Avaliação da aplicação de paralelismo em classificadores taxonômicos usando Qiime2

Caetano Müller, Junior Löff, Dalvan Griebler, Eduardo Eizirik

[1] Escola Politécnica, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brasil

{caemuller1,loffjh}@gmail.com, {dalvan.griebler, eduardo.eizirik}@pucrs.br

Resumo. *A classificação de sequências de DNA usando algoritmos de aprendizado de máquina ainda tem espaço para evoluir, tanto na qualidade do resultado quanto na eficiência computacional dos algoritmos. Nesse trabalho, realizou-se uma avaliação de desempenho em dois algoritmos de aprendizado de máquina da ferramenta Qiime2 para classificação de sequências de DNA. Os resultados mostram que o desempenho melhorou em até* $9,65$ *vezes utilizando* 9 *threads.*

1. Introdução

O estudo de genomas é importante para o entendimento dos seres vivos que convivem em nossa biosfera. Para organismos de maior tamanho, é possível identificar suas diferenças através das suas características visíveis. Para organismos microscópicos existe um grau de dificuldade maior apenas analisando suas aparências. Portanto, utiliza-se o DNA para diferenciá-los e ressaltar suas características através do código genético. Após uma amostra de DNA ser sequenciada, ou seja, determinar a sequência de nucleotídeos dessa amostra, é importante comparar essa sequência com outras já catalogadas em uma base de dados a fim de obter a sua taxonomia. A taxonomia pode ser vista como uma atribuição destes seres a um organismo semelhante e já conhecido. Obtendo a taxonomia é possível saber a qual grupo de seres-vivos aquela sequência de DNA pertence. Essa classificação é crucial para dar significado aos dados e conseguir avançar nas investigações sobre a relação entre os seres vivos.

Para obter a taxonomia de uma sequência de DNA, utiliza-se um algoritmo para classificação taxonômica. A metodologia mais eficiente para obter essa classificação é utilizando algoritmos de aprendizado de máquina. Com isso, as sequências amostrais são comparadas com dados genéticos de uma base de dados referência, tornando assim possível uma comparação entre as sequências e a assimilação de cada sequência a seu determinado grupo ecológico.

Existem diversas ferramentas e soluções para executar classificações taxonômicas. Por exemplo, o BLAST (*Basic Local Alignment Search Tool*) é uma das ferramentas mais antigas e muito utilizada em pesquisas na área de genomas. Outra solução é o Kraken2 [Lu and Salzberg 2020], que foi implementado em C++ e têm mostrado resultados positivos no quesito desempenho. Devido ao seu fácil acesso, qualidade dos resultados e constante melhora em seu sistema, nesse trabalho será utilizado o Qiime2 (*Quantitative Insights Into Microbial Ecology*). O Qiime2 representa o estado-da-arte nesse tipo de análise. Mais detalhes sobre a ferramenta são discutidos adiante neste artigo.

O trabalho faz parte de um projeto maior que unifica pesquisas da biologia com a computação. Na parte da biologia, coletou-se amostras de diferentes eco-sistemas de cisternas em bromélias na área de conservação ambiental PRO-MATA no Rio Grande do

Sul[1]. Na parte da computação, recebeu-se os dados referentes as amostras de DNA já sequenciadas para posterior classificação taxonômica. A contribuição desse estudo consiste em uma análise de desempenho utilizando dois algoritmos diferentes de aprendizado de máquina. Baseou-se em conceitos de computação de alto desempenho (HPC) e computação paralela para habilitar parâmetros já implementados na biblioteca do Qiime2, a fim de acelerar a execução dos algoritmos de classificação. Como resultado, o estudo pretende investigar e selecionar o melhor algoritmo considerando o tempo de execução e consumo de recursos computacionais no Qiime2 para executar as sequências de DNA que serão obtidas pelos pesquisadores da biologia.

O trabalho está organizado da seguinte forma. A Seção 2 apresenta trabalhos semelhantes que serviram de inspiração para o desenvolvimento desse artigo. A Seção 3 apresenta o software usado nas análises e um pouco de seu funcionamento e ferramentas incluídas. Na Seção 4 são apresentados e discutidos os resultados das análises executadas. Por fim, na Seção 5 são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Já existem na literatura diversos estudos para otimização de programas de classificação taxonômica. O trabalho [Bokulich et al. 2018] apresenta um conjunto de experimentos testando diferentes algoritmos de classificação. Os resultados mostram que o classificador do Qiime2 que utiliza o Algoritmo Naive-Bayes tem os melhores resultados em acurácia e precisão das classificações. O estudo [Loff et al. 2018] investiga otimizações no MASA (Multi-Platform Architecture for Sequence Aligner) e providencias duas versões paralelas alternativas a versão oficial. Os experimentos mostram que o desempenho das novas paralelizações foi até 4% maior que a versão original.

O trabalho [Lu and Salzberg 2020] apresenta análises de desempenho, uso de recursos e qualidade dos resultados. Porém, a comparação ocorre entre diferentes ferramentas, ou seja, compara o Qiime2 com outras soluções como o Kraken2. Os resultados mostram que o Kraken2 é eficiente e está em constante evolução e otimizando as suas funções. Diferente desses trabalhos, o objetivo deste artigo é conduzir uma avaliação de desempenho com versões mais recentes do Qiime2 e com foco em aspectos de alto desempenho e computação paralela.

3. O Software Qiime2

O software Qiime2 é amplo e não trata-se apenas de uma solução para classificação taxonômica de sequências de DNA, mas também da limpeza e geração de diversos outros dados relacionados a amostras com sequências. Toda a etapa de análise após a coleta e sequenciamento de genomas pode ser realizado pelo Qiime2 já que a plataforma integra diferentes algoritmos do estado-da-arte. É possível utilizar algoritmos e ferramentas como o DADA2, que é utilizado na remoção de ruídos dos dados de amostra de DNA. Ademais, também existem opções de visualização gráfica em 3D de dados junto de outras ferramentas de visualização para facilitar e tornar viável uma análise mais profunda da amostra trabalhada. Todas as ferramentas descritas podem ser encontradas em [Bolyen et al. 2019]

O programa é implementado na linguagem de programação Python e utiliza diversas bibliotecas disponíveis nessa linguagem. Por exemplo, o Qiime2 utiliza a biblioteca scikit-learn, utilizada pelos classificadores taxonômicos baseados em aprendizado de máquina. O Qiime2 é uma extensão do Anaconda e é constantemente utilizado na literatura

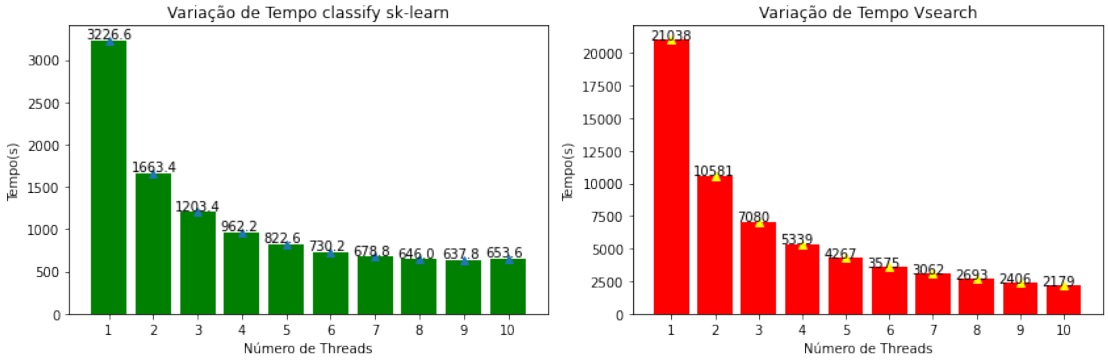
[1]https://www.pucrs.br/ima/pro-mata/

para classificações de amostras de meio-ambiente. O software é relevante para a pesquisa de microbiologia e traz excelentes resultados, além de apresentar uma interface simples e amigável, e que é fácil de instalar e utilizar.

O Qiime2 possui suporte para execução paralela já integrada em etapas de computação intensiva, que apresenta oportunidades de otimização. Neste trabalho, o objetivo é avaliar métodos de classificação, mas objetiva-se seguir com o estudo e aprimorar as estratégias paralelas existentes. Possíveis otimizações são referentes ao uso de memória, paralelismo com granularidade mais fina e suporte a execução em ambientes distribuídos.

4. Experimentos

Os experimentos foram executados em uma máquina com dois Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz, com 20 núcleos e 40 threads. Cada hyper-threaded core tem 64KB privada L1, 1MB privada L2 e 13.75MB de L3 compartilhada. A máquina possui 64 GB de RAM @ 2400 MHz e quatro HDD 3.5 @ 7200rpm usando SATA 3.1 com 6.0 Gb/s. O kernel é Linux 5.4.0-59-generic e usa OS Ubuntu 20.04.1 LTS. Foi utilizada a versão Qiime2-2021.8 com os parâmetros de classificação e limpeza de dados em múltiplas *threads* da plataforma, todas as partes precedentes à classificação taxonômica foram realizadas previamente e não foram avaliadas nesse artigo tal que buscamos uma comparação de desempenho entre dois algoritmos de classificação. Os experimentos foram executados cinco vezes e os valores de tempo de execução ilustrados nos gráficos representam a média dos tempos. Foram testados dois algoritmos de classificação: (1) O classificador Naive-bayes usa o algoritmo de classificação padrão da biblioteca do scikit-learn. (2) O classificador Vsearch utiliza o algoritmo de busca de sequências genômicas Vsearch para realizar a classificação taxonômica. A base de dados usada como referencia para a classificação de todos os experimentos realizados é o SILVA [Quast et al. 2012].



(a) Naive-Bayes **(b)** Vsearch

Figura 1. Resultado de desempenho dos algoritmos Naive-Bayes e Vsearch.

A Figura 1 ilustra os resultados do tempo de execução (em segundos) por número de *threads*. Os resultados mostram que ambos algoritmos escalam bem para graus de paralelismo maiores. No grau máximo de paralelismo, o tempo de execução do algoritmo Vsearch é $9,65$ vezes menor que o tempo sequencial. Contudo, o algoritmo Naive-Bayes escala somente até o grau 9 de paralelismo, obtendo um speed-up de $5,06$. Isso é explicado pela aumento linear no consumo de memória RAM por cada *thread* adicionada no sistema, levando ao uso de memória *swap*. Desta forma, utilizou-se apenas 10 *threads* ao

invés das 20 físicas disponíveis. Esse problema de ineficiência de memória também foi relatado e abordado por [Lu and Salzberg 2020].

Note que independente do número de *threads* utilizadas, o classificador Vsearch possui tempo de execução relativamente maior que o tempo de execução observado no algoritmo Naive-Bayes. Esse comportamento já havia sido observado por pesquisadores [Bokulich et al. 2018]. Além disso, os autores também destacaram a melhor qualidade do classificador Naive-Bayes em relação aos demais. Neste trabalho focou-se na análise de desempenho, mas como trabalho futuro, pretende-se considerar também a qualidade do resultado proporcionado pelos diferentes classificadores da literatura.

5. Conclusões

Esse trabalho apresentou uma análise de desempenho entre dois algoritmos de classificações taxonômicas usando o Qiime2. Os experimentos foram realizados com dados reais de amostras de DNA coletadas na natureza. Em suma para o caso de eficiência e velocidade o classificador Naive-Bayes leva uma grande vantagem especialmente em seu tempo de execução, tendo em vista que em sua execução sequencial o classificador Naive-Bayes demora até 6.5 vezes menos que o classificador Vsearch e mantém uma razão semelhante conforme o grau de paralelismo é aumentado. Além disso, o desempenho da versão paralela do Naive-Bayes é até $9,65$ vezes melhor que a versão paralela do Vsearch.

O tempo para obtenção de resultados que possam ser analisados é de grande importância em pesquisas, levando em consideração que o quanto maior for a velocidade da obtenção de resultados sobre um ecossistema, mais análises podem ser feitas sobre o mesmo ou outros ecossistemas conhecendo assim um pouco melhor a vida e o mundo em que vivemos por isso. Possíveis trabalhos futuros são a otimização do uso de recursos pelo classificador taxonômico, e buscar reduzir o seu alto custo de memória abordado em [Lu and Salzberg 2020]. Outra possibilidade é a comparação da qualidade dos resultados após a classificação taxonômica usando não apenas diferentes algoritmos mas também diferentes ferramentas como o Kraken2 e o BLAST mencionados previamente, buscando cada vez mais melhorar a qualidade de classificações taxonômicas em amostras ambientais.

Referências

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2's q2-feature-classifier plugin. *Microbiome*, 6(1):1–17.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, 37(8):852–857.

Loff, J., Griebler, D., Sandes, E., Melo, A., and Fernandes, L. G. (2018). Suporte ao Paralelismo Multi-Core com FastFlow e TBB em uma Aplicação de Alinhamento de Sequências de DNA. In *ERAD-RS*, page 2, Porto Alegre, BR. SBC.

Lu, J. and Salzberg, S. L. (2020). Ultrafast and accurate 16s rrna microbial community analysis using kraken 2. *Microbiome*, 8(1):1–11.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., et al. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596.