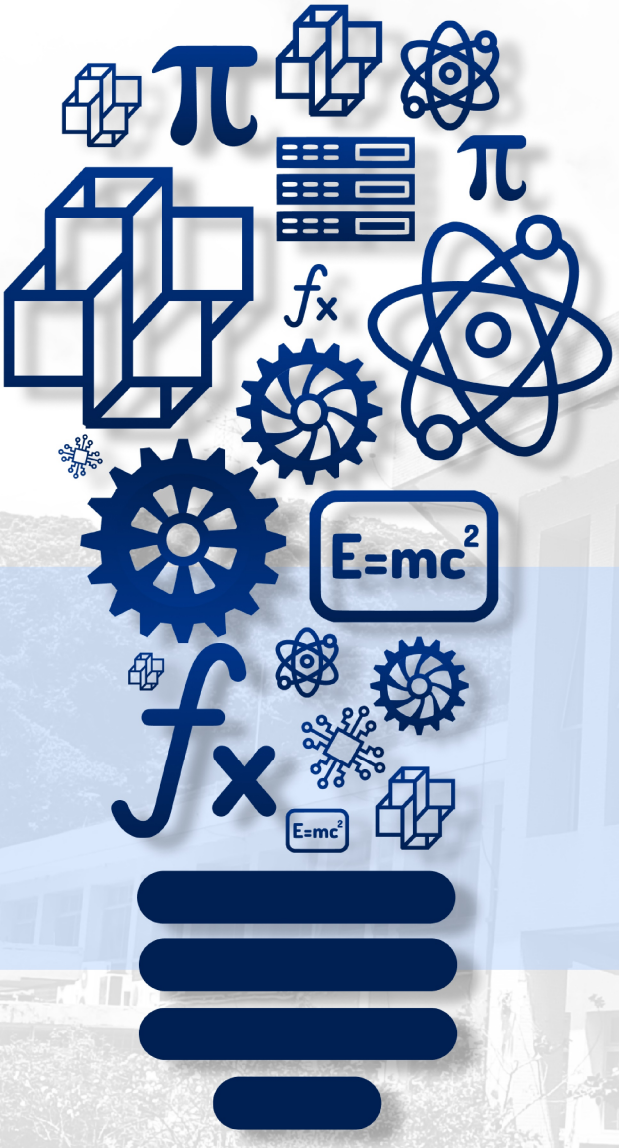
Jornada de Iniciação Científica e Tecnológica

2023



PIBIC · PIBITI

LNCC/MCTI

21 de agosto

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

LNCC
Laboratório Nacional de Computação Científica

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO

Jornada de Iniciação Científica e Tecnológica do LNCC

Petrópolis, 21 de agosto de 2023.

Laboratório Nacional de Computação Científica – LNCC

Diretor
Fabio Borges de Oliveira

Coordenação de Gestão e Administração - COGEA
Marcia Aparecida Almeida Pereira

Coordenação de Pós-Graduação e Aperfeiçoamento - COPGA
Sandra Mara Cardoso Malta

Programa Institucional de Bolsas de Iniciação Científica&
Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação
José Karam Filho

Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq

Presidente
Ricardo Magnus Osório Galvão

Coordenadora Geral do PIBIC/PIBITI
Lucimar Batista de Almeida

Jornada de Iniciação Científica e Tecnológica do LNCC

Comissão Interna do PIBIC/PIBITI-LNCC

José Karam Filho
Antônio Tadeu Azevedo Gomes
Eduardo Lucio Mendes Garcia
Fábio Lima Custódio
Jack Baczynski

Avaliadores Externos

Márcio Antônio de Andrade Bortoloti – UESB
Priscila Vanessa Zabala Capriles Goliatt UFJF

Apresentação

O LNCC realiza este ano a XX Edição da Jornada de Iniciação Científica e Tecnológica, que é um fórum de divulgação das pesquisas desenvolvidas no contexto dos Programas Institucionais de Bolsas de Iniciação Científica (PIBIC) e de Bolsas de Iniciação Tecnológica (PIBITI) fomentados pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). No período de setembro de 2022 a agosto de 2023, o PIBIC e PIBITI congregaram alunos de várias instituições de ensino e de diversas áreas do conhecimento. Este volume apresenta os resumos dos trabalhos desenvolvidos pelos bolsistas no período. Durante a Jornada, os trabalhos são apresentados pelos bolsistas oralmente e avaliados por um comitê científico externo.

Nesta XX Edição da Jornada, o Comitê Externo de Avaliação do PIBIC/PIBITI tem a seguinte composição:
Prof. Márcio Antônio de Andrade Bortoloti - UESB
Prof.ª. Priscila Vanessa Zabala Capriles Goliatt - UFJF

Destacamos o papel relevante do PIBIC/PIBITI do LNCC no desenvolvimento das pesquisas no LNCC e, principalmente, na formação complementar dos bolsistas, promovendo o aprimoramento do conhecimento, espírito criativo, reflexão crítica e ética. Estas características têm contribuído para suas inserções no mercado de trabalho e em programas de pós-graduação, como o PPG em Modelagem Computacional do LNCC. Este é o resultado do esforço e dedicação de todos os participantes.

Agradecimentos

Agradecemos ao CNPq pelas bolsas concedidas, à Direção do LNCC pelo apoio e à Comissão Interna do PIBIC e PIBITI no LNCC.

Agradecemos a disponibilidade e contribuição dos membros do Comitê Externo de Avaliação.

O sucesso desta Jornada, e do Programa como um todo, é o resultado da dedicação e do esforço de toda a comunidade do LNCC. Expressamos em particular nosso reconhecimento ao apoio concedido pela secretaria do PPG-LNCC, Sra. Roberta Machado e Sra. Ana Neri Fernandes e, especialmente, a Sra. Tatiane Ribeiro.

José Karam Filho
Coordenador do PIBIC/PIBITI - LNCC

Índice
Bolsistas PIBIC ativos

Previsão Meteorológica Utilizando Métodos de Inteligência Artificial
Bolsista: Julia Neumann Bastos
Orientadores: Fabio Andre Machado Porto e Rafael S. Pereira

Projeto e Implementação de Workflows Científicos Reprodutíveis de Alto Desempenho
Bolsista: Lucas da Cruz Silva
Orientadores: Luiz Manoel Rocha Gadelha Júnior, Carla Osthoff Ferreira de Barros e Kary Ann del Carmen Ocaña Gautherot

Modelagem e Integração de Bancos de Dados Relacionais na Arquitetura do Bioinfo-Portal
Bolsista: Marco Antônio Silva Cabral
Orientadores: Kary Ann del Carmen Ocaña Gautherot, Antônio Tadeu Azevedo Gomes e Marcelo Monteiro Galheigo

Programação Orientada a Objeto em um Método Numérico Escalável para escoamento Bifásico de Fluidos em Meios Porosos em Ambientes Computacionais de Alto Desempenho
Bolsista: Mariana Aguiar Ribeiro
Orientadores: Carla Osthoff Ferreira de Barros e Stiw Harrison Herrera Taipe

Inteligência Artificial Aplicada ao Diagnóstico por Imagem
Bolsista: Matheus Molina Alves Lima
Orientadores: Bruno Schulze e Fabio Lopes Licht

Implementação de Workflows Científicos de Biologia Computacional Reprodutíveis e Escaláveis de Alto Desempenho
Bolsista: Reiglan Soares Di Lourenço
Orientadores: Kary Ann del Carmen Ocaña Gautherot, Carla Osthoff Ferreira de Barros e Diego Moreira de Araujo Carvalho

Modelagem de Vigas de Euler Bernoulli e Timoshenko Equações Diferenciais Parciais
Bolsista: Tarsiane Ribeiro da Costa
Orientador: Jaime Edilberto Munoz Rivera

Técnicas de Ciência de Dados Aplicadas a Pesquisas de Dados em Larga Escala
Bolsista: Thiago Dutra da Silva
Orientador: Fábio André Machado Porto

Bolsistas PIBITI ativos

Estudo e Implementação de Sistema de Banco de Dados para Análise em Saúde Coletiva
Bolsista: Alan de Souza Mello
Orientadores: José Karam Filho e Paulo Cabral Filho

Uso do Padrão de Paralelismo de Linguagem em Esquema de Radiação da Atmosfera
Bolsista: Gabriel Thomaz do Nascimento
Orientadores: Roberto Pinto Souto e Eduardo Lucio Mendes Garcia

SEGMENTAÇÃO SEMÂNTICA PARA GERAÇÃO DE MAPAS DE COBERTURA E USO DO SOLO PARA OTERRITÓRIO DA APA PETRÓPOLIS

Bolsista:

Alan Daiki Suga

14 de junho de 2023

Relatório:
Período: 01/04/2023 a 14/06/2023

Orientador: Gilson Antonio Giraldi (gilson@lncc.br)
Co-Orientador: Paulo Sérgio Silva Rodrigues (psergio@fei.edu.br)

Área: Ciência da Computação: 1.03.00.00-7
Sub-área: Metodologia e Técnicas da Computação:
Processamento Gráfico (Graphics): 1.03.03.05-7

Laboratório Nacional de Computação Científica - LNCC/MCTIC

Bolsa de Iniciação Científica PIBIC-LNCC

Junho 2023

Resumo

A geração de um mapeamento espacial sobre o uso e a cobertura do solo é fundamental para fornecer subsídios necessários para tomada de decisões no que diz respeito ao correto manejo dos recursos naturais envolvidos. Cientes destas questões, os pesquisadores do Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) e do Laboratório Nacional de Computação Científica (LNCC), vêm desenvolvendo ações integradas para obter um levantamento atualizado das ações humanas referentes a utilização do solo e da distribuição espacial da vegetação da Área de Proteção Ambiental da região da cidade de Petrópolis, a qual compreende uma área de aproximadamente 68.000,00 hectares, no bioma da Mata Atlântica. Neste sentido, este projeto envolve atividades de pesquisa em processamento de imagens georreferenciadas extraídas da plataforma Google Earth para segmentação da cobertura e uso do solo da região de interesse. As atividades foram concentradas no estudo sobre os temas envolvidos, bem como estudo de bibliotecas para a geração de códigos para o projeto.

1 Introdução

Este projeto está inserido no contexto do Acordo de Cooperação número 6/2022, o qual tem por objetivo a cooperação técnica entre Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) e o Laboratorio Nacional de Computação Científica (LNCC), objetivando a cooperação mútua para que, através de ações integradas na obtenção, processamento, tratamento, classificação, modelagem, sistematização e disponibilização das informações de cobertura e uso do solo, sejam aprimorados os instrumentos de gestão da APA Petrópolis e do território no qual está inserida.

Especificamente, a atualização e aprimoramento dos mapas de cobertura e uso do solo da Área de Proteçao Ambiental da região da cidade de Petrópolis (APA Petrópolis) é uma demanda importante para o ICMBio porque foi identificado no Mapa de Zoneamento, estabelecido pelo Plano de Manejo da unidade, inúmeras inconsistências na caracterização da cobertura e uso do solo, tendo sido geradas classificações equivocadas, o que permite a utilização de áreas importantes à preservação e protege áreas sem atributos naturais. Estas inconsistências têm gerado dúvidas quanto à aplicação do Plano de Manejo da APA Petrópolis, tendo sido recomendada pelo ICMBio a sua revisão.

No contexto acima, a primeira fase do projeto é a aquisição de imagens aéreas da região de interesse para seu posterior processamento. Essa etapa foi resolvida com imagens RGB (Red-Green-Blue) extraídas da plataforma Google Earth, por meio da ferramenta ArcGis Pro. Um aluno de mestrado da Universidade Tecnológica Federal do Paraná (UTFPR) vem trabalhando com essas imagens.

As imagens exportadas têm resolução 2048 x 2048. Um subconjunto delas foi selecionado manualmente para ser anotado com o objetivo de gerar o padrão-ouro para treinamento de redes neurais que serão utilizadas para a segmentação automática de toda a região da APA Petrópolis, de acordo com as classes de interesse. As imagens selecionadas são importadas na ferramenta Computer Vision Annotation Tool (CVAT), a qual é uma ferramenta interativa de anotação de vídeo e imagem para visão computacional. As imagens foram então anotadas manualmente por meio da criação de polígonos ou da utilização da ferramenta *pincel* disponível no CVAT. Até o momento foram testadas redes neurais do tipo MPSegnet desenvolvidas pela equipe do projeto [1].

O presente plano de trabalho dará sequência as atividades já desenvolvidas. A experiência adquirida mostrou a necessidade de novas pesquisas nos seguintes temas:

1. Anotação semi-automática das imagens,
2. Desbalanceamento de classes,
3. Aumento de dados,
4. Testar novas arquiteturas de redes neurais.

Com relação ao item (1) trata-se de uma etapa tediosa e demorada, que precisa ser realizada de forma semi-automática para agilizar e melhorar padrão-ouro gerado. Por outro lado, é uma etapa fundamental para que possamos 'ensinar' as redes neurais, via processo de treinamento, a reconhecer automaticamente os padrões de interesse. Foram selecionadas oito classes de cobertura e uso do solo pela equipe do ICMBio. Em cada imagem RGB com 2048 x 2048 é necessário que um especialista identifique as regiões contendo cada uma das classes marcando as mesmas com cores distintas. Assim, é fundamental a pesquisa de técnicas de processamento de imagens e aprendizagem de máquina para diminuir o esforço manual desta etapa. Neste tema, encontramos na literatura abordagens baseadas em adaptação de domínio [2], utilização de dados sintéticos [3], anotações fracas ou esparsas [4, 5].

Outro ponto é o desbalanceamento de classes. Por exemplo, em geral, as imagens possuem muito mais regiões com floresta do que com solo exposto. Do ponto de vista do treinamento, isso implica que as redes neurais podem ter dificuldade para distinguir esse último padrão dos demais. Esse tipo de problema pode ser tratado com técnicas apropriadas de aumento de dados, como visto nos trabalho [6]. Outra possibilidade seria ponderar convenientemente os termos da função de perda, como realizado em [7], ou mesmo tentar diferentes funções de perda ou arquiteturas de redes que podem ser menos sensíveis ao desbalanceamento [8, 9].

Neste projeto vamos pesquisar e testar soluções para os temas (1)-(3), tendo em vista melhorar a acurácia das arquiteturas de rede a serem testadas no item 4. Esse é o objetivo geral deste projeto, como descrito na seção 2. Para atingir esse objetivo seguimos a metodologia apresentada na seção 3. As seções 4 e 5 apresentam os resultados e conclusões do trabalho desenvolvido pelo bolsista.

2 Objetivos

Considerado o exposto acima, o objetivo geral desta proposta é pesquisar estratégias mais eficientes para: (a) geração do padrão-ouro; (b) treinamento de redes neurais para segmentação de imagens RGB da região da APA Petrópolis. Para o item (a), serão tratadas questões envolvendo anotação semi-automática das imagens e para resolver o item (b) serão pesquisadas técnicas para balanceamento de classes via aumento de dados.

Assim, com relação aos objetivos específicos, temos:

1. Desenvolver metodologia de anotação semi-automática das imagens usando CVAT, super-pixel [10] e *support vector machine* (SVM) [11],
2. Implementar metodologia para aumento de dados direcionada para as classes com menor número de amostras,
3. Testar a eficiência dos itens acima no treinamento da rede neurais para segmentação semântica de imagens aéreas RGB da região de interesse.

3 Metodologia

Para atingir os objetivos descritos na seção 2, a primeira etapa do trabalho foi um estudo orientado sobre aprendizado de máquina, redes neurais e processamento de imagens. Isso foi feito usando a monografia [12]. Também foi realizado um curso online para o treinamento prático e aprofundado em relação às bibliotecas Numpy, Pandas, Matplotlib, Pytorch, Scikit-Learn e TensorFlow [13].

Os temas de redes neurais e SVM são descritos na monografia [12] que traz também referências clássicas e atuais para aprofundamento do estudo. Os estudos em processamento de imagens foram realizados usando a referência [14]. Em paralelo, o bolsista recebeu apoio das equipes envolvidas para aprender a implementação de redes neurais e SVM usando bibliotecas como TensorFlow, Pytorch e Scikit-learn.

O bolsista participou de reuniões remotas com a equipe da UTFPR para entender o procedimento para geração do padrão-ouro, o qual passa por duas etapas: extração das imagens aéreas e a anotação manual. Na primeira etapa, as imagens aéreas são extraídas do Google Earth por meio da ferramenta ArcGis Pro. Para isso, uma região do mapa é selecionada e a imagem é exportada em várias partes com tamanhos iguais de 2048 x 2048 pixels em formato RGB (Red-Green-Blue).

Dentre as imagens exportadas, algumas delas são selecionadas manualmente para serem anotadas na próxima etapa usando o CVAT. Uma amostra do padrão ouro pode ser visualizada na Figura 1. Especificamente, a imagem original é mostrada na Figura 1.(a), e os polígonos anotados manualmente contendo as diferentes classes de ocupação do solo, juntamente com regiões de sombra, são visualizados na Figura 1.(b). Finalmente, o padrão ouro gerado pode ser analisado na 1.(c). Cada cor utilizada corresponde a uma classe de acordo com a legenda da Figura 1. Pode-se notar que na Figura 1 temos uma grande área de agricultura e floresta. As demais classes constituem regiões menores na imagem o que gera o problema do desbalanceamento comentado anteriormente.

Outro aspecto muito importante será entender o pré-processamento realizado durante o treinamento das redes neurais no problema de interesse. Neste caso, cada imagem com resolução 2048 x 2048 é subdividida aleatoriamente em quadrados de tamanho 256 x 256. Esses quadrados são agrupados numa lista, denominada *mini-batch* cujo número de elementos é pré-definido pelo programador. O algoritmo de treinamento recebe como entrada um *mini-batch* a cada iteração, para montar a função objetivo (função de perda) que será processada por um método de otimização baseado no gradiente, computado em relaçao aos parâmetros internos da rede neural, denominados genericamente de *pesos* da rede.

(a) (b) (c)

Figura 1: (a) Imagem original extraída da plataforma Google Earth. (b) Polígonos indicando as diferentes classes observadas. (c) Resultado da segmentação manual realizada onde temos agricultura (em amarelo), floresta em verde, sombra em preto, solo exposto em marrom, área construída em vermelho e piscina em azul.

4 Resultados e Discussão

Como o trabalho realizado até o momento focou em estudo e aprofundamentos, o principal resultado obtido foi o conhecimento básico em processamento de imagens e sobre as bibliotecas relacionadas a aprendizado de máquina, com particular atenção ao Pytorch. Com esse objetivo atingido, será possível implementar código e seguir avançando nos estudos sobre as bibliorecas relevantes para o desenvolvimento deste projeto.

5 Conclusão

Considerando os objetivos deste projeto, foi concluída a primeira etapa de estudos da fundamentação teórica necessária. Os próximos passos seguirão o cronograma apresentado no projeto da bolsa, iniciando com a programação de redes neurais voltadas para a segmentação semântica. O bolsista tem realizado testes computacionais em seu computador pessoal (Intel Core i7-9750H, 16GB memória RAM, placa gráfica NVIDIA GeForce GTX 1650). Tem notado queda de desempenho em função do custo computacional das tarefas envolvidas. Isso demandará utilização de recursos de alto-desempenho, o que fará parte das tarefas seguintes.

Referências

[1] A. de Souza Brito, M. B. Vieira, M. L. S. C. de Andrade, R. Q. Feitosa, and G. A. Giraldi, "Combining max-pooling and wavelet pooling strategies for semantic image segmentation," *Expert Systems with Applications*, vol. 183, p. 115403, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421008253

[2] Y. Zhao, P. Guo, H. Gao, and X. Chen, "Depth-assisted residualgan for cross-domain aerial images semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2022.

[3] A. Kamilaris, C. van den Brink, and S. Karatsiolis, "Training deep learning models via synthetic data: Application in unmanned aerial vehicles," in *CAIP Workshops*, 2019.

[4] C. Fasana, S. Pasini, F. Milani, and P. Fraternali, "Weakly supervised object detection for remote sensing images: A survey," *Remote. Sens.*, vol. 14, p. 5362, 2022.

[5] Y. Hua, D. Marcos, L. Mou, X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, pp. 1–48, 2019.

[7] P. O. Bressan, J. M. Junior, J. A. Correa Martins, M. J. de Melo, D. N. Gonçalves, D. M. Freitas, A. P. Marques Ramos, M. T. Garcia Furuya, L. P. Osco, J. de Andrade Silva, Z. Luo, R. C. Garcia, L. Ma, J. Li, and W. N. Gonçalves, "Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, p. 102690, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243422000162

[8] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. J. Belongie, "Class-balanced loss based on effective number of samples," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, 2019.

[9] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2017.

[10] Y. Zhou, L. Wang, K. Jiang, L. Xue, F. An, B. Chen, and T. Yun, "Individual tree crown segmentation based on aerial image using superpixel and topological features," *Journal of Applied Remote Sensing*, vol. 14, no. 2, p. 022210, 2020. [Online]. Available: https://doi.org/10.1117/1.JRS.14.022210

[11] D. Zhang, F. Pan, B. Xing, Q. An, R. Wang, and D. Qi, "Research on the detection of the uav remote sensing chili images based on superpixel segmentation and svm," *2019 Chinese Control Conference (CCC)*, pp. 7828–7834, 2019.

[12] G. A. Giraldi, "Machine Learning and Pattern Recognition", LECTURE NOTES - COURSE GB-500 (http://virtual01.lncc.br/∼giraldi/NN-GB500/book.pdf)

[13] Curso online Data Science and Machine Learning Bootcamp https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/

[14] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Addison Wesley, 1992.

PROPOSTA DE PROJETO DE INICIAÇÃO CIENTÍFICA

BIOINFORMÁTICA, BANCO DE DADOS E ENGENHARIA DE COMPUTAÇÃO

Título do Projeto Proposto

Gerência de workflows escaláveis de redes genômicas de evolução para elucidar mecanismos evolutivos do Dengue no Brasil

Instituição

Laboratório Nacional de Computação Científica

Nome do Aluno

Albert Siqueira Cosme Emidio

Nome do Professor

D.Sc. Kary Ann del Carmen Ocaña Gautherot (Tecnologista Pleno – LABINFO/LNCC, Orientador)

D.Sc. Diego Moreira de Araujo Carvalho (Professor Associado – CEFET/RJ, Colaborador – LNCC, Coorientador)

D.Sc. Carla Osthoff Ferreira de Barros (Tecnologista Sênior – SEPAD/COTIC//LNCC, Coorientador)

Linha de Pesquisa

- Ciências Exatas e da Terra (1.00.00.00-3) – Ciência da Computação (1.03.00.00-7) – Metodologia e Técnicas da Computação (1.03.03.00-6) – Banco de Dados (1.03.03.03-0)
- Ciências Biológicas (2.00.00.00-6) – Biologia Geral (2.01.00.00-0) – Genética (2.02.00.00-5)

Palavras Chaves

Aprendizado de Máquinas, Bioinformática, Banco de Dados, Processamento de Alto Desempenho

Plano de Trabalho

Investimentos expressivos em pesquisas científicas e tecnológicas de bioinformática são necessários para atender às demandas crescentes do país como COVID-19 e as doenças tropicais negligenciadas (DTN), dentro delas a dengue que será alvo de pesquisa no presente projeto. Cientistas precisam de um grau de abstração, a fim de integrar eficientemente experimentos de bioinformática às várias tecnologias computacionais de sistemas de gerência, workflows científicos, aplicações de GPU, aprendizagem de máquina e profundo e arquiteturas de computação de alto desempenho (CAD), do supercomputador Santos Dumont (SDumont, https://sdumont.lncc.br/). O desafio em questão é processar, gerenciar e analisar grandes volumes de dados usando aplicações de bioinformática apoiadas por tecnologias e ambientes de CAD.

Workflows científicos de bioinformática são complexos e mais ainda se manipulam dados científicos voluminosos, heterogêneos ou multi-distribuídos [1] pelo que precisam de ambientes e tecnologias de CAD para execução, gerência, acesso e armazenamento de dados científicos e proveniência [2]. Os dados científicos resultantes de um determinado experimento científico [3] - do seu ciclo de vida, metadados de desempenho do workflow, atividades e programas - são diretrizes do bom desempenho de um workflow. Workflows científicos de bioinformática podem ser bem suportados por tecnologias e algoritmos de mineração de dados e aprendizado de máquinas [4-6] no apoio à análise dos dados resultantes, mas é uma gestão multitarefa e multidisciplinar ainda considerada um desafio em aberto na ciência de dados.

Muitas vezes para desenhar e usar um workflow é necessário compreender o problema biológico a ser desvendado, pois é a natureza desses dados que guiará a modelagem conceitual de um workflow e sua

adaptação na arquitetura computacional mais adequada. Um problema a ser atacado está relacionado ao tratamento, formatação e conversão de dados, motivado pela corrente diversificação de versões de programas usados pela comunidade científica nas atividades de um workflow. Dentre as bibliotecas mais usadas que oferecem alternativas tanto para a conversão e manipulação de formatos dos dados como para a gerência de aplicações de bioinformática estão BioPython (https://biopython.org/), BioConductor (https://www.bioconductor.org/), BioPerl (https://bioperl.org/) e mais recentemente o BioDocker. Essas bibliotecas oferecem suporte a ambientes de CAD e foram já acopladas no ambiente SDumont.

Os workflows científicos estão evoluindo simultaneamente para oferecer suporte à execução transparente e escalonável de uma variedade de análises. No entanto, integrar sistemas de workflows em ambientes de CAD pode ser desafiador, especialmente à medida que as análises se tornam mais interativas e dinâmicas, exigindo orquestração e gerenciamento sofisticados de aplicativos e dados, e personalização para ambientes de execução específicos. Parsl (Parallel Scripting Library, https://parsl-project.org/) é uma biblioteca Python para programar e executar fluxos de trabalho orientados a dados em paralelo, resolvendo esses problemas. Os desenvolvedores precisam anotar um script Python com diretivas Parsl envolvendo funções Python ou chamadas para aplicativos externos. Parsl gerencia a execução do script em clusters, nuvens, grades e outros recursos; orquestra a movimentação de dados necessária; e gerencia a execução de funções Python e aplicativos externos em paralelo. A biblioteca Parsl pode ser facilmente integrada a ambientes baseados em Python, permitindo gerenciamento e dimensionamento simples de fluxos de trabalho.

Nesse trabalho pretende-se explorar a gerência desses dados de uma forma organizada e eficiente em ambientes de CAD a fim de encontrar a melhor forma de armazenar e pesquisar dados científicos gerados em projetos de pesquisa na área de bioinformática. Respondendo à problemática de saúde relacionada à família Flaviviridae incluídos o dengue, febre amarela, e Zika, e considerando a problemática de infecciones reemergentes no Brasil, pretendemos realizar um estudo evolutivo do dengue, atendendo questões sobre sobre sua emergência e distribuição e mecanismos evolutivos e de recombinação.

A obtenção de dados anotados e de qualidade será realizado usando o pipeline chamado “Fast Loci Annotation of Viruses” (FLAVi; http://flavi-web.com/) que analizará genomas representativos dos Flavívirus disponíveis no GenBank. Dado o processo ser computacionalmente intensivo, FLAVi será acoplado no SDumont e a tecnologias de CAD como workflows, sistemas e linguagens e gerência e técnicas de aprendizado de máquina. Desta forma, o objetivo é propor uma série de soluções para a gerência de simulações computacionais e dados científicos de bioinformática requeridos pelo FLAVi. Se espera dar maior enfoque no tratamento e adequação de formatos de dados, que serão requeridos a posteriori por diferentes workflows de filogenia e redes filogenômicas, cada um requerendo um tipo de dados específico. Se viabilizará a adaptação do FLAVi no SDumont e se levantará uma proposta de integração com bibliotecas e containers tal que permita a interação entre a arquitetura do workflow com o SDumont. Se estudará a viabilidade da integração do workflow com o Parsl.

Na parte da implementação, o projeto terá como referência, os trabalhos previamente realizados e publicados pelo nosso grupo de pesquisa e colaboradores [7]–[9]. Feita a implementação da solução proposta, o passo seguinte será testar seu desempenho e avaliar suas funcionalidades frente aos benefícios propostos. Este projeto possui seis (6) etapas principais:

- Etapa 1: Revisar bibliografía sobre bioinformática, *workflows* científicos, Parsl, FLAVi e bibliotecas no SDumont;
- Etapa 2: Explorar tecnologias que explorem paralelismo e distribuição de tarefas em *workflows* de redes filogenômicas;
- Etapa 3: Implementar o esquema de execução e gerência de tarefas e dados para *workflows* de redes filogenômicas;
- Etapa 4: Explorar o ambiente de CAD do SDumont;
- Etapa 5: Análises de desempenho e escalabilidade das ferramentas propostas;
- Etapa 6: Elaboração de relatório final com descrição dos resultados.

Referências

[1] V. Marx, “Biology: The big challenges of big data,” *Nature*, vol. 498, no. 7453, pp. 255–260, Jun. 2013.

[2] J. Freire, D. Koop, and L. Moreau, Eds., *Provenance and Annotation of Data and Processes*, vol. 5272. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[3] M. Mattoso *et al.*, “Towards supporting the life cycle of large scale scientific experiments,” *International Journal of Business Process Integration and Management*, vol. 5, no. 1, pp. 79–92, 2010.

[4] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2. ed., At 7. printing. New York, NY: Springer, 2013.

[5] G. Da San Martino and A. Sperduti, “Mining Structured Data,” *IEEE Computational Intelligence Magazine*, vol. 5, no. 1, pp. 42–49, Feb. 2010.

[6] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, “Accomplishments and challenges in literature data mining for biology,” *Bioinformatics*, vol. 18, no. 12, pp. 1553–1561, Dec. 2002.

[7] Ocaña, K. A. C. S.; Galheigo, M.; Osthoff, C.; Gadelha, L. M. R.; Porto, F.; Gomes, A. T. A.; Oliveira, D.; Vasconcelos, A. T. BioinfoPortal: A scientific gateway for integrating bioinformatics applications on the Brazilian national high-performance computing network. Future Generation Computer Systems, v. 107, p. 192-214, 2020.

[8] M. Wilde, M. Hategan, J. M. Wozniak, B. Clifford, D. S. Katz, and I. Foster, “Swift: A language for distributed parallel scripting,” *Parallel Computing*, no. 37(9), pp. 633–652, 2011.

[9] L. M. R. Gadelha, M. Wilde, M. Mattoso, and I. Foster, “MTCProv: a practical provenance query framework for many-task scientific computing,” *Distrib Parallel Databases*, vol. 30, no. 5–6, pp. 351–370, 2012.

Programa Institucional de Bolsas de Iniciação Científica e Tecnológica
PIBIC/PIBITI

RELATÓRIO DE ATIVIDADES

1) Dados gerais

Título do projeto:
Identificação de padrões de problemas de performance e aceleração da execução de experimentos

Nome do bolsista:
Alexandre de Paiva Almeida

Nome do orientador:
Jauvane de Oliveira Cavalvante, PhD

Tipo de bolsa e período do relatório:
PIBITI – De 30 MAIO 2023 até 30 JUN 2023

2) Objetivos

Este projeto tem por objetivo o estudo, desenvolvimento e utilização da técnica de engenharia de software e ciência de dados para identificar padrões de bugs de software relacionados à performance. Os resultados serão úteis para acelerar a execução de experimentos científicos e simulações em vários tipos de projetos de desenvolvimento, inclusive os relacionados à natureza militar e de bioinformática.

3) Introdução

A aceleração da execução de experimentos científicos e simulações desempenha um papel fundamental no avanço da pesquisa em diversas áreas, incluindo a natureza militar e a bioinformática. No entanto, essas atividades computacionais muitas vezes enfrentam desafios relacionados à performance, resultando em processamentos lentos e ineficientes. A busca por soluções eficazes e economicamente viáveis para melhorar a velocidade de execução desses experimentos tem se mostrado uma tarefa complexa.

Neste contexto, este projeto de pesquisa tem como objetivo principal o estudo, desenvolvimento e utilização de técnicas de engenharia de software e ciência de dados para identificar padrões de bugs de software que afetam o desempenho. A identificação desses padrões permitirá uma abordagem mais eficiente na resolução dos gargalos de performance, possibilitando a escolha adequada dos recursos e estratégias necessários para otimizar o processamento e a simulação de experimentos.

Nesse contexto, a técnica de engenharia de software e ciência de dados combina princípios da engenharia de software com análise de dados e aprendizado de máquina. Essa abordagem multidisciplinar busca resolver problemas complexos de desenvolvimento e otimização de software, garantindo qualidade e eficiência. Ela utiliza ferramentas da engenharia de software para escalabilidade e técnicas de ciência de dados para extrair insights de grandes volumes de dados. Essa combinação é fundamental para identificar padrões, detectar bugs e melhorar a performance dos sistemas de computação.

A pesquisa proposta aborda a utilização da técnica de Mineração de Repositórios de Software (MSR) e Aprendizado de Máquina (IA) como meio de identificar os padrões de bugs relacionados à performance. Ao compreender a causa raiz dos problemas de performance, será possível direcionar os esforços de forma mais precisa e evitar soluções genéricas e onerosas, que podem não trazer os resultados desejados.

Além disso, este projeto também busca a formação de recursos humanos nas áreas de Mineração de Repositórios de Software, Aprendizado de Máquina e análise de gargalos de performance em simulações, softwares para ciência de dados e experimentos científicos. Pretende-se desenvolver uma metodologia e automação de scripts para a descoberta dos gargalos de performance, visando reduzir o tempo de processamento/simulação em pelo menos 10 vezes.

Espera-se que os resultados obtidos nessa pesquisa contribuam para a otimização da execução de experimentos científicos e simulações, possibilitando avanços significativos em diversas áreas do conhecimento. Os resultados também serão documentados em um relatório científico no formato de artigo acadêmico, com potencial para submissão em periódicos indexados, considerando a qualidade do trabalho e os resultados alcançados.

4) Material e Métodos ou Metodologia

Simulações, softwares para ciência de dados e experimentos científicos *in sílico* tem um processamento computacional muito lento, dadas as complexidades do mundo real, cuja verossimilhança é buscada ao máximo em tais sistemas. Para resolver estes problemas, os cientistas e centros de pesquisa lançam mão de recursos como supercomputadores, processamento via placas gráficas de alta performance (GPU), paralelismo de software, aumento de escala com computação em nuvem e otimizações computacionais nos programas. Porém, nem sempre tais soluções são eficientes ou justificam seu emprego, já que podem ser caras tanto do ponto de vista da compra dos recursos propriamente dito, quanto do ponto de vista da quantidade de horas necessárias para ajustar os experimentos e simulações a tais ambientes.

Diante deste dilema, uma das estratégias para empregar os recursos de forma eficiente e responsável é o estudo da causa raiz do gargalo de performance, o que pode ser feito através da implementação de probes e profiling ou através da identificação de padrões de bugs relacionados à performance. A pesquisa relacionada a este projeto tem por finalidade a utilização da técnica de Mineração de Repositórios de Software (MSR) e Aprendizado de Máquina (IA) para identificar padrões de bugs relacionados à performance. Uma vez identificado o padrão, o cientista pode atacar de forma mais eficiente os gargalos relacionados ao seu contexto, empregando o recurso mais apropriado para solucionar o problema da performance em vez de uma solução “força bruta” e que nem sempre trará o resultado desejado ou terá um custo muito alto para um benefício baixo.

5) Resultados e Discussão

Nesse mês de trabalho os resultados foram:

1. Formação de recursos humanos em mineração de repositórios de software (MSR), Aprendizado de Máquina (IA) e análise de gargalos de performance em simulações, softwares para ciência de dados e experimentos científicos.
2. Levantamento de projetos e funções para que as técnicas desenvolvidas, como o método de serialização e interpretador, pudessem ser aplicados com o objetivo de reduzir o tempo de processamento/simulação;
3. Definição de metodologia e automação de scripts para descoberta de gargalos de performance
4. Escrita de relatório de atividades

Os resultados apurados a partir da análise inicial mostraram relevante a automatização do procedimento para garantir que os experimentos sejam executados exatamente da mesma maneira e sob as mesmas condições para que a comparação seja precisa e justa. Para isso, está sendo desenvolvido um notebook Colab que comparará todas as versões do projeto com todos os métodos de serialização implementados (pickle, picklejson, simplejson e hdf5)

6) Conclusões

O projeto de pesquisa se mostra cada vez mais relevante no cenário atual e o progresso nesse mês de trabalho mostra que é possível fazer as análises esperadas

7) Referências bibliográficas

[1] SELAKOVIC, Marija; PRADEL, Michael. Performance issues and optimizations in javascript: an empirical study. In: Proceedings of the 38th International Conference on Software Engineering. 2016. p. 61-72.
[2] HU, Jingmei et al. Improving data scientist efficiency with provenance. In: 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE). IEEE, 2020. p. 1086-1097.
[3] PIMENTEL, João Felipe et al. A survey on collecting, managing, and analyzing provenance from scripts. ACM Computing Surveys (CSUR), v. 52, n. 3, p. 1-38, 2019.
[4] MORENO, Alexander; BALCH, Tucker. Speeding up large-scale financial recomputation with memoization. In: 2014 Seventh Workshop on High Performance Computational Finance. IEEE, 2014. p. 17-22.
[5] GUO, Philip J.; ENGLER, Dawson. Using automatic persistent memoization to facilitate data analysis scripting. In: Proceedings of the 2011 International Symposium on Software Testing and Analysis. 2011. p. 287-297.

Laboratório Nacional de Computação Científica

PIBIC/LNCC

CNPq

Comportamento Estrutural de Dutos Flexíveis no Transporte de Petróleo Offshore

Aluno: Alexandre Vitor Silva Braga - 201965501B

Professor orientador: Eduardo Lucio Mendes Garcia

Professores co-orientadores: Elson Magalhães Toledo / Marcos Vinicius Rodrigues

Juiz de Fora

Julho de 2023

Sumário

1. Resumo

O presente trabalho consiste no estudo e análise do comportamento estrutural de dutos flexíveis no transporte de petróleo offshore. Esses dutos são conhecidos também como risers e consistem em estruturas responsáveis pelo deslocamento do petróleo desde o leito marítimo até uma plataforma flutuante na superfície do oceano.

Desde o fim da década de 70, a utilização dos risers tem aumentado cada vez mais em escala global. No entanto, essa tecnologia ainda carece de estudos comportamentais mais aprofundados e precisos, justamente pela dificuldade de simular com exatidão o desempenho e os modos de falha de todas as camadas do riser e suas respostas a intempéries e outros estímulos do meio em que está inserido.

Desse modo, tendo em vista os desafios de validação dos modelos de dutos flexíveis, esse trabalho visa o estudo do modo de falha de fadiga nas armaduras de tração de aço e a posterior modelagem de uma secção transversal para validação.

2. Apresentação

Os risers são dutos suspensos utilizados para o transporte de petróleo desde estruturas subaquáticas até estruturas na superfície do oceano. Estão sujeitos, principalmente, a carregamentos dinâmicos, tais como tração, torção, flexão, etc.

A configuração da disposição dos dutos dependerá sempre de seu uso e de fatores como profundidade de operação, movimentação da estrutura flutuante, condições do ambiente e problemas de interferência. Algumas das configurações são chamadas de Steep-S(1), Lazy-S(2), Steep-wave(3), Lazy-wave(4) e Free-hanging(8).

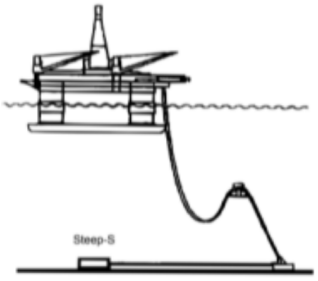


Figura 1. Steep-S [1]

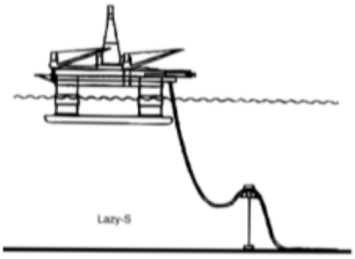


Figura 2. Lazy-S [1]

As seções críticas dos risers estão localizadas nas zonas de acoplamento nas estruturas flutuantes e no touchdown point, onde há grandes variações de tração (e grandes curvaturas), além da crista e da corcova de sua trajetória, onde há grandes curvaturas (e baixas trações) [1]. São justamente essas variações de tração que são responsáveis pela fadiga do componente.

As configurações "Lazy"são, muitas vezes, preferíveis em comparação às configurações

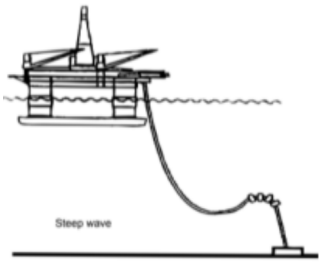


Figura 3. Steep Wave [1]

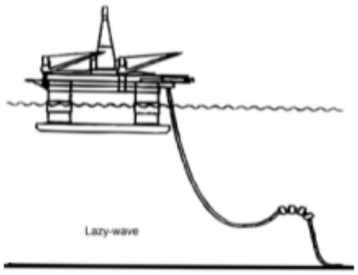


Figura 4. Lazy Wave [1]

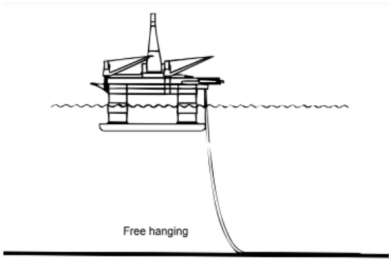


Figura 5. Free Hanging [1]

"Steep"pelo fato de aliviarem as trações no touchdown point (seção crítica inferior), ao introduzirem uma curva suave na extremidade do riser.

Dutos flexíveis são preferencialmente utilizados em relação a dutos rígidos, pois permitem uma conexão permanente entre o solo oceânico e estruturas flutuantes, sujeitas a movimentos translacionais e rotacionais causados por variações da maré, entre outros. Além disso, seu transporte e instalação são facilitados devido ao fato de serem fabricados e armazenados em rolos e bobinas, de fácil manuseio.

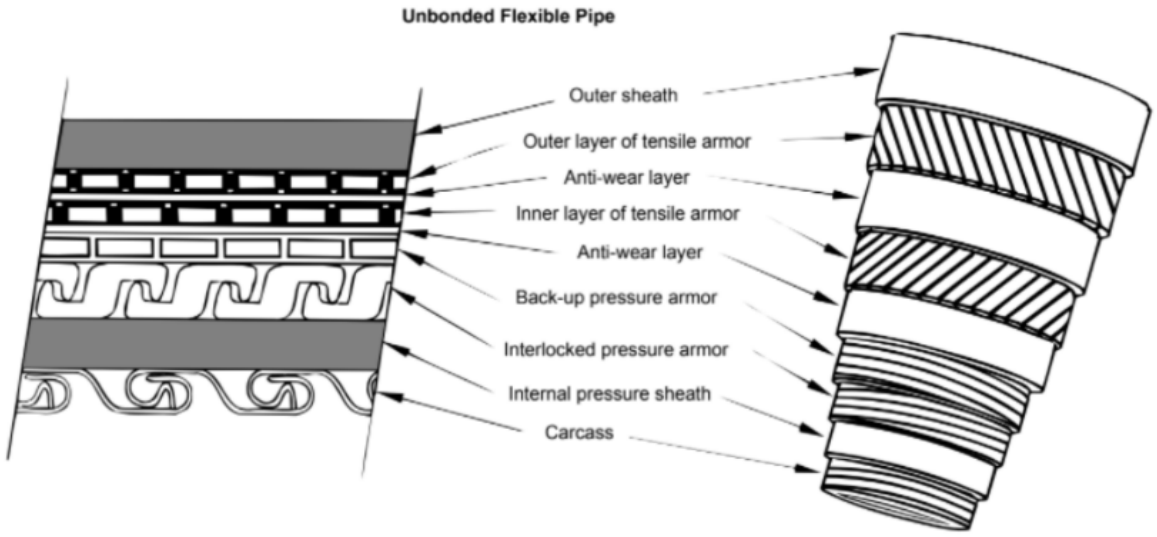


Figura 6. Esquema das secções transversais de um riser. [1]

Esses dutos flexíveis possuem uma estrutura complexa, de várias camadas, que visam garantir confiabilidade ao sistema de transporte e proteção contra carregamentos externos oriundos de intempéries. As camadas consistem em:

- Carcaça interna, responsável pelo revestimento interno protegendo o duto de entrar em colapso;

- Revestimento de pressão polimérico, que mantém o fluido dentro de seu calibre;
- Armaduras de pressão, que proveem resistência à compressão radial do riser;
- Armaduras de tração, oferecem resistência à tração. São duas camadas constituídas por arames de aço dispostos helicoidalmente em sentidos opostos, que oferecem rigidez ao duto;
- Revestimento polimérico externo, que impede a interação da água do mar com as armaduras internas de aço.

As armaduras de tração são intercaladas por camadas anti-desgaste que permitem que os arames da estrutura fiquem livres para o deslizamento relativo entre si e assim, fazem com que os estresses axiais provenientes de flexões possam ser liberados.

Neste trabalho será feita uma análise mais aprofundada nas armaduras de tração e nas suas falhas por fadiga. Para tanto, sua vida útil será calculada considerando os carregamentos cíclicos mais relevantes que agem sobre o riser.

3. Descrição de atividades

O presente trabalho consistiré no estudo do comportamento estrutural de dutos flexíveis com armaduras helicoidais e suas respostas quando expostas a carregamentos cíclicos de tração.

Baseando-se, respectivamente, nas obras de Svein Sævik e Nils Sødahl, as quais nos referiremos como Metodologia 1 e Metodologia 2 a partir de agora, faz-se uma distinção dos carregamentos presentes em duas categorias [3]:

- Carregamentos axissimétricos constituídos por tensões, torções e pressões internas e externas;
- Carregamentos por flexões.

Neste trabalho, um foco maior será dado à categoria de **carregamentos por flexões**.

3.1. Modelos para Carregamento de Flexão

Os mecanismos relevantes que podem levar ao desgaste, fadiga e colapso de componentes estruturais são diversos e podem ser atribuídos a vários fenômenos. Alguns dos mecanismos mais significativos incluem [2]:

- **Deslizamento de camadas:** Em sistemas estruturais compostos por várias camadas, o deslizamento entre as camadas pode ocorrer devido a forças de cisalhamento. Esse deslizamento pode levar a danos e desgaste progressivo nos materiais, comprometendo a integridade estrutural.
- **Mudança de ângulo de assentamento devido a rearranjo da estrutura:** Durante a operação ou sob certas condições de carregamento, a estrutura pode sofrer rearranjos ou ajustes que resultam em mudanças no ângulo de assentamento. Essas mudanças podem afetar a distribuição principalmente de **tensões** e a capacidade de suportar cargas, levando a problemas de fadiga e colapso.

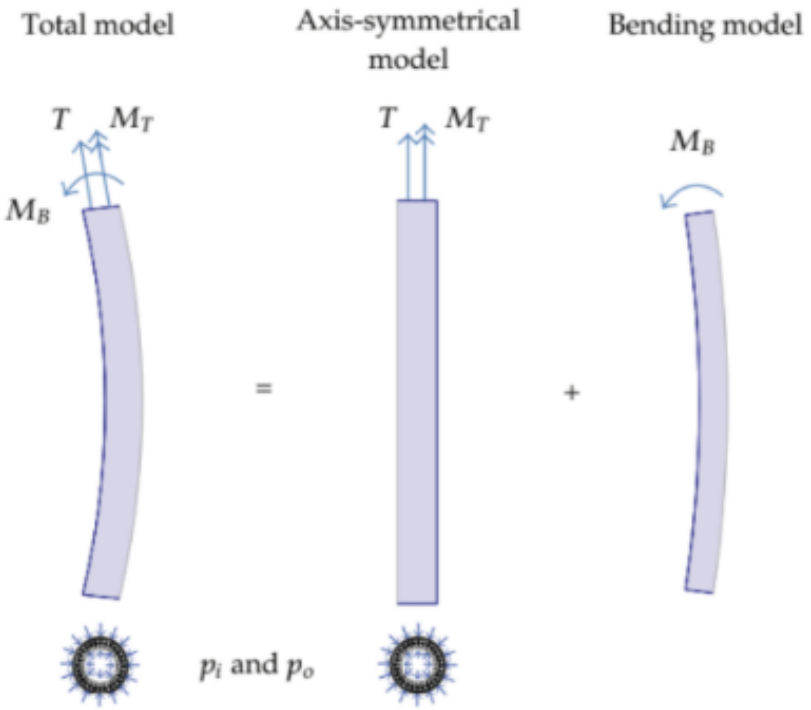


Figura 7. Modelos de resposta para análises transversais. [3]

- **Contato lateral das armaduras de tração:** Com o aumento da deformação e da curvatura K, as armaduras de tração podem chegar a uma situação de contato lateral. Esse contato lateral pode gerar tensões adicionais e acelerar os processos de fadiga e colapso.
- **Mudança de curvatura e ângulo de torção:** Variações na curvatura e no ângulo de torção ao longo da estrutura podem ocorrer devido a cargas cíclicas ou mudanças nas condições de operação. Essas variações podem levar a concentração de **tensões** em áreas específicas, resultando em fadiga e potencial colapso da estrutura.
- **Momento resistente no início da deformação devido a atrito interno:** Em alguns casos, a presença de atrito interno entre as camadas da estrutura pode levar a uma rigidez à flexão da tubulação. Esse momento resistente pode causar tensões elevadas e iniciar processos de fadiga prematura.

Portanto deverão ser estudados com base em [2] os seguinte fatores:

- Mecanismo de Deslizamento de Camadas
- Tensões induzidas por flexão
- Tensões induzidas por atrito
- Rigidez à flexão da tubulação

Esses mecanismos destacam a importância de considerar uma análise detalhada e abrangente das condições de carregamento e dos fenômenos envolvidos para garantir a integridade estru-

tural e a vida útil adequada dos componentes. A compreensão desses mecanismos é essencial para o projeto e a operação segura de estruturas sujeitas a esses fenômenos.

3.1.1. Mecanismo de Deslizamento entre Camadas

No contexto de um riser flexível, o deslizamento entre camadas refere-se ao movimento relativo que pode ocorrer entre as diferentes camadas que compõem a estrutura do riser durante o processo estático ou dinâmico de deformação por flexão. Um riser flexível é geralmente constituído por várias camadas, como revestimentos poliméricos, armaduras de tração e camadas de proteção.

Se o comportamento for dinâmico e houver um diferencial de pressão na tubulação, a variação de tensão pode resultar em fadiga nas armaduras de tração, e o atrito interno em desgaste das superfícies de contato. O rearranjo cilíndrico é a principal causa desse desgate mecânico.

Sob a ação do carregamento, as armaduras são tensionadas e se comportam elasticamente, tendendo a se rearranjarem na forma de linhas geodésicas sobre a superfície do tubo ou camada inferior [2].

A definição dessa geodésica pode ser feita segundo as teorias da geometria diferencial ou da análise funcional. Porém devido a resultarem em integrais de difícil solução, [2] propõe um procedimento simplificado, o qual será estudado.

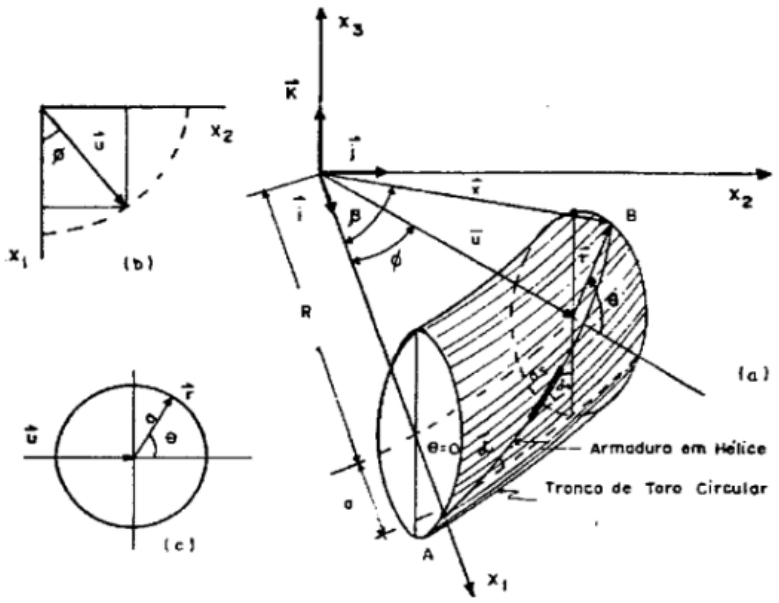


Figura 8. Hélice sobre a superfície de um toro (trecho de armadura hélica em uma tubulação flexível com curvatura circular). [2]

Referências

[1] American Petroleum Institute. Recommended Practice for Flexible Pipe: API RECOMMENDED PRACTICE 17B. Technical report, American Petroleum Institute, [s. l.], March 2002.

[2] J.R. Mendes de Souza. Análise numérica de risers flexíveis. Master's thesis, Instituto de Pesquisas Tecnológicas do Estado de São Paulo, São Paulo, Brasil, 1999.

[3] Geir Skeie, Nils Sødahl, and Oddrun Steinkjer. Efficient fatigue analysis of helix elements in umbilicals and flexible risers: Theory and applications. *Journal of Applied Mathematics*, 2012:1–23, April 11 2012.

Relatório de Atividades - PIBIC/LNCC

Título: **Modelagem Computacional para Análise de Dados de Tratamento de Câncer em Saúde Coletiva**

Bolsista PIBIC: **Ana Paula de Oliveira Souza**

Orientador: José Karam Filho
Coorientador: Paulo Cabral Filho

Período do relatório: 09/2022-01072023

1. Objetivos

Objetivo principal: Estudo e desenvolvimento de modelagem computacional para análise de dados de tratamento em saúde coletiva utilizando dados do DATASUS através de metodologia associada a um instrumento de extração e disseminação automática dos dados do SUS referentes a custos com câncer. Propõe-se analisar diversos aspectos desses gastos em certo período de anos, utilizando os dados disponibilizados pelo DATASUS, estudando sistemas de gerenciamento de banco de dados, de modo a criar uma metodologia que propicie, posteriormente, alimentar um sistema informacional amigável. Para materializar este trabalho, serão utilizados dados com os gastos em tratamento, pelo sistema público, de um tipo de câncer no Brasil.
Objetivos específicos: Adquirir e desenvolver conhecimentos de fundamentos de modelagem científica; **f**undamentos de estatística descritiva; sistema de banco de dados (MySQL e PostgreSQL); práticas de programação.

2. Introdução

O câncer é uma das principais causas de mortes no mundo, atingindo 11 milhões de indivíduos em 2020 [1], quase quatro vezes mais do que a pandemia de COVID-19 [2]. Os custos de tratamento são críticos, causando problemas econômicos e financeiros e pressionando o orçamento público em quase todos os países. Há previsões de que os gastos globais diretos com o tratamento oncológico devem quadriplicar entre 2011 e 2022, [3]. Além disso, as nações enfrentam perda de produtividade devido a mortes prematuras e afastamento do trabalho, aumentando a carga da doença para uma estimativa de US$ 1,16 trilhões, [2]. O câncer também representa uma parcela importante dos gastos em saúde no Brasil com custos diretos e indiretos com tratamento, mortes prematuras, faltas no trabalho e aposentadoria por invalidez, [3]. A estimativa do Instituto Nacional do Cancer (INCA) para a incidência de câncer no Brasil em 2022 foi de 625 mil novos casos, [4].

Apesar de todo o peso econômico e social do câncer, ainda são raros os estudos e pesquisas sobre o custo e a dinâmica de tratamento. A ausência de informações mais atualizadas impede um planejamento mais detalhado e eficiente do processo de gestão. O objetivo deste trabalho é gerar uma metodologia associada a um instrumento de extração e disseminação automático dos dados do SUS referentes ao custo mensal de radioterapia, quimioterapia e cirurgia por tipo de câncer.

Este trabalho propõe analisar diversos aspectos desses gastos em certo período de anos, utilizando os dados disponibilizados pelo DATASUS [5], estudando sistemas de gerenciamento de banco de dados, [6], de modo a criar uma metodologia que propicie, posteriormente, alimentar um sistema informacional amigável. Para

materializar este trabalho, serão utilizados dados com os gastos em tratamento, pelo sistema público, de um tipo de câncer no Brasil.

3. Metodologia

Para isso, a metodologia, neste projeto, segue a da modelagem computacional: observação/identificação; modelo fenomenológico; modelo matemático, metodologia de solução; obtenção de resultados; análise do modelo com retroalimentação; e seguiu as atividades descritas a seguir.

3.1. Fase 1:

Foi disponibilizada uma máquina no LNCC, com o sistema operacional Linux (Ubuntu) para a criação do banco com dados extraídos do Datasus.

Primeiramente, foi necessária uma pesquisa dos principais SGBDs (Sistema de Gerenciamento de Banco de Dados), para decidir qual se encaixaria melhor no projeto.

Foi concluído que o MySQL seria a melhor opção, tendo em vista ser gratuito e de código aberto; sua ampla compatibilidade com uma variedade de sistemas operacionais, incluindo Windows, Linux e macOS; também pelo seu desempenho, já que é projetado para lidar com altas cargas de trabalho e é capaz de lidar com grandes quantidades de dados e transações com baixa latência; sua escalabilidade, tanto horizontal como verticalmente, o que significa que é possível adicionar mais recursos ou máquinas para lidar com cargas de trabalho crescentes; por oferecer várias opções de segurança, incluindo autenticação de usuário, criptografia de dados e controle de acesso a dados; e também por ter uma ampla comunidade de usuários e desenvolvedores, o que significa que há muitos recursos e ferramentas disponíveis, incluindo documentação, fóruns de suporte e plug-ins.

Como os dados que serão transferidos para o banco de dados estão atualmente em arquivos .csv, foi necessário um estudo de como fazer a conversão de formatos, e ou transferência dos dados para uma tabela no banco. E como trata-se de um grande volume de dados e inúmeros arquivos, isso foi feito através de scripts para automatizar o processo de criação de banco de dados e tabelas, e também para as transferências dos dados.

Para estudar a elaboração de scripts no linux, a fonte foi um curso básico de bash (Curso básico de programação em Bash - 01 - Conceitos básicos - YouTube). Para o estudo de linguagem SQL via terminal, a fonte de pesquisa foi a documentação do MySQL (MySQL :: MySQL Documentation) e SQL Tutorial (w3schools.com)[7].

Foram feitos testes de transferências dos dados dos arquivos ‘.csv’ para tabelas no MySQL, e também foram testados alguns relacionamentos de tabelas, pois existem inúmeros campos da tabela principal que possuem tabelas associativa e, tanto os dados das tabelas associativas, como os dados de cada Estado, estão em arquivos ‘.csv’ distintos.

3.2. Fase 2:

Começou-se um estudo dos princípios básicos de modelagem, através da leitura de [8] e a introdução de [9], e foi dada continuidade às atividades anteriores de modelagem de banco de dados no que tange com a criação de novos campos: ipca, tipo de doença e estado.

Além da formatação das colunas VAL_SH,VAL_SP,VAL_TOT e VAL_UTI., com isso o banco foi remodelado.

Está-se trabalhando continuamente na documentação de todos os scripts a fim de registrar e explicar as funções de cada módulo para futuros grupos que venham a

utilizar este material. O padrão de documentação dos scripts está sendo seguindo conceitos de [10], contendo: um cabeçalho, campos localização (path) do interpretador, objetivo/descrição, versão, autor, data, licença e uso, etc.

4. Resultados

A primeira parte do banco de dados foi desenvolvida, com sucesso no teste de transferência de arquivo, com a consequente transferência dos dados para a tabela criada. O arquivo .csv transferido foi referente ao Estado do Amazonas, com 47 campos e aproximadamente 26 mil registros. Criamos a tabela 'CodSexo', com os campos 'codigo' (*Primary Key*) e 'nome', e foram feitas referências na coluna 'Sexo' da tabela principal, tornando esse campo *Foreign Key* e assim criando o relacionamento entre as tabelas. Com o sucesso destes testes, o ambiente tecnológico para a criação do banco de dados está bem avançado.

Com isso, foram consolidados conhecimentos de linguagem SQL através de linha de comando em ambientes linux, também em linha de comandos; conversões de arquivos .csv, Excel; e programação em scripts.

5. Conclusões:

Começou-se um estudo dos princípios básicos de modelagem, foram consolidados conhecimentos de linguagem SQL através de linha de comando em ambientes linux, também em linha de comandos; conversões de arquivos .csv, Excel; e programação em scripts. A primeira parte do banco de dados foi desenvolvida, com sucesso no teste de transferência de arquivo, com a consequente transferência dos dados para a tabela criada. Continuamente estão sendo desenvolvidos as documentações dos scripts elaborados.

6. Atividades Futuras:

- Continuar as atividades de sistema de banco de dados MysQL ;
- Realizar estudos de Métodos Numéricos e práticas de programação.
- Realizar estudos de Fundamentos de Estatística Descritiva;

7. Referências Bibliográficas

1. INCA. *Estimativa 2014 Incidência de Câncer no Brasil*. 1ª ed. Rio de Janeiro: Ministério da Saúde, 2014.
2. McGuire S. World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. *Adv Nutr*. 2016;7(2):418-419.
3. Associação da Indústria Farmacêutica de Pesquisa (Interfarma). Cancer in Brazil: The Patient's Journey in the Healthcare System and Its Social and Financial Impacts.; 2019. https://www.interfarma.org.br/app/uploads/2021/04/cancer-in-brazil-the-patient-5C-s-journey-in-the-healthcare-system-and-its-social-and-financial-impacts-interfarma.pdf
4. Revista Galileu, https://revistagalileu.globo.com/Ciencia/Saude/noticia/2022/04/42-dos-brasileiros-associam-o-cancer-sentimentos-negativos-e-morte.html.
5. BRASIL. DATASUS. Informações de Saúde (TABNET). [*S. l.*], 2021a. Available at: http://www2.datasus.gov.br/DATASUS/index.php?area=02. Acesso em: 19 ago. 2021.

6. Silberschatz, Abraham; Korth, Henry F.; Sudarshan, S. – Sistema de Banco de Dados – Editora Campus – Tradução Daniel Vieira – 5a. edição. Rio de Janeiro 2006.
7. w3schools.com. (s.d.). SQL Tutorial. Recuperado de https://www.w3schools.com/sql/ ;
8. Karam F., J., Princípios básicos de Modelagem, in Analista Cognitivo, S. Messeder e E. Cambuí orgs., EDUFBA, pp. 23-44, 2019.
9. Karam F., J. e Almeida, R. C., Introdução à Modelagem Matemática, Publicações da Pós-Graduação do LNCC, LNCC, 2003.
10. Villas, M.V. e Villasboas, L.F.P., Programação:conceitos,técnicas e linguagens, ed. Campos, 1987.

PROGRAMA DE BOLSAS PIBIC/LNCC

RELATÓRIO DE ATIVIDADES

1) Dados Gerais

Título do projeto: Assimilação de dados no modelo WRF para uso operacional em previsão de tempo e simulação
Bolsista: Gabriel Costa Chaves
Orientadores: Roberto Pinto Souto (LNCC); Juliana Aparecida Anochi (INPE), Helaine Cristina Moraes Furtado (UFOPA)
Tipo de bolsa: PIBIC/CNPq
Período do Relatório: Maio a Junho

2) Objetivos

Objetivos abordados durante o período da pesquisa:

- Implementar o WRF no Santos Dumont e no sistema computacional híbrido do Instituto Nacional de Pesquisas Espaciais (INPE):
  - WRF já instalado no SDumont.
- Ativar o módulo de assimilação de dados variacional do WRF (método 3DVar):
  - Em andamento para a preparação das rodadas do WRF para o domínio da região de interesse.

3) Introdução

Muitos problemas em geociências requerem a estimação ao longo do tempo do estado do sistema a partir de uma sequência de medidas ruidosas. Esses sistemas podem ser descritos por meio de modelos matemáticos, isto é, por um conjunto de equações diferenciais. No entanto, descrever fenômenos físicos por meio da modelagem matemática, é uma atividade passível de erro, uma vez que qualquer modelo é sempre uma aproximação imperfeita da realidade. Uma alternativa, para melhorar a descrição desses modelos é adicionar observações ao sistema. Essas observações consistem de dados medidos obtidos a partir de experimentos. A combinação dessas fontes de informação pode ser realizada de modo eficaz fazendo uso das ferramentas de assimilação de dados.

A descrição da maioria dos fenômenos físicos por meio de equações diferenciais envolve erros de modelagem. Para sistemas operacionais de previsão, uma estratégia para mitigar os erros de modelagem é adicionar alguma informação real do sistema físico ao modelo matemático. Esta informação adicional consiste de observações (valores medidos) do

fenômeno que se deseja modelar. No entanto, os dados observados devem ser inseridos apropriadamente, para evitar uma degradação no desempenho da previsão – "choque dos dados". Técnicas de assimilação de dados são ferramentas que combinam de modo eficaz observações com dados de modelos físico-matemáticos para a determinação da melhor condição inicial (análise), que é usado para executar o modelo de previsão. Esse processo é fundamental na prática operacional da previsão numérica do tempo (PNT).

O presente projeto tem por meta implantar um novo sistema de assimilação de dados baseado em redes neurais artificiais e árvores de decisão (algoritmos de aprendizagem de máquina) no modelo WRF (*Weather Research and Forecasting*) em meso-escala para a região Amazônica para a previsão numérica do tempo e a simulação de eventos ambientais na região Oeste do Pará. Este projeto propõe a ativação do método de assimilação de dados existente no sistema WRF e a implementação do método de assimilação de dados por aprendizagem de máquina.

Nesta primeira etapa do projeto, foram testados *scripts* de execução de uma instalação do WRF v4.2 no supercomputador Santos Dumont. É mostrado neste relatório,

4) Metodologia

Para cumprimento das atividades nos dois primeiros meses da pesquisa, fez-se necessário ter o entendimento da estruturação do ambiente no supercomputador Santos Dumont, para a execução do modelo WRF.

- Modelo WRF v4.2 já previamente compilado no supercomputador Santos Dumont
- Aquisição de dados de condição inicial (GFS) para o modelo WRF
- Preparação das rodadas de 48h de integração do WRF, para todos os dias de determinado ano.

Na continuação da pesquisa, serão realizadas tarefas a fim de obter redes neurais que irão emular a assimilação de dados, para reduzir o custo computacional da previsão, sem prejuízo da qualidade da análise. A seguir, as etapas da metodologia no prosseguimento da pesquisa:

- Aquisição de dados provenientes de estações meteorológicas do Banco de Dados Meteorológicos (BDMET) do Instituto Nacional de Meteorologia (INMET);
- Selecionar as variáveis: precipitação, temperatura máxima e mínima, evaporação, umidade relativa e temperatura média no período em que o modelo for executado;
- Coletar dados ambientais obtidos por equipamentos em solo, sondagens por balões, instrumentos de satélite (*Total Ozone Mapping Spectrometer* (TOMS), *Ozone Monitoring Instrument* (OMI), *Microwave Limb Sounder* (MLS)), dados de reanálise do *European Centre for Medium-Range Weather Forecast* (ECMWF – ERA - *Interim daily*) entre outros serão utilizados a fim de validar as simulações realizadas pelo modelo WRF;
- Analisar e interpretar dados meteorológicos;
- Produzir rotinas computacionais para seleção automática das séries temporais a serem visualizadas.
- Implementar uma rede neural artificial como uma técnica de assimilação de dados aplicada ao modelo WRF.

5) Resultados e Discussão

No decorrer dos meses de maio e junho, o bolsista participou das reuniões semanais com todos os integrantes do projeto para exposição do que foi realizado e definição de direcionamentos futuros. As atividades desenvolvidas neste período estão descritas a seguir.

Acesso ao supercomputador Santos Dumont

Para o desenvolvimento das atividades práticas, que consistem na execução do modelo de WRF, foi possível após a autorização do acesso ao supercomputador Santos Dumont. Neste sentido, o bolsista teve acesso às instruções de como acessar a máquina remotamente via VPN, acesso ao supercomputador por meio de uma ferramenta de acesso *shell* remoto chamada *ssh*.

Aquisição de dados de condição inicial e de contorno para o modelo WRF

Foi baixado um recorte de dados GFS na sua pasta reservada, por meio do seguinte comando:

```
curl -O
https://request.rda.ucar.edu/dsrqst/CHAVES642582/TarFiles/gfs.0p25.2023042800-25
.2023042818.f072.grib2.tar
```

E, para descompactar o arquivo descarregado, o bolsista usou o seguinte comando:

```
tar -xf gfs.0p25.2023042800-25.2023042818.f072.grib2.tar
```

Execução de rodadas do modelo WRF

A tarefa seguinte foi executar o modelo de previsão WRF usando um recorte de dados de teste, o que foi realizado exitosamente. Isso consistia em, primeiramente, executar rotinas de configuração do modelo e, então, executar as rodadas;

Para a execução dessa tarefa, foi acessado o diretório reservado para trabalho com dados GFS, cujo caminho absoluto é: `/scratch/g-assimila/gabriel.chaves`. Para acessá-lo, usa-se o seguinte comando:

```
cd $SCRATCH
```

À esse diretório, reservado ao usuário, foi colado um diretório contendo as rotinas de configuração e criação de rodadas. Estando na pasta reservada supracitada, o comando usado para copiar o diretório (que implica no uso do caminho do mesmo) foi este:

```
cp -r /scratch/g-assimila/sdbase/wrf-model/exemplo_wrf_rj .
```

Feito isso, foi executada a rotina de configuração e criação das rodadas, usando este comando:

```
./tudo_gassimila.sh 2018
```

tudo_gassimila.sh

```
#!/bin/bash

ano=${1}

./criardiretorios.sh ${ano}
./criarrodadas.sh ${ano}
./criarnamelist.sh ${ano}
./criarlinkgrib.sh ${ano}
```

Ao executar esse comando, as rodadas foram criadas na pasta 2018, agrupadas por meses, e por dias. Tais agrupamentos foram feitos por diretórios; diretórios para meses e diretórios para dias, estes dentro daqueles que lhes são respectivos.

Foi, então, executado o WRF, na rodada do dia primeiro de junho. Para isso, primeiramente foi acessado o diretório dessa rodada:

```
cd /scratch/g-assimila/gabriel.chaves/exemplo_wrf_rj/2018/JUN/01/rodada
```

Então, foi executada a rotina para iniciar o processamento dos dados, loteando-a para um escalonador de processos chamado Slurm:

```
sbatch ./submitjob.sh
```

submitjob.sh

```
#!/bin/bash
#SBATCH --nodes=1                       #Numero de Nós
#SBATCH --ntasks=1                     #Numero total de tarefas MPI
#SBATCH -p cpu_dev                      #Fila (partition) a ser utilizada
#SBATCH -J WRF       #Nome job

module load cmake/3.17.3
module load python/3.8.2
module load openmpi/gnu/4.1.2_gcc-9.3_ucx_1.12+cuda-11.12
. /scratch/g-assimila/sdbase/spack/v0.17.1/share/spack/setup-env.sh
export SPACK_USER_CONFIG_PATH=/scratch/g-assimila/sdbase/spack/.config/v0.17.1

spack load --only dependencies wrf@4.2%gcc@9.3.0

export LD_LIBRARY_PATH=/scratch/g-assimila/sdbase/usr/lib64:$LD_LIBRARY_PATH

export WPS_DIR=$SLURM_SUBMIT_DIR/WPS
export WRF_DIR=$SLURM_SUBMIT_DIR/WRF/run

cd $WPS_DIR
./geogrid.exe
./ungrib.exe
./metgrid.exe
```

```
#rm GRIBFILE* && rm FILE*

cd $WRF_DIR
mv $WPS_DIR/met_em* .
export OMPI_MCA_mpi_cuda_support=0
srun -n 1 ./real.exe
srun -n $SLURM_NTASKS ./wrf.exe
#rm met_em*
#gzip rsl.*
```

Foram gerados resultados na pasta '2018/JUN/01/rodada/WRF/run/'

```
2018/JUN/01/
└── rodada
    ├── slurm-10893601.out
    ├── slurm-10893607.out
    ├── slurm-10893660.out
    ├── submitjob.sh
    ├── verificatempojob.sh
    ├── WPS
    │   ├── geogrid.exe ->
/scratch/g-assimila/sdbase/wrf-model/v4.2/WPS/geogrid/src/geogrid.exe
    │   ├── metgrid.exe ->
/scratch/g-assimila/sdbase/wrf-model/v4.2/WPS/metgrid/src/metgrid.exe
    │   ├── ungrib.exe ->
/scratch/g-assimila/sdbase/wrf-model/v4.2/WPS/ungrib/src/ungrib.exe
    └── WRF
        └── run
            ├── namelist.input
            ├── ndown.exe ->
/scratch/g-assimila/sdbase/wrf-model/v4.2/WRF/run/ndown.exe
            ├── real.exe ->
/scratch/g-assimila/sdbase/wrf-model/v4.2/WRF/run/real.exe
            ├── tc.exe ->
/scratch/g-assimila/sdbase/wrf-model/v4.2/WRF/run/tc.exe
            ├── wrfbdy_d01
            ├── wrf.exe ->
/scratch/g-assimila/sdbase/wrf-model/v4.2/WRF/run/wrf.exe
            ├── wrfinput_d01
            ├── wrfout_d01_2018-06-01_00:00:00
            ├── wrfout_d01_2018-06-02_00:00:00
            └── wrfout_d01_2018-06-03_00:00:00
```

Resultados obtidos e esperados

- Foi realizada com êxito uma rodada de teste do WRF no supercomputador Santos Dumont.
- Estes primeiros passos foram importantes para a familiarização com o ambiente computacional, com o modo de submissão de jobs, e também com o modelo WRF.

- Os passos seguintes na pesquisa consistem em executar rodadas paralelas do WRF em umas das regiões de interesse do projeto, na região norte do país, abrangendo o estado do Pará.

Espera-se ainda, no decorrer do andamento do trabalho, chegar-se aos seguintes resultados finais:

- Novo método de assimilação de dados para o modelo WRF, baseado em aprendizado de máquina, para a região amazônica;
- Os algoritmos de aprendizado de máquina propostos devem garantir a qualidade da análise e apresentar grande ganho no desempenho computacional;
- Uso da metodologia de assimilação por aprendizado de máquina em previsão operacional para a região de interesse.

6) Conclusões

Neste período o bolsista pôde ter contato com o ambiente virtual do supercomputador Santos Dumont e realizar a execução do processamento de algumas rodadas com o modelo WRF em alguns recortes de dados presentes na máquina. O cumprimento desta etapa foi importante como aprendizado para a preparação das rodadas do modelo aos domínios de interesse neste projeto, tais como, a região norte do país, mais especificamente o estado do Pará.

7) Referências Bibliográficas

[1] E. Kalnay, Atmospheric modeling, data assimilation, and predictability, 2nd ed. New York: Cambridge university press, 2003.
[2] Ismail-Zadeh, A. and Tackley, P. J. (2010). Computational Methods for Geodynamics. Cambridge University Press, United States of America, first edition.
[3] Skamarock, W. C., Klemp, J. B., J., D., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D. M., and Huang, X.-Y. (2019). A description of the advanced research wrf version 4. Technical report, NCAR Technical Notes -NCAR/TN-556+STR, Colorado. USA. DOI: 10.5065/1dfh-6p97.

RELATÓRIO DE ATIVIDADES

Título do Projeto: Avaliação da aplicação BEAST no Ambiente do Supercomputador SDumont

Nome do bolsista:

Guilherme Freire da Silva Dornelas

Nome do orientador:

D.Sc. Carla Osthoff Ferreira de Barros (Tecnologista Sênior – CENAPAD/LNCC)

Nome do coorientador:

D.Sc. Kary Ann del Carmen Ocaña Gautherot (Tecnologista Pleno – LABINFO/LNCC)

B.Sc Micaella Coelho Valente de Paula (Analista de TI – SEPAD/COTIC/LNCC)

Tipo de bolsa: PIBIC

Período do relatório: 01/09/2022 a 03/07/2023

INTRODUÇÃO

O Portal de Bioinformática, Bioinfo-Portal (https://bioinfo.lncc.br/), é um ambiente computacional que gerencia a execução de aplicações e dados científicos de bioinformática em larga escala, e serve de apoio às pesquisas da comunidade científica de informática através de uma interface Web amigável e interativa. É gerenciado pelo Laboratório Nacional de Computação Científica, LNCC (https://lncc.br) e usa recursos e tecnologias de computação de alto desempenho, do Supercomputador Santos Dumont, SDumont (https://sdumont.lncc.br).

Este projeto faz parte da colaboração entre o CENAPAD e o LABINFO, ambos pertencentes ao LNCC, onde são realizadas diversas pesquisas no apoio às análises computacionais envolvendo bioinformática, biologia computacional e CAD.

A aplicação BEAST 1.10 passou por diversas análises junto ao ambiente computacional do SDumont. A disponibilidade de dados de sequência filogenômica motivou os experimentos, buscando melhorias adicionais para a eficiência computacional. Os dados foram executados no sistema de muitos núcleos, do supercomputador, esse sistema oferece um poder de computação muito maior. Nas seções a seguir, apresentaremos os parâmetros utilizados, dados e condições do ambiente para os experimentos. Analisamos o desempenho do BEAST 1.10 no SDumont, com a finalidade de garantir o uso eficiente dos recursos computacionais.

OBJETIVOS

- O projeto visa desenvolver estudos de desempenho das aplicações do Portal de Bioinformática (Bioinfo-Portal), de forma a obter a melhor alocação de recursos computacionais, segundo os parâmetros de entrada dos usuários e gerar scripts de submissão de *jobs* otimizados para o Portal.

- Analisar o desempenho das aplicações de filogenômica e evolução molecular computacional em GPU no ambiente de computação de alto desempenho (CAD), mais especificamente nos recursos computacionais do SDumont.
- Estabelecer uma análise comparativa do BEAST 1.10 acoplado a BEAGLE 3, nos ambientes CPU, GPU e ambiente híbrido executando em paralelo CPU/GPU.
- Realizar análises de desempenho e escalabilidade, visando otimizar o uso do ambiente computacional.

METODOLOGIA

O projeto está realizando a integração de ferramentas de bioinformática em clusters de supercomputadores nos sistemas de CPU e GPU.

O BEAST 1.10 é um programa de análise filogenômica baseada em inferência Bayesiana, multiplataforma de sequências moleculares utilizando os métodos de *Markov Chain Monte Carlo* (MCMC). Em estatística, os métodos MCMC compreendem uma classe de algoritmos para a amostragem de uma distribuição de probabilidade.

A BEAGLE 3 é uma biblioteca de alto desempenho, que faz uso de processadores altamente paralelos, como aqueles em placas gráficas (GPU). Melhorando o uso de *software* filogenéticos em clusters. A biblioteca está acoplada ao programa BEAST 1.10 para tornar mais eficiente a paralelização em escala fina de cálculos de probabilidade filogenética. Oferecer suporte à computação simultânea de matrizes de probabilidades, para aumentar o desempenho das análises de modelos de nucleotídeos com maior flexibilidade de particionamento de dados.

O BEAST 1.10 acoplado a BEAGLE 3 fazem parte do Bioinfo-Portal, do SDumont. O *software* foi executado em diferentes cenários, nas filas CPU, GPU e ambiente híbrido executando em paralelo CPU/GPU, levando em consideração as características e natureza dos dados e parametrização das aplicações. A combinação do BEAST 1.10 com a BEAGLE 3 [Ayres et al. 2019] permite paralelizar várias partições de dados em um único dispositivo de alto desempenho (GPU) para usar a capacidade total desses dispositivos, reduzindo as sobrecargas computacionais.

Nos experimentos fizemos uso de quatro conjuntos de dados filogenômicos, do vírus da Dengue (DENV), dados do diretório benchmark do BEAST 1.10 (Bench1 e Bench2) e do vírus da febre amarela (YFV), ambos os dados em formato (XML). Para os cálculos foi usado como variabilidade o parâmetro *chainLength* fixado em "100000" e "20000000", o incremento dos valores fornece consistência as análises bayesianas, mostrando-se proporcional ao tempo computacional. O parâmetro *chainLength* no BEAST 1.10, especifica o número de etapas que a cadeia MCMC fará antes de terminar. Esse número depende do tamanho do conjunto de dados, da complexidade do modelo e da precisão da resposta necessária.

As execuções foram realizadas usando 1 nó computacional (24 *threads*), compostos por duas CPUs *Ivy Bridge Intel Xeon E5 2695v2 (12cores @2.4GHz)* e 64 GB de memória RAM. Alternando os (24 *threads*) com 1CPU, 1 GPU, 1 CPU/1GPU, conforme a (Figura 2) e (Figura 3). Cada dado passou por cinco repetições junto ao ambiente do SDumont, obtendo a média total das execuções dos *jobs* no Santos Dumont.

RESULTADOS

Nesta seção apresentaremos a avaliação computacional do BEAST 1.10 e BEAGLE 3. Na (Figura 1) apresentamos a interface *web* do Bioinfo-Portal selecionando a aplicação BEAST 1.10. O uso do portal é prático e interativo junto as aplicações de Bioinformática e workflows científicos.

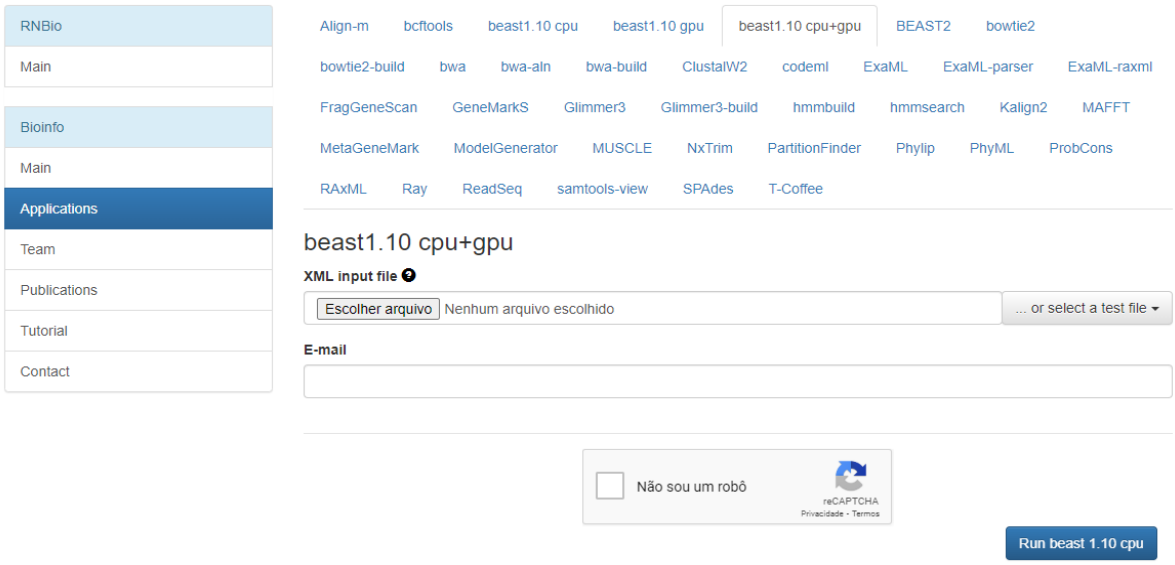


Figura 1. Aplicação BEAST no Bioinfo-Portal.

O conjunto de dados DENV [Ayres et al. 2019] é composto por 997 genomas abrangendo a diversidade global da dengue e 6.869 padrões de locais únicos em 10 subconjuntos baseados em genes. Os dados do Benchmark 1 (Bench1) é composto por 1.441 táxons de um alinhamento humano com 987 padrões de sítio únicos. Os dados do Benchmark 2 (Bench2) é composto por 62 táxons de um alinhamento de mamíferos com 5.565 padrões únicos de sítios.

A (Tabela 1) apresenta as principais características dos conjuntos de dados e parâmetros apresentados nos arquivos (XML) utilizado pelo BEAST 1.10. As duas últimas colunas referem-se a considerações (também sugeridas pela equipe do BEAST) de uso de recursos computacionais com base no tipo de conjuntos de dados e estudo evolutivo.

Conjunto de dados	Taxa número*	Alinhamento sites**	Alinhamento partição	Dados e sugestões recursos	Estudo evolutivo considerações de tipo
DENV - Genoma	997	10,188	10	Grandes dados; muitos taxa (centenas); GPU	Taxas de evolução e filogenético relacionamentos para carimbo de data/hora sequências
Bench1 - Genoma	1441	98	1	Dados pequenos; muitos taxa (centenas); multi-threads	Benchmark para tempo medido filogenias
Bench2 - Genoma	62	10,869	1	Grandes dados; alguns taxa (dezenas); GPU	Benchmark para tempo medido filogenias
YFV - Gene	71	654	1	Dados médios; alguns taxa (dezenas); multi-threads	Filogenética relacionamentos de sequências particionadas

Tabela 1. Recursos de dados genômicos e configuração de arquivo XML

Realizamos os experimentos com o *software* BEAST 1.10, utilizando nas análises 4 modelos de dados, sendo do vírus da Dengue (DENV), Bench1, Bench2 e vírus da febre amarela (YFV), todos em (XML), conforme a (Figura 2), apresenta essas execuções. Fixamos o parâmetro *chainLength* em

"100000", selecionamos para todos os testes 1 nó computacional (24 *threads*), alternando com 1CPU, 1GPU e 1CPU/1GPU. Os melhores desempenhos do Tempo Total de Execução (TTE) foi obtido usando o ambiente híbrido executando em paralelo 1CPU/1 GPU. Com exceção do arquivo DENV, pois nessa execução o melhor desempenho foi obtido com o uso de CPU, isso mostra que o tamanho dos dados e o *chainLength* podem interferir no resultado da execução.

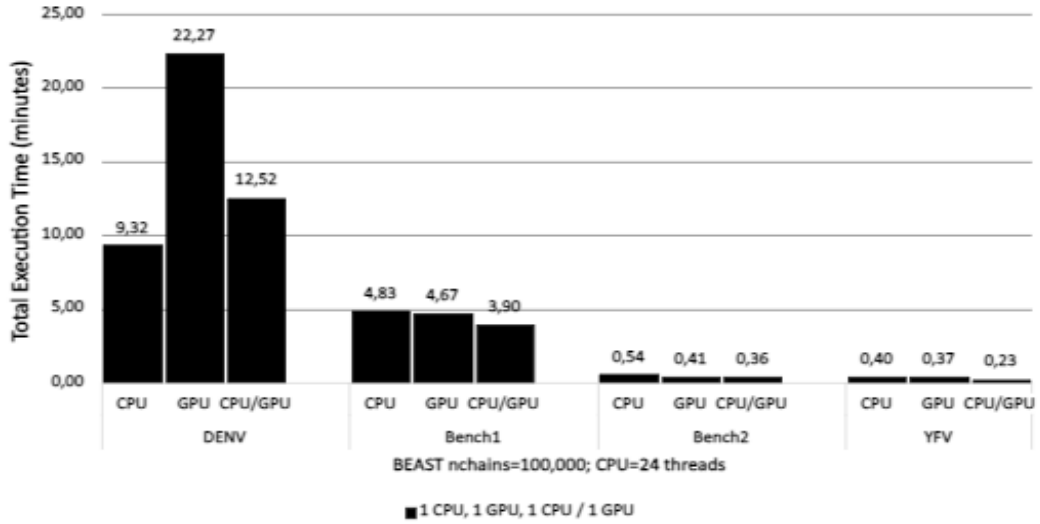


Figura 2. Análise (TTE) no BEAST 1.10 do DENV, Bench1, Bench2, YFV. c*hainLength*= "100000"

Na (Figura 3), o experimento com o *software* BEAST 1.10, foi utilizado nas análises 4 modelos de dados, sendo do vírus da Dengue (DENV), Bench1, Bench2 e vírus da febre amarela (YFV), todos em (XML). Fixamos o parâmetro *chainLength* em "20000000", selecionamos em ambos os testes 1 nó computacional (24 *threads*), alternando com 1CPU, 1GPU e 1CPU/1GPU. Os melhores desempenhos do Tempo Total de Execução (TTE) foi obtido usando o ambiente híbrido executando em paralelo 1CPU/1 GPU. No arquivo Bench2, o menor (TTE) foi do ambiente de GPU e no arquivo DENV, o menor (TTE) foi obtido com o uso de CPU.

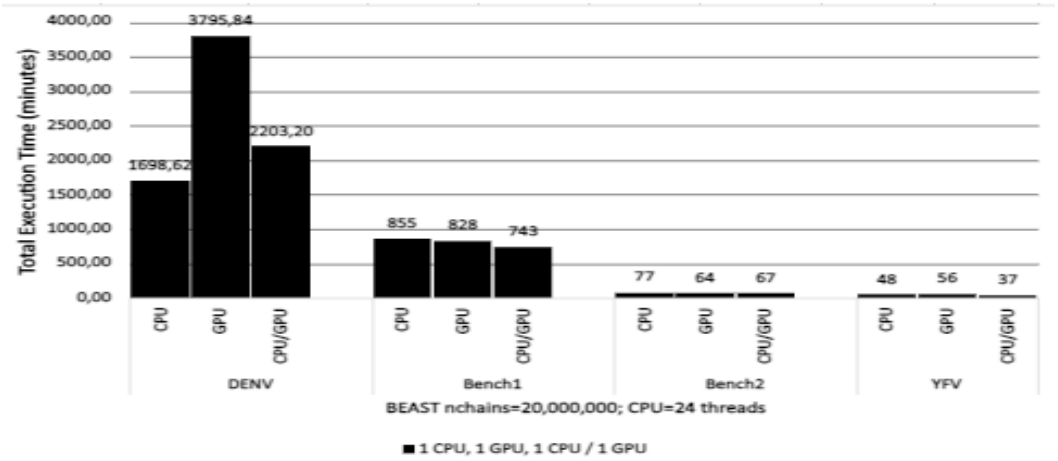


Figura 3. Análise (TTE) no BEAST 1.10 do DENV, Bench1, Bench2, YFV. c*hainLength*= "20000000"

DISCUSSÃO

Os experimentos sugerem que as características como tamanho dos dados e configuração de parâmetros no BEAST, como o *chainLength*, influenciam no tempo computacional. As execuções usando o parâmetro *chainLength* = "100000" é exploratório e requer menos gasto computacional. Quando utilizamos um valor maior de *chainLength* = "20000000" tende a gerar uma maior consistência nos resultados, os cálculos de probabilidade requeridos tornam-se mais exaustivos levando a um gasto computacional maior.

Nos resultados, o incremento do *chainLength* influência no melhor tempo computacional obtido, as execuções com alternância do parâmetro do *chainLength* apresentaram como melhor desempenho o uso dos recursos do ambiente híbrido executando em paralelo 1CPU/1GPU. Exceto para os dados do DENV com *chainLength* = "100000", nessa execução o melhor desempenho foi obtido com o uso de CPU e os experimentos com *chainLength* = "20000000", do DENV o menor (TTE) é do ambiente de CPU e do Bench2 o menor (TTE) foi com o ambiente de GPU.

CONCLUSÕES

O presente estudo viabiliza a exploração e análise de desempenho do BEAST 1.10 e a BEAGLE 3 em ambientes de CAD com a especificação do ambiente computacional que leve a um desempenho mais eficiente. Dessa maneira, permite que usuários possam usufruir dessas informações e realizar execuções garantindo um uso racional do ambiente do SDumont.

Análises de desempenho do BEAST 1.10 com a BEAGLE 3 em múltiplas configurações de CPU e GPU no SDumont sugerem o uso da configuração híbrida 1CPU/1GPU como a mais eficiente. Sobre a variabilidade no número de *chainLength* fixados em "100000" e "20000000" como esperado o melhor tempo computacional foi obtido com o ambiente híbrido executando em paralelo CPU/GPU. Exceto o arquivo DENV com *chainLength* = "100000", e com *chainLength* "20000000" os dados DENV e Bench2. Esses resultados indicam que fornecer mais threads para execução pode não trazer o benefício esperado, principalmente em relação ao tamanho dos arquivos de dados e *chainLength* estiverem envolvidos.

Este relatório apresenta estudos de desempenho de aplicações científicas BEAST 1.10 com a BEAGLE 3, em diversos ambientes computacionais do Santos Dumont: executado em paralelo em 1 nó com CPU, em 1 nó com arquitetura GPU, em um ambiente híbrido executando em paralelo na CPU e na GPU. Os estudos mostram que, para as configurações dos testes realizados para a aplicação BEAST 1.10 com BEAGLE 3, a execução de maior eficiência pelo portal é realizada ao utilizar 1 nó computacional (24 *threads*) com CPU/GPU.

Como trabalhos futuros iremos implementar testes com a aplicação BEAST, para um número maior de nós computacionais, alternância de *chainLength* e maiores arquivos de dados. Estabelecendo assim o desenvolvimento de estudos de desempenho das Aplicações de Bioinformática e dos *Workflows* Científicos do Bioinfo-Portal para que os mesmos possam vir a ser executados de forma eficiente, versátil, escalável e inteligente pelo portal.

REFERÊNCIAS BIBLIOGRÁFICAS

• Ocana, K., Coelho, M., Freire, G., and Osthoff, C. (2020). High-performance computing of beast/beagle in bayesian phylogenetics using sdumont hybrid resources. In 14º BreSci – Brazilian e-Science Workshop.

• C.-L. Hung, Y.-S. Lin, C.-Y. Lin, Y.-C. Chung, e Y.-F. Chung, “CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs,” Comput Biol Chem, vol. 58, pp. 62– 68, May 2015.

• Yin, Z., Lan, H., Tan, G., Lu, M., Vasilakos, A. V., and Liu, W. (2017). Computing platforms for big biological data analytics: Perspectives and challenges. Computational and Structural Biotechnology Journal, 15:403–411.

• Jin, Z. and Bakos, J. D. (2013). Extending the beagle library to a multi fpga platform. BMC Bioinformatics, 14(1):25.

• J. P. Huelsenbeck e F. Ronquist, “MRBAYES: Bayesian inference of phylogenetic trees,” Bioinformatics, vol. 17, no. 8, pp. 754–755, Aug. 2001.

• ERAD RJ - Exploração de Módulos Paralelo Híbrido de Bioinformática para Ambientes GPU de Supercomputação - 1 de dezembro de 2020 Autores: G Dornelas, M Coelho, K Ocaña, C Osthof .

• Ocaña K, Coelho M, Terra R, Freire G, Santos M, Cruz L, Galheigo M, Carneiro A, Fagundes B, Carvalho D, Cardoso D, Meneses E, Gadelha L, Osthoff C, DEVELOPING EFFICIENT SCIENTIFIC GATEWAYS FOR BIOINFORMATICS INSUPERCOMPUTER ENVIRONMENTS SUPPORTED BY ARTIFICIAL INTELLIGENCE. In: (ISC High Performance 2021).

• Ayres, D. L.; Cummings, M. P.; Baele, G.; Darling, A. E.; Lewis, P. O.; Swofford D. L.; Huelsenbeck, J. P.; Lemey, P.; Rambaut, A.; Suchard, M. A. (2019). “BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics”. Systematic Biology, Volume 68, Issue 6, Pages 1052–1061.

• Suchard, M. A.; Lemey P.; Baele G.; Ayres D. L.; Drummond A. J.; and Rambaut A. (2018). “Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10”. Virus Evolution, Volume 4, Issue 1, vey016.

• Hill, V. and Baele, G. (2019). “Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model”. Molecular Biology and Evolution, Volume 36, Issue 11, Pages 2620-2628.

RELATÓRIO DE ATIVIDADES - BOLSA PIBIC/LNCC

1) Dados Gerais

Título do projeto: Metodologia de auditoria de código e planejamento de otimização aplicada no núcleo dinâmico do modelo MONAN
Bolsista: Isabel de Freitas Barboza
Orientadores: Roberto Pinto Souto, Eduardo Lucio Mendes Garcia
Tipo de bolsa: PIBIC
Período do relatório: maio e junho de 2023
Início da Bolsa: maio 2023

2) Objetivos

Objetivos abordados durante o período da pesquisa:

1. Familiarização com as linguagens de programação selecionadas.
2. Desenvolvimento de habilidades na manipulação dos comandos Linux.
3. Compreensão dos conceitos e práticas da linguagem C.
4. Preparação para o progresso do projeto.
5. Capacidade de realizar experimentos mais complexos no futuro.
6. Estudos iniciais com aplicação didática `miniWeather` que se adere a proposta da pesquisa.

3) Introdução

Modelos de previsão numérica de tempo (PNT), devido à grande quantidade de cálculos que são realizados durante a sua execução, são aplicações que demandam o uso de computação de alto desempenho. Está em desenvolvimento um novo modelo comunitário unificado do sistema terrestre – o **MONAN** (**M**odel for **O**cean-**La**N**d**-**A**tmosphere predictio**N**), Modelo para Previsão dos Oceanos, Superfícies Terrestres e Atmosfera (na sigla em português), para produzir previsões com ênfase na região tropical e foco sobre a América do Sul, em diferentes escalas espaciais e de tempo, buscando incluir as necessidades dos setores produtivo e social, que substituirá os atuais modelos atmosféricos atualmente em operação. O MONAN é um programa institucional

do MCTI, coordenado pelo Instituto Nacional de Pesquisas Espaciais (INPE) e, nesta fase inicial do projeto, o LNCC está participando no desenvolvimento computacional, auxiliando na avaliação de desempenho dos códigos candidatos ao núcleo da dinâmica do modelo. Para auxílio a esta finalidade, é necessário um bom conhecimento de linguagem de programação a fim de bem entender como foi realizada a implementação dos modelos numéricos. No que diz respeito a desempenho computacional dos modelos, também é de grande importância o entendimento da estratégia de paralelização empregada nos seus códigos fonte. Portanto, este relatório contém os estudos de linguagem C, uma das linguagens mais utilizadas (além de Fortran) em módulos que formam um modelo numérico de previsão. Há também estudo preliminar de desempenho paralelo de uma aplicação (`miniWeather`) que implementa de maneira simplificada a dinâmica da atmosfera.

4) Material e Métodos

Etapas Iniciais de Estudos

Durante o período de maio e junho, foram implementadas atividades didáticas que trouxessem os conhecimentos necessários para a execução da pesquisa, abordando a importância das linguagens de baixo nível mais utilizadas no projeto como Fortran e C, além dos conhecimentos necessários para execução e desenvolvimento das atividades futuras, como a utilização do terminal para compilação e execução por linha de comando.

As primeiras atividades foram direcionadas ao desenvolvimento de programas na linguagem C **[1]**, aplicando e exercitando conceitos fundamentais relacionados a estruturas de dados e laços de repetição. Inicialmente, foram trabalhados códigos simples e conforme o progresso foi sendo alcançado, novas operações foram adicionadas. Também exploramos o uso de funções, o conceito de alocação dinâmica e estática e o conceito de Programação Modular, com isso foi introduzido o aprendizado da compilação de maneira automática com shell script, aprofundando os comandos do SO linux. Esses conhecimentos iniciais adquiridos permitiu o entendimento da instalação e execução do `miniWeather` **[2]**, repositório para estudos da pesquisa, para se analisar desempenhos computacionais, fazendo entender mais claramente esses processos.

Programação em C: Edição, Compilação, Execução

Durante as atividades iniciais do projeto, uma das áreas abordadas foi a programação em C, com foco na edição, compilação e execução dos programas desenvolvidos. Para esse propósito, utilizamos o ambiente Windows Subsystem for Linux (WSL) com o sistema operacional Ubuntu.

Através do WSL e do Ubuntu, realizamos estudos sobre a utilização da linha de comando para compilação e execução de programas em C entendendo as principais etapas da compilação, que são o pré-processamento, a compilação propriamente dita e a ligação.

Para o desenvolvimento das atividades trabalhamos com comandos Git e o GitHub para o controle de versão

- https://github.com/TempoHPC/codeaudit_ic (trabalhamos no repositório codeaudit_ic pertencente a organização TempoHPC)

Estudo de caso inicial: aplicação `miniWeather`

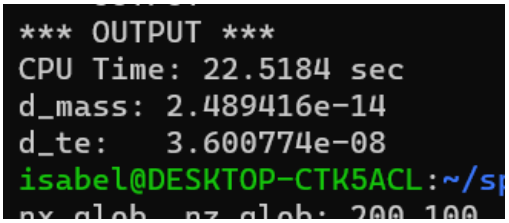
O código `miniWeather` tem como finalidade reproduzir a dinâmica básica observada na atmosfera. As equações neste código formam a espinha dorsal de praticamente todos os códigos de dinâmica de fluidos, que são a base dos modelos numéricos de previsão de tempo e clima.

- https://github.com/mrnorman/miniWeather (repositório original)
- https://github.com/TempoHPC/miniWeather (fork gerado a partir repositório original, para desenvolvimento da pesquisa na bolsa)

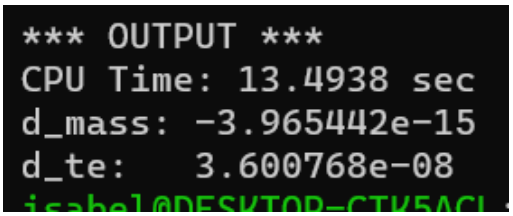
O fork do repositório `miniWeather` é um repositório acadêmico criado para fins didáticos da pesquisa. Na seção **5)Resultados e Discussão** deste relatório é mostrado o resultado da instalação e execução, com 1, 2 e 4 processadores da aplicação, em um ambiente Linux WSL - Ubuntu 20.04 LTS.

5) Resultados e Discussão

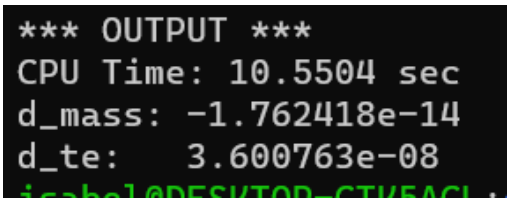
Nas próximas etapas do projeto, serão aplicados os conhecimentos adquiridos nessa fase inicial. O resultado dessa etapa é a compreensão da importância dos conceitos iniciais da computação abordados no projeto para o desenvolvimento da pesquisa. Além disso, é possível entender o objetivo científico de focar na compreensão dos processos, analisando cada passo e parte executados, em vez de se concentrar apenas no resultado final. Também é importante destacar a compreensão dos conceitos de linguagem, comandos e ambiente de trabalho.
Nestes dois primeiros meses de pesquisa, as últimas etapas foram a instalação e execução do `miniWeather` e foram obtidos os seguintes resultados de desempenho computacional com a versão MPI **[3]** do código:



Com um processo MPI



Com dois processos MPI



Com quatro processos MPI

Percebe-se uma consistente redução no tempo de processamento com o acréscimo de processos MPI. Esses resultados contribuem de forma significativa

para o alcance dos objetivos propostos, fornecendo uma base para o avanço da pesquisa e o desenvolvimento de habilidades na área da computação e do projeto.

6) Conclusões

As principais conclusões obtidas no decurso do trabalho realizado destacam

1. A importância dos conceitos iniciais da computação;
2. A análise detalhada dos processos;
3. A compreensão dos conceitos de linguagem, comandos e ambiente de trabalho;
4. O uso adequado das ferramentas disponíveis;
5. Observação de ganho de desempenho com paralelização da aplicação `miniWeather`;
6. As próximas etapas na pesquisa consistem em obtenção de perfil de execução da aplicação, bem como explorar outros níveis de paralelização, como o uso de múltiplas *threads* com OpenMP **[4]**;

Essas conclusões fornecem uma base para o prosseguimento do projeto, permitindo avançar com confiança e eficiência nas etapas subsequentes.

7) Referências Bibliográficas

[1] Notas de aulas, Linguagem C. Eduardo Lucio Mendes Garcia. Pós graduação em modelagem. 2023

[2] Norman, Matthew R, and USDOE. miniWeather. Computer software. https://www.osti.gov//servlets/purl/1631691. USDOE. 15 Mar. 2020. Web. doi:10.11578/dc.20201001.88.

[3] Gropp, William, et al. Using MPI: portable parallel programming with the message-passing interface. Vol. 1. MIT press, 1999.

[4] Chapman, Barbara, Gabriele Jost, and Ruud Van Der Pas. Using OpenMP: portable shared memory parallel programming. MIT press, 2007.

RELATÓRIO DE ATIVIDADES

1. Dados gerais

Título do projeto: Correlação da coesão das publicações em mídias sociais sobre a COVID-19 e os subeventos relacionados à pandemia

Nome do bolsista: João Matheus Nascimento Gonçalves

Nome do orientador: Fabio Porto (coorientação de Tiago Cruz de França e Jonice Oliveira)

Tipo de bolsa: PIBIC

Período: 01/12/2022 a 31/08/2023

2. Objetivos

O trabalho teve como objetivo desenvolver um método de análise de discurso em mídias sociais (MS) ao longo do tempo durante eventos extremos, através da análise de coesão textual. Dentro deste tema, um dos objetivos foi entender as métricas (FADIGAS e PEREIRA 2013) usadas para a análise de coesão textual. Considerando a coleta e análise realizada de textos de MS acerca da COVID-19, também foi objetivo deste trabalho correlacionar as variações temporais da coesão com acontecimentos reais que foram comentados nas MS.

3. Introdução

Coletivamente, os usuários das MS geram, rapidamente, um grande volume de dados sobre eventos que acontecem no mundo. Em grandes eventos (como a pandemia da COVID-19), o volume de interações sobre o assunto é muito grande e, devido à longa duração do evento, o conteúdo pode variar, por exemplo com o surgimento de subtemas relacionados a um mesmo grande evento. A análise do volume de dados requer soluções computacionais que viabilizem a identificação e a observação de aspectos do discurso.

As atividades desenvolvidas se voltaram à análise da coesão textual do discurso em MS. Para fazer a análise do discurso, foi construído um método para investigação de coesão textual ao longo do tempo de publicações relacionadas a um grande evento de longa duração (e com diferentes ocorrências/subeventos). Foi realizada uma análise do método proposto (o VERSATILE) com bases de dados sintéticas, criadas de forma que possuíssem características desejadas. Em seguida, realizou-se a análise de duas bases de tuítes sobre a COVID-19. A principal contribuição deste trabalho é o método VERSATILE para análise de coesão de bases textuais de documentos que representam discursos em MS.

Um artigo foi publicado no BraSNAM 2023, onde é feito o detalhamento do VERSATILE e se descreve sua aplicação em diferentes bases textuais (algumas bases sintéticas e uma base de tuítes).

4. Metodologia

4.1. Processamento dos dados coletados das duas bases do Twitter

Para o desenvolvimento da pesquisa, foi preciso processar os dados coletados do Twitter. Utilizou-se duas bases textuais: uma com tuítes dos principais perfis relacionados à resposta à pandemia no Brasil (NEVES *et. al,* 2020), no período de 13/03 a 31/05 de 2020; e outra como parte da base extraída por (de MELO e FIGUEIREDO, 2020), contendo tuítes em português com algumas palavras relacionadas à COVID-19. Desta segunda base, foram analisados cerca de 700.000 tuítes, no período de 01/01/2020 a 31/05/2020.

4.2. Desenvolvimento do VERSATILE

Para analisar as bases de dados, foi criado o método VERSATILE, o principal objetivo do trabalho realizado. O método é descrito com um diagrama de atividades na figura abaixo:

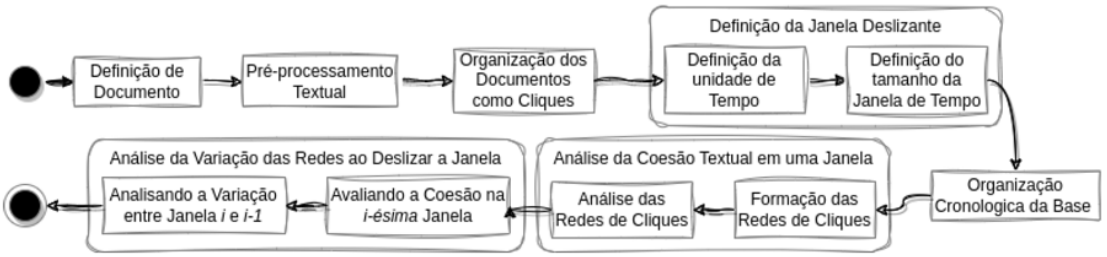


Figura 1. Diagrama de atividades com as etapas do método VERSATILE

Para abordar a questão da variação no tempo, foi elaborada uma estratégia de janelas (intervalos de tempo) de análise que deslizam em unidades de tempo. Os documentos utilizados para a análise precisam estar associados a estampas de tempo do momento da sua publicação.

Na etapa de **definição do documento**, um documento de texto (*d*) será utilizado como unidade a ser analisada. Todo *d* pertence a uma base textual (B) composta por N publicações coletadas de uma mídia social. O documento pode ser todo o texto da publicação, uma frase ou um parágrafo de uma publicação.

No **pré-processamento textual** foi feita: a conversão de letras para minúsculas, a remoção da pontuação, a tokenização, a remoção de *stopwords* (palavras como artigos ou conectivos textuais), e a extração de radicais com técnicas como *lemmatization* ou *stemming* (KHYANI *et al.,* 2021). Cada palavra será um token (n-gram igual 1). Os tokens foram agrupados e identificados como pertencendo a um documento.

A **organização dos documentos como cliques** sucede o pré-processamento, e é feita a partir da criação de cliques (grafos completos, inicialmente isolados), conforme proposto por (FADIGAS e PEREIRA, 2013). Os cliques representam os documentos pré-processados, sendo os tokens os nós do grafo, e as arestas definidas pela coexistência dos tokens em um documento.

Para a **definição da janela deslizante**, são estabelecidos os intervalos a serem analisados (janelas) como divisões no tempo de duração do evento estudado. O termo "deslizante" faz referência ao intervalo (*s*) entre o começo de uma janela e o começo da próxima, de modo que as janelas "deslizam" uma certa quantidade de tempo. Define-se a **unidade de tempo** (*t*) para organizar e analisar os dados em uma granularidade (meses, dias, horas, minutos, etc.) que possibilite a melhor investigação sobre o evento estudado. Uma **janela de tempo** (T) corresponde a um período de duração *p* múltiplo de *t*. Se *t* corresponde a 1 dia, pode-se definir janelas de 5*t*, então *p*=5*t*. Assim, temos $T_1 = \{t_1, t_2, t_3, t_4, t_5\}$ que corresponderá aos documentos publicados nesse intervalo. Com o deslize, se definirmos s=1t, a janela T_2 será dada por $T_2 = \{t_2, t_3, t_4, t_5, t_6\}$.

Na **organização cronológica**, os dados são ordenados de acordo com a estampa de tempo das mensagens. Essa etapa trata da preparação para que a análise seja realizada nas janelas de acordo com a unidade de tempo definida.

A **análise da coesão textual em uma janela** se baseou em (FADIGAS e PEREIRA, 2013). Ela se inicia após o processo da formação da rede de cliques. A conexão dos cliques acontecerá de acordo com as palavras (tokens) iguais entre eles. As conexões ocorrem por justaposição (dois ou mais cliques possuem um nó em comum) ou sobreposição (mais de um nó de dois ou mais cliques são iguais). Uma rede com poucas ocorrências desses processos seria pouco coesa, uma vez que isso indicaria menor ocorrência de formas de reiteração dos termos entre os documentos. As métricas utilizadas para análise da coesão foram: variação de densidade - $v(\Delta)$; variação do grau médio - $v(\langle k \rangle)$; coeficiente de clusterização - C; fragmentação - F; e fragmentação de cliques - F_{cliques} (FADIGAS e PEREIRA, 2013).

A **análise da variação das redes ao deslizar a janela** possibilita que se perceba a diferença entre os índices de coesão ao longo do tempo. Numa base textual relacionada a um longo período, haverá a oportunidade de observar a variação das métricas de coesão no tempo. Cada métrica possibilitará, em algum grau, a inferência de aspectos do discurso.

4.3. Aplicação do VERSATILE

O VERSATILE foi aplicado em bases sintéticas (com diferentes níveis de coesão preestabelecidos) que simulavam tuítes publicados ao longo do tempo; e em duas bases extraídas do Twitter, conforme detalhado na seção 4.1. Para todas as bases, o documento usado foi o texto integral do tuíte.

As bases sintéticas utilizadas foram as seguintes: uma "pouco conexa", uma "conexa" e uma "muito conexa". Cada mensagem nas bases recebeu uma estampa de tempo com data e hora da publicação (como acontece com os tuítes). Para a construção dessas bases, foi definido que as palavras nas mensagens se repetiriam em uma porcentagem x do total de palavras usadas. Para a base "pouco conexa", o valor de x foi definido empiricamente para estar entre 0% e 12%; para a base "conexa", x variou de 15% a 33%; e para a base "muito conexa", este valor variou entre 70% e 95%. Estas diferenças na construção das bases influenciam diretamente as redes de cliques geradas ao se aplicar o VERSATILE.

Foram analisados os índices de coesão das janelas de tempo correspondendo a uma transição gradual entre: i) uma base pouco conexa (B1) e uma base conexa (B2); e ii) duas bases conexas diferentes (B2 e B3). A unidade de tempo (t) foi de um dia (t=1), a janela p=10 dias, e o deslize s=1 dia. O propósito foi simular cenários de mudança na coesão do discurso em MS, variando de discursos mais ou menos coesos sobre diferentes assuntos ao longo do tempo. O VERSATILE também foi aplicado para cada uma das bases sintéticas, usando uma janela de 10 dias que correspondia à integralidade da base (sem deslizar).

Com a aplicação do VERSATILE nas bases de dados sintéticas, buscou-se entender melhor as métricas de coesão para diferentes bases textuais (um dos objetivos deste trabalho), bem como para cenários de transição entre diferentes níveis de coesão, motivados possivelmente (no caso real) por diferentes subeventos acerca do evento analisado vindo à tona nas MS. Tal análise dos índices de coesão agrega valor aos resultados para as bases com textos reais, do Twitter.

5. Resultados e discussão

5.1. Análise das bases sintéticas

A Figura 2 (a) apresenta o resultado das análises da transição da base pouco conexa para a base conexa, onde observou-se aumento gradual de índices de coesão como $v(\Delta)$ e $v(\langle k \rangle)$. O

índice da fragmentação não apresentou mudanças significativas, o que também foi observado para a transição entre duas bases conexas.

A Figura 2 (b) apresenta os resultados das análises quando duas bases conexas são colocadas adjacentes uma à outra. Mesmo sendo as duas bases conexas, as bases têm "assuntos" diferentes (por serem formadas de bancos de palavras diferentes). Assim, a conectividade na transição não é tão grande quanto em cada base individualmente. Com isso, $v(\Delta)$ e $v(\langle k \rangle)$ diminuem nas janelas intermediárias, e $F_{cliques}$ e C aumentam. Mas os índices se aproximam novamente dos valores iniciais na última janela, na qual predomina o texto da segunda base conexa.

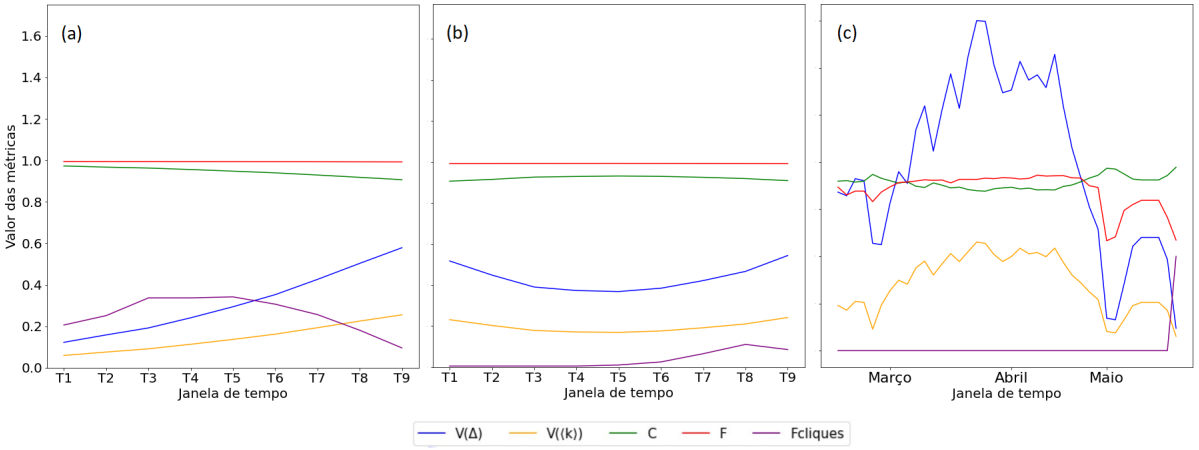


Figura 2. Variação de janelas em bases: a)Transição da base pouco conexa para a conexa; b) Transição entre duas bases conexas diferentes; c) Base do Twitter.

5.2. Análise das bases do Twitter

Para análise dos tuítes da COVID-19, definiu-se t=1 dia, p=6 dias e s=1 dia. Importante notar que, enquanto as janelas das bases sintéticas eram de 10 dias e tinham invariavelmente 200 tuítes (todos com 20 palavras), nas janelas da base do Twitter, os valores podem variar.

Nas bases de dados de (NEVES *et. al,* 2020) (Figura 3 (c)), pode-se observar um $v(\Delta)$ majoritariamente entre 0,5 e 1,6, sendo o índice mais baixo próximo daquele observado em bases sintéticas "conexas". Isto aponta para um nível de coesão médio de acordo com a análise das bases sintéticas.

O $v(\langle k \rangle)$ de 18/03 a 13/04 esteve, em geral, maior do que o valor observado para as bases sintéticas "conexas", mas sem chegar aos valores da base "muito conexa", sendo um indício de uma quantidade considerável reiterações entre os tuítes, colaborando para sua coesão. Foram poucas as janelas nas quais os valores estiveram abaixo daqueles encontrados para bases sintéticas conexas. Em 17/03 aconteceu a primeira morte por COVID-19 no Brasil. Nessa data, passou a ser crime desrespeitar as medidas de isolamento [Sanarmed 2020].

O valor de C das janelas ficou majoritariamente abaixo daquele em bases sintéticas conexas, mas com algumas ocorrências próximas de 1 (valor máximo) e mantendo-se estável ao longo do período estudado. Os valores mais baixos de C provavelmente se relacionam a uma maior coesão, visto que indicam um "distanciamento" da rede de cliques para o estado inicial de cliques isolados, no qual C sempre é igual a 1.

Diferente do caso para as bases sintéticas, F diminuiu consideravelmente nas últimas 3 janelas, se aproximando do índice predominante nas bases sintéticas "muito conexas". $F_{cliques}$ foi nula exceto por uma janela, o que significa que o estado final da rede de cliques, em

quase todos os casos, apresentou um número de componentes igual a 1, o que é um indicador positivo de coesão conforme a análise das bases sintéticas.

Ressalta-se a variação brusca das métricas em 5 janelas de tempo contíguas a partir de 13/04. Este período está relacionado a subeventos muito importantes (causando uma variação e diversidade de assuntos), como a demissão do ministro da saúde (16/04), o Brasil fica de fora da ACT Accelerator (05/05), chegou-se a 10.000 mortos (09/05), a contratação e queda do 2º Ministro da Saúde (15/05). Porém, não foi possível investigar tal suposição, porque a coleta teve problemas nas janelas começando em 14/04 até aquelas começando em 7/05.

Para a base de (de MELO e FIGUEIREDO, 2020), destaca-se o índice $v(\langle k \rangle)$ (Figura 4). Os picos observados esta métrica foram em datas próximas a alguns eventos importantes relacionados à COVID-19 (SANARMED, 2020): em 21/01 é confirmada a primeira transmissão do novo coronavírus entre humanos; em 26/02, é confirmado o primeiro caso de COVID-19 no Brasil; em 11/03, a OMS declarou oficialmente o início da pandemia De COVID-19; em 17/03, é confirmada a primeira morte no Brasil por conta do vírus; em 03/05, o Brasil passa de 100.000 mortes por COVID.

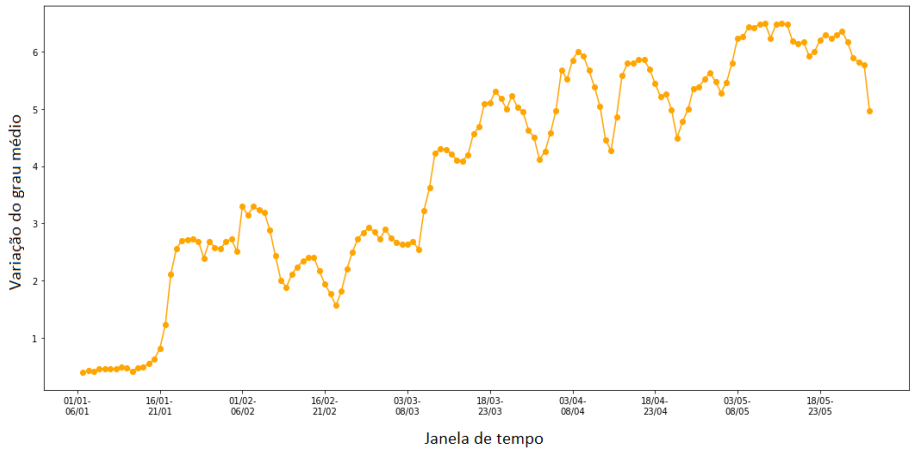


Figura 4. Variação do grau médio na base do Twitter analisada.

O valor de C, nos períodos onde $\nu(\Delta)$ e de $v(\langle k \rangle)$ foram muito altas, ficou majoritariamente abaixo de 0,8, um pouco menor do que o valor deste índice para as bases sintéticas muito conexas, sinalizando uma alta coesão textual. Já a fragmentação esteve invariavelmente próxima de 1, embora um pouco menor nas janelas de tempo antes do dia 16/01. $F_{cliques}$ foi majoritariamente nula ou muito próxima de zero, o que é um indicador positivo de coesão conforme a análise das bases sintéticas.

6. Demais Atividades Durante o Período de Iniciação Científica

- Apresentação do trabalho na Semana de Integração Acadêmica da Universidade Federal do Rio de Janeiro (UFRJ);
- Frequentar o LabCores (Laboratório de Computação Social e Análise de Redes Sociais) da UFRJ, participar de reuniões e interagir com outros alunos de graduação e pós-graduação;
- Participação no grupo de Sensoriamento e Monitoramento do LabCores, colaborando com outros trabalhos e apresentando as atividades para os colegas;
- Apresentar o trabalho para a turma de Análise de Redes Sociais;
- Colaborar na manutenção em códigos de *scraping* de dados Web;

- Estudar a área de processamento de linguagem natural.

7. Considerações Finais

Neste trabalho foi apresentado e aplicado o método VERSATILE para avaliação da coesão textual lexical entre mensagens publicadas por diferentes usuários de uma mídia social que estão relacionadas a um assunto em comum. Foram realizados testes com bases sintéticas e com duas bases da COVID-19 de tuítes escritos em português brasileiro. Foi possível perceber o nível de coesão das bases, o que demonstrou a diferença entre os termos usados nas publicações na época. A partir dos resultados, percebeu-se o potencial do método para análise automática da coesão de conteúdo textual publicado nas MS. Em especial para a base de (de MELO e FIGUEIREDO), foi possível correlacionar diversos subeventos relacionados à COVID-19 com as mudanças observadas nos índices de coesão.

Como trabalho futuro, vislumbra-se o aperfeiçoamento do método, levando em conta a presença de sinônimos ou a função sintagmática dos termos em orações ao analisar um texto. Também é possível explorar outras características do grafo formado para identificar o conteúdo do discurso. Um exemplo seria encontrar *clusters* nos grafos, indicando padrões e estruturas neles presentes.

Referências

FADIGAS, I.S. e PEREIRA, H.B.B. (2013) “A network approach based on cliques”. Physica A: Statistical Mechanics and its Applications.

NEVES, J. C. B. ; FRANÇA, Tiago Cruz ; BASTOS, M. P.; CARVALHO, P. V. R.; GOMES, J. O. Analysis of government agencies and stakeholders? twitter communications during the first surge of COVID-19 in Brazil. WORK-A Journal of Prevention Assessment & Rehabilitation, v. 73, p. 1-13, 2022.

de MELO, Tiago; FIGUEIREDO, Carlos M.S. (2020). “A first public dataset from Brazilian twitter and news on COVID-19 in Portuguese”, Volume 32, 106179,ISSN 2352-3409, https://doi.org/10.1016/j.dib.2020.106179.

KHYANI, D.; SIDDHARTHA, B. S.; NIVEDITHA, N. M.; DIVYA, B. M. (2021) “An interpretation of lemmatization and stemming in natural language processing”. Journal of University of Shanghai for Science and Technology, *22*(10), P. 350-357.

GONÇALVES, J.M.N.; OLIVEIRA, J.; PORTO, F.; FRANÇA, T.C. (2023) “Análise Temporal de Discurso em Mídia Social Durante Grandes Eventos”. BraSNAM 2023 - XII Brazilian Workshop on Social Network Analysis and Mining, XLIII Congresso da Sociedade Brasileira de Computação.

SANARMED (2020). “Linha do tempo do coronavirus no Brasil”, https://www.sanarmed.com/linha-do-tempo-do-coronavirus-no-brasil.

FRANÇA, Tiago Cruz de. "ANDARE: um framework para inclusão da análise de dados de mídias sociais no contexto da preparação e resposta à emergência em situações de manifestações de massa", 2019, Tese (Doutorado) - Curso de Pós-graduação em Informática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019, https://tinyurl.com/tmaydae4. Acesso em: 08 Mar. 2023.

CASTEIGTS, Arnaud; FLOCCHINI, Paola; QUATTROCIOCCHI, Walter; SANTORO, Nicola. (2010). Time-Varying Graphs and Dynamic Networks. International Journal of Parallel, Emergent and Distributed Systems. 27. 10.1007/978-3-642-22450-8_27.

Coletor de amostras de dossel embarcado em veículo aéreo não tripulado

Bolsa: PIBIT - LNCC, Período do relatório: 12 meses

Aluno: João Vitor Rosa Rebello[1]
Orientadores: Jauvane Cavalcante de Oliveira, PhD[2] **; Luis Claudio Batista da Silva, D.Sc.**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET-RJ) Campus Petrópolis

[2]Laboratório Nacional de Computação Científica (LNCC) - Petrópolis - RJ

1. Objetivo

O objetivo deste projeto é estudar e desenvolver atuadores e efetuadores robóticos aplicados no desenvolvimento de um robô do tipo veículo aéreo não tripulado (VANT) para coleta de amostras de folhagem no topo das árvores, conhecido como dossel. Sendo assim, para dar início a este processo fez-se necessário estudar cuidadosamente os temas relacionados a introdução a robótica e assuntos mais específicos que dizem respeito as tarefas do robô. Logo após foi iniciada a etapa de esquematização e modelagem do mecanismo. Dito isso, este relatório tratará de explicar e detalhar os a esquematização e modelagem ao longo de 12 meses, e como será a construção do mecanismo à partir de agora.

2. Introdução

No Brasil, a coleta de amostras de árvores em regiões de mata são geralmente feitas manualmente. Isto acaba ocasionando risco grave para o operador que realizará a coleta e também consequências para a fauna e flora local por conta da extração desse material ser feita de forma predatória. À fim de minimizar esses riscos e consequências, um dispositivo para realizar essa coleta de forma remota foi projetado.

A principal ideia do projeto é que estas aeronaves do tipo VANT substituam o uso de trabalho manual para coleta de amostras no topo das arvores, evitando assim riscos acidentais para o ser humano, visto que a prospecção atual de amostras desse tipo é feita manualmente. x Por tratar-se então de um robô, se fez necessário os estudos sobre temas relevantes na robótica, cujo progresso foi detalhado no relatório de atividades anterior, à partir disso então, foi iniciada a etapa de esquematização e modelagem do mecanismo.

3. Metodologia

Utilizando todo o arcabouço teórico estudado na primeira fase do projeto,seguiu-se para a etapa de desenvolvimento que inicialmente foi criar um esboço do mecanismo em papel: quais seriam suas funcionalidades, qual seria seu formato e quais componentes utilizaria.

Com isso, as especificações definidas para a amostra foram:

- Peso: até 1.8kg;
- Comprimento: mínimo de 25cm;
- Diâmetro: até 10cm;

Após isso, foi iniciada a modelagem do mecanismo esboçado na ferramenta 3D, que no caso, utilizou-se o software Inventor 2023 da AutoDesk com licença acadêmica gratuita.

Primeiramente foram criadas todas as partes do mecanismo separadamente, não focando em aparência da peça nem em desempenho dela, somente para parametrização do mecanismo completo.

Sendo assim, as primeiras peças passaram por um refinamento, para só então serem montadas no conjunto, que seria o próprio mecanismo. Neste conjunto, a ferramenta Inventor permite a criação de animações e simulações de movimentos das peças, o que foi de extrema importância para o entendimento do funcionamento das peças, e como cada alteração em uma peça específica, refletia-se na montagem geral.

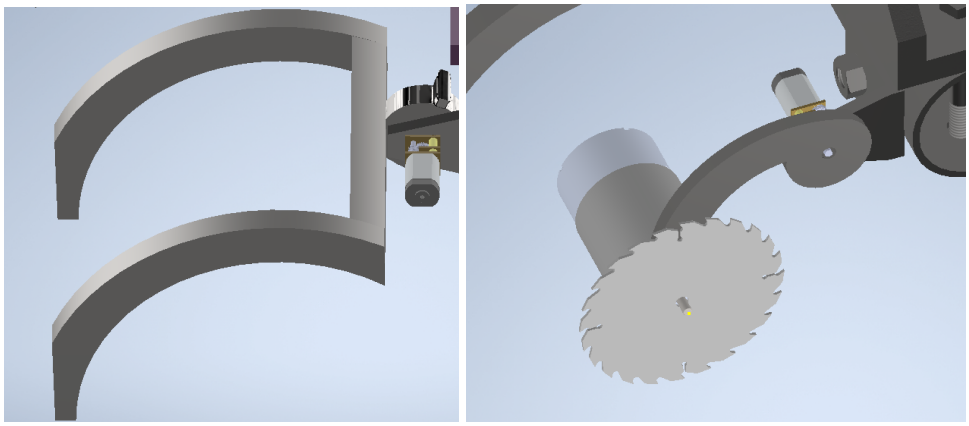
4. Resultados

4.1. Esquematização do sistema

4.1.1. Garra

As garras ficarão dispostas em sua base, que mede 25cm centímetros de comprimento. O mecanismo terá 2 tipos de garras (Figura 3):

1. Garra de sustentação: Garra na qual ficará apoiado o galho à ser retirado do topo da árvore, possui 2 hastes de sustentação de peso (Figura 1a);
2. Garra de serra: Esta garra terá comprimento menor e não participará da sustentação do galho, nela será acoplada a serra e seu motor (Figura 1b).



(a) Garra de sustentação **(b) Garra com serra acoplada**

Figura 1: Tipos de garras do mecanismo

As garras de sustentação não pressionarão o galho para mantê-lo preso ao mecanismo. Dada a característica de fechamento em tesoura do projeto, a amostra ficaria presa ao dispositivo pela força peso dela e suas ramificações que enganchariam na garra.

4.1.2. Suporte

A base com as garras ficaria ligada à um suporte de cabos que conteria o microcontrolador, a câmera IMX219-83 com ajuste de angulação, a trava de segurança e a célula de baterias

destes componentes (Figura 2). Esta ligação seria feita por 4 cabos, com pontos de apoio formando um retângulo. Isso foi feito para manter a estabilidade do suporte de garras para que não haja rotação do mecanismo no ar sem que o usuário queira.

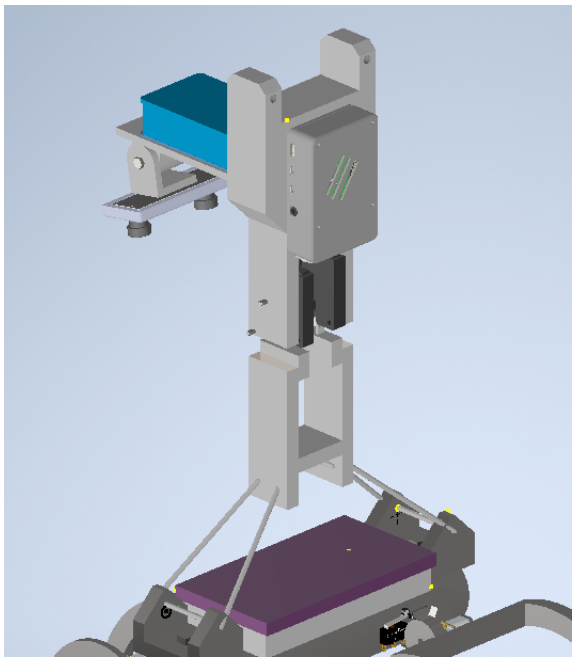


Figura 2: Suporte do mecanismo

Esses cabos também fornecem maior liberdade de movimentação do suporte de garras quando em contato com a amostra, pois pode compensar o desnível da amostra no momento da coleta utilizando a folga do cabo para tal feito.

4.1.3. Mecanismo de segurança

O mecanismo de segurança foi projetado para soltar as garras do drone caso fique preso em alguma parte da árvore (Figura 3b). Para isso, foram utilizadas duas travas elétricas SARY de 12V de 50mm com detecção de abertura (Figura 3a) para segmentar a parte de cima do suporte onde estão o microcontrolador, sua célula de bateria e a câmera, e a parte de baixo onde estão as garras e suas células de bateria.

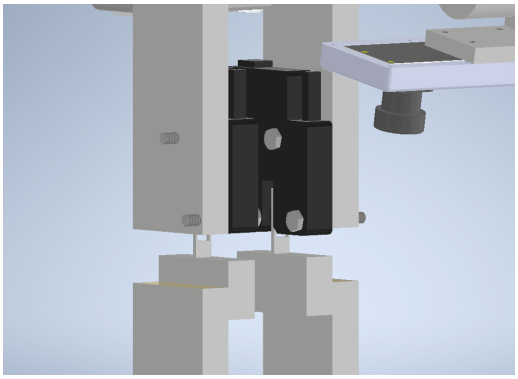
A trava seria então liberada caso o condutor do drone perceba que a aeronave não alça voo para além da árvore, separando o suporte de garras do mecanismo. A Figura 3b mostra o mecanismo de segurança

4.1.4. Microcontroladores

Para processamento de imagens e controle dos dispositivos mecânicos e eletrônicos, foi decidido que o microcontrolador a ser utilizado seria um Raspberry PI 4 modelo B (Figura 4a).

Este modelo de microcontrolador existe com quantidade de memória RAM variável, portanto, o escolhido para este projeto foi o de 4GB. No projeto, para protegê-lo

(a) Trava elétrica SARY 12v 50mm

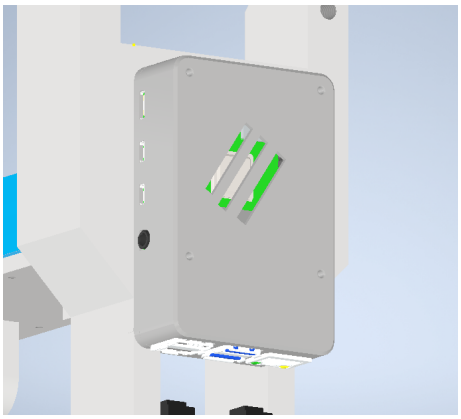


(b) Modelagem do mecanismo de segurança

Figura 3: Trava elétrica e seu uso no projeto

de fenômenos climáticos, o microcontrolador ficaria dentro de uma caixa protetora ou ”case”logo acima do mecanismo de segurança e em lado oposto à câmera (Figura).

(a) RaspBerry Pi 4 model B



(b) RaspBerry protegido pelo case

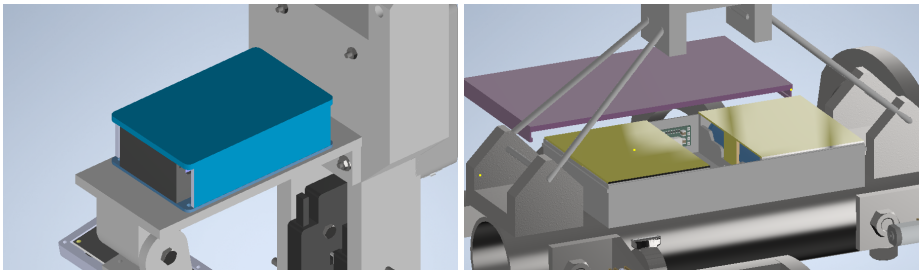
Figura 4: Microcontrolador e sua implementação no projeto

4.1.5. Energia

Foi decidido que o mecanismo teria seus componentes energéticos acoplados em sua estrutura, e não seria utilizada bateria do drone para nenhuma ação que não fosse o movimento dele. Então, segundo os cálculos de tensão e corrente total do mecanismo, as células seriam divididas em 2 unidades (Figura 5):

- Célula superior (Figura 5a): Tensão de 7,4V com 3A, designada para os dispositivos antes da trava elétrica, que seriam a câmera, o microcontrolador e a própria trava;
- Célula inferior (Figura 5b): Tensão de 22,2V com 3A, designada para os motores das garras e para o motor da serra.

Em cada célula de energia, são utilizadas baterias Li-ion 18650 de 3.7V, na superior, seriam usadas 2 baterias, e na inferior 6. Devido à volatilidade dessas baterias, as



(a) Célula de bateria superior (b) Célula de bateria inferior

Figura 5: Células de bateria do dispositivo

duas células de energia estariam dentro de uma caixa protetora, para que se reduza o risco de eventos climáticos ou físicos afetarem o funcionamento ou integridade das baterias.

É importante explicar que, o mecanismo foi projetado para que não seja necessário aplicar energia nos motores de garra, assim que eles já estiverem com o galho preso, pois o torque dos motores sem energia supera o torque aplicado pela amostra. Então, para que o motor da serra seja ligado, todos os outros motores das garras de sustentação estariam desligados e apenas a garra da serra e o próprio motor da serra estariam atuando, tendo então tensão nominal suficiente para acionamento dessas peças.

Também deve-se deixar explícito que para utilizar-se das células de energia, foi necessário acrescentar um módulo Mini Conversor De Tensão (também conhecido como step-down) para reduzir a voltagem que será utilizada para os motores, mesmo que os motores possam aguentar uma tensão mais elevada, à fim de preservar integridade desses componentes. Também foram incluídas relés no projeto para ativação de modo digital dos motores por meio do microcontrolador.

Com isso, para efeitos de entendimento do leitor, a figura 6 mostra o diagrama esquemático dos motores de sustentação da garra, e como seriam ativados. Na figura é exemplificado o uso para a ativação dos motores em uma direção de rotação, sendo a direção oposta também possível ser controlada por de relés de inversão.

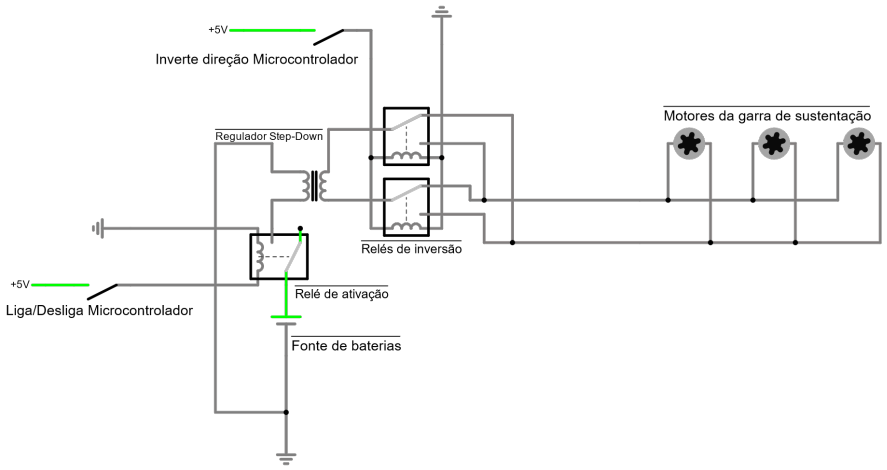


Figura 6: Diagrama esquemático de energia das garras de sustentação

O diagrama é bem similar para o uso da garra da serra e o motor da serra, mudando apenas o sinal de controle do microcontrolador, a tensão após o regulador e os próprios motores utilizados.

5. Conclusão

Com o estudo realizado e o modelo projetado, será possível à partir desta etapa, realizar testes com os primeiros protótipos do mecanismo. De acordo com as animações projetadas para o mecanismo, a proposta inicial é cumprida, à partir de agora então, será realizada uma bateria de testes.

Os primeiros testes serão feitos sem o uso de drone, utilizando apenas o mecanismo, e com movimentação manual do protótipo à ser montado. Após concluída a primeira etapa de testes e ajustes, o drone será introduzido ao projeto, e nova bateria de testes e ajustes será realizado.

Também foi estudado a possibilidade de implementação de um algoritmo de visão computacional para auxiliar o operador do drone ao retirar a amostra, porém, ainda não existem bancos de dados nacionais oficiais com imagens de galhos e topo de árvores para treinar um algoritmo de aprendizado de máquina deste tipo. Portanto, isso será adquirido ao longo do tempo com pesquisadores da área, e quando finalizado, será também introduzido a assistência de manuseabilidade por meio de inteligência artificial.

Referências

Arafat, M. Y., Alam, M. M., and Moh, S. (2023). Vision-based navigation techniques for unmanned aerial vehicles: Review and challenges. *Drones*, 7(2):89.

Bouabdallah, S., Noth, A., and Siegwart, R. (2004). Pid vs lq control techniques applied to an indoor micro quadrotor. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2451–2456. IEEE.

Charron, G., Robichaud-Courteau, T., La Vigne, H., Weintraub, S., Hill, A., Justice, D., Bélanger, N., and Lussier Desbiens, A. (2020). The deleaves: a uav device for efficient tree canopy sampling. *Journal of Unmanned Vehicle Systems*, 8(3):245–264.

da Aeronáutica, B. C. (2016). Sistemas de aeronaves remotamente pilotadas e o acesso ao espaço aéreo brasileiro.

Groover, M., Weiss, M., Nagel, R. N., and Odrey, N. G. (1986). *Industrial Robotics: Technology, Programming, and Applications*. McGraw-Hill Higher Education.

NETO, R. P. M. and BREUNIG, F. M. (2019). Drones nas ciências florestais. *Drones E Ciência*, page 68.

Sreenath, K., Lee, T., and Kumar, V. (2013). Geometric control and differential flatness of a quadrotor uav with a cable-suspended load. In *52nd IEEE Conference on Decision and Control*, pages 2269–2274. IEEE.

Tang, L. and Shao, G. (2015). Drone remote sensing for forestry research and practices. *Journal of Forestry Research*, 26:791–797.

Relatório de atividades

Bolsa PIBIC/LNCC
Projeto: Simulação 3D do voo inaugural de Santos Dumont no 14 bis
Bolsista: Jonatas Halliday Sant Anna
Orientador: Jauvane C. de Oliveira

- **Dados Gerais**

Projeto de Iniciação científica desenvolvido por Jonatas Halliday Sant Anna do Nascimento orientado por Jauvane C. de Oliveira com o título Simulação 3D do vôo inaugural de Santos Dumont no 14 bis através do PIBIC-LNCC que iniciou em Setembro de 2022 e deve seguir até a graduação do aluno.

- **Objetivos**

O projeto em questão tem por objetivo desenvolver uma simulação 3D do primeiro voo de Alberto Santos Dumont com o 14 bis, no campo de Bagatelle em 23 de Outubro de 1906. Tal simulação será usada posteriormente pelo Museu Casa de Santos Dumont, localizado em Petrópolis/RJ, em comemoração do sesquicentenário de Santos Dumont e do Dia do Aviador. A simulação incluirá, além do 14 bis e seu piloto, o local onde ocorreu o voo, bem como a reação da platéia lá presente.

- **Introdução**

Todo o projeto foi feito utilizando o ambiente Unity e Unity Hub para o desenvolvimento. Este ambiente é conhecido pela comunidade por ser capaz de criar simulações 2D e 3D. Neste ambiente utiliza-se a linguagem C# adaptada para unity para desenvolver os scripts utilizados no projeto. Usando um modelo do 14 bis, foi adicionado um terreno com algum grau de elevações, ou seja, não totalmente plano e vegetação característica - grama. O projeto procura ser fidedigno ao voo e portanto obedecendo a escala do ambiente de simulação utiliza-se a distância percorrida de 60 m e a uma altura de 2 m do solo. Para obter um maior grau de imersividade foi introduzida uma segunda câmera que simula o(a) observador(a) do local exato de onde Santos Dumont esteve no 14 bis durante o voo. A simulação obedece também os ângulos de elevação e descida do 14 bis original.

- **Métodos**

A implementação dos métodos foi feita através de classes, ou seja, orientada a objetos. Foi anexado ao objeto 14 bis as duas câmeras

existentes. Uma com a visão na parte de trás da aeronave e a segunda simulando o local onde Santos Dumont esteve durante o voo.

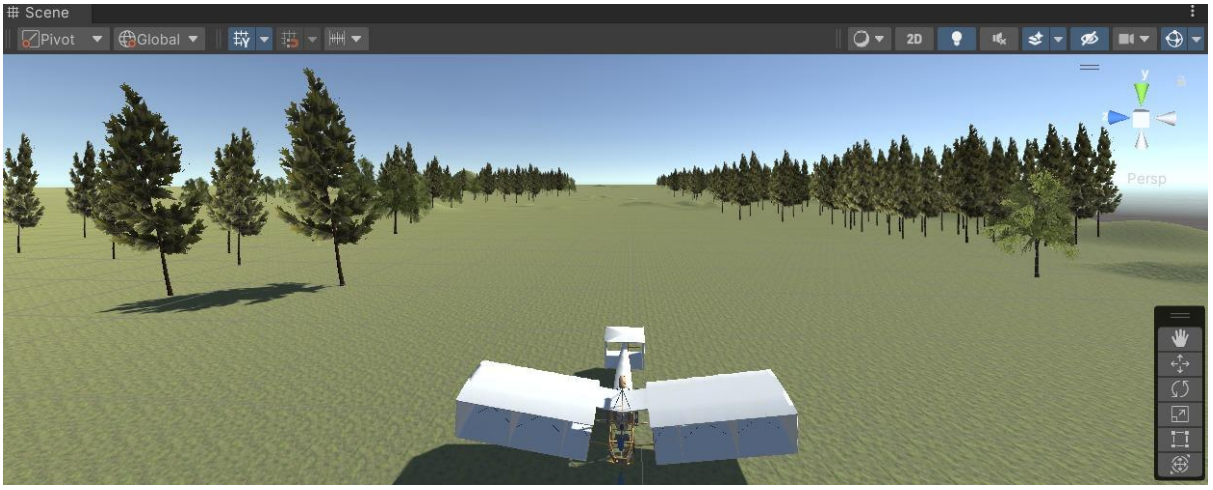


Figura 1: 14 bis antes de iniciar a animação

O terreno é composto de árvores de quatro tipos diferentes de árvores, sendo elas: "baum hd pine fbx", "baum ld0", "baum hd med fbx" e "baum ld2". Esses modelos foram adquiridos da biblioteca Unity. Para a movimentação do 14 bis optou-se por utilizar uma funcionalidade presente do ambiente Unity chamada de NavMesh. Essa funcionalidade usa inteligência artificial para escanear o terreno e detectar outros objetos. Especificamente para o voo foi utilizado um objeto Caminho que é composto de 3 planos, totalizando os 60 m percorridos. Foi removida a textura deles (o Mesh Renderer) e usando um outro objeto intitulado Objetivo- que simula o ponto final, ou seja onde o 14 bis deve pousar.

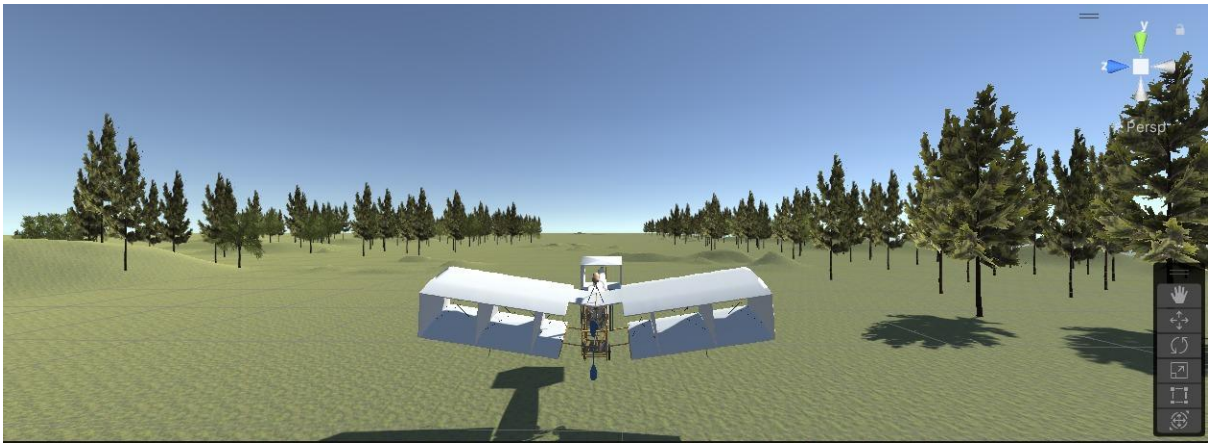


Figura 2: 14 bis com a simulação em execução

O script intitulado Colisao.cs tem por finalidade permitir que o 14 bis voe planando sobre os três planos sem a renderização de forma suave. Ele é

responsável pela movimentação do 14 bis e pela mudança de ângulo do mesmo, fazendo o eixo z aumentar sua inclinação a uma taxa de 0.04 graus, e a mesma quantidade para a descida. O script Movement.cs é responsável pelo aumento da velocidade gradual do 14 bis- assim como ocorreu o voo original- fazendo esta aumentar durante a maior parte do percurso,mais precisamente entre os 18 e 50 s de simulação.

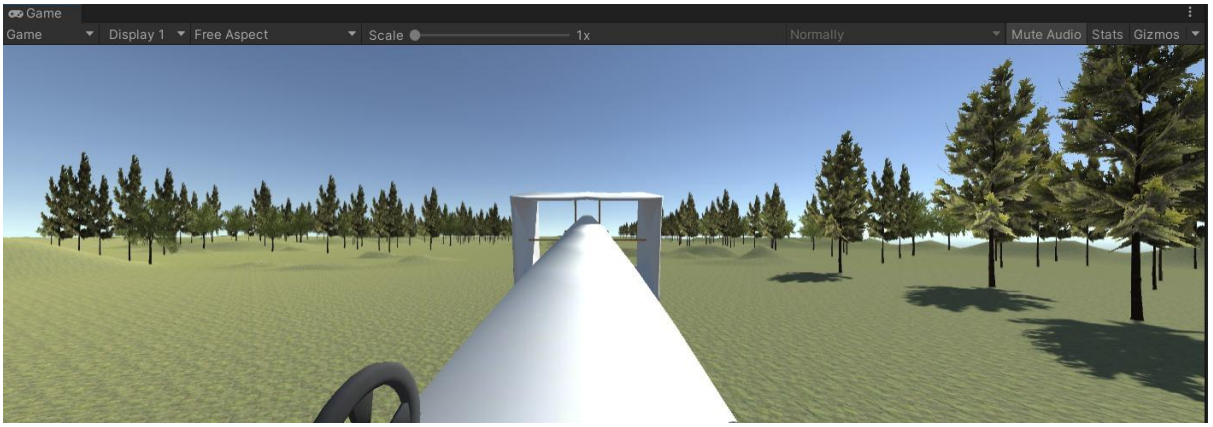


Figura 3: Simulação da visão de Santos Dumont durante o voo

- **Resultados e Discussão**
  Uma animação contundente e realista está sendo desenvolvida e finalizada no projeto em questão, cabendo refinações apenas na dinâmica do voo, da animação dos curiosos e da paisagem da época de Paris do início do século XX.

- **Conclusões**

O objetivo principal do projeto foi atingido: a animação do primeiro voo de Santos Dumont. Por fim, para obtermos um modelo fidedigno e mais realista é necessário um refinamento nas animações das pessoas presentes no campo de Bagatelle, na dinâmica do voo e uma paisagem que remonte mais a Paris da época. Tal desenvolvimento trouxe o conhecimento necessário ao aluno pra o desenvolvimento de outros projetos mais complexos, o que será explorado nos próximos períodos da bolsa PIBIC/LNCC.

Previsão meteorológica utilizando métodos de inteligência artificial

Júlia Neumann Bastos[1], **Rafael S. Pereira**[2], **Fabio Porto**[1]

[1]DEXL – Laboratorio Nacional de Computação Cientifica (LNCC)
Av Getulio Vargas 333 Petropolis Brasil

[2]Data Science Department – Just A Little Data
São Paulo, Brazil.

julia@lncc.br, rafael.pereira@just.bi, fporto@lncc.br

Bolsa de Iniciação Científica, no período de 1/9/2022 a 31/8/2023

1. Objetivos

O objetivo desse trabalho é avaliar a qualidade de modelos preditivos baseados em redes neurais profundas para o problema de previsão meteorológica. Com base nisso, avaliaremos o desempenho do modelo em estudo, para diferentes regiões do espaço. Caso haja regiões onde o desempenho se mostre insatisfatório, investigaremos suas limitações para predições nessas regiões tentando superar os problemas encontrados.

2. Introdução

As mudanças climáticas se referem a transformações de longo prazo nos padrões de temperatura e clima na terra. Como se tem observado, tais mudanças colocam em risco a vida no planeta, acarreta no desalojamento de de comunidades inteiras, assim como perdas em safras agrícolas, afetando financeiramente a população. Desta forma, acompanhar tais mudanças e seus impactos no meio-ambiente é uma tarefa muito importante.

Por esses motivos, a previsão meteorológica, principalmente envolvendo eventos extremos permite o planejamento de ações que minimizem os danos à população, às propriedades e aos ambientes públicos. Tendo em vista que a execução de modelos numéricos de previsão meteorológica atualmente utilizados é extremamente custosa, e depende do processamento custoso referente à aquisição de dados [Leticia Braga Berlandi e Analice Costacurta Brandi 2022], existe interesse em modelar esse problema via métodos de aprendizado profundo.

Optamos neste trabalho pela utilização de redes neurais profundas do tipo ConvLSTM como arquitetura para predição de chuva, como adotado em [Souto et al. 2018]. De forma a adequar a distribuição espacial dos dados, alocados de forma irregular na superfície, construímos um algoritmo de pre-processamento de dados que realiza uma interpolação espacial entre os dados de pluviômetros para alcançar uma grade regular a ser utilizada como entrada no treinamento da rede neural.

O restante deste relatório encontra-se estruturado da seguinte forma. Na seção 3 descrevemos a metodologia utilizada na solução deste problema. Na seção 4 apresentamos os resultados experimentais obtidos e na seção 5 terminamos com algumas conclusões.

3. Metodologia

Nesta seção discutimos a metodologia usada para atacar este problema. Primeiramente discutiremos sobre o processo de extração dos dados. Em seguida, apresentaremos a etapa de preprocessamento de dados. Na seção seguinte, discutimos o tratamento dado à natureza esparsa dos dados, permitindo sua utilização em um modelo espaço temporal, e por fim descrevemos o processo de treinamento e avaliação do modelo.

3.1. Extração e natureza dos dados

Os dados utilizados nesse projeto foram cedidos pelo Centro de Operações do Rio (COR), no contexto da parceria de pesquisa que [Prefeitura do Rio de Janeiro 2022], com a participação do Laboratório Nacional de Computação Científica (LNCC) e do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (Cefet/RJ). A parceria tem por objetivo o aprimoramento da qualidade preditiva das condições meteorológicas da cidade. A partir dessa parceria, tivemos acesso aos dados referentes aos registros de precipitação nas estações meteorológicas do COR espalhadas pela cidade do Rio de Janeiro.

O período escolhido para estudo foi do dia 15 de fevereiro de 2022, uma terça-feira, data que marcou o acontecimento de evento extremo na cidade de Petrópolis, acarretando em inúmeras mortes e perdas materiais que assolou a cidade, na região serrana. Os dados começaram a ser extraídos a partir de 13hrs e continuaram até as 23hrs deste mesmo dia. Foram realizadas medições a cada 2 minutos. O evento começou a ocorrer a partir do meio deste período, com duração de aproximadamente 3hrs e registrou mais de 250 milímetros de chuva [Nathan Lopes - UOL Notícias 2022].

Os dados foram disponibilizados por meio de um arquivo no formato csv, que consiste em um tipo de arquivo de texto fundamental para transferência de informações. O arquivo apresenta os seguintes atributos: o código de cada estação meteorológica presente na área geográfica de Petrópolis; a Unidade Federativa (UF); o nome de cada estação; sua localização em latitude e longitude; a data e hora de cada medida; e por fim, o valor de precipitação medido e os respectivos instantes de tempo. Sendo assim, ao todo foram localizadas 21 estações meteorológicas. Destas, nem todas estavam em funcionamento e não mediram em todos os instantes de tempo, não seguindo assim seguir um padrão. Por estes motivos o preprocessamento de dados foi de extrema importância, após o estudo dos dados.

3.2. Pré-processamento

Nesta seção discutiremos processos aplicados nos dados para podermos usá-los no processo de treinamento do modelo.

O primeiro ponto observado no dataset fornecido, é que as diferentes estações presentes na região de Petrópolis não realizam todas suas medidas ao mesmo tempo. Observando a variação entre elas e tendo uma frequência miníma de captura de uma hora, resolvemos adotar esta frequência para os dados. Ao tomar esta decisão, consideramos a coluna que contêm a data e hora da medição, e arrendondamos para a hora mais próxima, após isto tomamos a média das medidas pelo par (hora, estação), assim padronizando para as estações com frequência de captura maior. Após esta etapa, fazemos um teste de valores inválidos, removendo possíveis valores negativos na medida de chuva, pois estas

indicariam defeito no sensor na hora da medida. Por fim, calculamos o número de estações minímas que funcionaram durante o período analisado, pois este seria um limitador para o numero de vizinhos utilizado no processo de interpolação discutido a seguir

3.3. A natureza esparsa dos dados e o processo de interpolação

Após obter os dados descritos na seção 3.2, temos as medidas associadas a cada estação.Porém, como pode ser observado na figura 1, as estações meteorológicas estão localizadas em apenas para alguns pontos do espaço. Desta forma, para obtermos uma grade regular necessária ao treinamento de modelos convolucionais, utilizamos um algoritmo de interpolação derivado de [Souto et al. 2018] e baseado no algoritmo KNN (K-nearest neighbors). O passo a passo do algoritmo é descrito na figura 2. Foi utilizada uma discretização produzindo uma grade regular de 500 por 500 metros, definida após discussão com especialistas da área.

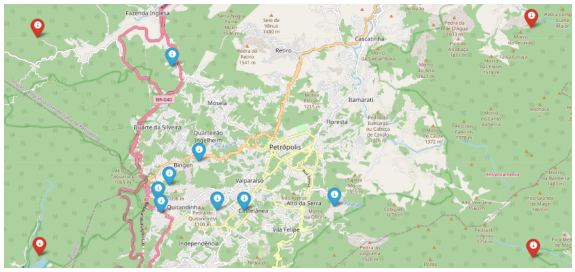


Figura 1. Grid utilizado, pontos vermelhos demonstram delimitação da cidade e pontos azuis localizações das estações

Fonte: Elaborada pelo autor.

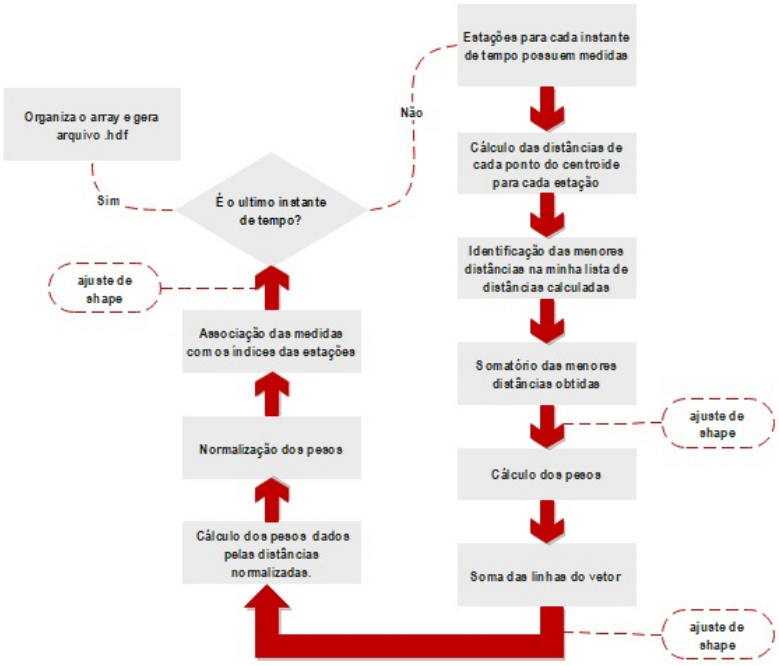


Figura 2. Processo de interpolação baseado em algoritmo KNN

Fonte: Elaborada pelo autor.

3.4. Construção do modelo preditivo

Para modelagem, nós utilizamos uma arquitetura baseada na arquitetura de redes neurais profunda ConvLSTM. A adoção desta arquitetura é inspirada no trabalho de [Souto et al. 2018], e tem como objetivo capturar as relações espaciais do fenômeno a partir da convolução enquanto captura a relação temporal utilizando a recorrência das camadas *Long Short Term memory*.

Neste tipo de arquitetura, o modelo processa não apenas o passado de múltiplas series com o modelo recorrente, mas também se utiliza da convolução para associar series localizadas espacialmente próximas para melhor entendimento do fenômeno que estamos modelando. Este tipo de modelo é normalmente utilizado em problemas espaço temporais. Como exemplo, podem-se citar problemas meteorológicos, processamento de vídeos, entre outros. Neste trabalho, utilizamos de uma segunda camada que é a convolução tridimensional. Esta tem como função colapsar as informações temporais adquiridas pela ConvLSTM para o 'espaço de medidas'. assim obtendo o tensor final na estrutura desejada. No estudo aqui apresentado, utilizamos de um modelo com duas camadas ConvLSTM, com 32 e 64 filtros, assim como, uma camada Conv3D [Tran et al. 2015], com número de filtros igual ao número de atributos modelados. Nós utilizamos de *zero padding* em todas as camadas para não ocorrer redução de dimensionalidade espacial. Uma vez que queremos projetar o futuro de toda esta região. A seguir descrevemos os processos para treinamento do modelo.

3.5. Treinamento e avaliação do modelo

Nesta seção discutimos o processo de treinamento do modelo. Ao final da etapa de interpolação, discutida na seção 3.3, produzimos um arquivo no formato *h5* [The HDF Group]. O tensor assim construído é estruturado nas seguintes dimensões: tempo, espaço, espaço e amostra, já separado em treino, validação e teste. O dado é passado para um processo *gerador* que reajusta o tensor utilizando as 10 últimas medidas para prever a próxima. Para resolver o problema de otimização dos pesos, utilizamos da biblioteca *keras* e otimizamos o modelo por 20 épocas, escolhendo o resultado que minimiza a perda de validação. Reportaremos os resultados iniciais na seção 4. Uma síntese do processo de treinamento do modelo pode ser vista na Figura 3.

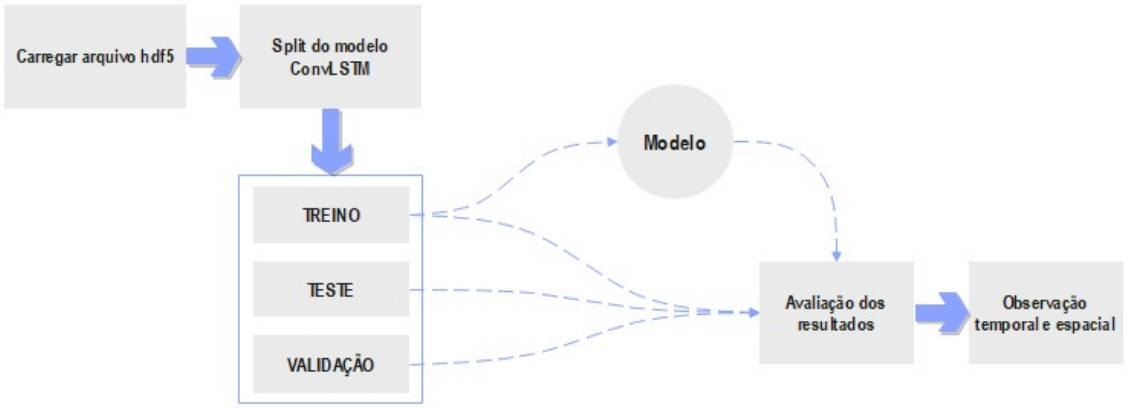


Figura 3. Processo de treinamento e avaliação do modelo

Fonte: Elaborada pelo autor.

4. Resultados e discussão

Nesta seção apresentamos os resultados iniciais obtidos a partir da avaliação do modelo. Na Figura 4 vemos a previsão inicial do modelo comparado ao valor interpolado a partir das medidas reais obtidas pelos pluviômetros utilizando as três estações mais próximas. Podemos observar que no intervalo de tempo analisado para o evento extremo, o modelo não conseguiu capturar e realizar a previsão do momento em que o evento ocorreu. No restante do tempo distribuído em um local do espaço, o modelo conseguiu seguir razoavelmente bem a curva dos valores reais observados. Isso mostra que o uso de apenas um atributo como base (por exemplo, volume de precipitação) para previsão de eventos extremos é insuficiente para se atingir uma previsão com boa qualidade, considerando-se a base de dados com os valores já interpolados e com uma medida de chuva como atributo.

Uma forma de resolver este problema pode ser incluir outros atributos que baseados em outras fontes de dados, como talvez a velocidade do vento. Eventos dessa magnitude muitas das vezes devem ser estudados por modelos multimodais, aos quais conseguiriam ver qual a relação de cada fonte no resultado obtido. A partir dos sinais de diferentes fontes de dados sobre o mesmo fenômeno poder-se-ia utilizar as medidas necessárias que tenham mais interferência sobre o evento extremo. Outra forma de observar o erro de predição se dá quanto à sua variação no espaço, como pode-se visualizar na figura 5. Podemos assim observar as limitações do modelo ao observar concentrações de erro em certas regiões do espaço, que em alguns cortes temporais fica evidente que a concentração do erro foi muito maior e cobriu toda a região, se tratando dos instantes de tempo que o evento extremo aconteceu.

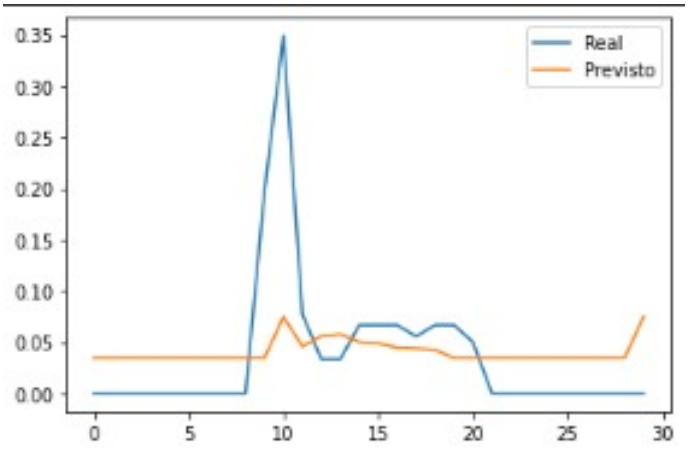


Figura 4. Previsão vs Valor Real em uma região do espaço

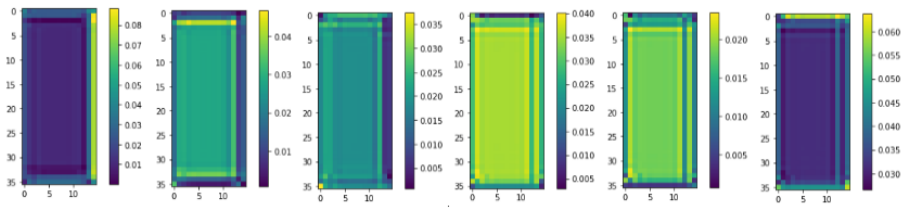


Figura 5. Distribuição espacial do erro em múltiplas janelas temporais

Fontes: Elaborada pelo autor.

5. Conclusão

Neste trabalho, todos os resultados e análises foram incluídas em um documento de trabalho de conclusão de curso. Além disso estudamos a utilização de redes neurais convolucionais na predição de eventos extremos. De maneira a preparar o dado para treinamento, uma série de atividades de pré-processamento foram executadas, incluindo o desenvolvimento de um algoritmo de interpolação espacial para dados esparsos, de forma a gerar uma estrutura em grade regular necessária ao treinamento do modelo. De forma a aprender as relações espaço-temporais das séries temporais capturadas pelos pluviômetros, foi utilizada uma arquitetura de redes *ConvLSTM*. Contudo, avaliamos que para obter boas métricas e resultados é preciso a utilização de um modelo multimodal com a inclusão de novos atributos de outras fontes de dados.

A partir disso, é possível concluir que um atributo não é suficiente para obtenção de uma boa predição, assim como uma maior quantidade de dados, uma vez que o experimento foi feito em uma pequena localidade da região serrana, onde a distribuição geográfica é composta por um território muito irregular com inúmeras montanhas formando um vale, não conseguindo ter a difusão do evento, ficando concentrado apenas naquela região. Isso mostra que inúmeras características e fontes de dados vão estar influenciando para conseguir lidar de forma ideal com os dados e eventos extremos, muito localizado.

Em trabalhos futuros, pode-se realizar a combinação de várias fontes de dados e maior estudo focalizado na região onde ocorreu o evento, com o objetivo de extrair o máximo de informações possíveis para estudo do problema, realizando assim, uma análise e auxilio na base de novos trabalhos.

Referências

Leticia Braga Berlandi e Analice Costacurta Brandi (2022). Comparação entre métodos numéricos computacionais na solução de um problema de valor inicial. Disponível em: `https://www.fc.unesp.br/Home/Departamentos/Matematica/revistacqd2228/v07a01-comparacao-entre-metodos-numericos.pdf`. Acesso em: 22 de junho 2022.

Nathan Lopes - UOL Notícias (2022). Notícia: Choveu mais que o esperado para fevereiro inteiro em 3 horas em petrópolis. Disponível em: `https://noticias.uol.com.br/cotidiano/ultimas-noticias/2022/02/16/chuva-petropolis-meteorologia.htm`. Acesso em: 18 de junho 2022.

Prefeitura do Rio de Janeiro (2022). Centro de operaÇÕes do rio. Disponível em: `http://cor.rio`. Acesso em: 13 de junho 2022.

Souto, Y. M., Porto, F., Moura, A. M., and Bezerra, E. (2018). A spatiotemporal ensemble approach to rainfall forecasting. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

The HDF Group. Arquivo hierárquico formato de dados 5. Disponível em: `https://ficheiros.com.br/extensao/h5/`. Acesso em: 20 de junho 2022.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Relatório Final de Atividades

1. Dados gerais

Título do projeto: Projeto e Implementação de *Workflows* Científicos Reprodutíveis de Alto Desempenho
Bolsista: Lucas da Cruz Silva
Orientador: Luiz Manoel Rocha Gadelha Júnior
Coorientadoras: Carla Osthoff e Kary Ocaña
Tipo de bolsa: Iniciação Científica
Período do relatório: 01/12/2019 - 03/07/2023

2. Objetivos

O presente trabalho se desenvolve a partir de um projeto multidisciplinar que envolve o Laboratório de Bioinformática (LABINFO) e o Centro Nacional de Processamento de Alto Desempenho do Rio de Janeiro (CENAPAD-RJ), no Laboratório Nacional de Computação Científica (LNCC), com o objetivo de acoplar ao Bioinfo-Portal, um portal de Bioinformática hospedado no LNCC, *workflows* científicos de Alto Desempenho. Dessa forma, através da motivação de facilitar o gerenciamento (definição, execução e monitoração) de atividades extensivas, complexas e computacionalmente custosas, este projeto tem por objetivo geral, modelar, executar e analisar *workflows* científicos que utilizam técnicas e sistemas de Computação de Alto Desempenho (CAD), tomando como caso de estudo experimentos de bioinformática.

3. Introdução

A evolução da ciência tem levado a humanidade a lidar com tarefas de alta complexidade que podem ser difíceis ou, até mesmo, impossíveis de serem realizadas manualmente. Além de tarefas complexas, hoje, há também uma crescente demanda por soluções rápidas e precisas para problemas complexos, como a compreensão de doenças e seus mecanismos ou para o desenvolvimento de novos medicamentos e antídotos [Liu et al. 2016]. No entanto, a ciência também incitou o desenvolvimento de técnicas sofisticadas, que fazem o uso de ambientes computacionais, para facilitar a criação de soluções para problemas complexos, incluindo a modelagem computacional, que permite aos cientistas a criação de modelos virtuais de sistemas complexos para estudar seu funcionamento sob diferentes condições [Brodland et al. 2015].

A modelagem computacional tem desempenhado um papel crucial no campo científico e industrial como, por exemplo, a física, a matemática, geologia, biologia, petróleo e gás, química, bioinformática e muitas outras [Mattoso et al. 2008]. Isso pois, se apresentam de forma fundamental para lidar com grandes quantidades de dados e realizar tarefas complexas, oferecendo eficiência, precisão, escalabilidade, redução de custos, flexibilidade e, até mesmo, inovação, pois permitem a criação de novas tecnologias e soluções para problemas complexos, aliando diferentes áreas de estudo. No entanto, há procedimentos experimentais, provindos de uma simulação real ou não, que processam um grande volume de dados e por essa razão exigem um alto poder computacional. A bioinformática, por exemplo, é uma área que se utiliza de algoritmos e técnicas de análise de dados para examinar grandes conjuntos de dados biológicos. Em específico, a análise de dados de sequenciamento RNA, pode ser usada para obtenção de informações valiosas sobre as características de uma doença e identificar possíveis alvos terapêuticos [Ness et al. 2022]. Em casos como esse, é que se faz necessário o uso de estratégias computacionais sofisticadas para gerar soluções de alta performance.

É fato que existem várias formas de realizar experimentos científicos em sistemas de alto desempenho e tudo depende da necessidade e complexidade do experimento. Há, por exemplo, procedimentos experimentais em que o cientista não tem dificuldades em lidar com o fluxo de dados, nem com as etapas experimentais ou com a utilização de recursos. Mas, há outros, em que o custo de administrar todas as etapas experimentais é tão alto que exige atenção redobrada do cientista. Para isso existem os *workflows* científicos, que são formas de descrever, totalmente ou parcialmente, o fluxo das etapas e a dependência de dados de um experimento computacional [Gadelha et al. 2012].

Nesse âmbito, *workflows* científicos podem ser utilizados para automatizar a execução de tarefas repetitivas e complexas, tornando o processo de análise mais eficiente e reprodutível. Eles permitem a aplicação de métodos computacionais de forma estruturada e organizada, usando estratégias de CAD, como paralelismo, para otimizar o uso dos recursos computacionais disponíveis, minimizando o tempo de execução e maximizando a utilização de outros recursos como processador e memória. Dessa forma, se faz necessário a utilização de *workflows* científicos de alta performance. São uma solução valiosa para a condução de experimentos computacionais em larga escala, tornando possível a realização de análises mais sofisticadas, com a obtenção de resultados de forma mais rápida e mais precisa.

O presente trabalho apresenta um *workflow* da área de bioinformática e se baseia em um experimento real de sequenciamento de Ácido Ribonucleico (*Ribo-Nucleic Acid*, também conhecido como RNA). Assim como sua descrição, os resultados obtidos desse estudo foram publicados e apresentados em anais de eventos, revistas, periódicos e foi objeto de estudo para o trabalho de conclusão de curso do bolsista em questão. Este relatório apresenta o ganho de desempenho e tempo de execução com a utilização de técnicas e estratégias de CAD, mostrando uma melhora de cerca de 76% no tempo de processamento da versão final do *workflow* desenvolvido, desde o início da pesquisa.

Este relatório está organizado da seguinte forma: a Seção 1, apresenta informações gerais sobre o relatório; a Seção 2 traz o resumo e objetivo geral do projeto de pesquisa; na Seção 3 a presente introdução, bem como um breve resumo do que será apresentado; a Seção 4 traz a metodologia adotada na modelagem do *workflow* de RNA-Seq desenvolvido e sua descrição; na Seção 5 são apresentados os resultados e análises experimentais; e, por fim, a Seção 6 traz as considerações finais e os trabalhos futuros.

4. Metodologia

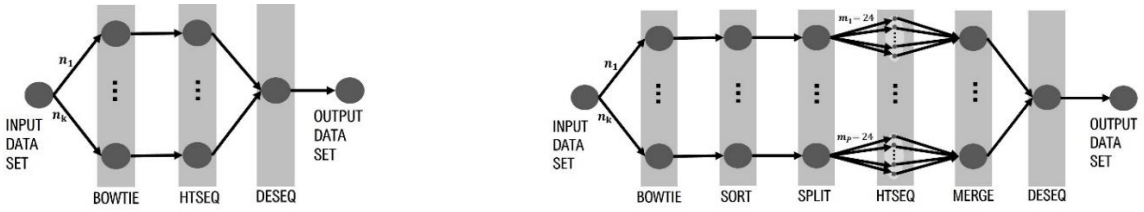
A modelagem computacional de um *workflow* científico requer o mapeamento bem definido das atividades ou etapas experimentais, a fim de fazer uma abstração de todos os procedimentos que seriam realizados por um cientista, sem a utilização de um ambiente computacional para automatizar os procedimentos. Em outras palavras, a modelagem requer o conhecimento não só do experimento que se realiza, como também de todo o fluxo, das atividades componentes utilizadas, das dependências existentes entre as atividades e dos dados gerados. Todo o desenvolvimento foi feito na linguagem de programação Python e utilizou o Parsl [Babuji et al. 2019], como principal biblioteca para descrever as atividades do *workflow*. Essa biblioteca foi desenvolvida para permitir a paralelização e encadeamento de tarefas, ou *tasks*, mantendo a dependência dos dados gerados durante a execução em ambientes de CAD.

4.1. Modelagem do *workflow* de RNA-Seq

O *workflow* de RNA-Seq desenvolvido, chamado ParslRNA-Seq, possui duas versões: a primeira, denominado neste trabalho como α-ParslRNA-Seq [Cruz et al. 2020], é composta de três atividades componentes (bowtie, htseq e deseq), que são consideradas as principais do *workflow*; e, a segunda versão, denominado neste trabalho como β-ParslRNA-Seq [Cruz et al. 2021], que se compõe de seis atividades componentes, o que inclui as três atividades da primeira versão (bowtie, htseq e deseq) e, adicionalmente, mais três atividades (sort, split e merge).

A distinção entre a primeira e a segunda versão do *workflow* não se restringe à quantidade de atividades incorporadas, mas também inclui a diferença de desempenho entre os dois modelos. O experimento empregado é idêntico em ambos os casos, utilizando os mesmos dados. No entanto, a primeira versão contém apenas os procedimentos indispensáveis para a realização do experimento de RNA-Seq, enquanto a segunda versão enfatiza a otimização do desempenho computacional da aplicação.

Cada atividade componente se refere a um *software*, algoritmo ou pacote de bioinformática, e, portanto, biologicamente, todas tem um papel a ser desempenhado no experimento: bowtie, executa o software Bowtie2 e mapeia as leituras curtas do genoma; sort, executa o software SAMTools e faz a ordenação das leituras; split, executa o programa Picard e divide os arquivos de leituras em várias subpartes; htseq, executa o HTSeq fazendo a contagem das leituras mapeadas de cada gene; merge, é um algoritmo usado indexar em um único arquivo as contagens das leituras relativas a uma única amostra; e, por fim, deseq, executa o pacote DESeq2 e aplica estatísticas em cima das contagens para a análise da Expressão Diferencial de Genes (EDG).



(a) α-ParslRNA-Seq (três atividades). (b) β-ParslRNA-Seq (seis atividades).

Figura 1. Modelagem conceitual do *workflow* científico de RNA-Seq.

4.2. Parametrização das atividades componentes do *workflow* de RNA-Seq

O tempo de execução de cada atividade componente é um fator importante que pode vir a determinar a eficiência do *workflow*. Por essa razão, a análise avalia também o desempenho dos *softwares*, algoritmos e pacotes de bioinformática

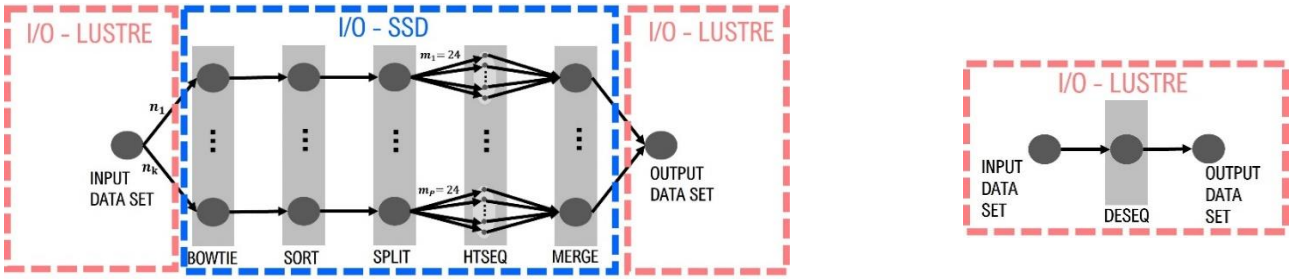
utilizados, de forma individual, antes de realizar a execução do *workflow*. Alguns deles possuem recurso de parametrização para execução paralela, são eles: Bowtie2 e sort (SAMTools), que utilizam threads; e, HTSeq, que realiza uma execução paralela em múltiplos cores. A função split, do Picard, o algoritmo Merge-HTSeq e o DESeq2 executam somente em modo serial.

4.3. Cenários de execução do *workflow* de RNA-Seq

Após a avaliação do desempenho das atividades componentes, finalmente, é possível seguir para a etapa de execução completa do *workflow* desenvolvido. As execuções acontecem em dois cenários: uma *single node*, com utilização de um nó de processamento e a utilização do *ThreadPoolExecutor* do Parsl, que possui recursos de paralelização de *tasks*; e, a outra *multiple nodes*, com a utilização de um conjunto de nós e a utilização do *HighThroughputExecutor* do Parsl, indicado para ambiente multinó. A utilização de um nó computacional é indicada para a etapa inicial de estudos em ambientes de CAD, uma vez que essa configuração oferece um ambiente de execução simplificado, com um conjunto menor de variáveis, o que facilita a compreensão e projeção do comportamento do *workflow* em um cenário distribuído em múltiplos nós. Além disso, essa é uma abordagem útil para avaliar e otimizar o desempenho pois possibilita a realização controlada do experimento e a identificação de gargalos de desempenho antes de executar o *workflow* em escala multinó.

Para auxiliar nas análises de desempenho nos cenários *single node* e *multiple nodes* foi utilizado o VTune *Profiler*, da Intel. Com ele é possível verificar a utilização dos recursos do sistema durante a execução, o que inclui uso de memória, tempo de execução por número de CPUs (*Central Process Units*) e distribuição do número de CPUs utilizadas durante a execução da aplicação.

Por fim, nas execuções com múltiplos nós, foi comparado o melhor formato de execução usando o Sistema de Arquivos Paralelo (*Parallel File System,* PFS), implementado no Lustre, e o armazenamento de disco local que utiliza tecnologia *Solid Sate*, para realização das operações de I/O (*Input/Output*). A proposta no formato de execução do *workflow* se refere a: forma de alocação dos nós computacionais, alocando um nó para executar cada *pipeline* de tarefas; a desacoplagem da atividade que detém a barreira de sincronização de dados (deseq); e, a utilização do armazenamento local, para não exigir que sejam feitas transferências de dados para o PFS durante todo o tempo de execução. Isso irá permitir que os recursos computacionais alocados para executar o *workflow* sejam liberados mais cedo.



(a) Primeira parte (cinco atividades). (b) Segunda parte (uma atividade).

Figura 2. Nova proposta para processamento do ParslRNA-Seq.

4.2. Configuração de ambiente computacional

O ambiente computacional utilizado para as execuções e as análises de desempenho é o supercomputador SDumont, um dos supercomputadores mais poderosos da América Latina. Ele possui uma capacidade de processamento de 5.1 Petaflop/s, com 34.688 CPU *multicores* distribuídas em 1.132 nós computacionais que são interligados por uma rede de interconexão *Infiniband* FDR/HDR. Os nós computacionais onde foram realizadas as execuções possuem duas CPUs Ivy Bridge Intel Xeon E5-2695v2 (12c @2.4GHz) e 64GB de memória RAM e uma GPU Nvidia K40 e SSD de 128 GB. O Lustre é implementado através do ClusterStor 9000 v3.3 da Cray/HPE.

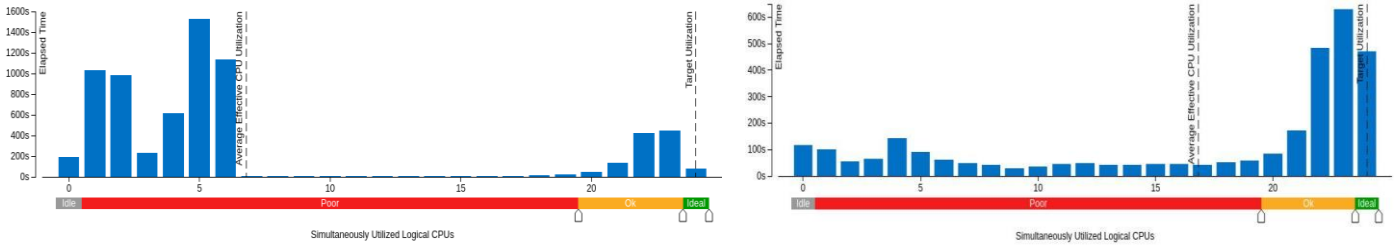
4.3. Dados de entrada

Os dados de entrada pertencem a um experimento real de RNA-Seq e foram extraídos do repositório público *Gene Expression Omnibus* (GEO), são eles: o genoma de referência; o arquivo de características genômicas no formato GTF (*Gene Transfer Format*); os arquivos de sequenciamento no formato FASTQ e as condições experimentais da amostra. Esses dados de RNA-Seq se referem à via metabólica de sinalização transcricional Wnt, a que tem sido reportada com um papel regulatório no coração. Um total de seis arquivos de sequenciamento, divididos em dois grupos de três, controle e condição, sequenciadas no *Mus Musculus* (rato) foram usados como dados de entrada pelo ParslRNA-Seq.

Há de se notar que o número de arquivos de sequenciamento e a divisão dos grupos podem variar conforme o experimento. Então, a análise de desempenho está ligada somente a esses seis arquivos, onde dois deles possuem o tamanho de 1.8 GB e os outros 1.9 GB, 2.0 GB, 2.5 GB e 3.0 GB, totalizando 13 GB.

5. Resultados e discussão

5.1 α-ParslRNA-seq e β-ParslRNA-Seq em execuções *single node*



(a) α-ParslRNA-seq. (b) β-ParslRNA-Seq.

Figura 3. Eficiência na utilização de CPUs utilizando um nó computacional.

Como destacado na subseção 4.1 é imprescindível levar em consideração que as atividades da nova versão foram incluídas somente em função da melhora de desempenho do *workflow*. E isso pode ser observado pela Figura 3, onde é apresentado o histograma da eficiência na utilização das CPUs. No cenário de execução proposto foi utilizado um nó computacional e com a utilização do Parsl foi possível realizar uma paralelização de tarefas dentro do nó. A grande diferença que se dá entre os histogramas da Figura 3 é que: na Figura 3(a) há somente paralelização *multithreading* dos processos do software Bowtie e o HTSeq está utilizando somente um núcleo de CPU para executar cada amostra. Então, apesar da execução paralelizada de tarefas realizadas pelo Parsl ainda existe um número considerável de CPUs ociosas na maior parte da execução do *workflow*; agora, na Figura 3(b) é possível observar uma maior distribuição na utilização no número de núcleos de CPUs, aumentando, consideravelmente, o número de CPUs utilizadas na maior parte do tempo de execução e levando a um fator ideal de utilização de recursos computacionais. Isso se deve ao fato de ter mais uma atividade fazendo execução *multithreading*, a atividade *Sort*, e principalmente pela execução paralelizada em múltiplos *cores* do HTSeq.

5.1 α-ParslRNA-seq e β-ParslRNA-Seq em execuções *multiple nodes*

A principal razão da execução desse *workflow* dentro de um cenário paralelo e distribuído é fazer com que cada amostra seja processada dentro de um nó, dessa forma, teoricamente, seria alcançado um melhor desempenho. No entanto, para a primeira versão do ParslRNA-Seq (α-ParslRNA-seq), não foram observados ganhos expressivos no TTE. Como destacado na Figura 4, a linha tracejada indica a execução da primeira versão do *workflow* e o que se constata é que nem mesmo aumentando o número de nós o TTE reduz. Isso ocorre, pois, apesar de haver distribuição das amostras entre os nós, não há uma execução paralela na atividade htseq e, além disso, há também uma barreira para sincronização de dados na altura da atividade DESeq. Esses fatores farão a amostra de maior tamanho segurar, praticamente, todo o tempo de execução do *workflow* na atividade htseq.

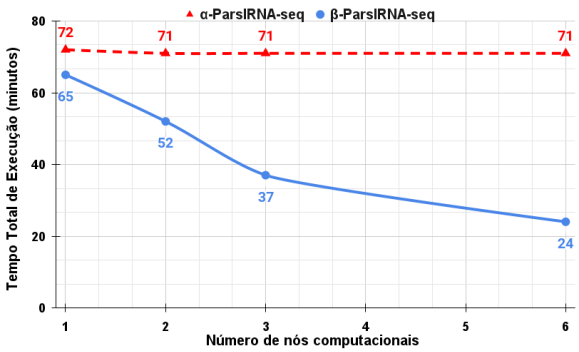
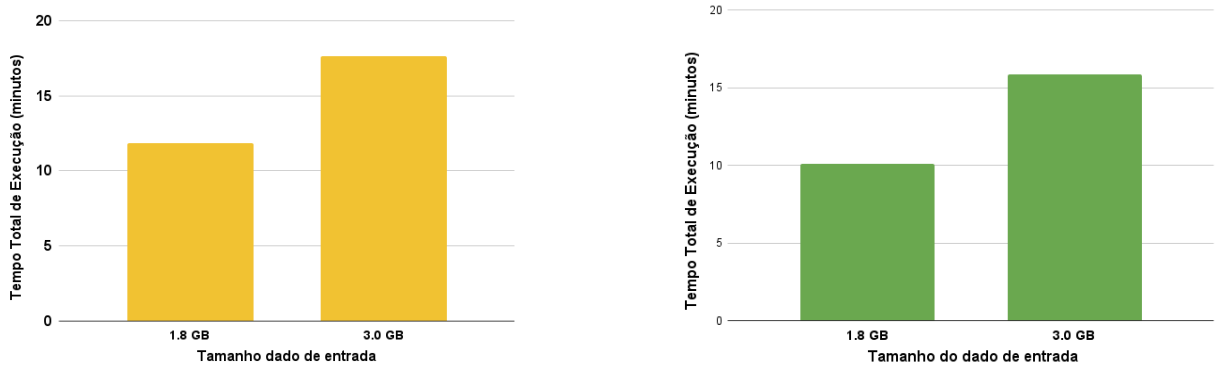


Figura 4. Execução das versões do *workflow* desenvolvido em múltiplos nós.

No β-ParslRNA-Seq, há uma distribuição de processamento entre os núcleos para o HTSeq. O que não impossibilita que a amostra de maior tamanho segure a execução do *workflow*, mas esse tempo é consideravelmente menor, como pode ser observado na linha contínua da Figura 4, que representa a execução da segunda versão do *workflow*. Com a redução de núcleos de CPUs ociosas, por consequência, há uma redução no TTE do *workflow* de 72 para 65 minutos,

quando o número de nós é igual a um, como pode ser observado na Figura 4. É imprescindível observar que a redução do TTE não ocorre de forma tão significativa pois o aumento da paralelização, aumenta também a competição para utilização dos recursos computacionais, para um cenário de um nó computacional. Com isso, o melhor TTE dentro do cenário multinó pode ser observado pela segunda versão do ParslRNA-Seq (β-ParslRNA-Seq), na Figura 4, em 6 nós, onde o TTE marca 24 minutos.

5.2 Análise comparativa de execuções entre uso de SSD e Lustre



(a) Sistemas de arquivos paralelos (PFS). (b) Armazenamento local;

Figura 5. Execuções pela estratégia no uso do PFS e do armazenamento local para operações de I/O.

Nesse cenário é possível ter como base o tempo do arquivo de maior tamanho, de 3 GB, que levará maior tempo para ser processado, e, portanto, o tempo dele determina o TTE do *workflow*. Pela Figura 5, o tempo médio de execução da primeira parte do *workflow* usando o sistema de arquivos paralelos é de cerca de 17 minutos. Já usando o armazenamento local, a execução dura em média cerca de 15 minutos. O tempo médio da atividade deseq é de cerca 1,4 minutos. Ou seja, usando o sistema de arquivos, o *workflow* leva cerca de 19 minutos para finalizar a execução e usando o armazenamento local, leva cerca de 17 minutos. É possível ainda notar, pelo tempo de execução do arquivo de menor tamanho, de 1.8 GB, que através da estratégia de alocações de nós, citada na seção 4.3, um nó computacional é liberado cerca de 6 minutos mais cedo em relação ao arquivo de maior tamanho.

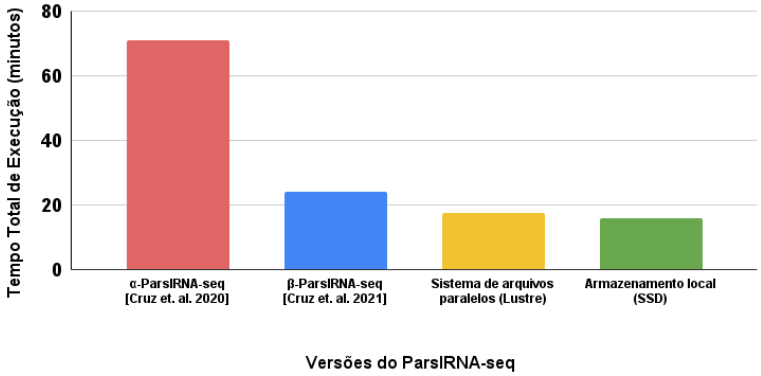


Figura 6. TTE alcançado das estratégias aplicadas no ParslRNA-seq.

Pela Figura 6, é possível comparar os ganhos em tempo de execução através da exploração de utilização dos recursos computacionais para execução do ParslRNA-seq. O uso do sistema de arquivos paralelos para realização das operações de I/O obteve ganho de cerca de 73% em relação ao α-ParslRNA-seq e 20% em relação ao β-ParslRNA-seq. Já o uso do armazenamento local para realização das operações de I/O obteve ganho de cerca de 76% em relação ao α-ParslRNA-seq e 29% em relação ao β-ParslRNA-seq.

6. Conclusão

O presente trabalho apresentou o desenvolvimento, otimizações e análises de desempenho de um *workflow* científico otimizado para ambientes computacionais de alta performance, chamado ParslRNA-seq. Para isso foi utilizado um caso de estudo de expressão diferencial de genes em experimentos de sequenciamento RNA, da bioinformática. As análises exploram o desempenho do *workflow* com o uso de técnicas de paralelismo como *threads*, paralelismo de *tasks* e paralelismo de dados.

Além disso, também é realizada uma exploração do desempenho no uso do sistema de arquivos paralelos disponível (Lustre) e no uso de armazenamento local, que conta com SSDs, para realização de escrita e leitura de dados. Os resultados da modelagem do *workflow* da primeira versão desenvolvida, α-ParslRNA-seq, deixam claro a necessidade do uso de ambientes de computação de alto desempenho, pois o mesmo fluxo no ambiente do Galaxy levou cerca de 3 dias para finalizar sua execução, enquanto com o uso do supercomputador Santos Dumont o tempo foi reduzido para cerca de 11 horas em uma execução sequencial. Já quando o *workflow*, α-ParslRNA-seq, é configurado para utilizar paralelismo de *threads* e de tarefas, o TTE passa para, aproximadamente, 72 minutos. O α-ParslRNA-seq não apresenta ganhos significativos em um cenário de execução distribuída.

Com a aplicação da abordagem de divisão e conquista para utilização do paralelismo de dados na atividade htseq, os resultados da modelagem do *workflow* da segunda versão desenvolvida, β-ParslRNA-seq, apresentaram uma melhora de cerca de 70% no tempo computacional, em relação ao α-ParslRNA-seq para uma execução serial, mesmo que tenham sido adicionadas mais três atividades. Em um cenário de execução paralela e distribuída o *workflow* passa de 65 minutos com apenas 1 nó computacional, para 24 minutos com 6 nós.

O β-ParslRNA-seq passou então, a se beneficiar com a estratégia de execução proposta para otimização, que se baseia no uso integrado do Lustre e do SSD disponível nos nós de computação. Ele também se beneficia do desacoplamento da atividade bloqueante (deseq) e na forma de alocação dos recursos computacionais, que agora, passa alocar um nó para processar uma pipeline. Essa estratégia, não só aumentou o *throughtput* do *workflow*, como também permitiu um uso racional dos recursos computacionais, evitando ociosidade no sistema. A exploração desses fatores, levaram a uma redução no TTE do *workflow* de cerca de 24 para 19 minutos usando o sistema de arquivos paralelos e 17 minutos usando o armazenamento local, com um total de 6 nós de computação. Além da liberação de recursos 6 minutos mais cedo antes do fim do processamento do total *workflow*.

Como passos futuros podem ser realizadas: Comparações de desempenho em arquiteturas distintas, pois todo o estudo apresentado neste trabalho foi apenas fazendo utilização dos nós Base do SDumont, outras máquinas podem admitir diferentes desempenhos para o ParslRNA-seq; Paralelismo para divisão dos dados na atividade split, pois ela não realiza processamento paralelo e, atualmente, apresenta o custo mais elevados quando comparada as demais atividades do *workflow*; Inclusão de métodos para realização de outros experimentos de RNA-seq, isso pode trazer maior robustez ao *workflow* e possibilidade de ser uma ferramenta generalizada, para outro tipo de experimento de RNA-seq; e por fim, melhorias na proveniência dos dados gerados pelo *workflow*, pois foi um ponto menos assistido no desenvolvimento do presente trabalho.

Por fim, o ParsRNA-Seq se encontra disponível no GitHub (https://github.com/lucruzz/RNA-seq) e no Docker Hub (https://hub.docker.com/r/lucruzz/parslrna-seq) para toda a comunidade científica e está sendo integrado ao Bioinfo-Portal, um portal de bioinformática hospedado no LNCC e acoplado ao ambiente computacional do SDumont.

7. Referências bibliográficas

Babuji, Y., Woodard, A., Li, Z., Katz, D. S., Clifford, B., Kumar, R., Lacinski, L., Chard, R., Wozniak, J., Foster, I., Wilde, M., and Chard, K. (2019). *Parsl: Pervasive parallel programming in python. In28th ACM International Symposium on High-Performance Parallel and Distributed Computing* (HPDC)

Brodland, G. W. *How computational models can help unlock biological systems. Seminars in Cell Developmental Biology*, v. 47-48, p. 62–73, 2015. ISSN 1084-9521. *Coding and non-coding RNAs Mammalian development.*

Cruz, L., Coelho, M., Gadelha, L., Ocaña, K., and Osthoff, C. (2020). Avaliação de desempenho de um *workflow* científico para experimentos de rna-seq no supercomputador santos dumont. *In* Anais Estendidos do XXI Simpósio em Sistemas Computacionais de Alto Desempenho, *pages* 86–93, Porto Alegre, RS, Brasil. SBC.

Cruz, L., Coelho, M., Terra, R., Carvalho, D., Gadelha, L., Osthoff, C., and Ocaña, K. (2021). *Workflows* científicos de rna-seq em ambientes distribuídos de alto desempenho: Otimização de desempenho e análises de dados de expressão diferencial de genes. *In* Anais do XV *Brazilian e-Science Workshop*, *pages* 57–64, Brasil. SBC.

Gadelha, L. Gerência de proveniência em *Workflows* científicos paralelos e distribuídos. Tese (Doutorado) - UFRJ,2012.

L. Cruz, M. Coelho, M. Galheigo, A. Carneiro, D. Carvalho, L. Gadelha, F. Boito, P. Navaux, C. Osthoff, K. Ocaña. Parallel Performance and I/O Profiling of HPC RNA-Seq Applications. Computación y Sistemas, vol. 26, n. 4, 2022.

Liu, T. et al. *Applying high-performance computing in drug discovery and molecular simulation. National science review, Oxford University Press*, v. 3, n. 1, p. 49–63, 2016

Mattoso, M. et al. Gerenciando experimentos científicos em larga escala. SBC-SEMISH, v. 8, p. 121–135, 2008.

Ness, T. E.; Dinardo, A.; Farhat, M. R. *High throughput sequencing for clinical tuberculosis: An overview. Pathogens*, MDPI, v. 11, n. 11, p. 1343, 2022.

RELATÓRIO DE ATIVIDADES

Título do Projeto: Modelagem e Integração de Bancos de Dados Relacionais na arquitetura do Bioinfo-Portal.
Nome: Marco Antonio Silva Cabral
Orientador: DSc. Kary Ann del Carmen Ocaña Gautherot
Coorientador: DSc. Antonio Tadeu Azevedo Gomes
Coorientador: Marcelo Monteiro Galheigo
Tipo de bolsa: PIBIC
Período do relatório: 04/07/2022 - 03/07/2023

O Projeto de Iniciação Científica (IC) se desenvolve sob coordenação da orientadora Kary Ocaña e coorientadores Antônio Tadeu Gomes e Marcelo Galheigo do Laboratório Nacional de Computação Científica, com bolsa de IC financiada pelo CNPq. Um objetivo importante da Rede Nacional de Bioinformática (RNBio) em conjunto ao LNCC é o desenvolvimento de ferramentas e arquiteturas para a bioinformática. Fruto dessas pesquisas em colaboração, foi desenvolvido o *gateway* Bioinfo-Portal (https://bioinfo.lncc.br/) que visa a execução de aplicações de bioinformática em larga escala, no apoio às pesquisas da comunidade científica de bioinformática [1]. Bioinfo-Portal está hospedado e gerenciado pelo LNCC e usa recursos e tecnologias de computação de alto desempenho, como o supercomputador Santos Dumont (SDumont, https://sdumont.lncc.br/) a fim de diminuir o grande tempo de processamento das execuções [2]. Bioinfo-Portal gerencia a execução automática de aplicações e dados científicos através de uma interface *Web* amigável e iterativa e das diversas camadas de *software* do *gateway*. Bioinfo-Portal utiliza, via serviços *Web* RESTful, o *middleware* CSGrid como *framework* de integração à arquitetura do SINAPAD. Atualizações e otimizações do Bioinfo-Portal na camada de banco de dados e de gerência de execuções irão fornecer uma melhor funcionalidade e escalabilidade de processos de execuções e armazenamento de dados de proveniência, tal que auxiliem na tomada de decisões inteligentes no uso de recursos computacionais [3].

OBJETIVOS

- Atualização das camadas de banco de dados e de gerência de execuções das aplicações de bioinformática em ambientes de CAD
- Desenvolvimento de serviços específicos para integrar informações científicas e dados de proveniência das diversas camadas da arquitetura do Bioinfo-Portal.

METODOLOGIA

Na primeira etapa, foi utilizado o PostgreSQL v10 como Sistema de Gerência de Banco de Dados (SGBD) relacional *Open Source* e o pgAdmin v5.2 como plataforma de gerência. O processo de conexão e implementação do banco de dados foi realizado via SSH a partir da liberação VPN do LNCC.

A segunda etapa envolve a utilização de serviços RESTful para o desenvolvimento de sistemas de inteligência na coleta de dados do *gateway*, sistemas de autenticação de usuário e mapeamento de dados das aplicações presentes na arquitetura do Bioinfo-Portal. A linguagem de programação utilizada é a PHP (*Hypertext Preprocessor*) e o *Visual Studio Code* como o editor de código-fonte usado para o desenvolvimento dos sistemas.

RESULTADOS E DISCUSSÃO

Baseado no referencial bibliográfico sobre gerência e modelos de banco de dados e na arquitetura do Bioinfo-Portal e do SINAPAD, foi implementado o modelo conceitual de banco de dados do Bioinfo-Portal. Iniciou-se a etapa de mapeamento dos dados da arquitetura do *gateway* para a implementação do modelo lógico. A Figura 1 apresenta a estrutura do banco de dados atualizada do Bioinfo-Portal que visa melhorar o armazenamento de dados de proveniência.

Estudo de Caso: Mapeamento do modelo conceitual ER

Atualmente, o banco de dados possui 15 entidades, *Files* e *Executions* são as únicas entidades originais do Bioinfo-Portal (Figura 1). Todas as demais entidades da nova versão desenvolvida do banco de dados estão disponibilizadas no servidor de desenvolvimento no SINAPAD.

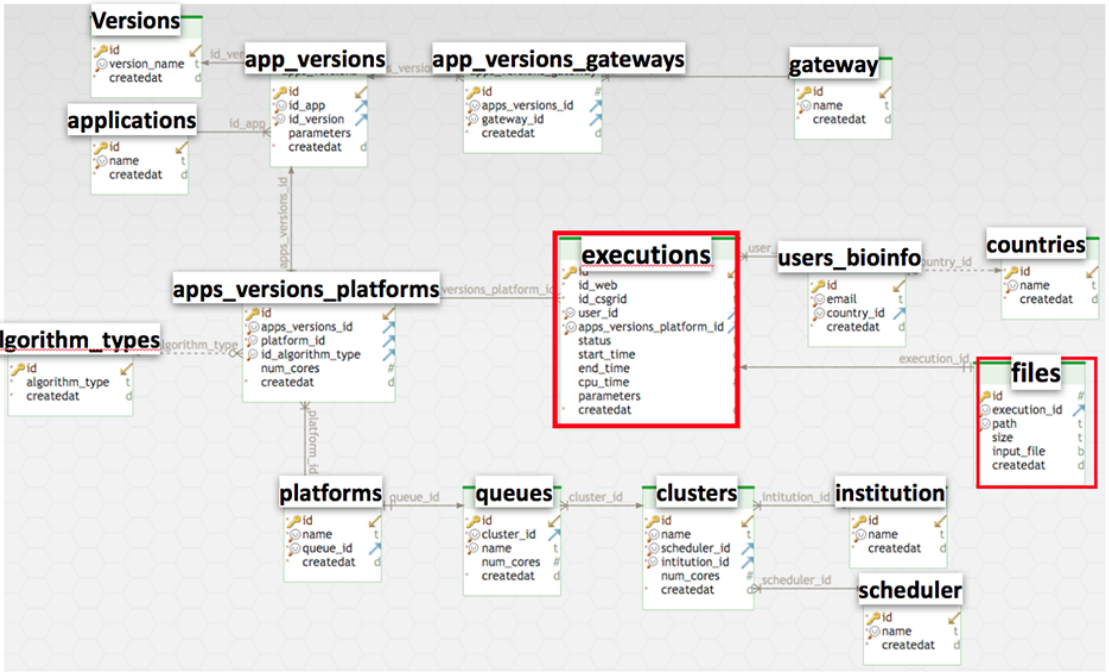


Figura 1. Diagrama Entidade-Relacionamento do Banco de Dados Bioinfo.

As entidades *Gateway*, *Institution* e *Scheduler* (mostrado na Figura 1) foram desenvolvidas para o melhor armazenamento de dados provenientes da estrutura do *gateway* que possuem esses tipos de informações.

A partir do acesso aprovado ao *VPN* do LNCC, foi possível acessar a máquina onde o PostgreSQL está instalado via SSH e implementar o modelo de banco de dados (mostrado na Figura 1) do Bioinfo-Portal. Após a implementação, foi realizado o desenvolvimento de sistemas utilizando serviços *Web* RESTFul, que interage dinamicamente com o *middleware* CSGrid do SINAPAD.

O primeiro sistema desenvolvido (Algoritmo 1) é o de autenticação, podendo tanto ser um método LDAP quanto *RSA*, armazenando dados de proveniência de usuários como nome e a identificação e servindo como base para futuros desenvolvimentos de outros sistemas.

```
$url=('http://.../rest/op***');
$data = array(
'username'=> '****',
'password'=> '*****',
'service'=> '****',
'uuid' => $uuid );
$headers = array(
'Accept: application/json');
$handle = curl_init();
curl_setopt($handle, CURLOPT_URL, $url);
curl_setopt($handle, CURLOPT_HTTPHEADER, $headers);
```

```
curl_setopt($handle, CURLOPT_RETURNTRANSFER, true);
curl_setopt($handle, CURLOPT_SSL_VERIFYHOST, false);
curl_setopt($handle, CURLOPT_SSL_VERIFYPEER, false);
curl_setopt($handle, CURLOPT_POST, true);
curl_setopt($handle, CURLOPT_POSTFIELDS, http_build_query($data));
$response = curl_exec($handle);
$obj = json_decode ( $response );
$b = $obj->{'uuid'};
```

Algoritmo 1: Sistema de autenticação LDAP.

O segundo sistema desenvolvido (Algoritmo 2) se refere ao mapeamento dos dados de nome, fila, plataforma, versões, *gateway*, *cluster*, escalonador das aplicações contidas na arquitetura do gateway. Dados esses que serão inseridos no Banco de dados do Bioinfo-Portal nas entidades *Applications*, *Queues*, *Plataforms*, *Versions*, *Gateways*, *Clusters* e *Scheduler* respectivamente. A integração desse mapeamento de dados permitirá extrair informações de execuções e melhor armazenamento de informações das aplicações do Bioinfo-Portal.

```
foreach ($objN["elements"]["element"] as $valor){
      $banconom=$valor["name"];
      pg_query($conn, "INSERT INTO teste.applications (name) VALUES
('$banconom')");

foreach ( $valor["versions"]["version"] as $j){
      $bancovers=$j["version"];
      pg_query($conn, "INSERT INTO teste.versions (version_name) VALUES
('$bancovers')");
      $f=pg_query($conn, "SELECT id FROM teste.applications WHERE name =
'$banconom';");
      $w=pg_query($conn, "SELECT id FROM teste.versions WHERE version_name =
'$bancovers';");
      $de=pg_fetch_row($f);
      $di=pg_fetch_row($w);
      pg_query($conn, "INSERT INTO teste.apps_versions (id_app, id_version)
VALUES ($de[0], $di[0])");

foreach ($j["queues"] as $chave){
       preg_match_all("/(.)-(.)-(.)-(.)-(.)-(.)$/U", $chave, $out,
PREG_PATTERN_ORDER);

      if ($out[3][0] != 'none' && $out[3][0] !=''){ // Caso o escalonador ou o
cluster apresente 'none', não amostra valores.
 $bancoplat=$out[0][0];
 $bancoinst=$out[2][0];
    pg_query($conn, "INSERT INTO teste.institution (name) VALUES
('$bancoinst')");
 $bancosh=$out[3][0];
 pg_query($conn, "INSERT INTO teste.scheduler (name) VALUES ('$bancosh')");
 $bancoclu=$out[4][0];
   $b=pg_query($conn, "SELECT id FROM teste.scheduler WHERE name =
'$bancosh';");
 $b=pg_fetch_row($b);
  pg_query($conn, "INSERT INTO teste.clusters (name, scheduler_id) VALUES
('$bancoclu', $b[0])");
$bancogate=$out[5][0];
```

```
 pg_query($conn, "INSERT INTO teste.gateway (name) VALUES ('$bancogate')");
 $bancoque=$out[6][0];
   $a=pg_query($conn,   "SELECT   id   FROM   teste.clusters   WHERE   name   =
'$bancoclu';");
 $a=pg_fetch_row($a);
  pg_query($conn,   "INSERT   INTO   teste.queues   (name,   cluster_id)   VALUES
('$bancoque', $a[0])");
   $c=pg_query($conn,   "SELECT   max(id)   FROM   teste.queues   WHERE   name   =
'$bancoque';");
 $c=pg_fetch_row($c);
  pg_query($conn,   "INSERT   INTO   teste.platforms   (name,   queue_id)   VALUES
('$bancoplat', $c[0])");
   $q=pg_query($conn,   "SELECT   id   FROM   teste.platforms   WHERE   name   =
'$bancoplat';");
  $k=pg_query($conn, "SELECT teste.apps_versions.id FROM teste.apps_versions
LEFT JOIN teste.versions on teste.versions.id = teste.apps_versions.id_version
LEFT      JOIN      teste.applications      on      teste.applications.id      =
teste.apps_versions.id_app  WHERE  teste.versions.version_name  =  '$bancovers'
and teste.applications.name = '$banconom';");
 $qe=pg_fetch_row($q);
 $ki=pg_fetch_row($k);
 pg_query($conn, "INSERT INTO teste.apps_versions_platforms (apps_versions_id,
platform_id) VALUES ($ki[0],$qe[0])");
```

Algoritmo 2: Sistema de mapeamento das aplicações contidas na arquitetura do *Gateway* Bioinfo-Portal e suas inserções no banco de dados Bioinfo.

Atualmente, estão sendo feitas atualizações no código fonte do *gateway* Bioinfo-Portal para a conexão com o banco de dados atualizado, descrito na figura 1. Através da conexão SVN, foi possível baixar as cópias dos códigos e fazer algumas alterações onde ocorrem ligações com banco de dados antigo do portal.

O algoritmo 3, é uma parte específica de um dos códigos, na qual se observa se o trabalho submetido gerou algum erro ou está sendo preparado para a execução no Bioinfo-Portal. No algoritmo 3, em negrito, os status de submissão serão armazenados no banco de dados atualizado junto com a identificação do trabalho.

```
if ($code == 200) {
      $job_id = $result->{'jobId'};
      file_put_contents("$GLOBALS[BIOINFO_HOME]/daemon/running/$bioinfo_id",
      "$job_id::$database_id::$email" );
      unlink ( "$GLOBALS[BIOINFO_HOME]/daemon/checking/" . $bioinfo_id );
      file_put_contents("$GLOBALS[BIOINFO_HOME]/jobs/$bioinfo_id/job_id",
      $job_id);
      pg_query($conn, "UPDATE teste.executions SET status = 'RUNNING',
      id_csgrid = '$job_id' WHERE id_web = '$database_id'");
} else if ($code == 400) {
      file_put_contents  (  "$GLOBALS[BIOINFO_HOME]/daemon/error/$bioinfo_id",
      "$job_id::$database_id::$email" );
```

```
        unlink ( "$GLOBALS[BIOINFO_HOME]/daemon/checking/" . $bioinfo_id );
        pg_query($conn, "UPDATE teste.executions SET status = 'FAILED' WHERE
        id_web = '$database_id'");}
else {
        file_put_contents("$GLOBALS[BIOINFO_HOME]/daemon/pending/$bioinfo_id",
              "$code::$database_id::$email" );
        unlink ( "$GLOBALS[BIOINFO_HOME]/daemon/checking/" . $bioinfo_id );}
```

Algoritmo 3B: Condição *Else,* parte do código de submissão, diz respeito ao trabalho se falhou ou obteve êxito no Bioinfo-Portal.

Então, dados de usuários e execuções depositadas e realizadas no portal, serão armazenadas em suas devidas tabelas no banco de dados centralizado, garantindo uma maior eficiência ao *gateway* em relação a armazenamentos de dados de usuários que utilizam o Bioinfo-Portal.

CONCLUSÕES

A integração de sistemas ao banco de dados centralizado e atualizado permitirá obter melhorias no desempenho relacionadas ao gerenciamento de arquivos, envio de trabalhos ou interfaces de contabilidade. Esse estudo irá sustentar o acesso aos dados para estudos realizados em paralelo no Bioinfo-Portal, análises de predição via aprendizado de máquina e etc.

Como passos próximos estão a alimentação do banco de dados com execuções do Bioinfo-Portal e o desenvolvimento de sistemas para o mapeamento de dados de localização do usuário como IP e País de origem.

REFERÊNCIAS

1. LESK, A. M.. Bioinformatics. **Britannica**, Pennsylvania, Fevereiro 2019.

2. OCAÑA, K.A.C.S., et al. BioinfoPortal: A scientific gateway for integrating bioinformatics applications on the Brazilian national high-performance computing network. **Elsevier**, Rio de Janeiro, v. 107, p. 23, Janeiro 2020.

3. ELMASRI, R., NAVATHE, SHAMKANT B. Sistemas de Banco de Dado. 1. ed. [S.l.]: **Pearson**, 2019.

PROPOSTA DE PROJETO DE INICIAÇÃO CIENTÍFICA

Título do Projeto Proposto

Programação orientada a objeto em um Método Numérico Escalável para o Escoamento Bifásico de Fluidos em Meios Porosos em ambientes Computacionais de Alto Desempenho

Instituição

Laboratório Nacional de Computação Científica

Nome do Aluno

Mariana Aguiar Ribeiro

Nome do Professor

D.Sc. Carla Osthoff Ferreira de Barros (Tecnologista Sênior – SEPAD/COTIC//LNCC, Orientador)

Stiw Harrison Herrera Taipe (Analista de TI- SEPAD;COTIC/LNCC- Coorientador)

Linha de Pesquisa

- Ciências Exatas e da Terra (1.00.00.00-3) – Ciência da Computação (1.03.00.00-7) – Metodologia e Técnicas da Computação (1.03.03.00-6)

Palavras Chaves

Processamento de Alto Desempenho, Dinâmica dos Fluídos, Programação orientada a objetos.

Plano de Trabalho

Dentro de diversas áreas da engenharia e ciências aplicadas, existe um grande interesse no desenvolvimento de modelos matemáticos e métodos computacionais para a simulação de escoamentos em meios porosos. Na engenharia de Petróleo, a otimização dos processos de recuperação de hidrocarbonetos está intimamente relacionada com a simulação. Para isto, diversos fatores devem ser adequados e considerados nos modelos fisico-matemático e numéricos, como por exemplo a troca de massa e momento linear das fases que escoam, suas relações de capilaridade e mobilidade, a estabilidade dos poços de produção e injeção, dentre inúmeros outros [Correa 2013]. Este problema é modelado por um sistema de equações diferenciais parciais, basicamente composto por um subsistema elıptico para a determinação do campo de velocidades e uma equação hiperbólica não linear para o transporte das fases que escoam (equação da saturação). Pretendemos construir aplicações científicas considerando as boas práticas da área de engenharia de software para computação de alto desempenho. Adotaremos processos da engenharia de software com um adequado modelos de ciclo de vida, desde a concepção, elaboração, construção, transição e abordagens ágeis durante a construção do software. O modelo computacional deste estudo apresenta uma metodologia numérica, proposta em [Corrêa], para a simulação do escoamento bifásico (água e óleo) em um reservatório rígido altamente heterogêneo. Do ponto de vista numérico, o modelo propõe a aplicação de um método de elementos finitos localmente conservativo para a velocidade da mistura e um método de volumes finitos não-oscilatório de alta ordem, baseado em esquemas centrais, para a equação hiperbólica não-linear que governa a saturação das fases. Do ponto de vista da engenharia de software este projeto visa desenvolver um ambiente computacional escalável em novas linguagens de programação para o estudo de novos métodos numéricos desenvolvidos pelo grupo de óleo e gás do LNCC no supercomputador Santos Dumont (SDumont, https://sdumont.lncc.br/).

O presente projeto tem como objetivo reestruturar o código atual, que foi desenvolvido na linguagem FORTRAN, para uma estrutura orientada a objetos na linguagem C++, conforme as metodologias apresentadas em [5], para a seguir desenvolver otimizações de paralelismo no código.

Na parte da implementação, o projeto terá como referência, os trabalhos previamente realizados e publicados pelo nosso grupo de pesquisa e colaboradores [7]–[9]. Feita a implementação da solução proposta, o passo seguinte será testar seu desempenho e avaliar suas funcionalidades frente aos benefícios propostos. Este projeto possui seis (6) etapas principais:

- Etapa 1: Revisar bibliografía sobre programação em FORTRAN, programação em C++, programação orientada o objeto
- Etapa 2: Desenvolver a modelagem do Código a ser desenvolvido em C++
- Etapa 3: Desenvolver o código originalmente escrito em FORTRAN para C++
- Etapa 5: Análises de desempenho e escalabilidade do novo código em relação ao código original
- Etapa 6: Elaboração de otimizações de paralelismo no código no ambiente computacional do Supercomputador Santos Dumont..
- Etapa 7: Elaboração de relatório final com descrição dos resultados

Referências

[1] Stiw Herrera, Thiago Teixeira, Weber Ribeiro, André Carneiro, Frederico L. Cabral, Matheus Serpa, Márcio Borges, Carla Osthoff, Sanderson L. Gonzaga de Oliveira, Philippe Navaux. Optimizations in a numerical method code for the biphasic fluid flow in porous media using the SDumont supercomputer.In: XLII Ibero-Latin-American Congress on Computational Methods in Engineering (CILAMCE-2021), 2021, Rio de Janeiro (online). Proceedings of XLII Ibero-Latin-American Congress on Computational Methods in Engineering (CILAMCE-2021), 2021.

[2] Correa, M. & Borges, M., 2013. A semi-discrete central scheme for scalar hyperbolic conservation laws with heterogeneous storage coefficient and its application to porous media flow. International Journal for Numerical Methods in Fluids, vol. 73, n. 3, pp. 205–224.

[3] Taipe, S. H. H., Teixeira, T, Ribeiro, W., Carneiro, A., Cabral, F. L., Gonzaga de Oliveira S.L., Serpa, M. S., Borges, M. R., Esteban Meneses, Osthoff, C., Navaux,P.SCALABLE NUMERICAL METHOD FOR BIPHASIC FLOWS IN HETEROGENEOUS POROUS MEDIA IN HIGH-PERFORMANCE COMPUTATIONAL ENVIRONMENTS. In: The Event for High Performance Computing, Machine Learning and Data analysis (ISC), 2021. The Event for High Performance Computing, Machine Learning and Data analysis (ISC), 2021

[4] Weber Ribeiro, Thiago Teixeira , Frederico L. Cabral, Carla Osthoff, Márcio Borges, Otimização para Ambientes Intel de um Método Numérico para o Escoamento Bifásico de Fluidos em Meios Porosos Através da Eliminação de Barreiras OpenMP. Workshop de Iniciação Científica do Simpósio Brasileiro de Computação de Alto Desempenho 2019(WSCAD-WIC).

[5] Farley, D. (2021). *Modern Software Engineering: Doing What Works to Build Better Software Faster*. Addison-Wesley Professional.

[6] Sommerville, I. (2020). *Engineering software products* (Vol. 355). London: Pearson.

[7] Sommerville, I. (2007). Sommerville: Software Engineering.

[8] Jacobson, I., Lawson, H., Ng, P. W., McMahon, P. E., & Goedicke, M. (2019). The Essentials of Modern Software Engineering. *Association for Computing Machinery*.

[9] Valente, M. T. (2020). Engenharia de software moderna. *Princípios e Práticas para Desenvolvimento de Software com Produtividade*, *1*, 24.

[10] Prikladnicki, R., Willi, R., & Milani, F. (2014). *Métodos ágeis para desenvolvimento de software*. Bookman Editora.

Laboratório Nacional de Computação Científica

Tipo de Bolsa e Período: Iniciação Científica — Dezembro 2022 - Novembro de 2023

Inteligência Artificial Aplicada ao Diagnóstico por Imagem

Bolsista: Matheus Molina Alves Lima

Orientador: Bruno Schulze

Co-Orientador: Fabio Lopes Licht

Petrópolis
2023

1 Objetivo

Este projeto tem como objetivo principal a aplicação de técnicas avançadas de Inteligência Artificial (IA) no diagnóstico médico, utilizando imagens de ressonâncias magnéticas[1], radiografias e/ou tomografias computadorizadas. Além disso, busca-se a integração da IA com a Computação Distribuída de Alto Desempenho, a fim de estabelecer uma infraestrutura computacional adequada para auxiliar na tomada de decisões médicas de forma ágil.

O foco central do projeto é proporcionar uma análise precisa e eficiente das imagens provenientes do banco de imagens, tanto de pacientes com comorbidades quanto de pacientes sem comorbidades. Isso inclui a avaliação de afecções, bem como o cálculo da área e percentual ocupado pelas mesmas.

2 Introdução

Nos dias atuais, vivemos em uma sociedade onde a presença da tecnologia é cada vez mais predominante, especialmente em um contexto pós-pandêmico de COVID-19. Além disso, um tema que tem recebido grande destaque é a Inteligência Artificial (IA), impulsionada pelo surgimento de tecnologias como o ChatGPT e o Google Bard, entre outros.

Nesse contexto, o objetivo deste projeto é desenvolver uma solução baseada em IA para análise de imagens e aplicá-la ao diagnóstico de exames médicos, como ressonâncias magnéticas, radiografias e/ou tomografias computadorizadas. O intuito é otimizar o tempo de diagnóstico médico, permitindo uma avaliação mais ágil e precisa das imagens, por meio do suporte da inteligência artificial e da utilização de uma infraestrutura computacional de alto desempenho.

3 Material e Métodos ou Metodologia

Geralmente, três modelos de exames são usados para obter imagens do corpo humano, entre eles temos:

- Ultrassonografia: técnica dependente de operador, utilizada para reproduzir imagens dos órgãos internos, tecidos, rede vascular e fluxo sanguíneo onde o transdutor é colocado sobre a área a ser examinada, transmitindo ondas sonoras refletidas em imagens
- Ressonância Magnética: técnica que não utiliza radiação e permite retratar imagens de alta definição dos órgãos do corpo humano através de campo magnético, onde as moléculas de hidrogênio do nosso corpo ficam alinhadas
- Tomografia Computadorizada: técnica que utiliza radiação ionizante para registrar imagens internas do corpo humano através de Raio -X

A abordagem convencional para seleção e identificação de áreas de interesse em imagens médicas não invasivas segue um processo que começa com a detecção, que consiste na identificação visual de discrepâncias entre imagens com padrão normal e aquelas com variações. Em seguida, há a etapa de classificação, onde a comparação é feita com base no conhecimento e na literatura existente, a fim de identificar e categorizar as discrepâncias visuais em relação aos fatores clínicos relevantes.[2]
Por fim, temos a segmentação, que envolve a subdivisão das imagens em regiões distintas, podendo ser realizada com base na descontinuidade da intensidade da imagem ou em similaridades nos padrões de textura dos diferentes tecidos.

Os métodos mais utilizados atualmente para a segmentação de imagens, são executados através de IA, mais comumente por meio de ferramentas de "Deep Learning"[3]. Tais ferramentas têm a aptidão de potencializar a capacidade de extração de maiores informações de imagens médicas. Dessa forma, a tomada de decisões demanda menos tempo, ampliando o desenvolvimento de análises nos exames, podendo fazer com que a permanência do paciente no equipamento seja menor, e, ainda assim, de maior qualidade e automatização complementar ao especialista médico, nos processos de detecção, classificação e segmentação.

De forma geral, as etapas que adotaremos no projeto são:

1. Recebimento e armazenamento de blocos de exames específicos conforme o ponto focal de atuação. Em nosso projeto, os dados foram pesquisado na plataforma Kaggle, sendo completamente anonimizados antes de serem de fato utilizados no treinamento de algoritmos de IA. Em alguns

casos, os dados são apenas as imagens dos exames; em outros, as imagens vêm acompanhadas de laudos.

2. Anotação dos dados com o uso de mão-de-obra humana, de acordo com as áreas de interesse do projeto. Isto pode significar emitir laudos a partir das imagens, quando ainda não há laudos; ou identificar e segmentar manualmente estruturas relevantes nas imagens. O produto desta etapa é um conjunto de bases de dados curadas e anonimizada de imagens de exames com respectivas anotações.

3. Desenvolvimento de métodos de IA a partir dos dados curados. Quando cabível, desenvolvimento e aplicação de algoritmos específicos para aumento de dados ("data augmentation")[4] ou para potencializar a definição/qualidade da imagem

4. Avaliação, validação e (quando cabível) quantificação de incerteza dos métodos propostos. Esta etapa envolve tanto testes matemáticos (em que alguma métrica é usada para se auferir a qualidade do método) quanto testes médicos/clínicos (em que especialistas avaliam o resultado do método). É natural que os resultados desta etapa levem a ajustes dos métodos proposto na etapa 3

5. Publicações cientificas e apresentações em congressos de Medicina, IA e áreas relacionadas, para atestar e/ou comprovar a precisão dos métodos desenvolvidos junto à comunidade cientifica

6. Eventuais produtos advindos dos resultados acima.

4 Resultados e Discussão

Para a obtenção dos resultados, foram utilizadas quatro redes neurais. A rede U-Net[5] Simples, executada com um total de 7.771.939 parâmetros; a rede Residual U-Net[6], executada com um total de 11.699.427 parâmetros; a rede Inseption U-Net[7], executada com um total de 48.819.683 parâmetros.; a rede Stacked U-Net[8]; executada com um total de 24.953.350 parâmetros.

Inicialmente havia 210 imagens de tumores malignos em mamas obtidas do portal Kaggle, com suas respectivas máscaras. Para obter mais imagens para treinamento foi utilizada a técnica de Data Augmentation em Imagens realizando rotações de 90^{o}, 180^{o} e 270^{o} nas imagens de ultrassonografia e suas respectivas máscaras. Com isso, foram usadas no total 840 imagens de tumores de mama com suas respectivas máscaras indicando o tumor maligno, sendo 640 para o treinamento das redes e 200 para validação. Foram feitos testes com 100, 200, 500, 1.000 e 10.000 épocas. Entretanto, percebeu-se que não houve ganho significativo acima de 200 épocas nos resultados, pois haviam alguns pontos de variações extremas de treinamento e validação, que causavam saltos, que sendo visualizados nos gráficos distorciam e aumentavam o valor da precisão no eixo das ordenadas e consequentemente a área de plotagem.

5 Conclusões

Os resultados indicaram que a U-Net Simples obteve o melhor desempenho em termos de tempo de execução, enquanto a Inception U-Net apresentou a melhor precisão, com 99% e Interseção média sobre a União (IoU) 98%. A escolha da rede mais adequada varia de acordo com as necessidades específicas do projeto, considerando fatores como tempo de treinamento, precisão e custo-benefício do uso de GPUs. A análise dos resultados sugere que o desempenho das redes está relacionado ao número de hiperparâmetros e à estabilidade da precisão ao longo do treinamento. E importante destacar que a escolha da rede convolucional mais adequada para cada projeto depende de diversos fatores, como o tipo de imagem a ser analisada, a quantidade de dados disponíveis, o tempo disponível para treinamento e a precisão desejada. No entanto, os resultados indicam que a U-Net Simples é uma opção eficiente para a segmentação de imagens médicas de ultrassom. Além disso, a análise comparativa mostrou que o uso de GPUs pode reduzir significativamente o tempo de treinamento das redes, especialmente em casos que envolvem grandes quantidades de dados.

Por fim, é importante ressaltar que a segmentação de imagens médicas com o uso de redes convolucionais e GPUs pode ter um impacto significativo na prática clínica, permitindo uma avaliação mais precisa de doenças e um planejamento mais adequado de tratamentos.

Referências

[1] D. Eun, R. Jang and W.S.e.a. Ha, Deep-learning-based image quality enhancement of compressed sensing magnetic resonance imaging of vessel wall: comparison of self- supervised and unsupervised approaches, *Scientific Reports* (2020).

[2] M.B.H. Moran, G.A. Giraldi, L.F. Bastos and A. Conci, Using super-resolution generative adversarial network models and transfer learning to obtain high resolution digital periapical radiographs., *Comput Biol Med* (2021).

[3] J. Kukačka, V. Golkov and D. Cremers, Regularization for Deep Learning: A Taxonomy, 2017.

[4] C. Shorten and T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data* **6**(60) (2019).

[5] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P.F. Jaeger, S. Kohl and J. Wasserthal, nnU-NET: Self-adapting framework for u-net-based medical image segmentation, *arXiv* (2018).

[6] C. Li, X. Song, H. Zhao, L. Feng, T. Hu, Y. Zhang, J. Jiang, J. Wang, J. Xiang and Y. Sun, n 8-layer residual u-net with deep supervision for segmentation of the left ventricle in cardiac ct angiography., *Computer Methods and Programs in Biomedicine* (2021).

[7] T. Zhang, Y. Xia and Y. Zhang, Inception-u-net: A deep convolutional neural network combined with inception and u-net for medical image segmentation, *Journal of Healthcare Engineering* (2018).

[8] X. Liu, J. Yang, C. Zhang, D. Zhang, W. Luo and Y. Zheng, Multi-modal brain tumor segmentation using stacked unet with integrated spatial constraints, *Neurocomputing* (2020), 266–274.

PROPOSTA DE PROJETO DE INICIAÇÃO CIENTÍFICA

BIOINFORMÁTICA, BANCO DE DADOS E ENGENHARIA DE COMPUTAÇÃO

Título do Projeto Proposto

Implementação de workflows científicos de biologia computacional reprodutíveis e escaláveis de alto desempenho

Instituição

Laboratório Nacional de Computação Científica

Nome do Aluno

Reiglan Soares Di Lourenço

Nome do Professor

D.Sc. Kary Ann del Carmen Ocaña Gautherot (Tecnologista Pleno – LABINFO/LNCC, Orientador)

D.Sc. Carla Osthoff Ferreira de Barros (Tecnologista Sênior – SEPAD/COTIC//LNCC, Coorientador)

D.Sc. Diego Moreira de Araujo Carvalho (Professor Associado – CEFET/RJ, Colaborador – LNCC, Coorientador)

Linha de Pesquisa

- Ciências Exatas e da Terra (1.00.00.00-3) – Ciência da Computação (1.03.00.00-7) – Metodologia e Técnicas da Computação (1.03.03.00-6) – Banco de Dados (1.03.03.03-0)
- Ciências Biológicas (2.00.00.00-6) – Biologia Geral (2.01.00.00-0) – Genética (2.02.00.00-5)

Palavras Chaves

Aprendizado de Máquinas, Bioinformática, Banco de Dados, Processamento de Alto Desempenho

Plano de Trabalho

O presente trabalho se desenvolve a partir de um projeto multidisciplinar que envolve o LABINFO, SEPAD e SDumont no LNCC com o objetivo de acoplar ao portal de Bioinformática Bioinfo-Portal (https://bioinfo.lncc.br/) hospedado no LNCC, um *workflow* científico de sequenciamento RNA de Alto Desempenho. Dessa forma, visamos modelar, executar e analisar *workflows* científicos com tecnologias a ambientes de alto desempenho, tomando como caso de estudo o experimento de bioinformática de sequenciamento RNA.

Em diversas áreas da ciência existem experimentos que podem ser computacionalmente intensivos, como é o caso dos experimentos de bioinformática de sequenciamento RNA (RNA-seq). Esses experimentos de sequenciamento são realizados com uma técnica para a análise da Expressão Diferencial de Genes (EDG) e sequenciamento genômico, ou seja, permite o estudo do comportamento de um conjunto de transcritos de uma célula que apresentam uma dada condição fisiológica ou que estão em fase de desenvolvimento, como o câncer. A tecnologia de RNA-seq demonstra um grande avanço nos estudos da transcriptômica, no entanto, existem ainda muitos desafios relacionados à complexidade, produção e consumo de grande volume de dados e no uso de programas que demandam altos custos computacionais. Tipicamente, podem levar muitas horas ou mesmo dias para que um estudo de caso de RNA-seq seja processado. Sob essa condição, diversas áreas da ciência têm se apoiado a procedimentos e uso de infraestruturas computacionais no decorrer dos anos.

Workflows científicos de bioinformática são complexos e mais ainda se manipulam dados científicos voluminosos, heterogêneos ou multi-distribuídos [1] pelo que precisam de ambientes e tecnologias de

Computação de Alto Desempenho (CAD) para execução, gerência, acesso e armazenamento de dados científicos e proveniência [2]. Os dados científicos resultantes de um determinado experimento científico [3] - do seu ciclo de vida, metadados de desempenho do workflow, atividades e programas - são diretrizes do bom desempenho de um *workflow*. Sistema de Gerência de *Workflows* Científicos (SGWfC) ou mesmo determinadas linguagens de programação são capazes de orquestrar a execução de workflows e até acoplar *workflows* computacionalmente intensivos em ambientes de CAD. Nesse âmbito existem SGWfC baseados em sistema web, como o Galaxy, e pacotes estatísticos do R e Bioconductor, como DESeq2 e EdgeR, que são usados nos estudos de EDG. Para automação das tarefas em ambientes distribuídos, existem SGWfC como o Kepler e Pegasus, além de linguagens de programação específica para ambientes distribuídos e computação científica, como o Swift e suas variantes (Swift/T e Swift/K) e Python, com a biblioteca Parsl.

O desenvolvimento deste trabalho é uma continuação da proposta de [7-9] para a otimização de desempenho do *workflow* desenvolvido, chamado ParslRNA-Seq. O que levou as execuções a alcançarem um ganho em tempo computacional maior do que 65% em relação a versão apresentada em [8]. Da versão do ParslRNA-Seq apresentada em [9], o workflow teve melhoras no Tempo Total de Execução (TTE) de, aproximadamente, 3 dias para 24 minutos dentro de um ambiente distribuído de alto desempenho. A seguinte fase da pesquisa se baseia em uma análise das operações de E/S que além de sugerir um novo formato de execução, visa outra melhora significativa no TTE do workflow.

Portanto, o presente trabalho apresenta um estudo computacional intensivo executado com a nova proposta de modelagem do *workflow*, feita sob uma análise comparativa baseada no ganho computacional das operações de E/S das atividades mais custosas do workflow científico ParslRNA-Seq. Com essa nova proposta, o workflow desenvolvido agora explora os dispositivos de armazenamento disponíveis para realização de operações de escrita e leitura. Os estudos serão realizados com os recursos computacionais disponibilizados através dos nós de base do supercomputador Santos Dumont (SDumont, https://sdumont.lncc.br/) e se baseiam em dois formatos de execução do ParslRNA-Seq: uma delas utilizando o Hard Disk Drive (HDD), ou seja, o Lustre; e, a segunda, o Solid State Drive (SSD).

Desta forma, o objetivo deste projeto é propor uma série de soluções para a gerência de simulações computacionais de workflows de transcriptômica para ambientes de alto desempenho. Espera-se conseguir uma maior escalabilidade usando técnicas de CAD, de paralelização, de distribuição de tarefas e granularidade considerando o uso de proveniência como base para as soluções propostas. Ainda dentre outros desafios considerados neste trabalho estão os estudos para executar múltiplos experimentos com grandes quantidades de dados de RNA-Seq em paralelo no SDumont. Pensamos também em propor uma biblioteca de Apps de bioinformática já pré-configurada para Parsl (similar aos módulos nf-core, BioContainers) e enriquecer metadados associados ao workflow (p.ex. ontologia para descrever tipos de entrada e saída de cada atividade) na direção de princípios FAIR.

Na parte da implementação, o projeto terá como referência, os trabalhos previamente realizados e publicados pelo nosso grupo de pesquisa e colaboradores [7-9]. Feita a implementação da solução proposta, o passo seguinte será testar seu desempenho e avaliar suas funcionalidades frente aos benefícios propostos. Este projeto possui seis (6) etapas principais:

- Etapa 1: Revisar bibliografía sobre bioinformática, análises de transcriptômicas, *workflows* científicos, Parsl, ParslRNA-Seq e bibliotecas no SDumont;
- Etapa 2: Explorar tecnologias que explorem paralelismo e distribuição de tarefas em *workflows* científicos;
- Etapa 3: Implementar o esquema de execução e gerência de tarefas e dados para *workflows* de análises de transcriptômicas;
- Etapa 4: Explorar o ambiente de CAD do SDumont;
- Etapa 5: Análises de desempenho e escalabilidade das ferramentas propostas;
- Etapa 6: Elaboração de relatório final com descrição dos resultados.

Referências

[1] V. Marx, “Biology: The big challenges of big data,” *Nature*, vol. 498, no. 7453, pp. 255–260, Jun. 2013.

[2] J. Freire, D. Koop, and L. Moreau, Eds., *Provenance and Annotation of Data and Processes*, vol. 5272. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
[3] M. Mattoso *et al.*, “Towards supporting the life cycle of large scale scientific experiments,” *International Journal of Business Process Integration and Management*, vol. 5, no. 1, pp. 79–92, 2010.
[4] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2. ed., At 7. printing. New York, NY: Springer, 2013.
[5] G. Da San Martino and A. Sperduti, “Mining Structured Data,” *IEEE Computational Intelligence Magazine*, vol. 5, no. 1, pp. 42–49, Feb. 2010.
[6] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, “Accomplishments and challenges in literature data mining for biology,” *Bioinformatics*, vol. 18, no. 12, pp. 1553–1561, Dec. 2002.
[7] Cruz, L.; Coelho, M.; Terra, R.S.; Carvalho, D.; Gadelha, L.M.R.; Osthoff, C.; Ocaña, K.A.C.S. Workflows Científicos de RNA-Seq em Ambientes Distribuídos de Alto Desempenho: Otimização de Desempenho e Análises de Dados de Expressão Diferencial de Genes. In: Brazilian e-Science Workshop (BreSci 2021), 2021, Florianópolis, Santa Catarina. Anais do XV Brazilian e-Science Workshop. 2021.
[8] Cruz, L.; Coelho, M.; Gadelha, L.M.R.; Ocaña, K.A.C.S.; Osthoff, C. Avaliação de Desempenho de um Workflow Científico para Experimentos de RNA-Seq no Supercomputador Santos Dumont. In: Workshop de Iniciação Científica em Arquitetura de Computadores e Computação de Alto Desempenho (WSCAD 2020 - WIC), 2020. Anais do Workshop de Iniciação Científica em Arquitetura de Computadores e Computação de Alto Desempenho, 2020.
[9] Ocaña, K.; Cruz, L.; Galheigo, M.; Coelho, M.; Carneiro, A.; Terra, R.; Gadelha, L.; Carvalho, D.; Boito, F.; Navaux, P.; Osthoff, C. ParslRNA-Seq: A scalable, efficient, and high-throughput RNAseq analysis workflow in supercomputers. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2023.

Modelagem de vigas de Euler Bernoulli e Timoshenko

Equações Diferenciais Parciais

Bolsista: Tarsiane Ribeiro da Costa

Orientador: Jaime Ediberto Munoz Rivera

Bolsista de iniciação cientifica – CNPq

Período do desenvolvimento: 01/09/22 à 01/08/2023

2022

I. Objetivo

O Projeto consiste em estudar o comportamento de tais estruturas experimentadas a diferentes cargas e condições. Muitas estruturas da engenharia civil tais como pontes são modeladas através de uma aprovação de vigas. Devido a problemas operacionais e considerações de segurança e desempenho os engenheiros utilizam vigas interconectadas uma com as outras para construir tais estruturas.

Nosso objetivo é desenvolver métodos numéricos (diferenças e elementos finitos) :

- Desenvolver uma equação de diferenças, levando em consideração as características do sistema em estudo.

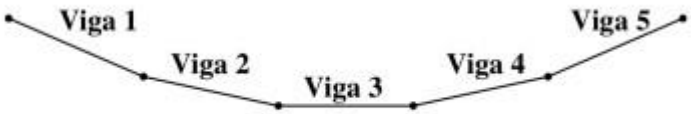
- Aplicar o método de diferenças finitas para discretizar a equação desenvolvida, permitindo uma abordagem numérica para a resolução do problema.

- Obter uma solução aproximada para o sistema por meio da implementação do método de diferenças finitas, considerando as condições iniciais e de contorno adequadas.

- Comparar a solução aproximada obtida com a solução exata do problema, analisando a precisão e o desempenho do método de diferenças finitas na modelagem do sistema em estudo.

II. Introdução

Muitas estruturas da Engenharia Civil, tais como pontes longas são modelados através de uma equação de vigas. Devidos a problemas operacionais e considerações de segurança e performance os engenheiros utilizam vigas interconectadas umas com outras para construir tais estruturas. Estas interconexões produzem os chamados problemas de transmissão.



Viga serialmente inteligada

O primeiro problema consiste em modelar as diferentes formas de interconexão entre as vigas. Consideraremos num primer momento o modelo estacionário, tanto para os modelos de vigas de Timoshenko como o modelo de Euler Bernoulli. Posteriormente consideraremos os correspondentes problemas de evolução primeiro no caso conservativo. Um segundo problema consiste em introduzir algum tipo de mecanismo

externos que possibilite a estabilização da estrutura . Estes mecanismos podem ser através de controles colocados entre as Inter conexões da viga, ou também devido as leis constitutivas de cada componente, por exemplo uma componente pode ser viscosa, outra pode ser termo elástica, outra pode ser um material com memória . Mostraremos que a variação de temperatura numa viga produz efeitos mecânicos (dilatação e contração) que são capazes de estabilizar uniformemente a estrutura.

III. Material e Métodos ou Metodologia

A metodologia adotada neste estudo visa estimar o erro decorrente da discrepância entre a solução numérica e a solução exata de um problema modelado numericamente. A discretização de problemas é uma prática amplamente empregada para simplificar a representação matemática e permitir a aplicação de métodos numéricos. No entanto, ao discretizar um problema, é possível introduzir erros que resultam em soluções aproximadas. Nesse contexto, é crucial investigar e estimar o erro associado à solução discreta, a fim de avaliar a precisão e confiabilidade do método utilizado. A compreensão da magnitude e da natureza desse erro fornecerá insights valiosos sobre a fidelidade das soluções discretas em relação à solução exata. Além disso, possibilitará a identificação de estratégias para mitigar e reduzir o erro, aprimorando a precisão das soluções numéricas obtidas.

IV. Resultado e Discussão

O objetivo deste estudo é realizar a aplicação de um método numérico para resolver uma equação de quarta ordem e, assim, obter uma compreensão mais profunda do comportamento desse sistema. Espera-se como resultado a obtenção do erro entre a

solução numérica e a solução exata, com o intuito de avaliar a precisão do método utilizado. Especificamente, pretende-se analisar a ordem do erro em relação ao parâmetro de discretização, denominado "h". Além disso, busca-se verificar se a discretização adotada pode ser considerada eficiente, considerando que a aprovação exata é válida para coeficientes constantes.

Essa análise mais aprofundada do comportamento da tolerância de quarta ordem e do erro numérico associado permitirá uma melhor compreensão das características do método utilizado, sua precisão e eficiência para resolver problemas dessa natureza.

V. Conclusões

Entre as conclusões do problema que estamos estudando podemos citar:

1) As diferenças centrais usadas nas aproximações produzem um erro estimado pelo quadrado das aproximações.

2) O método das diferenças finitas é eficiente para encontrar soluções aproximadas de segunda ordem.

3) As matrizes dos coeficientes no caso unidimensional são matrizes de 5 diagonais.

4) Estudo da Equação de quarta ordem.

VI. Referências Bibliográficas

[1]. M. Renardy, R.C. Rogers;
An Introduction to partial differential equations, Sringer-Verlag (1993) Vol. 1.

***[2]** R. V. Churchill;*
Fourier Series and boundary value problems, McGraw-Hill, New York (1988) Vol. I.

[3] J.E. Munoz Rivera; Introdução às Equações Diferenciais Parciais, LNCC Vol. 1, pages 1-230, (2004). ISBN 978-85-9996108-7

[4] J.E. Munoz Rivera; Teoria das distribuições e EDP, LNCC Vol. 1, pages 1- 244, (2004). ISBN 85-99961-06-3

Técnicas de Ciência de Dados Aplicadas a Pesquisas de Dados em Larga Escala

Bolsista: Thiago Dutra da Silva

Orientador: Fábio Porto

Bolsa de Iniciação Cientıfica, no perıodo de 01/02/2023 a 31/08/2023

1. Objetivos

O objetivo deste trabalho é prover dados meteorológicos para aprendizado de máquinas por Inteligência artificial (IA) com objetivo de previsões meteorológicas de altos riscos. Usando diversos scripts para particularidades de cada fonte de dados para criar um modelo de aprendizado que consiga captar condições meteorológicas extremas com mínimos erros.

2. Introdução

A evolução do Big Data e da Internet das Coisas (IoT) tem trazido desafios e oportunidades tanto para a indústria quanto para a ciência. Com o aumento exponencial no volume de dados gerados, torna-se crucial gerenciá-los, processá-los e analisá-los em tempo real. Nesse contexto, sistemas especializados desempenham um papel fundamental em diversas áreas. No entanto, é importante destacar que o desenvolvimento de sistemas eficientes capazes de lidar com dados semiestruturados, dinâmicos e de grande escala apresenta inúmeros desafios.

Para superar essas dificuldades, o uso de algoritmos de aprendizado de máquina e técnicas de inteligência artificial. Atualmente, as soluções existentes enfrentam dificuldades em atender plenamente a esses requisitos complexos para dados meteorológicos visto na vasta quantidade de dados e sua variedade, indo desde dados pluviométricos à bóias marítimas que atualmente ajudam na previsão meteorológica. A natureza dos dados, muitas vezes desorganizados e em constante mudança acabam por dificultar a criação de uma base para o aprendizado, visto que o clima da terra está em constante mudança. Além disso, a necessidade de processamento em tempo real para análise e tomada de decisões impõe desafios adicionais.

Com esses problemas em mente, optamos por desenvolver uma base de dados antigos que satisfaça a demanda pela IA de forma em que suas predições sejam o mais precisas e rápidas possíveis, tendo foco em diminuir danos de condições naturais extremas.

3. Metodologia

Com o objetivo em vista e utilizando o python, uma linguagem de programação com uma alta maleabilidade, faremos uma extração dos dados específica para as diversas fontes disponíveis para o projeto e armazenamos esses dados no data lake, um repositório centralizado que permite armazenar todos os seus dados estruturados e não estruturados em qualquer escala facilitando assim o tratamento e acesso a esses dados.

Neste relatório, destacamos os dados pluviométricos e meteorológicos disponibilizados pelo Sistema Alerta Rio da Prefeitura do Rio de Janeiro. Utilizando a biblioteca Selenium em Python, realizamos a extração desses dados diretamente do site do Alerta Rio e os armazenamos em um data lake. Isso permite um processamento posterior e possibilita o treinamento de modelos de Inteligência Artificial (IA).

Através do Sistema Alerta Rio, temos acesso a informações valiosas sobre as condições climáticas e os níveis de chuva na região do Rio de Janeiro. A biblioteca Selenium nos permite automatizar o processo de coleta de dados, acessando diretamente o site do Alerta Rio. Dessa forma, podemos obter os dados atualizados de forma eficiente e confiável. Em seguida, esses dados são armazenados em um data lake, facilitando o gerenciamento e a integração com outras fontes de dados.

Essa abordagem de coleta e armazenamento dos dados pluviométricos e meteorológicos proporciona uma base sólida para a análise posterior. Ao treinar modelos de IA com esses dados, podemos desenvolver algoritmos capazes de fazer previsões e análises mais precisas.

Em resumo, a utilização da biblioteca Selenium para extrair os dados do Sistema Alerta Rio e o armazenamento em um data lake proporcionam uma maneira eficiente e confiável de coletar e processar informações pluviométricas e meteorológicas. Esses dados têm o potencial de contribuir para uma melhor compreensão do clima na região do Rio de Janeiro e oferecer insights valiosos para diversas aplicações futuras.

4. Resultados e discussão

Com os dados pluviométricos e meteorológicos extraídos pelo Sistema Alerta Rio e armazenados no data lake, o próximo passo seria o tratamento e a utilização dessas informações. No entanto, é importante ressaltar que, atualmente, os resultados obtidos são limitados devido a algumas dificuldades enfrentadas ao longo do processo.

Um dos desafios enfrentados diz respeito ao versionamento de software. À medida que novas versões do sistema são lançadas, pode haver mudanças na estrutura e na forma como os dados são disponibilizados. Isso pode exigir ajustes no código de extração e processamento dos dados, tornando necessário acompanhar de perto essas atualizações e garantir a compatibilidade. No momento, o trabalho está em andamento para superar essas dificuldades e melhorar a utilização dos dados.

5. Conclusão

Com os objetivos e limitações do projeto, conclui-se que precisaria de mais tempo para os objetivos serem alcançados e com as dificuldades e variedades das fontes de dados, uma parametrização seja montada.

Referências

Sistema Alerta Rio da Prefeitura do Rio de Janeiro. Disponível em: http://alertario.rio.rj.gov.br Acesso em: 01 de julho 2023.

RELATÓRIO DE ATIVIDADES - BOLSA PIBITI

1. Dados Gerais

Projeto: Estudo e implementação de Sistema de Banco de Dados para Análise em Saúde Coletiva

Bolsista: Alan de Souza Mello

Orientador: José Karam Filho

Coorientador: Paulo Cabral Filho

2. Objetivo

Este projeto tem como objetivo ajudar na implementação da tese de doutorado do coorientador (em fase de defesa) no que tange a disseminação de informações em saúde com base nos sistemas oficiais do SUS, sobre internações por câncer no Brasil desde 2011.

Para isso a necessidade de estudar sistemas de gerenciamento de banco de dados relacionais, tais como MYSQL e Postgree, as técnicas disponíveis para conversão de formatos, geração de banco de dados, análise dos sistemas, desenvolvimento e implantação de um Sistema Piloto de Monitoramento dos dados extraídos do DataSus.

3. Introdução

Tem sido observada a necessidade de uma plataforma computacional flexível orientada a Web que permita a exportação de dados armazenados no DataSus para outros sistemas de análises, tais como Excel, Tableau ou Power BI (ferramentas de Business Intelligence) ou para um Sistema piloto a ser desenvolvido, onde será possível criar um Painel de Indicadores sobre diversas informações baseadas nos dados do SUS, de modo a deixá-los disponíveis via internet para a comunidade acadêmica e equipes de gestão dos governos Municipais, Estaduais e Federal.

4. Material e Métodos ou Metodologia

Inicialmente foi disponibilizada uma máquina virtual no LNCC, denominada Pantanal, com o sistema operacional Linux (Ubuntu) para criação do Sistema, com os dados extraídos do DATASUS.

Primeiramente, foi realizada uma pesquisa dos principais SGBDs (Sistema de Gerenciamento de Banco de Dados), para se decidir qual seria o melhor para o projeto.

Foi concluído que o MYSQL seria a melhor opção, tendo em vista ser Gratuito e de código aberto; sua ampla compatibilidade com uma variedade de sistemas operacionais, incluindo Windows, Linux e macOS; também pelo seu desempenho, já que é projetado para lidar com altas cargas de trabalho e é capaz de lidar com grandes quantidades de dados e transações com baixa latência; sua escalabilidade, tanto horizontal, como verticalmente, o que significa que é possível adicionar mais recursos ou máquinas para lidar com cargas de trabalho crescentes; por oferecer várias opções de segurança, incluindo autenticação de usuário, criptografia de dados e controle de acesso a dados; e também por ter uma ampla comunidade de usuários e desenvolvedores, o que significa que há muitos recursos e ferramentas disponíveis, incluindo documentação, fóruns de suporte e plug-ins.

Como os dados que serão transferidos para o banco de dados estão atualmente em arquivos .csv, foi necessário um estudo de como fazer essa conversão de formatos e transferência dos dados para uma tabela no banco. E por se tratar de um grande volume de dados e inúmeros arquivos, isso foi feito através de scripts para automatizar o processo de criação de banco de dados e tabelas, e também para as transferências dos dados.

Realizou-se testes de transferências dos dados dos arquivos ‘.csv’ para tabelas no MYSQL, e também foi testado alguns relacionamento de tabelas, pois existem inúmeros campos da tabela principal que estão vinculados a outras tabelas associativas.

Cabe ressaltar que tanto os dados das tabelas associativas, como os dados de cada Estado, estão em arquivos ‘.csv’ distintos, será necessário repetir o processo algumas vezes, razao pela qual a demanda por criação de Scripts.

Foi adicionada uma nova coluna chamada 'doença' para permitir a inclusão de dados de outras doenças além do câncer.Essa adição permite a inclusão de dados de outras doenças além do câncer no sistema, proporcionando maior flexibilidade e expansibilidade aos dados coletados. A coluna 'doença' está sendo utilizada atualmente para armazenar os dados relacionados ao câncer, conforme previamente estabelecido. Essa nova estrutura permitirá a inclusão de outras doenças no futuro, se necessário, sem comprometer a funcionalidade do banco de dados.

Foi realizada a documentação detalhada dos 61 arquivos de scripts desenvolvidos para o projeto. Cada arquivo foi devidamente comentado, descrevendo sua função e a forma correta de executá-lo. Além dos comentários nos scripts individuais, foi criado um ‘ReadMe’ abrangente que fornece instruções claras sobre a ordem de execução dos scripts e como lidar com possíveis erros que possam ocorrer durante o processo.

Foi realizado um estudo da linguagem SQL no contexto do MySQL para tratar a formatação de valores monetários. Observou-se que o MySQL utiliza o padrão de valores americanos, com o uso do ponto (.) ao invés da vírgula (,) para separar casas decimais em valores monetários. Com base nessa observação, foi identificada a necessidade de ajustar a formatação dos valores monetários para o padrão brasileiro (utilizando a vírgula como separador decimal). Para isso, utilizou-se a função FORMAT no SQL. Apesar do MySQL não oferecer o padrão 'pt_BR' para formatação, optou-se por utilizar o padrão 'pt_PT' (Portugal), pois obteve o mesmo resultado esperado no tratamento dos valores monetários.

Identificou-se a necessidade de ter os valores do banco de dados atualizados com base no IPCA (Índice Nacional de Preços ao Consumidor Amplo) para realizar análises futuras considerando a variação inflacionária. Para isso, foi criada a tabela ‘IPCA’ com as colunas 'ano', 'mes' e 'indice’. Em seguida, foi preparado um arquivo CSV contendo os índices do IPCA a partir de janeiro de 2012 até abril de 2023. Os dados do arquivo CSV foram transferidos para a tabela IPCA utilizando o comando SQL "LOAD DATA INFILE”. Essa operação permitiu que os índices de inflação ficassem disponíveis no banco de dados para serem utilizados nas análises posteriores. Para visualizar os valores atualizados com base no IPCA, foi desenvolvido um comando SQL e esse comando realiza a junção entre as tabelas

ATENDIMENTO e IPCA com base no ano e mês correspondentes, e calcula o novo valor ajustado pelo IPCA.

É importante destacar que essa estrutura de dados permite a criação de Views no contexto necessário para análises futuras. Views são objetos virtuais que podem ser criados a partir de uma ou mais tabelas existentes, oferecendo uma representação personalizada dos dados. Elas podem ser utilizadas para simplificar consultas complexas e fornecer resultados pré-processados ou agregados para análises específicas.

5. Resultados e Discussão

Após a conclusão das etapas de pesquisa, estudo e implementação, tenho o prazer de informar que o banco de dados foi criado com sucesso. Foram transferidos aproximadamente 5 milhões de dados do DataSUS referentes a internações por câncer no Brasil, no período de 2012 a 2021.

A criação do banco de dados envolveu a execução de scripts desenvolvidos especificamente para esse propósito. Esses scripts automatizaram o processo de criação das tabelas, transferência dos dados dos arquivos CSV para as tabelas correspondentes e estabelecimento dos relacionamentos necessários.

Ao longo desse processo, foram realizados testes para garantir a integridade e consistência dos dados. Foram verificadas as chaves primárias, as restrições de integridade referencial e a correta correspondência dos dados transferidos. Todos os testes foram bem-sucedidos, o que confirma a eficácia do sistema de banco de dados implementado.

Com o banco de dados totalmente criado e ‘populado’ com os dados de internações por câncer, torna-se possível realizar análises e extrair informações relevantes para a área de saúde coletiva. A disponibilidade desses dados em um formato estruturado e acessível possibilita a realização de estudos epidemiológicos, identificação de tendências, análise de fatores de risco, entre outros.

Agora, com base nesse banco de dados, acredito que poderemos avançar para a etapa de análise exploratória dos dados. Assim sendo, será possível realizar consultas SQL para extrair informações específicas, criar visualizações gráficas e desenvolver indicadores relevantes para a área de saúde.

Além disso, o desenvolvimento de um Painel de Indicadores se torna viável a partir dos dados disponíveis no banco de dados. Esse painel poderá ser acessado via web, permitindo que a comunidade acadêmica, equipes de gestão dos governos municipais, estaduais e federal tenham acesso fácil e rápido às informações sobre internações por câncer no Brasil.

É importante ressaltar que o banco de dados implementado não se limita apenas aos dados de internações por câncer. Com a inclusão da nova coluna 'doença', há a possibilidade de expansão para incluir dados de outras doenças, conforme mencionado anteriormente. Essa flexibilidade do sistema permite a realização de análises comparativas entre diferentes doenças e a geração de informações abrangentes sobre a saúde coletiva.

6. Conclusões

O projeto de estudo e implementação do Sistema de Banco de Dados para Análise em Saúde Coletiva obteve sucesso em sua etapa de criação do banco de dados. Foram transferidos aproximadamente 5 milhões de registros de internações por câncer no Brasil, no período de 2012 a 2021, para o banco de dados desenvolvido.

A utilização do MySQL como Sistema de Gerenciamento de Banco de Dados mostrou-se adequada, oferecendo recursos, desempenho e segurança necessários para o projeto. Através da automação por meio de scripts, foi possível agilizar o processo de criação das tabelas e transferência dos dados, garantindo a integridade e consistência dos mesmos.

Com o banco de dados criado, abre-se um vasto leque de possibilidades para análises e estudos no campo da saúde coletiva. Através de consultas SQL, é possível extrair informações relevantes, identificar padrões e tendências, desenvolver indicadores e criar um Painel de Indicadores acessível via web.

Agradeço ao meu orientador, José Karam Filho, e ao meu coorientador, Paulo Cabral Filho, pela orientação e suporte ao longo do projeto. Também agradeço à equipe do Laboratório Nacional de Computação Científica (LNCC) pela

disponibilização da máquina virtual e dos recursos necessários para o desenvolvimento do sistema.

Espero que o sistema de banco de dados implementado possa contribuir para a disseminação de informações em saúde, auxiliando no avanço da área de saúde coletiva e no aprimoramento das políticas públicas relacionadas ao câncer e outras doenças.

7. Referências bibliográficas

- w3schools.com. (s.d.). SQL Tutorial. Recuperado de https://www.w3schools.com/sql/ ;
- Shell Scripting Tutorial.Bóson Treinamentos (2018, setembro 6). Shell Scripting - Introdução e Conceitos Básicos [Vídeos]. Recuperado de https://www.youtube.com/watch?v=EOLPUc6oo-w&list=PLucm8g_ezqNrYgjXC8_CgbvHbvI7dDfhs ;
- Nield, T. (2016). Introdução à Linguagem SQL: Abordagem Prática Para Iniciantes. São Paulo: Novatec Editora;
- MySQL. (2023). MySQL 8.0 Reference Manual. Recuperado de https://dev.mysql.com/doc/refman/8.0/en/;
- Jargas, Aurelio Marinho (2008). Shell Script Profissional. São Paulo: Novatec Editora.

PROGRAMA DE BOLSAS PIBITI/LNCC

RELATÓRIO DE ATIVIDADES

1) Dados Gerais

Título do projeto: Uso do padrão de paralelismo de linguagem em esquema de radiação da atmosfera
Bolsista: Gabriel Thomaz do Nascimento
Orientadores: Roberto Pinto Souto, Eduardo Lucio Mendes Garcia
Tipo de bolsa: PIBITI/CNPq
Período do Relatório: Junho 2023
Início do período de bolsa: 01/06/2023

2) Objetivos

Objetivos abordados durante o período da pesquisa:

- Assimilar os conceitos de programação paralela, e dos seus diferentes paradigmas.

3) Introdução

A execução paralela de aplicações científicas pode ser obtida por meio do uso de bibliotecas específicas para esta finalidade, tais como *pthread* [1]e OpenMP [2], para aplicação em máquina de memória compartilhada contendo CPU com múltiplos núcleos computacionais (*multicore*), ou como a biblioteca MPI [3] em equipamentos de memória distribuída, por exemplo. Outra possível abordagem é a adoção de execução massivamente paralela em dispositivos aceleradores com GPU (*manycore*), em conjunto com a execução paralela multicore, configurando-se como um processamento heterogêneo, no qual combina-se a utilização de distintas arquiteturas para obtenção de melhor desempenho.

Um outra opção para alcançar a paralelização de código, é através da utilização de instruções paralelas implementadas no próprio padrão da linguagem, como por exemplo em Fortran ocorre com DO CONCURRENT, definido inicialmente na especificação Fortran 2008 da linguagem [4], estipulando que as iterações de um laço não possuem dependência entre si. O compilador Fortran da NVIDIA (**nvfortran**), disponível no NVIDIA HPC Software Development Kit (SDK) [5], provê suporte ao uso de DO CONCURRENT, permitindo o direcionamento da execução paralela das iterações de um laço em CPU multicore ou em GPU [6].

4) Metodologia

Foi gerado notebook jupyter, hospedado no Google Colab, para testar o DO CONCURRENT em uma aplicação (método Jacobi):

https://colab.research.google.com/github/TempoHPC/FortranSP/blob/main/notebook/english/DoConcurrent_Examples.ipynb

5) Resultados e Discussão

A utilização do DO CONCURRENT baseou-se numa aplicação relacionada a distribuição bidimensional de calor a partir da borda de uma placa quadrada, com a resolução sendo dada pelo método iterativo de Jacobi. Com a criação de uma matriz bidimensional, os valores de temperatura são armazenados em cada ponto e atualizados a cada processo de iteração.

Realizada a instalação do **NVIDIA HPC Software Development Kit** em sequência da compilação pelo **nvfortran**, 2 casos de teste foram enfatizados: em GPU e em CPU multicore:

- **GPU**

```
[ ]  %%bash
     su fortransp
     source /usr/share/modules/init/bash
     module use /opt/nvidia/hpc_sdk/modulefiles
     module load nvhpc/22.11

     cd
     cd examples/stdpar/jacobi
     nvfortran -stdpar -Minfo=accel -fast jacobi.f90 -o jacobi_gpu
```

Junto a compilação utilizaram-se os seguintes parâmetros:

- stdpar: aciona o paralelismo padrão direcionado à GPU.
- Minfo: flag que explicita o processo de paralelização do DO CONCURRENT.

```
%%bash
su fortransp
source /usr/share/modules/init/bash
module use /opt/nvidia/hpc_sdk/modulefiles
module load nvhpc/22.11

cd
cd examples/stdpar/jacobi
ls
./jacobi_gpu
```

```
jacobi.f90
jacobi_gpu
sm.mod
        16823  microseconds on parallel with do concurrent
        86812  microseconds on sequential
 Test PASSED
```

A aplicação processada paralelamente em uma GPU **NVIDIA Tesla T4** demonstrou uma maior eficiência em tempo de execução em comparação a sua execução serial.

- **CPU**

```
[ ] %%bash
    su fortransp
    source /usr/share/modules/init/bash
    module use /opt/nvidia/hpc_sdk/modulefiles
    module load nvhpc/22.11

    cd
    cd examples/stdpar/jacobi
    nvfortran -stdpar=multicore -Minfo=accel jacobi.f90 -o jacobi_multicore
```

Junto a compilação utilizaram-se os seguintes parâmetros:

- stdpar=multicore: aciona o paralelismo padrão multicore direcionado à CPU.
- Minfo: flag que explicita o processo de paralelização do DO CONCURRENT.

```
%%bash
su fortransp
source /usr/share/modules/init/bash
module use /opt/nvidia/hpc_sdk/modulefiles
module load nvhpc/22.11

cd
cd examples/stdpar/jacobi
ls
./jacobi_multicore
```

```
jacobi.f90
jacobi_gpu
jacobi_multicore
sm.mod
      4534803  microseconds on parallel with do concurrent
      4492391  microseconds on sequential
 Test PASSED
```

A aplicação processada demonstrou resultados parecidos tanto para com a execução paralela multicore em CPU quanto para a execução serial. Isso se deve aos recursos alocados pelo jupyter notebook.

- **Comparação Geral**

Junto à análise dos tempos de execução de processamento paralelo em GPU e em CPU multicore pode-se perceber um ganho geral de eficiência ao executar determinada aplicação em GPU, sobretudo quando processado paralelamente.

6) Conclusões

A análise dos dados obtidos junto às informações obtidas através de material didático esclarece o processo de funcionamento de uma aplicação executada paralelamente em diversas vertentes. Averigua-se a eficiência do processo quanto comparado ao processo serial (dadas as devidas condições) e confirma o suporte efetivo ao processamento paralelo através de uma linguagem padrão de programação (Fortran), em comparação a outras bibliotecas específicas. Enfatiza-se também através dessa análise o trabalho interdisciplinar como um grande integrador a compreensão dos assuntos abordados pela pesquisa em procedência com a resolução de suas respectivas problemáticas.

7) Referências Bibliográficas

[1] Nichols, Bradford, Dick Buttlar, and Jacqueline Farrell. Pthreads programming: A POSIX standard for better multiprocessing. " O'Reilly Media, Inc.", 1996.
[2] Chapman, Barbara, Gabriele Jost, and Ruud Van Der Pas. Using OpenMP: portable shared memory parallel programming. MIT press, 2007.

[3] Gropp, William, et al. Using MPI: portable parallel programming with the message-passing interface. Vol. 1. MIT press, 1999.
[4] Fortran 2008, ISO/IEC 1539:2010 standard, https://j3-fortran.org/doc/year/10/10-007r1.pdf
[5] NVIDIA HPC Software Development Kit (SDK) https://developer.nvidia.com/hpc-sdk
[6] Miko Stulajter, Ronald M. Caplan, Jeff Larkin and Jon Linker; Using Fortran Standard Parallel Programming for GPU Acceleration https://developer.nvidia.com/blog/using-fortran-standard-parallel-programming-for-gpu-acceleration/
[7] Guray Ozen, Graham Lopez; Accelerating Fortran DO CONCURRENT with GPUs and the NVIDIA HPC SDK https://developer.nvidia.com/blog/accelerating-fortran-do-concurrent-with-gpus-and-the-nvidia-hpc-sdk/