UM SOFTWARE PARA IDENTIFICAÇÃO DE REGIÕES CODANTES EM GENES DE FUNGOS FILAMENTOSOS E SUA TRADUÇÃO PARA OS POLIPEPTÍDEOS CORRESPONDENTES

Gustavo Henrique Ferreira Cruz (PIC-UEM), Vinícius Menossi (PIC-UEM), Josiane Melchiori Pinheiro (orientadora), João Alencar Pamphile (co-orientador), e-mails: {ra109895, ra108840, jmpferreira, japamphile}@uem.br

Universidade Estadual de Maringá / Centro de Tecnologia / Maringá, PR.

Ciência da Computação/Metodologia e Técnicas da Computação

Palavras-chave: aprendizado de máquina, identificação de regiões codantes, tradução proteica

Resumo:

Considerando a necessidade de identificar as regiões codantes e não codantes em sequências de DNA, este trabalho investiga uma solução computacional, que faz uso de Aprendizado de Máquina (AM) para tentar predizer estas regiões. Foram treinados modelos do algoritmo *Conditional Random Fields*, um modelo de AM supervisionado, para os fungos *Colletotrichum* e *Diaporthe.* O algoritmo foi estruturado em 3 módulos para preparar os dados do GenBank e treinar o modelo. O desempenho dos modelos treinados foram promissores nas medidas obtidas durante o treinamento (medida F1 acima de 0,84), mas na avaliação da precisão real da identificação das sequências codantes e não codantes, o modelo treinado para o fungo *Colletotrichum* na proteína actina foi melhor do que os resultados do modelo treinado do fungo *Diaporthe* na proteína beta tubulina.

Introdução

Atualmente, identificar as regiões codantes do mRNA que podem ser traduzidas para proteínas é uma tarefa pouco automatizada. Alguns trabalhos apontam que esta é uma área de pesquisa com bastante potencial para a aplicação de técnicas computacionais, como, por exemplo, o AM (RÄTSCH *et al.*, 2007).

Durante a síntese de proteínas, o DNA é utilizado para gerar o RNA mensageiro (mRNA), que após formado, sai do núcleo da célula, é liberado no citoplasma e direcionado até o encontro dos ribossomos, onde ocorre o processo de síntese de proteínas conhecido como tradução (WATSON et al., 2015).

Porém, nem todos os nucleotídeos do mRNA são codantes, ou seja, são utilizados na síntese de proteínas. Partes do mRNA são descartadas e não participam do processo de tradução, são os chamados íntrons. Essas partes

se alternam com as regiões codantes na sequência, os chamados éxons. Todos os organismos eucariontes possuem íntrons em suas estruturas genéticas e antes do processo de tradução é necessário remover essas partes do mRNA através de um mecanismo molecular chamado spliceossoma no processo conhecido como *splicing* de RNA (CHOW et al., 1977). Segundo Alberts et al. (2017), há padrões nos nucleotídeos quase invariantes no processo de identificação de íntrons, esses padrões são chamados de "sequências consenso", sendo um destes a presença das bases GU no início do íntron e AG no seu final.

Este trabalho busca continuar os estudos iniciados por Ferreira (2019), que modelou o problema de identificação das regiões codantes em genes de fungos filamentosos como um problema de AM supervisionada. Os exemplos da base de dados para o treinamento do modelo *Conditional Random Fields* (CRF) foram coletados de sequências oriundas do GenBank[1] e transformadas em sequências menores que poderiam ser classificadas em uma das duas classes: intron e not-intron.

Neste trabalho, o problema de identificação das regiões codantes em genes também foi modelado como um problema de AM supervisionada, mas uma nova base de dados foi criada, identificando três tipos de classes: íntron, éxon e *neither*. Uma estrutura de código modular foi implementada, para tornar o software mais manutenível nos estudos futuros. Também foi construído um módulo para a tradução das regiões codantes para a proteína e foram feitas comparações tanto da sequência codante com os dados vindos do GenBank, quanto das proteínas resultantes no BLAST[2].

Materiais e métodos

O algoritmo para preparar os dados e treinar o modelo CRF foi implementado em 3 módulos, utilizando a linguagem Python e as bibliotecas *Biopython*, *Sklearn* e *Sklearn-crfsuite*. Inicialmente, os dados são selecionados do GenBank e guardados em arquivos de texto.

Foram selecionadas do GenBank as sequências de DNA dos genes do fungo *Colletotrichum* correspondentes à proteína actina (5209 sequências) e dos genes do fungo *Diaporthe* correspondentes à proteína beta tubulina (1821 sequências).

O primeiro módulo extrai dos arquivos do GenBank as informações necessárias para a construção da base de dados. O resultado parcial é salvo em um novo arquivo, que é a entrada do segundo módulo, que prepara a base de dados no formato de entrada para o treinamento do CRF.

A base de dados foi estruturada de forma que os exemplos possuem apenas uma *feature* que os descreve, ou seja, a própria sequência, candidata a ser classificada como íntron, éxon ou *neither*. Os exemplos de sequências que formam a base foram construídos levando-se em conta os padrões de GU

1 Mantido e gerenciado pelo *National Center for Biotechnology Information*, o GenBank conta com milhares de sequências de diversos seres vivos e seus marcadores genômicos.

2 O BLAST é um algoritmo usado na comparação de informações biológicas, como sequências de aminoácidos ou nucleotídeos de sequências de DNA.

no início do íntron e AG no final. Neste módulo são geradas todas as combinações de possíveis íntrons e possíveis éxons de cada uma das sequências, sendo que as sequências que são verdadeiramente íntron ou éxon, são identificadas no arquivo inicial vindo do GenBank e as sequências restantes são classificadas como *neither*. Para não gerar uma base com uma quantidade desbalanceada de exemplos de cada classe, a quantidade de exemplos na classe *neither* é igual à média simples entre a quantidade de íntrons e éxons na base.

O terceiro módulo, chamado de módulo de treinamento, é alimentado com os dados gerados no segundo módulo, no formato de entrada do CRF. É neste módulo que o modelo é treinado e salvo para poder ser usado em predições futuras.

O modelo treinado é usado no módulo de aplicação, onde novos exemplos de dados são fornecidos ao modelo, que resulta em uma classificação para cada exemplo. A partir dessas classificações, a sequência inicial precisa ser "remontada" para determinar quais são as sequências de íntrons que precisam ser retiradas da sequência inicial, para que apenas as sequências codantes (éxons) sejam enviadas ao processo de tradução para a produção da proteína. Para gerar a proteína resultante o processo de tradução utiliza um dicionário de trincas (códons) para a proteína dentro de uma iteração nas trincas da sequência.

Para testar a precisão real da abordagem foi criado um algoritmo que recupera cada uma das sequências iniciais do GenBank, gera os exemplos de sequências baseadas no padrão GU e AG dos íntrons, submete os exemplos ao modelo treinado do CRF, e por fim, retira da sequência inicial as regiões classificadas como íntrons, e a compara com a sequência final correta do GenBank. Se a comparação de todos os caracteres for idêntica, o resultado é dado como correto, caso contrário, é dado como incorreto.

Resultados e Discussão

Foram treinados dois modelos do CRF, uma para cada fungo selecionado. Os resultados do modelo treinado para o fungo *Colletotrichum* (actina) foram satisfatórios, obtendo uma acurácia de 83,3% nas comparações com os resultados gerados pelo modelo em relação às sequências corretas do GenBank. O modelo treinado para o fungo *Diaporthe* (beta tubulina), não obteve resultados tão promissores, atingindo uma acurácia, para o mesmo teste aplicado ao *Colletotrichum*, de 27,3%. As diferenças nas acurácias se devem ao fato de que, para cada fungo, a proteína selecionada é de uma região diferente do gene, sendo que as sequências da beta tubulina são maiores quando comparadas às sequências da actina do *Colletotrichum*. As medidas de desempenho geradas pelo algoritmo de treinamento do modelo mostraram sempre valores maiores da medida F1 para a classificação dos íntrons, reforçando a afirmação de Alberts et al. (2017) sobre as sequências de consenso.

Conclusões

Com a análise dos resultados obtidos podemos notar que o modelo CRF treinado para o fungo *Colletotrichum* (proteína actina) apresentou bons resultados na identificação das regiões não codantes, gerando assim, sequências codantes com ótima precisão. Por outro lado, o modelo CRF treinado para o fungo *Diaporthe* (proteína beta tubulina), de maneira geral, não apresentou bons resultados, mostrando uma possível evidência de que o CRF pode não lidar bem com sequências maiores, como são as sequências da beta tubulina.

Agradecimentos

À Universidade Estadual de Maringá que possibilitou a execução deste trabalho através do Programa de Iniciação Científica - PIC. Ao prof. Dr. João Alencar Pamphile (*in memorian*), co-orientador deste trabalho, assim como Gustavo Luiz Furuhata Ferreira (egresso 2019), que foram cruciais para o desenvolvimento desse tema.

Referências

ALBERTS, B. et al. **Biologia molecular da célula.** 6. ed. [S.1]: Artmed Editora, 2017. ISBN 9788582714225.

CHOW, L.T., ROBERTS, J.M., LEWIS, J.B., BROKER, T.R. (1977) **A map of cytoplasmic RNA transcripts from lytic adenovirus type 2**, determined by electron microscopy of RNA:DNA hybrids.

FERREIRA, GUSTAVO L. F. **Identificando regiões não codantes em sequências de DNA utilizando técnicas de aprendizado de máquina.** 54p. TCC (Bacharelado). Universidade Estadual de Maringá. Maringá, 2019.

RÄTSCH, G. et al. **Improving the caenorhabditis elegans genome annotation using machine learning.** PLoS Computational Biology, Public Library of Science, v. 3, n. 2, p. e20, 2007.

WATSON, JAMES D. et al. **Biologia Molecular do Gene**. 7. ed. Artmed, 2015.