https://eventos.utfpr.edu.br//sicite/sicite2019

Estudo de Modelo Probabilístico para Sequências: Análise com sítios de splicing de lncRNA

Study of Probabilistic Model for Sequences: Analysis with lncRNA splicing sites

Aline Mara Rudsit Bini
alinebini@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

André Yoshiaki Kashiwabara
kashiwabara@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

RESUMO

Os RNA's são classificados em dois tipos: (i) codificadores de proteínas; (ii) não codificadores de proteínas. Os RNA mensageiros (RNAm) codificam para proteínas e atualmente foi identificado uma nova classe chamada de circular RNA (circRNA) que podem possuir potencial codante. Os RNA não-codificadores de proteínas são separados em muitas outras famílias. Em particular, os *long non-coding* RNA (lncRNA), por ter diversas funções, são amplamente estudados. Responsáveis por diversos aspectos de organismos complexos, os lncRNA ainda possuem restrições de anotações em relação aos RNAs codificantes. Isto acontece por serem fracamente amostrados, terem a associação sequência-função ruim e serem fracamente conservados durante a evolução. O trabalho estudou as anotações no genoma de A. thaliana dos lncRNA para verificar a hipótese de que os sítios de splicing de mRNA e de lncRNA são distinguíveis ou não. Foi utilizado um modelo probabilístico *Weight Array Model* (WAM) para representar os sítios de splicing de genes de lncRNA e os sítios de splicing de genes de mRNA. Utilizando *5-fold cross-validation* e com um classificador Bayesiano, o resultado sugere que há uma pouca diferença entre as duas classes.

PALAVRAS-CHAVE: Ácido ribonucleico. Modelos probabilísticos. Sítios de splicing, lncRNA, mRNA.

Recebido: 19 ago. 2019.
Aprovado: 01 out. 2019.

ABSTRACT

RNAS ARE CLASSIFIED INTO TWO TYPES: (I) PROTEIN-CODING; (II) NON PROTEIN-CODING. MESSENGER RNAS (MRNAS) ENCODE PROTEINS AND A NEW CLASS CALLED CIRCULAR RNA (CIRCRNA) HAS BEEN RECENTLY IDENTIFIED THAT CAN ALSO CODE TO PROTEIN. NON-PROTEIN CODING RNAS ARE SEPARATED INTO MANY OTHER FAMILIES. IN PARTICULAR, LNCRNA (LONG NON-CODING RNA), HAVING SEVERAL FUNCTIONS, IS WIDELY STUDIED. RESPONSIBLE FOR MANY ASPECTS OF COMPLEX ORGANISMS, LNCRNA STILL HAVE ANNOTATION RESTRICTIONS ON CODING RNAS. THIS IS BECAUSE THEY ARE POORLY SAMPLED, HAVE POOR SEQUENCE-FUNCTION ASSOCIATION, AND ARE POORLY CONSERVED DURING EVOLUTION. OUR WORK STUDIED THE ANNOTATIONS IN THE A. THALIANA GENOME OF LNCRNA TO VERIFY THE HYPOTHESIS THAT MRNA AND LNCRNA SPLICING SITES ARE DISTINGUISHABLE OR NOT. A WEIGHT ARRAY MODEL (WAM) WAS USED TO REPRESENT THE LNCRNA GENE SPLICING SITES AND MRNA GENE SPLICING SITES. USING 5-FOLD CROSS-VALIDATION AND A BAYESIAN CLASSIFIER, THE RESULT SUGGESTS THAT THERE IS LITTLE DIFFERENCE BETWEEN THE TWO CLASSES OF SPLICING SITES.

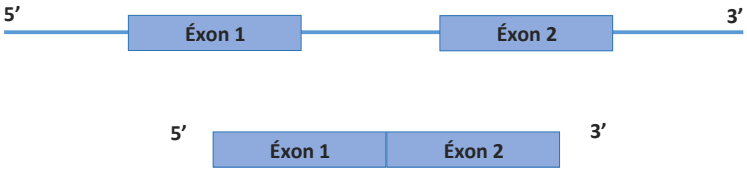
KEYWORDS: Ribonucleic acid. Probabilistic model. Splicing site, lncRNA, mRNA.

INTRODUÇÃO

A biologia molecular é fundamentada no Dogma Central da Biologia, a qual consiste no processo de transmissão da informação genética a partir do ácido desoxirribonucleico (DNA) (WATSON, 2015). Este fluxo de informação ocorre de três maneiras: duplicação; transcrição; e tradução. Através da duplicação, são geradas duas fitas a partir do DNA inicial, sendo ele seu próprio molde. Já no processo de transcrição, são expressas regiões gênicas por meio da síntese do ácido ribonucleico (RNA), chamado de RNA mensageiro (mRNA). Por último, a tradução, a qual caracteriza a síntese de proteínas, realizada a partir da leitura da fita de mRNA (WATSON, 2015).

A fita de mRNA é composta por éxons (regiões codificadoras) e íntrons (regiões não codificadoras). Como apenas os éxons detém de sequências que codificam proteínas, entre a transcrição e a tradução o mRNA submete-se a um processo de maturação chamado *splicing*, o qual retira os íntrons e mantém os éxons (LEVINE, 2001), processo ilustrado pela Figura 1. O limite entre éxon e íntron é chamado de sítio de *splicing*, sendo o sítio que inicia o íntron chamado de sítio doador, e o que termina o íntron de sítio aceitador. Uma característica destes sítios, é que mais de 99% apresentam um padrão, em que os íntrons iniciam com um dinucleotídeo GT e terminam com um dinucleotídeo AG (LEVINE, 2001).

Figura 1 – RNA antes e depois do processo de *splicing*



Fonte: Autoria própria (2019).

Além do mRNA, destacam-se no presente relatório os não codificadores de proteínas. Os genes codificadores de proteínas são muito semelhantes entre diferentes espécies, o que leva a inferir na importância dos não codificadores, os caracterizando como responsáveis por diversos aspectos de organismos complexos (USZCZYNSKA-RATAJCZAK, 2018). Os não codificadores são detentos de uma rica diversidade de unidades reguladoras e funcionais, sendo uma delas o loci que codifica *long non-coding* RNA (lncRNA); estes, possuem a característica de serem transcritos de RNA com um número maior ou igual a 200 nucleotídeos. Além disto, números crescentes de doenças humanas têm sido associadas a eles, porém, ainda são menos de 1% dos loci identificados (USZCZYNSKA-RATAJCZAK, 2018).

Ademais, os lncRNAs apresentam algumas resistências quanto à sua anotação em relação aos codificantes. Estes, por sua vez, são pouco expressos, o que significa que seus transcritos são fracamente amostrados; também, a associação sequência-função é ruim, o que dificulta encontrar novos lncRNAs, pela falta de características de sequências ou elementos funcionais; por último, são fracamente conservados durante a evolução, o que torna difícil identificar ortólogos ou parólogos por similaridade de sequência (USZCZYNSKA-RATAJCZAK, 2018).

Sendo assim, a proposta do presente trabalho foi acrescentar ao estado da arte dos lncRNA, buscando identificar se há diferença entre sítios de *splicing* de mRNA e lncRNA. Os resultados serão apresentados adiante.

MATERIAL E MÉTODOS

MATERIAL

Para a realização dos experimentos, utilizou-se o genoma na versão Tair 10 da planta *Aradopsis thaliana*, obtido através do banco de dados *Ensembl Genomes*. Ao todo foram três arquivos: o primeiro composto pelo genoma da planta, no formato fasta; e outros dois, no formato GTF, que contém um cabeçalho mapeador de regiões de transcritos, íntrons e éxons de mRNA e lncRNA.

A realização da análise destas sequências computacionalmente é um problema classificado como estatístico (DURBIN, 1998) e, uma das formas de efetuá-la é através de modelos probabilísticos. Estes, oferecem estruturas dependentes entre si, sendo possível executar a partir de uma base de dados, treinamentos para o amoldamento das probabilidades em um determinado problema, permitindo a previsão de futuros casos estatisticamente.

Dentre os tipos de modelos probabilísticos, há a Cadeia de Markov (CM). Esta é formada por um conjunto de estados, em que as transições entre tais estados e as suas emissões são parâmetros de probabilidade (DURBIN, 1998). Uma de suas características é ter como parâmetro para uma nova emissão apenas o seu último elemento. Entretanto em alguns casos, como o do presente experimento, é necessário considerar dependências entre posições adjacentes.

É possível generalizar uma CM para um tipo não homogênea - ou seja, dependente do tempo decorrido. Então, para se adequar às necessidades do problema em questão, foi utilizado o *Weight Array Model* (WAM), um tipo de CM de primeira ordem e não homogênea, idealizada para modelar sinais biológicos (BURGE, 1997). Sua vantagem está no fato de ser um modelo em que a posição dos nucleotídeos influencia nas probabilidades de transição entre os estados.

Para a classificação a partir dos modelos gerados, foi utilizado um classificador bayesiano, o qual é fundamentado no Teorema de Bayes, que calcula a probabilidade a posteriori de um evento, ou seja, as chances de um determinado evento ocorrer dado um outro já ocorrido. Este, tem boa aplicabilidade para estimar grandes conjuntos de parâmetros a partir de pouca quantidade de dados (DURBIN, 1998).

A execução dos treinamentos das WAMs e classificações bayesianas, foi realizada por meio da ferramenta ToPS (KASHIWABARA, 2013), que consiste em um *framework* para manipular modelos probabilísticos, fornecendo aplicabilidades para treinamento, simulação e decodificação a partir dos modelos de entrada.

Por fim, para a análise dos resultados obtidos a partir do ToPS, utilizou-se *scripts* em Python 3.6 juntamente com as bibliotecas PhyloPandas, para leitura dos arquivos fasta; Pandas e Numpy, para manipulação dos dados; e Scikit-Learn, para a geração das matrizes de confusão e métricas de aprendizado de máquina.

MÉTODOS

A partir do banco de dados, foram selecionadas sequências do genoma com janelas de tamanho igual a 20 nucleotídeos e com *offset* igual a 10, representando as posições dos dinucleotídeos AG nas sequências para sítios aceitadores e GT nas sequências para sítios doadores. Este processo resultou em quatro *datasets*, um para sítio doador e outros para aceitador de mRNA contendo 2000 sequências cada; e dois *datasets* de lncRNA, com 773 para o sítio doador e 816 para o sítio aceitador. A Figura 2 ilustra um exemplo do funcionamento destas sequências, porém com janelas de tamanho 10 e *offset* igual a 5.

Figura 2 – Janelas e seus *offsets* para sítio doador e aceitador, respectivamente

Fonte: Autoria própria (2019).

Com as sequências estabelecidas, executou-se a técnica de *cross-validation*. Este método consiste na divisão do *dataset* em k partes de mesmo tamanho, sendo $\frac{1}{k}$ utilizado para teste e $\frac{k-1}{k}$ para treinamento e, a partir desta divisão, a realização de k testes, os quais são medidos e analisados ao fim. Para o presente experimento, foi utilizado um k igual a 5.

Com os modelos WAM gerados a partir dos treinamentos e a aplicação de um classificador bayesiano através do ToPS, foram obtidos os resultados com a identificação entre codificadores e não codificadores, sendo geradas matrizes de confusão e tabelas para a análise. Este processo foi realizado para os sítios de *splicing* aceitadores e doadores.

RESULTADOS E DISCUSSÃO

RESULTADOS

Foram efetuados cinco testes para cada sítio de *splicing*, a partir dos treinamentos realizados com os conjuntos de dados. Com base nisto, para cada sítio foi gerada uma matriz de confusão, a qual contêm a relação entre os rótulos preditos e os rótulos verdadeiros, como é possível observar através da Tabela 1 e Tabela 2.

Tabela 1 – Matriz de confusão sítio aceitador

		Rótulos Preditos	
		Não codificante	Codificante
Rótulos Verdadeiros	Não codificante	0.49	0.51
	Codificante	0.35	0.65

Fonte: Autoria própria (2019).

Tabela 2 – Matriz de confusão sítio doador

		Rótulos Preditos	
		Não codificante	Codificante
Rótulos Verdadeiros	Não codificante	0.48	0.52
	Codificante	0.37	0.63

Fonte: Autoria própria (2019).

Também a partir das relações entre os rótulos, construiu-se uma tabela para cada sítio, a qual contém métricas comumente utilizadas em análises de aprendizado de máquina, entre eles o *F1 Score*, que é calculado a partir da precisão e *recall*, mostrando bons resultados com conjuntos de dados que apresentam classes desarmônicas. Os resultados podem ser observados na Tabela 3 e Tabela 4.

Tabela 3 – Resultados sítio aceitador

Métricas	Teste 1	Teste 2	Teste 3	Teste 4	Teste 5	Média	Desvio Padrão
Acurácia	0.57	0.60	0.64	0.54	0.65	0.60	0.047
Precisão	0.73	0.75	0.78	0.71	0.81	0.76	0.041
Recall	0.64	0.65	0.68	0.60	0.67	0.65	0.033
F1 Score	0.68	0.70	0.73	0.65	0.74	0.70	0.035

Fonte: Autoria própria (2019).

Tabela 4 – Resultados sítio doador

Métricas	Teste 1	Teste 2	Teste 3	Teste 4	Teste 5	Média	Desvio Padrão
Acurácia	0.61	0.55	0.69	0.58	0.59	0.59	0.031
Precisão	0.76	0.73	0.78	0.77	0.75	0.76	0.020
Recall	0.68	0.56	0.68	0.59	0.63	0.63	0.043
F1 Score	0.72	0.67	0.72	0.67	0.68	0.69	0.029

Fonte: Autoria própria (2019).

DISCUSSÃO

Com as matrizes de confusão geradas, descritas pelas Tabelas 1 e 2, tanto para os sítios doadores quanto para os aceitadores, foi possível observar uma relação de semelhança entre os RNAs codificantes e os não codificantes, com base nos valores de verdadeiros positivos e falsos negativos, que apresentaram resultados próximos de 50%.

Também, a partir das métricas registradas pelas Tabelas 3 e 4, os resultados obtidos mostraram uma qualidade do modelo de classificação dos sítios em torno de 70% $(\pm\ 0.032)$, provindo da média do *F1 Score*.

CONCLUSÃO

Obter as sequências dos lncRNAs representou uma tarefa desafiadora, devido à quantidade resultante de dados em um modo filtrado no genoma, em razão da análise feita apenas em regiões de *splicing*. Assim, a quantidade de sequências obtidas de lncRNA foi muito inferior à de mRNA, o que mostra a dificuldade em anotações de lncRNA.

Entretanto, tal dificuldade não impediu a realização dos testes, que foram realizados com sucesso. A partir destes, foram resultados valores de verdadeiros positivos e falsos negativos em torno de 50%, além de um *F1 Score* relativamente alto. Estes, mostraram pouca distinção entre os rótulos dos lncRNAs, sugerindo que há pouca diferença entre os sítios de *splicing* de lncRNAs e mRNAs, tanto em sítios aceitadores quanto sítios doadores, cumprindo o objetivo da presente pesquisa.

REFERÊNCIAS

BURGE, Christopher Boyce. **Identification of Genes in Human Genomic DNA**. 1997. Tese de Doutorado. Stanford University.

DURBIN, Richard et al. **Biological sequence analysis**. Cambridge University Press, 1996.

KASHIWABARA, André Yoshiaki et al. ToPS: a framework to manipulate probabilistic models of sequence data. **PLoS computational biology**, v. 9, n. 10, p. e1003234, 2013.

LEVINE, Aaron. Bioinformatics approaches to RNA splicing. 2001. Masters Thesis - University of Cambridge, UK.

USZCZYNSKA-RATAJCZAK, Barbara et al. Towards a complete map of the human long non-coding RNA transcriptome. **Nature Reviews Genetics**, v. 19, n. 9, p. 535, 2018.

WATSON, James D. et al. **Biologia molecular do gene**. Artmed Editora, 2015.

AGRADECIMENTOS

Agradeço a Deus, por nunca ter me desamparado; à minha família, que sempre acreditou nos meus sonhos; ao professor Dr. André Yoshiaki Kashiwabara, pela oportunidade e confiança; aos meus queridos amigos da UTFPR-CP; e à Fundação Araucária, pela concessão da bolsa de iniciação científica.