ESTUDO E TREINAMENTO DE UMA LSTM PARA A CLASSIFICAÇÃO DE REGIÕES CODANTES EM SEQUÊNCIAS DE DNA CONSIDERANDO BASES DEGENERADAS

Bruno Rafael Comin Scheffel (PIC/UEM), Josiane Melchiori Pinheiro (Orientador). E-mail: jmpferreira@uem.br.

Universidade Estadual de Maringá, Departamento de Informática, Maringá, PR.

Área e subárea do conhecimento: Ciência da Computação / Metodologia e Técnicas da Computação

Palavras-chave: Bioinformática; Aprendizado profundo; Anotação genômica.

RESUMO

A identificação de éxons e íntrons em sequências de DNA, tarefa de extrema relevância para anotação genômica, tem sido apoiada por aprendizado de máquina. No entanto, a maioria das abordagens computacionais lida apenas com as quatro bases nitrogenadas canônicas, descartando as bases degeneradas, símbolos de ambiguidade da IUPAC comuns na rotina de sequenciamento. Para contornar essa limitação, este trabalho adaptou uma rede LSTM bidirecional para classificação de splicing, incorporando nativamente essas ambiguidades ao substituir a camada de *embedding* original por uma codificação *one-hot* probabilística. Utilizando um corpus do GenBank com as sequências ambíguas preservadas, o modelo demonstrou robustez: atingiu 97,36% de acurácia (*F1-score* de 0,9801 para éxons) no cenário de controle e manteve o *F1-score* estável (entre 0,9816 e 0,9824) em cinco cenários de injeção de degeneração. Além disso, a reconstrução das fitas codantes registrou identidade média entre 0,9677 e 0,9701, superando o trabalho de referência. Conclui-se que é plenamente viável incluir dados ambíguos no treinamento desses modelos sem perda de desempenho, evitando descartes desnecessários e aproximando a ferramenta da realidade prática dos laboratórios.

INTRODUÇÃO

O progresso contínuo nas tecnologias de sequenciamento genético introduziu novos desafios para a anotação genômica. Para lidar com essa crescente complexidade de dados, a área tem recorrido cada vez mais às ferramentas de bioinformática e aos algoritmos de aprendizado de máquina visando a correta análise e classificação das cadeias de DNA.

Um dos focos de maior relevância nesse domínio é a predição das regiões de *splicing*, os limites entre éxons e íntrons ao longo da cadeia de DNA (Alberts et al., 2017). Entre os modelos já treinados para essa tarefa, destacam-se as redes recorrentes do tipo LSTM (*Long Short-Term Memory*) (Hochreiter; Schmidhuber,

1997), adequadas por capturarem dependências de longo alcance ao longo da sequência.
Contudo, grande parte dessas soluções processa exclusivamente o alfabeto genético canônico (A, C, T, G), apresentando falhas ao lidar com as bases degeneradas (símbolos de ambiguidade padronizados pela IUPAC - *International Union of Pure and Applied Chemistry*) rotineiramente encontradas nos resultados reais de laboratório.
Foi nesse contexto que Menossi (2023) desenvolveu um modelo LSTM para a classificação de regiões de *splicing* em sequências de DNA do gene da actina, tanto em animais quanto em fungos, alcançando bons resultados, considerando apenas as quatro bases canônicas. Seu modelo, no entanto, descartava totalmente os registros com bases degeneradas.
Partindo dessa mesma arquitetura como ponto de referência, este trabalho tem como objetivo estudar a importância dessas ambiguidades na detecção de regiões codantes em sequências de DNA e propor uma adaptação na arquitetura LSTM de Menossi (2023), capaz de incorporá-las nativamente ao aprendizado, sem descartar sequências reais por conterem bases degeneradas.

MATERIAIS E MÉTODOS

A metodologia caracteriza-se como pesquisa aplicada, experimental e quantitativa, desenvolvida em linguagem Python com o apoio das bibliotecas Biopython, Keras/TensorFlow e minineedle. Utilizou-se um corpus de sequências do gene da actina em animais e fungos, extraído do GenBank e mantido pelo NCBI, mas preservando os registros com bases degeneradas que a modelagem de Menossi (2023) havia descartado por completo.
Porém, muitas sequências do GenBank chegam "limpas" de bases degeneradas, tornando o corpus extraído ineficiente para o objetivo deste trabalho. Dessa forma, optou-se por simular essa ambiguidade, o que exigiu um método de injeção controlada para reproduzir a incerteza real do sequenciamento. Essa estratégia une duas frentes: uma de caráter evolutivo, ligada à conservação estrutural das cadeias genéticas (Castle, 2011), e outra de ordem técnica, pautada no perfil de erros característico das plataformas Illumina, atualmente as mais utilizadas em sequenciamento, onde erros de leitura ocorrem mais em trechos com bases repetidas (homopolímeros) e nas extremidades com maior taxa de GC (Nakamura et al., 2011). Para regular a influência de cada um desses fatores na taxa final de inserção, aplicou-se um parâmetro de mistura α, com valores variando entre 0 e 1. Um segundo parâmetro, a intensidade global de injeção (*rate*), escala uniformemente a taxa final já combinada pelos dois fatores, controlando a quantidade total de ambiguidade introduzida, independentemente da proporção entre eles.
A classificação foi estruturada por janela deslizante de 60 nucleotídeos, com cada base representada por uma codificação *one-hot* probabilística: bases canônicas concentram valor 1,0 em uma única dimensão, enquanto códigos degenerados distribuem essa massa entre as bases que admitem, o que dispensa a camada de

embedding usada na modelagem de Menossi (2023). A rede proposta é uma LSTM bidirecional com 60 unidades por sentido de leitura, seguida de camada densa (ReLU), *dropout* de 0,3 e neurônio de saída sigmóide, com 35.105 parâmetros treináveis (58.501 no modelo de referência). O treinamento usou *Binary Cross Entropy*, otimizador Adam (taxa de 5×10^{-4}) e 10 épocas.

Para a validação, estruturou-se um protocolo de testes com quatro diferentes cenários de intensidade de injeção (*rate* = 0,25; 0,50; 0,75; 1,00), com α fixado em 0,5, aplicados ao corpus de actina de animais e fungos. O particionamento dos dados ocorreu no nível da sequência, destinando 80% do volume para o treinamento e os 20% restantes para o teste. Além das métricas tradicionais computadas por nucleotídeo, realizou-se a remontagem da fita codante predita, aplicando-se antes uma etapa de suavização via voto majoritário onde, para cada posição, avalia-se uma janela de 7 predições ao seu redor, e a classe final é definida pela maioria dessa janela, corrigindo predições isoladas. Por fim, a sequência remontada foi contrastada com a versão original mediante a aplicação do algoritmo de alinhamento global de Needleman-Wunsch (Needleman; Wunsch, 1970).

RESULTADOS E DISCUSSÃO

No cenário de controle (0% de injeção), o modelo alcançou acurácia global de 97,36% e *F1-score* de 0,9801 para a classe éxon e 0,9607 para a classe íntron, valor próximo ao relatado por Menossi (2023) em sua modelagem de referência (0,97).

Nos quatro cenários de intensidade de injeção mista, o *F1-score* da classe éxon oscilou entre 0,9816 e 0,9824, sem sinal de degradação relevante associado ao aumento da intensidade de ambiguidade, mesmo no cenário de intensidade máxima (*rate* = 1,00), em que se esperava maior impacto. Essa estabilidade pode ser atribuída à leitura bidirecional da arquitetura Bi-LSTM, que permite inferir a classe do nucleotídeo central a partir do contexto vizinho mesmo diante de símbolos ambíguos, sem depender exclusivamente da exatidão de uma base isolada. As pequenas oscilações nas casas decimais provavelmente refletem o ruído da otimização de uma única rodada de treinamento, mais do que um efeito real do parâmetro *rate*.

No que diz respeito à reconstrução da cadeia codante, validada por meio do algoritmo de alinhamento global de Needleman-Wunsch, observou-se uma identidade média situada entre 0,9677 e 0,9701 ao longo dos cinco cenários testados. Da mesma forma, notou-se apenas uma leve tendência de degradação decorrente da intensidade de injeção de bases degeneradas, de pequena magnitude, mostrando-se ainda assim superior à similaridade média de 0,961 relatada no estudo de Menossi (2023).

CONCLUSÕES

Os resultados mostram que é possível incorporar nativamente as bases degeneradas ao treinamento de modelos de anotação genômica baseados em

LSTM, sem precisar descartar sequências reais por conterem ambiguidades. A arquitetura proposta manteve a acurácia acima de 97% e identidade de remontagem estável em todos os cenários testados. Contudo, a pesquisa apresenta como fator limitante o uso restrito a sequências do gene da actina, uma vez que, por codificarem uma proteína com alto grau de conservação evolutiva, a generalização destas conclusões para outros contextos genômicos fica comprometida. Diante disso, recomenda-se para trabalhos futuros a ampliação das análises para incluir genes com menor nível de conservação e amostras diretas provenientes da rotina de sequenciamento laboratorial.

AGRADECIMENTOS

À Universidade Estadual de Maringá (UEM), pelo apoio institucional que viabilizou o desenvolvimento desta pesquisa; e à orientadora, Profa. Dra. Josiane Melchiori Pinheiro, pela orientação ao longo de todo o projeto.

REFERÊNCIAS

ALBERTS, Bruce *et al.* **Biologia molecular da célula**. 6. ed. Porto Alegre: Artmed, 2017.

CASTLE, J. C. SNPs occur in regions with less genomic sequence conservation. *PLoS ONE*, v. 6, n. 6, e20660, 2011.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735-1780, 1997.

MENOSSI, V. **Estudo da aplicação de redes neurais recorrentes na classificação de regiões de splicing em sequências de DNA.** 2023. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) — Departamento de Informática, Centro de Tecnologia, Universidade Estadual de Maringá, Maringá, 2023.

NAKAMURA, K. et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, v. 39, n. 13, e90, 2011.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, v. 48, n. 3, p. 443-453, 1970.