# Otimizações para Pipelines de Pós-processamento Metagenômico

**Raquel Dias, César Augusto F. De Rose**

Programa de Pós Graduação em Ciência da Computação - PUCRS

Av. Ipiranga 6681, Porto Alegre - RS, Brasil

raquel.dias.001@acad.pucrs.br, cesar.derose@pucrs.br

## Resumo

Dados obtidos de pesquisas em metagenômica proporcionam uma grande quantidade de informações sobre a estrutura, a organização e a origem do DNA de organismos a partir de amostras ambientais ou clínicas. Essas informações podem ser utilizadas em benefício do meio ambiente e da saúde humana. O pós-processamento em metagenômica é indispensável para a interpretação de resultados. A sua dependência por ferramentas computacionais tem se tornado cada vez mais crítica, devido ao seu crescimento exponencial na geração de dados de sequenciamento genético [1].

Algumas fases de filtragem e interpretação dos dados genéticos foram agrupadas em certas ferramentas computacionais, conhecidas como pipelines metagenômicos [2]. Entretanto, estas ferramentas ainda necessitam intervenção manual dos pesquisadores e há demandas por desempenho computacional em várias de suas etapas. O presente trabalho descreve o projeto de pesquisa que tem o objetivo de melhorar as metodologias de pós-processamento para os dados gerados a partir do sequenciamento metagenômico. São discutidos os principais objetivos do projeto, seus avanços e seu estado atual.

## Objetivos

Os principais objetivos do presente projeto consistem no desenvolvimento de otimizações para o pós-processamento, que seja adaptável aos pipelines metagenômicos existentes. Ao decorrer do projeto esse objetivo resultou no desenvolvimento do protótipo de uma nova arquitetura de pós-processamento.

## Proposta de uma Nova Arquitetura

Em busca de atender as demandas de desempenho computacional, confiabilidade dos resultados e automatização, foi proposta uma nova arquitetura de pós-processamento metagenômico. Esta arquitetura é capaz de gerar um consenso a partir da comparação entre várias metodologias de classificação de espécies. O consenso é obtido da comparação entre diferentes métodos de classificação, sendo um resultado mais confiável do que método tradicional (adotar apenas uma metodologia).

Um ponto importante na viabilidade do projeto é o desempenho computacional. Alguns dos métodos adotados são paralelos ou estão sendo paralelizados. O paralelismo ocorre tanto pelo uso de memória compartilhada (OpenMP), quanto pelo uso de troca de mensagens (MPI). O diagrama da arquitetura proposta é demonstrado na Figura 1, com ferramentas adotadas de outros autores (branco) e implementadas atualmente (cinza).

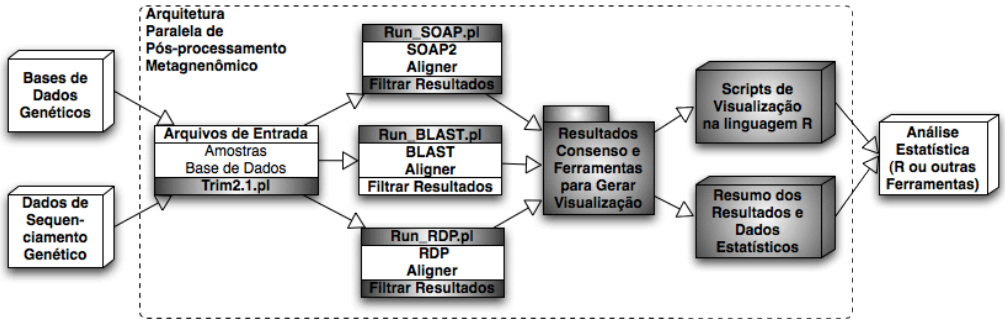


**Figura 1. Arquitetura de pós-processamento metagenômico**

A arquitetura foi construída sobre a estrutura do pipeline PANGEA [2]. PANGEA executa sequencialmente cada etapa de sua análise e usa apenas um método de classificação de espécies, também de forma sequencial. As funcionalidades da nova arquitetura estão classificadas com o seguinte conjunto de etapas:

(1) Reconhecimento e filtragem de dados de entrada. A sub-rotina Trim2.pl, modificada do programa PANGEA, faz a interpretação automatizada de novos formatos de entrada.

(2) Classificação de espécies: este passo está formado por três ferramentas de classificação [3][4][5]. Foram implementadas funções que convertem a saída do passo anterior e executam as rotinas de classificação de forma paralela.

(3) Classificação de espécies: nesta etapa é gerado um consenso a partir da comparação entre os métodos de classificação de espécies. O usuário pode definir os critérios para seleção dos resultados (correlação, qualidade, grau de ocorrência, etc.).

(4) Pós-processamento: após obter o consenso, os dados de saída são resumidos e convertidos em formatos para serem revisados pelo usuário. A revisão pode ser feita através de ferramentas de análise estatística, como o R Project Package [6].

Com as ferramentas de classificação paralelizadas, será possível obter resultados com uma qualidade maior em um período de tempo igual ou menor ao método tradicional utilizado (adotar apenas uma metodologia). O projeto está iniciando sua fase de avaliação de desempenho e validação. Para a validação os resultados serão comparados com dados experimentais disponíveis em bases de dados genéticos.

## Referências

[1] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz, "A bioinformatician's guide to metagenomics," *MMBR,* vol. 72, pp. 557-78, Dec 2008.

[2] A. Giongo, D. B. Crabb, A. G. Davis-Richardson, D. Chauliac *et al.*, "PANGEA: pipeline for analysis of next generation amplicons," *The ISME journal,* vol. 4, pp. 852-61, Jul 2010.

[3] M. D. Healy, "Using BLAST for performing sequence alignment," *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.],* vol. 6, Jan 2007.

[4] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang, "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics,* vol. 25, pp. 1966, 2009.

[5] J. R. Cole, Q. Wang, E. Cardenas, *et al.*, "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis," *Nuc. ac. res.,* vol. 37, pp. D141-5, Jan 2009.

[6] R Project for Statistical Computing. *www.r-project.org*. Acessado em 25/11/2011.