ESTUDO E OTIMIZAÇÃO DO FFGC PARA ACELERAR ANÁLISES GENÔMICAS

Autores(as): BRAGA, A. F.[1]; BARALDI, M. S.[1]; MARTINEZ, V. F.[2]

[1]Grupo PET-Computação, UFMS, Campus Cidade Universitária; [2]Tutor(a) do Grupo PET-Computação, UFMS, Campus Cidade Universitária
E-mail: fernando.braga@ufms.br, pet-comp.facom@ufms.br

RESUMO: A genômica comparativa, campo que investiga a evolução através da análise da estrutura e conteúdo de genomas, enfrenta desafios metodológicos como o "problema circular" na análise da ordem gênica. O software Family-Free Genome Comparison (FFGC) foi desenvolvido para superar essa limitação com uma abordagem "livre de famílias", porém seu fluxo de trabalho padrão apresenta um alto custo computacional na etapa de cálculo de similaridade, que realiza um alinhamento "todos contra todos". Este trabalho propõe e implementa uma otimização do FFGC através de um fluxo de trabalho alternativo que utiliza como entrada um arquivo de famílias de genes pré-computadas em nível de refinamento 0. A metodologia envolveu a modificação do script de inicialização do projeto e a criação de novas regras no workflow do Snakemake para restringir os alinhamentos apenas aos genes dentro de cada família. Os resultados demonstraram uma redução drástica no tempo de execução, com destaque para o alinhador BLAST+, que se tornou mais rápido que sua execução no fluxo padrão e superou significativamente o desempenho do DIAMOND na nova abordagem. A otimização valida a eficácia da estratégia em reduzir o custo computacional, tornando as análises genômicas em larga escala mais eficientes.

Palavras-chave: Ortologia; Refinamento; Genética.

STUDY AND OPTIMIZATION OF FFGC TO ACCELERATE GENOMIC ANALYSES

ABSTRACT : Comparative genomics, a field that investigates evolution through the analysis of genome structure and content, faces methodological challenges such as the 'circular problem' in gene order analysis. The Family-Free Genome Comparison (FFGC) software was developed to overcome this limitation using a 'family-free' approach; however, its standard workflow presents a high computational cost in the similarity calculation step, which performs an all-against-all alignment. This work proposes and implements an optimization for FFGC through an alternative workflow that uses a file of pre-computed gene families at refinement level 0 as input. The methodology involved modifying the project's initialization script and creating new rules in the Snakemake workflow to restrict alignments to only the genes within each family. The results showed a drastic reduction in execution time, with emphasis on the BLAST+ aligner, which became faster than its execution in the standard workflow and significantly surpassed DIAMOND's performance in the new approach. The

optimization validates the strategy's effectiveness in reducing computational cost, making large-scale genomic analyses more efficient.

Keywords: Orthology; Refinement; Genetics.

1 INTRODUÇÃO

A genômica comparativa é um campo da biologia que busca compreender as relações evolutivas e funcionais entre as espécies através da análise de seus genomas (Moreira, 2015). Uma de suas ferramentas mais poderosas é a análise da ordem gênica, que estuda como a disposição dos genes ao longo dos cromossomos é conservada ou modificada. Contudo, os métodos tradicionais enfrentam um "problema circular": eles exigem o agrupamento de genes em famílias de homólogos antes da comparação, um passo inicial propenso a erros que se propagam e comprometem a análise.

Para superar esse desafio, foi desenvolvido o software Family-Free Genome Comparison (FFGC) (Doerr, 2018). Esta ferramenta implementa uma abordagem "livre de famílias", que integra a inferência de homologia com a comparação da estrutura genômica em uma única etapa. No entanto, a etapa mais importante do FFGC, o cálculo da similaridade inicial entre os genomas, tem um custo computacional excessivamente elevado. O objetivo deste trabalho é otimizar o FFGC, propondo uma alteração em seu fluxo para contornar essa etapa de alto custo, visando uma redução drástica no tempo de execução.

2 MÉTODO

O FFGC utiliza o Snakemake (Mölder, 2021) como gerenciador de seu fluxo de trabalho, que se baseia na definição de regras para gerar arquivos de saída a partir de arquivos de entrada. A otimização consistiu em criar um fluxo de trabalho alternativo que contorna o método padrão de cálculo de similaridade.

2.1 FLUXO DE TRABALHO PADRÃO DO FFGC

O processo normal do FFGC representa o principal gargalo de desempenho e envolve três regras principais:

- Criação do Banco de Dados: Um banco de dados de sequências único e binário é construído a partir de todos os genes dos genomas analisados.

- Execução do Alinhamento: O alinhador (BLAST ou Diamond (Buchfink, 2021)) realiza uma comparação massiva de cada gene contra todos os outros, gerando uma tabela de similaridade para cada genoma.
- Processamento das Similaridades: As tabelas brutas são processadas e organizadas em arquivos de similaridade par a par (ex: G1_G2.sim).

2.2 IMPLEMENTANDO O FLUXO DE TRABALHO OTIMIZADO

O novo fluxo de trabalho é ativado por um novo parâmetro de linha de comando (-families-10) na criação do projeto, que sinaliza o uso de um arquivo de famílias de genes pré-calculadas. Este novo fluxo substitui as regras de alto custo por uma nova sequência:

Extração das Sequências das Famílias: O sistema lê o arquivo de famílias e cria múltiplos arquivos FASTA, um para cada família, contendo as sequências completas de todos os seus membros.

Cálculo de Similaridade Intra-Família: A comparação de similaridade é restrita apenas aos genes dentro da mesma família. Para o BLAST, foi utilizada uma funcionalidade que permite a comparação direta entre sequências sem a necessidade de um banco de dados, o que simplifica e acelera o processo. Para o Diamond, foi necessário criar um banco de dados para cada arquivo de família, o que gerou sobrecarga de processamento.

Consolidação dos Resultados: Os múltiplos arquivos de resultado gerados (um por família) são consolidados e reorganizados com base em seus genomas de origem, alinhando a saída com o formato esperado pelo restante do fluxo padrão do FFGC.

3 RESULTADOS E DISCUSSÃO

A implementação do fluxo de trabalho alternativo alterou drasticamente o desempenho computacional, com uma notável inversão na performance entre o BLAST+ e o DIAMOND.

Os tempos de execução em segundos foram:

Tabela 1 – Desempenho de cada fluxo de trabalho.

Fluxo de Trabalho	Alinhador	Tempo de Execução
Padrão	DIAMOND	101,51
Padrão	BLAST+	2038,82
Alternativo	DIAMOND	525,15
Alternativo	BLAST+	316,87

Fonte: Autoria própria (2025).

No fluxo padrão, o DIAMOND foi aproximadamente 20 vezes mais rápido que o BLAST+, devido à sua otimização para comparações "todos contra todos" a partir de um banco de dados único.

Contudo, no fluxo alternativo, o BLAST+ apresentou uma redução de tempo de execução de aproximadamente 84%, tornando-se a configuração mais rápida. A razão para este ganho de desempenho foi sua capacidade de realizar a comparação direta entre as sequências sem a necessidade de criar múltiplos bancos de dados. Por outro lado, o desempenho do DIAMOND foi inferior, pois a exigência de gerar um banco de dados para cada família introduziu um gargalo de processamento significativo, superando a velocidade de seu algoritmo de alinhamento neste cenário específico.

4 CONCLUSÃO

Este trabalho demonstrou com sucesso a otimização do software FFGC ao implementar um fluxo de trabalho alternativo que utiliza famílias de genes pré-computadas (nível 0). A estratégia contornou o principal gargalo de desempenho do método original, resultando em uma redução drástica no tempo de execução. Notavelmente, o BLAST+ obteve uma redução de aproximadamente 84% em seu tempo, superando o DIAMOND no novo fluxo. Conclui-se que a modificação torna o FFGC uma ferramenta mais ágil e eficiente para análises genômicas em larga escala, destacando também como as características das ferramentas de bioinformática influenciam o desempenho em diferentes workflows.

AGRADECIMENTOS

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo fomento a esta pesquisa através da concessão de bolsa. Agradecemos especialmente ao professor Diego Rubert por sua orientação precisa, pelo suporte contínuo e pelas discussões enriquecedoras que foram fundamentais para o desenvolvimento deste estudo.

REFERÊNCIAS

[1] DOERR, D.; FEIJÃO, P.; STOYE, J. Family-Free Genome Comparison. *In*: KUTTER, Claudia (ed.). **Comparative Genomics**. New York: Humana Press, 2018. p. 331-342. (Methods in Molecular Biology, v. 1704). DOI: 10.1007/978-1-4939-7463-4_1.

[2] MOREIRA, Leandro Marcio (org.). **Ciências genômicas**: fundamentos e aplicações. Ribeirão Preto: Sociedade Brasileira de Genética, 2015. 403 p. ISBN 978-85-89265-22-5.

[3] MÖLDER, F. et al. Sustainable data analysis with Snakemake. **F1000Research**, v. 10, art. 33, 2021. DOI: 10.12688/f1000research.29032.1.

[4] BUCHFINK, B.; REUTER, K.; DROST, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. **Nature Methods**, v. 18, n. 4, p. 366-368, 2021. DOI: 10.1038/s41592-021-01101-x.