RNA secondary structure prediction. In *Proc. 3rd Int. Conf. Computational Molecular Biology (RECOMB'99)*. ACM, Apr 1999.

[16] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, (29):1105, 1990.

[17] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algoritms for loop matchings. *SIAM J. appl. Math.*, (35):68–82, 1978.

[18] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, (285):2053–2068, 1999.

[19] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, (22):5112–5120, 1994.

[20] D. Sankoff. Simultaneous solution of the mRNA folding, alignment and protosequense problems. *SIAM J. Appl. Math.*, (5):1–35, 1985.

[21] J. C. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. IC-UNICAMP/PWS, 1997.

[22] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

[23] M. S. Waterman and T. F. Smith. Rapid dynamic programming algorithms for RNA secondary structure. *Advances in Applied Mathematics*, (7):455–464, 1986.

[24] S. Winker, R. Overbeek, C. R. Woese, G. J. Olsen, and N. Pfluger. Structure detection through automated covariance search. *Comput. Appl. Biosci.*, (6):365–371, 1990.

[25] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, (49):145–165, 1999.

[26] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, (244):48–52, 1989.

[27] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, (46):591–621, 1984.

[28] M. Zuker and C. D. H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. (333):333–344, 1999.