PELICAN: FERRAMENTA PARA INFERÊNCIA DE ANOTAÇÃO DE CONSENSO EM GENOMAS DE FAGOS

Pedro Rossi de Andrade Franco

Guilherme Wenceslau de Lima Cardoso, Guilhermo Uceda Campos, Fernando Pacheco Nobre Rossi

Prof. Dr. João Carlos Setubal

Universidade de São Paulo

pedro.andrade.franco@usp.br

Objetivos

A anotação de genomas de bacteriófagos (fagos), vírus que infectam bactérias, é uma etapa importante para a compreensão de sua biologia. Porém, as principais ferramentas computacionais de anotação frequentemente produzem resultados divergentes entre si. Nesse contexto, o presente projeto tem como objetivo principal desenvolver uma ferramenta capaz de gerar automaticamente uma anotação consenso para genomas de fagos. Objetivos específicos da ferramenta, intitulada PELICAN, são: agregar e comparar predições de algumas ferramentas existentes de anotação; avaliar os genes identificados por cada uma; e gerar uma predição final de forma unificada e mais acurada, baseada no consenso entre os diferentes métodos utilizados.

Métodos e Procedimentos

O *pipeline* de análise do PELICAN inicia com um genoma de bacteriófago em formato FASTA dado pelo usuário. Quatro ferramentas de anotação são então executadas em paralelo: Phanotate, PROKKA-virus, Prodigal e GLIMMER. O Consenso I é identificar os genes preditos por pelo menos duas das ferramentas empregadas. Os genes restantes, ou seja, aqueles que não entraram no primeiro consenso, são submetidos a uma busca por similaridade de sequência, por meio de BlastP, contra a base de dados de proteínas de fagos PHROGS. Os resultados com *hits* significativos formam o Consenso II [1]. Finalmente, os genes que ainda permanecem sem classificação são analisados por um modelo de *deep learning* (AutoEncoder) baseado em redes neurais convolucionais. Este modelo é treinado para reconhecer características de genes de fagos, e classifica as sequências restantes, gerando o Consenso III. A anotação final é a união dos três conjuntos de consenso, conforme exemplificado na Figura 1.

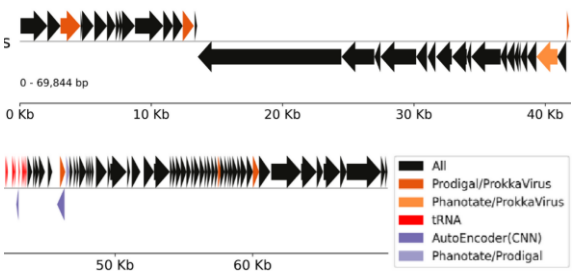

Figura 1: Consenso do genoma do bacteriófago NC_048638.1

A performance da ferramenta foi testada preliminarmente usando 20 genomas de fagos do projeto Sea-Phages [2], comparando resultados do PELICAN com os resultados da anotação manual e de outras ferramentas.

Resultados

Ao ser aplicada aos testes, a ferramenta PELICAN predisse um total de 2.467 genes, um número superior aos genes previstos por ferramentas como PROKKA-virus (2.188), PHAROKKA (2.414), GLIMMER (2.127). Em relação à anotação manual, PELICAN alcançou uma média de cobertura de 97,93% dos genes anotados. Este resultado é maior que o obtido pelo PROKKA-virus (94,19%), pelo PHAROKKA (81,04%) e pelo GLIMMER (82,71%), vide a Figura 2.

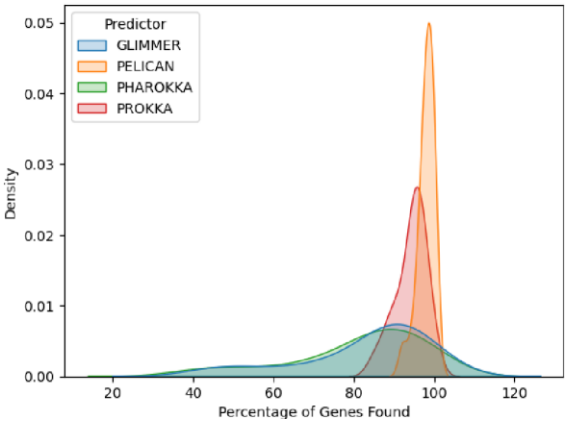

Figura 2: Distribuição da cobertura dos genomas por preditor

Além disso, PELICAN apresentou a menor quantidade de genes perdidos (genes não identificados pela ferramenta em relação aos genes identificados por anotação manual), com apenas 62 genes perdidos enquanto PROKKA-virus, GLIMMER e PHAROKKA perderam 182, 399 e 427, respectivamente.

Conclusões

A ferramenta PELICAN se mostrou robusta e eficaz na anotação dos genomas dos fagos testados até agora. A estratégia de combinar o consenso entre anotadores, a homologia com bancos de dados e o modelo de *deep learning* permitiu alcançar uma cobertura gênica maior e um menor número genes perdidos em relação com as ferramentas padrão de anotação. Os resultados preliminares sugerem que PELICAN consegue automatizar o processo de anotação com alta precisão, além de refinar anotações existentes, aproximando-se da qualidade da curadoria manual. Os próximos passos incluem a validação da ferramenta com dados de transcriptômica e com genomas de fagos recém-isolados pelo grupo de pesquisa.

Os colaboradores declaram não haver conflito de interesses. Rossi concebeu a ferramenta. Franco, Campos e Rossi construíram a ferramenta. Rossi, Cardoso, Campos e Rossi testaram a ferramenta. Todos os colaboradores aprovaram a versão final do resumo.

Agradecimentos

Agradeço à Prof.ª Dr.ª Aline Maria da Silva (*in memoriam*) pela orientação inicial no projeto e ao Prof. Dr. João Carlos Setubal por aceitar me orientar após a perda da Prof.ª Aline. Agradeço também aos colegas de projeto, de laboratório e à Prof.ª Dr.ª Regina Baldini por todo o apoio. Por fim, agradeço ao CNPq pela bolsa de ITI-A.

Referências

[1] TURNER, Dann et al. Phage annotation guide: guidelines for assembly and high-quality annotation. **Phage**, v. 2, n. 4, p. 170-182, 2021.
[2] RUSSELL, Daniel A.; HATFULL, Graham F. PhagesDB: the actinobacteriophage database. **Bioinformatics**, v. 33, n. 5, p. 784-786, 2017.