Aprimoramento de uma ferramenta de predição de ncRNAs baseada em estruturas secundárias

Felipe Lamberti Amado, orientado por Prof. Dr. Alan Mitchell Durham

Instituto de Matemática e estatística da USP, São Paulo

IME-USP

Introdução

Algumas sequências de DNA são transcritas para moléculas de RNA, mas que, por sua vez, não são traduzidas para proteínas. Tais moléculas são chamadas de RNA não codificantes (ncRNA) e são importantes, principalmente, na regulação de genes expressos.

Um microRNA é um tipo específico de ncRNA curto, com cerca de apenas 100 nucleotídeos, e com estrutura secundária bem característica. Estudos mostraram que os microRNAs podem estar relacionados com diversas doenças, entre elas o câncer de mama[6], e o primeiro passo para compreendermos sua atuação no organismo é identificá-los facilmente e de maneira confiável.

O ex-aluno de mestrado do IME-USP, Vitor Onuchi, juntamente com os professores doutores Alan Durham e Ariane Machado, desenvolveu um preditor de microRNAs que se baseia em estrutura secundária, isto é, o arranjo espacial da molécula. A técnica consiste em encontrar subsestruturas que se assemelhem a subestruturas típicas de microRNAs, reduzindo o problema a um problema conhecido da Computação. Trata-se do problema em encontrar um subgrafo isomorfo em um grafo maior, o qual foi provado ser NP-completo. No entanto, testes mostraram que, para as instâncias desse problema em especial, o algoritmo tipicamente leva tempo menor que o proporcional a uma função exponencial no comprimento da sequência.

Este trabalho consistiu na realização de comparativos justos entre uma das soluções já existentes para predição de microRNAs, no aprimoramento da ferramenta de predição — encontrando melhores parâmetros através de curvas ROC — e, por fim, no desenvolvimento de uma extensão dela, com o intuito de automatizar o processo de encontrar substruturas típicas.

Métodos

As sequências de microRNAs utilizadas nos experimentos foram obtidas da base de dados aberta do Rfam.[4] Nesta base, as sequências são separadas em famílias, que são definidas pelo grau de similaridade de suas estruturas secundárias. Cada família possui um modelo de covariância, que é gerado a partir de sequências verificadas e bem anotadas. Selecionamos apenas estas sequências para os experimentos.

A representação da estrutura secundária de uma molécula em forma de grafo direcionado foi feita da seguinte maneira: a sequência é alinhada a sua reversa utilizando o programa RNAfold, cada subsequência alinhada com *score* maior que um determinado limiar é representada por uma aresta ligando as duas regiões e estas são representadas por nós. Além disso, foram inseridas arestas, chamadas arestas de ordem, para cada par de nós (i,j) tal que o nó i representa uma região mais à esquerda na sequência. Cada nó e aresta foram rotulados com as posições inicial e final em que ocorrem na sequência.

Deste modo, a identificação de um microRNA consiste na verificação de existência de um subgrafo isomorfo induzido a um dos padrões típicos no grafo da sequência alvo, obedecendo algumas restrições nos comprimentos de hélices e de arcos. Para isso, utilizamos a biblioteca VFlib2 com algumas modificações.

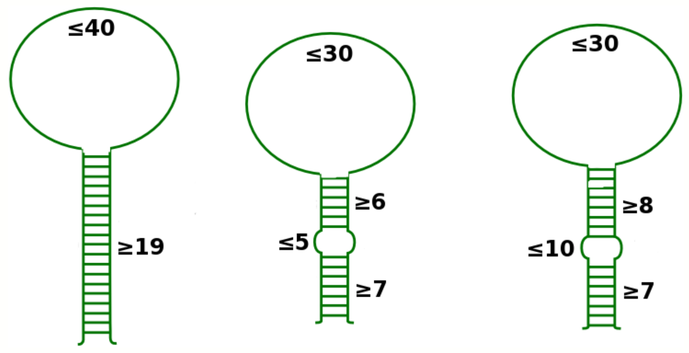


Figura 1: Subsestruturas típicas usadas inicialmente pelo preditor. Os números próximos às áreas alinhadas são os comprimentos mínimos de hélice enquanto os outros números são comprimentos máximos de arco.

Os seguintes parâmetros do preditor foram otimizados através de curvas ROC: limiar de probabilidade do alinhamento dado pelo RNAfold, comprimento de hélice mínimo para um alinhamento, comprimentos de hélices e arcos dos padrões utilizados.

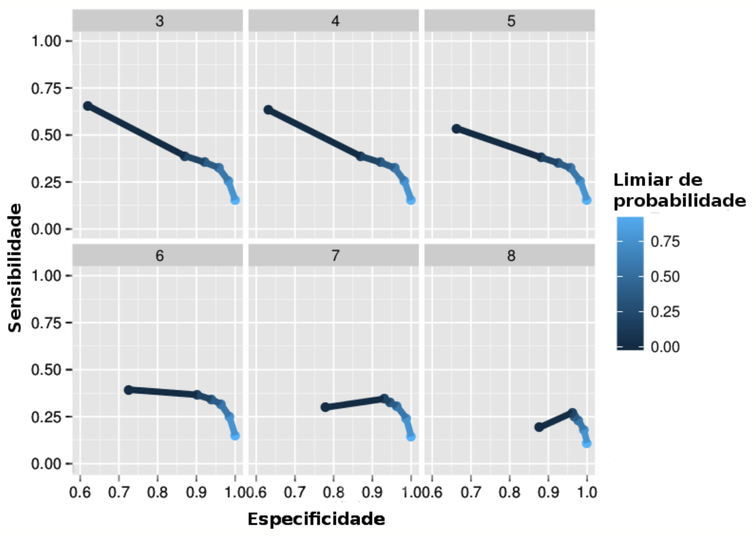


Figura 2: Curvas ROC ao variar o limiar de probabilidade (cor) e o comprimento mínimo de hélice.

Para encontrar os padrões (i.e. subestruturas típicas) nos grafos das sequências, utilizamos uma solução para o problema em encontrar o subgrafo isomorfo comum máximo a dois grafos, através de uma redução para o problema do clique máximo como descrita por Koch[3]. Algumas modificações ao algoritmo foram necessárias para a introdução de restrições de compatibilidade que levassem em conta os comprimentos dos nós e das arestas.

Resultados e discussão

Comparação com MiPred

Comparamos nosso preditor com um dos melhores preditores de microRNAs segundo Hu [1], chamado miPred[5]. Para compará-los, foi escolhida uma amostra aleatória de 1300 sequências de microRNAs, entre aquelas mais bem anotadas do Rfam, para extrair a medida de sensibilidade e utilizamos o conjunto dessas mesmas sequências porém embaralhadas para extrair a medida de especificidade.

Tabela 1: Comparação entre os preditores.

Medida	MiPred	Nosso preditor
Especificidade	88,076%	84,154%
Sensibilidade	80,615%	82,692%
Acurácia	84,346%	83,423%

Otimização dos parâmetros

Com os novos parâmetros, utilizamos apenas dois padrões em vez de três. O que, intuitivamente, faz com que o preditor fique menos específico e mais generalista. Apesar de perdermos cerca de 3% em especificidade, ganhamos 11% em sensibilidade, resultando em um aumento de quase 4% em acurácia.

Tabela 2: Antes e depois da otimização dos parâmetros.

Medida	Antes	Depois
Especificidade	87,538%	84,154%
Sensibilidade	71,692%	82,692%
Acurácia	79,615%	83,423%

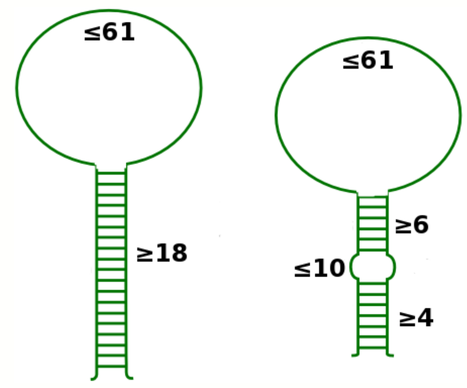


Figura 3: Subsestruturas típicas otimizadas.

Treinamento do preditor

Em um dos testes realizados, duas famílias A e B foram selecionadas. O treinamento foi realizado em 21 sequências da família A, encontrando 4 subestruturas recorrentes.

Executando o preditor em 106 sequências dessa mesma família A, o programa foi capaz de identificar 57 delas, ou seja, uma sensibilidade de aproximadamente 53,7%.

Executando o preditor com as mesmas subestruturas anteriores, mas dessa vez, em 438 sequências da familia B, apenas 12 foram identificadas, i.e. uma porcentagem de falso positivo de cerca de 2,7%.

Entretanto, houve uma alta variação dessas medidas nos testes realizados. Acredita-se que esta variação esteja relacionada com a variabilidade dos tamanhos médios das sequências de cada família, fazendo com que seja necessária uma nova configuração dos limiares para o critério de compatibilidade entre nós e entre arestas, e também com a dependência, que o algoritmo de treinamento possui atualmente, da ordem e do tamanho das sequências no conjunto de treinamento.

Conclusões

Com a otimização dos parâmetros do preditor, conseguimos uma discreta melhora, mas suficiente para alcançarmos níveis de acurácia comparáveis aos dos melhores preditores de microRNA atuais. Apesar de o algoritmo de treinamento ter mostrado medidas altas de sensibilidade e especificidade em alguns casos, ainda precisa de melhorias para que não seja necessária a configuração de novos parâmetros para cada nova família.

Referências

[1] Hu LL et al (2012). *Benchmark comparison of ab initio microRNA identification methods and software*. GMR.

[2] A Machado-Lima, HA del Portillo, AM Durham (2008). *Computational methods in noncoding RNA research.*. J Math Biol. 56(1-2):15-49

[3] Ina Koch (2001). *Enumerating all connected maximal common subgraphs in two graphs.* Elsevier.

[4] Sarah W. Burge et al (2013). *Rfam 11.0: 10 years of RNA families*. Nucleic Acids Res.

[5] Jiang P et al (2007). *MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features*. Nucleic Acids Res.

[6] Marilena V. Iorio et al (2005). *MicroRNA Gene Expression Deregulation in Human Breast Cancer*. AACR.